Inferential Statistics

Dr. Upendra Pratap Singh

LNMIIT

September 8, 2023

1. Introduction to Inferential Statistics

2. Motivation

Inferential Statistics

Introduction

1. Inferential statistics involves *drawing conclusions/making inferences* about a population based on a sample of data from that population.
2. Useful when it is not feasible to get the population data.

Inferential Statistics

Motivation

1. *Cost and Resource Constraints:* expensive data collection, logistic chain complexity.
2. *Time Efficiency:* conducting a study on an entire population might take an impractically long time.
3. *Destructive Testing:* in fields such as medical research, destructive testing might be involved, where samples are consumed or destroyed during the testing process.
4. *Population Variability:* populations are often diverse and heterogeneous. By studying a sample, researchers can capture a range of variation that is representative of the broader population.

Inferential Statistics

Motivation

1. *Infeasibility:* inferential statistics offers a way to work with manageable subsets while still making meaningful inferences.
2. *Ethical Considerations:* it might not be ethical to gather data from an entire population.
3. *Testing Hypotheses:* researchers often have specific hypotheses or questions they want to address. Inferential statistics provides a framework for testing these hypotheses using sample data and making conclusions about the population.

Inferential Statistics

Introduction

What it involves ?

1. *Sampling:* This is the process of selecting a subset of individuals or items from a larger population to represent it.
2. *Hypothesis Testing:* This involves making educated guesses about a population parameter and using sample data to test whether the hypotheses is likely to be true or not.
3. *Confidence Intervals:* finding a range around a sample statistic (like the sample mean) that is likely to contain the corresponding population parameter.

Inferential Statistics

Introduction

What it involves ?

1. *Regression Analysis:* model the relationship between dependent and independent variables and use that model to make predictions.
2. *Estimation:* involves estimating population parameters based on sample statistics.
3. *Significance Testing:* process of determining whether an observed effect in a sample is *statistically significant* or if it could have occurred by chance.
4. *Probability:* to quantify the likelihood of different outcomes or events occurring, given a certain set of assumptions.

Inferential Statistics

Statistic Vs Parameter

Statistic

1. A statistic is a numerical measure *calculated from a sample*.
2. May *differ* from sample to sample reflecting the inherent variability in the sampling process.
3. Examples include: sample mean $(\bar{X})$, sample standard deviation and sample median, etc.

Inferential Statistics

Statistic Vs Parameter

Parameter

1. A parameter is a numerical measure that describes a specific characteristic of an *entire population.*
2. *Fixed* value that may not be observable in its entirety.
3. Serve as targets of inference in statistical analysis.
4. Examples include: population mean (μ), population standard deviation (σ) and variance.

Inferential Statistics

Estimator

1. An estimator is a statistic or a method used to calculate an estimate of an unknown parameter.
2. In other words, an estimator is a formula, a calculation, or a procedure that takes observed sample data as input and produces an estimated value for a population parameter as output.

Inferential Statistics

Estimator Types

1. *Point Estimator:* A point estimator is a specific value calculated from the sample data that serves as an estimate for the true parameter value; example: sample mean $(\bar{x})$.
2. *Interval Estimator:* An interval estimator provides a range of values within which the true parameter value is likely to fall; example: confidence intervals.
3. *Maximum Likelihood Estimator (MLE):* finds parameter values that maximize the likelihood of observing the given sample data.

Inferential Statistics

Estimator Types

1. *Bayesian Estimator:* an estimator incorporates both prior beliefs and the likelihood of the data to calculate a posterior distribution.

$$\boxed{P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}} \tag{1}$$

2. *Robust Estimator:* is less sensitive to outliers or deviations.
3. An ideal estimator would have *low bias* and *low variability*.

Inferential Statistics

Properties of a good estimator

1. **Unbiased:** The center of the sampling distribution for the estimate is the same as that of the population.
2. The estimate has the *smallest standard error* when compared to other estimators.
3. **Standard error:** measures the *variability of sample statistics* across multiple samples.

$$\boxed{Standard \;\; Error = \frac{Standard \;\; Deviation}{\sqrt{N}}} \tag{2}$$

Inferential Statistics

Point Estimate: Mean

1. Estimating the unknown population parameter is straightforward.
2. Calculate the sample mean ($\bar{x}$) and use it as an estimate of true population mean μ.
3. In other words, $\hat{\mu} = \bar{x}$.

Inferential Statistics

Point Estimate: Standard Deviation

1. Sample needs to be reasonable sized. Why?. For example, in a sample sized 1, standard deviation will always be 0.
2. Minimum of *two* samples needed to observe any variability.
3. Sample deviation s is usually small compared to the population standard deviation σ. Moreover, sample deviation s increases with the size of the sample.

$$\boxed{s^2 = \frac{1}{N}\sum_{i=1}^{i=N}(X_i - \bar{X})^2} \tag{3}$$

Inferential Statistics

Point Estimate: Standard Deviation

1. In other words, sample variance s^2 is the biased estimator of true population variance σ^2
2. Solution: Bessels's correction

$$\boxed{\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2} \tag{4}$$

Homework: Find the point estimate for a median.

Inferential Statistics

Interval Estimate: Confidence Intervals

1. Confidence interval: an interval of values computed from the sample data that is likely to cover the true parameter.
2. For a good CI, the center of the interval should be the point estimate for the parameter of interest.

Inferential Statistics

Interval Estimate: Confidence Intervals

1. General form of confidence interval:

$$\boxed{Confidence\ interval = sample\ statistic \pm margin\ of\ \ error} \tag{5}$$

2. General form of margin of error:

$$\boxed{margin\ of\ \ error = M \times \hat{SE}} \tag{6}$$

   where $\hat{SE}$ is the standard error of sample statistic and M is a multiplier based on how confident we want to be in.

3. **Interpretation:** We are 95% confident that the true population parameter lies within the calculated confidence interval.

Inferential Statistics

Confidence Intervals: Advantages and Limitations

Advantages

1. An easy interpretation to quantifies uncertainty.
2. Incorporates the sample size.
3. Maybe adjusted for different confidence intervals (M needs to change accordingly).

Limitations

1. Doesn't ensure true parameter inclusion.
2. No information about distribution shape, bias.
3. Sensitive to outliers.

Inferential Statistics

Hypothesis Testing: P-Value

1. Hypothesis Testing: a more formal method for testing whether a given value *(hypothesized value)* is a reasonable value of a population parameter.
2. Fundamental steps:
   - Compare data from a sample to a hypothesized parameter.
   - Compute the probability that a population with the specified parameter would produce a sample statistic as extreme or more extreme to the one we observed in our sample.
   - The probability value computed in the above step is the $p - value$.
3. **Goal:** to determine whether there is enough evidence to support a particular claim.

Inferential Statistics

Hypothesis Testing: Null and Alternate Hypothesis

1. The null hypothesis ($\mathbf{H}_0$) and the alternative hypothesis ($\mathbf{H}_1$) are two *competing* statements used to make inferences about a population parameter based on sample data.
2. The above statements play a central role in the process of hypothesis testing to determine whether there is enough evidence to support a particular claim.

Inferential Statistics

Hypothesis Testing: Null Hypothesis

1. Represents the status quo, a *default* assumption, or a lack of an expected effect.
2. In other words, the null hypothesis is a *statement of no effect*
3. Mathematically, null hypothesis often *includes equality* statements; for example, $\textit{population mean} = 25.5$

Inferential Statistics

Hypothesis Testing: Alternate Hypothesis

1. *Contradicts* the null hypothesis.
2. It represents what we are trying to *find evidence for* — a claim of an effect, a difference, or a change in the population parameter.
3. It can be one-sided (greater than or less than) or two-sided (not equal to) depending on the research question we are trying to answer.

Inferential Statistics

Hypothesis Testing: Type I Error

Type I Error (False Positive)

1. Occurs when we *reject* the null hypothesis when it is actually true; in essence, we conclude that there is an effect or difference when there isn't one in the population.
2. The probability of making a Type I error is denoted by the significance level (α).
3. Example: convicting an innocent person in a court trial.

Inferential Statistics

Hypothesis Testing: Type II Error

Type II Error (False Negative)

1. Occurs when we fail to reject the null hypothesis when it is actually false; in essence, we conclude that there is no effect or difference when there is one in the population.
2. The probability of making a Type II error is denoted by β; power given by $(1-\beta)$
3. Example: Not diagnosing a disease in a person who actually has the disease.

Inferential Statistics

Hypothesis Testing: Type II Error

Type II Error (False Negative)

1. Occurs when we fail to reject the null hypothesis when it is actually false; in essence, we conclude that there is no effect or difference when there is one in the population.
2. The probability of making a Type II error is denoted by β; power given by $(1-\beta)$
3. Example: Not diagnosing a disease in a person who actually has the disease.

In practice, there's often a trade-off between Type I and Type II errors. Adjusting the significance level (α) can impact the balance between these errors.

Inferential Statistics

Hypothesis Testing: Significance level

1. **Significance level** (α)**:** a predetermined threshold used in hypothesis testing to determine the level of evidence required to reject the $\mathbf{H}_0$.
    - A lower significance level ($e.g., 0.01$) requires stronger evidence to reject $\mathbf{H}_0$; reduces the risk of falsely rejecting a true null hypothesis.
    - A higher significance level ($e.g., 0.10$) requires less evidence to reject the null hypothesis but it also increases the risk of making a Type I error.
2. **Interpretation:** If the result is statistically significant, it means that the observed data is unlikely to have occurred under the assumption that the null hypothesis is true.

Inferential Statistics

Hypothesis Testing: Test Statistic

1. Test statistic: is a numerical value calculated from sample data that is used to assess the strength of evidence against the null hypothesis.
2. In other words, it quantifies the difference between the sample data and the null hypothesis's expected values.
3. The test statistic's behavior helps determine whether the observed data provides enough evidence to either reject or fail to reject $\mathbf{H}_0$.
4. Choice of test statistic depends on the data itself; common examples include **Z**- score, **T**- score and χ^2-square test.

Inferential Statistics

Hypothesis Testing: Steps Involved

1. State $\mathbf{H}_0$ and $\mathbf{H}_1$
2. Choose a significance level α
3. Collect the data and then compute the test statistic
4. Calculate the p-value
    - A p-value smaller than the significance level (α) suggests stronger evidence against the null hypothesis; the null hypothesis is rejected in favor of the alternative hypothesis.
    - Alternatively, a p-value greater than α means that the null hypothesis is not rejected; there is not enough evidence to support the claim made by the alternative hypothesis.

Inferential Statistics

Hypothesis Testing: One Sample Test

One Sample T Test

1. AKA *single sample t-test.*
2. Is a statistical hypothesis test used to determine whether the mean calculated from sample data collected from a single group is different from a designated value.
3. The designated value does not come from the sample itself.
4. The different hypothesis are:
   - $\mathbf{H}_0$: The population mean equals the specified mean value
   - $\mathbf{H}_1$: The population mean is different from the specified mean value
5. Test determines if there is enough evidence to reject $\mathbf{H}_0$ in favor of $\mathbf{H}_1$.

Inferential Statistics

Hypothesis Testing: One Sample Test

One Sample T Test: Use Case

1. **Assumptions:**
   - Sample size must be large $N \geq 30$.
   - Population must be known to be normally distributed.
   - α is predetermined.
2. **Hypothesis**:
   - $\mathbf{H}_0 : \mu = \mu_0$
   - $\mathbf{H}_1 : \mu \neq \mu_0$
3. **Compute Test Statistic**: one group mean

$$\boxed{t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}} \tag{7}$$

where s is the sample standard deviation n is the sample size.

Inferential Statistics

Hypothesis Testing: One Sample Test

One Sample T Test: Use Case

1. **Compute** $p - value$**:** use a t-distribution to find a $p - value$.
2. **Decision Making:**
   - If $p \leq \alpha$, reject the null hypothesis $\mathbf{H}_0$.
   - If $p > \alpha$, fail to reject the null hypothesis.
3. State the *real world research finding* based on the above step.

Inferential Statistics

Hypothesis Testing: Two Sample Test for Mean

Two Sample Test:

1. Determines if two samples correspond to the same population mean.
2. Requires some extra steps:
   - If samples are independent.
   - Inherent variability in each sample.

Inferential Statistics

Hypothesis Testing: Two Sample Test for Mean

Independent and Dependent Samples

1. **Independent samples:** if the samples selected from one of the populations have no relationship with the samples selected from the other population.
2. **Dependent samples:** if each measurement in one sample is matched or paired with a particular measurement in the other sample

Inferential Statistics

Hypothesis Testing: Two Sample Test for Mean

1. Assumption: samples are reasonable sized and population follows a normal distribution.
    - If $\mu_1 - \mu_2 = 0$: then there is no difference between two population means.
    - If $\mu_1 - \mu_2 \neq 0$: then the two population means are different.

Inferential Statistics

Hypothesis Testing: Two Sample Test for Mean

Computation of $t-$statistics

1. The $t-$statistic is computed as:

$$\boxed{t^* = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_{x_1}^2}{n_{x_1}} + \frac{s_{x_2}^2}{n_{x_2}}}}} \tag{8}$$

where $s_{x_1}^2$ and $s_{x_2}^2$ are the sample variances and n_{x_1} and n_{x_2} are the corresponding sample sizes.

Inferential Statistics

Regression

Simple Linear Regression

1. Statistical method that allows us to *summarize* and *study relationships* between predictor x and response $\mathbf{y}$.
2. *Simple:* only one predictor variable is considered; *multiple* linear regression also exists.
3. The relationship we study is *statistical* and not *deterministic*.
4. Visualization tool: scatter plot is suited.

Inferential Statistics

Simple Linear Regression

Best Fit Curve

1. To determine the best fit curve, the following notations are useful:
   - y_i: observed response
   - $\hat{y}_i$: predicted response
   - x_i: predictor value
2. With the above notation in place, the best fit line may be written as:

$$\boxed{\hat{y}_i = \beta_0 + \beta_1 x_i} \tag{9}$$

3. Clearly, the predictions made may not be correct, so we define the **residual error** as:

$$\boxed{e_i = y_i - \hat{y}_i} \tag{10}$$

Inferential Statistics

Simple Linear Regression

Best Fit Curve

1. The best-fit line should have minimum residual error $e_i \; \forall i = 1, 2, \ldots N$.
2. In other words, the best-fit line minimizes the sum of total squared residuals given by:

$$\boxed{Q(\beta_0, \beta_1) = \sum_{i=1}^{i=N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{i=N} (y_i - (\beta_0 + \beta_1 x_i))^2} \qquad (11)$$

3. **Method of Least Squares:** *Squared differences* are considered so that they don't cancel the effect of one another.

Inferential Statistics

Simple Linear Regression

Best Fit Curve

1. In the determination of the best-fit line, the parameters β_0 and β_1 need to be estimated.
2. Using differential calculus, the estimates may be obtained by setting:

$$\boxed{\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} = 0} \tag{12}$$

and

$$\boxed{\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} = 0} \tag{13}$$

Inferential Statistics

Simple Linear Regression

Best Fit Curve

1. By solving the above partial differential equations, the optimal estimates β_0^* and β_1^* are obtained as:

$$\boxed{\beta_0^* = \bar{y} - \beta_1^* \bar{x}} \tag{14}$$

where

$$\boxed{\beta_1^* = \frac{\sum_{i=1}^{i=N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{i=N}(x_i - \bar{x})^2}} \tag{15}$$

2. The line represented by the coefficients, β_0^* and β_1^* is **least squares regression line**.

Inferential Statistics

Analysis of Variance

1. Also known as *ANOVA*.
2. It helps understand whether the differences observed in sample means are likely due to actual differences in the populations they represent or if they could have occurred by random chance.
3. **Categorization:**
   - *One-way:* for comparing means of groups.
   - *Two-way:* for examining the effect of two independent variables on a dependent variable.

Inferential Statistics

Analysis of Variance

Hypothesis

1. $\mathbf{H}_0 : \mu_1 = \mu_2 = ... = \mu_K$
2. $\mathbf{H}_1 : \mu_i \neq \mu_j$ *for some i and j where* $i \neq j$
3. **Interpretation:**
   - Null hypothesis suggests that the factor did not have any significant impact on the results obtained; in other words, the different groups may be considered to be the part of the same population.
   - Alternate hypothesis states that the means of at least one of the pairs is not equal.

Inferential Statistics

Analysis of Variance

Test Statistics: F-Ratio

1. For more than 2 groups, we use test statistic $F-$ ratio defined as the ratio of *between-group sample variance* and the *within-group-sample variance*

$$\boxed{F = \frac{\textit{between sample group variance}}{\textit{within sample group variance}}} \tag{16}$$

2. **Interpretation:**
   - Measures whether the means of different samples are significantly different.
   - $F-$ratio higher than the $F-$critical value suggests evidence against the null hypothesis.

Inferential Statistics

Analysis of Variance

Between-Group Sample Variability

1. Computation similar to standard deviation.
2. Sample deviations are weighted by sample sizes.
3. Between sample variability is computed as:

$$MS_{between} = \left(\frac{n_1(\bar{x}_1 - \bar{x}_G)^2 + n_2(\bar{x}_2 - \bar{x}_G)^2 + ... + n_K(\bar{x}_K - \bar{x}_G)^2}{K - 1} \right) \tag{17}$$

where $\bar{x_G}$ is the grand mean.

Inferential Statistics

Analysis of Variance

Within-Group Sample Variability

1. No interactions between the samples.
2. Measured by looking at how much each value in the sample differs from its respective sample mean
3. Within-group sample variability is computed as:

$$MS_{within} = \left(\frac{\sum(x_{ij} - \bar{x}_j)^2}{N - K}\right) \tag{18}$$

   where j corresponds to a group.
4. Conclude given $F - ratio$ computed and F-critical value.

Inferential Statistics

Analysis of Variance

Limitation

1. Tells us that at least two groups are different.
2. However, it does not tell us which groups are different.

Relation with $t - test$

1. With only two samples, $t-$test and ANOVA give same result;
2. $t-$test is not applicable for more than 2 samples.

Test of Independence

Introduction

1. AKA Chi-Square Test of independence.
2. Provides statistical evidence of an *association* or *relationship* between the two categorical variables.
3. Do not confuse it with correlation. Why?

Test of Independence

Introduction

1. **Assumptions:**
   - The sample must be been obtained *randomly*.
   - Variables must be *categorical*: nominal or ordinal.
2. **Data Representation:**
   - Contingency table (also known as **cross-tab).**
   - Tabulates the relationship between two categorical variables.

Test of Independence

Contingency Table

		Sport Preference			
		Archery	Boxing	Cycling	
Gender	Female	35	15	50	**100**
	Male	10	30	60	**100**
		45	**45**	**110**	**200**

Figure: Contingency Table

Test of Independence

Contingency Table

Steps:

1. **Hypothesis:**
   - $\mathbf{H}_0$: the two categorical variables are independent.
   - $\mathbf{H}_1$: the two categorical variables are not independent.
2. Fix α and hence, χ^2_α.
3. Now, given the observed counts in the contingency table, compute the expected counts under $\mathbf{H}_0$.

$$\boxed{E = \frac{(row\ total)(column\ total)}{total\ sample\ size}} \tag{19}$$

Test of Independence

Contingency Table

Steps:

1. What question is being answered ?
   - Are the observed counts so different from the expected counts that we can conclude a relationship exists between the two variables?
2. Compute χ^2 statistic: in a summary table of $r \times c$ entries:

$$\boxed{\chi^{2*} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \ldots + \frac{(O_{rc} - E_{rc})^2}{E_{rc}}} \tag{20}$$

3. Conclude by comparing χ^{2*} and χ^2_α.