ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Testing in models that are not true

Christian Hennig

1. Introduction

Frequentist statistical methods rely on model assumptions, e.g., want to test from a number of measurements whether water turbidity of a river is ≤ 25 NTU (common standard).

Test $H_0: \mu \leq 25$ against $H_1: \mu > 25$ using

$$T = \frac{\bar{X}_n - 25}{S_n/\sqrt{n}}$$

assuming $X_1, \ldots, X_n$ i.i.d. with $\mathcal{L}(X_1) = \mathcal{N}(\mu, \sigma^2)$.

$X_1, \ldots, X_n$ i.i.d. with $\mathcal{L}(X_1) = \mathcal{N}(\mu, \sigma^2)$

What about these assumptions?
Do they have to be fulfilled? Can this be checked?

$X_1, \ldots, X_n$ i.i.d. with $\mathcal{L}(X_1) = \mathcal{N}(\mu, \sigma^2)$

What about these assumptions?
Do they have to be fulfilled? Can this be checked?

But “all models are wrong”! (Though some are useful.)
Why should we check something that we know is wrong?

$X_1, \ldots, X_n$ i.i.d. with $\mathcal{L}(X_1) = \mathcal{N}(\mu, \sigma^2)$

What about these assumptions?
Do they have to be fulfilled? Can this be checked?

But “all models are wrong”! (Though some are useful.)
Why should we check something that we know is wrong?

This is often used as argument against frequentist methods.
“You have to believe the model is true, but it isn’t.”

An issue in testing:

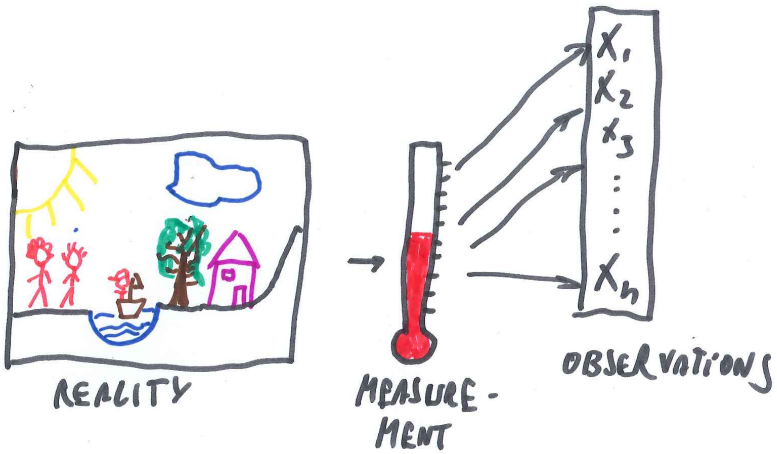
Greenland, Senn et al. (2016):
"In logical terms, the P value tests all the assumptions about how the data were generated, not just the targeted hypothesis it is supposed to test"

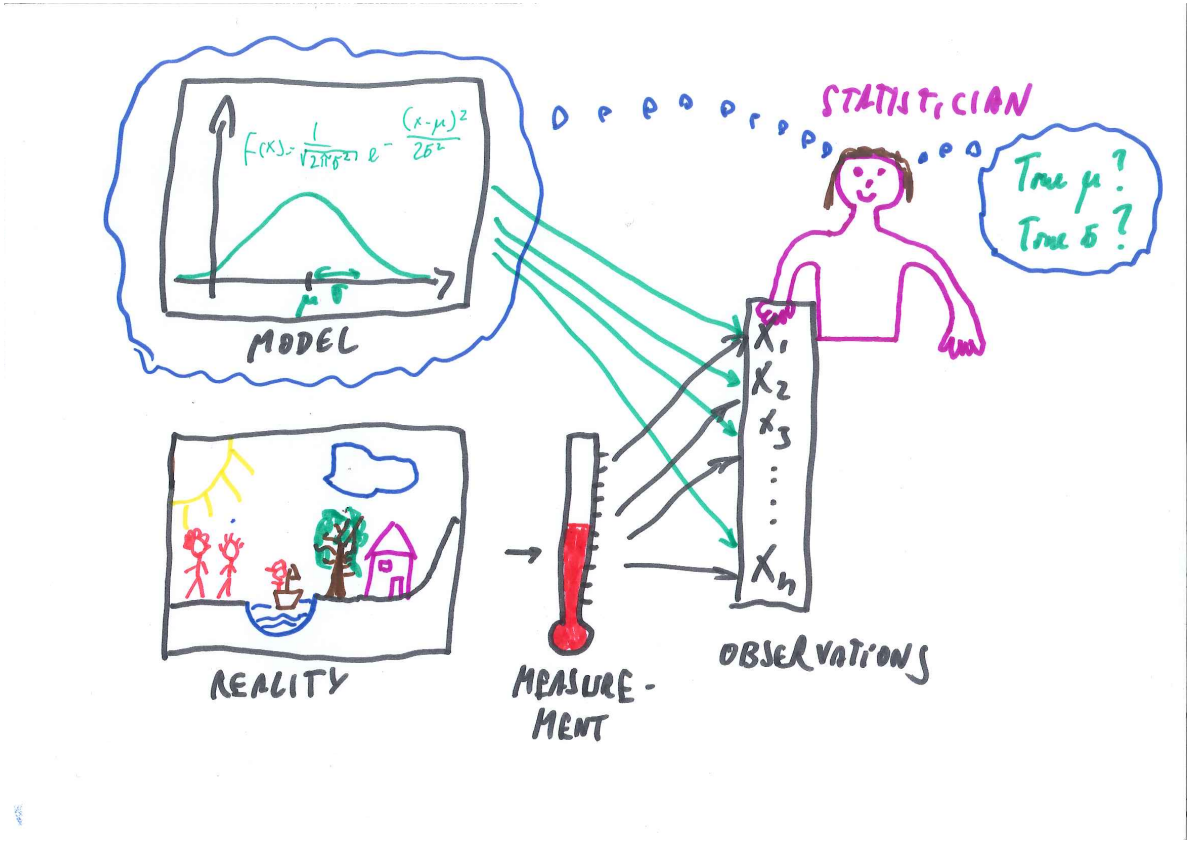
Trafimov (2020, NISS debate):
"I'll make a more general comment, which is that since the model is wrong, in the sense of not being exactly correct, whenever you reject it, you haven't learned anything."

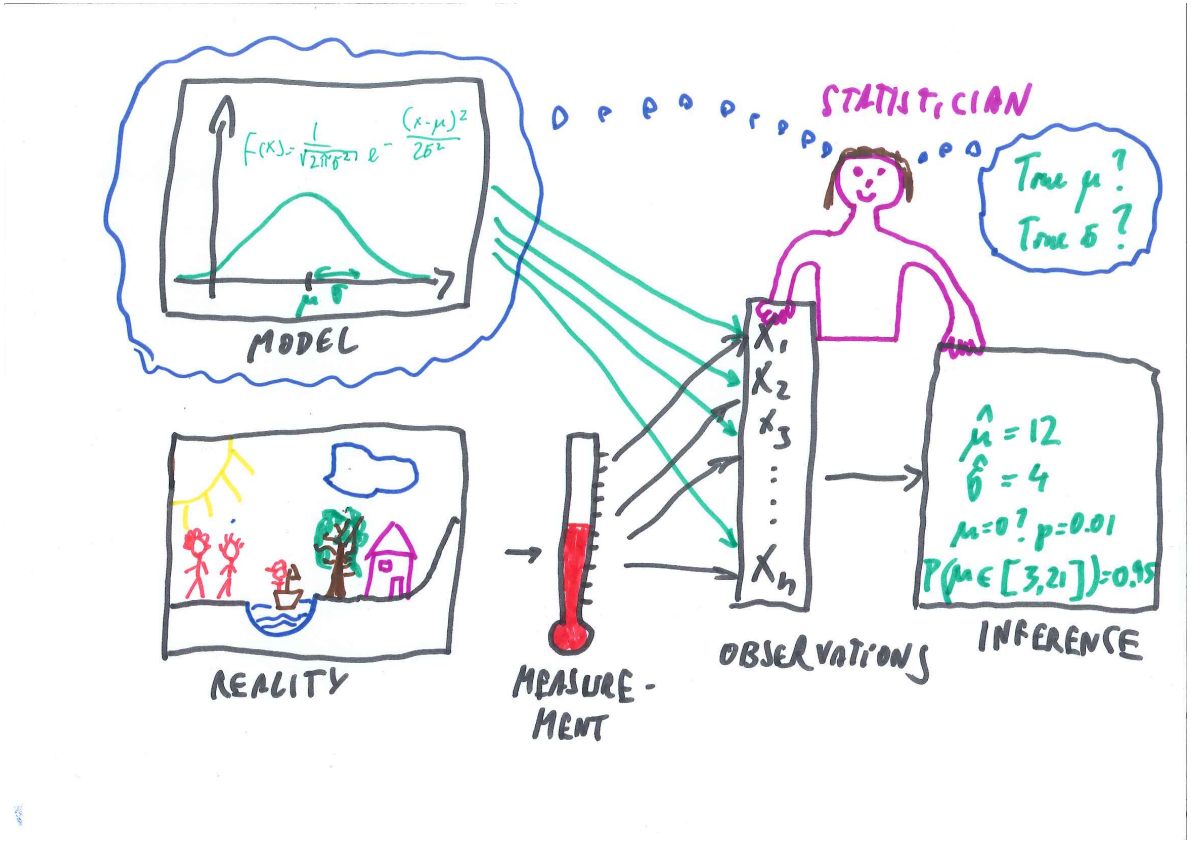
Some are more careful and say, “the model has to be valid”.
Box: “All models are wrong but some are useful.”

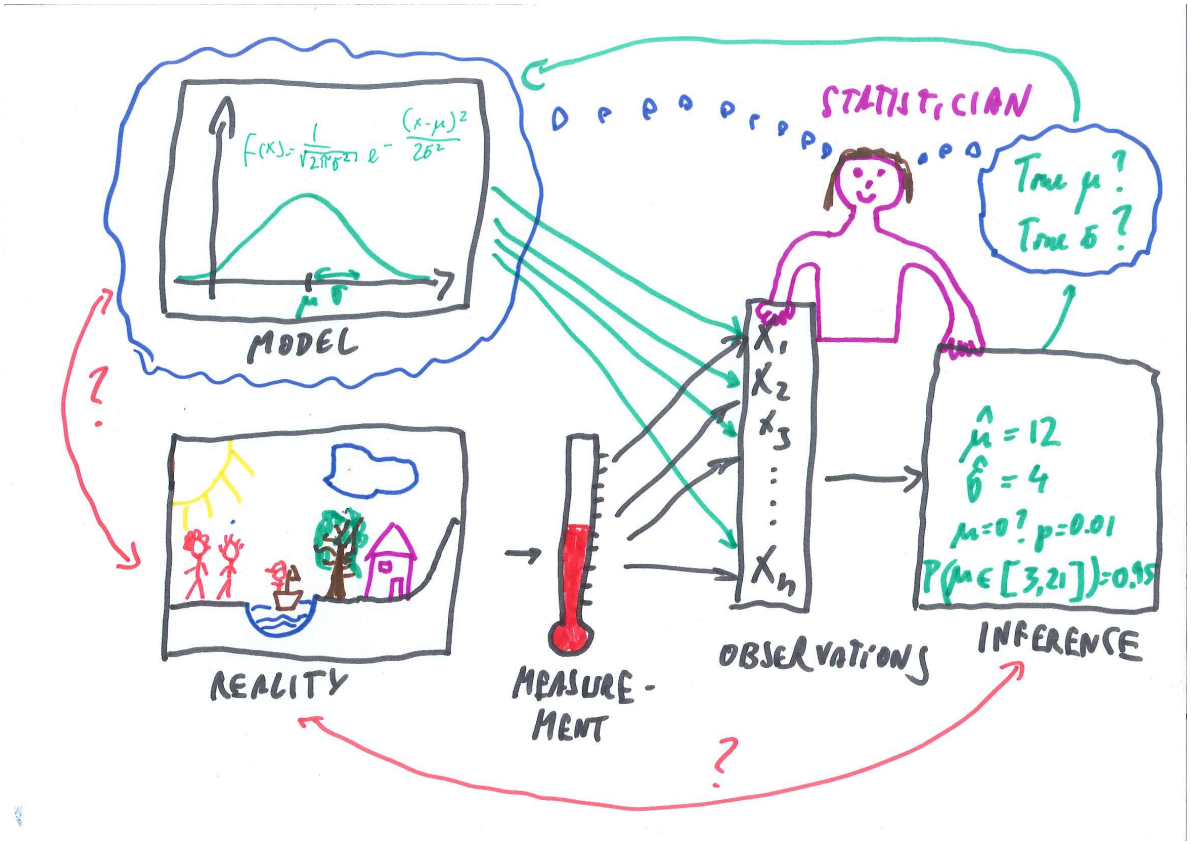
What does this mean, and can we check this?

What is going on?




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


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


STATISTICIAN
True μ ?
True σ ?
MODEL
REALITY
MEASURE-
MENT
OBSERVATIONS
INFERENCE
$\hat{\mu} = 12$
$\hat{\sigma} = 4$
$\mu = 0$? $p = 0.01$
$P(\mu \in [3,21]) = 0.95$

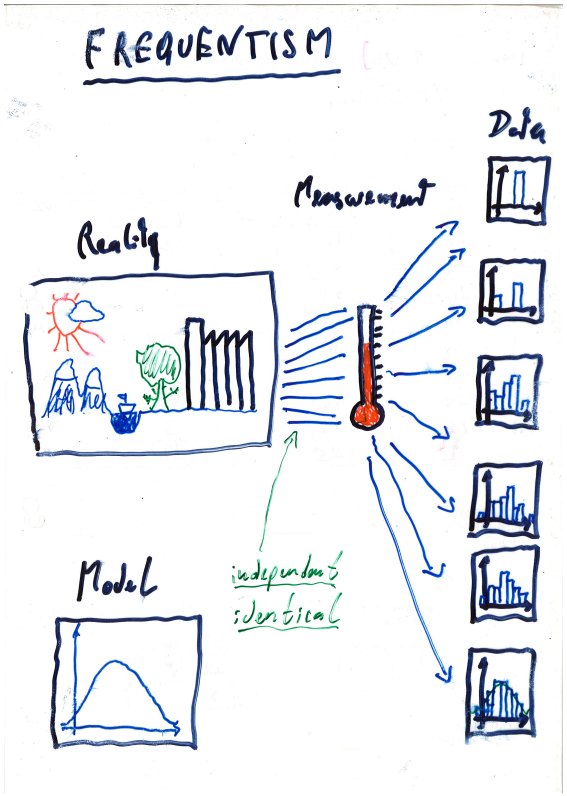
“Model world” and “real world” are separate -
it’s not the job of models to be “true”.
Models are tools for thinking.

"Model world" and "real world" are separate -
it's not the job of models to be "true".
Models are tools for thinking.

Benefits of "model thinking" (even if model not true):

- Predictions (testable)
- Quantification of uncertainty (often testable)
- Inspiration for methods and decisions
- Unambiguous communication of point of view
- Learn through mathematics
- Learn from objections and falsification

Frequentist interpretation of probability:



"We think (at least tentatively) 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 $P(A)$ $\forall A$.)

"We think (at least tentatively) 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 $P(A)$ $\forall A$.)

"I.i.d.":

Identity: We treat systematic differences as irrelevant.

Independence: We treat potential dependencies as irrelevant.

"We think (at least tentatively) 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 $P(A)$ $\forall A$.)

"I.i.d.":

Identity: We treat systematic differences as irrelevant.

Independence: We treat potential dependencies as irrelevant.

Of course need to discuss these for situation of interest.

Detour on epistemic probability (used by Bayesians):
"(Frequentist) probability does not exist" (de Finetti) - model subjective (or "objective") epistemic uncertainty instead.

But still same separation between "model world"
and "real epistemic uncertainty"
- no "solution" of "all models are wrong".

If we're interested in reality,
why not model reality directly,
rather than our thinking about reality?

2. What happens if assumptions are violated?

What does it mean
that a method requires model assumptions?

2. What happens if assumptions are violated?

What does it mean
that a method requires model assumptions?

It means there's a result stating that method will perform well or even optimal if model assumptions are fulfilled.

2. What happens if assumptions are violated?

What does it mean
that a method requires model assumptions?

It means there's a result stating that method will perform well or even optimal if model assumptions are fulfilled.

Benefit of model is that it inspires methods.
This doesn't mean we have to believe it's true.

2. What happens if assumptions are violated?

What does it mean
that a method requires model assumptions?

It means there's a result stating that method will perform well or even optimal if model assumptions are fulfilled.

Benefit of model is that it inspires methods.
This doesn't mean we have to believe it's true.

It doesn't say anything about what happens
if model assumptions are not fulfilled.
(In fact method may still do well.)

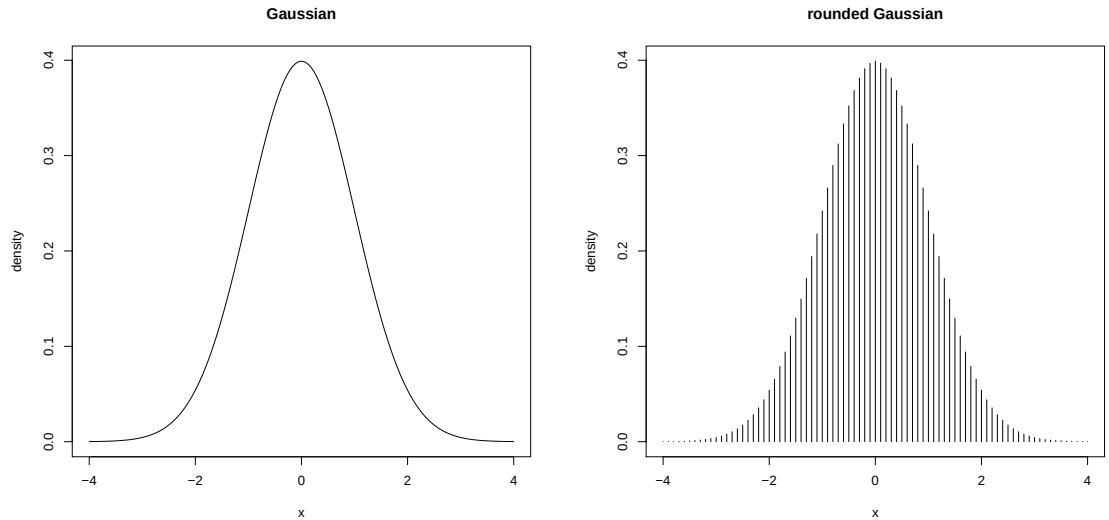
How can we know what happens if assumptions are violated?

We need to *model* violated model assumptions,
then theory or simulations.

Some examples:

Assume $X_1, \ldots, X_n$ i.i.d. with $\mathcal{L}(X_1) = \mathcal{N}(\mu, \sigma^2)$, $\sigma^2 = 1$, $n = 50$, test $H_0 : \mu = 0$ against $H_1 : \mu > 0$, more precisely $\mu = 0.5$ at $\alpha = 0.05$.

(a) Rounded Gaussian - as above but data rounded to full 0.1 (very realistic, but no continuous likelihood!)



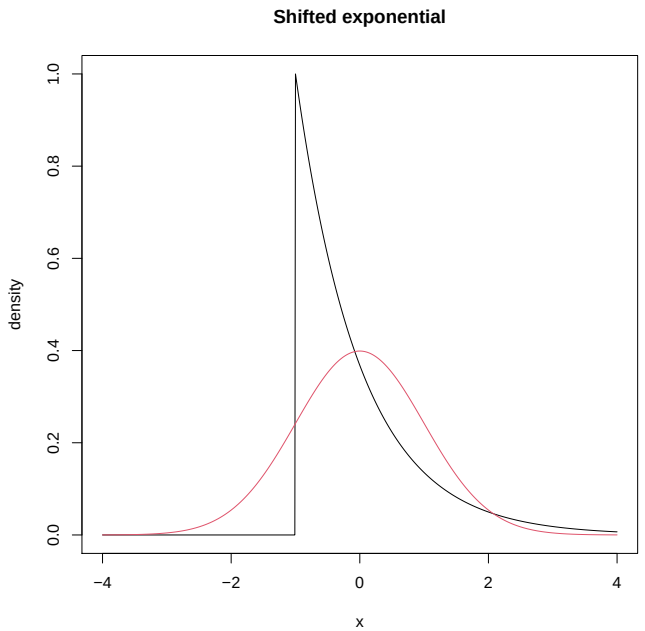
Performance of t-test of $H_0: \mu = 0$

Distribution	effective level	power
Gaussian	0.05	0.93
rounded Gaussian	0.05	0.94

Some examples:

$X_1, \ldots, X_{50}$ i.i.d. with $\mathcal{L}(X_1) = \mathcal{N}(\mu, 1)$,
test $H_0 : \mu = 0$ against $H_1 : \mu = 0.5$.

(b) (Shifted) exponential



Performance of t-test of $H_0 : \mu = 0$

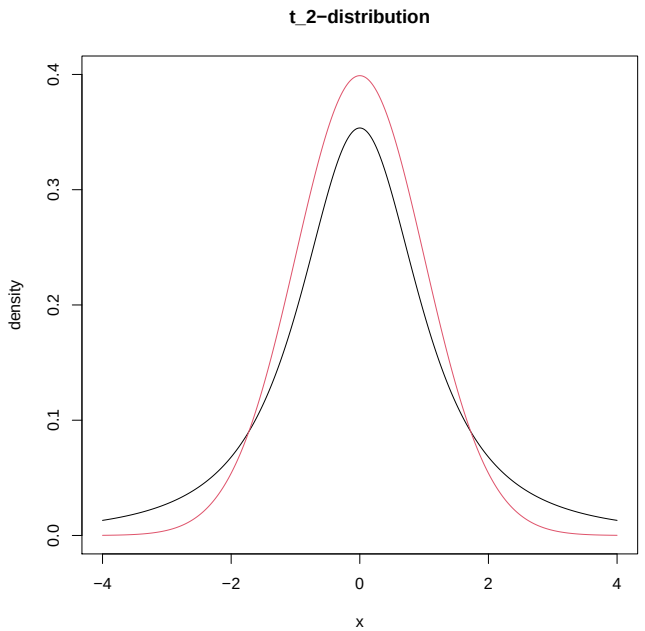
Distribution	effective level	power
Gaussian	0.05	0.93
rounded Gaussian	0.05	0.94
exponential	0.06	1

Central limit theorem:
For large n, as long as variances exist,
t-test for the mean under non-normality
behaves approximately as under normality.

More examples:

$X_1, \ldots, X_{50}$ i.i.d. with $\mathcal{L}(X_1) = \mathcal{N}(\mu, 1)$,
test $H_0 : \ \mu = 0$ against $H_1 : \ \mu = 0.5$.

(c) t_2 (non-existing variance, CLT doesn't hold)



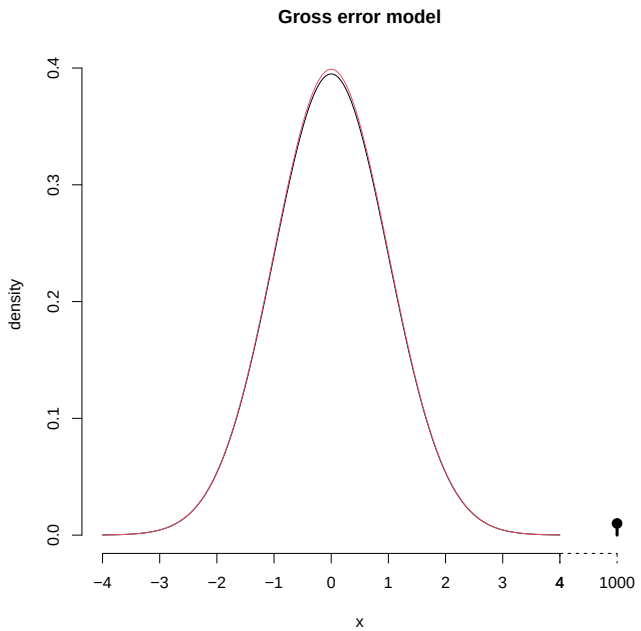
Performance of t-test of $H_0: \mu = 0$

Distribution	effective level	power
Gaussian	0.05	0.93
rounded Gaussian	0.05	0.94
exponential	0.06	1
t_2	0.04	0.39

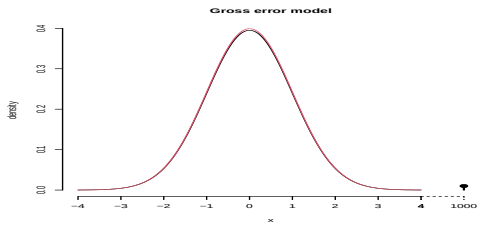
More examples:

$X_1, \ldots, X_{50}$ i.i.d. with $\mathcal{L}(X_1) = \mathcal{N}(\mu, 1)$,
test $H_0 : \mu = 0$ against $H_1 : \mu = 0.5$.

(d) Gross error model

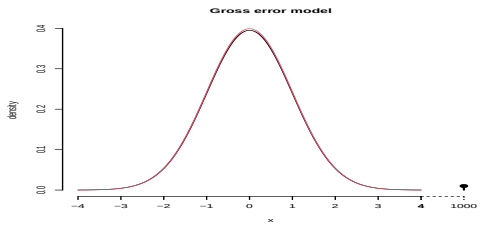


(d) Gross error model



Here $\mu = 0$ with prob. 0.99, but $E_P X = 10$!
Does this belong to H_0 or H_1 (compute level or power)?

(d) Gross error model



Here $\mu = 0$ with prob. 0.99, but $E_P X = 10$!
Does this belong to H_0 or H_1 (compute level or power)?

General issue: μ is defined within nominal model.
If model violated,
it's matter of interpretation how to "translate" H_0 and H_1.

(In fact also relevant for exponential;
do we want expected value, median, mode$= 0$?)

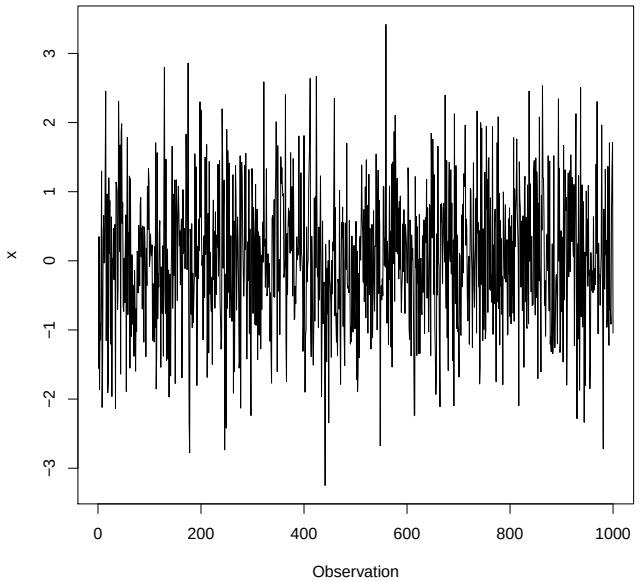
Performance of t-test of $H_0 : \mu = 0$

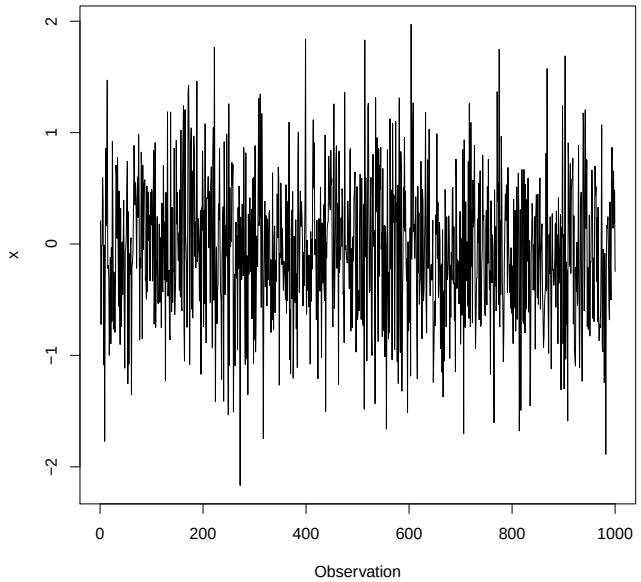
Distribution	effective level	power
Gaussian	0.05	0.93
rounded Gaussian	0.05	0.94
exponential	0.06	1
t_2	0.04	0.39
gross error ($E_P X = 10$)	0.03	0.56
gross error ($E_P X = 0$)	0.60	0.56

More examples:

$X_1, \ldots, X_{50}$ i.i.d. with $\mathcal{L}(X_1) = \mathcal{N}(\mu, 1)$,
test $H_0 : \mu = 0$ against $H_1 : \mu = 0.5$.

(e) Constant correlation. $X_1, \ldots, X_n$ marginally as above,
$\rho(X_i, X_j) = 0.1 \ \forall i, j$.





Performance of t-test of $H_0 : \mu = 0$

Distribution	effective level	power
Gaussian	0.05	0.93
rounded Gaussian	0.05	0.94
exponential	0.06	1
t_2	0.04	0.39
gross error	0.03	0.56
gross error ($E_P X = 0$)	0.60	0.56
correlated Gaussian	0.44	0.86

Some of these are dangerous, some are harmless.

3. Nominal and substantial hypotheses

Inference target parameter is defined in “model world”;
but we’re interested in real world.

E.g., test H_0, assuming normality of measurements,
“Water turbidity in river X at place Y is not larger than 25.”

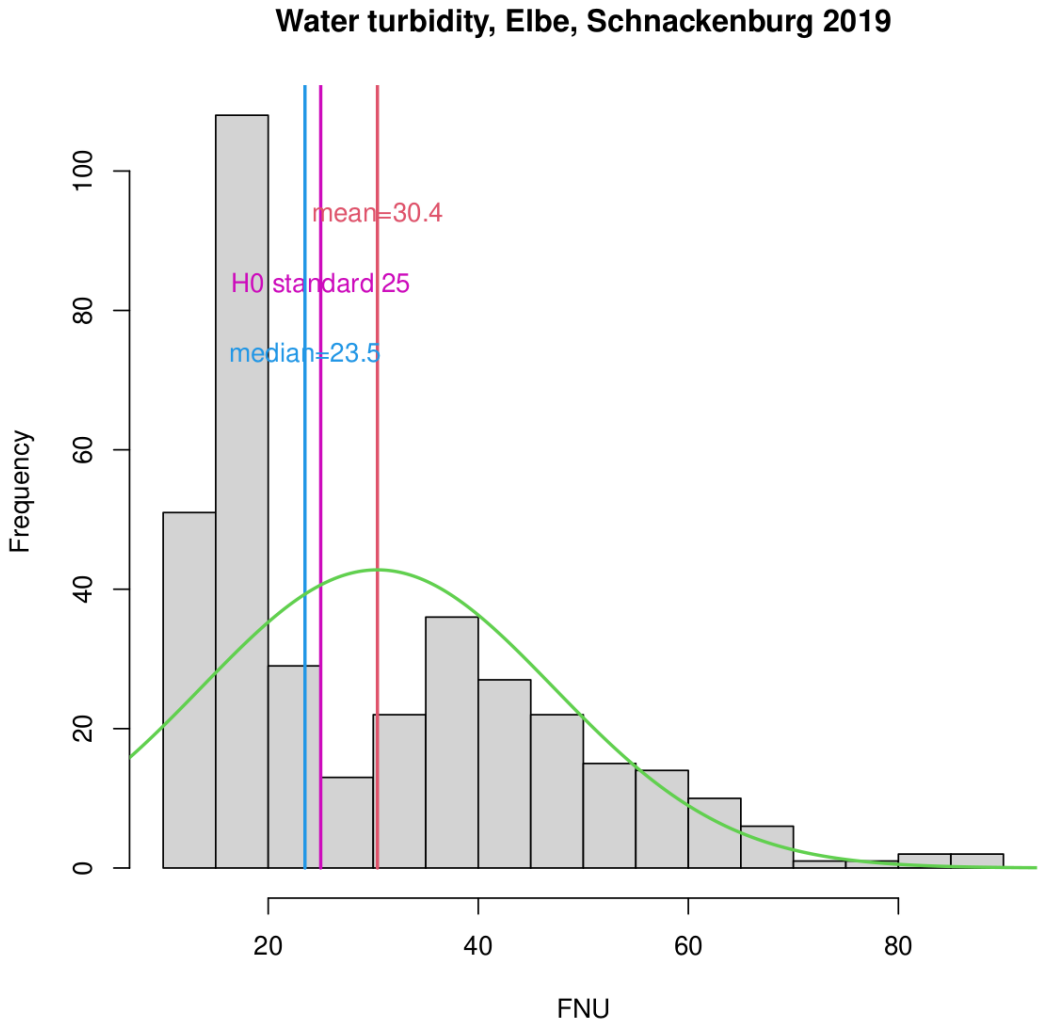
3. Nominal and substantial hypotheses

Inference target parameter is defined in "model world";
but we're interested in real world.

E.g., test H_0, assuming normality of measurements,
"Water turbidity in river X at place Y is not larger than 25."

If underlying distribution isn't the nominal one,
does it belong to *substantial* H_0, to H_1, or neither?

("*Should* we reject?" $\Rightarrow$ definition of "misleading")


Water turbidity, Elbe, Schnackenburg 2019
mean=30.4
H0 standard 25
median=23.5
Frequency
0
20
40
60
80
100
20
40
60
80
FNU

"Turbidity in river X at place Y not larger than 25."
Issues with "translation into model world":

- Assume unobserved "true" turbidity, implicitly defined by model.
- How to aggregate measurement distribution? (Skewness? Median? Mean?)
- Are turbidity peaks/outliers essentially important or to be ignored?

"Turbidity in river X at place Y not larger than 25."
Issues with "translation into model world":

- Assume unobserved "true" turbidity, implicitly defined by model.
- How to aggregate measurement distribution? (Skewness? Median? Mean?)
- Are turbidity peaks/outliers essentially important or to be ignored?

E.g. gross error model $0.99\mathcal{N}(25, 1) + 0.01\delta_{1025}$:
"Substantial μ"$= 25$ (H_0; of Gaussian) or $= 35$ (H_1; E-value)?

This needs judgement - data cannot decide this!

"Turbidity in river X at place Y not larger than 25."
Issues with "translation into model world":

- Assume unobserved "true" turbidity, implicitly defined by model.
- How to aggregate measurement distribution? (Skewness? Median? Mean?)
- Are turbidity peaks/outliers essentially important or to be ignored?

E.g. gross error model $0.99\mathcal{N}(25, 1) + 0.01\delta_{1025}$:
"Substantial μ"$= 25$ (H_0; of Gaussian) or $= 35$ (H_1; E-value)?

This needs judgement - data cannot decide this!

CLT holds for gross error model, but this doesn't help if E-value doesn't reflect substantial hypothesis!

Target of inference in much robust and nonparametric work:
functional of true distribution, e.g., $E_P(X)$;
Robustness problems do not only arise from estimators/statistics,
but also from the functionals they're estimating!

Nonparametric procedures "work" (e.g., be consistent)
for their inference target with weak assumptions,
but need think about whether inference target behaves
in line with "substantial (real) target".

Baseline:
When investigating procedures under non-nominal models, need specify how parameters of non-nominal models relate to nominal model in terms of interpretation.

Only then can we know whether procedure
"does the right thing" under violated assumptions.

What does the test actually do?

t-test with $T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$,
rejecting H_0 for $|T| > c_\alpha$
can be interpreted as testing *general nonparametric*
H_0 : P is such that $P\{|T| > c_\alpha\} \leq \alpha$ against
H_1 : P is such that $P\{|T| > c_\alpha\} > \alpha$
For this, the test is *unbiased by definition*.

What does the test actually do?

t-test with $T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$,
rejecting H_0 for $|T| > c_\alpha$
can be interpreted as testing *general nonparametric*
H_0 : P is such that $P\{|T| > c_\alpha\} \leq \alpha$ against
H_1 : P is such that $P\{|T| > c_\alpha\} > \alpha$
For this, the test is *unbiased by definition*.

The key issue then is:
Does definition of T indicate
the desired *direction* of deviation from the substantial H_0?

What does the test actually do?

t-test with $T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$,
rejecting H_0 for $|T| > c_\alpha$
can be interpreted as testing *general nonparametric*
H_0 : P is such that $P\{|T| > c_\alpha\} \leq \alpha$ against
H_1 : P is such that $P\{|T| > c_\alpha\} > \alpha$
For this, the test is *unbiased by definition*.

The key issue then is:
Does definition of T indicate
the desired *direction* of deviation from the substantial H_0?
Rather than "are the assumptions fulfilled"? (Which they aren't.)

This amounts to understanding whether $T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$ as aggregation of the information in the data is "substantially correct".

Need to understand properties of $\bar{X}_n$ and S_n such as breakdown under gross outliers,
$\bar{X}_n$ as distributing sum equally among observations.
(In given application appropriate for skew distributions?)

This amounts to understanding whether $T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$ as aggregation of the information in the data is "substantially correct".

Need to understand properties of $\bar{X}_n$ and S_n such as breakdown under gross outliers,
$\bar{X}_n$ as distributing sum equally among observations.
(In given application appropriate for skew distributions?)

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.

With this interpretation, it is *not* true that
"the P value tests all the assumptions about how the data were generated, not just the targeted hypothesis it is supposed to test".

With this interpretation, it is *not* true that
"the P value tests all the assumptions about how the data were generated, not just the targeted hypothesis it is supposed to test".

It doesn't automatically test the substantial hypothesis,
but in fact it tests
whether T is where it is expected to be under the H_0

With this interpretation, it is *not* true that
"the P value tests all the assumptions about how the data were generated, not just the targeted hypothesis it is supposed to test".

It doesn't automatically test the substantial hypothesis,
but in fact it tests
whether T is where it is expected to be under the H_0
(...and under many other distributions,
hopefully mostly formalising the substantial H_0).

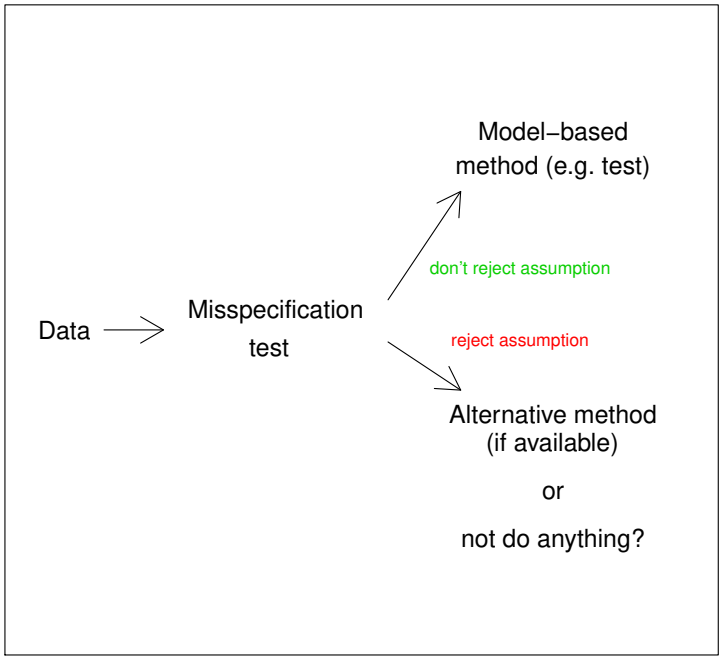
4. What can we do about the model assumptions?

Standard approaches:

- Misspecification testing
- Informal (visual) diagnosis
- “Translate” information about reality into model world, e.g., time dependence of water turbidity

Misspecification testing:

H_0 : Assumption holds, H_1 : Asumption violated.



Fisher (1922): *"For empirical as the specification of the hypothetical population may be, this empiricism is cleared of its dangers if we can apply a rigorous and objective test of its adequacy."*

Cox & Mayo (2006): *"An important part of frequentist theory is its ability to check model assumptions."*

Kass et al. (2016): *"Rule 8: Check your assumptions."*

Spanos (2018): *"The typicality of (observations)* $\mathbf{z}_0$ *(for the proposed model) can - and should - be assessed using trenchant misspecification testing."*

Example: Shapiro-Wilk test for normality:

Distribution	eff. level	power	S-W detection prob.
Gaussian	0.05	0.93	0.05
rounded Gaussian	0.05	0.94	0.05
exponential	0.06	1	0.99
t_2	0.04	0.39	0.86
gross error	0.03	0.56	0.42
gross error ($E_P X = 0$)	0.60	0.56	0.42
correlated Gaussian	0.44	0.86	(0.05)

Example: Shapiro-Wilk test for normality:

Distribution	eff. level	power	S-W detection prob.
Gaussian	0.05	0.93	0.05
rounded Gaussian	0.05	0.94	0.05
exponential	0.06	1	0.99
t_2	0.04	0.39	0.86
gross error	0.03	0.56	0.42
gross error ($E_P X = 0$)	0.60	0.56	0.42
correlated Gaussian	0.44	0.86	(0.05)

Least normal $\neq$ most dangerous!

Example: Shapiro-Wilk test for normality:

Distribution	eff. level	power	S-W detection prob.
Gaussian	0.05	0.93	0.05
rounded Gaussian	0.05	0.94	0.05
exponential	0.06	1	0.99
t_2	0.04	0.39	0.86
gross error	0.03	0.56	0.42
gross error ($E_P X = 0$)	0.60	0.56	0.42
correlated Gaussian	0.44	0.86	(0.05)

Least normal $\neq$ most dangerous!

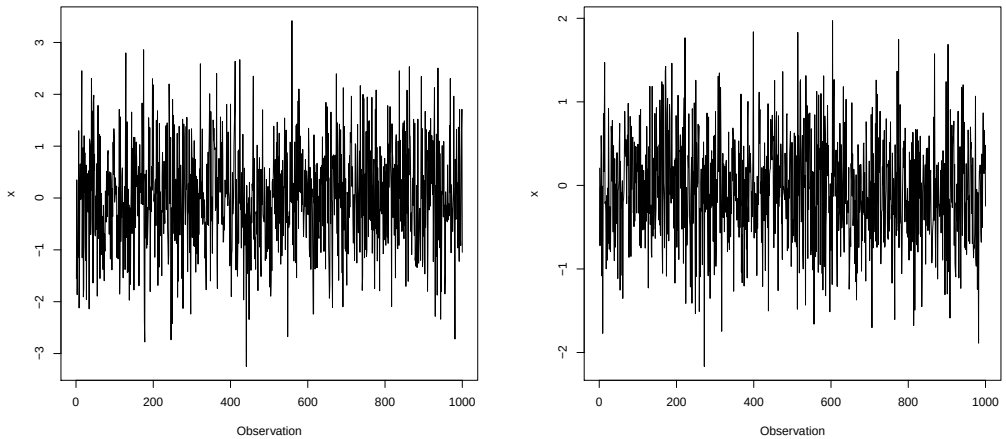
S-W test can't find rounded Gaussian.

This bug is actually a feature!

Don't *want* to find everything.

Untestable assumptions

Constant correlation. $X_1, \ldots, X_n$ marginally $\mathcal{N}(\mu, \sigma^2)$, $\rho(X_i, X_j) = 0.1 \; \forall i, j$.



This is pretty bad (see above)…
but it's indistinguishable from i.i.d.!

Why's that? Assume $X_1, \ldots, X_n$ as before with $\mathrm{Cor}(X_i, X_j) = \rho$.

Lemma 1 (H, 2021): For $Y_1, \ldots, Y_n$ iid, $\mathcal{L}(Y_1) = \mathcal{N}(\mu, (1-\rho)\sigma^2)$:

$$\mathcal{L}(X_1, \ldots, X_n | \bar{X}_n) = \mathcal{L}(Y_1, \ldots, Y_n | \bar{Y}_n).$$

Proof: Elementary calculations on conditional multivariate normals.

Why's that? Assume $X_1, \ldots, X_n$ as before with $\mathrm{Cor}(X_i, X_j) = \rho$.

Lemma 1 (H, 2021): For $Y_1, \ldots, Y_n$ iid, $\mathcal{L}(Y_1) = \mathcal{N}(\mu, (1-\rho)\sigma^2)$:

$$\mathcal{L}(X_1, \ldots, X_n | \bar{X}_n) = \mathcal{L}(Y_1, \ldots, Y_n | \bar{Y}_n).$$

Proof: Elementary calculations on conditional multivariate normals.

Theorem: No set A, $0 \leq \beta \leq 1$, sample size n exist so that $P_n(A) \leq \beta$ for P_n with $\rho = 0$ (i.i.d.), all μ, σ^2 and $Q_n(A) > \beta$ for Q_n with fixed $\rho > 0$, all μ, σ^2.

$\Rightarrow \rho = 0$ is indistinguishable from $\rho > 0$ from any finite data set.

Generally, can only test dependence
assuming regularly repeated dependence pattern
(such as in time series, within random effect levels).

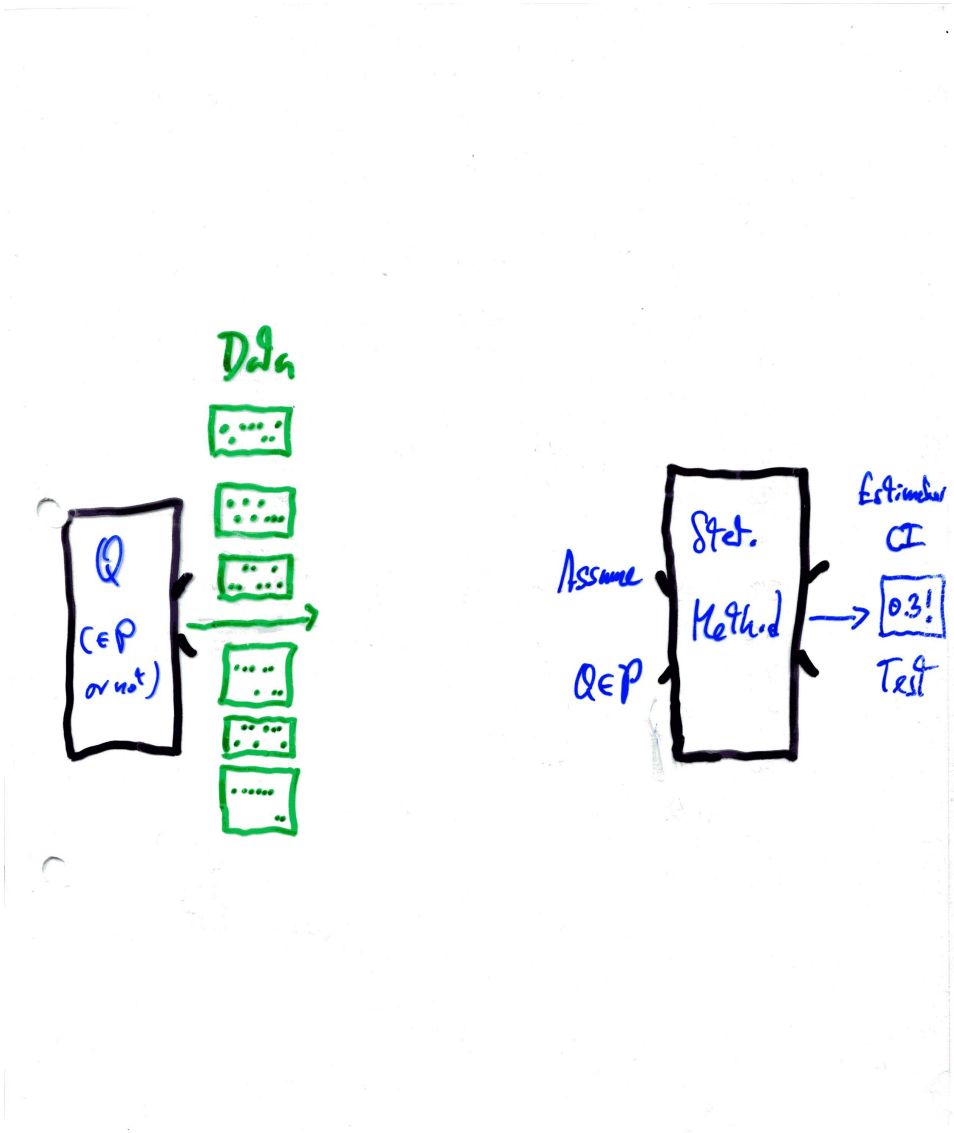
Dependence can only be found
if we can specify how observation order is informative for it.

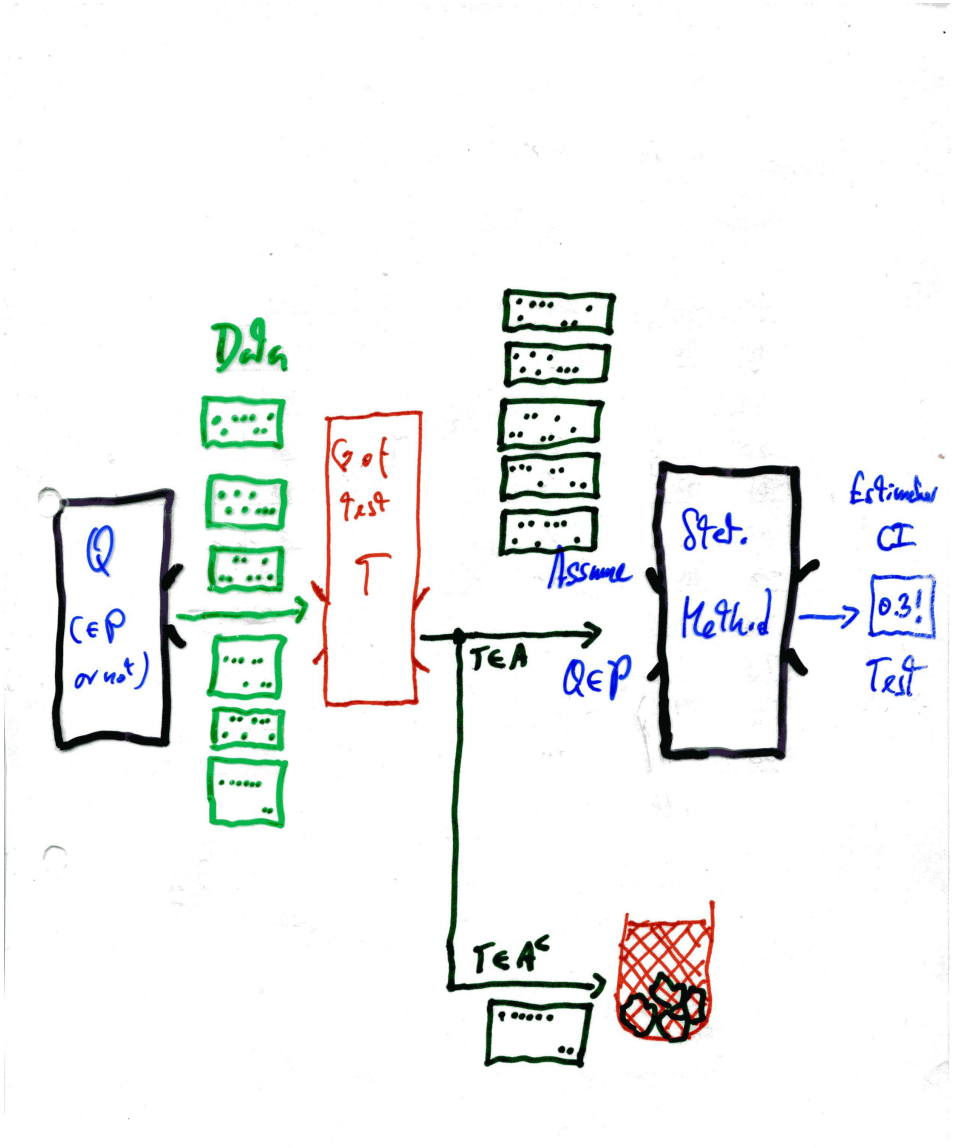
Other dependence patterns
can only be excluded by assumption.
The best we can do is to think very hard about the situation.

Further issue with misspecification testing:
The misspecification (goodness-of-fit) paradox
(H, 2007)

Checking the model assumptions violates them automatically!

(Known in literature for long,
e.g., Bancroft 1944, Chatfield 1995)

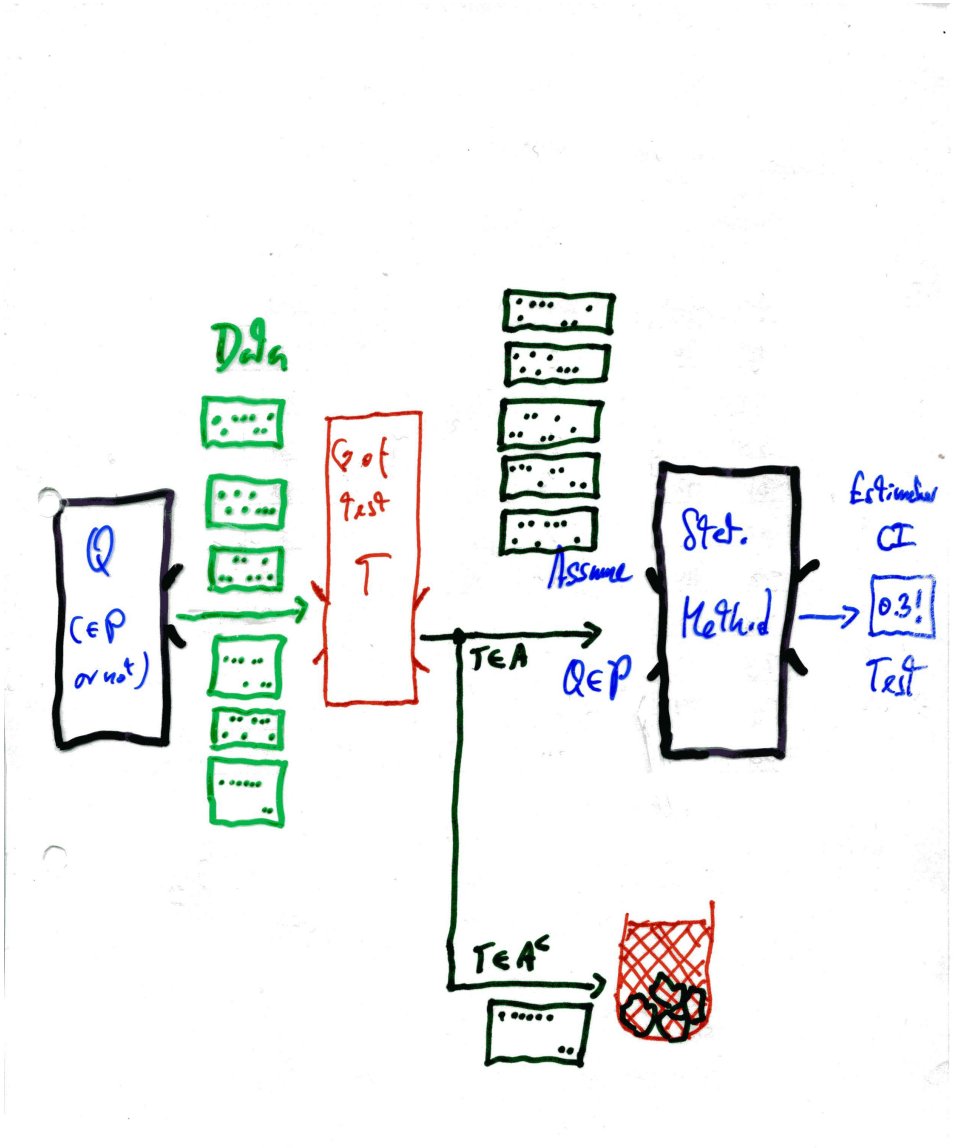

Data
Q
(∈P or not)
Assume
Q∈P
Stat.
Method
Estimator
CI
0.3!
Test


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

But is this a problem?
A. Spanos (2018): *"No, we learn that model is valid for data. (MS test and main test) "pose very different questions to data"*.
MS test tests whether data *"constitutes truly typical realization of mechanism described by model"*.

But is this a problem?
A. Spanos (2018): *"No, we learn that model is valid for data. (MS test and main test) "pose very different questions to data"*.
MS test tests whether data *"constitutes truly typical realization of mechanism described by model"*.

In fact, if MS test and main test are *independent*, misspecification paradox does not affect distribution of main test statistic.
(E.g., Gaussian linear regression model checking based on residuals.)


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

But independence is often not fulfilled.

Statistics literature from Bancroft (1944) investigates distribution of result
conditionally on not rejecting assumption.

E.g., will test level be kept, power decline?
Also, does MS testing help if model is violated?

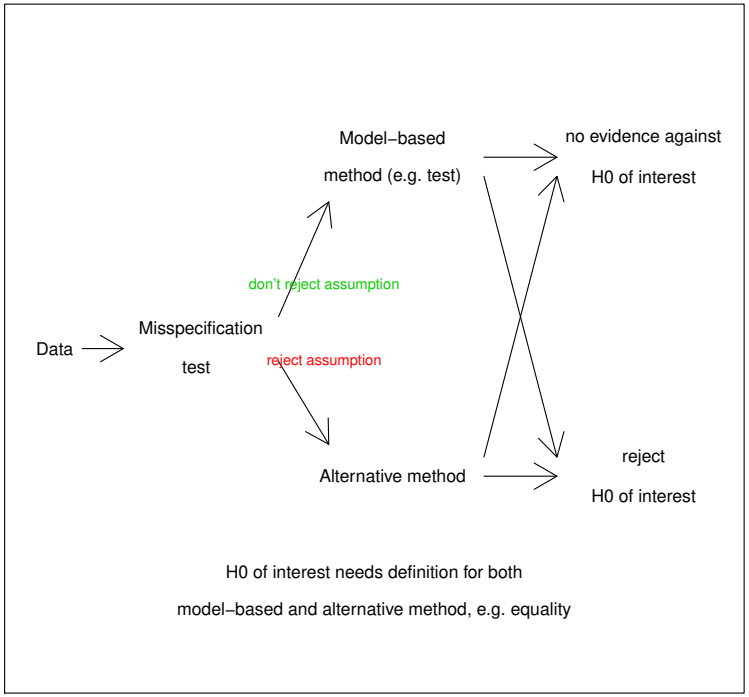
But independence is often not fulfilled.

Statistics literature from Bancroft (1944) investigates distribution of result
conditionally on not rejecting assumption.

E.g., will test level be kept, power decline?
Also, does MS testing help if model is violated?

Again: model violation of assumption, and what is done, and see what happens.

5. Combined procedures



Analyse under nominal model and violated assumptions what these procedures deliver.

Some results

Authors who investigated specific combined procedures:

Easterling and Anderson (1978): *"The results given here (. . .) are not supportive of the notion that preliminary testing is the proper thing to do."*

Freeman (1989): *"In the light of the results in this paper, the two-stage analysis is so unsatisfactory as to be ruled out of future use."*

Moser and Stevens (1992): *"Is the current practice of preliminary variance tests appropriate? The answer is no."*

Fay and Proschan (2010): *"The choice between t- and Wilcoxon-Mann-Whitney should not be based on a test of normality."*

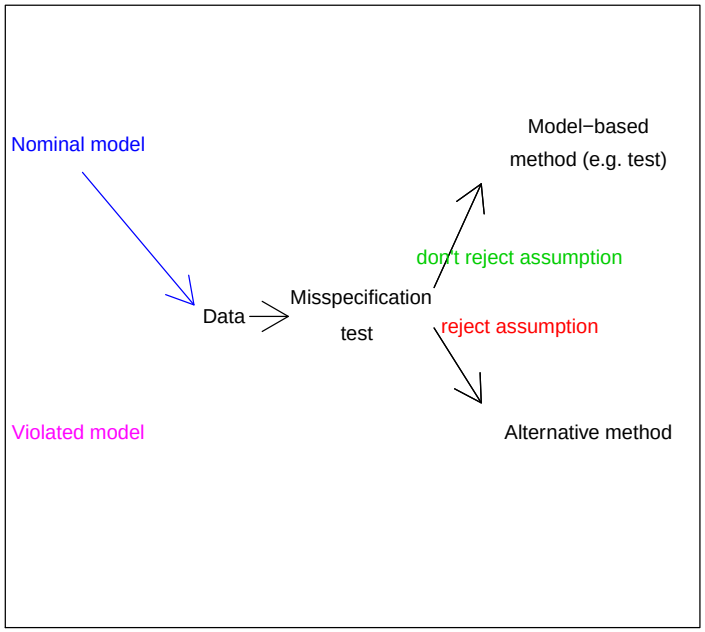
Rochon, Gondan and Kieser (2012): *"From a formal perspective, preliminary testing for normality is incorrect and should therefore be avoided."*

Overall disturbing, given
preference for assumption checking in general literature.

. . . but at least King and Giles (1984): *"We find that overall, pre-testing is preferable to pure OLS regression techniques and generally compares favourably with the strategy of always correcting for possible autocorrelation."*

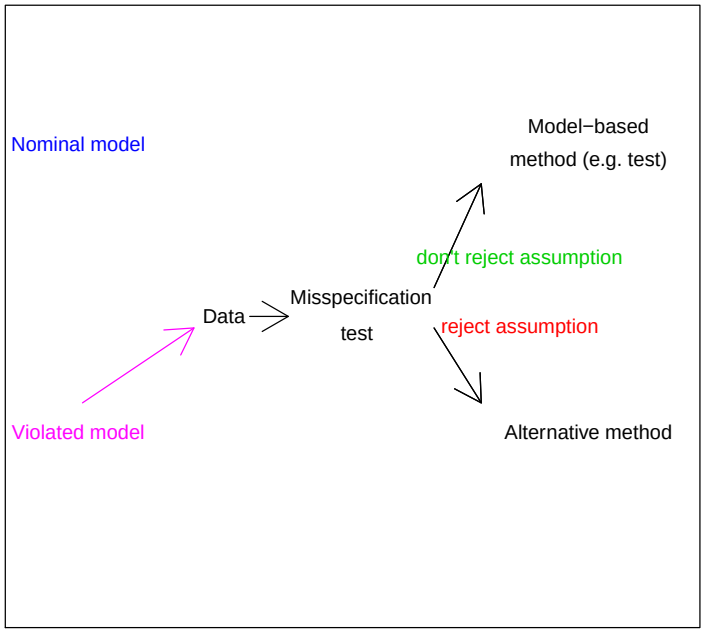
"Mixed" setups

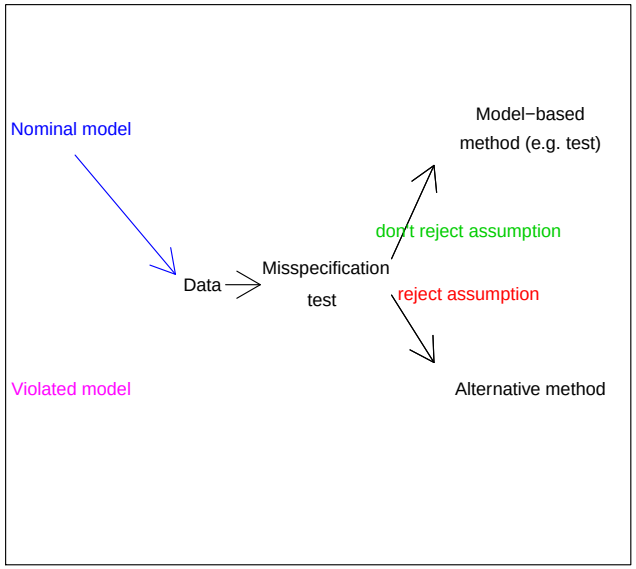
Literature looks at either fulfilled or violated assumptions



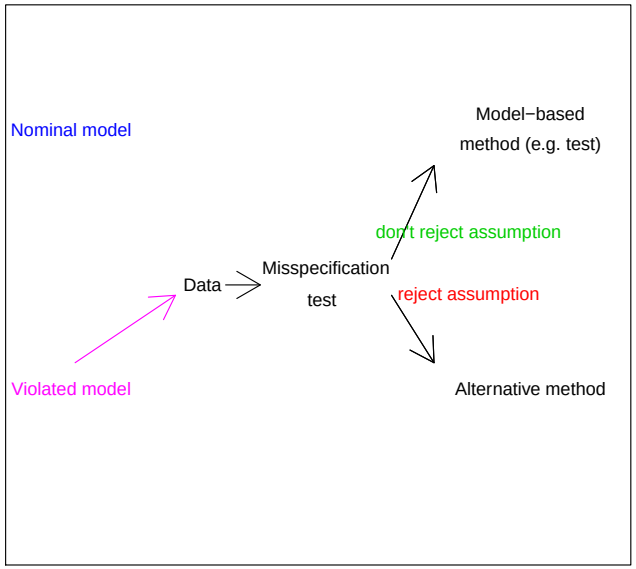
"Mixed" setups

Literature looks at either fulfilled or violated assumptions

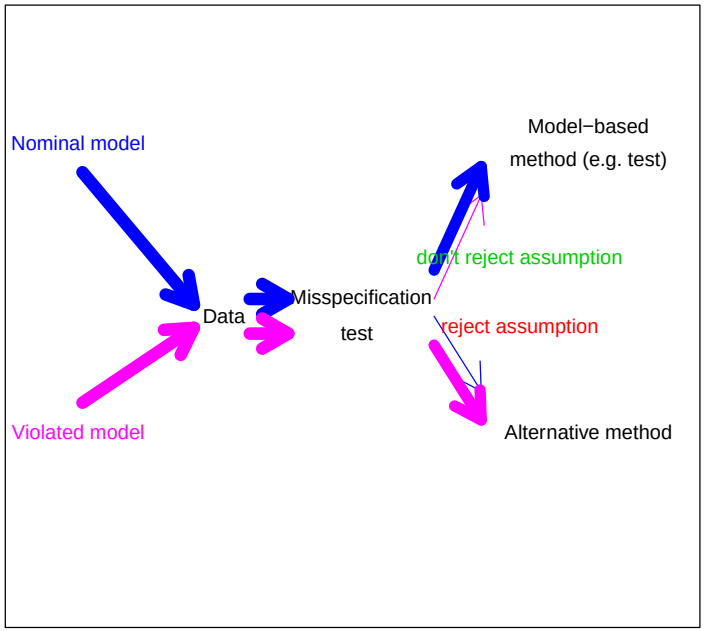




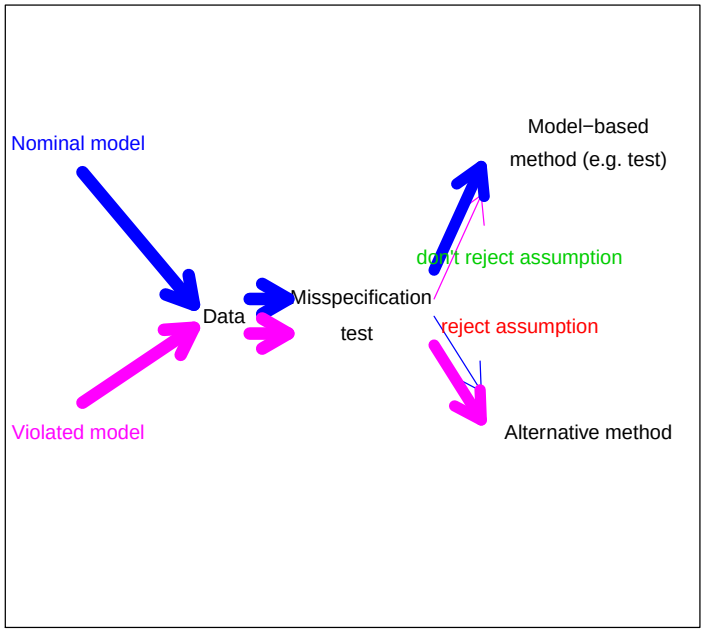
MS testing may not help for nominal model. . .



. . . may not help if assumptions violated. . .



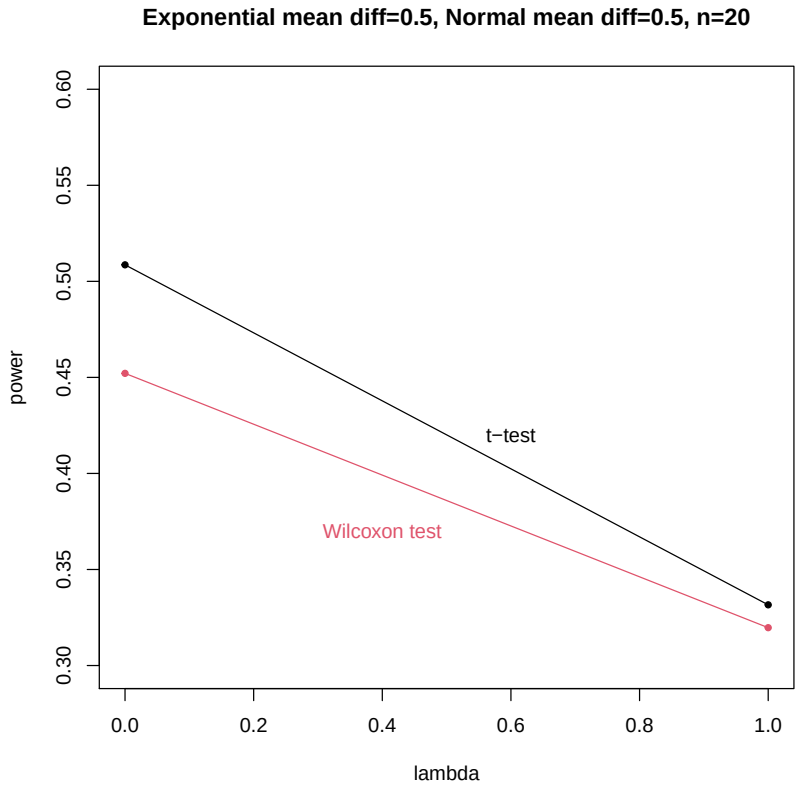
. . . but can help if both are mixed.



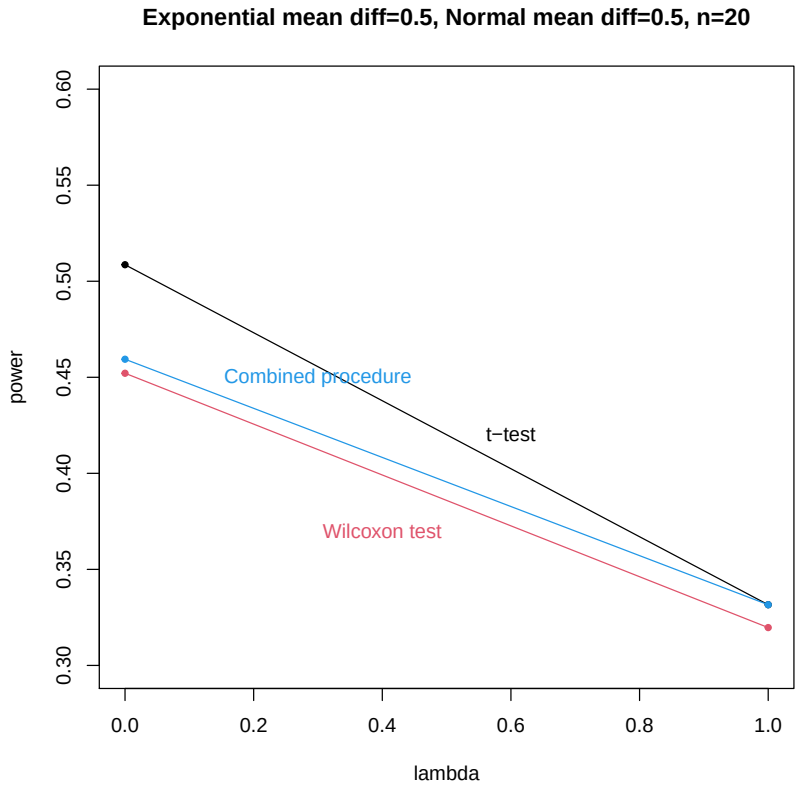
Looking at nominal or violated model *in isolation* will hide ability of MS test to make a difference.

PhD thesis of Iqbal Shamsudheen:
Look at **"mixed" setups**
in which with probability $\lambda \in [0, 1]$
model assumption fulfilled or not (for whole data set).

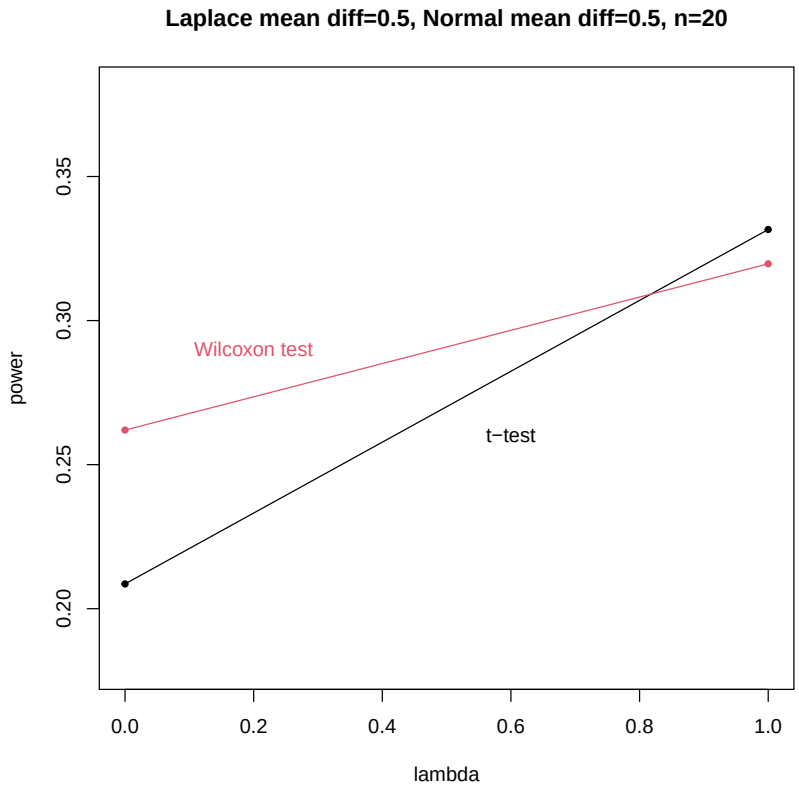
(Two two-sample test examples,
look at power only here;
type I error probability also relevant
but level not significantly violated
by any procedure in these examples.)

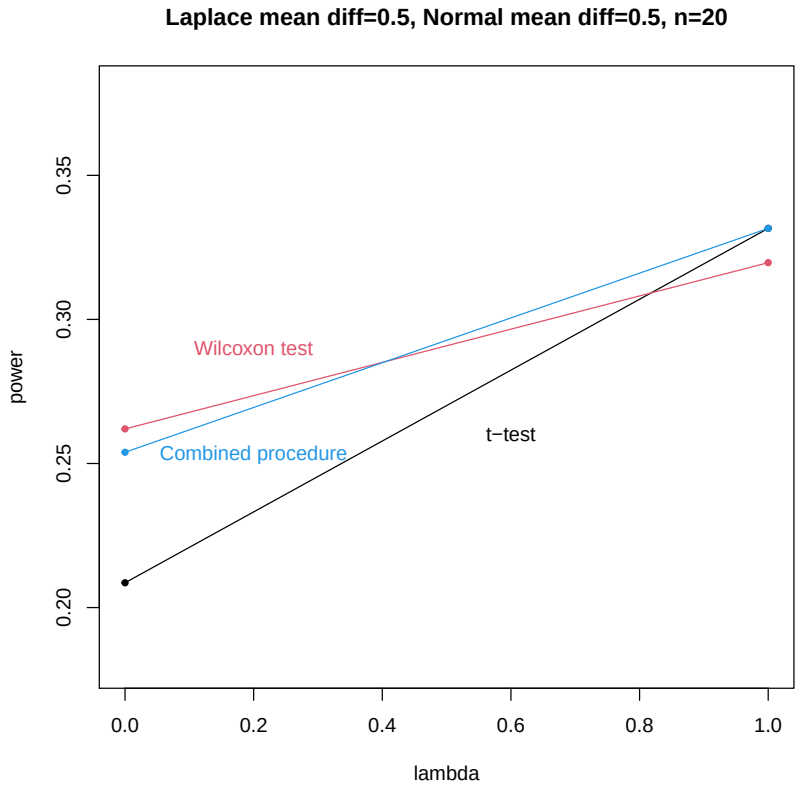


Setup from Rochon et al. (2012) -
note that t-test is more superior for exp than for normal.



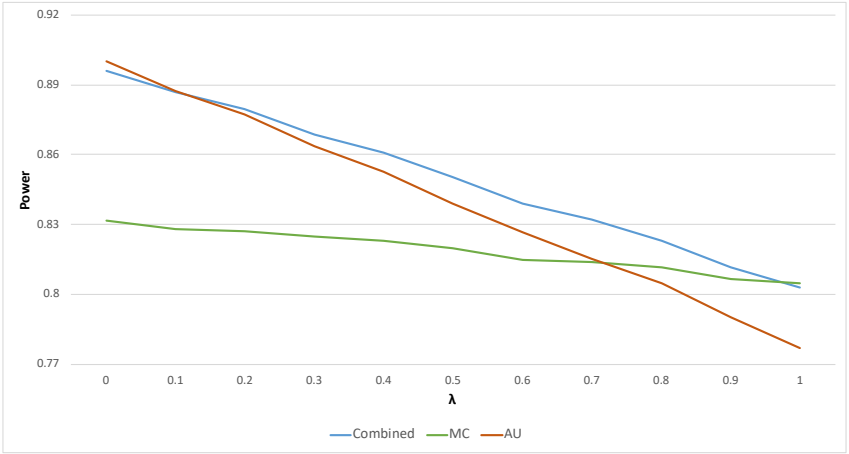
. . . and combined procedure is quite competitive under normal.


Laplace mean diff=0.5, Normal mean diff=0.5, n=20
power
0.20
0.25
0.30
0.35
Wilcoxon test
t−test
0.0
0.2
0.4
0.6
0.8
1.0
lambda

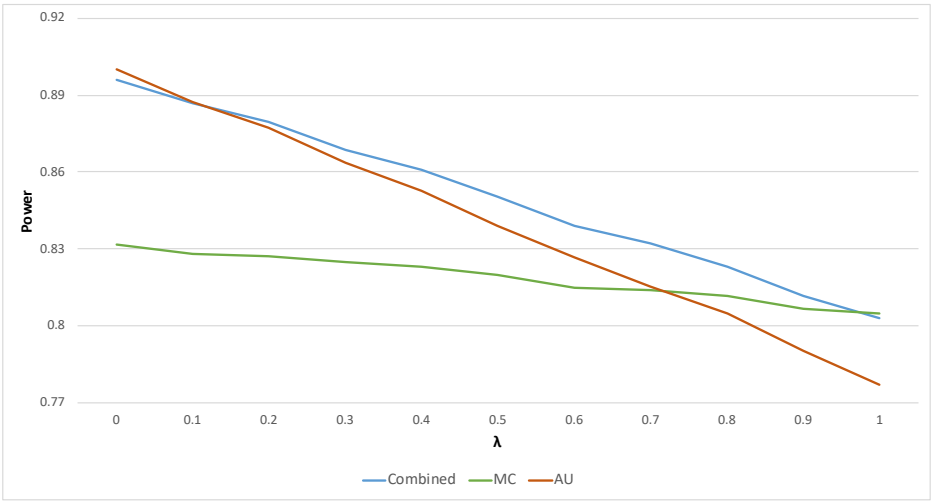


. . . but combined procedure can better them both for much of λ-range.

Many follow this pattern:



A general theoretical result



Lemma 2, Shamsudheen & H. (2020):

Look at probability λ for fulfilled assumptions P,
otherwise violated assumptions Q.

Assume Φ_{MS} approx. independent of both Φ_{MC} and Φ_{AU}.

Assume MS test “better than useless”.

Assume model-based method has higher power under P,
alternative higher power under Q.

Lemma 2, Shamsudheen & H. (2020):

Look at probability λ for fulfilled assumptions P, otherwise violated assumptions Q.

Assume Φ_{MS} approx. independent of both Φ_{MC} and Φ_{AU}.

Assume MS test "better than useless".

Assume model-based method has higher power under P, alternative higher power under Q.

Then combined procedure has higher power than both Φ_{MC} and Φ_{AU} for $[\lambda_1, \lambda_2]$, $0 < \lambda_1 < \lambda_2 < 1$.

Are MS testing/combined procedures advisable?

No, if model-based test is robust (good overall).

No, if alternative test is good also under nominal model.

No, if good robust/alternative approaches are preferred.

Are MS testing/combined procedures advisable?

No, if model-based test is robust (good overall).

No, if alternative test is good also under nominal model.

No, if good robust/alternative approaches are preferred.

Yes, if MS test is sensitive to violations that matter,

and MS test is approximately independent of main tests,

and main tests have “complementary qualities”,

and both close-to-nominal and violated assumptions seem realistic.

Details matter!

Major issue with current MS testing:

Focus on testing whether model assumptions hold -
but focus should be to distinguish
problematic from unproblematic violations!

Much research potential!

Discussion

More than one assumption needs checking.

More complicated combined procedures,
analyse easier cases first.

Discussion

More than one assumption needs checking.

More complicated combined procedures,
analyse easier cases first.

Is visual assumption checking better?

It may be, in the hands of good data analyst,
but it may also be worse, and
it cannot be analysed by theory or simulation!

Key take-aways

- Much communication about model assumptions is misleading.

Key take-aways

- Much communication about model assumptions is misleading.
- The issue is *not* whether assumptions are fulfilled, but rather whether they are violated in ways that mislead about substantial hypothesis.

Key take-aways

- Much communication about model assumptions is misleading.
- The issue is *not* whether assumptions are fulfilled, but rather whether they are violated in ways that mislead about substantial hypothesis.
- Whether assumption checking helps depends on many details.

Key take-aways

- Much communication about model assumptions is misleading.
- The issue is *not* whether assumptions are fulfilled, but rather whether they are violated in ways that mislead about substantial hypothesis.
- Whether assumption checking helps depends on many details.
- Some key assumptions cannot be checked against data.

Key take-aways

- Much communication about model assumptions is misleading.
- The issue is *not* whether assumptions are fulfilled, but rather whether they are violated in ways that mislead about substantial hypothesis.
- Whether assumption checking helps depends on many details.
- Some key assumptions cannot be checked against data.
- Judgment and interpretation are always involved.

Key take-aways

- Much communication about model assumptions is misleading.
- The issue is *not* whether assumptions are fulfilled, but rather whether they are violated in ways that mislead about substantial hypothesis.
- Whether assumption checking helps depends on many details.
- Some key assumptions cannot be checked against data.
- Judgment and interpretation are always involved.
- None of these issues is solved by Bayesian statistics.

References:

Bancroft, T. A. (1944) On biases in estimation due to the use of preliminary tests of significance. *Annals of Mathematical Statistics* 15(2), 190-204.

Chatfield, C. (1995) Model Uncertainty, Data Mining and Statistical Inference (with discussion). *Journal of the Royal Statistical Society, Series B* 158(3), 419-466.

Cox, D. R. (2006) *Principles of Statistical Inference*. Cambridge University Press.

Easterling, R. G., & Anderson, H. E. (1978) The effect of preliminary normality goodness of fit tests on subsequent inference, *Journal of Statistical Computation and Simulation* 8(1), 1-11.

Fay, M. P. and Proschan M. A. (2010) Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys* 4, 1-39.

Fisher, R. A. (1922) On the Mathematical Foundation of Theoretical Statistics, Philosophical *Transactions of the Royal Society of London A* 222, 309-368.

Freeman, P. (1989) The performance of the two-stage analysis of two-treatment, two-period cross-over trials. *Statistics in Medicine* 8, 1421-1432.

Greenland, S., Senn, S.J., Rothman, K.J. et al. (2016) Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiolgy* 31, 337-350.

Hennig, C. (2007) Falsification of propensity models by statistical tests and the goodness-of-fit paradox. *Philosophia Mathematica* 15(2), 166-192.

Hennig, C. (2020) Frequentism-as-model. *arXiv:2007.05748*

Hennig, C. (2021) Parameters not identifiable or distinguishable from data, including correlation between Gaussian observations. *arXiv:2108.09227*

Kass, R. E., Caffo, B. S., Davidian, M., Meng, X. L., Yu, B., & Reid, N. (2016) Ten simple rules for effective statistical practice. *PLoS Computational Biology* 12(6), e1004961.

King, M. L. and Giles, D. E. A. (1984) Autocorrelation pre-testing in the linear model: Estimation, testing and prediction, *Journal of Econometrics* 25 (1), 35-48.

Moser, B. K., & Stevens, G. R. (1992) Homogeneity of variance in the two-sample means test. *The American Statistician* 46(1), pp. 19-21.

Rochon, J., Gondan, M., & Kieser, M. (2012) To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC Medical Research Methodology* 12(1), 81-91.

Shamsudheen, M. I. and Hennig, C. (2020) Should we test the model assumptions before running a model-based test? *arXiv:1908.02218*.

Spanos, A. (2018) Mis-specification Testing in Retrospect. *Journal of Economic Surveys* 32(2), 541-577.

Appendix on Bayesian modelling

In epistemic (Bayesian) probability modelling,
“all models are wrong” as well,
and similar issues arise.

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 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.

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

Exchangeability implies that $P\{1\}$ in 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).

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

Exchangeability implies that $P\{1\}$ in 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).

If in fact all observations are correlated,
analysis based on exchangeability
may give hugely misleading results.

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.

Models are thought constructs and
operate on domain different from observer-dependent reality,
be it frequentist models vs. real data generation,
or epistmic models vs. degree of belief.

Appendix on identifiability of Gaussian correlation

$X_1, \ldots, X_n$ id. $\sim \mathcal{N}(\mu, \sigma^2)$ with $\text{Cor}(X_i, X_j) = \rho$
defines n-variate Gaussian; ρ is identifiable.

How can it be that $\rho_1 = 0,\ \rho_2 \neq 0$ cannot be distinguished from data?

Answer: In fact they *can* be distinguished from data
if several *independent* n-vectors $(X_1, \ldots, X_n)$ are observed.

But if all observations are correlated with all others,
observe only one n-vector (effective sample size 1).

Dependence within n-vector can only be observed
as contrasted against independence between them.