Biostatisics

Lecture12

Analysis of Variance

Analysis of Variance

- Extension of T-Test to multiple group means
  - Reason for a single 'global' test is to avoid problems with Type I error if test all possible pairs using a t-test
- Compare a continuous outcome among groups
  - One-way classification
    - Treatment groups
      - Morphine, midazolam, placebo
    - SES groups
      - Education: < HS, HS+, College+
  - Two-way classification
    - Treatment by age
      - Six groups
    - Can look at factors separately
      - Differences among treatment groups and among age groups

ANOVA Models - I

- Now dealing with outcome variables and predictor or explanatory variables
- Ex. for one-way ANOVA
  - BMI = $\alpha + \beta(\text{income}) + \varepsilon$
  - BMI is predicted by an overall mean (intercept) + an effect of income level β(* income level) + a random effect
  - Typically, income would be scaled from a reference group
    - e.g., lowest income = 0, middle income = 1, high income = 2

ANOVA Models - I

- Thus, to predict BMI at lowest income, model would simply be
  - BMI = $\alpha + \varepsilon$
  - Implication: α is the mean BMI for lowest income group with a random effect for each participant
    - Generalization: α is the mean when all other factors are set to 0

ANOVA Models - I

- To predict BMI at middle income level (income=1), model would be:
  - BMI = $\alpha + \beta*1 + \varepsilon$
  - Note: coding of factor is important in this situation
    - Some statisticians argue that multi-level categorical variables should be converted to dummy variables for modeling purposes
    - Implication of structure is that income level 2 has twice the effect of income level 1
      - » Difficult to demonstrate

ANOVA Models – II

- Two-way ANOVA is analogous to one-way ANOVA with an additional term
- Ex. for two-way ANOVA
  - BMI = $\alpha + \beta_1$(income) + β_2(race) + ε
  - Interpretation is very similar
  - BMI is predicted by an overall mean (intercept) + an effect of income level (β_1 * income level) + an effect of race (β_2 * race) + a random effect
  - Income is coded 0, 1, 2 as before; race coded 0=white, 1=black
  - Now, to predict BMI at middle income level for black participants, model would be:
    - BMI = $\alpha + \beta_1 * 1 + \beta_2 * 1 + \varepsilon$
    - Note: now, α is the mean of white participants at the low income group

ANOVA Models – III

- Two-way ANOVA with interaction allows us to look at differential (non-additive) effects of factors
- Ex. for two-way ANOVA with interaction
  - BMI = α + β_1(income) + β_2(race) + β_3(income*race) + ε
  - Interpretation is very similar with the additional aspect of the income-race interaction
  - Income-race interaction for each participant is constructed just as it looks – code for income multiplied by code for race

ANOVA Models – III

- BMI is predicted by an overall mean (intercept) + an effect of income level (β_1 * income level) + an effect of race (β_2 * race) + the non-additive effect of income and race + a random effect
- Now, to predict BMI at middle income level for black participants, model would be:
  - $BMI = \alpha + \beta_1 * 1 + \beta_2 * 1 + \beta_3 * 1 + \varepsilon$
  - Note: β_3 can be any magnitude and can be tested just like any other factor

ANOVA – Further testing

- ANOVA gives you the ‘global’ test for differences among groups
  - Does not tell you where the differences are
- If the ANOVA yields a significant result for the factor, then need to do futher testing to find out where the differences are
  - Use techniques such as Tukey’s HSD, Bonferroni, LSD, etc. to identify pairs that are significantly different

Anatomy of ANOVA table

- Total Sums of Squares (SST)
  - Use sums of squares as they represent the deviations (variance) and they can be manipulated (i.e., added, subtracted, etc.)
  - SST is total SS for the entire sample – analogous to the variance
- Error Sums of Squares (SSE)
  - Also called ‘within group’ or ‘residual’ sums of squares
  - Represents the inherent variability within each group
    - SSE is the 📪 in the models above
    - Variation around group mean

Anatomy of ANOVA table

- Model Sums of Squares (SSM)
  - Also called ‘among group’ sums of squares
  - Represents the variability in the sample accounted for by differences among the group means
  - This is the part of the ANOVA Table that is of most interest
    - It tells you if the factor that you are interested in explains a significant amount of the variance in the sample
    - If the group means have little deviation from the overall mean, they will explain very little of the variance in the sample and that factor will not have a significant F-test in the table (non-significant p-value)

ANOVA F-Tests

- F-tests are constructed as the ratio of two variances with n_1-1 df for the numerator and n_2-1 df in the denominator
- In ANOVA, we are primarily testing estimates of variances (called mean square in ANOVA tables)
- Primarily, the mean square for each factor in the model is tested against the mean square error (an estimate of the sample variance)
  - If the mean square for a factor is significantly greater than the MSE compared to the tabled values for an F distribution, then we say that the factor explains a significant amount of the variance
  - MSE is a true estimate of the sample variance only if the population variances are equal

Assumptions for ANOVA

- The k sets of data are k independent random samples from the respective populations
- Populations are normally distributed
- Populations have same variance
- 👓’s are unknown and sum of 👓’s is zero
  - Not always clear from computer printouts since software selects one group as reference group and that effect is ‘buried’ in the 💻
- ε_{ij} have a mean of 0 since deviations from overall mean
- ε_{ij} have a variance equal to sample variance
- ε_{ij} are normally (and independently) distributed

Study Design Issues

- Completely randomized design
  - One-way ANOVA with ‘treatments’ assigned at random
- Randomized complete block design
  - Two-way ANOVA with ‘treatments’ assigned at random within ‘blocks’ of participants
    - Blocks chosen to group homogeneous participants together; e.g., sex, age, race…
  - Typically, treatment assigned to one member of block

Study Design Issues

- Repeated measures design
  - One- or Two-way ANOVA with repeated measurements of the outcome to test for differences across time
  - Assumptions are very restrictive
  - Better approaches for this situation is to use techniques for correlated data (i.e., GEE or random effects)
- Factorial design
  - Frequently used in RCT because can be used to test interactions
  - Treatments assigned to several members of a block
    - Thus, somewhat different from the randomized complete block design
    - Can estimate mean treatment affect within block and, thus, can test interactions