Error en la investigación Oncológica:

Error sistemático y error aleatorio

Juan José Yepes Nuñez. MD, MSc, PhD

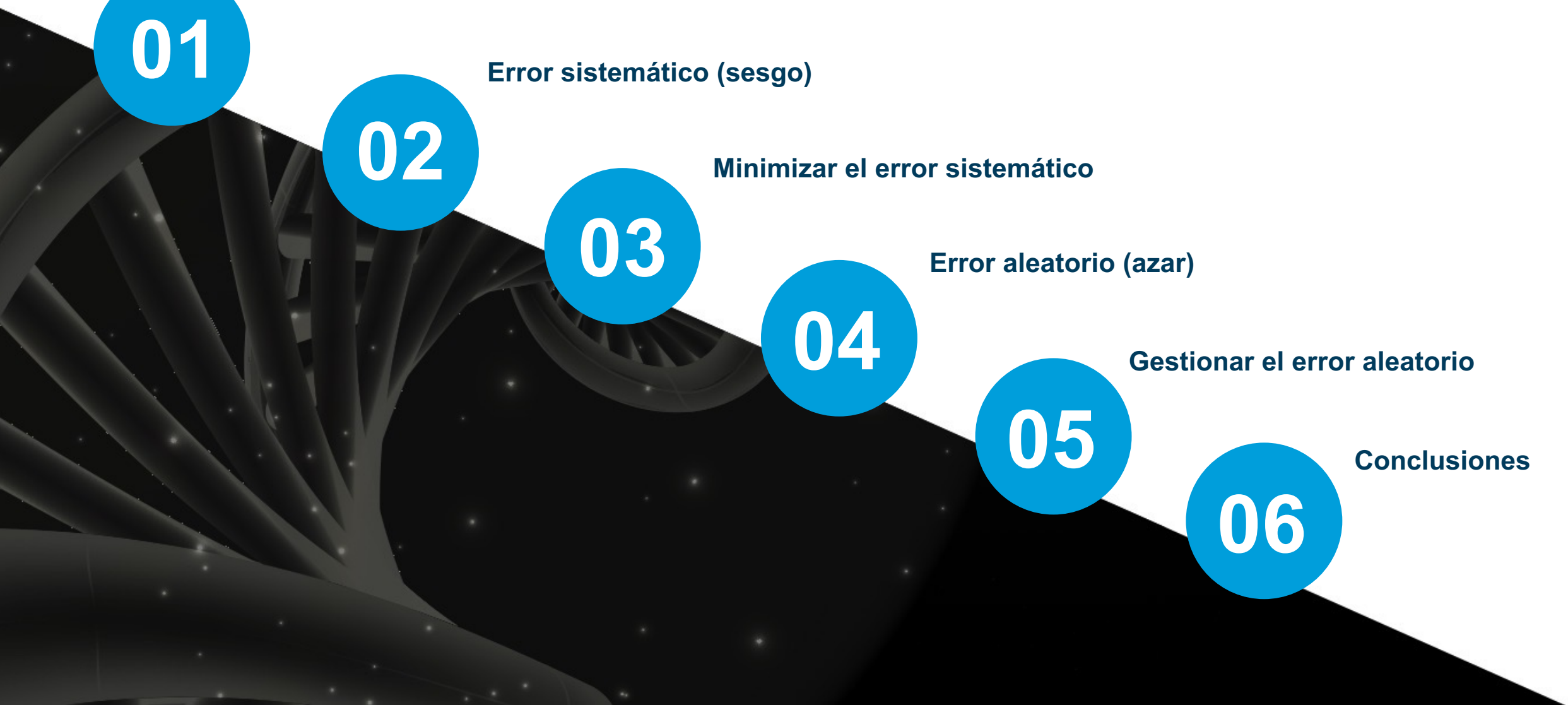
Epidemiólogo clínico I Alergólogo clínico

Mayo, 2024

PP-NUB-CO-0483-1

DISCOVER
Driving Scientific Cancer Research

Tabla de contenidos



DISCOVER
Driving Scientific Cancer Research

BAYER

01 Error en la investigación

1. Error en la investigación

Definición AZAR

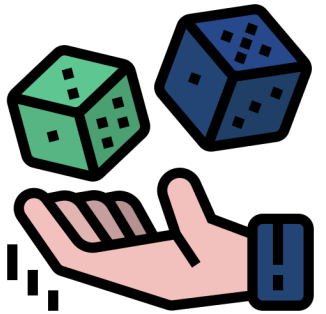
- Deriva del árabe "az-zahr", que es el dado utilizado en el juego
- DRAE: Casualidad, caso fortuito, desgracia imprevista.
  - Azar como encuentro accidental
  - Azar como desorden o como complejidad
  - Azar como proceso que carece de finalidad.

- Rothman, et al.:
  - Uno se refiere al resultado de un proceso aleatorio (experimento) que no puede predecirse y el otro es un resultado que tampoco se puede predecir fácilmente pero que no constituye un fenómeno aleatorio.

DISCOVER
Driving Scientific Cancer Research

1. Error en la investigación

AZAR y juego



- Probabilidad (cara o sello) al lanzar una moneda es 0,5.
  - Si se lanza 10 veces podría ocurrir que no se obtengan 5 caras y 5 sellos como seria lo esperado, y obtener combinaciones divergentes como 8:2 o 9;1
- El azar es responsable para esta variabilidad en los resultados.
- Lanzar dos dados no sesgados (dados con igual probabilidad de caer en un numero entre 1 y 6) dará dos unos o dos seis

DISCOVER
Driving Scientific Cancer Research

1. Error en la investigación

AZAR e investigación clínica

- El azar no se restringe al mundo del lanzamiento de monedas, dados o juego de cartas.

  - Si tomamos una muestra de pacientes de una comunidad, el azar puede resultar en una inusual distribución de una enfermedad crónica.
  - El azar también puede ser responsable de un desequilibrio en la tasa de eventos en dos grupos de pacientes que se les ha dado diferentes tratamientos que son igualmente efectivos.
  - Se requiere herramientas estadísticas para determinar la extensión en la cual la distribución desequilibrada es atribuida al azar o a otra explicación

1. Error en la investigación

AZAR y variación aleatoria

- Como las observaciones sobre las enfermedades se realizan generalmente en **muestras de pacientes** más que en todos los individuos (población), podrían representar **erróneamente** la situación, incluso si no están sesgadas, **debido al azar**.

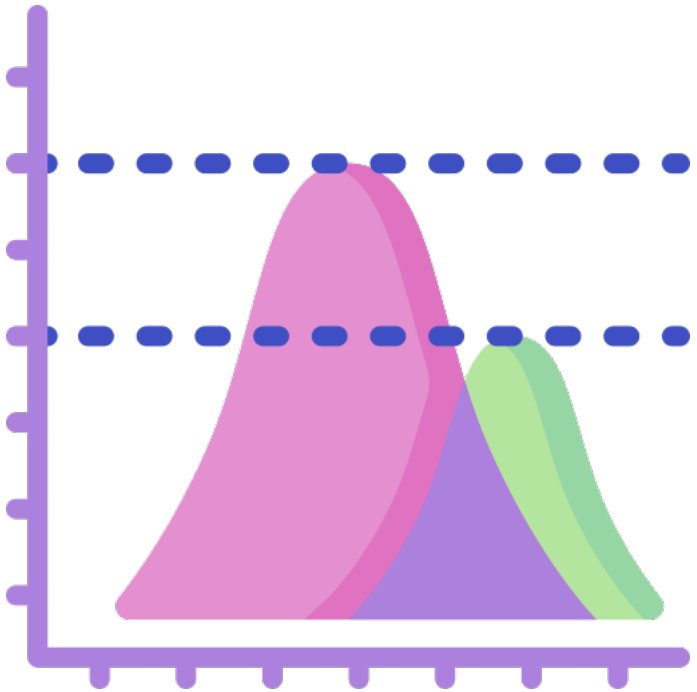
- **La diferencia** de una observación de una muestra con respecto al **valor real de la población**, debida únicamente al azar, se llama ***variación aleatoria***.

- Si las observaciones **se repiten numerosas veces** en muestras de pacientes, los resultados para la muestra variaran alrededor del valor real.

La divergencia entre una observación de una muestra y el valor real de la población, debida únicamente al azar es la ***variación aleatoria***

1. Error en la investigación

SESGO y variación aleatoria



- A diferencia del sesgo, que desvía los valores en una dirección o en otra, la variación aleatoria **tiene tantas probabilidades de sobreestimar o subestimar el valor real.**

- Como consecuencia la media de observaciones no sesgadas tiene tendencia a corresponder al valor real de la población. Aunque los resultados de los valores individuales no se correspondan.

1. Error en la investigación

Variación aleatoria y estadística

- La *variación aleatoria* se produce en el **muestreo de pacientes para el estudio**, en la selección de grupos de tratamiento y en las mediciones llevadas a cabo en cada grupo.

- La variación aleatoria **nunca puede eliminarse**, de tal manera que cuando se evaluen los resultados de las observaciones clínicas siempre debe considerarse el azar.

- La estadística puede contribuir a estimar la probabilidad de que el azar justifique los resultados clínicos y también a *disminuir* dicha probabilidad durante las etapas de diseño y análisis.

1. Error en la investigación

El error es inevitable en la investigación. Este puede surgir de:

Herramientas de medición

Inexactitudes en los instrumentos de medición, como un tensiómetro defectuoso o un microscopio mal calibrado.

Participantes del estudio

Variaciones en la forma en que los individuos responden a las preguntas o participan en los experimentos.

Investigadores

Subjetividad en la observación o el registro de datos, incluso involuntariamente.

Factores externos

Eventos imprevistos o cambios en el entorno que influyen en el estudio.

1. Error en la investigación

Error de medición

- Un atributo implícito a toda variable es la susceptibilidad de ser medida.
- La diferencia entre el valor obtenido al medir una variable con relación a su valor real y objetivo es el error de medición.
- Hay dos tipos de errores:

DISCOVER
Driving Scientific Cancer Research

1. Error en la investigación

Existen dos tipos principales de error

Error Sistemático (sesgo)

- Una desviación constante y direccional del valor real
- Normalmente ocurre debido a defectos en el diseño del estudio, la medición o el muestreo

Error Aleatorio (azar)

- Fluctuaciones Impredecibles en las mediciones
- Ocurre debido a variaciones aleatorias en la muestra o el proceso de medición

DISCOVER
Driving Scientific Cancer Research

02

Error sistemático (sesgo)

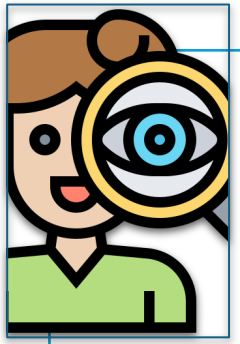
2. Error sistemático (sesgo)

Sesgos

- ¿Qué significa cuando se dice que un estudio es válido o creíble?
  - La validez es el grado en el cual un estudio responde en forma apropiada las preguntas que están siendo solicitadas o mide apropiadamente lo que intenta medir. El sesgo lleva a una desviación sistemática, el error del sesgo tiene dirección.
  - En un ensayo clínico el sesgo lleva a subestimación o sobreestimación del efecto benéfico o perjudicial.
  - Al iniciar el estudio el sesgo puede ser origen de diferencias adicionales a las que se derivan de la intervención experimental.

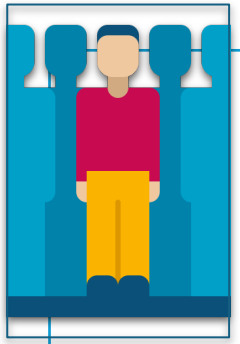
2. Error sistemático (sesgo)

Fuentes de variabilidad en los estudios



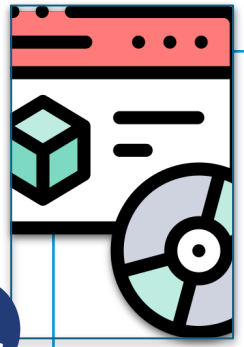
Variabilidad debida al observador

- Tiene que ver con la apreciación del observador, si son dos observadores pueden no estar midiendo exactamente lo mismo



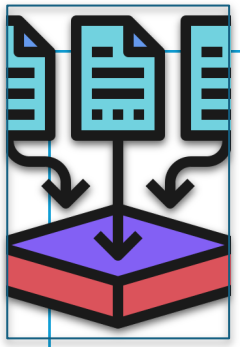
Variabilidad debida al individuo

- La relacionada con el individuo que se estudia, factores como animo o humor, ritmo circadiano entre otros.



Variabilidad debida al instrumento de medición

- La precisión de un tensiómetro o de diferentes encuestas puede variar



Variabilidad debida a errores en el registro de los datos

- Puede ser en el momento del registro o de la introducción en la base de datos.

DISCOVER
Driving Scientific Cancer Research

2. Error sistemático (sesgo)

Existen diferentes tipos de errores sistemáticos en la investigación oncológica, estos son:

Sesgo de Selección

- Diferencias sistemáticas entre los seleccionados y no seleccionados
  - *Ejemplo: reclutamiento exclusivo de pacientes con cáncer en un hospital privado*

Sesgo de Medición

- Errores en la forma en que se miden las variables, esto lleva a resultados constantemente inexactos
  - *Ejemplo: cuestionario que evalúa el dolor mal redactado o es culturalmente no sensible*

Sesgo de Información

- Errores en la recogida, registro o manejo de datos.
  - *Ejemplo: Investigadores conscientes de la asignación del grupo de tratamiento de pacientes*

Sesgo de Recuerdo

- Errores en la forma en que los participantes recuerdan eventos pasados, particularmente relevantes en los estudios retrospectivos
  - *Ejemplo: Pacientes con cáncer más propensos a recordar exposiciones o factores de riesgo específicos*

Sesgo de Publicación

- Tendencia a que los estudios con resultados positivos o estadísticamente significativos tengan más probabilidades de ser publicados que aquellos con hallazgos negativos o no concluyentes
- Esto puede distorsionar la base de evidencia general en oncología.

DISCOVER
Driving Scientific Cancer Research

DISCOVER
Driving Scientific Cancer Research

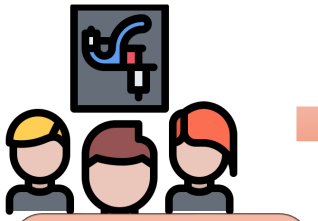
BAYER

03

Minimizar el error sistemático

3. Minimizar el error sistemático

Estrategias para minimizar el error sistemático

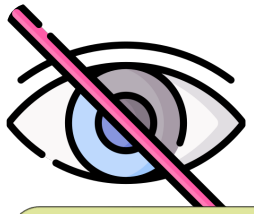


Diseño cuidadoso del estudio

Define claramente tu pregunta de investigación, la población de estudio y los criterios de elegibilidad para evitar el sesgo de selección

Protocolos estandarizados

Utiliza herramientas de medición validadas y procedimientos estandarizados para la recogida y el manojo de datos para reducir los sesgos de medición e Información



Cegamientos

Mantén a los investigadores y/o a los participantes sin conocimiento de las asignaciones de los grupos de tratamiento (si procede) para minimizar el sesgo del observador

Pruebas piloto

Prueba tus métodos y herramientas de recogida de datos en un estudio piloto para identificar posibles fuentes de error desde el principio

Formación y calibración

Asegúrate de que todo el personal de investigación está adecuadamente formado y de que los instrumentos de medición se calibran periódicamente.

Revisiones sistemáticas y metaanálisis

Al revisar la literatura existente, ten en cuenta el posible sesgo de publicación y utiliza métodos rigurosos para evaluar la calidad de los estudios incluidos.

DISCOVER
Driving Scientific Cancer Research

BAYER

04 Error aleatorio (azar)

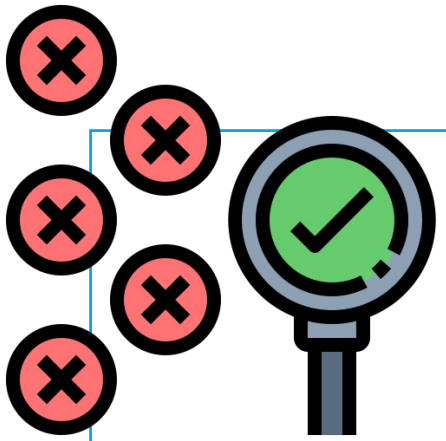
4. Error aleatorio (azar)

Fuentes de error aleatorio

- ***Error de muestreo***
  - La diferencia entre la muestra y el valor real de la población
  - Las muestras más grandes reducen el error de muestreo
- ***Error de medición***
  - Fluctuaciones aleatorias en la forma en que se miden las variables, incluso cuando se utilizan herramientas fiables

DISCOVER
Driving Scientific Cancer Research

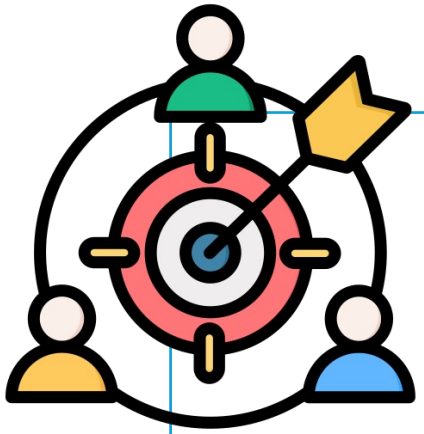
4. Error aleatorio (azar)



Impacto del error aleatorio

- **Reduce la precisión**
  - Dificulta la detección de efectos reales
- **Amplía los intervalos de confianza**
  - Aumenta la incertidumbre en torno a nuestras estimaciones.

4. Error aleatorio (azar)



¿Como se determina la precisión?

- La precisión es contraria a la variabilidad.
- Una forma de medir la precisión de una variable es midiendo la variabilidad para lo cual se emplea la desviación estándar de una serie de medidas repetidas.
- También se puede emplear el coeficiente de variación (CV) cuando se desea comparar dos variables cuantitativas.

DISCOVER
Driving Scientific Cancer Research

BAYER

05

Gestionar el error aleatorio

5. Gestionar el error aleatorio

Estrategias para gestionar el error aleatorio

Estandarizar los métodos de medición

Elabora un manual sobre el proceso de medición para que se pueda realizar de manera consistente.

Entrenar a los observadores

Entrena los diferentes participantes en la realización de las mediciones para que sea lo más uniforme posible.

Refinamiento de los instrumentos de medición

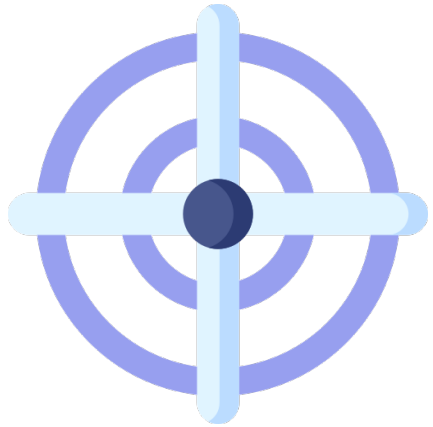
Los instrumentos mecánicos y electrónicos deben ser calibrados de manera que la variabilidad sea mínima.

Automatización de los instrumentos

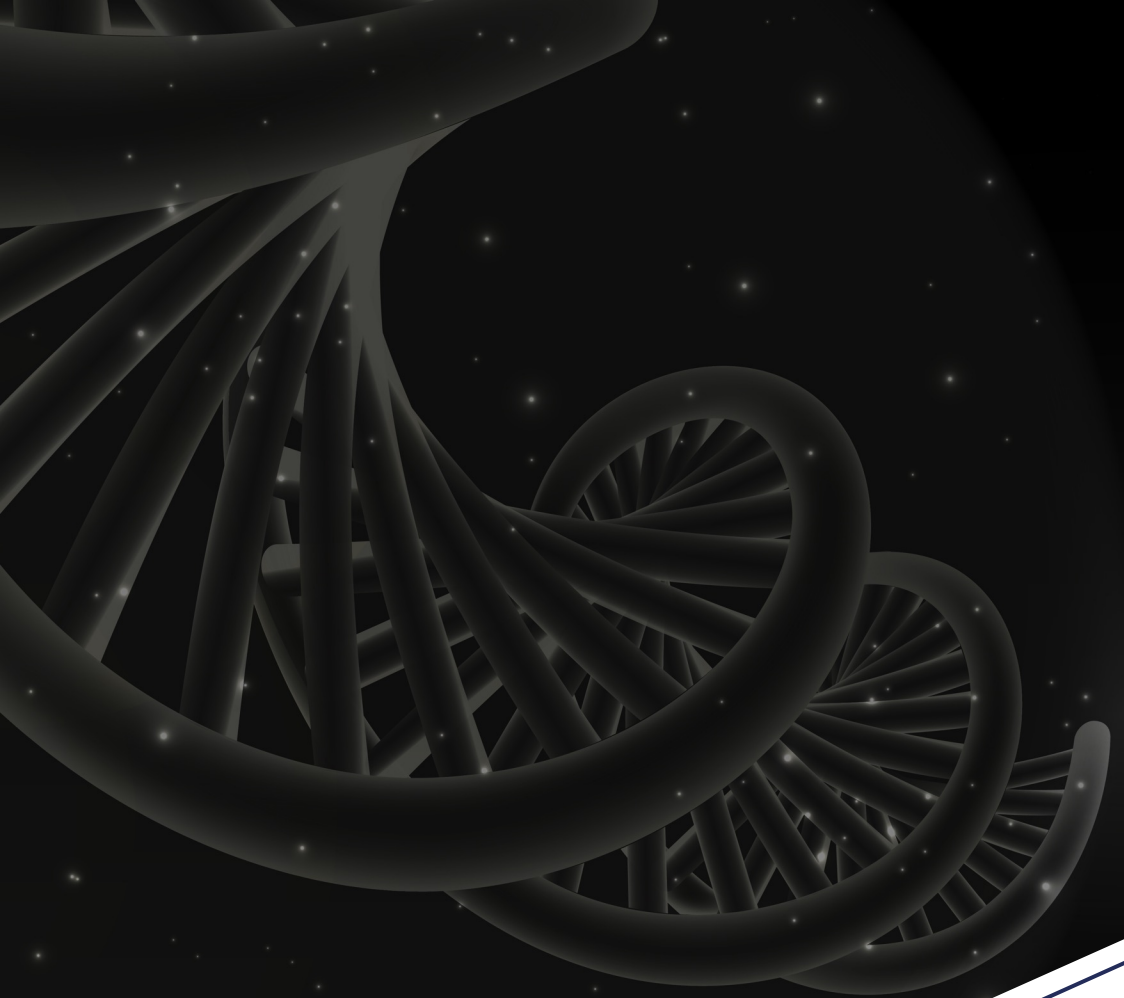
Evitar la variabilidad reduciendo la manipulación humana.

Repetición

La repetición de la medida reduce el error aleatorio.

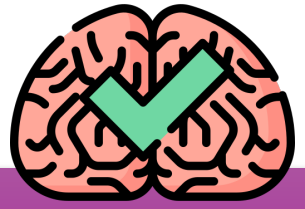


DISCOVER
Driving Scientific Cancer Research



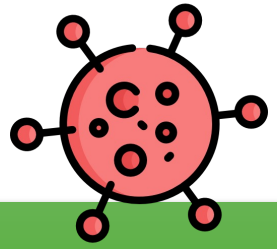
06 Conclusiones

6. Conclusiones



Comprender tanto el error sistemático como el error aleatorio es crucial para llevar a cabo una investigación rigurosa y fiable en oncología

Al tomar medidas para minimizar el sesgo y gestionar el error aleatorio, mejoramos la precisión de nuestros hallazgos.



Pero también contribuimos a una base de evidencia más sólida para una mejor atención al cáncer

La precisión es lo contrario de error aleatorio y la exactitud es lo contrario de error sistemático.

Referencias

Rothman KJ, Greenland S, Lash TL. Modern epidemiology. Lippincott Williams & Wilkins; 2008. Available from: https://shop.lww.com/Modern-Epidemiology/p/9781451193282

Szklo M, Nieto FJ. Epidemiology: Beyond the Basics: Jones & Bartlett Learning; 2014.

Daniel WW, Cross CL. Biostatistics: A Foundation for Analysis in the Health Sciences, 10th Edition: A Foundation for Analysis in the Health Sciences: Wiley; 2012.

Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.4 (updated August 2023). Cochrane, 2023. Available from www.training.cochrane.org/handbook.

Altman DG. Practical Statistics for Medical Research: Taylor & Francis; 1990.

3

Describing Data

3.1 INTRODUCTION

If there is one key concept underlying the subject of statistics, it is that of variability. In medicine we can see this most obviously in the way people differ in their physiological, biochemical and other characteristics and also in their variable responses to disease and to therapy. We also often encounter variability between machines that are supposed to be identical, and between different observers. There are sometimes many sources of variability present at once. For example, if I have my blood pressure measured the value recorded by my GP will depend greatly on some unknown underlying 'true' value, but it will also relate to the time of day, whether I was late and had to run to the surgery, the type of sphygmomanometer being used, whether I was anxious about the outcome, and so on. When many people have their blood pressure measured other factors will affect between-subject variability, such as age, sex and race.

In general we can divide variability into that due to known causes and that which is unexplained. Thus, for example, in a study of men aged 25 to 65 part of the variability in their blood pressures may be ascribed to their age, but most of the rest is unexplained. We often refer to this unexplained variability as **random variation**.

In any study we will usually want to summarize some of the data in a simple way. Sometimes this will be as far as the statistical analysis goes, but often it is a first step. For categorical variables, such as sex and blood group, it is straightforward to present the number in each category, usually indicating the frequency or percentage of the total number of patients. When shown graphically this is called a **bar diagram**. Figure 3.1 shows a bar diagram of general aviation accident rates in 1974 by occupation (Booze, 1977). A similar diagram can also be used to relate frequencies (or rates) to values of another variable. For example, Figure 3.2 shows perinatal mortality per 1000 births in England and Wales in 1979 by day of the week. The higher mortality rates at the weekend are clearly seen. It is very important that the vertical axis of a bar diagram starts at zero, otherwise the visual impression is misleading, with the differences between groups being exaggerated.

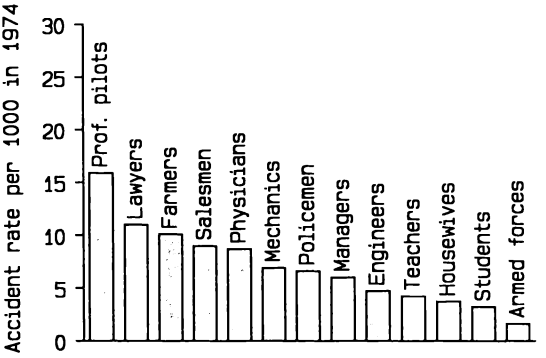


Figure 3.1 Bar diagram showing general aviation accident rates (per 1000) in 1974 by occupation (Booze, 1977).

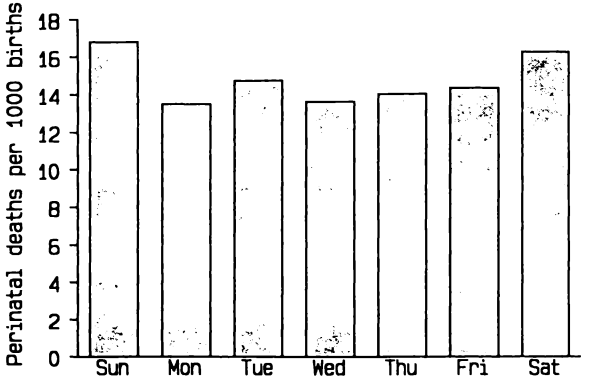


Figure 3.2 Perinatal mortality in England and Wales in 1979 by day of the week (Macfarlane and Mugford, 1984).

For continuous variables, such as age and serum bilirubin, there will be a large number of different observed values, so an alternative approach is needed. The remainder of this chapter concentrates on ways of describing and summarizing such data both numerically and graphically.

In this chapter I shall introduce some mathematical notation for the first

time. Further explanation of this notation can be found in Appendix A at the end of the book.

3.2 AVERAGES

The obvious first step when describing a set of observations of a continuous variable is to calculate the average value. In colloquial use the word 'average' does not have a precise meaning, but in statistics there are several so-called 'measures of central tendency' that are precisely defined and which can be taken as the average or typical value.

The most common of these is the **arithmetic mean**, usually just called the **mean**, which is the sum of all the observations divided by the number of observations. Table 3.1 shows age and lung function data for 25 patients with cystic fibrosis. The variable shown is the maximal static inspiratory

Table 3.1 Age and PImax in 25 patients with cystic fibrosis (O'Neill *et al.*, 1983)

Subject	Age (years)	PImax (cm H_2O)
1	7	80
2	7	85
3	8	110
4	8	95
5	8	95
6	9	100
7	11	45
8	12	95
9	12	130
10	13	75
11	13	80
12	14	70
13	14	80
14	15	100
15	16	120
16	17	110
17	17	125
18	17	75
19	17	100
20	19	40
21	19	75
22	20	110
23	23	150
24	23	75
25	23	95

pressure (PImax) and is an index of respiratory muscle strength. The sum of the PImax values is 2315, so the mean is 2315/25 = 92.6 cm H_2O. The mean is the value usually meant when talking about 'the average'. The mean is sometimes indicated by $\bar{x}$ (pronounced 'x bar'), but this shorthand notation is best avoided other than in equations.

The other frequently used measure is the **median**. This is the value that comes half-way when the data are ranked in order. For the PImax data in Table 3.1 there are 25 observations, so the median is the 13th value in order. If we rank the PImax values in ascending order we get

Rank	1	2	3	4	5	6	7	8	9	10	11	12	13
PImax	40	45	70	75	75	75	75	80	80	80	85	95	**95**
Rank	14	15	16	17	18	19	20	21	22	23	24	25	
PImax	95	95	100	100	100	110	110	110	120	125	130	150	

and we can see that the median is 95 cm H_2O. More easily, we can see immediately from Table 3.1 that the median age of these patients was 14 years. When there is an even number of observations the median is defined as the average of the two central values: if we had 24 observations the median would be the average of the 12th and 13th values in an ordered listing of the observations. There are usually equal numbers of observations above and below the median. However, when there is more than one observation equal to the median, as for the PImax data, this may not be exactly true.

The median is especially useful when some extreme data values are censored. If observations are not recorded precisely when they are above a certain level or below a level of detection, we cannot calculate the mean, but we can calculate the median if we have definite values for over half the subjects. The median is also valuable in the analysis of survival times, which is considered in Chapter 13.

The mean and the median are both widely used to describe the average or typical value of a set of data. The mean is much more frequently used because this ties in well with the most common types of statistical analysis, but the median is in no way inferior as a descriptive statistic and in some circumstances it is much more useful than the mean, as we shall see later. In some situations we calculate another measure known as the **geometric mean**, which is usually close to the median. Its use is described in section 3.4.4.

A final indicator of the centre of a set of data is the **mode** which is simply the most common value observed. The mode is rarely of any practical use for continuous data.

3.3 DESCRIBING VARIABILITY

The second aspect of describing a set of observations of a continuous

variable is to assess the variability of the observations in some way. Any set of data will contain many different values, for example the PImax data shown above. We are interested in the way these values are distributed – are they all similar or do they vary a lot? There are several ways of tackling this problem. I shall look first at graphical methods, and then consider numerical methods.

3.3.1 Histogram

A simple graphical way of depicting a complete set of observations is by means of the **histogram** in which the number (or frequency) of observations is plotted for different values or groups of values. Table 3.2 shows the **frequency distribution** of the immunoglobulin IgM in 298 healthy children aged 6 months to 6 years, and Figure 3.3 shows a histogram of

Table 3.2 Concentrations of serum IgM in 298 children aged 6 months to 6 years (Isaacs *et al.*, 1983)

IgM (g/l)	Number of Children
0.1	3
0.2	7
0.3	19
0.4	27
0.5	32
0.6	35
0.7	38
0.8	38
0.9	22
1.0	16
1.1	16
1.2	6
1.3	7
1.4	9
1.5	6
1.6	2
1.7	3
1.8	3
2.0	3
2.1	2
2.2	1
2.5	1
2.7	1
4.5	1

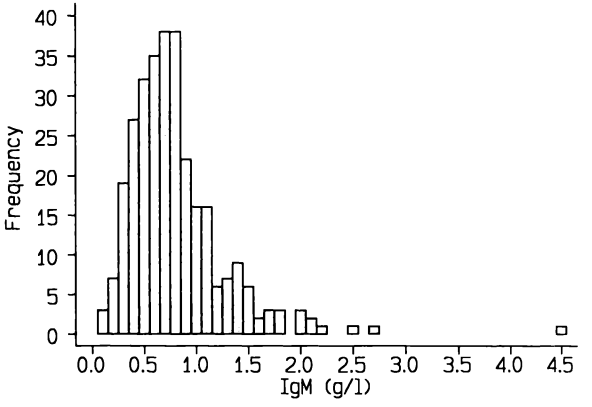


Figure 3.3 Frequency histogram of IgM concentrations in 298 children aged 6 months to 6 years (Isaacs *et al.*, 1983).

these values. If there are many different values it is often desirable to group observations before constructing a histogram in order to get a better visual impression. Unless the sample is very large somewhere around 8 to 15 groups will usually suffice for a satisfactory display. This will depend upon the actual data, for it is desirable to keep the groupings simple. Although we could group the IgM data in intervals of, say, 0.25, this goes beyond the precision of the data. Better is the grouping in intervals of 0.2 shown in Figure 3.4. Note that the width of each vertical bar covers the range of values that have been grouped. So, for example, when we group 0.1 and 0.2 we are actually including values between 0.05 and 0.25 even though the data were not recorded that accurately. A histogram is similar to a bar diagram, but because the frequencies relate to a continuous variable adjacent bars of a histogram should touch.

The bars in histograms are usually all the same width, because the groupings are the same size. If the groups are not the same size this should be allowed for by remembering that it is the *area* of each bar that is proportional to the frequency, not its height. This principle is illustrated on data showing the age distribution of road accident casualties in the London borough of Harrow in 1985. Table 3.3 shows the data as presented. Most of the casualties were adults, with the greatest number in the age range 25 to 59. Clearly the widths of the groupings vary considerably, from 1 to 35 years in fact, and this must be taken account of in a histogram of the data. Note that in order to include the 60+ age group in a histogram we have to assume a reasonable upper age limit – here it will be taken as 80.

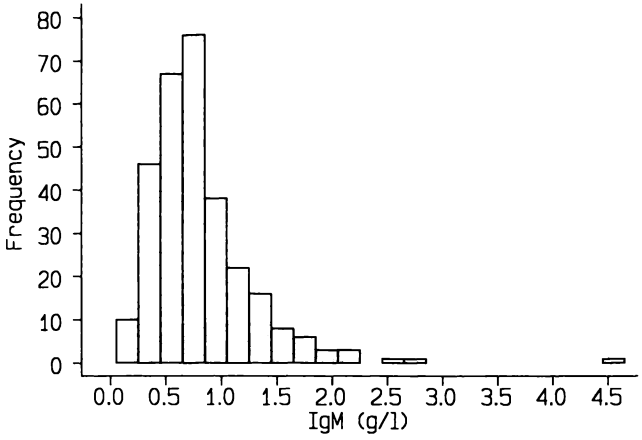


Figure 3.4 As Figure 3.3 but data grouped in intervals of 0.2 g/l.

Table 3.3 Road accident casualties in the London Borough of Harrow in 1985 (excluding 65 with unknown age)

Age	Frequency
0– 4	28
5– 9	46
10–15	58
16	20
17	31
18–19	64
20–24	149
25–59	316
60+	103
Total	815

First, consider what happens if we ignore the above warning and draw a histogram where, for each age group, the height indicates the frequency shown in Table 3.3 and the width shows the age range – this is shown in Figure 3.5. This histogram suggests that accident victims are much less likely to be 16 and 17 year olds than adults, whereas we would probably expect the opposite to be true. We get the correct picture by making the frequencies correspond to the area of each bar rather than its height, as is shown in Figure 3.6. What we have done is consider the number of casualties *per year of age* – where we don't have this explicitly we take the

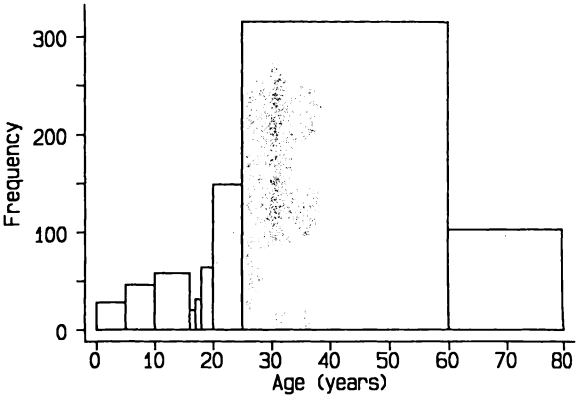


Figure 3.5 Incorrect histogram of road accident data of Table 3.3.

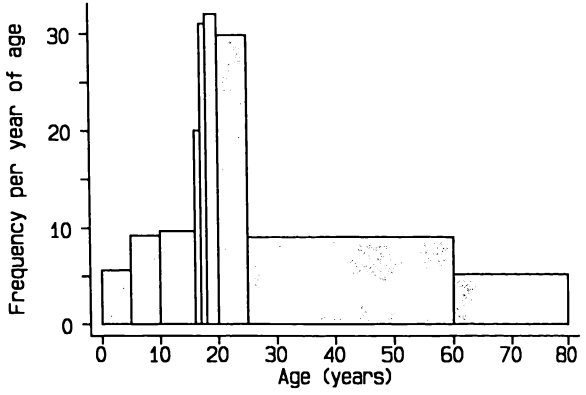


Figure 3.6 Correct histogram of road accident data.

average value in that age group. Figure 3.6 shows a true impression of the data, from which we can see that road accident casualties are more likely to be aged 16 to 24 than any other age group.

Note that this histogram just shows the observed numbers of casualties. It does not indicate the *risk* of a road accident for people of varying age – for this we would also need to know the age distribution of the population, and would need to assume that all casualties lived in Harrow and that no Harrow residents had accidents elsewhere.

It is sometimes more useful to show the proportion of the sample in each interval. All the frequencies are converted into percentages by dividing by the sample size and multiplying by 100. Figure 3.7(a) shows the resulting **relative frequency histogram** for the IgM data, which differs from Figure 3.3 only in the way the vertical axis is labelled. An alternative way of plotting the data is to join the mid-points of the tops of all the vertical bars of the histogram; this is called a **frequency polygon**. Figure 3.7(b) shows such a plot for the same data.

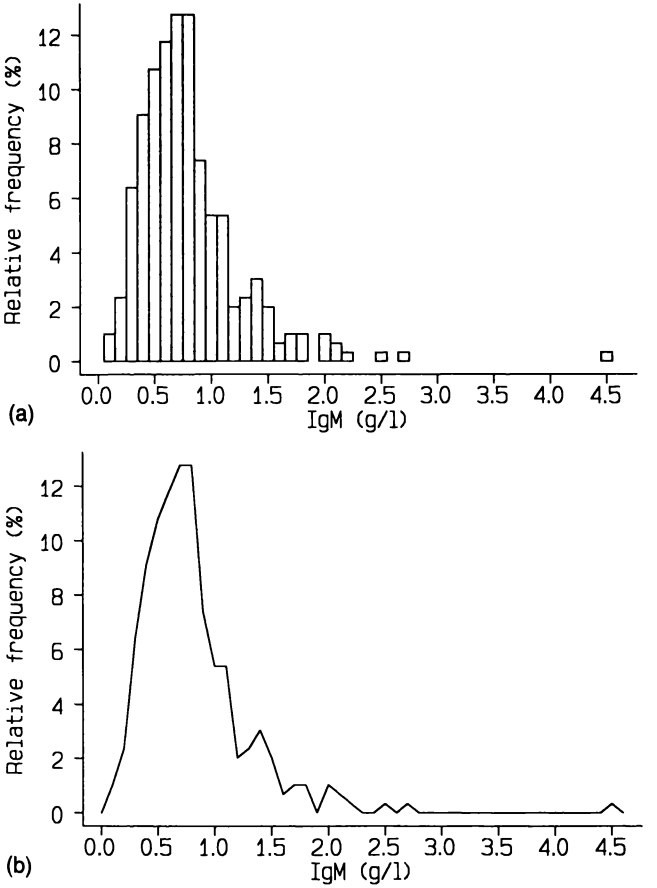


Figure 3.7 IgM data in Figure 3.3 shown as (a) Relative frequency histogram, (b) Relative frequency polygon.

The vertical axis of a histogram must start at zero, and there should not be any breaks in the scale. Otherwise the visual impression will be misleading. Likewise three-dimensional effects should not be used.

3.3.2 Stem-and-leaf diagram

A clever modification of the histogram called a **stem-and-leaf** diagram allows all the actual observations to be shown too. Figure 3.8 shows the PImax data from Table 3.1 redrawn as a stem-and-leaf diagram. The raw data can be reconstructed by joining the numbers on the left (the stems) to each of the numbers on the right (the leaves) on the same row. This is a very economical way of reproducing the raw data, and is more useful than a simple list of the data.

```
 4  05
 5
 6
 7  05555
 8  0005
 9  5555
10  000
11  000
12  05
13  0
14
15  0
```

Figure 3.8 Stem-and-leaf diagram of PImax data in Table 3.1.

The stem-and-leaf diagram works well in many circumstances, especially where there are many different values, but the best format depends on the nature of the data and the sample size. The IgM data in Table 3.2 cannot be made into a successful stem-and-leaf diagram using five 'stems' (0, 1, 2, 3, 4), but we can split each group to get a useful diagram, as in Figure 3.9.

```
0  111
0  22222223333333333333333
0  44444444444444444444444444455555555555555555555555555555
0  66666666666666666666666666666677777777777777777777777777777
0  88888888888888888888888888888888888899999999999999999
1  00000000000000000000011111111111111
1  2222223333333
1  4444444455555
1  66777
1  888
2  00011
2  2
2  5
2  7
2
3
3
3
3
3
4
4
4  5
```

Figure 3.9 Stem-and-leaf diagram of IgM data in Table 3.2.

3.3.3 Cumulative frequencies

We saw earlier how the distribution of a sample of observations can be shown as the percentage of the sample with values in each of several small ranges. This was shown in the relative frequency histogram in Figure 3.7. We can take this idea a stage further by considering for each group the proportion of subjects in that group *or a lower one*. Thus we calculate the **cumulative frequency** at each level – the proportion of observations less than or equal to each value. The calculations are shown in Table 3.4. The cumulative relative frequencies can be plotted in a histogram, as in Figure 3.10(a). However, for cumulative frequencies there is no need to group the data like this because we can plot the cumulative frequencies directly, as in Figure 3.10(b). This plot can be used either to see what percentage of

Table 3.4 Cumulative frequency distribution of 298 IgM values

IgM g/l	Frequency	Relative Frequency %	Cumulative Frequency	Cumulative Relative Frequency %
0.1	3	1.0	3	1.0
0.2	7	2.3	10	3.4
0.3	19	6.4	29	9.7
0.4	27	9.1	56	18.8
0.5	32	10.7	88	29.5
0.6	35	11.7	123	41.3
0.7	38	12.8	161	54.0
0.8	38	12.8	199	66.8
0.9	22	7.4	221	74.2
1.0	16	5.4	237	79.5
1.1	16	5.4	253	84.9
1.2	6	2.0	259	86.9
1.3	7	2.3	266	89.3
1.4	9	3.0	275	92.3
1.5	6	2.0	281	94.3
1.6	2	0.7	283	95.0
1.7	3	1.0	286	96.0
1.8	3	1.0	289	97.0
2.0	3	1.0	292	98.0
2.1	2	0.7	294	98.7
2.2	1	0.3	295	99.0
2.5	1	0.3	296	99.3
2.7	1	0.3	297	99.7
4.5	1	0.3	298	100.0
Total	298	99.9		

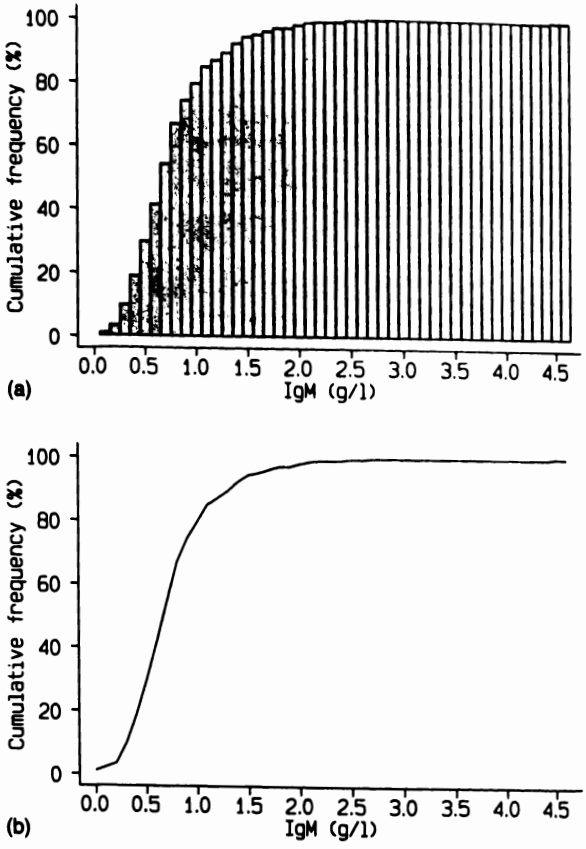


Figure 3.10 IgM data shown as (a) Cumulative relative frequency histogram, (b) Cumulative distribution.

observations lie above or below any chosen level, or to find the values which a given percentage of children's IgM values lie above or below. For example, we can easily see that the median IgM concentration was 0.7 g/l. This information cannot be obtained from a histogram or cumulative histogram if values have been grouped.

Cumulative frequencies are especially useful for comparing the distribution of values in two or more different groups of individuals. Figure 3.11(a) shows relative frequency histograms for the age at first tooth eruption of 1568 children of smokers and 1576 non-smokers. Figure 3.11(b) shows cumulative histograms of the same data. Figure 3.11(c) shows cumulative frequency polygons of the same data. Because we are considering cumulative frequencies we join the right-hand points of the vertical bars rather than the mid-points as in Figure 3.7(b). This plot shows that the difference between the groups is not as great as was suggested in Figure 3.11(b) – the two groups were side by side in the previous plot, which can lead to a misleading visual impression. We can easily see from Figure 3.11(c) that the median age at first tooth eruption was about one week earlier in the children of smokers.

3.4 QUANTIFYING VARIABILITY

Graphical methods are important for examining the variability of data, but it is necessary also to have a numerical way of summarizing the amount of variability. Used in conjunction with the mean, this would provide an informative but brief summary of a set of observations. There are three main approaches to quantifying the variability of a set of data. We can either quote the range of all the values, specific values derived from the cumulative frequency distribution, or we can obtain a numerical measure of the dispersion of the observations around the mean.

3.4.1 Range

The simplest way to describe the spread of a set of data is to quote the lowest and highest values. These values are known as the **range**. The range of the IgM data was 0.1 to 4.5 g/l. This is not a satisfactory summary, because it takes account of only the most extreme (and perhaps most peculiar) values at each end of the data, and the way the intermediate values are distributed will not influence the range. Thus for the IgM data we have no idea that 4.5 was considerably more than the second highest value of 2.7 g/l. Mainly for this reason the range is not widely used.

3.4.2 Centiles

By specifying two values that encompass *most* rather than all of the data values we get round much of the difficulty. For example, we could calculate the values between which 90% of the observations lie. The value below which a given percentage of the values occur is called a **centile** or **percentile**, and corresponds to a value with a specified cumulative relative frequency.

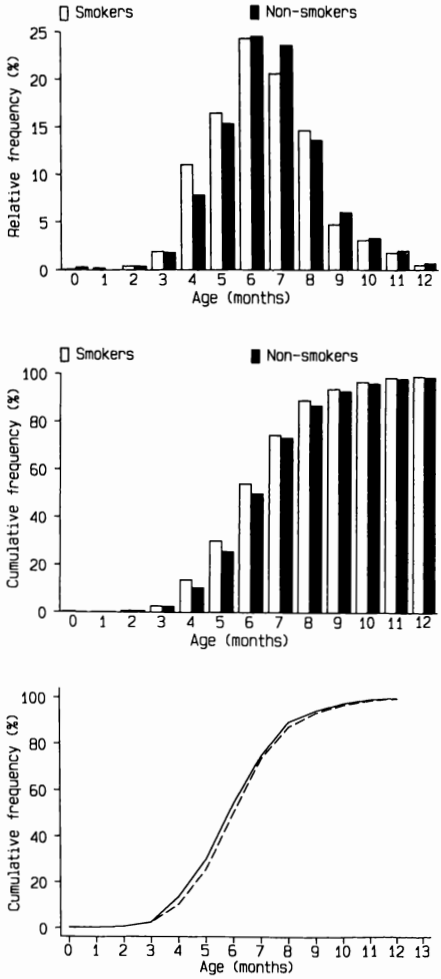


Figure 3.11 Age at first tooth eruption of children born to smokers (——) and non-smokers (- - - - - -) (Rantakallio and Mäkinen, 1984): (a) Relative frequency histogram; (b) Cumulative relative frequency histogram; (c) Cumulative relative frequency polygon.

We require the 5th and 95th centiles of the distribution of IgM values. From the last column of Table 3.4 we can see that the cumulative relative frequency passes 5% somewhere in the group of IgM values of 0.3 g/l, and 95% is reached at the value of 1.6 g/l.

A more correct general approach is to calculate the ranks of the required observations, which we do by taking the necessary percentages of the sample size plus one. Here we need the values with ranks $0.05 \times 299 = 14.95$ and $0.95 \times 299 = 284.05$. This calculation usually leads to non-integer values, so we may need to interpolate. For example we want the value of IgM 0.95 of the way between the 14th and 15th observations in rank order. As these are, from Table 3.4, both equal to 0.3 g/l the 5th centile is 0.3 g/l, and likewise the 95th centile is 1.7 g/l. However, if we want the 10th centile, we would need the IgM value corresponding to a rank of $0.10 \times 299 = 29.9$. The observations with ranks 29 and 30 are 0.3 and 0.4 g/l and we take the value nine-tenths of the way between these values, by calculating $0.3 + 0.9(0.4 - 0.3) = 0.39$ g/l. The values 0.3 and 1.7 are thus the 5th and 95th centiles of the observed distribution of IgM in this sample of children and these two values thus specify what we can call a 90% **central range**—the range within which the central 90% of values lie (i.e. excluding 5% at each end of the distribution).

Other centiles can be quoted rather than the 5th and 95th. The most common alternative is to quote a 95% central range ($2\frac{1}{2}$th and $97\frac{1}{2}$th centiles), but an 80% central range (10th and 90th centiles) is sometimes used. The 50th centile is another name for the median, as half of the observations are less than (and greater than) this value. The 25th and 75th centiles are known as **quartiles**; these values together with the median divide the data into four equally populated subgroups. The numerical difference between the 25th and 75th centiles is the **inter-quartile range**, and is occasionally used to describe variability.

A simple but useful semi-graphical way of summarizing data using centiles is the **box-and-whisker plot**. Figure 3.12 shows a box-and-whisker plot for the IgM data. The box indicates the lower and upper quartiles and the central line is the median. The points at the ends of the 'whiskers' are the $2\frac{1}{2}$% and $97\frac{1}{2}$% values, although the whiskers sometimes indicate the extreme values. For a single set of data a histogram is more informative, but several sets of data can be summarized economically using the box-and-whisker plot. Sometimes any values outside the range of the whiskers are plotted individually.

3.4.3 Standard deviation

The alternative approach to quantifying variability is based on the idea of averaging the distance each value is from the mean. For an individual with

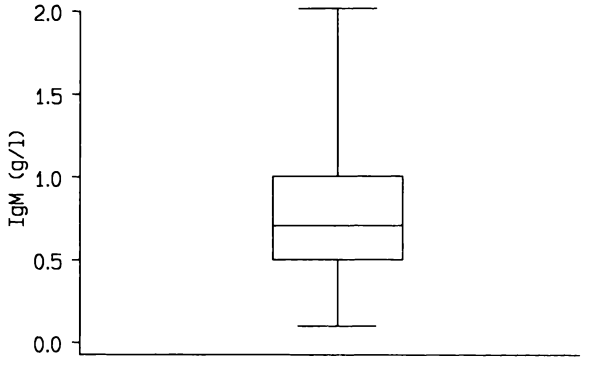


Figure 3.12 Box-and-whisker plot of the IgM data, showing the $2\frac{1}{2}$, 25, 50, 75 and $97\frac{1}{2}$% cumulative relative frequencies (centiles).

an observed value x_i the distance from the mean $\bar{x}$ is $x_i - \bar{x}$, and if we have n observations we have a set of n such distances, one for each individual. For observations below the mean the difference will be negative. We can calculate the average distance between the observations and their mean, but the sum of these distances, $\Sigma(x_i - \bar{x})$, is always zero because of the way the mean is calculated from the individual observations. However, if we square the distances before we sum them we get a quantity that must be positive. The average of these squared differences thus gives a measure of individual deviations from the mean. This quantity is called the **variance**, and is defined as

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}.$$

Note that we divide by $n - 1$ rather than the more obvious n. Dividing by n gives the variance of the observations around the sample mean, but we virtually always consider our data as a sample from some larger population, and wish to use the sample data to estimate the variability in the population. Dividing by $n - 1$ gives us a better estimate of the population variance, although clearly for large samples the difference is negligible.

The variance will turn up in later chapters, notably when discussing the technique known as **analysis of variance**. For our present purpose, the

variance is not a suitable measure for describing variability because it is not in the same units as the raw data. We do not, for example, wish to express the variability of a set of blood pressure measurements in *square* mm Hg. The obvious solution to this problem is to take as our measure the square root of the variance. We call this quantity the **standard deviation**. The standard deviation is usually abbreviated to sd or SD or s or σ (the Greek letter sigma), and is defined as

$$\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

Standard deviation is not a good name for this statistic as there is nothing 'standard' about it. It may more reasonably be thought of as approximately the average deviation (or distance) of the observations from the mean.

Many calculators can calculate the standard deviation, by means of a key marked s or σ. (The use of the Greek σ here rather than s is not strictly correct, as will be explained in the next chapter. If there are keys marked σ_n and σ_{n-1} the latter should be used.)

However, should we wish to do the calculation ourselves there is a much easier formula to use, which is mathematically equivalent:

$$s = \sqrt{\frac{\Sigma x^2 - (\Sigma x)^2/n}{n-1}}.$$

(Note the simplification of the Σ notation, as described in Appendix A.) Using this formula we can calculate the standard deviation from the sum of the observations, Σx, and the sum of the squares of the observations, Σx^2. We do not need to calculate the individual distances from the mean.

For example, for the PImax data shown in Table 3.1 the sum of the data and the sum of the squares of the data are

$$\sum x = 2315 \quad \text{and} \quad \sum x^2 = 229275$$

so the mean PImax is 2315/25 = 92.60 cm H_2O and the standard deviation is

$$s = \sqrt{\frac{229275 - 2315^2/25}{24}}.$$

$$= 24.92 \text{ cm } H_2O.$$

Note that I shall keep an extra decimal place at present for the mean and standard deviation because I shall be doing some further calculations. One decimal place would be sufficient when reporting these results.

The standard deviation has an important role in data analysis, but here we are concerned with its value as a descriptive statistic. In fact, although the standard deviation is widely used for this purpose it is useful only indirectly for describing the variability of a set of data. We can say, for example, that in many circumstances **the large majority (about 95%) of a set of observations will be within two standard deviations of the mean**. The appropriateness of this statement depends on the *shape* of the distribution of the data. If the distribution is reasonably symmetric then the above statement will usually be true.

For the PImax data in Figure 3.8 the mean was 92.60 and the standard deviation was 24.92 cm H_2O. The values that are two standard deviations either side of the mean are $92.60 - 2(24.92) = 42.76$ cm H_2O and $92.60 + 2(24.92) = 142.44$ cm H_2O. (We often use the expression 'mean ±2SD' to mean both of these values, i.e. the mean 'plus or minus' twice the standard deviation.) All but two of the 25 observations were within this range; we would expect to find on average one observation outside the range mean ±2SD (i.e. about 5% of 25).

3.4.4 Skewed distributions

For data which do not have a symmetric distribution we need to be careful when using the standard deviation in the way just described. For example, the IgM data in Figure 3.3 clearly have an asymmetric distribution—there is a long right-hand 'tail'. This is called a **skewed** distribution. The mean and standard deviation of the IgM data are 0.80 and 0.47 g/l respectively. Calculating the mean ±2SD gives the values −0.14 and 1.74. The lower value is negative, which is not a possible value of IgM. The upper value of 1.74 is exceeded by 12 of the observations, 4% of the total. The two values clearly do not describe the range of the bulk of the data very well. Although they still include about 95% of the observations, the exclusions are all in one tail.

For measurements that cannot be negative, which is usually the case, we can infer that the data have a skewed distribution if the standard deviation is more than half the mean. There is no guarantee that the converse is true, however, but a histogram will quickly reveal whether the data are skewed or not. Skewness like that of the IgM data is called **positive** skewness and is common. The opposite phenomenon, with an extended left hand tail, is called **negative** skewness and is rare.

In general, when we have data with a skewed distribution we use other ways of describing the data. There are two main possibilities. The first is to **transform** the data mathematically so that the transformed data have a more nearly symmetric distribution. The most frequent device is to take logarithms (logs) of the data. The rationale for this approach will be

discussed in Chapter 7. We can see that it works well here, however, from Figure 3.13 which shows a histogram of $\log_{10}$ IgM values. The mean and SD of the log data are −0.158 and 0.238 respectively, so that the values mean ±2SD are −0.63 and +0.32. These values are indicated in Figure 3.13. They cut off 10 values in the lower tail of the distribution and 6 in the upper tail, and thus give a range of values encompassing 282/298 or 94.6% of the observations. The cut-off values can be 'back-transformed' to the original scale giving 0.23 and 2.08, and reference to Table 3.2 shows the 16 values outside these limits. If we back-transform (or 'antilog') the mean of the log data we get a quantity known as the **geometric mean**. The geometric mean of the IgM data is thus $10^{-0.158} = 0.695$ g/l. Where log transformation successfully removes skewness the geometric mean will be similar to the median, and will be less than the mean of the raw data. The standard deviation of the log data cannot be meaningfully back-transformed.

Note that log data can be negative, and that it does not matter whether logs to base e or base 10 are used. In this example, logs to base 10 were used, with the function 10^x used for the back-transformation. Log transformation is only useful for removing positive skewness.

The alternative approach to describing the distribution of skewed data is to calculate the centiles corresponding to a chosen central range. For example, to get the values that enclose 95% of the observations we need to calculate the $2\frac{1}{2}$th and $97\frac{1}{2}$th centiles. Using the method described in the

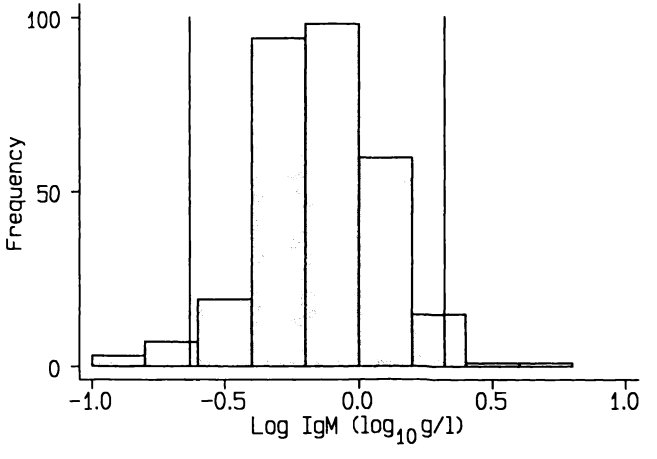


Figure 3.13 Frequency histogram of $\log_{10}$IgM showing the values of mean ± 2SD.

previous section, these values are obtained by interpolation as 0.2 and 2.0 g/l. These values of 0.2 and 2.0 g/l are called **empirical** centiles as opposed to the earlier values of 0.23 and 2.08 (obtained from the mean ±2SD of the log data), which are estimated centiles. The two methods agree well for these data. Likewise the median IgM value is 0.7 g/l, which is very close to the geometric mean.

3.4.5 Comment

The standard deviation is one of the key quantities in statistical analysis. Its value for describing variability is conditional on the distribution of the data. Although it is always valid to calculate the standard deviation we can infer that about 95% of the observations were in the interval mean ±2SD only if we know (or assume) that the distribution of the data was reasonably symmetric. In fact, as happens with the IgM data, the range mean ±2SD may include about 95% of the observations even when the distribution is skewed. However, while we may reasonably use just the mean and SD to summarize such data, the skewness will be hidden. For skewed data, it is preferable to use the median and a 90% or 95% central range to summarize a set of observations. However, it is not practical to quote centiles for small samples, so the range can be given. Otherwise, the standard deviation can be used. It has the advantage of using each observation directly and it is easier to calculate (by computer) for large amounts of data.

The question of the shape of the distribution of one's data is of fundamental importance when choosing a method of analysis, as will be seen in later chapters.

3.5 TWO VARIABLES

3.5.1 Describing data in two or more groups

In many studies comparisons are made between different groups. For example, two groups of patients may be given different treatments and the outcomes observed. It is desirable in such studies to demonstrate that the characteristics of the two groups of subjects were comparable at the start of the study. As an example, Table 3.5 shows the characteristics of the groups of subjects in a clinical trial comparing short-wave diathermy treatment, osteopathic treatment, and an ineffective placebo treatment in patients with non-specific low back pain (Gibson *et al.*, 1985). The characteristics of the three groups at the start of the study (often called 'baseline' values) are shown as numbers and percentages for categorical variables, and as means and standard deviations for the two continuous variables. This information

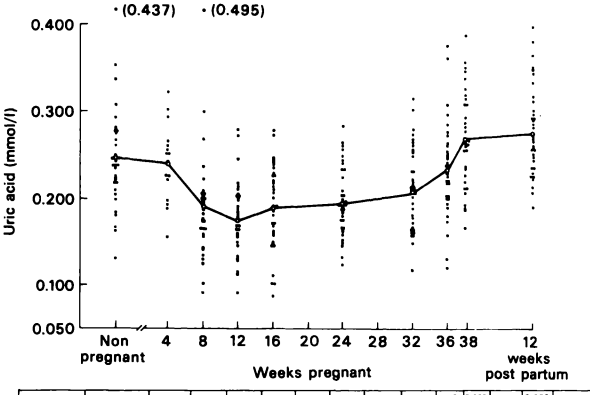
is usually sufficient to judge the comparability of the groups. I shall consider how we assess whether they *are* comparable in Chapter 15. For the moment we can see that the mean duration of pain had a skewed distribution as the mean is a lot less than twice the standard deviation in all three groups.

Table 3.5 Details of patients in each treatment group in a study of low back pain (Gibson *et al.*, 1985)

	Treatment group		
	Short-wave diathermy	Osteopathy	Placebo
Number of patients	34	41	34
Sex	16F/18M	21F/20M	11F/23M
Mean age (SD)	35 (16)	34 (14)	40 (16)
Mean duration of pain in weeks (SD)	18 (11)	16 (14)	17 (11)
Median pain score at presentation (range)*	45 (5–82)	35 (4–90)	48 (10–96)
Radiological abnormalities of the spine	12 (34%)	12 (29%)	11 (32%)

*Visual analogue scale

Sometimes we wish to show graphically the distribution of a continuous variable in two or more groups. This can be done by means of a separate histogram for each group, these being aligned vertically, but there is a rather clearer format that shows all the observations. Figure 3.14 shows the distribution of uric acid in a group of women before, during and after pregnancy (Lind *et al.*, 1984). All the data are shown in the graph, and the authors have also given the mean, standard deviation and number of observations at each stage. This informative figure thus effectively incorporates a table while using little extra space. Bar diagrams are often used to show means and standard deviations in each group. This is not a good format – this information is better in a table, or else a more informative display, such as that in Figure 3.14 or a box-and-whisker diagram, should be used.



Mean	0.246	0.240	0.190	0.173	0.189		0.196		0.216	0.232	0.269		0.275
SD	0.059	0.044	0.070	0.047	0.048		0.040		0.053	0.056	0.056		0.057
Number of observations	33	16	32	32	35		35		36	32	32		35

Figure 3.14 Distribution of serum uric acid in a group of healthy women before, during and after pregnancy (reproduced from Lind *et al.*, 1984, with permission).

3.5.2 Relation between two continuous variables

The relation between two continuous variables may be shown graphically in a **scatter diagram**. This is a simple graph in which the values of one variable are plotted against those of the other. For example, Figure 3.15 shows a scatter diagram of the PImax data of Table 3.1 related to age. Scatter diagrams are very simple to produce using statistical computer programs. When there are two (or more) individuals with identical values of both variables this should be shown, preferably by moving one point slightly. Some software packages print the actual number of coincident points up to 9, so that '9' means '9 or more'. It is easy to indicate subgroupings by using different plotting symbols. For example, in Figure 3.15 males and females could have been indicated by closed and open circles. The scatter diagram is a very useful descriptive tool, and is often valuable as a prelude to formal statistical analysis. The graph in Figure 3.14 is really a scatter diagram relating a continuous and a categorical variable.

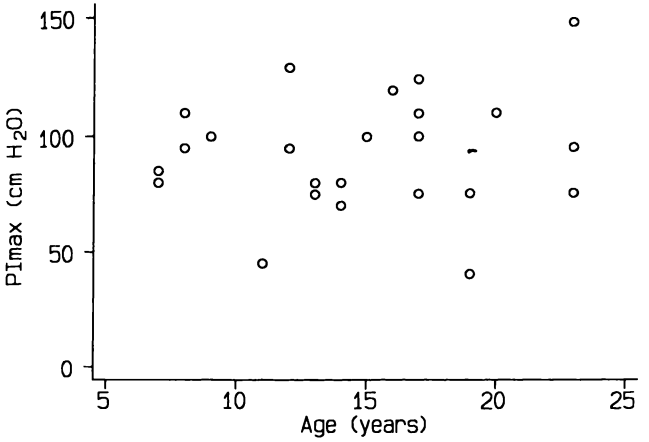


Figure 3.15 Scatter diagram of PImax by age.

3.6 THE EFFECT OF TRANSFORMING THE DATA

If we change our data in some way we will inevitably change the mean and standard deviation too. In some situations we alter, or transform, a complete set of data, in which case the effect on the mean and standard deviation may be predicted.

The simplest case to consider is where we alter the units of measurement. If we change the IgM data from values recorded as g/l to mg/l each observation will be 1000 times as large. It is easy to see that the mean will also be 1000 times bigger, and inspection of the formula for the standard deviation shows that it too will be 1000 times bigger. In contrast, if we add or subtract a constant value from all the observations, the mean of the new data is obtained by the same subtraction or addition but the standard deviation is unaffected. Thus to the mean of a set of temperatures recorded as degrees Celsius we must add 273.15 to give the mean of the equivalent thermodynamic temperature on the Kelvin scale.

Any transformation based on multiplication, division, subtraction or addition is called a **linear** transformation, because if we plot the new values against the original values we get a straight line. The mean and standard deviation of the transformed values are obtained in a simple manner. For other, non-linear transformations, however, we cannot obtain the mean and standard deviation of the transformed data in this way. Examples of non-linear transformation are taking logarithms (illustrated in section 3.4.4) or square roots. Thus the mean of the log data is not the same as the log

of the mean of the raw data. The reasons for transforming data are considered in Chapter 7.

3.7 DATA PRESENTATION

3.7.1 Numerical presentation

Data summary should not be by the mean (or median) alone, but some indication of variability should also be provided. It is common to put the SD in brackets after the mean. When these values are quoted in text the format mean ±SD, as in 'their mean diastolic blood pressure was 102.3 ± 11.9 mm Hg', should be avoided. (Indeed several medical journals no longer allow this notation.) It is much better to write 102.3 mm Hg (SD 11.9) because this format makes it clear what the second number is and also avoids the implication that the range of values from mean −SD to mean +SD is of specific importance. As we have seen, it is the range mean ±2SD which can often be used to describe the spread of the large majority (about 95%) of a set of observations.

It is not possible to give absolute rules for numerical presentation, but the following guidelines will generally be reasonable. It is usually appropriate to quote the mean to one extra decimal place compared with the raw data. The mean should not be presented to ridiculous (and spurious) 'accuracy'. For example, it is clearly absurd to quote the mean length of gestation of a group of babies to the nearest 10 minutes. This is done when quoting weeks of gestation to 3 decimal places. The standard deviation should usually be given to the same accuracy as the mean, or with one extra decimal place.

3.7.2 Tables

Whether or not to put descriptive data in tables will depend on the number of variables and groups of subjects. Table 3.5 shows a recommended way of presenting descriptive data, both continuous and categorical. In general it is preferable to put data of a like kind in *columns* rather than *rows* as the eye can scan columns more easily, but this is not always possible. For example, in Table 3.5 the means of the same variables in the three treatment groups are shown in rows, as it is usually more natural that way. However, means and SDs are clearly distinguished side by side, with the latter in brackets for clarity.

Tables can also be used to show raw data, although this is only reasonable when there are not too many observations. Where possible, it is helpful to order the data by one of the variables – after all, there is usually nothing special about the order in which the patients were seen. Many of the tables in this book, such as Table 3.1, have been ordered in this way.

3.7.3 Graphs

It is difficult to offer much general advice about when it is appropriate to use a graph rather than a table. Graphs offer the opportunity to show much more data than could be shown in a table, and are thus probably most suited to data that cannot easily be displayed in a table. There is no point in using a graph to show, for example, the means and standard deviations of one variable in two or three groups. Some displays, such as histograms, are in essence graphical – Figure 3.3 is a much clearer display than Table 3.2. It is possible to combine the best features of a table and a figure, and an example was given in Figure 3.14. This form of display should be used more often.

Scatter diagrams are particularly useful for showing the relation between two variables. It is important that all the data points should be shown, which can pose difficulties when there are coincident points (see section 6.7). Different symbols can be used to indicate subgroups of the data.

Graphs are a very powerful way of getting a message across, but the same data can be portrayed in many ways, with a variety of visual effects. For example, Figure 3.16 shows three alternative displays of the data in Table 3.6 showing average amounts of bread consumed per person per week in London from 1960 to 1980. Features visible in one or more figures include a gradual reduction in total bread consumption, a more than proportionate fall in consumption of white bread, and a rise in consumption of brown and wholemeal bread in the last five year period. These features are probably more easily seen in Table 3.6.

Table 3.6 Amounts of bread consumed in London from 1960 to 1980 (g per person per week) (Sivell and Wenlock, 1983)

Type of bread	Year				
	1960	1965	1970	1975	1980
White	1040	975	915	785	620
Brown	70	80	70	75	115
Wholemeal	25	20	15	20	45
Other	155	80	85	75	105
Total	1290	1155	1080	955	880

An excellent book on graphical methods in general is that by Tufte (1983), and graphs for statistics are discussed by Moses (1987). Many innovative ideas for descriptive methods are described by Tukey (1977).

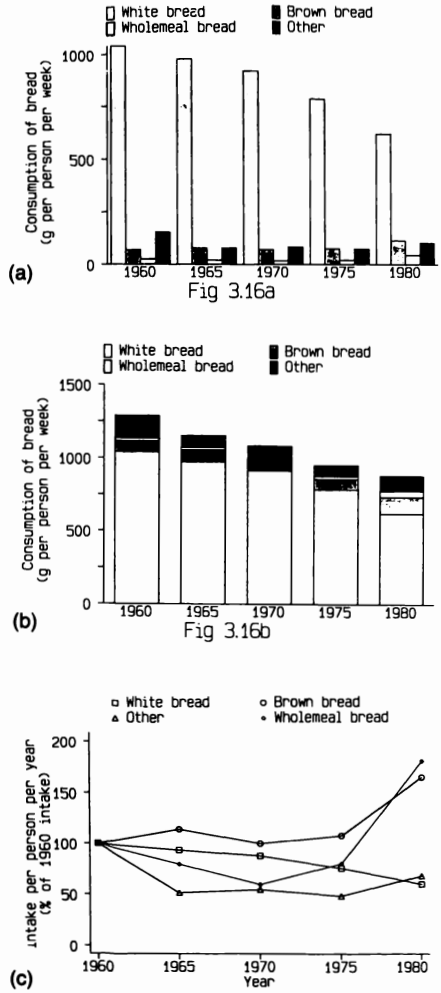


Figure 3.16 Average amount of bread consumed per person per week in London from 1960 to 1980; three alternative graphs of the data in Table 3.6: (a) adjacent bars, (b) stacked bars, (c) graph of relative changes since 1960.

EXERCISES

3.1 The table overleaf shows some data for 65 patients with rheumatoid arthritis treated with sodium aurothiomalate (SA) (Ayesh *et al.*, 1987). The total dose of SA is shown, together with values of the sulphoxidation index (SI), which measures the capacity to convert organic divalent alkyl sulphide to its corresponding sulphoxide form. The patients have been separated into 28 without and 37 with major adverse reactions to the drug.

(a) Some values of SI are given as '>80.0'. What is the name given to observations like this?
(b) What is the difficulty about drawing histograms of SI in each group? What shape are the distributions?
(c) Give two reasons why it is preferable to calculate the median rather than the mean to describe the average SI value.
(d) Obtain the median SI for each group of patients. (This should take less than ten seconds.)
(e) Obtain the median total dose of SA for the group with adverse reactions.
(f) Produce stem-and-leaf diagrams to compare the age distributions in the two groups.
(g) Do the data support the idea that patients experiencing adverse reactions were on average older than those without adverse reactions?

	Without adverse reactions				With adverse reactions		
	Age	Total dose of SA (mg)	SI		Age	Total dose of SA (mg)	SI
1	44	1560	1.0	1	53	360	2.0
2	65	1310	1.2	2	74	2010	2.0
3	58	850	1.2	3	29	1390	2.0
4	57	1250	1.7	4	53	660	3.0
5	51	950	1.8	5	67	1135	3.5
6	64	850	1.8	6	67	510	5.3
7	33	1200	1.9	7	54	410	5.7
8	61	1390	2.0	8	51	910	6.5
9	49	1450	2.3	9	57	360	13.0
10	67	3300	2.8	10	62	1260	13.0
11	39	2760	2.8	11	51	560	13.9
12	42	860	3.4	12	68	1135	14.7
13	35	1810	3.4	13	50	1410	15.4
14	31	1310	3.8	14	38	1110	15.7
15	37	1250	3.8	15	61	960	16.6
16	43	1210	4.2	16	59	1310	16.6
17	39	1460	4.9	17	68	910	16.6
18	53	2310	5.4	18	44	1235	22.0
19	44	1360	5.9	19	57	2950	22.3
20	41	1910	6.2	20	49	360	33.2
21	72	910	12.0	21	49	1935	47.0
22	61	1410	18.8	22	63	1660	61.0
23	48	2460	47.0	23	29	435	65.0
24	59	1350	70.0	24	53	310	65.0
25	72	810	>80.0	25	53	310	>80.0
26	59	1460	>80.0	26	49	410	>80.0
27	71	760	>80.0	27	42	690	>80.0
28	53	910	>80.0	28	44	910	>80.0
				29	59	1260	>80.0
				30	51	1260	>80.0
				31	46	1310	>80.0
				32	46	1350	>80.0
				33	41	1410	>80.0
				34	39	1460	>80.0
				35	62	1535	>80.0
				36	49	1560	>80.0
				37	53	2050	>80.0

3.2 (a) Does Figure 3.1 indicate that professional pilots are more likely to have an aviation accident than other groups?

The following table shows the data that were plotted in Figure 3.1, together with the aviation accident rates per 100 000 hours of recent flight time (Booze, 1977).

	Number of accidents	Rate per 1000*	Rate per 100 000 hr
Professional pilots	1302	15.9	0.2
Lawyers	57	11.0	1.5
Farmers	166	10.1	1.3
Sales representatives	137	9.0	1.2
Physicians	76	8.7	1.8
Mechanics and repairmen	44	6.9	1.5
Policemen and detectives	48	6.6	1.8
Managers and administrators	643	6.0	0.7
Engineers	125	4.7	1.1
Teachers	43	4.2	1.1
Housewives	29	3.7	3.2
Academic students	188	3.2	3.7
Armed Forces Members	111	1.6	0.7

*in the specified occupation

(b) The rates per 100 000 hours can also be made into a bar diagram. From such a diagram, or from the figures shown in the table, which two groups of pilots had most accidents? Why do the two sets of figures give different answers? (A scatter diagram is useful to see the relation between the two.)

3.3 Calculate the centiles used to construct the box-and-whisker plot in Figure 3.12 using the method of calculation given in section 3.4.2.

repair capabilities as a genetic marker for susceptibility to ultraviolet light–induced nonmelanoma skin cancer in young cases and controls.[24] Regrettably, however, most environmental exposures that can be assessed by means of objective biologic markers represent current or very recent rather than past exposures, such as the levels of serum cotinine to indicate exposure to cigarette smoking, and are thus of limited usefulness.[25]

When a well-defined cohort is available, nested case-control or case-cohort studies (see Chapter 1, Section 1.4.2) allow the evaluation of certain hypotheses free of recall bias (or temporal bias; see Section 4.3.3). Typically, in these case-control studies, information on exposure and confounders is collected at baseline (i.e., before the incident cases occur), thus reducing the likelihood of systematic recall differences between cases and controls. The study discussed previously examining the relationship of tanning ability to melanoma[18] (see previously here and Section 4.3.3) is an example of a case-control study within a cohort (the Nurses' Health Study cohort); the application of the premelanoma diagnosis questionnaire avoids the recall bias that was observed when the analysis was based on information obtained from the postmelanoma questionnaire.[18]

Although exposure recall bias is typically a problem of case-control studies, it may also occur in cohort studies. In the latter type of study, it may be present at the outset of the study when categorization of individuals by level of exposure relies on recalled information from the distant or recent past, as when attempts are made to classify cohort participants at baseline by duration of exposure.

Interviewer Bias

When data collection in a case-control study is not masked with regard to the disease status of study participants, *observer bias* in ascertaining exposure, such as *interviewer bias*, may occur. Interviewer bias may be a consequence of trying to "clarify" questions when such clarifications are not part of the study protocol and failing to follow either the protocol-determined probing or skipping rules of questionnaires. Although it is often difficult to recognize interviewer bias, it is important to be aware of it and to implement procedures to minimize the likelihood of its occurrence. Attempts to prevent interviewer bias involve the careful design and conduct of quality assurance and control activities (see Chapter 8), including development of a detailed manual of operations, training of staff, standardization of data collection procedures, and monitoring of data collection activities. Even when these methods are in place, however, subtle deviations from the protocol (e.g., emphasizing certain words when carrying out the case but not the control interviews or vice versa) might be difficult to identify. Additional measures to recognize and prevent this bias are the performance of reliability/validity substudies and the masking of interviewers with regard to case-control status.

Reliability and validity substudies in samples are described in more detail in Chapter 8, Section 8.3. They constitute an important strategy that needs to be carried out systematically with quick feedback to interviewers who do not follow the protocol or who have encountered problems. Reliability substudies of interviews are not as straightforward as those aimed at assessing the reproducibility of laboratory measurements, such as those described in many of the examples in Chapter 8. Assessing the reliability of interview data is difficult because of intraparticipant variability and because, when interviews are done at separate points in time, interviewees or interviewers may recall previous responses, with the resultant tendency to provide/record the same, albeit mistaken, responses.

As for recall bias, validity studies using independent sources (e.g., medical charts) can be conducted to assess accuracy of data collection by interviewers.

Masking of interviewers with regard to case-control status of study participants is difficult, but when feasible, it may remove an important source of bias, particularly when the interviewer is

familiar with the study hypothesis. On occasion, by including a health question for which a frequent affirmative response is expected from both cases and controls, it is possible to mask the interviewers with regard to the main study hypothesis and have them believe that the hypothesis pertains to the "misleading" question. Such a strategy was employed in a case-control study of psychosocial factors and myocardial infarction in women in which questions about hysterectomy, which were often answered positively in view of the high frequency of this intervention in the United States, led the interviewers to believe that the study was testing a hormonal hypothesis.[26]

A mistake made in an early study of lung cancer and smoking conducted by Doll and Hill[27] in which some controls were erroneously classified as cases provided an unplanned opportunity to assess the possible occurrence of interviewer bias. In this study, the odds of exposure to smoking in the misclassified controls was very similar to that of the nonmisclassified controls and much lower than that of cases, thus confirming the absence of interviewer bias. This example suggests the possibility of assessing interviewer bias by using "phantom" cases and controls and/or purposely misleading interviewers to believe that some cases are controls and vice versa.

4.3.2 Outcome Identification Bias

Outcome (e.g., disease) *identification bias* may occur in both case-control and cohort studies. This bias may result from either differential or nondifferential misclassification of disease status, which in turn may be due to an imperfect definition of the outcome or to errors at the data collection stage.

Observer Bias

In a cohort study, the decision as to whether the outcome is present may be affected by knowledge of the exposure status of the study participant. This may happen particularly when the outcome is "soft" or subjective, such as, for example, when reporting migraine episodes or psychiatric symptoms. There may be *observer bias* at different stages of the ascertainment of the outcome, including at the stage of applying pathologic or clinical criteria. A fairly crude example of observer bias is the assignment of a histologic specimen to a diagnosis of "alcoholic cirrhosis" when the pathologist knows that the patient is an alcoholic. A documented example of observer bias is the effect of the patient's race on the diagnosis of hypertensive end-stage renal disease (ESRD). In a study conducted by Perneger et al.,[28] a sample of nephrologists were sent case histories of seven patients with ESRD. For each case history, the simulated race of each patient was randomly assigned to be "black" or "white." Case histories that identified the patient's race as black were twice as likely to result in a diagnosis of hypertensive ESRD as case histories in which the patient's race was said to be white.

This type of observer bias occurs when the ascertainment of outcome is not independent from the knowledge of the exposure status and results in *differential* misclassification of the outcome. Thus, measures aimed at *masking observers in charge of deciding whether the outcome is present by exposure status* would theoretically prevent observer bias. When masking of observers by exposure status is not practical, observer bias can be assessed by stratifying on certainty of diagnosis. For example, exposure levels can be assessed in relationship to incidence of "possible," "probable," or "definite" disease. Observer bias should be suspected if an association is seen for only the "softer" categories (e.g., possible disease).

Another strategy to prevent observer bias is to perform diagnostic classification with *multiple observers*. For example, two observers could independently classify an event, and if disagreement occurred, a third observer would adjudicate; that is, decision on the presence or absence of the

outcome would have to be agreed on by at least two of three observers. This is the strategy used to classify events such as acute myocardial infarction and stroke in both the Atherosclerosis Risk in Communities (ARIC) Study and the Multi-Ethnic Study of Atherosclerosis (MESA).[21,29]

Respondent Bias

Recall and other informant biases are usually associated with identification of exposure in case-control studies; however, outcome ascertainment bias may occur during follow-up of a cohort when information on the outcome is obtained by participant response, for example, when collecting information on events for which it is difficult to obtain objective confirmation, such as episodes of migraine headaches.

Whenever possible, information given by a participant on the possible occurrence of the outcome of interest should be confirmed by more objective means, such as hospital chart review. Objective confirmation may, however, not be possible, for example, for nonhospitalized events or events in which laboratory verification is impossible, such as pain or acute panic attacks. For these types of outcomes, detailed information not only on presence versus absence of a given event but also on related symptoms that may be part of a diagnostic constellation may be of help in preventing *respondent bias*. For example, the questionnaire on the occurrence of an episode of migraine headaches in a study by Stewart et al.[30] included questions not only on whether a severe headache had occurred but also on the presence of aura, nausea, and fatigue accompanying the headache. This strategy allowed more objectivity in classifying migraines than the simple determination of the presence or absence of pain. For several outcomes, such as angina pectoris and chronic bronchitis, standardized questionnaires are available (see Chapter 8). Other soft outcomes are diagnosed by using symptom scales, such as the Center for Epidemiologic Studies Depression Scale, which has a good correlation with psychiatrist-diagnosed depression.[31] The validity and limitations of some of these instruments, such as, for example, the Rose Questionnaire for the diagnosis of angina pectoris,[32] have been assessed.[33-35]

4.3.3 The Result of Information Bias: Misclassification

Information bias leads to *misclassification* of exposure and/or outcome status. For example, when there is recall bias in a case-control study, some exposed subjects are classified as unexposed and vice versa. In a cohort study, a positive outcome may be missed. Alternatively, a pseudo-event may be mistakenly classified as an outcome (a "false positive"). The examples of both differential and nondifferential misclassification in this section refer to exposure levels in case-control studies. Misclassification of case-control status in case-control studies and of exposure and outcome in cohort studies can be readily inferred, although they are not specifically discussed to avoid repetition.

There are two types of misclassification bias: nondifferential and differential.

Nondifferential Misclassification

In a case-control study, *nondifferential misclassification* occurs when the degree of misclassification of exposure is independent of case-control status (or vice versa).

Nondifferential Misclassification When There Are Two Categories. A simplistic hypothetical example of nondifferential misclassification of (dichotomous) exposure in a case-control study is shown in **EXHIBIT 4-2**. In this example, misclassification of exposed subjects as unexposed occurs

EXHIBIT 4-2 Hypothetical example of the effect of nondifferential misclassification of two categories of exposure with 30% of both exposed cases and exposed controls misclassified as unexposed.

No misclassification		
Exposure	**Cases**	**Controls**
Yes	50	20
No	50	80

$$OR = \frac{\left(\frac{50}{50}\right)}{\left(\frac{20}{80}\right)} = 4.0$$

30% Exposure misclassification in each group		
Exposure	**Cases**	**Controls**
Yes	50 – **15** = 35	20 – **6** = 14
No	50 + **15** = 65	80 + **6** = 86

$$OR = \frac{\left(\frac{35}{65}\right)}{\left(\frac{14}{86}\right)} = 3.3$$

Effect of nondifferential misclassification with two exposure categories: to bias the OR toward the null value of 1.0. (It "dilutes" the association.)

Bold numbers represent misclassified individuals.

in 30% of cases and 30% of controls. In this simple situation when there are only two exposure categories (for instance, "yes" or "no"), nondifferential misclassification *tends to bias the association toward the null hypothesis*.

In the hypothetical example shown in Exhibit 4-2, misclassification occurs in only one direction: Exposed individuals are misclassified as unexposed. Often, however, misclassification occurs in both directions; that is, exposed individuals are classified as unexposed or "false negatives" (i.e., the correct classification of the truly exposed, or sensitivity, is less than 100%), and unexposed individuals are classified as exposed or false positives (i.e., the correct classification of the unexposed, or specificity, is less than 100%). In a case-control study, nondifferential misclassification occurs when *both*

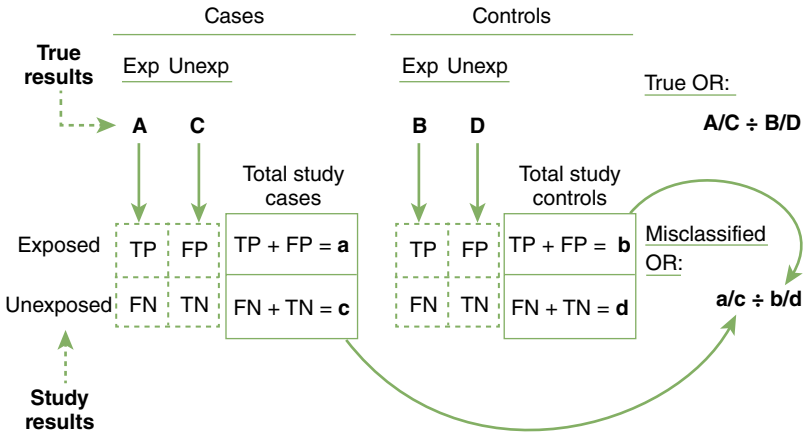


FIGURE 4-4 Application of sensitivity/specificity concepts in misclassification of exposure: schematic representation of true and misclassified relative odds. Sensitivity of exposure ascertainment = TP ÷ (TP + FN); specificity of exposure ascertainment = TN ÷ (TN + FP).

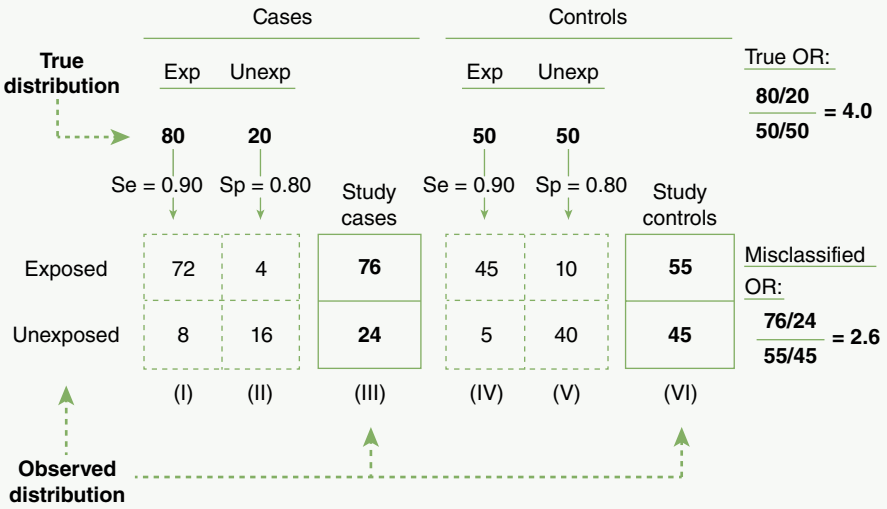
Exp, exposed; FN, false negative; FP, false positive; TN, true negative; TP, true positive; Unexp, unexposed.

the sensitivity *and* specificity of the classification of exposure are the same for cases and controls but either (or both) is less than 100%. Estimation of the total numbers of individuals classified as exposed or unexposed by using a study's data collection procedures and exposure level definitions is akin to the estimation of "test-positive" and "test-negative" individuals when applying a screening test. Thus, the notions of sensitivity and specificity, schematically represented in **FIGURE 4-4**, can be used to explore the issue of misclassification in more depth.

A hypothetical example showing nondifferential misclassification of exposure in a case-control study in both directions—that is, when exposed subjects are misclassified as unexposed and unexposed subjects are misclassified as exposed—is presented in **EXHIBIT 4-3**. The exhibit shows the effects of nondifferential misclassification resulting from an exposure ascertainment with a sensitivity of 90% and a specificity of 80%. The fact that these sensitivity and specificity values are the same for cases and controls identifies this type of misclassification as nondifferential.

The net effect of misclassifying cases at a sensitivity of 90% and a specificity of 80% is shown in column (III) of Exhibit 4-3. The totals in column (III) indicate the numbers of cases classified as exposed or unexposed in the study and reflect the misclassification due to the less-than-perfect sensitivity and specificity values. Thus, cases classified as exposed include both the 72 persons truly exposed (true positives) and the 4 cases which, although unexposed, are misclassified as exposed (false positives) due to a specificity less than 100% (see also Figure 4-4). Similarly, cases classified in the study as unexposed include both the 16 truly unexposed cases (true negatives) and the 8 exposed cases misclassified as unexposed (false negatives) because the sensitivity is less than 100%. Exhibit 4-3 also shows similar data for *controls*. The net effect of the classification of controls by exposure at the same sensitivity (90%) and specificity (80%) levels as those of cases is shown in column (VI). The observed (biased) odds ratio of 2.6 in the study underestimates the true odds ratio of 4.0, as expected when misclassification of a dichotomous exposure is nondifferential between cases and controls.

EXHIBIT 4-3 Effects of nondifferential misclassification on the odds ratio (sensitivity = 0.90; specificity = 0.80).

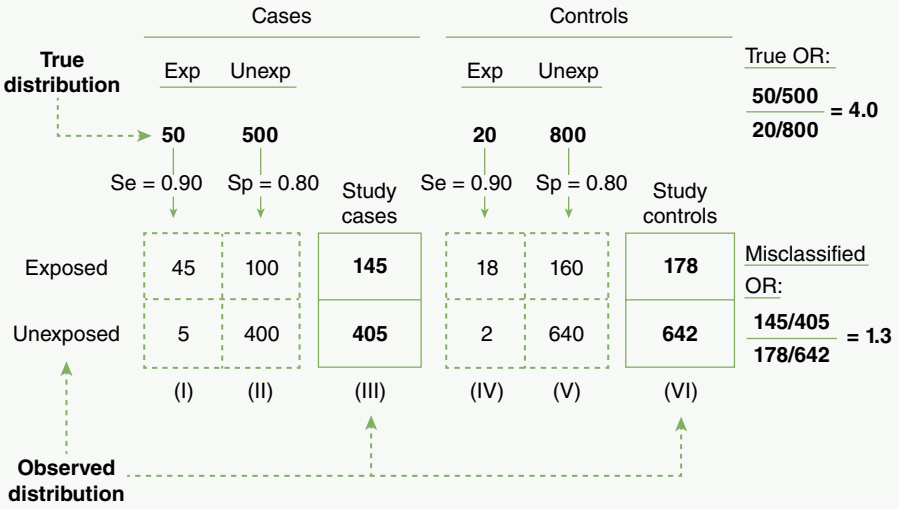


Exp, exposed; OR, odds ratio; Se, sensitivity; Sp, specificity; Unexp, unexposed.

In the example shown in Exhibit 4-3, nondifferential misclassification of a dichotomous exposure is shown to be affected by sensitivity and specificity levels, such that the net effect is to bias the odds ratio toward 1.0. In addition to reflecting sensitivity and specificity of the procedures for exposure definition and ascertainment, the magnitude of the bias also depends on the exposure prevalence, particularly in the presence of a large control group. For example, **EXHIBIT 4-4** shows a hypothetical situation where the true strength of the association between exposure and disease is identical to that in Exhibit 4-3 (odds ratio = 4.0), as are the sensitivity and specificity of exposure measurement (90% and 80%, respectively). However, because of the lower prevalence of exposure (i.e., 20/820 or 2.4% among controls, compared to 50% in Exhibit 4-3), the bias is substantially more pronounced (biased odds ratio = 1.3 versus 2.6 in Exhibit 4-3). In general, low exposure prevalence tends to be associated with a higher degree of bias when the specificity is low. If specificity is high but sensitivity is low, however, a higher degree of bias will result from a situation in which exposure is common. The complex relationships between bias and sensitivity/specificity of exposure definition and its prevalence are illustrated in **TABLE 4-6**, showing examples of the effects of sensitivity, specificity, and exposure prevalence in controls on the observed odds ratio in several hypothetical situations where the true odds ratio is 4.0.

Nondifferential Misclassification When There Are More Than Two Exposure Categories. The rule that the direction of a nondifferential misclassification bias dilutes the strength of the association may not hold in certain nondifferential misclassification situations involving more than two exposure categories. A hypothetical example involving three exposure levels in a case-control study ("none," "low," and "high") is discussed by Dosemeci et al.[36] (**TABLE 4-7**). In this example, 40% of both cases and controls in the "high" exposure category were misclassified as belonging to the

EXHIBIT 4-4 Effects of nondifferential misclassification on the odds ratio when the exposure prevalence in controls is low.



Exp, exposed; OR, odds ratio; Se, sensitivity; Sp, specificity; Unexp, unexposed.

TABLE 4-6 Nondifferential misclassification: hypothetical examples of the effects of sensitivity and specificity of exposure identification and of exposure prevalence in controls on a study's odds ratio when the true odds ratio is 4.0.

Sensitivity*	Specificity†	Prevalence of exposure in controls	Observed odds ratio
0.90	0.85	0.200	2.6
0.60	0.85	0.200	1.9
0.90	**0.95**	0.200	3.2
0.90	**0.60**	0.200	1.9
0.90	0.90	**0.368**	3.0
0.90	0.90	**0.200**	2.8
0.90	0.90	**0.077**	2.2

Bold figures represent the factor (sensitivity, specificity, or exposure prevalence) that is allowed to vary for fixed values of the other two factors.
*Sensitivity of the exposure identification is defined as the proportion of all truly exposed correctly classified by the study.
†Specificity of the exposure identification is defined as the proportion of all truly unexposed correctly classified by the study.

TABLE 4-7 Examples of the effects of nondifferential misclassification involving three exposure categories; misclassification of 40% between "high" and "low" (A) and between "high" and "none" (B).

Case-control status	True exposure status		
	None	Low	High
Cases	100	200	600
Controls	100	100	100
Odds ratio	1.00	2.00	6.00
	Misclassified exposure status (in situations A and B)		
A. Adjacent categories: 40% of cases and controls in "high" misclassified as "low"			
Cases	100	200 CC + 240 MC = 440	600 CC – 240 MC = 360
Controls	100	100 CC + 40 MC = 140	100 CC – 40 MC = 60
Odds ratio	1.00	3.14	6.00
B. Nonadjacent categories: 40% of cases and controls in "high" misclassified as "none"			
Cases	100 CC + 240 MC = 340	200	600 CC – 240 MC = 360
Controls	100 CC + 40 MC = 140	100	100 CC – 40 MC = 60
Odds ratio	1.00	0.82	2.47

CC, correctly classified; MC, misclassified.
Data from Dosemeci M, Wacholder S, Lubin JH. Does nondifferential misclassification of exposure always bias a true effect toward the null value? *Am J Epidemiol.* 1990;132:746-748.[36]

adjacent category, "low"; the net effect was an increase in the odds ratio for the "low" category without a change for the "high." Misclassification for nonadjacent categories of exposure in the example—that is, between "high" and "none"—resulted in the disappearance of the truly graded relationship and, assuming no random error, the emergence of a J-shaped pattern. Additionally, as shown by Dosemeci et al.,[36] misclassification of nonadjacent exposure categories may invert the direction of the graded relationship.

Differential Misclassification

Differential misclassification occurs when the degree of misclassification differs between the groups being compared; for example, in a case-control study, the sensitivity and/or the specificity of the

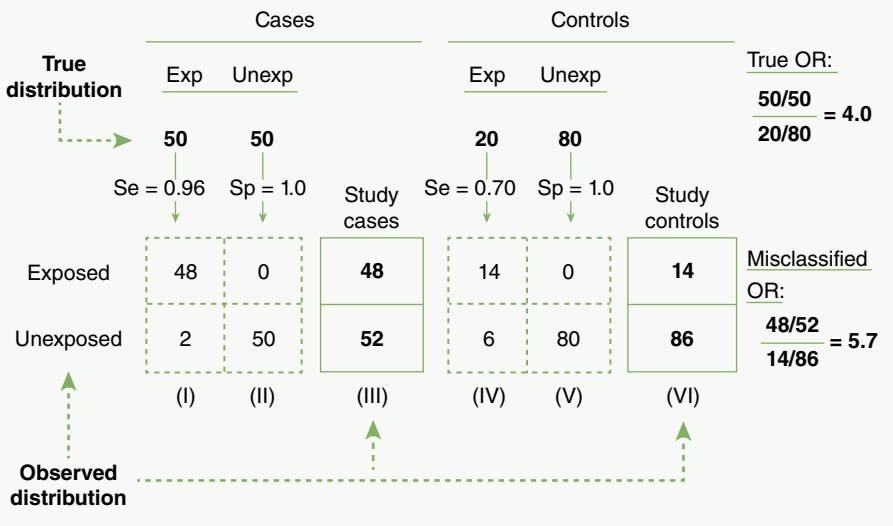
classification of exposure status is different between cases and controls. (Note that differential misclassification may occur even when only one of these validity indices differs.) In a cohort study, differential misclassification will occur when the accuracy of outcome definition differs between exposed and nonexposed.

Whereas the general tendency of nondifferential misclassification of a dichotomous exposure factor is to weaken a true association, differential misclassification may bias the association either toward or away from the null hypothesis. Thus, it is difficult to predict the direction of the bias when differential misclassification occurs, as it is the result of a complex interplay involving differences between cases and controls in sensitivity, specificity, and prevalence of exposure.

A hypothetical example of differential misclassification in a case-control study is given in **EXHIBIT 4-5** in which the sensitivity of capturing the exposure in cases is 96% and that in controls is only 70%. Specificity in the example is 100% for both cases and controls. The better sensitivity among cases leads to a higher proportion of truly exposed subjects being identified in cases than in controls, yielding a biased odds ratio further away from 1.0 than the true odds ratio (true odds ratio = 4.0; biased odds ratio = 5.7). To underscore the difficulties in predicting results when there is differential misclassification, if the same calculations are done using a higher specificity in cases (100%) than in controls (80%), the odds ratio is biased toward the null hypothesis (**EXHIBIT 4-6**), as a poorer specificity in controls offsets the higher sensitivity in cases.

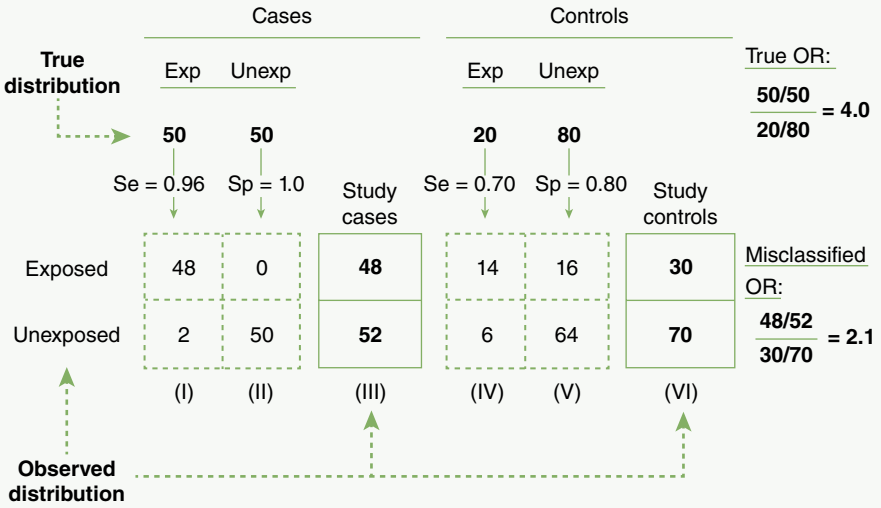
EXHIBIT 4-7 shows a hypothetical example that illustrates a shortcut to the calculation of misclassified odds ratios. The table shows the complements to sensitivity and specificity values and their application to the relevant cells in cases and controls. In this example, misclassification is differential, and yet the misclassified odds ratio is biased toward the null hypothesis.

EXHIBIT 4-5 Hypothetical example of the effects of differential misclassification on the odds ratio in which, for sensitivity, cases > controls and, for specificity, cases = controls.



Exp, exposed; OR, odds ratio; Se, sensitivity; Sp, specificity; Unexp, unexposed.

EXHIBIT 4-6 Hypothetical example of the effects of differential misclassification on the odds ratio in which, for both sensitivity and specificity, cases > controls.



Exp, exposed; OR, odds ratio; Se, sensitivity; Sp, specificity; Unexp, unexposed.

EXHIBIT 4-7 Shortcut calculation of misclassified odds ratios in a case-control study. Exposure information sensitivity: cases = 0.96; controls = 0.85; specificity: cases = 0.80 and controls = 0.70. Application of complements of sensitivity and specificity values estimates the number of false negatives and false positives in each exposure category. For example, 1 – sensitivity (0.04) for cases results in 2 exposed cases being misclassified as unexposed (false negatives); 1 – specificity for cases (0.20) results in 10 unexposed cases being misclassified as exposed (false positives). Similar calculations are done for controls.

	Cases (n = 100)				Controls (n = 100)			
		1 – Se	1 – Sp			1 – Se	1 – Sp	
	True distribution	0.04	0.20	Total misclassified	True distribution	0.15	0.30	Total misclassified
Exposed	50	–2	+10 (FP)	58	20	–3	+24 (FP)	41
Unexposed	50	+2 (FN)	–10	42	80	+3 (FN)	–24	59

This differential misclassification biases the odds ratios toward the null hypothesis.

True odds ratio : $\frac{50/50}{20/20} = 4.0$

Misclassified odds ratio: $\frac{58/42}{41/59} = 1.98$

FN, false negative; FP, false positive; Se, sensitivity; Sp, specificity.

Examples of the isolated effects of sensitivity (for a specificity of 100%) or specificity (for a sensitivity of 100%) on the odds ratio in a hypothetical case-control study with differential misclassification of exposure and a control exposure prevalence of 10% are shown in **TABLE 4-8**.

An example of differential misclassification of exposure was documented by Weinstock et al.[18] This example was used previously here to illustrate the concept of recall bias (see Section 4.3.1). In this study, melanoma cases and controls selected from participants of the Nurses' Health Study cohort and matched for duration of follow-up were compared with regard to their report of "hair color" and "tanning ability" both at baseline and after the case was diagnosed. In this example, the differential misclassification in the postdiagnosis interview probably occurred because the disease status was known to the case and had the potential to affect recall of exposure. Therefore, the premelanoma diagnosis interview is assumed to accurately reflect the true association. The main results of the study are summarized in **TABLE 4-9**. The discussion that follows focuses on the "exposure" categories that were found associated with an increase in odds using the case-control data obtained *after* the occurrence of melanoma ("red or blond" and "no tan or light tan" for hair color and tanning ability, respectively).

Compared with the predisease development data, the odds for hair color among cases did not change when the postmelanoma interview data were used (11:23 in both interviews) and increased only slightly among controls (from 37:197 to 41:193); as a result, the odds ratio changed relatively little (prediagnosis odds ratio = 2.5; postdiagnosis odds ratio = 2.3). The effect of differential misclassification of tanning ability, however, was severe, leading to a reversal of the direction of the association. Assuming no random variability, the true association (i.e., that detected using the premelanoma diagnosis information) suggests a protective effect (odds ratio = 0.7), whereas the observed postdiagnosis association (odds ratio = 1.6) indicates a greater melanoma odds associated with a low tanning ability. It is of interest that the misclassification of exposure as

TABLE 4-8 Examples of the effects of differential sensitivity and specificity of exposure ascertainment on the odds ratio (OR) for a true OR of 3.86 and a control exposure prevalence of 0.10.

Exposure ascertainment				Odds ratio
Sensitivity*		Specificity†		
Cases	Controls	Cases	Controls	
0.90	0.60	1.00	1.00	5.79
0.60	0.90	1.00	1.00	2.22
1.00	1.00	0.90	0.70	1.00
1.00	1.00	0.70	0.90	4.43

*Sensitivity of the exposure identification is defined as the proportion of all truly exposed correctly classified by the study.
†Specificity of the exposure identification is defined as the proportion of all truly unexposed correctly classified by the study.

TABLE 4-9 Reported hair color and tanning ability among incident cases and controls in a case-control study of melanoma within the Nurses' Health Study cohort.

	Premelanoma diagnosis information (gold standard)		Postmelanoma diagnosis information	
	Cases	Controls	Cases	Controls
Hair color				
Red or blond (exposed)	11	37	11	41
Brown or black (unexposed)	23	197	23	193
Odds ratio	2.5		2.3	
Tanning ability				
No tan, practically no tan, or light tan (exposed)	9	79	15	77
Medium, average, deep, or dark tan (unexposed)	25	155	19	157
Odds ratio	0.7		1.6	

Data from Weinstock MA, Colditz GA, Willett WC, et al. Recall (report) bias and reliability in the retrospective assessment of melanoma risk. *Am J Epidemiol.* 1991;133:240-245.[18]

measured by tanning ability seems to have resulted in only a slight change in odds of exposure in controls (from 79:155 to 77:157). In cases, however, the misclassification effect was substantial, with the number of individuals classified as exposed increasing from 9 to 15 between the first and second interviews.

The cross-tabulation of the premelanoma and postmelanoma diagnosis data enables a more detailed analysis of this situation by the calculation of sensitivity and specificity of tanning ability ascertainment in cases.* As shown in **TABLE 4-10**, the sensitivity of 89% of the postmelanoma diagnosis interviews resulted in the correct classification of eight of the nine truly exposed cases. However, a specificity of only 72% led to a relatively large number of unexposed persons in the false-positive cell and thus to a marked increase in the postdiagnosis exposure odds (true exposure

*In the paper by Weinstock et al.,[18] sensitivity and specificity of postdiagnosis responses on tanning ability were provided for only cases.

TABLE 4-10 Distribution of incident cases in the Nurses' Health Study cohort, 1976 to 1984, according to responses given with regard to tanning ability prior to the development of melanoma and after diagnosis was made.

Postmelanoma diagnosis information	Premelanoma diagnosis information (gold standard)		
	No tan, practically no tan, or light tan (exposed)	Medium, average, deep, or dark tan (unexposed)	Total (case-control classification)
No tan, practically no tan, or light tan (exposed)	8 (TP)	7 (FP)	15
Medium, average, deep, or dark tan (unexposed)	1 (FN)	18 (TN)	19
Total (true classification)	9 Sensitivity: 8/9 = 89%	25 Specificity: 18/25 = 72%	34

FN, false negatives; FP, false positives; TN, true negatives; TP, true positives.
Data from Weinstock MA, Colditz GA, Willett WC, et al. Recall (report) bias and reliability in the retrospective assessment of melanoma risk. *Am J Epidemiol.* 1991;133:240-245. [18]

odds in cases, 9:25 or 0.36:1.0; biased exposure odds, 15:19 or 0.79:1.0). Such change resulted in an odds ratio in the postdiagnosis study in a direction opposite to that of the true value. As mentioned previously (Section 4.3.1), differential misclassification in the study by Weinstock et al.[18] probably occurred because of recall bias. Additional misclassification may have occurred because the questions on hair color and tanning ability were not exactly the same in the interviews conducted before and after diagnosis. The latter, however, would be expected to result in nondifferential misclassification (equally affecting cases and controls). (In addition, if the misclassification had been nondifferential, that is, if the sensitivity and specificity values observed among the cases had been the same in controls, the odds ratio would have changed from the true [premelanoma diagnosis] value of 0.7 to a misclassified value of 0.83, that is, an estimate of the association biased toward the null value.)

An example of both differential and nondifferential misclassification is given by a case-control study of childhood acute lymphoblastic leukemia conducted by Infante-Rivard and Jacques.[37] Four hundred ninety-one cases and two sets of age-, sex-, and broad geographic area–matched controls were chosen, one set from a population sample and the other from among hospital patients. For each individual in the study, the authors measured the actual distance between the residence and the nearest power line and compared this distance with the parent's answer to the question, "Within a radius of 1 km of your house, was there a high-voltage power line?" The authors classified cases into those living in a geographic area where people were concerned about an excess of the disease ("GA" cases) and "other" cases. When comparing GA cases with either population or hospital controls, substantial differential misclassification was detected, with a higher sensitivity but lower specificity

TABLE 4-11 Sensitivity and specificity of response by parent of childhood (age 9 years or less) acute lymphoblastic leukemia cases to question, "Within a radius of 1 km (1000 m) of your house, was there a high-voltage power line?" Montreal Island, Quebec, Canada, 1980–1993.

	Sensitivity (%)*	Specificity (%)*
GA cases†	61.9	54.4
Other cases	34.9	90.6
Population controls	22.2	89.4
Hospital controls	35.8	90.2

*Gold standard: measured distance.
†Cases living in a geographic area where people were concerned about an excess of acute lymphoblastic leukemia cases.
Data from Infante-Rivard C, Jacques L. Empirical study of parental recall bias. *Am J Epidemiol*. 2000;152:480-486.[37]

seen for GA cases (**TABLE 4-11**). If, to calculate the odds ratio, analyses were limited to "other" cases versus hospital controls, however, nondifferential misclassification would have resulted, as their sensitivity and specificity values were found to be almost the same.

Effect of Misclassification of a Confounding Variable

Misclassification also affects the efficiency of adjustment for confounding effects. Whereas a nondifferential misclassification of a potential risk factor tends to bias the measure of association toward the null hypothesis, nondifferential misclassification of a confounding variable results in an imperfect adjustment when that variable is matched or controlled for in the analyses (see Chapters 5 and 7).[38] This imperfect adjustment results in residual confounding (see Chapter 5, Section 5.5.4, and Chapter 7, Section 7.6).

Prevention of Misclassification

Misclassification has been extensively discussed in the epidemiologic literature,[39-42] reflecting its importance in epidemiologic studies. As seen in the examples described in this section, misclassification may severely distort the magnitude of an association between a risk factor and a disease. If the true relative risk or odds ratio is close to 1.0, a nondifferential misclassification may completely mask the association. For example, for an exposure with a prevalence as high as 16% (i.e., in a range not unlike that of many risk factors), if the true odds ratio is approximately 1.3, the observed odds ratio may be virtually 1.0 if a nondifferential misclassification resulted from a measurement procedure with both sensitivity and specificity levels of approximately 70%. Differential misclassification of a confounding variable, on the other hand, may either dilute or strengthen an association or even produce a spurious one. When the exposure is common, failing to demonstrate a real relationship or inferring that an association exists when it is spurious may have serious public health consequences (see Chapter 10).

Data are usually not available to allow a comparison between correctly classified and misclassified individuals in terms of available characteristics (e.g., educational level), but when they are, they may be informative. As seen in Table 4-10, of the 34 incident cases included in the case-control study on melanoma nested in the Nurses' Health Study cohort, 26 were correctly classified (8 true positives and 18 true negatives),[18] and 8 were misclassified (7 false positives and 1 false negative). A comparison could be made, for example, between the false positives and true negatives on the one hand (addressing the issue of specificity) and between the false negatives and true positives on the other (addressing the issue of sensitivity). In the Nurses' Health Study, the authors reported no important differences between the correctly and incorrectly classified cases. (When studying tanning ability, it would not be unreasonable to postulate that recall of tanning ability could be influenced by factors such as family history of skin diseases or involvement in outdoor activities.) Similarity in pertinent characteristics of correctly classified and misclassified persons may perhaps indicate that recall bias is not a probable explanation for the misclassification and raises the possibility that the information bias originated from problems related to the instrument or the observer. Thus, the comparison between misclassified and nonmisclassified subjects need not be limited to respondent characteristics and should also include aspects of the data collection procedures. When interviews are taped, adherence to the protocol by interviewers can be compared. Additionally, information should be obtained on the reliability and validity of the instrument (e.g., a questionnaire), as discussed in Chapter 8.

A more general approach to assess information bias is based on the evaluation of the odds of "inaccurate self-reporting" as the outcome of interest (i.e., without specification of sensitivity or specificity). An example is given by a study of the validity of self-reported AIDS-specific diagnoses (such as esophageal candidiasis) vis-à-vis AIDS diagnoses documented by AIDS surveillance registries, with the latter used as the gold standard.[43] In this study, when compared with former smoking and no smoking, current smoking was found to be strongly related to inaccurate self-reporting of any AIDS-specific diagnoses, as expressed by an odds ratio of 2.6 (95% confidence interval, 1.2, 5.6). On the other hand, the odds of inaccurate self-reporting in this study did not appear to be related to age, ethnic background, education, or time since the patient had first tested positive for the human immunodeficiency virus (HIV).

Prevention of misclassification of exposure and outcome is a function of the state-of-the-art measurement techniques that can safely be applied to the large number of subjects participating in epidemiologic studies. The use of objective (e.g., biological) markers of exposure and more accurate diagnostic techniques for ascertainment of outcomes, such as the use of ultrasound or computerized tomography (CT) scan to diagnose asymptomatic atherosclerosis,[20,44] constitutes the most efficient approach for ameliorating the problems related to misclassification bias. In the meantime, if sensitivity and specificity of outcome or exposure measurements are known, it is possible to correct for misclassification; for example, in a case-control study, this can be done by using available formulas that estimate a "corrected odds ratio" as a function of the "observed odds ratio" and the estimated sensitivity and specificity of the exposure classification.[40-43,45] Furthermore, correction methods that can be applied to situations in which measurement errors affect both exposure variables and covariates (either categorical or continuous variables) have been described.[46] Similarly to the imputation methods used to attempt to correct for selection bias (see Section 4.2.1), when misclassification parameters are unknown, sensitivity analysis could be used to obtain a range of plausible "corrected" estimates under different assumptions about the levels of misclassification (see Chapter 10, Section 10.3).

4.4 Combined Selection/Information Biases

This section discusses biases that have both selection and information components. They include biases related to medical surveillance, cross-sectional studies, and evaluation of screening. The sections on cross-sectional and screening evaluation biases may seem somewhat repetitious vis-à-vis previous discussions on selection and information biases in this chapter. They have, however, been included here because they reflect examples specific to these areas and thus may be of special value to those especially interested in cross-sectional and screening intervention studies.

4.4.1 Medical Surveillance (or Detection) Bias

Medical surveillance bias occurs when a presumably medically relevant exposure leads to a closer surveillance of study outcomes that may result in a higher probability of detection in exposed individuals (i.e., when the identification of the outcome is not independent of the knowledge of the exposure). This type of bias is particularly likely when the exposure is a medical condition or therapy—such as diabetes or use of oral contraceptives—that leads to frequent and thorough checkups and the outcome is a disease that is characterized by a high proportion of subclinical cases and thus likely to be diagnosed during the frequent medical encounters resulting from the need to monitor the exposure. For example, although there may be no basis for believing that oral contraceptive use can lead to renal failure, a spurious association would be observed if women taking oral contraceptives were more likely than other women to have medical checkups that included repeated measurements of glomerular filtration rate.

Depending on the study design, medical surveillance bias can be regarded as a type of either selection bias or information bias. In the context of a case-control study, medical surveillance bias can occur if cases are more likely to be identified (or selected into the study) if they are exposed (see Figure 4-2). In a cohort study, medical surveillance bias may be akin to information bias if, for example, the exposed individuals undergo a more thorough examination than the unexposed individuals.

Medical surveillance bias is more likely to occur when the outcome is ascertained through regular healthcare channels (e.g., electronic health records). Alternatively, *when the outcome is assessed systematically, regardless of exposure* in a concurrent cohort design, medical surveillance bias is less likely to occur.[3] Thus, meticulously standardized methods of outcome ascertainment are routinely used in most major cohort studies, such as the classic Framingham Study[47] or the Atherosclerosis Risk in Communities Study.[20] Another strategy to prevent medical surveillance bias that can be used when conducting cohort studies is to *mask exposure status when ascertaining the presence of the outcome.*

The strategies mentioned heretofore may not be feasible, however, when carrying out a case-control study in which the case diagnosis may have already been affected by the presence of the exposure. When this occurs, for analytical purposes, *information should be obtained on the frequency, intensity, and quality of medical care received by study participants.* For example, to assess the relationship between use of hormone replacement therapy and a given disease with a subclinical component (e.g., non–insulin-dependent diabetes) using a traditional case-control design, it is important to take into consideration medical care indicators, such as the frequency of medical visits in the past and whether the individual has medical insurance. Because education and socioeconomic status are usually related to availability and use of medical care, they too should be taken into consideration when trying to assess surveillance bias.

It is also possible to *obtain information on variables that indicate awareness of health problems*, such as compliance with screening exams and knowledge of subclinical disease or of results of blood measurements. For example, in a prospective study of the relationship of vasectomy to the risk of clinically diagnosed prostate cancer (that is, not through systematic examination), the possibility of surveillance bias was assessed by examining variables that might reflect greater utilization of medical care.[48] In this study, no differences were found between subjects who had and those who had not had vasectomy with regard to their knowledge of their blood pressure or serum cholesterol levels. The proportions of study participants who had had screening sigmoidoscopy were also similar, leading the authors to conclude that vasectomized men were not under a greater degree of medical surveillance than those who had not been vasectomized. In addition, the frequency of digital rectal examinations was similar between the vasectomized (exposed) and the nonvasectomized (unexposed) groups, implying equal access to a procedure that may lead to the diagnosis of the study outcome (prostate cancer).

Finally, when medical surveillance bias occurs, the disease tends to be diagnosed earlier in exposed than in unexposed individuals; as a result, the proportion of less advanced disease in a cohort study is higher in the exposed group. In a case-control study, the bias is denoted by the fact that the association is found to be stronger or present only for the less advanced cases. In the cohort study discussed previously, Giovannucci et al.[48] found that the histologic severity staging of prostate cancer was similar for vasectomized and nonvasectomized men, a finding inconsistent with what would be expected if medical surveillance had been more intensive in the vasectomized group. *Stratification by disease severity at diagnosis* is thus an additional strategy for examining and taking into consideration the possibility of surveillance bias.

4.4.2 Cross-Sectional Biases

Cross-sectional biases can be classified as incidence–prevalence bias and temporal bias. The former is a type of selection bias, whereas the latter can be regarded as an information bias.

Incidence–Prevalence Bias

Incidence–prevalence bias may result from the inclusion of prevalent cases in a study when the goal is to make inferences in relation to disease *risk*. As discussed in Chapter 3, Section 3.3, the strength of an association is sometimes estimated using the prevalence rate ratio rather than the relative risk, as when analyzing data from a cross-sectional survey or when assessing cross-sectional associations at baseline in a cohort study. If the investigator is interested in assessing potentially causal associations, the use of the prevalence rate ratio as an estimate of the incidence ratio is subject to bias. Equation 2.3, described in Chapter 2, Section 2.3, shows the dependence of the point prevalence odds [Prev/(1.0 – Prev)] on cumulative incidence (Inc) and disease duration (Dur), assuming that incidence and duration are approximately constant:

$$\frac{\text{Prev}}{1.0 - \text{Prev}} = \text{Inc} \times \text{Dur}$$

Equation 2.3 can be rewritten as Equation 2.4,

$$\text{Prev} = \text{Inc} \times \text{Dur} \times (1.0 - \text{Prev})$$

thus demonstrating that, in addition to incidence and duration, prevalence is a function of the term (1.0 – Prev) (which, in turn, obviously depends on the magnitude of the point prevalence rate).

As a corollary of Equation 2.4, the point prevalence rate ratio comparing exposed (denoted by subscript "+") and unexposed (denoted by subscript "–") individuals, obtained in cross-sectional studies, will be a function of (1) the relative risk, (2) the ratio of the disease duration in exposed individuals to that in unexposed individuals, and (3) the ratio of the term (1.0 – Prev) in exposed individuals to the same term in unexposed individuals. Ratios 2 and 3 represent two types of incidence–prevalence bias, that is, the *duration ratio bias* and the *point prevalence complement ratio bias*, respectively, when the prevalence rate ratio (PRR) is used to estimate the relative risk (see Chapter 3, Section 3.3),

$$\text{PRR} = \left(\frac{q_+}{q_-}\right) \times \left(\frac{\text{Dur}_+}{\text{Dur}_-}\right) \times \left(\frac{1.0 - \text{Prev}_+}{1.0 - \text{Prev}_-}\right)$$

where q_+ and q_- are the cumulative incidence values for exposed and unexposed individuals, respectively.

Duration Ratio Bias. This type of bias (which can be thought of as a type of selection bias) occurs when the prevalence rate ratio is used as a measure of association and the duration of the disease after its onset is different between exposed and unexposed persons. (Because duration of a chronic disease is so often related to survival, this type of bias may also be designated as *survival bias*.) For diseases of low prevalence, when the duration (or prognosis) of the disease is independent of the exposure (i.e., the same in exposed and unexposed), the prevalence rate ratio is a virtually unbiased estimate of the relative risk. On the other hand, when the exposure of interest affects the prognosis of the disease, bias will be present, as shown in the examples later in this chapter.

Point Prevalence Complement Ratio Bias. Even if duration is independent of exposure regardless of the direction of the effect of the factor on the outcome, the prevalence rate ratio tends to underestimate the *strength* of the association between the exposure and the outcome (i.e., it biases the relative risk toward 1.0). The magnitude of this bias depends on both the prevalence rate ratio and the absolute magnitude of the point prevalence rates. When the point prevalence rate is higher in exposed than in unexposed individuals (prevalence rate ratio > 1.0), the point prevalence complement ratio [or $(1.0 - \text{Prev}_+)/(1 - \text{Prev}_-)$] is less than 1.0. It is close to 1.0 when the point prevalence rates are low in both exposed and unexposed even if the prevalence rate ratio is relatively high. For example, if the prevalence of the disease in exposed subjects is 0.04 and in unexposed subjects 0.01, the prevalence rate ratio is high ($0.04/0.01 = 4.0$), but the bias resulting from the point prevalence complement ratio is merely $0.96/0.99 = 0.97$ (i.e., still < 1.0 but close enough to 1.0 to result in a practically negligible bias). On the other hand, when the prevalence is relatively high in exposed individuals, the point prevalence complement ratio can be markedly less than 1.0, thus resulting in important bias. For example, if the prevalence of the disease in exposed subjects is 0.40 and in unexposed subjects 0.10, the prevalence rate ratio is the same as in the previous example (4.0); however, the point prevalence complement ratio is $0.6/0.90 = 0.67$ (i.e., the prevalence rate ratio underestimates the relative risk by 33%—even in the absence of duration ratio bias). The influence of the magnitude of prevalence is sometimes felt even for a low prevalence rate ratio. For example, if the point prevalence rates are 0.40 in exposed and 0.25 in unexposed subjects, the prevalence rate ratio is fairly small (1.6), but the bias factor is 0.80 (i.e., the prevalence rate ratio underestimates the relative risk by at least 20%—and more if there is also duration ratio bias). Obviously, the bias will be greatest when both the prevalence rate ratio and the prevalence rate in one of the groups (exposed or unexposed) are high. For studies of factors that

decrease the prevalence of the disease (i.e., prevalence rate ratio < 1.0), the reciprocal reasoning applies; that is, $(1.0 - \text{Prev}_+)/(1 - \text{Prev}_-)$ will be greater than 1.0, and the magnitude of the bias will also be affected by the absolute rates.

Examples of Incidence–Prevalence Biases. In the examples that follow, it is assumed that the incidence and duration according to the exposure have remained stable over time.

- *Gender and acute myocardial infarction in U.S. whites:* White U.S. males have a much higher risk of myocardial infarction than white females. Some studies, however, have shown that, even after careful age adjustment, females have a shorter average survival than males.[49] Thus, the ratio $(\text{Dur}_{\text{males}}/\text{Dur}_{\text{females}})$ tends to be greater than 1.0, and as a consequence, the prevalence rate ratio expressing the relationship of sex to myocardial infarction overestimates the relative risk.
- *Current smoking and emphysema:* Smoking substantially increases the risk of emphysema. In addition, survival (and thus duration of the disease) in emphysema patients who continue to smoke after diagnosis is shorter than in those who quit smoking. As a result, prevalence rate ratios estimated in cross-sectional studies evaluating the association between current smoking and emphysema tend to underestimate the relative risk.
- *Tuberculin purified protein derivative (PPD) reaction and clinical tuberculosis:* In assessments of the relationship between the size of the PPD skin test reaction and clinical tuberculosis, prevalence rate ratios were shown to underestimate relative risks in a population-based study carried out by G. Comstock et al. (unpublished observations) a few decades ago (**FIGURE 4-5**). This underestimation was likely due to the relatively high prevalence of clinical tuberculosis in this population at the time the study was carried out and thus to the occurrence of prevalence complement ratio bias.

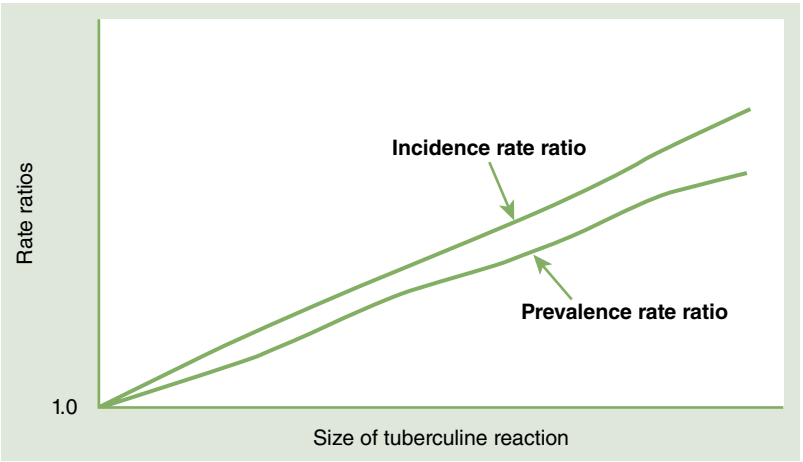


FIGURE 4-5 Schematic representation of the results of the study by Comstock et al. (unpublished) evaluating the relationship of size of PPD reaction to clinical tuberculosis. After an initial cross-sectional survey conducted in 1946, the cohort was followed over time for determination of incidence rates.

Data from GW Comstock et al., personal communication.

Prevention of Incidence–Prevalence Bias. If the goal is to evaluate potential disease determinants, whenever possible, incident cases should be used to avoid incidence–prevalence bias. Incidence–prevalence bias, although more easily conceptualized by comparing incidence with prevalence (cross-sectional) rate ratio data, may also occur in case-control studies when prevalent rather than only newly developed (incident) cases are used. For example, if smoking decreases survival after diagnosis, thereby decreasing the disease's duration (as in myocardial infarction), a case-control study based on prevalent cases may include a higher proportion of nonsmoking cases (as smokers would have been selected out by death) than would a study based on incident cases, thus diluting the strength of the association (see Chapter 1, Figure 1-19).

Another problem in case-control studies is that newly *diagnosed* cases are used as proxies for newly *developed* cases. Thus, for diseases that may evolve subclinically for many years before diagnosis, such as chronic lymphocytic leukemia, diabetes, or renal insufficiency, presumed incident cases are, in fact, a mix of incident and prevalent cases, and incidence–prevalence bias may occur unbeknownst to the investigator. A cohort study efficiently prevents incidence–prevalence bias if its procedures include careful ascertainment and exclusion of all prevalent cases at baseline (clinical and subclinical) as well as a systematic and periodic search of newly developed clinical and subclinical outcomes.

Temporal Bias

In cross-sectional studies, the proper temporal sequence needed to establish causality, risk factor → disease, cannot be firmly established. In other words, it is difficult to know which came first, the exposure to the potential risk factor or the disease. *Temporal bias* occurs when the inference about the proper temporal sequence of cause and effect is erroneous. For example, results from a prevalence survey may establish a statistical association between high serum creatinine levels and the occurrence of high blood pressure. Because the time sequence cannot be established, however, a cross-sectional association between these variables may mean either that high serum creatinine (a marker of kidney failure) leads to hypertension or vice versa. A prospective study in which blood pressure levels are measured in persons with normal serum creatinine levels who are then followed over time for ascertainment of hypercreatininemia can obviously identify the proper temporal sequence and thus lend support to the conclusion that high blood pressure predicts incipient renal insufficiency.[50]

Temporal bias may also occur in case-control studies—even those including only newly developed (incident) cases—when the suspected exposure is measured after disease diagnosis in cases. For example, because hepatitis B virus (HBV) is myelotoxic, it has been suggested that HBV may be an etiologic factor for the so-called idiopathic aplastic anemia (AA).[51] Temporal bias, however, could explain the relationship between HBV and AA in a case-control study if serum samples for determination of HBV antibody and antigen levels had been collected after AA onset, as individuals with AA may receive transfusions of blood contaminated with HBV even before a diagnosis is made. Thus, the erroneously inferred sequence is

$$\text{HBV} \rightarrow \text{AA},$$

but the true sequence is

$$\text{undiagnosed AA} \rightarrow \text{blood transfusion} \rightarrow \text{diagnosed AA}.$$

An example of the reasoning underlying the possibility of temporal bias is given by the association between estrogen replacement therapy (ERT) in postmenopausal women and endometrial

cancer.[52] Although the causal nature of this association is currently well established, it was initially disputed on the grounds that a higher likelihood of using ERT resulted from symptoms occurring as a consequence of incipient, undiagnosed endometrial cancer.[53] Thus, instead of the sequence

ERT → endometrial cancer,

the alternative sequence would be

undiagnosed endometrial cancer → symptoms → ERT → diagnosed endometrial cancer.

Another example of temporal bias is given by a cross-sectional study of Dutch children in which negative associations were found of pet ownership with allergy, respiratory symptoms, and asthma.[54] As aptly postulated by the study's investigators, these results may have resulted from the fact that families are likely to remove from the home (or not acquire) pets after such manifestations occur. This study also underscores why the term *reverse causality* is occasionally used in connection with a temporal bias of this sort.

A further example of this type of bias was suggested by Nieto et al.,[55] who found that the relationship of current smoking to prevalent clinical atherosclerosis (defined by self-reported physician-diagnosed heart attack or cardiac surgery) was much stronger when using longitudinal data than when using cross-sectional data (in contrast to the association between smoking and subclinical atherosclerosis, which was of similar strength for the longitudinal and cross-sectional data). One possible explanation for these findings was that the occurrence of a heart attack (but not the presence of subclinical atherosclerosis) may lead to smoking cessation and thus to a dilution of the association when using prevalent cases.* Similarly, prevalent cases of clinical coronary heart disease, as part of their medical care, are more likely to be engaged in physical exercise than normal controls. This type of bias may occur even in prospective analyses when the outcome of interest is mortality. For example, the short-term mortality from lung cancer can be higher in individuals who stopped smoking recently than in current smokers because of the tendency of symptomatic individuals or those for whom a diagnosis has been made to quit smoking.[56] Epidemiologists usually handle this bias by excluding from the analysis the deaths that occur within a specified period after the beginning of the study.

To prevent temporal bias in a cross-sectional survey, it is occasionally possible to improve the information on temporality when obtaining data through questionnaires. Temporality pertaining to potential risk factors, such as smoking, physical activity, and occupational exposures, can be ascertained in cross-sectional samples by means of questions such as, "When were you first exposed to . . . ?" For some chronic diseases, such as angina pectoris, it is also possible to obtain information on the date of onset.

The investigators can then establish the temporal sequence between risk factor and disease, assuming, of course, that the information from surveyed individuals is accurate. (Obviously, even if temporality can be established in a cross-sectional study, the investigator will still have the incidence–prevalence bias to contend with.) When the date of the beginning of the exposure is unknown, as in the example of viral hepatitis and aplastic anemia, the only solution is to use prospective data on exposure and outcome (a formidable challenge in the aplastic anemia example, given the rarity of this disease).

*Survival bias is, of course, another explanation, resulting from a poor prognosis of myocardial infarction in smokers.

Finally, it may be possible to assess temporal bias occurring because the presumed exposure is a consequence of undiagnosed disease—as in the example of ERT and endometrial cancer mentioned previously in this chapter—by considering why the exposure occurred. In the study of Antunes et al.,[52] for instance, data can be stratified according to indication for ERT use, such as bleeding; if temporal bias is not a likely explanation for the relationship of estrogen to endometrial cancer, the association will be observed for both individuals who were prescribed estrogens because they were bleeding and those who were given estrogens for other reasons (e.g., prevention of osteoporosis).

4.4.3 Biases Related to the Evaluation of Screening Interventions

Like any other epidemiologic studies, evaluation of screening interventions are also prone to biases, of which five types are particularly relevant: selection bias, incidence–prevalence bias, length bias, lead time bias, and overdiagnosis bias. (For a better understanding of these types of biases, the reader should review the concepts underlying the natural history of disease; see, e.g., Gordis.[17])

Selection Bias

Selection bias stems from the fact that when the evaluation of screening relies on an observational design, the screened group may differ substantially from the nonscreened group. Thus, for example, persons who attend a screening program may be of a higher socioeconomic status than those who do not and may therefore have a better prognosis regardless of the effectiveness of the screening program. *Prevention* of this type of selection bias is best carried out by using an experimental design (i.e., by randomly assigning screening status to study participants). While improving internal validity, however, experimental studies to evaluate screening programs are typically conducted in selected populations, thus potentially limiting their external validity.

Incidence–Prevalence Bias, Length Bias

Also known as *survival bias, incident–prevalence bias* results from comparing prognosis in prevalent cases detected in the first screen, which is akin to a cross-sectional survey, with that in incident cases detected in subsequent screenings. This bias occurs because prevalent cases include long-term survivors who have a better average survival than that of incident cases in whom the full spectrum of severity is represented. This type of bias may occur in "pre–post" studies, for example, when comparing a screening strategy used in a screening exam (pre) that identifies prevalent cases with a different strategy in subsequent screens identifying incident cases (post).

A related bias is the so-called *length bias*, which occurs when a better prognosis for cases detected directly by the screening procedure (e.g., occult blood test for colorectal cancer) than for cases diagnosed between screening exams is used as evidence that the screening program is effective. To understand this type of bias, it is important to briefly review some key concepts related to the natural history of a disease and screening.

The effectiveness of screening is positively related to the length of the detectable preclinical phase (DPCP; see **FIGURE 4-6** and, for definitions, **TABLE 4-12**), which in turn reflects the rate at which the disease progresses. This means that for diseases with a rapid progression, it is difficult, if not outright impossible, to improve prognosis by means of early detection. For example, a short average DPCP and its attending poor survival characterize most cases of lung cancer, for which screening generally is not effective. On the other hand, the long DPCP of *in situ* cervical cancer (or high-grade squamous intraepithelial lesions) explains why treatment after an abnormal Pap smear is related to a cure rate of virtually 100%.

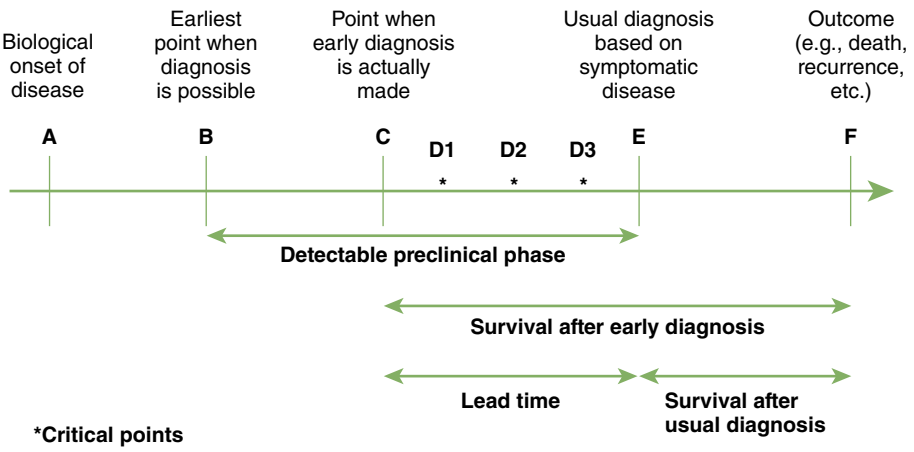


FIGURE 4-6 Natural history of a disease.

See Table 4-12 for definitions.
Modified from Gordis L. *Epidemiology.* 5th ed. Philadelphia, PA: Elsevier Saunders; 2014.[17]

TABLE 4-12 Natural history of a disease: definitions of components represented in Figure 4-6.

Component	Represented in Figure 4-6 as . . .	Definition
Detectable preclinical phase	The interval between points B and E	Phase that starts when early diagnosis becomes possible and ends with the point in time when usual diagnosis based on symptomatic disease would have been made.
Critical points	D1, D2, and D3	Points beyond which early detection and treatment are less and less effective vis-à-vis treatment following usual diagnosis. Treatment is totally ineffective after the last critical point (point D3 in the figure).
Lead time	The interval between points C and E	Period between the point in time when early diagnosis was made and the point in time when the usual diagnosis (based on symptoms) would have been made.

Data from Gordis L. *Epidemiology.* 5th ed. Philadelphia, PA: Elsevier Saunders; 2014.[17]

Even for the same disease, regardless of screening, it can be shown that patients whose disease has a longer DPCP have a better prognosis than those whose disease has a shorter DPCP (e.g., postmenopausal versus premenopausal breast cancer, respectively). For example, in the Health Insurance Plan Study of the effectiveness of screening for breast cancer, the so-called interval cases—that is, cases who were clinically diagnosed during the interval between the screening exams—had, on average, a higher case fatality rate than subclinical cases diagnosed as a result of the screening exam.[57] Although some of these cases may have been false negatives missed by the previous screening exam and therefore not true interval cases, many were probably characterized by rapidly growing tumors, that is, by a short DPCP (**FIGURE 4-7**). It follows that when evaluating a screening program, one must take into careful consideration the fact that cases detected by the screening procedure (e.g., mammography), which thus tend to have a longer DPCP, have an inherently better prognosis than the interval cases *regardless of the effectiveness of screening*. Failure to do so results in *length bias*, which occurs when a better prognosis for screening-detected than for interval cases is used as evidence that the screening program is effective when in reality it may be due to the longer DPCP of the former cases, reflecting a slower growing disease than that of interval cases.

Prevention of length bias can be accomplished by using an experimental approach and comparing the prognosis of *all* cases—which include cases with both short and long DPCPs—occurring in individuals randomly assigned to a screening program with that of *all* cases occurring in randomly assigned controls who do not undergo the screening exams. (The distribution of patients with long DPCPs versus those with short DPCPs is expected to be the same in the randomly assigned screening and control groups.)

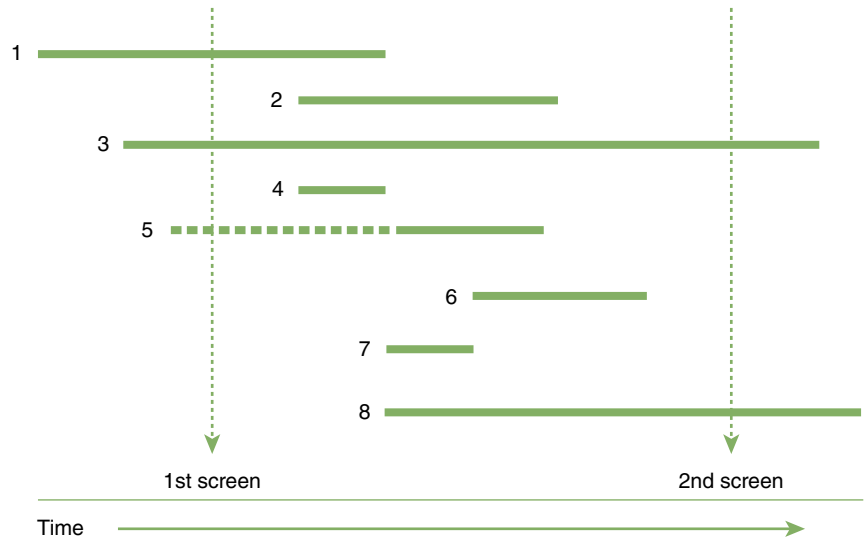


FIGURE 4-7 Schematic representation of the length of the detectable preclinical phase (DPCP) in cases occurring during a screening program. Cases with a longer DPCP (cases 1, 3, and 8) have a higher probability of identification at each screening exam. Cases with a shorter DPCP occurring between screening exams are the interval cases (cases 2, 4, 6, and 7). Case 5 is a false negative (missed by the first exam).

Lead Time Bias

Lead time is the time by which diagnosis can be advanced by screening. It is the time between early diagnosis (Figure 4-6, point C) and the usual time when diagnosis would have been made if an early diagnostic test(s) had not been applied to the patient (Figure 4-6, point E; see also Table 4-12). The lead time, therefore, is contained within the DPCP.

When evaluating effectiveness of screening, lead time bias occurs when survival (or recurrence-free time) is counted from the point in time when early diagnosis was made. Thus, even if screening is ineffective, the early diagnosis adds lead time to the survival counted from the time of usual diagnosis. Survival may then be increased from time of early diagnosis but not from the biological onset of the disease (**FIGURE 4-8**).[58]

Lead time bias occurs only when estimating survival (or time-to-event) from time of diagnosis. Thus, lead time bias can be avoided by calculating the mortality risk or rate among all screened and control subjects rather than the cumulative probability of survival (or its complement, the cumulative case fatality probability) from diagnosis among cases.[17] If survival from diagnosis is chosen as the strategy to describe the results of the evaluation of a screening approach, the average duration of lead time must be estimated and taken into account when comparing survival after diagnosis between screened and nonscreened groups. For survival to be regarded as increased from the biological onset, it must be greater than the survival after usual diagnosis plus lead time (**FIGURE 4-9**). It is, thus, important to estimate average lead time.

If the disease for a given individual is identified through screening, it is impossible to know when "usual" diagnosis would have been made if screening had not been carried out. Thus, it is not possible to estimate the lead time for individual patients, only an average lead time. What follows is a simplified description of the basic approach used to estimate average lead time. A more detailed account of lead time estimation is beyond the scope of this intermediate methods text and can be found elsewhere.[58]

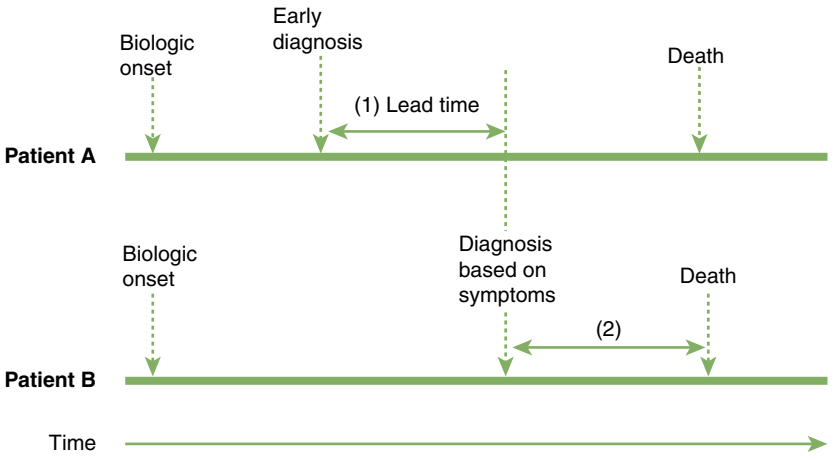


FIGURE 4-8 Schematic representation of lead time bias: In spite of the early diagnosis by screening, survival of patient A is the same as survival of patient B, whose disease was diagnosed because of clinical symptoms, because survival (A) = (1) lead time + (2) survival (B).

Modified from Gordis L. *Epidemiology*. 5th ed. Philadelphia, PA: Elsevier Saunders; 2014.[17]

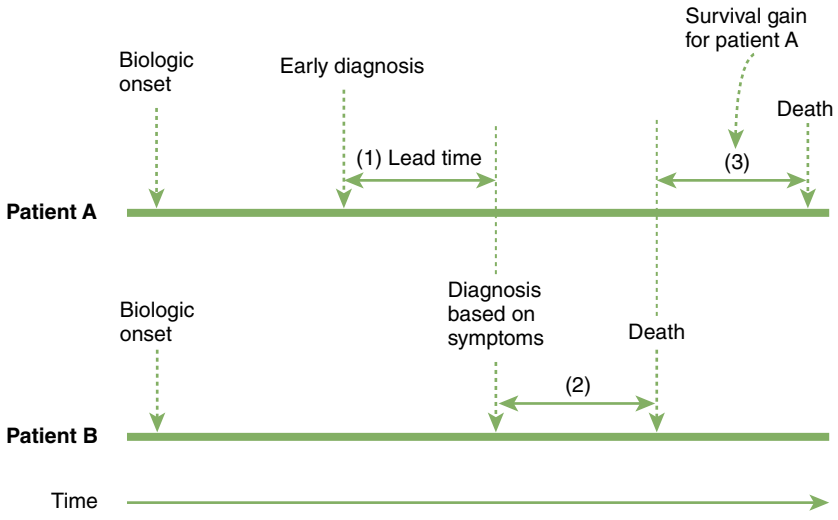


FIGURE 4-9 Schematic representation of lead time bias: Survival of patient A from early diagnosis is better than survival of patient B because survival (A) > (1) lead time + (2) survival (B).

Modified from Gordis L. *Epidemiology*. 5th ed. Philadelphia, PA: Elsevier Saunders; 2014.[17]

As mentioned previously, the lead time is a component of the DPCP. Thus, to estimate the average lead time, it is first necessary to estimate the average duration of the DPCP (Dur_{DPCP}) using the known relationship between prevalence ($Prev_{DPCP}$) and incidence (Inc_{DPCP}) of preclinical cases, that is, cases in the DPCP (see also Chapter 2, Section 2.3, Equation 2.4):

$$Prev_{DPCP} = Inc_{DPCP} \times Dur_{DPCP} \times (1.0 - Prev_{DPCP})$$

The duration of the DPCP can then be easily derived as

$$Dur_{DPCP} = \frac{Prev_{DPCP}}{Inc_{DPCP} \times (1.0 - Prev_{DPCP})}$$

If the prevalence of the disease is not too high (e.g., no greater than about 5%), $1.0 - Prev_{DPCP}$ will be close to 1.0, and thus, this equation can be simplified as follows:

$$Dur_{DPCP} \approx \frac{Prev_{DPCP}}{Inc_{DPCP}}$$

To apply this formula, the $Prev_{DPCP}$ is estimated using data from the first screening exam of the target population, which is equivalent to a cross-sectional survey. The Inc_{DPCP} can be estimated in successive screening exams among screenees found to be disease free at the time of the first screening. An alternative way to estimate Inc_{DPCP}, and one that does not require follow-up with the screenees, is to use the incidence of clinical disease in the reference population if available.

The rationale for this procedure, and an important assumption justifying screening, is that, if left untreated, preclinical cases would necessarily become clinical cases; thus, there should not be a difference between the incidence of clinical and preclinical disease. When using available clinical disease incidence (e.g., based on cancer registry data), however, it is important to adjust for differences in risk factor prevalence, expected to be higher in screenees than in the reference population from which clinical incidence is obtained. Thus, for example, a family history of breast cancer is likely to be more prevalent in individuals screened for breast cancer than in the female population at large.

Next, using the duration of the DPCP estimate, the estimation of the average lead time needs to take into account whether early diagnosis by screening is made at the first screening exam or in subsequent screening exams.

The estimation of the lead time of point prevalent preclinical cases detected at the first screening exam relies on certain assumptions regarding the distribution of times of early diagnosis during the DPCP. For example, if the distribution of early diagnosis by screening can be assumed to be homogeneous throughout the DPCP—that is, if the sensitivity of the screening test is independent of time within the DPCP (**FIGURE 4-10A**)—the lead time of point prevalent preclinical cases can be simply estimated as

$$\text{Lead time} = \frac{\text{DPCP}}{2}$$

The latter assumption, however, may not be justified in many situations. For most diseases amenable to screening (e.g., breast cancer), the sensitivity of the screening test, and thus the probability of early diagnosis, is likely to increase during the DPCP (**FIGURE 4-10B**) as a result of the

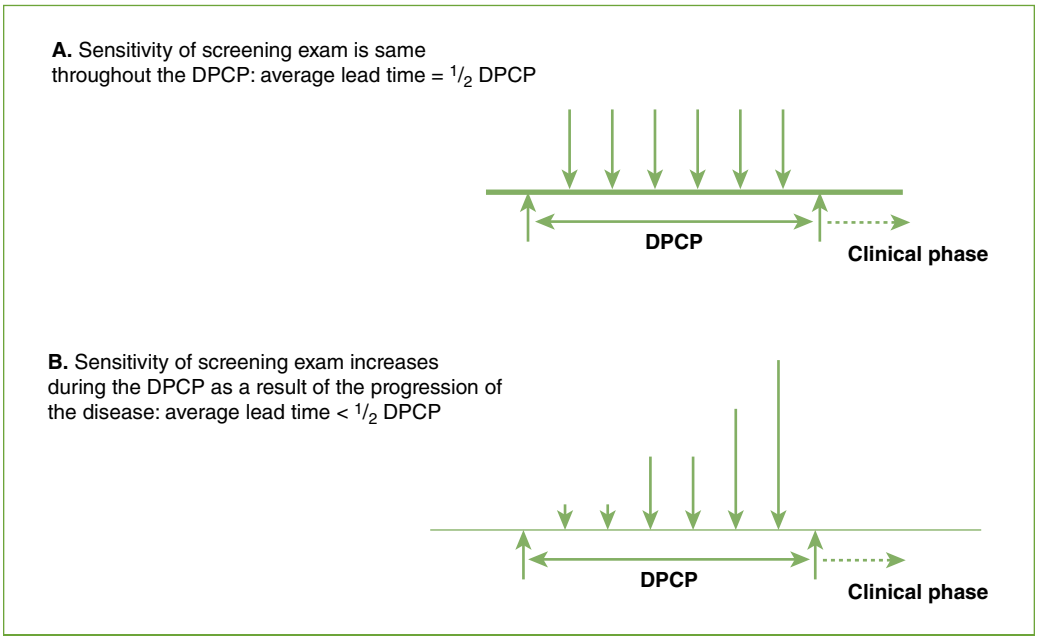


FIGURE 4-10 Estimation of lead time as a function of the variability of the sensitivity of the screening exam during the detectable preclinical phase (DPCP).

progression of the disease as it gets closer to its symptomatic (clinical) phase. If this is the case, a more reasonable assumption would be that the average lead time is less than one-half of the DPCP. Obviously, the longer the DPCP, the longer the lead time under any distributional assumption. Also, because the DPCP and thus the average lead time are dependent on the validity of the screening exam, they become longer as more sensitive screening tests are developed.

The duration of the lead time for *incident* preclinical cases identified in a program in which repeated screening exams are carried out is a function of how often the screenings are done (i.e., the length of the interval between successive screenings). The closer in time the screening exams are, the greater the probability that early diagnosis will occur closer to the onset of the DPCP, and thus, the more the lead time will approximate the DPCP.

FIGURE 4-11 schematically illustrates short and long between-screening intervals and their effects on the lead time, assuming that the sensitivity of the test does not vary throughout the DPCP. Considering, for example, two persons with similar DPCPs whose diseases start soon after the previous screening, the person with the shorter between-screening interval (*a* to *b*, patient A) has his or her newly developed preclinical disease diagnosed nearer the beginning of the DPCP than the person with the longer between-screening interval (*a* to *c*, patient B). Thus, the duration of the

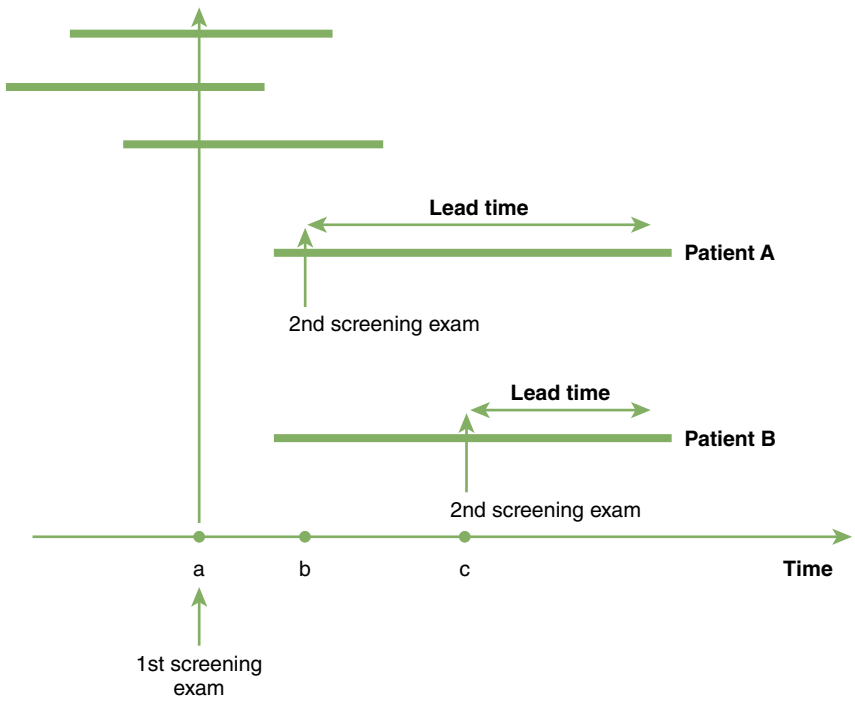


FIGURE 4-11 Relationship between frequency of screening and duration of lead time. Horizontal lines represent duration of the detectable preclinical phase (DPCP). In patient A, the second screening exam is carried out soon after the first screening exam: lead time ≈ DPCP. In patient B, the between-screening interval is longer: lead time ≈ (1/2) × DPCP.

TABLE 4-13 Range of prevalence rates of prostate cancer by age.

Age (years)	Prevalence ranges of prostate cancer (%)
50–59	10–42%
60–69	17–38%
70–79	25–66%
≥ 80	18–100%

Data from Franks LM. Latent carcinoma of the prostate. *J Pathol Bacteriol.* 1954;68:603-616[65]; Bostwick DG, Cooner WH, Denis L, Jones GW, Scardino PT, Murphy GP. The association of benign prostatic hyperplasia and cancer of the prostate. *Cancer.* 1992;70:291-301[66]; Breslow, Chan CW, Dhom G, et al. Latent carcinoma of prostate at autopsy in seven areas. *Int J Cancer.* 1977;20:680-688[67]; Baron E, Angrist A. Incidence of occult adenocarcinoma of the prostate after fifty years of age. *Arch Pathol.* 1941;32:787-793[68]; Edwards CN, Steinthorsson E, Nicholson D. An autopsy study of latent prostatic cancer. *Cancer.* 1953;6:531-554[69]; Halpert B, Schmalhorst WR. Carcinoma of the prostate in patients 70 to 79 years old. *Cancer.* 1966;695-698[70]; Scott R, Mutchnik DL, Laskowski TZ, Schmalhorst WR. Carcinoma of the prostate in elderly men: incidence, growth characteristics and clinical significance. *J Urol.* 1969;101:602-607.[71]

lead time is closer to the duration of the DPCP for patient A than for patient B. The maximum lead time obviously cannot be longer than the DPCP.[59]

Overdiagnosis Bias

Overdiagnosis bias occurs when screening identifies patients whose early subclinical disease does not evolve to more advanced stages and the analysis is based on survival of patients. Consider, for example, the natural history of prostate cancer: It has been estimated that as many as one-third of men younger than 70 years and between two-thirds to 100% of older men may have prostate cancer but often in a microscopic, noninvasive form (see **TABLE 4-13**).[60] In the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute, the number of cases of prostate cancer per year was estimated at close to 221,000 from 1975 through 2011, but only 27,540 deaths (≈ 12.5%) occurred in the same period.[61] Thus, it is likely that many individuals—particularly as they age—die *with* prostate cancer rather than *from* prostate cancer. Because it is not currently possible to identify cases that would not evolve to more invasive stages leading to death, if they represent a relative high proportion of all cases, analysis of survival after diagnosis would favor the diagnosis made by screening, which would include many such cases.*

There have been two recent trials of screening for prostate cancer with prostate-specific antigen (PSA): one in the United States and the other in Europe. In the United States-based trial,[62]

*Note that the efficiency of PSA screening would be diluted by overdiagnosis (i.e., by the inclusion of noninvasive cases). For example, if in a hypothetical population of 10,000, there were 2000 potentially lethal cases, without screening, they would all die. With screening and assuming that the effectiveness of the treatment is very high (e.g., 97.5%), only 50 deaths would occur. If the potentially lethal cases could be identified and screened, the number needed to screen to prevent 1 death would be 40 (i.e., 2000/50). On the other hand, if these potentially lethal cases could not be identified, the whole target population would have to be screened, and the number needed to screen to prevent 1 death would be 200 (i.e., 10,000/50). Thus, screening of only the potentially "lethal" subgroup (those who would have died without screening) would be much more efficient than screening the whole target population.

the cumulative hazards of death for individuals tested with PSA and for the control group were virtually the same for most of the follow-up and at the end of the trial. In the European trial, a significant difference in mortality could not initially be found between the screened and nonscreened groups.[63] However, upon further follow-up, cumulative mortality was significantly reduced in the PSA group.[64] In the two trials, however, by evaluating mortality for all individuals (and not just for those with a diagnosis of prostate cancer), overdiagnosis bias was avoided, as, in each trial, the distributions of noninvasive and invasive cases were the same or very similar in the two fairly large randomly allocated samples.

Unfortunately, it is currently impossible to identify patients with prostate cancer who, if left unscreened, would die. A similar problem may exist with regard to *in situ* breast cancer.[72] This problem suggests expanding the definition of a false-positive test to also include individuals with cancer who do not progress to an invasive, potentially lethal phase.

4.4.4 Additional Types of Bias

In addition to the biases extensively discussed previously in this chapter, the "natural history" of a study provides an appropriate framework to understand other types of bias. More specifically, different types of bias are related to different phases in the natural history of a study (**FIGURE 4-12**). In addition to selection and information, the epidemiologist should be concerned with the plausibility of his or her hypothesis or hypotheses and with the analysis, interpretation, and dissemination of study results. Plausibility bias occurs when the hypothesis lacks biological, sociological, or psychological

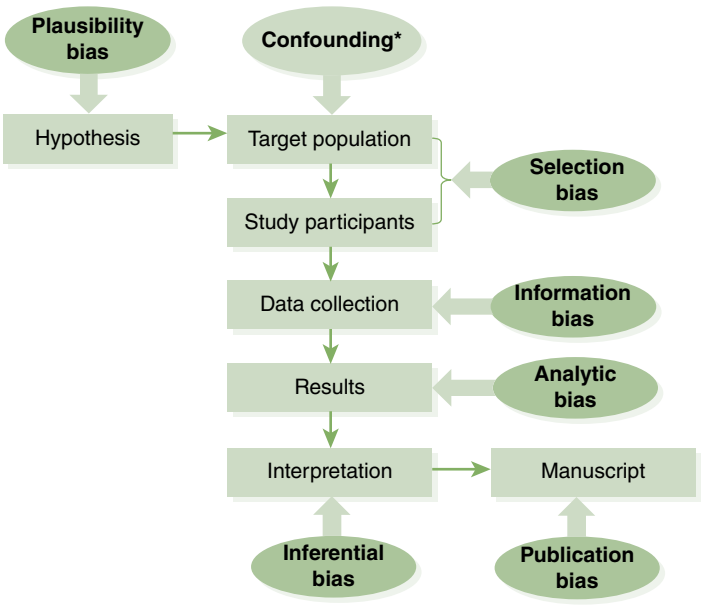


FIGURE 4-12 Natural history of a study and corresponding biases.

*Found in the population: only a bias for causal inference purposes.
Data from J. Samet, personal communication.

plausibility, for example, when postulating that chewing gum causes kidney cancer. An example of analytic bias is the use of a single analytic unit (e.g., 1 standard deviation) when the function is not linear (see Chapter 9, Section 9.3.4). Inferential bias may occur when causality is inferred from results of a single observational study.* [73] Publication bias, which has been mentioned previously, is discussed in Chapter 10, Section 10.5.

References

1. Sackett DL. Bias in analytical research. *J Chron Dis*. 1979;32:51-63.
2. Porta, M. *A Dictionary of Epidemiology*. 5th ed. New York, NY: Oxford University Press; 2008.
3. Lilienfeld DE, Stolley PD. *Foundations of Epidemiology*. 3rd ed. New York, NY: Oxford University Press; 1994.
4. Schlesselman J. *Case Control Studies: Design, Conduct, Analysis*. New York, NY: Oxford University Press; 1982.
5. Rothman K, Greenland S. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Wolters Kluwer Health/Lippincott; 2008.
6. Byrne C, Brinton LA, Haile RW, Schairer C. Heterogeneity of the effect of family history on breast cancer risk. *Epidemiology*. 1991;2:276-284.
7. MacMahon B, Yen S, Trichopoulos D, Warren K, Nardi G. Coffee and cancer of the pancreas. *N Engl J Med*. 1981;304:630-633.
8. Hsieh CC, MacMahon B, Yen S, Trichopoulos D, Warren K, Nardi G. Coffee and pancreatic cancer (Chapter 2). *N Engl J Med*. 1986;315:587-589.
9. Report from the Boston Surveillance Collaborative Program. Reserpine and breast cancer. *Lancet*. 1974;ii:669-671.
10. Labarthe DR, O'Fallon WM. Reserpine and breast cancer: a community-based longitudinal study of 2,000 hypertensive women. *J Am Med Assoc*. 1980;243:2304-2310.
11. Horwitz RI, Feinstein AR. Exclusion bias and the false relationship of reserpine and breast cancer. *Arch Intern Med*. 1985;145:1873-1875.
12. Heederik K. Micro-epidemiology of the healthy worker effect? *Occup Environ Med*. 2006;63:83.
13. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ: John Wiley & Sons; 1987.
14. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiologic and clinical research: potential and pitfalls. *Br Med J*. 2009:338:b2393.
15. Biering K, Hjollund NH, Frydenberg M. Using multiple imputation to deal with missing data and attrition in longitudinal studies with repeated measures of patient-reported outcomes. *Clin Epidemiol*. 2015;7:91-106.
16. Kmetic A, Lawrence J, Berger C, Tenenhouse A. Multiple imputation to account for missing data in a survey: estimating the prevalence of osteoporosis. *Epidemiology*. 2002;13:437-444.
17. Gordis L. *Epidemiology*. 5th ed. Philadelphia, PA: Elsevier Saunders; 2014.
18. Weinstock MA, Colditz GA, Willett WC, Stampfer MJ, Rosner B, Speizer FE. Recall (report) bias and reliability in the retrospective assessment of melanoma risk. *Am J Epidemiol*. 1991;133:240-245.
19. Brinton LA, Hoover RN, Szklo M, Fraumeni JF Jr. Menopausal estrogen use and risk of breast cancer. *Cancer*. 1981;47:2517-2522.
20. Chambless LE, Heiss G, Folsom AR, et al. Association of coronary heart disease incidence with carotid arterial wall thickness and major risk factors: the Atherosclerosis Risk in Communities (ARIC) Study, 1987-1993. *Am J Epidemiol*. 1997;146:483-494.
21. Bild D, Bluemke DA, Burke GL, et al. Multi-Ethnic Study of Atherosclerosis. *Am J Epidemiol*. 2002;156:871-881.
22. Mele A, Szklo M, Visani G, et al. Hair dye use and other risk factors for leukemia and pre-leukemia: a case-control study: Italian Leukemia Study Group. *Am J Epidemiol*. 1994;139:609-619.
23. Lilienfeld AM, Graham S. Validity of determining circumcision status by questionnaire as related to epidemiological studies of cancer of the cervix. *J Natl Cancer Inst*. 1958;21:713-720.
24. Wei Q, Matanoski GM, Farmer ER, Hedayati MA, Grossman L. DNA repair and susceptibility to basal cell carcinoma: a case-control study. *Am J Epidemiol*. 1994;140:598-607.
25. *The Health Effects of Passive Smoking*. Canberra: Australian Government Publishing Service. Commonwealth of Australia, National Health and Medical Research Council; November 1997.
26. Szklo M, Tonascia J, Gordis L. Psychosocial factors and the risk of myocardial infarctions in white women. *Am J Epidemiol*. 1976;103:312-320.

*Possible exceptions to the notion that causal inference is not appropriate when interpreting results from a single observational study include studies using Mendelian randomization or instrumental variables when the rather stringent assumptions related to these approaches are met (see Chapter 7, Section 7.5.1).

27. Doll R, Hill AB. Smoking and carcinoma of the lung: preliminary report. *Br Med J.* 1950;2:739-748.
28. Perneger TV, Whelton PK, Klag MJ, Rossiter KA. Diagnosis of hypertensive end-stage renal disease: effect of patient's race. *Am J Epidemiol.* 1995;141:10-15.
29. The ARIC investigators. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. *Am J Epidemiol.* 1989;129:687-702.
30. Stewart WF, Linet MS, Celentano DD, Van Natta M, Ziegler D. Age- and sex-specific incidence rates of migraine with and without visual aura. *Am J Epidemiol.* 1991;134:1111-1120.
31. Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. *Appl Psychol Meas.* 1977;3:385-401.
32. Rose GA. Chest pain questionnaire. *Milbank Mem Fund Q.* 1965;43:32-39.
33. Bass EB, Follansbee WP, Orchard TJ. Comparison of a supplemented Rose Questionnaire to exercise thallium testing in men and women. *J Clin Epidemiol.* 1989;42:385-394.
34. Garber CE, Carleton RA, Heller GV. Comparison of "Rose Questionnaire Angina" to exercise thallium scintigraphy: different findings in males and females. *J Clin Epidemiol.* 1992;45:715-720.
35. Sorlie PD, Cooper L, Schreiner PJ, Rosamond W, Szklo M. Repeatability and validity of the Rose Questionnaire for angina pectoris in the Atherosclerosis Risk in Communities Study. *J Clin Epidemiol.* 1996;49:719-725.
36. Dosemeci M, Wacholder S, Lubin JH. Does nondifferential misclassification of exposure always bias a true effect toward the null value? *Am J Epidemiol.* 1990;132:746-748.
37. Infante-Rivard C, Jacques L. Empirical study of parental recall bias. *Am J Epidemiol.* 2000;152:480-486.
38. Greenland S. The effect of misclassification in the presence of covariates. *Am J Epidemiol.* 1980;112:564-569.
39. Wacholder S. When measurement errors correlate with truth: surprising effects of nondifferential misclassification. *Epidemiology.* 1995;6:157-161.
40. Flegal KM, Brownie C, Haas JD. The effects of exposure misclassification on estimates of relative risk. *Am J Epidemiol.* 1986;123:736-751.
41. Flegal KM, Keyl PM, Nieto FJ. Differential misclassification arising from nondifferential errors in exposure measurement. *Am J Epidemiol.* 1991;134:1233-1244.
42. Willett W. An overview of issues related to the correction of non-differential exposure measurement error in epidemiologic studies. *Stat Med.* 1989;8:1031-1040; discussion 1071-1033.
43. Hessol NA, Schwarcz S, Ameli N, Oliver G, Greenblatt RM. Accuracy of self-reports of acquired immunodeficiency syndrome and acquired immunodeficiency syndrome-related conditions in women. *Am J Epidemiol.* 2001;153:1128-1133.
44. Osawa K, Nakanishi R, Budoff M. Coronary artery calcification. *Glob Heart.* 2016;11:287-293.
45. Thomas D, Stram D, Dwyer J. Exposure measurement error: influence on exposure-disease: relationships and methods of correction. *Annu Rev Public Health.* 1993;14:69-93.
46. Armstrong BG. The effects of measurement errors on relative risk regressions. *Am J Epidemiol.* 1990;132:1176-1184.
47. Kannel WB. CHD risk factors: a Framingham Study update. *Hosp Pract (Off Ed).* 1990;25:119-127, 130.
48. Giovannucci E, Ascherio A, Rimm EB, Colditz GA, Stampfer MJ, Willett WC. A prospective cohort study of vasectomy and prostate cancer in US men. *J Am Med Assoc.* 1993;269:873-877.
49. Goldberg RJ, Gorak EJ, Yarzebski J, et al. A communitywide perspective of sex differences and temporal trends in the incidence and survival rates after acute myocardial infarction and out-of-hospital deaths caused by coronary heart disease. *Circulation.* 1993;87:1947-1953.
50. Perneger TV, Nieto FJ, Whelton PK, Klag MJ, Comstock GW, Szklo M. A prospective study of blood pressure and serum creatinine: results from the "Clue" Study and the ARIC Study. *J Am Med Assoc.* 1993;269:488-493.
51. Szklo M. Aplastic anemia. In: Lilienfeld AM, ed. *Reviews in Cancer Epidemiology.* Vol 1. New York, NY: Elsevier North-Holland; 1980:115-119.
52. Antunes CM, Strolley PD, Rosenshein NB, et al. Endometrial cancer and estrogen use: report of a large case-control study. *N Engl J Med.* 1979;300:9-13.
53. Horwitz RI, Feinstein AR. Estrogens and endometrial cancer: responses to arguments and current status of an epidemiologic controversy. *Am J Med.* 1986;81:503-507.
54. Brunekreef B, Groot B, Hoek G. Pets, allergy and respiratory symptoms in children. *Int J Epidemiol.* 1992;21:338-342.
55. Nieto FJ, Diez-Roux A, Szklo M, Comstock GW, Sharrett AR. Short- and long-term prediction of clinical and subclinical atherosclerosis by traditional risk factors. *J Clin Epidemiol.* 1999;52:559-567.
56. Hammond EC, Horn D. Smoking and death rates: report on forty-four months of follow-up of 187,783 men. 2. Death rates by cause. *J Am Med Assoc.* 1958;166:1294-1308.
57. Shapiro S, Venet W, Strax P, Venet L, Roeser R. Selection, follow-up, and analysis in the Health Insurance Plan Study: a randomized trial with breast cancer screening. *Natl Cancer Inst Monogr.* 1985;67:65-74.
58. Hutchison GB, Shapiro S. Lead time gained by diagnostic screening for breast cancer. *J Natl Cancer Inst.* 1968;41:665-681.

59. Morrison AS. *Screening in Chronic Disease*. New York, NY: Oxford University Press; 1992.
60. Stamey TA, McNeal JE, Yemoto CM, Sigal BM, Johnstone IM. Biological determinants of cancer progression in men with prostate cancer. *J Am Med Assoc.* 1999;281:1395-1400.
61. Howlader N, Noone AM, Krapcho M, et al., eds. *SEER Cancer Statistics Review, 1975-2011*. Bethesda, MD: National Cancer Institute; 2014.
62. Andriole GL, Crawford ED, Grubb RL III, et al. Mortality results from a randomized prostate-cancer screening trial. *New Eng J Med.* 2009;360:1310-1319.
63. Schröder FH, Hugosson J, Roobol MJ, et al. Screening and prostate-cancer mortality in a randomized European study. *New Eng J Med.* 2009;360:1320-1328.
64. Schröder FH, Hugosson J, Roobol MJ, et al. Screening and prostate cancer mortality: results of the European Randomised Study of Screening for Prostate Cancer (ERSPC) at 13 years of follow-up. *Lancet*. 2014;384(9959):2027-2035.
65. Franks LM. Latent carcinoma of the prostate. *J Pathol Bacteriol.* 1954;68:603-616.
66. Bostwick DG, Cooner WH, Denis L, Jones GW, Scardino PT, Murphy GP. The association of benign prostatic hyperplasia and cancer of the prostate. *Cancer.* 1992;70:291-301.
67. Breslow, Chan CW, Dhom G, et al. Latent carcinoma of prostate at autopsy in seven areas. *Int J Cancer.* 1977;20:680-688.
68. Baron E, Angrist A. Incidence of occult adenocarcinoma of the prostate after fifty years of age. *Arch Pathol.* 1941;32:787-793.
69. Edwards CN, Steinthorsson E, Nicholson D. An autopsy study of latent prostatic cancer. *Cancer.* 1953;6:531-554.
70. Halpert B, Schmalhorst WR. Carcinoma of the prostate in patients 70 to 79 years old. *Cancer.* 1966;695-698.
71. Scott R, Mutchnik DL, Laskowski TZ, Schmalhorst WR. Carcinoma of the prostate in elderly men: incidence, growth characteristics and clinical significance. *J Urol.* 1969;101:602-607.
72. Schonberg MA, Marcantonio ER, Ngo L, Li D, Silliman RA, McCarthy EP. Causes of death and relative survival of older women after a breast cancer diagnosis. *J Clin Oncol.* 2011;29:1570-1577.
73. Petitti D. Associations are not effects. *Am J Epidemiol.* 1991;133:101-102.

CHAPTER 15

Precision and Study Size

Kenneth J. Rothman, and Timothy L. Lash

We use the term *accuracy* to describe an estimate of an epidemiologic measure that is close to the estimand. Two types of error, systematic and random, detract from accuracy. In earlier chapters, we have considered sources of systematic error, including biases related to selection of study participants, measurement errors of study variables, and confounding. In this chapter, we discuss methods to measure, limit, and account for random error in an epidemiologic study and how to interpret these methods properly.

RANDOM ERROR AND STATISTICAL PRECISION

What is random error? It is often equated with chance or random variation, which itself is rarely well defined. Many people believe that chance plays a fundamental role in all physical and, by implication, biologic phenomena. For some, the belief in chance is so dominant that it vaults random occurrences into an important role as component causes of all we experience. Others believe that causality may be viewed as deterministic, meaning that a full elaboration of the relevant factors in a set of circumstances will lead, on sufficient analysis, to a perfect prediction of effects resulting from these causes. Under the latter view, all experience is predestined to unravel in a theoretically predictable way that follows from the previous pattern of actions. Even with this extreme deterministic view, however, one must face the fact that only rarely could one acquire sufficient knowledge to predict effects perfectly, and then only for trivial cause-effect patterns. The resulting incomplete predictability of determined outcomes makes their residual variability indistinguishable from random occurrences.

A unifying description of incomplete predictability can thus be forged. In this description, random variation equates with a component of ignorance about causes of study outcomes, an ignorance that is inevitable whether the outcome is deterministic or is partly random. For example, predicting the outcome of a tossed coin represents a physical problem, the solution of which is feasible through the application of physical laws. Whether the sources of variation that we cannot explain are actually chance phenomena makes little difference. We treat such variation as being random until we can explain it, and thereby reduce it, by relating it to known factors.

In an epidemiologic study, random variation has many sources, but a major contributor is the process of selecting the specific study participants. This process is usually referred to as *sampling*; the

attendant random variation is known as *sampling variation* or *sampling error*. Case-control studies sometimes involve a physical sampling process, whereas cohort studies often do not. Nevertheless, it is a standard practice to treat all epidemiologic studies, including cohort studies, as having sampling error. In this view, the subjects in a study, whether physically sampled or not, constitute a figurative sample of possible people who could have been included in the study or of the different possible experiences the study subjects could have had. Even if all the individuals in a population were included in a study, the study subjects are viewed as a sample of the potential biologic experience of an even broader conceptual population. Under this view, the statistical dictum that there is no sampling error if an entire population (as opposed to a sample of it) is studied does not apply to epidemiologic studies, even if an entire population is included in the study. Conceptually, the actual subjects are always considered a sample of a broader experience of interest—although they seldom actually satisfy the definition of a random sample that underpins the statistical models ordinarily used to measure random variation.[1, 2]

Sampling is only one source of random error that contributes to unpredictable inaccuracies in epidemiologic studies. Another source is the unexplained variation in occurrence measures, such as observed incidence rates or prevalence proportions. Sources of systematic errors also abound. For example, when exposure status is not randomly assigned, confounding (see Chapter 12) may lead to deviations of estimated associations from target effects that far exceed what standard statistical models assume probable. Mismeasurement of key study variables also contributes to the overall inaccuracy, in both random and systematic ways. As a result of these extra sources of variation, and because of the weak theoretical underpinnings for conceptualizing study subjects as a sample of a broader experience, the usual statistical tools that we use to measure random variation at best provide minimum estimates of the actual uncertainty we should have about the estimand. One

elementary way to improve the quantification of our uncertainty is through *bias analysis,* which we discuss in Chapter 27.

A common measure of random variation in a measurement or estimation process is the *variance* of the process. The *statistical precision of* (or *statistical information in*) a measurement or process is often taken to be the inverse of the variance of the measurements or estimates that the process produces. In this sense, precision is the opposite of random error. Precision of estimation can be improved (which is to say, variance can be reduced) by increasing the size of the study. Precision can also be improved by modifying the design of the study to decrease the variance, given a fixed total number of subjects; this process is called improving the *statistical efficiency* of the study. Perhaps, the most common epidemiologic example of such design improvement is the use of a case-control study rather than a cohort study, because for a fixed study size, the variance of an effect estimate is heavily dependent on the proportion of subjects in the study that are cases, and case-control studies increase this proportion by design.

APPROACHES TO EVALUATING RANDOM ERROR

Statistics and its role in data analysis have undergone a gradual but profound transformation in recent times. There is an essential distinction between a qualitative study objective (to answer a question "yes" or "no") and a quantitative one (to measure something). The recent transformation reflects a growing preference for the latter objective and for statistical methods consistent with it. Until the 1970s, most applications of statistics in epidemiology focused on deciding whether "chance" or "random error" could be solely responsible for an observed association, or equivalently, whether exposure x was related to outcome y. The methods used for this decision were those of classical *significance testing*, predominant in British applications, and those of Neyman-Pearson *hypothesis testing*, predominant in American applications.[3, 4] Because of their similarities, the term *significance testing* is often applied to both collections of methods.

These testing applications, which were subject to some early criticism,[5-8] came under growing criticism by epidemiologists and statisticians during the late 20th century, which intensified in the 2010s.[9-12] The critics pointed out that most, if not all, epidemiologic applications need more than a decision as to whether chance alone could have produced an association. More important is the estimation of the magnitude of the association, including an assessment of the precision of the estimation method. The estimation tool used by most authors is the confidence interval, which provides a range of values for the association, under the hypothesis that only random variation has created discrepancies between the true value of the association under study and the value observed in the data.[13] Other authors, while favoring the move toward interval estimation, point out that confidence intervals suffer from some of the flaws associated with significance testing and favor other approaches to interval estimation.[14_17]

Significance Testing and Hypothesis Testing

Nearly 70 years ago, Berkson[6] wrote:

> It is hardly an exaggeration to say that statistics, as it is taught at present in the dominant school, consists almost entirely of tests of significance, though not always presented as such, some comparatively simple and forthright, others elaborate and abstruse.

The ubiquitous use of *P*-values and references to "statistically significant" findings in the current medical literature demonstrates the dominant role that statistical hypothesis testing still plays in data analysis in some branches of biomedical sciences. Many researchers still believe that it would be fruitless to submit for publication any paper that lacks statistical tests of significance. Until recently, their belief was not entirely ill-founded, because many journal editors and referees relied on tests of significance as indicators of sophisticated and meaningful statistical analysis as well as the primary means of assessing sampling variability in a study.[18] *Statistical significance* is usually based on the *P*-value (described below): results are considered "significant" or "not significant" according to whether the *P*-value is less than or greater than an arbitrary cutoff value, usually 0.05, which is called the *alpha level* of the test.

The preoccupation with significance testing derives from the research interests of the statisticians who pioneered the development of statistical theory in the early 20th century. Their research problems were primarily industrial and agricultural, and they typically involved randomized experiments or random-sample surveys that formed the basis for a choice between two or more alternative courses of action. Such studies were designed to produce results that would enable a decision to be made, and the statistical methods employed were intended to facilitate decision-making. The concepts that grew out of this heritage are today applied in clinical

and epidemiologic research, and they strongly reflect this background of decision-making.

Statistical significance testing of associations usually focuses on the *null hypothesis*, which is usually formulated as a hypothesis of no association between two variables in a *superpopulation*, the population from which the observed study groups were purportedly sampled in a random fashion. For example, one may test the hypothesis that the risk difference (RD) in the superpopulation is 0 or, equivalently, that the risk ratio (RR) is 1. Note that this hypothesis is about the superpopulation, *not* about the observed study groups. Testing may alternatively focus on any other specific hypothesis, *e.g.*, that the RD is 0.1 or the RR is 2. For non-null hypotheses, tests about one measure (*e.g.*, RD) are not usually equivalent to tests about another measure (*e.g.*, an RR), so one must choose a measure of interest to perform a non-null test.

A common misinterpretation of significance tests[10] is to claim that there is no difference between two observed groups because the null test is not statistically significant, in that *P* is greater than the cutoff for declaring statistical significance (again, usually .05). This interpretation confuses a descriptive issue (whether the two observed groups differ) with an inference about the superpopulation. The significance test refers only to the superpopulation, not the observed groups. To say that the difference is not statistically significant means only that one cannot reject the null hypothesis that the superpopulation groups are the same; it does *not* imply that the two observed groups are the same.

One needs only to look at the two observed groups to see whether they are different. Significance testing concerns instead whether the observed difference should lead one to infer that there is a difference between the corresponding groups in the superpopulation. Furthermore, even if the observed difference is not statistically significant, the superpopulation groups may be different (*i.e.*, the result does not imply that the null is correct). Rather, the nonsignificant observed difference means only that one should not

rule out the null hypothesis if one accepts the statistical model used to construct the test.

Conversely, it is a misinterpretation to claim that an association exists in the superpopulation because the observed difference is statistically significant. First, the test may be significant only because the model used to compute it is wrong (*e.g.*, there may be many sources of uncontrolled bias). Second, the test may be significant because of chance alone; for example, even under perfect conditions, a test using a 0.05 alpha level will yield a statistically significant difference 5% of the time if the null hypothesis is correct.

As we emphasize, the alpha cutoff point is an arbitrary and questionable convention; it can be dispensed simply by reporting the actual *P*-value from the test, which we now discuss in detail. We will then further explore and criticize the theory that led to the widespread use of arbitrary testing cutoffs in research.

P-Values

P-values come in two types: one-tailed and two-tailed. Further, there are two types of one-tailed *P*-values: upper and lower. An *upper* one-tailed *P*-value is the probability that a corresponding quantity computed from the data, known as the *test statistic* (such as a *t*-statistic or a chi-square statistic), will be greater than or equal to its observed value, assuming that (a) the test hypothesis is correct and (b) there is no source of bias in the data collection or analysis processes. Similarly, a *lower* one-tailed *P*-value is the probability that the corresponding test statistic will be less than or equal to its observed value, again assuming that (a) the test hypothesis is correct and (b) there is no source of bias in the data collection or analysis processes (sometimes described by saying that the underlying statistical model is correct). The two-tailed *P*-value is usually defined as twice the smaller of the upper and lower *P*-values, although more complicated definitions have been used. Being a probability, a one-tailed *P*-value must fall between 0 and 1; the two-tailed *P*-value as just defined, however, may exceed 1. The following comments apply to all types of *P*-values. Some authors refer to *P*-values as "levels of

significance,"[19] but the latter term is best avoided because it has been used by other authors to refer to alpha levels.

In the classical significance testing paradigm, small *P*-values are supposed to indicate that at least one of the assumptions used to derive it is incorrect, that is, either or both the test hypothesis (assumption a) or the statistical model (assumption b) is incorrect. All too often, the statistical model is taken as given, so that a small *P*-value is taken as indicating a low degree of compatibility between the test hypothesis and the observed data. This incompatibility derives from the fact that a small *P*-value represents a low probability of getting a test statistic as extreme as or more extreme than the observed statistic if the test hypothesis is true and no bias is operative. Small *P*-values, therefore, are supposed to indicate that the test hypothesis is not an acceptable explanation for the association observed in the data. This common interpretation has been extensively criticized because it does not account for alternative explanations and their acceptability (or lack thereof); for example, refer to the study of Berkson[6] and later epidemiologic criticisms by Goodman and Royall,[15] Greenland,[1] Goodman,[4] and Gigerenzer.[3] A less hypothetical and more cautious interpretation is then that a small *P*-value indicates that there is a problem with the test hypothesis or with the study, or with both.[20]

A common but naive misinterpretation of *P*-values is that they represent probabilities of test hypotheses. In many situations, one can compute a Bayesian probability, or credibility for the test hypothesis, but it will almost always be far from the two-tailed *P*-value.[21, 22] A one-tailed *P*-value can be used to put a lower bound on the Bayesian probability of certain compound hypotheses,[23] and under certain conditions, it will approximate the Bayesian probability that the true association is the opposite of the direction observed.[24] Nonetheless, a *P*-value for a simple test hypothesis (for example, exposure and disease are unassociated) is not a probability of that hypothesis. That *P*-value is usually much smaller than such a Bayesian probability and so can easily mislead one into inappropriately rejecting the test hypothesis.[15, 22]

Another incorrect interpretation is that the *P*-value is the probability of the observed data under the test hypothesis. This probability is known as the likelihood of the test hypothesis, see Goodman and Royall,[15] Royall,[17] Edwards,[25] and the following discussion. The likelihood of a hypothesis is usually much smaller than the *P*-value for the hypothesis, because the *P*-value includes not only the probability of the observed data under the test hypothesis, but also the probabilities for all other possible data configurations in which the test statistic was more extreme than that observed.

A subtle and common misinterpretation of a *P*-value for testing the null hypothesis is that it represents the probability that the data would show as strong an association as observed or stronger if the null hypothesis were correct. This misinterpretation can be found in many methodologic articles and textbooks. The nature of the misinterpretation can be seen in a study of a risk difference (RD). The study might produce an estimate of RD of 0.33 with an estimated standard error (or standard deviation) of 0.20, which would produce a standard normal test statistic of $z = 0.33/0.20 = 1.65$ and a two-tailed $P = 0.10$. The same study, however, might have instead estimated an RD of 0.30 and a standard deviation of 0.15, which would produce a standard normal test statistic of $z = 0.30/0.15 = 2.00$ and $P = 0.05$. The result with the association nearer the null would then produce a smaller *P*-value. The point is that the *P*-value refers to the size of the test statistic (which in this case is the estimate divided by its estimated standard deviation), not to the strength or size of the estimated association.

It is crucial to remember that *P*-values are calculated from statistical models, which are assumptions about the form of study-to-study data variation. Every *P*-value, even "nonparametric" and "exact" *P*-values, depends on a statistical model; it is only the strength of the model assumptions that differs.[26, 27] A major problem with the *P*-values and tests in common use (including all commercial software) is that the assumed models make no allowance for sources of bias, apart from confounding by controlled covariates.

Neyman-Pearson Hypothesis Tests

A *P*-value is a continuous measure of the compatibility between a hypothesis and data. Although its utility as such a measure can be disputed,[15, 17] a worse problem is that it is often used to force a qualitative decision about rejection of a hypothesis. As introduced earlier, a fixed cutoff point or alpha level, often denoted by the Greek letter α (alpha), is selected as a criterion with which the *P*-value is judged. This point is then used to classify the observation either as "significant at level α" if $P \leq \alpha$, in which case the test hypothesis is rejected, or "not significant at level α" if $P > \alpha$, in which case the test hypothesis is accepted (or, at least, not rejected).

The use of a fixed cutoff α is a hallmark of the Neyman-Pearson form of statistical hypothesis testing. Both the alpha level[28] and the *P*-value[4, 29] have been called the "significance level" of the test. This usage has led to misinterpretation of the *P*-value as the alpha level of a statistical hypothesis test. To avoid the error, one should recall that the *P*-value is a quantity computed from the data, whereas the alpha level is a fixed cutoff (usually 0.05) that can be specified without even seeing the data. (As a technical aside, Neyman and Pearson actually avoided use of *P*-values in their formulation of hypothesis tests and instead defined their tests based on whether the value of the test statistic fell within a "rejection region" for the test.)

An incorrect rejection is called a *Type I error*, or alpha error. A hypothesis testing procedure is said to be *valid* if, whenever the test hypothesis is true, the probability of rejection (*i.e.*, the probability that $P \leq \alpha$) does not exceed the alpha level (provided there is no bias and all test assumptions are satisfied). For example, a valid test with $\alpha = 0.01$ (a 1% alpha level) will lead to a Type I error with no more than 1% probability, provided there is no bias or incorrect assumption.

If the test hypothesis is false but is not rejected, the incorrect decision not to reject is called a *Type II*, or beta error. If the test hypothesis is false, so that rejection is the correct decision, the probability (over repetitions of the study) that the test hypothesis is rejected is called the *power* of the test. The probability of a Type II

error is related to the power by the equation *Pr* (Type II error) = 1 − power.

There is a trade-off between the probabilities of a Type I and a Type II error. This trade-off depends on the chosen alpha level. Reducing the Type I error when the test hypothesis is true requires a smaller alpha level, for with a smaller alpha level a smaller *P*-value will be required to reject the test hypothesis. Unfortunately, a lower alpha level increases the probability of a Type II error if the test hypothesis is false. Conversely, increasing the alpha level reduces the probability of Type II error when the test hypothesis is false, but increases the probability of Type I error if it is true.

The concepts of alpha level, Type I error, Type II error, and power stem from a paradigm in which data are used to decide whether to reject the test hypothesis, and therefore follow from a qualitative study objective. The extent to which decision-making dominates research thinking is reflected in the frequency with which the *P*-value, a continuous measure, is reported or interpreted only as an inequality (such as $P < 0.05$ or $P > 0.05$) or else not at all, with the evaluation focusing instead on "statistical significance" or its absence.

When a single study forms the sole basis for a choice between two alternative actions, as in industrial quality-control activities, a decision-making mode of analysis may be justifiable. Even then, however, a rational recommendation about which of two actions is preferable will require consideration of the costs and benefits of each action. These considerations are rarely incorporated into statistical tests. In most scientific and public health settings, it is presumptuous if not absurd for an investigator to act as if the results of his or her study will form the sole basis for a decision. Such decisions are inevitably based on results from a collection of studies, and proper combination of the information from the studies requires more than just a classification of each study into "significant" or "not significant". Thus, degradation of information about an effect into a simple dichotomy is counterproductive, even for decision-making, and can be misleading.

In a classic review of 71 clinical trials that reported no "significant" difference between the compared treatments, Freiman et al.[30] found that in the great majority of such trials, the data either indicated or at least were consistent with a moderate or even reasonably strong effect of the new treatment (Figure 15-1). In all of these trials, the original investigators interpreted their data as indicative of no effect because the *P*-value for the null hypothesis was not "statistically significant." The misinterpretations arose because the investigators relied solely on hypothesis testing for their statistical analysis rather than on estimation. On failing to reject the null hypothesis, the investigators in these 71 trials inappropriately accepted the null hypothesis as correct, which probably resulted in Type II error for many of these so-called negative studies.

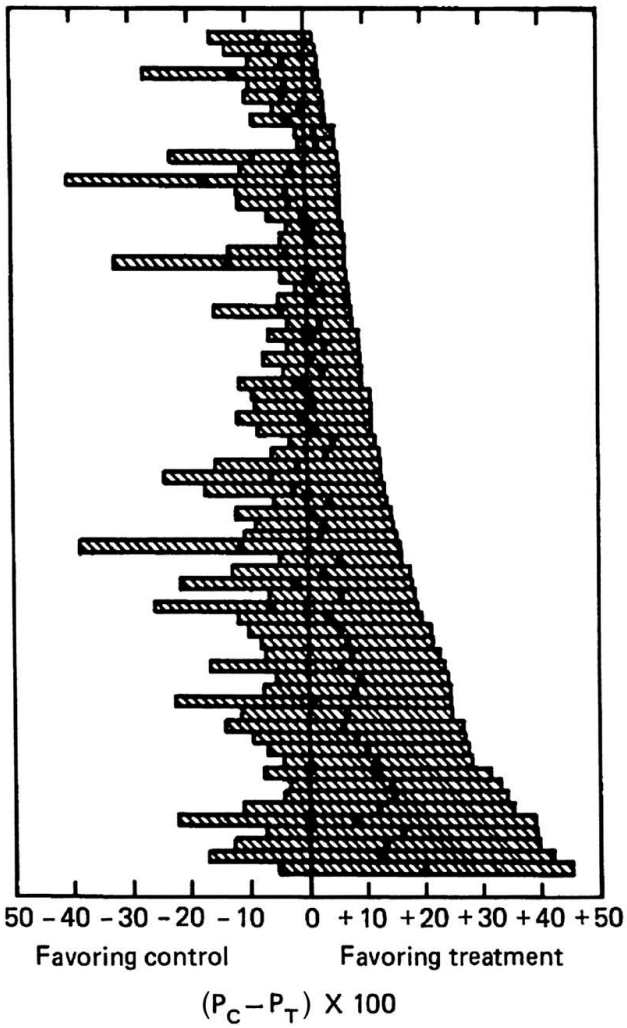


Figure 15.1 Ninety percent confidence limits for the true percentage difference for the 71 trials. The vertical bar at the center of each interval indicates the observed difference, P_C-P_T, for each trial. (Reproduced with permission from Freiman JA, Chalmers TC, Smith H, et al. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. *N Engl J Med*. 1978;299:690-694.)

Type II errors result when the magnitude of an effect, biases, and random variability combine to give results that are insufficiently inconsistent with the null hypothesis to reject it. This failure to reject the null hypothesis can occur because the effect is small, the observations are too few, or both, as well as from biases. More to the point, however, is that Type I and Type II errors arise because the

investigator has attempted to dichotomize the results of a study into the categories "significant" or "not significant." Because this degradation of the study information is unnecessary, an "error" that results from an incorrect classification of the study result is also unnecessary.

Why has such an unsound practice as Neyman-Pearson (dichotomous) hypothesis testing become so ingrained in scientific research? Undoubtedly, much of the popularity of hypothesis testing stems from the apparent objectivity and definitiveness of the pronouncement of significance.[31] Declarations of significance or its absence can supplant the need for more refined interpretations of data; the declarations can serve as a mechanical substitute for thought, promulgated by the inertia of training and common practice. The neatness of an apparent clear-cut result may appear more gratifying to investigators, editors, and readers than a finding that cannot be immediately pigeonholed.

The unbridled authority given to statistical significance in the social sciences has also been attributed to the apparent objectivity that the pronouncement of significance can convey[32]:

> "Let us look and see what is significant" is not too far from the approach of some researchers, and when the data involve perhaps several hundred variables, the practical temptations to use a ready-made decision rule are enormous. . . . [T]he pressure to *decide,* in situations where the very use of probability models admits the uncertainty of the inference, has certain consequences for the presentation of knowledge. The significance test appears to guarantee the objectivity of the researcher's conclusions, and may even be presented as providing crucial support for the whole theory in which the research hypothesis was put forward. As we have seen, tests of significance cannot do either of these things—but it is not in the interests of anyone involved to admit this too openly.

The origin of the nearly universal acceptance of the 5% cutoff point for significant findings is tied to the abridged form in which the chi-square table was originally published.[27] Before computers and calculators could easily give quick approximations to the chi-square distribution, tables were used routinely. Because there is a different chi-square distribution corresponding to every possible value for the degrees of freedom, the tables could not give many points for any one distribution. The tables typically included values at 1%, 5%, and a few other levels, encouraging the practice of checking the chi-square statistic calculated from one's data to see if it exceeded the cutoff levels in the table. In the original formulation of the Neyman and Pearson hypothesis testing, the alpha level was supposed to be determined from contextual considerations, especially the cost of Type I and Type II errors. This more thoughtful aspect of their theory was rapidly lost when the theory entered common scientific use. In fact, the default values assigned to acceptable Type I and Type II errors often run contrary to public health values. Imagine, for example, a study of the effect of a pollution source on the health of a nearby community, designed with acceptable Type I error rate of $\alpha = 0.05$ and with acceptable Type II error rate of $\beta = 0.2$ (80% power). These imply that a Type II error is four times worse than a Type I error. Limiting the Type I error protects the polluter by keeping low the probability of a false-positive result, which would be to declare that it is harmful to the community when it is not. Limiting the Type II error protects the community by keeping low the probability of a false-negative result, declaring no harm when in fact there is harm. Setting a Type II error to be 0.20, four times greater than the Type I error of 0.05, suggests that protecting the pollution source is worth four times as much as protecting the community. It would be a valuable exercise for investigators to write out statements such as these at the design stage to clarify whether they comport with their personal and professional values.

The Alternative Hypothesis

Another hallmark of Neyman-Pearson hypothesis testing, and perhaps one that most distinguishes it from earlier significance-testing paradigms, is that if the test hypothesis is rejected, it is supposed to be rejected in favor of some alternative hypothesis. The alternative hypothesis may be very specific, but more often it is implicit and very broad. For example, if the test hypothesis postulates that there is no association, then the usual (implicit) alternative hypothesis is that there is an association. Such nonspecific alternatives lead to nondirectional tests based on comparing a two-tailed *P*-value from a directional test statistic against the alpha level. Because this *P*-value is sensitive to violations of the test hypothesis in either direction, it is often called a *two-sided P*-value.

Nonetheless, the test and alternative hypotheses can instead be one-sided (directional). For example, the test hypothesis could state that an association is not positive (that is, either null or negative). The alternative is then that the association is positive. Such an alternative leads to use of a *one-sided* test based on comparing an *upper-tailed P*-value from a directional test statistic against alpha. Because this one-tailed *P*-value is sensitive to violations of the test hypothesis in only one direction, it is often called a *one-sided P*-value. An analogous one-sided test that the association was not negative would employ the lower-tailed *P*-value; the alternative for this test is that the association is negative.

Another form of the alternative hypothesis is a finite interval of "equivalence" about the null, for example, that the RD is between −0.1 and +0.1. This alternative is found in comparisons of two treatments (so that the "exposed" are those given one treatment and the "unexposed" are those given another treatment). The bounds of the interval are selected so that any value within the interval is considered close enough to the null for practical purposes. The test hypothesis is then that the two treatments are not equivalent (RD is outside the interval) and is rejected if *P* is less than alpha for all values outside the interval of equivalence. This approach is called *equivalence testing,* and it corresponds to rejecting the test hypothesis

when the $1 - \alpha$ confidence interval falls entirely within the equivalence interval.[33]

Note that the alternative hypothesis in all these examples comprises a range of values. For a two-sided test, the alternative comprises every possible value except the one being tested. For epidemiologic effect measures, this two-sided alternative hypothesis will range from absurdly large preventive effects to absurdly large causal effects and include everything in between except the test hypothesis. This hypothesis will be compatible with any observed data. The test hypothesis, on the other hand, corresponds to a single value of effect and therefore is readily consistent with a much narrower range of possible outcomes for the data. Statistical hypothesis testing amounts to an attempt to falsify the test hypothesis. It is natural to focus on a test hypothesis that is as specific as possible because it is easier to marshal evidence against a specific hypothesis than a broad one. The equivalence-testing example shows, however, that in some cases, the alternative may be more specific than the test hypothesis, and the test hypothesis may range from absurdly large preventive effects to absurdly large causal effects.

A major defect in the way all the above alternatives are usually formulated is that they assume the statistical model is correct. Because the model is never exactly correct and is often grossly incorrect, a scientifically more sound formulation of the alternative to the null hypothesis (for example) would be "either the null is false or else the statistical model is wrong."[20] By adding the warning "or else the statistical model is wrong" to the alternative, we allow for the possibility that uncontrolled systematic errors were responsible for the rejection.

Statistical Estimation

If Neyman-Pearson hypothesis testing is misleading, how should results be interpreted and presented? In keeping with the view that science is based on measurement—which leads in turn to quantitative study objectives—the analysis of epidemiologic data can

be conceptualized as a measurement problem rather than as a problem in decision-making. Measurement requires more detailed statistics than the simple dichotomy produced by a statistical hypothesis testing. Whatever the parameter that is the target of inference in an epidemiologic study—usually an effect measure, such as a ratio of rates or risks, but it can also be an incidence rate or any other epidemiologic measure—it will be measured on a continuous scale, with a theoretically infinite number of possible values.

The data from a study can be used to generate an estimate of the target parameter. An estimate may be presented as a single value on the measurement scale of the parameter; this value is referred to as a *point estimate.* A point estimate may be viewed as a measure of the extent of the association, or (in causal analyses) the magnitude of effect under study. There will be many forces that will determine the final data values, such as confounding, measurement error, selection biases, and "random" error. It is thus extremely unlikely that the point estimate will equal the true value of the parameter.

It might be tempting to believe that an inferential emphasis on estimation can coexist with an inferential emphasis on testing, but they cannot. Selecting results for attention based on statistical significance distorts the ability to achieve accurate measurement.[11] For example, when power of a study is less than 100%, statistically significant results are more likely to overestimate than to underestimate the true value for the parameter, and results that are not statistically significant are more likely to underestimate than to overestimate the true value for the parameter. This distortion, a mathematical fact, is an inevitable consequence of the selection pressures exerted by the significance testing. Selecting results for attention on the basis of statistical significance also fosters analytic manipulations to achieve statistical significance. "Significance questing" and "P-hacking" describe unplanned changes to study design or analysis implemented to achieve a statistically significant result.[34-36] These changes hinder accurate estimation. In short, selecting results for attention based on statistical significance

introduces distortions that preclude the goal of accurate estimation, so the two inferential paradigms cannot coexist.

Confidence Intervals and Confidence Limits

One way to account for random error in the estimation process is to compute *P*-values for a broad range of possible parameter values (in addition to the null value). If the range is broad enough, we will be able to identify an interval of parameter values for which the test *P*-value exceeds a specified alpha level (typically 0.05). All parameter values within the range are compatible with the data under the standard interpretation of significance tests. The range of values is called a *confidence interval*, and the endpoints of that interval are called *confidence limits.* The process of calculating the confidence interval is an example of the process of *interval estimation.*

The width of a confidence interval depends on the amount of random variability inherent in the data-collection process (as estimated from the underlying statistical model and the data). It also depends on an arbitrarily selected alpha level that specifies the degree of compatibility between the limits of the interval and the data. One minus this alpha level (0.95 if alpha is 0.05) is called the *confidence level* of the interval and is usually expressed as a percentage.

If the underlying statistical model is correct and there is no bias, a confidence interval derived from a valid test will, over unlimited repetitions of the study, contain the true parameter with a frequency no less than its confidence level. This definition specifies the coverage property of the method used to generate the interval, not the probability that the true parameter value lies within the interval. For example, if the confidence level of a valid confidence interval is 90%, the frequency with which the interval will contain the true parameter will be at least 90%, if there is no bias. Consequently, under the assumed model for random variability (*e.g.*, a binomial model) and with no bias, we should expect the confidence interval to include the true parameter value in at least 90% of replications of the process of obtaining the data. Unfortunately, this interpretation for

the confidence interval is based on probability models and sampling properties that are seldom realized in epidemiologic studies; consequently, it is preferable to view the confidence limits as only a rough estimate of the uncertainty in an epidemiologic result due to random error alone. Even with this limited interpretation, the estimate depends on the correctness of the statistical model, which may be incorrect in many epidemiologic settings.[1]

Relation of Confidence Intervals to Significance Tests and Hypothesis Tests

Consider now the relation between the confidence level and the alpha level of hypothesis testing. The confidence level equals 1 minus the alpha level ($1 - \alpha$) of the test used to construct the interval. To understand this relation, consider the diagram in Figure 15-2. Suppose that we performed a test of the null hypothesis with $\alpha = 0.10$. The fact that the 90% confidence interval does not include the null point indicates that the null hypothesis would be rejected for $\alpha = 0.10$. On the other hand, the fact that the 95% confidence interval includes the null point indicates that the null hypothesis would not be rejected for $\alpha = 0.05$. Because the 95% interval includes the null point and the 90% interval does not, it can be inferred that the two-sided *P*-value for the null hypothesis is greater than 0.05 and less than 0.10.

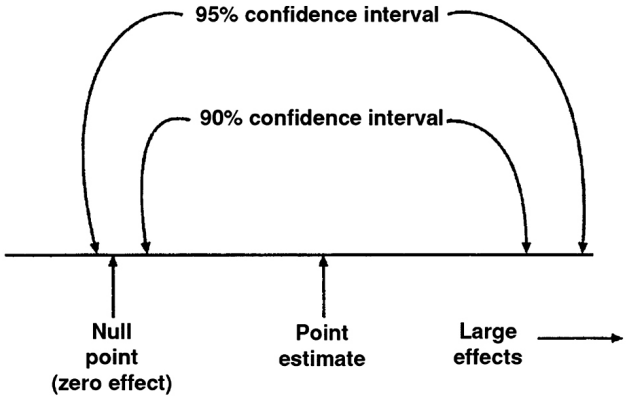


Figure 15.2 Two nested confidence intervals, with the wider one including the null hypothesis.

The point of the preceding example is not to suggest that confidence limits should be used as surrogate tests of significance. Although they can be and often are used this way, doing so defeats all the advantages that confidence intervals have over hypothesis tests. An interval-estimation procedure does much more than assess the extent to which a hypothesis is compatible with the data. It provides simultaneously an idea of the likely direction and magnitude of the underlying association and the random variability of the point estimate. The two-sided *P*-value, on the other hand, indicates only the degree of consistency between the data and a single hypothesis and thus reveals nothing about the magnitude or even the direction of the association or the random variability of the point estimate.[37]

For example, consider the data in Table 15-1. An exact test of the null hypothesis that the exposure is not associated with the disease gives a two-sided *P*-value of 0.14. (The methods used to calculate this *P*-value are described in Chapter 17.) This result might be reported in several ways. The least informative way is to report that the observed association is not significant. Somewhat more information can be given by reporting the actual *P*-value; to express the *P*-value as an inequality such as $P > 0.05$ is not much better than reporting the results as not significant, whereas reporting $P = 0.14$ at least gives the *P*-value explicitly rather than degrading it into a

dichotomy. An additional improvement is to report $P_2 = 0.14$, denoting the use of a two-sided rather than a one-sided *P*-value.

TABLE 15-1

Hypothetical Data From a Cohort Study, Corresponding to the *P*-Value Function in Figure 15-3

	Exposure	
	Yes	**No**
Cases	9	2
Person-Years	186	128

Any one *P*-value, no matter how explicit, fails to convey the descriptive finding that exposed individuals had about three times the rate of disease as unexposed subjects. Furthermore, exact 95% confidence limits for the true rate ratio are 0.7 and 13. The fact that the null value (which, for the rate ratio, is 1.0) is within the interval tells us the outcome of the significance test: The estimate would not be statistically significant at the $1 - 0.95 = 0.05$ alpha level. The confidence limits, however, indicate that these data, although statistically compatible with no association, are even more compatible with a strong association—assuming that the statistical model used to construct the limits is correct. Stating the latter assumption is important because confidence intervals, like *P*-values, do nothing to address biases that may be present.

P-Value Functions

Although a confidence interval can be much more informative than a single *P*-value, it is subject to the misinterpretation that values inside the interval are equally compatible with the data, and all values outside it are equally incompatible. Like the alpha level of a test, however, the specific level of confidence used in constructing a

confidence interval is arbitrary; values of 95% or, less often, 99% or 90% are those most frequently used.

A given confidence interval is only one of an infinite number of ranges nested within one another. Points nearer the center of these ranges are more compatible with the data than points farther away from the center. To see the entire set of possible confidence intervals, one can construct a *P-value function*.[38-40] This function, also known as a consonance function[41] or confidence-interval function,[42] reflects the connection between the definition of a two-sided *P*-value and the definition of a two-sided confidence interval (*i.e.*, a two-sided confidence interval comprises all points for which the two-sided *P*-value exceeds the alpha level of the interval).

The *P*-value function gives the two-sided *P*-value for the null hypothesis, as well as every alternative to the null hypothesis for the parameter. A *P*-value function from the data in Table 15-1 is shown in Figure 15-3. Figure 15-3 also provides confidence levels on the right and so indicates all possible confidence limits for the estimate. The point at which the curve reaches its peak corresponds to the point estimate for the rate ratio, 3.1. The 95% confidence interval can be read directly from the graph as the function values where the right-hand ordinate is 0.95, and the 90% confidence interval can be read from the graph as the values where the right-hand ordinate is 0.90. The *P*-value for any value of the parameter can be found from the left-hand ordinate corresponding to that value. For example, the null two-sided *P*-value can be found from the left-hand ordinate corresponding to the height where the vertical line drawn at the hypothesized rate ratio = 1 intersects the *P*-value function.

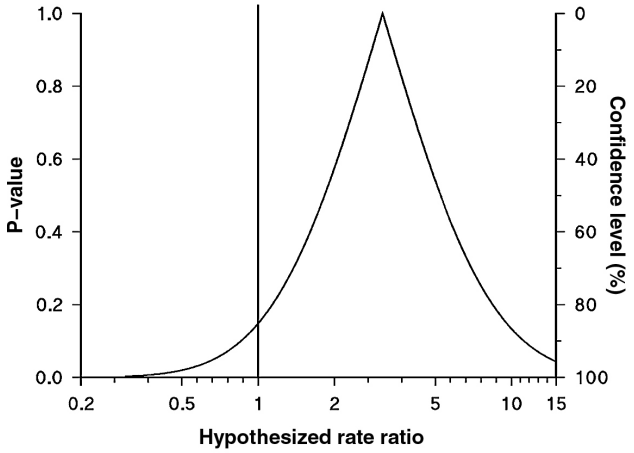


Figure 15.3 *P*-value function, from which one can find all confidence limits, for a hypothetical study with a rate ratio estimate of 3.1 (see Table 15-1).

A *P*-value function offers a visual display that neatly summarizes two key components of the estimation process. The peak of the curve indicates the point estimate, and the concentration of the curve around the point estimate indicates the precision of the estimate. A narrow *P*-value function would result from a study with high precision, which derives from a combination of the study size and its efficiency (see Study Efficiency, below). Conversely, a broad *P*-value function corresponds to a study that had low precision.

A confidence interval represents only one possible horizontal slice through the *P*-value function, but the single slice is enough to convey the two essential messages: Confidence limits usually provide enough information to locate the point estimate and to indicate the precision of the estimate. In large-sample epidemiologic statistics, the point estimate will usually be either the arithmetic or the geometric mean of the lower and upper limits. The distance between the lower and upper limits indicates the spread of the full *P*-value function.

The message of Figure 15-3 is that the example data are more compatible with a moderate to strong association than with no association, assuming the statistical model used to construct the function is correct. The confidence limits, when taken as indicative of the *P*-value function, summarize the size and precision of the

estimate.[43, 44] A single *P*-value, on the other hand, gives no indication of either the size or the precision of the estimate, and, if it is used merely as a hypothesis test, might result in a Type II error if there indeed is an association between exposure and disease.

Evidence of Absence of Effect or Incompatible Results

Confidence limits and *P*-value functions convey information about size and precision of the estimate simultaneously, keeping these two features of measurement in the foreground. The use of a single *P*-value—or (worse) dichotomization of the *P*-value into significant or not significant ones—obscures these features so that the focus on measurement is lost. A study cannot be reassuring about the safety of an exposure or treatment if only a statistical test of the null hypothesis is reported. As we have already seen, results that are not significant may be compatible with substantial effects. Lack of significance alone provides no evidence against such effects.[45]

Standard statistical advice states that when the data indicate a lack of significance, it is important to consider the power of the study to detect as significant a specific alternative hypothesis. The power of a test, however, is only an indirect indicator of precision, and it requires an assumption about the magnitude of the effect. In planning a study, it is reasonable to make conjectures about the magnitude of an effect to compute study-size requirements or power (see below). In analyzing data, however, it is always preferable to use the information in the data about the effect to estimate it directly, rather than to speculate about it with study-size or power calculations.[46-50] Confidence limits and (even more so) *P*-value functions convey much more of the essential information by indicating the range of values that are reasonably compatible with the observations (albeit at a somewhat arbitrary alpha level), assuming the statistical model is correct. They can also show that the data do not contain the information necessary for reassurance about an absence of effect.

Freiman et al.[30] used confidence limits for the RDs to reinterpret the findings from 71 negative clinical trials. These confidence limits indicated that many of the treatments under study were probably beneficial, as seen in Figure 15-1. The inappropriate interpretations of the authors in most of these trials could have been avoided by focusing their attention on the confidence limits rather than on the results of a statistical test.

For a study to provide evidence of lack of an effect, the confidence limits must be near the null value and the statistical model must be correct. In equivalence-testing terms, the entire confidence interval must lie within the zone around the null that would be considered practically equivalent to the null. Consider Figure 15-4, which depicts the *P*-value function from Figure 15-3 on an expanded scale, along with another *P*-value function from a study with a point estimate of 1.05 and 95% confidence limits of 1.01 and 1.10.

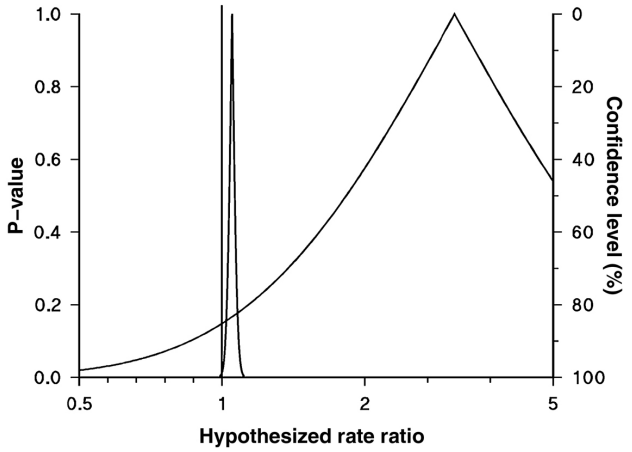


Figure 15.4 A *P*-value function from a precise study with a relative risk estimate of 1.05 and the *P*-value function from Figure 15-3.

The study yielding the narrow *P*-value function must have been large and information dense to generate such precision. The precision enables one to infer that, provided any strong biases or other serious problems with the statistical model are absent, the study provides evidence against a strong effect. The upper confidence limit (with any reasonable level of confidence) is near the

null value, indicating that the data are not readily compatible with large or even moderate effects. Or, as seen from the *P*-value function, the curve is a narrow spike close to the null point. The spike is not centered exactly on the null point, however, but slightly above it. In fact, the data from this large study would be judged as statistically significant by conventional criteria, because the (two-sided) *P*-value testing the null hypothesis is about 0.03. In contrast, the other *P*-value function in Figure 15-4 depicts data that, as we have seen, are readily compatible with large effects but are not statistically significant by conventional criteria.

Figure 15-4 illustrates the dangers of using statistical significance as the primary basis for inference. Even if one assumes no bias is present (*i.e.*, that the studies and analyses are perfectly valid), the two sets of results differ in that one result indicates there may be a large effect, while the other offers evidence against a large effect. The irony is that it is the statistically significant finding that offers evidence *against* a large effect, while it is the finding that is not statistically significant that raises concern about a possibly large effect. In these examples, statistical significance gives a message that is opposite of the appropriate interpretation. Focusing on interval estimation and proper interpretation of the confidence limits avoids this problem.

Numerous real-world examples demonstrate the problem of relying on statistical significance for inference. One such example occurred in the interpretation of a large randomized trial of androgen blockade combined with the drug flutamide in the treatment of advanced prostate cancer.[51] This trial had been preceded by 10 similar trials, which in aggregate had found a small survival advantage for patients given flutamide, with the pooled results for the 10 studies producing a summary odds ratio of 0.88, with a 95% confidence interval of 0.76, 1.02.[52, 53] In their study, Eisenberger et al. reported that flutamide was ineffective, thus contradicting the results of the 10 earlier studies, despite their finding an odds ratio of 0.87 (equivalent in their study to a mortality rate ratio of 0.91), a result not very different from that of the earlier

10 studies. The *P*-value for their finding was above their predetermined cutoff for "significance," which is the reason that the authors concluded that flutamide was an ineffective therapy. But the 95% confidence interval of 0.70 to 1.10 for their odds ratio showed that their data were readily compatible with a meaningful benefit for patients receiving flutamide. Furthermore, their results were similar to those from the summary of the 10 earlier studies. The *P*-value functions for the summary of the 10 earlier studies, and the study by Eisenberger et al., are shown in Figure 15-5. The figure shows how the findings of Eisenberger et al. reinforce rather than refute the earlier studies. They misinterpreted their findings because of their focus on statistical significance.

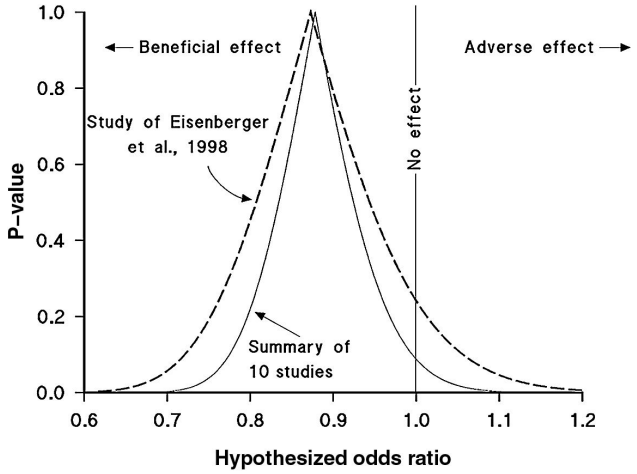


Figure 15.5 *P*-value functions based on 10 earlier trials of flutamide (solid line) and the trial by Eisenberger et al. (dashed line), showing the similarity of results and revealing the fallacy of relying on statistical significance to conclude, as did Eisenberger et al., that flutamide has no meaningful effect.

Another example was a headline-generating study reporting that women who consumed moderate amounts of alcohol retained better cognitive function than nondrinkers.[54] For moderate drinkers (up to 15 g of alcohol per day), the authors reported an RR for impaired cognition of 0.81 with 95% confidence limits of 0.70 and 0.93, indicating that moderate drinking was associated with a benefit with

respect to cognition. In contrast, the authors reported, “There were no significant associations between higher levels of drinking (15 to 30 g/d) and the risk of cognitive impairment or decline,” implying no benefit for heavy drinkers, an interpretation repeated in widespread news reports. Nevertheless, the finding for women who consumed larger amounts of alcohol was essentially identical to the finding for moderate drinkers, with a risk-ratio estimate of 0.82 instead of 0.81. It had a broader confidence interval, however, with limits of 0.59 and 1.13. Figure 15-6 demonstrates how precision, rather than different effect size, accounted for the difference in statistical significance for the two groups. The *P*-value function for moderate drinkers was narrower than the estimate for heavy drinkers. There is more information about moderate drinkers because the prevalence of moderate drinking is higher than the prevalence of heavy drinking. From the data, there is no basis to infer that the effect size differs for moderate and heavy drinkers; in fact, the hypothesis that is most compatible with the data is that the effect is about the same in both groups. Furthermore, the lower 95% confidence limit for the ratio of the RR in the heavy drinkers to the RR in the moderate drinkers is 0.71, implying that the data are also quite compatible with a much lower (more protective) RR in the heavy drinkers than in the moderate drinkers.

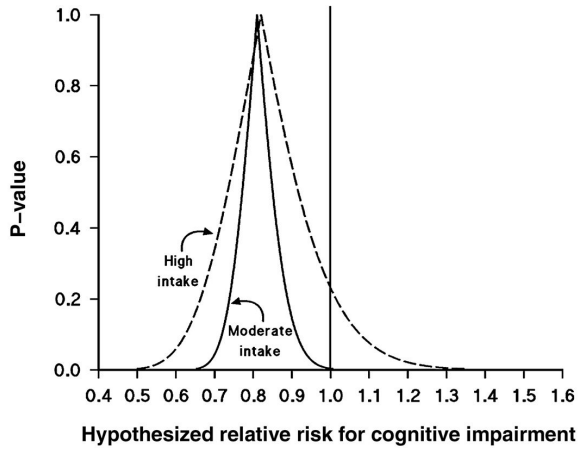


Figure 15.6 *P*-value functions for moderate and heavier drinkers of alcohol showing essentially identical negative associations with decline in cognitive function. The authors incorrectly reported that there was an association with moderate drinking, but not with heavier drinking, because only the finding for moderate drinking was statistically significant. (From Stampfer MJ, Kang JH, Chen J, et al. Effects of moderate alcohol consumption on cognitive function in women. *N Engl J Med*. 2005;352:245-253.)

These types of inferential errors have contributed to the perceived crisis[55, 56] in the reproducibility of scientific research.[11, 57] The literature on this perceived crisis has sometimes assessed whether the results of two studies are concordant based on whether their respective results are both statistically significant or not.[58] This approach has been used to assess reproducibility despite the fact that the idea that two results—one statistically significant and the other not—are necessarily different from one another is a well-known fallacy,[10, 59] as described above. Nonetheless, examples of claims of irreproducible results based on *P*-values falling on opposite sides of the commonly accepted Type I error rate are easy to find. In one example, subcutaneous heparin was reported to reduce the risk of deep vein thrombosis compared with intravenous heparin (OR = 0.62, 95% CI 0.39, 0.98); a reanalysis disagreed with the conclusion of a protective effect (OR = 0.61, 95% CI 0.30, 1.25).[60] In a second example, authors concluded that their results (OR = 0.75, 95% CI 0.48, 1.17) did not support previous sparse evidence of a

protective effect of statins use against glioma (previous study results were reported to be OR = 0.72; 95% CI 0.52, 1.00 and OR = 0.76; 95% CI 0.59, 0.98).[61] Finally, in a study of the association between antidepressant use during pregnancy and autism spectrum disorder in offspring, the authors reported a meta-analysis of earlier studies (OR = 1.7; 95% CI 1.1, 2.6), a multivariate adjusted hazards ratio in their study (1.59; 95% CI 1.17, 2.17), and an inverse probability of treatment weighted hazards ratio in their study (1.61; 95% CI 0.997, 2.59).[62] They concluded "antidepressant exposure compared with no exposure was not associated with autism spectrum disorder in the child." Examples of this type of misinterpretation abound and likely contribute to the perception that epidemiologic results are poorly reproducible when, at least in these examples, the evidence base is entirely consistent. It is impossible to estimate the degree to which this common misinterpretation has distorted the impression of a reproducibility crisis.[57]

Guidelines for Good Practice

Good data analysis does not demand that *P*-value functions be plotted routinely. They are especially useful when comparing two or more results. Overlap between the *P*-value functions will often forewarn against interpreting two results as different when they are not so different at all. For most results, though, it is sufficient to use conventional confidence limits to generate the proper mental visualization for the underlying *P*-value function. In fact, for large studies, only one pair of limits and their confidence level is needed to sketch the entire function, and one can easily learn to visualize the function that corresponds to any pair of limits. If, however, one uses the limits only to determine whether the null point lies inside or outside the confidence interval, one is only performing a significance test, and all of the biases and inferential errors described above will ensue. It is lamentable to go to the trouble to calculate confidence limits and then use them for nothing more than classifying the study finding as statistically significant or not. One should instead

remember that the precise locations of confidence limits are not important for proper interpretation. Rather, the limits should serve to give one a mental picture of the location and spread of the entire *P*-value function.

The main thrust of the preceding sections has been to argue the inadequacy of statistical significance testing for inference about effects. The view that estimation is preferable to testing has been argued by many scientists in a variety of disciplines, including, for example, economics, social sciences, environmental science, and accident research. There has been a particularly heated and welcome debate in psychology. In the overall scientific literature, hundreds of publications have addressed the concerns about statistical hypothesis testing. Some selected references include Rozeboom,[63] Morrison and Henkel,[64] Wulff,[65] Cox and Hinkley,[19] Rothman,[66] Salsburg,[67] Simon and Wittes,[68] Langman,[69] Gardner and Altman,[70] Walker,[71] Oakes,[72] Ware et al.,[73] Pocock et al.,[74] Poole,[40, 43] Thompson,[75] Evans et al.,[76] Anscombe,[77] Cohen,[78] Hauer,[79] Gigerenzer,[3] Ziliak and McCloskey,[80] Batterham and Hopkins,[81] Marshall,[82] the American Statistical Association[9] statement, Lash,[11] and the commentaries in Supplement 1 to Volume 73 (2019) of *The American Statistician*. To quote Atkins and Jarrett[32]:

> Methods of estimation share many of the problems of significance tests—being likewise based on probability model assumptions and requiring "arbitrary" limits of precision. But at least they do not require irrelevant null hypotheses to be set up nor do they force a decision about "significance" to be made—the estimates can be presented and evaluated by statistical *and other* criteria, by the researcher or the reader. In addition, the estimates of one investigation can be compared with others. While it is often the case that different measurements or methods of investigation or theoretical approaches lead to "different" results, this is not a disadvantage; these differences reflect

> important theoretical differences about the meaning of the research and the conclusions to be drawn from it. And it is precisely those differences which are obscured by simply reporting the significance level of the results.

Indeed, because statistical hypothesis testing promotes so much misinterpretation, we recommend avoiding its use in epidemiologic presentations and research reports. Such avoidance requires that *P*-values (when used) be presented without reference to alpha levels or "statistical significance" and that careful attention be paid to the confidence interval, especially its width and its endpoints (the confidence limits).[13, 44]

Problems With Confidence Intervals

Because they can be derived from *P*-values, confidence intervals and *P*-value functions are themselves subject to some of the same criticisms as significance tests.[1, 15, 16] One problem that confidence intervals and *P*-value functions share with statistical hypothesis tests is their very indirect interpretations, which depend on the concept of "repetition of the study in a manner identical in all respects except for random error." Interpretations of statistics that appeal to such a concept are called repeated-sampling or *frequentist* interpretations, because they refer to the frequency of certain events (rejection by a test or coverage by a confidence interval) in a series of repeated experiments.

An astute investigator may properly ask what frequency interpretations have to do with the single study under analysis. It is all very well to say that an interval estimation procedure will, in 95% of repetitions, produce limits that contain the true parameter. But in analyzing a given study, the relevant scientific question is this: Does the single pair of limits produced from this one study contain the true parameter? The ordinary (frequentist) theory of confidence intervals does not answer this question. The question is so important that many (perhaps most) users of confidence intervals mistakenly

interpret the confidence level of the interval as the probability that the answer to the question is "yes."[83] It is quite tempting to say that the 95% confidence limits computed from a study contain the true parameter with 95% probability. Unfortunately, this interpretation can be correct only for Bayesian interval estimates (discussed later and in Chapter 23), which often diverge from ordinary confidence intervals.

There are several alternative types of interval estimation that attempt to address these problems. We will discuss two of these alternatives in the next two subsections.

Likelihood Intervals

To avoid interpretational problems, a few authors prefer to replace confidence intervals with likelihood intervals, also known as support intervals.[15, 17, 25] In ordinary English, "likelihood" is just a synonym for "probability." In the likelihood theory, however, a more specialized definition is used: The *likelihood* of a specified parameter value given observed data is defined as the probability of the observed data, given that the true parameter equals the specified parameter value. This concept is covered in depth in many statistics textbooks; for example, see Berger and Wolpert,[84] Clayton and Hills,[85] Edwards,[25] and Royall.[17] Here, we will describe the basic definitions of the likelihood theory.

To illustrate the definition of likelihood, consider again the population in Table 15-1, in which 186/(186 + 128) = 59% of person-years were exposed. Under standard assumptions, it can be shown that, if there is no bias and the true rate ratio is 10, there will be a 0.125 chance of observing nine exposed cases, given 11 total cases and 59% exposed person-years. (The calculation of this probability is beyond the present discussion.) Thus, by definition, 0.125 is the likelihood for a rate ratio of 10, given the data in Table 15-1. Similarly, if there are no biases and the true ratio is 1, there will be a 0.082 chance of observing 9 exposed cases given 11 total and 59%

exposed person-years; thus, by definition, 0.082 is the likelihood for a rate ratio of 1, given the data in Table 15-1.

When one parameter value makes the observed data more probable than another value and hence has a higher likelihood, it is sometimes said that this parameter value has higher support from the data than the other value.[17, 25] For example, in this special sense, a rate ratio of 10 has higher support from the data in Table 15-1 than a rate ratio of 1, because those data have a greater chance of occurring if the rate ratio is 10 than if it is 1.

For most data, there will be at least one possible parameter value that makes the chance of getting those data highest under the assumed statistical model. In other words, there will be a parameter value whose likelihood is at least as high as that of any other parameter value and so has the maximum possible likelihood (or maximum support) under the assumed model. Such a parameter value is called a *maximum-likelihood estimate* (MLE) under the assumed model. For the data in Table 15-1, there is just one such value, and it is the observed rate ratio (9/186)/(2/128) = 3.1. If there are no biases and the true rate ratio is 3.1, there will be a 0.299 chance of observing 9 exposed cases, given 11 total and 59% exposed person-years, so 0.299 is the likelihood for a rate ratio of 3.1, given the data in Table 15-1. No other value for the rate ratio will make the chance of these results higher than 0.299, and so 3.1 is the MLE. Thus, in the special likelihood sense, a rate ratio of 3.1 has the highest possible support from the data.

As has been noted, Table 15-1 yields a likelihood of 0.125 for a rate ratio of 10; this value (0.125) is 42% of the likelihood (of 0.299) for 3.1. Similarly, Table 15-1 yields a likelihood of 0.082 for a rate ratio of 1; this value (0.082) is 27% of the likelihood for 3.1. Overall, a rate ratio of 3.1 maximizes the chance of observing the data in Table 15-1. Although rate ratios of 10 and 1 have less support (lower likelihood) than 3.1, they are still among values that likelihoodists regard as having enough support to warrant further consideration; these values typically include all values with a likelihood above one-seventh of the maximum.[15, 17, 25] Under a normal model for random

errors, such one-seventh likelihood intervals are approximately equal to 95% confidence intervals.[17]

The maximum of the likelihood is the height of the likelihood function at the MLE. A likelihood interval for a parameter (here, the rate ratio) is the collection of all possible values whose likelihood is no less than some specified fraction of this maximum. Thus, for Table 15-1, the collection of all rate ratio values with a likelihood no less than 0.299/7 = 0.043 (one-seventh of the highest likelihood) is a likelihood interval based on those data. Upon computing this interval, we find that all rate ratios between 0.79 and 20 imply a probability for the observed data at least one-seventh of the probability of the data when the rate ratio is 3.1 (the MLE). Because the likelihoods for rate ratios of 1 and 10 exceed 0.299/7 = 0.043, 1 and 10 are within this interval.

Analogous to confidence limits, one can graph the collection of likelihood limits for all fractions of the maximum (1/2, 1/4, 1/7, 1/20, etc.). The resulting graph has the same shape as one would obtain from simply graphing the likelihood for each possible parameter value. The latter graph is called the *likelihood function* for the data. Figure 15-7 gives the likelihood function for the data in Table 15-1, with the ordinate scaled to make the maximum (peak) at 3.1 equal to 1 rather than 0.299 (this is done by dividing all the likelihoods by the maximum, 0.299). Thus, Figure 15-7 provides all possible likelihood limits within the range of the figure.

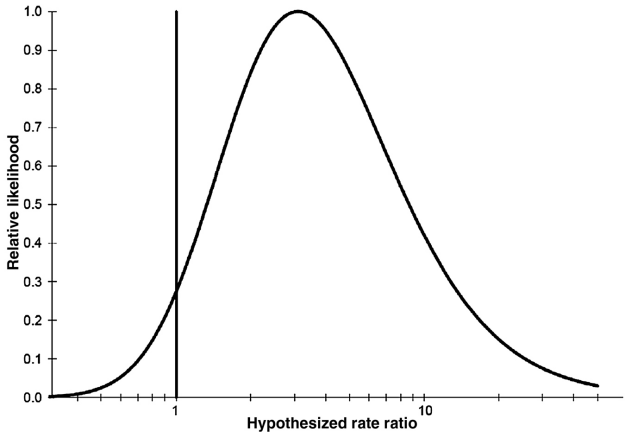


Figure 15.7 Relative likelihood function based on Table 15-1.

The function in Figure 15-7 is proportional to

$$\left(\frac{186(\mathrm{IR})}{186(\mathrm{IR})+128}\right)^{9}\left(\frac{128}{186(\mathrm{IR})+128}\right)^{2}$$

where IR is the hypothesized incidence rate ratio (the abscissa). Note that this function is broader and less sharply peaked than the *P*-value function in Figure 15-3, reflecting the fact that, by likelihood standards, *P*-values and confidence intervals tend to give the impression that the data provide more evidence against the test hypothesis than they actually do.[15]

Some authors prefer to use the natural logarithm of the likelihood function, or *log-likelihood function,* to compare the support given to competing hypotheses by the data.[15, 17, 25] These authors sometimes refer to the log-likelihood function as the support function generated by the data. Although we find log-likelihoods less easily interpretable than likelihoods, log-likelihoods can be useful in constructing confidence intervals.

Bayesian Intervals

As with confidence limits, the interpretation of likelihood limits is indirect, in that it does not answer the question: "Is the true value

between these limits?" Unless the true value is already known (in which case there is no point in gathering data), it can be argued that the only rational answer to the question must be a *subjective* probability statement, such as "I am 95% sure that the true value is between these limits."[86, 87] Such subjective probability assessments, or *certainties,* are common in everyday life, as when a weather forecaster predicts 80% chance of rain tomorrow, or when one is delayed while traveling and thinks that there is a 90% chance of arriving between 1 and 2 hours after the scheduled arrival time. If one is *sure* that the true arrival time will be between these limits, this sureness represents a subjective assessment of 100% probability (complete certainty) that arrival will be 1 to 2 hours late. In reality, however, there is always a chance (however small) that one will be delayed longer or may never arrive, so complete certainty is never warranted.

Subjective Bayesian analysis is concerned with producing realistic and rationally coherent probability assessments, and it is especially concerned with updating these assessments as data become available. *Rationally coherent* means only that assessments are free of logical contradictions and do not contradict the axioms of the probability theory (which are also used as axioms for frequentist probability calculations).[86-89]

All statistical methods require a model for data probabilities. The Bayesian analysis additionally requires a *prior probability distribution.* In theory, this means that one must have a probability assessment available for every relevant interval; for example, when trying to study a rate ratio, before seeing the data one must be able to specify one's certainty that the rate ratio is between 1 and 2, and between ½ and 4, and so on. This prior-specification requirement demands that one has a probability distribution for the rate ratio that is similar in shape to Figure 15-3 *before* seeing the data. This is a daunting demand, and it was enough to have impeded the use and acceptance of Bayesian methods for most of the 20th century.

Suppose, however, that one succeeds in specifying in advance a prior probability distribution that gives prespecified certainties for

the target parameter. Bayesian analysis then proceeds by combining this prior distribution with the likelihood function (such as in Figure 15-7) to produce a new, updated set of certainties, called the *posterior probability distribution* for the target parameter based on the given prior distribution and likelihood function. This posterior distribution in turn yields posterior probability intervals (posterior certainty intervals). Suppose, for example, one accepts the prior distribution as a good summary of previous information about the parameter and similarly accepts the likelihood function as a good summary of the data probabilities, given various possible values for the parameter. The resulting 95% posterior interval is then a range of numbers that one can be 95% certain contains the true parameter.

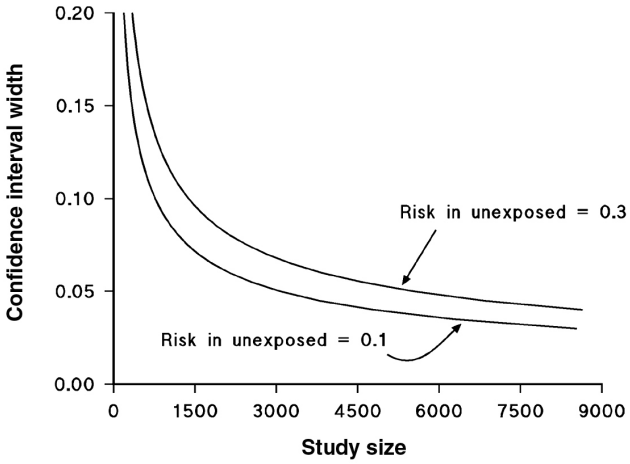


Figure 15.8 Study size in relation to 95% confidence interval width for two cohorts of equal size with a risk difference of 0.1.

The technical details of computing exact posterior distributions can be quite involved and were also an obstacle to the widespread adoption of the Bayesian methods. Modern computing advances have all but eliminated this obstacle as a serious problem; also, the same approximations used to compute conventional frequentist statistics can be used to compute approximate Bayesian statistics.

Another obstacle to Bayesian methods has been that the intervals produced by a Bayesian analysis refer to subjective probabilities rather than objective frequencies. Some argue that, because

subjective probabilities are just one person's opinion, they should be of no interest to objective scientists. Unfortunately, in nonexperimental studies, there is (by definition) no identified random mechanism to generate objective frequencies over study repetitions; thus, in such studies, the so-called objective frequentist methods (such as significance tests and confidence intervals) lack the objective repeated-sampling properties usually attributed to them.[1, 2, 16, 26, 27, 88, 90] Furthermore, scientists do routinely offer their opinions and are interested in the opinions of colleagues. Therefore, it can be argued that a rational (if subjective) certainty assessment may be the only reasonable inference we can get out of a statistical analysis of observational epidemiologic data. Some argue that this conclusion applies even to perfect randomized experiments.[14, 87, 91]

At the least, Bayesian statistics provide a probabilistic answer to questions such as "Does the true rate ratio lie between 1 and 4?" (to which one possible Bayesian answer is "In light of the data and my current prior information, I can be 90% certain that it does"). A more general argument for the use of Bayesian methods is that they can provide point and interval estimates that have better objective frequency (repeated-sampling) properties than ordinary frequentist estimates. These calibrated Bayesian statistics include Bayesian confidence intervals that are narrower (more precise) than ordinary confidence intervals with the same confidence level. Because the advantages of procedures with Bayesian justification can be so dramatic, some authors argue that only methods with a clear Bayesian justification should be used, even though repeated-sampling (objective frequency) properties are also desirable (such as proper coverage frequency for interval estimates).[92-94]

In addition to providing improved analysis methods, the Bayesian theory can be used to evaluate established or newly proposed statistical methods. For example, if a new confidence interval is proposed, we may ask: "What prior distribution do we need to get this new interval as our Bayesian posterior probability interval?" It is often the case that the prior distribution one would need to justify a conventional confidence interval is patently absurd; for example, it

would assign equal probabilities to rate ratios of 1 and 1,000,000.[16, 88, 95] In such cases, it can be argued that one should reject the proposed interval because it will not properly reflect any rational opinion about the parameter after a careful data analysis.[16, 93]

Under certain conditions, ordinary (frequentist) confidence intervals and one-sided *P*-values can be interpreted as approximate posterior (Bayesian) probability intervals.[19, 24] These conditions typically arise when little is known about the associations under study. Frequentist intervals cease to have Bayesian utility when much is already known or the data under analysis are too limited to yield even modestly precise estimates. The latter situation arises not only in small studies, but also in large studies that must deal with many variables at once, or that fail to measure key variables with sufficient accuracy.

Summary

Statistics can be viewed as having many roles in epidemiology. Data description is one role, and statistical inference is another. The two are sometimes mixed, to the detriment of both activities, and are best distinguished from the outset of an analysis.

Different schools of statistics view statistical inference as having different roles in data analysis. The hypothesis-testing approach treats statistics as chiefly a collection of methods for making decisions, such as whether an association is present in a source population or "superpopulation" from which the data are randomly drawn. This approach has been declining in the face of criticisms that estimation, not decision-making, is the proper role for statistical inference in science. Within the latter view, frequentist approaches derive estimates by using probabilities of data (either *P*-values or likelihoods) as measures of compatibility between data and hypotheses, or as measures of the relative support that data provide hypotheses. In contrast, the Bayesian approach uses data to improve existing (prior) estimates in light of new data. Different approaches can be used in the course of an analysis. Nonetheless, proper use of

any approach requires more careful interpretation of statistics than has been common.

PRECISION AND STRATIFICATION

In many epidemiologic analyses, the crude data are divided into strata to examine effects in subcategories of another variable or to control confounding. The efficiency of a study can be affected dramatically by stratifying the data. A study that has an overall apportionment ratio that is favorable for precision (which will be a ratio of 1.0 if there is no effect and no confounding) may nevertheless have apportionment ratios within strata that vary severely from low to high values. It is common to see some strata with the extreme apportionment ratios of 0 and infinity (*e.g.*, no cases in some strata and no controls in others). The smaller the numbers within strata, the more extreme the variation in the apportionment ratio across strata is likely to be. The extreme values result from zero subjects or person-time units for one group in a stratum. Small numbers within strata result from having too few subjects relative to the number of strata created. This sparse-data problem can develop even with large studies, because the number of strata required in the analysis increases geometrically with the number of variables used for stratification. Indeed, sparse data are a major limitation of stratified analysis, although the same problem negatively affects regression modeling as well.

When comparisons within strata will be essential and much variation in the apportionment ratio is expected across strata, then matching on the stratification variables (Chapter 6) is one way to improve the efficiency of the apportionment ratio within strata and to reduce sparsity problems without increasing the study size. When matching on all stratification variables is not feasible, increasing the overall number of subjects will at least reduce data sparsity and improve precision, even if only one group (*e.g.*, the controls in a case-control study) can be expanded.

PLANNING STUDY SIZE

Enlarging the size of a study is one of the key ways to reduce random error in an epidemiologic estimate. Practical constraints on resources inevitably limit study size, so one must plan accordingly. One method that is used to plan the size of a study is to calculate the study size based on conventional statistical "sample-size" formulas.[96-100] These formulas relate the size of a study to the study design, study population, and the desired power or precision.

Study-size formulas, being purely mathematical, do not account for anything that is not included as a variable in the formula. At best, they serve only to provide rough guidelines, and in some situations, they may be misleading from a broader perspective. For example, conventional formulas do not weigh the value of the information obtained from a study against its use of resources. Yet a focal problem in planning the study size is determining how to balance the value of greater precision in study results against the greater cost. Solving the problem thus involves a cost-benefit analysis of expending greater effort or funds to gain greater precision. Greater precision has a value to the beneficiaries of the research, but the value is indeterminate because it is always uncertain how many beneficiaries there will be. Furthermore, the potential benefits of the study involve intricacies of many social, political, and biologic factors that are almost never quantified. Consequently, only informal guesses as to a cost-efficient size for an epidemiologic study are feasible. Although study-size determination can be aided by conventional formulas, the final choice must also incorporate unquantified practical constraints and implications of various study sizes.

In this section, we will discuss the considerations that the researcher should weigh in planning the size of a study. The term "sample size" is often used to describe the study size, possibly borrowed from the vocabulary of survey sampling design. We prefer to use "study size" to avoid confusing a study of causal effects with a survey to describe a large population based on a sample from it.

A study size is the primary determinant of the precision of the estimates that come from the study, but better precision—the reduction of random error—is affected not only by the study size but by the study efficiency as well, as described above. Furthermore, the measurement of study efficiency or study informativeness depends on assumptions that are often absent or implicit in statistical discussions, such as what constitutes the most relevant hypotheses and costs.

Study Efficiency

Study efficiency can be thought of as the amount of information that a study produces in relation to its size or cost. Efficiency usually depends on issues such as the ratio of the number of subjects or person-time units across categories of exposure or disease. A study with 1,000,000 people may seem large, but if only 100 of them are exposed and 999,900 are unexposed, it will have considerably less information than a study with an even balance of exposed and unexposed subjects. When the study factor has no effect and no adjustment for confounding is needed, equal apportionment into exposure groups is the most efficient cohort design.[101] In case-control studies, the study efficiency will depend on the balance between case and control groups. As described in Chapter 8, case-control designs are best conceptualized as efficient cohort designs that include all cases and a subset of the cohort giving rise to the cases. This conceptualization rests on the idea of optimizing apportionment of study subjects: cases are usually few in relation to the size of the population giving rise to the cases, so over-representing cases by design is efficient. On the other hand, in cohort studies, it may be costly to engineer a desirable balance of exposure and may lead to inefficiencies in trying to use the data to study a wider range of exposures. In short, if the apportionment of study subjects by categories of exposure or outcome can be manipulated in the study design, it may be advantageous to design the study with a good balance between groups.

The considerations involved in designing an efficient study usually pit the costs of doing so against the benefits, which may be difficult to assess. Nonetheless, it is worth knowing that when the study exposure has no effect and no adjustment is needed, equal apportionment into exposure groups is the most efficient cohort design.[101] For example, when no association is expected for any reason (which is to say, in the absence of any source of association, whether bias or exposure effect), a cohort study of 2,400 persons will be most efficient statistically if it comprises 1,200 exposed and 1,200 unexposed persons for study (a 1:1 exposed-unexposed allocation ratio). Similarly, in a case-control study, when no association is expected, it will be most efficient to have an equal number of cases and controls (a 1:1 case-control allocation ratio). Nonetheless, since we do not know whether an association is present, translating these statistical facts into a directive for study design assumes that the primary goal of the analysis is to evaluate the hypothesis of no association (or, with additional assumptions, no effect).

In the presence of an effect, the statistically most efficient allocation ratio will differ from equal apportionment (1:1) by an amount that depends on the magnitude of the effect.[101] In the cohort example, if priority were instead given to testing the hypothesis that there is a doubling of risk, parallel derivations lead to allocating more of the total to the unexposed, for whom risk is lower than the exposed. For a disease that would occur in no more than a few percent of the cohort, this could lead to a 1:2 allocation ratio (800 exposed and 1,600 unexposed) as more efficient than 1:1 allocation. If equal weight is given to the hypotheses of no association and doubling of risk, the efficient allocation would be nearer 1,000 exposed and 1,400 unexposed than either 1:1 or 1:2 allocation.

Efficiency of a study may be difficult to modify. For example, if exposure is rare, population samples will show a preponderance of unexposed subjects unless special populations are sought. Both cohort studies and case-control studies drawn from a general population tend to be inefficient if exposure has low prevalence. Two-stage (two-phase) sampling designs that account for both

exposure and disease frequency are available, although the data they produce requires special analysis methods.

Adjustments for biases such as confounding will also influence study precision, usually diminishing it. In many epidemiologic analyses, the crude data are divided into strata to examine exposure effects within subcategories of another variable, or to control confounding. The efficiency of a study can be affected dramatically by stratifying the data. A study that has an overall apportionment ratio that is favorable for precision, with close to overall balance between the main groups to be compared, may nevertheless have apportionment ratios within strata that vary severely from low to high values. It is not uncommon to see some strata with the extreme apportionment ratios of 0 and infinity (*e.g.*, a case-control study with no cases in some strata and no controls in others). The smaller the numbers within strata, the more extreme the variation in the apportionment ratio across strata is likely to be.

The most extreme values occur when there are zero subjects or person-time units for one group in a stratum. The chance of zero subjects or person-time units for a group in a stratum is increased when there are many strata in relation to the number of study subjects. This sparse-data problem can develop even with studies based on large numbers of subjects. It not only affects study efficiency but can introduce a bias in the estimation of ratio measures. This sparse data bias can be substantial, and is more common than realized, because the number of strata required increases geometrically with the number of variables used for stratification.[102] Indeed, sparse data are a major limitation of stratified analysis, although the problem is not limited to stratified analysis, and can affect regression modeling as well, where it may go unnoticed.[103, 104]

When comparisons within strata will be essential and substantial variation in the apportionment ratio is expected across strata, then matching on the stratification variables (see Chapter 6) is one way to improve the efficiency of the apportionment ratio within strata and to reduce sparse-data problems without increasing the overall study

size. It is, however, not guaranteed to improve efficiency and may even harm it, especially in case-control studies and especially when matching on variables that are unrelated to the outcome variable. When matching on all stratification variables is not feasible or advisable, increasing the overall number of subjects may mitigate sparse-data problems and improve overall study precision, even if only one group (*e.g.*, the control series) can be expanded.

Study Size

Study size is only adjustable within the constraints of available data and budget. Even when it is not adjustable (*e.g.*, one has only a fixed data base for use), however, study size is usually the largest determinant of study precision. Considerations of study size in the planning stage of a research project are thus essential for investigators and reviewers to assess the potential informativeness of a study.

In planning a study, assessing the informativeness may be addressed in various ways. The typical approaches involve judging the informativeness of a potential study by postulating one or more possible study sizes and calculating a measure of informativeness for each. Often this amounts to calculating the statistical power of the study for a statistical test of a targeted hypothesis (usually but not necessarily a "null" hypothesis) for a range of possible study sizes.

To review briefly these terms and concepts, and their relationships to the study size, we note that conventional statistical hypothesis tests can usually be described as producing a *P*-value that will be compared to a critical cutoff α. As described above, and subject to all of the limitations already mentioned, the event of $P \leq \alpha$ is usually taken as "rejection" or "statistical significance" at level α, although actual rejection or significance should depend on many other considerations besides the result of a statistical test. The false-positive rate of the test is the probability that $P \leq \alpha$ if the tested hypothesis is correct and is also known as the Type I error rate, alpha error rate, or false-rejection rate. The *P*-value from the test is said to be valid if the probability that $P \leq \alpha$ equals α (*i.e.*, if the false-

positive rate is α). The power of the test for detecting a specific alternative to the tested hypotheses is the probability that $P \leq \alpha$ when the alternative is correct; the value of 1 − power is called the false-negative rate, Type II error rate, or beta-error rate, is often denoted by β.

The Imbalance of Power in Traditional Study-Size Computations

There are many formulas that relate study size to power, taking into account design features such as the apportionment ratios of the exposure or outcome groups, whether subjects are clustered by some variable, and what covariates will be controlled analytically. The following problems apply to them all.

The test cutoff α is usually chosen to be 5%, although that choice is rarely explicitly justified. The theory says that α should be chosen to reflect the actual cost of false-positive errors,[105, 106] with α being smaller if false positives are costly compared with false negatives and larger in the reverse situation. For a valid test, the chance α of false-positive error and the chance β of false-negative error are inversely related to each other, so in choosing α and β, there is an inevitable trade-off between the risks of the two errors. This alignment of acceptable error rates with values is discussed above but seldom implemented in practice.

Traditional study-size requirements assume that a power of only 80% is acceptable when conducting a test with α = 5%, although (as with α = 5%) that choice is rarely explicitly justified. Although 80% may at first sound high, it means that the false-negative rate β is 20%, which would be wholly unjustified if false negatives were very costly and false positives were not. Adopting such a gross imbalance, allowing the false-negative rate to be four times the false-positive rate, may reflect nothing more than the fact that it results in a much smaller study-size requirement than if both the acceptable false-positive and false-negative rates were set to be 5%.

Complicating matters is that the costs may be radically different for different stakeholders in a problem; for example, in litigation claiming harms from an industrial chemical, false positives are typically very costly for the defendant (*e.g.*, the chemical manufacturer) but beneficial for the plaintiff demanding compensation (*e.g.*, the exposed); conversely false negatives are beneficial for the defendant but costly for the plaintiff. This difference in acceptable error rates means that the tradition of accepting a higher false-negative rate than a false-positive rate when designing studies to test a null hypothesis is a tradition favoring the defendant and more generally favors the null hypothesis over the alternative.

The traditional imbalance in favor of the null in the study design and testing can lead to apparently paradoxical results in which a study with "high power" (*e.g.*, 90%) may fail to reject the null at the 5% level, yet exhibit data that statistically favor the alternative according to other conventional criteria.[107] To avoid such imbalance and its consequences, one may instead seek equal rates of both errors, *e.g.*, by designing a study to have 95% power ($\beta = 5\%$) when using $\alpha = 5\%$. A justification for inequality in terms of error costs may instead be sought, but the justification will not apply to those whose costs differ from the costs assumed by the justification.

Other Drawbacks of Power Calculations

A glaring drawback of power calculations is that they are based on dichotomous statistical testing (technically, Neyman-Pearson hypothesis testing for statistical decisions), and as such they promulgate the "dichotomania" that is characteristic of significance testing, classifying the results of a quantitative exercise into two ultimate categories, significant or not significant. This type of thinking allows divergent conclusions to be drawn from possible study results that might differ little, but with the two results falling on different sides of the demarcation for significance. This critical limitation has been described in detail above.

Relying on power calculations can also lead to overestimation of a study's informativeness. For example, if a study is planned to have 90% power with $\alpha = 5\%$ and the effect is postulated to be a rate ratio of 3.0, the power calculations imply that the study will give statistically significant results with 90% probability, assuming that the statistical model used applies (a tall assumption) and that all other relevant factors, such as control of confounding, are sufficiently taken into account. Imagine that the actual effect is just what was assumed, a rate ratio of 3.0, and the study is conducted at the size that corresponds to 90% power. If the estimated rate ratio had been instead 1.9 or less (which would occur 9% of the time with a true rate ratio of 3), that result would have $P > 0.05$ by the null test and so be easily mistaken for supporting the null (a Type II error). But if the estimate were 1.9, the confidence interval around it would include 3 as well as 1, showing that the result is inconclusive according to the $\alpha = 5\%$ criterion. More generally, a study will be incapable of discriminating between the null and the alternative at a rate equal to the false-negative rate, not the false-positive rate, and thus a study "powerful" by the usual weak standards will often produce results that are ambiguous when interpreted correctly.

Another drawback of power calculations is that they are highly dependent on the alternative chosen for the calculation. Those wanting to claim high power need only use a large value for the alternative, at least up to the point before it becomes obvious they are "gaming" the calculation. Conversely, those wishing to condemn a study as "underpowered" need to only select a small alternative. In reality, the informativeness of a study grows progressively with increasing association or effect size and should be viewed on a continuum. One step toward this goal when planning studies is to plot power against alternatives and do so for different possible study sizes. A power curve graphs study power against effect size and shows the continuous relation between the two.

For all the above reasons and more, power calculations can be misleading when *analyzing study designs and study* data.[47-49, 107] These

problems lead to considering study precision directly for design as well as for analysis,[100, 108] as described next.

Factors Influencing Study Precision

Various factors affect precision of an effect estimate in a study; these are related to study design features and the analytic methods used. For that reason, there are numerous different formulas, mostly based on statistical power, but referred to as "sample size" formulas, that are used to calculate the study size. We do not cover the range of formulas that apply to the entire spectrum of epidemiologic research designs, but below we give a simple general formula to illustrate the inputs needed. Suppose we are planning a cohort study that is intended to measure and compare risk in two groups, so that the denominators are the number of persons in each exposure group. This situation will also include randomized trials that measure outcomes as risks. A simple formula for study size for this type of study was given by Kelsey et al.[109]:

$$N_1 = \frac{\left(Z_{\alpha/2} + Z_{\beta}\right)^2 p(1-p)(R+1)}{D^2 R}$$

where

N_1 = size of exposed cohort

$R = N_0/N_1$ = ratio of size of unexposed cohort, N_0, to size of exposed cohort, N_1

$Z^{\alpha/2}$ = standard normal deviate corresponding to the alpha level of the hypothesis test

$Z\beta$ = standard normal deviate corresponding to the desired study power

p_1 = proportion of exposed cohort hypothesized to develop disease

p_0 = proportion of unexposed cohort expected to develop disease

$p = (p_1 + Rp_0)/(R+1)$
$D = p_1 - p_0$

Notice what must be postulated to compute the study size:

1. the desired power of the study ($1 - \beta$),
2. an alpha level for a significance test,
3. the relative size of the unexposed and exposed cohorts, R,
4. the risk among unexposed, p_0, and
5. either p_1 or the hypothesized RD, D

If the investigator is interested in estimating an RR from a study instead of an RD, then Equation 15-1 can still be used, by solving $p_1 = \text{RR}p_0$. If the cohort study is aiming to measure rates rather than risks, this formula can still be used as a rough approximation by converting the amount of person-time to person denominators N_1 and N_0 by multiplying the person-time by the anticipated average amount of time followed.

Equation 15-1 can also be used to calculate the size of a case-control study. To do so, one would redefine N_1 and N_0 as the size of the case group and control group, respectively, and p_1 and p_0 will be the proportion of cases and controls, respectively, that are exposed. If the study is planned based on an anticipated odds ratio, $\text{OR} = p_1(1 - p_0)/[p_0(1 - p_1)]$, then one needs to specify p_0 and the OR and solve for p_1:

$$P_1 = \frac{\text{OR}p_0}{1 - p_0 + \text{OR}p_0}$$

When case-control studies involve individual matching retained in the analysis, any matched set that is completely concordant for

exposure (that is, if the case and all matched controls have the same exposure value) is effectively lost to the study, as that set contributes no information about the conditional odds ratio. Thus, study size calculations for matched case-control studies are affected by the degree of correlation between the matching factors and the exposure, as well as the ratio of controls to cases. Power and study size formulas for matched case-control data were presented by Miettinen.[110]

Estimating Study Size Based on the Confidence Interval Width

Instead of statistical power, one can plan a study by anticipating the study precision directly.[111] Specifically, one can postulate the desired width of the study confidence interval and examine how that varies with the study size. If we start with the formulas that are used to obtain confidence intervals and set them equal to a desired width, we can solve these equations for the study size. The particular formula to use will depend on which of the three types of data will be involved in the study: (1) risk data with person denominators; (2) rate data with person-time denominators; or (3) case-control data. For each type of data, there will be a formula for a difference measure of effect and another formula for a ratio measure of effect. Case-control studies are an exception, as there is no difference measure.

For difference measures, the asymptotic confidence interval is obtained from

$$\widehat{\mathrm{RD}} \pm Z \cdot \mathrm{SD}\left(\widehat{\mathrm{RD}}\right)$$

where $\widehat{\mathrm{RD}}$ is the point estimate of the risk or rate difference, Z is the value from a standard normal distribution corresponding to the

level of confidence (*e.g.*, 1.96 for 95% CI), and $SD(\widehat{RD})$ is the standard deviation (also referred to as the standard error) of the estimate.

The corresponding formula for the risk, rate, or odds ratio is

$$e^{\ln\left(\widehat{RR}\right)+Z\cdot SD\left[\ln\left(\widehat{RR}\right)\right]}$$

For a simple approach to estimating study size from these expressions, we can assume that crude data will be the basis for the analysis. This assumption may overstate the precision of the result, as control for confounding often costs some precision. We will also assume that there are no missing data. The effects of these assumptions might need to be considered in the final study planning.

To obtain the study size that corresponds to a given confidence interval width, some design features and other factors must be specified. In a cohort study, these specifications include the following:

1. the risk or rate among unexposed (p_0 for risk, I_0 for rate),
2. the risk or rate among exposed (p_1 or I_1),
3. the relative size of the unexposed and exposed cohorts, R,
4. the desired level of precision, and
5. the confidence level.

For item 2 above, if the rate difference or rate ratio is specified, p_1 or I_1 can be calculated from the difference or ratio and from p_0 or I_0. For item 3, the relative size of the cohorts, R, will be expressed as the ratio of the size of the unexposed cohort to that of the exposed cohort. N_1 will be the size of the exposed cohort and N_0 the size of the unexposed cohort, with $R = N_0/N_1$. For risk data, N_1 will represent people and for rate data, N_1 will represent person-time. The desired level of precision can be expressed in various ways; here

we will express precision as the absolute width of the confidence interval for difference measures of effect, and as the ratio of the upper confidence limit to the lower confidence limit for ratio measures of effect. The level of confidence corresponds to Z in the confidence interval formulas above; Z is the value of the standard normal distribution such that the area under the curve from $-Z$ to $+Z$ equals the confidence level. Z is 1.96 for a 95% confidence interval.

For case-control studies, we use a slight variation in the above list. Rather than risk or rates, p_1 and p_0 refer to the respective exposure prevalences among cases and controls. Alternatively, we can specify p_0 and the odds ratio, OR, from which p_1 can be calculated as

$$p_1 = \mathrm{OR}p_0 \Big/ \left(1 - p_0 + \mathrm{OR}p_0\right).$$

Standard error formulas used for risk and rate differences, risk and rate ratios, and odds ratio are given in Table 15-2. From these, the study size formulas can readily be derived (Table 15-3). As an example, consider a study based on risk data and focused on RD as the measure of interest. Denoting F as the absolute width of the RD that we desire to achieve for a study's confidence interval, we can solve for N_1 as:

$$N_1 = \frac{4Z^2\left[Rp_1\left(1-p_1\right) + p_0\left(1-p_0\right)\right]}{RF^2}$$

TABLE 15-2

Standard Deviation Formulas for Crude Measures of Epidemiologic Effect

Risk difference

a = exposed cases
b = unexposed cases
N_1 = total exposed people
N_0 = total unexposed people

Risk ratio (on log scale)
a = exposed cases
b = unexposed cases
N_1 = total exposed people
N_0 = total unexposed people

Incidence rate difference
a = exposed cases
b = unexposed cases
N_1 = total exposed person-time
N_0 = total unexposed person-time

Incidence rate ratio (log scale)
a = exposed cases
b = unexposed cases
N_1 = total exposed person-time
N_0 = total unexposed person-time

Odds ratio (case-control study, log scale)
a = exposed cases
b = unexposed cases
c = exposed controls
d = unexposed controls

Adapted from Rothman KJ. *Epidemiology, an Introduction*. 2nd ed. New York, NY: Oxford University Press; 2012:chap 9. Formulas 9-2 to 9-6.

TABLE 15-3

Study Size Formulas Based on the Width of Confidence Interval

Risk data, estimating risk difference
Risk data, estimating risk ratio 

Rate data, estimating rate difference

$$N_1 = \frac{4Z^2\left(R I_1 + I_0\right)}{R I_0 \left[\ln(F)\right]^2}$$

Rate data, estimating rate ratio

$$N_1 = \frac{4Z^2\left(R I_1 + I_0\right)}{R I^2}$$

Case-control data, estimating odds ratio

$$M_1 = \frac{4Z^2\left(R p_1 (1-p_1) + p_0 (1-p_0)\right)}{\left[\ln(F)\right]^2 \left(R p_1 p_0 (1-p_1)(1-p_0)\right)}$$

N_1 = size of the exposed cohort (persons or person-time).

M_1 = size of case group in the case-control study.

R = size of the unexposed cohort divided by size of the exposed cohort; in the case-control study, size of the control group divided by size of the case group.

p_1 = risk in the exposed cohort; in the case-control study, exposure prevalence in cases.

p_0 = risk in the unexposed cohort; in the case-control study, exposure prevalence in controls.

I_1 = rate in the exposed cohort.

I_0 = rate in the unexposed cohort.

F = width of the desired confidence interval.

Suppose that $p_1 = 0.4$ and $p_0 = 0.3$, corresponding to an RD of 0.1. If we plan a study with three times as many unexposed as exposed people ($R = 3$), and we wish the 90% confidence interval to span a distance of 0.08 (so $Z = 1.645$), Equation 15-5 gives a value for N_1 of 524 people and a total for the study size of 524 + 1,573 = 2,097. With a study of this size, the SD(RD) would be 0.0243, and half of a 90% confidence interval would be equal to 1.645 × 0.0243 ≈ 0.04, giving a full confidence interval with a width of about 0.08.

Similarly, the same study could be planned with respect to desired precision for the RR. For ratio measures, it is convenient to specify the precision in terms of the magnitude of the ratio of the upper bound to the lower bound, which leads to a constant width on the log scale. For example, the following confidence intervals all have the same precision, with a ratio of upper/lower bound of 2:1.0 to 2.0,

1.5 to 3.0, 0.8 to 1.6, etc. The formula for study size of a cohort study with person denominators giving a fixed value for the ratio of the upper to lower bound for RR confidence interval is

$$N_1 = \frac{4Z^2\left[Rp_0(1-p_1)+p_1(1-p_0)\right]}{Rp_1p_0\left[\ln(F)\right]^2}$$

where F is the desired ratio of upper to lower bound for the confidence interval. If we wish F to be 2, and assuming again that $p_1 = 0.4$ and $p_0 = 0.3$, $R = 3$, and this time using a 95% confidence interval, so that $Z = 1.96$, Equation 15-6 gives a value for N_1 of 73 people and a total study size of 291 people. This size should produce 95% confidence intervals that on the average will have an upper bound that is approximately twice the magnitude of the lower bound. If a study of this size produced results that were equal to the expected risks in the exposed and unexposed groups, we would have 29 exposed cases and 66 unexposed cases, and the 95% confidence interval for the RR would be about 0.93 to 1.86.

Table 15-3 gives all five study size formulas based on the width of the confidence interval, according to the type of data and type of effect measure. Figure 15-8 illustrates the use of the formulas with a graph showing the relation between size of a cohort study measuring risks and the width of a confidence interval for two values of p_0.

Estimating Study Size Based on the Probability That Upper Confidence Bound Stays Below a Level of Concern

Another way to use precision to plan the size of a study arises when the aim is to provide reassurance about the absence of a strong effect. No study can provide evidence for the absence of a small effect, but it is feasible and reasonable to plan a study aimed at

indicating the low compatibility between study results and strong effects, when there is a strong prior of little or no effect. In this situation, one can choose an effect level of practical concern and aim to design the study to produce a confidence interval with an upper bound below that demarcation point. Equation 15-7 and equations in Table 15-3, can be used for this calculation, with two small modifications: (1) $4Z^2$ must be replaced with $(Z' + Z)^2$, where Z' is the value of the cumulative normal distribution that corresponds to the desired probability that the upper confidence bound is below the demarcation point chosen and (2) the value of F in the formulas should be the demarcation point, rather than the width of the confidence interval. For example, if one wishes to have 90% probability that the upper bound for a rate ratio in a cohort study is below 2.0, assuming that the exposure has no effect, Z' would be 1.282, F would be 2, and the study size for a rate-data study using 95% confidence limits with a rate of 10 cases per 1,000 person-years among unexposed and among exposed, and equal sized exposed and unexposed cohorts, would be 4,374 person-years in each of the two cohorts, for a total of 8,748 person-years. If a study of this size had results equal to the expected value (44 cases in each of the two cohorts), the 95% confidence interval would be about 0.66 to 1.52, with an upper bound well below 2. However, the point estimate will be below 1.3 with 90% probability under the conditions assumed, and the upper bound of the 95% confidence interval would be 2.0 when the point estimate is 1.3.

Summary

Power as a planning tool for study size perpetuates the drawbacks of statistical significance testing. These drawbacks include the temptation to dichotomize study results into qualitative categories that notoriously have led to misinterpreting nonsignificant findings to be support for the null hypothesis and misinterpreting small, precise significant findings to be strong evidence against the null.

The formulas presented here for planning study size are keyed to the anticipated precision of the study and are consistent with the

objective of accurate estimation and the inferential goal of interpreting findings as continuous measures that are estimated with varying degrees of precision. Study size is a central determinant of the precision of study results, and it is natural to consider a study's precision in determining or anticipating the size of a study. Given the many unknown elements in implementing an epidemiologic study, the formulas here provide rough approximations, based on a simple crude analysis of data. Consequently, the results are likely to be underestimates of the study size needed for the intended precision in the light of actual data. In particular, control of confounding will usually widen the confidence interval, and missing data will also reduce precision. The degree of loss of precision from these factors will depend on the circumstances of a particular study.

Planning the size of a study requires as input some information that the study itself is intended to elucidate, such as the risk among unexposed and the effect size. This input is needed regardless of whether one is computing study size based on power or on precision. Uncertainty about these values can be addressed by graphing curves for different assumed values of these parameters, as in Figure 15-8. Another possible approach to addressing uncertainty about risks and effects would be to postulate a distribution for these parameters and to use Monte Carlo simulation, drawing repeated samples from the assumed distribution, to estimate study precision. The Monte Carlo method could also be used to extend these formulas, which apply only to dichotomous exposures. Using a Monte Carlo approach, one could assess the precision of more complicated analyses, such as precision in estimating trends in effect over a range of exposure levels, results from stratified analyses, or results affected by systematic errors (see Chapter 29).

References

1. Greenland S. Randomization, statistics, and causal inference. *Epidemiology*. 1990;1:421-429.
2. Greenland S. Multiple-bias modeling for analysis of observational data (with discussion). *J R Stat Soc Ser A*. 2005;168:267-308.

3. Gigerenzer G. Mindless statistics. *J Socioeconomics*. 2004;33:567-606.
4. Goodman SN. P Values, hypothesis tests, and likelihood: *implications for epidemiology of a neglected historical debate*. *Am J Epidemiol*. 1993;137:485-496.
5. Berkson J. Some difficulties of interpretation encountered in the application of the chi-square test. *J Am Stat Assoc*. 1938;33:526-536.
6. Berkson J. Tests of significance considered as evidence. *J Am Statist Assoc* 1942;37:325-335. Reprinted in *Int J Epidemiol*. 2003;32:687-691.
7. Boring EG. Mathematical versus statistical importance. *Psychol Bull*. 1919;16:335-338.
8. Hogben L. *Statistical Theory*. London: Allen and Unwin; 1957.
9. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. *Am Stat*. 2016;70(2):129-133. doi:10.1080/00031305.2016.1154108.
10. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: *a guide to misinterpretations (2016)*. *Eur J Epidemiol*. 2016;31:337-350. doi:10.1007/s10654-016-0149-3.
11. Lash TL. The harm done to reproducibility by the culture of null hypothesis significance testing. *Am J Epidemiol*. 2017;186:627-635.
12. McShane BB, Gal D. Statistical significance and the dichotomization of evidence. *JASA*. 2017;112:885-908.
13. Altman DG, Machin D, Bryant TN, Gardner MJ, eds. *Statistics With Confidence*. 2nd ed. London: BMJ Books; 2000.
14. Berger JO, Berry DA. Statistical analysis and the illusion of objectivity. *Am Scientist*. 1988;76:159-165.
15. Goodman SN, Royall R. Evidence and scientific research. *Am J Public Health*. 1988;78:1568-1574.
16. Greenland S. Bayesian perspectives for epidemiologic research. I. Foundations and basic methods (with comment and reply). *Int J Epidemiol*. 2006;35:765-778.
17. Royall R. *Statistical Inference*: *A Likelihood Paradigm*. New York, NY: Chapman and Hall; 1997.

18. Stang A, Deckert M, Poole C, Rothman KJ. Statistical inference in abstracts of major medical and epidemiology journals 1975-2014: *a systematic review*. *Eur J Epidemiol*. 2017;32:21-29.

19. Cox DR, Hinkley DV. *Theoretical Statistics*. New York, NY: Chapman and Hall; 1974.

20. Fisher RA. Note on Dr. Berkson's criticism of tests of significance. *J Am Statist Assoc*. 1943;38:103-104. Reprinted in *Int J Epidemiol*. 2003;32:692.

21. Berger JO, Delampady M. Testing precise hypotheses (with discussion). *Stat Sci*. 1987;2:317-352.

22. Berger JO, Sellke T. Testing a point null hypothesis: *the irreconcilability of P values and evidence (with discussion)*. *J Am Stat Assoc*. 1987;82:112-139.

23. Casella G, Berger RL. Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J Am Stat Assoc*. 1987;82:106-111.

24. Greenland S, Gustafson P. Adjustment for independent nondifferential misclassification does not increase certainty that an observed association is in the correct direction. *Am J Epidemiol*. 2006;164:63-68.

25. Edwards AWF. *Likelihood*. 2nd ed. Baltimore, MD: Johns Hopkins University Press; 1992.

26. Freedman DA. Statistics and the scientific method. In: Mason W, Feinberg SE, eds. *Cohort Analysis and Social Research*. New York, NY: Springer-Verlag; 1985:345-390.

27. Freedman DA, Pisani R, Purves R. *Statistics*. 4th ed. New York, NY: Norton; 2007.

28. Lehmann EL. *Testing Statistical Hypotheses*. 2nd ed. New York, NY: Wiley; 1986.

29. Goodman SN. A comment on replication, p-values and evidence. *Stat Med*. 1992;11:875-879.

30. Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error and sample size in the design and

interpretation of the randomized control trial. *N Engl J Med*. 1978;299:690-694.

31. Greenland S. Invited commentary: *the need for cognitive science in methodology*. *Am J Epidemiol*. 2017;186:639-645.
32. Atkins L, Jarrett D. The significance of "significance tests". In: Irvine J, Miles I, Evans J, eds. *Demystifying Social Statistics*. London: Pluto Press; 1979.
33. Blackwelder WC. Equivalence trials. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. New York, NY: John Wiley and Sons, Inc; 1998.
34. Motulsky HJ. Common misconceptions about data analysis and statistics. *Br J Pharmacol*. 2015;172:2126-2132.
35. Rothman KJ. Significance questing. *Ann Intern Med*. 1986;105:445-447.
36. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: *undisclosed flexibility in data collection and analysis allows presenting anything as significant*. *Psychol Sci*. 2011;22:1359-1366.
37. Bandt CL, Boen JR. A prevalent misconception about sample size, statistical significance, and clinical importance. *J Periodontol*. 1972;43:181-183.
38. Birnbaum A. A unified theory of estimation, I. *Ann Math Stat*. 1961;32:112-135.
39. Miettinen OS. *Theoretical Epidemiology*. New York, NY: Wiley; 1985.
40. Poole C. Beyond the confidence interval. *Am J Public Health*. 1987;77:195-199.
41. Folks JF. *Ideas of Statistics*. New York, NY: Wiley; 1981.
42. Sullivan KM, Foster DA. Use of the confidence interval function. *Epidemiology*. 1990;1:39-42.
43. Poole C. Confidence intervals exclude nothing. *Am J Public Health*. 1987;77:492-493.

44. Poole C. Low P-values or narrow confidence intervals: *which are more durable?*. *Epidemiology*. 2001;12:291-294.

45. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *Br Med J*. 1995;311:485.

46. Cox DR. *The Planning of Experiments*. New York, NY: Wiley; 1958:161.

47. Goodman SN, Berlin J. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann Intern Med*. 1994;121:200-206.

48. Hoenig JM, Heisey DM. The abuse of power: *the pervasive fallacy of power calculations for data analysis*. *Am Stat*. 2001;55:19-24.

49. Senn S. Power is indeed irrelevant in interpreting completed studies. *Br Med J*. 2002;325:1304.

50. Smith AH, Bates MN. Confidence limit analyses should replace power calculations in the interpretation of epidemiologic studies. *Epidemiol*. 1992;3:449-452.

51. Eisenberger MA, Blumenstein BA, Crawford ED, et al. Bilateral orchiectomy with or without flutamide for metastatic prostate cancer. *N Engl J Med*. 1998;339:1036-1042.

52. Prostate Cancer Trialists' Collaborative Group. Maximum androgen blockade in advanced prostate cancer: *an overview of 22 randomised trials with 3283 deaths in 5710 patients*. *Lancet*. 1995;346:265-269.

53. Rothman KJ, Johnson ES, Sugano DS. Is flutamide effective in patients with bilateral orchiectomy? *Lancet*. 1999;353:1184.

54. Stampfer MJ, Kang JH, Chen J, Cherry R, Grodstein F. Effects of moderate alcohol consumption on cognitive function in women. *N Engl J Med*. 2005;352:245-253.

55. Collins FS, Tabak LA. Policy: *NIH plans to enhance reproducibility*. *Nature*. 2014;505(7485):612-613.

56. Nosek BA, Alter G, Banks GC, et al. Promoting an open research culture. *Science*. 2015;348(6242):1422-1425.

57. Lash TL, Collin LJ, Van Dyke ME. The replication crisis in epidemiology: *snowball, snow job, or winter solstice?*. *Curr Epidemiol Rep*. 2018;5:175-183.

58. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*. 2015;349(6251):aac4716.

59. Gelman A, Stern H. The difference between "significant" and "not significant" is not itself statistically significant. *Am Statistician*. 2006;60:328-331.

60. Rothman KJ, Lanes S, Robins J. Causal inference. *Epidemiology*. 1993;4(6):555-556.

61. Seliger C, Meier CR, Becker C, et al. Statin use and risk of glioma: *population-based case–control analysis*. *Eur J Epidemiol*. 2016;31(9):947-952.

62. Brown HK, Ray JG, Wilton AS, Lunsky Y, Gomes T, Vigod SN. Association between serotonergic antidepressant use during pregnancy and autism spectrum disorder in children. *J Am Med Assoc*. 2017;317(15):1544-1552.

63. Rozeboom WM. The fallacy of null-hypothesis significance test. *Psych Bull*. 1960;57:416-428.

64. Morrison DE, Henkel RE, eds. *The Significance Test Controversy*. Chicago, IL: Aldine; 1970.

65. Wulff HR. Confidence limits in evaluating controlled therapeutic trials. *Lancet*. 1973;2:969-970.

66. Rothman KJ. A show of confidence. *N Engl J Med*. 1978;299:1362-1363.

67. Salsburg DS. The religion of statistics as practiced in medical journals. *Am Statist*. 1985;39:220-223.

68. Simon R, Wittes RE. Methodologic guidelines for reports of clinical trials. *Cancer Treat Rep*. 1985;69:1-3.

69. Langman MJS. Towards estimation and confidence intervals. *Br Med J*. 1986;292:716.

70. Gardner MA, Altman DG. Confidence intervals rather than P values: *estimation rather than hypothesis testing*. *Br Med J*.

1986;292:746-750.

71. Walker AM. Reporting the results of epidemiologic studies. *Am J Public Health*. 1986;76:556-558.

72. Oakes M. *Statistical Inference*. Chestnut Hill, MA: ERI; 1990.

73. Ware JH, Mosteller F, Ingelfinger JA. *P values*. In: *Medical Uses of Statistics*. Waltham, MA: NEJM Books; 1986.

74. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. *N Eng J Med*. 1987;317:426-432.

75. Thompson WD. Statistical criteria in the interpretation of epidemiologic data. *Am J Public Health*. 1987;77:191-194.

76. Evans SJW, Mills P, Dawson J. The end of the P-value? *Br Heart J*. 1988;60:177-180.

77. Anscombe FJ. The summarizing of clinical experiments by significance levels. *Stat Med*. 1990;9:703-708.

78. Cohen J. The earth is round ($P < 0.05$). *Am Psychol*. 1994;47:997-1003.

79. Hauer E. The harm done by tests of significance. *Accid Anal Prev*. 2003;36:495-500.

80. Ziliak ST, McCloskey DN. Size matters: *the standard error of regressions in the American Economic Review*. *J Socio-Economics*. 2004;33:527-546.

81. Batterham AM, Hopkins WG. Making meaningful inferences about magnitudes. *Int J Sports Physiol Perform*. 2006;1:50-57.

82. Marshall SW. Commentary on making meaningful inferences about magnitudes. *Sportscience*. 2006;9:43-44.

83. Lash TL, Fox MP, Greenland S, et al. Re: *promoting healthy skepticism in the news. Helping journalists get it right*. *J Natl Cancer Inst*. 2010;102:829-830.

84. Berger JO, Wolpert RL. *The Likelihood Principle*. 2nd ed. Hayward, CA: Institute of Mathematical Statistics; 1988.

85. Clayton D, Hills M. *Statistical Models in Epidemiology*. New York, NY: Oxford University Press; 1993.

86. DeFinetti B. *The Theory of Probability*. Vol 1. New York, NY: Wiley; 1974.

87. Howson C, Urbach P. *Scientific Reasoning*: *The Bayesian Approach*. 2nd ed. LaSalle, IL: Open Court; 1993.

88. Greenland S. Probability logic and probabilistic induction. *Epidemiology*. 1998;9:322-332.

89. Savage LJ. *The Foundations of Statistics*. New York, NY: Dover; 1972.

90. Freedman DA. As others see us: *a case study in path analysis (with discussion)*. *J Educ Stat*. 1987;12:101-223.

91. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. New York, NY: Wiley; 2004.

92. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. 2nd ed. New York, NY: Chapman and Hall/CRC; 2003.

93. Rubin DB. Bayesianly justifiable and relevant frequency calculations. *Ann Stat*. 1984;12:1151-1172.

94. Rubin DB. Practical implications of modes of statistical inference for causal effects, and the critical role of the assignment mechanism. *Biometrics*. 1991;47:1213-1234.

95. Greenland S. A semi-Bayes approach to the analysis of correlated multiple associations, with an application to an occupational cancer-mortality study. *Stat Med*. 1992;11:219-230.

96. Schlesselman JJ. Sample size requirements in cohort and case-control studies of disease. *Am J Epidemiol*. 1974;99(6):381-384.

97. Rothman KJ, Boice JD. *Epidemiologic Analysis with a Programmable Calculator*. 2nd ed. Newton MA: Epidemiology Resources; 1982.

98. Greenland S.Power, sample size and smallest detectable effect determination for multivariate studies. *Stat Med*. 1985;4(2):117-127.

99. Greenland S. On sample-size and power calculations for studies using confidence intervals. *Am J Epidemiol*. 1988;128(1):231-237.

100. Greenland S. On sample-size and power calculations for studies using confidence intervals. *Am J Epidemiol*. 1988;128:231-237.

101. Walter SD. Determination of significant relative risks and optimal sampling procedures in prospective and retrospective comparative studies of various sizes. *Am J Epidemiol*. 1977;105:387-397.

102. Rothman KJ, Mosquin PL: Sparse-data bias accompanying overly fine stratification in an analysis of beryllium exposure and lung cancer risk. *Ann Epidemiol*. 2013;23(2):43-48. doi:10.1016/j.annepidem.2012.11.005.

103. Fink AK, Lash TL. A null association between smoking during pregnancy and breast cancer using Massachusetts registry data (United States). *Cancer Causes Control*. 2003;14:497-503.

104. Greenland S, Mansournia MA, Altman DG. Sparse-data bias: *a problem hiding in plain sight*. *Br Med J*. 2016;353:i1981. doi:10.1136/bmj.i1981.

105. Neyman J. Frequentist probability and frequentist statistics. *Synthese*. 1977;36:97-131.

106. Lakens D, Adolfi FG, Albers CJ, et al. Justify your alpha. *Nat Hum Behav*. 2018;2:168-171.

107. Greenland S. Nonsignificance plus high power does not imply support for the null over the alternative. *Ann Epidemiol*. 2012;22:364-368.

108. Bland JM. The tyranny of power: *is there a better way to calculate sample size? Br Med J*. 2009;339:b3985.

109. Kelsey JL, Thompson WD, Evans AS. *Methods in Observational Epidemiology*. New York, NY: Oxford University Press; 1986:chap 10.

110. Miettinen OS. Individual matching with multiple controls in the case of all or none responses. *Biometrics*. 1969;25:339-355.

111. Rothman KJ, Greenland S. Planning study size based on precision rather than power. *Epidemiology*. 2018;29:599-603.