Introduction to statistical inference

Statistical inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

Statistical inference defined

Statistical inference is the process of drawing formal conclusions from data.

In our class, we wil define formal statistical inference as settings where one wants to infer facts about a population using noisy statistical data where uncertainty must be accounted for.

Motivating example: who's going to win the election?

In every major election, pollsters would like to know, ahead of the actual election, who's going to win. Here, the target of estimation (the estimand) is clear, the percentage of people in a particular group (city, state, county, country or other electoral grouping) who will vote for each candidate.

We can not poll everyone. Even if we could, some polled may change their vote by the time the election occurs. How do we collect a reasonable subset of data and quantify the uncertainty in the process to produce a good guess at who will win?

Motivating example: is hormone replacement therapy effective?

A large clinical trial (the Women's Health Initiative) published results in 2002 that contradicted prior evidence on the efficacy of hormone replacement therapy for post menopausal women and suggested a negative impact of HRT for several key health outcomes. **Based on a statistically based protocol, the study was stopped early due an excess number of negative events.**

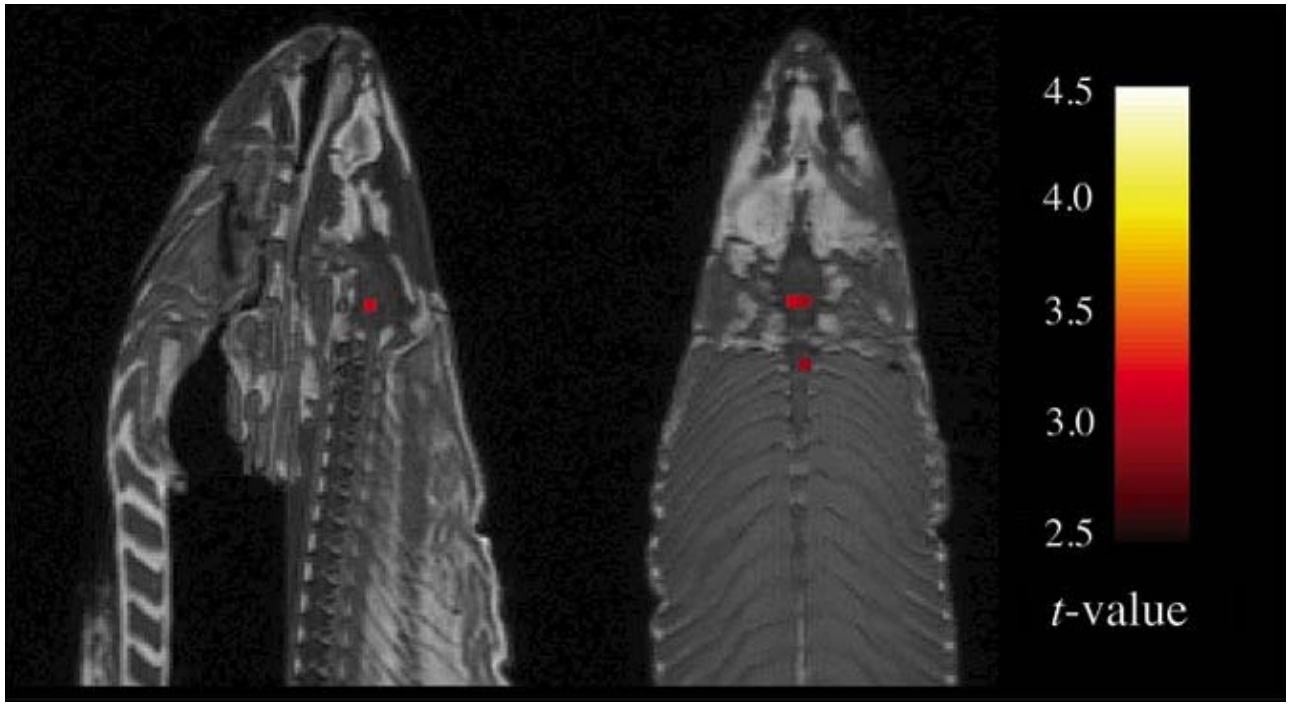
Here's there's two inferential problems.

1. Is HRT effective?
2. How long should we continue the trial in the presence of contrary evidence?

See WHI writing group paper JAMA 2002, Vol 288:321 - 333. for the paper and Steinkellner et al. Menopause 2012, Vol 19:616 621 for adiscussion of the long term impacts

Motivating example

Brain activation



http://www.wired.com/2009/09/fmrisalmon/

Summary

- These examples illustrate many of the difficulties of trying to use data to create general conclusions about a population.
- Paramount among our concerns are:
  - Is the sample representative of the population that we'd like to draw inferences about?
  - Are there known and observed, known and unobserved or unknown and unobserved variables that contaminate our conclusions?
  - Is there systematic bias created by missing data or the design or conduct of the study?
  - What randomness exists in the data and how do we use or adjust for it? Here randomness can either be explicit via randomization or random sampling, or implicit as the aggregation of many complex uknown processes.
  - Are we trying to estimate an underlying mechanistic model of phenomena under study?
- Statistical inference requires navigating the set of assumptions and tools and subsequently thinking about how to draw conclusions from data.

Example goals of inference

1. Estimate and quantify the uncertainty of an estimate of a population quantity (the proportion of people who will vote for a candidate).
2. Determine whether a population quantity is a benchmark value ("is the treatment effective?").
3. Infer a mechanistic relationship when quantities are measured with noise ("What is the slope for Hooke's law?")
4. Determine the impact of a policy? ("If we reduce polution levels, will asthma rates decline?")
5. Talk about the probability that something occurs.

Example tools of the trade

1. Randomization: concerned with balancing unobserved variables that may confound inferences of interest
2. Random sampling: concerned with obtaining data that is representative of the population of interest
3. Sampling models: concerned with creating a model for the sampling process, the most common is so called "iid".
4. Hypothesis testing: concerned with decision making in the presence of uncertainty
5. Confidence intervals: concerned with quantifying uncertainty in estimation
6. Probability models: a formal connection between the data and a population of interest. Often probability models are assumed or are approximated.
7. Study design: the process of designing an experiment to minimize biases and variability.
8. Nonparametric bootstrapping: the process of using the data to, with minimal probability model assumptions, create inferences.
9. Permutation, randomization and exchangeability testing: the process of using data permutations to perform inferences.

Different thinking about probability leads to different styles of inference

We won't spend too much time talking about this, but there are several different styles of inference. Two broad categories that get discussed a lot are:

1. Frequency probability: is the long run proportion of times an event occurs in independent, identically distributed repetitions.
2. Frequency inference: uses frequency interpretations of probabilities to control error rates. Answers questions like "What should I decide given my data controlling the long run proportion of mistakes I make at a tolerable level."
3. Bayesian probability: is the probability calculus of beliefs, given that beliefs follow certain rules.
4. Bayesian inference: the use of Bayesian probability representation of beliefs to perform inference. Answers questions like "Given my subjective beliefs and the objective information from the data, what should I believe now?"

Data scientists tend to fall within shades of gray of these and various other schools of inference.

In this class

- In this class, we will primarily focus on basic sampling models, basic probability models and frequency style analyses to create standard inferences.
- Being data scientists, we will also consider some inferential strategies that rely heavily on the observed data, such as permutation testing and bootstrapping.
- As probability modeling will be our starting point, we first build up basic probability.

Where to learn more on the topics not covered

1. Explicit use of random sampling in inferences: look in references on "finite population statistics". Used heavily in polling and sample surveys.
2. Explicit use of randomization in inferences: look in references on "causal inference" especially in clinical trials.
3. Bayesian probability and Bayesian statistics: look for basic itroductory books (there are many).
4. Missing data: well covered in biostatistics and econometric references; look for references to "multiple imputation", a popular tool for addressing missing data.
5. Study design: consider looking in the subject matter area that you are interested in; some examples with rich histories in design:
    1. The epidemiological literature is very focused on using study design to investigate public health.
    2. The classical development of study design in agriculture broadly covers design and design principles.
    3. The industrial quality control literature covers design thoroughly.

Probability

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

Probability

- In these slides we will cover the basics of probability at low enough level to have a basic understanding for the rest of the series
- For a more complete treatment see the class Mathematical Biostatistics Boot Camp 1
  - Youtube: www.youtube.com/playlist?list=PLpl-gQkQivXhk6qSyiNj51qamjAtZISJ-
  - Coursera: www.coursera.org/course/biostats
  - Git: http://github.com/bcaffo/Caffo-Coursera

Probability

Given a random experiment (say rolling a die) a probability measure is a population quantity that summarizes the randomness.

Specifically, probability takes a possible outcome from the expertiment and:

- assigns it a number between 0 and 1
- so that the probability that something occurs is 1 (the die must be rolled) and
- so that the probability of the union of any two sets of outcomes that have nothing in common (mutually exclusive) is the sum of their respective probabilities.

The Russian mathematician Kolmogorov formalized these rules.

Rules probability must follow

- The probability that nothing occurs is 0
- The probability that something occurs is 1
- The probability of something is 1 minus the probability that the opposite occurs
- The probability of at least one of two (or more) things that can not simultaneously occur (mutually exclusive) is the sum of their respective probabilities
- If an event A implies the occurrence of event B, then the probability of A occurring is less than the probability that B occurs
- For any two events the probability that at least one occurs is the sum of their probabilities minus their intersection.

Example

The National Sleep Foundation (www.sleepfoundation.org) reports that around 3% of the American population has sleep apnea. They also report that around 10% of the North American and European population has restless leg syndrome. Does this imply that 13% of people will have at least one sleep problems of these sorts?

Example continued

Answer: No, the events can simultaneously occur and so are not mutually exclusive. To elaborate let:

If you want to see the mathematics

$$A_1 = \{\text{Person has sleep apnea}\}$$
$$A_2 = \{\text{Person has RLS}\}$$

Then

$$\begin{aligned} P(A_1 \cup A_2) &= P(A_1) + P(A_2) - P(A_1 \cap A_2) \\ &= 0.13 - \text{Probability of having both} \end{aligned}$$

Likely, some fraction of the population has both.

Going further

Probability calculus is useful for understanding the rules that probabilities must follow.

However, we need ways to model and think about probabilities for numeric outcomes of experiments (broadly defined).

Densities and mass functions for random variables are the best starting point for this.

Remember, everything we're talking about up to at this point is a population quantity not a statement about what occurs in the data.

- We're going with this is: use the data to estimate properties of the population.

Random variables

- A **random variable** is a numerical outcome of an experiment.
- The random variables that we study will come in two varieties, **discrete** or **continuous**.
- Discrete random variable are random variables that take on only a countable number of possibilities and we talk about the probability that they take specific values
- Continuous random variable can conceptually take any value on the real line or some subset of the real line and we talk about the probability that they line within some range

Examples of variables that can be thought of as random variables

Experiments that we use for intuition and building context

- The $(0-1)$ outcome of the flip of a coin
- The outcome from the roll of a die

Specific instances of treating variables as if random

- The web site traffic on a given day
- The BMI of a subject four years after a baseline measurement
- The hypertension status of a subject randomly drawn from a population
- The number of people who click on an ad
- Intelligence quotients for a sample of children

PMF

A probability mass function evaluated at a value corresponds to the probability that a random variable takes that value. To be a valid pmf a function, p, must satisfy

1. It must always be larger than or equal to 0.
2. The sum of the possible values that the random variable can take has to add up to one.

Example

Let X be the result of a coin flip where $X = 0$ represents tails and $X = 1$ represents heads.

$$p(x) = (1/2)^x (1/2)^{1-x} \quad \text{for} \quad x = 0, 1$$

Suppose that we do not know whether or not the coin is fair; Let θ be the probability of a head expressed as a proportion (between 0 and 1).

$$p(x) = \theta^x (1-\theta)^{1-x} \quad \text{for} \quad x = 0, 1$$

PDF

A probability density function (pdf), is a function associated with a continuous random variable

Areas under pdfs correspond to probabilities for that random variable

To be a valid pdf, a function must satisfy

1. It must be larger than or equal to zero everywhere.
2. The total area under it must be one.

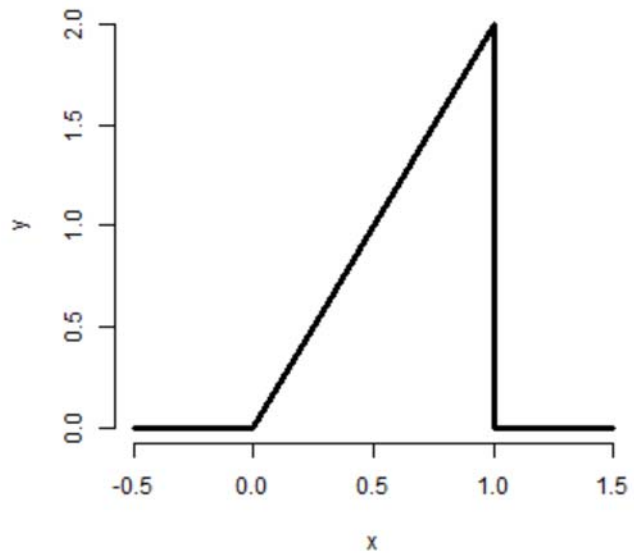
Example

Suppose that the proportion of help calls that get addressed in a random day by a help line is given by

$$f(x) = \begin{cases} 2x & \text{for} \quad 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

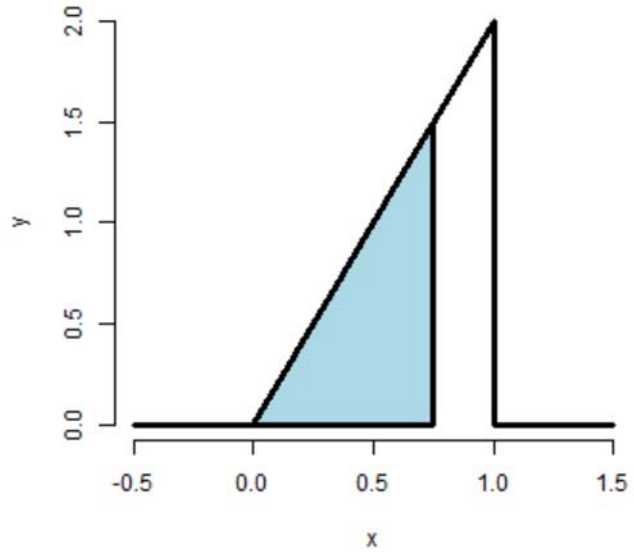
Is this a mathematically valid density?

```r
x <- c(-0.5, 0, 1, 1, 1.5)
y <- c(0, 0, 2, 0, 0)
plot(x, y, lwd = 3, frame = FALSE, type = "l")
```



Example continued

What is the probability that 75% or fewer of calls get addressed?



```
1.5 * 0.75/2
```

```
## [1] 0.5625
```

```
pbeta(0.75, 2, 1)
```

```
## [1] 0.5625
```

CDF and survival function

Certain areas are so useful, we give them names

- The **cumulative distribution function** (CDF) of a random variable, X, returns the probability that the random variable is less than or equal to the value x

$$F(x) = P(X \leq x)$$

  (This definition applies regardless of whether X is discrete or continuous.)

- The **survival function** of a random variable X is defined as the probability that the random variable is greater than the value x

$$S(x) = P(X > x)$$

- Notice that $S(x) = 1 - F(x)$

Example

What are the survival function and CDF from the density considered before?

For $1 \geq x \geq 0$

$$F(x) = P(X \leq x) = \frac{1}{2} \, Base \times Height = \frac{1}{2} \, (x) \times (2x) = x^2$$

$$S(x) = 1 - x^2$$

```
pbeta(c(0.4, 0.5, 0.6), 2, 1)
```

```
## [1] 0.16 0.25 0.36
```

Quantiles

You've heard of sample quantiles. If you were the 95th percentile on an exam, you know that 95% of people scored worse than you and 5% scored better. These are sample quantities. Here we define their population analogs.

Definition

The α^{th} **quantile** of a distribution with distribution function F is the point x_α so that

$$F(x_\alpha) = \alpha$$

- A **percentile** is simply a quantile with α expressed as a percent
- The **median** is the 50^{th} percentile

For example

The 95^{th} percentile of a distribution is the point so that:

- the probability that a random variable drawn from the population is less is 95%
- the probability that a random variable drawn from the population is more is 5%

Example

What is the median of the distribution that we were working with before?

- We want to solve $0.5 = F(x) = x^2$
- Resulting in the solution

```
sqrt(0.5)
```

```
## [1] 0.7071
```

- Therefore, about 0.7071 of calls being answered on a random day is the median.

Example continued

R can approximate quantiles for you for common distributions

```
qbeta(0.5, 2, 1)
```

```
## [1] 0.7071
```

Summary

- You might be wondering at this point "I've heard of a median before, it didn't require integration. Where's the data?"
- We're referring to are **population quantities**. Therefore, the median being discussed is the **population median**.
- A probability model connects the data to the population using assumptions.
- Therefore the median we're discussing is the **estimand**, the sample median will be the **estimator**

Conditional Probability

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

Conditional probability, motivation

- The probability of getting a one when rolling a (standard) die is usually assumed to be one sixth
- Suppose you were given the extra information that the die roll was an odd number (hence 1, 3 or 5)
- *conditional on this new information*, the probability of a one is now one third

Conditional probability, definition

- Let B be an event so that $P(B) > 0$
- Then the conditional probability of an event A given that B has occurred is

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

- Notice that if A and B are independent (defined later in the lecture), then

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

Example

- Consider our die roll example
- $B = \{1, 3, 5\}$
- $A = \{1\}$

$$
\begin{aligned}
P(\text{one given that roll is odd}) &= P(A \mid B) \\
&= \frac{P(A \cap B)}{P(B)} \\
&= \frac{P(A)}{P(B)} \\
&= \frac{1/6}{3/6} = \frac{1}{3}
\end{aligned}
$$

Bayes' rule

Baye's rule allows us to reverse the conditioning set provided that we know some marginal probabilities

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid B^c)P(B^c)} \, .$$

Diagnostic tests

- Let $+$ and $-$ be the events that the result of a diagnostic test is positive or negative respectively
- Let D and D^c be the event that the subject of the test has or does not have the disease respectively
- The **sensitivity** is the probability that the test is positive given that the subject actually has the disease, $P(+ \mid D)$
- The **specificity** is the probability that the test is negative given that the subject does not have the disease, $P(- \mid D^c)$

More definitions

- The **positive predictive value** is the probability that the subject has the disease given that the test is positive, $P(D \mid +)$
- The **negative predictive value** is the probability that the subject does not have the disease given that the test is negative, $P(D^c \mid -)$
- The **prevalence of the disease** is the marginal probability of disease, $P(D)$

More definitions

- The **diagnostic likelihood ratio of a positive test**, labeled DLR_+, is $P(+ \mid D)/P(+ \mid D^c)$, which is the

$$sensitivity/(1 - specificity)$$

- The **diagnostic likelihood ratio of a negative test**, labeled DLR_-, is $P(- \mid D)/P(- \mid D^c)$, which is the

$$(1 - sensitivity)/specificity$$

Example

- A study comparing the efficacy of HIV tests, reports on an experiment which concluded that HIV antibody tests have a sensitivity of 99.7% and a specificity of 98.5%
- Suppose that a subject, from a population with a .1% prevalence of HIV, receives a positive test result. What is the positive predictive value?
- Mathematically, we want $P(D \mid +)$ given the sensitivity, $P(+ \mid D) = .997$, the specificity, $P(- \mid D^c) = .985$, and the prevalence $P(D) = .001$

Using Bayes' formula

$$
\begin{aligned}
P(D \mid +) &= \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)} \\
&= \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + \{1 - P(- \mid D^c)\}\{1 - P(D)\}} \\
&= \frac{.997 \times .001}{.997 \times .001 + .015 \times .999} \\
&= .062
\end{aligned}
$$

- In this population a positive test result only suggests a 6% probability that the subject has the disease
- (The positive predictive value is 6% for this test)

More on this example

- The low positive predictive value is due to low prevalence of disease and the somewhat modest specificity
- Suppose it was known that the subject was an intravenous drug user and routinely had intercourse with an HIV infected partner
- Notice that the evidence implied by a positive test result does not change because of the prevalence of disease in the subject's population, only our interpretation of that evidence changes

Likelihood ratios

- Using Bayes rule, we have

$$P(D \mid +) = \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)}$$

and

$$P(D^c \mid +) = \frac{P(+ \mid D^c)P(D^c)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)} .$$

Likelihood ratios

- Therefore

$$\frac{P(D \mid +)}{P(D^c \mid +)} = \frac{P(+ \mid D)}{P(+ \mid D^c)} \times \frac{P(D)}{P(D^c)}$$

  ie

$$\text{post-test odds of } D = DLR_+ \times \text{pre-test odds of } D$$

- Similarly, DLR_- relates the decrease in the odds of the disease after a negative test result to the odds of disease prior to the test.

HIV example revisited

- Suppose a subject has a positive HIV test
- $DLR_+ = .997/(1 - .985) \approx 66$
- The result of the positive test is that the odds of disease is now 66 times the pretest odds
- Or, equivalently, the hypothesis of disease is 66 times more supported by the data than the hypothesis of no disease

HIV example revisited

- Suppose that a subject has a negative test result
- $DLR_{-} = (1 - .997)/.985 \approx .003$
- Therefore, the post-test odds of disease is now $.3\%$ of the pretest odds given the negative test.
- Or, the hypothesis of disease is supported $.003$ times that of the hypothesis of absence of disease given the negative test result

Independence

- Two events A and B are **independent** if

$$P(A \cap B) = P(A)P(B)$$

- Equivalently if $P(A \mid B) = P(A)$
- Two random variables, X and Y are independent if for any two sets A and B

$$P([X \in A] \cap [Y \in B]) = P(X \in A)P(Y \in B)$$

- If A is independent of B then
  - A^c is independent of B
  - A is independent of B^c
  - A^c is independent of B^c

Example

- What is the probability of getting two consecutive heads?
- $A = \{\text{Head on flip } 1\}$ ~ $P(A) = .5$
- $B = \{\text{Head on flip } 2\}$ ~ $P(B) = .5$
- $A \cap B = \{\text{Head on flips } 1 \text{ and } 2\}$
- $P(A \cap B) = P(A)P(B) = .5 \times .5 = .25$

Example

- Volume 309 of Science reports on a physician who was on trial for expert testimony in a criminal trial
- Based on an estimated prevalence of sudden infant death syndrome of 1 out of $8,543$, the physician testified that that the probability of a mother having two children with SIDS was $\left(\frac{1}{8,543}\right)^2$
- The mother on trial was convicted of murder

Example: continued

- Relevant to this discussion, the principal mistake was to *assume* that the events of having SIDs within a family are independent
- That is, $P(A_1 \cap A_2)$ is not necessarily equal to $P(A_1)P(A_2)$
- Biological processes that have a believed genetic or familiar environmental component, of course, tend to be dependent within families
- (There are many other statistical points of discussion for this case.)

IID random variables

- Random variables are said to be iid if they are independent and identically distributed
  - Independent: statistically unrelated from one and another
  - Identically distributed: all having been drawn from the same population distribution
- iid random variables are the default model for random samples
- Many of the important theories of statistics are founded on assuming that variables are iid
- Assuming a random sample and iid will be the default starting point of inference for this class

Expected values

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

Expected values

- Expected values are useful cor characterizing a distribution
- The mean is a characterization of its center
- The variance and standard deviation are characterizations of how spread out it is
- Our sample expected values (the sample mean and variance) will estimate the population versions

The population mean

- The **expected value** or **mean** of a random variable is the center of its distribution
- For discrete random variable X with PMF $p(x)$, it is defined as follows

$$E[X] = \sum_x xp(x).$$

  where the sum is taken over the possible values of x
- $E[X]$ represents the center of mass of a collection of locations and weights, $\{x, p(x)\}$

The sample mean

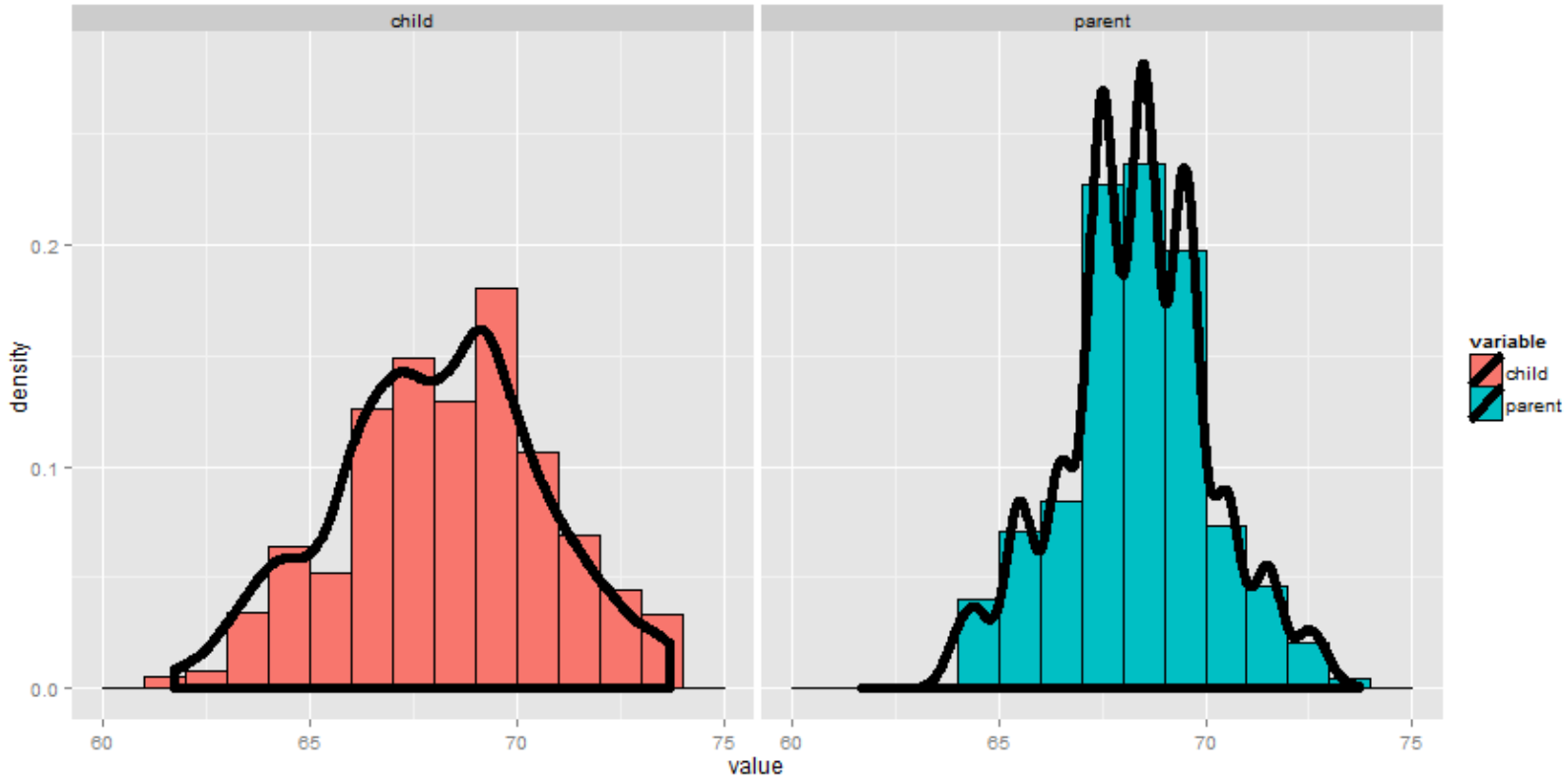
- The sample mean estimates this population mean
- The center of mass of the data is the empirical mean

$$\bar{X} = \sum_{i=1}^{n} x_i p(x_i)$$

where $p(x_i) = 1/n$

Example

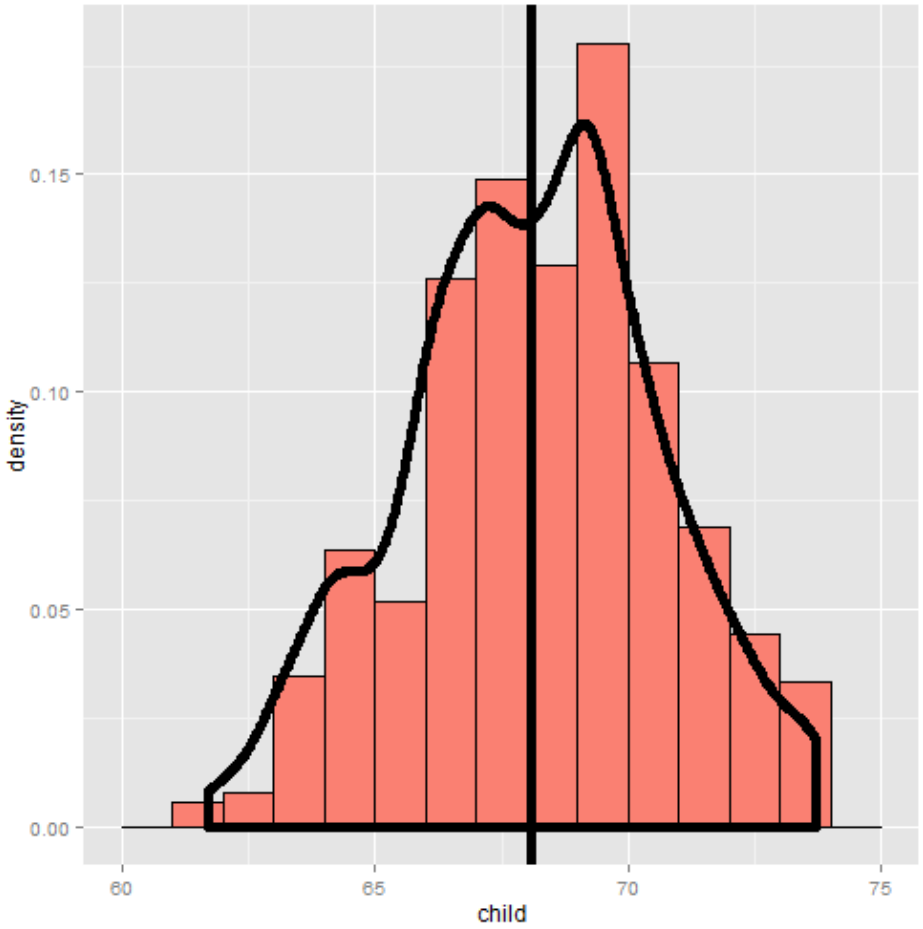
Find the center of mass of the bars



Using manipulate

```
library(manipulate)
myHist <- function(mu){
    g <- ggplot(galton, aes(x = child))
    g <- g + geom_histogram(fill = "salmon",
      binwidth=1, aes(y = ..density..), colour = "black")
    g <- g + geom_density(size = 2)
    g <- g + geom_vline(xintercept = mu, size = 2)
    mse <- round(mean((galton$child - mu)^2), 3)
    g <- g + labs(title = paste('mu = ', mu, ' MSE = ', mse))
    g
}
manipulate(myHist(mu), mu = slider(62, 74, step = 0.5))
```

The center of mass is the empirical mean

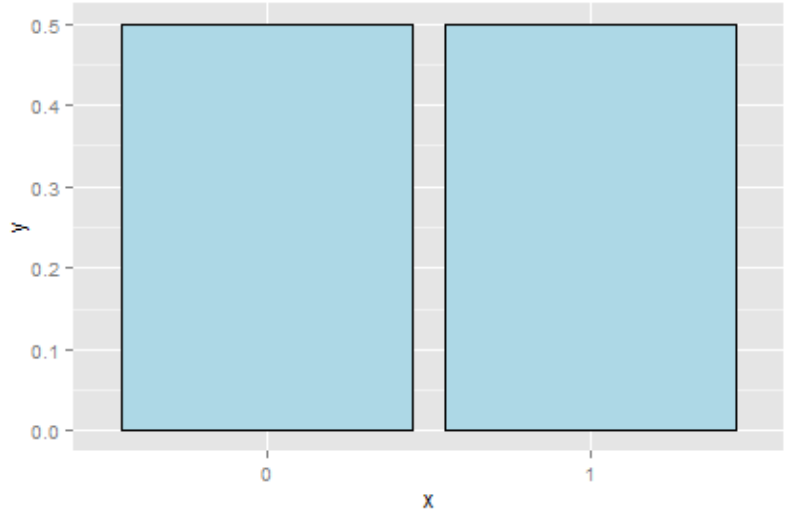


Example of a population mean

- Suppose a coin is flipped and X is declared 0 or 1 corresponding to a head or a tail, respectively
- What is the expected value of X?

$$E[X] = .5 \times 0 + .5 \times 1 = .5$$

- Note, if thought about geometrically, this answer is obvious; if two equal weights are spaced at 0 and 1, the center of mass will be $.5$



What about a biased coin?

- Suppose that a random variable, X, is so that $P(X = 1) = p$ and $P(X = 0) = (1 - p)$
- (This is a biased coin when $p \neq 0.5$)
- What is its expected value?

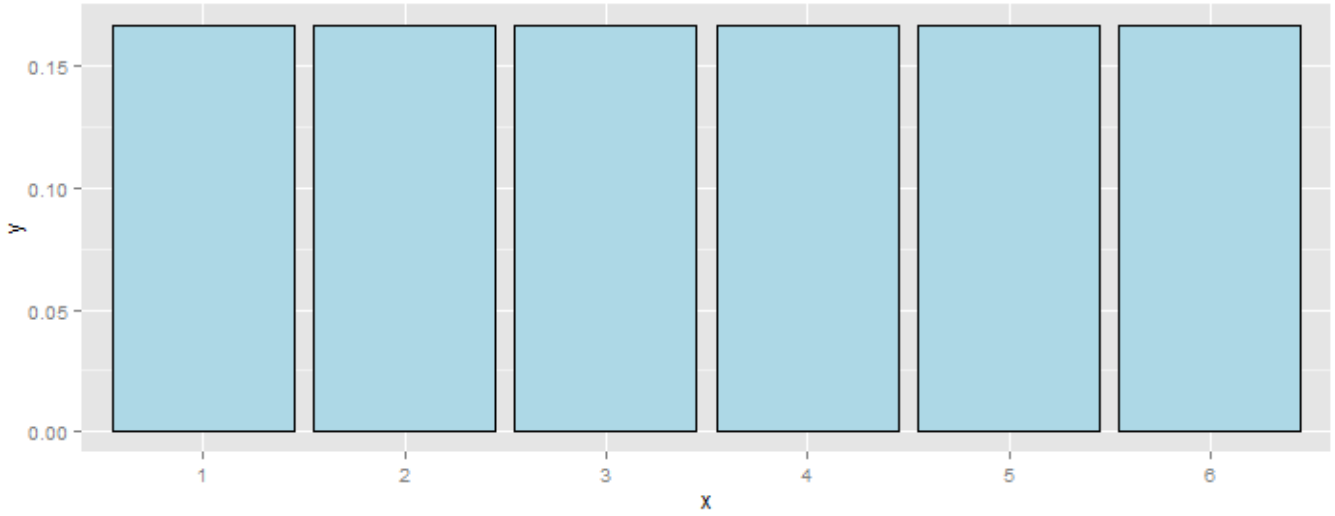
$$E[X] = 0 * (1 - p) + 1 * p = p$$

Example

- Suppose that a die is rolled and X is the number face up
- What is the expected value of X?

$$E[X] = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5$$

- Again, the geometric argument makes this answer obvious without calculation.

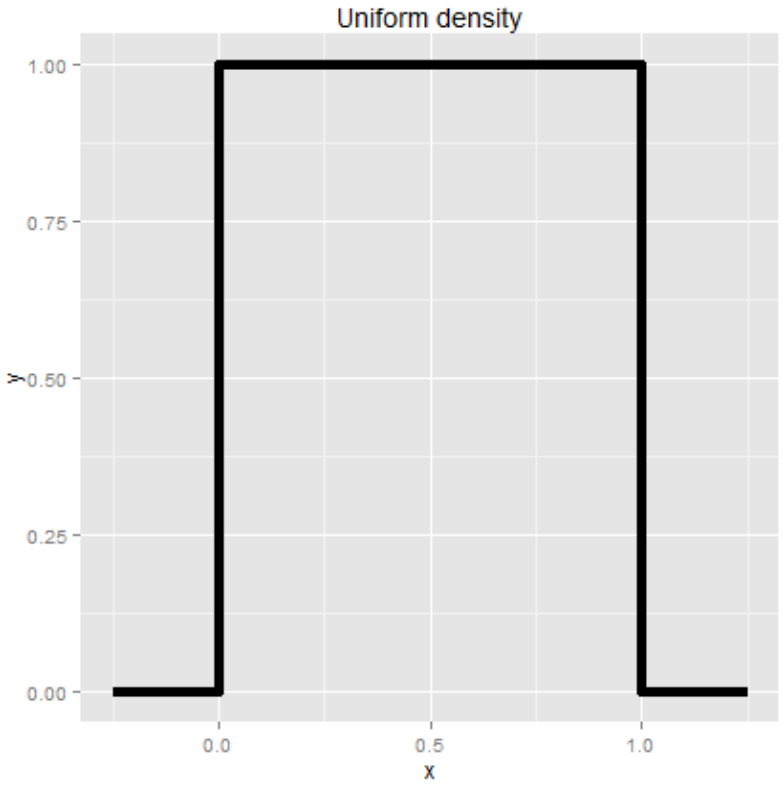


Continuous random variables

- For a continuous random variable, X, with density, f, the expected value is again exactly the center of mass of the density

Example

- Consider a density where $f(x) = 1$ for x between zero and one
- (Is this a valid density?)
- Suppose that X follows this density; what is its expected value?

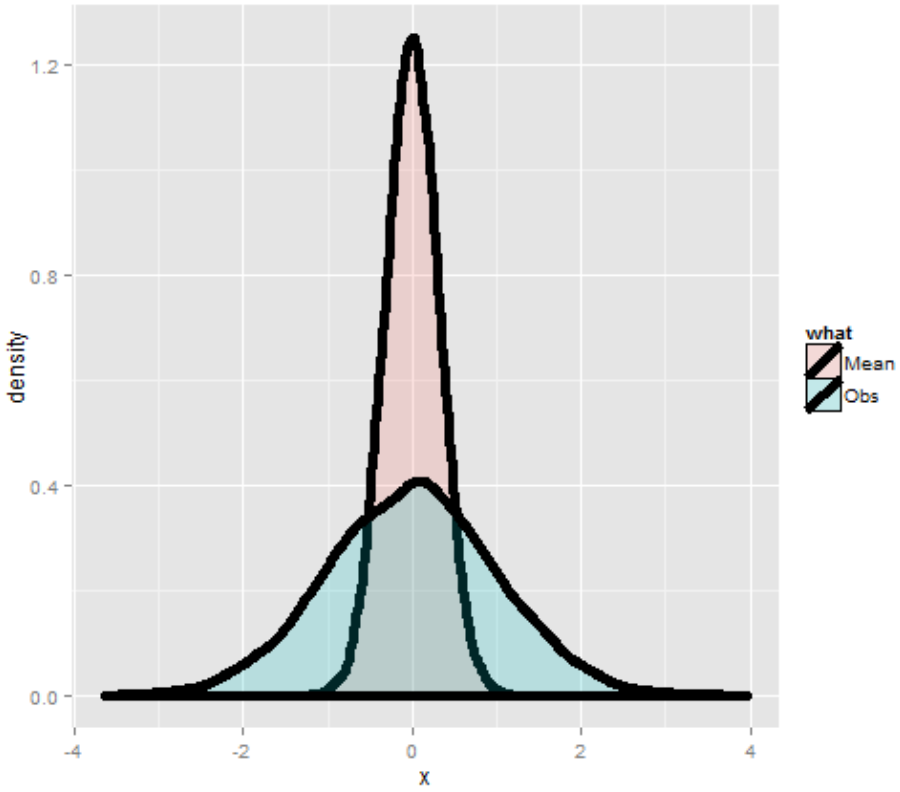


Facts about expected values

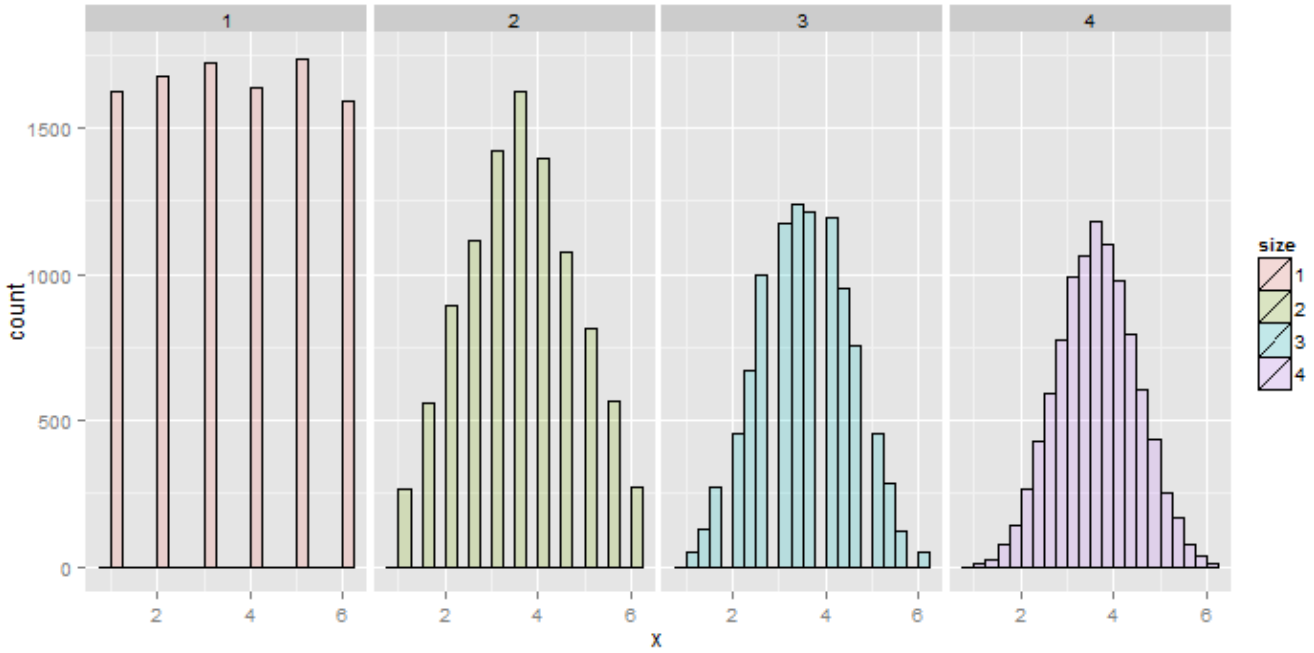
- Recall that expected values are properties of distributions
- Note the average of random variables is itself a random variable and its associated distribution has an expected value
- The center of this distribution is the same as that of the original distribution
- Therefore, the expected value of the **sample mean** is the population mean that it's trying to estimate
- When the expected value of an estimator is what its trying to estimate, we say that the estimator is **unbiased**
- Let's try a simulation experiment

Simulation experiment

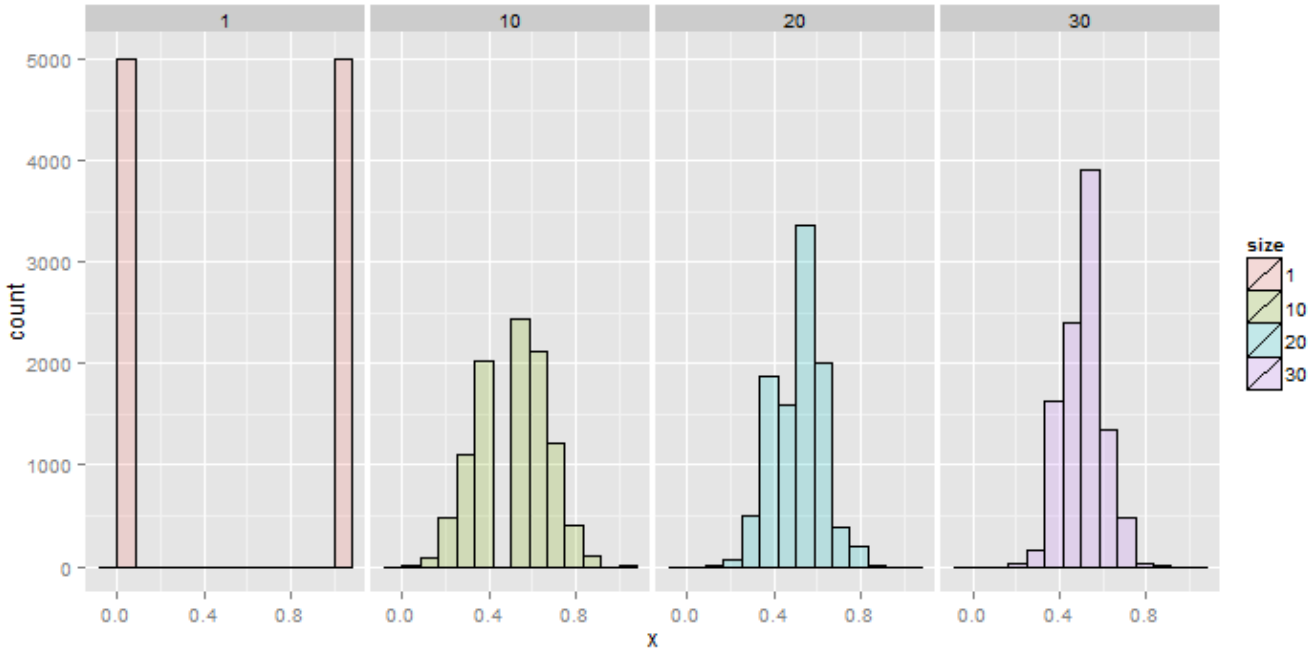
Simulating normals with mean 0 and variance 1 versus averages of 10 normals from the same population



Averages of x die rolls



Averages of x coin flips



Sumarizing what we know

- Expected values are properties of distributions
- The population mean is the center of mass of population
- The sample mean is the center of mass of the observed data
- The sample mean is an estimate of the population mean
- The sample mean is unbiased
  - The population mean of its distribution is the mean that it's trying to estimate
- The more data that goes into the sample mean, the more concentrated its density / mass function is around the population mean

The variance

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

The variance

- The variance of a random variable is a measure of *spread*
- If X is a random variable with mean μ, the variance of X is defined as

$$Var(X) = E[(X - \mu)^2] = E[X^2] - E[X]^2$$

- The expected (squared) distance from the mean
- Densities with a higher variance are more spread out than densities with a lower variance
- The square root of the variance is called the **standard deviation**
- The standard deviation has the same units as X

Example

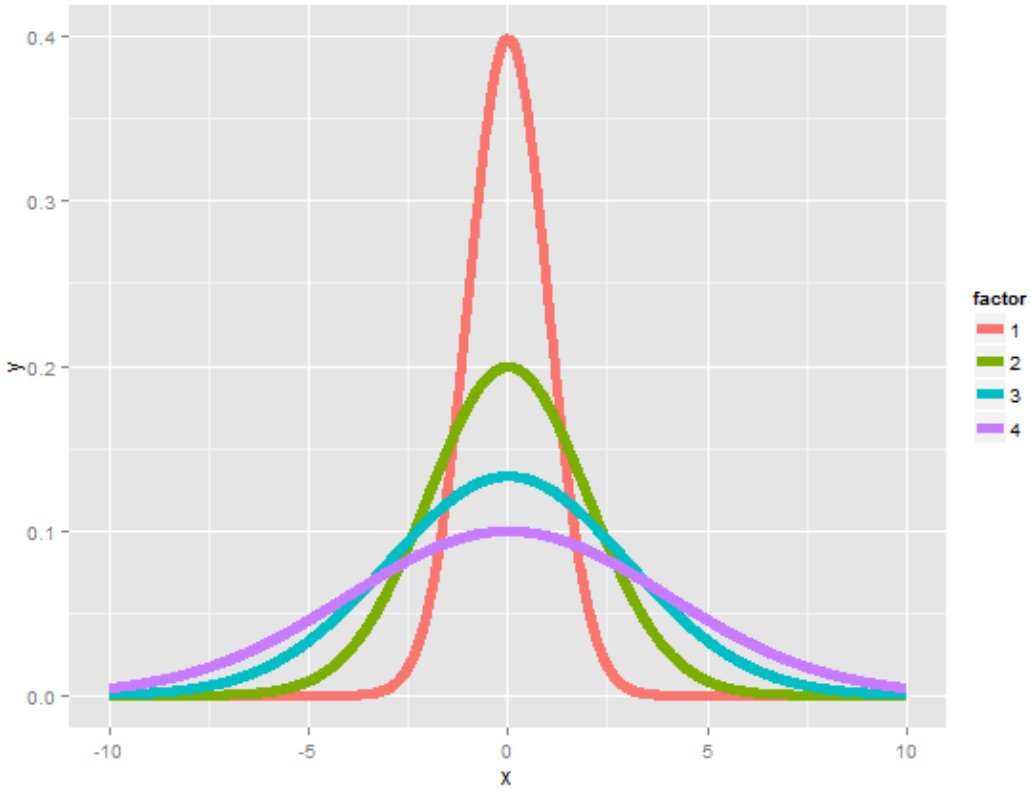
- What's the variance from the result of a toss of a die?
  - $E[X] = 3.5$
  - $E[X^2] = 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + 3^2 \times \frac{1}{6} + 4^2 \times \frac{1}{6} + 5^2 \times \frac{1}{6} + 6^2 \times \frac{1}{6} = 15.17$
- $Var(X) = E[X^2] - E[X]^2 \approx 2.92$

Example

- What's the variance from the result of the toss of a coin with probability of heads (1) of p?
  - $E[X] = 0 \times (1-p) + 1 \times p = p$
  - $E[X^2] = E[X] = p$

$$Var(X) = E[X^2] - E[X]^2 = p - p^2 = p(1-p)$$

Distributions with increasing variance



The sample variance

- The sample variance is

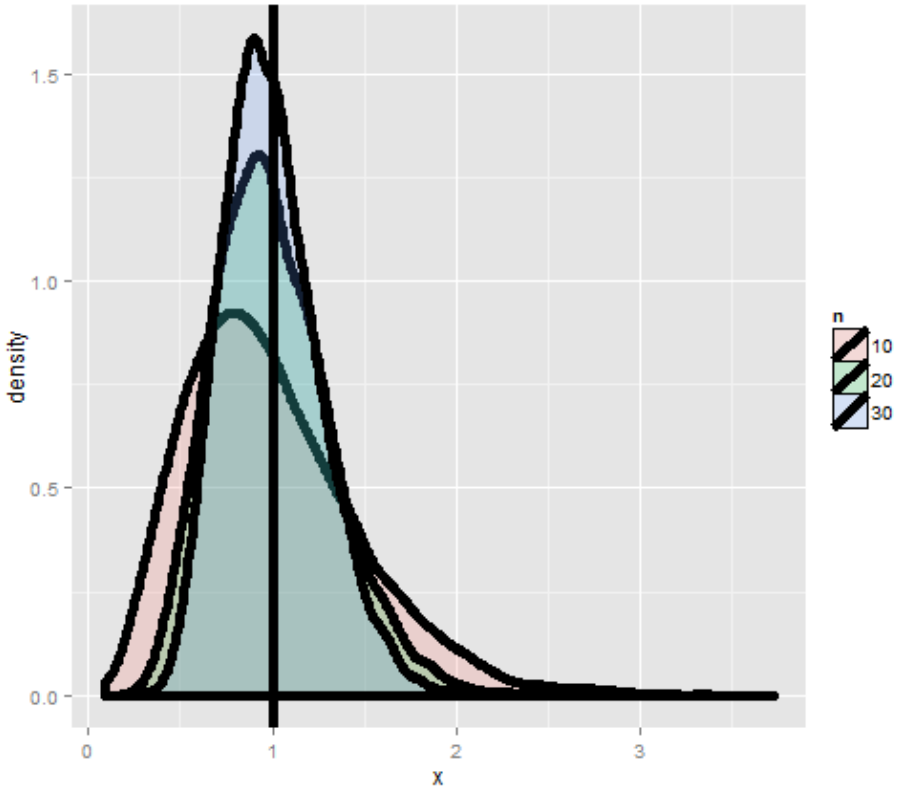
$$S^2 = \frac{\sum_{i=1}(X_i - \bar{X})^2}{n-1}$$

  (almost, but not quite, the average squared deviation from the sample mean)

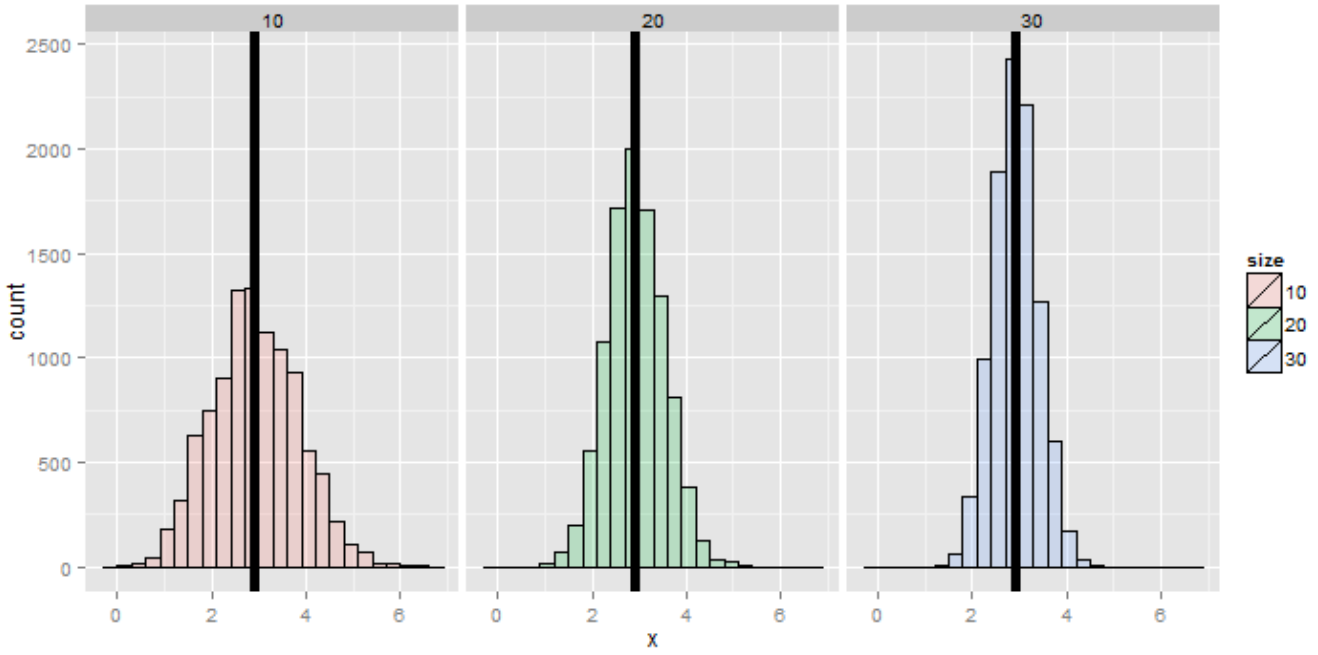
- It is also a random variable
  - It has an associate population distribution
  - Its expected value is the population variance
  - Its distribution gets more concentrated around the population variance with mroe data
- Its square root is the sample standard deviation

Simulation experiment

Simulating from a population with variance 1



Variances of x die rolls



Recall the mean

- Recall that the average of random sample from a population is itself a random variable
- We know that this distribution is centered around the population mean, $E[\bar{X}] = \mu$
- We also know what its variance is $Var(\bar{X}) = \sigma^2/n$
- This is very useful, since we don't have repeat sample means to get its variance; now we know how it relates to the population variance
- We call the standard deviation of a statistic a standard error

To summarize

- The sample variance, S^2, estimates the population variance, σ^2
- The distribution of the sample variance is centered around σ^2
- The the variance of sample mean is σ^2/n
  - Its logical estimate is s^2/n
  - The logical estimate of the standard error is $S/\sqrt{n}$
- S, the standard deviation, talks about how variable the population is
- $S/\sqrt{n}$, the standard error, talks about how variable averages of random samples of size n from the population are

Simulation example

Standard normals have variance 1; means of n standard normals have standard deviation $1/\sqrt{n}$

```
nosim <- 1000
n <- 10
sd(apply(matrix(rnorm(nosim * n), nosim), 1, mean))
```

```
## [1] 0.3156
```

```
1 / sqrt(n)
```

```
## [1] 0.3162
```

Simulation example

Standard uniforms have variance $1/12$; means of random samples of n uniforms have sd $1/\sqrt{12 \times n}$

```
nosim <- 1000
n <- 10
sd(apply(matrix(runif(nosim * n), nosim), 1, mean))
```

```
## [1] 0.09017
```

```
1 / sqrt(12 * n)
```

```
## [1] 0.09129
```

Simulation example

Poisson(4) have variance 4; means of random samples of n Poisson(4) have sd $2/\sqrt{n}$

```
nosim <- 1000
n <- 10
sd(apply(matrix(rpois(nosim * n, 4), nosim), 1, mean))
```

```
## [1] 0.6219
```

```
2 / sqrt(n)
```

```
## [1] 0.6325
```

Simulation example

Fair coin flips have variance 0.25; means of random samples of n coin flips have sd $1/(2\sqrt{n})$

```
nosim <- 1000
n <- 10
sd(apply(matrix(sample(0 : 1, nosim * n, replace = TRUE),
                nosim), 1, mean))
```

```
## [1] 0.1587
```

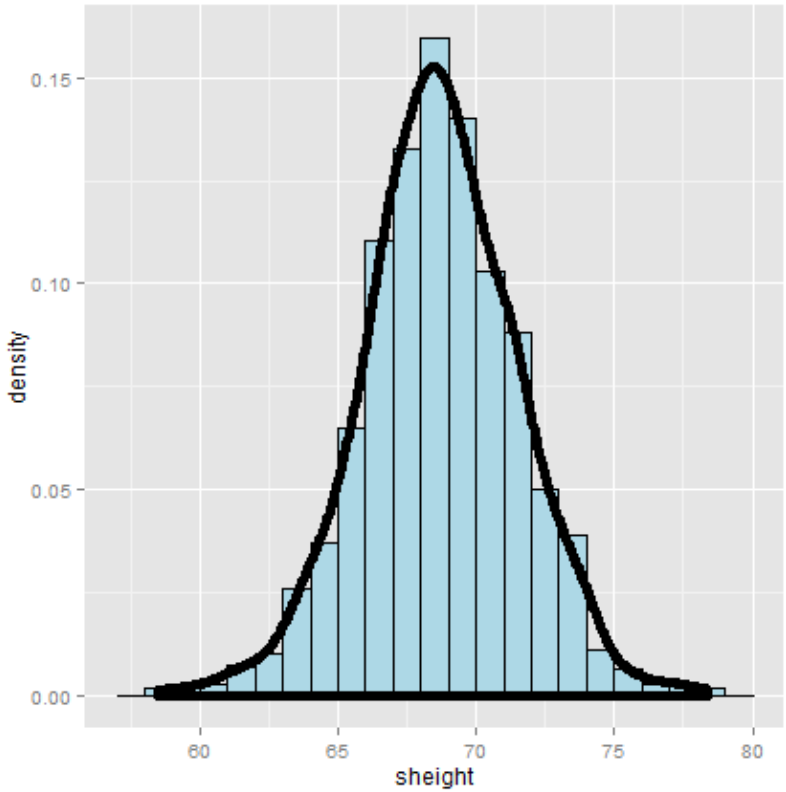
```
1 / (2 * sqrt(n))
```

```
## [1] 0.1581
```

Data example

```
library(UsingR); data(father.son);
x <- father.son$sheight
n<-length(x)
```

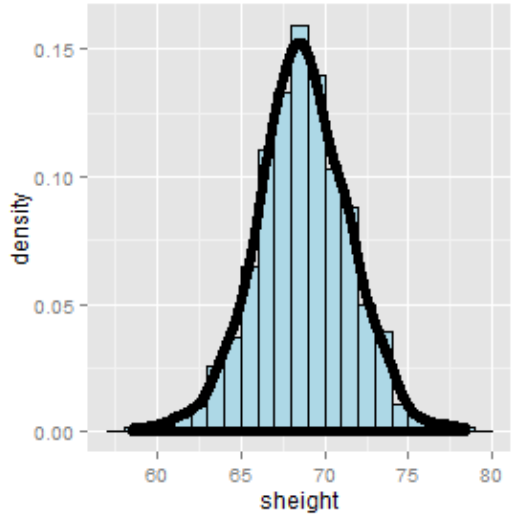
Plot of the son's heights



Let's interpret these numbers

```
round(c(var(x), var(x) / n, sd(x), sd(x) / sqrt(n)),2)
```

```
## [1] 7.92 0.01 2.81 0.09
```



Summarizing what we know about variances

- The sample variance estimates the population variance
- The distribution of the sample variance is centered at what its estimating
- It gets more concentrated around the population variance with larger sample sizes
- The variance of the sample mean is the population variance divided by n
  - The square root is the standard error
- It turns out that we can say a lot about the distribution of averages from random samples, even though we only get one to look at in a given data set

Some Common Distributions

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

The Bernoulli distribution

- The **Bernoulli distribution** arises as the result of a binary outcome
- Bernoulli random variables take (only) the values 1 and 0 with probabilities of (say) p and $1-p$ respectively
- The PMF for a Bernoulli random variable X is

$$P(X=x) = p^x(1-p)^{1-x}$$

- The mean of a Bernoulli random variable is p and the variance is $p(1-p)$
- If we let X be a Bernoulli random variable, it is typical to call $X=1$ as a "success" and $X=0$ as a "failure"

Binomial trials

- The *binomial random variables* are obtained as the sum of iid Bernoulli trials
- In specific, let $X_1, \ldots, X_n$ be iid Bernoulli(p); then $X = \sum_{i=1}^{n} X_i$ is a binomial random variable
- The binomial mass function is

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

for $x = 0, \ldots, n$

Choose

- Recall that the notation

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

(read "n choose x") counts the number of ways of selecting x items out of n without replacement disregarding the order of the items

$$\binom{n}{0} = \binom{n}{n} = 1$$

Example

- Suppose a friend has 8 children (oh my!), 7 of which are girls and none are twins
- If each gender has an independent 50% probability for each birth, what's the probability of getting 7 or more girls out of 8 births?

$$\binom{8}{7}.5^7(1-.5)^1 + \binom{8}{8}.5^8(1-.5)^0 \approx 0.04$$

```
choose(8, 7) * 0.5^8 + choose(8, 8) * 0.5^8
```

```
## [1] 0.03516
```

```
pbinom(6, size = 8, prob = 0.5, lower.tail = FALSE)
```

```
## [1] 0.03516
```

The normal distribution

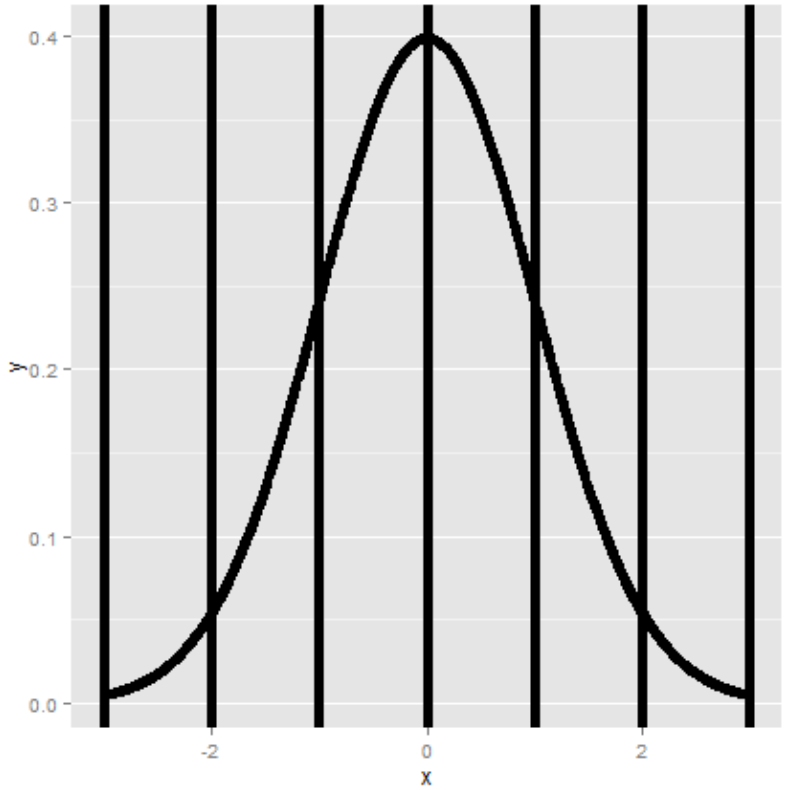
- A random variable is said to follow a **normal** or **Gaussian** distribution with mean μ and variance σ^2 if the associated density is

$$(2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/2\sigma^2}$$

  If X a RV with this density then $E[X] = \mu$ and $Var(X) = \sigma^2$

- We write $X \sim \mathrm{N}(\mu, \sigma^2)$
- When $\mu = 0$ and $\sigma = 1$ the resulting distribution is called **the standard normal distribution**
- Standard normal RVs are often labeled Z

The standard normal distribution with reference lines



Facts about the normal density

If $X \sim \mathrm{N}(\mu, \sigma^2)$ then

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

If Z is standard normal

$$X = \mu + \sigma Z \sim \mathrm{N}(\mu, \sigma^2)$$

More facts about the normal density

1. Approximately 68%, 95% and 99% of the normal density lies within 1, 2 and 3 standard deviations from the mean, respectively
2. -1.28, -1.645, -1.96 and -2.33 are the 10^{th}, 5^{th}, 2.5^{th} and 1^{st} percentiles of the standard normal distribution respectively
3. By symmetry, 1.28, 1.645, 1.96 and 2.33 are the 90^{th}, 95^{th}, 97.5^{th} and 99^{th} percentiles of the standard normal distribution respectively

Question

- What is the 95^{th} percentile of a $N(\mu, \sigma^2)$ distribution?
  - Quick answer in R `qnorm(.95, mean = mu, sd = sd)`
- Or, because you have the standard normal quantiles memorized and you know that 1.645 is the 95th percentile you know that the answer has to be

$$\mu + \sigma 1.645$$

- (In general $\mu + \sigma z_0$ where z_0 is the appropriate standard normal quantile)

Question

- What is the probability that a $\mathrm{N}(\mu, \sigma^2)$ RV is larger than x?

Example

Assume that the number of daily ad clicks for a company is (approximately) normally distributed with a mean of 1020 and a standard deviation of 50. What's the probability of getting more than 1,160 clicks in a day?

Example

Assume that the number of daily ad clicks for a company is (approximately) normally distributed with a mean of 1020 and a standard deviation of 50. What's the probability of getting more than 1,160 clicks in a day?

It's not very likely, 1,160 is 2.8 standard deviations from the mean

```
pnorm(1160, mean = 1020, sd = 50, lower.tail = FALSE)
```

```
## [1] 0.002555
```

```
pnorm(2.8, lower.tail = FALSE)
```

```
## [1] 0.002555
```

Example

Assume that the number of daily ad clicks for a company is (approximately) normally distributed with a mean of 1020 and a standard deviation of 50. What number of daily ad clicks would represent the one where 75% of days have fewer clicks (assuming days are independent and identically distributed)?

Example

Assume that the number of daily ad clicks for a company is (approximately) normally distributed with a mean of 1020 and a standard deviation of 50. What number of daily ad clicks would represent the one where 75% of days have fewer clicks (assuming days are independent and identically distributed)?

```
qnorm(0.75, mean = 1020, sd = 50)
```

```
## [1] 1054
```

The Poisson distribution

- Used to model counts
- The Poisson mass function is

$$P(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

  for $x = 0, 1, \ldots$
- The mean of this distribution is λ
- The variance of this distribution is λ
- Notice that x ranges from 0 to ∞

Some uses for the Poisson distribution

- Modeling count data
- Modeling event-time or survival data
- Modeling contingency tables
- Approximating binomials when n is large and p is small

Rates and Poisson random variables

- Poisson random variables are used to model rates
- $X \sim Poisson(\lambda t)$ where
  - $\lambda = E[X/t]$ is the expected count per unit of time
  - t is the total monitoring time

Example

The number of people that show up at a bus stop is Poisson with a mean of 2.5 per hour.

If watching the bus stop for 4 hours, what is the probability that 3 or fewer people show up for the whole time?

```
ppois(3, lambda = 2.5 * 4)
```

```
## [1] 0.01034
```

Poisson approximation to the binomial

- When n is large and p is small the Poisson distribution is an accurate approximation to the binomial distribution
- Notation
  - $X \sim \text{Binomial}(n, p)$
  - $\lambda = np$
  - n gets large
  - p gets small

Example, Poisson approximation to the binomial

We flip a coin with success probablity 0.01 five hundred times.

What's the probability of 2 or fewer successes?

```
pbinom(2, size = 500, prob = 0.01)
```

```
## [1] 0.1234
```

```
ppois(2, lambda = 500 * 0.01)
```

```
## [1] 0.1247
```

A trip to Asymptopia

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

Asymptotics

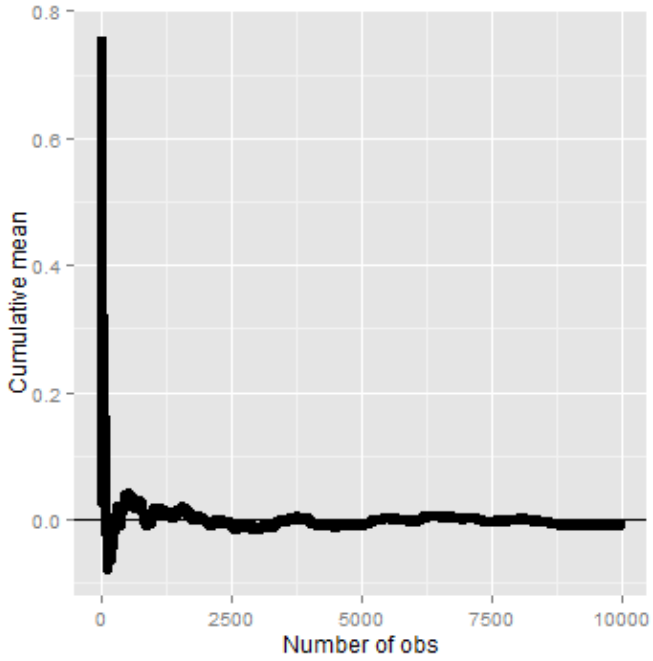
- Asymptotics is the term for the behavior of statistics as the sample size (or some other relevant quantity) limits to infinity (or some other relevant number)
- (Asymptopia is my name for the land of asymptotics, where everything works out well and there's no messes. The land of infinite data is nice that way.)
- Asymptotics are incredibly useful for simple statistical inference and approximations
- (Not covered in this class) Asymptotics often lead to nice understanding of procedures
- Asymptotics generally give no assurances about finite sample performance
- Asymptotics form the basis for frequency interpretation of probabilities (the long run proportion of times an event occurs)

Limits of random variables

- Fortunately, for the sample mean there's a set of powerful results
- These results allow us to talk about the large sample distribution of sample means of a collection of iid observations
- The first of these results we inuitively know
  - It says that the average limits to what its estimating, the population mean
  - It's called the Law of Large Numbers
  - Example $\bar{X}_n$ could be the average of the result of n coin flips (i.e. the sample proportion of heads)
  - As we flip a fair coin over and over, it evetually converges to the true probability of a head The LLN forms the basis of frequency style thinking

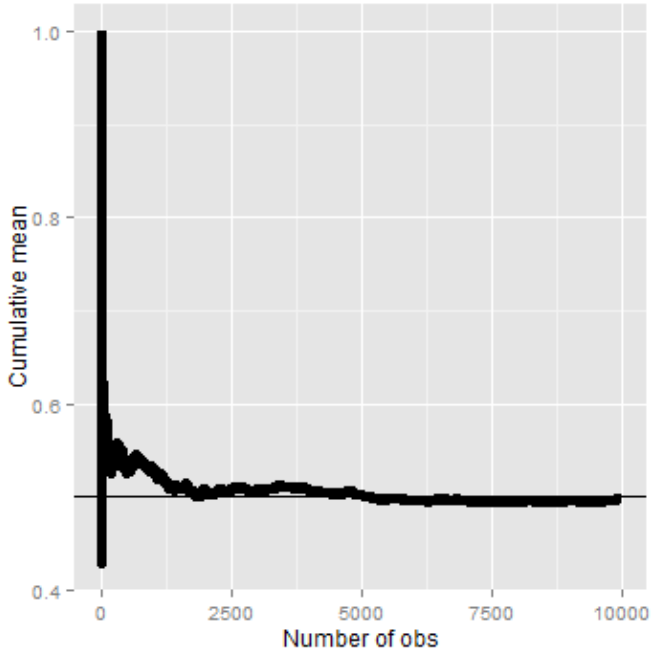
Law of large numbers in action

```
n <- 10000
means <- cumsum(rnorm(n))/(1:n)
library(ggplot2)
g <- ggplot(data.frame(x = 1:n, y = means), aes(x = x, y = y))
g <- g + geom_hline(yintercept = 0) + geom_line(size = 2)
g <- g + labs(x = "Number of obs", y = "Cumulative mean")
g
```



Law of large numbers in action, coin flip

```
means <- cumsum(sample(0:1, n, replace = TRUE))/(1:n)
g <- ggplot(data.frame(x = 1:n, y = means), aes(x = x, y = y))
g <- g + geom_hline(yintercept = 0.5) + geom_line(size = 2)
g <- g + labs(x = "Number of obs", y = "Cumulative mean")
g
```



Discussion

- An estimator is **consistent** if it converges to what you want to estimate
  - The LLN says that the sample mean of iid sample is consistent for the population mean
  - Typically, good estimators are consistent; it's not too much to ask that if we go to the trouble of collecting an infinite amount of data that we get the right answer
- The sample variance and the sample standard deviation of iid random variables are consistent as well

The Central Limit Theorem

- The **Central Limit Theorem** (CLT) is one of the most important theorems in statistics
- For our purposes, the CLT states that the distribution of averages of iid variables (properly normalized) becomes that of a standard normal as the sample size increases
- The CLT applies in an endless variety of settings
- The result is that

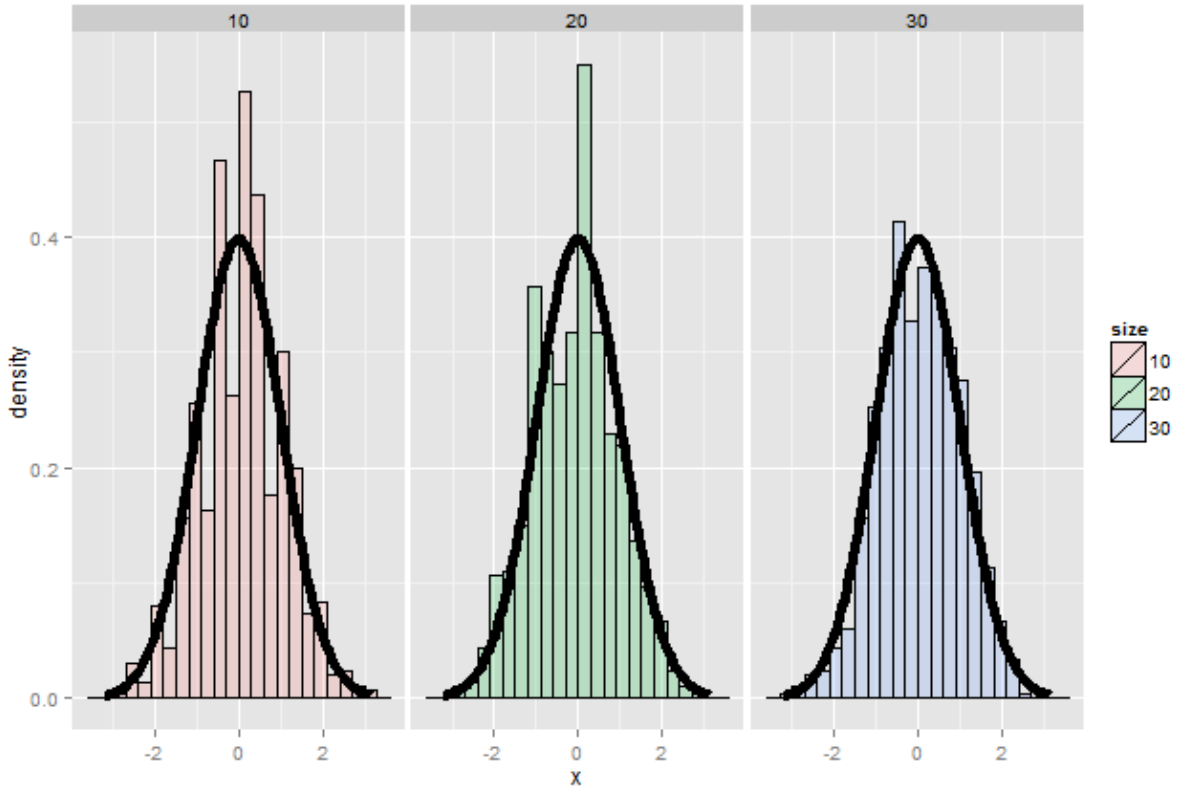
$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{\text{Estimate} - \text{Mean of estimate}}{\text{Std. Err. of estimate}}$$

  has a distribution like that of a standard normal for large n.
- (Replacing the standard error by its estimated value doesn't change the CLT)
- The useful way to think about the CLT is that $\bar{X}_n$ is approximately $N(\mu, \sigma^2/n)$

Example

- Simulate a standard normal random variable by rolling n (six sided)
- Let X_i be the outcome for die i
- Then note that $\mu = E[X_i] = 3.5$
- $Var(X_i) = 2.92$
- SE $\sqrt{2.92/n} = 1.71/\sqrt{n}$
- Lets roll n dice, take their mean, subtract off 3.5, and divide by $1.71/\sqrt{n}$ and repeat this over and over

Result of our die rolling experiment



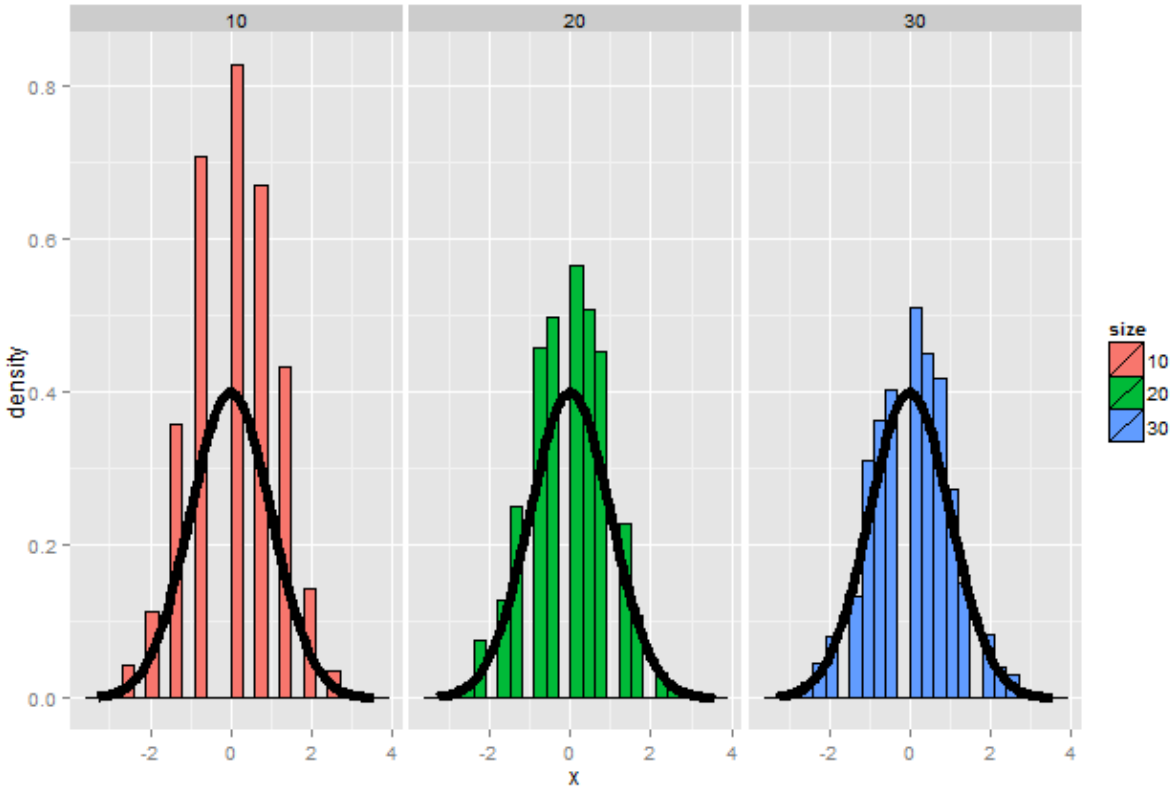
Coin CLT

- Let X_i be the 0 or 1 result of the i^{th} flip of a possibly unfair coin
  - The sample proportion, say $\hat{p}$, is the average of the coin flips
  - $E[X_i] = p$ and $Var(X_i) = p(1-p)$
  - Standard error of the mean is $\sqrt{p(1-p)/n}$
  - Then

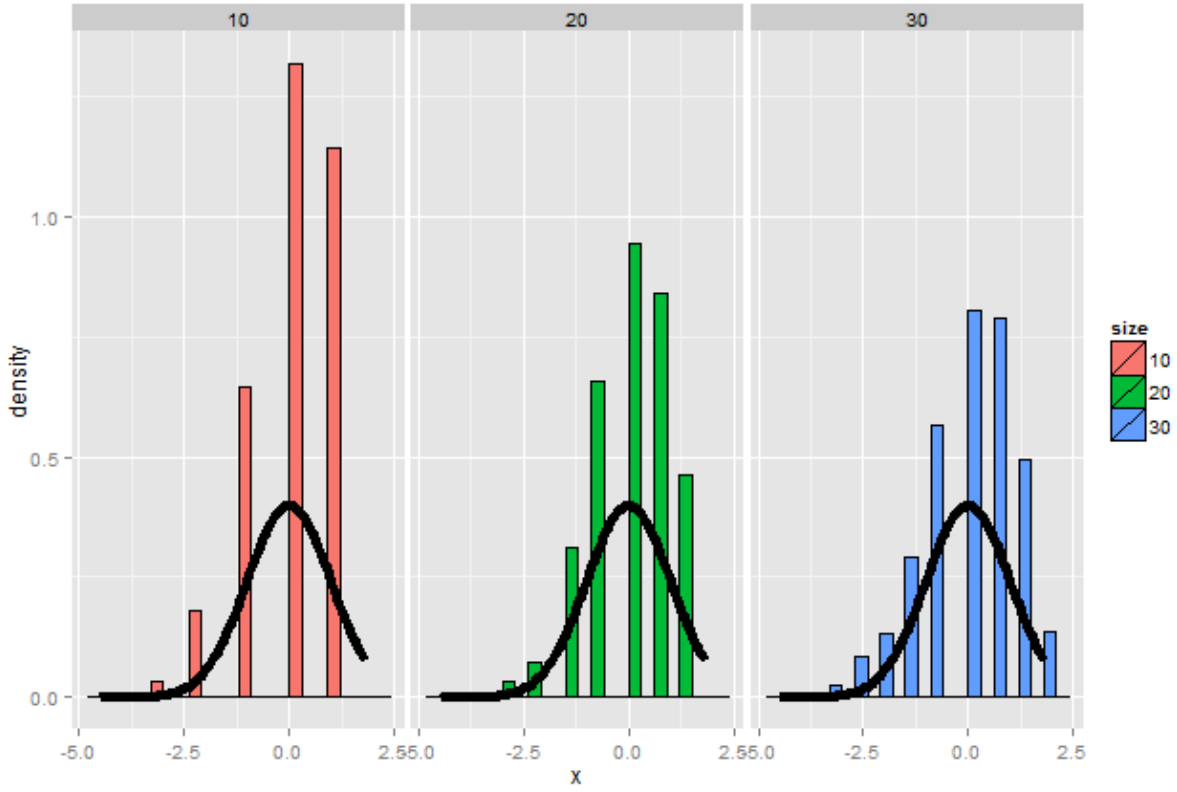
$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

    will be approximately normally distributed
  - Let's flip a coin n times, take the sample proportion of heads, subtract off .5 and multiply the result by $2\sqrt{n}$ (divide by $1/(2\sqrt{n})$)

Simulation results

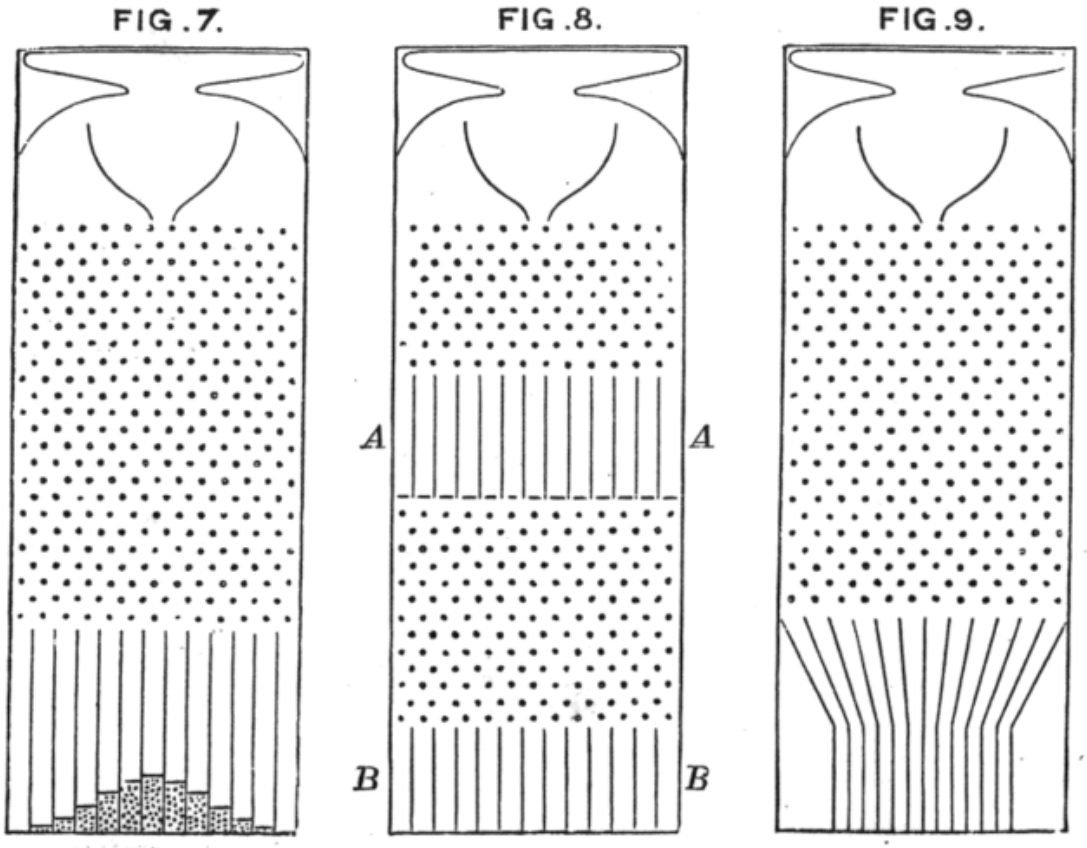


Simulation results, $p = 0.9$



Galton's quincunx

http://en.wikipedia.org/wiki/Bean_machine#mediaviewer/File:Quincunx_(Galton_Box)_-_Galton_1889_diagram.png



Confidence intervals

- According to the CLT, the sample mean, $\bar{X}$, is approximately normal with mean μ and sd $\sigma/\sqrt{n}$
- $\mu + 2\sigma/\sqrt{n}$ is pretty far out in the tail (only 2.5% of a normal being larger than 2 sds in the tail)
- Similarly, $\mu - 2\sigma/\sqrt{n}$ is pretty far in the left tail (only 2.5% chance of a normal being smaller than 2 sds in the tail)
- So the probability $\bar{X}$ is bigger than $\mu + 2\sigma/\sqrt{n}$ or smaller than $\mu - 2\sigma/\sqrt{n}$ is 5%
  - Or equivalently, the probability of being between these limits is 95%
- The quantity $\bar{X} \pm 2\sigma/\sqrt{n}$ is called a 95% interval for μ
- The 95% refers to the fact that if one were to repeatly get samples of size n, about 95% of the intervals obtained would contain μ
- The 97.5th quantile is 1.96 (so I rounded to 2 above)
- 90% interval you want (100 - 90) / 2 = 5% in each tail
  - So you want the 95th percentile (1.645)

Give a confidence interval for the average height of sons

in Galton's data

```
library(UsingR)
data(father.son)
x <- father.son$sheight
(mean(x) + c(-1, 1) * qnorm(0.975) * sd(x)/sqrt(length(x)))/12
```

```
## [1] 5.710 5.738
```

Sample proportions

- In the event that each X_i is 0 or 1 with common success probability p then $\sigma^2 = p(1-p)$
- The interval takes the form

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

- Replacing p by $\hat{p}$ in the standard error results in what is called a Wald confidence interval for p
- For 95% intervals

$$\hat{p} \pm \frac{1}{\sqrt{n}}$$

  is a quick CI estimate for p

Example

- Your campaign advisor told you that in a random sample of 100 likely voters, 56 intent to vote for you.
  - Can you relax? Do you have this race in the bag?
  - Without access to a computer or calculator, how precise is this estimate?
- `1/sqrt(100)=0.1` so a back of the envelope calculation gives an approximate 95% interval of `(0.46, 0.66)`
  - Not enough for you to relax, better go do more campaigning!
- Rough guidelines, 100 for 1 decimal place, 10,000 for 2, 1,000,000 for 3.

```
round(1/sqrt(10^(1:6)), 3)
```

```
## [1] 0.316 0.100 0.032 0.010 0.003 0.001
```

Binomial interval

```
0.56 + c(-1, 1) * qnorm(0.975) * sqrt(0.56 * 0.44/100)
```

```
## [1] 0.4627 0.6573
```

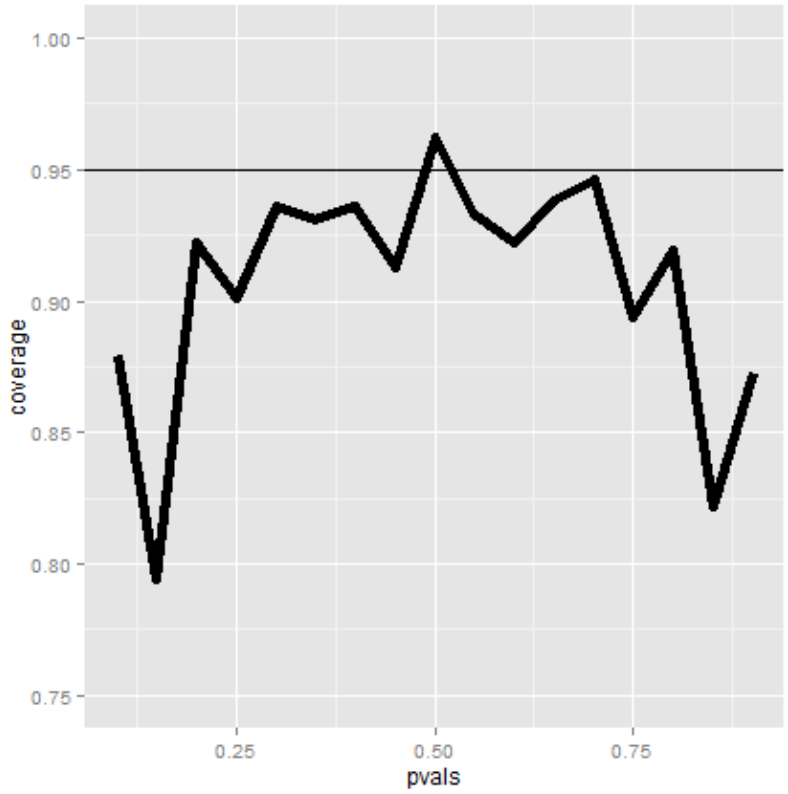
```
binom.test(56, 100)$conf.int
```

```
## [1] 0.4572 0.6592
## attr(,"conf.level")
## [1] 0.95
```

Simulation

```
n <- 20
pvals <- seq(0.1, 0.9, by = 0.05)
nosim <- 1000
coverage <- sapply(pvals, function(p) {
    phats <- rbinom(nosim, prob = p, size = n)/n
    ll <- phats - qnorm(0.975) * sqrt(phats * (1 - phats)/n)
    ul <- phats + qnorm(0.975) * sqrt(phats * (1 - phats)/n)
    mean(ll < p & ul > p)
})
```

Plot of the results (not so good)



What's happening?

- n isn't large enough for the CLT to be applicable for many of the values of p
- Quick fix, form the interval with

$$\frac{X+2}{n+4}$$

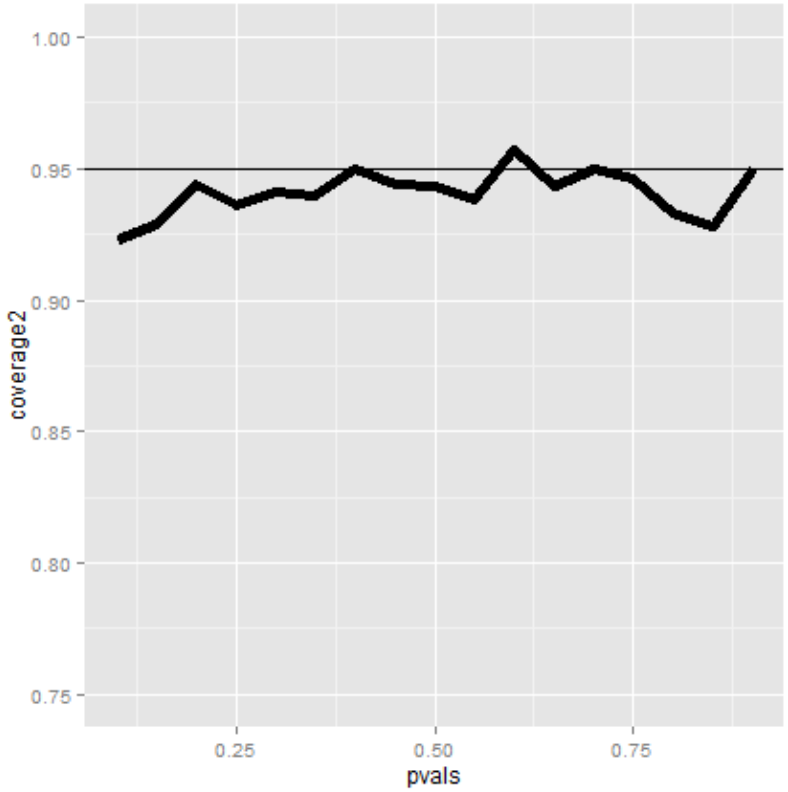
- (Add two successes and failures, Agresti/Coull interval)

Simulation

First let's show that coverage gets better with n

```
n <- 100
pvals <- seq(0.1, 0.9, by = 0.05)
nosim <- 1000
coverage2 <- sapply(pvals, function(p) {
    phats <- rbinom(nosim, prob = p, size = n)/n
    ll <- phats - qnorm(0.975) * sqrt(phats * (1 - phats)/n)
    ul <- phats + qnorm(0.975) * sqrt(phats * (1 - phats)/n)
    mean(ll < p & ul > p)
})
```

Plot of coverage for $n = 100$



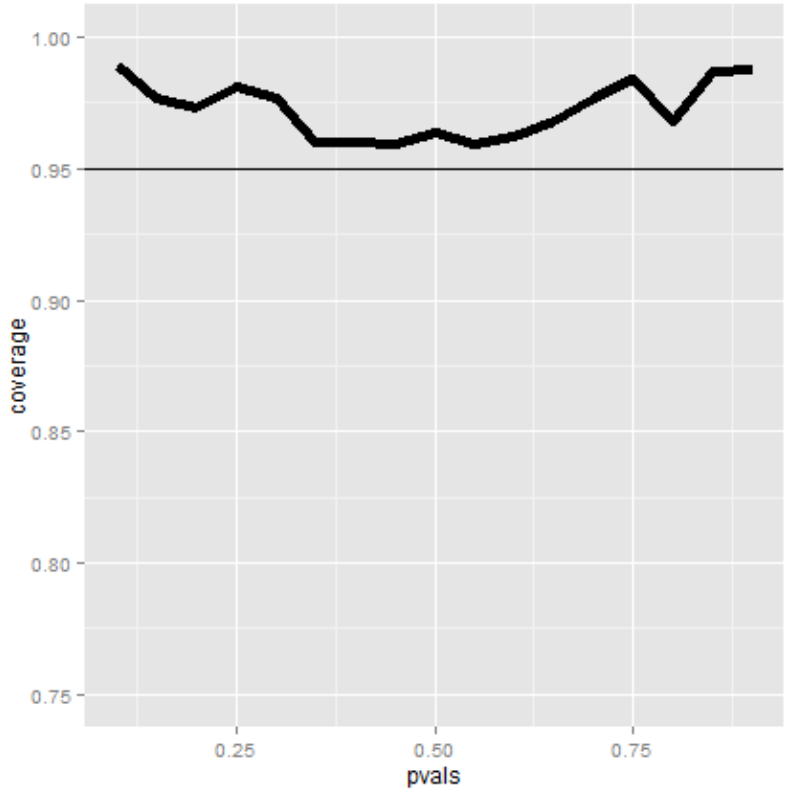
Simulation

Now let's look at $n = 20$ but adding 2 successes and failures

```
n <- 20
pvals <- seq(0.1, 0.9, by = 0.05)
nosim <- 1000
coverage <- sapply(pvals, function(p) {
    phats <- (rbinom(nosim, prob = p, size = n) + 2)/(n + 4)
    ll <- phats - qnorm(0.975) * sqrt(phats * (1 - phats)/n)
    ul <- phats + qnorm(0.975) * sqrt(phats * (1 - phats)/n)
    mean(ll < p & ul > p)
})
```

Adding 2 successes and 2 failures

(It's a little conservative)



Poisson interval

- A nuclear pump failed 5 times out of 94.32 days, give a 95% confidence interval for the failure rate per day?
- $X \sim Poisson(\lambda t)$.
- Estimate $\hat{\lambda} = X/t$
- $Var(\hat{\lambda}) = \lambda/t$
- $\hat{\lambda}/t$ is our variance estimate

R code

```
x <- 5
t <- 94.32
lambda <- x/t
round(lambda + c(-1, 1) * qnorm(0.975) * sqrt(lambda/t), 3)
```

```
## [1] 0.007 0.099
```

```
poisson.test(x, T = 94.32)$conf
```

```
## [1] 0.01721 0.12371
## attr(,"conf.level")
## [1] 0.95
```

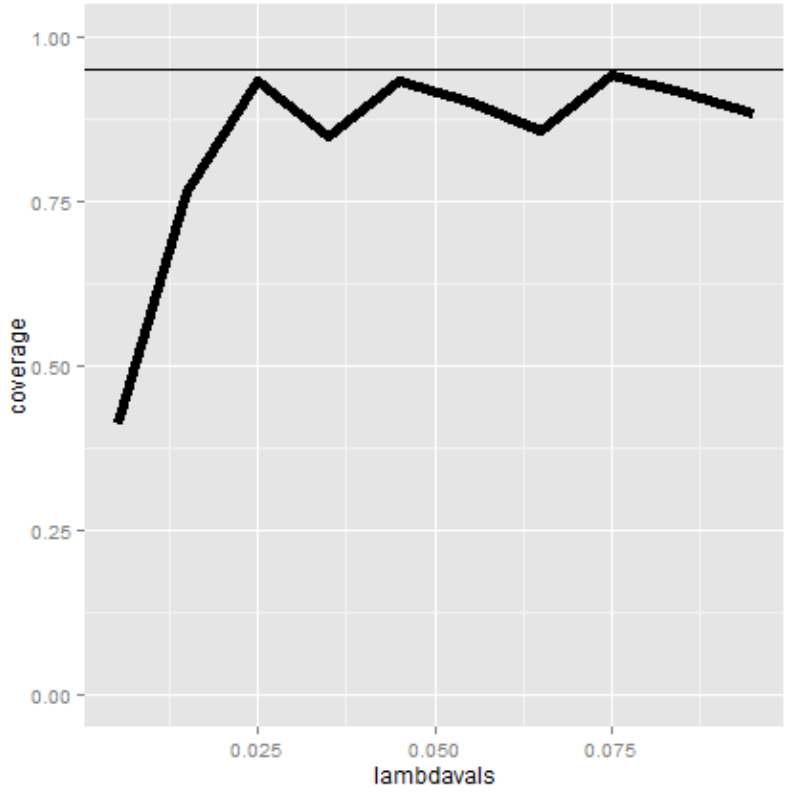
Simulating the Poisson coverage rate

Let's see how this interval performs for lambda values near what we're estimating

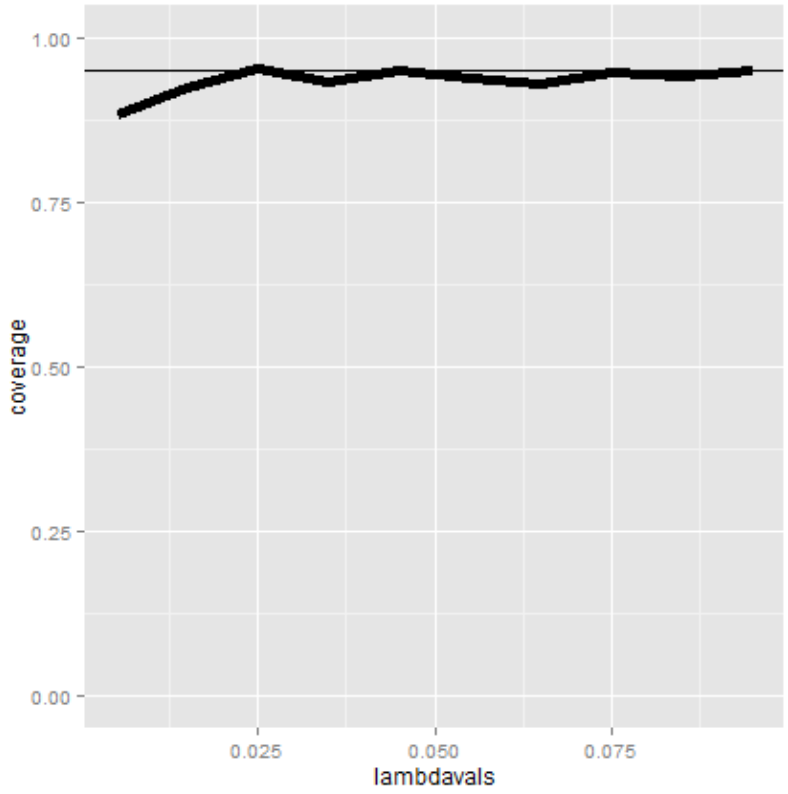
```
lambdavals <- seq(0.005, 0.1, by = 0.01)
nosim <- 1000
t <- 100
coverage <- sapply(lambdavals, function(lambda) {
    lhats <- rpois(nosim, lambda = lambda * t)/t
    ll <- lhats - qnorm(0.975) * sqrt(lhats/t)
    ul <- lhats + qnorm(0.975) * sqrt(lhats/t)
    mean(ll < lambda & ul > lambda)
})
```

Covarage

(Gets really bad for small values of lambda)



What if we increase t to 1000?



Summary

- The LLN states that averages of iid samples converge to the population means that they are estimating
- The CLT states that averages are approximately normal, with distributions
  - centered at the population mean
  - with standard deviation equal to the standard error of the mean
  - CLT gives no guarantee that n is large enough
- Taking the mean and adding and subtracting the relevant normal quantile times the SE yields a confidence interval for the mean
  - Adding and subtracting 2 SEs works for 95% intervals
- Confidence intervals get wider as the coverage increases (why?)
- Confidence intervals get narrower with less variability or larger sample sizes
- The Poisson and binomial case have exact intervals that don't require the CLT
  - But a quick fix for small sample size binomial calculations is to add 2 successes and failures

T Confidence Intervals

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

T Confidence intervals

- In the previous, we discussed creating a confidence interval using the CLT
  - They took the form $Est \pm ZQ \times SE_{Est}$
- In this lecture, we discuss some methods for small samples, notably Gosset's t distribution and t confidence intervals
  - They are of the form $Est \pm TQ \times SE_{Est}$
- These are some of the handiest of intervals
- If you want a rule between whether to use a t interval or normal interval, just always use the t interval
- We'll cover the one and two group versions

Gosset's t distribution

- Invented by William Gosset (under the pseudonym "Student") in 1908
- Has thicker tails than the normal
- Is indexed by a degrees of freedom; gets more like a standard normal as df gets larger
- It assumes that the underlying data are iid Gaussian with the result that

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

  follows Gosset's t distribution with $n - 1$ degrees of freedom
- (If we replaced s by σ the statistic would be exactly standard normal)
- Interval is $\bar{X} \pm t_{n-1} S/\sqrt{n}$ where t_{n-1} is the relevant quantile

Code for manipulate

```
k <- 1000
xvals <- seq(-5, 5, length = k)
myplot <- function(df){
  d <- data.frame(y = c(dnorm(xvals), dt(xvals, df)),
                  x = xvals,
                  dist = factor(rep(c("Normal", "T"), c(k,k))))
  g <- ggplot(d, aes(x = x, y = y))
  g <- g + geom_line(size = 2, aes(colour = dist))
  g
}
manipulate(myplot(mu), mu = slider(1, 20, step = 1))
```

Easier to see

```
pvals <- seq(.5, .99, by = .01)
myplot2 <- function(df){
  d <- data.frame(n= qnorm(pvals),t=qt(pvals, df),
                  p = pvals)
  g <- ggplot(d, aes(x= n, y = t))
  g <- g + geom_abline(size = 2, col = "lightblue")
  g <- g + geom_line(size = 2, col = "black")
  g <- g + geom_vline(xintercept = qnorm(0.975))
  g <- g + geom_hline(yintercept = qt(0.975, df))
  g
}
manipulate(myplot2(df), df = slider(1, 20, step = 1))
```

Note's about the t interval

- The t interval technically assumes that the data are iid normal, though it is robust to this assumption
- It works well whenever the distribution of the data is roughly symmetric and mound shaped
- Paired observations are often analyzed using the t interval by taking differences
- For large degrees of freedom, t quantiles become the same as standard normal quantiles; therefore this interval converges to the same interval as the CLT yielded
- For skewed distributions, the spirit of the t interval assumptions are violated
  - Also, for skewed distributions, it doesn't make a lot of sense to center the interval at the mean
  - In this case, consider taking logs or using a different summary like the median
- For highly discrete data, like binary, other intervals are available

Sleep data

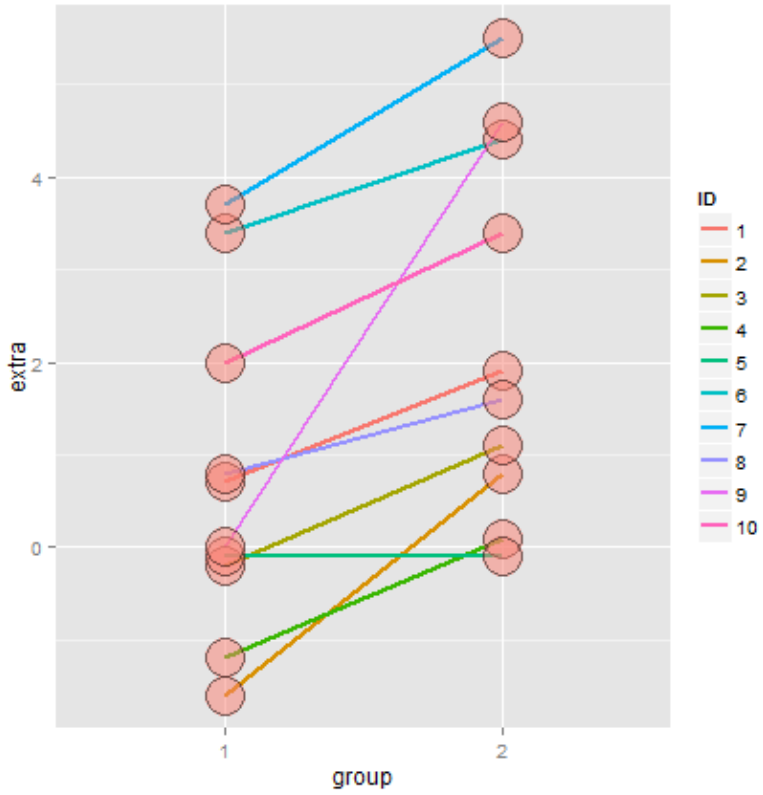
In R typing `data(sleep)` brings up the sleep data originally analyzed in Gosset's Biometrika paper, which shows the increase in hours for 10 patients on two soporific drugs. R treats the data as two groups rather than paired.

The data

```
data(sleep)
head(sleep)
```

```
##   extra group ID
## 1   0.7     1  1
## 2  -1.6     1  2
## 3  -0.2     1  3
## 4  -1.2     1  4
## 5  -0.1     1  5
## 6   3.4     1  6
```

Plotting the data



Results

```
g1 <- sleep$extra[1 : 10]; g2 <- sleep$extra[11 : 20]
difference <- g2 - g1
mn <- mean(difference); s <- sd(difference); n <- 10
```

```
mn + c(-1, 1) * qt(.975, n-1) * s / sqrt(n)
t.test(difference)
t.test(g2, g1, paired = TRUE)
t.test(extra ~ I(relevel(group, 2)), paired = TRUE, data = sleep)
```

The results

(After a little formatting)

```
##        [,1] [,2]
## [1,] 0.7001 2.46
## [2,] 0.7001 2.46
## [3,] 0.7001 2.46
## [4,] 0.7001 2.46
```

Independent group t confidence intervals

- Suppose that we want to compare the mean blood pressure between two groups in a randomized trial; those who received the treatment to those who received a placebo
- We cannot use the paired t test because the groups are independent and may have different sample sizes
- We now present methods for comparing independent groups

Confidence interval

- Therefore a $(1-\alpha)\times 100\%$ confidence interval for $\mu_y - \mu_x$ is

$$\bar{Y} - \bar{X} \pm t_{n_x+n_y-2,1-\alpha/2} S_p \left(\frac{1}{n_x} + \frac{1}{n_y} \right)^{1/2}$$

- The pooled variance estimator is

$$S_p^2 = \{(n_x - 1)S_x^2 + (n_y - 1)S_y^2\}/(n_x + n_y - 2)$$

- Remember this interval is assuming a constant variance across the two groups
- If there is some doubt, assume a different variance per group, which we will discuss later

Example

Based on Rosner, Fundamentals of Biostatistics

(Really a very good reference book)

- Comparing SBP for 8 oral contraceptive users versus 21 controls
- $\bar{X}_{OC} = 132.86$ mmHg with $s_{OC} = 15.34$ mmHg
- $\bar{X}_{C} = 127.44$ mmHg with $s_{C} = 18.23$ mmHg
- Pooled variance estimate

```
sp <- sqrt((7 * 15.34^2 + 20 * 18.23^2) / (8 + 21 - 2))
132.86 - 127.44 + c(-1, 1) * qt(.975, 27) * sp * (1 / 8 + 1 / 21)^.5
```

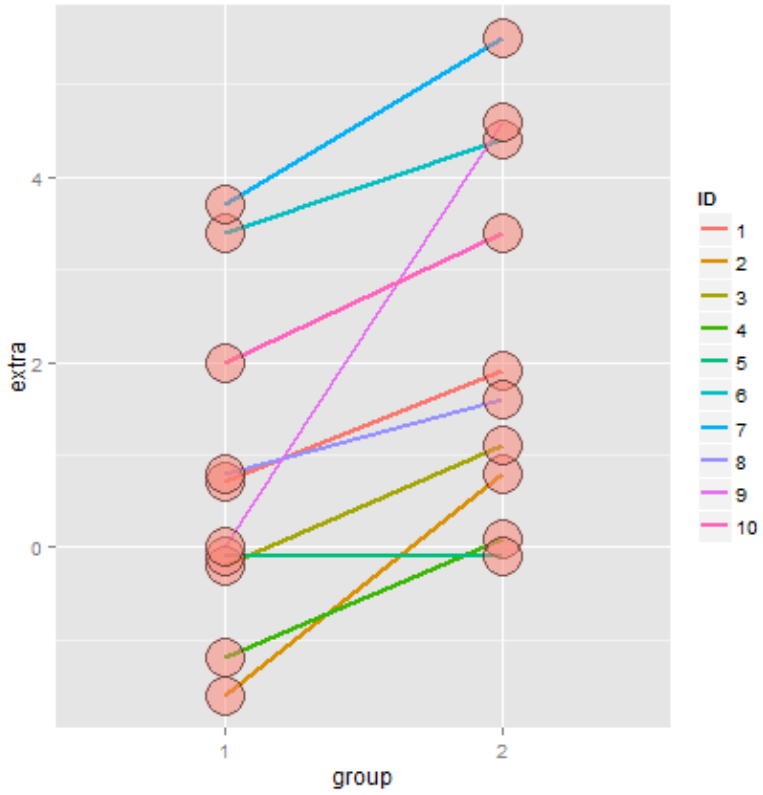
```
## [1] -9.521 20.361
```

Mistakenly treating the sleep data as grouped

```
n1 <- length(g1); n2 <- length(g2)
sp <- sqrt( ((n1 - 1) * sd(x1)^2 + (n2-1) * sd(x2)^2) / (n1 + n2-2))
md <- mean(g2) - mean(g1)
semd <- sp * sqrt(1 / n1 + 1/n2)
rbind(
md + c(-1, 1) * qt(.975, n1 + n2 - 2) * semd,
t.test(g2, g1, paired = FALSE, var.equal = TRUE)$conf,
t.test(g2, g1, paired = TRUE)$conf
)
```

```
##         [,1]  [,2]
## [1,] -0.2039 3.364
## [2,] -0.2039 3.364
## [3,]  0.7001 2.460
```

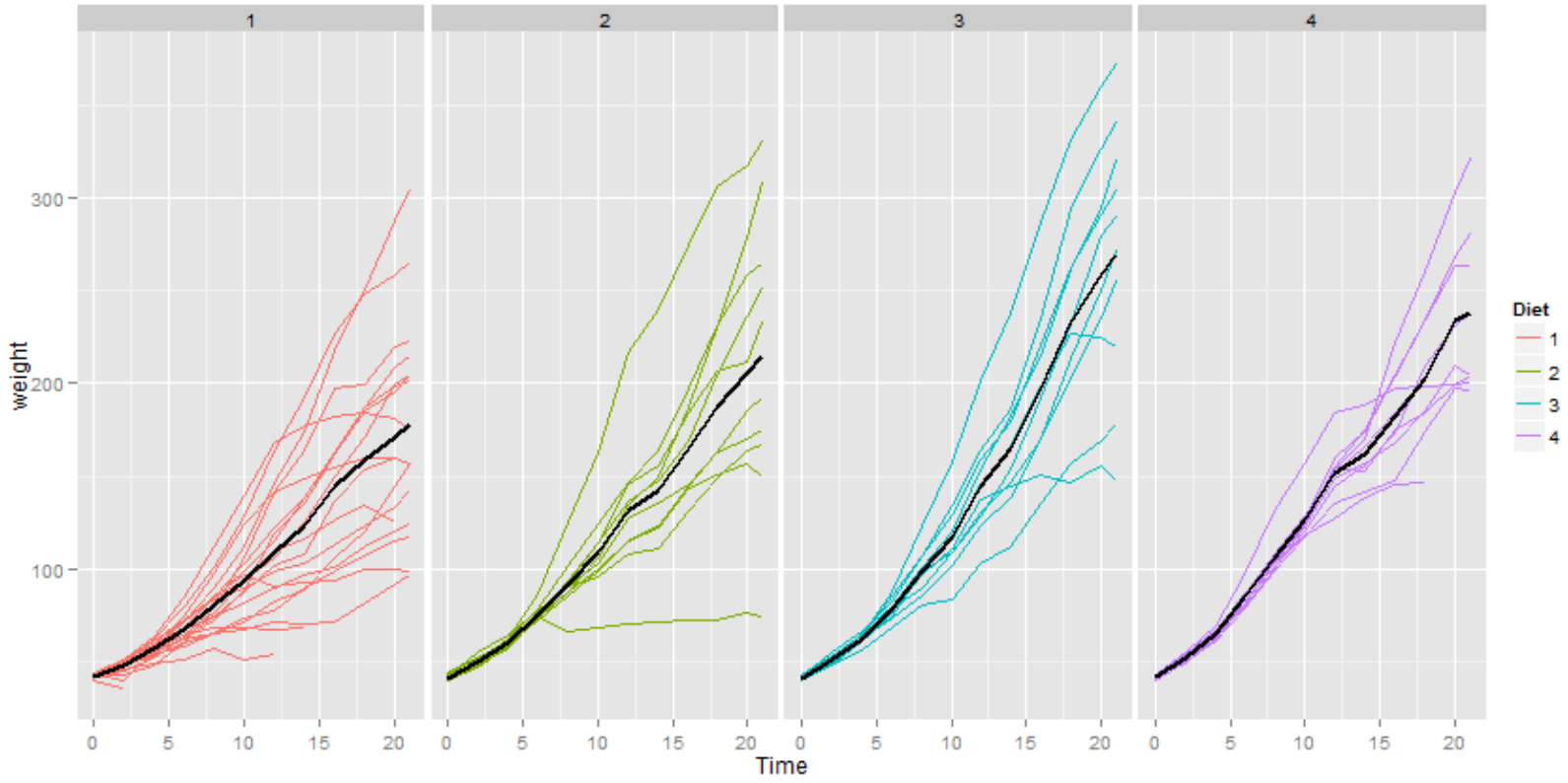
Grouped versus independent



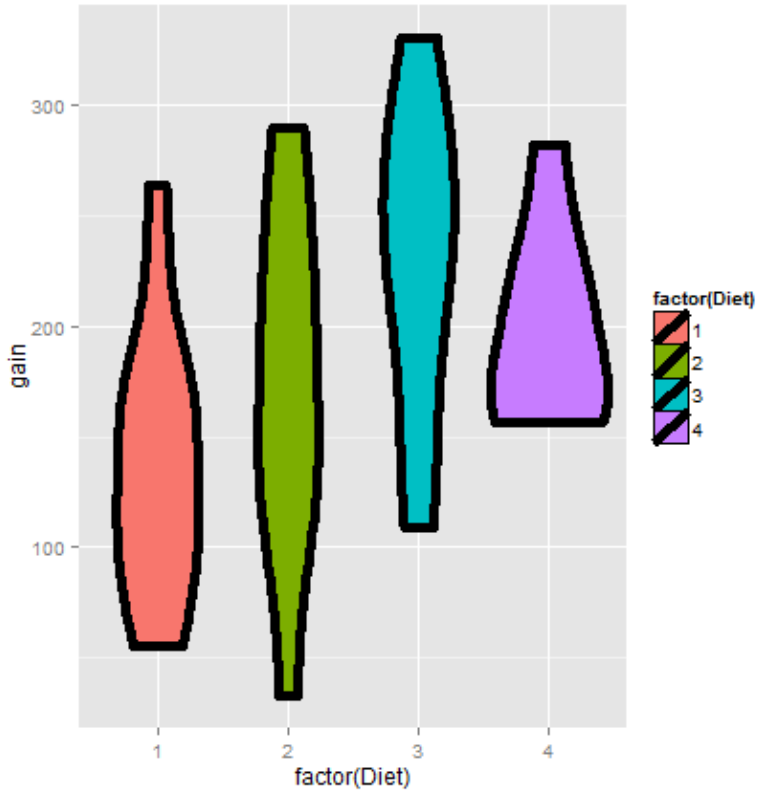
`ChickWeight` data in R

```
library(datasets); data(ChickWeight); library(reshape2)
##define weight gain or loss
wideCW <- dcast(ChickWeight, Diet + Chick ~ Time, value.var = "weight")
names(wideCW)[-(1 : 2)] <- paste("time", names(wideCW)[-(1 : 2)], sep = "")
library(dplyr)
wideCW <- mutate(wideCW,
  gain = time21 - time0
)
```

Plotting the raw data



Weight gain by diet



Let's do a t interval

```
wideCW14 <- subset(wideCW, Diet %in% c(1, 4))
rbind(
t.test(gain ~ Diet, paired = FALSE, var.equal = TRUE, data = wideCW14)$conf,
t.test(gain ~ Diet, paired = FALSE, var.equal = FALSE, data = wideCW14)$conf
)
```

```
##        [,1]   [,2]
## [1,] -108.1 -14.81
## [2,] -104.7 -18.30
```

Unequal variances

- Under unequal variances

$$\bar{Y} - \bar{X} \pm t_{df} \times \left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y} \right)^{1/2}$$

where t_{df} is calculated with degrees of freedom

$$df = \frac{\left(S_x^2/n_x + S_y^2/n_y \right)^2}{\left(\frac{S_x^2}{n_x} \right)^2 /(n_x - 1) + \left(\frac{S_y^2}{n_y} \right)^2 /(n_y - 1)}$$

will be approximately a 95% interval

- This works really well
  - So when in doubt, just assume unequal variances

Example

- Comparing SBP for 8 oral contraceptive users versus 21 controls
- $\bar{X}_{OC} = 132.86$ mmHg with $s_{OC} = 15.34$ mmHg
- $\bar{X}_{C} = 127.44$ mmHg with $s_{C} = 18.23$ mmHg
- $df = 15.04$, $t_{15.04,.975} = 2.13$
- Interval

$$132.86 - 127.44 \pm 2.13\left(\frac{15.34^2}{8} + \frac{18.23^2}{21}\right)^{1/2} = [-8.91, 19.75]$$

- In R, `t.test(..., var.equal = FALSE)`

Comparing other kinds of data

- For binomial data, there's lots of ways to compare two groups
  - Relative risk, risk difference, odds ratio.
  - Chi-squared tests, normal approximations, exact tests.
- For count data, there's also Chi-squared tests and exact tests.
- We'll leave the discussions for comparing groups of data for binary and count data until covering glms in the regression class.
- In addition, Mathematical Biostatistics Boot Camp 2 covers many special cases relevant to biostatistics.

Hypothesis testing

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

Hypothesis testing

- Hypothesis testing is concerned with making decisions using data
- A null hypothesis is specified that represents the status quo, usually labeled H_0
- The null hypothesis is assumed true and statistical evidence is required to reject it in favor of a research or alternative hypothesis

Example

- A respiratory disturbance index of more than 30 events / hour, say, is considered evidence of severe sleep disordered breathing (SDB).
- Suppose that in a sample of 100 overweight subjects with other risk factors for sleep disordered breathing at a sleep clinic, the mean RDI was 32 events / hour with a standard deviation of 10 events / hour.
- We might want to test the hypothesis that
  - $H_0 : \mu = 30$
  - $H_a : \mu > 30$
  - where μ is the population mean RDI.

Hypothesis testing

- The alternative hypotheses are typically of the form $<$, $>$ or $\neq$
- Note that there are four possible outcomes of our statistical decision process

TRUTH	DECIDE	RESULT
H_0	H_0	Correctly accept null
H_0	H_a	Type I error
H_a	H_a	Correctly reject null
H_a	H_0	Type II error

Discussion

- Consider a court of law; the null hypothesis is that the defendant is innocent
- We require a standard on the available evidence to reject the null hypothesis (convict)
- If we set a low standard, then we would increase the percentage of innocent people convicted (type I errors); however we would also increase the percentage of guilty people convicted (correctly rejecting the null)
- If we set a high standard, then we increase the the percentage of innocent people let free (correctly accepting the null) while we would also increase the percentage of guilty people let free (type II errors)

Example

- Consider our sleep example again
- A reasonable strategy would reject the null hypothesis if $\bar{X}$ was larger than some constant, say C
- Typically, C is chosen so that the probability of a Type I error, α, is $.05$ (or some other relevant constant)
- α = Type I error rate = Probability of rejecting the null hypothesis when, in fact, the null hypothesis is correct

Example continued

- Standard error of the mean $10/\sqrt{100} = 1$
- Under H_0 $\bar{X} \sim N(30, 1)$
- We want to chose C so that the $P(\bar{X} > C; H_0)$ is 5%
- The 95th percentile of a normal distribution is 1.645 standard deviations from the mean
- If $C = 30 + 1 \times 1.645 = 31.645$
  - Then the probability that a $N(30, 1)$ is larger than it is 5%
  - So the rule "Reject H_0 when $\bar{X} \geq 31.645$" has the property that the probability of rejection is 5% when H_0 is true (for the μ_0, σ and n given)

Discussion

- In general we don't convert C back to the original scale
- We would just reject because the Z-score; which is how many standard errors the sample mean is above the hypothesized mean

$$\frac{32 - 30}{10/\sqrt{100}} = 2$$

  is greater than 1.645
- Or, whenever $\sqrt{n}(\bar{X} - \mu_0)/s > Z_{1-\alpha}$

General rules

- The Z test for $H_0 : \mu = \mu_0$ versus
  - $H_1 : \mu < \mu_0$
  - $H_2 : \mu \neq \mu_0$
  - $H_3 : \mu > \mu_0$
- Test statistic $TS = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$
- Reject the null hypothesis when
  - $TS \leq Z_\alpha = -Z_{1-\alpha}$
  - $|TS| \geq Z_{1-\alpha/2}$
  - $TS \geq Z_{1-\alpha}$

Notes

- We have fixed α to be low, so if we reject H_0 (either our model is wrong) or there is a low probability that we have made an error
- We have not fixed the probability of a type II error, β; therefore we tend to say ``Fail to reject H_0'' rather than accepting H_0
- Statistical significance is no the same as scientific significance
- The region of TS values for which you reject H_0 is called the rejection region

More notes

- The Z test requires the assumptions of the CLT and for n to be large enough for it to apply
- If n is small, then a Gossett's T test is performed exactly in the same way, with the normal quantiles replaced by the appropriate Student's T quantiles and $n-1$ df
- The probability of rejecting the null hypothesis when it is false is called *power*
- Power is a used a lot to calculate sample sizes for experiments

Example reconsidered

- Consider our example again. Suppose that $n = 16$ (rather than 100)
- The statistic

$$\frac{\bar{X} - 30}{s/\sqrt{16}}$$

  follows a T distribution with 15 df under H_0
- Under H_0, the probability that it is larger that the 95th percentile of the T distribution is 5%
- The 95th percentile of the T distribution with 15 df is 1.7531 (obtained via `qt(.95, 15)`)
- So that our test statistic is now $\sqrt{16}(32 - 30)/10 = 0.8$
- We now fail to reject.

Two sided tests

- Suppose that we would reject the null hypothesis if in fact the mean was too large or too small
- That is, we want to test the alternative $H_a : \mu \neq 30$
- We will reject if the test statistic, 0.8, is either too large or too small
- Then we want the probability of rejecting under the null to be 5%, split equally as 2.5% in the upper tail and 2.5% in the lower tail
- Thus we reject if our test statistic is larger than `qt(.975, 15)` or smaller than `qt(.025, 15)`
  - This is the same as saying: reject if the absolute value of our statistic is larger than `qt(0.975, 15)` = 2.1314
  - So we fail to reject the two sided test as well
  - (If you fail to reject the one sided test, you know that you will fail to reject the two sided)

T test in R

```
library(UsingR); data(father.son)
t.test(father.son$sheight - father.son$fheight)
```

```
>
>   One Sample t-test
>
> data:  father.son$sheight - father.son$fheight
> t = 11.79, df = 1077, p-value < 2.2e-16
> alternative hypothesis: true mean is not equal to 0
> 95 percent confidence interval:
>  0.831 1.163
> sample estimates:
> mean of x
>     0.997
```

Connections with confidence intervals

- Consider testing $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$
- Take the set of all possible values for which you fail to reject H_0, this set is a $(1-\alpha)100\%$ confidence interval for μ
- The same works in reverse; if a $(1-\alpha)100\%$ interval contains μ_0, then we *fail to* reject H_0

Two group intervals

- First, now you know how to do two group T tests since we already covered indepedent group T intervals
- Rejection rules are the same
- Test $H_0 : \mu_1 = \mu_2$
- Let's just go through an example

`chickWeight` data

Recall that we reformatted this data

```
library(datasets); data(ChickWeight); library(reshape2)
##define weight gain or loss
wideCW <- dcast(ChickWeight, Diet + Chick ~ Time, value.var = "weight")
names(wideCW)[-(1 : 2)] <- paste("time", names(wideCW)[-(1 : 2)], sep = "")
library(dplyr)
wideCW <- mutate(wideCW,
  gain = time21 - time0
)
```

Unequal variance T test comparing diets 1 and 4

```
wideCW14 <- subset(wideCW, Diet %in% c(1, 4))
t.test(gain ~ Diet, paired = FALSE,
       var.equal = TRUE, data = wideCW14)
```

```
>
>   Two Sample t-test
>
>  data:  gain by Diet
>  t = -2.725, df = 23, p-value = 0.01207
>  alternative hypothesis: true difference in means is not equal to 0
>  95 percent confidence interval:
>   -108.15  -14.81
>  sample estimates:
>  mean in group 1 mean in group 4
>            136.2           197.7
```

Exact binomial test

- Recall this problem, *Suppose a friend has* 8 *children,* 7 *of which are girls and none are twins*
- Perform the relevant hypothesis test. $H_0 : p = 0.5$ $H_a : p > 0.5$
  - What is the relevant rejection region so that the probability of rejecting is (less than) 5%?

REJECTION REGION	TYPE I ERROR RATE
[0 : 8]	1
[1 : 8]	0.9961
[2 : 8]	0.9648
[3 : 8]	0.8555
[4 : 8]	0.6367
[5 : 8]	0.3633
[6 : 8]	0.1445
[7 : 8]	0.0352
[8 : 8]	0.0039

Notes

- It's impossible to get an exact 5% level test for this case due to the discreteness of the binomial.
  - The closest is the rejection region [7 : 8]
  - Any alpha level lower than 0.0039 is not attainable.
- For larger sample sizes, we could do a normal approximation, but you already knew this.
- Two sided test isn't obvious.
  - Given a way to do two sided tests, we could take the set of values of p_0 for which we fail to reject to get an exact binomial confidence interval (called the Clopper/Pearson interval, BTW)
- For these problems, people always create a P-value (next lecture) rather than computing the rejection region.

P-values

Statistical inference

Brian Caffo, Jeffrey Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

P-values

- Most common measure of "statistical significance"
- Their ubiquity, along with concern over their interpretation and use makes them controversial among statisticians
  - http://warnercnr.colostate.edu/~anderson/thompson1.html
  - Also see *Statistical Evidence: A Likelihood Paradigm* by Richard Royall
  - *Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy* by Steve Goodman
  - The hilariously titled: *The Earth is Round (p < .05)* by Cohen.
- Some positive comments
  - simply statistics
  - normal deviate
  - Error statistics

What is a P-value?

Idea: Suppose nothing is going on - how unusual is it to see the estimate we got?

Approach:

1. Define the hypothetical distribution of a data summary (statistic) when "nothing is going on" (*null hypothesis*)
2. Calculate the summary/statistic with the data we have (*test statistic*)
3. Compare what we calculated to our hypothetical distribution and see if the value is "extreme" (*p-value*)

P-values

- The P-value is the probability under the null hypothesis of obtaining evidence as extreme or more extreme than would be observed by chance alone
- If the P-value is small, then either H_0 is true and we have observed a rare event or H_0 is false
- In our example the T statistic was 0.8.
    - What's the probability of getting a T statistic as large as 0.8?

```
pt(0.8, 15, lower.tail = FALSE)
```

```
## [1] 0.2181
```

- Therefore, the probability of seeing evidence as extreme or more extreme than that actually obtained under H_0 is 0.2181

The attained significance level

- Our test statistic was 2 for $H_0 : \mu_0 = 30$ versus $H_a : \mu > 30$.
- Notice that we rejected the one sided test when $\alpha = 0.05$, would we reject if $\alpha = 0.01$, how about 0.001?
- The smallest value for alpha that you still reject the null hypothesis is called the *attained significance level*
- This is equivalent, but philosophically a little different from, the *P-value*

Notes

- By reporting a P-value the reader can perform the hypothesis test at whatever α level he or she choses
- If the P-value is less than α you reject the null hypothesis
- For two sided hypothesis test, double the smaller of the two one sided hypothesis test Pvalues

Revisiting an earlier example

- Suppose a friend has 8 children, 7 of which are girls and none are twins
- If each gender has an independent 50% probability for each birth, what's the probability of getting 7 or more girls out of 8 births?

```
choose(8, 7) * 0.5^8 + choose(8, 8) * 0.5^8
```

```
## [1] 0.03516
```

```
pbinom(6, size = 8, prob = 0.5, lower.tail = FALSE)
```

```
## [1] 0.03516
```

Poisson example

- Suppose that a hospital has an infection rate of 10 infections per 100 person/days at risk (rate of 0.1) during the last monitoring period.
- Assume that an infection rate of 0.05 is an important benchmark.
- Given the model, could the observed rate being larger than 0.05 be attributed to chance?
- Under $H_0 : \lambda = 0.05$ so that $\lambda_0 100 = 5$
- Consider $H_a : \lambda > 0.05$.

```
ppois(9, 5, lower.tail = FALSE)
```

```
## [1] 0.03183
```

Power

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

Power

- Power is the probability of rejecting the null hypothesis when it is false
- Ergo, power (as its name would suggest) is a good thing; you want more power
- A type II error (a bad thing, as its name would suggest) is failing to reject the null hypothesis when it's false; the probability of a type II error is usually called β
- Note Power $= 1 - \beta$

Notes

- Consider our previous example involving RDI
- $H_0 : \mu = 30$ versus $H_a : \mu > 30$
- Then power is

$$P\left(\frac{\bar{X} - 30}{s/\sqrt{n}} > t_{1-\alpha,n-1} \; ; \; \mu = \mu_a \right)$$

- Note that this is a function that depends on the specific value of μ_a!
- Notice as μ_a approaches 30 the power approaches α

Calculating power for Gaussian data

- We reject if $\frac{\bar{X}-30}{\sigma/\sqrt{n}} > z_{1-\alpha}$
  - Equivalently if $\bar{X} > 30 + Z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$
- Under H_0 : $\bar{X} \sim N(\mu_0, \sigma^2/n)$
- Under H_a : $\bar{X} \sim N(\mu_a, \sigma^2/n)$
- So we want

```
alpha = 0.05
z = qnorm(1 - alpha)
pnorm(mu0 + z * sigma/sqrt(n), mean = mua, sd = sigma/sqrt(n), lower.tail = FALSE)
```

Example continued

- $\mu_a = 32, \mu_0 = 30, n = 16, \sigma = 4$

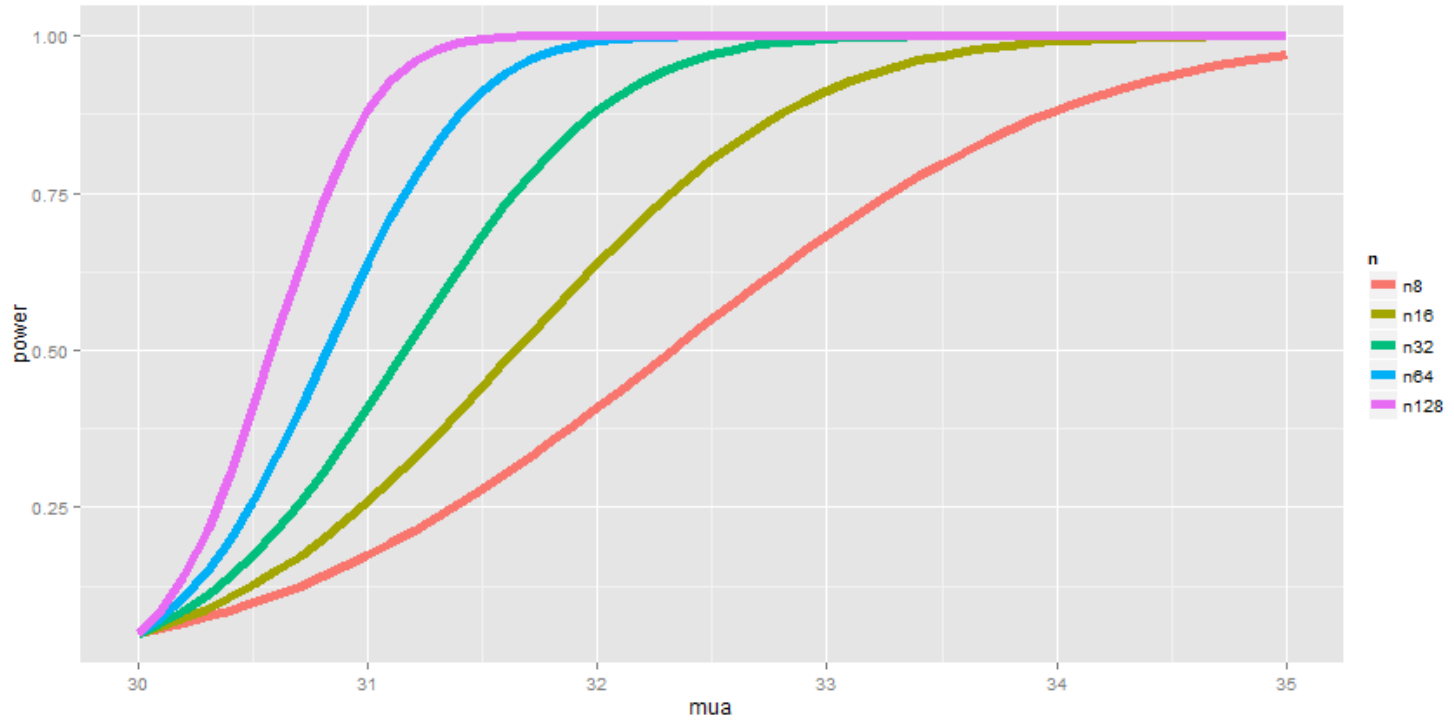
```
mu0 = 30
mua = 32
sigma = 4
n = 16
z = qnorm(1 - alpha)
pnorm(mu0 + z * sigma/sqrt(n), mean = mu0, sd = sigma/sqrt(n), lower.tail = FALSE)
```

```
## [1] 0.05
```

```
pnorm(mu0 + z * sigma/sqrt(n), mean = mua, sd = sigma/sqrt(n), lower.tail = FALSE)
```

```
## [1] 0.6388
```

Plotting the power curve



Graphical Depiction of Power

```
library(manipulate)
mu0 = 30
myplot <- function(sigma, mua, n, alpha) {
    g = ggplot(data.frame(mu = c(27, 36)), aes(x = mu))
    g = g + stat_function(fun = dnorm, geom = "line", args = list(mean = mu0,
        sd = sigma/sqrt(n)), size = 2, col = "red")
    g = g + stat_function(fun = dnorm, geom = "line", args = list(mean = mua,
        sd = sigma/sqrt(n)), size = 2, col = "blue")
    xitc = mu0 + qnorm(1 - alpha) * sigma/sqrt(n)
    g = g + geom_vline(xintercept = xitc, size = 3)
    g
}
manipulate(myplot(sigma, mua, n, alpha), sigma = slider(1, 10, step = 1, initial = 4),
    mua = slider(30, 35, step = 1, initial = 32), n = slider(1, 50, step = 1,
        initial = 16), alpha = slider(0.01, 0.1, step = 0.01, initial = 0.05))
```

Question

- When testing $H_a : \mu > \mu_0$, notice if power is $1 - \beta$, then

$$1 - \beta = P\left(\bar{X} > \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\,; \mu = \mu_a\right)$$

- where $\bar{X} \sim N(\mu_a, \sigma^2/n)$
- Unknowns: μ_a, σ, n, β
- Knowns: μ_0, α
- Specify any 3 of the unknowns and you can solve for the remainder

Notes

- The calculation for $H_a : \mu < \mu_0$ is similar
- For $H_a : \mu \neq \mu_0$ calculate the one sided power using $\alpha/2$ (this is only approximately right, it excludes the probability of getting a large TS in the opposite direction of the truth)
- Power goes up as α gets larger
- Power of a one sided test is greater than the power of the associated two sided test
- Power goes up as μ_1 gets further away from μ_0
- Power goes up as n goes up
- Power doesn't need μ_a, σ and n, instead only $\frac{\sqrt{n}(\mu_a - \mu_0)}{\sigma}$
  - The quantity $\frac{\mu_a - \mu_0}{\sigma}$ is called the effect size, the difference in the means in standard deviation units.
  - Being unit free, it has some hope of interpretability across settings

T-test power

- Consider calculating power for a Gossett's T test for our example
- The power is

$$P\left(\frac{\bar{X} - \mu_0}{S/\sqrt{n}} > t_{1-\alpha,n-1} \; ; \; \mu = \mu_a\right)$$

- Calcuting this requires the non-central t distribution.
- `power.t.test` does this very well
  - Omit one of the arguments and it solves for it

Example

```
power.t.test(n = 16, delta = 2/4, sd = 1, type = "one.sample", alt = "one.sided")$power
```

```
## [1] 0.604
```

```
power.t.test(n = 16, delta = 2, sd = 4, type = "one.sample", alt = "one.sided")$power
```

```
## [1] 0.604
```

```
power.t.test(n = 16, delta = 100, sd = 200, type = "one.sample", alt = "one.sided")$power
```

```
## [1] 0.604
```

Example

```
power.t.test(power = 0.8, delta = 2/4, sd = 1, type = "one.sample", alt = "one.sided")$n
```

```
## [1] 26.14
```

```
power.t.test(power = 0.8, delta = 2, sd = 4, type = "one.sample", alt = "one.sided")$n
```

```
## [1] 26.14
```

```
power.t.test(power = 0.8, delta = 100, sd = 200, type = "one.sample", alt = "one.sided")$n
```

```
## [1] 26.14
```

Multiple testing

Statistical Inference

Brian Caffo, Jeffrey Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

Key ideas

- Hypothesis testing/significance analysis is commonly overused
- Correcting for multiple testing avoids false positives or discoveries
- Two key components
  - Error measure
  - Correction

Three eras of statistics

The age of Quetelet and his successors, in which huge census-level data sets were brought to bear on simple but important questions: Are there more male than female births? Is the rate of insanity rising?

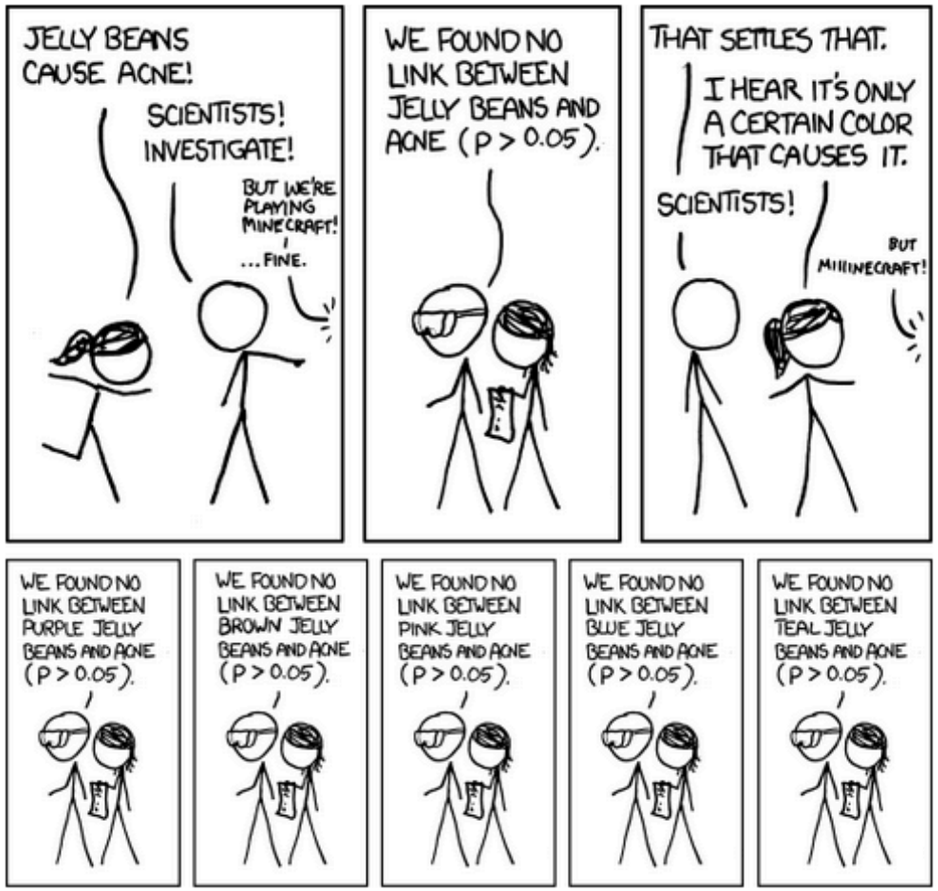
The classical period of Pearson, Fisher, Neyman, Hotelling, and their successors, intellectual giants who **developed a theory of optimal inference capable of wringing every drop of information out of a scientific experiment**. The questions dealt with still tended to be simple Is treatment A better than treatment B?

The era of scientific mass production, in which new technologies typified by the microarray allow a single team of scientists to produce data sets of a size Quetelet would envy. But now the flood of data is accompanied by a deluge of questions, perhaps thousands of estimates or hypothesis tests that the statistician is charged with answering together; not at all what the classical masters had in mind. Which variables matter among the thousands measured? How do you relate unrelated information?

http://www-stat.stanford.edu/~ckirby/brad/papers/2010LSIexcerpt.pdf

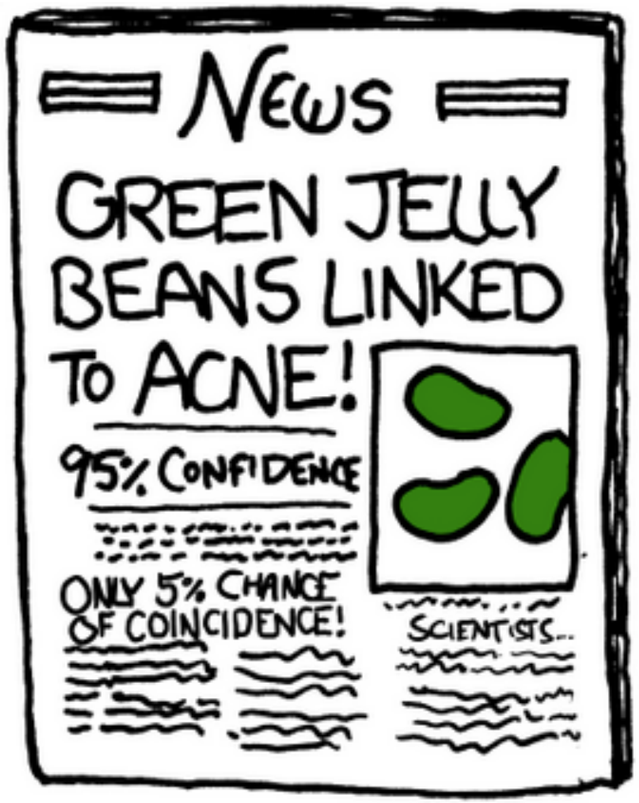
Reasons for multiple testing

Why correct for multiple tests?



http://xkcd.com/882/

Why correct for multiple tests?



http://xkcd.com/882/

Types of errors

Suppose you are testing a hypothesis that a parameter β equals zero versus the alternative that it does not equal zero. These are the possible outcomes.

	$\beta = 0$	$\beta \neq 0$	HYPOTHESES
Claim $\beta = 0$	U	T	$m - R$
Claim $\beta \neq 0$	V	S	R
Claims	m_0	$m - m_0$	m

Type I error or false positive (V) Say that the parameter does not equal zero when it does

Type II error or false negative (T) Say that the parameter equals zero when it doesn't

Error rates

False positive rate - The rate at which false results ($\beta = 0$) are called significant: $E\left[\frac{V}{m_0}\right]$ *

Family wise error rate (FWER) - The probability of at least one false positive $\Pr(V \geq 1)$

False discovery rate (FDR) - The rate at which claims of significance are false $E\left[\frac{V}{R}\right]$

- The false positive rate is closely related to the type I error rate http://en.wikipedia.org/wiki/False_positive_rate

Controlling the false positive rate

If P-values are correctly calculated calling all $P < \alpha$ significant will control the false positive rate at level α on average.

Problem: Suppose that you perform 10,000 tests and $\beta = 0$ for all of them.

Suppose that you call all $P < 0.05$ significant.

The expected number of false positives is: $10,000 \times 0.05 = 500$ false positives.

How do we avoid so many false positives?

Controlling family-wise error rate (FWER)

The Bonferroni correction is the oldest multiple testing correction.

Basic idea:

- Suppose you do m tests
- You want to control FWER at level α so $Pr(V \geq 1) < \alpha$
- Calculate P-values normally
- Set $\alpha_{fwer} = \alpha/m$
- Call all P-values less than α_{fwer} significant

Pros: Easy to calculate, conservative **Cons**: May be very conservative

Controlling false discovery rate (FDR)

This is the most popular correction when performing *lots* of tests say in genomics, imaging, astronomy, or other signal-processing disciplines.

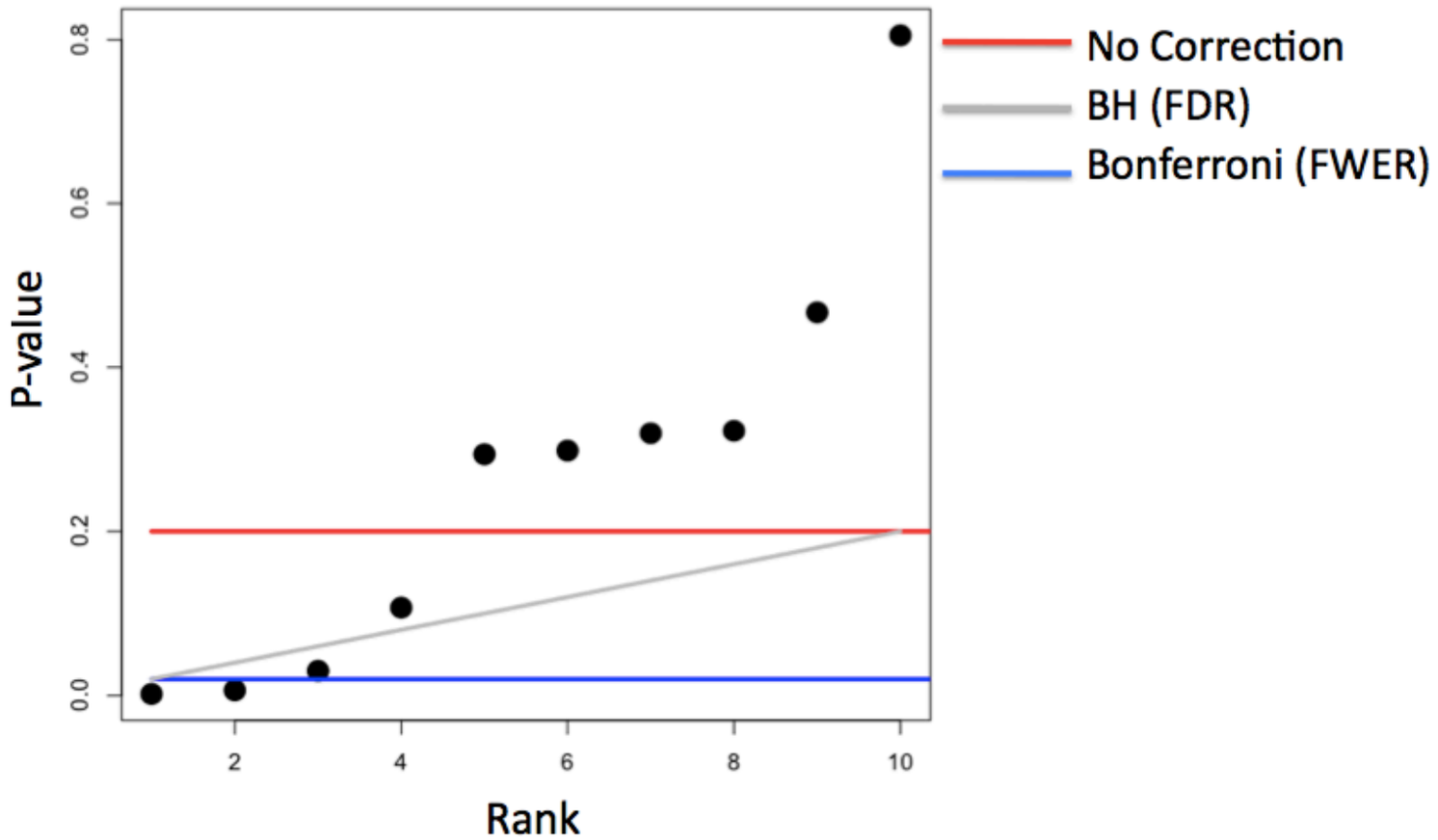
Basic idea:

- Suppose you do m tests
- You want to control FDR at level α so $E\left[\frac{V}{R}\right]$
- Calculate P-values normally
- Order the P-values from smallest to largest $P_{(1)}, \ldots, P_{(m)}$
- Call any $P_{(i)} \leq \alpha \times \frac{i}{m}$ significant

Pros: Still pretty easy to calculate, less conservative (maybe much less)

Cons: Allows for more false positives, may behave strangely under dependence

Example with 10 P-values



Controlling all error rates at $\alpha = 0.20$

Adjusted P-values

- One approach is to adjust the threshold α
- A different approach is to calculate "adjusted p-values"
- They *are not p-values* anymore
- But they can be used directly without adjusting α

Example:

- Suppose P-values are $P_1, \ldots, P_m$
- You could adjust them by taking $P_i^{fwer} = \max m \times P_i, 1$ for each P-value.
- Then if you call all $P_i^{fwer} < \alpha$ significant you will control the FWER.

Case study I: no true positives

```
set.seed(1010093)
pValues <- rep(NA, 1000)
for (i in 1:1000) {
    y <- rnorm(20)
    x <- rnorm(20)
    pValues[i] <- summary(lm(y ~ x))$coeff[2, 4]
}

# Controls false positive rate
sum(pValues < 0.05)
```

```
## [1] 51
```

Case study I: no true positives

```
# Controls FWER
sum(p.adjust(pValues, method = "bonferroni") < 0.05)
```

```
## [1] 0
```

```
# Controls FDR
sum(p.adjust(pValues, method = "BH") < 0.05)
```

```
## [1] 0
```

Case study II: 50% true positives

```
set.seed(1010093)
pValues <- rep(NA, 1000)
for (i in 1:1000) {
    x <- rnorm(20)
    # First 500 beta=0, last 500 beta=2
    if (i <= 500) {
        y <- rnorm(20)
    } else {
        y <- rnorm(20, mean = 2 * x)
    }
    pValues[i] <- summary(lm(y ~ x))$coeff[2, 4]
}
trueStatus <- rep(c("zero", "not zero"), each = 500)
table(pValues < 0.05, trueStatus)
```

```
##        trueStatus
##         not zero zero
##   FALSE        0  476
```

Case study II: 50% true positives

```
# Controls FWER
table(p.adjust(pValues, method = "bonferroni") < 0.05, trueStatus)
```

```
##        trueStatus
##         not zero zero
##   FALSE       23  500
##   TRUE       477    0
```

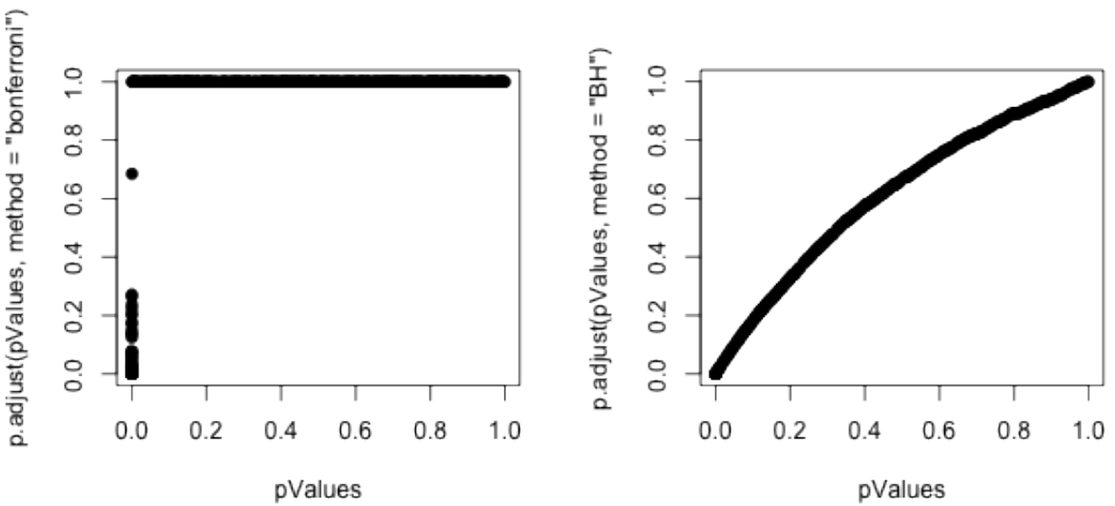
```
# Controls FDR
table(p.adjust(pValues, method = "BH") < 0.05, trueStatus)
```

```
##        trueStatus
##         not zero zero
##   FALSE        0  487
##   TRUE       500   13
```

Case study II: 50% true positives

P-values versus adjusted P-values

```
par(mfrow = c(1, 2))
plot(pValues, p.adjust(pValues, method = "bonferroni"), pch = 19)
plot(pValues, p.adjust(pValues, method = "BH"), pch = 19)
```



Notes and resources

Notes:

- Multiple testing is an entire subfield
- A basic Bonferroni/BH correction is usually enough
- If there is strong dependence between tests there may be problems
    - Consider method="BY"

Further resources:

- Multiple testing procedures with applications to genomics
- Statistical significance for genome-wide studies
- Introduction to multiple testing

Resampled inference

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

The jackknife

- The jackknife is a tool for estimating standard errors and the bias of estimators
- As its name suggests, the jackknife is a small, handy tool; in contrast to the bootstrap, which is then the moral equivalent of a giant workshop full of tools
- Both the jackknife and the bootstrap involve *resampling* data; that is, repeatedly creating new data sets from the original data

The jackknife

- The jackknife deletes each observation and calculates an estimate based on the remaining $n-1$ of them
- It uses this collection of estimates to do things like estimate the bias and the standard error
- Note that estimating the bias and having a standard error are not needed for things like sample means, which we know are unbiased estimates of population means and what their standard errors are

The jackknife

- We'll consider the jackknife for univariate data
- Let $X_1, \ldots, X_n$ be a collection of data used to estimate a parameter θ
- Let $\hat{\theta}$ be the estimate based on the full data set
- Let $\hat{\theta}_i$ be the estimate of θ obtained by *deleting observation* i
- Let $\bar{\theta} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_i$

Continued

- Then, the jackknife estimate of the bias is

$$(n-1)\left(\bar{\theta}-\hat{\theta}\right)$$

(how far the average delete-one estimate is from the actual estimate)

- The jackknife estimate of the standard error is

$$\left[\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_i-\bar{\theta})^2\right]^{1/2}$$

(the deviance of the delete-one estimates from the average delete-one estimate)

Example

We want to estimate the bias and standard error of the median

```
library(UsingR)
data(father.son)
x <- father.son$sheight
n <- length(x)
theta <- median(x)
jk <- sapply(1:n, function(i) median(x[-i]))
thetaBar <- mean(jk)
biasEst <- (n - 1) * (thetaBar - theta)
seEst <- sqrt((n - 1) * mean((jk - thetaBar)^2))
```

Example test

```
c(biasEst, seEst)
```

```
## [1] 0.0000 0.1014
```

```
library(bootstrap)
temp <- jackknife(x, median)
c(temp$jack.bias, temp$jack.se)
```

```
## [1] 0.0000 0.1014
```

Example

- Both methods (of course) yield an estimated bias of 0 and a se of 0.1014
- Odd little fact: the jackknife estimate of the bias for the median is always 0 when the number of observations is even
- It has been shown that the jackknife is a linear approximation to the bootstrap
- Generally do not use the jackknife for sample quantiles like the median; as it has been shown to have some poor properties

Pseudo observations

- Another interesting way to think about the jackknife uses pseudo observations
- Let

$$\text{Pseudo Obs} = n\hat{\theta} - (n-1)\hat{\theta}_i$$

- Think of these as ``whatever observation i contributes to the estimate of θ''
- Note when $\hat{\theta}$ is the sample mean, the pseudo observations are the data themselves
- Then the sample standard error of these observations is the previous jackknife estimated standard error.
- The mean of these observations is a bias-corrected estimate of θ

The bootstrap

- The bootstrap is a tremendously useful tool for constructing confidence intervals and calculating standard errors for difficult statistics
- For example, how would one derive a confidence interval for the median?
- The bootstrap procedure follows from the so called bootstrap principle

The bootstrap principle

- Suppose that I have a statistic that estimates some population parameter, but I don't know its sampling distribution
- The bootstrap principle suggests using the distribution defined by the data to approximate its sampling distribution

The bootstrap in practice

- In practice, the bootstrap principle is always carried out using simulation
- We will cover only a few aspects of bootstrap resampling
- The general procedure follows by first simulating complete data sets from the observed data with replacement
    - This is approximately drawing from the sampling distribution of that statistic, at least as far as the data is able to approximate the true population distribution
- Calculate the statistic for each simulated data set
- Use the simulated statistics to either define a confidence interval or take the standard deviation to calculate a standard error

Nonparametric bootstrap algorithm example

- Bootstrap procedure for calculating confidence interval for the median from a data set of n observations

  i. Sample n observations **with replacement** from the observed data resulting in one simulated complete data set

  ii. Take the median of the simulated data set

  iii. Repeat these two steps B times, resulting in B simulated medians

  iv. These medians are approximately drawn from the sampling distribution of the median of n observations; therefore we can

  - Draw a histogram of them
  - Calculate their standard deviation to estimate the standard error of the median
  - Take the 2.5^{th} and 97.5^{th} percentiles as a confidence interval for the median

Example code

```
B <- 1000
resamples <- matrix(sample(x, n * B, replace = TRUE), B, n)
medians <- apply(resamples, 1, median)
sd(medians)
```

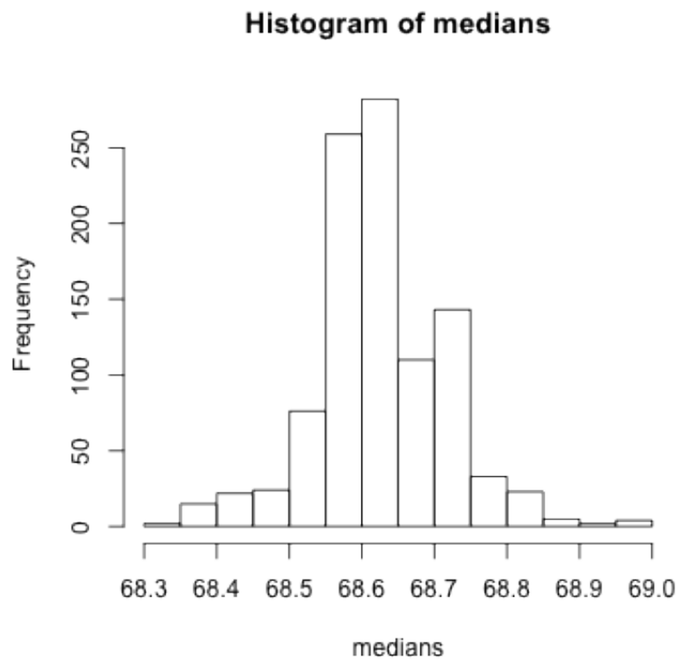
```
## [1] 0.08834
```

```
quantile(medians, c(0.025, 0.975))
```

```
##  2.5% 97.5%
## 68.41 68.82
```

Histogram of bootstrap resamples

```
hist(medians)
```



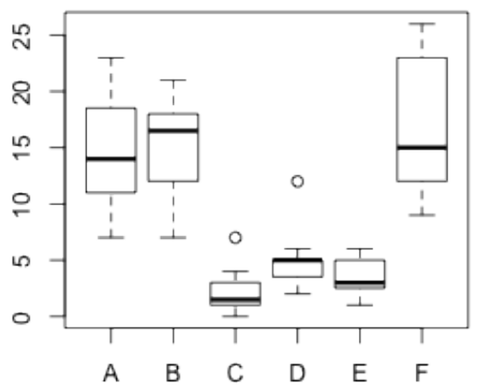
Notes on the bootstrap

- The bootstrap is non-parametric
- Better percentile bootstrap confidence intervals correct for bias
- There are lots of variations on bootstrap procedures; the book "An Introduction to the Bootstrap"" by Efron and Tibshirani is a great place to start for both bootstrap and jackknife information

Group comparisons

- Consider comparing two independent groups.
- Example, comparing sprays B and C

```
data(InsectSprays)
boxplot(count ~ spray, data = InsectSprays)
```



Permutation tests

- Consider the null hypothesis that the distribution of the observations from each group is the same
- Then, the group labels are irrelevant
- We then discard the group levels and permute the combined data
- Split the permuted data into two groups with n_A and n_B observations (say by always treating the first n_A observations as the first group)
- Evaluate the probability of getting a statistic as large or large than the one observed
- An example statistic would be the difference in the averages between the two groups; one could also use a t-statistic

Variations on permutation testing

DATA TYPE	STATISTIC	TEST NAME
Ranks	rank sum	rank sum test
Binary	hypergeometric prob	Fisher's exact test
Raw data		ordinary permutation test

- Also, so-called *randomization tests* are exactly permutation tests, with a different motivation.
- For matched data, one can randomize the signs
  - For ranks, this results in the signed rank test
- Permutation strategies work for regression as well
  - Permuting a regressor of interest
- Permutation tests work very well in multivariate settings

Permutation test for pesticide data

```
subdata <- InsectSprays[InsectSprays$spray %in% c("B", "C"), ]
y <- subdata$count
group <- as.character(subdata$spray)
testStat <- function(w, g) mean(w[g == "B"]) - mean(w[g == "C"])
observedStat <- testStat(y, group)
permutations <- sapply(1:10000, function(i) testStat(y, sample(group)))
observedStat
```

```
## [1] 13.25
```

```
mean(permutations > observedStat)
```

```
## [1] 0
```

Histogram of permutations