Data analysis methods in weather and climate research

Dr. David B. Stephenson
Department of Meteorology
University of Reading
www.met.rdg.ac.uk/cag

6. Statistical hypothesis testing

- The basic idea
- A legal example
- The procedure
- What not to do!
- Examples of tests

6. The basic idea

Use data to make a rational decision between two hypotheses about population parameters:

- The alternative (“research”) hypothesis H1
- The null (“chance”) hypothesis H0

Use the data to REJECT H0 in favour of H1

Key idea: try to minimise the number of incorrect decisions

6. Example 1: Do meteorologists have different heights to everyone else??

Does the evidence from our sample of meteorologists show that the population mean height of meteorologists is different to that of everyone else (assume this is known to be 170cm)?

$$H_0 : \mu = \mu_0 = 170$$

$$H_1 : \mu \neq \mu_0$$

6. Example 2: A legal case

H0: suspect is innocent

H1: suspect is guilty

We try to use the evidence/data (e.g. murder weapon) to REJECT H0 rather than to assume H1 and then try to find data that proves innocence.

Example: Did Erik murder his last office mate?

Assume Erik is guilty → then try to find no bloody keyboard in Erik's office

Assume Erik is innocent → then try to find bloody keyboard in Erik's office

6. Hypothesis testing procedure

1. Set up a simple null hypothesis $H_0 : X \sim f(\theta_0)$
2. Specify the "**level of significance**" α that you are prepared to accept (e.g. α=0.05)
3. Calculate the sampling distribution $T \sim f_T(\theta_0)$ of the test statistic
4. Calculate the "p-value" p=Pr(|T|>=t) for the sample value t you measured for T
5. If p<α then reject the null hypothesis (not chance) otherwise don't reject ("data not inconsistent with chance sampling")

6. Sampling distribution

A sample statistic T[X] is distributed with

its *sampling distribution*:

$$T \sim f_T(n,\theta)$$

The sampling distribution depends on:

- Choice of sample statistic;
- Sample size n;
- Parameters of the original distribution $X \sim f_X(\theta)$

6. Example: Mean of normally distributed variable

$$X \sim N(\mu, \sigma^2)$$

$$\Rightarrow \overline{X} \sim N(\mu, \sigma^2 / n)$$

$$E(\overline{X}) = \mu$$

$$Var(\overline{X}) = \sigma^2 / n$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

6. Example: Mean of any random variable

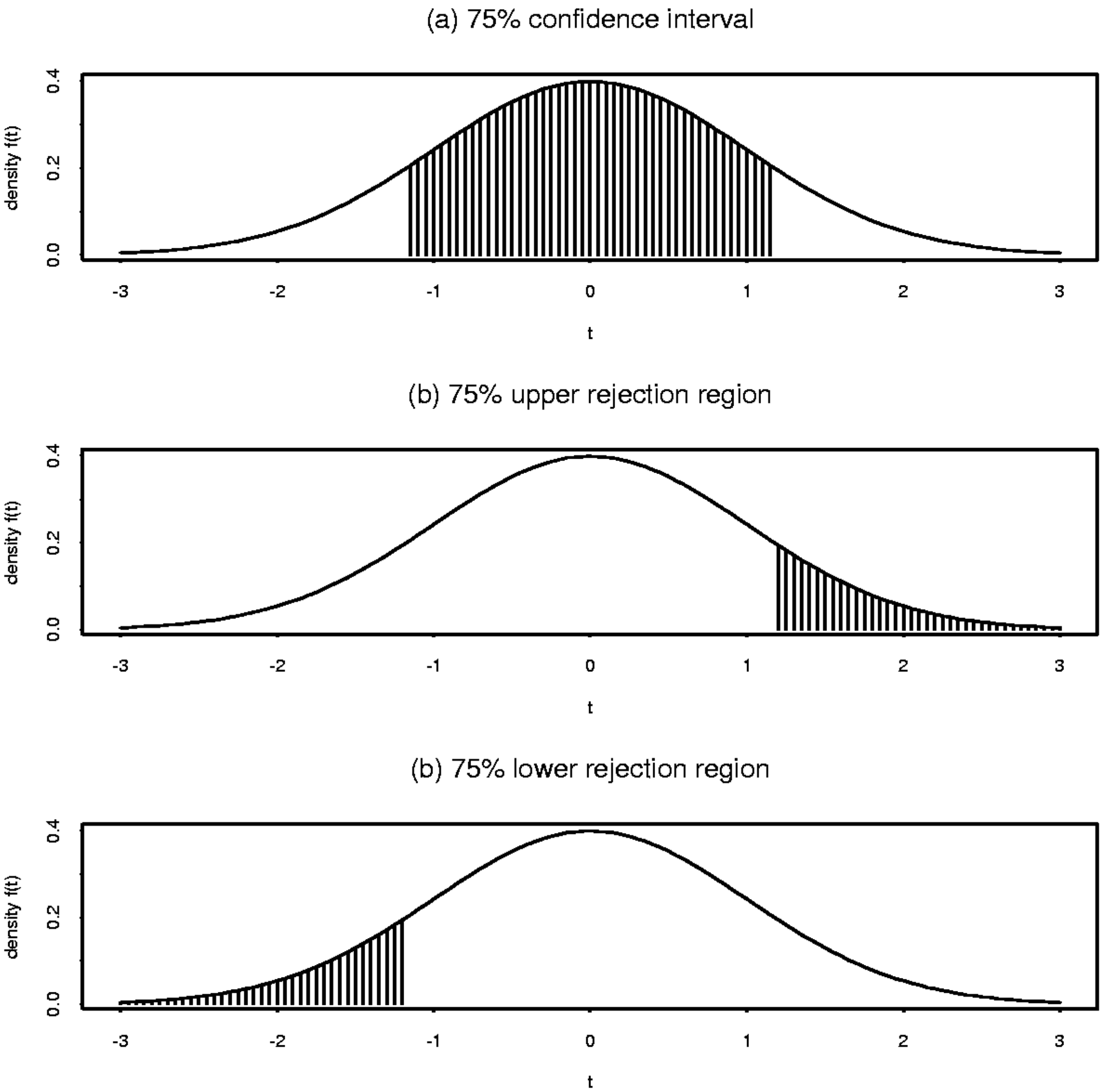
Central Limit Theorem

$$X \sim f_X(\theta) \;\; \textit{and independent}$$

$$\Rightarrow \lim_{n\to\infty} \overline{X} \sim N(\mu_X, \sigma_X^2 / n)$$

This works for ANY f() with finite mean and variance and explains why we see so many variables that are normally distributed e.g. measurement errors due to many random effects.

6. Critical regions of sampling distribution



Example 1: Do meteorologists have different heights?

$$H_0 : \mu = \mu_0 = 170$$

$$H_1 : \mu \neq \mu_0$$

$$H_0 \Rightarrow \overline{X} \sim N(\mu_{0,} \sigma^2 / n)$$

$$\Rightarrow Z = (\overline{X} - \mu_0)/(\sigma / \sqrt{n}) \sim N(0,1)$$

$$\overline{X} = 174.3$$

$$n = 11$$

$$\sigma = 30$$

$$\Rightarrow z = 0.48$$

$$\Rightarrow p = 2(1 - \Phi(z)) = 2(1 - \Phi(0.48)) = 0.63$$

$p > 0.05 \Rightarrow H_0$ can not be rejected at 0.05 level

6. Definitions

Level of significance

= probability of rejecting H0 even if it is true

e.g. probability of convicting an innocent person.

p-value

= probability of finding a chance data sample less consistent with the null hypothesis than the current sample.

e.g. the *burden of evidence*

Type 1 and 2 errors

Decision	H0 True	H1 true
P>alpha Don’t reject H0	<u>Correct non-rejection</u> Rate=1-alpha	<u>Missed rejection</u> Rate=beta Type 2 error
P<=alpha Reject H0	<u>False rejection</u> Rate=alpha Type 1 error	<u>Correct rejection</u> Rate=1-beta =“power” of the test

Type 1 error: rejection of null hypothesis when it is true
Type 2 error: failure to reject null hypothesis when it is false

Legal example:
Type 1 error: Conviction of an innocent person
Type 2 error: Failure to convict a guilty person

6. Bad and good practice

Examples of bad practice:

- “*The results are statistically significant*”
- *“… and are 95% significant”*
- *“Not significant at 0.05 level but results are significant at 0.10 level”*
- *“Some of the samples are significant”*
- *“The null hypothesis can’t be rejected and so is true”*

Good practice:

- State clearly the hypotheses that you are testing
- State which test you are using (e.g. 2-sided t-test on means)
- Give the level of significance and p-value

  e.g. *… statistically significant at the 5% level (p-value 0.044).*
- Don’t quote p-values of zero! Instead of p=0.000 write p<0.001

6. Some commonly used tests

One sample tests: **> # R commands**

- z test on mean (variance known) `> pnorm(…)`
- t test on mean (variance estimated) `> t.test(…)`
- z test for non-zero correlation `> cor.test(…)`

Two sample tests:

- t test for means of unpaired data `> t.test(…)`
- t test for means of paired data `> t.test(…)`
- F test for unequal variances `> var.test(…)`
- z test for unequal correlations `> pnorm(…)`

The "Student" t-test

'Student' in 1908

"Student" was the nom-de-plume of statistician W.S. Gosset who was working at Guinness when he wrote his famous t distribution article.

The t-test is useful when the variance has to be estimated from the data (rather than being known):

$$T = (\bar{X} - \mu)/(s/\sqrt{n}) \sim t_\nu$$

$$\Rightarrow f(t) \propto (1 + t^2)^{-\nu/2}$$

6. More robust tests

It is important to be aware of the distributional assumuptions on which the tests are based. More robust non-parametric tests make less distributional assumptions (but can have the disadvantage of having less power):

Examples:

- Wilcoxon (Mann-Whitney) test on means **`> wilcox.test(…)`**
- Kruskal-Wallis test on means of >2 samples **`> kruskal.test(…)`**
- Spearman correlation **`> cor.test(…,method="spearman")`**

See this book for more non-parametric testing methods:

M. Hollander & D.A. Wolfe (1999), *Nonparametric statistical inference*. New York: John Wiley & Sons.

Summary

- Statistical hypothesis testing uses the data to make a binary decision between 2 hypotheses
- The decision-making fails in two ways: type 1 errors and type 2 errors
- The level of significance controls the rate of type 1 errors and the power of the test determines the rate of type 2 errors
- Statistical tests should be clearly reported!
- Robust (distribution free) tests can also be used if needed

Probability problem

You meet a person who tells you that they have 2 children and that 1 (or more) of them is a girl.

Q1: What is the probability that the other child is also a girl?

The person now tells you that the girl has a very rare name with probability p close to zero.

Q2: Would you revise your probability estimate? If so, what would be your new estimate?