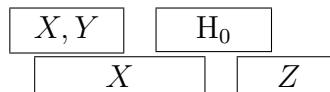


Emmanuel College
MA 200 – Statistics
(linked with KN401)
Course Material

The following is a summary of course material. Blue links lead to more detail; red links lead back.

contents
[preliminaries](#)
[basic concepts](#)
[graphs](#)
[measures of central tendency](#)
[measures of dispersion](#)
[other measures of descriptive statistics](#)
[the Standard Normal distribution](#)
[correlation](#)
[regression and prediction](#)
[hypothesis testing](#)
[examples, proofs, and additional discussion](#)



©2008 Jason Colwell. All rights reserved.

preliminaries

order of operations

summation notation

discrete and continuous scales

basic concepts

“population” The population is the set of things examined to collect the data. The number of these things is called n .

“random variable” A random variable X is an attribute measured for each of the n members of the population.

notation We shall use italic lower-case letters to indicate variables taking a single value (e.g. \bar{x} , s), and italic upper-case letters to denote random variables (which take multiple values) (e.g. \bar{X} , S).

“distribution” A distribution consists of the values, of a random variable X , obtained from the population. It is a collection of n numbers, one for each member of the population. (A number in a distribution may occur more than once.)

examples

“frequency table” Sometimes, a distribution is given, not as a list of values of the random variable X , but in terms of the frequencies of certain values of the random variable (that is, frequencies of the various numbers appearing in the distribution). We often use the notation where the distribution consists of values $X = x_1, x_2, \dots, x_k$, with respective frequencies f_1, f_2, \dots, f_k .

example

graphs

“line graph” This depicts a distribution written as a list, as place in the list (on the horizontal axis) versus value of the random variable Y (on the vertical axis). Often the list is in chronological order.

“bar graph” This depicts values of a discrete random variable X versus frequency in the distribution (or percent of the distribution). In other words, it is a plot of a frequency table.

“histogram” A “histogram” depicts specified intervals of a continuous random variable X versus frequency in the distribution (or percent of the distribution).

“scatterplot” A “scatterplot depicts the value of one random variable X versus the value of another random variable Y for a single population. There is one point for each member of the population, with coordinates (X, Y) .

measures of central tendency

The three measures of central tendency we will consider are the mode, the median, and the mean.

“mode” This is most frequently occurring number in the distribution.

“median” If the distribution is listed from smallest to largest, the median is the number at the “middle” place in the list; that is:

$$\text{median} = \left(\frac{n+1}{2}\right)\text{th number in the list}$$

If n is even, the median will be the average of two entries. For example, if $n = 14$, then $\frac{n+1}{2} = \frac{14+1}{2} = 7.5$, and the median is the average of the 7th and 8th numbers in the list.

“mean” The mean is the sum of all numbers in the distribution, divided by the size of the distribution (which is also the size of the population). This can be written

$$\bar{x} = \frac{\sum X}{n},$$

where X is the random variable and n is the size of the distribution (and of the population).

Suppose the distribution is given by the frequencies of certain values of the random variable (that is, frequencies of the various numbers appearing in the distribution). If the distribution consists of values x_1, x_2, \dots, x_k , with respective frequencies f_1, f_2, \dots, f_k , then the mean of the distribution is

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}.$$

example

linearly transformed random variables

Simpson’s Paradox provides motivation for the following concept.

“weighted mean” Let X_1, X_2, \dots, X_k be random variables, with **“weights”** w_1, w_2, \dots, w_k (that is, numbers such that $0 \leq w_1, w_2, \dots, w_k \leq 1$ and $w_1 + w_2 + \dots + w_k = 1$.) Then the weighted mean of the n random variables is

$$\bar{X}_w = w_1 X_1 + w_2 X_2 + \dots + w_k X_k = \sum_{i=1}^k w_i X_i.$$

The weighted mean is used often to calculate an “average” of things that have different significance, like assignment and test scores within a course. The weighted mean also provides a **resolution of Simpson’s Paradox.**

measures of dispersion

The two measures of dispersion we will consider are the range and the standard deviation.

“range” The range of a distribution is the highest number minus the lowest number.

$$\text{range}_X = \text{highest value of } X - \text{lowest value of } X$$

“variance” The average squared deviation from the mean is called the “variance”. The variance of a distribution is

$$\text{var}_X = \frac{\sum (X - \bar{x})^2}{n},$$

as written for a list, and

$$\text{var}_X = \frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i},$$

as written for a frequency table.

discussion

“standard deviation” The standard deviation of a distribution is the square-root of the variance:

$$s = \sqrt{\text{var}_X}.$$

note As a result of the definition of s , the variance of X is often written

$$s^2.$$

example

linearly transformed random variables

other measures of descriptive statistics

“skew” The “skew” of a distribution is

$$\boxed{\text{skew}_X = \left(\frac{\sum (X - \bar{x})^3}{n} \right) / s^3}, \quad \boxed{\text{skew}_X = \left(\frac{\sum f_i (x_i - \bar{x})^3}{\sum f_i} \right) / s^3},$$

as written for a list and for a frequency table, respectively.

A distribution is said to be **“skewed”** if the skew is > 0 or < 0 . If the distribution is **“positively skewed”**, the values are clustered toward the lower end. If the distribution is **“negatively skewed”**, the values are clustered at the upper end. In a positively skewed distribution, the mean is usually higher than the median. In a negatively skewed distribution, the mean is usually lower than the median.

counter-example

example

linearly transformed random variables

“kurtosis” The “kurtosis” of a distribution is

$$\boxed{\text{kurt}_X = \left(\frac{\sum (X - \bar{x})^4}{n} \right) / s^4}, \quad \boxed{\text{kurt}_X = \left(\frac{\sum f_i (x_i - \bar{x})^4}{\sum f_i} \right) / s^4},$$

as written for a list and for a frequency table, respectively.

The kurtosis is always ≥ 0 . It measures the “peakedness” of the distribution. A distribution is described as **“mesokurtic”** if it has kurtosis 3 (or close to it). A distribution with a low kurtosis is **“leptokurtic”** – bunched up around the mean. A distribution with a high kurtosis is **“platykurtic”** – spread out more over the range (relative to the dispersion).

example

linearly transformed random variables

the Standard Normal distribution

“z-score” Suppose X is a random variable with mean \bar{x} and standard deviation s . Then the random variable

$$Z = \frac{X - \bar{x}}{s}$$

is called the “z-score”. No matter what the mean and standard deviation of X are, the mean of Z is 0 and the standard deviation of Z is 1.

reason

There is a particular shape to which many z-score distributions conform. The reason for this will be discussed later, but we introduce this shape now. **“Standard Normal” random variable** A random variable Z is said to be “Standard Normal” if the graph of its distribution is given by

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-Z^2/2}.$$

A random variable X is said to be **“Normal”** if its z-score $Z = \frac{X - \bar{x}}{s}$ is a Standard Normal random variable.

facts Suppose Z is a Standard Normal random variable. Then:

- the total area under the curve giving the distribution of Z is 1. (Think of the region under the curve as representing 100% of the population.)
- by the symmetry of the Standard Normal distribution,

$$P(Z \leq 0) = P(Z \geq 0) = 0.5.$$

Equivalently, the median of Z is 0.

- $\bar{z} = 0$. (This, also, results from the symmetry of the Standard Normal distribution.)
- $s_Z = 1$.

- $\text{skew}_Z = 0$. (This is because the distribution is symmetric about the mean.)
- $\text{kurt}_Z = 3$. (This is why 3 is considered the “normal amount” of kurtosis, and a distribution with kurtosis 3 is called “mesokurtic”.)

“percentile rank” The “percentile rank” of a value of the random variable is the percentage of the population that has values of the random variable below that number. That is, if x has percentile rank n , this means that $X < x$ for $n\%$ of the population, and $X > x$ for $(100 - n)\%$ of the population. In this course, we will discuss finding percentile ranks only for Normal distributions.

“percentile” The “ n th percentile” is the number such that $n\%$ of the population has values of the random variable below that number, and $(100 - n)\%$ of the population has values above that number. That is, $n\%$ of the distribution is below the n th percentile, and $(100 - n)\%$ of the distribution is above it. In this course, we will discuss finding percentiles (other than the 50th percentile – the median) only for Normal distributions.

the usefulness of the Standard Normal distribution

proportions for the Standard Normal distribution This is the table giving, for $z > 0$ a value of the Standard Normal variable Z , the proportion

$$P(Z > z)$$

of the population having values of Z greater than z (the area under the Standard Normal curve to the right of z).

specific cases of proportion of population

specific cases of percentile

example

example

correlation

Sometimes we wish to consider, for a given population, two different random variables X and Y . For example, X might be the height of residents of Franklin Springs, and Y might be the weights of residents of Franklin Springs. Note that the population is the same for X and Y .

“scatterplot” To each member of the population is associated a pair (x, y) , where $X = x$ and $Y = y$ for that member. The set of all such pairs, depicted on the (X, Y) -plane, is called a “scatterplot”.

“covariance” Suppose X and Y are two random variables for the same population. The covariance of X and Y is defined to be

$$\text{cov}_{X,Y} = \frac{\sum(X - \bar{x})(Y - \bar{y})}{n}.$$

linearly transformed random variables, etc.

note

“correlation coefficient” Suppose X and Y are two random variables for the same population. The “correlation coefficient” of X and Y is

$$\text{corr}_{X,Y} = \frac{\text{cov}_{X,Y}}{s_X s_Y}.$$

meaning

example

linearly transformed random variables, etc.

correlation coefficient and z-score

limitations One must be aware of the limitations of what one can conclude from the correlation coefficient.

regression and prediction

linear regression Ideally, if all the points in a scatterplot lay on a straight line, we could find the equation $y = mx + b$ for the line, and thus have a precise description for the relationship between X and Y . (That is, $Y = mX + b$.)

In most situations, though, the points in a scatterplot do not lie on a line. We may still wish to find the best linear approximation for the relationship between X and Y . That is, we want to find the line such that the points of the scatterplot are as close to it as possible. Put another way, we want to find m and b such that the random variable $Y' = mX + b$ has values as close as possible to the values of Y .

What do we mean, “as close as possible”? Precisely, we want to minimize

$$\sum(Y' - Y)^2.$$

That is, we want to make the sum of the squares of the differences between the values of Y and Y' , as small as possible. If $Y' = mX + b$ is found so that $\sum(Y' - Y)^2$ is as small as possible, then the line $y = mx + b$ is called the “**regression line of Y on X** ”. In other words, the regression line is the line such that the sum of the squares of the vertical distances from the the points of the scatterplot to the line, is minimized.

Suppose that we are interested in finding values of Y for a population, but that it is much easier to measure X . The regression line of Y on X (based on past observation) would allow us to estimate values of Y based on values of X . We would use the estimation $Y \approx Y' = mX + b$.

calculation *This provides justification for the formula for the regression line, but does not need to be memorized.*

formula for the regression line Suppose X and Y are random variables for the same population. The regression line of Y on X is given by

$$y - \bar{y} = \text{corr}_{X,Y} \frac{s_Y}{s_X} (x - \bar{x}).$$

example

hypothesis testing

Suppose X is a random variable for a large population – too large for us to find the value of X for every member of the population.

“null hypothesis” The “null hypothesis” is a statement

$$\boxed{H_0 : \bar{x} = \bar{x}_0}$$

that the mean of X has a particular value \bar{x}_0 . We want to try to provide evidence against the null hypothesis.

note

“sample” A sample consists of n randomly selected individuals from the population.

“sample mean” The “sample mean” has essentially the same formula as the population mean. It is

$$\boxed{\bar{X} = \frac{\sum X}{n}}, \quad \boxed{\bar{X} = \frac{\sum f_i x_i}{\sum f_i}},$$

as written for a list and for a frequency table, respectively. It is the mean of all the values of X for the members of the sample. The sample mean is itself a random variable.

note

note

how to reject H_0 We gather data by taking a sample of the population, and provide evidence against H_0 as follows.

We suppose that H_0 is true. We observe that the sample mean \bar{X} for our sample differs from the value \bar{x}_0 claimed by H_0 .

Then we show (under the assumption that H_0 is true) that we were unlikely to get a sample whose value of \bar{X} was as far from \bar{x}_0 as ours was.

But since we did in fact did get a sample with this value of \bar{X} , the evidence suggests that our supposition H_0 was false.

“confidence” and “significance” We have in mind how “extreme” (that is – how much \bar{X} differs from \bar{x}_0) our sample has to be, for us to reject H_0 . This is specified by a number α (the Greek letter “alpha”), called the “level of significance”. The number α is between 0 and 1, normally close to 0. Common values for α are 0.1 (10%), 0.05 (5%), and 0.01 (1%).

If we find (assuming H_0) that the probability of getting a sample as extreme as the one we did is less than α , then we say that we can reject H_0 **“at the α significance level”**. Another way of saying this is that we reject H_0 **“with $1 - \alpha$ confidence”**. (For example, if $\alpha = 0.05$ (5%), we would reject the null hypothesis H_0 “with 95% confidence”.)

Because of the following remarkable fact, we assume that the random variable \bar{X} is Normal.

Central Limit Theorem

statistical inference about \bar{x} when s is known Specifically, we suppose that H_0 is true, and that \bar{x} has a particular value \bar{x}_0 . Since \bar{X} is assumed to be normal, we use the z-score of \bar{X} – a Standard Normal random variable – to measure how the sample differs from what the null hypothesis H_0 would predict.

$$Z = \frac{\bar{X} - \bar{x}_0}{s/\sqrt{n}}$$

Then we calculate the probability of obtaining a value of Z equal to (or more extreme than) the one we actually obtained from our sample. If this probability is $\leq \alpha$, then we can assert the alternative hypothesis with a confidence of $1 - \alpha$.

example (“left-tailed” test)

example (“right-tailed” test)

“one-tailed”, “two-tailed”

example (two-tailed test)

critical values for the Standard Normal distribution

“sample variance”

“sample standard deviation” The “sample standard deviation” of X is

$$S = \sqrt{\frac{\sum(X - \bar{X})^2}{n - 1}}, \quad S = \sqrt{\frac{\sum f_i(x_i - \bar{X})^2}{\sum f_i - 1}},$$

as written for a list and for a frequency table, respectively.

statistical inference about \bar{x} when s is unknown The same procedure is used, except that we cannot calculate the z-score because we do not know s , but only S . We perform the same calculation that we would use to find Z , except that S is used in place of s . The resulting number is called the **t-score**:

$$T = \frac{\bar{X} - \bar{x}_0}{S/\sqrt{n}}$$

Since S is a random variable instead of a constant, T is not Standard Normal like Z . The t-score T has a distribution known as a “**Student distribution**”. There is a different Student distribution for each value of the parameter $df = n - 1$, the “**degrees of freedom**”.

critical values for the Student distribution

example (left-tailed test)

example (right-tailed test)

example (two-tailed test)

t-scores vs. z-scores

It is important to understand hypothesis testing from both the abstract and the procedural points of view. We summarize each to conclude the section.

abstract summary of hypothesis testing

procedural summary of hypothesis testing

©2008 Jason Colwell. All rights reserved.

examples, proofs, and additional discussion

The following pages contain more detail on various items in the main text.

order of operations Expressions inside parentheses are evaluated first. The default order of operations is: first, exponentiation; second, multiplication and division; third, addition and subtraction.

summation notation The expression

$$\sum X$$

means the sum of all values of X . Exactly what these values are should be clear from the context. Sometimes we wish to specify more precisely the set of numbers to be added. For examples, the expression

$$\sum_{i=1}^5 i^2$$

means

$$1^2 + 2^2 + 3^2 + 4^2 + 5^2,$$

and the expression

$$\sum_{i=2}^4 \sqrt{1+i^2}$$

means

$$\sqrt{1+2^2} + \sqrt{1+3^2} + \sqrt{1+4^2}.$$

If the range of values of i to be used is clear, we may simply write

$$\sum i^2$$

or

$$\sum \sqrt{1+i^2}.$$

discrete and continuous scales A “**discrete**” scale of measurement is one where there are values between which no other values lie. For example, if we are considering the ranking of teams in the American League, then there is no ranking between 2nd and 3rd. (This is a discrete scale.) Or, if we are finding the number of pets owned by residents of Franklin Springs, there is no possible number of pets between 3 and 4.

By contrast, a “**continuous**” scale is one where values might be as close together as we want. For example, if we are measuring the weights of players on American League teams, there is no minimum gap between the measurements (as there is with team rankings). Or, if we are finding the heights of residents of Franklin Springs, there is no limit on how close to each other those measurements might be.

examples

- 1.** Suppose we were discussing the heights of players on the Lady Lions basketball team. Then the population would be the Lady Lions basketball team, the random variable would be height, and the distribution would be a collection of numbers – the measurements of the players’ heights.
 - 2.** Suppose we were discussing the lengths of pet cats in Royston. Then the population would be the set of pet cats in Royston, the random variable would be length, and the distribution would be a collection of numbers – the measurements of the cats’ lengths.
 - 3.** Suppose we were discussing the annual rainfall in Georgia over the past 10 years. Then the population would be the last 10 years, the random variable would be annual rainfall, and the distribution would be a collection of 10 numbers – the measurements of the annual rainfall for the last 10 years in Georgia.
-

example The following frequency table is another way of writing the distribution

$X : -2, -2, -1, -1, -1, 0, 0, 1, 2, 2, 2, 2, 3, 3.$

x_i	-2	-1	0	1	2	3
f_i	2	3	2	1	4	2

note The distribution could just as well have been written

$X : -1, -2, -1, 3, -1, 0, 1, 2, 2, 0, 2, 3, 2, -2.$

The frequency table would be the same.

example Consider this distribution:

$$X : 1, 5, 2, 1, 2, 1$$

It could also be written like this:

x_i	1	2	5
f_i	3	2	1

The mode of X is 1. This can be seen from the list – that 1 occurs more than any other number – or from the frequency table by noticing that 3 is the highest frequency f_i , and that the value x_i of X occurring that many times is 1.

As for the median of X , we calculate

$$\frac{n+1}{2} = \frac{6+1}{2} = 3.5.$$

The median is the “3.5th” number in the distribution (listed in increasing order 1, 1, 1, 2, 2, 5), which is the average of the 3rd and 4th numbers, $\frac{1+2}{2} = 1.5$. (This can also be seen from the frequency table.)

The mean of X is calculated

$$\bar{x} = \frac{\sum X}{n} = \frac{1+5+2+1+2+1}{6} = \frac{12}{6} = 2$$

from the list, or

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{(3)(1) + (2)(2) + (1)(5)}{3+2+1} = \frac{12}{6} = 2$$

from the frequency table.

linearly transformed random variables Suppose X is a random variable with mean \bar{x} , and a, b are constants. Then

$$V = a + X$$

and

$$W = bX$$

are random variables with respective means

$$\bar{v} = a + \bar{x}$$

and

$$\bar{w} = b\bar{x}.$$

That is, if a certain number a is added to all the numbers in a distribution, then the mean is increased by that certain number. Similarly, if all the numbers in a distribution are multiplied by a certain number b , then the mean is multiplied by that certain number.

Simpson's Paradox

In the 1995 baseball season, Derek Jeter had a batting average of .250, while David Justice had a batting average of .253. In the 1996 season, Jeter batted .314, while Justice batted .321. Could we conclude that when comparing combined batting averages for the 1995 and 1996 seasons, Justice would have a higher average than Jeter?

	Derek Jeter	David Justice
1995	.250	.253
1996	.314	.321
combined	.310	.270

Surprisingly, Jeter has the higher average for the two seasons combined. The following more detailed chart shows why.

	Derek Jeter	David Justice
1995	$\frac{12}{48} = .250$	$\frac{104}{411} = .253$
1996	$\frac{183}{582} = .314$	$\frac{45}{140} = .321$
combined	$\frac{195}{630} = .310$	$\frac{149}{551} = .270$

resolution of Simpson's Paradox. The weighted mean also provides a resolution of Simpson's Paradox. To see this, note that in the combined two seasons, Derek Jeter had

$$\frac{48}{48 + 582} = 0.07619$$

of his at-bats in 1995, and

$$\frac{582}{48 + 582} = 0.92381$$

of them in 1996. So to calculate a reasonable combined average, the batting averages for the two halves should be weighted accordingly:

$$w_1 = 0.07619, \quad w_2 = 0.92381$$

Then, the weighted average for Jeter is

$$\bar{X}_w = w_1X_1 + w_2X_2 = (0.07619)(0.250) + (0.92381)(0.314) = 0.310.$$

For David Justice, the weights used should be:

$$w_1 = \frac{411}{411 + 140} = 0.745917, \quad w_2 = \frac{140}{411 + 140} = 0.254083$$

Justice's weighted average is

$$\bar{X}_w = w_1X_1 + w_2X_2 = (0.745917)(0.253) + (0.254083)(0.321) = 0.270.$$

These calculations yield the accurate averages for the two seasons combined. The reason that Jeter can have a greater combined batting average than Justice is that the combined averages are calculated from the season averages using different weights. Both players do better in 1996, and that season average is weighted higher for Jeter than for Justice.

discussion

To measure the dispersion of the distribution more precisely than the range does, we wish to take into consideration every value, not just the highest and lowest. If we chose some p which was some “middle” number for the distribution, then we could measure the deviation $x_i - p$ of each value x_i of the random variable. We could square each deviation (which for one thing would give all positive numbers) and take the average of the squared deviations

$$\frac{\sum(X - p)^2}{n} = \frac{\sum_{i=1}^k f_i(x_i - p)^2}{\sum_{i=1}^k f_i}.$$

We may wonder which p we should use. We want to have a p that is “as close to all the numbers as possible”. Precisely, we want to choose p so as to minimize the total squared deviation

$$\sum(X - p)^2 = \sum_{i=1}^k f_i(x_i - p)^2.$$

The expression to be minimized is a quadratic expression in the variable p . The graph of $q = \sum(X - p)^2$ on the (p, q) -plane is a parabola opening upward. The minimum value of the expression occurs at the vertex of the parabola, at which $p = \bar{x}$.

This tells us that the sum of the squared deviations is the least about the mean. That is,

$$\sum(X - p)^2 = \sum_{i=1}^k f_i(x_i - p)^2$$

is a minimum when $p = \bar{x}$, the mean.

In other words, the expression for the variance would not be made smaller if \bar{x} were replaced with any other number.

example Consider this distribution:

$$X : 1, 5, 2, 1, 2, 1$$

It could also be written like this:

x_i	1	2	5
f_i	3	2	1

The range of X is $5 - 1 = 4$.

To calculate the standard deviation s , we must calculate the mean \bar{x} first. It is calculated to be

$$\bar{x} = 2.$$

The standard deviation of X is calculated

$$\begin{aligned} s &= \sqrt{\frac{\sum(X - \bar{x})^2}{n}} \\ &= \sqrt{\frac{(1 - 2)^2 + (5 - 2)^2 + (2 - 2)^2 + (1 - 2)^2 + (2 - 2)^2 + (1 - 2)^2}{6}} \\ &= \sqrt{\frac{(-1)^2 + (3)^2 + (0)^2 + (-1)^2 + (0)^2 + (-1)^2}{6}} \\ &= \sqrt{\frac{1 + 9 + 0 + 1 + 0 + 1}{6}} \\ &= \sqrt{\frac{12}{6}} \\ &= \sqrt{2} \end{aligned}$$

from the list, or

$$\begin{aligned} s &= \sqrt{\frac{\sum f_i(x_i - \bar{x})^2}{\sum f_i}} \\ &= \sqrt{\frac{\sum(3)(1-2)^2 + (2)(2-2)^2 + (1)(5-2)^2}{3+2+1}} \\ &= \sqrt{\frac{\sum(3)(-1)^2 + (2)(0)^2 + (1)(3)^2}{6}} \\ &= \sqrt{\frac{\sum(3)(1) + (2)(0) + (1)(9)}{6}} \\ &= \sqrt{\frac{12}{6}} \\ &= \sqrt{2} \end{aligned}$$

from the frequency table.

linearly transformed random variables Suppose X is a random variable with standard deviation s , and a, b are constants. Then

$$V = a + X$$

and

$$W = bX$$

are random variables with respective standard deviations

$$s_V = s_{a+X} = s$$

and

$$s_W = s_{bX} = bs.$$

That is, if a certain number a is added to all the numbers in a distribution, then the standard deviation is unchanged. If all the numbers in a distribution are multiplied by a certain number b , then the standard deviation is multiplied by that certain number.

counter-example However, this is not always the case. Consider

$$X : -2, -2, -2, 0, 1, 1, 1, 3.$$

The mean is 0 and the median is 0.5, but the skew is positive.

example Consider this distribution:

$$X : 1, 5, 2, 1, 2, 1$$

It could also be written like this:

x_i	1	2	5
f_i	3	2	1

To calculate the skew of X , we must calculate the mean \bar{x} and the standard deviation s first. We calculate

$$\bar{x} = 2$$

and

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

The skew of s of X is calculated

$$\begin{aligned} \text{skew}_X &= \left(\frac{\sum (X - \bar{x})^3}{n} \right) / s^3 \\ &= \left(\frac{(1 - 2)^3 + (5 - 2)^3 + (2 - 2)^3 + (1 - 2)^3 + (2 - 2)^3 + (1 - 2)^3}{6} \right) / (\sqrt{2})^3 \\ &= \left(\frac{(-1)^3 + (3)^3 + (0)^3 + (-1)^3 + (0)^3 + (-1)^3}{6} \right) / (\sqrt{2})^3 \\ &= \left(\frac{(-1) + 27 + 0 + (-1) + 0 + (-1)}{6} \right) / (\sqrt{2})^3 \\ &= \left(\frac{24}{6} \right) / (\sqrt{2})^3 \\ &= (4) / (\sqrt{2})^3 \\ &= (4) / (2\sqrt{2}) \\ &= \sqrt{2} \end{aligned}$$

from the list, or

$$\begin{aligned}\text{skew}_X &= \left(\frac{\sum f_i(x_i - \bar{x})^3}{\sum f_i} \right) / s^3 \\ &= \left(\frac{(3)(1-2)^3 + (2)(2-2)^3 + (1)(5-2)^3}{3+2+1} \right) / (\sqrt{2})^3 \\ &= \left(\frac{(3)(-1) + (2)(0) + (1)(27)}{6} \right) / (\sqrt{2})^3 \\ &= \left(\frac{24}{6} \right) / (\sqrt{2})^3 \\ &= (4) / (\sqrt{2})^3 \\ &= (4) / (2\sqrt{2}) \\ &= \sqrt{2}\end{aligned}$$

from the frequency table.

note To calculate whether skew_X is positive or negative, we need not perform the entire calculation above. The sign of skew_X is the same as that of

$$\sum (X - \bar{x})^3 = \sum f_i(x_i - \bar{x})^3,$$

since $n = \sum f_i$ and s are always positive. In the above example,

$$\sum (X - \bar{x})^3 = \sum f_i(x_i - \bar{x})^3 = 4,$$

and we could tell from just this that the distribution is positively skewed.

linearly transformed random variables Suppose X is a random variable, and a, b are constants. Then

$$V = a + X$$

and

$$W = bX$$

are random variables with respective skews

$$\text{skew}_V = \text{skew}_{a+X} = \text{skew}_X$$

and

$$\text{skew}_W = \text{skew}_{bX} = \text{skew}_X.$$

That is, if a certain number a is added to all the numbers in a distribution, then the skew is unchanged. If all the numbers in a distribution are multiplied by a certain number b , then the skew is unchanged.

example Consider this distribution:

$$X : 1, 5, 2, 1, 2, 1$$

It could also be written like this:

x_i	1	2	5
f_i	3	2	1

To calculate the kurtosis of X , we must calculate the mean \bar{x} and the standard deviation s first. We calculate

$$\bar{x} = 2$$

and

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

The kurtosis of s of X is calculated

$$\begin{aligned} \text{kurt}_X &= \left(\frac{\sum (X - \bar{x})^4}{n} \right) / s^4 \\ &= \left(\frac{(1 - 2)^4 + (5 - 2)^4 + (2 - 2)^4 + (1 - 2)^4 + (2 - 2)^4 + (1 - 2)^4}{6} \right) / (\sqrt{2})^4 \\ &= \left(\frac{(-1)^4 + (3)^4 + (0)^4 + (-1)^4 + (0)^4 + (-1)^4}{6} \right) / (\sqrt{2})^4 \\ &= \left(\frac{1 + 81 + 0 + 1 + 0 + 1}{6} \right) / (\sqrt{2})^4 \\ &= \left(\frac{84}{6} \right) / (\sqrt{2})^4 \\ &= (14) / (4) \\ &= \frac{7}{2} \end{aligned}$$

from the list, or

$$\begin{aligned}\text{skew}_X &= \left(\frac{\sum f_i(x_i - \bar{x})^4}{\sum f_i} \right) / s^4 \\ &= \left(\frac{(3)(1-2)^4 + (2)(2-2)^4 + (1)(5-2)^4}{3+2+1} \right) / (\sqrt{2})^4 \\ &= \left(\frac{(3)(1) + (2)(0) + (1)(81)}{6} \right) / (\sqrt{2})^4 \\ &= \left(\frac{84}{6} \right) / (\sqrt{2})^4 \\ &= (14) / (4) \\ &= \frac{7}{2}\end{aligned}$$

from the frequency table.

linearly transformed random variables Suppose X is a random variable, and a, b are constants. Then

$$V = a + X$$

and

$$W = bX$$

are random variables with respective kurtoses

$$\text{kurt}_V = \text{kurt}_{a+X} = \text{kurt}_X$$

and

$$\text{kurt}_W = \text{kurt}_{bX} = \text{kurt}_X.$$

That is, if a certain number a is added to all the numbers in a distribution, then the kurtosis is unchanged. If all the numbers in a distribution are multiplied by a certain number b , then the kurtosis is unchanged.

reason Since X has mean \bar{x} , then $X - \bar{x}$ has mean $\bar{x} - \bar{x} = 0$, and consequently $\frac{X - \bar{x}}{s}$ has mean $\frac{0}{s} = 0$. Since X has standard deviation s , then $X - \bar{x}$ has standard deviation s also, and consequently $\frac{X - \bar{x}}{s}$ has standard deviation $\frac{s}{s} = 1$.

the usefulness of the Standard Normal distribution If $z = \frac{x-\bar{x}}{s}$, then

$$P(X < x) = P(Z < z).$$

That is, the proportion of the population with x-scores below a certain value of X is the same as the proportion of the population with z-scores below the corresponding value of Z .

The latter can be determined from a table.

If a random variable X is Normal, we can calculate the percentile rank of a given value x of X by calculating the corresponding value $z = \frac{x-\bar{x}}{s}$ of $Z = \frac{X-\bar{x}}{s}$, and determining the percentile rank of that z-score using the table.

On the other hand, if we want to find the p th percentile of a Normal distribution, we can find the p th percentile z of the Standard Normal distribution, and find the corresponding value x of X , by solving the equation $z = \frac{X-\bar{x}}{s}$ for X .

proportions for the Standard Normal distribution The following table shows, for $z > 0$ a value of the Standard Normal variable Z , the proportion

$$P(Z > z)$$

of the population having values of Z greater than z . (the area under the Standard Normal curve to the right of z).

$\pm z$	Area Beyond $\pm z$	$\pm z$	Area Beyond $\pm z$	$\pm z$	Area Beyond $\pm z$	$\pm z$	Area Beyond $\pm z$	$\pm z$	Area Beyond $\pm z$	$\pm z$	Area Beyond $\pm z$	$\pm z$	Area Beyond $\pm z$	$\pm z$	Area Beyond $\pm z$
0.00	0.5000	0.36	0.3594	0.72	0.2358	1.08	0.1401	1.44	0.0749	1.80	0.0359	2.16	0.0154	2.52	0.0059
0.01	0.4960	0.37	0.3557	0.73	0.2327	1.09	0.1379	1.45	0.0735	1.81	0.0351	2.17	0.0150	2.53	0.0057
0.02	0.4920	0.38	0.3520	0.74	0.2296	1.10	0.1357	1.46	0.0721	1.82	0.0344	2.18	0.0146	2.54	0.0055
0.03	0.4880	0.39	0.3483	0.75	0.2266	1.11	0.1335	1.47	0.0708	1.83	0.0336	2.19	0.0143	2.55	0.0054
0.04	0.4840	0.40	0.3446	0.76	0.2236	1.12	0.1314	1.48	0.0694	1.84	0.0329	2.20	0.0139	2.56	0.0052
0.05	0.4801	0.41	0.3409	0.77	0.2206	1.13	0.1292	1.49	0.0681	1.85	0.0322	2.21	0.0136	2.57	0.0051
0.06	0.4761	0.42	0.3372	0.78	0.2177	1.14	0.1271	1.50	0.0668	1.86	0.0314	2.22	0.0132	2.58	0.0049
0.07	0.4721	0.43	0.3336	0.79	0.2148	1.15	0.1251	1.51	0.0655	1.87	0.0307	2.23	0.0129	2.59	0.0048
0.08	0.4681	0.44	0.3300	0.80	0.2119	1.16	0.1230	1.52	0.0643	1.88	0.0301	2.24	0.0125	2.60	0.0047
0.09	0.4641	0.45	0.3264	0.81	0.2090	1.17	0.1210	1.53	0.0630	1.89	0.0294	2.25	0.0122	2.61	0.0045
0.10	0.4602	0.46	0.3228	0.82	0.2061	1.18	0.1190	1.54	0.0618	1.90	0.0287	2.26	0.0119	2.62	0.0044
0.11	0.4562	0.47	0.3192	0.83	0.2033	1.19	0.1170	1.55	0.0606	1.91	0.0281	2.27	0.0116	2.63	0.0043
0.12	0.4522	0.48	0.3156	0.84	0.2005	1.20	0.1151	1.56	0.0594	1.92	0.0274	2.28	0.0113	2.64	0.0041
0.13	0.4483	0.49	0.3121	0.85	0.1977	1.21	0.1131	1.57	0.0582	1.93	0.0268	2.29	0.0110	2.65	0.0040
0.14	0.4443	0.50	0.3085	0.86	0.1949	1.22	0.1112	1.58	0.0571	1.94	0.0262	2.30	0.0107	2.66	0.0039
0.15	0.4404	0.51	0.3050	0.87	0.1922	1.23	0.1093	1.59	0.0559	1.95	0.0256	2.31	0.0104	2.67	0.0038
0.16	0.4364	0.52	0.3015	0.88	0.1894	1.24	0.1075	1.60	0.0548	1.96	0.0250	2.32	0.0102	2.68	0.0037
0.17	0.4325	0.53	0.2981	0.89	0.1867	1.25	0.1056	1.61	0.0537	1.97	0.0244	2.33	0.0099	2.69	0.0036
0.18	0.4286	0.54	0.2946	0.90	0.1841	1.26	0.1038	1.62	0.0526	1.98	0.0239	2.34	0.0096	2.70	0.0035
0.19	0.4247	0.55	0.2912	0.91	0.1814	1.27	0.1020	1.63	0.0516	1.99	0.0233	2.35	0.0094	2.71	0.0034
0.20	0.4207	0.56	0.2877	0.92	0.1788	1.28	0.1003	1.64	0.0505	2.00	0.0228	2.36	0.0091	2.72	0.0033
0.21	0.4168	0.57	0.2843	0.93	0.1762	1.29	0.0985	1.65	0.0495	2.01	0.0222	2.37	0.0089	2.73	0.0032
0.22	0.4129	0.58	0.2810	0.94	0.1736	1.30	0.0968	1.66	0.0485	2.02	0.0217	2.38	0.0087	2.74	0.0031
0.23	0.4090	0.59	0.2776	0.95	0.1711	1.31	0.0951	1.67	0.0475	2.03	0.0212	2.39	0.0084	2.75	0.0030
0.24	0.4052	0.60	0.2743	0.96	0.1685	1.32	0.0934	1.68	0.0465	2.04	0.0207	2.40	0.0082	2.76	0.0029
0.25	0.4013	0.61	0.2709	0.97	0.1660	1.33	0.0918	1.69	0.0455	2.05	0.0202	2.41	0.0080	2.77	0.0028
0.26	0.3974	0.62	0.2676	0.98	0.1635	1.34	0.0901	1.70	0.0446	2.06	0.0197	2.42	0.0078	2.78	0.0027
0.27	0.3936	0.63	0.2643	0.99	0.1611	1.35	0.0885	1.71	0.0436	2.07	0.0192	2.43	0.0075	2.79	0.0026
0.28	0.3897	0.64	0.2611	1.00	0.1587	1.36	0.0869	1.72	0.0427	2.08	0.0188	2.44	0.0073	2.80	0.0026
0.29	0.3859	0.65	0.2578	1.01	0.1562	1.37	0.0853	1.73	0.0418	2.09	0.0183	2.45	0.0071	2.81	0.0025
0.30	0.3821	0.66	0.2546	1.02	0.1539	1.38	0.0838	1.74	0.0409	2.10	0.0179	2.46	0.0069	2.82	0.0024
0.31	0.3783	0.67	0.2514	1.03	0.1515	1.39	0.0823	1.75	0.0401	2.11	0.0174	2.47	0.0068	2.83	0.0023
0.32	0.3745	0.68	0.2483	1.04	0.1492	1.40	0.0808	1.76	0.0392	2.12	0.0170	2.48	0.0066	2.84	0.0023
0.33	0.3707	0.69	0.2451	1.05	0.1469	1.41	0.0793	1.77	0.0384	2.13	0.0166	2.49	0.0064	2.85	0.0022
0.34	0.3669	0.70	0.2420	1.06	0.1446	1.42	0.0778	1.78	0.0375	2.14	0.0162	2.50	0.0062	2.86	0.0021
0.35	0.3632	0.71	0.2389	1.07	0.1423	1.43	0.0764	1.79	0.0367	2.15	0.0158	2.51	0.0060	2.87	0.0021

note Because of the symmetry of the Standard Normal Distribution, this is the same as $P(Z \leq -z)$ (the area under the Standard Normal curve to the left of $-z$).

specific cases of proportion of population

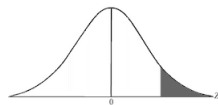
1. Suppose $z_1 < 0$. Then $P(Z < z_1)$ may be found directly from the table:



And $P(Z > z_1) = 1 - P(Z < z_1)$:



2. Suppose $z_2 > 0$. Then $P(Z > z_2)$ may be found directly from the table:



And $P(Z < z_2) = 1 - P(Z > z_2)$:



3. Suppose $z_1 < z_2 < 0$. Then $P(z_1 < Z < z_2) = P(Z < z_2) - P(Z < z_1)$:

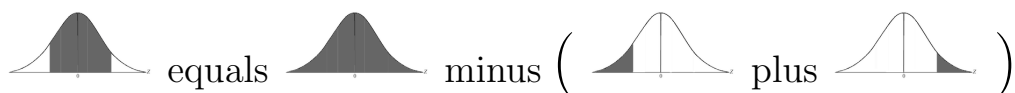


4. Suppose $0 < z_1 < z_2$. Then $P(z_1 < Z < z_2) = P(Z > z_1) - P(Z > z_2)$:



5. Finally suppose $z_1 < 0 < z_2$.

Then $P(z_1 < Z < z_2) = 1 - (P(Z < z_1) + P(Z > z_2))$:



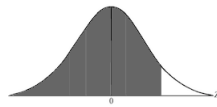
specific cases of percentile

1. Suppose $0 < p < 50$.

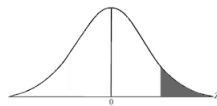


Then the table may be used to find the p th percentile of the Standard Normal distribution (the value $z_1 < 0$ of Z such that $P(Z < z_1) = p\%$). Note that a negative sign will have to be prepended to the value found in the table (in the “ $\pm z$ ” column).

2. Suppose $50 < p < 100$.



Then the table may be used to find the p th percentile. It is the value $z_1 > 0$ of Z such that $P(Z < z_1) = p\%$, or, equivalently, $P(Z > z_1) = (100 - p)\%$:



Note that the proportion to be looked up in the table (in the “area beyond” column) is $1 - p$.

example Suppose X is a Normal random variable with mean 47 and standard deviation 5. What is the percentile rank of the value $X = 55$?

We first calculate the z-score corresponding to the x-score 55:

$$\begin{aligned} Z &= \frac{X - \bar{x}}{s} \\ &= \frac{55 - 47}{5} \\ &= \frac{8}{5} \\ &= 1.6. \end{aligned}$$

From the table “proportions for the Standard Normal distribution”, we find that the proportion of the population with z-scores greater than 1.6 is 0.0548, or 5.48%. It follows that $1 - 0.0548 = 0.9452$ of the population – or $100\% - 5.48\% = 94.52\%$ – has z-scores less than 1.6. So the value $X = 55$ is the 94.52th percentile.

example

Suppose X is a Normal random variable with mean 47 and standard deviation 5. What is the 61st percentile?

We first calculate the 61st percentile of the Standard Normal distribution. We observe that the z such that 61% of the population has values of Z lower than z , is the z such that 39% of the population has values of Z greater than z . From the table “proportions for the Standard Normal distribution”, we find this value to be approximately $Z = 0.28$. Now we calculate the x-score corresponding to this z-score:

$$\begin{aligned}Z &= \frac{X - \bar{x}}{s} \\ \Rightarrow 0.28 &= \frac{X - 47}{5} \\ \Rightarrow 1.4 &= X - 47 \\ \Rightarrow 48.4 &= X\end{aligned}$$

The 61st percentile of X is thus 48.4.

linearly transformed random variables, etc. Suppose X and Y are two random variables for the same population. Then

$$\text{COV}_{X,Y} = \text{COV}_{Y,X}.$$

Now suppose a, b are constants. Then

$$V = a + X$$

and

$$W = bX$$

are random variables with covariances

$$\text{COV}_{V,Y} = \text{COV}_{a+X,Y} = \text{COV}_{X,Y}$$

$$\text{COV}_{W,Y} = \text{COV}_{bX,Y} = b \cdot \text{COV}_{X,Y}$$

with Y .

note The covariance measures the degree to which two random variables “vary together” from their respective means. But this is affected by the degree to which each variable separately varies from its mean; that is, by the standard deviation of each variable. So we divide by the the standard deviations of the two variables to compensate for this. The resulting quantity, the “correlation coefficient”, measures the degree to which two random variables “vary together”, from their respective means, relative to their respective standard deviations.

meaning The correlation coefficient measures the extent to which two random variables have a linear relationship. The correlation coefficient always satisfies

$$-1 \leq \text{corr}_{X,Y} \leq 1.$$

If $\text{corr}_{X,Y} = 1$, then the random variables X and Y are perfectly correlated. If $\text{corr}_{X,Y} = -1$, then X and Y are perfectly anti-correlated. If $\text{corr}_{X,Y} = 0$, then there is no linear relationship between X and Y .

example

Suppose we have the following bivariate distribution:

X	0	8	4	6	2
Y	2	4	1	5	3

First, we calculate the means of X and Y :

$$\begin{aligned}\bar{x} &= \frac{\sum X}{n} = \frac{0 + 8 + 4 + 6 + 2}{5} = \frac{20}{5} = 4 \\ \bar{y} &= \frac{\sum Y}{n} = \frac{2 + 4 + 1 + 5 + 3}{5} = \frac{15}{5} = 3\end{aligned}$$

Next, we calculate the standard deviations of X and Y :

$$\begin{aligned}s_X &= \sqrt{\frac{\sum (X - \bar{x})^2}{n}} \\ &= \sqrt{\frac{(0 - 4)^2 + (8 - 4)^2 + (4 - 4)^2 + (6 - 4)^2 + (2 - 4)^2}{5}} \\ &= \sqrt{\frac{(-4)^2 + (4)^2 + (0)^2 + (2)^2 + (-2)^2}{5}} \\ &= \sqrt{\frac{16 + 16 + 0 + 4 + 4}{5}} = \sqrt{\frac{40}{5}} = \sqrt{8}\end{aligned}$$

$$\begin{aligned}s_Y &= \sqrt{\frac{\sum (Y - \bar{y})^2}{n}} \\ &= \sqrt{\frac{(2 - 3)^2 + (4 - 3)^2 + (1 - 3)^2 + (5 - 3)^2 + (3 - 3)^2}{5}} \\ &= \sqrt{\frac{(-1)^2 + (1)^2 + (-2)^2 + (2)^2 + (0)^2}{5}} \\ &= \sqrt{\frac{1 + 1 + 4 + 4 + 0}{5}} = \sqrt{\frac{10}{5}} = \sqrt{2}\end{aligned}$$

Now, it's time to start looking at the relationship of the variables to each other. We calculate the covariance of X and Y :

$$\begin{aligned}\text{cov}_{X,Y} &= \frac{\sum(X - \bar{x})(Y - \bar{y})}{n} \\ &= \frac{(0 - 4)(2 - 3) + (8 - 4)(4 - 3) + (4 - 4)(1 - 3) + (6 - 4)(5 - 3) + (2 - 4)(3 - 3)}{5} \\ &= \frac{(-4)(-1) + (4)(1) + (0)(-2) + (2)(2) + (-2)(0)}{5} \\ &= \frac{4 + 4 + 0 + 4 + 0}{5} = \frac{12}{5} = 2.4\end{aligned}$$

Finally, we can calculate the correlation of X and Y :

$$\begin{aligned}\text{corr}_{X,Y} &= \frac{\text{cov}_{X,Y}}{s_X s_Y} \\ &= \frac{2.4}{(\sqrt{8})(\sqrt{2})} \\ &= 0.6\end{aligned}$$

linearly transformed random variables, etc. Suppose X and Y are two random variables for the same population. Then

$$\text{corr}_{X,Y} = \text{corr}_{Y,X}.$$

Now suppose a, b are constants. Then

$$V = a + X$$

and

$$W = bX$$

are random variables with correlations

$$\text{corr}_{V,Y} = \text{corr}_{a+X,Y} = \text{corr}_{X,Y}$$

$$\text{corr}_{W,Y} = \text{corr}_{bX,Y} = \text{corr}_{X,Y}$$

with Y .

correlation coefficient and z-score Let

$$Z_X = \frac{X - \bar{x}}{s_X}$$

be the z-score for X and let

$$Z_Y = \frac{Y - \bar{y}}{s_Y}$$

be the z-score for Y .

Since $\frac{\text{cov}_{X,Y}}{s_X s_Y} = \frac{\left(\frac{\sum(X-\bar{x})(Y-\bar{y})}{n}\right)}{s_X s_Y} = \frac{\sum\left(\frac{X-\bar{x}}{s_X}\right)\left(\frac{Y-\bar{y}}{s_Y}\right)}{n} = \frac{\sum Z_X Z_Y}{n}$, we may also write

$$\boxed{\text{corr}_{X,Y} = \frac{\sum Z_X Z_Y}{n}}$$

limitations One must be aware of the limitations of what one can conclude from the correlation coefficient.

linearity The correlation coefficient reflects only the linear relationship between two random variables. In other words, it only measures the extent to which a straight line fits the scatterplot of two variables. The assumption that the two random variables have a linear relationship is required to justify the use of the correlation coefficient as a measure of their relationship.

outliers Outliers can have a strong effect on the correlation coefficient, and may result in misleading conclusions about the relationship between two random variables.

correlation and causation Just because there is a high correlation between two measurable phenomena, does not mean that one causes the other.

reliability and validity Reliability is the extent to which a test produces consistent results. It is measured by the correlation between the results obtained from two different applications of the same test to the same population. The closer the correlation to 1, the greater the evidence of reliability.

Validity, on the other hand, is the extent to which a test measures what it purports to measure. It is measured by the correlation between the results obtained, and the results obtained from an accepted test applied to the same population. The closer the correlation to 1, the greater the evidence of validity.

calculation This provides justification for the formula for the regression line, but does not need to be memorized. We have

$$\sum(Y' - Y)^2 = \sum((mX + b) - Y)^2.$$

The expression to be minimized is a quadratic expression in the variable b . The graph of $c = \sum((mX + b) - Y)^2$ on the (b, c) -plane is a parabola opening upward. The minimum value of $\sum((mX + b) - Y)^2$ occurs at the vertex of the parabola, at which $b = \bar{y} - m\bar{x}$. This means that b should be chosen to be $\bar{y} - m\bar{x}$, once m is chosen.

Because of this, the sum of squares to be minimized is actually

$$\begin{aligned}\sum(Y' - Y)^2 &= \sum((mX + b) - Y)^2 \\ &= \sum(m(X - \bar{x}) - (Y - \bar{y}))^2 \\ &= n(\text{var}_X m^2 - 2\text{cov}_{X,Y} m + \text{var}_Y).\end{aligned}$$

This last expression is quadratic in the variable m . The graph of $q = n(\text{var}_X m^2 - 2\text{cov}_{X,Y} m + \text{var}_Y)$ on the (m, q) -plane is a parabola opening upward. The minimum value of $n(\text{var}_X m^2 - 2\text{cov}_{X,Y} m + \text{var}_Y)$ occurs at the vertex of the parabola, at which

$$m = \frac{\text{cov}_{X,Y}}{\text{var}_X} = \frac{\text{cov}_{X,Y}}{s_X^2} = \frac{\text{cov}_{X,Y}}{s_X s_Y} \cdot \frac{s_Y}{s_X} = \text{corr}_{X,Y} \frac{s_Y}{s_X}.$$

This proves that the sum of the squares $\sum(Y' - Y)^2$ is minimized when

$$\begin{aligned}Y' &= mX + b \\ &= mX + (\bar{y} - m\bar{x}) \\ &= m(X - \bar{x}) + \bar{y} \\ &= \left(\text{corr}_{X,Y} \frac{s_Y}{s_X}\right) (X - \bar{x}) + \bar{y},\end{aligned}$$

or equivalently

$$Y' - \bar{y} = \left(\text{corr}_{X,Y} \frac{s_Y}{s_X}\right) (X - \bar{x}).$$

example

Suppose we have the following bivariate distribution:

X	0	8	4	6	2
Y	2	4	1	5	3

First, we calculate the means and standard deviations of X and Y :

$$\bar{x} = 4, \quad \bar{y} = 3$$

Next, we calculate the correlation of X and Y :

$$\text{corr}_{X,Y} = 0.6$$

We can now find the equation of the regression line of Y on X :

$$\begin{aligned}y - \bar{y} &= \text{corr}_{X,Y} \frac{s_Y}{s_X} (x - \bar{x}) \\ \Rightarrow y - 3 &= (0.6) \frac{(\sqrt{8})}{(\sqrt{2})} (x - 4) \\ \Rightarrow y - 3 &= (1.2)(x - 4) \\ \Rightarrow y - 3 &= 1.2x - 4.8 \\ \Rightarrow y &= 1.2x - 1.8\end{aligned}$$

note Our goal is to reject the null hypothesis H_0 . Note that if we cannot marshal enough evidence to reject the null hypothesis H_0 , that does not mean that we accept H_0 , only that we fail to reject H_0 .

note The average value of the sample mean is the mean of the whole population. In other words, the mean of \bar{X} is \bar{x} .

note The variance of the sample mean is given by

$$\text{var}_{\bar{X}} = \text{var}_X/n.$$

The standard deviation of the sample mean is given by

$$s_{\bar{X}} = s/\sqrt{n},$$

where s is the standard deviation of X .

Central Limit Theorem No matter what the distribution of X for the population, the distribution of sample means \bar{X} becomes, as n increases, more and more like a Normal distribution.

Here we will assume, because of the Central Limit Theorem, that \bar{X} is Normal.

example (“left-tailed” test) Suppose a company claims that its light bulbs last an average of 1000 hours. By some means we know that the standard deviation of the lives of the bulbs is 200 hours. (This is in fact a bit unrealistic; we will repeat the example later, without this assumption.) We buy 4 light bulbs; they last 850 hours, 900 hours, 650 hours, and 800 hours. We suspect that the company’s claim about the mean life of their bulbs is untrue, and want to test it at the 5% significance level. We choose this claim as H_0 , the null hypothesis:

$$H_0 : \bar{x} = \bar{x}_0 = 1000$$

We suspect that the mean life of the bulbs is in fact less than 1000 hours:

$$\bar{x} < \bar{x}_0 = 1000$$

What we’d really like to do is find evidence that H_0 is false; and if we do, we’ll conclude that $\bar{x} < \bar{x}_0 = 1000$. This is called a “left-tailed” test. But to find the evidence, we’ll start by assuming that H_0 is true, and see what happens.

Assume that $\bar{x} = \bar{x}_0 = 1000$. The value of \bar{X} for our sample is

$$\bar{X} = \frac{\sum X}{n} = \frac{850 + 900 + 650 + 800}{4} = 800.$$

Calculate the z-score of this value of \bar{X} :

$$Z_{\bar{X}} = \frac{\bar{X} - \bar{x}_0}{s/\sqrt{n}} = \frac{800 - 1000}{200/\sqrt{4}} = \frac{-200}{100} = -2.$$

Now we use a table to find the probability that we would pick a random sample with an z-score at least as low as -2 (or lower). We find

$$P(Z_{\bar{X}} \leq -2) \approx 0.0228.$$

What does this mean? It means that if the company’s claim (H_0) were true, there would be a probability of only 0.0228 that we would obtain a random sample with a sample mean as low as 800. Since we in fact did obtain such a sample, this casts serious doubt on the company’s claim. Precisely, we say that we reject H_0 at the 5% significance level. That is, we reject H_0

with 95% confidence. Put another way, we can say that we have (at least) 95% confidence that the mean life of the company's light bulbs is in fact less than 1000 hours.

note In the above example, we reject H_0 with 95% confidence – that is, at the 5% significance level. We would also reject H_0 at the 2.5% significance level – but not at the 1% significance level. That is, we could assert that $\bar{x} < 1000$ with 97.5% confidence, but not with 99% confidence.

example (“right-tailed” test) Suppose a company claims that its low-calorie cookies contain 100 calories each. By some means we know that the standard deviation of the energy per cookie is 10. (As before, this is in fact a bit unrealistic; we will repeat the example later, without this assumption.) We buy 4 low-calorie cookies and measure their energy content; they contain 120 calories, 95 calories, 115 calories, and 110 calories. We suspect that the company’s cookies contain an average of more than 100 calories each, and want to test this at the 5% significance level. We choose

$$H_0 : \quad \bar{x} = \bar{x}_0 = 100$$

If we reject this hypothesis, we will conclude that the mean number of calories per cookie is in fact more than 100 calories:

$$\bar{x} > \bar{x}_0 = 100.$$

This is called a right-tailed test.

Assume H_0 – that $\bar{x} = \bar{x}_0 = 100$. The value of \bar{X} for our sample is

$$\bar{X} = \frac{\sum X}{n} = \frac{120 + 95 + 115 + 110}{4} = 110.$$

Calculate the z-score of this value of $\bar{X} = 110$:

$$Z_{\bar{X}} = \frac{\bar{X} - \bar{x}_0}{s/\sqrt{n}} = \frac{110 - 100}{10/\sqrt{4}} = \frac{10}{5} = 2.$$

Then

$$P(2 \leq Z_{\bar{X}}) \approx 0.0228$$

(from a table).

So, as before, we reject H_0 at the 5% significance level. That is, we can conclude that $\bar{x} > 100$ is true with 95% confidence.

“one-tailed”, “two-tailed” In the above two examples, our conclusion, upon rejecting H_0 , was that the (population) mean \bar{x} was less than a certain value ($\bar{x} < \bar{x}_0 = 1000$), or a statement that the (population) mean was greater than a certain value ($\bar{x} > \bar{x}_0 = 100$). The hypothesis test is said to be “one-tailed” in such a case.

Sometimes, however, it is not that we want to show that \bar{x} is larger than a specific value, or smaller than a specific value, but merely unequal to a specific value (that is, larger or smaller). In the above two examples, it was only one inequality that we were concerned about. We don’t care if the light bulbs last longer than advertised; we are only concerned if they last less time than advertised. We don’t care if the low-calorie cookies have even fewer calories than advertised; we are only concerned if they have more calories than advertised. But there are situations where we would be concerned with \bar{x} being higher or lower than a particular value. In this case our hypothesis test will be “two-tailed”: upon rejecting H_0 we will conclude simply that \bar{x} is unequal to a particular value. Here we will consider the probability that the z-score of the sample mean \bar{X} is greater than 0 by at least certain amount or less than 0 by at that same amount.

example (two-tailed test) Suppose an exercise equipment company makes 10-kilogram barbell plates. By some means we know that the standard deviation of the weight of the plates is 0.1kg. (As before, this is in fact a bit unrealistic; we will repeat the example later, without this assumption.) (Implicit is their claim that the plates weigh an average of 10 kilos.) We inspect 4 plates, and find their weights to be 9.9kg, 10.0kg, 9.8kg, and 9.9kg. We suspect that the company's "10kg" barbell plates do not weigh an average of 10kg, and want to test this at the 5% significance level. This time, we are concerned that the claimed weights are not accurate; we would be care both if the mean weight were higher than 10kg, and if the mean weight were lower than 10kg. We choose

$$H_0 : \bar{x} = \bar{x}_0 = 10$$

Our conclusion upon rejecting the null hypothesis will be that the mean weight is not equal to 10kg. (that is, more than 10kg or less than 10kg.):

$$\bar{x} \neq \bar{x}_0 = 10.$$

(This is a two-tailed hypothesis test.)

Assume H_0 – that $\bar{x} = 10$. The value of \bar{X} for our sample is

$$\bar{X} = \frac{\sum X}{n} = \frac{9.9 + 10.0 + 9.8 + 9.9}{4} = 9.9.$$

This value differs from the assumed mean by 0.1. Calculate the z-score of this value of $\bar{X} = 9.9$:

$$Z_{\bar{X}} = \frac{\bar{X} - \bar{x}_0}{s/\sqrt{n}} = \frac{9.9 - 10}{0.1/\sqrt{4}} = \frac{0.1}{0.05} = -2.$$

We have

$$\begin{aligned} P(Z_{\bar{X}} \leq -2 \text{ or } Z_{\bar{X}} \geq 2) &= P(Z_{\bar{X}} \leq -2) + P(Z_{\bar{X}} \geq 2) \\ &\approx 0.0228 + 0.0228 \text{ (from a table)} \\ &= 0.0456. \end{aligned}$$

We therefore reject H_0 at the 5% significance level. That is, we conclude that $\bar{x} \neq 10$ with 95% confidence.

note In the above example, we reject H_0 with 95% confidence – that is, at the 5% significance level. We would also reject H_0 at the 10% significance level – but not at the 2% significance level. That is, we could assert $\bar{x} \neq 10$ with 90% confidence, but not with 98% confidence.

critical values for the Standard Normal distribution Notice that in using the z-score table in the above examples, we are really only checking, for our sample's value z of Z , whether or not $P(Z \geq |z|) \leq \alpha$ (for a one-tailed alternative hypothesis) or whether or not $2P(Z \geq |z|) \leq \alpha$ (for a two-tailed alternative hypothesis). (Here α is the significance level.) We can construct a more concise table containing just such information. The value z_0 given in the table is the critical value.

	level of significance α of one-tailed test					
	0.10	0.05	0.025	0.01	0.005	0.0005
	level of significance α of two-tailed test					
	0.20	0.10	0.05	0.02	0.01	0.001
z_0	1.282	1.645	1.960	2.326	2.576	3.291

If, for our sample, $|Z|$ is greater than this value, we reject the null hypothesis with confidence $1 - \alpha$.

“sample variance” The “sample variance” of X is

$$S^2 = \frac{\sum(X - \bar{X})^2}{n - 1}.$$

question Why does the formula for sample variance use $n - 1$ in place of the n in the formula for population variance?

answer The average value of the sample variance (a random variable) is the variance of the whole population. In other words, the mean of $\frac{\sum(X - \bar{X})^2}{n - 1}$ (where n is the size of the sample) is s^2 .

proof *This is for reference only and does not need to be memorized.* First note that the average value of $X - \bar{x}$ for the whole population is 0, while the average value of $(X - \bar{x})^2$ for the whole population is s^2 . Let X_1, \dots, X_n be independent random variables distributed identically to X . These represent the values of X for a sample of size n . Then $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$ is the sample mean. Note that the average value of $(X_i - \bar{x})(X_j - \bar{x})$ for the whole population is 0 if $i \neq j$ and s^2 if $i = j$. The random variable

$$\begin{aligned} S^2 &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} \\ &= \frac{\sum_{i=1}^n \left(X_i - \left(\frac{X_1 + \dots + X_n}{n} \right) \right)^2}{n - 1} \\ &= \frac{\sum_{i=1}^n \left((X_i - \bar{x}) - \left(\frac{(X_1 - \bar{x}) + \dots + (X_n - \bar{x})}{n} \right) \right)^2}{n - 1} \\ &= \frac{\sum_{i=1}^n \left((X_i - \bar{x})^2 - 2(X_i - \bar{x}) \left(\frac{(X_1 - \bar{x}) + \dots + (X_n - \bar{x})}{n} \right) + \left(\frac{(X_1 - \bar{x}) + \dots + (X_n - \bar{x})}{n} \right)^2 \right)}{n - 1} \end{aligned}$$

has average value

$$\frac{\sum_{i=1}^n \left(s^2 - \frac{2}{n}s^2 + \frac{1}{n^2}(ns^2) \right)}{n - 1} = \frac{n \left(s^2 - \frac{2}{n}s^2 + \frac{1}{n^2}(ns^2) \right)}{n - 1} = s^2.$$

□

critical values for the Student distribution As with the situation when s is known, we are really only checking, for our sample's value t of T , whether or not $P(T \geq |t|) \leq \alpha$ (for a one-tailed test) or whether or not $2P(T \geq |t|) \leq \alpha$ (for a two-tailed test). (Here α is the significance level.) As we did for z -scores, we can construct a table for hypothesis testing with t -scores, (to be used when the population standard deviation s is unknown). The entries in the main part of the table are the critical values of the Student distribution with $df = n - 1$ degrees of freedom. That is, they are the values of t_0 such that $P(T \geq t_0) = \alpha$ (for a one-tailed test) or $2P(T \geq t_0) = \alpha$ (for a two-tailed test).

Level of Significance for One-Tailed Test						Level of Significance for One-Tailed Test						Level of Significance for One-Tailed Test					
Level of Significance for Two-Tailed Test						Level of Significance for Two-Tailed Test						Level of Significance for Two-Tailed Test					
df	0.10	0.05	0.02	0.01	0.001	df	0.10	0.05	0.02	0.01	0.001	df	0.10	0.05	0.02	0.01	0.001
1	6.31	12.71	31.82	63.66	636.58	32	1.69	2.04	2.45	2.74	3.62	63	1.67	2.00	2.39	2.66	3.45
2	2.92	4.30	6.96	9.92	31.60	33	1.69	2.03	2.44	2.73	3.61	64	1.67	2.00	2.39	2.65	3.45
3	2.35	3.18	4.54	5.84	12.92	34	1.69	2.03	2.44	2.73	3.60	65	1.67	2.00	2.39	2.65	3.45
4	2.13	2.78	3.75	4.60	8.61	35	1.69	2.03	2.44	2.72	3.59	66	1.67	2.00	2.38	2.65	3.44
5	2.02	2.57	3.36	4.03	6.87	36	1.69	2.03	2.43	2.72	3.58	67	1.67	2.00	2.38	2.65	3.44
6	1.94	2.45	3.14	3.71	5.96	37	1.69	2.03	2.43	2.72	3.57	68	1.67	2.00	2.38	2.65	3.44
7	1.89	2.36	3.00	3.50	5.41	38	1.69	2.02	2.43	2.71	3.57	69	1.67	1.99	2.38	2.65	3.44
8	1.86	2.31	2.90	3.36	5.04	39	1.68	2.02	2.43	2.71	3.56	70	1.67	1.99	2.38	2.65	3.43
9	1.83	2.26	2.82	3.25	4.78	40	1.68	2.02	2.42	2.70	3.55	71	1.67	1.99	2.38	2.65	3.43
10	1.81	2.23	2.76	3.17	4.59	41	1.68	2.02	2.42	2.70	3.54	72	1.67	1.99	2.38	2.65	3.43
11	1.80	2.20	2.72	3.11	4.44	42	1.68	2.02	2.42	2.70	3.54	73	1.67	1.99	2.38	2.64	3.43
12	1.78	2.18	2.68	3.05	4.32	43	1.68	2.02	2.42	2.70	3.53	74	1.67	1.99	2.38	2.64	3.43
13	1.77	2.16	2.65	3.01	4.22	44	1.68	2.02	2.41	2.69	3.53	75	1.67	1.99	2.38	2.64	3.42
14	1.76	2.14	2.62	2.98	4.14	45	1.68	2.01	2.41	2.69	3.52	76	1.67	1.99	2.38	2.64	3.42
15	1.75	2.13	2.60	2.95	4.07	46	1.68	2.01	2.41	2.69	3.51	77	1.66	1.99	2.38	2.64	3.42
16	1.75	2.12	2.58	2.92	4.01	47	1.68	2.01	2.41	2.68	3.51	78	1.66	1.99	2.38	2.64	3.42
17	1.74	2.11	2.57	2.90	3.97	48	1.68	2.01	2.41	2.68	3.50	79	1.66	1.99	2.37	2.64	3.42
18	1.73	2.10	2.55	2.88	3.92	49	1.68	2.01	2.40	2.68	3.50	80	1.66	1.99	2.37	2.64	3.42
19	1.73	2.09	2.54	2.86	3.88	50	1.68	2.01	2.40	2.68	3.50	81	1.66	1.99	2.37	2.64	3.41
20	1.72	2.09	2.53	2.85	3.85	51	1.68	2.01	2.40	2.68	3.49	82	1.66	1.99	2.37	2.64	3.41
21	1.72	2.08	2.52	2.83	3.82	52	1.67	2.01	2.40	2.67	3.49	83	1.66	1.99	2.37	2.64	3.41
22	1.72	2.07	2.51	2.82	3.79	53	1.67	2.01	2.40	2.67	3.48	84	1.66	1.99	2.37	2.64	3.41
23	1.71	2.07	2.50	2.81	3.77	54	1.67	2.00	2.40	2.67	3.48	85	1.66	1.99	2.37	2.63	3.41
24	1.71	2.06	2.49	2.80	3.75	55	1.67	2.00	2.40	2.67	3.48						
25	1.71	2.06	2.49	2.79	3.73	56	1.67	2.00	2.39	2.67	3.47	90	1.66	1.99	2.37	2.63	3.40
26	1.71	2.06	2.48	2.78	3.71	57	1.67	2.00	2.39	2.66	3.47	100	1.66	1.98	2.36	2.63	3.39
27	1.70	2.05	2.47	2.77	3.69	58	1.67	2.00	2.39	2.66	3.47	120	1.66	1.98	2.36	2.62	3.37
28	1.70	2.05	2.47	2.76	3.67	59	1.67	2.00	2.39	2.66	3.46	140	1.66	1.98	2.35	2.61	3.36
29	1.70	2.05	2.46	2.76	3.66	60	1.67	2.00	2.39	2.66	3.46	160	1.65	1.97	2.35	2.61	3.35
30	1.70	2.04	2.46	2.75	3.65	61	1.67	2.00	2.39	2.66	3.46	200	1.65	1.97	2.35	2.60	3.34
31	1.70	2.04	2.45	2.74	3.63	62	1.67	2.00	2.39	2.66	3.45	220	1.65	1.97	2.34	2.60	3.34

If, for our sample, $|T|$ is greater than this value, we reject the null hypothesis with confidence $1 - \alpha$.

example (left-tailed test) Suppose a company claims that its light bulbs last an average of 1000 hours. (Realistically) we do not know what the standard deviation of the lives of the bulbs is. We buy 4 light bulbs; they last 850 hours, 900 hours, 650 hours, and 800 hours. We suspect that the company's claim about the mean life of their bulbs is untrue, and want to test it at the 5% significance level. We choose this claim as H_0 , the null hypothesis:

$$H_0 : \bar{x} = \bar{x}_0 = 1000$$

If we reject this hypothesis, we will find that the mean life of the bulbs is in fact less than 1000 hours:

$$\bar{x} < \bar{x}_0 = 1000$$

What we'd really like to do is find evidence that H_0 is false; and if we do, we'll conclude that $\bar{x} < \bar{x}_0 = 1000$. But to find this evidence, we'll start by assuming that H_0 is true, use a left-tailed test, and see what happens.

Assume that $\bar{x} = \bar{x}_0 = 1000$. The value of \bar{X} for our sample is

$$\bar{X} = \frac{\sum X}{n} = \frac{850 + 900 + 650 + 800}{4} = 800.$$

The value of S_X for our sample is

$$\begin{aligned} S_X &= \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}} \\ &= \sqrt{\frac{(850 - 800)^2 + (900 - 800)^2 + (650 - 800)^2 + (800 - 800)^2}{4 - 1}} \\ &\approx 234.52. \end{aligned}$$

Calculate the t-score of this sample:

$$T = \frac{\bar{X} - \bar{x}_0}{S_X/\sqrt{n}} = \frac{800 - 1000}{234.52/\sqrt{4}} \approx -1.706.$$

We look up the critical value of T for the Student distribution with $d = n - 1 = 4 - 1 = 3$ degrees of freedom. We find a critical value of -2.35 , so that we cannot reject the null hypothesis at the 5% level of significance (i.e. with 95% confidence).

example (right-tailed test) Suppose a company claims that its low-calorie cookies contain 100 calories each. We buy 4 low-calorie cookies and measure their energy content; they contain 120 calories, 95 calories, 115 calories, and 110 calories. We suspect that the company's cookies contain an average of more than 100 calories each, and want to test this at the 5% significance level. We choose

$$H_0 : \bar{x} = \bar{x}_0 = 100.$$

If we reject this hypothesis, we will conclude that the mean number of calories per cookie is in fact more than 100 calories:

$$\bar{x} > \bar{x}_0 = 100$$

We do not know the standard deviation of the whole population of cookies, so we will use t-scores, and a right-tailed test.

Assume H_0 – that $\bar{x} = \bar{x}_0 = 100$. The value of \bar{X} for our sample is

$$\bar{X} = \frac{\sum X}{n} = \frac{120 + 95 + 115 + 110}{4} = 110.$$

The value of S_X for our sample is

$$\begin{aligned} S_X &= \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}} \\ &= \sqrt{\frac{(120 - 110)^2 + (95 - 110)^2 + (115 - 110)^2 + (110 - 110)^2}{4 - 1}} \\ &\approx 10.80. \end{aligned}$$

Calculate the t-score of this sample:

$$T = \frac{\bar{X} - \bar{x}_0}{S_X/\sqrt{n}} = \frac{110 - 100}{10.80/\sqrt{4}} \approx 1.852.$$

We consult the table of critical t-scores. We would reject H_0 at the 10% significance level, but not at the 5% significance level. That is, we can conclude that $\bar{x} > 100$ with 90% confidence, but not with 95% confidence.

example (two-tailed test) Suppose an exercise equipment company makes 10-kilogram barbell plates. (Implicit is their claim that the plates weigh an average of 10 kilos.) We inspect 4 plates, and find their weights to be 9.9kg, 10.0kg, 9.8kg, and 9.9kg. We suspect that the company's "10kg" barbell plates do not weigh an average of 10kg, and want to test this at the 10% significance level. This time, we are concerned that the claimed weights are not accurate; we would care both if the mean weight were higher than 10kg, and if the mean weight were lower than 10kg. We choose

$$H_0 : \bar{x} = \bar{x}_0 = 10$$

If this hypothesis were rejected, we would conclude that the mean weight is not equal to 10kg (that is, more than 10kg or less than 10kg):

$$\bar{x} \neq \bar{x}_0 = 10$$

(This is a two-tailed hypothesis test.)

Assume H_0 – that $\bar{x} = \bar{x}_0 = 10$. The value of \bar{X} for our sample is

$$\bar{X} = \frac{\sum X}{n} = \frac{9.9 + 10.0 + 9.8 + 9.9}{4} = 9.9.$$

Since we do not know the standard deviation of the whole population of barbell plates, we will use the sample standard deviation, and the Student distribution with $d = n - 1 = 3$ degrees of freedom. The value of S_X for our sample is

$$\begin{aligned} S_X &= \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}} \\ &= \sqrt{\frac{(9.9 - 9.9)^2 + (10.0 - 9.9)^2 + (9.8 - 9.9)^2 + (9.9 - 9.9)^2}{4 - 1}} \\ &\approx 0.0816. \end{aligned}$$

Calculate the t-score of this sample:

$$T = \frac{\bar{X} - \bar{x}_0}{S_X / \sqrt{n}} = \frac{9.9 - 10}{0.0816 / \sqrt{4}} \approx -2.450.$$

We consult the table of critical t-values. Finding critical values of -2.35 and 2.35 , we reject H_0 at the 10% significance level. That is, we can conclude that $\bar{x} \neq 10$ with 90% confidence.

note We cannot reject H_0 at the 5% significance level. That is, we cannot conclude that $\bar{x} \neq 10$ is true with 95% confidence.

t-scores vs. z-scores T-scores allow us to reject null hypotheses with less confidence than the same numbers as z-scores would. This makes sense, since we use the t-scores when we have less information; that is, when we do not know the standard deviation of X for the whole population. And this is the case in most real-life situations, since if we do not know the mean of X for the whole population, it is unlikely that we know the standard deviation of X for the whole population.

abstract summary of hypothesis testing

- X is a random variable for a large population; its mean is \bar{x}_0 and its standard deviation is s .
- \bar{X} is the sample mean for samples of size n from the population; it is Normal; its mean is \bar{x}_0 and its standard deviation is s/\sqrt{n} .
- S is the sample standard deviation; its mean is s .
- Z is the z-score of \bar{X} ; it is Standard Normal.
- T is the t-score of \bar{X} ; it is Student.
- We reject H_0 at significance level α if the following occurs: the probability of getting a value of the test statistic (Z or T) at least as extreme as the one we obtained from the sample, is less than α .

procedural summary of hypothesis testing

- X is a random variable for a large population; we do not know its mean.
 - There is a claim H_0 saying that the mean \bar{x} is a certain number \bar{x}_0 .
 - We take a sample of size n of the population.
 - The sample mean \bar{X} differs from \bar{x}_0 ; we suspect H_0 is false, and decide whether to use a left-tailed, right-tailed, or two-tailed test.
 - We assume that H_0 is true – that $\bar{x} = \bar{x}_0$.
 - If we do not know the population standard deviation s , we calculate the sample standard deviation S . (If we know s , this is unnecessary.)
 - We calculate the test statistic for our sample – Z if s is known, and T if s is unknown.
 - Find the critical value of the test statistic from the appropriate table. Put a negative sign on the number from the table for a left-tailed test, use the positive number for a right-tailed test, and use both positive and negative numbers for a two-tailed test.
 - If our value of the test statistic is more extreme than the critical value (i.e. higher for a positive number, or lower for a negative number) then we reject H_0 . Otherwise, we fail to reject H_0 .
-
-

©2008 Jason Colwell. All rights reserved.
