MDA102 – Statistical methods using R
Inferential Statistics Units 4,5

Statistical inference

- Definition

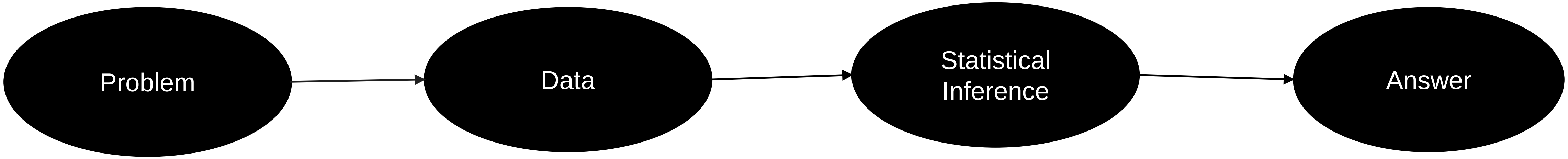
statistical inference

noun

the theory, methods, and practice of forming judgements about the parameters of a population and the reliability of statistical relationships, typically on the basis of random sampling.
"the problem is fundamental to statistical inference"

Data and Statistical Inference

Statistical inference is a branch of statistics which deals with drawing conclusions about population characteristics (Parameter) from scientifically collected data.



Population and parameter

- Population is a collection of all objects under study.
- All units share some common characteristic.

- A parameter defines characteristic of population. (say average, median, etc.)
- Parameter can be considered as function of population values.

Statistical Inference in application

Example 1: who’s going to win the election?

- In every major election, pollsters would like to know, ahead of the actual election, who’s going to win.
- percentage of people in a particular group (city, state, county, country or other electoral grouping) who will vote for a particular candidate
- collect a reasonable subset of population
- Count the number of people in the subset who will vote for a particular candidate
- to produce a good guess of actual percentage of people who will vote for a particular candidate.

Statistical Inference in application- Continued...

Example 2: Has pollution control policy impacted on number of asthma patients?

- Some random cities in the country will be selected
- Number asthma patients in these cities will be noted (percentage) before and after policy implementation
- The question is to check whether new pollution control policy has reduced number of asthma patients

Statistical Inference - Classification

Parametric
Sample is drawn from a particular probability distribution

Non-parametric
No such assumption

- Estimation theory: Point estimation and Interval estimation (Example 1)
- Testing of hypothesis (Example 2)

Statistical Inference - continued

- Estimation theory

Suggest a value or an interval for the parameter based on sample observations

- Testing of hypothesis

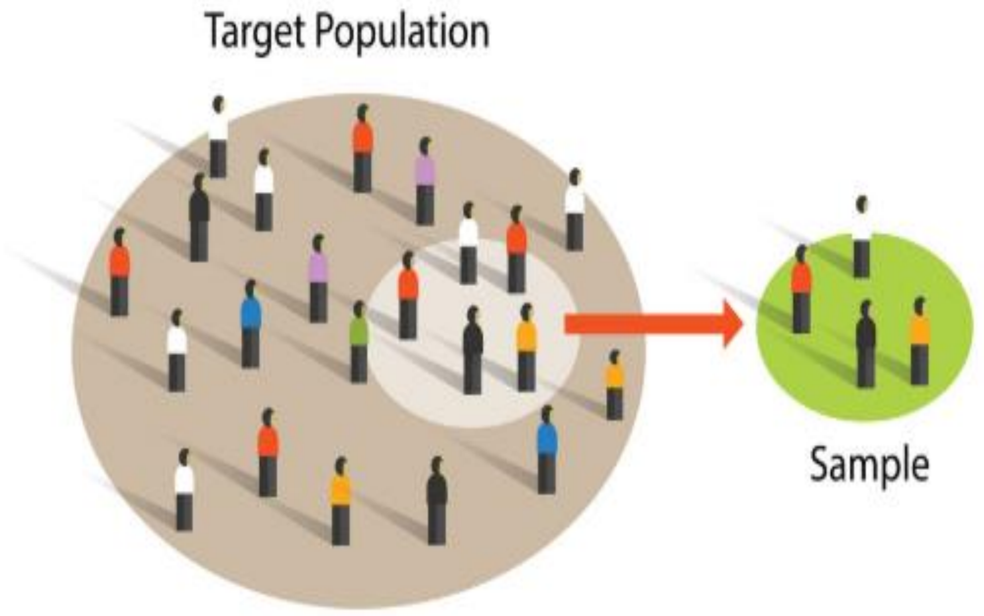
Test a statement about parameter based on sample observations

Statistical Inference - continued

Sample

- Sample is a representative subset of population.
- In mathematical notation, population can be considered as universal set and sample as the subset of the universal set.

e.g. blood test, group of voters in example 1, cities in example 2



Statistical Inference - continued

Independently and identically distributed (iid) Sample

- All observations are independently drawn and follow same probability distribution.
- To ensure iid sample, we make use of simple random sampling with replacement.

Statistical Inference - continued

Statistic

- Function of sample values is known as statistic.
- It is a quantity calculated from the sample.

e.g.

Sample mean, sample proportion, etc.

Statistical Inference - continued

Example 1
Population – Voters in the country
Parameter – proportion (percentage) of voters voting for a particular candidate
Sample- subset of voters from whom the response is collected
Statistic – sample proportion of voters who will vote for a particular candidate
Nature of inference problem – Estimation

Example 2
Population – entire population in the country
Parameter – proportion (percentage) of asthma patients in the country before and after asthma patients
Sample - subset of population (people from some districts or states)
Statistic – sample proportions of asthma patients before and after policy implementation
Nature of inference problem – Testing of hypothesis

Estimation theory

- Point Estimation

Find a value for parameter based on statistic known as estimator
Usually sample mean is used as the estimator for population mean

- Confidence interval

Find a an interval such that $P(U_1 \leq \theta \leq U_2) = 1 - \alpha$

Testing of hypothesis

- To test pre defined statement about parameters and these statements are known as statistical hypothesis
- Mainly we want to test a conjecture. For example, a heart disease is more prevalent in men than in women.
- The hypothesis is that our conjecture is false is called as the null hypothesis denoted by H_0.
- It is the hypothesis of equality or no difference (null).
- The hypothesis under which our conjecture is true is known as alternative hypothesis denoted by $H_{1.}$
- A test procedure is conducted to determine whether we should reject or accept null hypothesis based on sample.
- A test statistic from a sample is computed and based on the value of that test statistics we may reject or do not reject $H_{0.}$

Errors in testing of hypothesis

- We have four scenarios when we take a decision

- Reject H_0 when H_0 is actually true
- Do not reject H_0 when H_0 is actually true
- Reject H_0 when H_0 is not true
- Do not reject H_0 when H_0 is not true

Here 2 and 3 are correct decisions and 1 and 4 are wrong decisions. 1 is known as type 1 error, 4 is type 2 error.

Level of significance and power of the test

- Since type 1 error is more serious, we will fix probability of type 1 error before testing a hypothesis
- Maximum value of probability of type 1 error of a test procedure is known as level of significance represented by α. Usually α is fixed at 5% or at 0.05. 1- α is known as confidence level.
- 1-Probability of type 2 error is known as the power of the test, β. It is the power to reject a wrong hypothesis
- So our intention is to fix α at 0.05 or 0.01 and maximize β, when conducting a test
- P-Value is the smallest level of significance at which we should reject null hypothesis for the data we observe. That is we reject null hypothesis if P-values is less than α.

Testing of hypothesis - algorithm

- Data collection
- Choose a conjecture
- Determine the null hypothesis
- Choose a test
- Compute the test statistic from the data
- Compute the P- value and compare it with α and reject the null hypothesis if P-value is less than α.

Inference for single normal population mean μ

- We want to estimate or test hypothesis on average of height ,marks, production, income, etc. is equal to a specified value.
- Let $X_1, X_2, \dots, X_n$ are iid random sample from a normal population with mean μ and σ^2. We would like to test $H_0: \mu = \mu_0$. Against $H_1: \mu \neq \mu_0$ or $H_1: \mu < \mu_0$ $H_1: \mu > \mu_0$
- We have two cases

Case1: when σ^2 is known (z-test)

$$Z = (\bar{X} - \mu_0)/(\frac{\sigma}{\sqrt{n}})$$

Z follows standard normal distribution
Case 2: when σ^2 is unknown (t-test)

$$T = (\bar{X} - \mu_0)/(\frac{s}{\sqrt{n}})$$

Where T follows t distribution with n-1 degrees of freedom and s is the sample variance

Doing in R

Suppose we have a sample from a normal population with known variance and at level 0.05, we want to test whether the population mean is equal to 0. i.e., $H_0 : \mu = 0(\mu_0)$ VS $H_1 : \mu \neq 0$

```
X<-rnorm(37,mean=2,sd=3)#generates normal random sample from N(2,3)
library(TeachingDemos)#for z.test()
z.test(X,mu=0,sd=3,conf.level = 0.95)#conf.level=1-level of significance,mu is mu_0 here
```

```
One Sample z-test

data:  X
z = 2.7916, n = 37.0000, Std. Dev. = 3.0000, Std. Dev.
of the sample
mean = 0.4932, p-value = 0.005244
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.4101818 2.3434784
sample estimates:
mean of X
  1.37683
```

Since p-value is less than 0.05, level of significance, we reject H_0P