Now that we know how to compute power as a function of sample size, this can be easily turned around to give sample size for a given power.

E.g., for a one-sided alternative, the power of a z-test of $H_0 : p = p_0$ when p_1 is the true value of p is

$$\text{power} = P(Z < -\tilde{z}) = P\left(Z < \underbrace{-z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} + \frac{|p_0 - p_1|}{\sqrt{p_1(1-p_1)/n}}}_{=(*)} \right)$$

Therefore, to achieve power equal to $1-\beta$, say, $(*)$ must be the $100(1-\beta)$th percentile of the $N(0,1)$ distribution. That is, the necessary sample size is given by the solution of the equation

$$z_{1-\beta} = -z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} + \frac{|p_0 - p_1|}{\sqrt{p_1(1-p_1)/n}}$$

Solving this equation leads to the following general result:

General result for sample size to achieve power equal to $1-\beta$ in one sample z test of $H_0 : p = p_0$, when the true value of p is p_1:

$$n = \begin{cases} \dfrac{p_1(1-p_1)\left\{z_{1-\beta} + z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}}\right\}^2}{(p_0-p_1)^2} & \text{for a one-sided alternative} \\ \dfrac{p_1(1-p_1)\left\{z_{1-\beta} + z_{1-\alpha/2}\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}}\right\}^2}{(p_0-p_1)^2} & \text{for a two-sided alternative} \end{cases}$$

Breast Cancer Prevalence

- How many women should be sampled to achieve at least 90% power in the breast cancer study described previously?

Assuming a one-sided test of

$$H_0 : p = p_0 = .02 \quad \text{versus} \quad H_A : p > .02$$

conducted at level $\alpha = .05$ when the true prevalence is $p_1 = .05$, we want power $= 1 - \beta = .90$. Therefore,

$$\begin{aligned} n &= \frac{p_1(1-p_1)\left\{z_{1-\beta} + z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}}\right\}^2}{(p_0 - p_1)^2} \\ &= \frac{.05(1-.05)\left\{\overbrace{z_{.90}}^{1.282} + \overbrace{z_{1-.05}}^{1.645}\sqrt{\frac{.02(1-.02)}{.05(1-.05)}}\right\}^2}{(.02-.05)^2} \\ &= 288.67 \end{aligned}$$

which we round up to $n = 289$.

Two-Sample Inference for Proportions

The extension of normal-approximation-based inference methods for proportions from the one to the two sample case follows along the same lines as when we studied means of continuous random variables.

Example — Comparison of Drug Therapies

- In patients with congestive heart failure, two or more drugs were prescribed in 257 of 437 American patients.
- In Scotland, 39 of 179 patients were prescribed two or more drugs.

Is there a difference in the proportion of such patients prescribed two or more drugs between Scotland and the US?

In this example, there are two populations, patients with congestive heart failure (CHF) in the US, and patients with CHF in Scotland.

We are interested in the population proportions p_1 and p_2, the population proportions of CHF patients prescribed more than 1 drug in the respective populations.

We have a sample from each of these populations, and corresponding sample proportions:

$\hat{p}_1 =$ sample proportion from population 1 (US) $= x_1/n_1 = 257/437$

$\hat{p}_2 =$ sample proportion from population 2 (US) $= x_2/n_2 = 39/179$

where

$x_1 =$ number of patients prescribed > 1 drug out of n_1 patients in pop. 1

$x_2 =$ number of patients prescribed > 1 drug out of n_2 patients in pop. 2

- We assume that the two samples are independent of each other (not paired).

We wish to test

$$H_0 : p_1 = p_2 \quad \text{versus either} \quad \begin{cases} H_A : p_1 \neq p_2 & \text{(two-sided alternative) or} \\ H_A : p_1 < p_2 & \text{(“less than” alternative), or} \\ H_A : p_1 > p_2 & \text{(“greater than” alternative).} \end{cases}$$

We've already seen that, based on the normal approximation to the binomial, a single sample proportion is approximately normal: $\hat{p} \dot{\sim} N(p, p(1-p)/n)$, with mean the corresponding population proportion.

It can be shown that the difference between two independent sample proportions is approximately normal too. That is,

$$\hat{p}_1 - \hat{p}_2 \dot{\sim} N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right) \qquad (*)$$

- That is, the difference in sample proportions $\hat{p}_1 - \hat{p}_2$ has expected value equal to (i.e., is an unbiased estimate of) the corresponding population proportion $p_1 - p_2$.
- And the variance of $\hat{p}_1 - \hat{p}_2$ is $\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$ (the sum of the variances of $\hat{p}_1$ and $\hat{p}_2$).
- Result (*) is the basis of normal-approximation-based methods of inference on $p_1 - p_2$.

Testing:

Under $H_0 : p_1 = p_2$, the mean in (*) becomes $p_1 - p_2 = 0$ and the variance in (*) simplifies to

$$\text{var}(\hat{p}_1 - \hat{p}_2) = \frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2} = p(1-p)\left\{\frac{1}{n_1} + \frac{1}{n_2}\right\} \quad \text{if } H_0 : p_1 = p_2 \text{ is true}$$

where $p = p_1 = p_2$ is the common value of p under the null hypothesis.

In that case, the standardized version of (*), becomes

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)\left\{\frac{1}{n_1} + \frac{1}{n_2}\right\}}} \dot{\sim} N(0,1) \quad \text{if } H_0 : p_1 = p_2 \text{ is true}$$

We are interested in testing whether the population proportions are equal. The quantity above examines how similar the corresponding the sample proportions are, and it has a simple, known distribution under H_0.

Therefore, it is a perfect test statistic! ... Except for one thing: We don't know p, the common population proportion in the denominator.

Solution: replace p by its sample estimate.

Since we are estimating the common population proportion of "successes" (e.g., patient prescribed > 1 drug), it makes sense to estimate this quantity based on the combined sample. So, replace p by

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{\text{total no. of successes}}{\text{total no. of trials}} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Thus, our test statistic for $H_0 : p_1 = p_2$ is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left\{\frac{1}{n_1} + \frac{1}{n_2}\right\}}} \dot{\sim} N(0,1) \quad \text{if } H_0 : p_1 = p_2 \text{ is true}$$

- (Fortunately, replacing p by $\hat{p}$ doesn't change the distribution of this quantity. It is still standard normal.)

General Two-Sample z test for Proportions:

An approximate α-level test of $H_0 : p_1 = p_2$ has test statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left\{\frac{1}{n_1} + \frac{1}{n_2}\right\}}}$$

and we reject H_0 if

$$|z| > \begin{cases} z_{1-\alpha/2} & \text{if alternative is } H_A : p_1 \neq p_2 \\ z_{1-\alpha} & \text{if } z < 0 \text{ and alternative is } H_A : p_1 < p_2 \\ z_{1-\alpha} & \text{if } z > 0 \text{ and alternative is } H_A : p_1 > p_2 \end{cases}$$

or, equivalently, if the p-value is less than α where the p-value is given by

$$\begin{cases} 2P(Z > |z|) & \text{if alternative is } H_A : p_1 \neq p_2 \\ P(Z < z) & \text{if alternative is } H_A : p_1 < p_2 \\ P(Z > z) & \text{if alternative is } H_A : p_1 > p_2 \end{cases}$$

- The rule of thumb for validity of the normal approximation in this context is that $n_1p_1, n_1(1-p_1), n_2p_2, n_2(1-p_2)$ all be ≥ 5.

Example — Comparison of Drug Therapies

- In this example, the sample proportions are

$$\hat{p}_1 = \frac{x_1}{n_1} = \frac{257}{437} = .5881$$
$$\hat{p}_2 = \frac{x_2}{n_2} = \frac{39}{179} = .2179$$

  and the estimate of the common population proportion (assumed under H_0) would be

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{257 + 39}{437 + 179} = \frac{296}{616} = .4805$$

- Therefore, our test statistic is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left\{\frac{1}{n_1} + \frac{1}{n_2}\right\}}} = \frac{.5881 - .2179}{\sqrt{.4805(1-.4805)\left\{\frac{1}{437} + \frac{1}{179}\right\}}} = 8.3503$$

- For a two-sided alternative hypothesis $H_A : p_1 \neq p_2$, the critical value for an $\alpha = .05$-level test would be

$$z_{1-\alpha/2} = z_{1-.05/2} = z_{.975} = 1.96$$

  and the p-value here would be

$$2P(Z > |z|) = 2P(Z > 8.3503) = 2(.0000) = .0000$$

- Conclusion: we reject $H_0 : p_1 = p_2$ at level .05 and conclude that there is a difference in the population proportion of CHF patients who are prescribed > 1 drugs.

- The normal approximation based test can be justified in this example because $n_1 p_1 = 257, n_1(1-p_1) = 180, n_2 p_2 = 39, n_2(1-p_2) = 140$ are all ≥ 5.

- Note that a continuity-corrected version of the z test is available, but we will not study it. In addition, if the sample size is not sufficient to justify the z test, an exact method is available known as *Fisher's Exact Test*.

Confidence Intervals:

Instead of testing $H_0 : p_1 = p_2$, we may be more interested in estimating $p_1 - p_2$ and forming a confidence interval for this quantity.

Methods for doing this based upon the normal approximation to the binomial are closely related to the methods we have just described for hypothesis testing.

Standardizing (*), we have

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \dot{\sim} N(0,1)$$

Therefore, we can make statements like

$$P\left(-1.96 \leq \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \leq 1.96\right) \approx .95$$

or, more generally,

$$P\left(-z_{1-\alpha/2} \leq \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \leq z_{1-\alpha/2}\right) \approx 1 - \alpha$$

Rearranging, we get the following general result:

General formula for $100(1-\alpha)\%$ CI for $p_1 - p_2$ (normal approximation):

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- Note the difference in the standard error of $\hat{p}_1 - \hat{p}_2$ here versus in the test of $H_0 : p_1 = p_2$. Here we use $\hat{p}_1$ and $\hat{p}_2$ rather than the combined estimate $\hat{p}$.
- This difference is because the standard error is computed under H_0 for the test statistic, but H_0 is not assumed true when computing the standard error for a confidence interval.

Example — Comparison of Drug Therapies

- Suppose we want a 90% confidence interval for $p_1 - p_2$, the difference between the population proportions of CHF patients receiving > 1 drug in the US versus Scotland.

Answer:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$= (.5881 - .2179) \pm \underbrace{z_{1-.1/2}}_{=1.645}\sqrt{\frac{.5881(1-.5881)}{437} + \frac{.2179(1-.2179)}{179}}$$

$$= (.306, .434)$$

- In Minitab, normal approximation-based testing of $H_0 : p_1 = p_2$ and confidence intervals for $p_1 - p_2$ are obtained by selecting

Stat $\rightarrow$ Basic Statistics $\rightarrow$ 2 Proportions...

- Note that for testing, go into "Options" and place a check next to "Use pooled estimate of p for test". Whether or not this option is checked has no effect on the confidence interval computations.

Contingency Tables*

In the previous section (Ch.14) we presented a z test of $H_0 : p_1 = p_2$, for population proportions p_1, p_2, based upon independent samples.

We now revisit that problem from another perspective, based upon organizing and conceptualizing the data using **contingency tables**.

We will not only revisit the z test of $H_0 : p_1 = p_2$, but will develop other methods for analyzing categorical/discrete data using contingency tables.

Example — Comparison of Drug Therapies

Consider again the data from Scotland and the US on the number of CHF patients prescribed > 1 drug for treatemnt of their condition. In sample data, we found that 257 of 437 US patients and 39 of 179 Scottish patients were prescribed > 1 drug.

A convenient way to display these data is by using a contingency table, where we tabulate the number of patients in each of the four groups defined by the combinations of country (US vs Scotland) and drug prescription (≤ 1 vs. > 1):

		Number of Drugs		
		> 1	≤ 1	
Country	US	257	180	437
	Scotland	39	140	179
		296	320	616

- The above table is a 2×2 **contingency table** because it has 2 rows and 2 columns. Later, we will discuss more general $r \times c$ tables, which are contingency tables with $r \geq 2$ rows and $c \geq 2$ columns.

* Read Ch.15 of our text.

- Also displayed in the above table are
  - the row marginal totals, which are usually called **row margins**;
  - the column marginal totals, which are usually called **column margins**;
  - the grand total, which is the sum of all the cells in the table, or euivalently, the sum of the row margins, or of the column margins.

In general, we can represent a 2×2 table as

	Column 1	Column 2	
Row 1	a	b	$a+b$
Row 2	c	d	$c+d$
	$a+c$	$b+d$	n

- Here, the grand total is $n = a + b + c + d$.

Another Example — Oral Contraceptive Use and Myocardial Infarction

- A prospective cohort study was conducted in which 5000 oral contraceptive (OC) users and 10,000 OC non-users were followed for 3 years.
- It was found that 13 of the OC users and 7 of the non-users experienced a myocardial infarction (MI) during the period of the study.

Here are the data:

		MI		
		Yes	No	
OC Use	Yes	13	4987	5000
	No	7	9993	10000
		20	14980	15000

Is there an association between OC use and MI?

- In both this example and the last, we have two independent samples, corresponding to the rows of the contingency table.
- In each case, the two sample sizes are the row margins, and these quantities are fixed by the design of the study.
  - We went out and selected 5000 OC users and 10000 non-users. How many OC users and non-users we selected was a design decision.

In such cases, the natural hypothesis to test is $H_0 : p_1 = p_2$, where

$p_1 =$ pop. proportion with response given by col 1
for the population corresponding to row 1

$p_2 =$ pop. proportion with response given by col 1
for the population corresponding to row 2

- This hypothesis is sometimes called the hypothesis of homogeneous binomial proportions.
  - If there are two rows, then it is just $p_1 = p_2$. However, more generally there might be three, four, or more rows (e.g., we might look at proportion prescribing > 1 drugs in the US, Scotland, Canada, France, etc.). In that case, homogeneous binomial proportions would be $p_1 = p_2 = \cdots = p_r$.
- Note that under the alternative hypothesis, the population proportion falling in column 1 depends upon the row, and under the null hypothesis, the population proportion in column 1 does not depend upon the row.
  - That is, under the alternative hypothesis there is an association between the variable defining the rows and the variable defining the columns. The null hypothesis is that there is no association between these variables (i.e., the row and column variables are independent).
- So, in the 2×2 contingency table with fixed row totals, testing for homogeneous (i.e., equal) proportions among the rows, is equivalent to testing independence versus association between rows and columns.

- There are many types of contingency tables where the idea of testing homogeneity of binomial proportions does not make sense (e.g., tables with fixed row totals but > 2 columns, tables involving paired data, etc.).

- However, the idea of testing association between rows and columns is almost always sensible for contingency table data, so from now on, we emphasize testing for association/independence rather than testing homogeneous (equal) proportions.

Chi-Square Test of Independence in a Two-way Contingency Table:

The general tool for testing independence versus association between the rows and columns of a two-way contingency table is the **Chi-Square Test.**

- The name of this test comes from the distribution of the test statistic, which is called the chi-square of χ^2 distribution.

- Like the normal (z), the t, and the F distributions, the chi-square distribution is a parametric distribution. That is, the chi-square distribution is completely determined by its parameter, which is called its *degrees of freedom.*

  - We will write $\chi^2(d)$ to denote the chi-square with d degrees of freedom.

  - E.g., $Y \sim \chi^2(3)$ means the random variable Y follows a chi-square distribution on 3 degrees of freedom.

  - We will sometimes want percentiles/critical values of the chi-square. We will denote the $100(1-\alpha)$th percentile (i.e., the upper αth critical value) of the $\chi^2(d)$ distribution as

    $\chi^2_{1-\alpha}(d) = 100(1-\alpha)$th percentile of the chi-square on d d.f.

In all of the tests we have looked at so far, our test statistics have been based on comparing a sample statistic (e.g., $\bar{x}$) to its expected value under the null hypothesis (e.g., μ_0).

- The same is true for testing association in a contingency table.

Specifically, we will compare the **observed cell counts** in the table (a, b, c, d) to their expected values, or **the expected cell counts**, computed under the null hypothesis of no association between rows and columns.

How do we compute the expected cell count?

Example — Comparison of Drug Therapies

Recall the drug prescriptions for CHF data again:

		Number of Drugs		
		> 1	≤ 1	
Country	US	$a = O_{11} = 257$	$b = O_{12} = 180$	$a + b = 437$
	Scotland	$c = O_{21} = 39$	$d = O_{22} = 140$	$c + d = 179$
		$a + c = 296$	$b + d = 320$	$n = 616$

- Here, we've used both a, b, c, d to denote the observed cell counts, but we've also used the notation O_{ij} to denote the observed count in the i, jth cell.

Under the null hypothesis of no association between country and whether or not > 1 drug has been prescribed, the two samples from the US and Scotland can be thought of as samples from a common, single population, where the population proportion of patients prescribed > 1 drug is $p = p_1 = p_2$.

Therefore, the two samples can be pooled together, and the best estimate of p is the sample proportion prescribed > 1 drug in both countries combined:

$$\hat{p} = \frac{a + c}{n} = \frac{296}{616}$$

Applying that estimated probability to the number of US patients, we would expect that

$$\hat{p}(a+b) = \frac{(a+c)}{n}(a+b) = \frac{(296)(437)}{616} = 209.99$$

US CHF patients would be prescribed > 1 drug, and the remainder,

$$437 - 209.99 = 227.01$$

would be prescribed ≤ 1 drug under the null hypothesis.

Similarly, applying $\hat{p}$ to the number of Scottish patients, we would expect that

$$\hat{p}(c+d) = \frac{(a+c)}{n}(c+d) = \frac{(296)(179)}{616} = 86.01$$

Scottish CHF patients would be prescribed > 1 drug, and the remainder,

$$179 - 86.01 = 92.99$$

would be prescribed ≤ 1 drug under the null hypothesis.

Therefore, our expected cell counts are

	Number of Drugs > 1	Number of Drugs ≤ 1	
US	$E_{11} = \frac{(a+b)(a+c)}{n} = 209.99$	$E_{12} = \frac{(a+b)(b+d)}{n} = 227.01$	$a+b = 437$
Scotland	$E_{21} = \frac{(c+d)(a+c)}{n} = 86.01$	$E_{22} = \frac{(c+d)(b+d)}{n} = 92.99$	$c+d = 179$
	$a+c = 296$	$b+d = 320$	$n = 616$

- Notice the form of all of these expected cell counts is the same:

  the expected cell count is the product of the corresponding row and column margins, divided by n.

With observed and expected cell counts now computed, we are ready to form our test statistic.

The chi-square test of no association between rows and columns has test statistic which we will denote by X^2.

X^2 is computed by comparing observed and expected cell counts for each cell via the formula $\frac{(O-E)^2}{E}$, and then summing these quantities up over all cells.

That is, X^2 is given by

$$\begin{aligned} X^2 &= \frac{(O_{11}-E_{11})^2}{E_{11}} + \frac{(O_{12}-E_{12})^2}{E_{12}} + \frac{(O_{21}-E_{21})^2}{E_{21}} + \frac{(O_{22}-E_{22})^2}{E_{22}} \\ &= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \end{aligned}$$

where r and c are the number of rows and columns in the table, respectively.

- Under the null hypothesis of no association,

$$X^2 \dot{\sim} \chi^2(d), \quad \text{where} \quad d = \begin{cases} (r-1)(c-1) & \text{in general, for an } r \times c \text{ table} \\ 1 & \text{for a } 2 \times 2 \text{ table} \end{cases}$$

- So, we reject the null hypothesis of independence between rows and columns at level α if the test statistic X^2 exceeds $\chi^2_{1-\alpha}(d)$.
  - The alternative hypothesis here is a two sided one, of general association between rows and columns.
- The (approximate) p-value of this test is given by the area under the $\chi^2(d)$ distribution to the right of X^2, which can be computed from a computer program such as Minitab.

In our drug prescriptions for CHF example, our test statistic is

$$\begin{aligned} X^2 &= \frac{(O_{11}-E_{11})^2}{E_{11}} + \frac{(O_{12}-E_{12})^2}{E_{12}} + \frac{(O_{21}-E_{21})^2}{E_{21}} + \frac{(O_{22}-E_{22})^2}{E_{22}} \\ &= \frac{(257-209.99)^2}{209.99} + \frac{(180-227.01)^2}{227.01} + \frac{(39-86.01)^2}{86.01} + \frac{(140-92.99)^2}{92.99} \\ &= 69.73 \end{aligned}$$

Table A.8 in the back of our book contains critical values of the $\chi^2(d)$ distributions. From that table we can find that

$$\chi^2_{1-.05}(1) = 3.84$$

- Therefore, since $X^2 = 69.73$ exceeds the $\alpha = .05$-level critical value 3.84, we reject the null hypothesis of independence, and conclude that there the population proportion, or probability of prescribing > 1 drug depends upon country.
- The p-value for this test can be obtained from Minitab. It is very small in this case; less than .0001.
- Note that this test is equivalent to the z test we conducted back on p.206. In fact, the X^2 statistic here is just the square of the z statistic ($69.73 = 8.3503^2$) and the p values are identical.
  - However, the chi-square test applies to a much broader class of data types that can be summarized in terms of contingency tables.
- From the equivalence between the chi-square test and the z test for proportions, it should be clear that the chi-square test is based upon a normal approximation, which comes from the CLT.
- Therefore, the chi-square test should only be used when the sample size is reasonably large. The usual rule of thumb requires
  - 2×2 table: all of the expected cell counts should be ≥ 5.
  - $r \times c$ table: (i) No more than 20% of the cells have expected cell counts < 5 and (ii) no expected celll count should be < 1.

- In the 2×2 case, the formula for the X^2 statistic simplifies considerably from how we have presented it above. The simplified formula is

$$X^2 = \frac{n(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

- A continuity correction (also known as *Yates' correction*) is sometimes used with the X^2 test which improves the χ^2 approximation to its distribution.

  The continuity corrected version of the test statistic is given by

$$X^2 = \begin{cases} \frac{n(|ad-bc|-n/2)^2}{(a+c)(b+d)(a+b)(c+d)} & \text{for the } 2 \times 2 \text{ case} \\ \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(|O_{ij}-E_{ij}|-0.5)^2}{E_{ij}} & \text{for the general } r \times c \text{ case.} \end{cases}$$

Example — OC use and MI

		MI		
		Yes	No	
OC Use	Yes	13	4987	5000
	No	7	9993	10000
		20	14980	15000

To test whether there is an association between OC use and incidence of MI, we compute the test statistic:

$$X^2 = \frac{n(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)} = \frac{15000[13(9993)-7(4987)]^2}{(20)(14980)(5000)(10000)} = 9.04$$

The $\alpha = .05$th critical value for this test is

$$\chi^2_{1-.05}(1) = 3.84$$

so since $X^2 = 9.04 > 3.84$, we reject the hypothesis of independence and conclude that there is an association between OC use and MI incidence.

- The p value for this test is

$$P(\chi^2(1) > 9.04) = .0026 \quad \text{(from Minitab)}$$

- In this example, the sample size is very large but the expected cell counts in column 1 are fairly small, so we might expect the continuity corrected version of the X^2 statistic to give appreciably different results.

The continuity corrected version is

$$X^2 = \frac{n(|ad - bc| - n/2)^2}{(a+c)(b+d)(a+b)(c+d)}$$
$$= \frac{15000(|13(9993) - 7(4987)| - 15000/2)^2}{(20)(14980)(5000)(10000)} = 7.67$$

which has p value

$$P(\chi^2(1) > 7.67) = .0056$$

- The use of the continuity correction in a chi-square test of independence is somewhat controversial and there's not consensus on whether or not it should be used.

- In this course, I will not expect you to use the continuity correction.

- Minitab implements the chi-square test without the continuity correction.

Example — Seat Belt Use

- In a study of seatbelt use, Robertson (1974) reports the frequency of seatbelt use by ownership of the vehicle for a sample of drivers observed at 138 sites around the country in 1973–1974.

- Seatbelt use was categorized by type: lap and shoulder belt used (L/S), lap belt only used (Lap), and no belt used (None). Ownership of the car was determined from motor vehicle records.

- The data are as follows:

		Belt Use			
		L/S	Lap	None	
Ownership	Individual	583	139	524	1246
	Lease	86	24	74	184
	Company/Other	182	31	145	358
	Rental	145	24	59	228
		996	218	802	2016

Is there an association between seat belt use and car ownership?

The Pearson chi-square test can be used to address this question. Here, we use Minitab to generate the following results:

```
Tabulated statistics: ownership, belt Use
Using frequencies in freq
Rows: ownership   Columns: belt Use
                   L/S     Lap     None     All
individual         583     139      524    1246
                 615.6   134.7    495.7  1246.0
                1.7247  0.1349   1.6180       *
lease               86      24       74     184
                  90.9    19.9     73.2   184.0
                0.2646  0.8462   0.0088       *
other              182      31      145     358
                 176.9    38.7    142.4   358.0
                0.1488  1.5365   0.0468       *
rental             145      24       59     228
                 112.6    24.7     90.7   228.0
                9.2947  0.0174  11.0806       *
All                996     218      802    2016
                 996.0   218.0    802.0  2016.0
                     *       *        *       *
Cell Contents:      Count
                    Expected count
                    Contribution to Chi-square
Pearson Chi-Square = 26.722, DF = 6, P-Value = 0.000
Likelihood Ratio Chi-Square = 27.584, DF = 6, P-Value = 0.000
```

- The Minitab output contains the observed and expected counts for each cell in the table, and each cell's contribution to the X^2 statistic.

- E.g., in the $(1,1)^{\text{th}}$ cell, the observed count is $O_{11} = 583$, the expected cell count is the product of the margins divided by n:

$$E_{11} = \frac{(1246)(996)}{2016} = 615.6$$

and this cell's contribution to X^2 is

$$\frac{(O_{11} - E_{11})^2}{E_{11}} = \frac{(583 - 615.6)^2}{615.6} = 1.7247$$

- Summing up each cell's contribution, we get our test statistic:

$$X^2 = \sum_{i=1}^{4}\sum_{j=1}^{3} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 1.7247 + 0.1349 + \cdots + 11.0806 = 26.722$$

- For this $r \times c = 4 \times 3$ table, under the null hypothesis

$$X^2 \dot{\sim} \chi^2((r-1)(c-1)) = \chi^2((4-1)(3-1)) = \chi^2(6)$$

so the $\alpha = .05$ level critical value is

$$\chi^2_{1-.05}(6) = 12.59 \quad \text{(Table A.8 in our text)}$$

- The p-value is

$$p = P(\chi^2(6) > 26.722) = .00016 \quad \text{(Minitab)}$$

so we reject the null hypothesis and conclude that there is an association between seatbelt use and car ownership.

- A comparison of the observed and expected cell counts can tell us more about the nature of this association. Notice that the biggest contributions to X^2 come from the $(4,1)^{\text{th}}$ and $(4,3)^{\text{th}}$ cells. Apparently, many more people than expected wear lap and shoulder belts in rental cars and many fewer than expected wear no belts in rental cars.

McNemar's Test

When we studied two-sample tests of $H_0 : \mu_1 = \mu_2$ (equal population means), we distinguished the paired samples case (paired t-test) from the independent samples case (two-sample t test).

However, when we switched over to the study of inference for proportions, we only considered the independent samples case for testing $H_0 : p_1 = p_2$ (equal population proportions).

Of course, samples can be paired regardless of whether the outcome is a 0-1 variable (leading to a proportion) or a continuous variable (leading to a mean, μ), so we now return to the question of how to test $H_0 : p_1 = p_2$ based upon paired samples.

- One way in which paired data arise is from repeated observations taken on the same subject at two different time points or via two different measuring devices/methods or observers.

Example — DKA and Insulin Pump Therapy

- Improvement in control of blood-glucose levels is an important motivation for the use of insulin pumps for diabetic patients. However, certain side-effects have been reported with pump therapy. One such side effect is the occurrence of diabetic ketoacidosis (DKA).

- 161 diabetic patients who went on insulin pump therapy were assessed for the presence of DKA both before beginning therapy and after onset of therapy (each patient's DKA status was measured twice). The following contingency table contains the results:

		After onset of pump therapy: DKA	After onset of pump therapy: No DKA	
Before Therapy	DKA	7	7	14
	No DKA	19	128	147
		26	135	161

- These data are paired in the sense that the presence/absence of DKA is measured twice on the same subjects.

- Another common way for paired data pertaining to proportions can arise is through the use of **matched pairs.**
  - Matched pairs are often used in case-control studies in which cases are identified, and then each case is matched to a control subject who is similar to that case in terms of age, sex, race, and other relevant characteristics.

Example — Thromboembolism and Oral Contracptives

- A retrospective matched pair case-control study of thromboembolism and oral contraceptive (OC) use was reported by Sartwell et al. (1969).
- The cases were 175 women of reproductive age (15–44), discharged alive from 43 hospitals in five cities after initial attacks of idiopathic (i.e., of unknown cause) thrombophlebitis,* pulmonary embolism,† or cerebral thrombosis or embolism.
- The controls were matched to the cases on the basis of hospital, residence, time of hospitalization, race, age, marital status, parity‡, and hospital pay status (ward, semi-private, private).
- The data from this study are as follows:

		Control use of OCs: Yes	Control use of OCs: No	
Case use of OCs	Yes	10	57	67
	No	13	95	108
		23	152	175

- Again, these are paired data because each case is matched to a control. There are 350 subjects in this study, but only 175 independent pairs here.

---

* blood clots in the veins with inflammation in the vessel walls

† a clot carried through the blood and obstructing lung blood flow

‡ number of prior pregnancies

McNemar's Test:

In general, paired data appropriate for McNemar's test of equal proportions can be displayed as follows:

Response for 1st member of pair	Response for 2nd member of pair: 1	0
1	O_{11}	O_{12}
0	O_{21}	O_{22}

McNemar's test is based upon examining the numbers of discordant pairs. First, let's define what we mean by that:

- A **concordant pair** is a pair of observations of a dichotomous (0-1) random variable, where the outcomes are the same for the members of the pair.
  - There are $O_{11} + O_{22}$ concordant pairs in the table above.
  - It turns out that concordant pairs provide no information concerning the difference between the proportions of 1's in the two samples (e.g., between cases and controls, or between time 1 (pre-therapy) and time 2 (post-therapy)).
- A **discordant pair** is a pair of observations of a dichotomous random variable, where the outcomes are different for the members of the pair.

  There are two types of discordant pairs:
  - In the table above, O_{12} is the number of pairs where the responses are $(1, 0)$ (first member of pair has the response 1, second member has the response 0). This is denoted by r in our text.
  - In the table above, O_{21} is the number of pairs where the responses are $(0, 1)$. This is denoted by s in our text.

Logic of McNemar's test: If the proportion of 1's doesn't differ from one member of the pair to the other (equal proportions, or no association) then we would expect O_{12} and O_{21} to be about the same.

Therefore, base test statistic on $|O_{12} - O_{21}|$.

- E.g., if there is a positive association between being a case and taking OCs, we would expect there would be more (case, control) pairs of the form $(1, 0)$ than pairs of the form $(0, 1)$ (1 denotes use of OCs).
  - I.e., we'd expect $O_{12} > O_{21}$.
- Similarly, if there is a positive association between insulin pump therapy and occurrence of DKA, we'd expect that we would see more (before, after) pairs of the form $(0, 1)$ than of the form $(1, 0)$.
  - I.e., we'd expect people to switch from no DKA to DKA after onset of pump therapy more often than we'd expect people to switch from DKA to no DKA after onset of pump therapy. That is, we'd expect $O_{21} > O_{12}$.
  - Under no association between DKA and insuling pump therapy, we'd expect people to switch from no DKA to DKA about as often as we'd expect people to switch from DKA to no DKA ($O_{12} \approx O_{22}$).

McNemar's test statistic is computed as

$$X^2 = \frac{(|O_{12} - O_{21}| - 1)^2}{O_{12} + O_{21}} \dot{\sim} \chi^2(1) \quad \text{under } H_0$$

- We denote it by X^2 because its approximate distribution under the null hypothesis of equal population proportions is $\chi^2(1)$
- The alternative hypothesis for this test is two-sided: non-equal proportions of 1's for the two populations forming the pairs. I.e., association between the variable defining the pairs and the response variable.
- The -1 in the numerator of X^2 is a continuity correction, which is not used in some computer packages.

- The critical value for an α level test is $\chi^2_{1-\alpha}$. I.e, we reject the hypothesis of equal population proportions/no association at level α if
$$X^2 > \chi^2_{1-\alpha}(1).$$

- The p-value is
$$p = P(\chi^2(1) > X^2)$$

- Like Pearson's chi-square test, McNemar's test is based on a normal approximation, so it should only be used for large sample sizes in the following sense:
  - Rule of thumb: use the above version of McNemar's test only if $O_{12} + O_{21} \geq 20$.

- There is an exact version of McNemar's test which is implemented in some computer packages. The exact version can be thought of as the paired-sample version of Fisher's exact test, and can be used for any sample size, no matter how small.

Example — DKA and Insulin Pump Therapy

- In this case, $O_{12} = 7, O_{21} = 19$ and $O_{12} + O_{21} = 7 + 19 = 26 \geq 20$, so we have sufficient sample size to proceed with McNemar's test.

- The test statistic is
$$X^2 = \frac{(|O_{12} - O_{21}| - 1)^2}{O_{12} + O_{21}} = \frac{(|7 - 19| - 1)^2}{26} = 4.654$$

- The critical value for an $\alpha = .05$-level test is $\chi^2_{1-.05}(1) = 3.84$, so since $X^2 = 4.654 > 3.84$ we reject the hypothesis of no association between pump therapy and DKA.

- The p-value for this test is
$$p = P(\chi^2(1) > 4.654) = .0310 \quad \text{(from Minitab)}$$

Example — Thromboembolism and Oral Contracptives

- In this case, $O_{12} = 57, O_{21} = 13$ and $O_{12} + O_{21} = 57 + 13 = 70 \geq 20$, so we have sufficient sample size to proceed with McNemar's test.
- The test statistic is

$$X^2 = \frac{(|O_{12} - O_{21}| - 1)^2}{O_{12} + O_{21}} = \frac{(|57 - 13| - 1)^2}{70} = 26.41$$

- The critical value for an $\alpha = .05$-level test is $\chi^2_{1-.05}(1) = 3.84$, so since $X^2 = 26.41 > 3.84$ we reject the hypothesis of no association between pump therapy and DKA.
- The p-value for this test is

$$p = P(\chi^2(1) > 26.41) = .0000 \quad \text{(from Minitab)}$$

Warning: One must be careful to set up the contingency table for paired proportion data correctly for McNemar's test.

For example, in the Thromboembolism and Oral Contracptives example, it is tempting to set up the contingency table as

Presence of thromboembolism		Use of OCs: Yes	Use of OCs: No	
	Case	67	108	175
	Control	23	152	175
		90	260	350

- This is the table organization that we used for independent samples.
- The samples are not independent here, however, and an analysis of the table as given above via a chi-square test or Fisher exact test will be incorrect!
- In the matched pair context, it is important that the overall margin for the table be the number of independent pairs (175 in the example), not the total number of subjects (350 in the example).

The Odds Ratio

The Pearson chi-square test and McNemar's test allow us to test for association between two categorical variables in the independent- and paired-samples contexts, respectively.

However, in addition to (or instead of) testing whether there's no association, we might want to measure the strength and direction of the association that exists.

We have already seen that the risk difference, relative risk and odds ratio are all useful measures of association in a contingency table.

However, the odds ratio is the most broadly appropriate measure of association (unlike the relative risk, it applies in both cohort and case-control designs) and it has the best statistical properties, so we concentrate on it.

Recall that the odds of an event A is the probability of A divided by the probability of A^c (not A).

Suppose the event A we are interested in is that a subject has a disease. Then the odds of disease are

$$\text{odds(disease)} = \frac{P(\text{diseased})}{P(\text{not diseased})}$$

The *odds ratio* is the ratio of the odds of disease (or whatever event we're interest in) in one population divided by the odds of disease in a comparison population.

That is, if the two populations we are interested in are an exposed group and an unexposed group, then the odds ratio (OR) for the exposed versus unexposed populations is

$$\begin{aligned} OR &= \frac{\text{odds(disease}|\text{exposed)}}{\text{odds(disease}|\text{unexposed)}} \\ &= \frac{P(\text{diseased}|\text{exposed})/P(\text{not diseased}|\text{exposed})}{P(\text{diseased}|\text{unexposed})/P(\text{not diseased}|\text{unexposed})} \end{aligned}$$

- The property that makes the OR useful for both cohort and case control studies is that the formula above is mathematically equivalent to

$$OR = \frac{\text{odds(exposure}|\text{diseased)}}{\text{odds(exposure}|\text{not diseased)}}$$
$$= \frac{P(\text{exposed}|\text{diseased})/P(\text{not exposed}|\text{diseased})}{P(\text{exposed}|\text{not diseased})/P(\text{not exposed}|\text{not diseased})}$$

- In cohort studies, we have information from which we can estimate $P(\text{disease}|\text{exposure status})$ but we don't have information from which to estimate $P(\text{exposure}|\text{disease status})$.

- In case-control studies, its just the opposite: we can estimate $P(\text{exposure}|\text{disease status})$ but not $P(\text{disease}|\text{exposure status})$.

- However, since the OR can be estimated from either $P(\text{exposure}|\text{disease status})$ or $P(\text{disease}|\text{exposure status})$, without altering its interpretation as the relative odds of disease in two groups, it can be used in either design.

Both study designs generate data that can be displayed as follows:

		Exposure Status		
		Exposed	Not Exposed	
Disease Status	Diseased	a	b	$a+b$
	Not Diseased	c	d	$c+d$
		$a+c$	$b+d$	n

For either design, the OR can be estimated as the cross-product ratio:

$$\hat{OR} = \frac{ad}{bc}$$

Example — Breast Cancer and Age at First Birth

- A hypothesis has been proposed that breast cancer in women is caused in part by events that occur between the age at menarche (the age at which menstruation begins) and the age of first childbirth.

- In particular, the hypothesis is that the risk of breast cancer increases as the length of this time interval increases. If this theory is correct, then an important risk factor for breast cancer is age at first birth.

- An international study was conducted to test this hypothesis. Breast cancer cases were identified in selected hospitals in several countries. Controls without breast cancer were chosen from women of comparable age from the same hospitals, but they were not individually matched to cases (data not paired).

- The subset of women with at least one birth were (somewhat arbitrarily) divided into two categories: those whose first birth occurred before age 30, and those whose first birth came at age 30 or later. The resulting data are as follows:

		Age at 1st Birth		
		≥ 30	< 30	
Breast Cancer Status	Case	683	2537	3220
	Control	1498	8747	10245
		2181	11284	13465

- Here, the "exposure" is being ≥ 30 years of age when giving birth for the first time.

The estimated odds ratio for the odds of disease in the exposed versus unexposed group is

$$\hat{OR} = \frac{ad}{bc} = \frac{(683)(8747)}{(2537)(1498)} = 1.57$$

- Interpretation: the *odds** of breast cancer is 1.57 times higher (57% higher) for women with first birth at age ≥ 30 years, then for women with first birth earlier than that.

OK, so the estimate of the odds ratio is 1.57, or whatever it happens to be. But this point estimate of the population OR is almost certainly at least slightly wrong.

Can we quantify the uncertainty in our odds ratio estimate?

That is, can we form a confidence interval for OR based on $\hat{OR}$*?*

The answer is yes. But the generally preferred method of doing it is slightly different than we've used before:

- Instead of forming a CI on OR directly, we will form a CI on $\ln(OR)$† (a transformation of OR) and then "back-transform" the endpoints of our interval, by applying the exp() function‡, to obtain a CI on OR.

- The reason for this indirect approach is that, as usual, we're going to use a normal-based interval (that is, an interval of the form estimate $\pm\, z_{1-\alpha/2}$standard errors), but this approach works best if our estimate is approximately normally distributed.

- It turns out that, generally speaking, our estimate $\hat{OR} = (ad)/(bc)$ does not follow a normal distribution as closely as $\ln(\hat{OR}) = \ln\{(ad)/(bc)\}$ does.

Therefore, we first form an interval for $\ln(OR)$ as

$$\ln(\hat{OR}) \pm z_{1-\alpha/2}\text{s.e.}\{\ln(\hat{OR})\} = (L, U)$$

where

$$\text{s.e.}\{\ln(\hat{OR})\} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

---

* remember: odds, not probability. Just think of odds as an alternative way to quantify chance.

† The natural logarithm of OR

‡ which is the inverse function for ln()

Then, the corresponding interval for OR is given by $(\exp(L), \exp(U)) = (e^L, e^U)$.

General Result: An approximate $100(1-\alpha)\%$ CI for the population odds ratio OR is given by

$$(e^L, e^U) \quad \text{where} \quad \begin{aligned} L &= \ln(\hat{OR}) - z_{1-\alpha/2}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \\ U &= \ln(\hat{OR}) + z_{1-\alpha/2}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \end{aligned}$$

Example — Breast Cancer and Age at First Birth

- Recall we found that $\hat{OR} = 1.57$, so $\ln(\hat{OR}) = \ln(1.57) = 0.4523$.
- The standard error of $\ln(\hat{OR})$ is

$$\text{s.e.}\{\ln(\hat{OR})\} = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{683}+\frac{1}{2537}+\frac{1}{1498}+\frac{1}{8747}} = .05138$$

- Therefore, an approximate 95% CI for $\ln(OR)$ is

$$\ln(\hat{OR}) \pm z_{1-.05/2}\text{s.e.}\{\ln(\hat{OR})\} = .4523 \pm 1.96(.05138) = (.3516, .5530)$$

- The corresponding approximate 95% CI for OR is then

$$(e^L, e^U) = (e^{.3516}, e^{.5530}) = (1.42, 1.74)$$

- We can be (approximately) 95% confident that the population OR for breast cancer for the exposed versus unexposed group lies between 1.42 and 1.74.
  - I.e., we can be 95% confident that the exposed group has an odds of breast cancer that is between 42% and 74% higher than that of the unexposed group.

- The previous method of estimating a placing a CI around OR was for non-paired data.

For paired data, that are summarized in a contingency table of the form

Response for 1st member of pair	Response for 2nd member of pair: 1	Response for 2nd member of pair: 0
1	O_{11}	O_{12}
0	O_{21}	O_{22}

the OR is estimated as

$$\hat{OR} = \frac{O_{12}}{O_{21}}$$

A confidence interval for the OR is again formed by taking exp() of the endpoints of an interval for $\ln(OR)$.

Specifically, an approximate $100(1-\alpha)\%$ CI for $\ln(OR)$ based on paired data is

$$(L, U) = \ln(\hat{OR}) \pm z_{1-\alpha/2}\text{s.e.}\{\ln(\hat{OR})\} \quad \text{where} \quad \text{s.e.}\{\ln(\hat{OR})\} = \sqrt{\frac{O_{12}+O_{21}}{O_{12}O_{21}}}$$

The corresponding approximate $100(1-\alpha)\%$ CI for OR from paired data is

$$(e^L, e^U) \quad \text{where} \quad \begin{aligned} L &= \ln(\hat{OR}) - z_{1-\alpha/2}\sqrt{\frac{O_{12}+O_{21}}{O_{12}O_{21}}} \\ U &= \ln(\hat{OR}) + z_{1-\alpha/2}\sqrt{\frac{O_{12}+O_{21}}{O_{12}O_{21}}} \end{aligned}$$

Example — DKA and Insulin Pump Therapy

Recall these data:

Before Therapy		After onset of pump therapy: DKA	After onset of pump therapy: No DKA	
	DKA	7	7	14
	No DKA	19	128	147
		26	135	161

- The estimated OR here is

$$\hat{OR} = \frac{O_{12}}{O_{21}} = \frac{7}{19} = .368$$

- Interpretation: the odds of DKA before insulin pump therapy was estimated to be .368 times that of the odds of DKA after insulin pump therapy (odds of DKA before were 36.8% as large as odds of DKA after).

- $\ln(\hat{OR}) = \ln(.368) = -.9985$ and

$$\text{s.e.}\{\ln(\hat{OR})\} = \sqrt{\frac{O_{12} + O_{21}}{O_{12}O_{21}}} = \sqrt{\frac{7+19}{(7)(19)}} = .4421$$

so an approximate 95% CI for $\ln(OR)$ is

$$\ln(\hat{OR}) \pm z_{1-.05/2}\text{s.e.}\{\ln(\hat{OR})\} = -.9985 \pm 1.96(.4421) = (-1.865, -.1319)$$

- Therefore, an approximate 95% CI for OR is

$$(e^L, e^U) = (e^{-1.865}, e^{-.1319}) = (.155, .876)$$

Some Other Important Methods of Biostatistics

Correlation (Ch.17 in our book):

- The **Pearson correlation coefficient** is a number between -1 and +1 that measures the strength of the linear association between two continuous variables X and Y.
  - A correlation near +1 or -1 indicates that the points on a scatterplot of X versus Y would cluster closely around a straight line.
  - Positive values indicate that this line slopes up (Y tends to go up when X goes up), and negative values indicates that this line slopes down (Y tends to go down as X goes up).
  - A correlation near 0 indicates no linear association between X and Y — no systematic trend for Y to increase (decrease) by a constant amount for every unit increase in Y.
  - This often means no association at all, but can also mean there is an association, but it is nonlinear.
- The sample Pearson correlation coefficient is usually denoted r. This statistic is computed from sample data that contain measurements of both X and Y on each subject. It estimates the corresponding population correlation ρ.
- The sample correlation coefficient is often computed to quantify the degree of (linear) association between pairs of variables.
- It is also common to test the hypothesis that $\rho = 0$ (no linear association), by computing a test statistic that examines how far r is from 0.

- For data with outliers, **Spearman's correlation coefficient** is often preferred over Pearson correlation coefficient.
  - Spearman's correlation coefficient is just Pearson's coefficient computed on the ranks of the data rather than the data themselves.
  - Spearman's coefficient is sometimes called the rank correlation coefficient.
- Correlation makes no explanatory variable/response variable (independent/dependent varaible) distinction between X and Y.

Simple Linear Regression (Ch.18 in our book):

- Similar to correlation in that we examine the linear relationship between two variables X and Y.
- However, regression distinguishes between the response variable (Y) and the explanatory variable (X).
  - Goal is to *predict* Y from X, or *explain* how Y changes with X, using a straight-line relationship.
- Simple linear regression involves a model for how Y changes with X, which is based on the equation of a line: $Y = \text{intercept} + (\text{slope})X$. Since Y can't be expected to lie exactly on a line with X, the model adds error:

$$Y = \text{intercept} + (\text{slope})X + \text{error}$$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Interpretation: for every unit increase in X (every time we add 1 to X) we can expect the mean of Y (the mean response) to increase by β_1 units.
- The intercept β_0 and slope β_1 of this relationship have to be estimated from sample data on Y and X.
  - The preferred method of doing this estimation is called the meathod of least squares, and produces the straight line that goes through the points in a scatterplot of Y vs. X which is closest (in some sense) to all of the points.

- In simple linear regression, we are often interested in testing H_0 : $\beta_1 = 0$ (whether the slope is 0). This test gets at whether or not Y depends (linearly) on X. That is, is there a linear trend in Y as we increase X.

- Also often interested in the point estimate of β_1. This tells how much Y increases (on average) for a given increase in X.

Multiple Linear Regression (Ch.19 in our book):

- Multiple regression extends simple regression to the case when we have multiple explanatory variables $X_1, X_2, \ldots, X_p$ from which we'd like to predict/explain Y.

- Model now looks like

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

- Each explanatory variable has its own slope describing the linear effect of that variable on Y above and beyond the effects of the other explanatory variables in the model.

- Multiple regression is useful for improving prediction of Y, explaining how Y depends upon many other variables, controlling for the effects of $X_2, X_3, \ldots, X_p$ when assessing relationship between Y and X_1, etc.

Analysis of Variance (ANOVA) (Ch.12 in our book):

- We learned that the two-sample t test was the appropriate method for testing equality of two means:

  $$H_0 : \mu_1 = \mu_2$$

  (in the usual case where the sample variance is unknown).

- The ANOVA is designed for the more general problem of testing equality of g group means. It generalizes the two sample t test from $g = 2$ to $g \geq 2$.

- That is, the ANOVA is designed to test

  $$H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$$

  where $\mu_1, \ldots, \mu_g$ are the population means corresponding to g populations (e.g., g different treatments).

- The ANOVA leads to an F test which generalizes the two-sample t test.

  - Like the "case 2" two-sample t test, the ANOVA F test assumes unknown but equal variances in the g groups, although this assumption can be relaxed.

- The ANOVA is based upon a model which has a close relationship to the multiple linear regression model.

- If we combine a regression model with an ANOVA model, we get what's known as the analysis of covariance or ANCOVA. The ANCOVA allows the comparison of g population means while controlling for one or more additional explanatory variables (also known as covariates).

Methods for Multiple 2×2 Tables (Ch.16 in our book):

- We studies methods for quantifying and testing association in a 2×2 table; e.g., for a table that describes the relationship between exposure and disease.

- Often there are other variables that influence the probability of disease, the probability of exposure or both, which should be controlled when examining the association between exposure and disease.

- One way that such confounding variables can be controlled, is to look at the two-way relationship between exposure and disease, separately at each level of the confounding variable. This leads to the construction of multiple 2×2 tables.

  - E.g., if we want to look at the association between exposure and disease while controlling for the effect of age, we can form a separate 2×2 table of exposure vs disease for individuals in each of several age categories.

- Methods of analyzing multiple 2×2 tables were developed by researchers Cochran, Mantel, and Haenszel, and these methods go by the name Mantel-Haenszel or (sometimes) Cochran-Mantel-Haenszel.

- To measure the association between exposure and disease while controlling for a *stratification* variable (like age), the **Mantel-Haenszel odds ratio estimator** takes a weighted average of the odds ratio in each of the 2×2 tables formed at each level of the variable (at each stratum).

- A test of association between exposure and disease in stratified 2×2 tables is provided by the **Mantel-Haenszel test,** which can be thought of as an extension of the Pearson chi-square test form 1 2×2 table to multiple 2×2 tables.

Nonparametric Methods (Ch.13 in our book):

- Many of the most commonly used statistical methods (e.g., t and z tests, the ANOVA, etc.) are based upon an assumption that the data follow (at least approximately) a normal distribution or some other parametric probability distribution.

- Nonparametric methods are statistical techniques that do not depend upon such parametric distributional assumptions and are, therefore, more broadly appropriate.

  - This "robustness" comes at some cost, however. If the assumptions of a parametric analysis are satisfied and we do a nonparametric analysis instead, we typically lose power/efficiency.

- Many of the basic parametric statistical methods we have studied have nonparametric alternatives.

  - E.g., the Sign Test and the Wilcoxon Signed-Rank Test are both nonparametric alternatives to the paired t test.

  - The Wilcoxon Rank Sum Test is the nonparametric version of the two-sample t test.

  - The Kruskal-Wallis test is the nonparametric version of the ANOVA F test.

- Many nonparametric methods are based upon analyzing the signs and/or ranks of the data rather than the original data themselves.

- Nonparametric counterparts to simple methods exist and are attractive, but nonparametric solutions to difficult/complex statistical problems are often difficult to find and necessitate the use of parametric procedures.

Logistic Regression (Ch.20 in our book):

- Multiple linear regression is useful for predicting/explaining a *continuous* response variable Y in terms of explanatory variables $X_1, \ldots, X_p$.

- Logistic regression does the same thing for a binary response variable (e.g., subject lives/dies) or a response variable that can be expressed as a proportion.

Survival Analysis (Ch.21 in our book):

- For response variables which are survival times, or times until failure, classical normal-theory methods are not well suited.

- There are two reasons for this:

  - Survival times are rarely normally distributed;

  - Survival times are often *censored*.

- If the response is the time until some event occurs (e.g., death), the response is said to be censored if the study ends before the event occurs. In that case, we don't know the exact survival time, only that it was greater than the time at which the subject was last observed.

- Survival analysis includes methods for analyzing time to event/survival time data that are based upon non-normal (in some cases nonparametric) distributional assumptions, and which can account for censoring in an appropriate way.