Inference for a single proportion

```r
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
theme_set(theme_bw())
```

Set up data, make sure everything checks out

```r
correct <- 571
incorrect <- 99
N <- correct + incorrect


p <- correct/N
p
```

```
## [1] 0.8522388
```

Let's build our estimate of what the sampling distribution would be like

```r
empirical_sample <- c(rep("Correct", correct), rep("Incorrect", incorrect))

head(empirical_sample)
```

```
## [1] "Correct" "Correct" "Correct" "Correct" "Correct" "Correct"
```

```r
sample_func <- function(){
  hypothetical_sample <- sample(empirical_sample, N, replace = TRUE)
  mean(hypothetical_sample == "Correct")
}

sample_func()
```

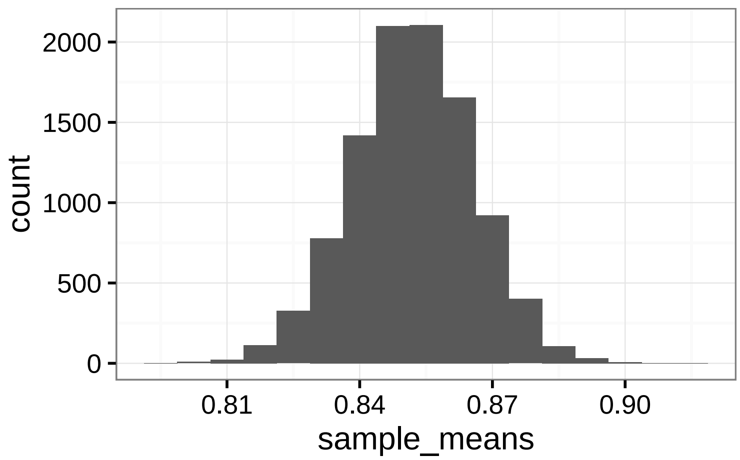
```
## [1] 0.8597015
```

We expect from CLT that p_hat ~ Normal(p, sqrt((p*1-p)/n))

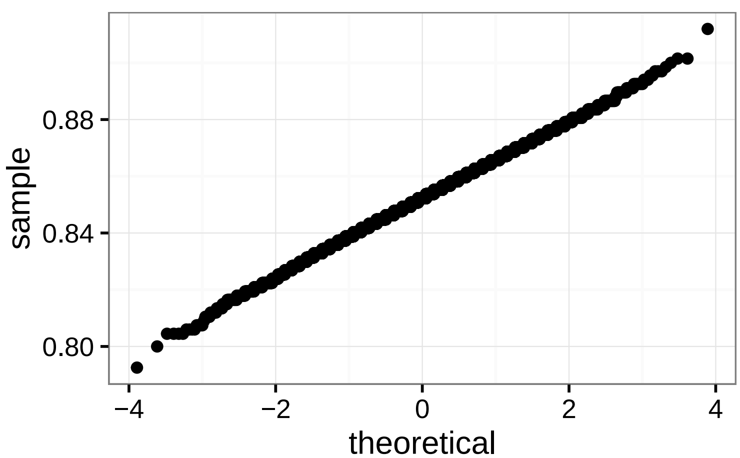
What does it's shape look like?

```r
sample_means <- replicate(10000,sample_func())

qplot(sample_means, binwidth = .0075)
```




```r
ggplot(data.frame(x = sample_means), aes(samples = x)) +
  stat_qq()
```



What do it's parameters look like?

```r
sample_means <- replicate(10000,sample_func())

pop_mean_estimate <- mean(sample_means)
pop_mean_estimate
```

```
## [1] 0.8521412
```

```r
sample_mean_sd <- sd(sample_means)
sample_mean_sd
```

```
## [1] 0.01377049
```

```r
predicted_mean_sd <- sqrt((correct/N)*(incorrect/N)/N)
predicted_mean_sd
```

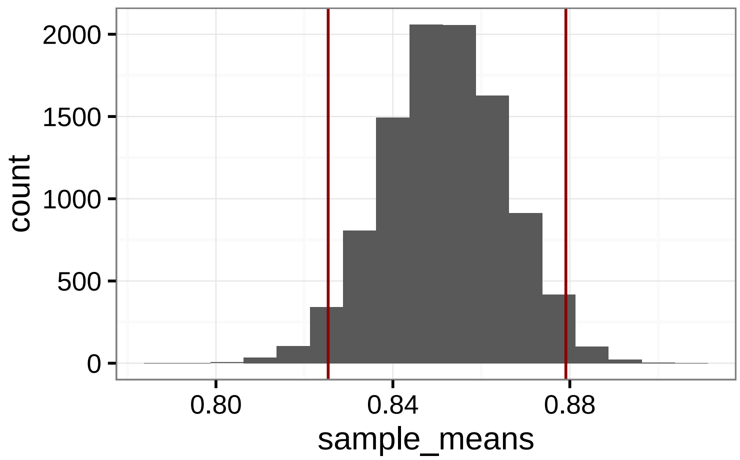
```
## [1] 0.01370956
```

Let's look at confidence intervals

```r
sample_means <- replicate(10000,sample_func())

qplot(sample_means, binwidth = .0075) +
  geom_vline(aes(xintercept = quantile(sample_means,.025)), color = "darkred") +
```

```r
  geom_vline(aes(xintercept = quantile(sample_means,.975)), color = "darkred")
```



Let's get confidence intervals

```r
sample_means <- replicate(10000,sample_func())

sample_mean_quantiles <- quantile(sample_means, probs = c(.025, .975))
sample_mean_quantiles
```

```
##      2.5%     97.5%
## 0.8253731 0.8791045
```

```r
predicted_mean_quantiles <- c((correct/N) - 1.96 * sqrt((correct/N) *(incorrect/N)/N),
                              (correct/N) + 1.96 * sqrt((correct/N) *(incorrect/N)/N))

predicted_mean_quantiles
```

```
## [1] 0.8253681 0.8791095
```

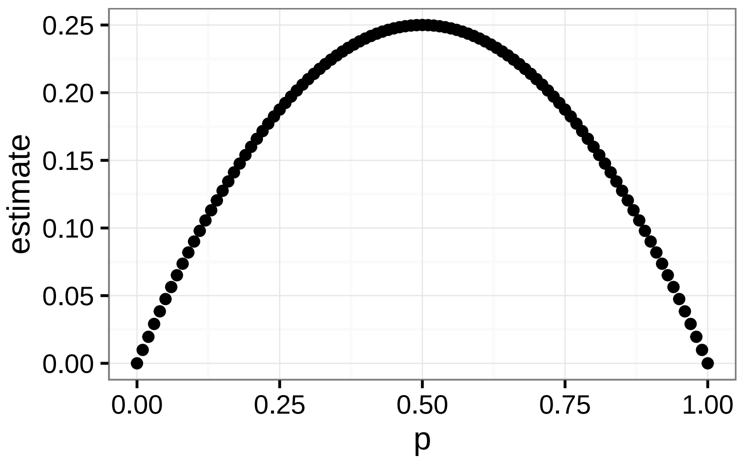
Why p_hat = .5 is conservative

```r
p_check <- data.frame(p = seq(0, 1, .01)) %>%
  mutate(estimate = p * (1-p))

head(p_check)
```

```
##      p estimate
## 1 0.00   0.0000
## 2 0.01   0.0099
## 3 0.02   0.0196
## 4 0.03   0.0291
## 5 0.04   0.0384
## 6 0.05   0.0475
```

```r
qplot(data = p_check, x = p, y = estimate)
```



Hypothesis testing with CIs. First use Sampling.

```
chance_p <- .5

null_correct <- chance_p * N
null_incorrect <- (1-chance_p) * N

null_sample <- c(rep("Correct", null_correct), rep("Incorrect", null_incorrect))

head(null_sample)
```

```
## [1] "Correct" "Correct" "Correct" "Correct" "Correct" "Correct"
```

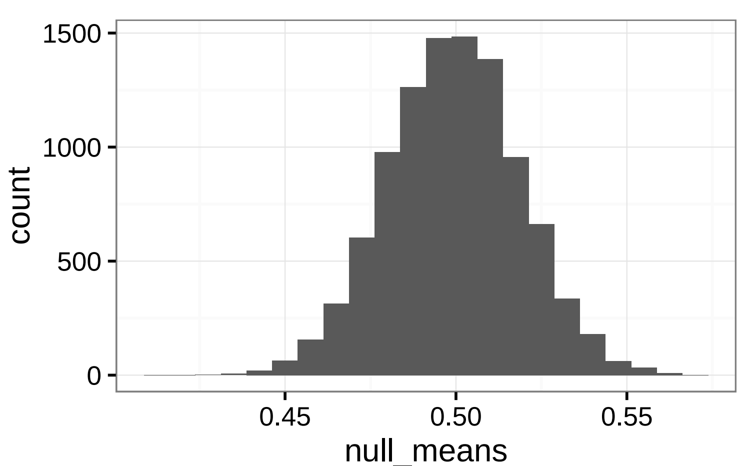
```
null_func <- function(){
  hypothetical_sample <- sample(null_sample, N, replace = TRUE)
  mean(hypothetical_sample == "Correct")
}

null_func()
```

```
## [1] 0.5014925
```

```
null_means <- replicate(10000,null_func())


qplot(null_means, binwidth = .0075)
```



```r
null_quantiles <- quantile(null_means, probs = c(.025, .975))
null_quantiles
```

```
##      2.5%     97.5%
## 0.4611940 0.5373134
```

Now use CLT

```r
chance_p <- .8

null_correct <- chance_p * N
null_incorrect <- (1-chance_p) * N


predicted_null_quantiles <- c((null_correct/N) - 1.96 *
                                sqrt((null_correct/N) *(null_incorrect/N)/N),
                              (null_correct/N) + 1.96 *
                                sqrt((null_correct/N) *(null_incorrect/N)/N))

predicted_null_quantiles
```

```
## [1] 0.7697114 0.8302886
```