Lecture 4

Criminology 250

Prof Maria Cuellar

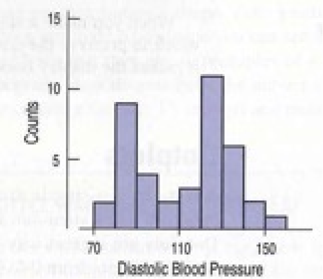
University of Pennslyvania

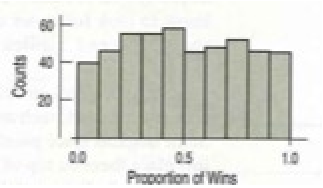
Shape: Modes

Does the histogram have a single, central hump, or several separated humps?

The peaks or humps in a histogram are called *modes*. If a histogram has one peak it's called *unimodal*, and if it has two peaks it's called *bimodal*. If it's more than that it's often called *multimodal*.

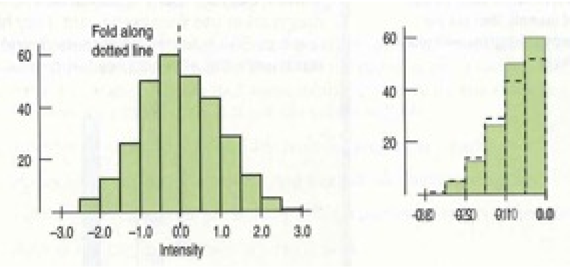
If it doesn't appear to have a mode because all the bars are approximately at the same height, then it's called *uniform*.



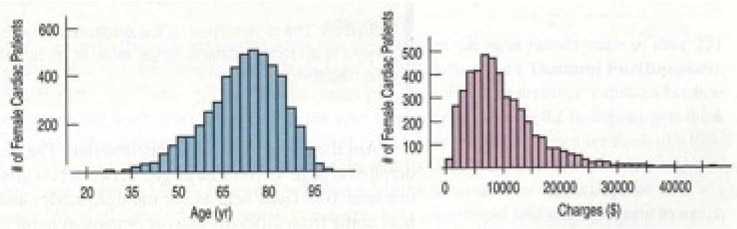


Shape: Symmetry

Is the histogram symmetric? Can you fold it along a vertical line through the middle and have the edges match pretty closely, or are more of the values on one side?



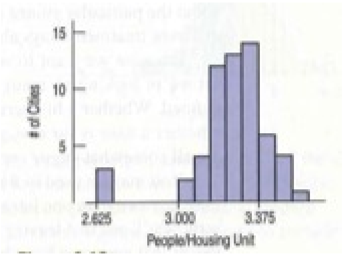
The (usually) thinner ends of a distribution are called the *tails*. If one tail stretches out farther than the other, the histogram is said to be *skewed* to the side of the longer tail.



Shape: Outliers

Do any unusual features stick out?

You should always mention any stragglers or *outliers* that stand away from the body of the distribution. An outlier might be the most informative part of your data, or it might just be an error. **But don't throw it away without comment.**

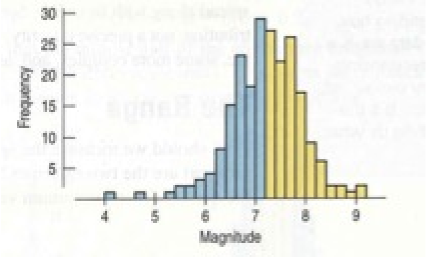


Center: Median

When we think of a typical value, we usually look for the center of the distribution. The center is easy to find with a symmetric unimodal distribution. Where is the center for skewed or multimodal distributions?

How do we discuss the "center" of a graph? We can talk about a *median*, the value that has half the values above it and half below it.

For a histogram, the median is the middle value that divides the histogram into two equal areas. The median has the same units as the data.



Spread: Range

How much do the values vary around the center? How spread out are they? There are several measures of spread.

The *range* of the data is defined as the difference between the maximum and minimum values:

$$Range = max - min.$$

Perhaps we want to ignore the extremes and concentrate on the middle of the data.

Spread: Interquartile range

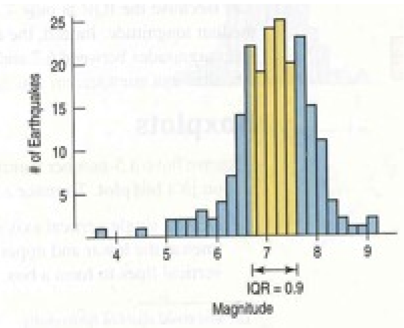
Suppose we divide the data in half at the median. Now divide both halves in half again, cutting the data into four quarters. We call these new dividing points *quartiles*.

One quarter of the data lies below the *lower quartile*, and one quarter above the *upper quartile*, so half (50%) the data lies between them.

The *interquartile range* is:

$$IQR = upper\ quartile - lower\ quartile.$$

For any percentage of the data there is a corresponding percentile, the value that leaves that percentage of the data below it. The lower and upper quartiles are also known as the 25th and 75th percentiles of the data, respectively. The median is the 50th percentile.

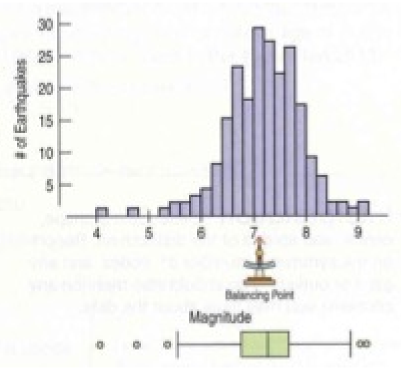


Center of symmetric distributions: The mean

Medians do a good job of summarizing the center of a distribution, even when the shape is skewed or when there is an outlier. But we often use the *mean*:

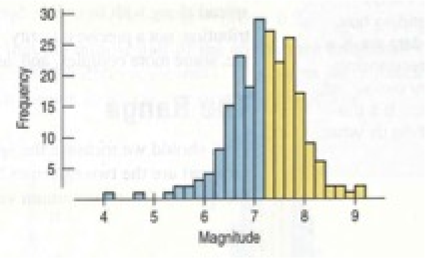
$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

The mean is the point where the histogram balances.



Which to use, the mean or the median?

Using the center of balance makes sense when the data are symmetric. But data are not always this well behaved. If the distribution is skewed or has outliers, the center is not well defined and the mean may not be what we want.



For example, the mean of the flight cancellations shown here doesn't give a very good idea of the typical percentage of cancellations.

The mean is 2.01%, but nearly two-thirds of months had cancellation rates below that, so the mean doesn't feel like a good overall summary. Why is the balancing point so high? The large outlying value pulls it to the right. For data like these, the median is a better summary of the center.

Because the median considers only the order of the values, it is resistant to values that are extraordinarily large or small: it simply notes that they are one of the "big ones" or "small ones", and ignores their distance from the center.

Why not just use the median all the time?

The spread of a symmetric distribution: The standard deviation

IQR is a good summary of spread, but it ignores much of the information about how individual values vary. A more powerful approach uses the *standard deviation* (or sd), which takes into account how far each value is from the mean.

The standard deviation is appropriate only for symmetric data.

The *variance* is the square of the standard deviations:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}.$$

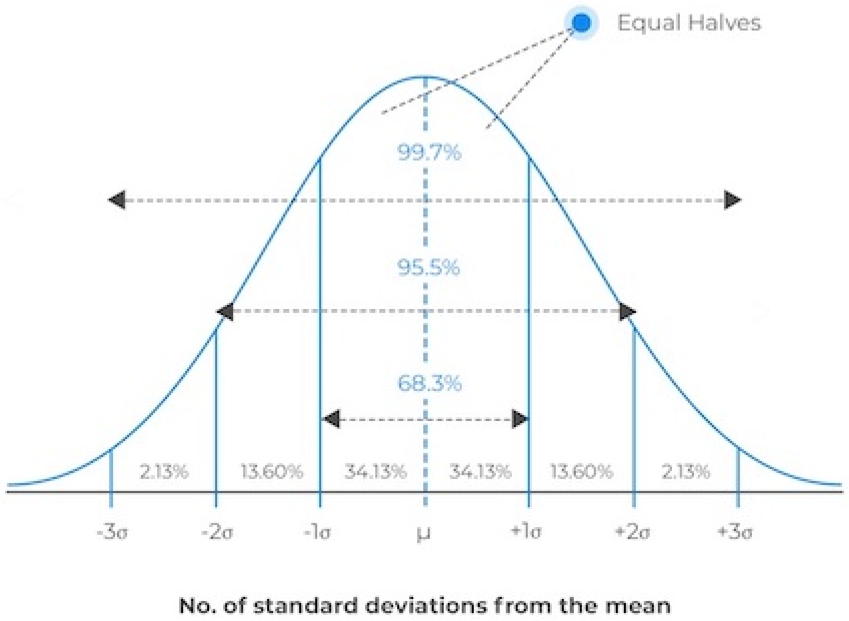
The definition of sd is:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}.$$

Normal model

The normal model has the following attributes:

- Unimodal
- Symmetric
- Bell-shaped
- Follows the 68/95/99 rule


Equal Halves
-3σ
-2σ
-1σ
μ
+1σ
+2σ
+3σ
No. of standard deviations from the mean