Hypotheses and Data

Stephen B. Holt

ROCKEFELLER COLLEGE
OF PUBLIC AFFAIRS & POLICY
UNIVERSITY AT ALBANY State University of New York

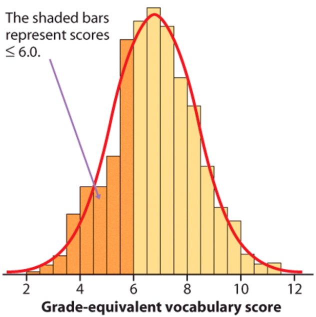
October 30, 2022

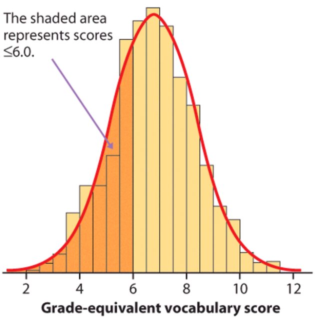
Basics of Hypotheses

- Hypotheses are statements about theoretical relationships between two variables
- Generally, hypotheses flow logically from a theoretical argument
- Example: if wages are determined purely by the marginal product of labor, a higher price for labor will increase unemployment
- When using empirical data to test a hypothesis, we use thresholds to determine the likelihood our observed patterns would occur if the underlying data generating process had no pattern (or a dissimilar pattern)
- Empirical hypothesis testing rests does not tell you anything about what relationships between variables to expect, what those relationships should be, or what hypothesized relationships to even test - that is the role of theory

Density Curves and Distributions

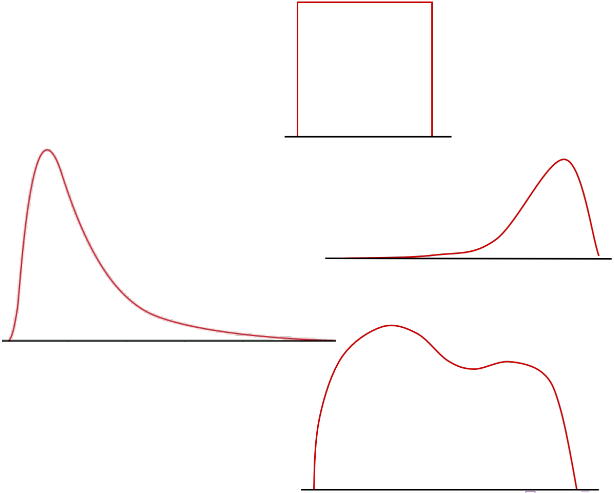
- Distributions describe the range and frequency of observations of a particular variable.
- While a histogram can document the number of observations at a given value of a variable, a density curve plots the area of a range of values the represents the proportion of all values within that range.
- The full area under a curve, consequently, is 100% of observations.





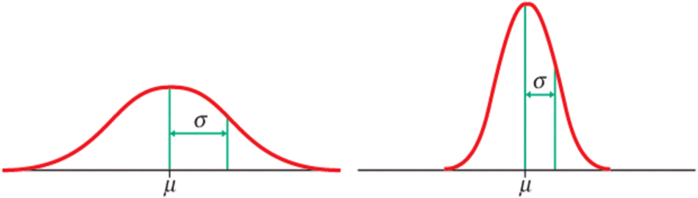
Density Curves and Distributions

Density curves can come in all shapes and sizes, some better understood than others.



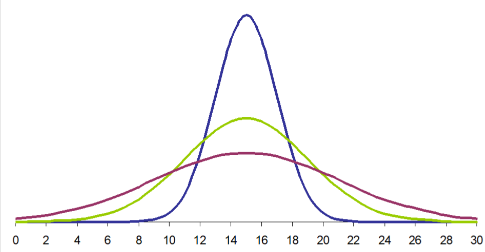
Normal Distributions

Normal – or Gaussian – distributions are a family of symmetrical, bell-shaped density curves defined by a mean μ (mu) and a standard deviation σ (sigma): $N(\mu, \sigma)$.

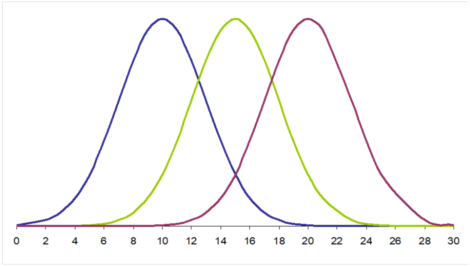


Normal Distributions

Distributions with the same mean and different spreads:

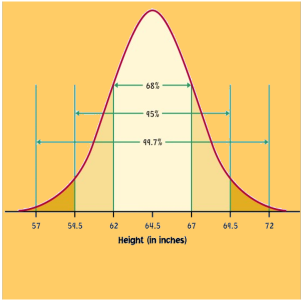


Distributions with different means and the same spreads:



Special Properties of Normal Distributions

- Normal distributions have mathematical properties that we use a lot in statistical inference. The 68-95-99.7 rule allows us to assess relative points in a distribution in terms of standard deviation. The rule is:
- 68% of observations fall within 1 standard deviation (σ) of the mean (μ)
- 95% of observations fall within 2 standard deviations (σ) of the mean (μ)
- 99.7% of observations fall within 3 standard deviations (σ) of the mean (μ)
- Note that μ and σ will always be used to refer to the (usually unknown) population mean and s.d., while $\overline{X}$ and s_x will always refer to a sample mean and s.d.



What are sampling distributions?

The **sampling distribution of a statistic** is the distribution of all possible values taken by the statistic when all possible samples of a fixed size n are taken from the population. It is a theoretical idea — we do not actually build it.

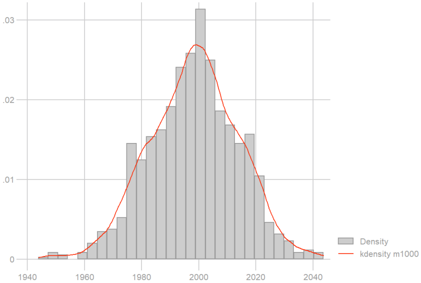
The sampling distribution of a statistic is the **probability distribution** of that statistic.

Sampling distribution of the sample mean

If we **were** to build the sampling distribution, we would take many, many samples of a given sample size (n) from a population with mean μ and standard deviation σ.
Some of these samples will have a mean above the population mean μ and some will be below. If we plot the mean of each sample in a distribution, we will have the sampling distribution of μ. Note, this is different than a distribution of x from a single sample.

Say we have a population of 80,000 people and $\mu = 2000$ sq. ft. houses. We draw a bunch of simple random samples (SRS) and calculate the mean sq. ft.:

1. SRS size 1000 $\rightarrow \overline{x} = 1985$
2. SRS size 1000 $\rightarrow \overline{x} = 1999$
3. SRS size 1000 $\rightarrow \overline{x} = 2010$



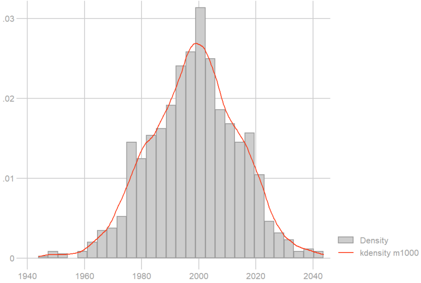
For any population with mean μ and standard deviation σ:

- The mean (or center) of the sampling distribution of $\overline{x}$ is equal to the population mean μ (the **law of large numbers**)
- The standard deviation of the sampling distribution is $\frac{\sigma}{\sqrt{n}}$, where n is the sample size (the **central limit theorem**).

These two properties of the sampling distribution are dictated by the **law of large numbers** and the **central limit theorem**.

Say we have a population of 80,000 people and $\mu = 2000$ sq. ft. houses. We draw a bunch of simple random samples (SRS) and calculate the mean sq. ft.:

1. SRS size 1000 → $\overline{x} = 1985$
2. SRS size 1000 → $\overline{x} = 1999$
3. SRS size 1000 → $\overline{x} = 2010$



Law of Large Numbers

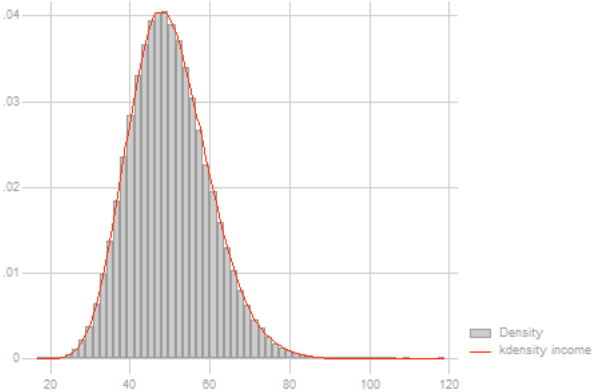
Here, I will use data from a simulation to demonstrate the law of large numbers and central limit theorem. The **law of large numbers** dictates that as a sample size gets larger and larger, the average of the sample will be the same as the average of the population.

In other words, the larger the sample we collect information from, the more likely the samples will have roughly the same average. This is referred to as an estimate's **consistency**.

Since sampling distributions are theoretically built with infinite samples, or very large numbers of samples of the same variable from the same population, the law of large numbers dictates that the mean of their distribution will be equal to the mean of the population.

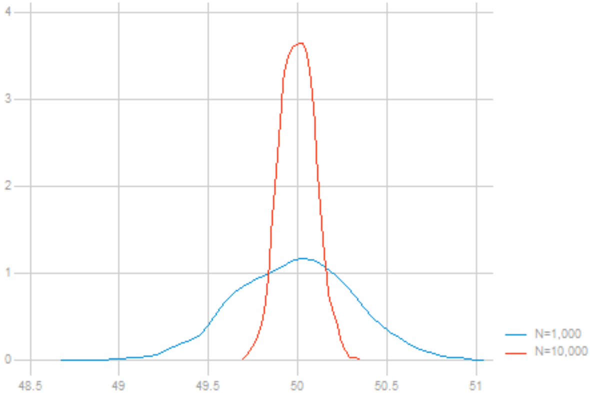
Example of Law of Large Numbers

I created a fake population of 800,000 people and data on their incomes (in \$1000s) with a mean of \$50K. Unlike most real-world situations, this means we **know** the true value of μ and σ. In this case, $\mu = 49.99$ and $\sigma = 9.99$. Below is the distribution from the population:



Example of Law of Large Numbers

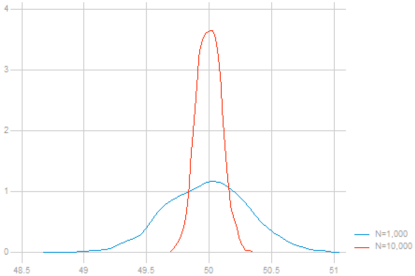
I simulated taking a thousand random samples with sample sizes of 1,000 observations and 10,000 observations and calculated the average of each sample. The distributions below show the sample means of all 1,000 samples for each sample size:



Example of Law of Large Numbers

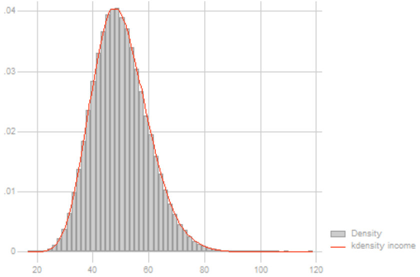
A few things to note about what the simulation is showing:

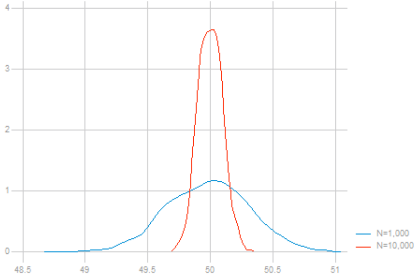
1. The range for the population is from 20 to 120 while the sample distributions both range from 1.3 below to 1.01 above the μ of 50 (with a smaller range for the larger sample).
2. The high peak for samples of 10,000 people shows that a much larger proportion of samples have an estimated mean ($\overline{x}$) equal to the underlying population mean (μ).
3. Larger sample sizes yield means that converge closer and closer to the underlying population average.



Example of Central Limit Theorem

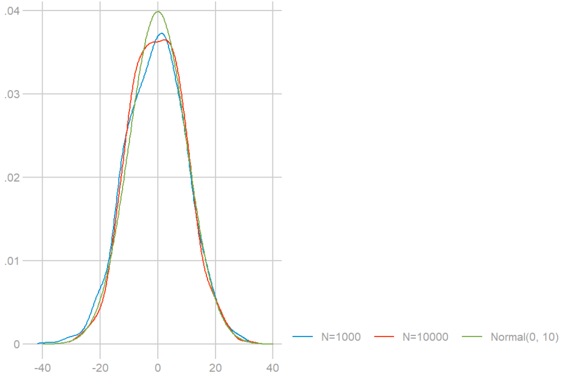
The central limit theorem provides the basis for estimating standard deviations from samples and performing hypothesis tests. The central limit theorem states that the sampling distribution of means will always be approximately normally distributed, even when the variable is not normally distributed in the population. Notice that our population in the previous example had a long tail to the right, but the sampling distributions did not:





Example of Central Limit Theorem

Rescaling and recentering both distributions shows that the sampling distribution for both 1,000 observation samples and 10,000 observations samples is nearly identical to a normal distribution.



Statistical Confidence

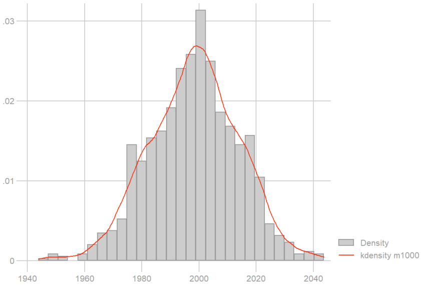
Although the sample mean is a unique number for any particular sample, if you pick a different sample you will probably get a different sample mean.

In fact, you could get many different values for the sample mean, and virtually none of them would actually equal the true population mean, μ. But due to the Law of Large Numbers, we know it will be close, particularly when we take larger samples.

$\mu = ?$; $\sigma = 500$. We draw a bunch of simple random samples (SRS) and calculate the mean sq. ft.:

1. SRS size 1000 $\rightarrow \overline{x} = 1985$
2. SRS size 1000 $\rightarrow \overline{x} = 1999$
3. SRS size 1000 $\rightarrow \overline{x} = 2010$

Sampling distribution of $\overline{x}$



Confidence Intervals

We can use our knowledge that the sampling distribution will be 1) centered around the population mean because of the law of large numbers and 2) will have a normal distribution because of the central limit theorem to help us assess the quality and confidence we have that our sample average accurately represents the population average. We can do this using the standard deviation of the sampling distribution to calculate what's known as a **confidence interval**.

The **confidence interval** is a range of values with an associated probability or confidence level C. The probability quantifies the chance that the interval contains the true population parameter.

$\mu = ?$; $\sigma = 500$; $\sigma_{\overline{x}} = \frac{500}{\sqrt{1000}} = 15.81$. We draw a bunch of simple random samples (SRS) and calculate the mean sq. ft.:

1. SRS size 1000 $\rightarrow \overline{x} = 1985 \rightarrow \overline{x} \pm 31.62 \rightarrow 1985 \pm 31.62$
2. SRS size 1000 $\rightarrow \overline{x} = 1999 \rightarrow \overline{x} \pm 31.62 \rightarrow 1999 \pm 31.62$
3. SRS size 1000 $\rightarrow \overline{x} = 2010 \rightarrow \overline{x} \pm 31.62 \rightarrow 2010 \pm 31.62$

Confidence Intervals

A confidence interval can be calculated as $\overline{x} \pm m$.
m refers to the **margin of error** or the z-score the analyst is choosing for C confidence levels. Thus, $m = z * \frac{\sigma}{\sqrt{n}}$.

As an example, mean of 120 and margin of error of 6:
$120 \pm 6 \rightarrow$ confidence interval ranges from 114 to 126.

You interpret this as: 95% of sample means of samples of n size will be between 114 and 126 or I am 95% confident that the true population mean is between 114 and 126.

A confidence level C (in %) indicates the probability that the μ falls within the interval. It represents the area under the normal curve within $\pm m$ of the center of the curve.

Link between Confidence Level and Margin of Error

The confidence level C determines the value of z*.
The margin of error also depends on z*.

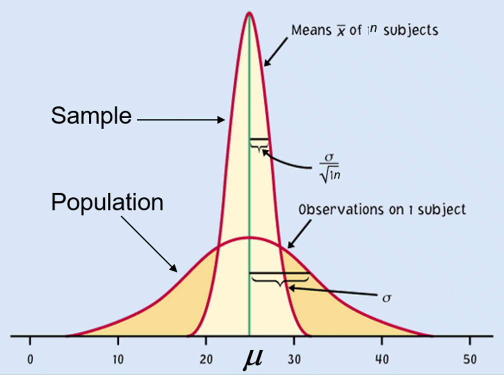
The C is the percent of the distribution you would like to fall between the two end points you plan to calculate. That percent provides a measure of how likely you are to have an estimated average that is unrealistic or far from the true average. Once you've decided on the percentage for C, you choose the z-score associated with C and use the formula $m = z * \frac{\sigma}{\sqrt{n}}$ to calculate the end points of the confidence interval.

A higher C means more confidence that the true mean is within the interval, but it also means a wider interval (less precise estimates).

A lower C means less confidence in our estimate, but a narrower interval (more precise estimates).
We generally prefer to be cautious. A rule of thumb is to use 95% confidence intervals (i.e., a z-score of 2).

Implications

We *don't need* to take endless random samples to "rebuild" the sampling distribution and find μ at its center. We only need one SRS of size n and we can use the properties of the sampling distribution of means to infer the population mean μ. The central limit theorem and law of large numbers let us have pre-existing knowledge of the sampling distribution without building it from scratch each time.


Means $\bar{x}$ of 1n subjects
Sample
$\frac{\sigma}{\sqrt{1n}}$
Population
Observations on 1 subject
σ
0
10
20
μ
30
40
50

Reasoning of Significance Tests

- We have seen that the properties of the sampling distribution of $\overline{x}$ help us estimate a range of likely values for population mean μ
    - Centered on μ
    - Normal distribution with a narrower measure of spread than the population
- Example: You are in charge of ensuring safe streets. You randomly sample speeds of drivers on 4 parts of a main avenue.
- The average speed in your sample was 48 mph. Obviously, we cannot expect every section of the avenue to have the same travel speeds. Thus,
    - Is the somewhat higher speed in your sample due to chance variation?
    - Is it evidence that the city should consider more aggressive enforcement or changes to the streetscape?

Stating Hypotheses

A test of statistical significance tests a specific hypothesis using sample data to decide on the validity of the hypothesis.

In statistics, a hypothesis is an assumption or a theory about the characteristics of one of more variables in one or more populations.

Example: What you want to know: Does the street need more attention for safety reasons?

That same question reframed statistically: Is the population mean μ for the distribution of speeds traveled on the road equal to 35 mph (i.e., the speed limit)?

Stating Hypotheses

The null hypothesis is a very specific statement about a parameter of the population(s). It is labeled H_0.

The alternative hypothesis is a more general statement about a parameter of the population(s) that is exclusive of the null. It is labeled H_a.

Example: Travel speeds on main avenue:

H_0: $\mu = 35 mph$ (μ is the average speed of travelers on the road)
H_a: $\mu \neq 35 mph$ (μ is either larger or smaller)

One-sided and Two-sided Tests

- A two-tail or two-sided test of the population mean has these null and alternative hypotheses:
  - $H_0 : \mu =$ [a specific number] $H_a : \mu \neq$ [a specific number]
- A one-tail or one-sided test of a population mean has these null and alternative hypotheses:
  - $H_0 : \mu =$ [a specific number] $H_a : \mu <$ [a specific number]
  - $H_0 : \mu =$ [a specific number] $H_a : \mu >$ [a specific number]

The FDA tests whether a generic drug has an absorption extent similar to the known absorption extent of the brand-name drug it is copying. Higher or lower absorption would both be problematic, thus we test:
$H_0 : \mu_{generic} = \mu_{brand}$ $H_a : \mu_{generic} \neq \mu_{brand}$ two-sided

How to Choose?

What determines the choice of a one-sided versus a two-sided test is what we know about the problem before we perform a test of statistical significance.

Example: A health advocacy group tests whether the mean nicotine content of a brand of cigarettes is greater than the advertised value of 1.4 mg.

Here, the health advocacy group suspects that cigarette manufacturers sell cigarettes with a nicotine content higher than what they advertise in order to better addict consumers to their products and maintain revenues. Thus, this is a one-sided test: $H_0 : \mu = 1.4\text{mg}$ $H_a : \mu > 1.4\text{mg}$

It is important to make that choice before performing the test or else you could make a choice of “convenience” or fall into circular logic.

In practice, **we want to exercise caution** - a two-sided t-test will thus be preferred in most instances.

The P-Value

The speed of drivers in your city has a known standard deviation of 10 mph.

H_0: $\mu = 35 mph$ versus H_a: $\mu \neq 35 mph$

Tests of statistical significance quantify the chance of obtaining a particular random sample result if the null hypothesis were true. This quantity is the **P-value**.

This is a way of assessing the “believability” of the null hypothesis, given the evidence provided by a random sample.

Interpreting The P-Value

With a small p-value we reject H_0. The true property of the population is significantly different from what was stated in H_0.

Thus, small P-values are strong evidence AGAINST H_0

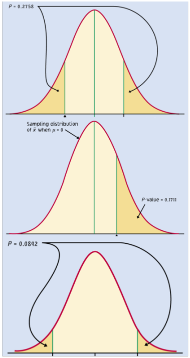
But how small is small...?

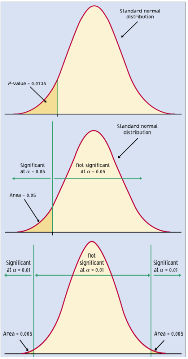
Interpreting The P-Value

P = 0.2758

P = 0.1711

P = 0.0892





P = 0.0735

P = 0.05

P = 0.01

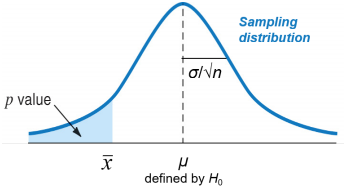
When the shaded area becomes very small, the probability of drawing such a sample at random gets very slim. Oftentimes, a P-value of 0.05 or less is considered significant: The phenomenon observed is unlikely to be entirely due to chance event from the random sampling.

Tests for a Population Mean

The p-value is the area under the sampling distribution for values at least as extreme, in the direction of H_a, as that of our random sample.

Again, we first calculate a z-value and then use a z-table:

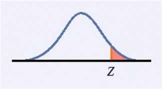
$$z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$



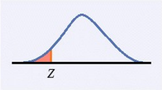
P-value in one-sided and two-sided tests

One-sided
(one-tailed) test

$H_a\colon \mu > \mu_0$ is $P(Z \geq z)$

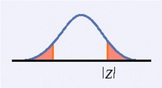


$H_a\colon \mu < \mu_0$ is $P(Z \leq z)$



Two-sided
(two-tailed) test

$H_a\colon \mu \neq \mu_0$ is $2P(Z \geq |z|)$



To calculate the P-value for a two-sided test, use the symmetry of the normal curve. Find the P-value for a one-sided test and double it.

Does the street need attention for speeding?

- H_0: $\mu = 35mph$ versus H_a: $\mu \neq 35mph$
- What is the probability of drawing a random sample such as yours if H_0 is true?
  $\overline{x} = 48mph$ $\sigma = 10mph$ $n = 4$
  $$z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow \frac{48 - 35}{\frac{10}{\sqrt{4}}} \rightarrow 2.4$$
  From a z-table, the area under the standard normal curve to the left of z is 0.9918.
  To the right, this would be 1 - 0.9918 or 0.0082.
  For a two-sided test, we would multiply by 2 (2×0.0082) for a p-value of 0.0164.
  The probability of getting a random sample average this far above μ is so low that we can safely reject H_0.
  We would conclude that the street does need some safety attention.

Steps for Tests of Significance

1. State the null hypotheses H_0 and the alternative hypothesis H_a.
2. Calculate value of the test statistic.
3. Determine the P-value for the observed data.
4. State a conclusion.

The significance level: α

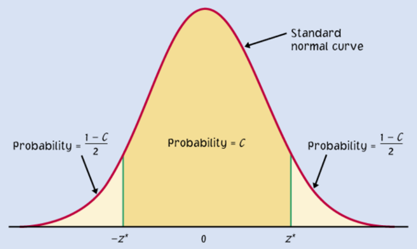
The significance level, α, is the largest P-value tolerated for rejecting a true null hypothesis (how much evidence against H_0 we require). This value is decided arbitrarily before conducting the test.

- If the P-value is equal to or less than α ($P \leq \alpha$), then we reject H_0.
- If the P-value is greater than α ($P > \alpha$), then we fail to reject H_0.

Example: The speed sample p-value was 0.0164. If α had been set to 1%, we would fail to reject the null and the p-value would be insignificant. If α had been set to 5%, we would reject the null and the p-value would be significant.

Confidence intervals and Inference

Because a two-sided test is symmetrical, you can also use a confidence interval to test a two-sided hypothesis. In a two-sided test, C = 1 - α.



Example: $\sigma = 10$ mph: H_0: $\mu = 35 mph$ versus H_a: $\mu \neq 35 mph$
Sample average 48 mph. 95% CI for μ = 48 mph ±
$1.96 \times \frac{10}{\sqrt{4}} \rightarrow 48 mph = \pm 9.8 mph$
35 mph is not in the 95% CI (38.2 to 57.8 mph). Thus, we reject H_0.