

Summary of Formulas

• **Some Properties of Probability.**

1. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

2. $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

• **Descriptive Statistics.** Let $\{x_1, x_2, \dots, x_n\}$ be a sample of size n , then

1. $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

2. $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{\sum x^2 - \frac{1}{n}(\sum x)^2}{n-1} = \frac{SS_{xx}}{n-1}$.

• **Discrete Random Variable.**

1. $\mu_X = E(X) = \sum_{i=1}^k x_i p_i = x_1 p_1 + x_2 p_2 + x_3 p_3 + \dots + x_k p_k$.

2. $\sigma_X^2 = Var(X) = \sum_{i=1}^k (x_i - \mu)^2 p_i = (x_1 - \mu)^2 p_1 + (x_2 - \mu)^2 p_2 + \dots + (x_k - \mu)^2 p_k$.

• **Binomial Distribution.** $X \sim \text{bin}(n, p) \Rightarrow P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$, for $x = 0, 1, \dots, n$.

1. $E(X) = np$

2. $V(X) = np(1-p)$.

• **(Continuous) Uniform Distribution.** $X \sim \text{unif}[c, d] \Rightarrow f(x) = \frac{1}{d-c}$, for $c \leq x \leq d$.

1. $E(X) = \frac{1}{2}(c+d)$

2. $V(X) = \frac{1}{12}(d-c)^2$.

• If $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$.

1. The $(1-\alpha)100\%$ confidence interval for μ is $[\bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$.

2. For a specified margin of error M , the required sample size is $n = \left(\frac{Z_{\frac{\alpha}{2}} \sigma}{M} \right)^2$.

3. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$, reject the null if $Z_{\text{obs}} = \frac{\bar{X}_{\text{obs}} - \mu}{\sigma/\sqrt{n}} > Z_{\alpha}$, or if the $p\text{-value} = P(Z \geq Z_{\text{obs}})$ is $\leq \alpha$.

4. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu < \mu_0$, reject the null if $Z_{\text{obs}} < -Z_{\alpha}$, or if the $p\text{-value} = P(Z \leq Z_{\text{obs}})$ is $\leq \alpha$.

5. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$, reject the null if $|Z_{\text{obs}}| > Z_{\frac{\alpha}{2}}$, or if the $p\text{-value} = 2 * P(Z \geq |Z_{\text{obs}}|)$ is $\leq \alpha$.

• If $t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{(n-1)}$

1. The $(1-\alpha)100\%$ confidence interval for μ is $[\bar{X} - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}]$.

2. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$, reject the null if $t_{\text{obs}} = \frac{\bar{X}_{\text{obs}} - \mu}{s/\sqrt{n}} > t_{\alpha}$.

3. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu < \mu_0$, reject the null if $t_{\text{obs}} < -t_{\alpha}$.

4. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$, reject the null if $|t_{\text{obs}}| > t_{\frac{\alpha}{2}}$.

• If $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{(n-1)}^2$.

1. To test $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 > \sigma_0^2$, reject the null if $X_{\text{obs}}^2 = \frac{(n-1)S_{\text{obs}}^2}{\sigma_0^2} > \chi_{\alpha}^2$.

2. To test $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 < \sigma_0^2$, reject the null if $X_{\text{obs}}^2 = \frac{(n-1)S_{\text{obs}}^2}{\sigma_0^2} < \chi_{1-\alpha}^2$.

3. To test $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 \neq \sigma_0^2$, reject the null if $X_{\text{obs}}^2 > \chi_{\frac{\alpha}{2}}^2$ or if $X_{\text{obs}}^2 < \chi_{1-\frac{\alpha}{2}}^2$.

- If $Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0, 1)$.

1. The $(1 - \alpha)100\%$ confidence interval for p is $[\hat{p} - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$.
2. For a specified margin of error M , the required sample size is $n = \frac{Z_{\alpha/2}^2 [p(1-p)]}{M^2} \leq \frac{Z_{\alpha/2}^2 (0.25)}{M^2}$.
3. To test $H_0 : p = p_0$ vs. $H_1 : p > p_0$, reject the null if $Z_{\text{obs}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > Z_{\alpha}$.
4. To test $H_0 : p = p_0$ vs. $H_1 : p < p_0$, reject the null if $Z_{\text{obs}} < -Z_{\alpha}$.
5. To test $H_0 : p = p_0$ vs. $H_1 : p \neq p_0$, reject the null if $|Z_{\text{obs}}| > Z_{\frac{\alpha}{2}}$.

- If $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$.

1. The $(1 - \alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is $(\bar{X}_1 - \bar{X}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.
2. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 > d_0$, reject the null if $Z_{\text{obs}} > Z_{\alpha}$.
3. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 < d_0$, reject the null if $Z_{\text{obs}} < -Z_{\alpha}$.
4. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 \neq d_0$, reject the null if $|Z_{\text{obs}}| > Z_{\frac{\alpha}{2}}$.

- If $t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx t_{(k)}$, where $k \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \left(\frac{s_2^2}{n_2}\right)^2}$.

1. The $(1 - \alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is $(\bar{X}_1 - \bar{X}_2) \pm t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.
2. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 > d_0$, reject the null if $t_{\text{obs}} > t_{\alpha}$.
3. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 < d_0$, reject the null if $t_{\text{obs}} < -t_{\alpha}$.
4. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 \neq d_0$, reject the null if $|t_{\text{obs}}| > t_{\frac{\alpha}{2}}$.

- If $t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$, where $s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$.

1. The $(1 - \alpha)100\%$ C. I. for $\mu_1 - \mu_2$ is $(\bar{X}_1 - \bar{X}_2) \pm t_{\frac{\alpha}{2}} * s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$.
2. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 > d_0$, reject the null if $t_{\text{obs}} > t_{\alpha}$.
3. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 < d_0$, reject the null if $t_{\text{obs}} < -t_{\alpha}$.
4. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 \neq d_0$, reject the null if $|t_{\text{obs}}| > t_{\frac{\alpha}{2}}$.

- If $Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \approx N(0, 1)$.

1. The $(1 - \alpha)100\%$ confidence interval for $p_1 - p_2$ is $\left[(\hat{p}_1 - \hat{p}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right]$.
2. To test $H_0 : p_1 - p_2 = 0$ vs. $H_1 : p_1 - p_2 > 0$, reject the null if $Z_{\text{obs}} = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} > Z_{\alpha}$,

where, $\hat{p} = \frac{Y_1 + Y_2}{n_1 + n_2}$, the pooled sample proportion.

3. To test $H_0 : p_1 - p_2 = 0$ vs. $H_1 : p_1 - p_2 < 0$, reject the null if $Z_{\text{obs}} < -Z_{\alpha}$.
4. To test $H_0 : p_1 - p_2 = 0$ vs. $H_1 : p_1 - p_2 \neq 0$, reject the null if $|Z_{\text{obs}}| > Z_{\frac{\alpha}{2}}$.

- **Chi-Square Statistic.** The *chi-square statistic* is a measure of how much the observed cell counts diverge from the expected cell counts. The formula for the statistic is

1. Goodness-of-fit Test.

$$X^2 = \sum_{i=1}^k \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} \quad (1)$$

$$= \sum_{i=1}^k \frac{(n_i - n * p_i)^2}{n * p_i} \sim \chi_{k-1}^2. \quad (2)$$

This X^2 statistic follows **approximately** the χ^2 distribution with $k - 1$ degrees of freedom.

2. Testing equality of several proportions, independence, and homogeneity.

$$X^2 = \sum_{\text{all cells}} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} \quad (3)$$

$$= \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - r_i * \hat{p}_j)^2}{r_i * \hat{p}_j} \quad (4)$$

$$= \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - r_i * c_j / n)^2}{r_i * c_j / n} \sim \chi_{(r-1)(c-1)}^2. \quad (5)$$

where, r_i and c_j are the total of the i th row and j th column, respectively. This X^2 statistic follows **approximately** the χ^2 distribution with $(r - 1)(c - 1)$ degrees of freedom.

• ANOVA Table

Source	DF	Sum of Squares	Mean Square	F
Groups	$k - 1$	$SSG = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2$	$MSG = \frac{SSG}{k - 1}$	$F_{\text{obs}} = \frac{MSG}{MSE}$
Error	$n - k$	$SSE = \sum_{i=1}^k (n_i - 1) s_i^2$	$MSE = \frac{SSE}{n - k}$	
Total	$n - 1$	$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$		

Note: $SST = SSG + SSE$. Under H_0 , $F_{\text{obs}} \sim f_{(k-1, n-k)}$.

- **Equation of the Least-Squares Regression Line .** Suppose we have data on an explanatory variable x and a response variable y for n individuals. The means and standard deviations of the sample data are $\bar{x}$ and s_x for x and $\bar{y}$ and s_y for y , and the correlation between x and y is r . The equation of the least-squares regression line of y on x is

$$\hat{y} = \hat{a} + \hat{b}x$$

with *slope*

$$\hat{b} = \frac{SS_{xy}}{SS_{xx}} = \frac{(\sum xy) - \frac{1}{n}(\sum x)(\sum y)}{(\sum x^2) - \frac{1}{n}(\sum x)^2} = r \frac{s_y}{s_x} \quad (6)$$

and *intercept*

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \quad (7)$$

- The **fitted** (or **predicted**) values $\hat{y}_i$'s are obtained by successively substituting the x_i 's into the estimated regression line: $\hat{y} = \hat{a} + \hat{b}x_i$. The **residuals** are the vertical deviations, $e_i = y_i - \hat{y}_i$, from the estimated line.
- The **error sum of squares**, (equivalently, **residual sum of squares**) denoted by SSE , is

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum [y_i - (\hat{a} + \hat{b}x_i)]^2 \quad (8)$$

$$= \sum y_i^2 - \hat{a} \sum y_i - \hat{b} \sum x_i y_i \quad (9)$$

and the estimate of σ^2 is

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n - 2} = \frac{(n - 1)s_y^2(1 - r^2)}{n - 2}. \quad (10)$$

- **Linear Correlation.** The *linear correlation coefficient* r measures the strength of the linear relationship between the paired x - and y -quantitative values in a sample.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (11)$$

$$= \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} \quad (12)$$

where,

$$SS_{xx} = \sum x^2 - \frac{1}{n}(\sum x)^2 = (n-1)s_x^2 \quad (13)$$

$$SS_{yy} = \sum y^2 - \frac{1}{n}(\sum y)^2 = (n-1)s_y^2 \quad (14)$$

$$SS_{xy} = \sum xy - \frac{1}{n}(\sum x)(\sum y) \quad (15)$$

- The **coefficient of determination**, denoted by r^2 , is the amount of the variation in y that is explained by the regression line.

$$r^2 = 1 - \frac{SSE}{SST}, \quad \text{where, } SST = SS_{yy} = \sum(y_i - \bar{y})^2 \quad (16)$$

$$= \frac{SST - SSE}{SST} = \frac{\text{explained variation}}{\text{total variation}} \quad (17)$$

- **Inference for b .**

1. Test statistic:

$$\frac{\hat{b} - b}{SE_{\hat{b}}} \sim t_{(n-2)} \quad SE_{\hat{b}} = \frac{s}{s_x \sqrt{n-1}} = \frac{\hat{b} \sqrt{1-r^2}}{r \sqrt{n-2}}$$

2. Confidence Interval: $\hat{b} \pm t_{\alpha/2} SE_{\hat{b}}$

- **Mean Response of Y at a specified value x^* , $(\mu_{Y|x^*})$.**

1. **Point Estimate.** For a specific value x^* , the estimate of the **mean** value of Y is given by

$$\hat{\mu}_{Y|x^*} = \hat{a} + \hat{b}x^*$$

2. **Confidence Interval.** For a specific value x^* , the $(1-\alpha)100\%$ confidence interval for $\mu_{Y|x^*}$ is given by

$$\hat{\mu}_{Y|x^*} \pm t_{\alpha/2; (n-2)} SE_{\hat{\mu}}$$

$$\text{where, } SE_{\hat{\mu}} = s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}, \text{ and } s = s_y \sqrt{\frac{(n-1)(1-r^2)}{n-2}}$$

- **Prediction of Y at a specified value x^* .**

1. **Point Estimate.** For a specific value x^* , the predicted value of Y is given by

$$\hat{y} = \hat{a} + \hat{b}x^*$$

2. **Prediction Interval.** For a specific value x^* , the $(1-\alpha)100\%$ prediction interval is given by

$$\hat{y} \pm t_{\alpha/2; (n-2)} SE_{\hat{y}}$$

$$\text{where, } SE_{\hat{y}} = s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}, \text{ and } s = s_y \sqrt{\frac{(n-1)(1-r^2)}{n-2}}$$