Grant MacEwan University

STAT 151 – Formula Sheet – Final Exam

Dr. Karen Buro

Descriptive Statistics

- Sample Variance: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\Sigma_{i=1}^{n} x_i)^2}{n}}{n-1}$
- Sample Standard Deviation: $s = \sqrt{\text{Sample Variance}} = \sqrt{s^2}$
- Median: Order the data from smallest to largest. The median M is either the unique middle value or the mean of the two middle values.
- Lower Quartile: Order the data from smallest to largest. The lower quartile Q_1 is the median of the smaller half of the values.
- Upper Quartile: Order the data from smallest to largest. The upper quartile Q_3 is the median of the upper half of the values.
- Interquartile Range (iqr) = Upper Quartile – Lower Quartile $=Q_3 - Q_1$
- Outliers: lower fence $=Q_1 - 1.5iqr$ and upper fence$=Q_3 + 1.5iqr$

Probability Theory

- Addition Rule: $P(A \text{ or } B) = P(A) + P(B) - P(A\&B)$
- Complement Rule: $P(\text{A does not occur}) = P(\text{not } A) = 1 - P(A)$
- Multiplication Rule: $P(A\&B) = P(A|B)P(B)$
- Multiplication Rule for **Independent** Events:
  If A and B are **independent**, then $P(A\&B) = P(A)P(B)$
- Conditional Probability of A given B, if $P(B) > 0:\ P(A|B) = \frac{P(A\&B)}{P(B)}$

Population Distributions

- The mean (expected value) of a discrete random variable: $\mu = \sum xp(x)$.
- The variance of a discrete random variable: $\sigma^2 = \sum(x-\mu)^2 p(x)$
- The standard deviation of a discrete random variable: $\sigma = \sqrt{\sigma^2}$

Binomial Distribution

- Probability to observe k successes in n trials: $p(k) = P(x = k) = \binom{n}{k} p^k (1-p)^{n-k}$
- $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
- Mean and standard deviation of a binomial distribution: $\mu = np$ and $\sigma = \sqrt{np(1-p)}$

Sampling Distributions

- Sampling Distribution of a Sample Mean, $\bar{x}$:

$$\mu_{\bar{x}} = \mu, \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- Sampling Distribution of the difference of two Sample Means, $\bar{x}_1 - \bar{x}_2$:

$$\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2, \quad \sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- Sampling Distribution of a Sample Proportion, $\hat{p}$:

$$\mu_{\hat{p}} = p, \quad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

- Sampling Distribution of the difference of two Sample Proportions, $\hat{p}_1 - \hat{p}_2$:

$$\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2, \quad \sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Estimation

Parameter	Estimator	SE(Estimator)	Approximate Confidence Interval
μ	$\bar{x}$	$\frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$
p	$\hat{p}$	$\sqrt{\frac{p(1-p)}{n}}$	$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
$\mu_1 - \mu_2$	$\bar{x_d}$	$\frac{\sigma_d}{\sqrt{n}}$	$\bar{x_d} \pm t_{\alpha/2}\frac{s_d}{\sqrt{n}}$
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$	$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Choosing the Sample Size (formulas)

- Estimate a mean μ with a $(1-\alpha)$ confidence interval within an amount of m.

$$n \geq \left(\frac{z_{(1-\alpha/2)}\sigma}{m}\right)^2$$

- Estimate a proportion p with a $(1-\alpha)$ confidence interval within an amount of m.

$$n \geq \left(\frac{z_{(1-\alpha/2)}}{m}\right)^2 p(1-p)$$

Test Statistics

- Test Statistic for large sample z-Test concerning p

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

- Test Statistic for Large-Sample z Test for comparing p_1 and p_2:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n_1} + \frac{\hat{p}_c(1-\hat{p}_c)}{n_2}}}, \text{ with } \hat{p}_c = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

- Test Statistic for 1-sample t-Test concerning μ if σ is unknown

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \qquad df = n - 1$$

- Test Statistic for two–sample t-Test for comparing two population means:

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

- Test Statistic for paired t-Test for comparing two population means:

$$t = \frac{\bar{x}_d}{(s_d/\sqrt{n})} \qquad df = n - 1$$

Goodness-of-Fit Tests and Test for Independence of two categorical variables

- Test Statistic for Goodness–of–Fit Test:

$$\chi^2 = \sum_{\text{all categories}} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

Expected cell count $= n \cdot$(hypothesized value of corresponding population proportion)
$df = k - 1$ where k is the number of categories.

- χ^2 Test for Independence: The test statistic is

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

Expected cell count $= \frac{\text{(row total)(column total)}}{\text{grand total}}$

df =(number of rows - 1)(number of columns - 1)

Regression Analysis

- Sum of Squares

$$SS_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}, \quad SS_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}, \quad SS_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

- Correlation Coefficient (r), Coefficient of Determination (R^2)

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}, \quad R^2 = r^2$$

- Least Squares Regression line $\hat{y} = b_0 + b_1 x$, with

$$b_1 = \frac{SS_{xy}}{SS_{xx}} \text{ , and } b_0 = \bar{y} - b_1\bar{x}$$

- Estimation of σ

$$s_e = \sqrt{\frac{SSE}{n-2}}, \text{ with } SSE = \sum(\hat{y}_i - y_i)^2 = SS_{yy} - b_1 SS_{xy}$$

- Confidence interval for β_1

$$b_1 \pm t^* \frac{s_e}{\sqrt{SS_{xx}}}$$

- test statistic for a test about β_1

$$t_0 = \frac{b_1}{s_e/\sqrt{SS_{xx}}}, \quad df = n-2$$

- Confidence interval for the mean of y, $E(y)$, at $x = x_p$

$$\hat{y} \pm t^* \ s_e \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

- Prediction Interval for y at $x = x_p$

$$\hat{y} \pm t^* \ s_e \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

The Analysis of Variance (ANOVA)

- Total Sum of Squares

$$SST = \sum_{ij}(x_{ij} - \bar{x})^2 = \sum_{ij} x_{ij}^2 - CM \quad (df = n - 1)$$

with $\bar{x}$ = sample mean of all measurements, $G = \sum_{ij} x_{ij}$ and $CM = \dfrac{G^2}{n}$

- Sum of Squares for groups

$$SSTR = \sum_{i} n_i(\bar{x}_i - \bar{x})^2 = \sum_{i} \frac{T_i^2}{n_i} - CM \quad (df = k - 1)$$

with $\bar{x}_i$ = sample mean of observations in sample i,
T_i = Total of observations in sample i.

- Sum of Squares for Error

$$SSE = \sum_{i}(n_i - 1)s_i^2 = SST - SSG \quad (df = n - k)$$

with s_i is the standard deviation of observation from sample i.

- ANOVA–Table

Source	df	SS	MS=SS/df	F
Treatments/Groups	$k-1$	$SSTR$	$MSTR = SSTR/(k-1)$	$MSTR/MSE$
Error	$n-k$	SSE	$MSE = SSE/(n-k)$	
Total	$n-1$	SST		