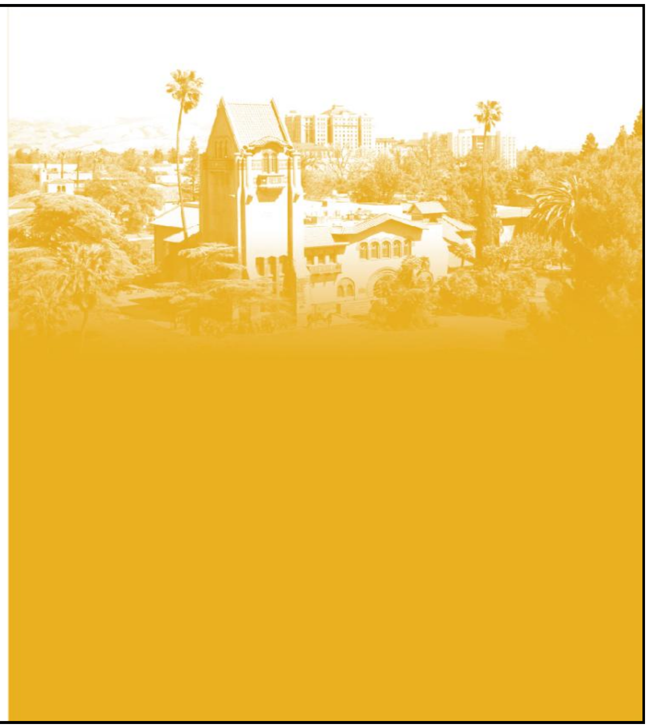
Parameter Estimation &
Central Limit Theorem



1

Agenda

- Sampling Distributions & Parameter Estimations
- Central Limit Theorem

2

2

SJSU SAN JOSÉ STATE UNIVERSITY

Sampling Distributions

3

SJSU SAN JOSÉ STATE UNIVERSITY

Example of a Population Distribution

- Suppose a certain movie has a bipolar distribution of ratings, that in a 1 to 10 scale, of those having watched the movie, 13 gave 9 points, 13 gave 2 points, and the remaining 13 gave 1 points.
- So, the population distribution is:

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

Population Distribution

4

4

SJSU SAN JOSÉ STATE UNIVERSITY

Histogram of Samples

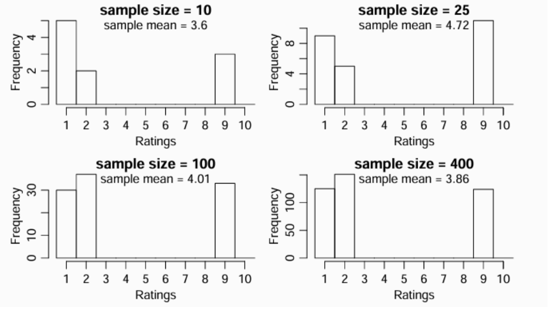
- In practice, since the population are difficult (or impossible) to examine completely, we take a sample to learn about the population. Will the makeup of the sample mimic the makeup of the population?
- First, the sampling method must be appropriate. A biased sample won't give us the correct information 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?

5

5

SJSU SAN JOSÉ STATE UNIVERSITY

Histogram of Samples



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

6

6

SJSU SAN JOSÉ STATE UNIVERSITY

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?**

SJSU SAN JOSÉ STATE UNIVERSITY

Variability of the Sample Means

To determine 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

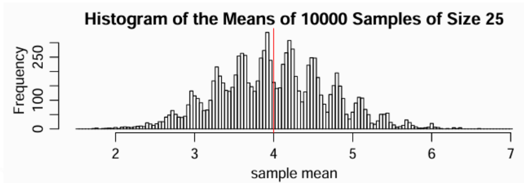
Then, we perform the following simulations:

- We take a random sample of size n = 25 from the population, compute and record the sample mean, and the put the sample back.
- We repeat the previous step 10,000 times, and then obtain 10,000 sample means.
- What will the histogram of the 10,000 (n = 25) samples means look like?

SJSU SAN JOSÉ STATE UNIVERSITY

Variability of the Sample Means

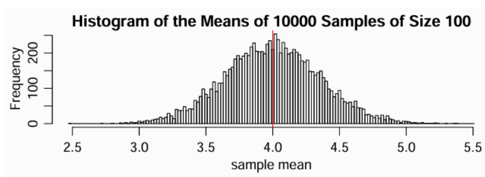
- For a sample of size of n = 25, the distribution of the sample means is <u>not very normal</u>, with a number of hills and valleys.



SJSU SAN JOSÉ STATE UNIVERSITY

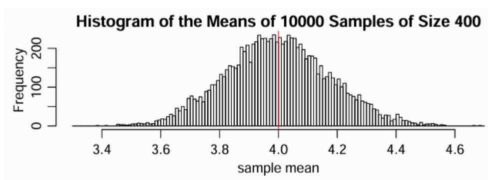
Variability of the Sample Means

- But when increase to n = 100, then…



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

- And n = 400 :



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

SJSU SAN JOSÉ STATE UNIVERSITY

Sampling Distributions

- 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.

A few observations:

- 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.

11

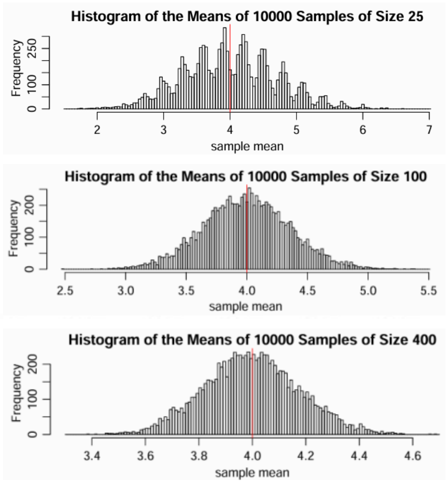
11

SJSU SAN JOSÉ STATE UNIVERSITY

Sampling Distributions

What are the SDs of the sample means?

Sample Size	Mean	Std Dev
25	3.998	0.707
100	4.001	0.358
400	3.999	0.177



12

12

SJSU SAN JOSÉ STATE UNIVERSITY

Expected Value and Standard Deviation of the Sample Mean

- Given random variables $X_1, X_2, \cdots, X_n$ from a population with mean μ and SD σ^2 that are independent and identical probability distributions (aka i.i.d. – independent & identically distributed), the sample mean is simply:

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

- The expected value and standard deviation of the sample mean are:

$$E(\bar{X}_n) = \mu \qquad SD(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

- Observations in a simple random sample is nearly i.i.d. if the sample size is less than 10% of the population size.
- Standard deviation of the sample mean is also called the standard error.

SJSU SAN JOSÉ STATE UNIVERSITY

Example: Movie Rating Revisited

- 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:

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

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

- The sample means and SD's are:

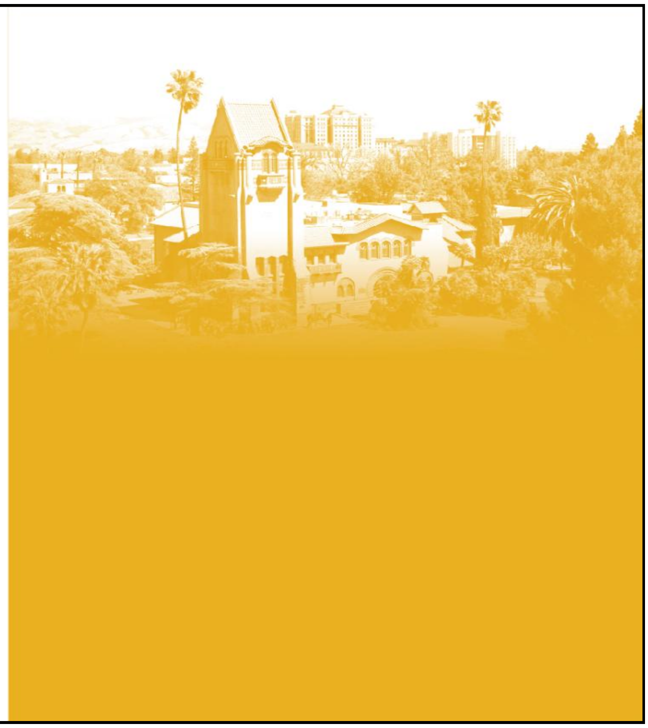
$$E(\bar{X}_n) = \mu$$

$$SD(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

Sample Size	$E(\bar{X}_n)$	$SD(\bar{X}_n)$
25	4	$3.56/\sqrt{25} = 0.712$
100	4	$3.56/\sqrt{100} = 0.356$
400	4	$3.56/\sqrt{400} = 0.178$

SJSU SAN JOSÉ STATE UNIVERSITY

Central Limit Theorem



15

SJSU SAN JOSÉ STATE UNIVERSITY

Central Limit Theorem (CLT)

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

- the distribution of the sample mean

$$\bar{X}_n = \frac{(X_1 + X_2 + \cdots + X_n)}{n} \quad \text{is approximately} \quad N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

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

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

16

16

SJSU SAN JOSÉ STATE UNIVERSITY

Example: Movie Rating Revisited

- For the movie rating example, recall that:

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

$\mu = 4, \qquad \sigma = 3.56$

- The sampling distribution of $\bar{X}_{100}$ is approximately:

$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) = N(4, 0.356)$$

- The probability of $\bar{X}_{100} > 4.5$ is:

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

- In the simulation 804 of the 10,000 simulated $\bar{X}_{100}$ exceeds 4.5, which agrees with the CLT approximation that $\bar{X}_{100}$ exceeds 4.5 for about 8% of the time.

SJSU SAN JOSÉ STATE UNIVERSITY

What's the Sample Size Required for CLT?

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

Example: Central Limit Theorem

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.

1) Can we find an approximate probability that a randomly chosen house in Topanga costs more than $1.4 million using the normal distribution?
2) 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.

What Does the CLT Say?

<u>True or False</u>: 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. **Explain.**

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

SJSU SAN JOSÉ STATE UNIVERSITY

Python CLT Demo!!!