Normal Distributions

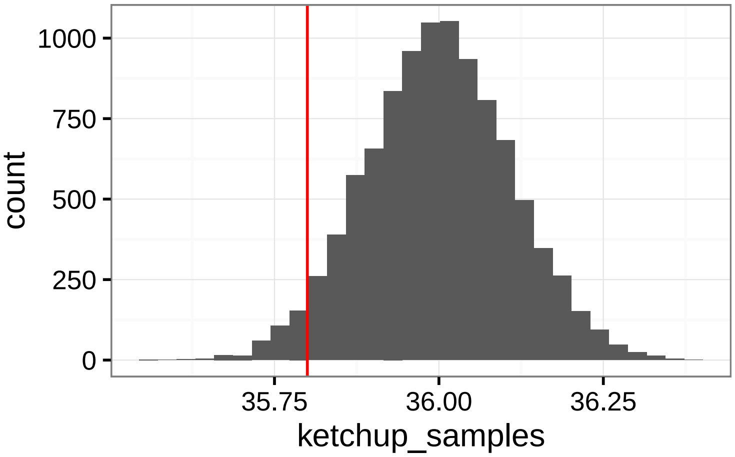
Load libraries

```
library(ggplot2)
library(dplyr)
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
```

Let's take a quick look at the empirical data

```
#Generate 10000 samples from the appropriate normal distribution
ketchup_samples <- rnorm(10000, mean = 36, sd = .11)

# Where is 35.8oz on this plot?
qplot(ketchup_samples) +
  theme_bw() +
  geom_vline(xintercept = 35.8, color = "red")
```



Now let's check the percentile of 35.8.

First, let's use sampling:

```
#Generate 10000 samples from the appropriate normal distribution
ketchup_samples <- rnorm(10000, mean = 36, sd = .11)

# What percent are below 35.8oz?
sum(ketchup_samples < 35.8)/length(ketchup_samples)
```

```
## [1] 0.0356
```

Now, let's use the normal distribution directly

```
#Let's check this directly:
pnorm(35.8, mean = 36, sd = .11)
```

```
## [1] 0.03451817
```

Now let's convert 35.8 to it's Z-score and check against the standard normal distribution

```
#Now in z-scores!
z_score <- (35.8 - 36)/.11
z_score
```

```
## [1] -1.818182
```

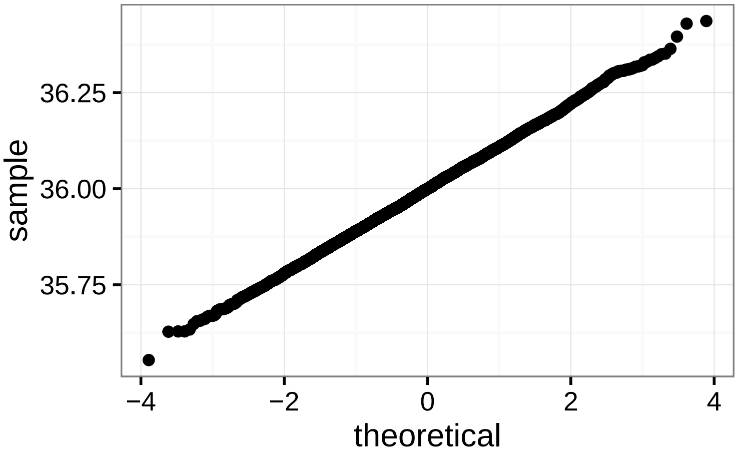
```r
pnorm(z_score, mean = 0, sd = 1)
```

```
## [1] 0.03451817
```

QQPlots

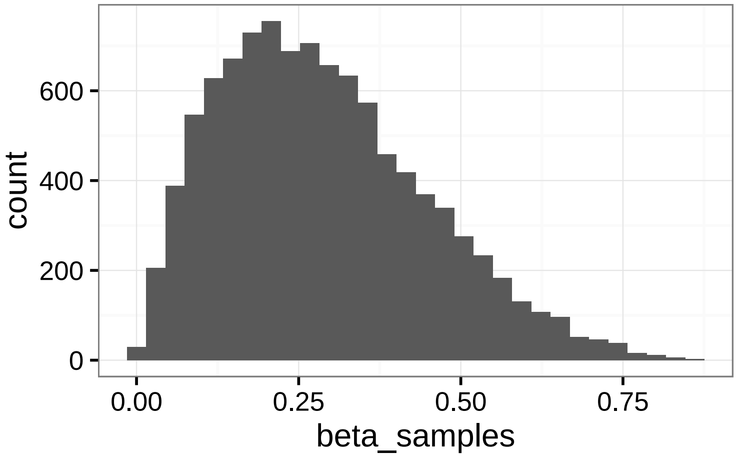
```r
ketchup_df <- data.frame(y = ketchup_samples)

ggplot(ketchup_df, aes(sample = y)) +
  geom_qq() +
  theme_bw()
```

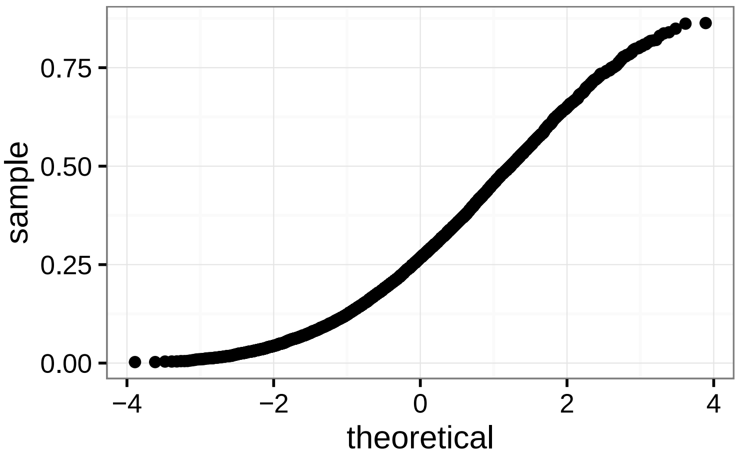


```r
beta_samples <- data.frame(sample = rbeta(10000, 2, 5))

qplot(beta_samples) +
  theme_bw()
```

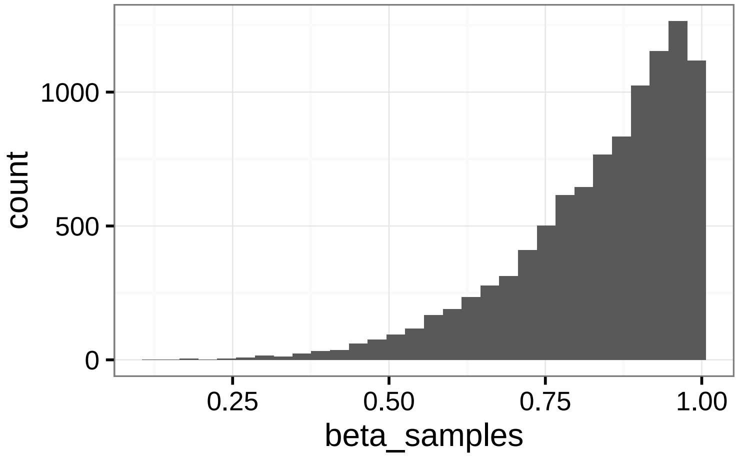


```r
ggplot(beta_samples, aes(sample = sample)) +
  geom_qq() +
  theme_bw()
```

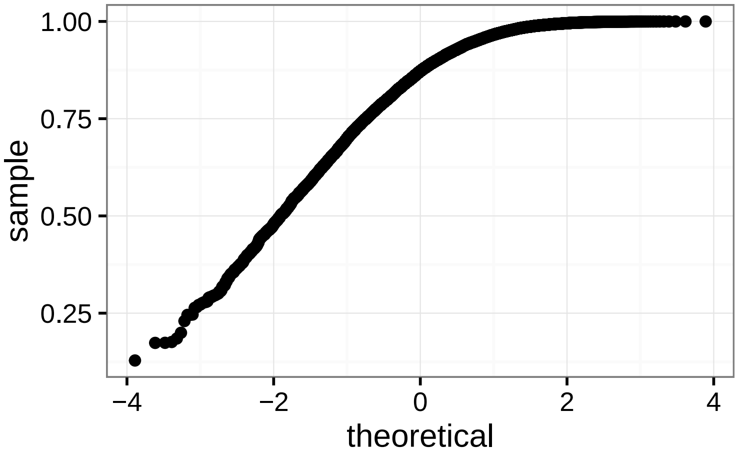


```r
beta_samples <- data.frame(sample = rbeta(10000, 5, 1))

qplot(beta_samples) +
  theme_bw()
```



```r
ggplot(beta_samples, aes(sample = sample)) +
  geom_qq() +
  theme_bw()
```

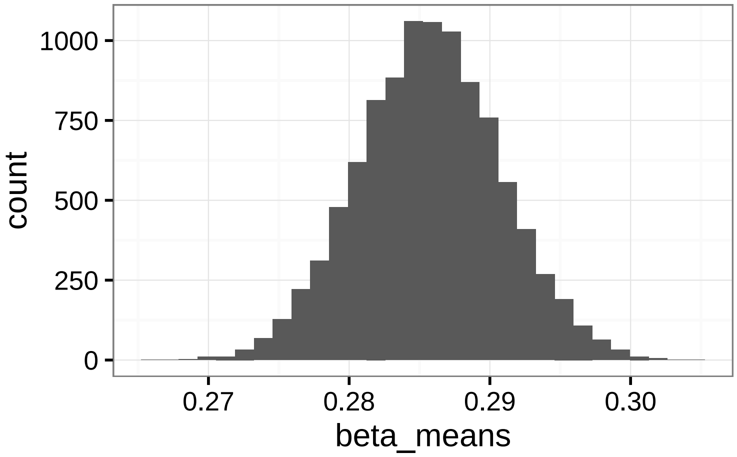


Let's use QQ Plots to check the central limit theorem

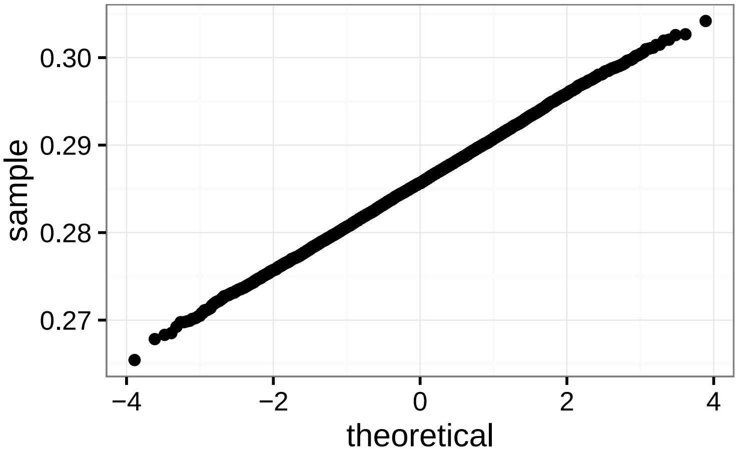
```
num_per_sample <- 1000
num_samples <- 10000

beta_means <- data.frame(mean = replicate(num_samples,
                                          mean( rbeta(num_per_sample, 2, 5))))


qplot(beta_means) +
  theme_bw()
```



```
ggplot(beta_means, aes(sample = mean)) +
  geom_qq() +
  theme_bw()
```



Three Different distributions

```
num_per_sample <- c(10, 30, 100, 1000, 1000)

get_sample_stats <- function(num_per_sample) {
  samples <- rbeta(num_per_sample, 2, 5)
  data.frame(num_per_sample = num_per_sample,
             mean = mean(samples), sd = sd(samples),
             median = median(samples))
}
```

```r
lapply(num_per_sample, get_sample_stats) %>%
  bind_rows()
```

```
##   num_per_sample      mean        sd    median
## 1             10 0.2699872 0.1468595 0.2663796
## 2             30 0.2808902 0.1688676 0.2281687
## 3            100 0.2955892 0.1628067 0.2669570
## 4           1000 0.2851330 0.1605109 0.2655098
## 5           1000 0.2906032 0.1644631 0.2697304
```

```r
get_stats_of_sample_stats <- function(num_per_sample) {

  samples <- replicate(1000, mean(rbeta(num_per_sample, 2, 5)))

 data.frame(num_per_sample = num_per_sample,
             mean = mean(samples), sd = sd(samples),
            median = median(samples))
}


lapply(num_per_sample, get_stats_of_sample_stats) %>%
  bind_rows()
```

```
##   num_per_sample      mean          sd    median
## 1             10 0.2855383 0.050326646 0.2822082
## 2             30 0.2864496 0.028796517 0.2864116
## 3            100 0.2856003 0.016149171 0.2850690
## 4           1000 0.2859725 0.005090558 0.2859964
## 5           1000 0.2854758 0.005164720 0.2853403
```