2.0 Lesson Plan

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

2. 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, since most of this is review from elementary school.

2.1 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 one observes the following data:

1, 0, 2, -2, 1, -2, 5, -1

The mean is $\bar{X} = \frac{1}{8}(1 + 0 + 2 - 2 + 1 - 2 + 5 - 1) = 0.5.$

The modes are 1 and -2.

The median is the average of 0 and 1, or 0.5.

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.

2.2 Measures of Dispersion

To measure how spread out a sample is, we mostly use the **standard deviation** (or sd). This is:

$$sd = \sqrt{\frac{1}{n-1}[(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2]} \tag{1}$$

$$= \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2\right) - \frac{n}{n-1}\bar{X}^2}. \tag{2}$$

The sd is the square root of the average squared deviation of each observation from the mean. The square of the sd is the **variance**.

Formula (1) is better for understanding, but (2) is better for calculation.

Note Bene: When one observes the whole population rather than just a sample, then the formula for the population sd divides by n instead of $n-1$.

The majority of observations are usually within 1 sd of the mean. But it can happen that none of the data are less than 1 sd from the mean. Tchebyshev proved that for all datasets, the proportion of data that lie within a standard deviations of the mean must be at least $1-\frac{1}{a^2}$. In terms of probability, if you pick an observation X at random,

$$\mathbf{P}[\,|X-\mathbf{mean}|<a*sd\,]\geq 1-\frac{1}{a^2}.$$

Thus

- At least 75% of the observations must always be less than 2 standard deviations from the population mean.
- At least 89% of the observations must always be less than 3 standard deviations of the population mean.

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).

- The 25th percentile is the number u such that at least 25% of the sample is less than or equal to u and at least 75% of the sample is greater than or equal to u. (The u need not be a sample value; and if there are many numbers u that satisfy this definition, we take the middle value.)
- The 75th percentile is the number v such that at least 75% of the sample is less than or equal to v and at least 25% of the sample is greater than or equal to v. (If this is not unique, take the middle value.)

The interquartile range is not strongly influenced by outliers.

Example: Suppose you have the following sample: 1, 0, 2, -2, 1, 5, -2, -1.

It helps to order the data first:

-2, -2, -1, 0, 1, 1, 2, 5

The range is 5 - (-2) = 7. There are eight values, so the 25th percentile is any number between -2 and -1; we take -1.5. Similarly, the 75th percentile is 1.5. The IQR is the difference of these, or 1.5 - (-1.5) = 3.

The standard deviation is

$$sd = \sqrt{\frac{1}{7}[(1-0.5)^2 + \cdots + ((-1)-0.5)^2]} = ?$$

but it is faster to calculate

$$sd = \sqrt{\frac{1}{7}[1^2 + \cdots + (-1)^2] - (8/7)(0.5)^2} = 2.32993.$$

2.3 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: C = 5/9 * F - 17.778.

Then

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

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

2.4 The Histogram

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

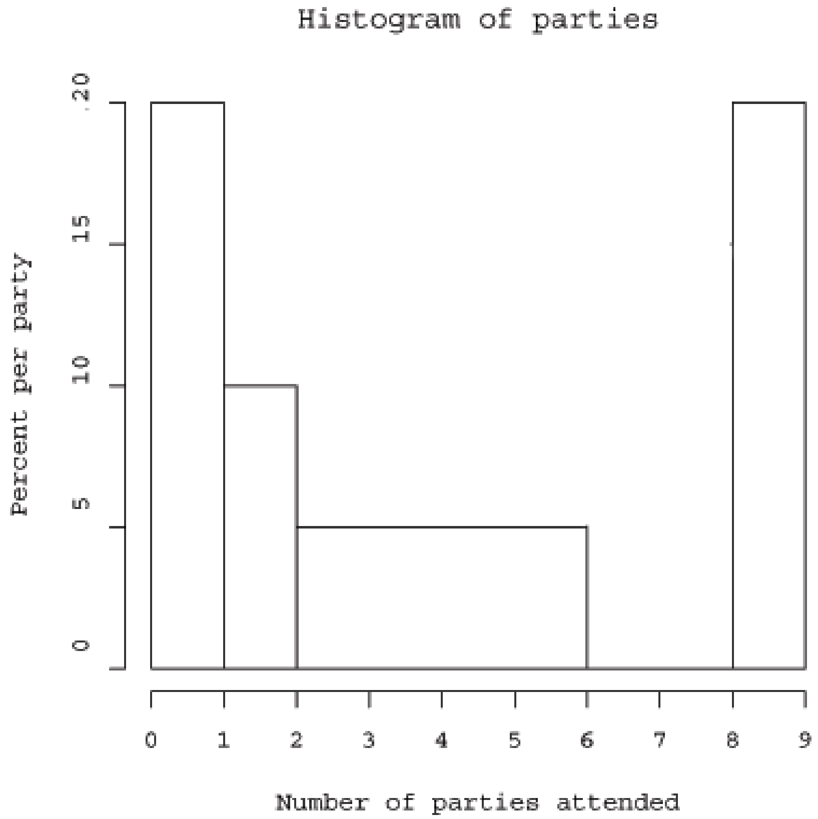
The x-axis gives the sample value, and the y-axis is the percent per x-value. (This is different from a bar chart.)

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.

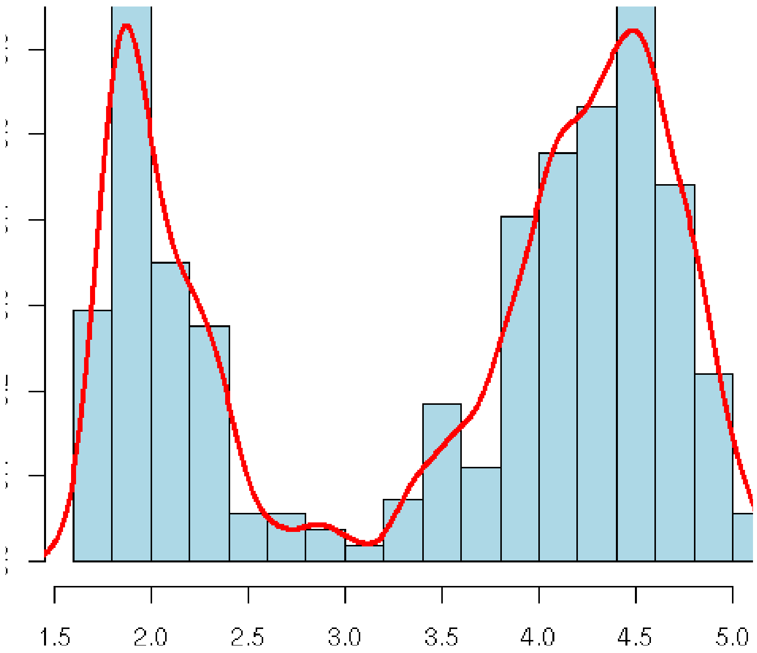
This incomplete histogram shows the number of parties attended in one week by Duke freshmen.

- What is the height of the missing bar?
- What percentage go to 1 or fewer parties?



As the sample size gets large and as the bin-width gets small at the appropriate relative rates, then the histogram becomes smooth.

This limiting smooth curve is called a **probability density function**.



2.5 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 distribution was named after Carl Friedrich Gauss, the greatest mathematician in history. He proved the fundamental theorem of algebra four ways, inventing a new branch of mathematics each time. He worked in number theory, co-invented the telegraph, and discovered non-Euclidean geometry, but did not publish, fearing controversy.

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.

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

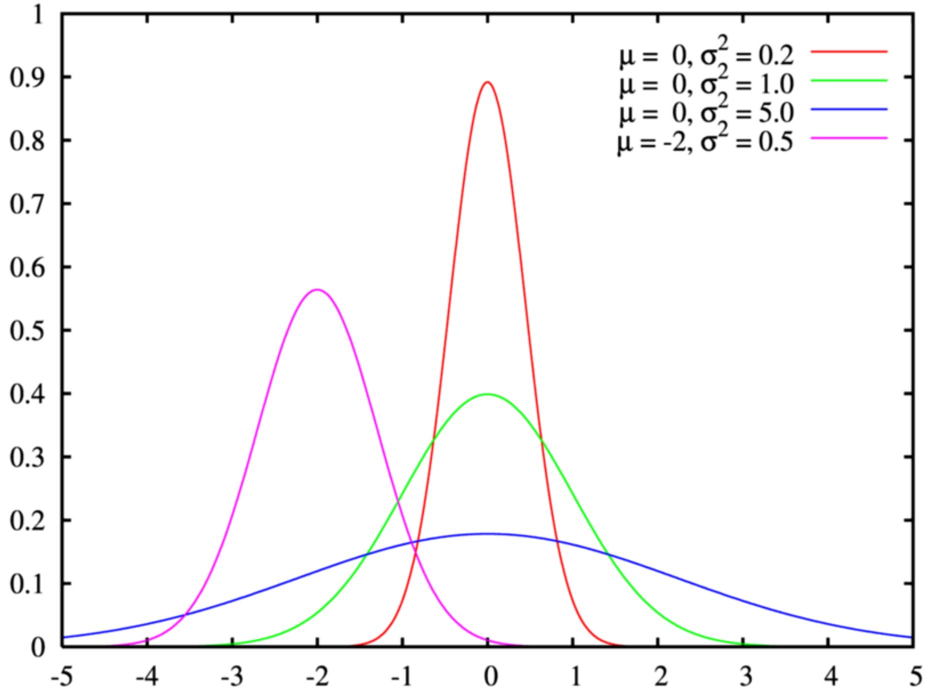
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$$

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.



2.6 The Standard Normal Distribution

Practice reading the standard normal table in the handout. A copy is also posted on the FAQ site. You may bring your table to class and use whatever notes you have on the back of the table during quizzes.

What proportion of a standard normal population has values between -1.5 and 1.5? **From the table, the proportion is 1 - 2 * 0.067 = 0.866 (i.e., the total area under the curve is one, and we subtract the upper tail area and the symmetric lower tail area).**

Now go the other way. About 80% of the population lies between what two values that are centered at 0? **The answer is about -1.28 and 1.28 (the table shows that 10% of the area is above 1.28, and symmetrically, 10% is below -1.28).**

Some problems to think about:

- What is the value of z such that 25% of a standard normal population is larger than that value? (Ans: about 0.67)
- What is the value for which about 90% of the population is smaller? (Ans: about 1.28)
- What proportion of the population has a value larger than -1? (Ans: about 0.841)
- What proportion of the population has a value less than -0.3? (Ans: about 0.382)

In this class, use the nearest value in the table. If you interpolate, or use a number from your calculator, it will confuse the graders.

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, order the observations from smallest to largest; denote the ordered observations by

$$X_{(1)}, X_{(2)}, \ldots, X_{(n)}.$$

For observation $X_{(i)}$, find the z-value such that $(i - 0.5)/n * 100\%$ of the area under the standard normal curve is to the left. Call this z-value Y_i. Then plot $(X_{(i)}, Y_i)$ for all $i = 1, \ldots n$. If this looks pretty much like a straight line, then the data are approximately normal. (There is a Stata command for this.)