STAT 4033

Final Notes Name: Cooper Morris

C1S3:

Histogram: Number of classes should be smallest whole number K that makes $2^K \geq$ number of measurements. For large data sets either $\log_2(n)$ or $2n^{1/3}$

Unimodal: One major peak

Bimodal: Two major peaks

Symmetric: Symmetric

Right Skewed: Long right tail, short left tail

Boxplots: Outliers are outside 1.5×IQR. Box goes from Q_1 to Q_3, horizontal line at median, whiskers to largest data point inside 1.5×IQR, X's for outliers

Five Number Summary: Min, Max, Median, Q1, and Q3

C2S6:

Jointly Distributed Random Variable: Two or more random variables that are related when considering "individuals" in a population.

Joint Probability Mass Function:

$P(X = x, Y = y) = p(x, y)$ $P(X = x \cap Y = y)$

Marginal Probability Mass Function:

$P_x(x) = \sum_y p(x, y)$

$f_x(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$

Summing or integrating out the opposite variable gives you the marginal probability mass function for the variable you desire.

C6S1: Large Sample Tests for Population Mean

$$z_{test} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Alternative Hypothesis: The claim about the population that we are trying to find evidence for.

Null Hypothesis: What is assumed to be true.

Reject the null hypothesis if P-Value is less than α

C6S3: Tests for a Population Proportion

$$z_{test} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 \cdot (1-p_0)}{n}}}$$

C6S4: Small Sample Tests for μ

$$t_{test} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Use Table 2

C6S7: Small Sample Tests for Difference Between Two Means

If $\sigma_1 = \sigma_2$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 + 2}$$

$$t_{test} = \frac{(\bar{x_1} - \bar{x_2}) - D_0}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

If $\sigma_1 \neq \sigma_2$

$$t_{test} = \frac{(\bar{x_1} - \bar{x_2}) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$\nu = \frac{(\frac{s_1^2}{n^1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2 - 1}}$$

C6S8: Tests with Paired Data

$$t_{test} = \frac{\bar{d} - D_0}{\frac{s_d}{\sqrt{n}}}$$

Where $\bar{d}$ is the difference between sample means and s_d is the difference between standard deviations.

	Null Hypothesis	
Our Decision	True	False
Reject H_0	**X** Type 1 error (false positive)	✓
Fail to reject H_0	✓	**X** Type 2 error (false negative)

C7S1: Linear Correlation

Correlation Coefficient: The direction and strength of a linear relationship.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$-1 \leq r \leq 1$

C7S2: Least Squares Line

$y = \beta_0 + \beta_1 x + \epsilon$

$$b_1 = \frac{SS_{xy}}{SS_{xy}}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

Sum of Squared Error (SSE):

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

Estimate Standard Deviation:

$\hat{\sigma}_\epsilon = s = \sqrt{\frac{SSE}{n-2}}$

C7S3: Uncertainties in the Least-Squares Coefficients

$$s_{b0} = s \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}}$$

$$s_{b1} = \frac{s}{\sqrt{SS_{xx}}}$$

$$b_i \pm t_{\alpha/2,n-2} \cdot s_{bi}$$

$$t_{test} = \frac{b_i}{s_{bi}}$$

S is residual standard error if given R output.

Use a t-distribution with n-2 degrees of freedom.

Confidence Intervals for the Mean Value of y:

$$D = \frac{1}{n} + \frac{(x - \bar{x})^2}{SS_{xx}}$$

$$s_{\hat{y}} = s\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{SS_{xx}}} = s \cdot \sqrt{D}$$

Confidence for mean value of y at given value x:

$$\hat{y} \pm t_{\alpha/2,\nu} \cdot s_{\hat{y}}$$

$\nu = n - 2$

Prediction for mean values of y:

$$s_{pred} = s \cdot \sqrt{1 + D}$$

$$\hat{y} \pm t_{\alpha/2,\nu} \cdot s_{pred}$$

$$SSTotal = \sum_{i=1}^{n}(y_i - \bar{y_i})^2$$

$$SSR = \sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

SSTotal = SSR + SSE

$$r^2 = \frac{SS_{xy}^2}{SS_{xx} \cdot SS_{yy}} = \frac{SSR}{SSTotal}$$

$r = \sqrt{r^2}$ when $b_1 > 0$ or $r = -\sqrt{r^2}$ when $b_1 \leq 0$

C7S4: Checking Assumptions and Transforming Data

Residual Plots:

Assumption 1: At any given value of x, the mean of potential errors is 0. Even number above and below zero line and don't follow a trend.

Assumption 2: The vaiance of potential errors is always the same, no matter what the value of x is. Spread neither increasing nor decreasing.

Assumption 3: At any given value of x, the distribution of potential errors is normal. Dont see deviance from line on a Normal Q-Q plot.

$p(i) = P(Z \leq z_{(i)} = \frac{3i-1}{3n-1}$

Assumption 4: The error terms are independent of each other. Shouldn't see a trend in residuals with time.

If assumptions are invalid we can transform data (square root, ln, power, etc.)