# Statistical Inference<br>What? Why? and How?

by

Prof. Sarath Banneheka

Department of Statistics

University of Sri Jayewardenepura

June 09, 2023

## 1 What is Inference?

Inference (general meaning):

- a conclusion reached on the basis of evidence and reasoning.
- a guess that you make or an opinion that you form based on the information that you have
- the process of inferring something

## 2 What is Statistical Inference?

- Statistical inference is a **guess or a conclusion about a population**, reached **based on a sample** from that population **using statistical reasoning so that a measure of accuracy is available**.
- The process of making statistical inference.

Mainly, there are three types of statistical inference problems

1. Point estimation of parameters
2. Interval Estimation for parameters
3. Tests of hypotheses concerning parameters

We use the term ‘parameter’ to mean a constant, whose value is unknown. Usually, parameters are population characteristics or unknown constants in various models. For example,

- Population mean ($\mu$)
- Population proportion ($\theta$)
- Regression coefficients in a regression model like $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

- Parameters in a time series model like

ARMA(p,q) model:

$$y_t = c + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q}$$

- Factor loading in a Factor analysis model like

$$\begin{aligned} X_1 - \mu_1 &= \ell_{11}F_1 + \ell_{12}F_2 + \cdots + \ell_{1m}F_m + \varepsilon_1 \\ X_2 - \mu_2 &= \ell_{21}F_1 + \ell_{22}F_2 + \cdots + \ell_{2m}F_m + \varepsilon_2 \\ &\vdots \\ X_p - \mu_p &= \ell_{p1}F_1 + \ell_{p2}F_2 + \cdots + \ell_{pm}F_m + \varepsilon_p \end{aligned}$$

It is very important to understand that

- when we estimate parameters based on samples, almost always there are errors in our estimates (error=difference between the estimate and the parameter)
- We never know the actual size of the error of our estimate.
- When we test a hypothesis concerning parameter(s) based on a sample, there is a possibility of making an erroneous conclusion.

| | Decision | |
|---|---|---|
| Reality | Reject $H_0$ | Do not reject $H_0$ |
| $H_0$ is correct | Type I error | Correct decision |
| $H_0$ is wrong | Correct decision | Type II error |

- We never know whether we have made an erroneous conclusion or not.

# 3 Why Statistical Inference is Required?

1. to obtain an estimate/predication/forecast/conclusion with a measure of accuracy.

   Eg:

   - Standard error of the estimator
   - Given bound ($b$) on the error with known level of confidence.
     i.e., Pr(|Estimator − parameter| $< b) = 0.95$
   - Confidence interval satisfying Pr($T_1 <$ parameter $< T_2) = 0.95$

2. to make conclusions so that the probability of type I error is less than a pre-specified (given) level of significance ($\alpha$) and a given power.

   Eg:

   - Pr(Type I error) $\leq 0.05$

# 4 How to Make Statistical Inference?

- Obtain the data from a **random sample** of units (Simple random sampling, stratified random sampling, cluster sampling, systematic sampling, multistage random sampling) or from a well designed experiment.
- Formulate the problem (identify the purpose)
- Find out statistical techniques/methods/procedures suitable for your purpose
- Different techniques for different purposes
  - Regression analysis (simple linear, multiple linear, generalized linear, non-linear etc)
  - ANOVA (Analysis of variace)
  - Time series Analysis
  - Multivariate methods (cluster analysis, factor analysis, PCA, discriminant analysis etc.)
- Different types of methods
  - Parametric methods
  - Non-parametric methods
  - Bayesian methods
- Find out the underline assumptions of those techniques/methods/procedures
- Explore which of those assumptions are reasonable for your data.
- Select a best technique/method/procedure accordingly.

# 5 Some Examples

## 5.1 Mean SBP

- A study is to be conducted about the systolic blood pressure (SBP) of people with glaucoma, in a certain age group. From a previous study, it is known that the mean SBP of normal people in the same age group is 130 mm Hg.
  1. What is the required sample size?
  2. What is the mean SBP of people with glaucoma?
  3. Is there evidence that the mean SBP of people with glaucoma is higher than the mean SBP of normal people in the same age group?
- Let $\mu$ be the mean SBP of the people with glaucoma in the population. If the sample is selected using the simple random sampling method, the sample size required to estimate $\mu$ with bound $b$ on the error and 95% level of confidence is given by
  $n \approx \frac{Ns^2}{\frac{(N-1)b^2}{4}+s^2}$,
  where $s^2$ is the sample variance of SBP obtained from a pilot survey.

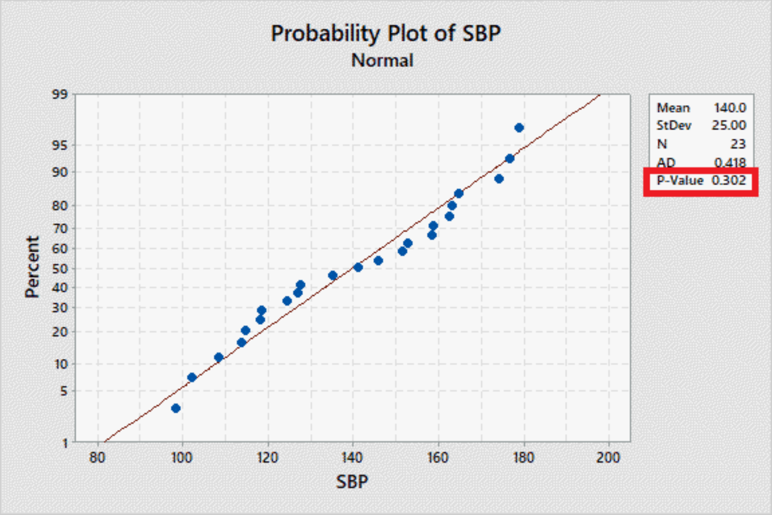
- In other words, the above $n$ is the required sample size to satisfy
  $\Pr(|\text{Estimator} - \mu| < b) = 0.95$

- For example, if $N = 1000$, $s^2 = 144$, and $b = 5$, then, $n = 22.54 \approx 23$.

The study includes 23 people with glaucoma. Suppose that the sample mean SBP is 140 mm Hg with a standard deviation of 25 mm Hg.

- Estimate for $\mu = \bar{x} = 140$.

- Test of normality of SBP
  $H_0$: SBP is normally distributed vs. $H_1$: SBP is not normally distributed.



P-value of the Anderson-Darling normality test= 0.302.

Decision rule: Reject $H_0$ if p-value $\leq$ specified level of significance.

At 0.05 level of significance, we conclude that SBP is normally distributed.

- 95% confidence interval for $\mu$ is (129.19, 150.8)

- $H_0 : \mu \leq 130$ vs $H_1 : \mu > 130$.

```
One-Sample T

Test of   = 130 vs > 130


 N    Mean  StDev  SE Mean  95% Lower Bound     T      P
23  140.00  25.00     5.21           131.05  1.92  0.034
```

At 0.05 level of significance, we conclude that the mean SBP of people with glaucoma is higher than the mean SBP of normal people in the same age group.

## 5.2 Proportion of pre-diabetic students

- The medical center of the USJ wants to find out the proportion of currently registered students who are pre-diabetic (Prediabetes is a serious health condition where blood sugar levels are higher than normal, but not high enough yet to be diagnosed as type 2 diabetes)

  1. What is the population of interest?
  2. What is the quantity (parameter) of interest?
  3. How to select a sample?
  4. How accurate the answer should be (bound on the error)?
  5. How confident you want to be about the answer?
  6. What is the required sample size?

- Let $\theta$ be the the proportion of currently registered students who are pre-diabetic. If simple random sampling is used, the sample size required to estimate $\theta$ with bound $b$ on the error and 95% level of confidence is given by

  $n \approx \frac{Np(1-p)}{\frac{(N-1)b^2}{4}+p(1-p)}$,

  where $p$ is the sample proportion of registered students who are pre-diabetic in a pilot sample. If such a value is not available, replace $p$ by 0.5 to get a conservative value for $n$.

- In other words, the above $n$ is the required sample size to satisfy

  $\Pr(|\text{Estimator} - \theta| < b) = 0.95$

- For example, if $N = 10000$, $p = 0.05$, and $b = 0.01$, then, $n = 1596.77 \approx 1597$.

## 5.3 Statistical process control

- A production process of bulbs is considered as 'in control' if the the proportion of defective bulbs produced by the process is not more than 5%.

1. Was the process 'in control' or 'out of control' during the last hour?

   Let $\theta$ be the proportion of defective bulbs produced by the process in the last hour.

   We want to test $H_0 : \theta \leq 0.05$ vs. $H_1 : \theta > 0.05$.

2. What is the required sample size to satisfy that

   - the probability of deciding that process is out of control when it is actually in control should be less than 0.05, and
   - the probability of deciding that process is out of control if the actual proportion of defective bulbs produced by the process is 0.1 should be at least 0.75?

## 5.4 Comparison of mean cholesterol levels

- A dietitian hopes to reduce a person's cholesterol level by using a special diet supplemented with a combination of vitamin pills. Five (5) subjects were pre-tested and then placed on diet for two weeks. Their cholesterol levels were checked after the two week period. The results are shown in the following table. Cholesterol levels are measured in milligrams per deciliter.

| Subject | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Before ($x$) | 192 | 210 | 206 | 200 | 192 |
| After ($y$) | 172 | 215 | 198 | 195 | 193 |

Table 5.1: Cholesterol level

The dietitian asks the following questions:

1. What is the difference of Cholesterol levels (Before-After)?
2. Does the special diet supplemented with a combination of vitamin pills reduce the cholesterol level?

The statistician may suggest to modify the questions as follows:

1. What is the average difference of Cholesterol levels (Before-After)?
2. Does the special diet supplemented with a combination of vitamin pills reduce the cholesterol level on average?

- Derive a point estimate and a 95% confidence interval for the mean or median difference of Cholesterol levels
- Test the hypothesis

  $H_0 : \mu_d \leq 0$ vs $H_1 : \mu_d > 0$

  ($\mu_d$= mean difference of Cholesterol levels)

  or

  $H_0 : \nu_d \leq 0$ vs $H_1 : \nu_d > 0$

  ($\nu_d$= medain difference of Cholesterol levels)
- Point estimate for the mean difference of cholesterol levels= 5.4
- Test of normality of differences:

  Shapiro-Wilk normality test

  data: d W = 0.9561, p-value = 0.781
- At 0.05 level of significance, we conclude that the differences between cholesterol levels are normally distributed.
-

  ```
  95% CI for mean difference: (-6.53, 17.33)
  ```
-

  ```
  T-Test of mean difference = 0 (vs > 0):

  T-Value = 1.26  P-Value = 0.139
  ```
- There is no evidence to conclude that the special diet supplemented with a combination of vitamin pills reduces the cholesterol level on average.