

Formulas involving data

- Mean:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum x_i$$

- Standard Deviation:

$$s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

- Standardized value, for x from a distribution with mean μ , s.d. σ :

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

- Correlation:

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

- Regression formulas:

– Fitted values: $\hat{y} = a + bx$

– Coefficients

$$b = r \frac{s_y}{s_x}, \quad a = \bar{y} - b\bar{x}$$

– Residuals (observed - fit):

$$r_i = y_i - \hat{y}_i$$

Probability Rules

- For any event A , $0 \leq P(A) \leq 1$
- Sample space S : $P(S) = 1$
- Addition rule

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

When A , B disjoint events, note that $P(A \text{ and } B) = 0$.

- Complementation: $P(A^c) = 1 - P(A)$
- Conditional probability

$$P(B | A) = \frac{P(A \text{ and } B)}{P(A)} \quad \text{or} \quad P(A \text{ and } B) = P(A)P(B | A)$$

Note that $P(B | A) = P(B)$ when A , B are independent events

Distributions

- Sampling dist. for mean of sample of size n from population with mean μ , s.d. σ

$$\bar{X} \sim \text{Norm}(\mu, \sigma / \sqrt{n})$$

- Binomial: For $X \sim \text{Binom}(n, p)$,

$$\mu = np, \quad \sigma = \sqrt{np(1-p)}$$

When $np \geq 10$, $n(1-p) \geq 10$, X has approx. dist. $\text{Norm}(np, \sqrt{np(1-p)})$

Inference Procedures

- Level C Confidence Intervals (general):

(estimate) \pm (critical value)(std. error)

- 1-sample t : test statistic when $\mathbf{H}_0: \mu = \mu_0$

$$t = \frac{\bar{x} - \mu_0}{\text{SE}}, \quad \text{SE} = \frac{s}{\sqrt{n}}, \quad df = n - 1$$

- 2-sample t : test statistic when $\mathbf{H}_0: \mu_1 = \mu_2$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\text{SE}}, \quad \text{SE} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \quad df = \min(n_1 - 1, n_2 - 1)$$

- 1-sample proportion:

– CIs for p use $\text{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

– Hyp. tests: test statistic when $\mathbf{H}_0: p = p_0$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

- 2-sample proportion:

– Confidence intervals for $p_1 - p_2$ use

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

– Hyp. tests: test statistic when $\mathbf{H}_0: p_1 = p_2$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\text{SE}}, \quad \text{SE} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)},$$

where \hat{p} is the “pooled sample proportion” (see p. 521)

Inference Procedures (cont.)

- Chi-square

- For two-way tables:

Under H_0 : "no relationship between variables",

$$\text{cell's expected count} = \frac{(\text{row total})(\text{column total})}{\text{table total}}$$

$$\text{Test statistic: } \chi^2 = \sum \frac{[(\text{observed count}) - (\text{expected count})]^2}{\text{expected count}}, \quad df = (\#rows - 1)(\#cols - 1)$$

- Goodness of Fit:

Under H_0 : $p_1 = p_{10}, p_2 = p_{20}, \dots, p_k = p_{k0}$,

$$\chi^2 = \sum \frac{[(\text{count of outcome } i) - np_{i0}]^2}{np_{i0}}, \quad df = k - 1$$

- Regression with p predictor variables ($p = 1$ for simple linear regression)

- Model utility test:

test statistic: F (from ANOVA table), $df_{\text{numer}} = p, df_{\text{denom}} = n - p - 1$

- Hypothesis tests for parameter values, with $H_0: \beta_i = 0$:

$$\text{test statistic: } t = \frac{b_i}{SE_{b_i}}, \quad df = n - p - 1$$

- 1-way ANOVA

For I groups/populations, $H_0: \mu_1 = \mu_2 = \dots = \mu_I$

test statistic: F (from ANOVA table), $df_{\text{numer}} = I - 1, df_{\text{denom}} = n - I$