Statistical Inference

Lecturer: Abdol-Hossein Vahabie

Spring Semester 1401-1402

Writing Assignment I Deadline 1402/01/09

1 Sampling Strategies

In the following examples classify the sampling method as Census, Simple Random Sample, Cluster, Stratified, or Multistage. Please explain why you chose the classification you did. Additionally, do you think this is the correct approach? If not, what approach would you suggest?

1. There is a myth that science students possess a stronger memory than students in other academic fields. Sara wants to test this hypothesis. Therefore, she designs a creativity test and selects thirty students from each department of the university of Tehran as participants. (3 pts)

   Stratified sampling: The population is divided into homogeneous strata (different majors), then Sara randomly samples from within each stratum. This is a suitable sampling strategy for this study.

2. Suppose you are conducting a study on cognitive abilities among undergraduate students in a large city. First, you randomly select ten universities from a list of all universities in that city. Then, within each selected university, you randomly select fifty students. (3 pts)

   Multistage sampling: In the first stage, the universities are selected randomly from a list of all universities in the city, creating clusters of universities. In the second stage, a random sample of students is selected from within each of the selected universities. Keep in mind this method is not Stratified sampling because the universities are not selected based on any predetermined characteristics or variables related to the study, and the sample of students within each university is also not selected based on any specific characteristics or variables. Also, in stratified sampling, each stratum should be sampled.

3. We want to investigate the effects of a new mindfulness-based intervention on the stress levels of employees in a large organization. we could randomly select five departments from the organization, and then invite all employees in those selected departments to participate in the study. We would administer the mindfulness intervention to the selected employees and compare their stress levels to employees in other departments that did not receive the intervention. (3 pts)

   Cluster sampling: In cluster sampling, the population is divided into clusters or groups based on geographic location, administrative divisions, or some other natural grouping. Then, a random sample of clusters is selected, and all individuals within each selected cluster are included in the sample. However, If the clusters that represent the complete population were created based on biased perspectives, then the conclusions drawn about the entire population would also be biased.

To reduce the risk of bias and increase the generalizability of the findings, a better sampling method would be to randomly select a representative sample of employees from all departments in the organization. This can be achieved by using stratified sampling.

4. A team of researchers distributes flyers on the street for partaking in a memory test. (3 pts)

   Simple Random Sample: This method would ensure that every person in the population has an equal opportunity to participate in the study, which would increase the representativeness of the sample and reduce bias. In the scenario described, the sample will be made up of individuals who are willing to stop and participate in the memory test. Since the sample is not selected randomly, it may not be representative of the entire population, and there is a risk of selection bias. One way to conduct random sampling in this scenario would be to generate a list of all eligible individuals in the population, such as residents of a particular area, and then use a random number generator to select a sample of participants from the list.

5. Conducting a survey among the entire student of the University of Tehran to inquire about their predisposition towards participating in a cognitive assessment. (3 pts)

   Census: census is recording and calculating population information about the members of a given population, which in this case, is all the students of University of Tehran. Census is expensive to conduct. Gathering data using this method is labor-intensive and time-consuming, requiring significant manpower. Furthermore, conducting a census investigation presents numerous opportunities for errors to occur. A better method for this experiment could be stratified sampling. We could stratify the student population by academic department, and then randomly select a certain number of students from each department to participate in the survey. This would ensure that the sample includes representation from all departments while also being more manageable and cost-effective than surveying the entire population.

2 Probability Question

The distribution of hospital patients with brain injuries, categorized by age range and the specific area of brain injury within the four main lobes of the brain, is as follows:

	Damaged cerebral lobe				
Age Range	Insular/Limbic	Frontal	Temporal	Parietal	Occipital
00-02 yrs	0.1%	0.7%	1%	3.2%	8.8%
03-12 yrs	0.5%	1.2%	1.5%	1.7%	2%
13-25 yrs	0.3%	6.7%	2.9%	0.5%	0.3%
26+ yrs	4%	36%	28.2%	0.3%	0.1%

1. Given the information that the patient brought to the hospital unconscious, has a damaged working memory, what is the probability that they are in 13-25 age range? (3 pts)

   Damaged working memory is mostly caused by a dysfunction in the temporal lobe of the brain.

So, to answer the question we need to calculate the probability:

$$P(A|B) = \text{The probablity of a patient being in } 13-25 \text{ age range given he/she has dysfunction in temporal lobe}$$

To calculate the above probability, using Bayes' rule we have:

$$P(A|\ B)\ =\ P(A,B)\ /\ P(B)$$

According to the table, we have the probability of a patient being in 13-25 age range and having dysfunction in the temporal lobe:

$$P(A,\ B)\ =\ 2.9\%$$

To calculate the probability that a patient has dysfunction in the temporal lobe, we need to sum the probabilities across all the age ranges for that lobe:

$$P(\text{dysfunction in temporal lobe }) = 1\% +\ 1.5\% +\ 2.9\% +\ 28.2\% =\ 33.6\% =\ 33.6\%$$

So, we get:

$$P(A|B)\ =\ 2.9/33.6\ =\ 0.086 = 8.6\%$$

If your reference implicates the frontal lobe as the source of damaged working memory. The probability of a patient between the ages of 13-25 having damaged working memory is 0.15.

2. What is the probability that a patient will have no dysfunction in the temporal lobe of the brain? (3 pts)

$$P(T|A) = \text{The probablity of a patient having an injury to the temporal lobe if he/she is in the age range } (0-2)$$

$$P(T|B) = \text{The probablity of a patient having an injury to the temporal lobe if he/she is in the age range } (3-12)$$

$$P(T|C) = \text{The probablity of a patient having an injury to the temporal lobe if he/she is in the age range } (13-25)$$

$$P(T|D) = \text{The probablity of a patient having an injury to the temporal lobe if he/she is in the age range } 26+$$

$$P(T) = \text{The probablity of a patient having injury in the Temporal lobe}$$

$$P(\bar{T}) = \text{The probablity of a patient not having injury in the Temporal lobe}$$

according to the table, we can calculate the probability of a patient having an injury to the temporal lobe as below:

$$P(T) = P(T|A)\ +\ P(T|B)\ +\ P(T|C)\ +\ P(T|D)\ =\ 1\%\ +\ 1.5\%\ +\ 2.9\%\ +\ 28.2\%\ =\ 33.6\%$$

So, the probability that a patient will have no dysfunction in the temporal lobe of the brain is:

$$P(\bar{T}) = 1\ -\ P(T)\ =\ 1\ -\ 33.6\% =\ 66.4\%$$

3. What is the probability that an adolescent patient will have Frontotemporal damage? Is this probability

feasible to calculate? (3 pts)

The probability that an adolescent patient will have Frontotemporal damage is not feasible cause we don't have the joint probability of frontal and temporal lobe .

4. What is the probability that a patient will have no injury in any of the cortical lobes of the brain? (3 pts)

   According to the table, the only lobe that is not cortical, is the Insular/Limbic lobe. So, we need to find the complement of the probability that a patient will have an injury in at least one of the cortical lobes, which is calculated as below:

   $$P(A) = P(frontal \cup temporal \cup parietal \cup occipital) = 44.6\% + 33.6\% + 5.7\% + 11.2\% = 95.1\%$$

   $$P(A') = 1 - P(A) = 1 - 95.1\% = 4.9\%$$

3 Confounding Variables

Define the term ”confounding variable” and assess whether there is one present in the following observations: (3pts).

An extraneous variable that affects both the explanatory and the response variable and that makes it seem like there is a relationship between the two is called a confounder or confounding variable.

1. On January 29, 2023, it was observed that the sales of A.S. Roma Football Club's uniforms were several times higher than normal days. On the previous day, a picture of a celebrity wearing Roma's uniform was published on social networks, and many users attributed the rise in sales to this celebrity. (5pts)

   The rise in sales could be due to a significant football match that took place on that day or a promotional campaign launched by the team.

2. On October 27, 2022, Elon Musk was forced to buy Twitter after a long conflict with the former owners. After taking over the company, Elon Musk fired top Twitter execs, including the CEO, as well as many talented engineers and employees. However, on November 17, Musk claimed that Twitter had hit an all-time high in usage. Do you believe this happened because of his ”great” management skills? (5pts)

   It is possible that the increase in usage was due to external factors such as a major news event or viral trend, rather than the changes in leadership at Twitter (i.e., The record-breaking hashtag that originated from Iran's social events.

3. Some studies suggest that the suicide rate is lower during wartime. Do you suspect that there is a confounding factor in these studies? (5pts)

   Yes, it is possible that there is a confounding factor in studies that suggest that the suicide rate is lower during wartime.

For example [3], showed that the decline in suicide rate in the US during World War II was spurious for failing to take into account the decline of the unemployment rate during wartime.
Another possible factor is that families may be hesitant to report suicides during times of war due to cultural and social pressures. Suicide may be stigmatized in some societies, and families may choose not to report a suicide to avoid social ostracism or dishonoring their loved ones.
Another potential factor is that during wartime, people may prioritize survival over seeking help for mental health issues, including suicidal thoughts. Additionally, the sense of purpose and belonging that comes with wartime may act as a protective factor against suicide. It is important to note that any observed decrease in suicide rates during wartime may be temporary, and there could be an increase in suicide rates among veterans and survivors after the conflict ends. This increase may be due to factors such as trauma and other experiences during the war.

4. Some studies have found that overweight or obese patients with heart disease have a lower risk of mortality compared to those who are of normal weight. This paradoxical relationship has also been observed in other conditions such as chronic kidney disease, chronic obstructive pulmonary disease (COPD), and some types of cancer. (5pts)

   The use of BMI as a measure of weight status has limitations, as it does not differentiate between muscle mass and body fat, nor does it take into account differences in body composition or distribution of fat. This could lead to misclassification of individuals into weight categories and potential bias in the results. Another possible confounding factor is that overweight or obese individuals may have a different distribution of body fat than individuals who are of normal weight, with more fat located in subcutaneous areas rather than around organs. This distribution of body fat could impact the risk of mortality for certain conditions.

4 The Marshmallow Experiment

Walter Mischel of Stanford University set out to study whether deferred gratification can be an indicator of future success. During the Marshmallow Experiment conducted in 1972, children between the ages of four and six were brought into a room and presented with a marshmallow on a table in front of them. Prior to leaving the room, the researcher informed each child that they would be given a second marshmallow if they refrained from eating the first one for fifteen minutes until the researcher returned. The researcher recorded the amount of time each child was able to resist eating the marshmallow and also noted whether this ability was associated with success in adulthood. Out of the six hundred children tested, a small proportion immediately consumed the marshmallow, while one-third delayed gratification long enough to receive the second marshmallow. In follow-up studies, Mischel discovered that those who exhibited delayed gratification were notably more competent and achieved higher SAT scores compared to their peers.

1. Define the explanatory variable and the response variable in this experiment: (3 pts)

   The explanatory variable: The ability to delay gratification, i.e., the amount of time each child was able to resist eating the marshmallow.
   The response variable: Academic achievement in adulthood (SAT score)

2. Is this an experimental or an observational study? Why? (3 pts)

This is an experimental study. Experimental studies are ones where researchers introduce an intervention and study its effects. In this case, They wanted to investigate the effect of delayed gratification on academic achievement later in life. In this experiment, we have two groups of children: one group that ate the marshmallow before the researcher returned, and another group that waited for the researcher to come back before eating the marshmallow (control and effect group respectively).

3. Do you see any confounding factors in this study? explain your reasons. (3 pts)

   There could be many confounding factors in this study such as:
   Home environment, Cognitive ability, socioeconomic status

4. How can you improve this experiment? pay attention to every element of the study i.e. sampling strategy, confounding variables, etc. (3 pts)

5. Confounding variables: The marshmallow experiment does not control for several confounding variables that may influence children's ability to delay gratification, such as hunger, tiredness, and cognitive ability. Researchers could address these variables by conducting the study at different times of the day and ensuring that all children have had a similar meal beforehand. Additionally, researchers could measure children's cognitive ability and use it as a covariate in the analysis.

If you are interested, you can read this [4] article that investigates the flaws in the marshmallow experiment.

5 Misinformation Analysis

Select a piece of news that contains statistical information and critically analyze it. Look for sources of bias, statistical mistakes, fallacies, misinformation, misleading graphs, and other common deficiencies. Please provide your news source link. (Attach the news link from the news agency e.g. Kayhan website). (8pts)

Fox News tried to show the change in the top tax rate if the Bush tax cuts expire, so they showed the rate now and what'd it be in 2013. Wow, it'll be around five times higher. However, if you pay attention you find out that, The value axis starts at 34 percent instead of zero, which you don't do with bar charts because length is the visual cue. That is to say, when you look at this chart, you compare how high each bar is. Fox News might as well have started the vertical axis at 34.9 percent. That would've been more dramatic.
Fox News attempted to demonstrate the potential impact of the expiration of the Bush tax cuts on the top tax rate by comparing the current rate to what it would be in 2013. The resulting figure was shocking, showing a tax rate increase of almost five times the current rate. However, closer inspection reveals that the chart is misleading because its value axis starts at 34 percent instead of zero. This approach is problematic for bar charts, where length is the primary visual cue, as it causes viewers to compare the height of each bar rather than the difference between the starting and ending points of the bars. In other words, Fox News could have further exaggerated the effect by starting the vertical axis at 34.9 percent. For more examples of misleading data and visualizations, check out the website Flowing Data.

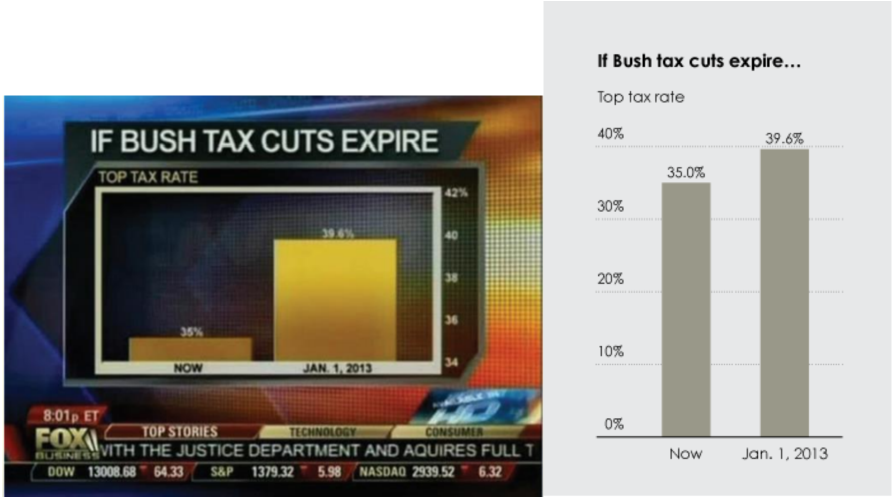

Figure 1: Example of Bar plot abuse

6 Cognitive Abilities (R)

In a research from Educational Psychology journal [1], cognitive abilities was classified into five categories: memory ability (MA), representational ability (RA), information processing ability (IPA), logical reasoning ability (LRA), and thinking conversion ability (TCA). The study analyzed how these five abilities impacted academic achievement and provided a dataset, which can be accessed [2].

1. Prior to creating histograms for each of the five cognitive abilities, could you make an educated guess regarding the distribution of the data? (6 pts)

   It is commonly assumed that human performance follows a normal distribution, also known as a Gaussian distribution or a bell curve. This means that most people fall near the average or mean performance, with fewer people performing either exceptionally well or exceptionally poorly. For example, in academic testing, it is often assumed that scores follow a normal distribution, with most students receiving scores around the average, and fewer students receiving scores that are much higher or much lower than the average. Similarly, in sports, it is assumed that most athletes fall within a range of average performance, with fewer athletes performing at the extremes.

   It is important to note that any observed decrease in suicide rates during wartime may be temporary, and there could be an increase in suicide rates among veterans and survivors after the conflict ends. This increase may be due to factors such as trauma and other experiences during the war.

2. Create a histogram to visualize the distribution of each attribute and analyze its skewness and modality in your discussion. (6 pts)

```
df<-read.csv("Table1.csv")
par(mfrow = c(4, 4),main="hists")
for(i in names(df))
  hist(df[[i]], main = i, xlab = "Value")
title("Hisogram for all the columns", line =
     -2, outer = TRUE)
```

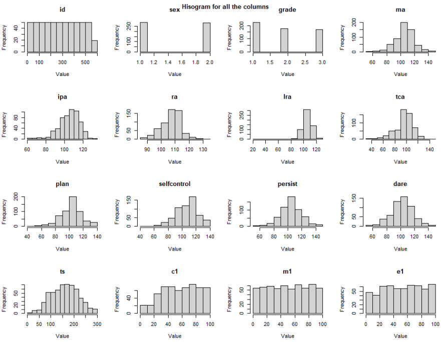

Figure 2: Sample Box-plot

id: In most datasets, an ”id” or ”ID” typically refers to a unique identifier assigned to each individual record or data point in the dataset. This identifier is used to distinguish and track each data point, and it enables the dataset to be organized and analyzed more effectively. Therefore, this data is not useful for us.
Sex: The distribution of both genders is almost equal (uniform) in our dataset. The histogram is bimodal and without skewness. However, this data is not important for our purpose.
Grade: The distribution of all grade is almost equal(uniform) in our dataset. The histogram is multi-modal and without skewness. However, this data is not important for our purpose.
MA: The distribution is almost normal. The histogram is unimodal and without skewness.
IPA: The distribution is almost normal. The histogram is unimodal and lef skewed.
RA: The distribution is almost normal. The histogram is unimodal and without skewness.
LRA: The distribution is almost normal. The histogram is unimodal and lef skewed.
TCA: The distribution is almost normal. The histogram is unimodal and lef skewed.
Plan: The distribution is almost normal. The histogram is unimodal and lef skewed.
Self-control: The distribution is almost normal. The histogram is unimodal and lef skewed.
Persist: The distribution is almost normal. The histogram is unimodal and without skewness.
Dare: The distribution is almost normal. The histogram is unimodal and without skewness.
e1: The distribution is almost uniform. The histogram is unimodal and without skewness.
m1: The distribution is almost uniform. The histogram is unimodal and without skewness.
c1: The distribution is almost uniform. The histogram is unimodal and without skewness.
ts: The distribution is almost normal. The histogram is unimodal and without skewness.

3. Create a pie chart to illustrate the frequency of each grade, ensuring that each category is represented by a unique color and percentage value. Additionally, include a legend for the chart. (6 pts)

```
grades_table <- table(df$grade)
grades_table_headers <- as.vector(names(
    grades_table))
grades_freq = as.vector(as.numeric(
    grades_table))
grades_precentages <- round((grades_freq/sum
    (grades_freq)*100), 1)
slices_labels <- paste(grades_precentages, "
    %", sep="")
pie(grades_freq, main="Pie chart of grades",
     col= c("#EC6B56",  "#FFC154", "#47B39C"
    ), labels=slices_labels, cex=1.2)
legend(1.0, .1, paste("grade = ",
    grades_table_headers, sep=""), cex = 1,
    fill = c("#EC6B56",  "#FFC154", "#47B39C
    "))
```

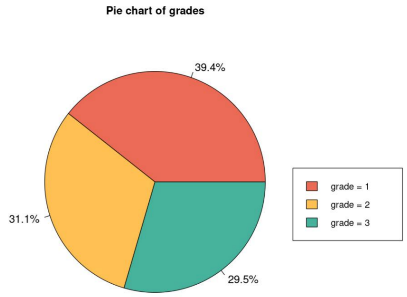

Figure 3: Pie chart

4. Construct a boxplot for the Total Score (ts) of each gender (sex) group, and use your visual analysis to determine if there are any gender-based differences in academic performance? Later, in this course, you will learn to statistically prove your conclusion. (6 pts)
   By visually examining the boxplot, it can be inferred that there is a negligible difference between the academic achievement of the two genders. Thus, it can be concluded that gender does not appear to have a significant impact on academic achievement.

```
boxplot(df$ts
    dfsex, df = data, col = c(”619CFF”, ”FF6666”), main =
    ”TotalScorebyGender”, xlab = ”TotalScore”, ylab =
    ”Gender”, horizontal = TRUE)
```

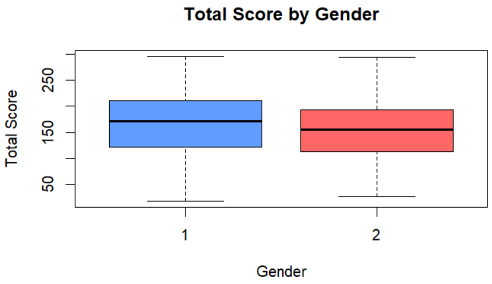

Figure 4: Box plot

5. The authors of the paper hypothesized that cognitive abilities positively influence academic achievement. Use visual analysis to explore this idea. Later, in this course, you will learn to statistically prove your conclusion. (6 pts)

   Although there appears to be a correlation between cognitive abilities and academic achievement, it is important to remember that correlation does not necessarily imply causation. Therefore, based on this plot alone, it would be wrong to draw any definitive conclusions.

```
plot(x =df$ts , y = df$ra , main = "
    influence of RA on academic achievement
    ", xlab = "RA", ylab = "TS", col ="#1
    f485b", pch = 16)
```

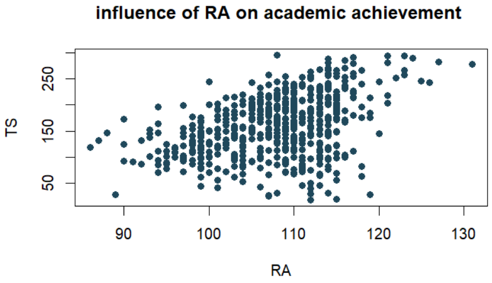

Figure 5: Influence of RA on academic achievement

Required Document

[1] Shi, Y., Qu, S. (2022). Analysis of the effect of cognitive ability on academic achievement: Moderating role of self-monitoring. In Frontiers in Psychology (Vol. 13). Frontiers Media SA. https://doi.org/10.3389/fpsyg.2022.996504

[2] Download dataset

[3] Marshall JR. Political Integration and the Effect of War on Suicide: United States, 1933–76. Soc. Forces. 1981;59(3):771–85.

[4] The original marshmallow test was flawed, researchers

General Rules

Please upload a file in ZIP format (~~not RAR~~) to the elearn platform(`https://elearn5.ut.ac.ir/course/view.php?id=14838).`

You are allowed a total grace period of 3 days to submit late assignments for all of your exercises.

It is prohibited to use ~~handwritten material~~ and only material produced through typing in the HW template is permissible.

Utilizing a LaTeXto compose the report will grant an additional 5 points.

Please use only built-in R packages for plotting, and ensure that each plot has a title and appropriate legends, if needed and do not use non-built-in packages such as ~~ggplot2~~

Deadline

Wednesday 23:59. 1402/01/09.

Contact Information

Please direct your questions regarding Homework 1 only to the teaching assistants, Mohammad Javad Ranjbar and Sara Rostami, through the course mail (statistical.inference.ut@gmail.com). Use ”HW1” as the subject line.

Good Luck