STAT 22000 Lecture Slides
Variability in Estimates & Central Limit Theorem

Yibi Huang
Department of Statistics
University of Chicago

Outline

This set of slides covers section 4.1 and 4.4 in the text, which includes

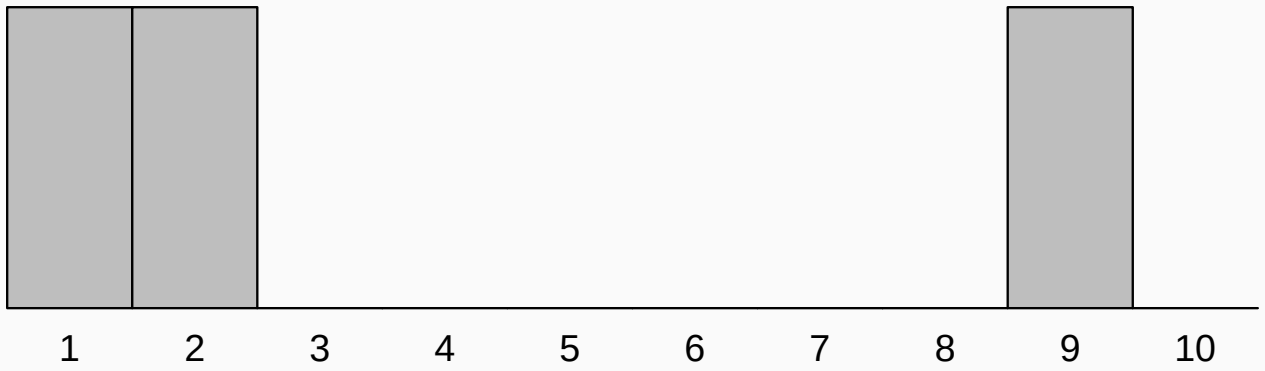
- Central Limit Theorem (CLT)
- Sampling distribution

Example — Rating of a Movie

Suppose a certain movie has a bipolar distribution of ratings, that in a 1 to 10 scale, of those having watched the movie, 1/3 gave 9 points, 1/3 gave 2 points, and the remaining 1/3 gave 1 points.

So the population distribution is

X	1	2	9
$P(X)$	1/3	1/3	1/3



Population Distribution

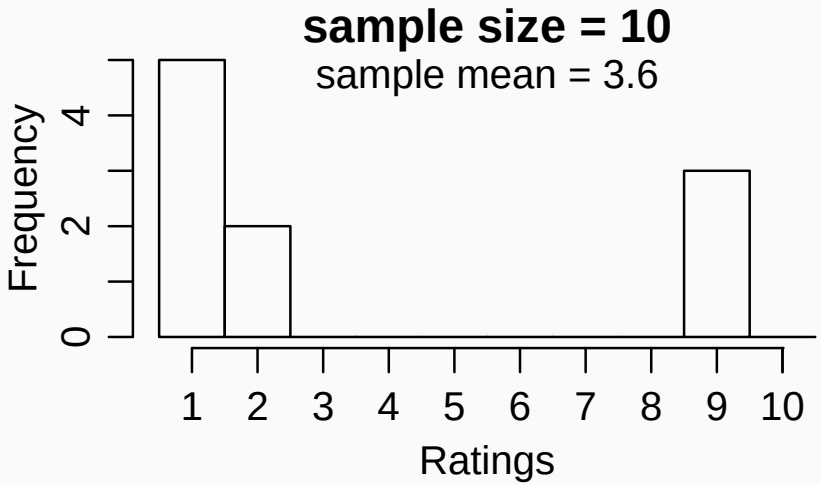
Histogram of the Sample

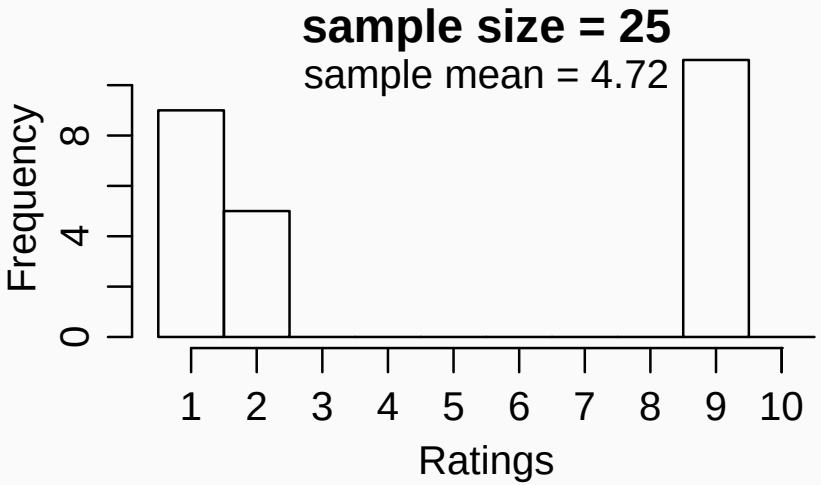
In practice, since the population are difficult (or impossible) to examine completely, we take a sample to learn about the population. Will the makeup of the sample mimic the makeup of the population?

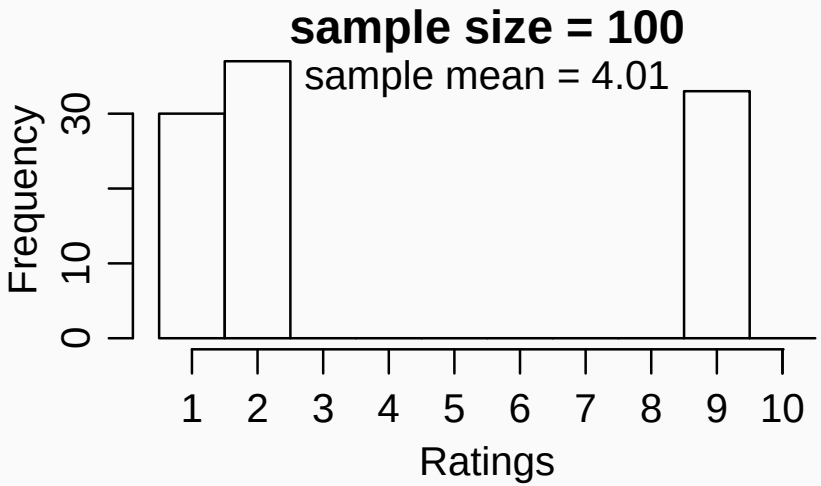
First, the sampling method must be appropriate. A biased sample won't give us the correct information about the population.

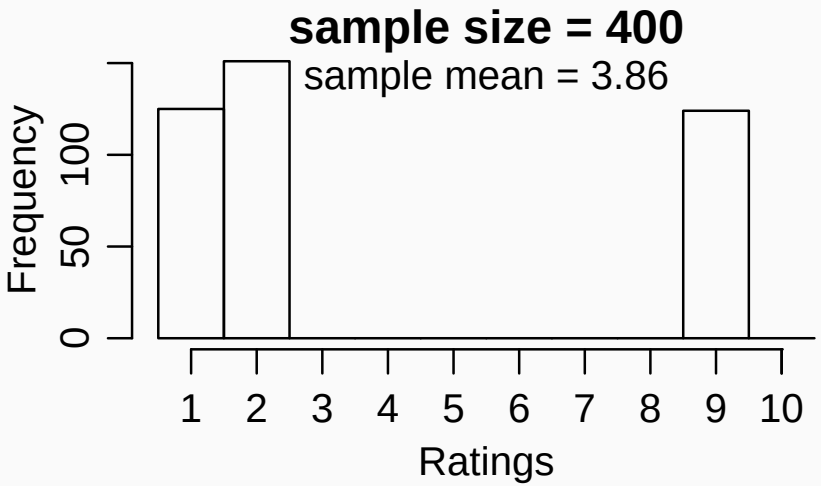
Suppose we take a **simple random sample** of size n (say $n = 400$) from the population. What will the histogram of the ratings of the movie given by subjects in the sample look like?

```
popratings = c(1,2,9)
s400 = sample(popratings, size = 400, replace=T, prob=c(1/3,1/3,1/3))
hist(s400, breaks=0:10+.5, xlab="Ratings", main="Sample Size = 400")
```









The histogram of the sample looks somewhat like the histogram of the population. The larger the sample size, the higher the resemblance.

Estimation of the Population Mean

In practice, the population distribution is usually unknown. We are often interested in *population parameters*, like the *population mean*.

- As all we know about the population is the sample, we can only use the sample to estimate the population parameter of interest, called *statistic*.
- A commonly used estimate of the population mean is the sample mean. Thus the sample mean is one of such statistic.
- Sample statistics vary from sample to sample.
- How close is the sample mean to the population mean?

Variability of the Sample Means

To know the variability of the sample mean of a sample of size $n = 25$, we pretend that we know the population

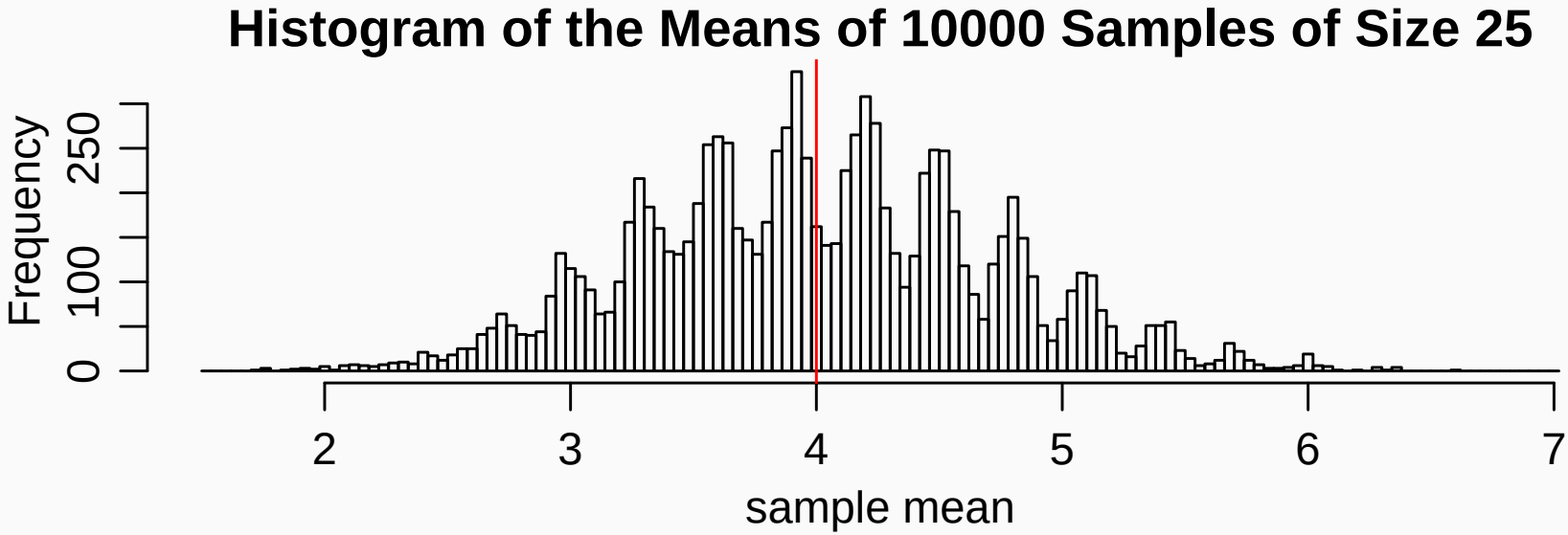
X	1	2	9
$P(X)$	1/3	1/3	1/3

and then to the following simulation.

1. We take a random sample of size $n = 25$ from the population, compute and record the sample mean, and the put the sample back.
2. We repeat the previous step 10000 times, and then obtain 10000 sample means.

What will the histogram of the 10000 sample means look like?

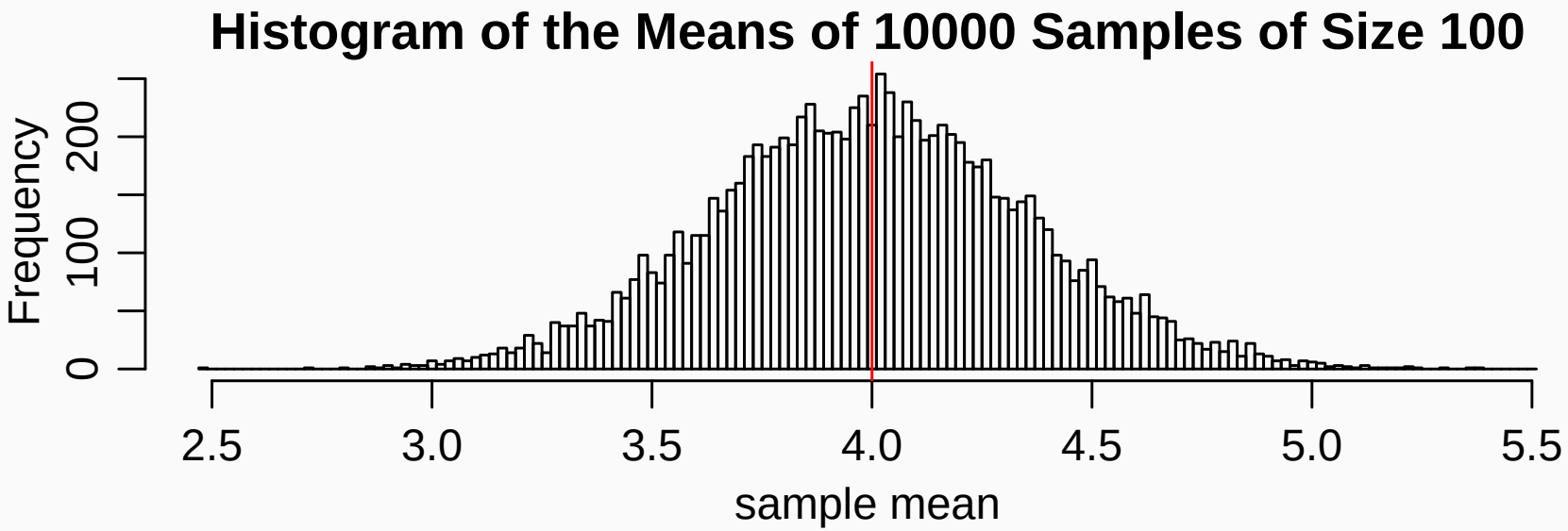
```
samplemean25 = vector("numeric", 10000)
for(i in 1:10000){
  samplemean25[i] = mean(sample(popratings, size = 25, replace=T,
                                prob=c(1/3,1/3,1/3)))
}
hist(samplemean25, breaks=seq(1.5,7.02,by=0.04),
     xlab="sample mean",
     main="Histogram of the Means of 10000 Samples of Size 25")
abline(v=4, col=2)
```



The red vertical line marks the position of the population mean $= 4$

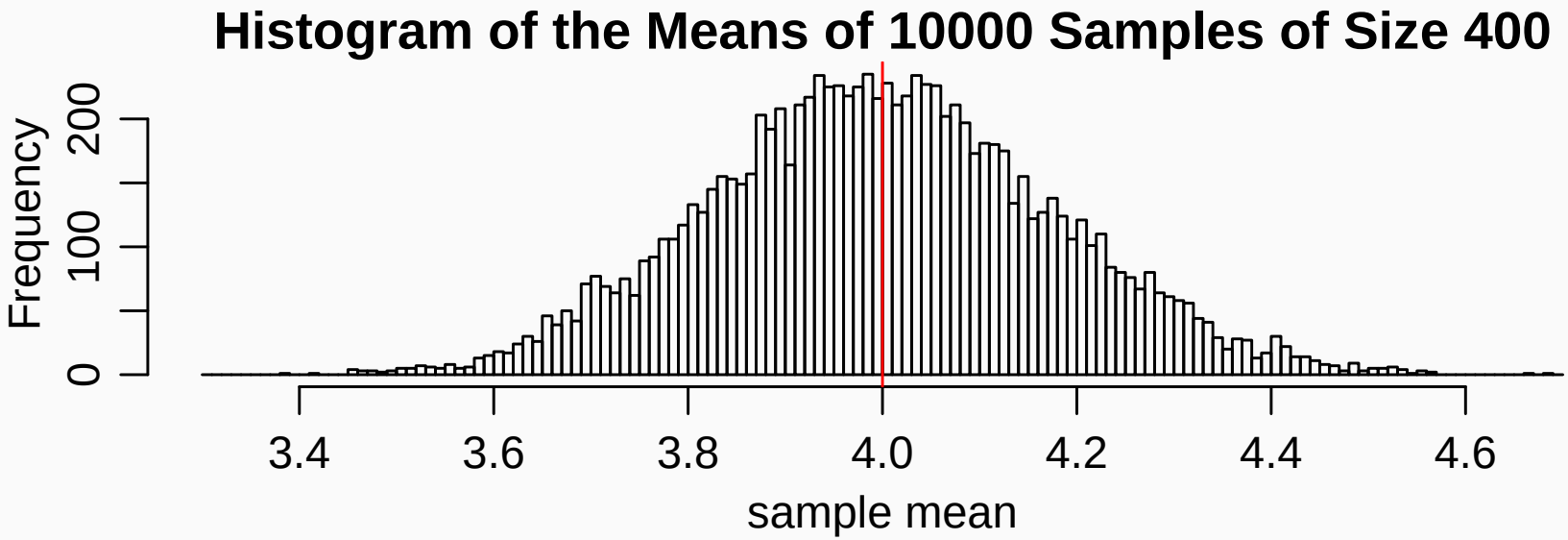
When we take a sample of size 25, the distribution of the sample means is not very normal, with a number of hills and valleys.

```
samplemean100 = vector("numeric", 10000)
for(i in 1:10000){
  samplemean100[i] = mean(sample(popratings, size = 100, replace=T,
                                 prob=c(1/3,1/3,1/3)))
}
hist(samplemean100, breaks=seq(2.51,5.51,by=0.02),
     xlab="sample mean",
     main="Histogram of the Means of 10000 Samples of Size 100")
abline(v=4, col=2)
```



The red vertical line marks the position of the population mean $= 4$

```
samplemean400 = vector("numeric", 10000)
for(i in 1:10000){
  samplemean400[i] = mean(sample(popratings, size = 400, replace=T,
                                 prob=c(1/3,1/3,1/3)))
}
hist(samplemean400, breaks=seq(3.3,4.7,by=0.01),
     xlab="sample mean",
     main="Histogram of the Means of 10000 Samples of Size 400")
abline(v=4, col=2)    # population mean
```



The red vertical line marks the position of the population mean $= 4$

When the sample size increases to 400, the distribution of the sample means looks very normal.

Sampling Distribution

- The probability distribution of a *statistic* is called the *sampling distribution* of the statistic.
- What we just constructed is the *sampling distribution* of the sample mean.

Observations for the Simulations Above

- The sampling distribution of the sample mean may not be normal when the sample size is small, but it gets more normal when the sample size gets larger.
- The sample mean may not be equal to the population mean, but its distribution centers at the population mean.
- With a larger sample, the variability sample mean around the population gets smaller.
    - What are the SDs of the sample means?

```
> mean(samplemean25)
[1] 3.99808
> mean(samplemean100)
[1] 4.001438
> mean(samplemean400)
[1] 3.99929
```

```
> sd(samplemean25)
[1] 0.7073244
> sd(samplemean100)
[1] 0.3577802
> sd(samplemean400)
[1] 0.1770972
```

Expected Value and SD of the Sample Mean

For **i.i.d.** random variables $X_1, X_2, \ldots, X_n$ from a population with mean μ and SD σ, the expected value and SD of the sample mean $\overline{X}_n = (X_1 + X_2 + \cdots + X_n)/n$ are respectively

$$E(\overline{X}_n) = \mu, \quad SD(\overline{X}_n) = \sigma/\sqrt{n}$$

- Here, "**i.i.d.**" = "**independent**, and **identically distributed**". which means $X_1, \ldots, X_n$ are independent and have identical probability distributions.
- Observations in a simple random sample is nearly i.i.d. if the sample size is less than 10% of the population size.
- SD of the sample mean is specifically call the *standard error*.

For the movie rating example, recall the population distribution is

X	1	2	9
$P(X)$	1/3	1/3	1/3

The mean, variance and SD of the population distribution are respectively

$$\mu = 1 \cdot \frac{1}{3} + 2 \cdot \frac{1}{3} + 9 \cdot \frac{1}{3} = 4$$

$$\sigma = \sqrt{(1-4)^2 \cdot \frac{1}{3} + (2-4)^2 \cdot \frac{1}{3} + (9-4)^2 \cdot \frac{1}{3}} = \sqrt{\frac{38}{3}} \approx 3.56.$$

sample size n	expected value of $\overline{X}_n$	SD of $\overline{X}_n$
25	4	$3.56/\sqrt{25} \approx 0.712$
100	4	$3.56/\sqrt{100} \approx 0.356$
400	4	$3.56/\sqrt{400} \approx 0.178$

```
> sd(samplemean25)
[1] 0.7073244
> sd(samplemean100)
[1] 0.3577802
> sd(samplemean400)
[1] 0.1770972
```

Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots$ be a sequence of **i.i.d.** random variables (discrete or continuous) with **mean μ and variance σ^2**. Then, when *n is large*,

- the distribution of the **sample mean**

$$\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

  is approximately

$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

- the distribution of the sum $S_n = X_1 + X_2 + \cdots + X_n$ is approximately

$$N(n\mu, \sqrt{n}\sigma).$$

Example

X_i's are i.i.d., with the distribution

X_i	1	2	9
$P(X_i)$	1/3	1/3	1/3

Recall that $\mu = 4, \quad \sigma \approx 3.56$. So the sampling distribution of $\overline{X}_{100}$ is approximately

$$N(\mu, \sigma/\sqrt{100}) = N(4, 0.356).$$

So

$$P(\overline{X}_{100} > 4.5) = P\left(Z > \frac{4.5 - 4}{0.356}\right) \approx P(Z > 1.40) \approx 0.08.$$

In the simulation 804 of the 10000 simulated $\overline{X}_{100}$ exceeds 4.5, which agrees with the CLT approximation that $\overline{X}_{100}$ exceeds 4.5 for about 8% of the time.

```
> sum(samplemean100 > 4.5)
[1] 804
```

Sample Size Required to Use CLT?

- Provided the sample size is large enough, the sampling distributions of the sample mean will be approximately normal, even when the population distribution is not normal.
- If the population distribution is normal, then so does the sampling distributions of the sample mean, regardless of the sample size.
- If population distribution is symmetric, then n should be at least 30 or so.
- If the population distribution is skewed or has outliers, then sample size n should be moderate (at least 100 or so), or even larger depending on how skewed or irregular the population distribution is.

Exercise 4.35 – Housing Prices (p.214)

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with an SD of $0.3 million. There were no houses listed below $0.3 million but a few houses above $3 million.

Can we find an approximate probability that a randomly chosen house in Topanga costs more than $1.4 million using the normal distribution?

No, because the population do not follow a normal distribution (it is right skewed), and a sample of size 1 is too small to use CLT.

Exercise 4.35 – Housing Prices (p.214)

Can we find an approximate probability that the mean of 60 randomly chosen houses in Topanga is more than \$1.4 million using the normal distribution? If yes, compute the approximate probability.

Yes, if the population distribution is not too skewed, the sampling distribution of the sample mean of a sample of size 60 might be normal by CLT.

$$\overline{X}_{60} \sim N\left(\mu = 1.3, SE = \frac{\sigma}{\sqrt{60}} = \frac{0.3}{\sqrt{60}}\right) = N(1.3, 0.0387).$$

So,

$$P(\overline{X}_{60} > 1.4) = P\left(Z > \frac{1.4 - 1.3}{0.0387}\right) \approx P(Z > 2.58) \approx 0.0049.$$

What Does the CLT Say?

True or False and explain: The central limit theorem says that as you take larger and larger samples from a population, the histogram of the sample values looks more and more normal.

False, as you take larger and larger samples, the histogram of the sample values looks more and more like the histogram of the population.

What is the thing that becomes more and more normal as the sample size gets larger and larger?

It is the distribution of the sample mean that get's more and more normal.