Lecture 15
The Central Limit Theorem
Sampling Distributions of $\bar{x}$ and $\hat{p}$

Review From Monday

Three Types of Distributions:

- **Population Distribution** –
- the probability distribution of a single observation of a random variable – shows the possible outcomes of the single observation and their probabilities.
  - Its properties are described by unknown parameters such as p or μ, σ^2, σ

- **Data Distribution** –
- This is the distribution(s) of variable(s) in our sample based on the observations we sampled.
  - Its properties are described by statistics such as $\bar{x}$ or $\hat{p}, s^2, s$
  - The data distribution of a variable will converge to the population distribution as $n \to N$

- **Sampling Distribution** –
- This is the distribution(s) of statistic(s) computed from the observations in the sample. This distribution arises from repeatedly sampling from the same population and computing statistics from those samples. It tells us how close a given estimate is to the true population parameter it is estimating (sampling error).
  - Its properties are described by the properties of the population distribution and sample size n

Central Limit Theorem

- The central limit theorem gives us some nice guarantees about the shape of the distribution of a statistic

Definition: if $X_1, X_2, \ldots X_n$ are independent and identically distributed random variables (all have the same distribution) such that

$E[X_i] = \mu$ and $E[X_i - \mu]^2 = \sigma^2 < \infty$ (have finite variance)

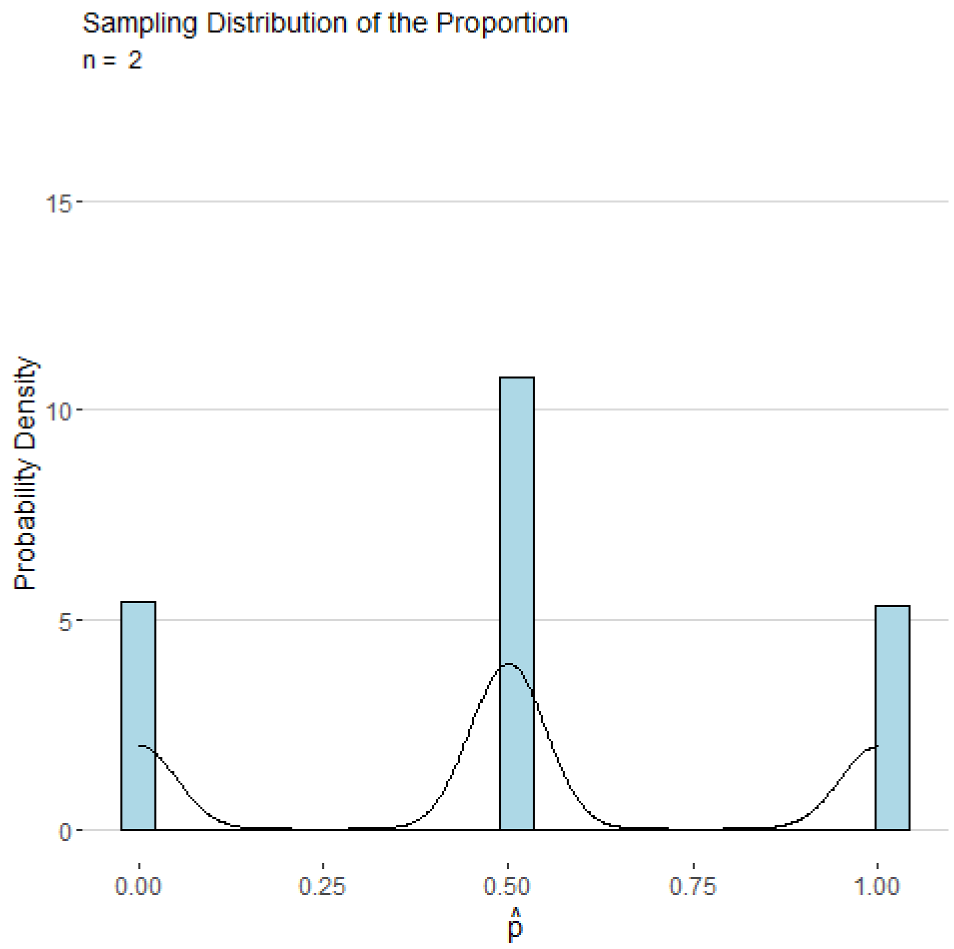
Then,

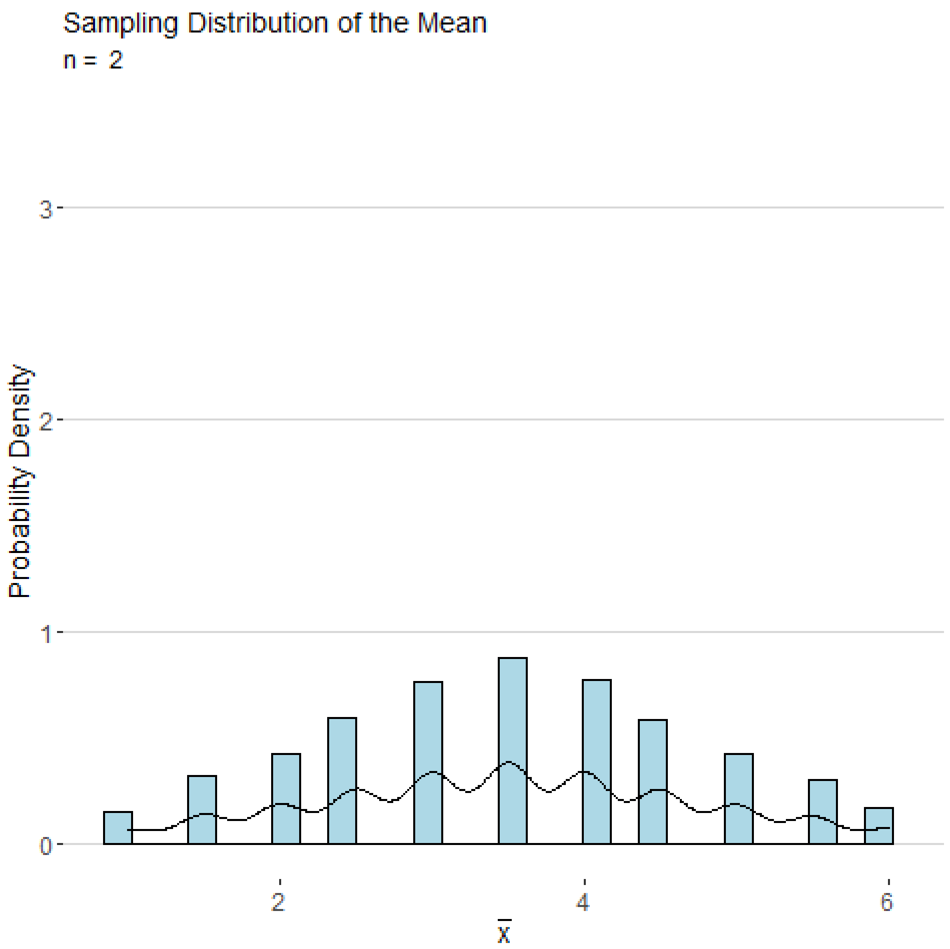
$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{d} N(0,1)$$

Where $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ and where $\xrightarrow{d}$ denotes convergence in distribution

(in layman's terms) the **central limit theorem** states that as the sample size increases the *shape* of a sampling distribution of $\bar{x}$ will "approach" that of a normal distribution

Central Limit Theorem





Sampling Distributions of $\bar{x}$ and $\hat{p}$

The sampling distribution of the mean

- The mean of $\bar{x}$ is: μ - the population mean
- The standard deviation of $\bar{x}$ is: ${}^{\sigma}/_{\sqrt{n}}$

The sampling distribution of the sample proportion

- The mean of $\hat{p}$ is: p – the population proportion
- The standard deviation of $\hat{p}$ is: $\sqrt{\frac{p(1-p)}{n}}$

California Gubernatorial Election

- Election polling is one of the few cases where we know p - the true proportion of voters (either voting for one candidate or another) - because all the votes are counted.

- From our example in week two about the California race for governor, the true population proportion of voters who cast a vote for Democrat Jerry Brown was 54.8% while the sample proportion measured from 3,889 voter interviews was 53.1%.
- What are the mean and standard deviation of $\hat{p}$?

- Why is the standard deviation so small?

Applying The CLT

- Recall that the **Empirical Rule** tells us how observations are distributed for approximately symmetric bell-shaped (normal) distributions.

- Since 95% of observations in a normal distribution fall within 2 standard deviations of the mean.

- Adapting the rule for probability distributions means that there is a 95% probability that random variable will fall within ∓ 2 standard deviations of the mean of the distribution.

Applying The CLT

The probability that $\bar{x}$ will be between $\mu - {}^{2\sigma}/_{\sqrt{n}}$ and $\mu + {}^{2\sigma}/_{\sqrt{n}}$ is approximately 0.95

- The probability that $\hat{p}$ will be between $p - 2\sqrt{{}^{p(1-p)}/_{n}}$ and

$p + 2\sqrt{{}^{p(1-p)}/_{n}}$ is approximately 0.95

Estimation

Estimation is a type of statistical inference where we use our statistic to estimate a parameter

- We can use $\bar{x}$ (the means of a sample of n observations) to estimate the mean of single observation (i.e μ)

- We can use s (the standard deviation of the observations a sample of n observations) to estimate the standard deviation of a single observation (i.e σ)

- We can use $\hat{p}$ (the proportion of observations that are a "success" in a sample of n observations) to estimate the probability of success (i.e p)

Some Technical Points

Note that parameters μ, p, σ have a couple of interpretations.

- We use the sampling distribution of our statistics to determine how effective they are at estimating the parameters of interest.
  - Both $\bar{x}$ and p are **unbiased** estimators
  - A **standard error** is the standard deviation of a statistic
  - The central limit theorem implies that (unless n is very small) the shape of the sampling distributions of $\bar{x}$ and p are approximately normal

The above properties rely on some technical assumptions about how the data are collected which we will talk about in a few lectures

Practice: Crooked Casino

- A crooked casino uses loaded dice at all of their Craps tables to improve their earnings. The table to left gives the probability distribution for the sum of roll of two die for a pair of fair dice (denoted X_{fair}) and for a pair of loaded dice (denoted X_{loaded})

X	$P(X_{\text{loaded}})$	$P(X_{\text{fair}})$
2	0.03	0.028
3	0.06	0.056
4	0.08	0.083
5	0.10	0.111
6	0.14	0.139
7	0.28	0.167
8	0.12	0.139
9	0.10	0.111
10	0.05	0.083
11	0.02	0.056
12	0.03	0.028

Practice: Crooked Casino

- Suppose a gambler at the casino is suspects that the casino is using loaded dice so he observes the proportion of "sums of 7" rolled in the next 30 turns at the Craps table. He computes the proportion of rolls that summed to 7 to be 0.33

- Assuming the dice are fair, Compute the interval that has a probability of approximately 0.95 of containing estimated proportion of rolls that sum to 7

$$\hat{p} \approx N\left(0.167, \sqrt{\frac{0.167(1-0.167)}{30}}\right) \approx N(0.167, 0.068)$$

$$P(0.167 - 2 \times 0.068 < \hat{p} < 0.167 + 2 \times 0.068) = 0.95$$

$$P(0.031 < \hat{p} < 0.303) = 0.95$$

X	$P(X_{\text{loaded}})$	$P(X_{\text{fair}})$
2	0.03	0.028
3	0.06	0.056
4	0.08	0.083
5	0.10	0.111
6	0.14	0.139
7	0.28	0.167
8	0.12	0.139
9	0.10	0.111
10	0.05	0.083
11	0.02	0.056
12	0.03	0.028

Practice: Crooked Casino

- Suppose a gambler at the casino is suspects that the casino is using loaded dice so he observes the proportion of "sums of 7" rolled in the next 30 turns at the Craps table. He computes the proportion of rolls that summed to 7 to be 0.33

- Assuming the dice are fair, what is the probability of observing a proportion greater than the gamblers estimate?

$$SE(\hat{p}) = \sqrt{\frac{0.167(1-0.167)}{30}} = 0.068$$

$$z = \frac{0.33-0.167}{0.068} = 1.89$$

$$P(z > 2.39) = 1 - P(z \leq 2.39) = 0.0084$$

X	$P(X_{\text{loaded}})$	$P(X_{\text{fair}})$
2	0.03	0.028
3	0.06	0.056
4	0.08	0.083
5	0.10	0.111
6	0.14	0.139
7	0.28	0.167
8	0.12	0.139
9	0.10	0.111
10	0.05	0.083
11	0.02	0.056
12	0.03	0.028

Example

- Recall from the tiger trout example on Wednesday that the probability of catching a tiger trout in a single cast was 5%. Suppose a fisherman makes 450 casts in an afternoon and marks any time he catches a tiger trout as a success. Compute the interval for which the probability of $\hat{p}$ is approximately 0.95

- $n = 450$
- $p = 0.05$
- $P\left(p - 2\sqrt{p(1-p)/n} < \quad \hat{p} \quad < p + 2\sqrt{p(1-p)/n}\right) = 0.95$
- $SD = 0.01$

$$[0.05 \ - 2 \times 0.01, 0.05 + 2 \times 0.01]$$

$$P(0.03 < \hat{p} < 0.07) = 0.95$$