Statistics Recap

M. Fuat Kına

Statistics recap

- Descriptive statistics
  - Collecting, presenting, and describing data

- Inferential statistics
  - Drawing conclusions and/or making decisions concerning a population based only on sample data

- A **Population** is the set of all items or individuals of interest.
- A **Sample** is a subset of the population.

Statistics recap

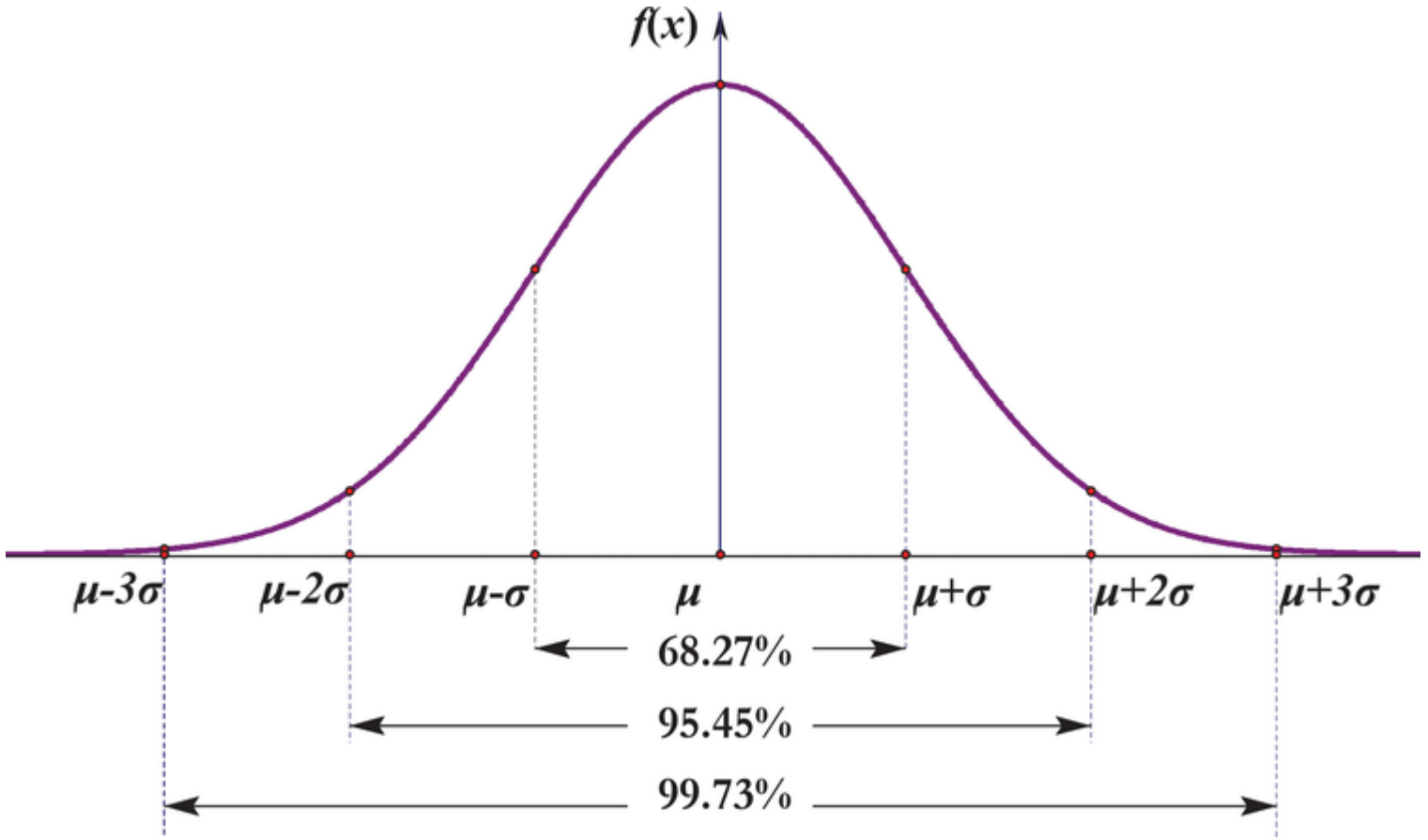
- **Univariate**: mean, mode, median, quartiles, range, standard deviation, variance, frequency, etc.
- **Bivariate**: crosstabs, t-test, anova, correlations, etc.
- **Multivariate**: regressions, etc.

- Data types matter:
  - Categorical, nominal
  - Ordinal, hierarchical
  - Continuous, numeric

Bivariate

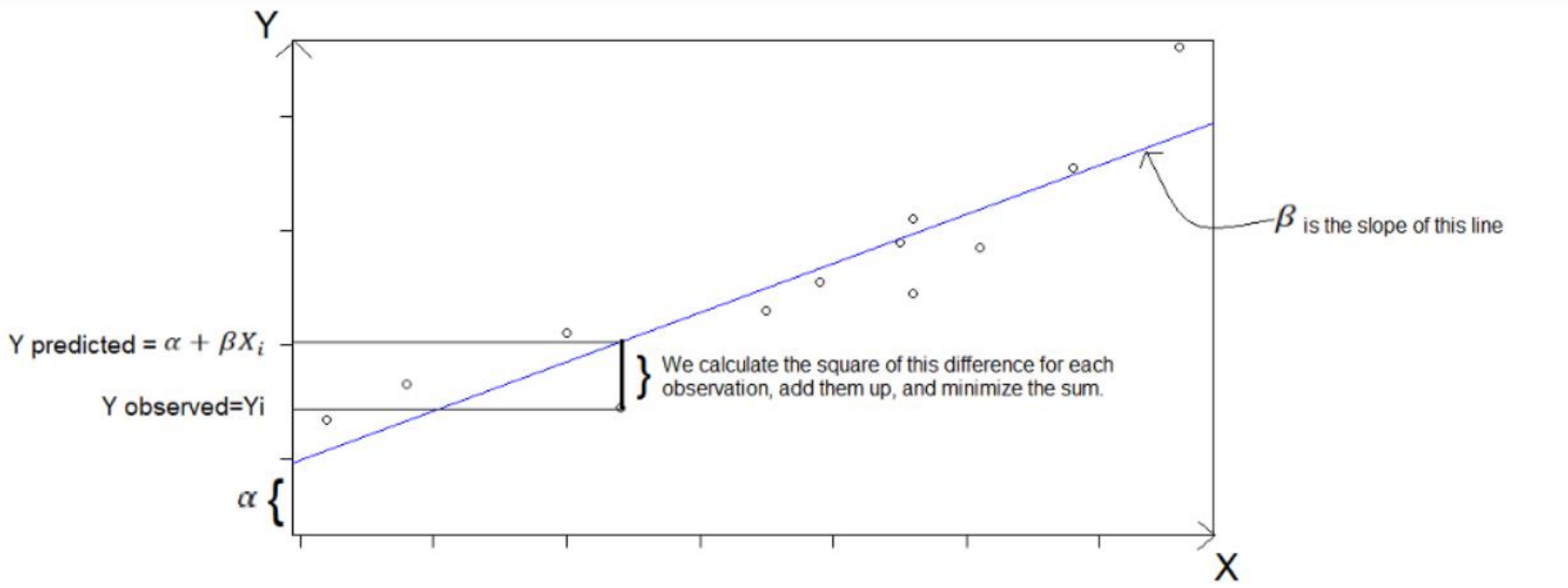
- **Cross-tabulation** analysis (crosstab) is the initial level for bivariate analyses.
  - It presents a descriptive relationship between two categorical (nominal or ordinal) variables. Note that it is not common to use crosstabs for numeric (interval or ratio) variables.
- **T-test** analysis is an inferential statistical test that determines whether there is a statistically significant difference between the means in two unrelated groups.
  - For a t-test analysis we need one categorical (nominal or ordinal) variable and one continuous viariable. However note that categorical variable should have only two values, which we will compare.
- **One-way ANOVA** analysis is an inferential statistical test that determines whether there is a statistically significant difference between the means in more than two unrelated groups. It is very much similar to T-test analysis. The only difference is the number of values.
  - For a one-way ANOVA analysis, we need one categorical (nominal or ordinal) variable and one continuous variable. However, note that the categorical variable should have more than two values, which we will compare.
- **Correlation** basically refers “a mutual relationship or connection between two or more things.” It is a statistical measure that indicates the extent to which two or more variables fluctuate together. A positive correlation indicates the extent to which those variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable increases as the other decreases.
  - Positive Correlation – when the value of one variable increases with respect to another.
  - Negative Correlation – when the value of one variable decreases with respect to another.
  - No Correlation – when there is no linear dependence or no relation between the two variables.

Reminder: Confidence intervals for normal distribution



Linear regression (OLS)

- OLS is a good use-case, and will be important while learning basic prediction mechanism behind Machine Learning models
- Derivation of OLS



Predicting the DV in real life

- According to DV structure
  - Logistic regression: Binary
  - Mutinomial regression: Ordinal/categorical (including more than two values)
  - Poisson regression: Right skewed
  - Negative binomial regression: Right skewed + count variable

- Multilevel regressions
- Structural equation models
- Working with Panel Data structures (time series cross sectional data)
- Spatial and temporal autocorrelations
- Clustered standard errors, robust standard errors

Homework hints

- Calculate n, k
- $\hat{\beta} = (X'X)^{-1}X'y$
- Covariance matrix $= \sigma^2(X'X)^{-1}$
- $\hat{\sigma}^2 = \frac{e'e}{n-k}$
- X’ = X.T
- @ is matrix multiplication
- numpy.linalg

Practical session for basic statistics