Statistical Inference

- Generating conclusions about a population from a noisy sample
- Will pick the most frequently taught paradigm in stats classes
    - Focus on Frequency Style Inference

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

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

- 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.

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.

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

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

Probability Mass Function

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).

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

Probability Density Function

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.

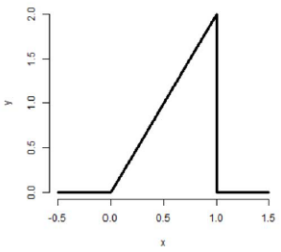
Example

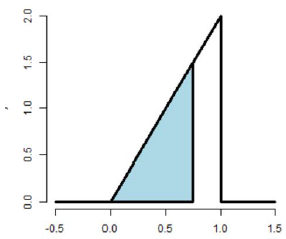
Suppose that the proportion of help calls that get addressed in a random day by a help line is given by

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

Is this a mathematically valid density?

```
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")
```






```
1.5 * 0.75/2
```

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

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

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

Cumulative 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.

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

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 #2

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.

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

- 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

- 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

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

- 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, $P(- \mid D^c)$
- 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, $P(D)$
- 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 $P(D) = .001$

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

- 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)

- 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

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

- 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

- 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?
- $A = \{\text{Head on flip } 1\}$ ~ $P(A) = .5$
- $B = \{\text{Head on flip } 2\}$ ~ $P(B) = .5$
- $A \cap B = \{\text{Head on flips } 1 \text{ and } 2\}$
- $P(A \cap B) = P(A)P(B) = .5 \times .5 = .25$

- 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

- 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

- 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

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

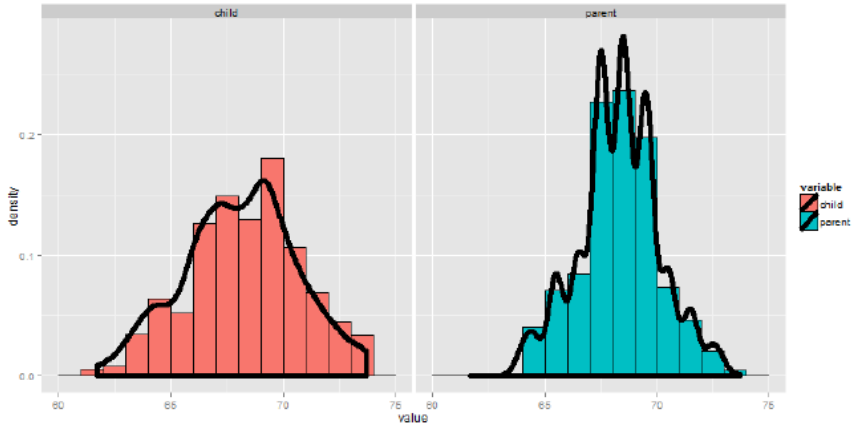
- The sample mean estimates this population mean
- The center of mass of the data is the empirical mean

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

where $p(x_i) = 1/n$

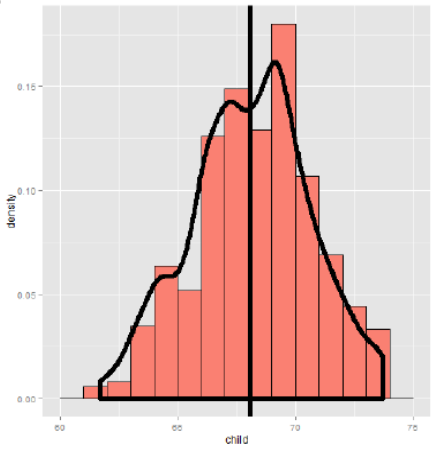
Center of Mass Example

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))
```

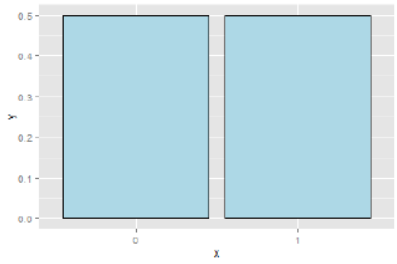


Example of a Population Mean

- 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?

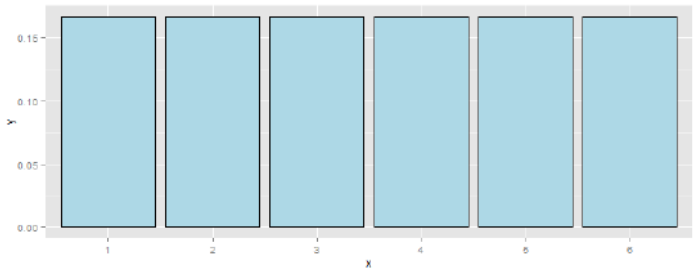
$$E[X] = 0 * (1 - p) + 1 * p = p$$

What about a Die

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

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

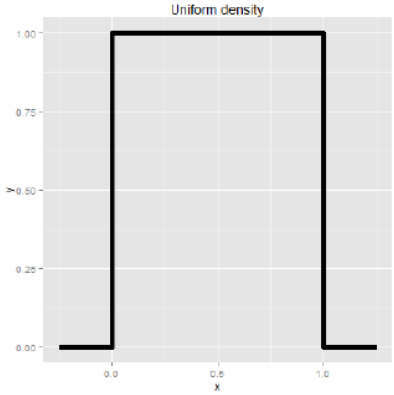
- Again, the geometric argument makes this answer obvious without calculation.



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

- 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?

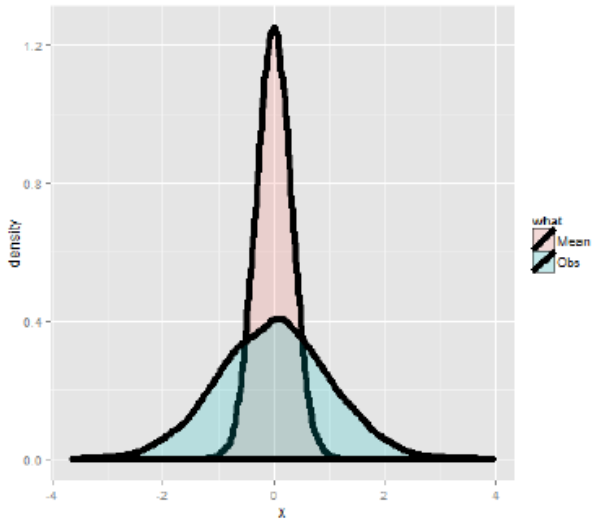


Facts about Expected Values

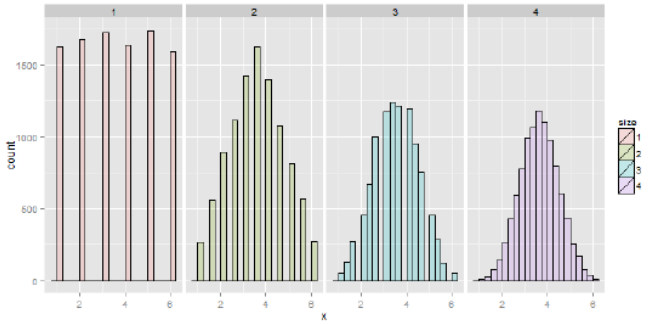
- Recall that expected values are properties of distributions
- Note the average of random variables is itself a random variable and its associated distribution has an expected value
- The center of this distribution is the same as that of the original distribution
- Therefore, the expected value of 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

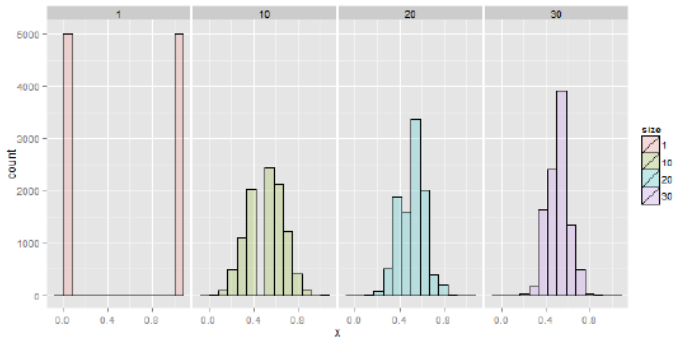
Simulating normals with mean 0 and variance 1 versus averages of 10 normals from the same population



Average of x Die Rolls



Average of x Coin Flips



Summary

- 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

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

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

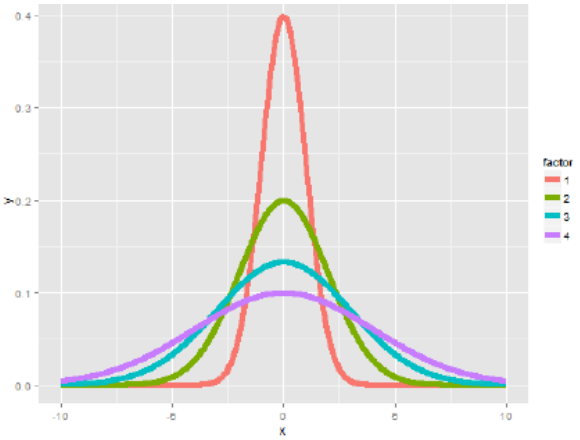
- The expected (squared) distance from the mean
- Densities with a higher variance are more spread out than densities with a lower variance
- The square root of the variance is called the **standard deviation**
- The standard deviation has the same units as X

Example

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



Sample Variance

- The sample variance is

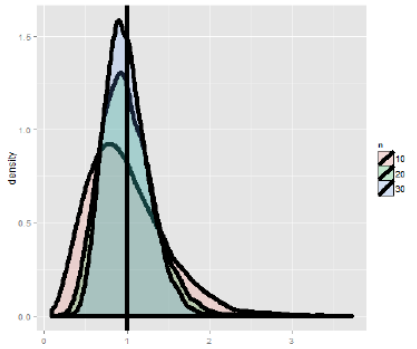
$$S^2 = \frac{\sum_{i=1}(X_i - \bar{X})^2}{n-1}$$

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

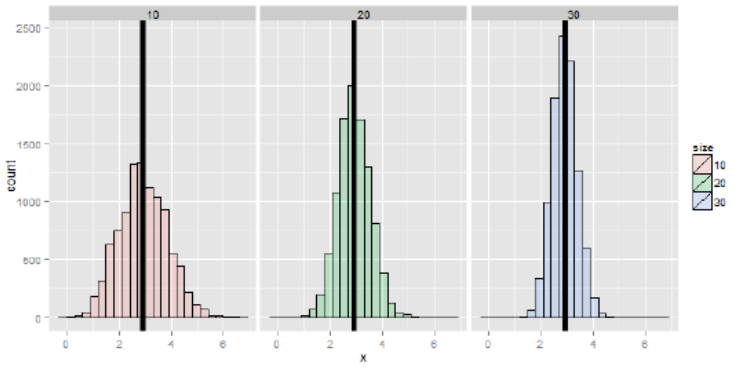
- It is also a random variable
  - It has an associate population distribution
  - Its expected value is the population variance
  - Its distribution gets more concentrated around the population variance with mroe data
- Its square root is the sample standard deviation

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

Summary

- 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
```

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
```

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
```

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))
```

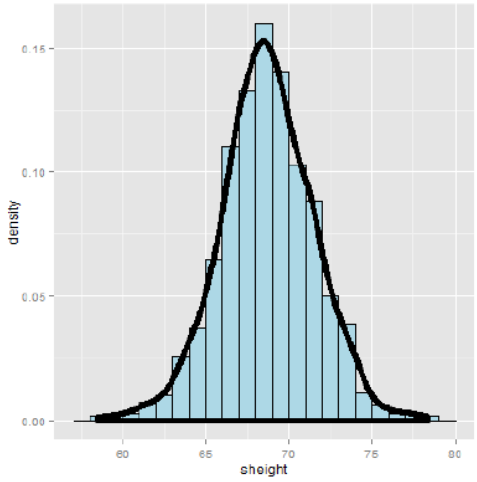
```
## [1] 0.1587
```

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

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

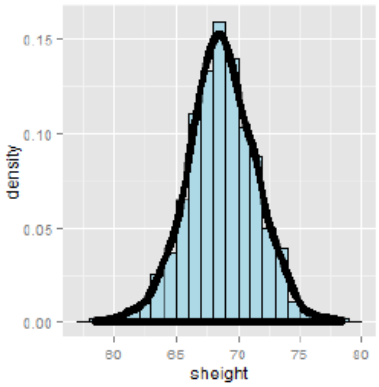
Data Example

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




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

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



Summarizing 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

Common Distributions

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)
```

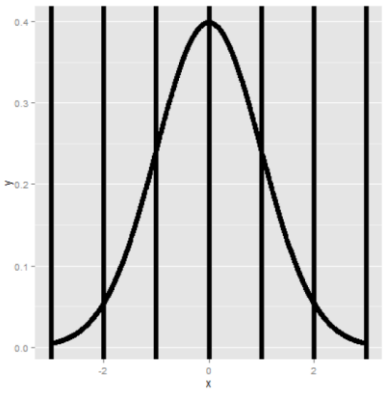
```
## [1] 0.03516
```

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



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

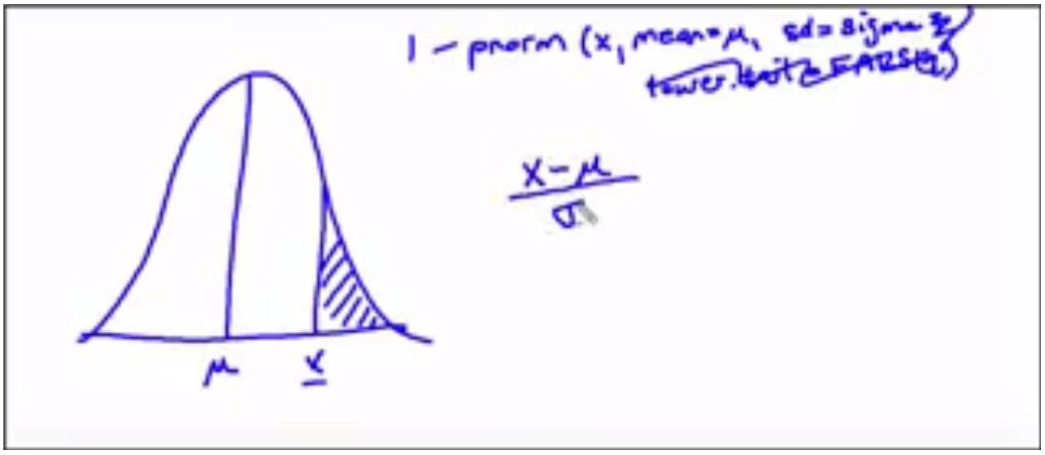
$$X = \mu + \sigma Z \sim \mathrm{N}(\mu, \sigma^2)$$

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

- 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)
- 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?

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
```

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
```

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 ∞
- 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

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

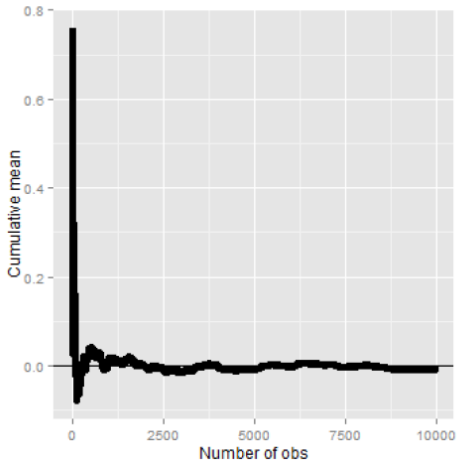
- Asymptotics is the term for the behavior of statistics as the sample size (or some other relevant quantity) limits to infinity (or some other relevant number)
- (Asymptopia is my name for the land of asymptotics, where everything works out well and there's no messes. The land of infinite data is nice 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

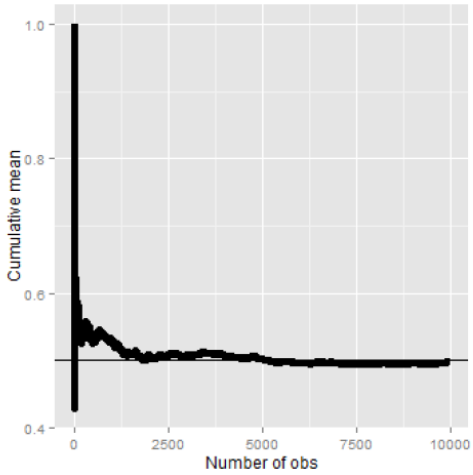
- Fortunately, for the sample mean there's a set of powerful results
- These results allow us to talk about the large sample distribution of sample means of a collection of iid observations
- The first of these results we inuitively know
  - It says that the average limits to what its estimating, the 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 #s 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
```




```
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

Central Limit Theorem

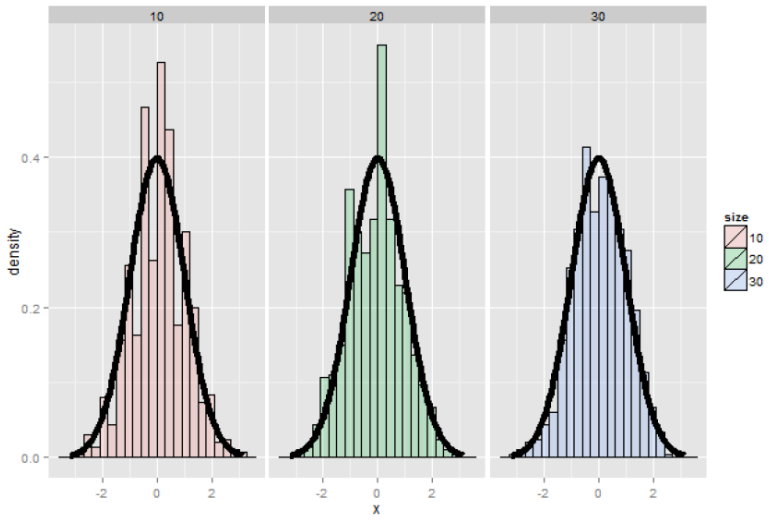
- The **Central Limit Theorem** (CLT) is one of the most important theorems in statistics
- For our purposes, the CLT states that the distribution of averages of iid variables (properly normalized) becomes that of a standard normal as the sample size increases
- The CLT applies in an endless variety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

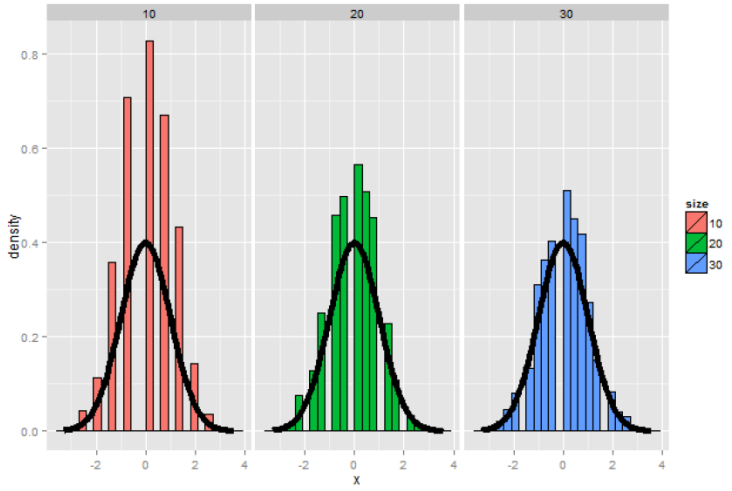


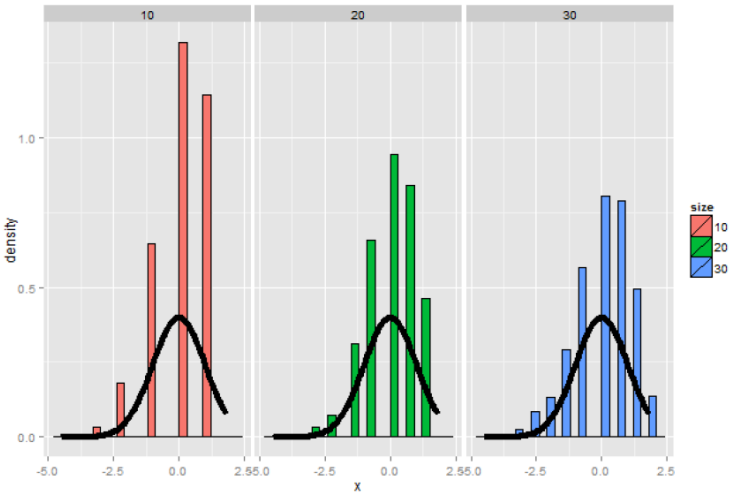
Coin CLT

- 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

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

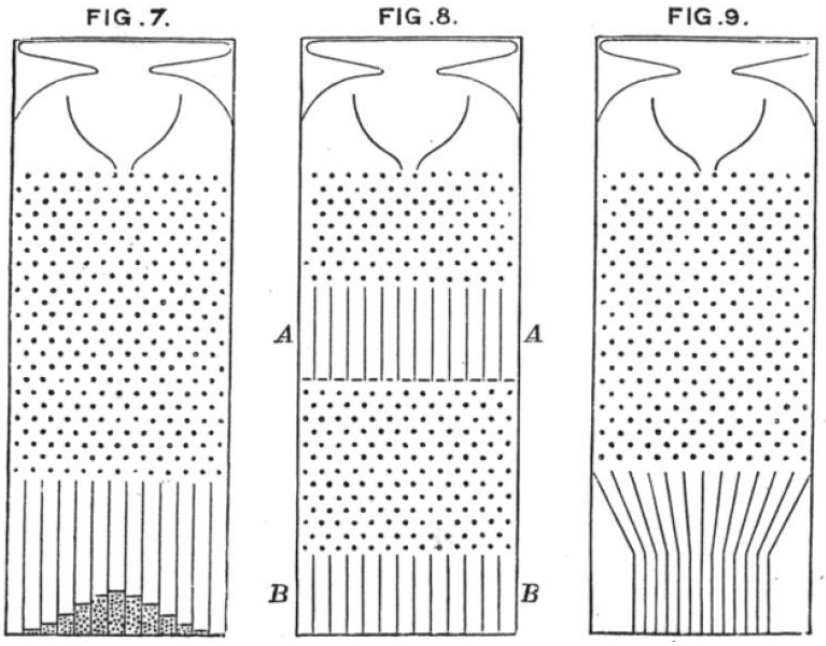
    will be approximately normally distributed
  - Let's flip a coin n times, take the sample proportion of heads, subtract off .5 and multiply the result by $2\sqrt{n}$ (divide by $1/(2\sqrt{n})$)





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)

Confidence Interval for Avg. Height of Sons

```
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)
```

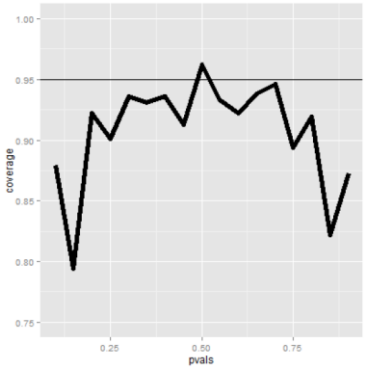
```
## [1] 0.4627 0.6573
```

```
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)
})
```



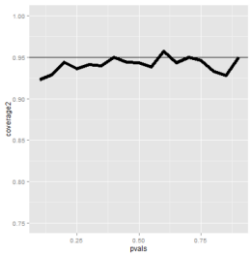
- n isn't large enough for the CLT to be applicable for many of the values of p
- Quick fix, form the interval with

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

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

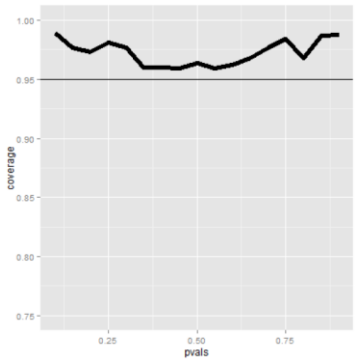
First let's show that coverage gets better with n

```
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)
})
```



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)
})
```



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

```
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
```

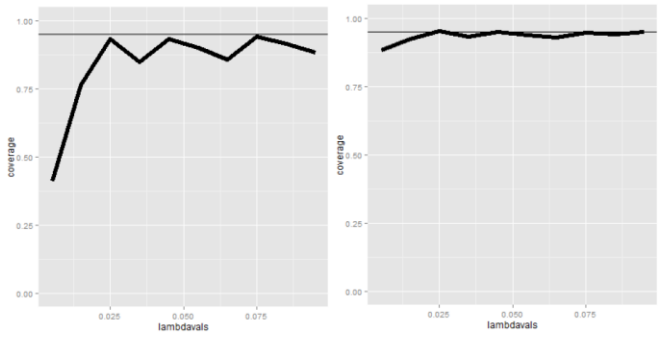
```
## [1] 0.01721 0.12371
## attr(,"conf.level")
## [1] 0.95
```

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)
})
```

(Gets really bad for small values of lambda)



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