Tables and Formulas

for Sullivan, *Fundamentals of Statistics, 2e.*

CHAPTER 2 Organizing and Summarizing Data

- Relative frequency $= \dfrac{\text{frequency}}{\text{sum of all frequencies}}$
- Class midpoint: The sum of consecutive lower class limits divided by 2.

CHAPTER 3 Numerically Summarizing Data

- Population Mean: $\mu = \dfrac{\sum x_i}{N}$
- Sample Mean: $\bar{x} = \dfrac{\sum x_i}{n}$
- Range = Largest Data Value − Smallest Data Value
- Population Variance: $\sigma^2 = \dfrac{\sum(x_i - \mu)^2}{N} = \dfrac{\sum x_i^2 - \dfrac{(\sum x_i)^2}{N}}{N}$
- Sample Variance: $s^2 = \dfrac{\sum(x_i - \bar{x})^2}{n-1} = \dfrac{\sum x_i^2 - \dfrac{(\sum x_i)^2}{n}}{n-1}$
- Population Standard Deviation $\sigma = \sqrt{\sigma^2}$
- Sample Standard Deviation: $s = \sqrt{s^2}$
- ***Empirical Rule:*** If the shape of the distribution is bell-shaped, then
  - Approximately 68% of the data lie within 1 standard deviation of the mean
  - Approximately 95% of the data lie within 2 standard deviations of the mean
  - Approximately 99.7% of the data lie within 3 standard deviations of the mean
- ***Chebyshev's Inequality:*** For any data set, regardless of the shape of the distribution, at least $\left(1 - \dfrac{1}{k^2}\right)100\%$ of the observations will lie within k standard deviations of the mean where k is any number greater than 1.
- Population Mean from Grouped Data: $\mu = \dfrac{\sum x_i f_i}{\sum f_i}$
- Sample Mean from Grouped Data: $\bar{x} = \dfrac{\sum x_i f_i}{\sum f_i}$
- Weighted Mean: $\bar{x}_w = \dfrac{\sum w_i x_i}{\sum w_i}$
- Population Variance from Grouped Data

$$\sigma^2 = \frac{\sum(x_i - \mu)^2 f_i}{\sum f_i} = \frac{\sum x_i^2 f_i - \dfrac{(\sum x_i f_i)^2}{\sum f_i}}{\sum f_i}$$

- Sample Variance from Grouped Data:

$$s^2 = \frac{\sum(x_i - \mu)^2 f_i}{\left(\sum f_i\right) - 1} = \frac{\sum x_i^2 f_i - \dfrac{(\sum x_i f_i)^2}{\sum f_i}}{\sum f_i - 1}$$

- Population Z-score: $z = \dfrac{x - \mu}{\sigma}$
- Sample Z-score: $z = \dfrac{x - \bar{x}}{s}$
- Percentile of $x = \dfrac{\text{Number of data values less than } x}{n} \cdot 100$
- Determining the kth percentile: $i = \left(\dfrac{k}{100}\right)(n + 1)$. If i is not an integer, find the mean of the observations on either side of i.
- Interquartile Range: $\text{IQR} = Q_3 - Q_1$
- Lower and Upper Fences: Lower fence $= Q_1 - 1.5(\text{IQR})$; Upper fence $= Q_3 + 1.5(\text{IQR})$
- Five-Number Summary

  Minimum, Q_1, M, Q_3, Maximum

Tables and Formulas

for Sullivan, *Fundamentals of Statistics*, 2e.

CHAPTER 4 Describing the Relation between Two Variables

- Correlation Coefficient: $r = \dfrac{\sum\left(\dfrac{x_i - \bar{x}}{s_x}\right)\left(\dfrac{y_i - \bar{y}}{s_y}\right)}{n - 1}$
- The equation of the least-squares regression line is $\hat{y} = b_1 x + b_0$, where $\hat{y}$ is the predicted value, $b_1 = r \cdot \dfrac{s_y}{s_x}$ is the slope, and $b_0 = \bar{y} - b_1\bar{x}$ is the intercept.
- Residual = observed y − predicted $y = y - \hat{y}$
- Coefficient of Determination: R^2 = the percent of total variation in the response variable that is explained by the least-squares regression line.
- $R^2 = r^2$ for the least-squares regression model $\hat{y} = b_1 x + b_0$

CHAPTER 5 Probability

- Empirical Probability

$$P(E) \approx \frac{\text{frequency of } E}{\text{number of trials of experiment}}$$

- Classical Probability

$$P(E) = \frac{\text{number of ways that } E \text{ can occur}}{\text{number of possible outcomes}} = \frac{N(E)}{N(S)}$$

- Addition Rule for Disjoint Events

$$P(E \text{ or } F) = P(E) + P(F)$$

- Addition Rule for n Disjoint Events

$$P(E \text{ or } F \text{ or } G \text{ or} \cdots) = P(E) + P(F) + P(G) + \cdots$$

- General Addition Rule

$$P(E \text{ or } F) = P(E) + P(F) - P(E \text{ and } F)$$

- Complement Rule

$$P(E^c) = 1 - P(E)$$

- Multiplication Rule for Independent Events

$$P(E \text{ and } F) = P(E) \cdot P(F)$$

- Multiplication Rule for n Independent Events

$$P(E \text{ and } F \text{ and } G \cdots) = P(E) \cdot P(F) \cdot P(G) \cdot \ldots$$

- Conditional Probability Rule

$$P(F|E) = \frac{P(E \text{ and } F)}{P(E)} = \frac{N(E \text{ and } F)}{N(E)}$$

- General Multiplication Rule

$$P(E \text{ and } F) = P(E) \cdot P(F|E)$$

- Factorial

$$n! = n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot 3 \cdot 2 \cdot 1$$

- Permutation of n objects taken r at a time: ${}_nP_r = \dfrac{n!}{(n-r)!}$
- Combination of n objects taken r at a time: ${}_nC_r = \dfrac{n!}{r!(n-r)!}$
- Permutations with Repetition: n_1 of one type, n_2 of a second type, …, with $n_1 + n_2 + \cdots + n_k = n$

$$\frac{n!}{n_1! \cdot n_2! \cdot \cdots \cdot n_k!}$$

Tables and Formulas

for Sullivan, *Fundamentals of Statistics*, 2e.

Chapter 6 Discrete Probability Distributions

- Mean (Expected Value) of a Discrete Random Variable

$$\mu_X = \Sigma x \cdot P(X = x)$$

- Variance of a Discrete Random Variable

$$\sigma_X^2 = \Sigma(x - \mu)^2 \cdot P(x) = \Sigma x^2 P(x) - \mu x^2$$

- Binomial Probability Distribution Function

$$P(x) = {}_nC_x p^x (1 - p)^{n-x}$$

- Mean of a Binomial Random Variable

$$\mu_X = np$$

- Standard Deviation of a Binomial Random Variable

$$\sigma_X = \sqrt{np(1 - p)}$$

Chapter 7 The Normal Distribution

- Standardizing a Normal Random Variable

$$Z = \frac{X - \mu}{\sigma}$$

- Finding the Score:

$$X = \mu + Z\sigma$$

Chapter 8 Sampling Distributions

- Mean and Standard Deviation of the Sampling Distribution of $\bar{x}$

$$\mu_{\bar{x}} = \mu \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- Sample Proportion:

$$\hat{p} = \frac{x}{n}$$

- Mean and Standard Deviation of the Sampling Distributon of $\hat{p}$

$$\mu_{\hat{p}} = p \text{ and } \sigma_{\hat{p}} = \sqrt{\frac{p(1 - p)}{n}}$$

- Standardizing a Normal Random Variable

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1 - p)}{n}}}$$

Chapter 9 Estimating the Value of a Parameter Using Confidence Intervals

Confidence Intervals

- A $(1 - \alpha) \cdot 100\%$ confidence interval about μ with σ known is $\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$, provided the population from which the sample was drawn is normal or the sample size is large ($n \geq 30$).
- A $(1 - \alpha) \cdot 100\%$ confidence interval about μ with σ unknown is $\bar{x} \pm t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$, provided the population from which the sample was drawn is normal or the sample size is large ($n \geq 30$). *Note*: $t_{\frac{\alpha}{2}}$ is computed using $n - 1$ degrees of freedom.
- A $(1 - \alpha) \cdot 100\%$ confidence interval about p is $\hat{p} \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$, provided $n\hat{p}(1 - \hat{p}) \geq 10$.

Sample Size

- To estimate the population mean with a margin of error E at a $(1 - \alpha) \cdot 100\%$ level of confidence requires a sample of size $n = \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{E}\right)^2$ rounded up to the next integer.
- To estimate the population proportion with a margin of error E at a $(1 - \alpha) \cdot 100\%$ level of confidence requires a sample of size $n = \hat{p}(1 - \hat{p})\left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2$ rounded up to the next integer, where $\hat{p}$ is a prior estimate of the population proportion.
- To estimate the population proportion with a margin of error E at a $(1 - \alpha) \cdot 100\%$ level of confidence requires a sample of size $n = 0.25\left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2$ rounded up to the next integer when no prior estimate of p is available.

Tables and Formulas

for Sullivan, *Fundamentals of Statistics*, 2e.

CHAPTER 10 Hypothesis Tests Regarding a Parameter

Test Statistics

- $z_0 = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$, provided that the population from which the sample was drawn is normal or the sample size is large $(n \geq 30)$.
- $t_0 = \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}}$ follows Student's t-distribution with $n - 1$ degrees of freedom, provided that the population from which the sample was drawn is normal or the sample size is large $(n \geq 30)$.
- $z_0 = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}}$, provided that $np_0(1 - p_0) \geq 10$ and the sample size is less than 5% of the population size $(n < 0.05N)$.

CHAPTER 11 Inferences on Two Samples

- Test Statistic for Matched-Pairs data

$$t_0 = \frac{\bar{d}}{s_d / \sqrt{n}}$$

where $\bar{d}$ is the mean and s_d is the standard deviation of the differenced data.

- Confidence Interval for Matched-Pairs data:

$$\text{Lower bound: } \bar{d} - t_{\frac{\alpha}{2}} \cdot \frac{s_d}{\sqrt{n}}$$

$$\text{Upper bound: } \bar{d} + t_{\frac{\alpha}{2}} \cdot \frac{s_d}{\sqrt{n}}$$

Note: $t_{\frac{\alpha}{2}}$ is found using $n - 1$ degrees of freedom.

- Test Statistic Comparing Two Means (Independent Sampling):

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

- Confidence Interval for the Difference of Two Means (Independent Samples):

$$\text{Lower bound: } (\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\text{Upper bound: } (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Note: $t_{\frac{\alpha}{2}}$ is found using the smaller of $n_1 - 1$ or $n_2 - 1$ degrees of freedom.

- Test Statistic Comparing Two Population Proportions

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

where $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$.

- Confidence Interval for the Difference of Two Proportions

$$\text{Lower bound: } (\hat{p}_1 - \hat{p}_2) - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

$$\text{Upper bound: } (\hat{p}_1 - \hat{p}_2) + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

Tables and Formulas

for Sullivan, *Fundamentals of Statistics, 2e.*

Chapter 12 Additional Inferential Procedures

- Expected Counts (when testing for goodness of fit)

$$E_i = \mu_i = np_i \quad \text{for} \quad i = 1, 2, \ldots, k$$

- Expected Frequencies (when testing for independence or homogeneity of proportions)

$$\text{Expected frequency} = \frac{(\text{row total})(\text{column total})}{\text{table total}}$$

- Chi-Square Test Statistic

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum \frac{(O_i - E_i)^2}{E_i}$$

$i = 1, 2, \ldots, k$

(1) All expected frequencies are greater than or equal to 1 and (2) no more than 20% of the expected frequencies are less than 5.

Use $k - 1$ degrees of freedom for goodness of fit.

Use $(r - 1)(c - 1)$ degrees of freedom when testing for independence or homogeneity of proportions (r is the number of rows, c is the number of columns).

- Standard Error of the Estimate

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}} = \sqrt{\frac{\sum \text{residuals}^2}{n - 2}}$$

- Standard error of b_1

$$s_{b_1} = \frac{s_e}{\sqrt{\sum (x_i - \overline{x})^2}}$$

- Test statistic for the Slope of the Least-Squares Regression Line

$$t_0 = \frac{b_1 - \beta_1}{s_e \big/ \sqrt{\sum (x_i - \overline{x})^2}} = \frac{b_1 - \beta_1}{s_{b_1}}$$

- Confidence Interval for the Slope of the Regression Line
A $(1 - \alpha) \cdot 100\%$ confidence interval for the slope of the true regression line, β_1, is given by:

$$\text{Lower bound:} \quad b_1 - t_{\frac{\alpha}{2}} \cdot \frac{s_e}{\sqrt{\sum (x_i - \overline{x})^2}}$$

$$\text{Upper bound:} \quad b_1 + t_{\frac{\alpha}{2}} \cdot \frac{s_e}{\sqrt{\sum (x_i - \overline{x})^2}}$$

where $t_{\frac{\alpha}{2}}$ is computed with $n - 2$ degrees of freedom.

- Confidence Interval about the Mean Response of y, $\hat{y}$
A $(1 - \alpha) \cdot 100\%$ confidence interval for the mean response of y, $\hat{y}$, is given by

$$\text{Lower bound:} \quad \hat{y} - t_{\frac{\alpha}{2}} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum (x_i - \overline{x})^2}}$$

$$\text{Upper bound:} \quad \hat{y} + t_{\frac{\alpha}{2}} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum (x_i - \overline{x})^2}}$$

where x^* is the given value of the explanatory variable and $t_{\frac{\alpha}{2}}$ is the critical value with $n - 2$ degrees of freedom.

- Prediction Interval about an Individual Response, $\hat{y}$
A $(1 - \alpha) \cdot 100\%$ prediction interval for the individual response of y, $\hat{y}$, is given by

$$\text{Lower Bound:} \quad \hat{y} - t_{\frac{\alpha}{2}} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum (x_i - \overline{x})^2}}$$

$$\text{Upper Bound:} \quad \hat{y} + t_{\frac{\alpha}{2}} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum (x_i - \overline{x})^2}}$$

where x^* is the given value of the explanatory variable and $t_{\frac{\alpha}{2}}$ is the critical value with $n - 2$ degrees of freedom.