Lecture 16
Central Limit Theorem

Review From Monday:

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 2 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 $\hat{p}$?

$$\text{mean of } \hat{p} = 0.548$$

$$\text{SD of } \hat{p} = \sqrt{\frac{0.531 \times (1 - 0.531)}{3889}} = \sqrt{6.4e^{-5}} = 0.008$$

- Why is the standard deviation so small?

Central Limit Theorem

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

Definition: if $\mathrm{X}_1, \mathrm{X}_2, \ldots \mathrm{X}_\mathrm{n}$ 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)

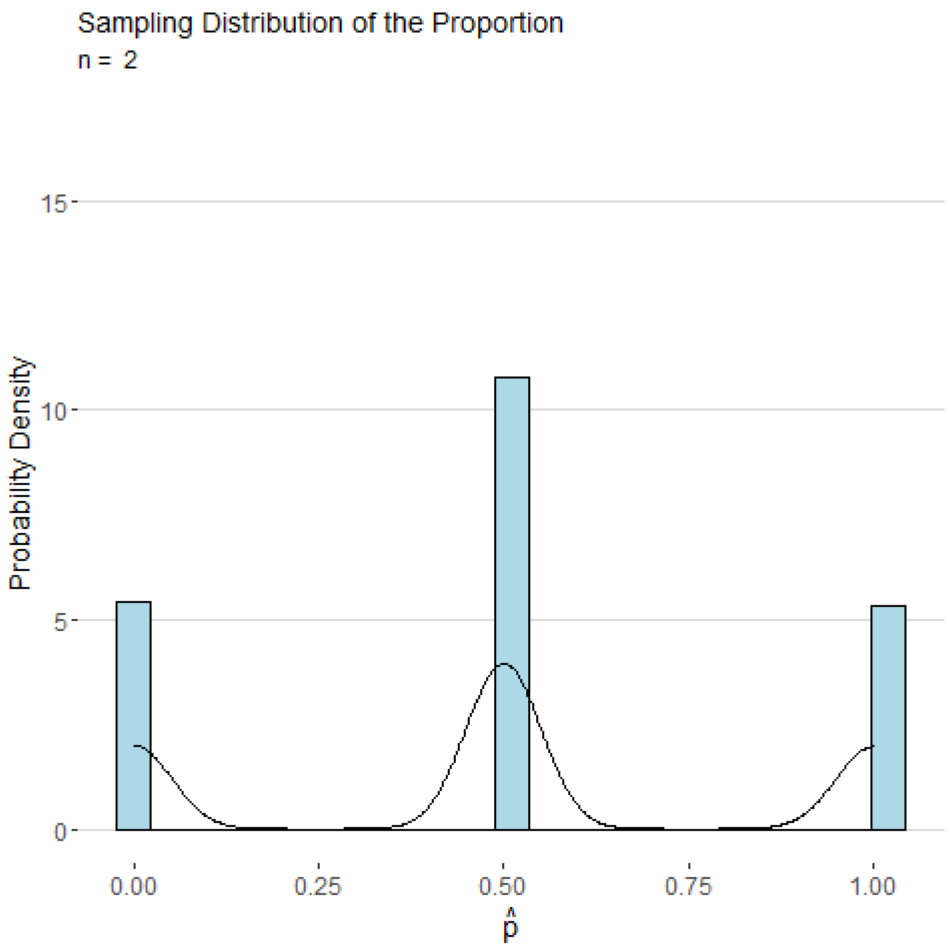
Then,

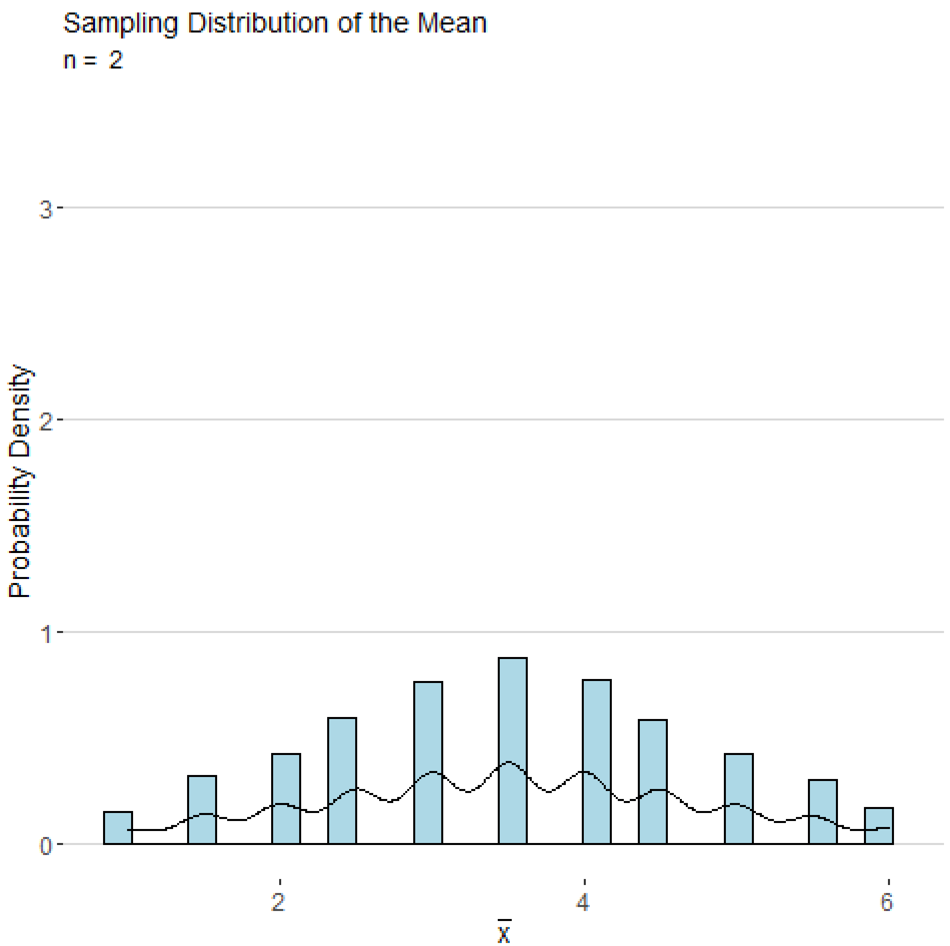
$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \overset{d}{\rightarrow} N(0,1)$$

- $E[X]$ is the expected value (mean) of a random variable
- $E[X \ - \mu]^2$ is the variance of a random variable
- $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$
- $\overset{d}{\rightarrow}$ 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





Applying The CLT

- The central limit theorem tells us that for moderate to large n

$$\bar{X} \sim N\left(\mu, {}^{\sigma}/_{\sqrt{n}}\right)$$

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

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

- 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

- A **standard error (SE)** is the standard deviation of a statistic

- In statistical Inference, we can place bounds on estimation by computing the interval of the sampling distribution that has a probability of 0.95 of containing our estimate

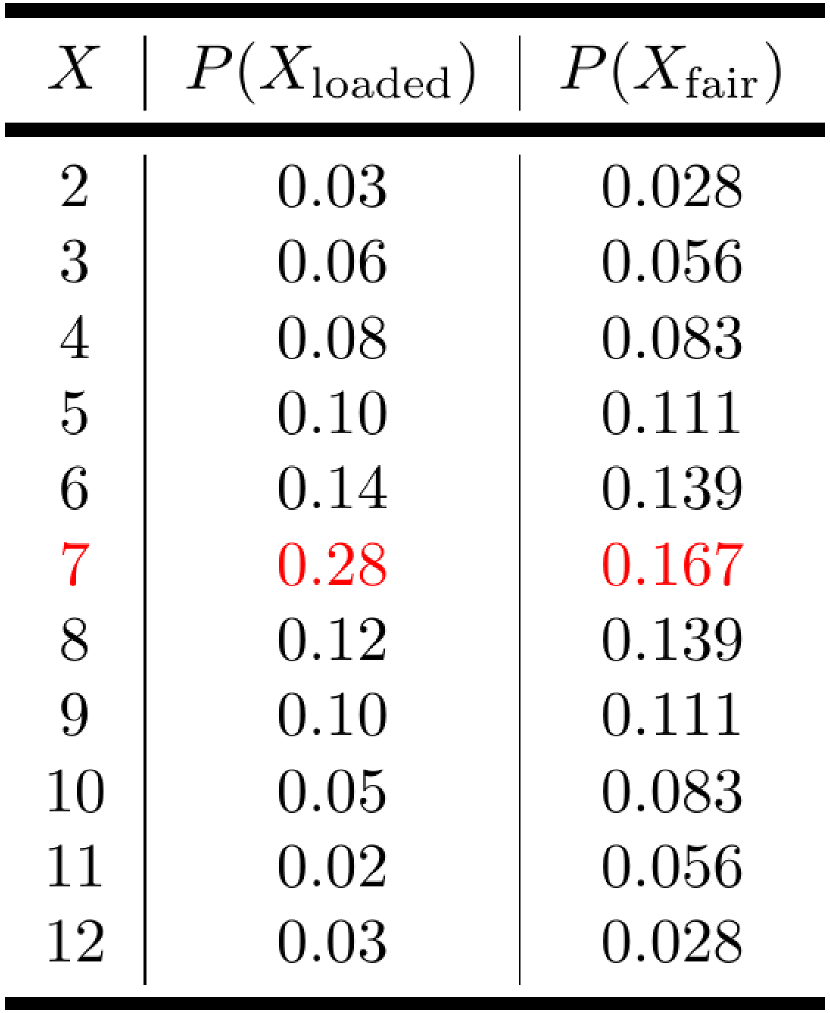
- That interval is defined as $\pm 2 \times \mathrm{SE}$ from the mean:

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

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

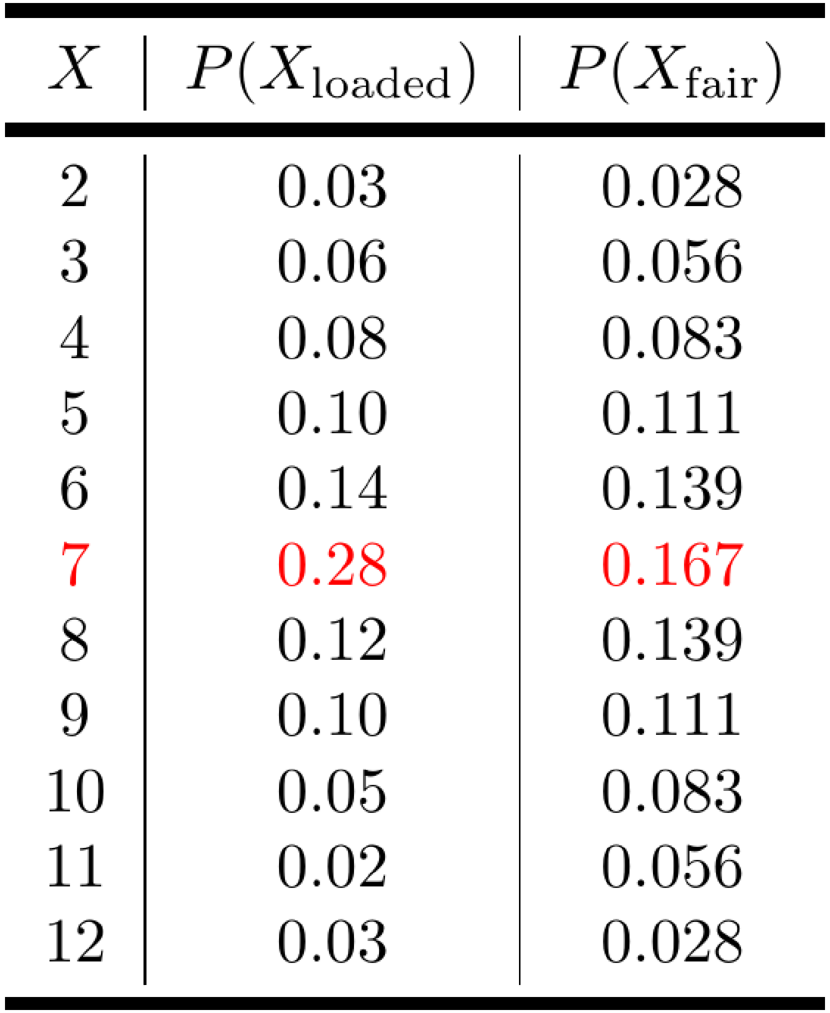
- Suppose a gambler at the casino is suspects that the casino is using loaded dice so he observes the proportion of “sums of 7” rolled in the next 30 turns at the Craps table. He computes the proportion of rolls that summed to 7 to be 0.33

- 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	$P(X_{\mathrm{loaded}})$	$P(X_{\mathrm{fair}})$
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

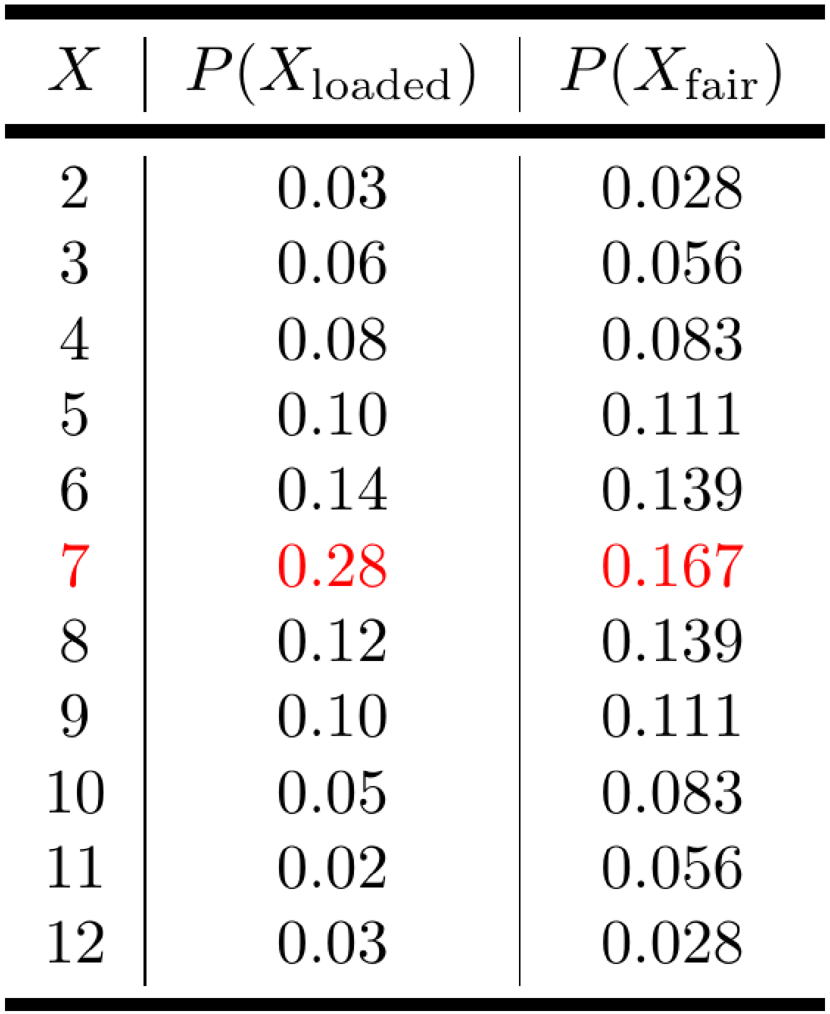
- Suppose a gambler at the casino is suspects that the casino is using loaded dice so he observes the proportion of “sums of 7” rolled in the next 30 turns at the Craps table. He computes the proportion of rolls that summed to 7 to be 0.33

- 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