Chapter 9

Statistical Inference : Significance Tests about Hypotheses (One Population)

Steps for Performing a Significance Test

The Steps of a Significance Test

A *significance test* is a method of using data to summarize the evidence about a hypothesis.

A *significance test* about a hypothesis has *five steps.*

1. Hypotheses
2. Assumptions
3. Test Statistic (sample evidence)
4. P-value
5. Conclusion

Step 1: Hypothesis

A *hypothesis* is a statement about a population, usually of the form that a certain parameter takes a particular numerical value or falls in a certain range of values.

The main goal in many research studies is to check whether the data support certain hypotheses.

Each significance test has two *hypotheses*:

1. The *null hypothesis* is a statement that the parameter takes a particular value. It has a single parameter value.

2. The *alternative hypothesis* states that the parameter falls in some alternative range of values.

Null and Alternative Hypotheses

The value in the null hypothesis usually represents *no effect.*

The symbol H_0 denotes null hypothesis.

The value in the alternative hypothesis usually represents *an effect of some type.*

The symbol H_a denotes alternative hypothesis.

The alternative hypothesis should express what the researcher hopes to show. The hypotheses should be formulated before viewing or analyzing the data!

Step 2: Assumptions

A significance test assumes that the data production used randomization.

Other assumptions may include:

- Assumptions about the sample size.
- Assumptions about the shape of the population distribution.

Step 3: Test Statistic

A *test statistic* describes how far the point estimate falls from the parameter value given in the null hypothesis (usually in terms of the number of standard errors between the two).

If the *test statistic* falls **far** from the value suggested by the null hypothesis in the direction specified by the alternative hypothesis, it is evidence against the null hypothesis and **in favor of the alternative hypothesis**.

We use the test statistic to assesses the evidence against the null hypothesis by giving a probability, the

P-Value.

Step 4: P-value

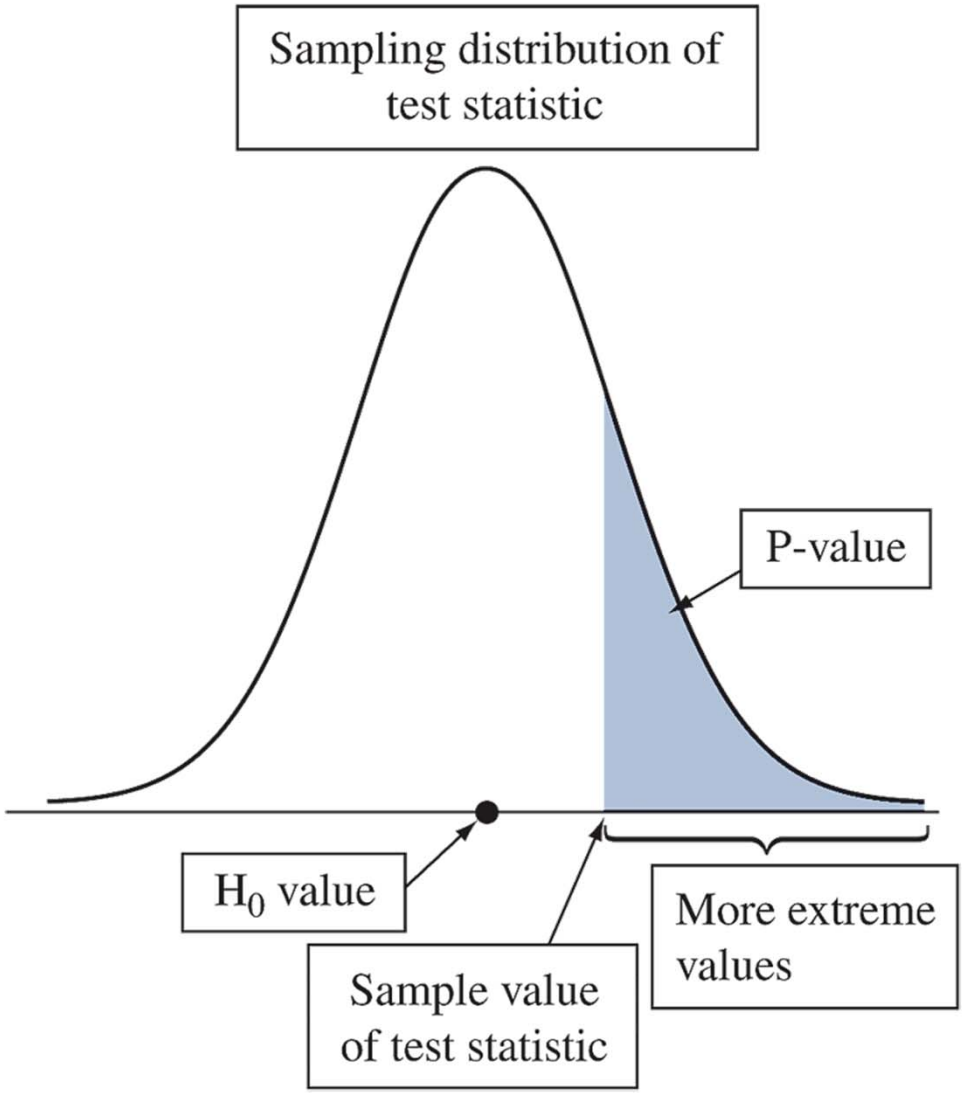
To interpret a test statistic value, we use a probability summary of the evidence *against* the null hypothesis, H_0 .

- First, we presume that H_0 is true.
- Next, we consider the sampling distribution from which the test statistic comes.

We summarize how far out in the tail the *test statistic* falls by the tail probability of that value and values even more extreme. This probability is called a *P-value.*

The smaller the P-value, the stronger the evidence the data provide against the null hypothesis. That is, a small *P*-value indicates a small likelihood of observing the sampled results if the null hypothesis were true.

Step 4: P-value



Step 5: Conclusion

The **conclusion** of a significance test reports the *P*-value and *interprets* what it says about the question that motivated the test.

May includes a decision about the *validity* of the null hypothesis H_0.

Significance Tests About Proportions

Steps of a Significance Test about a Population Proportion

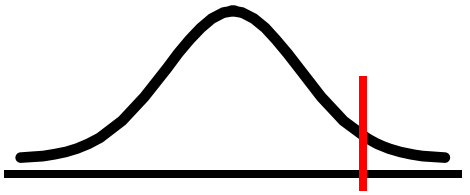
Step 1: Hypotheses

The **null hypothesis** has the form:

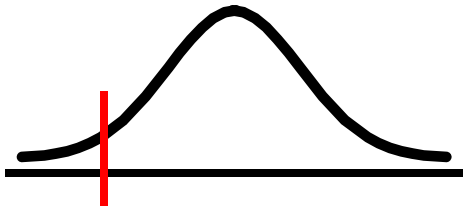
$$H_0 : p = p_0$$

The **alternative hypothesis** has the form:

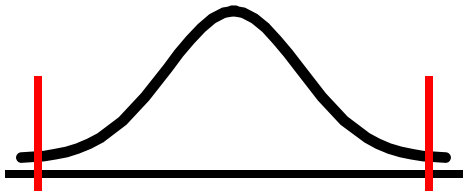
$$H_a : p > p_0 \left(one - sided\ test\right) or$$



$$H_a : p < p_0 \left(one - sided\ test\right) or$$



$$H_a : p \neq p_0 \left(two - sided\ test\right)$$



Steps of a Significance Test about a Population Proportion

Step 2: Assumptions

- The variable is categorical
- The data are obtained using randomization
- The sample size is sufficiently large that the sampling distribution of the sample proportion is approximately normal:

$$np_0 \geq 15 \text{ and } n(1-p_0) \geq 15$$

Steps of a Significance Test about a Population Proportion

Step 3: Test Statistic

The test statistic is: $z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}}$

Step 4: *P*-value

P-value = P(Z > test statistic) for right-tailed test

= P(Z < test statistic) for left-tailed test

= 2 X P(Z > |test statistic|) for two-tailed test

Step 5: Conclusion

We summarize the test by reporting and interpreting the *P*-value.

Example: Are Astrologers' Predictions Better Than Guessing?

An astrologer prepares horoscopes for 116 adult volunteers. Each subject also filled out a California Personality Index (CPI) survey. For a given adult, his or her horoscope is shown to the astrologer along with their CPI survey as well as the CPI surveys for two other randomly selected adults. The astrologer is asked which survey is the correct one for that adult.

- With random guessing, $p = 1/3$
- The astrologers' claim: $p > 1/3$
- The hypotheses for this test:

Example: Are Astrologers' Predictions Better Than Guessing?

Step 1: Hypotheses

The null hypothesis has the form

The alternative hypothesis has the form

Step 2: Assumptions

- The data is categorical – each prediction falls in the category "correct" or "incorrect".
- Each subject was identified by a random number. Subjects were randomly selected for each experiment.

Example: Are Astrologers' Predictions Better Than Guessing?

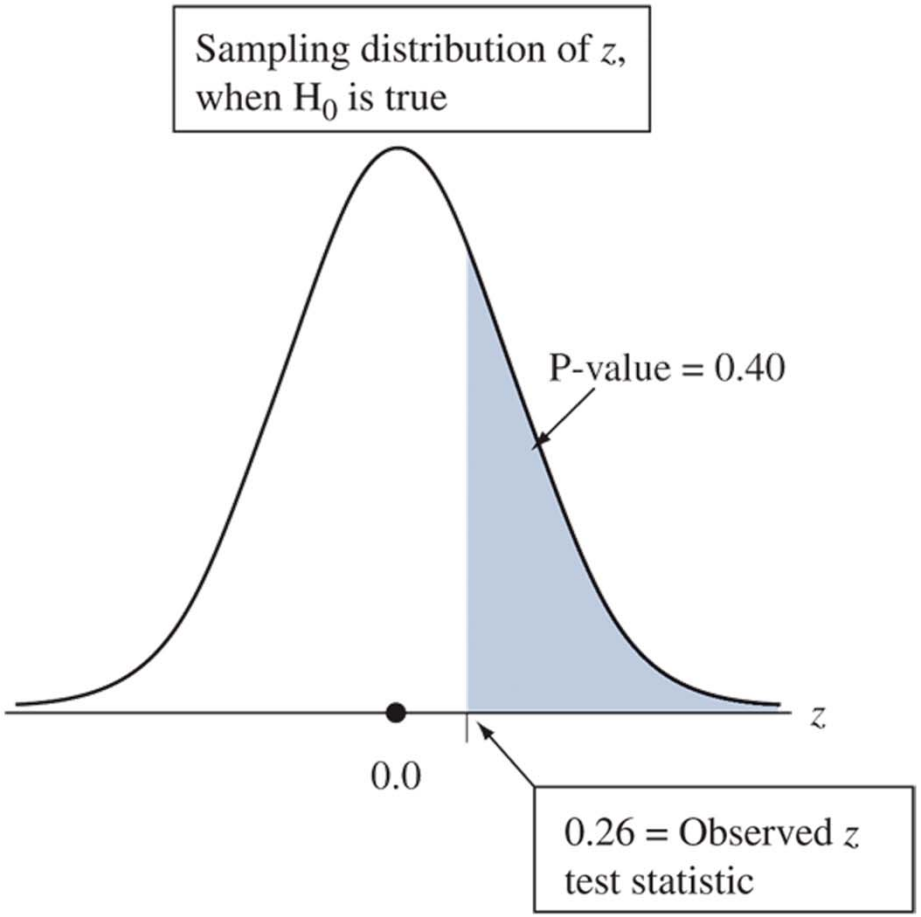
Step 3: Test Statistic:

In the actual experiment, the astrologers were correct with 40 of their 116 predictions (a success rate of 0.345).

Example: Are Astrologers' Predictions Better Than Guessing?

Step 4: P-value

P-value =



Example: Are Astrologers' Predictions Better Than Guessing?

Step 5: Conclusion

The *P*-value of 0.40 is *not* especially small.

It does *not* provide strong evidence against $H_0 : p = 1/3$.

There is *not* strong evidence that astrologers have special predictive powers.

Example:
Dogs Detecting Cancer by Smell

Study:

Investigate whether dogs can be trained to distinguish a patient with bladder cancer by smelling compounds released in the patient's urine.

Experiment:

Each of 6 dogs was tested with 9 trials.

In each trial, one urine sample from a bladder cancer patient was randomly place among 6 control urine samples.

Results:

In a total of 54 trials with the six dogs, the dogs made the correct selection 22 times (a success rate of 0.407).

Question to Explore:

Does this study provide strong evidence that the dogs' predictions were better or worse than with random guessing?

Example:
Dogs Detecting Cancer by Smell

Step 1: Hypotheses

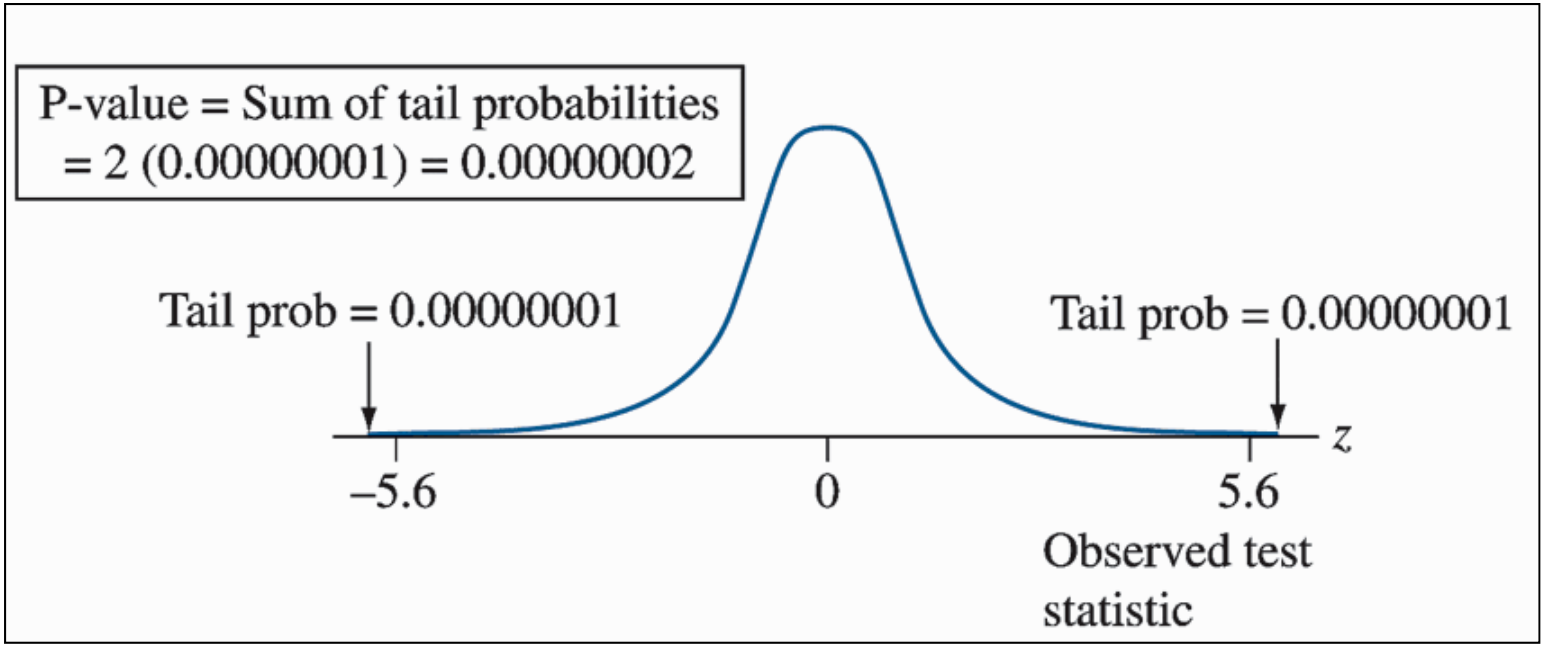
Step 2: Assumptions:

We will see that the two-sided test is robust even when this assumption is not satisfied.

Example:
Dogs Detecting Cancer by Smell

Step 3: Test Statistic

Step 4: P-value

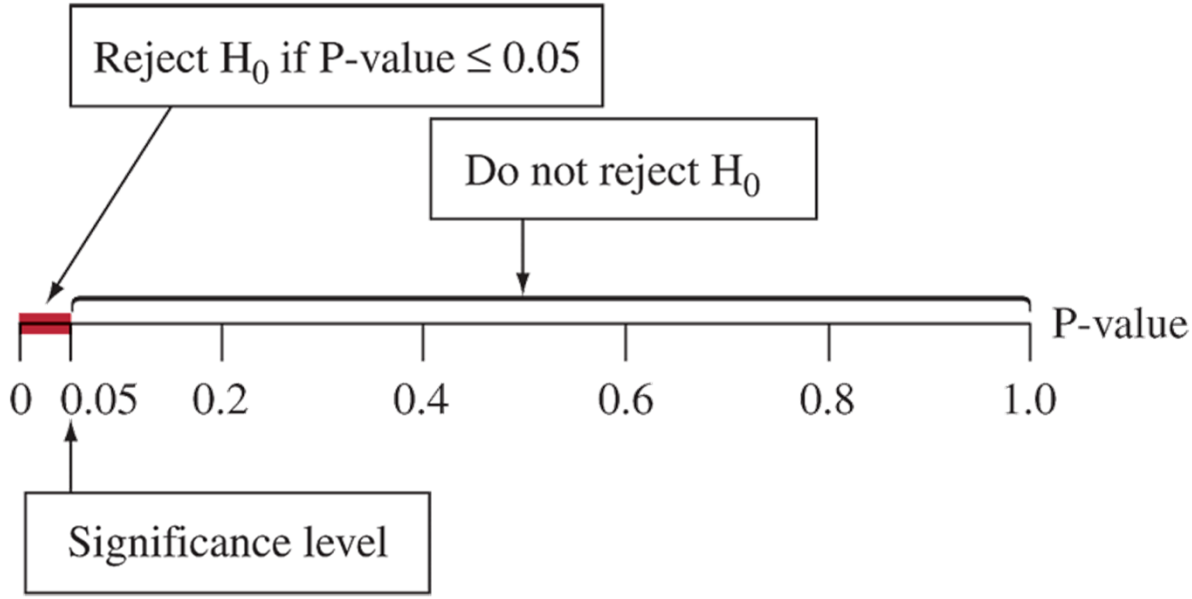


Example:
Dogs Detecting Cancer by Smell

Step 5: Conclusion

Since the P-value is very small and the sample proportion is *greater than* 1/7, the evidence strongly suggests that the dogs' selections are *better* than random guessing.

The Significance Level Tells Us How Strong the Evidence Must Be



The *significance level* is a number such that we reject H_0 if the P-value is less than or equal to that number.

In practice, the most common significance level is 0.05.

When we reject H_0 we say the results are ***statistically significant***.

Report the *P*-value

Learning the actual *P*-value is more informative than learning only whether the test is “statistically significant at the 0.05 level”.

The *P*-values of 0.01 and 0.049 are both *statistically significant* in this sense, but the first *P*-value provides much stronger evidence against H_0 than the second.

The Binomial Test for Small Samples

The significance test about a proportion assumes **normal** sampling distributions for $\hat{p}$ and the z-test statistic**.**

the expected numbers of successes and failures are at least 15.

In practice, the large-sample *z*-test still performs quite well in two-sided alternatives even for small samples.

Warning:

For one-sided tests, when p_0 differs from 0.50, the large-sample test does not work well for small samples.

Significance Tests About Means

Steps of a Significance Test About a Population Mean

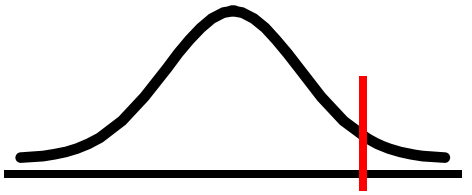
Step 1: Hypotheses:

The null hypothesis has the form:

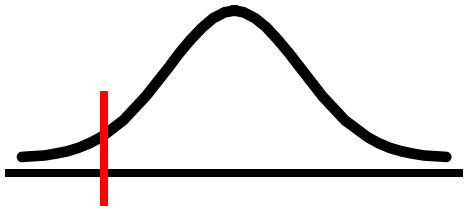
$$H_0 : \mu = \mu_0$$

The alternative hypothesis has the form:

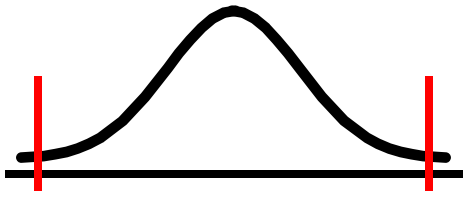
$$H_a : \mu > \mu_0 \left(one - sided\ test\right) or$$



$$H_a : \mu < \mu_0 \left(one - sided\ test\right) or$$



$$H_a : \mu \neq \mu_0 \left(two - sided\ test\right)$$



Steps of a Significance Test About a Population Mean

Step 2: Assumptions

The variable is quantitative.

The data are obtained using randomization.

The population distribution is approximately normal. This is most crucial when *n* is small and H_a is one-sided.

Steps of a Significance Test About a Population Mean

Step 3: Test Statistic

The test statistic is: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Step 4: *P*-value

P-value	= P(t > test statistic)	for right-tailed test
	= P(t < test statistic)	for left-tailed test
	= 2 X P(t > |test statistic|)	for two-tailed test

Step 5: Conclusion

We summarize the test by reporting and interpreting the *P*-value.

Example: Mean Weight Change in Anorexic Girls

A study compared different psychological therapies for teenage girls suffering from anorexia.

The variable of interest was each girl's weight change: weight at the end of the study minus weight at the beginning of the study. Change = weight after – weight before

The weight change was positive if the girl gained weight and negative if she lost weight.

In this study, 29 girls received the cognitive therapeutic treatment.

Results:
The weight changes for the 29 girls had a sample mean $\overline{x}$ of 3.00 pounds and standard deviation *s* of 7.32 pounds.

Example: Mean Weight Change in Anorexic Girls

Weight				Weight				Weight			
Girl	Before	After	Change	Girl	Before	After	Change	Girl	Before	After	Change
1	80.5	82.2	1.7	11	85.0	96.7	11.7	21	83.0	81.6	−1.4
2	84.9	85.6	0.7	12	89.2	95.3	6.1	22	76.5	75.7	−0.8
3	81.5	81.4	−0.1	13	81.3	82.4	1.1	23	80.2	82.6	2.4
4	82.6	81.9	−0.7	14	76.5	72.5	−4.0	24	87.8	100.4	12.6
5	79.9	76.4	−3.5	15	70.0	90.9	20.9	25	83.3	85.2	1.9
6	88.7	103.6	14.9	16	80.6	71.3	−9.3	26	79.7	83.6	3.9
7	94.9	98.4	3.5	17	83.3	85.4	2.1	27	84.5	84.6	0.1
8	76.3	93.4	17.1	18	87.7	89.1	1.4	28	80.8	96.2	15.4
9	81.0	73.4	−7.6	19	84.2	83.9	−0.3	29	87.4	86.7	−0.7
10	80.5	82.1	1.6	20	86.4	82.7	−3.7				

Example: Mean Weight Change in Anorexic Girls

How can we frame this investigation in the context of a *significance test* that can detect whether the therapy was effective?

Null hypothesis: “no effect”

Alternative hypothesis: therapy is “effective”

Step 1: Hypotheses

population mean weight change is μ .

Example: Mean Weight Change in Anorexic Girls

Step 2: Assumptions

The variable (weight change) is quantitative.

The subjects were a *convenience sample*, rather than a random sample. The question is whether these girls are a good representation of all girls with anorexia.

The population distribution is approximately normal.

Step 3: Test Statistic

Example: Mean Weight Change in Anorexic Girls

Step 4: *P*-value

The ***P*-value** is in (0.01, 0.025).

Using software, P-value = 0.018

This is the probability of obtaining a sample this extreme if the treatment had no effect.

Step 5: Conclusion

The **small** *P*-value of 0.018 provides considerable evidence against the null hypothesis (the hypothesis that the therapy had no effect).

"The diet had a statistically significant positive effect on weight. The effect, however, may be small in practical terms.

95% CI for μ : (0.2, 5.8) pounds

Results of Two-Sided Tests and Results of Confidence Intervals Agree

Conclusions about means using two-sided significance tests are consistent with conclusions using confidence intervals.

If *P*-value ≤ 0.05 in a two-sided test,

the 95% confidence interval does not contain the H_0 value specified by the null hypothesis.

If *P*-value > 0.05 in a two-sided test,

the 95% confidence interval does contain the H_0 value specified by the null hypothesis.

When the Population Does Not Satisfy the Normality Assumption

For large samples (roughly about 30 or higher), this assumption is usually not important.

The sampling distribution of $\bar{x}$ is approximately normal regardless of the population distribution.

In the case of small samples, we cannot assume that the sampling distribution of $\bar{x}$ is approximately normal.

Two-sided inferences using the *t* distribution are robust against violations of the normal population assumption. They still usually work well if the actual population distribution is not normal.

The test does not work well for a one-sided test with small *n* when the population distribution is highly skewed.

Regardless of Robustness, Look at the Data

Whether *n* is small or large, you should look at the data to check for *severe skew* or for *outliers* that occur primarily in **one** direction.

They could cause the **sample mean** to be a misleading measure.

Limitations of Significance Tests

Statistical Significance vs. Practical Significance

When the sample size is very large, tiny deviations from the null hypothesis (with little practical consequence) may be found to be statistically significant.

When the sample size is very small, large deviations from the null hypothesis (of great practical importance) might go undetected (statistically insignificant).

Statistical significance is not the same thing as practical significance.

Misinterpretations of Results of Significance Tests

Results of significance tests are often *misinterpreted*.

- **"Do Not Reject H_0" does *not* mean "Accept H_0".** A *P*-value above 0.05 when the significance level is 0.05, does not mean that H_0 is correct. A test merely indicates whether a particular parameter value is *plausible.*

  A **confidence interval** shows that there is a range of plausible values, not just a single one.

Misinterpretations of Results of Significance Tests

- **The *P*-value cannot be interpreted as the probability that H_0 is true.**

  The *P*-value is

  P(test statistic takes observed value or beyond in tails | H_0 is true)

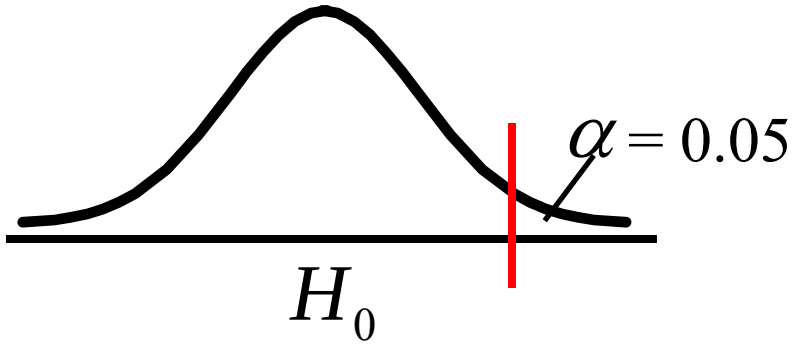
- Some tests may be statistically significant just by chance.

- It is misleading to report results only if they are "statistically significant".

- True effects may not be as large as initial estimates reported by the media.

Decision Errors

Suppose H_0 is true



At $\alpha = 0.05$, we reject the null hypothesis with probability at most 0.05. (The extremity of the test statistic is due to chance)

If the null hypothesis *is true* , P(type I error) is at most 0.05.

About 5% of all samples from this population will lead us to incorrectly reject the null hypothesis and conclude significance.

P(Type I Error) = Significance Level, α

Suppose H_0 is true. The probability of rejecting H_0, thereby committing a Type I error, equals the significance level, α, for the test.

We can control the probability of a Type I error by our *choice* of the significance level.

The more serious the *consequences* of a Type I error, the smaller α should be.

P(Type II Error) = β

If we fail to reject H_0 when in fact H_a is true, this is a **Type II** error.

The probability of this incorrect decision is denoted by β.

Significance Test Results

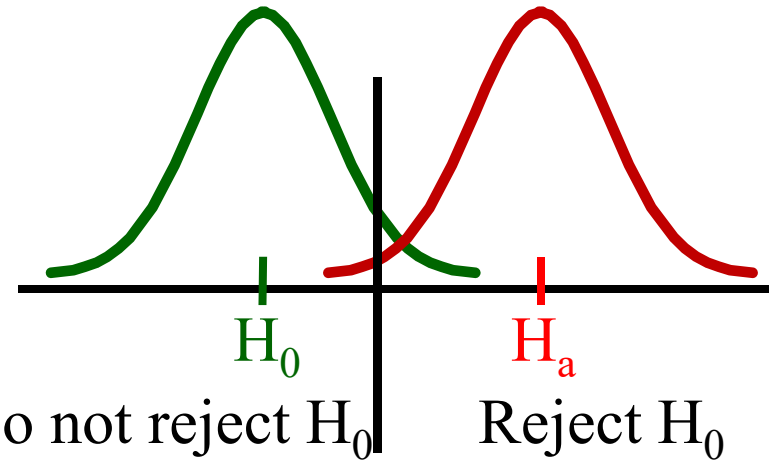
Reality About H_0	Decision: Do not reject H_0	Decision: Reject H_0
H_0 true	Correct decision	Type I error
H_0 false	Type II error	Correct decision

Type I error occurs if we reject H_0 when it is actually true.

Type II error occurs if we do not reject H_0 when it is actually false.

As P(Type I Error), α goes *Down*, P(Type II Error), β goes *Up.*

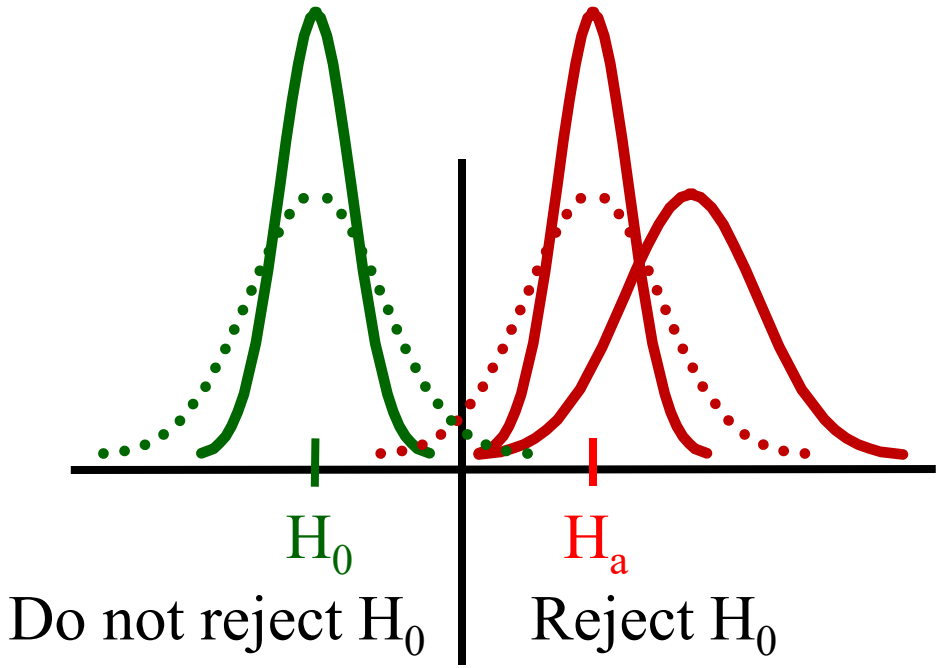
The two probabilities are **inversely** related.



Type II Error

For a fixed significance level α, P(Type II error) decreases

- as the parameter value moves farther into the H_a values and away from the H_0 value.
- as the sample size increases.



Power of a Test

When H_0 is false, you want the probability of rejecting it to be high.

The probability of rejecting H_0 when it is false is called the power of the test.

Power = 1 – P(Type II error)

The higher the power, the better.

In practice, it is ideal for studies to have high power while using a relatively small significance level.