Goals for the Week
Linear Combinations
Sampling Distributions
Central Limit Theorem

STAT 140: Week 6

My hope is that after Week 6 (10/04-10/08), you will be able to:

After this week, you should be able to

- Calculate expected values and variances for combinations of random variables.
- Understand and use the Central Limit Theorem to identify sampling distributions and calculate probabilities.

Random Variables 1

- **Random Variable**
  - A **random variable** is a variable or process with a numerical outcome.
  - We usually represent random variables with a capital letter, such as X, Y, or Z.
- Each random variable we think about has:
  - An associated probability density function, which gives us theoretical information about the likelihood of a particular outcome (or outcomes).
  - An expected value, denoted $E(X)$.
  - A variance or standard deviation, denoted $Var(X)$ or $SD(X)$ (rememeber, the standard deviation is the square root of the variance.)

Linear Combinations

- A **linear combination** is a mathematical phrase to describe a combination $aX + bY + c$.
  - The expected value of $aX + bY + c$ is

$$aE(X) + bE(Y) + c$$

- The variance of $aX + bY + c$ is

$$a^2 \times Var(X) + b^2 \times Var(Y).$$

Practicing Linear Combinations 1

- Let's say the time it takes my washing machine to run a cycle is a random variable represented by W, and the time it takes my dryer to run a cycle is a random variable represented by D.
- Let's say it always takes me three minutes to move my laundry from the washing machine to the dryer.
  - Represent the total time it takes me to do a load of laundry as a linear combination.
  - If $E(W) = 45$ and $E(D) = 65$, how long should I expect it to take to do a load of laundry?
  - If $Var(W) = 5$ and $Var(D) = 3$, what is the variance for the length of time it takes me to do a load of laundry?

Practicing Linear Combinations 2

- It takes me
- The time it takes me to do a load of laundry can be expressed as

$$W + D + 3$$

- Then, I should expect it to take $E(W + D + 3) = E(W) + E(D) + 3 = 45 + 65 + 3 = 113$ minutes to do a load of laundry.
- The variance for the length of time it takes me to do a load of laundry is $Var(W + D + 3) = 1^2 \times Var(W) + 1^2 \times Var(D) = 1 \times 5 + 1 \times 3$ is 8 minutes (standard deviation of $\sqrt{8}$).

Practicing Linear Combinations 3

In lab last week, we assumed that the yield of cake batter from a box of cake mix followed the normal distribution with a mean of 4.75 cups and a standard deviation of 0.25 cups. Let's call the amount of cake batter a random variable C.

- What is the expected value of the amount of cake batter from four boxes of mix? The variance?
- A two-layer cake requires at least 20 cups of cake batter. Knowing that the sum of normal random variables is also normal, what is the probability that I have enough cake batter from four boxes of mix for a two-layer cake?

Practicing Linear Combinations 4

- The expected value of four boxes of mix is $E(4C) = 4E(C) = 4 \times 4.75 = 19$.
- The variance of four boxes of mix is $Var(4C) = 4^2 \times Var(C) = 16 \times (0.25^2) = 1$.

```
1 - pnorm(20, mean = 19, sd = 1)
```

```
## [1] 0.1586553
```

Random Variables 2

- Remember, random variables are mostly theoretical! In these problems, we have been given parameters, like μ, σ, or p, and then we calculate probabilities based on what *should* be happening.
- In reality, what we actually observe gets compared to what we think should happen–i.e., we are using *statistics* to estimate *parameters* and then drawing conclusions about the population.
- We can think of statistics, like $\bar{x}$ and $\hat{p}$, as linear combinations of random variables.
- Then, a statistic has a distribution–this distribution is called the **sampling distribution**.

Sample Average, $\bar{x}$

Remember,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}x_1 + \frac{1}{n}x_2 + ... + \frac{1}{n}x_n$$

- Then, if $x_1, x_2, ...x_n$ have mean μ,

$$\begin{aligned} E(\bar{x}) &= E\left(\frac{1}{n}x_1\right) + E\left(\frac{1}{n}x_2\right) + ... + E\left(\frac{1}{n}x_n\right) \\ &= \frac{1}{n}E(x_1) + \frac{1}{n}E(x_2) + ... + \frac{1}{n}E(x_n) \\ &= \frac{1}{n}\mu + \frac{1}{n}\mu + ... + \frac{1}{n}\mu = \mu \end{aligned}$$

Standard Deviation of $\bar{x}$

- Then, if $x_1, x_2, ...x_n$ have standard deviation σ,

$$\begin{aligned} Var(\bar{x}) &= Var\left(\frac{1}{n}x_1\right) + Var\left(\frac{1}{n}x_2\right) + ... + Var\left(\frac{1}{n}x_n\right) \\ &= \frac{1}{n^2}Var(x_1) + \frac{1}{n^2}Var(x_2) + ... + \frac{1}{n^2}Var(x_n) \\ &= \frac{1}{n^2}\sigma^2 + \frac{1}{n^2}\sigma^2 + ... + \frac{1}{n^2}\sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

The standard deviation would be $\frac{\sigma}{\sqrt{n}}$.

Example

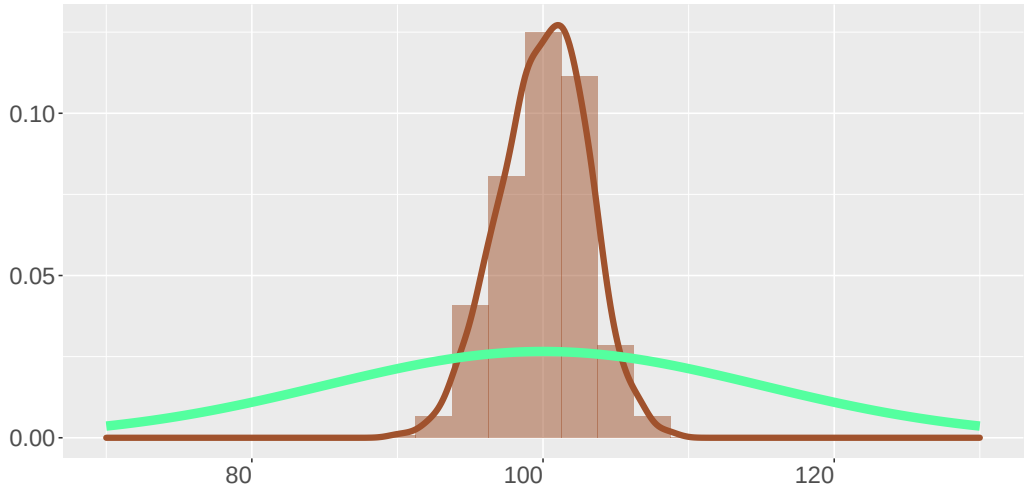
We can randomly generate observations from different distributions in R.

```
sample1 <- rnorm(25, mean = 100, sd = 15)
mean(sample1)
```

```
## [1] 106.9211
```

- We would expect the mean to be 100, and it is 106.9211–fairly close!
- If we repeated this process an infinite number of times, we would build the sampling distribution of the mean.

Sampling Distributions 1



Practicing Sampling Distributions 1

Describe both of these graphs, making sure to compare their centers, shapes, and spreads.

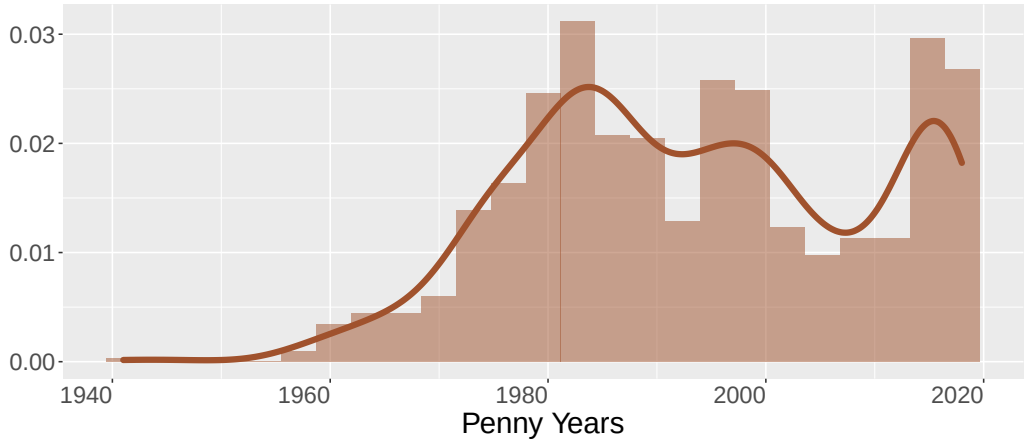
Practicing Sampling Distributions 2

- The **centers** of the two graphs are approximately equal at 100.
- The **spread** of the first graph is much larger than the spread of the second graph (ranging from about 75 to 125, as compared to 93 to 107).
  - When we are talking about a sampling distribution (bottom graph), we typically use the term **standard error** rather than standard deviation.
- The **shapes** of both graphs are approximately normal (unimodal, symmetric, bell-shaped).

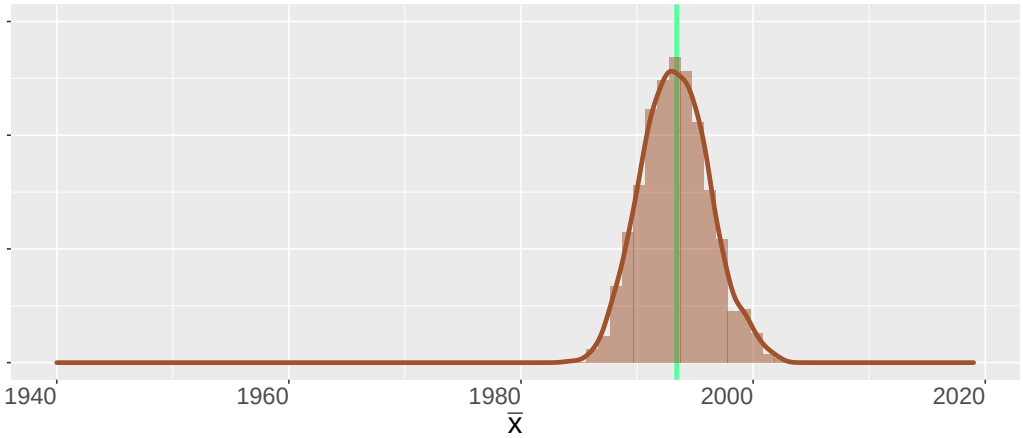
Sampling Distributions 2

- We know that the sum of normal random variables also follows the normal distribution, but what about those variables that are not normal? What happens to their mean or sum?
- Let's carry out an example with a "population" of pennies.

Sampling Distributions 3



Sampling Distributions 4



Practicing Sampling Distributions 3

Describe both of these graphs, making sure to compare their centers, shapes, and spreads.

Practicing Sampling Distributions 4

- The **centers** of the two graphs are approximately equal, around 1993-1994.
- The **spread** of the first graph is much larger than the spread of the second graph (ranging from 1941 to 2018, as compared to 1985 to 2003).
- The **shapes** of the graphs are quite different–even though the first is left-skewed and has many peaks, the second graph is approximately normal (unimodal, symmetric, bell-shaped).

Central Limit Theorem 1

When we collect a sufficiently large sample of n independent observations from a population with mean μ and standard deviation σ, the sampling distribution of the sample mean, $\bar{x}$, is nearly normal with

$$\text{mean, or } E(X) = \mu, \text{ and}$$

$$\text{standard error, or } SE = \frac{\sigma}{\sqrt{n}}$$

Central Limit Theorem 2

- The distribution is normal!!
- The center of a sampling distribution is always at the population mean μ.
- For a particular population, the variability in the sampling distribution decreases as the sample size n becomes larger. This will likely align with your intuition: an estimate based on a larger sample size will tend to be more accurate.

Central Limit Theorem: Conditions

- Independence: The sample observations must be independent!
  - The most common way to satisfy this condition is to use a simple random sample from the population.
  - The data can also come from a random process (a.k.a., rolling a die).
- Normality:
  - When a sample is small, we require that the sample observations come from a normal population.
  - We can relax this condition as the sample size grows large.
  - Rule of Thumb: $n \geq 30$

Practicing the Central Limit Theorem 1

Assume that cholesterol levels for women aged 20-34 are approximately normally distributed with a mean of 185 and a standard deviation of 39.

- Write the mathematical shorthand for this distribution.
- What is the probability that a randomly chosen woman's cholesterol is between 125 and 180?
- What is the probability that a randomly chosen woman's cholesterol is above 220?

Practicing the Central Limit Theorem 2

$$N(185, 39)$$

```
pnorm(180, mean = 185, sd = 39) - pnorm(125, 185, 39)
```

```
## [1] 0.3870254
```

```
1 - pnorm(220, mean = 185, sd = 39)
```

```
## [1] 0.1847432
```

Practicing the Central Limit Theorem 3

Now, assume that we have taken a random sample of 9 women aged 20-34.

- Write the mathematical shorthand for this *sampling* distribution.
- What is the probability that the average of the 9 women's cholesterol is above 220?

Practicing the Central Limit Theorem 4

$$\sigma_9 = \frac{39}{\sqrt{9}} = \frac{39}{3} = 13$$

$$N(185, 13)$$

```
1 - pnorm(220, mean = 185, sd = 13)
```

```
## [1] 0.003547972
```

Central Limit Theorem: Implications

- The Central Limit Theorem implies that regardless of what distribution the original population has, the distribution of the sample mean can almost always be treated as normal.
- For samples with large n, there will be less variability than for samples with small n.
- Extreme results are almost always easier to achieve in smaller sample sizes!