# Significance Level as a Long-Run Error Rate

When implanting a hypothesis test, you set the **significance level (α)** before performing any calculations.

What typical value(s) do researchers select for α?

**0.05 is the most common value. Other common values include 0.01 and 0.1. Typically, α is chosen to be between .1% and 10%.**

How does a researcher determine what value of α is appropriate?

**The researcher needs to decide how often, in the long run, they are comfortable with making the mistake of rejecting H0 when H0 is true. A lot of sources default to a 5% rate of this type of error (α=0.05). However, the field of research can and implications of such an error can help the researcher determine if 5% is appropriate, or a lower/higher rate is feasible.**

What is the definition of significance level?

**α = P(Type I error) = P( rejecting H0 | H0 true)**

**Remember, the significance level (α) could be interpreted as a long run error rate.** For example,

Suppose we had the ability to gather 100 independent samples, each sample of size n, from the population of interest. When the null hypothesis is true, we expect approximately (100α) of the hypothesis tests to result in a decision where the null hypothesis is incorrectly rejected.

1. As the researcher, you set α=0.01 and you (somehow) can gather 10,000 independent samples, each sample of size 6, from the population of interest. If the null hypothesis were true, approximately how many of your hypothesis test results should result in an incorrect decision?
2. 1
3. 6
4. 10
5. 60
6. **100**
7. 600
8. Cannot be determined
9. In a simulation study with α=0.05, 1,000 independent samples, each of size 10, were generated from a population distribution where the null hypothesis is true. Which of the following numbers might be associated with the number of times the null hypothesis was rejected, if the statistical methodology used to implement the test works as expected?
10. 48
11. 52
12. 47
13. 54
14. **Any of the above is highly plausible**

# Power: A Long-Run Rate of Correct Decisions

The significance level describes the probability of rejecting the null hypothesis, H0, when the null hypothesis is true. This is describes one kind of mistake that could occur when implementing a hypothesis test. There is a second kind of mistake that could occur.

**Type I Error:** Reject H0 when H0 is true

**Type II Error:** Do not reject H0 when H0 is false

Example: You are testing hypotheses about a population mean, µ:

H0: µ ≤ 8  
 H1: µ > 8

You gather a sample of size 11, conduct a hypothesis test at significance level α=0.05 with a resulting p-value of 0.07. The null hypothesis was not rejected. However, the true population mean is µ = 9.1. Your hypothesis test resulted in a Type II error.

**β** is common notation used to denote the probability that H0 is not rejected when H0 is false.

1. For the hypotheses below, what value(s) of the true population mean could result in a hypothesis test with a Type II error?  
    H0: µ ≥ -5  
    H1: µ < -5
   1. (-5, 5)
   2. [-5, ∞)
   3. **(-∞, -5]**
2. For the hypotheses below, what value(s) of the true population mean could result in a hypothesis test with a Type II error?  
    H0: µ = 2  
    H1: µ ≠ 2
   1. (2, ∞)
   2. (-∞, 2)
   3. **Both choices above**
   4. None of the choices above

Instead of discussing β = Pr(Type II error), most researchers are interested in the probability of rejecting H0 when H0 is false – that is, what is the probability of making a correct decision that we have support for H1 when, in reality, H1 is true?

**Power (1-β)**: Probability of rejecting H0 when H0 is false.

Similar to the significance level, power can be interpreted as a **long-run rate of making a correct decision** when H0 is false. For example,

Suppose we had the ability to gather 100 independent samples, each sample of size n, from the population of interest. When the null hypothesis is false and we conduct a hypothesis test with significance level, α, we expect approximately (100(1-β)) of the hypothesis tests to result in a decision where the null hypothesis is correctly rejected.

1. As the researcher, you set α=0.05 and you (somehow) can gather 10,000 independent samples, each sample of the same size which allows your specified test to have power of .90, from the population of interest. If the null hypothesis were false, approximately how many of your hypothesis test results should result in a correct decision (e.g. H0 is rejected)?
2. 10
3. 90
4. 100
5. 900
6. 1000
7. **9000**
8. Cannot be determined
9. In a simulation study with α=0.05, 1,000 independent samples, each of the same size which allows your specified test to have power of .78, were generated from a population distribution where the null hypothesis is not true. Which of the following numbers might be associated with the number of times the null hypothesis was rejected, if the statistical methodology used to implement the test works as expected?
10. 33
11. 52
12. 77
13. 325
14. **779**
15. Any of the above is highly plausible

# Applet: Exploring Power

**Power of a statistical test depends on several items:**

* Items related to the unknown population distribution
  + How far the true population mean, µ, is from the hypothesized value, k
  + Standard deviation of the population distribution, σ
  + Shape of the population distribution
* Items related to the question of interest
  + The format of the hypotheses tested -- 1-sided vs. 2-sided
  + The significance level, α, for the test – often, this is determined by industrial standards
* Items the researcher might have control over
  + Sample size, n
  + Statistical method used in calculations

We will use an applet to explore power, but under a couple of constraints:

* The standard deviation of the population distribution, σ, is 1, no matter which population we examine.
* Hypotheses are 2-sided (e.g. H0: µ = k H0: µ - k = 0)

A qr code with a black and red background

AI-generated content may be incorrect.

<https://meganheyman.shinyapps.io/PowerCompApp/>

The “Compare Methods” tab provides the ability to explore the long-run behavior of power for several population distribution shapes, sample sizes, significance levels, and statistical methods.

## Describing the Power Curve

Select settings in the applet:

* Select a symmetric population distribution (Uniform, Normal, or LaPlace).
* Select a sample size of n=8
* Select a significance level of α=0.05
* For now, only select “t-test” as the statistical method (unselect the other methods)

1. In the “Compare Methods” tab, what is represented on the vertical axis?
   1. Probability of rejecting H0 when H0 is true
   2. **Probability of rejecting H0 when H0 is not true**
   3. Probability of not rejecting H0 when H0 is true
   4. Probability of not rejecting H0 when H0 is not true
2. What is represented on the horizontal axis?
   1. The value of the population mean
   2. The hypothesized value
   3. **The difference between the true population mean and the hypothesized value**
3. Using your own words, describe the shape of the plotted curve.

**The graph depicts a curve that has a “U” shape. The minimum value occurs near 0 on the horizontal axis. Both tails approach value of 1 from below.**

1. Based on the plotted curve, which statement is most justified?
   1. As the true population mean becomes closer to the hypothesized value in size/magnitude, the power increases.
   2. **As the true population mean becomes closer to the hypothesized value in size/magnitude, the power decreases.**
   3. As the true population mean becomes closer to the hypothesized value in size/magnitude, the power remains fairly constant.
2. Can power be any larger than 1 or smaller than 0? Briefly justify your answer.

**No, power must be strictly bounded between 0 and 1. Power is a probability, and probabilities must take values within [0, 1].**

1. What value of power for a hypothesis test do you think would be most desirable? Why?

**Obtaining a value of power = 1 would be most desirable. This implies that the null hypothesis is always rejected when the null hypothesis is, in truth, false (e.g. a correct decision is always made).**

## Exploring how the power curve changes

1. Keeping the significance level and population distribution shape the same, examine how the power curve changes for any single statistical method as you increase the sample size. In your own words, describe the behavior you observed.

**As the sample size increases, the curve still appears to maintain the “U” shape, with the minimum near 0 on the horizontal axis. However, the curve approaches a power of 1 more quickly in both tails.**

1. When everything except sample size is held constant, which statement do you find most justified by simulation results displayed in the applet?
   1. As sample size increases, power appears to be unaffected.
   2. As sample size increases, power appears to decrease, pointwise, as the horizontal axis variable becomes farther from 0.
   3. **As sample size increases, power appears to increase, pointwise, as the horizontal axis variable becomes farther from 0.**
2. Keeping the sample size and population distribution shape the same, examine how the power curve changes for any statistical method as you increase the significance level. In your own words, describe the behavior you observed.

**As the significance level increases, the curve still appears to maintain the “U” shape, with the minimum near 0 on the horizontal axis. However, the curve approaches a power of 1 more quickly in both tails. Moreover, the value of the minimum in the curve is close to the selected significance level, thus the value of the curve minimum also increase as sig. level increases.**

1. When everything except significance level is held constant, which statement do you find most justified by simulation results displayed in the applet?
   1. As significance level increases, power appears to be unaffected.
   2. As significance level increases, power appears to decrease, pointwise, as the horizontal axis variable becomes farther from 0.
   3. **As significance level increases, power appears to increase, pointwise, as the horizontal axis variable becomes farther from 0.**
2. Now, display the power curves for all four statistical methods for a symmetric population distribution with any sample size and significance level you would like. In your own words, describe how the power curves compare across methods.

**This is somewhat dependent on the population distribution selected. In this set of simulation results, the residual bootstrap tends to be least powerful, point to point, while wild bootstrap appears as most powerful. However, the wild bootstrap does not maintain the significance level (at a true parameter – hypothesized value) difference of 0, thus the test is called “biased.” This might be due to the choice of wild bootstrap distribution used within the simulations.**

**The t-test and randomization test often seem to have similar power results.**

# Reflection

1. Is there any behavior that you observed in the applet that was unexpected? If so, please briefly describe.
2. The applet that we explored contains simulation results. Of the items that were explored today, which aspects could a researcher control when conducting a hypothesis test?

**The researcher can potentially set the significance level. Depending on available resources, the researcher can select a larger sample size to achieve higher power. Moreover, the researcher selects the statistical method used within the hypothesis test calculations, based on conditions that seem reasonable.**

1. Suppose that you gather a single sample of size n from an unknown population distribution. What is the probability that H0 is correctly rejected for a test that you conduct? Briefly justify.

**The probability is either 0 or 1 – the null hypothesis will either correctly be rejected or not for a given sample.**

1. Optionally, explore how power results depend on the population distribution shape. (Remember, the preferred statistical method that you use for analysis can depend on the underlying population distribution shape.) If you notice any unexpected behavior or have questions about what you observe, please describe here.