9

Inference Based on Two Samples

Chapter 9 Stat 4570/5570

Material from Devore's book (Ed 8), and Cengage

Difference Between Two Population Means

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 cig brand A have the same amount of nicotine as brand B?
3. Is this class of students more statistically savvy than the last?

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 $\mu_1 = \mu_2$.

Example: Difference Between Two Population Means

Forty patients were randomly assigned to a treatment group (m = 20) or a control group (n = 20).

One patient in the treatment group and four in the control group dropped out because of complications.

The data analysis is then based on two random samples, *i.e.* the treatment group (m = 19) and control group (n = 16).

We need some basic assumptions

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

2. Y_1, Y_2,…..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.

Note – no distribution form assumed (for now)

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 $\overline{X} - \overline{Y}$.

Difference Between Two Population Means

Proposition

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 $\sigma_{\overline{X}-\overline{Y}}$.

Test Procedures for Normal Populations with Known Variances

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

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

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

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

A second random sample of 25 two-sided galvanized steel specimens gave a sample average strength of
$\bar{y} = 34.7$ 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 $\alpha = 0.01$

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. WHY?

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:

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

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 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.

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 $\bar{x} - \bar{y} - z_{\alpha/2} \cdot \sigma_{\overline{X}-\overline{Y}}$ and upper limit $\bar{x} - \bar{y} + z_{\alpha/2} \cdot \sigma_{\overline{X}-\overline{Y}}$, where $\sigma_{\overline{X}-\overline{Y}}$ is the square-root expression.

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 $\mu_1 - \mu_2$

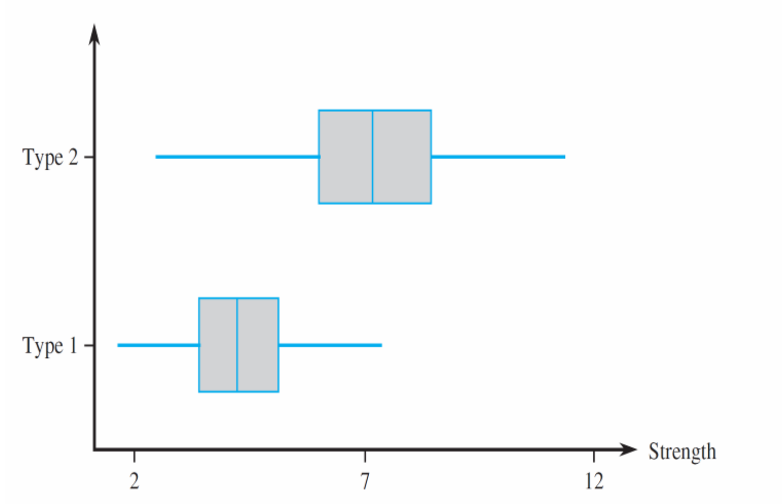
Provided that m and n are both large, a CI for $\mu_1 - \mu_2$ with a confidence level of approximately $100(1 - \alpha)\%$ is

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

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.



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

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%.

Not-so-large Sample Tests

The Two-Sample *t* Test and Confidence Interval

For large samples, the CLT allows us to use these methods we have discussed even when the two populations of interest are not normal.

In practice, 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

When the population distribution are both normal, the standardized variable

$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1^2}{m} + \dfrac{S_2^2}{n}}} \qquad \textbf{(9.2)}$$

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

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

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

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

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

Test statistic value: t =

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

The Two-Sample t Test and Confidence Interval

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

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 $(S_1^2 + S_2^2)/2$. Why might this be a problem?

Pooled *t* Procedures

The following *weighted* average of the two sample variances, called the **pooled** (i.e., combined) **estimator of** $\boldsymbol{\sigma^2}$,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 total of $\boldsymbol{m + n - 2}$ df.

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 quite sensitive to the assumption of normal population distributions—much 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. The density function is complicated and will not be used explicitly, so it's not shown.

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

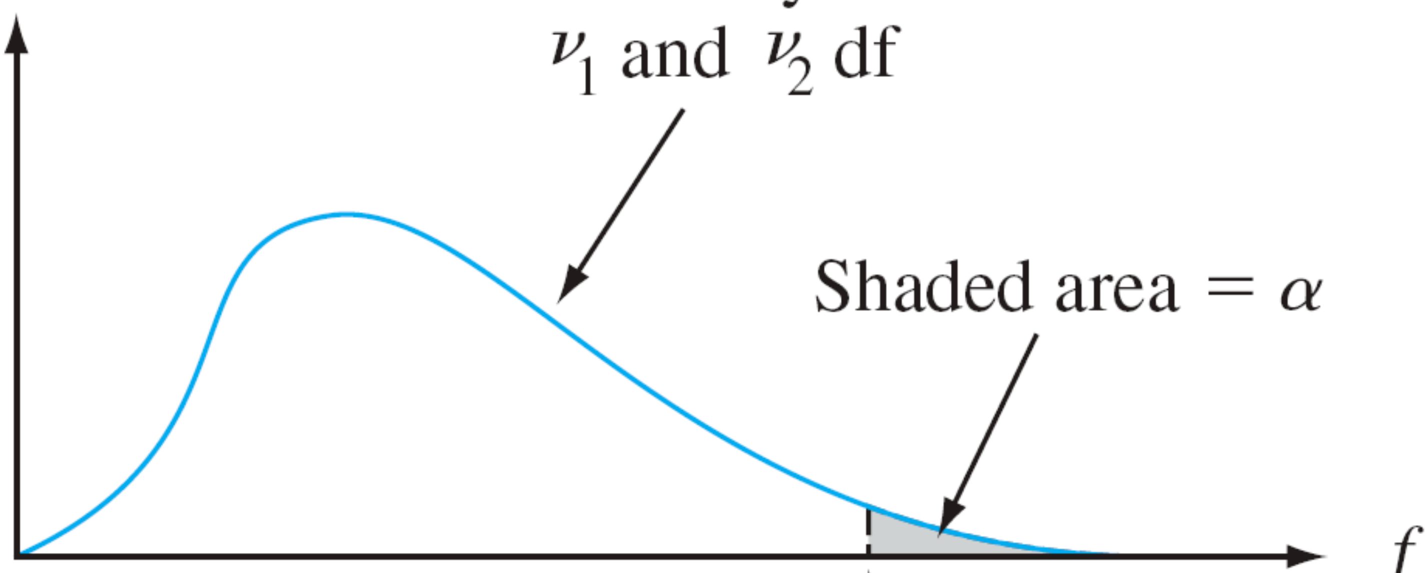
$$F = \frac{X_1/v_1}{X_2/v_2}$$

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

Recall that a chi-squared distribution was obtain by summing squared standard Normal variables (such as squared deviations for example). 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.



The *F* Distribution

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 it would seem that both upper- and lower-tail critical values must be tabulated. This is not necessary, though, because of the fact that

$$F_{1-\alpha,v_1,v_2} = 1/F_{\alpha,v_2,v_1}.$$

The *F* Distribution

F-testing is used in a lot in statistics for scaled ratios of "squared" (often "sums of squares") quantities.

Appendix Table A.9 gives F_{α,v_1,v_2} for α = .10, .05, .01, and .001, and various values of v_1 (in different columns of the table) and v_2 (in different groups of rows of the table).

For example, $F_{.05,6,10} = 3.22$ and $F_{.05,10,6} = 4.06$. The critical value $F_{.95,6,10}$, which captures .95 of the area to its right (and thus .05 to the left) under the *F* curve with $v_1 = 6$ and $v_2 = 10$, is $F_{.95,6,10} = 1/F_{.05,10,6} = 1/4.06 = .246$.

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: $H_0\colon \sigma_1^2 = \sigma_2^2$

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

Alternative Hypothesis	**Rejection Region for a Level α Test**
$H_a\colon \sigma_1^2 > \sigma_2^2$	$f \geq F_{\alpha,m-1,n-1}$
$H_a\colon \sigma_1^2 < \sigma_2^2$	$f \leq F_{1-\alpha,m-1,n-1}$
$H_a\colon \sigma_1^2 \neq \sigma_2^2$	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? Use alpha = .01.

9.4 Inference Concerning a Difference Between Population 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

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.

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

$$V(\hat{p}_1 - \hat{p}_2) = \frac{p_1q_1}{m} + \frac{p_2q_2}{n} \quad \text{(where } q_i = 1 - p_i\text{)}$$

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

$$\hat{p} = \frac{X + Y}{m + n} = \frac{m}{m + n} \cdot \hat{p}_1 + \frac{n}{m + n} \cdot \hat{p}_2 \qquad (9.5)$$

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* in (9.4) 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.

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 – α)% 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 – α)% 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.

What is the 99% confidence interval for this difference in proportions?

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.