Demonstration of Central Limit Theorem

David Gerard

2017-10-02

Learning Objectives

- Sample Means/Sample Proportions converge to Normal Distribution.
- Section 3.4.2, 3.4.3, 4.4 of DBC

Means of of Bernoulli's

Recall: Bernoulli distribution

Recall that X is Bernoulli if its pmf is

$$f_X(X) = \begin{cases} p^x(1-p)^{1-x} & \text{if } x \in \{0,1\} \\ 0 & \text{otherwise,} \end{cases}$$

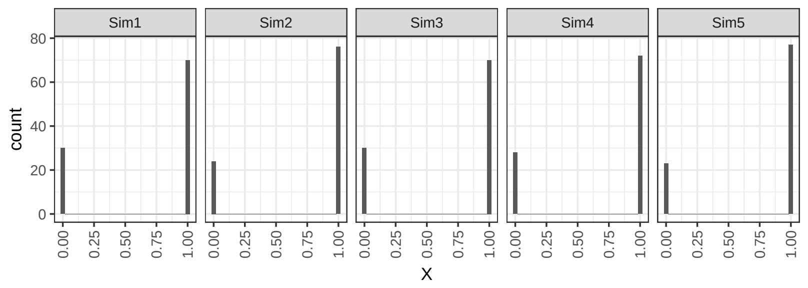
for some $p \in [0,1]$. That is, X is 1 with probability p and 0 with probability $1-p$.

E.g., have a box with six 1's and two 0's and we draw a number, then $p = 6/8 = 3/4$.

Sample

Suppose we sample 100 numbers from this box with six 1's and two 0's *with* replacement. We can do this multiple times (say 5000):

```
p <- 3/4
samp <- replicate(n = 5000, sample(c(0, 1), 100, TRUE,
c(1 - p, p)))
```



CTL Visual

Now draw 5000 samples of size 900, 2500. Compute the means

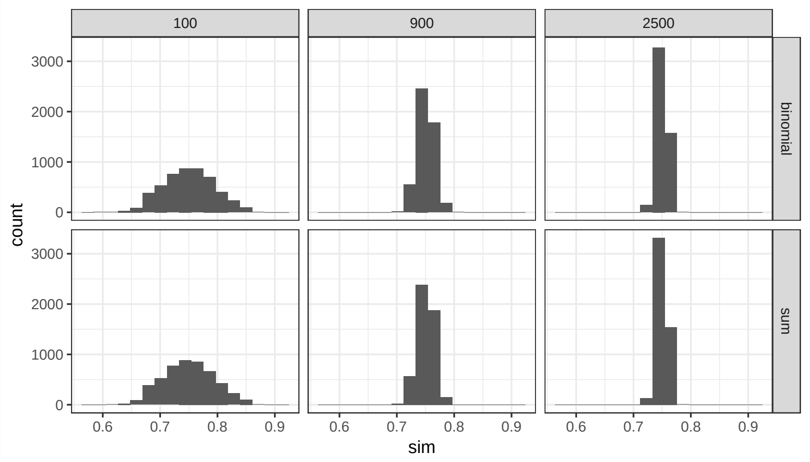
```
samp900 <- replicate(n = 5000, sample(c(0, 1), 900, TRUE, c(1 - p, p)))
samp2500 <- replicate(n = 5000, sample(c(0, 1), 2500, TRUE, c(1 - p, p)))

sum100 <- colSums(samp)
sum900 <- colSums(samp900)
sum2500 <- colSums(samp2500)
```

Same as drawing from a binomial

```
b100 <- rbinom(5000, 100, 3/4)
b900 <- rbinom(5000, 900, 3/4)
b2500 <- rbinom(5000, 2500, 3/4)
```

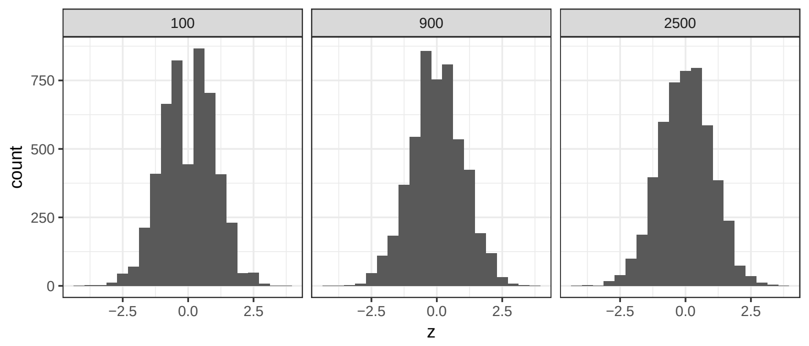
Dividing sums or binomials by number of samples (100, 900, or 2500), we get the following histograms:



Look the same because they are from the same distribution.

Center and Scale

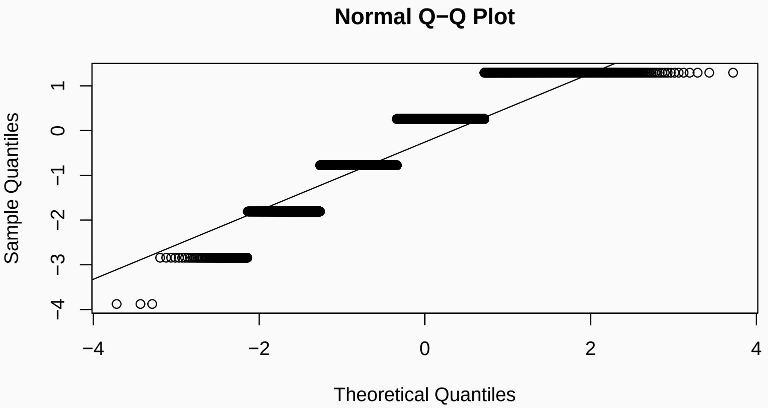
Let's subtract the means (np for $n = 100, 900, 2500$) and divide by the standard deviations ($n\sqrt{p(1-p)}$) and replot the histograms



Now the histograms all have the same spread and are centered at zero, but they are looking more and more like the normal distribution.

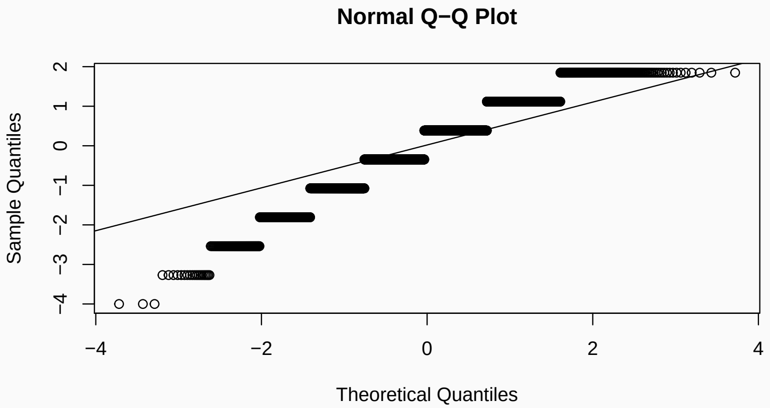
$n = 5$

```r
x <- scale(rbinom(n = 5000, size = 5, prob = p)) qqnorm(x)
qqline(x)
```



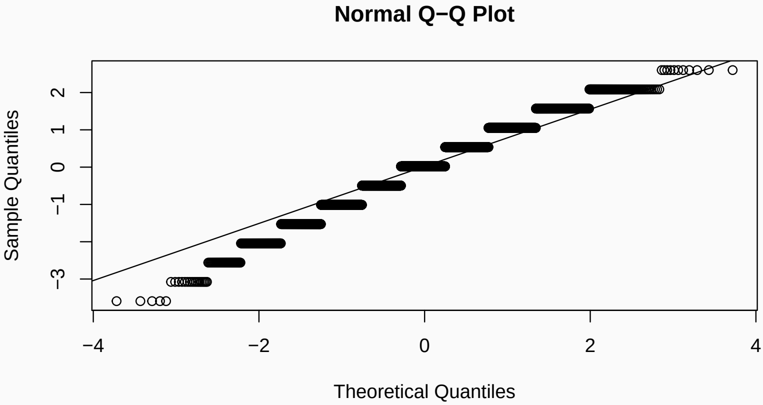
$n = 10$

```
x <- scale(rbinom(n = 5000, size = 10, prob = p)) qqnorm(x)
qqline(x)
```



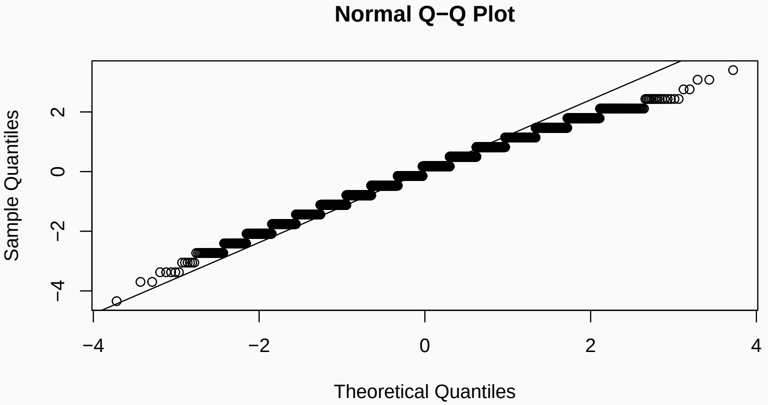
$n = 20$

```
x <- scale(rbinom(n = 5000, size = 20, prob = p)) qqnorm(x)
qqline(x)
```



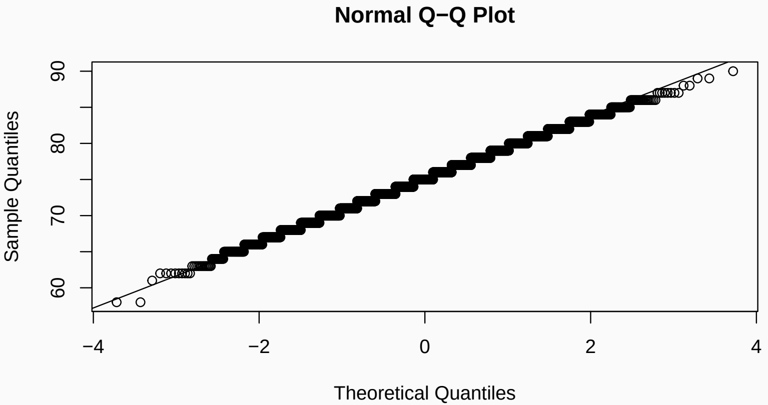
$n = 50$

```
x <- scale(rbinom(n = 5000, size = 50, prob = p)) qqnorm(x)
qqline(x)
```



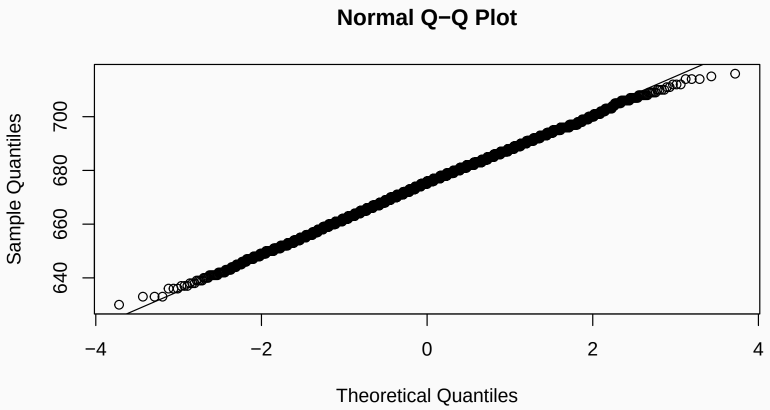
$n = 100$

```
qqnorm(sum100) qqline(sum100)
```



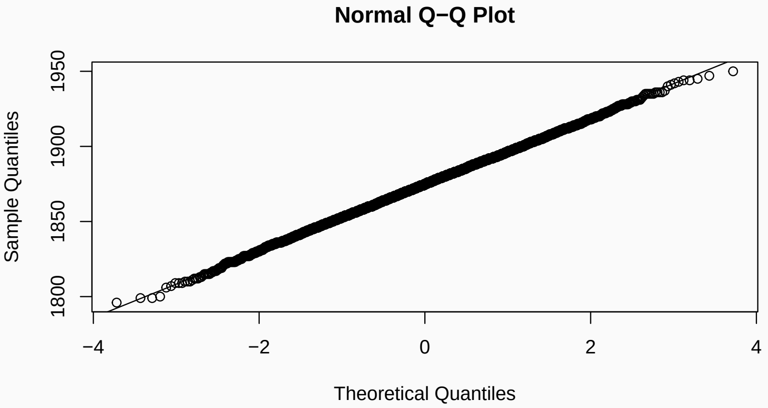
$n = 900$

```r
qqnorm(sum900) qqline(sum900)
```



$n = 2500$

```
qqnorm(sum2500) qqline(sum2500)
```



Central Limit Theorem

That sums/means of a large number of random variables are well approximated by the normal distribution is a general result that we will prove using the chalk board.