Analysis of Variance (ANOVA)

Statistical Research Methods I

Seongsoo Choi (최성수)

Estimating how a continuous variable and a categorical variable are associated

- We've learned about how to measure the association between
  - two categorical variables (chi-squared test, odds ratios)
  - two continuous variables (covariance, correlation coefficient)
- Then, how can we gauge the association between a continuous variable and a categorical variable?
  - Comparing groups
    - Comparing means (the two-group t-test)
    - Analysis of variance (ANOVA) & Analysis of Covariance (ANCOVA)

Comparing means across groups

- Suppose there are two groups (defined by a categorical variable, e.g., men and women) and we'd like to compare the means of a continuous variable y (e.g., math scores) between these groups
- The parameter of our interest is the gap in x, $\mu_1 - \mu_2$, and its sample estimator is $\bar{y}_1 - \bar{y}_2$
- The standard error of $\bar{y}_1 - \bar{y}_2$ is

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{(SE_1)^2 + (SE_2)^2} = \sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}$$

- With this estimated SE, we can draw a statistical inference about the mean difference in y between two groups
  - T-test with the p-value or confidence intervals

Example: Heart Surgery Recovery and Prayer

- Outcome of interest (y): recovery from heart surgery with no complications
- Two groups: For Group A, Christian volunteers prayed for a successful surgery with a quick and healthy recovery. Group B did not have volunteers praying for them

TABLE 7.2: Whether Complications Occurred for Heart Surgery Patients Who Did or Did Not Have Group Prayer

	Complications (y)		
Prayer	Yes (0)	No (1)	Total
Yes A	315	289	604
No B	304	293	597

Example: Heart Surgery Recovery and Prayer

TABLE 7.2: Whether Complications Occurred for Heart Surgery Patients Who Did or Did Not Have Group Prayer

	Complications (Y)		
Prayer	Yes (0)	No (1)	Total
Yes A	315	289	604
No B	304	293	597

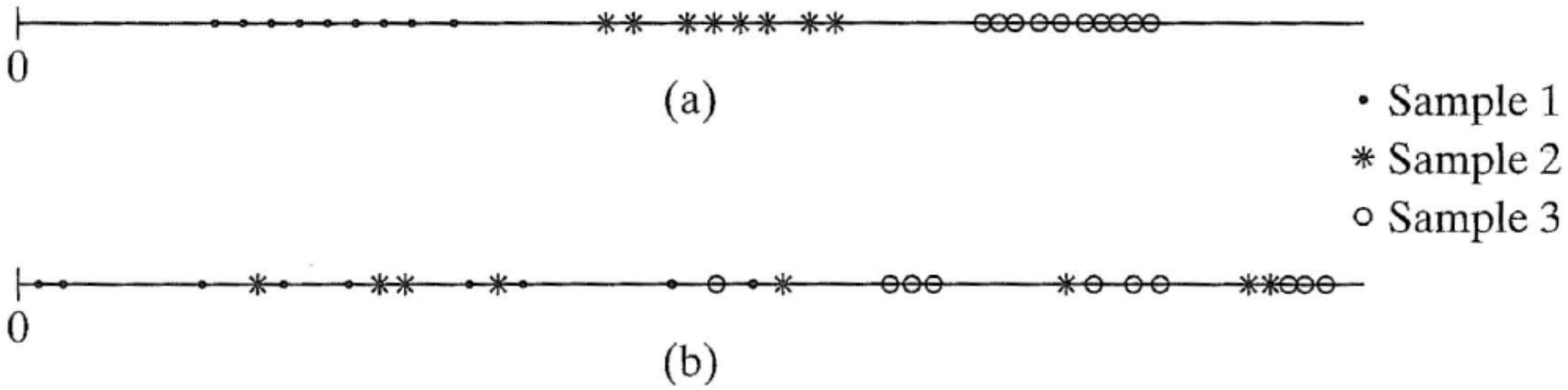
- Note that the SD of the proportion $\hat{\pi}$ is: $s = SD(\hat{\pi}) = \sqrt{\hat{\pi}(1-\hat{\pi})}$
- What is the standard error? What is the null hypothesis, the t-statistic, and the p-value? What is the 95% CI?
- Were the prayers effective for successfuly recovery?

Example with Stata: Residential Areas and Commute Time in Japan

- Tabulate
- Two group t-test
- What if we'd like to examine group differences in y across more than two groups?

Analysis of Variance (ANOVA)

- What if there are more than two groups for a comparison?
  - e.g., wage differences across four race groups



Analysis of Variance (ANOVA)

- Analysis of Variance (ANOVA): a model for comparing the means of y across multiple groups
  - ANOVA decomposes the variance of y into the *between-group* component and the *within-group* component
    - Within Sum of Squares (WSS) = $\sum_i^{n_g} (y_i - \bar{y}_g)^2$
    - Between Sum of Squares (BSS) = $\sum_g^G n_g (\bar{y}_g - \bar{y})^2$
    - Total Sum of Squares (TSS) = $\sum_i^N (y_i - \bar{y})^2 = WSS + BSS$
    - where G: number of groups, n_g: number of observations in group g, $\bar{y}_g$: the mean of y in group g

Analysis of Variance (ANOVA)

- Total variance: $\frac{TSS}{N-1}$, Within-group variance: $\frac{WSS}{N-G}$, Between-group variance: $\frac{BSS}{G-1}$
- The F-statistic of the linear regression model where y is regressed on the categorical variable x is:

$$F_{G-1,N-G} = \frac{\text{Between-group variance}}{\text{Within-group variance}}$$

- The F-statistic approaches to 1 when the means of y becomes equal between groups (e.g., $\mu_{g1} = \mu_{g2} = \ldots \mu_{gG}$) and increasingly exceeds 1 as groups differ in their means of y

Analysis of Variance (ANOVA) in practice

- In practice, ANOVA is a linear regression analysis with a categorical variable without no other covariates
  - The F-statistic indicates a statistical inference to figure out if the between-group variance (explained by our model; ESS) is greater than the within-group variance (e.g., remains unexplained; RSS) statistically significantly
  - *reg commute i.size*
- What if we would like to do ANOVA after controlling for some covariates $\Longrightarrow$ Analysis of Covariance (ANCOVA)
- What if we would like to do ANOVA with two categorical variables $\Longrightarrow$ two-way ANOVA