Collaborative Statistics: Projects: Continuous Distributions & Central Limit Theorem*

Susan Dean
Barbara Illowsky, Ph.D.

Abstract

In this project, students will identify and analyze a continuous data set, determine which distribution model most closely describes the data, and calculate probabilities.

1 Student Learning Objectives

- The student will collect a sample of continuous data.
- The student will attempt to fit the data sample to various distribution models.
- The student will validate the Central Limit Theorem.

2 Instructions

As you complete each task below, check it off. Answer all questions in your summary.

3 Part I: Sampling

_____ Decide what **continuous** data you are going to study. (Here are two examples, but you may NOT use them: the amount of money a student spends on college supplies this term or the length of a long distance telephone call.)
_____ Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random (using a random number generator) sampling. Do not use convenience sampling. What method did you use? Why did you pick that method?
_____ Conduct your survey. Gather **at least 150 pieces of continuous quantitative data**.
_____ Define (in words) the random variable for your data. $X =$ _____
_____ Create 2 lists of your data: (1) unordered data, (2) in order of smallest to largest.
_____ Find the sample mean and the sample standard deviation (rounded to 2 decimal places).

*Version 1.9: Feb 19, 2009 2:50 pm US/Central
†http://creativecommons.org/licenses/by/2.0/

1. $\bar{x}$ =
2. s =

_____ Construct a histogram of your data containing 5 - 10 intervals of equal width. The histogram should be a representative display of your data. Label and scale it.

4 Part II: Possible Distributions

_____ Suppose that X followed the theoretical distributions below. Set up each distribution using the appropriate information from your data.
_____ Uniform: X ~ U _____ Use the lowest and highest values as a and b.
_____ Exponential: X ~ Exp _____Use $\overline{x}$ to estimate μ .
_____ Normal: X ~ N _____ Use $\overline{x}$ to estimate for μ and s to estimate for σ.
_____ **Must** your data fit one of the above distributions? Explain why or why not.
_____ **Could** the data fit 2 or 3 of the above distributions (at the same time)? Explain.
_____ Calculate the value k(an X value) that is 1.75 standard deviations above the sample mean. k = _____ (rounded to 2 decimal places) Note: $k = \overline{x} + (1.75) * s$
_____ Determine the relative frequencies (RF) rounded to 4 decimal places.

1. RF = $\frac{\text{frequency}}{\text{total number surveyed}}$
2. RF $(X < k)$ =
3. RF $(X > k)$ =
4. RF $(X = k)$ =

Use a separate piece of paper for EACH distribution (uniform, exponential, normal) to respond to the following questions.

NOTE: You should have one page for the uniform, one page for the exponential, and one page for the normal

_____ State the distribution: X ~ _____
_____ Draw a graph for each of the three theoretical distributions. Label the axes and mark them appropriately.
_____ Find the following theoretical probabilities (rounded to 4 decimal places).

1. P(X < k) =
2. P(X > k) =
3. P(X = k) =

_____ Compare the relative frequencies to the corresponding probabilities. Are the values close?
_____ Does it appear that the data fit the distribution well? Justify your answer by comparing the probabilities to the relative frequencies, and the histograms to the theoretical graphs.

5 Part III: CLT Experiments

_____ From your original data (before ordering), use a random number generator to pick 40 samples of size 5. For each sample, calculate the average.
_____ On a separate page, attached to the summary, include the 40 samples of size 5, along with the 40 sample averages.
_____ List the 40 averages in order from smallest to largest.
_____ Define the random variable, $\overline{X}$, in words. $\overline{X}$ =
_____ State the approximate theoretical distribution of $\overline{X}$. $\overline{X}$~

_ _ _ _ _ _ _ Base this on the mean and standard deviation from your original data.
_ _ _ _ _ _ _ Construct a histogram displaying your data. Use 5 to 6 intervals of equal width. Label and scale it.
Calculate the value $\overline{k}$ (an $\overline{X}$ value) that is 1.75 standard deviations above the sample mean. $\overline{k}$= _ _ _ _ _ (rounded to 2 decimal places)
Determine the relative frequencies (RF) rounded to 4 decimal places.

1. RF($\overline{X} < \overline{k}$) =
2. RF($\overline{X} > \overline{k}$) =
3. RF($\overline{X} = \overline{k}$) =

Find the following theoretical probabilities (rounded to 4 decimal places).

- $P(\overline{X} < \overline{k})$ =
- $P(\overline{X} > \overline{k})$ =
- $P(\overline{X} = \overline{k})$ =

_ _ _ _ _ _ _ Draw the graph of the theoretical distribution of X.
_ _ _ _ _ _ _ Answer the questions below.
_ _ _ _ _ _ _ Compare the relative frequencies to the probabilities. Are the values close?
_ _ _ _ _ _ _ Does it appear that the data of averages fit the distribution of $\overline{X}$ well? Justify your answer by comparing the probabilities to the relative frequencies, and the histogram to the theoretical graph.
_ _ _ _ _ _ _ In 3 - 5 complete sentences for each, answer the following questions. Give thoughtful explanations.
_ _ _ _ _ _ _ In summary, do your original data seem to fit the uniform, exponential, or normal distributions? Answer why or why not for each distribution. If the data do not fit any of those distributions, explain why.
_ _ _ _ _ _ _ What happened to the shape and distribution when you averaged your data? **In theory,** what should have happened? In theory, would "it" always happen? Why or why not?
_ _ _ _ _ _ _ Were the relative frequencies compared to the theoretical probabilities closer when comparing the X or $\overline{X}$ distributions? Explain your answer.

6 Assignment Checklist

You need to turn in the following typed and stapled packet, with pages in the following order:

_ _ _ _ _ **Cover sheet**: name, class time, and name of your study
_ _ _ _ _ **Summary pages**: These should contain several paragraphs written with complete sentences that describe the experiment, including what you studied and your sampling technique, as well as answers to all of the questions above.
_ _ _ _ _ **URL** for data, if your data are from the World Wide Web.
_ _ _ _ _ **Pages, one for each theoretical distribution**, with the distribution stated, the graph, and the probability questions answered
_ _ _ _ _ **Pages of the data requested**
_ _ _ _ _ **All graphs required**