Descriptive Statistics

$$\bar{X} = \frac{\sum x}{n} \qquad \bar{X} = \frac{\sum[f \cdot x]}{\sum f} \text{ (grouped)} \qquad \mu = \frac{\sum x}{N}$$

$$s = \sqrt{\frac{\sum(x-\bar{X})^2}{n-1}} \qquad \sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

$$Z = \frac{X-\bar{X}}{s} \qquad Z = \frac{X-\mu}{\sigma} \qquad CV = \frac{s}{\bar{X}} \times 100\% \qquad CV = \frac{\sigma}{\mu} \times 100\%$$

Outliers

Upper Fence $= Q_3 + 1.5 \times IQR$

Lower Fence $= Q_1 - 1.5 \times IQR$

Empirical Rule

If a variable X follows a bell-shaped distribution, then:

- 68% of data values of X are within one standard deviation of the mean
- 95% of data values of X are within two standard deviations of the mean
- 99.7% of data values of X are within three standard deviations of the mean

Chebyshev's Theorem

For any random variable X, the percentage of values lying within k standard deviations of the mean is at least:

$$\left(1 - \frac{1}{k^2}\right) \times 100\%$$

Probability Rules

$P(A \cup B) = P(A) + P(B)$ if events A and B are mutually exclusive

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ for any events A and B

$P(A \cap B) = P(A) \cdot P(B)$ if events A and B are independent

$P(A \cap B) = P(A) \cdot P(B \mid A)$ for any events A and B

$P(\bar{A}) = 1 - P(A)$ for any event A

Permutations and Combinations

$$P(n,r) = \frac{n!}{(n-r)!} \qquad C(n,r) = \frac{P(n,r)}{r!} = \frac{n!}{(n-r)!r!}$$

Bayes' Rule

If events $A_1, A_2, \ldots, A_k$ are mutually exclusive and exhaustive, then for any event B:

$$P(B) = \sum_{j=1}^{k} P(A_j) \cdot P(B \mid A_j)$$

$$P(A_i \mid B) = \frac{P(A_i) \cdot P(B \mid A_i)}{\sum_{j=1}^{k} P(A_j) \cdot P(B \mid A_j)} \quad \text{(where } A_i \text{ is any one of the events } A_1, A_2, \ldots, A_k\text{)}$$

Discrete Probability Distributions

For any discrete random variable X,

$$\mu = \sum[x \cdot P(x)] \qquad \sigma^2 = \sum[(x-\mu)^2 \cdot P(x)] \quad \text{or} \quad \sigma^2 = \sum[x^2 \cdot P(x)] - \mu^2$$

Binomial Distribution

$$P(x) = C(n,x)p^x q^{n-x} \qquad \mu = np \qquad \sigma^2 = npq$$

Geometric Distribution

$$P(x) = q^{x-1} \cdot p \qquad \mu = \frac{1}{p} \qquad \sigma^2 = \frac{1-p}{p^2}$$

Hypergeometric Distribution

$$P(x) = \frac{C(K,\ x) \cdot C(N-K,\ n-x)}{C(N,\ n)} = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} \qquad \mu = n \cdot \frac{K}{N} \qquad \sigma^2 = n \cdot \frac{K}{N} \cdot \frac{N-K}{N} \cdot \frac{N-n}{N-1}$$

Poisson Distribution

$$P(x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!} \qquad \mu = \lambda \qquad \sigma^2 = \lambda$$

Continuous Probability Distributions

General

$$\mu = \int_{-\infty}^{+\infty} x \cdot f(x)\,dx \qquad \sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot f(x)\,dx = \int_{-\infty}^{+\infty} x^2 \cdot f(x)\,dx - \mu^2$$

Uniform

$$f(x) = \frac{1}{b-a},\ a \le x \le b \qquad F(x) = \frac{x-a}{b-a},\ x \ge a \qquad \mu = \frac{a+b}{2} \qquad \sigma^2 = \frac{1}{12}(b-a)^2$$

Exponential

$$f(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}},\ x \ge 0 \qquad F(x) = 1 - e^{-\frac{x}{\beta}} \qquad \mu = \beta \qquad \sigma^2 = \beta^2$$

Normal

$$f(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad Z = \frac{X-\mu}{\sigma}$$

Sampling Distributions and Central Limit Theorem

If X is a *numerical* random variable with mean μ and standard deviation σ, then for the sampling distribution of $\bar{X}$ for samples of size n:

- $\mu_{\bar{X}} = \mu$
- $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
- $\bar{X}$ is normally distributed as long as:
    - X is normally distributed, *or*
    - $n > 30$

If p is the proportion of a population that satisfies a certain *categorical* condition, then for the sampling distribution of $\hat{p}$ for samples of size n:

- $\mu_{\hat{p}} = p$
- $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$
- $\hat{p}$ is normally distributed as long as:
    - $np \ge 5$, and
    - $nq \ge 5$

Confidence Intervals (with conditions)

$\bar{X} - E < \mu < \bar{X} + E$ $\quad E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ $\quad \sigma$ is known, X normal or $n > 30$

$E = t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$ $\quad \sigma$ unknown, X normal or $n > 30$

$\text{df } = n - 1$

$\hat{p} - E < p < \hat{p} + E$ $\quad E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}$ $\quad n\hat{p} \geq 5, n\hat{q} \geq 5$

$(\bar{X}_2 - \bar{X}_1) - E < \mu_2 - \mu_1 < (\bar{X}_2 - \bar{X}_1) + E$

$E = z_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n^2}}$ $\quad$ independent samples, $n_1 > 30, n_2 > 30$

$E = t_{\frac{\alpha}{2}} \cdot S_p\sqrt{\frac{1}{n_1} + \frac{1}{n^2}}$ $\quad$ independent samples, X_1 and X_2 normal,

$\sigma_1^2 = \sigma_2^2,\ \text{df} = n_1 + n_2 - 2$

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

$\bar{d} - E < \mu_d < \bar{d} + E$ $\quad E = t_{\alpha/2} \cdot \frac{s_d}{\sqrt{n}}$ $\quad n$ dependent pairs, $\text{df} = n - 1$

$(\hat{p}_2 - \hat{p}_1) - E < p_2 - p_1 < (\hat{p}_2 - \hat{p}_1) + E$ $\quad n_1\hat{p}_1 \geq 5, n_1\hat{q}_1 \geq 5, n_2\hat{p} \geq 5, n_2\hat{q} \geq 5$

$E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$

Sample Sizes

$n = \left[z_{\alpha/2} \cdot \frac{\sigma}{E}\right]^2$ $\quad$ sample size for conf int for μ

$n = \frac{1}{4}\left[z_{\alpha/2} \cdot \frac{1}{E}\right]^2$ $\quad$ sample size for conf int for p

$n = pq\left[z_{\alpha/2} \cdot \frac{1}{E}\right]^2$ $\quad$ sample size for conf int for p (p and q known)

Hypothesis Testing – Test Statistics

$$z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$ one mean (σ known)

$$t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$$ one mean (σ unknown)

$$z = \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}}$$ one proportion

$$z = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$ two means, independent samples, σ^2 known

$$t = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{S_p^2}{n_1}+\frac{S_p^2}{n_2}}}$$ two means, independent samples, X_1 and X_2 normal, $\sigma_1^2 = \sigma_2^2$,

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

$$t = \frac{\bar{d}-\mu_d}{s_d/\sqrt{n}}$$ matched pairs (dependent samples), $\mathrm{df} = n - 1$

Conclusions for Hypothesis Testing

If the original claim was H_0 and we *reject* H_0:

“There is sufficient evidence to reject the claim that …”

If the original claim was H_0 and we *fail to reject* H_0:

“There is insufficient evidence to reject the claim that …”

If the original claim was H_1 and we *reject* H_0:

“There is sufficient evidence to accept the claim that …”

If the original claim was H_1 and we *fail to reject* H_0:

“There is insufficient evidence to accept the claim that …”

Regression and Correlation

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \cdot \sqrt{n(\sum y^2) - (\sum y)^2}} \qquad t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \text{ (test statistic for } \rho, \text{df} = \text{n} - 2)$$

$$\hat{y} = a + bx$$

$$b = r \cdot \frac{s_Y}{s_X} = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$a = \bar{Y} - b\bar{X} = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

Prediction Intervals

$$S_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}} = \sqrt{\frac{(\sum y^2) - a(\sum y) - b(\sum xy)}{n-2}}$$

$$\hat{y} - E < y < \hat{y} + E \qquad E = t_{\alpha/2} \cdot S_e\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{(n-1)s_X^2}} \qquad \text{df} = n - 2$$

Confidence Intervals for Regression Coefficients $\hat{Y} = a + bX$

$$a - E < \alpha < a + E \qquad E = t_{\alpha/2} \cdot S_e\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_X^2}} \qquad \text{df} = n - 2$$

$$b - E < \beta < b + E \qquad E = t_{\alpha/2} \cdot S_e\frac{1}{\sqrt{(n-1)s_X^2}} \qquad \text{df} = n - 2$$