Sampling and Sampling Distributions

EDP 613

Week 7

A Note About The Slides

Currently the equations do not show up properly in Firefox. Other browsers such as Chrome and Safari do work.

A Note About Probability

We're going to touch on this now but come back to more of it later in the term when talking about Bayesian Statistics.

For Now

- An **event** E is a set of outcomes of an experiment.
- The **probability** P of an event describes how likely it will occur.
- A **sample space** contains all possible outcomes.
- A **probability distribution** gives a probability for each value in a sample space.

Example

- What is the sample space and probability distribution created by tossing a fair quarter?
- Sample space: $\{Heads, Tails\}$
- Probability distribution: $\left\{\frac{1}{2}, \frac{1}{2}\right\}$

Notions

> - The probability of an event is ALWAYS between 0 and 1.

–

> - Assuming all outcomes are likely, the probability P of an event E can be found

$$P(E) = \frac{\text{Number of times an event will happen}}{\text{Total number of events}}$$

Example

- Assume that a standard fair six sided die is rolled. Find the (a) sample space and then (b) the probability that someone will roll a 2.
- (a) The sample space of event $E =$ six sided dice is rolled is $P(E) = \{1, 2, 3, 4, 5, 6\}$
- (b) The probability that someone will roll a 2 is $P(2)$ which can be found by

$$P(2) = \frac{1}{6}$$

On Your Own (More of a Challenge!)

- Assume that a standard fair six sided die is rolled. Find the (a) the probability that someone will roll a 7 and (b) the probability that someone will roll less than a 3.

- (a) The probability that someone will roll a 7 is $P(7)$ which can be found by

$$P(7) = \frac{0}{6}$$

since the sample space is $P(E) = \{1, 2, 3, 4, 5, 6\}$

- (b) The probability that someone will roll less than a 3 is $P(< 3)$ which can be found by

$$P(< 3) = P(1) + P(2)$$

$$= \frac{1}{6} + \frac{1}{6}$$

$$= \frac{2}{6}$$

$$= \frac{1}{3}$$

Note: Reducing Fractions

- Rule: *Always reduce your fractions*
- But why?
- $\frac{2}{6} = \frac{1}{3}$ but what do you lose by reducing?
- The sample size information which seems sort of important!
- New rule: *Only reduce your fractions if it makes sense in context*

Sampling Distributions

> - If several samples are drawn from a population, they are likely to have different values for for the mean $\overline{Y}$. The probability distribution of those means (aka all of the $\overline{Y}$s) is called the **sampling distribution**.

Sampling Distributions: Words and Notation - The Mean

- The mean is calculated the ***exact same way*** as always but
  - is called the *mean of the sampling distribution*
  - has special variables:
    - represented by $\mu_{\overline{Y}}$
    - sample size is specifically for probabilities and represented by M
  - given by the formula:

$$\mu_{\overline{Y}} = \frac{\overline{Y}}{M}$$

Sampling Distributions: Words and Notation - The Standard Deviation

- The standard deviation is calculated the ***exact same way*** as always but
    - is called the *standard error of the mean*
    - has special variables:
        - represented by $\sigma_{\overline{Y}}$
        - sample size is specifically for probabilities and represented by N
    - given by the formula:

$$\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{N}}$$

Central Limit Theorem (CLT)

> - Officially: If $\overline{Y}$ is the mean of a large SRS ($N > 30$) from a population with mean μ and standard deviation σ, as M increases, the distribution becoems normal.
> - In better terms: As you take more samples, especially big ones, your graph of the sample means will look more like a normal distribution.

> - Implications
>   - If you add up the means from all of your samples and find the average, that number will be your *actual population mean*.
>   - If you add up the standard deviations from all of your samples and find the average, that number will be your *actual population standard deviation*.
>   - Helps you predict characteristics os a population.

Procedure for Calculating the CLT

1. Be sure $N > 30$
2. Find $\mu_{\overline{Y}}$ and $\sigma_{\overline{Y}}$
3. Sketch a normal curve and shade in the area to be found.
4. Find the area using The Standard Normal Table (Appendix B).

Example

According to the Nielsen Company, the mean number of TV sets in a U.S. household in 2008 was 2.83. Assume the standard deviation is 1.2. A sample of 85 households is drawn. What is the probability that the sample mean number of TV sets is between 2.5 and 3?

–

1. Clearly $85 > 30$, so we may assume a normal curve.
2. We have

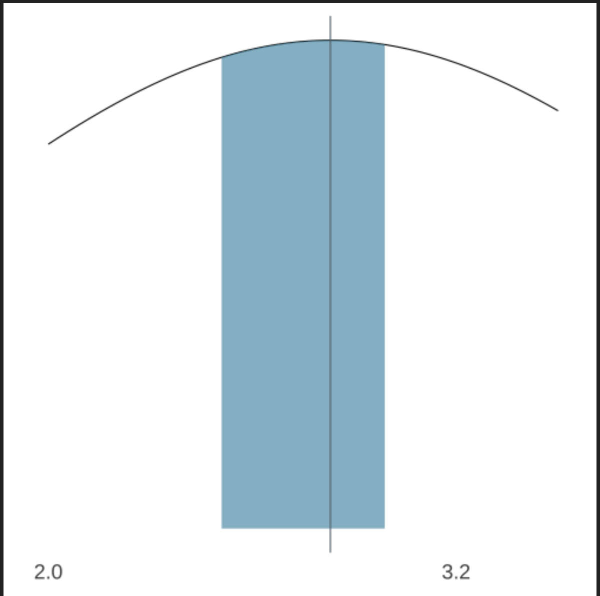
$$\mu_{\overline{Y}} = 2.83$$

   with

$$\sigma_{\overline{Y}} = \frac{1.2}{\sqrt{85}}$$

$$\approx 0.130158$$

3.



4. We have z-scores `

$$
\begin{aligned}
z &= \frac{3 - 2.83}{0.130158} \\
&\approx 1.31
\end{aligned}
$$

$$
\begin{aligned}
z &= \frac{2.5 - 2.83}{0.130158} \\
&\approx -2.54
\end{aligned}
$$

`

So The Standard Normal Table tells us that this is 0.8994. So there was about a 90% chance that a random household had between 2.5 and 3 TVs in 2008.

Example

It is estimated that the mean number of TV sets in a U.S. household in 2020 is 2.00. Assume the standard deviation is 0.8. A sample of 180 households is drawn. What is the probability that the sample mean number of TV sets is still between 2.5 and 3?

–

1. Clearly $180 > 30$, so we may assume a normal curve.

2. We have

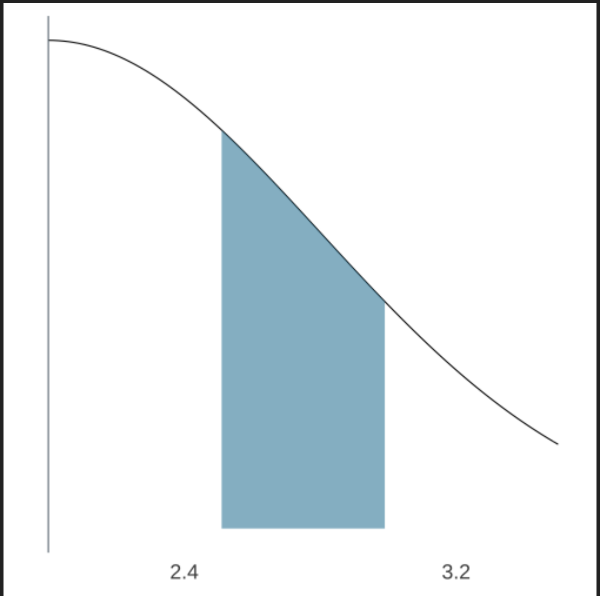
$$\mu_{\overline{Y}} = 2.00$$

   with

$$\sigma_{\overline{Y}} = \frac{0.8}{\sqrt{180}}$$

$$\approx 0.059628$$

3.



4. We have z-scores `

$$
\begin{aligned}
z &= \frac{3 - 2.00}{0.059628} \\
&\approx 16.77
\end{aligned}
$$

$$
\begin{aligned}
z &= \frac{2.5 - 2.00}{0.059628} \\
&\approx 8.38
\end{aligned}
$$

`

```r
pnorm(16.77) - pnorm(8.38)
```

```
## [1] 0
```

R tells us that this is 0. So there is nearly a 0% chance that a random household has between 2.5 and 3 TVs in 2020.

That’s it for sampling today!