HW 06: Foundations for Inference

Pushpak Sunil Rane

Study Design (IMS Ch 2.5)

1 Parameters and statistics.

Identify which value represents the sample mean and which value represents the claimed population mean.

a. American households spent an average of about $52 in 2007 on Halloween merchandise such as costumes, decorations, and candy. To see if this number had changed, researchers conducted a new survey in 2008 before industry numbers were reported. The survey included 1,500 households and found that average Halloween spending was $58 per household.

Claimed population mean: $52 (the average spent of American households in 2007)

Sample mean: $58 (the average spent in the 2008 survey of 1,500 households)

b. The average GPA of students in 2001 at a private university was 3.37. A survey on a sample of 203 students from this university yielded an average GPA of 3.59 a decade later.

Claimed population mean: 3.37 (the average GPA of students at the private university in 2001)

Sample mean: 3.59 (the average GPA from the survey on a sample of 203 students a decade later)

6 Stealers, scope of inference.

In a study of the relationship between socio-economic class and unethical behavior, 129 University of California undergraduates at Berkeley were asked to identify themselves as having low or high social class by comparing themselves to others with the most (least) money, most (least) education, and most (least) respected jobs. They were also presented with a jar of individually wrapped candies and informed that the candies were for children in a nearby

laboratory, but that they could take some if they wanted. After completing some unrelated tasks, participants reported the number of candies they had taken. It was found that those who were identified as upper-class took more candy than others. (Piff et al. 2012)

a. Identify the population of interest and the sample in this study.

Population of interest :- The researchers are interested in all university students or young adults to study how socio-economic class relates to unethical behavior.

Sample :- The study involved 129 undergraduates from the University of the California, Berkeley.

b. Comment on whether the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

The study's results may not apply generalization to all university students since it only has UC Berkeley undergraduate students. The unique environment and characteristics of Berkeley students might affect the results, we need to be careful when we apply these results to larger data or groups.

I think it is hard to derive causes in this study. While there might be link between socio-economic class and candy they take, correlation doesn't mean one causes the other. The study did not count all factors that affects the behavior. In-order to prove causes, a more experiment or long term study is required.

9 Course satisfaction across sections.

A large college class has 160 students. All 160 students attend the lectures together, but the students are divided into 4 groups, each of 40 students, for lab sections administered by different teaching assistants. The professor wants to conduct a survey about how satisfied the students are with the course, and he believes that the lab section a student is in might affect the student's overall satisfaction with the course.

a. What type of study is this?

This is an observational study.

b. Suggest a sampling strategy for carrying out this study.

Stratified random sampling strategy will appropriate to carry out this study.

15 Haters are gonna hate, study confirms.

A study published in the Journal of Personality and Social Psychology asked a group of 200 randomly sampled participants recruited online using Amazon's Mechanical Turk to evaluate how they felt about various subjects, such as camping, health care, architecture, taxidermy, crossword puzzles, and Japan in order to measure their attitude towards mostly independent stimuli. Then, they presented the participants with information about a new product: a microwave oven. This microwave oven does not exist, but the participants didn't know this, and were given three positive and three negative fake reviews. People who reacted positively to the subjects on the dispositional attitude measurement also tended to react positively to the microwave oven, and those who reacted negatively tended to react negatively to it. Researchers concluded that "some people tend to like things, whereas others tend to dislike things, and a more thorough understanding of this tendency will lead to a more thorough understanding of the psychology of attitudes." (Hepler and Albarracín 2013)

a. What are the cases?

The cases are 200 participants who were recruited online using Amazon's Mechanical Turk.

b. What is (are) the response variable(s) and explanatory variable(s) in this study?

Response Variable :- The microwave oven.

Explanatory variable :- Camping, Health care, Architecture, Taxidermy, Crossword puzzles, and Japan

c. Does the study employ random sampling? Explain. How could they have obtained participants?

Yes, the study uses random sampling where participants were chosen from people on Amazon's Mechanical Turk which helps researchers get a diverse group.

d. Can we establish a causal link between the explanatory and response variables?

We cannot prove cause and effect relationship from this study. It shows a link between attitudes and reactions to the microwave but it does not count for all other factors or use an experimental method is much needed to establish cause.

e. Can the results of the study be generalized to the population at large?

The study results may not apply to everyone. As, participants were randomly chosen from Amazon's Mechanical Turk, they may not reflect the general population. People on Mechanical Turk may have different traits, like better internet skills or varying income levels. So, we have to be careful when we apply these results to the large population.

Sampling Distribution Theory (IMS Ch 13.1)

Describe what a sampling distribution is, what the central limit theorem states, and what are the technical conditions that must be met to use the normal model for inference.

:warning: Come prepared to class to share out these answers. You can use chat GPT to aid your understanding, there are also videos in the topic overview webpage. But your writing must be your own.

1. Sampling Distribution

The sampling distribution is a probability distribution that is based on large number of samples of size from given dataset.

It is generally mean of several samples we take from large group of data. It gives an idea about variation of distribution from large data.

The shape, spread, and center of the sampling distribution depend on the sample size and the population distribution

Example: Rolling die multiple times and taking the average of it and you start seeing different average for each rolling die.

2. Central Limit Theorem

The central limit theorem (CLT) tells us if there is change of sample size and number of samples from a population distribution then the mean and standard deviation of samples changes every-time we update the graph and there is variance in height and width of the samples based on different sample size.

As per statistics, it we take many random samples from distribution and calculate averages, then it will form a normal distribution in the form of **"bell curve"**.

3. Conditions for a normal approximation to provide valid inference

1. **Variability of the statistics :** The data points should be independent and identically distributed but not on certain criteria.
2. **Random sampling :** The data must be collected using simple random or stratified random sampling methods to provide valid inference.

The Normal Model (IMS 13.2)

R commands for working with the Normal Model

If $X \sim N(\mu, \sigma^2)$, and P means "the probability of"

1. you can find $P(X < x)$ using the R command `pnorm(x, mu, sigma)`
2. you can find a such that $P(X < a) = q$ using the R command `qnorm(q, mu, sigma)`

1. **What are the two parameters that govern the shape and location of a Normal distribution? Write the symbol, and what it means in words**

   μ is first parameter that is mean and second parameter is σ^2 that is sigma which is variance/standard deviation.

2. **What does $X \sim N(\mu, \sigma^2)$ mean in English? What is this type of notation called? (You may have to Google this)**

$X \sim N(\mu, \sigma^2)$ means that X is normally distributed with mean μ and variance σ

μ is mean notation (mu).

σ is notation for variance/standard deviation(sigma).

3. **What is a Z score? (in a sentence)**

   Z score tells how far a specific number is from the average of group of numbers in terms of standard deviation/variance.

   It is calculated using formula:

   $Z = (X - \mu) / \sigma$

4. **Example: Head lengths of brushtail possums follow a nearly normal distribution with mean 92.6 mm and standard deviation 3.6 mm. Compute the Z scores for possums with head lengths of 95.4 mm and 85.8 mm.**

```
# Z = (x - mu) / sigma
z1 = (95.4 - 92.6) / 3.6
z1
```

```
[1] 0.7777778
```

```
z2 = (85.8 - 92.6) / 3.6
z2
```

```
[1] -1.888889
```

5. **What does the area under the Normal distribution correspond to?**

   The area under Normal Distribution represents probabilities of an event occurring and it always add to 1 which tells that total probability of all possible outcomes is 100%

6. **Let** X be the head length of a brushtail possum, where $X \sim N(92.6, 3.6^2)$ Use R to calculate the probability that a possum will have a head length lower than 85mm. That is, find $P(X < 85)$.

```
pnorm(85,92.6,3.6)
```

```
[1] 0.01738138
```

7. **Would you consider it to be unusual if you found a possum with a head smaller than 85mm? Explain your answer.**

   The probability of a possum having head length of smaller than 85mm is around 0.01738138, which is around 1.7% which indicates that only about 1.7% of possums have head length smaller than 85mm. As per this probability it can be considered unusual to find a possum with a head length less then 85mm.

8. **Suppose test scores are distributed normally with mean of 115 and standard deviation of 15.**

   a. **Let X be the value of the test scores. Write the distributional notation for X.**

      The distributional notation for X :

      $X \sim N(115,\ 15^2)$

   b. **What values bound the middle 75% of test scores? Give a lower and an upper value.**

```
lower_value <- qnorm(0.125,115,15)
lower_value
```

```
[1] 97.74476
```

```
upper_value <- qnorm(0.875,115,15)
upper_value
```

```
[1] 132.2552
```

The values bounding the middle 75% of test scores are approximately 97.75 (lower bound) and 132.26 (upper bound)

c. **What percent of students have test scores less than 83? Show your code and answer**

```
students <- pnorm(83,115,15)
students
```

```
[1] 0.0164487
```

Approximately 1.6% of students have test score less than 83.

9. **A study found that the mean amount of time cars spent in drive-through of a certain fast-food restaurant was 154.6 seconds. Assume drive-through times are normally distributed with a standard deviation of 27 seconds**

   a. **What is the probability that someone spends more than 230 seconds in the drive through?**

```
probability <- 1 - pnorm(230, 154.6, 27)
probability
```

```
[1] 0.002614375
```

The probability that someone spends more than 230 seconds in the drive through is around 0.2%

   b. **How long do 75% of the cars typically have to wait in the drive through? (That is, what is the 75th percentile?)**

```
percentile <- qnorm(0.75, 154.6, 27)
percentile
```

```
[1] 172.8112
```

172.8 is the 75th percentile of the cars typically have to wait in drive through.

11. **Bonita computed a z-score after receiving the results of her exam in order to compare her results with the rest of the class. Her z-score was 2.1. Interpret her Z score in a sentence.**

    Bonita's z-score 2.1 shows that her exam result is 2.1 above standard deviations above the class average. So it means she performed better than the average students in her class.