Hypothesis Test Notes
P-value, Significance Levels & Simulation

Hypothesis Test: Using Random Sample data to decide between two possible views about the population.

Review: Test Statistics

A test statistic allows us to measure whether the sample data significantly disagrees with the null hypothesis or not.

- Test Statistic falls in the tail determined by critical value → Sample data significantly disagrees with null hypothesis.
- Test Statistic does NOT fall in the tail determined by critical value → Sample data does NOT significantly disagree with null hypothesis.

Big Problem: Sampling Variability!!

Principle of Sampling Variability: Random samples are usually different and random sample statistics are usually very different than the population parameter.

Note: Some people refer to Sampling Variability as "random chance".

Deciding Between Two Options

Option 1: Is our random sample data different than the population parameter (null hypothesis) because all random samples are different (sampling variability)? In which case the population parameter and null hypothesis might be correct.

OR

Option 2: Is our random sample data different than the population parameter (null hypothesis) because the population parameter and the null hypothesis is wrong.

Dealing with the Two Options

Key Question: Could my sample data be different than H_0 because of sampling variability?
(Could the sample data have occurred by random chance?)

Think of sampling variability (random chance) as a confounding variable. In order to show that the population parameter and the null hypothesis is wrong (option 2), we have to make sure that the reason the sample is different is not sampling variability.

In other words we have to make sure option 1 is not correct (or at least highly unlikely), to be able to say that population parameter and the null hypothesis is probably wrong (option 2). In that case, we "Reject the Null Hypothesis".

If we cannot rule out option 1, we will never know for sure which option is probably correct. The sample data disagrees with the null hypothesis so H_0 might be wrong. But the sample data might just be different because of sampling variability indicating H_0 might be correct.

What a Mess!!!!

P-value to the rescue!!

P-value can help us understand sampling variability and decide between the two options.

Definition

P-value : The probability of getting the sample data or more extreme because of sampling variability if the null hypothesis is true.

Some Stat Books write the definition this way.

P-value : The probability of getting the sample data or more extreme by random chance if the null hypothesis is true.

Probability & Logic principle: If the probability of an event happening is very low, but the event keeps happening, then we should look for a different explanation. Our assumption about how that event works might be wrong.

Assumption: Suppose the population parameter in H_0 is correct
The P-value calculates the probability of getting the sample data because of sampling variability based on that assumption.

Low P-value *(Sampling variability is unlikely.)*

If the P-value is very low (close to zero) , then the sample data probably did not happen by sampling variability (random chance). A low P-value rules out sampling variability. Since the sample data probably did not occur by sampling variability, the only other option is that the null hypothesis must be wrong. When that happens we say we "Reject the Null Hypothesis". It also implies that the alternative hypothesis is probably correct.

High P-value *(Could be sampling variability)*

If the P-value is high, then the sample data could have occurred just because of sampling variability. Since sampling variability might or might not be involved, we will not be able to decide whether the null hypothesis is right or wrong. Which means we also will not be able to decide whether the alternative hypothesis is right or wrong. When this happens, we say we "Fail to Reject the Null Hypothesis". We cannot decide between the null and alternative hypotheses.

Important Notes:

- *"Failing to reject H_0" does not mean that H_0 is true!! It means we cannot tell if the population parameter in H_0 is right or wrong. Sampling Variability has struck again.*

- *A low P-value occurs when the sample value significantly disagrees with the population value in the null hypothesis. In other words a low P-value corresponds with a large test statistic. Both mean that the population parameter in the null hypothesis is probably wrong.*

- *A high P-value occurs when the sample value is pretty close to the population value in the null hypothesis. In other words a high P-value corresponds with a small test statistic. The population parameter might be correct, but because of sampling variability we cannot tell.*

Significance Levels

Sometimes a P-value might be borderline. Remember we want the P-value to be low (close to zero) to insure that the sample data did not occur because of sampling variability. But how low do we need it?

Significance Levels *(also called "alpha levels")*

α (Greek Letter Alpha)

Significance levels (α) are a number we can compare the P-value too. We will also see later they are also associated with avoiding certain types of errors in statistics.

Remember confidence levels? Significance levels (α) are the opposite of confidence levels $(1-\alpha)$. If you want to be 95% confident for example the significance level would be 100%-95% = 5%. This is the most common significance level used.

Common Confidence Levels and Significance Levels.

Confidence Level $(1-\alpha)$	Significance Level (α)
90% (0.90)	10% (0.10)
95% (0.95)	5% (0.05)
99% (0.99)	1% (0.01)

So before you do your hypothesis test you should choose which significance level you want to use. If you are unsure, use 5% as this is the most common.

Using Signficance Levels

If the P-value ≤ significance level, Reject the null hypothesis.
(P-value is low enough to rule out sampling variability)

If the P-value > significance level, Fail to reject the null hypothesis.
(P-value is too high. Sample data may have occurred because of sampling variability.)

P-value Summary

Low P-value (Less than or equal to the Significance Level)

- Sample data significantly disagrees with the null hypothesis.
  (Sample data significantly disagrees with the population parameter.)
- Sample data probably did <u>not</u> happen because of sampling variability.
  (The sample data probably did not happen by random chance.)
- Reject H_0

High P-value (Higher than the Significance Level)

- Sample data does NOT significantly disagree with the null hypothesis.
  (Sample data close to the population parameter.)
- Sample data could have happened because of sampling variability.
  (The sample data could have happened by random chance.)
- Fail to reject H_0 *(This does NOT mean H_0 is correct! It means we don't know.)*

Example 1

We used to think that the population mean average typing speed for all U.S. adults is about 40 (words per minute), but now we think the average typing speed has decreased. We took a large random sample in order to test this claim. Our sample mean was 38 (words per minute). The P-value was 0.216 and the significance level was 5%.

H_0 : μ = 40
H_A : μ < 40 (CLAIM)

Convert the P-value into a percentage.

Compare the P-value to the significance level. Is the P-value large or small?

Does the sample data significantly disagree with the null hypothesis?

Could the sample data have occurred because of sampling variability?

Should we reject H_0 or fail to reject H_0 ? Explain why.

Convert the P-value into a percentage. P-value = 0.216 = 21.6%

Compare the P-value to the significance level. Is the P-value large or small?
P-value (21.6%) higher than significance level (5%). This is a high P-value!! (Bad)

Does the sample data significantly disagree with the null hypothesis?
Sample data does NOT significantly disagree. They are relatively close.
Remember a high P-value means that the test statistic will be smaller than the critical value.

Could the sample data have occurred because of sampling variability?
Yes. The P-value is high, meaning if H_0 is true, there was a 21.6% probability of getting the sample data or more extreme because of sampling variability.

Should we reject H_0 or fail to reject H_0 ? Explain why.
Fail to reject H_0. P-value is high and sampling variability might be involved.
We will not be able to tell if the null is right or wrong.

Example 2

A pharmaceutical company is developing a new medicine to to help people with diabetes. They want to see if the medicine will help more than 50% of people that take it. Unbiased random sample data gave the following P-value. They used a 1% significance level.

P-value = 8.74×10^{-4}

H_0: $\pi \leq 0.5$
H_A: $\pi > 0.5$ (CLAIM)

Convert the P-value into a percentage.

Compare the P-value to the significance level. Is the P-value large or small?

Does the sample data significantly disagree with the null hypothesis?

Could the sample data have occurred because of sampling variability?

Should we reject H_0 or fail to reject H_0 ? Explain why.

Convert the P-value into a percentage. Notice the P-value is in scientific notation. Move the decimal four places to the left. P-value = 8.74×10^{-4} = 0.000874 = 0.0874%

Compare the P-value to the significance level. Is the P-value large or small?
P-value (0.0874%) lower than significance level (1%). This is a low P-value!! (Happy!!)

Does the sample data significantly disagree with the null hypothesis?
Sample data significantly disagrees with H_0. Remember a low P-value means that the test statistic will be larger than the critical value.

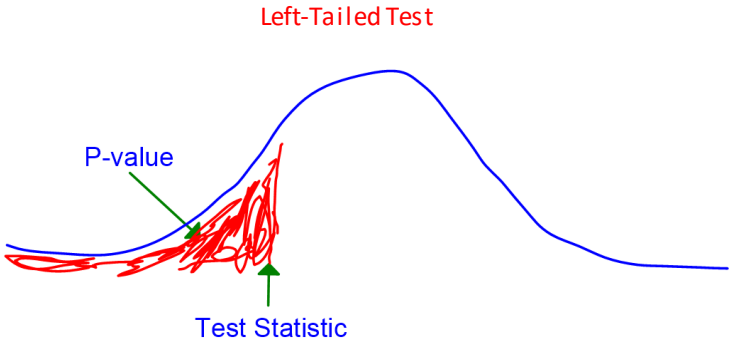
Could the sample data have occurred because of sampling variability?
Probably not. The P-value is low, meaning if H_0 is true, there was only 0.0874% (close to zero) probability of getting the sample data or more extreme because of sampling variability.

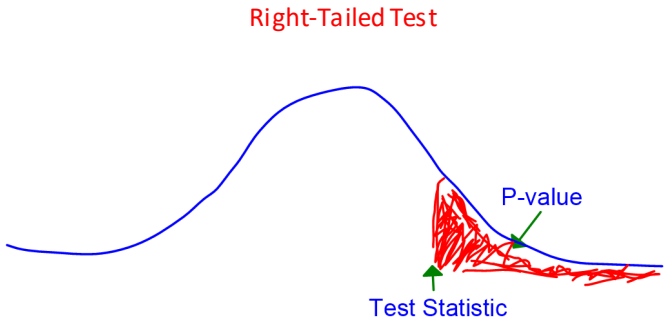
Should we reject H_0 or fail to reject H_0 ? Explain why.
Reject H_0. P-value is low and rules out sampling variability. Yes!!! Since sampling variability is not involved, the null hypothesis is probably wrong.

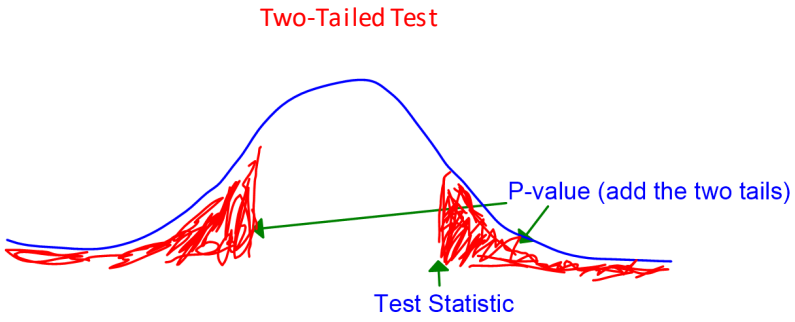
Calculating P-values: Simulation and Traditional Approaches

Traditional Approaches

Before computers were invented, statisticians used curves and test statistics to calculate approximate P-values. Finding what curve best fits the sampling distribution was very important. They had to insure the shape of the sampling distribution would fit the curve they were using, so traditional approaches are tied to assumptions. The one-population and two-population hypothesis test assumptions are about the same as what we learned for confidence interval estimates. This technique of using the test statistic to find the P-value is common to this day in computer programs.





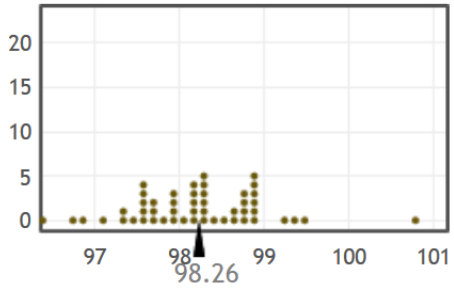


Example: The population mean average body temperature has long thought to be 98.6°F. Scientists now claim that the population mean average is less than 98.6°F. Use the following random sample data to test the claim. Use a 5% significance level. Assume the data met all of the assumptions.

H_0 : μ = 98.6 °F
H_A : μ < 98.6 °F (CLAIM)

Original Sample

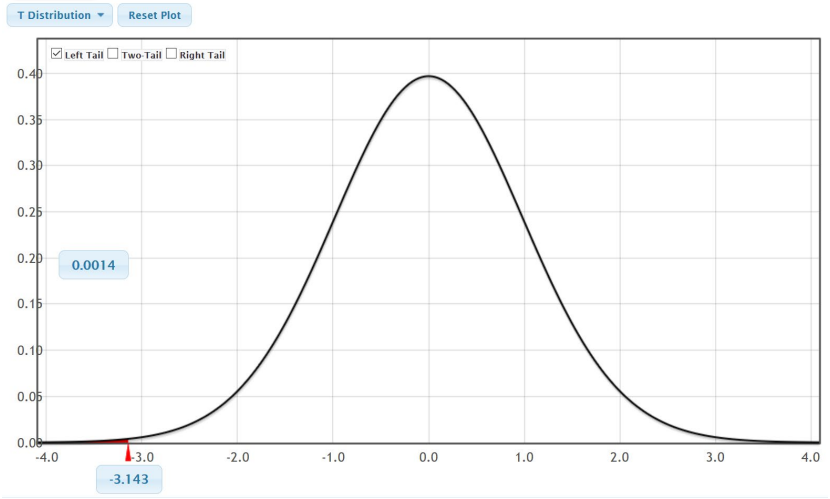
n = 50, mean = 98.26
median = 98.2, stdev = 0.765



Traditionally, a statistician would first calculate the T-test statistic.

$$T = \frac{(\bar{x} - \mu)}{\left(\frac{s}{\sqrt{n}}\right)} = \frac{(98.26 - 98.6)}{\left(\frac{0.765}{\sqrt{50}}\right)} = -3.143$$

They would then use a normal T-curve with degrees of freedom 49 to calculate the approximate P-value. Remember this is a left-tailed test.



Notice the P-value is about 0.0014 or 0.14%.

Randomized Simulation (Randomization) Techniques

Computer technology has progressed to the point where we can measure the sampling variability in a situation directly instead of trying to estimate it with a curve and test statistic. The P-value is the probability of getting the sample data or more extreme if the null hypothesis is true. Randomized Simulation or Randomization is a technique for calculating P-value and exploring sampling variability in a hypothesis test. The idea is to have the computer take thousands of random samples from a "simulated" population under the premise that the null hypothesis is true. We can then compare the sample data to the simulated distribution and calculate the P-value.

Simulation "Tail" Principle

- If the sample data or test statistic falls in the tail of the simulation, then the sample data significant disagrees with H_0 and the P-value will be low.
- If the sample data or test statistic does NOT fall in the tail of the simulation, then the sample data does NOT significant disagree with H_0 and the P-value will be high.

Example: Let's use simulation to calculate the P-value for the previous body temperature example.
Use a 5% significance level.

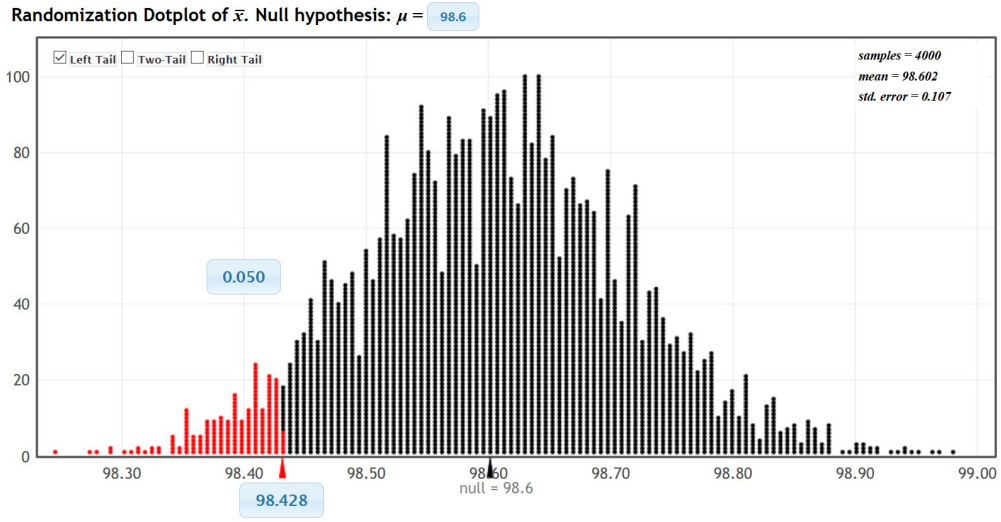
H_0 : $\mu = 98.6$ °F
H_A : $\mu < 98.6$ °F (CLAIM)

Go to www.lock5stat.com and click on "StatKey". Under the "Randomized Hypothesis Tests" menu, click on "Test for Single Mean". The body temperature sample data is the first of the preloaded data sets. Notice the null hypothesis is set at $\mu = 98.6$, but can be changed if we had a different null hypothesis.

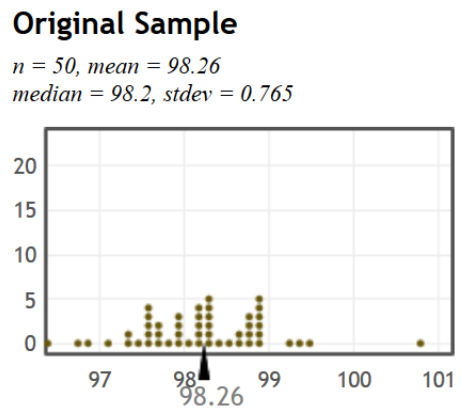
Click the "generate 1000 samples" a few times. The computer is taking thousands of random samples with the same sample size as the original sample data from the simulated population where the population mean is 98.6 °F.

Remember this is a left tail test (H_A is less than which points to the left.)

Click the "left tail" button and change the tail proportion to 5% (0.05) significance level. This is now a picture of the tail. Notice that the computer indicates that sample temperatures of 98.428°F or below significant disagree with the null hypothesis. Does the original real sample data fall in the tail?

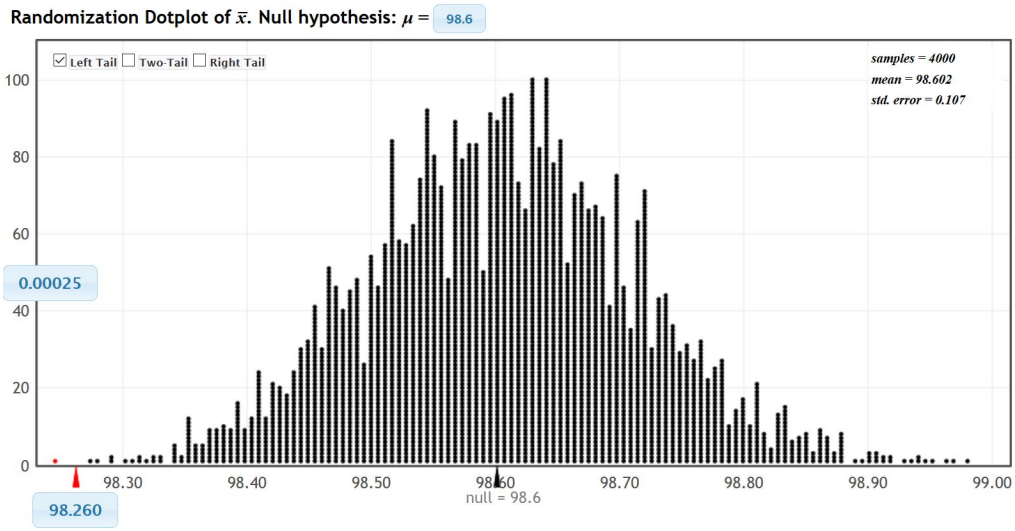


Notice the original real sample mean is 98.26°F and that is in the tail. The sample data significantly disagrees with the null hypothesis. (We did not need the test statistic to determine this.)



What about the P-value? How do you calculate the P-value?

The P-value is the probability of getting the sample data or more extreme. In the bottom box, simply put in the original real sample mean (98.26). The probability in the tail is the P-value. The simulated P-value is 0.00025 or 0.025%. This is lower than our 5% significance level so we will reject the null hypothesis.



Notes about Randomized Simulation Techniques

- This is sampling variability, so you will get slight differences in the tail numbers and P-values. They will all be very close as well.
- The P-value for a left-tailed test is in the left tail. The P-value for a right-tailed test is in the right tail. To calculate the P-value for a two-tailed test we will need to add the proportions from the two tails.
- Because you are calculating the P-value directly, you do not need as many assumptions about insuring that the sampling distribution fits a certain curve perfectly.
- For one and two-population tests we do not need to calculate a test statistic to determine significance. We can check if the sample data falls in the tail or not.

- For more advanced simulation tests involving more than two populations, we will compare the original sample data test statistic to simulated test statistics. If the original test statistic falls in the tail, the sample data significantly disagrees with the null hypothesis.

P-Value, Test Statistic & Simulation Summary Table

		Significant Test Statistic		**Test Statistic NOT Significant**
		(test stat falls in tail)		**(test stat does not fall in tail)**
		OR		**OR**
		Small P-value		**Large P-value**
		(P-value ≤ significance level)		**(P-value > significance level)**
		OR		**OR**
		Sample Data in Tail		**Sample Data NOT in Tail**
		(when simulating Ho)		**(when simulating Ho)**
Is the sample data significantly		Yes. Significantly different		Not Significantly different
different than H_0?				
Could the sample data happen		Unlikely		Could happen
by random chance				
(sampling variability)?				
Reject H_0 or		Reject H_0		Fail to Reject H_0
Fail to Reject H_0?				
Is there significant		Yes. Is evidence		No evidence
Evidence?				