t-tests (cont'd)

Research Methods for Political Science

Thomas Chadefaux

Trinity College Dublin

Comparing *two* samples

So far

So far: test for ONE sample: is the sample mean different from e.g. 0?

Now: test for TWO samples. Two cases:

- Paired samples $\rightarrow$ very easy, same as one sample case
- Independent samples:

Paired samples

Paired samples: the theory

Paired samples: same unit, observed twice.

E.g., ask Bob, Jane, Nathan and Claire how they feel about xyz at time t. Then ask them again at time $t + 1$. Same people, different time.

Paired samples: the theory

E.g.:

	Before	After
Bob	30	31
Jane	42	44
Nathan	76	76.5
Claire	13	12.8

Hypothesis: There is no change

Paired samples: the theory

E.g.:

	Before	After	*Difference*
Bob	30	31	1
Jane	42	44	2
Nathan	76	76.5	0.5
Claire	13	12.8	-0.2

Paired samples: the theory

The t-test will be conducted on the *difference*:

$$t_{paired} = \frac{\overline{x}_D - 0}{SE_D}$$

where $\overline{x}_D$ is the average difference and SE_D is the standard error of the difference.

Note that it is EXACTLY the same as a standard t-test, except that we are using the mean *difference*, not the simple mean.

Paired samples: an example

	Before	After	*Difference*
Bob	30	31	1
Jane	42	44	2
Nathan	76	76.5	0.5
Claire	13	12.8	-0.2

```
##
##  One Sample t-test
##
## data:  c(1, 2, 0.5, -0.2)
## t = 1.7836, df = 3, p-value = 0.1725
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.6470606  2.2970606
## sample estimates:
```

Independent samples

t-tests for independent samples

This is slightly trickier, but the logic is exactly the same.

Independent samples: different units (e.g, different people, individuals, etc.)

there is no matching of the units in the two samples, and the two samples may be of different sizes.

- The natural estimator of the difference $\mu_1 - \mu_2$ is $\bar{x}_1 - \bar{x}_2$, and so we need to know its sampling distribution.

t-tests for independent samples

- The variance of the difference $\bar{x}_1 - \bar{x}_2$ is the sum of their variances:

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

- So the two-sample t-test is:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

two-sample t-test: an example

1. state the hypothesis:

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

two-sample t-test: an example

Step 2: choose a significance level α. Let's say $\alpha = 0.05$

two-sample t-test: an example

Step 3: gather what we need, i.e., the mean and the standard deviation of each sample:

```
mean(x1)
```

```
## [1] 2.5
```

```
mean(x2)
```

```
## [1] 5.25
```

```
var(x1)
```

```
## [1] 1.666667
```

```
var(x2)
```

```
## [1] 2.916667
```

two-sample t-test: an example

Now apply the formula:

$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\bar{\mu_1} - \bar{\mu_2})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{1}$$

$$\approx \frac{(2.5 - 5.25) - 0}{\sqrt{1.66/4 + 2.99/4}} \tag{2}$$

$$\approx -2.57 \tag{3}$$

two-sample t-test: an example

Let's check with R:

```
##
##  Welch Two Sample t-test
##
## data:  x1 and x2
## t = -2.569, df = 5.5846, p-value = 0.04522
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  -5.41722062 -0.08277938
## sample estimates:
## mean of x mean of y
##      2.50      5.25
```

(NB: degrees of freedom for a two-sample t-test)

In case you are wondering how R calculated the degrees of freedom here. Because we are not assuming equal variance, the distribution of the test statistic is approximated as an ordinary Student's t-distribution with the degrees of freedom calculated using

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

This is known as the Welch–Satterthwaite equation.

Inference for Proportions

Inference for a single Proportion

- Count data rather than measurements
- When you have at least 15 observations
- The sample proportion is

$$\hat{p} = \frac{X}{n},$$

  where X is the number of successes.

- The standard error of $\hat{p}$ is

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

(NB: How did we derive this?)

Too long to cover here, but see slides at the end if you are dying to know (not required).

Click *here*

Inference for a single Proportion: example

- In a sample of TCD students, researchers found that 4,000 were supportive of the CETA agreement, whereas 8,000 were opposed to it.
- The proportion of supporters is therefore

$$\hat{p} = \frac{4000}{12000} = 1/3$$

Inference for a single Proportion: example

- To find the 95% CI, first compute the SE:

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(1/3)(1-1/3)}{12,000}} = 0.00430$$

  Approximately 95% of the time, $\hat{p}$ will be within 1.96 standard errors of the true p. So the CI is:

$$\hat{p} \pm z \times SE_{\hat{p}} = 1/3 \pm 1.96 \times 0.00430 = (0.3249, 0.3418)$$

- So we estimate with 95% confidence that between 32.5% and 34.2% of students support the CETA agreement.

Inference for a single Proportion: example

- Check with R:

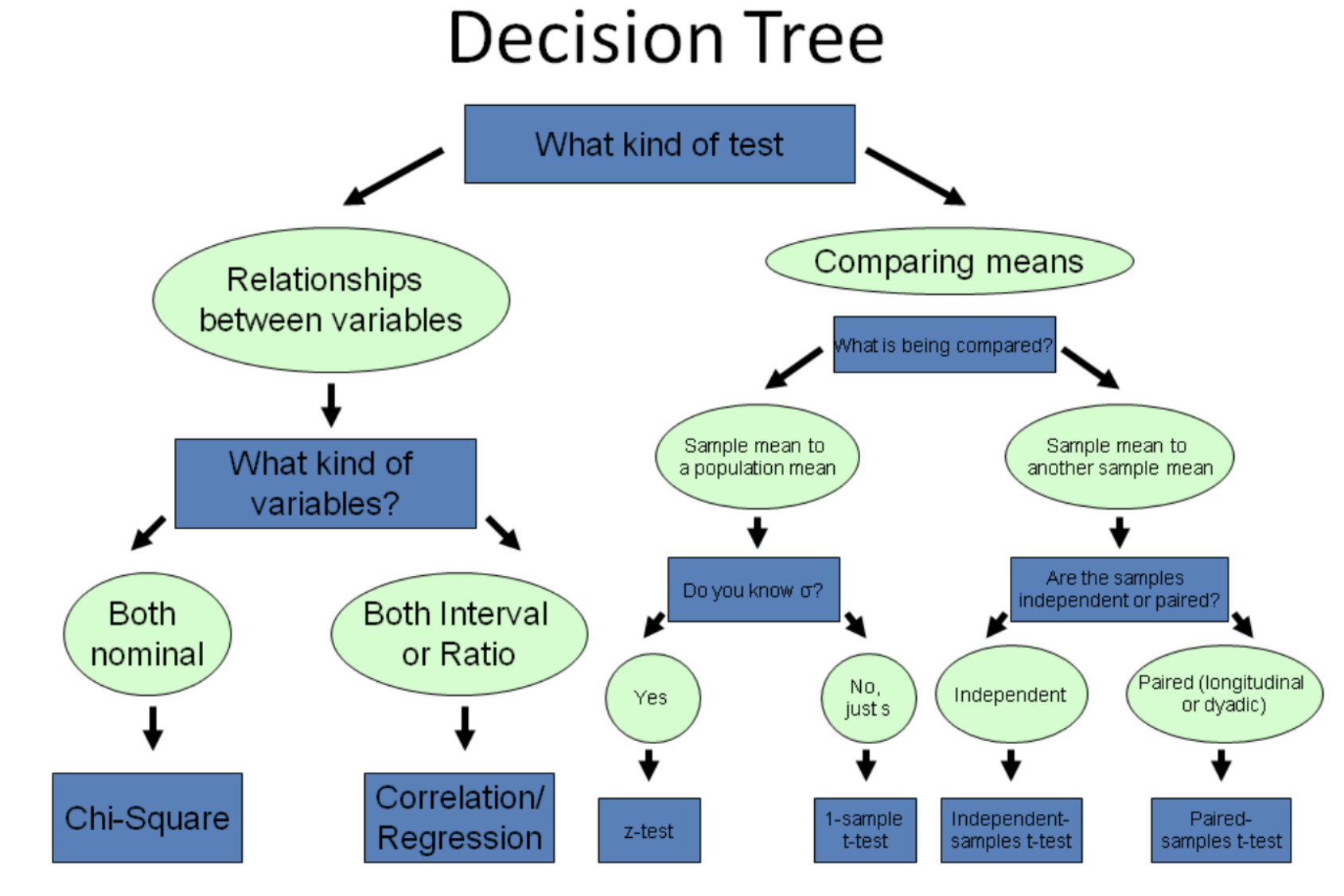
```
##
##  1-sample proportions test with continuity correction
##
## data:  4000 out of 12000, null probability 0.5
## X-squared = 1332.7, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3249121 0.3418617
## sample estimates:
##         p
## 0.3333333
```

Significance test for a single proportion

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Overview: Where we are

Overview: Where we are



Appendix

(NB: How did we derive this?)

First note that we are dealing with a sample in which there are 1s and 0s. Suppose the number of 1s is m, and the number of 0s is $n - m$. I.e. the mean is $\overline{x} = \hat{p} = m/n$

(NB: How did we derive this?)

Then the standard deviation will be:

$$\sum_i (x - \overline{x}) \tag{4}$$

$$= m(1 - \overline{x})^2 + (n - m)(0 - \overline{x})^2 \tag{5}$$

$$= m(1 - m/n)^2 + (n - m)(0 - m/n)^2 \tag{6}$$

$$= m(1 + m^2/n^2 - 2m/n) + (n - m)(m^2/n^2) \tag{7}$$

$$= m + -m^2/n \tag{8}$$

$$= m(1 - m/n) \tag{9}$$

$$= n\hat{p}(1 - \hat{p}) \tag{10}$$

(NB: How did we derive this?)

Divide all this by n to get the variance:

$$var(\hat{p}) = \frac{n\hat{p}(1-\hat{p})}{n} = \hat{p}(1-\hat{p})$$

(NB: How did we derive this?)

Then the SE, just like for the mean, will be:

$$SE = \frac{s_{\hat{p}}}{\sqrt{n}} \tag{11}$$

$$= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{12}$$

What if you data is not normal AND small sample?

- Use another distribution.
- Try to normalize the data. (Esp. for skewed data): take the log
- Use a nonparametric procedure. These don't assume that the distribution of the population has any specific form. More on this in the next lecture!

An aside: taking the log

A logarithm is the power to which a number must be raised in order to get some other number

For example, the base ten logarithm of 100 is 2, because ten raised to the power of two is 100:

$$log_{10}(100) = 2.$$

similarly

$$log_{10}(10000) = 4$$

and

$$log_e(e) = 1$$

$$log_e(2e) = 2$$

etc.

What happens when you log a skewed variable?