STAT 515 Lec 09 slides

Sampling distributions and the Central Limit Theorem

Karl B. 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 variable

Random sample

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

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

independent

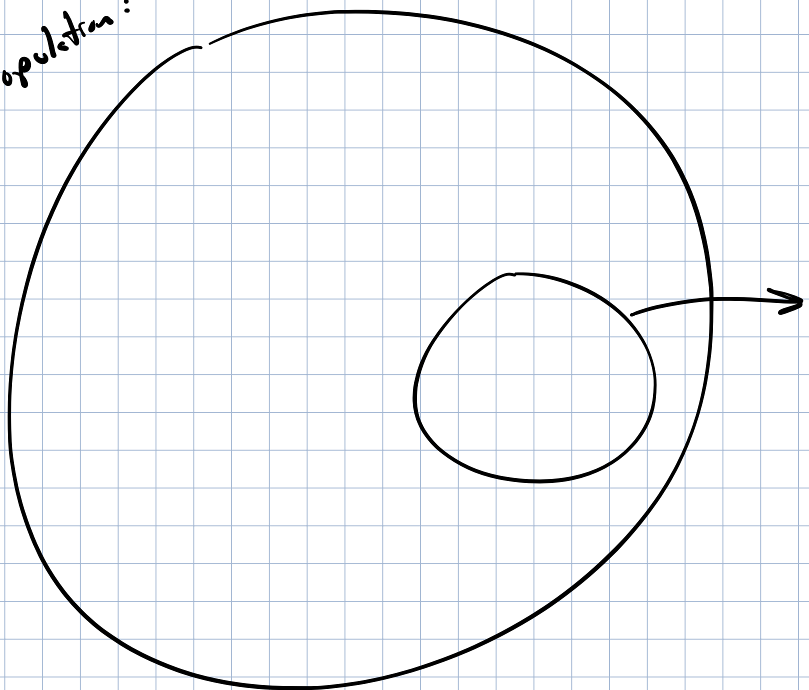
random sample

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

population mean: μ
population variance: σ^2

These are called "parameters"

population:

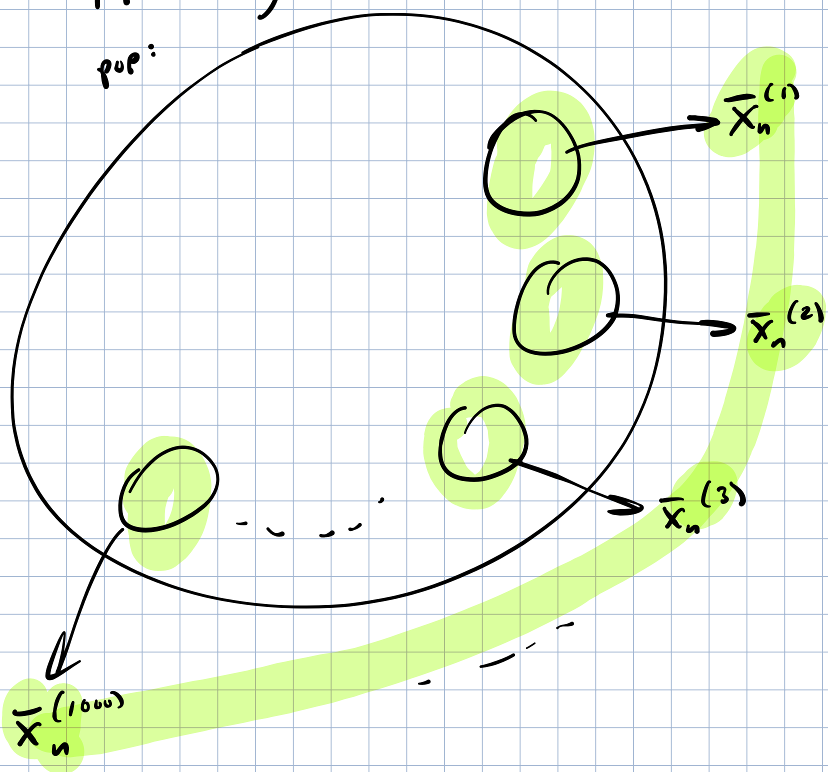


sample: $X_1, \dots, X_n$

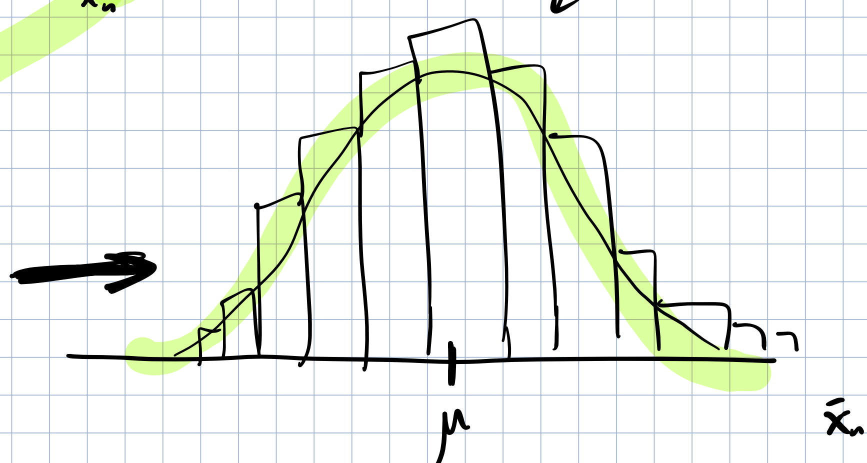
sample mean: $\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$

pop mean: μ

pop:



Histogram of the 1000 values of $\bar{X}_n$



Called the <u>sampling distribution</u> of $\bar{X}_n$.

Expected value and variance of the sample mean

Let $X_1, \dots, X_n$ be a rs from a population with mean μ and variance σ^2. Then

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

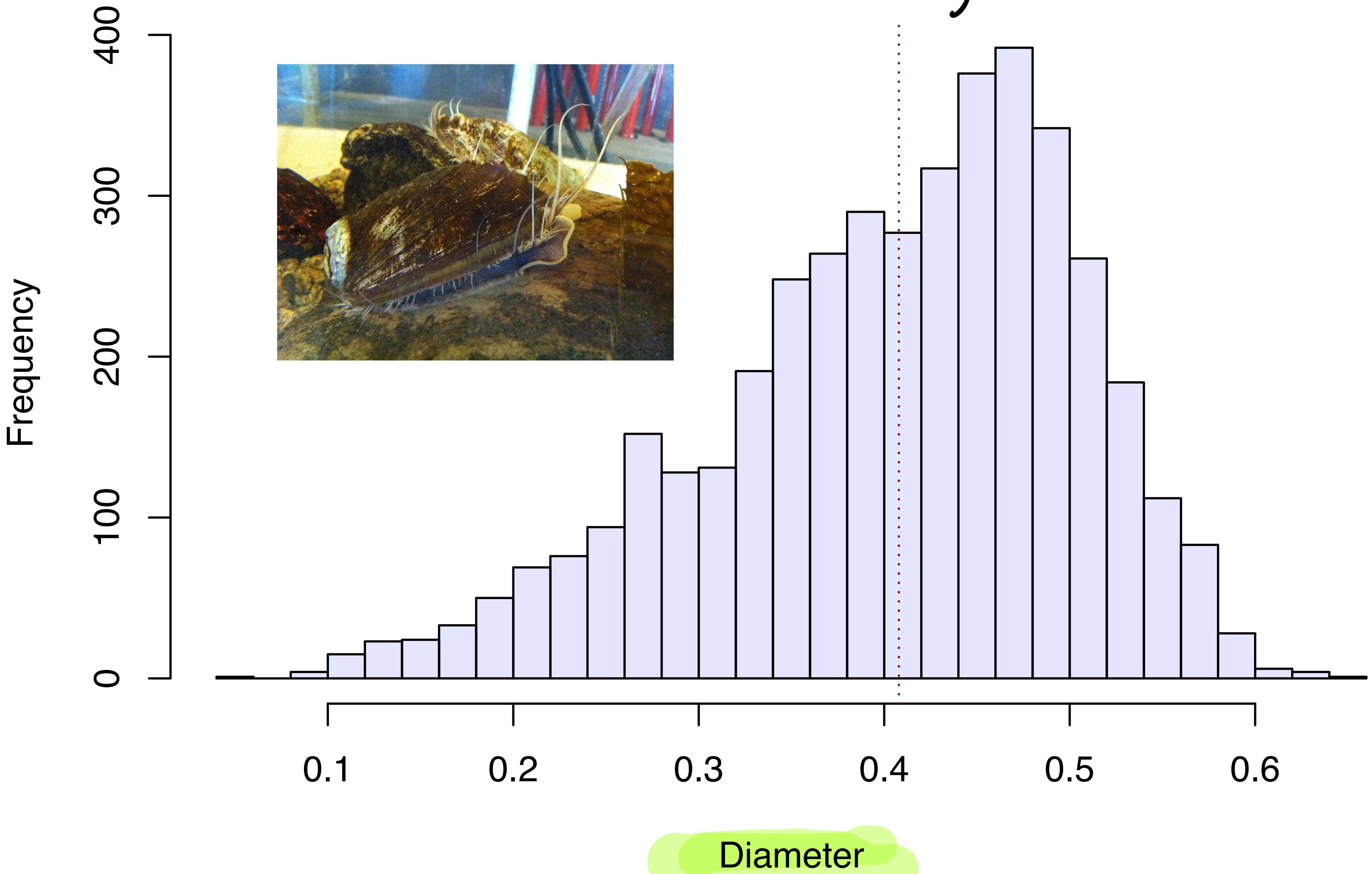
Examples:

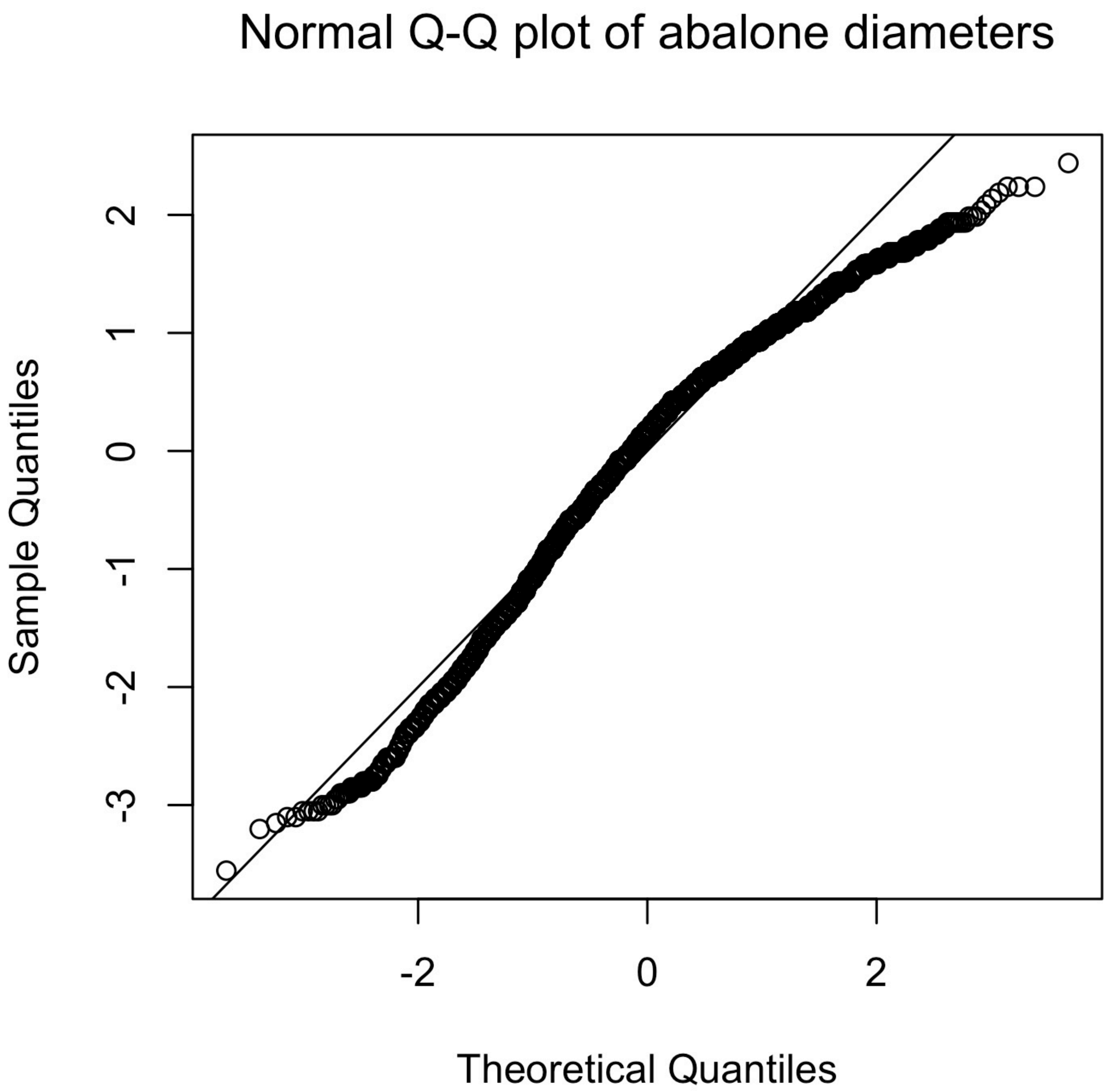
mean variance

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

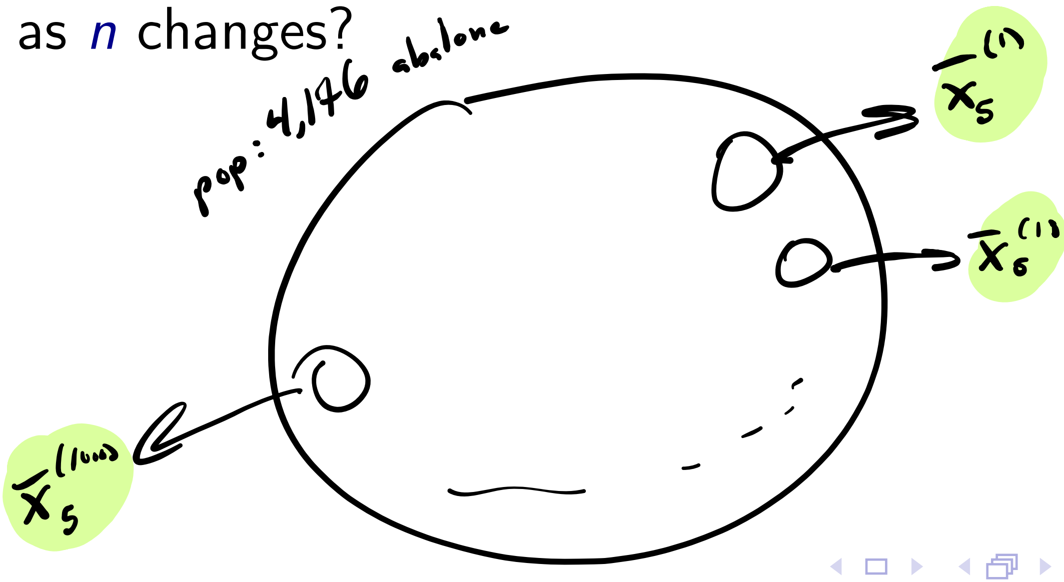
$\mu = 0.4078915$

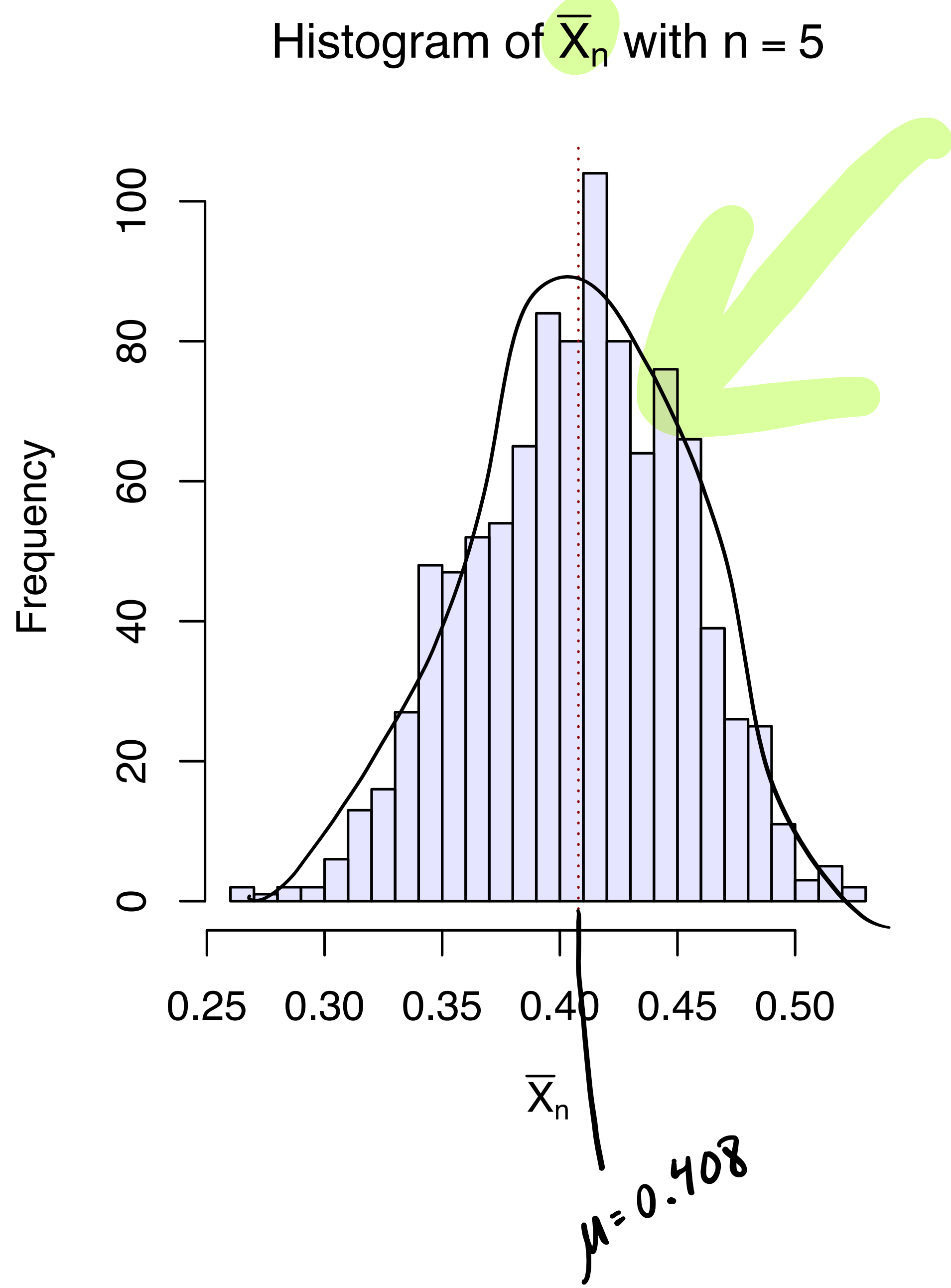


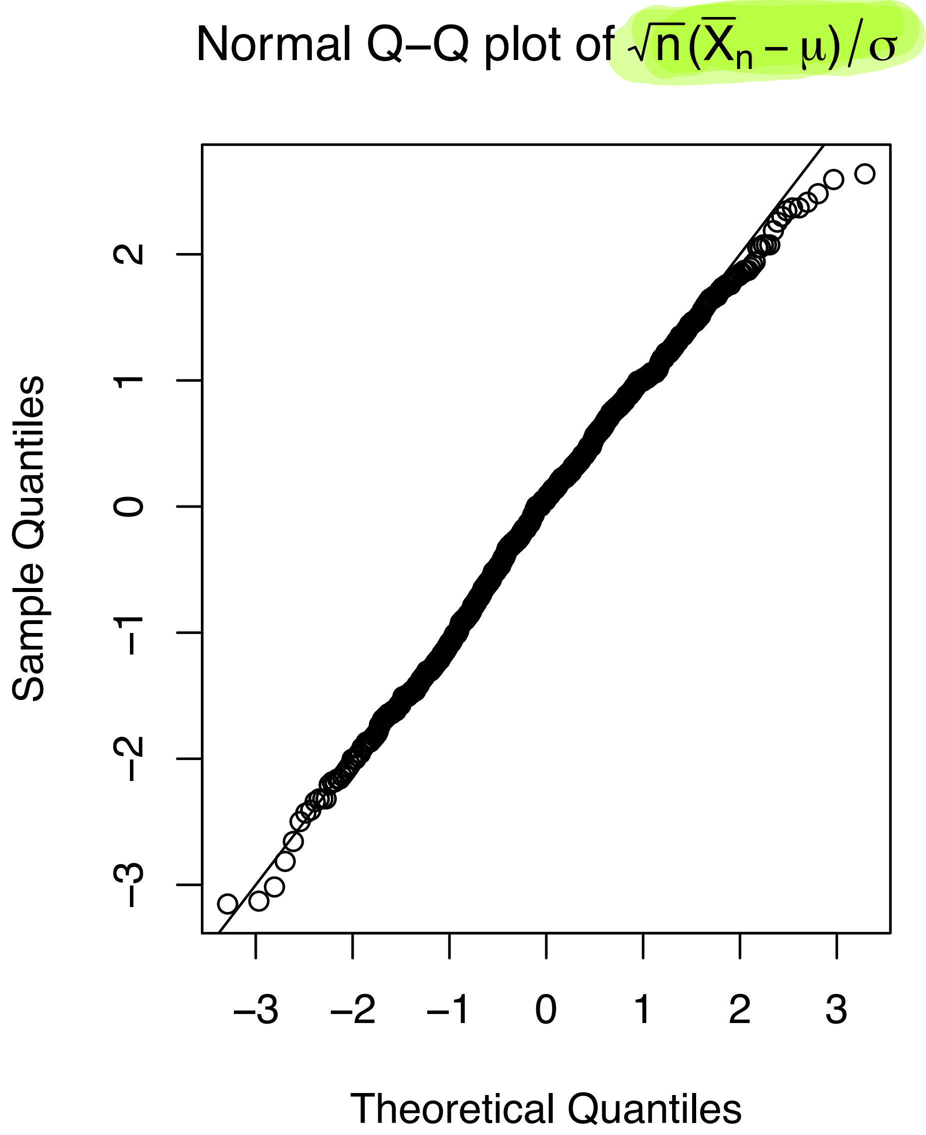

Normal Q-Q plot of abalone diameters
Sample Quantiles
-3
-2
-1
0
1
2
Theoretical Quantiles
-2
0
2

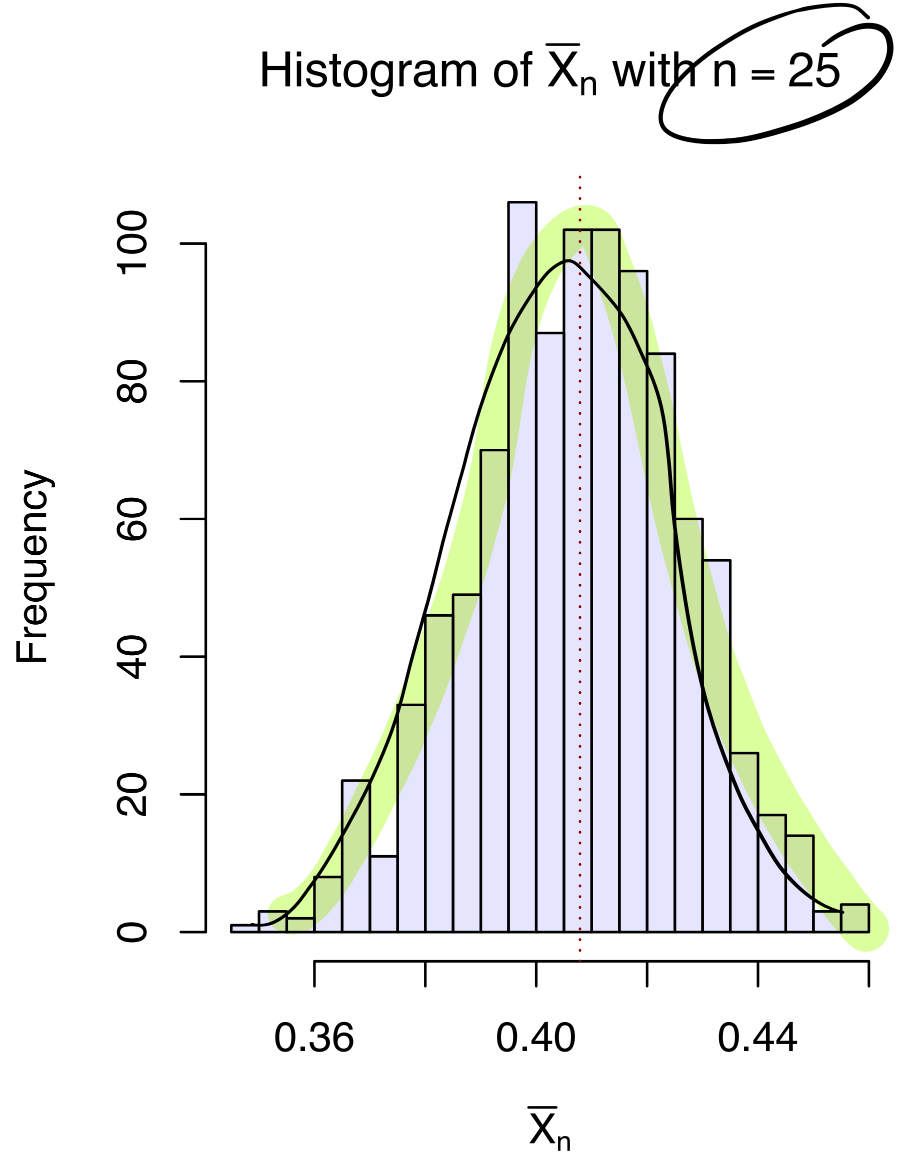
Exercise: Treat the 4,176 abalone as a population. The mean diameter is $\mu = 0.408$. Let $\bar{X}_n$ be the mean diameter from a sample of abalone.

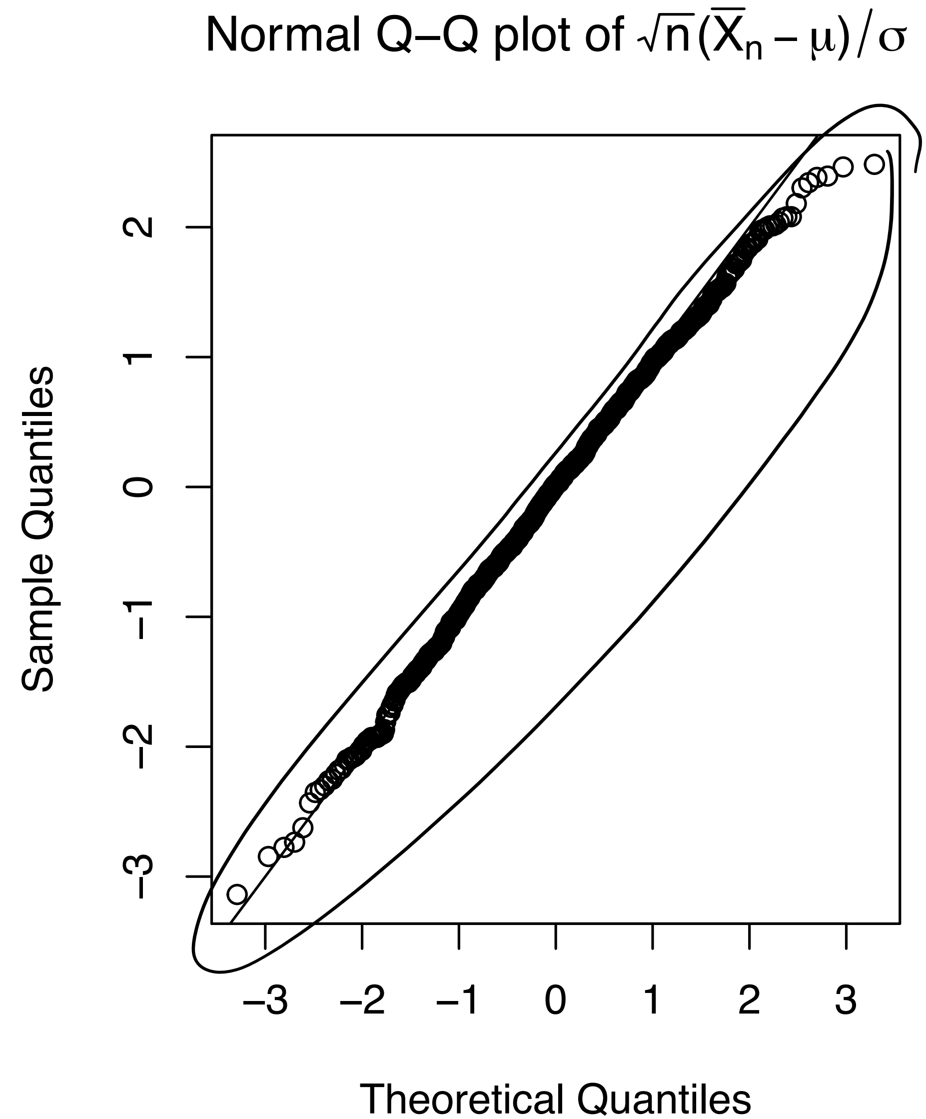
1. For the sample sizes $n = 5, 25, 100$, draw $1{,}000$ samples and
   1. Make a histogram of the $\bar{X}_n$ values.
   2. Make a Normal Q-Q plot of the $\bar{X}_n$ values.
2. Around what value are the values of $\bar{X}_n$ centered?
3. What changes as n changes?

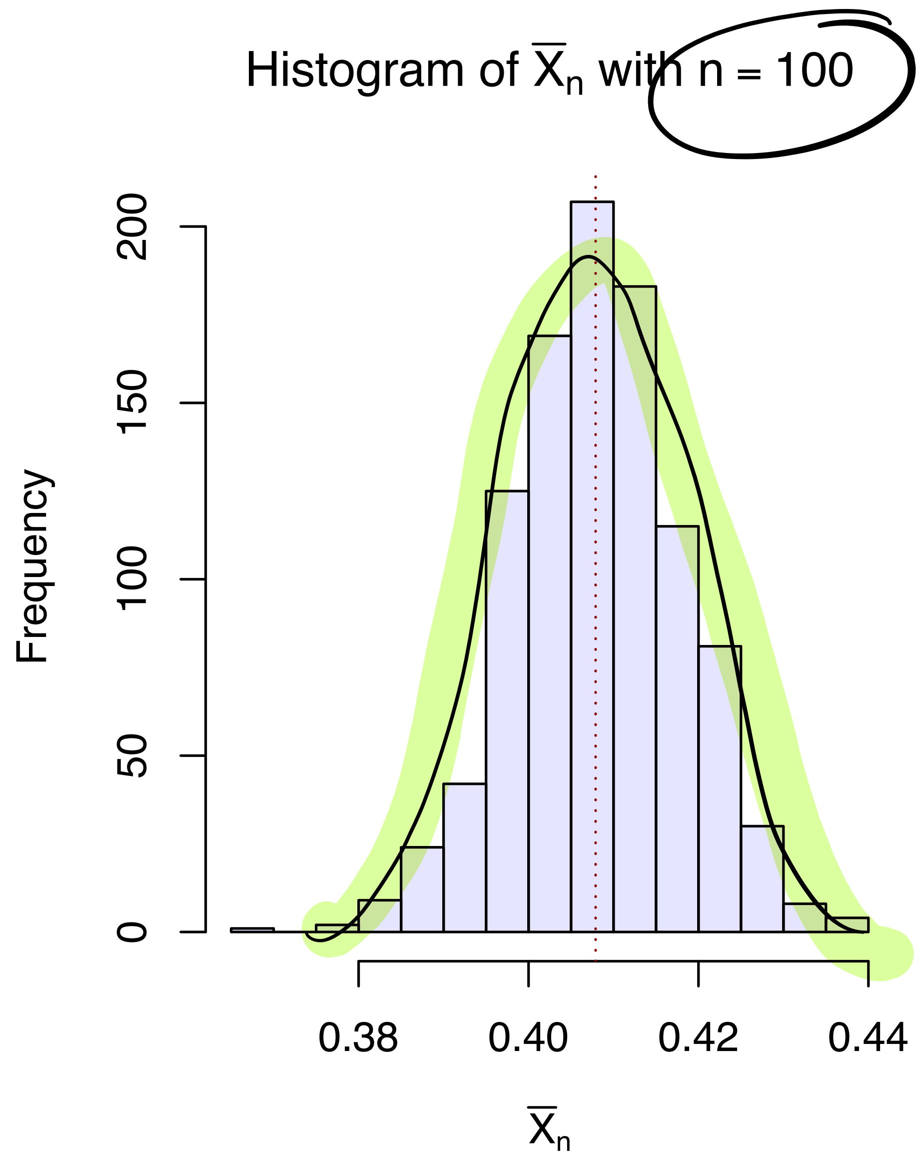


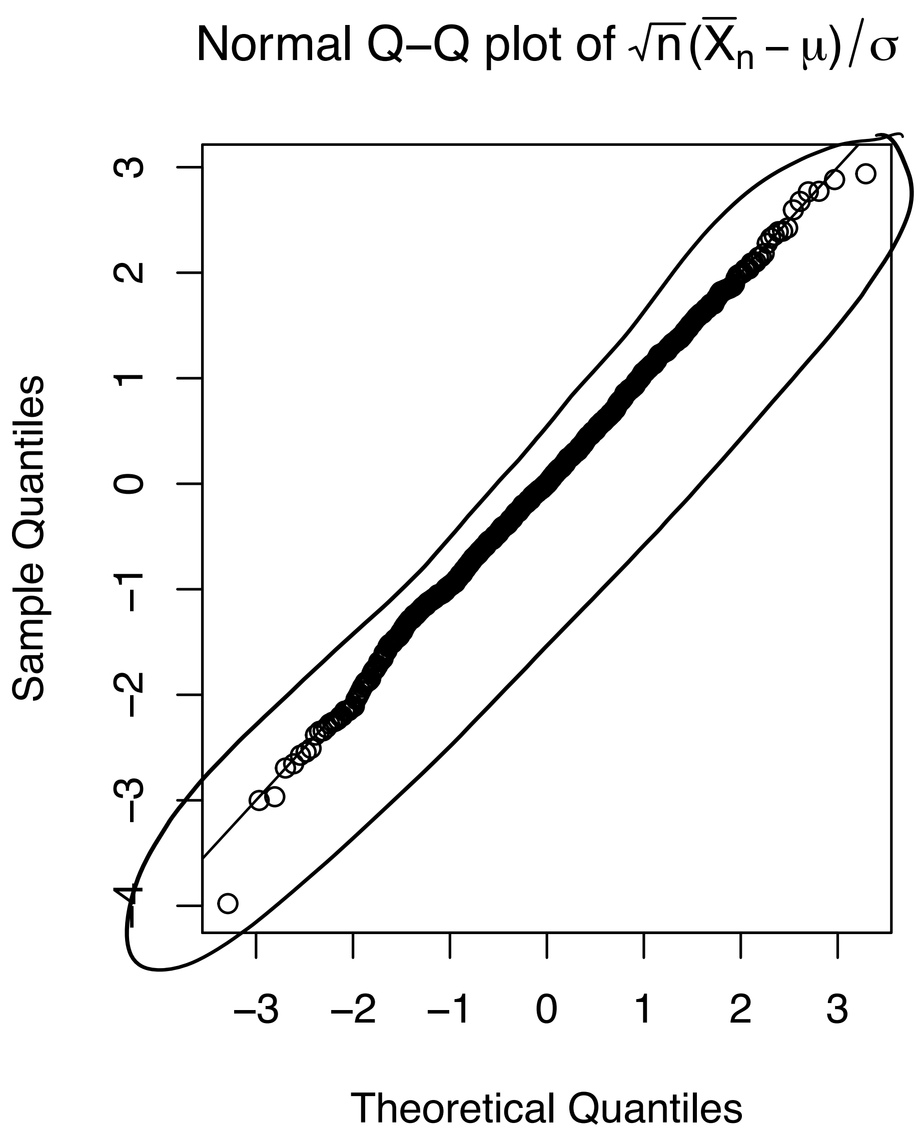

Histogram of $\bar{X}_n$ with n = 5
Frequency
0
20
40
60
80
100
0.25
0.30
0.35
0.40
0.45
0.50
$\bar{X}_n$
$\mu = 0.408$


Normal Q–Q plot of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$
Sample Quantiles
−3
−2
−1
0
1
2
3
Theoretical Quantiles

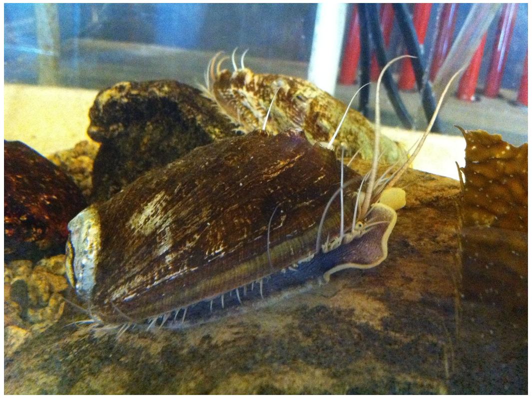

Histogram of $\overline{X}_n$ with n = 25
Frequency
0
20
40
60
80
100
0.36
0.40
0.44
$\overline{X}_n$

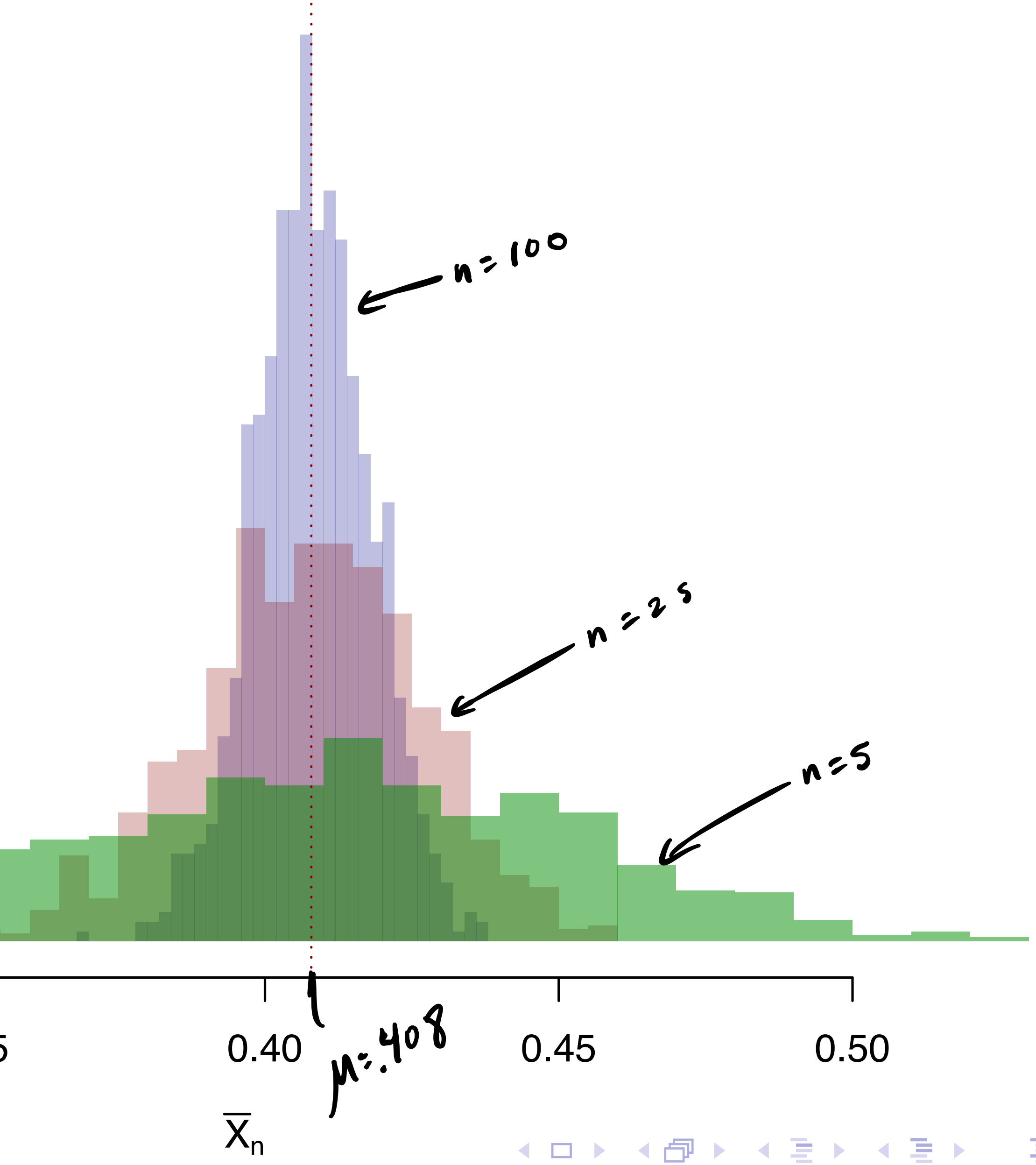

Normal Q–Q plot of $\sqrt{n}(\overline{X}_n - \mu)/\sigma$
Sample Quantiles
−3
−2
−1
0
1
2
Theoretical Quantiles


Histogram of $\bar{X}_n$ with n = 100
Frequency
0
50
100
150
200
0.38
0.40
0.42
0.44
$\bar{X}_n$


Normal Q–Q plot of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$
Sample Quantiles
Theoretical Quantiles
−4
−3
−2
−1
0
1
2
3

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



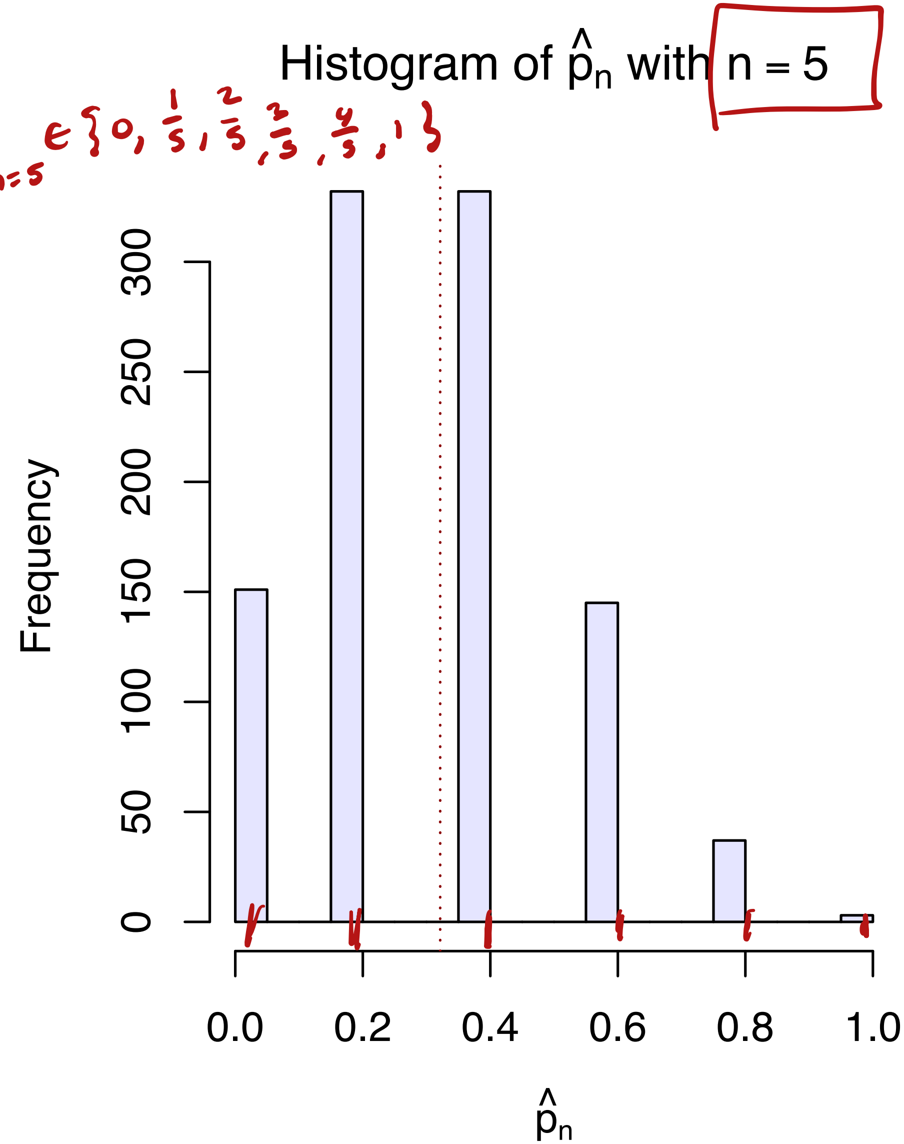


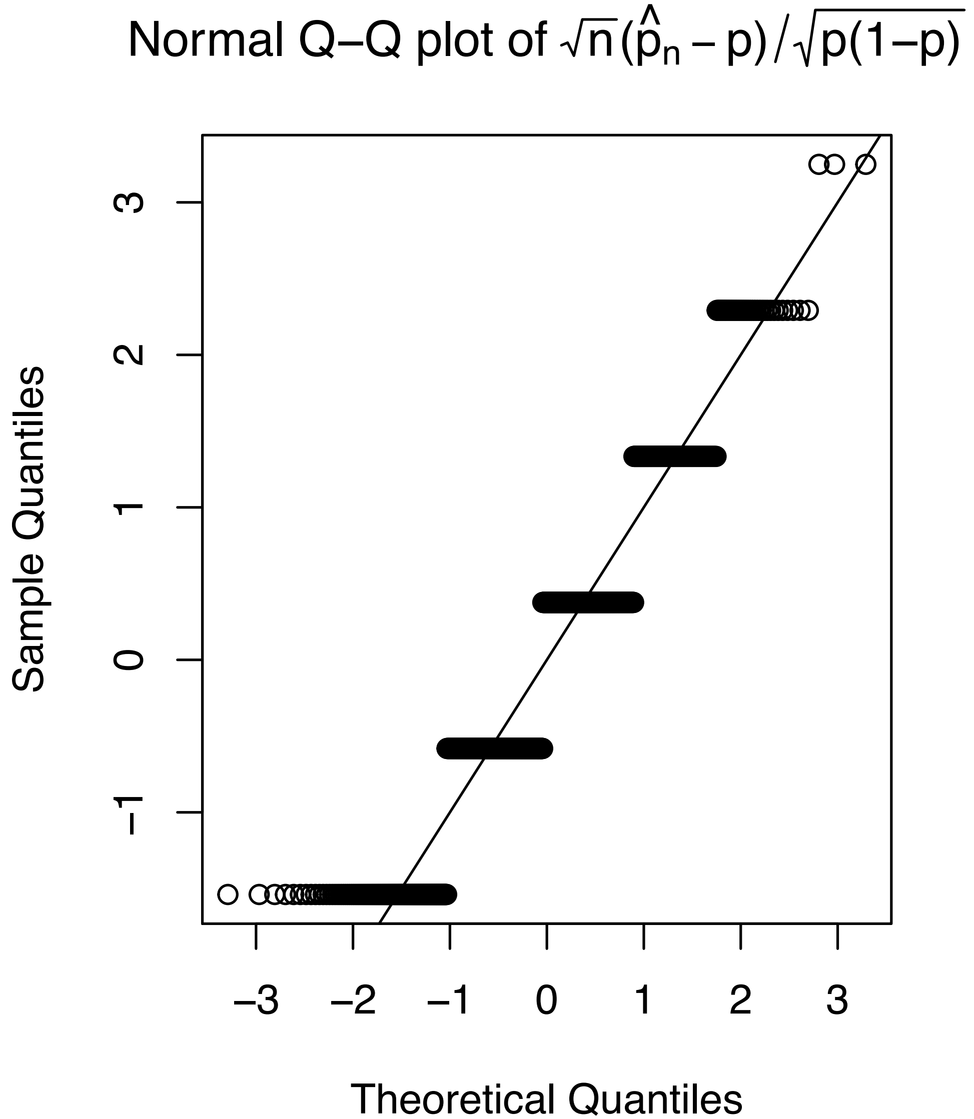
Exercise: Treat the 4,176 abalone as a population. The proportion classified as infants among the abalone is $p = 0.321$; let $\hat{p}_n$ represent the proportion of infants in a random sample of abalone.

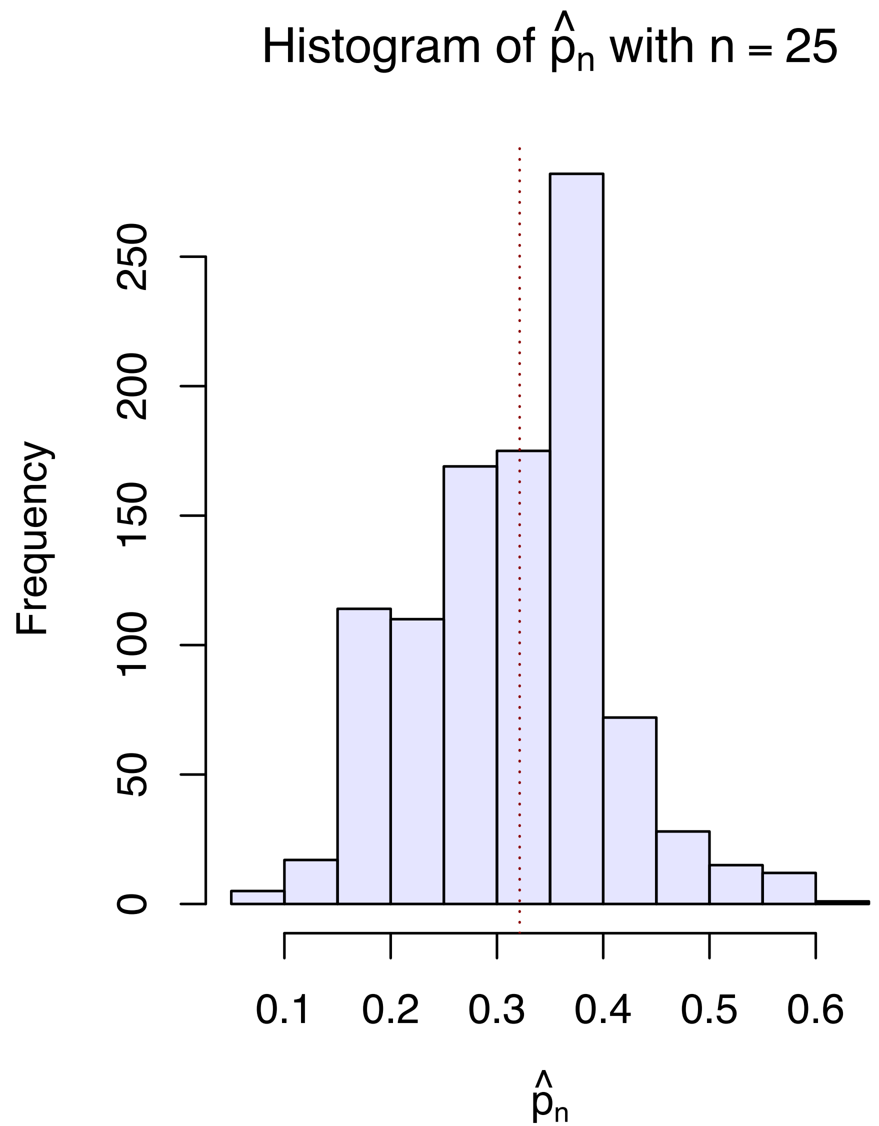
1. For the sample sizes $n = 5, 25, 100$, draw 1,000 samples and
   1. Make a histogram of the $\hat{p}_n$ values.
   2. Make a Normal Q-Q plot of the $\hat{p}_n$.
2. Around what value are the values of $\hat{p}_n$ centered?
3. What changes as n changes?

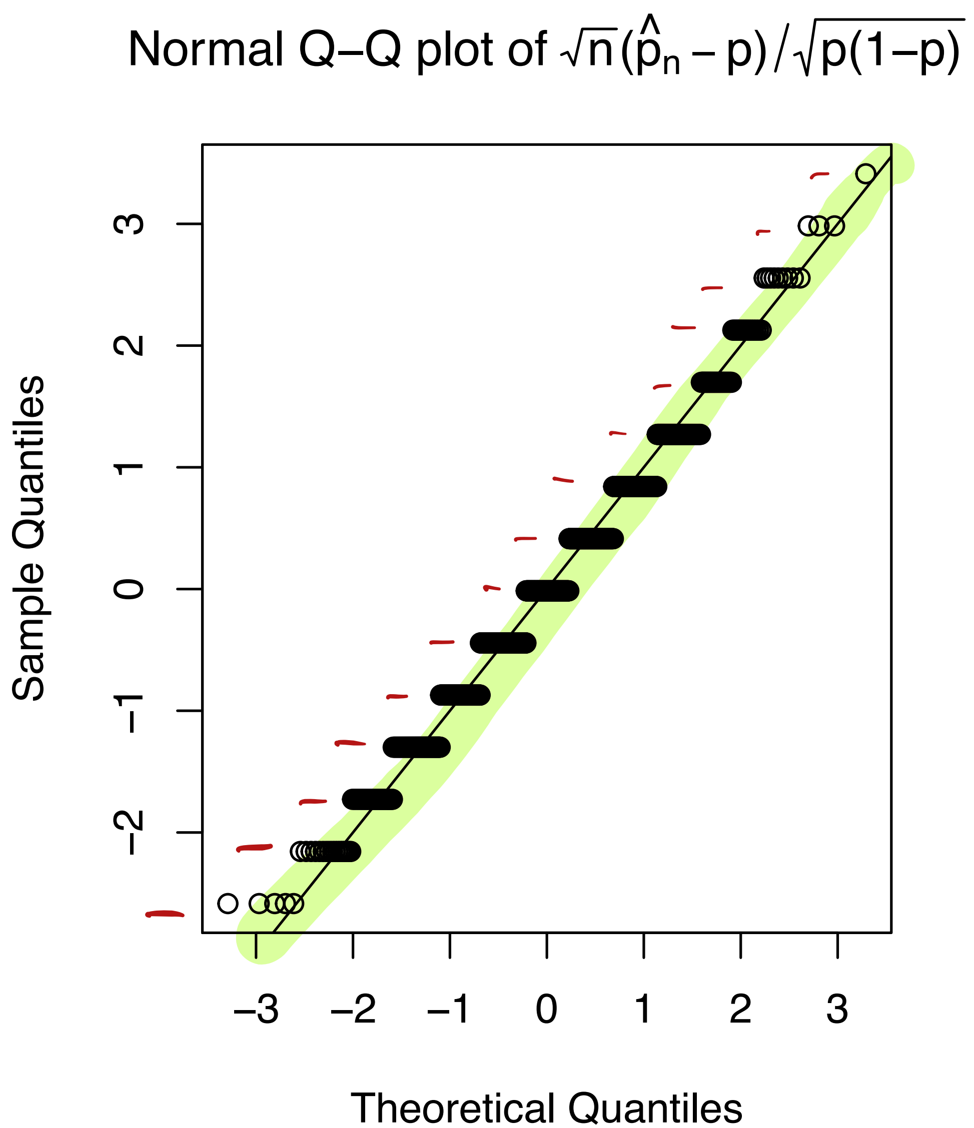
$p = 0.321$

$\hat{p}_{n=5} \in \{0, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, 1\}$

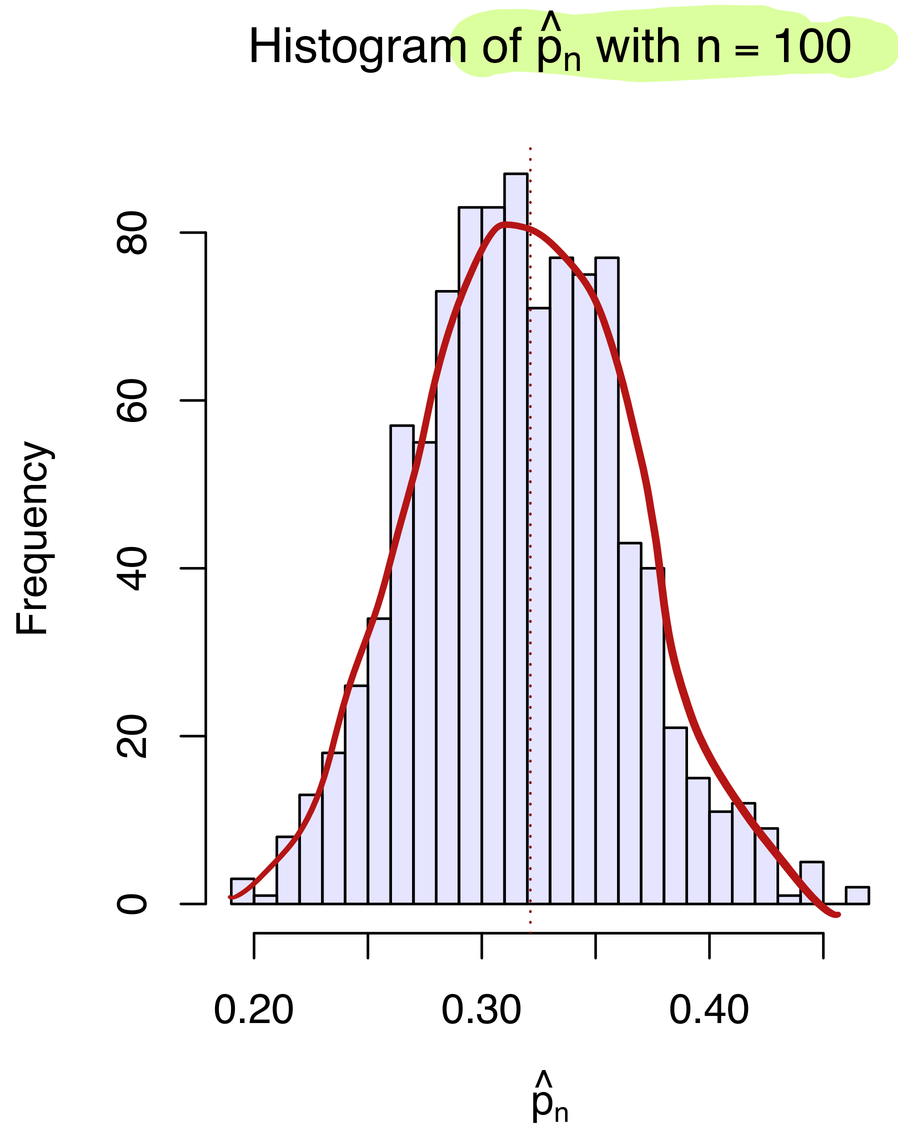


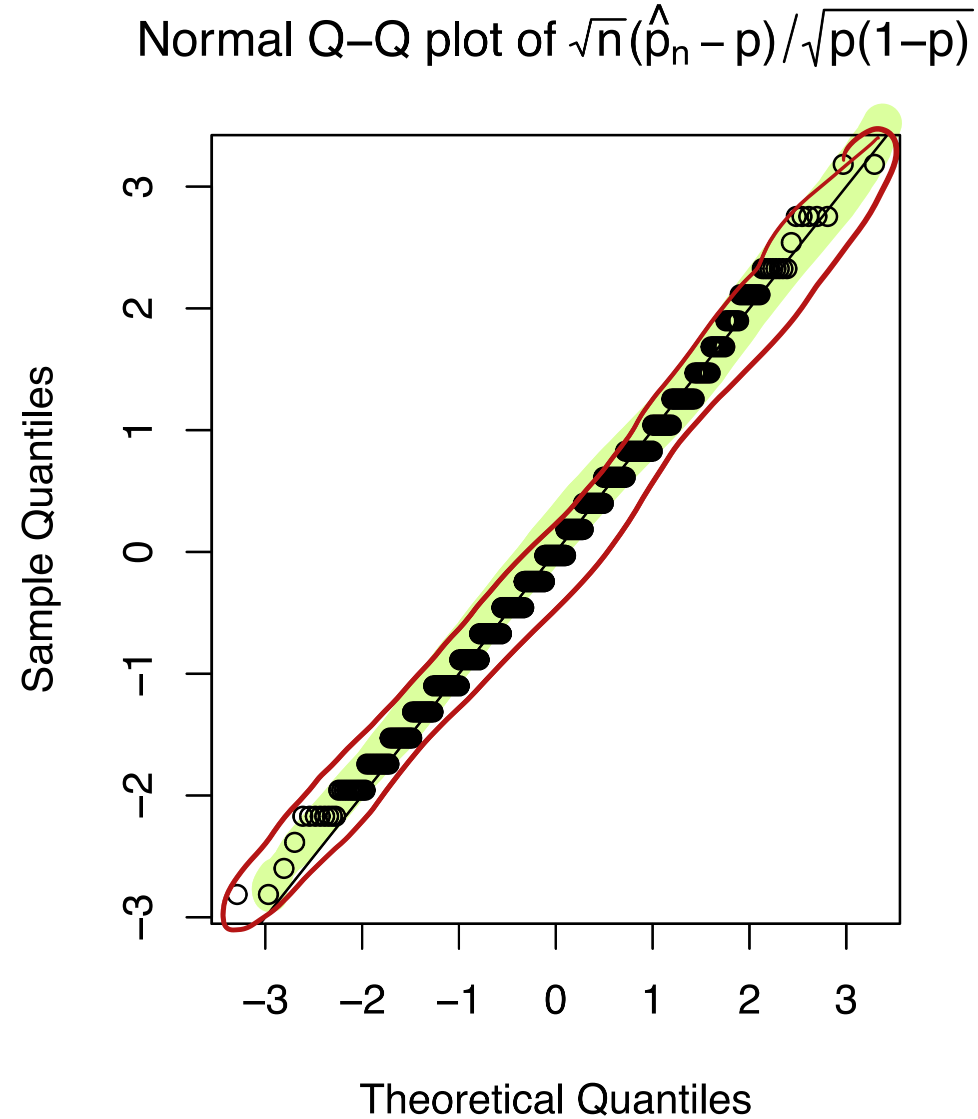



Histogram of $\hat{p}_n$ with n = 25
Frequency
0
50
100
150
200
250
0.1
0.2
0.3
0.4
0.5
0.6
$\hat{p}_n$


Normal Q–Q plot of $\sqrt{n}(\hat{p}_n - p)/\sqrt{p(1-p)}$
Sample Quantiles
−2
−1
0
1
2
3
−3
−2
−1
0
1
2
3
Theoretical Quantiles

$\hat{p}_n = \{0, \frac{1}{100}, \frac{2}{100}, \ldots, \frac{99}{100}, 1\}$





Distribution of sample mean when population is Normal

Let $X_1, \dots, X_n \overset{\text{ind}}{\sim} \text{Normal}(\mu, \sigma^2)$. 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:

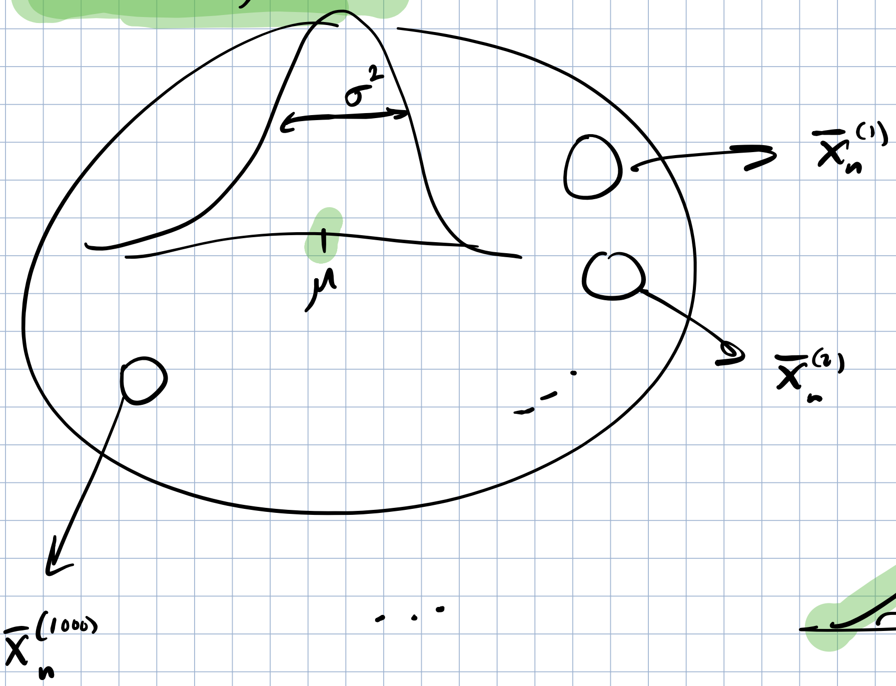
1. Transform a and b to the Z-world (# of standard deviations world):

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

2. Find

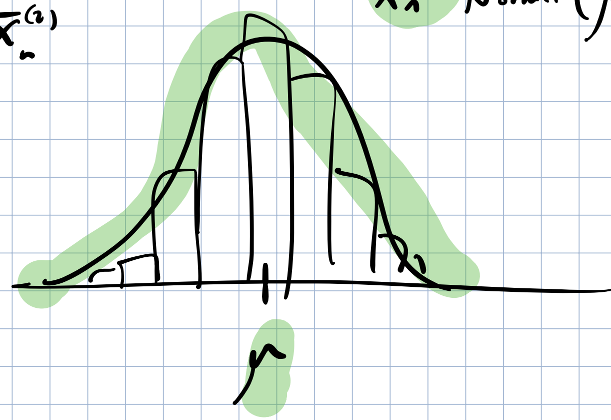
$$P\left(\frac{a - \mu}{\sigma/\sqrt{n}} < Z < \frac{b - \mu}{\sigma/\sqrt{n}}\right).$$

Pop: $\text{Normal}(\mu, \sigma^2)$

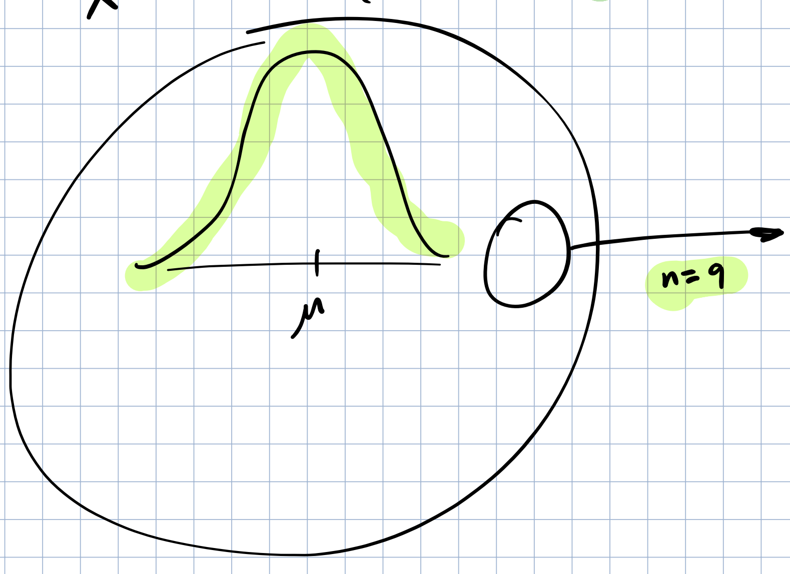


Variance gets divided by n.

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



$X \sim \text{Normal}(\mu = 450, \sigma^2 = 50^2)$



$\bar{X}_n \sim \text{Normal}\left(\mu = 450, \frac{50^2}{9}\right)$

Exercise: Let X = minutes talking on phone in last month of a randomly selected USC student. Assume $X \sim \text{Normal}(\mu = 450, \sigma^2 = 50^2)$.

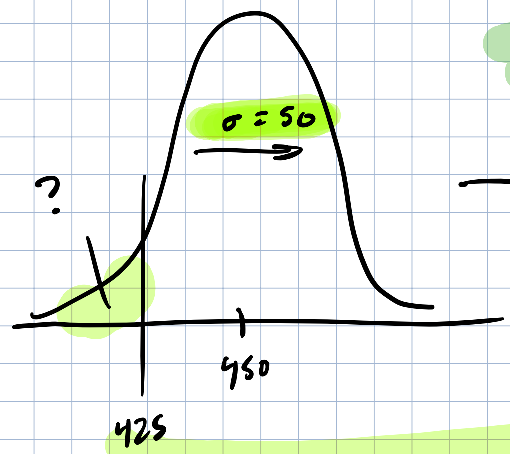
1. a)
- Find $P(|X - 450| > 50)$.
- b) Find $P(X < 425)$.

Now let $\bar{X}_n$ be the mean talk time from $n = 9$ randomly selected students.

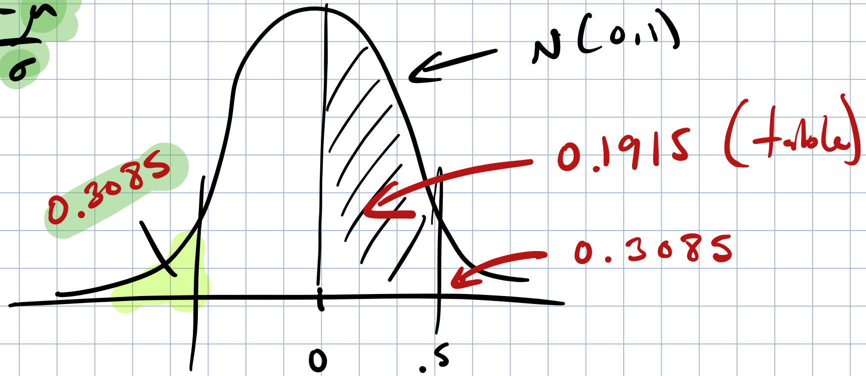
2.
- a) Find $P(|\bar{X}_n - 450| > 50)$.
- b) Find $P(\bar{X}_n < 425)$.

1. b) $P(X < 425) = 0.3085$

$X \sim \text{Normal}(\mu = 450, \sigma^2 = 50^2)$



$$z = \frac{x-\mu}{\sigma}$$



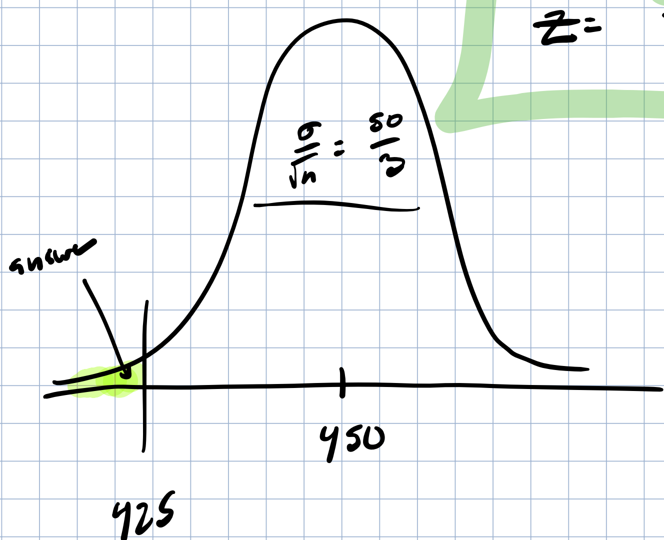
$$\frac{425-450}{50} = -0.50$$

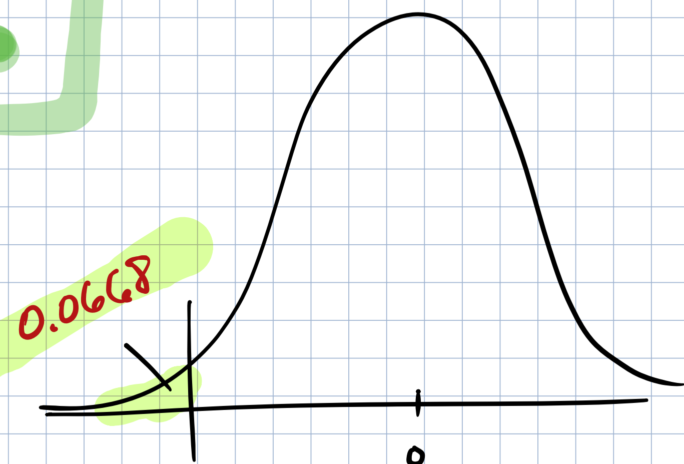
2. b) $n = 9.$

$P(\bar{X}_n \le 425)$, $\bar{X}_n \sim \text{Normal}\left(\mu = 450, \frac{\sigma^2}{n} = \frac{50^2}{9}\right)$

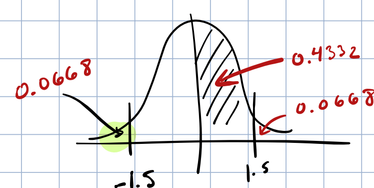
$$\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} = \frac{50}{3}$$

$$\boxed{z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}}$$





$$\frac{425-450}{50/3} = -1.5$$



$$\boxed{\begin{aligned} P(X \le 425) &= 0.3085 \\ P(\bar{X}_n \le 425) &= 0.0668 \end{aligned}}$$

random sample

Central Limit Theorem

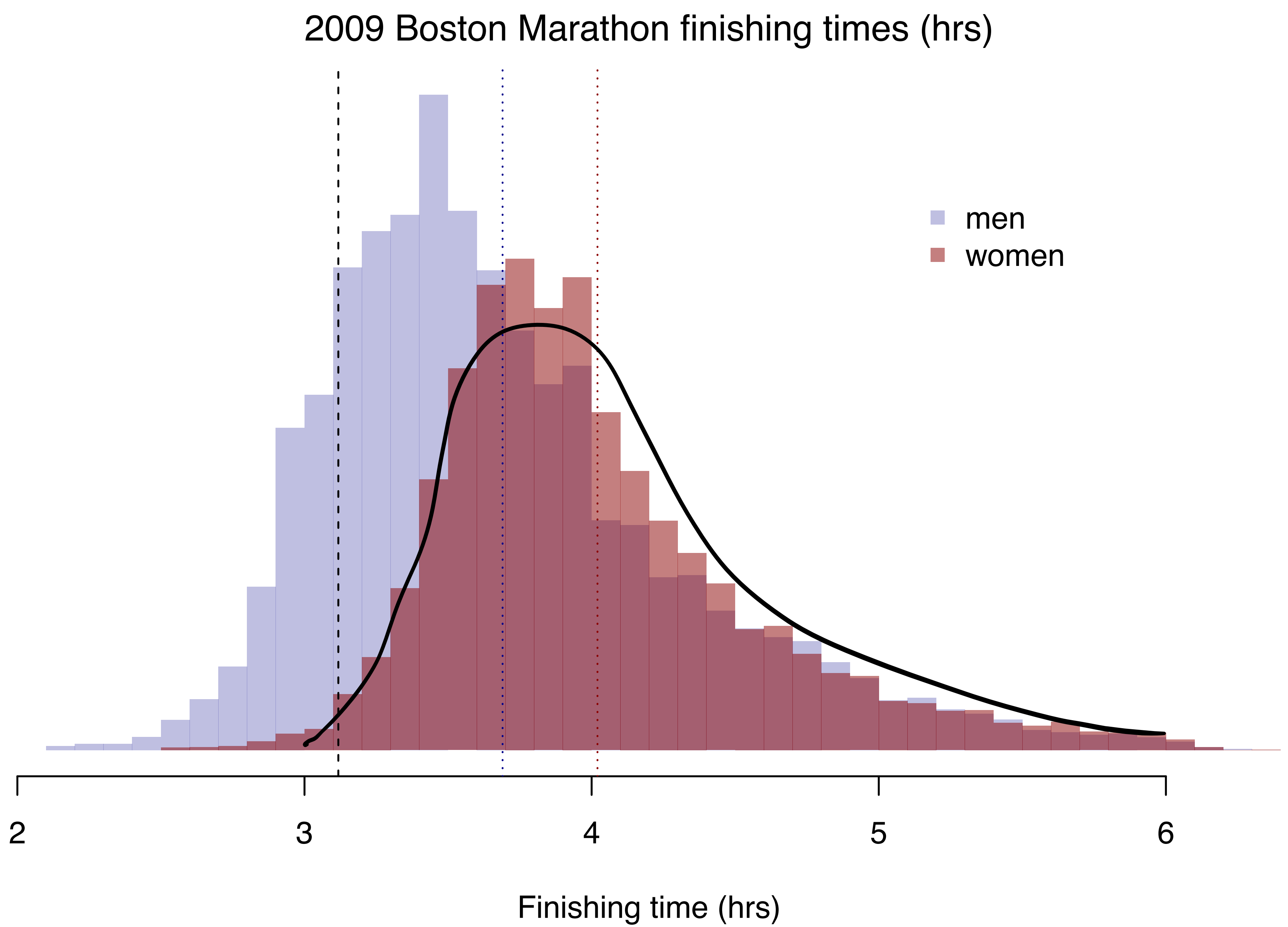
Let $X_1, \ldots, X_n$ be a rs from a dist. with mean $\boxed{\mu}$ and variance $\boxed{\sigma^2 < \infty.}$ Then

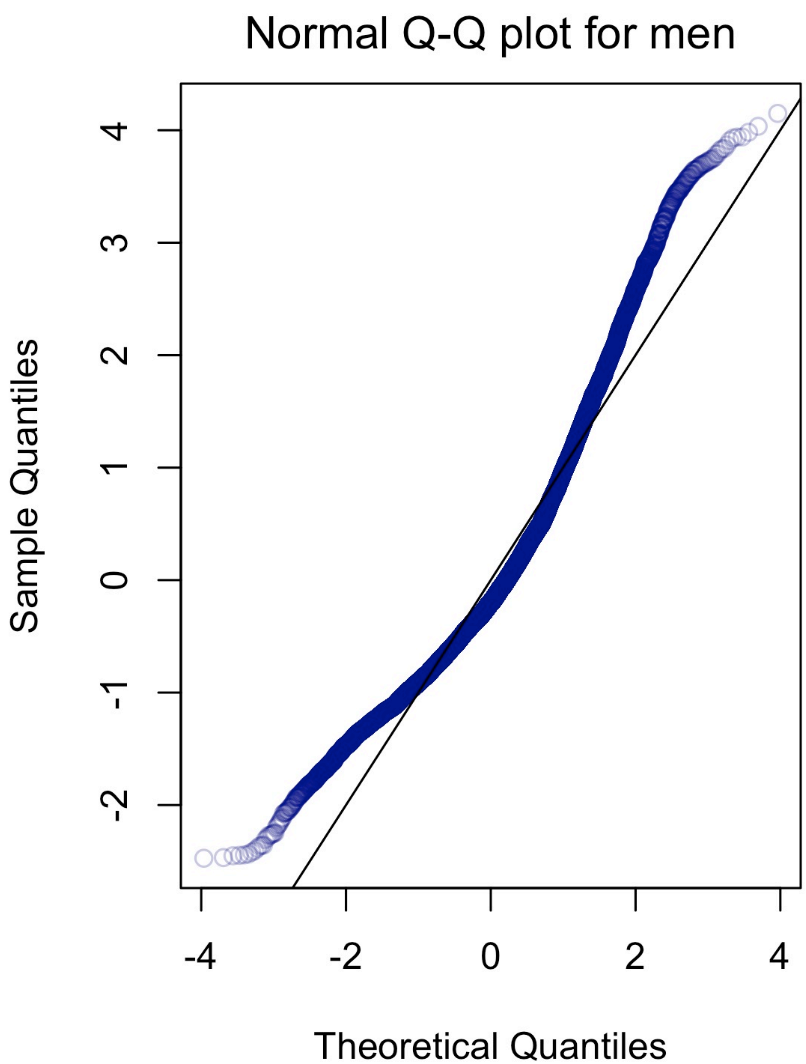
$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \text{ behaves more and more like } Z \sim \text{Normal}(0, 1)$$

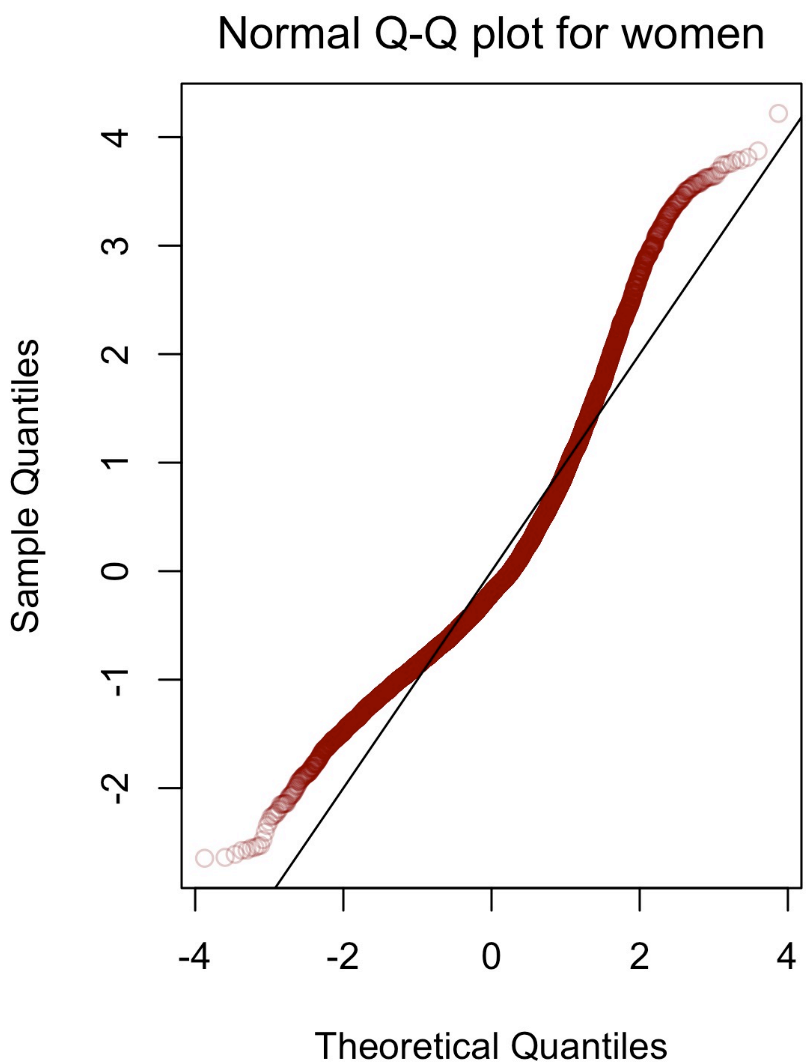
for larger and larger n.

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)
men
women
2
3
4
5
6
Finishing time (hrs)


Normal Q-Q plot for men
Sample Quantiles
Theoretical Quantiles
-4
-2
0
2
4
-2
-1
0
1
2
3
4


Normal Q-Q plot for women
Sample Quantiles
Theoretical Quantiles
-4
-2
0
2
4
-2
-1
0
1
2
3
4

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.

X has mean $\mu = 4.02$, variance $\sigma^2 = (0.555)^2$.

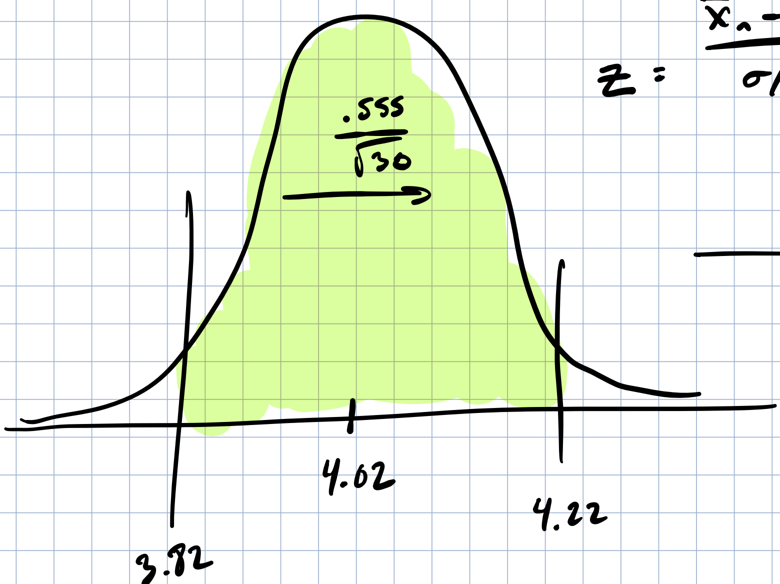
Let $n = 30$.

Find (approximately) $P(|\bar{X}_n - 4.02| < 0.2) = 0.9512$

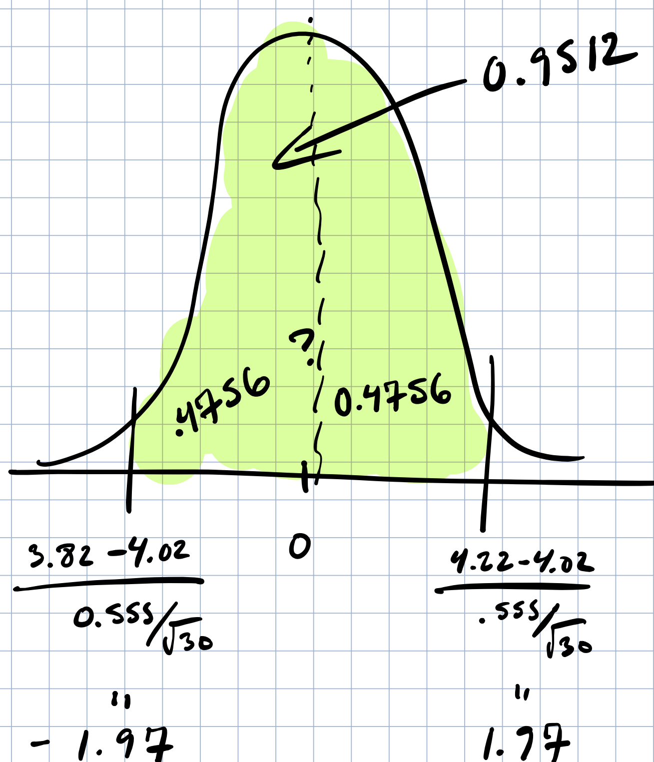
$|a| < b \iff -b < a < b$

$$\begin{aligned} &= P(-0.2 < \bar{X}_n - 4.02 < 0.2) \\ &= P(4.02 - 0.2 < \bar{X}_n < 4.02 + 0.2) \\ &= P(3.82 < \bar{X}_n < 4.22) \end{aligned}$$

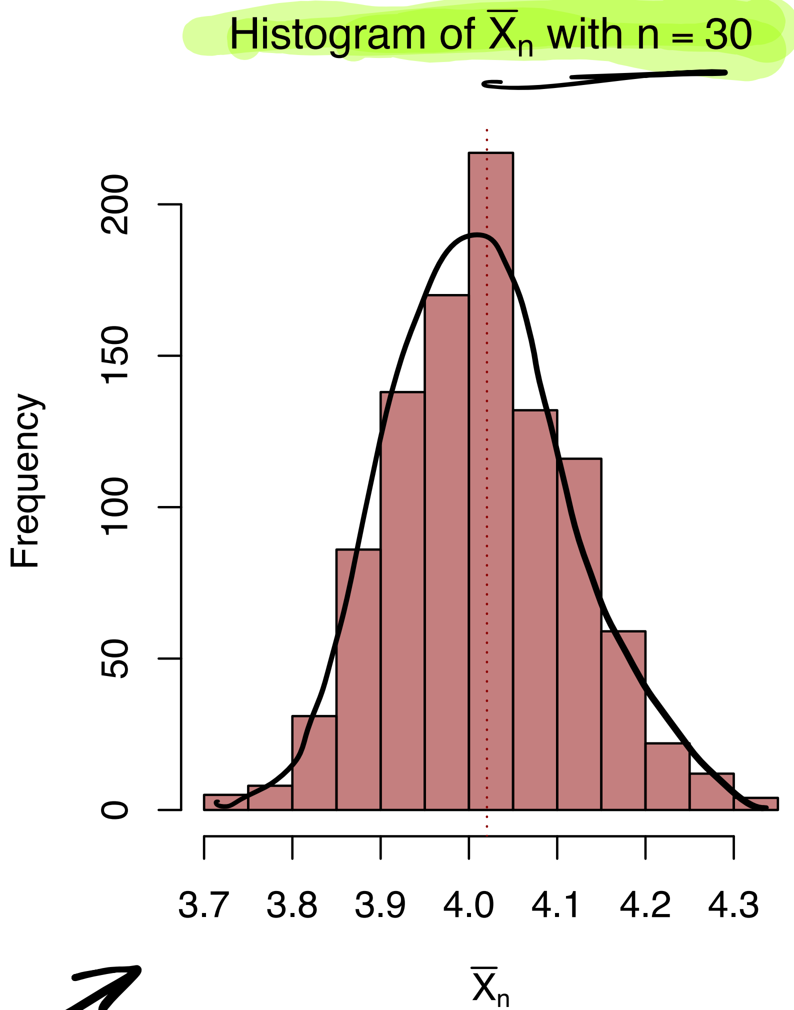
Assume $\bar{X}_n \sim N\left(4.02, \dfrac{(0.555)^2}{30}\right)$.

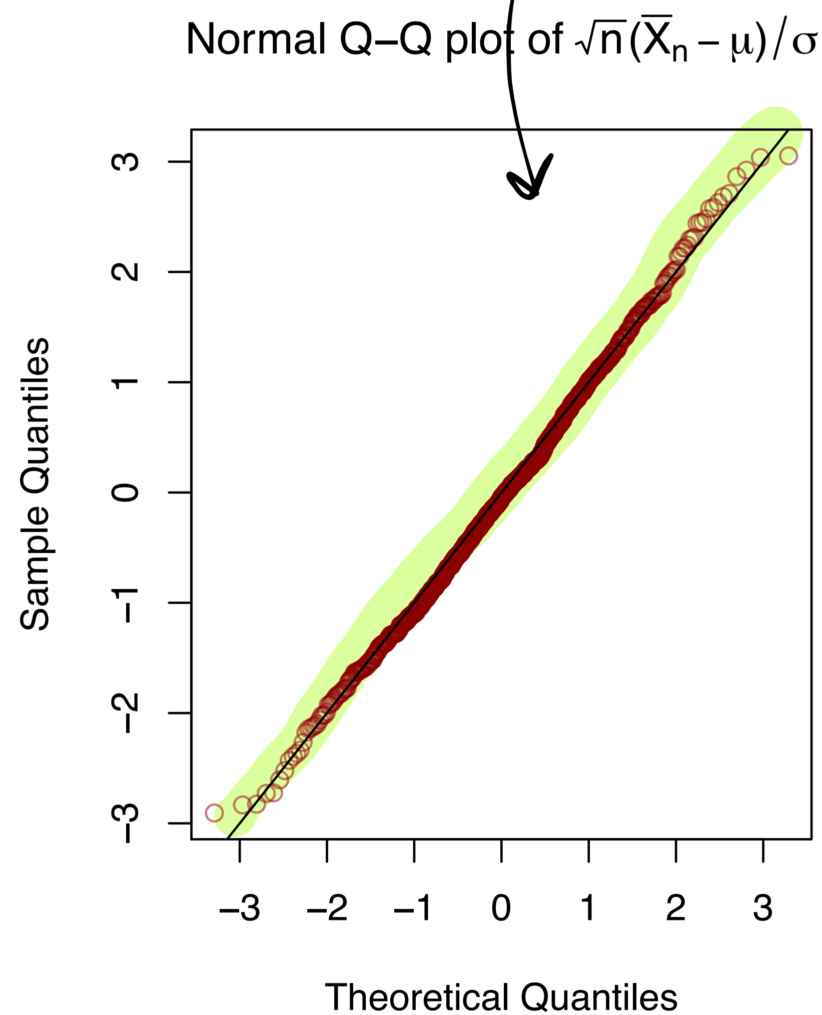


$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

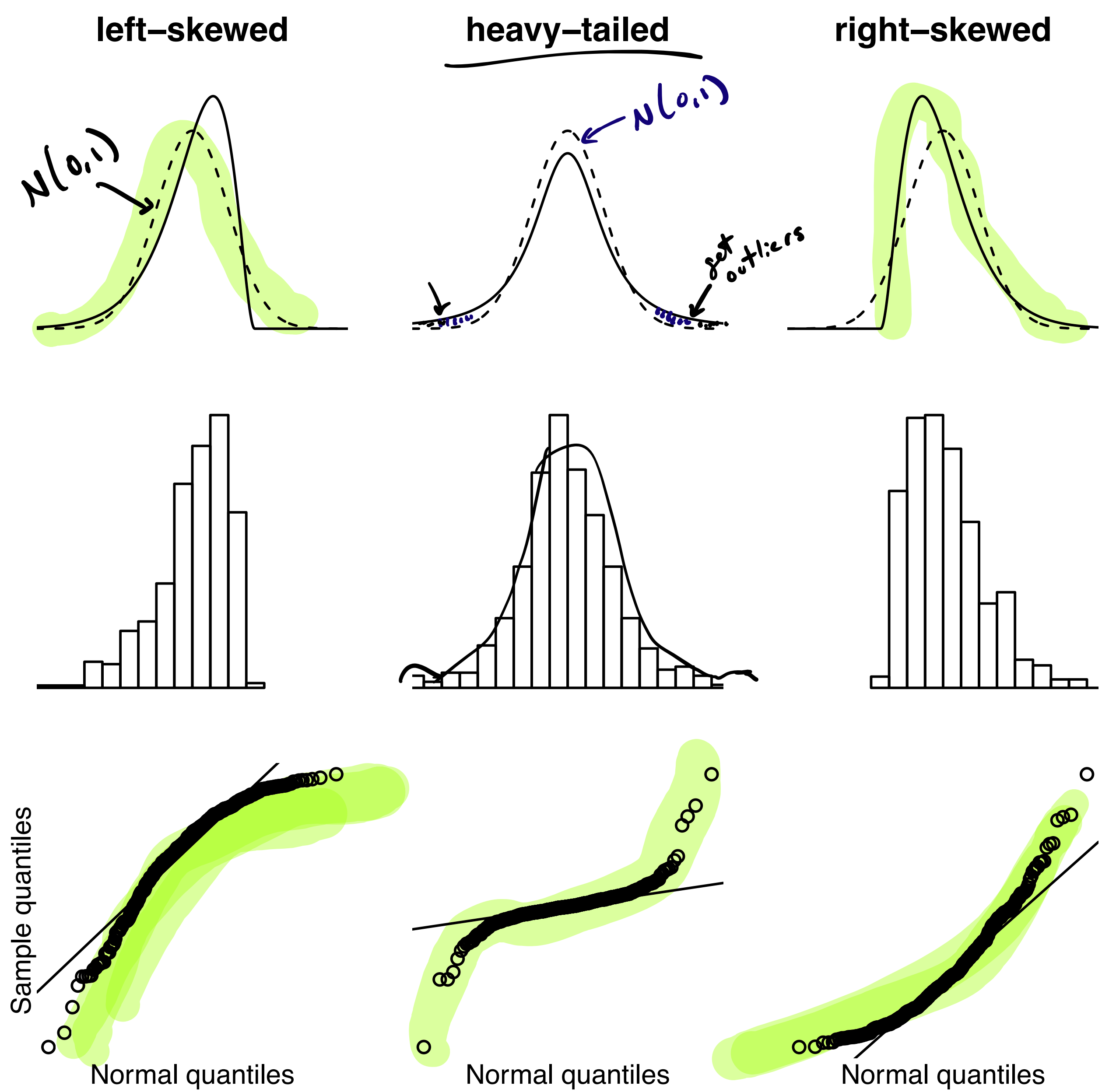


$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$





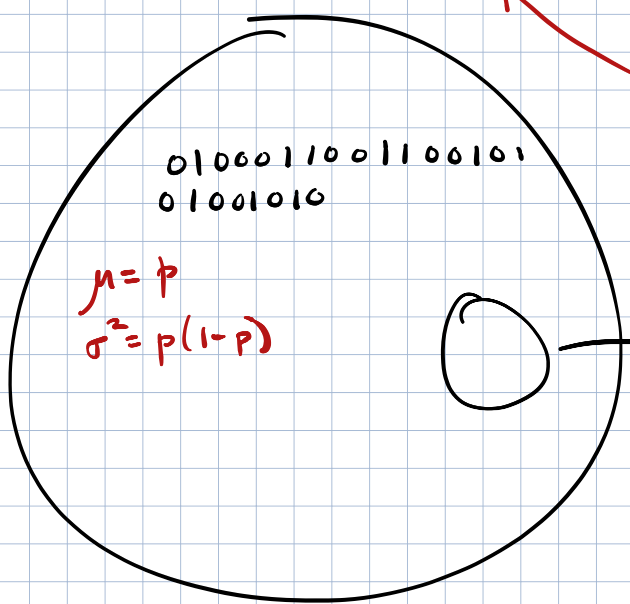
1000 values of $\bar{X}_n$ from Women's finishing times


left–skewed
heavy–tailed
right–skewed
N(0,1)
N(0,1)
fat outliers
Sample quantiles
Normal quantiles
Normal quantiles
Normal quantiles

Population of 0s and 1s, with proportion p of 1's.

$X \sim \text{Bernoulli}(p)$

population proportion (of 1's)



$$\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$$

$$= \frac{\#\{1\text{'s in the sample}\}}{n}$$

$$= \hat{p}_n$$, the sample proportion.

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

$\mu = p$, $\sigma^2 = p(1-p)$

Central Limit Theorem for the sample proportion

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

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

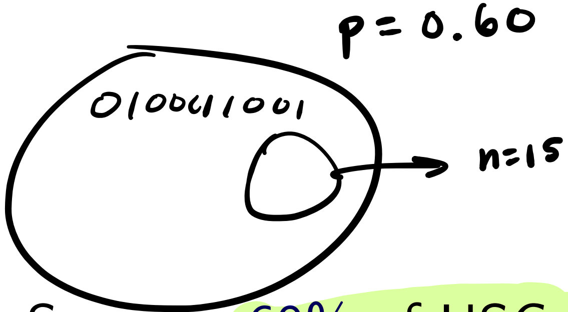
just $\dfrac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$

for larger and larger n.

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

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

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



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 $n = 100$.

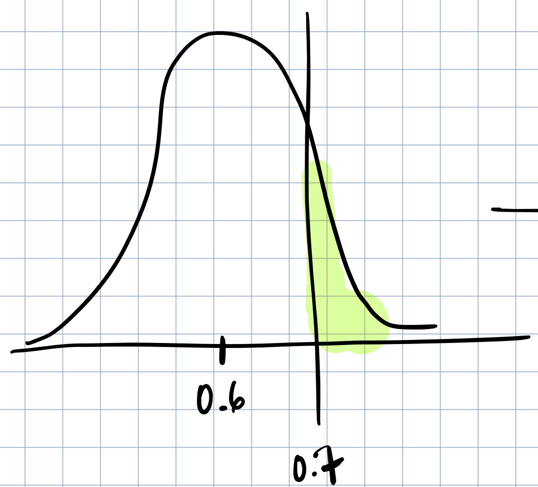
① $p = 0.60$, $n = 15$

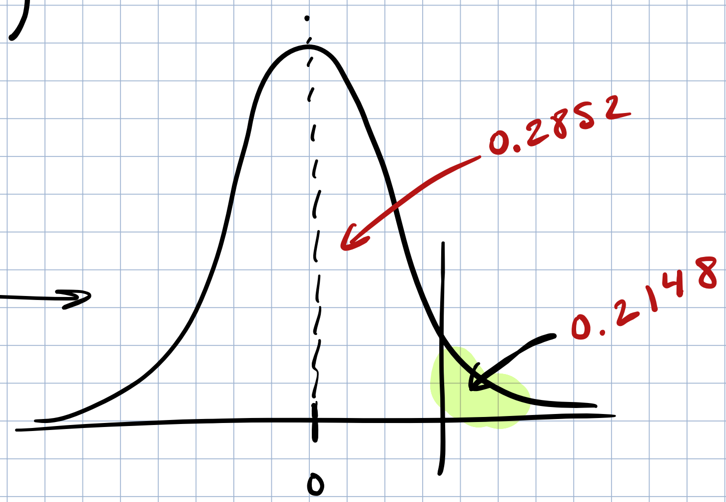
$$\hat{p}_n \overset{\text{approx}}{\sim} \text{Normal}\left(p, \underbrace{\frac{p(1-p)}{n}}\right).$$

$$P(\hat{p}_n > 0.70)$$

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

$$\hat{p}_n \overset{\text{approx}}{\sim} \text{Normal}\left(0.60, \frac{0.6(1-0.6)}{15}\right)$$





$$\frac{0.7 - 0.60}{\sqrt{\frac{0.6(1-0.6)}{15}}} = 0.79$$

$$P(\hat{p}_n > 0.70) \approx 0.2148$$

approximation not to bad...

② Find exact value of $P(\hat{p}_n > 0.70)$.

$$P(\hat{p}_n > 0.70) = P(\underbrace{15 \cdot \hat{p}_n}_{n} > 15 \cdot 0.70)$$

$$= P(Y > 15(0.70)), \quad Y \sim \text{Binomial}(n=15, p=0.60)$$

$$= 1 - P(Y \le 15(0.70))$$

$= 1 - \text{pbinom}\left(15(0.70), 15, \text{prob} = 0.60\right)$

$= 0.2173.$

$$\hat{p}_n = \frac{\#\{1\text{'s in sample}\}}{15}$$

$$15 \cdot \hat{p}_n = \#\{1\text{'s in sample}\}$$

$$\sim \text{Binomial}(n = 15, p = 0.60)$$

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

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

pop. dist. not Normal

X non - Normal

$n \geq 30$: $\dfrac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \overset{\texttt{approx}}{\sim} \texttt{Normal}(0,1)$ ✓

$n < 30$: $\dfrac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ non-Normal ...

$X \sim N(\mu, \sigma^2)$

Pop. dist. is Normal

$\dfrac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \texttt{Normal}(0,1)$ ✓

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

pop is Bernoulli(p)

np = Expected # "successes"
$n(1-p)$ = Expected # "failures".

small sample size
$\min\{np, n(1-p)\} < 5$

$$n\hat{p}_n \sim \texttt{Binomial}(n, p)$$

$\min\{np, n(1-p)\} \geq 5$
large sample size

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

$$\frac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{\texttt{approx}}{\sim} \texttt{Normal}(0, 1)$$

$$n\hat{p}_n \overset{\texttt{approx}}{\sim} \texttt{Normal}(np, np(1-p))$$

$$n\hat{p}_n \sim \texttt{Binomial}(n, p)$$