STAT 135 Lecture 13

Henry Liev

23 July 2025

Remark 0.1 (Paired Designs)

$X_1, \ldots, X_n$ iid
$Y_1, \ldots, Y_n$ iid
$\mu_X - \mu_Y$, X_i and Y_i are related

Example 0.2 (Test Prep Course)

Student	Before	After
1	65	80
2	73	50
⋮	⋮	⋮
100		

Solution. $H_0 : \mu_X = \mu_Y$
$H_1 : \mu_X \neq \mu_Y$
$\theta = \mu_Y - \mu_X$
$H_0 : \theta = 0$
$H_1 : \theta \neq 0$
$D_i = Y_i - X_i$
$\bar{D} \pm 1.96 SE(\bar{D})$ ■

Method 0.3 (Paired Designs are Good (11.3))

Let $X_1, \ldots, X_n$ (Baseline) with Variance σ^2
Define $Y_i = X_i + D_i$ D_i random variable with $\mu = \Delta, \sigma^2 = \tau^2$ Observe X_i, Y_i
$D_i = Y_i - X_i$
Estimate of $\hat{\Delta} = \bar{D} = \frac{1}{n} \sum_{i=1}^{n} D_i$
$\bar{D} \pm 1.96 SE(\bar{D})$

Remark 0.4 (Forget that data is paired)

Analyze as if X_i and Y_i are independent.
Still use $\bar{Y} - \bar{X} = \bar{D}$
$Var(\bar{D}) = Var(\bar{X}) + Var(\bar{Y})$
$Var(\bar{D}) = \frac{\sigma^2}{n} + \frac{\sigma^2+\tau^2}{n}$
$SE(\bar{D}) = \sqrt{\frac{2\sigma^2+\tau^2}{n}}$
If we used pairing, would estimate $SE(\sigma) = \frac{\tau}{\sqrt{n}}$
$\sqrt{\frac{\tau^2}{2\sigma^2+\tau^2}}$ much tighter estimate of confidence interval using pairing
Quantifies why paired design is better

Method 0.5 (Chi Squared Goodness of Fit Tests)

Testing does my data follow some prescribed discrete (Multinomial) Distribution

25	150	6	4	19	5
P_1	P_2	P_3	P_4	P_5	P_6

Example 0.6 (5-side die)

H_0 : Die is fair
H_1 : Die is not fair
Roll 100 times

Side	1	2	3	4	5
Count	17	31	16	18	18

Method 0.7 (Pearson's Chi-Square)

$$T = \sum_{k=1}^{n} \frac{(E_k - O_k)^2}{E_k} \qquad \text{E = expected counts, O = observed counts}$$

Measure of discrepancy from expected under H_0
$T \approx \chi^2_{5-1}$ "n-1 degrees of freedom" since last value is determined by the fixed number of observations
Observe $T = t$
$P(T \geq t)$ gives our p-value $T \sim \chi^2_4$

Example 0.8

5 race categories in the US population
$P_1, P_2, \ldots, P_5$
H_0 : Employee race distribution = US race distribution
H_1 : Employee race distribution $\neq$ US race distribution

Category	1	2	3	4	5
Count	100	270	40	55	75

$T = \frac{(500p_1-110)^2}{500p_1} + \cdots + \frac{(500p_5-75)^2}{500p_5}$

Example 0.9 (Goodness of Fit with parametric)

Cells $k = 1, \ldots, m$
$H_0 : P_k = f_k(\theta_1, \theta_2, \ldots, \theta_r)$

Method 0.10 (Goodness of Fit with Parametric Model Playbook)

1. Estimate θ's with MLE
2. $E_k = n\hat{p}_k$, $\hat{p}_k = f(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_r)$
3. $T = \sum_{k=1}^{m} \frac{(E_k - O_k)^2}{E_k}$
4. $T \sim \chi^2_{m-1-r}$

Example 0.11 (Hardy-Weinberg Equilibrium)

M, MN, N
$(1-\theta)^2, 2\theta(1-\theta), \theta^2$

M X_1	MN X_2	N X_3
342	500	187

$\hat{\theta}_{MLE} = \frac{2X_3 + X_2}{2n}$
χ^2_{3-1-1} H_0 : Hardy-Weinberg Equilibrium
H_1 : Does not hold

Method 0.12 (General parameterized multinomial distribution)

Multinomial θ
$X_1, X_2, X_2, \ldots, X_m$ counts of outcome k

$$\Lambda = \frac{\max_{p \in \omega_0} \left(\frac{n!}{x_1! x_2! \cdots x_m!} \right) P_1(\theta)^{x_1} P_2(\theta)^{x_2} \cdots P_m(\theta)^{x_m}}{\max_{p \in \Omega} \left(\frac{n!}{x_1! x_2! \cdots x_m!} \right) P_1(\theta)^{x_1} P_2(\theta)^{x_2} \cdots P_m(\theta)^{x_m}}$$

$$\Lambda = \frac{P_1(\hat{\theta})^{x_1} P_2(\hat{\theta})^{x_2} \cdots P_m(\hat{\theta})^{x_m}}{\hat{p}_1^{x_1} \hat{p}_2^{x_2} \cdots \hat{p}_m^{x_m}} = \prod_{i=1}^{m} \left(\frac{P_i(\hat{\theta})}{\hat{p}_i} \right)^{x_i}$$

$$\hat{p}_k = \frac{X_k}{n}$$

$$-2 \log \Lambda = -2 \sum_{i=1}^{m} x_i \log \left(\frac{P_i(\hat{\theta})}{\hat{p}_i} \right) = -2n \sum_{i=1}^{m} \hat{p}_i \log \left(\frac{P_i(\hat{\theta})}{\hat{p}_i} \right) = 2n \sum_{i=1}^{m} \hat{p}_i \log \left(\frac{\hat{p}_i}{P_i(\hat{\theta})} \right)$$

$$f(x) \approx (x - x_0) + \frac{1}{2}(x - x_1)^2 \cdots$$

$$-2 \log \Lambda = 2n \sum_{i=1}^{m} (\hat{p}_i - P_i(\hat{\theta})) + 2n \frac{1}{2} \sum_{i=1}^{m} \frac{(\hat{p}_i - P_i(\hat{\theta}))^2}{P_i(\theta)} = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$