MET CS555 B1

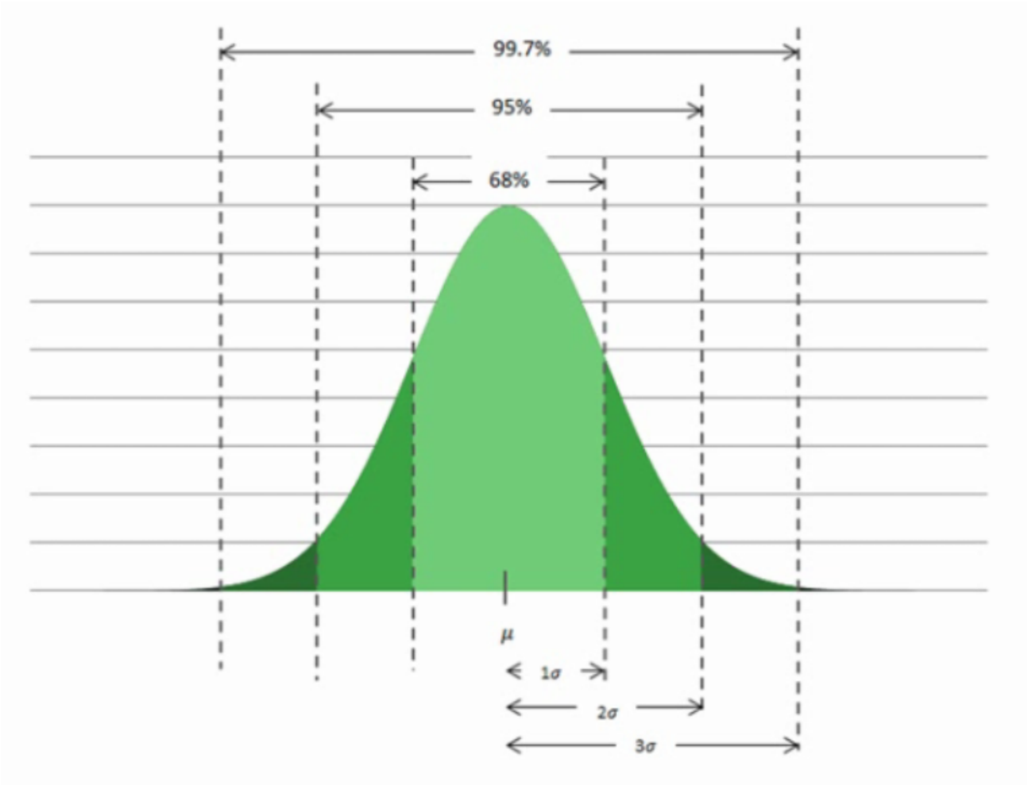
Data Analysis and Visualization

Final Exam Preparation

Kia Teymourian

Key Concepts – Normal Distribution

- Properties of the Normal Distribution
  - Symmetric
  - Area sums to calculate the probabilities



Key Concepts – Central Limit Theorem

- Central Limit Theorem
  - If you have a sample that is sufficiently large, then the distribution of the sample mean will be approximately normally distributed with
    - A mean of μ mu (equal to the population mean)
    - Standard deviation of $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

      (equal to the population standard deviation divided by the square root of the sample size n)
  - As the **sample size increase**, the **variability decreases**

Key Calculations – Z-score

- How we calculated the z-score

$$z = \frac{\bar{x} - \mu}{\sigma}$$

Or

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Key Concepts

- Null and alternative hypotheses for each type of analysis
- Null value of each statistic (**correlation coefficient, beta coefficient, OR, risk difference, risk ration**)

- When to use which analysis
  - **One-sample test for means**
  - **Two-sample test for means**
  - **Correlation/Simple Linear Regression**
  - **Multiple Linear Regression**
  - **ANOVA**
  - **Logistic Regression**

Key Concepts

- **Interpretation of confidence intervals**
  - We are 95% confident that the underlying …
  - Relationship with testing (across different types of analysis)
    - A level **α** alpha significance test **rejects the null hypothesis**

      H_0: $\mu = \mu_0$ when value of **μ_0** is **not included** in the **1-α** confidence interval for **μ**
    - A level **α** alpha significance test fails to reject **the null hypothesis** **H_0: $\mu = \mu_0$** when value of **μ_0** is **included** in the **1-α** confidence interval for **μ**
    - The conclusion of a **two-sided significance test** (whether or not the null hypothesis is rejected is rejected) at the **α** alpha significance can be determined by **checking if the “null” value** as specified by the null hypothesis **is contained within the 1-α confidence interval**

Key Calculations

- **Z-Score**
- Prediction using regression equation
  - **Linear Regression**

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

  - **Logistic Regression**

$$\hat{p} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k}}$$

- **Residuals**

$$\hat{y} - y$$

Linear Regression

- Relationship between beta estimate and correlation coefficient

  - **Positive Association ($r>0$, $\beta > 0$)**
    - As one variable goes **up**, the other goes **up**
    - As one variable goes **down**, the other goes **down**

  - **Negative Association ($r< 0$, $\beta < 0$)**
    - As one variable goes **up**, the other goes **down**
    - As one variable goes **down**, the other goes **up**

Linear Regression

- Interpretation of beta estimates (in **SLR vs. MLR**)
- Calculation of **Confidence Intervals**
- Purpose of diagnostic plots
  - **Residual Plots:**
    - **Residual Plot: Asses regression assumptions** (**Linearity, Constant Variance)** and to identify outliers
    - **Histogram of the Residuals: Asses regression assumptions** (**normality**) and to identify observations with large residuals

Key Concepts

- **Calculate p-value**
  - One-sided vs. Two-sided
  - Using tables, based on the z, t or F distributions

- **ANOVA table dependencies**
  - How df is calculated for each type of analysis (regression, ANOVA)

	SS (Sum of Squares)	df (degrees of freedom)	MS (Mean Square)	F-statistic	p-value
Regression	Reg SS	Reg df$= k$	Reg MS = Reg SS/Reg df	F=Reg MS/Res MS	$P(F_{\text{Reg df,Res df},\,\alpha} > F)$
Residual	Res SS	Res df$= n - k - 1$	Res MS = Res SS/Res df		
Total	Total SS = Reg SS + Res SS				

Key Concepts

- **Multiple comparison procedures**
  - When we do lots of comparison, we increase our **risk of a Type I error**

  - To avoid this. **We need to make it harder to reject.** This can be accomplished in different ways (all equivalent):
    - Increase the p-value
    - Decrease the significance level used for each test
    - Increase the critical value used for the decision rule

  - Various methods for doing this, most simple is **Bonferroni**
  - The **Bonferroni methodology** suggests that individual tests should be performed at the $\alpha_* = \alpha/c$ level of significance,

    like $\alpha_* = \alpha/c = 0.05/3 \approx 0.0167$

Logistic Regression

- **Odds ratio for 1-unit or x-unit increase in explanatory variable**

$$\widehat{\mathrm{OR}}_{x_a\,versus\,x_b} = e^{\hat{\beta}_1 x_a - \hat{\beta}_1 x_b} = e^{\hat{\beta}_1 (x_a - x_b)}$$

- Odds Ratio **Confidence Interval:**

$$\widehat{\mathrm{OR}}_{x_a\,versus\,x_b} = e^{\left(\hat{\beta}_1 \pm z_{\frac{\alpha}{2}} * SE_{\hat{\beta}_1}\right)(x_a - x_b)}$$