Ch10: T-tests

- *Please download: 08-TTests.xls*
- *HW5 this week*
- *Projects*

6 Mar 2012
Dr. Sean Ho

busi275.seanho.com

Trinity Western University

Outline for today

- Preview of statistical tests for your projects
- T-tests (comparing two groups of values):
  - Standard error
    - When σ_1, σ_2 are known
    - When s_1, s_2 are known, heteroscedastic
    - When s_1, s_2 are known, homoscedastic
  - Using Excel's TTEST() function on data
  - Types of t-test
    - Independent groups
    - Binomial proportions (σ known)
    - Paired data

Exploratory analysis

- Choosing good research questions:
- Start with the outcome variable (DV)
  - e.g., sales volume
- Research background (prior literature) on the DV to find likely predictors
  - e.g., marketing budget, consumer trends, new products from competitors, etc.
- Select some effect/predictor(s) to examine
  - In your analysis, control for other covariates
- Correlation ≠ causation: look for hidden vars
  - e.g., ice cream correlates with drownings!
    - Why? What are they both correlated with?

Analysis Types by IV/DV

- DV quantitative, IV categorical:
  - IV dichotomous (two groups): t-test
  - IV has many groups: ANOVA
  - Multiple categorical IVs: Factorial ANOVA
    - Controlling for covariates: ANCOVA
- DV quantitative, IV quantitative:
  - One IV: Simple Regression
  - Multiple IVs: Multiple Regression
    - Also if mix of categorical/quant IVs
- DV dichotomous: Logistic Regr. (survival an.)
- DV ordinal: Ordinal Regr.
  - … and much more!

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 to calculate:
  - Point estimate ($\bar{x}_1 - \bar{x}_2$),
  - t-score, and
  - Standard error
  - T-distribution also needs a df

How to approach a t-test

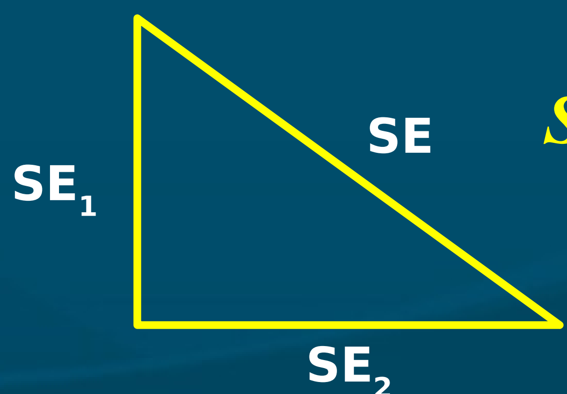
- What format do you want the output in?
  - Hypothesis test (*p*-value) or conf. interval?
- What info on the data do you have?
  - Full dataset: use Excel's TTEST() function
  - Only means/SD: calculate manually
    - Standard error (SE) is key ingredient
- What type of t-test?
  - Independent groups
    - Homoscedastic or heteroscedastic?
  - Binomial proportions
  - Paired data

Outline for today

- Preview of statistical tests for your projects
- T-tests (comparing two groups of values):
    - Standard error
        - When σ_1, σ_2 are known
        - When s_1, s_2 are known, heteroscedastic
        - When s_1, s_2 are known, homoscedastic
    - Using Excel's TTEST() function on data
    - Types of t-test
        - Independent groups
        - Binomial proportions (σ known)
        - Paired data

Standard error: σ known

- 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
- If we have σ_1 and σ_2: the SE is a combination of SE_1 and SE_2 from each of the two groups:



$$SE = \sqrt{SE_1^2 + SE_2^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

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 are 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:

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

Outline for today

- Preview of statistical tests for your projects
- T-tests (comparing two groups of values):
  - Standard error
    - When σ_1, σ_2 are known
    - When s_1, s_2 are known, heteroscedastic
    - When s_1, s_2 are known, homoscedastic
  - Using Excel's TTEST() function on data
  - Types of t-test
    - Independent groups
    - Binomial proportions (σ known)
    - Paired data

Example: household income

- RQ: did US household income decrease between 2001 and 2004?
- Data: income for each of 100 hholds in 2001; another sample of 100 households in 2004

- What format output? (p-value or conf. int.?)
- What format input? (raw data or just mean/SD?)
- What kind of t-test?
  - Indep. groups, proportions, or paired data?
  - Homoscedastic or heteroscedastic?
    - How can we check?
- See “Income” in 08-TTests.xls

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

Outline for today

- Preview of statistical tests for your projects
- T-tests (comparing two groups of values):
  - Standard error
    - When σ_1, σ_2 are known
    - When s_1, s_2 are known, heteroscedastic
    - When s_1, s_2 are known, homoscedastic
  - Using Excel's TTEST() function on data
  - Types of t-test
    - Independent groups
    - Binomial proportions
    - Paired data

T-test on proportions

- e.g., customer satisfaction vs. bank branch:
  - See “Banks” in 08-TTests.xls
  - At Langley, 160/200 customers satisfied
  - At Abbt., 210/300 satisfied
  - Is there a significant difference?
- Use normal approximation to binomial:
  - When both np, nq > 5 (for both groups)
- For confidence intervals, SE = $\sqrt{SE_1^2 + SE_2^2}$
  - Where each $SE_i = \sqrt{p_i q_i / n_i}$
- For hypothesis tests, use a different SE
  - Uses pooled proportion:

SE for hyp. tests on proportion

- The textbook offers a second form of the SE for hypothesis tests on binomial proportions:

$$SE = \sqrt{\bar{p}\,\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

- Where $\bar{p}$ is the pooled proportion:

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

- This is equivalent to the "χ^2 test of goodness-of-fit" we will learn in ch13
  - Most stats software uses this method

Proportions: bank example

- Langley: $x_L = 160$, $n_L = 200$
- Abbt.: $x_A = 210$, $n_A = 300$
- Pooled $\overline{p} = (160+210) / (200+300) = 74\%$
- SE $= \sqrt{\overline{pq}\,(1/n_L + 1/n_A)} \approx 4.0042\%$
- Sample difference of proportions is $p_L - p_A = (160/200) - (210/300) = 10\%$
- This means a z-score of $z = ((p_L - p_A) - 0) / SE$
  - $\approx 10\% / 4.0042\% \approx 2.497$
- Find the p-value (2-tailed):
  - =2*NORMSDIST(-2.497) → 0.0125
  - Reject H_0: yes, there is a difference

Outline for today

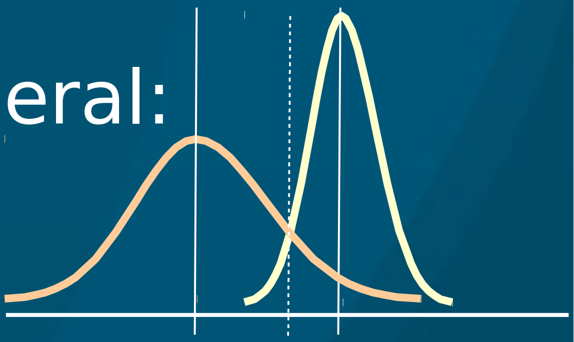
- Preview of statistical tests for your projects
- T-tests (comparing two groups of values):
    - Standard error
        - When σ_1, σ_2 are known
        - When s_1, s_2 are known, heteroscedastic
        - When s_1, s_2 are known, homoscedastic
    - Using Excel's TTEST() function on data
    - Types of t-test
        - Independent groups
        - Binomial proportions
        - Paired data

Repeated measures

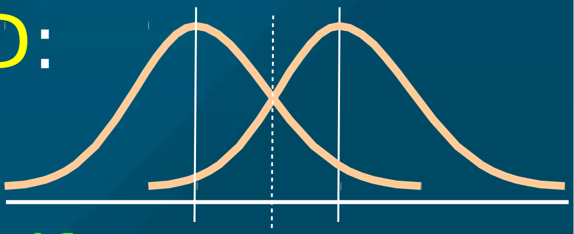
- Apply same measurement to same subjects, but at different points in time:
  - e.g., annual revenue, 2000-2010
  - Time series / longitudinal data
- Or under different conditions:
  - e.g., highway vs. city mileage (on same car!)
  - e.g., wife's income, husband's income
    - (What is the unit of observation?)
- The measurements are linked to each other
  - Not independent
- Paired data is the simplest repeated measure
  - Use a t-test on the pairwise differences

Types of t-test (as in Excel)

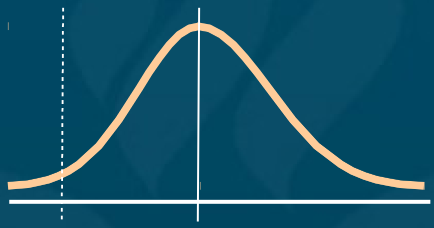
- Type 3: two indep groups, most general:
  - H_A: $\mu_1 - \mu_2 \neq 0$ (or >0)
  - SE = $\sqrt{(SE_1^2 + SE_2^2)}$, df is messy



- Type 2: two indep groups, similar SD:
  - H_A: $\mu_1 - \mu_2 \neq 0$ (or >0)
  - SE = $s_p\sqrt{(1/n_1 + 1/n_2)}$, df = $df_1 + df_2$



- Type 1: paired observations:
  - Form pairwise diffs: n = # pairs
  - H_A: $\mu_d \neq 0$ (or >0)
  - SE = $s_d / \sqrt{n}$, df = $n-1$



Paired data t-test

- e.g., "Mileage" in 08-TTests.xls
- Calculate the pairwise differences: =A2-B2, fill
- Find n, mean ($\bar{d}$), and SD (s_d) of pairs:
  - COUNT(), AVERAGE(), STDEV()
  - SD of diffs is not the same as diff of SDs!
- Calculate standard error: SE = $s_d/\sqrt{n}$
- Find t-score: $(\bar{d} - 0) / SE$
- Use TDIST() to find p-value, compare w/α
  - TDIST(*t*, *n*-1, *tails*)
- Or use all-in-one Excel function:
  TTEST(*before*, *after*, *tails*, 1)

TODO

- HW5 due Thu
- Projects: be pro-active and self-led
  - If waiting on REB approval:
    generate fake (reasonable) data and
    move forward on analysis, presentation
  - Remember your potential clients:
    what questions would they like answered?
  - Tell a story/narrative in your presentation