Unit Schedule: Modules

Module	Week	Content	Ross
1.	1	introduction to modelling	1,2
2.	2	probability refresher	3
	3	random vars & expected values	4
	4	special distributions	5
3.	5	**statistical inference**	6&7
	6	confidence intervals	7
	7	hypothesis testing	8
4.	8	dependence & linear regression	9
	9	classification, clustering & mixtures	
5.	10	random numbers & simulation	15(bits)
	11	basic machine learning	
6.	12	modelling, validation and review	

FIT5197 Statistical Data Modelling

Module 3

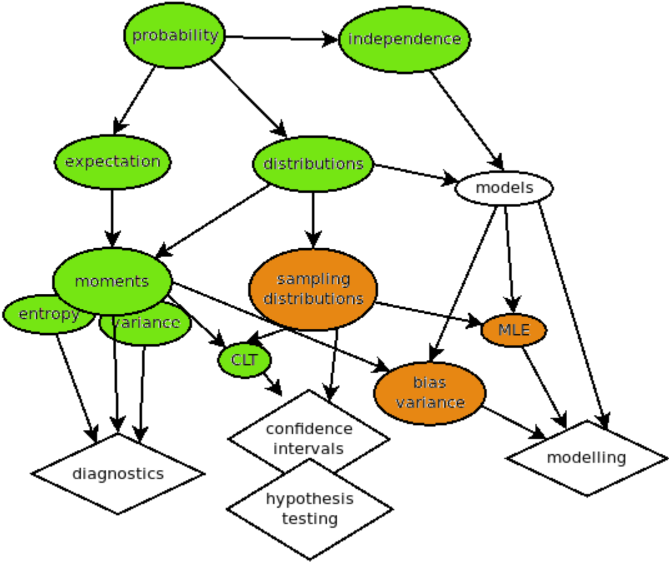
Statistical Inference

2020 Lecture 5

Monash University

polls at *https://flux.qa/43FMK4*

Concept Map for This Unit



Statistical Inference

(ePub sections 3.1,3.2,3.4,3.5

Ross 5.8, 6.1-6.6, 7.1, 7.2, 7.7)

Outline

Overview of Statistical Inference

Sampling Distributions

Distributions From the Gaussian

Point Estimation

Minimising Squared Error

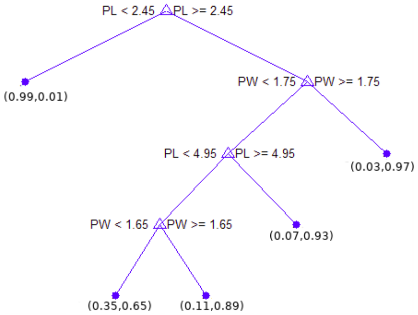
Maximising Likelihood

Saline Use for Patients

See
Saline use on decline at Vanderbilt following landmark studies.

- 28,000 patients received either a "saline drip" or a "balanced fluids drip" when under care
- 1% more died or had serious kidney damage with the saline drip
- we don't know the base rate, what percentage died or had serious kidney damage
- Questions:
  - is this 1% improvement "significant" or is it due to statistical chance?
  - what might the "true" percentage improvement be? perhaps just 0.25% or up to 2%?

Estimating Tree Probabilities

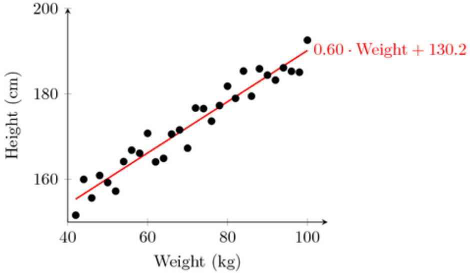


- how do we estimate the probability vectors at leaves
- how reliable are these estimates?

NB. the internal test nodes have been selected to make the class probabilities more extreme, so there is selection bias in getting to the leaf node

  - the leaf nodes may be more extreme than the "truth"

Linear Regression Coefficients



from *PGFplots.net*

- how good is our estimate of the y-intercept (130.2)?
    - if we assume the slope is given, the y-intercept has the Gaussian sample $\{ \textit{height}_i - 0.60\textit{weight}_i \; : \; i = 1, ..., 30 \}$
- how do we estimate the error of our estimate?

Statistical Inference

Statistical inference: the use of observed data to make inferences about the unknown parameters of a model.

- an excellent review in *Wikipedia*
- the process sometimes referred to as estimation
- in the hard sciences, a related concept is *inverse problem*, inferring from a set of observations the causal factors that produced them
- there is no correct or “objective” answer in general
  - from finite observations, we can rarely know the “truth”
- statistical inference is the first step on the way to full-scale model building

Outline

Overview of Statistical Inference

Sampling Distributions

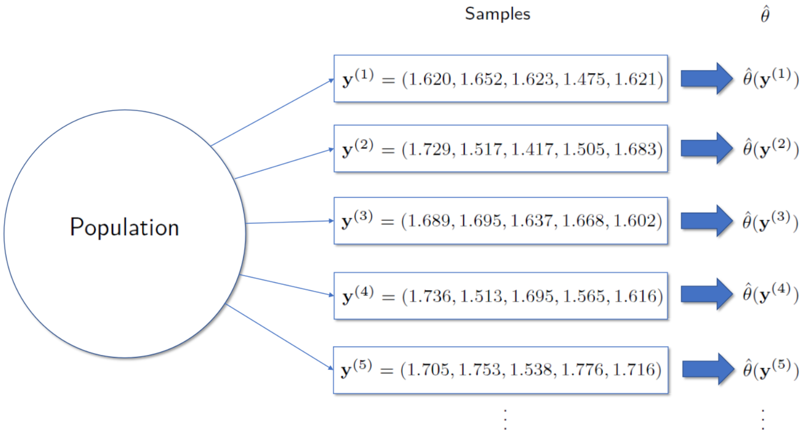
Distributions From the Gaussian

Point Estimation

Minimising Squared Error

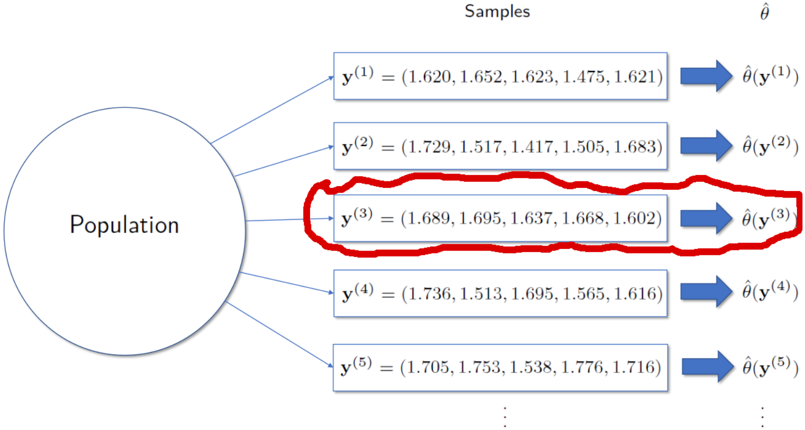
Maximising Likelihood

Samples from the Population


Samples
$\hat{\theta}$
Population
$\mathbf{y}^{(1)} = (1.620, 1.652, 1.623, 1.475, 1.621)$
$\hat{\theta}(\mathbf{y}^{(1)})$
$\mathbf{y}^{(2)} = (1.729, 1.517, 1.417, 1.505, 1.683)$
$\hat{\theta}(\mathbf{y}^{(2)})$
$\mathbf{y}^{(3)} = (1.689, 1.695, 1.637, 1.668, 1.602)$
$\hat{\theta}(\mathbf{y}^{(3)})$
$\mathbf{y}^{(4)} = (1.736, 1.513, 1.695, 1.565, 1.616)$
$\hat{\theta}(\mathbf{y}^{(4)})$
$\mathbf{y}^{(5)} = (1.705, 1.753, 1.538, 1.776, 1.716)$
$\hat{\theta}(\mathbf{y}^{(5)})$
$\vdots$
$\vdots$

An (infinite) number of different random samples can be drawn from a population. Each sample would lead to a potentially different estimate $\hat{\theta}$ of a population parameter θ.

Our One Sample



Samples

$\hat{\theta}$

Population

$\mathbf{y}^{(1)} = (1.620, 1.652, 1.623, 1.475, 1.621)$ $\hat{\theta}(\mathbf{y}^{(1)})$

$\mathbf{y}^{(2)} = (1.729, 1.517, 1.417, 1.505, 1.683)$ $\hat{\theta}(\mathbf{y}^{(2)})$

$\mathbf{y}^{(3)} = (1.689, 1.695, 1.637, 1.668, 1.602)$ $\hat{\theta}(\mathbf{y}^{(3)})$

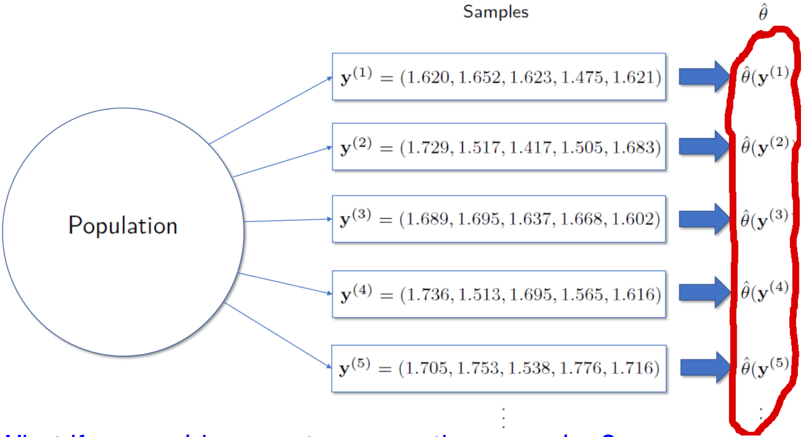
$\mathbf{y}^{(4)} = (1.736, 1.513, 1.695, 1.565, 1.616)$ $\hat{\theta}(\mathbf{y}^{(4)})$

$\mathbf{y}^{(5)} = (1.705, 1.753, 1.538, 1.776, 1.716)$ $\hat{\theta}(\mathbf{y}^{(5)})$

$\vdots$ $\vdots$

We have just one sample though, circled in red. So we cannot know anything about the other samples.

A Sampling Distribution?



$\mathbf{y}^{(1)} = (1.620, 1.652, 1.623, 1.475, 1.621)$ → $\hat{\theta}(\mathbf{y}^{(1)})$

$\mathbf{y}^{(2)} = (1.729, 1.517, 1.417, 1.505, 1.683)$ → $\hat{\theta}(\mathbf{y}^{(2)})$

$\mathbf{y}^{(3)} = (1.689, 1.695, 1.637, 1.668, 1.602)$ → $\hat{\theta}(\mathbf{y}^{(3)})$

$\mathbf{y}^{(4)} = (1.736, 1.513, 1.695, 1.565, 1.616)$ → $\hat{\theta}(\mathbf{y}^{(4)})$

$\mathbf{y}^{(5)} = (1.705, 1.753, 1.538, 1.776, 1.716)$ → $\hat{\theta}(\mathbf{y}^{(5)})$

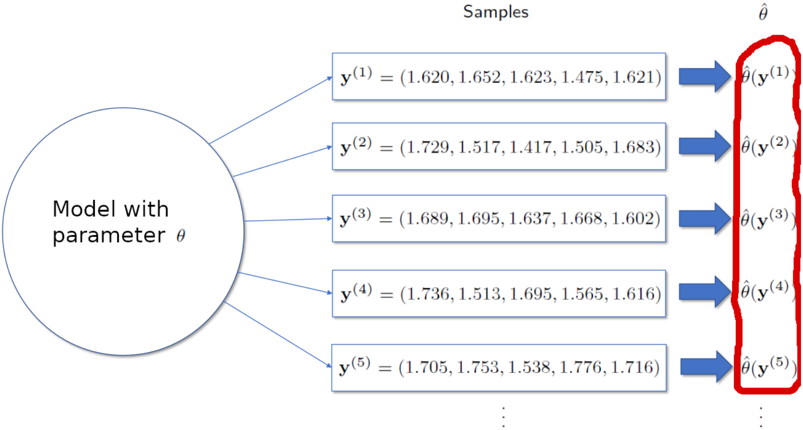
$\vdots$

What if we could generate many other samples?

Then we could understand how estimation works.

But we cannot!!

Model-based Sampling



Samples

$\hat{\theta}$

Model with parameter θ

$\mathbf{y}^{(1)} = (1.620, 1.652, 1.623, 1.475, 1.621)$ $\hat{\theta}(\mathbf{y}^{(1)})$

$\mathbf{y}^{(2)} = (1.729, 1.517, 1.417, 1.505, 1.683)$ $\hat{\theta}(\mathbf{y}^{(2)})$

$\mathbf{y}^{(3)} = (1.689, 1.695, 1.637, 1.668, 1.602)$ $\hat{\theta}(\mathbf{y}^{(3)})$

$\mathbf{y}^{(4)} = (1.736, 1.513, 1.695, 1.565, 1.616)$ $\hat{\theta}(\mathbf{y}^{(4)})$

$\mathbf{y}^{(5)} = (1.705, 1.753, 1.538, 1.776, 1.716)$ $\hat{\theta}(\mathbf{y}^{(5)})$

$\vdots$ $\vdots$

If we assume a particular model is true, then we can generate many samples. This lets us do “what if” experiments.

Sampling Distributions

- This is model-based reasoning for doing “what if” thought experiments.
- We assume a particular model family holds with parameter θ.
- We can then investigate properties of the sampling distribution and see how well inference works:
  - Quantifying accuracy of an estimate (confidence intervals)
  - Determining how unlikely a statistic is (hypothesis testing)
  - Comparing and evaluating quality of estimators
- But it all assumes a particular model.
- Now look at the examples at the *Sampling demo*

Outline

Overview of Statistical Inference

Sampling Distributions

Distributions From the Gaussian

Point Estimation

Minimising Squared Error

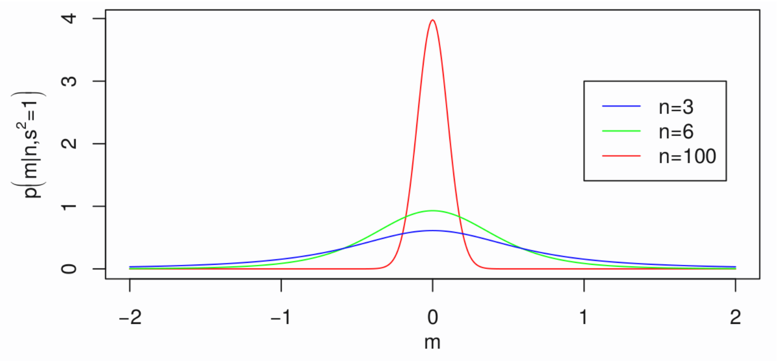
Maximising Likelihood

Distributions from the Gaussian

- the Gaussian has a lot of special properties for its sampling distribution and estimators
- these are unique, and for us best rote learned
- usually they are proven via so-called *Moment-Generating Functions* (MGFs) or using multivariate calculus with change of variables
- but we wont do this:
  - to do more advanced statistics, you should learn about MGFs
  - and about multivariate calculus for change of variables

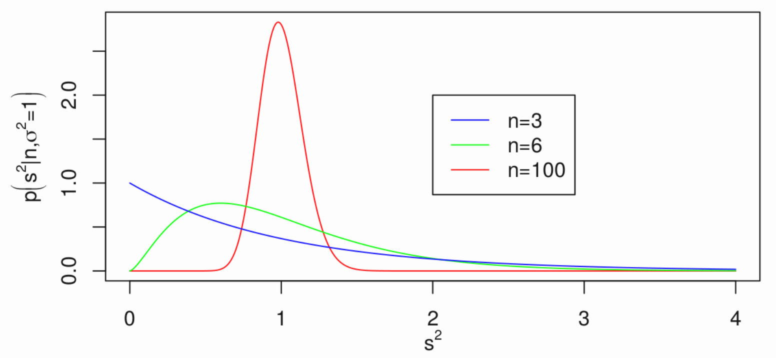
Samples from Gaussian: mean

- draw samples from a Gaussian for different sample size *n*
- assuming $s^2 = 1$, what is your distribution over mean *m*?



Samples from Gaussian: s^2

- draw samples from a Gaussian for different sample size n
- assuming $\sigma^2 = 1$, what is your distribution over sample variance s^2?



Dist. from the Gaussian

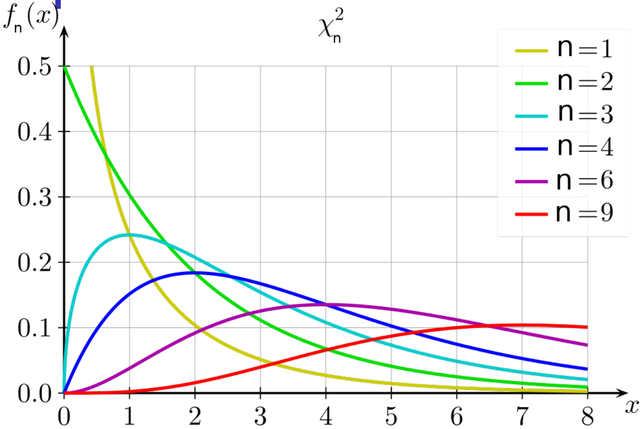
- m has distribution $N\left(\mu, \frac{1}{n}\sigma^2\right)$.
  - the central limit theorem is exact for Gaussians
- $s^2 = \frac{\sigma^2}{n-1}\chi^2_{n-1}$ where χ^2_{n-1} is a RV distributed as a chi-squared distribution with $n-1$ degrees of freedom.
  - we don't know σ^2 but we can use this to infer properties of it independent of μ
- $m = \mu + \frac{s}{\sqrt{n}} t_{n-1}$ where t_{n-1} is a RV distributed as Student's t distribution with $(n-1)$ degrees of freedom.
  - we don't know μ but we can use this to infer properties of it, independent of σ^2

Chi-Squared Distribution

Chi-Squared Distribution: Let $Z_1, ..., Z_n$ be n independent standard normal distributions. Then $X = Z_1^2 + ... + Z_n^2$ is said to have the chi-squared distribution with n degrees of freedom. Denote this by $X \sim \chi_n^2$.

- its always positive
- adding n squared RVs with unit variance should average to n in the long run
  - indeed $\mathbb{E}_{\chi_n^2}[X] = n$, $\mathbb{V}_{\chi_n^2}[X] = 2n$
- chi-squared variables add if you also add their degrees of freedom
- thus if $X_1 \sim \chi_{n_1}^2$ and $X_2 \sim \chi_{n_2}^2$ then $X_1 + X_2 \sim \chi_{n_1+n_2}^2$

Chi-Squared Plot



by Geek3 (Own work) [GFDL or CC-BY 3.0], via Wikimedia Commons

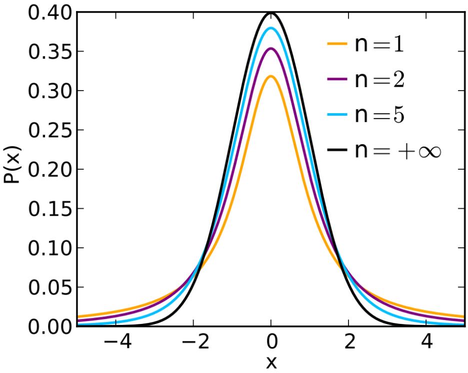
(optional) $$p(x|\chi^2_n) = \frac{1}{2^{n/2}} x^{n/2-1} e^{-x/2}$$

Student's t-Distribution

Student's t-Distribution: Let Z be a standard normal RV and let X be a chi-squared variable with n degrees of freedom. Then $T = \frac{Z}{\sqrt{X/n}}$ is said to have the Student's t-distribution with n degrees of freedom. Denote this by $T \sim \mathrm{Stu}_n$.

- it looks like a standard normal as $n \to \infty$
- is symmetric about 0
- has mean $\mathbb{E}_{\mathrm{Stu}_n}[T] = 0$ for $n > 1$
  - mean undefined for $n = 1$
- has variance $\mathbb{V}_{\mathrm{Stu}_n}[T] = \frac{n}{n-2}$ for $n > 2$
  - variance undefined for $n \leq 2$

Student's *t* Plot



by Skbkekas (Own work) [CC BY 3.0], via Wikimedia Commons

(optional)
$$p(x|\mathrm{Stu}_n) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\,\Gamma(n/2)} \left(1 + \frac{1}{n}x^2\right)^{-(n+1)/2}$$

Distributions from the Gaussian

Gaussian Sampling Distributions:
Let $X_1, ..., X_n$ be n identical RVs with distribution $N\left(\mu, \sigma^2\right)$. Let m and s^2 be the usual sample mean and variance. Then the following hold:

- m has distribution $N\left(\mu, \frac{1}{n}\sigma^2\right)$.
- $s^2 = \frac{\sigma^2}{n-1}\chi^2_{n-1}$ where χ^2_{n-1} is a RV distributed as a chi-squared distribution with $n-1$ degrees of freedom.
- $m = \mu + \frac{s}{\sqrt{n}}t_{n-1}$ where t_{n-1} is a RV distributed as Student's t distribution with $(n-1)$ degrees of freedom.

Outline

Overview of Statistical Inference

Sampling Distributions

Distributions From the Gaussian

Point Estimation

Minimising Squared Error

Maximising Likelihood

Problem Statement

Point estimation is perhaps the most basic statistical inference: trying to determine the value of a population parameter

Examples:

- What percentage of voters prefer one given political party to its main rival?
- What's 10th percentile of income in Australia?
- What's the standard deviation of income in Australia?
- What is the slope and y-intercept when regressing height onto weight?

Point Estimation Examples

Which of the hypothetical estimates below would you have the most confidence in?

- the proportion of the British electorate favouring Theresa May is 34%
- the proportion of the British electorate favouring Theresa May is 51%
- the mean distance to the sun is $1.49x10^8$ km
- Carlos Slim has assets allegedly worth $53.5B and is the richest person on Earth
- Mt. Everest is 8.8km high

Point Estimation

Simple Approach:

1. perform unbiased sampling of the quantity,
2. get a sample $x_1, ..., x_n$,
3. compute the sample mean, $m = \frac{1}{n}\sum_{i=1}^{n} x_i$

Example for Proportion: what proportion of patients θ die or have serious kidney failure when using a saline drip?

1. give a saline drip to 14,000 patients
2. for each record $x_i = 1$ if they die or have kidney failure, and $x_i = 0$ otherwise
3. compute the sample mean, $\hat{\theta} = \frac{1}{14000}\sum_{i=1}^{14000} x_i$

NB. estimators are usually identified by putting a hat on the quantity we are estimating

Point Estimation Example

Example for Variance: what is the variance of our Gaussian data?

1. collect the sample $x_1, ..., x_n$
2. compute their mean $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$
3. compute the mean of the squared errors, $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$

Point Estimation Theory (1)

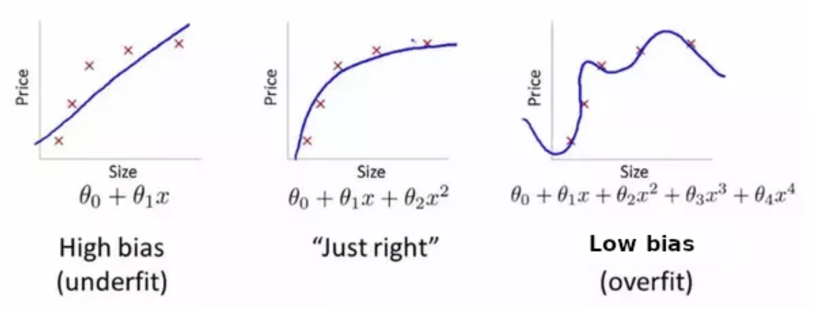
What makes a good point estimate? Hard to know. But:

Unbiased point estimate: Let $\hat{\theta}(\vec{x})$ be a point estimate for model parameter θ based on sample $\vec{x}$.
Then $\hat{\theta}(\vec{x})$ is **unbiased** if $\mathbb{E}_{\vec{x}}\left[\hat{\theta}(\vec{x})\right] = \theta$, where the expectation is taken over samples $\vec{x}$.

Note if the model parameter is computed as a pure population mean of some kind, then a sample average is always an unbiased estimator.

- for a Gaussian, $\mu = \mathbb{E}\left[x\right]$
- for a Poisson, $\lambda = \mathbb{E}\left[x\right]$
- note $\sigma^2 = \mathbb{E}\left[(x - \mathbb{E}\left[x\right])^2\right]$ is not a pure mean, it is a mean with a mean inside, so this doesn't work!

Point Estimation Example



See *"Making Sense of the Bias / Variance Trade-off"*

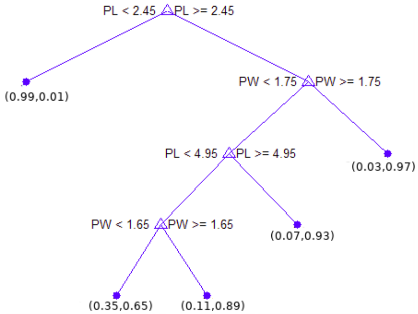
Estimating Proportions

Consider the following estimates for proportion θ from Boolean data $x_1, ..., x_n$.

Alternative point estimates $\hat{\theta}(\vec{x})$ could be

1. p_0 : a fixed value, this ignores the data, and is certainly not unbiased!
2. $\frac{1}{n}\sum_{i=1}^{n} x_i$: sample average, is unbiased
3. $\frac{1}{n+1}\left(p_0 + \sum_{i=1}^{n} x_i\right)$: offset the average with a initial value p_0 weighted by 1, not unbiased, but maybe OK if p_o is a good guess!
4. $\frac{2}{n}\sum_{i=1}^{n/2} x_{2i}$: only look at every second value, is unbiased, but wastes half the data, might be OK if you are streaming massive data

Estimating Proportions



Consider the estimate

$$\hat{\theta}(\vec{x}) = \frac{1}{n+1}\left(p_0 + \sum_{i=1}^{n} x_i\right)$$

at each leaf.

- the tree is grown to make probabilities more extreme, so pushing back a bit to centre may be OK
- could use $p_0 = 0.5$
- could set p_0 to be the proportion exhibited at a higher node
- both work well, and empirically better than the unbiased average!

Unbiasedness is not necessarily a good thing!

Is Sample Variance Unbiased?

Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ then

$$\begin{aligned}
\mathbb{E}_{\vec{x}}\left[s^2\right] &= \mathbb{E}_{\vec{x}}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2\right] \\
\ldots & \quad \text{“a lot of math to simplify”} \\
&= \sigma^2
\end{aligned}$$

So the reason we use the $\frac{1}{n-1}$ term (and not $\frac{1}{n}$) is to make the estimate unbiased!

Point Estimation Theory (2)

Characterising a point estimate: Let $\hat{\theta}(\vec{x})$ be a point estimate for model parameter θ based on sample $\vec{x}$.
Then **the bias of the estimator** is

$$b_\theta(\hat{\theta}) = \mathbb{E}_{\vec{x}}\left[\hat{\theta}(\vec{x})\right] - \theta$$

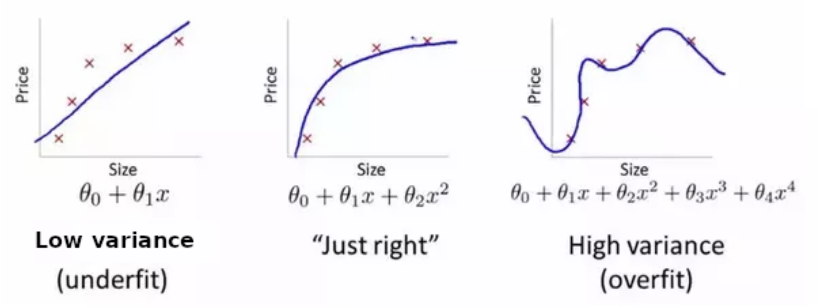
and the **variance of the estimator** is

$$\mathbb{V}_\theta\left[\hat{\theta}\right] = \mathbb{E}_{\vec{x}}\left[\left(\hat{\theta}(\vec{x}) - \mathbb{E}_{\vec{x}}\left[\hat{\theta}(\vec{x})\right]\right)^2\right]$$

and the **mean square error (MSE) of the estimator** is

$$\mathrm{MSE}_\theta\left[\hat{\theta}\right] = \mathbb{E}_{\vec{x}}\left[\left(\hat{\theta}(\vec{x}) - \theta\right)^2\right]$$

Variance Example



See *"Making Sense of the Bias / Variance Trade-off"*

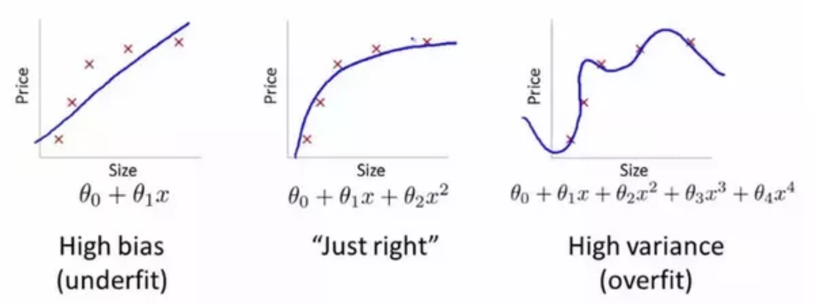
Point Estimation Measures

- the bias of the estimator, $b_\theta(\hat{\theta})$, is the average difference to the true value
  - if its non-zero, the estimator is systematically biased
- the variance of the estimator, $\mathbb{V}_\theta\left[\hat{\theta}\right]$ is its variance in the usual sense
  - how much it varies from itself
- the MSE of the estimator, $\mathrm{MSE}_\theta\left[\hat{\theta}\right]$, is its mean square error with the true value
  - how much it varies from the "truth"
- we would hope bias and variance both go to zero as the sample size n goes to infinity
  - there is a formal version of this known as consistency

$$\boxed{\mathrm{MSE}_\theta\left[\hat{\theta}\right] = b_\theta(\hat{\theta})^2 + \mathbb{V}_\theta\left[\hat{\theta}\right]}$$

Bias-Variance Tradeoff

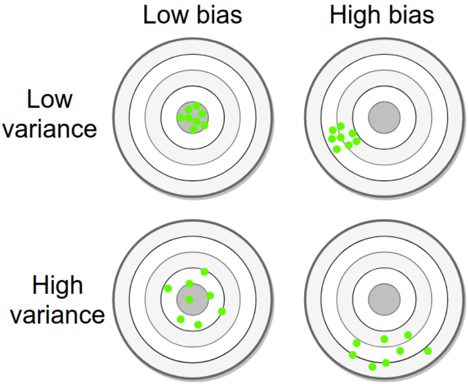
$$\mathrm{MSE}_\theta\left[\hat{\theta}\right] = b_\theta(\hat{\theta})^2 + \mathbb{V}_\theta\left[\hat{\theta}\right]$$



See *"Making Sense of the Bias / Variance Trade-off"*

Bias-Variance Tradeoff

$$\boxed{\mathbb{MSE}_\theta\left[\hat{\theta}\right] = b_\theta(\hat{\theta})^2 + \mathbb{V}_\theta\left[\hat{\theta}\right]}$$



See *“The bias-variance tradeoff”*

Estimating Proportions

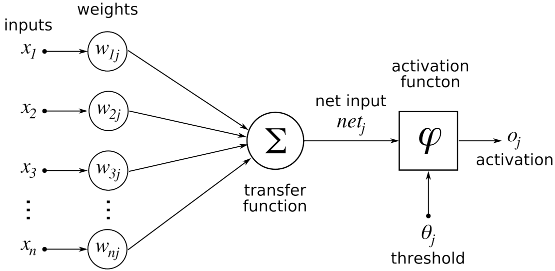
Boolean data $x_1, ..., x_n$. Alternative point estimates $\hat{\theta}(\vec{x})$ could be:

code	form	bias	variance
(A)	p_0	$(p_0 - \theta)$	0
(B)	$\frac{1}{n}\sum_{i=1}^{n} x_i$	0	$\frac{1}{n}\theta(1-\theta)$
(C)	$\frac{1}{n+1}\left(p_0 + \sum_{i=1}^{n} x_i\right)$	$\frac{p_0-\theta}{n+1}$	$\left(\frac{n}{n+1}\right)^2 \frac{1}{n}\theta(1-\theta)$
(D)	$\frac{2}{n}\sum_{i=1}^{n/2} x_{2i}$	0	$\approx \frac{2}{n}\theta(1-\theta)$

- form (C) gets a decrease in variance but an increase in bias over the simple mean, form (B)
- calculation shows form (C) beats form (B) MSE when $p_0 \in [\theta - 2\sqrt{\theta(1-\theta)}, \theta + 2\sqrt{\theta(1-\theta)}]$

  i.e. when p_0 is within $\sqrt{2}$ sd.s of θ
- (B), (C) and (D) are consistent

What About Neural Networks?



(from Wikimedia Commons by Chrislb, 2005)

How do we do point estimates for w_{ij} and θ_j give a set of data values in the form $(x_1, \ldots, x_n, o_1, \ldots, o_m)$?

No simple means appear to work.

We need another scheme (other than averages) for doing point estimates!

Outline

Overview of Statistical Inference

Sampling Distributions

Distributions From the Gaussian

Point Estimation

Minimising Squared Error

Maximising Likelihood

Problem Statement

- Imagine we have observed some data $\mathbf{y} = (y_1, \ldots, y_n)$
  $\Longrightarrow$ **y** commonly used to denote data
- For example, heights of people in a classroom

$$\mathbf{y} = (1.75, 1.64, 1.81, 1.55, 1.51, 1.67, 1.83, 1.63, 1.72)$$

- We would like to model these using a normal distribution

$$p(y \,|\, \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} \exp\left(-\frac{(y-\mu)^2}{\sigma^2} \right)$$

  but of course, the *population* μ and σ^2 are unknown.

- Estimation: How to use the *data* to select values of μ and σ^2

Sum of Squared Errors

- Let's focus first on the mean, μ
  $\Longrightarrow$ This represents the centre of the normal distribution
- One heuristic approach might be to choose a μ that is close to all the data points
- How do we measure closeness?
  $\Longrightarrow$ A mathematically convenient measure is squared error:

$$\mathrm{sse}(\mu) = \sum_{i=1}^{n} (y_i - \mu)^2$$

- Squared-error is obviously always greater than zero
- To estimate μ using this approach: adjust μ until $\mathrm{sse}(\mu)$ attains its minimum

SSE Plot

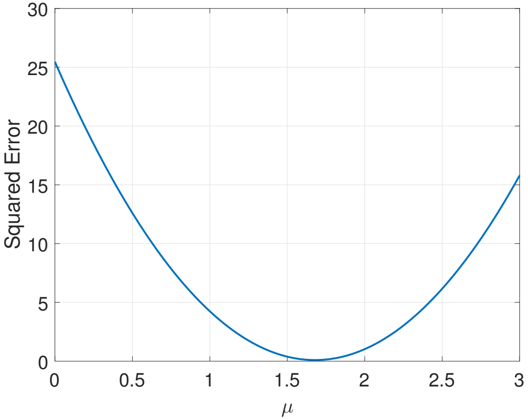


Figure: Sum of squared errors (sse) as a function of the parameter μ for our example data set. There is one clear minimum.

Minimising SSE

- Formally we can write this process as

$$\hat{\mu} = \operatorname*{arg\,min}_{\mu} \left\{ \sum_{i=1}^{n} (y_i - \mu)^2 \right\},$$

where

- $\operatorname*{arg\,min}_{x} \{f(x)\}$ means find the value of x that minimises $f(x)$

Minimising SSE, cont.

Due to choice of squared error, the estimate is easy to find:

1. First, differentiate $\mathrm{sse}(\mu)$ with respect to μ

$$\begin{aligned}\frac{d\,\mathrm{sse}(\mu)}{d\mu} &= \sum_{i=1}^{n} \frac{d}{d\mu}(y_i - \mu)^2 = -2\sum_{i=1}^{n}(y_i - \mu) \\ &= -2\sum_{i=1}^{n} y_i + 2n\mu\end{aligned}$$

2. Then set the derivative to zero, and solve for μ, yielding:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

which is readily identified as the **sample mean.**

SSE Example

- Recall our example data set:

$$\mathbf{y} = (1.75, 1.64, 1.81, 1.55, 1.51, 1.67, 1.83, 1.63, 1.72)$$

- In this case, $\hat{\mu} = 1.6789$ if we minimise squared error
- But the normal distribution has two parameters: μ and σ ...
  $\Longrightarrow$ How do we estimate σ?
- The minimum error approach offers no obvious measure of goodness-of-fit for σ
- A more general approach is required

Outline

Overview of Statistical Inference

Sampling Distributions

Distributions From the Gaussian

Point Estimation

Minimising Squared Error

Maximising Likelihood

Maximum Likelihood

- One solution: *maximum likelihood estimation*, we call MLE
- This is a very general procedure for estimating parameters of statistical models
  - Was first proposed in the 1920s by Ronald Fisher (1890–1962)
  - Heuristic procedure that has been shown to have many strong properties
  - Widely applicable to many models

MLE Context

- Let us consider a probability model with parameter(s) Θ
- We measure the goodness-of-fit of a model to data by the probability it assigns to the data, i.e.,

$$p(\mathbf{y} \mid \Theta)$$

- The larger the probability, the more likely the observed data would be under that model
- For many models we will examine, the probabilities of $y_1, y_2, \ldots, y_n$ are independent so that:

$$p(\mathbf{y} \mid \Theta) = \prod_{i=1}^{n} p(y_i \mid \Theta)$$

MLE Definition

Method of Maximum Likelihood: The method of maximum likelihood says we should use the model that assigns the greatest probability to the data we have observed.

Formally, the maximum likelihood (ML) estimator is found by solving

$$\hat{\Theta} = \arg\max_{\Theta}\{p(\mathbf{y} \,|\, \Theta)\}$$

where $p(\mathbf{y} \,|\, \Theta)$ is called the likelihood function.

MLE Computation

- In practice it is mathematically easier to solve the equivalent problem:

  $$\hat{\Theta} = \arg\min_{\Theta}\{-\log p(\mathbf{y} \,|\, \Theta)\}$$

  where

  $$-\log(\mathbf{y} \,|\, \Theta)$$

  is known as the negative log-likelihood.

- Sometimes we use $L(\mathbf{y} \,|\, \Theta)$ to denote the negative log-likelihood
- Sometimes, the log-likelihood is used instead.

ML Estimation of Gaussian

- Let's return to the problem of estimating μ and σ for a normal distribution
- For the normal distribution $\Theta = (\mu, \sigma)$.
- Given data $\mathbf{y} = (y_1, \ldots, y_n)$ the likelihood is

$$
\begin{aligned}
p(\mathbf{y} \,|\, \mu, \sigma) &= \prod_{i=1}^{n} \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} (\mu - y_i)^2 \right) \\
&= \prod_{i=1}^{n} \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2} (\mu - y_i)^2 \right) \\
&= \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\mu - y_i)^2 \right)
\end{aligned}
$$

- by the fact that $e^{-a} e^{-b} = e^{-a-b}$.

ML Estimation of Gaussian, cont.

- The negative log-likelihood function is then:

$$\begin{aligned} L(\mathbf{y} \mid \mu, \sigma) &= -\log p(\mathbf{y} \mid \mu, \sigma) \\ &= \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \qquad (1) \end{aligned}$$

- To minimise this for μ and σ we need to differentiate equation (1) with respect to μ and σ and find the values that set the (partial) derivatives to zero, i.e., we need to solve the simultaneous equations:

$$\begin{aligned} \partial L(\mathbf{y}|\mu, \sigma)/\partial\mu &= 0, \\ \partial L(\mathbf{y}|\mu, \sigma)/\partial\sigma &= 0. \end{aligned}$$

- It turns out for this problem, this is actually quite easy

ML Estimation of Mean

- Partial derivative with respect to μ (proof optional):

$$\begin{aligned}\frac{\partial L(\mathbf{y} \mid \mu, \sigma)}{\partial \mu} &= -\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu) \\ &= -\frac{1}{\sigma^2}\sum_{i=1}^{n} y_i + \frac{n\mu}{\sigma^2}\end{aligned}$$

  which is similar to our minimum squared error estimator.

- In fact, setting this equation to zero and solving for μ yields

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

  which is again, just the sample mean.

ML Estimation of Variance

(optional)

- However, ML also gives us a clear recipe for estimating σ
- Plugging $\hat{\mu}$ into $L(\mathbf{y}|\mu,\sigma)$ removes μ from the equation
- Partial derivative with respect to σ:

$$\begin{aligned}
\frac{\partial L(\mathbf{y}\,|\,\hat{\mu},\sigma)}{\partial\sigma} &= \frac{\partial}{\partial\sigma}\frac{n}{2}\left[\log\sigma^2+\log(2\pi)\right]+\sum_{i=1}^{n}(y_i-\hat{\mu})^2\frac{\partial}{\partial\sigma}\frac{1}{2\sigma^2} \\
&= \frac{n}{\sigma}-\frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i-\hat{\mu})^2 \qquad (2)
\end{aligned}$$

where we use the facts that

- $\log(ab)=\log b+\log b$;
- $\frac{\partial}{\partial x}Kg(z)f(x)=Kg(z)\frac{\partial}{\partial x}f(x)$; and
- $\frac{\partial}{\partial x}f(z)=0$.

MLE of Variance, cont.

(optional)

- Solving for σ:

$$
\begin{aligned}
& \frac{n}{\sigma} - \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \hat{\mu})^2 = 0 \\
\Rightarrow \quad & \sigma^3\left[\frac{n}{\sigma} - \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \hat{\mu})^2\right] = 0 \\
\Rightarrow \quad & n\sigma^2 - \sum_{i=1}^{n}(y_i - \hat{\mu})^2 = 0 \\
\Rightarrow \quad & n\sigma^2 = \sum_{i=1}^{n}(y_i - \hat{\mu})^2 \\
\Rightarrow \quad & \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu})^2
\end{aligned}
$$

MLE of Variance, cont.

- The ML estimator for σ is:

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu})^2}$$

  which can be identified as the sample standard deviation, but with n instead of $n-1$

- So, for the normal distribution, the ML estimators are:
  - sample mean for μ;
  - (modified) sample standard deviation for σ

Gaussian Example

- Recall our example data set:

$$\mathbf{y} = (1.75, 1.64, 1.81, 1.55, 1.51, 1.67, 1.83, 1.63, 1.72)$$

- Using maximum likelihood to fit a normal distribution to this data, we have:

$$\hat{\mu} = 1.6789$$

and

$$\begin{aligned}\hat{\sigma} &= \sqrt{\frac{1}{9}\sum_{i=1}^{9}(y_i - 1.6789)^2} \\ &= 0.1032\end{aligned}$$

Gaussian Example, Fit

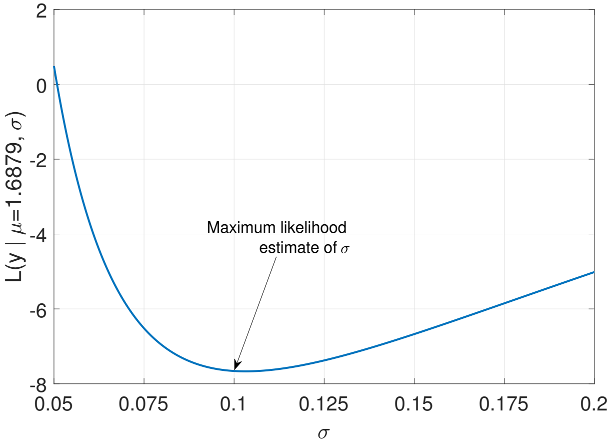


Figure: Negative log-likelihood $L(\mathbf{y} \mid \mu = \hat{\mu}, \sigma)$ as a function of σ with μ fixed at the maximum likelihood estimate $\hat{\mu} = 1.6789$.

Gaussian Example, Plot

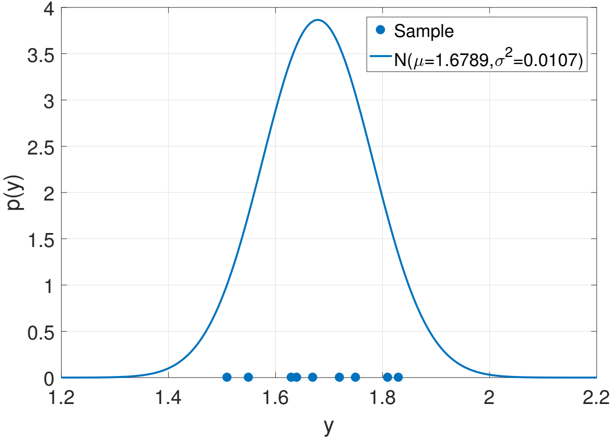


Figure: Data samples and the normal distribution fitted by maximum likelihood with $\hat{\mu} = 1.6879$ and $\hat{\sigma} = 0.1032$. Note that the bulk of the samples lie within $(\hat{\mu} - 2\hat{\sigma}, \hat{\mu} + 2\hat{\sigma}) \approx (1.47, 1.88)$

How to Use Estimates?

- We can use these estimates to make statements/predictions about the population
- To do this, we use them in the distribution, i.e., $p(y \mid \hat{\mu}, \hat{\sigma}^2)$ $\Longrightarrow$ this is called the plug-in distribution
- Can use plug-in distribution to make probability statements
- In our example, we could ask “what is the probability a person from our population has a height between $1.6m$ and $1.8m$”?, which is estimated by

$$p(1.6 < X < 1.8 \mid \hat{\mu} = 1.6879, \hat{\sigma}^2 = 0.1032) \approx 0.664$$

$\Longrightarrow$ The better our estimates, the more accurate the answers

Properties of ML

- The original maximum likelihood proposal was heuristic
- But a large body of research since has shown ML has many good theoretical properties
- For Poisson $\mathrm{Poi}(\lambda)$ with count data $\vec{y} = (y_1, ..., y_n)$, the MLE is

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

- For Bernoulli $\mathrm{Be}(\theta)$ with binary data $\vec{y} = (y_1, ..., y_n)$, the MLE is

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

- For Binomial $\mathrm{Bin}(\theta, m)$ with count data $\vec{y} = (y_1, ..., y_n)$, the MLE is

$$\hat{\theta} = \frac{1}{nm} \sum_{i=1}^{n} y_i$$

End of Week 5