Applied Statistics

Lectures 8 & 9

David Barton & Sabine Hauert

Department of Engineering Mathematics

Outline

- Difference of two means
- Paired t test
- Meaning of hypothesis testing
- Confidence intervals

OpenIntro Statistics

Chapter 5, particularly §5.3

Are Trump supporters criminals?

As with all real-world questions, converting this into a statistical question is a matter of (mis-)interpretation...

Two datasets:

- Crime in the USA (FBI; 2013) — `ucr.fbi.gov/crime-in-the-u.s/`
- Red and blue states (Wikipedia; 2017) — `en.wikipedia.org/wiki/Red_states_and_blue_states`

Is there a higher rate of violent crime in states which supported Trump in the last election?

The data

State	Population	Violent crimes	Per 100,000	Voted
Alabama	4,833,722	20,826	430.8	Trump
Alaska	735,132	4,708	640.4	Trump
Arizona	6,626,624	27,599	416.5	Trump
Arkansas	2,959,373	13,621	460.3	Trump
California	38,332,521	154,129	402.1	Clinton
Colorado	5,268,367	16,226	308.0	Clinton
Connecticut	3,596,080	9,440	262.5	Clinton
Delaware	925,749	4,549	491.4	Clinton
Florida	19,552,860	91,986	470.4	Trump
Georgia	9,992,167	36,541	365.7	Trump
Hawaii	1,404,054	3,533	251.6	Clinton
Idaho	1,612,136	3,498	217.0	Trump
Illinois	12,882,135	48,974	380.2	Clinton

Summary statistics

Extremes

- Vermont (Clinton) lowest at 121.1 violent crimes per 100,000 people
- Alaska (Trump) highest at 640.4 violent crimes per 100,000 people

Means

- 20 states voted Clinton, mean of 328.8, sd of 139.5
- 30 states voted Trump, mean of 362.5, sd of 114.5

T-scores

- $\mathrm{T} = \frac{328.8 - 362.5}{139.5/\sqrt{20}} = -1.08$
- $\mathrm{T} = \frac{362.5 - 328.8}{114.5/\sqrt{30}} = 1.61$

Which one is right?

Normal probabilities

Assume

- $\bar{x}_C \sim N(\mu_C, \sigma_C^2/n_C)$ (sample mean for Clinton's states)
- $\bar{x}_T \sim N(\mu_T, \sigma_T^2/n_T)$ (sample mean for Trump's states)

Calculate the difference (properties of normal distributions)

$$\bar{x}_\Delta = \bar{x}_C - \bar{x}_T \sim N(\mu_C - \mu_T, \sigma_C^2/n_C + \sigma_T^2/n_T)$$

If we test the hypothesis that $H_0 : \bar{x}_\Delta = 0$ we get

$$\frac{\bar{x}_C - \bar{x}_T}{\sqrt{\frac{\sigma_C^2}{n_C} + \frac{\sigma_T^2}{n_T}}} \sim N(0, 1)$$

Assumes we know the population variances!

Difference of two sample means

As with the t test, we substitute in the estimates for the population variances

$$\frac{\bar{x}_C - \bar{x}_T}{\sqrt{\frac{s_C^2}{n_C} + \frac{s_T^2}{n_T}}} \sim T_{n-1}$$

But what is the number of degrees-of-freedom n?

For an accurate answer, use statistical software (it depends slightly on the data).

For a simple (conservative) answer, use

$$n = \min(n_C, n_T)$$

Difference of two sample means

Answering the question (Clinton versus Trump) we have

$$\frac{328.8 - 362.5}{\sqrt{\frac{139.5^2}{20} + \frac{114.5^2}{30}}} = -0.897$$

This should be compared against $T_{19}(0.05) = 2.09$

Hence we cannot reject the null hypothesis that the violence rates of both sets of supporters is the same. . .

Is America getting more violent?

State	Rate in 2013	Rate in 2014
Alabama	430.8	427.4
Alaska	640.4	635.8
Arizona	416.5	399.9
Arkansas	460.3	480.1
California	402.1	396.1
Colorado	308.0	309.1
Connecticut	262.5	236.9
Delaware	491.4	489.1
Florida	470.4	540.5
Georgia	365.7	377.3
Hawaii	251.6	259.2
Idaho	217.0	212.2
Illinois	380.2	370.0

Is America getting more violent?

This is *not* the same type of problem — the data are not independent samples!

This is *paired data*

Summary statistics

- Biggest decrease: Rhode Island (-38 per 100,000 people)
- Biggest increase: Montana (+71 per 100,000 people) narrowly beating Florida (+70 per 100,000 people)

Paired data

This time calculate the differences in pairs for each state

$$\begin{aligned} X_{\Delta,\text{Alabama}} &= X_{2014,\text{Alabama}} - X_{2013,\text{Alabama}} \\ X_{\Delta,\text{Alaska}} &= X_{2014,\text{Alaska}} - X_{2013,\text{Alaska}} \\ X_{\Delta,\text{Arizona}} &= X_{2014,\text{Arizona}} - X_{2013,\text{Arizona}} \\ &\vdots \end{aligned}$$

Each X_Δ is now independent of each other

Test to see if the sample mean is different to zero

$$H_0 : \bar{x}_\Delta = 0$$

Paired data

Use the standard sample means test statistic

$$T = \frac{\bar{x}_\Delta}{s_\Delta/\sqrt{n}} \sim T_{n-1}$$

where

$$\bar{x}_\Delta = \frac{1}{n}\sum X_\Delta, \quad s_\Delta^2 = \frac{1}{n-1}\sum (X_\Delta - \bar{x}_\Delta)^2$$

and n is the number of pairs

Crime data 2013–2014 gives $\bar{x}_\Delta = 2.30$, $s = 20.9$, and $n = 50$

$$T = \frac{2.30}{20.9/\sqrt{50}} = 0.78$$

The critical value is $T_{49}(0.05) = 2.01$ (two-tailed) — we cannot reject H_0

If you are comparing. . .

a sample mean to a constant then use

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad \text{or} \quad T = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim T_{n-1}$$

a sample mean compared to an independent sample mean then use

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \sim T_{n-1} \quad \text{with } n = \min(n_1, n_2)$$

pairs of dependent data then use

$$T = \frac{\bar{x}_\Delta}{s_\Delta/\sqrt{n}} \sim T_{n-1} \quad \text{with } n \text{ pairs of data}$$

Exercise

A survey is done on a random sample of gifted children and the IQ of each parent is measured.

	Mother	Father	Δ
Mean	118.2	114.8	3.4
S.D.	6.5	3.5	7.5
n	36	36	36

(Δ is the pairwise difference between the two parents.)

- Is this a difference of sample means or a paired test?
- Is there a difference between the IQs of mothers versus fathers?

Hypothesis testing

Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say 20 subjects in each sample). Further, suppose you use a simple independent means t-test and your result is 2.7 (18 degrees of freedom, 1% significance).

Please mark each of the statements below as “true” or “false”. “False” means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct.

Hypothesis testing

1. You have absolutely disproved the null hypothesis (that is, there is no difference between the population means).
2. You have found the probability of the null hypothesis being true.
3. You have absolutely proved your experimental hypothesis (that there is a difference between the population means).
4. You can deduce the probability of the experimental hypothesis being true.
5. You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.
6. You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.

Hypothesis testing

Statements made by students as to what a hypothesis test means.

1. The improbability of observed results being due to error.
2. The probability that an observed difference is real.
3. If the probability is low, the null hypothesis is improbable.
4. The statistical confidence . . . with odds of 95 out of 100 that the observed difference will hold up in investigations.
5. The degree to which experimental results are taken ‘seriously’.
6. The danger of accepting a statistical result as real when it is actually due only to error.
7. The degree of faith that can be placed in the reality of the finding.
8. The investigator can have 95 percent confidence that the sample mean actually differs from the population mean.

Confidence intervals

Problems with hypothesis testing:

- A binary indicator based on an arbitrary significance level.
- Statistical significance doesn't equal practical significance.
- "Bad" results often get discarded rather than reported.
- Easy to misuse. . .
  - Some research journals now refusing hypothesis tests on results and require confidence intervals instead!

Problems with estimation

- Point estimates are useful but not the whole story.
- There can be a range of estimates that are very nearly as good.

  For example, there will be 5 mm of rain tomorrow versus there will be between 2 mm and 8 mm of rain tomorrow.

Confidence intervals

Generating confidence intervals requires an underlying model (just like hypothesis testing, linear regression, etc).

- If the model is rubbish, the results will be rubbish!

Example (Coefficient of lift)

An experiment measures the coefficient of lift for a particular aerofoil design. Due to wind tunnel imperfections, there is a high degree of variance in the results. The measured results are

$$[1.17, 1.52, 1.35, 1.67, 1.46]$$

What is the possible range of values for the coefficient of lift to 95% confidence assuming the measurement errors/disturbances are normally distributed?

Example: coefficient of lift

The underlying model is that the data is normally distributed — nothing has been said about the mean or variance!

The best estimate of the true value of the coefficient of lift for this aerofoil is the sample mean

$$\bar{x} = \frac{1}{5}(1.17 + 1.52 + 1.35 + 1.67 + 1.46) = 1.434$$

For a normal distribution we know that the sample mean follows the distribution

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

or, in this case since we don't know the population standard deviation σ,

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \sim T_{n-1}$$

Example: coefficient of lift

We're interested in the possible values for the true mean μ and so we also need to estimate the sample standard deviation s

$$s^2 = \frac{1}{4}\left((1.17-1.434)^2+(1.52-1.434)^2+(1.35-1.434)^2 \right.$$
$$\left. +(1.67-1.434)^2+(1.46-1.434)^2\right) = 0.03513$$

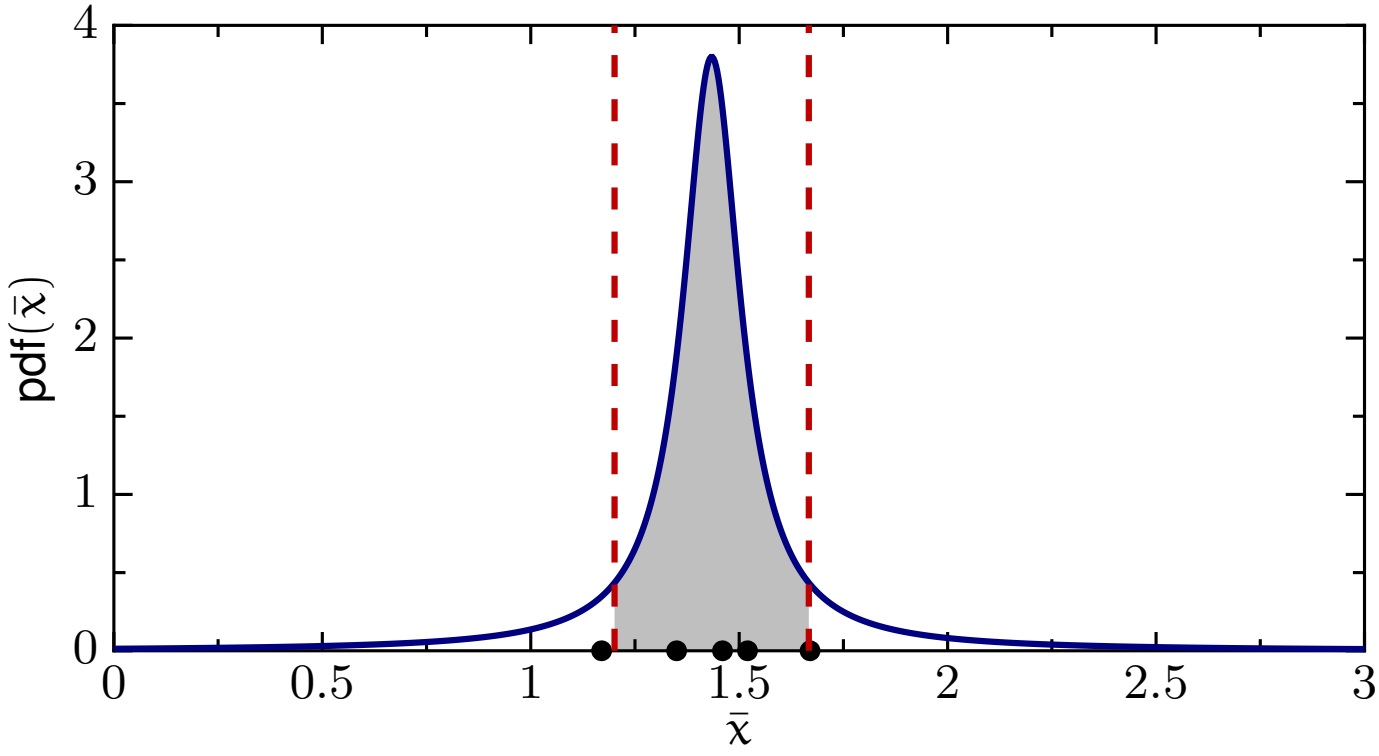
The critical values for the t distribution at 95% confidence (two-tailed) with 4 degrees of freedom are ± 2.776. Hence we have

$$-2.776 \leqslant \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{1.434-\mu}{\sqrt{0.03513/5}} \leqslant 2.776$$

and so

$$1.434 - 2.776\sqrt{0.03513/5} \leqslant \mu \leqslant 1.434 + 2.776\sqrt{0.03513/5}$$

Example: coefficient of lift



Finally, 95% of the possible values for the sample mean, based on the data provided, lie in the interval

$$1.201 \leqslant \mu \leqslant 1.667$$

Exercise

For both exercises, assume that samples are taken from a normally distributed population.

1. Calculate the 95% confidence interval for the Young's modulus of a particular sample based on the measurements

$$[222.4, 187.7, 206.2, 192.5, 190.9]$$

   (units are MPa).
2. From previous studies, the measuring device used is known to be accurate to ± 20 MPa on individual samples with 95% confidence. How does this information change the confidence interval?

Confidence intervals in general

The general formula for a confidence interval is

$$\bar{x} - C \cdot SE \leqslant \mu \leqslant \bar{x} + C \cdot SE$$

where C is the appropriate critical value, and SE is the standard error.

When you know the population variance σ^2 the standard error is

$$SE = \sigma/\sqrt{n}$$

and critical values are taken from the normal distribution.

When you only know the sample variance s^2 the standard error is

$$SE = s/\sqrt{n}$$

and critical values are taken from the T-distribution.

Exercise: Binomial distribution

The same procedure can be used to estimate confidence intervals for the probability of a Binomial distribution

Success	Failure	Total
154	46	200

The estimated probability of success is simply $\mathrm{p} = 154/200$. We could use the exact distribution to calculate the confidence interval but the normal approximation is more convenient

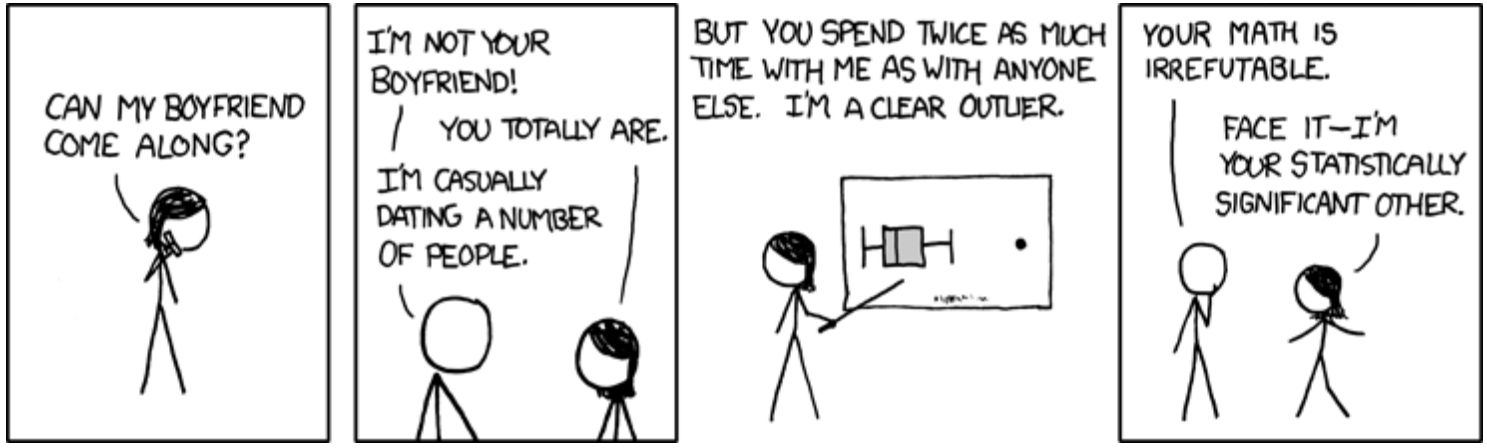
$$\text{Number of successes} \sim \mathrm{N}(\mathrm{np}, \mathrm{np}(1-\mathrm{p}))$$

What are the minimum/maximum values of p for which 154 successes out of 200 lies within the 95% confidence interval?

Quote of the day

Anon in reply to G.B. Shaw

If all the statisticians in the world were laid head to toe, they wouldn't be able to reach a conclusion



Exercises

- A selection from §5.6 in OpenIntro Statistics