MATHEMATICS E-156, SPRING 2014
MATHEMATICAL FOUNDATIONS OF STATISTICAL SOFTWARE
Module #2 (Exploring Data)

Last modified: February 3, 2014

Reading from Chihara and Hesterberg

- Chapter 2
- Appendix A.3
- Theorem 6.2 and Definition 6.3 on pp. 149-150. This theorem explains why the R function var() does not compute what you might have expected.
- Appendix A.7, Definition A.9. This is mathematical background for Section 2.7.

Optional Reading from Haigh

- Section 4.3 introduces continuous random variables.
- Section 5.2 is another version of the Proof of the Week.
- The normal, exponential and gamma distributions, which we are using for examples, appear in several places – check the index. We will explore these in great detail later in the course. For the moment, you need not understand how they might arise.

Proof of the Week

- Let $X_1, X_2, \cdots X_n$ be independent random variables from a distribution with $\text{Var}[X_i] = \sigma^2 < \infty$.

  Prove that

$$E[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2] = \frac{n-1}{n}\sigma^2.$$

  This is theorem 6.2 on page 149 but can be done much earlier.

R scripts

- 0B-PlotTutorial.R
  This script includes all the commands from the second of the three tutorials that the textbook authors have posted on their Web site. It is on the "Installing R" page of the course Web site. By using it, you can work through the tutorial without having to type in all the commands. There are fancier ways to do graphics in R. The book R Graphics Cookbook by Winston Chang uses `ggplot2`. The `lattice` package is another popular add-on. We will stick to basic R grahics, and so should you in your homework. Using an alternative graphics package in your term project would be fine.

- 0C-DistributionTutorial.R This script includes all the commands from the third of the three tutorials. You will meet the collection of four functions, e.g. `dnorm(), pnorm(), qnorm(), rnorm()` that R provides for many important probability distributions.

- 2A-DataSets.R This covers sections 2.1 through 2.3 of the textbook. It concentrates on using R graphics to explore thr Flight Delays dataset.

- 2B-Quantiles.R This covers sections 2.4 and 2.5 of the textbook, including the thorny issue of how to define quantiles for discrete random variables. You may already be familiar with quantiles as "percentiles."

- 2C-Moments.R This script explores *skewness* and *kurtosis*, which go beyond mean and variance in describing a distribution. The integrals in the textbook are done by standard calculus techniques, but R can replicate the results using numerical integration.

- 2P-Proof2.R This script includes examples to demonstrate that "sample variance is a biased estimator of population variance," which is the Proof of the Week.

Mathematical notes

1. Variance of the mean of n independent random variables

   This is a theorem of probability.

   Suppose that $X_1, X_2, \cdots X_n$ are independent random variables, all with the same expectation μ and variance σ^2. Their mean is

   $$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots X_n).$$

   Prove that

   $$E[\overline{X}] = \mu$$

   and that

   $$\mathrm{Var}[\overline{X}] = \frac{1}{n}\sigma^2.$$

2. Proof of the week

   This is a theorem of statistics.

   Let $X_1, X_2, \cdots X_n$ be independent random variables from a distribution with $\mathrm{Var}[X_i] = \sigma^2 < \infty$.

   We do not know the expectation μ, although we know from the previous result that the expectation of $\overline{X}$ is equal to μ.

   We also do not know the variance. We try to estimate it by using the usual formula but, not knowing μ, we can do no better than to use $\overline{X}$ in its place.

$$\text{Prove that } E[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2] = \frac{n-1}{n}\sigma^2.$$

   It follows that $S^2 = E[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2] = \sigma^2$. This is what var() computes.

3. Inverse functions and quantiles

- For a continuous random variable described by a density function $f(x)$, the distribution function

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

is a continuous, nondecreasing function that goes from 0 at $-\infty$ to 1 at ∞. If $f(x) > 0$ then $F(x)$, with codomain (0,1), is both surjective and injective and so has an inverse function F^{-1}, the quantile function. The domain of F^{-1} is (0,1); the codomain is $(-\infty, \infty)$.
Example in R: f is `dnorm()`, F is `pnorm()`, F^{-1} is `qnorm()`.

- If the density $f(x)$ is zero on some intervals, then $F(x)$ is still surjective but ceases to be injective. In this case there are a variety of conventions about whether to use the left endpoint, the middle, or the right endpoint of an interval on which $f(x) = 0$ and $F(x)$ is constant.

- If the probability distribution is discrete, then $F(x)$ is still nondecreasing, but it is neither injective nor surjective. In this case `quantile(q)` appears to interpolate between the largest value for which $F(x_1) \leq q$ and the smallest value for which $F(x_2) > q$.

4. Moments, skewness, and kurtosis

The nth moment of random variable X is just the expectation of X^n.

$$\mu'_n = E[X^n]$$

The nth central moment of X is the expectation of $(X - \mu)^n$.

$\mu_n = E[(X - \mu)^n]$ (zero for odd n if the distribution is symmetrical)

The skewness is rescaled so that multiplying X by a constant has no effect. Positive skewness means a long tail to the right.

$$\gamma_1 = \frac{\mu_3}{\sigma^3}.$$

The kurtosis is also rescaled so that multiplying X by a constant has no effect. Then the kurtosis of a normal distribution (3) is usually subtracted.

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3.$$

Positive kurtosis comes from a sharp central peak and "fat tails." Negative kurtosis comes from thin or nonexistent tails. By definition a normal distribution has zero kurtosis.

Section problems These are optional, but they may help to focus the discussion during section. You should work through the second and third R tutorials before trying them.

1. The lower right-hand plot of figure 2.15 appears to be the density function for a gamma distribution with shape = 2, rate = 0.5. The density function in R is `dgamma(x, 2, 0.5)`. Plot this function to confirm the identification, then see if you get the same skewness and kurtosis that are shown on the plot. Why would you expect positive values for both these quantiities?

2. For the same gamma distribution, find the 0.2 and 0.8 quantiles. Illustrate your answers by drawing appropriate horizontal or vertical lines on graphs of `dgamma(x, 2, 0.5)`, `pgamma(x, 2, 0.5)`, and `rgamma(x, 2, 0.5)`, putting all three plots into a 2x2 grid by the technique at the end of the plot tutorial.

3. Problem 5 on page 31 of the textbook (General Social Survey Case Study).

Homework assignment Again, this assignment should be submitted as a single R script. Upload it to the dropbox on the Class 2 page of the Web site.

It is OK to paste and edit lines from the scripts on the course Web site. It is not OK to paste lines from your classmates' solutions!

1. Work through the second R tutorial, on plotting. After completing the tutorial, do the following:

   - For the data in States03.csv, make a scatter plot of percentage of high school graduates against teacher pay, with a solid red vertical line at mean teacher pay and a dashed blue horizontal line at mean percentage of high school graduates.

2. Problem 6 on page 31 of the textbook (Black Spruce Seedlings)

3. Work through the third R tutorial, on distribution functions. After completing the tutorial, do the following:

   (a) Plot a graph of the Student's t density function for six degrees of freedom `dt(x,6)` on the interval [-3,3]. This is a classic example of a distribution with "fat tails." On the same plot overlay, in green, a graph of the density function for the standard normal distribution `dnorm(x)`

   (b) Determine the 0.1 and 0.9 quantiles for each of these distributions and add vertical lines (in black and green respectively) that mark off the interval on which 80% of the area under the graph lies for each distribution.

   (c) Plot a graph of the Student's t distribution function for six degrees of freedom `pt(x,6)` on the interval [-3,3]. On the same plot overlay, in green, a graph of the quantile function for the standard normal distribution `dnorm(x)` Add horizontal lines (in black and green respectively) at 0.1 and 0.9 and use them to estimate the 0.1 and 0.9 quantiles.

   (d) Plot a graph of the Student's t quantile function for six degrees of freedom `qt(x,6)` on the interval [-3,3]. On the same plot overlay, in green, a graph of the distribution function for the standard normal distribution `qnorm(x)` Add vertical lines (in black and green respectively) at 0.1 and 0.9 and use them to estimate the 0.1 and 0.9 quantiles. (Since `dnorm()` and `qnorm()` are inverse functions, this is the same plot as in part (c), with the axes reversed!)

   (e) Compare the Student's t distribution for six degrees of freedom with the normal distribution by using `qqnorm()` and `qqline()`.

4. (a) Because a histogram of flight delays has a long tail to the right, we would expect a positive skewness.
   Load FlightDelays.csv, and by using the same approach that was used for the Poisson distribution in script 2C, determine the skewness of the delays. You will need to start by calculating the mean and variance. (One can argue about whether the data are a population or a sample from a larger population, but there are so many data points that it really doesn't matter!)

   (b) Since the Student t distribution has fatter tails than the normal distribution, we would expect a positive kurtosis. The integrals are a pain to evaluate by using calculus, even though $\mu = 0$. By having R evaluate the second and fourth moments by numerical integration (consult script 2C for the tricky syntax of `integrate`), confirm that the kurtosis of the Student's t distribution for six degrees of freedom is 3.

5. (a) Problem 17 on page 33. By using the technique from the last section of the plot tutorial, you can place the graphs side by side in the top row of a 2x2 grid.

   (b) In the bottom row, plot the quantiles for 100 values ranging from 0.01 to 0.99. You will get essentially the same graphs, but with the axes interchanged.