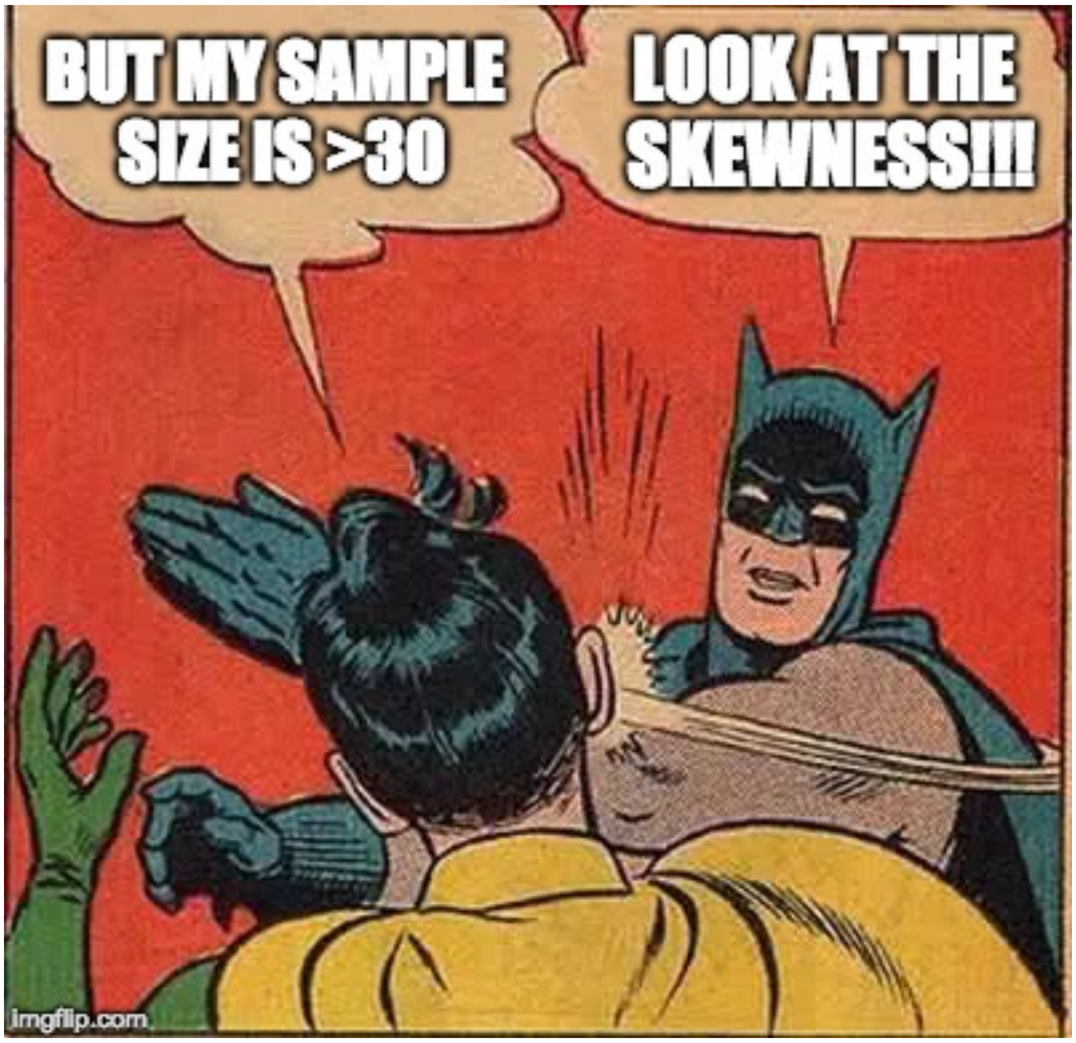
Unit 6 Live Session

W203 Instructional Team

Sampling Distributions and the Central Limit Theorem



(https://imgflip.com/i/2b7ve4)

Class Announcements

1. Lab 1 Grades
2. Research Question

1 Statistic Review

1.1 Statistic: Recall from your study materials that a statistic is some function of a sample $\{X_i\}_{i=1}^n$,

$$\text{Statistic } = f(X_1, X_2, \cdots, X_n)$$

1.2 Sampling Regime: A sample $\{X_i\}_{i=1}^n$ is just a collections of realization from the random variable X collected according to a sampling regime.

Example sampling regimes,

- Choose random person from the phone book and ask their education level, if they have a college degree collect the education level of this persons extended family, if not choose another person at random.
- Choose random person from the phone book and ask their education level, if they have less than a college degree collect the education level of this persons extended family, if not choose another person at random.

1.3 Sample Distribution of $f(\{X_i\}_{i=1}^n)$: The probability of the statistic taking on any value dependes on X, f, and the way $\{X_i\}$ is gathered.

$$\text{Sampling Distribution of} f \ = \ \text{Distribution of } X \ + \ \text{Sampling Regime of } \{X_i\}_{i=1}^n + \text{Properties of} f$$

Change any of these three and the sampling distribution will change.

2 Sampling Distributions

You are conducting an experiment involving flipping a coin several times. You have been assured that the probability of getting a heads is $p = 0.7$. Based on the outcome of that coin flip you compute the value of the following random variable.

$$X = \begin{cases} 0 & \text{if heads} \\ 1 & \text{if tails} \end{cases}$$

2.1 Compute the distribution of the random variable X, then compute the sampling distribution of the average of two observations from X.

In [ ]:

```

```

2.2 In words, what is the difference between the distribution of a random variable X and the sampling distribution of a statistic based on n observations from X?

In [ ]:

2.3 Why do we want to know things about the sampling distribution of a statistic?

In [ ]:

2.4 If you drew 10 observations and got a value of $\overline{X}_{10} = 0.9$ would you be suspicious about whether $p = 0.7$ is in fact true? What if you calculated $\overline{X}_{1000} = 0.9$? What is the difference if any?

In [ ]:

3 Sampling from Bernoulli Distribution

Recall that a Bernoulli random variable with parameter p takes on just two values: 1, with probability p; and 0, with probability $1 - p$. We choose this variable because (1) it's very simple, and (2) its distribution is distinctly non-normal.

Oddly it turns out that (base) R doesn't have a Bernoulli function. To simulate draws from a Bernoulli variable, you can either ...

3.1 Use R's 'sample' command to select values from {0,1}

In [1]:

```
n=3
p = 0.5
sample(c(0,1), 3, prob = c(1-p,p), replace = TRUE)
```

```
0   1   0
```

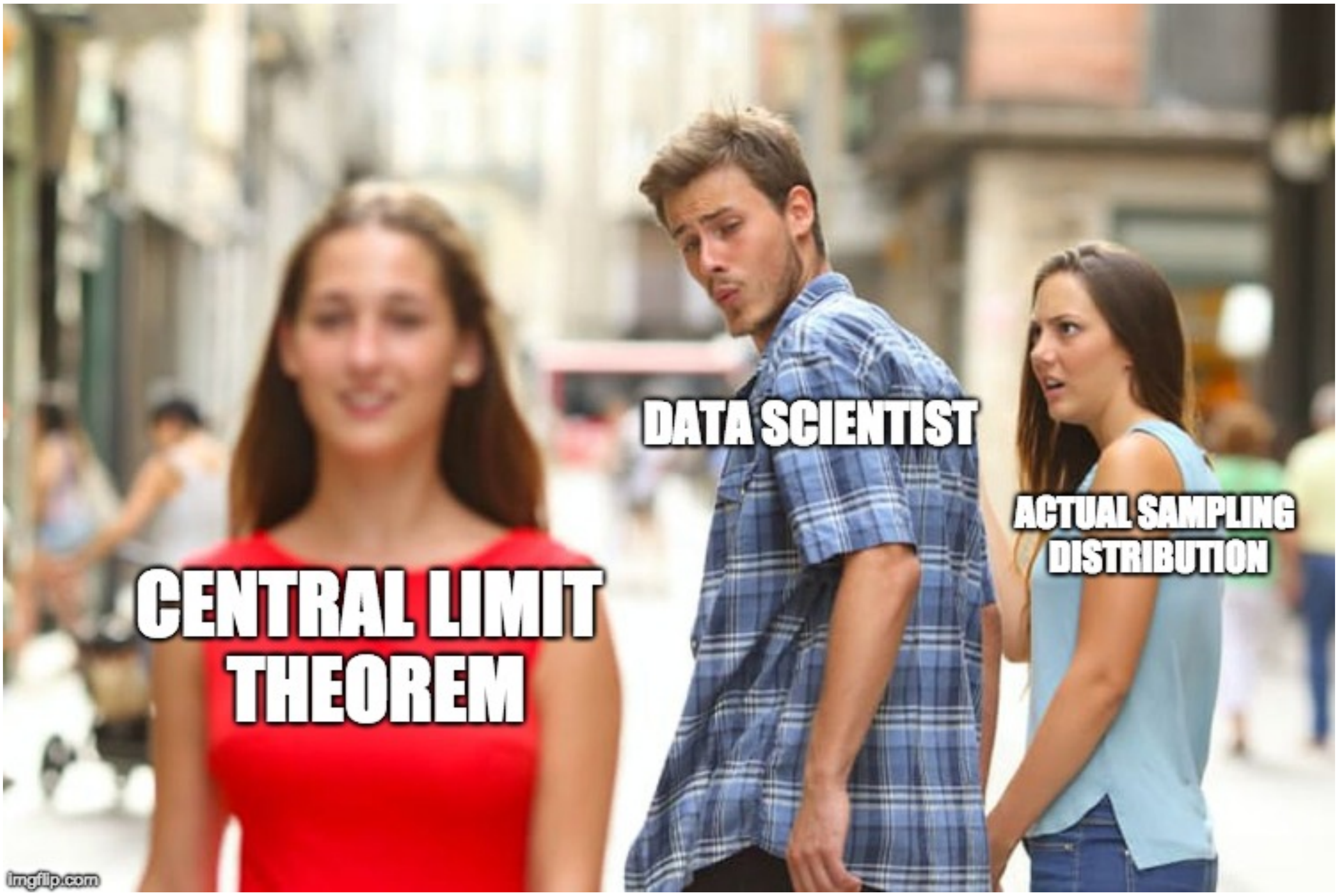
3.2 Note that the Bernoulli distribution is a special case of the more general binomial distribution, with the binomial size parameter set to 1. R has an rbinom function that lets you draw from this distribution.

In [2]:

```
rbinom(3, size=1, prob=p)
```

```
0   0   0
```

4 CLT Review



(https://imgflip.com/i/2b8myf)

By now you may begin to understand that even for the average of observations from a discrete distribution, once the number of observations is larger than 20 or 30 calculating its true sampling distribution can get real complicated/time consuming real quick.

But never fear, we have a result which allows us to approximate the sampling distribution of averages of any sample of observations which satisfy 3 important conditions

4.1 i.i.d. Observations

- *independent observations*: $P\Big([X_i = a] \cap [X_j = b]\Big) = P(X_i = a)P(X_j = b)$ for all $i \neq j$
- *identically distributed*: The distribution of the values of each observation are the same which implies that for all $i \neq j$

$$E(X_i) = E(X_j) = E(X) \quad \text{and} \quad V(X_i) = V(X_j) = V(X)$$

4.2 Lindeberg–Lévy CLT if

- $E(X) = \mu$ and $\infty < \mu < \infty$
- $V(X) = \sigma^2 < \infty$
- $\{X_i\}_{i=1}^{n}$ are i.i.d observations

then,

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

or of more practical use, for n "large"

$$\overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

where $\sim$ means " is approximately distributed as"

5 A Fair Coin

Here we recreate the demonstration of the CLT seen in the async. However instead of using the Old Faithful data, you are to take random draws from a Bernoulli distribution.

5.1 First, set p = 0.5 so your population distribution is symmeteric. Use a variable n to represent your sample size. Initially, set $n = 3$.

In [ ]:

5.2 Simulate n draws from a Bernoulli variable with parameter p, then compute the sample mean.

In [ ]:

5.3 Write code to replicate the above experiment 100,000 times, storing all of the resulting sample means. Create a histogram of your result. Compute the standard deviation of the result.

In [ ]:

5.4 Increase n to 30, replicate the experiment 100,000 times, storing all of the resulting sample means. Create a histogram of your result. Compute the standard deviation of the result.

In [ ]:

5.5 Mathematically, what effect will increasing the number of observations n in each sample have on standard deviation of the sampling distribution of the mean? Do the results of **5.3** and **5.4** reflect this?

In [ ]:

5.6 Experiment with different values of n and note the point at which the sampling distribution of the mean 'looks' normal to you.

In [ ]:

6 A Skewed Distribution

Now let p = 0.001. This result is a highly skewed Bernoulli variable. We are especially interested to see how skewed the sampling distribution of the mean will be for different sample sizes.

For this activity, you can simply assess the skew of a distribution visually. If you prefer, you can also use the skewness command in the moments package. You may hear a rule of thumb that a skewness less than -1 or greater than 1 is considered substantially skewed.

6.1 Rerun the previous exercise for n = 3, p = 0.001 and note the shape of the sampling distribution.

In [ ]:

6.2 Increase n to 30, and note the shape of the sampling distribution.

In [ ]:

6.3 As before, experiment with different values of n and note the point at which the sampling distribution looks normal.

In [ ]:

7.0 Discussion

Here we discuss the result of our analysis.

7.1 How does the skewness of the distribution of the underlying varaible X affect the applicability of the Central Limit Theorem to sample mean $\overline{X}_n$?

In [ ]:

7.2 Why do we care about this / why is this fact important?

In [ ]:

7.3 Name a variable you would be interested in measuring that has a substatially skewed distribution.

In [ ]:

7.4 The Cauchy Distribution is a well-known distribution with some interesting mathematical properties. In particular, it has "infinite" variance That is, the variance does not exist because the tails are too spread out.

Would exercises **5.0** and **6.0** work if you took draws from a Cauchy distribution?

In [ ]: