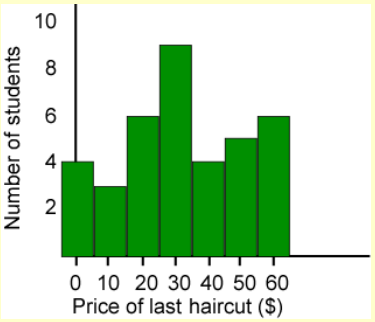
Graphs, and measures of central tendency and spread

9.07

9/13/2004

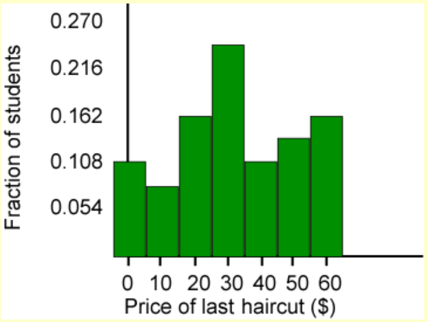
Histogram



If discrete or categorical, bars don't touch. If continuous, can touch, should if there are lots of bins.

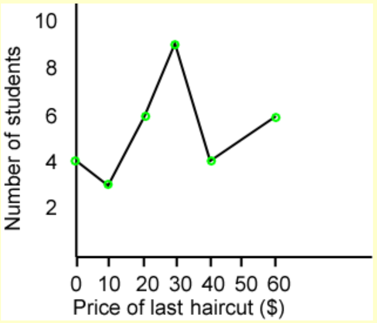
Sum of bin heights = N

Alternative: density (or "relative frequency") plot

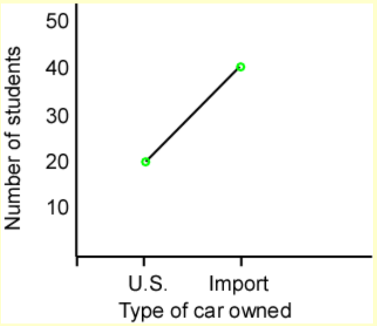


Sum of bin heights = 1

Alt: for lots of bins, continuous variable, draw histogram as a "frequency polygon"



Don’t do this for categorical variables



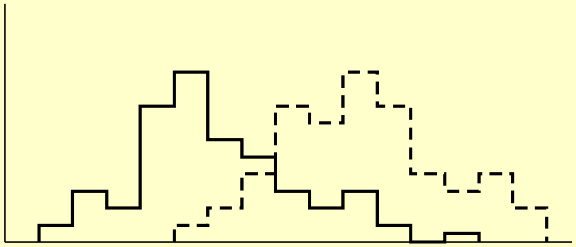
Stem-and-Leaf Plots

- A quick way of examining the distribution of data
- Like histograms on their sides, but with more information.
- Data: 7 5 5 10 11 10 10 15 15 14 20 25 20 35 35 40
- Stem-and-Leaf Plot:

```
0755                          0557
10100554                      10001455
2050            --->          2005
355                           355
40                            40
```

We can plot two histograms in the same plot, to compare them

- E.G. distribution of heights for women (solid) vs. men (dashed)

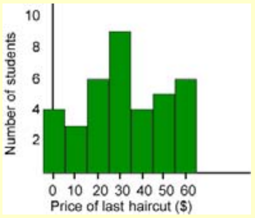


Summarizing the distribution

- You might not want to look at the whole distribution in that last case. You get lots of information about the shape of the distribution, but it’s kind of visually noisy.
- Summarize.
- Central tendency intuition: capture the impression that the heights for the men tend to be higher than the heights for the women.

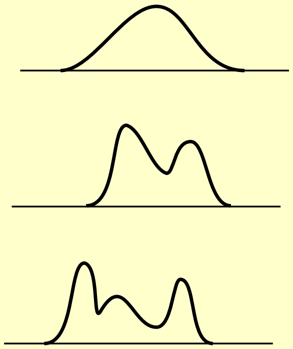
Measures of central tendency

- Mode
  - Where's the peak of the histogram?
  - Can be used for any data, but typically only used for categorical data, where the other measures we'll discuss don't apply.



Data can have one or more modes

- Unimodal vs. bimodal vs. trimodal
- Here a mode is the highest point *locally*

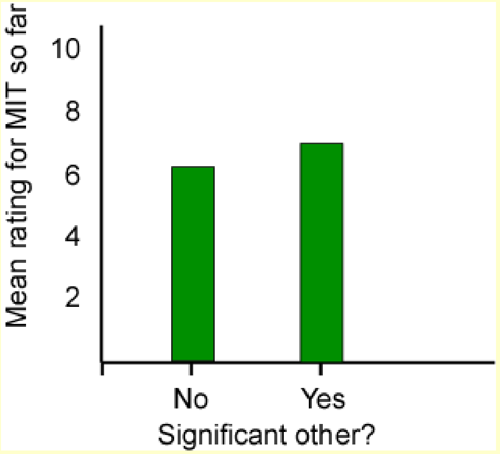


Average

- Notation: observations x={x_1,x_2,x_3,…,x_N}

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- Also known as the *mean*. But note this is the *sample mean*. Soon we will talk about the *population mean*. Don't get confused!
- Only makes sense for interval or ratio data, but will often see it used for rank-order or rating data, too



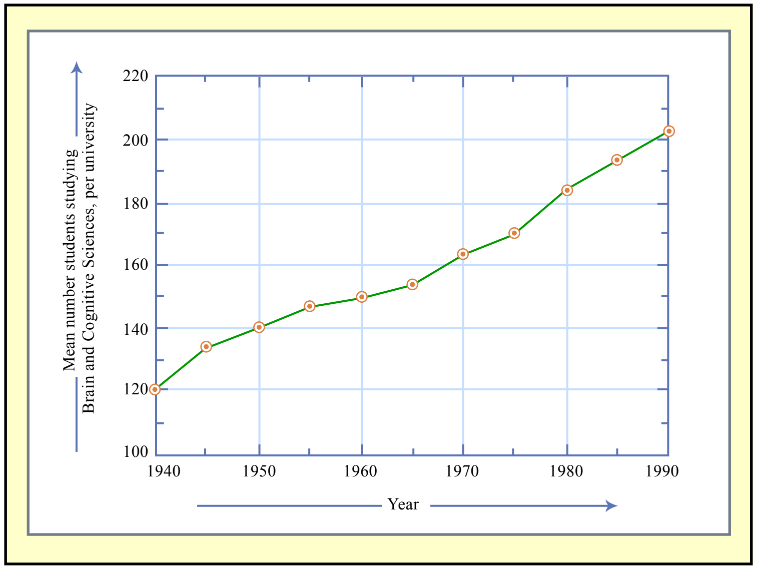
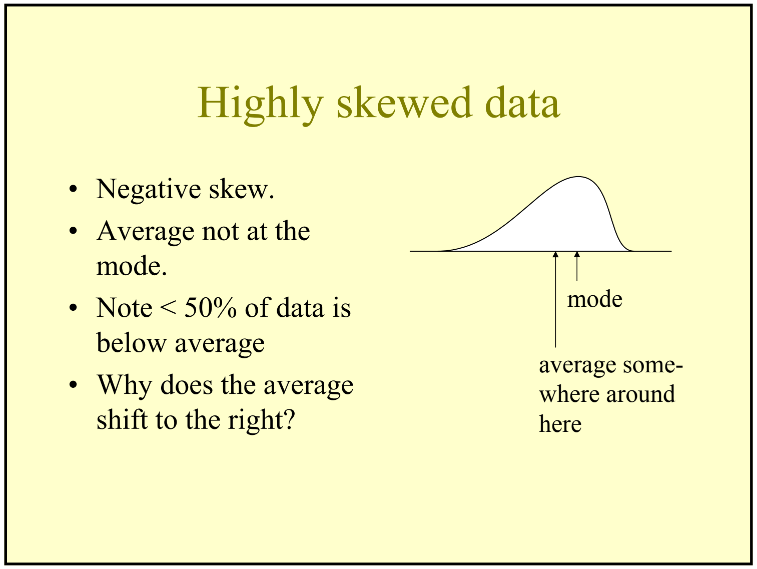


Figure by MIT OCW.



Highly skewed data

- Negative skew.
- Average not at the mode.
- Note < 50% of data is below average
- Why does the average shift to the right?

mode

average somewhere around here

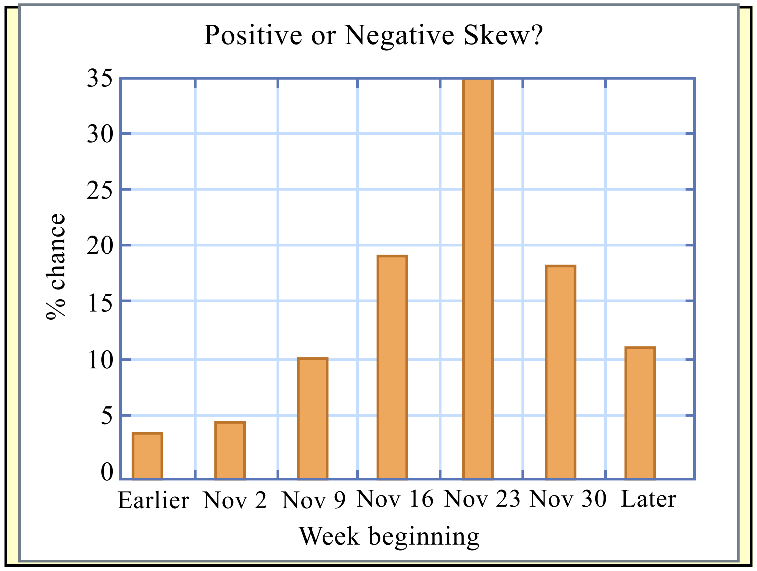
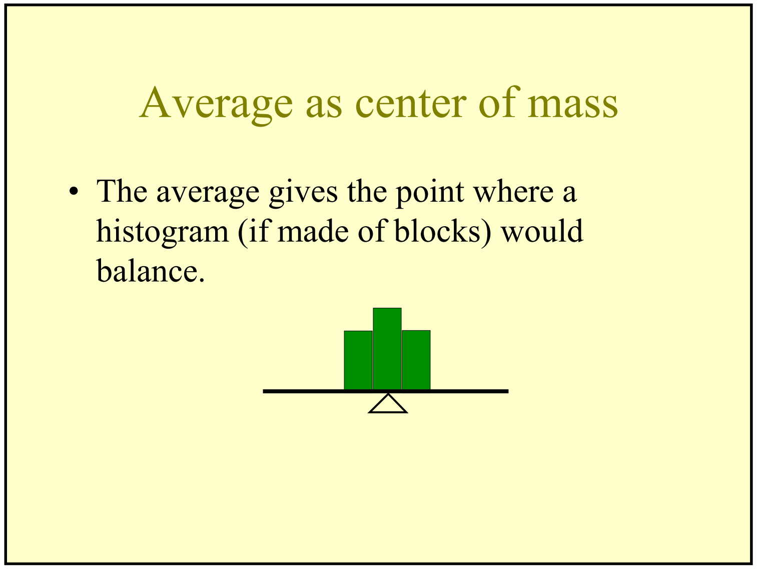


Figure by MIT OCW.

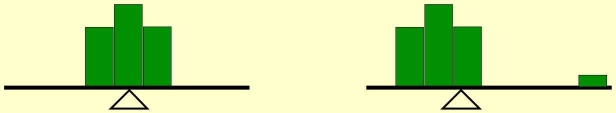


Average as center of mass

- The average gives the point where a histogram (if made of blocks) would balance.

What an outlier does to the mean

- Outlier shifts balance to right, average to right.



Another measure: the median

- Median = value with ½ of the data points to the left, ½ to the right.
- E.G.

  1, 2, 4.7, 6, 8, 9.2, 10 → median=6
- E.G.

  8, 15.2, 18, 19.2, 21.3, 25 → median=(18+19.2)/2 =18.6

The median is robust to outliers

- # of hours of TV watched per week:

  3, 5, 7, 7, 140
- Mean = 44.4!
- Median = 7.

When to use which measure?

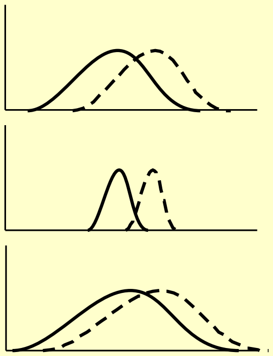
- Mode
  - Variables are categorical.
  - You want a quick and easy measure for ordinal/quantitative data.
  - You want to report the most common score
- Median
  - Variables are measured at the ordinal level.
  - Variables measured at the interval-ratio level have highly skewed distributions or lots of outliers
  - You want to report the central score. The median lies at the exact center of the distribution

When to use which measure?

- Mean
  - Quantitative variables (except for highly skewed distributions).
    - Income distributions are highly skewed – be wary of anyone talking about the "average tax cut". They should be using the median.
  - You want to report the typical score (except for highly skewed distributions). The mean is the "fulcrum that exactly balances all of the scores."
  - You anticipate additional statistical analysis

The notion of "spread"

- Plots look pretty different even when there's the same central tendencies.
- Also need to capture a notion of spread.



Measures of spread

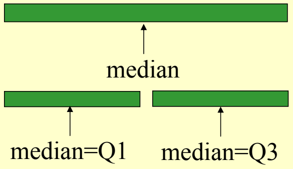
- Range
- Interquartile range
- Standard deviation

Range

- Biggest value observed – smallest value observed.
- Pretty sensitive to outliers in the tails of the distribution.

Interquartile range

- Divide data into 4 groups, see how far about the extreme groups are.



median

median=Q1

median=Q3

- Q3-Q1 = IQR.

Box-and-whiskers plot

- Draw box ends = Q1 & Q3
- Draw median through box.
- Plot as points the outliers = any points > 1.5 IQR from an end of the box.
- Plot "whiskers" out to the farthest points that are not outliers.

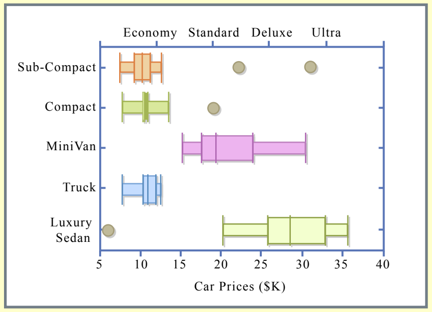


Figure by MIT OCW.

Standard deviation

- Measures spread about average.
- Desired: How far are data points from the average, on average?
- But,

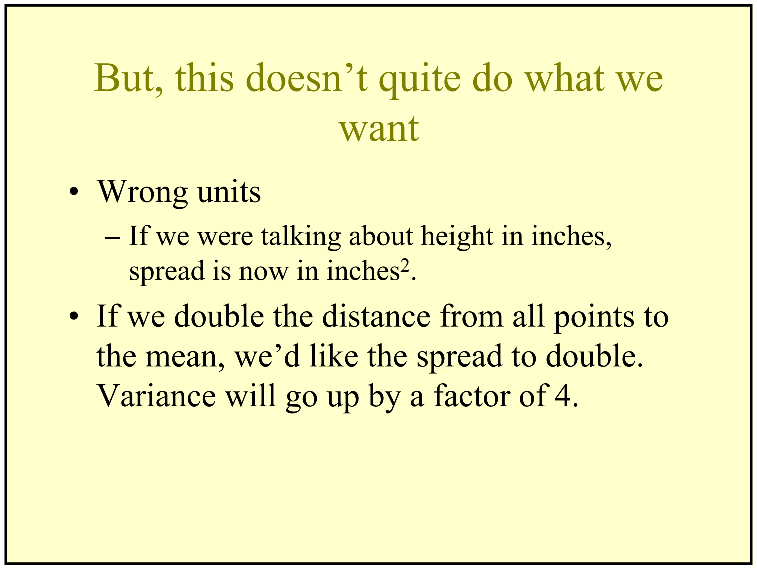
$$\frac{\sum_{i=1}^{N}(x_i-\overline{x})}{N} = \frac{\sum x_i}{N} - \frac{\sum \overline{x}}{N} = \overline{x} - \frac{N\overline{x}}{N} = 0$$

- Clearly, this is not going to be useful…

Try squaring the difference, before taking the average

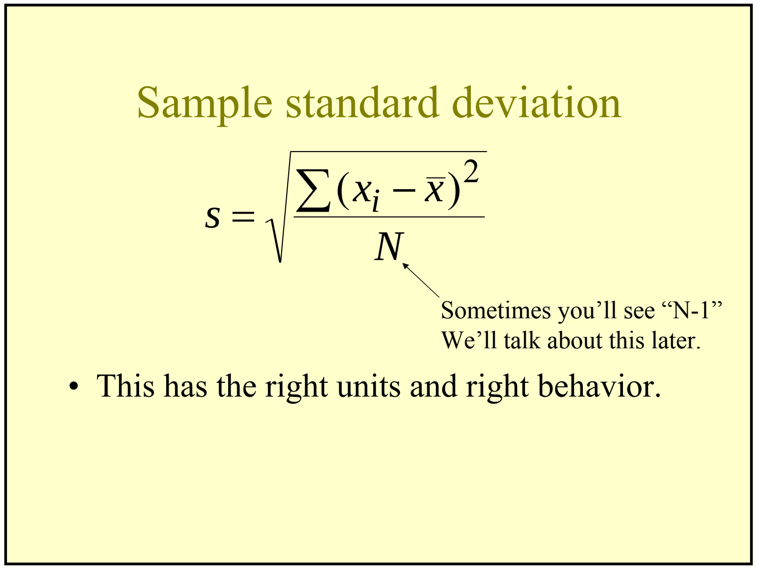
- Sample variance =

$$s^2 = \frac{\sum_{i=1}^{N}(x_i-\overline{x})^2}{N} = \frac{\sum x_i^2}{N} - \overline{x}^2$$



But, this doesn't quite do what we want

- Wrong units
  - If we were talking about height in inches, spread is now in inches^2.
- If we double the distance from all points to the mean, we'd like the spread to double. Variance will go up by a factor of 4.

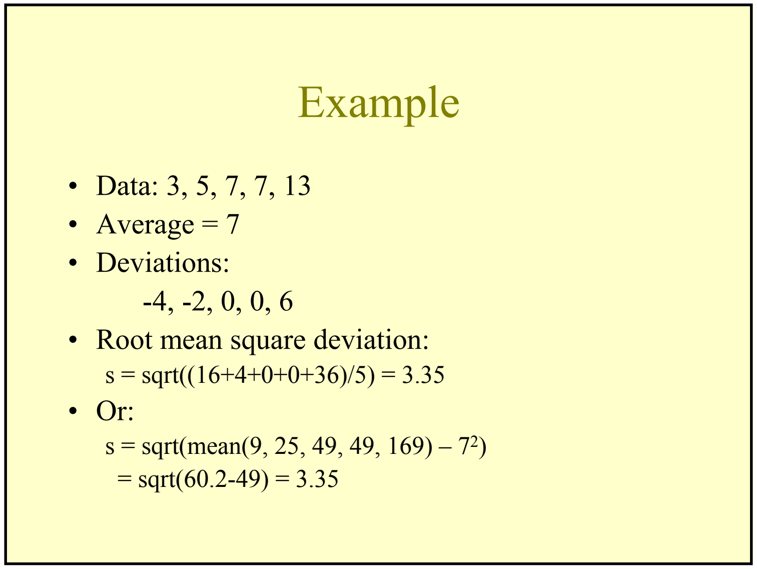


Sample standard deviation

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}}$$

Sometimes you'll see "N-1"
We'll talk about this later.

- This has the right units and right behavior.



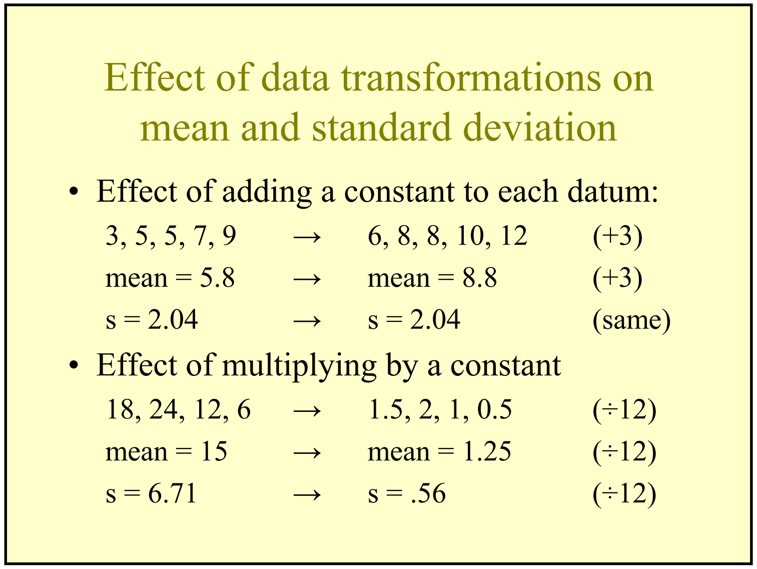
Example

- Data: 3, 5, 7, 7, 13
- Average = 7
- Deviations:

  -4, -2, 0, 0, 6
- Root mean square deviation:

  s = sqrt((16+4+0+0+36)/5) = 3.35
- Or:

  s = sqrt(mean(9, 25, 49, 49, 169) – 7^2)
  = sqrt(60.2-49) = 3.35



Effect of data transformations on mean and standard deviation

- Effect of adding a constant to each datum:

3, 5, 5, 7, 9	→	6, 8, 8, 10, 12	(+3)
mean = 5.8	→	mean = 8.8	(+3)
s = 2.04	→	s = 2.04	(same)

- Effect of multiplying by a constant

18, 24, 12, 6	→	1.5, 2, 1, 0.5	(÷12)
mean = 15	→	mean = 1.25	(÷12)
s = 6.71	→	s = .56	(÷12)

Effect of transformation on mean and standard deviation

- Adding a constant shifts the mean by that amount, and leaves the standard deviation unchanged.
- Multiplying by a constant multiplies both the mean and standard deviation by that constant.

The z-score is robust to these changes

$$z_i = \frac{(x_i - \bar{x})}{s}$$

How many standard deviations above the mean is score x_i?

- x = 18, 24, 12, 6 → x = 1.5, 2, 1, 0.5
- z = .4, 1.3, -.4,-1.3 → z = .4, 1.3, -.4, -1.3
- We'll use this transformation to z a lot in this class…

Example use of z-scores

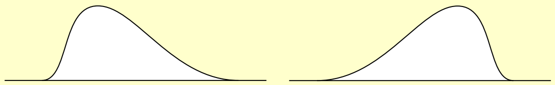
- How do you combine scores from different people?
- Mary, Jeff, and Raul see a movie.
- Their ratings of the movie, on a scale from 1 to 10:
  - Mary = 7, Jeff = 9, Raul = 5
- Average score = 7?
- It's more meaningful to see how those scores compare to how they typically rate movies.

Example use of z-scores

- Recent ratings from Mary, Jeff, & Raul:
  - Mary: 7, 8, 7, 9, 8, 9 — 7 is pretty low…
  - Jeff: 8, 9, 9, 10, 8, 10 — 9 is average…
  - Raul: 1, 2, 2, 4, 5 — 5 is pretty good…
- z-scores:
  - Mary: m=8, s=.8 — z(7) = -1.2
  - Jeff: m=9 — z(9) = 0
  - Raul: m=2.8, s=1.5 — z(5) = 1.5
- mean(z) = 0.1 = # of standard deviations above the mean
  - It's probably your average movie, nothing outstanding.

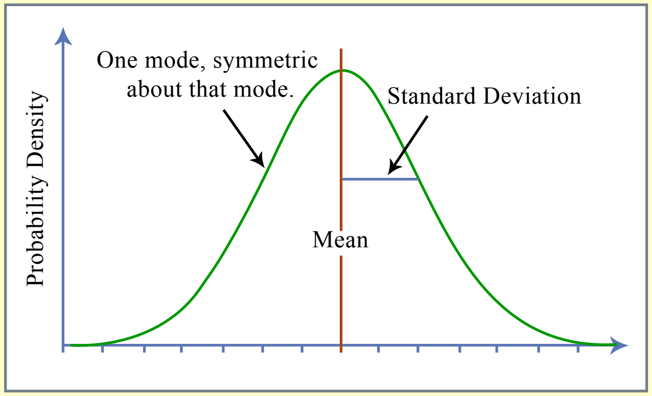
Describing Distributions

- We could use more parameters to describe the distribution. E.G. skew:



- For a normal distribution only two parameters are necessary to completely specify the distribution.
  - Mean (average)
  - Standard deviation

Normal Distribution



$$p(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

Figure by MIT OCW.

Note: About your book!

- MANY of the pictures of normal distributions in your book are *not normal*!!!
- They seem to be splines -- someone’s attempt to draw a normal distribution using a graphics program.
- Normal distributions really look like the picture on the previous slide.

Normal distribution

- If x is distributed according to a normal distribution with mean μ and standard deviation σ, we write:

  $x \sim N(\mu, \sigma)$ or $x \sim N(\mu, \sigma^2)$

- Notation:

sample (Roman)	population (Greek)
$\overline{x}$, s or m, s	μ, σ

Some Terminology

- Density curve to describe a population
  - Area under a density curve = 1
  - "Density" because the actual probability of any value is zero
    - There are an infinite number of events -- the probability of any one of them is 0.

Some things that might be normally distributed

- Heights and weights
- Arrival times to class
- Exam scores
- Gas mileage
- Sizes of anything that grows
- Number of raisins in a box of *Raisin Bran*
- Lots of phenomena have approx. this distribution – later we'll talk about one reason why.

Mean

- Describes the center of a normal distribution
- Because the normal distribution is symmetrical, the median and mean are equal
- Changing the mean just changes the position of the distribution, not the shape

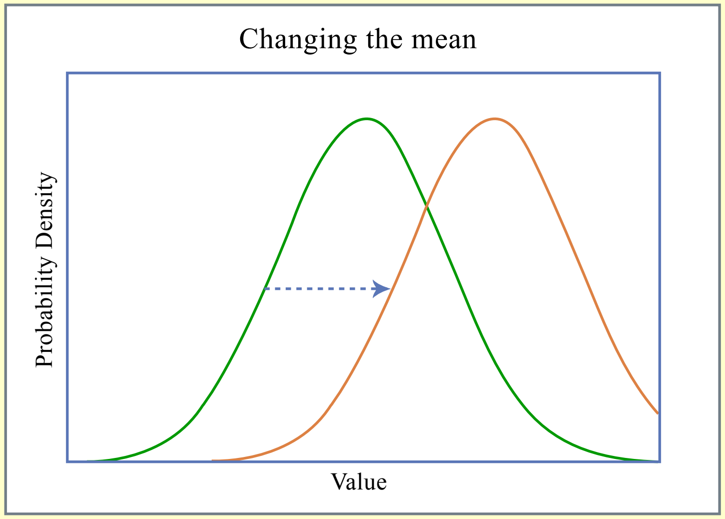


Figure by MIT OCW.

Standard Deviation

- Describes the width (spread) of the distribution
- Can be read directly from the density curve at the inflection point (point where the curvature changes)
- Changing the standard deviation changes the width and height of a normal curve, but not the position

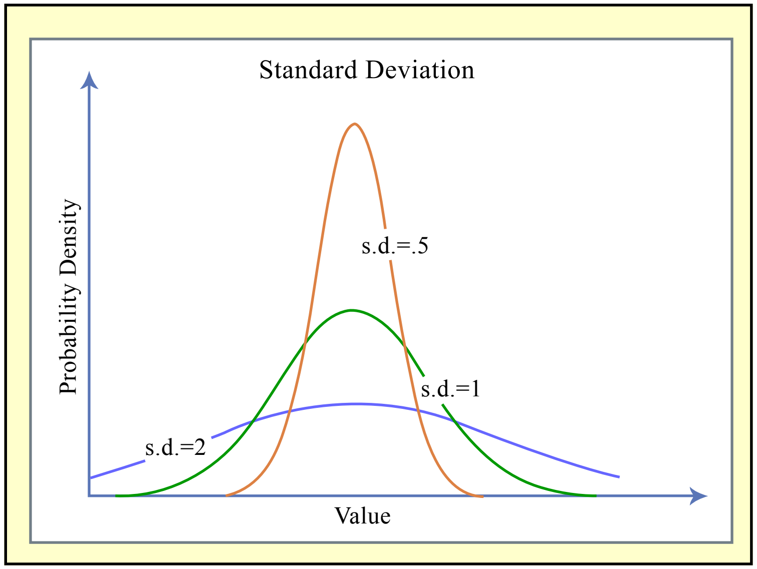


Figure by MIT OCW.

Using the Standard Deviation

- The 68-95-99.7 rule for **any** normal distribution
- 68% of the population fall within one standard deviation of the mean
- 95% of the population fall within two standard deviations of the mean
- 99.7% of the population fall within three standard deviations of the mean

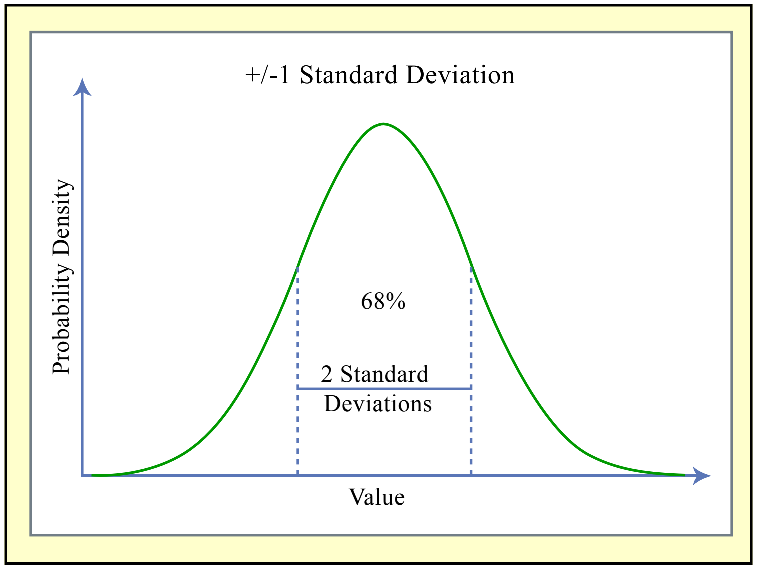


Figure by MIT OCW.

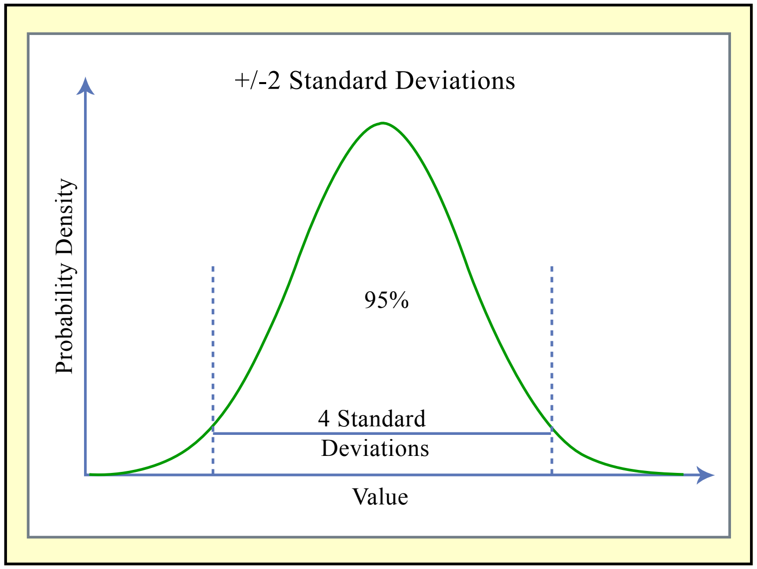


Figure by MIT OCW.

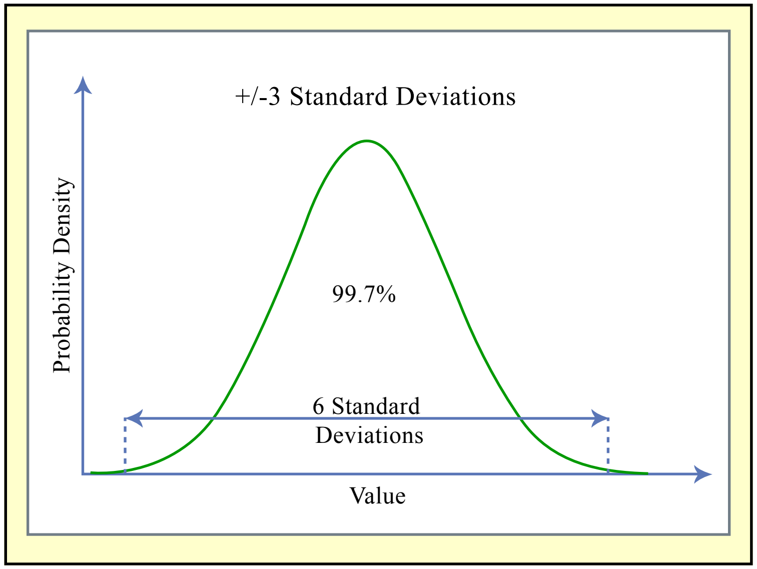


Figure by MIT OCW.

The normal curve approximation for data

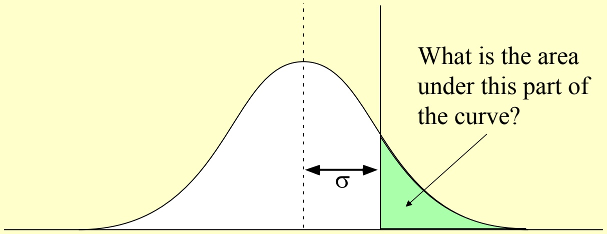
- For many types of data, the normal curve is a good approximation to the distribution of the data.
- Use this approximation to generalize from the sample to the larger population.
- First:
  - Estimate the population mean, μ, from the sample mean $\bar{x}$.
  - Estimate the population std. deviation, σ, from the sample standard deviation, s.

Example question you can ask

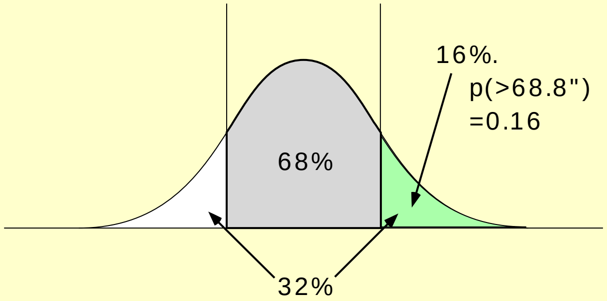
- Mean height for this class is 65.7 inches.
- Std. dev. = 3.1 inches
- Assume the broader population we're interested in is MIT students in general.
- What's the probability that a student will be taller than 68.8 inches?

Using what we know about normal distributions

- Well, 68.8” is one std. dev. above the mean height.



What is the area under this part of the curve?



16%.
p(>68.8")
=0.16

One little detail…

- s^2 is not a good estimate for σ^2, for small sample size N. It is *biased*, i.e. it is consistently too small.
- Why: we compute $\overline{x}$ and s^2 at the same time. $\overline{x}$ follows the samples around, so the deviation from $\overline{x}$ tends to be smaller than the population σ^2.
- Fix: *when estimating* σ^2, $s^2 = \frac{(x-\overline{x})^2}{N-1}$

Makes it a little bigger…