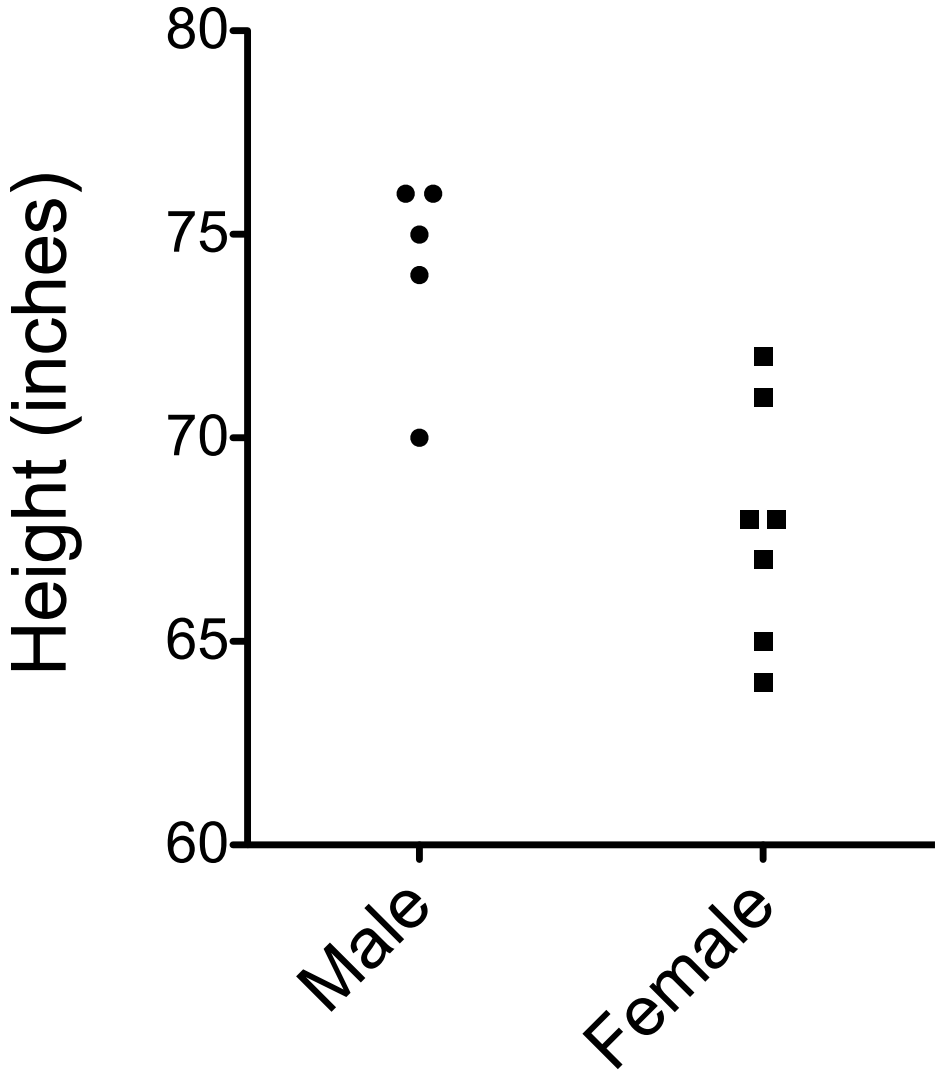
Introductory statistics for biological engineers

Module 2, Lecture 7
20.109 Spring 2014
Agi Stachowiak, Shannon Hughes, and Leona Samson
Lecture by Zac Nagel, 4/8/14

What is an experiment?

- A test of a hypothesis
- A comparison of two conditions
- Seeing what happens when we change a variable
- Asking a question about the world

Typical Experimental Data:



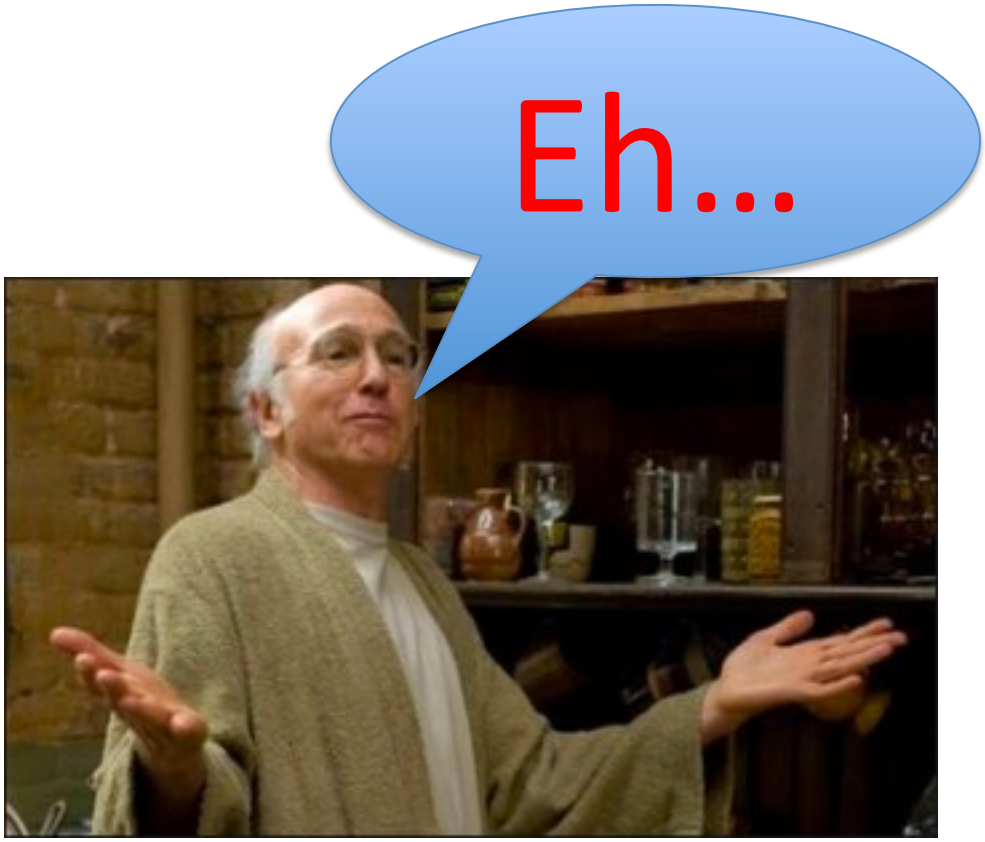
Did my experiment tell me anything?

- Did I get an answer?
- How sure am I of the answer?
- Can we all agree on how sure of the answer we are?

Experimental data speak with limited authority

http://jimbuchan.com/

Experimental data speak with limited authority



http://tvbythenumbers.zap2it.com/

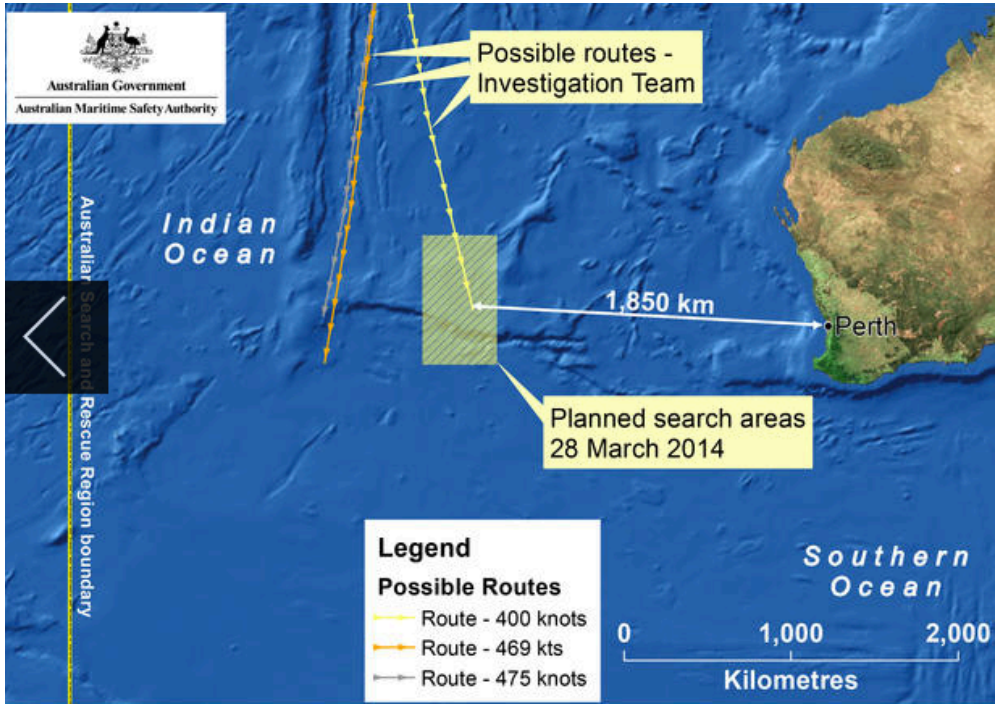
All Science is an approximation

http://www.harborfreight.com/

- Measure twice, cut once

http://www.neontommy.com/

All Science is an approximation

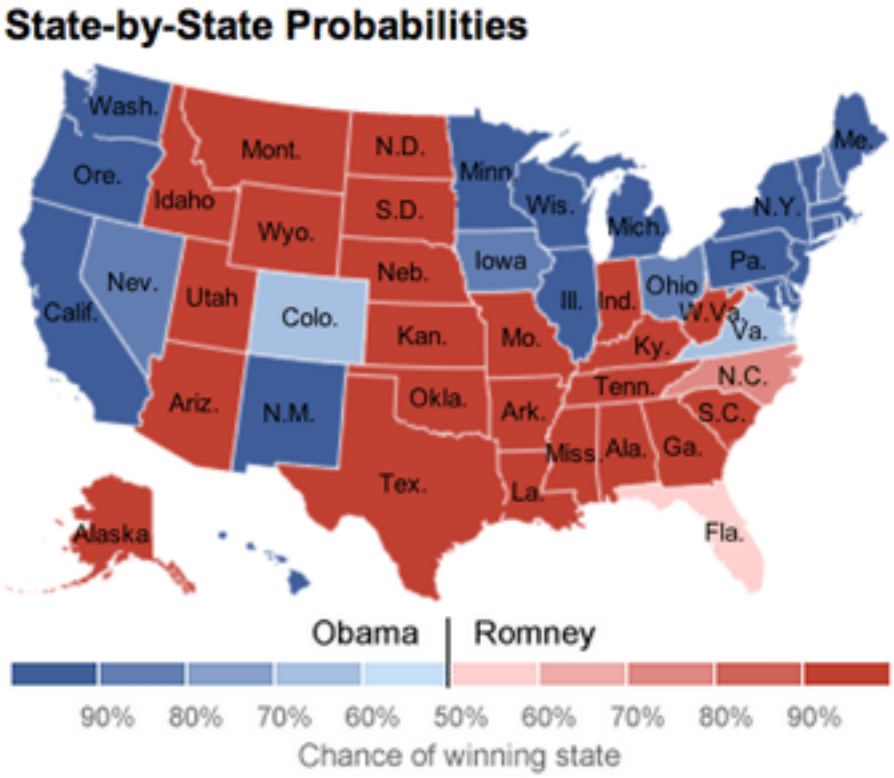


- Measure **even more than twice**? Why?
- → **We never have all the data!**

Very limited radar, satelite and blackbox pinger data had been used (as of the date of this lecture) together with mathematical modeling to obtain a sort of confidence interval delineating the most likely location of Malasia Airlines flight 370.

http://www.cbsnews.com/news/malaysia-airlines-flight-370-no-new-black-box-like-sounds-heard/

All Science is an approximation

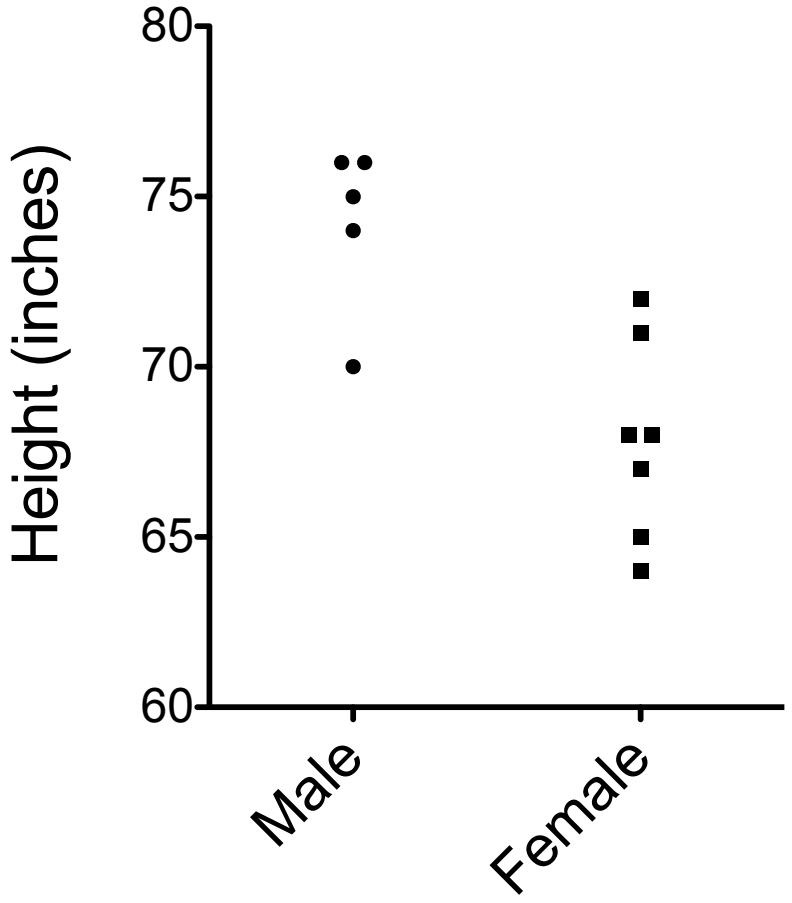


http://www.laprogressive.com/election-2012-watch/

- Measure **even more than twice**? Why?
- → **We almost never have all the data!**

Nate Silver has used statistics to predict the most likely outcome in elections with great success; at left, his last prediction before the 2012 presidential election. He happened to get 50/50 this time: red states to Mitt Romney, blue states to Barack Obama. Note the predictions amounted to probabilities based on a sample of polling data.

Can a sample of 12 people be used to draw conclusions about the global human population?

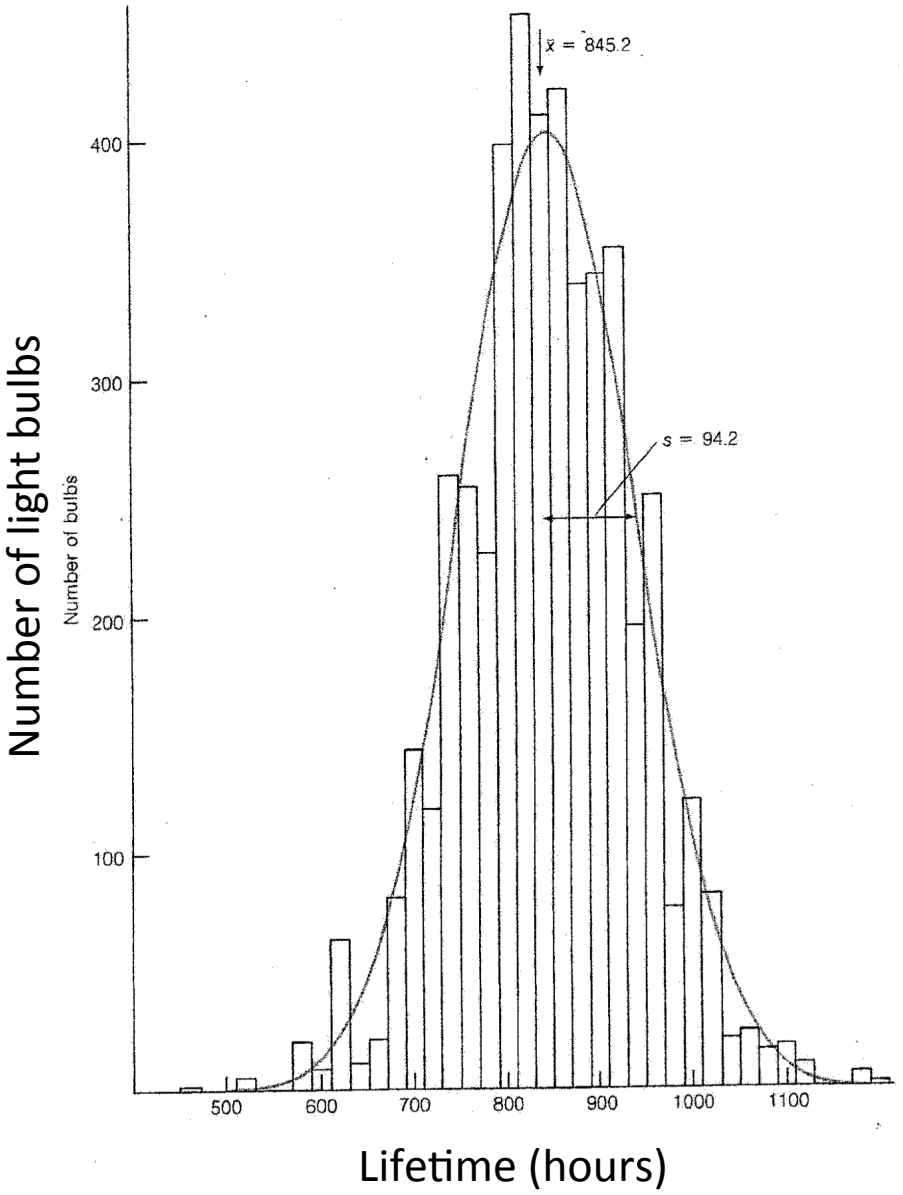


Statistics is a powerful tool for deciding what can be concluded from data

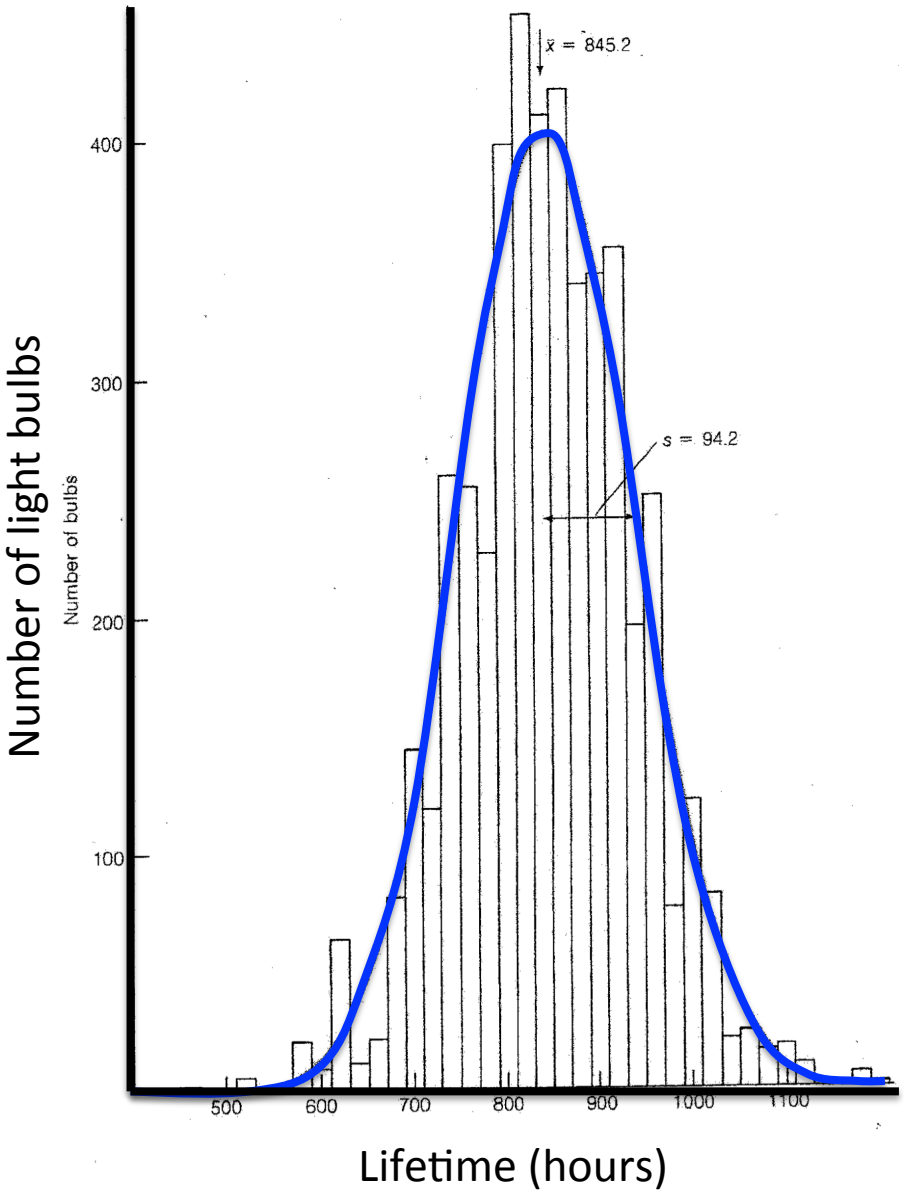
An example: how long does a light bulb last?

http://www.clipartbest.com/clipart-Kijxd4riq

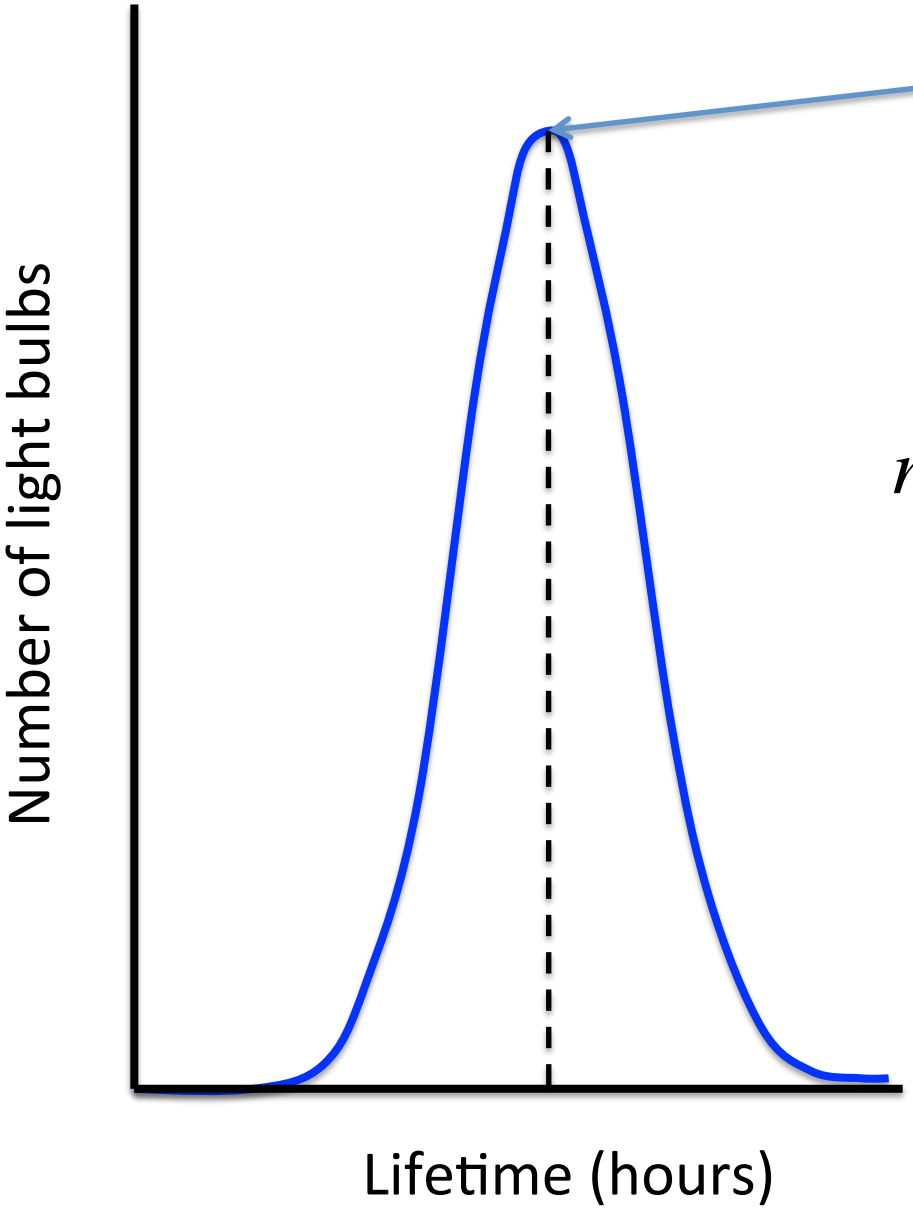
What real data look like:



What real data look like:



Mean is the average lifetime:



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

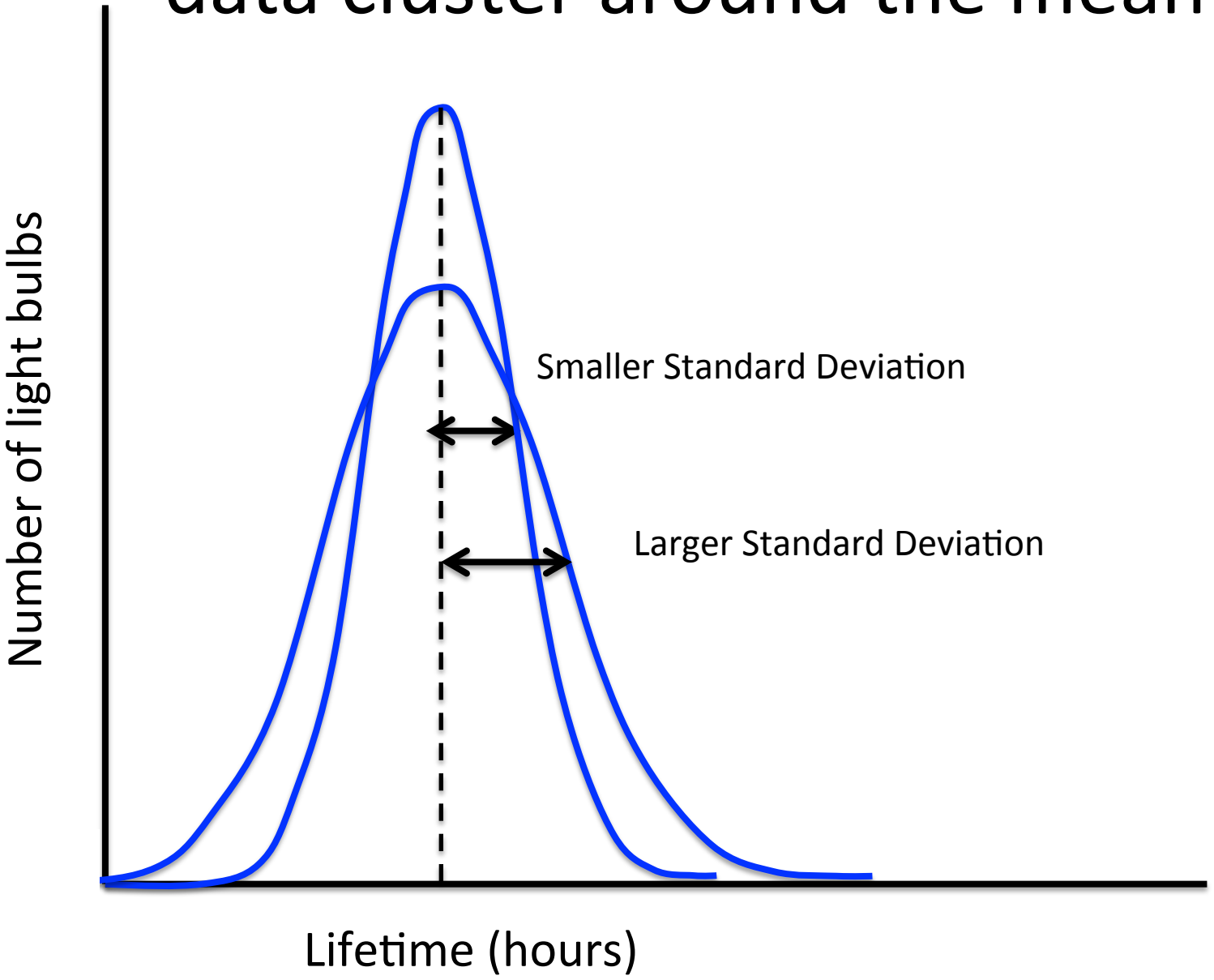
Where:

$\bar{x}$ = Average value for xi

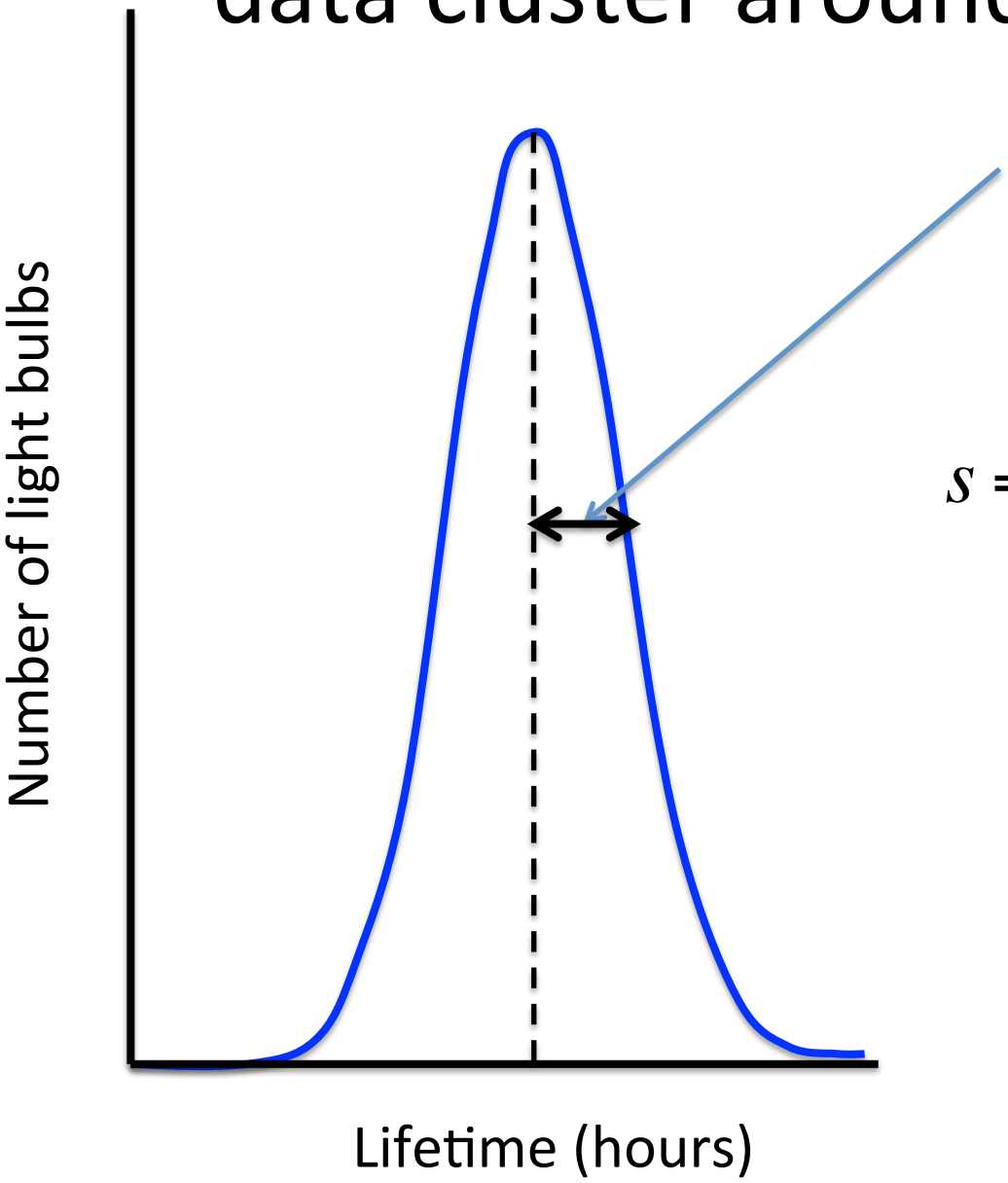
x_i = Lifetime for a particular bulb

n = Total number of bulbs

Standard deviation measures how closely data cluster around the mean



Standard deviation measures how closely data cluster around the mean



Standard Deviation

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

Where:

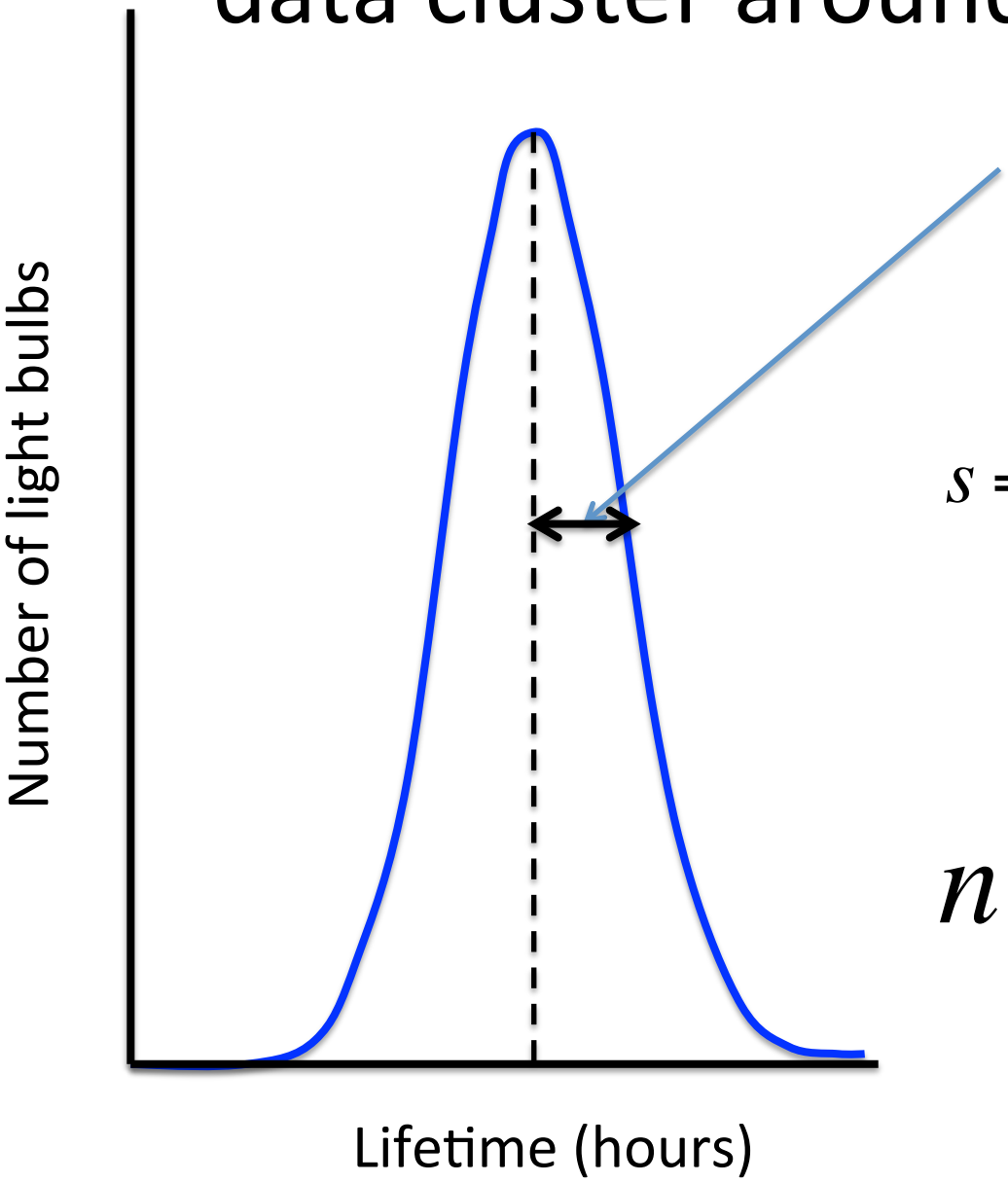
$s =$ Standard deviation

$\overline{x} =$ Average value for xi

$x_i =$ Lifetime for a particular bulb

$n =$ Total number of bulbs

Standard deviation measures how closely data cluster around the mean



Standard Deviation

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

$$n - 1 = \text{Degrees of freedom}$$

If we *had* all the data

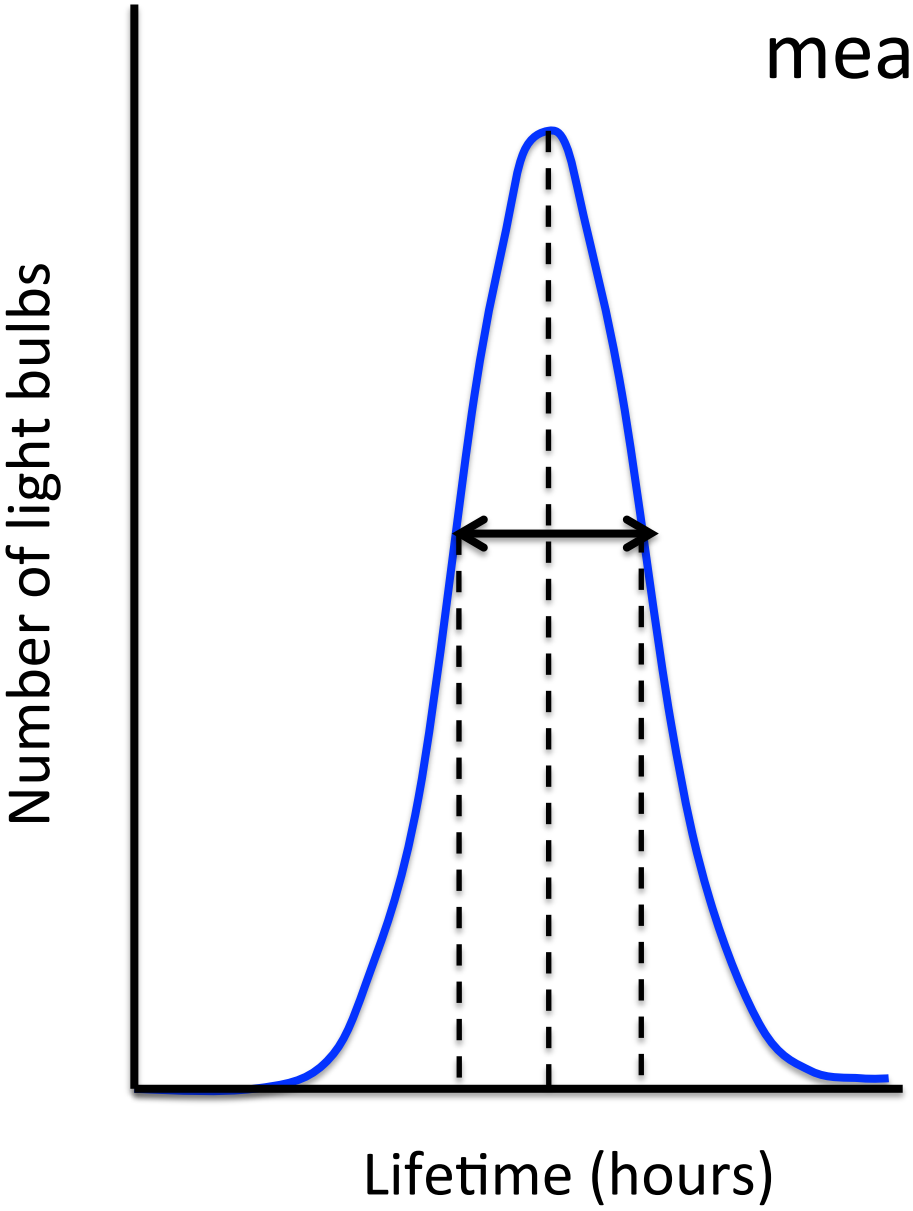
$\overline{x}$ → True mean, mu

s → True Standard deviation, sigma

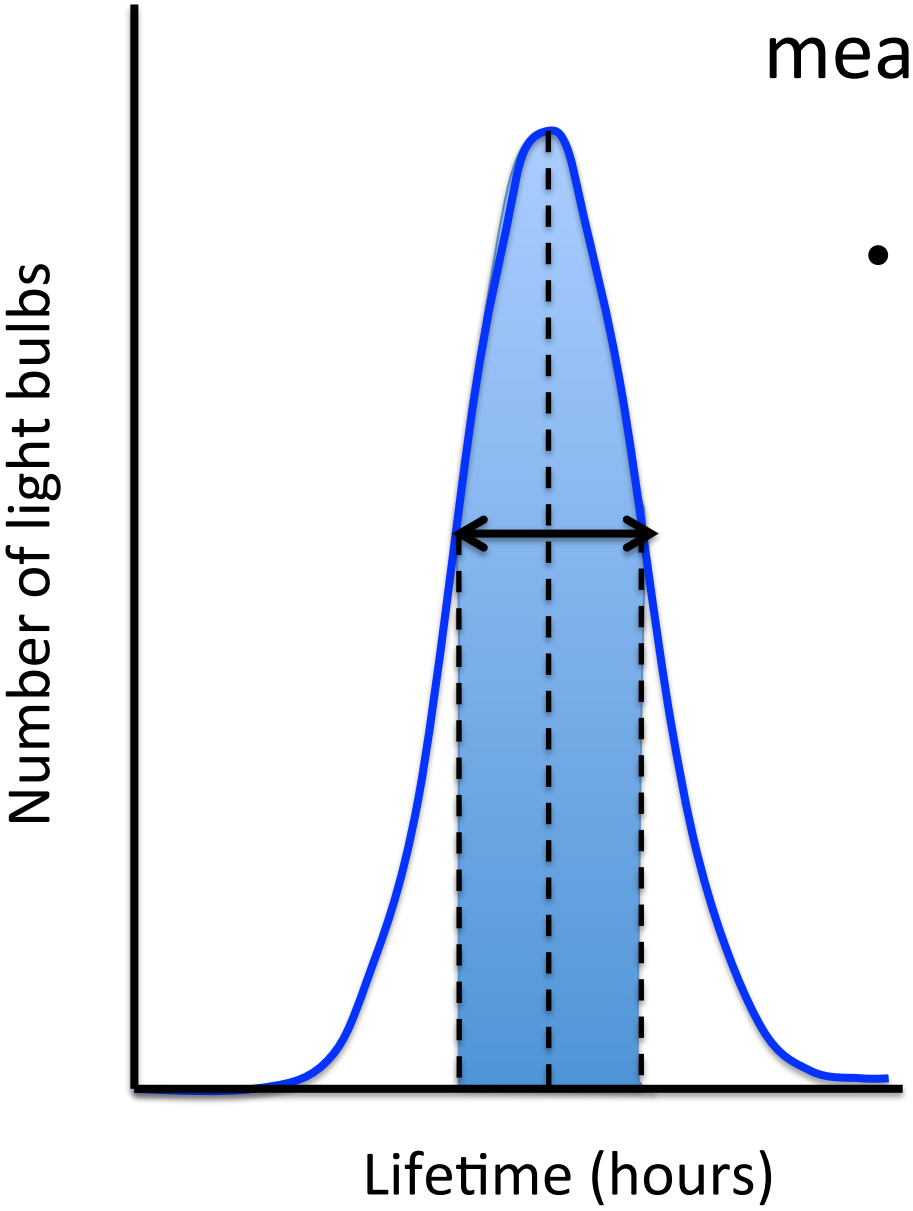
Since we almost never have all the data, how confident are we about our estimate of the mean?

Confidence Interval: a range of values within which there is a specified probability of finding the true mean

Confidence Interval: a range of values within which there is a specified probability of finding the true mean



Confidence Interval: a range of values within which there is a specified probability of finding the true mean



- For a Gaussian curve, a 68.3% of the data fall within 1 sigma of the mean, mu

Confidence Interval: a range of values within which there is a specified probability of finding the true mean

$$\mu = \bar{x} \pm \frac{ts}{\sqrt{n}}$$

$\mu =$ The true mean

$t =$ Student's t

$s =$ Standard deviation

$n =$ Number of observations

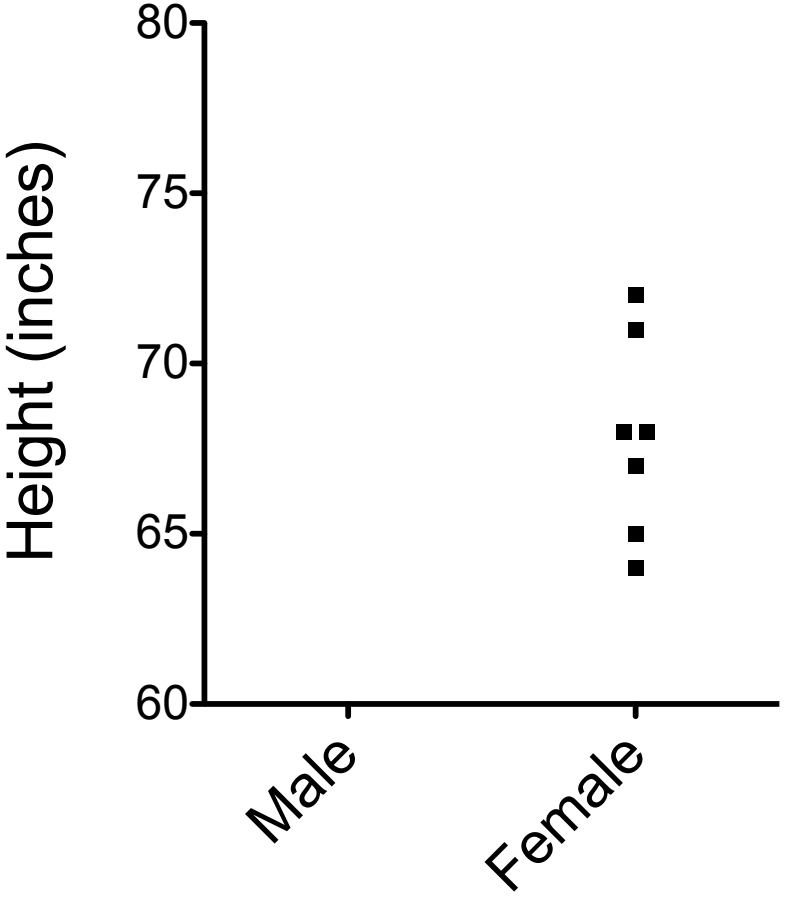
Where do we get t?

TABLE 4-2 **Values of Student's *t***

	Confidence level (%)					
Degrees of freedom	*50*	*90*	*95*	*98*	*99*	*99.5*
1	1.000	6.314	12.706	31.821	63.657	127.32
2	0.816	2.920	4.303	6.965	9.925	14.089
3	0.765	2.353	3.182	4.541	5.841	7.453
4	0.741	2.132	2.776	3.747	4.604	5.598
5	0.727	2.015	2.571	3.365	4.032	4.773
6	0.718	1.943	2.447	3.143	3.707	4.317
7	0.711	1.895	2.365	2.998	3.500	4.029
8	0.706	1.860	2.306	2.896	3.355	3.832
9	0.703	1.833	2.262	2.821	3.250	3.690
10	0.700	1.812	2.228	2.764	3.169	3.581
15	0.691	1.753	2.131	2.602	2.947	3.252
20	0.687	1.725	2.086	2.528	2.845	3.153
25	0.684	1.708	2.068	2.485	2.787	3.078
30	0.683	1.697	2.042	2.457	2.750	3.030
40	0.681	1.684	2.021	2.423	2.704	2.971
60	0.679	1.671	2.000	2.390	2.660	2.915
120	0.677	1.658	1.980	2.358	2.617	2.860
∞	0.674	1.645	1.960	2.326	2.576	2.807

Note: In calculating confidence intervals, σ may be substituted for s in Equation 4-3 if you have a great deal of experience with a partici and have therefore determined its "true" population standard deviation. If σ is used instead of s, the value of t to use in Equation 4-3 cor the bottom row of Table 4-2.

Confidence Interval Example:



$$\mu = \bar{x} \pm \frac{ts}{\sqrt{n}}$$

$\bar{x} = 67.9$

$s = 2.9$

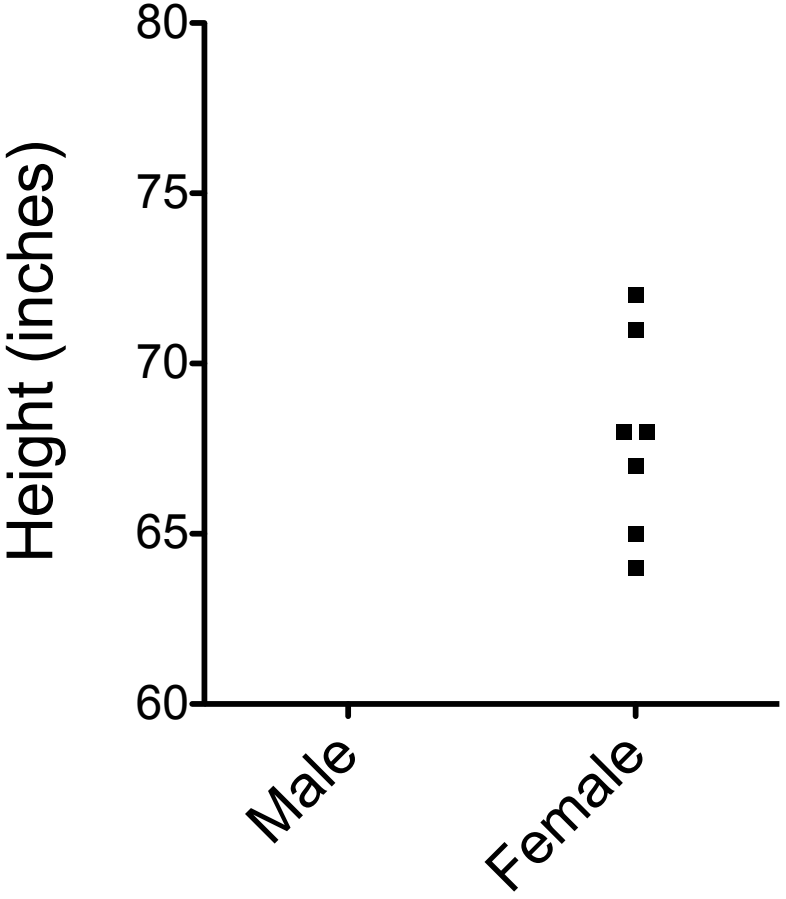
$n = 7$

$DOF = 6$

$t_{50\%} = 0.718$

Confidence Interval Example:

$$\mu = \bar{x} \pm \frac{ts}{\sqrt{n}}$$



$$\mu_{50\%} = 67.9 \pm 0.8$$

$\bar{x} = 67.9$

$s = 2.9$

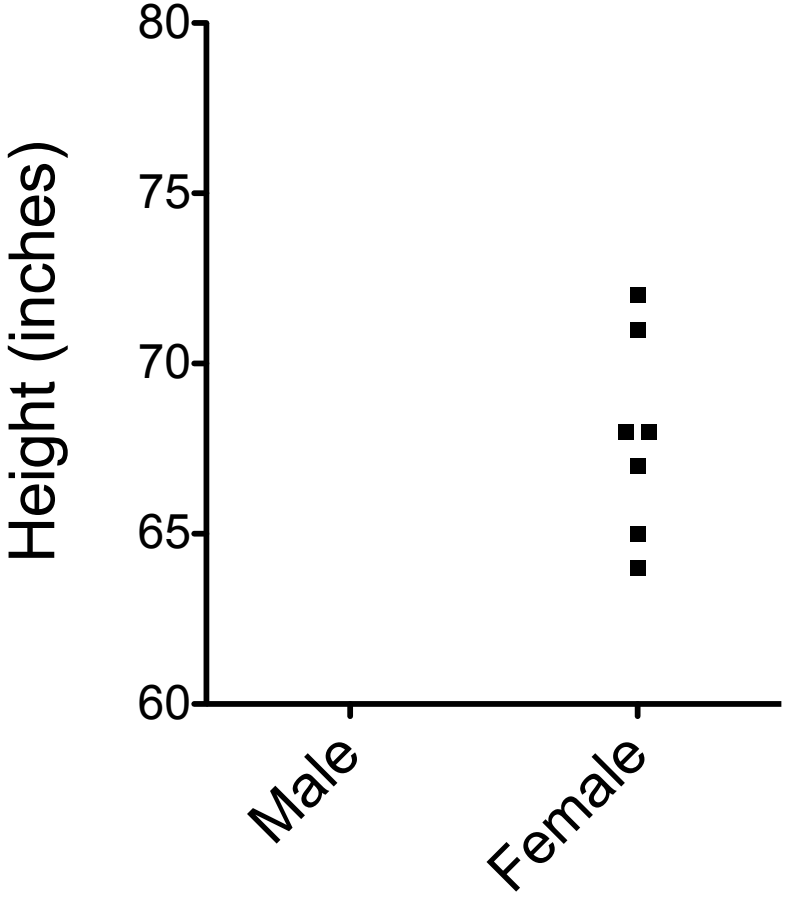
$n = 7$

$DOF = 6$

$t_{50\%} = 0.718$

99.9% Confidence interval is larger:

$$\mu = \bar{x} \pm \frac{ts}{\sqrt{n}}$$



$$\mu_{99.9\%} = 67.9 \pm 6.5$$

$$\bar{x} = 67.9$$

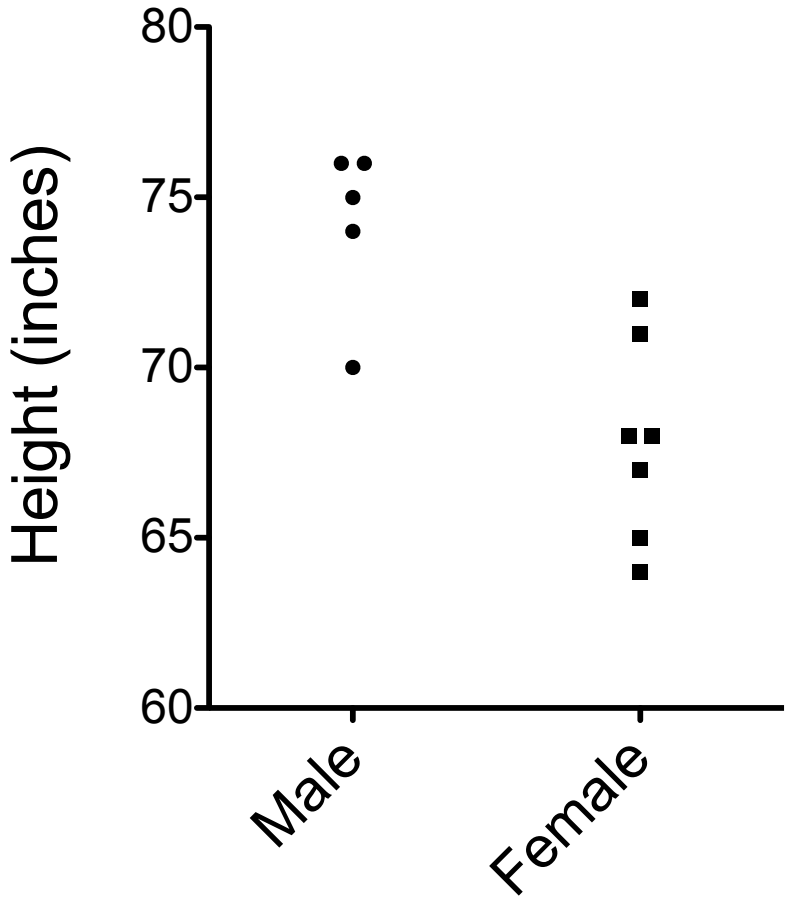
$$s = 2.9$$

$$n = 7$$

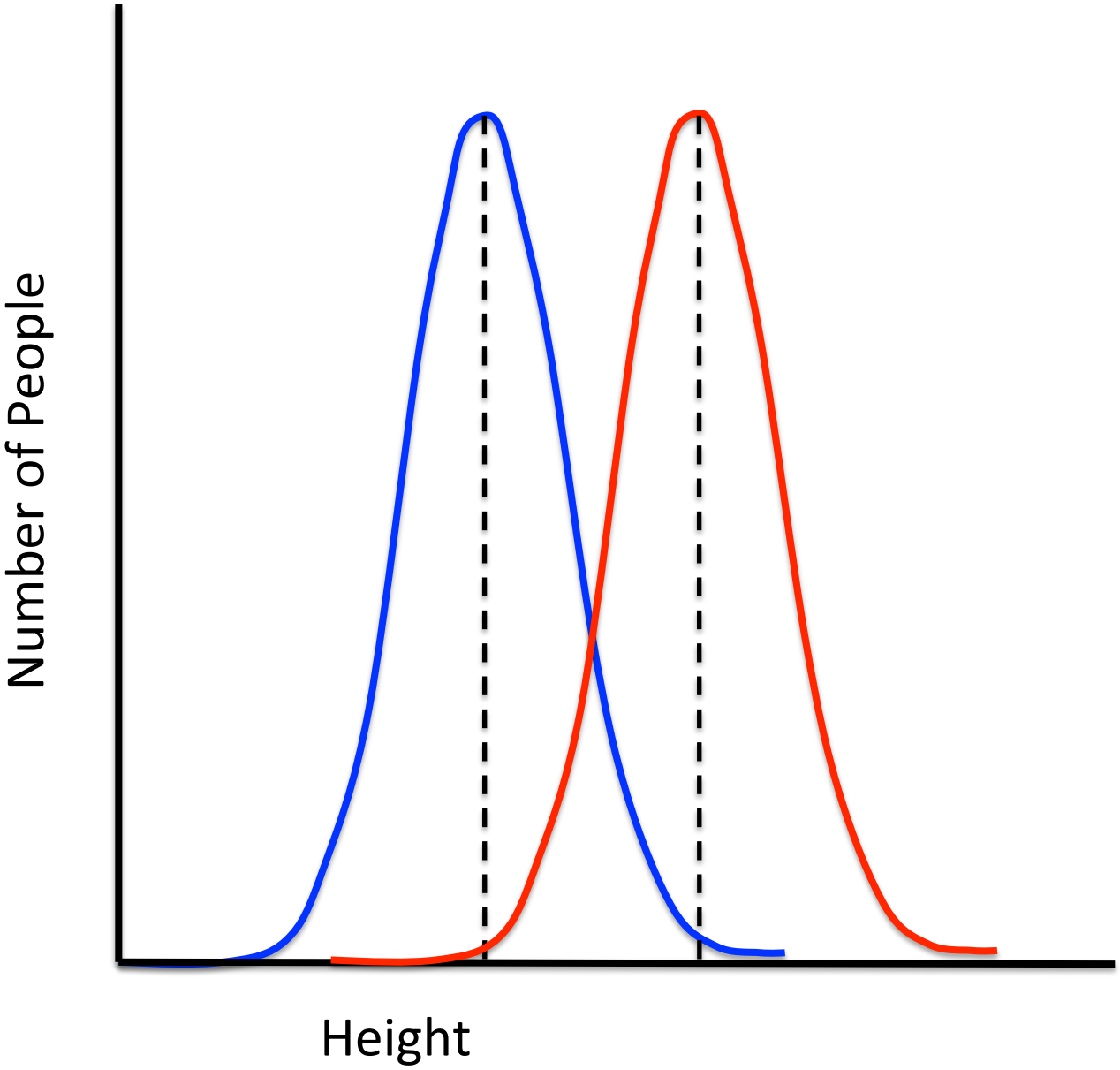
$$DOF = 6$$

$$t_{99.9\%} = 5.959$$

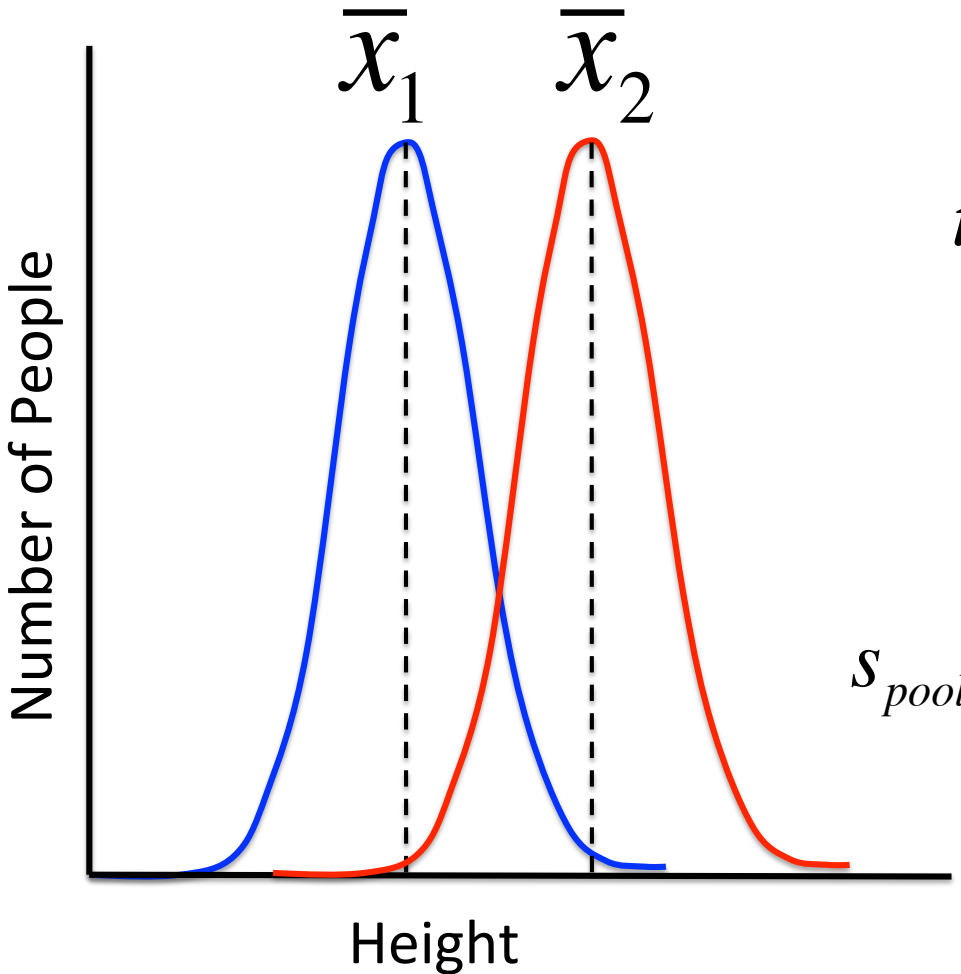
How do we know whether two means are different from one another?



How do we know whether two means are different from one another?



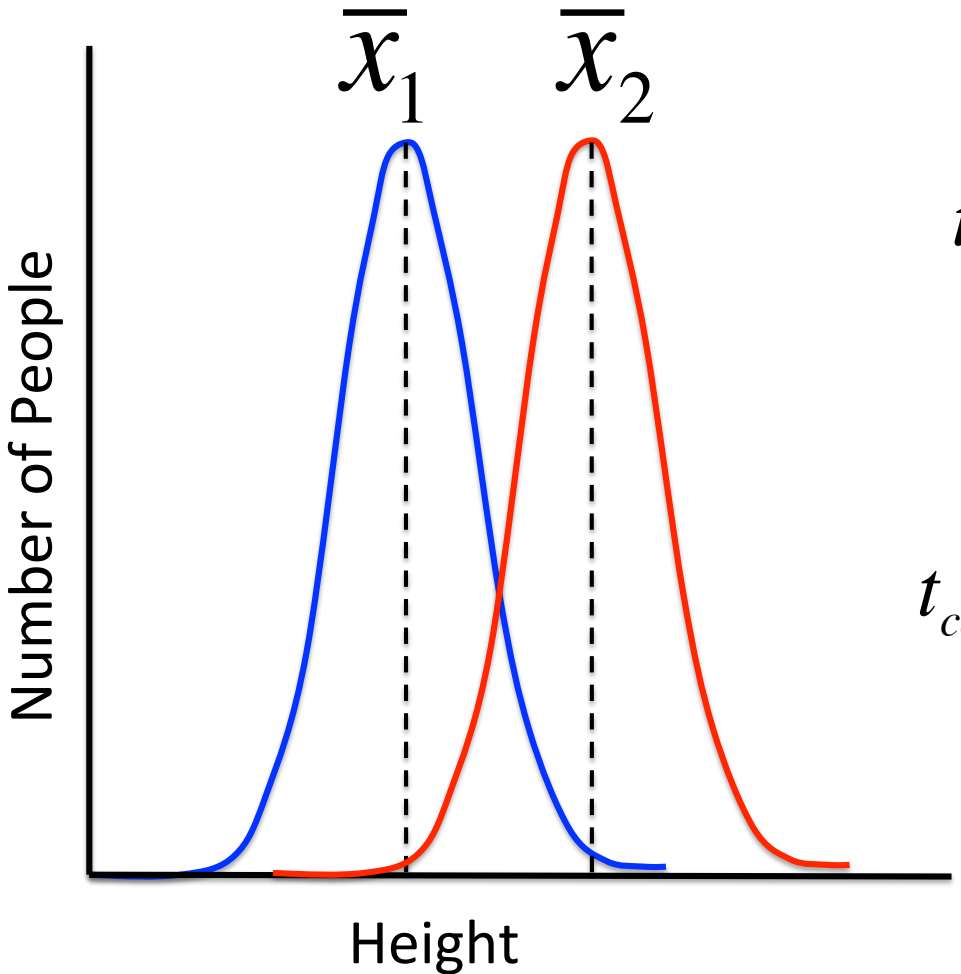
T-test for comparison of means:



$$t_{calc} = \frac{\left|\overline{x}_1 - \overline{x}_2\right|}{s_{pooled}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$s_{pooled} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

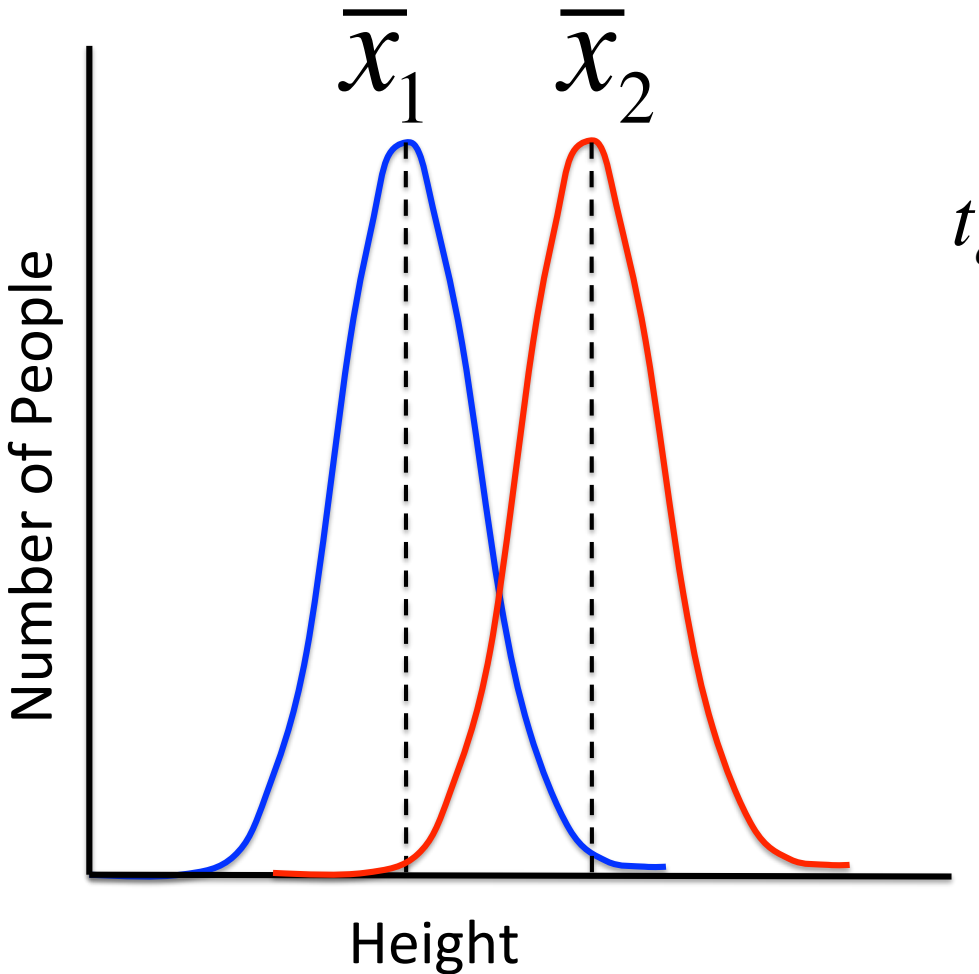
T-test for comparison of means:



$$t_{calc} = \frac{\left|\overline{x}_1 - \overline{x}_2\right|}{s_{pooled}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$t_{calc} = \frac{74.2 - 67.9}{7.6} \sqrt{\frac{5 \times 7}{5 + 7}}$$

T-test for comparison of means:



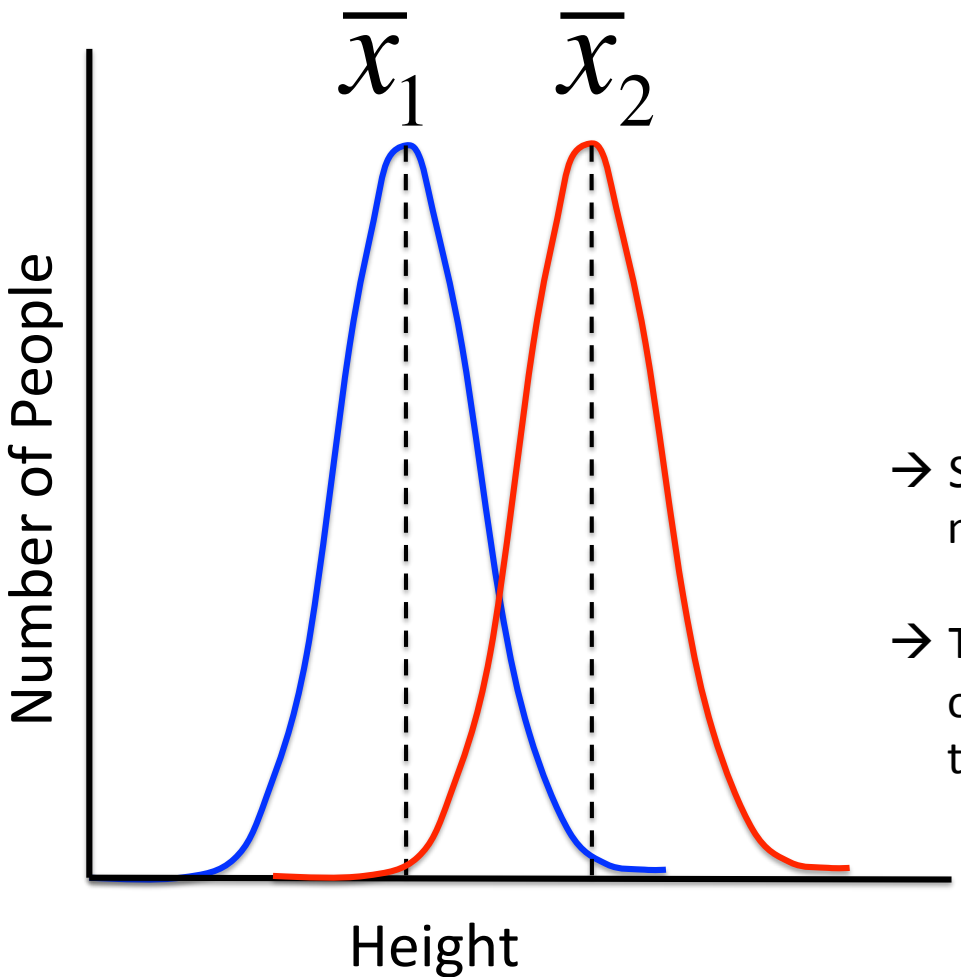
$$t_{calc} = \frac{\left|\overline{x}_1 - \overline{x}_2\right|}{s_{pooled}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$t_{calc} = 1.41$$

$$t_{table\,50\%} = 0.7$$

$$t_{table\,90\%} = 1.8$$

T-test for comparison of means:



$$t_{calc} = 1.41$$

$$t_{table\,50\%} = 0.7$$

$$t_{table\,90\%} = 1.8$$

→ Significant at the 50% confidence level, not at the 90% confidence level

→ There is a chance **greater than 10%** of observing a difference this large if the two populations were the same

Be aware of assumptions

- The data adopt a normal distribution
- The data correspond to a representative sample of the total population
- The (self-reported) data are accurate

What can we do to increase our confidence?

$$t_{calc} = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{s_{pooled}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

- Seek larger t values by increasing n
  - *For a given total number of observations ($n_1 + n_2$), t is maximized when $n_1 = n_2$*
- More precise experimental technique
- Avoiding sampling bias, reporting bias
- Perform a second, complementary experiment

Tools for computing statistical significance

- Excel TTEST function (spreadsheet available on course page)
- Graphing software has built in tools to perform t tests

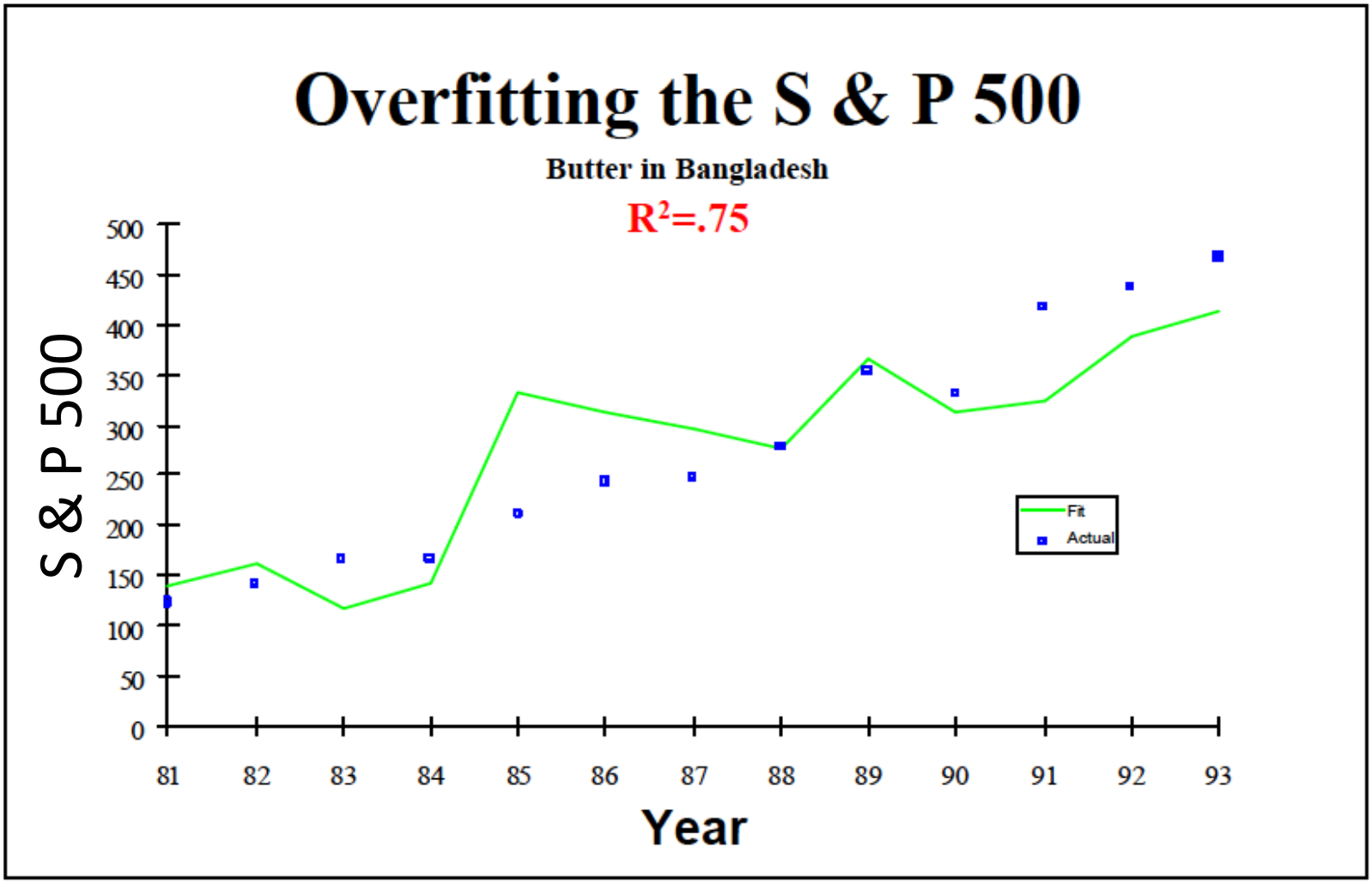
Conclusions

- We are never totally certain of a result
- Statistics provide a way to tell others how certain we are
- Increasing the number of data points can increase certainty
- Doing a second independent experiment can be very helpful

Lies, Damn Lies, and Statistics

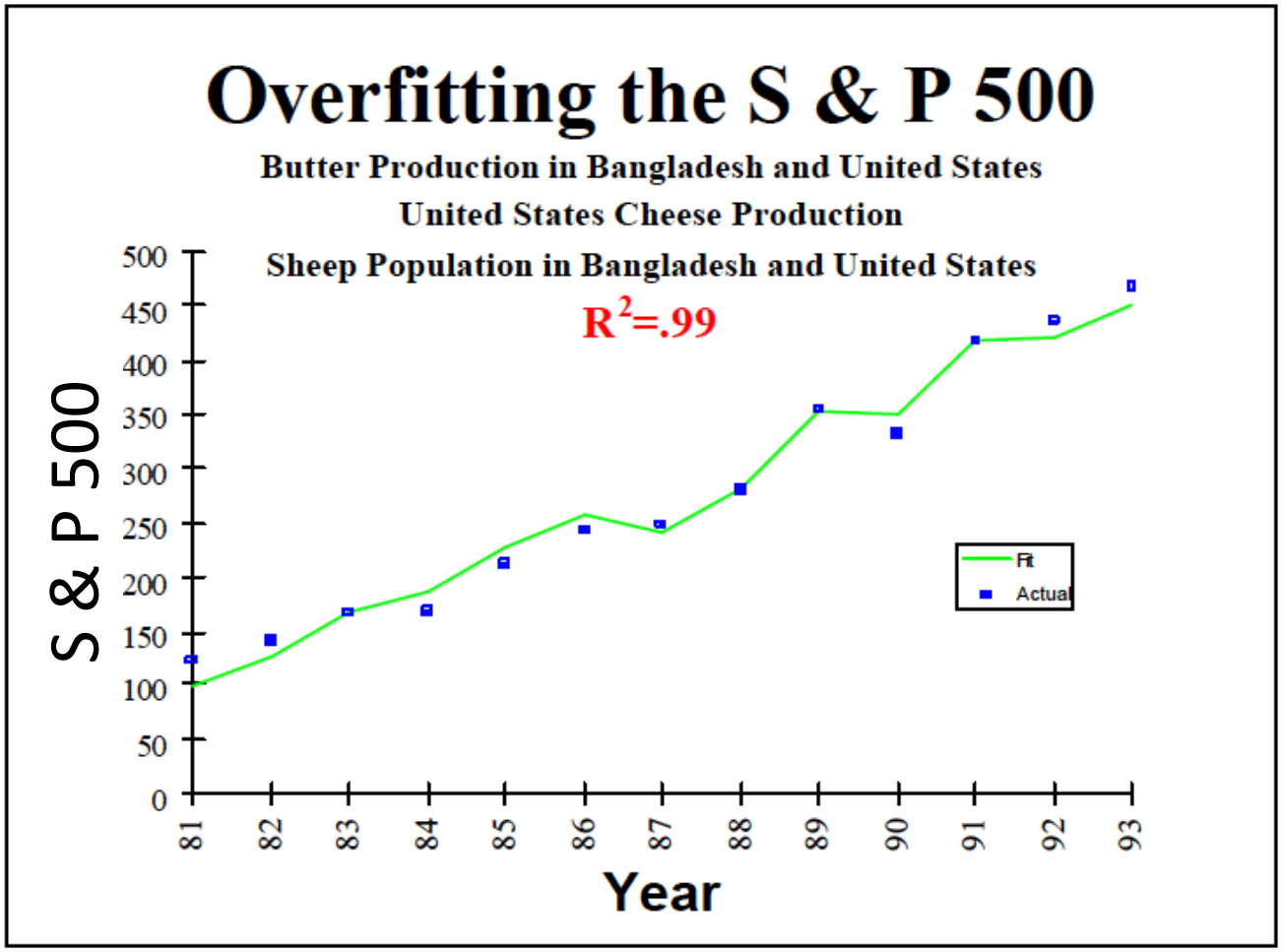
- What can go wrong?
  - Ill-defined question
  - Insufficient data
  - Over-interpretation of Data
    - The following slides show how data mining can result in spurious correlations; this illustrates a pitfall of working with large data sets. We often call a result significant if there is only a 5% chance of obtaining the result by chance; but if you "roll the dice" 1000 times, you will find quite a few examples of that 5% chance.

Data Fail



he Journal of Investing Spring 2007, Vol. 16, No. 1: pp. 15-22

Data Fail



he Journal of Investing Spring 2007, Vol. 16, No. 1: pp. 15-22

Examples of Gaussian Curves forming before your eyes

- Exhibit at the Boston Museum of Science

https://www.youtube.com/watch?v=KtPr7iipUso

- Patterns of snow accumulation around an obstruction appear to generate a similar effect