Statistics
Probability
Set theory

random variable
- discrete
  - binomial
  - hyper
  - poisson
- continuous
  - Normal
  - Exponential

STAT 515 fa 2023 Lec 09 slides

Sampling distributions and the Central Limit Theorem

Karl B. Gregory

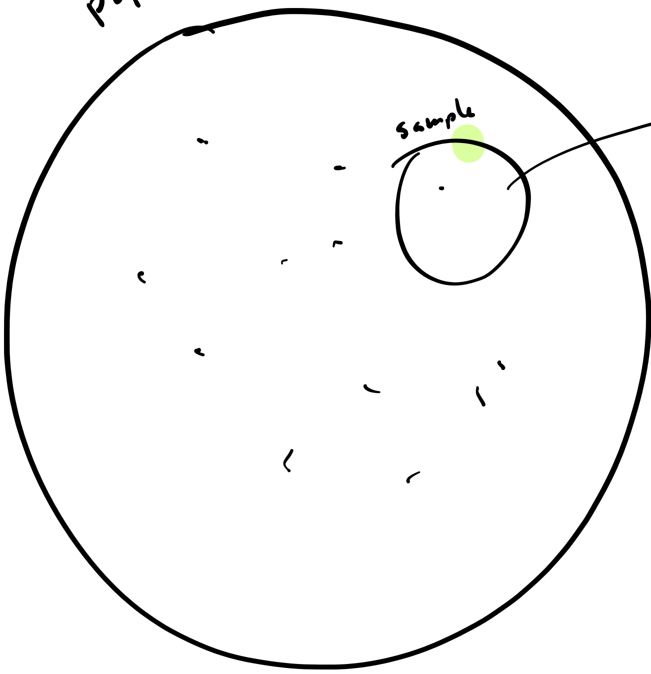
University of South Carolina

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.

population

μ = population mean

σ^2 = population variance



$$\overbrace{X_1, X_2, \dots, X_n}^{\text{random sample}}$$

n = sample size

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

"X bar"

Sample mean

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

= avg. squared deviation from the mean in our sample

Sample variance

$X \sim \text{Binom}(n,p)$

Random sample

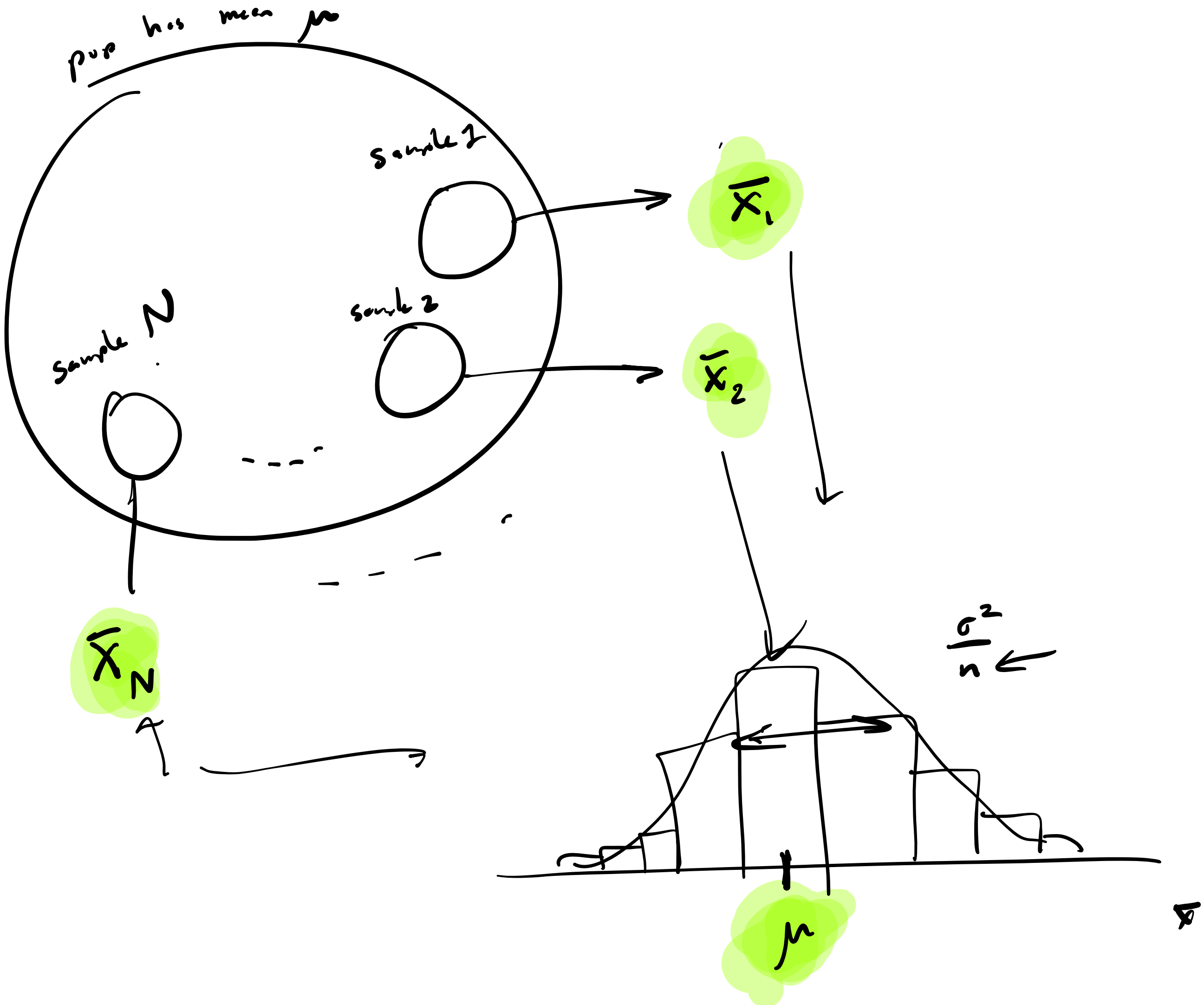
random variable

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*: independent and identically distributed.
- Common distribution of $X_1, \ldots, X_n$ called the *population distribution*.
- Can write $X_1, \ldots, X_n \overset{\text{ind}}{\sim} F$ if a rs from a distribution F.

$X_1, \ldots, X_n \overset{\text{ind}}{\sim} \text{Binom}(n,p)$

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


pop has mean μ
sample 1
sample 2
sample N
$\bar{x}_1$
$\bar{x}_2$
$\bar{x}_N$
$\frac{\sigma^2}{n}$
μ
$\bar{x}$

random sample

$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

Expected value and variance of the sample mean

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

λ λ

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

Examples:

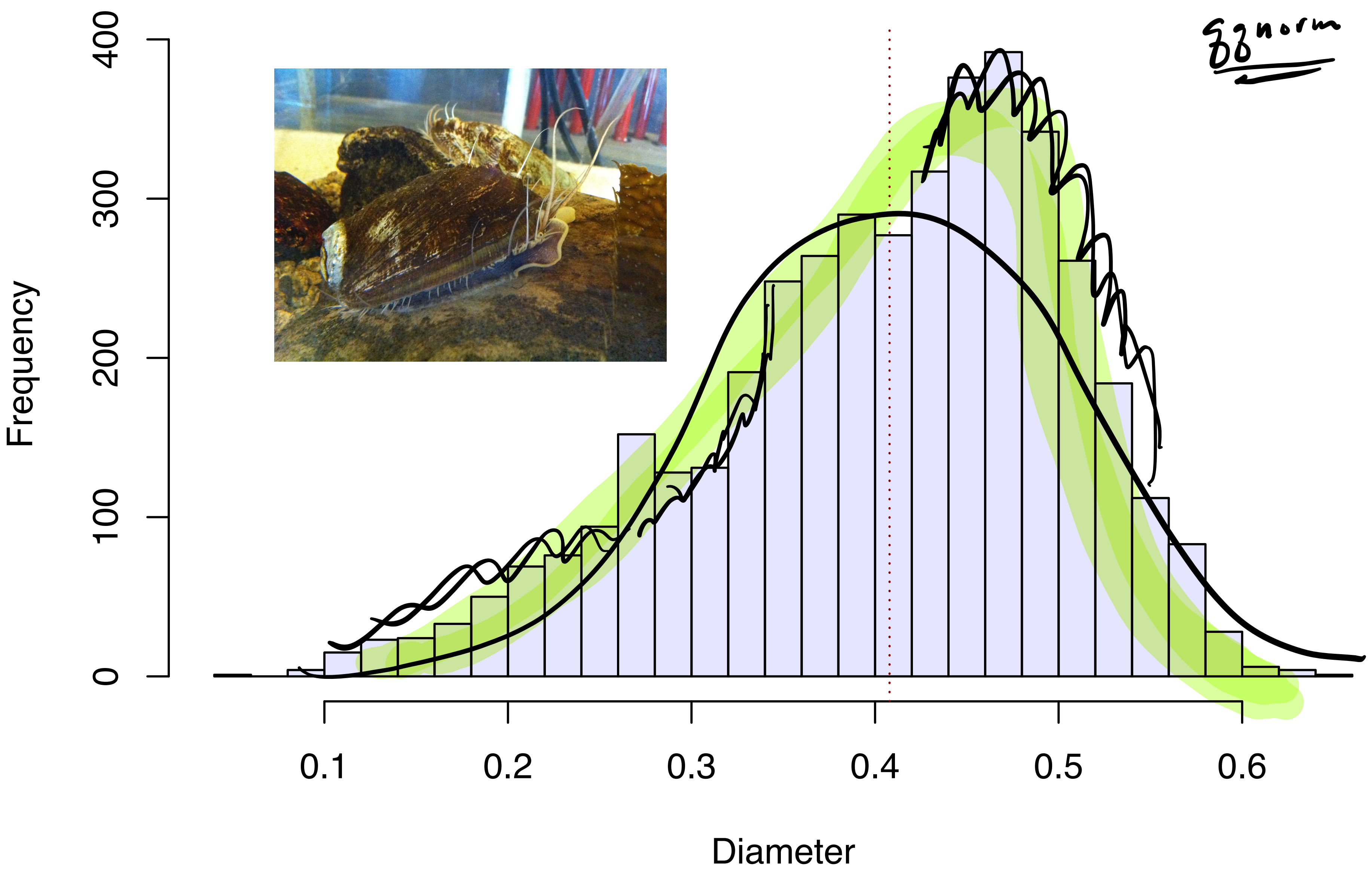
1. If $X_1, \dots, X_n \overset{\text{ind}}{\sim}$ Normal(μ, σ^2), then $\mathbb{E}\bar{X}_n = \mu$ and Var $\bar{X}_n = \sigma^2/n$.
2. If $X_1, \dots, X_n \overset{\text{ind}}{\sim}$ $\boxed{\text{Bernoulli}(p)}$, then $\boxed{\mathbb{E}\bar{X}_n = p}$ and Var $\bar{X}_n = p(1-p)/n$. $\mu = p$, $\sigma^2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

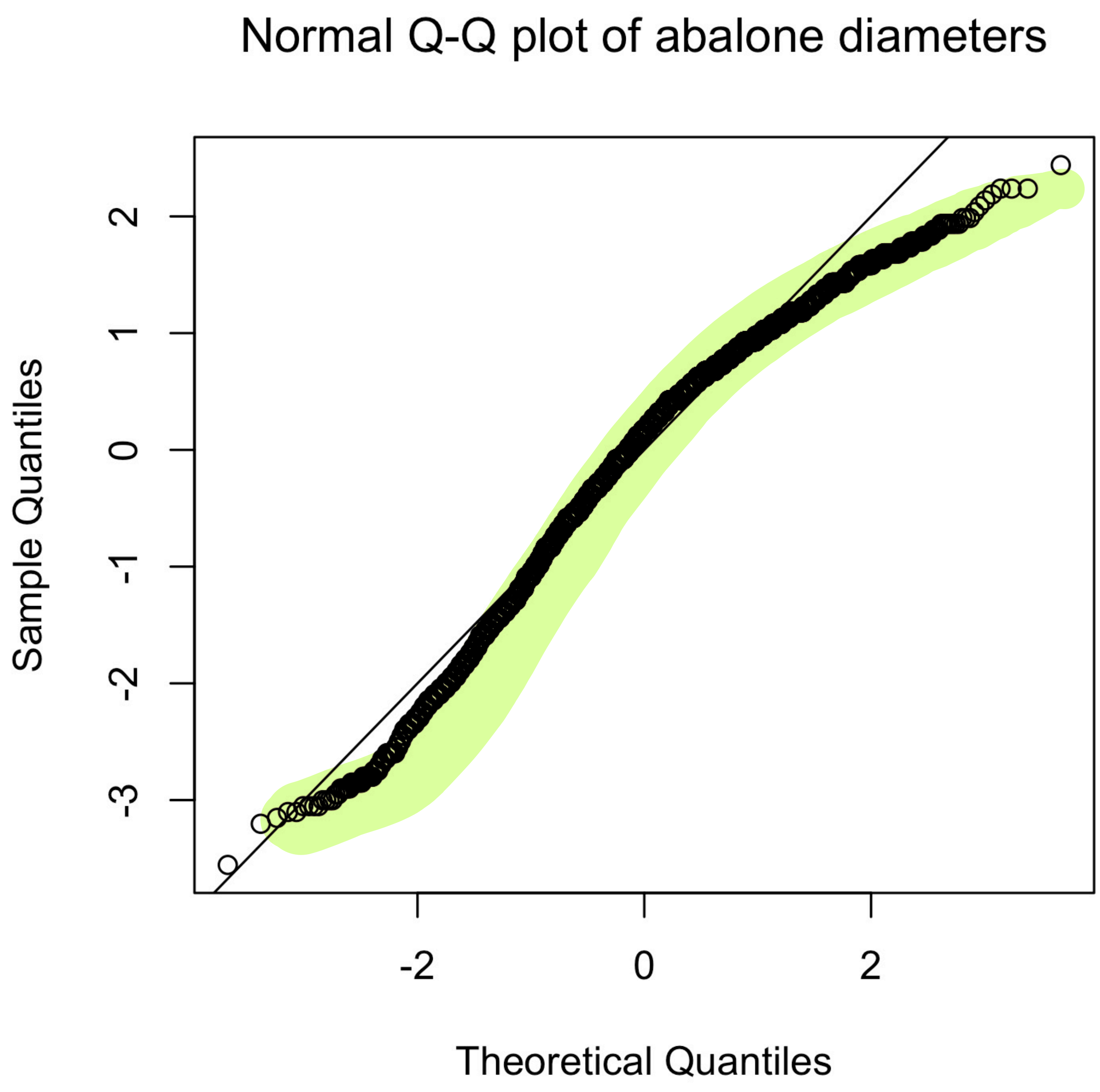
$\mu = \frac{1}{\lambda}$

$\sigma^2 = \frac{1}{\lambda^2}$

← population ← μ

Consider the diameters of 4,176 abalones with mean 0.4078915. link to data

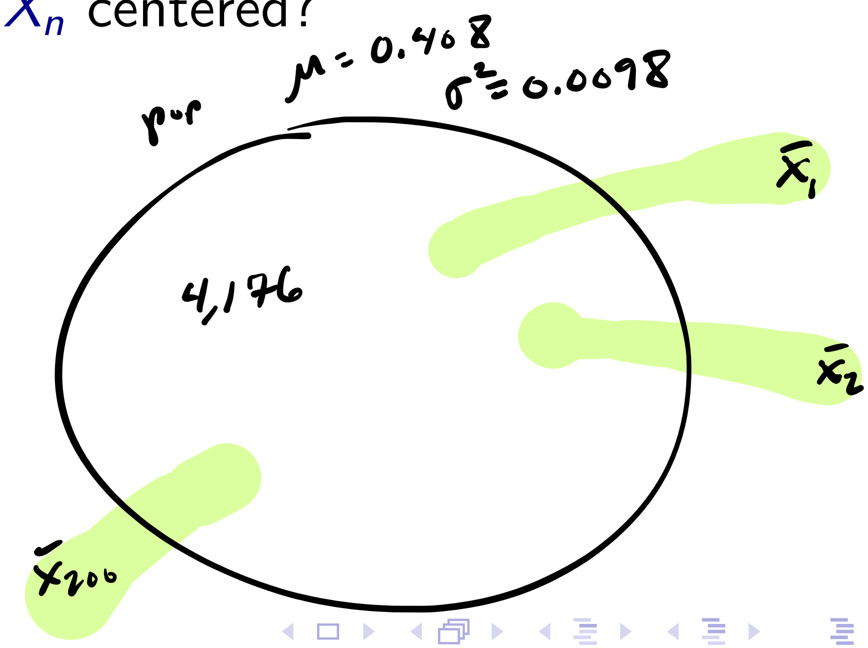



Normal Q-Q plot of abalone diameters
Sample Quantiles
Theoretical Quantiles
-3
-2
-1
0
1
2
-2
0
2

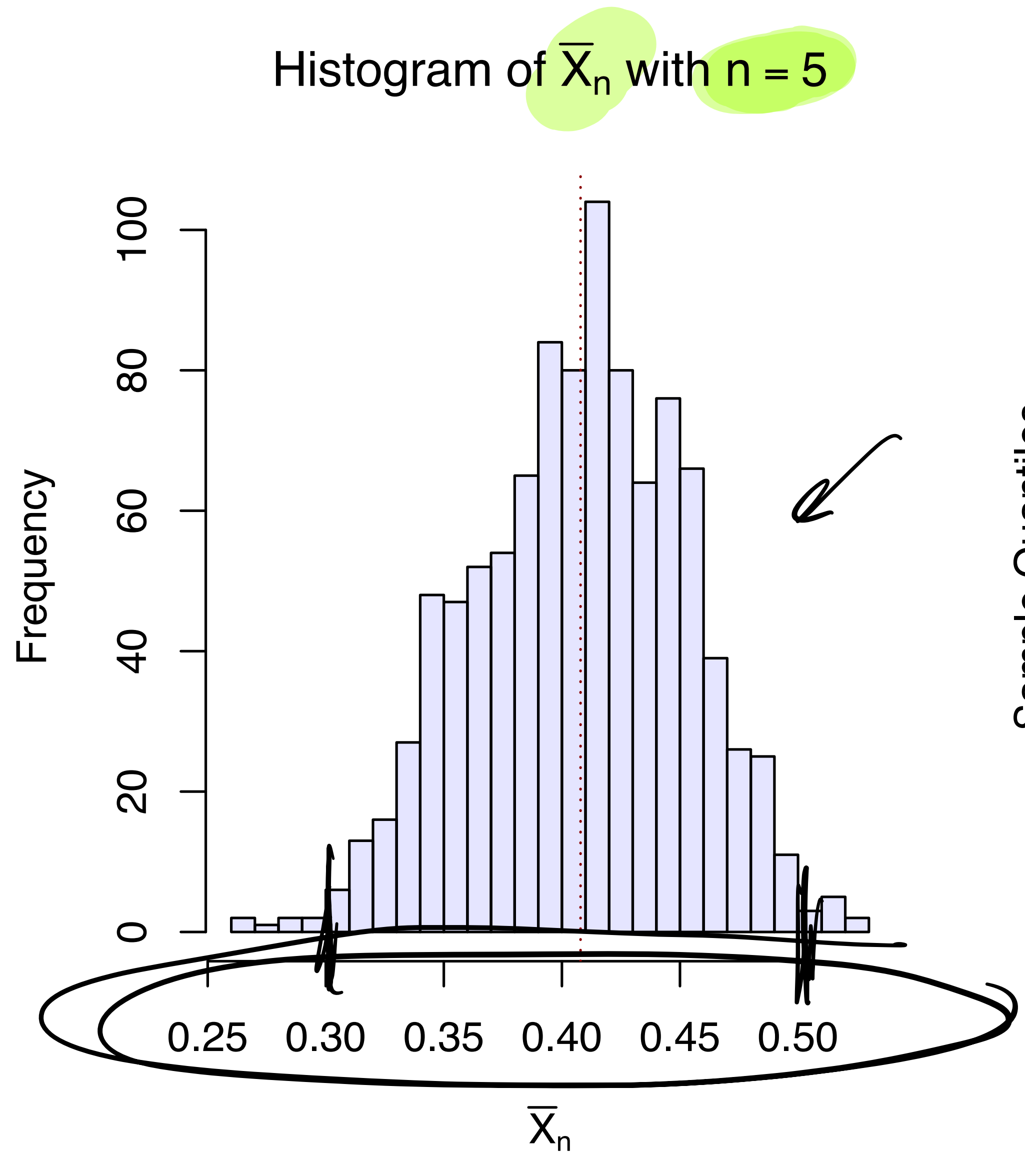
Exercise: Treat the 4,176 abalone as a population. The mean diameter is $\mu = 0.408$. Let $\bar{X}_n$ be the mean diameter from a sample of abalone.

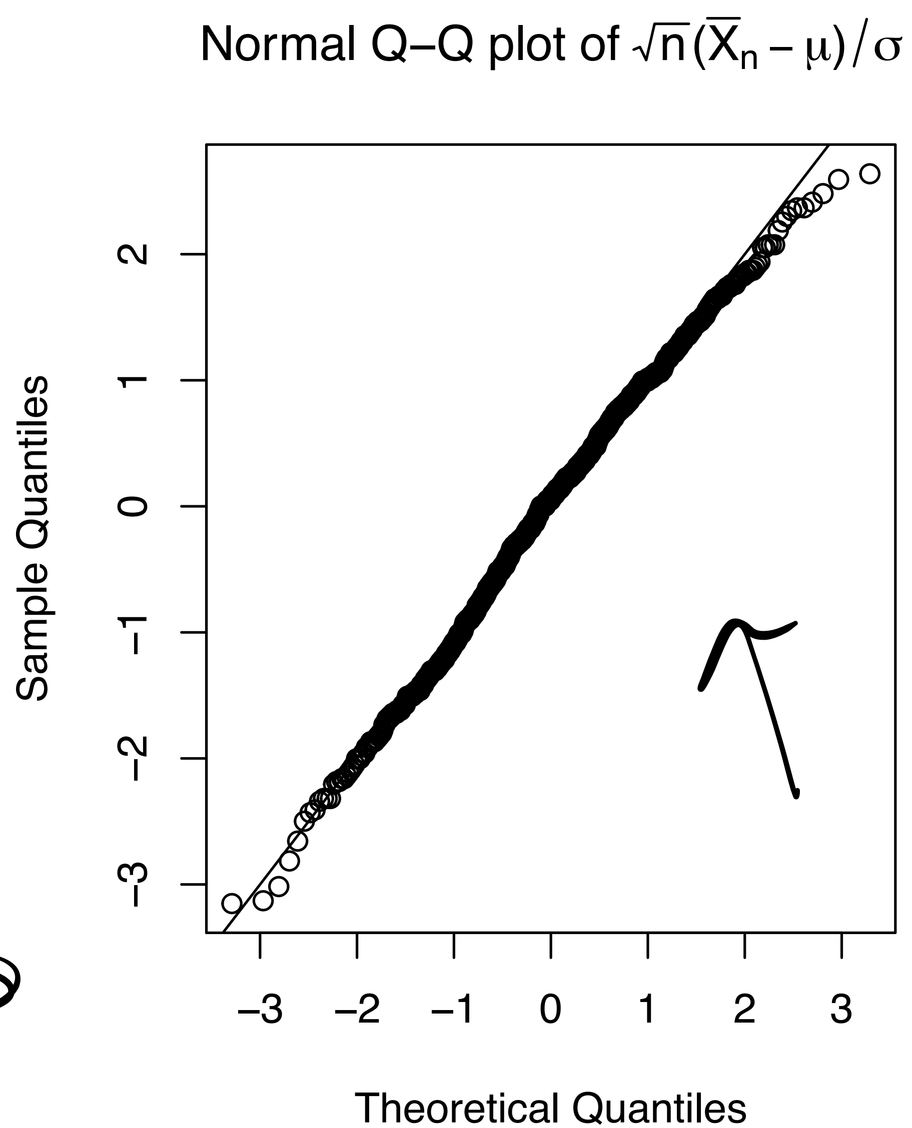
1. For the sample sizes $n = 5, 25, 100$, draw 1,000 samples and
   1. Make a histogram of the $\bar{X}_n$ values.
   2. Make a Normal Q-Q plot of the $\bar{X}_n$ values.
2. Around what value are the values of $\bar{X}_n$ centered?
3. What changes as n changes?

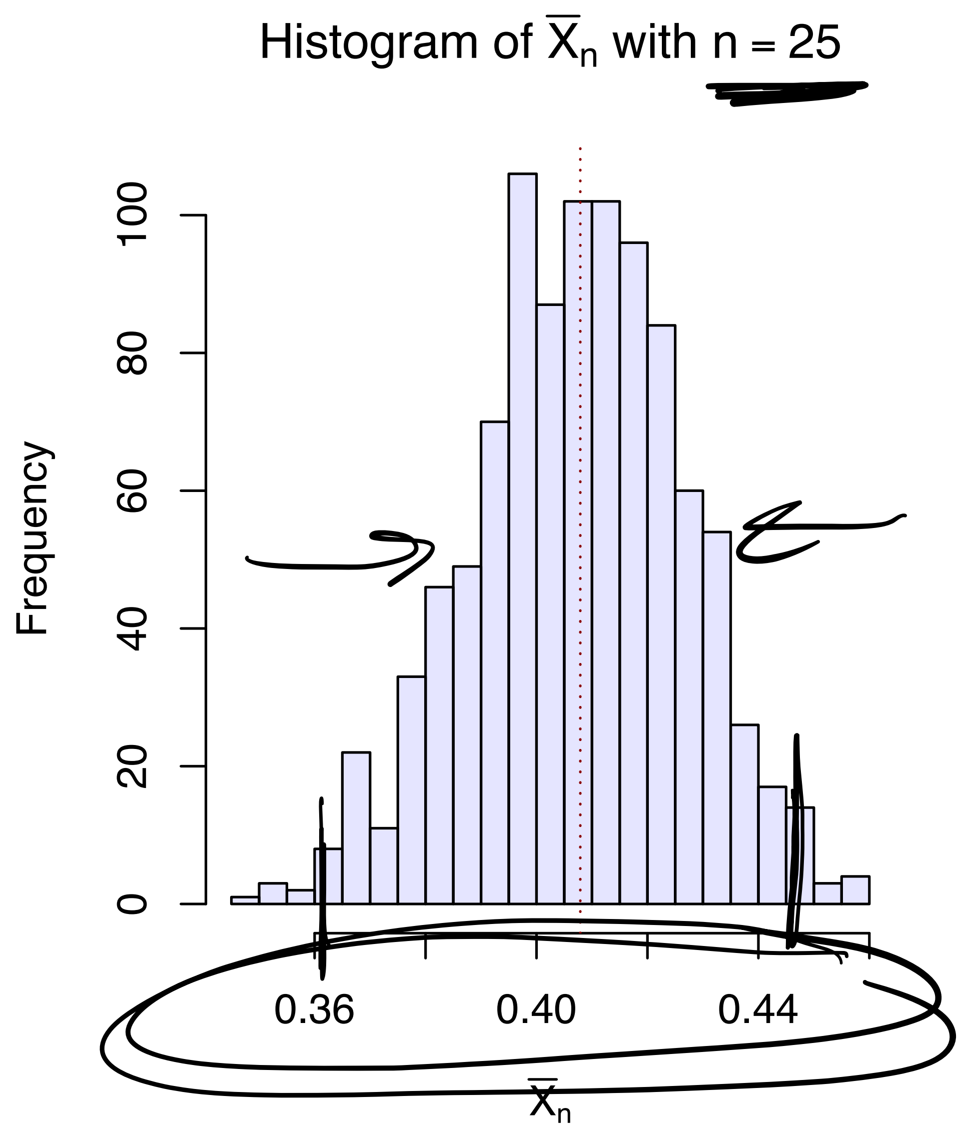
n	Var $\bar{X}_n$	$\frac{\sigma^2}{n}$
5	0.002	0.002
25	0.00036	0.00039
↓		

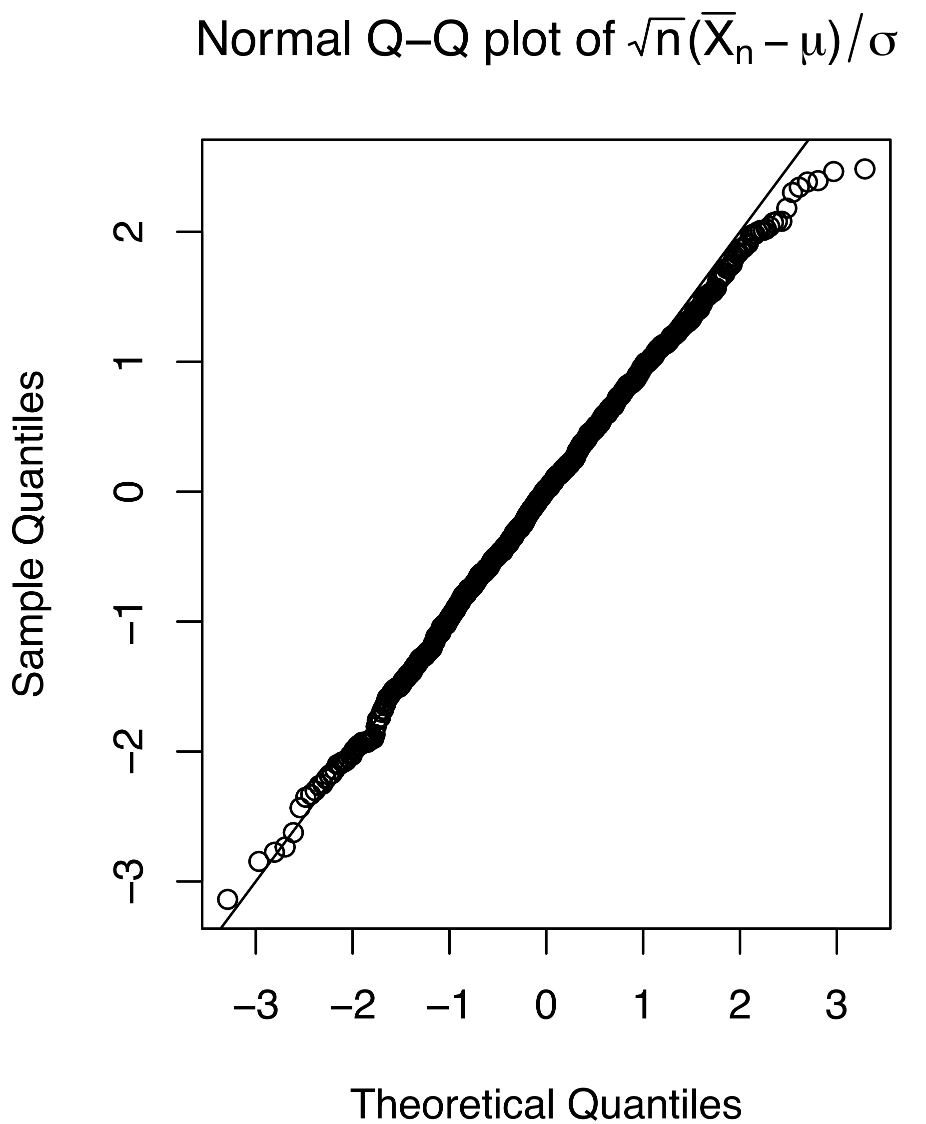


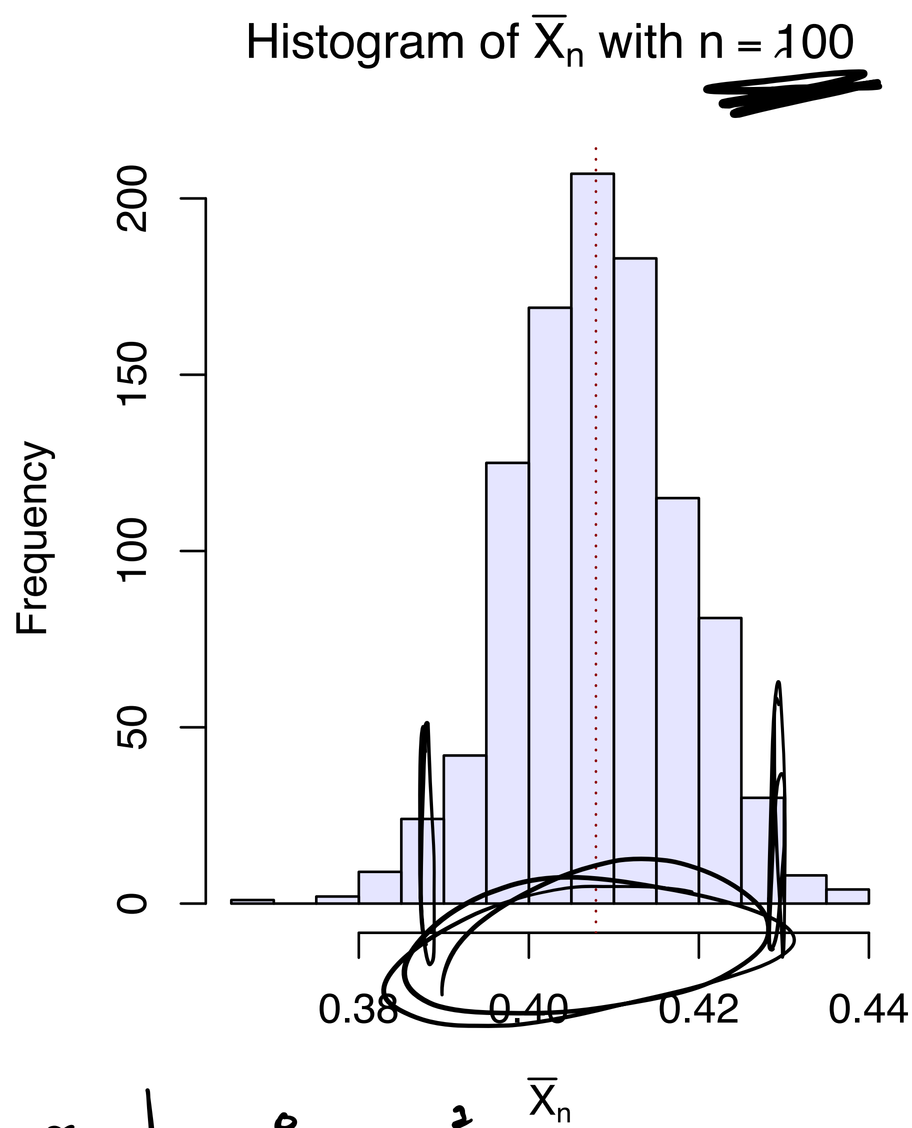
```r
abalones <- read.csv("/Users/karlgregory/Desktop/abalone/abalone.data")

diam <- abalones$V3

mean(diam)
var(diam)

hist(diam)

n <- 5
xbar <- numeric(200)
for(i in 1:200){

  # draw a random sample of size n from the population of diameters:
  X <- sample(diam,n,replace = FALSE)
  xbar[i] <- mean(X)

}

mean(xbar)
var(xbar)
var(diam)/n
```

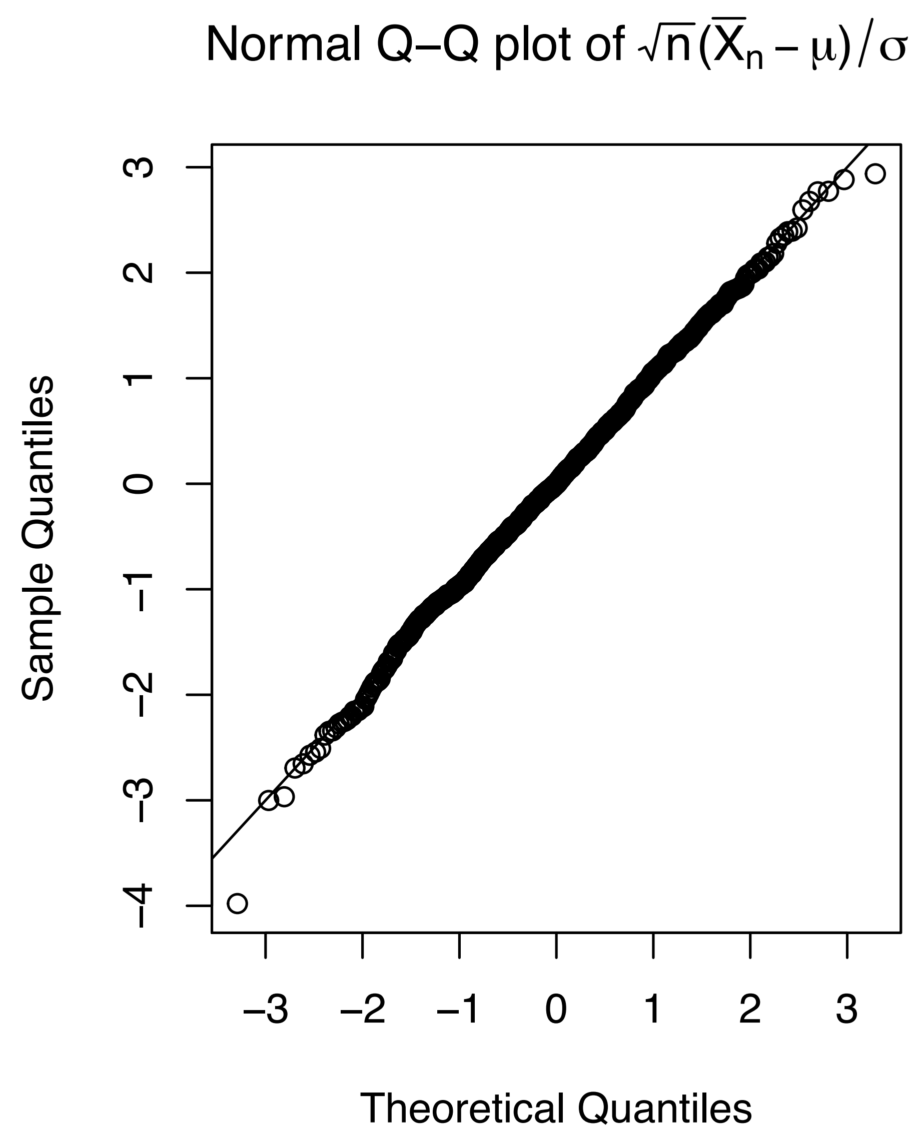

Histogram of $\overline{X}_n$ with $n = 5$
Frequency
0
20
40
60
80
100
0.25
0.30
0.35
0.40
0.45
0.50
$\overline{X}_n$


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





x	0	1
$P(X=x)$	$1-p$	p

$E X = 1 \cdot p + 0(1-p) = p$

$X_1, \dots, X_n \overset{ind}{\sim} \text{Bernoulli}(p)$

$\mu = p$

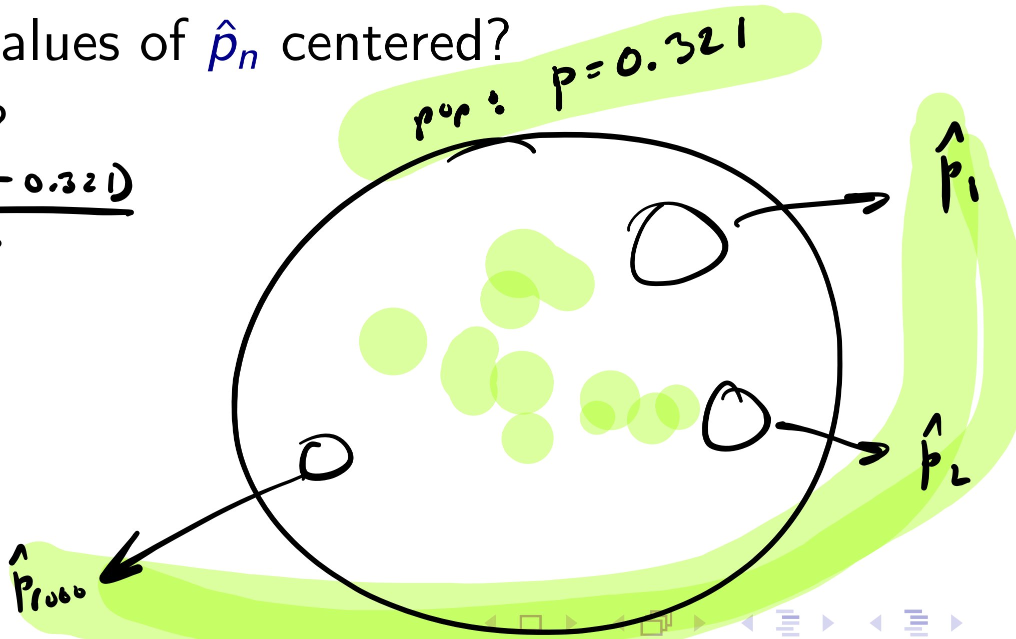
$\sigma^2 = p(1-p)$

$\hat{p}_n = \bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n) = \frac{\#\{1s\}}{n} =$ proportion of 1s.

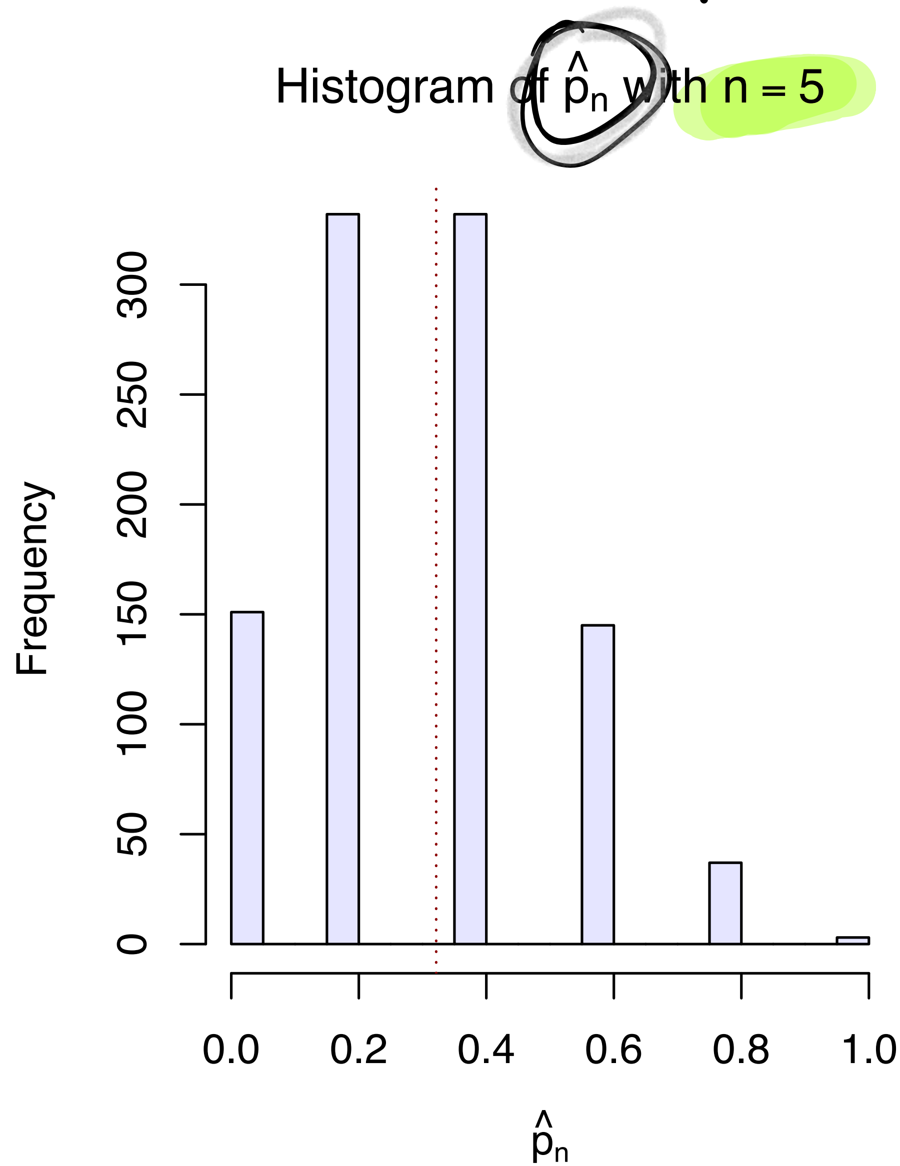
Exercise: Treat the 4,176 abalone as a population. The proportion classified as infants among the abalone is $p = 0.321$; let $\hat{p}_n$ represent the proportion of infants in a random sample of abalone.

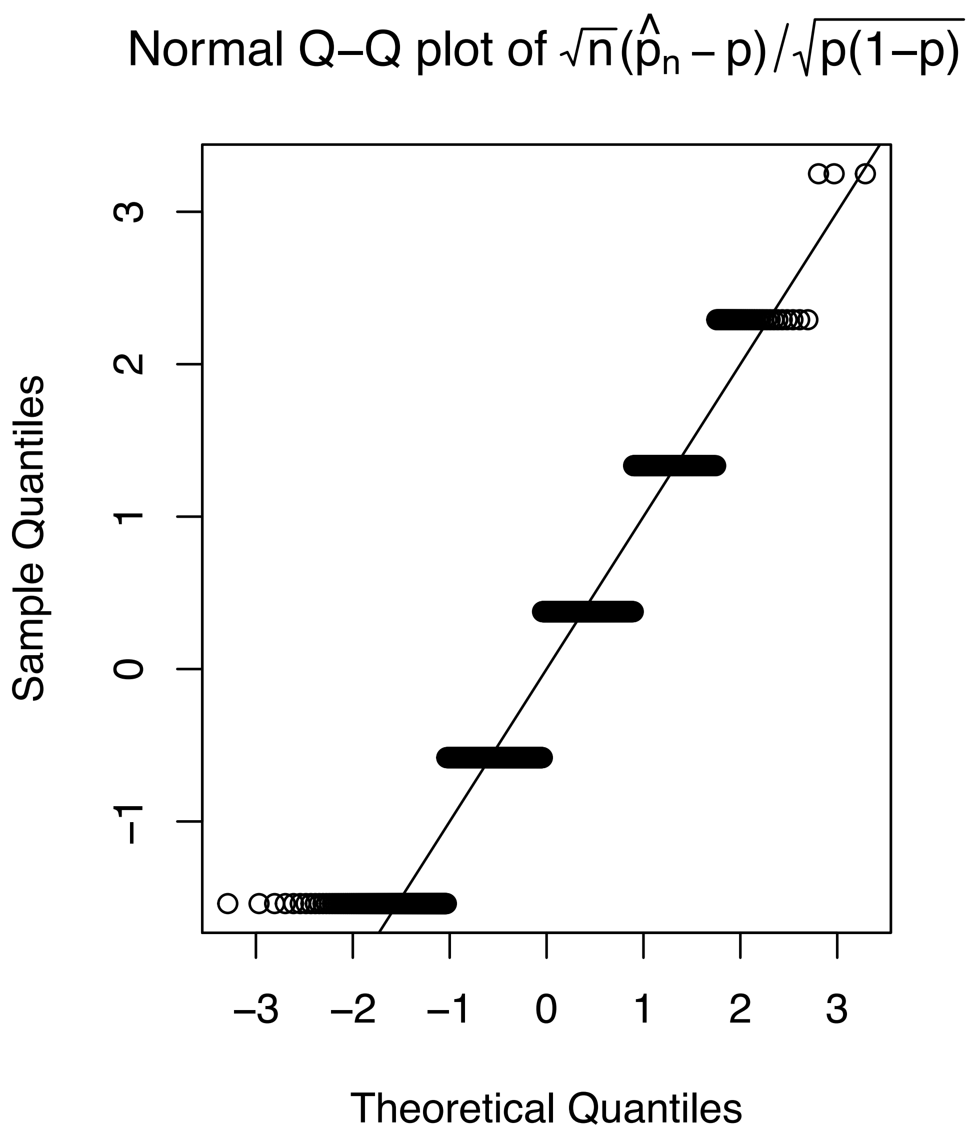
1. For the sample sizes $n = 5, 25, 100$, draw 1,000 samples and
   1. Make a histogram of the $\hat{p}_n$ values.
   2. Make a Normal Q-Q plot of the $\hat{p}_n$.
2. Around what value are the values of $\hat{p}_n$ centered?
3. What changes as n changes?

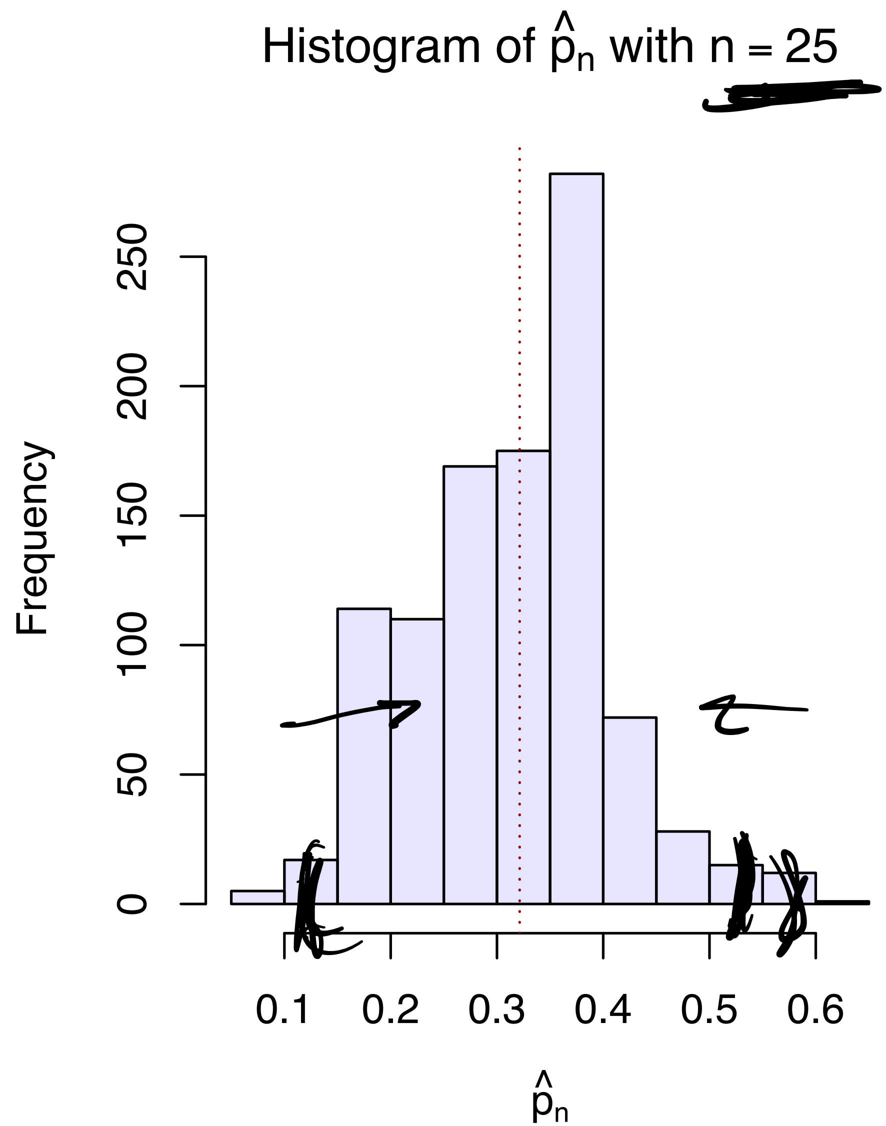
n	Var $\hat{p}_n$	$\frac{p(1-p)}{n} = \frac{0.321(1-0.321)}{n}$
5	0.0416	0.0435
25	0.0083	0.0087

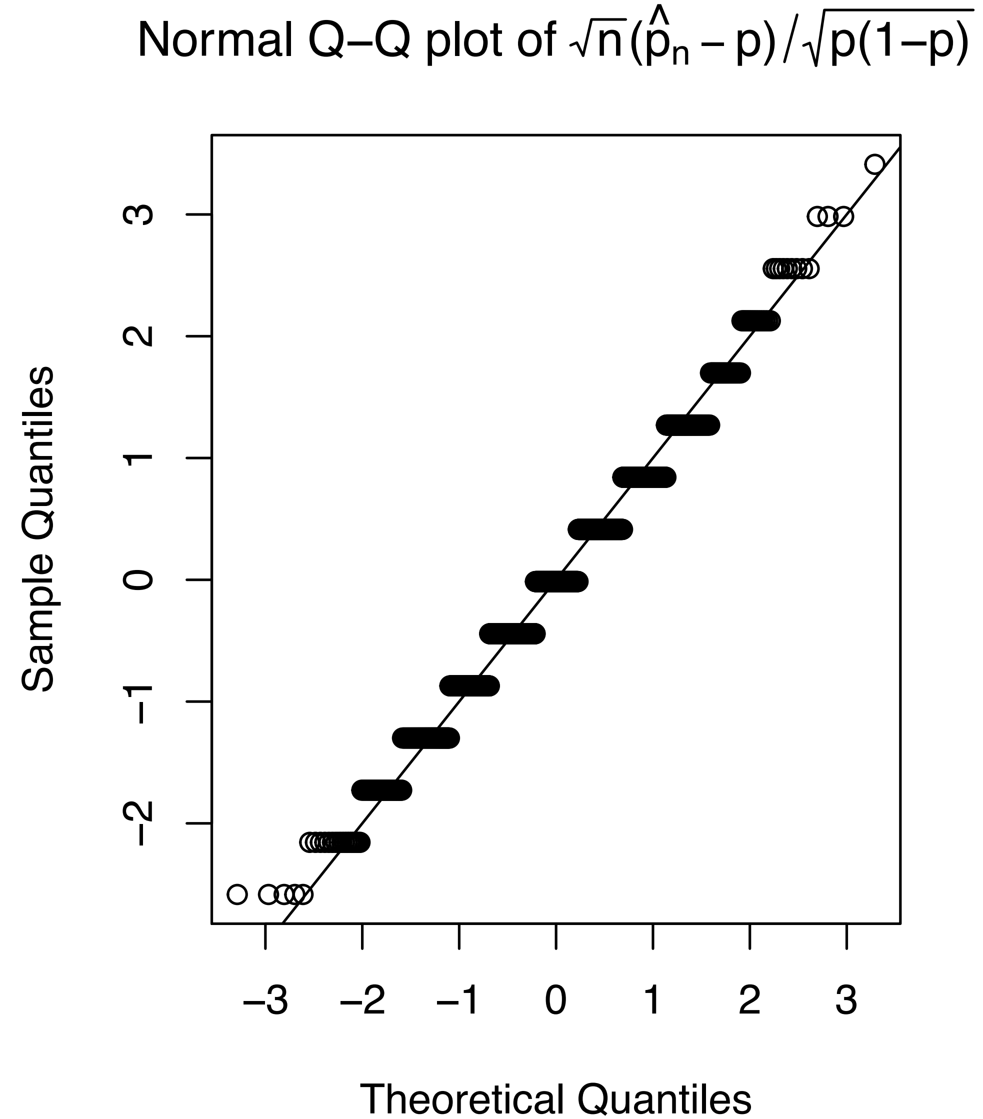


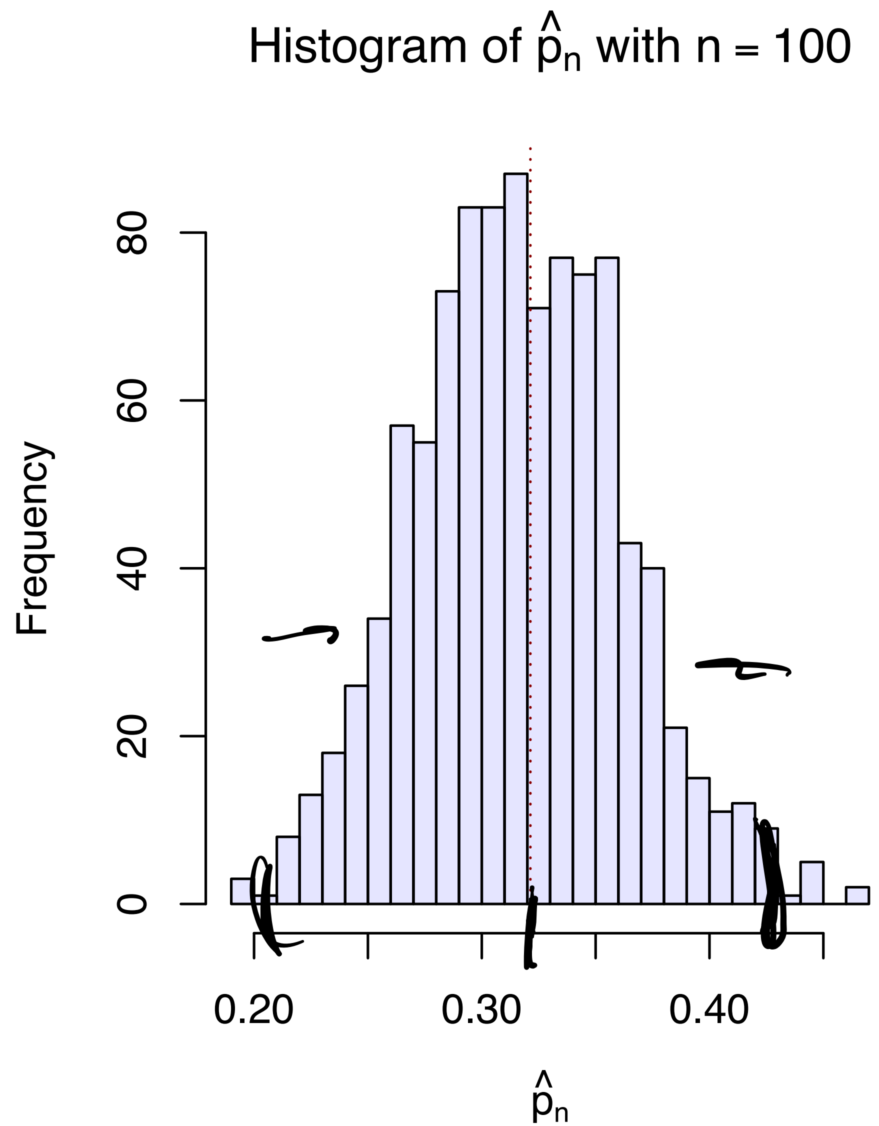
n = 5

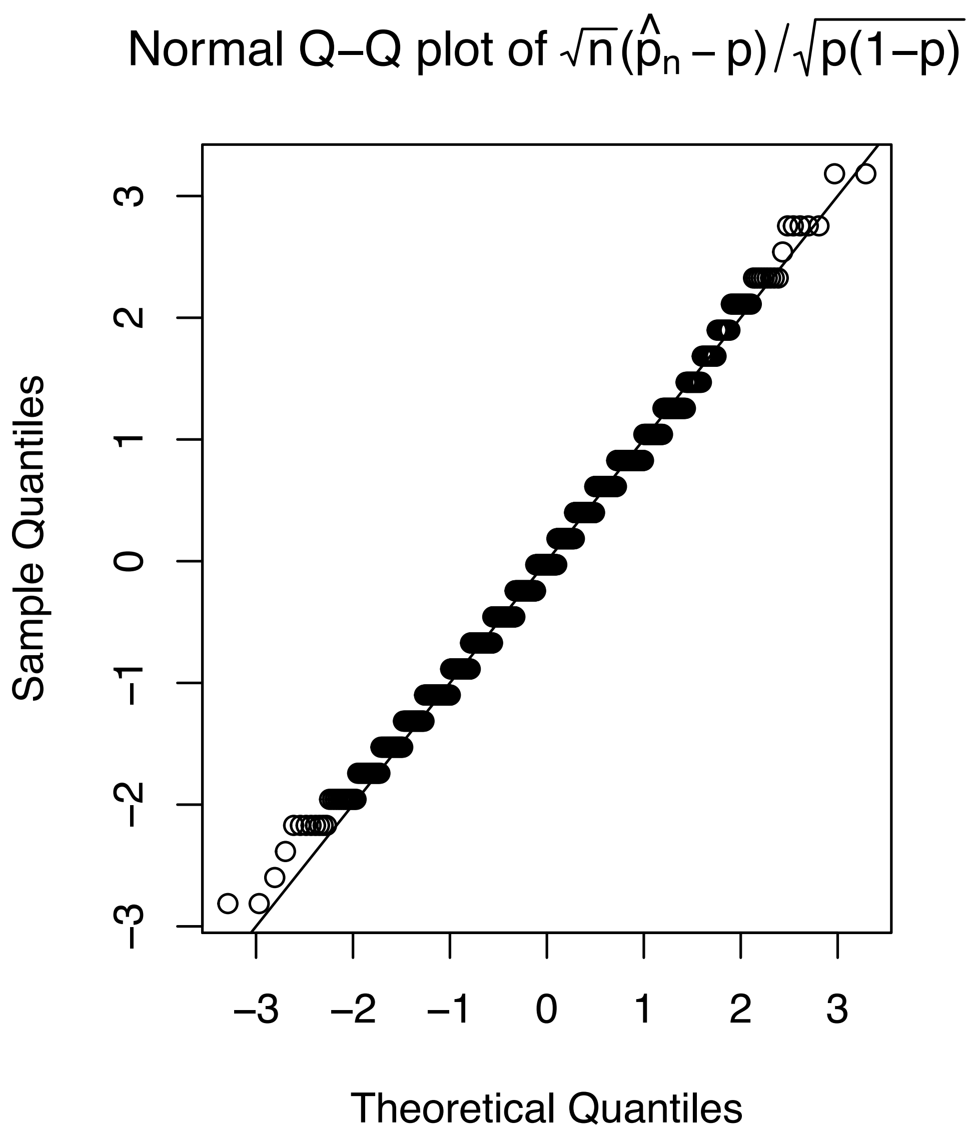

Histogram of $\hat{p}_n$ with n = 5
Frequency
0
50
100
150
200
250
300
0.0
0.2
0.4
0.6
0.8
1.0
$\hat{p}_n$


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

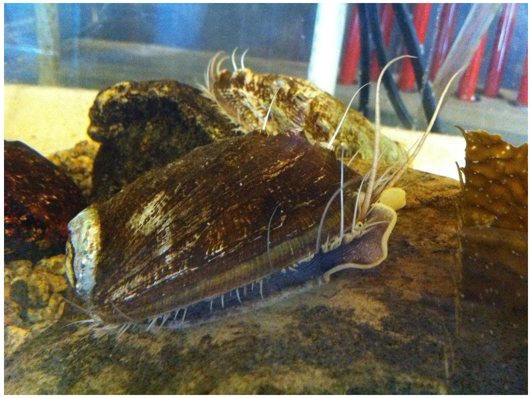

Histogram of p̂n with n = 25
Frequency
0
50
100
150
200
250
0.1
0.2
0.3
0.4
0.5
0.6
p̂n

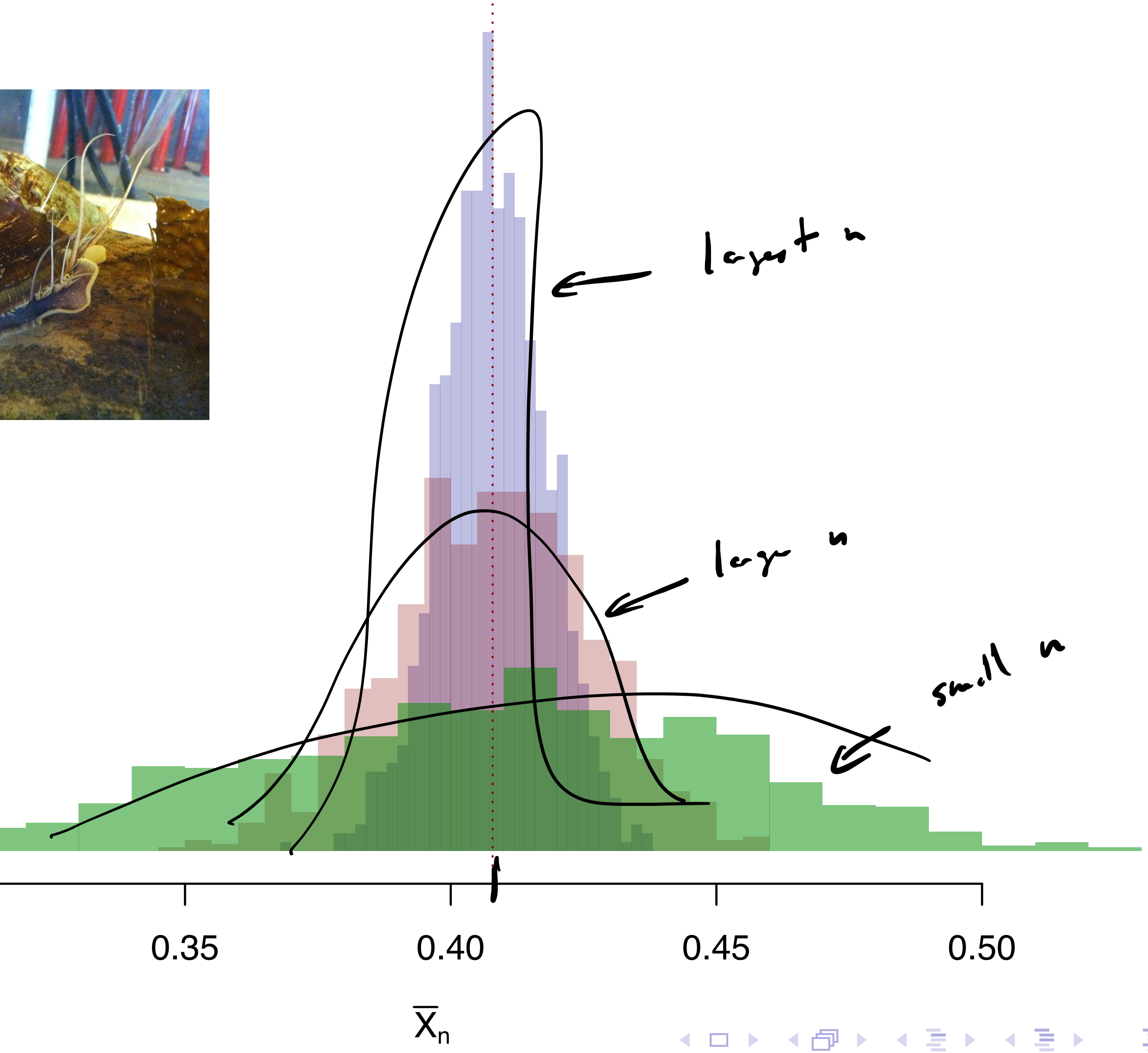

Normal Q–Q plot of √n(p̂n – p)/√p(1–p)
Sample Quantiles
−2
−1
0
1
2
3
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Histogram of $\hat{p}_n$ with n = 100
Frequency
0
20
40
60
80
0.20
0.30
0.40
$\hat{p}_n$


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

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





$\bar{X}_n$ has standard deviation $\frac{\sigma}{\sqrt{n}}$.

Distribution of sample mean when population is Normal

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

$Z = \frac{X-\mu}{\sigma}$

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

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

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

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

$\sigma^2 = 2500$

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

1. Find $P(|X - 450| > 50)$. $= 0.3174$
2. Find $P(X < 425)$. $= 0.3085$

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

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

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

(1) $P(\underbrace{|X-450|}_{\text{distance of } X \text{ from } 450} > 50) = P(X \text{ is more than } 50 \text{ minutes from } 450)$

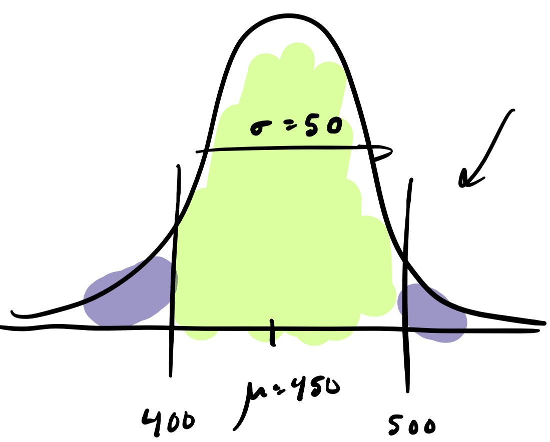
$X \sim \text{Normal}(\mu = 450, \sigma^2 = 50^2)$

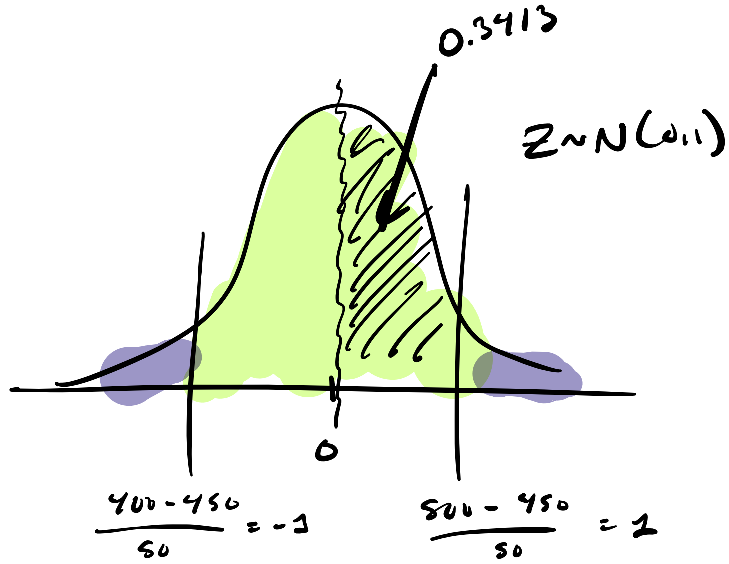
$= 1 - P(|X-450| \le 50)$

$= 1 - P(-50 \le X - 450 \le 50)$

$= 1 - \underbrace{P(400 \le X \le 500)}_{0.6826}$

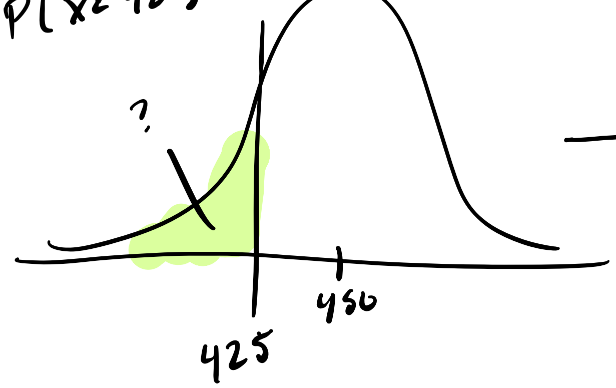
$= .3174$

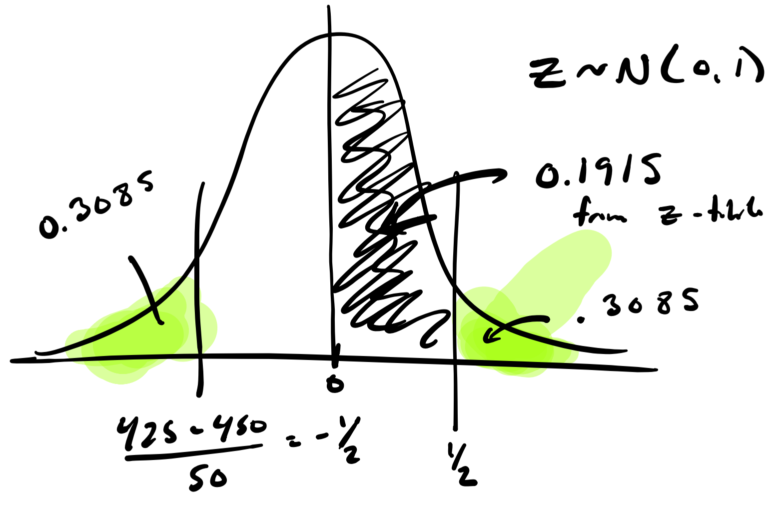




$P(|X-450| > 50) = 2(.3413) = .6826.$

$P(X < 425) = 0.3085$





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

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

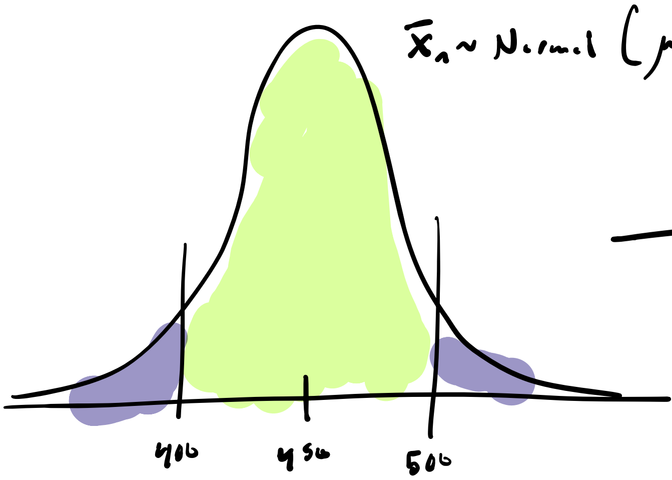
$\bar{X}_n \sim \text{Normal}\left(\mu = 450, \frac{50^2}{9}\right)$

(1) $P(|\bar{X}_n - 450| > 50) = P(\bar{X}_n$ is more than 50 minutes from the mean$)$

$= 1 - P(|\bar{X}_n - 450| \le 50)$

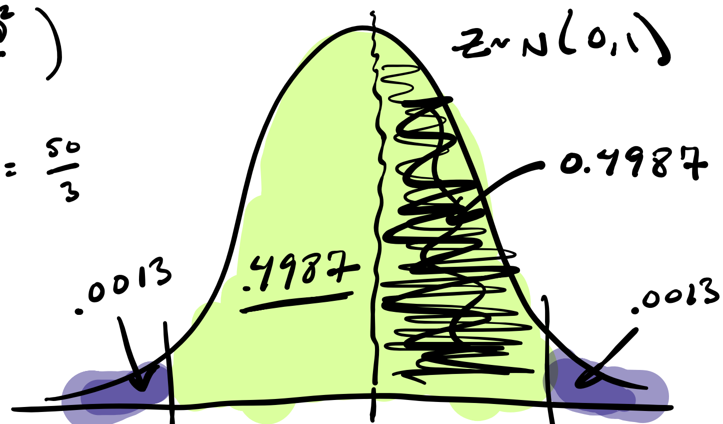
$= 1 - P(400 \le \bar{X}_n \le 500)$

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



$\sqrt{\frac{50^2}{9}} = \frac{50}{3}$

$Z \sim N(0,1)$



$\frac{400-450}{50/3} = -3$

$\frac{500-450}{50/3} = 3$

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

$P(|\bar{X}_n - 450| > 50) = 2(0.0013) = 0.0026.$

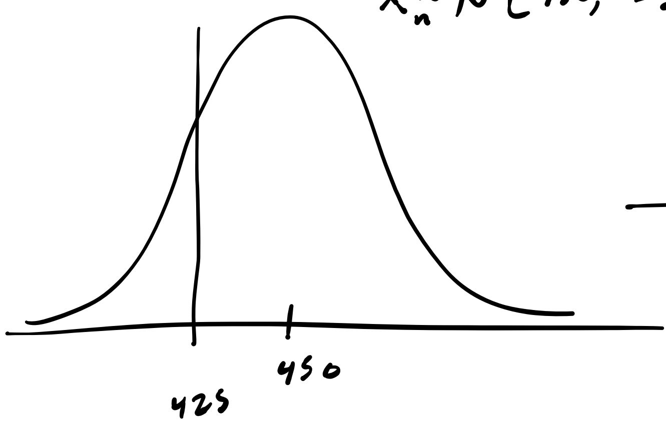
$\mu = 450$

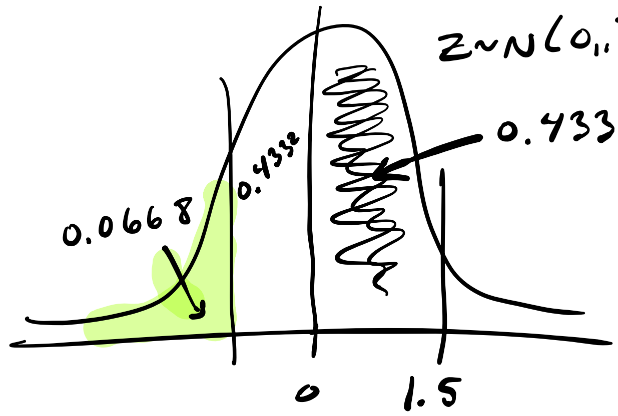
② $P(\bar{X}_n < 425) = 0.0668.$

Before

$P(X < 425) = 0.3085$

$\bar{X}_n \sim N\left(450, \frac{50^2}{9}\right)$

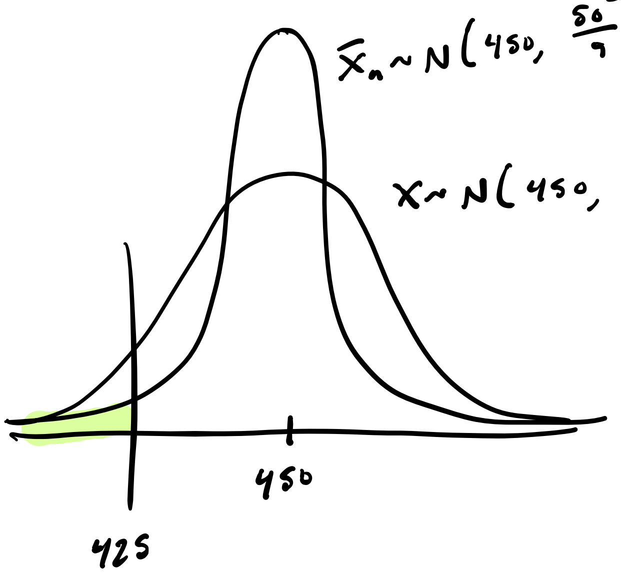




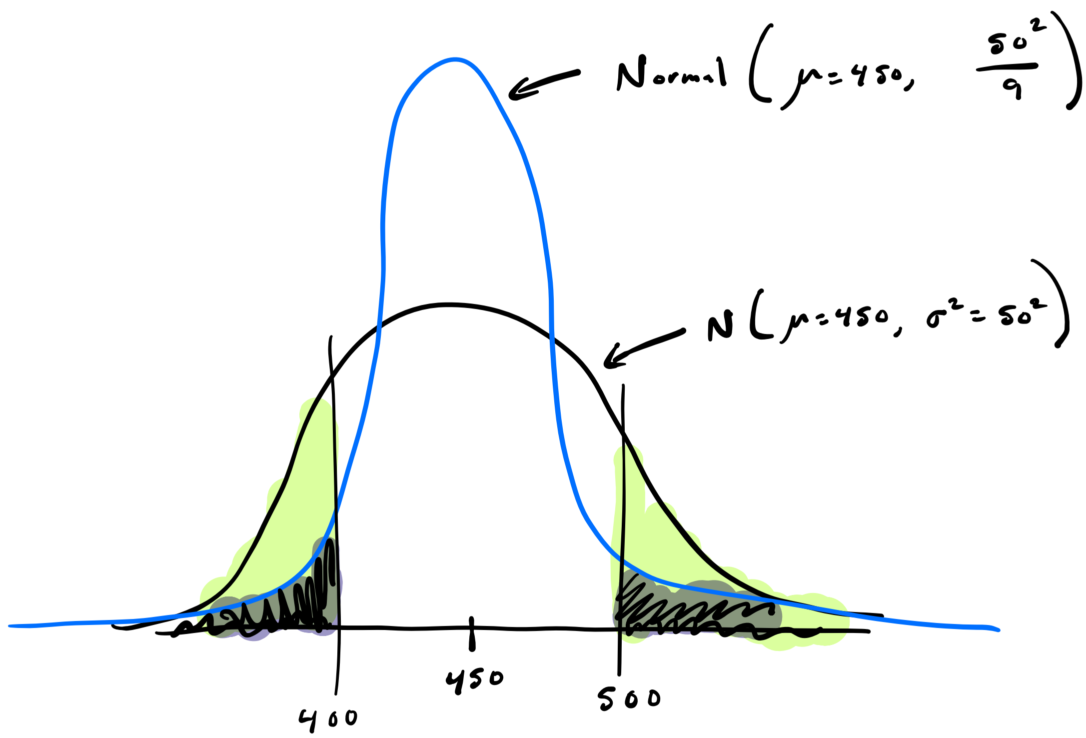
$$\frac{425-450}{50/3} = -\frac{3}{2} = -1.5$$

$\bar{X}_n \sim N\left(450, \frac{50^2}{9}\right)$

$X \sim N(450, \sigma^2 = 50^2)$



$$P(|X - 450| > 50) = 0.3174$$

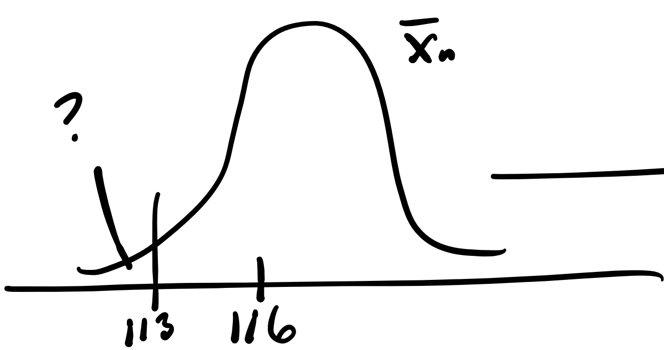


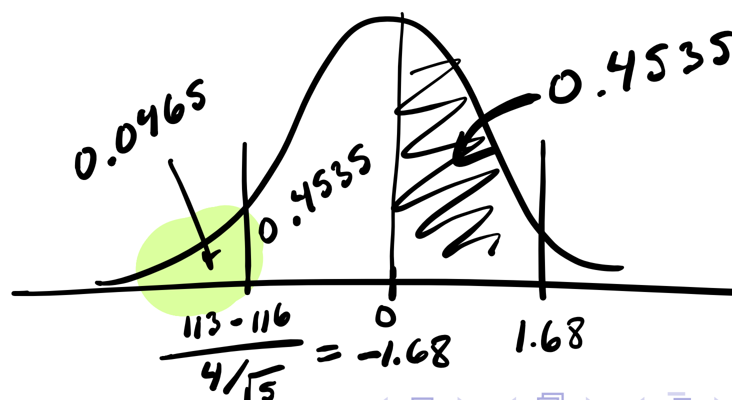
Exercise: You sell jars of baby food labelled as weighing 4oz $\approx$ 113g. Suppose your process results in jar weights with the Normal($\mu = 116, \sigma^2 = 4^2$) distribution. A regulator will sample 5 jars and fine you if the average weight is less than 113g.

1. With what probability will you get fined?
2. To what must you increase μ so that you are fined with prob. at most 0.01?
3. Keeping $\mu = 116$g, to what must you reduce σ so that you are fined with probability at most 0.01?

① $P(\bar{X}_n < 113)$, $n = 5$

$\bar{X}_n \sim N\left(116, \frac{4^2}{5}\right)$



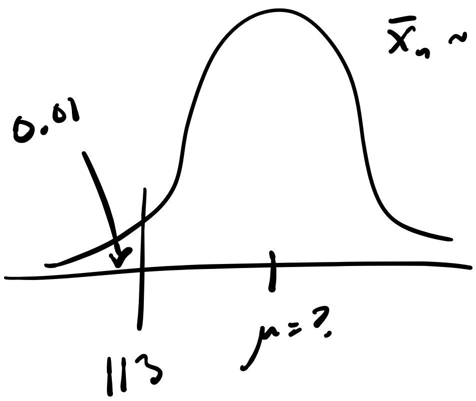


$\sqrt{\frac{4^2}{5}} = \frac{4}{\sqrt{5}}$

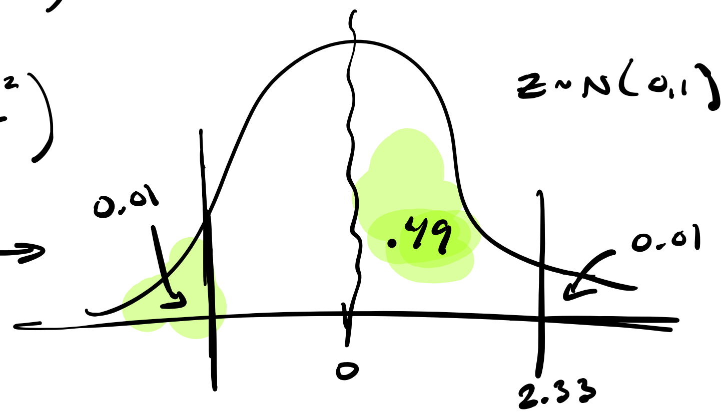
$P(\bar{X}_n < 113) = 0.0465.$

② $P(\bar{X}_n < 113) \overset{set}{=} 0.01$, μ is unknown.

$\bar{X}_n \sim N\left(\mu = ?, \frac{4^2}{5}\right)$



→



$$\frac{113-\mu}{4/\sqrt{5}} = z_{0.01} = -2.33$$

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

$$\frac{113-\mu}{4/\sqrt{5}} = -2.33$$

$\iff$

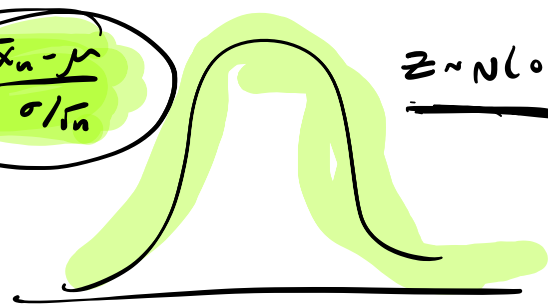
$$113 - \mu = (-2.33)\frac{4}{\sqrt{5}}$$

$\iff$ $113 + 2.33\frac{4}{\sqrt{5}} = \mu$

$\iff$ $\mu = 117.17.$

$\cancel{\bar{X}_n \sim N\left(\ , \frac{\sigma^2}{n}\right)}$

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



$Z \sim N(0,1)$

If not normal, but n is large, can still do the z stuff.

Central Limit Theorem

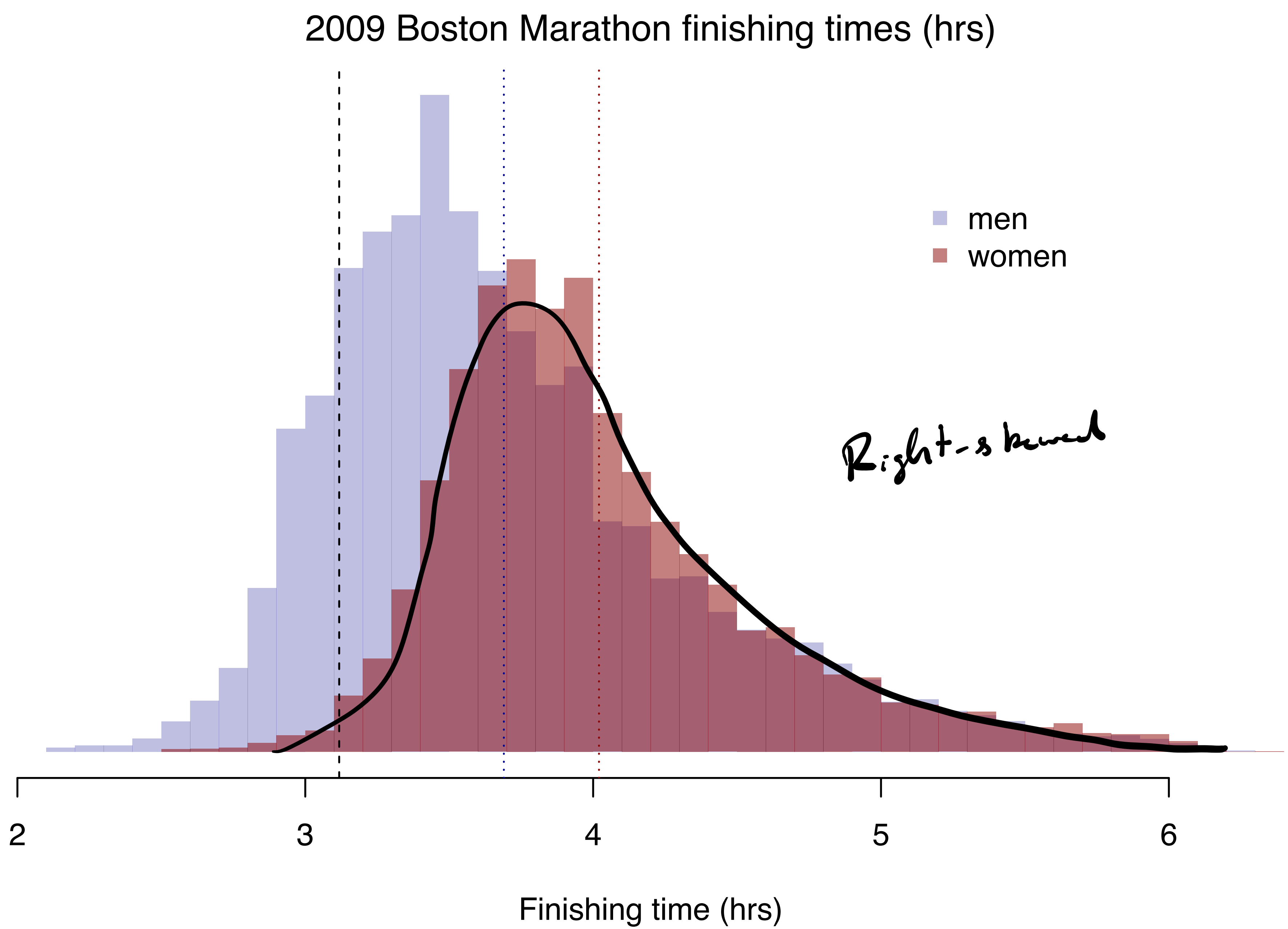
Let $X_1, \ldots, X_n$ be a rs from a dist. with mean μ and variance $\sigma^2 < \infty$. Then

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

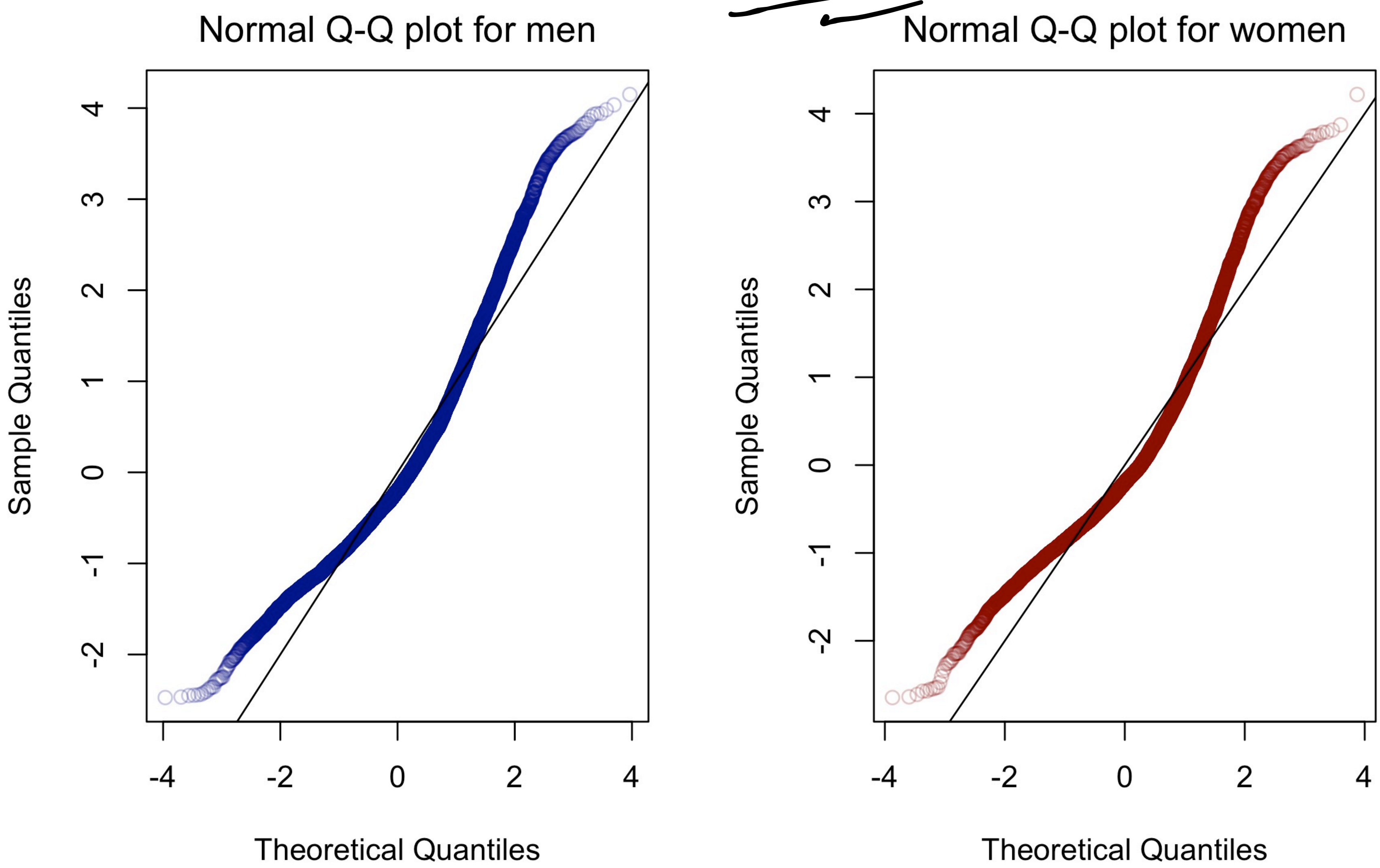
for larger and larger n.

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
Right-skewed
2
3
4
5
6
Finishing time (hrs)

Not Normal.


Normal Q-Q plot for men
Sample Quantiles
Theoretical Quantiles
Normal Q-Q plot for women
Sample Quantiles
Theoretical Quantiles

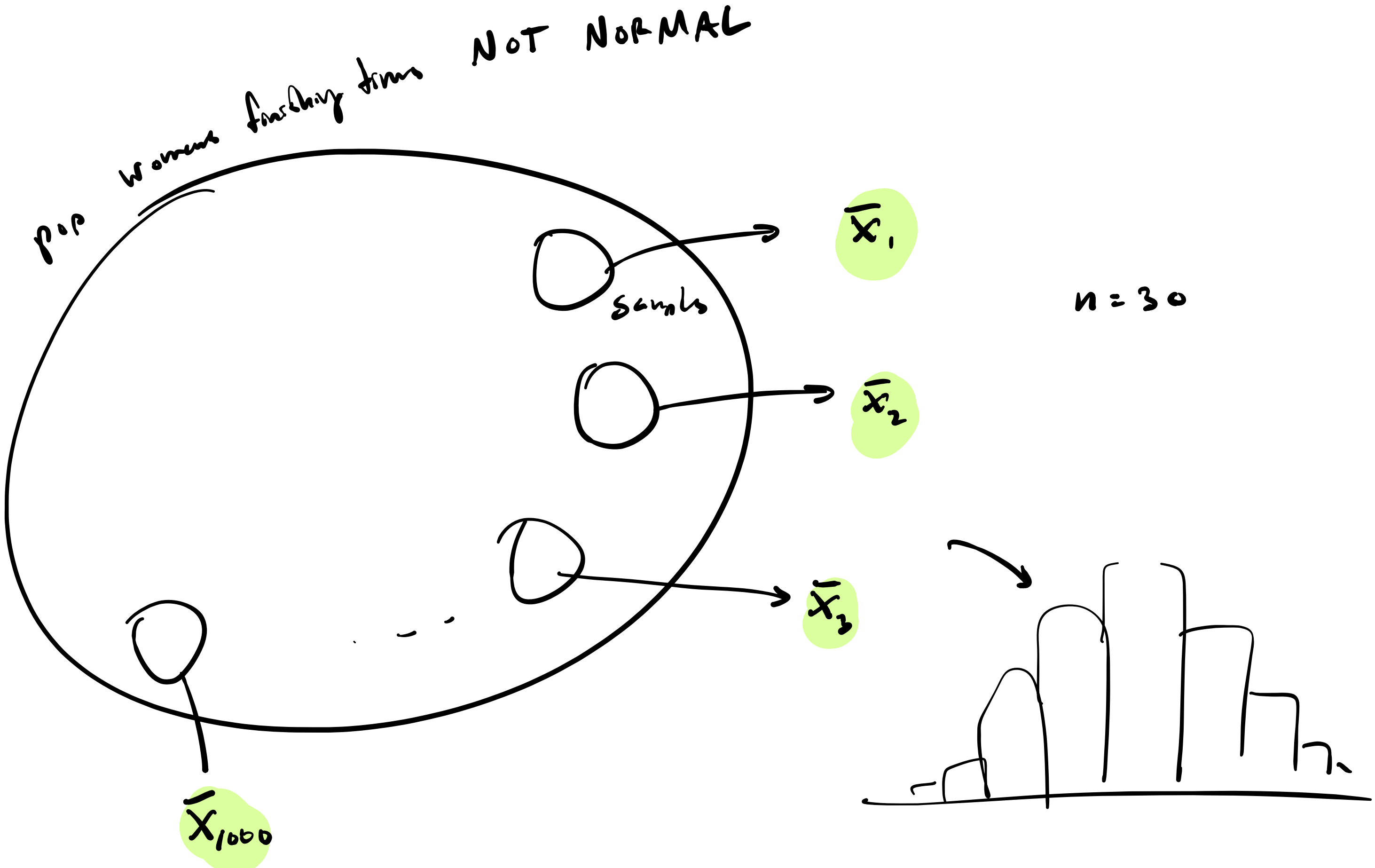
Exercise: Women's finishing times for the 2009 Boston Marathon had mean 4.02 hours and standard deviation 0.555 hours.

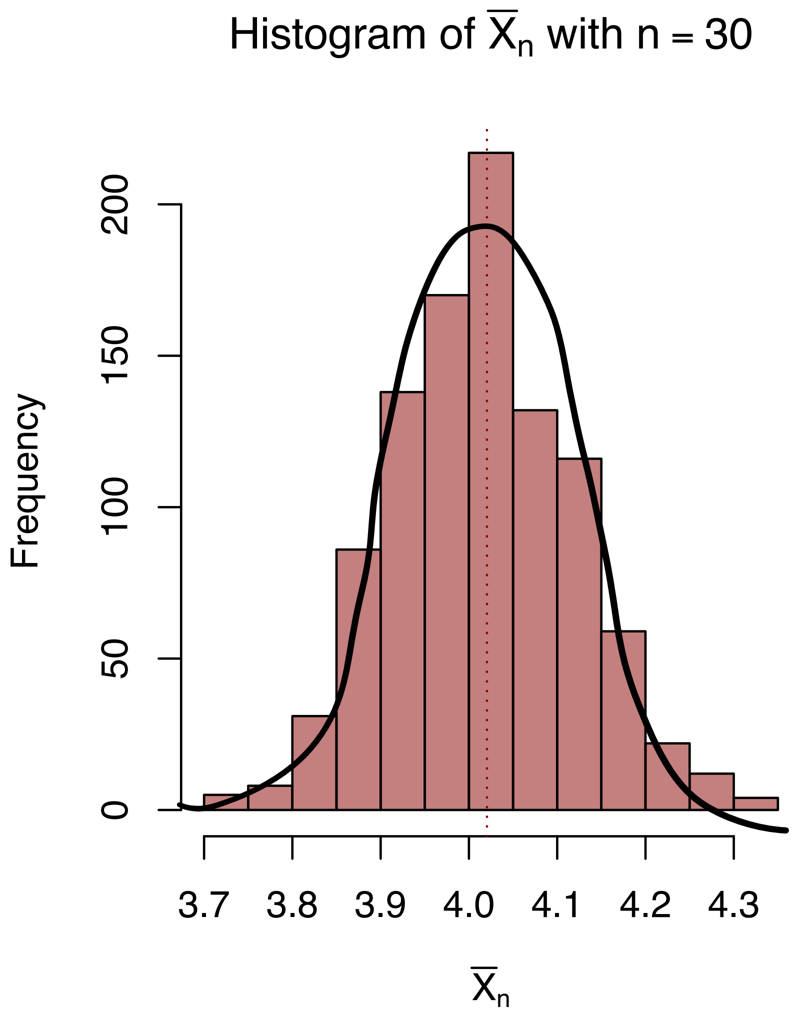
Consider sampling $n = 30$ women and let $\bar{X}_n$ be the mean of their finishing times.

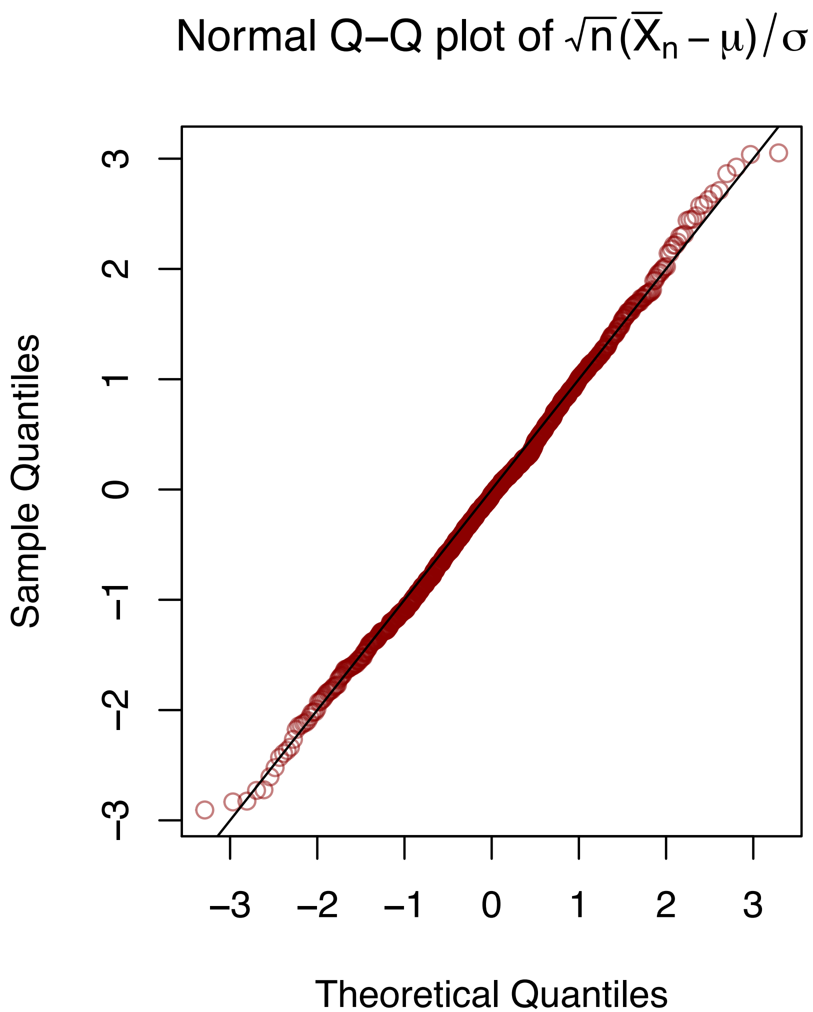
1. Find an approximation to $P(\bar{X}_n < 3.90)$.
2. Find an approximation to $P(\bar{X}_n > 4.25)$.
3. Find an approximation to $P(|\bar{X}_n - 4.02| < 0.2)$.

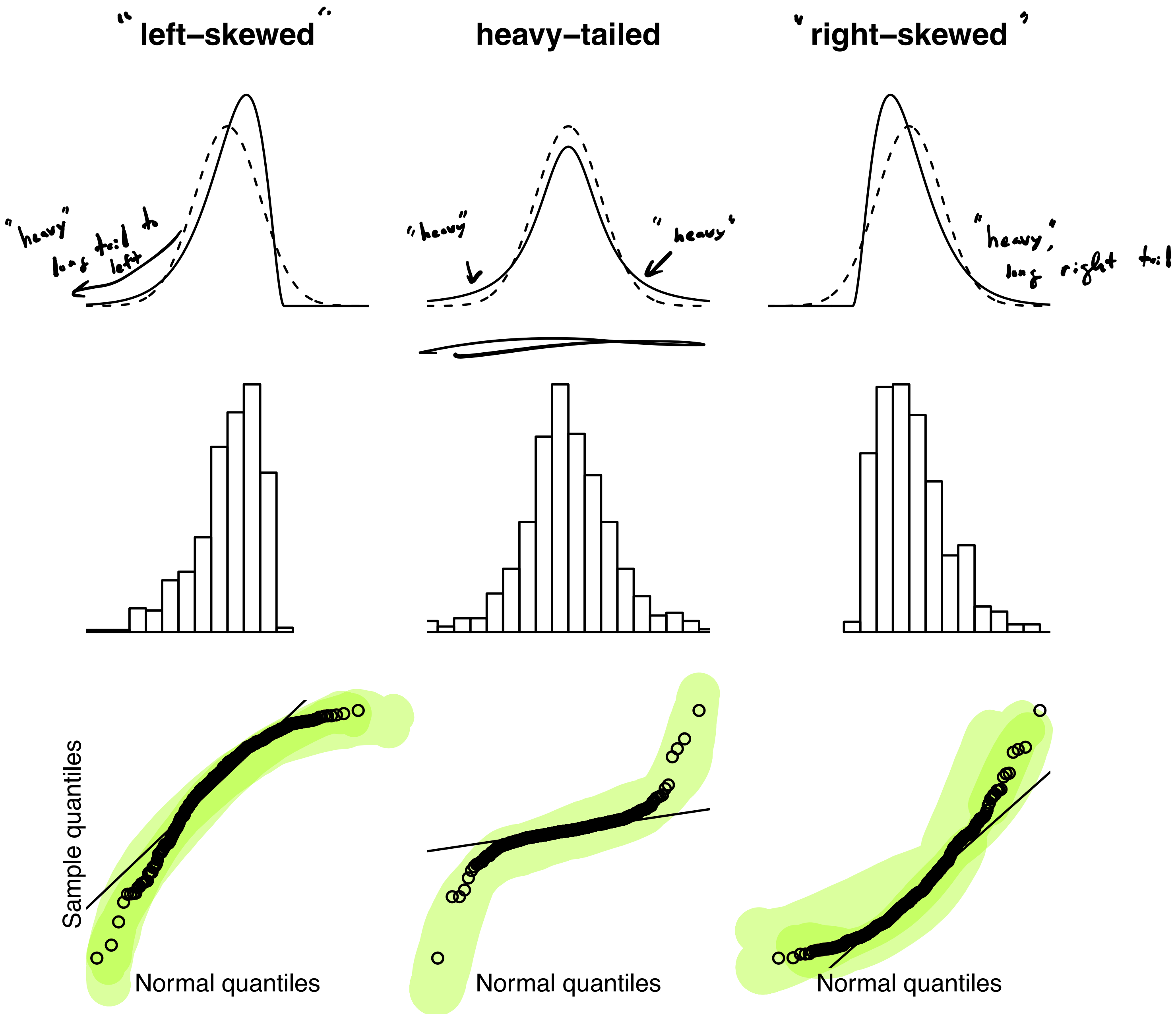
Now use R to draw 1,000 samples of size $n = 30$. `link to women's data`.

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


pop womens finishing times
NOT NORMAL
samples
$\bar{x}_1$
$\bar{x}_2$
$\bar{x}_3$
$\bar{x}_{1000}$
n = 30


Histogram of $\bar{X}_n$ with n = 30
Frequency
0
50
100
150
200
3.7
3.8
3.9
4.0
4.1
4.2
4.3
$\bar{X}_n$


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


"left–skewed"
heavy–tailed
"right–skewed"
"heavy"
long tail to left
"heavy"
"heavy"
"heavy"
long right tail
Sample quantiles
Normal quantiles
Normal quantiles
Normal quantiles

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \overset{\text{approx}}{\sim} Z \sim N(0,1) \quad \text{and} \quad \bar{X}_n \overset{\text{approx}}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right).$$

$$X \sim \text{Bernoulli}(p) \Rightarrow \mathbb{E}X = \underset{\uparrow \atop \mu}{p}, \quad \text{Var}\, X = \underbrace{p(1-p)}_{\uparrow \atop \sigma^2}.$$

Central Lim. Thm

$$\frac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{\text{approx}}{\sim} Z \sim N(0,1)$$

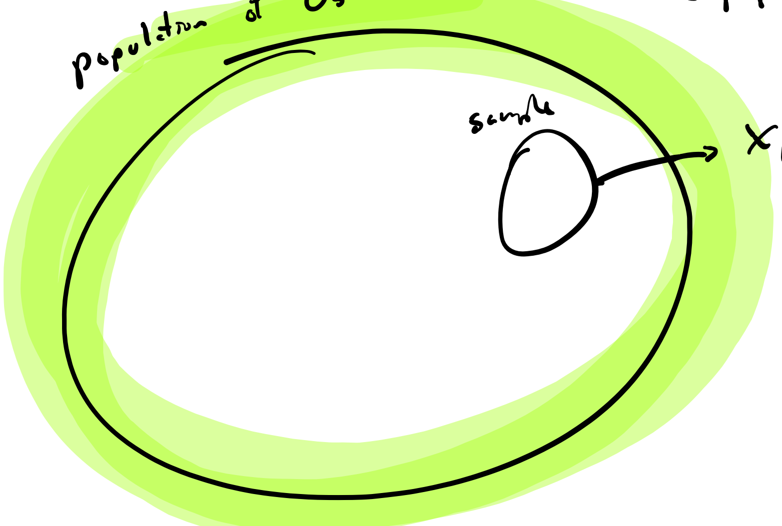
for large n.

Also

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

p is proportion of 1s in the population.

population of 0s and 1s.



$X_1, \ldots X_n$

$$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n}$$

$=$ proportion of 1s

$= \hat{p}_n$

We can apply the Central Limit theorem to proportions.

Central Lim Thm

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \longrightarrow Z \sim N(0,1)$$

as $n \to \infty$.

Central Limit Theorem for the sample proportion

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

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

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\hat{p}_n - p}{\sqrt{\underbrace{p(1-p)}_{\sigma^2}/n}}$$

behaves more and more like $Z \sim$ Normal$(0,1)$ for larger and larger n.

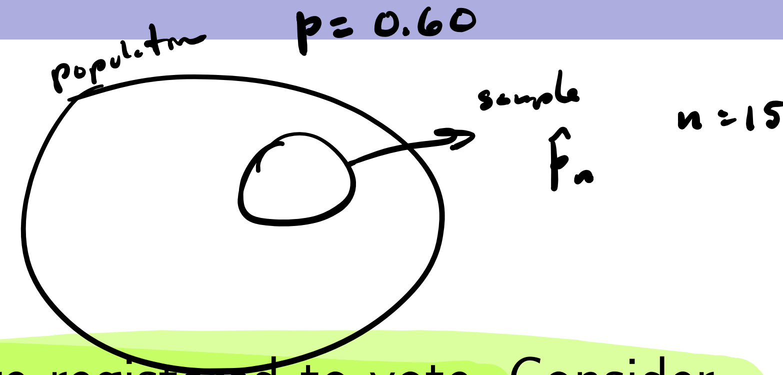
This means that for large n (say $np \geq 5$ and $n(1-p) \geq 5$), we have

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

for large n.

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

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

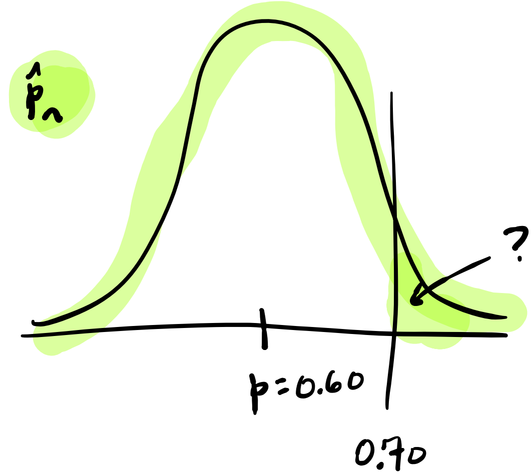


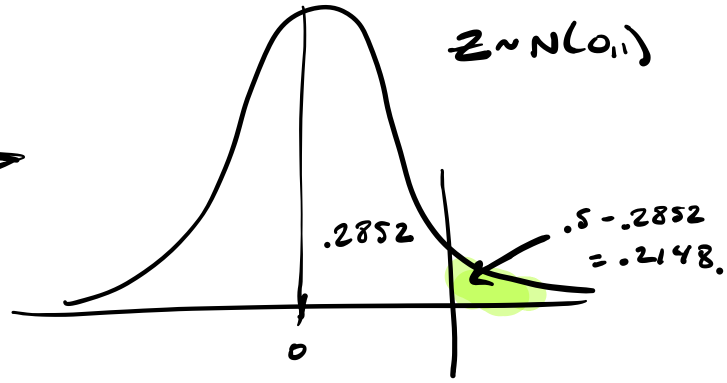
n = 15 $\quad X_1, \dots, X_n \overset{ind}{\sim} \text{Bernoulli}(p = 0.60)$

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) $P(\hat{p}_n > .70) \approx .2148$





$$\frac{.70 - 0.60}{\sqrt{\frac{.6(1-.6)}{15}}} = 0.79$$

$$z = \frac{X - \mu}{\sigma} \qquad z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \qquad z = \frac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}}$$

② Find $P(\hat{p}_n > 0.70)$ using Binomial.

$$\hat{p}_n = \frac{X_1 + \dots + X_n}{n} = \frac{\#\{\text{successes}\}}{n}$$

$$\iff \quad n\hat{p}_n = \#\{\text{successes}\} \sim \text{Binomial}(n = 15, \ p = 0.60)$$

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

$$= P(\underbrace{15 \cdot \hat{p}_n} > 15(.7))$$

15(.7) 10.5

$$= P(Y > 10.5) \qquad \underline{Y \sim \text{Binomial}(n=15, \ p=0.60)}$$

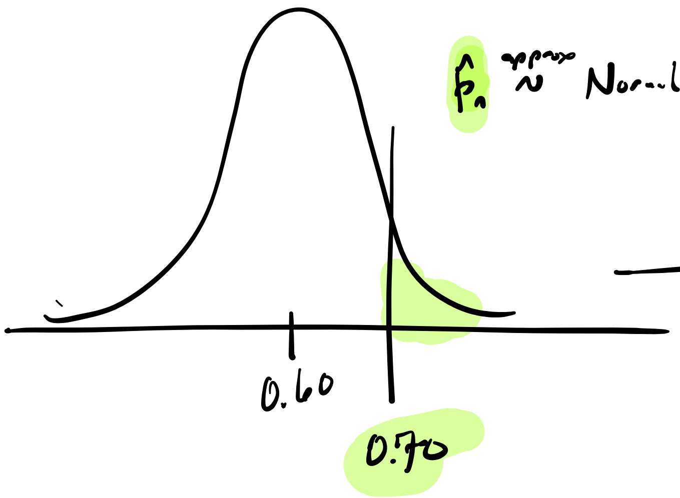
$$= P(Y > 10)$$

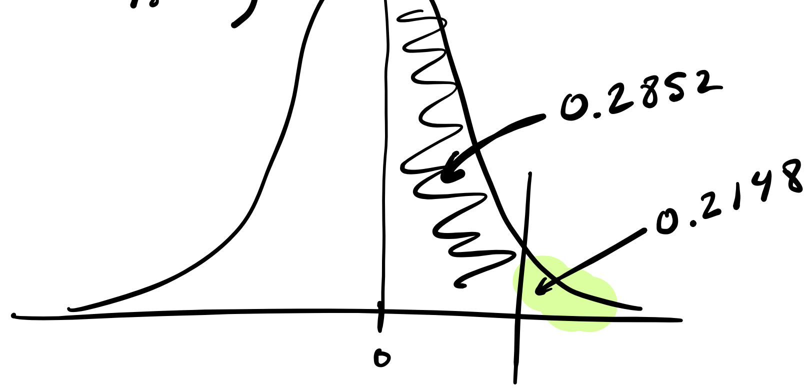
$$= 1 - P(Y \le 10)$$

$$= 1 - \text{pbinom}(\underline{10}, 15, 0.60)$$

$$= 0.2173$$

$\hat{p}_n \overset{\text{approx}}{\sim} \text{Normal}\left(0.60, \frac{0.6(1-0.6)}{15}\right)$ $\qquad Z \sim N(0,1)$





$$\frac{0.70 - 0.60}{\sqrt{\frac{0.60(1-0.60)}{15}}} = 0.79$$

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

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

② Find $P(\hat{p}_n > 0.7)$ using binomial.

$$P(\hat{p}_n > 0.7) = P(\bar{X}_n > 0.7)$$

$$= P\left(\frac{X_1 + \dots + X_n}{n} > 0.7\right)$$

$$= P\big(\underbrace{X_1 + \dots + X_n}_{\#\text{ successes in } n \text{ trials}} > n(0.7)\big)$$

$\sim \text{Binomial}(n, 0.7)$

GRADE DISTRIBUTION

Greater than 100	4
90 - 100	4
80 - 89	8
70 - 79	4
60 - 69	3
50 - 59	4
40 - 49	3
30 - 39	2
20 - 29	2
10 - 19	0
0 - 9	0
Less than 0	0

14 - 15

Workflow for success

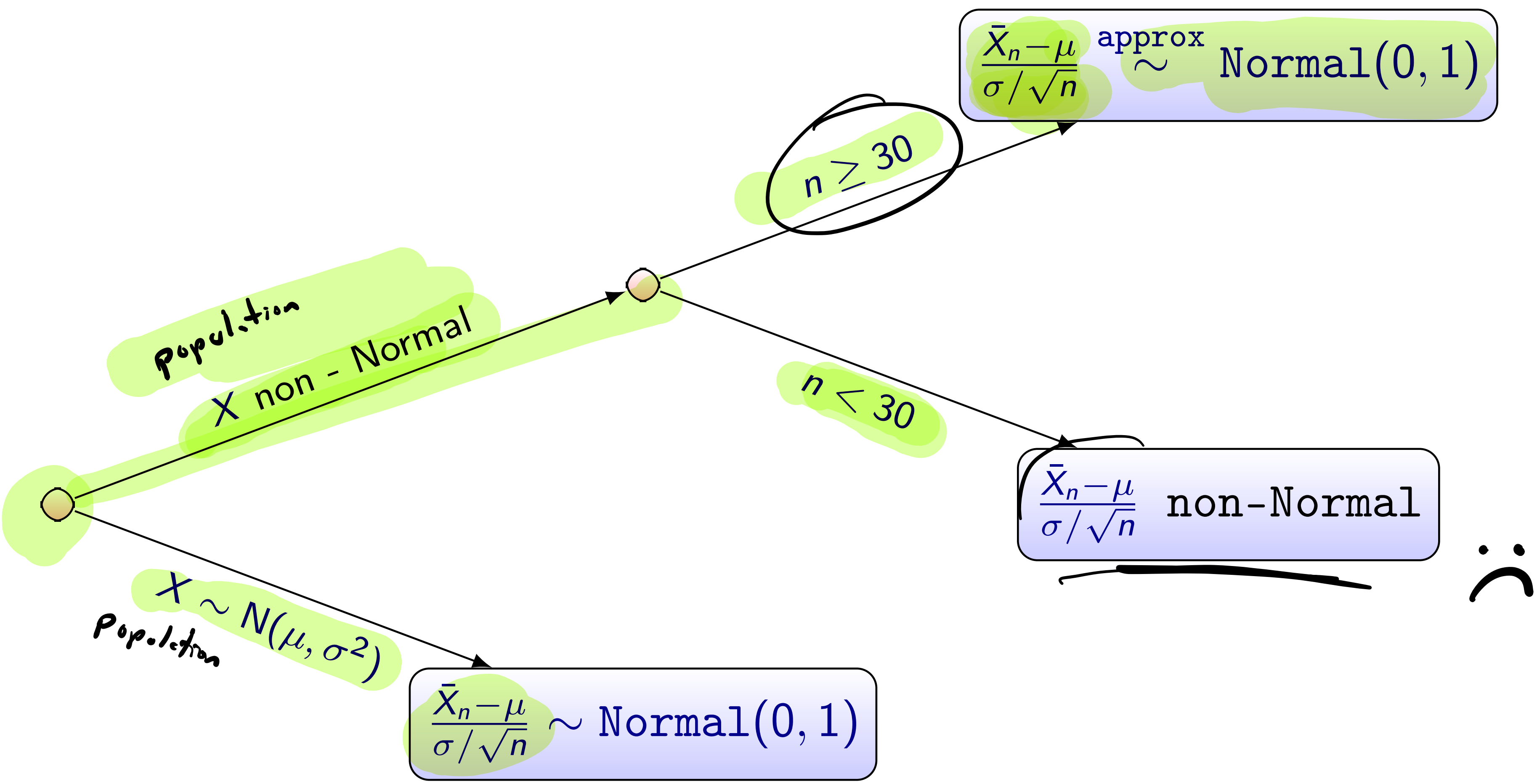
2.5

M - 30 minutes - redo all examples from class.

W - 30 min.

If needed read notes.

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



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

if n is small

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

1s in the samples

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

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

if n is large

From central limit theorem.

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

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

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