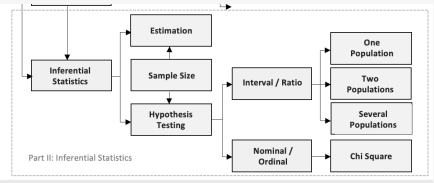
Inferential Statistics Pathway



Inferential Statistics

Concept:

Making generalization about population parameter based on the sample statistic

Examples

Is there a difference in participation in local decision-makings between low-income and middle-income groups in New Jersey?

What are reactions of New Brunswick residents to new investments of Rutgers University in development of College Avenue?

What is the level of effectiveness of the Corona vaccine developed by Johnson and Johnson pharmaceutical company?

Primary Date: Sampling Methods

1) Probability Sampling

Concept:

Sample is selected randomly (Equal Probability of Selection Method – EPSEM).

Methods:

Simple random / Systematic / Cluster / Stratified / Convenience

Application

Diverse population / Generalization is required

Primary Date: Sampling Methods (cont)

2) Non-Probability Sampling

Concept:

Sample selection is based on the subjective judgment of researcher.

Methods:

Judgement / Snowball / Quota / Consecutive

Application

Homogenous population / Pilot study

Probability Sampling Methods

Simple Sampling

Select a simple random sample through drawing or use of random number methods

Systematic Sampling

1) Number all members of population sequentially. 2) From a starting point select every nth individual.

Cluster Sampling

1) Divide population into non-overlapping clusters. 2) Sample all in some clusters (Ex. Sample all nurses in 5 hospitals in New Jersey).

Stratified Sampling

1) Divide population into non-overlapping clusters. 2) Sample some in clusters (Ex. Sample some nurses in every hospitals in New Jersey).

Convenience Sampling

Create a sample by using data from population members that are readily available.

Non-Probability Sampling Methods

Judgement Sampling

Select samples based on researcher's knowledge and if they fit to participate in the research - some subjects are more fit for the research compared to other individuals.

Snowball Sampling

Once the researcher finds suitable subjects, he asks them for assistance to seek similar subjects to form a considerably good size sample - good or small population.

Quota Sampling

Select equal or proportionate subjects depending on basis of the quota, which usually are age, gender, education, race, religion and socioeconomic status. (Ex. Sample size of 100, researcher can select 25 1st year students, 25 2nd year, 25 3rd year and 25 4th year students).

Consecutive Sampling

Is very similar to convenience sampling except that it seeks to include ALL accessible subjects as part of the sample.

Data Biases

Moderator / Interviewer bias

The moderator's facial expressions, body language, tone, manner of dress, and style of language may introduce bias. Similarly, the moderator's age, social status, race, and gender can produce bias.

Data Biases (cont)

Biased Questions

A biased question influences respondents' answers. And the way you ask a question, or "vague wording" can bias a question.

Biased Answers

A biased answer is an untrue or partially true statement, like 1) Nonresponse in survey, when individuals can't or refuse to respond, 2) Truthfulness of response, 3) Faulty recall or not remember accurately, 4) Voluntary response: individuals with strong feelings about a subject are more likely than other to respond

Biased Samples

Poor screening and recruiting causes biased samples. Examples are biases of time and location

Biased Reporting

Experiences, beliefs, feelings, wishes, attitudes, culture, views, state of mind, reference, error, and personality can bias analysis and reporting.

Biased Questions

Estimation of Population Parameter

Sample Statistics:

Because of several constraints of collecting data from all individuals in population, including limitations of time, money, and labor, we usually depend on the sample information.

Estimation of Population Parameter (cont)

Point of Estimate:

The value of the parameters are unknown, but sample statistic is available, and can be used to estimate the population value

Characteristics of Sample Statistics:
Sample should represent the population value
Sampling techniques
Sample size

Estimation of Population Value

Concept:

Estimation of population parameter based on a sample statistic

Example:

Rutgers Parking Authority collected data from a random sample of 220 Rutgers University students in order to estimate commuting time of all the university students.

What percent of New Brunswick residents are aware of the advocacy planning efforts of the local governments in Middlesex County?

Study of 937 adults in NJ shows mean cholesterol level of 196 and standard deviation of 18 points. What can be concluded about the cholesterol level of all adults in NJ?

Estimation: Confidence Interval

Concept of the Level of Confidence:	Confidence level reflects probability that interval of estimate presents actual value of population

Estimation: Confidence Interval (cont)

Structure:	Confidence Interval is a method of inferential statistics that helps us to use sample statistic to estimate population parameter. It is based on point of estimate and margin of error.
Confidence Interval =	Point of Estimate ± Margin of Error
Translation of the Level of Confidence to z score	z score in confidence interval formula presents level of confidence, such that area under normal curve between $-z$ and +z is equal to level of confidence
Requirements:	Sampling distribution is normal: ▪ Population distribution is normal ▪ Sample size is large (n>100)

Estimation: Confidence Interval (cont)

Population standard deviation (σ) is known, Sample is taken randomly

Central Limit Theorem

Concept:

Set of ideas about the relationship between sample and population (sampling distribution and normality, and difference)

Population mean equals to the mean of the sampling distribution

If population distribution is normal, sampling distribution is also normal.

Even if the population distribution is not normal, distribution of sample means approaches normal distribution as sample size is increased, and sampling distribution is almost normal if sample size is large (n >100).

Sampling Error:

is the difference between a sample statistics and corresponding population parameter.

Sampling error can not be determined

Sampling error most probably decreases as the sample size is increased.

Standard Error:

measures the accuracy with which a sample distribution represents a population by using standard deviation.

Central Limit Theorem (cont)

$\sigma/\sqrt{n}$

Sample Size

Concept:

What is the minimum sample size required for study

Examples:

A researcher wants to collect data from a random sample of Rutgers University students in order to estimate commuting time of all the university students. How many students he must study?

What is the proper sample size if you want to check a planner's claim that more than 75% of East Brunswick residents support increase of property tax to investment on development of renewable energy projects for the city?

How large a sample must be in order to be 95% sure about the outcome of our study?

Sample Size: Quantitative Data (Mean)

Sample size for mean: $n = (\frac{z * \sigma}{E})^2$

- z = Standard score for the Level of Confidence
- σ = Population standard deviation
- E = Acceptable error

* Always round up the answer.

Sample Size: Qualitative Data (Proportion)

Sample size for proportion: $n = P\ (1 - P)(\frac{z}{E})^2$

- z = Standard score for the Level of Confidence
- p = Estimated population proportion
- E = Acceptable error

* Always round up the answer.
* Use 50% for p if estimated proportion is unknown.
* For computation use decimal format not % format.

//

Inferential Statistics: Hypothesis Testing

Concept:

Making inference about population parameter based on a sample statistic

Example:

An officer at Rutgers Parking Authority collected data from a random sample of 220 Rutgers University students in order to check the idea that daily commuting time of Rutgers students is more than 45 minutes.

A researcher claims that less than one quarter of New Brunswick residents are aware of the advocacy planning efforts of the local governments in Middlesex County? (Qualitative)

Study of 937 adults in NJ shows mean cholesterol level of 196 and standard deviation of 18 points. Can we conclude that comparing to two year ago, the mean cholesterol level of NJ adults has increased?

Inferential Statistics: Hypothesis Testing (cont)

(Confidence Interval)

> Make an estimation about population parameter

(Hypothesis Testing)

> Test validity of a claim about population parameter

Hypothesis Testing: Factors Affecting Decision

Sampling Error

> Greater the gap between sample statistics and population parameter, greater is probability of rejecting the null hypothesis

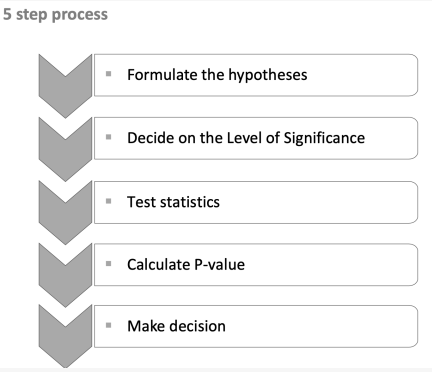
Sampling Size

> Greater the sample size, greater is probability of rejecting the null hypothesis

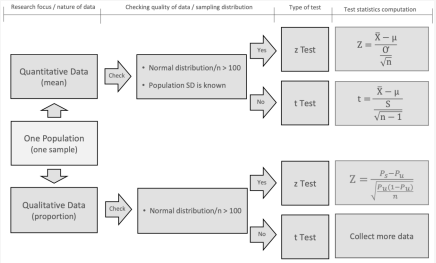
The Level of Significance

> Greater the level of significance, greater is probability of rejecting the null hypothesis

Hypothesis Testing: Steps



Hypothesis Testing: One Population



Concept: Research involves study of one population

Hypothesis Testing: Two Populations

Concept:

> Making inference about comparison of two population parameters based on two sample statistics, one from each population

Example:

> Does gender makes a difference in duration of commuting time to work?

> Is Tylenol more popular than Advil?

> Does GPA vary between undergraduate and graduate students?

> A politician claims that when comparing democrats and republicans, a greater percentage of democrats support idea of dialogue among civilizations. Is this a valid claim?

Independent vs Dependent Populations

Hypothesis Testing: Two Populations (cont)

Independent Populations / Samples

> The sample selected from one population is not related to the sample selected from the second population. // Changes within one population aren't related to changes within another population // Samples are randomly selected from these populations.

> With independent populations, we directly consider the differences in data points.

Example:

> Comparing drug A with drug B // Comparing degree of spread of corona-virus in NJ vs CA

Dependen Populations / Samples (Paired or Matched)

> Each member of one sample corres-ponds to a member of the other sample. // Data Pairs occur naturally, most often with one data point occurring "before" and another data point occurring "after" an event.

> With dependent populations, we pair the data points then consider the differences in data points.

Example:

> Blood pressure of students before and after exam

Hypothesis Testing: Qualitative Data

Hypothesis Testing for population proportion

$$Z = \frac{P_s - P_u}{\sqrt{\frac{P_u(1-P_u)}{n}}}$$

- P_s = Sample proportion
- P_u = Population proportion – Hypothesized value
- n = Sample size

Concept: Making inference about population parameter based on a sample statistic for qualitative data (nominal and ordinal - proportion)

Ex: Are proportion of households headed by single parent in the lower-income neighb-orhoods significantly different from the general population?

Ex: Are the police arrest rates in Middlesex county significantly less than the statewide rate?

Independent Populations: Qualitative Data (Z)

Step 3: Test Statistics

$$z = \frac{P_{s1} - P_{s2}}{\sqrt{(\bar{P}(1-\bar{P}))(\frac{1}{n_1} + \frac{1}{n_2})}}$$

- $\bar{P}$ is referred to as the pooled estimate of proportions = $(x_1 + x_2) / (n_1 + n_2)$
- n_1, n_2 : the sample sizes of sample 1 and 2, respectively

Requirements for qualitative data (Propo-rtion, z): Independent populations

- One sample is taken randomly from each population
- To use a normal distribution, for each population:
- Sample size is large (n > 100)

Independent Populations: Quantitative Data (Z)

Step 3: Test Statistics

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}}$$

- $\bar{x}_1, \bar{x}_2$: the sample means from sample 1 and 2, respectively
- σ_1, σ_2 : the population standard deviations from population 1 and 2, respectively
- n_1, n_2 : the sample sizes of sample 1 and 2, respectively

Requirements: Independent populations

- One sample is taken randomly from each population
- Sampling distribution is normal for each population:
- Population distribution is normal
- Sample size is large
- Standard deviation of each population (σ) is known

Independent Populations: Quantitative Data (T)

Step 3: Test Statistics

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(s_1)^2}{n_1 - 1} + \frac{(s_2)^2}{n_{2-1}}}}$$

- $\bar{x}_1, \bar{x}_2$: the sample means of samples 1 and 2, respectively
- s_1, s_2 : the sample standard deviations of populations 1 and 2, respectively
- n_1, n_2 : the sample sizes of samples 1 and 2, respectively

Requirements for quantitative data (Mean, t) : Independent populations

- One sample is taken randomly from each population
- Sampling distribution is normal for each population:
- Population distribution is normal
- Sample size is small
- Standard deviation of two populations (σ) is unknown

// (copy)

Confidence Interval- Quantitative (Z)

Confidence Interval for population mean = $\bar{x} \pm z(\frac{\sigma}{\sqrt{n}})$

- $\bar{x}$ = Sample mean
- z = Standard score for the Level of Confidence
- σ = Population standard deviation
- n = Sample size
- Margin of Error = The magnitude of the difference between the sample point estimate and the true population parameter value.
- Confidence Level = Degree of reliability of the estimate

Quantitative (ie. Normal): Confidence interval for estimation of mean

- Ex: What percentage of Rutgers University students study more than 25 hours per week?

T Distribution

$$CI = \bar{x} \pm t(\frac{s}{\sqrt{n-1}})$$

Confidence Interval: Quantitative Data (t)

T-Dist-rib-ution	used when estimating the mean of a normally distributed population in situations where the sample size is small, and/or the population standard deviation is unknown.
	The shape of distribution depends on the degrees of freedom (df = n-1).
	As the degrees of freedom increase, the t distribution approaches the standard normal distribution.

Confidence Interval: Qualitative Data (z)

Confidence Interval for population proportion = $P_s \pm z * \sqrt{\frac{P_u(1-P_u)}{n}}$

- P_s = Sample proportion
- P_u = Population proportion – Use 50% because it is un-known
- z = Standard score for the Level of Confidence
- n = Sample size
- Margin of Error = The magnitude of the difference between the sample point estimate and the true population parameter value.
- Confidence Level = Degree of reliability of the estimate

Qualitative (ie.Binomial): Confidence interval for estimation of proportion

▪ Ex: A researcher claims that at least 72% of Americans support Green Solutions for sustainable urban renewal.

// (copy) (copy)

Analysis of Variance (ANOVA)

Applications:

Mean differences across several populations

Potentials:

Works for comparison of three or more populations

Population with significant variation

Post Hoc Test:

The alternative hypothesis is not specific – it only states that at least one of the population means differs from the others. To find which category is different and how much, use post hoc (or after the fact) techniques.

Analysis of Variance (ANOVA) (cont)

Series of two Populations t Test:

To find the difference or critical group, you can use several two populations population analysis as needed (pick the important one, and then do the important one with another, and so on.

Different Types of ANOVA

One-way ANOVA:

One-way ANOVA is used to test effects of a single independent variable (categorical) on a single dependent variable (numerical).

Example:

Does income vary among African-Americans, Whites, Hispanic, and Other ethnic groups? This example includes only ONE dependent variable (income) and ONE independent variable (ethnicity).

Two-way ANOVA:

Two-way ANOVA is used to test effects of TWO independent variables on a single dependent variable

Example:

Does income vary in relation to ethnicity and personality?

Factorial ANOVA:

Factorial ANOVA is used to test effects of SEVERAL (two or more) independent variables on a single dependent variable.

ANOVA is appropriate for situations in which? we are comparing more than two populations

ANOVA

- **Step 3:** Test statistics

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares (MS)	F
Within	$SS_w = \sum_{j=1}^{k}\sum_{j=1}^{i}(X-\bar{X}_j)^2$	$df_w = k-1$	$MS_w = \frac{SS_w}{df_w}$	$F = \frac{MS_b}{MS_w}$
Between	$SS_b = \sum_{j=1}^{k}(\bar{X}_j-\bar{X})^2$	$df_b = n-k$	$MS_b = \frac{SS_b}{df_b}$	
Total	$SS_t = \sum_{j=1}^{n}(\bar{X}_j-\bar{X})^2$	$df_t = n-1$		

- K = Number of categories
- n = Total sample size

Hypothesis Testing: Analysis of Variance (ANOVA)

Concept:

Test of difference in the mean of **several populations** (Analysis of Variance)

Think of ANOVA as extension of t test for more than two populations

Example:

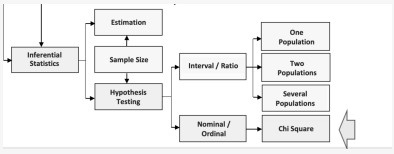
Is there a difference among Protestants, Catholics and Jews in terms of number of children?

How do Republicans, Democrats, and Independents vary in terms of income?

How do older, middle-aged, and younger people vary in terms of supermarket shopping duration?

// (copy)

Hypothesis Testing: Chi Square



Chi Square Formula

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

- f_0 = Observed frequency (data)
- f_e = Expected frequency (data)

Chi Square: Test of Independence

Concept:

Test the relationship between 2 categories through their sub-categories

Example:

We want to know if gender and type of food people prefer are associated.

We want to know if education level and political affiliation are associated.

Is type of sport students like associated with their ethnic background?

Application:

A Chi-Square (χ2) Test of Independence is used to determine existence of a significant association between two categorical variables.

Is type of sport students like associated with their ethnic background?

Chi Square: Test of Goodness-of-Fit

Concept:

Test validity of a particular distribution of population subcategories (how well sample data fit a distribution from a population with a normal distribution).

Chi Square: Test of Goodness-of-Fit (cont)

Example

A professor claims that in metropolitan areas, 19% of college students are Black, 42% are White, 16% are Hispanic, and the rest have a different ethnic background. Is this a valid claim?

Isthisvalidtoassumethatlocalgymshave-theirhighest attendance on Mondays, Tuesdays and Saturdays, average attendance on Wednesdays and Thursdays, and lowest attendance on Fridays and Sundays.

A researcher claims that season makes no difference in number of gun violations in large urban centers.

Applications

Study pattern of distribution of subgroups

// (copy)

// (copy)