12.1: One-way ANOVA

8 Nov 2011
BUSI275
Dr. Sean Ho

- **HW7 due tonight**
- **Please download: 18-Delivery.xls**

TRINITY WESTERN UNIVERSITY

Outline for today

- One-way ANOVA (1 nominal predictor)
  - Assumptions of ANOVA
  - Concepts of ANOVA: between vs. within
  - Global F-test
  - Example: Delivery.xls
- Follow-up Analysis
  - Post-hoc test using Tukey-Kramer
  - Example: Delivery.xls
- ANOVA and regression

ANOVA: Analysis of Variance

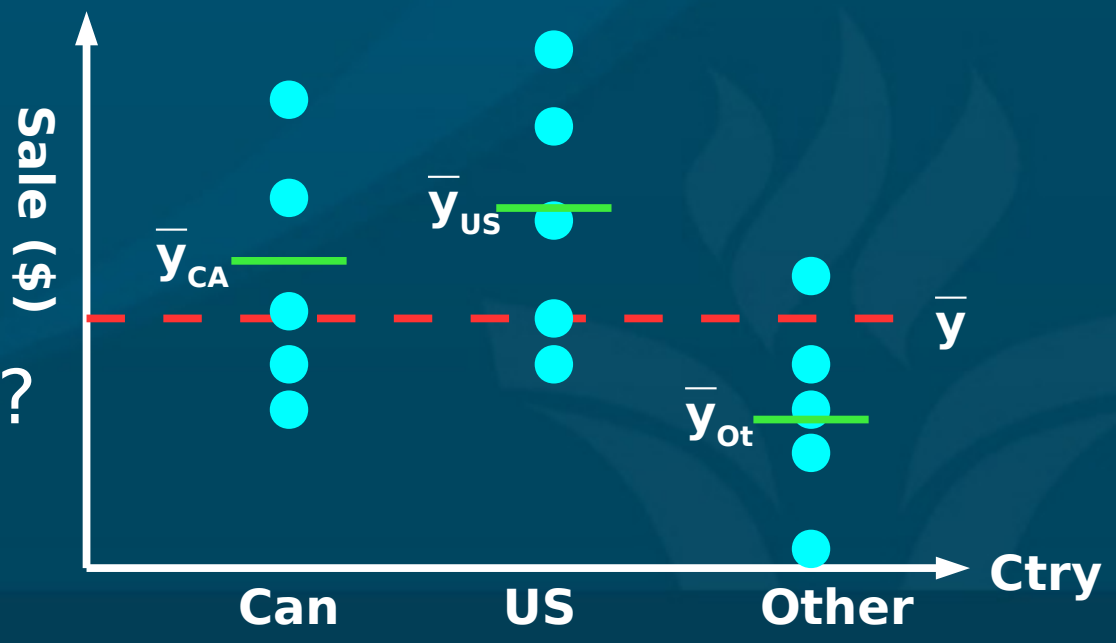
- 1 DV (scale) and one or more IVs (nominal)
  - One-way ANOVA: just one IV, with *k* levels
  - e.g., does country affect avg purchase amt?
    - Groups: Canada, US, China, UK, etc.
- The independent-groups t-test is a special case
  - One IV that is dichotomous
- ANOVA performs one global F-test to assess if the predictor has any effect on the outcome
  - H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$
  - Omnidirectional (generalization of 2-tailed)
- Follow-up tests then identify which groups differ

Assumptions: parametricity

- DV is continuous
  - If DV is dichotomous, try Logistic Regression
  - If all vars are nominal, try Log-Linear analysis
- Observations are independent, and Groups (levels of the IV) are independent
- DV is normally distributed within each group
  - If not, try transforming the DV
- Variance (SD) of DV in each group is roughly similar across all the groups (homoscedasticity)
  - Not crucial if n in each group is large and if balanced design: similar n in each group

ANOVA concepts

- How much of variability in purchase amount is due to country of origin?
  - $SS_{tot} = SS_{Country} + SS_{residual}$
  - $SS_{Country}$ is "between-group" variation (SSB)
  - $SS_{residual}$ is "within-group" variation (SSW)
- Do the group means differ significantly?
  - *F*-test, *p*-value
- Fraction of variability explained by country?
  - η^2 (equiv. to R^2)



ANOVA table

- Model: Y = (offset due to group) + (residual ε)

-	Group (Between)	Residual (Within)
SS	$SSB = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$	$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$
df	k - 1	n - k
MS = SS/df	MSB = SSB / (k - 1)	MSW = SSW / (n - k)

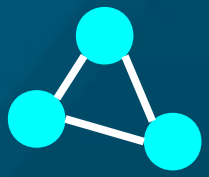
- Test statistic is F = MSB / MSW
  - Model vs. residual (as in regression!)
  - Use FDIST() with two dfs to get p-value

Example: Delivery minivans

- Dataset: 18-Delivery.xls (see p.496, #12-15)
- DV: operating cost per mile
  - IV: manufacturer (3 companies)
  - Unit of observation: one minivan (total n=13)
- ANOVA table: df = (2, 9)
  - SS = (6.07, 3.45), MS = (3.04, 0.38)
  - ⇒ F = 7.91, so p = 0.010
- Reject H_0: operating costs per mile do differ significantly depending on manufacturer

Follow-up analysis

- ANOVA's global F test is an omnibus test:
  - Just says there is a difference somewhere
  - Doesn't tell us which groups differ!
- There may be sets of groups that don't differ significantly from each other



- Follow-up analysis tries to find these
  - Post-hoc: try all pairs of groups
    - The multiple comparisons problem: "shotgun" approach leads to inflated Type I error

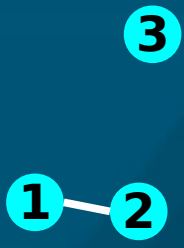
  - Planned contrasts: if theory guides us to try certain comparisons of groups

Post-hoc: Tukey-Kramer

- Considers all possible pairings of groups
  - (Can vs. US), (Can vs. Other), (US vs. Other)
  - In general, k*(k-1) pairings!
- From table in Appendix J, find critical value for q
  - Test statistic for studentized range (like F)
  - Use α (.05 or .01) and both dfs to look up
- For each pairing (group i vs. group j):
  - Find standard error: $SE = \sqrt{\frac{MSW}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$
  - Compare difference of means: $|x_i - x_j|$ against critical range: (q)*(SE)
  - If larger, then these groups differ significantly

Tukey-Kramer: Delivery.xls

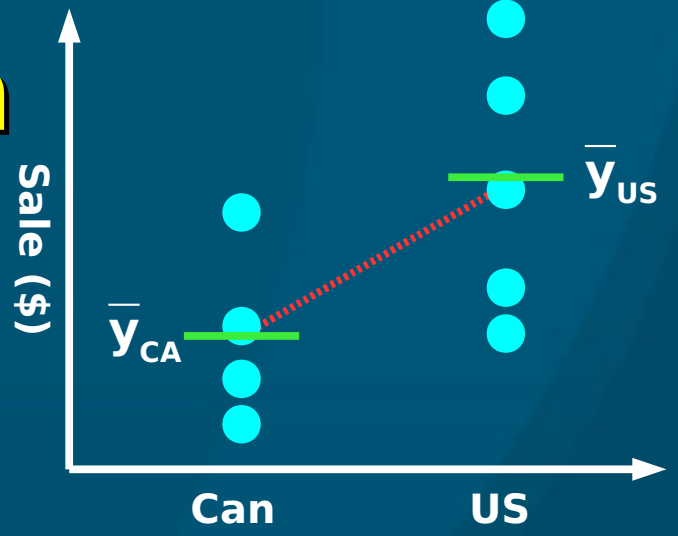
- Which manufacturers differ significantly?
- Appendix J (p.867): α=0.05 (95% conf)
  - df = (2, 9) ⇒ q = 3.20
- Calculate SE for each pairing
- Calculate critical range for each pair: q*SE
- Compare against mean differences:



q	3.20		
Pair:	1 vs 2	1 vs 3	2 vs 3
SE	0.303	0.279	0.317
Crit Range:	1.024	0.940	1.071
Mean diff:	0.633	1.175	1.808
Result	FALSE	TRUE	TRUE

- Conclusion: manufacturer 3 is the odd one out, with significantly higher operating costs

ANOVA and regression



- With only 1 dichotomous IV:
  - ANOVA = t-test = regression
  - Code the IV as 0/1
    - Intercept b_0 = mean of group 0 ($\bar{y}_0$)
    - Slope b_1 = difference of means
  - Effect size $\eta^2 = R^2$
- If the IV has multiple levels, use dummy coding:
  - Choose a reference level
  - Make k-1 dummy variables, for each of the other levels: each coded 0/1
  - Use multiple regression

Cty	US	Ot
Ca	0	0
US	1	0
Ot	0	1

TODO

- HW7 (ch10,14): due tonight
- Projects:
  - Acquire data if you haven't already
    - If waiting for REB: try making up toy data so you can get started on analysis
  - Background research for likely predictors of your outcome variable
  - Read up on your chosen method of analysis (regression, time-series, logistic, etc.)