Statistics 350 Help Card

Summary Measures

Sample Mean

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

Sample Standard Deviation

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

Probability Rules

- **Complement rule**

  $P(A^C) = 1 - P(A)$
- **Addition rule**

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

  For independent events:

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

  For mutually exclusive events: $P(A \text{ or } B) = P(A) + P(B)$
- **Multiplication rule**

  General: $P(A \text{ and } B) = P(A)P(B \mid A)$

  For independent events: $P(A \text{ and } B) = P(A)P(B)$

  For mutually exclusive events: $P(A \text{ and } B) = 0$
- **Conditional Probability**

  General: $P(A \mid B) = \dfrac{P(A \text{ and } B)}{P(B)}$

  For independent events: $P(A \mid B) = P(A)$

  For mutually exclusive events: $P(A \mid B) = 0$

Discrete Random Variables

Mean

$$E(X) = \mu = \sum x_i p_i = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k$$

Standard Deviation

$$s.d.(X) = \sigma = \sqrt{\sum (x_i - \mu)^2 p_i} = \sqrt{\sum \left(x_i^2 p_i\right) - \mu^2}$$

Binomial Random Variables

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where $\dbinom{n}{k} = \dfrac{n!}{k!(n-k)!}$

Mean

$$E(X) = \mu_X = np$$

Standard Deviation

$$s.d.(X) = \sigma_X = \sqrt{np(1-p)}$$

Normal Random Variables

- $z - \text{score} = \dfrac{\text{observation} - \text{mean}}{\text{standard deviation}} = \dfrac{x - \mu}{\sigma}$
- Percentile: $x = z\sigma + \mu$
- If X has the $N(\mu, \sigma)$ distribution, then the variable $Z = \dfrac{X - \mu}{\sigma}$ has the $N(0,1)$ distribution.

Normal Approximation to the Binomial Distribution

If X has the *B(n, p)* distribution and the sample size n is large enough (namely $np \geq 10$ and $n(1-p) \geq 10$), then X is approximately $N\left(np, \sqrt{np(1-p)}\right)$.

Sample Proportions

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

Mean

$$E(\hat{p}) = \mu_{\hat{p}} = p$$

Standard Deviation

$$\text{s.d.}(\hat{p}) = \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Sampling Distribution of $\hat{p}$

If the sample size n is large enough (namely, $np \geq 10$ and $n(1-p) \geq 10$)

then $\hat{p}$ is *approximately* $N\left(p, \sqrt{\dfrac{p(1-p)}{n}}\right)$.

Sample Means

Mean

$$E(\overline{X}) = \mu_{\overline{X}} = \mu$$

Standard Deviation

$$s.d.(\overline{X}) = \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

Sampling Distribution of $\overline{X}$

If X has the $N(\mu, \sigma)$ distribution, then $\overline{X}$ is

$$N\left(\mu_{\overline{X}}, \sigma_{\overline{X}}\right) \Leftrightarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

If X follows *any* distribution with mean μ and standard deviation σ and n is large,

then $\overline{X}$ is *approximately* $N\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$.

This last result is **Central Limit Theorem**

.

Population Proportion

Parameter p

Statistic $\hat{p}$

Standard Error

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Confidence Interval

$$\hat{p} \pm z^* \text{s.e.}(\hat{p})$$

Conservative Confidence Interval

$$\hat{p} \pm \frac{z^*}{2\sqrt{n}}$$

Large-Sample *z*-Test

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

Sample Size

$$n = \left(\frac{z^*}{2m}\right)^2$$

Two Population Proportions

Parameter $p_1 - p_2$

Statistic $\hat{p}_1 - \hat{p}_2$

Standard Error

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

Confidence Interval

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \text{s.e.}(\hat{p}_1 - \hat{p}_2)$$

Large-Sample *z*-Test

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

where $\hat{p} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$

Population Mean

Parameter μ

Statistic $\bar{x}$

Standard Error

$$\text{s.e.}(\bar{x}) = \frac{s}{\sqrt{n}}$$

Confidence Interval

$\bar{x} \pm t^* \text{s.e.}(\bar{x})$ df $= n - 1$

Paired Confidence Interval

$\bar{d} \pm t^* \text{s.e.}(\bar{d})$ df $= n - 1$

One-Sample *t*-Test

$$t = \frac{\bar{x} - \mu_0}{\text{s.e.}(\bar{x})} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \qquad \text{df} = n - 1$$

Paired *t*-Test

$$t = \frac{\bar{d} - 0}{\text{s.e.}(\bar{d})} = \frac{\bar{d}}{s_d/\sqrt{n}} \qquad \text{df} = n - 1$$

Two Population Means

General

Parameter $\mu_1 - \mu_2$

Statistic $\bar{x}_1 - \bar{x}_2$

Standard Error

$$\text{s.e.}(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Confidence Interval

$(\bar{x}_1 - \bar{x}_2) \pm t^*(\text{s.e.}(\bar{x}_1 - \bar{x}_2))$ df $= \min(n_1 - 1, n_2 - 1)$

Two-Sample *t*-Test

$$t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\text{s.e.}(\bar{x}_1 - \bar{x}_2)} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad \text{df} = \min(n_1 - 1, n_2 - 1)$$

Pooled

Parameter $\mu_1 - \mu_2$

Statistic $\bar{x}_1 - \bar{x}_2$

Standard Error

$$\text{pooled s.e.}(\bar{x}_1 - \bar{x}_2) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

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

Confidence Interval

$(\bar{x}_1 - \bar{x}_2) \pm t^*(\text{pooled s.e.}(\bar{x}_1 - \bar{x}_2))$ df $= n_1 + n_2 - 2$

Pooled Two-Sample *t*-Test

$$t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\text{pooled s.e.}(\bar{x}_1 - \bar{x}_2)} = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \text{df} = n_1 + n_2 - 2$$

One-Way ANOVA

$$\text{SS Groups} = \text{SSG} = \sum_{\text{groups}} n_i(\bar{x}_i - \bar{x})^2$$

$$\text{MS Groups} = \text{MSG} = \frac{\text{SSG}}{k - 1}$$

$$\text{SS Error} = \text{SSE} = \sum_{\text{groups}} (n_i - 1)s_i^2$$

$$\text{MS Error} = \text{MSE} = s_p^2 = \frac{\text{SSE}}{N - k}$$

$$\text{SS Total} = \text{SSTO} = \sum_{\text{values}} (x_{ij} - \bar{x})^2$$

$$F = \frac{\text{MS Groups}}{\text{MS Error}}$$

ANOVA Table

Source	SS	DF	MS	*F*
Groups	SS Groups	$k - 1$	MS Groups	F
Error	SS Error	$N - k$	MS Error	
Total	SSTO	$N - 1$		

Confidence Interval $\bar{x}_i \pm t^* \dfrac{s_p}{\sqrt{n_i}}$ df $= N - k$

Under H_0, the F statistic follows an $F(k - 1, N - k)$ distribution.

Regression

Linear Regression Model

Population Version:

Mean: $\mu_Y(x) = E(Y) = \beta_0 + \beta_1 x$

Individual: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

where ε_i is $N(0, \sigma)$

Sample Version:

Mean: $\hat{y} = b_0 + b_1 x$

Individual: $y_i = b_0 + b_1 x_i + e_i$

Standard Error of the Sample Slope

$$\text{s.e.}(b_1) = \frac{s}{\sqrt{S_{XX}}} = \frac{s}{\sqrt{\sum (x - \bar{x})^2}}$$

Confidence Interval for β_1

$b_1 \pm t^* \text{s.e.}(b_1)$ df $= n - 2$

t-Test for β_1

To test $H_0 : \beta_1 = 0$

$$t = \frac{b_1 - 0}{\text{s.e.}(b_1)}$$ df $= n - 2$

or $F = \dfrac{MSREG}{MSE}$ df $= 1, n - 2$

Parameter Estimators

$$b_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{\sum (x - \bar{x}) y}{\sum (x - \bar{x})^2}$$

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

Confidence Interval for the Mean Response

$\hat{y} \pm t^* \text{s.e.(fit)}$ df $= n - 2$

where $\text{s.e.(fit)} = s\sqrt{\dfrac{1}{n} + \dfrac{(x - \bar{x})^2}{S_{XX}}}$

Residuals

$e = y - \hat{y} =$ observed y – predicted y

Prediction Interval for an Individual Response

$\hat{y} \pm t^* \text{s.e.(pred)}$ df $= n - 2$

where $\text{s.e.(pred)} = \sqrt{s^2 + (\text{s.e.(fit)})^2}$

Correlation and its square

$$r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}$$

$$r^2 = \frac{SSTO - SSE}{SSTO} = \frac{SSREG}{SSTO}$$

where $SSTO = S_{YY} = \sum (y - \bar{y})^2$

Estimate of σ

$s = \sqrt{MSE} = \sqrt{\dfrac{SSE}{n - 2}}$ where $SSE = \sum (y - \hat{y})^2 = \sum e^2$

Standard Error of the Sample Intercept

$$\text{s.e.}(b_0) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}}$$

Confidence Interval for β_0

$b_0 \pm t^* \text{s.e.}(b_0)$ df $= n - 2$

t-Test for β_0

To test $H_0 : \beta_0 = 0$

$$t = \frac{b_0 - 0}{\text{s.e.}(b_0)}$$ df $= n - 2$

Chi-Square Tests

Test of Independence & Test of Homogeneity

Expected Count

$$E = \text{expected} = \frac{\text{row total} \times \text{column total}}{\text{total } n}$$

Test Statistic

$$\mathrm{X}^2 = \sum \frac{(O - E)^2}{E} = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

df $= (r - 1)(c - 1)$

Test for Goodness of Fit

Expected Count

$E_i = \text{expected} = np_{i0}$

Test Statistic

$$\mathrm{X}^2 = \sum \frac{(O - E)^2}{E} = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

df $= k - 1$

If Y follows a $\chi^2(df)$ distribution, then E(Y) = df and Var(Y) = 2(df).