Parameter estimation, confidence intervals & hypothesis testing

http://imgs.xkcd.com/comics/null_hypothesis.png

Lecture 2
Biological Statistics III
Ayco Tack

1

Outline

- *Parameter estimation*
  - *Mean*
  - *Variance*
  - *Standard error*
- *Confidence intervals*
  - *A good alternative to the standard error*
  - *Special confidence intervals*
- *Hypothesis testing*
  - *Tests for simple hypotheses*
  - *Comparing two means*
- *Advice*
  - *What to use when*

2

Parameter estimation

Suppose $y_1, y_2, \dots, y_n$ is a sample from a normal distribution with mean μ and standard deviation σ; we now can estimate some parameters:

The sample average is:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- and is an estimate of μ

The sample variance is:

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

- and is an estimate of σ^2

The sample standard deviation:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

- and is the unbiased estimate of σ

But how precise are our estimates?

The sample average $\bar{y}$ is normally distributed with mean μ and standard deviation $\sigma_{\bar{y}}$

The standard deviation of the average is:

$$s_{\bar{y}} = \frac{s}{\sqrt{n}}$$

- this is the standard error!
- and is an estimate of the 'true standard error' $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$

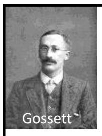
Bootstrap is a modern general method of computing estimates and standard errors (see pages 349-350 and 385-387 in the R book)

Confidence interval:

In statistics, a confidence interval (CI) is a type of interval estimate of a population parameter. It is an observed interval (i.e., it is calculated from the observations), in principle different from sample to sample, that frequently includes the value of an unobservable parameter of interest if the experiment is repeated. How frequently the observed interval contains the parameter is determined by the confidence level or confidence coefficient. (Wikipedia)

3

4



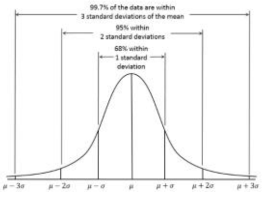
Confidence interval of the mean

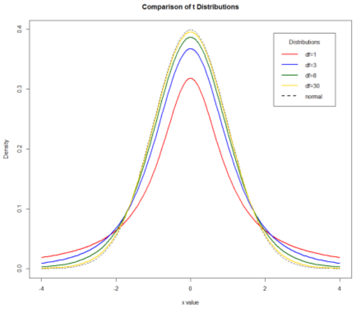
Lower and upper 95% confidence limits L_1 and L_2 for μ:

$\Pr[L_1 \leq \mu \leq L_2]$ =0.95

- $L_1 = \bar{y} - t_{0.05\,[n-1]} s_{\bar{y}}$
- $L_2 = \bar{y} + t_{0.05\,[n-1]} s_{\bar{y}}$

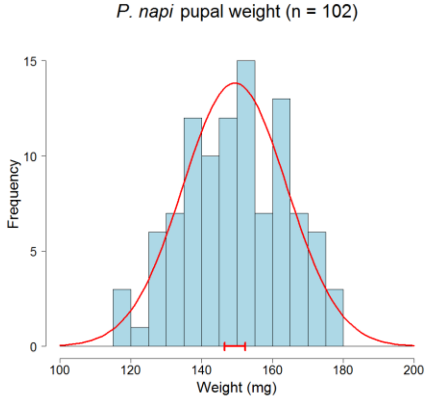
The interval is stochastic and has a 95% probability to contain the true mean μ





The **t-distribution** is a sampling distribution.
The *sampling distribution* of a statistic is the *distribution* of that statistic, considered as a random variable, when derived from a random *sample* of size n. It may be considered as the *distribution* of the statistic for all possible samples from the same population of a given *sample* size. (Wikipedia)

Example: *Pieris napi* pupal weights



Compute lower and upper 95% confidence limits for μ given:

- $n = 102$
- $\bar{y} = 149.37$
- $s = 14.7$
- $s_{\bar{y}} = \frac{s}{\sqrt{n}} = 1.456$
- $t_{0.05\,[102-1]} = 1.984$

We can then compute:

- $L_1 = 146.48$
- $L_2 = 152.26$

So the 95% confidence interval for μ is

- $146.48 \leq \mu \leq 152.26$

Test for simple hypothesis when both mean and variance are unknown

One-sample t-test is used to compare the mean of one sample to a known standard (or theoretical/hypothetical) mean (µ).

Situation: y is normal with mean μ and SD σ, where both μ and σ are unknown

Two-tailed test with the hypotheses

H_0: $\mu = \mu_0$

H_A: $\mu \neq \mu_0$

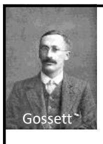
If H_0 is true, then we know that

$$t_{\bar{y}} = \frac{\bar{y} - \mu_0}{s_{\bar{y}}}$$

- Follows a t-distribution with $df = n - 1$
- Reject H_0 if
  - $t_{\bar{y}} > t_{0.05[n-1}$
  - $t_{\bar{y}} < -t_{0.05[n-1]}$

- We reject H_0 if the observation seems `too extreme´ for H_0 to be true
- The paired t-test is an example (in which case $\mu_0 = 0$)

7



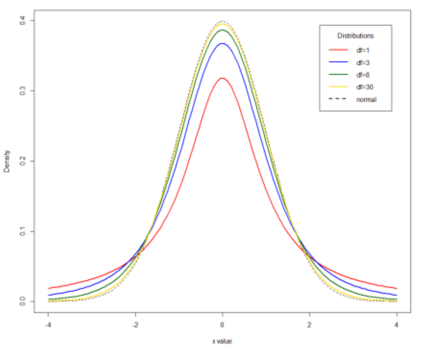
The t-statistic

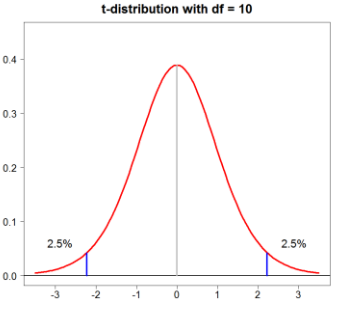
https://www.youtube.com/watch?v=Uv6nGIgZMVw

The t-statistic:

$$t_{\bar{y}} = \frac{\bar{y} - \mu}{s_{\bar{y}}}$$

- is t-distributed with $df = n - 1$
- follows a Student's t distribution if the null hypothesis is supported





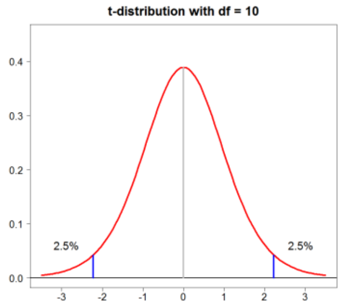
Critical value at level $\alpha = 5\%$: $t_{0.05[10]} = 2.228$

This critical value is for a **two-tailed test**

8

Hypothesis test for comparing two means

Situation: y_1 and y_2 are normal with mean μ_1 and μ_2. The SD σ is unknown but is the same for y_1 and y_2.



Critical value at level $\alpha = 5\%$: $t_{0.05[10]} = 2.228$

This critical value is for a **two-tailed test**

- *Two-tailed test with the hypotheses*
  H_0: $\mu_1 = \mu_2$
  H_A: $\mu_1 \neq \mu_2$
- If H_0 is true, then we know that
  $$t_{\overline{y_1}-\overline{y_2}} = \frac{\overline{y_1} - \overline{y_2}}{s_{\overline{y_1}-\overline{y_2}}}$$
- is t-distributed with $df = n_1 + n_2 - 2$
- The SE of $\overline{y_1} - \overline{y_2}$ is
  - $s_{\overline{y_1}-\overline{y_2}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- Reject H_0 if
  - $t_{\overline{y_1}-\overline{y_2}} > t_{0.05[df]}$
  - $t_{\overline{y_1}-\overline{y_2}} < -t_{0.05[df]}$
- **This is the t-test!**
- Assumptions:
  - Normality
  - Equal variances

9

What motivated Student?

In a letter dated May 12,1907 to a colleague at Guinness in Dublin, William Gosset, who was at Pearson's Biometric Laboratory in London at the time, wrote of his working day:

"… and on other days work at small numbers; a greater toil than I had expected, but I think absolutely necessary if the Brewery is to get all possible benefit from statistical processes."

(in McMullen and Pearson, 1939)

"… and the Experimental Brewery which concerns such things as the connection between the analysis of malt or hops, and the behaviour of beer, and which takes a day to each unit of the experiment, thus limiting the numbers …"

(in Pearson, 1990)

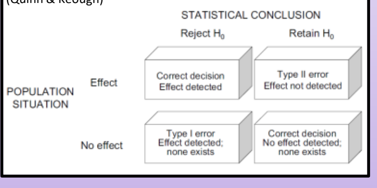
10

Most text from Bellhouse, 2005

Errors in accepting or rejecting the null hypothesis

Type 1 error:
- Reject a true H_0 (false positive)
- $\Pr[type\ I\ error] = \alpha$
- α is called the level of the test

Type 2 error:
- Accept a false H_o (false negative)
- $\Pr[type\ II\ error] = \beta$
- $Power = 1 - \beta = \Pr[reject\ false\ H_0]$



- Power is the probability of actually getting a significance given that some alternative hypothesis is true
- The power of a test depends on which particular alternative hypothesis one considers
- It is bad practice to do tests with too little power (say, less than 50%)
- A good recommendation is to have at least 80% power

Confidence interval for a proportion

A common special case where it is tempting to use the standard formula for the confidence interval of a mean is when one estimates a probability or proportion:

Example:
- Survival: $y = 1$
- Death: $y = 0$
- The sample average estimates the probability of survival
- For very small samples ($n = 3$)
  - $y_1 = 0, y_2 = 1, y_3 = 0$
  - $\bar{y} = \frac{1}{3}, s^2 = \frac{1}{3}, s = 0.577, s_{\bar{y}} = 0.333$
  - $t_{crit} = t_{0.05[2]} = 4.303$
  - $L_1 = 0.333 - 4.303\ x\ 0.333 = -1.10$
  - $L_2 = 0.333 + 4.303\ x\ 0.333 = 1.77$

These confidence intervals are meaningless, since we know that the probability of survival is between 0 and 1. The problem is that the sample average does not follow a normal distribution very well:

- There is a special procedure for getting a more meaningful confidence interval for a proportion, which makes use of the binomial distribution. For the example, the procedure gives the 95% confidence limits
  - $L_1 = 0.0084, L_2 = 0.9057$
- These limits were computed with the binom.test function in R

For large samples sizes and intermediate proportions, the `ordinary´ confidence interval is okay.

Special confidence intervals

*For other statistics than the mean, we need to modify the method of computing confidence intervals. Sometimes, we can use the formula with standard error and critical t-value, but we need a special way to compute the standard error. For instance, for **skew** and **kurtosis** estimated from a sample of size n, we can use the standard errors:*

The SE of skew is:

$$s_{skew} = \frac{6n(n-1)}{(n-2)(n+1)(n+3)} \approx \sqrt{\frac{6}{n}}$$

The SE of kurtosis is:

$$s_{kurtosis} = \frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)} \approx \sqrt{\frac{24}{n}}$$

and is an assumption of df=infinity to get the confidence intervals (the R book p. 352 suggests $df = n - 2$).

For the variance σ^2 there is a special formula:

$$\Pr[X^2_{0.025[n-1]} \leq \frac{(n-1)s^2}{\sigma^2} \leq X^2_{0.975[n-1]}]$$
$$\Pr[\frac{(n-1)s^2}{X^2_{0.975[n-1]}} \leq \sigma^2 \leq \frac{(n-1)s^2}{X^2_{0.025[n-1]}}]$$

- which makes use of the chi-square distribution (p. 288 in the R book)

Calculating σ^2 for our *Pieris napi* sample

- $s^2 = 216.21$
- $n = 102$
- $L_1 = \frac{101x216.2}{130.70} = 167.1$
- $L_2 = \frac{101x216.2}{75.08} = 290.8$

13

SD, SE or confidence interval?

*When should we report data as **mean ± SD**, as **mean ± SE** or as a confidence interval?*

- Give mean ± SD when you want to give an idea about what a distribution looks like
- Give mean ± SE when you want to give an idea of how accurately you have estimated a mean and
  - The variable is normally distributed
  - The distribution of the variable is not `too extreme´ (e.g. very skewed) and the sample size is rather big
- Otherwise give a (correctly computed) confidence interval to give an idea of how accurately you have estimated a parameter
- Using bootstrap for SE and confidence intervals can be recommended

14

When do we use simple hypothesis tests?

A typical situation is a paired comparison: for each unit we have two values and we examine the difference between them

- For the paired t-test to be OK, we need that either
  - the distribution of differences is normal
  - or the sample size is fairly large and the distribution is not very skewed
- Otherwise, a non-parametric test might be OK
  - but there are requirements also for nonparametric tests
- Use a two-tailed test unless you have very good reasons to use a one-tailed test

15

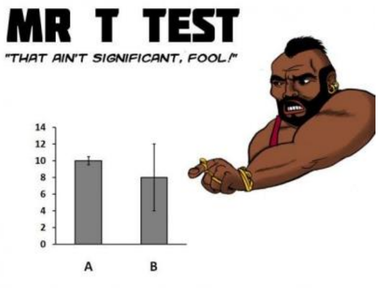
How should we test for a difference between the means of two groups?

We can choose between the ordinary t-test, the Welch t-test (t´-test), a nonparametric test, and a randomization test

- Use the ordinary t-test when the variable is normally distributed in each group and the variance is the same in each group
  - You can test whether the variances are the same, for instance with an *F*-test or with Bartlett's test
- Use the Welch *t*-test if the variables are normally distributed but the variances differ between groups
- With fairly large sample sizes and distributions that are not too skewed you can still use the *t*-tests
- Otherwise, a non-parametric or randomization test might be OK
- Use a two-tailed test unless you have very good reasons to use a one-tailed test

16

Related reading and information



- **Quinn & Keough:** Chapter 2.2 – 2.4 & 3.1 – 3.3
- **Crawley:** Sections 8.1, 8.3, 8.4 and 8.5

17