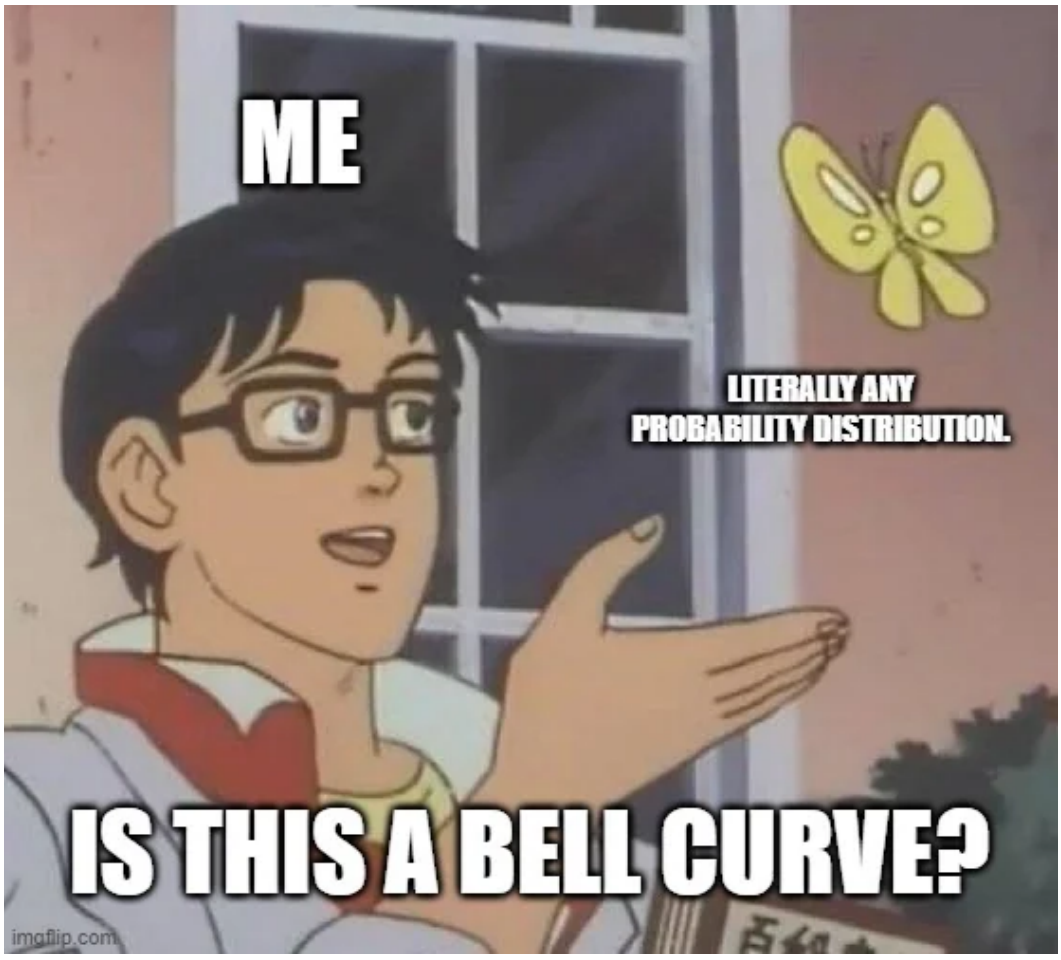
Day 12 - Sampling distributions

Learning outcomes and terms to know

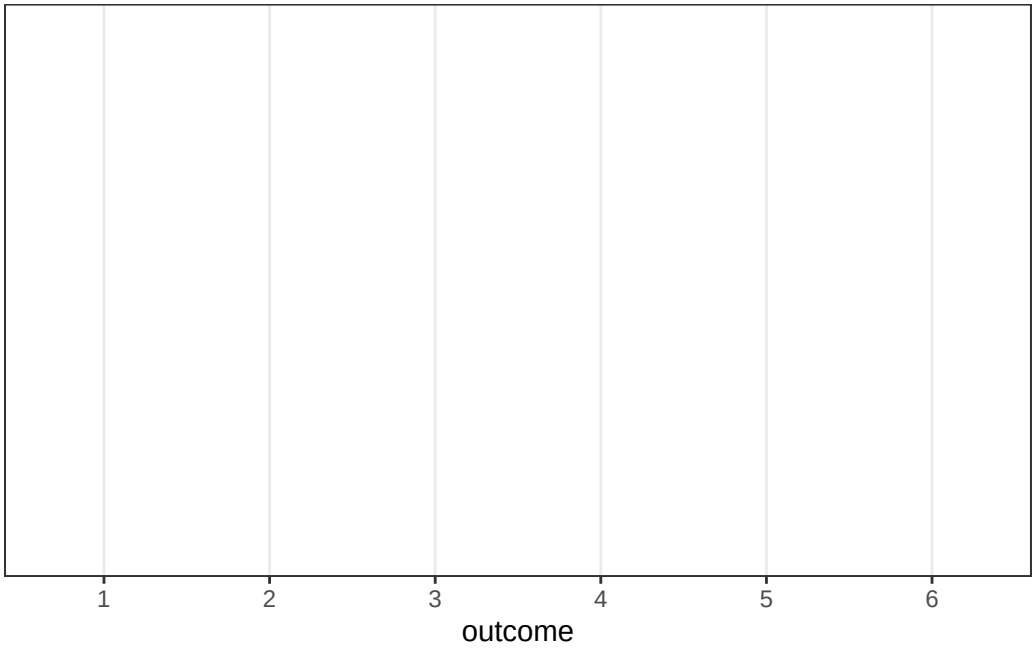
In many ways, understanding sampling distributions is the single most important learning objective of this course; **the purpose of this discipline is to understand the variability of statistics.** In today's activity, we continue to build intuition with sampling distributions and demonstrate the Central Limit theorem in action.



Review: population distributions, data distributions, and sampling distributions

> **Note**
>
> A _____ is the probability distribution of the **entire population** of data. The _____ describes the shape, center, and variability of the observed data.

Sketch the **population distribution** for rolling a single die:



> **Note**
>
> A _____ is the probability distribution of the **observed data**. The _____ describes the shape, center, and variability of the observed data.

> **Instructions**
>
> 1. On your own, roll a die 10 times and record the result of each roll.

Outcomes:

Use the code below to create a plot of the values you recorded in your sample.

> **Important**
>
> You should remove the `eval = F` from the code chunk in each of the code chunks below!

```
df <- data.frame(roll = c()) # enter your rolls here!

df %>%
  ggplot(aes(x = roll, y = after_stat(count)/sum(after_stat(count)))) +
  geom_histogram() +
  scale_x_continuous(breaks = 1:6, limits = c(.5, 6.5)) +
  labs(
    x = "outcome",
    y = "Proportion"
  ) +
  theme_bw()
```

Note

The purpose of statistics is to use data to make inferences about population-level quantities based on sample-level quantities. Suppose we are interested in make inferences about μ, the true mean value of the roll of a die.

Instructions

2. Calculate the mean of your 10 dice rolls, $\bar{x}$, and record it.

Mean:

Use the code below to plot your sample mean.

```
df <- tibble(mean = ) # enter your mean here!

df %>%
  ggplot(aes(x = mean, y = after_stat(count)/sum(after_stat(count)))) +
  geom_histogram(binwidth = 0.5) +
  labs(
    x = bquote("Sample mean (" ~ bar(x) ~ ")"),
    y = "Proportion"
  ) +
  scale_x_continuous(limits = c(0.5, 6.5), breaks = 1:6) +
  theme_bw()
```

> **Note**
>
> The _____ of a statistic is the probability distribution of the statistic. The _____ describes the central tendency and variability of statistics resulting from infinitely many samples of size n from the population.

In plain terms, the sampling distribution of $\bar{x}$ represents all of the possible values of the sample mean, and how likely each value is to occur. Since we are just rolling dice, we can repeat our study to get a sense of the **sampling distribution**.

> **Instructions**
>
> 3. Repeat *Step 1* and *Step 2* four more times: rolling the die 10 times and calculating the mean of the 10 rolls. Record each of your means below:

Use the code chunk below to plot your 5 sample means. What is the shape of this distribution?

```
df <- tibble(mean = c()) # enter your 5 sample means here!

df %>%
  ggplot(aes(x = mean, y = after_stat(count)/sum(after_stat(count)))) +
  geom_histogram(binwidth = 0.5) +
  labs(
    x = bquote("Sample mean (" ~ bar(x) ~ ")"),
    y = "Proportion"
  ) +
  scale_x_continuous(limits = c(0.5, 6.5), breaks = 1:6) +
  theme_bw()
```

Instructions

4. Pair up with 1 other student and combine your data (you should now have 10 sample means to plot). What is the shape of this distribution?

```
df <- tibble(mean = c()) # enter your 10 sample means here!

df %>%
  ggplot(aes(x = mean, y = after_stat(count)/sum(after_stat(count)))) +
  geom_histogram(binwidth = 0.5) +
  labs(
    x = bquote("Sample mean (" ~ bar(x) ~ ")"),
    y = "Proportion"
  ) +
  scale_x_continuous(limits = c(0.5, 6.5), breaks = 1:6) +
  theme_bw()
```

Instructions

5. Create a group of at least 4 students, combine your data, and plot all of your sample means. What is the shape of this distribution?

```
df <- tibble(mean = c()) # enter your sample means here!

df %>%
  ggplot(aes(x = mean, y = after_stat(count)/sum(after_stat(count)))) +
  geom_histogram(binwidth = 0.5) +
  labs(
    x = bquote("Sample mean (" ~ bar(x) ~ ")"),
    y = "Proportion"
  ) +
  scale_x_continuous(limits = c(0.5, 6.5), breaks = 1:6) +
  theme_bw()
```

Instructions

6. Let's combine our data as a class. Use this Google sheet link to add your five means to the class data. What is the shape of the distribution of all sample means from our class?

```
# might need to install the googlesheets4 package!
df <- googlesheets4::read_sheet("https://docs.google.com/spreadsheets/d/1fFXpjAP2D5IK9pybFp7e

df %>%
  ggplot(aes(x = mean, y = after_stat(count)/sum(after_stat(count)))) +
  geom_histogram(binwidth = 0.5) +
  labs(
    x = bquote("Sample mean (" ~ bar(x) ~ ")"),
    y = "Proportion"
  ) +
  scale_x_continuous(limits = c(0.5, 6.5), breaks = 1:6) +
  theme_bw()
```

Note

For this simple example, we are able to replicate the experiment to produce more samples to understand the sampling distribution of the sample mean. But what do we do if we cannot replicate the experiment and produce more samples?

In such a case, we can use **either** 1) computer simulation **or** 2) rely on the theoretical result suggested by the Central Limit Theorem (next class).

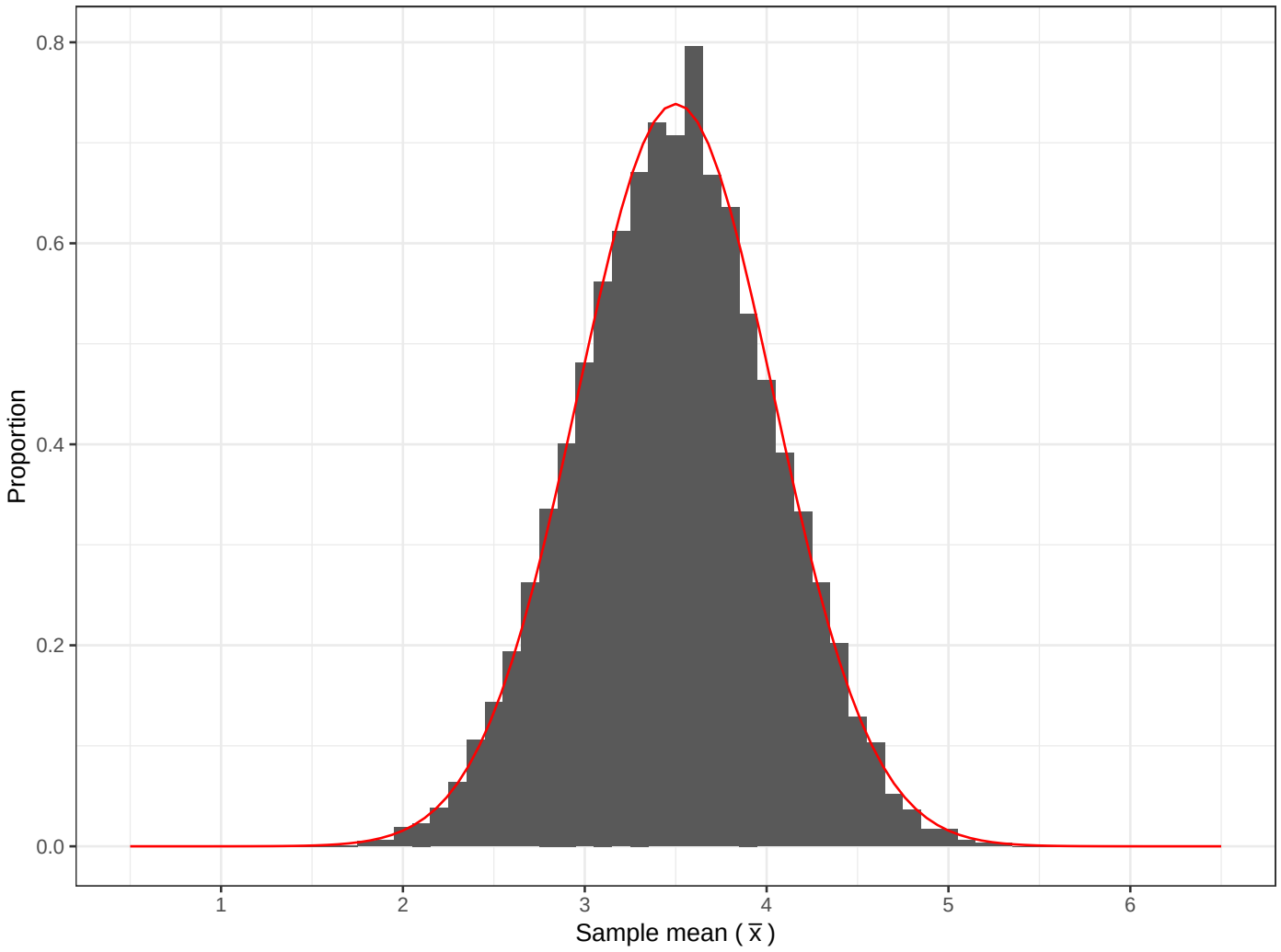
Instructions

7. The code below uses computer simulation to produce 10,000 sample means of 10 dice rolls, and plots that distribution.

```
# simulate means
nsims <- 10000
mean <- NULL
for(i in 1:nsims){
  mean[i] <- sample(1:6, 10, replace = T) %>%
    mean()
}
df <- tibble(mean = mean)

# do some math stat - STAT 310
mu <- sum(1:6 * rep(1/6, 6))
sigma <- sqrt(sum(rep(1/6, 6) * (1:6 - mu)^2))

df %>%
  ggplot() +
  geom_histogram(aes(x = mean, y = after_stat(density)), binwidth = 0.1) +
  stat_function(
    fun = dnorm,
    args = list(
      mean = mu, # theoretical mean - take STAT 310
      sd = sigma / sqrt(10) # theoretical sd - take STAT 310
    ),
    col = "red"
  ) +
  labs(
    x = bquote("Sample mean (" ~ bar(x) ~ ")"),
    y = "Proportion"
  ) +
  scale_x_continuous(limits = c(0.5, 6.5), breaks = 1:6) +
  theme_bw()
```


0.8
0.6
0.4
0.2
0.0
Proportion
1
2
3
4
5
6
Sample mean (x̄)