Data analysis methods in weather and climate research

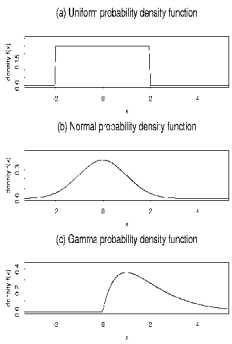
Dr. David B. Stephenson
Department of Meteorology
University of Reading
www.met.rdg.ac.uk/cag

5. Parameter estimation
- Fitting probability models
- The sampling distribution of an estimator
- Error bars and Confidence intervals
- Types of estimator
- Accuracy, bias, and efficiency of estimators

5. Probability modelling in 6 steps

Modelling strategy:

1. Explore the data sample (EDA)
2. Identify a suitable distribution
3. Fit the distribution to the data by *estimation* of the parameters
4. Check the goodness-of-fit
5. Make out-of-sample predictions
6. Go back to 1 or 2 if needed



5. Parameter estimation

Estimate the values of population parameter(s) θ that give the *best fit* of the probability model

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

to the observed sample of data.

Note: we fit the model to the data NOT the data to the model!

5. Sample statistics and estimators

Parameters are estimated by using sample statistics T[X] of the original random variables. Such sample statistics are known as "estimators".

For example, the population mean parameter μ in the Normal distribution $N(\mu,\sigma^2)$ can be estimated by the sample mean $\hat{\mu} = \bar{x}$. This is known as a *point estimate*.

The hat symbol denotes "estimate of".

5. 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 the standard deviation in this estimate:

$$T \pm \sigma_T$$

The standard deviation quantifies the amount of uncertainty in the estimate caused by sampling.

5. Sampling distribution

Each sample statistic T[X] is distributed with its own "*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 $X \sim f_X(\theta)$

5. Mean of iid normally distributed variables

For iid ("independent and identically distributed") normally distributed random variables:

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

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

Sampling distribution of sample mean

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

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

5. Central Limit Theorem

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

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

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

5. Definition of *standard error*

The "standard error" is the standard deviation of the sample statistic. i.e.

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

e.g. for sample mean of iid variables:

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

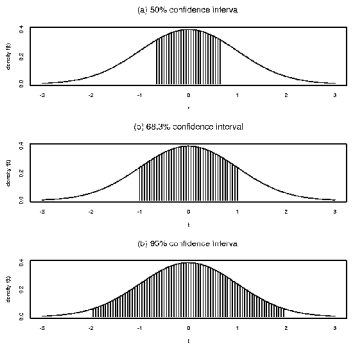
5. Confidence intervals (C.I.'s)

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

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

There is probability $1-\alpha$ that the interval will overlap the true value.

5. Examples of confidence intervals



5. 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

5. Choice of estimator

Many ways to choose the estimators such as:

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.

5. 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.

5. Summary

- Model is fit to the data by using sample statistics (estimators) to estimate the true model parameters
- Interval estimates give a range of probable values rather than a single point estimate
- Each estimator has its own probability distribution known as a sampling distribution
- The standard deviation of the estimator is known as the standard error
- The sampling distribution can be used to construct confidence intervals in which the true value is most likely to be found.
- There are several methods for estimating parameters: moment method, robust estimation, maximum likelihood estimation.
- Different estimators have different accuracies, bias, and efficiency.