Basic Stats and Probability

Everything Data

CompSci 216 Spring 2015

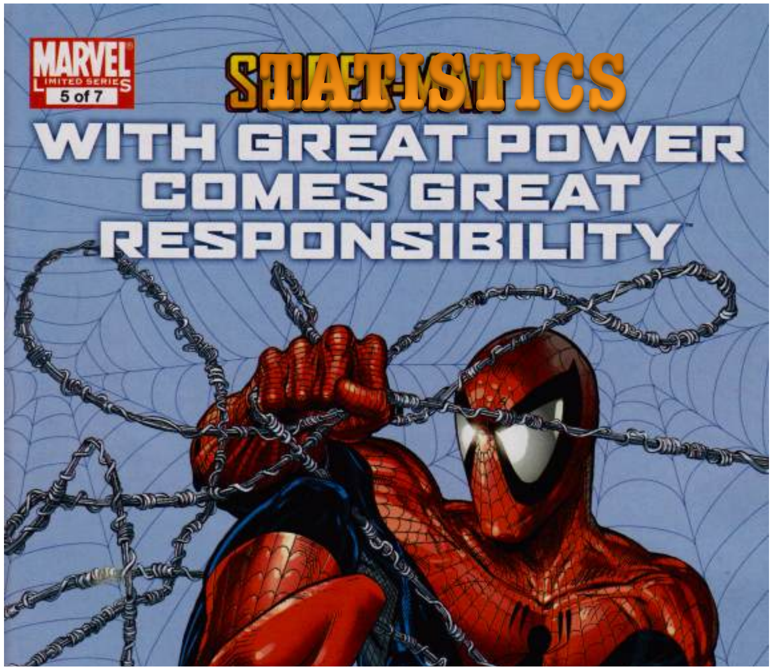
DUKE
COMPUTER SCIENCE

Announcements (Wed. Feb. 4)

- **Homework #4** will be posted by tomorrow morning
  - Due midnight Sunday
- Thinking about revising our office hours…

Disclaimer

- This lecture is no substitute for a real course on statistics
- We intend it to be a “teaser,” to illustrate some basic concepts + potential power and pitfalls



http://marvel.wikia.com/Spider-Man:_With_Great_Power_Comes_Great_Responsibility_Vol_1_5

Say you are buying a house…

Your agent could tell you, with perfect "honesty," that the "average" annual income in the neighborhood is

- $150,000
- $35,000
- $10,000

Example from Rick Ord of UCSD: http://cseweb.ucsd.edu/~ricko/CSE3/Lie_with_Statistics.pdf
Image: http://yourfinancialblog.com/wp-content/uploads/2013/01/buy-house.jpg

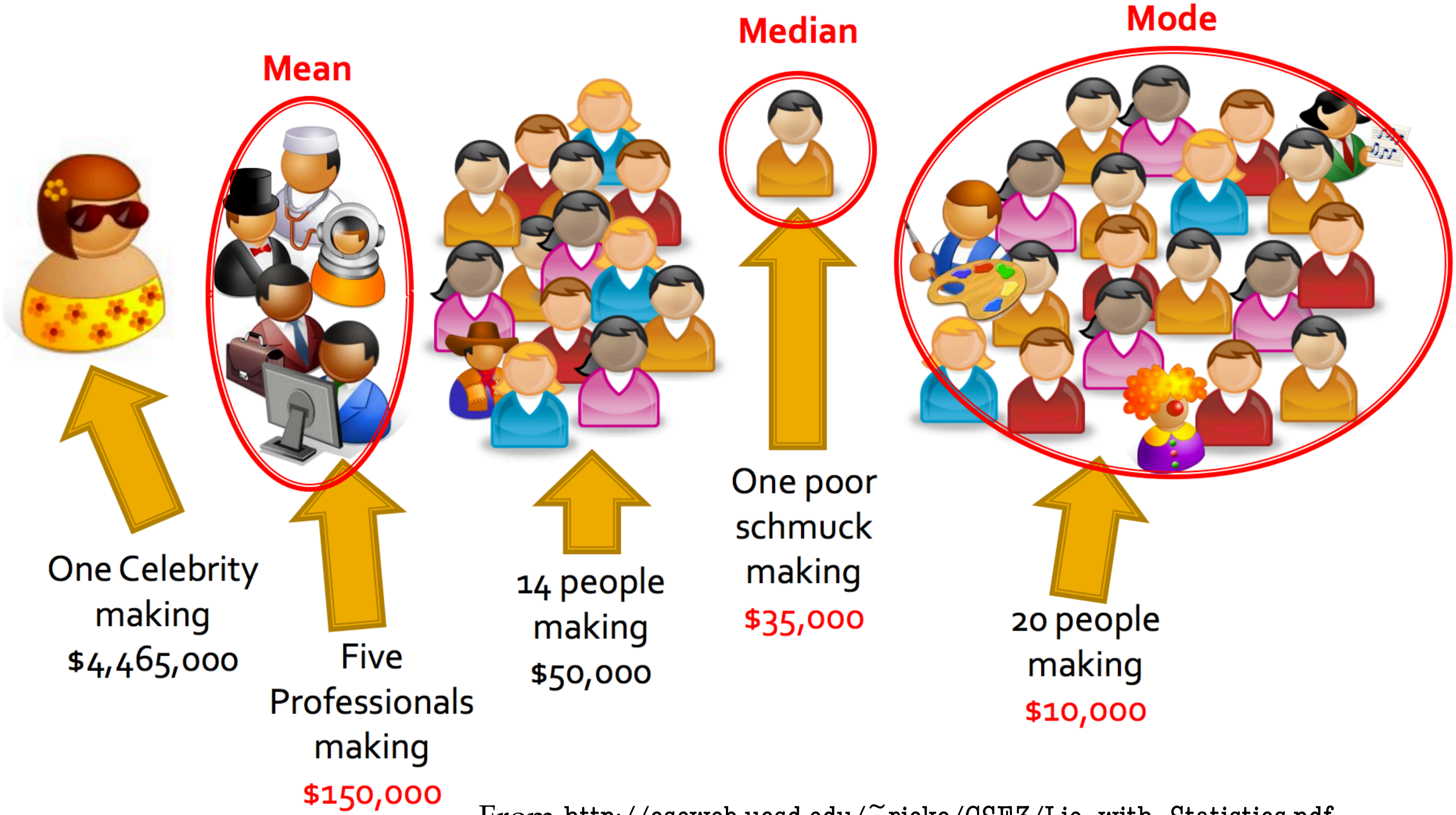
Measuring the “center”

Give a collection of values:

- ***Mean***: the arithmetic average of all values
  - = sum / count
- ***Median***: the “middle” number when all values are arranged in order
- ***Mode***: the most common value

Which is which? $10,000, $35,000, $150,000

The reality is…



Mean

Median

Mode

One Celebrity making $4,465,000

Five Professionals making $150,000

14 people making $50,000

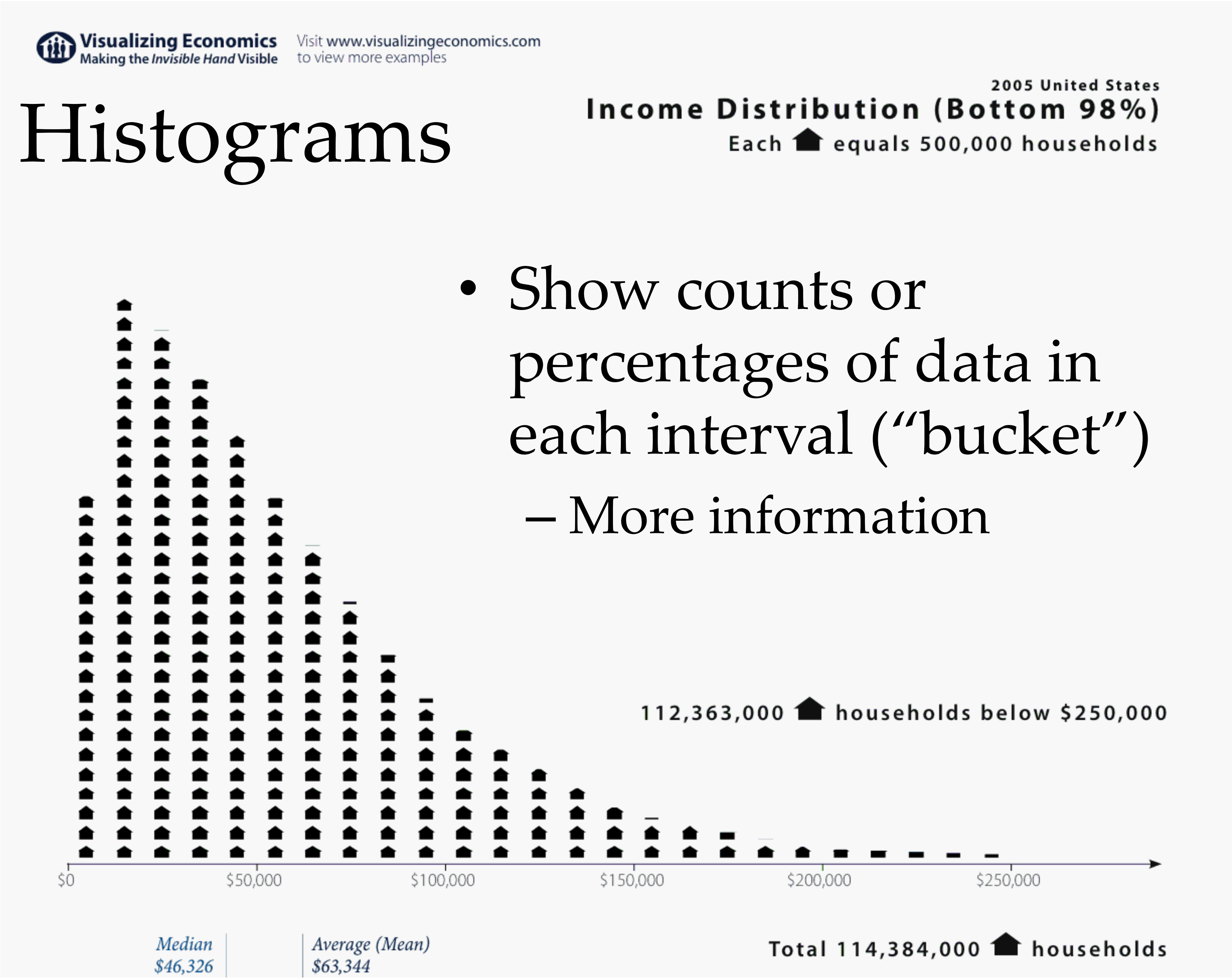
One poor schmuck making $35,000

20 people making $10,000

From http://cseweb.ucsd.edu/~ricko/CSE3/Lie_with_Statistics.pdf

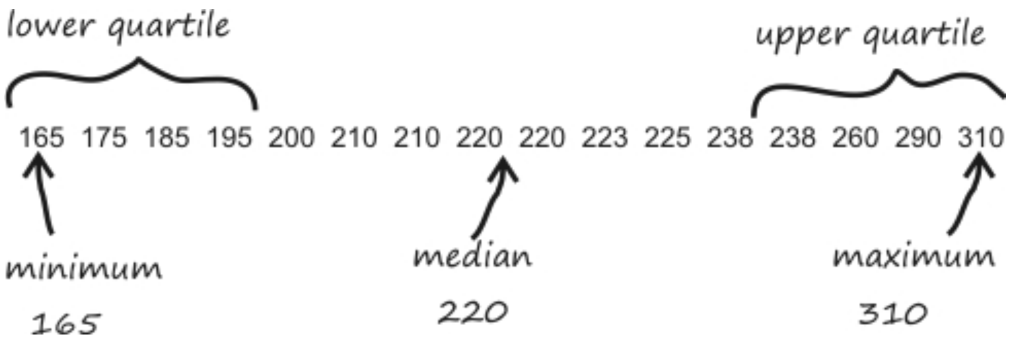
Histograms

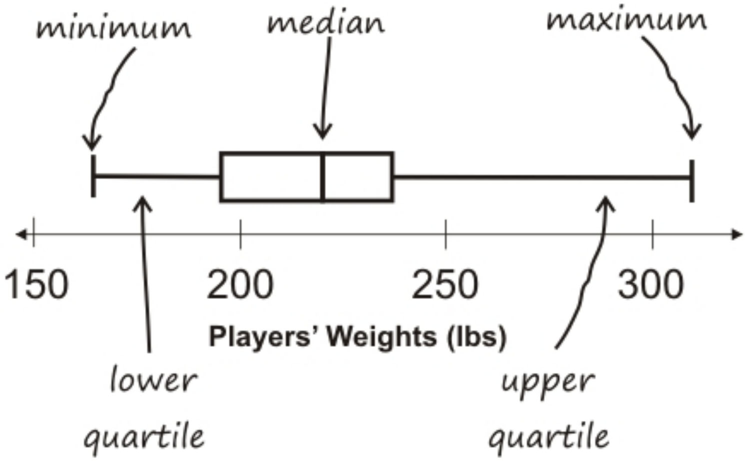
- Show counts or percentages of data in each interval ("bucket")
  - More information



Illustrating data distribution

- Histogram
- *5-number summary*: min, lower quartile, median, upper quartile, max
  - Lower (upper) *quartile* is marked by the middle value below (above) median
  - Box(-and-whisker) plot





Images: http://illuminations.nctm.org/Lesson.aspx?id=2643

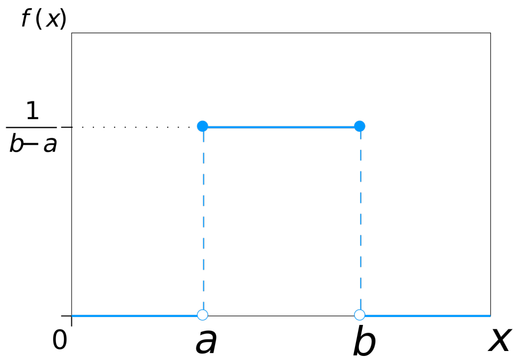
Measuring the “spread”

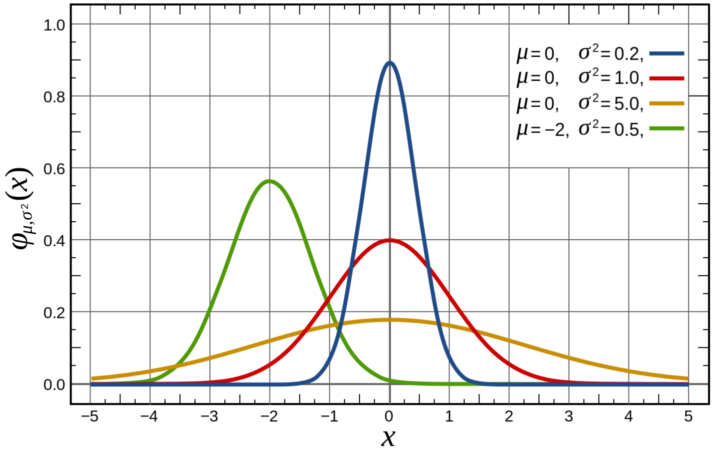
- ***Variance***: average ***squared*** deviation from the mean
- ***Standard deviation (σ)***: the square root of variance

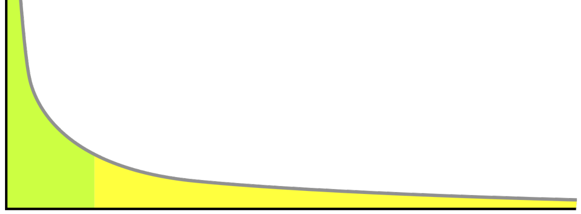
- *Chebyshev’s Inequality: No matter how data is distributed, no more than* $1/k^2$ *of the values can be more than* k *standard deviation away from the mean*
  - *E.g., no more than 25% of the values can be 2 standard deviations away from the mean*
  - *Provided that mean and standard deviation are defined*

Example distributions

- ***Uniform*** distribution
  - E.g., outcome of a dice, rand()
- ***Normal*** distributions
  - aka *Gaussian*, *bell curve*
  - E.g., test scores, people's blood pressure
- ***Power-law*** distributions
  - "80/20" rule
  - E.g., income in US, # page views, # friends







http://en.wikipedia.org/wiki/File:Uniform_Distribution_PDF_SVG.svg
http://en.wikipedia.org/wiki/File:Normal_Distribution_PDF.svg
http://en.wikipedia.org/wiki/File:Long_tail.svg

The *Ferengi*

An alien species in Star Trek notorious for extreme sexism

http://en.memory-alpha.org/wiki/The_Magnificent_Ferengi_(episode)

A Google interview question

- Ferengi want boys, so every family keeps on having children until a boy is born
  - If the newborn is a girl, have another child
  - If the newborn is a boy, stop

Can their strategy influence the composition of their population?

Probabilities come to rescue

- We need them to quantify uncertainty so that we can make better decisions
  - Human intuitions about uncertainty are sometimes ***terrible***!

A SWEDISH STUDY IN 2009 SHOWED THAT PEOPLE WITH BAD BOSSES HAD 40% MORE HEART ATTACKS.
AAAK!!!
I SHOULD WARN YOU THAT I'LL PROBABLY TELL THIS STORY A FEW TIMES.
Dilbert.com DilbertCartoonist@gmail.com
11-6-10 ©2010 Scott Adams, Inc./Dist. by UFS, Inc.

Probabilities

Probability of event A: a number $\mathrm{P}(A)$ between 0 and 1, indicating the likelihood of A happening

E.g., P(newborn is a boy) = P(newborn is girl) = 0.5

$$\mathrm{P}(A) + \mathrm{P}(\bar{A}) = 1$$

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$$

E.g., x = roll of a 6-sided die;

$\mathrm{P}(x \text{ is even} \cup x > 3)$

$= \mathrm{P}(x \text{ is even}) + \mathrm{P}(x > 3) - \mathrm{P}(x \text{ is even} \cap x > 3)$

$= 0.5 + 0.5 - 1/3 = 2/3$

Conditional probabilities

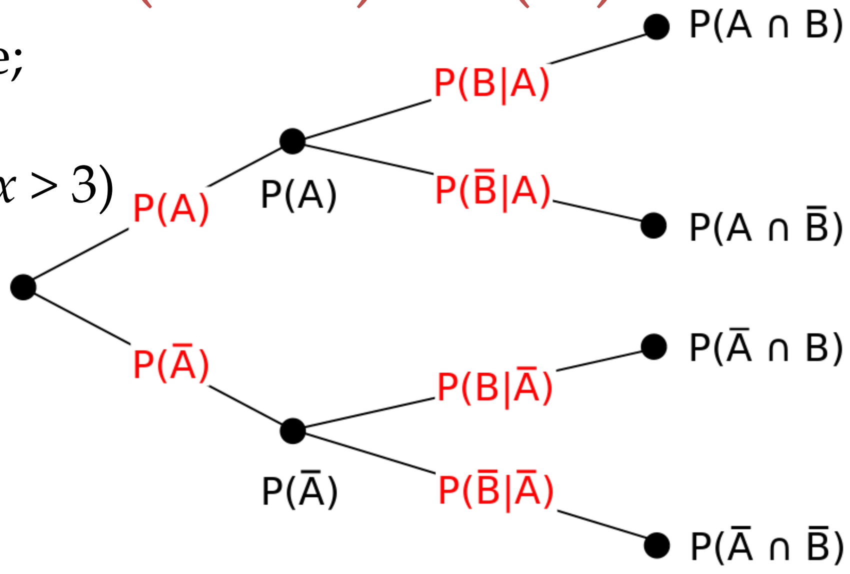
Given that an event A has occurred, the probability that event B also occurs is denoted by $P(B \mid A) = P(A \cap B) \,/\, P(A)$

E.g., x = roll of a 6-sided die;

$P(x \text{ is even} \mid x > 3)$

$= P(x \text{ is even} \cap x > 3) \,/\, P(x > 3)$

$= (1/3) \,/\, 0.5 = 2/3$



http://en.wikipedia.org/wiki/File:Probability_tree_diagram.svg

Independent events

Two events are independent if the occurrence (or non-occurrence) of one event does not change the probability that the other will occur, i.e.: $P(A \mid B) = P(A)$

- Or equivalently: $P(B \mid A) = P(B)$
- Or equivalently: $P(A \cap B) = P(A)\, P(B)$

E.g., *x* and *y* are two newborns;

“*x* is girl” and “*y* is a boy” are independent;

so P(first girl, then boy) = (1/2)(1/2) = 1/4

Random variables

A random variable assigns a numerical value to each possible outcome

E.g.: possible outcomes in a Ferengi family, and random variables B (# boys) and G (# girls)

Prob	Outcome	B	G
1/2	boy	1	0
$(1/2)^2$	girl, boy	1	1
$(1/2)^3$	girl, girl, boy	1	2
$(1/2)^4$	girl, girl, girl, boy	1	3
…	…	…	…

What's the probability of each outcome?

Probability distributions

For a discrete random variable, we can specify its distribution by a *probability mass function* (*pmf*) that assigns each value a probability

Prob	Outcome	*B*	*G*
1/2	boy	1	0
$(1/2)^2$	girl, boy	1	1
$(1/2)^3$	girl, girl, boy	1	2
$(1/2)^4$	girl, girl, girl, boy	1	3
…	…	…	…

G	Prob
0	1/2
1	$(1/2)^2$
2	$(1/2)^3$
3	$(1/2)^4$
	…

B	Prob
1	1

Expectation

For a discrete random variable *X*, the expected value of *X*, denoted E[*X*], is the average of all possible *X* values, each weighted by its probability

- Intuitively, E[*X*] = average of *X* values that we would observe if we draw from *X*'s distribution an infinite number of times

Back to the Ferengi…

Expected # of boys in a family?

- E[B] = 1

Prob	B
1	1

Expected # of girls in a family?

- E[G] = $1\times\left(\frac{1}{2}\right)^2+2\times\left(\frac{1}{2}\right)^3+3\times\left(\frac{1}{2}\right)^4+\cdots=1$

Prob	G
1/2	0
$(1/2)^2$	1
$(1/2)^3$	2
$(1/2)^4$	3
…	

Expected # of kids in a family?

- E[B+G] = E[B] + E[G] = 2
  - *Linearity of expectation*

How about…

Expected % of boys in a family?

- Hint: quite a bit more than 50%
- E[B/(B+G)] does ***not*** equal E[B] / E[B+G]

Expected % of boys in a population?

- Hint: not the same as expected % of boys in a family—averaging percentages doesn’t always make sense

What can this stuff *really* do?

(Besides nailing Google interviews?)

Just two examples here:

- Hypothesis testing
- Bayesian inference

Hypothesis testing

A simple example

A Ferengi government official showed you census data asserting that in the 10,000 families surveyed, there are just 5,000 girls

What's going on?

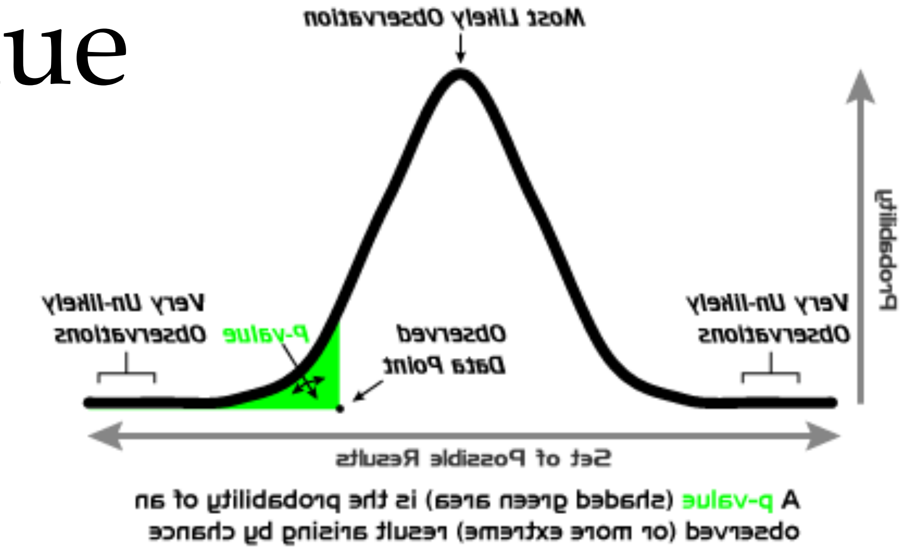
Commander *Data* of Star Trek

http://www.fanpop.com/clubs/lt-commander-data/images/381182/title/sherlock-data-photo

Playing Sherlock Holmes

- We have a model of how Ferengi families reproduce
  - Which implies that # girls/# families should follow a normal distribution centered at 1, by the ***Central Limit Theorem*** (details omitted)
- ***Null hypothesis***: census data is consistent with the model
- What's the probability of seeing ≤ 5,000 girls?
  - Known as the "***p-value***"
  - A very small p-value allows us to reject the null hypothesis

Calculating p-value


Most Likely Observation
Probability
Very Un-likely Observations
p-value
Observed Data Point
Very Un-likely Observations
Set of Possible Results
A p-value (shaded green area) is the probability of an observed (or more extreme) result arising by chance

Via a statistical formula

- Requires you to know which one to use and work out some math

Or, you can just “simulate” computationally

- Run many experiments, and compute the fraction with ≤ 5,000 girls

p-value for this example is less than 1 in 10^{250}

http://en.wikipedia.org/wiki/File:P-value_Graph.png

So?

In this case we reject the null hypothesis

- Either our model about Ferengi family planning is wrong, or
- The census data is bogus—data collection procedure was flawed, or data has been tampered with

Bayesian inference

A quiz

- The probability that a woman 40 to 50 years old has breast cancer is 0.8% P(cancer) = 0.008
- If a woman has breast cancer, the probability is 90% that she will have a positive mammogram P(pos|cancer) = 0.9
- If a woman does not have breast cancer, the probability is 7% that she will still have a positive mammogram P(pos|no cancer) = 0.07

Imagine a woman who has a positive mammogram. *What is the probability that she actually has breast cancer?* P(cancer|pos) = ?

Source: http://opinionator.blogs.nytimes.com/2010/04/25/chances-are/
Image: http://hcpress2.healthcommunities.com/wp-content/uploads/2008/08/stethoscope.jpg

To a doctor…

Reaction from a department chief at a German university teaching hospital with more than 30 years of experience:

> *"[He] was visibly nervous while trying to figure out what he would tell the woman. After mulling the numbers over, he finally estimated the woman's probability of having breast cancer, given that she has a positive mammogram, to be* ***90 percent****. Nervously, he added, 'Oh, what nonsense. I can't do this. You should test my daughter; she is studying medicine.'"*

Source: http://opinionator.blogs.nytimes.com/2010/04/25/chances-are/

Bayesian inference

- $P(\text{cancer}) = 0.008$
- $P(\text{pos} \mid \text{cancer}) = 0.9$
- $P(\text{pos} \mid \text{no cancer}) = 0.07$
- $P(\text{cancer} \mid \text{pos}) = ?$

$$= \frac{P(\text{pos} \mid \text{cancer})\ P(\text{cancer})}{P(\text{pos})}$$

$$= \frac{P(\text{pos} \mid \text{cancer})\ P(\text{cancer})}{P(\text{pos} \mid \text{cancer})\ P(\text{cancer}) + P(\text{pos} \mid \text{no cancer})\ P(\text{no cancer})}$$

$$= \frac{0.9 \times 0.008}{0.9 \times 0.008 + 0.07 \times (1 - 0.008)}$$

$$\approx 9.4\%$$

Bayes' Theorem:

$$P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B)}$$

$P(A)$: "prior,"
initial degree of belief in A

$P(A \mid B)$: "posterior,"
degree of belief having accounted for B

As for the American doctors…

… 95 out of 100 estimated the woman's probability of having breast cancer to be somewhere around ***75 percent****.*

Source: http://opinionator.blogs.nytimes.com/2010/04/25/chances-are/
Image: http://cutiepicture.files.wordpress.com/2007/06/catpop1.jpg