207 Xavier Hall Annex seating arrangement

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tahani Al-Rajeh | Yit Mui Khoo | Nathan House | Laura Gardner | Diana Ramirez |
| Joe Fetter | X | Malcolm Townes | Darren O’Brien | X |
| James Tillis | X | Ashley | Destiny Brooks | Courtney |
| Renee Jonas | Saeed Asiri | X | X | X |

Recommend text: SPSS Survival Manual

Data for final project

* The data set for the project must include nominal, ordinal, and ratio variables and a minimum sample size of n=100.
* Clean the data before using it.

Article review assignment

* The article for the article review assignment must use regression analysis.
* Pair up for the article review assignment and presentation.
* Decide on pairs in the next couple of weeks.

Nominal and ordinal data are both referred to as categorical data.

Mean, median, and mode

* Three moments in statistics
* Measures of central tendency
* We’re not concerned with outliers in statistical analysis

Confidence interval is the level of confidence that the real mean falls between two points.

It’s a good practice to include descriptive statistics in data analysis.

This course is concerned with Type 1 errors; Type 2 errors concern Bayesian statistics.

Significance levels are 0.05 (\*), 0.01 (\*\*), and 0.001 (\*\*\*)

Degrees of freedom = sample size - number of variables

Chi-Square Test

* Useful for nominal variables
* Requires a minimum count of 5 for each variable

Standard coefficient is equivalent to the Pearson correlation only in simple regression

Data set requirements for final project

* Minimum of 5 variables
* Minimum of 3 ratio variables
* Cross-sectional data is fine (i.e., longitudinal data is not necessary)

Options for identifying a group for the journal article review and presentation assignment

* Work with other students in the Public and Social Policy (PSP) program
* Work with Nathan House and Saeed Asiri \*

Technically ordinal data should not be used for regression analysis but is sometime acceptable.

General Social Survey (GSS) is conducted by the National Opinion Research Center (NORC)

Residual is the difference from the target value (e.g., mean value)

Null hypothesis: The percent of male in the sample is the same as the percent of female in the sample

Alternative hypothesis: The pct of male in the sample is NOT the same as the pct of female in the sample

H0: pM - pF and HA = pM ≠ pF

Null hypothesis: The percent of each race in the sample is the same as the percent of each race in the population

Alternative hypothesis: The percent of each race in the sample is the same as the percent of each race in the population H0: white = pwhite , etc. and HA = white ≠ pwhite , etc.

Degrees of freedom for two variables = (n1-1)( n2-1)

Generally, hypotheses are structured as follows

* H0: There is NOT an association
* HA: There is an association

H0: μM = μF 🡪 The population mean of males is the same as the population mean of females

HA1: μM ≠ μF 🡪 The population mean of males is NOT the same as the pop. mean of females (2-tail test)

HA2: μM > μF 🡪 The population mean of males is greater than the pop. mean of females (1-tail test)

To use T-test, the test variable must be a ratio variable and there must only be a total of two variables.

ANOVA is used for testing 3 or more variables

Paired T-test is used for comparing two sets of observations on the same unit of analysis (i.e., before and after tests, twins, married couples, etc.)

Data set for class

* NBER patent citation data is okay, per Dr. Matsuo
  + Use data from 1994-1999
  + Between 100 and 200,000 observations is good
  + Data must have at least three continuous variables, not necessarily ratio variables

Assignment #1

* Due by 12 noon on Thursday, September 20th.

Journal article review and presentation

* Submit chosen article to Dr. Matsuo for pre-approval

The more objective the data the more likely to get a high R2 value if a strong correlation exists.

In statistics, the term “parsimonious” means economical.

Sample distribution is NOT the same as sampling distribution.

Example…

Population, N=10,000

Sample size, n=100

Number of samples possible (of size n=100), k=100 (i.e., 10,000 / 100)

Sample means are : = V1 … = Vk

Sample distribution is the distribution of x values for each sample (i.e., frequencies as in a histogram)

Sampling distribution is the distribution of the sample means, which should be approximately normally distributed according to the Central Limit Theorem

Standard error = standard deviation of the sampling distribution

Ideally, errors (residuals) should not correlate with each other.

Autocorrelation is when the residuals of a single variable correlate with one another.

Autocorrelation is about how the residual of one data point influences the residual of other data points in a sample for one variable.

Multicollinearity is about whether there is a linear relationship between two independent variables.

Standardized coefficient beta

* Measures the increase in the contribution of a variable that results from a one standard deviation change in the value of the variable.
* |t-value| ≥ 2.0 🡪 likely significant