Difference Between Two Population Means

We are now moving to more refined questions:

How do two (several) sub-populations compare? In particular, are their means the same?

For example,

1. Is this drug's effectiveness the same in children and adults?
2. Does brand A have the same amount of nicotine as brand B?

The way we answer these is to collect samples from both (all) subpopulations, and perform a two-sample test (ANOVA).

Statistically speaking, for two samples, we want to test whether $\mu_1 - \mu_2 = 0$ that is, whether $\mu_1 = \mu_2$.

We need some basic assumptions

1. $X_1, X_2, \ldots X_m$ is a **random sample** from a distribution with mean μ_1 and variance σ_1^2

2. $Y_1, Y_2, \ldots Y_n$ is a **random sample** from a distribution with mean μ_2 and variance σ_2^2.

3. The X and Y samples are **independent** of one another.

Example: Difference Between Two Population Means

40 patients were randomly assigned to either the Professional Oral Care (POC) group ($m = 20$) or the control group ($n = 20$).

…1 patient in the POC group and 4 in the control group dropped out because of complications

The data analysis is then based on two random sample POC ($m = 19$) and Control ($n = 16$)

Difference Between Two Population Means

The data analysis can be based on two samples with uneven sample sizes.

The natural estimator of $\mu_1 - \mu_2$ is $\overline{X} - \overline{Y}$, the difference between the corresponding sample means.

Inferential procedures are based on standardizing this estimator, so we need expressions for the expected value and standard deviation of $\overline{X} - \overline{Y}$.

$$E(\overline{X} - \overline{Y}) = E(\overline{X}) - E(\overline{Y}) = \mu_1 - \mu_2$$
$$Var(\overline{X} - \overline{Y}) = Var(\overline{X}) + (-1)^2\, Var(\overline{Y})$$

Difference Between Two Population Means

The expected value of $\overline{X} - \overline{Y}$ is $\mu_1 - \mu_2$, so $\overline{X} - \overline{Y}$ is an unbiased estimator of $\mu_1 - \mu_2$.

The standard deviation of $\overline{X} - \overline{Y}$ is

$$\sigma_{\overline{X}-\overline{Y}} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

The sample variances must be used to estimate this when population variances are unknown.

Normal Populations with Known Variances

If both of the population distributions are normal, both $\overline{X}$ and $\overline{Y}$ have normal distributions.

Furthermore, independence of the two samples implies that the two sample means are independent of one another.

Thus the difference $\overline{X} - \overline{Y}$ is normally distributed, with expected value $\mu_1 - \mu_2$ and standard deviation $\sigma_{\overline{X}-\overline{Y}}$

$$\sigma_{\overline{X}-\overline{Y}} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

Test Procedures for Normal Populations with Known Variances

Standardizing $\overline{X} - \overline{Y}$ gives the standard normal variable

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$$

In a hypothesis-testing problem, the null hypothesis states that $\mu_1 - \mu_2$ has a specified value.

If we are testing equality of the two means, then $\mu_1 - \mu_2$ will be 0 under the null hypothesis.

Test Procedures for Normal Populations with Known Variances

In general:

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$

Test statistic value: $z = \dfrac{\overline{x} - \overline{y} - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{m} + \dfrac{\sigma_2^2}{n}}}$

Test Procedures for Normal Populations with Known Variances

Null hypothesis: $H_0: \mu_1 - \mu_2 = \Delta_0$

Alternative Hypothesis   Rejection Region for Level α Test

$H_a: \mu_1 - \mu_2 > \Delta_0$ $z \geq z_\alpha$ (upper-tailed)

$H_a: \mu_1 - \mu_2 < \Delta_0$ $z \leq -z_\alpha$ (lower-tailed)

$H_a: \mu_1 - \mu_2 \neq \Delta_0$ either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (two-tailed)

Because these are z tests, a P-value is computed as it was for the z tests [e.g., P-value $= 1 - \Phi(z)$ for an upper-tailed test].

Example 1

Analysis of a random sample consisting of $m = 20$ specimens of cold-rolled steel to determine yield strengths resulted in a sample average strength of $\bar{x} = 29.8 \text{ ksi}$.

A second random sample of $n = 25$ two-sided galvanized steel specimens gave a sample average strength of $\bar{y} = 34.7 \text{ ksi}$.

Assuming that the two yield-strength distributions are normal with $\sigma_1 = 4.0$ and $\sigma_2 = 5.0$, does the data indicate that the corresponding true average yield strengths μ_1 and μ_2 are different?

Let's carry out a test at significance level $\alpha = 0.01$

Example 1
cont'd

1. The parameter of interest is $\mu_1 - \mu_2$, the difference between the true average strengths for the two types of steel.

2. The null hypothesis is $H_0 : \mu_1 - \mu_2 = 0$

3. The alternative hypothesis is $H_a : \mu_1 - \mu_2 \neq 0$

if H_a is true, then μ_1 and μ_2 are different.

4. With $\Delta_0 = 0$, the test statistic value is

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\dfrac{\sigma_1^2}{m} + \dfrac{\sigma_2^2}{n}}}$$

Example 1
cont'd

5. The inequality in H_a implies that the test is two-tailed. For

$\alpha = .01$, $\alpha/2 = .005$, and $z_{\alpha/2} = z_{.005} = 2.58$,

H_0 will be rejected if $z \geq 2.58$ or if $z \leq -2.58$.

6.

$$z = \frac{29.8 - 34.7}{\sqrt{\dfrac{16.0}{20} + \dfrac{25.0}{25}}} = \frac{-4.90}{1.34} = -3.66$$

That is, the observed value of $\bar{x} - \bar{y}$ is more than 3 standard deviations below what would be expected were H_0 true.

Example 1

7. Since –3.66 < –2.58, z does fall in the lower tail of the rejection region. H_0 is therefore rejected at level .01 in favor of the conclusion that $\mu_1 \neq \mu_2$. The sample data strongly suggests that the true average yield strength for cold-rolled steel differs from that for galvanized steel.

The *P*-value for this two-tailed test is

$$2(1 - \Phi(3.66)) \approx 2(1 - 1) = 0,$$

So H_0 would have been rejected at *any* reasonable significance level.

Large-Sample Tests

Large-Sample Tests

The assumptions of normal population distributions and known values of σ_1 and σ_2 are fortunately unnecessary when both sample sizes are sufficiently large. In this case, the Central Limit Theorem guarantees that $\overline{X} - \overline{Y}$ has approximately a normal distribution regardless of the underlying population distributions.

Furthermore, using S_1^2 and S_2^2 in place of σ_1^2 and σ_2^2 gives a variable whose distribution is approximately standard normal:

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1^2}{m} + \dfrac{S_2^2}{n}}}$$

Large-Sample Tests

A large-sample test statistic results from replacing $\mu_1 - \mu_2$ by Δ_0, the expected value of $\overline{X} - \overline{Y}$ when H_0 is true.

This statistic Z then has approximately a standard normal distribution when H_0 is true.

Tests with a desired significance level are obtained by using z critical values exactly as before.

Large-Sample Tests

Use of the test statistic value

$$z = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\dfrac{s_1^2}{m} + \dfrac{s_2^2}{n}}}$$

along with the previously stated upper-, lower-, and two-tailed rejection regions based on z critical values gives large-sample tests whose significance levels are approximately α.

These tests are usually appropriate if both $m > 40$ and $n > 40$.

Example

Data on daily calorie intake both for a sample of teens who said they did not typically eat fast food and another sample of teens who said they did usually eat fast food.

Eat Fast Food	Sample Size	Sample Mean	Sample SD
No	663	2258	1519
Yes	413	2637	1138

Does this data provide strong evidence for concluding that true average calorie intake for teens who typically eat fast food exceeds by more than 200 calories per day the true average intake for those who don't typically eat fast food?

Let's investigate by carrying out a test of hypotheses at a significance level of approximately .05.

Example
cont'd

The parameter of interest is $\mu_1 - \mu_2$, where μ_1 is the true average calorie intake for teens who don't typically eat fast food and μ_2 is true average intake for teens who do typically eat fast food.

The hypotheses of interest are
$H_0 : \mu_1 - \mu_2 = -200$ versus $H_a : \mu_1 - \mu_2 < -200$

The alternative hypothesis asserts that true average daily intake for those who typically eat fast food exceeds that for those who don't by more than 200 calories.

Example
cont'd

The test statistic value is

$$z = \frac{\bar{x} - \bar{y} - (-200)}{\sqrt{\dfrac{s_1^2}{m} + \dfrac{s_2^2}{n}}}$$

The inequality in H_a implies that the test is lower-tailed; H_0 should be rejected if $z \leq -z_{0.5} = -1.645$.

The calculated test statistic value is

$$z = \frac{2258 - 2637 + 200}{\sqrt{\dfrac{(1519)^2}{663} + \dfrac{(1138)^2}{413}}} = \frac{-179}{81.34} = -2.20$$

Example

Since –2.20 ≤ –1.645, the null hypothesis is rejected. At a significance level of .05, it does appear that true average daily calorie intake for teens who typically eat fast food exceeds by more than 200 the true average intake for those who don't typically eat such food.

The *P*-value for the test is

      P-value = area under the *z* curve to the left of

    –2.20 = $\Phi(-2.20)$ = .0139

0.0139 ≤ .05, we reject the null hypothesis at significance level .05.

Confidence Intervals for $\mu_1 - \mu_2$

When both population distributions are normal, standardizing $\overline{X} - \overline{Y}$ gives a random variable Z with a standard normal distribution.

Since the area under the *z* curve between $-z_{\alpha/2}$ and $z_{\alpha/2}$ is $1 - \alpha$, it follows that

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{m} + \dfrac{\sigma_2^2}{n}}} < z_{\alpha/2} \right) = 1 - \alpha$$

Confidence Intervals for $\mu_1 - \mu_2$

Manipulation of the inequalities inside the parentheses to isolate $\mu_1 - \mu_2$ yields the equivalent probability statement

$$P\left(\overline{X} - \overline{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} < \mu_1 - \mu_2 < \overline{X} - \overline{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} \right) = 1 - \alpha$$

This implies that a 100(1 – α)% CI for $\mu_1 - \mu_2$ has lower limit $\overline{x} - \overline{y} - z_{\alpha/2} \cdot \sigma_{\overline{X}-\overline{Y}}$ and upper limit $\overline{x} - \overline{y} + z_{\alpha/2} \cdot \sigma_{\overline{X}-\overline{Y}}$,

This interval is a special case of the general formula

$$\hat{\theta} \pm z_{\alpha/2} \cdot \sigma_{\hat{\theta}}.$$

Confidence Intervals for $\mu_1 - \mu_2$

If both *m* and *n* are large, the CLT implies that this interval is valid even without the assumption of normal populations; in this case, the confidence level is *approximately* 100(1 – α)%.

Furthermore, use of the sample variances S_1^2 and S_2^2 in the standardized variable Z yields a valid interval in which S_1^2 and S_2^2 replace σ_1^2 and σ_2^2.

Confidence Intervals for $\mu_1 - \mu_2$

Provided that m and n are both large, a CI for $\mu_1 - \mu_2$ with a confidence level of approximately $100(1 - \alpha)\%$ is

$$\bar{x} - \bar{y} \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

where $-$ gives the lower limit and the upper limit of the interval. An upper or a lower confidence bound can also be calculated by retaining the appropriate sign (+ or $-$) and replacing $z_{\alpha/2}$ *by* z_α

Our standard rule of thumb for characterizing sample sizes as large is $m > 40$ and $n > 40$.

Example

An experiment carried out to study various characteristics of anchor bolts resulted in 78 observations on shear strength (kip) of 3/8-in. diameter bolts and 88 observations on the strength of 1/2-in. diameter bolts.

Summary quantities follow, and a comparative box plot is presented in Figure 9.1.
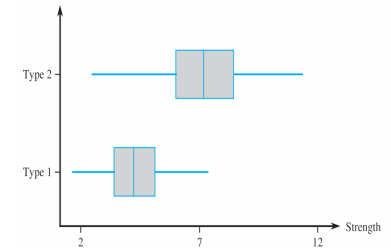
A comparative box plot of the shear strength data

Example

The sample sizes, sample means, and sample standard deviations agree with values given in the article "Ultimate Load Capacities of Expansion Anchor Bolts" (*J. of Energy Engr.,* 1993: 139–158).

The summaries suggest that the main difference between the two samples is in where they are centered.

```
Variable        N       Mean    Median    TrMean    StDev    SEMean
diam 3/8       78      4.250     4.230     4.238    1.300     0.147
Variable      Min        Max        Q1        Q3
diam 3/8    1.634      7.327     3.389     5.075
Variable        N       Mean    Median    TrMean    StDev    SEMean
diam 1/2       88      7.140     7.113     7.150    1.680     0.179
Variable      Min        Max        Q1        Q3
diam 1/2    2.450     11.343     5.965     8.447
```

Example

cont'd

Let's now calculate a confidence interval for the difference between true average shear strength for 3/8-in. bolts (μ_1) and true average shear strength for 1/2-in. bolts (μ_2) using a confidence level of 95%:

$$4.25 - 7.14 \pm (1.96)\sqrt{\frac{(1.30)^2}{78} + \frac{(1.68)^2}{88}} = -2.89 \pm (1.96)(.2318)$$

$$= -2.89 \pm .45$$

Not-so-large Sample Tests

The Two-Sample t Test and Confidence Interval

Up to now, we worked with large sample sizes and used a z-test and CI in which the sample variances were used in place of the population variances.

In fact, for large samples, the CLT allows us to use these methods even when the two populations of interest are not normal.

In practice, though, it will often happen that at least one sample size is small and the population variances have unknown values.

Without the CLT at our disposal, we proceed by making specific assumptions about the underlying population distributions.

The Two-Sample t Test and Confidence Interval

We could, for example, assume that both population distributions are members of the Weibull family or that they are both Poisson distributions. It shouldn't surprise you to learn that normality is typically the most reasonable assumption.

Assumptions

Both population distributions are normal, so that $X_1, X_2,\ldots, X_m$ is a random sample from a normal distribution and so is $Y_1,\ldots,Y_n$ (with the X's and Y's independent of one another).

The Two-Sample t Test and Confidence Interval

When the population distribution are both normal, the standardized variable

$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}$$

has approximately a t distribution with df v estimated from the data by

$$\nu = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{(s_1^2/m)^2}{m-1} + \frac{(s_2^2/n)^2}{n-1}}$$

The Two-Sample t Test and Confidence Interval

The **two-sample t confidence interval for $\mu_1 - \mu_2$** with confidence level $100(1 - \alpha)$ % is then

$$\bar{x} - \bar{y} \pm t_{\alpha/2, \nu}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

A one-sided confidence bound can be calculated as described earlier.

The **two-sample t test** for testing $H_0: \mu_1 - \mu_2 = \Delta_0$ is as follows:

Test statistic value: $t = \dfrac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\dfrac{s_1^2}{m} + \dfrac{s_2^2}{n}}}$

The Two-Sample t Test and Confidence Interval

Alternative Hypothesis	**Rejection Region for Approximate Level α Test**
$H_a: \mu_1 - \mu_2 > \Delta_0$	$t \geq t_{\alpha, \nu}$ (upper-tailed)
$H_a: \mu_1 - \mu_2 < \Delta_0$	$t \leq - t_{\alpha, \nu}$ (lower-tailed)
$H_a: \mu_1 - \mu_2 \neq \Delta_0$	either $t \geq t_{\alpha/2, \nu}$ or $t \leq -t_{\alpha/2, \nu}$ (two-tailed)

Example
cont'd

Consider the following data on two different types of plainweave fabric:

Fabric Type	Sample Size	Sample Mean	Sample Standard Deviation
Cotton	10	51.71	.79
Triacetate	10	136.14	3.59

Assuming that the porosity distributions for both types of fabric are normal, let's calculate a confidence interval for the difference between true average porosity for the cotton fabric and that for the acetate fabric, using a 95% confidence level

Example 6
cont'd

Assuming that the porosity distributions for both types of fabric are normal, let's calculate a confidence interval for the difference between true average porosity for the cotton fabric and that for the acetate fabric, using a 95% confidence level.

Before the appropriate t critical value can be selected, df must be determined:

$$\text{df} = \frac{\left(\dfrac{.6241}{10} + \dfrac{12.8881}{10}\right)^2}{\dfrac{(.6241/10)^2}{9} + \dfrac{(12.8881/10)^2}{9}} = \frac{1.8258}{.1850}$$

Example 6

Thus we use $v = 9$; $t_{.025,9} = 2.262$. The resulting interval is

$$51.71 - 136.14 \pm (2.262)\sqrt{\frac{.6241}{10} + \frac{12.8881}{10}} = -84.43 \pm 2.63$$

With a high degree of confidence, we can say that true average porosity for triacetate fabric specimens exceeds that for cotton specimens by between 81.80 and 87.06

Pooled *t* Procedure

Pooled *t* Procedures

Alternatives to the two-sample *t* procedures just described:

what if you know that the two populations are normal, AND also that they have equal variances? $(\sigma_1^2 = \sigma_2^2)$

That is, the two population distribution curves are assumed normal with equal spreads, the only possible difference between them being where they are centered.

Pooled *t* Procedures

Let σ^2 denote the common population variance. Then standardizing $\overline{X} - \overline{Y}$ gives

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma^2}{m} + \dfrac{\sigma^2}{n}}} = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma^2\left(\dfrac{1}{m} + \dfrac{1}{n}\right)}}$$

which has a standard normal distribution. Before this variable can be used as a basis for making inferences about $\mu_1 - \mu_2$, the common variance must be estimated from sample data.

Pooled *t* Procedures

One estimator of σ^2 is S_1^2, the variance of the *m* observations in the first sample, and another is S_2^2, the variance of the second sample. Intuitively, a better estimator than either individual sample variance results from combining the two sample variances.

A first thought might be to use $(S_1^2 + S_2^2)/2.$

However, if *m > n*, then the first sample contains more information about σ^2 than does the second sample, and an analogous comment applies if *m < n*.

Pooled *t* Procedures

The following *weighted* average of the two sample variances, called the **pooled** (i.e., combined) **estimator of** σ^2, adjusts for any difference between the two sample sizes:

$$S_p^2 = \frac{m-1}{m+n-2} \cdot S_1^2 + \frac{n-1}{m+n-2} \cdot S_2^2$$

The first sample contributes *m* – 1 degrees of freedom to the estimate of σ^2, and the second sample contributes *n* – 1 df, for a <u>total of **m + n – 2** df.</u>

Pooled *t* Procedures

It has been suggested that one could carry out a preliminary test of $H_0: \sigma_1^2 = \sigma_2^2$ and use a pooled *t* procedure if this null hypothesis is not rejected. This is the "*F* test" of equal variances.

Note the F-test is rather sensitive to the assumption of normal population distributions— more so than *t* procedures. We need normally distributed samples here.

The *F* Distribution

The *F* Distribution

The *F* probability distribution has two parameters, denoted by v_1 and v_2. The parameter v_1 is called the *numerator degrees of freedom*, and v_2 is the *denominator degrees of freedom*.

A random variable that has an *F* distribution cannot assume a negative value. Since the density function is complicated and will not be used explicitly, we omit the formula.

There is an important connection between an *F* variable and chi-squared variables.

The *F* Distribution

If X_1 and X_2 are independent chi-squared rv's with v_1 and v_2 df, respectively, then the rv

$$F = \frac{X_1/v_1}{X_2/v_2}$$

(the ratio of the two independent chi-squared variables divided by their respective degrees of freedom), can be shown to have an *F* distribution.

Mathematically, a chi-squared distribution can be obtained by summing **squared standardized Normal variables**. So a scaled ratio of two variances is a ratio of two scaled chi-squared variables.

The *F* Distribution

Figure below illustrates a typical *F* density function.

F density curve with ν_1 and ν_2 df

Shaded area $= \alpha$

F_{α, ν_1, ν_2}
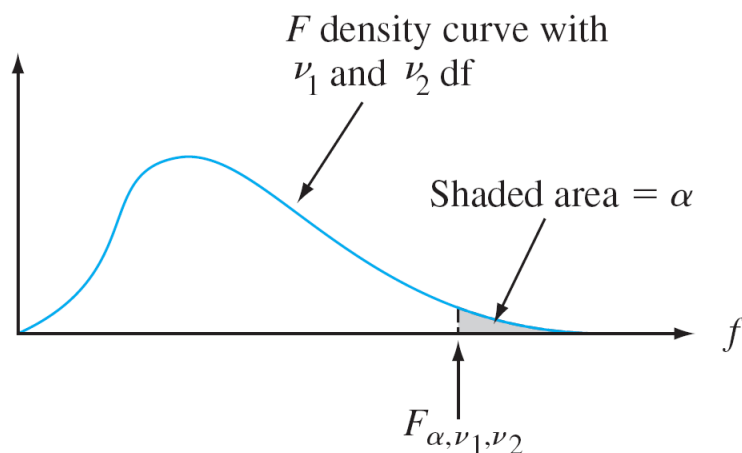
The *F* Distribution

Analogous to the notation $t_{\alpha,v}$ and $\chi^2_{\alpha,v}$, we use F_{α,v_1,v_2} for the value on the horizontal axis that captures α of the area under the *F* density curve with v_1 and v_2 df in the upper tail.

The density curve is not symmetric, so both upper- and lower-tail critical values must be found separately.

.

The *F* Test for Equality of Variances

The *F* Test for Equality of Variances

A test procedure for hypotheses concerning the ratio σ_1^2/σ_2^2 is based on the following result.

Theorem

Let $X_1,\ldots, X_m$ be a random sample from a normal distribution with variance σ_1^2, let $Y_1,\ldots, Y_n$ be another random sample (independent of the X_i's) from a normal distribution with variance σ_2^2, and let S_1^2 and S_2^2 denote the two sample variances. Then the rv

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

has an *F* distribution with $v_1 = m - 1$ and $v_2 = n - 1$.

The *F* Test for Equality of Variances

This theorem results from combining the fact that the variables $(m - 1)S_1^2/\sigma_1^2$ and $(n - 1)S_2^2/\sigma_2^2$ each have a chi-squared distribution with $m - 1$ and $n - 1$ df, respectively.

Because *F* involves a ratio rather than a difference, the test statistic is the ratio of sample variances.

The claim that $\sigma_1^2 = \sigma_2^2$ is then rejected if the ratio differs by too much from 1.

The *F* Test for Equality of Variances

Null hypothesis: $H_0: \sigma_1^2 = \sigma_2^2$

Test statistic value: $f = s_1^2/s_2^2$

Alternative Hypothesis	**Rejection Region for a Level α**
$H_a: \sigma_1^2 > \sigma_2^2$	$f \geq F_{\alpha,m-1,n-1}$
$H_a: \sigma_1^2 < \sigma_2^2$	$f \leq F_{1-\alpha,m-1,n-1}$
$H_a: \sigma_1^2 \neq \sigma_2^2$	either $f \geq F_{\alpha/2,m-1,n-1}$ or $f \leq F_{1-\alpha/2,m-1,n-1}$

Example

On the basis of data reported in the article "Serum Ferritin in an Elderly Population" (*J. of Gerontology*, 1979: 521–524), the authors concluded that the ferritin distribution in the elderly had a smaller variance than in the younger adults. (Serum ferritin is used in diagnosing iron deficiency.)

For a sample of 28 elderly men, the sample standard deviation of serum ferritin (mg/L) was $s_1 = 52.6$; for 26 young men, the sample standard deviation was $s_2 = 84.2$.

Does this data support the conclusion as applied to men?

Example
cont'd

Let σ_1^2 and σ_2^2 denote the variance of the serum ferritin distributions for elderly men and young men, respectively. The hypotheses are $H_0: \sigma_1^2 = \sigma_2^2$ and $H_a: \sigma_1^2 < \sigma_2^2$.

At level .01, H_0 will be rejected if $f \le F_{.99, 27, 25}$. To obtain the critical value, we use `qf(0.01, 27, 25)` function in R, which yields 0.394.

The computed value of F is $(52.6)^2/(84.2)^2 = 0.390$.
As $0.390 \le 0.394$, H_0 is rejected at level .01 in favor of H_a: the variability appears greater in young than in elderly men.

A Confidence Interval for σ_1/σ_2

A Confidence Interval for σ_1/σ_2

The CI for σ_1^2/σ_2^2 is based on the probability statement

$$P(F_{1-\alpha/2, v_1, v_2} < F < F_{\alpha/2, v_1, v_2}) = 1 - \alpha$$

by the F variable and manipulating the inequalities to isolate σ_1^2/σ_2^2.

An interval for σ_1/σ_2 results from taking the square root of each limit.

Two proportions

Difference Between Population Proportions

Having presented methods for comparing the means of two different populations, we now turn attention to the comparison of two population proportions.

Regard an individual or object as a success S if some characteristic of interest is present ("graduated from college", a refrigerator "with an icemaker", etc.).

Let
  p_1 = the true proportion of S's in population # 1
  p_2 = the true proportion of S's in population # 2

Inferences Concerning a Difference Between Population Proportions

Alternatively, p_1 can be regarded as the probability that a randomly selected individual or object from the first population is a success.

Suppose that a sample of size m is selected from the first population and independently a sample of size n is selected from the second one.

Let X denote the number of S's in the first sample and Y be the number of S's in the second.

Independence of the two samples implies that X and Y are independent.

Inferences Concerning a Difference Between Population Proportions

Provided that the two sample sizes are much smaller than the corresponding population sizes, X and Y can be regarded as having binomial distributions.

The natural estimator for $p_1 - p_2$, the difference in population proportions, is the corresponding difference in sample proportions $X/m - Y/n$.

Proposition

Let $\hat{p}_1 = X/m$ and $\hat{p}_2 = Y/n$, where $X \sim \text{Bin}(m, p_1)$ and $Y \sim \text{Bin}(n, p_2)$ with X and Y independent variables. Then

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

So $\hat{p}_1 - \hat{p}_2$ is an unbiased estimator of $p_1 - p_2$, and

$$V(\hat{p}_1 - \hat{p}_2) = \frac{p_1 q_1}{m} + \frac{p_2 q_2}{n} \quad \text{(where } q_i = 1 - p_i)$$

A Large-Sample Test Procedure

A Large-Sample Test Procedure

The most general null hypothesis an investigator might consider would be of the form $H_0: p_1 - p_2 = \Delta_0$.

Although for population means the case $\Delta_0 \neq 0$ presented no difficulties, for population proportions $\Delta_0 = 0$ and $\Delta_0 \neq 0$ must be considered separately.

Since the vast majority of actual problems of this sort involve $\Delta_0 = 0$ (i.e., the null hypothesis $p_1 = p_2$). we'll concentrate on this case.

When $H_0: p_1 - p_2 = 0$ is true, let p denote the common value of p_1 and p_2 (and similarly for q).

A Large-Sample Test Procedure

Then the standardized variable

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{pq\left(\frac{1}{m} + \frac{1}{n}\right)}}$$

has approximately a standard normal distribution when H_0 is true.

However, this Z cannot serve as a test statistic because the value of p is unknown—H_0 asserts only that there is a common value of p, but does not say what that value is.

A Large-Sample Test Procedure

A test statistic results from replacing p and q in by appropriate estimators.

Assuming that $p_1 = p_2 = p$, instead of separate samples of size m and n from two different populations (two different binomial distributions), we really have a single sample of size $m + n$ from one population with proportion p.

The total number of individuals in this combined sample having the characteristic of interest is $X + Y$.

The natural estimator of p is then

$$\hat{p} = \frac{X + Y}{m + n} = \frac{m}{m + n} \cdot \hat{p}_1 + \frac{n}{m + n} \cdot \hat{p}_2$$

A Large-Sample Test Procedure

The second expression for $\hat{p}$ shows that it is actually a weighted average of estimators $\hat{p}$ and $\hat{p}_2$ obtained from the two samples.

Using $\hat{p}$ and $\hat{q} = 1 - \hat{p}$ in place of p and q gives a test statistic having approximately a standard normal distribution when H_0 is true.

Null hypothesis: $H_0: p_1 - p_2 = 0$

Test statistic value (large samples): $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\dfrac{1}{m} + \dfrac{1}{n}\right)}}$

A Large-Sample Test Procedure

Alternative Hypothesis	Rejection Region for Approximate Level α Test
$H_a: p_1 - p_2 > 0$	$z \geq z_a$
$H_a: p_1 - p_2 < 0$	$z \leq -z_a$
$H_a: p_1 - p_2 \neq 0$	either $z \geq z_{a/2}$ or $z \leq -z_{a/2}$

A P-value is calculated in the same way as for previous z tests.

The test can safely be used as long as $m\hat{p}_1, m\hat{q}_1, n\hat{p}_2,$ and $n\hat{q}_2$ are all at least 10.

Example 11

The article "Aspirin Use and Survival After Diagnosis of Colorectal Cancer" (*J. of the Amer. Med. Assoc.,* 2009: 649–658) reported that of 549 study participants who regularly used aspirin after being diagnosed with colorectal cancer, there were 81 colorectal cancer-specific deaths, whereas among 730 similarly diagnosed individuals who did not subsequently use aspirin, there were 141 colorectal cancer-specific deaths.

Does this data suggest that the regular use of aspirin after diagnosis will decrease the incidence rate of colorectal cancer-specific deaths? Let's test the appropriate hypotheses using a significance level of .05.

Example 11

The parameter of interest is the difference $p_1 - p_2$, where p_1 is the true proportion of deaths for those who regularly used aspirin and p_2 is the true proportion of deaths for those who did not use aspirin.

The use of aspirin is beneficial if $p_1 < p_2$ which corresponds to a negative difference between the two proportions.

The relevant hypotheses are therefore

$H_0: p_1 - p_2 = 0$ versus $H_a: p_1 - p_2 < 0$

Example 11

Parameter estimates are $\hat{p}_1 = 81/549 = .1475$, $\hat{p}_2 = 141/730 = .1932$ and $\hat{p} = (81 + 141)/(549 + 730) = .1736$.

A z test is appropriate here because all of $m\hat{p}_1, m\hat{q}_1, n\hat{p}_2$, and $n\hat{q}_2$ are at least 10. The resulting test statistic value is

$$z = \frac{.1475 - .1932}{\sqrt{(.1736)(.8264)\left(\frac{1}{549} + \frac{1}{730}\right)}} = \frac{-.0457}{.021397} = -2.14$$

The corresponding P-value for a lower-tailed z test is $\Phi(-2.14) = .0162$.

Example 11

Because $.0162 \leq .05$, the null hypothesis can be rejected at significance level .05.

So anyone adopting this significance level would be convinced that the use of aspirin in these circumstances is beneficial.

However, someone looking for more compelling evidence might have selected a significance level .01 and then would not have been persuaded.

A Large-Sample Confidence Interval

A Large-Sample Confidence Interval

As with means, many two-sample problems involve the objective of comparison through hypothesis testing, but sometimes an interval estimate for $p_1 - p_2$ is appropriate.

Both $\hat{p}_1 = X/m$ and $\hat{p}_2 = Y/n$ have approximate normal distributions when m and n are both large.

If we identify θ with $p_1 - p_2$, then $\hat{\theta} = \hat{p}_1 - \hat{p}_2$ satisfies the conditions necessary for obtaining a large-sample CI.

In particular, the estimated standard deviation of $\hat{\theta}$ is $\sqrt{(\hat{p}_1\hat{q}_1/m) + (\hat{p}_2\hat{q}_2/n)}$.

A Large-Sample Confidence Interval

The general $100(1 - \alpha)\%$ interval $\hat{\theta} \pm z_{\alpha/2} \cdot \hat{\sigma}_{\hat{\theta}}$ then takes the following form.

A CI for $p_1 - p_2$ with confidence level approximately $100(1 - \alpha)\%$ is

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{m} + \frac{\hat{p}_2\hat{q}_2}{n}}$$

This interval can safely be used as long as $m\hat{p}_1, m\hat{q}_1, n\hat{p}_2,$ and $n\hat{q}_2$ are all at least 10.

Example 13

The authors of the article "Adjuvant Radiotherapy and Chemotherapy in Node- Positive Premenopausal Women with Breast Cancer" (*New Engl. J. of Med.,* 1997: 956–962) reported on the results of an experiment designed to compare treating cancer patients with chemotherapy only to treatment with a combination of chemotherapy and radiation.

Of the 154 individuals who received the chemotherapy-only treatment, 76 survived at least 15 years, whereas 98 of the 164 patients who received the hybrid treatment survived at least that long.

Example 13

cont'd

With p_1 denoting the proportion of all such women who, when treated with just chemotherapy, survive at least 15 years and p_2 denoting the analogous proportion for the hybrid treatment, $\hat{p}_1 = 76/154 = .494$ and $98/164 = .598$.

A confidence interval for the difference between proportions based on the traditional formula with a confidence level of approximately 99% (qnorm(0.995) = 2.58) is

$$.494 - .598 \pm (2.58)\sqrt{\frac{(.494)(.506)}{154} + \frac{(.598)(.402)}{164}} = -.104 \pm .143$$

$$= (-.247, .039)$$

Example 13
cont'd

We are 99% confident that (−0.247 , 0.039) covers the true difference between the two proportions.

Notice that 0 is one of the plausible values of $p_1 - p_2$, so it's tempting to use this CI to conclude that neither treatment can be judged superior to the other.

However in the case of two sample proportions, a better test can be used for testing H0 of equality, using combined p, as in the slide 71.

Example 13
cont'd

Null hypothesis: $H_0: p_1 - p_2 = 0$ versus $H_a: p_1 - p_2 =/= 0$

Test statistic value (large samples): $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\dfrac{1}{m} + \dfrac{1}{n}\right)}}$

Here, the combined p = (76+98)/(154+164) = .55

So,
z = -0.104 / sqrt(0.55*0.45*(1/154 + 1/164)) = -1.86

Thus, since the p-value (2*pnorm(-1.86) = 0.06) of this test statistic is greater than 1%, we cannot reject this null at the significance level of 1%.

Example
cont'd

H0: p1=p2
Ha: p1=/=p2

Under H0:
p.hat = (76+98)/(154+164)
se.hat = sqrt(p.hat*(1-p.hat)*(1/154+1/164))

p1.hat = 76/154
p2.hat = 98/164
P-value: 2*pnorm((p1.hat-p2.hat – 0)/se.hat)
[1] 0.06 => Do not reject H0

Small-Sample Inferences

Small-Sample Inferences

On occasion an inference concerning $p_1 - p_2$ may have to be based on samples for which at least one sample size is small.

Appropriate methods for such situations are not as straightforward as those for large samples, and there is more controversy among statisticians as to recommended procedures.

One frequently used test, called the Fisher–Irwin test, is based on the hypergeometric distribution.

Your friendly neighborhood statistician can be consulted for more information.

R code example

t.test

The t.test( ) function performs several t-tests for normal data, or large sample(s)
Default assumes unequal variances

```
# independent 2-group t-test
t.test(y~x) # where y is numeric and x is a binary factor

# independent 2-sample t-test
t.test(y1,y2) # where y1 and y2 are numeric

# paired t-test
t.test(y1,y2,paired=TRUE) # where y1 & y2 are numeric

# one sample t-test
t.test(y,mu=3) # Ho: mu=3
```

You can use the var.equal = TRUE option to ask for a pooled variance.

You can use the alternative="less" or alternative="greater" option to specify a one tailed test.

F test: Compare Two Variances in R

The var.test( ) function performs the equality of variances test for normal data

```
# When samples are in a single dataset, with 2 labels
var.test( values ~ label, data,  alternative = "two.sided")

# when 2 separate samples
var.test(Xsample, Ysample, alternative = "two.sided")
```

KS test: Kolmogorov-Smirnov Tests

ks.test() function performs a one- or two-sample Kolmogorov-Smirnov test

The two-sample test is for the null hypothesis that x and y were drawn from the same continuous distribution

The one sample test is for the null hypothesis that x is drawn from the specified continuous distribution. It needs "y" to be a character string naming a continuous (cumulative) distribution function, like pnorm

```
ks.test(x, y, alternative = c("two.sided", "less", "greater"))

ks.test(x, "pnorm" …)
# note: … is where you specify parameters
```

Example

```
#Pesticide trees:
Pesticide = c(55.57109, 36.50319, 47.80090, 33.34822, 36.16251,
35.28337, 41.50154, 44.18931, 40.81439, 33.88648, 44.90427, 49.97089,
22.85414, 27.84301, 38.49843)

#Ladybug trees:
Ladybug = c(45.44505, 35.52320, 46.97865, 45.76921, 41.66216,
54.69599, 58.77678, 49.08538, 48.53812, 70.17137, 51.86253, 39.59365,
42.10194, 47.39945, 39.04648)
```

First, let's test for normality of each sample:
```
ks.test(Pesticide, "pnorm", mean(Pesticide), sqrt(var(Pesticide)))
ks.test(Ladybug, "pnorm", mean(Ladybug), sqrt(var(Ladybug)))
```

Then, the var.test( ) function performs the equality of variances test for normal data
```
var.test(Ladybug,Pesticide)
```

Finally, we can run the appropriate t.test( )
```
t.test(Ladybug,Pesticide, var.equal = TRUE)
```

Testing difference of two means with non-independent samples

How do we think about a test statistic when the two sample means are not independent?

The covariance between the two sample means is not 0

How do we estimate it?

Sometimes, in a special situation when data are "matched" or "paired", we can compute differences between the matched (or paired) elements of two samples.

Then, we would end up with a single sample of differences. We would then proceed doing one-sample inference, and test the true mean difference being equal to 0.

Testing difference of two means with non-independent samples

Ten pregnant women were given an injection of pitocin to induce labor. Their systolic blood pressures immediately before and after the injection were:

Patient	Before	After
1	134	140
2	122	130
3	132	135
4	130	126
5	128	134
6	140	138
7	118	124
8	127	126
9	125	132
10	142	144

a. Perform the appropriate hypothesis test at the $\alpha = .05$ level to determine whether the average systolic blood pressure of pregnant women changes after the administering of pitocin.

Testing difference of two means with non-independent samples

Before	After	Difference
134	140	-6
122	130	-8
132	135	-3
130	126	4
128	134	-6
140	138	2
118	124	-6
127	126	1
125	132	-7
142	144	-2

```
before = c(134,122,132,130,128,140,118,127,125,142)
after  = c(140,130,135,126,134,138,124,126,132,144)

diff = before-after
t.test(diff)

    One Sample t-test

data:  diff
t = -2.333, df = 9, p-value = 0.04453
alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:
 -6.10582345 -0.09417655

sample estimates:
mean of x   -3.1
```