Data Analysis Methods in Weather and Climate Research

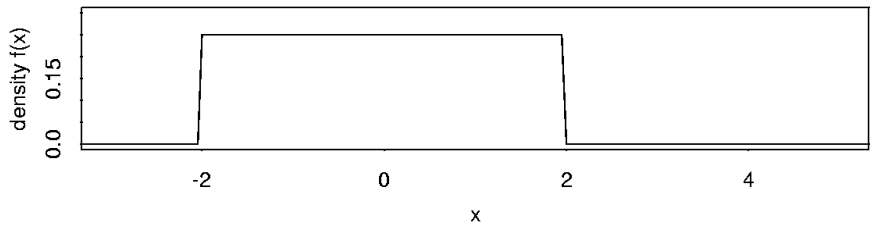
Dr. David B. Stephenson
Climate Analysis Group
Department of Meteorology
University of Reading
Room 3L36

D.B.Stephenson @ reading.ac.uk
www.met.rdg.ac.uk/cag/courses

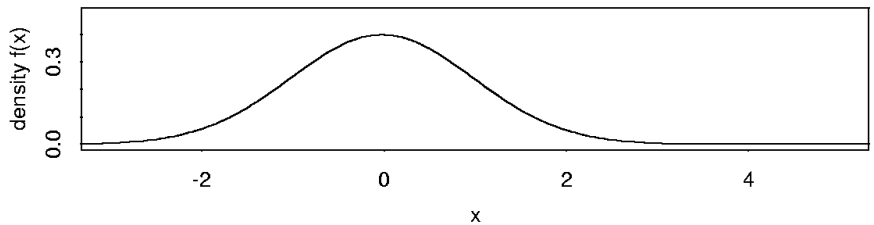
North Male' max altitude 1.8

Continuous distributions

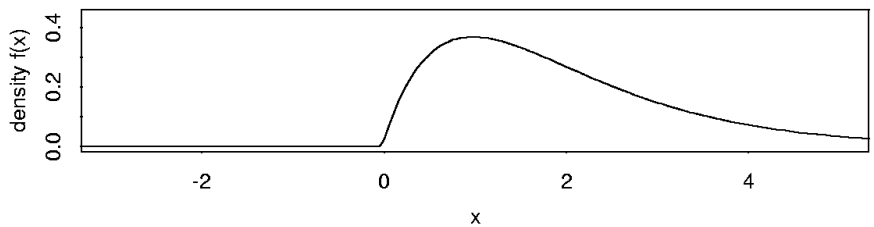
(a) Uniform probability density function



(b) Normal probability density function



(c) Gamma probability density function



5. Parameter estimation

The problem:
Estimate the population parameter(s) $\hat{\theta}$
that give the best fit of the probability model

$$X \sim f(x;\theta)$$

to the observed sample of data.

- **The sampling distribution f(T) of an estimator T(X)**
- **Error bars and Confidence intervals**
- **Types of estimator**
- **Accuracy, bias, and efficiency of estimators**

Sample statistics and estimators

Parameters are estimated by using sample statistics T[X] based on the original random variables. For example, the population mean μ can be estimated by the sample mean $\hat{\mu} = \overline{x}$. Such sample statistics are known as "estimators"

Sampling distribution

A sample statistic T[X] is distributed with a "sampling distribution":

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

Sampling distribution depends on:

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

Example: Mean of normally distributed variable

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

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

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

Central Limit Theorem

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

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

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

“Standard error”

The “standard error” is the standard deviation of a sample statistic:

$$\sigma_T = \sqrt{Var(T)}$$

e.g. for sample mean of normal variables:

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

Interval estimates

Rather than just give a single best estimate of a parameter ("point estimate"), it is more informative to give a likely range of possible values – in other words, an "interval estimate".

The simplest way to do this is to quote the best estimate plus/minus its standard error:

$$T \pm \sigma_T$$

The standard error quantifies the amount of uncertainty due to sampling.

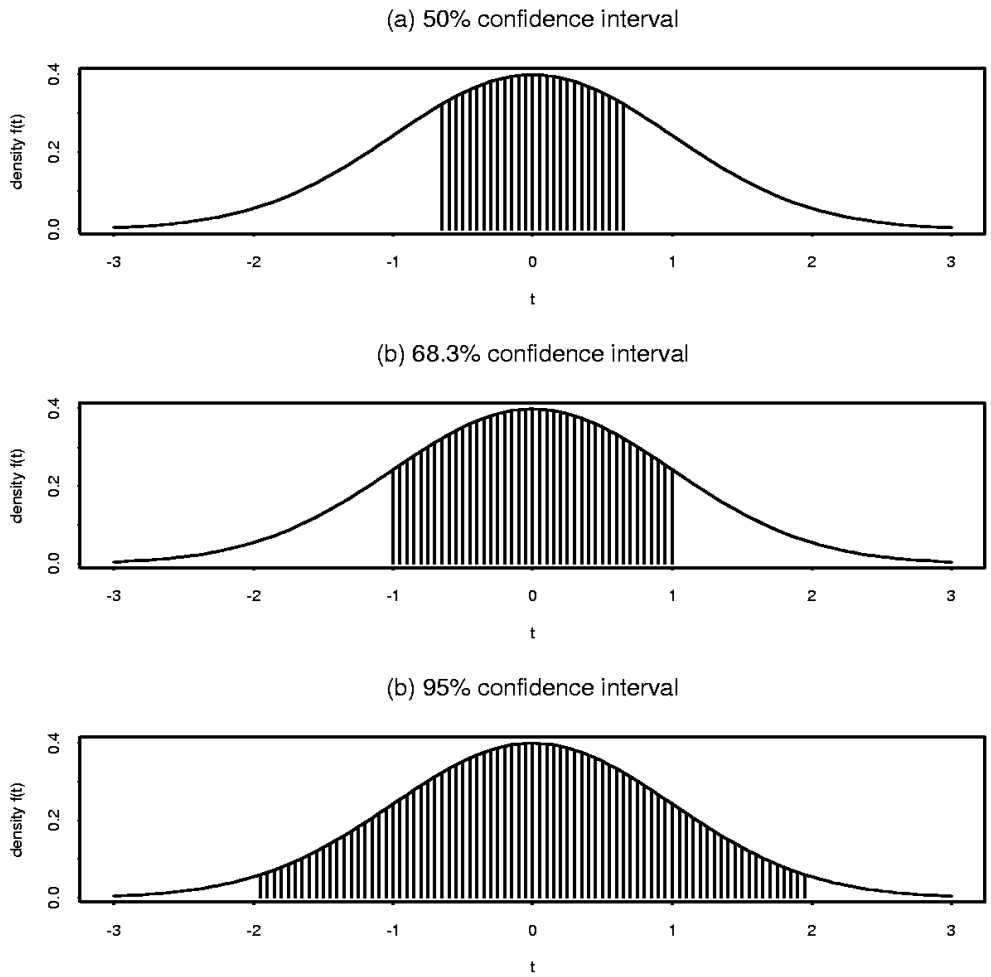
Confidence intervals (C.I.'s)

The (1-alpha)100% confidence interval of a sample statistic T is the interval between the t(alpha/2) and the t(1-alpha/2) quantiles of the sampling distribution.

$$\Pr\{t_{\alpha/2} \le T \le t_{1-\alpha/2}\} = 1-\alpha$$

$$= confidence\ level$$

Confidence intervals of t[x]



Some commonly used C.I.'s

Alpha	1-alpha	Zc	Description
0.50	0.50	0.68	50% C.I. +/- probable error
0.32	0.68	1.00	68% C.I. +/- 1 std. errors
0.05	0.95	1.96~2	95% C.I. ~+/- 2 std. errors
0.001	0.999	3.29	99.9% C.I. ~+/- 3 std. errors

Choice of estimator?

- "Method of moments" – use sample moments e.g. mean, variance, skewness, etc.
- Robust estimation – use rank statistics such as the median, IQR, etc. instead.
- Maximum Likelihood Estimation – choose estimator so that it maximises the likelihood of our data sampling occurring.

Accuracy, bias, and efficiency

The accuracy of an estimator can be quantified as follows:

$$Mean\ Squared\ Error\ E((\hat{\theta}-\theta)^2)$$

$$=(E(\hat{\theta})-\theta)^2\ squared\ "bias"$$

$$+Var(\hat{\theta})"efficiency"$$

There is invariably a trade-off between bias and efficiency.

6. Statistical hypothesis testing

1. The basic idea
2. A legal example
3. The procedure
4. What not to do !
5. Examples of tests

6. The basic idea

Use data to decide between two hypotheses about population parameters:

The “alternative” hypothesis H1

The null (“chance”) hypothesis H0

Try to use data to REJECT H0

Example1: 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?

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

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

6. Example 2: A legal case

H0: suspect is innocent

H1: suspect is guilty

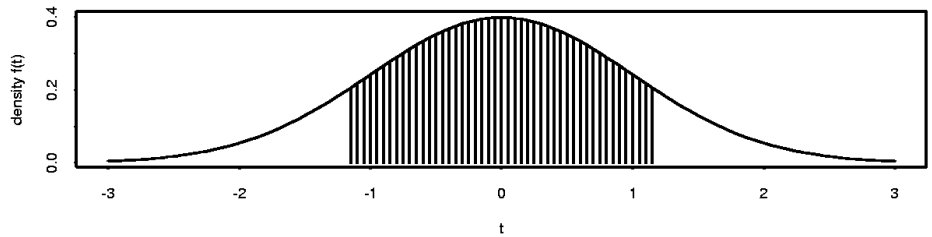
We try to REJECT H0 using available evidence (e.g. fingerprints on the murder weapon) rather than assume H1 and try and prove innocence.

6. Testing procedure

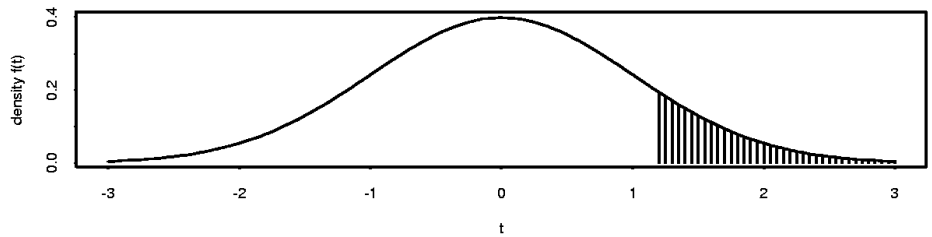
1. Set up a simple null hypothesis $H_0 : X \sim f(\theta_0)$
2. Specify the "level of significance" alpha that you are prepared to accept (e.g. alpha=0.05)
3. Calculate the sampling distribution $T \sim f_T(\theta_0)$ of the test statistic
4. Calculate the "p-value" p=Pr(|T|>=t) for the sample value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

Critical regions of sampling distribution

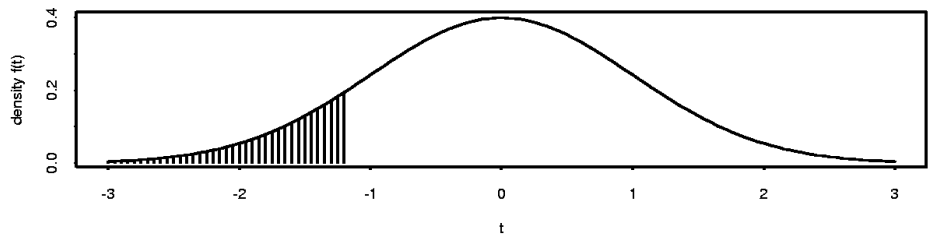
(a) 75% confidence interval



(b) 75% upper rejection region



(b) 75% lower rejection region



Example1: Do meteorologists have different heights to everyone else??

$$H_0 : \mu = \mu_0 = 170$$
$$H_1 : \mu \neq \mu_0$$

$$H_0 \Rightarrow \overline{X} \sim N(\mu_0, \sigma^2 / n)$$
$$\Rightarrow Z = (\overline{X} - \mu_0) / (\sigma / \sqrt{n}) \sim N(0,1)$$

$$\overline{X} = 174.3$$
$$n = 11$$
$$\sigma = 30$$
$$\Rightarrow z = 0.48$$
$$\Rightarrow p = 2(1 - \Phi(z)) = 2(1 - \Phi(0.48)) = 0.63$$
$$p > 0.05 \Rightarrow H_0 \text{ can not be rejected at 0.05 level}$$

6. Definitions

Level of significance alpha

= probability of rejecting H0 even if it is true

e.g. probability of convicting an innocent person.

P-value = Probability of finding a data sample less consistent with the null hypothesis than the current sample.

Type 1 and 2 errors

	H0 True	H1 true
P>alpha Don't reject H0	Correct non-rejection Rate=1-alpha	Missed rejection Rate=beta Type 2 error
P<=alpha Reject H0	False rejection Rate=alpha Type 1 error	Correct rejection Rate=1-beta ="power" of the test

6. 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 can't be rejected and so is true"

6. Common tests

- One sample z test on mean (variance known)
- One sample t test on mean (variance estimated)
- One sample z test for non-zero correlation
- Two sample t test for means of unpaired data
- Two sample t test for means of paired data
- Two sample F test for unequal variances
- Two sample Z test for unequal correlations

“Student” aka W.S. Gosset

'Student' in 1908

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

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

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