# Continuous Probability Distribution and Confidence Interval

**Instructions:**

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name: Vishvash C Batch ID:** 23012024

**Topic: Continuous Probability Distribution and Confidence Interval**

**Guidelines:**

**1. An assignment submission is considered complete only when the correct and executable code(s) and documentation explaining the method and results are submitted. Failing to submit either of those will be considered an invalid submission and not a correct submission.**

**2. Ensure that you submit your assignments correctly and in full. Resubmission is not allowed.**

**3. Post the submission you can evaluate your work by referring to the keys provided. (will be available only post the submission).**

**Hints:**

1. Business Problem
   1. Objective
   2. Constraints (if any)
2. For each assignment the solution should be submitted in the below format
3. Research and Perform all possible steps for obtaining a solution.
4. For Basic Statistics explanation of the solutions should be documented in black and white along with the codes.

One must follow these guidelines as well:

* 1. Be thorough with the concepts of Probability, and Central Limit Theorem and Perform the calculation stepwise.
  2. For True/False Questions, the explanation is a must.
  3. Python code for Univariate Analysis (histogram, box plot, bar plots, etc.) for data distribution to be attached.

1. All the codes (executable programs) should execute without errors.
2. For each of the following statements, indicate whether it is True/False. If false, explain why.

The sample size of the survey should at least be a fixed percentage of the population size to produce representative results.

**False**. The sample size required for a survey depends on various factors such as the population size, the desired level of confidence, and the acceptable margin of error. There's no fixed percentage rule; instead, sample size determination involves statistical calculations based on these factors.

The sampling frame is a list of every item that appears in a survey sample, including those that did not respond to questions.

**True**. The sampling frame includes all elements in the population from which the sample is drawn, regardless of whether they respond to the survey or not.

Larger surveys convey a more accurate impression of the population than smaller surveys.

**True**. Generally, larger surveys tend to provide more precise estimates of population parameters compared to smaller surveys, assuming that both are well-designed and representative of the population.

1. *PC Magazine* asked all of its readers to participate in a survey of their satisfaction with different brands of electronics. In the 2004 survey, which was included in an issue of the magazine that year, more than 9000 readers rated the products on a scale from 1 to 10. The magazine reported that the average rating assigned by 225 readers to a Kodak compact digital camera was 7.5. For this product, identify the following:

Population: Readers of PC Magazine

Parameter of interest: Average rating for the Kodak compact digital camera, 7.5

Sampling frame: All readers who participated in the survey

Sample size: 225

Sampling design: Convenience sampling (since all readers were asked to participate)

Potential sources of bias: Self-selection bias (readers who chose to participate may have different opinions than those who did not), non-response bias (if only a subset of readers responded), social desirability bias (readers may provide socially desirable ratings).

1. Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate 94%,98%,96% confidence interval?

import scipy.stats as stats

# Given data

sample\_mean = 200 # in pounds

sample\_std\_dev = 30 # in pounds

sample\_size = 2000

# Calculate confidence intervals

conf\_levels = [0.94, 0.98, 0.96]

for conf\_level in conf\_levels:

# Find the critical value for the t-distribution

t\_value = stats.t.ppf((1 + conf\_level) / 2, sample\_size - 1)

# Calculate margin of error

margin\_of\_error = t\_value \* (sample\_std\_dev / (sample\_size \*\* 0.5))

# Calculate confidence interval

lower\_bound = sample\_mean - margin\_of\_error

upper\_bound = sample\_mean + margin\_of\_error

print(f"{conf\_level \* 100}% confidence interval: ({lower\_bound}, {upper\_bound})")

**Output:**

94.0% confidence interval: (198.7376089443071, 201.2623910556929)

98.0% confidence interval: (198.4381860483216, 201.5618139516784)

96.0% confidence interval: (198.6214037429732, 201.3785962570268)

1. What are the chances that ?
2. ¼
3. ½
4. ¾
5. 1

﻿

In a normal distribution, the sample mean (xbar) and the population mean (μ) lie close to each other most frequently. Therefore, the probability that xbar> μ is equivalent to the probability that a random sample mean is greater than the population mean.

Since the sample mean and population mean lies close to each other in a normal distribution, the probability that xbar > μ approximately 50%. This means that the chances that > μ is:

B. 1⁄2

1. A book publisher monitors the size of shipments of its textbooks to university bookstores. For a sample of texts used at various schools, the 95% confidence interval for the size of the shipment was 250 ± 45 books. Which, if any, of the following interpretations of this interval are correct?

A. Incorrect. The confidence interval only tells us about the range of values that the population mean is likely to fall within, not about individual shipments.

B. Incorrect. The confidence interval does not provide information about individual shipments; it's about the population mean.

C. Correct. This is the correct interpretation of a 95% confidence interval. It means that if we were to take many samples and calculate a confidence interval for each one, approximately 95% of those intervals would contain the true population mean.

D. Incorrect. The confidence interval is about the population mean, not about individual sample means.

E. Incorrect. This interpretation is incorrect as it provides a different range than the one specified (205 to 295) and misrepresents the confidence level associated with the interval.

1. Which is shorter: a 95% *z*-interval or a 95% *t*-interval for *μ* if we know that σ =s?
2. The z-interval is shorter
3. The t-interval is shorter
4. Both are equal
5. We cannot say
6. z-interval is shorter

This is because when the population standard deviation is the sample standard deviation, the z- distribution or t-distribution can be employed. When t-distribution is employed, it has heavier tails compared to the normal distribution (z-distribution). The heavier tails of the t-distribution result in wider intervals, making the t-interval longer than the z-interval in this scenario.

1. How many randomly selected employers (minimum number) must we contact to guarantee a margin of error of no more than 4% (at 95% confidence)?

A. 600

B. 400

C. 550

D. 1000

import math

# Given data

Z = stats.norm.ppf(0.975) # Z-score for 95% confidence level

p = 0.5 # Assuming maximum variability

E = 0.04 # Margin of error

# Calculate minimum sample size

minimum\_sample\_size = math.floor((Z \*\* 2 \* p \* (1 - p)) / E \*\* 2)

print("Minimum number of employers:", minimum\_sample\_size)

**Output:**

Minimum number of employers: 600

Questions 8 and 9 are based on the following: To prepare a report on the economy, analysts need to estimate the percentage of businesses that plan to hire additional employees in the next 60 days.

1. Suppose we want the above margin of error to be based on a 98% confidence level. What sample size (minimum) must we now use?

A. 1000

B. 757

C. 848

D. 54

import scipy.stats as stats

import math

# Given data

confidence\_level = 0.98 # 98% confidence level

Z = stats.norm.ppf(0.99) # Z-score for 98% confidence level

p = 0.5 # Assumed proportion for maximum variability

margin\_of\_error = 0.04 # Desired margin of error

# Calculate minimum sample size

minimum\_sample\_size = math.floor((Z \*\* 2 \* p \* (1 - p)) / margin\_of\_error \*\* 2)

print("Minimum sample size:", minimum\_sample\_size)

**Output:**

Minimum sample size: 848

1. Examine the following normal Quantile plots carefully. Which of these plots indicates that the data?
2. Are nearly normal?

Plot C appears to be the closest to normal. The points follow a relatively straight diagonal line, which indicates that the observed quantiles are consistent with the expected quantiles from a normal distribution.

1. Have a bimodal distribution? (One way to recognize a bimodal shape is a “gap” in the

spacing of adjacent data values.)

Plot B and D might be bimodal distributions. They show a clear “gap” in the middle of the plot, where the points fall away from the diagonal line in both the upper and lower tails.

1. Are skewed (i.e. not symmetric)?

Plot A appears skewed to the left. The points deviate from the line in a way that curves downwards as the quantiles increase. This suggests that the data has a longer tail on the left side.

1. Have outliers on both sides of the center?

Plot A and Plot B seem to have outliers but not both sides of the centre.



1. For each of the following statements, indicate whether it is True/False. If false, explain why.

The manager of a warehouse monitors the volume of shipments made by the delivery team. The automated tracking system tracks every package as it moves through the facility. A sample of 25 packages is selected and weighed every day. Based on current contracts with customers, the weights should have μ = 22 lbs. and σ = 5 lbs.

1. Before using a normal model for the sampling distribution of the average package weights, the manager must confirm that the weights of individual packages are normally distributed.
2. The standard error of the daily average SE(𝑥̅) = 1

i) True. Before the manager assumes that the average weight of packages behaves like a normal bell curve, they should check if the individual package weights are already close to a bell curve shape. If the weights are already shaped like a bell curve, it's easier to trust the averages. But if the weights are all over the place, it might not be safe to assume that the averages will follow a nice bell curve pattern.

ii) True. The standard error tells us how much the average weight might vary from day to day. In this case, the standard error is like a measure of how much the average weight might bounce around if we took many samples of 25 packages each day. With a standard error of 1 lb, it means the average weight might typically vary by about 1 lb [(Sigma/root of n) = (5/root of 25) = (5/5) =1] from day to day, which is useful information for the manager to know.

1. An educational startup that helps MBA aspirants write their essays is targeting individuals who have taken GMAT in 2012 and have expressed interest in applying to FT top 20 b-schools. There are 40000 such individuals with an average GMAT score of 720 and a standard deviation of 120. The scores are distributed between 650 and 790 with a very long and thin tail towards the higher end resulting in substantial skewness. Which of the following is likely to be true for randomly chosen samples of aspirants?
2. The standard deviation of the scores within any sample will be 120.

False. While the population standard deviation is 120, the standard deviation of scores within any sample will depend on the sample itself and may not necessarily be 120. It will likely be close to 120, especially for large samples, but it can vary.

1. The standard deviation of the mean of across several samples will be 120.

False. The standard deviation of the sample means across several samples, also known as the standard error of the mean, is calculated as the population standard deviation divided by the square root of the sample size. It is not the same as the population standard deviation. Therefore, the standard deviation of the mean across several samples will be less than 120.

1. The mean score in any sample will be 720.

False. The mean score in any sample will vary from sample to sample. While the population mean is 720, individual sample means may differ due to random sampling variability.

1. The average of the mean across several samples will be 720.

True. The law of large numbers states that as the number of samples increases, the average of the sample means will converge to the population mean. Therefore, the average of the mean across several samples is expected to be approximately equal to the population mean of 720.

1. The standard deviation of the mean across several samples will be 0.60

False. The standard deviation of the mean across several samples, also known as the standard error of the mean, is calculated using a formula that involves the population standard deviation and the sample size. It will depend on these factors, but it will not be 0.60. The value of 0.60 seems arbitrary and unrelated to the given data.