AMS 315/576 Lecture Notes

Chapter 10. CATEGORICAL DATA

10.1 Inferences About one population Proportion π

The public opinion polls are becoming increasingly popular and vital in our societies nowadays. Almost daily, the news media report the results of some poll. There are polls aimed to determine the percentage of people in favor of the President's various policies, the fraction of voters in favor of a certain candidate, the percentage of customers who plan to buy a certain product within the next year, and the proportion of college students who smoke cigarettes. In each case, we are interested in estimating the percentage (or proportion) of a population with a certain characteristic. In this section we consider methods for making inferences about one population proportion when the sample is considered large.

1. Sample. The sample is summarized entirely by 2 statistics: (1). the sample size n, and (2). the total number of subjects (or objects) in the sample that possess the characteristic of interest, x.

2. Point estimator. The population proportion π is estimated by the sample proportion $\hat{\pi} = x/n$. When the sample is considered large, namely $n\pi \geq 5$ and $n(1-\pi) \geq 5$, the distribution of $\hat{\pi}$ can be approximated by a normal distribution with mean π and variance $\pi(1-\pi)/n$.

3. Large sample confidence interval.

Confidence Interval for π, with Confidence Coefficient of $(1-\alpha)$

$$\hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n}},$$

where

$$\hat{\pi} = \frac{x}{n}.$$

Note: the large sample confidence interval is valid when $n\widehat{\pi} = x \geq 5$ and $n\,(1-\widehat{\pi}) = n - x \geq 5$.

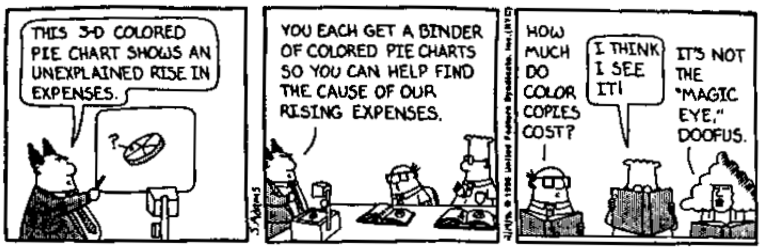
4. Sample size calculation.

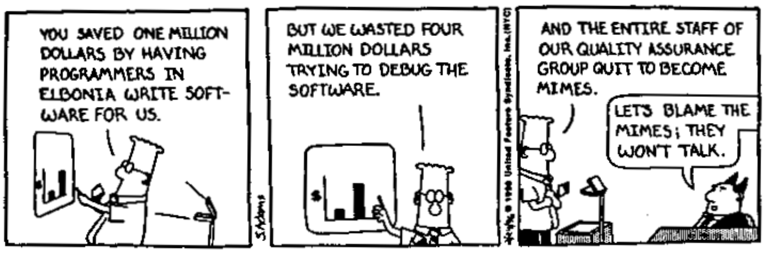
Sample Size Required for a $100(1-\alpha)\%$ Confidence Interval for π of the Form $\hat{\pi} \pm E$

$$n = \frac{z^2_{\alpha/2}\pi(1-\pi)}{E^2} \quad \pi \text{ is known}$$

$$n = \frac{z^2_{\alpha/2}}{4E^2} \quad \pi \text{ is unknown}$$

Note. The above sample size will ensure a $100(1-\alpha)\%$ Confidence Interval for π to be of the required length 2E, or equivalnetly, a maximum error of E with probability $100(1-\alpha)\%$.





5. Hypothesis testing.

Summary of a Statistical Test for π

H_0: $\pi = \pi_0$ (π_0 is specified)

H_a:
1. $\pi > \pi_0$
2. $\pi < \pi_0$
3. $\pi \neq \pi_0$

T.S.:

$$z_0 = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

R.R.: For a probability α of a Type I error

1. reject H_0 if $z_0 > z_\alpha$
2. reject H_0 if $z_0 < -z_\alpha$
3. reject H_0 if $|z_0| > z_{\alpha/2}$

Note: the large sample test is valid when $n\pi_0 \geq 5$ and $n\,(1 - \pi_0) \geq 5$.

EXAMPLE 10.1

Sports car owners in a town complain that their cars are judged differently from family-style cars at the state vehicle inspection station. Previous records indicate that 30% of all passenger cars fail the inspection on the first time through. In a random sample of 150 sports cars, 60 failed the inspection on the first time through. Is there sufficient evidence to indicate that the percentage of first failures for sports cars is higher than the percentage for all passenger cars? Use $\alpha = .05$.

SOLUTION $n = 150$, $x = 60$, $\hat{\pi} = \frac{60}{150} = .4$

H$_0$: $\pi = .30$

H$_a$: $\pi > .30$

Since

$$n\pi_0 = 150(.3) = 45 \geq 5$$

and

$$n(1-\pi_0) = 150(.7) = 105 \geq 5$$

we can use the large sample test.

T.S. :

$$z_0 = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.4 - 0.3}{\sqrt{\frac{(.3)(.7)}{150}}} = 2.7$$

R.R. : Since $z_0 = 2.7 > z_{0.05} = 1.645$, we reject H_0 at $\alpha = .05$ and conclude that sports cars at the vehicle inspection station have a first-failure rate greater than .3.

EXAMPLE 10.2

Response to an advertising display was measured by counting the number of people who purchased the product out of the total number exposed to the display. If 330 purchased the product out of a total of 870 exposed, estimate the proportion of all persons exposed who will buy the product. Use a 90% confidence interval.

SOLUTION $n = 870$, $x = 330$,

$$\hat{\pi} = \frac{330}{870} = .38$$

Since $x = 330 \geq 5$ and $n - x = 520 \geq 5$, we can use the large sample confidence interval for π:

$$\hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n}},$$

Simple calculations show that the 90% confidence interval on the proportion of persons who will purchase the product after exposure to this display is

$$.38 \pm 1.645(.016) \quad \text{or} \quad .38 \pm .026.$$

10.2 Comparing Two Population Proportions

Many practical problems involve the comparison of two population proportions. For example, social scientists may wish to compare the proportions of women who take advantage of prenatal health services for two communities representing different socioeconomic backgrounds. Or, the director of marketing may wish to compare the public awareness of a new product recently launched and that of a competitor's product.

1. Samples. For comparisons of this type, we assume that independent random samples are drawn from two binomial populations with unknown parameters designated by π_1 and π_2. If x_1 successes are observed for the random

sample of size n_1 from population 1 and x_2 successes are observed for the random sample of size n_2 from population 2, then the point estimates of π_1 and π_2 are the observed sample proportions $\hat{\pi}_1$ and $\hat{\pi}_2$, respectively:

$$\hat{\pi}_1 = \frac{x_1}{n_1} \quad \textit{and} \quad \hat{\pi}_2 = \frac{x_2}{n_2}$$

This notation is summarized here.

Notation for Comparing Two Binomial Proportions

	Population 1	 2
Population proportion	π_1	π_2
Sample size	n_1	n_2
Number of successes	x_1	x_2
Sample proportion	$\hat{\pi}_1 = \frac{x_1}{n_1}$	$\hat{\pi}_2 = \frac{x_2}{n_2}$

2. Point estimator. Inferences about two binomial proportions are usually phrased in terms of their difference $\pi_1 - \pi_2$, and we use the difference in sample proportions $\hat{\pi}_1 - \hat{\pi}_2$ to estimate the corresponding population difference. The sampling distribution for $\hat{\pi}_1 - \hat{\pi}_2$ can be approximated by a normal distribution with mean and standard error given by

$$\mu_{\hat{\pi}_1 - \hat{\pi}_2} = \pi_1 - \pi_2$$

and

$$\sigma_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}.$$

This approximation is appropriate if we apply the same requirements to both binomial populations that we did in recommending a normal approximation to a binomial (see chapter 4). Thus, the normal approximation to the distribution of $\hat{\pi}_1 - \hat{\pi}_2$ is appropriate if both $n\pi$ and $n(1-\pi)$ are 5 or more for each sample.

3. Large samples confidence interval for $\pi_1 - \pi_2$.

$100(1-\alpha)\%$ Confidence Interval for $\pi_1 - \pi_2$

$$\hat{\pi}_1 - \hat{\pi}_2 \pm z_{\alpha/2}\sigma_{\hat{\pi}_1-\hat{\pi}_2},$$

where

$$\sigma_{\hat{\pi}_1-\hat{\pi}_2} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}.$$

Note: Substitute $\hat{\pi}_1$ and $\hat{\pi}_2$ for π_1 and π_2 in the formula for $\sigma_{\hat{\pi}_1-\hat{\pi}_2}$. When the normal approximation is valid for $\hat{\pi}_1-\hat{\pi}_2$, very little error will result from this substitution.

$$\hat{\pi}_1 - \hat{\pi}_2 \pm z_{\alpha/2}\sigma_{\hat{\pi}_1-\hat{\pi}_2},$$

where

$$\sigma_{\hat{\pi}_1-\hat{\pi}_2} = \sqrt{\frac{\widehat{\pi}_1(1-\widehat{\pi}_1)}{n_1} + \frac{\widehat{\pi}_2(1-\widehat{\pi}_2)}{n_2}}.$$

Note: This CI is valid when $x_1 \geq 5$, $n_1 - x_1 \geq 5$, $x_2 \geq 5$, and $n_2 - x_2 \geq 5$.

EXAMPLE 10.3

In a survey to analyze the funeral expenditure for various social classes, a random sample of 162 families from the working (blue-collar) class was taken to determine the funeral expenses for a recent death in the family. Of the 162 families contacted, 61 spend over \$800 on the funeral. A similar survey was conducted within the middle/upper classes. Of 189 families contacted, 106 spend more than \$800. Estimate $\pi_1 - \pi_2$, the difference in the proportions of families who have spent more than \$800 for a recent family death. Use a 95% confidence interval to interpret your findings.

SOLUTION The point estimate of $\pi_1 - \pi_2$ is the difference in sample proportions, $\hat{\pi}_1 - \hat{\pi}_2$:

$$\hat{\pi}_1 - \hat{\pi}_2 = \frac{61}{162} - \frac{106}{189} = .376 - .561 = -.185.$$

Note also that $n\hat{\pi}$ and $n(1-\hat{\pi})$ are 5 or more for both samples, implying that the normal approximation to the binomial is appropriate.

The standard error for $\hat{\pi}_1 - \hat{\pi}_2$ is estimated by

$$\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} = \sqrt{\frac{.376(.624)}{162} + \frac{.561(.439)}{189}} = .052.$$

A 95% confidence interval for $\pi_1 - \pi_2$ has $z_{\alpha/2} = 1.96$ and is of the form

$$\text{Point estimate} \pm z_{\alpha/2}(\text{standard error}).$$

Substituting into this formula we have

$$-.185 \pm 1.96(.052) \quad \text{or} \quad -.185 \pm .102,$$

This interval indicates that π_2 is larger than π_1; we are 95% confident that the difference in the proportions of families paying more that \$800 per funeral for the working class (π_1) and the middle/upper class (π_2) lies in the interval $-.287$ to $-.083$.

We can readily formulate a statistical test for the equality of two binomial parameters. The test statistic for testing $H_0 : \pi_1 - \pi_2 = 0$ is a z statistic having the familiar form

$$z = \frac{\text{point estimate}}{\text{standard error}} = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sigma_{\hat{\pi}_1 - \hat{\pi}_2}}.$$

The standard error is slightly different from what we used for a confidence interval. When H_0 is true, $\pi_1 = \pi_2$; we'll call the common value π. Then

$$\sigma_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}} = \sqrt{\pi(1-\pi)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

The best estimate of π, the proportion of successes common to both populations, is

$$\hat{\pi} = \frac{\text{total number of successes}}{\text{total number of trials}} = \frac{x_1 + x_2}{n_1 + n_2}.$$

We have summarized the test procedure here.

10.3 Large Samples Test.

Statistical Test for Comparing Two Binomial Proportions

H$_0$: $\pi_1 - \pi_2 = 0$

H$_a$:
1. $\pi_1 - \pi_2 > 0$
2. $\pi_1 - \pi_2 < 0$
3. $\pi_1 - \pi_2 \neq 0$

T.S.:

$$z_0 = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\widehat{\pi}(1-\widehat{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}},$$

where

$$\hat{\pi} = \frac{\text{total number of successes}}{\text{total number of trials}} = \frac{x_1 + x_2}{n_1 + n_2}$$

R.R.: For a probability α of a Type I error

1. reject H_0 if $z_0 > z_\alpha$
2. reject H_0 if $z_0 < -z_\alpha$
3. reject H_0 if $|z_0| > z_{\alpha/2}$

Note: $n\hat{\pi}$ and $n(1-\hat{\pi})$ must be greater than or equal to 5 for both populatons in order for the normal approximation (and hence for this test) to hold. That is, the above test is valid when $x_1 \geq 5$, $n_1 - x_1 \geq 5$, $x_2 \geq 5$, and $n_2 - x_2 \geq 5$.

EXAMPLE 10.4

In a recent survey of county high school students ($n_1 = 100$ males and $n_2 = 100$ females), 58 of the males and 46 of the females sampled said they consume alcohol on a regular basis. Use the sample data to conduct a test $H_0 : \pi_1 - \pi_2 = 0$ against the one-sided alternative $H_a : \pi_1 - \pi_2 > 0$, that a higher proportion of males than females consume alcohol on a regular basis. Use $\alpha = .05$.

SOLUTION The four parts of the statistical test are shown here:

H$_0$: $\pi_1 - \pi_2 = 0$

H$_a$: $\pi_1 - \pi_2 > 0$

T.S.:

$$z_0 = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\widehat{\pi}(1-\widehat{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

R.R.: For $\alpha = .05$, reject H_0 if $z > 1.645$.

From the sample data we find

$$\hat{\pi}_1 = \frac{58}{100} = .58, \quad \hat{\pi}_2 = \frac{46}{100} = .46, \quad \text{and} \quad \hat{\pi} = \frac{58+46}{100+100} = .52,$$

Note also that $n\hat{\pi}$ and $n(1-\hat{\pi})$ are 5 or more for both samples, validating the normal approximations to the binomial.

Substituting into the test statistic, we obtain

$$z = \frac{.58 - .46}{\sqrt{.52(.48)\left(\frac{1}{100}+\frac{1}{100}\right)}} = \frac{.12}{.071} = 1.69.$$

Conclusion: Since $z = 1.69$ exceeds 1.645, we reject $H_0 : \pi_1 - \pi_2 = 0$; we have shown that a higher proportion of high school males than females in the county studies consumes alcohol on a regular basis.

10.4 The Multinomial Experiment and Chi-Square Goodness-of-fit Test

The Multinomial Experiment

1. The experiment consists of n identical trials.
2. Each trial results in one of k outcomes.
3. The probability that a single trial will result in outcome i is π_i, $i = 1, 2, \ldots, k$, and remains constant from trial to trial. (Note: $\sum_i \pi_i = 1.$)
4. The trials are independent.
5. We are interested in n_i, the number of trials resulting in outcome i. (Note: $\sum_i n_i = n$.)

The probability distribution for the number of observations resulting in each of the k outcomes, called the **multinomial distribution**, is given by the formula

$$P(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1!n_2!\ldots n_k!}\pi_1^{n_1}\pi_2^{n_2}\cdots\pi_k^{n_k}$$

Recall from Chapter 4, where we discussed the binomial probability distribution, that

$$n! = n(n-1)\cdots 1$$

and

$$0! = 1.$$

We can use the formula for the multinomial distribution to compute the probability of particular events.

Note, when $k = 2$, the multinomial experiment reduces to the binomial experiment. Equivalently, the multinomial distribution reduces to the binomial distribution.

EXAMPLE 10.5 (only required for AMS graduate students)

Previous experience with the breeding of a particular herd of cattle suggests that the probability of obtaining one healthy calf from a mating is .83. Similarly, the probabilities of obtaining zero or two healthy calves are, respectively, .15 and .02. If a farmer breeds three dams from the herd, find the probability of obtaining exactly three healthy calves.

SOLUTION Assuming the three dams are chosen at random, this experiment can be viewed as a multinomial experiment with $n = 3$ trials and $k = 3$ outcomes. These outcomes are listed below with the corresponding probabilities.

Outcome	**Number of Progeny**	**Probability, π_i**
1	0	.15
2	1	.83
3	2	.02

Note that outcomes 1,2, and 3 refer to the events that a dam produces zero, one, or two healthy calves, respectively. Similarly, n_1, n_2, and n_3 refer to the number of dams producing zero, one, or two healthy progeny, respectively. To obtain exactly three healthy progeny, we must observe one of the following possible events.

$$\text{A} \quad : \quad \begin{cases} \text{1 dam gives birth to no healthy progeny} & : \quad n_1 = 1 \\ \text{1 dam gives birth to 1 healthy progeny} & : \quad n_2 = 1 \\ \text{1 dam gives birth to 2 healthy progeny} & : \quad n_3 = 1 \end{cases}$$

$$\text{B} \quad : \quad \text{3 dams give birth to 1 healthy progeny:} \quad \begin{cases} n_1 = 0 \\ n_2 = 3 \\ n_3 = 0 \end{cases}$$

For event A with $n = 3$ and $k = 3$.

$$P(n_1 = 1, n_2 = 1, n_3 = 1) = \frac{3!}{1!1!1!}(.15)^1(.83)^1(.02)^1 \approx .015.$$

Similarly, for event B,

$$P(n_1 = 0, n_2 = 3, n_3 = 0) = \frac{3!}{0!3!0!}(.15)^0(.83)^3(.02)^0 = (.83)^3 \approx .572.$$

Thus, the probability of obtaining exactly three healthy progeny from three dams is the sum of the probabilities for events A and B; namely, $.015 + .572 \approx .59$.

Chi-Square Goodness-of-Fit Test

Null hypothesis: $\pi_i = \pi_{i0}$ for categories $i = 1, \ldots, k$. π_{i0} are specified probabilitis or proportions.

Altenative hypothesis: At least one of the cell probabilities differes from the hypothesized value.

Test statistic: $\chi^2 = \sum_i \left[\frac{(n_i - E_i)^2}{E_i}\right]$, where n_i is the observed number in category i and $E_i = n\pi_{i0}$ is the expected number under H_0.

Rejection region: Reject H_0 if χ^2 exceeds the tabulated critical value for $a = \alpha$ and df $= k - 1$.

Some researchers (see, for example, Siegel (1956) and Dixon and Massey (1969)) recommend that all the E_is should be 5 or more before performing this test. This requirement is perhaps too stringent. Cochran (1954) indicates that the approximation should be quite good if no E_i is less than 1 and no more than 20% of the E_is are less than 5. We recommend applying Chochran's guidelines for determining whether χ^2 can be approximated with a chi-square distribution. We can combine categories if some of the E_is are too small, but care should be taken so that the combination of categories does not change the nature of the hypothesis to be tested.

EXAMPLE 10.6

A test drug is to be compared against a standard drug preparation useful in the maintenance of patients suffering from high blood pressure. Over many clinical trials at many different locations, patients suffering from comparable hypertension (as measured by the New York Heart Association (NYHA Classification) have been administered the standard therapy. Responses therapy for this large patient group were classified into one of four responses categories. Table 10.1 lists the categories and percentages of patients treatment on the standard preparation who have been classified in each category.

Category	**Percentage**
Marked decrease in blood pressure	50%
Moderate decrease in blood pressure	25%
Sight decrease in blood pressure	10%
Stationary or slight increase in blood pressure	15%

A clinical trial is conducted with a random sample of 200 patients suffering from high blood pressure. All patients are required to be listed according to the same hypertensive categories of the NYHA Classification as those studied under the standard preparation. Use the sample data in the following table to test the hypothesis that the cell probabilities associated with the test preparation are identical to those for the standard. Use $\alpha = .05$.

Category	Observed Cell Counts
1	120
2	60
3	10
4	10

SOLUTION This experiment possesses the characteristics of a multinomial experiment, with $n = 200$ and $k = 4$ outcomes.

Outcome 1: A person's blood pressure will decrease markedly after treatment on the test drug.

Outcome 2: A person's blood pressure will decrease moderately after treatment on the test drug.

Outcome 3: A person's blood pressure will decrease slightly after treatment on the test drug.

Outcome 4: A person's blood pressure will remain stationary or increase slightly after treatment on the test drug.

The null and alternative hypotheses are then

$$H_0 : \pi_1 = .50, \pi_2 = .25, \pi_3 = .10, \pi_4 = .15$$

and

H_a : At least one of the cell probabilities is different from the hypothesized value.

Before computing the test statistic, we must determine the expected cell numbers. These data are given in Table 10.3.

Category	**Observed Cell Number**, n_i	**Expected Cell Number**, E_i
1	120	200(.50) = 100
2	60	200(.25) = 50
3	10	200(.10) =20
4	10	200(.15) = 30

Since all the expected cell numbers are large, we may calculate the chi-square statistic and compare it to a tabulated value of the chi-square distribution

$$\begin{aligned}
\chi^2 &= \sum_i \left[\frac{(n_i - E_i)^2}{E_i}\right] \\
&= \frac{(120-100)^2}{100} + \frac{(60-50)^2}{50} + \frac{(10-20)^2}{20} + \frac{(10-30)^2}{30} \\
&= 4 + 2 + 5 + 13.33 = 24.33.
\end{aligned}$$

For the probability of a Type I error set at $\alpha = .05$, we look up the value of the chi-square statistic for $a = .05$ and df $= k - 1 = 3$. The critical value from Table 5 in the Appendix is 7.815.

$$\text{R.R. : Reject } H_0 \quad \text{if} \quad \chi^2 > 7.815$$

Conclusion: Since the computed value of χ^2 is greater than 7.815, we reject the null hypothesis and conclude that at least one of the cell probabilities differs from that specified under H_0. Practically, it appears that a much higher proportion of patients treated with the test preparation fall into the moderate and marked improvement categories.

EXAMPLE 10.7

Gregor Mendel (1822-1884) was an Austrian monk whose genetic theory is one of the greatest scientific discovery of all time. In his famous experiment with garden peas, he proposed a genetic model that would explain the occurrence of hereditary characteristics. In particular, he studied how the shape (smooth or wrinkled) and color (yellow or green) of pea seeds are transmitted through generations. According to his model, the second generation of peas from a certain ancestry should have the following distribution:

$wrinkled - green$	$wrinkled - yellow$	$smooth - green$	$smooth - yellow$
1/16	3/16	3/16	9/16

The outcome of his experiment was as follows:

$wrinkled - green$	$wrinkled - yellow$	$smooth - green$	$smooth - yellow$
31	102	108	315

Does his experiment support his theory? Test at $\alpha = 0.05$.
SOLUTION This is a multinomial experiment. We will use the chi-square

goodness-of-fit test to see whether the data approve of Mendel's theory.
$H_0 : \pi_1 = 9/16,\ \pi_2 = 3/16,\ \pi_3 = 3/16,\ \pi_4 = 1/16$. H_a : at least one of the previous equations is not true. If the null hypothesis is true, we would expect the following distribution of the 556 $(315 + 108 + 102 + 31)$ pea seeds in each of the four categories:

smooth-yellow $e_1 = 556 \times 9/16 = 312.75$	smooth-green $e_2 = 556 \times 3/16 = 104.25$
wrinkled-yellow $e_3 = 556 \times 3/16 = 104.25$	wrinkled-green $e_4 = 556 \times 1/16 = 34.75$

The test statistics under the null hypothesis is

$$\chi_0^2 = \frac{(315 - 312.75)^2}{312.75} + \frac{(108 - 104.25)^2}{104.25} + \frac{(102 - 104.25)^2}{104.25} + \frac{(31 - 34.75)^2}{34.75} \approx 0.604$$

Since $\chi_0^2 < \chi_{3,.05}^2 = 7.815$, we can not reject H_0 at $\alpha = .05$. The small value of $\chi_0^2 = .0604$ suggests that the Mendel theory is true.