SO5041 Unit 13: Hypothesis testing

Brendan Halpin, Sociology

Autumn 2019/0

Topics

- Three important topics today:
  - Hypothesis testing
  - Significance
  - t-test for paired samples

Confidence intervals assess imprecision

- Confidence intervals allow us to present sample information appropriately
  - Point estimate, e.g., mean or other sample statistic: our "best guess" of the true value
  - Confidence interval: range in which we are "confident" true value (the population parameter) lies
- In combination, the point estimate and CI give us the answer with a measure of its precision

Hypothesis testing and CIs

- Statistical inference proceeds by hypothesis testing: a formalisation of the Confidence Interval approach
- Easier to disprove something than to prove something true
- Thus if we wish to test if a variable has an effect on another (e.g., does a switch to a 4-day week change productivity) we set up a hypothesis, for instance

  H_1: *Productivity after is not the same as productivity before*, $X_a \neq X_b$

Recent example

- Microsoft recently experimented with a 4-day week in Tokyo:

https://www.theguardian.com/technology/2019/nov/04/microsoft-japan-four-day-work-week-productivity

- They found productivity increased, as well as worker satisfaction

Null hypothesis

- To test this turn it around, and set up a null hypothesis that says the opposite:

  H_0: *Productivity after is equal to productivity before,* $X_a = X_b$

- If we can *reject the null hypothesis* on the basis of our sample data, we can say the data support the main hypothesis

Claims about population

- Hypothesis testing is a way of using the reasoning behind CIs to make specific claims about the population
- Say we want to know if there is a relationship between two variables, e.g., whether switching to a 4-day week has an effect on productivity (positive or negative)
- We look at a sample of workers to make inferences about the population
- We begin with a hypothesis: $X_a \neq X_b$ or $X_a > X_b$
- We negate the hypothesis to form a null hypothesis, called H_0: $H_0 : X_a = X_b$ which is equivalent to $H_0 : D_x = X_a - X_b = 0$
- That is, on average, the productivity difference is zero

Testing the null hypothesis

- We then test the null hypothesis:
  - First we calculate a sample mean productivity difference, $\hat{D}_x$
  - Then we construct a confidence interval at our chosen level of confidence (e.g., 95%): $\hat{D}_x \pm z_{0.95} \times SE$
  - The CI gives us a band around the point estimate within which we are 95% sure the true value lies
  - If zero lies outside the interval, we are at least 95% sure the true population value is not zero, and we can reject the null hypothesis
  - If zero lies within the interval, then zero is in the range of plausible values, so we cannot reject the null hypothesis
  - We don't say we "accept the null hypothesis" because zero is just in the range of plausible values, and other values in this range are approximately as likely

Reject or fail to reject

- Rejecting the null hypothesis constitutes support for the initial or "alternative" hypothesis
- Failing to reject the null hypothesis means the data fail to support the initial hypothesis: "there is no evidence that the switch to a 4-day week affects productivity"
- Failure to support the initial hypothesis may be because
  - It is actually false, i.e., $D_x = 0$
  - The effect is small and/or very variable, and thus the sample is too small to detect it

Example: Productivity before and after an intervention

ID	Before	After
1	3.5	4.6
2	1.8	2.3
3	2.4	2.5
4	3.3	4.4
5	1.7	2.1
6	3.7	5.1
7	4.4	5.6
8	3.4	4.1
9	1.8	1.4
10	2.3	1.7

How do we "test" the null hypothesis?

- In this example we calculate the difference in productivity before and after, in the sample data
- Some may be negative, some may be positive, but we are interested in the average: is it systematically different from zero?
- Strategy: calculate the mean of the differences, and construct a CI around it (say at 95%)
- If zero lies outside the CI, then we are at least 95% sure the true value lies in a range that does not include zero
- If zero within the CI, then the range within which we think the true value lies does include zero

Reject or not?

- In the former case (zero outside interval), we can reject the null hypothesis: we are at least 95% sure that zero is not the true value
- We can therefore say the data support the initial hypothesis that the switch to a 4-day week affects productivity
- In the latter case (interval contains zero), we cannot reject the null hypothesis: zero is in the range where we feel the true value lies
- In this case there is no evidence in the sample data of an effect of the 4-day week on productivity
- This is not the same as evidence of no effect!
- That zero lies within the CI is not the same as zero being the true value!

Statistical significance

- Significance: let's say we do a hypothesis test with a 95% confidence level, and we find the zero is way outside the CI
- We can try again with a 99% confidence level:
  - If it is still outside the interval we are not "at least 95%" but "at least 99%" sure that zero is not the true value
- If we keep trying with CIs with higher confidence levels we will eventually find one where zero is just outside the interval
- If that is at confidence level C we can say that we are C% sure (not "at least" any more) that zero is not the true value

The chance of being wrong?

- There is then a C% chance that the true value is in the range that doesn't include zero, or a 100% - C% chance that the true value is outside the CI, and therefore could include zero
- This p = 100%-C% value is the probability that we get a sample statistic as different from zero as we did, even though the true value was zero
- This is known as the significance of the sample estimate, or its p-value
- We want it to be as small as possible, typically under 5% (0.05)
- p-values are widely used – stats programs report them in many places
- In general the interpretation is "what's the probability of getting this result by chance if the null hypothesis was true?"

App: confidence intervals and significance

http://teaching.sociology.ul.ie:3838/apps/cislide

t-test

- Rather than use the CI we can set this up as a "t-test"
- We can find the t corresponding to the CI just touching zero thus:

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

- If that t is larger than the critical value, then the CI using the critical value is smaller and doesn't overlap zero
- The significance is the exact p-value of that t
- This example is a "paired sample t-test"

t-test example in Stata

```
. gen diff = after-before

. ttest diff == 0

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    diff |      10    .5499999    .2166667    .6851601    .0598659    1.040134
------------------------------------------------------------------------------
    mean = mean(diff)                                             t =   2.5385
Ho: mean = 0                                     degrees of freedom =        9

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 0.9841         Pr(|T| > |t|) = 0.0318          Pr(T > t) = 0.0159
```

t-test – paired

```
. ttest before==after

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  before |      10        2.83     .299648      .94757    2.152149    3.507851
   after |      10        3.38    .4859812    1.536808    2.280634    4.479366
---------+--------------------------------------------------------------------
    diff |      10   -.5499999    .2166667    .6851601   -1.040134   -.0598659
------------------------------------------------------------------------------
     mean(diff) = mean(before - after)                            t =  -2.5385
 Ho: mean(diff) = 0                              degrees of freedom =        9

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.0159         Pr(|T| > |t|) = 0.0318          Pr(T > t) = 0.9841
```

χ^2 test and significance

- Another example of significance occurs in the χ^2 (chi-sq) test for association in a table
- Here the initial hypothesis is that the two variables are associated
- Thus the null hypothesis is that they are not associated (the "independence hypothesis")
- When we calculate the χ^2 statistic ($\sum \frac{(O-E)^2}{E}$) we compare its value with the range of possible values we would get if H_0 were true
- This is what we read from the table of the χ^2 distribution

χ^2 example in Stata

```
. tab race region, chi
```

race of respondent	region of the united states north eas	south eas	west	Total
white	582	307	375	1,264
black	82	94	28	204
other	15	14	20	49
Total	679	415	423	1,517

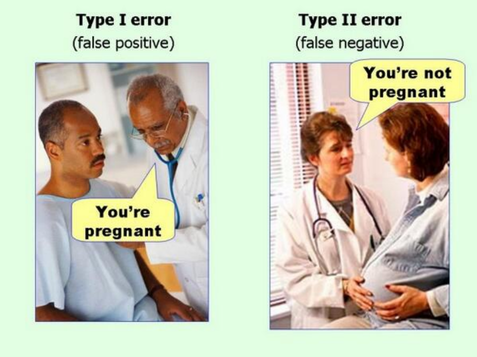
```
Pearson chi2(4) =  53.1683   Pr = 0.000
```

Signficance and error

- Another way of looking at significance is "the chance we would be wrong if we believe the initial hypothesis"
- For instance, if there is one chance in twenty ($p = 0.05$) that the true value is outside the CI, then by basing our decision on the CI we will be wrong one time in twenty
- This is known as Type I Error: rejecting the null hypothesis when it is true
  - e.g., the true value might be zero but a small number of possible samples generate CIs that don't include zero
  - e.g., there may be no association but a small number of possible samples yield high χ^2 statistics

Type I and Type II error

- If very important to avoid Type I error, use high confidence levels (e.g., 99.5% instead of 95%) or insist on low p-values (e.g., 0.005 instead of 0.05)
- However, there is a second type of error, Type II
  - Failing to reject the null hypothesis when it is false
- That is, failing to support the initial hypothesis even though it is true



Type I and Type II

- If we raise the confidence level we reduce the risk of Type I error but raise the risk of Type II error
- That is, if we make a special effort not to accept an initial hypothesis unless there is very clear evidence, we necessarily fail to accept it where there is only fairly clear evidence
- For a given p-value, we can only reduce the Type II error by increasing the sample size