Midterm Practice Exam Solutions

Midterm Practice Exam Solutions

Introduction to Data Science

Question 1a: Difference Between Population and Sample

A **population** refers to the entire set of individuals or items of interest in a study. It includes every possible subject or item under investigation. For example, if you are studying the voting behavior of all registered voters in a country, the population would be all registered voters.

A **sample** is a subset of the population that is actually observed or measured. The goal of using a sample is to draw conclusions about the entire population without having to collect data from every single individual. For example, in the voting behavior study, a sample could be 1,000 randomly selected voters. Researchers use samples because it is often impractical to study the entire population due to time or cost constraints.

Question 1b: Difference Between Simple Random Sampling and Stratified Random Sampling

Simple Random Sampling (SRS): - In simple random sampling, each member of the population has an equal chance of being selected. There is no consideration of any underlying structure in the population. - For example, if we wanted to select a sample of voters, we could randomly pick names from a list, and each voter would have the same probability of being chosen.

Stratified Random Sampling: - Stratified random sampling involves dividing the population into distinct subgroups (called strata) based on a certain characteristic (such as age, gender, or income level). Then, samples are taken from each subgroup, often proportionally to the size of the group. - For example, if we believe that voters from different age groups may have different voting patterns, we could divide the population into age groups and select a random sample from each group.

Which is better? - **Stratified random sampling** is often better when there are distinct subgroups within the population that are important to represent in the sample. It ensures that all subgroups are included, which often leads to more accurate estimates and smaller sampling errors. In contrast, simple random sampling can sometimes miss important subgroups or result in imbalanced samples.

Question 2a: Hypothetical Data Record Table

Plant ID	Type of Plant	Survival Time (days)	Alive/Dead
1	A	45	Dead
2	A	45	Dead
3	A	43	Dead
4	B	40	Dead
5	B	44	Dead
6	B	45	Dead

This table represents the outcome of a study where six plants (three of type A and three of type B) were subjected to water deprivation, and their survival times were recorded.

Question 2b: Statistical Units and Variables

- **Statistical units:** The individual plants. Each plant represents a unit of observation.
- **Variables:**
    1. **Type of plant:** This variable indicates whether the plant is of type A or type B. It is a nominal categorical variable because there is no inherent order to the categories.
    2. **Survival time:** This is the number of days the plant survived during the study. It is a continuous numerical variable because it can take any value within a range (e.g., 40, 43, or 45 days).
    3. **Alive/Dead status:** This indicates whether the plant was alive or dead at the end of the study. It is a nominal categorical variable (alive or dead).

Question 2c: Types of Data and Censoring/Truncation

1. **Type of plant:** This is a **nominal categorical** variable, as there is no order or ranking between plant types A and B.
2. **Survival time:** This is a **continuous numerical** variable. In this case, the data may also be **censored** because some plants were still alive at the end of the study (45 days). This means their actual survival time could not be fully observed.
3. **Alive/Dead:** This is a **nominal categorical** variable indicating the final status of the plant.

Question 3a: Distributions of X, Xc, and Xn

- $X \sim Binomial(n = 10,000, p)$ represents the number of people with perfect pitch in Pollster 1's simple random sample. The distribution of X is binomial because we are counting the number of successes (people with perfect pitch) out of 10,000 independent trials, where each trial has the same probability p of success.
- $X_c \sim Binomial(n = 5, p_c)$ represents the number of choir members with perfect pitch in Pollster 2's stratified sample.
- $X_n \sim Binomial(n = 995, p_n)$ represents the number of non-choir members with perfect pitch in Pollster 2's stratified sample.

Question 3b: Estimates of p

- The simple random sample estimate of the fraction of people with perfect pitch is:

$$\hat{p_1} = \frac{X}{10,000}$$

- The stratified random sample estimate of the fraction of people with perfect pitch is:

$$\hat{p_2} = \frac{X_c + X_n}{1000}$$

Question 3c: Mean and Standard Deviation

Mean:

- The mean of $\hat{p_1}$ is p as $E[p_1] = \frac{10000 \times p}{10000}$
- The mean of $\hat{p_2}$ is $\frac{5 \times p_c + 995 \times p_n}{1000}$ by just expanding the expected value of $ p_2$.

Standard Deviation:

- The standard deviation of $\hat{p_1}$ is given by the formula:

$$\sigma_{\hat{p_1}} = \sqrt{\frac{p(1-p)}{10,000}}$$

- The standard deviation of $\hat{p_2}$ is given by:

$$\sigma_{\hat{p_2}} = \sqrt{\frac{5 \cdot p_c(1-p_c) + 995 \cdot p_n(1-p_n)}{1000^2}}$$

Question 3f: Calculation of Confidence Intervals and Estimates

(a) Estimates of Pollster 1 and Pollster 2:

- Pollster 1's estimate is $\hat{p_1} = \frac{19}{10,000} = 0.0019.$
- Pollster 2's estimate is $\hat{p_2} = \frac{4+2}{1000} = 0.006.$

(b) Standard deviations for both estimators:

- For Pollster 1, using $\hat{p_1} = 0.0019$:

$$\sigma_{\hat{p_1}} = \sqrt{\frac{0.0019 \cdot (1-0.0019)}{10,000}} = 0.000435$$

- For Pollster 2, using $\hat{p_c} = 0.8$ and $\hat{p_n} = 0.00201$:

$$\sigma_{\hat{p_2}} = \sqrt{\frac{5 \cdot 0.8 \cdot (1-0.8) + 995 \cdot 0.00201 \cdot (1-0.00201)}{1000^2}} = 0.00167$$

(c) Pollster 1's 90% confidence interval:

For Pollster 1:

$$\mathbf{E}[\hat{p_1}] \pm 1.645 \cdot \sigma_{\hat{p_1}} = 0.0019 \pm 1.645 \cdot 0.000435 = (0.001185, 0.002615)$$

(d) Pollster 2's 90% confidence interval:

For Pollster 2:

$$\mathbf{E}[\hat{p_2}] \pm 1.645 \cdot \sigma_{\hat{p_2}} = 0.006 \pm 1.645 \cdot 0.00167 = (0.0032, 0.00879)$$