MAT 2379, Introduction to Biostatistics

Final Exam Formula Sheet

- Addition Rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Conditional probability of A given B:

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

- Total probability rule: $P(A) = P(A \cap B) + P(A \cap B') = P(A|B)P(B) + P(A|B')P(B')$
- Bayes' rule

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B')P(B')}$$

- Diagnostic tests:

$$\begin{aligned} \text{false-positive-rate} &= P(\text{Test}+|\text{True}-) \\ \text{false-negative-rate} &= P(\text{Test}-|\text{True}+) \\ \text{specificity} &= P(\text{Test}-|\text{True}-) \\ \text{sensitivity} &= P(\text{Test}+|\text{True}+) \\ \text{positive predictive value} &= P(\text{True}+|\text{Test}+) \\ \text{negative predictive value} &= P(\text{True}-|\text{Test}-) \end{aligned}$$

- Events A and B are independent if $P(A \cap B) = P(A)P(B)$
- Expected value of a discrete random variable X:

$$\mu = E(X) = \sum_x x f(x), \quad \text{where} \quad f(x) = P(X = x)$$

- Variance of a discrete random variable X:

$$\sigma^2 = \text{Var}(X) = \sum_x (x-\mu)^2 f(x) = \sum_x x^2 f(x) - \mu^2, \quad \text{where} \quad f(x) = P(X = x)$$

- Cumulative distribution function of a random variable X: $F(x) = P(X \leq x)$
- If X is a binomial random variable with n trials and probability p of success, then

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n$$

- Standardization: If X is a normal random variable with mean μ and variance σ^2, then

$$Z = \frac{X - \mu}{\sigma} \quad \text{has a standard normal distribution}$$

- Sample mean of the observations $x_1, \dots, x_n$: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Sample variance of the observations $x_1, \dots, x_n$:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]$$

- To calculate the median and the quartiles, we arrange the data $x_1, x_2 \dots, x_n$ in ascending order:

$$y_1 \le y_2 \le \dots \le y_n$$

a) The median is given by:

$$\tilde{x} = \begin{cases} y_{(n+1)/2}, & \text{if } n \text{ is odd} \\ \left(y_{n/2} + y_{n/2+1}\right)/2, & \text{if } n \text{ is even.} \end{cases}$$

b) To compute the first quartile, we write

$$\frac{n+1}{4} = r + \frac{a}{4}, \quad r = \text{integer}, a = 0, 1, 2 \text{ or } 3.$$

The first quartile is:

$$q_1 = \begin{cases} y_r, & \text{if } (n+1)/4 = r \\ (0.75)\, y_r + (0.25)\, y_{r+1}, & \text{if } (n+1)/4 = r + 1/4 \\ (0.5)\, y_r + (0.5)\, y_{r+1}, & \text{if } (n+1)/4 = r + 2/4 \\ (0.25)\, y_r + (0.75)\, y_{r+1}, & \text{if } (n+1)/4 = r + 3/4 \end{cases}$$

c) To compute the third quartile, we write

$$\frac{3(n+1)}{4} = r + \frac{a}{4}, \quad r = \text{integer}, a = 0, 1, 2 \text{ or } 3.$$

The third quartile is:

$$q_3 = \begin{cases} y_r, & \text{if } (3(n+1)/4 = r \\ (0.75)\, y_r + (0.25)\, y_{r+1}, & \text{if } 3(n+1)/4 = r + 1/4 \\ (0.5)\, y_r + (0.5)\, y_{r+1}, & \text{if } 3(n+1)/4 = r + 2/4 \\ (0.25)\, y_r + (0.75)\, y_{r+1}, & \text{if } 3(n+1)/4 = r + 3/4 \end{cases}$$

- Statistic used for confidence intervals and tests for the mean μ of a normal population with known variance σ^2:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \text{ has a standard normal distribution}$$

- Statistic used for confidence intervals and tests for the mean μ of a normal population with unknown variance σ^2:

$$T_0 = \frac{\bar{X} - \mu}{S/\sqrt{n}} \text{ has a } T \text{ distribution with } n-1 \text{ d.f.}$$

- Statistic used for confidence intervals for a proportion p:

$$Z = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \text{ has a standard normal distribution}$$

- Statistic used for tests of the hypothesis $H_0 : p = p_0$ for a proportion p:

$$Z_0 = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \text{ has a standard normal distribution}$$

- Statistic used for confidence intervals and tests for the means μ_1 and μ_2 of two independent normal populations with equal variances:

$$T_0 = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S_p^2(1/n_1 + 1/n_2)}} \text{ has a } T \text{ distribution with } n_1 + n_2 - 2 \text{ d.f.}$$

where

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

- Statistic used for confidence intervals and tests for the means μ_1 and μ_2 of two independent normal populations with unequal variances:

$$T_0 = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \text{ has a } T \text{ distribution with } \nu \text{ d.f.}$$

where

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

- The equation of the estimated regression line for points (x_i, y_i) with $i = 1, \ldots, n$ is

$$\hat{y} = \hat{\alpha} + \hat{\beta}x,$$

where

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

and

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{n\sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$