Mean

Same as ordinary average

Sum all the data values and divide by the sample size *n*.

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + ... + x_n)$$

Using summation notation, we write this as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i} x_i = \frac{1}{n}\sum x_i$$

Mean is only appropriate for interval or ratio scales, not ordinal or nominal.

Median

Value that divides the sample so that an equal number of cases are above and below.

- Sort the cases by magnitude of *x*: $x_{[1]}\ x_{[2]}...x_{[n]}$
- If *n* is odd, the median is the middle value:

  e.g. *x* = 1, 2, 5, 10, 11. Median is 5.
- If *n* is even, median is the average of the two middle values:

  e.g., *x* = 1, 2, 5, 10. Median is (2+5)/2 = 3.5
  - For test score example, sorted values are

    13 18 19 20 22 22 22 24 24 27 27 28 | 28 28 28 28 29 29 31 32 32 33 33 35

    So median is 28.

Median vs. Mean

Unlike the median, the mean is sensitive to extreme scores (outliers):

1, 2, 3, 4, 5 => mean=3, median=3

1, 2, 3, 4, 100 => mean=22, median=3

In symmetrical distributions, mean and median will be the same. In *skewed* distributions, they will be different. (more later).

So the median is often preferred for variables like income which have a relatively small number of extremely high scores.

Variance

How different is each score from the mean? $x_i - \bar{x}$

What's the average of these differences?

$$\frac{1}{n}\sum_i (x_i - \bar{x}) = 0$$

Positive deviations cancel out the negative

Mean is the only score for which this is true.

How to fix?

1. Take absolute values before averaging: **Mean absolute deviation**.

$$\frac{1}{n}\sum_i |x_i - \bar{x}|$$

2. Square the deviations before averaging: **Variance**.

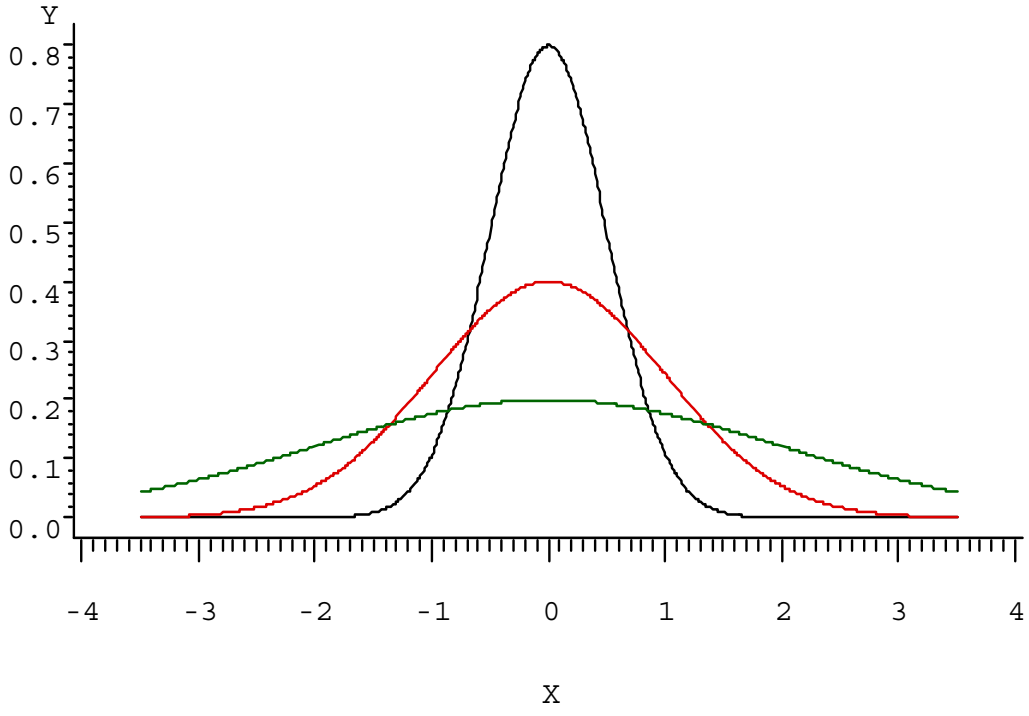
$$\frac{1}{n}\sum_i (x_i - \bar{x})^2 = s'^2$$

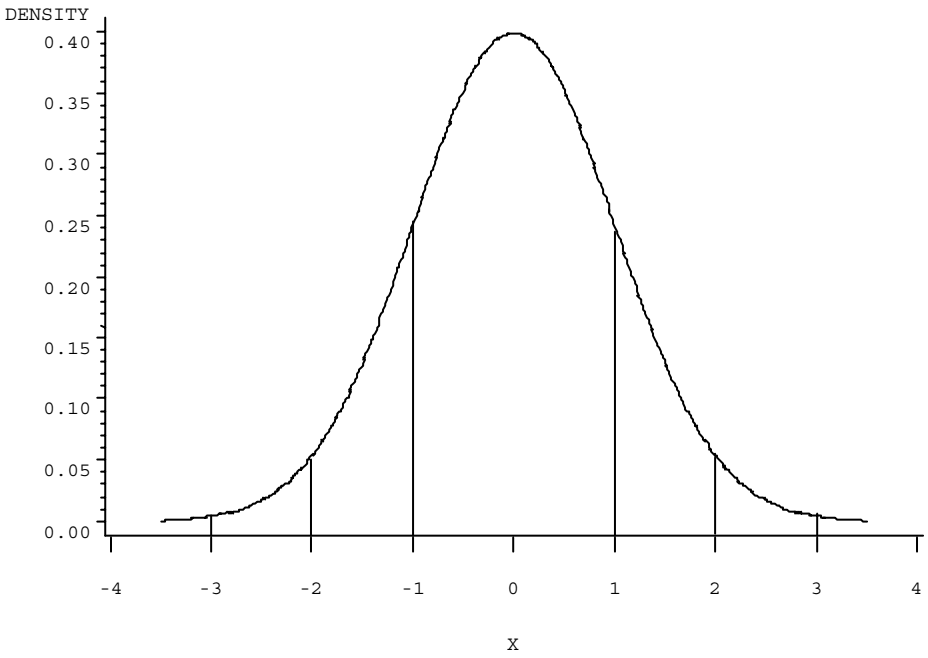
The square root of this is called the **standard deviation**:

$$s' = \sqrt{\frac{1}{n}\sum_i (x_i - \bar{x})^2}$$

For any normal distribution, the following rule holds:

- 68% of the cases fall within 1 s.d. of the mean.
- 95% fall within 2 s.d.s of the mean.
- 99.7% fall within 3 s.d.s of the mean.


Y
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4
-3
-2
-1
0
1
2
3
4
X


DENSITY
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-4
-3
-2
-1
0
1
2
3
4
X

Standard Error

Every statistic has a standard error associated with it.

- Not always reported and not always easy to calculate.
- Examples: Waiting times, companies

A measure of the (in)accuracy of the statistic.

- A standard error of 0 means that the statistic has *no* random error.
- The bigger the standard error, the less accurate the statistic.

Implicit in this the idea that anything we calculate in a sample of data is subject to random errors.

- The mean we calculated for the waiting times is not the *true* mean, but only an estimate of the true mean.
- Even if we could perfectly replicate our study, we would get a different value for the mean.

What are the sources of error?

- Classic approach in statistics: our data set may be only a random sample from some larger population.
- We may make errors of measurement.
- There are lots of other random factors affecting our outcome that we can't control.

The standard error of a statistic is the standard deviation of that statistic across hypothetical repeated samples.

- Example: 100 replications of waiting time study.
- In theory, need to replicate an infinite number of times.

The standard errors that are reported in computer output are only estimates of the true standard errors.

- Remarkably, we can estimate the variability across repeated samples by using the variability within samples.
- The more variability within the sample, the more variability between samples.
- The formula for the standard error of the mean is $\frac{s}{\sqrt{n}}$, i.e., the standard deviation divided by the square root of the sample size.

In general, the bigger the sample, the smaller the standard error.

- Why? Big samples give us more information to estimate the quantity we're interested in.
- The standard error generally goes down with the square root of the sample size. Thus, if you quadruple the sample size, you cut the standard error in half.

Confidence Intervals

The standard error is often used to construct confidence intervals.

- To construct a 95 percent confidence interval around the mean, add two standard errors and subtract two standard errors.
- E.g., for the waiting time example, the mean was approx. 20 and its standard error was 1. Then the upper confidence limit is 22 and the lower confidence limit is 18.
- Interpretation: we can be 95 percent confident that the true mean is somewhere between 18 and 22.
- Further interpretation: Suppose we could replicate our study many times. For each replication we could construct a 95 percent confidence interval by adding and subtracting 2 standard errors from the mean. Then 95 percent of those confidence intervals would contain the true mean.

Why two standard errors? Remember our rule for normal distributions: 95% of the cases fall within two standard deviations of the mean.

- Even though the original distribution of waiting times was not well approximated by a normal distribution, the distribution of means across repeated samples *is* approximately normal.

- Why? *Central limit theorem*: Whenever you average a bunch of things together, the resulting average tends to be approximately normally distributed. The more things you add together, the closer the approximation.

- In large samples, most statistics have approximately a normal distribution across repeated samples.

Correlation

A measure of the strength of the relationship between two variables.

There are many measures of correlation. The most common is Pearson's product-moment correlation coefficient, usually denoted by r.

Facts about the correlation:

- When r is 0, there is no correlation between the two variables. When r is 1 or –1, there is a perfect *linear* relationship.

- r cannot be greater than 1 or less than –1.

- The correlation measures the degree of scatter around a straight line.

- Correlation only measures the *linear* relationship between two variables.

- Correlation is symmetric: the correlation between x and y is the same as the correlation between y and x.