Statistical Inference

Module 5

Statistical Inference

- Statistical inference is the process of deducing properties of an underlying distribution by analysis of data.
- Inferential statistical analysis infers properties about a population: this includes testing hypotheses and deriving estimates.
- The population is assumed to be larger than the observed data set; in other words, the observed data is assumed to be sampled from a larger population.
- Inferential statistics can be contrasted with descriptive statistics.
- Descriptive statistics is solely concerned with properties of the observed data, and does not assume that the data came from a larger population.

Statistical Inference

- Statistical inference makes propositions about a population, using data drawn from the population via some form of sampling.
- Given a hypothesis about a population, for which we wish to draw inferences, statistical inference consists of (firstly) selecting a statistical model of the process that generates the data and (secondly) deducing propositions from the model.

Estimation

- In most statistical research studies, population parameters are usually unknown and have to be estimated from a sample.
- Estimators = random variables used to estimate population parameters (mean, variance)
- Estimates = specific values of the population parameters

Types of estimates

- Point estimate = estimate that specifies a single value of the population
- Interval estimate = estimate that specifies a range of values

Example

A poll may seek to estimate the proportion of adult residents of a city that support a proposition to build a new sports stadium.

Out of a random sample of 200 people, 106 say they support the proposition.

Thus in the sample, 0.53 of the people supported the proposition.

This value of 0.53 is called a point estimate of the population proportion.

It is called a point estimate because the estimate consists of a single value or point.

Interval Estimate

- Point estimates are usually supplemented by interval estimates called confidence intervals.
- Confidence intervals are intervals constructed using a method that contains the population parameter a specified proportion of the time.
- For example, if the pollster used a method that contains the parameter 95% of the time it is used, he or she would arrive at the following 95% confidence interval: $0.46 < \pi < 0.60$.
- The pollster would then conclude that somewhere between 0.46 and 0.60 of the population supports the proposal.
- The media usually reports this type of result by saying that 53% favor the proposition with a margin of error of 7%.

Properties of a good estimator

Let θ = a population parameter

Let = $\hat{\theta}$ a sample estimate of that parameter. Desirable properties of $\hat{\theta}$ are:

1. <u>Unbiased:</u>

   Expected value = the true value of the parameter

   $E(\hat{\theta}) = \theta$.

   For example, $E(\bar{X}) = \mu$, $E(s^2) = \sigma^2$.

Properties of a good estimator

2. Efficiency:

   The most efficient estimator among a group of unbiased estimators is the one with the smallest variance.

   For example, both the sample mean and the sample median are unbiased estimators of the mean of a normally distributed variable. However, $\bar{X}$ *has the smallest variance.*

Properties of a good estimator

3. Sufficiency:

   An estimator is said to be sufficient if it uses all the information about the population parameter that the sample can provide.

   The sample median is not sufficient, because it only uses information about the ranking of observations. The sample mean is sufficient.

Properties of a good estimator

4. Consistency

   An estimator is said to be consistent if it yields estimates that converge in probability to the population parameter being estimated as N becomes larger.

   That is, as N tends to infinity,

   $E(\hat{\theta}) = \theta$, $V(\hat{\theta}) = 0$.

   For example, as N tends to infinity,

   $V(\overline{X}) = \sigma^2/N = 0$.

Estimating the population mean μ

- Different samples could yield different values for $\overline{X}$
- For example, if we take a random sample of 5 exam scores, we might get a mean of 94; if we take a different random sample of 5 cases, we might get a mean of 92; and so on.
- That is, $\overline{X}$ is itself a random variable, which has its own mean and variance.
- If we take all possible samples of size N and determine for each sample, the resulting distribution is the probability distribution for X.
- The probability distribution of X is called a sampling distribution.

Estimating the population mean μ

- Sample mean $\overline{X}$ is the best estimator of the population mean , and,
- its sampling distribution approximates the normal distribution as long as the sample is sufficiently large.
- Even if the population is not normal , sample means are dispersed around the parameter in a distribution that is close to normal.
- The mean of the distribution of sample means is equal to population mean.

Estimating the population mean μ

$$\sigma_{\bar{X}} = \frac{\sigma_p}{\sqrt{n}}$$

Where $\sigma_{\bar{X}}$ = standard error of mean of a given sample

σ_p = standard deviation of population

n = size of sample

$$\sigma_{\bar{X}} = \frac{\sigma_s}{\sqrt{n}} \qquad \sigma_s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$

A manufacturer takes 25 measurements of the weight of a product (Kgs) with the measurement results presented in table (sample data). Calculate an unbiased estimate of the mean, and the standard deviation and standard error of your estimate of the mean.

5.02	5.32	5.14	5.12	4.97
5.11	4.89	5.23	4.92	4.86
5.06	4.97	4.84	5.03	4.79
5.30	4.75	4.85	5.27	4.56
4.46	4.93	5.29	4.82	5.30

$s^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$ 0.05245

Interval Estimates

- Suppose we take a sample of 36 items whose mean $\bar{X}$=6.20 and standard deviation $\sigma_{\bar{s}}$= 3.8
- Population mean μ= 6.20
- Standard error of mean $\sigma_{\bar{X}}$ =3.8/sqrt(36)=0.663
- Interval estimate of μ

$$\overline{X} \pm 1.96(\sigma_{\overline{X}})$$

  4.96 to 7.44
- 95% chance that population is within this range
- To reduce this interval

1) smaller degree of confidence

2) Increase Sample size

Interval Estimates

- When population standard deviation is not known and the sample size is small, normal distribution is not appropriate.
- Use t-distribution
- There is a different t-distribution for each of the possible degrees of freedom.
- Degrees of freedom = n – 1
- Look for critical value of ‘t’ in the t-distribution table for appropriate degrees of freedom at a given level of significance.

Example

In a random selection of 64 of the 2400 intersections in a small city, the mean number of scooter accidents per year was 3.2 and the sample standard deviation was 0.8.

(1) Make an estimate of the standard deviation of the population from the sample standard deviation.

(2) Work out the standard error of mean for this finite population.

(3) If the desired confidence level is .90, what will be the upper and lower limits of the confidence interval for the mean number of accidents per intersection per year?

$$\hat{\sigma}_p = \sigma_s = 0.8$$

N=2400

n=64

$$\sigma_{\bar{X}} = \frac{\sigma_s}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}} \quad =0.097$$

$$\overline{X} = 3.2$$

$$\sigma_s = 0.8$$

$$\overline{X} + z\left\{\frac{\sigma_s}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}}\right\} \quad =3.2+\_1.6*0.097=3.3$$

finite population multiplier or finite correction factor.

Estimating the population proportion

- Sample proportion (p) of units that have a particular characteristic is the best estimator of the population proportion $\hat{p}$
- Its sampling distribution, so long as the sample is sufficiently large, approximates the normal distribution.

Estimating the population proportion

- *If we take a random sample of 50 items and find that 10 per cent of these are defective*

  $p = .10$

- *we can use this sample proportion (p = .10) as best estimator of the population proportion*
- $\hat{p} = p = .10$
- *To construct confidence interval to estimate a population proportion, we use the binomial distribution*
- *mean of population* $\mu = n \times p$ *,*
  - *n = number of trials*
  - *p = probability of a success in any of the trials population*
- *standard deviation = sqrt(npq)*

Estimating the population proportion

As the sample size increases

- The mean of the sampling distribution of the proportion of successes

  $\mu_p = p$

- *standard deviation for the proportion of* successes, also known as the standard error of proportion

  sqrt(pq/n)

Estimating the population proportion

- When population proportion is unknown, then we can estimate the population parameters by substituting the corresponding sample statistics p and q

$$\sigma_p = \sqrt{\frac{pq}{n}}$$

- Confidence Interval

$$p \pm z.\sqrt{\frac{pq}{n}}$$

Example

A market research survey in which 64 consumers were contacted states that 64 per cent of all consumers of a certain product were motivated by the product's advertising. Find the confidence limits for the proportion of consumers motivated by advertising in the population, given a confidence level equal to 0.95.

Example

n = 64

p = 64% or .64

q = 1 – p = 1 – .64 = .36

Z= 1.96

$$p \pm z.\sqrt{\frac{pq}{n}} \qquad 0.64 \pm 1.96.\sqrt{\frac{0.64*0.36}{64}}$$

- lower confidence limit is 52.24%
- upper confidence limit is 75.76%

	Infinite population	Finite population
Estimating population mean μ when we know σ_p	$\bar{X} \pm z.\left\{\frac{\sigma_p}{\sqrt{n}}\right\}$	$\bar{X} + z\left\{\frac{\sigma_p}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}}\right\}$
Estimating population mean μ when we know σ_p and use σ_s as the best estimate of σ_p and sample is large (*n > 30)*	$\bar{X} \pm z.\left\{\frac{\sigma_s}{\sqrt{n}}\right\}$	$\bar{X} + z\left\{\frac{\sigma_s}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}}\right\}$
Estimating population mean μ when we know σp and use σs as the best estimate of σp and sample is small (n <= 30)	$\bar{X} \pm t.\left\{\frac{\sigma_s}{\sqrt{n}}\right\}$	$\bar{X} + t.\left\{\frac{\sigma_s}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}}\right\}$
Estimating the population proportion p when *p is not* known but the sample is large	$p \pm z.\sqrt{\frac{pq}{n}}$	$p \pm z.\sqrt{\frac{pq}{n}} * \sqrt{\frac{N-n}{n-1}}$

Estimation of Population Mean and Population Variance

- One of the main objectives after the selection of a sample is to know about the tendency of the data to cluster around the central value and the scatterdness of the data around the central value in the population.
- Popular choices are arithmetic mean and variance.
- Population mean is generally measured by arithmetic mean (or weighted arithmetic mean)

Estimation of Population Mean and Variance: Notations

$Y_1, Y_2, \ldots, Y_N$: **Population**

$y_1, y_2, \ldots, y_n$: **Sample**

$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$: **Population mean**

$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$: **Sample mean**

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} Y_i^2 - N\overline{Y}^2\right)$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \overline{Y})^2 = \frac{1}{N}\left(\sum_{i=1}^{N} Y_i^2 - N\overline{Y}^2\right)$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right)$$

Estimation of Population Mean

Let us consider the sample arithmetic mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ as an estimator of population mean $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$

Estimate population mean $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ by sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

$\bar{y}$ is an unbiased estimator of $\bar{Y}$ under SRSWR and SRSWOR cases.

$$E(\bar{y}) = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{Y}.$$

Estimation of Population Mean

Population: $X1 = 1$, $X2 = 3$, $X3 = 5$

Population mean = 3

Number of Samples of size 2= ${}^{3}C_{2}$ n!/(n-r)!*r!

=3

Suppose the population mean is unknown.

Use sample arithmetic mean to estimate the population mean.

Estimation of Population Mean

Sample arithmetic mean is an unbiased estimator of population mean.

Sample 1=(1,3) Sample mean $\overline{x1} = 2$

Sample 2=(3,5) Sample mean $\overline{x2} = 4$

Sample 3=(1,5) Sample mean $\overline{x3} = 3$

$$\overline{x} = \frac{\overline{x}_1 + \overline{x}_2 + \overline{x}_3}{3} = \frac{2+4+3}{3} = 3 = \textbf{Population mean}$$

- $\bar{y}$ is an unbiased estimator of $\bar{Y}$ in SRSWOR

$$E(\bar{y}) = \frac{1}{n}\sum_{j=1}^{n} E(y_j)$$

- $\bar{y}$ is an unbiased estimator of $\bar{Y}$ in SRSWR

$$E(\bar{y}) = \frac{1}{n} E\left(\sum_{i=1}^{n} y_i\right)$$

Sample Mean Example

Y: Height of students in a class
N = 10 : Number of students in the class (Population size)
n = 3 : Number of students in the sample (Sample size)

Name of Student Y_i = Height of students (in Centimeters)

A	Y_1= 151
B	Y_2= 152
C	Y_3 = 153
D	Y_4= 154
E	Y_5 = 155
F	Y_6= 156
G	Y_7 = 157
H	Y_8= 158
I	Y_9 = 159
J	Y_{10}= 160

Sample 1: 3rd , 7th and 9th student
$y_1 = Y_3$ = 153 cms., $y_2 = Y_7$ = 157 cms., $y_3 = Y_9$ = 159 cms
Sample mean 1($\bar{y}_1$) = (153 + 157 + 159)/3= 156.33 cms

Sample 2: 2nd , 5th and 4th student
$y_1 = Y_2$ = 152 cms., $y_2 = Y_5$ = 155 cms., $y_3 = Y_4$ = 154 cms
Sample mean 2 ($\bar{y}_2$) = (152 + 155 + 154)/3= 153.66 cms

Sample 3: 1st , 6th and 10th student
$y_1 = Y_1$ = 151 cms., $y_2 = Y_6$ = 156 cms., $y_3 = Y_{10}$ = 160 cms
Sample mean 3 ($\bar{y}_3$) = (151 + 156 + 160)/3= 155.66 cms

Population mean = $\bar{Y}$ 155.5

The total number of samples = $10C_3$=120

Variance of Sample Mean

- Population variability is generally measured by variance.
- Several sample can be drawn by SRSWR as well as SRSWOR from a population.
- Each sample will have different sample mean.
- Sample mean is a statistic, i.e., a function of random variables.
- So sample mean will also have variance.

Variance of Sample Mean

Variance of sample mean under SRSWOR

$$V(\bar{y}_{WOR}) = E(\bar{y} - \bar{Y})^2 = \frac{N-n}{Nn} S^2$$

Variance of sample mean under SRSWR

$$V(\bar{y}_{WR}) = E(\bar{y} - \bar{Y})^2 = \frac{N-1}{Nn} S^2$$

Example

For the following population, consider all SRSWOR samples of size 3

i	1	2	3	4	5
y_i	5	8	3	11	9

1. Show that $\bar{y}$ is unbiased estimator of $\bar{Y}$
2. Show that s^2 is unbiased estimator of S^2
3. Calculate sampling variation $\bar{y}$ and show that it agrees with the formula $(N-n/nN)\ S^2$
4. Verify that $Var_{srswr}(\bar{y}) > Var_{srswor}(\bar{y})$

N=5

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

(Population Mean)

$\bar{Y}$=(5+8+3+11+9)=7.2

(Population Mean Square)

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2$$

Y_i	1	2	3	4	5	
$(y_i-\bar{Y})^2$	$(5-7.2)^2$	$(8-7.2)^2$	$(3-7.2)^2$	$(11-7.2)^2$	$(9-7.2)^2$	40.80

S^2=1/4*40.80=10.2

From a population of 5 a sample of size 3 can be drawn in $5C_3$ ways= 10 ways

No	Values	Mean	S^2	$(y_i-\bar{Y})$	$(y_i-\bar{Y})^2$
1	1,2,3	16/3	19/3	-1.87	3.48
2					
3					
4					
5					
6					
7					
8					
9					
10					

mean=(5+8+3)=16/3

Y_i	1	2	3			
$(y_i-\bar{Y})^2$	$(5-16/3)^2$	$(8-16/3)^2$	$(3-16/3)^2$			12.66

$S^2=1/2*12.66=6.33=19/3$

$(16/3-7.2)=-1.87$

No	Values	Mean	S^2	$(y_i-\bar{Y})$	$(y_i-\bar{Y})^2$
1	1,2,3	16/3	19/3	-1.87	3.48
2					
3					
4					
5					
6					
7					
8					
9					
10					
		216/3	306/3		13.60

$E(\bar{y})= \frac{1}{NC_n} \sum_{i=1}^{N} \bar{y}_i$ =1/10*216/3=7.2=$\bar{Y}$

$E(s^2)= \frac{1}{NC_n} \sum_{i=1}^{N} s_i^2$ =1/10*306/3=10.2= S^2

In SRSWOR

Var($\bar{y}$)= $\frac{1}{NC_n}\sum_{i=1}^{N}(y_i-\bar{Y})^2$ =1/10*13.6=1.36

Var($\bar{y}$)= $\frac{N-n}{Nn}S^2$ =(5-3/5*3) *10.2=(2/15)*10.2=1.36

In SRSWR

Var($\bar{y}$)= $\frac{N-1}{Nn}S^2$ =(5-1/5*3) *10.2=(4/15)*10.2=2.72

$Var_{srswr}(\bar{y}) > Var_{srswor}(\bar{y})$

- Explain why a random sample of size 25 is to be preferred to a random sample of size 20 to estimate population mean.

$\text{var}(\bar{y})=\sigma^2/n$

$\text{S.E.}(\bar{y})=\sigma/\text{sqrt}(n)$

Larger the sample smaller is the error

Sample Size And Its Determination

- What should be the size of the sample or how large or small should be *'n'?*
- *n too small - may not serve to achieve the objectives*
- n too large- huge cost and waste resources
- The sample must be of an optimum size
- Technically, the sample size should be large enough to give a confidence interval of desired width and as such the size of the sample must be chosen by some logical process before sample is taken from the universe.

Sample size

1. Nature of universe:

   Homogenous - a small sample can serve the purpose

   Heterogeneous - a large sample

2. Number of classes proposed:

   many class-groups - large sample

3. Nature of study:

   intensive and continuous study- small sample

   general survey- large

   technical surveys - small sample

4. Type of sampling: A small random sample is apt to be much superior to a larger but badly selected sample.

Sample size

5. Standard of accuracy and acceptable confidence level:

   High - we shall require relatively larger sample.

   For doubling the accuracy for a fixed significance level, the sample size has to be increased fourfold.

6. Availability of finance:

   Large samples result in increasing the cost of sampling estimates.

7. Other considerations:

   Nature of units, size of the population, size of questionnaire, availability of trained investigators, the conditions under which the sample is being conducted, the time available for completion of the study

Sampling Approaches

- Two alternative approaches for determining the size of the sample.

1. To specify the precision of estimation desired and then to determine the sample size necessary to insure it
   - Mathematical solution
   - Frequently Used
   - Does not analyse the cost of gathering information
2. Use Bayesian statistics to weigh the cost of additional information against the expected value of the additional information
   - Theoretically optimal
   - seldom used because of the difficulty involved in measuring the value of information.

Precision based Sample Size

- Whenever a sample study is made, there arises some sampling error which can be controlled by selecting a sample of adequate size.
- Researcher will have to specify the precision that he wants in respect of his estimates concerning the population parameters.
- For instance, a researcher may like to estimate the mean of the universe within ± 3 of the true mean with 95 per cent confidence.
- Desired precision is ±3
- If the sample mean is Rs 100, the true value of the mean will be no less than Rs 97 and no more than Rs 103.
- Acceptable error, e =is equal to 3.

Sample size for estimating a mean

- Suppose, $\sigma_p = 48$
- If the difference between μ and $\overline{X}$ or the acceptable error is to be kept with in ±3 of the sample mean with 95% confidence

$$\overline{X} \pm z.\left\{\frac{\sigma_p}{\sqrt{n}}\right\} \qquad e = z.\left\{\frac{\sigma_p}{\sqrt{n}}\right\}$$

$$3 = 1.96.\left\{\frac{4.8}{\sqrt{n}}\right\}$$

$$n = \frac{(1.96)^2 * (4.8)^2}{3^2}$$

$$= 9.834$$

$$\cong 10$$

Sample Size

In general,

- Infinite Population

$$n = \frac{z^2 \sigma^2}{e^2}$$

- Finite Population

$$n = \frac{z^2 . N . {\sigma_p}^2}{(N-1)e^2 + z^2 . {\sigma_p}^2}$$

Example

- Determine the size of the sample for estimating the true weight of the cereal containers for the universe with *N = 5000 on the basis of the following information:*

  (1) the variance of weight = 4 ounces on the basis of past records.

  (2) estimate should be within 0.8 ounces of the true average weight with 99% probability.

  Will there be a change in the size of the sample if we assume infinite population in the given case? If so, explain by how much?

Example

N = 5000

σ_p = 2 ounces (variance of weight = 4 ounces)

e = 0.8 ounces

z = 2.57

$$n = \frac{z^2.N.\sigma_p^{\,2}}{(N-1)e^2 + z^2.\sigma_p^{\,2}}$$

= 41

$$n = \frac{z^2\sigma^2}{e^2}$$

= 41

A manufacturing concern wants to estimate the average amount of purchase of its product in a month by the customers. If the standard deviation is Rs. 10, find the sample size if the maximum error is not to exceed Rs 3.00 with a probability of 0.99.

Given $\sigma = 10$, $E = 3$, confidence level $= 99\%$ or $\alpha = 0.01$

We have to find n the sample size

Sample size

- If the standard deviation of the population is not available, and if we have an idea about the range of the population, we can use that to get a crude estimate of the standard deviation of the population for getting a working idea of the required sample size.
- Since 99.7 per cent of the area under normal curve lies within the range of ± 3 standard deviations, we may say that these limits include almost all of the distribution.
- Accordingly, we can say that the given range equals 6 standard deviations because of ±3.

$$6\hat{\sigma} = given\_range$$

$$\hat{\sigma} = \frac{given\_range}{6}$$

- If the range happens to be, say Rs 12, then $\hat{\sigma} = 2$

Sample size when estimating a percentage or proportion

$$p \pm z.\sqrt{\frac{pq}{n}}$$

$$e = z.\sqrt{\frac{pq}{n}}$$

Infinite Population

$$n = \frac{z^2.p.q}{e^2}$$

Finite Population

$$n = \frac{z^2.p.q.N}{e^2(N-1) + z^2.p.q}$$

Example

- Suppose a certain hotel management is interested in determining the percentage of the hotel's guests who stay for more than 3 days. The reservation manager wants to be 95 per cent confident that the percentage has been estimated to be within ± 3% of the true value. What is the most conservative sample size needed for this problem?

Example

e = .03

z = 1.96

most conservative sample size

p = .5 and q = .5.

$$n = \frac{z^2 . p . q}{e^2}$$

= 1067

Sample Size Through Bayesian Statistics

(i) Find the expected value of the sample information EVSI for every possible *n;*

(ii) Also workout reasonably approximated cost of taking a sample of every possible *n;*

(iii) Compare the EVSI and the cost of the sample for every possible *n.* Workout the expected net gain ENG for every possible *n as*

EVSI – Cost of sample = ENG

(iv) From (iii) above the optimal sample size, that value of *n which maximises the difference* between the EVSI and the cost of the sample, can be determined.