Ch10: Comparing Two Groups

- **REB forms due**
- **Midterm Thu**
- **Please download: Mileage.xls**

18 Oct 2011
BUSI275
Dr. Sean Ho

Outline for today

- Comparing two independent groups
  - DV and IV
  - Confidence interval vs. hypothesis test
  - Standard error
    - If we know the σ's
    - If we only have the s's
    - If the s's are similar (homoscedastic)
  - Example: by hand (risk tolerance)
  - Example: in Excel (mileage)

Comparing two groups

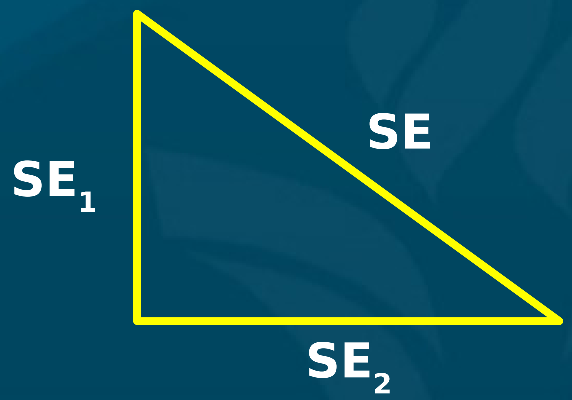
- Assume: quantitative DV
- Assume: two independent groups
  - IV is dichotomous (nominal w/ 2 categories)
  - Each participant goes in only one group
- Look at difference between pop means: $\mu_1 - \mu_2$.
- E.g., is CEO salary in US higher than in Can?
  - DV: salary. IV: country (US vs. Can)
  - H_A: $\mu_{US} - \mu_{Can} > 0$
- E.g., does gender affect invest. risk tolerance?
  - DV: risk tolerance. IV: gender (M vs. F)
  - H_A: $\mu_M - \mu_F \neq 0$

Hypothesis testing

- As before, we can either:
  - Estimate a confidence interval on $\mu_1 - \mu_2$
    - If 0 is not in the interval, then there is a significant difference between groups
  - Or do a hypothesis test on $\mu_1 - \mu_2$
    - $\bar{x}_1 - \bar{x}_2$ is a threshold: p-val is area in tail
- Key components:
  - Point estimate ($\bar{x}_1 - \bar{x}_2$), t-score, and standard error
  - T-distribution also needs a df

Standard error: σ known

- If we have σ_1, σ_2: $$SE = \sqrt{SE_1^2 + SE_2^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$
- SE is a "yardstick" by which we measure the group difference to see if it is significant
  - Larger SE ⇒ wider confidence interval, less precision in our estimate
- Here the SE is a combination of SE_1 and SE_2 from each of the two groups



Standard error: using s

- More realistically, we would only have s_1, s_2
  - As well as n_1, n_2, $\bar{x}_1$, $\bar{x}_2$
- SE is the same: $SE = \sqrt{SE_1^2 + SE_2^2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- But the t-dist needs a df, and it is messy:

$$df = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}$$

- In general, df is somewhere in between
  - $\min(n_1 - 1, n_2 - 1)$ (lower bound), and
  - $n_1 + n_1 - 2$ (upper bound)

Standard error: homoscedastic

- If s_1, s_2 are similar, we can try another method:
  - Homoscedasticity: same variance
  - Rule of thumb: s_1, s_2 within a factor of 2
- df is simpler: df = $df_1 + df_2 = n_1 + n_2 - 2$
- The pooled variance s_p^2 is a weighted sum:

$$s_p^2 = \left(\frac{df_1}{df}\right) s_1^2 + \left(\frac{df_2}{df}\right) s_2^2$$

- So the pooled SD is:

$$s_p = \sqrt{\frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2}}$$

- Then the SE simplifies to:

$$SE = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Example: risk tolerance

- RQ: do M have higher risk tolerance than F?
  - Data: 15 males, avg tol 7.8, SD=2
    12 females, avg tolerance 7.2, SD=2.5
- Point estimate: difference in tol is $\bar{x}_1 - \bar{x}_2 = 0.6$
- Standard error: using s, try heteroscedastic
  - $SE_1 = 2/\sqrt{15} \approx 0.5164$, $SE_2 = 2.5/\sqrt{12} \approx 0.7217$
    - $SE = \sqrt{(SE_1^2 + SE_2^2)} \approx 0.8874$
  - Messy df ≈ 20.8
- ⇒ t-score is $t = (0.6 - 0)/SE = 0.6/0.8874 \approx 0.68$
- p-val: TDIST(0.68, 20.8, 1) → 25.3%
- Fail to reject H_0: M tol. not significantly higher

TODO

- REB form due today
  - If approval by TWU's REB is required, also submit printed signed copy to me
- Midterm (ch1-8): Thu
- HW6 (ch9-10): next Thu 27 Oct