

# STAT 515 Lec 09 slides

## Sampling distributions and the Central Limit Theorem

Karl Gregory

University of South Carolina

These slides are an instructional aid; their sole purpose is to display, during the lecture, definitions, plots, results, etc. which take too much time to write by hand on the blackboard. They are not intended to explain or expound on any material.

## Random sample

A collection of independent rvs with the same distribution is a *random sample*.

- Often denote by  $X_1, \dots, X_n$ , where  $n$  is the *sample size*.
- In random sample,  $X_1, \dots, X_n$  are *iid*: independent and identically distributed.
- Common distribution of  $X_1, \dots, X_n$  called the *population distribution*.
- Can write  $X_1, \dots, X_n \stackrel{\text{ind}}{\sim} F$  if a rs from a distribution  $F$ .

Goal is to learn from  $X_1, \dots, X_n$  about the population distribution.

## Expected value and variance of the sample mean

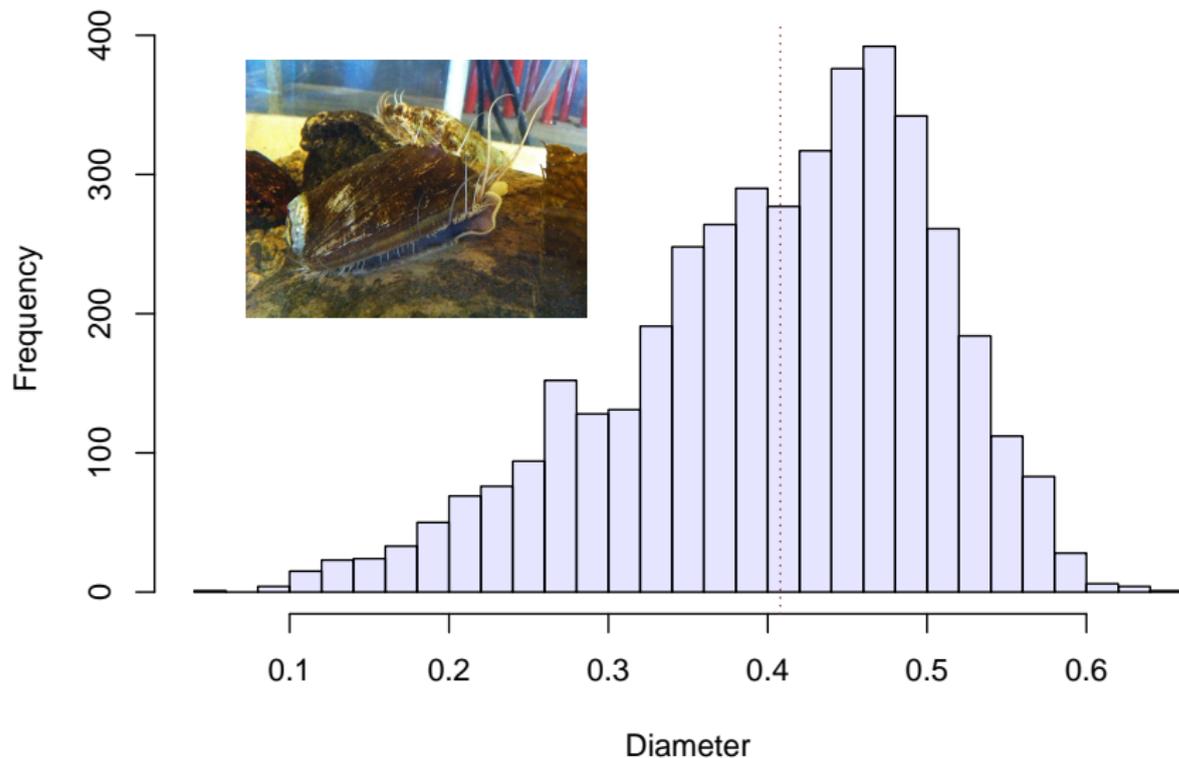
Let  $X_1, \dots, X_n$  be a rs from a population with mean  $\mu$  and  $\sigma^2$ . Then

$$\mathbb{E}\bar{X}_n = \mu \quad \text{and} \quad \text{Var } \bar{X}_n = \frac{\sigma^2}{n}.$$

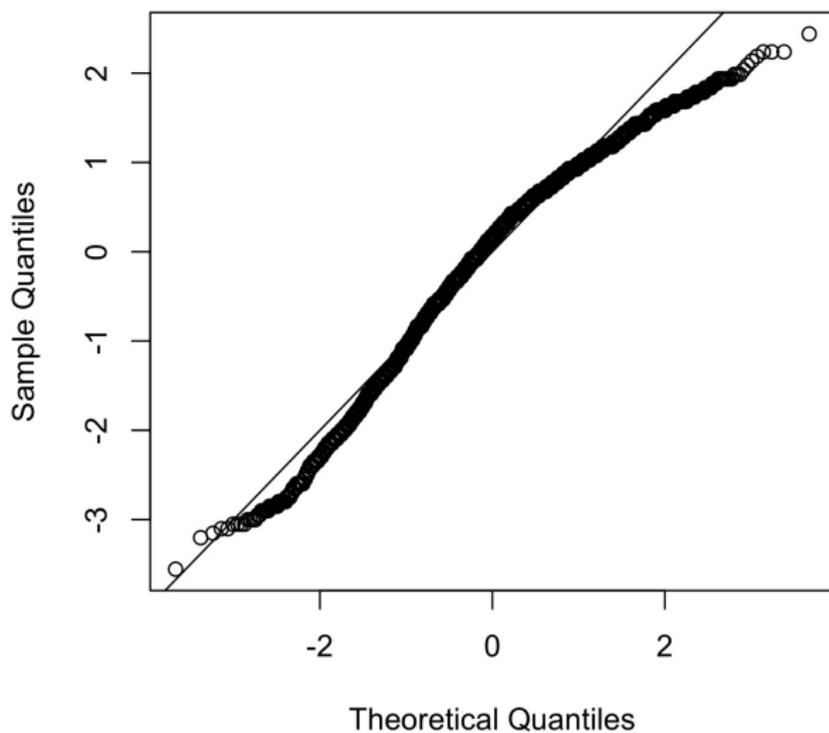
### Examples:

- 1 If  $X_1, \dots, X_n \stackrel{\text{ind}}{\sim} \text{Normal}(\mu, \sigma^2)$ , then  $\mathbb{E}\bar{X}_n = \mu$  and  $\text{Var } \bar{X}_n = \sigma^2/n$ .
- 2 If  $X_1, \dots, X_n \stackrel{\text{ind}}{\sim} \text{Bernoulli}(p)$ , then  $\mathbb{E}\bar{X}_n = p$  and  $\text{Var } \bar{X}_n = p(1-p)/n$ .
- 3 If  $X_1, \dots, X_n \stackrel{\text{ind}}{\sim} \text{Poisson}(\lambda)$ , then  $\mathbb{E}\bar{X}_n = \lambda$  and  $\text{Var } \bar{X}_n = \lambda/n$ .
- 4 If  $X_1, \dots, X_n \stackrel{\text{ind}}{\sim} \text{Exponential}(\lambda)$ , then  $\mathbb{E}\bar{X}_n = 1/\lambda$  and  $\text{Var } \bar{X}_n = 1/(n\lambda^2)$ .

Consider the diameters of 4,176 abalones with mean 0.4078915. [link to data](#)

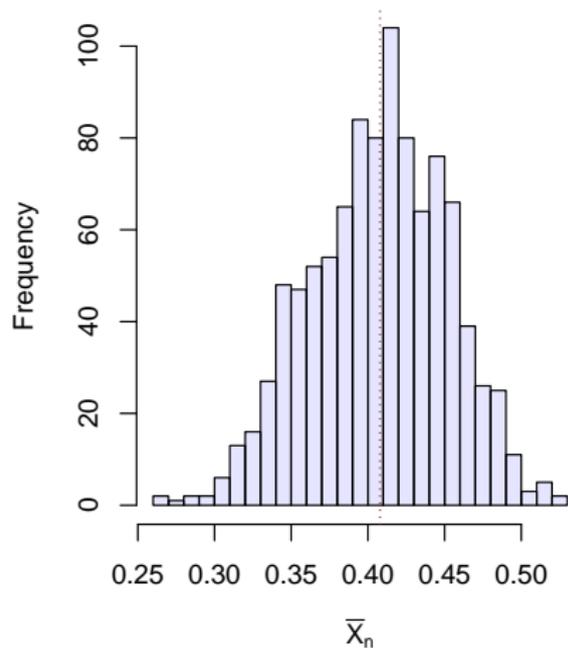
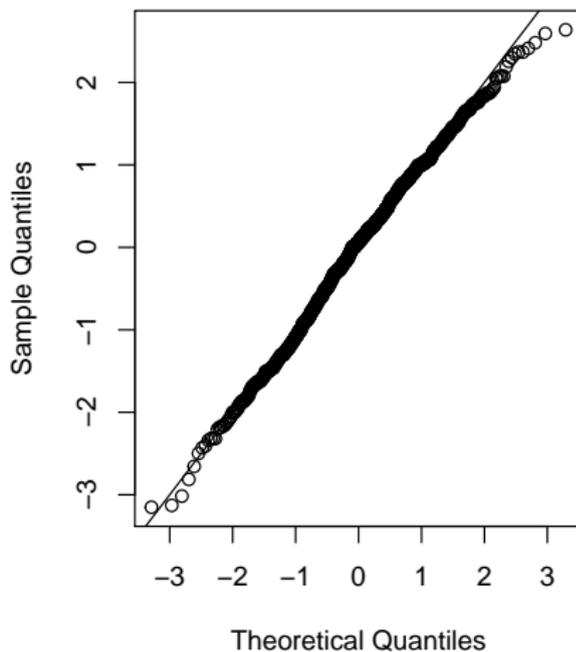


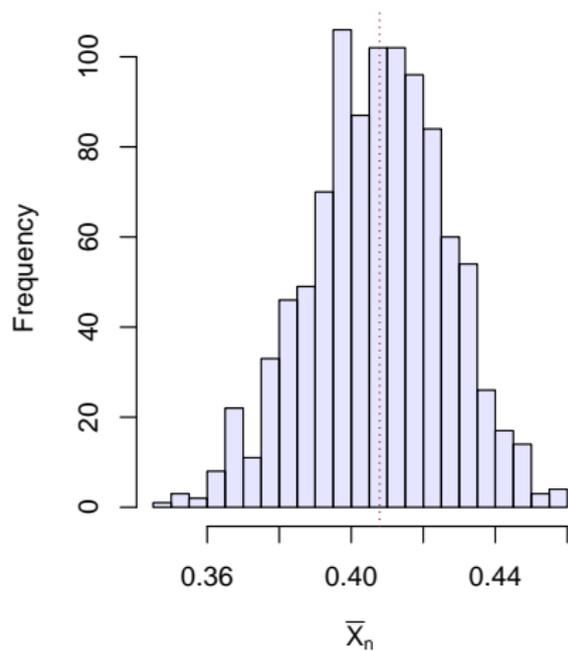
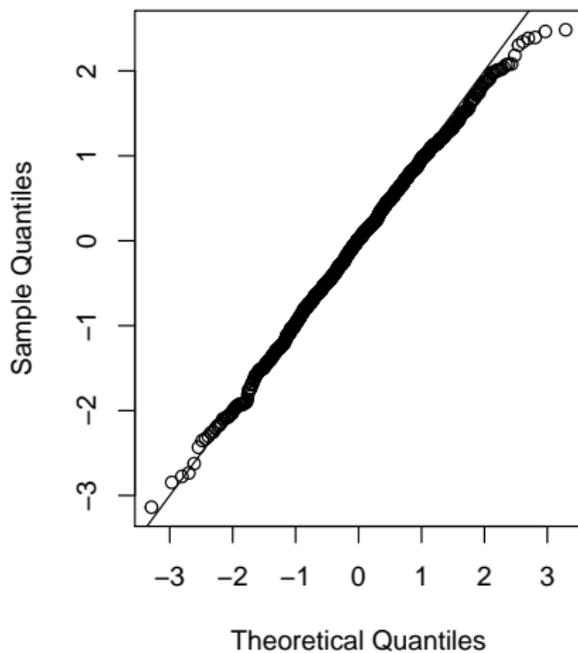
## Normal Q-Q plot of abalone diameters

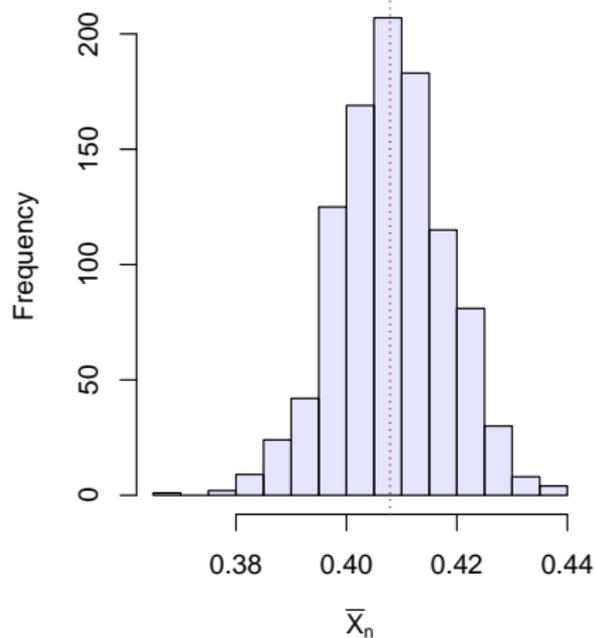
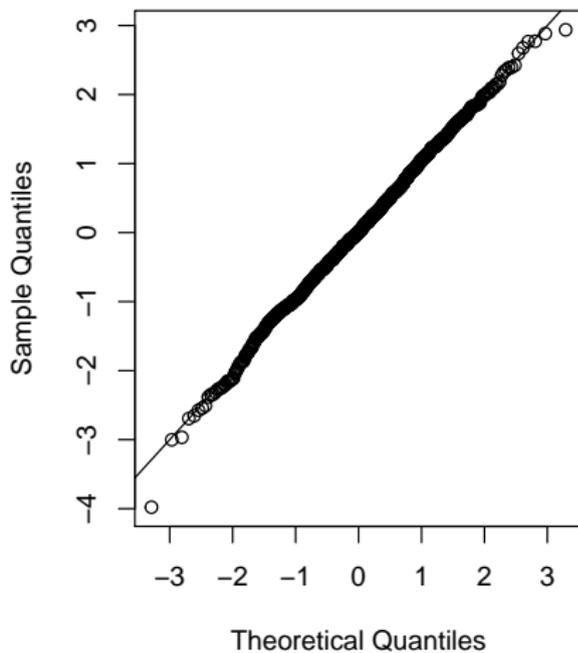


**Exercise:** Treat the 4,176 abalone as a population. The mean diameter is  $\mu = 0.408$ . Let  $\bar{X}_n$  be the mean diameter from a sample of abalone.

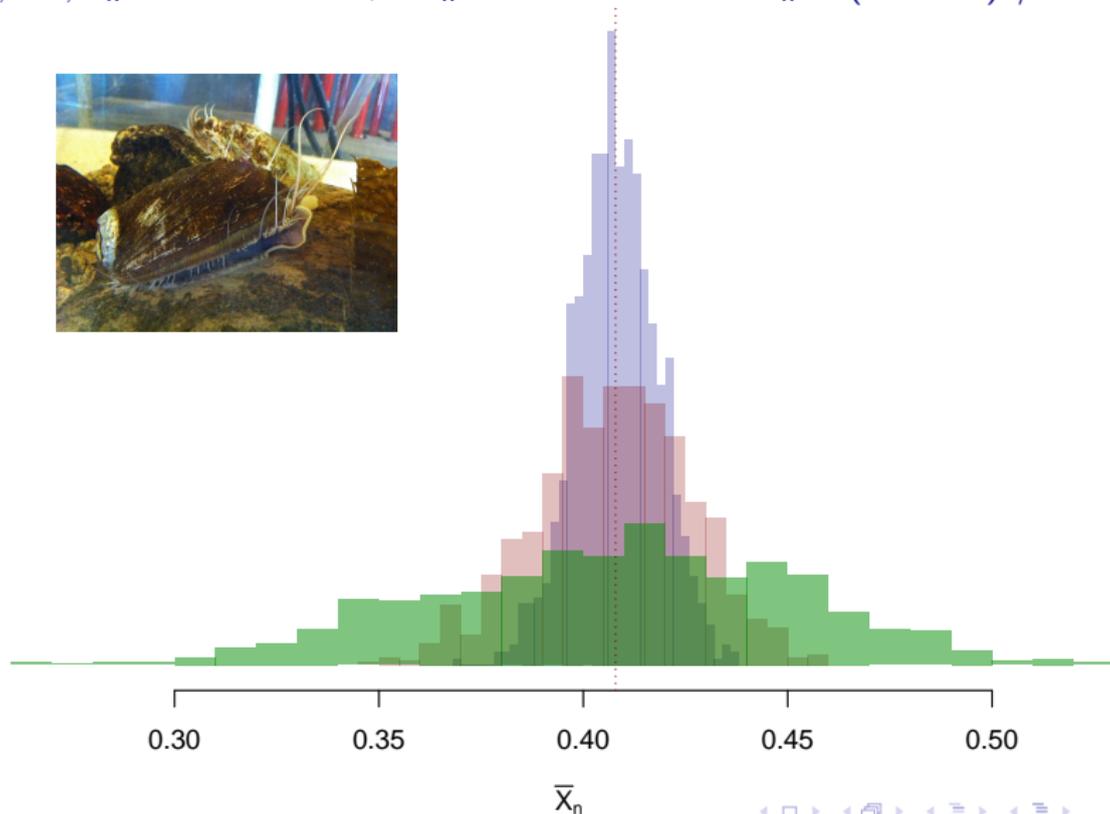
- 1 For the sample sizes  $n = 5, 25, 100$ , draw 1,000 samples and
  - 1 Make a histogram of the  $\bar{X}_n$  values.
  - 2 Make a Normal Q-Q plot of the  $\bar{X}_n$  values.
- 2 Around what value are the values of  $\bar{X}_n$  centered?
- 3 What changes as  $n$  changes?

Histogram of  $\bar{X}_n$  with  $n = 5$ Normal Q-Q plot of  $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ 

Histogram of  $\bar{X}_n$  with  $n = 25$ Normal Q-Q plot of  $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ 

Histogram of  $\bar{X}_n$  with  $n = 100$ Normal Q-Q plot of  $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ 

If  $X_1, \dots, X_n$  a rs of abalone,  $\mathbb{E}\bar{X}_n = 0.4079$  and  $\text{Var}\bar{X}_n = (0.09924)^2/n$ .



## Distribution of sample mean when population is Normal

Let  $X_1, \dots, X_n \stackrel{\text{ind}}{\sim} \text{Normal}(\mu, \sigma^2)$ . Then  $\bar{X}_n \sim \text{Normal}(\mu, \sigma^2/n)$ .

Can use this to get probabilities like  $P(a < \bar{X}_n < b)$  as follows:

- 1 Transform  $a$  and  $b$  to the  $Z$ -world (# of standard deviations world):

$$a \mapsto \frac{a - \mu}{\sigma/\sqrt{n}} \quad \text{and} \quad b \mapsto \frac{b - \mu}{\sigma/\sqrt{n}},$$

- 2 Find

$$P\left(\frac{a - \mu}{\sigma/\sqrt{n}} < Z < \frac{b - \mu}{\sigma/\sqrt{n}}\right).$$

**Exercise:** Suppose the number of minutes USC students spend talking on phone each month has the  $\text{Normal}(\mu = 450, \sigma^2 = 50^2)$  distribution.

- 1 Let  $X$  be the talk time of one randomly selected student.
  - a. Find  $P(X < 425)$ .
  - b. Find  $P(|X - 450| > 50)$ .
  
- 2 Now let  $\bar{X}_n$  be the mean talk time from  $n = 9$  randomly selected students.
  - a. Find  $P(\bar{X}_n < 425)$ .
  - b. Find  $P(|\bar{X}_n - 450| > 50)$ .

## Central Limit Theorem

Let  $X_1, \dots, X_n$  be a rs from a dist. with mean  $\mu$  and variance  $\sigma^2 < \infty$ . Then

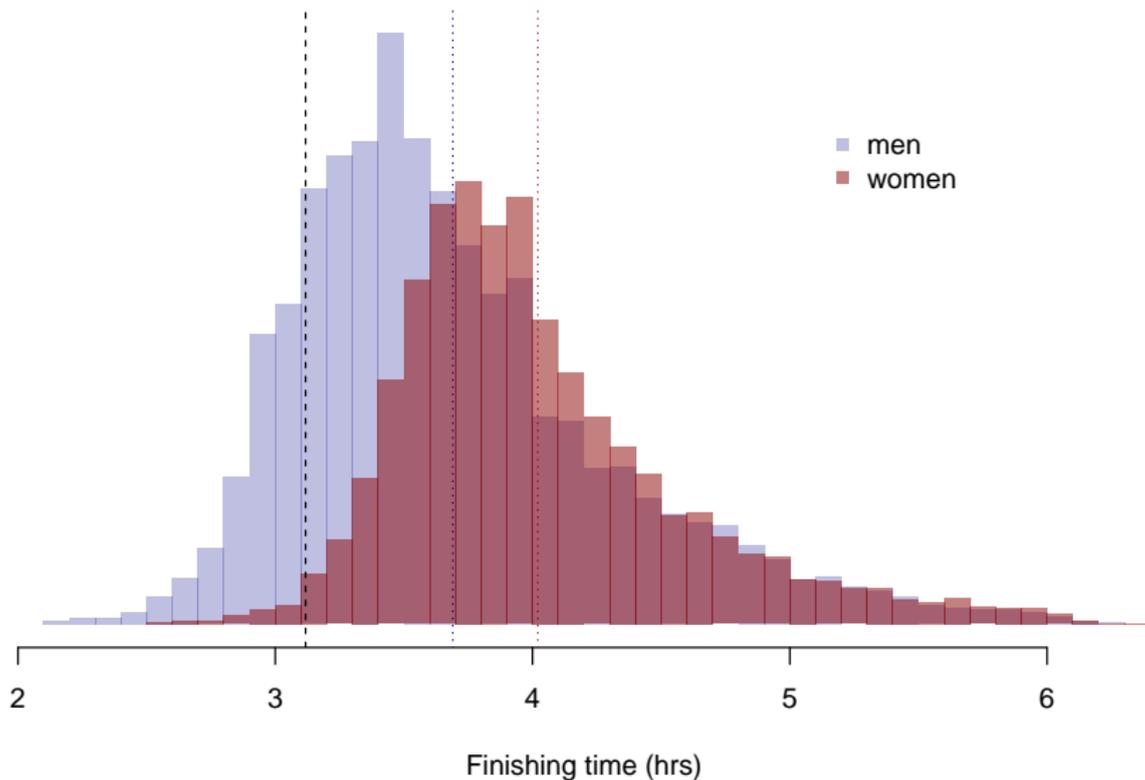
$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \text{ behaves more and more like } Z \sim \text{Normal}(0, 1)$$

for larger and larger  $n$ .

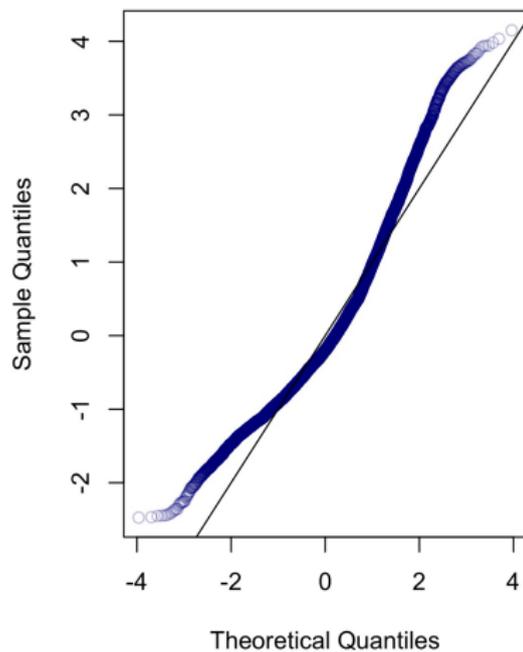
This means that for large  $n$  (say  $n \geq 30$ ), we have

$$\bar{X}_n \overset{\text{approx}}{\sim} \text{Normal} \left( \mu, \frac{\sigma^2}{n} \right).$$

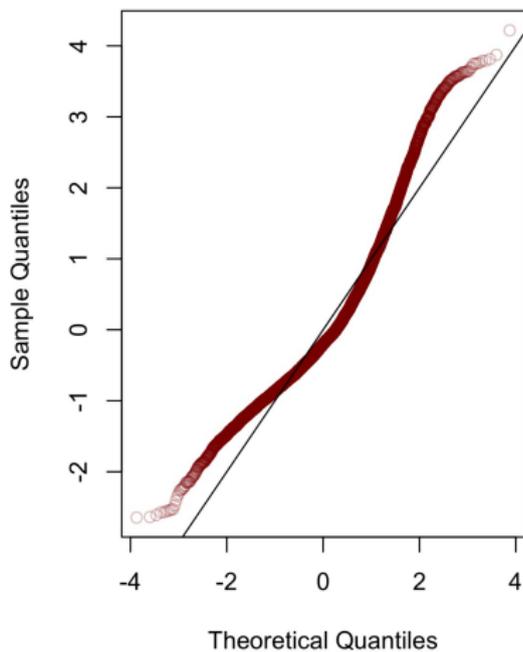
## 2009 Boston Marathon finishing times (hrs)



Normal Q-Q plot for men



Normal Q-Q plot for women



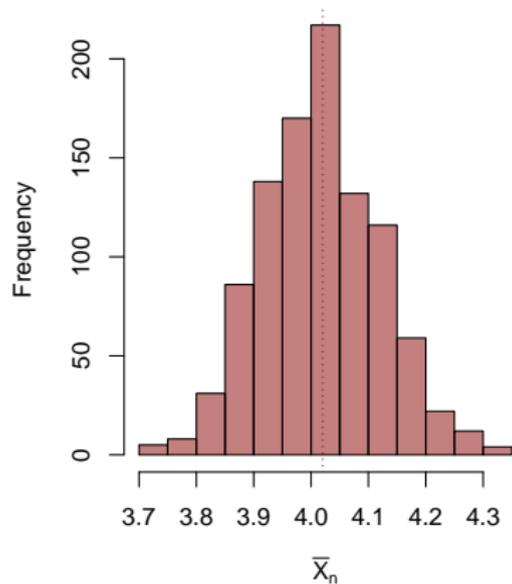
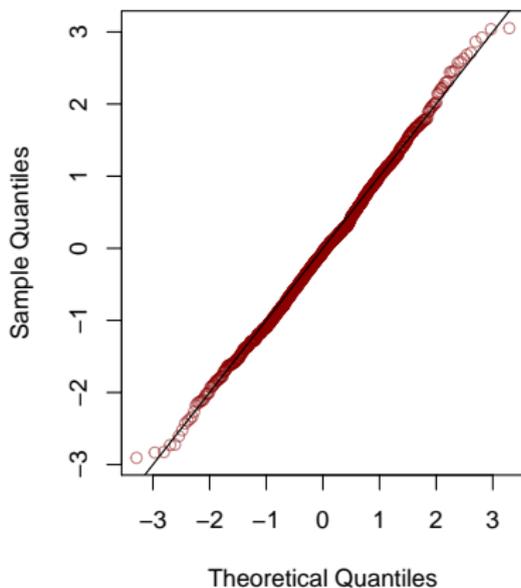
**Exercise:** Women's finishing times for the 2009 Boston Marathon had mean 4.02 hours and standard deviation 0.555 hours.

Consider sampling  $n = 30$  women and let  $\bar{X}_n$  be the mean of their finishing times.

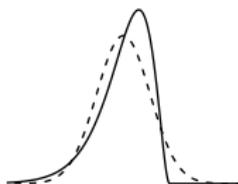
- 1 Find an approximation to  $P(\bar{X}_n < 3.90)$ .
- 2 Find an approximation to  $P(\bar{X}_n > 4.25)$ .
- 3 Find an approximation to  $P(|\bar{X}_n - 4.02| < 0.2)$ .

Now use R to draw 1,000 samples of size  $n = 30$ . [link to women's data](#).

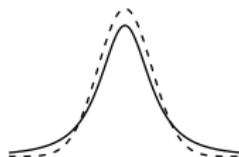
- 1 Make histogram and Normal Q-Q plot of  $\bar{X}_n$ .
- 2 Get the probabilities above using the output of the simulation.

Histogram of  $\bar{X}_n$  with  $n = 30$ Normal Q-Q plot of  $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ 

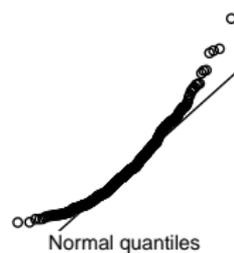
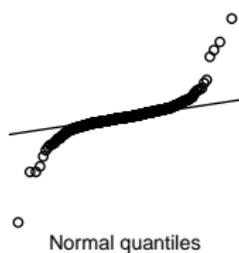
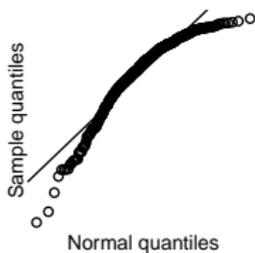
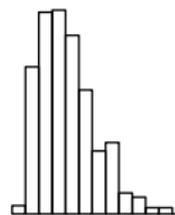
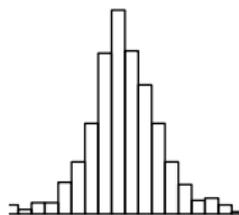
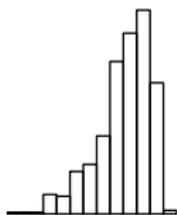
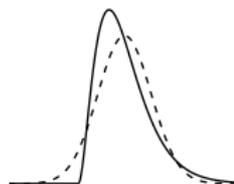
left-skewed



heavy-tailed



right-skewed



We can apply the Central Limit theorem to proportions. . .

## Central Limit Theorem for the sample proportion

Let  $X_1, \dots, X_n \stackrel{\text{ind}}{\sim} \text{Bernoulli}(p)$  and let  $\hat{p}_n = \bar{X}_n$ . Then

$$\frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}} \text{ behaves more and more like } Z \sim \text{Normal}(0, 1)$$

for larger and larger  $n$ .

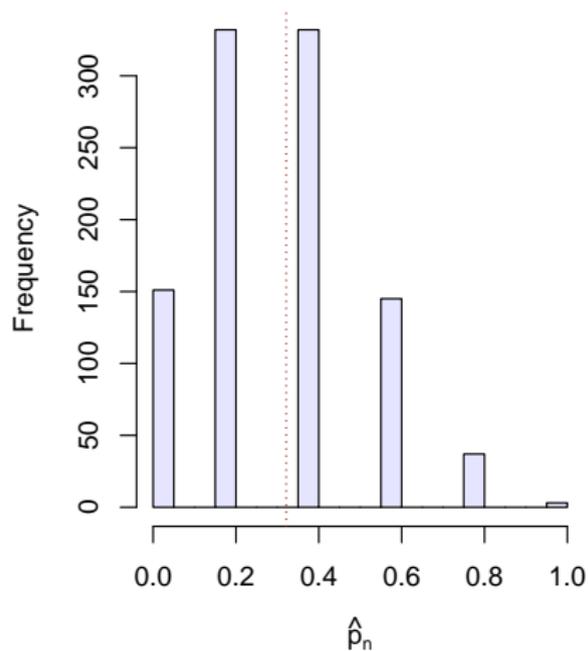
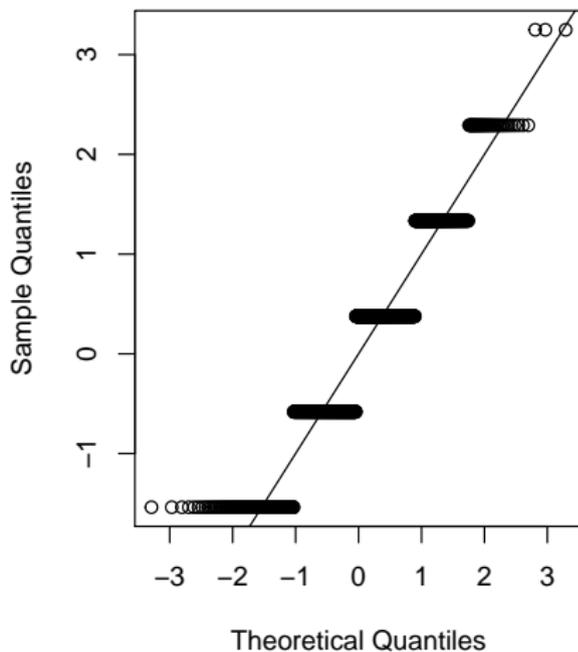
This means that for large  $n$  (say  $np \geq 5$  and  $n(1-p) \geq 5$ ), we have

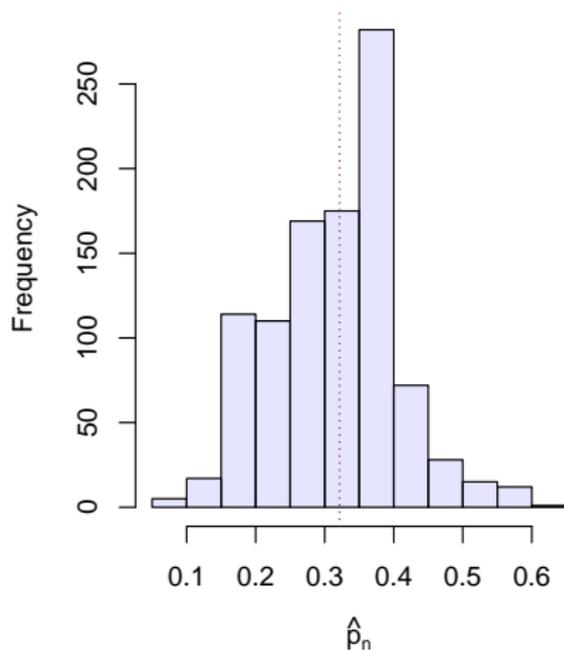
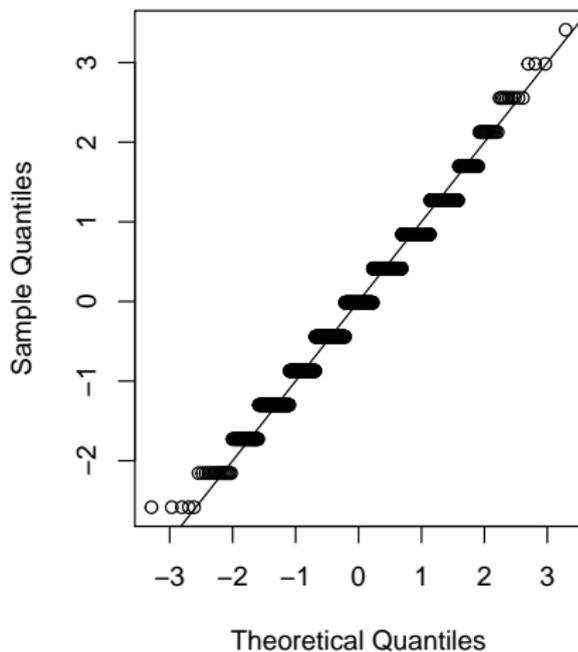
$$\hat{p}_n \stackrel{\text{approx}}{\sim} \text{Normal}\left(p, \frac{p(1-p)}{n}\right).$$

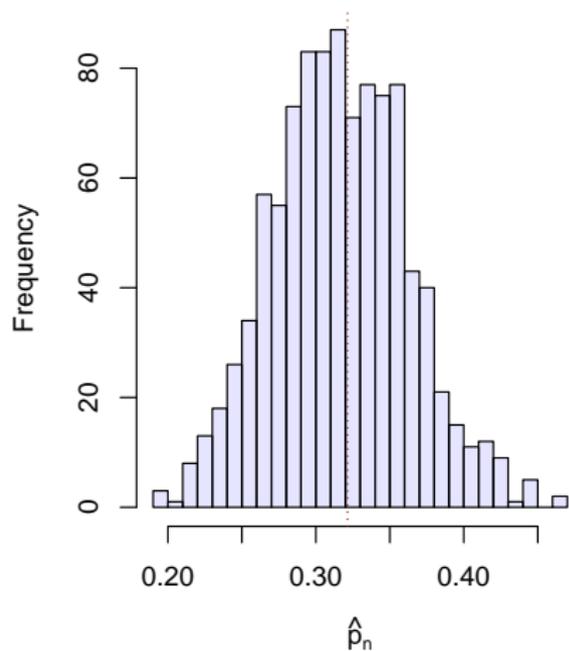
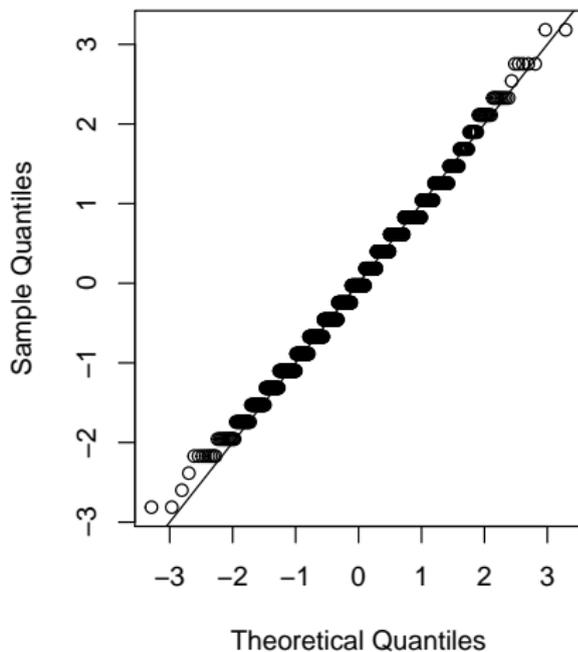
Also:  $\sum_{i=1}^n X_i = n\hat{p}_n \stackrel{\text{approx}}{\sim} \text{Normal}(p, np(1-p))$  for large  $n$ .

**Exercise:** Treat the 4,176 abalone as a population. The proportion classified as infants among the abalone is  $p = 0.321$ ; let  $\hat{p}_n$  represent the proportion of infants in a random sample of abalone.

- 1 For the sample sizes  $n = 5, 25, 100$ , draw 1,000 samples and
  - 1 Make a histogram of the  $\hat{p}_n$  values.
  - 2 Make a Normal Q-Q plot of the  $\hat{p}_n$ .
- 2 Around what value are the values of  $\hat{p}_n$  centered?
- 3 What changes as  $n$  changes?

Histogram of  $\hat{p}_n$  with  $n = 5$ Normal Q-Q plot of  $\sqrt{n}(\hat{p}_n - p) / \sqrt{p(1-p)}$ 

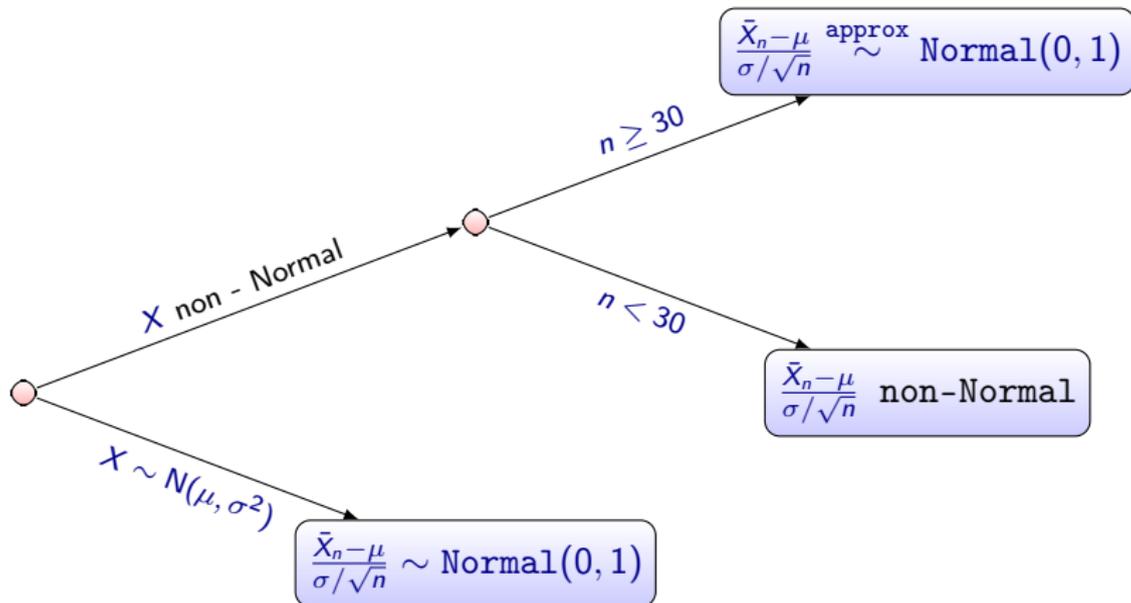
Histogram of  $\hat{p}_n$  with  $n = 25$ Normal Q-Q plot of  $\sqrt{n}(\hat{p}_n - p) / \sqrt{p(1-p)}$ 

Histogram of  $\hat{p}_n$  with  $n = 100$ Normal Q-Q plot of  $\sqrt{n}(\hat{p}_n - p) / \sqrt{p(1-p)}$ 

**Exercise:** Suppose 60% of USC undergraduates are registered to vote. Consider taking a sample of size  $n = 15$ . Let  $\hat{p}_n$  be the number in your sample who are registered to vote.

- 1 Find the approximate value of  $P(\hat{p}_n > 0.70)$  using the Normal distribution.
- 2 Find the exact value of  $P(\hat{p}_n > 0.70)$  using the Binomial distribution.
- 3 Find the approximate value of  $P(0.30 < \hat{p}_n < 0.80)$  using the Normal dist.
- 4 Find the exact value of  $P(0.30 < \hat{p}_n < 0.80)$  using the Binomial dist.
- 5 Repeat the above for a sample of size  $n = 100$ .

## Summary of sampling distribution results for $\bar{X}_n$ :



## Summary of sampling distribution results for $\hat{p}_n$ :

