8 Categorical Data Analysis

Single Proportion Problems

SW Section 6.6

Assume that you are interested in estimating the proportion p of individuals in a population with a certain characteristic or attribute based on a random or representative sample of size n from the population. The **sample proportion** $\hat{p}$ =(# with attribute in the sample)/n is the best guess for p based on the data.

This is the simplest **categorical data** problem. Each response falls into one of two exclusive and exhaustive categories, called success and failure. Individuals with the attribute of interest are in the success category. The rest fall into the failure category. Knowledge of the **population proportion** p of successes characterizes the distribution across both categories because the population proportion of failures is $1 - p$.

As an aside, note that the probability that a randomly selected individual has the attribute of interest is the population proportion p with the attribute, so the terms population proportion and probability can be used interchangeably with random sampling.

A CI for p

A two-sided CI for p is a range of plausible values for the unknown population proportion p, based on the observed data. To compute a two-sided CI for p:

1. Specify the confidence level as the percent $100(1-\alpha)\%$ and solve for the error rate α of the CI.

2. Compute $z_{crit} = z_{.5\alpha}$ (i.e. area under the standard normal curve to the left and to the right of z_{crit} are $1 - .5\alpha$ and $.5\alpha$, respectively). See the table in SW, page 643-4.

3. The $100(1-\alpha)\%$ CI for p has endpoints $L = \hat{p} - z_{crit}SE$ and $U = \hat{p} + z_{crit}SE$, respectively, where the "CI standard error" is
$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$
The CI is often written as $\hat{p} \pm z_{crit}SE$.

The length of the CI
$$U - L = 2z_{crit}SE$$
depends on the accuracy of the estimate $\hat{p}$, as measured by the standard error SE. For a given $\hat{p}$, this standard error decreases as the sample size n increases, yielding a narrower CI. For a fixed sample size, this standard error is maximized at $\hat{p} = .5$, and decreases as $\hat{p}$ moves towards either 0 or 1. In essence, sample proportions near 0 or 1 give narrower CIs for p. However, the normal approximation used in the CI construction is less reliable for extreme values of $\hat{p}$.

Example The 1983 Tylenol poisoning episode highlighted the desirability of using tamper-resistant packaging. The article "Tamper Resistant Packaging: Is it Really?" (Packaging Engineering, June 1983) reported the results of a survey on consumer attitudes towards tamper-resistant packaging.

A sample of 270 consumers was asked the question: "Would you be willing to pay extra for tamper resistant packaging?" The number of yes respondents was 189. Construct a 95% CI for the proportion p of all consumers who were willing in 1983 to pay extra for such packaging.

Here $n = 270$ and $\hat{p} = 189/270 = .700$. The critical value for a 95% CI for p is $z_{.025} = 1.96$. The CI standard error is given by

$$SE = \sqrt{\frac{.7 * .3}{270}} = .028,$$

so $z_{crit}SE = 1.96 * .028 = .055$. The 95% CI for p is $.700 \pm .055$. You are 95% confident that the proportion of consumers willing to pay extra for better packaging is between .645 and .755. (How much extra?).

Appropriateness of the CI

The standard CI is based on a **large sample** standard normal approximation to

$$z = \frac{\hat{p} - p}{SE}.$$

A simple rule of thumb requires $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$ for the method to be suitable. Given that $n\hat{p}$ and $n(1-\hat{p})$ are the observed numbers of successes and failures, you should have at least 5 of each to apply the large sample CI.

In the packaging example, $n\hat{p} = 270*(.700) = 189$ (the number who support the new packaging) and $n(1-\hat{p}) = 270*(.300) = 81$ (the number who oppose) both exceed 5. The normal approximation is appropriate here.

Hypothesis Tests on Proportions

The following example is typical of questions that can be answered using a hypothesis test for a population proportion.

Example Environmental problems associated with leaded gasolines are well-known. Many motorists have tampered with the emission control devices on their cars to save money by purchasing leaded rather than unleaded gasoline. A *Los Angeles Times* article on March 17, 1984 reported that 15% of all California motorists have engaged in emissions tampering. A random sample of 200 cars from L.A. county was obtained, and the emissions devices on 21 are found to be tampered with. Does this suggest that the proportion of cars in L.A. county with tampered devices differs from the statewide proportion?

Two-Sided Hypothesis Test for p

Suppose you are interested in whether the population proportion p is equal to a prespecified value, say p_0. This question can be formulated as a two-sided test. To carry out the test:

1. Define the null hypothesis $H_0 : p = p_0$ and alternative hypothesis $H_A : p \neq p_0$.

2. Choose the size or significance level of the test, denoted by α.

3. Using the standard normal probability table, find the critical value z_{crit} such that the areas under the normal curve to the left and right of z_{crit} are $1 - .5\alpha$ and $.5\alpha$, respectively. That is, $z_{crit} = z_{.5\alpha}$.

4. Compute the test statistic (often to be labelled z_{obs})

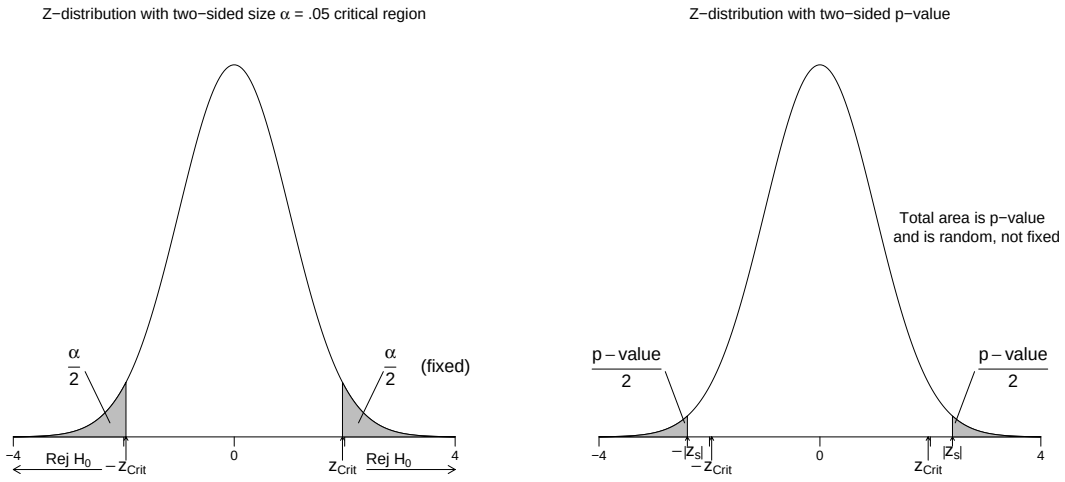
$$z_s = \frac{\hat{p} - p_0}{SE},$$

where the "test standard error" is

$$SE = \sqrt{\frac{p_0(1-p_0)}{n}}.$$

5. Reject H_0 in favor of H_A if $|z_{obs}| \geq z_{crit}$. Otherwise, do not reject H_0.

The rejection rule is easily understood visually. The area under the normal curve outside $\pm z_{crit}$ is the size α of the test. One-half of α is the area in each tail. You reject H_0 in favor of H_A if the test statistic exceeds $\pm z_{crit}$. This occurs when $\hat{p}$ is significantly different from p_0, as measured by the standardized distance z_{obs} between $\hat{p}$ and p_0.



The P-Value for a Two-Sided Test

To compute the p-value (not to be confused with the value of p!) for a two-sided test:

1. Compute the test statistic $z_s = z_{obs}$.

2. Evaluate the area under the normal probability curve outside $\pm|z_s|$.

Recall that the null hypothesis for a size α test is rejected if and only if the p-value is less than or equal to α.

Example (Emissions data) Each car in the target population (L.A. county) either has been tampered with (a success) or has not been tampered with (a failure). Let $p =$ the proportion of cars in L.A. county with tampered emissions control devices. You want to test $H_0 : p = .15$ against $H_A : p \neq .15$ (here $p_0 = .15$). The critical value for a two-sided test of size $\alpha = .05$ is $z_{crit} = 1.96$.

The data are a sample of $n = 200$ cars. The sample proportion of cars that have been tampered with is $\hat{p} = 21/200 = .105$. The test statistic is

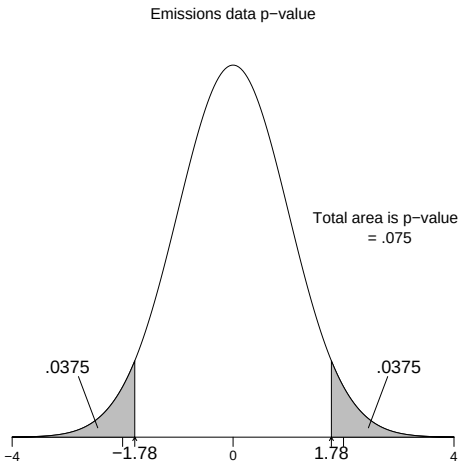
$$z_s = \frac{.105 - .15}{.02525} = -1.78,$$

where the test standard error satisfies

$$SE = \sqrt{\frac{.15 * .85}{200}} = .02525.$$

Given that $|z_s| = 1.78 < 1.96$, you have insufficient evidence to reject H_0 at the 5% level. That is, you have insufficient evidence to conclude that the proportion of cars in L.A. county that have been tampered with differs from the statewide proportion.

This decision is reinforced by the p-value calculation. The p-value is the area under the standard normal curve outside ± 1.78. This is $2 * .0375 = .075$, which exceeds the test size of .05.



REMARK: The SE used in the test and CI are different. This implies that a hypothesis test and CI could potentially lead to different decisions. That is, a 95% CI for a population proportion might cover p_0 when the p-value for testing $H_0 : p = p_0$ is smaller than 0.05. This will happen, typically, only in cases where the decision is "borderline."

Appropriateness of Test

The z-test is based on a large sample normal approximation, which works better for a given sample size when p_0 is closer to .5. The sample size needed for an accurate approximation increases dramatically the closer p_0 gets to 0 or to 1. A simple rule of thumb is that the test is appropriate when (the expected number of successes) $np_0 \geq 5$ and (the expected number of failures) $n(1-p_0) \geq 5$.

In the emissions example, $np_0 = 200 * (.15) = 30$ and $n(1-p_0) = 200 * (.85) = 170$ exceed 5, so the normal approximation is appropriate.

Minitab Implementation

1. CI and tests on a single proportion are obtained in Minitab by following the path: `Stat > Basic Statistics > 1 Proportion`.
2. The dialog box allows you to specify either raw data in columns (to be discussed later) or summarized data (number of trials, or sample size, and number of successes).
3. **Options**: the confidence level, the null proportion, and the direction of the alternative hypothesis. One-sided bounds are available. I do not how to generate a CI without a test, so I edit out the test output when it is not of interest.
4. You need to specify the normal approximation as an option. As default, Minitab computes an **exact** CI and test based on the **Binomial** distribution for the number of successes. The exact methods, which were described in conjunction with the sign test, are preferred, but not available in many packages. I will illustrate the exact methods later.

Minitab output for the emissions example is given below. The summary data, as provided in the example description, were directly entered in the `1 Proportion` dialog box. In the output, x is the number of successes. Note that the results of the 95% CI disagrees with the test done earlier. Exact methods will not contradict each other this way (neither do these asymptotic methods, usually).

```
Test and CI for One Proportion

Test of p = 0.15 vs p not = 0.15

Sample      X      N  Sample p          95.0% CI       Z-Value  P-Value
1          21    200  0.105000  (0.062515, 0.147485)    -1.78    0.075
```

One-Sided Tests and One-Sided Confidence Bounds

The mechanics of tests on proportions are similar to tests on means, except we use a different test statistic and a different probability table for critical values. This applies to one-sided and two-sided procedures. The example below illustrates a one-sided test and bound.

Example An article in the April 6, 1983 edition of *The Los Angeles Times* reported on a study of 53 learning impaired youngsters at the Massachusetts General Hospital. The right side of the

brain was found to be larger than the left side in 22 of the children. The proportion of the general population with brains having larger right sides is known to be .25. Does the data provide strong evidence for concluding, as the article claims, that the proportion of learning impaired youngsters with brains having larger right sides exceeds the proportion in the general population?

I will answer this question by computing a p-value for a one-sided test. Let p be the population proportion of learning disabled children with brains having larger right sides. I am interested in testing $H_0 : p = .25$ against $H_A : p > .25$ (here $p_0 = .25$).

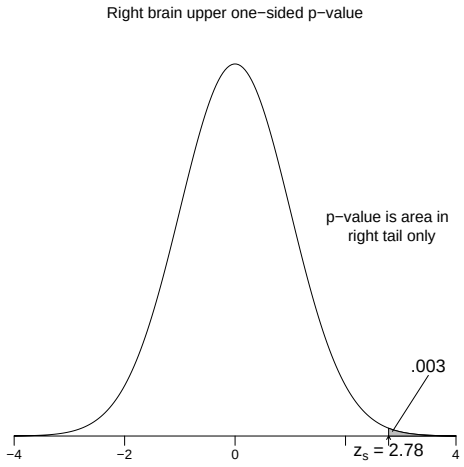
The proportion of children sampled with brains having larger right sides is $\hat{p} = 22/53 = .415$. The test statistic is

$$z_s = \frac{.415 - .25}{.0595} = 2.78,$$

where the test standard error satisfies

$$SE = \sqrt{\frac{.25 * .75}{53}} = .0595.$$

The p-value for an upper one-sided test is the area under the standard normal curve to the right of 2.78, which is approximately .003; see the picture below. I would reject H_0 in favor of H_A using any of the standard test levels, say .05 or .01. The newspaper's claim is reasonable.



A sensible next step in the analysis would be to compute a **lower confidence bound** $\hat{p} - z_{crit}SE$ for p. For illustration, consider a 95% bound. The CI standard error is

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{.415 * .585}{53}} = .0677.$$

The critical value for a one-sided 5% test is $z_{crit} = 1.645$, so a lower 95% bound on p is $.415 - 1.645 * .0677 = .304$. I am 95% confident that the population proportion of learning disabled children with brains having larger right sides is at least .304. Values of p smaller than .304 are not plausible.

You should verify that the sample size is sufficiently large to use the approximate methods in this example.

Minitab output for the right-side brain data is given below. An upper one-sided test and corresponding lower one-sided bound are given.

```
Test and CI for One Proportion

Test of p = 0.25 vs p > 0.25

                                95%
                              Lower
Sample   X   N  Sample p     Bound  Z-Value  P-Value
1       22  53  0.415094  0.303766     2.78    0.003
```

Small Sample Procedures

Large sample tests and CIs for p should be interpreted with caution in small sized samples because the true error rate usually exceeds the assumed (nominal) value. For example, an assumed 95% CI, with a nominal error rate of 5%, may be only an 80% CI, with a 20% error rate. The large sample CIs are usually overly optimistic (i.e. too narrow) when the sample size is too small to use the normal approximation.

SW use the following method developed by Alan Agresti for a 95% CI. The standard method computes the sample proportion as $\hat{p} = x/n$ where x is the number of successes in the sample and n is the sample size. Agresti suggested using the estimated proportion $\tilde{p} = (x+2)/(n+4)$ with the standard error

$$SE = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}},$$

in the "usual 95% interval" formula: $\tilde{p} \pm 1.96SE$. This appears odd, but amounts to adding two successes and two failures to the observed data, and then computing the standard CI.

This adjustment has little effect when n is large and $\hat{p}$ is not near either 0 or 1, as in the Tylenol example.

Example This example is based on a case heard before the U.S. Supreme Court. A racially segregated swimming club was ordered to admit minority members. However, it is unclear whether the club has been following the spirit of the mandate. Historically, 85% of the white applicants were approved. Since the mandate, only 1 of 6 minority applicants has been approved. Is there evidence of continued discrimination?

I examine this issue by constructing a CI and a test for the probability p (or population proportion) that a minority applicant is approved. Before examining the question of primary interest, let me show that the two approximate CIs are very different, due to the small sample size. One minority applicant ($x = 1$) was approved out of $n = 6$ candidates, giving $\hat{p} = 1/6$. A 95% large sample CI for p is (-.14, .46). Since a negative proportion is not possible, Minitab reports the CI as (.00,.46). Agresti's 95% CI (based on 3 successes and 7 failures) is (.02, .58). The big difference between the two intervals coupled with the negative lower endpoint on the standard CI suggests

that the normal approximation used with the standard method is inappropriate. This view is reinforced by the rule of thumb calculation for using the standard interval. Agresti's CI is wider, which is consistent with my comment that the standard CI is too narrow in small samples. As a comparison, the exact 95% CI is (.004,.64), which agrees more closely with Agresti's interval.

I should emphasize that the exact CI is best to use, but is not available in all statistical packages, so methods based on approximations may be required, and if so, then Agresti's method is clearly better than the standard normal approximation in small sized samples.

```
 Test and CI for One Proportion   <<<<<------- Standard Normal Approximation

Sample      X      N  Sample p          95.0% CI
1           1      6  0.166667  (0.000000, 0.464866)

* NOTE * The normal approximation may be inaccurate for small samples.

Test and CI for One Proportion    <<<<<------- Agresti's Method

Test of p = 0.5 vs p not = 0.5

Sample      X      N  Sample p          95.0% CI
1           3     10  0.300000  (0.015974, 0.584026)

* NOTE * The normal approximation may be inaccurate for small samples.

Test and CI for One Proportion   <<<<<-------- Exact CI

Sample      X      N  Sample p          95.0% CI
1           1      6  0.166667  (0.004211, 0.641235)
```

Returning to the problem, you might check for discrimination by testing $H_0 : p = .85$ against $H_A : p < .85$ using an **exact** test. The exact test p-value is .000 to three decimal places, and an exact upper bound for p is .582. What does this suggest to you?

```
 Test and CI for One Proportion

 Test of p = 0.85 vs p < 0.85

                              95%
                            Upper    Exact
 Sample  X  N  Sample p     Bound  P-Value
 1       1  6  0.166667  0.581803    0.000
```

Analyzing Raw Data

In most studies, your data will be stored in a spreadsheet with one observation per case or individual. For example, the data below give the individual responses to the applicants of the swim club.

```
 Data Display
```

```
outcome
   approved       not approved   not approved   not approved   not approved
   not approved
```

The data were entered as alphabetic strings, which are treated in Minitab as alphanumeric characters. As such, only certain tools can be used to summarize and analyze the data. A frequency distribution is a useful summary for categorical data, and is obtained in Minitab by following the steps: `Stat > Tables > Tally Individual Variables`. You can specify counts and percentages for each category.

```
Tally for Discrete Variables: outcome

     outcome  Count
    approved      1
not approved      5
           N=     6
```

You can compute a CI and test for a proportion using raw data (from `Stat > Basic Statistic > 1 Proportion`, data in column), provided the data column includes only two distinct values. The levels can be numerical or alphanumeric. Let us examine what happens if we do that here, though.

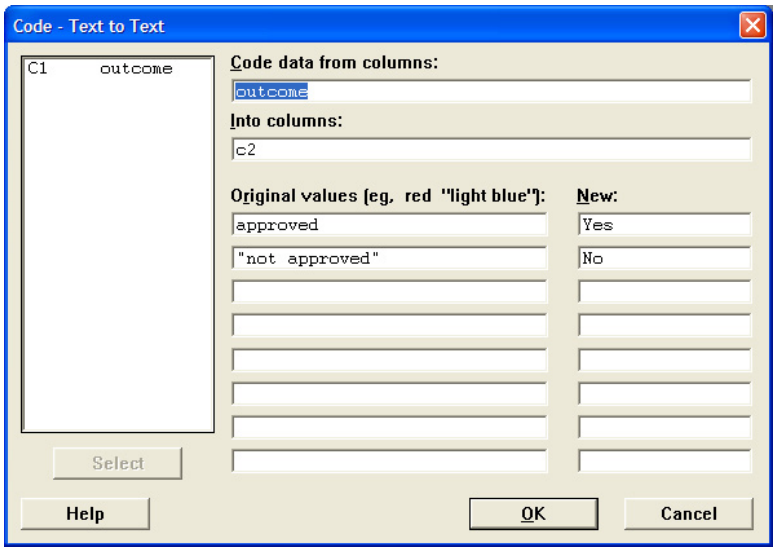
```
Test and CI for One Proportion: outcome

Test of p = 0.85 vs p < 0.85

Event = not approved         <<<<<<<<<---------- NOT what we wanted

                                95%
                               Upper    Exact
Variable  X  N  Sample p     Bound  P-Value
outcome   5  6  0.833333  0.991488    0.623
```

You must be very careful because Minitab defines the higher alphanumeric value to be success and the lower level to be failure. The above analysis actually is for $1-p$ the way we have structured the problem. Thus, in the swimming club problem if you define p to be the probability a minority applicant is approved and you wish to test $H_0 : p = .85$ against $H_A : p < .85$ then you need to relabel the variable values into another column (you can use the same column, but that is not a safe way to proceed in general).

If I create a new variable named "Approved", then I can reverse the alphabetic order.

```
Data Display

Row  outcome       Approved
  1  approved      Yes
  2  not approved  No
  3  not approved  No
  4  not approved  No
  5  not approved  No
  6  not approved  No
```

In order to create the new variable, follow the path `Data > Code > Text to Text` and fill in the dialog box appropriately. Note how I needed to use quotes to handle the embedded blank in the variable value. Since "Yes" follows "No" alphabetically, we get the correct analysis on the new variable.




```
Test and CI for One Proportion: Approved

Test of p = 0.85 vs p < 0.85

Event = Yes

                                  95%
                                Upper    Exact
Variable  X  N  Sample p      Bound  P-Value
Approved  1  6  0.166667   0.581803    0.000
```

In class we looked at the binomial distribution to obtain an exact Sign Test confidence interval for the median. Examine the following to see where the exact p-value for this test comes from. If we carried the p-value to a few more decimal places, what would we report?

```
Cumulative Distribution Function

Binomial with n = 6 and p = 0.85

x  P( X <= x )
0      0.00001
1      0.00040
2      0.00589
3      0.04734
4      0.22352
5      0.62285
6      1.00000
```

Goodness-of-Fit Tests

SW Section 10.1

Example: The following data set was used as evidence in a court case. The data represent a sample of 1336 individuals from the jury pool of a large municipal court district for the years 1975-1977. The fairness of the representation of various age groups on juries was being contested. The strategy for doing this was to challenge the representativeness of the pool of individuals from which the juries are drawn. This was done by comparing the age group distribution within the jury pool against the age distribution in the district as a whole, which was available from census figures.

Age group (yrs)	Obs. Counts	Obs. Prop.	Census Prop.
18-19	23	.017	.061
20-24	96	.072	.150
25-29	134	.100	.135
30-39	293	.219	.217
40-49	297	.222	.153
50-64	380	.284	.182
65-99	113	.085	.102

A statistical question here is whether the jury pool population proportions are equal to the census proportions across the age categories. This comparison can be formulated as a **goodness-of-fit test**, which generalizes the large sample test on a single proportion to a categorical variable (here age) with $r > 2$ levels. For $r = 2$ categories, the goodness-of-fit test and large sample test on a single proportion are identical. Although this problem compares two populations, only one sample is involved because the census data is a population summary!

In general, suppose each individual in a population is categorized into one and only one of r levels or categories. Let $p_1,\ p_2, ..., p_r$ be the population proportions in the r categories, where each $p_i \geq 0$ and $p_1 + p_2 + \cdots + p_r = 1$. The hypotheses of interest in a goodness-of-fit problem are $H_0 : p_1 = p_1^0, p_2 = p_2^0, ..., p_r = p_r^0$ and H_A : not H_0, where $p_1^0,\ p_2^0, ..., p_r^0$ are given category proportions.

The plausibility of H_0 is evaluated by comparing the hypothesized category proportions to estimated (i.e. observed) category proportions $\hat{p}_1,\ \hat{p}_2,\ ..., \hat{p}_r$ from a random or representative sample of n individuals selected from the population. The discrepancy between the hypothesized and observed proportions is measured by the Pearson chi-squared statistic:

$$\chi_s^2 = \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i},$$

where O_i is the **observed** number in the sample that fall into the i^{th} category ($O_i = n\hat{p}_i$), and $E_i = np_i^0$ is the number of individuals **expected** to be in the i^{th} category when H_0 is true.

The Pearson statistic can also be computed as the sum of the squared residuals:

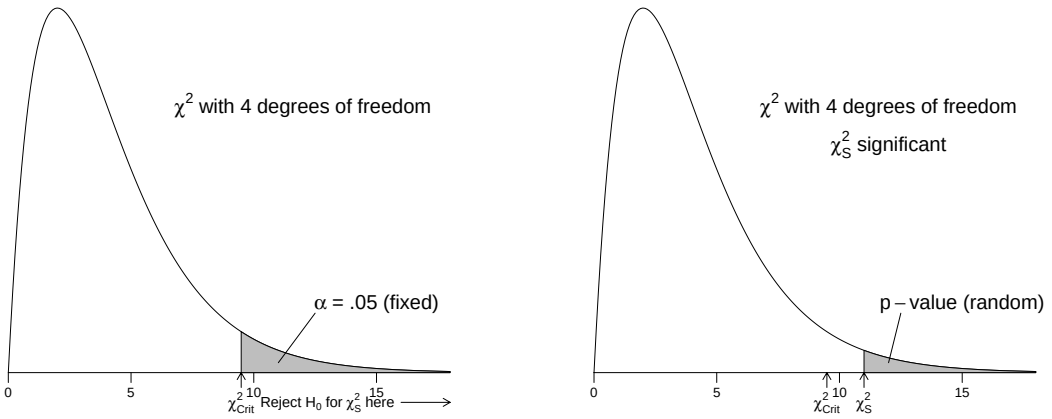
$$\chi_s^2 = \sum_{i=1}^{r} Z_i^2,$$

where $Z_i = (O_i - E_i)/\sqrt{E_i}$, or in terms of the observed and hypothesized category proportions

$$\chi^2_s = n \sum_{i=1}^{r} \frac{(\hat{p}_i - p_i^0)^2}{p_i^0}.$$

The Pearson statistic χ^2_s is "small" when all of the observed counts (proportions) are close to the expected counts (proportions). The Pearson χ^2 is "large" when one or more observed counts (proportions) differs noticeably from what is expected when H_0 is true. Put another way, large values of χ^2_s suggest that H_0 is false.

The critical value χ^2_{crit} for the test is obtained from a chi-squared probability table with $r-1$ degrees of freedom. A chi-squared table is given on page 686 of SW. The picture below shows the form of the rejection region. For example, if $r = 5$ and $\alpha = .05$, then you reject H_0 when $\chi^2_s \geq \chi^2_{crit} = 9.49$. The p-value for the test is the area under the chi-squared curve with $df = r - 1$ to the right of the observed χ^2_s value.



Example: (Jury pool problem) Let p_{18} be the proportion in the jury pool population between ages 18 and 19. Define p_{20}, p_{25}, p_{30}, p_{40}, p_{50} and p_{65} analogously. You are interested in testing $H_0 : p_{18} = .061$, $p_{20} = .150$, $p_{25} = .135$, $p_{30} = .217$, $p_{40} = .153$, $p_{50} = .182$ and $p_{65} = .102$ against H_A : not H_0, using the sample of 1336 from the jury pool.

The observed counts, the expected counts, and the category residuals are given in the table below. For example, $E_{18} = 1336 * (.061) = 81.5$ and $Z_{18} = (23 - 81.5)/\sqrt{81.5} = -6.48$ in the 18-19 year category.

The Pearson statistic is

$$\chi^2_s = (-6.48)^2 + (-7.38)^2 + (-3.45)^2 + .18^2 + 6.48^2 + 8.78^2 + (-1.99)^2 = 231.26$$

on $r - 1 = 7 - 1 = 6$ degrees of freedom. Here $\chi^2_{crit} = 12.59$ at $\alpha = .05$. The p-value for the goodness-of-fit test is less than .001, which suggests that H_0 is false.

Age group (yrs)	Obs. Counts	Exp. Counts	Residual
18-19	23	81.5	-6.48
20-24	96	200.4	-7.38
25-29	134	180.4	-3.45
30-39	293	289.9	0.18
40-49	297	204.4	6.48
50-64	380	243.2	8.78
65-99	113	136.3	-1.99

Adequacy of the Goodness-of-Fit Test

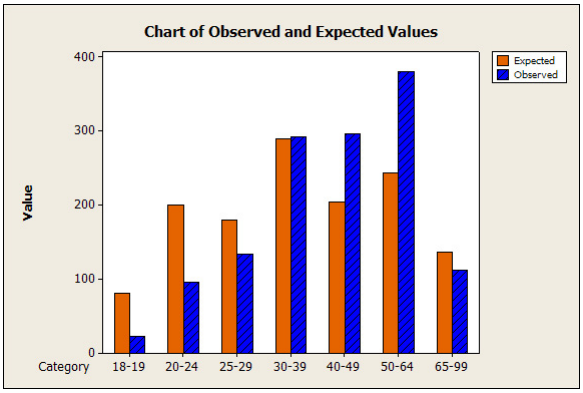
The chi-squared goodness-of-fit test is a large sample test. A conservative rule of thumb is that the test is suitable when each **expected** count is at least five. This holds in the jury pool example. There is no widely available alternative method for testing goodness-of-fit with smaller sample sizes. There is evidence, however, that the chi-squared test is **slightly conservative** (the p-values are too large, on average) when the expected counts are smaller. Some statisticians recommend that the chi-squared approximation be used when the minimum expected count is at least one, provided the expected counts are not too variable.

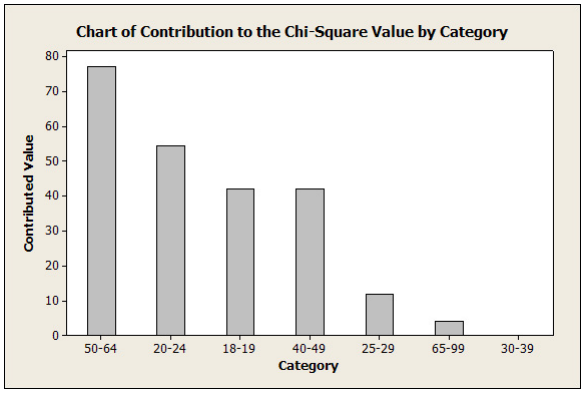
Minitab Implementation

Minitab will do a chi-squared goodness-of-fit test by following the menu path `Stat > Tables > Chi-Square Goodness-of-Fit Test (One Variable)`. Unlike the method we used for a single proportion of entering summarized data from a dialog box, the summarized data need to be entered into the worksheet (having counts for categories is summarized data). Following is the Minitab output for the jury pool problem:

```
Data Display

Row  Age    Count  CensusProp
  1  18-19     23       0.061
  2  20-24     96       0.150
  3  25-29    134       0.135
  4  30-39    293       0.217
  5  40-49    297       0.153
  6  50-64    380       0.182
  7  65-99    113       0.102

Chi-Square Goodness-of-Fit Test for Observed Counts in Variable: Count

Using category names in Age


                          Test            Contribution
Category  Observed  Proportion  Expected     to Chi-Sq
18-19           23       0.061    81.496       41.9871
20-24           96       0.150   200.400       54.3880
25-29          134       0.135   180.360       11.9164
30-39          293       0.217   289.912        0.0329
40-49          297       0.153   204.408       41.9420
50-64          380       0.182   243.152       77.0192
65-99          113       0.102   136.272        3.9743
```

```
   N  DF   Chi-Sq  P-Value
1336   6  231.260    0.000
```





The term "Contribution to Chi-Square" refers to the values of $\frac{(O-E)^2}{E}$ for each category. χ^2_s is the sum of those contributions.

Multiple Comparisons in a Goodness-of-Fit Problem

The goodness-of-fit test suggests that at least one of the age category proportions for the jury pool population differs from the census figures. A reasonable next step in the analysis would be to **separately** test the seven hypotheses: $H_0 : p_{18} = .061$, $H_0 : p_{20} = .150$, $H_0 : p_{25} = .135$, $H_0 : p_{30} = .217$, $H_0 : p_{40} = .153$, $H_0 : p_{50} = .182$ and $H_0 : p_{65} = .102$ to see which age categories led to this conclusion.

A Bonferroni comparison with a Family Error Rate $\leq .05$ will be considered for this multiple comparisons problem. The error rates for the seven individual tests are set to $\alpha = .05/7 = .0071$, which corresponds to 99.29% two-sided CIs for the individual jury pool proportions. The area under the standard normal curve to the right of 2.70 is .0035, about one-half the error rate for the individual CIs, so the critical value for the CIs, or for the tests, is $z_{crit} \approx 2.70$. The next table gives individual 99.29% CIs based on the large sample approximation. You can get the individual CIs in Minitab using the `1 Proportion` dialog box. For example, the CI for Age Group 18-19 is obtained by specifying 23 successes in 1336 trials.

The CIs for the 30-39 and 65-99 year categories contain the census proportions. In the other five age categories, there are significant differences between the jury pool proportions and the census proportions. In general, young adults appear to be underrepresented in the jury pool whereas older age groups are overrepresented.

Age group (yrs)	Lower limit	Upper limit	Census Prop.
18-19	.008	.027	.061
20-24	.053	.091	.150
25-29	.078	.122	.135
30-39	.189	.250	.217
40-49	.192	.253	.153
50-64	.251	.318	.182
65-99	.064	.105	.102

The residuals also highlight significant differences because the largest residuals correspond to the categories that contribute most to the value of χ^2_s. Some researchers use the residuals for the multiple comparisons, treating the Z_is as standard normal variables. Following this approach, you would conclude that the jury pool proportions differ from the proportions in the general population in every age category where $|Z_i| \geq 2.70$ (using a Bonferroni correction!) This gives the same conclusion as before.

The two multiple comparison methods are similar, but not identical. The residuals

$$Z_i = \frac{O_i - E_i}{\sqrt{E_i}} = \frac{\hat{p}_i - p_i^0}{\sqrt{\frac{p_i^0}{n}}}$$

agree with the large sample statistic for testing $H_0 : p_i = p_i^0$, except that the divisor in Z_i omits a $1 - p_i^0$ term. The Z_is are not standard normal random variables as assumed, and the value of Z_i underestimates the significance of the observed differences. Multiple comparisons using the Z_is will find, on average, fewer significant differences than the preferred method based on the large sample tests. However, the differences between the two methods are usually minor when all of the hypothesized proportions are small.

Comparing Two Proportions: Independent Samples

The New Mexico state legislature is interested in how the proportion of registered voters that support Indian gaming differs between New Mexico and Colorado. Assuming neither population proportion is known, the state's statistician might recommend that the state conduct a survey of registered voters sampled independently from the two states, followed by a comparison of the sample proportions in favor of Indian gaming.

Statistical methods for comparing two proportions using independent samples can be formulated as follows. Let p_1 and p_2 be the proportion of populations 1 and 2, respectively, with the attribute of interest. Let $\hat{p}_1$ and $\hat{p}_2$ be the corresponding sample proportions, based on independent random or representative samples of size n_1 and n_2 from the two populations.

Large Sample CI and Tests for $p_1 - p_2$

A large sample CI for $p_1 - p_2$ is $(\hat{p}_1 - \hat{p}_2) \pm z_{crit} SE_{CI}(\hat{p}_1 - \hat{p}_2)$, where z_{crit} is the standard normal critical value for the desired confidence level, and

$$SE_{CI}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

is the CI standard error.

A large sample p-value for a test of the null hypothesis $H_0 : p_1 - p_2 = 0$ against the two-sided alternative $H_A : p_1 - p_2 \neq 0$ is evaluated using tail areas of the standard normal distribution (identical to 1 sample evaluation) in conjunction with the test statistic

$$z_s = \frac{\hat{p}_1 - \hat{p}_2}{SE_{test}(\hat{p}_1 - \hat{p}_2)},$$

where

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

is the test standard error for $\hat{p}_1 - \hat{p}_2$. The **pooled proportion**

$$\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

is the proportion of successes in the two samples combined. The test standard error has the same functional form as the CI standard error, with $\bar{p}$ replacing the individual sample proportions.

The pooled proportion is the best guess at the common population proportion when $H_0 : p_1 = p_2$ is true. The test standard error estimates the standard deviation of $\hat{p}_1 - \hat{p}_2$ assuming H_0 is true.

Remark: As in the one-sample proportion problem, the test and CI SE's are different. This *can* (but usually does not) lead to some contradiction between the test and CI.

Example Two hundred and seventy nine French skiers were studied during two one-week periods in 1961. One group of 140 skiers receiving a placebo each day, and the other 139 receiving 1 gram of ascorbic acid (Vitamin C) per day. The study was double blind - neither the subjects nor the researchers knew who received what treatment. Let p_1 be the probability that a member of the ascorbic acid group contracts a cold during the study period, and p_2 be the corresponding probability for the placebo group. Linus Pauling and I are interested in testing whether $p_1 = p_2$. The data are summarized below as a two-by-two table of counts (a contingency table)

Outcome	Ascorbic Acid	Placebo
# with cold	17	31
# with no cold	122	109
Totals	139	140

The sample sizes are $n_1 = 139$ and $n_2 = 140$. The sample proportion of skiers developing colds in the placebo and treatment groups are $\hat{p}_2 = 31/140 = .221$ and $\hat{p}_1 = 17/139 = .122$, respectively. The pooled proportion is the number of skiers that developed colds divided by the number of skiers in the study: $\bar{p} = 48/279 = .172$.

The test standard error is:

$$SE_{test}(\hat{p}_1 - \hat{p}_2) = \sqrt{.172 * (1 - .172)\left(\frac{1}{139} + \frac{1}{140}\right)} = .0452.$$

The test statistic is

$$z_s = \frac{.122 - .221}{.0452} = -2.19.$$

The p-value for a two-sided test is twice the area under the standard normal curve to the right of 2.19 (or twice the area to the left of -2.19), which is $2 * (.014) = .028$ At the 5% level, we reject the hypothesis that the probability of contracting a cold is the same whether you are given a placebo or Vitamin C.

A CI for $p_1 - p_2$ provides a measure of the size of the treatment effect. For a 95% CI

$$z_{crit}SE_{CI}(\hat{p}_1 - \hat{p}_2) = 1.96\sqrt{\frac{.221 * (1 - .221)}{140} + \frac{.122 * (1 - .122)}{139}} = 1.96 * (.04472) = .088.$$

The 95% CI for $p_1 - p_2$ is $(.122 - .221) \pm .088$, or $(-.187, -.011)$. We are 95% confident that p_2 exceeds p_1 by at least .011 but not by more than .187.

On the surface, we would conclude that a daily dose of Vitamin C decreases a French skier's chance of developing a cold by between .011 and .187 (with 95% confidence). This conclusion was somewhat controversial. Several reviews of the study felt that the experimenter's evaluations of cold symptoms were unreliable. Many other studies refute the benefit of Vitamin C as a treatment for the common cold.

Example A case-control study was designed to examine risk factors for cervical dysplasia (Becker et al. 1994). All the women in the study were patients at UNM clinics. The 175 cases were women, aged 18-40, who had cervical dysplasia. The 308 controls were women aged 18-40 who did not have cervical dysplasia. Each women was classified as positive or negative, depending on the presence of HPV (human papilloma virus).

The data are summarized below.

HPV Outcome	Cases	Controls
Positive	164	130
Negative	11	178
Sample size	175	308

Let p_1 be the probability that a case is HPV positive and let p_2 be the probability that a control is HPV positive. The sample sizes are $n_1 = 175$ and $n_2 = 308$. The sample proportions of positive cases and controls are $\hat{p}_1 = 164/175 = .937$ and $\hat{p}_2 = 130/308 = .422$.

For a 95% CI

$$z_{crit} SE_{CI}(\hat{p}_1 - \hat{p}_2) = 1.96\sqrt{\frac{.937 * (1 - .937)}{175} + \frac{.422 * (1 - .422)}{308}} = 1.96 * (.03336) = .0659.$$

A 95% CI for $p_1 - p_2$ is $(.937 - .422) \pm .066$, or $.515 \pm .066$, or $(.449, .581)$. I am 95% confident that p_1 exceeds p_2 by at least .45 but not by more than .58.

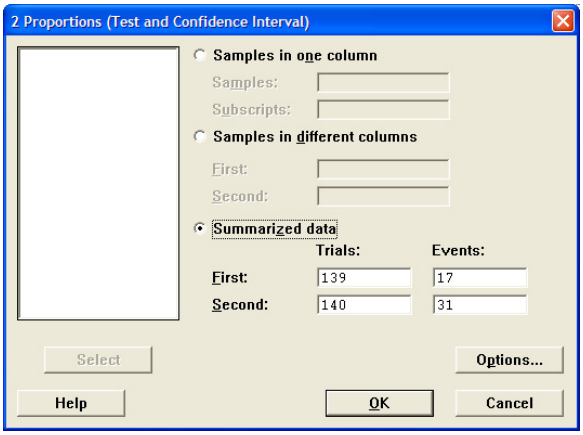
Not surprisingly, a two-sided test at the 5% level would reject $H_0 : p_1 = p_2$. In this problem one might wish to do a one-sided test, instead of a two-sided test. Let us carry out this test, as a refresher on how to conduct one-sided tests.

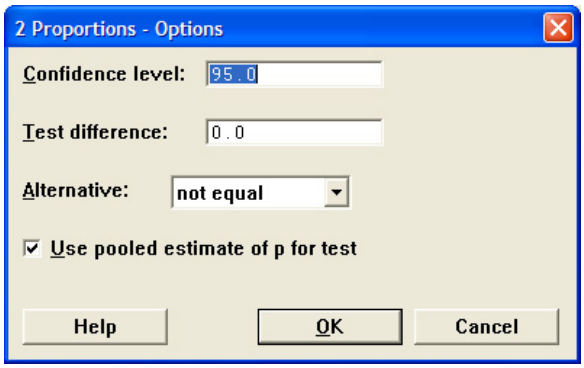
Appropriateness of Large Sample Test and CI

The standard two sample CI and test used above are appropriate when each sample is large. A rule of thumb suggests a minimum of at least five successes (i.e. observations with the characteristic of interest) and failures (i.e. observations without the characteristic of interest) in each sample before using these methods. This condition is satisfied in our two examples.

Minitab Implementation

For the Vitamin C example, in order to get Minitab to do all the calculations as presented, it is easiest to follow the menu path `Stat > Basic Statistics > 2 Proportions` and enter summary data as follows (you need to check the box for pooled estimate of p for test).






```
Test and CI for Two Proportions

Sample   X    N  Sample p
1       17  139  0.122302
2       31  140  0.221429


Difference = p (1) - p (2)
Estimate for difference:  -0.0991264
95% CI for difference:  (-0.186859, -0.0113937)
Test for difference = 0 (vs not = 0):  Z = -2.19  P-Value = 0.028
```

For the cervical dysplasia example, Minitab results are as follows:

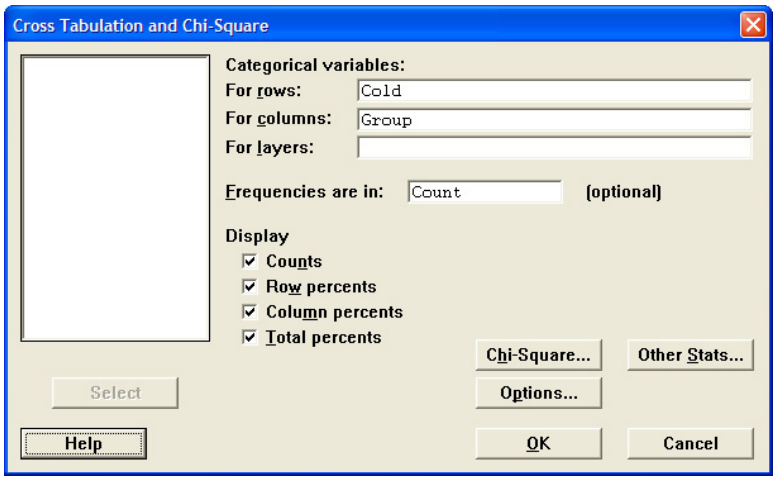
```
Test and CI for Two Proportions

Sample    X    N  Sample p
1       164  175  0.937143
2       130  308  0.422078


Difference = p (1) - p (2)
Estimate for difference:  0.515065
95% CI for difference:  (0.449221, 0.580909)
Test for difference = 0 (vs not = 0):  Z = 11.15  P-Value = 0.000
```

The above analyses are not the most common way to see data like this presented. The ability to get a confidence interval is particularly nice, and I do recommend including such an analysis. Usually, though, we present such data as a two-by-two contingency table. We need this structure in the rest of this section, so let us do that for these two examples.

The basic structure of data entry (it must be in the worksheet) is similar to our earlier use of stacked data. This is how SAS, Stata, and most other packages want it as well. For the Vitamin C example, the data are entered as follows (there are other options in Minitab - I will discuss those later):

```
Data Display

Row  Cold  Group      Count
  1  1Yes  1Vit C        17
  2  1Yes  2Placebo      31
  3  2No   1Vit C       122
  4  2No   2Placebo     109
```

The values for Cold could be entered as just Yes and No, but then Minitab alphabetizes in the presentation. What I have done is one way to get Minitab to present the table in the order we want it. Now we follow the menu path `Stat > Tables > Cross Tabulation and Chi-Square` and fill in the following box appropriately:



The various Display options and Other Stats are reflected in the following output. I structured this to present what I usually get out of SAS by default.

```
Tabulated statistics: Cold, Group

Using frequencies in Count


Rows: Cold   Columns: Group

         1Vit C  2Placebo     All

1Yes         17        31      48
           35.42     64.58  100.00
           12.23     22.14   17.20
            6.09     11.11   17.20
            23.9      24.1    48.0
          1.9990    1.9847       *

2No         122       109     231
           52.81     47.19  100.00
           87.77     77.86   82.80
           43.73     39.07   82.80
           115.1     115.9   231.0
         0.4154    0.4124       *

All         139       140     279
           49.82     50.18  100.00
          100.00    100.00  100.00
           49.82     50.18  100.00
           139.0     140.0   279.0
               *         *       *

Cell Contents:      Count
                    % of Row
                    % of Column
                    % of Total
                    Expected count
                    Contribution to Chi-square
```

```
Pearson Chi-Square = 4.811, DF = 1, P-Value = 0.028
Likelihood Ratio Chi-Square = 4.872, DF = 1, P-Value = 0.027

Fisher's exact test: P-Value =  0.0384925
```

The Pearson $\chi^2_s = 4.811$ is just the square of $Z_s = -2.19$, so for this case it's really an identical test (only for the two-sided hypothesis, though). The Likelihood Ratio Chi-Square is another large-sample test. Fisher's Exact test is another test that does not need large samples - I use it in practice very frequently. Minitab only performs this test for two-by-two tables — for more complicated tables, this is can be a very hard test to compute. SAS and Stata will at least try to compute it for arbitrary tables, though they do not always succeed. Let us examine the output to see what all these terms mean.

For the cervical dysplasia data, the results are:

```
Data Display

Row  HPV   Group    Count
  1  1Pos  Case       164
  2  1Pos  Control    130
  3  2Neg  Case        11
  4  2Neg  Control    178

Tabulated statistics: HPV, Group

Using frequencies in Count


Rows: HPV   Columns: Group

           Case  Control     All

1Pos        164      130     294
          55.78    44.22  100.00
          93.71    42.21   60.87
          33.95    26.92   60.87
          106.5    187.5   294.0
          31.01    17.62       *

2Neg         11      178     189
           5.82    94.18  100.00
           6.29    57.79   39.13
           2.28    36.85   39.13
           68.5    120.5   189.0
          48.25    27.41       *

All         175      308     483
          36.23    63.77  100.00
          100.00   100.00  100.00
          36.23    63.77  100.00
          175.0    308.0   483.0
              *        *       *

Cell Contents:      Count
                    % of Row
                    % of Column
                    % of Total
                    Expected count
                    Contribution to Chi-square
```

```
Pearson Chi-Square = 124.294, DF = 1, P-Value = 0.000
Likelihood Ratio Chi-Square = 144.938, DF = 1, P-Value = 0.000

Fisher's exact test: P-Value =  0.0000000
```

Effect Measures in Two-by-Two Tables

Consider a study of a particular disease, where each individual is either exposed or not-exposed to a risk factor. Let p_1 be the proportion diseased among the individuals in the exposed population, and p_2 be the proportion diseased among the non-exposed population. This population information can be summarized as a two-by-two table of population proportions:

Outcome	Exposed population	Non-Exposed population
Diseased	p_1	p_2
Non-Diseased	$1 - p_1$	$1 - p_2$

A standard measure of the difference between the exposed and non-exposed populations is the **absolute difference:** $p_1 - p_2$. We have discussed statistical methods for assessing this difference.

In many epidemiological and biostatistical settings, other measures of the difference between populations are considered. For example, the relative risk

$$\text{RR} = \frac{p_1}{p_2}$$

is commonly reported when the individual risks p_1 and p_2 are small. The odds ratio

$$\text{OR} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

is another standard measure. Here $p_1/(1-p_1)$ is the odds of being diseased in the exposed group, whereas $p_2/(1-p_2)$ is the odds of being diseased in the non-exposed group.

We will discuss these measures more completely next semester. At this time I will note that each of these measures can be easily estimated from data, using the sample proportions as estimates of the unknown population proportions. For example, in the vitamin C study:

Outcome	Ascorbic Acid	Placebo
# with cold	17	31
# with no cold	122	109
Totals	139	140

the proportion with colds in the placebo group is $\hat{p}_2 = 31/140 = .221$. The proportion with colds in the vitamin C group is $\hat{p}_1 = 17/139 = .122$.

The estimated absolute difference in risk is $\hat{p}_1 - \hat{p}_2 = .122 - .221 = -.099$. The estimated risk ratio and odds ratio are

$$\widehat{RR} = \frac{.122}{.221} = .55$$

and

$$\widehat{OR} = \frac{.122/(1-.122)}{.221/(1-.221)} = .49,$$

respectively.

Analysis of Paired Samples: Dependent Proportions

SW Section 10.8

Paired and more general **block analyses** are appropriate with longitudinal data collected over time and in medical studies where several treatments are given to the same patient over time. A key feature of these designs that invalidates the two-sample method discussed earlier is that repeated observations within a unit or individual are likely to be correlated, and not independent.

For example, in a random sample of $n = 1600$ voter-age Americans, 944 said that they approved of the President's performance. One month later, only 880 of the original 1600 sampled approved. The following two-by-two table gives the numbers of individuals with each of the four possible sequences of responses over time. Thus, 150 voter-age Americans approved of the President's performance when initially asked but then disapproved one month later. The row and column totals are the numbers of approvals and disapprovals for the two surveys.

(Obs Counts)	Second survey		
First Survey	Approve	Disapprove	Total
Approve	794	150	944
Disapprove	86	570	656
Total	880	720	1600

Let p_{AA}, p_{AD}, p_{DA}, p_{DD} be the population proportion of voter-age Americans that fall into the four categories, where the subscripts preserve the time ordering and indicate Approval or Disapproval. For example, p_{AD} is the population proportion that approved of the President's performance initially and disapproved one month later. The population proportion that initially approved is $p_{A+} = p_{AA} + p_{AD}$. The population proportion that approved at the time of the second survey is $p_{+A} = p_{AA} + p_{DA}$. The "+" sign used as a subscript means that the replaced subscript has been summed over.

Similarly, let $\hat{p}_{AA}$, $\hat{p}_{AD}$, $\hat{p}_{DA}$, $\hat{p}_{DD}$ be the sample proportion of voter-age Americans that fall into the four categories, and let $\hat{p}_{A+} = \hat{p}_{AA} + \hat{p}_{AD}$ and $\hat{p}_{+A} = \hat{p}_{AA} + \hat{p}_{DA}$ be the sample proportion that approves the first month and the sample proportion that approves the second month, respectively. The table below summarizes the observed proportions. For example, $\hat{p}_{AA} = 794/1600 = .496$ and $\hat{p}_{A+} = 944/1600 = .496 + .094 = .590$. The sample proportion of voting-age Americans that approve of the President's performance decreased from one month to the next.

(Obs Proportions)	Second survey		
First Survey	Approve	Disapprove	Total
Approve	.496	.094	.590
Disapprove	.054	.356	.410
Total	.550	.450	1.000

The difference in the population proportions from one month to the next can be assessed by a large sample CI for $p_{A+} - p_{+A}$, given by $(\hat{p}_{A+} - \hat{p}_{+A}) \pm z_{crit} SE_{CI}(\hat{p}_{A+} - \hat{p}_{+A})$, where the CI standard error satisfies

$$SE_{CI}(\hat{p}_{A+} - \hat{p}_{+A}) = \sqrt{\frac{\hat{p}_{A+}(1-\hat{p}_{A+}) + \hat{p}_{+A}(1-\hat{p}_{+A}) - 2(\hat{p}_{AA}\hat{p}_{DD} - \hat{p}_{AD}\hat{p}_{DA})}{n}}$$

One-sided bounds are constructed in the usual way.

The -2 term in the standard error accounts for the dependence between the samples at the two time points. If independent samples of size n had been selected for the two surveys, then this term would be omitted from the standard error, giving the usual two-sample CI.

For a 95% CI in the Presidential survey,

$$\begin{aligned} z_{crit} SE_{CI}(\hat{p}_{A+} - \hat{p}_{+A}) &= 1.96\sqrt{\frac{.590 * .410 + .550 * .450 - 2(.496 * .356 - .094 * .054)}{1600}} \\ &= 1.96 * (.0095) = .0186. \end{aligned}$$

A 95% CI for $p_{A+} - p_{+A}$ is $(.590 - .550) \pm .019$, or $(.021,\ .059)$. You are 95% confident that the population proportion of voter-age Americans that approved of the President's performance the first month was between .021 and .059 larger than the proportion that approved one month later. This gives evidence of a decrease in the President's approval rating.

A test of $H_0 : p_{A+} = p_{+A}$ can be based on the CI for $p_{A+} - p_{+A}$, or on a standard normal approximation to the test statistic

$$z_s = \frac{\hat{p}_{A+} - \hat{p}_{+A}}{SE_{test}(\hat{p}_{A+} - \hat{p}_{+A})},$$

where the test standard error is given by

$$SE_{test}(\hat{p}_{A+} - \hat{p}_{+A}) = \sqrt{\frac{\hat{p}_{A+}\hat{p}_{+A} - 2\hat{p}_{AA}}{n}}.$$

The test statistic is often written in the simplified form

$$z_s = \frac{n_{AD} - n_{DA}}{\sqrt{n_{AD} + n_{DA}}},$$

where the n_{ij}s are the observed cell counts. An equivalent form of this test, based on comparing the square of z_s to a chi-squared distribution with 1 degree of freedom, is the well-known **McNemar's test** for marginal homogeneity (or symmetry) in the two-by-two table.

For example, in the Presidential survey

$$z_s = \frac{150 - 86}{\sqrt{150 + 86}} = 4.17.$$

The p-value for a two-sided test is, as usual, the area under the standard normal curve outside ± 4.17. The p-value is less than .001, suggesting that H_0 is false.

Minitab does not have a built-in routine for paired analyses of categorical data. Small sample CI and tests for $p_{A+} - p_{+A}$ are available; see SW Section 10.8.

Testing for Homogeneity of Proportions

SW Section 10.5

Example The following two-way table of counts summarizes the location of death and age at death from a study of 2989 cancer deaths (Public Health Reports, 1983):

(Obs Counts)	Location of death			
Age	Home	Acute Care	Chronic care	Row Total
15-54	94	418	23	535
55-64	116	524	34	674
65-74	156	581	109	846
75+	138	558	238	934
Col Total	504	2081	404	2989

The researchers want to compare the age distributions across locations. A one-way ANOVA would be ideal if the actual ages were given. Because the ages are grouped, the data should be treated as categorical. Given the differences in numbers that died at the three types of facilities, a comparison of proportions or percentages in the age groups is appropriate. A comparison of counts is not.

The table below summarizes the proportion in the four age groups by location. For example, in the acute care facility 418/2081 = .201 and 558/2081 = .268. The **pooled proportions** are the Row Totals divided by the total sample size of 2989. The pooled summary gives the proportions in the four age categories, ignoring location of death.

The age distributions for home and for the acute care facilities are similar, but are very different from the age distribution at chronic care facilities.

To formally compare the observed proportions, one might view the data as representative sample of ages at death from the three locations. Assuming independent samples from the three locations (populations), a chi-squared statistic is used to test whether the population proportions of ages at death are identical (homogeneous) across locations. The **chi-squared test for homogeneity** of population proportions can be defined in terms of proportions, but is traditionally defined in terms of counts.

(Proportions)	Location of death			
Age	Home	Acute Care	Chronic care	Pooled
15-54	.187	.201	.057	.179
55-64	.230	.252	.084	.226
65-74	.310	.279	.270	.283
75+	.273	.268	.589	.312
Total	1.000	1.000	1.000	1.000

In general, assume that the data are independent samples from c populations (strata, groups, sub-populations), and that each individual is placed into one of r levels of a categorical variable. The raw data will be summarized as a $r \times c$ **contingency table** of counts, where the columns

correspond to the samples, and the rows are the levels of the categorical variable. In the age distribution problem, $r = 4$ and $c = 3$. (SW uses k to identify the number of columns.)

To implement the test:

1. Compute the (estimated) **expected** count for each cell in the table as follows:

$$\text{E} = \frac{\text{Row Total} * \text{Column Total}}{\text{Total Sample Size}}.$$

2. Compute the Pearson test statistic

$$\chi^2_s = \sum_{\text{all cells}} \frac{(O-E)^2}{E},$$

   where O is the **observed** count.

3. For a size α test, reject the hypothesis of homogeneity if $\chi^2_s \geq \chi^2_{crit}$, where χ^2_{crit} is the upper α critical value from the chi-squared distribution with $df = (r-1)(c-1)$.

The p-value for the chi-squared test of homogeneity is equal to the area under the chi-squared curve to the right of χ^2_s; see p.117.

For a two-by-two table of counts, the chi-squared test of homogeneity of proportions is identical to the two-sample proportion test we discussed earlier.

The (estimated) expected counts for the (15-54, Home) cell and for the (75+, Acute Care) cell in the age distribution data are $E = 535 * 504/2989 = 90.21$ and $E = 934 * 2081/2989 = 650.27$, respectively. The other expected counts were computed similarly, and are summarized below. The row and column sums on the tables of observed and expected counts always agree.

(Exp Counts)	Location of death			
Age	Home	Acute Care	Chronic care	Row Total
15-54	90.21	372.48	72.31	535
55-64	113.65	469.25	91.10	674
65-74	142.65	589	114.35	846
75-	157.49	650.27	126.24	934
Col Total	504	2081	404	2989

Why is a comparison of the observed and expected counts relevant for testing homogeneity? To answer this question, first note that the expected cell count can be expressed

$$\text{E} = \text{Col Total} * \text{Pooled proportion for category}.$$

For example, $E = 504 * (.179) = 90.21$ in the (15-54, Home) cell. A comparison of the observed and expected counts is a comparison of the observed category proportions in a location with the pooled proportions, taking the size of each sample into consideration. Thinking of the pooled proportion as a weighted average of the sample proportions for a category, the Pearson χ^2_s statistic is an aggregate measure of variation in the observed proportions across samples. If the category proportions are similar across samples then the category and pooled proportions are similar, resulting in a "small" value of χ^2_s. Large values of χ^2_s occur when there is substantial variation in the observed proportions across samples, in one or more categories. In this regard, the Pearson statistic is similar to the $F-$statistic in a one-way ANOVA.

Minitab Implementation

There are three standard ways to enter data for a chi-squared analysis. To illustrate, suppose we have the following summary table from a survey of 7 students, each classified by SEX and STATUS (undergrad, grad).

Status	Male	Female
undergrad	2	1
grad	1	3

The three ways to enter these data into the Minitab worksheet and conduct the chi-squared test are:

1. **Individual data:** Two columns are used to identify the row and column levels for each individual in the data set. A chi-squared test, with summary information (column, row and cell percentages, marginal percentages, cell residuals), is obtained via: `Stat > Tables > Cross Tabulation and Chi-Square`. You need to specify the two columns that contain the classification variables (here SEX and STATUS).

Obs	Sex	Status
1	m	u
2	m	u
3	f	u
4	m	g
5	f	g
6	f	g
7	f	g

2. **Frequencies:** Two columns are used to identify the levels for the rows and columns of the table. A third column gives the number of individuals in the sample for each combination of the row and columns. Then follow the procedure for the individual data, but specify the column that contains the frequencies in the dialog box (here FREQ).

Obs	Sex	Status	Freq
1	m	u	2
2	f	u	1
3	m	g	1
4	f	g	3

3. **Contingency Table:** The data are entered as a table of counts. The rows in the spreadsheet need not be labelled. In the table below, the rows correspond to undergrad and grads, respectively. A chi-squared test, with MINIMAL summary information is obtained by following: `STAT > TABLES > CHI-SQUARE TEST`. You must specify the columns that contain the counts (here MALE and FEMALE).

Obs	Male	Female
1	2	1
2	1	3

Location of Death Analysis

The location of data were entered into Minitab as a contingency table. I included an AGE column in the worksheet, but the chi-squared analysis is on the counts in the HOME, ACUTE CARE, and CHRONIC CARE columns.

In addition to the Pearson statistic, the output gives the expected counts (compare to table in notes) and the cell chi-squared values, which are the squared residuals, and can be interpreted as the contribution of the individual cells to the χ^2_s statistic (i.e $(O-E)^2/E$). The cell contributions provide insight into the categories that led to significant differences in locations. Recall our earlier discussion about the differences across locations, and then cross-reference this with the cell chi-squared values.

The output gives the Pearson statistic as 197.624 on 6 = (4-1)(3-1) df. The p-value is 0 to three places. The data strongly suggest that there are differences in the age distributions among locations.

```
Row    AGE    HOME  ACUTE CARE  CHRONIC CARE

  1  15-54     94         418            23
  2  55-64    116         524            34
  3  65-74    156         581           109
  4  75+      138         558           238

Chi-Square Test: HOME, ACUTE CARE, CHRONIC CARE

Expected counts are printed below observed counts

          HOME ACUTE CA  CHRONIC    Total
    1       94      418       23      535
         90.21   372.48    72.31

    2      116      524       34      674
        113.65   469.25    91.10

    3      156      581      109      846
        142.65   589.00   114.35

    4      138      558      238      934
        157.49   650.27   126.24

Total      504     2081      404     2989

Chi-Sq =  0.159 +  5.564 + 33.627 +
          0.049 +  6.388 + 35.789 +
          1.249 +  0.109 +  0.250 +
          2.412 + 13.092 + 98.937 = 197.624

DF = 6, P-Value = 0.000
```

Testing for Homogeneity in Cross-Sectional and Stratified Studies

Two-way tables of counts are often collected using either **stratified sampling** or **cross-sectional** sampling.

In a stratified design, distinct groups, strata, or sub-populations are identified. Independent samples are selected from each group, and the sampled individuals are classified into categories. The Indian gaming example is an illustration of a stratified design. Stratified designs provide

estimates for the strata (population) proportion in each of the categories. A test for **homogeneity of proportions** is used to compare the strata.

In a **cross-sectional design**, individuals are randomly selected from a population and classified by the levels of **two** categorical variables. With cross-sectional samples you can test homogeneity of proportions by comparing either the row proportions or by comparing the column proportions.

Example The following data (*The Journal of Advertising,* 1983, p. 34-42) are from a cross-sectional study that involved soliciting opinions on anti-smoking advertisements. Each subject was asked whether they smoked and their reaction (on a five-point ordinal scale) to the ad. The data are summarized as a two-way table of counts, given below:

	Str. Dislike	Dislike	Neutral	Like	Str. Like	Row Tot
Smoker	8	14	35	21	19	97
Non-smoker	31	42	78	61	69	281
Col Total	39	56	113	82	88	378

The row proportions (i.e. fix a row and compute the proportions for the column categories) are

(Row Prop)	Str. Dislike	Dislike	Neutral	Like	Str. Like	Row Tot
Smoker	.082	.144	.361	.216	.196	1.000
Non-smoker	.110	.149	.278	.217	.245	1.000

For example, the entry for the (Smoker, Str. Dislike) cell is: $8/97 = .082$.

Similarly, the column proportions are

(Col Prop)	Str. Dislike	Dislike	Neutral	Like	Str. Like
Smoker	.205	.250	.310	.256	.216
Non-smoker	.795	.750	.690	.744	.784
Total	1.000	1.000	1.000	1.000	1.000

Although it may be more natural to compare the smoker and non-smoker row proportions, the column proportions can be compared across ad responses. There is no advantage to comparing "rows" instead of "columns" in a formal test of homogeneity of proportions with cross-sectional data. The Pearson chi-squared test treats the rows and columns interchangeably, so you get the same result regardless of how you view the comparison. However, one of the two comparisons may be more natural to interpret.

Note that checking for homogeneity of proportions is meaningful in stratified studies only when the comparison is across strata! Further, if the strata correspond to columns of the table, then the column proportions or percentages are meaningful whereas the row proportions are not.

Question: How do these ideas apply to the age distribution problem?

Testing for Independence in a Two-Way Contingency Table

The row and column classifications for a population where each individual is cross-classified by two categorical variables are said to be **independent** if each **population** cell proportion in the two-way table is the product of the proportion in a given row and the proportion in a given column. One can show that independence is equivalent to homogeneity of proportions. In particular, the two-way table of population cell proportions satisfies independence if and only if the population column proportions are homogeneous. If the population column proportions are homogeneous then so are the population row proportions.

This suggests that a test for independence or **no association** between two variables based on a cross-sectional study can be implemented using the chi-squared test for homogeneity of proportions. This suggestion is correct. If independence is not plausible, I interpret the dependence as a deviation from homogeneity, using the classification for which the interpretation is most natural.

Example

Minitab output for testing independence between smoking status and reaction is given below. I entered the data as **frequencies** because I get more detailed summary information then when the data are entered as a **contingency table**. For example, Minitab gives the observed and expected cell counts, the cell residuals, and the percentage of all observations in the table, row, and column, respectively, found in a given cell. The row and column percentages agree with the summaries given earlier. Note that Minitab orders the rows and columns alphanumerically, which does not preserve the natural column ordering. How could this be fixed?

The chi-squared test of independence is not significant (p-value = .559). The observed association between smoking status and the ad reaction is not significant. This suggests, for example, that the smoker's reactions to the ad were not statistically significantly different from the non-smoker's reactions, which is consistent with the smokers and non-smokers attitudes being fairly similar.

Row	Smoke Stat	Reaction	Freq
1	Smoker	Str Dislike	8
2	Non	Str Dislike	31
3	Smoker	Dislike	14
4	Non	Dislike	42
5	Smoker	Neutral	35
6	Non	Neutral	78
7	Smoker	Like	21
8	Non	Like	61
9	Smoker	Str Like	19
10	Non	Str Like	69

Rows: Smoke St Columns: Reaction

	Dislike	Like	Neutral	Str Disl	Str Like	All
Non	42	61	78	31	69	281
	14.95	21.71	27.76	11.03	24.56	100.00
	75.00	74.39	69.03	79.49	78.41	74.34
	11.11	16.14	20.63	8.20	18.25	74.34
	41.63	60.96	84.00	28.99	65.42	281.00
	0.06	0.01	-0.65	0.37	0.44	--

```
Smoker      14       21       35        8       19       97
         14.43    21.65    36.08     8.25    19.59   100.00
         25.00    25.61    30.97    20.51    21.59    25.66
          3.70     5.56     9.26     2.12     5.03    25.66

         14.37    21.04    29.00    10.01    22.58    97.00
         -0.10    -0.01     1.11    -0.63    -0.75       --

All         56       82      113       39       88      378
         14.81    21.69    29.89    10.32    23.28   100.00
        100.00   100.00   100.00   100.00   100.00   100.00
         14.81    21.69    29.89    10.32    23.28   100.00

         56.00    82.00   113.00    39.00    88.00   378.00
           --       --       --       --       --       --

Chi-Square = 2.991, DF = 4, P-Value = 0.559

  Cell Contents --
                  Count
                  % of Row
                  % of Col
                  % of Tbl
                  Exp Freq
                  St Resid
```

Further Analyses in Two-Way Tables

The χ^2_s statistic is a **summary measure** of independence or homogeneity. A careful look at the data usually reveals the nature of the **association** or **heterogeneity** when the test is significant. There are numerous meaningful ways to explore two-way tables to identify sources of association or heterogeneity. For example, in the comparison of age distributions across locations, you might consider the 4×2 tables comparing all possible pairs of locations. Another possibility would be to compare the proportion in the 75+ age category across locations. For the second comparison you need a 2×3 table of counts, where the two rows correspond to the individuals less than 75 years old and those 75+ years old, respectively. (i.e. collapse the first three rows of the original 4×2 table). The possibilities are almost limitless in large tables. Of course, theoretically generated comparisons are preferred to data dredging.

Example: Testing for Homogeneity

A randomized double-blind experiment compared the effectiveness of several drugs in reducing postoperative nausea. All patients were anesthetized with nitrous oxide and ether. The following table shows the incidence of nausea during the first four postoperative hours of four drugs and a placebo. Compare the drugs to each other and to the placebo.

Drug	# with Nausea	# without Nausea	Sample Size
Placebo	96	70	166
Chlorpromazine	52	100	152
Dimenhydrinate	52	33	85
Pentobarbitol (100mg)	35	32	67
Pentobarbitol (150mg)	37	48	85

Let p_{PL} be the probability of developing nausea given a placebo, and define p_{CH}, p_{DI}, p_{PE100}, and p_{PE150} analogously. A simple initial analysis would be to test homogeneity of proportions: $H_0 : p_{PL} = p_{CH} = p_{DI} = p_{PE100} = p_{PE150}$ against H_A : not H_0.

The data were entered as frequencies. The Minitab output shows that the proportion of patients exhibiting nausea (see the **column** percents - the cell and row percentages are not interpretable, so they are omitted) is noticeably different across drugs. In particular, Chlorpromazine is the most effective treatment with $\hat{p}_{CH} = .34$ and Dimenhydrinate is the least effective with $\hat{p}_{DI} = .61$.

The p-value for the chi-squared test is .000 to three places, which leads to rejecting H_0 at the .05 or .01 levels. The data strongly suggest there are differences in the effectiveness of the various treatments for postoperative nausea.

```
Data Display

 Row    drug    reaction        freq

   1  placebo     Nausea          96
   2  placebo     noNausea        70
   3  chlorpr     Nausea          52
   4  chlorpr     noNausea       100
   5  dimenhy     Nausea          52
   6  dimenhy     noNausea        33
   7  pent100     Nausea          35
   8  pent100     noNausea        32
   9  pent150     Nausea          37
  10  pent150     noNausea        48

Tabulated Statistics: reaction, drug


 Rows: reaction     Columns: drug

           chlorpr  dimenhy  pent100  pent150  placebo      All

 Nausea         52       52       35       37       96      272
             34.21    61.18    52.24    43.53    57.83    49.01

 noNausea      100       33       32       48       70      283
             65.79    38.82    47.76    56.47    42.17    50.99

 All           152       85       67       85      166      555
            100.00   100.00   100.00   100.00   100.00   100.00

Chi-Square = 24.827, DF = 4, P-Value = 0.000
```

```
Cell Contents --
                Count
                % of Col
```

A sensible follow-up analysis is to identify which treatments were responsible for the significant differences. For example, the placebo and chlorpromazine can be compared using a test of $p_{PL} = p_{CH}$ or with a CI for $p_{PL} - p_{CH}$.

In certain experiments, specific comparisons are of interest, for example a comparison of the drugs with the placebo. Alternatively, all possible comparisons might be deemed relevant. The second case is suggested here based on the problem description. I will use a Bonferroni adjustment to account for the multiple comparisons. The Bonferroni adjustment accounts for data dredging, but at a cost of less sensitive comparisons.

There are 10 possible comparisons here. The Bonferroni analysis with an overall Family Error Rate of 0.05 (or less) tests the 10 individual hypotheses at the .05/10=.005 level. Alternatively, I can generate 99.5% CIs (which have .005 error rate) for the differences in two probabilities.

The following table gives 99.5% CIs for the differences between the ten pairs of probabilities. The only two CIs that do not cover zero correspond to $p_{PL} - p_{CH}$ and $p_{CH} - p_{DI}$. I am 99.5% confident that p_{CH} is between .084 and .389 less than p_{PL}, and you are 99.5% confident that p_{CH} is between .086 and .453 less than p_{DI}. The other differences are not significant.

Interval	Lower Limit	Upper Limit
$p_{PL} - p_{CH}$	.084	.389
$p_{PL} - p_{DI}$	-.217	.150
$p_{PL} - p_{PE100}$	-.146	.258
$p_{PL} - p_{PE150}$	-.042	.328
$p_{CH} - p_{DI}$	-.453	-.086
$p_{CH} - p_{PE100}$	-.383	.022
$p_{CH} - p_{PE150}$	-.279	.093
$p_{DI} - p_{PE100}$	-.137	.316
$p_{DI} - p_{PE150}$	-.035	.388
$p_{PE100} - p_{PE150}$	-.141	.315

Using ANOVA-type groupings, and arranging the treatments from most to least effective (low proportions to high), we get:

```
CH (.34)  PE150 (.44)  PE100 (.52)  PL (.58)  DI (.61)
----------------------------------
          ------------------------------------------
```

REMARK: How would you do these multiple comparisons in Minitab?

Adequacy of the Chi-Square Approximation

The chi-squared tests for homogeneity and independence are large sample tests. As with the goodness-of-fit test, a simple rule of thumb is that the approximation is adequate when the expected cell counts are 5 or more. This rule is conservative, and some statisticians argue that the approximation is valid for expected counts as small as one.

In practice, the chi-squared approximation to χ^2_s tends to be a bit conservative, meaning that statistically significant results would likely retain significance has a more accurate approximation been used. There are some subtle points about this issue that I will note in class.

Minitab prints out a warning message whenever a noticeable percentage of cells have expected counts less than 5. Ideally, one would use Fisher's exact test for tables with small counts, but as noted earlier, this test is not available in Minitab except for 2 X 2 tables.