DESCRIPTIVE STATISTICS II

(MEASURES OF SPREAD OF A SAMPLED VARIABLE

(DA_2022)

Andrea Giansanti
Dipartimento di Fisica, Sapienza Università di Roma

Andrea.Giansanti@roma1.infn.it

Lecture n. 5, Rome 16th March 2022

DIPARTIMENTO DI FISICA

SAPIENZA
UNIVERSITÀ DI ROMA

Summary

- The location of a distribution for a numerical variable can be measured by its mean or by its median. The mean gives the center of gravity of the distribution and is calculated as the sum of all measurements divided by the number of measurements. The median gives the middle value.
- The standard deviation measures the spread of a distribution for a numerical variable. It is a measure of the typical distance between observations and the mean. The variance is the square of the standard deviation.
- The quartiles break the ordered observations into four equal parts. The inter-quartile range, the difference between the first and third quartiles, is another measure of the spread of a frequency distribution.
- The mean and median yield similar information when the frequency distribution of the measurements is symmetric and unimodal. The mean and standard deviation become less informative about the location and spread of typical observations than the median and interquartile range when the data include extreme observations.
- The percentile of a measurement specifies the percentage of observations less than or equal to it. The quantile of a measurement specifies the fraction of observations less than or equal to it.
- All the quantiles of a sample of data can be shown using a graph of the cumulative frequency distribution.
- The proportion is the most important descriptive statistic for a categorical variable. It is calculated by dividing the number of observations in the category of interest by n, the total number of observations in all categories combined.

KEYWORDS OF LECTURE N. 5

DESCRIPTIVE STATISTICS II
(MEASURES OF SPREAD OF A SAMPLED VARIABLE)

- range
- Quantiles
- interquartile range
- sample variance and standard deviation
- coefficient of variation

FROM HISTOGRAMS TO PROBABILITY DISTRIBUTIONS

modes of science: deduction, induction, abduction
modes of presenting data:
scatter plots
bar graphs
pie charts
strip charts
box plots
frequency tables/histograms
Binning/resolution
sampling the distribution of estimates (statistics)
the mean of means
self-averaging/non self-averaging quantities

Back_issue n 2 the scientific method in a nutshell

- Deduction/ Induction/Abduction/
- The structure of a scientific paper (ad nauseam)
- (make your ideas clear for a brief discussion next monday)

-Introduction

-Materials and methods

-Results

-Discussion

Peirce's Abduction

DEDUCTION

Rule.—All the beans from this bag are white.
Case.—These beans are from this bag.
∴ Result.—These beans are white.

INDUCTION

Case.—These beans are from this bag.
Result.—These beans are white.
∴ Rule.—All the beans from this bag are white.

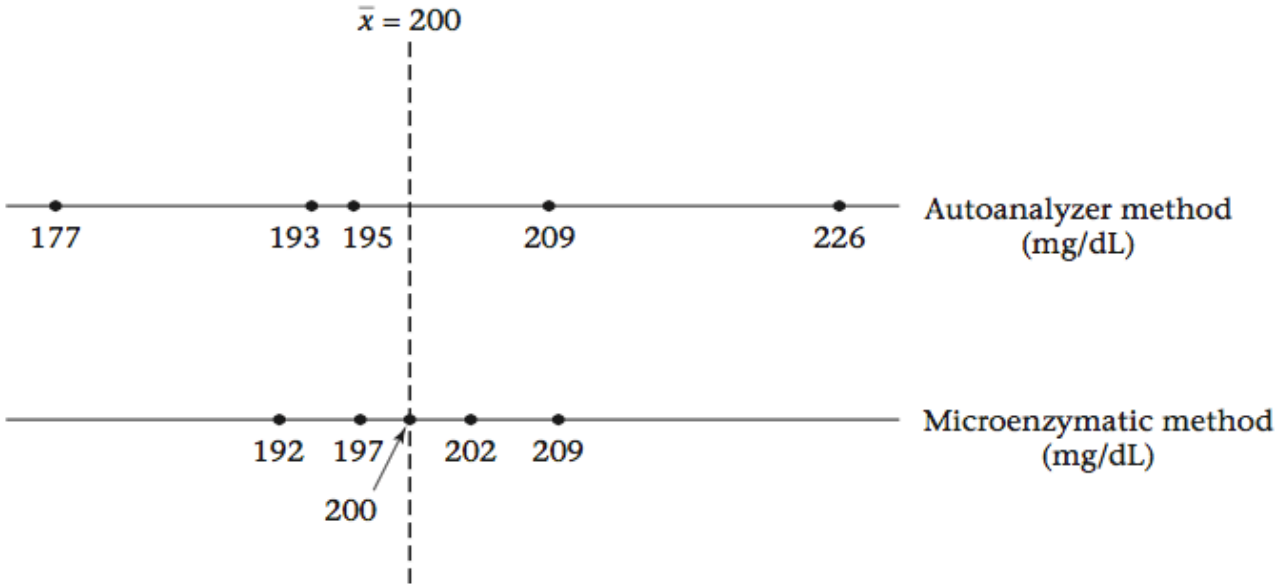
HYPOTHESIS

Rule.—All the beans from this bag are white.
Result.—These beans are white.
∴ Case.—These beans are from this bag.

Illustration of the Logic of Science, 1878, 1893

DEFINITION 2.5 The **range** is the difference between the largest and smallest observations in a sample.

FIGURE 2.4 **Two samples of cholesterol measurements on a given person using the Autoanalyzer and Microenzymatic measurement methods**



From:[**R**]Bernard Rosner - Fundamentals of Biostatistics-Brooks Cole (2015)

Quantiles

Another approach that addresses some of the shortcomings of the range in quantifying the spread in a data set is the use of **quantiles** or **percentiles.** Intuitively, the pth percentile is the value V_p such that p percent of the sample points are less than or equal to V_p. The median, being the 50th percentile, is a special case of a quantile. As was the case for the median, a different definition is needed for the pth percentile, depending on whether or not $np/100$ is an integer.

DEFINITION 2.6 The ***p*th percentile** is defined by

(1) The $(k + 1)$th largest sample point if $np/100$ is not an integer (where k is the largest integer less than $np/100$).

(2) The average of the $(np/100)$th and $(np/100 + 1)$th largest observations if $np/100$ is an integer.

Percentiles are also sometimes called **quantiles.**

The spread of a distribution can be characterized by specifying several percentiles. For example, the 10th and 90th percentiles are often used to characterize spread. Percentiles have the advantage over the range of being less sensitive to outliers and of not being greatly affected by the sample size (n).

From:[**R**]Bernard Rosner - Fundamentals of Biostatistics-Brooks Cole (2015)

The interquartile range

Quartiles are values that partition the data into quarters. The first quartile is the middle value of the measurements lying below the median. The second quartile is the median. The third quartile is the middle value of the measurements larger than the median. The **interquartile range** (*IQR*) is the span of the middle half of the data, from the first quartile to the third quartile:
Interquartile range=third quartile−first quartile. $\text{Interquartile range} = \text{third quartile} - \text{first quartile}.$

The ***interquartile range*** is the difference between the third and first quartiles of the data. It is the span of the middle 50% of the data.

Figure 3.2-1 shows the meaning of the median, first quartile, third quartile, and interquartile range for the spider data set (before amputation).

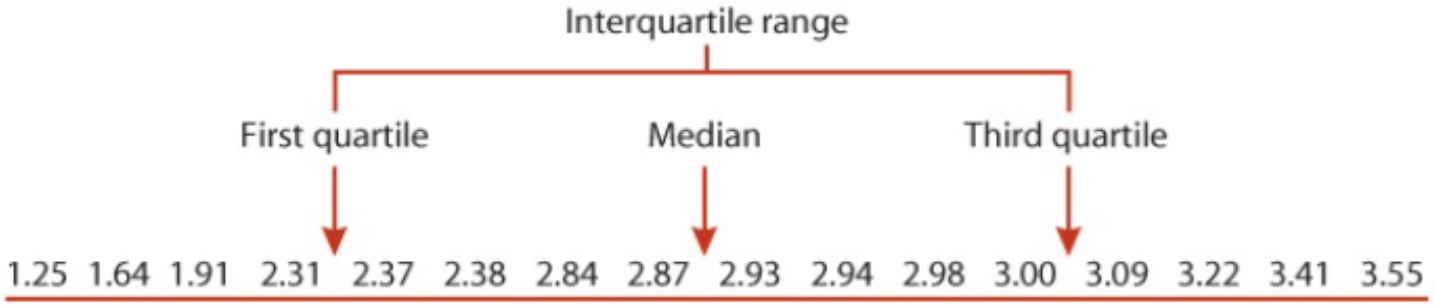


Figure 3.2-1
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

The Median REVISITED

An alternative measure of location, perhaps second in popularity to the arithmetic mean, is the **median** or, more precisely, the **sample median**.

Suppose there are n observations in a sample. If these observations are ordered from smallest to largest, then the median is defined as follows:

DEFINITION 2.2 The **sample median** is

(1) The $\left(\frac{n+1}{2}\right)$th largest observation if n is odd

(2) The average of the $\left(\frac{n}{2}\right)$th and $\left(\frac{n}{2}+1\right)$th largest observations if n is even

---

The rationale for these definitions is to ensure an equal number of sample points on both sides of the sample median. The median is defined differently when n is even and odd because it is impossible to achieve this goal with one uniform definition. Samples with an odd sample size have a unique central point; for example, for samples of size 7, the fourth largest point is the central point in the sense that 3 points are smaller than it and 3 points are larger. Samples with an even sample size have no unique central point, and the middle two values must be averaged. Thus, for samples of size 8 the fourth and fifth largest points would be averaged to obtain the median, because neither is the central point.

The Variance and Standard Deviation

The main difference between the Autoanalyzer- and Microenzymatic-method data in Figure 2.4 is that the Microenzymatic-method values are closer to the center of the sample than the Autoanalyzer-method values. If the center of the sample is defined as the arithmetic mean, then a measure that can summarize the difference (or deviations) between the individual sample points and the arithmetic mean is needed; that is,

$$x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$$

One simple measure that would seem to accomplish this goal is

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

Unfortunately, this measure will not work, because of the following principle:

From:[**R**]Bernard Rosner - Fundamentals of Biostatistics-Brooks Cole (2015)

The sum of the deviations of the individual observations of a sample about the sample mean is always zero.

Compute the sum of the deviations about the mean for the Autoanalyzer- and Microenzymatic-method data in Figure 2.4.

Solution: For the Autoanalyzer-method data,

$$\begin{aligned} d &= (177 - 200) + (193 - 200) + (195 - 200) + (209 - 200) + (226 - 200) \\ &= -23 - 7 - 5 + 9 + 26 = 0 \end{aligned}$$

For the Microenzymatic-method data,

$$\begin{aligned} d &= (192 - 200) + (197 - 200) + (200 - 200) + (202 - 200) + (209 - 200) \\ &= -8 - 3 + 0 + 2 + 9 = 0 \end{aligned}$$

Thus, d does not help distinguish the difference in spreads between the two methods.

A second possible measure is

$$\sum_{i=1}^{n} |x_i - \bar{x}| / n$$

which is called the **mean deviation**. The mean deviation is a reasonable measure of spread but does not characterize the spread as well as the standard deviation (see Definition 2.8) if the underlying distribution is bell-shaped.

A third idea is to use the average of the squares of the deviations from the sample mean rather than the deviations themselves. The resulting measure of spread, denoted by s^2, is

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

The more usual form for this measure is with $n - 1$ in the denominator rather than n. The resulting measure is called the *sample variance* (or *variance*).

From:[**R**]Bernard Rosner - Fundamentals of Biostatistics-Brooks Cole (2015)

The most common measures of spread: VARIANCE & SDEV

DEFINITION 2.7 The **sample variance**, or **variance**, is defined as follows:

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

A rationale for using $n - 1$ in the denominator rather than n is presented in the discussion of estimation in Chapter 6.

Another commonly used measure of spread is the sample standard deviation.

DEFINITION 2.8 The **sample standard deviation**, or **standard deviation**, is defined as follows:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\text{sample variance}}$$

From:[**R**]Bernard Rosner - Fundamentals of Biostatistics-Brooks Cole (2015)

Solution: Autoanalyzer Method

$$s^2 = \left[(177-200)^2 + (193-200)^2 + (195-200)^2 + (209-200)^2 + (226-200)^2\right]/4$$
$$= (529+49+25+81+676)/4 = 1360/4 = 340$$
$$s = \sqrt{340} = 18.4$$

Microenzymatic Method

$$s^2 = \left[(192-200)^2 + (197-200)^2 + (200-200)^2 + (202-200)^2 + (209-200)^2\right]/4$$
$$= (64+9+0+4+81)/4 = 158/4 = 39.5$$
$$s = \sqrt{39.5} = 6.3$$

Thus the Autoanalyzer method has a standard deviation roughly three times as large as that of the Microenzymatic method.

From:[**R**]Bernard Rosner - Fundamentals of Biostatistics-Brooks Cole (2015)

Two theorems on the variance (see [**R**]Rosner: par 2.5)

1) If a constant is added to a sample of data **then** the sample variance is not changed.

2) **If** a sample of data is multiplyied by a constant c **then** the sample variance is multiplyied by c^2

Let us make an exercise in Excel

Use Microsoft Excel to compute the mean and standard deviation for the Autoanalyzer and Microenzymatic-method data in Figure 2.4.

Solution: We enter the Autoanalyzer and Microenzymatic data in cells B3–B7 and C3–C7, respectively. We then use the Average and StDev functions to evaluate the mean and standard deviation as follows:

	Autoanalyzer Method	Microenzymatic Method
	177	192
	193	197
	195	200
	209	202
	226	209
Average	200	200
StDev	18.4	6.3

In Excel, if we make B8 the active cell and type = Average(B3:B7) in that cell, then the mean of the values in cells B3, B4, . . . , B7 will appear in cell B8. Similarly, specifying = Stdev(B3:B7) will result in the standard deviation of the Autoanalyzer Method data being placed in the active cell of the spreadsheet.

From:[**R**]Bernard Rosner - Fundamentals of Biostatistics-Brooks Cole (2015)

A dataset to practice with (you have also BMI.txt)

TABLE 2.1 **Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period**

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

From:[**R**]Bernard Rosner - Fundamentals of Biostatistics-Brooks Cole (2015)

Precision of a measurement (sampling):
relative uncertainty as measured by:

2.6 THE COEFFICIENT OF VARIATION

It is useful to relate the arithmetic mean and the standard deviation to each other because, for example, a standard deviation of 10 means something different conceptually if the arithmetic mean is 10 versus if it is 1000. A special measure, the coefficient of variation, is often used for this purpose.

DEFINITION 2.9 The **coefficient of variation** (*CV*) is defined by

$$100\% \times (s/\bar{x})$$

This measure remains the same regardless of what units are used because if the units change by a factor c, then both the mean and standard deviation change by the factor c; while the *CV*, which is the ratio between them, remains unchanged.

From:[**R**]Bernard Rosner - Fundamentals of Biostatistics-Brooks Cole (2015)

At this point go to [R] Rosner's textbook and illustrate the concepts in detail

Then connect to Di Leonardo's course Data Analysis:
lecture n. 3 where you can find a notebook you can download and put in the .ipynb format

HOW TO ORGANIZE DATA I
(see WS chapter 2 (read all)

THE SCATTER PLOT

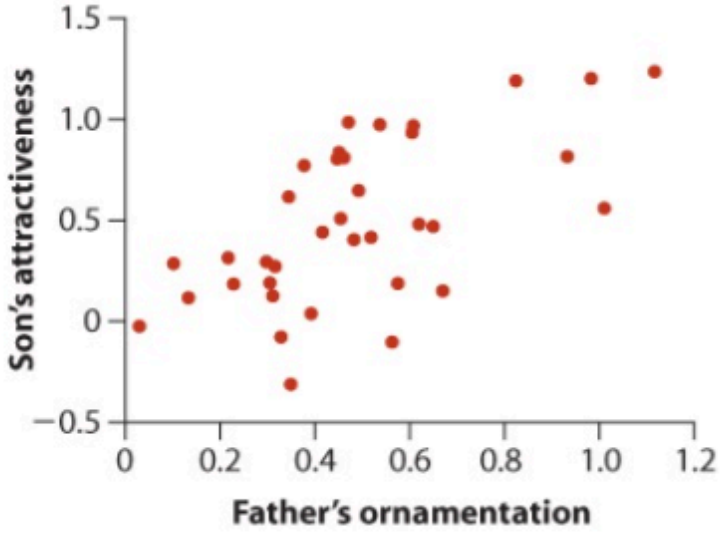


Figure 2.3-3
Whitlock et al., *The Analysis of Biological Data*, 2e,
© 2015 W. H. Freeman and Company

FIGURE 2.3-3 Scatter plot showing the relationship between the ornamentation of male guppies and the average attractiveness of their sons. Total number of families: $n = 36$.

BAR GRAPHS->histograms

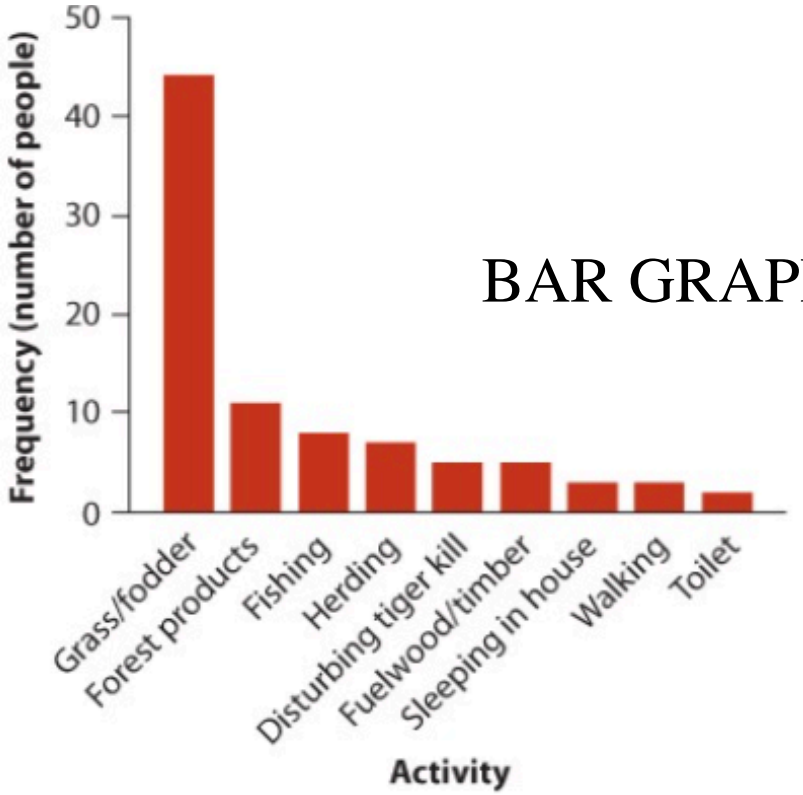


Figure 2.2-1
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

FIGURE 2.2-1 Bar graph showing the activities of people at the time they were attacked and killed by tigers near Chitwan National Park, Nepal, between 1979 and 2006. Total number of deaths: $n = 88$. The frequencies are taken from Table 2.2-1, which also gives more detailed labels of activities.

PIE CHARTS

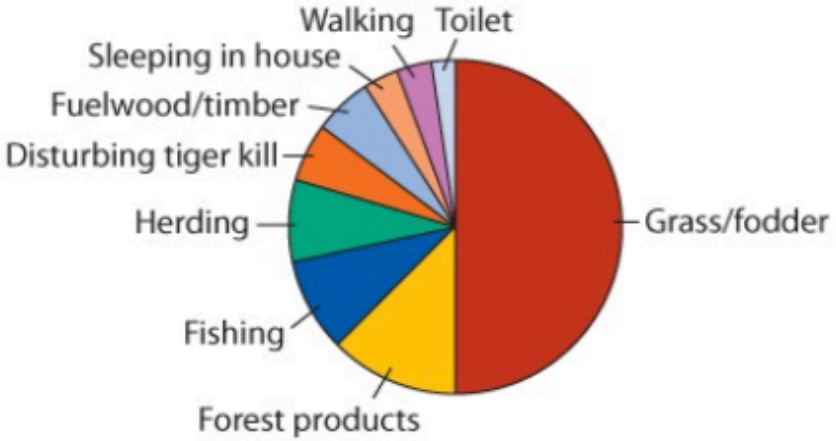


Figure 2.2-2
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015
W. H. Freeman and Company

FIGURE 2.2-2 Pie chart of the activities of people at the time they were attacked and killed by tigers near Chitwan National Park, Nepal. The frequencies are taken from Table 2.2-1. Total number of deaths: $n = 88$.

Let us start with a dirty exercise

Quick Formula Summary

Table of formulas for descriptive statistics

Quantity	Formula
Sample size	n
Mean	$\bar{Y} = \frac{\sum Y}{n}$
Variance	$s^2 = \frac{\sum (Y_i - \bar{Y})^2}{n-1}$
shortcut formula:	$s^2 = \frac{\sum Y_i^2 - n\bar{Y}^2}{n-1}$
Standard deviation	$s = \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{n-1}}$
shortcut formula:	$s = \sqrt{\frac{\sum (Y_i^2) - n\bar{Y}^2}{n-1}}$
Sum of squares	$\sum (Y_i - \bar{Y})^2 = \sum (Y_i^2) - n\bar{Y}^2$
Coefficient of variation	$CV = \frac{s}{\bar{Y}} \times 100\%$
Median	$Y_{([n+1]/2)}$ (if n is odd) $[Y_{(n/2)} + Y_{(n/2+1)}]/2$ (if n is odd) where $Y_{(1)}, Y_{(2)}, \ldots, Y(n)$ are the ordered observations
Proportion	$\hat{p} = \frac{\text{Number in category}}{n}$

Dirty exercise n.2

Effect of arithmetic operations on descriptive statistics

The table below lists the effect on the descriptive statistics of adding or multiplying all the measurements by a constant. The rules listed in the table are useful when converting measurements from one system of units to another, such as English to metric or degrees Fahrenheit to degrees Celsius.

Statistic	Value	Adding a constant c to all the measurements, $Y'=Y+c$	Multiplying all the measurements by a constant c, $Y'=cY$
Mean	$\bar{Y}$	$\bar{Y}'=\bar{Y}+c$	$\bar{Y}'=c\bar{Y}$
Standard deviation	s	$s' = s$	$s' = \|c\|s$
Variance	s^2	$s'^2 = s^2$	$s'^2 = c^2s^2$
Median	M	$M' = M + c$	$M' = cM$
Interquartile range	IQR	$IQR' = IQR$	$IQR' = \|c\|IQR$

HOW TO ORGANIZE DATA
(see WS chapter 2 (read all))

THE SCATTER PLOT

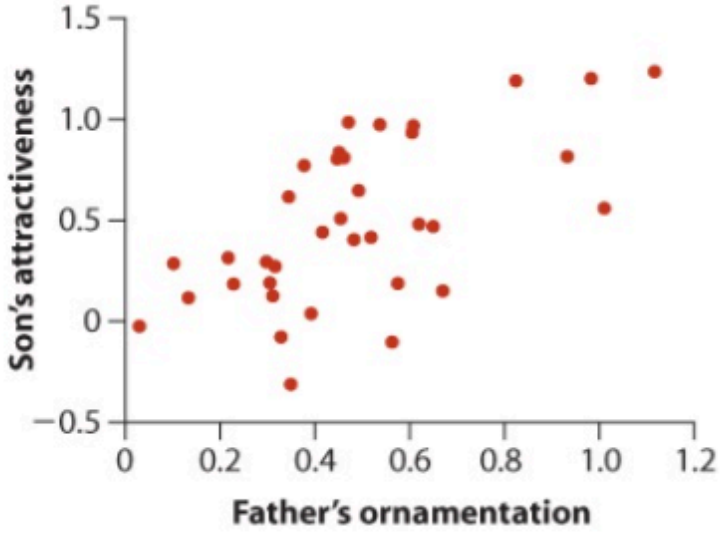


Figure 2.3-3
Whitlock et al., *The Analysis of Biological Data*, 2e,
© 2015 W. H. Freeman and Company

FIGURE 2.3-3 Scatter plot showing the relationship between the ornamentation of male guppies and the average attractiveness of their sons. Total number of families: $n = 36$.

BAR GRAPHS->histograms

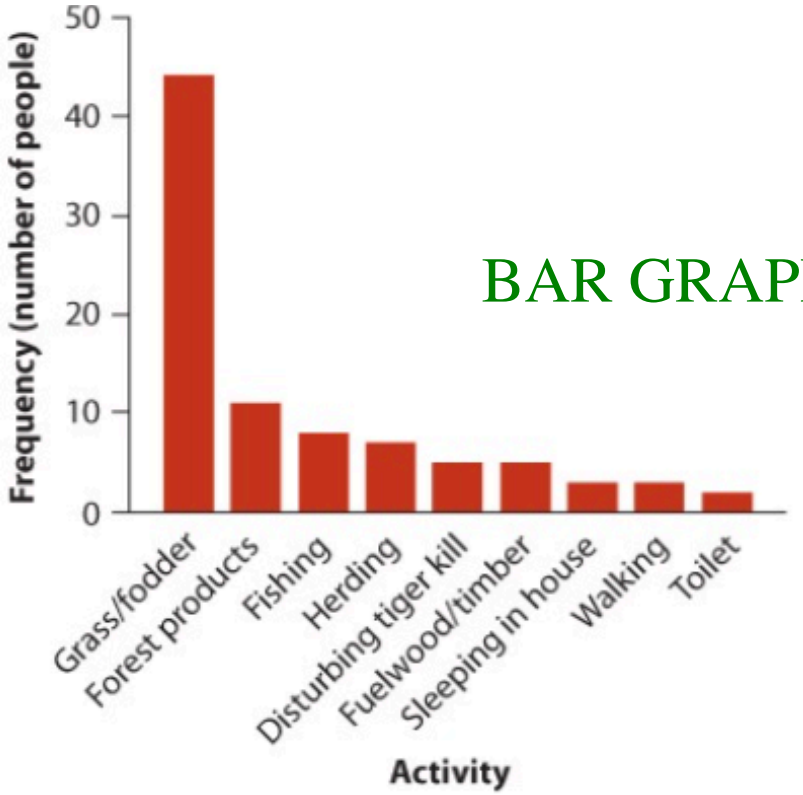


Figure 2.2-1
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

FIGURE 2.2-1 Bar graph showing the activities of people at the time they were attacked and killed by tigers near Chitwan National Park, Nepal, between 1979 and 2006. Total number of deaths: $n = 88$. The frequencies are taken from Table 2.2-1, which also gives more detailed labels of activities.

PIE CHARTS

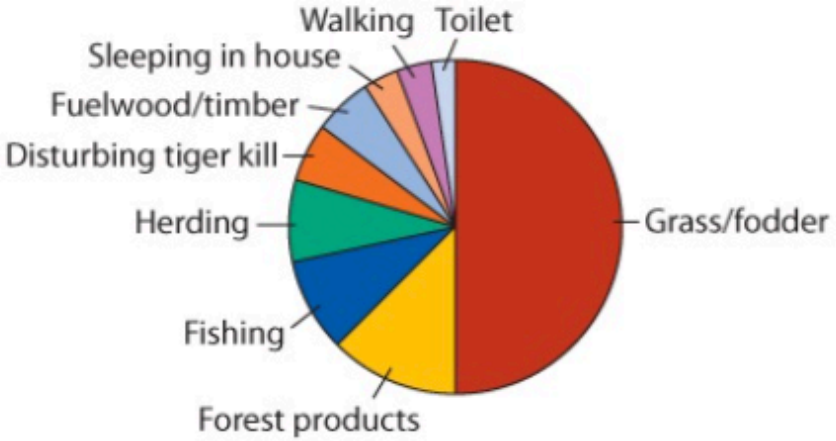
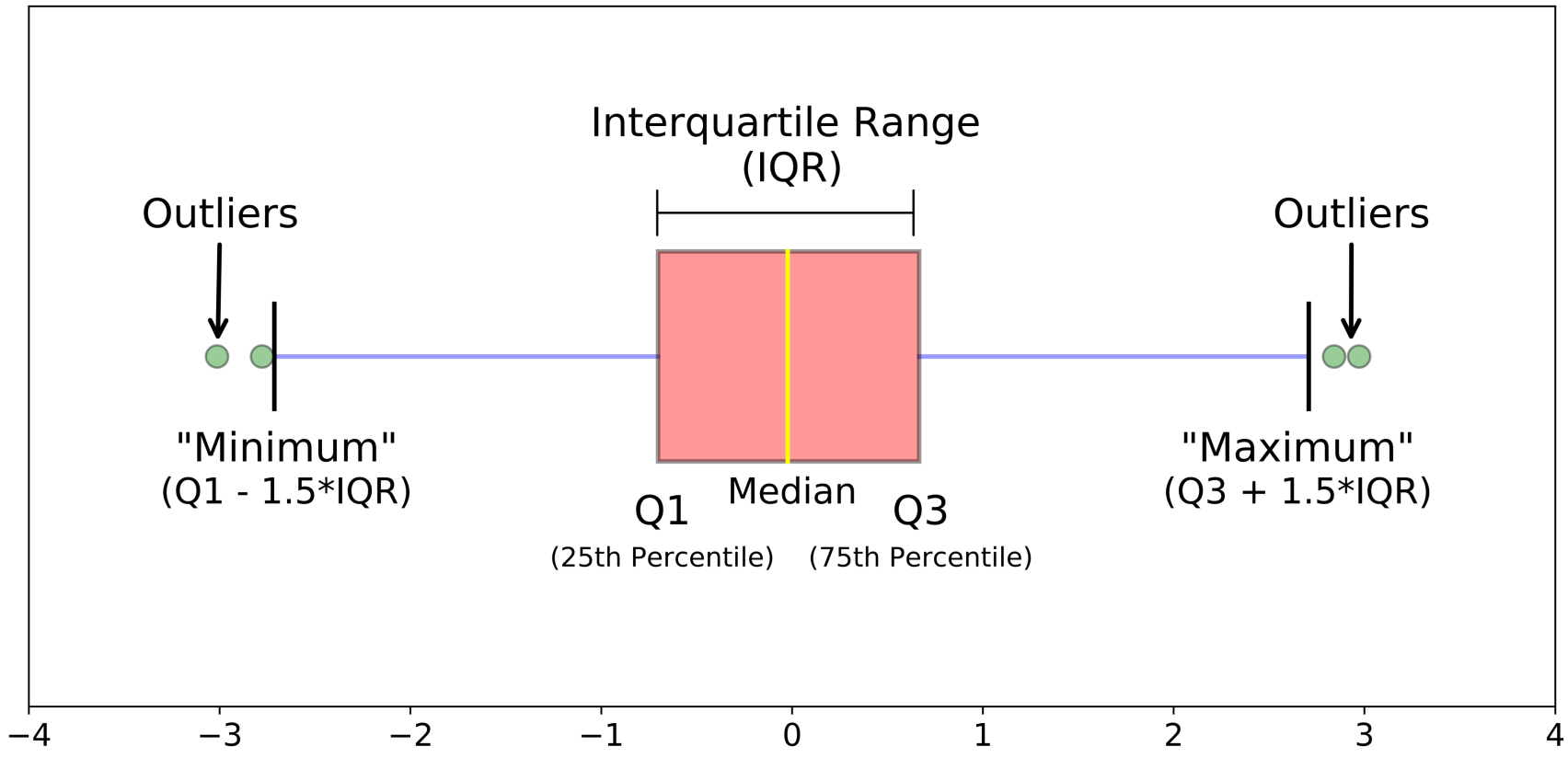


Figure 2.2-2
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015
W. H. Freeman and Company

FIGURE 2.2-2 Pie chart of the activities of people at the time they were attacked and killed by tigers near Chitwan National Park, Nepal. The frequencies are taken from Table 2.2-1. Total number of deaths: $n = 88$.

Anatomy of a BOXPLOT



https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

The ***strip chart*** is a graphical display of a numerical variable and a categorical variable in which each observation is represented as a dot.

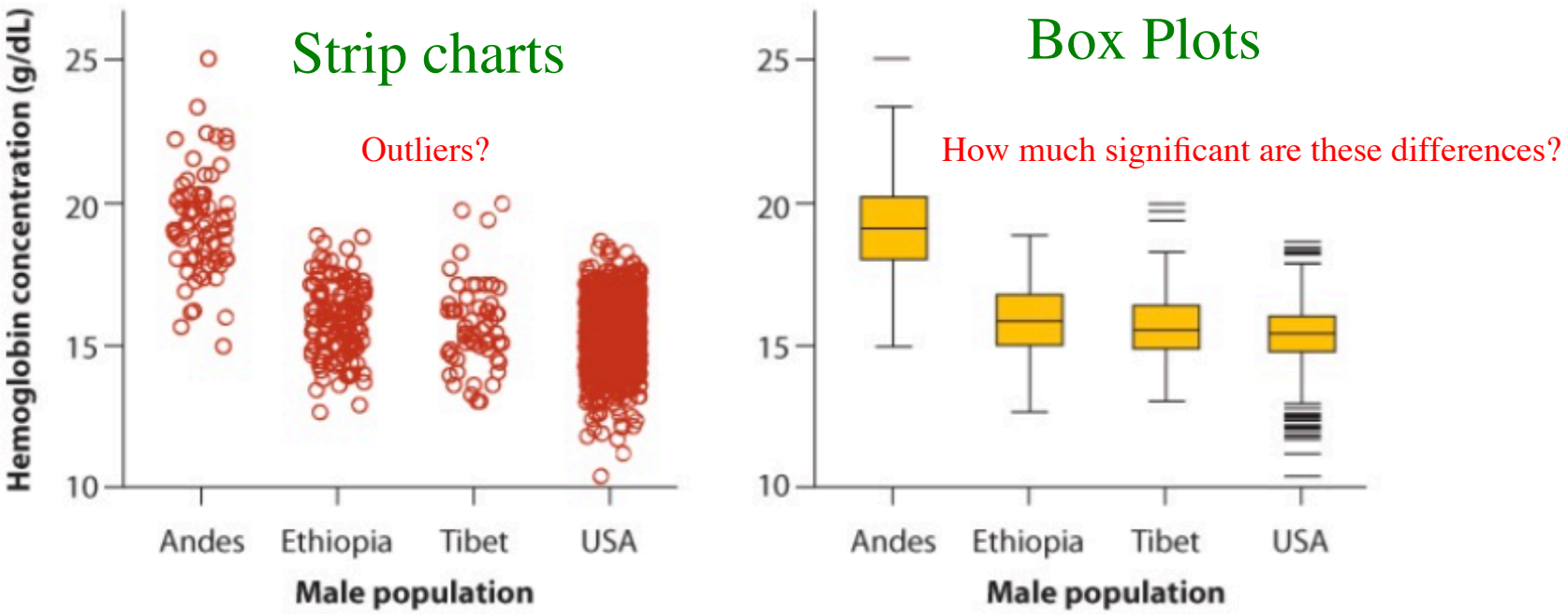


Figure 2.3-4
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

FIGURE 2.3-4 Strip chart (*left*) and box plot (*right*) showing hemoglobin concentration in males living at high altitude in three different parts of the world: the Andes (71), Ethiopia (128), and Tibet (59). A fourth population of 1704 males living at sea level (USA) is included as a control.

A ***box plot*** is a graph that uses lines and a rectangular box to display the median, quartiles, range, and extreme measurements of the data.

Showing numerical data: frequency table and histogram

A frequency distribution for a numerical variable can be displayed either in a frequency table or in a **histogram**. A histogram uses area of rectangular bars to display frequency. The data values are split into consecutive intervals, or "bins," usually of equal width, and the frequency of observations falling into each bin is displayed.

A ***histogram*** uses the area of rectangular bars to display the frequency distribution (or relative frequency distribution) of a numerical variable.

We discuss how histograms are made in greater detail using the data in Example 2.2B.

EXAMPLE 2.2B Abundance of desert bird species

How many species are common in nature and how many are rare? One way to address this question is to construct a frequency distribution of species abundance. The data in Table 2.2-2 are from a survey of the breeding birds of Organ Pipe Cactus National Monument in southern Arizona, USA. The measurements were extracted from the North American Breeding Bird Survey, a continent-wide data set of estimated bird numbers (Sauer et al. 2003).

Data: first part

TABLE 2.2-2 Data on the abundance of each species of bird encountered during four surveys in Organ Pipe Cactus National Monument.

Species	Abundance
Greater roadrunner	1
Black-chinned hummingbird	1
Western kingbird	1
Great-tailed grackle	1
Bronzed cowbird	1
Great horned owl	2
Costa's hummingbird	2
Canyon wren	2
Canyon towhee	2
Harris's hawk	3
Loggerhead shrike	3
Hooded oriole	4
Northern mockingbird	5
American kestrel	7
Rock dove	7
Bell's vireo	10
Common raven	12
Northern cardinal	13
House sparrow	14
Ladder-backed woodpecker	15
Red-tailed hawk	16
Phainopepla	18
Turkey vulture	23
Violet-green swallow	23
Lesser nighthawk	25
Scott's oriole	28
Purple martin	33
Black-throated sparrow	33
Brown-headed cowbird	59
Black vulture	64

Data: 2nd part (binning)

Lucy's warbler	67
Gilded flicker	77
Brown-crested flycatcher	128
Mourning dove	135
Gambel's quail	148
Black-tailed gnatcatcher	152
Ash-throated flycatcher	173
Curve-billed thrasher	173
Cactus wren	230
Verdin	282
House finch	297
Gila woodpecker	300
White-winged dove	625

We treated each bird species in the survey as the unit of interest and the abundance of a species in the survey as its measurement. The range of abundance values was divided into 13 intervals of equal width (0–50, 50–100, and so on), and the number of species falling into each abundance interval was counted and presented in a frequency table to help see patterns (Table 2.2-3).

Make a **frequency table** out of the data

TABLE 2.2-3 Frequency distribution of bird species abundance at Organ Pipe Cactus National Monument.

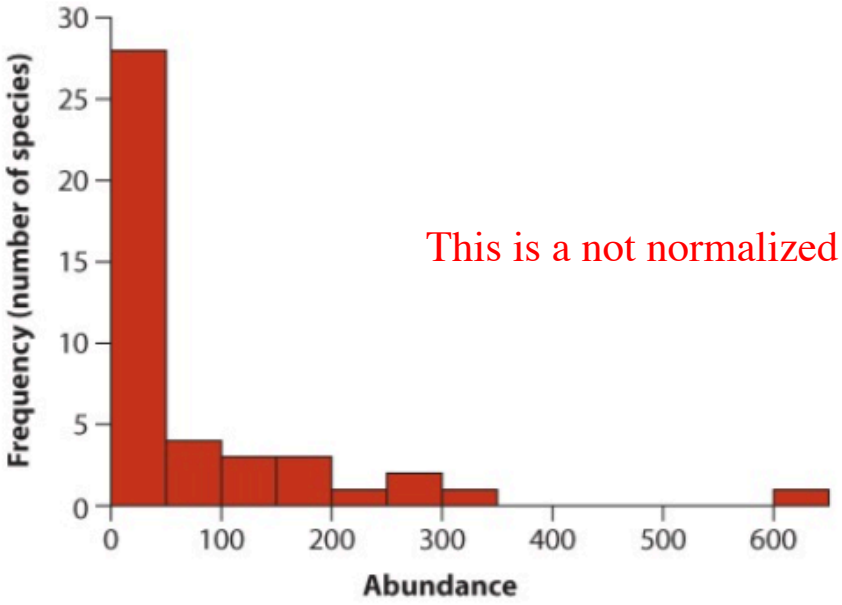
Abundance	Frequency (Number of species)
0–50	28
50–100	4
100–150	3
150–200	3
200–250	1
250–300	2
300–350	1
350–400	0
400–450	0
450–500	0
500–550	0
550–600	0
600–650	1
Total	43

Binning of the data

Source: Data are from Table 2.2-2.

Although the table shows the numbers, the shape of the frequency distribution is more obvious in a histogram of these same data (Figure 2.2-3). Here, frequency (number of species) in each abundance interval is perceived as bar area.

From table to histogram



This is a not normalized histogram!

Figure 2.2-3
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

FIGURE 2.2-3 Histogram illustrating the frequency distribution of bird species abundance at Organ Pipe Cactus National Monument. Total number of bird species: $n = 43$.

Describing the shape of a histogram

How is distributed the mass of the data?

The histogram reveals the shape of a frequency distribution. Some of the most common shapes are displayed in Figure 2.2-4. Any interval of the frequency distribution that is noticeably more frequent than surrounding intervals is called a peak. The **mode** is the interval corresponding to the highest peak. For example, a bell-shaped frequency distribution has a single peak (the mode) in the center of the range of observations. A frequency distribution having two distinct peaks is said to be **bimodal**.

The ***mode*** is the interval corresponding to the highest peak in the frequency distribution.

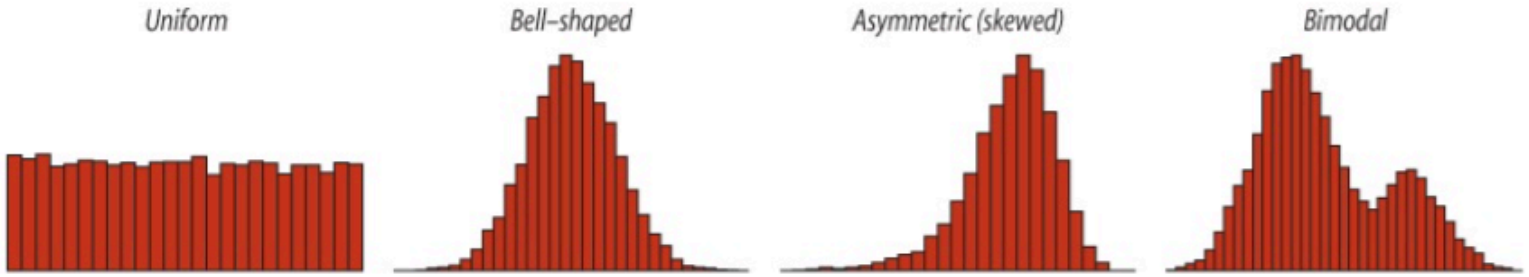


Figure 2.2-4
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

FIGURE 2.2-4 Some possible shapes of frequency distributions.

A frequency distribution is **symmetric** if the pattern of frequencies on the left half of the histogram is the mirror image of the pattern on the right half. The uniform distribution and the

Sample Skewness and Kurtosis (see e.g.Wikipedia)

It's a matter of **resolution ! Binning should match the information content in the data:…Trial & Error**

How to draw a good histogram

When drawing a histogram, the choice of interval width must be made carefully because it can affect the conclusions. For example, Figure 2.2-5 shows three different histograms that depict the body mass of 228 female sockeye salmon (*Oncorhynchus nerka*) from Pick Creek, Alaska, in 1996 (Hendry et al. 1999). The leftmost histogram of Figure 2.2-5 was drawn using a narrow interval width. The result is a somewhat bumpy frequency distribution that suggests the existence of two or even more peaks. The rightmost histogram uses a wide interval. The result is a smoother frequency distribution that masks the second of the two dominant peaks. The middle histogram uses an intermediate interval that shows two distinct body-size groups. The fluctuations from interval to interval within size groups are less noticeable.

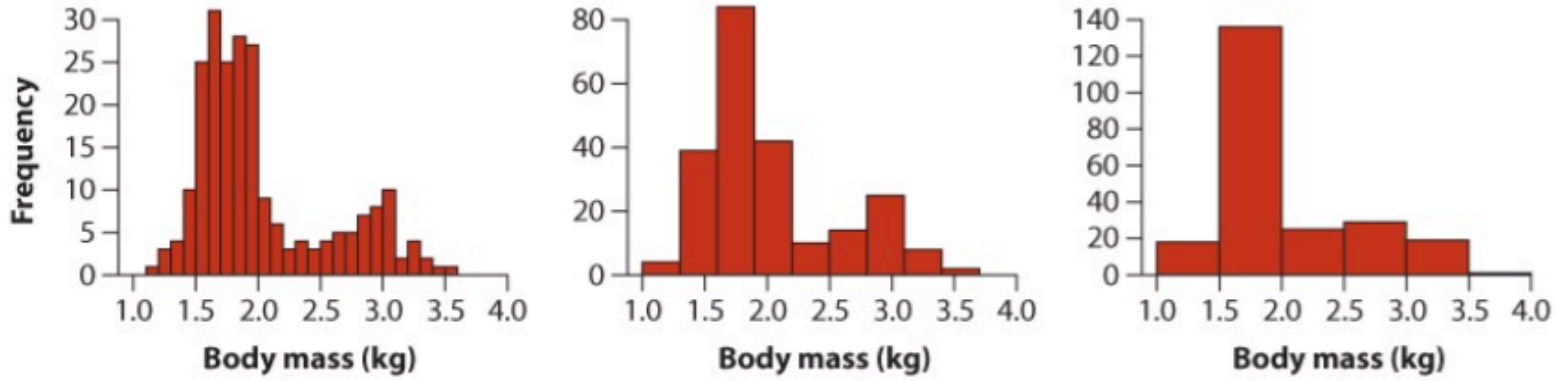


Figure 2.2-5
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

FIGURE 2.2-5 Body mass of 228 female sockeye salmon sampled from Pick Creek in Alaska (Hendry et al. 1999). The same data are shown in each case, but the interval widths are different: 0.1 kg (*left*), 0.3 kg (*middle*), and 0.5 kg (*right*).

The sampling distribution of an estimate

Estimation is the process of inferring a population parameter from sample data. The value of an estimate calculated from data is almost never exactly the same as the value of the population parameter being estimated, because sampling is influenced by chance. The crucial question is, "In the face of chance, how much can we trust an estimate?" In other words, what is its *precision*? To answer this question, we need to know something about how the sampling process might affect the estimates we get. We use the **sampling distribution** of the estimate, which is the probability distribution of all the values for an estimate that we *might* have obtained when we sampled the population. We illustrate the concept of a sampling distribution using samples from a known population, the genes of the human genome.

EXAMPLE 4.1 The length of human genes

The international Human Genome Project was the largest coordinated research effort in the history of biology. It yielded the DNA sequence of all 23 human chromosomes, each consisting of millions of nucleotides chained end to end.[2] These encode the genes whose products—RNA and proteins—shape the growth and development of each individual.

We obtained the lengths of all 20,290 known and predicted genes of the published genome sequence (Hubbard et al. 2005).[3] The length of a gene refers to the total number of nucleotides comprising the coding regions. The frequency distribution of gene lengths in the population of genes is shown in Figure 4.1-1. The figure includes only genes up to 15,000 nucleotides long; in addition, there are 26 longer genes. [4]

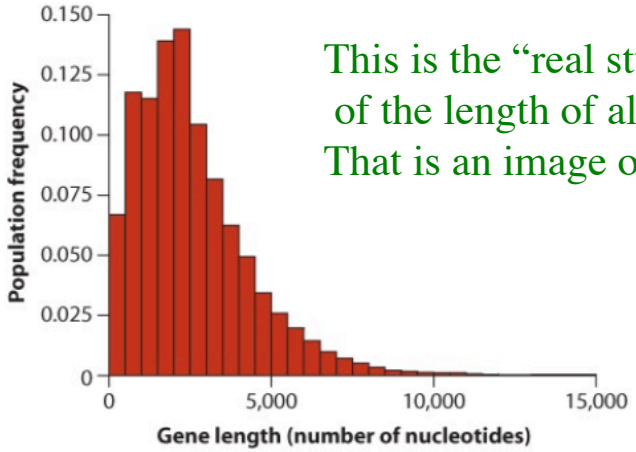


Figure 4.1-1
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

This is the "real stuff": an exaustive representation of the length of all recognized human genes. That is an image of the population of human genes

FIGURE 4.1-1 Distribution of gene lengths in the known human genome. The graph is truncated at 15,000 nucleotides; 26 larger genes are too rare to be visible in this histogram.

The histogram in Figure 4.1-1 is like those we have seen before, except that it shows the distribution of lengths in the *population* of genes, not simply those in a *sample* of genes.

Important remarks

Because it is the population distribution, the relative frequency of genes of a given length interval in Figure 4.1-1 represents the *probability* of obtaining a gene of that length when sampling a single gene at random. The probability distribution of gene lengths is positively skewed, having a long tail extending to the right.

The population mean and standard deviation of gene length in the human genome are listed in Table 4.1-1. These quantities are referred to as *parameters* because they are quantities that describe the population.

TABLE 4.1-1 Population mean and standard deviation of gene length in the known human genome.

Name	Parameter	Value (nucleotides)
Mean	μ	2622.0
Standard deviation	σ	2036.9

In real life, we would not usually know the parameter values of the study population, but in this case we do. We'll take advantage of this to illustrate the process of sampling.

HINT: relative frequency --> probability

Estimating mean gene length with a random sample

To begin, we collected a single random sample of $n = 100$ genes from the known human genome.[5] A histogram of the lengths of the resulting sample of genes is shown in Figure 4.1-2.

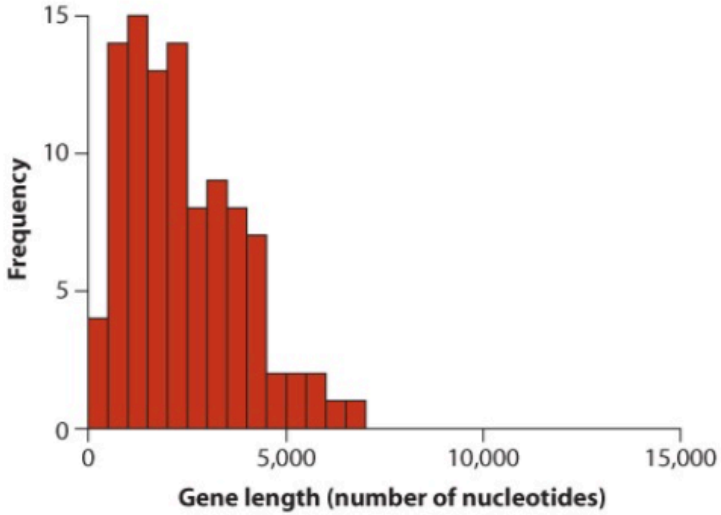


Figure 4.1-2
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

FIGURE 4.1-2 Frequency distribution of gene lengths in a unique random sample of $n = 100$ genes from the human genome.

The frequency distribution of the random sample (Figure 4.1-2) is not an exact replica of the population distribution (Figure 4.1-1), because of chance. The two distributions nevertheless share important features, including approximate location, spread, and shape. For example, the sample frequency distribution is skewed to the right like the true population distribution.

The sample mean and standard deviation of gene length from the sample of 100 genes are listed in Table 4.1-2. How close are these estimates to the population mean and standard

Further remarks

The sample mean and standard deviation of gene length from the sample of 100 genes are listed in Table 4.1-2. How close are these estimates to the population mean and standard deviation listed in Table 4.1-1? The sample mean is $\bar{Y}$=2411.8, $\bar{Y} = 2411.8$, which is about 200 nucleotides shorter than the true value, the population mean of μ = 2622.0. The sample standard deviation (s = 1463.5) is also different from the standard deviation of gene length in the population (σ = 2036.9). We shouldn't be surprised that the sample estimates differ from the parameter (population) values. Such differences are virtually inevitable because of chance in the random sampling process.

TABLE 4.1-2 Mean and standard deviation of gene length Y in our unique random sample of n = 100 genes from the human genome.

Name	Statistic	Sample value (number of nucleotides)
Mean	$\bar{Y}$	2411.8
Standard deviation	s	1463.5

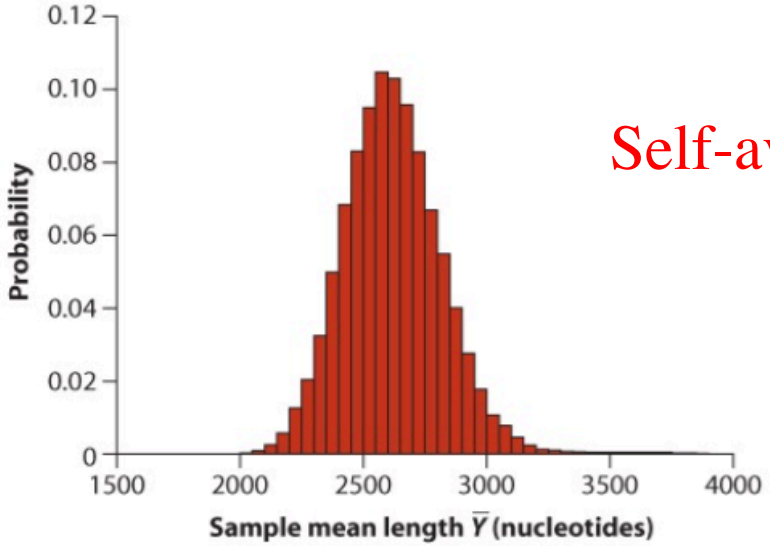
The sampling distribution of $\bar{Y}$

We obtained $\bar{Y}=2411.8$ nucleotides in our single sample, but by chance we might have obtained a different value. When we took a second random sample of 100 genes, we found $\bar{Y}=2643.5$. Each new sample will usually generate a different estimate of the same parameter. If we were able to repeat this sampling an infinite number of times, we could create the probability distribution of our estimate. The probability distribution of values we might obtain for an estimate make up the estimate's **sampling distribution.**

> The ***sampling distribution*** is the probability distribution of all values for an estimate that we might obtain when we sample a population.

The sampling distribution represents the "population" of values for an estimate. It is not a real population, like the squirrels in Muir Woods or all the retirees basking in the Florida sunshine. Rather, the sampling distribution is an imaginary population of values for an estimate. Taking a random sample of *n* observations from a population and calculating $\bar{Y}$ is equivalent to randomly sampling a *single* value of $\bar{Y}$ from its sampling distribution.

To visualize the sampling distribution for mean gene length, we used the computer to take a vast number of random samples of $n = 100$ genes from the human genome. We calculated the sample mean $\bar{Y}$ each time. The resulting histogram in Figure 4.1-3 shows the values of $\bar{Y}$ that might be obtained when randomly sampling 100 genes, together with their probabilities.



Self-averaging !

Figure 4.1-3
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

FIGURE 4.1-3 The sampling distribution of mean gene length, Y, when $n = 100$. Note the change in scale from Figure 4.1-2.

Figure 4.1-3 makes plain that, although the population mean μ is a constant (2622.0), its estimate $\bar{Y}$ is a variable. Each new sample yields a different $\bar{Y}$ value from the one before. We don't ever see the sampling distribution of $\bar{Y}$ because ordinarily we have only one sample, and therefore only one $\bar{Y}$. Notice that the sampling distribution for $\bar{Y}$ is centered exactly on the true mean, μ. This means that $\bar{Y}$ is an unbiased estimate of μ.[6]

The spread of the sampling distribution of an estimate depends on the sample size. The sampling distribution of $\bar{Y}$ based on $n = 100$ is narrower than that based on $n = 20$, and that based on $n = 500$ is narrower still (Figure 4.1-4). The larger the sample size, the narrower the sampling distribution. And the narrower the sampling distribution, the more precise the estimate. Thus, larger samples are desirable whenever possible because they yield more precise estimates. The same is true for the sampling distributions of estimates of other population quantities, not just $\bar{Y}$.

Study materials

- Rossner [R] chapter 2
- Whitlock&Sluter [**WS**] chapter 2
- RDL lecture n. 3 in Data Analysis Moodle Course
- Read Naomi Altman's article in Nature Methods *Importance of being uncertain* and write in word, latex… a 1 page essay with a resume of it.

From: [**JHZ**]Jerrold H Zar - Biostatistical Analysis_ Pearson New International Edition-Pearson (2014).

From:[**R**]Bernard Rosner - Fundamentals of Biostatistics-Brooks Cole (2015)

From:[**WS**] M.C. Whitlock and D. Schluter - The Analysis of Biological Data-W. H. Freeman and Company (2015).

SEE YOU NEXT MONDAY!

From Graziani2009