

# P8130 Biostatistics Methods I — Lecture-Based Cheat Sheet

## Basics

- Precision / Accuracy: Low sampling error / Unbiased
- Mode: the most frequently occurring value in the data
- Coefficient of Variation:  $\frac{\sigma}{\mu}$
- Independent Events:  $P(A|B) = P(A)$ ,  $P(A \cap B) = P(A)P(B)$
- Law of Total Probability:  $P(A) = \sum_{i=1}^k P(A \cap B_i) = \sum_{i=1}^k P(B_i)P(A|B_i)$
- Bayes's Theorem:  $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$

## Distributions

- **Bernoulli( $p$ ):**  $P(X = 1) = p$ ;  $E = p$ ;  $\text{Var} = p(1 - p)$
- **Binomial( $n, p$ ):**  $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$ ;  $E = np$ ;  $\text{Var} = np(1 - p)$
- **Poisson( $\lambda$ ):**  $P(X = k) = e^{-\lambda} \lambda^k / k!$ ;  $E = \lambda$ ;  $\text{Var} = \lambda$
- **Geometric( $p$ ):**  $P(X = k) = (1 - p)^{k-1} p$ ,  $k = 1, 2, \dots$ ;  $E = \frac{1}{p}$ ;  $\text{Var} = \frac{1-p}{p^2}$
- **Uniform( $a, b$ ):**  $f(x) = \frac{1}{b-a}$ ,  $a < x < b$ ;  $E = \frac{a+b}{2}$ ;  $\text{Var} = \frac{(b-a)^2}{12}$
- **Normal( $\mu, \sigma^2$ ):**  $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$ ,  $-\infty < x < \infty$ ;  $E = \mu$ ;  $\text{Var} = \sigma^2$
- **Gamma function:**  $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ ,  $x > 0$ ,  $\Gamma(x) = (x-1)!$  when  $x = 1, 2, 3, \dots$
- **Chi-squared( $\nu$ ):**  $f(x) = \frac{1}{2^{\nu/2} \Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}$ ,  $x > 0$ ;  $E = \nu$ ;  $\text{Var} = 2\nu$
- **Student's  $t(\nu)$ :**  $f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi} \Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$ ,  $-\infty < x < \infty$ ;  $E = 0$  ( $\nu > 1$ );  $\text{Var} = \frac{\nu}{\nu-2}$  ( $\nu > 2$ )

## Sampling & CLT

- $E[\bar{X}] = \mu$ ,  $\text{SE}(\bar{X}) = \sigma/\sqrt{n}$ ; for large  $n$ ,  $\bar{X} \approx N(\mu, \sigma^2/n)$
- $\hat{p} = X/n$ :  $E[\hat{p}] = p$ ,  $\text{SE}(\hat{p}) = \sqrt{p(1-p)/n}$

## Hypothesis tests

- **1 mean ( $\sigma$  known):**  $z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \Rightarrow N(0, 1)$ ,  $CI = \bar{X} \pm z_{\alpha/2} \sigma/\sqrt{n}$
- **1 mean ( $\sigma$  unknown):**  $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \Rightarrow t_{n-1}$ ,  $CI = \bar{X} \pm t_{\alpha/2, n-1} s/\sqrt{n}$
- **Paired means:**  $t = \frac{\bar{d} - 0}{s_d/\sqrt{n}} \Rightarrow t_{n-1}$ ,  $D = X_1 - X_2$ ,  $\bar{d} = E(D)$ ,  $s_d = \sqrt{\sum_{i=1}^n (d_i - \bar{d})^2 / (n-1)}$ ,  $CI = \bar{X} \pm t_{\alpha/2, n-1} s_d/\sqrt{n}$
- **2 means:**  $\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$ 
  - **Known equal variance:**  $z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$
  - **Unknown equal variance:**  $s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{\sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2 + \sum_{i=1}^{n_2} (X_{2i} - \bar{X}_2)^2}{n_1+n_2-2}$ ,  $t = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$
  - **Unknown unequal variance:**  $d'' = \text{floor}\left(\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2/(n_1-1) + \left(\frac{s_2^2}{n_2}\right)^2/(n_2-1)}\right)$ ,  $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \sim t_{d''}$
  - **Equality of Variances:**  $F = \frac{s_1^2}{s_2^2} \sim F_{n_1-1, n_2-1}$ , Reject  $H_0$  if  $F > F_{n_1-1, n_2-1, 1-\alpha/2}$  or  $F < F_{n_1-1, n_2-1, \alpha/2}$
- **1 proportion:**  $z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \Rightarrow N(0, 1)$ , **need to check**  $np(1-p) > 5$ 
  - CI of  $p_0 \in \left(\hat{p} - z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$
  - **Exact binomial** (when normal approx fails) if  $\hat{p} \leq p_0$ , use  $2 \sum_{k=0}^x \binom{n}{k} p_0^k (1-p_0)^{n-k}$ ; if  $\hat{p} > p_0$ , use upper tail.
- **2 proportions (pooled under  $H_0$ ):**  $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$ ,  $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$ 
  - CI of  $\hat{p}_1 - \hat{p}_2 \in (\hat{p}_1 - \hat{p}_2 \pm z_{1-\alpha/2} \sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2})$
- **1 variance:**  $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \Rightarrow \chi_{n-1}^2$

- **2 variances:**  $F = \frac{s_1^2}{s_2^2} \Rightarrow F_{n_1-1, n_2-1}$
- **One-way ANOVA ( $k$  groups):** with  $y_{ij} = \mu + \alpha_i + e_{ij}$ ,  $e_{ij} \sim N(0, \sigma^2)$ , define  $SS_B = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2$ ,  $SS_W = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$ ;  $MS_B = SS_B/(k-1)$ ,  $MS_W = SS_W/(n-k)$ ;
  - **Test stat:**  $F = \frac{MS_B}{MS_W} \Rightarrow F_{k-1, N-k}$ ;
  - **Bonferroni adjustment:**  $\alpha^* = \alpha / \binom{k}{2}$

## Chi-square for R×C (Session 12)

- **Expected counts:**  $E_{ij} = \frac{n_i \cdot n_j}{n}$ ;  $df = (r-1)(c-1)$
- **Pearson statistic:**  $\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \Rightarrow \chi_{df}^2$  (Yates correction optional: replace  $|O - E|$  by  $|O - E| - 0.5$  in numerator).

## Power

- **one mean**
  - $H_0 : \mu = \mu_0$  vs.  $H_1 : \mu < \mu_0$ : Power =  $P\left(Z < z_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right)$
  - $H_0 : \mu = \mu_0$  vs.  $H_1 : \mu > \mu_0$ : Power =  $P\left(Z > z_{1-\alpha} + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right)$
  - $H_0 : \mu = \mu_0$  vs.  $H_1 : \mu \neq \mu_0$ : Power =  $P\left(Z < -z_{1-\alpha/2} + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right) + P\left(Z < -z_{1-\alpha/2} + \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right)$
- **two means:**  $H_0 : \mu_1 = \mu_2$  vs.  $H_1 : \mu_1 \neq \mu_2$ : Power =  $P\left(Z < -z_{1-\alpha/2} + \frac{\Delta}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}\right)$
- **two proportions:**  $H_0 : p_1 = p_2$  vs.  $H_1 : p_1 \neq p_2$ : Power  $\approx P\left(Z < \frac{\Delta}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} - z_{1-\alpha/2} \sqrt{\frac{\bar{p}(1-\bar{p})(1/n_1 + 1/n_2)}{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}\right)$ ,  
 $\Delta = |p_1 - p_2|$  and  $\bar{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

## Required Sample Size

- **1 mean:**
  - $H_0 : \mu = \mu_0$  vs.  $H_1 : \mu < \mu_0$ :  $n = \left(\frac{\sigma}{\mu_0 - \mu_1} (z_{1-\beta} - z_\alpha)\right)^2$
  - $H_0 : \mu = \mu_0$  vs.  $H_1 : \mu \neq \mu_0$ :  $n = \left(\frac{\sigma}{\mu_0 - \mu_1} (z_{1-\beta} - z_{1-\alpha/2})\right)^2$
- **2 means:**  $H_0 : \mu_1 = \mu_2$  vs.  $H_1 : \mu_1 \neq \mu_2$ :  $n = \frac{(\sigma_1^2 + \sigma_2^2)(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2}$
- **1 proportion:**  $H_0 : p = p_0$  vs.  $H_1 : p < p_0$ :  $n = \frac{p_0(1-p_0) \left(z_{1-\alpha/2} + z_{1-\beta} \sqrt{\frac{p_1(1-p_1)}{p_0(1-p_0)}}\right)^2}{(p_1 - p_0)^2}$

## Common quantiles

- $z_{0.975} = 1.96$ ;  $t_{0.975, 30} \approx 2.04$