CHAPTER 13

Visualizing Goodness-of-Fit Tests

CONCEPTS

- Theoretical Distribution, Parameter, Fitted Distribution, Goodness-of-Fit Test, Chi-square Test, ECDF plot, Kolmogorov-Smirnov test, Probability Plot

OBJECTIVES

- Recognize the characteristics of data-generating situations that will suggest an appropriate theoretical distribution to be fitted

- Interpret common data displays that may reveal whether a specified theoretical distribution is compatible with a given data set

- Learn how the chi-square goodness-of-fit test works and how the number of histogram classes affects the results of the test

- Recognize limitations of the chi-square goodness-of-fit test and the alternatives that are available

- Use visual and analytical ECDF-based tests for goodness-of-fit and compare their results with chi-square tests

Overview of Concepts

Goodness-of-fit tests (or **GOF** tests) have many important applications. For example, a quality control analyst may need to test whether defects are Poisson-distributed, or whether a finished product dimension is normally distributed. In exploratory data analysis (EDA) statisticians feel that sample data should "tell its own story." But in a GOF test *a priori* logic plays a strong role. First, we consider the data generating situation and choose a **theoretical distribution** (perhaps more than one) that is consistent with the data type (continuous or discrete), the sampling method, and the data's magnitude and range. Second, we examine sample statistics and visual displays (such as a histogram) to assess central tendency, dispersion, shape (skewness and kurtosis), and unusual features (e.g., outliers). Third, we may proceed to formal tests to see whether the chosen theoretical distribution provides an acceptable fit to the data.

In this module, you may specify the **parameters** of the proposed distribution *a priori*, or you may use the sample data to create a **fitted distribution**. Setting the parameters from the sample data will give a "better fit" but may violate the logic of the data generating situation. Purists oppose fitting the parameters from the data, arguing that a GOF test is inappropriate when there is no *a priori* basis for specifying the parameters. Yet sometimes logic dictates a distribution but not its parameter values.

This module superimposes your proposed distribution on the histogram, so you can see how it looks. This visual test may rule out the proposed distribution. The next step usually is a **chi-square test** based on the number of histogram classes (bins) you have chosen. In this module, the histogram covers the data range exactly, to avoid introducing unnecessary imprecision in the end classes. Bin width is the data range divided by the number of classes, so class limits will not be "nice." To maintain a "nice" axis scale, the histogram is allowed to "float" on top of the horizontal axis (the actual class limits are shown in the chi-square test).

The **empirical cumulative distribution function** (or **ECDF plot**) graphs the cumulative proportion of observations against the data values. The proposed theoretical distribution may be superimposed on the ECDF to check its fit. More formally, the **Kolmogorov-Smirnov test** (or **K-S test**) is based on the largest vertical difference between the ECDF and the proposed theoretical distribution. The K-S calculations may also be shown in tabular format. Critical values for the K-S test are found in published tables. The **probability plot** is a visual test that compares actual and expected data values under the proposed distribution. It will resemble a straight 45^o line if the proposed distribution gives a good fit to the data.

Some words of caution are in order. In the chi-square test, expected frequencies should not be small (Cochran's Rule says at least 5, but other rules exist). To enlarge the frequencies, classes can be collapsed from the ends (this module tries to get frequencies of at least 2) but with discrete data this may not suffice. You can let the computer choose classes to maximize the expected frequencies (though this leads to unequal class sizes that no longer correspond to the histogram). You can also treat discrete data as if they are continuous.

Some distributions have logical approximations. For example, a binomial will resemble a normal if its mean is large (e.g., if $np \geq 5$ and $nq \geq 5$) and similarly for a Poisson (e.g., if $\lambda \geq 5$). A Poisson approximation can be used for a binomial (e.g., if $n \geq 20$ and $p < 0.05$). An exponential can approximate a geometric with a large mean. You can explore these possibilities in this module.

Illustration of Concepts

From a grocer's shelf, 22 D'Anjou pears are chosen at random and weighed. We expect the weights to be normally distributed, but since we have no *a priori* idea of the mean (μ) or standard deviation (σ) we must estimate the **parameters** from the sample. Figures 1, 2, and 3 show histograms using 4, 5, and 10 classes, along with a normal **fitted distribution** (μ = 248.1 grams and σ = 9.471 grams). The 10-class histogram is rather sparse (its **chi-square test** would have small expected frequencies) so we will consider only the first two cases.

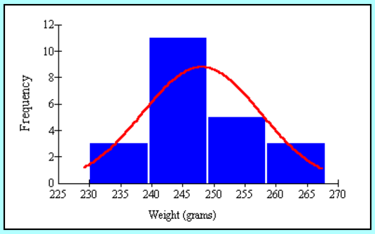


Figure 1: 4-Class Histogram

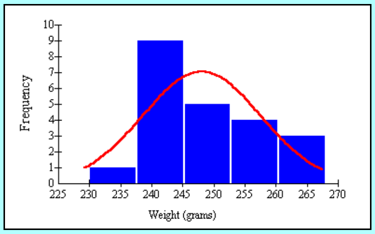


Figure 2: 5-Class Histogram

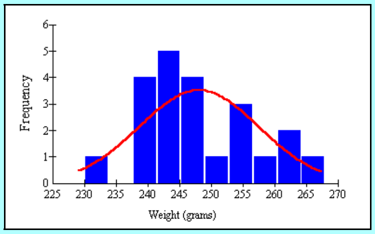
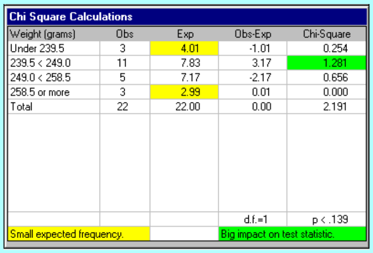


Figure 3: 10-Class Histogram

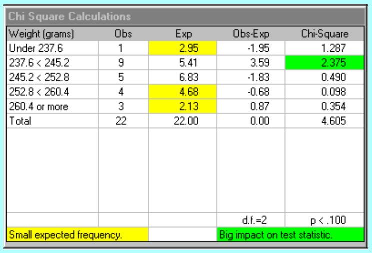
If the fit is good, the chi-square statistic will be near zero, using k–1–m degrees of freedom where m is the number of parameters estimated from the sample (m = 2 in this case) and k is the number of classes. In Figure 4 (for k = 4) the chi-square test statistic is 2.191 (p = 0.139), while Figure 5 (for k = 5) shows a chi-square test statistic of 4.605 (p = 0.100). The p-values indicate that we are unable to reject the hypothesis of normality at α = 0.05. In this example, the number of classes affects the test statistic and degrees of freedom, but has little impact on the p-value. Figure 6 shows that the second class contributes heavily to the chi-square statistic.



Chi Square Calculations

Weight (grams)	Obs	Exp	Obs-Exp	Chi-Square
Under 239.5	3	4.01	-1.01	0.254
239.5 < 249.0	11	7.83	3.17	1.281
249.0 < 258.5	5	7.17	-2.17	0.656
258.5 or more	3	2.99	0.01	0.000
Total	22	22.00	0.00	2.191
			d.f.=1	p < .139
Small expected frequency.			Big impact on test statistic.	

Figure 4: Chi-Square Test (k=4)



Chi Square Calculations

Weight (grams)	Obs	Exp	Obs-Exp	Chi-Square
Under 237.6	1	2.95	-1.95	1.287
237.6 < 245.2	9	5.41	3.59	2.375
245.2 < 252.8	5	6.83	-1.83	0.490
252.8 < 260.4	4	4.68	-0.68	0.098
260.4 or more	3	2.13	0.87	0.354
Total	22	22.00	0.00	4.605
			d.f.=2	p < .100
Small expected frequency.			Big impact on test statistic.	

Figure 5: Chi-Square Test (k=5)

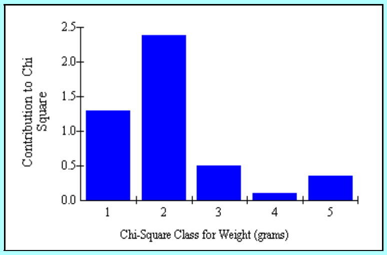
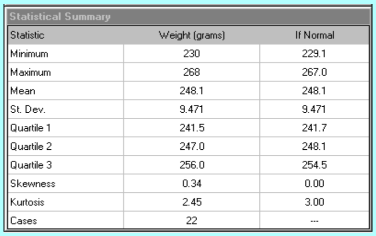


Figure 6: Chi-Square Cells (k = 5)

The descriptive statistics in Figure 7 show that the sample minimum and maximum (230 and 268) are almost exactly what we would expect in a sample of this size (229.1 and 267.0), calculated as the 0.5/n and 1-0.5/n fractiles of the **theoretical distribution**. The data are slightly right skewed (skewness 0.34) and platykurtic (kurtosis 2.45). The **ECDF plot** in Figure 8 resembles the fitted normal distribution, and the **K-S test** statistic of 0.14 shows no significant departure from normality. The **probability plot** In Figure 9 is close to a straight line. Overall, the normal distribution gives a reasonable fit in our **GOF** tests.



Statistical Summary

Statistic	Weight (grams)	If Normal
Minimum	230	229.1
Maximum	268	267.0
Mean	248.1	248.1
St. Dev.	9.471	9.471
Quartile 1	241.5	241.7
Quartile 2	247.0	248.1
Quartile 3	256.0	254.5
Skewness	0.34	0.00
Kurtosis	2.45	3.00
Cases	22	---

Figure 7: Statistical Comparison

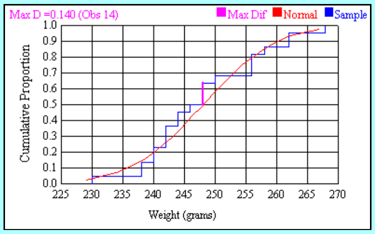


Figure 8: ECDF and K-S Test

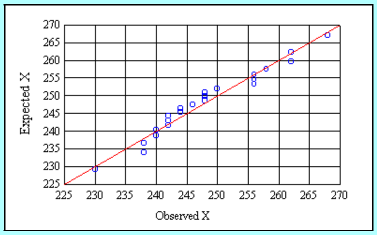


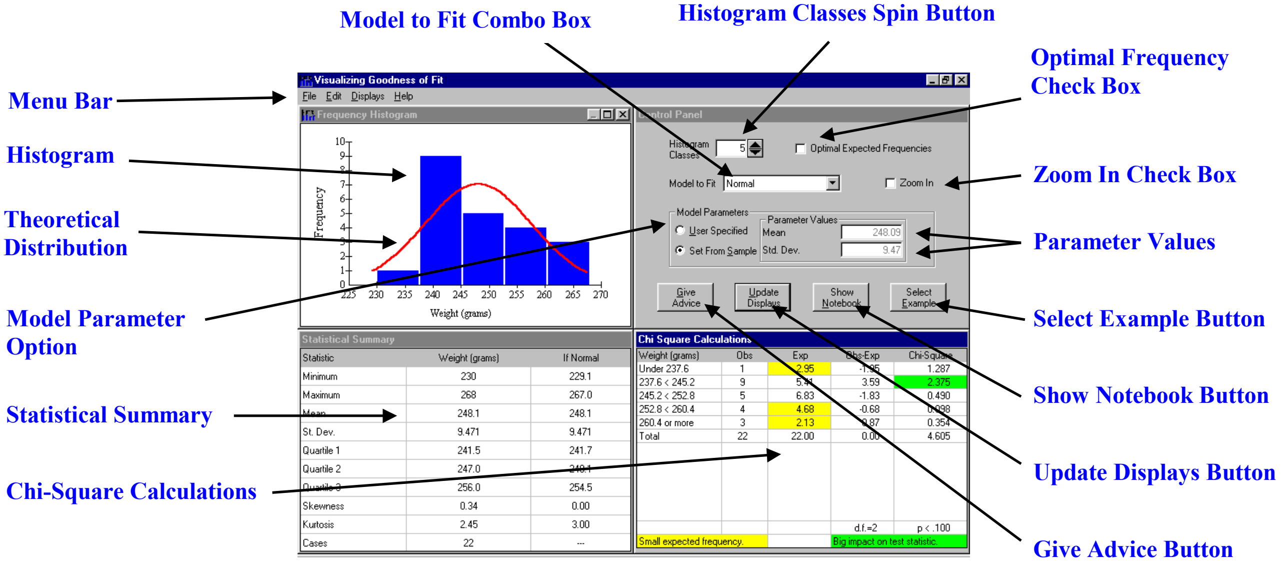
Figure 9: Probability Plot

Orientation to Basic Features

This module does goodness-of-fit tests. You can analyze a variety of different data sets by selecting them from the Notebook or create your own using the data editor.

1. Opening Screen

Start the module by clicking on the module's icon, title, or chapter number in the *Visual Statistics* menu and pressing the Run Module button. When the module is loaded, you will be on the introduction page of the Notebook. Click on the Introduction and Concepts tabs to see what will be covered in this module. Click on the Examples tab, select Class Projects, select Weight of D'Anjou Pears, and press OK. Read the Hint that appears in the middle of the display and press OK. The upper left shows a frequency histogram with a normal distribution (the default). The Control Panel appears on the right. On the bottom left is a statistical summary that compares the sample with the specified theoretical distribution. On the lower right is a table of calculations for the chi-square test. Other displays may be chosen from the menu bar at the top of the screen or by right-clicking a display and using the menu. The flashing Update Displays button will indicate when you have changed one or more control settings.



Histogram Classes 5 □ Optimal Expected Frequencies

Model to Fit Normal □ Zoom In

Model Parameters: ○ User Specified ● Set From Sample

Parameter Values: Mean 248.09 Std. Dev. 9.47

Statistical Summary

Statistic	Weight (grams)	If Normal
Minimum	230	229.1
Maximum	268	267.0
Mean	248.1	248.1
St. Dev.	9.471	9.471
Quartile 1	241.5	241.7
Quartile 2	247.0	248.1
Quartile 3	256.0	254.5
Skewness	0.34	0.00
Kurtosis	2.45	3.00
Cases	22	---

Chi Square Calculations

Weight (grams)	Obs	Exp	Obs-Exp	Chi-Square
Under 237.6	1	2.95	-1.95	1.287
237.6 < 245.2	9	5.41	3.59	2.375
245.2 < 252.8	5	6.83	-1.83	0.490
252.8 < 260.4	4	4.68	-0.68	0.098
260.4 or more	3	2.13	0.87	0.354
Total	22	22.00	0.00	4.605
			d.f.=2	p < .100
Small expected frequency.			Big impact on test statistic.	

2. Control Panel

a. The Histogram Classes spin button affects the number of histogram intervals and also the number of classes for the chi-square test. The number of classes is set initially using Sturges' Rule (k = 1 + $\log_2$ n). When too many classes are used, the chi-square test may generate a warning because the expected frequency in each class will become small.

b. When the Zoom In check box is chosen, the axis on the histogram is chosen to reflect the actual data range. By default, the histogram axis scale is based on the theoretical maximum and minimum (usually wider than the actual data range), though in some data sets the results will be the same. Watch the histogram scale as you select this option.

c. The Optimal Expected Frequencies check box forces the classes in the chi-square test to be chosen so that expected frequencies are maximized. For continuous data, this helps avoid small expected frequencies and may improve the power of the test. However, the unequal class sizes that arise may be less attractive. Note that this option is inactive for a discrete distribution since the “classes” are integers that cannot easily be converted to class limits. However, you can use this option for a discrete model if you also select the option to treat the data as continuous.

d. The Model Parameters default is to use Set From Sample to ensure a properly centered distribution. Choose the User Specified option if you want to specify the parameter values. If you make strange choices (e.g., specifying a distribution that is off the scale) you will receive a warning or, in extreme cases, be asked to choose different values.

e. The Model To Fit combo box lets you select another distribution. Select a Uniform distribution. If you specify an inappropriate distribution, you will be warned (or prevented if your choice would make computation impossible) and an explanation will be given.

f. Click the Give Advice button to see a commentary on your data. If you are selecting a model or setting its parameters, this button may also suggest options of interest.

g. Click the Select Example button to pick a new data set from the list of examples. This button changes to Select Databases or Edit Data depending upon the origin of the data you are analyzing. It acts as a shortcut to the Notebook tab previously selected.

h. Click the Show Notebook button to bring up the Notebook. There are two large databases that you can access with this module: U.S. States and World Nations. Select the Databases tab. Click on either U.S. States or World Nations. Each database is organized by categories. Click on the + symbol of any category to expand the category and list its variables. The + symbol will become a – symbol. Click on the – symbol to shrink a category and hide its variables. Within any category of interest, click on several variables and read their descriptions in the text window at the right. A complete discussion of the databases is given in the Introduction.

i. Click on the Show Notebook button to bring up the Notebook. You can use your own data by using the Data Editor. Select the Data Editor tab, and press OK. A simple two-column spreadsheet appears. You can enter data and corresponding labels directly into this spreadsheet. You can title each column by entering its label in the top row. You can copy data from another spreadsheet and paste it into the data editor. From File on the menu bar you can save the data in *Visual Statistics* format. When you are finished, click on File and choose Exit Editor and Use Data or Exit Editor and Discard Changes. A complete discussion of the Data Editor is given in the Introduction to this book.

3. Copying Graphs

To copy a display, select Copy from the Edit menu on the menu bar or the Copy option when you right click on a display. It can then be pasted into other applications, such as Word or WordPerfect. Graphs are copied as bitmaps, and tables as tab-delimited text.

4. Help

Click on Help on the menu bar at the top of the screen. Search for Help lets you search an index for this module, Contents shows a table of contents for this module, Using Help gives instructions on how to use Help, and About gives licensing and copyright information.

5. Exit

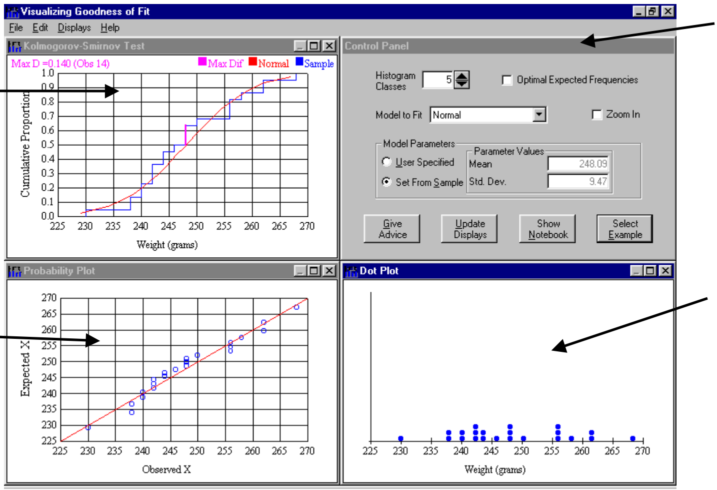
Close the module by selecting Exit in the File menu (or click ☒ in the upper right-hand corner of the window). You will be returned to the *Visual Statistics* main menu.

Orientation to Additional Features

This module offers seven different types of graphs, four types of tables, and a verbal interpretation. The Hint that was displayed as the module began said, “Click the right mouse button on a quadrant to select a different display.” This can also be done by selecting a quadrant and clicking on Displays in the menu bar. Only the Control Panel cannot be replaced.

1. Additional Graphs

Click on the upper left quadrant and choose Graphs and Kolmogorov-Smirnov Test. Click on the lower left quadrant and choose Graphs and Probability Plot. Click the lower left quadrant and choose Graphs and Dot Plot. The screen will look like the following:



Histogram Classes: 5
□ Optimal Expected Frequencies
Model to Fit: Normal
□ Zoom In
Model Parameters: ○ User Specified ● Set From Sample
Parameter Values: Mean 248.09, Std. Dev. 9.47

Control Panel

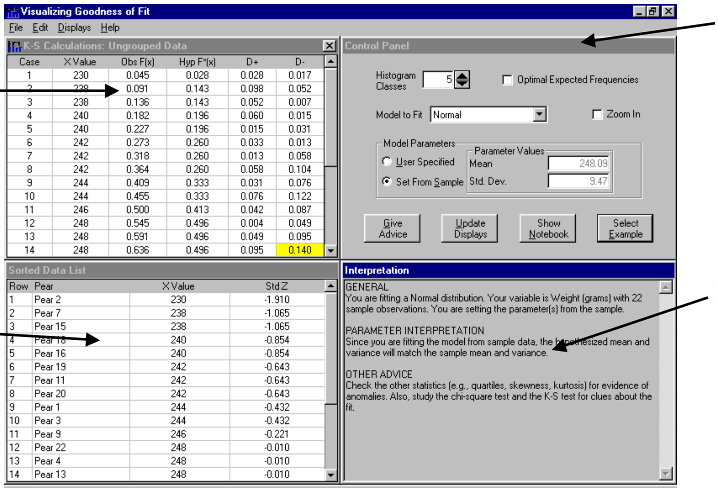
ECDF Plot and K-S Test

Probability Plot

Dot Plot

2. Additional Tables

Click on the upper left quadrant and choose Tables and K-S Calculations. Click on the lower left quadrant and choose Tables and Data List. Click the lower left quadrant and choose Interpretation. The screen will look like the following:



K-S Calculations: Ungrouped Data

Case	X Value	Obs F(x)	Hyp F*(x)	D+	D-
1	230	0.045	0.028	0.028	0.017
2	238	0.091	0.143	0.098	0.052
3	238	0.136	0.143	0.052	0.007
4	240	0.182	0.196	0.060	0.015
5	240	0.227	0.196	0.015	0.031
6	242	0.273	0.260	0.033	0.013
7	242	0.318	0.260	0.013	0.058
8	242	0.364	0.260	0.058	0.104
9	244	0.409	0.333	0.031	0.076
10	244	0.455	0.333	0.076	0.122
11	246	0.500	0.413	0.042	0.087
12	248	0.545	0.496	0.004	0.049
13	248	0.591	0.496	0.049	0.095
14	248	0.636	0.496	0.095	0.140

Sorted Data List

Row	Pear	X Value	Std Z
1	Pear 2	230	-1.910
2	Pear 7	238	-1.065
3	Pear 15	238	-1.065
4	Pear 18	240	-0.854
5	Pear 16	240	-0.854
6	Pear 19	242	-0.643
7	Pear 11	242	-0.643
8	Pear 20	242	-0.643
9	Pear 1	244	-0.432
10	Pear 3	244	-0.432
11	Pear 9	246	-0.221
12	Pear 22	248	-0.010
13	Pear 4	248	-0.010
14	Pear 13	248	-0.010

Histogram Classes: 5
□ Optimal Expected Frequencies
Model to Fit: Normal
□ Zoom In
Model Parameters: ○ User Specified ● Set From Sample
Parameter Values: Mean 248.09, Std. Dev. 9.47

Interpretation

GENERAL
You are fitting a Normal distribution. Your variable is Weight (grams) with 22 sample observations. You are setting the parameter(s) from the sample.

PARAMETER INTERPRETATION
Since you are fitting the model from sample data, the hypothesized mean and variance will match the sample mean and variance.

OTHER ADVICE
Check the other statistics (e.g., quartiles, skewness, kurtosis) for evidence of anomalies. Also, study the chi-square test and the K-S test for clues about the fit.

Control Panel

Kolmogorov-Smirnov Calculations

Data List

Interpretation

Basic Learning Exercises

Name ________________

Descriptive Statistics

Press the Show Notebook button and select the Examples tab. Click on Sports and select Kentucky Derby Winners. Read the description of the data set and click OK.

1. Give the exact definition of the variable. What are its units of measurement? Are the data continuous or discrete? Do you think the sample size is large enough to give insight into the shape of the distribution?

2. Click on OK. Click the right mouse button to select your displays. Right-click the display in the lower right quadrant and select Tables and Data List. From the sorted data list, which horse had the best winning time? The worst? What are their standardized z-values? Is either an outlier?

3. What is the value of each statistic shown below? Based on these measures, what can you say about centrality and skewness? If you are unfamiliar with any term, use Help.

   Mean _______ Median _______ 1st Quartile _______ 3rd Quartile _______

4. Record the value of each measure of shape. Click "Give Advice" to obtain the percent of observations within 1, 2, and 3 standard deviations. What do these statistics tell you?

	Skewness	Kurtosis	% within 1 SD	% within 2 SD	% within 2 SD
Sample					
If normal					

Histograms

5. Visually assess the histogram in comparison to the fitted normal distribution. Vary the number of classes (k) from 4 to 9. Does your impression of "fit" change as k changes?

Chi-Square Tests

6. Right-click the lower left display, select Tables, and choose Chi-Square Calculation. Consider the hypotheses

   H_0: Derby winning time follows a normal distribution
   H_1: Derby winning time does not follow a normal distribution

   Record the chi-square test statistic, degrees of freedom, and p-value as you vary the number of classes (k) from 4 to 9. For each test at $\alpha = 0.05$ state the conclusion about H_0 (accept, reject). What conclusion you would reach about H_0 overall?

	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9
Chi-square						
D.F.						
P-value						
Decision						

7. Repeat the previous exercise, but this time click Optimal Expected Frequencies. What is meant by "optimal expected frequencies"? **Hint**: Use Help. What conclusion you would reach overall? What was the effect of using equal expected frequencies on the apparent power of the test? **Hint**: Compare the p-values with those in Exercise 6.

	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9
Chi-square						
D.F.						
P-value						
Decision						

Intermediate Learning Exercises Name ________________

Chi-Square Tests

8. Press the Show Notebook button, select the Examples tab, click on Sports and select Kentucky Derby Winners. Read the description of the data set and click OK. Vary the number of classes from 4 to 9, noting the number of degrees of freedom in the chi-square test. Why did D.F. not always increase? **Hint**: How many end classes were combined?

9. Click the Optimal Expected Frequencies option. Why do you consistently get one extra D.F. each time you add a class? Contrast this with Exercise 8. **Hint**: Observe the expected frequencies and explain how they arise.

10. Does adding more classes lead to a better chi-square test (i.e., more powerful test)? Explain. **Hint**: Using optimal expected frequencies, keep adding classes (beyond 9) and see what happens to the expected frequencies.

11. Choose 7 classes and select Optimal Expected Frequencies. Right-click the display in the lower left and select Graphs and Chi-Square Test. Right-click the display in the upper left and select Graphs and Difference in Frequencies. What does the green highlighting in the chi-square table of calculations tell you? Relate it to the graphs.

12. Does the Chi-Square Test graph always agree with the Difference in Frequencies graph? **Hint**: Try varying k with and without Optimal Expected Frequencies.

Probability Plot

13. Right-click the display in the lower left and select Graphs and Probability Plot. What does this probability plot indicate? **Hint**: Use Help. Right-click the display in the lower right and select Tables and Data List. On the probability plot, locate the lowest data point in the data list (Secretariat in 1973). Estimate its coordinates on the graph and tell what the coordinates indicate. Repeat for the highest data point in the data list (Tiny Tam in 1958).

14. Right-click the display in the upper left and select Graphs and Frequency Histogram. Right-click the display in the lower right and select Tables and Chi-Square Calculation. Select Uniform Continuous on the control panel's Model to Fit. Use 7 classes. What does the chi-square test tell you? What does the probability plot tell you? Compare the two tests.

15. Again select Normal on the control panel's Model to Fit. Use 7 classes. Set Model Parameters to User Specified. Set the mean to 123.0 and the standard deviation to 1.0, and click Update Displays. What does the frequency histogram with overlaid normal distribution tell you? What does the chi-square test tell you? What does the probability plot tell you? Explain the problem that you have introduced.

Empirical Cumulative Distribution Function

16. Set Model Parameters to Set From Sample and click Update Displays. Right-click the display in the lower right and select Tables and Descriptive Statistics. Right-click the display in the upper left and select Graphs and Empirical Distribution Function. What does the ECDF tell you? Estimate the median from this graph, and check it against the descriptive statistics.

Advanced Learning Exercises Name ________________

Kolmogorov-Smirnov Test

Press the Show Notebook button and select the Examples tab. Click on Sports and select Kentucky Derby Winners. Read the description of the data set and click OK.

17. Right-click the display in the lower right and select Tables and Kolmogorov-Smirnov Calculation. Right-click the display in the lower left and select Graphs and Kolmogorov Smirnov Test. How does the K-S graph differ from the ECDF graph? Explain the meaning of the magenta line segment on the K-S graph, and relate it to the table of K-S calculations. Does this test support or disconfirm the hypothesis that the true distribution is normal?

18. Click Help and choose Search for Help On and Kolmogorov Smirnov and Equations. Using this information, try to calculate the critical value for $\alpha = 0.20$ and verify the conclusions of the K-S test. Was the decision close?

19. What does the help file say about the power of the Kolomogorov-Smirnov test? What does this mean?

20. Select Uniform Continuous on the control panel's Model to Fit. Use 7 classes. What does the K-S test tell you? Describe what the K-S graph shows.

Discrete Distributions

Press the Show Notebook button and select the Scenarios tab. Click on Binomial distribution and select True-false exam Scores. Read the scenario and click OK.

21. Right-click the display in the upper left and select Graphs and Frequency Histogram. Right-click the display in the lower left and select Graphs and Kolmogorov Smirnov Test. Right-click the display in the lower right and select Tables and Chi-Square Calculation. Do these tests support the hypothesis that the true distribution is binomial with p = 0.50 (i.e., the students were guessing)? Explain in your own words what the histogram suggests. How does the histogram appearance differ from the previous examples?

22. Leave the sample size at 50 but enter a probability of success somewhat greater than 0.50, and click Update Displays. Look at the histogram, K-S graph, and chi-square test. Use trial-and-error to vary the probability of success until you find a value that gives a good fit to the data, based on these visual displays. Explain in your own words what this experiment suggests.

23. Select Normal on the control panel's Model to Fit. Use 7 classes. Set Model Parameters to Set from Sample. Click Update Displays. Describe the results of this test. How does the appearance of the K-S graph and chi-square calculation differ from the previous discrete case?

Exponential Distribution

Press the Show Notebook button and select the Scenarios tab. Click on Exponential distribution and select Light bulb life. Read the scenario and click OK.

24. Is light bulb life exponentially distributed with a mean of 1,000 hours?

Individual Learning Projects

Write a report on one of the three topics listed below. Use the cut-and-paste facilities of the module to place the appropriate graphs and tables in your report.

1. Select a continuous variable from one of the normal scenarios. Create five histograms with various numbers of classes, and record the chi-square test statistic and its p-value for each. What effect does the number of classes have on the p-value? What do you consider the "best" number of classes? Do all the histograms appear normal? Discuss any anomalies that you notice. Repeat this exercise with one of the uniform scenarios. Discuss the similarities and differences in doing the chi-square test for the two variables.
2. Select a continuous variable of interest from one of the databases (U.S. States or World Nations) and test it for normality using the chi-square test, the K-S test, and the probability plot using the Set From Sample option for the parameters. Try more than one chi-square test. Do the tests agree? Examine the descriptive statistics (minimum, maximum, quartiles) and tell whether the sample and hypothesized normal values are similar. Repeat this exercise using another variable of interest. Tell which variable is closer to "normal." Discuss the similarities and differences in doing the K-S test for the two variables.
3. Select a Poisson scenario and carry out a hypothesis test using the chi-square test, the K-S test, and the probability plot. Click User Specified Values and try several parameter values to see whether the fit can be improved. How sensitive is the test to changes in the parameter? Finally, try the Set From Sample option. Was the fit improved? Choose a continuous distribution to fit to this data set (try more than one if appropriate). Do you get a reasonable fit? Compare and contrast the fit with the original Poisson distribution.

Team Learning Projects

Select one of the three projects listed below, and produce a team project that is suitable for an oral presentation. Use presentation software or a large poster board(s) to display your results. Graphs and tables should be large enough for your audience to see. Each team member should be responsible for producing some of the exhibits. Ask your instructor if a written report is also expected.

1. A team of two should select a continuous variable from any source and test the data for normality, using the chi-square test with various frequency histograms. Use 2 to 20 classes (make a display of each). Record each chi-square test statistic and its p-value. Describe whether the tests agree, and discuss any problems that arise. If the tests disagree, suggest reasons. Would you say that the chi-square test is robust to the number of bins (classes)? What does Sturges' Rule say about the number of classes? Which chi-square classification(s) would you recommend, and why? The objective of this project is to understand the effect of the number of classes on the results of a chi-square test.
2. Each member of a team of three should select a different variable from the normal-lognormal scenarios. Using the chi-square test, do a goodness-of-fit test first for a normal distribution, then for a lognormal distribution, and finally for a uniform continuous distribution. For each test, try at least three different classifications using different numbers of bins (classes). Which of the three proposed distributions is most compatible with your data? Would you say that more than one is acceptable? Explain any disagreements or difficulties that arise. The objective of the project is to explore the sensitivity of the chi-square test under various competing hypotheses about the population distribution.
3. This project is for a team of three to five. The team should agree on one of the discrete scenarios (binomial, Poisson, geometric, uniform) in the Notebook. Each team member should select a different variable and do various goodness-of-fit tests (chi-square, K-S, probability plot) to see whether the data are compatible with the hypothesized distribution and its parameters. Click User Specified Values and try several parameter values to see whether the fit can be improved. Finally, try the Set From Sample option. Compare the fit from the various tests for each set of parameter values. The objective of the project is to see whether the *a priori* distribution is reasonable, to see whether the hypothesized parameters were correct, and to see whether the tests agree. Discuss any problems that arise. If a different distribution might be better than the hypothesized one, explain why.

Self-Evaluation Quiz

1. Ideally, the choice of a theoretical distribution to be fitted to a sample is *not* based on
   a. the nature of the data-generating situation.
   b. the data type (continuous, discrete).
   c. the magnitude and range of the data.
   d. the sample statistics that give the best fit.
   e. the assumptions upon which the distribution is based.

2. The histogram is *least* likely to reveal
   a. the modal class(es).
   b. the mean and standard deviation.
   c. the general degree of skewness.
   d. the general shape of the distribution.
   e. the approximate range.

3. The chi-square GOF test on frequencies
   a. yields an inflated test statistic if the expected cell frequencies are small.
   b. usually requires that the parameters be estimated from the sample.
   c. requires a reasonably large sample size.
   d. is not essentially a visual test.
   e. has all of these characteristics.

4. Increasing the number of histogram classes with the exact range covered generally leads to
   a. more degrees of freedom in the chi-square test.
   b. decreasing class interval width.
   c. smaller expected frequencies in each class.
   d. a more powerful chi-square test.
   e. all of the above.

5. The probability plot is *primarily* used to
   a. check for randomness.
   b. check for normality.
   c. check for skewness.
   d. check the interquartile range.
   e. check more than one of the above.

6. To ascertain the approximate range of a sample we could *not* use which display(s)?
   a. Dot plot.
   b. Table of descriptive statistics.
   c. ECDF plot.
   d. Probability plot.
   e. Chi-square test with open-end classes.

7. The probability plot does *not* have which characteristic?
   a. It reveals outliers.
   b. It is easy to generate without a computer.
   c. It is approximately linear along the diagonal if the null hypothesis is true.
   d. It presents the data in sorted order.
   e. It is difficult to interpret if the data values are clustered.

8. The ECDF plot
   a. is a step function.
   b. works better for discrete data.
   c. presents the sample data in unsorted order.
   d. is generally U-shaped.
   e. has all of the above characteristics.

9. The Kolmogorov-Smirnov test
   a. requires the data to have a normal distribution.
   b. compares the ECDF and hypothesized distribution.
   c. is based on a minimum difference.
   d. has all of the preceding characteristics.
   e. has none of the above characteristics.

10. Which statement is *not* correct?
    a. The lognormal model is suitable when ln(x) is believed to be normal.
    b. The uniform distribution may be either discrete or continuous.
    c. The geometric distribution is skewed and has a very long right tail..
    d. The uniform distribution may be skewed either right or left.
    e. The geometric may be approximated by an exponential if the range is large.

11. Which is *not* a characteristic of the Poisson model?
    a. It is suitable only for discrete data.
    b. It can be approximated by a normal if the range is large.
    c. It is a versatile candidate for many kinds of samples.
    d. It is appropriate only for certain specific data-generating situations.
    e. It is useful in quality control.

12. The exponential distribution
    a. is right-skewed.
    b. is used as a model of waiting time for event arrivals.
    c. is appropriate only for non-negative data.
    d. is a continuous distribution.
    e. has all of the above characteristics.

Glossary of Terms

Binomial distribution Two-parameter distribution describing discrete data generated by a binary (success/failure) experiment with n independent trials and constant probability of success.

Chi-square test Non-parametric test for goodness-of-fit that groups the data into k classes and calculates the class frequency that would be expected under the hypothesized distribution. The parameters may be specified *a priori*, but often are estimated from sample data.

Degrees of freedom The number of independent pieces of information remaining after a statistical calculation. In a chi-square test with k classes, if m parameters are estimated the degrees of freedom will be k – 1 – m.

ECDF plot Acronym for *Empirical Cumulative Distribution Function*, a step function that plots the cumulative proportion of the sample against the corresponding sample X value. It is customary to superimpose the hypothesized distribution on the ECDF for comparison.

Expected frequency In a chi-square test, the number of observations that would be expected to fall within a defined interval if the data came from the hypothesized distribution.

Exponential distribution Continuous distribution used to characterize waiting time until a Poisson arrival (e.g., failure of a component). X must be non-negative. It has one parameter, the mean arrival rate.

Fitted distribution Distribution whose parameters have been estimated from sample data.

Geometric distribution Discrete distribution that describes the number of trials until the first "success" in a binomial experiment. It has one parameter (the probability of "success").

Goodness-of-fit test Any test that compares a hypothesized distribution with data from a sample. It may be visual (see **Probability plot**) but usually is a formal mathematical test.

Histogram Bar chart showing on the horizontal axis the values of X grouped into classes (intervals or bins) and on the vertical axis the frequency of occurrence within each class.

Kolmogorov-Smirnov test The K-S test statistic is the largest difference between the ECDF and the hypothesized distribution, measured in a vertical direction. It is a non-parametric test.

Kurtosis Measure of relative peakedness. For a unimodal, symmetric distribution K = 3 indicates a *mesokurtic* (normal, bell-shaped) distribution; K < 3 indicates a *platykurtic* distribution (flatter than normal, with short tails); and K > 3 indicates a *leptokurtic* distribution (more sharply peaked than normal, with long tails).

Lognormal distribution Continuous two-parameter distribution appropriate if ln(x) may be supposed to be normal. It is useful for certain kinds of positive, right-skewed data.

Nonparametric test Method of hypothesis testing that does not rely on an assumption about the population distribution from which the sample was drawn (e.g., normality).

Normal distribution "Bell-shaped" or Gaussian distribution with two parameters, the mean and variance. In a normal population, we expect 68.26% of the observations within 1 standard deviation, 95.44% within 2 standard deviations, and 99.73% within 3 standard deviations.

Observed frequency In a chi-square test, the actual frequency of sample data in a class interval.

Outlier Sample observation that is more than 3 standard deviations from the mean (suspected to come from a different population because it differs markedly from other sample data).

Parameter Numerical characteristic of a distribution that generally determine its probability density function, mean, variance, and other characteristics.

Probability density function (p.d.f.) Value f(x) assigned to every X in the domain of a distribution. The integral of f(x) must be unity (continuous distribution) or the f(x) values must sum to unity (discrete distribution).

Probability plot Comparison of each sample observation with the expected value of this observation assuming that the hypothesized distribution is correct (the sample is first sorted). If the hypothesized distribution is correct, the probability plot should be roughly linear. This provides a simple visual test for conformity between the sample and the hypothesis. The normal probability plot is most common.

Skewness Measure of relative symmetry. Zero indicates symmetry, positive values show a long right tail, and negative values show a long left tail.

Standardized Z value Obtained from an observation when we subtract the mean and divide by the sample standard deviation. These transformed data are called Z values because they may be used to see how closely the sample resembles a standard normal distribution, to spot outliers, and to check for asymmetry about the mean.

Sturges' Rule Rule of thumb suggesting that the number of histogram bins should be $1 + \log_2(n)$ where n is the sample size (e.g., 4 bins for 8 observations, 5 bins for 16 observations, 6 bins for 32 observations). It is a guideline to avoid too many or too few classes. If the data are skewed, Sturges' Rule may not provide enough classes to reveal adequate detail.

Theoretical distribution Hypothesized model of the population from which a sample is drawn, generally with specified parameter values.

Uniform distribution Continuous or discrete distribution describing a random variable whose p.d.f. is constant for all values in its domain. Its parameters are its upper and lower limits.

Solutions to Self-Evaluation Quiz

1. d Read the Overview of Concepts. Do Team Project 3.
2. b Do Exercise 5. Consult the Glossary. Read the Overview of Concepts.
3. e Do Exercises 6–12. Consult the Glossary. Read the Overview of Concepts.
4. e Do Exercises 5–12. Do Individual Project 1 or Team Projects 1 or 2.
5. b Do Exercises 12–14. Consult the Glossary.
6. e Do Exercises 2, 12, and 15. Read the Overview of Concepts.
7. b Do Exercises 13–15. Consult the Glossary. Read the Overview of Concepts.
8. a Do Exercise 16. Consult the Glossary. Read the Overview of Concepts.
9. b Do Exercises 17–20. Consult the Glossary.
10. d Do Team Project 2. Do Team Project 2.
11. c Do Individual Project 3. Consult the Glossary.
12. e Do Exercise 24. Consult the Glossary.