Descriptive statistics: central tendency and dispersion

Descriptive statistics

Descriptive statistics, or summary statistics, are quantities that capture important features of frequency distributions.

Whereas graphs reveal shapes and patterns in the data, descriptive statistics provide hard numbers.

The most important descriptive statistic for numerical data are those measuring the **location** of a frequency distribution and its **spread**.

The location tells us something about the average or typical individual—where the observations are centered.

The spread tells us how variable the measurements are from individual to individual—how widely scattered the observations are around the center.

The **proportion** is the most important descriptive statistic for a categorical variable, measuring the fraction of observations in a given category.

Descriptive statistics: central tendency and dispersion

Measures of the central tendency of the data (the mean, median, mode) and the dispersion and variability of the data.

1. Samples and populations

2. Measures of central tendency

Measures of central tendency		
Statistic	**Function**	**R Package**
Mean	mean()	base
Median	median()	base
Skewness	skewness()	e1071
Kurtosis	kurtosis()	e1071

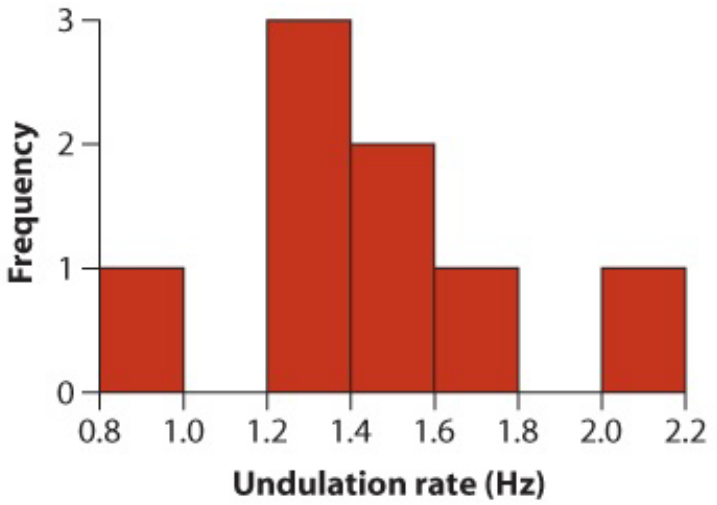
Descriptive statistics: central tendency and dispersion

2.1. The arithmetic mean

The arithmetic mean is the most common metric to describe the location of a frequency distribution. The sample *mean* is the arithmetic average of the data, and it is calculated by summing all the data and dividing it by the sample size, *n*. The mean, $\bar{x}$, is calculated thus:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

For data: 0.9, 1.4, 1.2, 1.2, 1.2, 2.0, 1.4, 1.6; the mean is 1.3625.



Descriptive statistics: central tendency and dispersion

2.1. The arithmetic mean

Let us take that the following table which shows the number of patients admitted in a hospital.

Week	Number of patients (X)
1	1
2	2
3	3
4	1
5	4
6	4
7	1

Then the number of patients admitted per week on an average is

$$Mean(X) = \mu = \frac{1+2+3+1+4+4+1}{7}$$

Descriptive statistics: central tendency and dispersion

2.1. The arithmetic mean

But, let us try to write this mean in a different way

$$\mu = \frac{1 \times 3 + 2 \times 1 + 3 \times 1 + 4 \times 2}{7}$$

Or the same can be written as

$$\mu = 1 \times \frac{3}{7} + 2 \times \frac{1}{7} + 3 \times \frac{1}{7} + 4 \times \frac{2}{7}$$

which is actually the value of random variable X multiplied by its frequency. This can be generalized as

$$\mu = \sum_{i=1}^{4} x_i \times f_i$$

Thus, for a discrete random variable X such that a ≤ X ≤ b and having a PMF: P(X=k), the expected value or mean of X

$$E(X) = \sum_{x_i=a}^{b} x_i \times P(X = x_i)$$

Descriptive statistics: central tendency and dispersion

2.1. The arithmetic mean

Similarly, if we wish to calculate the mean of square of the random variable, we can write

$$Mean(x^2) = \mu = \frac{1^2 + 2^2 + 3^2 + 1^2 + 4^2 + 4^2 + 1^2}{7}$$

Or the same can be written as

$$\mu = 1^2 \times \frac{3}{7} + 2^2 \times \frac{1}{7} + 3^2 \times \frac{1}{7} + 4^2 \times \frac{2}{7}$$

which is actually the value of random variable X multiplied by its frequency. This can be generalized as

$$\mu = \sum_{i=1}^{4} {x_i}^2 \times f_i$$

Thus, for a discrete random variable X such that a ≤ X ≤ b and having a PMF: P(X=k), the expected value or mean of X^2

$$E(X^2) = \sum_{x_i=a}^{b} {x_i}^2 \times P(X = x_i)$$

Descriptive statistics: central tendency and dispersion

2.1. The arithmetic mean

Thus, for a discrete random variable X such that a ≤ X ≤ b and having a PMF: P(X=k), the expected value or mean of X^m or the m-th moment of X:

$$\mu_m = E(X^m) = \sum_{x_i=a}^{b} {x_i}^m \times P(X = x_i)$$

Similarly, for a continuous random variable X such that a ≤ X ≤ b and having a PDF: $f_X(x)$, the expected value or mean of X^m or the m-th moment of X:

$$\mu_m = E(X^m) = \int_a^b x^m \times f_X(x)\,\mathrm{d}x$$

Descriptive statistics: central tendency and dispersion

2.1. The arithmetic mean

The mean is quite sensitive to the presence of outliers or extreme values in the data, and it is advised that its use be reserved for normally distributed data from which the extremes/outliers have been removed.

When extreme values are indeed part of our data and not simply ‘noise,’ then we have to resort to a different measure of central tendency: the median.

Descriptive statistics: central tendency and dispersion

2.2. The median

After the sample mean, the *median* is the next most common metric used to describe the location of a frequency distribution.

The median is therefore the value that separates the lower half of the sample data from the upper half. In normally distributed continuous data, the median is equal to the mean.

The **median** is the middle observation in a set of data, the measurement that partitions the ordered measurements into two halves.

If the number of observations is odd, the median lies on the position $\frac{n+1}{2}$th position of the sorted data.

For example:
For the data: 9, 14, 12, 13, 12, 20, 14, 16, 23. First sort it in the increasing order: 9, 12, 12, 13, 14, 14, 16, 20, 23. Since the number of data is 9, the median lies on the 5^{th} position in the sorted order of the data, i.e. 14.

Descriptive statistics: central tendency and dispersion

2.2. The median

If the number of observations is even, then the median is the average of the middle pair ($\frac{\frac{n}{2}\text{th} + \left(\frac{n}{2}+1\right)\text{th}}{2}$).

For example:
For the data: 9, 14, 12, 12, 12, 20, 14, 16. First sort it in the increasing order: 9, 12, 12, 12, 14, 14, 16, 20. Since the number of data is 8, the median is the average of 4^{th} and 5^{th} position in the sorted order of the data, i.e. (12+14)/2 = 13.

Descriptive statistics: central tendency and dispersion

2.2. The median

The advantage of the median over the mean is that it is unaffected (i.e. not skewed) by extreme values or outliers, and it gives an idea of the typical value of the sample.

The median is also used to provide a robust description of non-parametric data.

The median is often displayed in a box plot alongside the span between the first and third quartiles, or *interquartile range,* another measure of the spread of the distribution.

Descriptive statistics: central tendency and dispersion

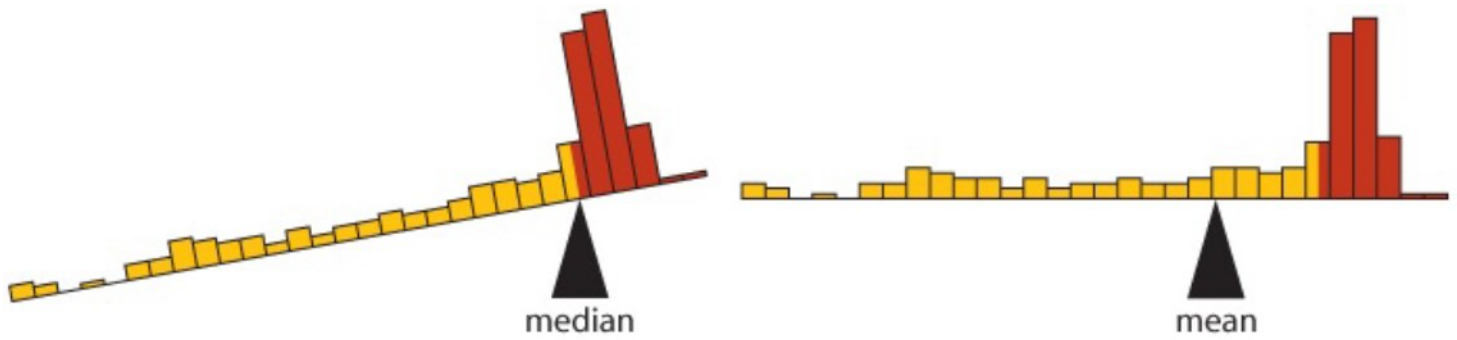
Mean versus median

Descriptive statistics for the number of lateral plates of the three genotypes of three spine sticklebacks

Genotype	n	Mean	Median	Standard deviation	Interquartile range
MM	82	62.8	63	3.4	2
Mm	174	50.4	59	15.1	21
mm	88	11.7	11	3.6	3

Why are the median and mean different from one another when the distribution is asymmetric?

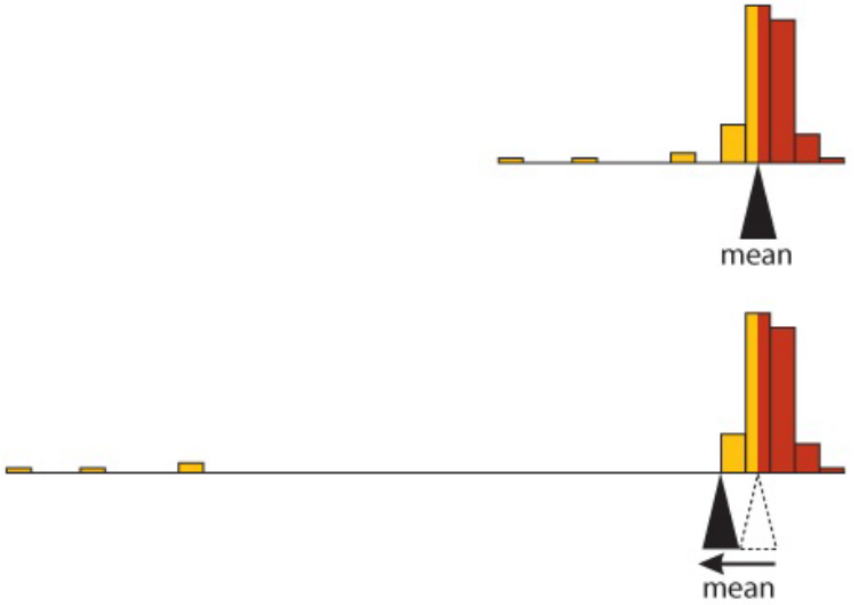
The answer is that the median is the middle measurement of a distribution, whereas the mean is the “center of gravity.”



Descriptive statistics: central tendency and dispersion

Mean versus median

The mean is sensitive to extreme observations.



Median and mean measure different aspects of the location of a distribution. The ***median*** is the middle value of the data, whereas the *mean* is its center of gravity. Thus, the mean is displaced from the location of the "typical" measurement when the frequency distribution is strongly skewed, particularly when there are extreme observations. The mean is still useful as a description of the data as a whole, but it no longer indicates where most of the observations are located. The median is less sensitive to extreme observations, and hence the median is the more informative descriptor of the typical observation in such instances. However, the mean has better mathematical properties, and it is easier to calculate measures of the reliability of estimates of the mean.

Descriptive statistics: central tendency and dispersion

2.3. Skewness

Skewness is a measure of symmetry, and it is best understood by understanding the location of the median relative to the mean.

A negative skewness indicates that the mean of the data is less than their median; the data distribution is left-skewed.

A positive skewness results from data that have a mean that is larger than their median; these data have a right-skewed distribution.

Descriptive statistics: central tendency and dispersion

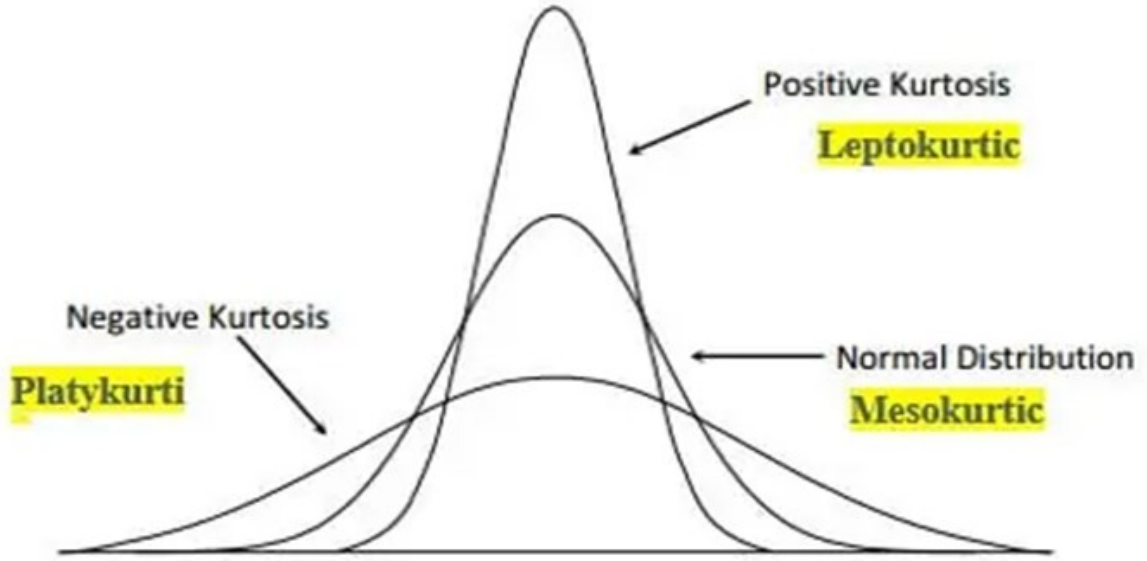
2.4. Kurtosis

Kurtosis describes the tail shape of the data's distribution.

A normal distribution has zero kurtosis and thus the standard tail shape (mesokurtic).

Negative kurtosis indicates data with a thin-tailed (platykurtic) distribution.

Positive kurtosis indicates a fat-tailed distribution (leptokurtic).



Descriptive statistics: central tendency and dispersion

3. Measures of variation and spread

Since the mean or median does not tell us everything there is to know about data, we will also have to determine some statistics that inform us about the variation (or spread or dispersal or inertia) around the central/mean value.

Measures of variation and spread	
Statistic	**Function**
Variance	var()
Standard deviation	sd()
Minimum	min()
Maximum	max()
Range	range()
Quantile	quantile()
Covariance	cov()
Correlation	cor()

Descriptive statistics: central tendency and dispersion

3. Measures of variation and spread

Given the patient data, the mean of X is 2.2857. By knowing this mean may not give a complete picture. So, we also need to know the variance of the data. That is how the number of patients per week are varying.

However, this also may not be very useful for the comparison of two data set e.g. patients of cancer and diabetes. Because both the data will have their means and variances around their means.

Thus, to nullify the effects of means on variances, we need to center the data around the mean s/t mean of the centered data is zero.

Week	Number of patients (X)	Centered data X-E(X)
1	1	1-2.2857
2	2	2-2.2857
3	3	3-2.2857
4	1	1-2.2857
5	4	4-2.2857
6	4	4-2.2857
7	1	1-2.2857

Now, the data is centered around the mean. So, any two data sets can be compared by comparing variances, as there is no variation due to means.

Descriptive statistics: central tendency and dispersion

3. Measures of variation and spread

The expected value or mean of the centered data (X-E(X)) of a random variable X such that a ≤ X ≤ b and having a PMF: P(X=k), is called the m^{th} Central Moment of X:

$$E([X - E(X)]^m) = \sum_{x_i=a}^{b} \left(\mathrm{x_i} - E(X)\right)^m \times P(X = x_i)$$

Thus, by definition, the variance is the 2^{nd} Central Moment of a random variable X

$$var(X) = E([X - E(X)]^2)$$

$$Var(X) = \sum_{x_i=a}^{b} \left(X - E(X)\right)^2 \times P(X = x_i)$$

$$Var(X) = E(X^2) - E(X)^2$$

Thus, variation is a measure of dispersion of a random variable around the mean. In other words, variance is mean or average of squared deviation. It is always a positive value.

Descriptive statistics: central tendency and dispersion

3. Measures of variation and spread

The standard deviation can be defined in terms of variance. And, this is always positive.

$$\sigma_X = \sqrt{var(X)}$$

Few rules of mean and variance

$$E(aX + b) = aE(X) + b$$

$$Var(aX + b) = a^2 Var(X)$$

Where a and b are constants.

If X and Y are independent random variables, then

$$E(XY) = E(X)E(Y)$$

$$Var(XY) = E(X^2)E(Y^2) - E(X)^2 E(Y)^2$$

Descriptive statistics: central tendency and dispersion

3. Measures of variation and spread

Let us take that the following table shows the number of patients admitted in a hospital.

Week	Number of patients (X)	Deviation from mean $X - \bar{X}$	$[X - \bar{X}]^2$
1	1	1-2.2857	1.653
2	2	2-2.2857	0.081
3	3	3-2.2857	0.510
4	1	1-2.2857	1.653
5	4	4-2.2857	2.938
6	4	4-2.2857	2.938
7	1	1-2.2857	1.653

$$Var(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n} = 11.426/7 = 1.632$$

Descriptive statistics: central tendency and dispersion

3. Measures of variation and spread

Variance can also be calculated as

$$Var(X) = E([X - E(X)]^2) \qquad Var(X) = E(X^2) - E(X)^2 \qquad Var(X) = \frac{\sum_{i=1}^{n} X_i^2}{n} - \bar{X}^2$$

Week	Number of patients (X)	X^2
1	1	1
2	2	4
3	3	9
4	1	1
5	4	16
6	4	16
7	1	1

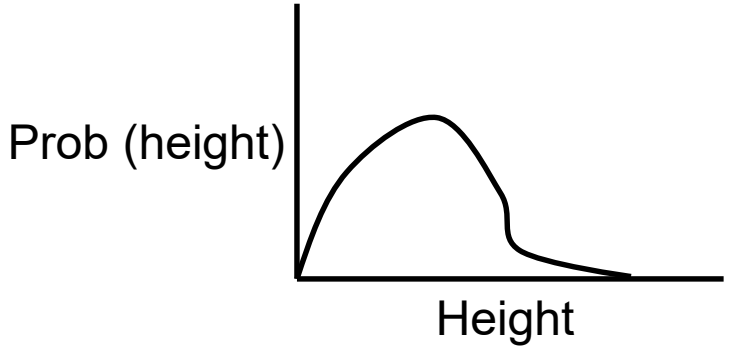
Var (X) = 48/7 – $(2.2857)^2$ = 6.8571 – 5.2244 = 1.632

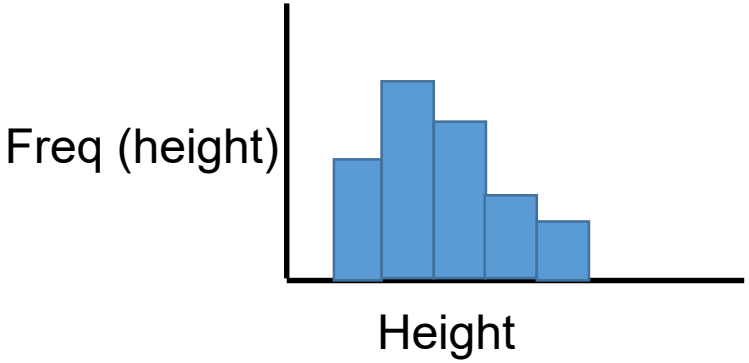
Descriptive statistics: central tendency and dispersion

3. Measures of variation and spread

However, there is a problem. That is, the way the variance has been calculated may not represent the real value of the population, i.e. the variance of the population. The reason is that we sample data from population to know the behavior (mean, variance, etc.) of the population. However, we do not know the probability distribution of population.

The probability distribution of population can be guessed by plotting the frequency distribution of the sampled data. However, due to sampling, there will be some variation in the mean and variance of the population as compared to the sampled data.





Descriptive statistics: central tendency and dispersion

3.1. The variance and standard deviation

So, actually we use the following formula to calculate variance of X

$$Var(X) = E([X - E(X)]^2)$$

Or by replacing E(X) by $\bar{X}$.

$$Var(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

Since we do not know the exact value of mean of population, i.e. E(X), we are replacing with the sample mean. Thus, each time, we are deviating from the exact value by making this adjustment. Thus, we are under estimating the value of population variance. To correct this, we multiply by (n/n-1) i.e.

$$Var(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n} \text{ x } (n/n\text{-}1) \quad \Rightarrow \quad Var(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

This is called Bessel's correction.

Descriptive statistics: central tendency and dispersion

3.1. The variance and standard deviation

The variance and standard deviation are examples of interval estimates.

The **sample variance**, S^2, may be calculated according to the following formula:

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

The **standard deviation** is a commonly used measure of the spread of a distribution. It measures how far from the mean the observations typically are. The standard deviation is large if most observations are far from the mean, and it is small if most measurements lie close to the mean.

The standard deviation is a more intuitive measure of the spread of a distribution (in part because it has the same units as the variable itself), but the variance has mathematical properties that make it useful sometimes as well.

To get the standard deviation, S, we take the square root of the variance, i.e.

$$s = \sqrt{s^2}$$

Descriptive statistics: central tendency and dispersion

3.1. The variance and standard deviation

So, let us calculate the variance (now it is called sample variance)

Week	Number of patients (X)	Deviation from mean $X - \bar{X}$	$[X - \bar{X}]^2$
1	1	1-2.2857	1.653
2	2	2-2.2857	0.081
3	3	3-2.2857	0.510
4	1	1-2.2857	1.653
5	4	4-2.2857	2.938
6	4	4-2.2857	2.938
7	1	1-2.2857	1.653

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$ = 11.426/6 = 1.904

Descriptive statistics: central tendency and dispersion

3.1. The variance and standard deviation

The interpretation of the concepts of mean and median are fairly straight forward and intuitive. Not so for the measures of variance.

What does S represent?

Firstly, the unit of measurement of S is the same as that of $\bar{x}$ (but the variance does not share this characteristic). If temperature is measured in ºC, then S also takes a unit of ºC.

Since S measures the dispersion around the mean, we write it as $\bar{x}$ ± S (note that often the mean and standard deviation are written with the letters $\mu \pm \sigma$).

The smaller S, the closer the sample data are to $\bar{x}$, and the larger the value is the further away they will spread out from $\bar{x}$. So, it tells us about the proportion of observations above and below $\bar{x}$.

Descriptive statistics: central tendency and dispersion

3.1. The variance and standard deviation

But what proportion?

We invoke the 68-95-99.7 rule: ~68% of the population (as represented by a random sample of *n* observations taken from the population) falls within 1S of $\bar{x}$ (i.e. ~34% below $\bar{x}$ and ~34% above $\bar{x}$), ~95% of the population falls within 2S, and ~99.7% falls within 3S.

Like the mean, S is affected by extreme values and outliers, so before we attach S as a summary statistic to describe some data, we need to ensure that the data are in fact normally distributed.

Descriptive statistics: central tendency and dispersion

3.1.1. The covariance

Let us take two random variable X and Y. How to measure the joint variability of X and Y.

$$Var(X) = E([X - E(X)]^2)$$

$$Var(X) = E\big[\big(X - E(X)\big) \times E\big(X - E(X)\big)\big]$$

What if X is replaced by in the 2nd term.

$$Cov(X, Y) = E\big[\big(X - E(X)\big) \times E\big(\mathrm{Y} - E(Y)\big)\big]$$

$$\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

Descriptive statistics: central tendency and dispersion

3.1.1. The covariance

X	**Y**	$X - \bar{x}$	$Y - \bar{y}$	$(X - \bar{x})\,(Y - \bar{y})$
x1	y1	x1-$\bar{x}$	y1-$\bar{y}$	(x1-$\bar{x}$)(y1-$\bar{y}$)
x2	y2	x2-$\bar{x}$	y2-$\bar{y}$	(x2-$\bar{x}$)(y2-$\bar{y}$)
x3	y3	x3-$\bar{x}$	y3-$\bar{y}$	(x3-$\bar{x}$)(y3-$\bar{y}$)
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
xi	yi	xi-$\bar{x}$	yi-$\bar{y}$	(xi-$\bar{x}$)(yi-$\bar{y}$)
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
xn	yn	xn-$\bar{x}$	yn-$\bar{y}$	(xn-$\bar{x}$)(yn-$\bar{y}$)

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(yi - \bar{y})}{n-1}$$

This is after Bessel's correction.

If X increases and Y decreases, Cov (X, Y) < 0.

If X increases and Y increases, Cov (X, Y) > 0.

It means that unlike Var (X, Y) which is always positive, Cov (X, Y) can be positive or negative.

If X and Y are independent, then

$$\mathrm{Cov}(X,Y) = E(X)E(Y) - E(X)E(Y) = 0$$

However, if Cov (X, Y) = 0, it does not imply that X and Y are independent.

Descriptive statistics: central tendency and dispersion

3.2. Quantiles

A more forgiving approach (forgiving of the extremes, often called ‘robust’) is to divide the distribution of ordered data into quarters, and find the points below which 25% (0.25, the first quartile), 50% (0.50, the median) and 75% (0.75, the third quartile) of the data are distributed.

These are called *quartiles* (for ‘quarter;’ not to be confused with *quantile*, which is a more general form of the function that can be used to divide the distribution into any arbitrary proportion from 0 to 1).

Descriptive statistics: central tendency and dispersion

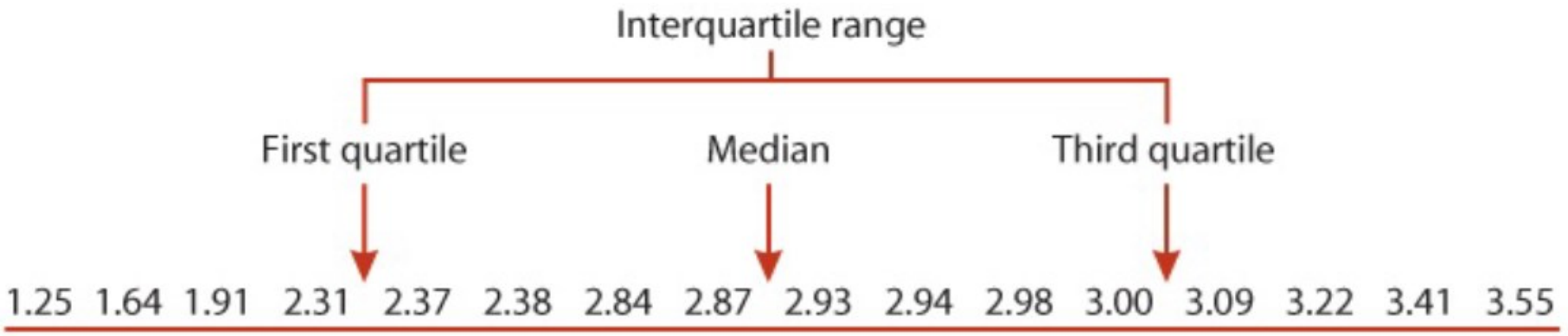
3.2. Quartile and interquartile range

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:

$$IQR = \text{third quartile} - \text{first quartile}.$$

It is the span of the middle 50% of the data.



Descriptive statistics: central tendency and dispersion

3.2. Quartile and interquartile range

The first step in calculating the interquartile range is to compute the first and third quartiles, as follows.

For the **first quartile**, calculate j=0.25n, where *n* is the number of observations. If *j* is an integer then the first quartile is the average of *Y*(*j*) and *Y*(*j*+1): First quartile=(Y(j)+Y(j+1))/2, where *Y*(*j*) is the *j*th sorted observation.

If *j* is not an integer, then convert *j* to an integer by replacing it with the next integer that exceeds it (i.e., round *j* up to the nearest integer). The first quartile is then First quartile=Y(j), where *j* is now the integer you rounded to.

The **third quartile** is computed similarly. Calculate k=0.75n. If *k* is an integer, then the third quartile is the average of *Y*(*k*) and *Y*(*k*+1): Third quartile=(Y(k)+Y(k+1))/2, where *Y*(*k*) is the *k*th sorted observation.

If *k* is not an integer, then convert *k* to an integer by replacing it with the next integer that exceeds it (i.e., round *k* up to the nearest integer). The third quartile is then Third quartile=Y(k), where *k* is the integer you rounded to.

Descriptive statistics: central tendency and dispersion

Cumulative frequency distribution

The median and quartiles are examples of percentiles, or quantiles, of the frequency distribution for a numerical variable. Plotting all the quantiles using the cumulative frequency distribution is another way to compare the shapes and positions of two or more frequency distributions.

Percentiles and quantiles

The *X*th **percentile** of a sample is the value below which *X* percent of the individuals lie. For example, the median, the measurement that splits a frequency distribution into equal halves, is the 50th percentile. Ten percent of the observations lie below the 10th percentile, and the other 90% of observations exceed it. The first and third quartiles are the 25th and 75th percentiles, respectively.

The ***percentile*** of a measurement specifies the percentage of observations less than or equal to it; the remaining observations exceed it. The *quantile* of a measurement specifies the fraction of observations less than or equal to it.

The same information in a percentile is sometimes represented as a **quantile**. This only means that the proportion less than or equal to the given value is represented as a decimal rather than as a percentage. For example, the 10th percentile is the 0.10 quantile, and the median is the 0.50 quantile. Be careful not to mix up the words *quantile* and *quartile* (note the difference in the fourth letters). The first and third quartiles are the 0.25 and 0.75 quantiles.

Descriptive statistics: central tendency and dispersion

Standard deviation versus interquartile range

Because it is calculated from the square of the deviations, the standard deviation is even more sensitive to extreme observations than is the mean. For this reason, the interquartile range is a better indicator of the spread of the main part of a distribution than the standard deviation when the data are strongly skewed to one side or the other, especially when there are extreme observations.

On the other hand, the standard deviation reflects the variation among all of the data points.

Descriptive statistics: central tendency and dispersion

Coefficient of variation

For many traits, standard deviation and mean change together when organisms of different sizes are compared. Elephants have greater mass than mice and also more variability in mass. For many purposes, we care more about the relative variation among individuals. A gain of 10 g for an elephant is inconsequential, but it would double the mass of a mouse.

On the other hand, an elephant that is 10% larger than the elephant mean may have something in common with a mouse that is 10% larger than the mouse mean. For these reasons, it is sometimes useful to express the standard deviation relative to the mean.

The **coefficient of variation** (CV) calculates the standard deviation as a percentage of the mean:

$$CV = \frac{s}{\bar{X}} \times 100\%$$

The ***coefficient of variation*** is the standard deviation expressed as a percentage of the mean. A higher CV means that there is more variability, whereas a lower CV means that individuals are more consistently the same.

Descriptive statistics: central tendency and dispersion

Proportions
The proportion is the most important descriptive statistic for a categorical variable.

Calculating proportions:

$$\hat{p} = \frac{\text{Number in category}}{n},$$

where the numerator is the number of observations in the category of interest, and *n* is the total number of observations in all categories combined

Descriptive statistics: central tendency and dispersion

Rounding means, standard deviations, and other quantities

To avoid rounding errors when carrying out calculations of means, standard deviations, and other descriptive statistics, always retain as many significant digits as your calculator or computer can provide. Intermediate results written down on a page should also retain as many digits as feasible. Final results, however, should be rounded before being presented.

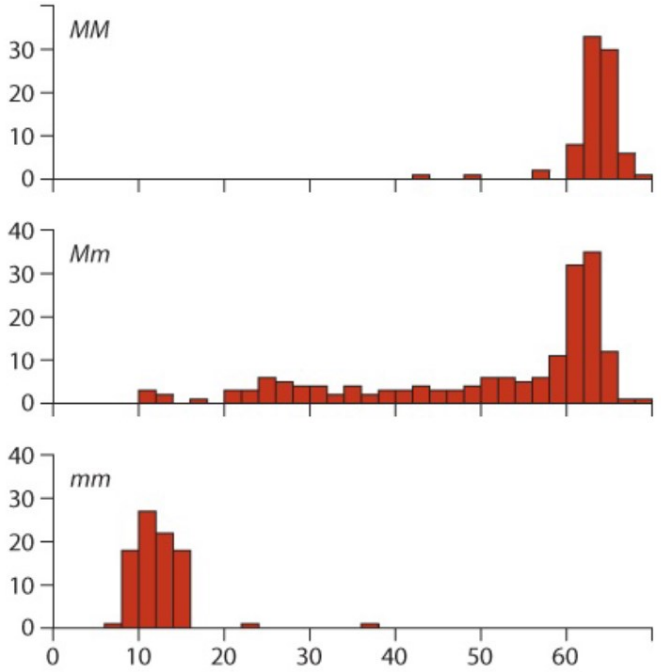
There are no strict rules on the number of significant digits that should be retained when rounding. A common strategy, which we adopt here, is to round descriptive statistics to one decimal place more than the measurements themselves.

Descriptive statistics: central tendency and dispersion

How measures of location and spread compare

Which measure of location, the sample mean or the median, is most revealing about the center of a distribution of measurements? And which measure of spread, the standard deviation or the interquartile range, best describes how widely the observations are scattered about the center?

The answer depends on the shape of the frequency distribution. These alternative measures of location and of spread yield similar information when the frequency distribution is symmetric and unimodal. The mean and standard deviation become less informative than the median and interquartile range when the data are strongly skewed or include extreme observations.



Descriptive statistics: central tendency and dispersion

3.3. The minimum, maximum and range

3.4. Covariance

3.5. Correlation

The correlation coefficient of two matched (paired) variables is equal to their covariance divided by the product of their individual standard deviations. It is a normalized measurement of how linearly related the two variables are.

Thank You