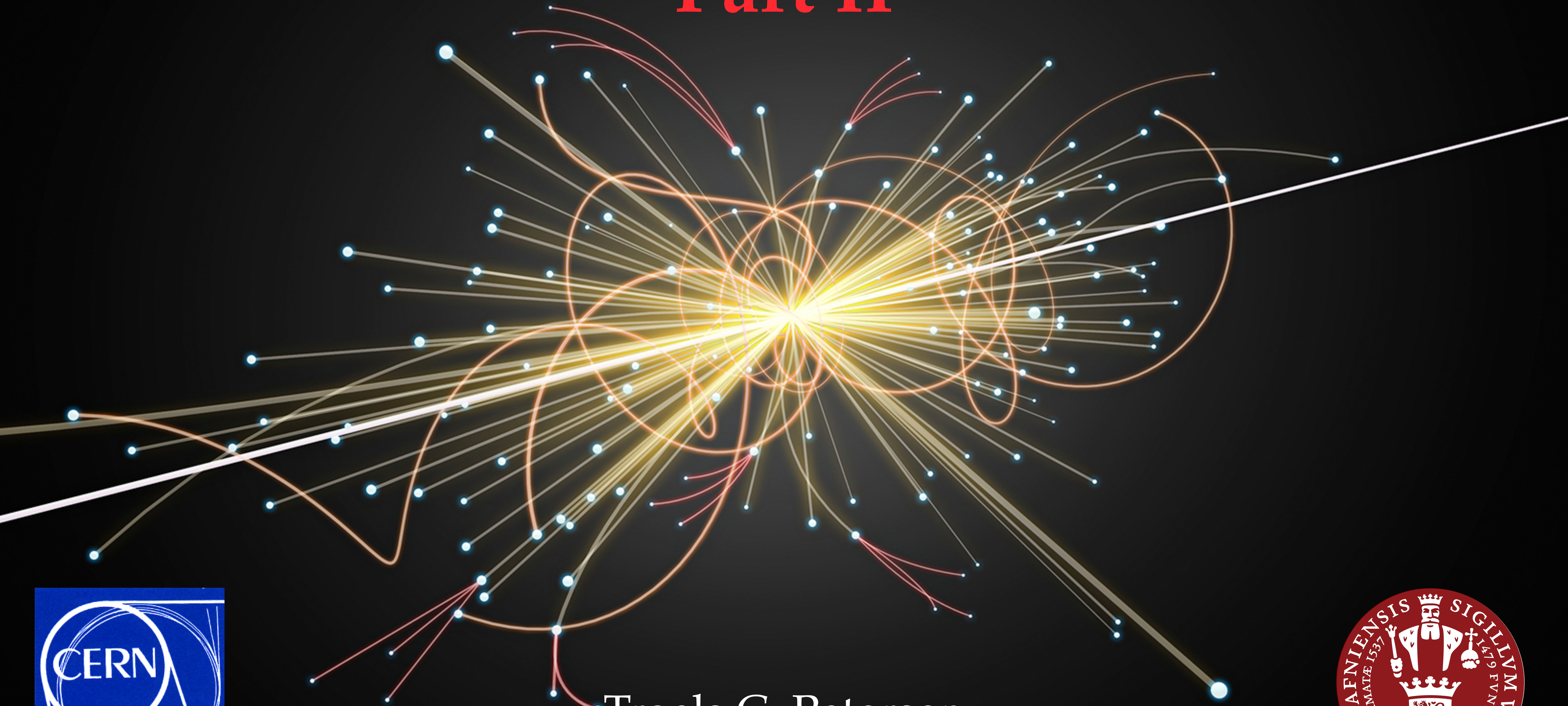
Practical Statistics

Part II



Troels C. Petersen

Niels Bohr Institute, Copenhagen

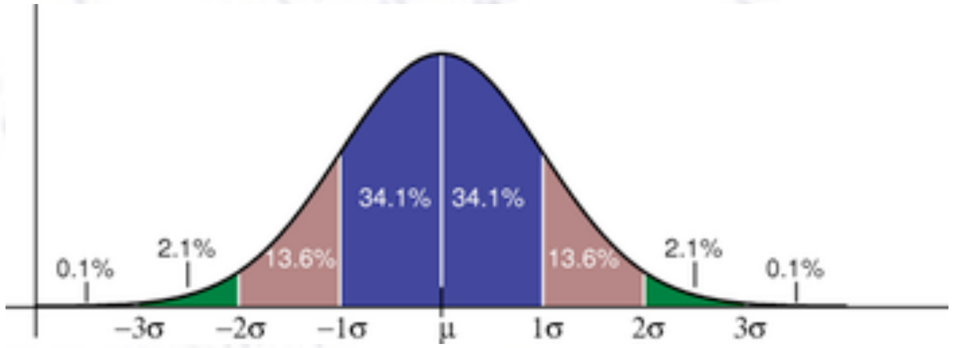
European School of High Energy Physics, 13th of September 2023

Practical Statistics

Part II - the necessities

Likelihood fitting, Hypothesis testing, and Systematic uncertainties

Troels C. Petersen (Niels Bohr Institute)



"Statistics is merely a quantisation of common sense"

European School of High Energy Physics, 13th of September 2023

Outline of lectures

Part I - the basics:

- Estimators
- Probability Density Functions
- ChiSquare & p-values
- Calibration
- Simpson’s Paradox

Part II - the necessities:

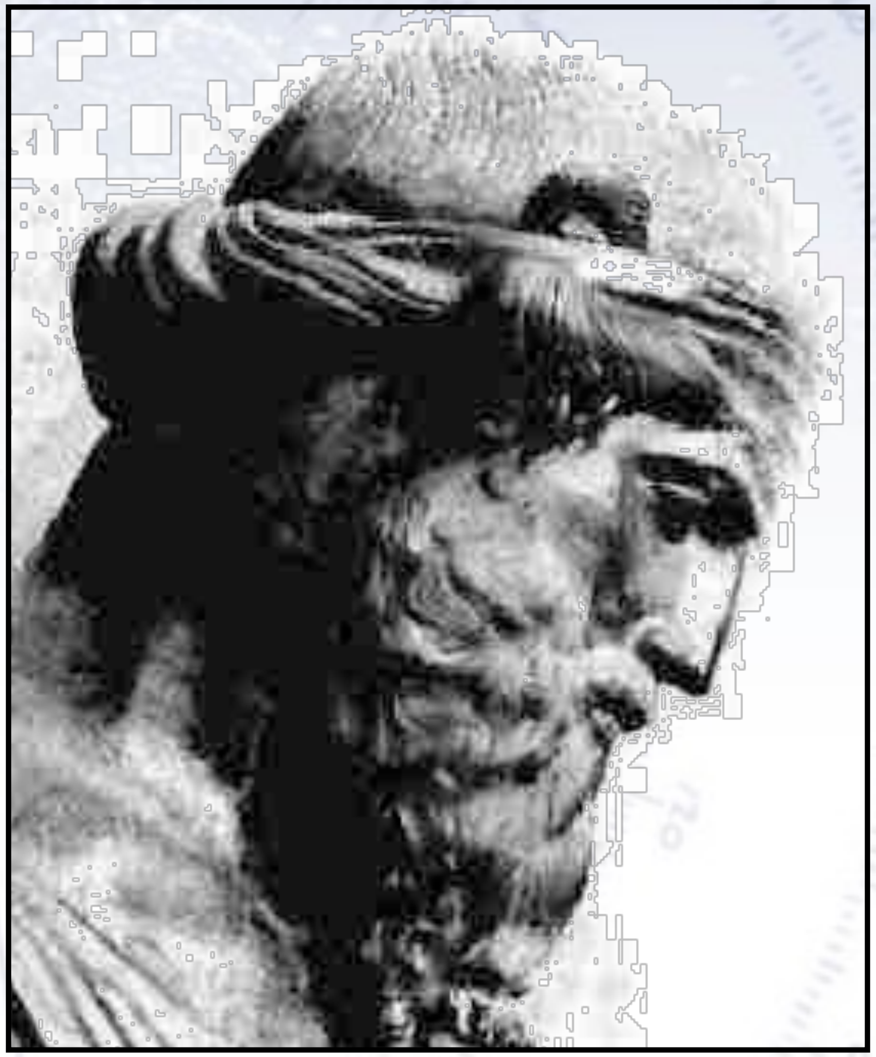
- Likelihood fitting
- Hypothesis testing
- Systematic uncertainties

Part III - the cool:

- Setting limits
- Look Elsewhere Effect
- The art of plotting
- The Fisher discriminant
- sPlots & sWeights

Likelihood Principle

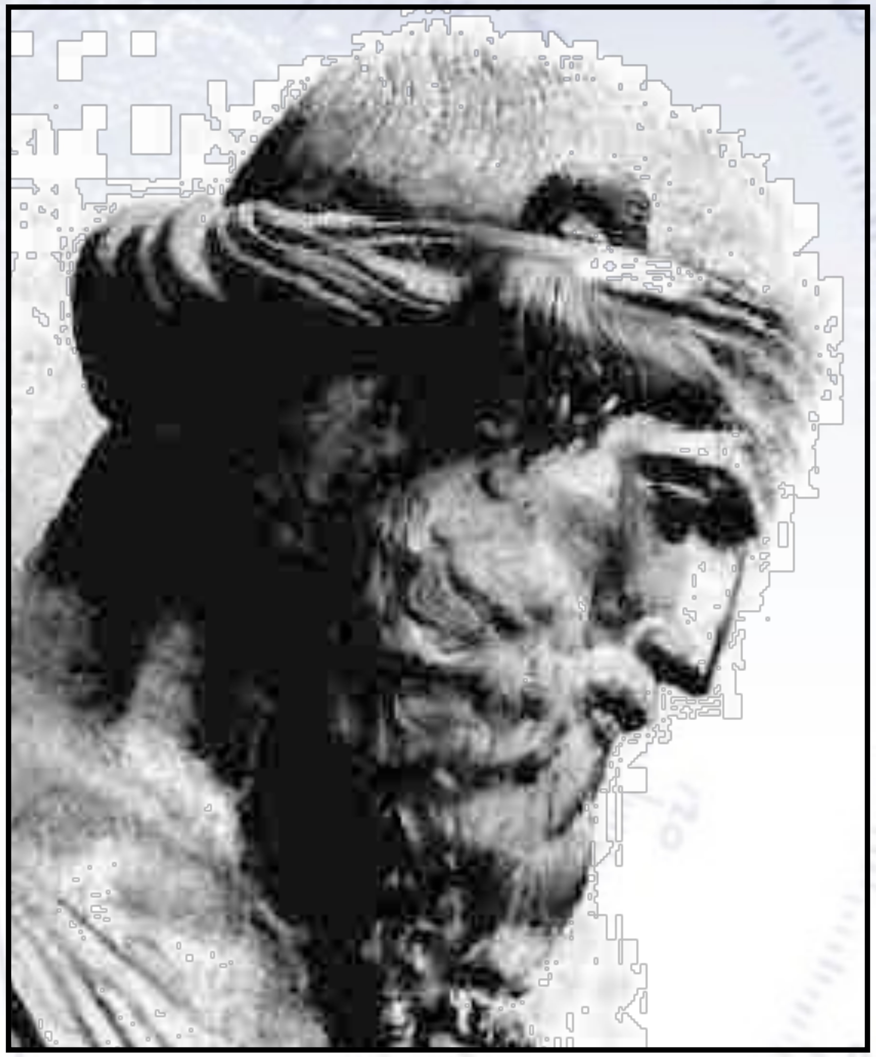
Likelihood function



"I shall stick to the principle of likelihood…"

[Plato, in Timaeus]

Likelihood function



Given a PDF f(x) with parameter(s) θ, what is the chance that with N observations, x_i falls in the intervals [x_i, $x_i + dx_i$]?

$$\mathcal{L}(\theta) = \prod_i f(x_i, \theta) dx_i$$

Likelihood function

Given a set of measurements **x**, and parameter(s) θ, the likelihood function is defined as:

$$\mathcal{L}(x_1, x_2, \ldots, x_N; \theta) = \prod_i p(x_i, \theta)$$

The **principle of maximum likelihood** for parameter estimation consist of maximising the likelihood of parameter(s) (here θ) given some data (here **x**).
There is nothing strange about this - it is exactly the same we do for the ChiSquare!

The likelihood function plays a central role in statistics, as it can be shown to be:
- Consistent (converges to the right value).
- Asymptotically normal (converges with Gaussian errors).
- Efficient (reaches the Minimum Variance Bound (MVB, Cramer-Rao) for large N).

$$V(\hat{a}) \geq \frac{1}{<(d \ln L/da)^2>}$$

To some extend, this means that the likelihood function is "optimal", that is, if it can be applied in practice.

Likelihood vs. Chi-Square

For computational reasons, it is often much easier to minimise the logarithm of the likelihood function:

$$\left.\frac{\partial \ln \mathcal{L}}{\partial \theta}\right|_{\theta=\bar{\theta}} = 0$$

In problems with Gaussian errors, it turns out that the **log likelihood function** boils down to the **Chi-Square** with a constant offset and a factor -2 in difference.

See Barlow 5.6

The likelihood function for fits comes in two versions:

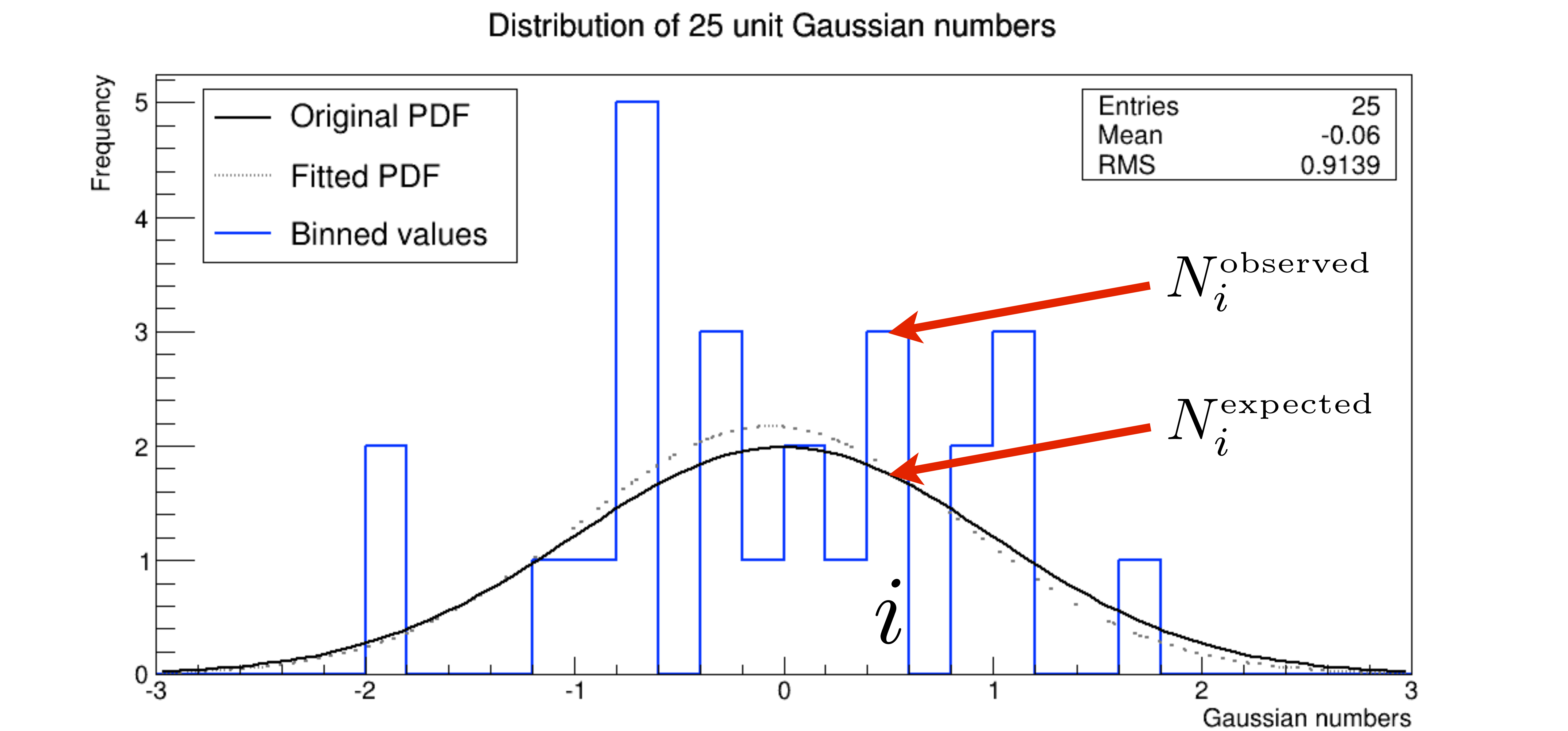
- Binned likelihood (using Poisson) for histograms.
- Unbinned likelihood (using PDF) for single values.

The “trouble” with the likelihood is, that it is unlike the Chi-Square, there is NO simple way to obtain a probability of obtaining certain likelihood value!

ChiSquare

Recall, the ChiSquare is a sum over bins in a histogram:

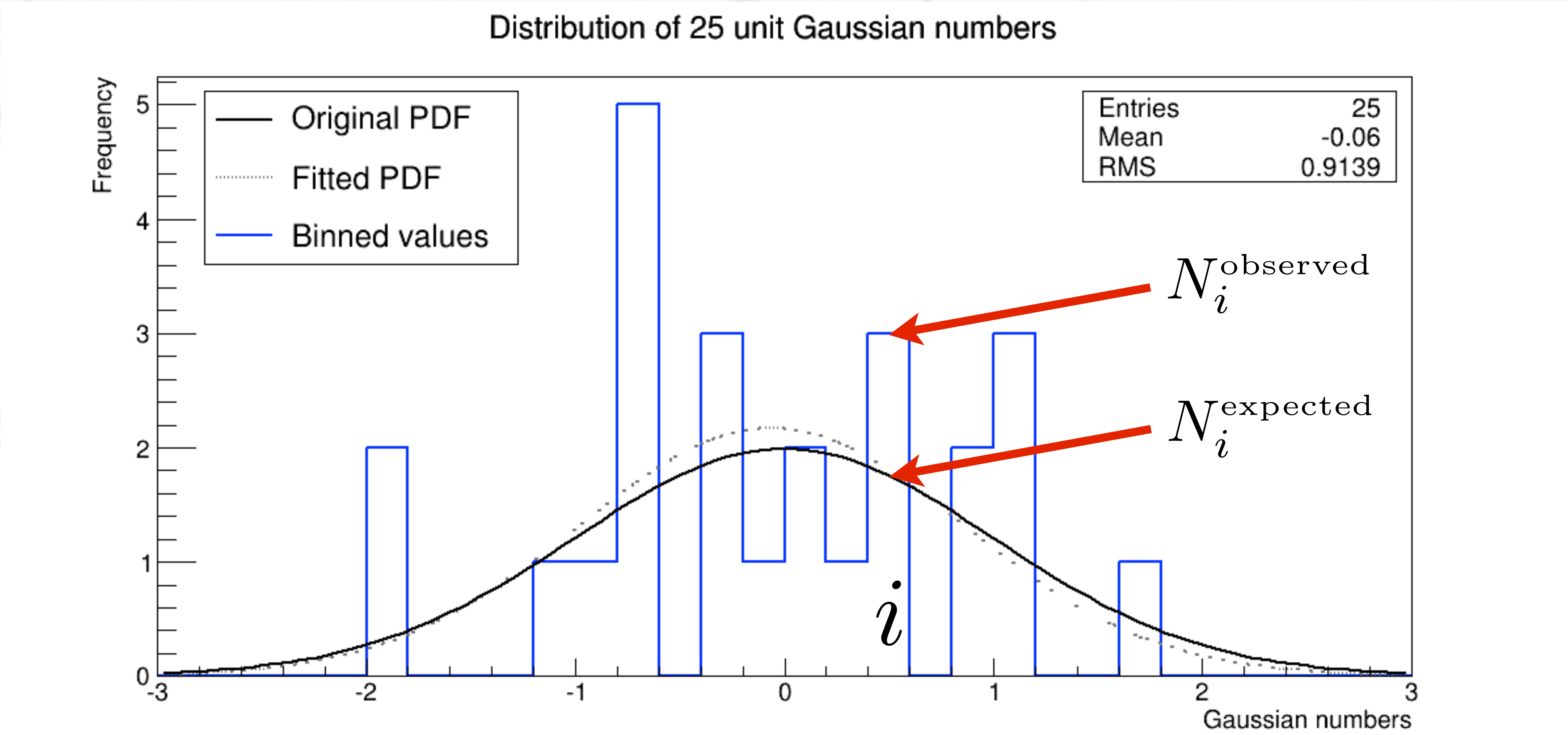
$$\chi^2(\theta) = \sum_{i}^{N_{\text{bins}}} \left(\frac{N_i^{\text{observed}} - N_i^{\text{expected}}}{\sigma(N_i^{\text{observed}})} \right)^2$$



ChiSquare

Recall, the ChiSquare is a sum over bins in a histogram:

$$\chi^2(\theta) = \sum_{i}^{N_{\text{bins}}} \frac{(N_i^{\text{observed}} - N_i^{\text{expected}})^2}{N_i^{\text{expected}}}$$

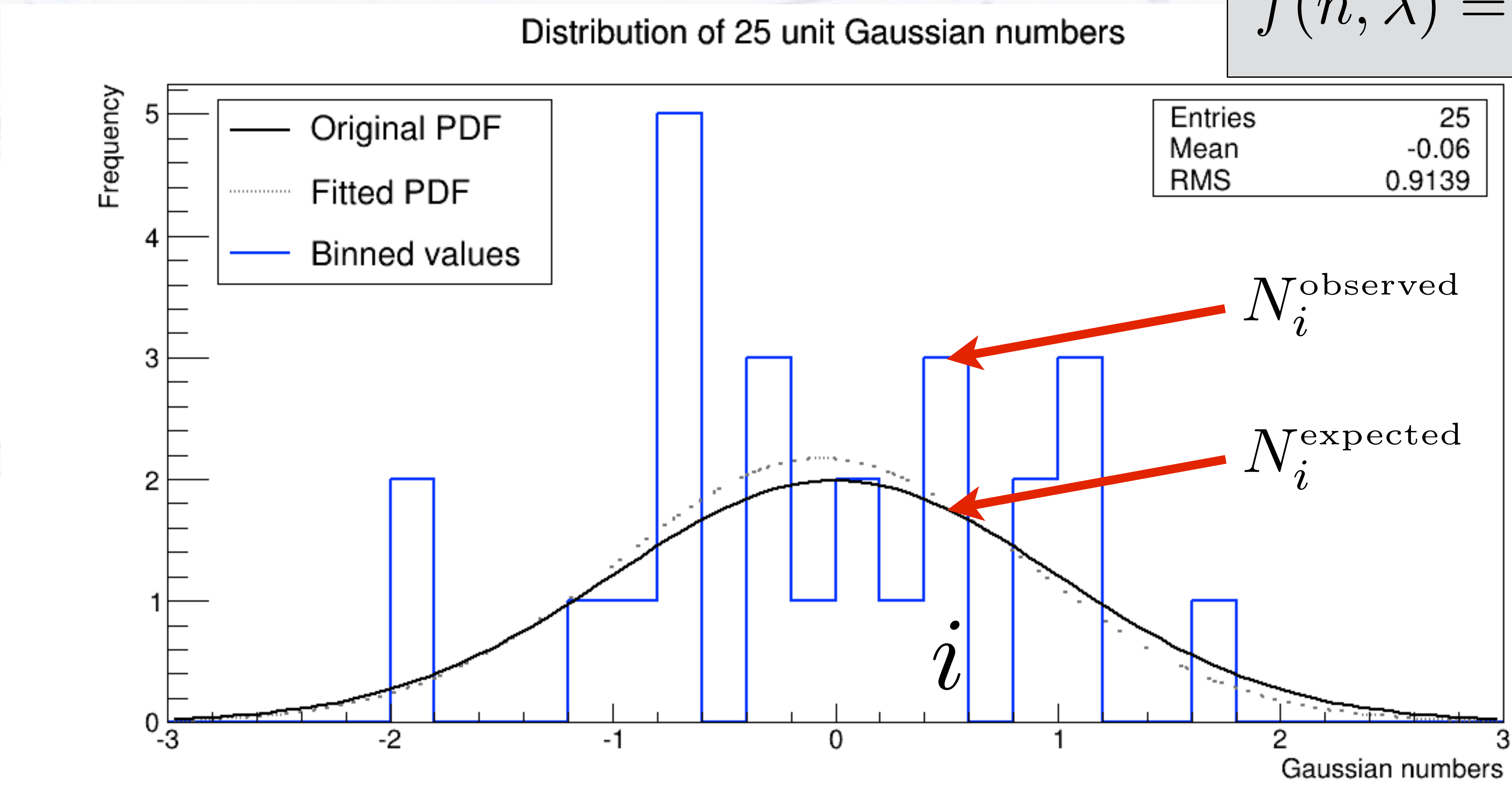


Binned Likelihood

The binned likelihood is a sum over bins in a histogram:

$$\mathcal{L}(\theta)_{\text{binned}} = \prod_{i}^{N_{\text{bins}}} \text{Poisson}(N_i^{\text{expected}}, N_i^{\text{observed}})$$

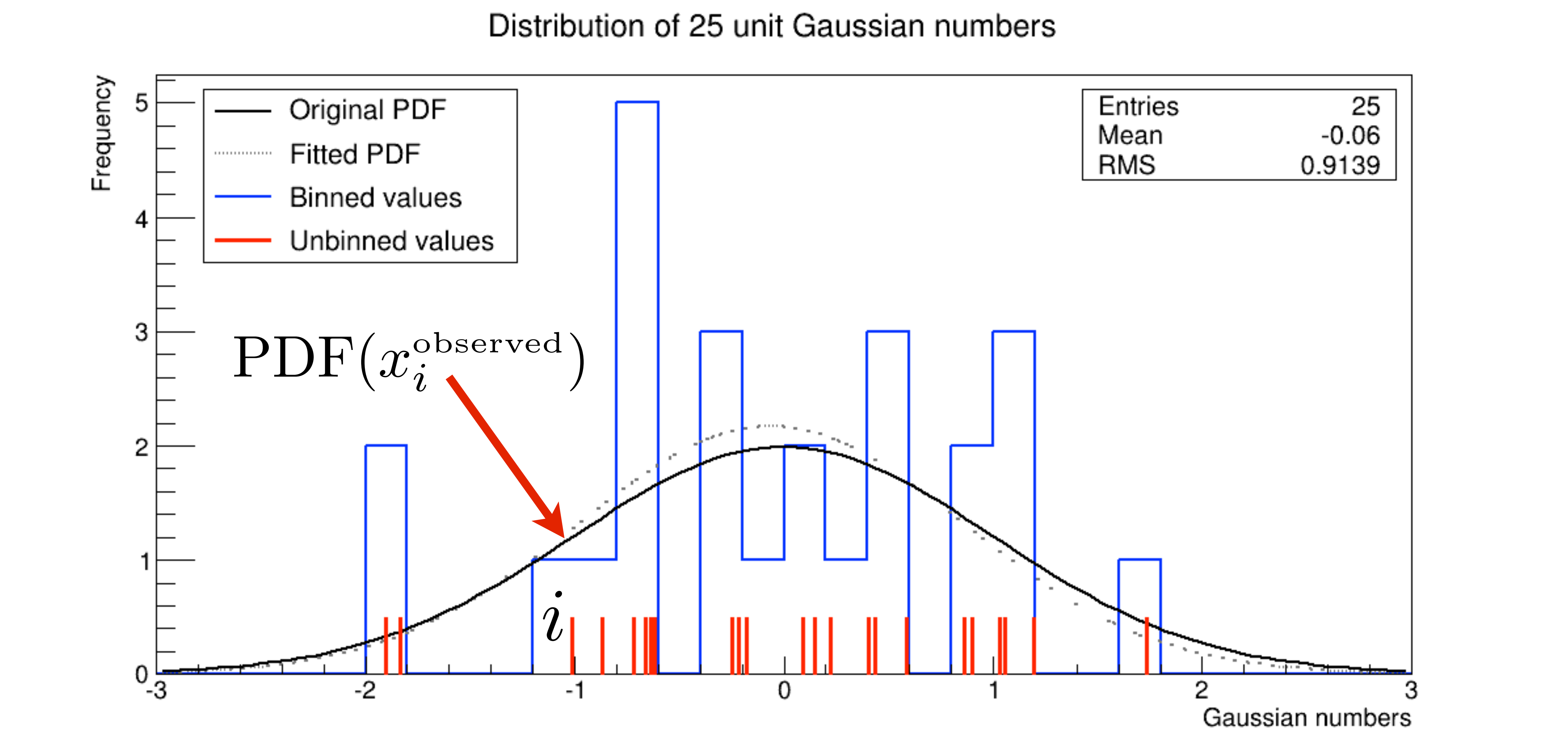
$$\boxed{f(n,\lambda) = \frac{\lambda^n}{n!}e^{\lambda}}$$



Unbinned Likelihood

The binned likelihood is a sum over single measurements:

$$\mathcal{L}(\theta)_{\text{unbinned}} = \prod_{i}^{N_{\text{meas.}}} \text{PDF}(x_i^{\text{observed}})$$



$\text{PDF}(x_i^{\text{observed}})$

Methods of fitting

In summary, there are four methods of fitting histograms with parameters θ, in order of increasing accuracy, but decreasing speed and convergence:

1. **Minimise the ("Neyman") Chi-Square:**
   Problem: Breaks in empty bins ($N^{obs} = 0$).
   Note: Minuit disregards empty bins!

$$\chi^2(\theta) = \sum_i^{N_\mathrm{bins}} \frac{(N_i^\mathrm{Obs.} - N_i^\mathrm{Exp.}(\theta))^2}{N_i^\mathrm{Obs.}}$$

2. **Minimise the ("Pearson") Chi-Square:**
   Minor problem: What range to include?
   Note on 1+2: Applies only to histograms.
   If errors are provided, these are used directly.

$$\chi^2(\theta) = \sum_i^{N_\mathrm{bins}} \frac{(N_i^\mathrm{Obs.} - N_i^\mathrm{Exp.}(\theta))^2}{N_i^{\mathrm{Exp.}(\theta)}}$$

3. **Minimise -2Ln(LLH) of each bin (Poisson):**
   Note: This can be used for low statistics binned data, avoiding the Gaussian approx.

$$-2\ln\mathcal{L}(\theta)_\mathrm{binned} =$$
$$-2\sum_{i\in N_\mathrm{bins}} \ln \mathrm{Pois}(N_i^\mathrm{Obs.}, N_i^\mathrm{Exp.}(\theta))$$

4. **Drop binning and minimise the unbinned -2Ln(LLH) likelihood.**
   Note: Sum runs over events not bins!
   Note: Fit parameters in PDF.

$$-2\ln\mathcal{L}(\theta)_\mathrm{unbinned} =$$
$$-2\sum_{i\in N_\mathrm{events}} \ln \mathrm{PDF}(N_i^\mathrm{Obs.})$$

The unbinned likelihood is generally the best method in case of low statistics.

Methods of fitting

The binned likelihood expression is as follows:

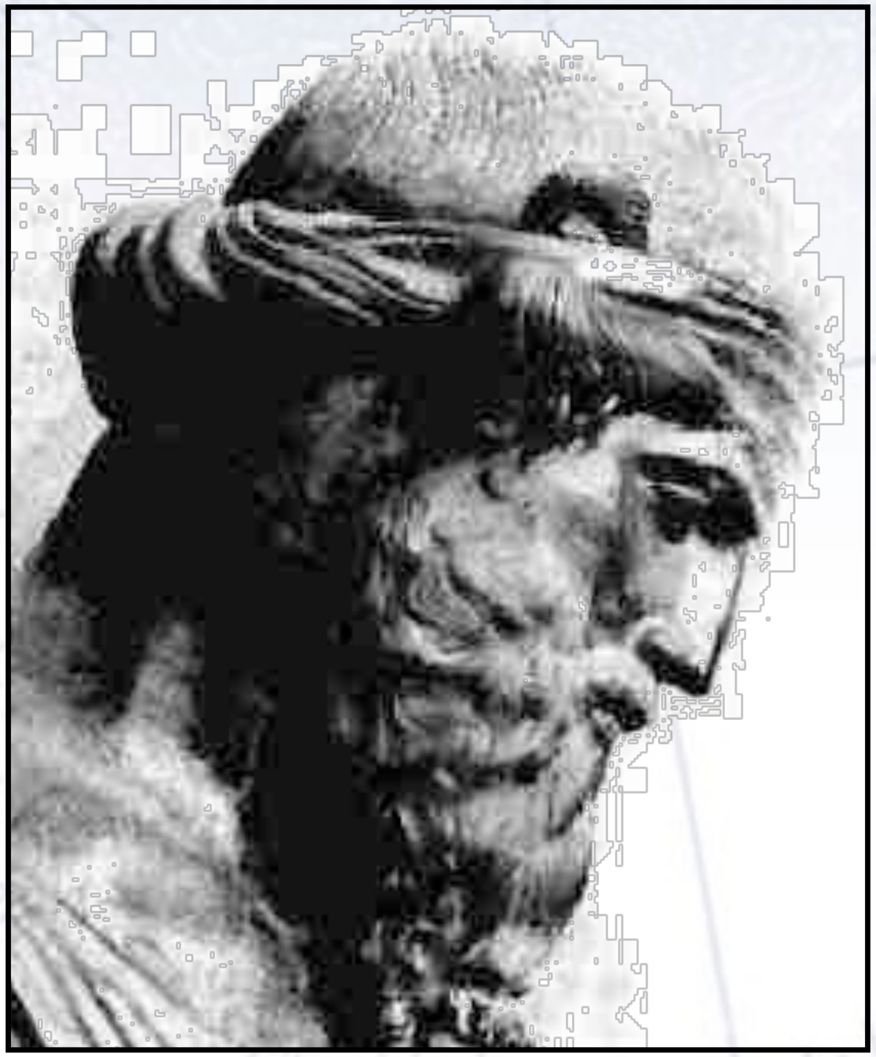
$$-2\ \ln \mathcal{L}(\theta) = -2 \sum_{i}^{N_{\text{bins}}} \ln \text{Pois}(N_i^{\text{Obs.}}, N_i^{\text{Exp.}}(\theta))$$

$$= -2 \sum_{i}^{N_{\text{bins}}} (N_i^{\text{Obs.}} \ln(N_i^{\text{Exp.}}(\theta)) - N_i^{\text{Exp.}}(\theta) - \ln(N_i^{\text{Obs.}}!))$$

$$= -2 \sum_{i}^{N_{\text{bins}}} N_i^{\text{Obs.}} \ln(N_i^{\text{Exp.}}(\theta)) + 2 \sum_{i}^{N_{\text{bins}}} N_i^{\text{Exp.}} + C$$

The middle term is simply the fitted content of all bins, which (along with the final term) is independent of the fit parameters θ, and can thus be dropped in the minimisation.

Notes on the likelihood



For a large sample, the likelihood is indeed unbiased and has the minimum variance - that is hard to beat! Also, the binned LLH approaches the unbinned version.

However, for the likelihood, unlike for the Chi-Square, **you get no goodness-of-fit measure to check it**!
And for large samples, the approximation that bin count uncertainties are Gaussian is good.

For small statistics, the likelihood is not necessarily unbiased, but still fares much better than the ChiSquare! **But be careful with small statistics.** The way to avoid this problem is also simulation.

Hypothesis Testing

Hypothesis testing

Suppose in a beer tasting, that someone gets 9 our of 10 right.

Does that prove that the person can taste difference between beers?

Hypothesis testing

Suppose in a beer tasting, that someone gets 9 our of 10 right.

Does that prove that the person can taste difference between beers?

NO!

What we can say is that the result is **inconsistent** (at some significance level) with the hypothesis that the person chooses at random.

This leaves us with the alternative hypotheses, that the person can taste the difference or have cheated (consciously or unconsciously).

In statistics one can never prove a hypothesis directly. However, one can set up alternative hypotheses and disprove these. That is how one works in statistics…

See Barlow Chapter 8, in particular 8.2.1 (p. 146)

Hypothesis testing

Hypothesis testing is like a criminal trial. The basic "null" hypothesis is **Innocent** (called H_0) and this is the hypothesis we want to test, compared to an "alternative" hypothesis, **Guilty** (called H_1).

Innocence ("negative") is initially assumed, and this hypothesis is only rejected, if enough evidence proves otherwise, i.e. that the probability of innocence is very small ("beyond reasonable doubt").

	Truly innocent **(H_0 is true)**	**Truly guilty** **(H_1 is true)**
Acquittal **(Accept H_0)**	**Right decision** **True Negative (TN)**	**Wrong decision** **False Negative (FN)**
Conviction **(Reject H_0)**	**Wrong decision** **False Positive (FP)**	**Right decision** **True Positive (TP)**

Hypothesis testing

Hypothesis testing is like a criminal trial. The basic "null" hypothesis is **Innocent** (called H_0) and this is the hypothesis we want to test, compared to an "alternative" hypothesis, **Guilty** (called H_1).

Innocence ("negative") is initially assumed, and this hypothesis is only rejected, if enough evidence proves otherwise, i.e. that the probability of innocence is very small ("beyond reasonable doubt").

	Truly innocent **(H_0 is true)**	**Truly guilty** **(H_1 is true)**
Acquittal **(Accept H_0)**	**Right decision** **True Negative (TN)**	Type II error, β **Wrong decision** **False Negative (FN)**
Conviction **(Reject H_0)**	Type I error, α **Wrong decision** **False Positive (FP)**	**Right decision** **True Positive (TP)**

The rate of type I/II errors are correlated, and one can only choose one of these!

Hypothesis terminology

H_0 = Null Hypothesis:

Definition: The initial/simplest hypothesis.

Examples: Data is background, data follows simple model, particle is a pion.

H_1 = Alternative Hypothesis:

Definition: The alternative to the null hypothesis, possibly more advanced.

Examples: Data is background + signal, data does not follows simple model, particle is an electron.

α = False Positive Rate (Significance):

Definition: Probability to **reject H_0**, even if it is **true** (aka. "False Positive").

Example: Finding guilty when innocent. Concluding no signal, even if there.

Note: The signal selection efficiency = $1 - \alpha$

β = False Negative Rate (1 - Power):

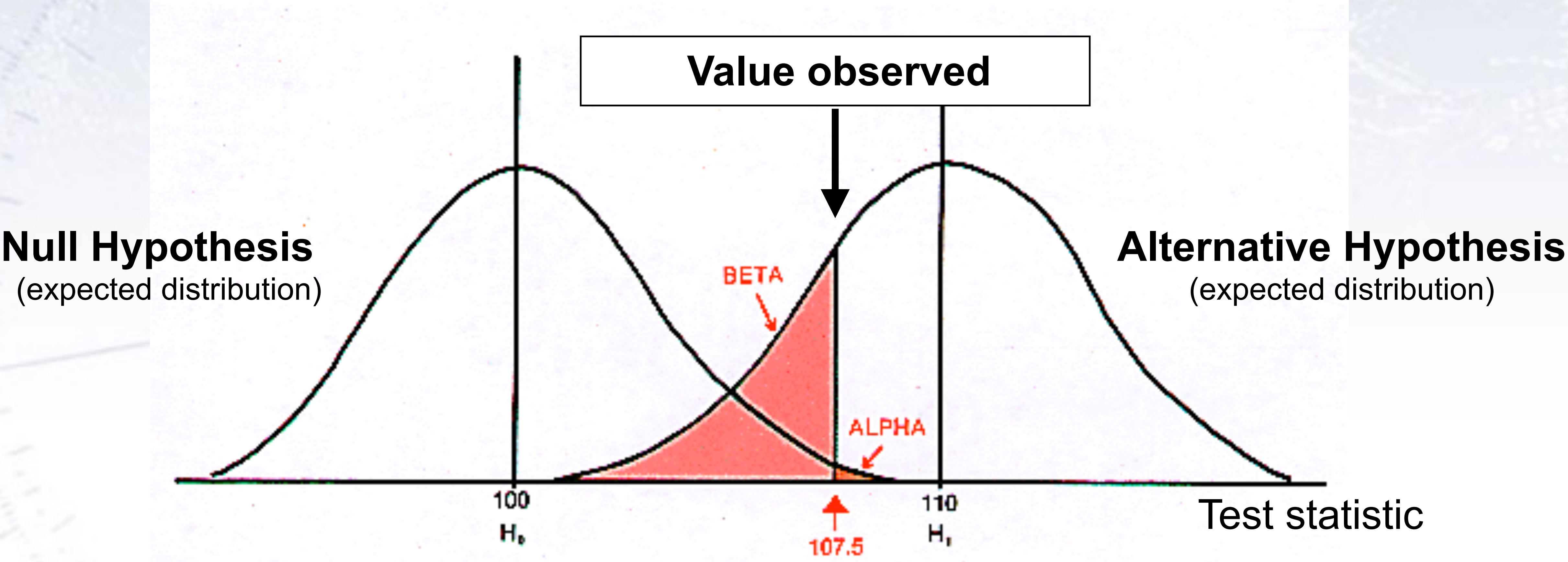
Definition: Probability to **accept H_0**, even if it is **false** (aka. "False Negative").

Example: Acquitting, when guilty. Concluding signal, even if not there.

Note: The misidentification probability = β

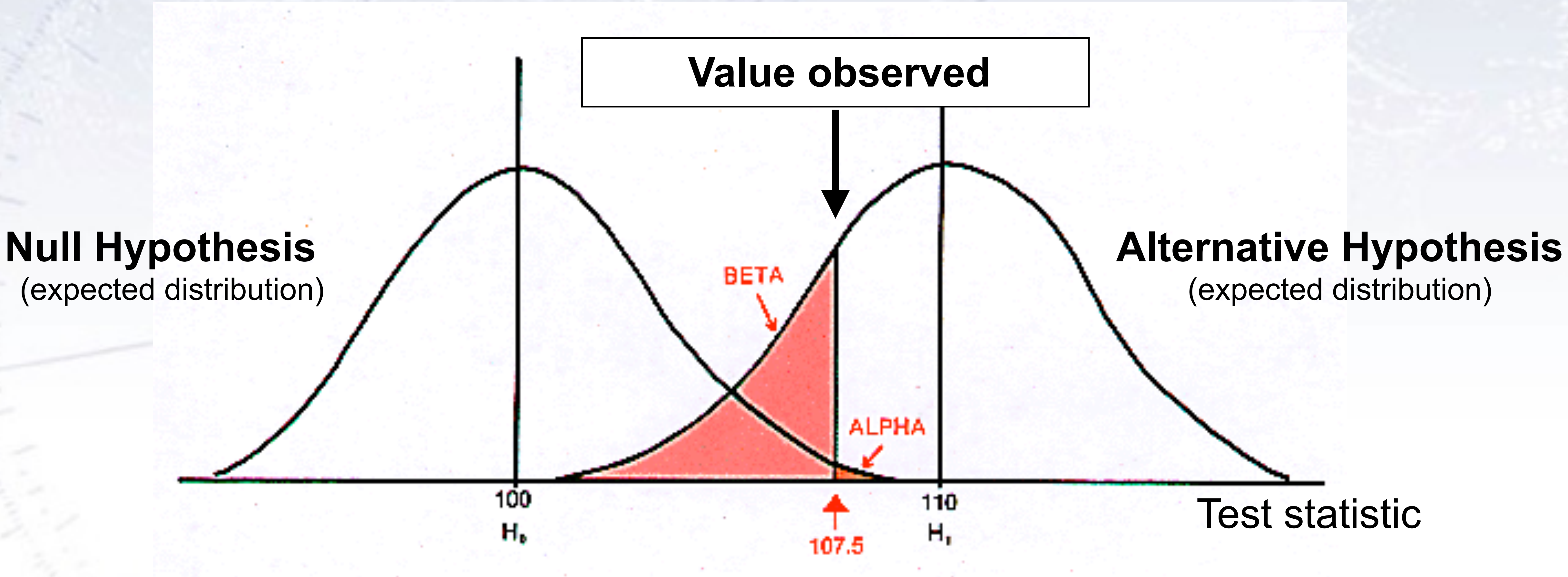
Taking decisions

You are asked to take a decision: **Given data - how to do that best?**



Taking decisions

You are asked to take a decision: **Given data - how to do that best?**



Value observed

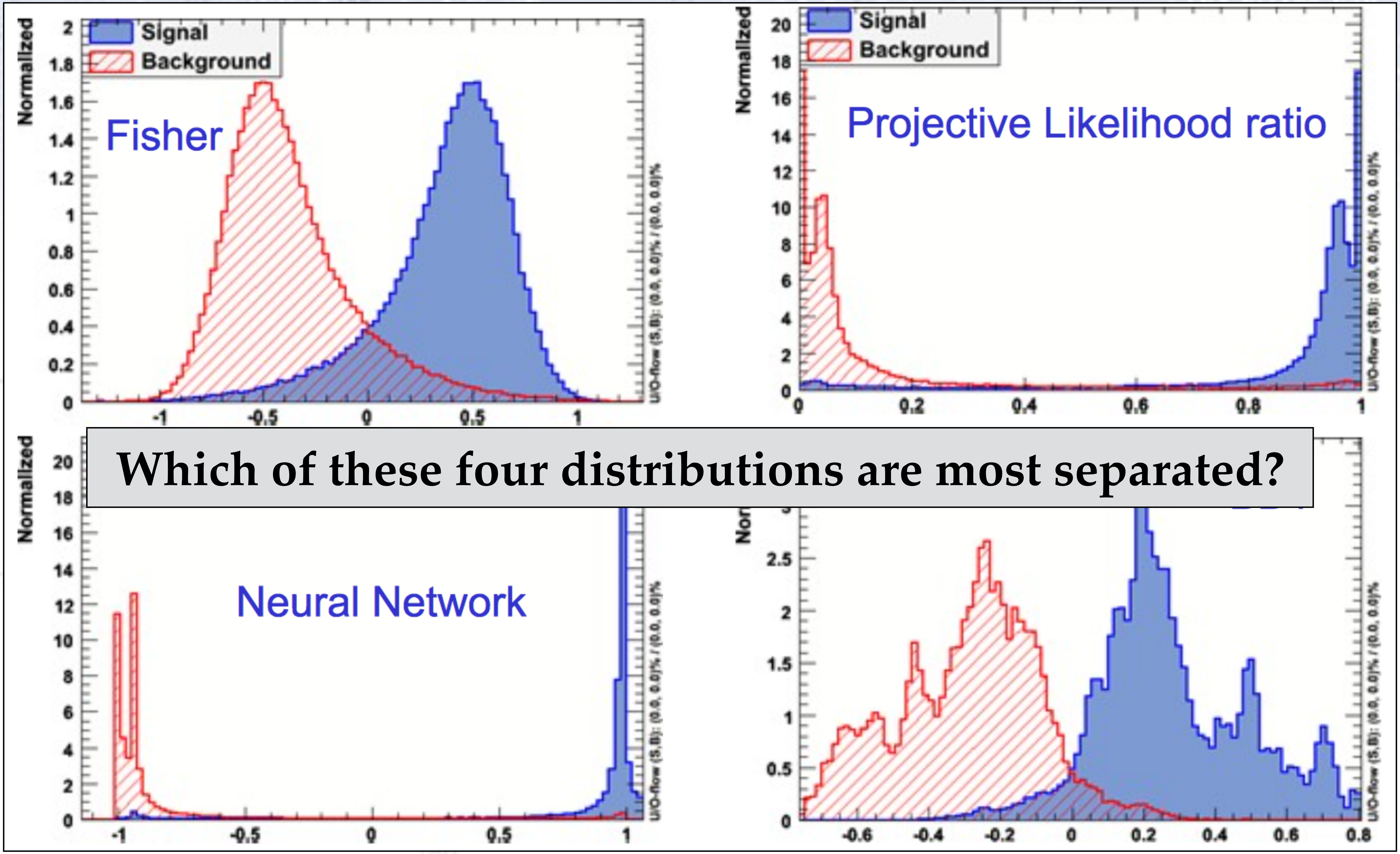
Null Hypothesis
(expected distribution)

Alternative Hypothesis
(expected distribution)

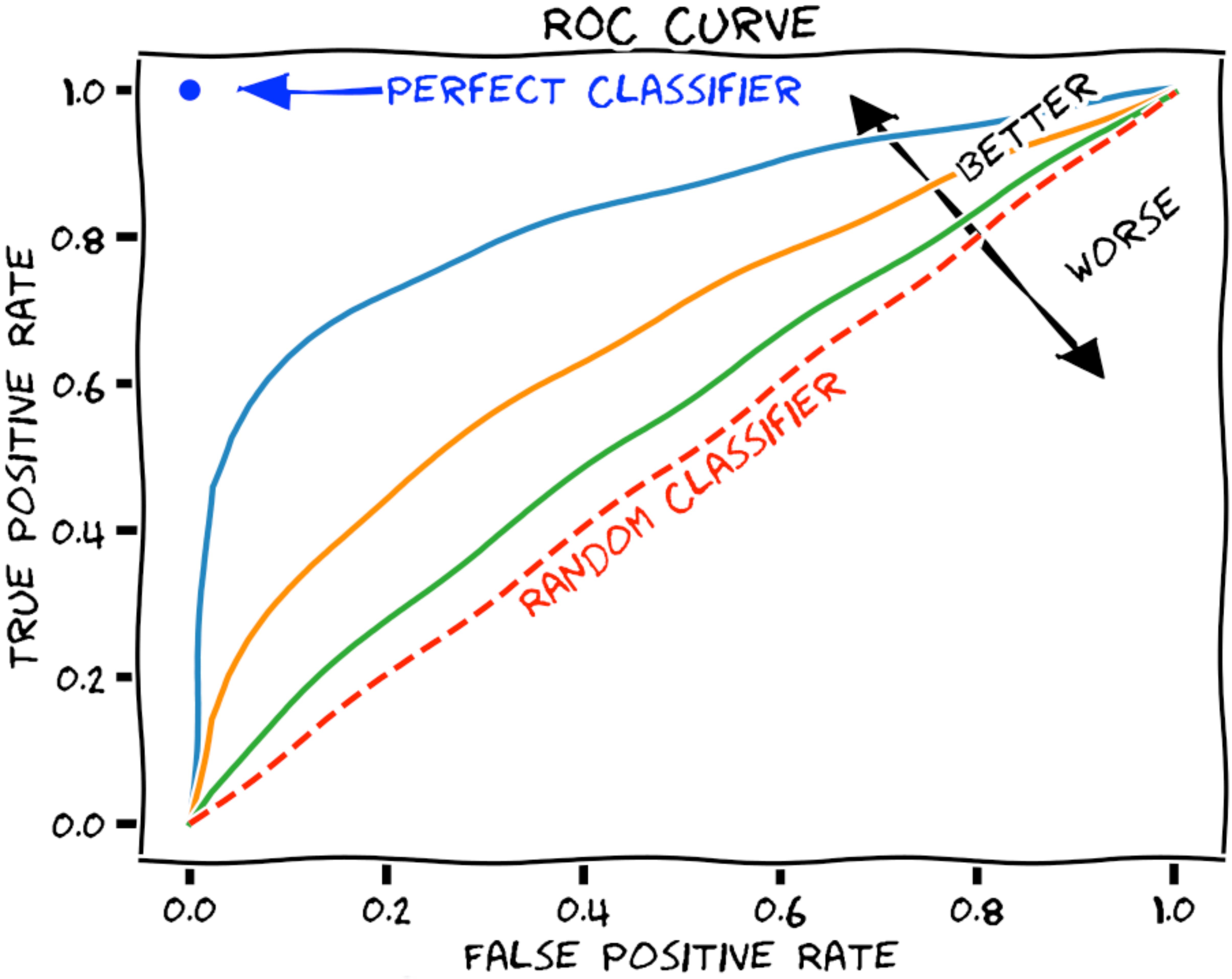
Test statistic

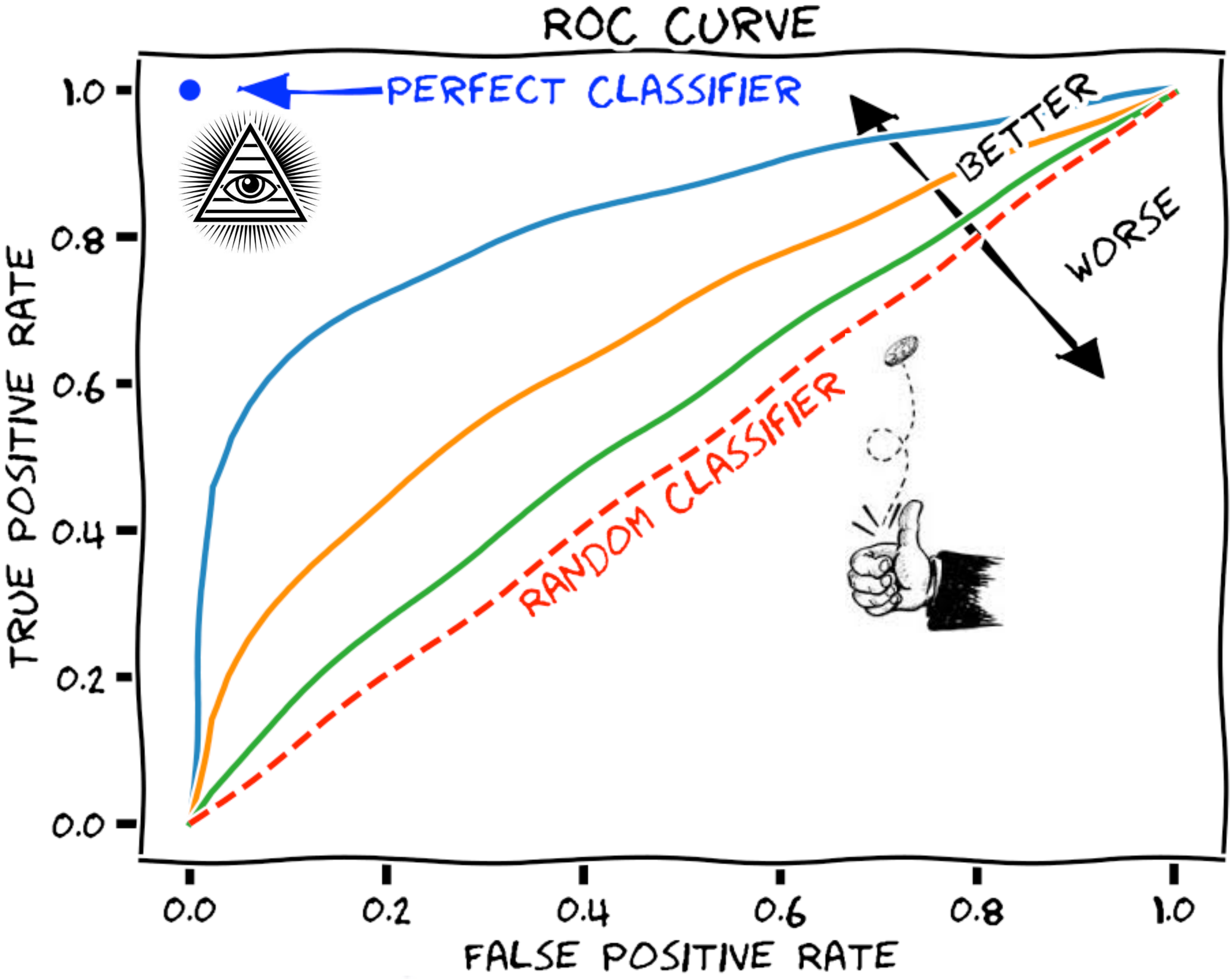
The purpose of a **test** is to yield (calculable/predictable) distributions for the **Null** and **Alternative hypotheses**, which are ***as separated from each other as possible*** (in order to minimise α and β).

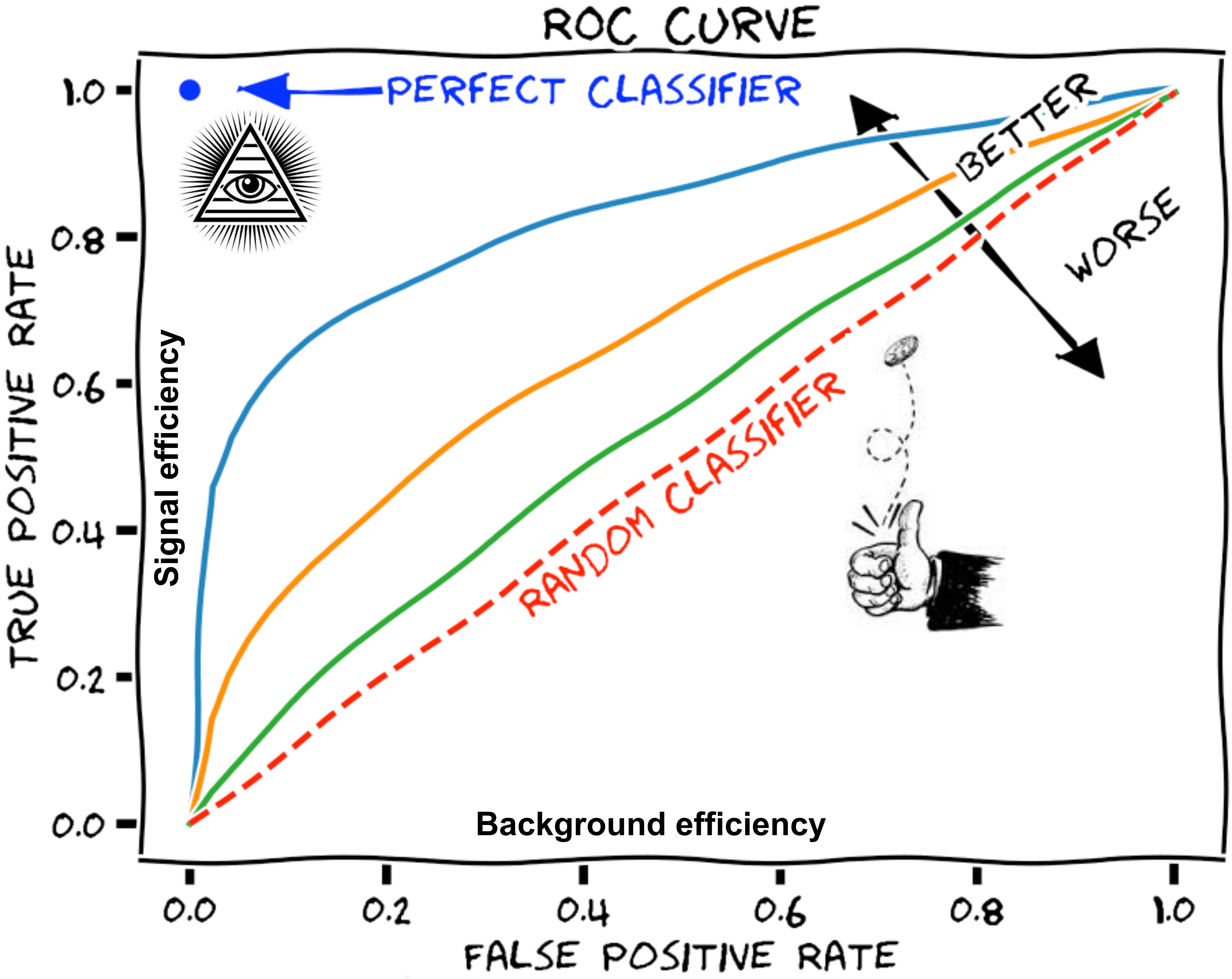
Measuring separation

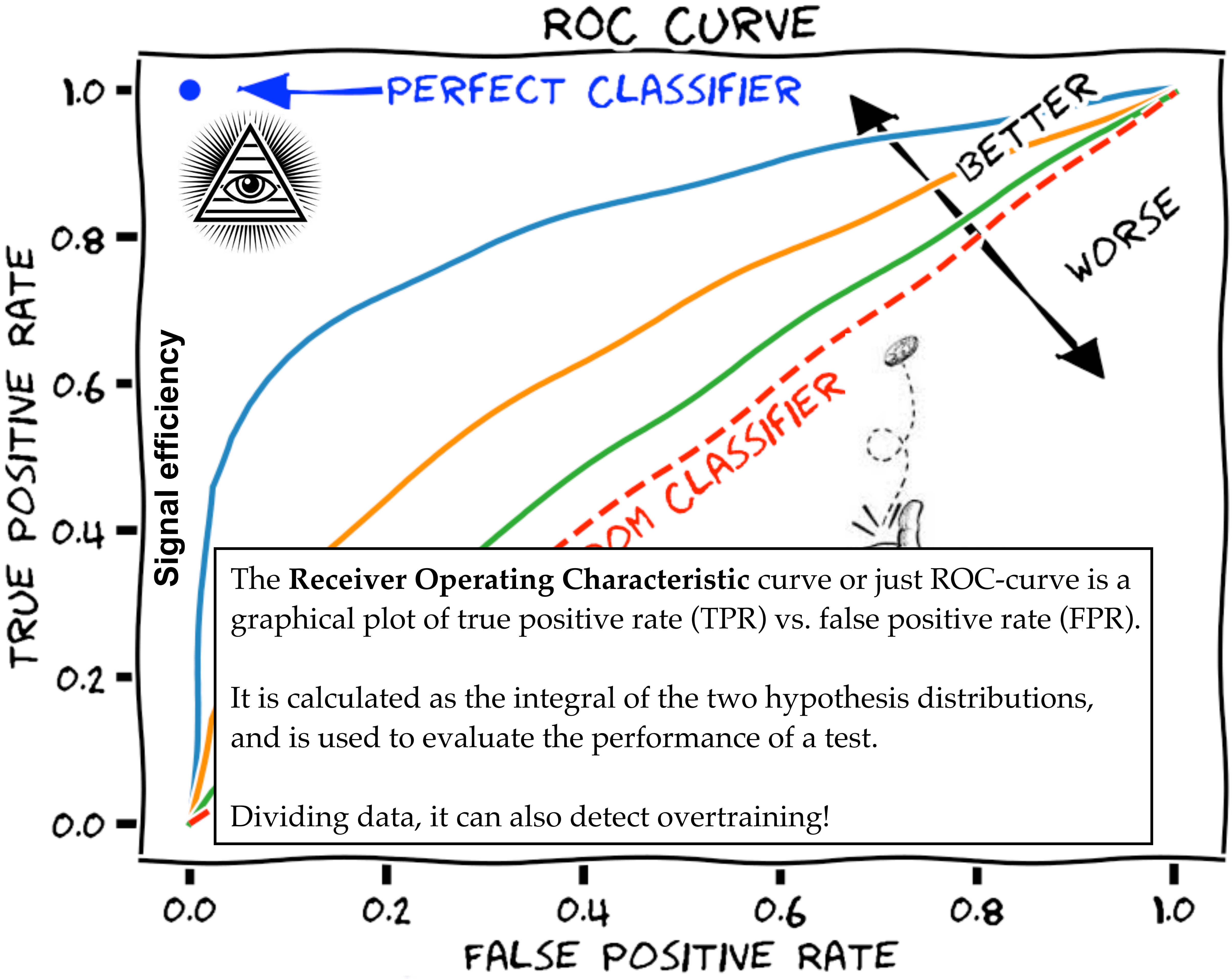


Which of these four distributions are most separated?


ROC CURVE
PERFECT CLASSIFIER
BETTER
WORSE
RANDOM CLASSIFIER
TRUE POSITIVE RATE
FALSE POSITIVE RATE
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


ROC CURVE
PERFECT CLASSIFIER
BETTER
WORSE
RANDOM CLASSIFIER
TRUE POSITIVE RATE
FALSE POSITIVE RATE
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


ROC CURVE
PERFECT CLASSIFIER
BETTER
WORSE
RANDOM CLASSIFIER
Signal efficiency
Background efficiency
TRUE POSITIVE RATE
FALSE POSITIVE RATE
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0



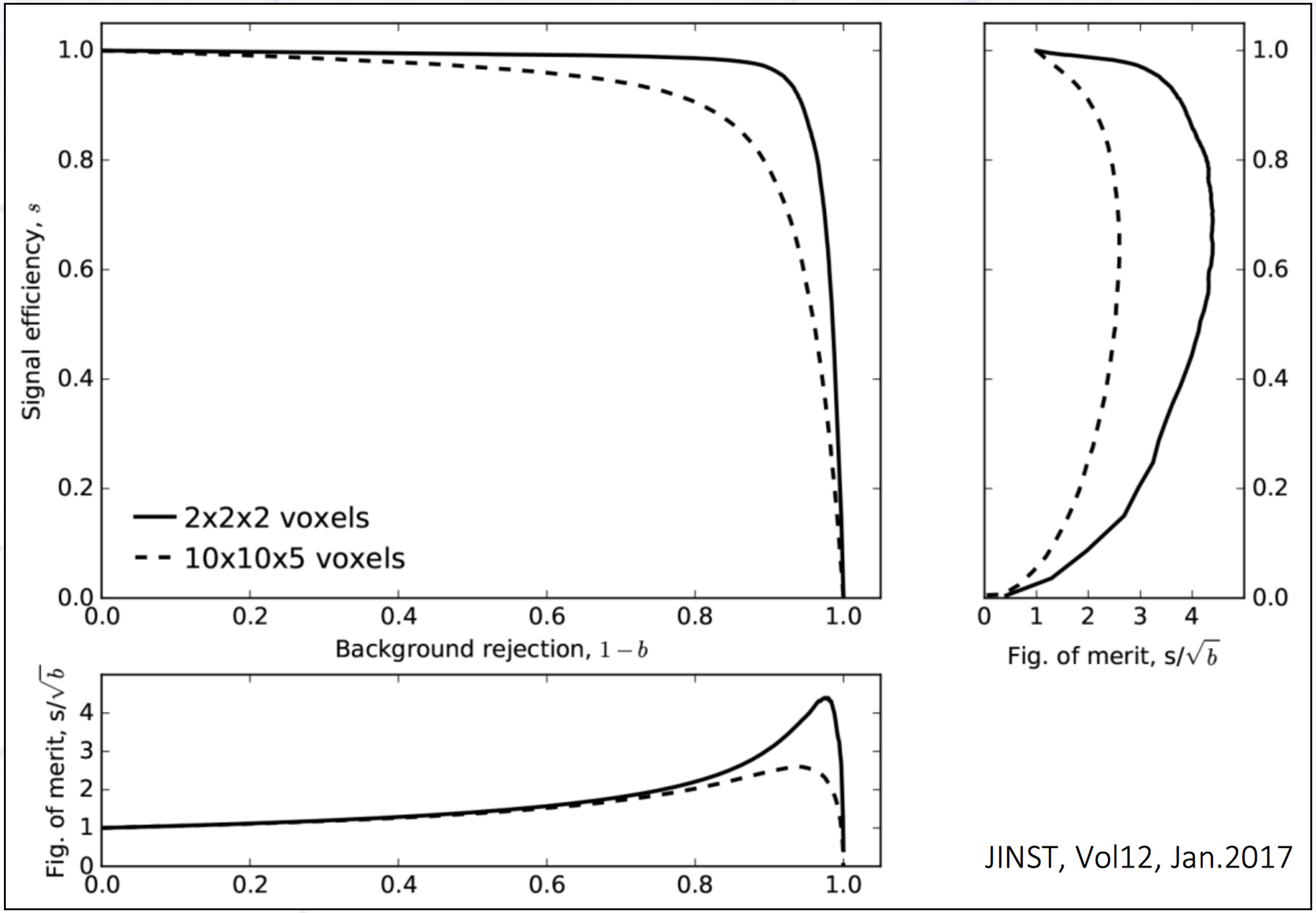
The **Receiver Operating Characteristic** curve or just ROC-curve is a graphical plot of true positive rate (TPR) vs. false positive rate (FPR).

It is calculated as the integral of the two hypothesis distributions, and is used to evaluate the performance of a test.

Dividing data, it can also detect overtraining!

Where to select?

The ROC curve does **not** tell you **where** to make your selection. You have to figure that out. In searches for signal (S) in background (B), optimising S/sqrt(B) or S/sqrt(S+B) is often used.

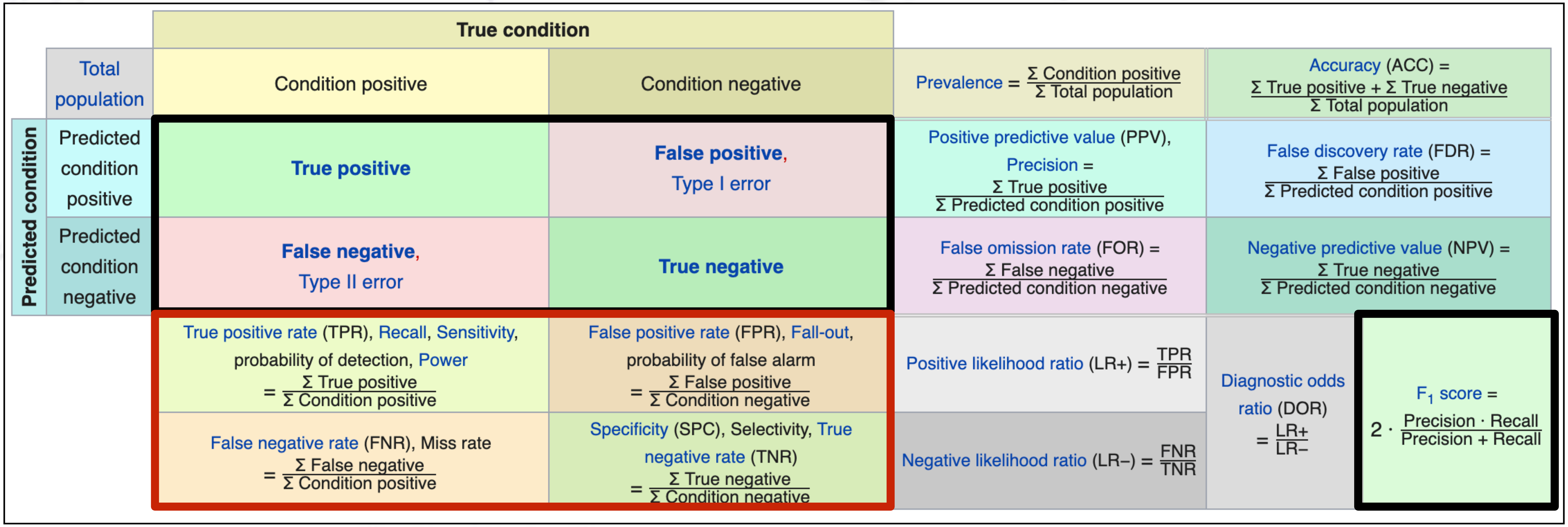


JINST, Vol12, Jan.2017

Which metric to use?

There are a ton of metrics in hypothesis testing, see below. However, those in the boxes below are the most central ones.

One metric - not mentioned here - is the Area Under the Curve (AUC), which is simply an integral of the ROC curve (thus 1 is perfect score). This is often used in Machine Learning to optimise performance (loss).

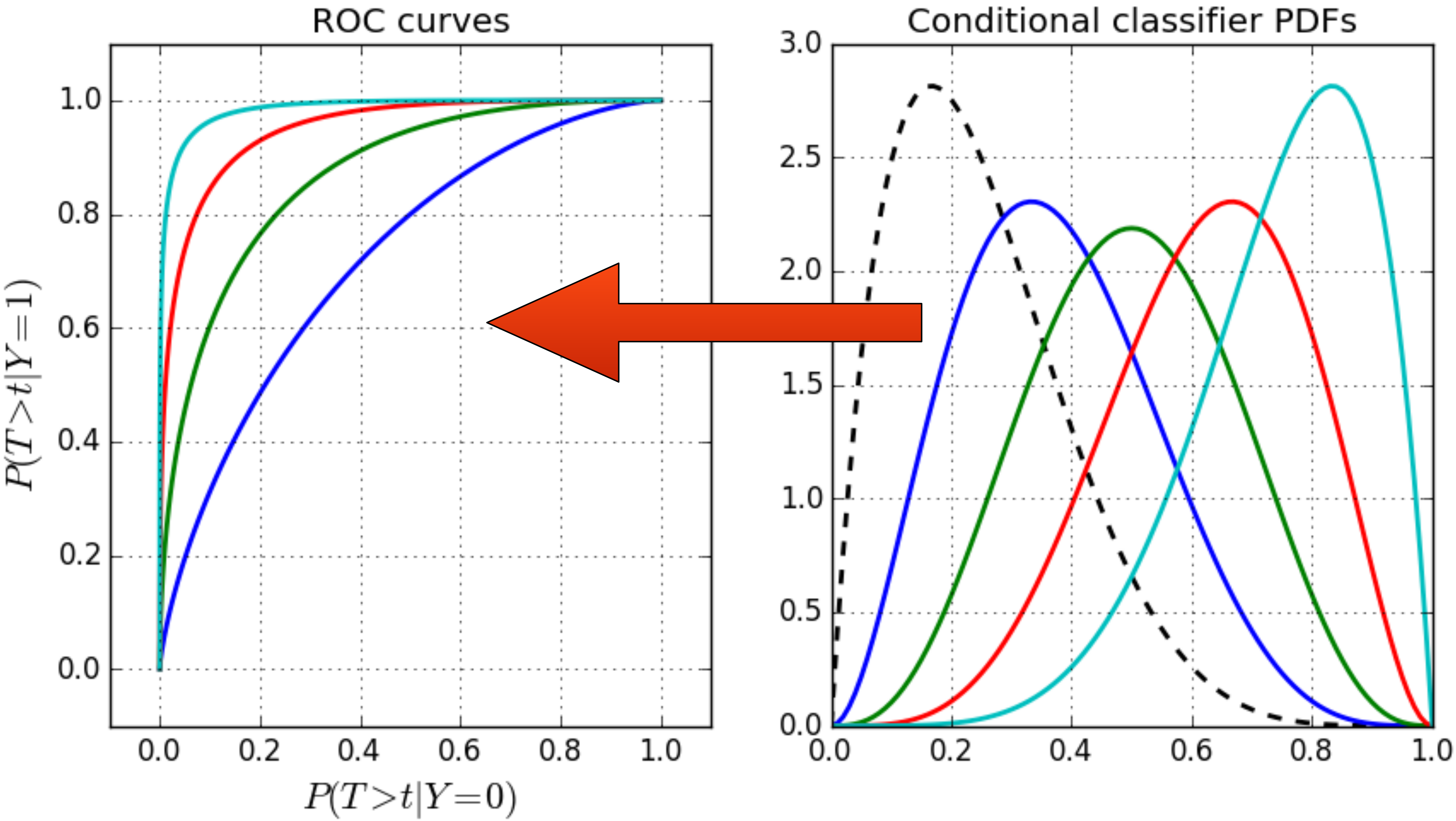


		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR–}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision + Recall}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR–) = $\frac{\text{FNR}}{\text{TNR}}$		

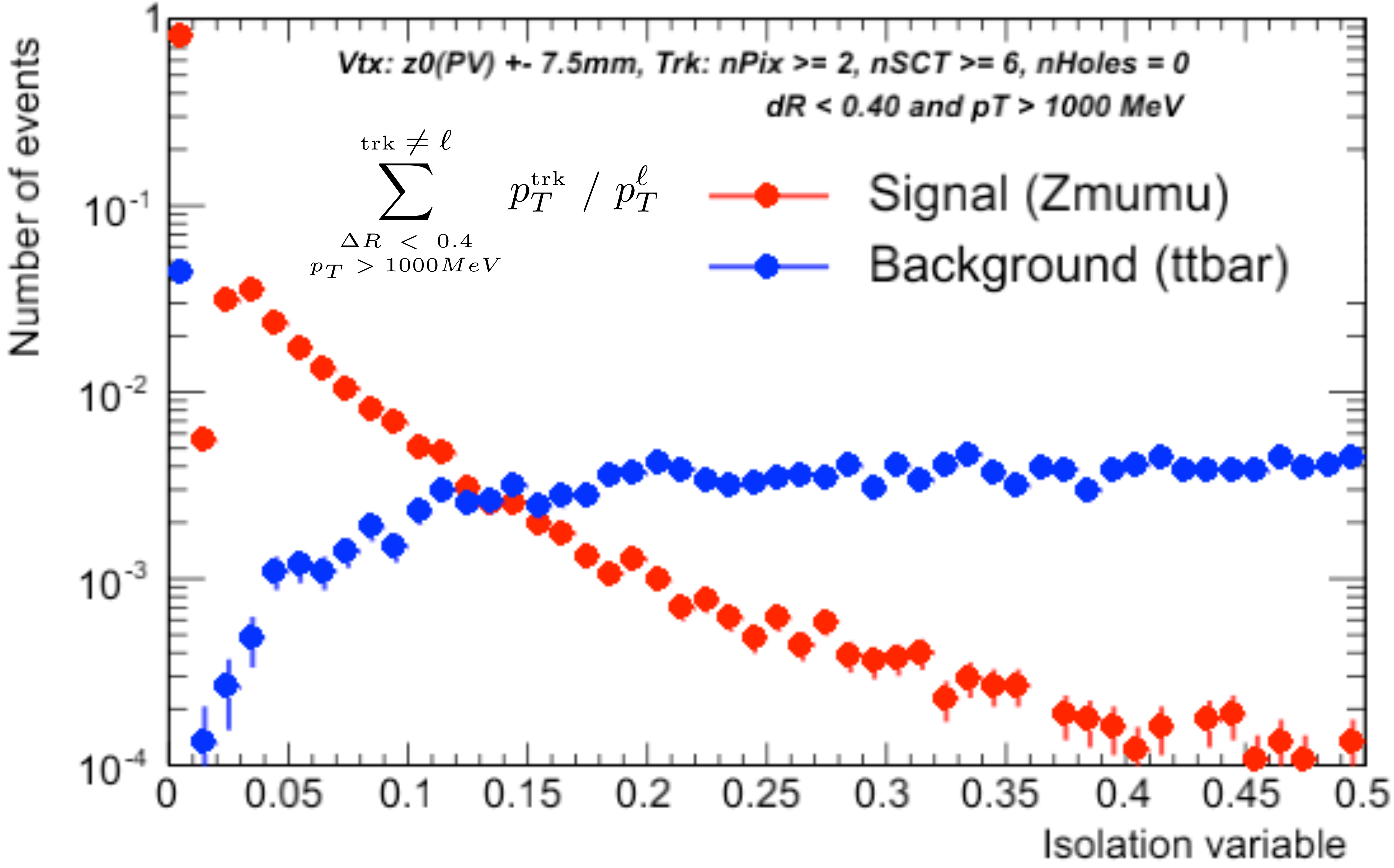
https://en.wikipedia.org/wiki/Receiver_operating_characteristic

Example of ROC curves in use

Simple case

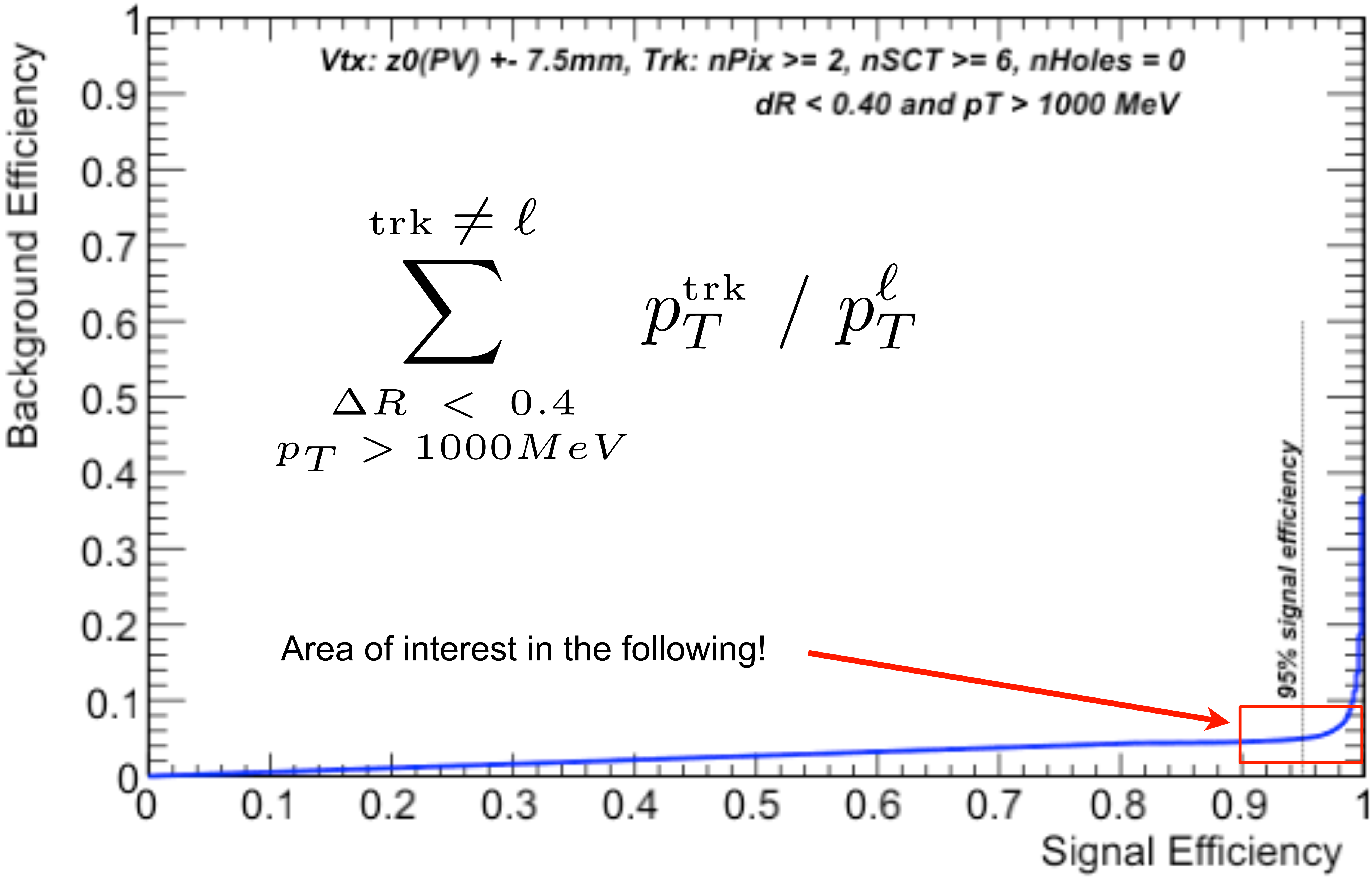

ROC curves
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
$P(T > t|Y=1)$
$P(T > t|Y=0)$
Conditional classifier PDFs
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0

Basic steps - distributions



$$\sum_{\substack{\Delta R\ <\ 0.4 \\ p_T\ >\ 1000 MeV}}^{\mathrm{trk}\ \neq\ \ell} p_T^{\mathrm{trk}}\ /\ p_T^{\ell}$$

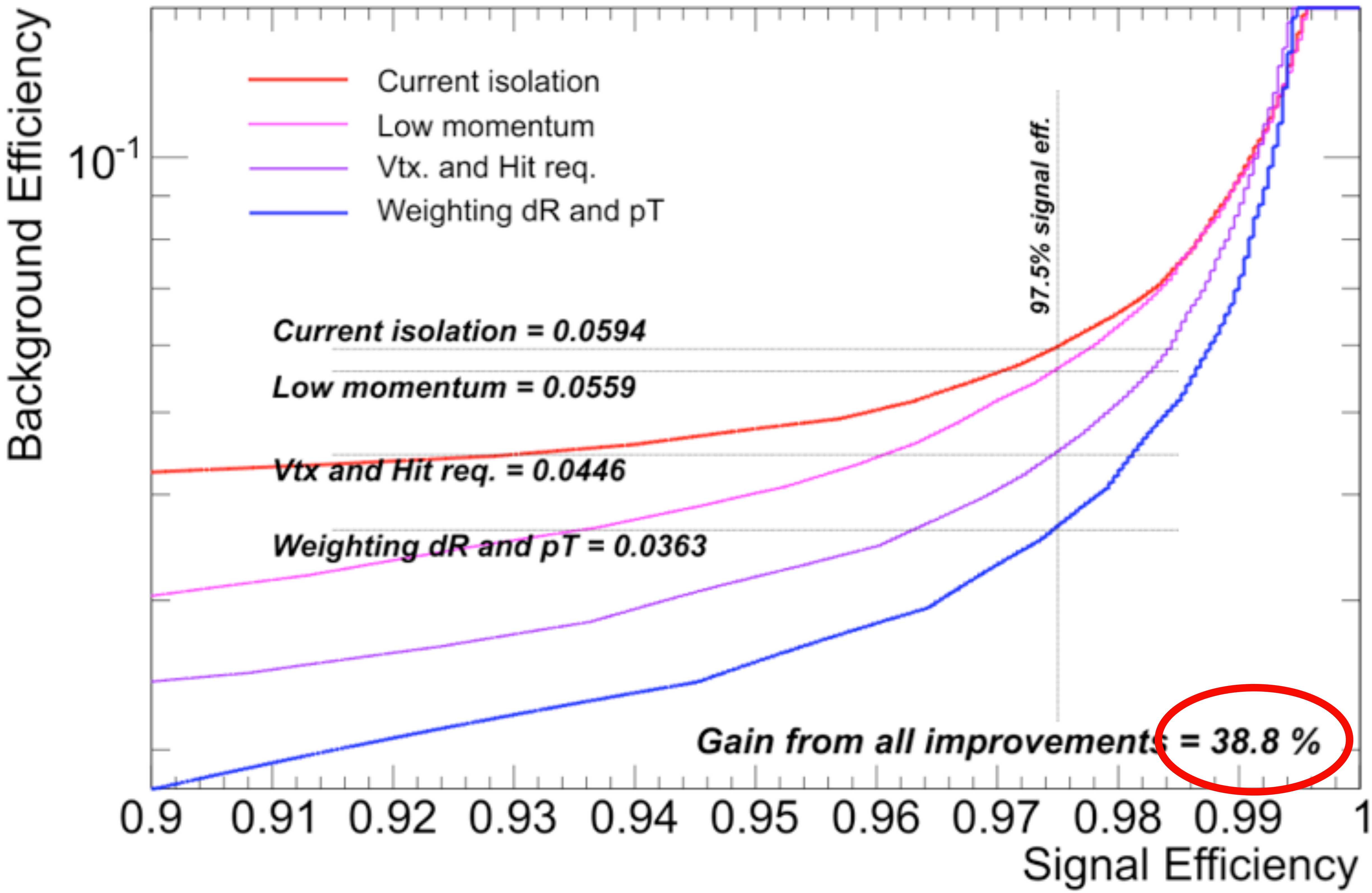
Basic steps - ROC curves



$$\sum_{\substack{\Delta R\ <\ 0.4 \\ p_T\ >\ 1000 MeV}}^{\mathrm{trk}\,\neq\,\ell} \quad p_T^{\mathrm{trk}} \ / \ p_T^{\ell}$$

Area of interest in the following!

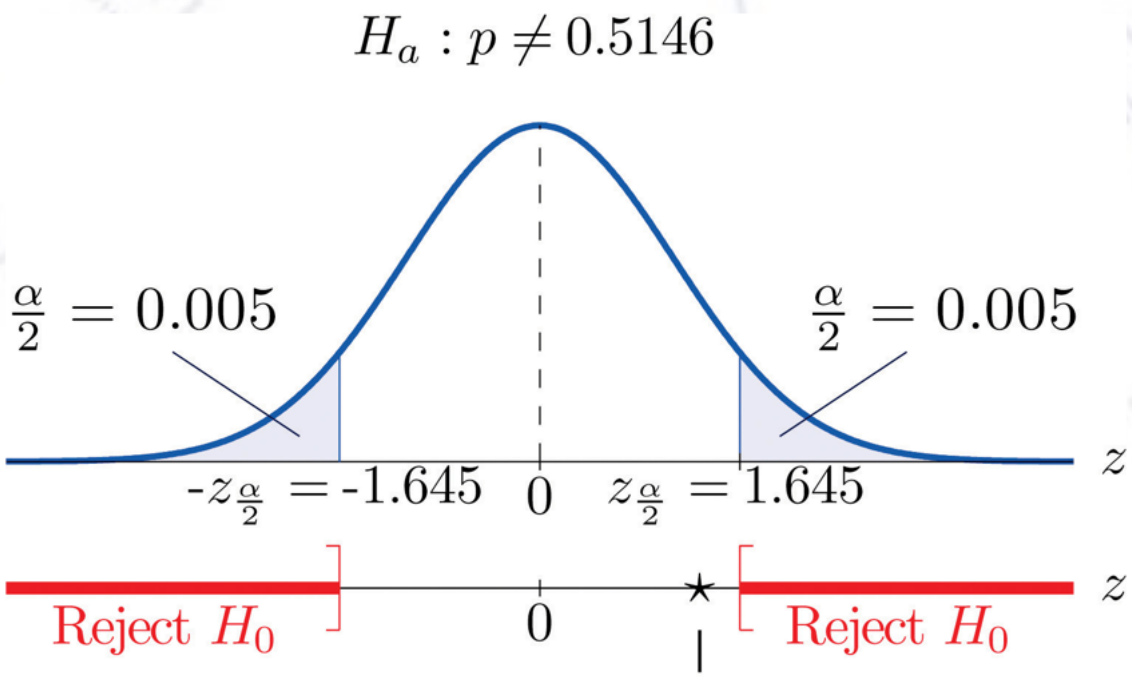
Overall improvement

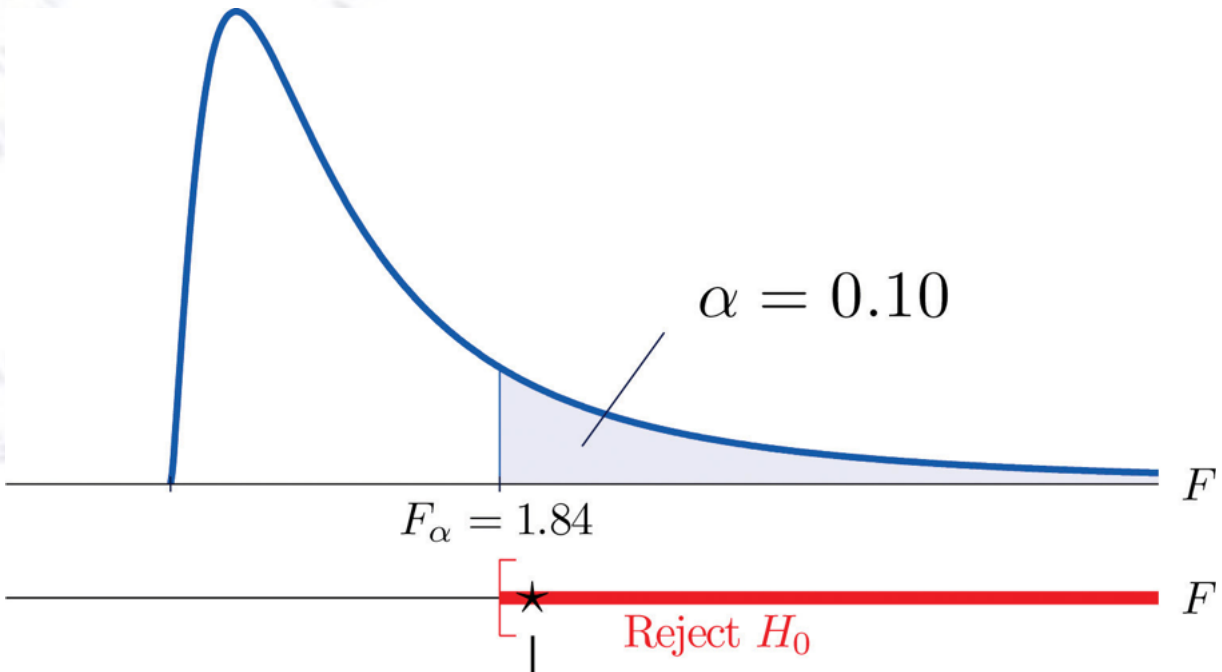


Testing procedure
&
Typical statistical tests

Testing procedure

1. Consider an **initial (null) hypothesis**, of which the truth is unknown.
2. State null and **alternative hypothesis**.
3. Consider statistical **assumptions** (independence, distributions, etc.)
4. Decide for appropriate test and state relevant **test statistic**.
5. **Derive the test statistic** distribution under null and alternative hypothesis.
   In standard cases, these are well known (Poisson, Gaussian, Student's t, etc.)
6. **Select a significance level** (α), that is a probability threshold below which null hypothesis will be rejected (typically from 5% (biology) and down (physics)).
7. Compute from (otherwise blinded) observations/data **value of test statistic** $\boldsymbol{t}$.
8. From t **calculate probability of observation** under null hypothesis (**p-value**).
9. **Reject null hypothesis** for alternative **if p-value is below significance level**.

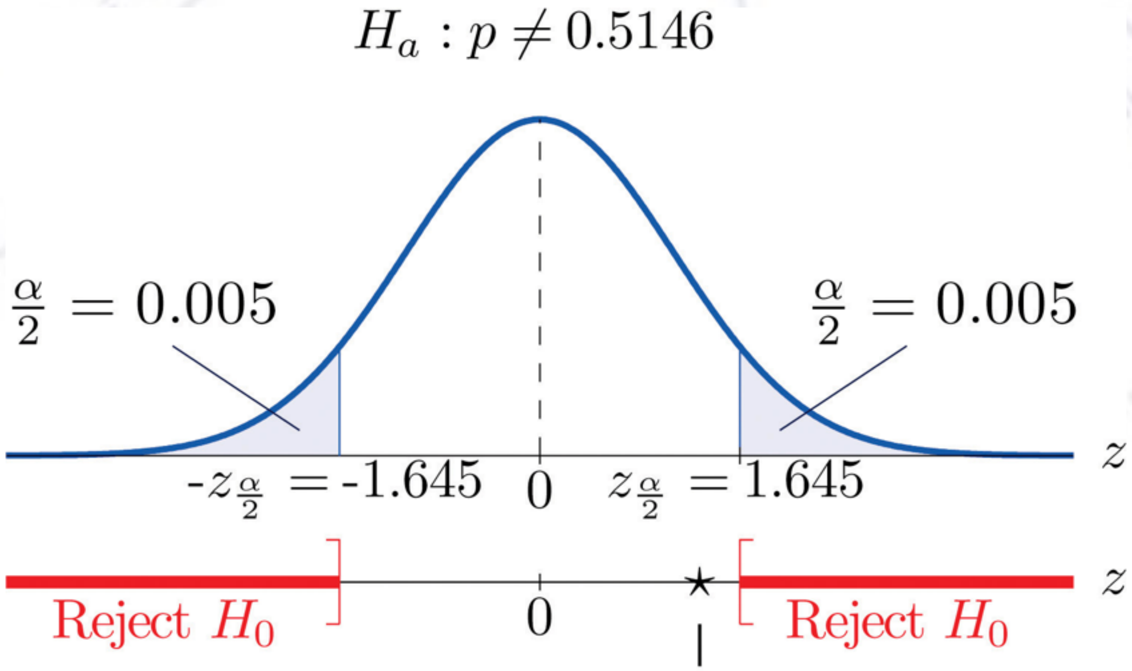


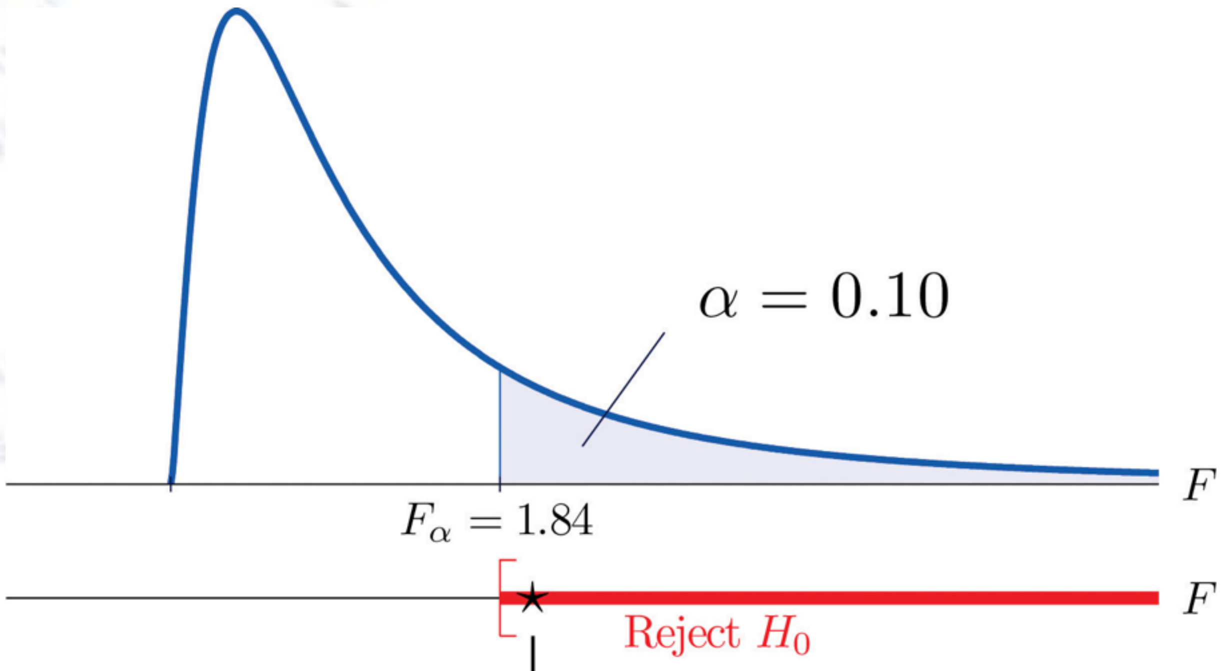


Testing procedure

1. Consider an **initial (null) hypothesis**, of which the truth is unknown.
2. State null and **alternative hypothesis**.
3. Consider statistical **assumptions** (independence, distributions, etc.)
4. Decide for appropriate test and state relevant **test statistic**.
5. **Derive the test statistic** distribution under null and alternative hypothesis.
   In standar , etc.)
6. **Select a si** ch null
   hypothesi sics)).
7. Compute f **istic** *t*.
8. From *t* **cal** **alue**).
9. **Reject nul** **level**.

1. State hypothesis.
2. Set the criteria for a decision.
3. Compute the test statistic.
4. Make a decision.





Hypothesis testing philosophy

In hypothesis testing, you can **never prove a hypothesis**.

You can **accept** a hypothesis, but this does not exclude accepting other hypothesis.

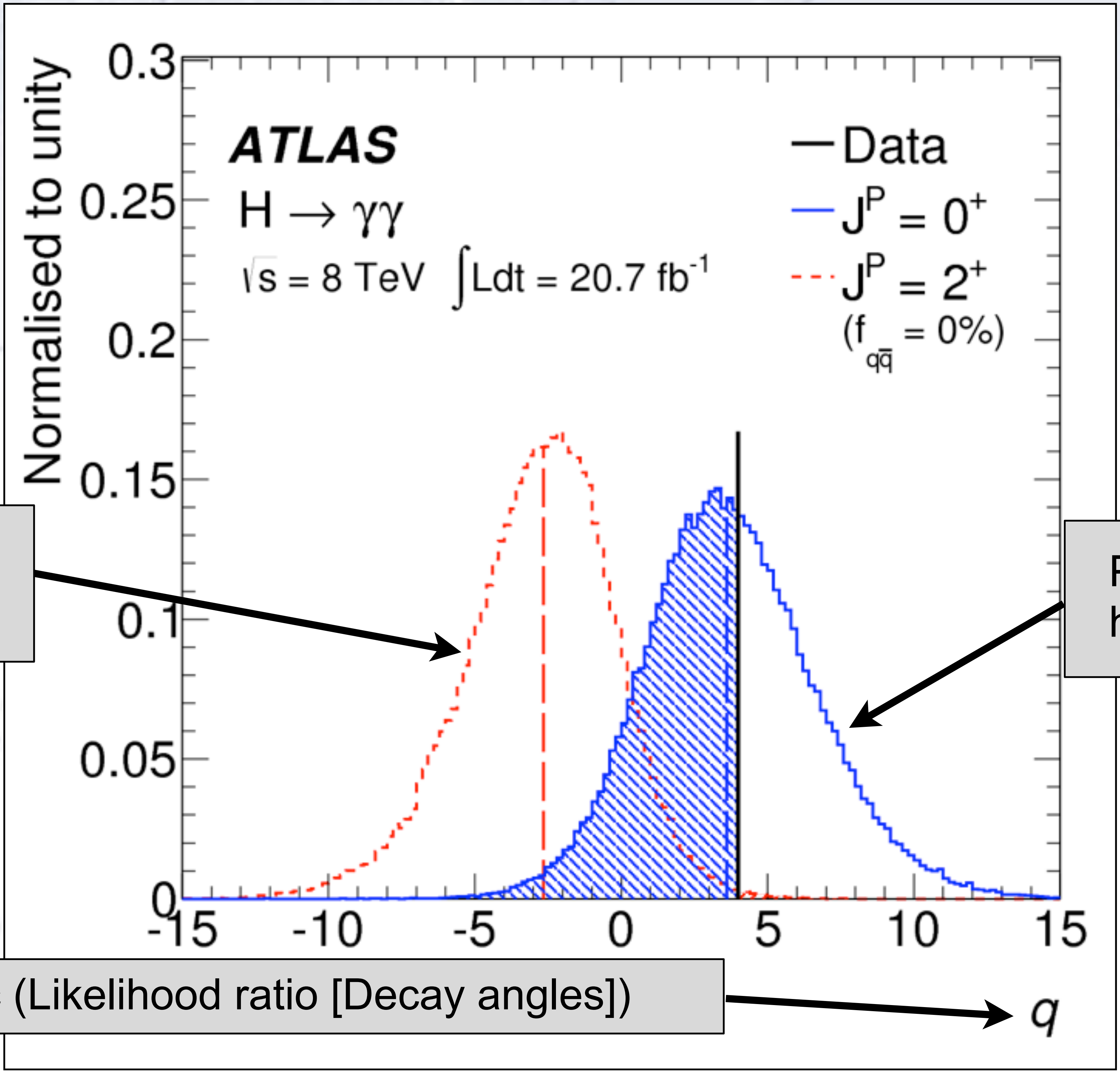
However, you can **reject** a hypothesis on the basis that it's probability of being correct (p-value) is too small.

Thus, in hypothesis testing, the line of reasoning is to state a hypothesis *opposite* of what you want to show, and then try to **reject** this hypothesis.

See Barlow 8.2.1 (p. 146)

Example of hypothesis test

The spin of the newly discovered "Higgs-like" particle (spin 0 or 2?):



PDF of spin 2 hypothesis

PDF of spin 0 hypothesis

Test statistic (Likelihood ratio [Decay angles])

The likelihood ratio test

While a single likelihood value says little, the likelihood ratio between two competing hypothesis can be compared (same offset constant/factor!).

As with the likelihood, one often takes the logarithm and multiplies by -2 to match the Chi-Square, thus the test statistic D becomes:

$$\begin{aligned} D &= -2\ln\left(\frac{\text{likelihood for null model, } H_0}{\text{likelihood for alternative model, } H_1}\right) \\ &= -2\ln\mathcal{L}(\text{null model, } H_0) + 2\ln\mathcal{L}(\text{alternative model, } H_1) \end{aligned}$$

If the two hypothesis are simple (i.e. no free parameters) then the **Neyman-Pearson Lemma** states (loosely) that **this is the best possible test one can make**.

If the alternative model is not simple but nested (i.e. contains the null hypothesis), then Wilk's Theorem states that this ratio approximately behaves like a Chi-Square distribution with $N_{dof} = N_{dof}$(alternative) - N_{dof}(null).

If errors are Gaussian (e.g. high statistics histograms) then this goes for Chi-Square differences also.

Wilk's Theorem

While the **likelihood ratio** is in principle both simple to write up and powerful:

$$D = -2\ln\left(\frac{\mathcal{L}\text{ for null model, }H_0}{\mathcal{L}\text{ for alternative model, }H_1}\right)$$

…it turns out that determining the expected distribution of the likelihood ratio is often very hard.
To know the two likelihoods one might use a Monte Carlo simulation, representing the distribution by an n-dimensional histogram (since our observable, x, can have n dimensions). But if we have M bins in each dimension, then we have to determine M^n numbers, which might be too much.

However, a convenient result (Wilk's Theorem) states that as the sample size approaches infinity, **the test statistic D will be χ^2-distributed with N_{dof} equal to the difference in dimensionality of the Null and the Alternative** (nested) **hypothesis**.
Alternatively, one can choose a simpler (and usually fully acceptable test)…

Common statistics tests

Common statistical tests

- **One-sample test** compares sample (e.g. mean) to known value:

  Example: Comparing sample to known constant ($\mu_{exp} = 2.91 \pm 0.01$ vs. $c = 2.99$).

$$z = \frac{\bar{x} - \mu_0}{\sigma(\bar{x})}$$

- **Two-sample test** compares two samples (e.g. means).

  Example: Comparing sample to control ($\mu_{exp} = 4.1 \pm 0.6$ vs. $\mu_{control} = 0.7 \pm 0.4$).

$$z = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\sigma(\bar{x_1})^2 + \sigma(\bar{x_2})^2}}$$

- **Paired test** compares paired member difference (to control important variables).

  Example: Testing environment influence on twins to control genetic bias ($\mu_{diff} = 0.81 \pm 0.29$ vs. 0).

- **Chi-squared test** evaluates adequacy of model compared to data.

  Example: Model fitted to (possibly binned) data, yielding p-value = Prob($\chi^2 = 45.9$, $N_{dof} = 36$) = 0.125

- **Kolmogorov-Smirnov test** compares if two distributions are compatible.

  Example: Compatibility between function and sample or between two samples, yielding p-value = 0.87

- **Wald-Wolfowitz runs test** is a binary check for independence.
- **Fisher's exact test** calculates p-value for contingency tables.
- **F-test** compares two sample variances to see, if grouping is useful.

From z- to p-value

- **One-sample test** compares sample (e.g. mean) to known value:

  Example: Comparing sample to known constant ($\mu_{exp} = 2.91 \pm 0.01$ vs. $c = 2.99$).

$$z = \frac{\bar{x} - \mu_0}{\sigma(\bar{x})}$$

- **Two-sample test** compares two samples (e.g. means).

  Example: Comparing sample to control ($\mu_{exp} = 4.1 \pm 0.6$ vs. $\mu_{control} = 0.7 \pm 0.4$).

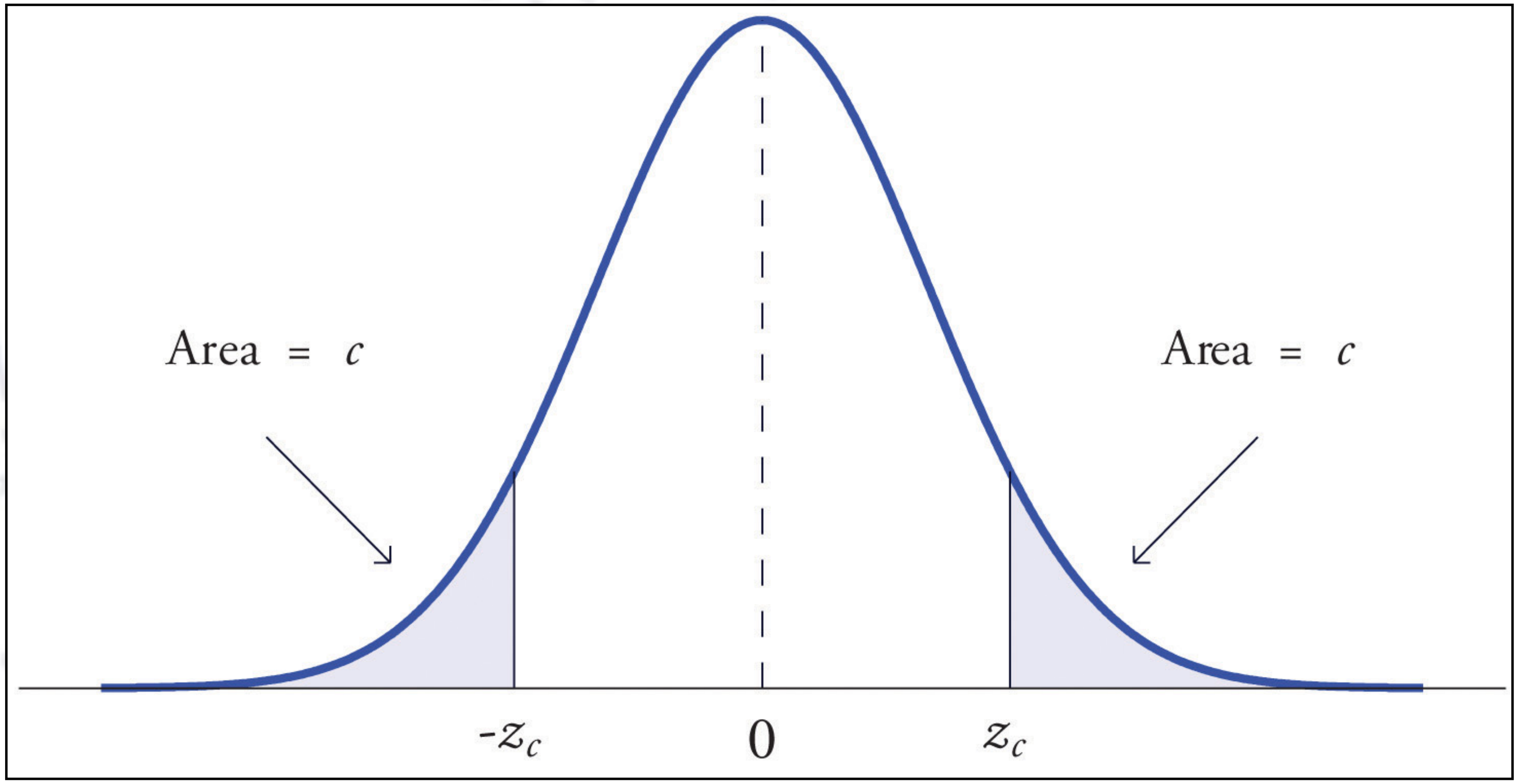
$$z = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\sigma(\bar{x_1})^2 + \sigma(\bar{x_2})^2}}$$

- **Paired test** compares paired member difference (to control important variables).

  Example: Testing environment influence on twins to control genetic bias ($\mu_{diff} = 0.81 \pm 0.29$ vs. 0).

The step from z-value to p-value consists of taking the integral of a Gaussian:

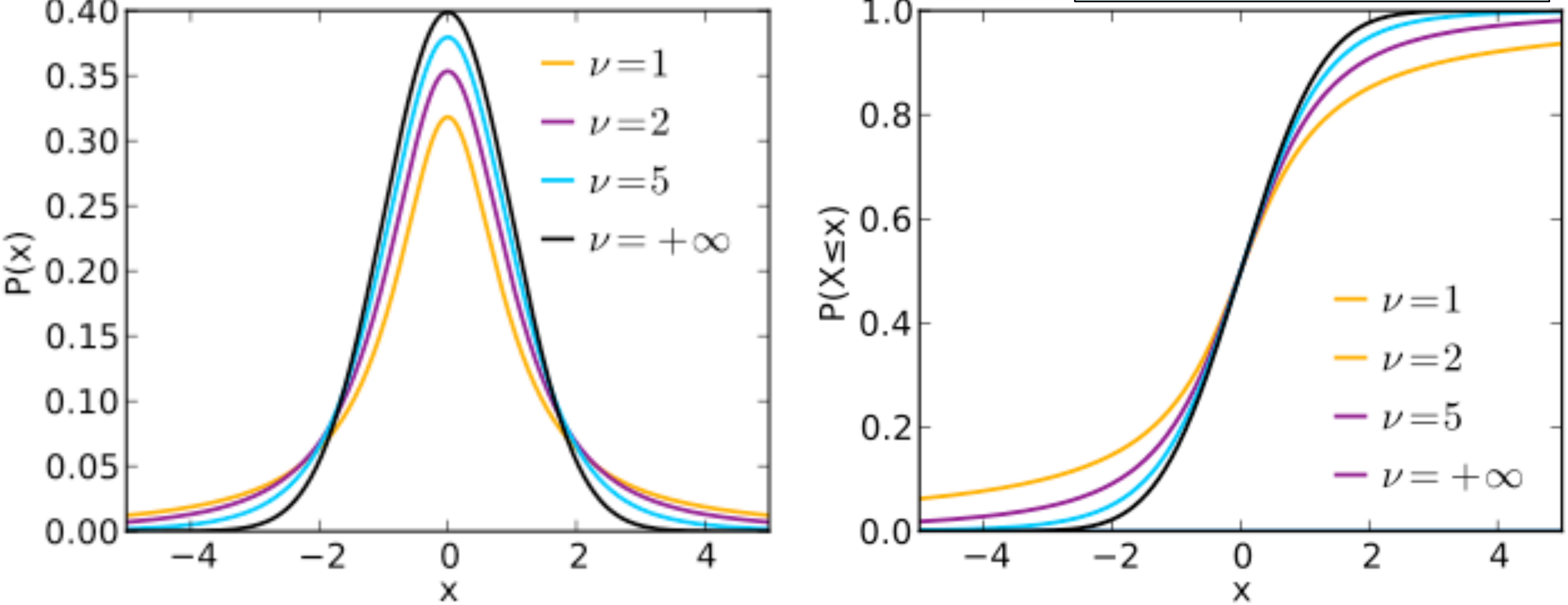
You ask yourself: "What is the probability of getting this result or worse?", and find the p-value from the integral of "this result" i.e. your z-value and "out" i.e. "worse".



Student's t-distribution

Discovered by William Gosset (who signed "student"), student's t-distribution takes into account lacking knowledge of the variance.

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$



When variance is unknown, estimating it from sample gives additional error:

Gaussian: $$z=\frac{x-\mu}{\sigma}$$

Student's: $$t=\frac{x-\mu}{\hat{\sigma}}$$

Simple tests (Z- or T-tests)

- **One-sample test** compares sample (e.g. mean) to known value:

  Example: Comparing sample to known constant ($\mu_{exp} = 2.91 \pm 0.01$ vs. $c = 3.00$).

$$z = \frac{\bar{x} - \mu_0}{\sigma(\bar{x})}$$

- **Two-sample test** compares two samples (e.g. means).

  Example: Comparing sample to control ($\mu_{exp} = 4.1 \pm 0.6$ vs. $\mu_{control} = 0.7 \pm 0.4$).

$$z = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\sigma(\bar{x_1})^2 + \sigma(\bar{x_2})^2}}$$

- **Paired test** compares paired member difference (to control important variables).

  Example: Testing environment influence on twins to control genetic bias ($\mu_{diff} = 0.81 \pm 0.29$ vs. 0).

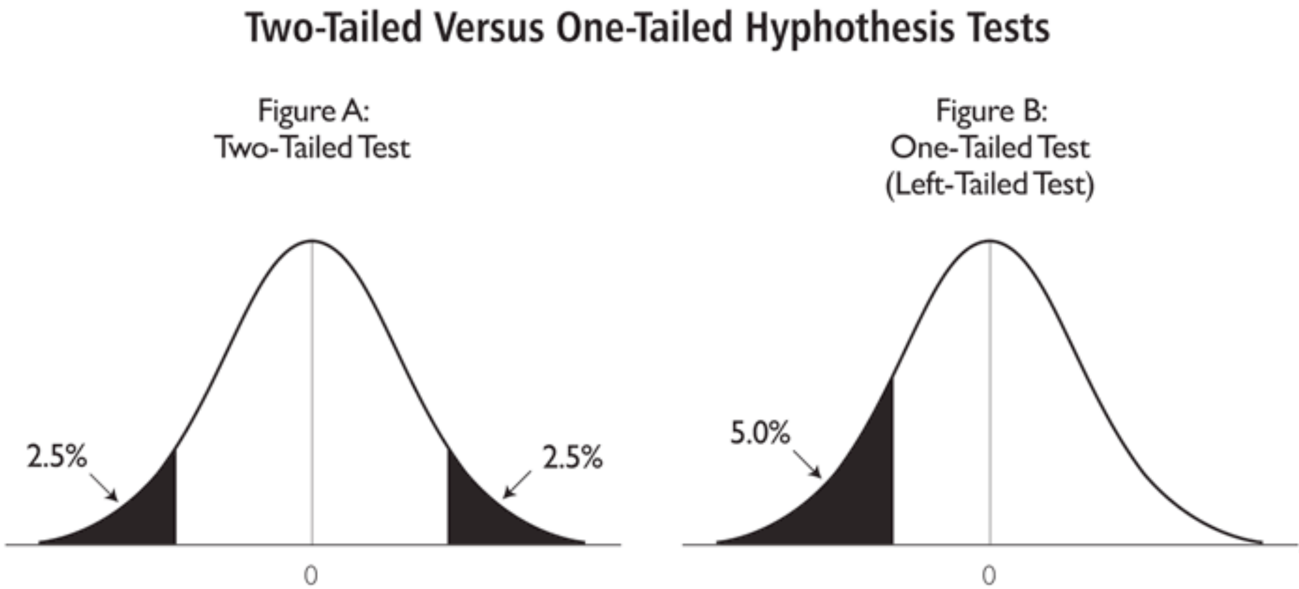
Things to consider:

- Variance known (Z-test) vs. Variance unknown (T-test).

  Rule-of-thumb: If N > 10-20 or σ known then Z-test, else T-test.

- One-sided vs. two-sided test.

  Rule-of-thumb: If you want to test for difference, then use two-sided. If you care about specific direction of difference, use one-sided.



Chi-squared test

Without any further introduction...

$$\chi^2(\bar{\theta}) = \sum_{i=1}^{N} \frac{(y_i - \lambda(x_i;\bar{\theta}))^2}{\sigma_i^2}$$

- **Chi-squared test** evaluates adequacy of model compared to data.

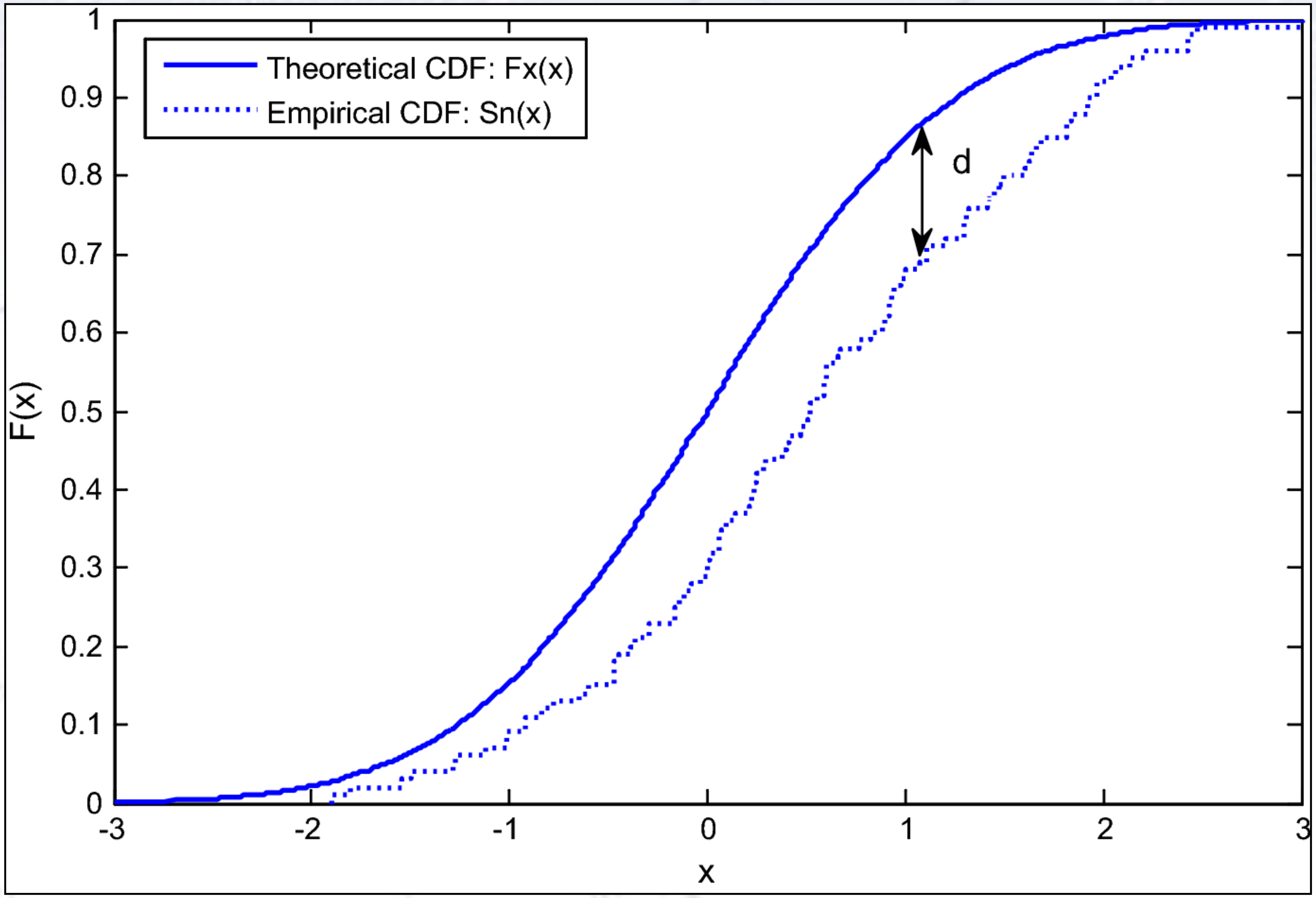
Example: Model fitted to (possibly binned) data, yielding p-value = Prob(χ^2 = 45.9, N_{dof} = 36) = 0.125

If the p-value is small, the hypothesis is unlikely...

Kolmogorov-Smirnov test

- **Kolmogorov-Smirnov test** compares if two distributions are compatible.

Example: Compatibility between function and sample or between two samples, yielding p-value = 0.87

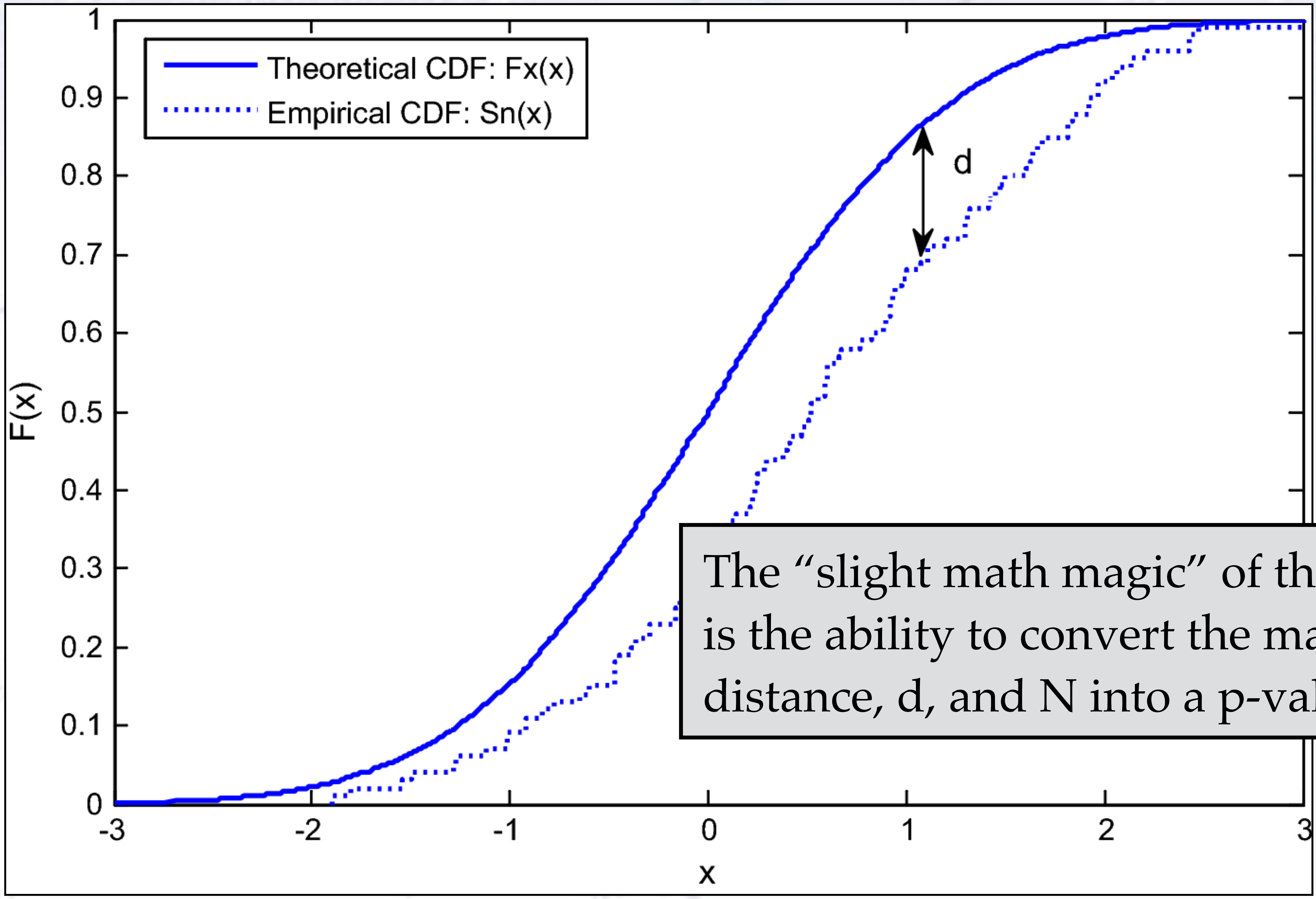


The Kolmogorov test measures the maximal distance between the integrals of two distributions and gives a probability of being from the same distribution.

Kolmogorov-Smirnov test

- **Kolmogorov-Smirnov test** compares if two distributions are compatible.

Example: Compatibility between function and sample or between two samples, yielding p-value = 0.87



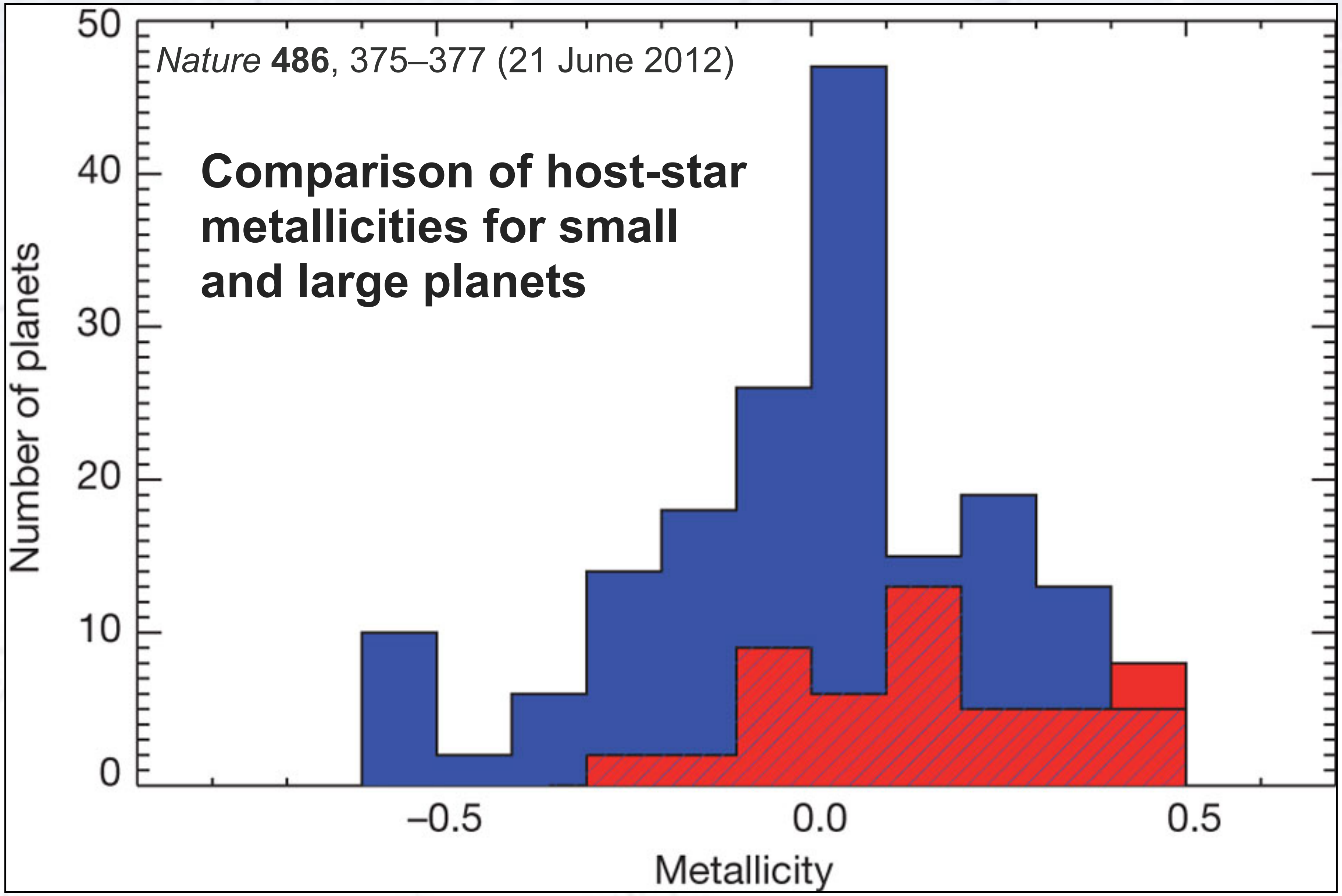
The "slight math magic" of the K-S test is the ability to convert the maximal distance, d, and N into a p-value.

The Kolmogorov test measures the maximal distance between the integrals of two distributions and gives a probability of being from the same distribution.

Kolmogorov-Smirnov test

- **Kolmogorov-Smirnov test** compares if two distributions are compatible.

Example: Compatibility between function and sample or between two samples, yielding p-value = 0.87

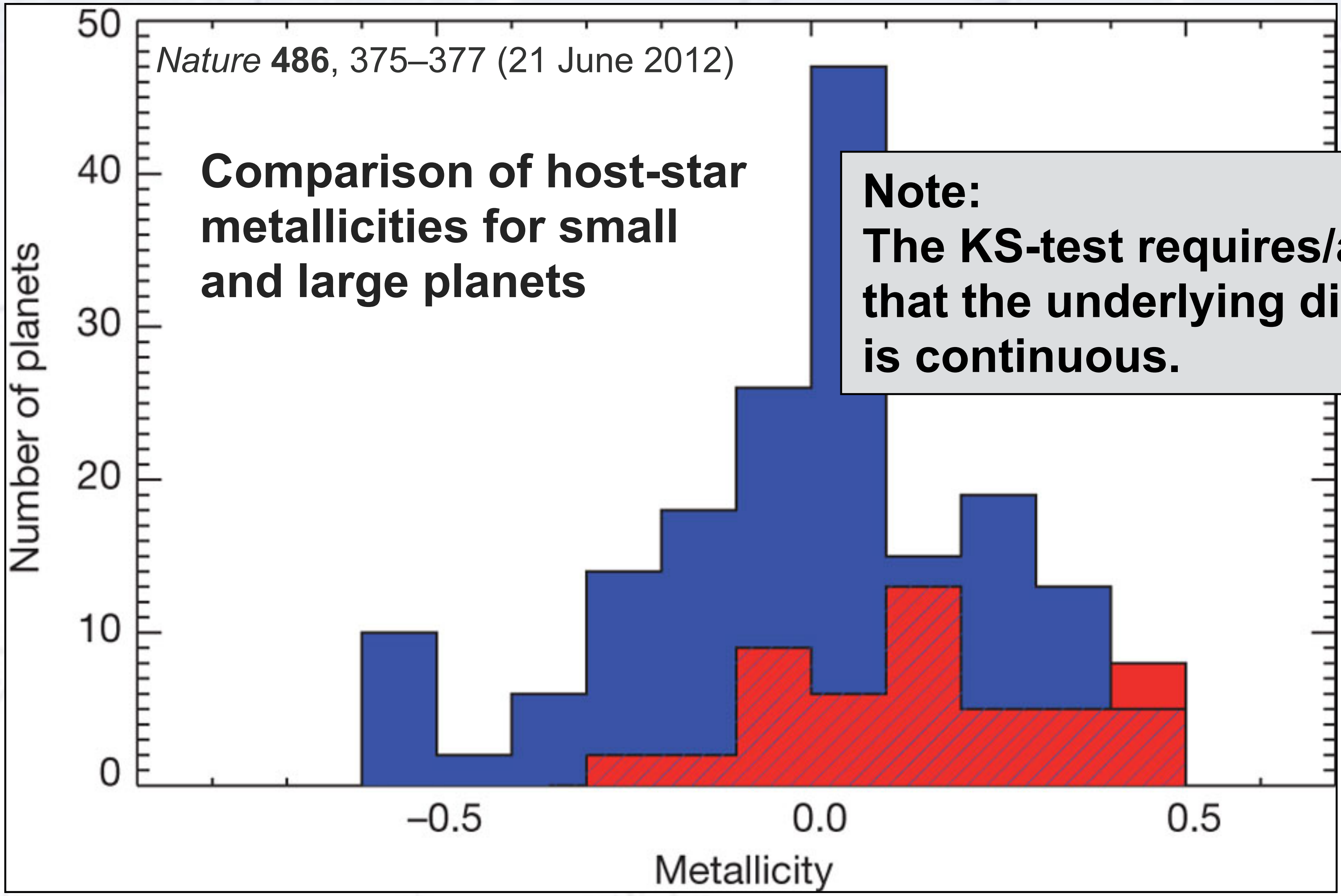


“A Kolmogorov–Smirnov test shows that the probability that the two distributions are not drawn randomly from the same parent population is greater than 99.96%; that is, the two distributions differ by more than 3.5σ”. [Quote from figure caption]

Kolmogorov-Smirnov test

- **Kolmogorov-Smirnov test** compares if two distributions are compatible.

Example: Compatibility between function and sample or between two samples, yielding p-value = 0.87



Nature **486**, 375–377 (21 June 2012)

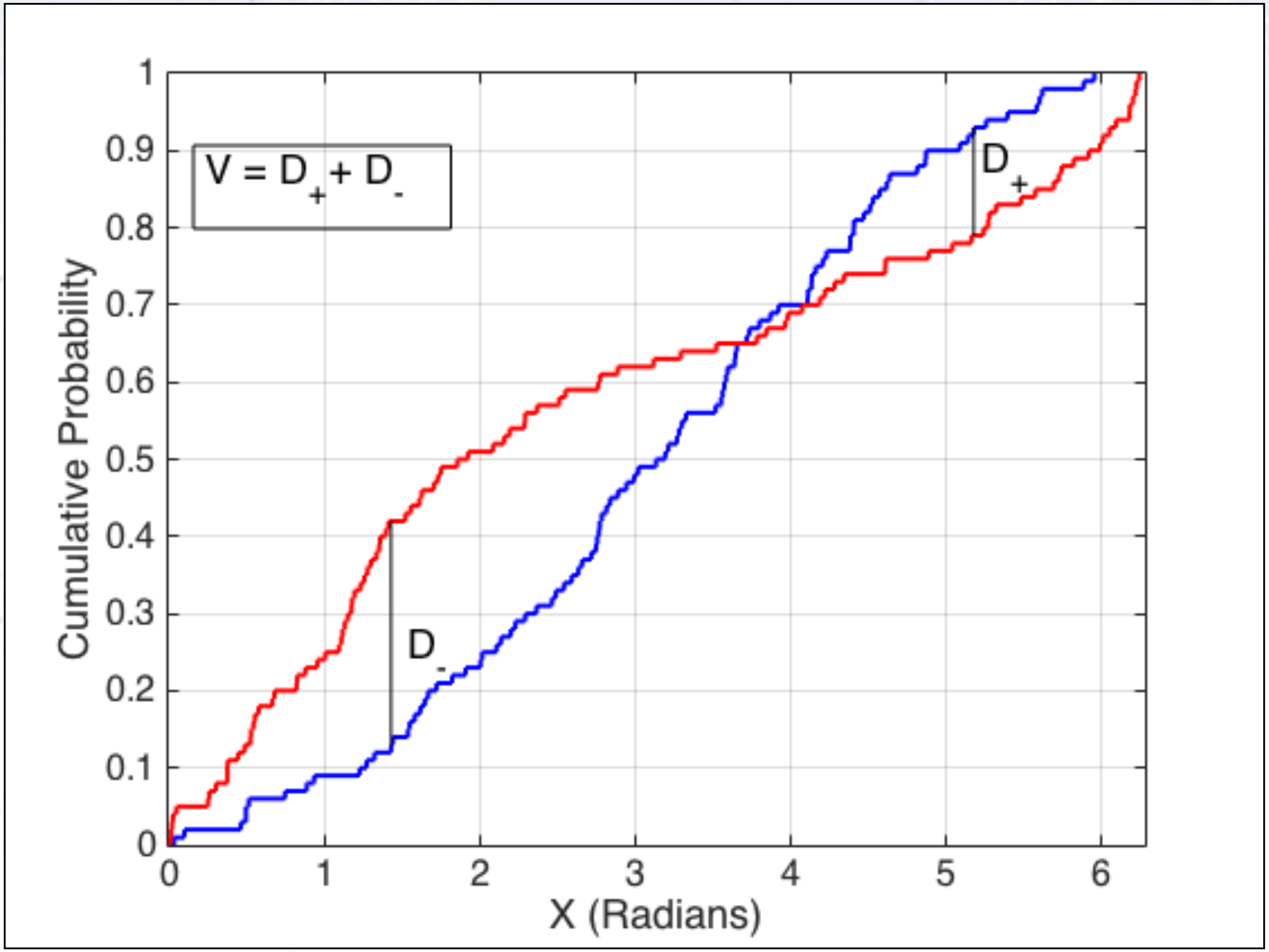
Comparison of host-star metallicities for small and large planets

**Note:
The KS-test requires/assumes, that the underlying distribution is continuous.**

“A Kolmogorov–Smirnov test shows that the probability that the two distributions are not drawn randomly from the same parent population is greater than 99.96%; that is, the two distributions differ by more than 3.5σ”. [Quote from figure caption]

Kuiper test

Is a similar test, but it is more specialised in that it is good to detect SHIFTS in distributions (as it uses the maximal signed distance in integrals).



Common statistical tests

- **One-sample test** compares sample (e.g. mean) to known value: $z = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n}$

  Example: Comparing sample to known constant ($\mu_{exp} = 2.91 \pm 0.01$ vs. $c = 3.00$).

- **Two-sample test** compares two samples (e.g. means). $z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n1} + \frac{\sigma_2^2}{n2}}}$

  Example: Comparing sample to control ($\mu_{exp} =$ [illegible] ± 0.6 vs. $\mu_{control} = 0.7 \pm 0.4$).

- **Paired test** compares paired member difference (to control important variables).

  Example: Testing environment influence on twins to control genetic bias ($\mu_{diff} = 0.81 \pm 0.29$ vs. 0).

- **Chi-squared test** evaluates adequacy of model compared to data.

  Example: Model fitted to (possibly binned) data, yielding p-value = Prob($\chi^2 = 45.9$, $N_{dof} = 36$) = 0.125

- **Kolmogorov-Smirnov test** compares if two distributions are compatible.

  Example: Compatibility between function and sample or between two samples, yielding p-value = 0.87

These tests you should know by heart!

Those below are for general education, and you should just know about them (and the last one is not curriculum).

- **Wald-Wolfowitz runs test** is a binary check for independence.
- **Fisher's exact test** calculates p-value for contingency tables.
- **F-test** compares two sample variances to see, if grouping is useful.

Which test to use?

In principle all statistical tests can be used on every problem, but they are not all equally powerful, and some might also be biased (low stat.) or otherwise unfit.
Finally, they may not all be equally easy to implement!

One figure of merit could be the **Power of a Test***, defined as (1 – β), complement of the false negative rate, β.

This is thus the test's probability of correctly rejecting the null hypothesis.

Example:
This is a powerful test: Thus, since the result is negative, we can confidently say that the null hypothesis is not rejected (e.g. the patient does not have the condition).

In medical science, it is typically important to have a powerful test (i.e. low β), while in criminal science it is a low type I error rate (i.e. low α), convicting innocents.
In the end, choosing a test comes down to **experience, importance of power, ease of use**, and even standards in the field of research in question.

* Power of a test is often termed sensitivity in biostatistics.

Wald-Wolfowitz runs test

Barlow, 8.3.2, page 153

A different test to the Chi2 (and in fact a bit orthogonal!) is the Wald-Wolfowitz runs test.

It measures the number of “runs”, defined as sequences of same outcome (only two types).

Example:

++++−−−+++−−++++++−−−−

If random, the mean and variance is known:

$$\mu = \frac{2\ N_+\ N_-}{N} + 1$$

$$\sigma^2 = \frac{2\ N_+\ N_-\ (2\ N_+\ N_- - N)}{N^2\ (N-1)} = \frac{(\mu-1)(\mu-2)}{N-1}.$$

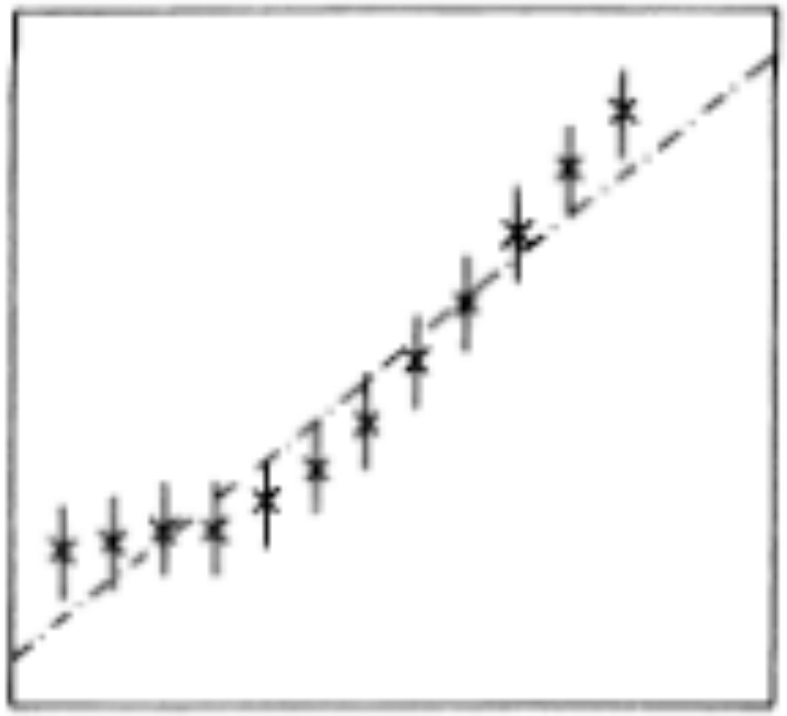


Fig. 8.3. A straight line through twelve data points.

$N = 12$, $N_+ = 6$, $N_- = 6$
$\mu = 7$, $\sigma = 1.76$
$(7-3)/1.65 = 2.4\ \sigma$ (~1%)

Note: The WW runs test requires N > 10-15 for the output to be approx. Gaussian!

Fisher’s exact test

When considering a **contingency table** (like below), one can calculate the probability for the entries to be uncorrelated. This is **Fisher’s exact test**.

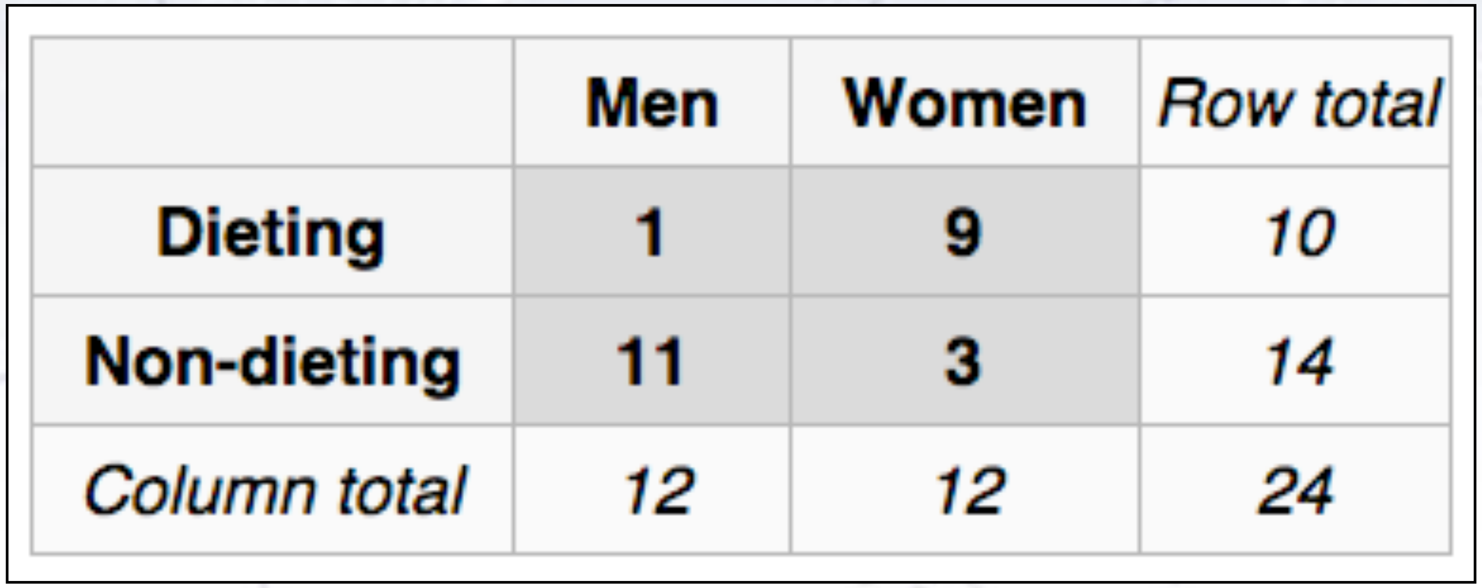
	Row 1	Row 2	Row Sum
Column 1	A	B	A+B
Column 2	C	D	C+D
Column Sum	A+C	B+D	N

$$p = \frac{\begin{pmatrix} A+C \\ A \end{pmatrix}\begin{pmatrix} B+D \\ B \end{pmatrix}}{\begin{pmatrix} N \\ A+B \end{pmatrix}} = \frac{(A+B)!\;(C+D)!\;(A+C)!\;(B+D)!}{A!\;B!\;C!\;D!\;N!}$$

Simple way to test categorial data (Note: Barnard’s test is “possibly” stronger).

Fisher's exact test - example

Consider data on men and women dieting or not. The data can be found in the below table:



	Men	**Women**	*Row total*
Dieting	1	9	*10*
Non-dieting	11	3	*14*
Column total	*12*	*12*	*24*

Is there a correlation between dieting and gender?

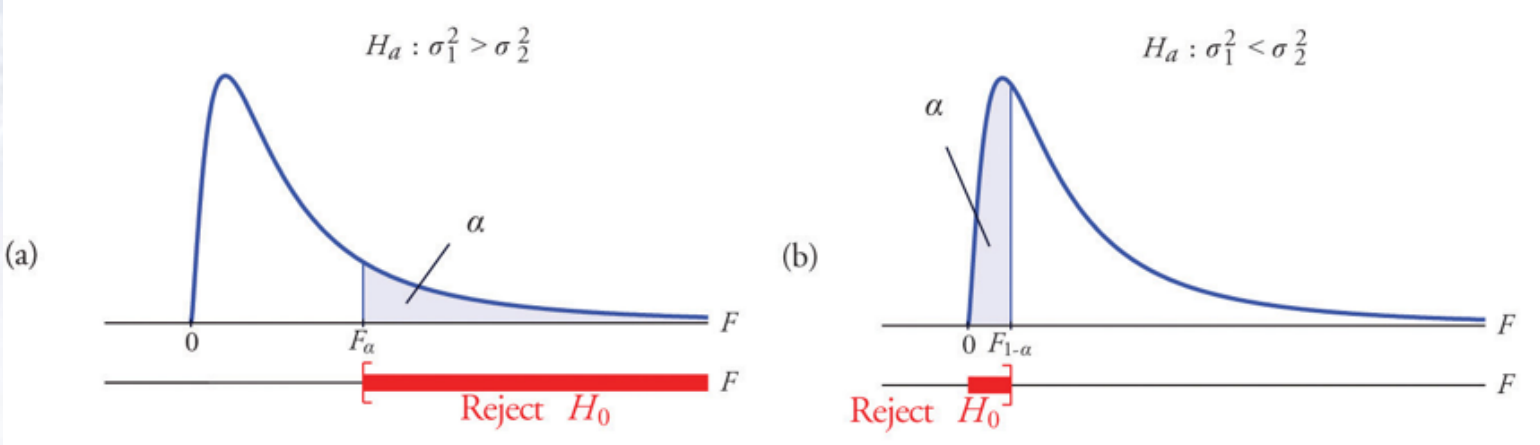
The Chi-square test is not optimal, as there are (several) entries, that are very low (< 5), but Fisher's exact test gives the answer:

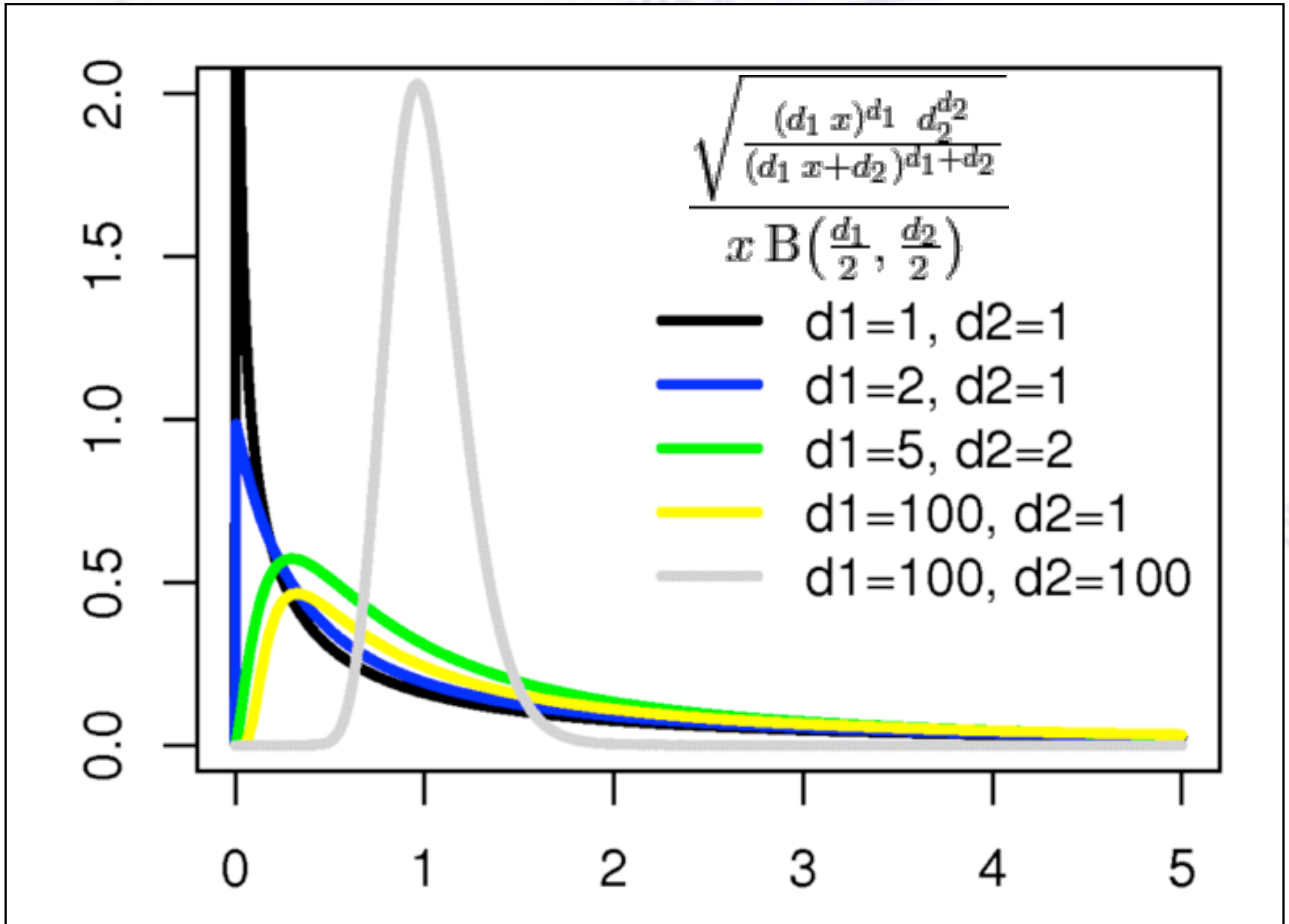
$$p = \left(\begin{array}{c} 10 \\ 1 \end{array}\right)\left(\begin{array}{c} 14 \\ 11 \end{array}\right) / \left(\begin{array}{c} 24 \\ 12 \end{array}\right) = \frac{10!\ 14!\ 12!\ 12!}{1!\ 9!\ 11!\ 3!\ 24!} \simeq 0.00135$$

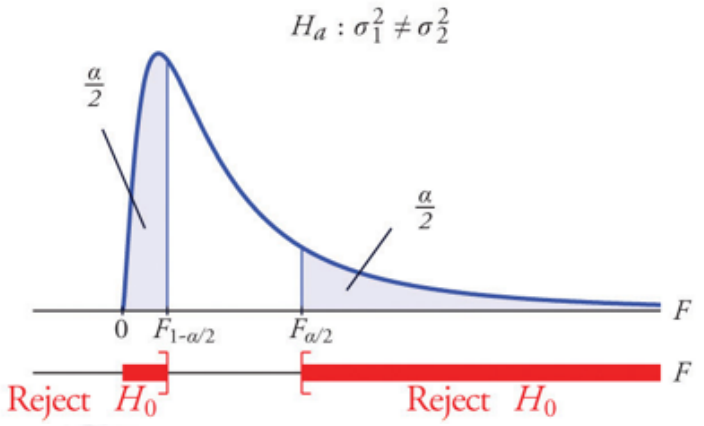
F-test

To test for differences between variances in two samples, one uses the F-test:

$$F = \frac{S_X^2}{S_Y^2}$$







Note that this is a two-sided test. One is generally testing, if the two variances are the same.

Anderson-Darling Test

A “simple” and powerful test between cumulative data F_n and distribution F is defined as:

$$n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \; w(x) \; dF(x)$$

Here, n is the number of elements in the sample and w(x) is a weighting function.

Choosing w(x) = F(x) (1-F(x)) yields the Anderson-Darling test statistic:

$$A^2 = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} \; dF(x)$$

which has more emphasis on the tails than the above (w(x) = 1, i.e. Cramer-von Mises) test statistic. An alternative is Shapiro-Wilks test, see here for comparison.

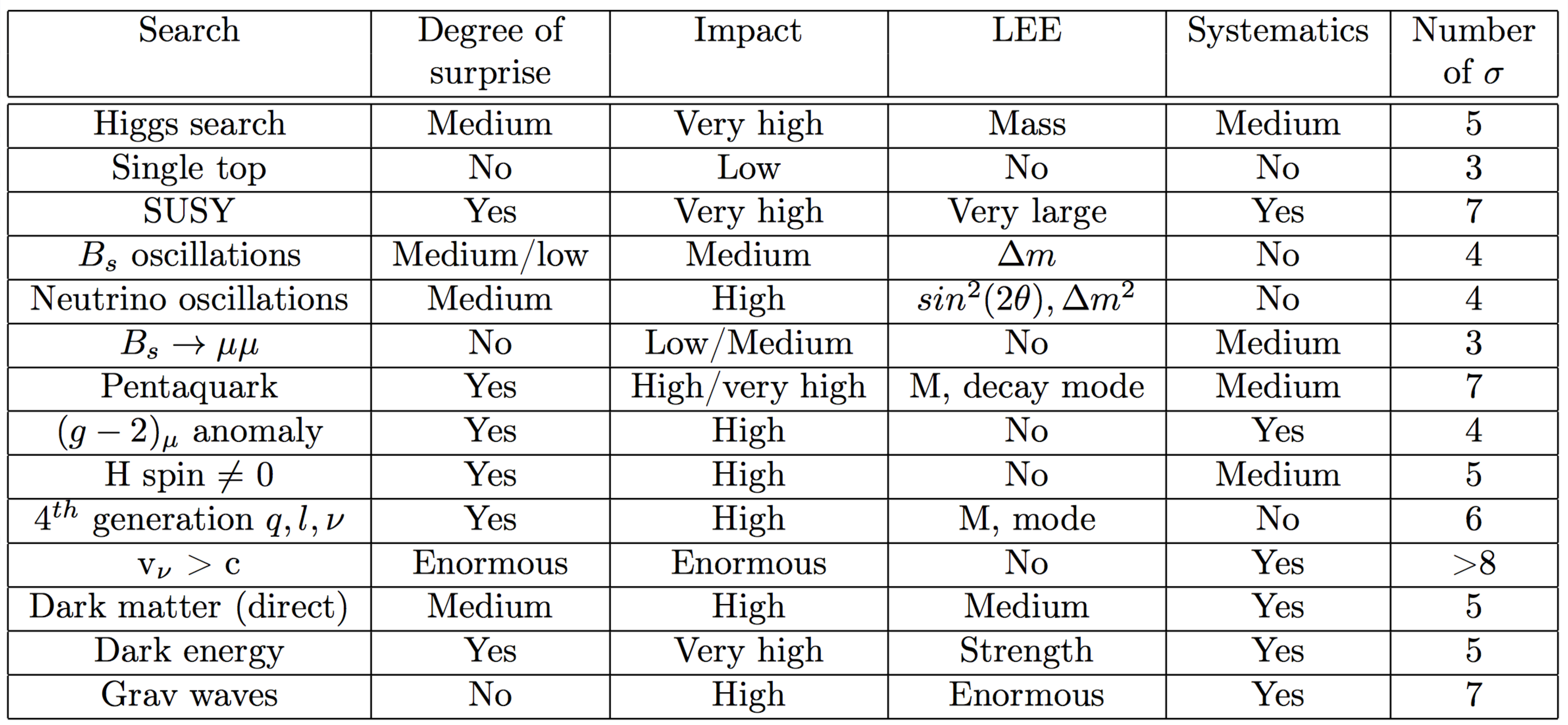
The test is implemented in the Python Statistics package (stats), with tests for the Gaussian, Exponential, Logistic & Gumbel distributions.

What value to decide at?

How many sigmas?

The number of sigmas (or p-value) required to make a claim should perhaps vary, according to the target of the data analysis.

Louis Lyons has below given his take on it (aimed at particle physics searches).



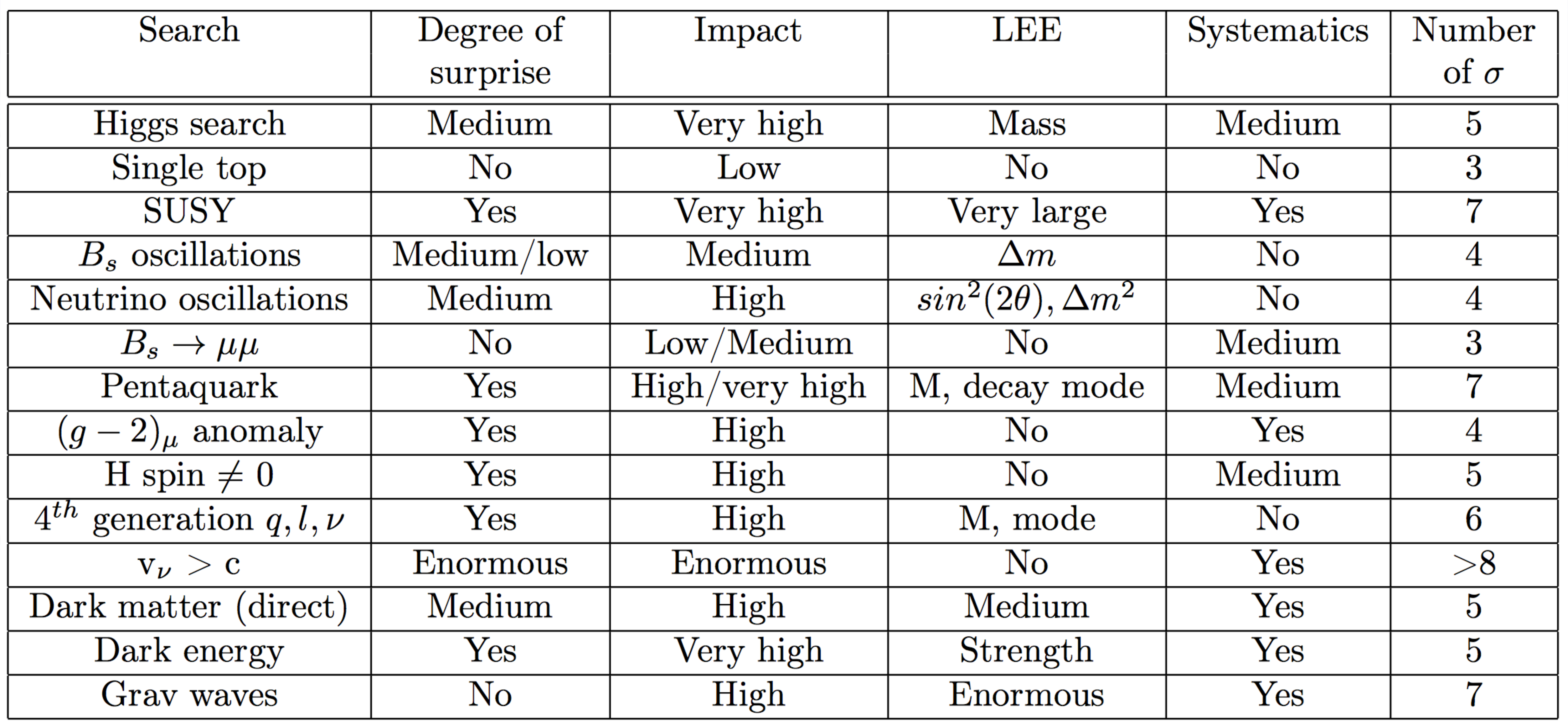
Search	Degree of surprise	Impact	LEE	Systematics	Number of σ
Higgs search	Medium	Very high	Mass	Medium	5
Single top	No	Low	No	No	3
SUSY	Yes	Very high	Very large	Yes	7
B_s oscillations	Medium/low	Medium	Δm	No	4
Neutrino oscillations	Medium	High	$sin^2(2\theta), \Delta m^2$	No	4
$B_s \to \mu\mu$	No	Low/Medium	No	Medium	3
Pentaquark	Yes	High/very high	M, decay mode	Medium	7
$(g-2)_\mu$ anomaly	Yes	High	No	Yes	4
H spin $\neq 0$	Yes	High	No	Medium	5
4^{th} generation q, l, ν	Yes	High	M, mode	No	6
$\mathrm{v}_\nu > \mathrm{c}$	Enormous	Enormous	No	Yes	>8
Dark matter (direct)	Medium	High	Medium	Yes	5
Dark energy	Yes	Very high	Strength	Yes	5
Grav waves	No	High	Enormous	Yes	7

From: "Discovering the Significance of 5 sigma", ArXiv: 1310.1284

How many sigmas?

The number of sigmas (or p-value) required to make a claim should perhaps vary, according to the target of the data analysis.

Louis Lyons has below given his take on it (aimed at particle physics searches).



Search	Degree of surprise	Impact	LEE	Systematics	Number of σ
Higgs search	Medium	Very high	Mass	Medium	5
Single top	No	Low	No	No	3
SUSY	Yes	Very high	Very large	Yes	7
B_s oscillations	Medium/low	Medium	Δm	No	4
Neutrino oscillations	Medium	High	$sin^2(2\theta), \Delta m^2$	No	4
$B_s \to \mu\mu$	No	Low/Medium	No	Medium	3
Pentaquark	Yes	High/very high	M, decay mode	Medium	7
$(g-2)_\mu$ anomaly	Yes	High	No	Yes	4
H spin $\neq 0$	Yes	High	No	Medium	5
4^{th} generation q, l, ν	Yes	High	M, mode	No	6
$v_\nu > c$	Enormous	Enormous	No	Yes	>8
Dark matter (direct)	Medium	High	Medium	Yes	5
Dark energy	Yes	Very high	Strength	Yes	5
Grav waves	No	High	Enormous	Yes	7

The more extraordinary the claim, the more extraordinary the evidence needed!

Systematic Uncertainties

Systematic Uncertainties

Caution! This contains material some people may find offensive.

Even my close colleagues and I disagree on some of the following!

Systematic uncertainties

"Everything is vague to a degree you do not realise till you have tried to make it precise."

[Bertrand Russell, 1872-1970]

Systematic uncertainties

"Systematic effects is a general category which includes effects such as background, scanning efficiency, energy resolution, variation of counter efficiencies with beam position, and energy, dead time, etc. *The uncertainty in the estimation of such a systematic effect is called a systematic error*"

[Jay Orear, 1958]

Very importantly, a *measurement error* is not a *mistake*. Systematically measuring something wrong is a mistake, if not corrected for. It is the uncertainties associated with the correction, that is the systematic uncertainty.
For this reason, it is also better to use the word *uncertainty* than *error*.

Systematic uncertainties

“Systematic effects is a general category which includes effects such as background, scanning efficiency, energy resolution, variation of counter efficiencies with beam position, and energy, dead time, etc. *The uncertainty in the estimation of such a systematic effect is called a systematic error*”

[Jay Orear, 1958]

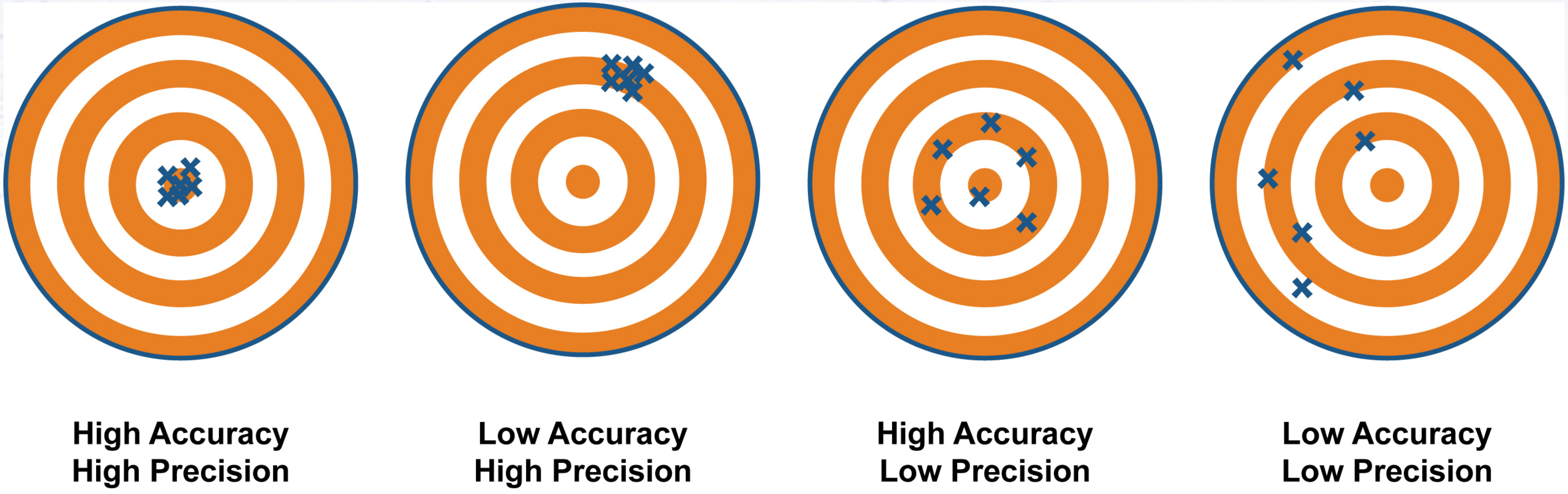
Very importantly, a *measurement error* is not a *mistake*. Systematically measuring something wrong is a mistake, if not corrected for. It is the uncertainties associated with the correction, that is the systematic uncertainty.
For this reason, it is also better to use the word *uncertainty* than *error.*

Example:
Measurements are taken with a steel ruler, the ruler was calibrated at 15°C, the measurements done at 22°C. This is a systematic **bias** and not only a systematic **uncertainty**! To neglect such an effect is a systematic **mistake**.
Effects can be corrected for! If we correct for effect, but corrections are not known exactly, then we have to introduce a systematic uncertainty (error propagation!).

How to find systematic errors?

Look for *ANY* effect that can have an influence on your results.

Divide your data in any way you can (space, period, condition, analysis, etc.).


High Accuracy High Precision
Low Accuracy High Precision
High Accuracy Low Precision
Low Accuracy Low Precision

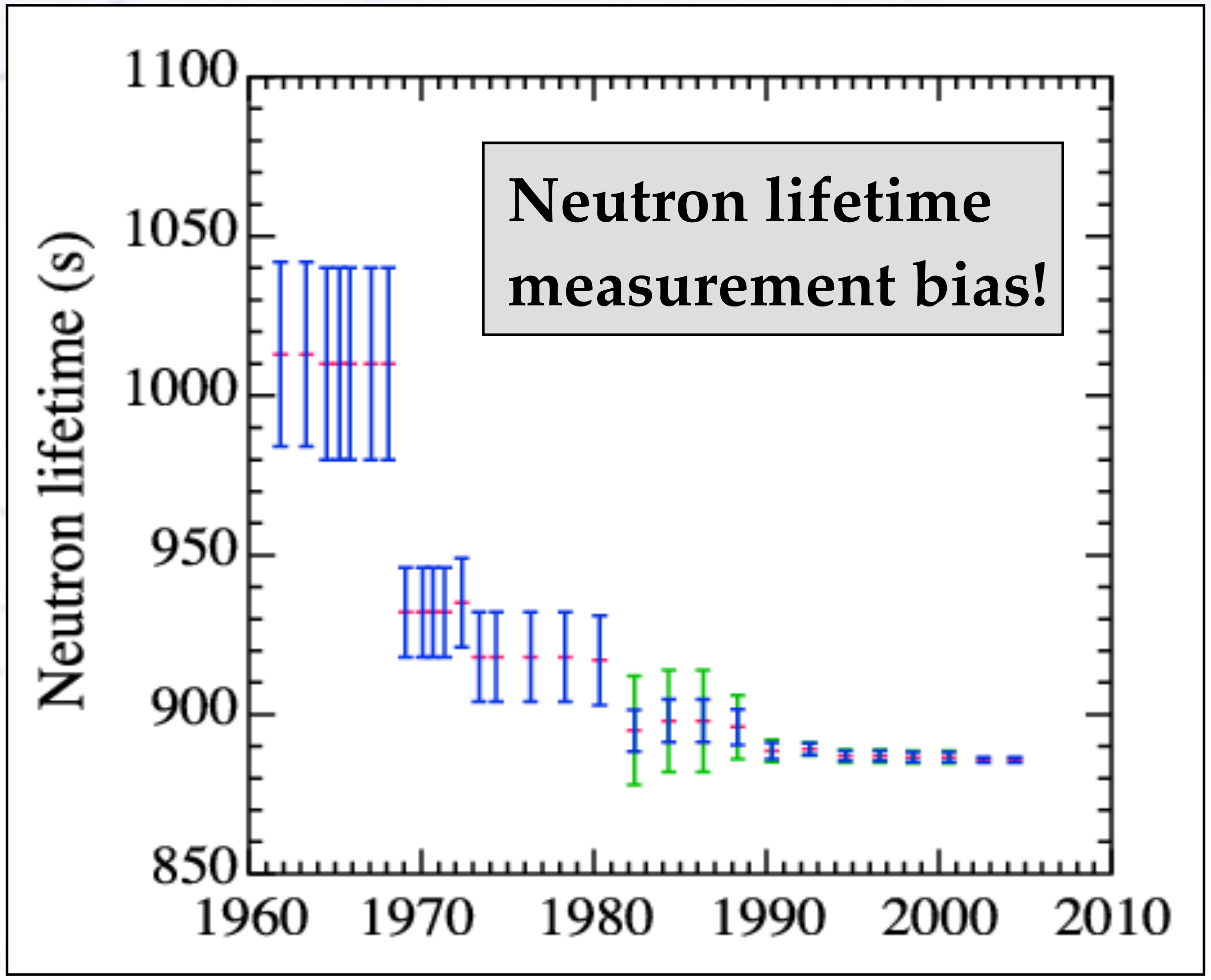
Often, systematic errors are also studied using simulation. However, this requires that the simulation is accurate! To check this, one studies known phenomena.

Biased measurements

Why does my experiment find a lower value than others?

It is questions like these, that makes you start looking for effects that could yield a higher value, leading to…

Biases!



When measuring a parameter for which there are already expectations/predictions, the result can be biased. Examples:

- Millikan's oil-drop experiment.
- Epsilon prime (CERN vs. FNAL).
- Most politically influenced decisions!

Those who forget good and evil and seek only the facts are more likely to achieve good, than those who view the world through the distorting medium of their own desires. [Bertrand Russell]

The charge of an electron

We have learned a lot from experience about how to handle some of the ways we fool ourselves. *One example: Millikan measured the charge on an electron by an experiment with falling oil drops, and got an answer which we now know not to be quite right. It's a little bit off because he had the incorrect value for the viscosity of air. It's interesting to look at the history of measurements of the charge of an electron, after Millikan. If you plot them as a function of time, you find that one is a little bit bigger than Millikan's, and the next one's a little bit bigger than that, and the next one's a little bit bigger than that, until finally they settle down to a number which is higher.*

Why didn't they discover the new number was higher right away? It's a thing that scientists are ashamed of—this history—because it's apparent that people did things like this: ***When they got a number that was too high above Millikan's, they thought something must be wrong—and they would look for and find a reason why something might be wrong.*** *When they got a number close to Millikan's value they didn't look so hard. And so they eliminated the numbers that were too far off, and did other things like that …*

[Richard Feynmann]

Blinding of results

To avoid experimenters biases, **blinding** has been introduced.

This means that the computer adds a random number to the result, which is not removed before the analysis has been thoroughly checked.

Example:

```
> ./FitSin2beta
Result is: sin(2beta) = x.xx +- 0.37
Do you wish to unblind (y/n)?
```

Emblem used by the BaBar experiment to label blinded analysis

This was first introduced by the French Academy of Science (1784), and has since become standard procedure in most science and medical experiments.

In this way experimenters bias is removed, and the results become truly independent and unaffected by wishful thinking and “common belief”.

Systematic Errors

Even with *infinite* statistics, the error on a result will never be zero!

Such "systematic uncertainties" have many origins, some of which are:

- Imperfect modeling/simulation
- Lacking understanding of experiment
- Uncertainty in parameters involved
- Uncertainty associated with corrections
- Theoretical uncertainties/limitations

While the ***statistical uncertainty*** is Gaussian and scales like $1/\sqrt{N}$, the ***systematic uncertainties*** do not necessarily follow this rule.

When **statistical** uncertainty dominate, more **data** will improve precision.
When **systematic** uncertainty dominate, more **understanding** will improve precision.

In these modern days of particle factories and huge data samples, systematic uncertainties play a significant role.

The finding/calculation of systematic errors is hard work.

Systematic uncertainty examples

Think of examples of systematic uncertainties on:

a. Track momentum
b. Cluster energy
c. Reconstruction efficiency

Systematic uncertainty

Think of examples of systematic uncertainties on:

a. Track momentum:

   Determined as $p = 0.3\ B\ \varrho$, thus uncertainties come from the magnetic field strength (B) and curvature (ϱ) uncertainties. Interactions with detector material could also be a source.

b. Cluster energy

c. Reconstruction efficiency

Systematic uncertainty

Think of examples of systematic uncertainties on:

a. Track momentum:
Determined as $p = 0.3\, B\, \varrho$, thus uncertainties come from the magnetic field strength (B) and curvature (ϱ) uncertainties. Interactions with detector material could also be a source.

b. Cluster energy:
Simplest case $E = \alpha S + \beta$, thus uncertainties come from calibration coefficients α and β. As a calorimeter is more complex, this can be expanded to many variables, and also as a function of E and angles.

c. Reconstruction efficiency

Systematic uncertainty

Think of examples of systematic uncertainties on:

a. Track momentum:
   Determined as $p = 0.3\ B\ \varrho$, thus uncertainties come from the magnetic field strength (B) and curvature (ϱ) uncertainties. Interactions with detector material could also be a source.

b. Cluster energy:
   Simplest case $E = \alpha S + \beta$, thus uncertainties come from calibration coefficients α and β. As a calorimeter is more complex, this can be expanded to many variables, and also as a function of E and angles.

c. Reconstruction efficiency
   Typically, $\varepsilon = (N_{pass} - Bkg_1) / (N_{total} - Bkg_2)$, where Bkg_x are the background numbers. These of course carry uncertainties.

Systematic uncertainty

Think of examples of systematic uncertainties on:

a. Track momentum:
Determined as $p = 0.3\ B\ \varrho$, thus uncertainties come from the magnetic field strength (B) and curvature (ϱ) uncertainties. Interactions with detector material could also be a source.

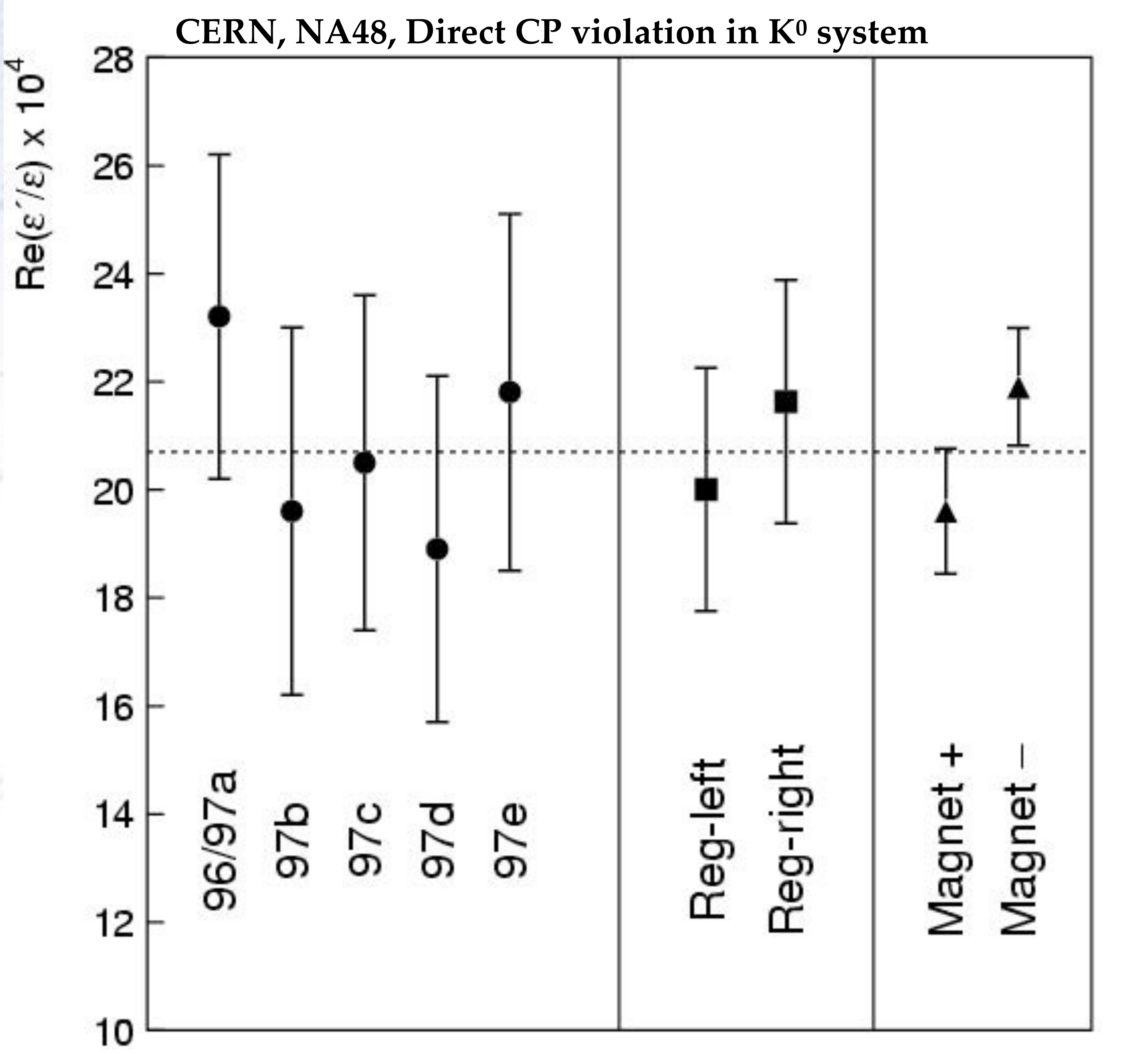
b. Cluster energy:
Simplest case $E = \alpha S + \beta$, thus uncertainties come from calibration coefficients α and β. As a calorimeter is more complex, this can be expanded to many variables, and also as a function of E and angles.

c. Reconstruction efficiency
Typically, $\varepsilon = (N_{pass} - Bkg_1) \ / \ (N_{total} - Bkg_2)$, where Bkg_x are the background numbers. These of course carry uncertainties.

Common for all cases are, that they can be written as "simple" formulae. When this is no so (e.g. ML output), one can resort to using simulation to propagate the uncertainties.

Cross check of data

CERN, NA48, Direct CP violation in K^0 system



Classic check of systematic errors, by dividing the data according to:

- Period of data taking
- Direction of regulator
- Direction of B-field

If any of these showed an inconsistency between the subsamples, one would know that this had an impact on the result.

This type of cross checks is at the heart of data analysis.

Checking your analysis

Do as many tests of your analysis as possible. You can’t prove that your analysis is correct, but the more tests and checks it passes, the more likely it is that your colleagues (and yourself!) will trust the result.

However, if the analysis passes a check, don’t include a possible (small) discrepancy as a systematic uncertainty!

1. It would penalise hard work and diligence
2. It is illogical (discrepancy is not significant!)
3. It inflates final error beyond its true size

However, re-check results, discuss with colleagues, vary all parameters and cuts, and *be your own worst critic*.

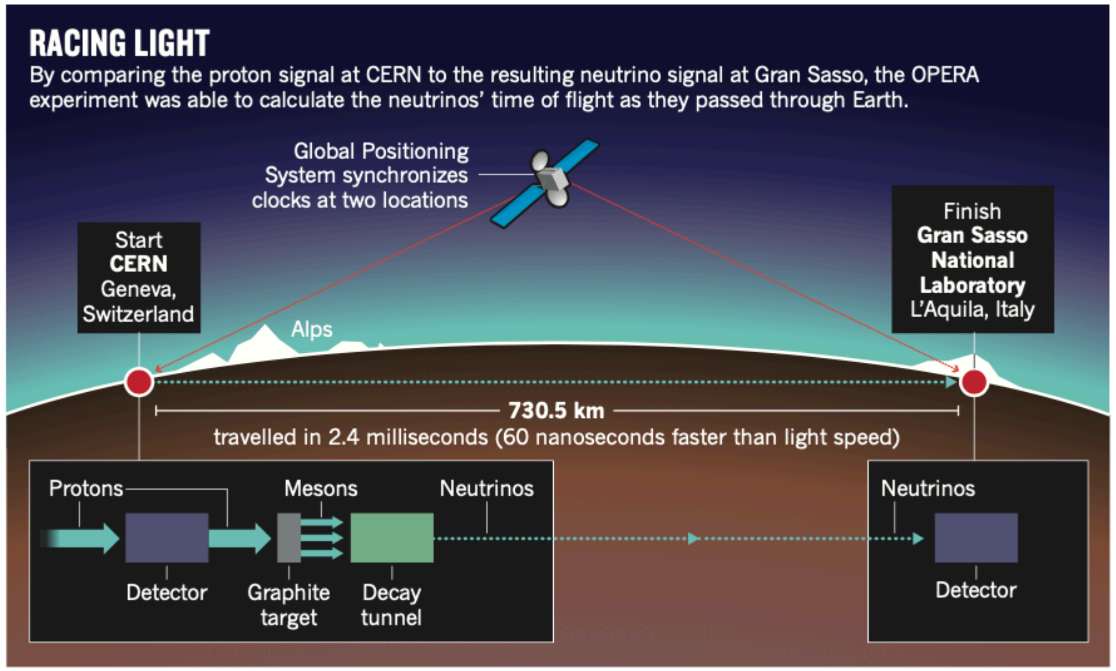
You may think that this is hard, but it carries great satisfaction and also peace of mind.

Example of systematic error

One of the best “recent” examples is the case of physicists measuring neutrinos to travel faster than speed of light.

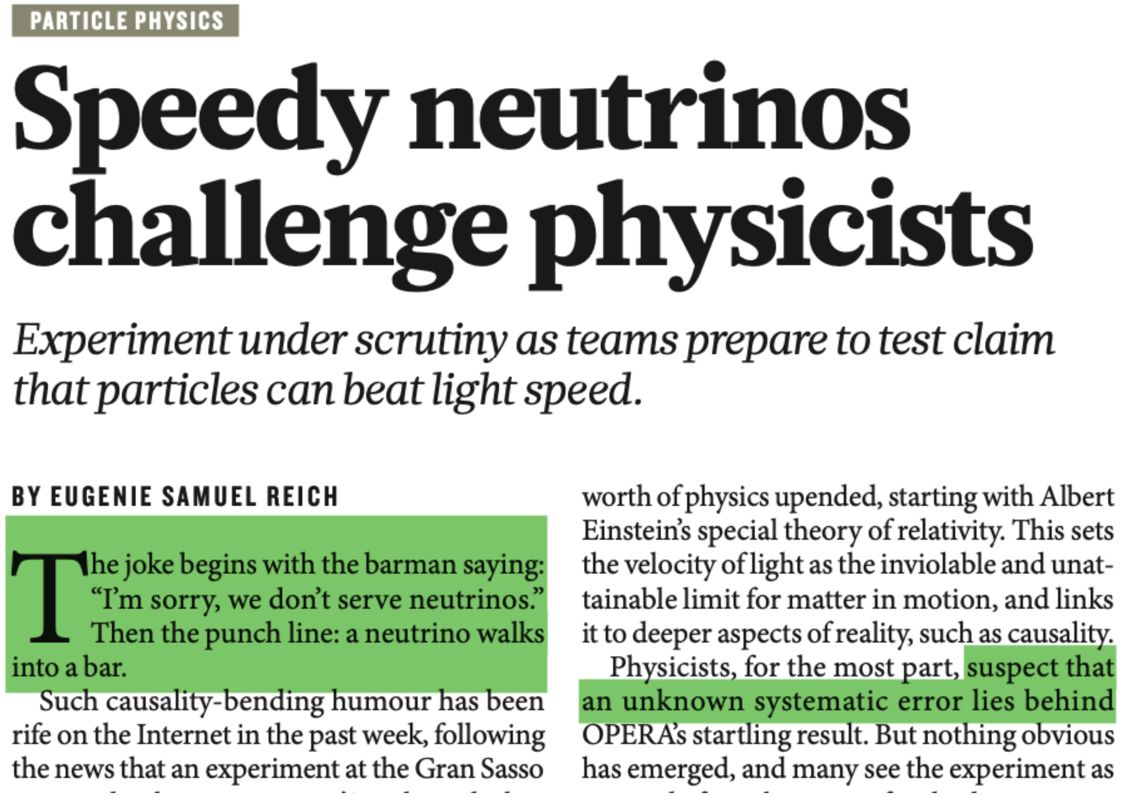
This would (if true) put the foundations of General Relativity in ruins - but be interesting! After 6 months of intense studies, the researchers found two possible systematic errors:

- A link from a GPS receiver to the OPERA master clock was loose, which increased the delay through the fiber.
- A clock on an electronic board ticked faster than its expected 10 MHz frequency, lengthening the reported flight-time of neutrinos, thereby somewhat reducing the seeming faster-than-light effect.


RACING LIGHT

By comparing the proton signal at CERN to the resulting neutrino signal at Gran Sasso, the OPERA experiment was able to calculate the neutrinos' time of flight as they passed through Earth.

Global Positioning System synchronizes clocks at two locations
Start CERN Geneva, Switzerland
Finish Gran Sasso National Laboratory L'Aquila, Italy
Alps
730.5 km
travelled in 2.4 milliseconds (60 nanoseconds faster than light speed)
Protons
Mesons
Neutrinos
Detector
Graphite target
Decay tunnel
Neutrinos
Detector


PARTICLE PHYSICS

Speedy neutrinos challenge physicists

Experiment under scrutiny as teams prepare to test claim that particles can beat light speed.

BY EUGENIE SAMUEL REICH

The joke begins with the barman saying: “I'm sorry, we don't serve neutrinos.” Then the punch line: a neutrino walks into a bar.

Such causality-bending humour has been rife on the Internet in the past week, following the news that an experiment at the Gran Sasso

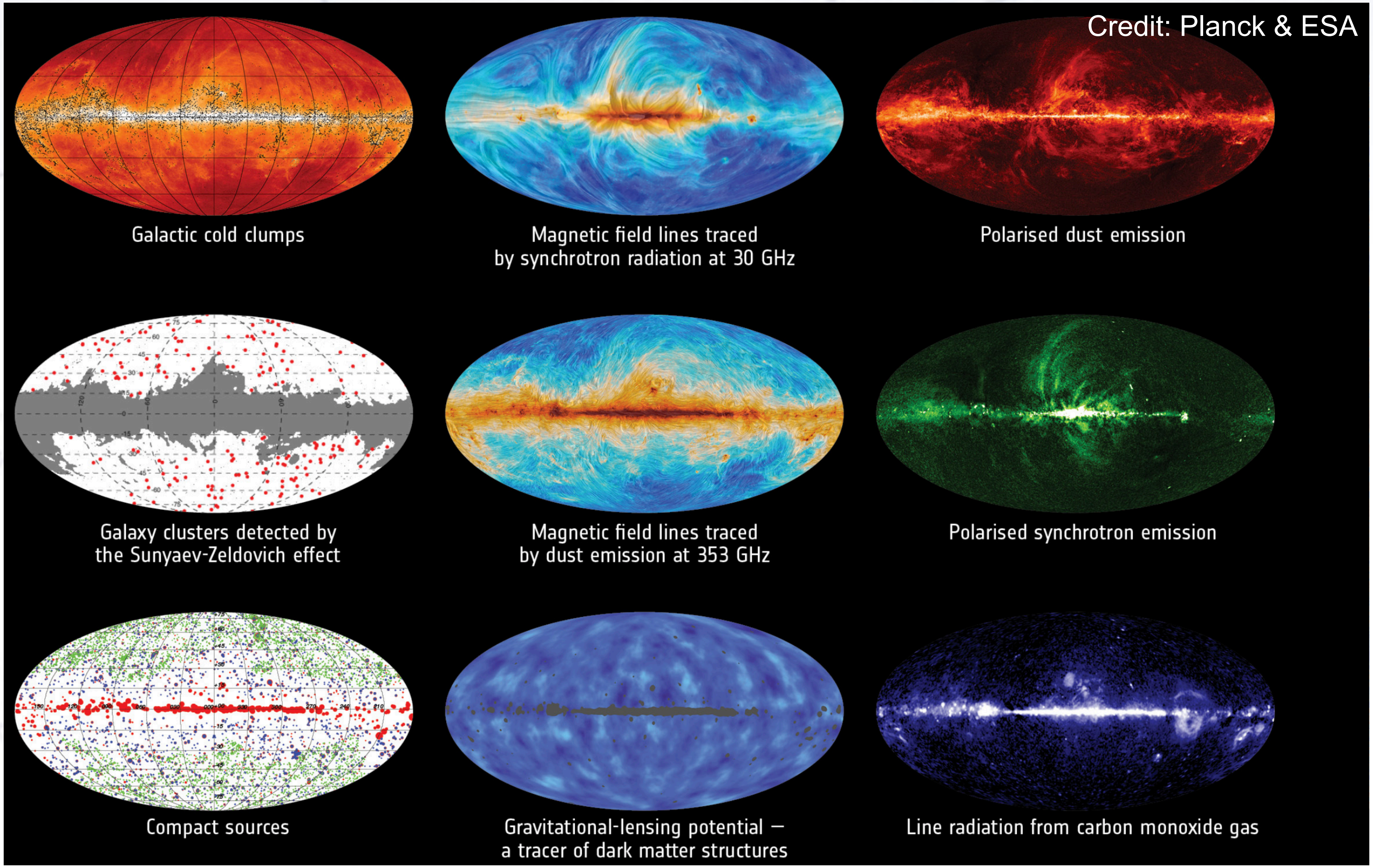
worth of physics upended, starting with Albert Einstein's special theory of relativity. This sets the velocity of light as the inviolable and unattainable limit for matter in motion, and links it to deeper aspects of reality, such as causality.

Physicists, for the most part, suspect that an unknown systematic error lies behind OPERA's startling result. But nothing obvious has emerged, and many see the experiment as

Unchecked biases

No method of checking for biases or systematics errors is fool proof. Overconfidence that all dominant systematic errors are included can result in wrong results.

Measuring the cosmic microwave background requires many subtractions of unwanted foregrounds. Missing a single systematic contribution ruins results.



Credit: Planck & ESA

Galactic cold clumps

Magnetic field lines traced by synchrotron radiation at 30 GHz

Polarised dust emission

Galaxy clusters detected by the Sunyaev-Zeldovich effect

Magnetic field lines traced by dust emission at 353 GHz

Polarised synchrotron emission

Compact sources

Gravitational-lensing potential — a tracer of dark matter structures

Line radiation from carbon monoxide gas

Evaluating systematic errors

Known sources:

- Error on factors in the analysis, energy calibration, efficiencies, corrections, ...
- Error on external input: theory error, error on temperature, masses, ...

Evaluate from varying conditions, and compute result for each. Error is RMSE.

For combining systematic uncertainties, the correlations (covariance matrix) is needed. Correlations typically come from common input (e.g. luminosity).

Unsuspected sources:

Repeating the analysis in different form helps to find such systematic effects.

- Use subset of data, or change selection of data used in analysis.
- Change histogram binning, change parameterisations, change fit techniques.
- Look for impossibilities.

If you do not a priori expect a systematic effect and if the deviation is not significant, then do not add this in the systematic uncertainties. If there is a deviation, try to understand, where the mistake is and fix it!

Only as a last resort include non-understood discrepancy as systematic error.

Discrete systematic errors

Discrete uncertainties are special. They typically arise from model choice.

Situation depends on status of model. Sometimes one model is preferred, sometimes all models are equal (more or less). Imagine two models each yielding estimate μ.

With 1 preferred model and one other, quote $\boldsymbol{\mu_1 \pm |\mu_1 - \mu_2|}$

With 2 models of equal status, quote $\boldsymbol{(\mu_1 + \mu_2)/2 \pm |\mu_1 - \mu_2| /}$ **sqrt(2)**

With N models, quote $\boldsymbol{\bar{\mu}}$ ± **sqrt(N / (N-1)) * Std.**

With 2 extreme models, quote $\boldsymbol{(\mu_1 + \mu_2)/2 \pm |\mu_1 - \mu_2| /}$ **sqrt(12)**

These are just rough estimates. Do not push them too hard.

If the difference is not small, you have a problem - which can be an opportunity to study model differences.

When all else fails

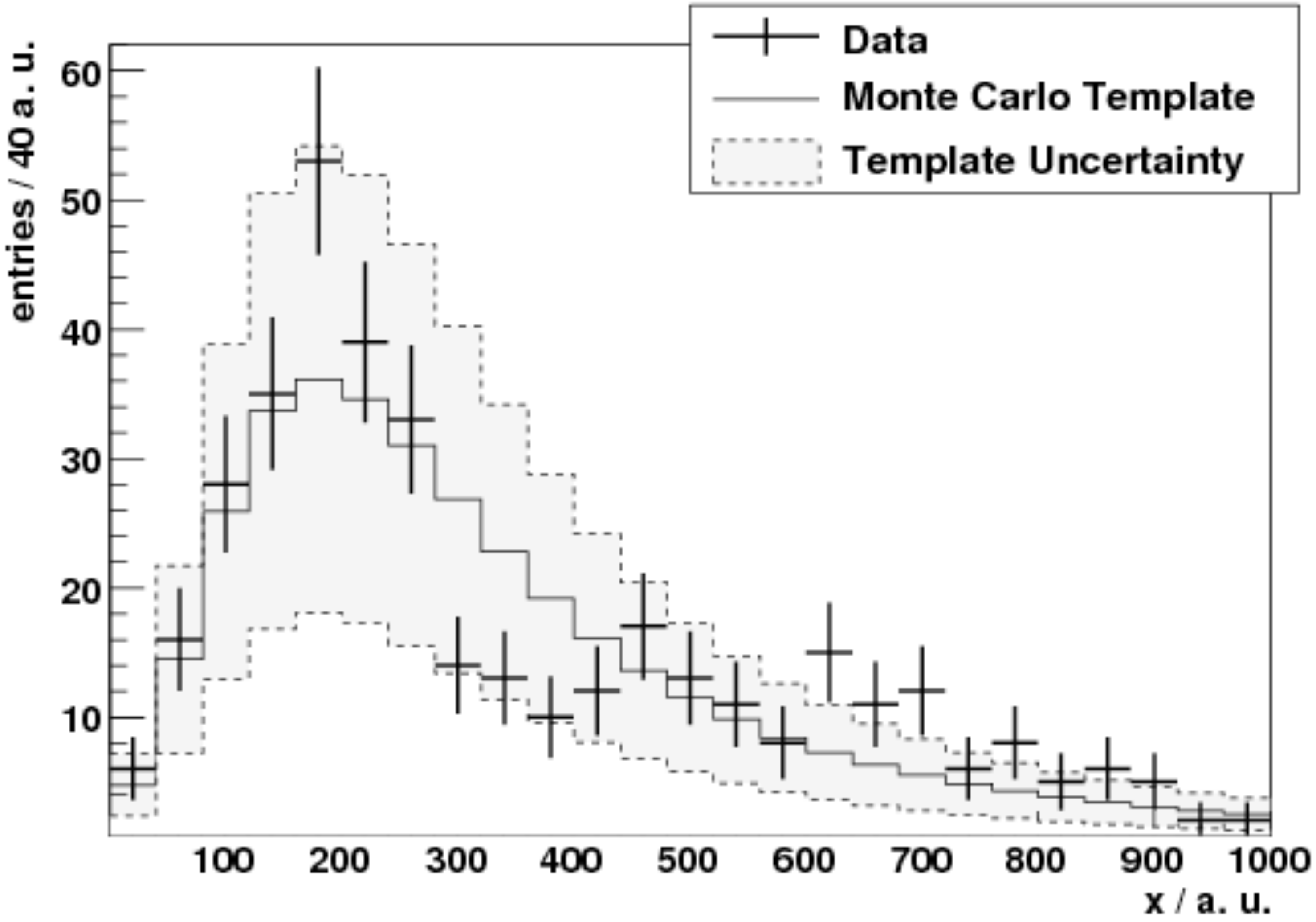
A sign of a systematic error (or bug), is that one can see in data, that "something" strange is going on.

It could be that a distribution between data and MC is very different, or that two data (e.g. control) channels disagree on the size of a systematic correction.

One should of course work hard to understand the effect, but occasionally one must give up, and suffer a large systematic uncertainty.

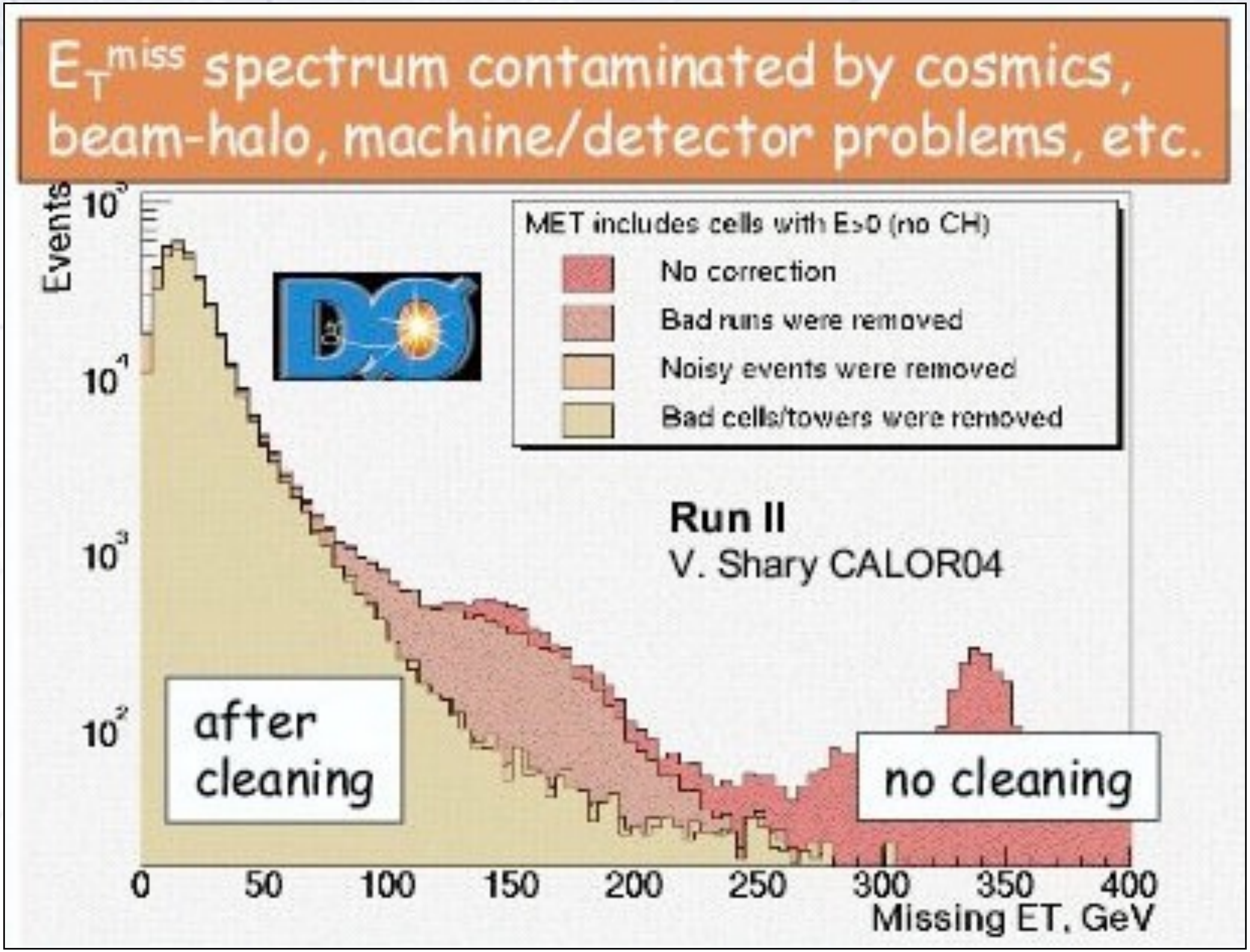
The example distribution shown is one such case. It looks as if some effect in data is at play around 280 GeV.

But what and why not in MC?



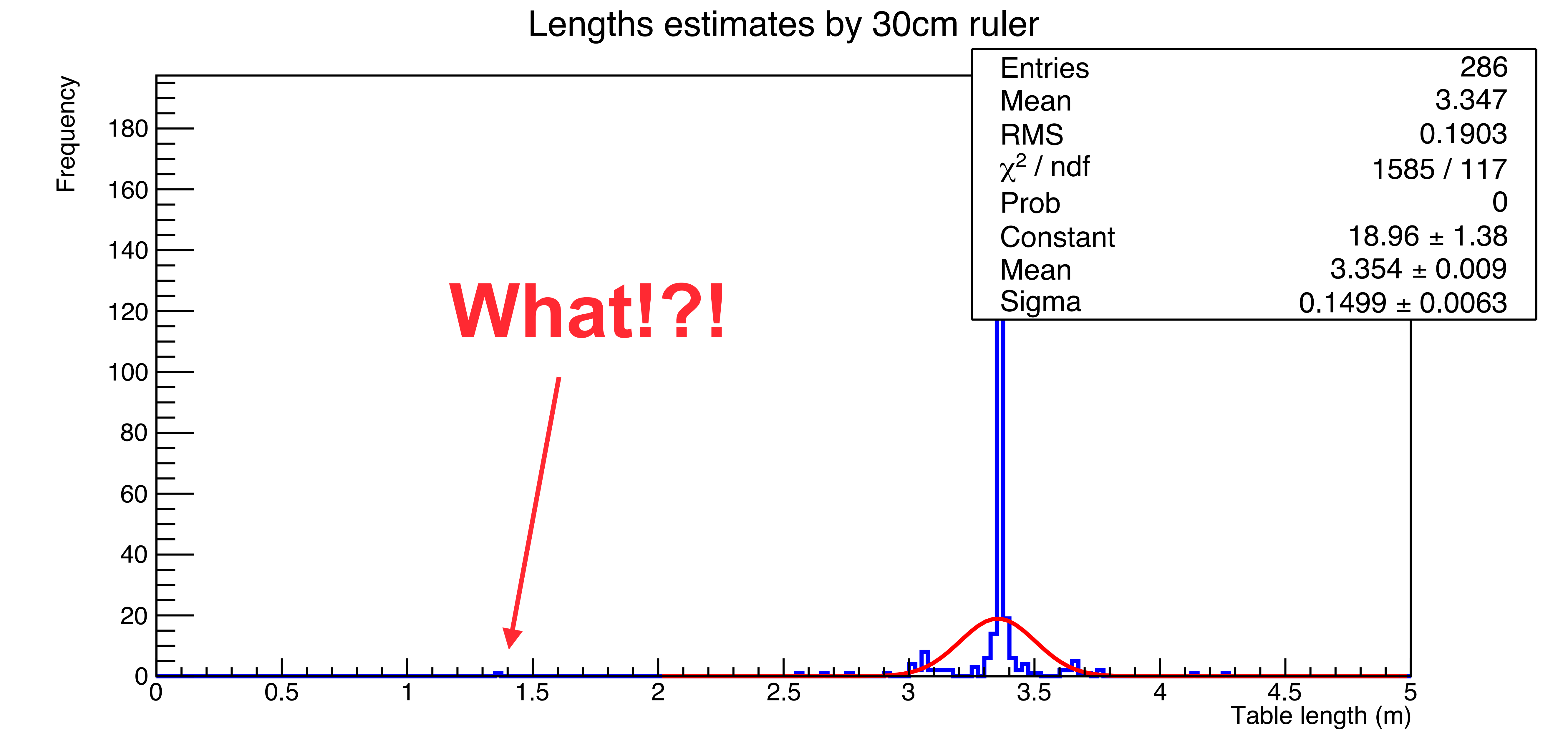
Cleaning data

Example of experimental error, which would be a disaster if not corrected for.



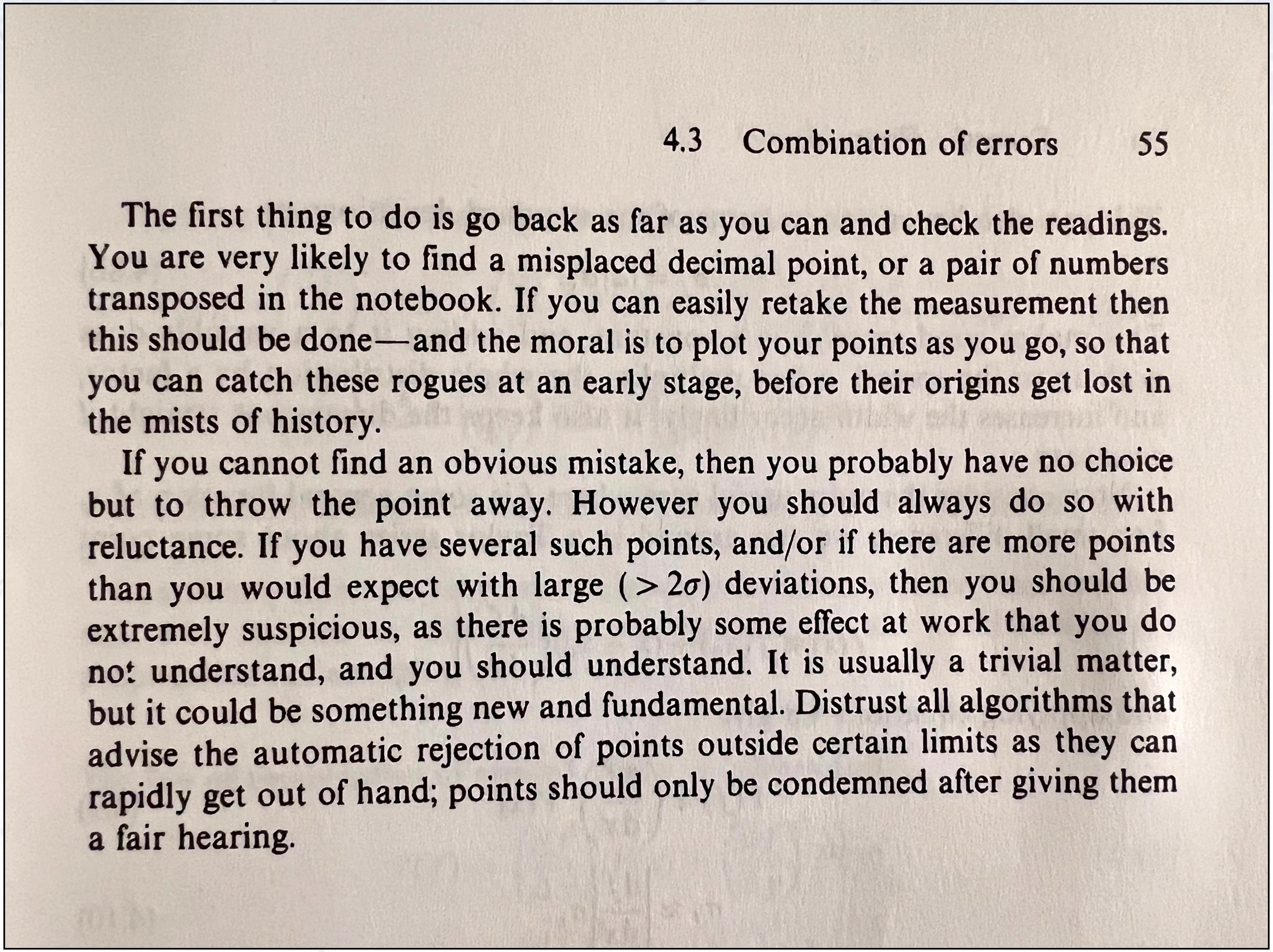
Removing data points

An example could be in some of the Table Measurement exercise data…



Caution discarding data!

The following passage (p. 55 of Barlow) is an intersting read:



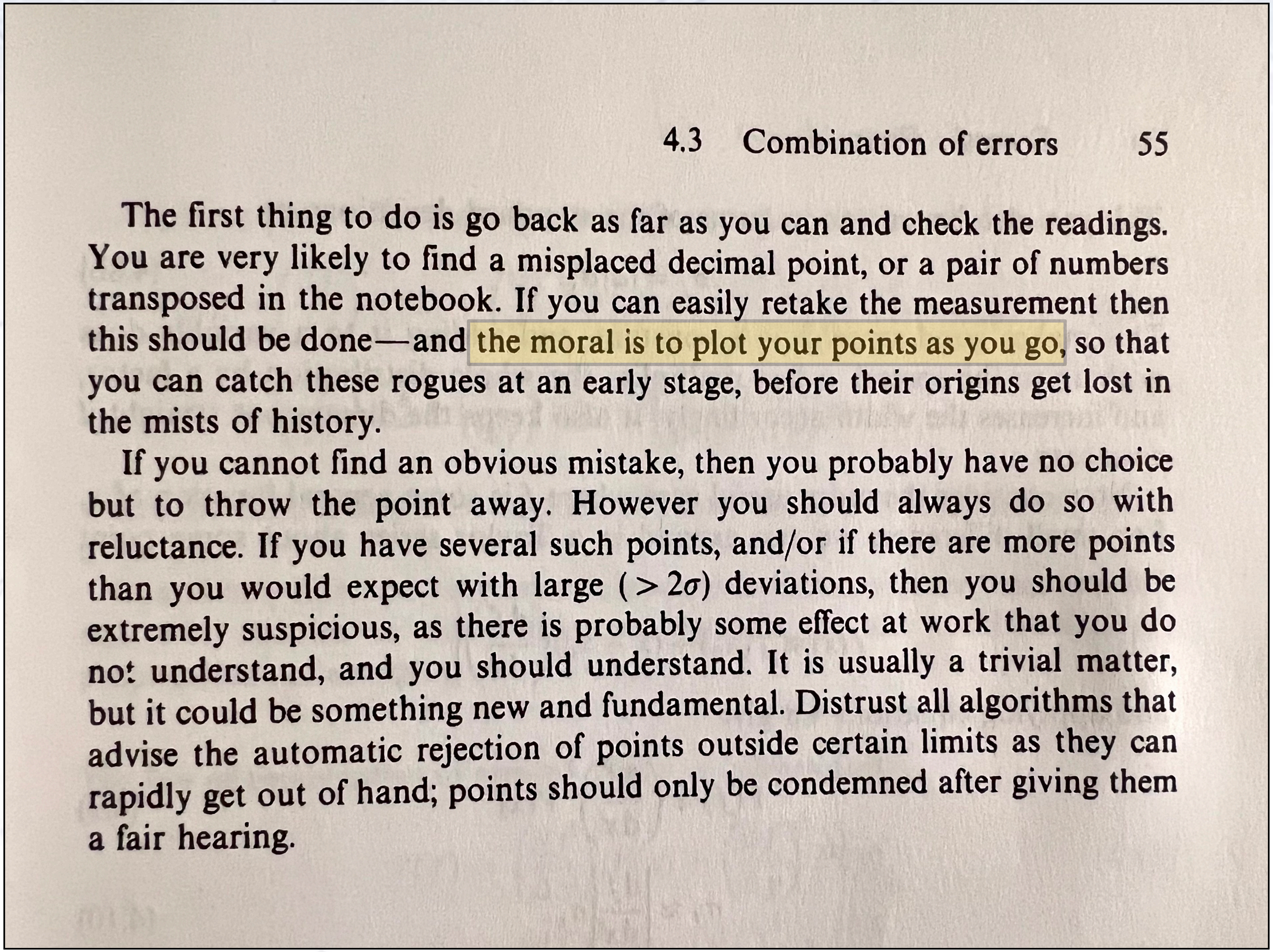
4.3 Combination of errors 55

The first thing to do is go back as far as you can and check the readings. You are very likely to find a misplaced decimal point, or a pair of numbers transposed in the notebook. If you can easily retake the measurement then this should be done—and the moral is to plot your points as you go, so that you can catch these rogues at an early stage, before their origins get lost in the mists of history.

If you cannot find an obvious mistake, then you probably have no choice but to throw the point away. However you should always do so with reluctance. If you have several such points, and/or if there are more points than you would expect with large ($>2\sigma$) deviations, then you should be extremely suspicious, as there is probably some effect at work that you do not understand, and you should understand. It is usually a trivial matter, but it could be something new and fundamental. Distrust all algorithms that advise the automatic rejection of points outside certain limits as they can rapidly get out of hand; points should only be condemned after giving them a fair hearing.

Caution discarding data!

The following passage (p. 55 of Barlow) is an intersting read:



4.3 Combination of errors 55

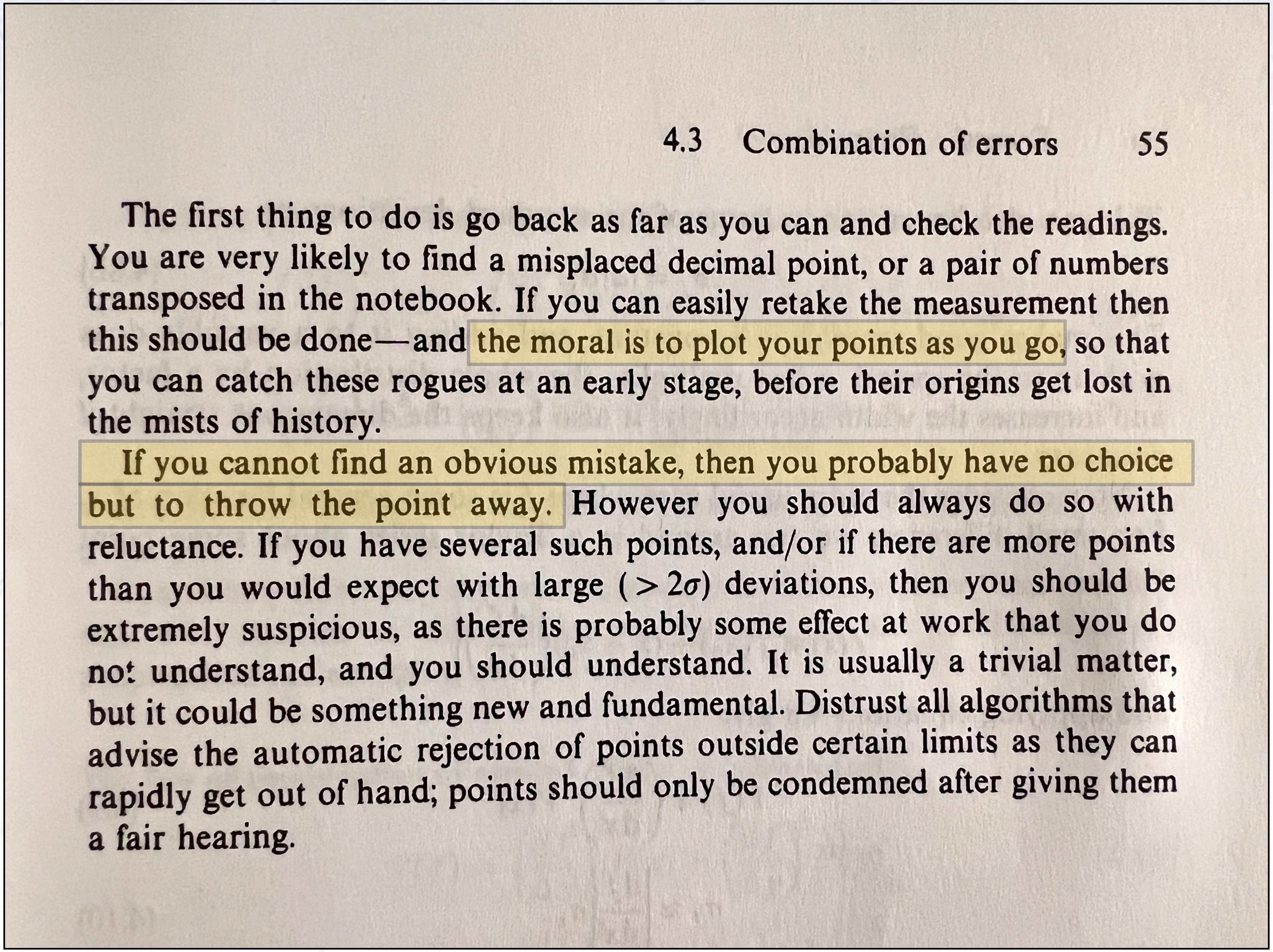
The first thing to do is go back as far as you can and check the readings. You are very likely to find a misplaced decimal point, or a pair of numbers transposed in the notebook. If you can easily retake the measurement then this should be done—and the moral is to plot your points as you go, so that you can catch these rogues at an early stage, before their origins get lost in the mists of history.

If you cannot find an obvious mistake, then you probably have no choice but to throw the point away. However you should always do so with reluctance. If you have several such points, and/or if there are more points than you would expect with large ($>2\sigma$) deviations, then you should be extremely suspicious, as there is probably some effect at work that you do not understand, and you should understand. It is usually a trivial matter, but it could be something new and fundamental. Distrust all algorithms that advise the automatic rejection of points outside certain limits as they can rapidly get out of hand; points should only be condemned after giving them a fair hearing.

Yes!

Caution discarding data!

The following passage (p. 55 of Barlow) is an intersting read:



4.3 Combination of errors 55

The first thing to do is go back as far as you can and check the readings. You are very likely to find a misplaced decimal point, or a pair of numbers transposed in the notebook. If you can easily retake the measurement then this should be done—and the moral is to plot your points as you go, so that you can catch these rogues at an early stage, before their origins get lost in the mists of history.

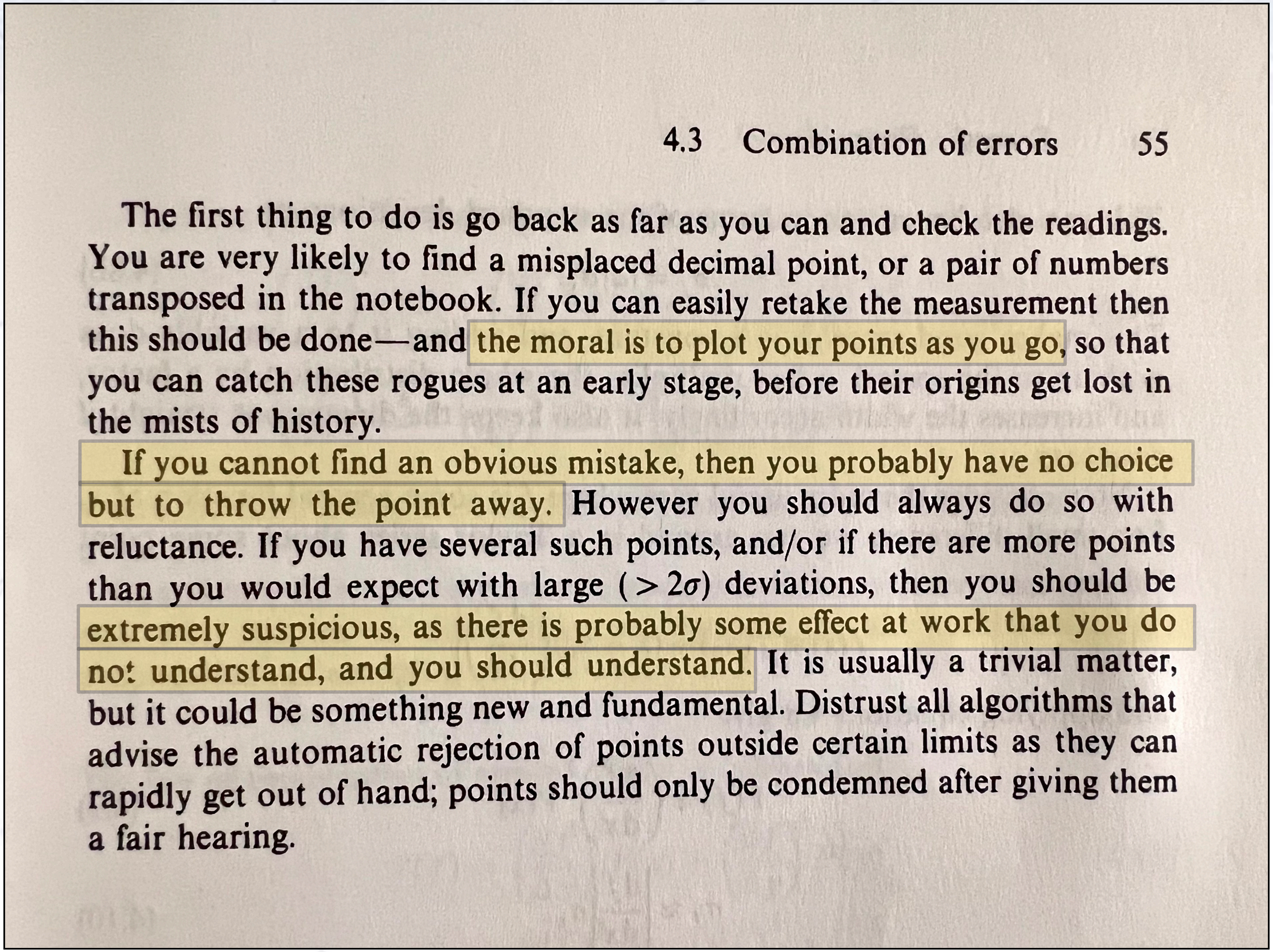
If you cannot find an obvious mistake, then you probably have no choice but to throw the point away. However you should always do so with reluctance. If you have several such points, and/or if there are more points than you would expect with large ($>2\sigma$) deviations, then you should be extremely suspicious, as there is probably some effect at work that you do not understand, and you should understand. It is usually a trivial matter, but it could be something new and fundamental. Distrust all algorithms that advise the automatic rejection of points outside certain limits as they can rapidly get out of hand; points should only be condemned after giving them a fair hearing.

Yes!

Hmm…

Caution discarding data!

The following passage (p. 55 of Barlow) is an intersting read:



4.3 Combination of errors 55

The first thing to do is go back as far as you can and check the readings. You are very likely to find a misplaced decimal point, or a pair of numbers transposed in the notebook. If you can easily retake the measurement then this should be done—and the moral is to plot your points as you go, so that you can catch these rogues at an early stage, before their origins get lost in the mists of history.

If you cannot find an obvious mistake, then you probably have no choice but to throw the point away. However you should always do so with reluctance. If you have several such points, and/or if there are more points than you would expect with large ($>2\sigma$) deviations, then you should be extremely suspicious, as there is probably some effect at work that you do not understand, and you should understand. It is usually a trivial matter, but it could be something new and fundamental. Distrust all algorithms that advise the automatic rejection of points outside certain limits as they can rapidly get out of hand; points should only be condemned after giving them a fair hearing.

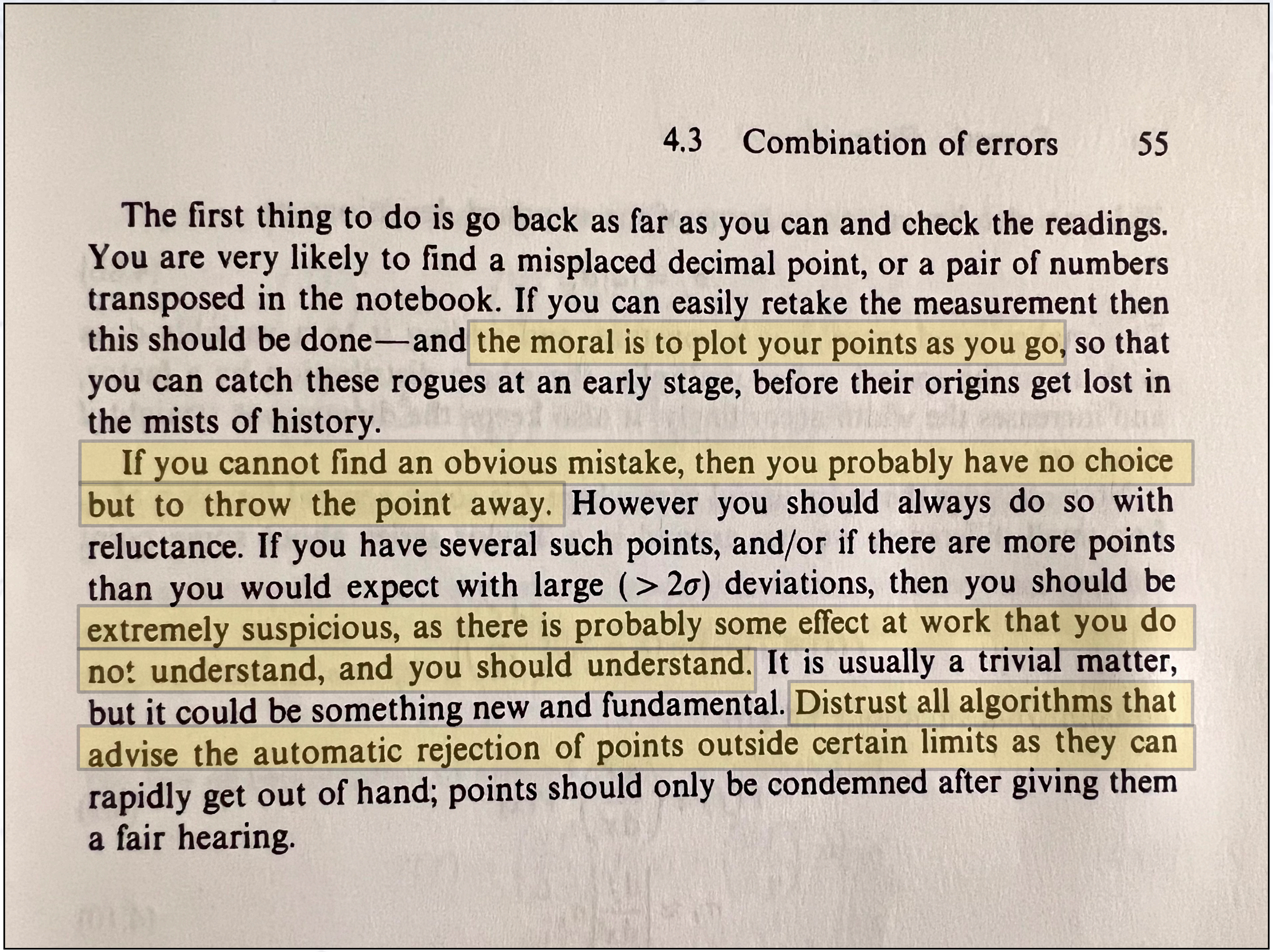
Yes!

Hmm…

Yes!

Caution discarding data!

The following passage (p. 55 of Barlow) is an intersting read:



4.3 Combination of errors 55

The first thing to do is go back as far as you can and check the readings. You are very likely to find a misplaced decimal point, or a pair of numbers transposed in the notebook. If you can easily retake the measurement then this should be done—and the moral is to plot your points as you go, so that you can catch these rogues at an early stage, before their origins get lost in the mists of history.

If you cannot find an obvious mistake, then you probably have no choice but to throw the point away. However you should always do so with reluctance. If you have several such points, and/or if there are more points than you would expect with large ($>2\sigma$) deviations, then you should be extremely suspicious, as there is probably some effect at work that you do not understand, and you should understand. It is usually a trivial matter, but it could be something new and fundamental. Distrust all algorithms that advise the automatic rejection of points outside certain limits as they can rapidly get out of hand; points should only be condemned after giving them a fair hearing.

Yes!

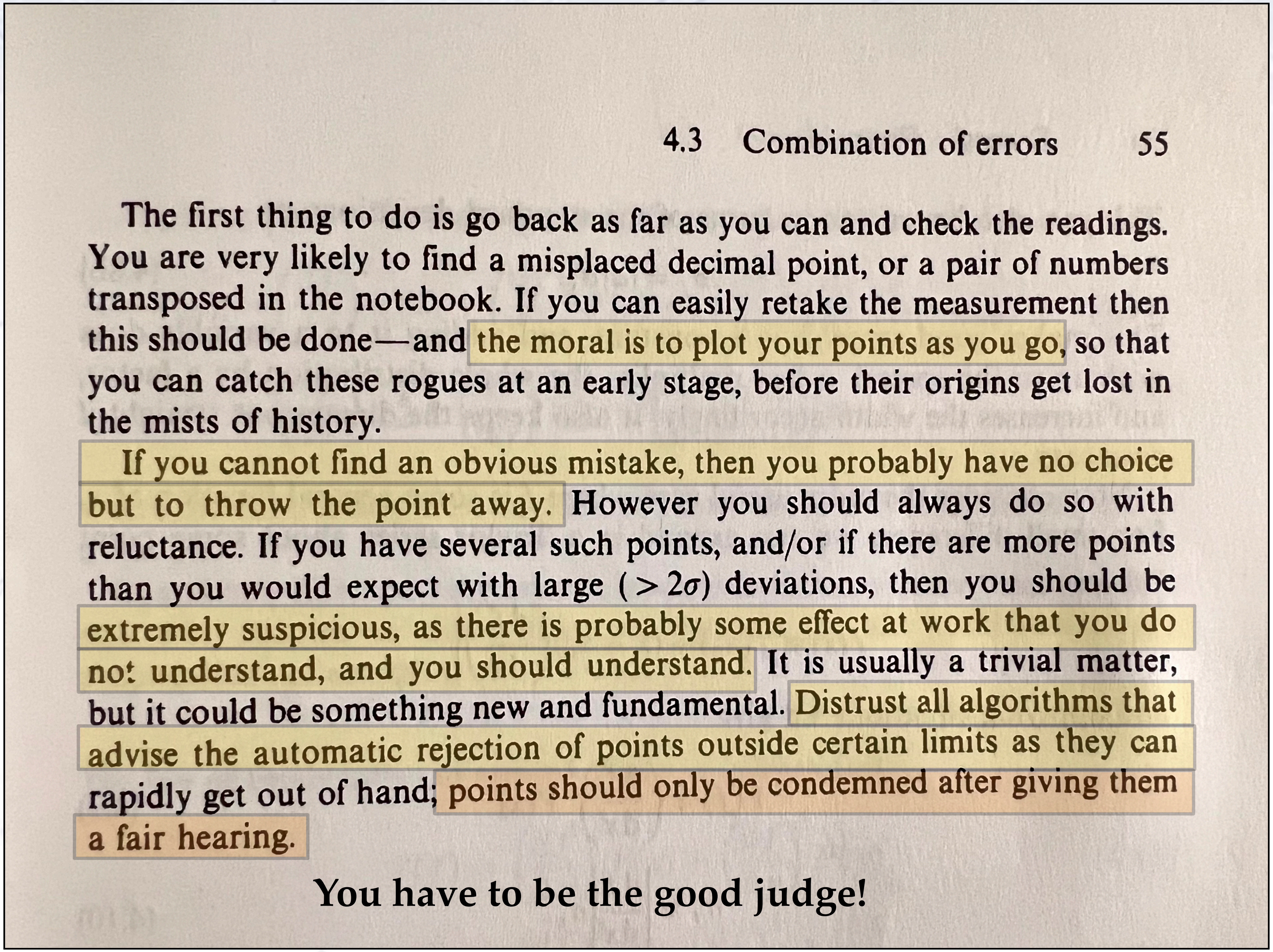
Hmm…

Yes!

YES!

Caution discarding data!

The following passage (p. 55 of Barlow) is an intersting read:



4.3 Combination of errors 55

The first thing to do is go back as far as you can and check the readings. You are very likely to find a misplaced decimal point, or a pair of numbers transposed in the notebook. If you can easily retake the measurement then this should be done—and the moral is to plot your points as you go, so that you can catch these rogues at an early stage, before their origins get lost in the mists of history.

If you cannot find an obvious mistake, then you probably have no choice but to throw the point away. However you should always do so with reluctance. If you have several such points, and/or if there are more points than you would expect with large ($> 2\sigma$) deviations, then you should be extremely suspicious, as there is probably some effect at work that you do not understand, and you should understand. It is usually a trivial matter, but it could be something new and fundamental. Distrust all algorithms that advise the automatic rejection of points outside certain limits as they can rapidly get out of hand; points should only be condemned after giving them a fair hearing.

Yes!

Hmm…

Yes!

YES!

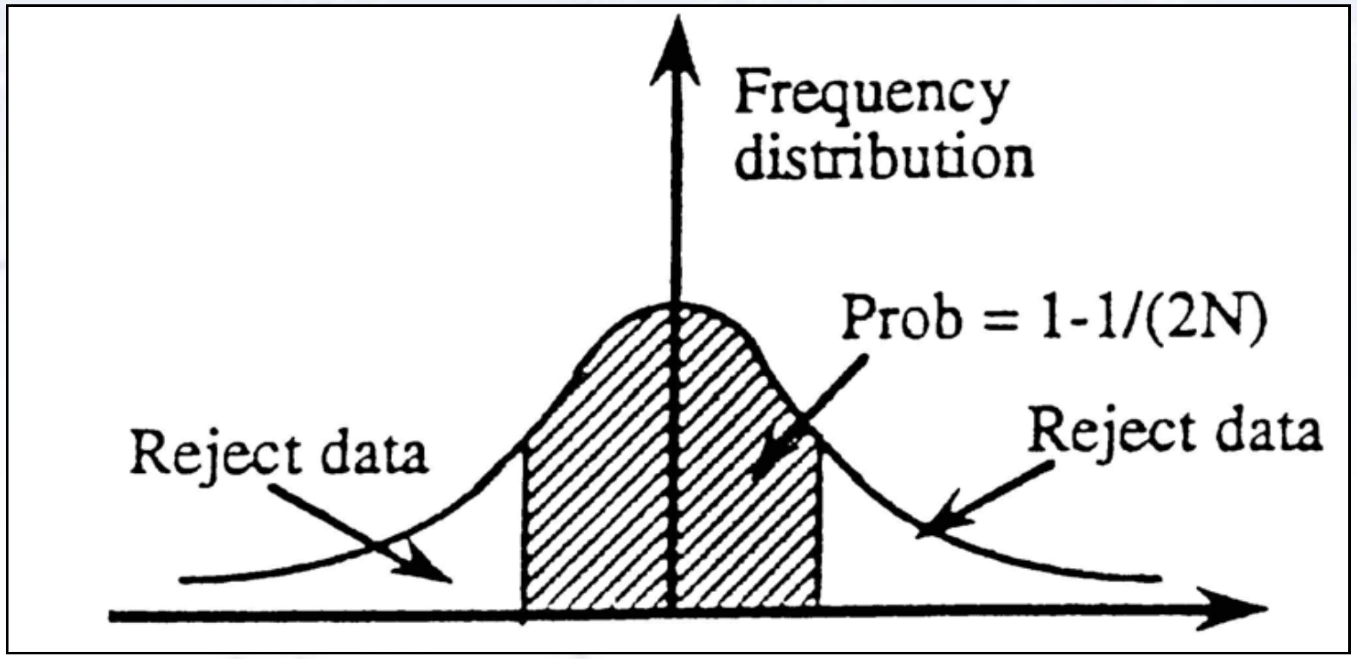
You have to be the good judge!

Removing data points

One should always be careful about removing data points, yet at the same to be willing to do so, if very good arguments can be found:

- It is an error measurement.
- Measurement is improbable.

Removing improbable data points is formalised in **Chauvenet's Criterion,** though many other methods exists (Pierce, Grubbs, etc.)



The idea is to assume that the distribution is Gaussian, and ask what the probability of the farthest point is. If it is below some value (which is preferably to be determined ahead of applying the criterion), then the point is removed, and the criterion is reapplied until no more points should be removed.

However, **ALWAYS keep a record of your original data**, as it may contain more effects than you originally thought.

Summary

Roger Barlow has six “commandments” to which I’ve added a seventh:

1. Thou shalt never say ‘systematic error’ when thou meanest ‘systematic effect’ or ‘systematic mistake’.
2. Thou shalt know at all times whether what thou performest is a check for a mistake or an evaluation of an uncertainty.
3. Thou shalt keep thy signal region out of sight until thy analysis has passed all checks and evaluated all uncertainties.
4. Thou shalt not incorporate successful check results into thy total systematic error and make thereby a shield to hide thy dodgy result.
5. Thou shalt not incorporate failed check results unless thou art truly at thy wits’ end.
6. Thou shalt not add uncertainties on uncertainties in quadrature. If they are larger than chickenfeed thou shalt generate more Monte Carlo until they shrink to become so.
7. Thou shalt say what thou doest, and thou shalt be able to justify it out of thine own mouth; not the mouth of thy supervisor, nor thy colleague who did the analysis last time, nor thy local statistics guru, nor thy mate down the pub.

Do these, and thou shalt flourish, and thine analysis likewise.