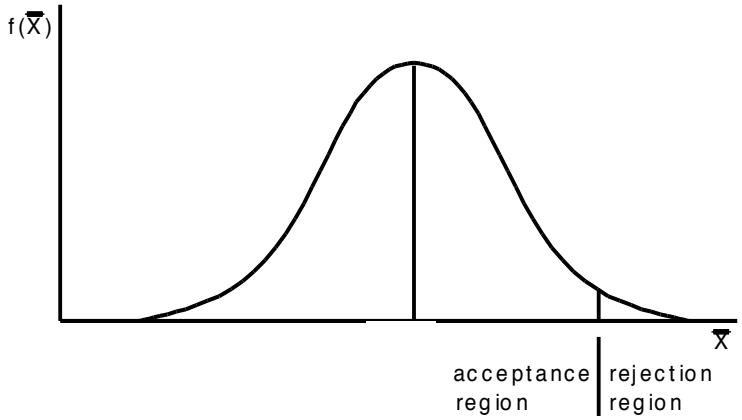
Hypothesis Testing

We have previously focused on estimating population means and variances with sample means and variances. This is termed inferential statistics. Another type of inference procedure is called hypothesis testing. We will be using hypothesis testing to evaluate questions where we need a yes/no answer. (Is the population mean equal to 30? Is this data distributed according to the negative exponential distribution?) We will be using hypothesis testing in order to make conclusions about whether or not there is a difference in means due to a process, or is it just randomness.

Hypothesis testing consists of a statistical test composed of five parts, and is based on proof by contradiction:

1. Define the null hypothesis, H_o
2. Develop the alternative hypothesis, H_a
3. Evaluate the test statistic
4. Define the rejection region (may be one or two-tailed)
5. Make a conclusion based on comparison of the value of the test statistic and the rejection region (accept or reject the null hypothesis).



*Keep in mind that acceptance or rejection of the null hypothesis does not assure you that your conclusion is correct. There is error inherent in every decision.

Reject H_o – this decision means that the observed data do not support the null hypothesis, that is, the observed data provides evidence that Ho is not true.

Accept H_o – this decision indicates that the data do not provide evidence that the null hypothesis is not true.

There are two types of errors which can result from a hypothesis test.

> Type I error (α = probability of type I error) – reject the null hypothesis when it is actually true.
>
> Type II error (β = probability of type II error**) –** accept the null hypothesis when it is actually false.

Typically, you define the amount of risk you are willing to take of making a type I error, and adjust the other factors to minimize the risk of making a type II error.

- Select level of significance based on how much risk you can tolerate of making a type I error.
  - Levels of significance are typically 0.01, 0.05, 0.10
  - Level of significance selected depends on the severity of the consequences of a type I error
- Factors effecting type II errors are
  - Design of sample selection (how samples are obtained)
  - Sample size
  - Choice of test statistic

We would ideally like to select a test procedure in which both type I and type II error probabilities are small.

The probability of committing both types of error can be reduced by increasing the sample size.

Single Sample Test of Hypothesis on μ:

For Normal Population and Known σ:

Ho: $\mu = \mu o$

Test Statistic: $z = \dfrac{\bar{x} - \mu_o}{\sigma / \sqrt{n}}$

Alternative Hypothesis:	$H_a: \mu \neq \mu_0$	$H_a: \mu > \mu_0$	$H_a: \mu < \mu_0$
Rejection Region: for level α test	RR: $\lvert z \rvert > z_{\alpha/2}$	RR: $z > z_{\alpha}$	RR: $z < -z_{\alpha}$

Example:

A company that produces bias-ply tires is considering a modification in the tread design. An economic feasibility study indicates that the modification can be justified only if the true average tire life under standard test condition exceeds 40,000 miles. A random sample of n = 16 prototype tires is manufactured and tested, resulting in a sample average tire life of $\bar{x}$ = 40,758 miles. Suppose the standard deviation for the current version of the tire is σ = 1500 miles and is not expected to change. Do the data suggest that the modification meets the condition required for changeover? Test the appropriate hypothesis using significance level α = 0.01.

Large Sample Tests, σ unkown:

If the population being sampled is known to be normally distributed ***and*** the sample size is large (typically $n \geq 30$), the sample standard deviation can be used as a surrogate for an unknown population standard deviation with little loss of accuracy:

H_o: $\mu = \mu_0$

Test Statistic: $$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Alternative Hypothesis:	H_a: $\mu \neq \mu_0$	H_a: $\mu > \mu_0$	H_a: $\mu < \mu_0$
Rejection Region:	RR: $\lvert z \rvert > z_{\alpha/2}$	RR: $z > z_\alpha$	RR: $z < -z_\alpha$*

* does not mean –z, means the value of z will be negative

Example

A certain type of brick is being considered for use in a particular construction project. The brick will be used unless sample evidence strongly suggests that the true average compressive strength is below 3200 psi. A random sample of 36 bricks is selected and each is tested to failure. The sample average compressive strength is 3109 psi with a standard deviation of 156 psi. At a level of significance of $\alpha = 0.05$, should the brick be used?

Small Sample Tests, σ Unknown:

If the population being sampled is known to be normally distributed but the sample size is small (typically n < 30), the sample standard deviation can still substitute for the population standard deviation, but the test statistic follows a t distribution instead of a z distribution:

H_o: $\mu = \mu_0$

Test Statistic: $$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Alternative Hypothesis:	H_a: $\mu \neq \mu_0$	H_a: $\mu > \mu_0$	H_a: $\mu < \mu_0$
Rejection Region:	RR: $\lvert t \rvert > t_{\alpha/2,n-1}$	RR: $t > t_{\alpha,n-1}$	RR: $t < -t_{\alpha,n-1}$

Example

In order to test gasoline mileage performance for a new version of one of its compact cars, an automobile manufacturer selected six nonprofessional drivers to drive test cars from Phoenix to Los Angeles. At the conclusion of the trip, the resulting gas mileage numbers for the six cars were:

32.2 29.3 31.5 28.7 30.2 30.0

The manufacturer wishes to advertise that this car gets 30 mpg or better on the highway. Do the sample data support the claim that the manufacturer would like to make? Assume $\alpha = 0.05$.

TWO SAMPLE HYPOTHESIS TESTS ON μ

Comparison of Two Means, σ_1 and σ_2 Known:

If the populations being sampled are known to be normally distributed or the sample sizes are large enough that the Central Limit Theorem holds (typically $n \geq 30$), a two-sample hypothesis test on the difference between two population means is as follows:

H_o: $\mu_1 - \mu_2 = \Delta_0$

Test Statistic:
$$z = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Alternative Hypothesis:	H_a: $\mu_1 - \mu_2 \neq \Delta_0$	H_a: $\mu_1 - \mu_2 > \Delta_0$	H_a: $\mu_1 - \mu_2 < \Delta_0$
Rejection Region:	RR: $\lvert z \rvert > z_{\alpha/2}$	RR: $z > z_\alpha$	RR: $z < -z_\alpha$

Example

A random sample of 20 specimens of cold-rolled steel had an average yield strength of 29.8 ksi. A second random sample of 25 galvanized steel specimens gave an average yield strength of 34.7 ksi. Assuming that the two yield strength distributions are normal with $\sigma_1 = 4.0$ and $\sigma_2 = 5.0$, do the data indicate that the true average yield strengths, μ_1 and μ_2, are different? Assume $\alpha = 0.01$.

Large Sample Comparison of Two Means, σ_1 and σ_2 Unknown:

If the populations being sampled are known to be normally distributed ***and*** the sample size is large (typically $n \geq 30$), the sample standard deviation can be substituted for unknown population standard deviations with little loss of accuracy:

H_o: $\mu_1 - \mu_2 = \Delta_0$

Test Statistic:
$$z = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Alternative Hypothesis:	H_a: $\mu_1 - \mu_2 \neq \Delta_0$	H_a: $\mu_1 - \mu_2 > \Delta_0$	H_a: $\mu_1 - \mu_2 < \Delta_0$
Rejection Region:	RR: $\lvert z \rvert > z_{\alpha/2}$	RR: $z > z_\alpha$	RR: $z < -z_\alpha$

Example

In a sample of 30 women who did not live near a freeway, the sample average blood lead level was 9.9 and the sample standard deviation was 4.9, while a second sample of 35 females who *did* live near a freeway had a sample average and sample standard deviation of 16.7 and 7.0, respectively. Does proximity to heavily traveled roads result in higher blood lead levels? Test at $\alpha = 0.01$.

Small Sample Comparison of Means, σ_1 and σ_2 Unknown but Equal (Pooled t test):

If the populations being sampled are known to be normally distributed but the sample sizes are small (typically n < 30), the sample standard deviations can still be substituted for the population standard deviations, but the test statistic follows a t distribution instead of a z distribution. If the population standard deviations can be presumed to be equal to each other, you can use a ***pooled t test***:

H_o: $\mu_1 - \mu_2 = \Delta_0$

Test Statistic: $$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where: $$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{\nu}}$$

$$\nu = n_1 + n_2 - 2 \text{ (degrees of freedom)}$$

Alternative Hypothesis:	H_a: $\mu_1 - \mu_2 \neq \Delta_0$	H_a: $\mu_1 - \mu_2 > \Delta_0$	H_a: $\mu_1 - \mu_2 < \Delta_0$
Rejection Region:	RR: $\lvert t \rvert > t_{\alpha/2,\nu}$	RR: $t > t_{\alpha,\nu}$	RR: $t < -t_{\alpha,\nu}$

Example

A random sample of 15 ceramic insulators doped in a certain manner yielded a sample average holdoff voltage of 110 kV and a sample standard deviation of 24 kV. A random sample of 76 undoped ceramic insulators produced a sample average holdoff voltage of 101 kV with a standard deviation of 22 kV. If we can assume that the actual population standard deviations should be the same, do the data suggest that the true average holdoff voltage for doped specimens exceeds that for plain specimens by more than 5 kV at a significance level of 0.10?

Comparison of Means, σ_1 and σ_2 Unknown and Unequal (the Smith-Satterthwaite Procedure):

If the populations being sampled are known to be normally distributed but the standard deviations are unknown and ***cannot*** be presumed to be equal to each other, you can use the ***Smith-Satterthwaite Procedure***, which uses a t distribution with the number of degrees of freedom calculated as a sort of variance-weighted average of the sample sizes:

H_o: $\mu_1 - \mu_2 = \Delta_0$

Test Statistic: $$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where: $$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}$$ **⇐ round *<u>down</u>*!**

Alternative Hypothesis:	H_a: $\mu_1 - \mu_2 \neq \Delta_0$	H_a: $\mu_1 - \mu_2 > \Delta_0$	H_a: $\mu_1 - \mu_2 < \Delta_0$
Rejection Region:	RR: $\lvert t \rvert > t_{\alpha/2,\nu}$	RR: $t > t_{\alpha,\nu}$	RR: $t < -t_{\alpha,\nu}$

Example

Dextroamphetamine is a drug commonly used to treat hyperkinetic children. A paper in the *Journal of Nervous and Mental Disorders* (1968, vol. 146, pp. 136-146) reported the following data on the percentage of the drug excreted within seven hours of its administration by children having organically related disorders and children with nonorganic disorders.

1. Organic:	17.53	20.60	17.62	28.93	27.10
2. Nonorganic:	15.59	14.76	13.32	12.45	12.79

The summary values are $\bar{x}_1 = 22.36$, $\bar{x}_2 = 13.78$, $s_1^2 = 28.63$, and $s_2^2 = 1.80$. The data suggest that there is much less variability in recovery percentage for children with organically related disorders. Compare the means at a significance level of 0.01.

Paired Samples (Paired *t* Test):

From the chapter on confidence intervals, we know that sometimes tests are made using paired data. In those instances, there is one set of n individuals or objects, and two observations (let's say "before" and "after") are made on each one. Rather than compare the "before" and "after" sample means, we instead compute the differences d_i in the "before" and "after" test measurements for each individual in the sample set, then test the mean value of the differences, μ_D:

H_o: $\mu_D = \Delta_0$

Test Statistic:
$$t_{paired} = \frac{\overline{d} - \Delta_0}{s_D / \sqrt{n}}$$

Where:
$$s_D^2 = \frac{\sum_{i=1}^{n}\left(d_i - \overline{d}\right)^2}{n-1} = \frac{\sum_{i=1}^{n} d_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} d_i\right)^2}{n-1}$$

Alternative Hypothesis:	H_a: $\mu_D \neq \Delta_0$	H_a: $\mu_D > \Delta_0$	H_a: $\mu_D < \Delta_0$
Rejection Region:	RR: $\lvert t_{pooled} \rvert > t_{\alpha/2,\nu}$	RR: $t_{pooled} > t_{\alpha,\nu}$	RR: $t_{pooled} < -t_{\alpha,\nu}$

Example

In an experiment designed to evaluate an additive to increase the strength of concrete, each of five batches of concrete was divided in half and the additive added to one half of each batch. The resulting compressive strength measurements (load in kips at failure) were shown below. Does the additive work? Test at $\alpha = 0.01$.

Treated	16.1	14.7	17.4	13.7	16.9
Untreated	14.8	13.2	15.5	12.3	15.9

Tests of Hypotheses on σ^2

Single Sample Test of Hypothesis on σ^2 From a Normal Population:

H_o: $\sigma^2 = \sigma_o^2$

Test Statistic: $$\chi^2 = \frac{(n-1)s^2}{\sigma_o^2}$$

Alternative Hypothesis:	H_a: $\sigma^2 > \sigma_o^2$	H_a: $\sigma^2 < \sigma_o^2$	H_a: $\sigma^2 \neq \sigma_o^2$
Rejection Region:	RR: $\chi^2 > \chi^2_{\alpha,(n-1)}$	RR: $\chi^2 < \chi^2_{(1-\alpha),(n-1)}$	RR: $\chi^2 > \chi^2_{\alpha/2,(n-1)}$ *or* $\chi^2 < \chi^2_{(1-\alpha/2),(n-1)}$

Example:

A manufacturer of liquid detergent is interested in the uniformity of the machine used to fill bottles. If the variance of fill volume exceeds 0.01 fluid oz^2, an unacceptable proportion of bottles will be underfilled. A random sample of 20 bottles results in a sample variance of fill volume of $s^2 = 0.0153$ fluid oz^2. Assuming $\alpha = 0.05$, is there evidence in the sample data to suggest the need for replacing or overhauling the machine?

Comparison of Variances:

H_o: $\sigma_1^2 = \sigma_2^2$

*may be used as a preliminary test for comparing the means of the two samples.

Test Statistic: $F = \frac{s_1^2}{s_2^2}$ where subscripts are such that $s_1^2 \geq s_2^2$

v1 = n_1 - 1, and v2 = n_2 - 1

Alternative Hypothesis:	H_a: $\sigma_1^2 > \sigma_2^2$	H_a: $\sigma_1^2 < \sigma_2^2$	H_a: $\sigma_1^2 \neq \sigma_2^2$
Rejection Region:	RR: $F > F_{\alpha,\nu_1,\nu_2}$	RR: $F > F_{\alpha,\nu_2,\nu_1}$*	RR: $F > F_{\alpha/2,\nu_1,\nu_2}$

*note reversal of signs and subscripts!!

Example:
A study of two types of materials used in electrical conduits is to be conducted. The purpose of the study is to compare the strength of one to the other by measuring the load required to crush a 6 inch long piece of material to 40% of its original diameter. The primary question to be answered is "Is μ1 > μ2?" However, before this is done, we must consider the question, "Is $\sigma_1^2 = \sigma_2^2$?" If the answer to this appears to be yes, then a pooled t procedure can be used. Otherwise, the Smith-Satterthwaite procedure should be employed.

Material 1	Material 2
n_1 = 25	n_2 = 16
x_1 = 380 lb	x_2 = 370 lb
s_1^2 = 100	s_2^2 = 400

P-VALUES

So far, our work in hypothesis testing simply rejects or does not reject the null hypothesis. There are two problems with this (maybe not huge ones, but we could do better). First, the conclusion does not tell us how close the test statistic was to the rejection region boundary. Did we barely reject H_o? Did we barely fail to reject H_o? Second, the level of significance is decided by the person performing the hypothesis test. People might have different opinions concerning the appropriate confidence level. A P-value conveys much information concerning the strength of evidence against H_0 and allows an individual to draw a conclusion at any specified level of α.

The P-value is the smallest level of significance at which H_0 would be rejected when a specified test procedure is used on a given set of data. Once the P-value has been determined, the conclusion at any particular level, α, results from comparing the P-value to α. Procedures exist for determining P-values for both z tests and t tests, although (as we will see shortly) P-values are not always easy to calculate. It has, fortunately, become common for statistical software to include P-values in their output.

The P-value for a *z* Test

$P = 2\left[1 - F(\lvert z \rvert)\right]$	for a two-tailed test ($H_0: \mu = \mu_0; H_a: \mu \neq \mu_0$)
$P = 1 - F(z)$	for an upper-tailed test ($H_0: \mu = \mu_0; H_a: \mu > \mu_0$)
$P = F(z)$	for a lower-tailed test ($H_0: \mu = \mu_0; H_a: \mu < \mu_0$)

Example

A certain type of brick is being considered for use in a particular construction project. The brick will be used unless sample evidence strongly suggests that the average compressive strength is below 3200 psi. A random sample of 36 bricks is selected and each is tested to failure. The sample average compressive strength is 3109 psi with a standard deviation of 156 psi. Use the P-value to test the hypothesis that the compressive strength is below 3200 psi.

The P-value for a *t* Test

The standard normal distribution gives critical values for many values of *z*. However, for any given number of degrees of freedom, the *t* table contains only 10 values. Therefore, the exact P-value cannot usually be determined. The procedure (unless you're using software that gives you exact P-values) is to use upper and lower bounds for the P-value.

Example

In order to test gasoline mileage performance for a new version of one of its compact cars, an automobile manufacturer selected six nonprofessional drivers to drive test cars from Phoenix to Los Angeles. At the conclusion of the trip, the resulting gas mileage numbers for the six cars were:

32.2 29.3 31.5 28.7 30.2 30.0

The manufacturer wishes to advertise that this car gets 30 mpg or better on the highway. Use P-values to determine whether or not it would be wise for the manufacturer to make this claim.

Solution

H_o: $\mu = 30$

H_a: $\mu > 30$

TS: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{30.32 - 30}{1.32/\sqrt{6}} = 0.59$ (from an earlier example)

The *t*-table in your textbook doesn't have a column for upper tail areas this large. An alternative approach is to use the TDIST function in Excel:

TDIST(t, ν, tails)

If *tails* = 1, this function returns the upper tail area of a *t* distribution with ν degrees of freedom corresponding to the specified *t* value. If *tails* = 2, this function returns the combined upper and lower tail areas corresponding to ± *t* instead.

If you enter $t = 0.59$, $\nu = 5$, and *tails* = 1, Excel returns a value of 0.29, which means the *P*-value for this hypothesis test is 0.29. Thus, the conclusion that $\mu = 30$ is only valid at a significance level of $\alpha = 0.29$ or greater. This means that there is a 29% chance that the conclusion is wrong! This would be unacceptable in the real world.