# EXAM CHEAT SHEET (PART II)

## I. HYPOTHESIS TESTING

Null Hypothesis ($H_0$): Always involves "=" sign.

Alternative Hypothesis ($H_1$): "<", ">" or "≠".

One-Tailed Test: $H_1$ involves "<" or ">".

Two-Tailed Test: $H_1$ involves "≠"".

Rejection Region: Reject $H_0$ if falling in this range.

$p$-value: Reject $H_0$ if $p$-value is very *small*.

Type I Error ($\alpha$): Reject $H_0$ when it is true.

Type II Error ($\beta$): Fail to reject $H_0$ when it is false.

Significance Level: $\alpha = P(\text{Type I Error})$

Testing $\mu$ ($\sigma^2$ known): $Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

Testing $\mu$ ($\sigma^2$ unknown): $T = \dfrac{\bar{X} - \mu_0}{s/\sqrt{n}}$ (Df. $= n-1$)

Testing Population Proportion $p$: $Z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

## II. COMPARING TWO POPULATIONS

Paired Samples: $T = \dfrac{\bar{X}_D - \mu_D}{s_D/\sqrt{n}}$

Independent Samples:

Testing $\mu_1 - \mu_2$ ($\sigma_1^2$, $\sigma_2^2$ known):

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Testing $\mu_1 - \mu_2$ ($\sigma_1^2$, $\sigma_2^2$ unknown):

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (\text{Df.} = n_1 + n_2 - 2)$$

*Pooled Sample Variance*:

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

Testing $\sigma_1^2 = \sigma_2^2$: $F = \dfrac{s_1^2}{s_2^2}$ ($s_1^2 > s_2^2$)

RR.: $F > F_{\frac{\alpha}{2}, n_1-1, n_2-1}$

Testing $p_1 - p_2 = D_0$: $Z = \dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$

Testing $p_1 - p_2 = 0$: $Z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

*Combined Proportion*: $\hat{p} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$

## III. ANALYSIS OF VARIANCE (ANOVA)

### A. One-Way ANOVA

Hypotheses: $\mu_1 = \mu_2 = \cdots = \mu_k$ ($k$ levels)

ANOVA Table:

| | Sum Sq. | Df. | Mean Sq. |
|---|---|---|---|
| Factor | $SST = \sum_j^k n_j(\bar{Y}_j - \bar{Y})^2$ | $k-1$ | $MST = \frac{SST}{k-1}$ |
| Error | $SSE = \sum_j^k \sum_i^{n_j} (Y_{ij} - \bar{Y}_j)^2$ | $n-k$ | $MSE = \frac{SSE}{n-k}$ |
| Total | $SS_{\text{Total}} = \sum_j^k \sum_i^{n_j} (Y_{ij} - \bar{Y})^2$ | $n-1$ | |

Test Statistic: $F = \frac{MST}{MSE}$ (RR.: $F > F_{\alpha,k-1,n-k}$)

### B. Two-Way ANOVA

Main Effects Hypotheses: Population means at different levels of *a factor* are all equal.

Interaction Hypotheses: There is no *interaction* between factors.

Interaction: Two factors *interact* when the effect of one factor on the response variable is altered by the level of the other factor.

Complete Factorial Experiment: There are data collected for all treatments.

Balanced Experiment: Number of observations for each treatment (also called *replicates*) are the same.

ANOVA Table:

| | Sum Sq. | Df. | Mean Sq. |
|---|---|---|---|
| Factor A | $SS_A$ | $a-1$ | $MS_A = \frac{SS_A}{a-1}$ |
| Factor B | $SS_B$ | $b-1$ | $MS_B = \frac{SS_B}{b-1}$ |
| Interaction | $SS_{AB}$ | $(a-1)(b-1)$ | $MS_{AB} = \frac{SS_{AB}}{(a-1)(b-1)}$ |
| Error | $SSE$ | $n-ab$ | $MSE = \frac{SSE}{n-ab}$ |
| Total | $SS(\text{Total})$ | $n-1$ | |

Test Statistics:

$F_A = \frac{MS_A}{MSE}$ (RR.: $F_A > F_{\alpha,a-1,n-ab}$)

$F_B = \frac{MS_B}{MSE}$ (RR.: $F_B > F_{\alpha,b-1,n-ab}$)

$F_{AB} = \frac{MS_{AB}}{MSE}$ (RR.: $F_{AB} > F_{\alpha,(a-1)(b-1),n-ab}$)

## IV. CHI-SQUARED TEST

Goodness-of-Fit Test:

$\chi^2 = \sum_i^k \dfrac{(f_i - e_i)^2}{e_i}$ (RR.: $\chi^2 > \chi^2_{\alpha,k-1}$)

Test of Contingency Table:

$\chi^2 = \sum_i^r \sum_j^c \dfrac{(f_{ij} - e_{ij})^2}{e_{ij}}$ (RR.: $\chi^2 > \chi^2_{\alpha,(r-1)(c-1)}$)

$$e_{ij} = \frac{i^{\text{th}} \text{ row total} \times j^{\text{th}} \text{ column total}}{\text{sample size}}$$

## V. LINEAR REGRESSION

### A. Simple Linear Regression

Regression Model: $Y = \beta_0 + \beta_1 X + \epsilon$

Model Assumption: $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2_\epsilon)$

Errors $\epsilon_i$ are assumed to be: (i) normally distributed; (ii) with mean equal to 0; (iii) with constant variance $\sigma^2_\epsilon$ regardless of $X$; (iv) independent with each other.

Fitted Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

Parameter Estimation:

$\hat{\beta}_1 = \dfrac{s_{XY}}{s^2_X}$, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

Residual: $e_i = Y_i - \hat{Y}_i$

ANOVA Table for Regression:

| | Sum Sq. | Df. | Mean Sq. |
|---|---|---|---|
| Regression | $SSR = \sum_i^n (\hat{Y}_i - \bar{Y})^2$ | 1 | $MSR = SSR$ |
| Residual | $SSE = \sum_i^n (Y_i - \hat{Y}_i)^2$ | $n-2$ | $MSE = \frac{SSE}{n-2}$ |
| Total | $SS_{\text{Total}} = \sum_i^n (Y_i - \bar{Y})^2$ | $n-1$ | |

Testing Overall Significance: $\beta_1 = 0$ or $\rho = 0$

Testing $\beta_1 = c$: $T = \dfrac{\hat{\beta}_1 - c}{s_{\hat{\beta}_1}}$ (Df. $= n-2$)

$$s_{\hat{\beta}_1} = \sqrt{\frac{\frac{1}{n-2}\sum_i^n e_i^2}{(n-1)s^2_X}} = \frac{s_\epsilon}{\sqrt{(n-1)s^2_X}}$$

Testing $\beta_0 = c$: $T = \dfrac{\hat{\beta}_0 - c}{s_{\hat{\beta}_0}}$ (Df. $= n-2$)

$$s_{\hat{\beta}_0} = s_{\hat{\beta}_1} \times \sqrt{\frac{\sum_i^n X_i^2}{n}}$$

Testing $\rho = 0$: $T = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ (Df. $= n-2$)

Coefficient of Determination:

$$R^2 = \frac{s^2_{XY}}{s^2_X s^2_Y} = \frac{SSR}{SS(\text{Total})}$$

Estimating $\sigma_\epsilon$: $s_\epsilon = \sqrt{\dfrac{\sum_i^n e_i^2}{n-2}} = \sqrt{\dfrac{SSE}{n-2}}$

Point Estimation: $\hat{y}_g = \hat{\beta}_0 + \hat{\beta}_1 x_g$

Prediction Interval for *expected* (or *particular*) value:

$$\hat{y}_g \pm t_{\frac{\alpha}{2},n-2} \times s_\epsilon \sqrt{\frac{1}{n} + \frac{(x_g - \bar{X})^2}{(n-1)s^2_X} \bigg( + 1 \bigg)}$$

### B. Multiple Linear Regression

Regression Model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$

Fitted Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k$

ANOVA Table for Regression:

| | Sum Sq. | Df. | Mean Sq. |
|---|---|---|---|
| Regression | $SSR = \sum_i^n (\hat{Y}_i - \bar{Y})^2$ | $k$ | $MSR = \frac{SSR}{k}$ |
| Residual | $SSE = \sum_i^n (Y_i - \hat{Y}_i)^2$ | $n-k-1$ | $MSE = \frac{SSE}{n-k-1}$ |
| Total | $SS_{\text{Total}} = \sum_i^n (Y_i - \bar{Y})^2$ | $n-1$ | |

Testing Overall Significance: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$

Test Statistic: $F = \frac{MSR}{MSE}$ (RR.: $F > F_{\alpha,k,n-k-1}$)

Testing Individual Coefficient Parameters: $\beta_j = c$

Interpretation: All other independent variables being considered, $X_j$ has a significant linear relationship with $Y$.

Test Statistic: $T = \dfrac{\hat{\beta}_j - c}{s_{\hat{\beta}_j}}$ (Df. $= n-k-1$)

Estimating $\sigma_\epsilon$: $s_\epsilon = \sqrt{\dfrac{\sum_i^n e_i^2}{n-k-1}}$

Multiple $R^2$: $R^2 = \frac{SSR}{SS(\text{Total})}$

Adjusted $R^2$: $R^2_{adj} = 1 - \dfrac{(n-1)(1-R^2)}{n-k-1} = 1 - \dfrac{\frac{SSE}{n-k-1}}{\frac{SS(\text{Total})}{n-1}}$

Multicollinearity: Independent variables are correlated with each other.

Interaction Model: $Y = \beta_0 + \beta_1 X + \beta_2 Y + \beta_3 (X \times Y) + \epsilon$