Lecture 8

Last time

We looked at specific SAMPLING DISTRIBUTIONS for count variables and proportions.

What is the distribution for count variables?

What are the properties of this distribution?

When would we use this type of distribution? An example situation:

Also, looked at the mean and standard deviation:

$$\mu_X = np \tag{1}$$
$$\sigma_X = \sqrt{np(1-p)} \tag{2}$$

$$\mu_p = p \tag{3}$$
$$\sigma_p = \sqrt{\frac{p(1-p)}{n}} \tag{4}$$

The Sampling Distribution for a Sample Mean

Before we select a sample, the sample mean $\bar{X}$ is a random variable - we don't know what value it will take. The value varies from sample to sample. By contrast, the population mean μ is a single fixed number. At a given time for a particular scenario, there can be many different sample but there's a single population. For random sampling, the sample mean $\bar{X}$ can fall either above or below the population mean μ. *In fact, the mean of the sampling distribution of the sample mean $\bar{X}$ equals the population mean μ.*

For random sampling, the standard deviation of $\bar{X}$ depends on the sample size n and the spread of the probability distribution from which the observations are sampled - *the population distribution.*

$$\mu_{\bar{X}} = \mu \tag{5}$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \tag{6}$$

Let's consider the formula for the standard deviation of $\bar{X}$: $\frac{\sigma}{\sqrt{n}}$. Notice that as the sample size n increases, the denominator increases, so the standard deviation of $\bar{X}$ decreases. Therefore, with larger samples, the sample mean is more likely to fall close to the population mean.

Example:

Many US states, Canadian provinces and countries have lottery games. Most lotteries use "number games". They are usually several-digit generalizations of the following game: You bet a dollar on a number between 0 and 9 that you pick at random. If you are correct, you win \$5, and otherwise, you win nothing.

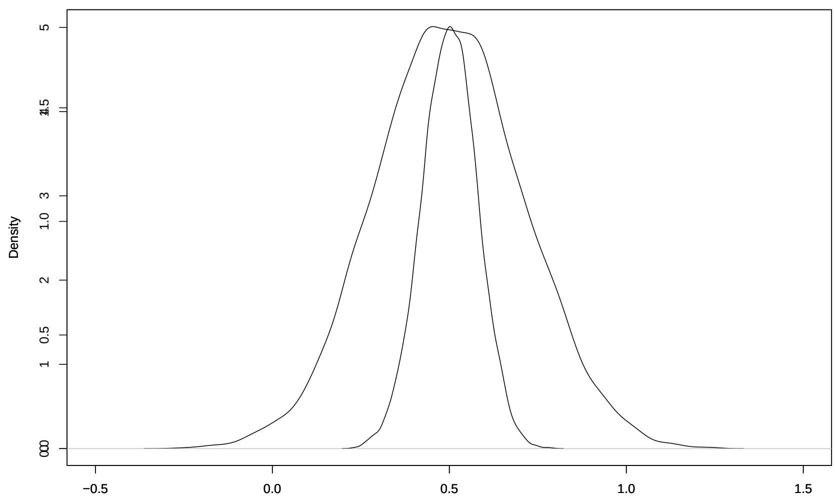
Let X denote your winnings for a single play of this lottery. The possible values for X are \$0 with probability $\frac{9}{10} = .90$ and \$5 with the probability $\frac{1}{10} = .10$. This probability distribution has a mean μ=0.50, 50 cents. This is the expected value of the outcome X for a single play. Naturally, the expected return is less than you pay to play because the government wants to make money. It can be shown that the standard deviation of this probability distribution is σ=1.50.

Find the mean and standard deviation of the sampling distribution of your mean winnings if you play this game:

(i) once a week for the next year (52 times)

(ii) once a day for the next year (365 times)

Which distribution is for (i) and (ii)?



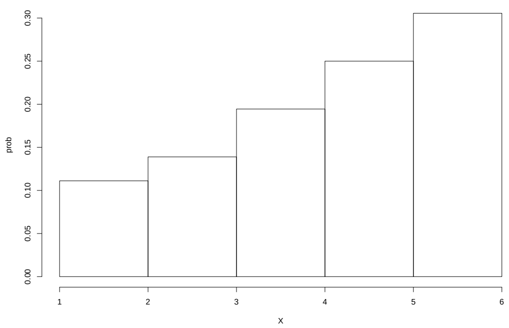
Central Limit Theorem

Sample means have NORMAL sampling distributions for large n

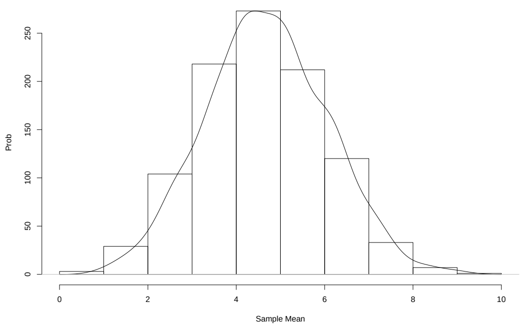
Example: Many times before beginning a game we roll a die to determine who gets to go first. Suppose there are two players playing a game of *Monopoly*. Each player rolls the die once and the player with the highest number goes first. Here is a probability distribution of X= highest number out of two rolls of a die:

X value	Probability
1	1/36
2	3/36
3	5/36
4	7/36
5	9/36
6	11/36

Here is the same information displayed on a bar graph:



This distribution is left-skewed with a mean of μ=4.5 and a standard deviation of σ=1.4. Suppose we repeatedly sample from this distribution with random sample sizes of n=30. Below is the histogram of the sample means for a huge number of simulated samples from this skewed distribution.

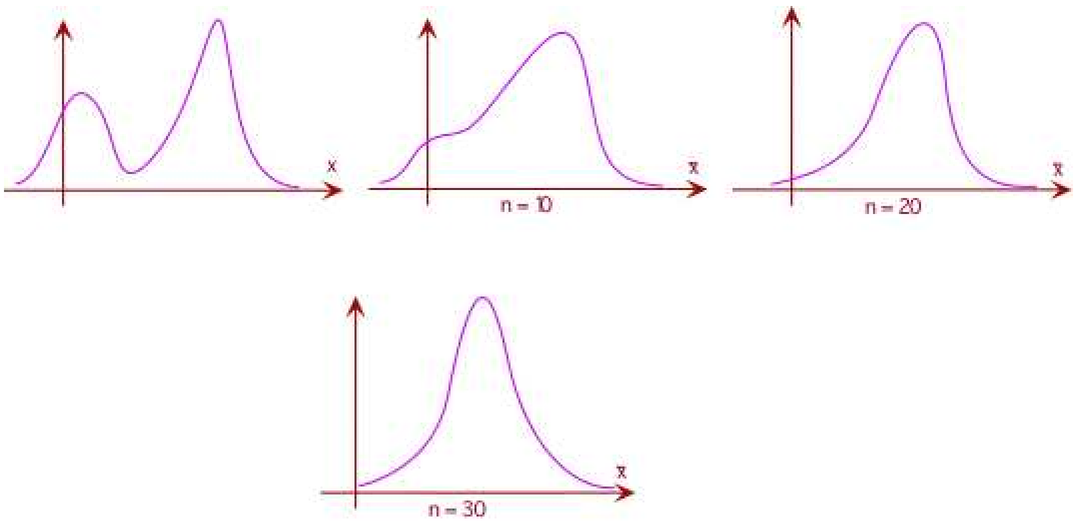


Even though the probability distribution from which we took the samples is highly skewed, this sampling distribution has a bell shape. The histogram above has a normal distribution superimposed that has the same mean (μ=4.5) as the sampling distribution and a standard deviation of $\frac{\sigma}{\sqrt{n}} = \frac{1.4}{30} = 0.26$. This phenomenon is described by the **central limit theorem**.

Central Limit Theorem states that for a random sample with a large sample size n, the sampling distribution of the sample mean $\bar{X}$ is approximately a normal distribution.

This result applies no matter what the shape of the underlying population distribution from which the samples are taken.

How large is "large"?
The more heavily skewed the underlying population distribution is, the greater n must be for the approximation to be good. Usually, at least n=30 is required but sometimes even more than that.



Example:
A local Lube Express, a franchised oil changing company, wants to know how long its customers must wait to have their oil changed. The mean wait time μ=32 minutes and σ=12 minutes is the standard for Lube Expresses nationwide with an exponential distribution (i.e. something nonnormal). During one week, the manager takes an SRS of 50 customers.

First, what does the Central Limit Theorem say about the approximate distribution of $\bar{X}$?

What is the probability that the average wait time is greater than 35 minutes?