Statistical Models, Estimation and Reality

Christian Hennig

February 13, 2012

1. Distributions: generators of observations

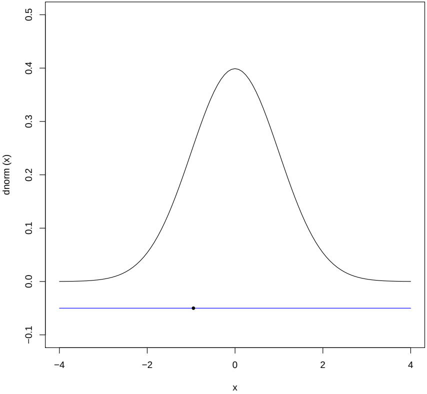
Statistical modelling is based on probability distributions.

Distributions are mathematical models,
and therefore artificial, abstract constructs.

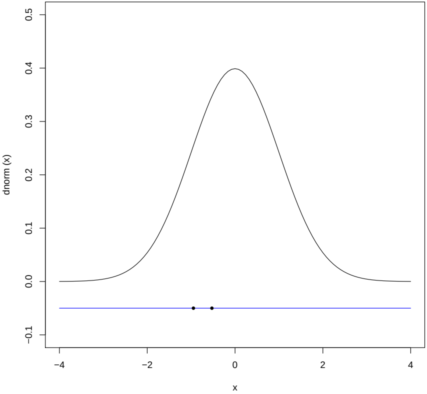
A distribution can be understood as
generator of observations.

If distribution tends to bring forth observations
that look (*in some sense*) like those
encountered in a real situation,
people may use it as a *model* for this situation.

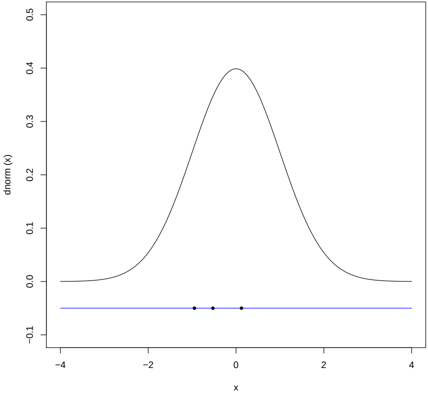
The normal distribution



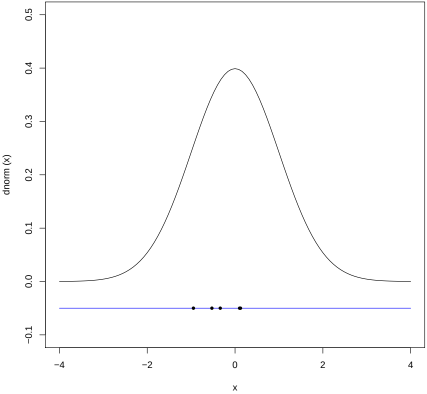
The normal distribution



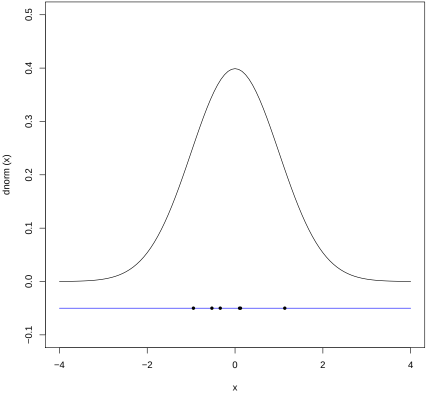
The normal distribution



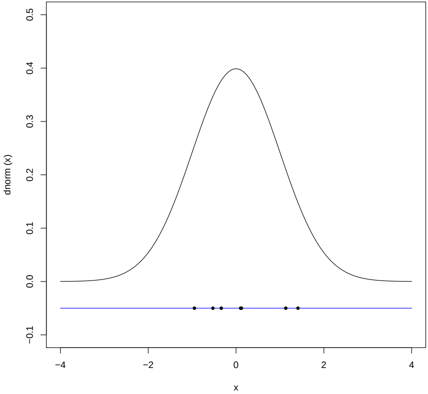
The normal distribution



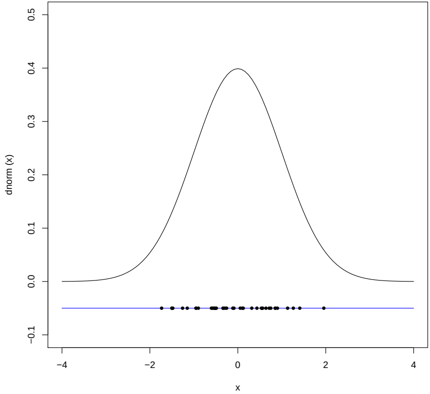
The normal distribution



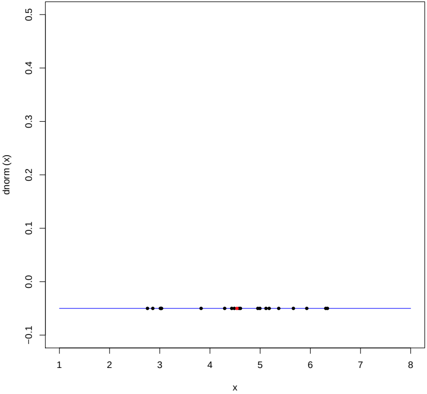
The normal distribution



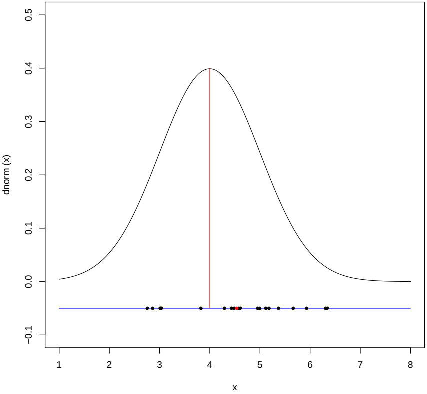
The normal distribution



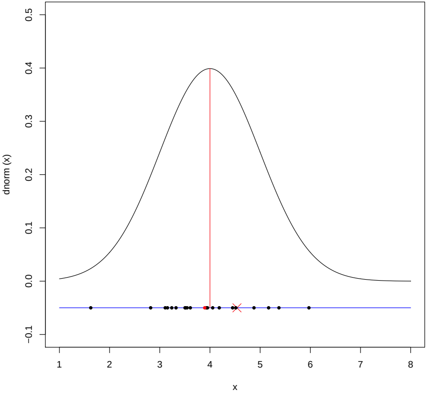
The mean as an *estimator* of the “true” centre



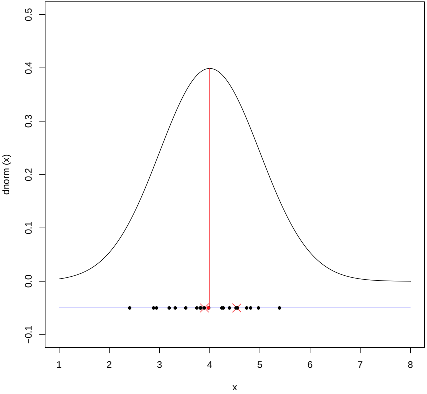
The mean as an *estimator* of the “true” centre



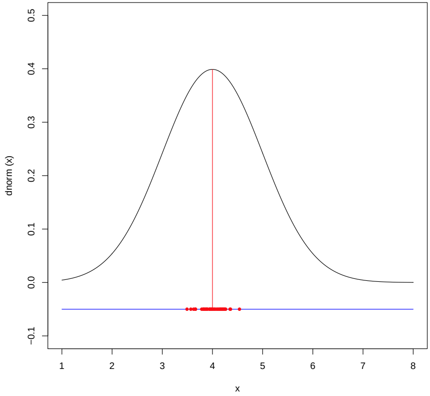
A new dataset will yield a different mean.



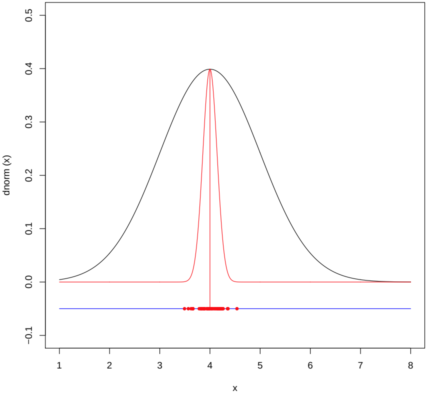
A new dataset will yield a different mean.



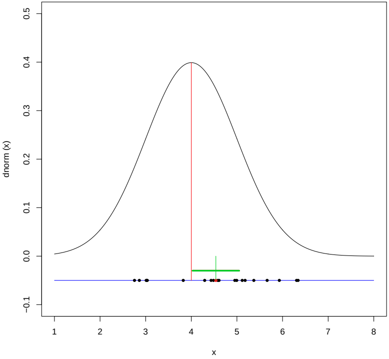
All means are ”erroneous”
but they are not bad and *on average* correct.



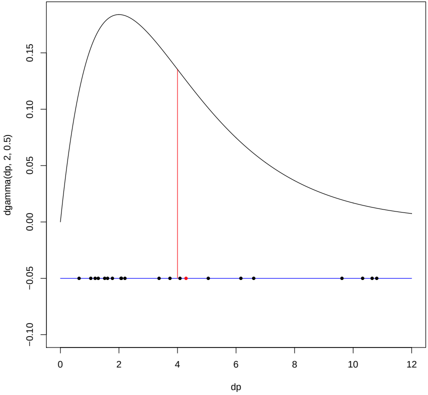
The means follow their own (normal) distribution.



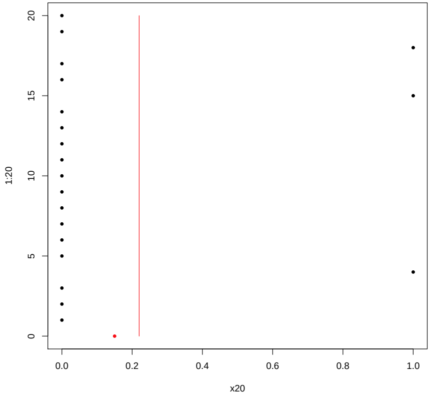
This allows to estimate “standard error”
(sd of the distribution of means)
and to construct a (95%) “confidence interval”.



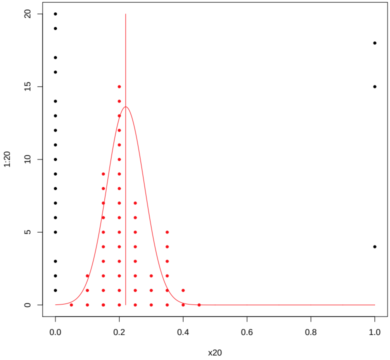
The same thing works for other distributions.



The same thing works for other distributions.



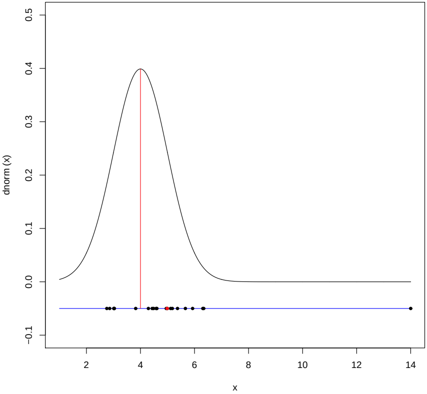
The mean is (often)
approximately *normally* distributed
even if the underlying distribution is different.



The mathematical theory of distributions ensures that things “work” in the abstract mathematical world.

Not only “true means” can be estimated, also other quantities such as variances, regression coefficients and functions (quantifying how a y depends on some x).

Things may go wrong, though,
for example with outliers.



Problems that can be analysed
within the mathematical world:

Error: an estimator deviates from the true value.

Problems that can be analysed
within the mathematical world:

Error: an estimator deviates from the true value.
Bias: estimators deviate
from the true value *on average*
(e.g., because of asymmetric contamination)

Problems that can be analysed
within the mathematical world:

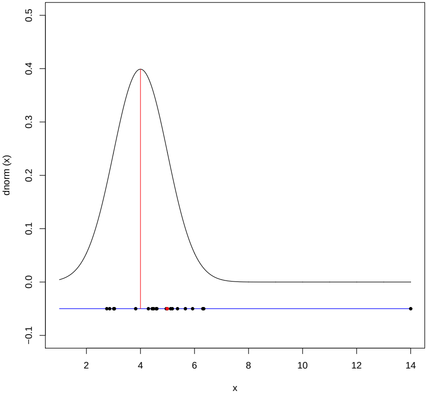
Error: an estimator deviates from the true value.
Bias: estimators deviate
from the true value *on average*
(e.g., because of asymmetric contamination)
Sensitivity/instability: estimators can change a lot
if data are changed little (ie, by adding outliers).

Problems that can be analysed
within the mathematical world:

Error: an estimator deviates from the true value.
Bias: estimators deviate
from the true value *on average*
(e.g., because of asymmetric contamination)
Sensitivity/instability: estimators can change a lot
if data are changed little (ie, by adding outliers).
Violation of model assumptions:
Theory derived from certain distribution is applied
where a different distribution is “true”.

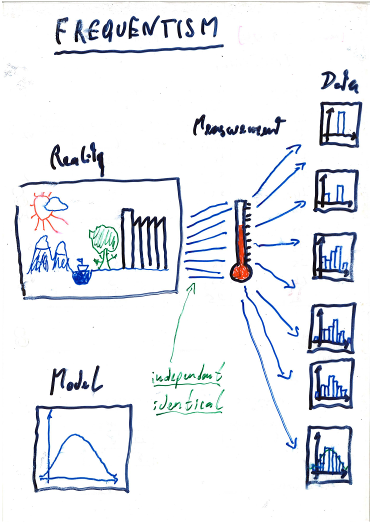
Error can be assessed in a routine manner;
other problems are complex to treat
because there are so many possibilities.

May look at data and spot the problem.



2. What about reality?

(Frequentist) statistician would assume
that real data generation
"works like" a probability distribution.


FREQUENTISM
Data
Measurement
Reality
Model
independent
identical

Is this realistic?

Well, not exactly. It's an idealisation.
Ignores *specific differences between observations*
by treating them as "repetitions".

Is this realistic?

Well, not exactly. It's an idealisation.
Ignores *specific differences between observations*
by treating them as "repetitions".

Ignores dependence and trends over time
on some level
(these can actually be modelled
by making more complex assumptions).

"Observation $=$ Truth $+$ Error"
assumes well defined truth,
which is not directly observable.

"Observation $=$ Truth $+$ Error"
assumes well defined truth,
which is not directly observable.

Alternative view:
"Observation $=$ Signal $+$ Noise"

separates "important overall pattern"
from "variation not of interest".

That's very powerful and more "realistic"
than deterministic modelling ignoring "noise".
It enables to *quantify uncertainty*.
There is pretty much no other way.

That's very powerful and more "realistic"
than deterministic modelling ignoring "noise".
It enables to *quantify uncertainty*.
There is pretty much no other way.

The critical point is:
What do the model assumptions entail,
what assumptions do we want to make?

(Note that not modelling noise can be seen
as *even stronger assumption*.)

3. An example: homeopathy

A standard random effects meta-analysis
From Shang et al. “Are the clinical effects of homoeopathy placebo effects? Comparative study of placebo-controlled trials of homoeopathy and allopathy” (The Lancet, 2005)

Famous study,
prompted Lancet-editorial “The Death of Homeopathy”,
major reference Wikipedia-homeopathy, Singh & Ernst etc.

Note: many details suppressed!

$$L_i = \theta + \eta_i + e_i, \quad i = 1, \ldots, n,$$
$$\eta_i \sim \mathcal{N}(0, \sigma_0^2), \quad e_i \sim \mathcal{N}(0, \sigma_i^2) \text{ independent.}$$

$$L_i = \theta + \eta_i + e_i, \quad i = 1, \ldots, n,$$
$$\eta_i \sim \mathcal{N}(0, \sigma_0^2), \quad e_i \sim \mathcal{N}(0, \sigma_i^2) \text{ independent.}$$

$n = 8$. For study i : p_i est. prob. that “placebo works”, q_i est. prob. that “homeopathy works”, $L_i = \log \frac{p_i/(1-p_i)}{q_i/(1-q_i)}$,

$$L_i = \theta + \eta_i + e_i, \quad i = 1, \ldots, n,$$
$$\eta_i \sim \mathcal{N}(0, \sigma_0^2), \quad e_i \sim \mathcal{N}(0, \sigma_i^2) \text{ independent.}$$

$n = 8$. For study i : p_i est. prob. that "placebo works", q_i est. prob. that "homeopathy works", $L_i = \log \frac{p_i/(1-p_i)}{q_i/(1-q_i)}$,
L_i negative $\Rightarrow$ homeopathy better than placebo

$$L_i = \theta + \eta_i + e_i, \quad i = 1, \ldots, n,$$
$$\eta_i \sim \mathcal{N}(0, \sigma_0^2), \quad e_i \sim \mathcal{N}(0, \sigma_i^2) \text{ independent.}$$

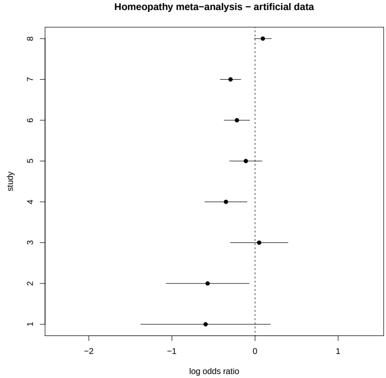
$n = 8$. For study i : p_i est. prob. that "placebo works", q_i est. prob. that "homeopathy works", $L_i = \log \frac{p_i/(1-p_i)}{q_i/(1-q_i)}$,
L_i negative $\Rightarrow$ homeopathy better than placebo
Aim: evidence for/against $\theta = 0$,
interpreted as "homeopathy equals placebo"?

$$L_i = \theta + \eta_i + e_i, \quad i = 1, \ldots, n,$$
$$\eta_i \sim \mathcal{N}(0, \sigma_0^2), \quad e_i \sim \mathcal{N}(0, \sigma_i^2) \text{ independent.}$$

$n = 8$. For study i : p_i est. prob. that "placebo works",
q_i est. prob. that "homeopathy works", $L_i = \log \frac{p_i/(1-p_i)}{q_i/(1-q_i)}$,
L_i negative $\Rightarrow$ homeopathy better than placebo
Aim: evidence for/against $\theta = 0$,
interpreted as "homeopathy equals placebo"?

η_i study specific effect, e_i within study variation.
σ_i standard error of L_i (small if n_i large),

$$L_i = \theta + \eta_i + e_i, \quad i = 1, \ldots, n,$$
$$\eta_i \sim \mathcal{N}(0, \sigma_0^2), \quad e_i \sim \mathcal{N}(0, \sigma_i^2) \text{ independent.}$$



$$L_i = \theta + \eta_i + e_i, \quad i = 1, \ldots, n,$$
$$\eta_i \sim \mathcal{N}(0, \sigma_0^2), \quad e_i \sim \mathcal{N}(0, \sigma_i^2).$$

Shang et al. find
overall estimate $\hat{\theta} = -0.13$,
95%-confidence interval $[-0.43, 0.17]$ for θ,
conclude (because 0 is in CI)
"no evidence for homeopathy better than placebo."

$$L_i = \theta + \eta_i + e_i, \quad i = 1, \ldots, n,$$
$$\eta_i \sim \mathcal{N}(0, \sigma_0^2), \quad e_i \sim \mathcal{N}(0, \sigma_i^2).$$

Model assumptions:

1. Independence of studies, η_i and e_i
2. Additive model for L_i.
3. Normal distribution for η_i.
4. Normal distribution for e_i.

$$L_i = \theta + \eta_i + e_i, \quad i = 1, \ldots, n,$$
$$\eta_i \sim \mathcal{N}(0, \sigma_0^2), \quad e_i \sim \mathcal{N}(0, \sigma_i^2).$$

Normal distribution for e_i.
Motivated approximately from a within-study model with "success probabilities" per patient.

$$L_i = \theta + \eta_i + e_i, \quad i = 1, \ldots, n,$$
$$\eta_i \sim \mathcal{N}(0, \sigma_0^2), \quad e_i \sim \mathcal{N}(0, \sigma_i^2).$$

Normal distribution for e_i.
Motivated approximately from a within-study model with “success probabilities” per patient.

σ_i^2 assumed known;
actually estimated from studies
(but probably reliable for big studies).

$$L_i = \theta + \eta_i + e_i, \quad i = 1, \ldots, n,$$
$$\eta_i \sim \mathcal{N}(0, \sigma_0^2), \quad e_i \sim \mathcal{N}(0, \sigma_i^2).$$

Independence of studies, η_i and e_i
Standard practice, usually from
"we don't know why they should be dependent".

Actually could be dependent by
"patients in different studies have read
same newspaper stories about homeopathy",
"fashionable diseases for studies" etc.

$$L_i = \theta + \eta_i + e_i, \quad i = 1, \ldots, n,$$
$$\eta_i \sim \mathcal{N}(0, \sigma_0^2), \quad e_i \sim \mathcal{N}(0, \sigma_i^2).$$

Independence of studies, η_i and e_i
Standard practice, usually from
"we don't know why they should be dependent".

Actually could be dependent by
"patients in different studies have read
same newspaper stories about homeopathy",
"fashionable diseases for studies" etc.

Use common sense to decide whether that's a problem.

$$L_i = \theta + \eta_i + e_i, \quad i = 1, \ldots, n,$$
$$\eta_i \sim \mathcal{N}(0, \sigma_0^2), \quad e_i \sim \mathcal{N}(0, \sigma_i^2).$$

Additive model for L_i.
Can check this by diagnostic plots
but hard with $n = 8$.

$$L_i = \theta + \eta_i + e_i, \quad i = 1, \dots, n,$$
$$\eta_i \sim \mathcal{N}(0, \sigma_0^2), \quad e_i \sim \mathcal{N}(0, \sigma_i^2).$$

Normal distribution for η_i.

This is the most problematic one.

- “Study effect” is not computed as “mean”, so “means are approximately normal” won’t work.
- $n = 8$ not enough to assess distributional shape.
- Researchers picked studies by “quality criteria”, study effect is “biased” as “random sample”.

Furthermore, studies apply homeopathy differently;
only 2 studies treat "classical homeopathy".

Modelling all as repetitions implies that
how homeopathy is applied is not "difference of interest".
$\Rightarrow$ "classical homeopathy" paradigm
is not tested by this study,

and homeopaths cannot accept that study effects
are identically distributed.

"Study effect" is treated as result
from independent repetition

$\Rightarrow$ a single study can't estimate θ at all,

"Study effect" is treated as result
from independent repetition

$\Rightarrow$ a single study can't estimate θ at all,
"Effective sample size" is $n = 8$, not $\sum n_i > 5,000$,
bad power, i.e.,
non-significance can easily happen under $\theta < 0$.

"Study effect" is treated as result
from independent repetition

$\Rightarrow$ a single study can't estimate θ at all,
"Effective sample size" is $n = 8$, not $\sum n_i > 5,000$,
bad power, i.e.,
non-significance can easily happen under $\theta < 0$.

$\Rightarrow$ the study's case against homeopathy
is quite weak, and was hugely overstated.

However, the study still gives a valuable summary
of existing evidence.

4. Conclusion

- Distributions are abstract data generators.
- They allow to quantify random variation, noise and error.

4. Conclusion

- Distributions are abstract data generators.
- They allow to quantify random variation, noise and error.
- Applying them to reality relies on idealisation (though arguably often not more than not applying them).
- Choosing a model in reality has implications for how reality is perceived.

4. Conclusion

- Distributions are abstract data generators.
- They allow to quantify random variation, noise and error.
- Applying them to reality relies on idealisation (though arguably often not more than not applying them).
- Choosing a model in reality has implications for how reality is perceived.
- Can discuss all the model assumptions and what they mean.
  This allows to decide whether model is appropriate,
  and how we think about the situation!