Grant MacEwan University
Stat 371 – Categorical Data Analysis
Formula Sheet for Midterm

- **Probability Theory**
  - Binomial distribution, with parameters n and π

$$P(X = x) = \frac{n!}{x!(n-x)!}\ \pi^x(1-\pi)^{n-x}$$

    $\mu = n \cdot \pi, \sigma = \sqrt{n\pi(1-\pi)}$.
  - Multinomial distribution

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_c = x_c) = \frac{n!}{x_1!x_2!, \ldots x_c!}\ \pi_1^{x_1}\pi_2^{x_2}\ldots\pi_c^{n_c}$$

  - Poisson distribution

$$P(X = x) = \frac{e^{-\mu}\mu^x}{x!}$$

    Mean=$\mu, \sigma = \sqrt{\mu}$
  - Conditional Probability

$$P(Y = i|X = j) = \frac{P(Y = i\ and\ X = j)}{P(X = j)}$$

  - Clinical tests Sensitivity = P(Test+|Disease+), Specificity = P(Test-|Disease-)

- **Inferential statistics for one probability, π**

Wald - Test Statistic and Confidence Interval:

$$z_0 = \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}} \quad \hat{\pi} \pm z_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n}$$

Score - Test Statistic:

$$z_0 = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

Likelihood-ratio - Test Statistic:
Let l_0 denote the maximal value of the likelihood function under the null hypothesis, and l_1 is the value of the likelihood function for the ML-estimator, then

$$\chi_0^2 = -2\ln(l_0/l_1), \quad df = 1$$

Score and Likelihood ratio Confidence Intervals
π_0 is in a $(1-\alpha)\times 100\%$ confidence interval if $H_o : \pi = \pi_0$ can not be rejected at significance level of α based on the related test.

- **Inferential statistics for Contingency Tables**

  Independence of two categorical random variables:

$$\pi_{ij} = \pi_{i+}\pi_{+j} \text{ for all } i, j$$

  Confidence Interval for $\pi_1 - \pi_2$.

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

  Test statistic for $\pi_1 - \pi_2$.

$$z_0 = \frac{(\hat{\pi}_1 - \hat{\pi}_2)}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}}$$

  Confidence Interval for $\ln(\pi_1/\pi_2)$.

$$\ln(\hat{\pi}_1/\hat{\pi}_2) \pm z_{\alpha/2}\sqrt{\frac{1-\hat{\pi}_1}{n_1\hat{\pi}_1} + \frac{1-\hat{\pi}_2}{n_2\hat{\pi}_2}}$$

  Confidence Interval for $\ln(\theta)$ (ln(odds ratio)).

$$\ln(\hat{\theta}) \pm z_{\alpha/2}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

  Pearson's χ^2 *statistic for test of independence*

$$\text{Expected cell count for cell } (i,j) = \hat{\mu}_{ij} = \frac{\text{row}_i\text{total} \times \text{column}_j\text{total}}{\text{Grand total}} = \frac{n_{i+}n_{+j}}{n}$$

  then

$$\chi_0^2 = \sum \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}},\ df = (I-1)(J-1)$$

  Likelihood-ratio χ^2 *statistic for test of independence*

$$\chi_0^2 = \sum n_{ij} \ln\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right),\ df = (I-1)(J-1)$$

  Standardized cell residuals

$$z_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1-\hat{\pi}_{i+})(1-\hat{\pi}_{+j})}}$$

- **Linear Association between Ordinal Variables**

  Pearson's Correlation Coefficient

$$r = \frac{\sum(u_i - \bar{u})(v_i - \bar{v})p_{ij}}{\sqrt{[\sum(u_i - \bar{u})^2 p_{i+}]\,[\sum(v_j - \bar{v})^2 p_{+j}]}}$$

  Spearman's Correlation Coefficient is Pearson's Correlation Coefficient for midranks

Test Statistic for testing for a linear association

$$\chi_0^2 = (n-1)r^2, df = 1$$

Kendall's τ

$$\hat{\tau} = 2\frac{n_c - n_d}{n(n-1)},$$

where $n_c(n_d)$ is the number of concordant (discordant) pairs of observations.

- **Link functions**
  - Identity: $g(\mu) = \mu$
  - Logit: $g(\mu) = \ln(\mu/(1-\mu))$
  - Probit: $g(\mu) = \Phi^{-1}(\mu)$
  - Log: $g(\mu) = \ln(\mu)$
- **Simple Logistic Regression**
  - $\ln(\pi/(1-\pi)) = \alpha + \beta x$
  - $\pi/(1-\pi) = e^{\alpha+\beta x}$
  - $\pi = e^{\alpha+\beta x}/(1+e^{\alpha+\beta x})$