Review midterm 1

Hanying

HW 4 common mistakes

Sample mean / Population mean

Sample mean Sample mean is a descriptive statistics. $\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$.

Population mean (Expectation) Population mean is a population parameter. Theoretically, it represents mean of an infinite number of realizations of X. Computationally, $\mu_x = E(X) = \sum_x x * P(X = x)$.

HW2 Exercise 2

The table below tabulates the number of errors detected on the 100 disks produced in a day.

Number of Defects	Number of Disks
0	41
1	31
2	15
3	8
4	5

Calculate the mean number of errors detected on the 100 disks.

Sample mean / Population mean

Sample mean Sample mean is a descriptive statistics. $\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$.

Population mean (Expectation) Population mean is a population parameter. Theoretically, it represents mean of an infinite number of realizations of X. Computationally, $\mu_x = E(X) = \sum_x x * P(X = x)$.

HW2 Exercise 1

There are $n = 12$ numbers in a sample, and the mean is $\bar{x} = 24$. The minimum of the sample is accidentally changed from 11.9 to 1.19. Is it possible to determine the direction (increase/decrease) in which the mean $\bar{x}$ changes? And how much the mean changes? If so, by how much does it change? If not, why not?

HW2 Exercise 3

A certain reaction was run several times using each of two catalysts, A and B. The catalysts are supposed to control the yield of an undesireable side product. Results, in units of percentage yield, for 25 runs of catalyst A and 23 runs of catalyst B are given below and also in `Catalysts.csv`.

In question (c), we obtained that the mean of observed catalyst A is 4.148, and the mean of observed catalyst B is 4.073913. Calculate the mean of combined data from the summary measures in part (c) along with the sample sizes.

Probability (HW 3 Exercise 1)

A geneticist is studying two genes. Each gene can be either dominant or recessive. A collection of 100 individuals is categorized and found to have 58 individuals with both genes dominant, 6 individuals with both genes recessive and a total of 70 Gene 2 dominant individuals.

X	Gene 2 Dominant	Gene 2 Recessive	Total
Gene 1 Dominant	58	24	82
Gene 1 Recessive	12	6	18
Total	70	30	100

Marginal probability $P(A) = \frac{n(A)}{n}$.

Example: What is the probability that a randomly sampled individual from this study has Gene 1 dominant?

$82/100 = 0.82$.

And $P(A \text{ and } B) = \frac{n(A \text{ and } B)}{n}$

Example: What is the probability that a randomly sampled individual from this group has Gene 1 and Gene 2 dominant?

$58/100 = 58/100$.

Or $P(A \text{ or } B) = \frac{n(A \text{ or } B)}{n} = P(A) + P(B) - P(A \text{ and } B)$

Example: What is the probability that a randomly sampled individual from this group has Gene 1 or Gene 2 dominant?

$(58 + 24 + 12)/100 = 94/100$.

$82/100 + 70/100 - 58/100 = 94/100$.

When $P(A \text{ or } B) = P(A) + P(B)$? - Mutually exclusive.

X	Gene 2 Dominant	Gene 2 Recessive	Total
Gene 1 Dominant	58	24	82
Gene 1 Recessive	12	6	18
Total	70	30	100

Conditional probability $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$.

$(P(A \text{ and } B) = P(B) * P(A|B).)$

Example: What is the probability that a randomly sampled individual from this study has Gene 2 dominant, given that we know they have Gene 1 dominant?

$P(Gene2Dom|Gene1Dom) = 58/82 = 0.7073.$

Independence

$$\begin{aligned} & A \text{ and } B \text{ are independent} \\ \iff & P(A|B) = P(A) \\ \iff & P(B|A) = P(B) \\ \iff & P(A \text{ and } B) = P(A) \times P(B) \end{aligned}$$

Example: The genes are said to be in linkage equilibrium if the event that Gene 1 is dominant is independent of the event that Gene 2 is dominant. Are these genes in linkage equilibrium in this group of 100 individuals?

in linkage equilibrium $\iff$ Gene 1 is dominant is independent of the event that Gene 2 $\iff$ $P(\text{Gene 2 Dominant}|\text{Gene 1 Dominant}) = P(Gene2Dominant)$.

$P(\text{Gene 2 Dominant}|\text{Gene 1 Dominant}) = 0.7073 \neq P(Gene2Dominant) = 0.70.$

Not in linkage equilibrium.

X	Gene 2 Dominant	Gene 2 Recessive	Total
Gene 1 Dominant	??		
Gene 1 Recessive		6	
Total	70		100

Independence

$$\begin{aligned} & A \text{ and } B \text{ are independent} \\ \iff & P(A|B) = P(A) \\ \iff & P(B|A) = P(B) \\ \iff & P(A \text{ and } B) = P(A) \times P(B) \end{aligned}$$

Example: . How many individuals in this sample would have both genes dominant if the event of Gene 1 dominant is independent of the event of Gene 2? Make sure to show how you calculated your answer.

Other topics

- Concepts: Parameter vs. statistic, population, type of variables, ...
- Know how to read a figure and table
  - Know how to interpret a figure (mainly histograms and boxplots)
  - Know how to comment a figure (compare the spread, the skewness, the outliers,...)
  - ...
- Sampling with and without replacement
- Understand binomial distribution
  - Understand how to identify the binomial variable in the context
  - Understand how to identify the parameter n and π in the context
  - Understand how to transfer the question into language of statistics (For example, the probability of at least two of the selected ... to $P(Y \geq 2)$, and is equivalent to $P(Y > 1)$ in binomial distribution.)
  - Understand how to compute the probability (either use the definition and a calculator or R with `pbinom` and `dbinom`)
- Transformations/Rescaling (table in P36 2023.09.26)
- CLT: understand that when sample size is large, we can use the normal distribution to approximate the distribution of the sample mean/sum. With Expectation and SD computed from previous rules, we can get the approximated distribution of the mean/sum.
- Understand Normal distribution
  - Know how to compute the probability and quantiles in the normal distirbution with the table of `pnorm` and `qnorm`.
  - Understand Z-score. Know how to use the Z score and the Z distribution to compute the probability and quantiles.
- ...