

DISCUSSION #13

1 Review

- Consider the 2 by 2 table

	D+	D-
E+	a	b
E-	c	d

$\widehat{OR} = \frac{ad}{bc}$, and $100 \times (1 - \alpha)\%$ confidence interval for $\ln(OR)$ is

$$\ln(\widehat{OR}) \pm z_{\alpha/2} \widehat{s.e.}, \text{ where } \widehat{s.e.} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}.$$

- **ANOVA table:** See lecture note problem 6-20.
- **Least square regression.** Some important formula about least square regression, they are:

(i). Sample linear correlation coefficient

$$r = \frac{S_{xy}}{S_x S_y},$$

where

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

$$s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} = \sqrt{SS_{Total}/(n-1)}.$$

(ii). (a). $SS_{Reg} = \sum_i (\hat{y}_i - \bar{y})^2$, where \bar{y} = sample mean of y_i 's.

(b). $SS_{Error} = \sum_i (y_i - \hat{y}_i)^2$.

(c). $SS_{Total} = SS_{Reg} + SS_{Error}$.

(iii). Least regression line (or fitted straight line) is $\hat{y} = b_0 + b_1 x$, where

$$b_1 = \frac{S_{xy}}{S_x^2}, \quad b_0 = \bar{y} - b_1 \bar{x}.$$

(iv). **Interpretation of r .** Notice that $r^2 = \frac{SS_{reg}}{SS_{Total}}$, this ratio indicates the proportion of total response variation that is accounted for by the least square regression model. If r^2 is close to one, then a linear trend in data is suggested, otherwise, nonlinear trend is indicated.

(v). Confidence limits for β_0 and β_1 .

- (a). %95 confidence interval for slope β_1 : $b_1 \pm t_{n-2,0.025} \cdot s_e \frac{1}{\sqrt{S_{xx}}}$, where $s_e = SS_{Error}/(n-2)$ and $S_{xx} = (n-1)s_x^2$.
- (b). %95 confidence interval for slope β_0 : $b_0 \pm t_{n-2,0.025} \cdot s_e \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}$.
- (vi). F-ratio = $\frac{SS_{Reg}/df_{Reg}}{SS_{Error}/df_{Error}}$. $df_{Reg} = k - 1$ is the degrees of freedom of treatment, and $df_{Error} = n - k$ is degrees of freedom of errors.
- (vii). Test $H_0 : \beta_1 = 0$:
 step(1). Find t-score = $\frac{b_1}{s_e} \sqrt{S_{xx}}$ with $df = n - 2$.
 step(2). Use p-value = $2P(T_{n-2} \geq |t|)$ to test H_0 .
- (viii). Test $H_0 : \beta_0 = 0$:
 step(1). Find t-score = $\frac{b_0}{s_e} \sqrt{\frac{nS_{xx}}{S_{xx} + n(\bar{x})^2}}$ with $df = n - 2$.
 step(2). Use p-value = $2P(T_{n-2} \geq |t|)$ to test H_0 .
- (ix). %95 confidence limits for $\mu_{Y|X=x^*}$:

$$(b_0 + b_1 x^*) \pm t_{n-2,0.025} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

2 Problems

- 6-15.** The following data are taken from a study that attempts to determine whether the use of electronic fetal monitoring (“exposure”) during labor affects the frequency of caesarian section deliveries (“disease”). Of the 5824 infants included in the study, 2850 were electronically monitored during labor and 2974 were not. Results are displayed in the 2×2 contingency table below.

		Caesarian Delivery		
		Yes	No	Totals
EFM Exposure	Yes	358	2492	2850
	No	229	2745	2974
Totals		587	5237	5824

- (a) Calculate a **point estimate** for the population odds ratio OR , and *interpret*.
- (b) Compute a **95% confidence interval** for the population odds ratio OR .
- (c) Based on your answer in part (b), show that the null hypothesis $H_0: OR = 1$ can be rejected in favor of the alternative $H_A: OR \neq 1$, at the $\alpha = .05$ significance level. **Interpret this conclusion:** What exactly has been demonstrated about the association between electronic fetal monitoring and caesarian section delivery? Be precise.

6-18. In a random sample of $n = 1200$ consumers who are surveyed about their ice cream flavor preferences, 416 indicate that they prefer vanilla, 419 prefer chocolate, and 365 prefer strawberry.

- (a) Conduct a **Chi-squared “Goodness-of-Fit” Test** of the null hypothesis of equal proportions $H_0: \pi_{\text{Vanilla}} = \pi_{\text{Chocolate}} = \pi_{\text{Strawberry}}$ of flavor preferences, at the $\alpha = .05$ significance level.

	Vanilla	Chocolate	Strawberry	
	416	419	365	1200

- (b) Suppose that the sample of $n = 1200$ consumers is equally divided between males and females, yielding the results shown below. Conduct a **Chi-squared Test** of the null hypothesis that flavor preference is not associated with gender, at the $\alpha = .05$ level.

	Vanilla	Chocolate	Strawberry	Totals
Males	200	190	210	600
Females	216	229	155	600
Totals	416	419	365	1200



Ismor Fischer, 1/7/2011

6-92

- 6-20.** Male patients with coronary artery disease were recruited from three different medical centers – the Johns Hopkins University School of Medicine, The Rancho Los Amigos Medical Center, and the St. Louis University School of Medicine – to investigate the effects of carbon monoxide exposure. One of the baseline characteristics considered in the study was pulmonary lung function, as measured by X = “Forced Expiratory Volume in one second,” or FEV_1 . The data are summarized below.

Johns Hopkins	Rancho Los Amigos	St. Louis
$n_1 = 21$	$n_2 = 16$	$n_3 = 23$
$\bar{x}_1 = 2.63$ liters	$\bar{x}_2 = 3.03$ liters	$\bar{x}_3 = 2.88$ liters
$s_1^2 = 0.246$ liters ²	$s_2^2 = 0.274$ liters ²	$s_3^2 = 0.248$ liters ²

Based on histograms of the raw data (not shown), it is reasonable to assume that the FEV_1 measurements of the three populations from which these samples were obtained are each approximately normally distributed, i.e., $X_1 \sim N(\mu_1, \sigma_1)$, $X_2 \sim N(\mu_2, \sigma_2)$, and $X_3 \sim N(\mu_3, \sigma_3)$. Furthermore, because the three sample variances are so close in value, it is reasonable to assume equivariance of the three populations, that is, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$. With these assumptions, answer the following.

- (a) Compute the pooled estimate of the common variance σ^2 “within groups” via the formula

$$s_{\text{within}}^2 = MS_{\text{Error}} = \frac{SS_{\text{Error}}}{df_{\text{Error}}} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \dots + (n_k - 1)s_k^2}{n - k}.$$

- (b) Compute the grand mean of the $k = 3$ groups via the formula

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + \dots + n_k \bar{x}_k}{n}, \quad \text{where the combined sample size } n = n_1 + n_2 + \dots + n_k.$$

From this, calculate the estimate of the variance “between groups” via the formula

$$s_{\text{between}}^2 = MS_{\text{Treatment}} = \frac{SS_{\text{Treatment}}}{df_{\text{Treatment}}} = \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \dots + n_k(\bar{x}_k - \bar{x})^2}{k - 1}.$$

- (c) Using this information, construct a complete ANOVA table, including the F -statistic, and corresponding p -value, relative to .05 (i.e., $< .05$, $> .05$, or $= .05$). **Infer** whether or not we can reject $H_0: \mu_1 = \mu_2 = \mu_3$, at the $\alpha = .05$ level of significance. **Interpret in context:** Exactly what has been demonstrated about the baseline FEV_1 levels of the three groups?