University of BRISTOL

spring semester 2019

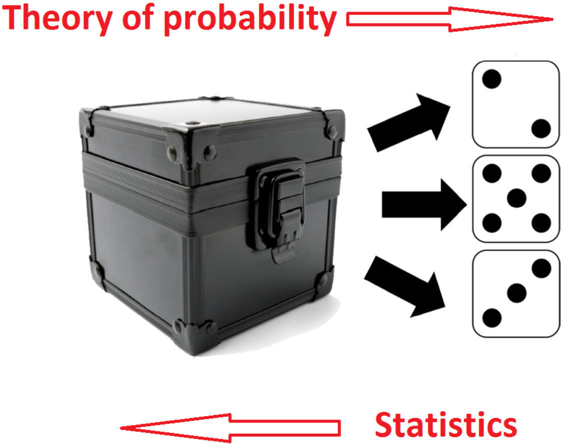
Estimating Parameters (1) - Lecture 1

APPLIED STATISTICS - EMAT 30007

Nikolai Bode and Ksenia Shalonova

Department Of Engineering Mathematics

What is Statistics?

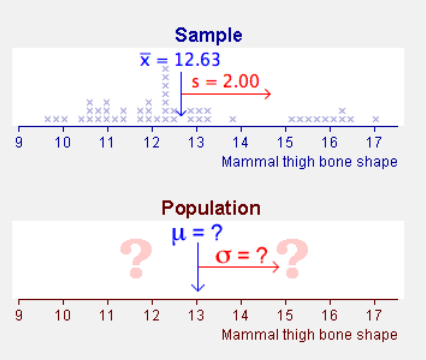


Theory of probability

Statistics

Population and samples

In statistical analysis we want to estimate a population parameter from a random sample. This is called inference about the parameter.



An archaeologist has discovered 47 bones that are believed to be thigh bones of a certain mammal. The measurements are the ratio of length to breadth.

The archaeologist is interested in the mean and standard deviation of the thigh bone shapes for this type of mammal in general in order to assess the type of mammal.

What is parameter estimation?

Distribution	Parameters
normal	mean, standard deviation
Poison	mean
Sudent's t	degrees of freedom
Weibull	shape
lognormal	mean, standard deviation
exponential	rate
uniform	minimum and maximum
logistic	location, scale
etc.	

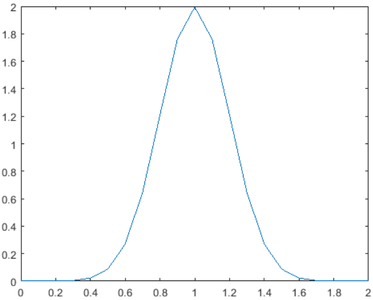
Random samples are used to provide information about parameters (θ) in an underlying population distribution.

Rather than estimating the full shape of the underlying distribution, we usually focus on one or two parameters.

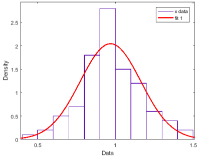
For the gallery of probability distributions and their parameters visit:
`http://www.itl.nist.gov/div898/handbook/eda/section3/eda366.htm`

Using Monte Carlo simulations for showing variability of sample means

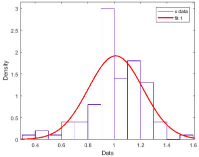
The Monte Carlo method uses repeated random sampling to generate simulated data to use with a mathematical model. Do not mix up with bootstrap! (Lecture 5)



Population Normal Distribution with population mean $\mu = 1$ and population standard deviation $\sigma = 0.2$

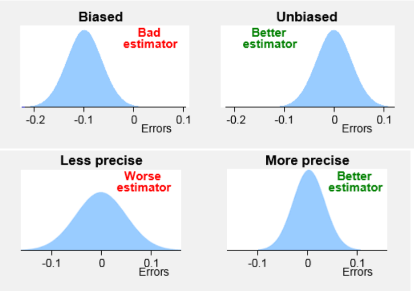


random sample (n=100) with sample mean $\bar{x} = 0.9687$



random sample (n=100) with sample mean $\bar{x} = 1.01128$

What is a good parameter estimator?



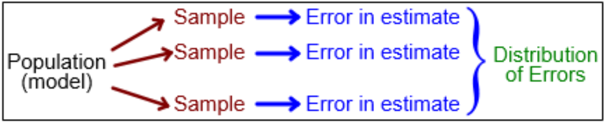
We want the error distribution to be centered on zero. Such an estimator is called unbiased. The biased estimator shown on the left tends to have negative errors, i.e.it usually underestimates the parameter that is being estimated.

We also want error distribution to be tightly concentrated on zero, i.e. to have a small spread.

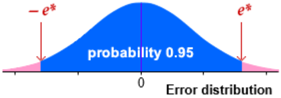
A good estimator should have a small bias and small standard error. These two criteria can be combined with into single value called the estimator's mean squared error. Most estimators that we will consider are unbiased, the spread of the error distribution is most important.

Estimating parameters - major points

- From a single small sample, there is a lot of uncertainty about the population distribution.
- When a population characteristic is estimated from a sample, there is usually a sampling error.
- The larger the sample size, the smaller the sampling error.

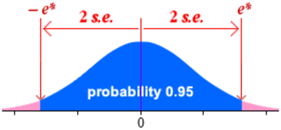


Error distribution of the mean and confidence interval



$\texttt{prob}(-e* < error < e*) = 0.95$

Mean error is normally distributed: Normal$(0, \frac{\sigma}{\sqrt{n}})$.



prob(estimate - 2SE<parameter<estimate + 2SE) = 0.95

95% confidence interval from standard error

Standard error of the mean estimator

Definition (Standard error)

The standard error (SE) of an estimator $\hat{\theta}$ of a parameter θ is defined to be its standard deviation.

Standard error of the mean:

- Bias (μ error) $= 0$, i.e. $E(\hat{\theta}) = \theta$
- When population standard deviation is known: SE = $\frac{\sigma}{\sqrt{n}}$
- When population standard deviation is unknown: SE = $\frac{s}{\sqrt{n}}$
- How to work out standard deviation of the sample s?

Do not confuse SD (sample standard deviation) and SE (standard deviation of the sample mean $\bar{x}$).

Confidence interval for μ with known σ

We can be $(1-\alpha)100\%$ confident that the estimate for μ will be in the intervals shown below. For 95% confidence interval $\alpha = 0.025$.

Confidence interval for μ when σ is known: $\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

Common `exact` values of $z_{\alpha/2}$ with critical values from normal distribution:

90%	1.645
95%	1.96
99%	2.575

Confidence interval for μ when σ is unknown

If we simply replace σ by its sample variance the confidence level will be lower than 95%. When the sample size is large, the confidence level is close to 95% but the confidence level can be much lower if the sample size is small.

Critical value comes from the Students `t distribution`. The value of $t_{\alpha/2}$ depends on the sample size through the use of degrees of freedom.
Confidence interval for μ when σ is unknown (use sample variance instead)

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

Sample size for estimating mean

- Consider estimation of a population mean, μ, from a random sample of size n. A confidence interval will be of the form $\bar{x} \mp t_{\alpha/2}\frac{s}{\sqrt{n}}$. If we want our estimate to be within k of μ, then we need n to be large enough so that $t_{\alpha/2}\frac{s}{\sqrt{n}} < k$.
- For 95% confidence interval if n is reasonably large the `t-value` in the inequality will be approximately 1.96: $1.96\frac{s}{\sqrt{n}} < k$ that can be re-written as $n > (\frac{1.96s}{k})^2$
- In practice, it is best to increase n a little over this value in case the sample standard deviation was wrongly guessed.

Example

If we expect that a particular type of measurement will have a standard deviation of about 8, and we want to estimate its mean, μ, to within 2 of its correct value with probability 0.95, the sample size should be... $n > (\frac{1.96 \times 8}{2})^2 = 61.5$
This suggests a sample size of at least 62. The more accurate trial-and-error method using a `t-value` would give a sample size of 64.

Confidence interval for proportion (1)

The sample proportion of successes is denoted by $\hat{p}$ and is an estimate of p.

The estimation error = $\hat{p} - p$.

$$\hat{p} = \frac{number\, of\, succeses\, in\, sample}{sample\, size}$$

Distribution of proportion

$\mu_{\hat{p}} = p$

$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

Distribution of estimation error

$bias = \mu_{error} = 0$

standard error = $\sigma_{error} = \sqrt{\frac{p(1-p)}{n}}$

Standard error from data

$bias = \mu_{error} = 0$

standard error = $\sigma_{error} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Normal approximation to error distribution

error = $\hat{p} - p = normal(0, \sqrt{\frac{p(1-p)}{n}})$

Confidence interval

A 95% confidence interval is:

$\hat{p} \mp 2 \times \sqrt{\frac{p(1-p)}{n}}$

Confidence interval for proportion (2)

Example

In a random sample of $n = 36$ values, there were $x = 17$ successes. We estimate the population proportion p with $p = 17/36 = 0.472$. A 95% confidence interval for p is 0.472 ± 0.166
We are therefore 95% confident that the population proportion of successes is between 30.6% and 63.8%. A sample size of $n = 36$ is clearly too small to give a very accurate estimate.

If the sample size n is small or p is close to either 0 or 1, this normal approximation is inaccurate and the confidence level for the interval can be considerably less than 95%.

Classical theory recommends to use the confidence interval for p only when
$n > 30$
$np > 5$
$n(1 - p) > 5$

Z-value or t-value?

If you know the variance of the population, then you should use the `Z-value` from normal distribution.

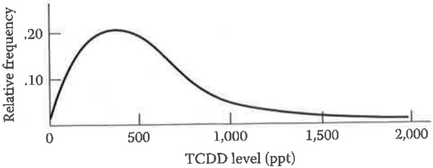
If you don't know the variance of the population or the population distribution is non-normal, then you should *formally* always use the `t-value`.

For most non-normal population distributions, the distribution of the sample mean becomes close to normal when the sample size increases (Central Limit Theorem).

Even for relatively small samples, the distributions are virtually the same. Therefore, it is common to approximate the `t-distribution` using the `normal distribution` for sufficiently large samples (e.g. $n > 30$).

Exercises (1) Sampling distribution

The National Institute for Occupational Safety and Health evaluated the level of exposure of workers to the chemical dioxin TCDD. The distribution of TCDD levels in parts per trillion (ppt) of production workers in New jersey chemical plant had a mean of 293 ppt and a standard deviation of 847 ppt. A graph of the distribution is shown below.



A random sample of 50 workers is selected at the New Jersey plant.

1. Find the mean and standard deviation of the sampling distribution.
2. Sketch the graph of the sampling distribution.

Exercises (2)

1. A pharmaceutical company routinely tests products to check that the concentration of an active ingredient is within tight limits. The results from one type of analysis vary around the true concentration with standard deviation $\sigma = 0.0068$ grams per litre. Six repeated analyses of the same product were conducted and $\bar{x} = 0.4234$ and $s = 0.4234$ grams. Find a 95% confidence interval for the true concentration of active ingredient in this product.

2. The annual rainfall in a region has been recorded in each of the last $n = 25$ years. The standard deviation of the 25 annual rainfalls is 10 mm. Approximately 95% of future years will have annual rainfall that is within $<$ .... $>$ mm.

3. The call centre for a bank samples $n = 38$ incoming phone calls and records the time taken to answer each. The statistics for call times for one day is the following: $\bar{x} = 99.2$ and $s = 24.3$ sec. Find a 95% confidence interval for the mean call time of all such calls.

Tolerance intervals for normal populations (1)

QUALITY APPROVED CONTROL

Definition (Tolerance interval)

A 100(1-α)% tolerance interval for 100γ% of the measurements in a normal population is given by $\bar{x} \pm Ks$ where K is a tolerance factor. Tolerance limits are the endpoints of the tolerance interval.

Do NOT mix up with confidence intervals! We focus on γ (a certain percentage of measurements) rather than on a population parameter.

What would be the value for the tolerance factor if we knew μ and σ?
Otherwise, tolerance factor K depends on the level of confidence 100(1-α)%, γ and the sample size n.

Tolerance intervals for normal populations (2)

Example

A corporation manufactures field rifles. To monitor the process, an inspector randomly selected fifty firing pins from the production line. The sample mean $\bar{x}$ for all observations is 0.9958 inch and standard deviation s is 0.0333. Assume that the distribution of pin lengths is normal. Find a 95% tolerance interval for 90% of the firing pin lengths. Interpret the result.

Given n=50, γ=0.9 and α=0.05, work out K (you can either use a special table or MATLAB function). K=1.996. The 95% tolerance interval is (0.9293, 1.0623).

Approximately 95 out of every 100 similarly constructed tolerance intervals will contain 90% of the firing pin lengths in the population.