SO5041 Unit 14: More t-tests

Brendan Halpin, Sociology

Autumn 2019/0

The basic t-test: sample versus reference-point

- The simplest t-test compares a sample mean against a fixed number:

$$H_0 : \mu = X_r$$

$$t = \frac{\bar{X} - X_r}{SE}$$

- If t is bigger than the critical value for 95%, or its p-value is below 0.05, we reject the null hypothesis

Paired-sample t-test

- Comparing "paired samples" is the same, with the diffence being compared with zero:

$$D = X_{after} - X_{before}$$

$$H_0 : \delta = 0$$

$$t = \frac{\bar{D} - 0}{SE}$$

"t-test" for a proportion

- Comparing a sample proportion against a fixed number such as 50% has a similar logic
- It doesn't use the t-distribution, but can use standard normal for "large" samples

$$H_0 : \pi = \pi_r$$

$$z = \frac{p - \pi_r}{SE}$$

Directional hypotheses

- A complication: some hypotheses are directional
- e.g., holidays make you *happier*, training raises your earning power

$$H_1 : W_{after} > W_{before}$$

$$H_o : W_{after} \leq W_{before}$$

$$H_o : D \leq 0$$

Direction ⇒ one-way t-test

- If zero is *below* the CI, reject the null hypothesis
- If zero is within or above the CI, cannot reject
- Net result: higher confidence for the same CI – $1.96 \times SE$ for 97.5% not 95%

Comparing means across groups

- We don't always have situations where we want to test something as simple as whether the true answer is zero
- However, we very often want to test whether a mean (e.g., income) is different according to values of another variable (e.g., sex)
- If sex affects wage, we would expect to see the mean wage for men (X_m) to be different from the mean wage for women (X_w)
- We can consider the sample difference ($\bar{X}_m - \bar{X}_w$) to be a point estimate of the population difference
- The null hypothesis is that $X_m = X_w$ or $X_m - X_w = 0$

Independent-samples t-test

- To construct a CI we need the SE, which is very like the normal one if both groups have the same population variance (or standard deviation):

$$\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- If this cannot be assumed, the SE is more complex, and depends on the separate standard deviations

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Two-sample t-test in Stata

```
. ttest grsearn, by(sex)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    male |     953    25.89192    1.584105    48.90243    22.78318    29.00066
  female |     978    27.92025    1.541657    48.21223     24.8949    30.94559
---------+--------------------------------------------------------------------
combined |   1,931    26.91921    1.104885     48.5521    24.75232    29.08611
---------+--------------------------------------------------------------------
    diff |           -2.028325    2.210045               -6.362652    2.306002
------------------------------------------------------------------------------
    diff = mean(male) - mean(female)                              t =  -0.9178
Ho: diff = 0                                     degrees of freedom =     1929

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.1794         Pr(|T| > |t|) = 0.3589          Pr(T > t) = 0.8206
```

Two-sample t-test, unequal variance

```
. ttest grsearn, by(sex) unequal
Two-sample t test with unequal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
male	953	25.89192	1.584105	48.90243	22.78318	29.00066
female	978	27.92025	1.541657	48.21223	24.8949	30.94559
combined	1,931	26.91921	1.104885	48.5521	24.75232	29.08611
diff		-2.028325	2.210451		-6.363455	2.306805

```
    diff = mean(male) - mean(female)                          t =  -0.9176
Ho: diff = 0                    Satterthwaite´s degrees of freedom =   1925.9

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.1795         Pr(|T| > |t|) = 0.3589          Pr(T > t) = 0.8205
```

Key concepts

- Hypothesis test
  - Null hypothesis
  - Initial or alternative hypothesis
- Types I and II error
- Statistical significance and p-values
- t-tests:
  - Single value compared with reference
  - Paired values compared with implicit zero reference
  - Independent samples t-test: comparing two groups
    - Equal vs unequal variance

Summarising inference

- For large samples, we use the normal distribution to construct confidence intervals around means of "quantitative" (interval/ratio) variables
- For small samples we use the t-distribution to construct the confidence interval
- For interval/ratio variables we usually estimate the mean, and use

$$SE = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{\sqrt{\frac{\sum(X-\bar{X})^2}{n-1}}}{\sqrt{n}}$$

sociology UNIVERSITY OF LIMERICK

Summarising inference: proportions

- For nominal variables like vote, sex, etc. we calculate proportions, not means (where we split in two)
- With large samples (at least 20 in each category) we can construct confidence intervals using the normal distribution and the formula $\sigma_{\hat{\pi}} = \sqrt{\frac{p(1-p)}{n}}$ for the standard error
- With this we can conduct hypothesis tests in exactly the same way as with interval/ratio data
- However, with small samples the approximation to the normal distribution no longer holds and we have to use another distribution, the binomial distribution

Summarising inference: χ^2 test

- For nominal variables with more categories, and for tables made from nominal variables, we can use the χ^2 test
- Again, this has "large sample" requirements – none of the expected values should be < 5