STATISTICS FOR ECONOMISTS: A BEGINNING

John E. Floyd
University of Toronto

July 2, 2010

PREFACE

The pages that follow contain the material presented in my introductory quantitative methods in economics class at the University of Toronto. They are designed to be used along with any reasonable statistics textbook. The most recent textbook for the course was James T. McClave, P. George Benson and Terry Sincich, *Statistics for Business and Economics*, Eighth Edition, Prentice Hall, 2001. The material draws upon earlier editions of that book as well as upon John Neter, William Wasserman and G. A. Whitmore, *Applied Statistics*, Fourth Edition, Allyn and Bacon, 1993, which was used previously and is now out of print. It is also consistent with Gerald Keller and Brian Warrack, *Statistics for Management and Economics*, Fifth Edition, Duxbury, 2000, which is the textbook used recently on the St. George Campus of the University of Toronto. The problems at the ends of the chapters are questions from mid-term and final exams at both the St. George and Mississauga campuses of the University of Toronto. They were set by Gordon Anderson, Lee Bailey, Greg Jump, Victor Yu and others including myself.

This manuscript should be useful for economics and business students enrolled in basic courses in statistics and, as well, for people who have studied statistics some time ago and need a review of what they are supposed to have learned. Indeed, one could learn statistics from scratch using this material alone, although those trying to do so may find the presentation somewhat compact, requiring slow and careful reading and thought as one goes along.

I would like to thank the above mentioned colleagues and, in addition, Adonis Yatchew, for helpful discussions over the years, and John Maheu for helping me clarify a number of points. I would especially like to thank Gordon Anderson, who I have bothered so frequently with questions that he deserves the status of mentor.

After the original version of this manuscript was completed, I received some detailed comments on Chapter 8 from Peter Westfall of Texas Tech University, enabling me to correct a number of errors. Such comments are much appreciated.

J. E. Floyd
July 2, 2010

Chapter 7

Inferences About Population Variances and Tests of Goodness of Fit and Independence

In the last chapter we made inferences about whether two population means or proportions differed based on samples from those populations. Integral in all those tests and in the inferences in the previous chapters about population means and population proportions was our use of the statistic

$$s^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} \tag{7.1}$$

as an unbiased point estimate of the population variance σ^2. A natural next step is to make inferences—set confidence intervals and test hypotheses—about σ^2 on the basis of the sample statistic s^2.

7.1 Inferences About a Population Variance

To proceed we must know the sampling distribution of s^2. This involves the *chi-square (χ^2) distribution*, the basis of which must now be explained. Suppose we have a set of independent random draws from a variable

$$X_1, X_2, X_3, \ldots\ldots$$

which is normally distributed with population mean μ and variance σ^2. Consider this sequence in standardised form

$$Z_1, Z_2, Z_3, \ldots\ldots$$

where, of course,

$$Z_i = \frac{X_i - \mu}{\sigma}.$$

Now square the Z_i to obtain

$$Z_i^2 = \frac{(X_i - \mu)^2}{\sigma^2}.$$

It turns out that the sum of n of these squared standardised independent normal variates,

$$Z_1^2 + Z_2^2 + Z_3^3 + \ldots + Z_n^2,$$

is distributed as a chi-square distribution. We can thus write

$$\sum_{i=1}^{n} Z_i^2 \quad = \quad \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \quad = \quad \chi^2(n) \tag{7.2}$$

where $\chi^2(n)$ is a chi-square random variable—that is, a random variable distributed according to the chi-square distribution—with parameter n, which equals the number of independent normal variates summed. This parameter is typically referred to as the degrees of freedom. Notice now that

$$\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}$$

differs from the expression above in that $\bar{X}$ replaces μ in the numerator. This expression is also distributed as χ^2—indeed

$$\sum_{i=1}^{n} Z_i^2 \quad = \quad \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \quad = \quad \chi^2(n-1) \tag{7.3}$$

where the parameter, the degrees of freedom, is now $n - 1$.

At this point is worth while to pay further attention to what we mean by *degrees of freedom.* The degrees of freedom is the number of independent pieces of information used in calculating a statistic. In the expression immediately above, the n deviations of the X_i from their sample mean contain only $n - 1$ independent pieces of information. The sample mean is constructed from the n sample values of X_i by summing the X_i and dividing by

n. Accordingly, the sum of the deviations around this mean must be zero. Hence, if we know any $n-1$ of the n deviations around the mean we can calculate the remaining deviation as simply the negative of the sum of the $n-1$ deviations. Hence, only $n-1$ of the deviations are freely determined in the sample. This is the basis of the term 'degrees of freedom'. Even though there are n deviations, only $n-1$ of them produce independent sum of squared deviations from the sample mean. This is in contrast to the sum of squared deviations about the *true* mean μ, which contains n independent pieces of information because μ is independent of all the sample observations. Information is not used up in calculating the population mean as it is in calculating $\bar{X}$. This is why the standardised sum of squared deviations of the sample values about the true mean is distributed as $\chi^2(n)$ whereas the sum of squared deviations of the sample values from the sample mean, standardised by the true variance σ^2, is distributed as $\chi^2(n-1)$.

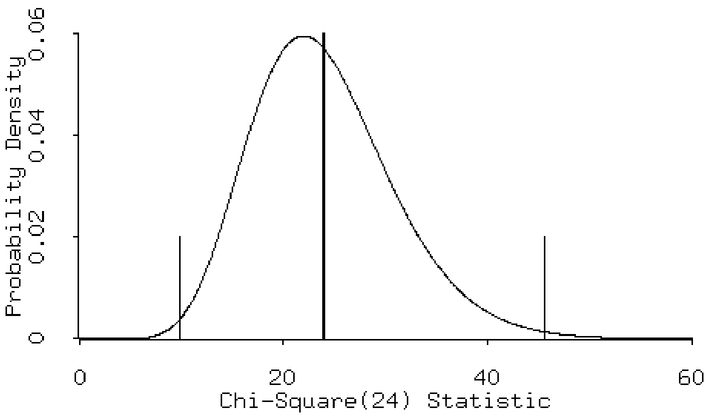

Figure 7.1: A chi-square distribution with 24 degrees of freedom. The thick vertical line shows the mean and the thin vertical lines the critical values for $\alpha = .99$.

Notice now that the expression for s^2, given in equation (7.1) above, can be rewritten

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad = \quad (n-1)\, s^2. \tag{7.4}$$

Substituting this into (7.3), we obtain

$$\frac{(n-1)\,s^2}{\sigma^2} \quad = \quad \chi^2(n-1). \tag{7.5}$$

The sampling distribution for this statistic is skewed to the right, with the skew being smaller the greater the degrees of freedom. Figure 7.1 shows a χ^2 distribution with 24 degrees of freedom. The thick vertical line gives the mean and the thin vertical lines the critical values for $\alpha = .99$. The mean of the χ^2 distribution is the number of degrees of freedom, usually denoted by v which in the above examples equals either n or $n-1$ or in Figure 7.1, 24. Its variance is $2\,v$ or twice the number of degrees of freedom. The percentiles of the χ^2 distribution (i.e., the fractions of the probability weight below given values of χ^2) for the family of chi-square distributions can be obtained from the chi-square tables at the back of any standard textbook in statistics.[1]

Now let us look at an example. Suppose a sample of 25 mature trout whose lengths have a standard deviation of 4.35 is taken from a commercial fish hatchery. We want a confidence interval for the true population variance σ^2, based on the two statistics $s^2 = 18.9225$ and $n = 25$. From a standard chi-square table we obtain the values of the χ^2 distribution with 24 degrees of freedom below which and above which the probability weight is .005,

$$\chi^2(\alpha/2; n-1) = \chi^2(.005; 24) = 9.89$$

and

$$\chi^2(1-\alpha/2; n-1) = \chi^2(.995; 24) = 45.56.$$

These are indicated by the thin vertical lines in Figure 7.1. Substituting these values into (7.5) after rearranging that expression to put σ^2 on the right-hand-side, we obtain

$$L = \frac{24s^2}{\chi^2(.995; 24)} = \frac{(24)(18.9225)}{45.56} = 9.968$$

and

$$U = \frac{24s^2}{\chi^2(.005; 24)} = \frac{(24)(18.9225)}{9.89} = 45.919$$

so that

$$9.968 \leq \sigma^2 \leq 45.919.$$

[1] Or calculated using XlispStat or another statistical computer program.

Now suppose we want to test whether the population variance of the lengths of trout in this hatchery differs from $\sigma_0^2 = 16.32$, an industry standard, controlling the α-risk at .01 when $\sigma = 16.32$. The null and alternative hypothesis then are

$$H_0\colon\ \sigma^2 = 16.32$$

and

$$H_1\colon\ \sigma^2 \neq 16.32.$$

From (7.5) the test statistic is

$$X = \frac{(n-1)s^2}{\sigma_0^2}$$

which we have shown to be distributed as $\chi^2(n-1)$. Its value is

$$X = \frac{(24)(18.9225)}{16.32} = 27.82$$

which can be compared the critical values 9.89 and 45.56 beyond which we would reject the null hypothesis of no difference between the variance of the lengths of trout in this hatchery and the industry standard. Clearly, the test statistic falls in the acceptance region so that we cannot reject the null hypothesis.

7.2 Comparisons of Two Population Variances

We are often interested in comparing the variability of two populations. For example, consider a situation where two technicians have made measurements of impurity levels in specimens from a standard solution. One technician measured 11 specimens and the other measured 9 specimens. Our problem is to test whether or not the measurements of impurity levels have the same variance for both technicians.

Suppose that we can assume that the technicians' sets of measurements are independent random samples from normal populations. The sample results are $s_1 = 38.6$ on the basis of the sample $n_1 = 11$ for technician number 1, and $s_2 = 21.7$ on the basis of the sample $n_2 = 9$ for technician number 2.

To proceed further we need a statistic based on the two values of s_i and n_i that is distributed according to an analytically tractable distribution. It turns out that the ratio of two chi-square variables, each divided by their

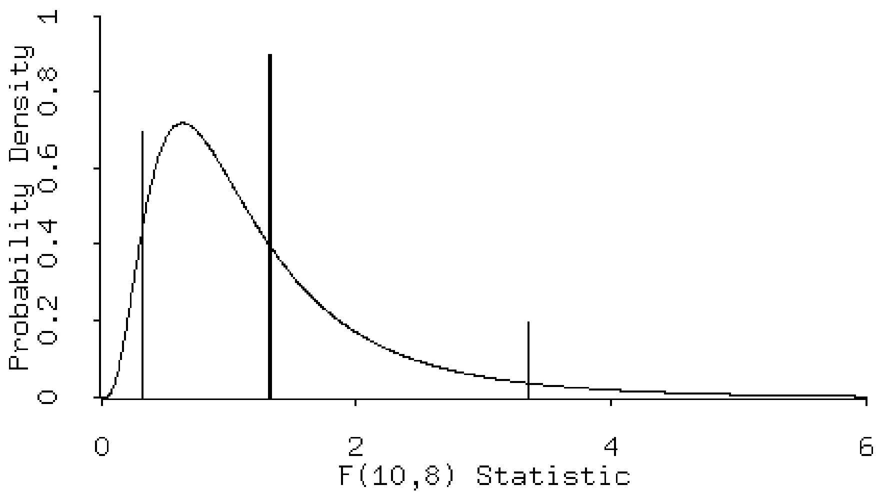

Figure 7.2: An F-distribution with 10 degrees of freedom in the numerator and 8 degrees of freedom in the denominator. The thick vertical line shows the mean and the thin vertical lines the critical values for $\alpha = .90$.

respective degrees of freedom, is distributed according to the *F-distribution*. In particular

$$\frac{\chi^2(v_1)/v_1}{\chi^2(v_2)/v_2} \;=\; F(v_1, v_2) \tag{7.6}$$

is distributed according to the F-distribution with parameters v_1 and v_2, which are the degrees of freedom of the respective chi-square distributions—v_1 is referred to as the degrees of freedom in the numerator and v_2 is the degrees of freedom in the denominator. The mean and variance of the F-distribution are

$$E\{F(v_1, v_2)\} = \frac{v_2}{(v_2 - 2)}$$

when $v_2 > 2$, and

$$\sigma^2\{F(v_1, v_2)\} = \frac{2\, v_2^2\, (v_1 + v_2 - 2)}{v_1\, (v_2 - 2)^2\, (v_2 - 4)}$$

when $v_2 > 4$. The probability density function for an F-distribution with 10 degrees of freedom in the numerator and 8 degrees of freedom in the denominator is plotted in Figure 7.2. The thick vertical line gives the mean and the two thin vertical lines give the critical values for $\alpha = .90$. The percentiles for this distribution can be found in the F-tables at the back of any textbook in statistics.[2] These tables give only the percentiles above 50 percent. To obtain the percentiles below 50 percent we must utilize the fact that the lower tail for the F-value

$$\frac{\chi^2(v_1)/v_1}{\chi^2(v_2)/v_2} = F(v_1, v_2)$$

is the same as the upper tail for the F-value

$$\frac{\chi^2(v_2)/v_2}{\chi^2(v_1)/v_1} = F(v_2, v_1).$$

This implies that

$$F(\alpha/2; v_1, v_2) = \frac{1}{F(1-\alpha/2; v_2, v_1)}.$$

Equation (7.5) can be written more generally as

$$\frac{v\,s^2}{\sigma^2} \quad = \quad \chi^2(v) \tag{7.7}$$

which implies that

$$\frac{s^2}{\sigma^2} = \frac{\chi^2(v)}{v}.$$

This expression can be substituted appropriately into the numerator and denominator of equation (7.6) to yield

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \quad = \quad F(v_1, v_2) \quad = \quad F(n_1 - 1, n_2 - 1). \tag{7.8}$$

To establish confidence intervals for the technician problem, we can manipulate (7.8) to yield

$$\begin{aligned} \frac{\sigma_2^2}{\sigma_1^2} \quad &= \quad F(n_1 - 1, n_2 - 1)\,\frac{s_2^2}{s_1^2} \quad = \quad F(10, 8)\,\frac{s_2^2}{s_1^2} \\ = \quad \frac{21.7^2}{38.6^2}\,F(10,8) \quad &= \quad \frac{470.89}{1489.96}\,F(10,8) \quad = \quad .31604\,F(10,8). \end{aligned} \tag{7.9}$$

[2] Or calculated using XlispStat or another statistical computer program.

To calculate a 90 percent confidence interval we find the values of $F(10, 8)$ at $\alpha/2 = .05$ and $1 - \alpha/2 = .95$. These are

$$F(.95; 10, 8) = 3.35$$

and

$$F(.05; 10, 8) = \frac{1}{F(.95; 8, 10)} = \frac{1}{3.07} = .3257$$

and are indicated by the thin vertical lines in Figure 7.2. The confidence intervals are thus

$$L = (.3257)(.31604) = .1029$$

and

$$U - (3.35)(.31604) = 1.057$$

so that

$$.1029 \leq \frac{\sigma_2^2}{\sigma_1^2} \leq 1.057.$$

Note that this confidence interval is based on the assumption that the two populations of measurements from which the sample variances are obtained are normally distributed or approximately so.

Since the above confidence interval straddles 1.0, it is clear that there is no indication that the variance of the measurements made by one technician exceeds the variance of the measurements made by the other. Nevertheless, we can test the hypothesis that the variances of the measurements of the two technicians are the same. The null and alternative hypotheses are

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

and

$$H_1\colon\ \sigma_1^2 \neq \sigma_2^2.$$

We want to control the α-risk at 0.1 when $\sigma_1^2 = \sigma_2^2$. Imposing the equal variance assumption on (7.8) we can extract the relationship

$$\frac{s_1^2}{s_2^2} = F(n_1 - 1, n_2 - 1).$$

The statistic on the left of the equality,

$$\frac{s_1^2}{s_2^2} = \frac{38.6^2}{21.7^2} = \frac{1489.96}{470.29} = 3.168$$

is thus distributed as $F(10, 8)$ and is greater than unity. We therefore need only look at the upper critical value $F(.95; 10, 8) = 3.35$ to see that the statistic falls in the acceptance region. We cannot reject the null hypothesis that the variances of the measurements of the two technicians are the same. When performing this test it is always easiest to manipulate the expression to put the largest variance in the numerator and thereby ensure that the sample statistic is bigger than unity. The decision to accept or reject the null hypothesis can then be made on the basis of the easy-to-calculate rejection region in the upper tail of the F-distribution.

7.3 Chi-Square Tests of Goodness of Fit

Statistical tests frequently require that the underlying populations be distributed in accordance with a particular distribution. Our tests of the equality of variances above required, for example, that both the populations involved be normally distributed. Indeed, any tests involving the chi-square or F-distributions require normally distributed populations. A rough way to determine whether a particular population is normally distributed is to examine the frequency distribution of a large sample from the population to see if it has the characteristic shape of a normal distribution. A more precise determination can be made by using a chi-square test on a sample from the population.

Consider, for example, the reported average daily per patient costs for a random sample of 50 hospitals in a particular jurisdiction. These costs were

257	274	319	282	253
315	313	368	306	230
327	267	318	326	255
392	312	265	249	276
318	272	235	241	309
305	254	271	287	258
342	257	252	282	267
308	245	252	318	331
384	276	341	289	249
309	286	268	335	278

with sample statistics $\bar{X} = 290.46$ and $s = 38.21$. We want to test whether the reported average daily costs are normally distributed, controlling the α-risk at .01. The null hypothesis is that they are normally distributed.

The chi-square test is based on a comparison of the sample data with the expected outcome if H_0 is really true. If the hypothesized distribution of the population was a discrete one, we could calculate the probability that each population value X_i will occur and compare that probability with the relative frequency of the population value in the sample. Since the normal distribution is a continuous one, however, the probability that any particular value X_i will occur is zero. So we must compare the probabilities that the X_i could lie in particular intervals with the frequency with which the sample values fall in those intervals.

The standard procedure is to select the intervals or classes to have equal probabilities so that the expected frequencies in all classes will be equal. Also, it is considered desirable to have as many classes as possible consistent with the expected frequencies in the classes being no less than 5. In the above example we therefore need $50/5 = 10$ classes.

To obtain the class intervals we find the values of z in the table of standardised normal values which divide the unit probability weight into 10 equal portions. These will be the z-values for which the cumulative probability density is respectively .1, .2, .3, .4, .5, .6, .7, .8, and .9. The values of X that fall on these dividing lines are thus obtained from the relationship

$$z = \frac{X - \bar{X}}{s}$$

which can be rearranged as

$$X = s\,z + \bar{X} = 38.21\,z + 290.46.$$

This gives us the intervals of z and X in the second and third columns of the table below.

i	z	X	f_i	F_i	$(f_i - F_i)^2$	$(f_i - F_i)^2/F_i$
1	$-\infty$ to -1.28	< 242	3	5	4	0.80
2	-1.28 to -0.84	242 to 258	11	5	36	7.20
3	-0.84 to -0.52	259 to 270	4	5	1	0.20
4	-0.52 to -0.25	271 to 280	6	5	1	0.20
5	-0.25 to -0.00	281 to 290	5	5	0	0.00
6	-0.00 to 0.25	291 to 300	0	5	25	5.00
7	0.25 to 0.52	301 to 310	5	5	0	0.00
8	0.52 to 0.84	311 to 322	7	5	4	0.80
9	0.84 to 1.28	323 to 339	4	5	1	0.20
10	1.28 to ∞	> 339	5	5	0	0.00
		Total	50	50		14.40

Column four gives the actual frequencies with which the sample observations fall in the ith category and column five gives the theoretically expected frequencies. The remaining two columns give the squared differences between the actual and expected frequencies and those squared differences as proportions of the expected frequencies. It turns out that the sum of the right-most column is distributed as χ^2 with 7 degrees of freedom. In general, when there are k classes with equal expected frequencies F_i in all classes and observed frequencies f_i in the ith class,

$$\sum_{k=1}^{k} \frac{(f_i - F_i)^2}{F_i}$$

is distributed as $\chi^2(k-m-1)$ where m is the number of parameters estimated from the sample data. As noted earlier in the definition of the chi-square distribution, the expression $(k-m-1)$ is the number of degrees of freedom. The 10 squared relative deviations give us potentially 10 degrees of freedom, but we have to subtract $m = 2$ because two parameters, $\bar{X}$ and s were estimated from the data, and a further degree of freedom because once we know the frequencies in nine of the ten classes above we can calculate the tenth frequency so only nine of the classes are independent. This leaves us with 7 degrees of freedom.

If the fit were perfect—i.e., the average daily per patient hospital costs were normally distributed—the total at the bottom of the right-most column in the table above would be zero. All the observed frequencies would equal their expected values—i.e., five of the sample elements would fall in each of the 10 classes. Clearly, the greater the deviations of the actual frequencies from expected, the bigger will be the test statistic. The question is then whether the value of the test statistic, 14.4, is large enough to have probability of less than 1% of occurring on the basis of sampling error if the true relative frequencies in the population equal the expected relative frequencies when the population is normally distributed. The critical value for $\chi^2(.99; 7)$ is 18.48, which is substantially above 14.4, so we cannot reject the null hypothesis that the population from which the sample was chosen is normally distributed.

It is interesting to note that the residuals indicate very substantial deviations from normality in two of the classes, 242–258 and 291–300 with the squared deviations from expected frequencies being 36 in the first of these classes and 25 in the second. We might be wise to examine more detailed data for certain of the hospitals to determine whether any reasons for deviations of these two specific magnitudes can be uncovered before we dismiss these observations as the result of sampling error. Finally, we should keep in mind that in the above test there is only a 1 percent chance that we would reject normality on the basis of sampling error alone if the population is in fact normal. This means that there is up to a 99 percent probability that we will accept normality if the population deviates from it—the β-risk is very high and the power of test is low for small departures from normality. Since it is usually crucial to our research conclusions that the population be normal, the more serious risk would appear to be the risk of accepting normality when it is not true rather than the risk of rejecting normality when it is true. One would like to make H_0 the hypothesis of non-normality and see if the data will lead us to reject it. Unfortunately, this is not possible because there are infinitely many ways to characterize a situation of non-normality. This suggests the importance of using large samples to make these inferences.

7.4 One-Dimensional Count Data: The Multinomial Distribution

Consider a manufacturer of toothpaste who wants to compare the marketability of its own brand as compared to the two leading competitors, A and B. The firm does a survey of the brand preferences of a random sample of 150 consumers, asking them which of the three brands they prefer. The results are presented in the table below.

Brand A	Brand B	Firm's Own Brand
61	53	36

The firm wants to know whether these data indicate that the population of all consumers have a preference for a particular brand.

Notice that the binomial distribution would provide the proper basis for the statistical analysis required here had the question stated "Do you prefer the firm's own brand to its competitors? Yes or No?" Each person's answer—i.e., each random trial—will yield an outcome $X_i = 1$ if the answer is 'yes' and $X_i = 0$ if it is 'no'. If the answer of each consumer in the survey is independent of the answer of all others, and if the probability that the answer of a random person picked will be 'yes' is the same for any person picked at random, then the total number of 'yes' answers,

$$X = \sum_{i=1}^{n} X_i$$

will be distributed as a binomial distribution with parameters p and n. The parameter p is the unknown probability that a person picked at random from the population will say 'yes'.

In the actual example above, the consumer surveyed has to choose between not two options (which would be a simple yes/no comparison) but three—she can prefer either brand A, brand B, or the firm's brand. Each random trial has 3 outcomes instead of 2. There are now three probabilities, p_1, p_2 and p_3, the probabilities of selecting A, B, or the firm's own brand, which must sum to unity. And the firm is interested in the counts n_1, n_2 and n_3 of consumers preferring the respective brands. This experiment is a *multinomial experiment* with k, the number of possible outcomes of each trial, equal to 3. The probabilities of observing various counts n_1, n_2 and n_3, given p_1, p_2 and p_3, is a *multinomial probability distribution.* In the case at hand, p_1, p_2 and p_3 are unknown and we want to make an inference about them on the basis of a sample n. The observed counts will be

$$n_1 + n_2 + n_3 = n.$$

To decide whether the population of consumers prefers a particular brand, we set up the null hypothesis of no preference and see if the data will prompt us to reject it. The null hypothesis is thus

$$H_0\colon\ p_1 = p_2 = p_3.$$

If the null-hypothesis is true we would expect an equal number of the sampled consumers to choose each brand—that is

$$E\{n_1\} = E\{n_2\} = E\{n_3\} = \frac{n}{3} = 50.$$

Notice the similarity of the problem here to the test of normality above. We have three classes each with an expected frequency of 50 and an actual frequency that differs from 50.

i	f_i	F_i	$(f_i - F_i)^2$	$(f_i - F_i)^2/F_i$
A	61	50	121	2.42
B	53	50	9	.18
Own Brand	36	50	196	3.92
Total	150	150		6.52

As in the normality test would expect

$$\sum_{k=1}^{k} \frac{(f_i - F_i)^2}{F_i}$$

to be distributed as $\chi^2(k-m-1)$. The number of classes here is $k = 3$, and no parameters were estimated from the sample data so $m = 0$. The statistic is thus distributed as $\chi^2(3-1) = \chi^2(2)$. From the chi-square table at the back of any textbook in statistics the critical value for $\chi^2(2)$ for $(\alpha = .05)$ will be found to be 5.99147. Since the total in the right-most column in the table above is 6.52, we can reject the null hypothesis of no brand preference when the α-risk is controlled at .05. The P-value of the statistic is .038. Does this imply a positive or negative preference for the firm's brand of toothpaste as compared to brands A and B? We want now to test whether consumers' preferences for the firm's own brand are greater or less than their preferences for brands A and B. This problem is a binomial one—consumers either prefer the firm's brand or they don't.

We can now use the techniques presented earlier—using a normal approximation to the binomial distribution—to set up a confidence interval for the proportion of the population of consumers choosing the firm's brand of toothpaste. Our sample estimate of p, now the proportion preferring the firm's brand, is

$$\bar{p} = \frac{36}{150} = .24.$$

Using the results in section 9 of Chapter 4, the standard deviation of $\bar{p}$ is

$$s_{\bar{p}} = \sqrt{\frac{\bar{p}\,(1-\bar{p})}{n-1}} = \sqrt{\frac{(.24)(.76)}{149}} = \sqrt{.00122316} = .03497.$$

The 95 percent confidence interval for p is thus (using the critical value $z = 1.96$ from the normal distribution table)

$$.24 \pm (1.96)(.03497) = .24 \pm .0685412,$$

or

$$.17 \leq p \leq .30984.$$

It is clear from this confidence interval that less than 1/3 of consumers prefer the firm's own brand of toothpaste, contrary to what one would expect under the null hypothesis of no differences in consumer preference. Indeed, we can test the hypothesis of no difference in preference between the firm's brand of toothpaste and other brands by setting up the null and alternative hypotheses

$$H_0\colon\ p = \frac{1}{3}$$

and

$$H_1\colon\ p \neq \frac{1}{3}$$

and calculating

$$z^* = \frac{\bar{p} - p}{s_p}$$

where

$$s_p = \sqrt{\frac{p\,(1-p)}{n}} = \sqrt{\frac{(\frac{1}{3})(\frac{2}{3}}{150}} = \sqrt{.00148148148} = .03849.$$

Notice that we use the value of p under the null hypothesis here instead of $\bar{p}$. Thus we have

$$z^* = \frac{\bar{p} - p}{s_p} = \frac{.24 - .333}{.03849} = \frac{.09333}{.03849} = 2.42.$$

The critical value of z for a two-tailed test with $\alpha = .05$ is 1.96. Clearly, we are led to reject the null hypothesis of no difference in preference. Indeed we could reject the null hypotheses that the population of consumers prefers the firm's brand to other brands with an α-risk of less than .01 because the P-value for a one-tailed test is .00776.

7.5 Contingency Tables: Tests of Independence

In the multinomial distribution above the data were classified according to a single criterion—the preferences of consumers for the three brands of toothpaste. Now we turn to multinomial distributions involving data that are classified according to more than one criterion.

Consider, for example, an economist who wants to determine if there is a relationship between occupations of fathers and the occupations of their sons. She interviewed 500 males selected at random to determine their occupation and the occupation of their fathers. Occupations were divided into four classes: professional/business, skilled, unskilled, and farmer. The data are tabulated as follows:

		Occupation of Son				
		Prof/Bus	Skilled	Unskilled	Farmer	Total
	Prof/Bus	55	38	7	0	100
Occupation of Father	Skilled	79	71	25	0	175
	Unskilled	22	75	38	10	145
	Farmer	15	23	10	32	80
	Total	171	207	80	42	500

The problem the economist faces is to determine if this evidence supports the hypothesis that sons' occupations are related to their fathers'. We can visualize there being a joint probability density function over all father-occupation, son-occupation pairs giving the probability that each combination of father and son occupations will occur. Treated as a table of probabilities, the above table would appear as

		Occupation of Son				
		Prof/Bus	Skilled	Unskilled	Farmer	Total
	Prof/Bus	p_{11}	p_{12}	p_{13}	p_{14}	p_{r1}
Occupation of Father	Skilled	p_{21}	p_{22}	p_{23}	p_{24}	p_{r2}
	Unskilled	p_{31}	p_{32}	p_{33}	p_{34}	p_{r3}
	Farmer	p_{41}	p_{42}	p_{43}	p_{44}	p_{r4}
	Total	p_{c1}	p_{c2}	p_{c3}	p_{c4}	1.00

where the probabilities along the right-most column p_{ri} are the marginal probabilities of fathers' occupations—i.e., the sum of the joint probabilities p_{ij} in the ith row and jth column over the j columns—and the probabilities along the bottom row p_{cj} are the marginal probabilities of sons' occupations—i.e., the sum of the joint probabilities p_{ij} over the i rows.

The count data, since they indicate the frequencies for each cell, can be thought of as providing point estimates of these probabilities. The marginal probabilities along the bottom row and the right-most column are the cell entries divided by 500 as shown in the table below. We know from the definition of statistical independence that if events A and B are independent,

$$P(A|B) = P(A)$$

which implies that

$$P(A \cap B) = P(A|B)P(B) = P(A)P(B).$$

Hence the joint probabilities in each cell of the table below should equal the product of the marginal probabilities for that particular row and column. The joint probabilities under the null hypothesis that fathers' occupations and sons' occupations are independent are as given below.

		Occupation of Son				
		Prof/Bus	Skilled	Unskilled	Farmer	Total
Occupation of Father	Prof/Bus	.0684	.0828	.0320	.0168	.20
	Skilled	.1197	.1449	.0560	.0294	.35
	Unskilled	.0992	.1201	.0464	.0244	.29
	Farmer	.0547	.0662	.0256	.0134	.16
	Total	.342	.414	.16	.084	1.00

This means that if the occupations of sons were independent of the occupations of their fathers the number or frequency of fathers and sons who were both in the professional or business category would be the joint probability of this outcome (.0684) times the number of sons sampled (500). Accordingly, we can calculate the expected number or expected count in each cell by multiplying the joint probability for that cell by 500. This yields the following table of actual and expected outcomes, with the expected outcomes in brackets below the actual outcomes.

		Occupation of Son				
		Prof/Bus	Skilled	Unskilled	Farmer	Total
	Prof/Bus	55	38	7	0	100
		(34.2)	(41.4)	(16.0)	(8.4)	
Occupation	Skilled	79	71	25	0	175
of		(59.85)	(72.45)	(28.0)	(14.7)	
Father	Unskilled	22	75	38	10	145
		(49.6)	(60.05)	(23.2)	(12.2)	
	Farmer	15	23	10	32	80
		(27.34)	(33.10)	(12.8)	(6.7)	
	Total	171	207	80	42	500

From this point the procedure is the same as in the test of normality. The tabulation, working from left to right, row by row, is as follows:

Father–Son	f_i	F_i	$(f_i - F_i)^2$	$(f_i - F_i)^2/F_i$
Prof/Bus–Prof/Bus	55	34.20	432.64	12.65
Prof/Bus–Skilled	38	41.40	11.56	0.28
Prof/Bus–Unskilled	7	16.00	81.00	5.06
Prof/Bus–Farmer	0	8.40	70.56	8.40
Skilled–Prof/Bus	79	59.85	366.72	6.13
Skilled–Skilled	71	72.45	2.10	0.03
Skilled–Unskilled	25	28.00	9.00	0.32
Skilled–Farmer	0	14.70	216.09	14.70
Unskilled–Prof/Bus	22	49.60	761.76	15.35
Unskilled–Skilled	75	60.05	223.50	3.72
Unskilled–Unskilled	38	23.20	219.04	9.44
Unskilled–Farmer	10	12.20	4.84	0.40
Farmer–Prof/Bus	15	27.34	152.28	5.57
Farmer–Skilled	23	33.10	102.01	3.08
Farmer–Unskilled	10	12.80	7.84	0.61
Farmer–Farmer	32	6.70	640.09	95.54
Total	500	500.00		181.28

It turns out that the total sum of squared relative deviations from expected values, represented by the number 181.28 at the bottom of the right-most column,

$$\sum_{k=1}^{k} \frac{(f_i - F_i)^2}{F_i},$$

Table 7.1: Percentage of Sons' Occupations by Father's Occupation

		Father's Occupation				
		Prof/Bus	Skilled	Unskilled	Farmer	Total
Son's Occupation	Prof/Bus	55	45	15	19	34
	Skilled	38	41	52	29	42
	Unskilled	7	14	26	12	16
	Farmer	0	0	7	40	8
	Total	100	100	100	100	100

is distributed according to a chi-square distribution with degrees of freedom equal to the product of the number of rows minus one and the number of columns minus one—i.e., $\chi^2((nr-1)(nc-1))$, where nr and nc are, respectively, the number of rows and columns in the contingency table. In the case at hand, the total is distributed as $\chi^2(9)$. The critical value for α-risk = .01 from the chi-square table for 9 degrees of freedom is 21.6660. Since the total in the right-most column of the table vastly exceeds that critical value, we must reject the hypothesis of independence and conclude that sons' occupations depend on the occupations of their fathers.

The pattern of dependence can be seen more clearly when we take the percentage of sons in each category of father's occupation and compare them with the overall percentage of sons in each occupation. This is done in the table immediately above. Each column of the table gives the percentage of sons of fathers in the occupation indicated at the top of that column who are in the various occupations listed along the left margin of the table. The right-most column gives the percentage of all sons in the respective occupations.

If sons' occupations were independent of their fathers', 34 percent of the sons in each father's-occupation category would be in professional/business occupations. As can be seen from the table, 55 percent of the sons of professional/business fathers and 45 percent of the sons of fathers in skilled trades are in professional/business occupations. Yet only 15 and 19 percent, respectively, of sons of unskilled and farmer fathers work in the professions and business. If sons' occupations were unrelated to their fathers' occupations, 42 percent would be in skilled occupations, regardless of the occupation of the father. It turns out from the table that 52 percent of sons of unskilled fathers are in skilled trades and less than 42 percent of the sons of fathers in each of the other categories are skilled workers. Judging from this and from the 45 percent of sons of skilled fathers who are in professional/business occupations, it would seem that the sons of skilled fathers tend either to move up into the business/professional category or fall back into the unskilled category, although the percentage of sons of skilled fathers who are also skilled is only slightly below the percentage of all sons who are skilled. If there were no occupational dependence between fathers and sons, 16 percent of sons of unskilled fathers would also be in unskilled work. As we can see from the table, 26 percent of the sons of unskilled workers are themselves unskilled and less than 16 percent of the sons of unskilled fathers are in each of the other three occupational categories. Finally, if the occupations of fathers and their sons were statistically independent we would expect that 8 percent of the sons of farmers would be in each occupational category. In fact, 40 percent of the sons of farmers are farmers, 7 percent of the sons of unskilled fathers are farmers, and none of the sons of fathers in the skilled and professional/business occupations are farmers.

The dependence of son's occupation on father's occupation can also be seen from the table by drawing a wide diagonal band across the table from top left to bottom right. The frequencies tend to be higher in this diagonal band than outside it, although there are exceptions. This indicates that sons' occupations tend to be the same or similar to their fathers'. Sons' occupations and the occupations of their fathers are statistically dependent.

7.6 Exercises

1. Independent random samples were selected from each of two normally distributed populations. The sample results are summarized as follows:

Sample 1	Sample 2
$n_1 = 10$	$n_2 = 23$
$\bar{X}_1 = 31.7$	$\bar{X}_2 = 37.4$
$s_1^2 = 3.06$	$s_2^2 = 7.60$

Setting the α-risk at 0.05, test the null hypothesis H_0: $\sigma_1^2 = \sigma_2^2$ against the alternative hypothesis H_1: $\sigma_1^2 \neq \sigma_2^2$,

2. A financial analyst is exploring the relationship between the return earned by stocks and the return earned by bonds. For a period of $n = 25$ months, the analyst records the return on a particular stock, denoted X, and the return on a particular bond, denoted Y. The relevant sample statistics are recorded below:

Monthly Returns	Stock (X)	Bond (Y)
Mean	1.5	1.2
Standard Deviation	1.0	0.8

Assume that X and Y are uncorrelated and perform hypotheses tests to determine whether the two population variances are equal. Then perform a test to determine whether the two population means are equal. How would your answer change if it turned out that the sample correlation between X and Y was $r_{xy} = -0.20$.

3. A labour economist studied the durations of the most recent strikes in the vehicles and construction industries to see whether strikes in the two industries are equally difficult to settle. To achieve approximate normality and equal variances, the economist worked with the logarithms (to the base 10) of the duration data (expressed in days). In the vehicle industry there were 13 strikes having a mean log-duration of 0.593 and a standard deviation of log-duration of 0.294. In the construction industry there were 15 strikes with a mean log-duration was 0.973 and a standard deviation of log-duration of 0.349. The economist believes that it is reasonable to treat the data as constituting independent random samples.

a) Construct and interpret a 90 percent confidence interval for the difference in the mean log-durations of strikes in the two industries.

b) Test whether the strikes in the two industries have the same log-durations, controlling the α risk at 0.10. State the alternatives, the decision rule, the value of the test statistic and the conclusion.

c) Test the economist's assumption that the log-durations of strikes in the two industries have the same variance controlling the α risk at 0.10. State the alternatives, the decision rule, the value of the test statistic and the conclusion.

4. An industrial machine has a 1.5-meter hydraulic hose that ruptures occasionally. The manufacturer has recorded the location of these ruptures for 25 ruptured hoses. These locations, measured in meters from the pump end of the hose, are as follows:

1.32	1.07	1.37	1.19	0.13
1.14	1.21	1.16	1.43	0.97
0.33	1.36	0.64	1.42	1.12
1.46	1.27	0.27	0.80	0.08
1.46	1.37	0.75	0.38	1.22

Using the chi-square procedure, test whether the probability distribution of the rupture locations is uniform with lowest value $a = 0$ and highest value $b = 1.5$.

5. A city expressway utilizing four lanes in each direction was studied to see whether drivers prefer to drive on the inside lanes. A total of 1000 automobiles was observed during the heavy early-morning traffic and their respective lanes recorded. The results were as follows:

Lane	Observed Count
1	294
2	276
3	238
4	192

Do these data present sufficient evidence to indicate that some lanes are preferred over others? Use $\alpha = .05$ in your test.

6. It has been estimated that employee absenteeism costs North American companies more than \$100 billion per year. As a first step in addressing the rising cost of absenteeism, the personnel department of a large corporation recorded the weekdays during which individuals in a sample of 362 absentees were away from work over the past several months:

	Number Absent
Monday	87
Tuesday	62
Wednesday	71
Thursday	68
Friday	74

Do these data suggest that absenteeism is higher on some days of the week than others?

7. The trustee of a company's pension plan has solicited the opinions of a sample of the company's employees about a proposed revision of the plan. A breakdown of the responses is shown in the accompanying table. Is there evidence at the 10% level to infer that the responses differ among the three groups of employees?

	Blue-Collar	White Collar	Managers
Responses	Workers	Workers	
For	67	32	11
Against	63	18	9

8. A study of the amount of violence viewed on television as it relates to the age of the viewer showed the accompanying results for 81 people. Each person in the study could be classified according to viewing habits as a low-violence or high-violence viewer.

	16–34 yrs. old	35–54 yrs. old	55 yrs. and over
Low Violence	8	12	21
High Violence	18	15	7

Do the data indicate that viewing of violence is not independent of age of viewer at the 5% significance level?

9. To see if there was any dependency between the type of professional job held and one's religious affiliation, a random sample of 638 individuals belonging to a national organization of doctors, lawyers and engineers were chosen in a 1968 study. The results were as follows:

	Doctors	Lawyers	Engineers
Protestant	64	110	152
Catholic	60	86	78
Jewish	57	21	10

Test at the 5 percent level of significance the hypothesis that the profession of individuals in this organization and their religious affiliation are independent. Repeat at the 1 percent level.

10. To study the effect of fluoridated water supplies on tooth decay, two communities of roughly the same socio-economic status were chosen. One of these communities had fluoridated water while the other did not. Random samples of 200 teenagers from both communities were chosen and the numbers of cavities they had were determined. The results were as follows:

Cavities	Fluoridated Town	Nonfluoridated Town
0	154	133
1	20	18
2	14	21
3 or more	12	28

Do these data establish, at the 5 percent level of significance, that the number of dental cavities a person has is not independent of whether that person's water supply is fluoridated? What about at the 1% level?