Sampling and Estimation

Ms Swapnil Shrivastava,
C-DAC Bangalore

Content

- What is Sampling?
- Sampling Plan
- Sampling Methods
- Estimation of Population Parameters
- Point Estimates
- Interval Estimates
- Summary

Sampling

- Sampling is a process of selecting subset of observations from a population to make inference about various population parameters such as mean, standard deviation etc.
- Sampling is necessary when it is difficult or expensive to collect data on the entire population.
- Application: Statistical Inference
  - Incorrect sample → wrong inference about the population.

Sampling Plan

- It is a description of the approach that is used to obtain samples from a population prior to any data collection activity. A sampling plan states
  - The objective of the sampling activity
  - The target population
  - The population frame (the list from which the sample is selected)
  - The method of sampling
  - The operation procedures for collecting the data
  - The statistical tools that will be used to analyze the data
- An effective sampling plan will yield representative samples of the population under study.

Example: Market Research Study

Suppose that a company wants to understand how golfers might respond to membership program that provides discounts at golf courses in the golfers locality as well as across the country.

Objective: estimate the proportion of golfers who would likely subscribe to this program.

Target population: might be all golfers over 25 years old

Population frame: list of golfers who have purchased equipment from national golf or sporting goods companies through which the discount cards will be sold.

Operational procedures: e-mail link to a survey site or direct—mail questionnaire. Data might be stored in Excel.

Statistical tools: Pivot Table and simple descriptive statistics would be used to segment the respondents into different demographic groups and estimate their likelihood of responding positively.

Method of sampling: ?

Sampling Methods

- Subjective Methods
  - Judgement Sampling: expert judgement is used to select the sample.
  - Convenience Sampling: samples are selected based on the ease at which data can be selected.
  - The accuracy of estimation may be biased.
- Probabilistic Sampling
  - Individual observations in the sample are selected according to a random procedure.
  - If N=100 and n=30 there will be more than 2.93×10^{25} possible samples.
  - Example: Simple random sampling

Probabilistic Sampling

- Simple Random Sampling
  - Every case in the population has equal probability of getting selected in a sample.
- Periodic Sampling
  - It selects every nth item from the population.
  - If a period of 5 is used, observations 5,10,15 …. will be selected as samples.
- Stratified Sampling:
  - The population can be divided into mutually exclusive groups called stratum using some factors.
  - The size of the sample in each strata should be proportional to the proportion of the strata in the population.

Probabilistic Sampling

- Cluster Sampling
  - The population is divided into mutually exclusive clusters.
  - The clusters are randomly selected and then all units within the selected clusters are included in the sample.
- Sampling from a continuous process: two ways
  - Select time at random, then select the next n items produced after that time
  - Select n times at random, then select the next item produced after each of these items.

Some Terminologies

- Population parameters or estimation: assessing the value of measures an unknown population parameter like population mean and population standard deviation using sample data.
  - μ represents population mean
  - α represents population standard deviation
- Estimators are the measures used to estimate population parameters.
  - $\bar{x}$ for sample mean
  - S or s for sample standard deviation

Estimation of Population Parameters

- Two types of Estimates
  - Point Estimate is a single value calculated from sample data that is used to estimate the value of a population parameter.
    - Sample mean $\bar{x}$ to estimate a population mean μ
  - Interval Estimate: is said to lie in an interval between a and b with certain probability.
    - We are 95% confident that the interval we obtain from sample data contain the true population.

Unbiased Estimator

- Sample mean should provide a good point estimate for the population mean.
- Repeatedly sample from a population and compute a point estimate for a population parameter.
- Each individual point estimate will vary from the population parameter.
- However it is possible that expected value of all possible point estimates would equal the population parameter.
- If the expected value of an estimator equals the population parameter it is intended to estimate, the estimator is said to be unbiased.

Errors in Point Estimation

- Point estimates do not provide any indication of the magnitude of the potential error in the estimate.
- Example: salary of college professors
- The estimators are random variables that are characterized by some distribution
- Knowing this distribution, probability theory can be used to quantify the uncertainty associated with the estimator.

Sampling Error

- Different samples from the same population have different characteristics.
  - Variations in the mean, standard deviation, frequency distribution and so on.
- Sampling error occurs because samples are only a subset of the total population.
- Sampling error is inherent in any sampling process, and although it can be minimized it cannot be totally avoided.
- Sampling error depends upon the size of the sample relative to the population.

Nonsampling Error

- Nonsampling error occur when the sample does not represent the target population adequately.
- Reasons for nonsampling error
  - Poor sample design, using convenience sampling instead of simple random sampling.
  - Choosing the wrong population frame.
  - Owing to inadequate data reliability.
- Eliminate nonsampling error and understand the nature of sampling error.

Understanding Sampling Error

Problem Statement: Estimate the mean of a population using the sample mean. Choose a population uniformly distributed between a=0 and b=10. The expected value is 5 and variance is 8.33. Using Sampling Experiment Excel file.

- Generate 25 samples each of size 10 from this population. Compute mean for each sample.
- Average of all sample means is close to true population mean 5.0
- Increase the sample size. The expected value is still close to 5. However standard deviation become smaller for increasing sample size.
- The distribution of sample mean appears to assume the shape of a normal distribution for larger sample size.

Estimating Sampling Error

Sample Size	Average of 25 Sample Means	Standard Deviation of 25 Sample Means
10	5.0108	0.816673
25	5.0779	0.451351
100	4.9173	0.301941
500	4.9754	0.078993

- Apply empirical rules to estimate the sampling error associated with one sample size.
- Empirical rule for three standard deviations
  - S=10, sample mean distribution 2.55 to 7.45
  - S=25, sample mean distribution between 3.65 to 6.35
  - S=100, sample mean distribution between 4.09 and 5.91
  - S=500, sample mean distribution between 4.76 and 5.24

Sampling Distribution

- To quantify the sampling error in estimating the mean for any unknown population.
- Sampling Distribution of the mean is the distribution formed by the means of all possible samples of a fixed size n from some population.
- Standard error of the mean is the standard deviation of the sampling distribution of the mean and is $\sigma / \sqrt{n}$. where σ is the standard deviation of the population and n is the sample size.
    - Compute for Sampling Experiment Excel file

Central Limit Theorem

Central limit theorem states that if the sample size is large enough, the sampling distribution of the mean is approximately normally distributed, regardless of the distribution of the population and that the mean of the sampling distribution will be the same as that of the population. If the population is normally distributed, then the sampling distribution of mean will also be normal for any sample size.

Central Limit Theorem

- Let S_1, S_2,…S_k be samples of size n drawn from an independent and identically distributed population with mean μ and standard deviation σ.
- Let $\overline{X}_1$, $\overline{X}_2$ …, $\overline{X}_K$ be the sample means of the samples S_1, S_2,…S_k respectively.
- The Central Limit Theorem says, the distribution of $\overline{X}_1$, $\overline{X}_2$ …, $\overline{X}_K$ follows a normal distribution with mean μ and standard deviation $\sigma / \sqrt{n}$ for large value of n.

Example: Using the Standard Error in Probability Calculations

Suppose that the size of individual customer orders (in dollars), X, from a major discount book publisher Web site is normally distributed with a mean of $36 and standard deviation of $8. The probability that the next individual who places an order at the Website will make purchase of more than $40 can be found by calculating

```
from scipy.stats import norm
print(1-norm.cdf(40, loc=36, scale=8))
```

Suppose a sample of 16 customers is chosen. What is the probability that the mean purchase for these 16 customers will exceed $40?

```
print(1-norm.cdf(40, loc=36, scale=2))
```

Interval Estimates

- It provides a range for a population characteristics based on a sample.
- Intervals specify a range of plausible values for the characteristic of interest and a way of assessing how plausible they are.
- A 100(1-α) % probability interval is any interval [A,B] such that the probability of falling between A and B is 1- α.
- Probability intervals are often centered around mean or median.

Interval Estimates in the News

Interval estimates are often constructed by taking a point estimate and adding and subtracting a margin of error that is based on the sample size.

A Gallup poll might report that 56% of voters support a certain candidate with a margin of error of +/- 3%

The poll showed a 52% level of support with a margin of error of +/- 4%

Confidence Intervals

- It is a range of values between which the value of the population parameters is believed to be, along with a probability that the interval correctly estimates true(unknown) population parameter.
- This probability is called the level of confidence denoted by $1 - \alpha$, where α is a number between 0 and 1.
  - If level of confidence is 90% then $\alpha=0.1$

Example

In a production process for filling bottles of liquid detergent, historical data have shown that the variance in the volume is constant, however clogs in the filling machine often affect the average volume. The historical standard deviation is 15ml. In filling 800ml bottles, a sample of 25 found an average volume of 796 ml.

Confidence Interval for the Mean with Known Population Standard Deviations

```
from scipy.stats import norm
Import math

m=796
n=25
alpha=0.05
std_err=15/math.sqrt(n)
zval=norm.ppf(1-alpha/2)
h1=std_err*zval

print( m - h1)
print ( m + h1)
```

T-distribution

- It is a family of probability distributions with a shape similar to the standard normal distribution.
  - Degree of freedom (df) is additional parameter.
- The t-distribution has a larger variance than the standard normal, thus making confidence intervals wider than those obtained from the standard normal distribution, in essence correcting for the uncertainty about the true standard deviation, which is not known.
- As the number of degrees of freedom increases, the t-distribution converges to the standard normal distribution.
- For large sample size people use z-values to establish confidence interval

Confidence Interval for the Mean with Unknown Population Standard Deviations

In the Excel file Credit Approval Decisions, a large bank has sample data used in making credit approval decisions.

```
from scipy.stats import t
import math
confidence = 0.95
m=12630
n=27
std_dev=5393.38
std_err=std_dev/math.sqrt(n)

h = std_err * t.ppf((1 + confidence) / 2, n - 1)

print( m - h)
print ( m + h)
```

Using Confidence Interval for Decision Making

In packaging a commodity product such as laundry detergent, the manufacturer must ensure that the packages contain the stated amount in order to meet government regulations.

The required volume is 800 millimeters.

The 95% confidence interval for the mean is [790.12, 801.88]

Is it a problem if sample average is 796?

Population Proportion(p)

- p is the proportion of cases in the data belonging to a specific category.
- Assume three categories: 1 Low, 2 Medium and 3 High
- p_1, p_2, and p_3 are proportions of population belonging to categories 1,2 and 3.
- n_1, n_2 and n_3 are the number of cases under categories 1,2 and 3.
- The estimates of proportions $\hat{p}_1$, $\hat{p}_2$ and $\hat{p}_3$ are given by

$$\hat{p}_1 = \frac{n_1}{n_1 + n_2 + n_3} \qquad \hat{p}_2 = \frac{n_2}{n_1 + n_2 + n_3} \qquad \hat{p}_3 = \frac{n_3}{n_1 + n_2 + n_3}$$

Confidence Interval for a Proportion

- For categorical variables, the proportion of observations in a sample is of interest.
- The confidence interval is the point estimate plus or minus some margin of error.

```
from scipy.stats import norm
import math

alpha=0.05
prop=0.25
samp_size=24
zval=norm.ppf(1-alpha/2)
pval=zval*math.sqrt(prop*(1-prop)/samp_size)

print(prop+pval)
print(prop-pval)
```

Prediction Interval

- It provides a range for predicting the value of a new observation from the same population.
- It is associated with the distribution of the random variable.
- The size of prediction is wider than that of the confidence interval.

Example

In estimating the revolving balance in the Excel file Credit Approval Decisions, compute a 95% prediction interval for the revolving balance of a new homeowner.

```
from scipy.stats import t
import math

m=12630.37
n=27
perr=t.ppf((1 + confidence) / 2, n - 1)*std_dev*math.sqrt(1+1/n)
print(m+perr)
print(m-perr)
```

Confidence Intervals and Sample Size

- As the sample size increases, the width of the confidence interval decreases.
- We can determine the sample sizes to achieve a given margin of error.
- The width of the confidence interval on either side of the mean (i.e. the margin of error) could be fixed to be at most E.

Sample size determination for the Mean

```
from scipy.stats import norm
import math

alpha=0.05
sd=15
n=97
m=796
err=3
zval=norm.ppf(1-alpha/2)
print(math.pow(zval,2)*math.pow(sd,2)/math.pow(err,2))
```

Sample Size Determination for a proportion

```
from scipy.stats import norm
import math

alpha=0.05
sd=15
n=97
m=796
pi=0.5
err=0.02
print(math.pow(zval,2)*pi*(1-pi)/math.pow(err,2))
```

Summary

- Sampling is a process of selecting subset of observations from a population to make inference about various population parameters.
  - Point estimates and interval estimates
- Applications
  - Provide values for inputs in decision models
  - Understand customer satisfaction
  - Reach a conclusion as to which of several sales strategies is more effective
  - Understand if a change in process resulted in an improvement.

References

- James Evans, Business Analytics: Methods, Models and Decisions, Second Edition, Pearson Publication, 2017.
- Manaranjan Pradhan and U Dinesh Kumar, Machine Learning using Python, Wiley Publication, 2019.
- U Dinesh Kumar, Business Analytics- The Science of Data-Driven Decision Making, Wiley Publication, 2017.
- Scipy- Statistical Computation www.scipy.org