DATA70121: Statistics and Machine Learning 1

Lecture 4: Estimation

Diego Perez Ruiz
School of Social Sciences
University of Manchester

19 October 2023

So you want to be a *data scientist*? There is no widely accepted definition of who a data scientist is. [1] My personal viewpoint is that a

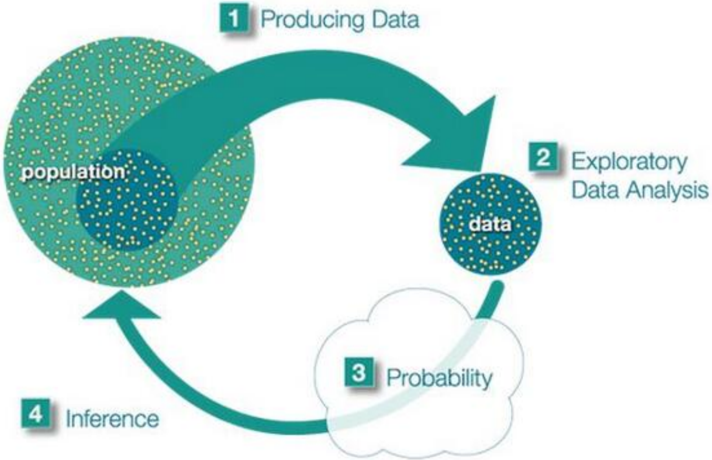
data scientist is someone who asks unique, interesting questions of data based on formal or informal theory, to generate rigorous and useful insights.

[1] The term "data scientist" was coined by D.J. Patil. He was the Chief Scientist for LinkedIn. In 2011 Forbes placed him second in their Data Scientist List, just behind Larry Page of Google.

It is likely to be an individual with multi-disciplinary training in computer science, business, economics, statistics, and armed with the necessary quantity of domain knowledge relevant to the question at hand.

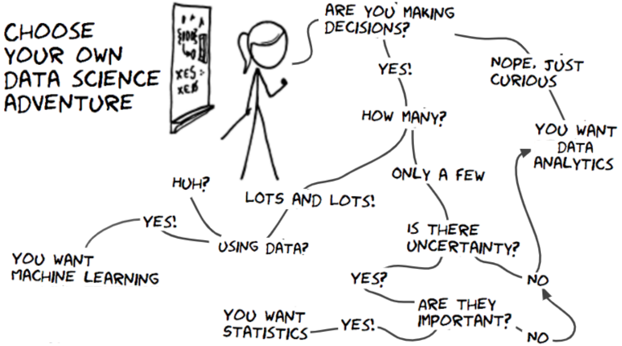
The potential of the field is enormous for just a few well-trained data scientists armed with big data have the potential to transform organisations and societies.

Part 1


1 Producing Data
population
2 Exploratory Data Analysis
data
3 Probability
4 Inference

exploratory versus inference

looking at the data and reporting what's there
(*just go with the estimate!*)
or
concluding w.r.t the underlying population from which data is from
(*hypothesis testing*)



basic concepts of estimation

estimator

statistic (function of observations) whose calculated value is used to estimate a parameter, θ

estimate

specific realization of an estimator

$$\hat{\theta} = g(X_1, X_2, \ldots, X_n)$$

where $X_1, X_2, \ldots, X_n$ are

independent and identically distributed

with pdf/pmf denoted $f(x_i|\theta)$

basic concepts of estimation

estimator

statistic (function of observations) whose calculated value is used to estimate a parameter, θ

estimate

specific realization of an estimator

$$\hat{\theta} = g(X_1, X_2, \dots, X_n)$$

where $X_1, X_2, \dots, X_n$ are

independent and identically distributed

with pdf/pmf denoted $f(x_i|\theta)$

$\hat{\theta}$ is a random variable (r.v.) and has a **sampling distribution**

basic concepts of estimation

estimator

statistic (function of observations) whose calculated value is used to estimate a parameter, θ

estimate

specific realization of an estimator

$$\hat{\theta} = g(X_1, X_2, \ldots, X_n)$$

where $X_1, X_2, \ldots, X_n$ are

independent and identically distributed

with pdf/pmf denoted $f(x_i|\theta)$

$\hat{\theta}$ is a random variable (r.v.) and has a **sampling distribution**

- **point estimate**
  single number regarded as the most plausible value of θ
- **interval estimate**
  a range of numbers, called a confidence interval, and likely to contain the true value of θ

basic concepts of estimation

example. if $X_1, X_2, \ldots, X_n$ is a random sample from some population distribution with mean μ and variance σ^2, then the sample average

$$\hat{\mu} = g(X_1, X_2, \ldots, X_n) = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

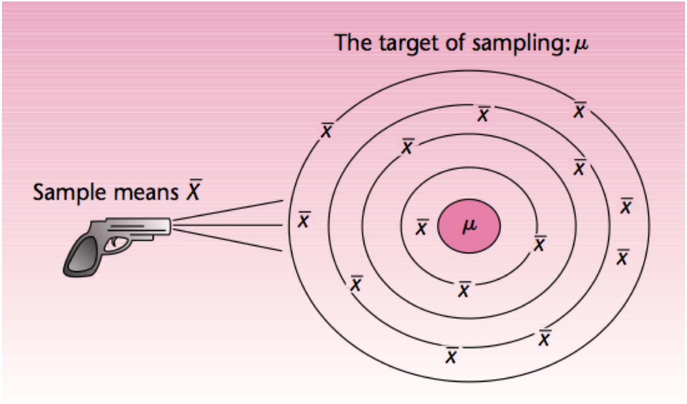
is an estimator of the population mean, and

$$\hat{\sigma}^2 = g(X_1, X_2, \ldots, X_n) = s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

is an estimator of the population variance.

assessing point estimators

there are many potential estimators for a population parameter



what are good properties of estimators?

a good estimator is unbiased

- is the mean of the estimator close to the actual parameter?

- recall: $\hat{\theta} = g(X_1, X_2, \ldots, X_n)$ is a r.v. with a sampling distribution

- $\hat{\theta} = g(X_1, X_2, \ldots, X_n)$ is a good estimator for θ
  if the values it typically takes are close θ

a good estimator is unbiased

- is the mean of the estimator close to the actual parameter?

- recall: $\hat{\theta} = g(X_1, X_2, \ldots, X_n)$ is a r.v. with a sampling distribution

- $\hat{\theta} = g(X_1, X_2, \ldots, X_n)$ is a good estimator for θ
  if the values it typically takes are close θ

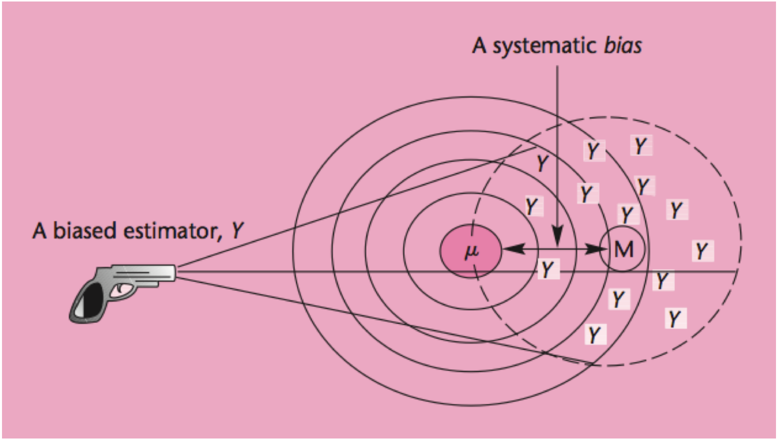
- look at a central value from a distribution: the expectation $\mathbb{E}[\hat{\theta}]$

- concept of bias:
$$\text{bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

- if bias$(\hat{\theta}) = 0$, then
$$\mathbb{E}[\hat{\theta}] = \theta$$

  and estimator is said to be **unbiased**

example: biased estimator



a good estimator is precise

- is the st. dev. of the estimator close to the actual parameter?
- when we calculate an estimate of θ we have some random error
- aim: the magnitude of random error should be small on average
- mean square error of $\hat{\theta} = g(X_1, X_2, \ldots, X_n)$

$$MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$$

- if $\hat{\theta}$ is unbiased, then

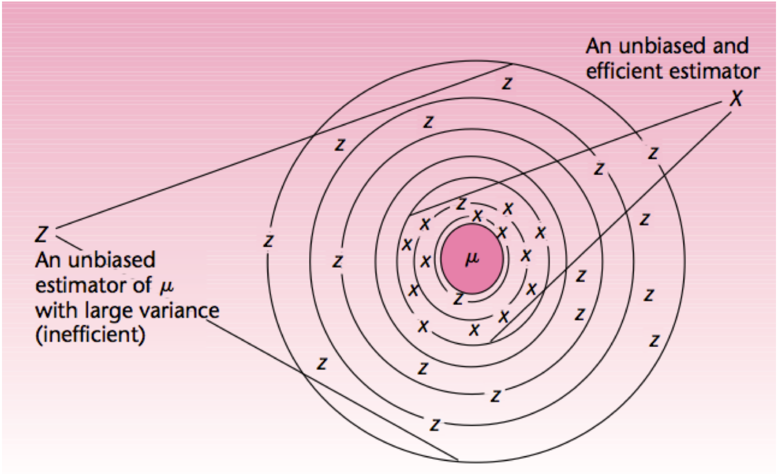
$$\mathbb{E}[\hat{\theta}] = \theta \quad \text{and} \qquad MSE(\hat{\theta}) = \mathrm{Var}(\hat{\theta})$$

  and we have an **efficient estimator**

- it can also be shown that

$$MSE(\hat{\theta}) = \mathrm{bias}^2(\hat{\theta}) + \mathrm{Var}(\hat{\theta})$$

example: efficient estimator



a good estimator is consistent

- is the probability distribution of the estimator concentrated on the parameter as the sample sizes increases?
- as the random sample of size n increases, $\hat{\theta}$ gets closer to θ

$$\hat{\theta} \to \theta \quad \text{as} \quad n \to \infty$$

- we say that $\hat{\theta} = g(X_1, X_2, \ldots, X_n)$ is **consistent** if, for all ϵ

$$\lim_{n \to \infty} P(|\hat{\theta} - \theta| > \epsilon) = 0$$

a good estimator is consistent

- is the probability distribution of the estimator concentrated on the parameter as the sample sizes increases?

- as the random sample of size n increases, $\hat{\theta}$ gets closer to θ

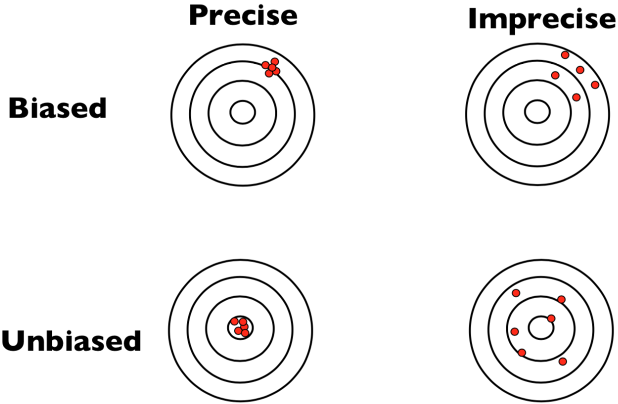
$$\hat{\theta} \to \theta \quad \text{as} \quad n \to \infty$$

- we say that $\hat{\theta} = g(X_1, X_2, \ldots, X_n)$ is **consistent** if, for all ϵ

$$\lim_{n \to \infty} P(|\hat{\theta} - \theta| > \epsilon) = 0$$

* curiosity: $\bar{X} + \frac{1}{n}$ is a biased but consistent estimator

summary



Part 1A

the sample mean

the sample mean is a consistent and unbiased estimator of the mean of the underlying distribution

the sample mean

the sample mean is a consistent and unbiased estimator of the mean of the underlying distribution

$$\theta = \mu \qquad\qquad \hat{\theta} = g(X_1, X_2, \ldots, X_n) = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

where X_i's are iid with $\mathbb{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$

the sample mean

the sample mean is a consistent and unbiased estimator of the mean of the underlying distribution

$$\theta = \mu \qquad\qquad \hat{\theta} = g(X_1, X_2, \ldots, X_n) = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

where X_i's are iid with $\mathbb{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$

expected value of $\hat{\theta}$:

$$\begin{aligned}
\mathbb{E}(\overline{X}) &= \mathbb{E}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu \\
&= \frac{1}{n} n\mu = \mu \quad \textbf{unbiased}
\end{aligned}$$

the sample mean

the sample mean is a consistent and unbiased estimator of the mean of the underlying distribution

$$\theta = \mu \qquad\qquad \hat{\theta} = g(X_1, X_2, \ldots, X_n) = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

where X_i's are iid with $\mathbb{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$

expected value of $\hat{\theta}$:

$$\begin{aligned}
\mathbb{E}(\overline{X}) &= \mathbb{E}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu \\
&= \frac{1}{n} n\mu = \mu \quad \textbf{unbiased}
\end{aligned}$$

variance of $\hat{\theta}$:

$$\begin{aligned}
\mathrm{Var}(\overline{X}) &= \mathrm{Var}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \\
&= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \quad \textbf{consistent}
\end{aligned}$$

Part 2

method of point estimation

the likelihood principle

example. assume we have two biased dice 

the frequencies for which numbers 1–6 appear on each die

die face	die nr 1	die nr 2
1	0.1	0.5
2	0.1	0.1
3	0.1	0.1
4	0.1	0.1
5	0.1	0.1
6	0.5	0.1

method of point estimation

the likelihood principle

example. assume we have two biased dice 

the frequencies for which numbers 1–6 appear on each die

die face	die nr 1	die nr 2
1	0.1	0.5
2	0.1	0.1
3	0.1	0.1
4	0.1	0.1
5	0.1	0.1
6	0.5	0.1

we roll one die and it shows a 6, which die do you think was rolled?

method of point estimation

the likelihood principle

example. assume we have two biased dice 

the frequencies for which numbers 1–6 appear on each die

die face	die nr 1	die nr 2
1	0.1	0.5
2	0.1	0.1
3	0.1	0.1
4	0.1	0.1
5	0.1	0.1
6	0.5	0.1

we roll one die and it shows a 6, which die do you think was rolled?

the idea underlying maximum likelihood estimation:
estimate $\hat{\theta}$ to be the value that makes the data most likely

method of point estimation

the likelihood principle

before an experiment

- outcome is **unknown**
- **probability** allows us to predict unknown outcomes based on **known** parameters

$$P(\text{data}|\theta)$$

method of point estimation

the likelihood principle

before an experiment

- outcome is **unknown**
- **probability** allows us to predict unknown outcomes based on **known** parameters

$$P(\text{data}|\theta)$$

after an experiment

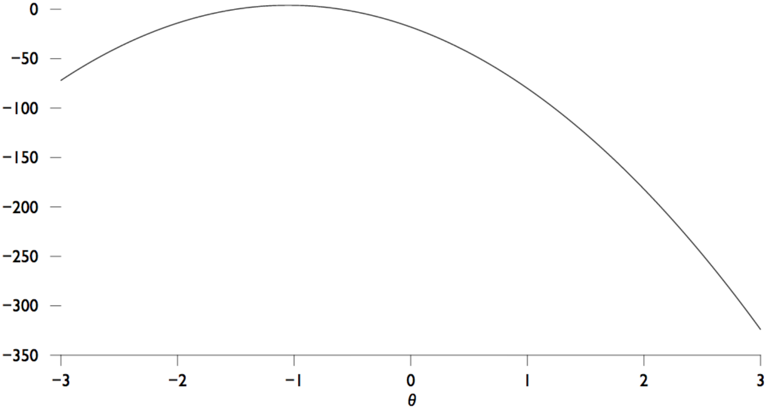
- outcome is **known**
- we can consider the **likelihood** L that a parameter would generate the observed data

$$L(\theta|\text{data})$$

- $L(\theta)$ is a surface in θ space that shows which parameter values are more likely than others

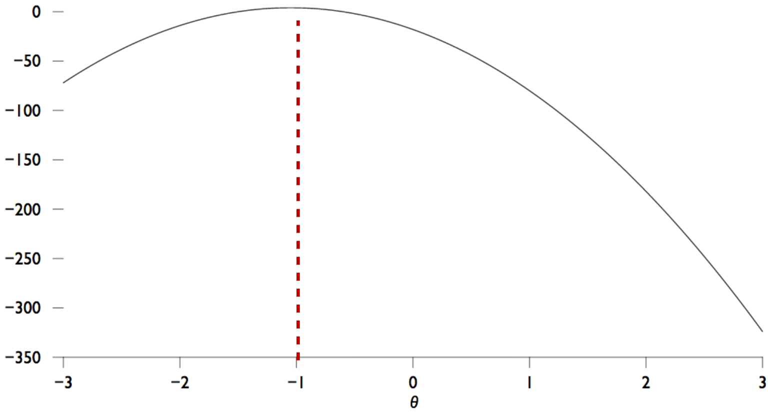
maximum likelihood estimation

likelihood that θ generated the data



maximum likelihood estimation

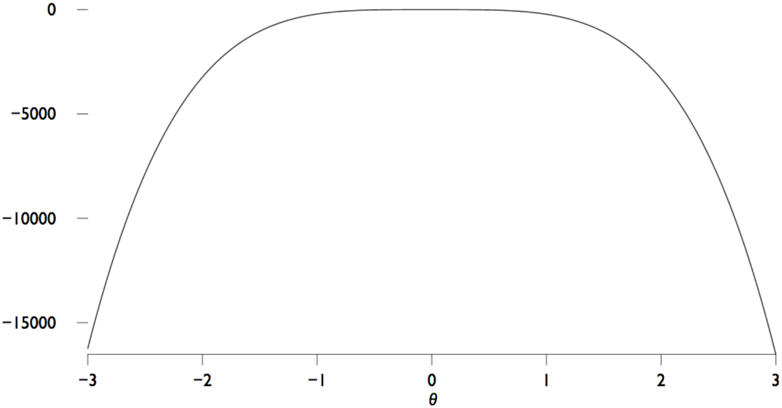
likelihood that θ generated the data



the most likely θ produces the largest likelihood

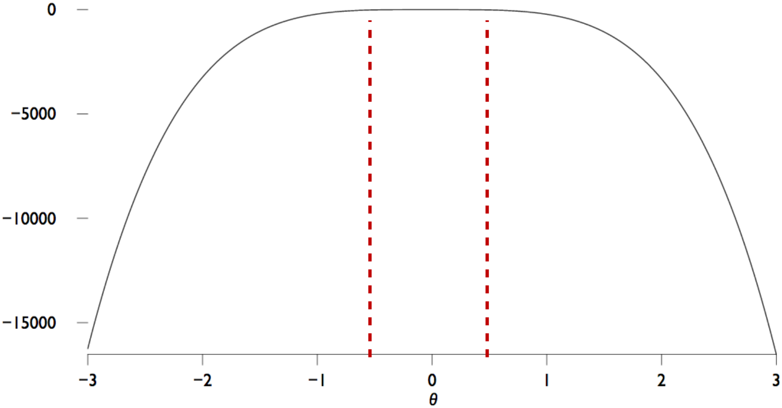
maximum likelihood estimation

likelihood that θ generated the data



maximum likelihood estimation

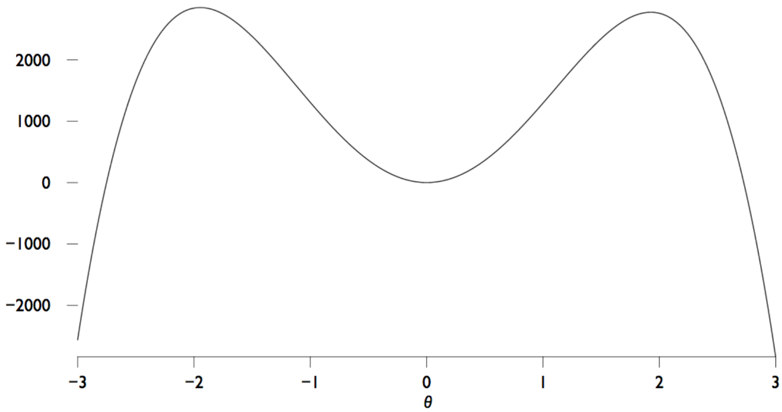
likelihood that θ generated the data



a flat maximum likelihood makes θ uncertain

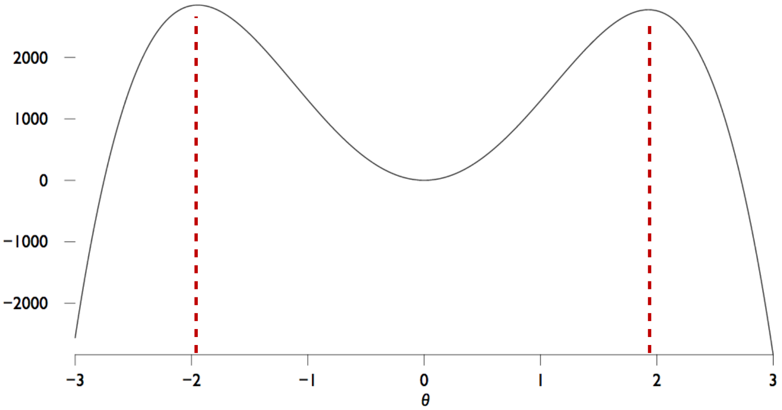
maximum likelihood estimation

likelihood that θ generated the data



maximum likelihood estimation

likelihood that θ generated the data



multimodal surfaces are more complex...

maximum likelihood estimation

precision of $\hat{\theta}_{mle}$ depends on the shape of likelihood near $\hat{\theta}_{mle}$

- if the likelihood is very curved or 'steep' around $\hat{\theta}_{mle}$:
  - θ will be precisely estimated
  - we have a lot of information about θ

maximum likelihood estimation

precision of $\hat{\theta}_{mle}$ depends on the shape of likelihood near $\hat{\theta}_{mle}$

- if the likelihood is very curved or 'steep' around $\hat{\theta}_{mle}$:
  - θ will be precisely estimated
  - we have a lot of information about θ
- if the likelihood is not curved or 'flat' around $\hat{\theta}_{mle}$:
  - θ will not be precisely estimated
  - we don't have much information about θ

maximum likelihood estimation

precision of $\hat{\theta}_{mle}$ depends on the shape of likelihood near $\hat{\theta}_{mle}$

- if the likelihood is very curved or 'steep' around $\hat{\theta}_{mle}$:
  - θ will be precisely estimated
  - we have a lot of information about θ
- if the likelihood is not curved or 'flat' around $\hat{\theta}_{mle}$:
  - θ will not be precisely estimated
  - we don't have much information about θ
- if the likelihood is completely flat:
  - the sample contains no information about θ
  - every value of θ produces same value of likelihood function
  - we say that θ can not be identified

maximum likelihood estimation

how to derive mle's

- let the vector $\mathbf{x} = (x_1, \dots, x_n)$ be the observed sample
- $\mathbf{x}$ is iid with pdf/pmf $f(x_i|\theta)$ where θ is a vector of parameters
- the joint density of the sample, by independence, is equal to

$$f(\mathbf{x}|\theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

- the likelihood function is equal to

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

- the mle is the value that maximizes $L(\theta|\mathbf{x})$

$$\hat{\theta}_{mle} = \arg\max_{\theta} L(\theta|\mathbf{x})$$

maximum likelihood estimation

how to derive mle's

- how do find the maximum?
- easier with the **log likelihood**

$$\ell(\theta|\mathbf{x}) = \ln L(\theta|\mathbf{x}) = \ln\left(\prod_{i=1}^{n} f(x_i|\theta)\right) = \sum_{i=1}^{n} \ln f(x_i|\theta)$$

- also, the following simplifies working with log likelihood:

$$\ln(x \cdot y) = \ln(x) + \ln(y)$$

$$\ln\left(\frac{x}{y}\right) = \ln(x) - \ln(y)$$

$$\ln(x^y) = y \ln x$$

$$\ln(e^x) = x$$

maximum likelihood estimation

how to derive mle's

- differentiate $\ell(\theta|\mathbf{x})$ w.r.t θ
- the derivative of the log-likelihood is the **score function**
- to find mle, set score function to 0 and solve

$$\frac{\partial \ell(\hat{\theta}_{mle}|\mathbf{x})}{\partial \theta} = 0$$

- use second derivatives to prove that an estimator is maximum

$$\frac{\partial^2 \ell(\hat{\theta}_{mle}|\mathbf{x})}{\partial \theta^2} < 0$$

Part 2A

maximum likelihood estimation

Bernoulli distribution

assume iid sample with Bernoulli random variables $\mathbf{x} = x_1, \ldots, x_n$

$$P(X = x) = p^x (1-p)^{1-x}, \quad x \in \{0, 1\}$$

Likelihood is the binomial pmf with two parameters:

- n (number of trials)
- p (probability of success)

$$P(Y = y) = \frac{n!}{y!\,(n-y)!} p^y (1-p)^{n-y}$$

$$y \in \{0, 1, \ldots, n\}$$

maximum likelihood estimation

Bernoulli distribution

assume iid sample with Bernoulli random variables $\mathbf{x} = x_1, \ldots, x_n$

$$P(X = x) = p^x (1-p)^{1-x}, \quad x \in \{0, 1\}$$

Likelihood is the binomial pmf with two parameters:

- n (number of trials)
- p (probability of success)

$$P(Y = y) = \frac{n!}{y!\,(n-y)!} p^y (1-p)^{n-y}$$

$$y \in \{0, 1, \ldots, n\}$$

likelihood and log likelihood:

$$L(p|\mathbf{x}) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i}$$

$$= p^{\sum_{i=1}^{n} x_i} (1-p)^{n - \sum_{i=1}^{n} x_i}, \qquad y = \sum_{i=1}^{n} x_i$$

maximum likelihood estimation

Bernoulli distribution

log likelihood:

$$\ell(p|\mathbf{x}) = \sum_{i=1}^{n} x_i \ln(p) + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p)$$

maximum likelihood estimation

Bernoulli distribution

log likelihood:

$$\ell(p|\mathbf{x}) = \sum_{i=1}^{n} x_i \ln(p) + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p)$$

first derivative set to zero:

$$\frac{\partial \ell(p|\mathbf{x})}{\partial p} = \frac{1}{p} \sum_{i=1}^{n} x_i - \frac{1}{1-p} \left(n - \sum_{i=1}^{n} x_i\right) = 0$$

maximum likelihood estimation

Bernoulli distribution

log likelihood:

$$\ell(p|\mathbf{x}) = \sum_{i=1}^{n} x_i \ln(p) + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p)$$

first derivative set to zero:

$$\frac{\partial \ell(p|\mathbf{x})}{\partial p} = \frac{1}{p} \sum_{i=1}^{n} x_i - \frac{1}{1-p} \left(n - \sum_{i=1}^{n} x_i\right) = 0$$

the mle:

$$\hat{p}_{mle} = \frac{\sum_{i=1}^{n} x_i}{n}$$

maximum likelihood estimation

Bernoulli distribution

log likelihood:

$$\ell(p|\mathbf{x}) = \sum_{i=1}^{n} x_i \ln(p) + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p)$$

first derivative set to zero:

$$\frac{\partial \ell(p|\mathbf{x})}{\partial p} = \frac{1}{p} \sum_{i=1}^{n} x_i - \frac{1}{1-p} \left(n - \sum_{i=1}^{n} x_i\right) = 0$$

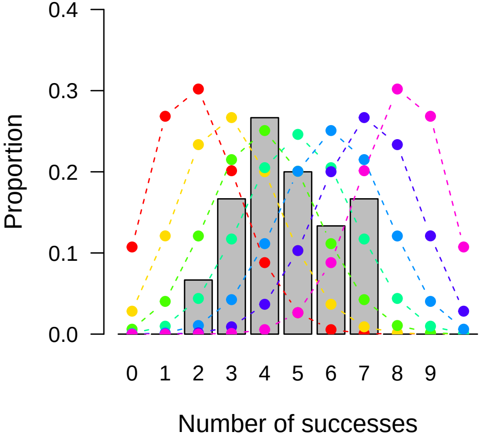
the mle:

$$\hat{p}_{mle} = \frac{\sum_{i=1}^{n} x_i}{n}$$

is the estimator unbiased?
is it consistent?

maximum likelihood estimation

binomial distribution



maximum likelihood estimation

normal distribution

assume iid sample with normal random variables $\mathbf{x} = x_1, \ldots, x_n$

the normal pdf has two parameters:

- mean μ
- variance σ^2

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

maximum likelihood estimation

normal distribution

assume iid sample with normal random variables $\mathbf{x} = x_1, \ldots, x_n$

the normal pdf has two parameters:

- mean μ
- variance σ^2

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

after some derivation we get

$$\frac{\partial \ell(\mu, \sigma^2 | \mathbf{x})}{\partial \mu} = 0 \implies \hat{\mu}_{mle} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

$$\frac{\partial \ell(\mu, \sigma^2 | \mathbf{x})}{\partial \sigma^2} = 0 \implies \hat{\sigma}^2_{mle} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

maximum likelihood estimation

normal distribution

assume iid sample with normal random variables $\mathbf{x} = x_1, \ldots, x_n$

the normal pdf has two parameters:

- mean μ
- variance σ^2

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

after some derivation we get

$$\frac{\partial \ell(\mu, \sigma^2 | \mathbf{x})}{\partial \mu} = 0 \implies \hat{\mu}_{mle} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

$$\frac{\partial \ell(\mu, \sigma^2 | \mathbf{x})}{\partial \sigma^2} = 0 \implies \hat{\sigma}^2_{mle} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

are the two estimators unbiased?
are they consistent?

maximum likelihood estimation

properties of mle's

maximum-likelihood estimators are

- consistent
  - mle spikes over true parameter values as $n \to \infty$
- asymptotically unbiased
  - although they may be biased in finite samples
- asymptotically efficient
  - as $n \to \infty$, mle tends to be the estimator with lowest error
  - no asymp. unbiased estimator has smaller asymptotic variance
- asymptotically normally distributed
  - for large n, sampling distribution of $\hat{\theta}$ becomes normal
  - easy calculation of standard errors, confidence intervals, etc

DATA70121: Statistics and Machine Learning 1
Lecture 4: Estimation

Part 3: interval estimation

19 October 2023

interval estimation

example

- **Research question**

interval estimation

example

- **Research question**
  - What is the average annual income of full-time students in the UK?

central limit theorem

- **Population** – clearly defined elements (full time students in the UK) that share some characteristic (income)
- **Random sample** – a selection of the elements from the population; each element has the same chance of being selected

central limit theorem

- **Population** – clearly defined elements (full time students in the UK) that share some characteristic (income)
- **Random sample** – a selection of the elements from the population; each element has the same chance of being selected

If we draw **repeated random samples** of size N from the population, the **means of the samples** are approximately **normally distributed** with

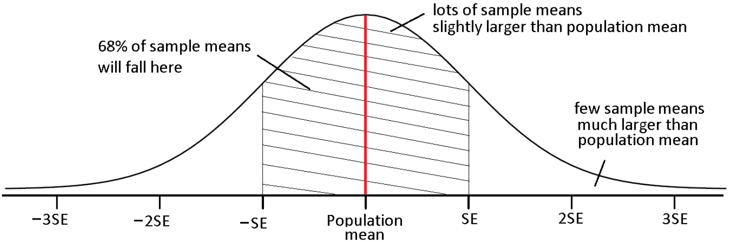
$$\text{Mean} = \text{Population mean}$$

and

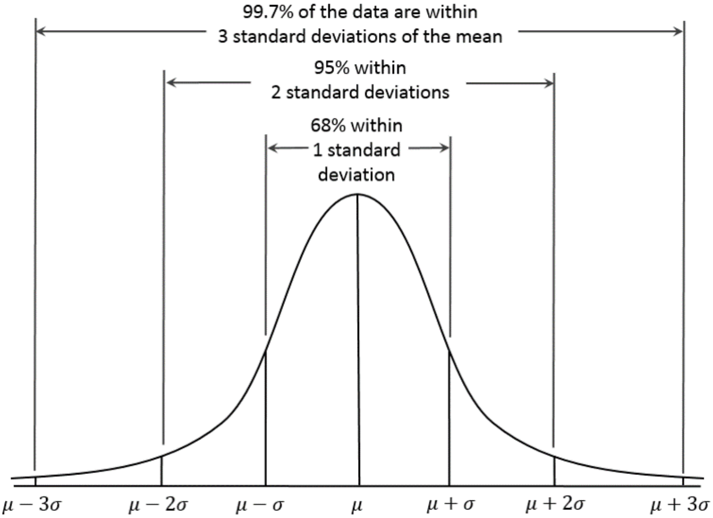
$$\text{SD} = \frac{\text{Population SD}}{\sqrt{N}}$$

This standard deviation of the sample means is known as the **standard error** (SE).

central limit theorem



central limit theorem

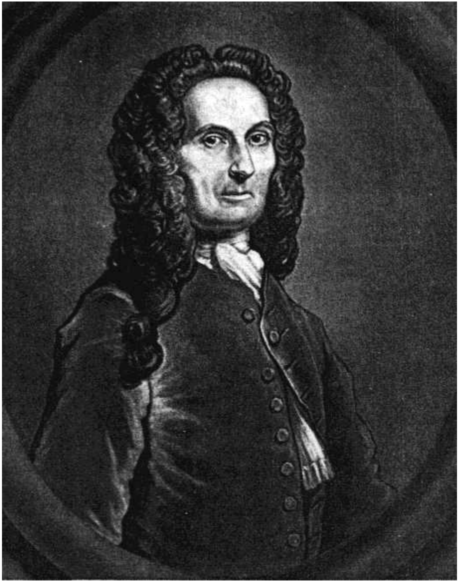


Source: `https://en.wikipedia.org/wiki/Normal_distribution`

central limit theorem

Carl Friedrich Gauss, 1777 – 1855

central limit theorem

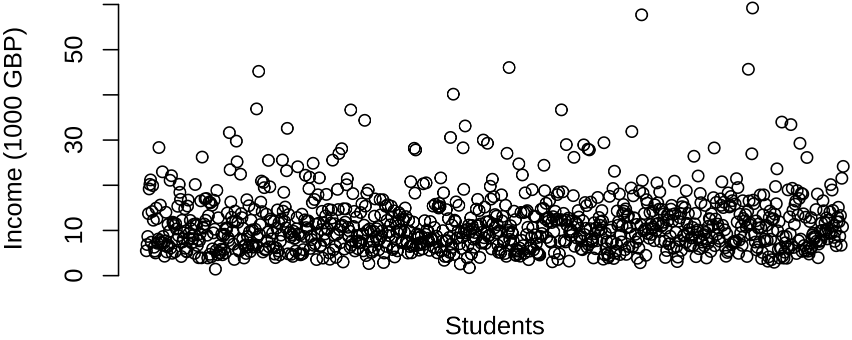


Abraham de Moivre, 1667 – 1754

sampling from the population

Thought experiment:

imagine all 2.3M students in the UK (statistical population)

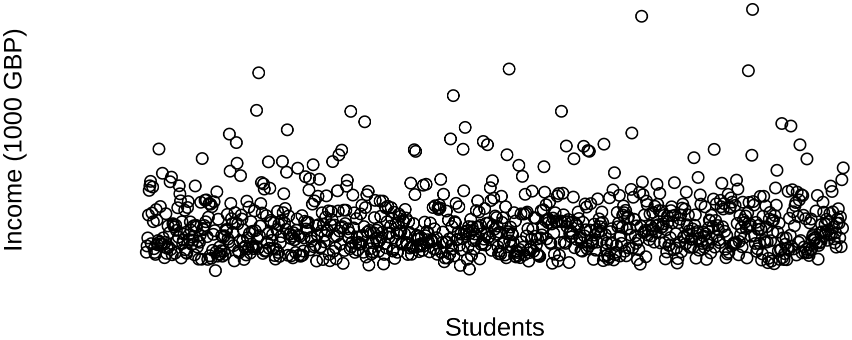


Population mean income=11.33

Population SD=6.00, (in 1000 GBP)

sampling from the population

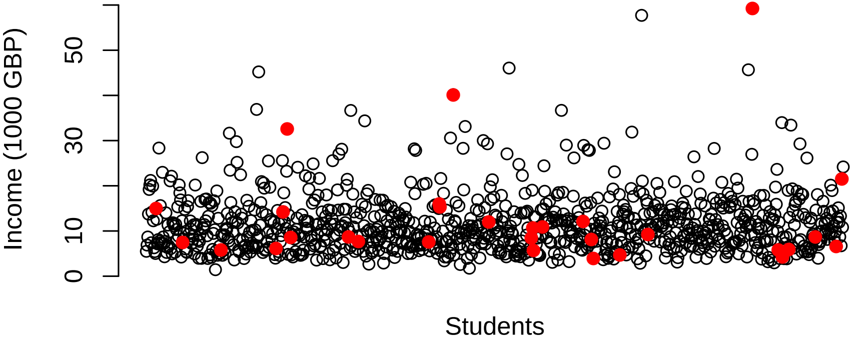
Thought experiment:
but what if we don't know their income?



Population mean income=$(? + ? + \ldots)/2.3\text{M} = ?$

sampling from the population

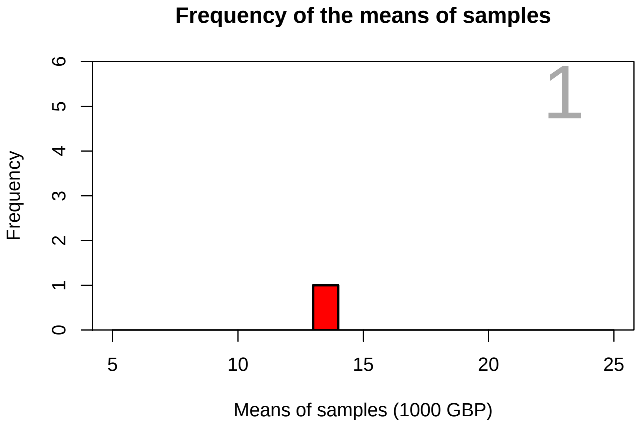
We draw a sample of size $N = 30$ students from this population (2.3M) and compute their mean income



Mean income=$(8 + 11 + 9 + 12 + 33 + 13 + 10 + 4 + \ldots + 5)/30 = \mathbf{13.6}$

sampling from the population

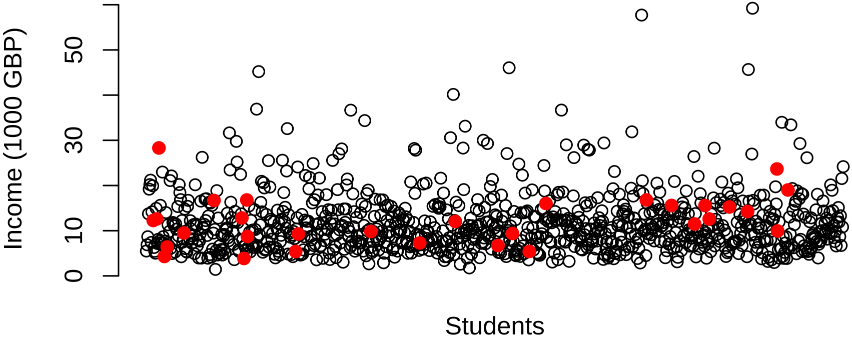
We draw a sample of size $N = 30$ students from this population (2.3M) and compute their mean income



Mean income=$(8 + 11 + 9 + 12 + 33 + 13 + 10 + 4 + \ldots + 5)/30 = \mathbf{13.6}$

sampling from the population

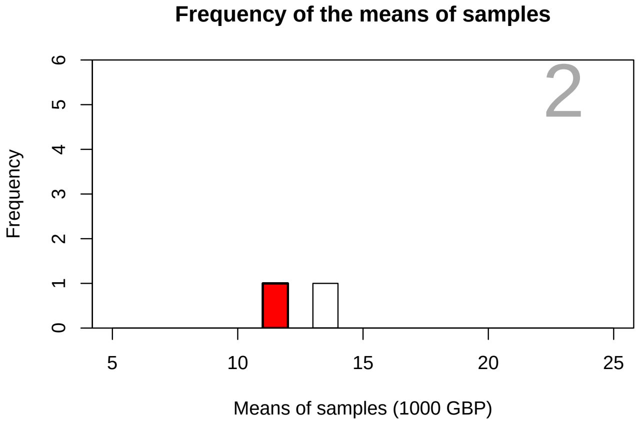
... and then we draw another sample of size $N = 30$ and compute its mean, and another... and another...



Mean income=$(4 + 22 + 16 + 3 + 4 + 7 + 5 + 15 + \ldots + 10)/30 = \mathbf{11.3}$

sampling from the population

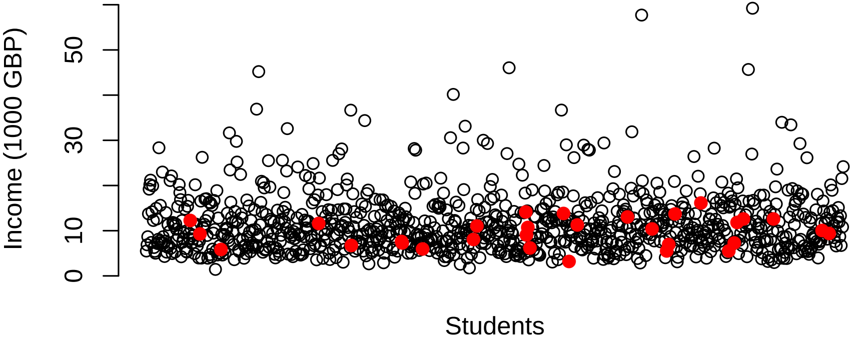
... and then we draw another sample of size $N = 30$ and compute its mean, and another... and another...



Mean income=$(4 + 22 + 16 + 3 + 4 + 7 + 5 + 15 + \ldots + 10)/30 = \mathbf{11.3}$

sampling from the population

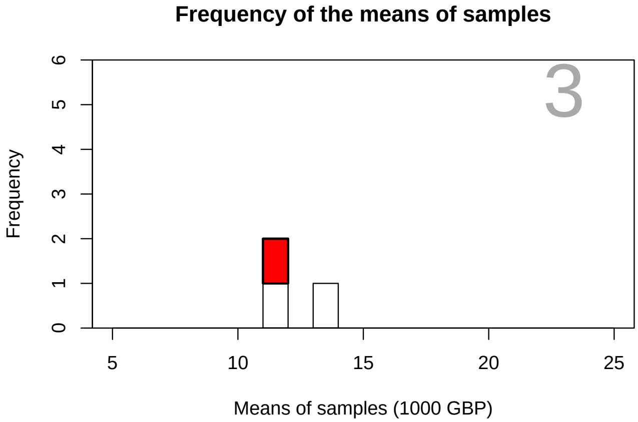
... and then we draw another sample of size $N = 30$ and compute its mean, and another... and another...



Mean income=$(6 + 8 + 5 + 41 + 16 + 7 + 8 + 8 + \ldots + 21)/30 = \mathbf{11.9}$

sampling from the population

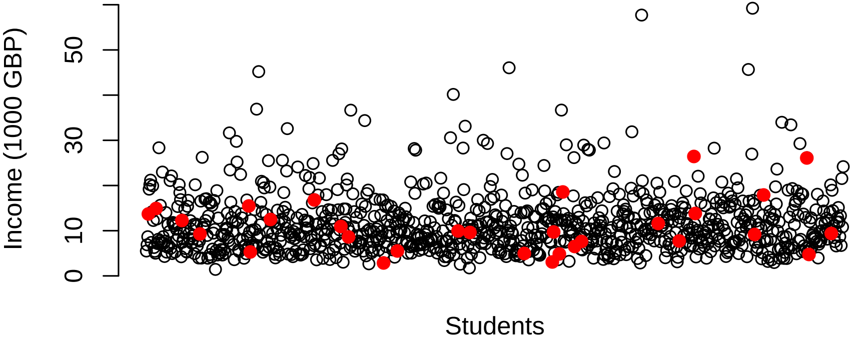
... and then we draw another sample of size $N = 30$ and compute its mean, and another... and another...



Mean income=$(6 + 8 + 5 + 41 + 16 + 7 + 8 + 8 + \ldots + 21)/30 = \mathbf{11.9}$

sampling from the population

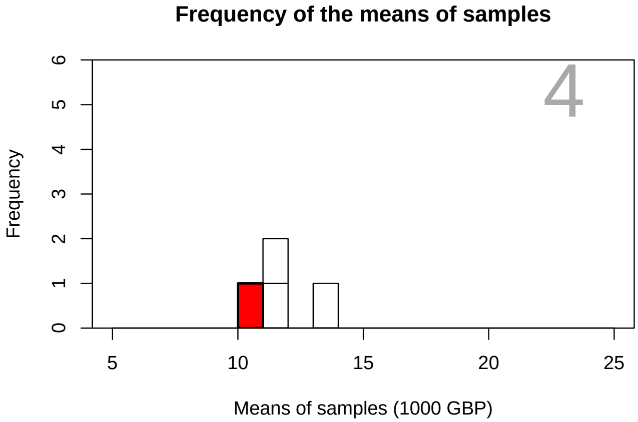
... and then we draw another sample of size $N = 30$ and compute its mean, and another... and another...



Mean income=$(13 + 9 + 16 + 12 + 6 + 22 + 22 + 26 + \ldots + 5)/30 = \mathbf{10.9}$

sampling from the population

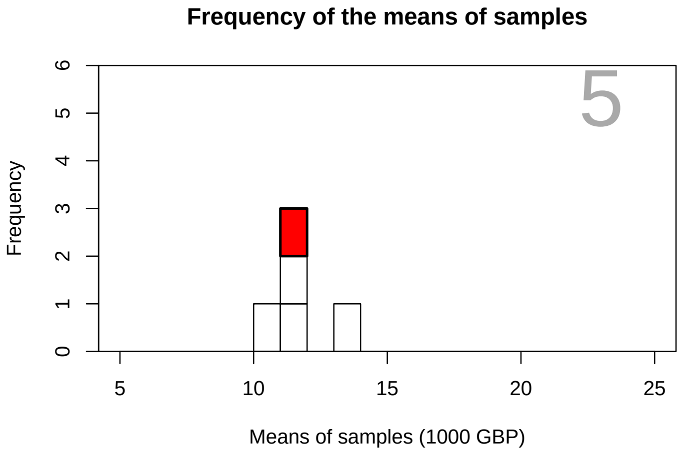
... and then we draw another sample of size $N = 30$ and compute its mean, and another... and another...



Mean income=$(13 + 9 + 16 + 12 + 6 + 22 + 22 + 26 + \ldots + 5)/30 = \mathbf{10.9}$

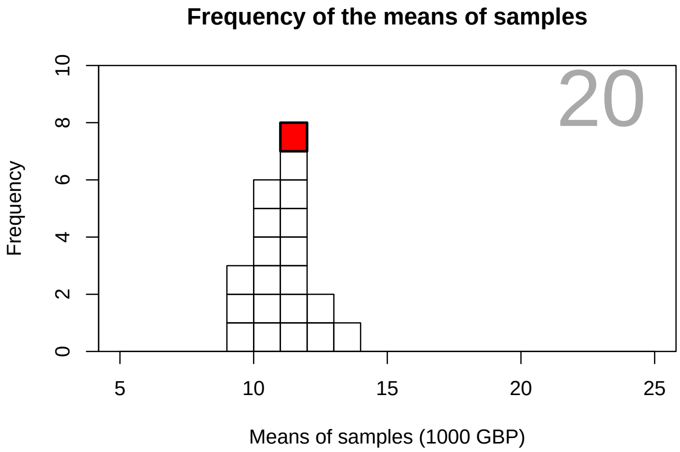
drawing samples – continuation

Sample size is $N = 30$ students



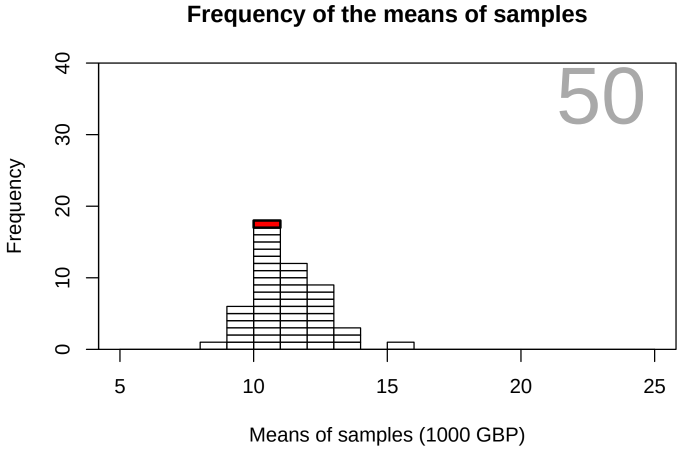
drawing samples – continuation

Sample size is $N = 30$ students



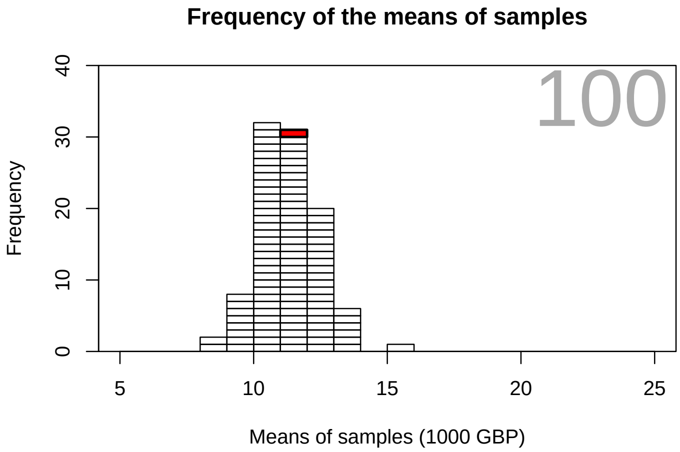
drawing samples – continuation

Sample size is $N = 30$ students



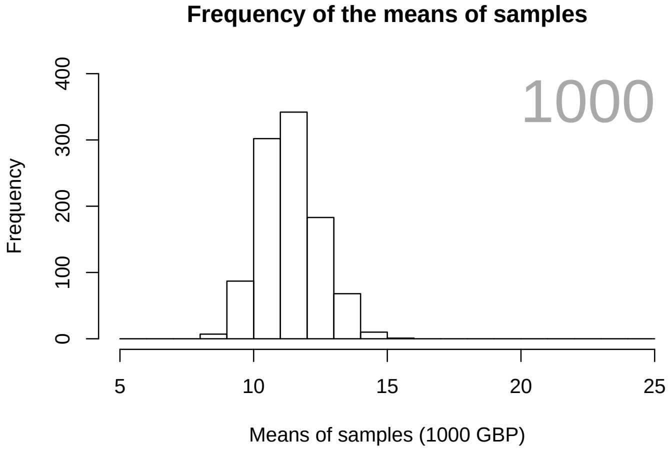
drawing samples – continuation

Sample size is $N = 30$ students



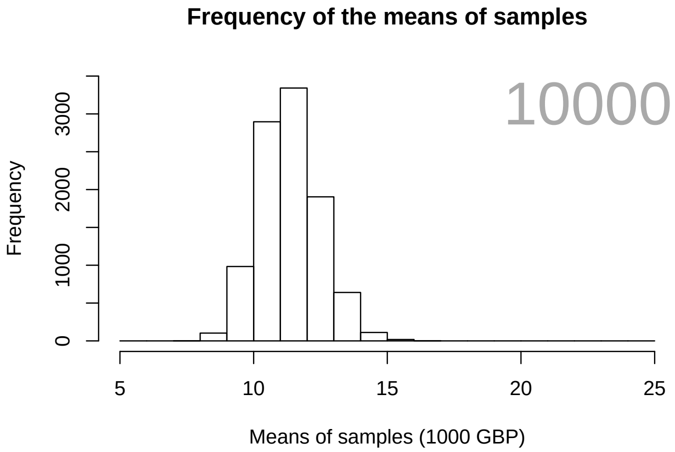
drawing samples – continuation

Sample size is $N = 30$ students



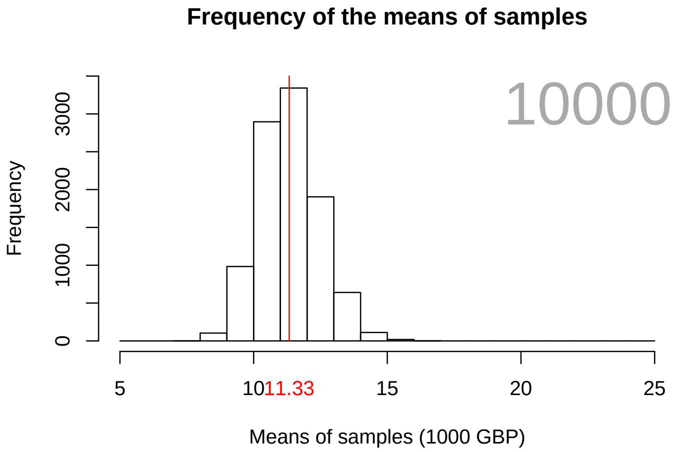
drawing samples – continuation

Sample size is $N = 30$ students



drawing samples – continuation

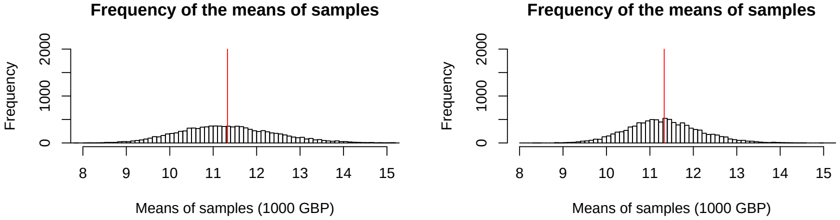
Sample size is $N = 30$ students



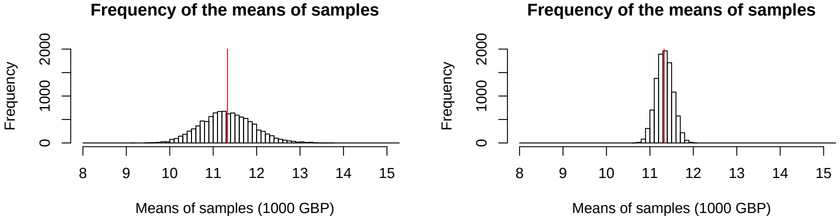
drawing samples 10 thousand times

Sample size:

$N = 30$ $\qquad$ $N = 50$



$N = 100$ $\qquad$ $N = 1000$



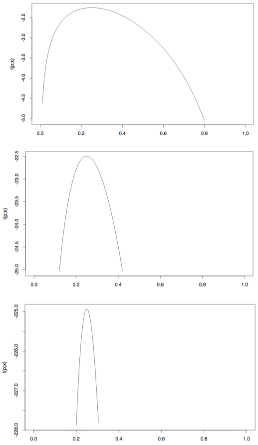
interval estimation

mle's are asymptotically normally distributed

assume (Bernoulli):

- we observe $X = 1$ from binomial($n = 4$, p)
- we observe $X = 10$ from binomial($n = 40$, p)
- we observe $X = 100$ from binomial($n = 400$, p)

log likelihood:



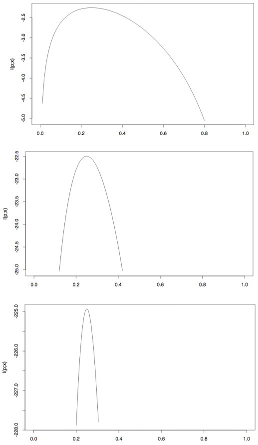
interval estimation

mle's are asymptotically normally distributed

assume (Bernoulli):

log likelihood:

- we observe $X = 1$ from binomial($n = 4$, p)
- we observe $X = 10$ from binomial($n = 40$, p)
- we observe $X = 100$ from binomial($n = 400$, p)



$$\hat{p}_{mle} = 0.25$$

interval estimation

mle's are asymptotically normally distributed

as n gets larger, we note the following

- log likelihood spikes around $\hat{p}_{mle}$
  - more confident that the true p lies close to $\hat{p}_{mle}$
- log likelihood becomes more symmetric around $\hat{p}_{mle}$
  - allows for constructing asymptotic confidence intervals for p

interval estimation

mle's are asymptotically normally distributed

as n gets larger, we note the following

- log likelihood spikes around $\hat{p}_{mle}$
  - more confident that the true p lies close to $\hat{p}_{mle}$
- log likelihood becomes more symmetric around $\hat{p}_{mle}$
  - allows for constructing asymptotic confidence intervals for p

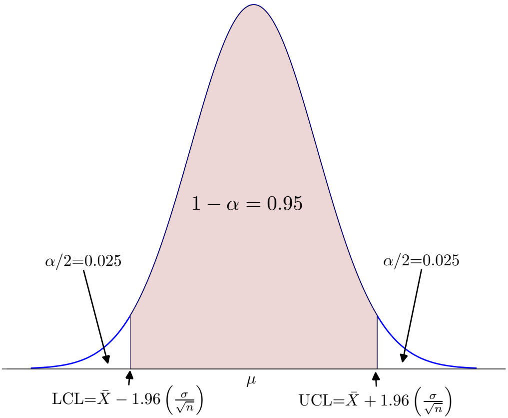
central limit theorem

- as $n \to \infty$, the log likelihood approaches a quadratic function (parabola) centered at the mle
- the parabola is the log likelihood for a normal distribution
- thus, we can form approximate confidence intervals for θ

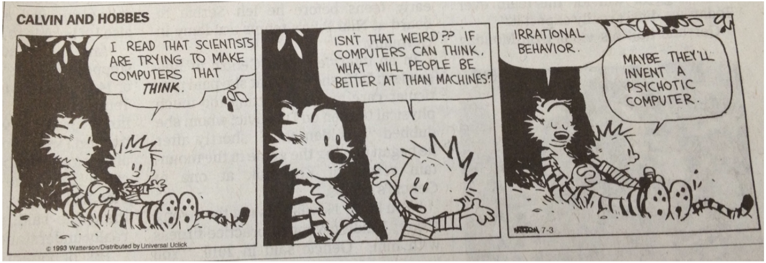
interval estimation

a sample from normal distribution with known variance σ^2
95% confidence interval for the mean μ is

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$



Calvin and Hobbes



reading

Agresti A., 2018, Statistical Methods for the Social Sciences, Fifth Edition, **Chapter 5**

link to the book via Manchester library