Probability and Statistics

Kristel Van Steen, PhD[2]

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

kristel.vansteen@ulg.ac.be

CHAPTER 5: PARAMETER ESTIMATION

1 Estimation Methods

1.1 Introduction

1.2 Estimation by the Method of Moments

1.3 Estimation by the Method of Maximum Likelihood

2 Properties of Estimators

2.1 Introduction

2.2 Unbiased

2.3 Efficiency

2.4 Consistency

3 Confidence Intervals

3.1 Introduction – understanding the concept

3.2 Finding confidence intervals in practice

Pivotal method

3.3 One-sample problems

Confidence Intervals for σ^2

Confidence Intervals for μ

3.4 Two-sample problems

Confidence Interval for σ_1^2/σ_2^2

Confidence Interval for $\mu_1 - \mu_2$

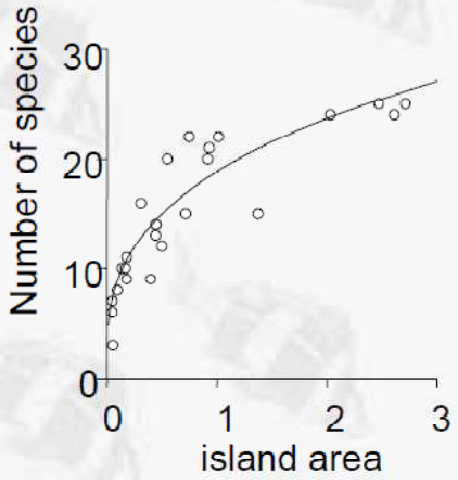
3.5 Summary

1 Estimation Methods

1.1 Introduction

Aims of biological research

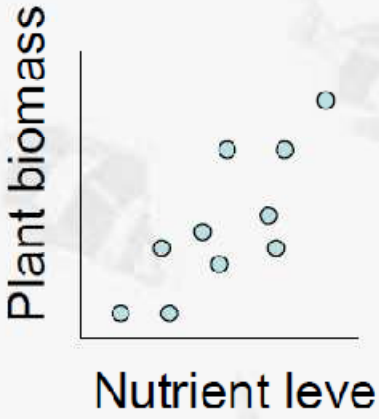
- Describe patterns
  - more species on bigger islands
- Develop predictive models, i.e. prediction
  - species number = α * (area)$^{\beta}$
- Determine cause-effect relationships, i.e. explanation
  - does area cause species number?
  - other factors (perimeter, habitat complexity)?



(Quinn and Keough 2002)

Association

- Measured statistically by correlations and associations
  - plant biomass in plots correlated with soil nutrient levels
  - dispersal mechanism associated with mode of regeneration for plant species



	Regeneration	
	Seed	Veg
Dispersal		
Ant	25	9
Vert	6	21

(Quinn and Keough 2002)

Inferring cause from correlation

- Aim:
  - which cause determines which effect?
- Correlation between one potential cause and one potential effect
- Problems of inferring causality
  - cause ↔ effect
  - multiple effects and alternate causes
- Can't easily rule out confounding

(Quinn and Keough 2002)

Cause versus effect

- There is a correlation between the number of roads built in Europe and the number of children born in the United States.
  - Does that mean that if we want fewer children in the U.S., we should stop building so many roads in Europe?
  - Or, does it mean that if we don't have enough roads in Europe, we should encourage U.S. citizens to have more babies?
  - Of course not. (At least, I hope not).
- While there is a relationship between the number of roads built and the number of babies, we don't believe that the relationship is a causal one.

Cause versus effect

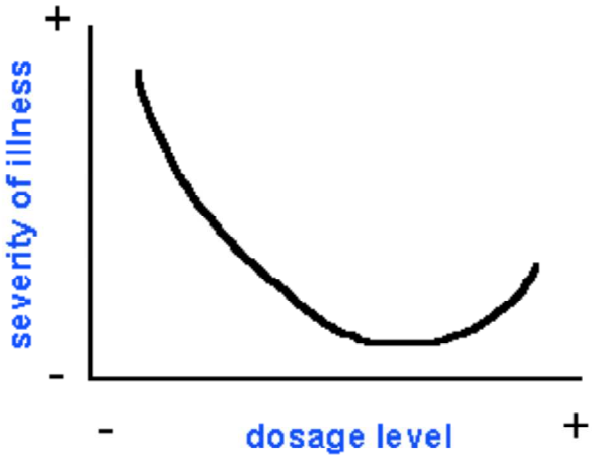
- This leads to consideration of what is often termed the ***third variable problem***.
- In the example above, it may be that there is a third variable that is causing both the building of roads and the birthrate that is causing the correlation we observe.
- For instance, perhaps the general world economy is responsible for both. When the economy is good more roads are built in Europe and more children are born in the U.S.
- The key lesson here is that you have to be careful when you interpret correlations

Cause versus effect

- Likewise, if you observe a correlation between the number of hours students use the computer to study and their grade point averages (with high computer users getting higher grades), you *cannot* assume that the relationship is *causal*: that computer use improves grades.
- In this case, the third variable might be socioeconomic status -- richer students who have greater resources at their disposal tend to both use computers and do better in their grades.
- It's the resources that drives both use and grades, not computer use that causes the change in the grade point average.

Types of relationships

- No relationship, positive relationship, negative relationship
- These are the simplest types of relationships we might typically estimate in research.
- The pattern of a relationship can be more complex than this, and one can try to find the most appropriate "model" to capture the observed pattern.



Natural experiments

- Use pre-existing or naturally occurring treatment groups
  - compare growth of species A in areas with and without species B to test for competition
  - compare photosynthesis of a species at extremes of range to test climate effects
- Correlation with only two levels of possible cause
  - can't easily rule out alternative explanations (i.e. confounding)

(Quinn and Keough 2002)

Manipulative experiments

- Controlled manipulation rules out some alternative explanations (causes)
  - appropriate controls, replication and randomisation reduce likelihood of confounding
- Multiple possible causes and their interactions can be investigated
  - manipulate two or more factors
- Problems with manipulative experiments
  - usually small spatial and temporal scales (relevance?)
  - controls for artefacts not always possible

(Quinn and Keough 2002)

Statistical inference

- Uncertainty caused by
  - sampling from populations
  - measurement error
- Conclusions about causes based on data
  - e.g. what causes variability in our data?
  - e.g. what causes these two means to be different?
- Statistical inference
  - probabilistic conclusions from our sample data addressing question (hypothesis) of interest

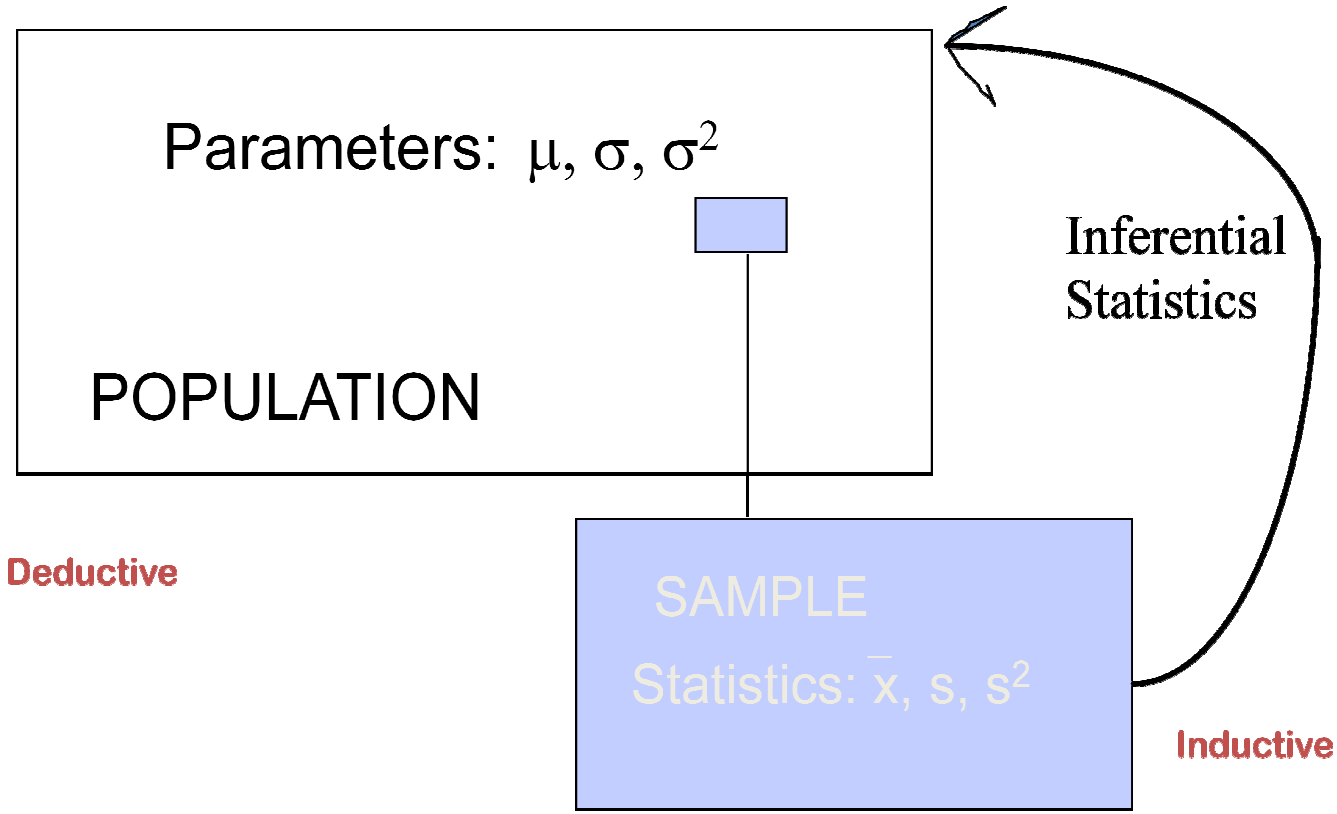
(Quinn and Keough 2002)

Estimating parameters

- Parameters of population
  - mean, variance, regression slope, effects of treatments etc.
- Use sample data to estimate most likely values of those parameters
  - sample statistics estimate population parameters
- Calculate confidence in those parameter estimates
  - standard errors and confidence intervals

(Quinn and Keough 2002)

Estimating parameters



Using "statistics" (derived from samples) to do the job

Suppose that we have a random sample $X_1, X_2, \ldots, X_n$ from a distribution with mean μ and variance σ^2.

1. $\overline{X} = \sum_{i=1}^{n} X_i/n$ is called the **sample mean.**
2. $S^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2/(n-1)$ is called the **sample variance.**
3. $S = \sqrt{S^2}$ is called the sample standard deviation.
4. $M_r = \sum_{i=1}^{n} X_i^r/n$ is called the **rth sample moment** about the origin.

Point and interval estimates

- Point estimate
  - single value estimate of parameter, e.g. $\bar{y}$ is point estimate of μ, s is point estimate of σ
- Interval estimate
  - range within which parameter lies known with some degree of confidence, e.g. 95% confidence interval is interval estimate of μ

(Quinn and Keough 2002)

Testing hypotheses – Chapter 6

- Specific research hypothesis
- Null hypothesis about population parameters
  - population parameter equals zero
  - no effect of predictor/treatment/groups
- Some probability statement about truth or otherwise of hypotheses
  - evidence for or against null hypotheses

(Quinn and Keough 2002)

There are several inferential frameworks

- Each of these offer their own machinery and set of tools to make inferences
- Recall (Chapter 4):

  - For classical (traditional) analysis, the sequence is
    Problem => Data => Model => Analysis => Conclusions
  - ~~For EDA, the sequence is~~
    ~~Problem => Data => Analysis => Model => Conclusions~~
  - For Bayesian, the sequence is
    Problem => Data => Model => Prior Distribution
    => Analysis => Conclusions

Inferential framework 1: parametric (frequentist) analysis

"The data collection is followed by the imposition of a model (normality, linearity, etc.) and the analysis, estimation, and testing that follows are focused on the parameters of that model."

Both systematic and random components are represented by a mathematical model and the model is a function of parameters which are estimated from the data. For example

$$y_{ij} = \beta_0 + \beta_1 x_i + \epsilon_{ij} \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$

is a parametric model where the parameters are

- the coefficients of the systematic model, β_0, β_1
- the variance of the random model, σ^2.

A rough description of frequentist methods is that population values of the parameters are unknown and based on a sample (x, y) we get estimates of the true, but unknown values. These are denoted as $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\sigma}^2$ in this case.

Inferential framework 1: parametric (frequentist) analysis

- Traditional statistical analysis
- Estimation of parameters
  - least squares or maximum likelihood
- Hypothesis testing
  - based on *P* values
  - null versus alternative hypotheses
  - test statistic *cf.* theoretical distribution (e.g. *t*, *F*, χ^2)
- Assumes data follow specific distribution
  - normal, log-normal, poisson etc.
- Frequentist interpretation of probability

(Quinn and Keough 2002)

Inferential framework 1: Bayesian analysis

"For a **Bayesian analysis**, the analyst attempts to incorporate scientific/engineering knowledge/expertise into the analysis by imposing a data-independent distribution on the parameters of the selected model; the analysis thus consists of formally combining both the prior distribution on the parameters and the collected data to jointly make inferences and/or test assumptions about the model parameters."

Or formulated differently (making links to "Bayesian theorem" and conditionality)

> Whereas in frequentist inference the data are considered a random sample and the parameters fixed, Bayesian statistics regards the data as fixed and the parameters as random samples. The exercise is that given the data, what are the distributions of the parameters such that the observed sample from those distributions could give rise to the observed data.

Inferential framework 1: Bayesian analysis

- Treats parameters as random variables with probability distributions
- Estimation of parameters
  - based on posterior probability distributions
  - usually relies on complex resampling procedures
- Hypothesis testing
  - possible but often not used in Bayesian stats
- Subjective interpretation of probability
- Must specify distributions of parameters
- Inclusion of prior information about parameters
  - prior probability distributions

(Quinn and Keough 2002)

Inferential framework 3: non-parametric analysis

This philosophy does not assume that a mathematical form (with parameters) should be imposed on the data and the model is determined by the data themselves. The techniques include

- premutation tests, bootstrap, Kolmogorov-Smirnov tests etc.
- Kernel density estimation, kernel regression, smoothing splines etc.

This seems a good idea to not impose any predetermined mathematical form on the data. However, the "limitations" are

- the data are not summarized by parameters and so interpretation of the data requires whole curves etc. There is not a ready formula to plug in values to derive estimates.

- Requires sound computing skills and numerical methods.
- The statistical method may be appropriate only when there is sufficient data to reliably indicate associations etc. without the assistance of a parametric model.

Monte Carlo (resampling) analysis

- Randomly resample or reshuffle sample data
  - no comparison to theoretical distribution
- Estimation of parameters
  - jackknife, bootstrap
- Hypothesis testing
  - randomisation test
  - simply evaluates null
- Frequentist interpretation of probability
- Less restrictive assumptions *cf*. parametric
- Limited to relatively simple hypotheses
  - simple univariate analyses

(Quinn and Keough 2002)

1.2 Estimation by the Method of Moments

Recall that, for a random variable X, the rth moment about the origin is $\mu'_r = \mathrm{E}(X^r)$ and that for a random sample $X_1, X_2, \ldots, X_n$, the rth sample moment about the origin is defined by

$$M_r = \sum_{i=1}^{n} X_i^r / n, \quad r = 1,\ 2,\ 3,\ \ldots$$

and its observed value is denoted by

$$m_r = \sum_{i=1}^{n} x_i^r / n \ .$$

Note that the first sample moment is just the sample mean, $\overline{X}$.

We will first prove a property of sample moments.

Theorem
Let $X_1, X_2, \ldots, X_n$ be a random sample of X. Then

$$E(M_r) = \mu'_r, \quad r = 1,\ 2,\ 3,\ \ldots$$

Proof

$$E(M_r) = \frac{1}{n} E\left(\sum_{i=1}^{n} X_i^r\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^r) = \frac{1}{n}\sum_{i=1}^{n} \mu'_r = \mu'_r \ .$$

This theorem provides the motivation for estimation by the method of moments (with the estimator being referred to as the method of moments estimator or MME). The sample moments, M_1, M_2, ..., are random variables whose means are μ'_1, μ'_2, ... Since the population moments depend on the parameters of the distribution, estimating them by the sample moments leads to estimation of the parameters.

We will consider this method of estimation by means of 2 examples, then state the general procedure

Example 1

Given $X_1, X_2, \ldots, X_n$ is a random sample from a $U(0, \theta)$ distribution, find the method of moments estimator (MME) of θ.

Solution: Now, for the uniform distribution ($f(x) = \frac{1}{\theta} I_{[0,\theta]}(x)$),

$$\begin{aligned} \mu = E(X) &= \int_0^\theta x \times \frac{1}{\theta} dx \\ &= \frac{\theta}{2} \end{aligned}$$

Using the Method of Moments we proceed to estimate $\mu = \theta/2$ by m_1. Thus since $m_1 = \overline{x}$ we have

$$\frac{\tilde{\theta}}{2} = \overline{x}$$

and,

$$\tilde{\theta} = 2\overline{x}.$$

Then, $\tilde{\theta} = 2\,\overline{x}$ and the MME of θ is $2\,\overline{X}$.

Example 2

In this example the distribution has two parameters.

Given $X_1, \ldots, X_n$ is a random sample from the $N(\mu, \sigma^2)$ distribution, find the method of moments estimates of μ and σ^2.

Solution:

For the normal distribution, $E(X) = \mu$ and $E(X^2) = \sigma^2 + \mu^2$.

Using the Method of Moments:

Equate $E(X)$ to m_1 and $E(X^2)$ to m_2 so that, $\tilde{\mu} = \overline{x}$ and $\tilde{\sigma}^2 + \tilde{\mu}^2 = m_2$.

That is, estimate μ by $\overline{x}$ and estimate σ^2 by $m_2 - \overline{x}^2$. Then,

$$\tilde{\mu} = \overline{x}, \text{ and } \tilde{\sigma^2} = \frac{1}{n}\sum x_i^2 - \overline{x}^2 .$$

The latter can also be written as $\tilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$

General procedure

Let $X_1, X_2, \ldots, X_n$ be a random sample from $F(x : \theta_1, \ldots, \theta_k)$. That is, suppose that there are k parameters to be estimated. Let μ'_r, m_r $(r = 1, 2, \ldots, k)$ denote the first k population and sample moments respectively, and suppose that each of these population moments are certain known functions of the parameters. That is,

$$\begin{aligned} \mu'_1 &= g_1(\theta_1, \ldots, \theta_k) \\ \mu'_2 &= g_2(\theta_1, \ldots, \theta_k) \\ &\vdots \\ \mu'_k &= g_k(\theta_1, \ldots, \theta_k) \ . \end{aligned}$$

Solving simultaneously the set of equations,

$$\mu'_r = g_r(\tilde{\theta}_1, \ldots, \tilde{\theta}_k) = m_r, \ r = 1, 2, \ldots, k$$

gives the required estimates, $\tilde{\theta}_1, \ldots, \tilde{\theta}_k$.

1.2 Estimation by the Method of Maximum Likelihood

Likelihood of a sample

First the term **likelihood of the sample** must be defined. This has to be done separately for discrete and continuous distributions.

Definition

Let $x_1, x_2, \ldots, x_n$ be sample observations taken on the random variables $X_1, X_2, \ldots, X_n$. Then the likelihood of the sample, $L(\theta|x_1, x_2, \ldots, x_n)$, is defined as:

(i) the joint probability of $x_1, x_2, \ldots, x_n$ if $X_1, X_2, \ldots, X_n$ are discrete, and

(ii) the joint probability density function of $X_1, \ldots, X_n$ evaluated at $x_1, x_2, \ldots, x_n$ if the random variables are continuous.

Likelihood of a sample

In general the value of the likelihood depends not only on the (fixed) sample $x_1, x_2, \ldots, x$ but on the value of the (unknown) parameter θ. and can be thought of as a function of θ.

The **likelihood function** for a set of n identically and independently distributed (iid) random variables, $X_1, X_2, \ldots, X_n$, can thus be written as:

$$\mathrm{L}(\theta; x_1, \ldots, x_n) = \begin{cases} P(X_1 = x_1).P(X_2 = x_2)...P(X_n = x_n) & \text{for X discrete} \\ f(x_1; \theta).f(x_2; \theta)...f(x_n; \theta) & \text{for X continuous.} \end{cases}$$

Likelihood of a sample

- Likelihood function
  - based on sample data only
  - likelihood of sample data for different values of parameter or different hypotheses
- Proportional to *P*(data|parameter)
  - sum of all likelihoods does not equal 1, so not true probability

(Quinn and Keough 2002)

Maximum likelihood estimators

For the discrete case, $L(\theta; x_1 \ldots, x_n)$ is the probability (or likelihood) of observing $(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$ It would then seem that a sensible approach to selecting an estimate of θ would be to find the value of θ which maximizes the probability of observing $(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$, (the event which occured).

The **maximum likelihood estimate** (MLE) of θ is defined as that value of θ which maximizes the likelihood. To state it more mathematically, the MLE of θ is that value of θ, say $\hat{\theta}$ such that

$$L(\hat{\theta}; x_1, \ldots, x_n) > L(\theta'; x_1, \ldots, x_n).$$

where θ' is any other value of θ.

Before we consider particular examples of MLE's, some comments about notation and technique are needed.

Maximum likelihood estimators: some remarks

1. It is customary to use $\hat{\theta}$ to denote both estimator (random variable) and estimate (its observed value). Recall that we used $\tilde{\theta}$ for the MME.

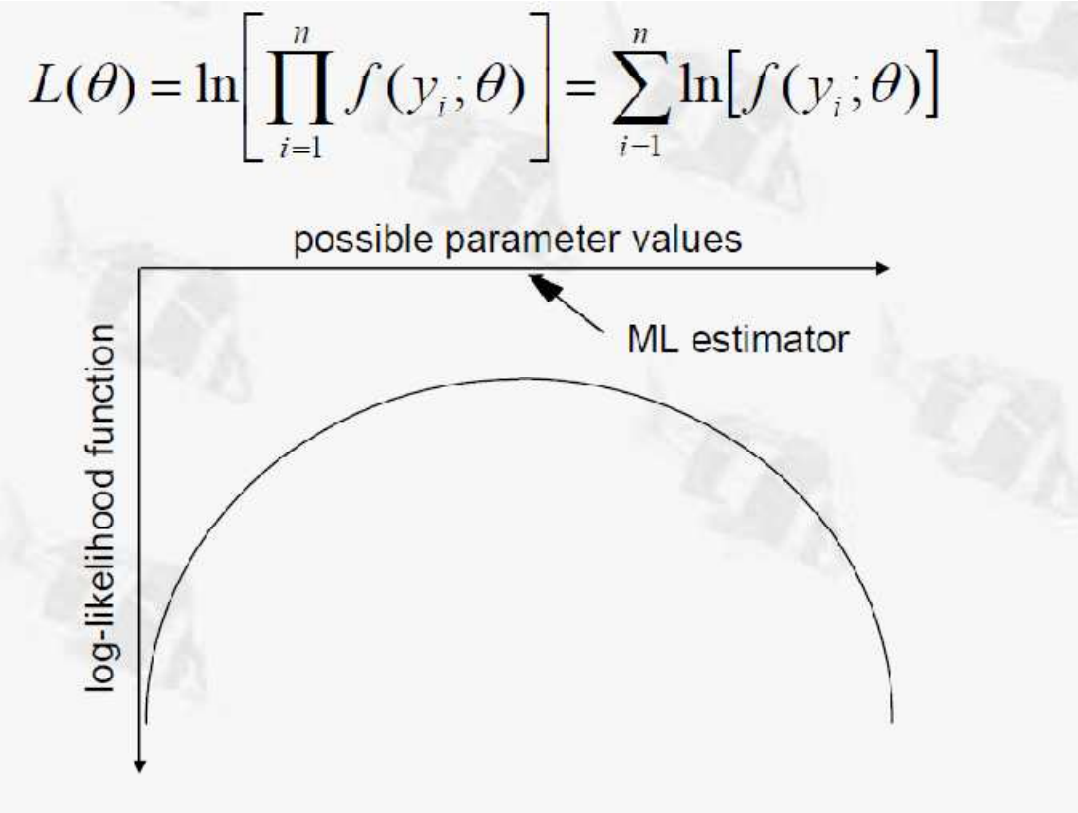
2. Since $L(\theta; x_1, x_2, \ldots, x_n)$ is a product, and sums are usually more convenient to deal with than products, it is customary to maximize $\log L(\theta; x_1, \ldots, x_n)$ which we usually abbreviate to $l(\theta)$. This has the same effect. Since $\log L$ is a strictly increasing function of L, it will take on its maximum at the same point.

3. In some problems, θ will be a vector in which case $L(\theta)$ has to be maximized by differentiating with respect to 2 (or more) variables and solving simuiltaneously 2 (or more) equations.

4. The method of differentiation to find a maximum only works if the function concerned actually has a turning point.

Maximum likelihood estimators

$$L(\theta) = \ln\left[\prod_{i=1}^{n} f(y_i;\theta)\right] = \sum_{i-1}^{n} \ln\left[f(y_i;\theta)\right]$$



(less reliable for small sample sizes and unusual distributions)

Example 1

Given X is distributed bin$(1, p)$ where $p \in (0, 1)$, and a random sample $x_1, x_2, ..., x_n$, find the maximum likelihood estimate of p.

Solution: The likelihood is,

$$\begin{aligned} L(p; x_1, x_2, \ldots, x_n) &= P(X_1 = x_1)P(X_2 = x_2)...P(X_n = x_n) \\ &= \prod_{i=1}^{n} \binom{1}{x_i} p^{x_i}(1-p)^{1-x_i} \\ &= p^{x_1+x_2+\cdots+x_n}(1-p)^{n-x_1-x_2-\cdots-x_n} \\ &= p^{\sum x_i}(1-p)^{n-\sum x_i} \end{aligned}$$

So

$$\log L(p) = \sum x_i \log p + (n - \sum x_i) \log(1-p)$$

Differentiating with respect to p, we have

$$\frac{d\ logL(p)}{dp} = \frac{\sum x_i}{p} \quad - \quad \frac{n - \sum x_i}{1-p}$$

This is equal to zero when $\sum x_i(1-p) = p(n - \sum x_i)$, that is, when $p = \sum x_i/n$.
This estimate is denoted by $\hat{p}$.
Thus, if the random variable X is distributed bin$(1, p)$, the MLE of p derived from a sample of size n is

$$\hat{p} = \overline{X}.$$

Example 2

Given $x_1, x_2, \ldots, x_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution, where both μ and σ^2 are unknown, find the maximum likelihood estimates of μ and σ^2.

Solution: Write the likelihood as:

$$\begin{aligned} L(\mu, \sigma^2; x_1, \ldots, x_n) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2} \end{aligned}$$

So

$$\log L(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n}(x_i - \mu)^2/2\sigma^2$$

To maximize this w.r.t. μ and σ^2 we must solve simultaneously the two equations

$$\partial \log L(\mu, \sigma^2)/\partial\mu = 0$$

$$\partial \log L(\mu, \sigma^2)/\partial\sigma^2 = 0.$$

These equations become, respectively,

$$-\frac{1}{2}\cdot\left(-\frac{2}{\sigma^2}\right)\sum_{i=1}^{n}(x_i-\mu)=0 \tag{1.5}$$

$$\frac{-n}{2\sigma^2}+\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^4}=0 \tag{1.6}$$

From (1.5) we obtain $\sum_{i=1}^{n} x_i = n\mu$, so that $\hat{\mu}=\overline{x}$. Using this in equation (1.6), we obtain

$$\hat{\sigma^2}=\sum_{i=1}^{n}(x_i-\overline{x})^2/n\ .$$

Thus, if X is distributed $N(\mu,\sigma^2)$, the MLE's of μ and σ^2 derived from a sample of size n are

$$\hat{\mu}=\overline{X} \qquad \text{and} \qquad \hat{\sigma^2}=\sum_{i=1}^{n}(X_i-\overline{X})^2/n.$$

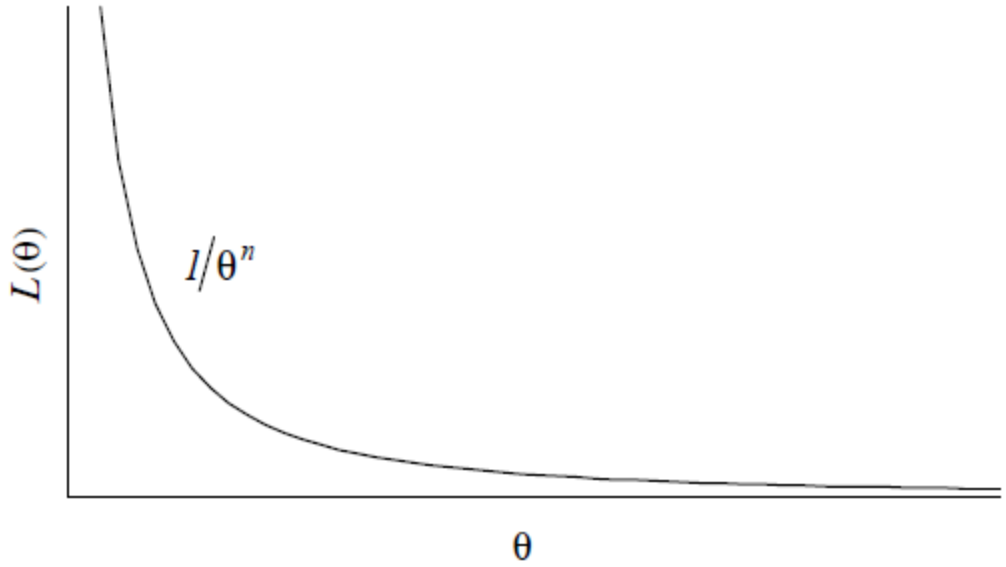
Note that these are the same estimators as obtained by the method of moments.

Example 3

Given random variable X is distributed uniformly on $[0, \theta]$, find the MLE of θ based on a sample of size n.

Solution: Now $f(x_i; \theta) = 1/\theta,\ x_i \in [0, \theta],\ i = 1, 2, \ldots, n$. So the likelihood is

$$L(\theta; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} (1/\theta) = 1/\theta^n \ .$$



Example 3 – continued

When we come to find the maximum of this function we note that the slope is not zero anywhere, so there is no use finding $\frac{dL(\theta)}{d\theta}$ or $\frac{d\log L(\theta)}{d\theta}$.

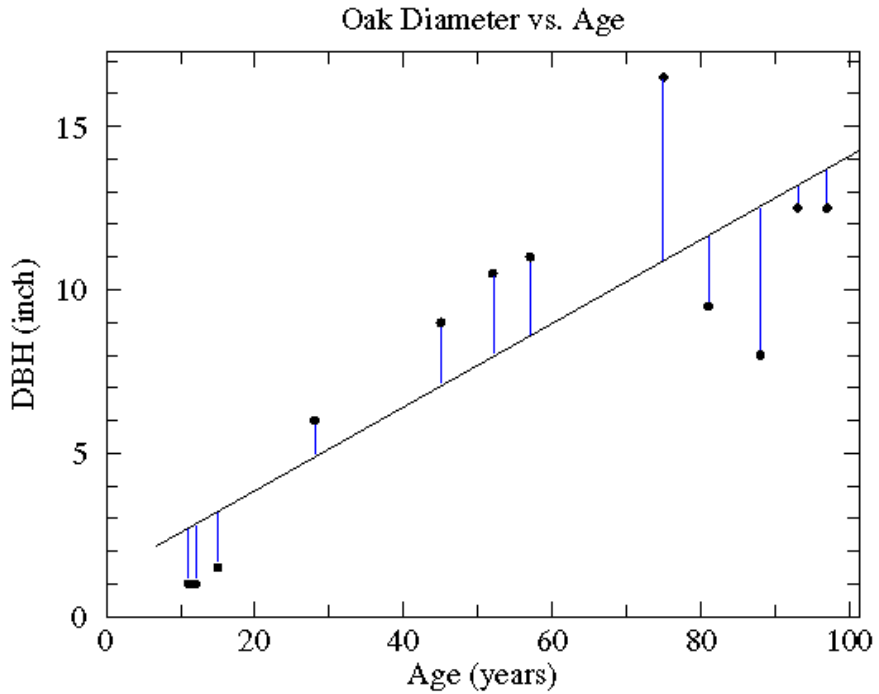
Note however that $L(\theta)$ increases as $\theta \to 0$. So $L(\theta)$ is maximized by setting θ equal to the *smallest* value it can take. If the observed values are $x_1, \ldots, x_n$ then θ can be no smaller than the *largest* of these. This is because $x_i \in [0, \theta]$ for $i = 1, \ldots, n$. That is, each $x_i \leq \theta$ or $\theta \geq$ each x_i.

Thus, if X is distributed U(0, θ), the MLE of θ is

$$\hat{\theta} = \max(X_i).$$

Ordinary least squares estimators

- OLS estimators minimize the sum of squared deviations between each observation and estimated parameter



$1-r^2$ =

$$\frac{\text{(deviations from trendline)}}{\text{(standard deviation of } y \text{ data)}}$$

(reliable for linear models with normal distributions)

Example

Suppose $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d) random variables with expectation μ and variance σ^2. If the sample mean and uncorrected sample variance are defined as

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

then S^2 is a biased estimator of σ^2, because

$$\begin{aligned}
\mathrm{E}[S^2] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left((X_i - \mu) - (\overline{X} - \mu)\right)^2\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - 2(\overline{X} - \mu)\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) + (\overline{X} - \mu)^2\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\overline{X} - \mu)^2\right] = \sigma^2 - \mathrm{E}\left[(\overline{X} - \mu)^2\right] < \sigma^2.
\end{aligned}$$

Example

In other words, the expected value of the uncorrected sample variance does not equal the population variance σ^2, unless multiplied by a normalization factor. The sample mean, on the other hand, is an unbiased estimator of the population mean μ.

The reason that S^2 is biased stems from the fact that the sample mean is an ordinary least squares (OLS) estimator for μ: It is such a number that makes the sum $\Sigma(X_i - \mu)^2$ as small as possible. That is, when any other number is plugged into this sum, the sum can only increase. In particular, the choice $m = \mu$ gives, first (or most outcomes)

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 < \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2,$$

and then

$$\mathrm{E}[S^2] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right] < \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right] = \sigma^2.$$

2 Properties of Estimators

2.1 Introduction

Using different methods of estimation can lead to different estimators. Criteria for deciding which are *good* estimators are required. Before listing the qualities of a good estimator, it is important to understand that they are random variables.

The behaviour of an estimator for different random samples will be described by a probability distribution. The actual distribution of the estimator is not a concern here and only its mean and variance will be considered. As a first condition it seems reasonable to ask that the distribution of the estimator be *centered* around the parameter it is estimating. If not it will tend to overestimate or underestimate θ. A second property an estimator should possess is *precision*. An estimator is precise if the dispersion of its distribution is small. These two concepts are incorporated in the definitions of *unbiasedness* and *efficiency* below.

2.2 Unbiased

In the following, $X_1, X_2, \ldots, X_n$ is a random sample from the distribution $F(x;\theta)$ and $H(X_1, \ldots, X_n) = \hat{\theta}$ will denote an estimator of θ (not necessarily the MLE).

Definition

An estimator $\hat{\theta}$ of θ is **unbiased** if

$$E(\hat{\theta}) = \theta \text{ for all } \theta.$$

If an estimator $\hat{\theta}$ is biased, the **bias** is given by

$$b = E(\hat{\theta}) - \theta \ .$$

There may be large number of unbiased estimators of a parameter for any given distribution and a further criterion for choosing between all the unbiased estimators is needed.

2.3 Efficiency

Definition Efficiency

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be 2 unbiased estimators of θ with variances $\text{Var}(\hat{\theta}_1)$, $\text{Var}(\hat{\theta}_2)$ respectively, We say that $\hat{\theta}_1$ is **more efficient** than $\hat{\theta}_2$ if

$$\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2) \ .$$

That is, $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ if it has a smaller variance.

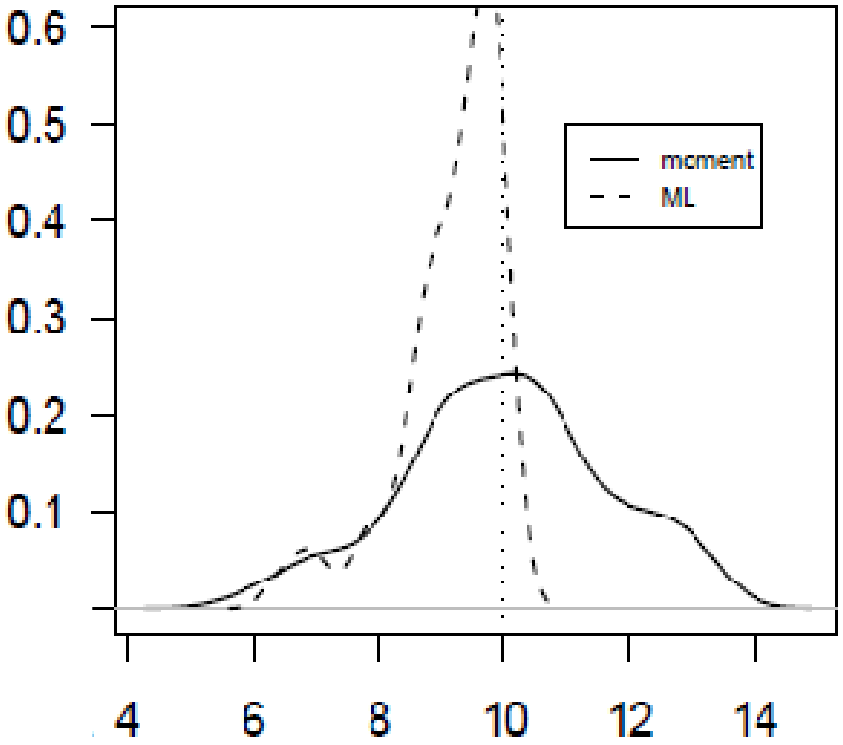
Definition Relative Efficiency

The **relative efficiency** of $\hat{\theta}_2$ with respect to $\hat{\theta}_1$ is defined as

$$\text{efficiency} = \text{Var}(\hat{\theta}_1)/\text{Var}(\hat{\theta}_2) \ .$$

Practical

Generate 100 random samples of size 10 from a U(0,10) distribution. For each of the 100 samples generated calculate the MME and MLE for μ and graph the results.



- The MMEs give unbiased estimates which may or may not be in the range space.
- The MLEs are all less than 10 and hence biased.
- What if the sample size is increased?

2.4 Consistency

It will now be useful to indicate that the estimator is based on a sample of size n by denoting it by $\hat{\theta}_n$.

Definition **Consistency** $\hat{\theta}_n$ is a **consistent** estimator of θ if

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0 \text{ for all } \epsilon > 0 \ .$$

We then say that $\hat{\theta}_n$ **converges in probability** to θ as $n \to \infty$. Equivalently,

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \epsilon) = 1 \ .$$

This is a large-sample or *asymptotic* property. Consistency has to do only with the limiting behaviour of an estimator as the sample size increases without limit and does not imply that the observed value of $\hat{\theta}$ is necessarily close to θ for any specific size of sample n. If only a relatively small sample is available, it would seem immaterial whether a consistent estimator is used or not.

Estimator	**Property**			
	Unbiased	**Efficient**	**Consistent**	**Invariant**
MLE	They become minimum variance *unbiased estimators as the sample size increases*	They become *minimum variance* unbiased *estimators as the sample size increases*	☺	☺

- MLEs are ***invariant*** with respect to reparameterizations :
  if you have found the MLE for a parameter θ, then you have found one for $g(\theta)$, namely the g(MLE), where g is an invertible function
- MLEs have approximate normal distributions and approximate sample variances that can be used to generate confidence bounds and hypothesis tests for the parameters.

3 Confidence intervals

3.1 Introduction – understanding the concept

From point estimators to interval estimators

In the earlier part of this chapter we have been considering **point estimators** of a parameter. By *point estimator* we are referring to the fact that, after the sampling has been done and the observed value of the estimator computed, our end-product is the single number which is hopefully a good approximation for the unknown true value of the parameter. If the estimator is good according to some criteria, then the estimate should be reasonably close to the unknown true value. But the single number itself does not include any indication of how high the probability might be that the estimator has taken on a value close to the true unknown value. The method of **confidence intervals** gives both an idea of the actual numerical value of the parameter, by giving it a *range* of possible values, and a measure of how confident we are that the true value of the parameter is in that range. To pursue this idea further consider the following example.

Example

Consider a random sample of size n for a normal distribution with mean μ (unknown) and known variance σ^2. Find a 95% confidence interval for the unknown mean, μ.

Solution: We know that the best estimator of μ is $\overline{X}$ and the sampling distribution of $\overline{X}$ is $\mathrm{N}(\mu, \frac{\sigma^2}{n})$. Then from the standard normal,

$$P\left(\frac{|\overline{X}-\mu|}{\sigma/\sqrt{n}} < 1.96\right) = .95 \ .$$

The event $\frac{|\overline{X}-\mu|}{\sigma/\sqrt{n}} < 1.96$ is equivalent to the event

$$\mu - \frac{1.96\sigma}{\sqrt{n}} < \overline{X} < \mu + \frac{1.96\sigma}{\sqrt{n}} \ ,$$

which is equivalent to the event

$$\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}} \ .$$

.

Hence

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = .95$$

The two statistics $\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}$, $\overline{X} + 1.96\frac{\sigma}{\sqrt{n}}$ are the endpoints of a 95% confidence interval for μ. This is reported as

The 95% CI for μ is $\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$

Definition

An interval, at least one of whose endpoints is a random variable is called a **random interval**.

With

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = .95$$

we are saying that the probability is 0.95 that the random interval $(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}})$ contains μ. A confidence interval (CI) has to be interpreted carefully. For a particular sample, where $\overline{x}$ is the observed value of $\overline{X}$, a 95% CI for μ is

$$\left(\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}\ ,\ \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)\ ,$$

but the statement

$$\overline{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < 1.96\frac{\sigma}{\sqrt{n}}$$

is either true or false. The parameter μ is a constant and either the interval contains it in which case the statement is true, or it does not contain it, in which case the statement is false.

Correct interpretation of "probabilities associated to intervals"

How then is the probability 0.95 to be interpreted? It must be considered in terms of the relative frequency with which the indicated event will occur "in the long run" of similar sampling experiments.

Each time we take a sample of size n, a different $\overline{x}$, and hence a different interval

$$\left(\overline{x} - 1.96\frac{\sigma}{\sqrt{n}} \ , \ \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

would be obtained. Some of these intervals will contain μ as claimed, and some will not. In fact, if we did this many times, we'd expect that 95 times out of 100 the interval obtained would contain μ. The measure of our confidence is then 0.95 because **before a sample is drawn** there is a probability of 0.95 that the confidence interval to be constructed will cover the true mean.

A statement such as $P(3.5 < \mu < 4.9) = 0.95$ is incorrect and should be replaced by :
A 95% confidence interval for μ is $(3.5,\ 4.9)$.

We can generalize the above as follows: Let $z_{\alpha/2}$ be defined by

$$\Phi(z_{\alpha/2}) = 1 - (\alpha/2) .$$

That is, the area under the normal curve **above** $z_{\alpha/2}$ is $\alpha/2$. Then

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) - 1 - \alpha.$$

So a $100(1-\alpha)\%$ CI for μ is

$$\left(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) .$$

Commonly used values of α are 0.1, 0.05, 0.01.

Confidence intervals are not unique

Confidence intervals for a given parameter are not unique. For example, we have considered a **symmetric, two–sided** interval, but

$$\left(\overline{x} - z_{2\alpha/3}\frac{\sigma}{\sqrt{n}}\ ,\ \overline{x} + z_{\alpha/3}\frac{\sigma}{\sqrt{n}}\right)$$

is also a $100(1-\alpha)\%$ CI for μ. Likewise, we could have one-sided CI's for μ. For example,

$$\left(-\infty,\ \overline{x} + z_{\alpha}\frac{\sigma}{\sqrt{n}}\right) \text{ or } \left(\overline{x} - z_{\alpha}\frac{\sigma}{\sqrt{n}}\ ,\ \infty\right) .$$

[The second of these arises from considering $P\left(\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha}\right) = 1-\alpha$.]

3.2 Finding confidence intervals in practice

The pivotal method

We will describe a general method of finding a confidence interval for θ from a random sample of size n. It is known as the **pivotal method** as it depends on finding a pivotal quantity that has 2 characteristics:

(i) It is a function of the sample observations and the unknown parameter θ, say $H(X_1, X_2, \ldots, X_n; \theta)$ where θ is the only unknown quantity,

(ii) It has a probability distribution that does not depend on θ.

Any probability statement of the form

$$P(a < H(X_1, X_2, \ldots, X_n; \theta) < b) = 1 - \alpha$$

will give rise to a probability statement about θ.

Example

Given $X_1, X_2, \ldots, X_{n_1}$ from $\mathrm{N}(\mu_1,\sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_{n_2}$ from $\mathrm{N}(\mu_2, \sigma_2^2)$ where σ_1^2, σ_2^2 are known, find a symmetric 95% CI for $\mu_1 - \mu_2$.

Solution: Consider $\mu_1 - \mu_2$ $(= \theta$, say) as a single parameter. Then $\overline{X}$ is distributed $\mathrm{N}(\mu_1, \sigma_1^2/n_1)$ and $\overline{Y}$ is distributed $\mathrm{N}(\mu_2, \sigma_2^2/n_2)$ and further, $\overline{X}$ and $\overline{Y}$ are independent. It follows that $\overline{X} - \overline{Y}$ is normally distributed, and writing it in standardized form,

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}} \text{ is distributed as } N(0,\ 1)\ .$$

So we have found the pivotal quantity which is a function of $\mu_1 - \mu_2$ but whose distribution does not depend on $\mu_1 - \mu_2$. A 95% CI for $\theta = \mu_1 - \mu_2$ is found by considering

$$P\left(-1.96\ <\ \frac{\overline{X} - \overline{Y} - \theta}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}\ <\ 1.96\right) = .95\ ,$$

which, on rearrangement, gives the appropriate CI for $\mu_1 - \mu_2$. That is,

$$\left(\overline{x} - \overline{y} - 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\ ,\ \overline{x} - \overline{y} + 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)\ .$$

3.3 One-sample problem

Confidence Intervals for σ^2

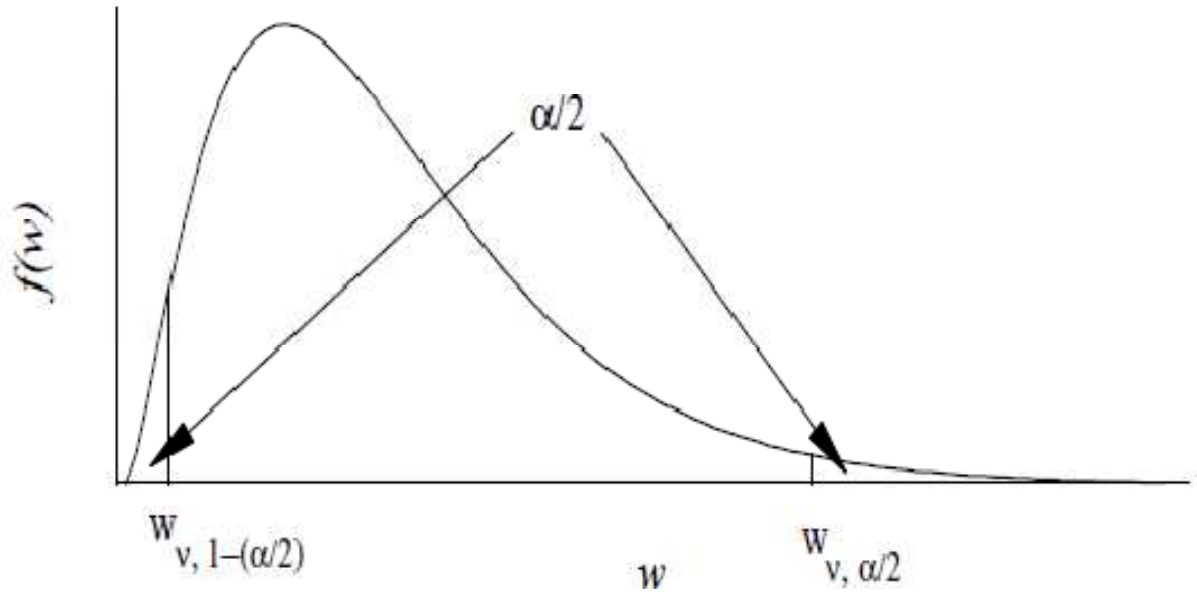
Case (ii)

Case (i)

Population μ	Estimation of σ^2	Test Statistic & Distribution
μ Known	$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu\right)^2$	$\frac{ns^2}{\sigma^2} \sim \chi^2_n$
	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \mu\right)^2$	$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_n$
μ Unknown	$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$	$\frac{ns^2}{\sigma^2} \sim \chi^2_{n-1}$
	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$	$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$

Case (i)

Let $X_1, X_2, \ldots, X_n$ be a random sample from $\mathrm{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown. It has been shown that S^2 is an unbiased estimate of σ^2 and we can find a confidence interval for σ^2 using the χ^2 distribution. Recall that $W = \nu S^2/\sigma^2 \sim \chi^2_\nu$. By way of notation, let $w_{\nu,\alpha}$ be defined by $P(W > w_{\nu,\alpha}) = \alpha$, where $W \sim \chi^2_\nu$.



The event $w_{\nu,1-(\alpha/2)} < W < w_{\nu,\alpha/2}$ occurs if and only if the events

$$\sigma^2 < \nu S^2/w_{\nu,1-(\alpha/2)}, \quad \sigma^2 > \nu S^2/w_{\nu,\alpha/2}$$

occur. So

$$P\left(w_{\nu,1-(\alpha/2)} < W < w_{\nu,\alpha/2}\right) - P\left(\nu S^2/w_{\nu,\alpha/2} < \sigma^2 < \nu S^2/w_{\nu,1-(\alpha/2)}\right)$$

and thus

A $100(1-\alpha)\%$ CI for σ^2 is $(\nu s^2/w_{\nu,\alpha/2}, \nu s^2/w_{\nu,1-(\alpha/2)})$

Case (ii)

Suppose now that $X_1, X_2, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$ where μ is known and we wish to find a CI for the unknown σ^2. Recall that the maximum likelihood estimator of σ^2 (which we'll denote by S^{*2}) is

$$S^{*2} = \sum_{i=1}^{n} (X_i - \mu)^2 / n.$$

We can easily show that this is unbiased.

$$E(S^{*2}) = \sum_{i=1}^{n} \frac{E(X_i - \mu)^2}{n} = n\frac{1}{n}\sigma^2 = \sigma^2$$

The distribution of S^{*2} is found by noting that $nS^{*2}/\sigma^2 = \sum_{i=1}^{n}(X_i - \mu)^2/\sigma^2$ is the sum of squares of n independent N(0,1) variates and is therefore distributed as χ^2_n. Proceeding in the same way as in Case (i) we find

A $100(1-\alpha)\%$ CI for σ^2 when μ is known is $\left(\frac{ns^{*2}}{w_{n,\alpha/2}}, \frac{ns^{*2}}{w_{n,1-(\alpha/2)}} \right)$

Confidence Intervals for μ

	Population σ^2	Estimation of μ	Test statistic and distribution
Case (ii)	σ^2 **Known**	$\bar{X} = 1/n \sum_1^n x_i$	$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$ $\frac{S}{\sqrt{n}} = \hat{sd}(\bar{X})$, S^2 unbiased for σ^2
Case (i)	σ^2 **Unknown**	$\bar{X} = 1/n \sum_1^n x_i$	$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$

Definition

A random variable has a t-distribution on ν degrees of freedom (or with parameter ν) if it can be expressed as the ratio of Z to $\sqrt{W/\nu}$ where $Z \sim N(0,1)$ and W (independent of Z) $\sim \chi^2_\nu$.

Case (i)

Given $X_1, X_2, \ldots, X_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution where σ^2 is unknown , then

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Then defining $t_{\nu,\alpha}$ as

$$P(T > t_{\nu,\alpha}) = \alpha \quad \text{where} \quad T \sim t_\nu,$$

we have

$$P\left(t_{\nu,\frac{\alpha}{2}} < \frac{\overline{X} - \mu}{S/\sqrt{n}} < t_{\nu,1-\frac{\alpha}{2}}\right) = 1 - \alpha.$$

That is,

$$P(\overline{X} - t_{\nu,\alpha/2}\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\nu,\alpha/2}\frac{S}{\sqrt{n}}) = 1 - \alpha.$$

Now rearrange the terms

$$
\begin{aligned}
&P\left(t_{\nu,\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{S/\sqrt{n}} < t_{\nu,1-\frac{\alpha}{2}}\right) \\
&\quad = P\left(\frac{S}{\sqrt{n}} \times t_{\nu,\frac{\alpha}{2}} < \overline{X}-\mu < \frac{S}{\sqrt{n}} \times t_{\nu,1-\frac{\alpha}{2}}\right) \\
&\quad = P\left(-\overline{X}+\frac{S}{\sqrt{n}} \times t_{\nu,\frac{\alpha}{2}} < -\mu < -\overline{X}+\frac{S}{\sqrt{n}} \times t_{\nu,1-\frac{\alpha}{2}}\right) \\
&\quad = P\left(\overline{X}-\frac{S}{\sqrt{n}} \times t_{\nu,\frac{\alpha}{2}} > \mu > \overline{X}-\frac{S}{\sqrt{n}} \times t_{\nu,1-\frac{\alpha}{2}}\right) \quad \text{inequality directions not conventional} \\
&\quad = P\left(\overline{X}-\frac{S}{\sqrt{n}} \times t_{\nu,1-\frac{\alpha}{2}} < \mu < \overline{X}-\frac{S}{\sqrt{n}} \times t_{\nu,\frac{\alpha}{2}}\right) \quad \text{inequality directions conventional}
\end{aligned}
$$

A $100(1-\alpha)\%$ confidence interval for μ is

$$\left(\overline{x} - t_{\nu,1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \ , \ \overline{x} - t_{\nu,\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right).$$

Remark

$$\left(\overline{x} - t_{\nu,1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \ , \ \overline{x} - t_{\nu,\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right).$$

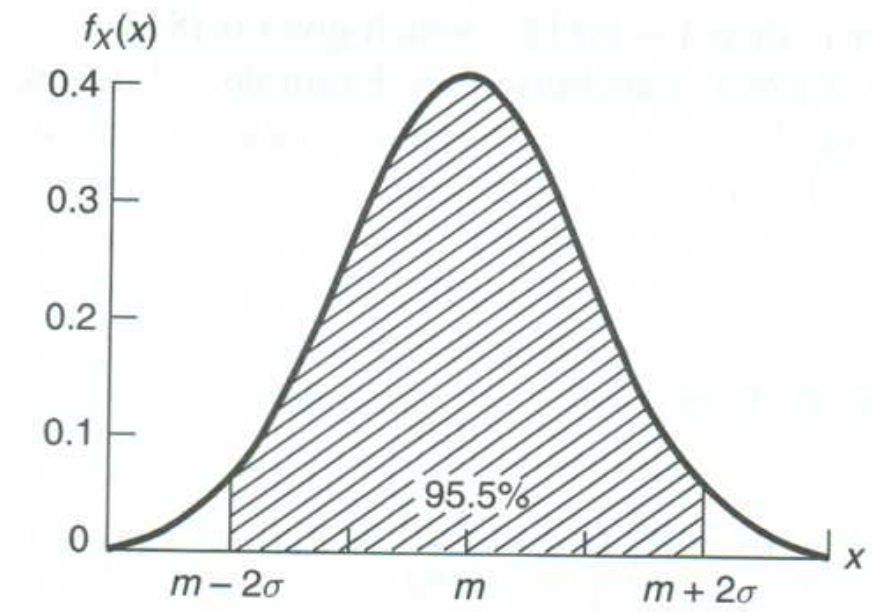
By the symmetry of the t-distribution, $t_{\nu,\frac{\alpha}{2}} = -t_{\nu,1-\frac{\alpha}{2}}$ and the lower tail quantile is a negative number, the upper tail quantile is the same magnitude but positive. So you would get the same result if you calculated

$$\left(\overline{x} - t_{\nu,1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \ , \ \overline{x} + t_{\nu,1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right)$$

Case (ii)

A symmetric 95% confidence interval for μ is $\overline{x} \pm 1.96\sigma/\sqrt{n}$ which arose from considering the inequality

$$-1.96 < \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} < 1.96$$



3.4 Two-sample problems

Confidence Interval for σ_1^2/σ_2^2 (note that nothing is said about the means)

Given s_1^2, s_2^2 are unbiased estimates of σ_1^2, σ_2^2 derived from samples of size n_1, n_2 respectively, from two normal populations, find a $(1-\alpha\%)$ confidence interval for σ_1^2/σ_2^2.

Now $\nu_1 S_1^2/\sigma_1^2$ and $\nu_2 S_2^2/\sigma_2^2$ are distributed as independent $\chi^2_{\nu_1}$, $\chi^2_{\nu_2}$ variates, and

$$\frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} \sim \frac{W_2/\nu_2}{W_1/\nu_1} \sim F(\nu_2, \nu_1).$$

Definition

Suppose S_1^2 and S_2^2 are the sample variances for two samples of sizes n_1, n_2 drawn from normal populations with variances σ_1^2 and σ_2^2, respectively. The random variable F is then defined as

$$F = S_1^2/S_2^2.$$

$$\frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} \sim \frac{W_2/\nu_2}{W_1/\nu_1} \sim F(\nu_2, \nu_1)$$

So

$$P\left(F_{1-\frac{\alpha}{2}}(\nu_2, \nu_1) < \frac{S_2^2}{S_1^2}\frac{\sigma_1^2}{\sigma_2^2} < F_{\frac{\alpha}{2}}(\nu_2, \nu_1)\right) = 1 - \alpha.$$

That is

$$P\left(\frac{S_1^2}{S_2^2}F_{1-\frac{\alpha}{2}}(\nu_2, \nu_1) < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}F_{\frac{\alpha}{2}}(\nu_2, \nu_1)\right) = 1 - \alpha.$$

Thus a $100(1-\alpha)\%$ confidence interval for σ_1^2/σ_2^2 is

$$\left(\frac{s_1^2}{s_2^2}F_{\frac{1-\alpha}{2}}(\nu_2, \nu_1),\ \frac{s_1^2}{s_2^2}F_{\frac{\alpha}{2}}(\nu_2, \nu_1)\right).$$

Confidence intervals for means in different populations

Setting	Estimate	Standard Error	Confidence Interval	Test Statistic	Distribution
Population Mean					
— σ known	$\bar{x}$	$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm z^* \mathrm{SE}$	$z = \frac{\bar{x}-\mu_0}{\mathrm{SE}}$	$\mathrm{Normal}(0,1)$
— σ unknown	$\bar{x}$	$\widehat{\mathrm{SE}} = \frac{s}{\sqrt{n}}$	$\bar{x} \pm t^* \widehat{\mathrm{SE}}$	$t = \frac{\bar{x}-\mu_0}{\widehat{\mathrm{SE}}}$	$\mathrm{t}(n-1)$
Difference Between Population Means					
— σ_1 and σ_2 known	$\bar{x}_1 - \bar{x}_2$	$\mathrm{SE} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$\bar{x}_1 - \bar{x}_2 \pm z^* \mathrm{SE}$	$z = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\mathrm{SE}}$	$\mathrm{Normal}(0,1)$
— $\sigma_1 = \sigma_2$ unknown	$\bar{x}_1 - \bar{x}_2$	$\widehat{\mathrm{SE}} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$	$\bar{x}_1 - \bar{x}_2 \pm t^* \widehat{\mathrm{SE}}$	$t = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\widehat{\mathrm{SE}}}$	$\mathrm{t}(n_1+n_2-2)$
— $\sigma_1 \neq \sigma_2$ unknown	$\bar{x}_1 - \bar{x}_2$	$\widehat{\mathrm{SE}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$\bar{x}_1 - \bar{x}_2 \pm t^* \widehat{\mathrm{SE}}$	$t = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\widehat{\mathrm{SE}}}$	$\mathrm{t}(f)$

Confidence Interval for $\mu_1 = \mu_2$ (known variances)

Given independent random samples $X_1, X_2, \ldots, X_{n_1}$ from a normal population with unknown mean μ_1 and known variance σ_1^2 and $Y_1, Y_2, \ldots, Y_{n_2}$ from a normal population with unknown mean μ_2 and known variance σ_2^2, derive a test for the hypothesis H: $\mu_1 = \mu_2$ against one-sided and two-sided alternatives.

Solution: Note that the hypothesis can be written as $H : \mu_1 - \mu_2 = 0$. An unbiased estimator of $\mu_1 - \mu_2$ is $\overline{X} - \overline{Y}$ so this will be used as the test statistic. Its distribution is given by

$$\overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2,\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

or, in standardized form, **if H_0 is true**

$$\frac{\overline{X} - \overline{Y}}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}} \sim N(0,\ 1).$$

Confidence Interval for $\mu_1 - \mu_2$ (unknown variances)

- Either the population variances are unequal
- Either the population variances are equal
  - o If it is reasonable to assume that $\sigma_1=\sigma_2$, we can estimate the standard error more efficiently by combining the sample.
  - o Assume equal variances when $S_2 / S_1 < 2$
  - o If you are unsure, the unequal variance formula will be the conservative choice (less power, but less likely to be incorrect).

Pooling variances

Given 2 unbiased estimates of σ^2, s_1^2 and s_2^2, it is often useful to be able to combine them to obtain a single unbiased estimate. Assume the new estimator, S^2, is linear combination of s_1^2 and s_2^2 so that S^2 has the smallest variance of all such linear, unbiased estimates (that is it is said to have *minimum variance*). Let

$$S^2 = a_1 S_1^2 + a_2 S_2^2, \text{ where } a_1, a_2 \text{ are positive constants.}$$

Firstly, to be unbiased,

$$E(S^2) = a_1 E(S_1^2) + a_2 E(S_2^2) = \sigma^2(a_1 + a_2) = \sigma^2$$

which implies that

$$a_1 + a_2 = 1.$$

Secondly, if it is assumed that S_1^2 and S_2^2 are independent then

$$\begin{aligned} \text{Var}(S^2) &= a_1^2 \text{Var}(S_1^2) + a_2^2 \text{Var}(S_2^2) \\ &= a_1^2 \text{Var}(S_1^2) + (1 - a_1)^2 \text{Var}(S_2^2) \end{aligned}$$

The variance of S^2 is minimised when, (writing $V(.)$ for Var(.)),

$$\frac{dV(S^2)}{da_1} = 2a_1\,V(S_1^2) \;-\; 2(1-a_1)V(S_2^2) = 0.$$

That is when,

$$a_1 = \frac{V(S_2^2)}{V(S_1^2)+V(S_2^2)},\;\; a_2 = \frac{V(S_1^2)}{V(S_1^2)+V(S_2^2)}$$

In the case where the X_i are normally distributed, $V(S_j^2) = 2\sigma^4/(n_j-1)$ (see Assignment 3, Question 1). Then the pooled sample variance is

$$s^2 = \frac{\frac{(n_1-1)s_1^2}{2\sigma^4}+\frac{(n_2-1)s_2^2}{2\sigma^4}}{\frac{n_1-1}{2\sigma^4}+\frac{n_2-1}{2\sigma^4}}$$

$$= \frac{\nu_1 s_1^2+\nu_2 s_2^2}{\nu_1+\nu_2}$$

where $\nu_1 = n_1-1$, $\nu_2 = n_2-1$.

With the pooled variance, the key expression to remember is

$$-t_{\alpha/2} < \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{(1/n_1 + 1/n_2)s^2}} < t_{\alpha/2},$$

with s^2 the pooled variance as described on the previous slide, naturally leading to a $(1 - \alpha)$ percent confidence interval.

Main reference:

STAT261 Statistical inference notes – School of mathematics, statistics and computer science. University of New England, Oct 4, 2007