Applied Mechanism Design and Big Data

→ sub-topic: Statistical Data Analysis

Max Baak

(with many thanks to Wouter Verkerke)

Roadmap for this course

1. Statistics basics

2. Parameter estimation
   - Maximum likelihood fits

3. Pitfalls in (big) data analysis
   - Spurious correlations
   - Data quality assessment

4. Hypothesis Testing
   - Hypothesis tests
   - Analysis pitfalls – reprise
   - Discussion on “proper” data analysis

Brief summary of material

Lecture 1: Advantages of modeling by hand

- Full control (over contents of the model).
  - You determine all ingredients of the model yourself.
    - E.g. What you want when underlying model is known.

- Assess the uncertainty on the fitted model
  - Can do error propagation of uncertainties on fit parameters.

- Simulation
  - Easy to test impact of parameters in model.
  - Often need for knobs that can be tuned.
    - E.g. test impact of policy changes on a population.

- Hypothesis testing
  - Likelihood fits to data have high(-est) sensitivity for classification.
  - Well-defined formulas to calculate p-values.

Lecture 2: Maximum likelihood

- The *likelihood* is the value of a probability mass/density function evaluated at the measured value of the observable(s)

$$L(\vec{p}) \quad = \quad F(\vec{x} \equiv \vec{x}_{data} \mid H(\vec{p}))$$

- *The likelihood proportional to the total probability.*

- Define the maximum likelihood (ML) estimator(s) to be the parameter value(s) for which the likelihood is maximum.

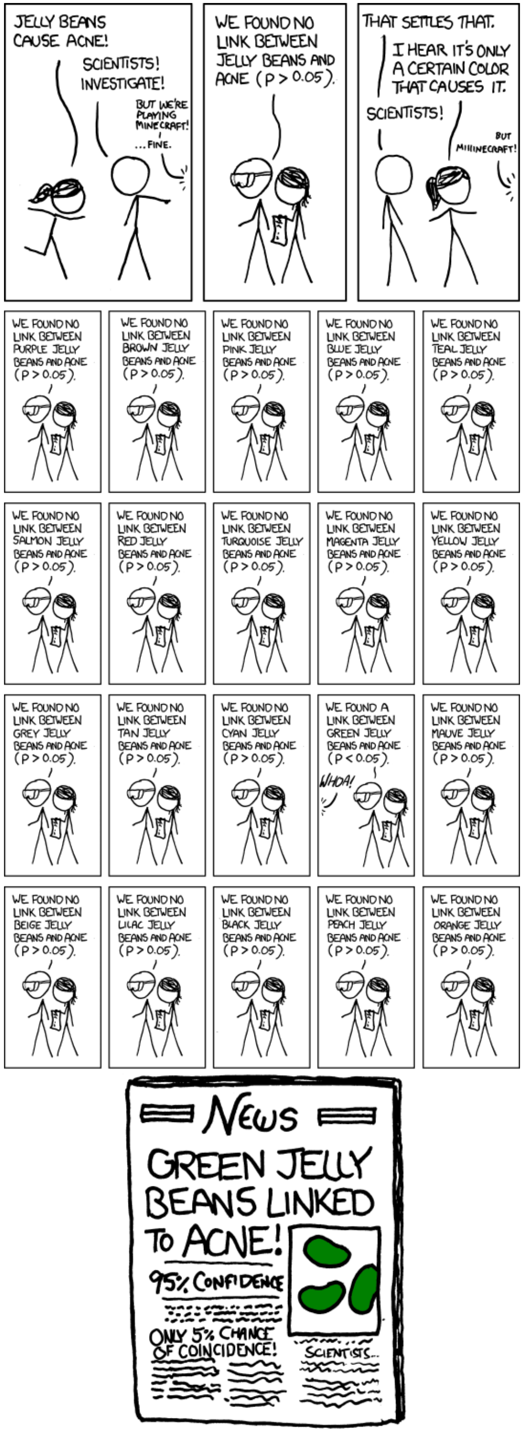
- Maximizing Likelihood is equivalent to *minimizing* the χ^2

$$\chi^2 = -2 \ln L(\vec{p})$$

Lecture 3:
Look elsewhere effect

- Also known as:
  'Multiple comparisons problem'

- Quoting the significance of of a single experiment …
- … even though many experiments have been performed …
  - (= wrong.)
- … such that the single experiment may well have been a statistical fluctuation in the whole set.


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

https://xkcd.com/882/

Hypothesis Testing

Hypothesis Testing

- An hypothesis *H* specifies the probability for the dataset *x* under a given model.
  - *P(x|H)* = the probability for *x* given model *H*
  - Also called the likelihood of the hypothesis, written as: $L(x|H)$

- The dataset *x* can represent:
  - a single number (“single observation”),
  - a single record of numbers (“event”),
  - or an entire collections of datasets (“experiment”).

- An hypothesis *test* then needs to decide whether to **reject** or **accept** the hypothesis H, given the data *x*.

Hypothesis Testing

- Hypothesis testing is one of the most common uses of statistics.

- Usually assumed one has a model for the data under two competing hypotheses:
  - Null hypothesis (H_0) = e.g. accepted, “reigning” model
  - Alternate hypotheses (H_1) = e.g. competing model

- One makes a measurement *x* and then needs to decide whether to accept or reject H_0
  - Additionally: possibly in favor of H_1

Example Hypothesis Tests

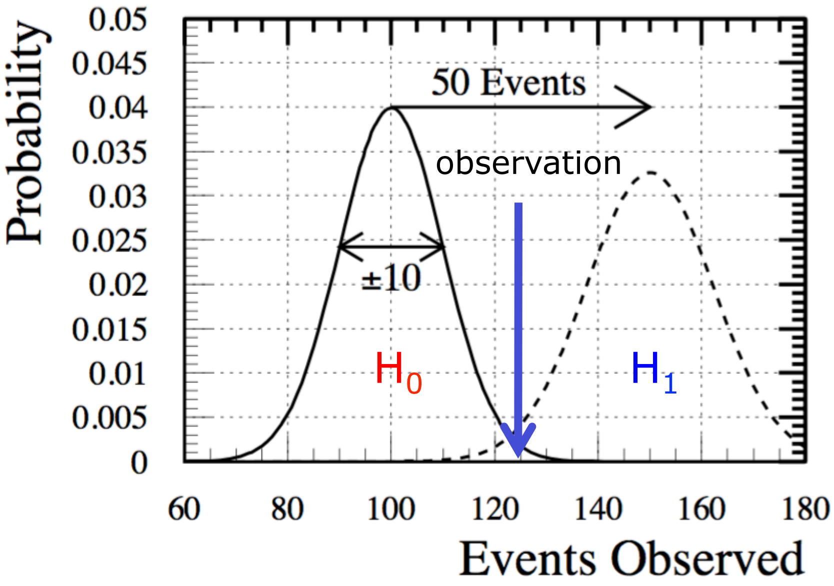
- Causal relationship / correlation:
  Is there a relationship between effect A (e.g. eating meat) and effect B (e.g. becoming aggressive) ?
  - H_0: no correlation
  - H_1: yes, there is a (significant) correlation.

- Classification of records:
  Given all the facts, is client guilty of crime?
  - H_0: no = low-risk client
  - H_1: yes = high-risk client

- Signal over background
  SETI: Seen radio signal over bkg noise from E-T life?
  - H_0: no = background only (described by known sources)
  - H_1: yes = signal peak over background (sign of ET)

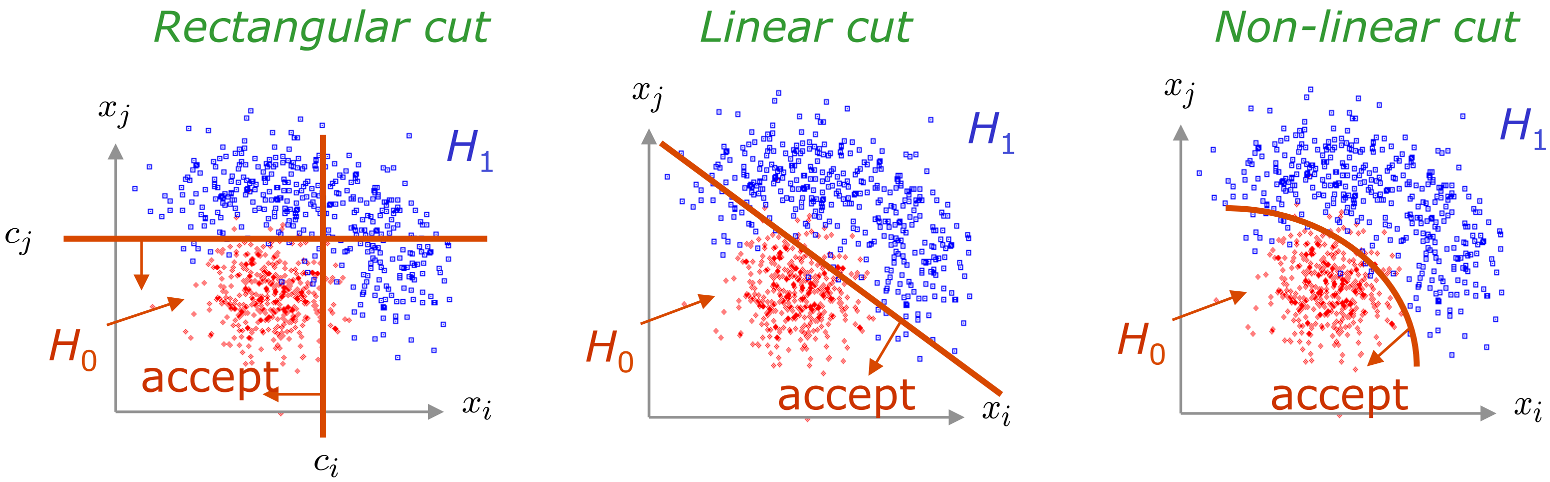
Example for single numbered dataset

- Experiment: count the number of yellow cars passing by in 1 hour on the highway.
  - Observation: 125 cars have passed by.
- Two hypotheses:
  - H_0: on average 100 yellow cars / hour
  - H_1: on average 150 yellow cars / hour
- To accept or reject H_0, that is the question …
- How do you conclude statistically (or not) that a model is ruled out?
  - What value is the decision boundary?



Example for single, multi-dimensional records

- Classification of individual records.
- What is the best decision boundary between H_0 and H_1 ?



Rectangular cut

Linear cut

Non-linear cut

$$t(x) = \theta(x_j - c_j)\theta(x_i - c_i)$$

$$t(x) = a_j \cdot x_j + a_i \cdot x_i$$

$$t(x) = \vec{a} \cdot \vec{x} + \vec{x} A \vec{x} + \ldots$$

- How to set the optimal decision boundary from statistics perspective?
  - In higher dimensions it is not so easy!

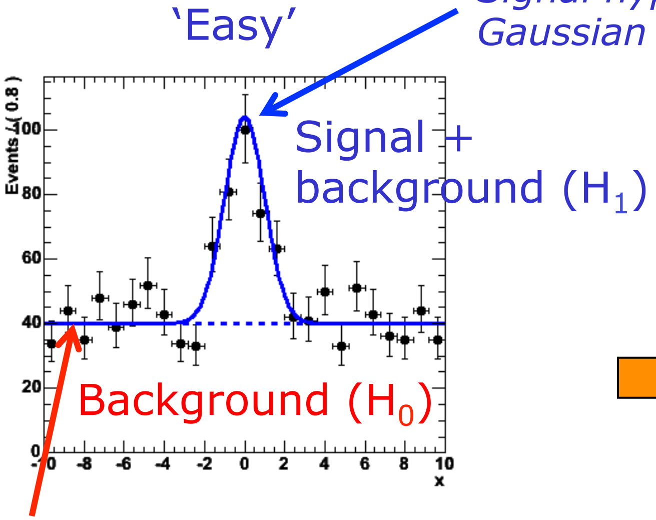
Example for entire dataset

- How to establish the presence of signal in the data (at a certain confidence level) ?
  - "What is the decision boundary for discovery?"
  - Alternative: how to set limits in absence of convincing signal.

'Easy'

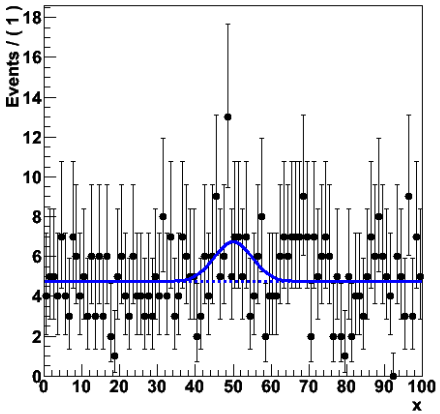
Signal hypothesis
Gaussian distribution in x

'Difficult'



Signal + background (H_1)

Background (H_0)



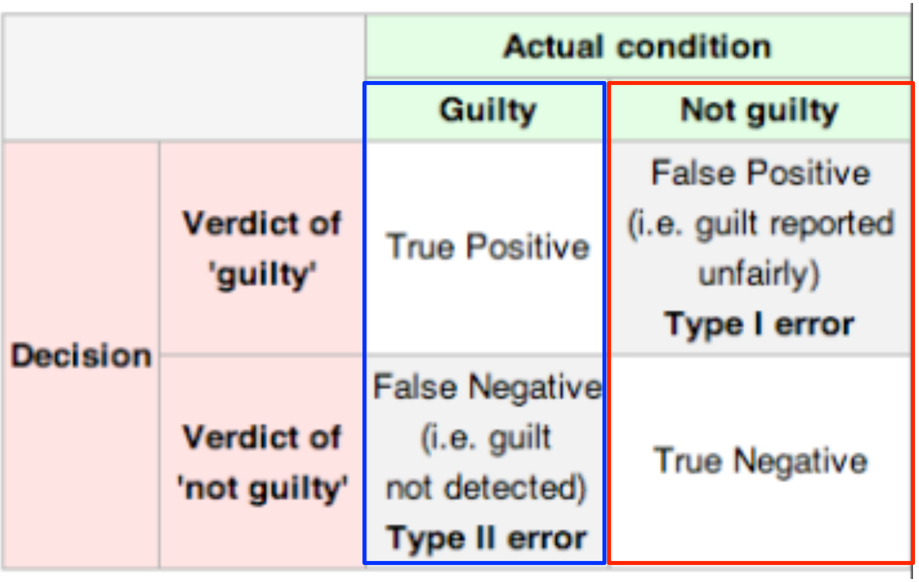
Background hypothesis
Flat distribution in x

- (Of course can fit multiple observables simultaneously.)

Hypothesis Testing – some more definitions

- Before we can make much progress with statistics, we need to decide what it is that we want to do …

- Definition of terms
  - Rate of type-I error = α
    - “Significance level”
    - “Confidence”
  - Rate of type-II error = β
  - Power of test = $1-\beta$
    - “Power”



		Actual condition	
		Guilty	Not guilty
Decision	Verdict of 'guilty'	True Positive	False Positive (i.e. guilt reported unfairly) **Type I error**
	Verdict of 'not guilty'	False Negative (i.e. guilt not detected) **Type II error**	True Negative

H_1 H_0

- α and β intimately related to discovery and exclusion of hypotheses. (More on this later.)

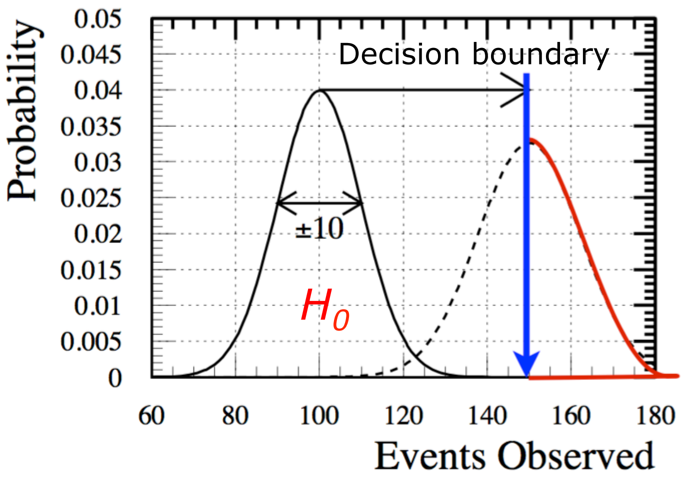
Confidence and Power

- Definitions unclear? T-shirts for sale @
  - https://www.etsy.com/listing/153369986/varsity-statistics-t-shirt

- Definition: power = 1 - β ☺

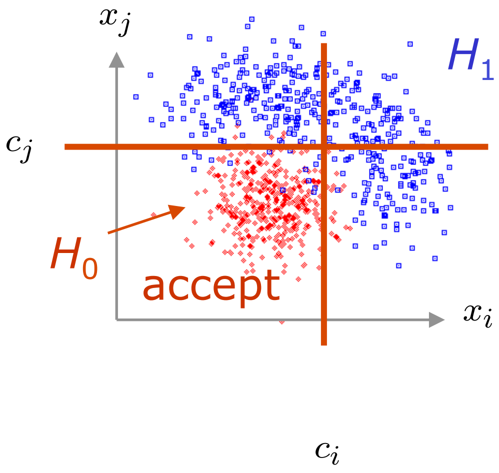
Hypothesis Testing

- In simple case of number counting it is obvious where to set decision boundary for smaller type-I error (α)
  - I.e. move to high number of events.

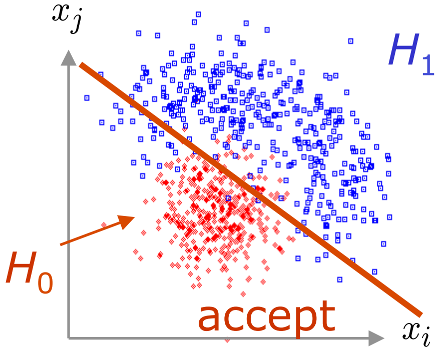


- But again, in higher dimensions it is trivial!

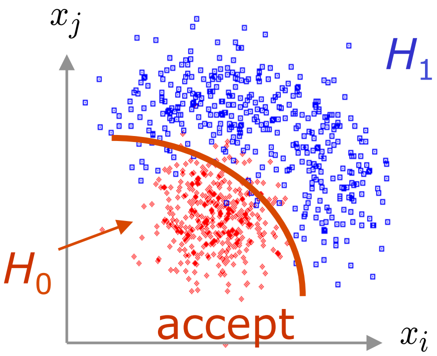
Rectangular cuts?



Linear cut?



Non-linear cut?



Test Statistic

- The ***test statistic***, *t(x)*, is a mapping of your data *x* onto a single, real number.
  - Scalar function: t (x) → $\mathbb{R}$

- Remember, the “data” vary.
  - Can mean: single number, single record of numbers, collection of records, collection of data sets …
  - “Data” are obtained from experiments.

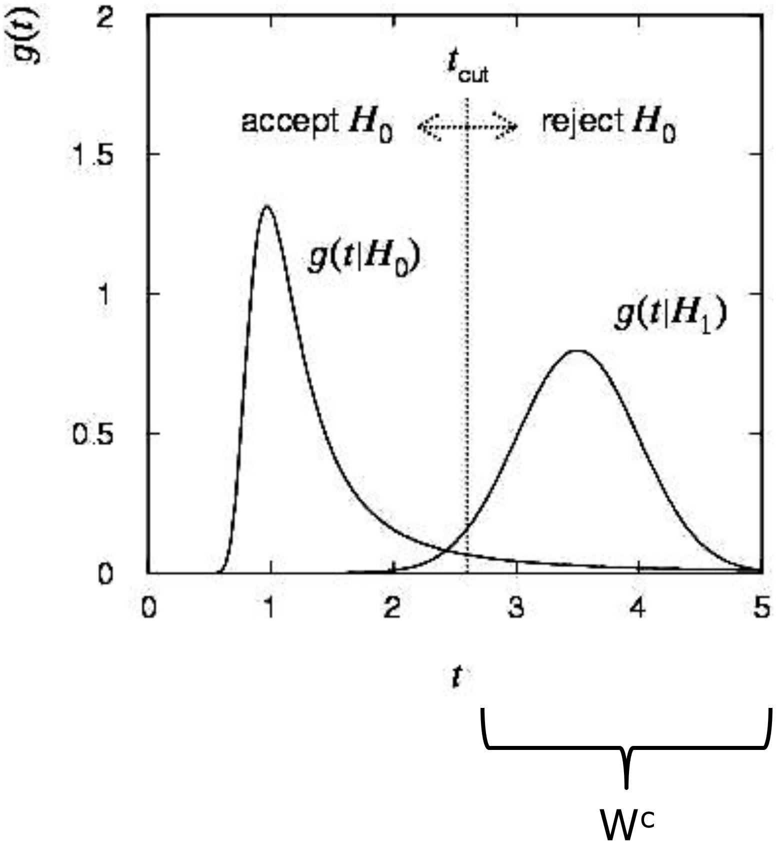
data *x*

- In case of multiple datasets in a testing sample, you obtain the distribution of *t(x)* …

- … can work out two distributions: $g(t|H_0)$ and $g(t|H_1)$
  - In case data can be split into corresponding two classes. (e.g. high- and low-risk clients.)

Test Statistic

- Decision boundary can be defined as a single 'cut' on the test statistic t
  - t_{cut}
- This cut defines the so-called "critical region" W^c.
- *For an n-dimensional problem we have a corresponding 1-d problem.*



Test Statistic

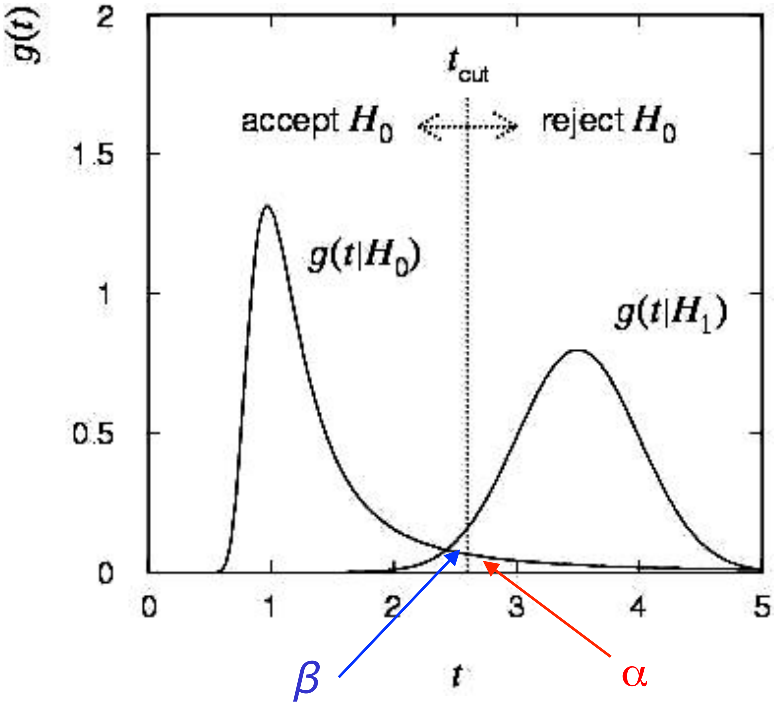
- Probability to falsely reject H_0 if it is true (type-I error):

$$\alpha = \int_{t_{\text{cut}}}^{\infty} g(t|H_0)\,dt$$

- Probability to accept H_0 if H_1 is true (type-II error):

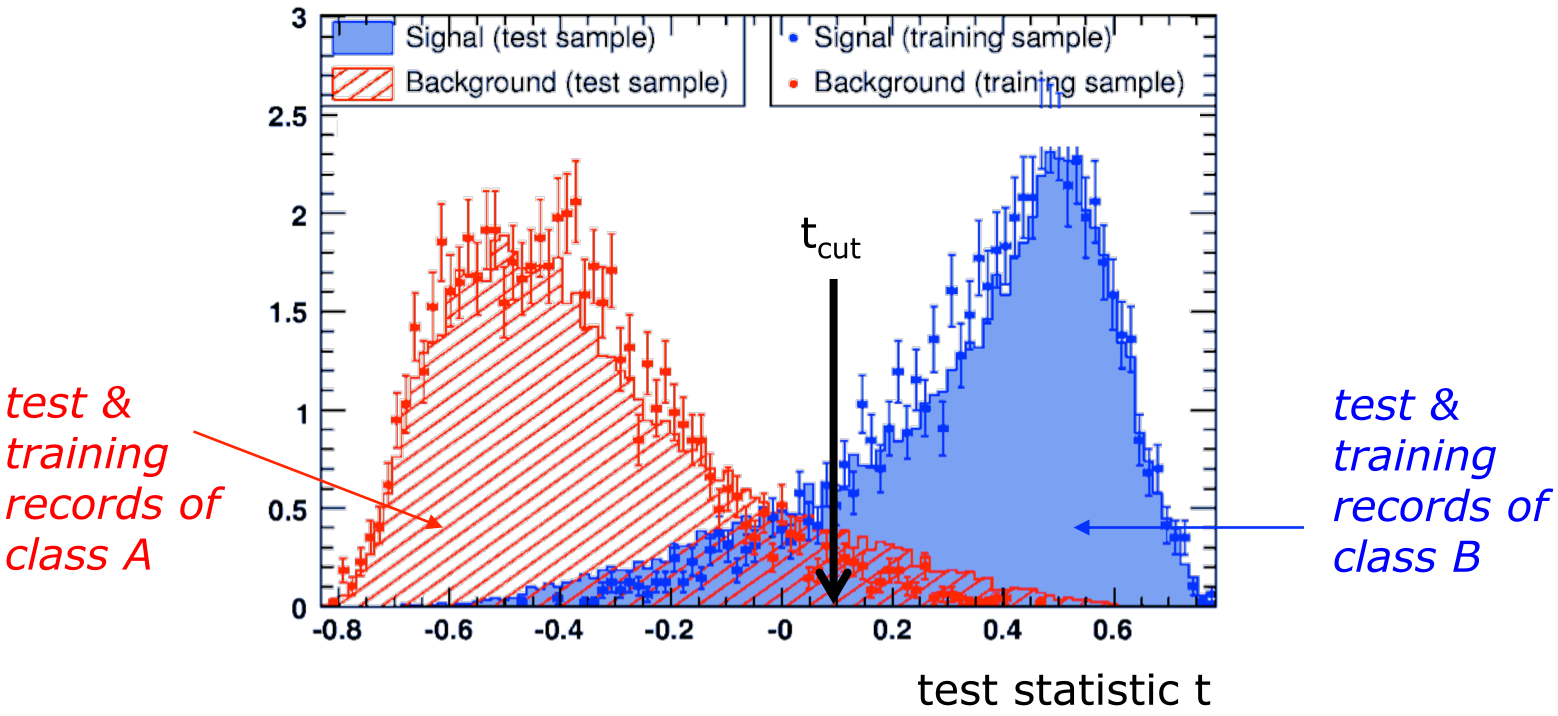
$$\beta = \int_{-\infty}^{t_{\text{cut}}} g(t|H_1)\,dt$$

  - Power = 1 - β



Test Statistic Example – ML classification

- Classification problem: sort records into classes A and B
- ML alg results in “score” value per record.



- Value of *t* relative to t_{cut} determines the assigned label.

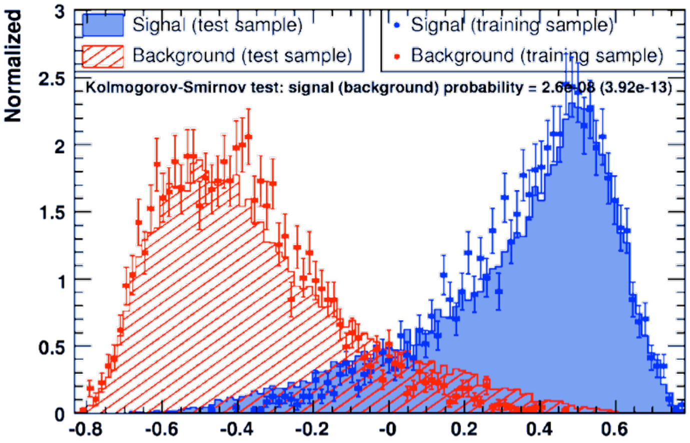
Decision Boundary Optimization

- Treat the two hypotheses asymmetrically
  - Null hypothesis is “special”; the established model.
    - I.e. don’t reject it easily

  → Fix rate of Type I error (= α) to a small value
    - Recall: “the size of the test”

- Now can define a well stated optimization goal:
  Maximize the power for fixed, small value of α

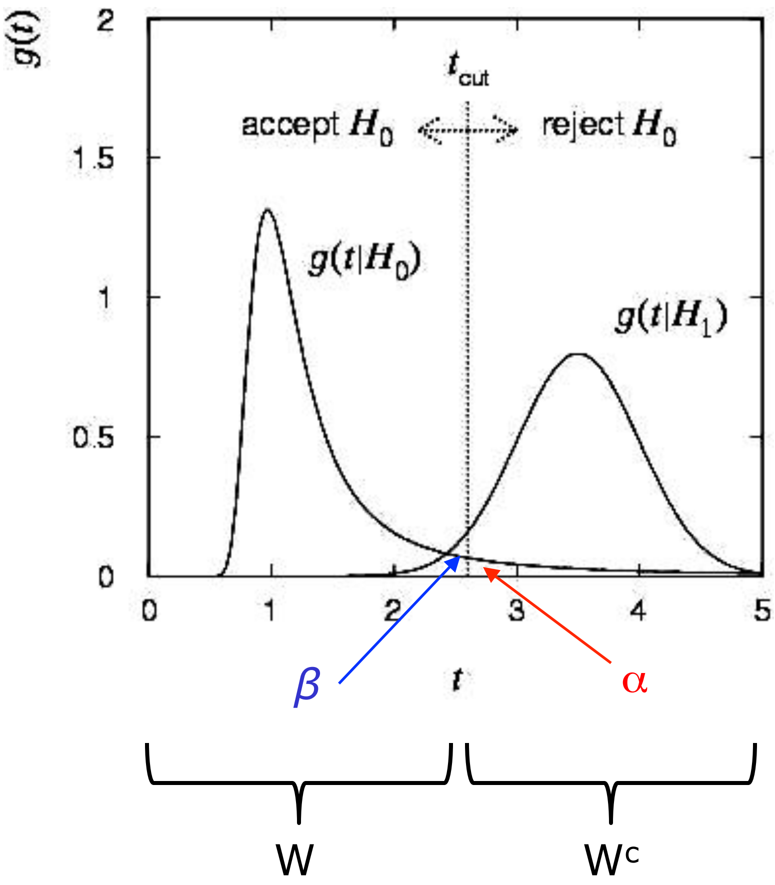
- In other words: obtain best possible separation between distributions of red (H_0) and blue (H_1) classes.



Neyman-Pearson lemma

- How to choose a test's critical region in an 'optimal way' ?
- In 1928-1938 Neyman & Pearson developed a theory for this with two competing hypotheses (H_0 and H_1).

- Given a fixed probability α that we wrongly reject the null hypothesis …
  - Convention: if data *x* falls in W then we accept H_0
  - $\alpha = P(x \notin W | H_0)$
- … find the region W such that we minimize the probability of wrongly accepting H_0 (when H_1 is true), i.e. that maximizes the power 1-*β*
  - $\beta = P(x \in W | H_1)$



Neyman-Pearson lemma

- The region W that minimizes the rate of the type-II error is a contour of the Likelihood Ratio:

$$\frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)} > c$$

- … where *c* is a constant which determines the power

- Any other region of the same size will have less power.

- Equivalently → the optimal test-statistic is the likelihood ratio!

$$\boxed{t(\mathbf{x}) = \frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)}}$$

Outline of Proof

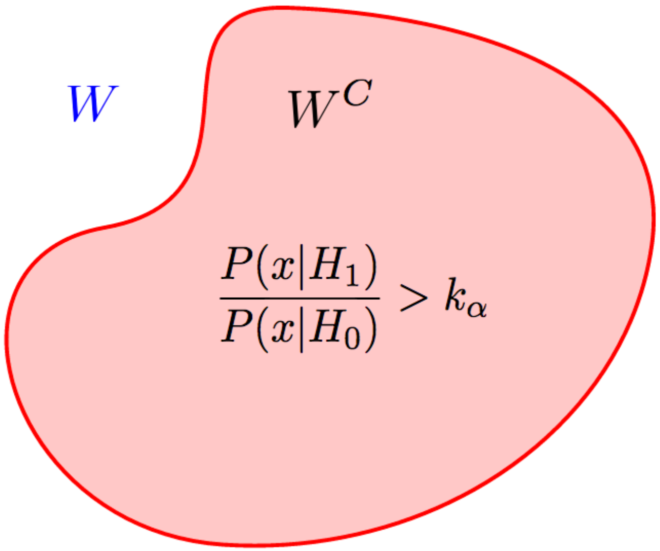
- We want to maximize the power (1-*β*) over α

$$\frac{1-\beta}{\alpha} = \frac{\int_{\mathbf{W}^c} \frac{P(x|H_1)}{P(x|H_0)} P(x|H_0)\,dx}{\int_{\mathbf{W}^c} P(x|H_0)\,dx}$$

  - Note: W^c is the complement of W

- = the average of the likelihood ratio $P(x|H_1) / P(x|H_0)$ over the critical region W^c.

- This is maximized if W^c contains the part of the sample space with the largest values of the likelihood ratio.



Summarizing

1. Data samples can be described with probabity mass functions (PMFs) or probability density functions (PDFs).
2. The *likelihood* is the value of a probability density/mass function evaluated at the measured value of the observable(s) in the dataset.
3. The maximum likelihood is the likelihood value after the fit to the data.

- The best separation between two hypotheses is obtained when cutting on the ratio of the two (maximum) likelihoods describing the two hypotheses.

The likelihood ratio – interpretation

- Likelihood ratio (LR) is Optimal test statistic: $\dfrac{L(x \mid H_1)}{L(x \mid H_0)}$

- NB any monotonic function of LR leads to the same test.

- Often used instead:

$$
\begin{aligned}
q &= -2\log\left[\frac{L(x \mid H_1)}{L(x \mid H_0)}\right] \\
&= -2\log L(x \mid H_1) + 2\log L(x \mid H_0) \\
&\cong \chi^2(x \mid H_1) - \chi^2(x \mid H_0) \\
&= \Delta\chi^2
\end{aligned}
$$

- Meaning: perform fit to data twice, for both hypotheses. Then take the *difference* in the -2log(likelihood) values.
  - This is (approx.) equal to the difference in X^2 between the fits.

Translation for 3 approaches

- Counting experiment:
  - t(x) is number of counts N.
  - Optimal decision boundary is defined by cut on t(x)
- MVA
  - Compactification of N observables to 1.
    - NB: this step looses information! (= loss of some sensitivity)
  - E.g. the ‘score’ predicted by MVA per record. t(x) = MVA score.
  - optimal decision boundary is defined by cut on t(x)
- Fit based on parametric models:
  - Data x contains N observables
  - Construct the N-dim pdfs describing both signal S(x) and bkg B(x).
    - All information retained → full sensitivity.
  - Cut on ratio of signal and background likelihoods, both fit to the data.

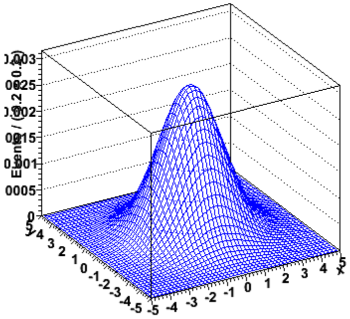
What is the best you can do?

- Neyman-Pearson lemma, in practice:
- For a problem described by a continuous signal distribution f(x|s) and a continuous bkg distribution f(x|b) the optimal acceptance region is defined by

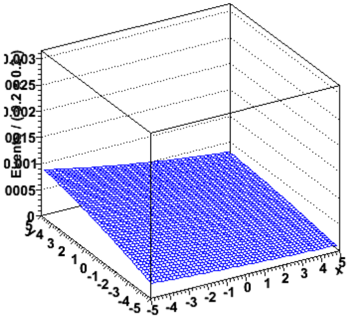
$$\boxed{\frac{f(\vec{x}|\text{s})}{f(\vec{x}|\text{b})} > c}$$

- Where:

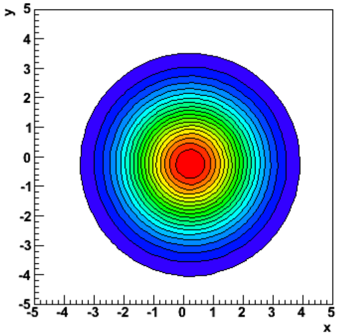
$f(\vec{x}|\text{s})$



$f(\vec{x}|\text{b})$



$\frac{f(\vec{x}|\text{s})}{f(\vec{x}|\text{b})} > c$



Practical notes on Neyman-Pearson

- The problem is that we usually don't usually have explicit formulae for the pdfs $f(\vec{x}|\mathrm{s}),\ f(\vec{x}|\mathrm{b})$.

- And these are hard to build in case of N-dim distributions of x (N>3).
  - Compromises are possible …
  - ... In this case multi-var algs have a clear advantage!
    - N can be O(100)-O(1000) observables!

However:

- If a good way can be found to describe the pdfs of x that enter into likelihood then the maximum likelihood ratio is hard to beat in practice.

Next topics:

p-values
Discovery and Exclusion
Wilk’s theorem

p-value joke

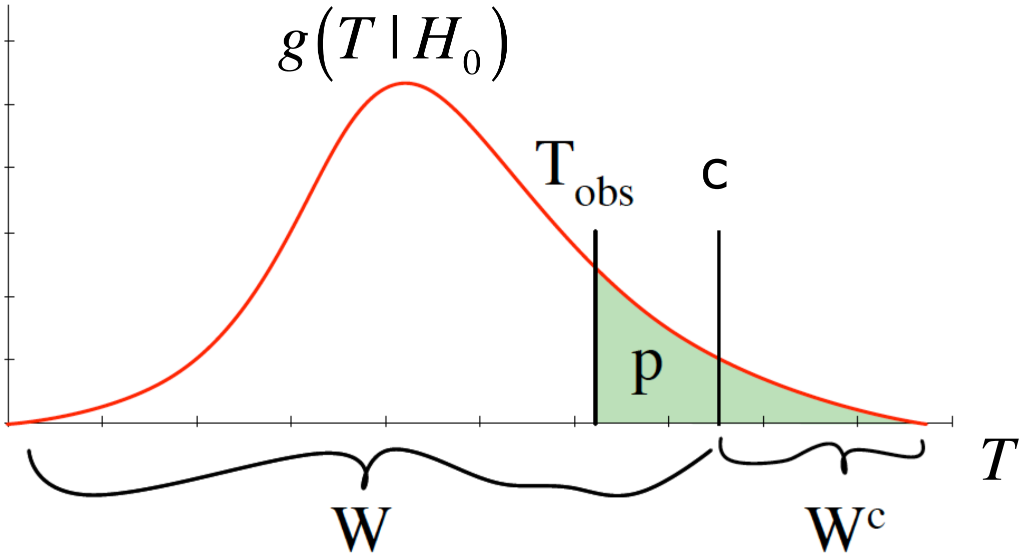
- How many statisticians does it take to ensure at least 50 percent chance of a disagreement about p-values?

p-value definition

- p-value =
  the probability of getting results at least as extreme as the ones you observed (T_{obs}), given that the null hypothesis is correct.

$$p = p_0 = \int_{T_{obs}}^{\infty} g\left(T \mid H_0\right) dT$$

- Defined in terms of the test statistic (T) distribution for H_0 →



- This is not the probability that H_0 is true!

Practical example of p-value

- Imagine that you have a coin that you suspect is weighted toward heads.
  - (Your null hypothesis is then that the coin is fair.)

- You flip it 100 times and get more heads than tails.

- The p-value will tell you the probability that you'd get at least as many heads as you did if the coin was fair.

- The p-value will *not* tell you whether the coin is fair.

- (That's it — nothing more!)

Using a p-value to define "discovery" test of H_0

- Rejecting the null-hypothesis is called "discovery". We do this when $p_0 \leq \alpha$.

- The probability to find a p-value of H_0, p_0, less than α is:

$$P(p_0 \leq \alpha | H_0) = \alpha$$

- The "critical region" (W^c) of a test of H_0 with size α is defined as: the set of data phase-space where $p_0 \leq \alpha$.

At what p-value do we claim discovery?

- Remember, the null hypothesis is special.
  In general don't reject this lightly.
  - Small α = small chance of false claim of discovery.

- Note: cost of type-I error (false claim of discovery) can be high!
  - Reputation damage! E.g. social psychology studies.
  - Or remember the 'discovery' of cold nuclear fusion

- The significance level is very subjective and depends on the phenomenon in question, e.g.,

phenomenon	reasonable α for discovery	
Sociology: relation b/n events A & B	~0.05	
Discovery new elementary particle	$\sim 10^{-7}$	Small, to play safe!
Life on Mars	$\sim 10^{-10}$	
Astrology is real	$\sim 10^{-20}$	

Significance from p-value

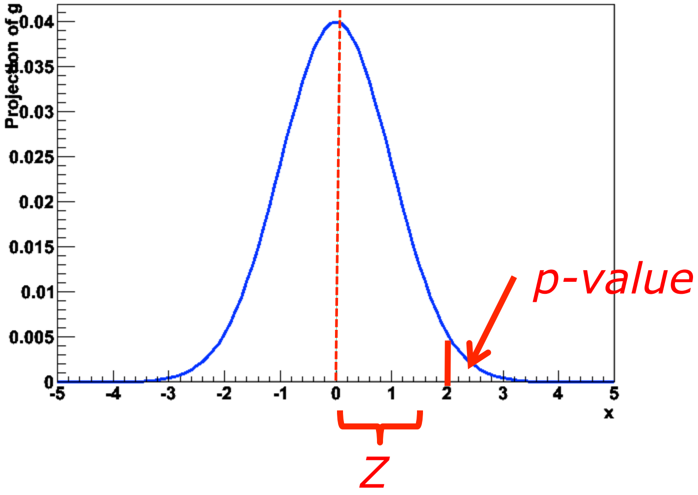
- p-Value vs Z-value (significance)
  - Often defines significance *Z* as the number of standard deviations that a Gaussian variable would fluctuate *in one direction* to give the same *p*-value.

TMath::Erfc

$$p = \int_Z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx = 1 - \Phi(Z)$$

$$Z = \Phi^{-1}(1-p)$$

TMath::NormQuantile



- We say, e.g.: “p-value is 5σ”

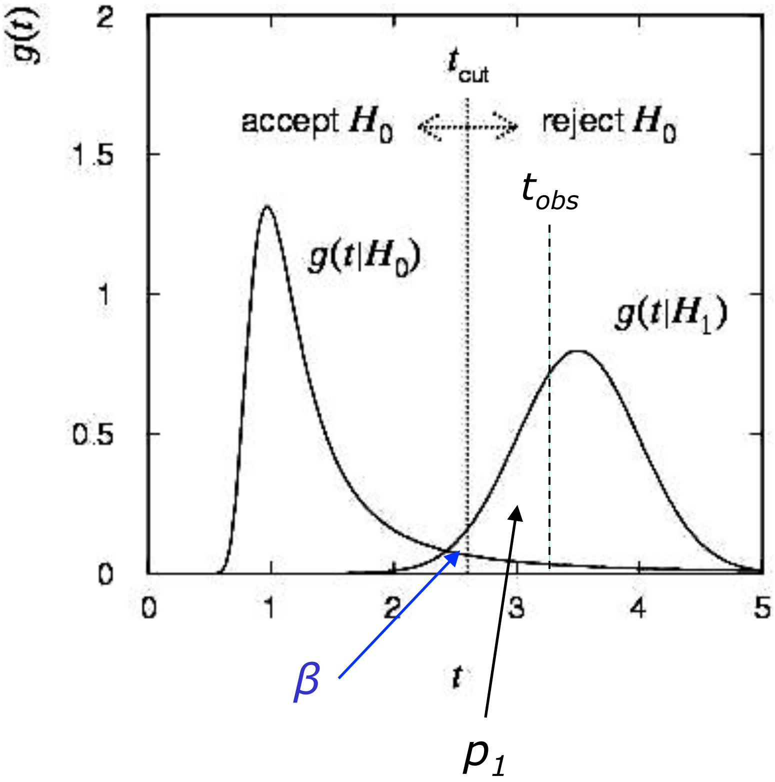
Exclusion test of H_1 hypothesis

- Define p_1 as:

$$p_1 = CL_{S+B} = \int_{-\infty}^{T_{obs}} g(T \mid H_1)\, dT$$

- And remember that

$$\beta = \int_{-\infty}^{T_{cut}} g(T \mid H_1)\, dT$$



- Exclude H_1 when: $p_1 \leq \beta$.
- In general less strict for model exclusion than discovery.
- β regularly set at 5%.
  - $p_1 \leq 5\%$ one says: "We exclude model H_1 at 95% confidence level."

Exclusion based on CL_S

- What if the models for H_0 and H_1 are very similar?

- To the point where the data really has no sensitivity to differentiate between the two hypotheses?

- (Illustration on black board.)

- For this we introduce the “CL_S probability”.
  - (Continued on next slide.)

Exclusion based on $\mathrm{CL_S}$

- Recall:

$$CL_{S+B} = \int_{-\infty}^{T_{obs}} g(T \mid H_1)\, dT$$

- Define:

$$CL_B = \int_{-\infty}^{T_{obs}} g(T \mid H_0)\, dT$$

- Then:

$$CL_S = {CL_{S+B}}\big/{CL_B}$$

E.g. only exclude $\mathrm{H_1}$ when $\mathrm{CL_S} < 5\%$.

- $\mathrm{CL_S}$ has property that $\mathrm{H_1}$ is not excluded when $\mathrm{H_0}$ and $\mathrm{H_1}$ hypotheses are very similar.
- $\mathrm{CL_S}$ is a conservative, on average: $CL_S = 2 \times p_1$

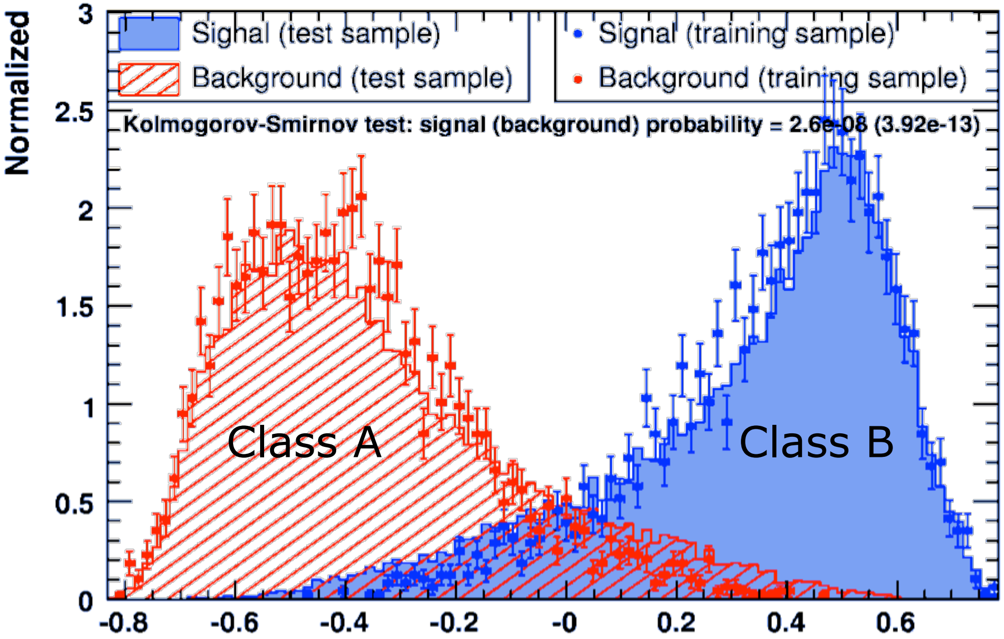
Practical evaluation of p-values

- P-values assigned to single records / experiments.

- This assumes we have enough experiments/records available to evaluate the underlying test-statistic distributions, as needed to calculate any (small) probabilities!

- E.g. in standard classification problem, need enough records to evaluate red and blue distributions.

- … which are needed to evaluate integrals for p-value calculation of single record.

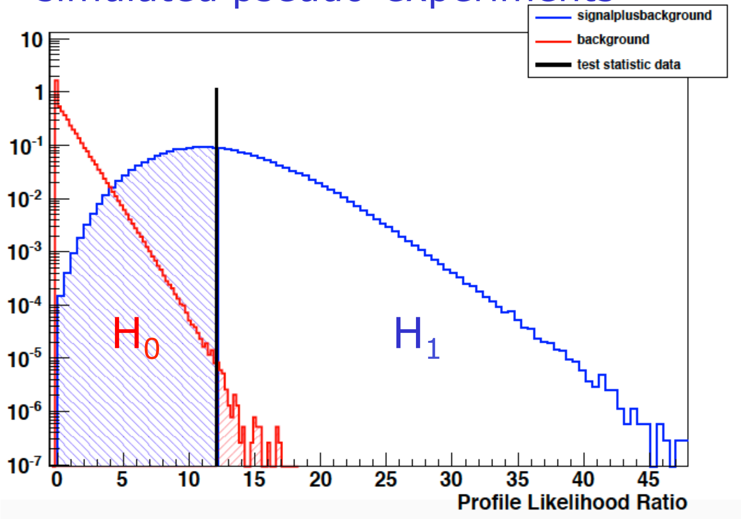


Practical evaluation of p-values

- Of course all this assumes we have / or can perform enough experiments/records to evaluate very small probabilities!

- Sometime one can repeat experiments, or simulate them …
  - If you repeat an experiment many times, a given fraction of experiments will result in more extreme t.s. than (original) observed value.
  - This can be very time consuming.
- … But often this is totally impossible!

Plot below uses millions of simulated pseudo-experiments



- E.g. to claim a 5σ discovery one needs ~10^8 (!) pseudo-experiments for reliable test statistic comparison!
- How can one ever claim such significances in practice?

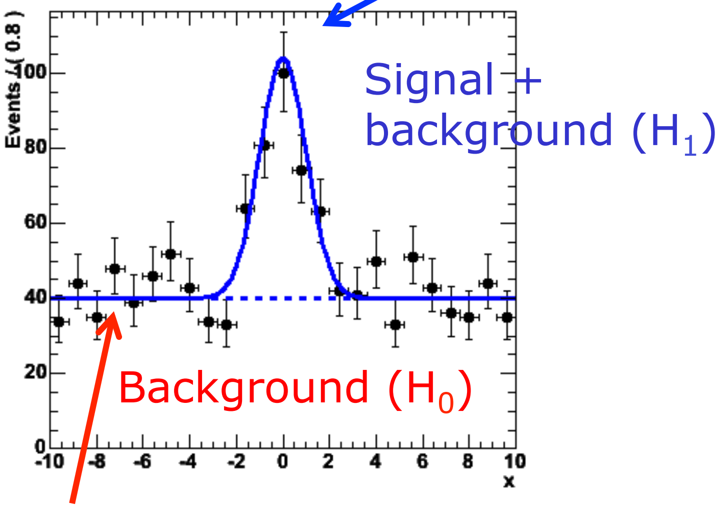
Wilk’s theorem

- One can show that, under H_0, the distribution of -2logLR follows a χ^2 distribution, with k degrees of freedom.
  - Certain special conditions apply, such as requirement of the “large sample limit” (which we will not go into here)
- k is the difference in the number of free parameters in the models for H_1 and H_0.
- Extension of Wilk’s theorem to H_1 exist as well:
  - “Asymptotic formulae for likelihood-based tests”
  - https://arxiv.org/abs/1007.1727

Wilk’s Example

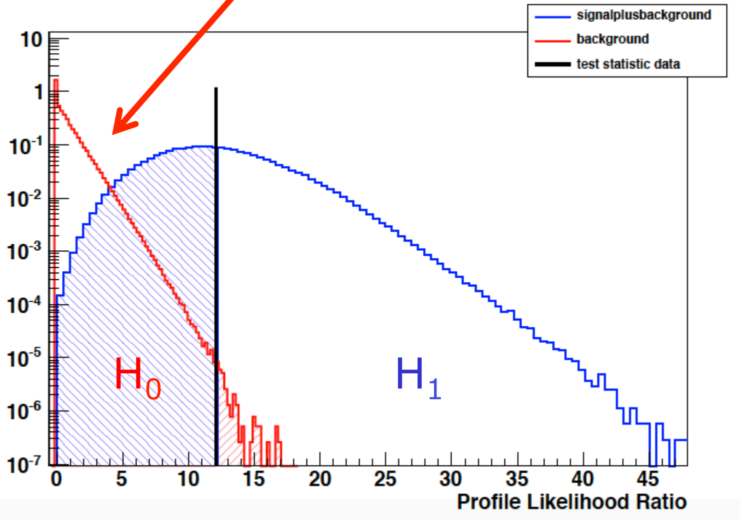
- Example: two fit models below differ by one parameter
  - Parameter turns signal peak on or off in the fit to data.
- Red distribution (H_0) follow an exponential distribution.
- Which is the χ^2 distribution with 1 degree of freedom.

Signal model is a Gaussian distribution in x



Signal + background (H_1)

Background (H_0)



H_0

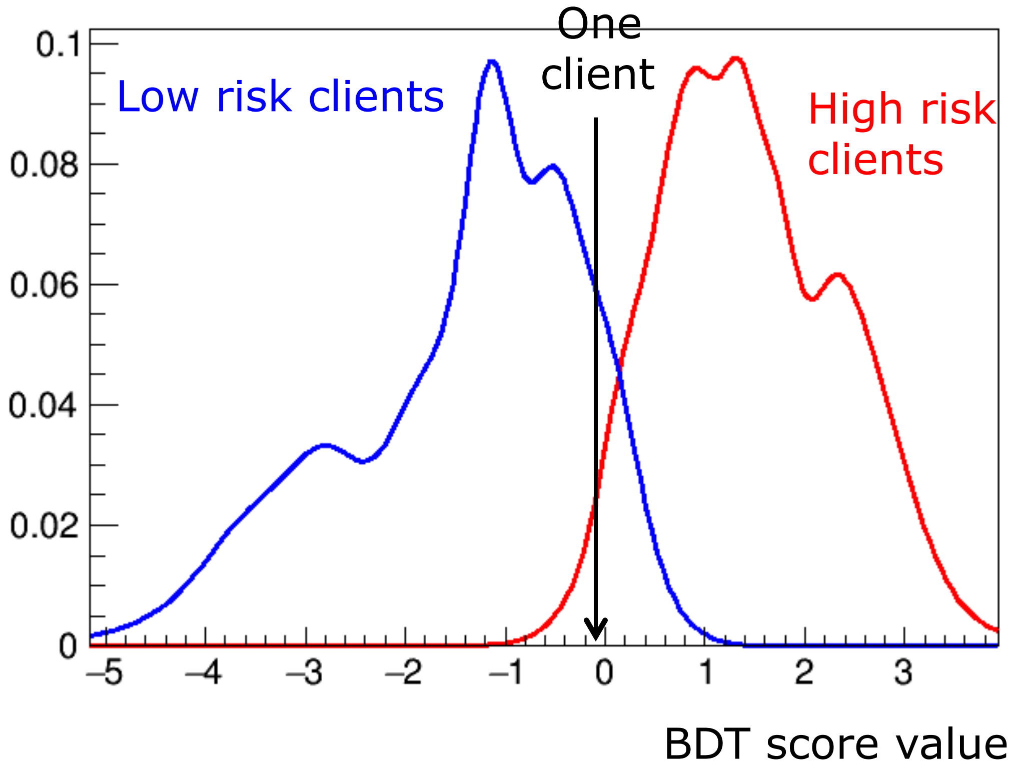
H_1

Background hypothesis
Flat distribution in x

- You can calculate the proper p-value without the need for all other records.

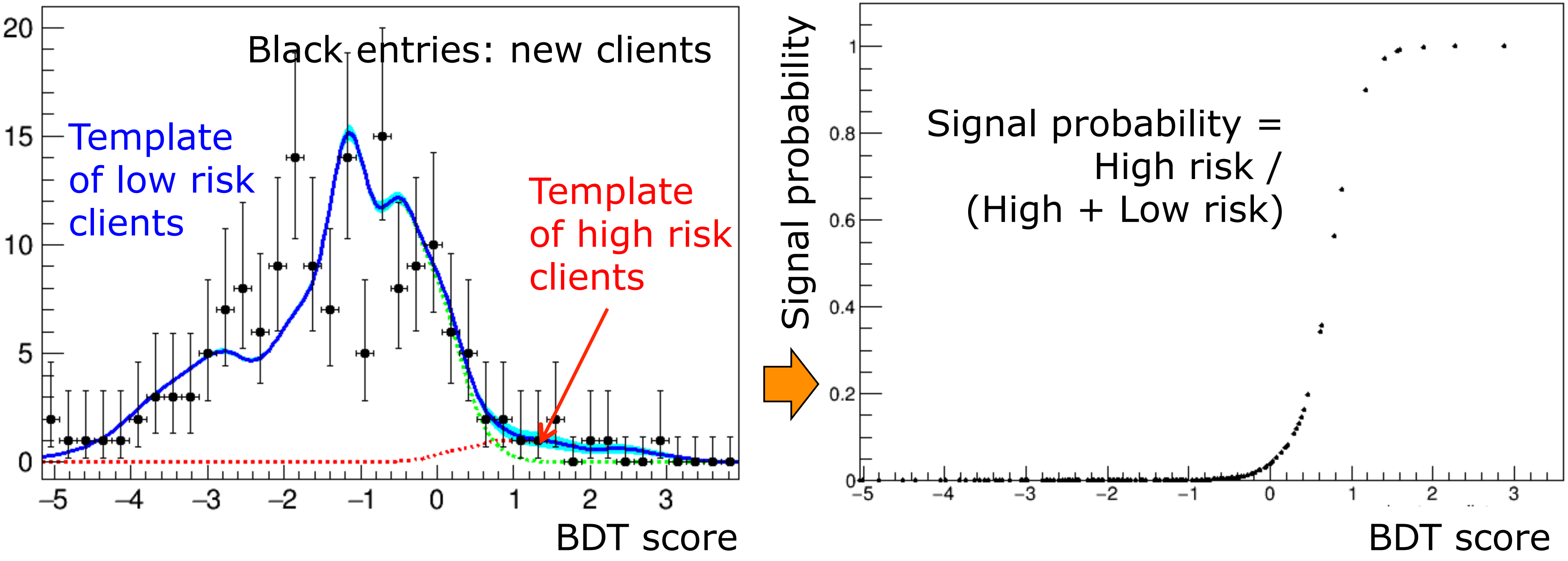
Experiment versus Records

- Example: MVA separates records (clients) into two classes: "low risk" and "high risk".
- Test statistic distributions below (from "testing sample")
- Can now assign p-value per single client:
  - … using statistics machinery just discussed.
  - … based purely on the characteristics of the single client.
- This ignores info of how often high-risk cases occur with other clients.
- *What if this is known?*



Probability including frequency fit information

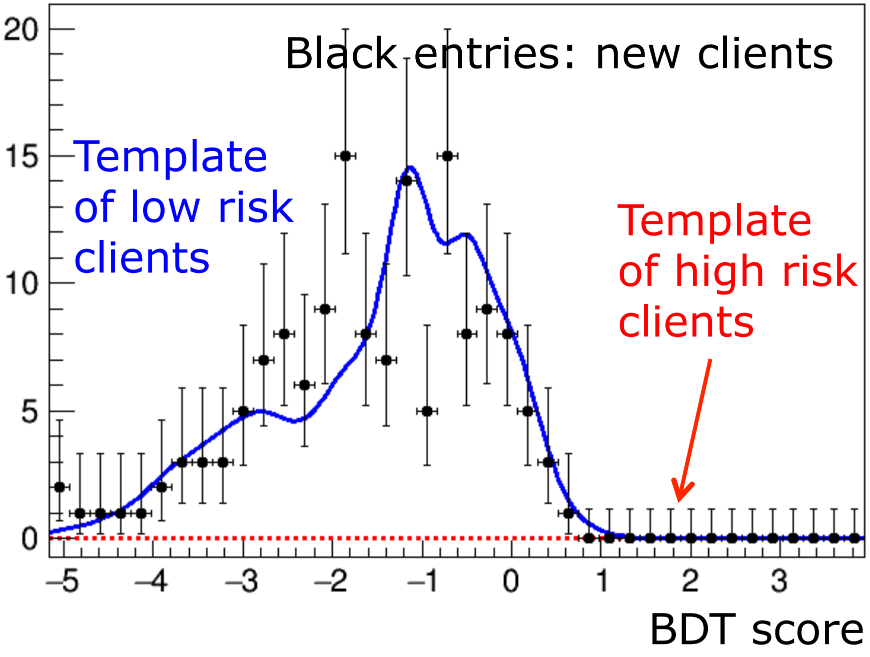
- One can fit the scoring distribution of new records to templates of low-risk and high-risk clients.
- Example below: dataset contains 10% high-risk clients.



- → A signal probability based on fit to BDT score, which *includes* chance of how frequently high-risk clients occur in total dataset.

Probability including frequency fit information

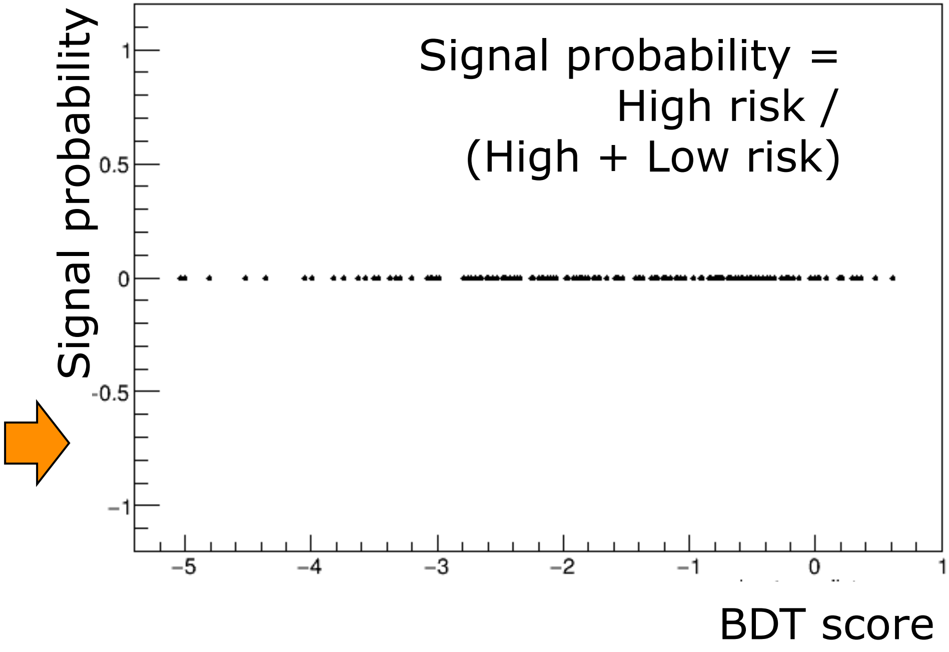
- Example below: dataset contains 0% high-risk clients.
- Fitting templates finds no evidence for high-risk clients.



Black entries: new clients

Template of low risk clients

Template of high risk clients



Signal probability = High risk / (High + Low risk)

- → Signal probability based on BDT score & fit to data assigns 0% signal probability for all clients.

Drawing conclusions from data – *the right way*

Six analysis pitfalls

Solutions

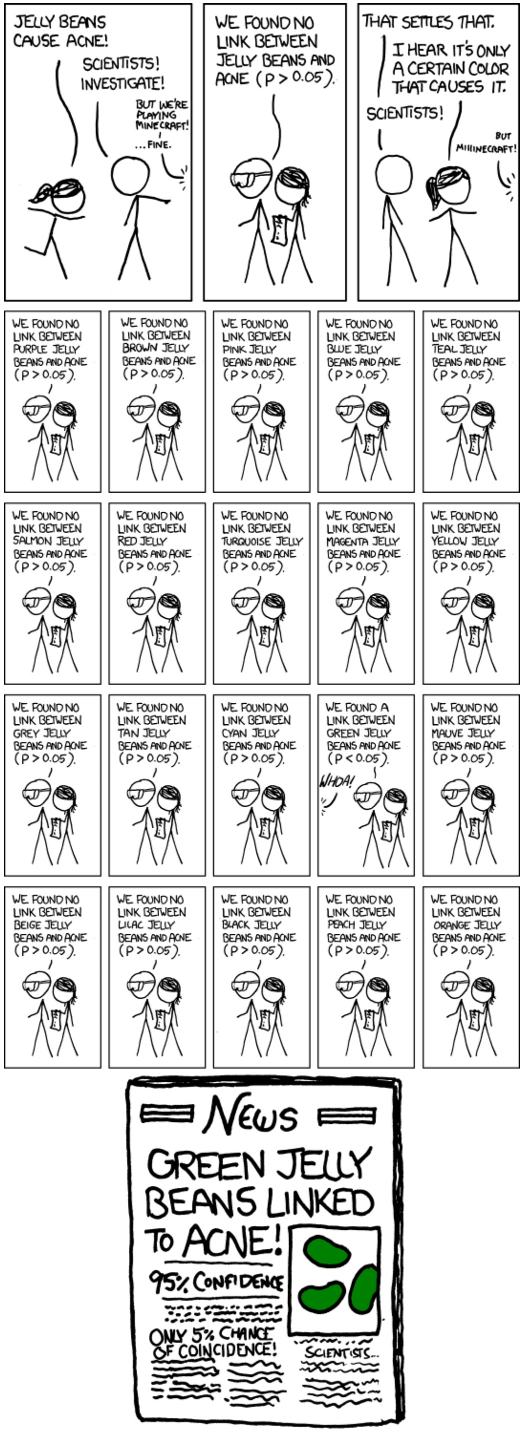
1. Look elsewhere effect

- Also known as:
  Multiple comparisons problem

- Quoting the significance of of a single experiment …

- … even though many experiments have been performed …
  - (is wrong!)

- … so the single experiment may well have been a statistical fluctuation in the whole set.



JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

https://xkcd.com/882/

2. Simpson’s paradox – confounding observables

- Other (extreme) example of dependence on confounding variables: Simpson’s paradox
- Conclusion of a study is reversed when confounding variables are taken into account

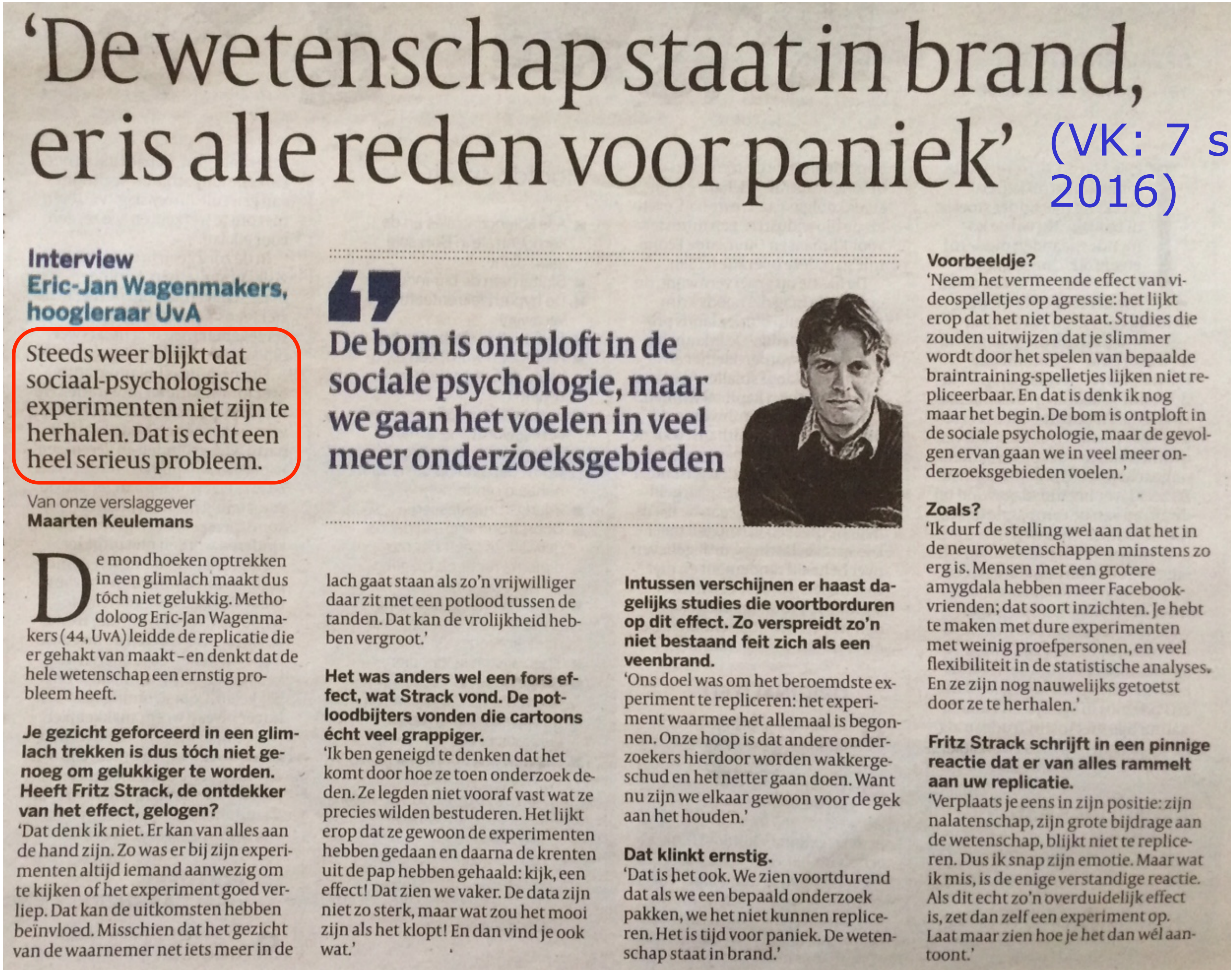
Example: study of success rate in removing kidney stones

Two methods were compared:

- Open surgery: 78% of treatments successful
- Small puncture: 83% of treatments successful

Conclusion: “Small puncture method” more successful. Or is it? →

BMJ 1994;309:1480 - http://dx.doi.org/10.1136/bmj.309.6967.1480
BMJ 1986;292:879-82 - http://dx.doi.org/10.1136/bmj.292.6524.879



‘De wetenschap staat in brand, er is alle reden voor paniek’

(VK: 7 sep 2016)

Interview
Eric-Jan Wagenmakers, hoogleraar UvA

Steeds weer blijkt dat sociaal-psychologische experimenten niet zijn te herhalen. Dat is echt een heel serieus probleem.

Van onze verslaggever
Maarten Keulemans

De mondhoeken optrekken in een glimlach maakt dus tóch niet gelukkig. Methodoloog Eric-Jan Wagenmakers (44, UvA) leidde de replicatie die er gehakt van maakt – en denkt dat de hele wetenschap een ernstig probleem heeft.

De bom is ontploft in de sociale psychologie, maar we gaan het voelen in veel meer onderzoeksgebieden

Je gezicht geforceerd in een glimlach trekken is dus tóch niet genoeg om gelukkiger te worden. Heeft Fritz Strack, de ontdekker van het effect, gelogen?
‘Dat denk ik niet. Er kan van alles aan de hand zijn. Zo was er bij zijn experimenten altijd iemand aanwezig om te kijken of het experiment goed verliep. Dat kan de uitkomsten hebben beïnvloed. Misschien dat het gezicht van de waarnemer net iets meer in de lach gaat staan als zo’n vrijwilliger daar zit met een potlood tussen de tanden. Dat kan de vrolijkheid hebben vergroot.’

Het was anders wel een fors effect, wat Strack vond. De potloodbijters vonden die cartoons écht veel grappiger.
‘Ik ben geneigd te denken dat het komt door hoe ze toen onderzoek deden. Ze legden niet vooraf vast wat ze precies wilden bestuderen. Het lijkt erop dat ze gewoon de experimenten hebben gedaan en daarna de krenten uit de pap hebben gehaald: kijk, een effect! Dat zien we vaker. De data zijn niet zo sterk, maar wat zou het mooi zijn als het klopt! En dan vind je ook wat.’

Intussen verschijnen er haast dagelijks studies die voortborduren op dit effect. Zo verspreidt zo’n niet bestaand feit zich als een veenbrand.
‘Ons doel was om het beroemdste experiment te repliceren: het experiment waarmee het allemaal is begonnen. Onze hoop is dat andere onderzoekers hierdoor worden wakkergeschud en het netter gaan doen. Want nu zijn we elkaar gewoon voor de gek aan het houden.’

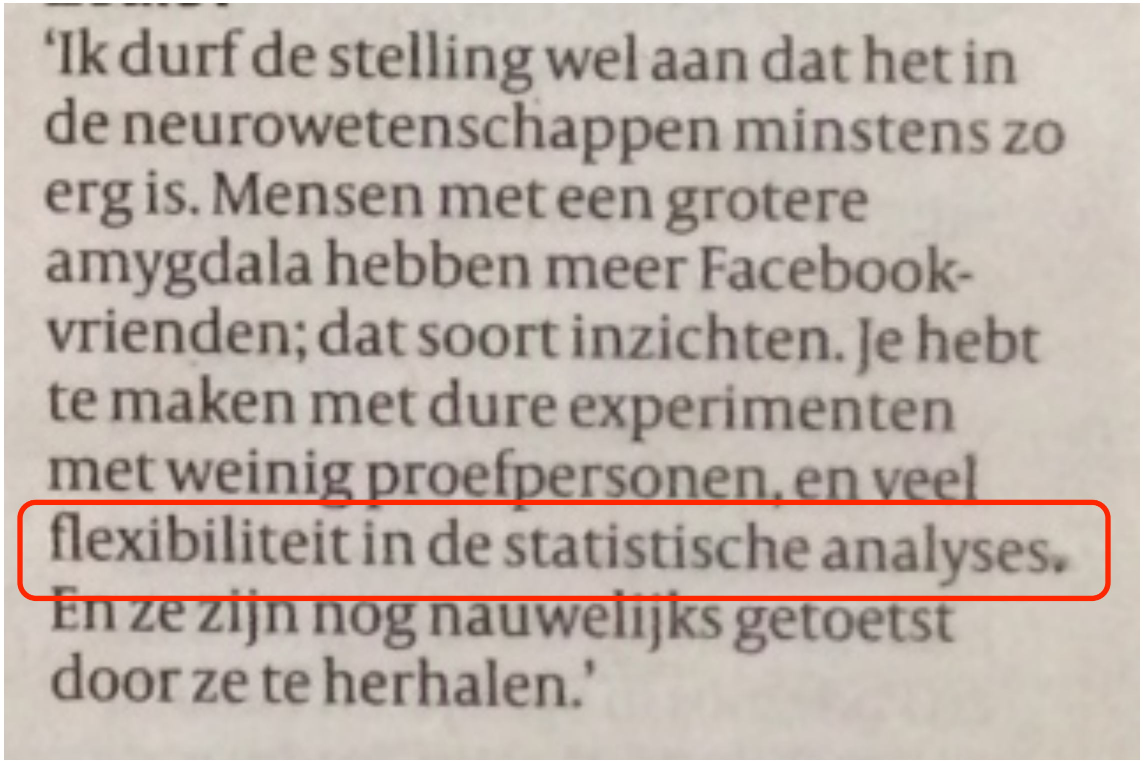
Dat klinkt ernstig.
‘Dat is het ook. We zien voortdurend dat als we een bepaald onderzoek pakken, we het niet kunnen repliceren. Het is tijd voor paniek. De wetenschap staat in brand.’

Voorbeeldje?
‘Neem het vermeende effect van videospelletjes op agressie: het lijkt erop dat het niet bestaat. Studies die zouden uitwijzen dat je slimmer wordt door het spelen van bepaalde braintraining-spelletjes lijken niet repliceerbaar. En dat is denk ik nog maar het begin. De bom is ontploft in de sociale psychologie, maar de gevolgen ervan gaan we in veel meer onderzoeksgebieden voelen.’

Zoals?
‘Ik durf de stelling wel aan dat het in de neurowetenschappen minstens zo erg is. Mensen met een grotere amygdala hebben meer Facebook-vrienden; dat soort inzichten. Je hebt te maken met dure experimenten met weinig proefpersonen, en veel flexibiliteit in de statistische analyses. En ze zijn nog nauwelijks getoetst door ze te herhalen.’

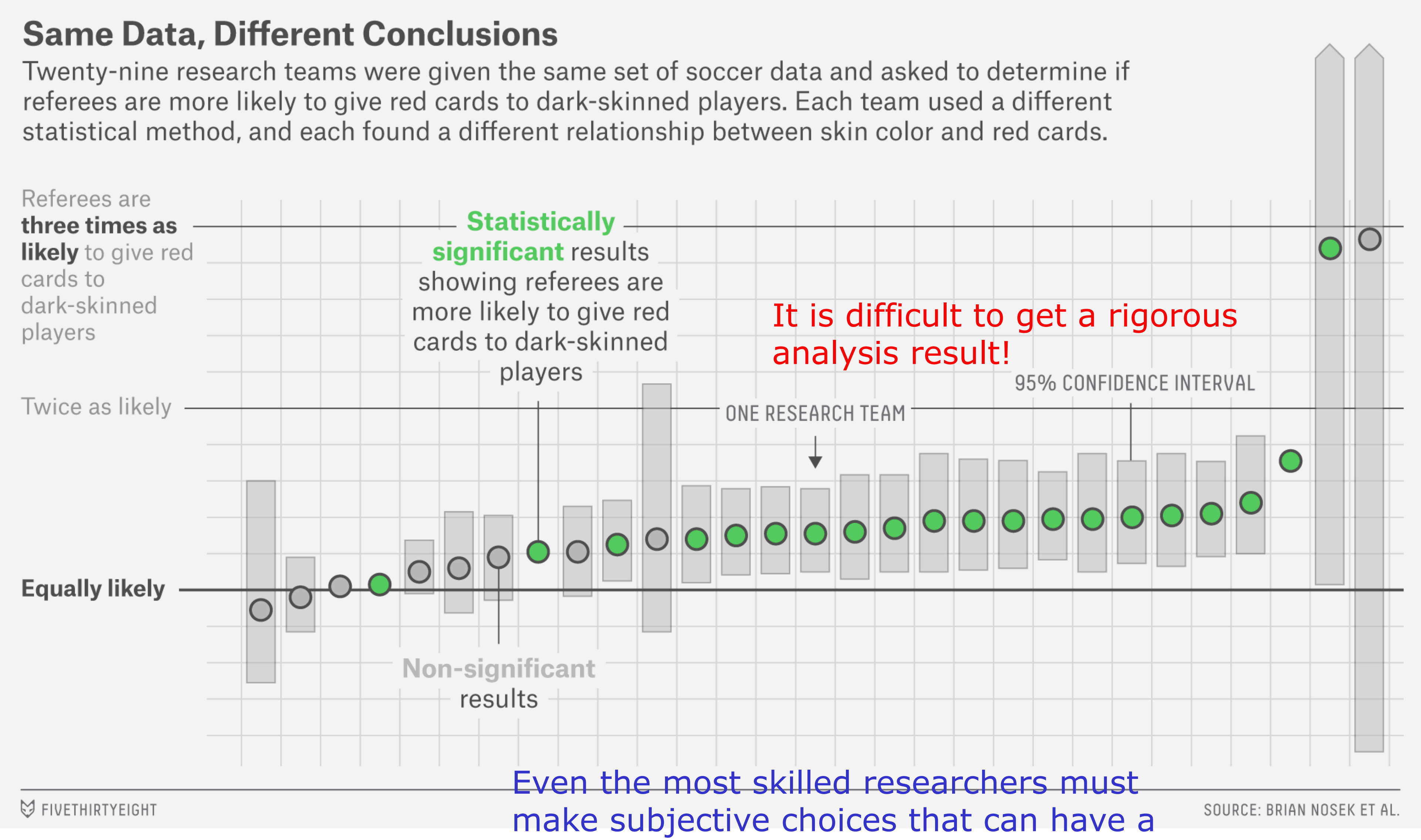
Fritz Strack schrijft in een pinnige reactie dat er van alles rammelt aan uw replicatie.
‘Verplaats je eens in zijn positie: zijn nalatenschap, zijn grote bijdrage aan de wetenschap, blijkt niet te repliceren. Dus ik snap zijn emotie. Maar wat ik mis, is de enige verstandige reactie. Als dit echt zo’n overduidelijk effect is, zet dan zelf een experiment op. Laat maar zien hoe je het dan wél aantoont.’

VK: 7 Sep 2016 – zoom in



'Ik durf de stelling wel aan dat het in de neurowetenschappen minstens zo erg is. Mensen met een grotere amygdala hebben meer Facebook-vrienden; dat soort inzichten. Je hebt te maken met dure experimenten met weinig proefpersonen, en veel flexibiliteit in de statistische analyses. En ze zijn nog nauwelijks getoetst door ze te herhalen.'

3. Flexibility: subjective choices in analysis setup



Same Data, Different Conclusions

Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.

SOURCE: BRIAN NOSEK ET AL.

It is difficult to get a rigorous analysis result!

Even the most skilled researchers must make subjective choices that can have a huge impact on the result they find …

4. Human bias in data analysis

- Working towards a desired analysis result (and similar types of manipulations) often arise from human biases.
  - Often human biases can be unconscious!
- Example biases:
  - Anchoring bias – (demonstrated last week)
  - Confirmation bias – confirm a pre-existing result

- E.g.: one really believe one's hypothesis, and once getting the data there is ambiguity about how to analyze it …
- … When the first analysis tried does not give the desired result, keep trying until one finds a result that is.
  - By tweaking the data sample, fit model, fit ranges, etc.

5. Misuse of p-value drives bad science.

- In some scientific fields, p-values have become a litmus test for deciding which studies are worthy of publication.
  - E.g. social psychology: p-value < 0.05
  - → Research that produces p-values that surpass an arbitrary threshold are more likely to be published.

1. → The research goal shifts from seeking "the truth" to obtaining a p-value that clears the threshold.
2. → Researchers tend to dredge around in their data and keep trying different analyses until they find something with the right p-value.
3. This last phenomenon is called: ***p-value hacking.***
   - Clearly the p-value result has become meaningless!

Goodhart’s law

“**When a measure become a target, it is no longer a measure.**“

Translated to data analysis:

- When you start working towards a desired analysis result (read: desired p-value), the outcome becomes meaningless.

Example p-value hacking

- The question is: Which political party is best for the economy?
  - Play the game: “Hack Your Way To Scientific Glory”
  - http://fivethirtyeight.com/features/science-isnt-broken

- (~1000 out of ~1800 combinations yield a publishable p-value.)

- … That does not mean these tests showed that which party (Democrats / Republicans) was in office had a strong effect on the economy!
- Clearly, the p-value reveals almost nothing about the strength of the evidence.

Six ways to p-hack successfully!

1. Analyze many analysis measures, but report only those with $p<0.05$.
2. Collect and analyze many conditions, but only report those with $p<0.05$.
3. Stop collecting data once $p<0.05$
4. Cut on strongly correlated variable to get $p<0.05$.
5. Exclude participants to get $p<0.05$.
6. Transform the data to get $p<0.05$

6. Hypothesizing after results are known

- Avoid post-hoc theorizing at all times.
  - A.k.a. “Hark-ing”.

Meaning:

1. Something seems true in the (limited) data set available, therefore we hypothesize that it is true in general.
   - (Effect might be caused by statistical fluctuation.)
2. Then one (wrongly) tests it on the same (limited) data set, which seems to confirm that it is true!

- Circular reasoning! (A.k.a. “double dipping”)
- Clearly any conclusion drawn has become meaningless!

Harking example

1. Throwing a coin 10 times, with a result of 2 heads and 8 tails, might lead one to hypothesize that the coin favors tails by 4/5 to 1/5.
2. If this hypothesis is then tested on the *existing* data set, it is obviously confirmed!

- Clearly the confirmation is meaningless!
- And so is any p-value deduced!

So what is the proper procedure?

1. The proper procedure is to form in advance a hypothesis of what the tails probability is.
2. Then throw the coin various times to see if the hypothesis is rejected or not.

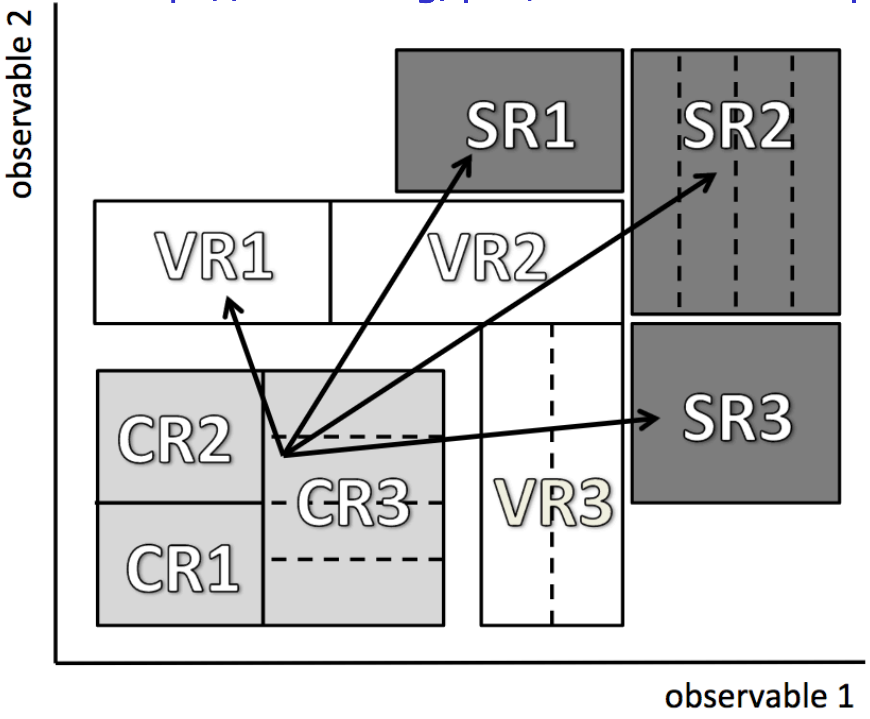
- If 8 tails and 2 heads are observed, another hypothesis could be formed: the tails probability is 4/5.
- Also this can only be tested by a new set of coin tosses.

Solutions to analysis pitfalls (1/3)

- Strict separation of training, validation and data samples.

http://arxiv.org/pdf/1410.1280v1.pdf

1. Develop your model on control samples.
2. Validate / test your mode on validation samples.
3. Apply your model to data sample(s) (“signal regions”) to obtain p-value.



- “Blind” your (signal region) dataset:
  Agree to only study it after analysis setup is complete.

Solutions to analysis pitfalls (2/3)

1. Strive at all costs to obtain an objective analysis result.

2. Never “optimize” your model on the same dataset used to obtain the p-value result.
   - If so the final result becomes meaningless.

3. Do not fall back on hypothesizing after the results are known.

4. Report everything you have done with the data:
   - how you determined your sample size,
   - all data exclusions (if any),
   - all manipulations and all measures in the study,
   - the exact definition of your (fit) model.

Solutions to analysis pitfalls (3/3)

- In general: one should not decide to publish based purely on p-value.

- Registration of all studies, including failed ones that do not get published.

Final remarks

- Significance tests do not protect against data dredging, p-value hacking, harking, etc !

- Important to realize that the statistical significance under the incorrect procedure(s) is completely spurious.

- Strive at all costs to obtain an objective analysis result.
- So be awareness of analysis pitfalls!
  - They are plentiful and deceptive.

- It is your responsibility to recognize them, to avoid them, and to point them out wherever you see them.

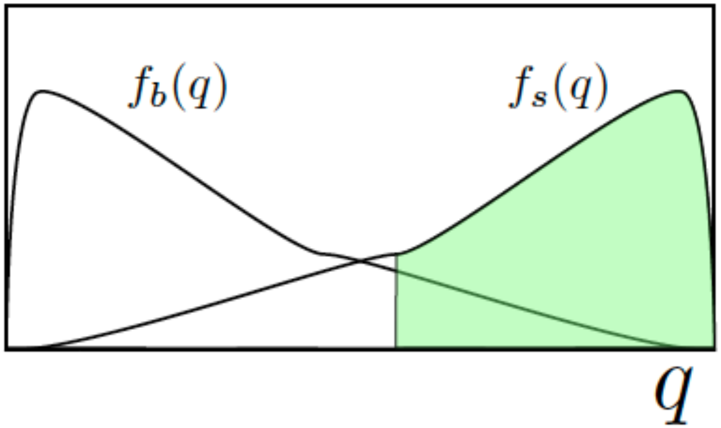
Le Fin

Merci beaucoup!

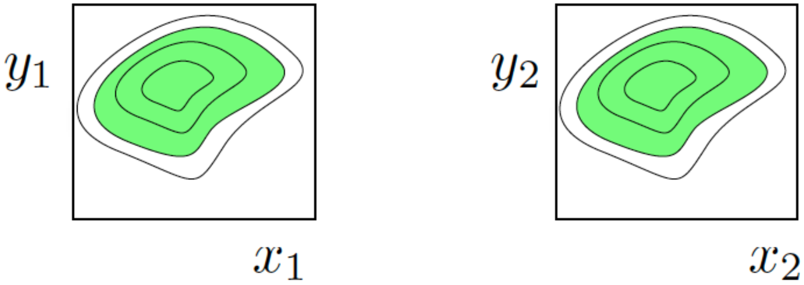
Backup

Two discriminating observables

- Often one uses the output of a neural network or multivariate algorithm *q* in place of a true likelihood ratio.
  - That's fine, but what do you do with it?



- If you have a fixed cut for all events ($q>q_{cut}$), this is what happens:
  - Example of two records:



Each records is treated separately.

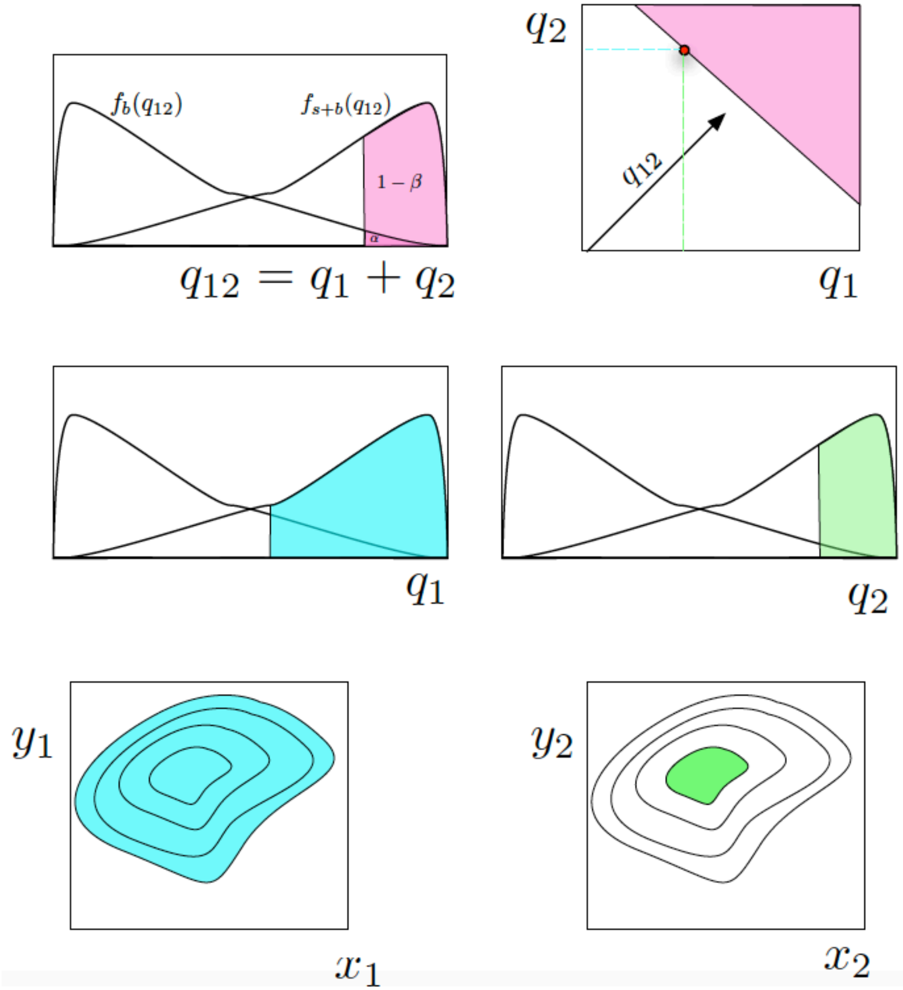
- *→ Always selecting events from within the same phase-space.*

Experiment versus Records

- Ideally, wish to cut on likelihood ratio for your entire experiment.
- ... instead of per record.

- Construct test statistic for experiment as sum over records:
  - $q_{12} = q_1 + q_2$ (etc)
  - Equivalent to a sum of LLRs

- Easy to see that includes experiments where one event had a high likelihood and the other one was relatively small.


$f_b(q_{12})$
$f_{s+b}(q_{12})$
$1-\beta$
α
$q_{12} = q_1 + q_2$
q_2
q_{12}
q_1
q_1
q_2
y_1
x_1
y_2
x_2