Lesson Plan

- Discuss Quizzes/Answer Questions
- Summary Statistics
- Histograms
- The Normal Distribution
- Using the Standard Normal Table

3. Summary Statistics

Given a collection of data, one needs to find representations of the data that facilitate understanding and insight. Three standard tools for this are:

- Measures of Central Tendency (mean, median, mode)
- Measures of Dispersion (standard deviation, range, IQR)
- Visualizations (histograms, other graphics)

We shall cover this quickly, so be sure to read the book.

3.1 Measures of Central Tendency

The mean is just the average of the data. Suppose one has a sample of n observations with values $X_1, \ldots, X_n$. Then the mean is just

$$\begin{aligned} \bar{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ &= \frac{1}{n}(X_1 + \cdots + X_n). \end{aligned}$$

The mode is just the value that occurs most frequently in the sample. There can be many modes.

The median is the middle largest value among the n observations, if n is odd. If there are an even number of observations, then we average the two middle-largest values.

Example: Suppose on observes the following data:

$$1,\ 0,\ 2,\ \text{-}2,\ 1,\ \text{-}2,\ 5,\ \text{-}1$$

The mean is $\bar{X} = \frac{1}{8}(1 + 0 + 2 - 2 + 1 - 2 + 5 - 1) = .5$.

The modes are 1 and -2.

The median is the average of 0 and 1, or .5.

The mean can be pulled in misleading directions if there are outliers. A single large or small datum will have a large influence on the mean, but not on the median.

An outlier is an incorrect or unrepresentative observation that is very different from the others in the sample.

3.2 Measures of Dispersion

To measure how spread out a sample is, we mostly use the standard deviation. This is:

$$\begin{aligned} sd &= \sqrt{\frac{1}{n}[(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2]} \\ &= \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X})^2}. \end{aligned}$$

This is the square root of the average of the squared deviations between each observation and the mean.

The first formula is better for understanding, and the second is better for calculation.

Note Bene: In some books, one divides by $n-1$ rather than n, but we shall not do that.

The majority of observations are usually within 1 sd of the mean. But in some extreme cases it can happen that none of the data are within 1 sd of the mean.

However, one can prove that:

- At least 75% of the observations must always be within 2 standard deviations of the mean.
- At least 89% of the observations must always be within 3 standard deviations of the mean.

This is a result of Tchebyshev's Theorem.

The range is the largest observation minus the smallest. As a measure of dispersion, it is strongly influenced by outliers.

The interquartile range is 75th percentile of the data minus the 25th percentile (the median is the 50th percentile). For our purposes, we shall calculate the 75th percentile is as the median of all observations strictly greater than the median, and the 25th percentile as the median of all observations strictly less than the median. (This is not perfectly correct, but is good enough for our purposes.)

The interquartile range is not strongly influenced by outliers.

Example: Suppose on observes the following data:

$$1,\ 0,\ 2,\ \text{-}2,\ 1,\ \text{-}2,\ 5,\ \text{-}1$$

The range is 5 - (-2) = 7.

The interquartile range is the median of all values greater than .5, minus the median of all values less than .5, or median$\{1, 2, 1, 5\}$ minus median$\{0, -2, -2, -1\}$, or 1.5 - (-1.5) = 3.

The standard deviation is

$$sd = \sqrt{\frac{1}{8}[(1 - .5)^2 + \cdots + ((-1) - .5)^2]} =$$

but it is faster to calcuate

$$sd = \sqrt{\frac{1}{8}[1^2 + \cdots + (-1)^2] - (\bar{.5})^2} = .$$

3.3 Properties of $\bar{X}$ and the sd

Suppose we have n observations $X_1, \ldots, X_n$ and we use these to create a new sample $Y_1, \ldots, Y_n$ where $Y_i = aX_i + b$. (This often arises when converting units of measurement, such has changing Fahrenheit data into the Centigrade scale.)

Then

$$\begin{aligned} \bar{Y} &= a\bar{X} + b \\ sd_Y &= |a| sd_X. \end{aligned}$$

Can you guess formulae for the median, mode, range, and interquartile range of the Y values?

3.4 The Histogram

The histogram shows where sample values are located and where they concentrate.

The x-axis gives the sample value, and the y-axis is the percent per x-value. (This is different from a bar chart.)

In a histogram, the areas under a block represent percentages.

By convention, the left endpoint of a histogram bar is included in the interval, but not the right.

This incomplete histogram represents monthly wages for part-time employees.

- What is the height of the missing bar?
- Estimate the percentage of part-time employees who make between \$100 and \$110.
- Does the second bar include people who make exactly \$100/month?
- Does the second bar include people who make exactly \$200/month?

.

As the sample size goes to infinity ($n \to \infty$) and as the bin-width of the histogram goes 0 ($h \to 0$) at the appropriate relative rates, then the histogram becomes smooth.

This limiting smooth curve is called a distribution.

.

3.5 The Normal Distribution

Some limiting histograms are famous and have names. The most famous distribution is the normal distribution (a/k/a the Gaussian distribution or the bell-shaped curve).

The term "Gaussian" refers to Carl Friedrich Gauss. Who was he?

People believe the normal distribution describes IQ, height, rainfall, measurement error, and many other features. This is only approximately true. But it is a good approximation for features that are the sum of many separate increments.

How can you decide if data are a random sample from a normal distribution?

- Inspect the histogram.
- Make a normal probability plot.

To make a normal probability plot, find the sample mean $\bar{X}$ and the sample standard deviation sd. For each observation X_i, find the area under the standard normal curve (from the z-table in our book, page A-105) to left of $(X_i - \bar{X})/sd$. Plot this area against X_i for all i. Perfect normality corresponds to a straight line.

A normal distribution with mean μ and standard deviation σ has the equation:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{1}{2\sigma^2}(x-\mu)^2\right]$$

for $-\infty < \mu < \infty$ and $\sigma \geq 0$.

The μ is the mean of the entire population, whereas $\bar{X}$ is used to denote the mean of a sample from the population. Similarly, σ is the standard deviation of the entire population, whereas sd is used to denote the standard deviation of a sample.

The standard normal has $\mu = 0$ and $\sigma = 1$.

The mean of a normal distribution shows where it is centered. The standard deviation of a normal distribution shows how spread out the normal is.

.

3.6 The Standard Normal Distribution

Any question about a normal distribution can be converted into an equivalent question about the standard normal, and vice-versa.

First we practice reading the book's table (page A-105). Assume you have a population that is standard normal.

What proportion of the population has values between -1.5 and 1.5?

.

From the table, the proportion is .8664, or 86.64% of the population lies between -1.5 and 1.5.

Now go the other way. 80% of the population lies between what two values that are centered at 0?

From the table, the answer is about -1.3 and 1.3. So approximately 80% of the standard normal values are within 1.3 of the mean (recall, the mean is 0).

From the table, the answer is about -1.3 and 1.3. So approximately 80% of the standard normal values are within 1.3 of the mean (recall, the mean is 0).

Some problems to think about in preparation for the quiz on Monday.

- What is the value of z such that 25% of a standard normal population is larger than that value? (Ans: about .7)
- What is the value for which about 90% of the population is smaller? (Ans: about 1.3)
- What proportion of the population has a value larger than -1? (Ans: about .8413)
- What proportion of the population has a value less than .3? (Ans: about .3821)