LDI Seminar Series in Biostatistics: Lecture 2

Kevin McGregor

September 25, 2017

Info

- Slides, tutorial, video will be available evenutally through LDI
- For now I will upload slides (before the lecture) to my personal Github page:
- `http://github.com/kevinmcgregor/LDI-Biostatistics-Seminar`
- Tutorials and instructions posted there as well.
- Email: kevin.mcgregor@mail.mcgill.ca

Last time

- Outlined difference between population parameters and statistics.
- Defined some simple descriptive statistics.
- Showed some simple types of plots.

Significance

- We've seen some tools for doing an initial data exploration.
- We know how to calculate estimates of several population parameters.
- We don't yet know how to determine if our results are "significant".
- Can do this through *hypothesis testing*.

Research question

- Before performing a statistical test, we need to define what parameter we're investigating, and in what way it's being tested
- Consider two hypotheses corresponding to two possible states of the world: the null hypothesis (H_0) and the alternate hypothesis (H_1 or H_a). Usually boils down to:

$$H_0 : \text{Nothing interesting to see}$$
$$H_1 : \text{Hey, there could be something here!}$$

Hypothesis testing examples

- A sample of grad students at McGill is taken and an IQ test is administered... do McGill grad students have a different IQ on average than the general population (=100)?

  $$\begin{aligned} H_0 &: \text{Average IQ in McGill grad students equal to 100} \\ H_1 &: \text{Average IQ in McGill grad students not equal to 100} \end{aligned}$$

- Mean survival time in patients taking drug A vs. B:

  $$\begin{aligned} H_0 &: \text{Mean survival time is the same between groups} \\ H_1 &: \text{Mean survival time differs between groups} \end{aligned}$$

Hypothesis testing

- Always define the hypotheses with respect to parameters... never with respect to statistics!
- In the IQ example, if $\overline{x}$ is the mean IQ in the sample, and μ is the mean IQ among McGill grad students:

$$H_0 : \mu = 100$$
$$H_1 : \mu \neq 100$$

- Never this!

$$H_0 : \overline{x} = 100$$
$$H_1 : \overline{x} \neq 100$$

Types of errors

		Reality	
		H_0 **False**	H_0 **True**
Test	**Reject** H_0	Correct rejection H_0 = Power = 1 - β	Type I error = α
	Accept H_0	Type II error	Correct acceptance of H_0

https://commons.wikimedia.org/wiki/File:Inferential_Statistics_Decision_Making_Table.png

The types of errors that could be made in a hypothesis test.

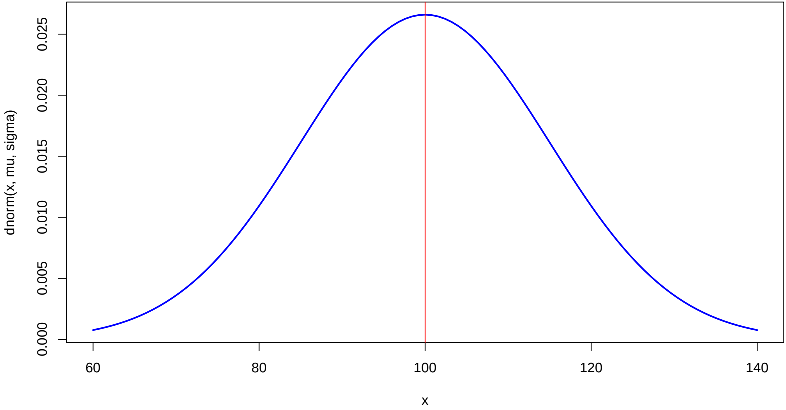
Deciding on the hypothesis

- Start by assuming null hypothesis
- Collect data and calculate "test statistic"
- If there's enough evidence based on the test statistic, reject the null hypothesis. Otherwise, do not reject the null hypothesis.
- *Never* accept either hypothesis.

Test statistic

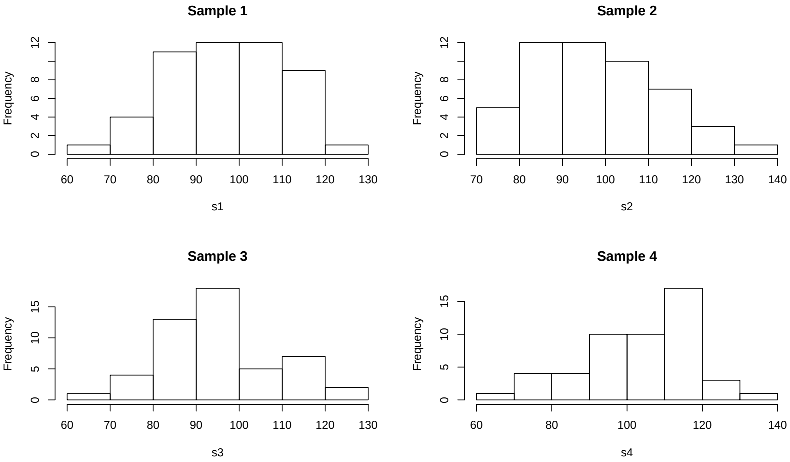
- A *test statistic* usually involves three main components:
  - Estimate of the parameter of interest
  - Measure of variation of estimate (how reliable is the estimate?)
  - Sample size
- The measure of variation is the tricky part. Let's illustrate using the sample mean.

Population normal distribution



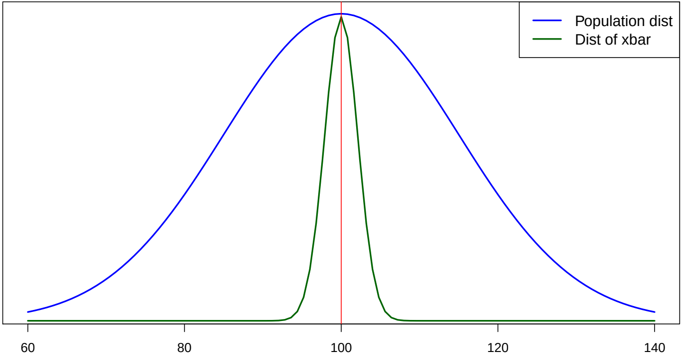
Population normal distribution: $\mu = 100$, $\sigma = 15$

Possible samples



Four samples of size 50: Sample means are: 97.34, 98.10, 95.41, 102.91

Normal distribution and $\overline{x}$ distribution



All possible values of $\overline{x}$ from sample size $n = 50$ form a normal distribution too!

Standard error of the mean

- If we could take repeated samples from a normally distributed population, the sample means would also form a normal distribution.
- A large amount of variation in this distribution means we're less confident about the estimated value of $\overline{x}$.
- If the population distribution has standard deviation σ, and we take samples of size n, then the standard deviation of all possible sample means is:

$$sd(\overline{x}) = \frac{\sigma}{\sqrt{n}}$$

- Known as the *standard error of the mean*.

Estimated standard error of the mean

- Usually, we don't know the true population standard deviation σ
- So instead, we use the sample standard deviation s to get an estimate:

$$s = \sqrt{\frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}}$$

- Then the estimate for the standard error of the mean is:

$$\hat{sd}(\overline{x}) = \frac{s}{\sqrt{n}}$$

- The estimated value for the standard error of the mean is also known as the *standard error of the mean* (yeah, seriously)

One-sample t-test

Simple statistical test

- One of the simplest statistical tests we can perform is testing whether the population mean is equal to some fixed quantity c (assuming normal distribution).
- Hypotheses:

$$H_0 : \mu = c$$
$$H_1 : \mu \neq c$$

- If we don't know the true population standard deviation σ, this test is called the *one-sample t-test*.

Assumptions of the one-sample t-test

- Assume that the underlying observations $x_1, \ldots, x_n$ come from a normal distribution.
  - For a large sample size n, don't necessarily need this assumption!
- Data observations must be independent.
- No significant outliers.

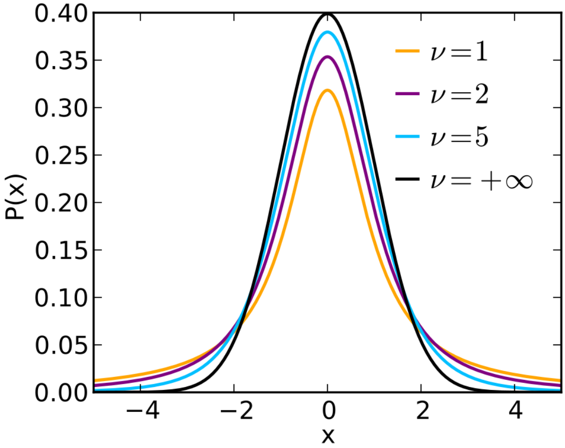
One-sample t-test

- Test statistic for the one-sample t-test:

$$t = \frac{\overline{x} - c}{s/\sqrt{n}}$$

- Reject null hypothesis if this is large in absolute value
  - Means sample mean $\overline{x}$ is far away from null hypothesis value c relative to overall amount of variation.
- Need to define threshold to make decision on whether to reject or not reject null hypothesis.

Student's t-distribution

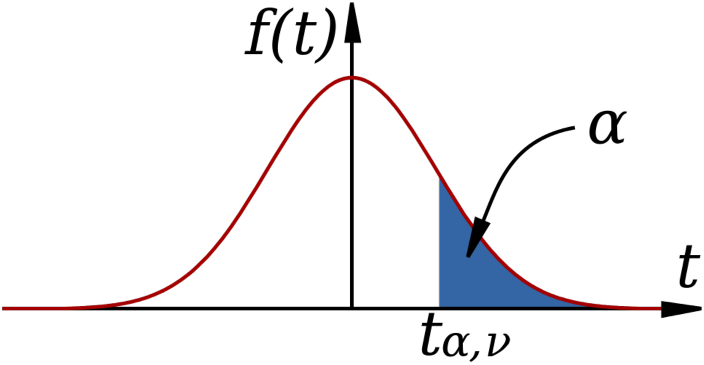


Under H_0: t statistic follows a t-distribution with $n-1$ degrees of freedom.

Extreme values in t-distribution

- We reject the null hypothesis if there were a small probability (say, $< \alpha$) of getting the observed value or a more extreme value, of t.
  - α is called the significance level of the test. Smaller α means a more strict threshold.
  - By convention we often choose $\alpha = 0.05$. However **this is completely arbitrary**. There's nothing special about this value.
- We're interested in extreme values in both tails of the distribution. But the t-distribution is symmetric, so we can just consider the absolute value of the t statistic and compare to the upper tail.

Tail of t-distribution



$t_{\alpha,\nu}$ is the t-value corresponding to a tail probability of α in a distribution with ν $(= n - 1)$ degrees of freedom. Found in a table or using software.

Performing the one-sample t-test

- Start by calculating the sample mean $\overline{x}$ and sample standard deviation s.
- Then calculate the test statistic:

$$t = \frac{\overline{x} - c}{s/\sqrt{n}}$$

- Choose significance level $0 < \alpha < 1$.
- Compare $|t|$ to the critical value $t_{\alpha,n-1}$
  - If $|t| > t_{\alpha,n-1}$, then reject the null hypothesis.
  - Otherwise, do not reject the null hypothesis.

t-test example

Systolic blood pressure calcluated for a sample of 50 medical residents at the JGH. Sample average turns out to be $\overline{x} = 128$ and standard deviation $s = 9$. Assume that the average SBP for medical residents across Canada is 120. Is the average SBP in this group significantly different from the population mean?

- $H_0 : \mu = 120$, $H_1 : \mu \neq 120$
- Test at level $\alpha = 0.05$
- Calculate t-statistic: $\frac{\overline{x}-120}{s/\sqrt{n}} = \frac{128-120}{9/\sqrt{50}} = 6.29$
- Calculate critical value: $t_{\alpha,n-1} = t_{0.05,49} = 1.68$
- Our t-statistic is greater than the critical value ($6.29 > 1.68$), so we reject H_0.

p-value

- The alternate way of determining significance is through the infamous p-value.
- The p-value can be used to determine significance in pretty much every statistical test, which is part of the reason why it's so popular.
- Have to be very careful with p-values!

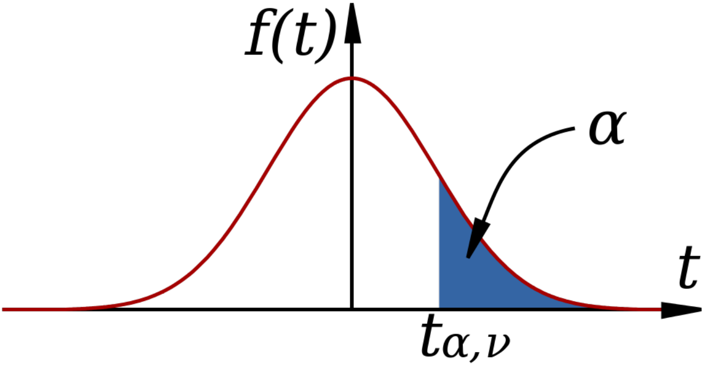
Coin flip example

- Want to test whether a coin is "fair" (i.e. has equal chance of coming up heads vs. tails)
- Can do an experiment to test this hypothesis. Say we flip the coin 100 times and we get 55 heads.
- If the coin were fair, then the probability of getting 55 or more heads is about 13%.
- If we had observed 65 heads, the same probability would be about 0.08%.

Definition of p-value

- The calculated probability is a p-value.
- A p-value is defined as the probability, **under the null hypothesis**, that we would obtain the result we got or a more extreme result.
- If that probability is small, then we have evidence against the null hypothesis.
- Can compare the p-value against some significance level α.

p-value t-distribution



In a one-sample t-test, we just calculate t as before, and calculate the tail probability. If it's less than α, we reject the null hypothesis.

Systolic blood pressure example

- In our blood pressure example from before, we calculated $t = 6.29$.
- The tail probability to the right of 6.29 in a t-distribution with 49 degrees of freedom is $p = 8.34 \times 10^{-8}$
- Significant at $\alpha = 0.05$, since $8.34 \times 10^{-8} < 0.05$
- This procedure is exactly equivalent to testing against the critical value as we did before.

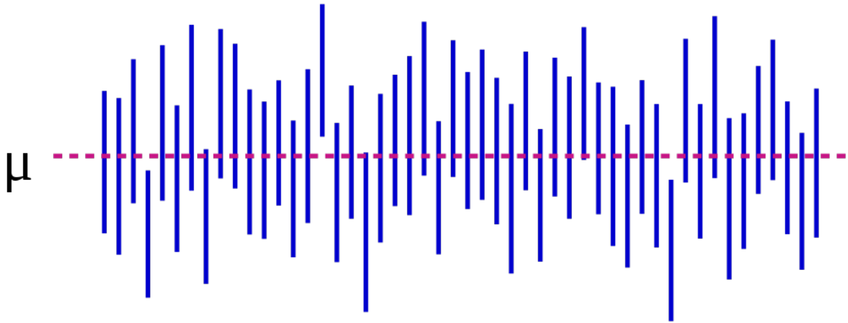
What *p*-values are **not**

- The *p*-value is *not* the probability that the null hypothesis is true.
- The *p*-value does *not* tell us whether either hypothesis is true.
- The *p*-value does *not* tell us anything about effect size. An extremely small *p*-value does not mean that something is important. It only tells us how strong the evidence is against the null hypothesis.
- $p < 0.05$ does *not* mean anything special. The choice of the significance threshold is arbitrary.

Confidence intervals

- Another important measure in hypothesis testing is the *confidence interval*.
- Gives an interval of the form (lower value, upper value) which we're "confident" contains the true population parameter.
- Our confidence is expressed through a percentage. E.g. a 95% confidence interval would contain the true parameter approximately 95% of the time if samples from the population had been taken repeatedly.

Confidence intervals



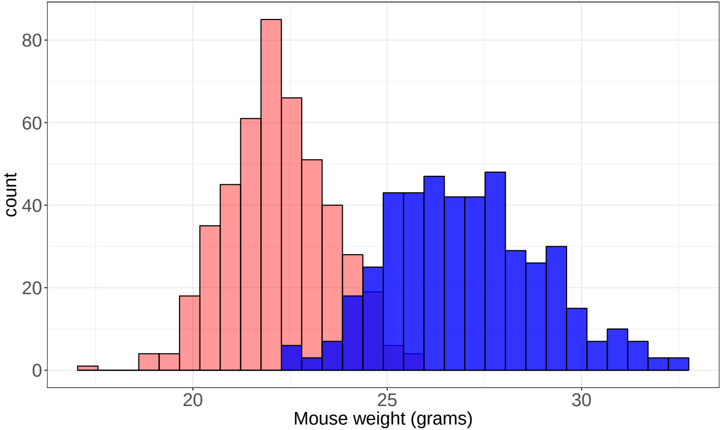
Taking repeated samples would result in the confidence interval containing the true parameter most of the time.

Confidence interval for systolic blood pressure example

- In the t-test, the $100 \times (1 - \alpha)\%$ confidence interval is $\overline{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$
- In the blood pressure example we can calculate a 95% confidence interval as (125.44, 130.56).
- Can compare the confidence interval to the value 120 to determine significance. Since our $100 \times (1 - \alpha)\% = 95\%$ confidence interval does not overlap with the value 120, then the result is significant at level $\alpha = 0.05$.
- Another equivalent method of testing significance.

Independent-samples t-test

Mouse weight



Remember the mouse weight histogram? Have male/female mice. How to test whether the mean weight differs between male/female mice?

Independent-samples t-test

- We can test the difference between the means of two groups through the independent-samples t-test.
- Assume the true mean for groups 1 and 2 are μ_1 and μ_2, respectively. Then our hypotheses are:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

- Equivalently,

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

Assumptions of the independent-samples *t*-test

- Assume that the underlying observations in each group come from a normal distribution.
- The variance should be equal in each of the two groups.
- Data observations must be independent within groups and across groups.
- No significant outliers.

Formula for the independent-samples t-test

Sample size in each group is n_1 and n_2, the sample means in each group are $\overline{x}_1$ and $\overline{x}_2$, and the sample variances in each group are . Then the test statistic is:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Mouse weight data example

- 467 female mice, 454 male mice. Mean in each group is: males 27.04g, females 22.22g. Difference = 4.81
- Resulting p-value is less than 1×10^{-16}. Implies significance difference in means at level $\alpha = 0.05$.
- 95% confidence interval is (4.59, 5.03). When comparing means we look to see whether the interval covers the value 0.

Take-home message

- Always need to be clear about what parameter you're testing and how you're testing it.
- Determining significance involves looking at both the parameter estimate, and an estimation of the theoretical variance of that estimate if we could take repeated samples.
- Don't worry too much about the calculations... more important to know how to properly interpret p-values and confidence intervals.

Thank you! - Merci!

Questions?