Inference

Inference

We are going to have a competition. There is a prize.

You have to guess the proportion of blue beads in the jar. You can take a sample (with replacement) from the jar here:

https://dgrtwo.shinyapps.io/urn_drawing/

Each sample will cost you 10 cents ($0.10). So if you take a sample size of 100 you will have to pay me $10 to collect your prize.

Once you take your sample, please enter your guess, sample size, and an interval size. We will create an interval of this size with your guess in the middle. If the true percentage is not in your interval you are eliminated. The size of the interval will serve as a tie breaker (smaller wins).

http://goo.gl/forms/5bv4cRQWKA

The deadline to enter the competition is Monday February 22, 2016 at 9:00 AM.

Introduction

This chapter introduces the statistical concepts necessary to understand margins of errors, p-values and confidence intervals. These terms are ubiquitous in data science. Let's use polls as an example:

```
##                      pollster diff Obama Romney margin_of_error    n
## 1                      Gallup   -1    48     49            2.00 2551
## 2                   Rasmussen   -1    48     49            3.00 1500
## 3                    Monmouth    0    48     48            2.60 1417
## 4                         ARG    0    49     49            3.00 1200
## 5  Politico/GWU/Battleground    0    47     47            3.10 1000
## 6            Gravis Marketing    0    48     48            3.30  872
## 7                         CNN    0    49     49            3.50  693
## 8                     NBC/WSJ    1    48     47            2.55 1475
## 9                    IBD/TIPP    1    50     49            3.70  712
## 10     DailyKos/SEIU/PPP (D)    2    50     48            2.70 1300
## 11                        PPP    2    50     48            2.80 1200
## 12                        Pew    3    48     45            2.20 2709
## 13                  ABC/Post    3    50     47            2.50 2345
```

What does *margin of error* mean? The first step is define and understand random variables. To do this, we will generate our own imaginary election with 1 million voters of which proportion p are democrats and $1-p$ are republicans. To keep it interesting we will keep generate p at random and not peek.

```
n <- 10^6 ##number of voters
set.seed(1) ##so we all get the same answer
##pick a number between 0.45 and 0.55 (don't peek!):
p <- sample( seq(0.45,0.55,0.001),1)
x <- rep( c("D","R"), n*c( p, 1-p))
x <- sample(x) ##not necessary, but shuffle them
```

The population is now fixed. There is a true proportion p but we don't know it.

One election day we will do the following to decide who wins (don't ruin the fun by doing it now!):

```
prop.table(table(x))
```

Pollsters try to *estimate* p but asking 1 million people and actually unnecessary so instead. They take a poll. To do this they take a random sample. They pick N random voter phone numbers, call, and ask. Assuming everybody answers and every voter has a phone, we can mimic a random sample of 100 people with:

```
poll <- sample(x, 25, replace = TRUE)
```

The pollster then looks at the proportion of democrats in the sample and uses this information to predict p. In Statistics we say try to *estimate* p.

```
table(poll)
```

```
## poll
##  D  R
## 13 12
```

So our poll predicts a democrat win! Is this a good *estiamte*? We will see how powerful mathematical statistics is at informing us about exactly how good it is.

Notation: we use lower case x for the population of all voters and capital letters X for the random sample.

Random Variables

Random variable are outcomes of random process. The number of democrats out of the 25 we picked at random is an example. Let's repeat the exercise above a few times. Let's run a few other polls and see what happens:

```
X <- sample(x, 25, replace = TRUE)
sum( X=="D")
```

```
## [1] 14
```

```
X <- sample(x, 25, replace = TRUE)
sum( X=="D")
```

```
## [1] 15
```

```
X <- sample(x, 25, replace = TRUE)
sum( X=="D")
```

```
## [1] 11
```

Note how the observed number varies. We refer to this as *random* or *chance* variation. How does this random variable relate to our quantity of interest p? Statistical theory has a lot to teach about this and is the main tool used by pollsters and poll aggregators.

Sampling Models

Many procedures studied and used by data scientist can be modeled quite well as a sum of draws from a jar. For example, we modeled the process of polling likely voters as drawing 0s (republicans) and 1s (democrats) from a jar. To be more specific we can turn the `x` into 0s and 1s:

```
x <- as.numeric(x=="D")
sample(x, 25, replace = TRUE)
```

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

Many other examples can be constructed. For example to model our winning after betting $1 on red (on a roulette wheel) 10 times we can use:

```
color <- rep(c("Black","Red","Green"), c(18,18,2))
X <- sample( ifelse( color=="Red", 1,-1),  10, replace = TRUE)
sum(X)
```

```
## [1] 2
```

Because we know the proportions we can do this a bit quicker by simply using:

```
X <- sample( c(-1,1), 10, replace = TRUE, prob=c(10/19, 9/19))
sum(X)
```

```
## [1] -2
```

The Expected Value and Standard Error

Now that we have described sampling models, now we can describe properties of the sum of draws. The first important concept to learn is the *expected value*. The random variable will vary around this expected value. The *standard error* gives us an idea of how the size of the variation around the expected value.

The expected value of the sum of draws is the

$$\text{number of draws} \times \text{average of the numbers in the jar}$$

Using the definition of average:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

We can see that the average of values inside a jar with 0s and 1s is the number of 1s divided by n. This is the proportion of 1s. In general if a jar has two values a and b with proportions p and $(1-p)$ respectively, the average is $ap + b(1-p)$.

Assessment: What are the averages for the roulette examples? - Betting on black - Betting on 7

In our example with the election we know that the average of the box is p, we just don't know what p is just yet.

If our draws are independent, then the standard error of our sum is

$$\sqrt{\text{number of draws}} \times \text{ standard deviation of the numbers in the jar}$$

Using the definition of standard deviation

$$\sigma = \sqrt{\frac{1}{n}\sum_{i}^{n}(x_i - \mu)^2}$$

we can derive, with a bit of math, that if a jar has two values a and b with proportions p and $(1-p)$ respectively, the standard deviation is $\mid b - a \mid \sqrt{p(1-p)}$.

Assessment What are the expected value and standard errors of the number of democrats. Hint: The answers are functions of p, the proportion of democrats.

Let S be the number of democrats in our sample, or the sum of the 25 draws. Our expected value is

$$\mathrm{E}(S) = 25p$$

and standard error is

$$\mathrm{SE}(S) = \sqrt{25p(1-p)}$$

Two more intuitive results. If we divide S by constant, the expected value and standard error are also divided by this constant. This implies that

Now we now that the expected value of the proportion in our sample $S/25$, has expected value p.

$$\mathrm{E}(S/25) = p$$

which implies that $S/25$ will be p plus some chance error.

The SE,

$$\mathrm{SE}(S/25) = \sqrt{p(1-p)}/\sqrt{N}$$

gives us an idea of how large the error is.

We do notice that by making N larger we can make our standard error smaller. The *Law of Large Numbers* tells us that the bigger N the closer the sample average gets to the average of the jar which is the quantity we want to estimate.

Estimates

Because $S/25$ is p plus or minus some chance error we use it as our *estimate* of p. We usually put hats on top of our estimates: $\hat{p}$. We now now that when N is large

$$\hat{p} \approx p$$

.

Suppose that after we take our poll we want to give a best guess for p. One way we could do this is to report the observed value and it's standard error. Unfortunately, we need to know p to report an expected value. It turns out that plugging in the observed proportion, works out pretty well. So we can report:

```
p_hat <- mean(poll=="D")
cat("Our estimate of the percent of decomrats is",p_hat,
    "plus or minus", round( sqrt(p_hat*(1-p_hat)/25), 2))
```

```
## Our estimate of the percent of decomrats is 0.52 plus or minus 0.1
```

Which is not terribly useful. However, now we know that we can be more precise by taking a larger poll.

Assessment If $p = 0.5$, how large should the poll be to have a standard error of 2% or less ? Then take a poll and report expected value and standard error

$$\sqrt{0.25/N} = 0.02$$

implies

$$N = 0.25/0.02^2 = 625$$

.

```
N <- 625
X <- sample(x, N, replace = TRUE)
p_hat <- mean(X)
cat("Our estimate of the percent of decomrats is",p_hat,
    "plus or minus", round( sqrt(p_hat*(1-p_hat)/N), 2))
```

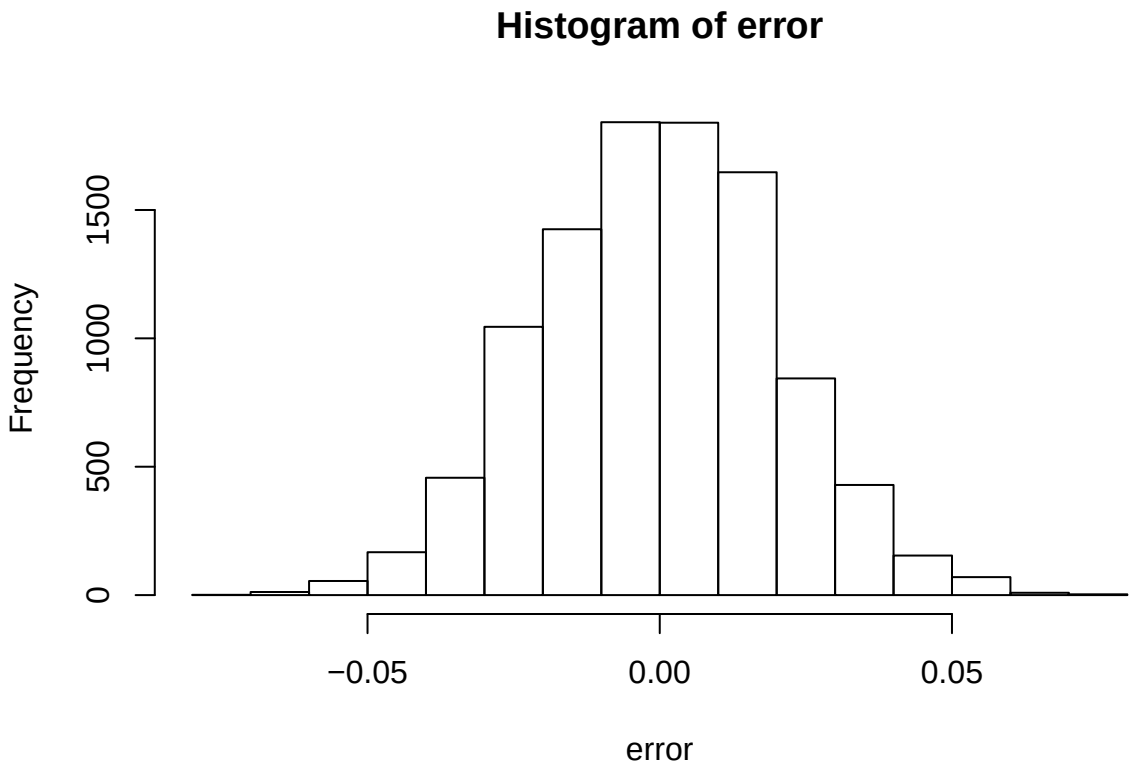
```
## Our estimate of the percent of decomrats is 0.4976 plus or minus 0.02
```

Probability Distribution for Random Variables

The plus or minus statements above are not very precise. Is it possible to say more? Can we, for example, compute the probability that $\hat{p}$ is within 1% of the actual p ? It turns out that we can. First let's start with a Monte Carlo simulation.

Assessment: Repeat the exercise of taking a poll of 625 likely voters, 10,000 times. Without peaking at p study the distribution of $\hat{p} - p$. Have we seen this distribution before? How often was our error smaller larger than 0.02?

```
N <- 625
B <- 10^4
error <- replicate(B,{
  X <- sample(x, N, replace = TRUE)
  mean(X)-p
})
hist(error)
```




```
mean(abs(error) > 0.02)
```

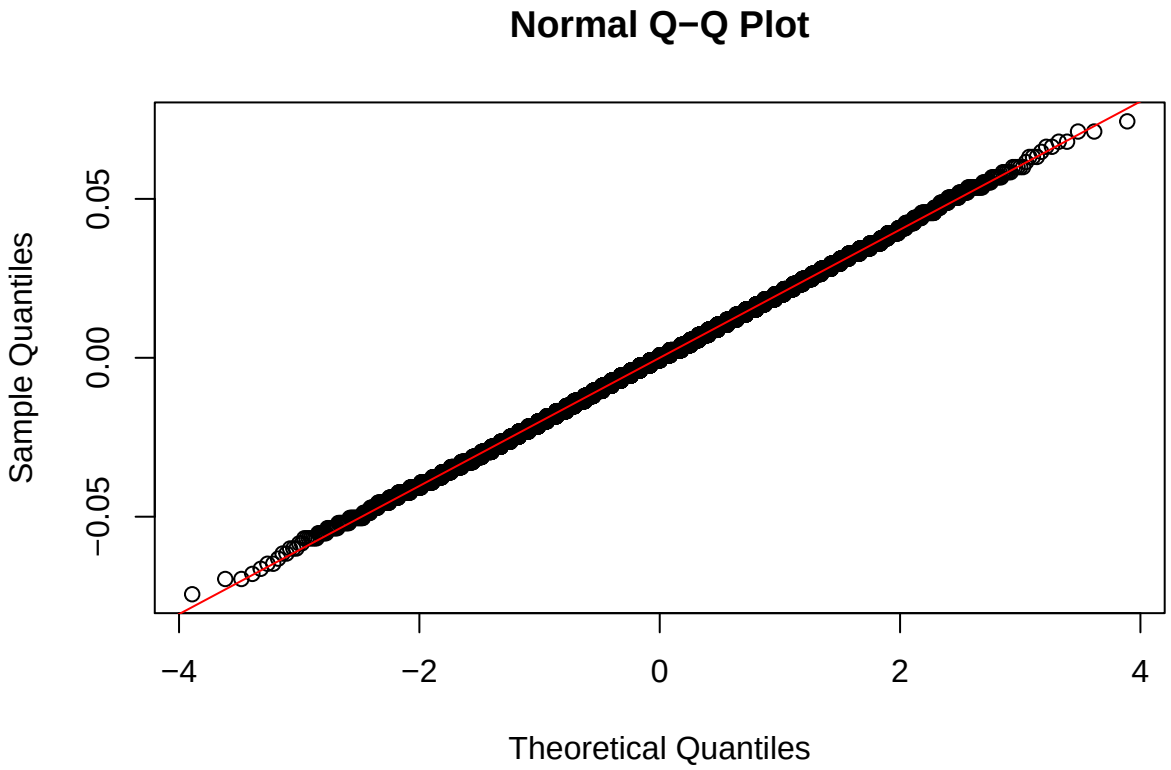
```
## [1] 0.3246
```

The distribution we see here is the *probability distribution* of our random variable. Knowing this distribution will be extremely helpful because we can, for example, report the probability that our error is smaller than a particular quantity. In the next section we describe a powerful mathematical result that helps simplify this.

Central Limit Theorem

We can see that the distribution of $\hat{p}$ is very well approximated with the normal distribution.

```
qqnorm(error)
qqline(error,col=2)
```



The fact that the distribution is centered at 0 confirms that the expected value of $\hat{p}$ is p. The fact that the standard deviation of the errors is

```
sqrt(mean( (error)^2))
```

```
## [1] 0.02024796
```

confirms that the standard error is in fact 0.02. Furthermore, we could have predicted that 32% of errors were larger than 0.02 using the normal approximation. What proportion of the normal curve is more than 1 SD away from average?

```
##compare
mean(abs(error) > 0.02)
```

```
## [1] 0.3246
```

```
##to
pnorm(-1) + (1-pnorm(1))
```

```
## [1] 0.3173105
```

Notice that the Monte Carlo simulation is like conducting 10,000 polls. This of course is practically impossible. But the mathematical theory saves us from having to do this! Without just one poll we are able to say that our estimate is $\hat{p}$ and there is 32% chance that our error is larger than 2%.

Assessment:

Suppose you are consulting for a casino. They want to know how many $1 bets a person needs to make so that the probability of negative earnings is less than 1 in 1,000.

1. We know that the average of the sampling model is -1/19 and the standard deviation is $\sigma = 2\sqrt{10/19 \times 9/19}$. Run a Monte Carlo simulation with $N = 400$ to learn about the distribution of the errors. Remember the error is the average winning minus $\mu = -1/19$.Confirm that the expected value of the error is 0 and that the standard error is $\sigma/\sqrt{N}$.

```
N <- 400
B <- 10^5
mu <- -1/19
sigma <- 2*sqrt(9/19*10/19)
error <- replicate(B,{
  X <- sample( c(-1,1), N, replace=TRUE, prob=(c(10/19,9/19)) )
  mean(X) - mu
})
mean(error)
```

```
## [1] -4.132105e-05
```

```
sqrt( mean( error^2))
```

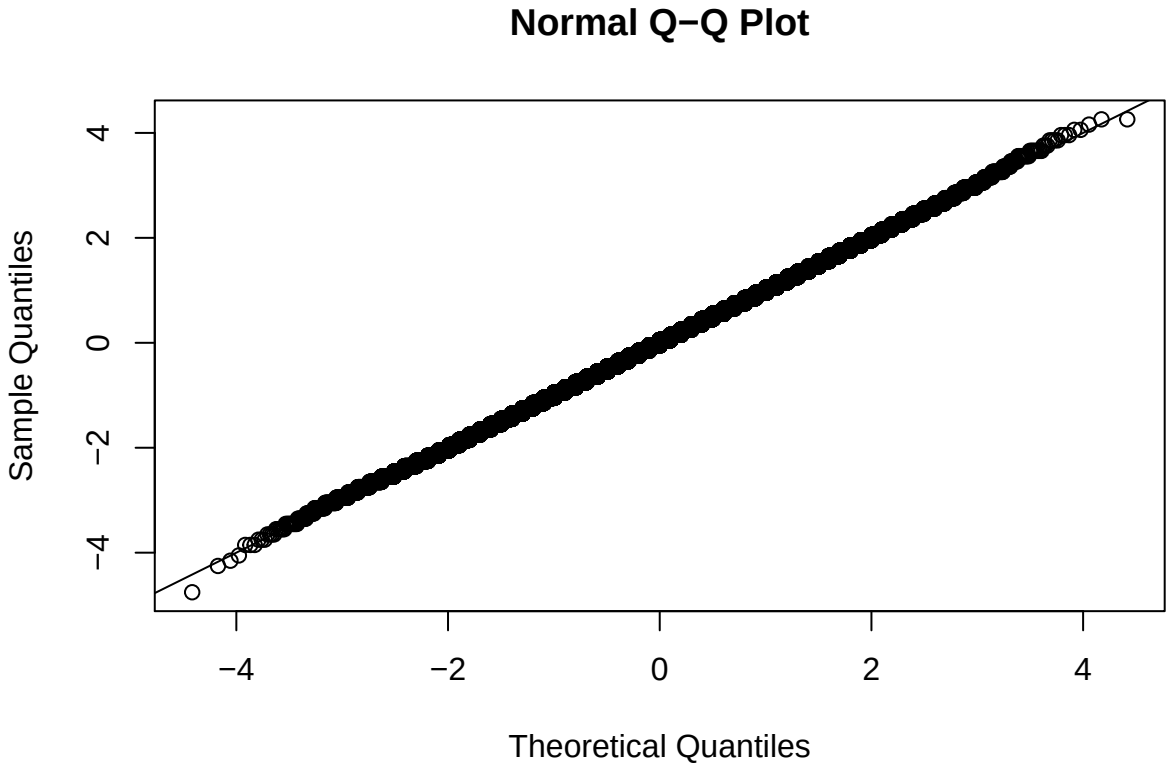
```
## [1] 0.05006588
```

```
sigma/sqrt(N)
```

```
## [1] 0.0499307
```

2. Confirm that the distribution is approximately normal with mean 0 and standard deviation σ/N

```
z <- error/ (sigma / sqrt(N)  )
qqnorm(z)
abline(0,1)
```



3. How large does the error have to be for the casino to lose money?

```
## if x is 0, the the error is:
0 - mu
```

```
## [1] 0.05263158
```

4. Use the CLT to compute the probability of a positive earnings. Confirm with the simulation results.

Let S be the sum of the draws.

$$Z = \sqrt{N}\frac{S/N - \mu}{\sigma}$$

is standard normal. We want to know

$$\Pr(\text{S/N} > 0)$$

We rewrite both sides so that Z is on the left:

$$\Pr(\sqrt{N}\frac{S/N - \mu}{\sigma} > -\sqrt{N}\frac{\mu}{\sigma})$$

which is

$$\Pr(Z > -\sqrt{N}\frac{\mu}{\sigma})$$

and we can compute

```r
1 - pnorm( -sqrt(N)*mu/sigma )
```

```
## [1] 0.1459203
```

```r
##and compare to
mean(error > -mu)
```

```
## [1] 0.13458
```

5. With $N = 400$ the probability of negative earnings is too high. What does N need to be for the probability of positive earnings is 0? Confirm with a simulation.

We know that if $z = \Phi^{-1}(1 - 10^{-3})$ then

$$\Pr(Z > z) = 10^{-3}$$

So we need

$$-\sqrt{N}\frac{\mu}{\sigma} = \Phi^{-1}(1 - 10^{-3})$$

or

$$N = (-\Phi^{-1}(1 - 10^{-3})\sigma/\mu)^2$$

```r
N <- ceiling( -qnorm(1-10^-3)*sigma/mu )^2
```

Lets' confirm:

```r
N <- ceiling( -qnorm(1-10^-3)*sigma/mu )^2
B <- 10^5
mu <- -1/19
sigma <- 2*sqrt(9/19*10/19)
profit <- replicate(B,{
  X <- sample( c(-1,1), N, replace=TRUE, prob=(c(10/19,9/19)) )
  mean(X)
})
mean(profit>0)
```

```
## [1] 0.00104
```

Assessment

In R we have access to a large population of male heights. We can access them like this

```r
data(father.son, package="UsingR")
y <- father.son$sheight
```

Suppose you are demographer and want to get an estimate of the average height of this population. We can use sampling to do this as well. We will sample 25 men and measure them.

1. What is the average and standard deviation of our sampling model? What is the expected value and standard deviation.

```
mu <- mean(y)
sigma <- sqrt(mean((y-mu)^2))
```

2. Consider the random variable defined by taking a sample of size N and then computing the average. What are the expected value and standard error of this random variable:

```
N <- 25
mu
```

```
## [1] 68.68407
```

```
sigma/sqrt(N)
```

```
## [1] 0.5626792
```

3. What is the probability that our estimate is within one inch of the population average?

Let's call the sample average

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

Then we have

$$\Pr(\mid \bar{Y} - \mu \mid \leq 1) = \Pr(\sqrt{N} \mid \bar{Y} - \mu \mid /\sigma \leq \sqrt{N}1/\sigma)$$

```
(pnorm(sqrt(N)/sigma) - pnorm(-sqrt(N)/sigma))
```

```
## [1] 0.9244666
```

Note that if we wanted to compute this last quantity in practice we need to know σ. But we don't.

A common approach is to use the sample standard deviation as a stand-in. We define the sample standard deviation as

$$s = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

The function `sd` computes this quantity

```
Y <- sample(y, N, replace=TRUE)
sd(X)
```

```
## [1] 0.5003947
```

```
1 - pnorm(sqrt(N)/sd(Y))
```

```
## [1] 0.009681883
```

```
#is an approximation of
1 - pnorm(sqrt(N)/sigma)
```

```
## [1] 0.0377667
```

```
##that is completly derived from data
```

The t-distribution (optional)

If you have heard of the t-distribution, you should know that here is where we would use it. If the original data is normal, as height data is, then there is a better approximation for

$$\sqrt{N}\frac{\bar{X}-\mu}{s}$$

than the normal distribution and it is the t-distribution. Note these are closer

```
1 - pnorm(sqrt(N)/sigma)
```

```
## [1] 0.0377667
```

```
#is an approximation of
1 - pt(sqrt(N)/sd(X), N-1)
```

```
## [1] 2.49593e-10
```

```
1 - pnorm(sqrt(N)/sd(X))
```

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

Confidence Intervals

If we create intervals that miss the target 32% of the time, we will not be considered good forecasters. We can always make less bold predictions, such as "the percent of democrats will be between 0% and 100%" but that will not help our reputation either. Can we find a better balance?

Because there is chance variation, it is impossible to hit the target 100% of time unless we state the obvious "between 0% and 100%". So let's compromise a bit. If we hit the target 95% of the time, chances are we will be considered good forecasters. Can we use the theory above to construct such interval?

What we want to do is use the result of our poll to construct an interval, say, $[A, B]$ such that

$$\Pr(A \leq p \text{ and } B \geq p) \geq 0.95$$

we want to make this interval as small as possible. How do we choose A and B?

We know $\hat{p}$ follows a normal distribution with expected value p and standard error approximately equal to $\sqrt{\hat{p}(1-\hat{p})}/\sqrt{N}$. This implies that the following random variable

$$Z = \sqrt{N}\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})}}$$

is approximately normal with expected 0 and standard deviation 1.

Assessment What value can we make z so that $\Pr(\mid Z \mid \leq z) = 0.95$?

Answer:

$$z = \Phi^{-1}(0.975)$$

Because:

```
z = qnorm(0.975)
pnorm(z) - pnorm(-z)
```

```
## [1] 0.95
```

So we have our interval. We need to actually define it in terms of $\hat{p}$, N, and z because these are the only quantity we can actually construct.

Assessment: Practice the math How do we rewrite the event

$$\mid Z \mid \leq z$$

so that it can be written as an interval of quantities we know.

$$\mid Z \mid = \sqrt{N}\frac{\mid \hat{p}-p \mid}{\sqrt{\hat{p}(1-\hat{p})}} \leq z$$

$$\mid \hat{p}-p \mid \leq z\sqrt{\hat{p}(1-\hat{p})}\sqrt{N}$$

Assuming z is positive,

$$\hat{p} - z\sqrt{\hat{p}(1-\hat{p})}/\sqrt{N} \leq p \leq \hat{p} + z\sqrt{\hat{p}(1-\hat{p})}/\sqrt{N}$$

```
p_hat + c(-1,1)*qnorm(0.975)*sqrt(p_hat*(1-p_hat))/sqrt(N)
```

```
## [1] 0.3016059 0.6935941
```

This interval has a 95% chance of *covering* the true p

Did we actually make a correct prediction? It's election night and we we will find out

```
ci <- p_hat + c(-1,1)*qnorm(0.975)*sqrt(p_hat*(1-p_hat))/sqrt(N)
p
```

```
## [1] 0.476
```

```
ci[1] <= p & ci[2] >= p
```

```
## [1] TRUE
```

Assessment Construct a 99% confidence interval. Is it more or less specific than the 95%?

```
p_hat + c(-1,1)*qnorm(0.995)*sqrt(p_hat*(1-p_hat))/sqrt(N)
```

```
## [1] 0.24002 0.75518
```

Assessment You have a coin. You want to figure out if it's fair or not. To create your coin run the following code:

```
set.seed(2016)
prob_heads <- sample(c(0.5, 0.6, 0.4),1)
```

1. Now the probability is set. The coin is either fair or not, but we don't know which. Use a Monte Carlo simulation to toss your coin 100 times. Report the proportion of heads

```
N <- 100
X <- sample(c(0,1), N, replace=TRUE,
            prob=c(1-prob_heads,prob_heads))
ph_hat <- mean(X)
ph_hat
```

```
## [1] 0.59
```

2. Is it fair? Use CLT to construct a 95% confidence based on this estimate.

```
conf_int <- ph_hat +
  c(-1,1)*qnorm(0.975)*sqrt(ph_hat*(1-ph_hat))/
  sqrt(N)
conf_int
```

```
## [1] 0.4936024 0.6863976
```

3. Does the interval cover the true value of `prob_head`?

```
prob_heads>=ci[1] & prob_heads<=ci[2]
```

```
## [1] TRUE
```

4. No use a Monte Carlo simulation to re run this experiment (keeping `prob_hear` fixed) 10000 times. Do we confirm that we cover 95% of the time?

```
B <- 10000
cover <- replicate(B,{
  X <- sample(c(0,1), N, replace=TRUE,
              prob=c(prob_heads,1-prob_heads))
  ph_hat <- mean(X)
  ci <- ph_hat +
    c(-1,1)*qnorm(0.975)*sqrt(ph_hat*(1-ph_hat))/
    sqrt(N)
  prob_heads>=ci[1] & prob_heads<=ci[2]
})
mean(cover)
```

```
## [1] 0.9463
```

Confidence Intervals: Monte Carlo Simulaion

Note that Nate Silver stated that Obama had a 90% chance of wining. Was this a confidence interval?

The description we have given up to know says nothing about the probability of winnings. In fact, the statement does not even make sense because p is fixed. It is not random. We should not make probability statements.

Later we will learn about Bayesian statistics were data-based statements such as "Obama has a 90% chance of winning" make sense. Here we clarify how we interpret margins of error and confidence intervals.

The 95% confidence interval

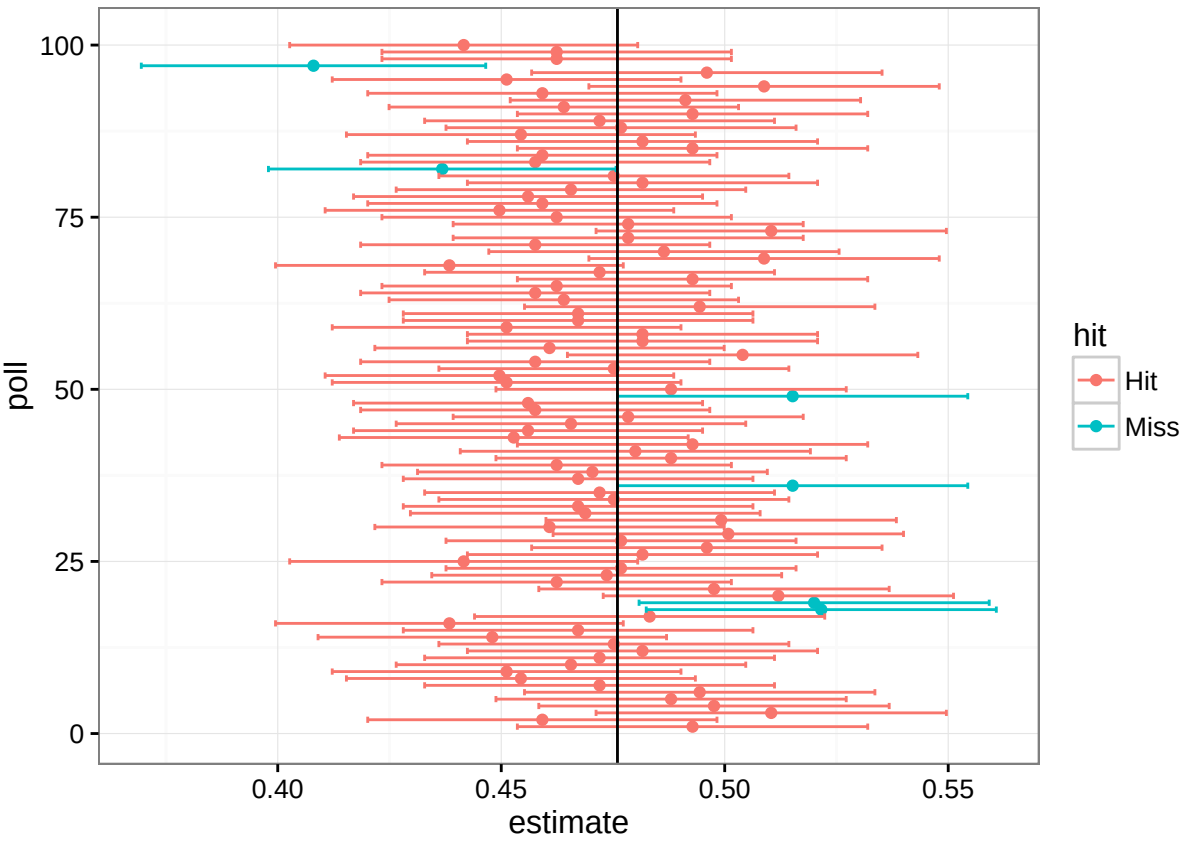
```
p_hat + c(-1,1)*qnorm(0.975)*sqrt(p_hat*(1-p_hat))/sqrt(N)
```

```
## [1] 0.3996029 0.5955971
```

is random because p_hat is random. Let's use a Monte Carlo simulation to see this:

```
library(dplyr)
library(ggplot2)
N <- 625
B <- 100
tab  <- replicate(B,{
  X <- sample(x, N, replace = TRUE)
  p_hat <- mean(X)
  ci <- p_hat + c(-1,1)*qnorm(0.975)*sqrt(p_hat*(1-p_hat))/sqrt(N)
  hit <- ci[1] <= p  & ci[2] >= p
  c(p_hat,ci,hit)
})

tab <- data.frame(poll=1:ncol(tab), t(tab))
names(tab)<-c("poll","estimate","low","high","hit")
tab <- mutate(tab, hit=ifelse(hit, "Hit","Miss") )
ggplot(tab, aes(poll,estimate,ymin=low,ymax=high,col=hit)) +geom_point(aes())+geom_errorbar() + coord_fl
```



Notice how the interval is different after each poll: it's random. So the 95% relates to the probability that this random interval falls on top of p, not that $p < .50$ or whatever other statements related to the probability of winning. In the figure above, that shows the interval for 100 polls, you can see that interval fall on p about 95% of the time.

Power

Pollsters are not successful for providing correct confidence intervals but rather for predicting who will win. Unfortunately, our confidence interval included 0.5 so if forced to declare a winner, we would have to say it was a "toss-up". A problem with our poll results is that, given the sample size and the value of p, we would have to sacrifice on the probability of an incorrect call to create an interval that does not include 50. In this particular case, the reported margin of error would have had to be about 3.5%

Assessment

What percent of confidence intervals with margins of error of 0.03 would correctly predict the election? Check with a Monte Carlo simulation.

Define

$$\delta = 0.035$$

We want to know

$$\Pr(-\delta < \mid p - \hat{p} \mid < \delta)$$

$$\Pr(|Z| < \sqrt{N}\delta / \sqrt{p(1-p)})$$

Because Z is approximately standard normal this probability is approximately

```
a = sqrt(N)*.035/sqrt(p_hat*(1-p_hat))
pnorm(a)-pnorm(-a)
```

```
## [1] 0.9198852
```

This a .91% confidence interval. Which is not bad, but it's not the 95% we set out for.

```
N <- 625
B <- 10^5
success  <- replicate(B,{
  X <- sample(x, N, replace = TRUE)
  p_hat <- mean(X)
  ci <- p_hat + c(-0.035,0.035)
  ci[1] <= p  & ci[2] >= p
})
mean(success)
```

```
## [1] 0.92121
```

Assessment

If we know the percent of democrats can be as close to 50% as, say, 48%, how large does the sample size have to be so that the margin of error is about 2%.

Solution we need: We add $z_{95}\sqrt{p(1-p)}\sqrt{N}$ and we want this to be 0.02. So

```
(qnorm(0.975)*sqrt(0.48*0.52)/0.02)^2
```

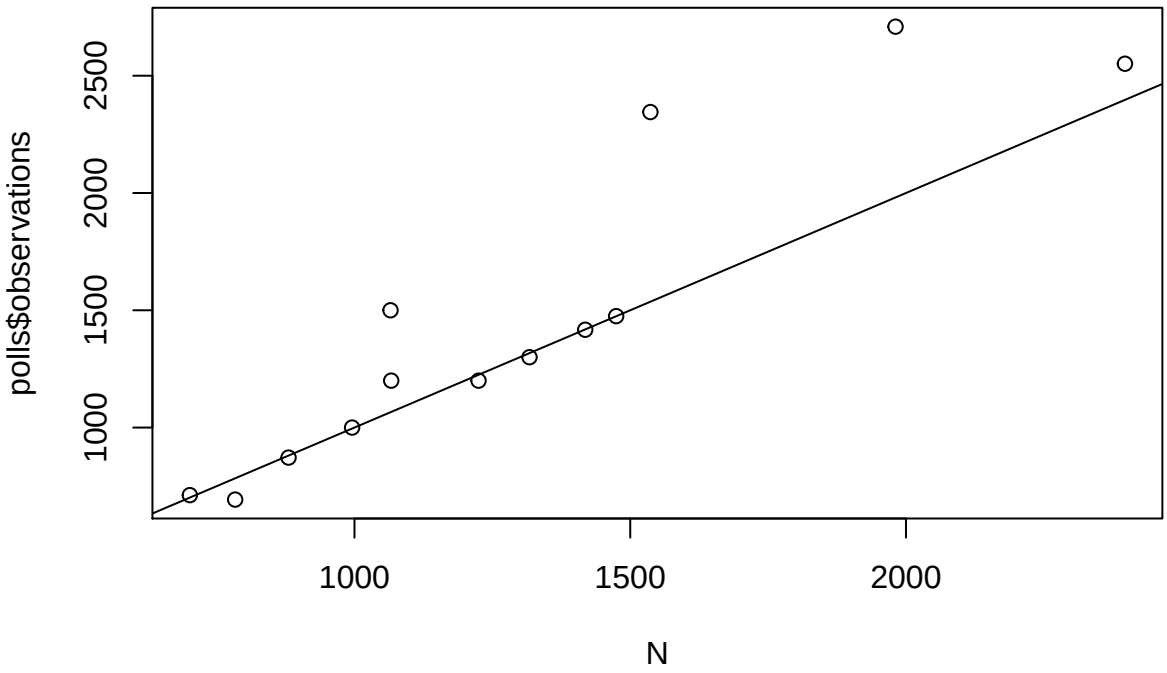
```
## [1] 2397.07
```

If you go back to the table with which we started, you notice that many polls have sample sizes between 500 and 2500.

Assessment From the `polls` object determine the sample size from the margin of error. The plot it against the reported sample size.

What do you see? A - A flat line B - Exactly a straight line C - Almost a straight line with some points off D - The function $f(x) = \sqrt{(}x)$

```
p_hat <- polls$Obama/100
N <- ( qnorm(0.975) * sqrt(p_hat*(1-p_hat)) /
         (polls$margin_of_error/100))^2
plot(N, polls$observations)
abline(0,1)
```



In-class questions: 2/29/2016

We're going to try something new today. If you have small technical questions during class, go to this link and ask away.

p-values

p-values are ubiquitous in the scientific literature. They are related to confidence interval so we introduce the concept here.

Let's consider the blue and red beads. Suppose that rather than wanting an estimate of the percent of blue beads I am more interested in the question are there more blue beads or red beads.

Suppose we take a random sample of $N = 100$ and we observe 53 blue beads. This seems to be pointing to their being more blue than red. However, as data scientists we need to be skeptical. We know there is chance involved in this process and we could get a 53 even when the proportions of red and blue are the same. We call this a *null hypothesis*. The null hypothesis is the skeptics hypothesis: the proportion of blue beads p is 0.5. We have observed a random variable $\hat{p} = 0.53$ and the p-value is the answer to the question how likely is it to see a value this large, when the null hypothesis is true. So we write

$$\Pr(\mid \hat{p} - 0.5 \mid > 0.03)$$

assuming the $p = 0.5$. Under the null we know that

$$\sqrt{N}\frac{\hat{p} - 0.5}{\sqrt{0.5(1-0.5)}}$$

is standard normal. So we can compute the probability above, which is the p-value.

$$\Pr\left(\sqrt{N}\frac{\mid \hat{p} - 0.5 \mid}{\sqrt{0.5(1-0.5)}} > \sqrt{N}\frac{0.03}{\sqrt{0.5(1-0.5)}}\right)$$

```
N=100
z <- sqrt(N)*0.03/0.5
1 - (pnorm(0.6) - pnorm(-0.6))
```

```
## [1] 0.5485062
```

So we do in fact have reason to be a skeptics. By constructing a p-value we see that

Assessment: If $\hat{p} = 51$ and $N = 10000$. What is the p-value?

We can show that if a $x \times 100\%$ confidence interval does not include 0, then the p-value must be smaller than $1 - x$. So they provide related information. However, the confidence interval is always more informative as it gives information of the size of the estimate.

Assessment: Suppose you are comparing average test scores from two school districts. You have exams from a random sample of 10000. The p-value for the difference in average scores is 0.01. Should we change the curriculum of the district with highly significant statistical differences?

Setting the random seed Before we continue, we briefly explain the following important line of code:

```
set.seed(1)
```

Throughout this book, we use random number generators. This implies that many of the results presented can actually change by chance, including the correct answer to problems. One way to ensure that results do not change is by setting R's random number generation seed. For more on the topic please read the help file:

```
?set.seed
```