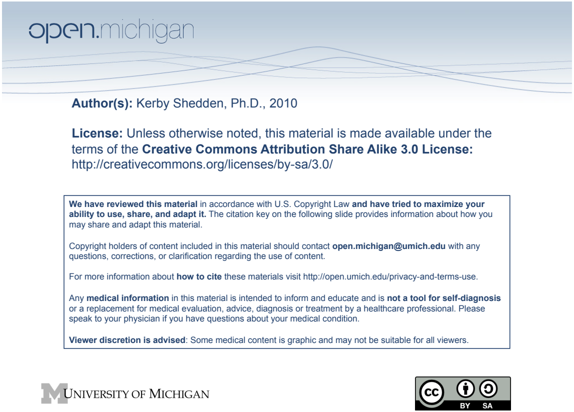
Topics in comparing means of one or two samples From "Research methods: background and review" by Kerby Shedden

Prepared by Edward Ionides

Department of Statistics, University of Michigan

2018-02-11

License and copyright for the complete document



Log transforms (E.g., Homework 6 in STATS 401 W18)

Some quantities that vary over several orders of magnitude are best analyzed on the log scale.

For example, if we observe these values:

$$14, 28, 3, 60, 39, 13, 1, 9, 3, 55$$

We can take $\log_2$ to get their approximate values as powers of 2:

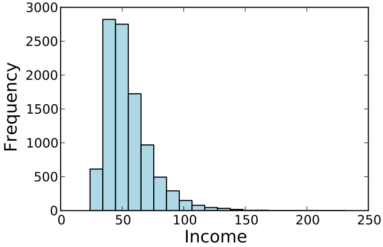
$$3.8, 4.8, 1.6, 5.9, 5.3, 3.7, 0, 3.2, 1.6, 5.8.$$

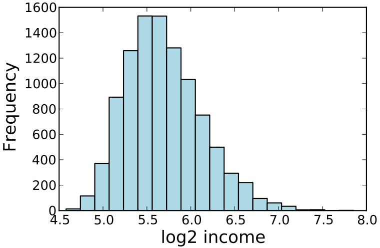
It usually doesn't matter what base is used, since we can convert from one base to another by scaling:

$$\log_b(x) = \log_a(x)/\log_a(b)$$

Symmetrizing effect of log transforms

The log transform symmetrizes right-skewed distributions:





It's common to transform data to make it more symmetric, and usually that's the right thing to do (but don't overdo it...).

Properties of log transforms

Remember the key properties of logarithms:

$$\log(ab) = \log(a) + \log(b) \qquad\qquad \log(a^b) = b\log(a).$$

As a consequence, if we take data $X_1, \ldots, X_n$ and scale it to get $Z_i = cX_i$, then

$$\log(Z_1), \ldots, \log(Z_n) = \log(c) + \log(X_1), \ldots, \log(c) + \log(X_n)$$

Thus changing the units of the original data becomes a shift by $\log(c)$ units for the log-transformed data.

Mean values and log transforms

If we observe data $X_1, \ldots, X_n$ and take a log transform to get $Y_i = \log X_i$, then the mean value of the logged data is:

$$\begin{aligned}
\bar{Y} &= n^{-1} \sum_i Y_i \\
&= n^{-1} \sum_i \log X_i \\
&= n^{-1} \log(X_1 \cdot X_2 \cdots X_n) \\
&= \log \left((X_1 \cdot X_2 \cdots X_n)^{1/n} \right).
\end{aligned}$$

$(X_1 \cdot X_2 \cdots X_n)^{1/n}$ is called the **geometric mean** of the X_i, so we see that the usual (arithmetic) mean of the log transformed data is the log of the geometric mean of the untransformed data.

Log transforms

We generally take the log of positive data that is substantially right skewed. If the data are roughly symmetrically distributed, there is no need to take a log transform, and you cannot take a log transform if any of the data values are less than or equal to zero.

Examples: We generally would log-transform income but not age.

Sampling behavior of the sample mean

Suppose we observe n independent and identically distributed (iid) data points $X_1, \ldots X_n$. Since the X_i are identically distributed, they have a common expected value μ and a common variance σ^2.

The sample mean is

$$\bar{X} = (X_1 + \cdots + X_n)/n = \sum_i X_i/n.$$

The expected value of the sample mean is

$$E\bar{X} = (EX_1 + \cdots + EX_n)/n = (\mu + \cdots + \mu)/n = n\mu/n = \mu,$$

thus the sample mean is an **unbiased** estimate of the population mean.

Sampling behavior of the sample mean

The variance of the sample mean of independent data is

$$\mathrm{var}(\bar{X}) = (\mathrm{var}(X_1)+\cdots+\mathrm{var}(X_n))/n^2 = (\sigma^2+\cdots+\sigma^2)/n^2 = n\sigma^2/n^2 = \sigma^2/n.$$

The standard deviation of the sample mean of independent data is

$$\mathrm{SD}(\bar{X}) = \sigma/\sqrt{n}.$$

This important formula tells us how our precision for estimating μ increases as the sample size increases.

Precision of the sample mean

Terminology: A **statistic** is a summary of raw data. For example, $\bar{X}$ is a summary of $X_1, \ldots, X_n$. The standard deviation of a statistic is sometimes called its **standard error**. So $\sigma/\sqrt{n}$ is the standard error of $\bar{X}$.

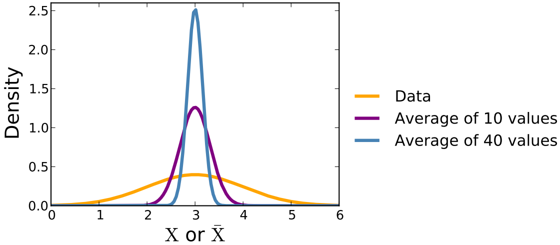
The key standard error result $\mathrm{SD}(\bar{X}) = \sigma/\sqrt{n}$ allows us to understand how the precision of our estimate of the expected value is influenced by different factors:

- Sample size: Since the sample size n occurs in the standard error as $1/\sqrt{n}$, we need to increase the sample size by a factor of four to cut the standard error in half.
- Data variability: The variability of the data is given by σ. We may have the opportunity to reduce σ, say by using a more accurate measurement instrument.

Precision of the sample mean

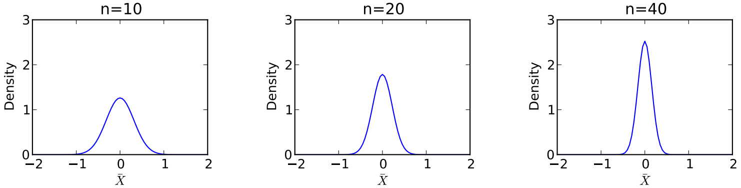
It is very important to distinguish between the variability of the individual data values, and the variability of the average of several data values.

The average has the same expected value as the individual data points, but is less variable than an individual data point.



Sampling behavior of the sample mean

The following graphs show sampling distributions for $\bar{X}$ for a certain population of X_i values. The value of $\mu = EX_i$ is zero.



Note that this is for a particular distribution for X_i. Different distributions will give different sampling distributions, but the reduction in scale as n increases happens for many (but not all) distributions.

The normal distribution

The normal (or Gaussian) distribution is a continuous, symmetric distribution

The **standard normal distribution**, denoted $N(0,1)$ is a normal distribution with mean 0 and variance 1. The normal distribution $N(\mu,\sigma^2)$ has mean μ and variance σ^2.

From the properties of expected values and standard deviations given earlier:

- If Z is standard normal, then $\mu + \sigma Z$ is $N(\mu,\sigma^2)$.
- If Z is $N(\mu,\sigma^2)$, then $(Z-\mu)/\sigma$ is standard normal.

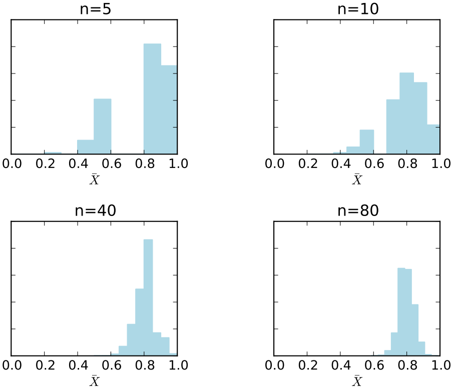
The central limit theorem

The normal distribution plays an important role in statistics because the average of independent values is usually approximately normal for large sample sizes, even if the individual data values are strongly non-normal.

This fact is called the **central limit theorem**.

The central limit theorem

As an illustration of the central limit theorem, suppose $X_1, \ldots, X_n$ are iid with sample space $\{0, 1\}$ and $P(X = 1) = 0.8$. The following histograms show the distributions of 10,000 values of $\bar{X}$ based on sample sizes $n = 5, 10, 40$ and 80.



Calculating normal probabilities

Suppose we have a random variable T which has mean μ and variance σ^2. If we are willing to assume T is normal, how can we calculate $P(T > c)$ for some constant c?

Normal probability tables are available on the web and in most statistics software packages. Often only a table for the standard normal distribution is available, but this is sufficient, since

$$\begin{aligned} P(T > c) &= P((T - \mu)/\sigma > (c - \mu)/\sigma) \\ &= 1 - P(Z \leq (c - \mu)/\sigma). \end{aligned}$$

Thus we can look up the value of $(c - \mu)/\sigma$ in a standard normal probability, table or use a software package, e.g. in R we would use

```
1 - pnorm((c-mu)/s)
```

Normal distribution rules of thumb

- The normal distribution is symmetric – around half of a normal sample lies below the mean and half lies above the mean.
- Around 68% of a normal sample lies within one standard deviation of the mean. For example, if we have 1000 points from a normal distribution with mean 10 and variance 4, around 680 of the points will lie between 8 and 12.
- Around 95% of a normal sample lies within two standard deviations of the mean. Continuing with the previous example, around 950 of the points will lie between 6 and 14.
- Around 99% of a normal sample lies within three standard deviations of the mean. Continuing with the previous example, around 990 of the points will lie between 4 and 16.

Standardizing the sample mean

Suppose we have an iid sample of n observations from a population with mean μ and variance σ^2.

We know that $\bar{X}$ has mean μ and variance σ^2/n (so the standard deviation is $\sigma/\sqrt{n}$).

The standardized sample mean is

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\frac{\bar{X} - \mu}{\sigma}.$$

Example calculation with the normal distribution

Suppose we have a sample $X_1, \ldots, X_{20}$ from a normal population with mean zero. We are told that the probability that $|\bar{X}|$ is greater than 1 is 0.2. What is the standard deviation of the X_i?

Using the fact that the normal distribution is symmetric,

$$\begin{aligned}
0.2 &= P(|\bar{X}| > 1) \\
&= P(\bar{X} > 1) + P(\bar{X} < -1) \\
&= 2P(\bar{X} > 1) \\
&= 2P(\sqrt{20}\bar{X}/\sigma > \sqrt{20}/\sigma) \\
&= 2P(Z > \sqrt{20}/\sigma).
\end{aligned}$$

So $P(Z > \sqrt{20}/\sigma) = 0.1$. Now if we look at a table of the normal distribution, we see that the probability of a standard normal value being bigger than 1.28 is 0.1, so $\sqrt{20}/\sigma = 1.28$, and $\sigma = \sqrt{20}/1.28 \approx 3.49$.

Exercises

1. Suppose we observe a sample of 100 values from a normal population with mean 100 and standard deviation 10. Around how many of the values will be greater than 110?
2. Suppose we observe a sample of 150 values from a normal population with mean 80 and standard deviation 12. Write an R expression that will give the approximate number of values between 75 and 85.
3. Suppose we observe a sample of 200 values from a normal distribution with mean zero. Around 20 of the values that we observe are greater than 50. Approximately what is the standard deviation of the population we are sampling from?

Confidence intervals for the population mean

Suppose we are interested in the expected value EX of a population X. We can estimate EX using $\bar{X}$, but we also would like to know how much uncertainty is present in this estimate.

A confidence interval is an interval of the form $(\mathrm{LB}, \mathrm{UB})$ (where LB and UB stand for the "lower bound" and "upper bound"). This interval will surround the point estimate $\bar{X}$. Since LB and UB depend on the data, they are random quantities.

The coverage probability of this confidence interval is the probability that the target value EX is contained in the interval:

$$P(\mathrm{LB} \le EX \le \mathrm{UB}).$$

Confidence intervals for the population mean

Suppose we have iid data $X_1, \ldots, X_n$ and use $\bar{X}$ to estimate the population mean. We know that $E\bar{X} = \mu$ and $\mathrm{SD}(\bar{X}) = \sigma/\sqrt{n}$. Thus

$$\sqrt{n}(\bar{X} - \mu)/\sigma$$

is standardized, and by the central limit theorem it is approximately standard normal if n is moderate or large. Thus

$$P(-1.96 \le \sqrt{n}(\bar{X} - \mu)/\sigma \le 1.96) = 0.95$$

where 1.96 is the 97.5^{th} percentile of the standard distribution (so the interval (-1.96,1.96) includes 95% of the mass of the standard normal distribution).

Confidence intervals for the population mean

Rearranging terms in

$$P(-1.96 \leq \sqrt{n}(\bar{X} - \mu)/\sigma \leq 1.96) = 0.95$$

yields

$$P(\bar{X} - 1.96\sigma/\sqrt{n} \leq \mu \leq \bar{X} + 1.96\sigma/\sqrt{n}) = 0.95,$$

so $\bar{X} \pm 1.96\sigma/\sqrt{n}$ is a 95% confidence interval (CI) for μ.

Put another way, $\mathrm{LB} = \bar{X} - 1.96\sigma/\sqrt{n}$ and $\mathrm{UB} = \bar{X} + 1.96\sigma/\sqrt{n}$ are the endpoints of the confidence interval.

Confidence intervals for the population mean

If σ is not known, it must be estimated from the data as $\hat{\sigma}$. Then we use the fact that

$$\sqrt{n}(\bar{X} - \mu)/\hat{\sigma}$$

has a t distribution with $n - 1$ degrees of freedom. The 97.5^{th} percentile for this distribution is larger than 1.96, and depends on n (e.g. for n=10 it is 2.26). Thus the CI becomes

$$\bar{X} \pm T_{n-1}\hat{\sigma}/\sqrt{n},$$

where T_{n-1} is the 97.5^{th} percentile of the t_{n-1} distribution.

If the sample size is not too small, the t distribution gives very similar results as the normal distribution.

Confidence intervals for the population mean

The width of a confidence interval is the distance from its lower bound to its upper bound.

A wider confidence interval indicates that we have less knowledge about the true value of an unknown quantity.

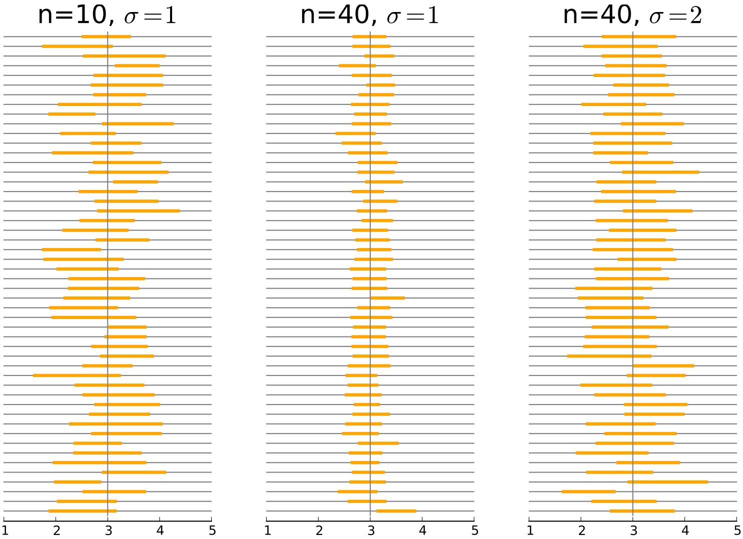
For the approximate 95% CI $\bar{X} \pm 2 \cdot \mathrm{SE}$, the width is

$$\bar{X} + 2 \cdot \mathrm{SE} - (\bar{X} - 2 \cdot \mathrm{SE}) = 4 \cdot \mathrm{SE}.$$

The width of a CI for the population mean when σ is known is approximately $4\sigma/\sqrt{n}$ (since the SE is $\sigma/\sqrt{n}$) and the width is slightly larger when σ is not known (since we would use a larger constant than 2 in that case).

Confidence intervals for the population mean

150 simulated data sets where $EX = 3$:



Exercises

Suppose we observe a data set containing 8 values:

$$8, 4, 5, 4, 6, 3, 4, 5$$

For these values, the sample mean $\bar{X}$ is 4.875.

1. If we are given that the population standard deviation is 1.3, construct a 95% confidence interval for EX.
2. If we do not know the population standard deviation, but we compute the sample standard deviation as 1.55, construct a 95% confidence interval for EX.
3. Suppose we have a data set with 20 values having exactly the same sample mean and standard deviation as the data given above. Construct a 95% confidence interval for EX.

Solutions

1. The interval is (4.875-2*1.3/sqrt(8), 4.875+2*1.3/sqrt(8)), or (3.956, 5.794), or 4.875 +/- 0.919.
2. If the standard deviation is estimated, instead of a multiplier of 2, we should use the 0.975 percentile of a t-distribution with 7 degrees of freedom, which is 2.36. Thus the interval is 4.875 +/- 2.3*1.55/sqrt(8), or 4.875 +/- 1.26.
3. If the SD is treated as known, as in exercise 1, we would get 4.875 +/- 2*1.3/sqrt(20), or 4.875 +/- 0.58. If the SD is treated as an estimate, as in number 2, the interval would be 4.875 +/- 2.09*1.55/sqrt(20), or 4.875 +/- 0.72 (note that 2.09 is the 0.975 percentile of the t-distribution with 19 degrees of freedom).

Exercises

1. Suppose we have two data sets with identical sample means and identical sample variances. One of the data sets contains twice as many values as the other. What is the ratio between the lengths of the 95% confidence intervals for EX in the two data sets?
2. Supppose we have two data sets with identical sample means and identical sample sizes. One of the data sets has twice the sample standard deviation of the other. What is the ratio between the lengths of the 95% confidence intervals for EX in the two data sets?

Exercises

1. The widths of the two intervals will be 4*sigma/sqrt(n) and 4*sigma/sqrt(2*n), so the smaller data set will have a CI that is sqrt(2) times wider than the CI of the larger data set.
2. The widths of the two intervals are 2*sigma/sqrt(n) and 2*(2*sigma)/sqrt(n). Thus the data set with the greater standard deviation will have a CI that is two times wider than the CI of the other data set.

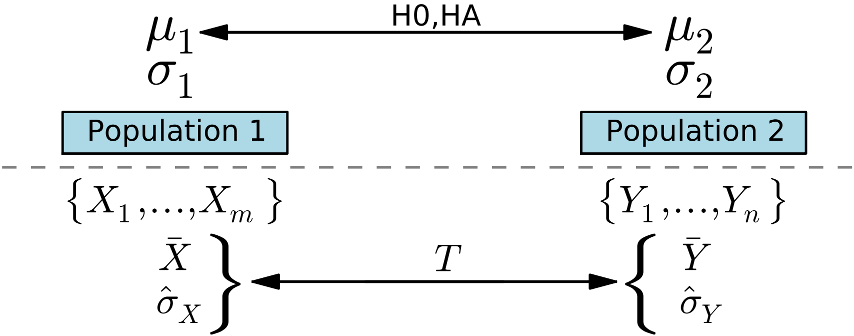
Two-sample comparisons

Suppose we are comparing samples from two populations, and we are interested in whether the two population means are equal.

This is a very common setting. Here are two specific examples:

- We are interested in comparing two treatments for high blood pressure. Each treatment lowers blood pressure on average by a certain amount. We would like to know the difference between the two population "treatment effects."
- Visitors to a web site are shown one of two different advertisements for the same product. We are able to obtain the fraction of the time that the user clicks on the ad (the "click rate"). We would like to know the population value of the difference in click rates.

Two-sample comparisons



Two-sample comparisons

Let $X_1, \ldots, X_n$ denote the data from the first population, and $Y_1, \ldots, Y_m$ denote the data from the second population. Let $\bar{X}$ and $\bar{Y}$ denote the two sample means.

We can estimate the difference in population means using $\bar{X} - \bar{Y}$. If the population means are μ_X and μ_Y, then $\bar{X} - \bar{Y}$ is an unbiased estimate of the mean difference $\mu_X - \mu_Y$, i.e.

$$E(\bar{X} - \bar{Y}) = \mu_X - \mu_Y.$$

Two-sample comparisons

For testing, we can consider the null hypothesis $\mu_X = \mu_Y$.

As a test statistic, we can start with $D = \bar{X} - \bar{Y}$, but we will need to standardize it.

Under the null hypothesis, $ED = 0$. The variance of D is

$$\begin{aligned} \mathrm{var}(\bar{X} - \bar{Y}) &= \mathrm{var}(\bar{X}) + \mathrm{var}(-\bar{Y}) \\ &= \mathrm{var}(\bar{X}) + \mathrm{var}(\bar{Y}) \\ &= \sigma_X^2/n + \sigma_Y^2/m. \end{aligned}$$

Thus the **two-sample Z-statistic**

$$T \equiv \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}}$$

has a standardized distribution under the null hypothesis.

Two-sample comparisons

A test statistic value of $T = 0$ is perfectly consistent with the null hypothesis. Larger values of $|T|$ indicate increasing levels of evidence against the null hypothesis.

The p-value is the null distribution probability that as much evidence or more evidence against the null is observed (hypothetically, if the null hypothesis were true) than was actually observed:

$$p = P_0(|T| \geq |T_{\mathrm{obs}}|)$$

This is the **2-sided p-value** – there are similar expressions for one-sided (right and left tail) p-values.

Two-sample comparisons

Treating the test statistic T as having a standard normal distribution under the null hypothesis, the p-value can be computed as

$$\begin{aligned} P_0(|T| \geq |T_{\mathrm{obs}}|) &= P_0(T \geq |T_{\mathrm{obs}}|) + P_0(T \leq -|T_{\mathrm{obs}}|) \\ &= 2 \cdot P(T \leq -|T_{\mathrm{obs}}|), \end{aligned}$$

where $P_0(T \geq |T_{\mathrm{obs}}|) = P_0(T \leq -|T_{\mathrm{obs}}|)$ due to the symmetry of the normal distribution.

Acceptance/rejection of hypotheses

An alternative framework (to p-values) for hypothesis testing is based on the idea that a null hypothesis is rejected if $|T|$ is sufficiently large, and otherwise is accepted.

In this setting, there are two ways we can make the correct decision and two ways we can make the wrong decision:

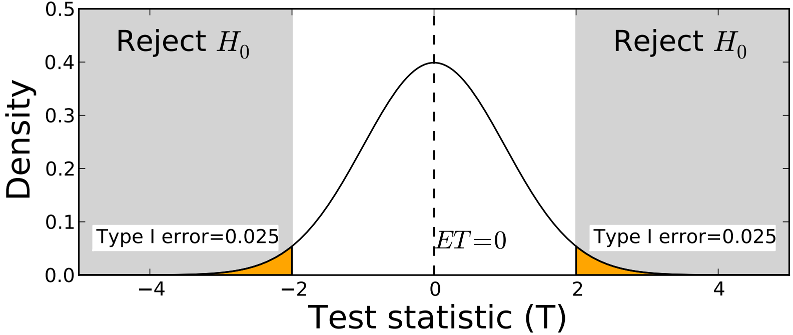
	Accept H_0	Reject H_0
H_0 is true	Correct	Type I error
H_A is true	Type II error	Correct

The level of a hypothesis test is the probability that a type I error occurs.

Acceptance/rejection of hypotheses

Suppose the sampling distribution of T under the null hypothesis is standard normal and we reject the null hypothesis when $|T| > 2$.

There are two ways we can reject – either when $T > 2$ or when $T < -2$. Each of these events has probability 0.025, so the type I error is 0.05.



Two-sample tests with unknown variances

If the variances in the X and Y populations are unknown, we can use the estimated variance

$$\hat{\sigma}_X^2/n + \hat{\sigma}_Y^2/m.$$

In this case the test statistic is approximately t-distributed with the following degrees of freedom:

$$\frac{(\sigma_X^2/n + \sigma_Y^2/m)^2}{(\sigma_X^2/n)^2/(n-1) + (\sigma_Y^2/m)^2/(m-1)}.$$

Power of hypothesis tests

The power of a hypothesis test is the probability of rejecting the null hypothesis when the alternative hypothesis is true.

Here are two equivalent ways of describing the power:

- The power is the probability of getting a p-value below a defined value (usually 0.05) when the alternative hypothesis is true.
- If the test statistic T is standardized under the null hypothesis, the power is the probability that T exceeds the critical value (usually 2) when the alternative hypothesis is true.

The power depends on the sample size and the effect size, which is a measure of how distinguishable the null and alternative hypotheses are from each other.

Power of two-sample hypothesis tests

The effect size is related to two other quantities:

- The raw (unstandardized) effect size is the difference in population means of the two groups being compared.
- The response variability is a summary of the differences among individuals that are not related to their treatment status.

Power is positively related to the raw effect size and is inversely related to response variability.

Example: We have more power to detect a given treatment effect if everyone's blood pressure drops by 5 units, than if half of the subjects have a 10 unit decline and half of the subjects have no decline at all (even though the average decline is 5 units in both cases).

Power of two-sample hypothesis tests

The two-sample Z-test statistic can be rewritten as

$$T = \sqrt{m+n}\frac{\bar{X}-\bar{Y}}{\sqrt{\sigma_X^2/q_X + \sigma_Y^2/q_Y}},$$

where $q_X = n/(n+m)$ and $q_Y = m/(n+m)$ are the proportions of the overall sample drawn from each of the two populations.

Most test statistics can be written in this form

$$\sqrt{\text{total sample size}} \times \text{a “stable value}''$$

Here the total sample size is $m+n$ and the “stable value” is

$$\frac{\bar{X}-\bar{Y}}{\sqrt{\sigma_X^2/q_X + \sigma_Y^2/q_Y}}.$$

Power of two-sample hypothesis tests

By the central limit theorem, we know that under the null hypothesis, as m and n grow, T becomes increasingly well approximated by a standard normal distribution.

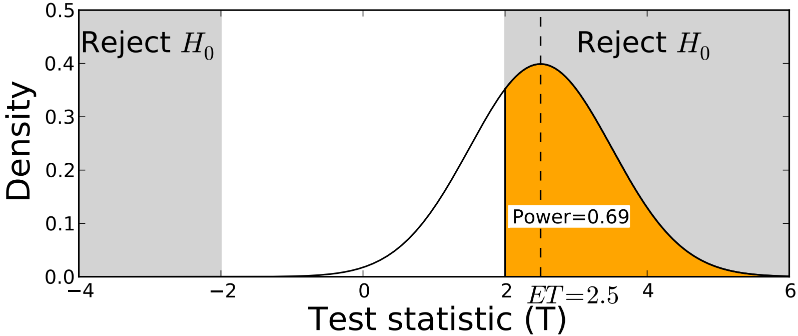
Under the alternative hypothesis, we have

$$E\frac{\bar{X}-\bar{Y}}{\sqrt{\sigma_X^2/q_X+\sigma_Y^2/q_Y}} = \frac{\mu_X-\mu_Y}{\sqrt{\sigma_X^2/q_X+\sigma_Y^2/q_Y}} \equiv \Delta.$$

Thus the expected test statistic is approximately equal to $\sqrt{m+n}\cdot\Delta$ – it continues to grow (at “rate” $\sqrt{m+n}$) as the sample size grows.

Power of two-sample hypothesis tests

Here we see the density function of the test statistic, in a case where its expected value is $ET = 2.5$:



If the test statistic falls in the grey region, we reject the null hypothesis. The power is the probability of this happening, indicated by the orange region.

Note that the standard deviation of the test statistics is one.

Power of two-sample hypothesis tests

For the two-sample problem, we can use $\delta = \mu_X - \mu_Y$ as the raw (unstandardized) effect size.

The power is

$$P(|T| \geq 2) = P(T \geq 2) + P(T \leq -2).$$

The first summand is

$$\begin{aligned}
P(T \geq 2) &= P\left(\sqrt{m+n}\,\frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_X^2/q_X + \sigma_Y^2/q_Y}} \geq 2\right) \\
&= P\left(\sqrt{m+n}\,\frac{\bar{X} - \bar{Y} - \delta}{\sqrt{\sigma_X^2/q_X + \sigma_Y^2/q_Y}} \geq \right. \\
&\qquad \left. 2 - \delta\sqrt{(m+n)/(\sigma_X^2/q_X + \sigma_Y^2/q_Y)}\right).
\end{aligned}$$

Power of two-sample hypothesis tests

Under the alternative hypothesis, $\bar{X} - \bar{Y} - \delta$ has mean zero, so

$$\sqrt{m+n}\frac{\bar{X} - \bar{Y} - \delta}{\sqrt{\sigma_X^2/q_X + \sigma_Y^2/q_Y}}$$

can be treated as being standard normal. Thus

$$P(T > 2) = 1 - P\left(Z \leq 2 - \delta\sqrt{(m+n)/(\sigma_X^2/q_X + \sigma_Y^2/q_Y)}\right),$$

which can be obtained from a standard normal probability table.

A similar calculation can be used to get an expression for $P(T \leq -2)$.

Note that as δ grows and/or $m+n$ grows and/or σ_X and σ_Y shrink, the power gets closer and closer to 1.

Power (determining sample size)

We often need to assess what sample size would be required to detect an effect in a particular situation.

For example, we may be interested in detecting a treatment effect in which a drug lowers a particular quantity by two units on average. We may be also be willing to assume (for the purposes of power analysis) that this quantity varies with a standard deviation of 3 units (for both treated and untreated subjects). Suppose also that we intend to carry out our study using equal numbers of treated and untreated subjects.

In the notation of the preceding slides, we have

- $\mu_X - \mu_Y = 2$ (This is the raw, or unstandardized treatment effect, where X is the untreated group and Y is the treated group.)
- $\sigma_X = \sigma_Y = 3$ (This is the variability of individual subjects' blood pressures around the mean of the group they belong to (either treated or untreated.)
- $q_X = q_Y = 1/2$ (This is our "design decision" to use equal numbers of treated and untreated subjects.)

Power (determining sample size)

So the “stable value” Δ is

$$\Delta = \frac{2}{\sqrt{9/(1/2) + 9/(1/2)}} = 1/3$$

Thus the test-statistic will have mean value equal to

$$\sqrt{m+n} \cdot \Delta = \sqrt{2n}/3$$

To reject the null hypothesis at the usual (two-sided $\alpha = 0.05$) level, we need the test statistic to be greater than 2. Thus we have $\sqrt{2n}/3 > 2$, or $n > 18$ to get 50% power.

If we want 80% power, then we need to solve

$$0.8 = P(T > 2) = P(T - ET > 2 - ET) = P(Z > 2 - \sqrt{2n}/3).$$

Since the $20\mathrm{th}$ percentile of a standard normal distribution is -0.84, this gives us $2 - \sqrt{2n}/3 = -0.84$, so we need $n = 36$ to get 80% power.

Power (determining effect size)

A different type of power analysis comes up when the sample size is fixed, and we are asked to determine what effects can be detected at that sample size.

Suppose we are given that a comparison of two treatments will involve 20 treated subjects, and 40 untreated subjects, and we are willing to assume (for the purposes of power analysis), that the standard deviation within the treated subjects is 1 and the standard deviation within the untreated subjects is 2.

We have

- $\sigma_X = 1$, $\sigma_Y = 2$
- $q_X = 1/3$, $q_Y = 2/3$
- $\sqrt{m+n} = \sqrt{60} \approx 7.75$.

We now need to determine what values of $\delta = \mu_x - \mu_y$ would allow us to reject the null hypothesis with reasonably high frequency.

Power (determining effect size)

The expected value of the test statistic is

$$\sqrt{m+n} \cdot \Delta \approx 7.75 \cdot \frac{\mu_x - \mu_y}{\sqrt{1/(1/3) + 2/(2/3)}} \approx 2.6\delta.$$

To get 80% power, we need

$$0.8 = P(T > 2) = P(T - ET > 2 - ET) = P(Z > 2 - 2.6\delta).$$

Thus we have $2 - 2.6\delta = -0.84$, so $\delta = 1.09$ is required.

Correlation

The correlation coefficient between X and Y is

$$r \equiv \frac{\mathrm{cov}(X, Y)}{\mathrm{SD}(X) \cdot \mathrm{SD}(Y)}.$$

This is sometimes called the “Pearson correlation coefficient.”

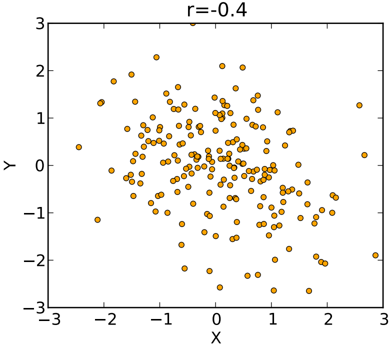
It is a fact that $-1 \leq r \leq 1$, and that $|r| = 1$ only if X and Y are linearly related.

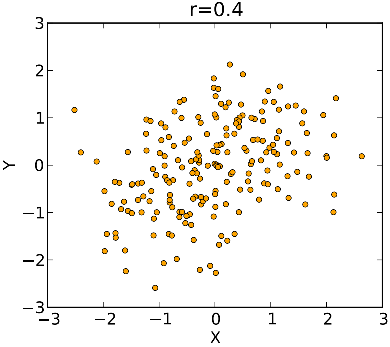
Since r has the same sign as the covariance, it also tends to be positive in situations where X and Y tend to be on the same side of their respective mean values.

If $\mathrm{SD}(X) = 0$ or $\mathrm{SD}(Y) = 0$, r is undefined.

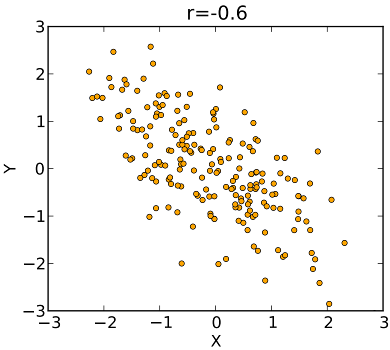
Correlation

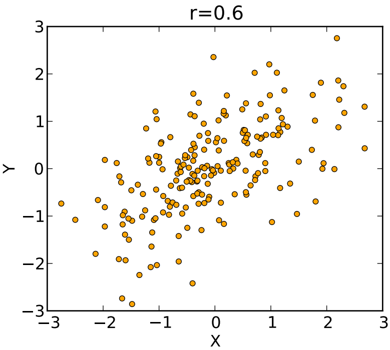
The main application of the correlation coefficient is to measure the degree of linear association between X and Y. The following plots illustrate what is captured by the correlation coefficient.



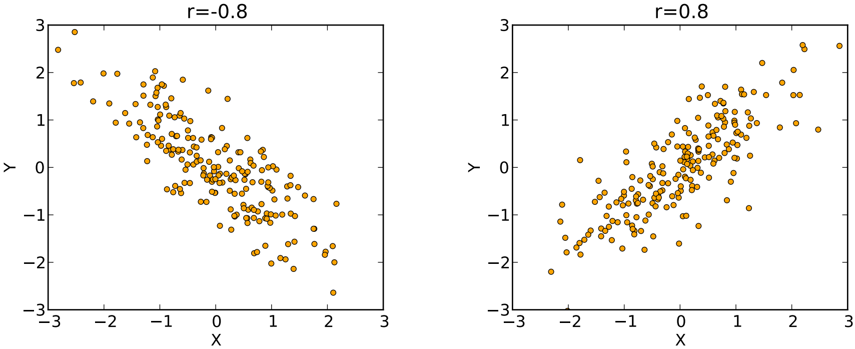


Correlation



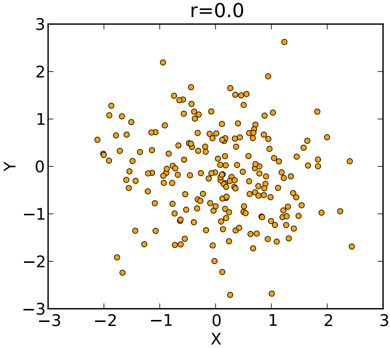


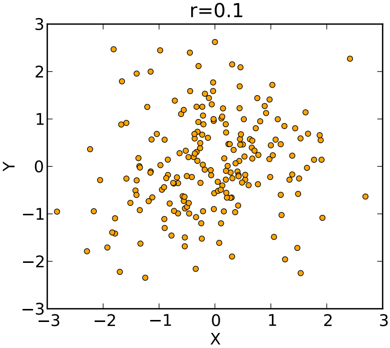
Correlation


r=-0.8

r=0.8

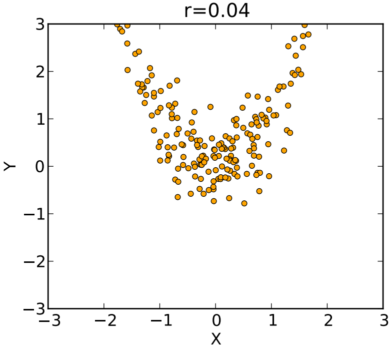
Correlation



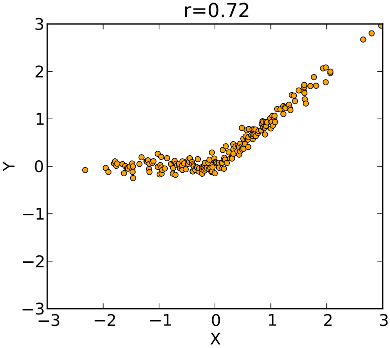


Correlation

Correlation coefficients are mainly useful for picking up a linear trend. The plot on the left shows a strong relationship between X and Y, but the correlation is nearly zero. The plot on the right is not linear, but since it is non-decreasing it still gives a substantial correlation.



r=0.04



r=0.72

The variance of the sample mean for data with unequal variances

If the data are uncorrelated, but have different variances, the covariance matrix has the form:

$$\begin{pmatrix} \mathrm{var}(Y_1) & 0 & 0 & \cdots \\ 0 & \mathrm{var}(Y_2) & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \mathrm{var}(Y_n) \end{pmatrix}$$

In this case, the formula

$$\mathrm{var}(\bar{Y}) = n^{-2} \sum_{ij} \Sigma_{ij}$$

simplifies to

$$\mathrm{var}(\bar{Y}) = n^{-2} \sum_{i} \mathrm{var}(Y_i) = \overline{\sigma_i^2}/n \qquad \sigma_i = \mathrm{SD}(Y_i).$$

The variance of the sample mean for correlated data

If the covariance matrix for a sample of size 3 is

$$\begin{pmatrix} 3 & 3 & 2 \\ 3 & 4 & 3 \\ 2 & 3 & 5 \end{pmatrix},$$

and we calculate the sample mean $\bar{Y} = (Y_1 + Y_2 + Y_3)/3$, the variance of $\bar{Y}$ is

$$\mathrm{var}(\bar{Y}) = n^{-2} \sum_{ij} \Sigma_{ij} = 28/9.$$

Thus the standard error of the sample mean is $\sqrt{28/9} \approx 1.76$.

Loss of precision due to positive dependencies

If the covariance matrix for a sample of size 3 is

$$\left(\begin{array}{ccc} 3 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 5 \end{array}\right),$$

the variance of $\bar{Y}$ is $4/3$, so the standard error is 1.15.

Compare this to the previous slide to see that positive dependencies result in more uncertainty about EY, even when the variances are the same.

Confidence intervals and hypothesis tests for non-iid data

Once we have the variance of $\bar{Y}$, we can form an approximate 95% confidence interval for EY as

$$\bar{Y} \pm 2\widehat{\mathrm{SD}}(\bar{Y}).$$

If we have two samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ such that the X_i are independent of the Y_i, then we can use

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\mathrm{var}(\bar{X}) + \mathrm{var}(\bar{Y})}}$$

as a test statistic, with the remaining steps of the analysis being the same as in the independent case.