ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Understanding statistical inference based on models that aren't true

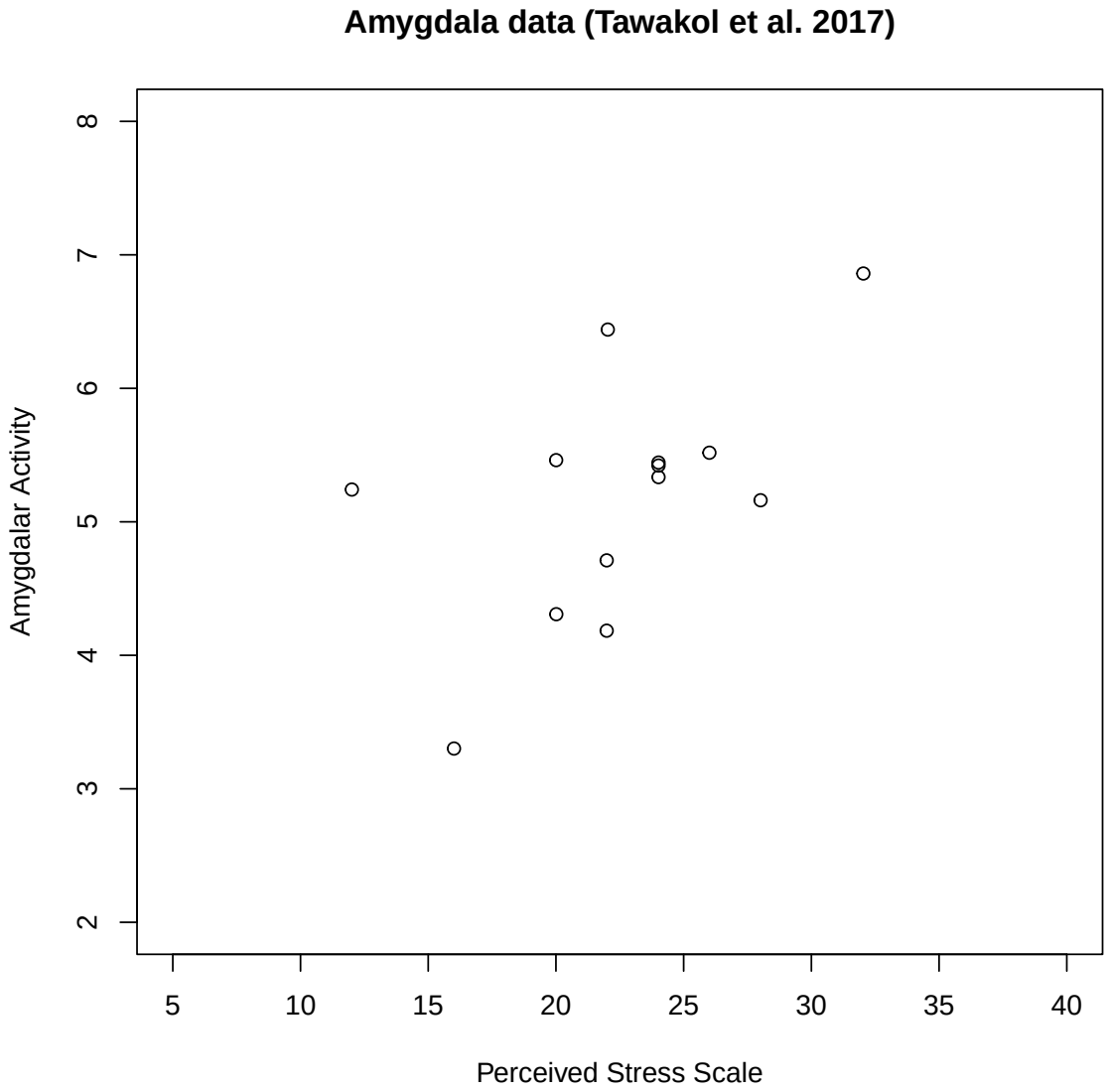
Christian Hennig

1. Introduction: some data analysis

Amygdala data from Tawakol et al. (2017),
analysed by four groups in van Dongen et al. (2019).

Is perceived stress positively associated
with resting activity in the amygdala?

Measurements from $n = 13$ individuals (volunteers)
with post-traumatic stress disorder.


Amygdala data (Tawakol et al. 2017)
Amygdalar Activity
Perceived Stress Scale
2
3
4
5
6
7
8
5
10
15
20
25
30
35
40

Tawakol et al. test $H_0: \rho = 0$ vs. $H_1: \rho \neq 0$.

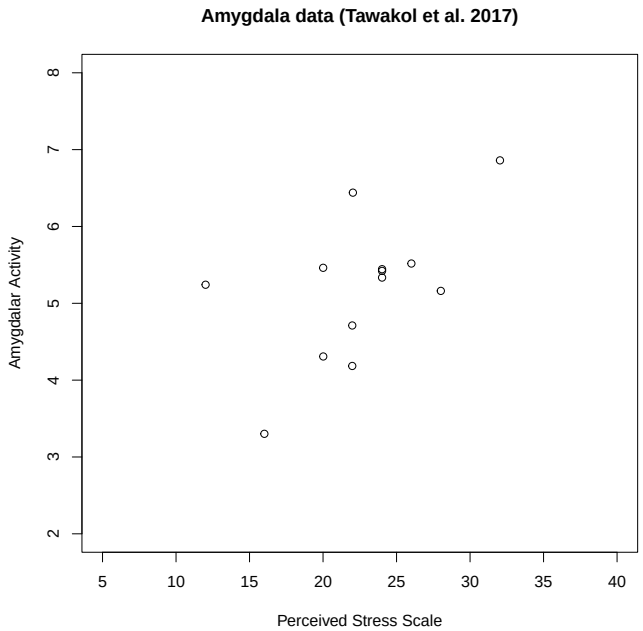
They use Pearson correlation

$$R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

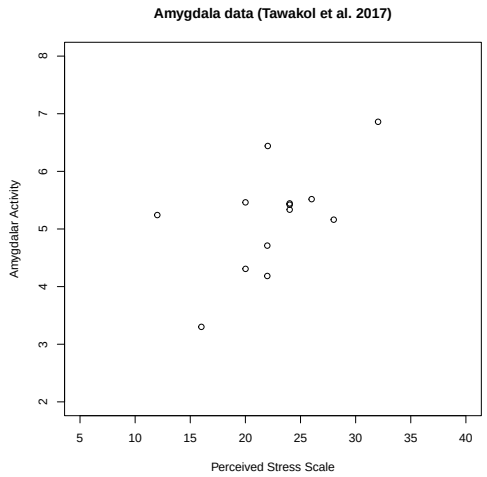
Assuming $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. bivariate normal (independently identically distributed),
test can be based on $R\sqrt{\frac{n-2}{1-R^2}} \sim t_{n-2}$.

This yields $r = 0.56,\ p = 0.048$,
95% confidence interval $[0.01, 0.85]$ for ρ.

Tawakol et al. state:
“Perceived stress was associated with amygdalar activity.”



Based on "$(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. bivariate normal"
- Tawakol et al. state these variables were "normal",
suggesting in turn that model assumptions were fulfilled.



OK, with $n = 13$ data show
no specific indication against normality,
but with $n = 13$ not much can be seen in the first place.

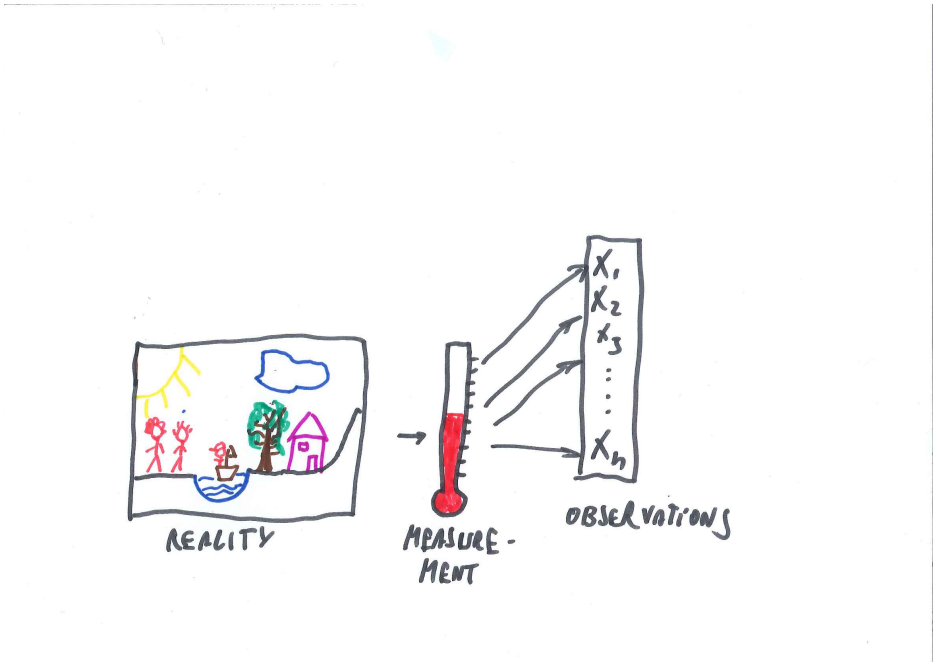
No information about i.i.d.
(Do some test persons know each other?
Any meaningful grouping?)

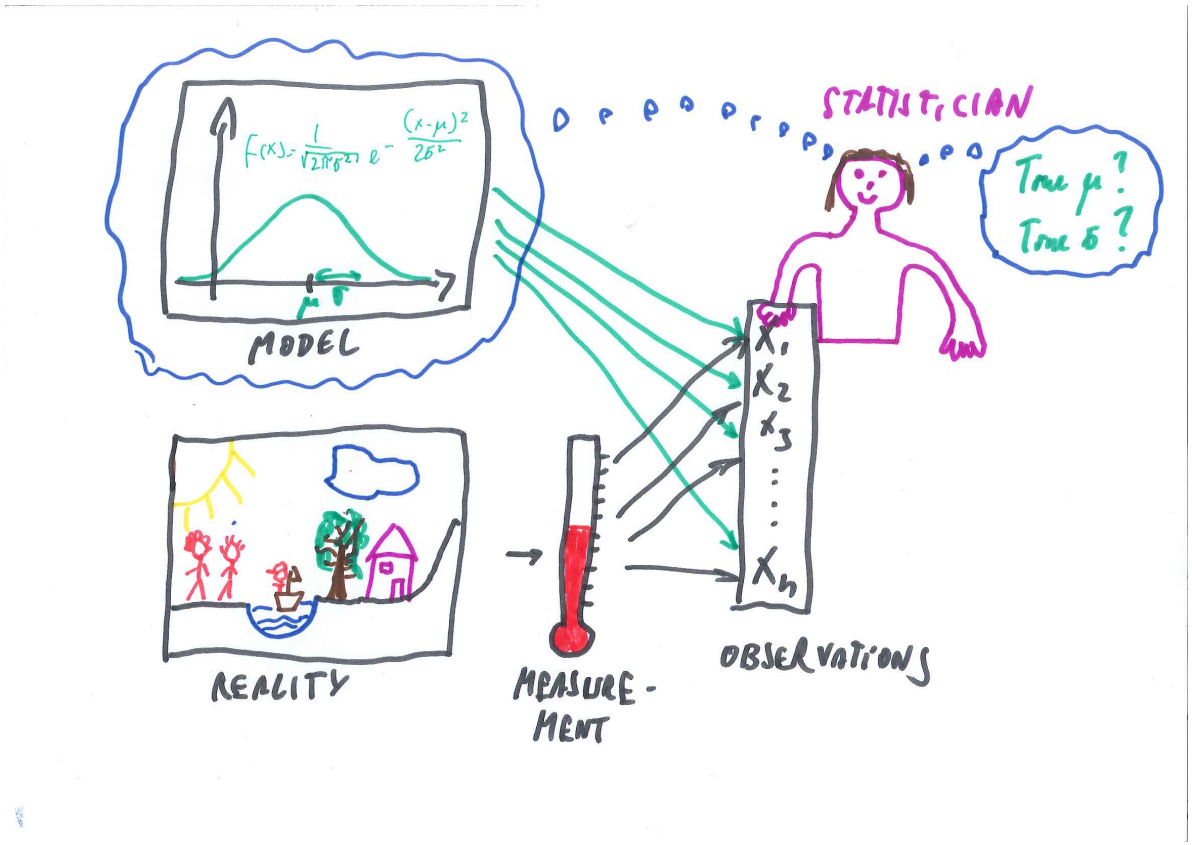
But in fact we know normality is violated!
Variables can only be positive;
normal distribution has nonzero probability
for negative observations.

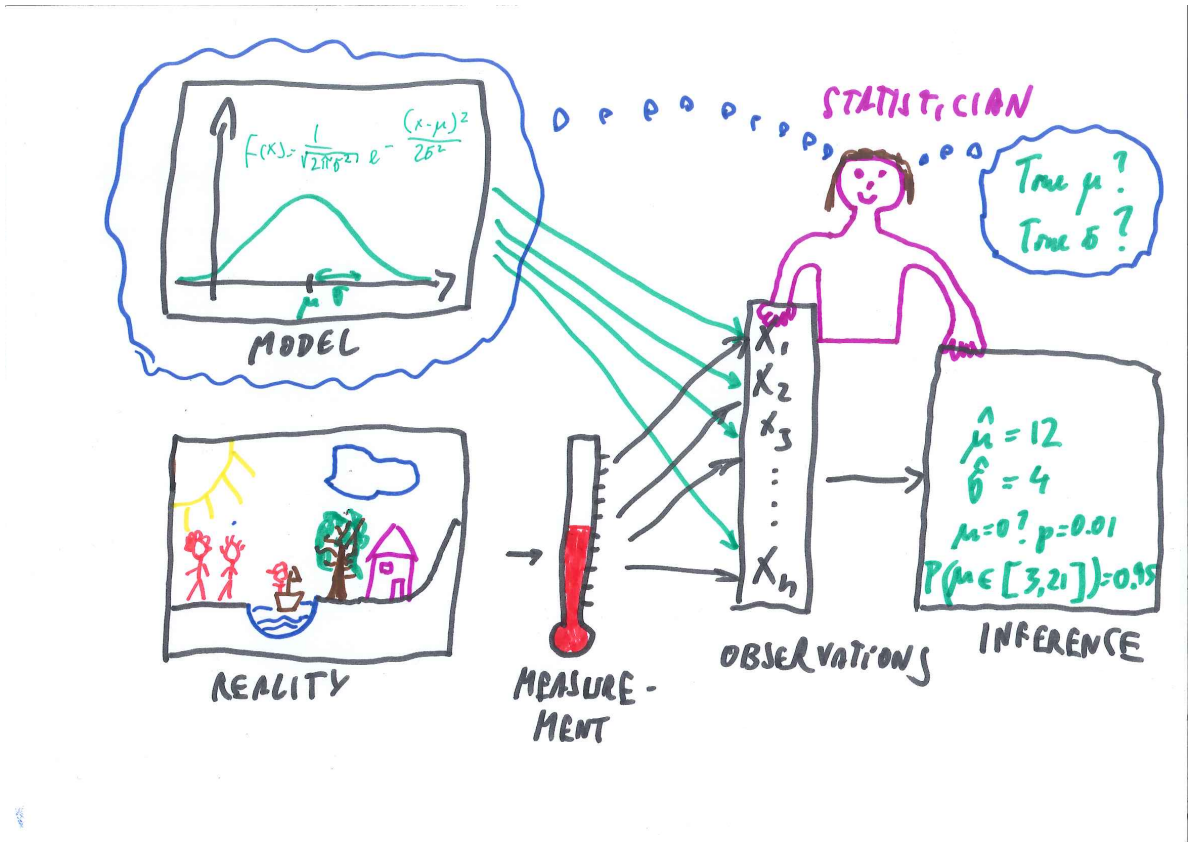
Thich Nhat Hanh said:
"Everything depends on everything else."

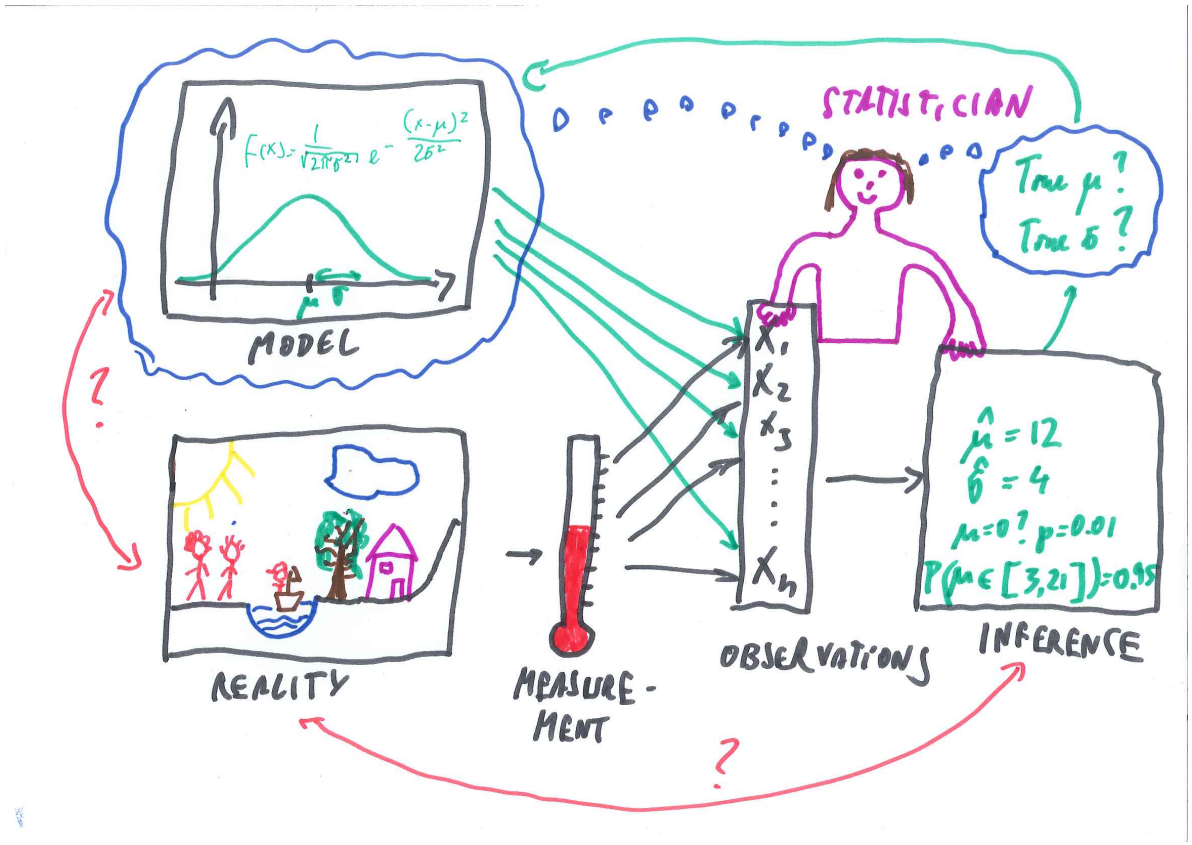
And George Box:
"All models are wrong but some are useful."
(But this one?)

2. Statistical modelling
What is going on?




MODEL
STATISTICIAN
True μ ?
True σ ?
REALITY
MEASURE-
MENT
OBSERVATIONS


MODEL
REALITY
MEASURE-
MENT
OBSERVATIONS
STATISTICIAN
True μ ?
True σ ?
INFERENCE


MODEL
STATISTICIAN
True μ ?
True σ ?
REALITY
MEASURE-
MENT
OBSERVATIONS
INFERENCE
?
?

Statistical inference is based on mathematical reasoning in the “model world”.

The model world is essentially different from the real world.

Data connect model world and real world,
but it is far from trivial to understand
what model world results mean for the real world.

"Model-based statistical inference is valid if and only if the model is true."

"Model-based statistical inference is valid if and only if the model is true."

This is misleading!

It's not the job of models to be "true". Models are tools for thinking.

What is the job of models?

- Required for handling observations mathematically
- Proved mathematical theory can be used
- Quantify uncertainty
- Quantitative predictions
- Assessment of method performance (controlling “truth”)
- Inspire method choice
- Unambiguous communication of point of view
- Learning from deviations from models
  (using that it’s wrong!)

Amygdala data: $r = 0.56,\ p = 0.048$,
95% confidence interval $[0.01, 0.85]$ for ρ.

These relate to i.i.d. bivariate normal model.

What do they mean in practice?

As long as we are willing to think in terms of model,
it can mean quite something!

At 95%-level, model with true $\rho = 0.01$ compatible with data,
no evidence of meaningfully large effect.

Amygdala data: $r = 0.56,\ p = 0.048$,
95% confidence interval $[0.01, 0.85]$ for ρ.

These relate to i.i.d. bivariate normal model.

What do they mean in practice?

As long as we are willing to think in terms of model,
it can mean quite something!

At 95%-level, model with true $\rho = 0.01$ compatible with data,
no evidence of meaningfully large effect.
But model doesn't hold. Implications?

3. Modelling what happens outside the model

Key concepts:

- *Interpretative hypotheses*
- *Effective hypotheses*

Starting point for *interpretative hypotheses*:
Informal research problem.

3.1 Interpretative hypotheses

Informal H_0:
"Amygdala activity is not associated with stress levels",
Informal H_1:
"Amygdala activity is notably associated with stress levels"

3.1 Interpretative hypotheses

Informal H_0:
"Amygdala activity is not associated with stress levels",
Informal H_1:
"Amygdala activity is notably associated with stress levels"

Translated into probability models:
Interpretative H_0/H_1: All distributions that model real (informal) null/alternative hypothesis of interest.

3.1 Interpretative hypotheses

Informal H_0:
"Amygdala activity is not associated with stress levels",
Informal H_1:
"Amygdala activity is notably associated with stress levels"

Translated into probability models:
Interpretative H_0/H_1: All distributions that model real (informal) null/alternative hypothesis of interest.

No normality and no i.i.d. is implied here!

In principle need to decide for *any* distribution whether it belongs to interpretative H_0, H_1, or none.

Can then investigate for any given test how it behaves ("interpretative" type I/II error probabilities).

In principle need to decide for *any* distribution whether it belongs to interpretative H_0, H_1, or none.

Can then investigate for any given test how it behaves ("interpretative" type I/II error probabilities).

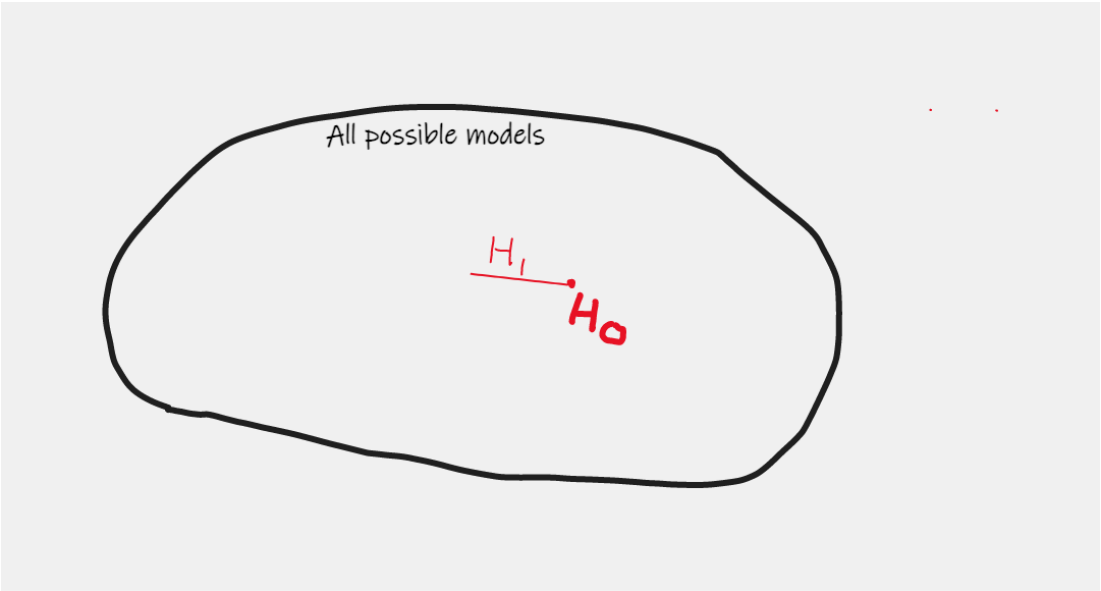
Proviso:
All kinds of models qualify (dependence, non-identity,. . .), but in practice (and theory) we can't handle them all.
It's about understanding the logic and its limitations.

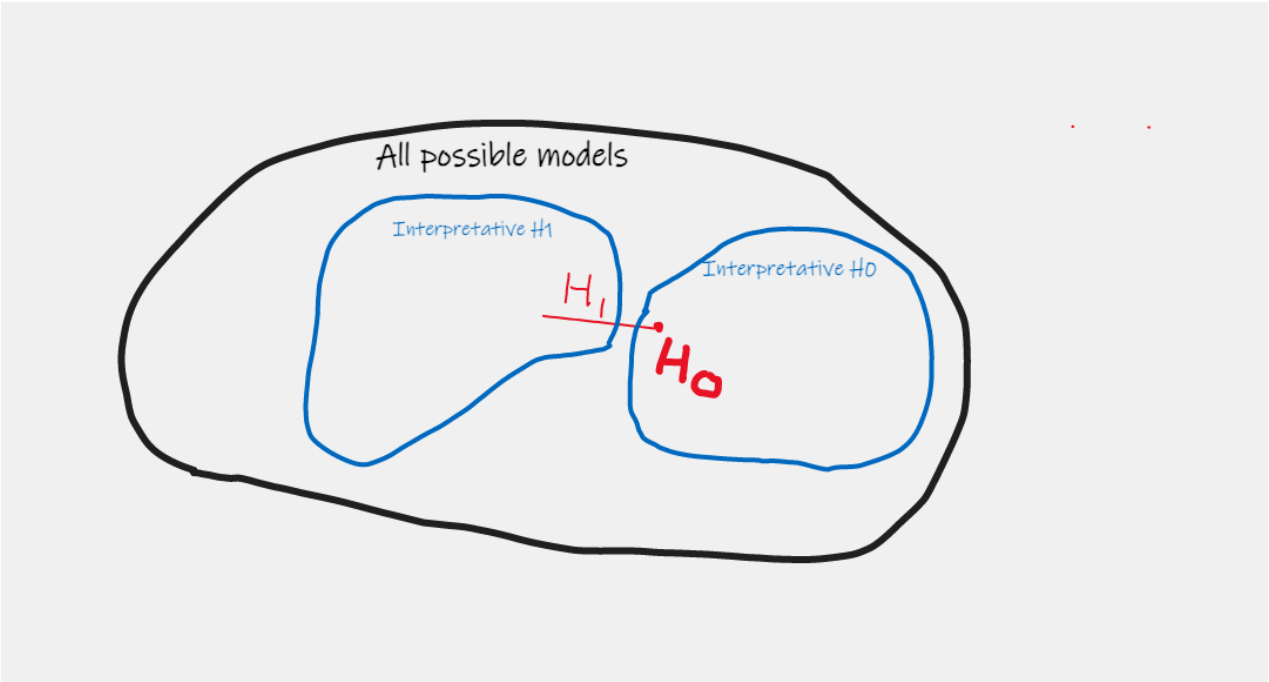
Promote awareness that real hypotheses are informal and could be modelled by many distributions.

Promote awareness that real hypotheses are informal and could be modelled by many distributions.

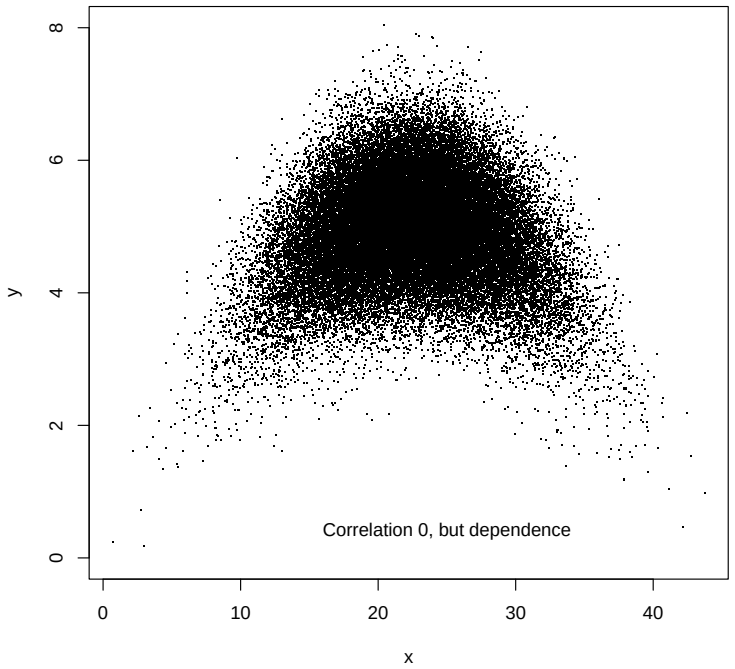
This amounts to modelling that the model doesn't hold.

May question whether *any* (frequentist) model holds.
Use this for benefitting from modelling,
not because belief in true model.


All possible models
H_1
H_0

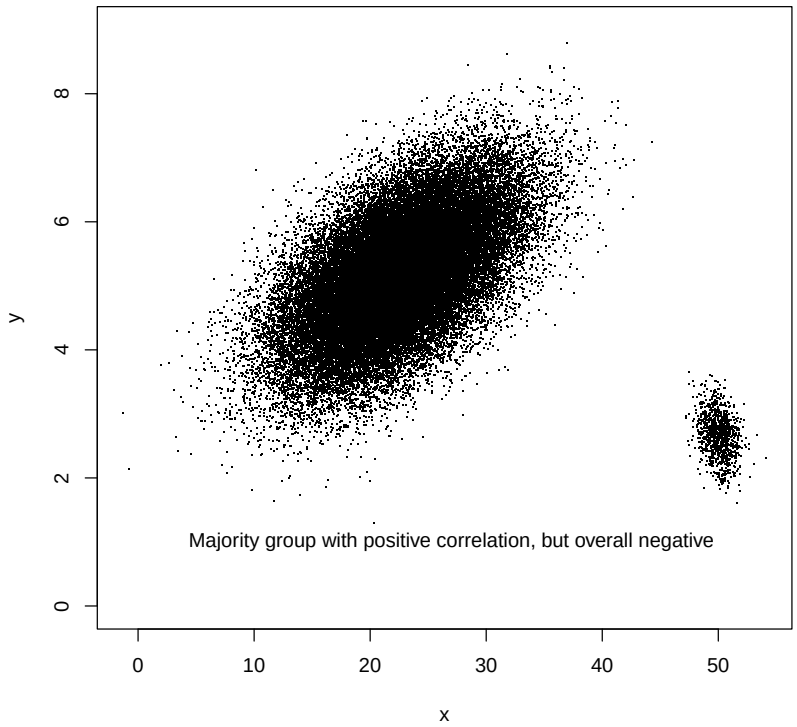

All possible models
Interpretative H1
Interpretative H0
H1
H0

Does this belong to interpretative H_0?



Correlation test will not normally reject, independence test will.

. . . and this?



Interested in majority group? Overall correlation, or is it really ambiguous?

This needs judgment
- data cannot decide this, neither can mathematics!

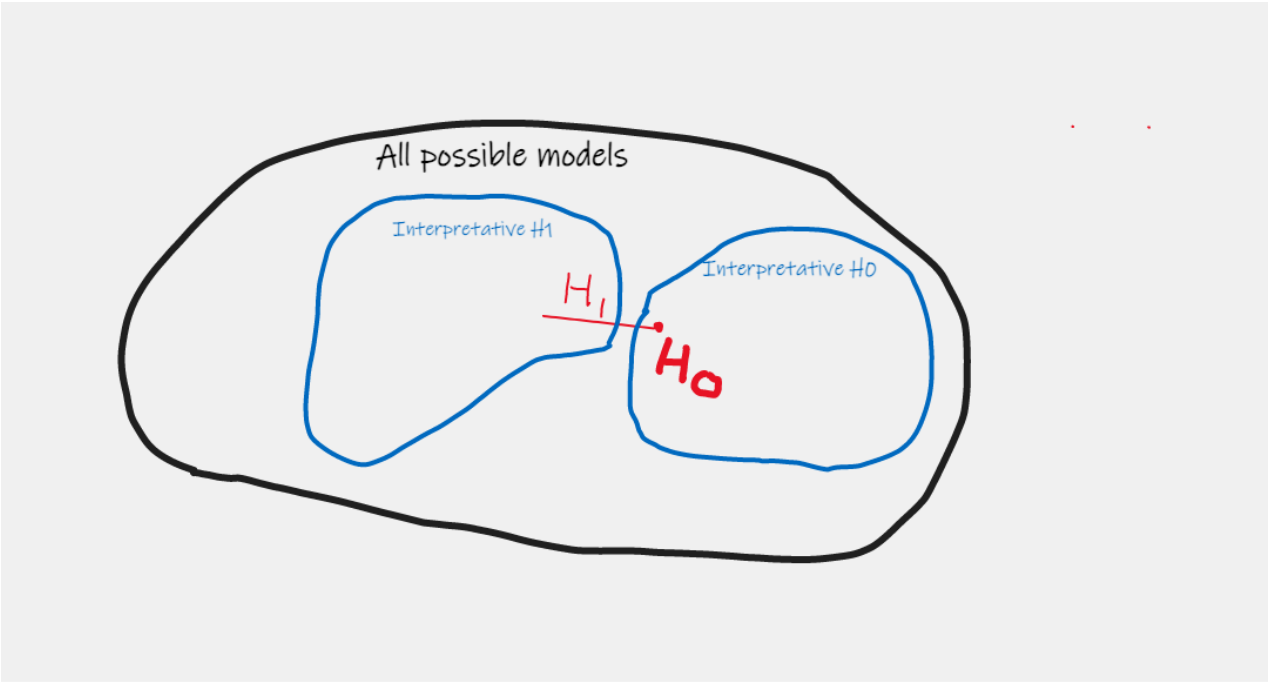
What kind of dependence are we interested in?
Is this appropriately expressed by R?

Interpretative similarity under nominal model:
Tests of point null hypotheses are often criticised for rejecting H_0 for too large n in presence of substantially meaningless deviations from H_0.

In fact even with $n = 13$ authors reject $\rho = 0$, despite confidence interval $[0.01, 0.85]$.

This is a problem because test ignores that parameter values very close to H_0 are often *interpretatively more similar* to H_0 than H_1.

(Need consider effect size etc.
to not misinterpret rejection of formal H_0.)


All possible models
Interpretative H1
Interpretative H0
H1
H0

3.2 Effective hypotheses

What do tests actually do
if we don't take model assumptions for granted?

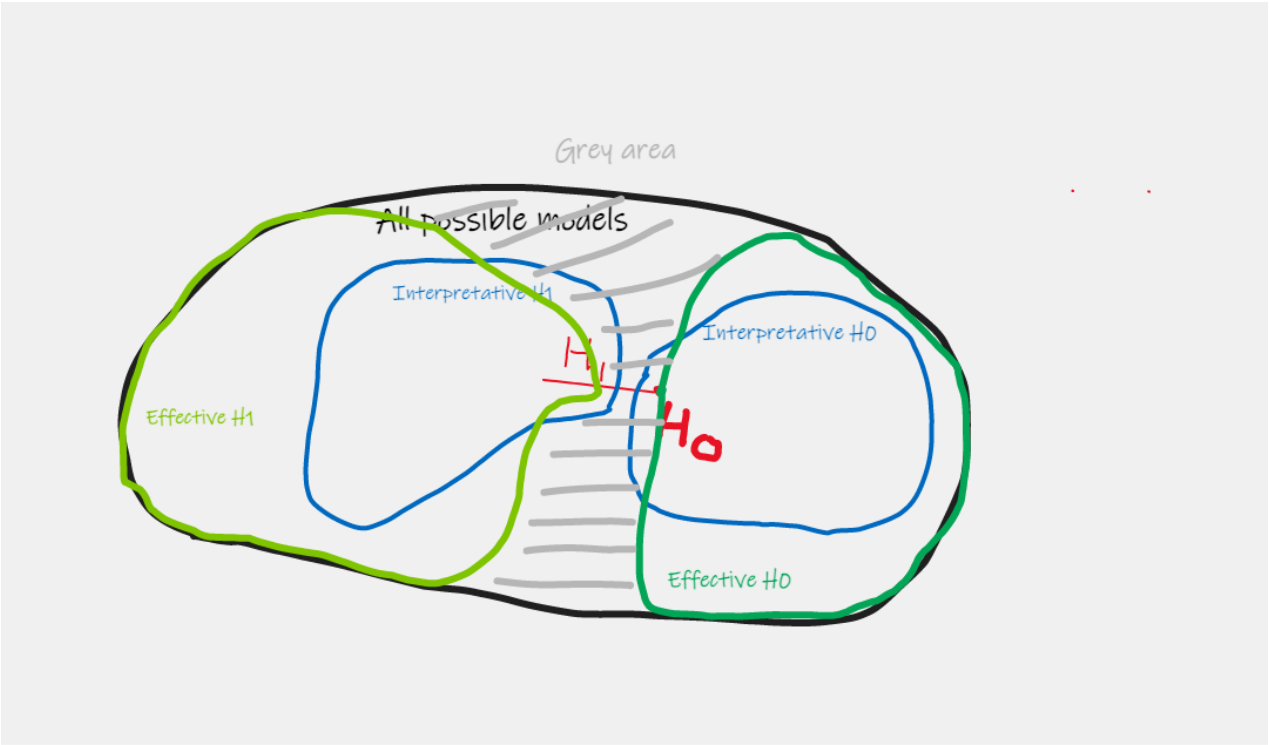
Rejection region $Q \Rightarrow$ tests

- "effective H_0:" any P for which $P(Q) \leq \alpha$ against
- "effective H_1:" any P for which $P(Q)$ large.

This provides a nonparametric definition of a test that originally might well be parametric.

Note that under P with $\alpha < P(Q)$ but $P(Q)$ not large,
the test will reject more easily than under H_0,
but can't be expected to reject.

Such distributions are in a "grey area" w.r.t. the test.


Grey area
All possible models
Interpretative H1
Interpretative H0
H1
H0
Effective H1
Effective H0

t-test with

$$R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

rejecting H_0 for $|R| > c_\alpha$
can be interpreted as testing *general nonparametric*
effective H_0 : P is such that $P\{|R| > c_\alpha\} \leq \alpha$ against
effective H_1 : P is such that $P\{|R| > c_\alpha\}$ large.

t-test with

$$R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

rejecting H_0 for $|R| > c_\alpha$
can be interpreted as testing *general nonparametric*
effective H_0 : P is such that $P\{|R| > c_\alpha\} \leq \alpha$ against
effective H_1 : P is such that $P\{|R| > c_\alpha\}$ large.

The key issue then is:
Does definition of R indicate the relevant *direction*
of deviation from the interpretative H_0?

t-test with

$$R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

rejecting H_0 for $|R| > c_\alpha$
can be interpreted as testing *general nonparametric*
effective H_0 : P is such that $P\{|R| > c_\alpha\} \leq \alpha$ against
effective H_1 : P is such that $P\{|R| > c_\alpha\}$ large.

The key issue then is:
Does definition of R indicate the relevant *direction*
of deviation from the interpretative H_0?

Rather than “are the assumptions fulfilled”? (Which they aren’t.)

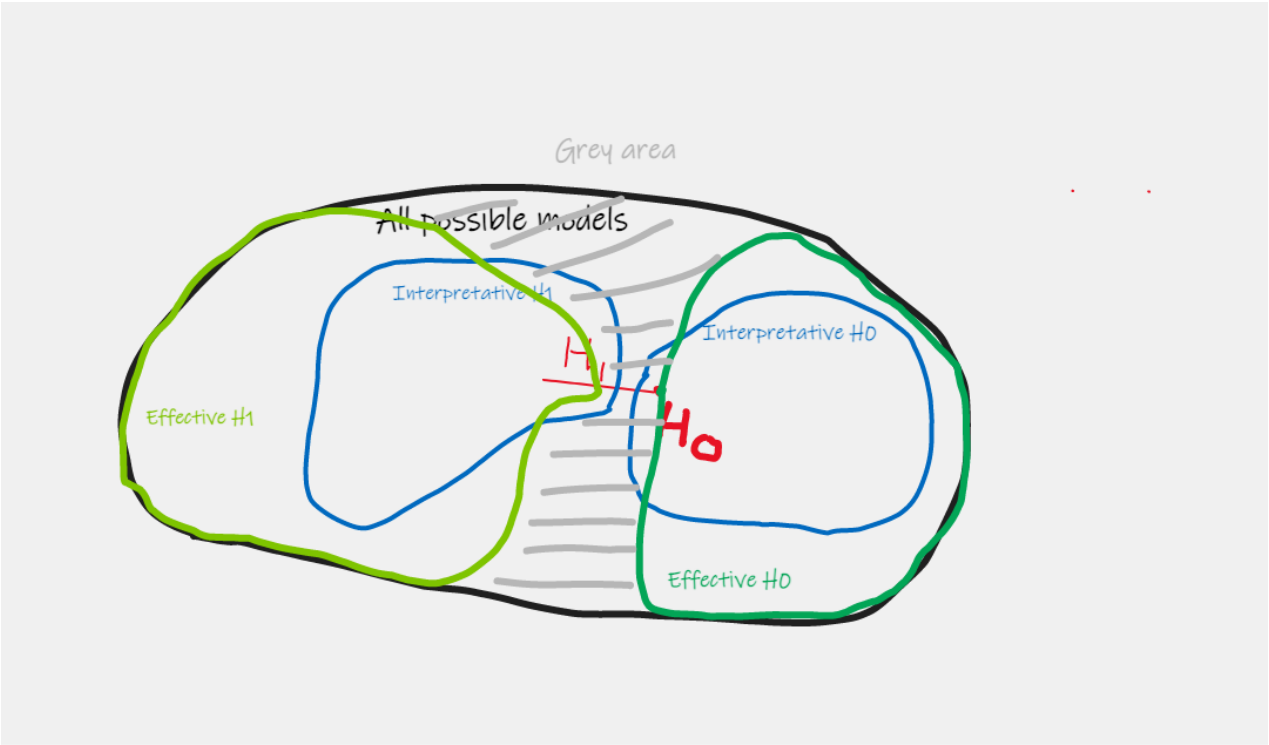
This amounts to understanding whether R
as aggregation of the information in the data
is "interpretatively correct";
effective H_0/H_1 correspond well to interpretative H_0/H_1.

Need to understand properties of R such as
breakdown under gross outliers,
behaviour under relevant dependence patterns.

This amounts to understanding whether R
as aggregation of the information in the data
is "interpretatively correct";
effective H_0/H_1 correspond well to interpretative H_0/H_1.

Need to understand properties of R such as
breakdown under gross outliers,
behaviour under relevant dependence patterns.

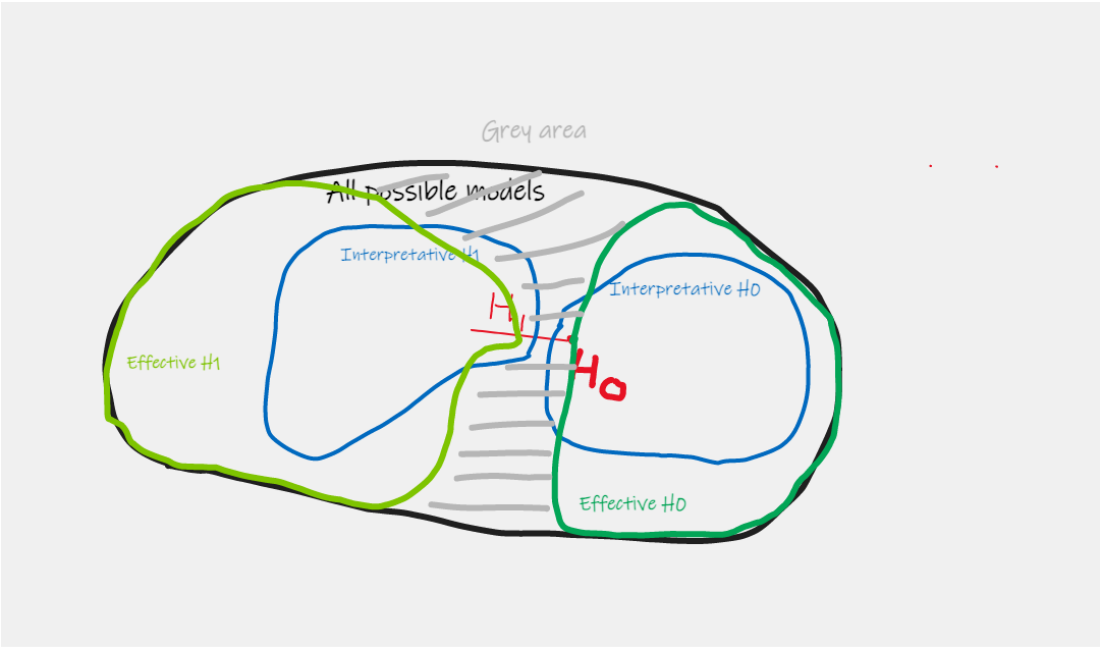
Statisticians tend to think of these statistics as
optimal under certain models,
but they have a *data analytic meaning* on top of it,
and this is crucial to understand for use in inference
without taking model for granted.


Grey area
All possible models
Interpretative H1
Interpretative H0
H1
H0
Effective H1
Effective H0

3.3 Evaluating what happens if the model is not true

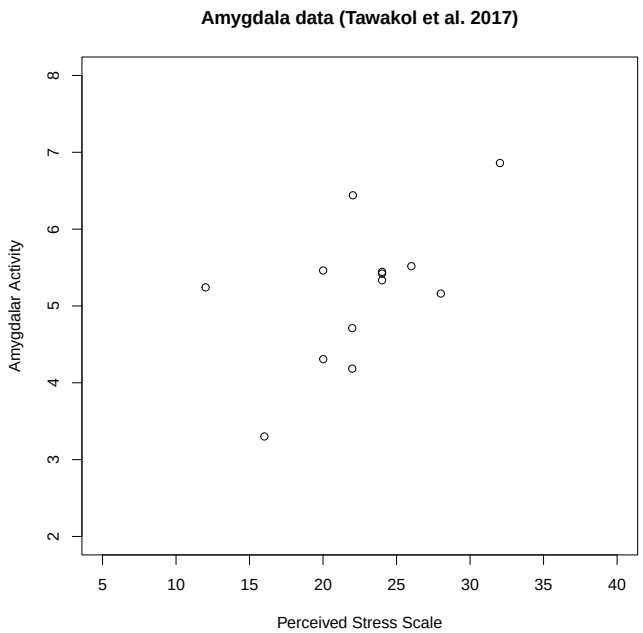
What happens to the methods if the model is not true?

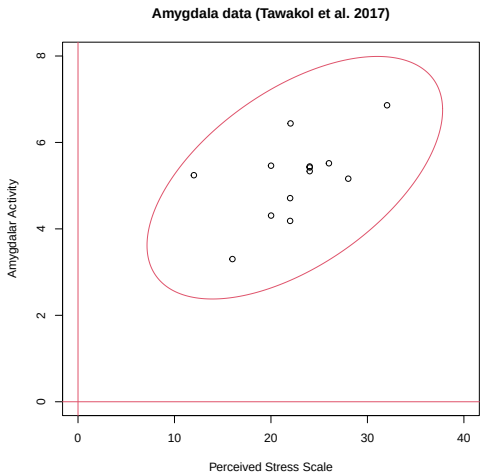
Mathematics and simulation can tell us -
if we *model* deviations from assumed nominal model,
then derive what our method will deliver.
(Even though a modelled deviation from nominal model
isn't really true either.)


Grey area
All possible models
Interpretative H1
Interpretative H0
Effective H1
Effective H0

E.g. model data as truncated normal without negative values, or with correlation *between* observations
compute distribution of R.

Is $P\{|R| > c_\alpha\}$ small/large for P from interpretative H_0/H_1? (Theory/simulation)





Truncation of negative values is harmless
if most of distribution far away from zero.

Correlation *between observations*
can increase variation of R a lot,
and cannot be detected from the data! (Hennig 2023).

5. Final remarks

- **Interpretative/effective hypotheses:**
  explain workings and interpretation of inference
  in fully general model class.
- *Aim:* Choose test so that
  effective and interpretative hypotheses match.
- Tests are only an example;
  can think like this about other model-based inference
- "Binary/discrete" concept of H_0, H_1, "grey area"
  for communication,
  but keep in mind transitions are smooth.

- *Frequentism is not the issue!*
  The "model is not true" issue is just the same
  for Bayesian statistics.
- Robust and nonparametric methods
  can improve link effective $\leftrightarrow$ interpretative hypotheses,
  but can't solve all issues (still assume i.i.d.;
  perform worse in some situations;
  user shouldn't feel too safe)

Practical implications

- Acknowledge that there is model uncertainty
  (sharp distinction between $p = 0.048$ and $p = 0.052$ is meaningless;
  may want to see "seriously low" p for confident rejection
  also because of data dependent decision making.)
- "Set of potentially relevant models" too big to analyse,
  but can simulate inference method behaviour
  under selected potentially relevant models
  that violate assumptions.
- Do model diagnostics, but for learning what goes on,
  not for "making sure assumptions hold".
- Use all background knowledge available
  for spotting potential issues such as unmodelled
  dependence.

References

Bancroft, T. A. (1944) On biases in estimation due to the use of preliminary tests of significance. *Annals of Mathematical Statistics* 15, 190-204.

Hand, D.J. (1997) Scientific and statistical hypotheses: Bridging the gap. In: McKenzie, G., Powell, J. & Usher, R. (Eds.) *Understanding Social Research: perspectives on methodology and practice*. Falmer Press, pp. 124-136.

Hennig, C. (2010) Mathematical models and reality: A constructivist perspective. *Foundations of Science* 15, 29-48.

Hennig, C. (2023) Probability Models in Statistical Data Analysis: Uses, Interpretations, Frequentism-as-Model. In: Sriraman, B. (eds) *Handbook of the History and Philosophy of Mathematical Practice.* Springer, Cham.

Hennig, C. (2023) Parameters not identifiable or distinguishable from data, including correlation between Gaussian observations. *Statistical Papers*, https://doi.org/10.1007/s00362-023-01414-3

Shamsudheen, M. I. and Hennig, C. (2023) Should we test the model assumptions before running a model-based test? *Journal of Data Science, Statistics, and Visualisation*, 3(3).

Tawakol, A. et al. (2017) Relation Between Resting Amygdalar Activity and Cardiovascular Events: A Longitudinal and Cohort Study. *The Lancet*, 389: 834-845.

van Dongen, N. N. N. et al. (2019) Multiple Perspectives on Inference for Two Simple Statistical Scenarios, *The American Statistician*, 73:sup1, 328-339.