UCL

Model Assumptions and Truth in Statistics

Christian Hennig

4 February 2015

1. Statistics and Philosophy

Statistics is about getting knowledge from data.

What can we know from data?

How can we find out?

Issues

- What is probability? - frequentist vs. Bayesian vs. ?
  Observers see what happens but probability
  is about “what could have happened
  other than what happened”.

Issues

- What is probability? - frequentist vs. Bayesian vs. ?
  Observers see what happens but probability
  is about "what could have happened
  other than what happened".
- Assessing evidence - testing hypotheses,
  p-values, probabilities of hypotheses...

Issues

- What is probability? - frequentist vs. Bayesian vs. ?
  Observers see what happens but probability
  is about "what could have happened
  other than what happened".
- Assessing evidence - testing hypotheses,
  p-values, probabilities of hypotheses...

  often confused with interpretations of probability,
  but Bayesian methods can work with frequentist
  probabilities.

Issues

- What is probability? - frequentist vs. Bayesian vs. ?
  Observers see what happens but probability
  is about "what could have happened
  other than what happened".
- Assessing evidence - testing hypotheses,
  p-values, probabilities of hypotheses...

  often confused with interpretations of probability,
  but Bayesian methods can work with frequentist
  probabilities.
- Modelling and model assumptions -
  where does the model come from?
  How can assumptions be assessed?

Issues

- Statistics vs. Data Analysis - our own “PoS vs. STS”.
  Data Analysis is about much more than probability models and classical inference.
  Prediction, exploratory analysis, data compression, pattern recognition, data visualisation, stability,. . .

Issues

- Statistics vs. Data Analysis - our own "PoS vs. STS".
  Data Analysis is about much more than probability models and classical inference.
  Prediction, exploratory analysis, data compression, pattern recognition, data visualisation, stability,...

  Data Analysts: Only use statistical models where they help, often other methods are better.

  Statisticians: St. models improve pretty much everything. (And we have experimental design.)

Issues

- Statistics vs. Data Analysis - our own “PoS vs. STS”.
  Data Analysis is about much more than probability models and classical inference.
  Prediction, exploratory analysis, data compression, pattern recognition, data visualisation, stability,. . .

  Data Analysts: Only use statistical models where they help, often other methods are better.

  Statisticians: St. models improve pretty much everything. (And we have experimental design.)

  (Parts of the field claimed by Machine Learning now, role of “black box prediction machines”?)

Issues

- Objectivity vs. subjectivity - automatic/standardised decisions vs. researcher's decisions

Issues

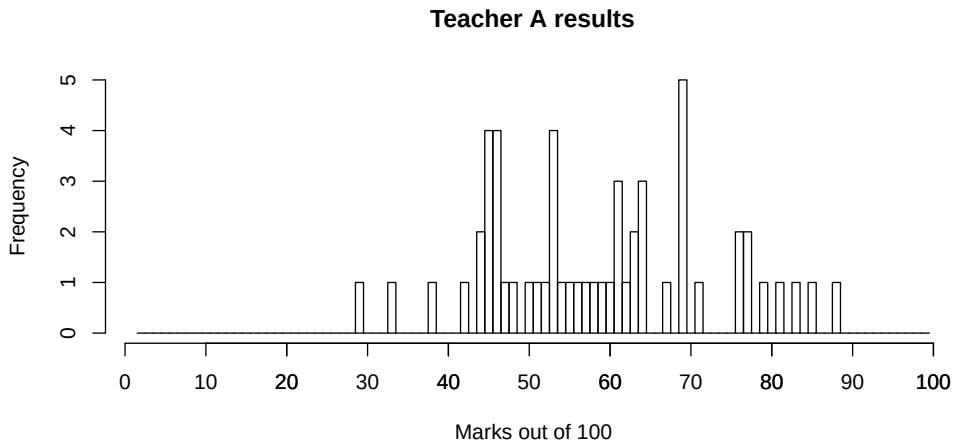
- Objectivity vs. subjectivity - automatic/standardised decisions vs. researcher's decisions
- Measuring quality of methodology - assumed truth (what is the truth we want to find)? prediction error? replication? interpretability?

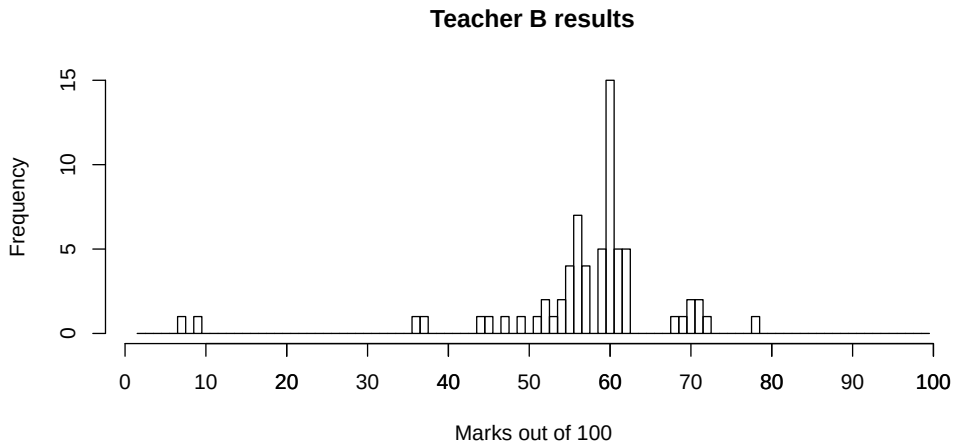
Issues

- Objectivity vs. subjectivity - automatic/standardised decisions vs. researcher's decisions
- Measuring quality of methodology - assumed truth (what is the truth we want to find)? prediction error? replication? interpretability?
- . . . and more - theory of measurement, role of visualisation, statistics communication. . .

Overview

2. Mathematical models and reality - some data





Standard model:

$$\begin{aligned} x_1, \ldots, x_n &\sim \mathcal{N}(\mu_1, \sigma_1^2) \text{ i.i.d.,} \\ y_1, \ldots, y_m &\sim \mathcal{N}(\mu_2, \sigma_2^2) \text{ i.i.d.} \end{aligned}$$

Estimate μ_1, μ_2 by means, $\bar{x}_1 = 58.6$, $\bar{x}_2 = 56.9$, teacher A students do better, but not significantly so ($p = 0.455$ under $\mu_1 = \mu_2$).

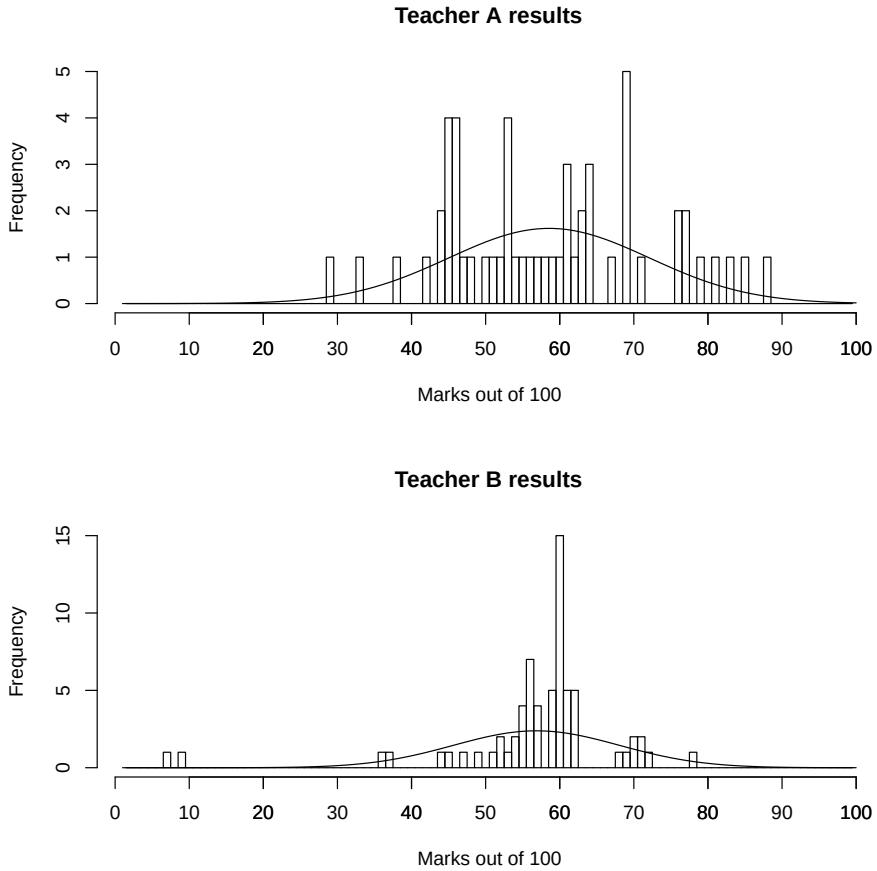
Standard model:

$$\begin{aligned} x_1, \ldots, x_n &\sim \mathcal{N}(\mu_1, \sigma_1^2) \text{ i.i.d.,} \\ y_1, \ldots, y_m &\sim \mathcal{N}(\mu_2, \sigma_2^2) \text{ i.i.d.} \end{aligned}$$

Estimate μ_1, μ_2 by means, $\bar{x}_1 = 58.6$, $\bar{x}_2 = 56.9$, teacher A students do better, but not significantly so ($p = 0.455$ under $\mu_1 = \mu_2$).

What do the model assumptions mean, and do they make sense?

With fitted distributions $\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)$



We actually *know* that the model assumptions are violated.
The normal distribution has an unlimited value range,
and produces integer values with probability zero.
Actually for such reasons we know that
no data observed by humans can ever be “truly normal”.

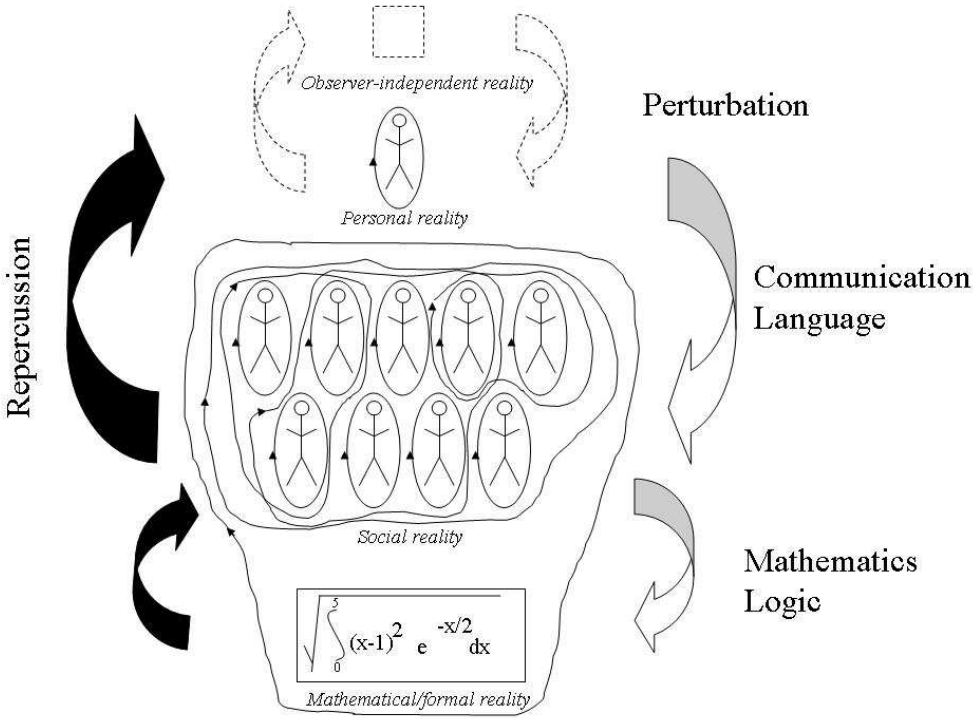
But how much harm does this do?

General mathematical modelling

Identification of items of perceived reality
with mathematical objects
and interpretation of results of mathematical operations
in terms of the items.

“Trivial” manipulation of counts and measurements,
deterministic and stochastic models,
analytical and computer models,
data fitting and mechanistic models.

Constructivist view (H., 2010, Foundations of Science)



Implications for mathematical modelling

- Science: establishing agreement in open exchange.

Implications for mathematical modelling

- Science: establishing agreement in open exchange.
- Mathematics is about creating a system that makes absolute agreement possible
  (but only *within mathematics*).

Implications for mathematical modelling

- Science: establishing agreement in open exchange.
- Mathematics is about creating a system that makes absolute agreement possible
  (but only *within mathematics*).
- Mathematical modelling is not about how things are, but about how we think and communicate about them.
  (Models communicate and change views of reality.)

Implications for mathematical modelling

- Science: establishing agreement in open exchange.
- Mathematics is about creating a system that makes absolute agreement possible
  (but only *within mathematics*).
- Mathematical modelling is not about how things are, but about how we think and communicate about them.
  (Models communicate and change views of reality.)
- It cannot be formally analysed how formal models are related to informal reality.

Implications for mathematical modelling

- Science: establishing agreement in open exchange.
- Mathematics is about creating a system that makes absolute agreement possible
  (but only *within mathematics*).
- Mathematical modelling is not about how things are, but about how we think and communicate about them.
  (Models communicate and change views of reality.)
- It cannot be formally analysed how formal models are related to informal reality.
- Pragmatist attitude: what do we get out of it?

Realism?

Models are about personal and social perceptions.
Science: can check models against
observations from agreed measurement procedures.

Realism?

Models are about personal and social perceptions.
Science: can check models against
observations from agreed measurement procedures.

Compatible with Chang's **Active Scientific Realism**:
"I take reality as whatever is not subject to one's will, and knowledge as an ability to act without being frustrated by resistance from reality."
We acknowledge reality's "resistance"
and that this is what science is about -
without accessing "truth"
about observer-independent reality.

Antony Gormley - Allotment

3. Frequentist probabilities (e.g., von Mises)

3.1 What do the model assumptions mean?

For example,

$$\begin{aligned} x_1, \ldots, x_n &\sim \mathcal{N}(\mu_1, \sigma_1^2) \text{ i.i.d.,} \\ y_1, \ldots, y_m &\sim \mathcal{N}(\mu_2, \sigma_2^2) \text{ i.i.d.} \end{aligned}$$

What do the model assumptions mean?

"We think of the situation as . . . "

- Potentially infinite repetition (of experimental conditions)
- $P(A)$: relative frequency limit of occurrence of A (e.g., normal distribution is defined by $P(A)$ $\forall A$.)

What do the model assumptions mean?

"We think of the situation as . . . "

- Potentially infinite repetition (of experimental conditions)
- $P(A)$: relative frequency limit of occurrence of A
  (e.g., normal distribution is defined by $P(A)\ \forall A$.)

This is obviously an idealisation -
and what constitutes a "repetition"?

"Whatever can be distinguished cannot be identical."
(B. de Finetti)

3.2 Independent repetitions ("i.i.d.")

Frequentism relies on "repetitions of experiments", e.g. results from different patients in clinical study, different students in the exam.

"$x_1, \ldots, x_n \sim \mathcal{N}(\mu_1, \sigma_1^2)$ i.i.d." defines a probability distribution for the full dataset, e.g., $P(x_1 \in A_1, x_2 \in A_2) = P(x_1 \in A_1)P(x_2 \in A_2)$.

3.2 Independent repetitions ("i.i.d.")

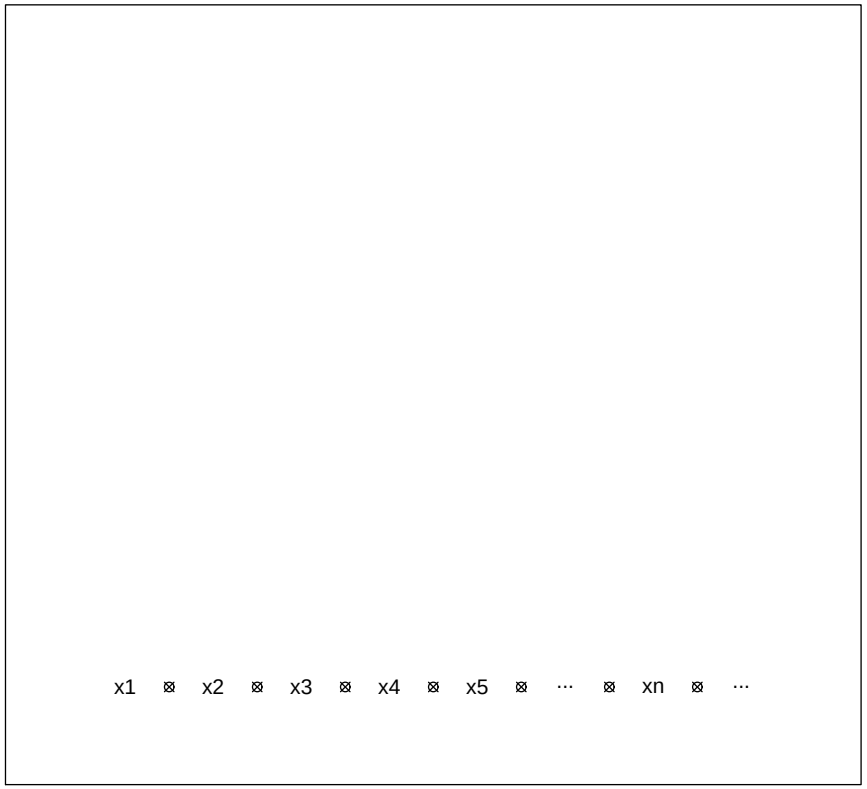
Frequentism relies on "repetitions of experiments", e.g. results from different patients in clinical study, different students in the exam.

"$x_1, \ldots, x_n \sim \mathcal{N}(\mu_1, \sigma_1^2)$ i.i.d." defines a probability distribution for the full dataset, e.g., $P(x_1 \in A_1, x_2 \in A_2) = P(x_1 \in A_1)P(x_2 \in A_2)$.

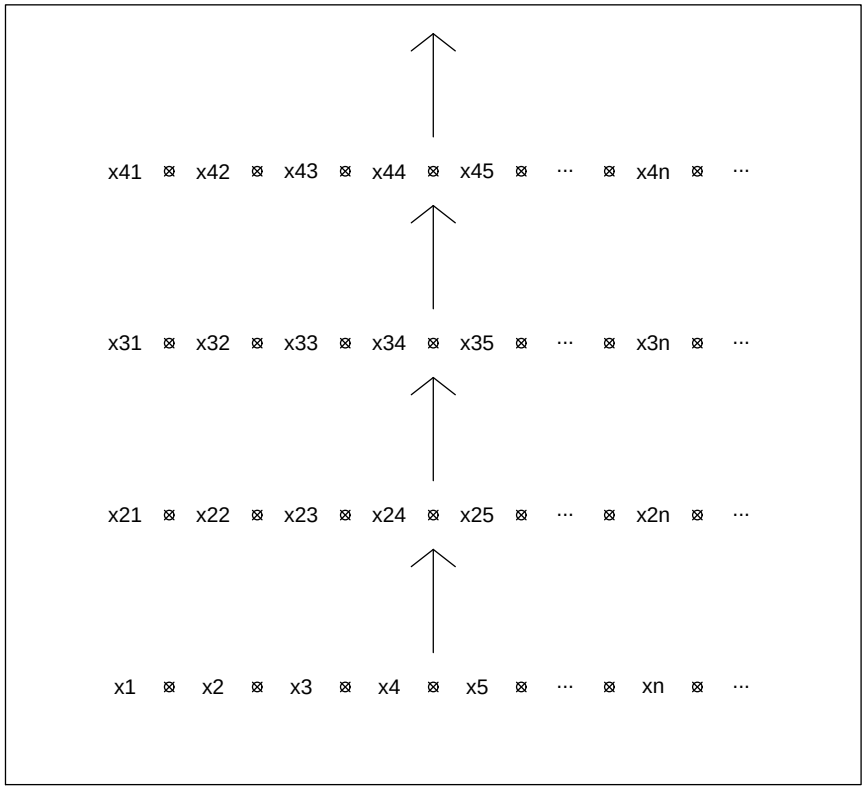
"i.i.d." is defined in terms of probabilities.

Probabilities cannot be defined in terms of i.i.d. repetitions.

In order to define “i.i.d.” sequences, i.i.d. repetitions and defining repetitions are required on different levels.



In order to define “i.i.d.” sequences, i.i.d. repetitions and defining repetitions are required on different levels.



In practice, there’s only one level of repetition.

The effective sample size for
(in-)dependence assumptions is usually 1.

But independent repetition *of some kind*
is always required in order to learn from data.

In practice, there’s only one level of repetition.

The effective sample size for
(in-)dependence assumptions is usually 1.

But independent repetition *of some kind*
is always required in order to learn from data.

Independent repetition is constructed by conscious decision
to ignore potential dependencies and differences.

3.3 Can frequentist model assumptions be checked?

Goodness-of-fit/misspecification tests (Mayo & Spanos)
If something modelled as very unlikely happens,
the model is interpreted to be *falsified*.

3.3 Can frequentist model assumptions be checked?

Goodness-of-fit/misspecification tests (Mayo & Spanos)
If something modelled as very unlikely happens,
the model is interpreted to be *falsified*.

For example, could “falsify” independence
from data where students at same table
tend to have similar results.

3.3 Can frequentist model assumptions be checked?

Goodness-of-fit/misspecification tests (Mayo & Spanos)
If something modelled as very unlikely happens,
the model is interpreted to be *falsified*.

For example, could “falsify” independence
from data where students at same table
tend to have similar results.

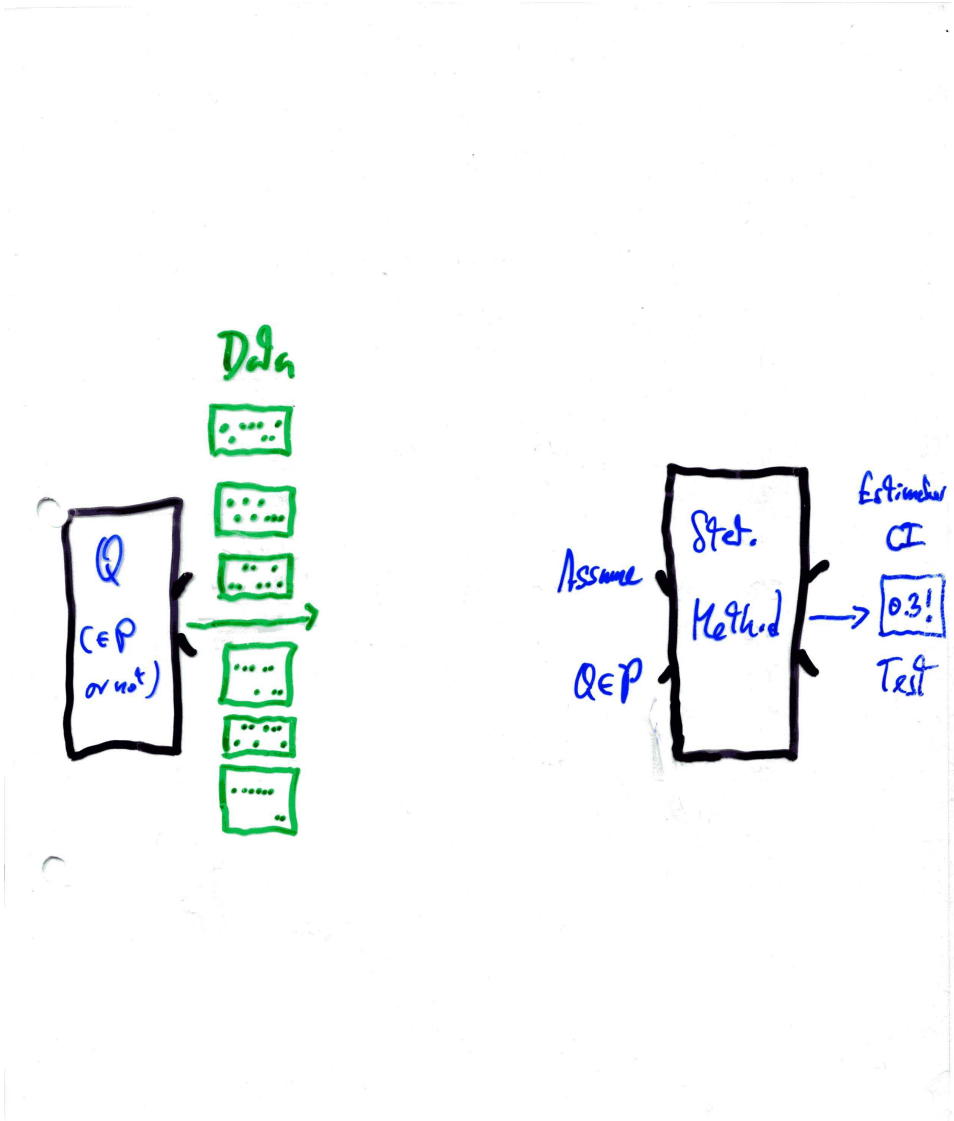
Implicitly assumes *tables* to be i.i.d.

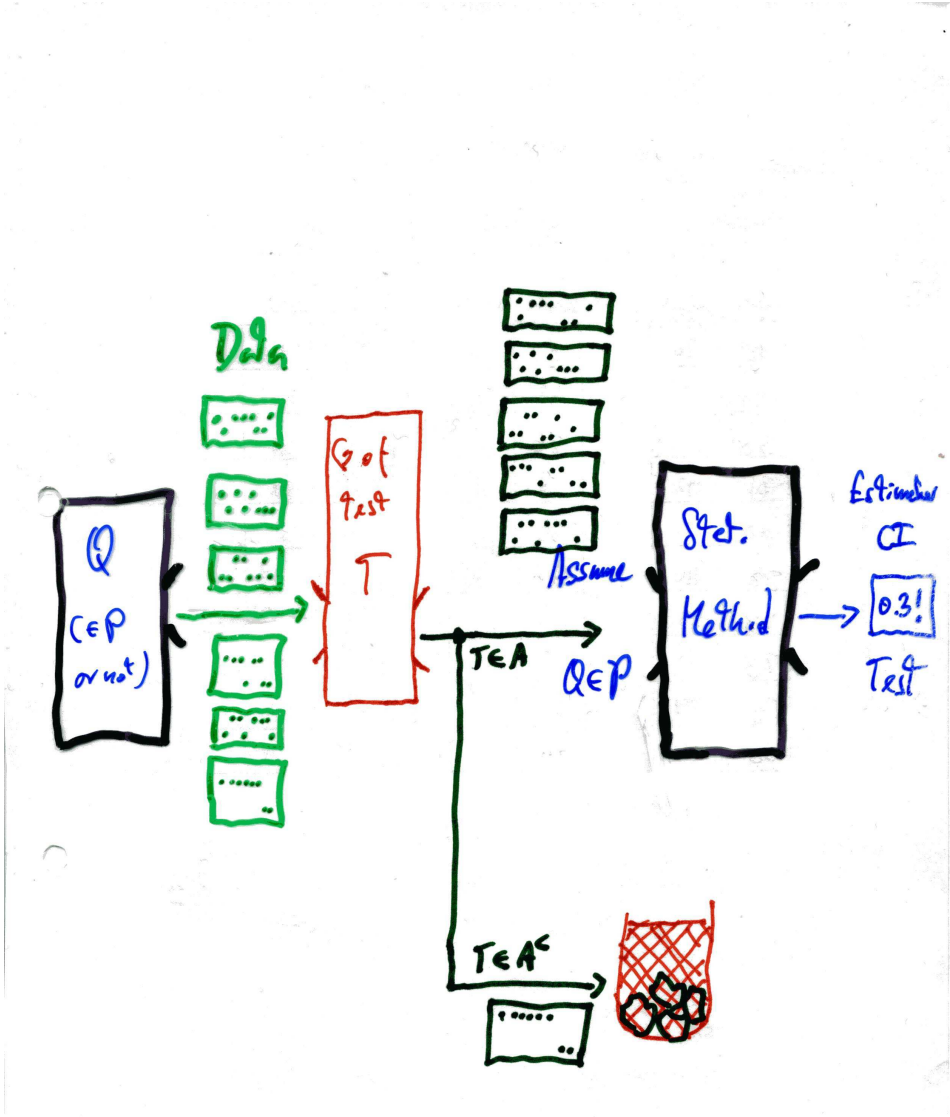
Dependence (between what’s “close”) can only be found by contrast to independence between what’s further away.

If everything’s dependent on everything else in same manner (or in irregularly different manners), this cannot be found.

3.4 The goodness-of-fit paradox (H, 2007)

Checking the model assumptions violates them automatically
because the *possibility* of unlikely events
is constitutive part of the models.


Data
Q
(∈P or not)
Assume
Q∈P
Stat. Method
Estimator
CI
0.3!
Test


Data
Q
(∈P or not)
Got test
T
Assume
Stat. Method
Estimator
CI
0.3!
Test
T∈A
Q∈P
T∈A^c

3.5 “Approximately true?”

Model assumptions are violated -
is it good enough if they are “approximately true”?

"The model assumptions hold approximately" - precise meaning could only be:
"We assume $Q \in \mathcal{P}$ and there is a true P so that $\min_{Q \in \mathcal{P}} d(P, Q)$ is small."

"The model assumptions hold approximately" -
precise meaning could only be:
"We assume $Q \in \mathcal{P}$ and there is a true P so that $\min_{Q \in \mathcal{P}} d(P, Q)$ is small."

Dissimilarity measure d is required.

Distances for data analysis (L. Davies):
$d(P, Q)$ small $\Leftrightarrow$ difficult to distinguish data from P, Q.

Problem: In every small d-neighbourhood of Q there are distributions that behave very differently.

Problem: In every small d-neighbourhood of Q there are distributions that behave very differently.

$Q = \mathcal{N}(\mu, \sigma^2)$ has expectation μ.
Expectation of P may be anywhere or not exist.
Likelihood ratio of Q and P may take any value.

If $P \neq \mathcal{N}(\mu, \sigma^2)$, what is estimated estimating μ?

Median, expectation, mode are identical for $\mathcal{N}(\mu, \sigma^2)$, but different for asymmetric P.

If $P \neq \mathcal{N}(\mu, \sigma^2)$, what is estimated estimating μ?

Median, expectation, mode are identical for $\mathcal{N}(\mu, \sigma^2)$, but different for asymmetric P.

Robustness theory: what characteristics of distributions don't change much in d-neighbourhoods?

If $P \neq \mathcal{N}(\mu, \sigma^2)$, what is estimated estimating μ?

Median, expectation, mode are identical for $\mathcal{N}(\mu, \sigma^2)$,
but different for asymmetric P.

Robustness theory: what characteristics of distributions don't change much in d-neighbourhoods?

In any case, cannot rule out too irregular distributions (such as complex dependence patterns).
Some assumptions need to be imposed without checking.

Should at least reflect our perception.

3.6 “Frequentism-as-model”, a new perspective

(Frequentist) models are still useful to...

- communicate researcher’s perception of situation,
- inspire methodology,
- check quality of methodology
  in situations with known (made up) truth.
  (Tukey, L. Davies)
- give an idea of potential variation to be expected.

Re-formulation of "model checking":
Find out whether data could lead
statistical method (derived from model) astray
(requires statistical theory/simulations).

Re-formulation of "model checking":
Find out whether data could lead
statistical method (derived from model) astray
(requires statistical theory/simulations).

Some violations are *not* problematic
(e.g., g.o.f.-testing is rarely problematic,
discreteness of data assumed normal doesn't hurt much).

Re-formulation of "model checking":
Find out whether data could lead
statistical method (derived from model) astray
(requires statistical theory/simulations).

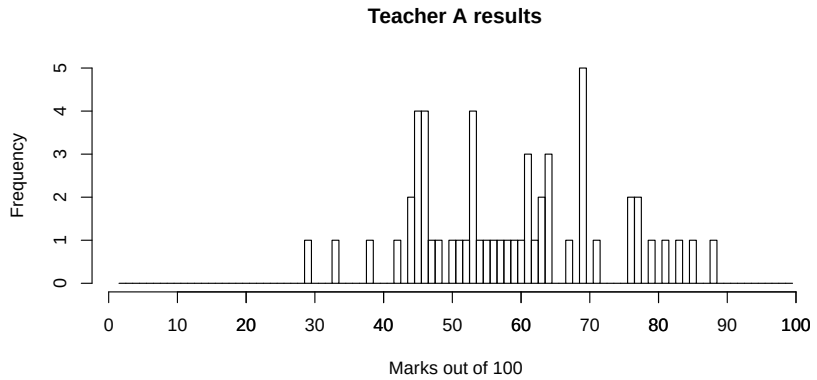
Some violations are *not* problematic
(e.g., g.o.f.-testing is rarely problematic,
discreteness of data assumed normal doesn't hurt much).

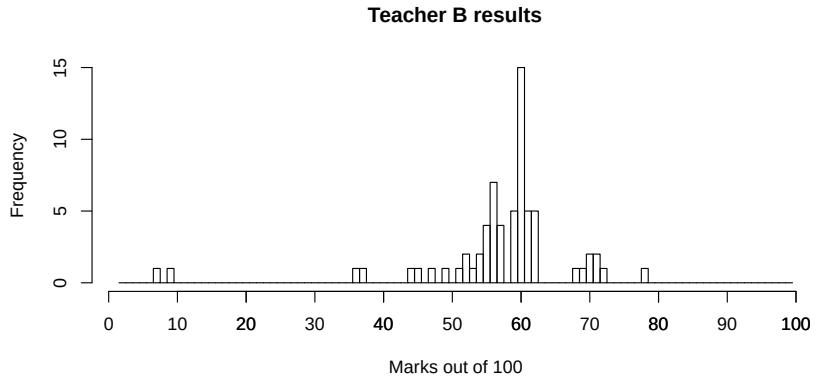
Depends on aim of analysis (researcher's construct)
which features of model are important.

Major motivation of methodology should come from
desired interpretation of results.

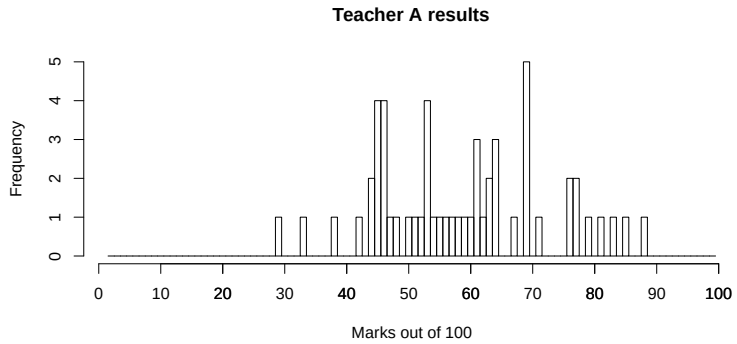
Students from same year: clearly not independent
(but may not show dependence pattern),
"identical" only by ignoring background
(*iid is constructed by selective ignorance*).

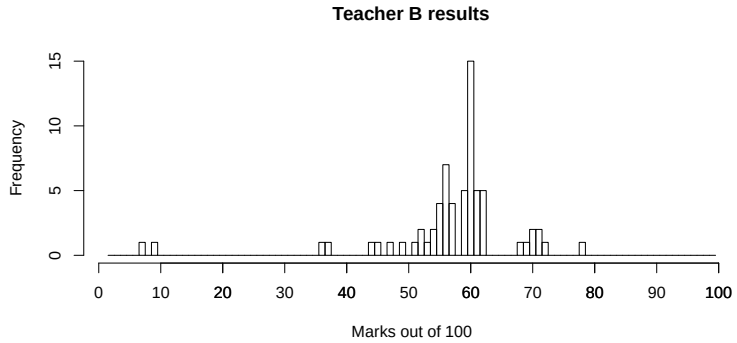
(That's the view the model carries.)



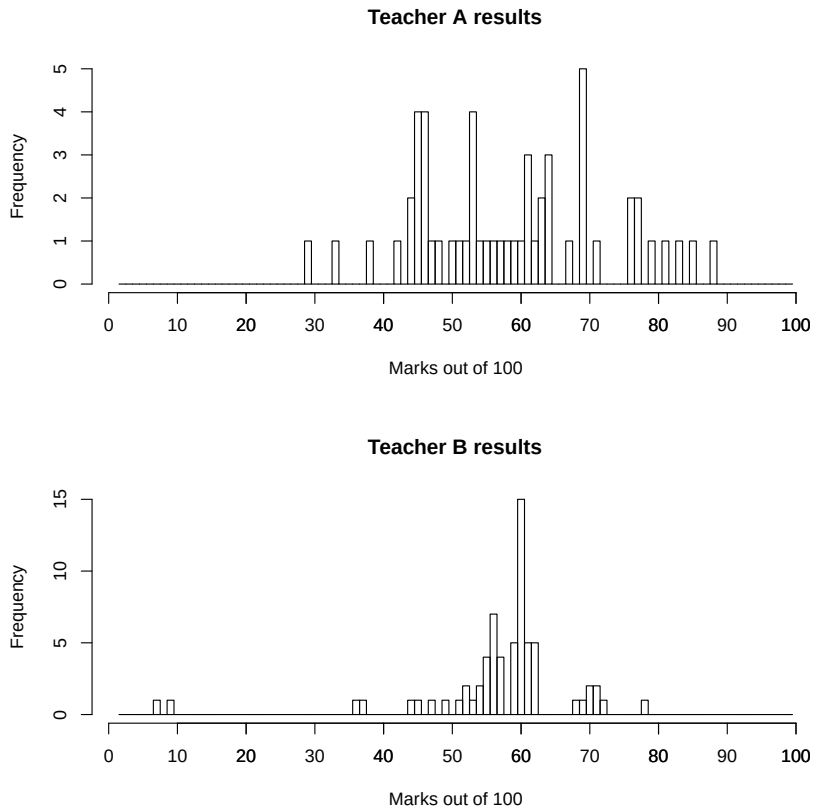


Means: 58.6 (A), 56.9 (B)
Medians: 58 (A), 59 (B).

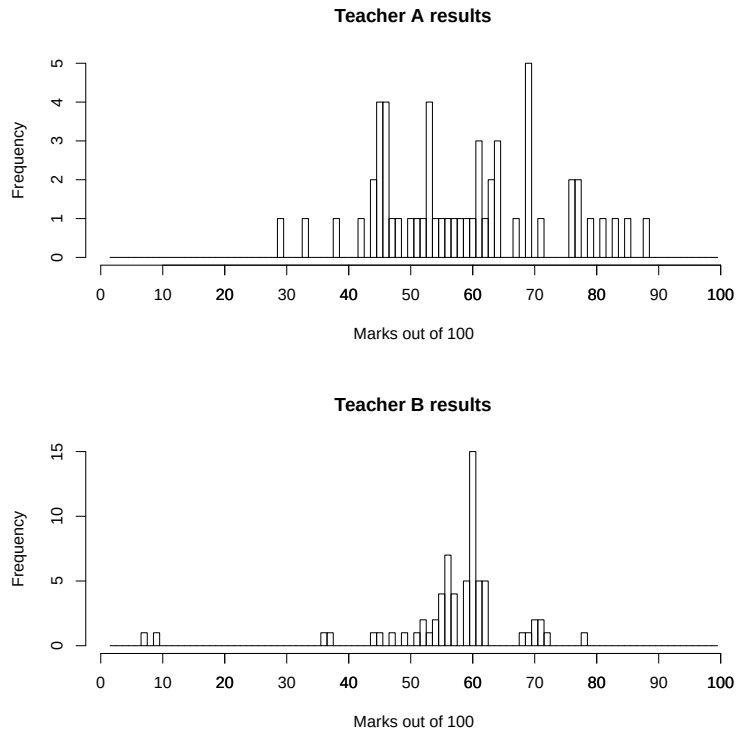




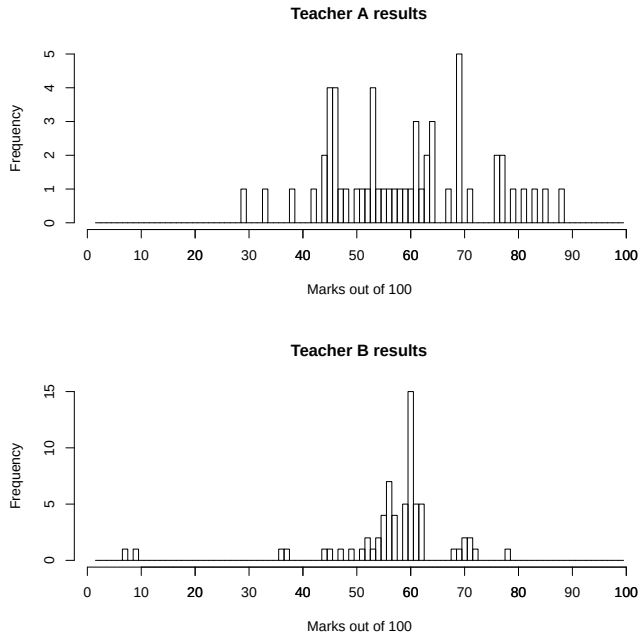
Can reject
equality of distributions (Kolmogorov-Smirnov $p = 0.012$),
neither means nor medians differ significantly.



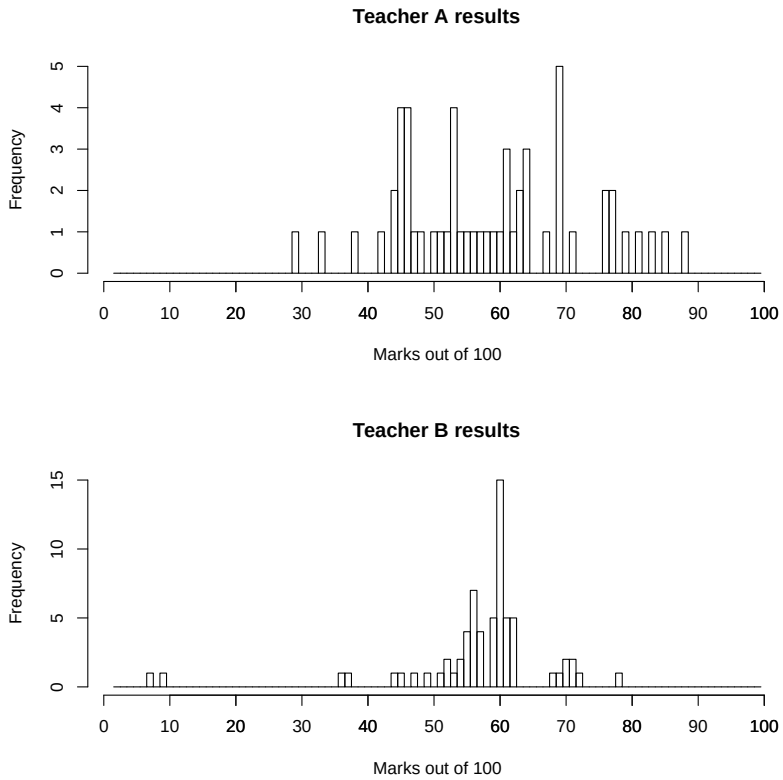
Lower outliers in B suggest median.



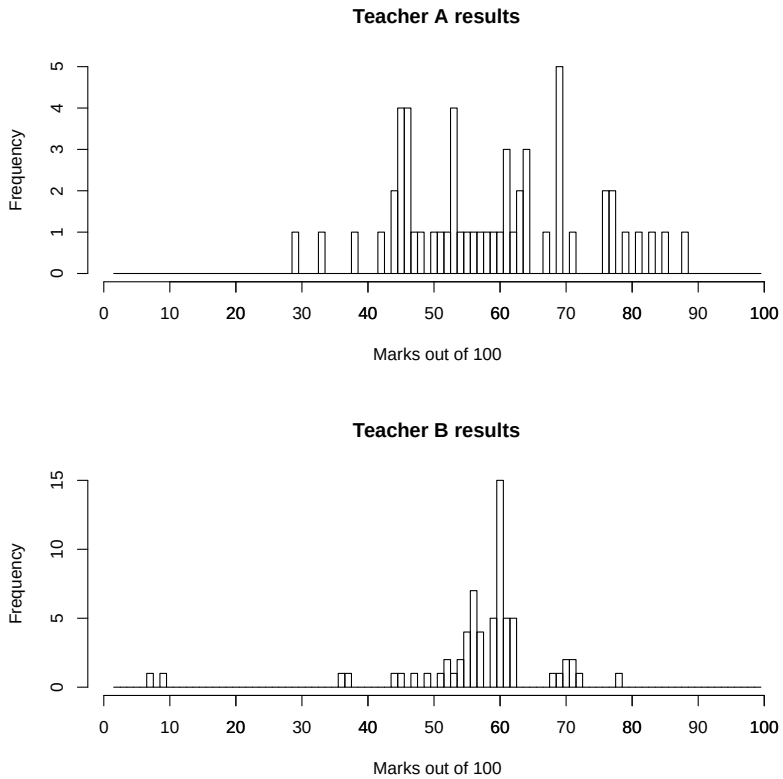
Favour mean: observations are not erroneous, so *all* observations should contribute!?



But it may not be seen as very relevant
how bad the clearly failed students exactly are.
(Would be different for upper outliers.)



Could look at pass ($\geq$ 50) rate (B clearly better) and other meaningful statistics (rank sum).



"Which aspect is of interest?"/meaning of data dominates "Which model is close to the truth."

Data analytic approach:
evaluated several characteristics,
got more comprehensive picture.

Data analytic approach:
evaluated several characteristics,
got more comprehensive picture.

Warning:
I ran several significance tests,
but this compromises test error probabilities.

Don't necessarily expect that
"significant different pass rate" will reproduce.

"Exploratory" interpretation of tests.

May check on independent data.

***p*-values**: highly controversial!
I use them routinely in exploratory manner
for checking whether what I spot
can be explained by random variation alone.

***p*-values**: highly controversial!
I use them routinely in exploratory manner
for checking whether what I spot
can be explained by random variation alone.

Relying on *p*-values as dominating tool
to “secure” scientific findings is dangerous.

It’s not the idea that is at fault
but people want too strong results too easily.

4. The Bayes-frequentist controversy

Long-standing controversy about foundations of statistics: Should a different probability interpretation be applied?

Can problems with frequentism be resolved by (subjective or objective) **Bayesian** approaches?

Bayes's theorem:

$$P(H_1|\text{data}) = \frac{P(\text{data}|H_1)P(H_1)}{\int_{H\in\mathcal{H}} P(\text{data}|H)P(H)}$$

Needed: sampling distribution $P(\text{data}|H)$,
prior $P(H)$ on set of sampling models $\mathcal{H}$.

4.1 Bayesian interpretations of probability

- Subjective Bayes (de Finetti) Probabilities measure individual's strength of belief in future outcomes, formalised as betting rates (epistemic probabilities).
- Objective Bayes (Jaynes) Similar, but believe that objectively rational strengths of belief can be found.
- Hidden frequentist/ "falsificationist" (Gelman)
  Use Bayesian methodology
  with sampling distribution interpreted frequentist.

Only the latter can check models against data.

The big question: Where does the prior come from?

Objective Bayesians want either non-informative or strong “objective” justification from background - rarely available.

The big question: Where does the prior come from?

Objective Bayesians want either non-informative or strong “objective” justification from background - rarely available.

“Falsificationist”: could use
frequentist idea even for prior (“all similar studies”);
could make sense for course mark data
but needs strong backing with past data.

The big question: Where does the prior come from?

Objective Bayesians want either non-informative or strong “objective” justification from background - rarely available.

“Falsificationist”: could use
frequentist idea even for prior (“all similar studies”);
could make sense for course mark data
but needs strong backing with past data.

Subjective impact cannot be suppressed.

4.2 Exchangeability

If you observe $x_1, \ldots, x_n$, future probabilities don't depend on order of observations.

It's essential for most Bayesian data analysis (at some level) - constructs "Bayesian repetition".

De Finetti's theorem:
Exchangeability $\Leftrightarrow P(\text{data}) = \int_{H \in \mathcal{H}} P(\text{data}|H)P(H)$ with i.i.d. model $P(\text{data}|H)$.

Exchangeability constructs “Bayesian repetition”; similar role as i.i.d. for frequentists.

Exchangeability constructs “Bayesian repetition”;
similar role as i.i.d. for frequentists.

Exchangeability implies that
$P\{1\}$ in the next go doesn’t depend on
whether you observe
0,0,1,0,1,1,1,0,0,1,0,1,1,0,1,1,0,0,1,0,1,0,0,1 or

0,0,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0.

Seems counterintuitive as epistemic assumption,
rather conscious decision to ignore deviations (as iid).

From my point of view,
the major philosophical problems
with Bayesian statistics
are about the same as with frequentism.

From my point of view,
the major philosophical problems
with Bayesian statistics
are about the same as with frequentism.

General problems of mathematical modelling,
“creation” of repetition by i.i.d./exchangeability.

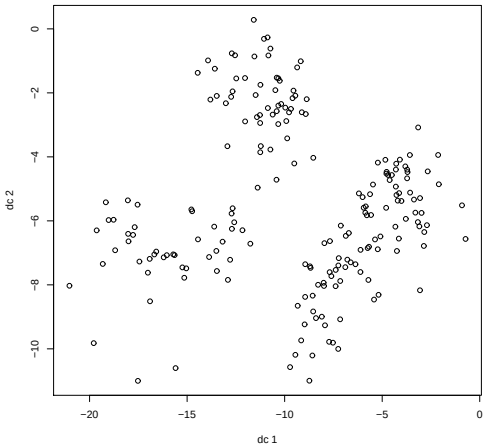
From my point of view,
the major philosophical problems
with Bayesian statistics
are about the same as with frequentism.

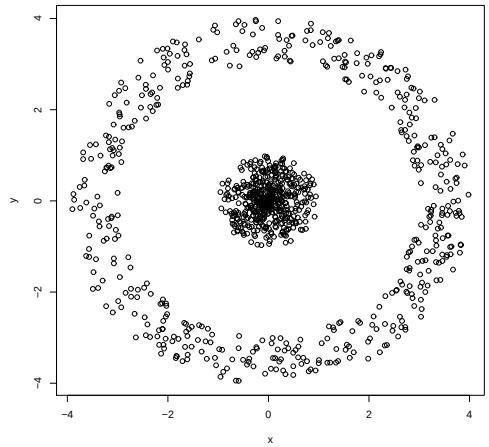
General problems of mathematical modelling,
“creation” of repetition by i.i.d./exchangeability.

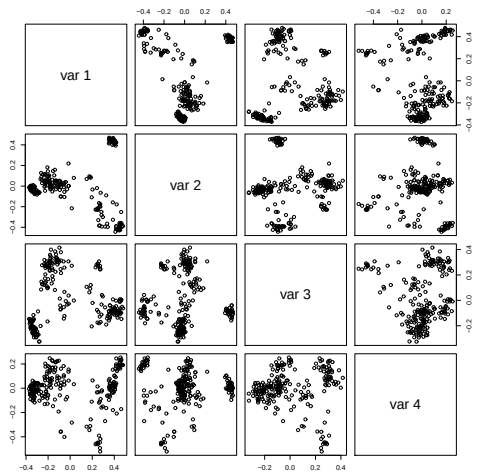
Bayes & frequentism
formalise different concepts of probability
(data generating mechanism vs. epistemic)
that can make sense in different situations.

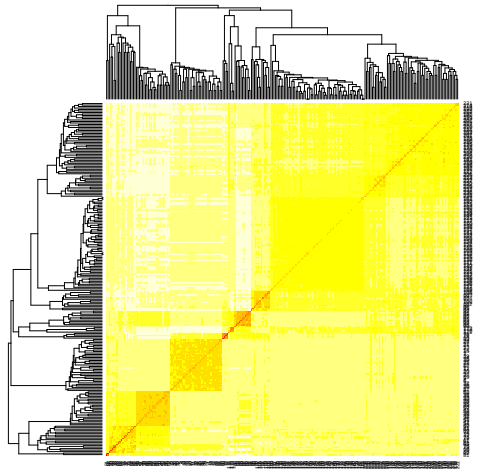
5. Cluster analysis and truth

Cluster analysis: methods to group data ("unsupervised classification").


dc 2
dc 1
0
−2
−4
−6
−8
−10
−20
−15
−10
−5
0


y
x
4
2
0
−2
−4
−4
−2
0
2
4


var 1
var 2
var 3
var 4

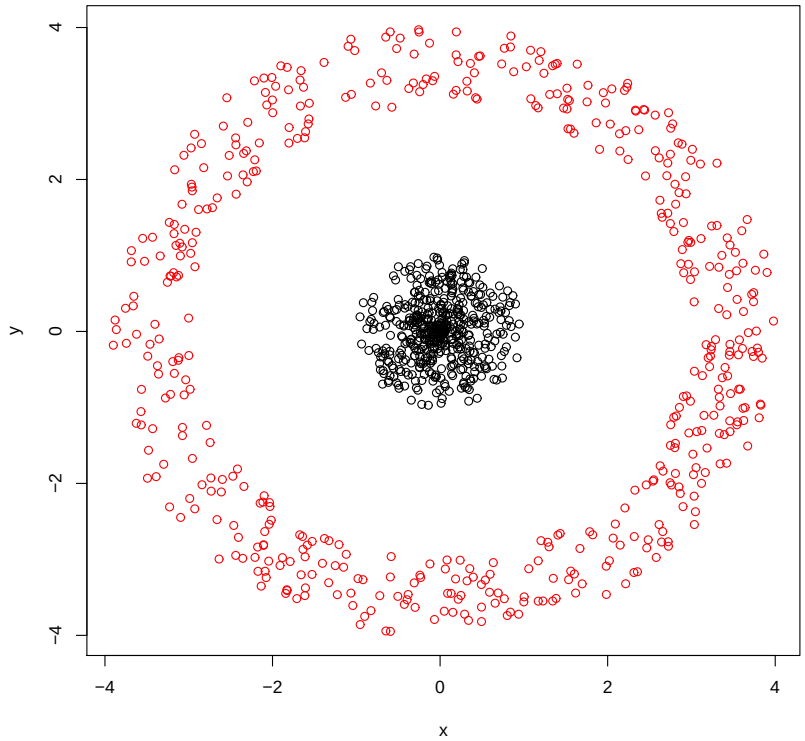


What are the “true” clusters?

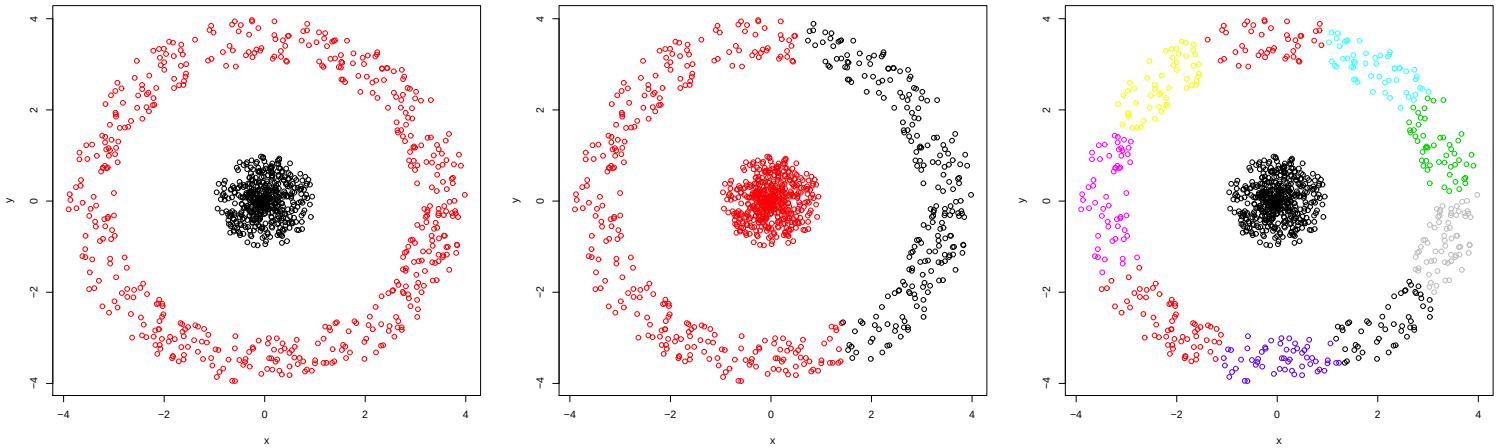
Need to connect substantial requirements to characteristics of cluster analysis methods.

Need a formal definition to measure method quality.

Intuitive clusters? (Often in literature)

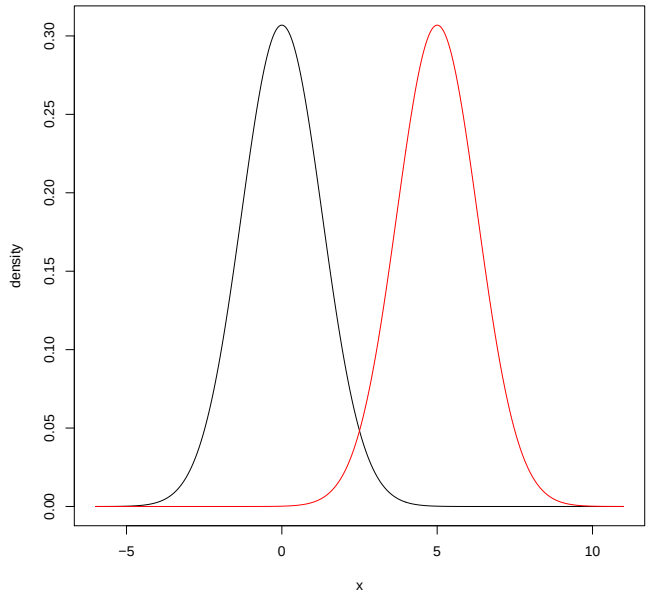


Intuitive clusters?

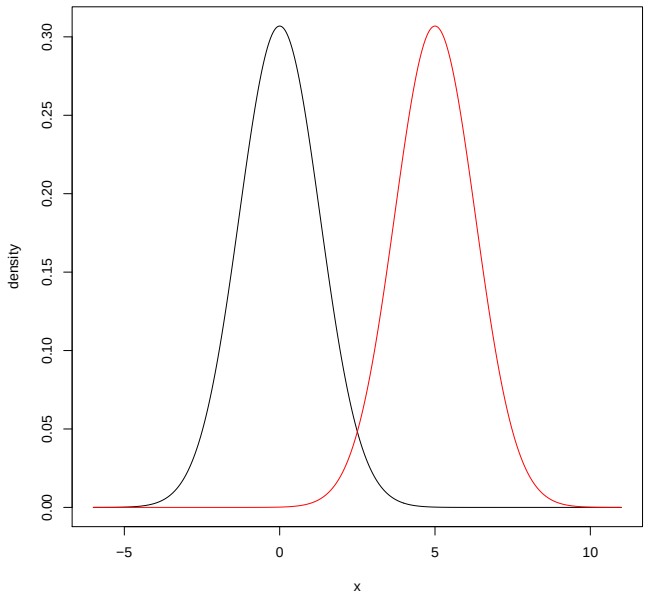


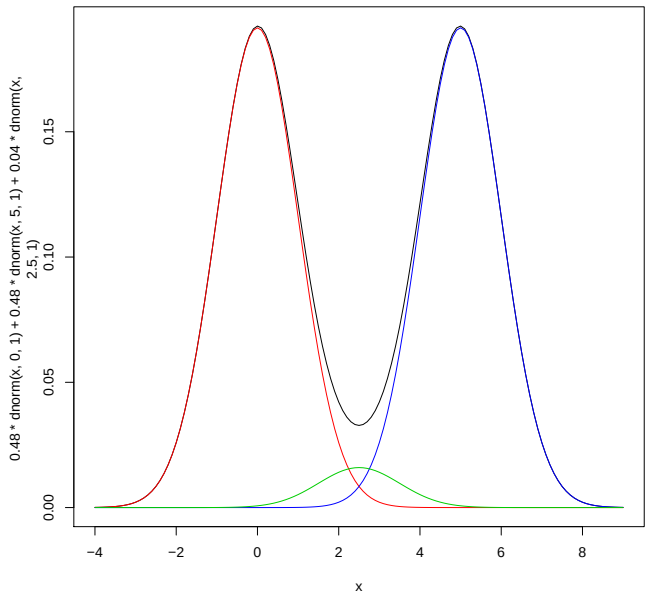
. . . but may be inappropriate, e.g.,
if small within-cluster distances required.

(Normal) mixture components in mixture distribution?



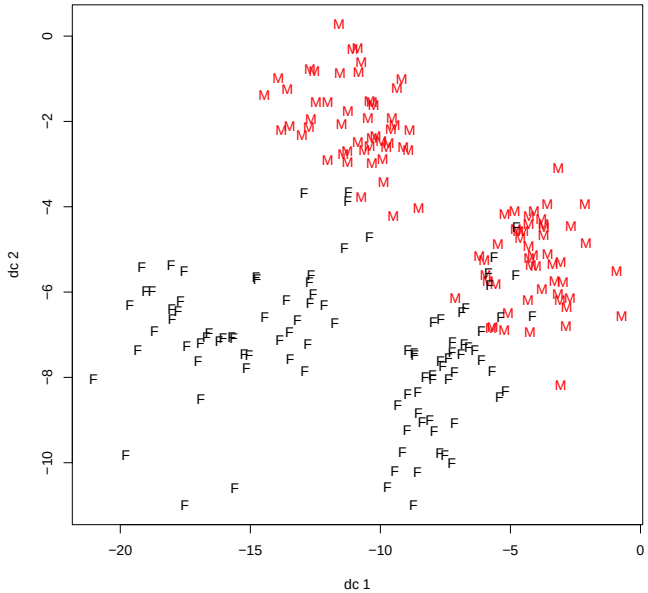
(Normal) mixture components in mixture distribution?





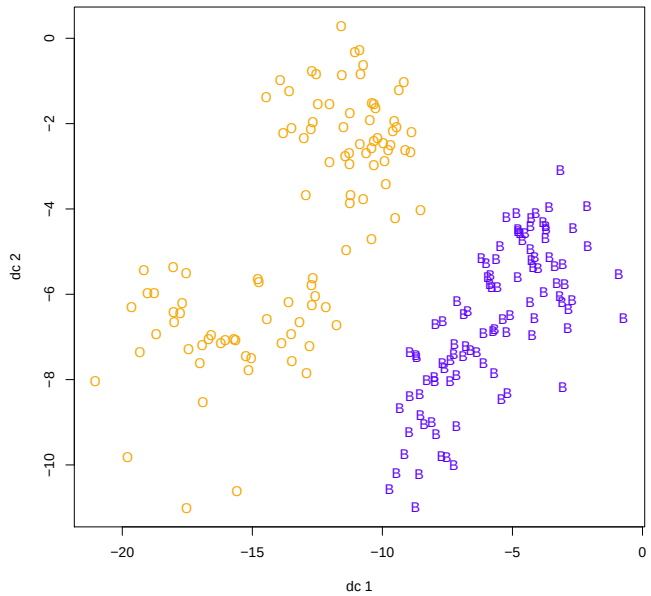
. . . but may not “cluster”.

Sometimes we know “true” clusters, don’t we?



(These look quite “unclustery”.)

Actually we know some more:



In reality there may be many valid classifications and only one may be known to us.

In literature, methods are advertised as finding the “natural” clusters, assumed unique.

In literature, methods are advertised as finding the “natural” clusters, assumed unique.

Literature is not open about researcher’s need to *decide* what kind of clusters are required in given application.
Suggests data can make all the decisions.

Researcher needs to “construct” cluster concept:
by what kind of principle should clusters be separated?

(Small within-cluster distances, separation,
probability mixture components, centroids etc.)

E.g., need to specify how the idea of a “species”
is connected to genetic or phenotype measurements.

6. Conclusion

- Foundations of statistics are quite shaky
  for those who hope to discover “objective truth”.
- Acknowledge what has to be decided subjectively
  (hopefully well informed).
- Acknowledge basic problems and limits
  of mathematical modelling.
- Acknowledge need for assumptions
  that cannot be checked.
- Motivate chosen method and (model) checks
  from desired interpretation.