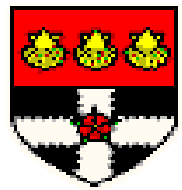
Data analysis methods in weather and climate research

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

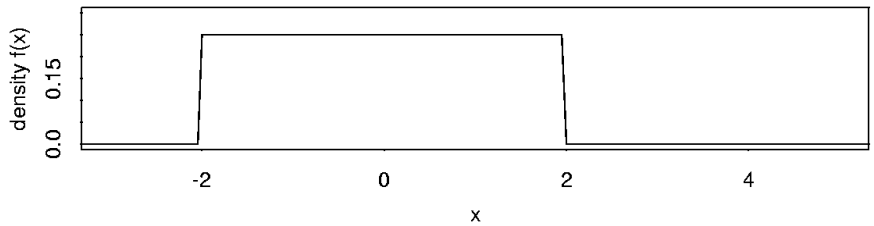


5. Parameter estimation

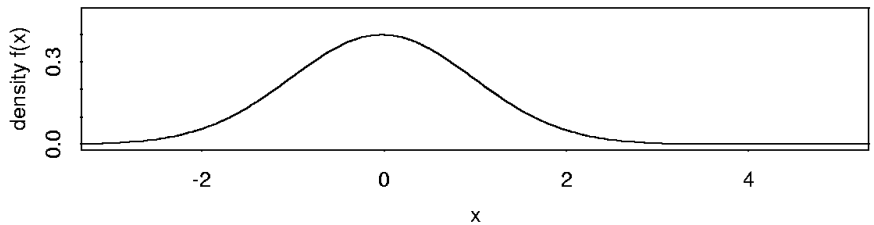
- ??
- ??
- ??

Continuous distributions

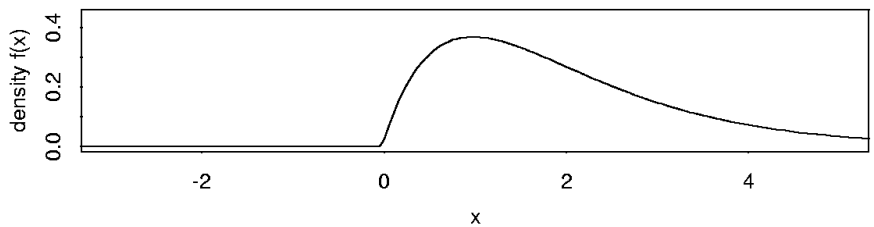
(a) Uniform probability density function



(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 can be estimated by the sample mean . Such sample statistics are known as “estimators”

μ

$$\hat{\mu} = \overline{x}$$

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 variables

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

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

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

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:

e.g. for sample mean of normal variables:

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

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

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:

$$T \pm \sigma_T$$

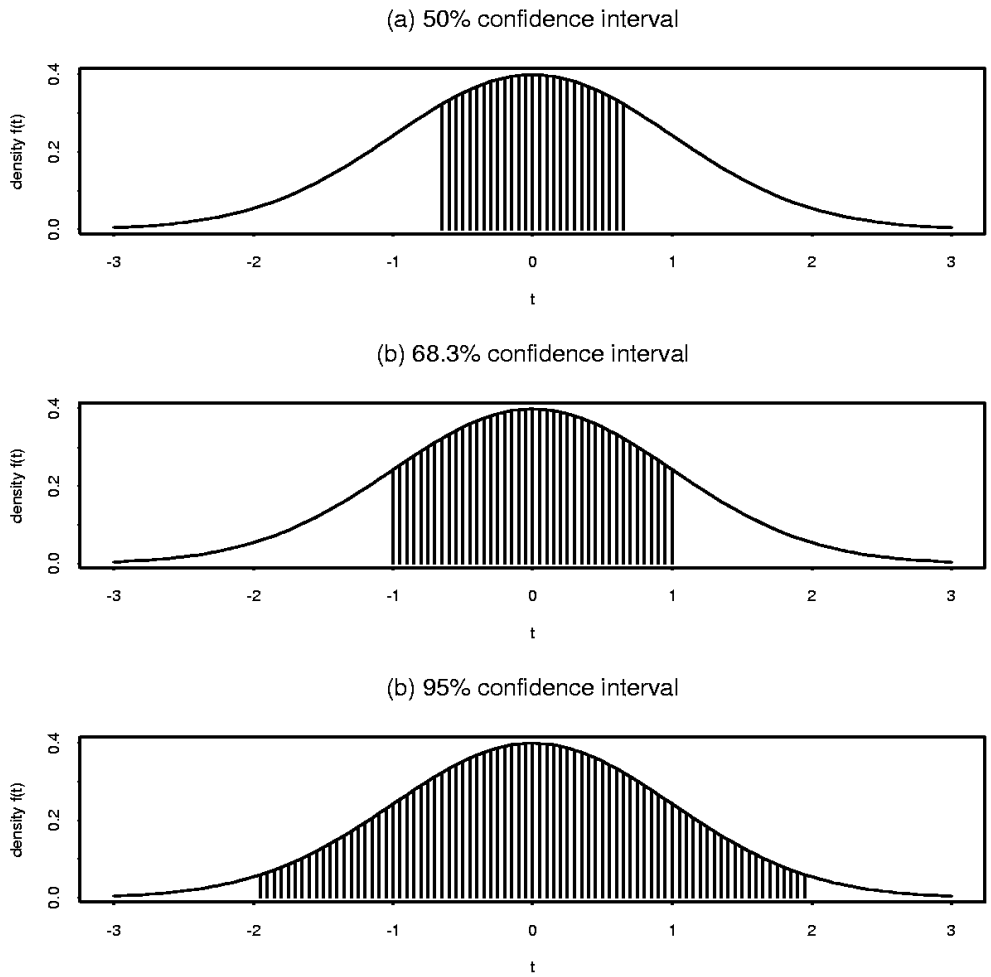
The standard error quantifies the amount of uncertainty due to sampling.

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.