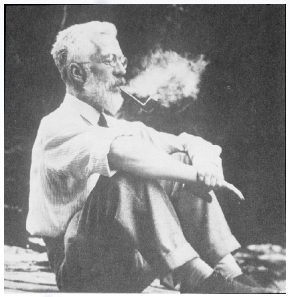
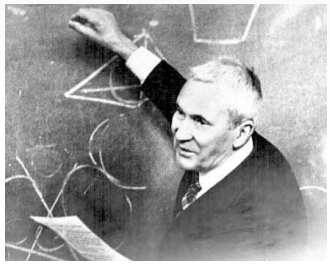
Applied Statistics

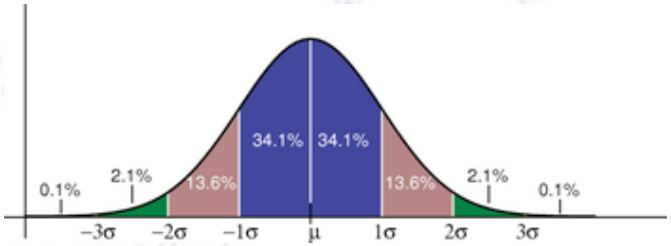
Hypothesis Testing





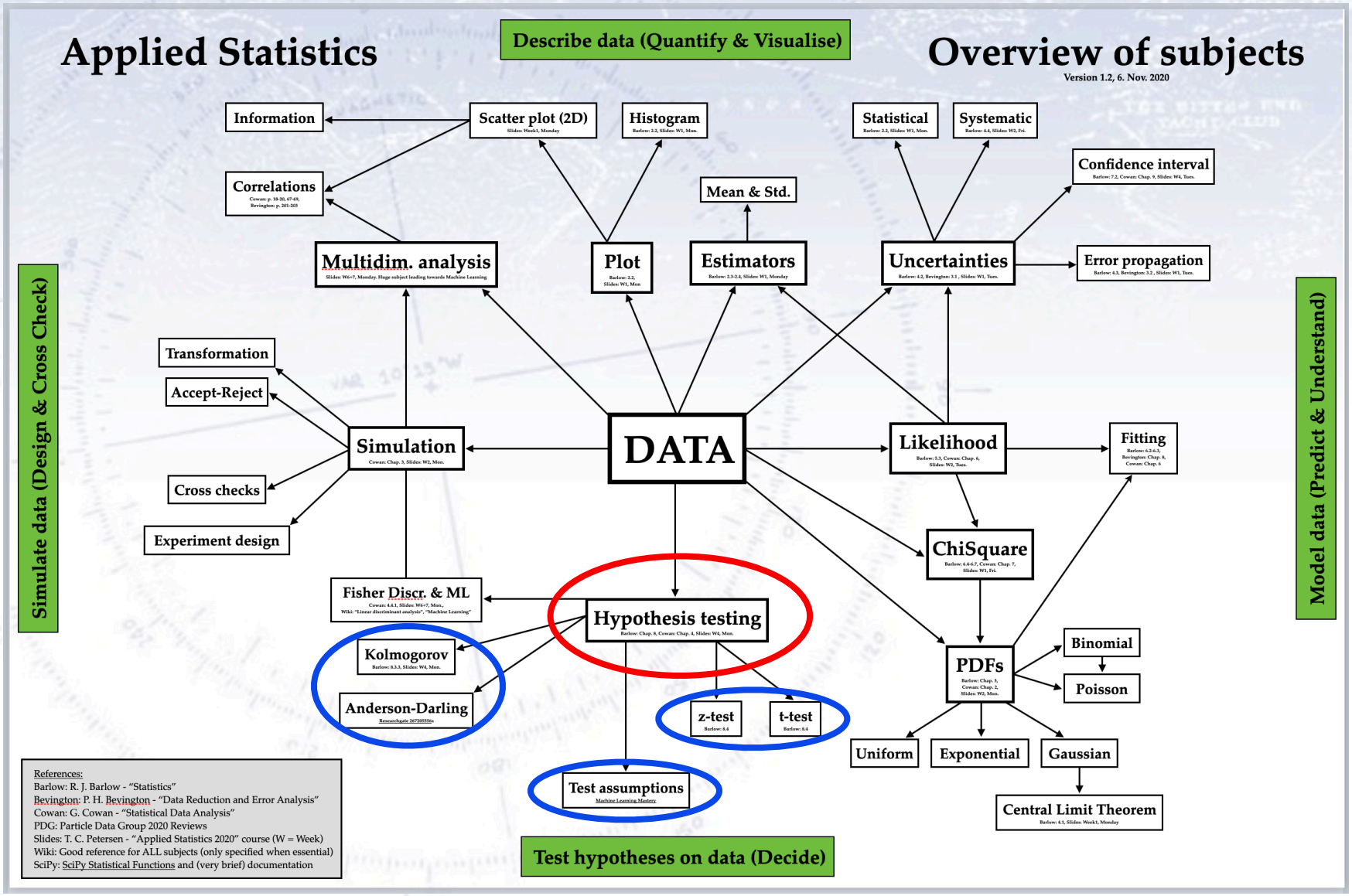



Troels C. Petersen (NBI)



"Statistics is merely a quantisation of common sense"

Probability Density Functions



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…

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 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**	**Wrong decision Type II error**
Conviction (Reject H_0)	**Wrong decision Type I error**	**Right decision**

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 = 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 = 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 or give judgement - it is yes-or-no.

Given data - how to do that best?

That is the basic question in hypothesis testing.

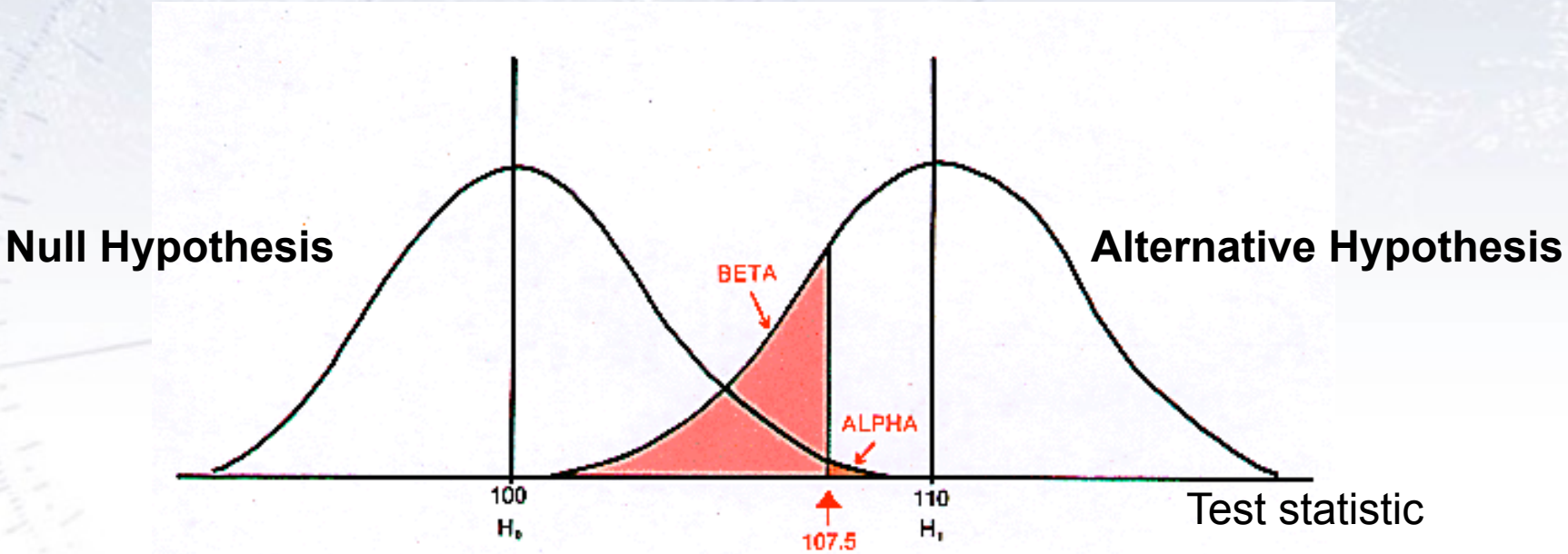
Trouble is, you may take the wrong decision, and there are TWO errors:

- The hypothesis is **true**, but you **reject** it (Type I).
- The hypothesis is **wrong**, but you **accept** it (Type II).

		REALITY	
		Null is True	Null is False
STATISTICAL DECISION:	Do Not Reject Null	$1 - \alpha$ Correct	β Type II error
	Reject Null	α Type I error	$1 - \beta$ Correct

Taking decisions

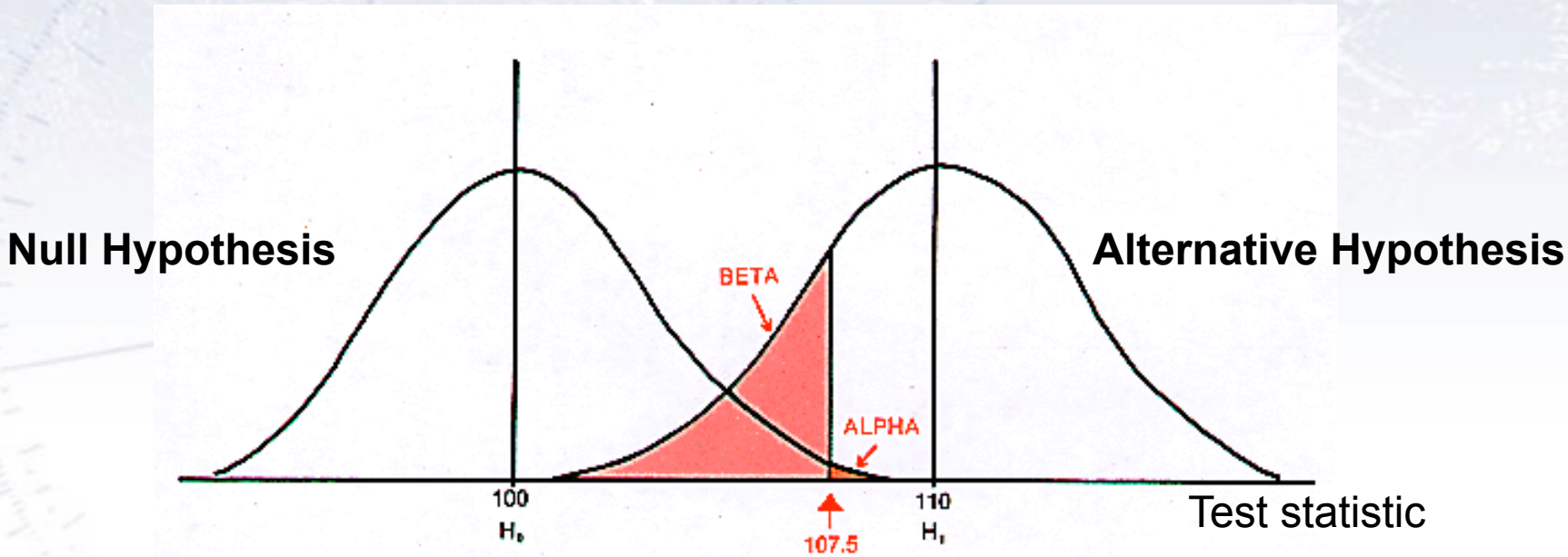
You are asked to take a decision: **Given data - how to do that best?**



		REALITY	
		Null is True	Null is False
STATISTICAL DECISION:	Do Not Reject Null	$1 - \alpha$ Correct	β Type II error
	Reject Null	α Type I error	$1 - \beta$ Correct

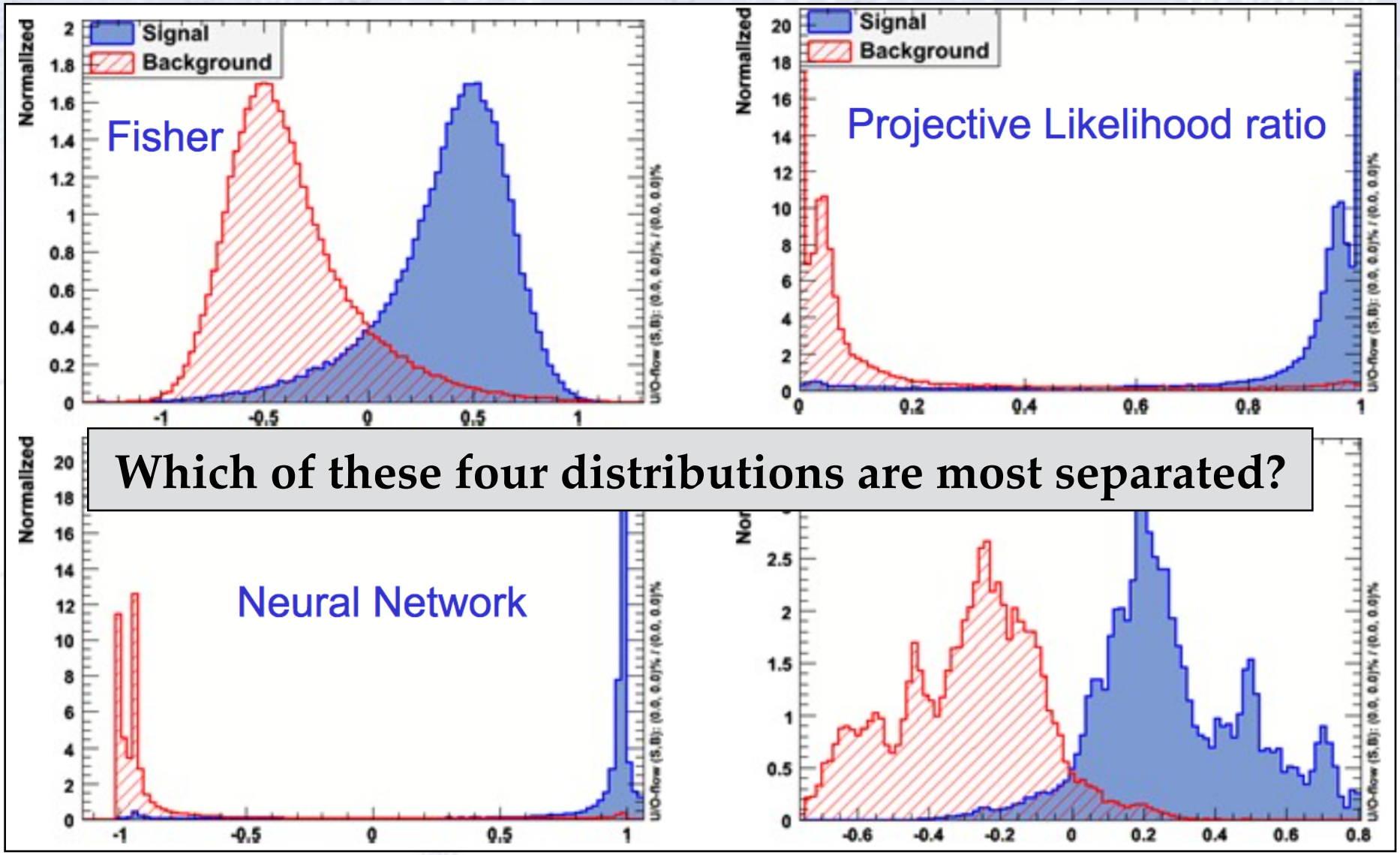
Taking decisions

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



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?

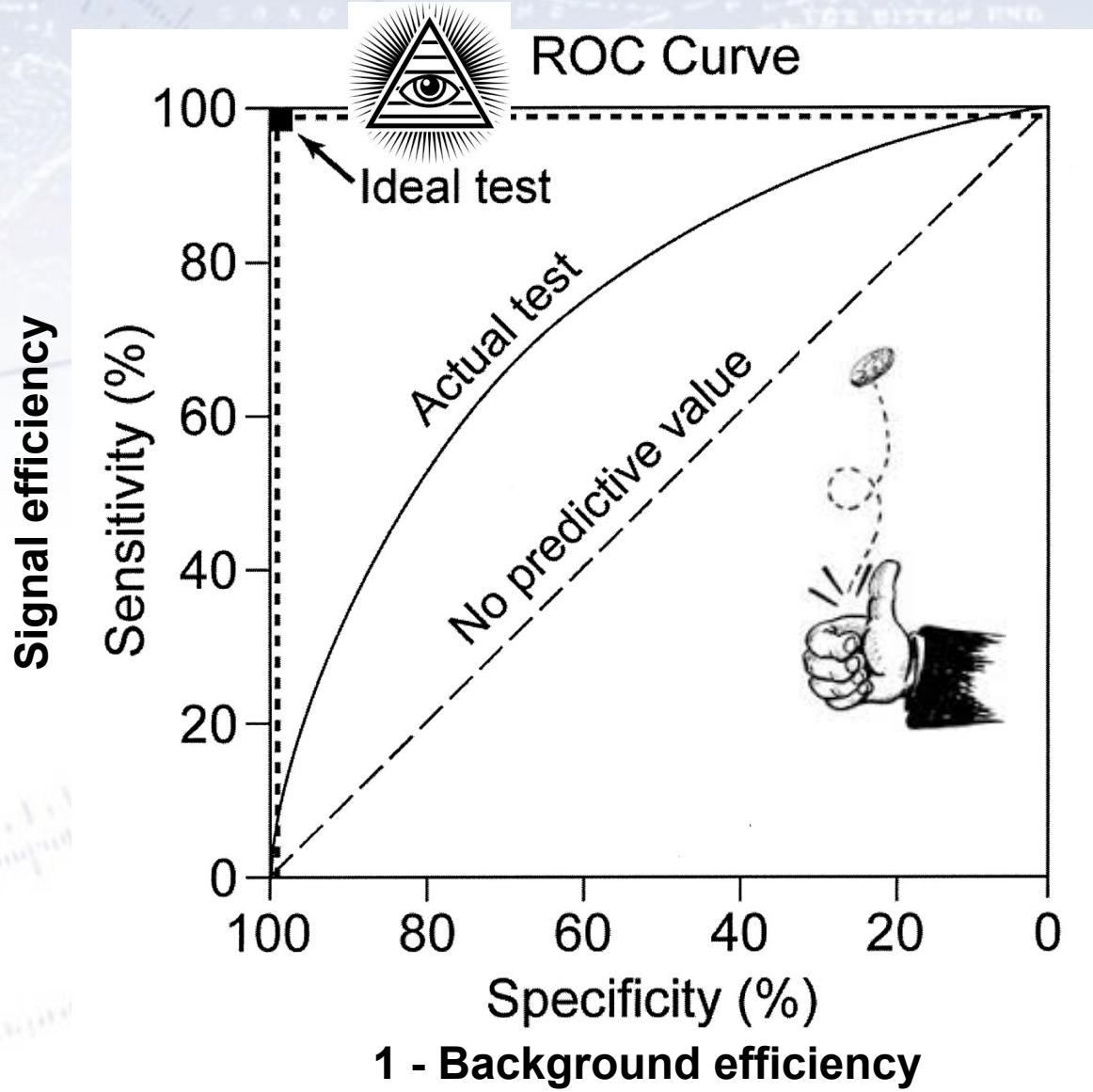
ROC curves

The **Receiver Operating Characteristic** or just ROC-curve is a graphical plot of the sensitivity, or true positive rate, vs. false positive rate.

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

Often, it requires a testing data set to actually see how well a test is performing.

Dividing data, it can also detect overtraining!



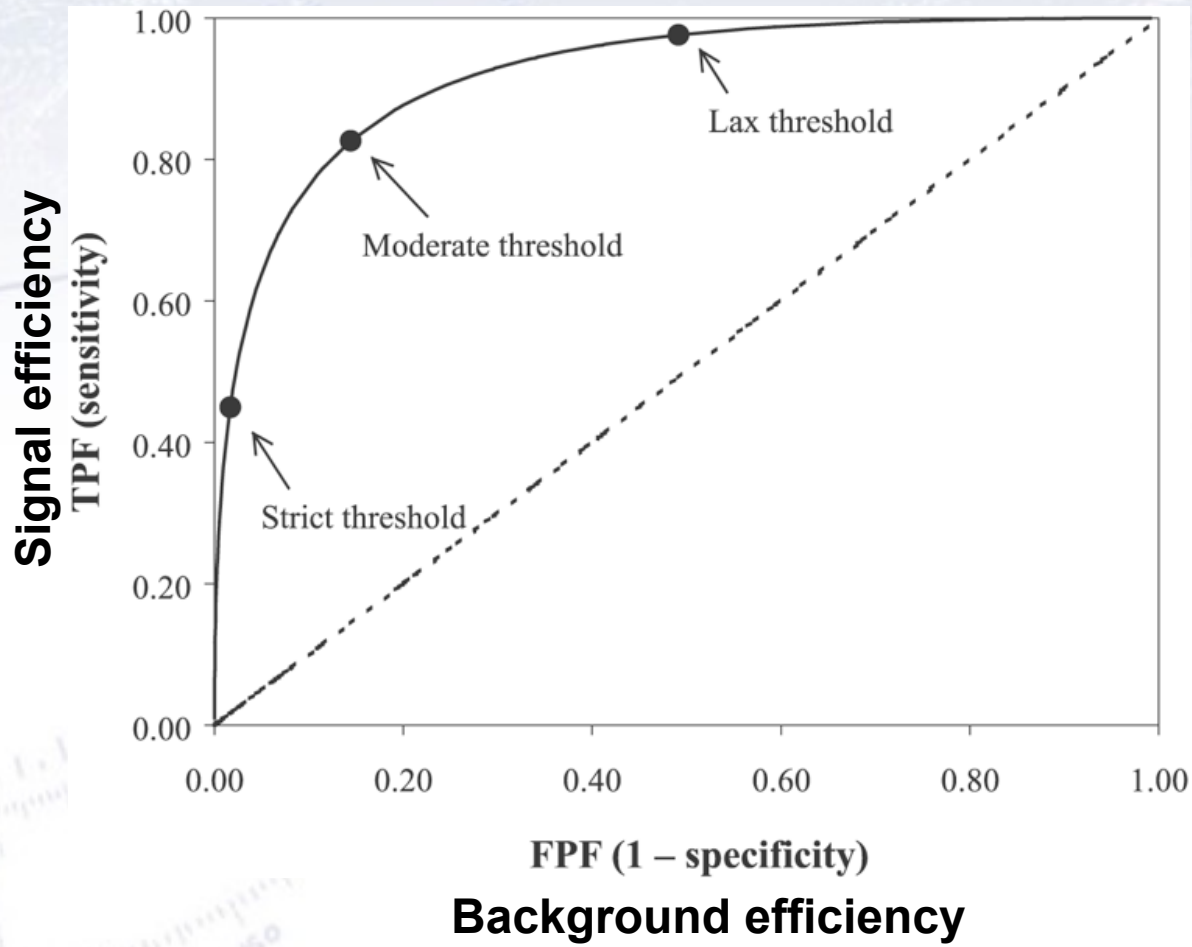
ROC curves

The **Receiver Operating Characteristic** or just ROC-curve is a graphical plot of the sensitivity, or true positive rate, vs. false positive rate.

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

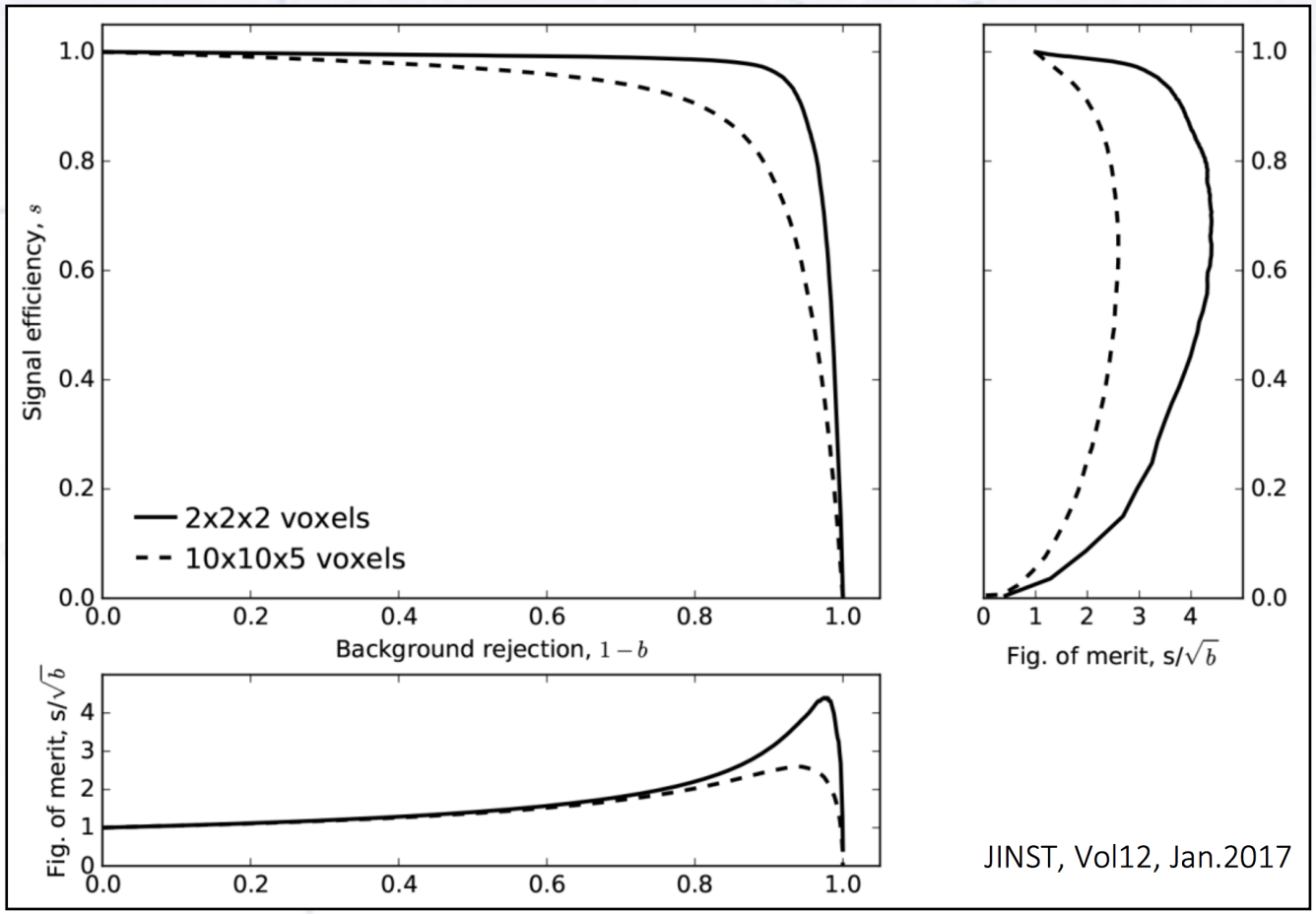
Often, it requires a testing data set to actually see how well a test is performing.

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) 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.

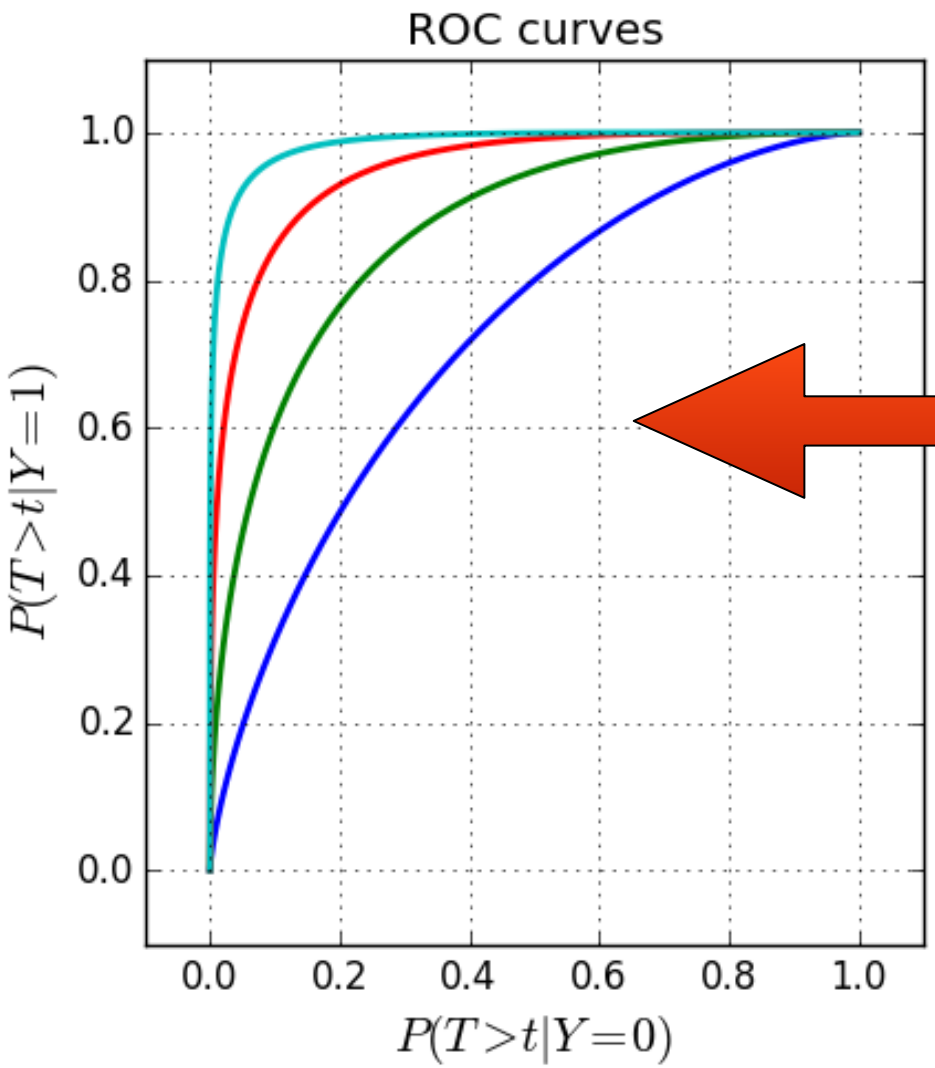
One metric - not mentioned here - is the Area Under the Curve (AUC), which is simply an integral of the ROC curve (thus 1 is perfect score). This is often used in Machine Learning to optimise performance (loss).

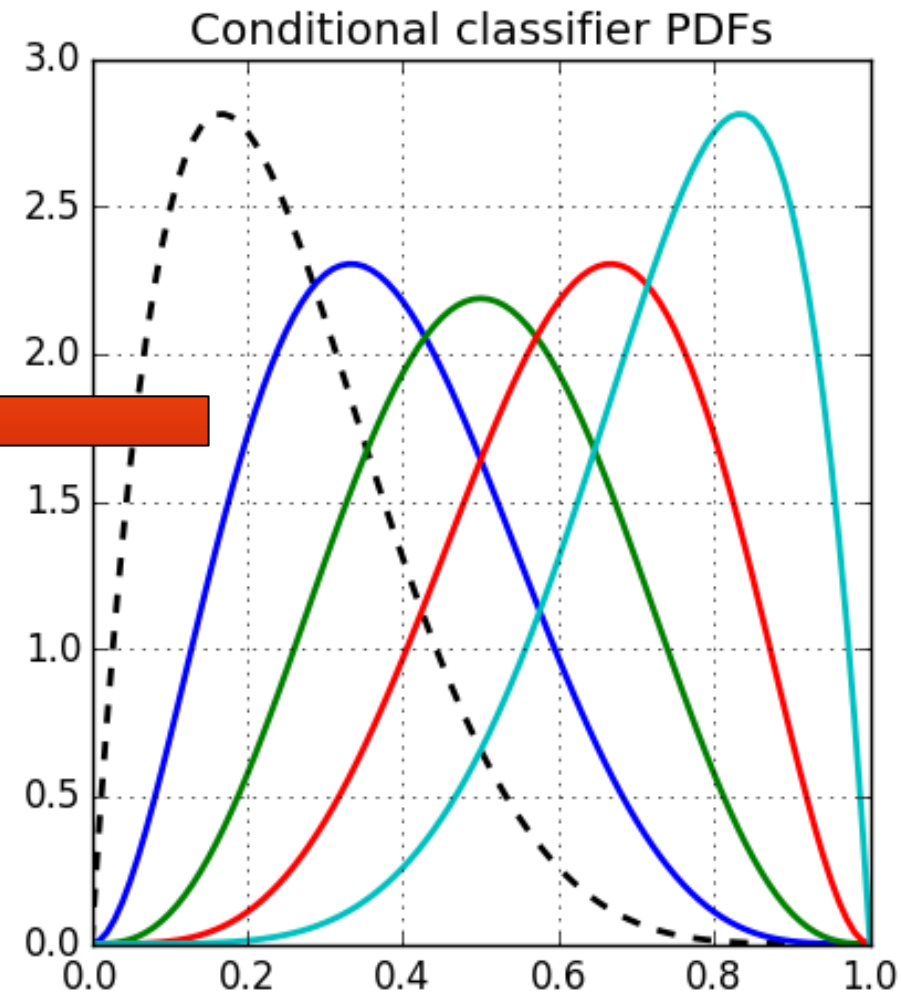
		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	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 = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		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

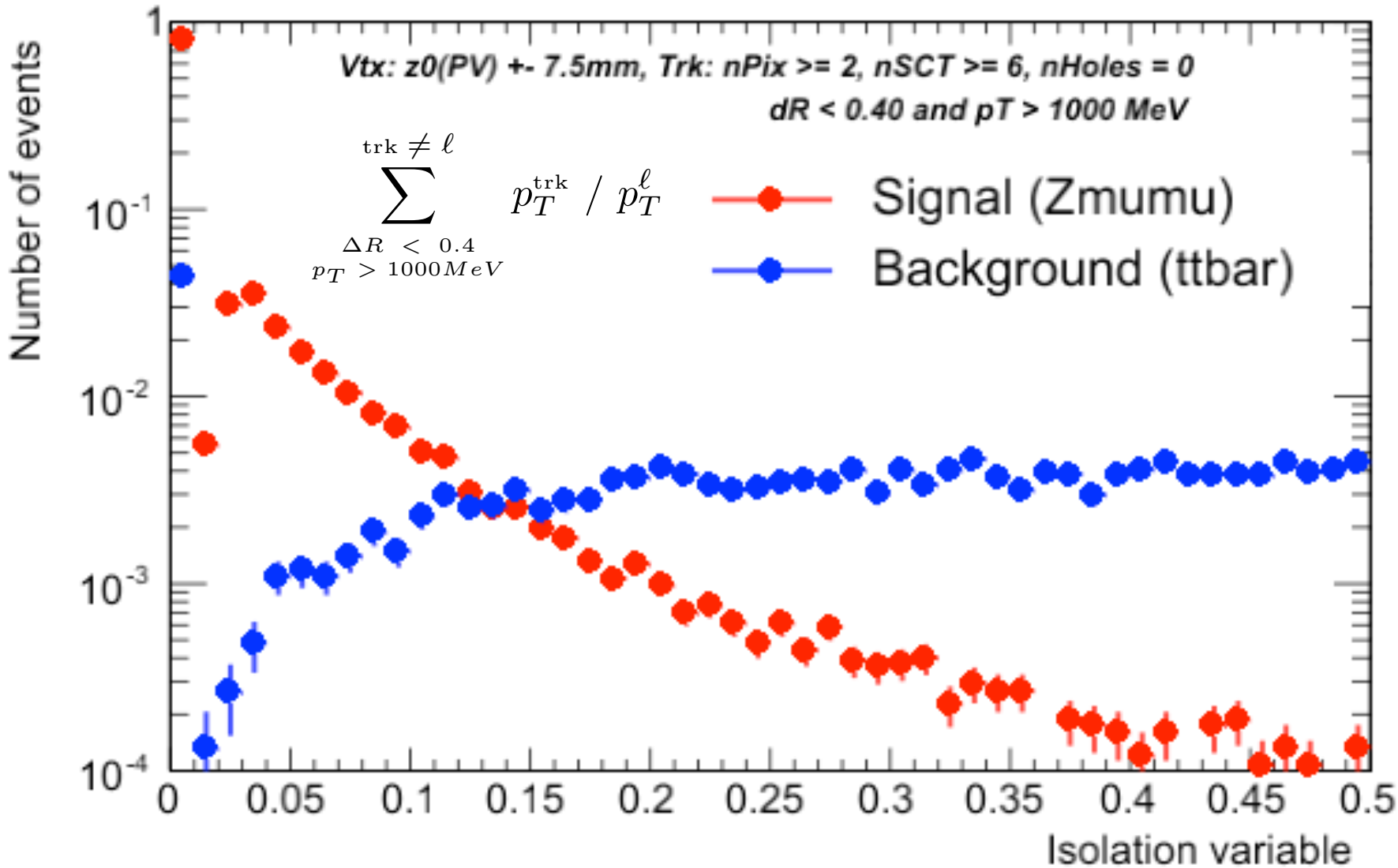
Example of ROC curves in use

Simple case

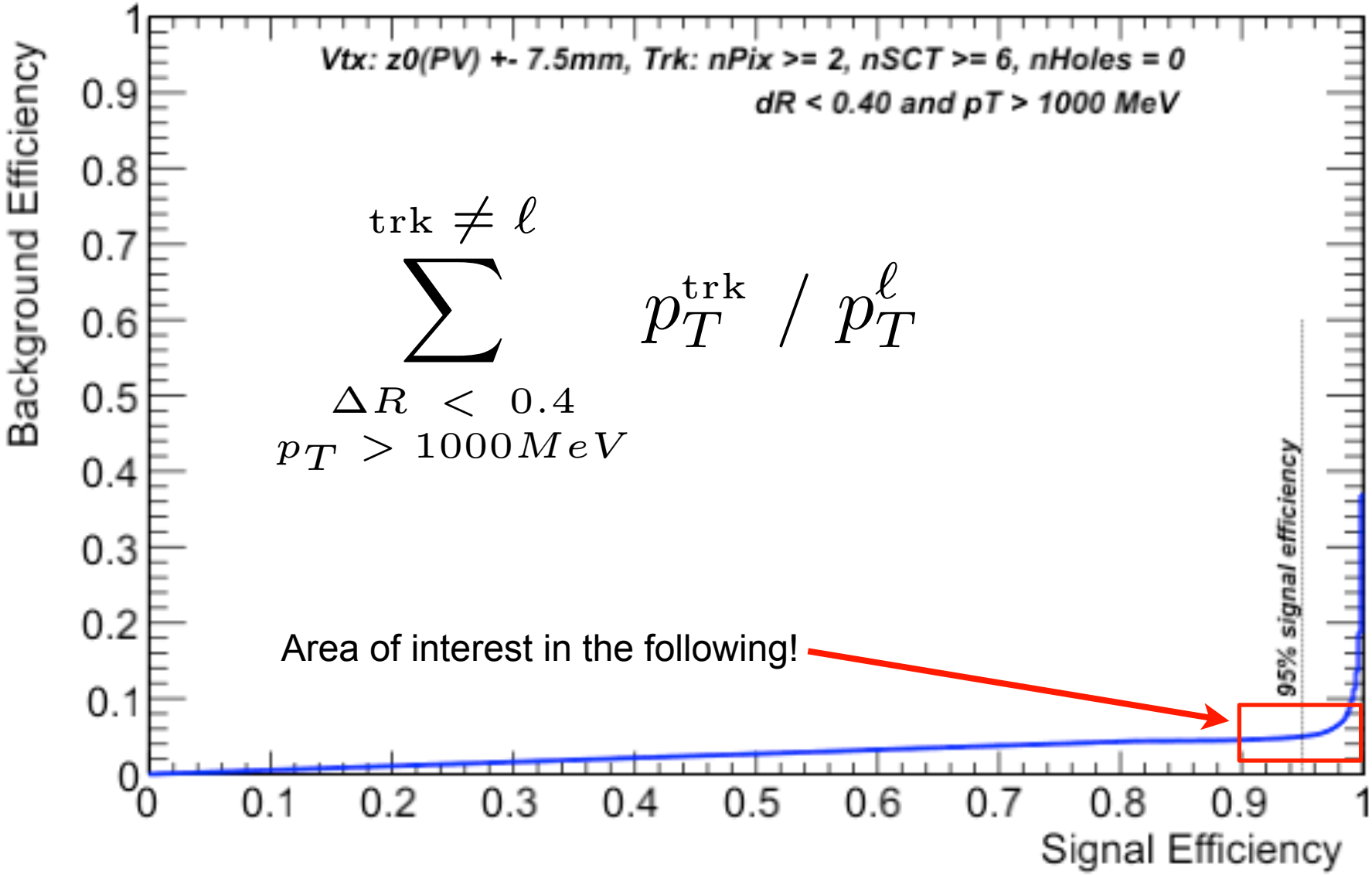

ROC curves
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
$P(T>t|Y=1)$
$P(T>t|Y=0)$


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

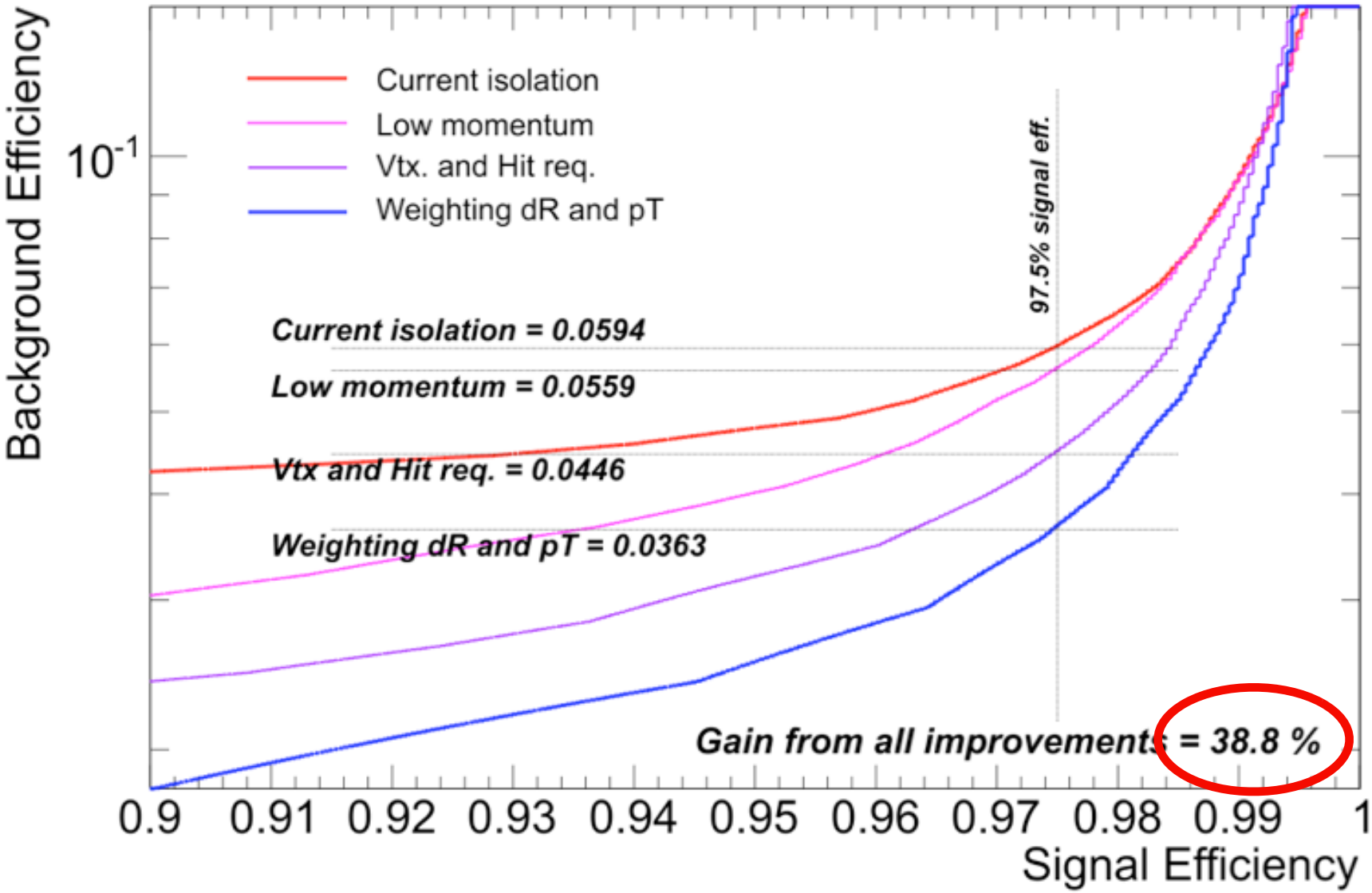
Basic steps - distributions



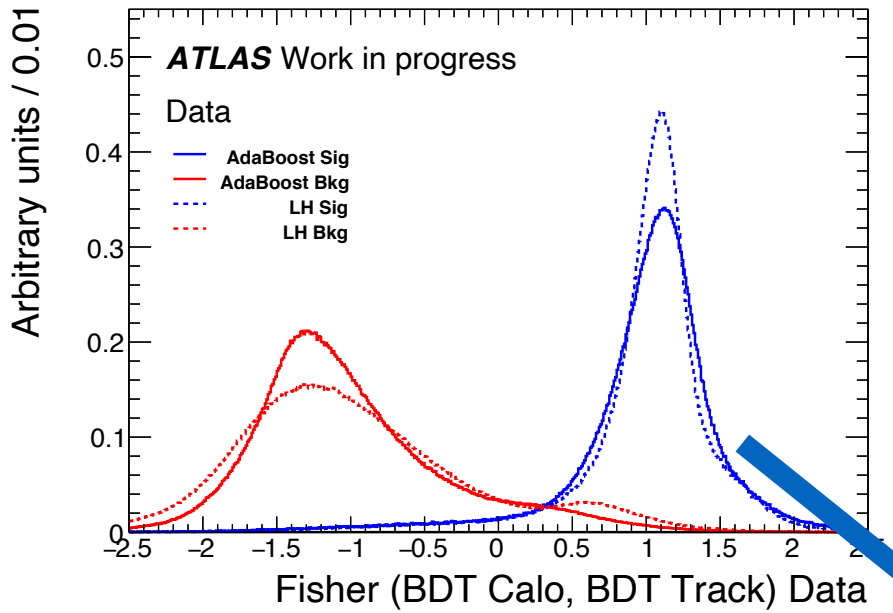
Basic steps - ROC curves



Overall improvement

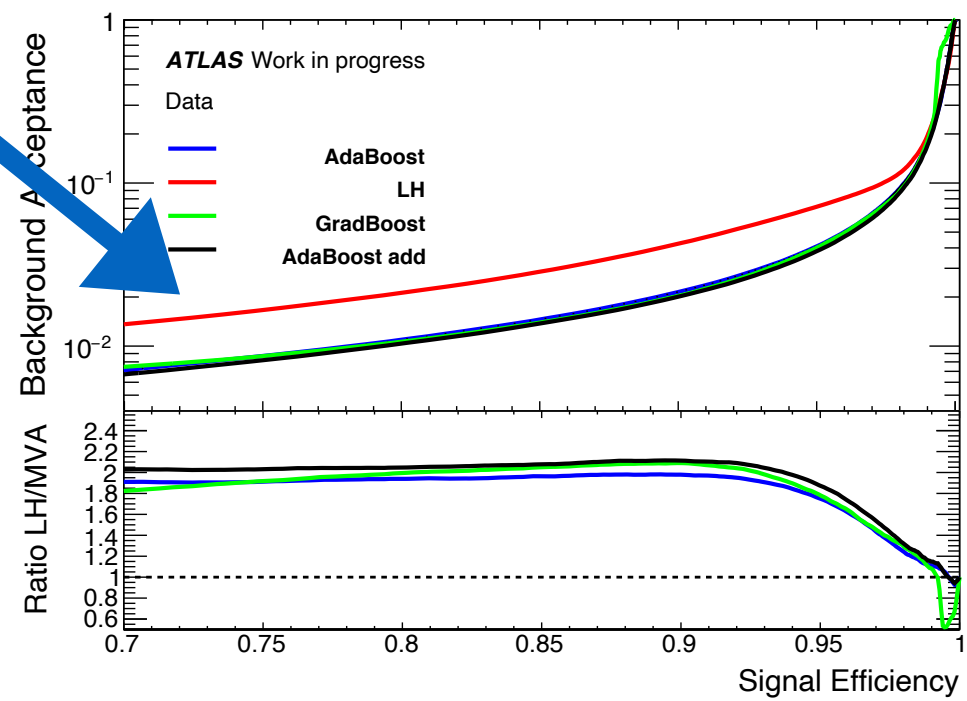


Recent example (electron PID)

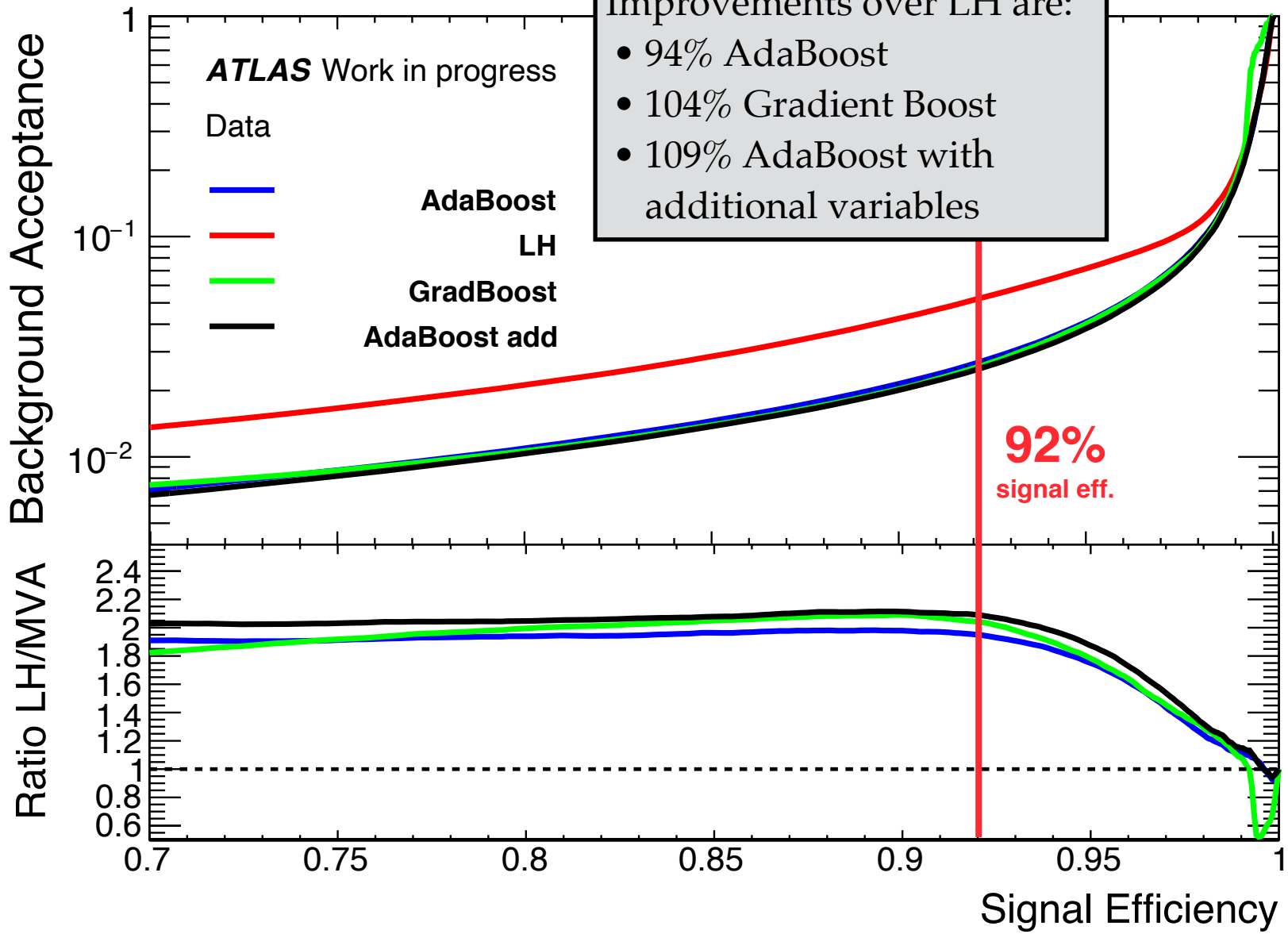


This example is from identifying electrons using Machine Learning in the ATLAS experiment.
It is the result of applying ML on data, which solves the problem of differences between data and simulation.

In addition to the ROC curves, the ratio of these are also shown, to illustrate the improvement as a function of operating point.
The three new methods clearly improve on the existing method.



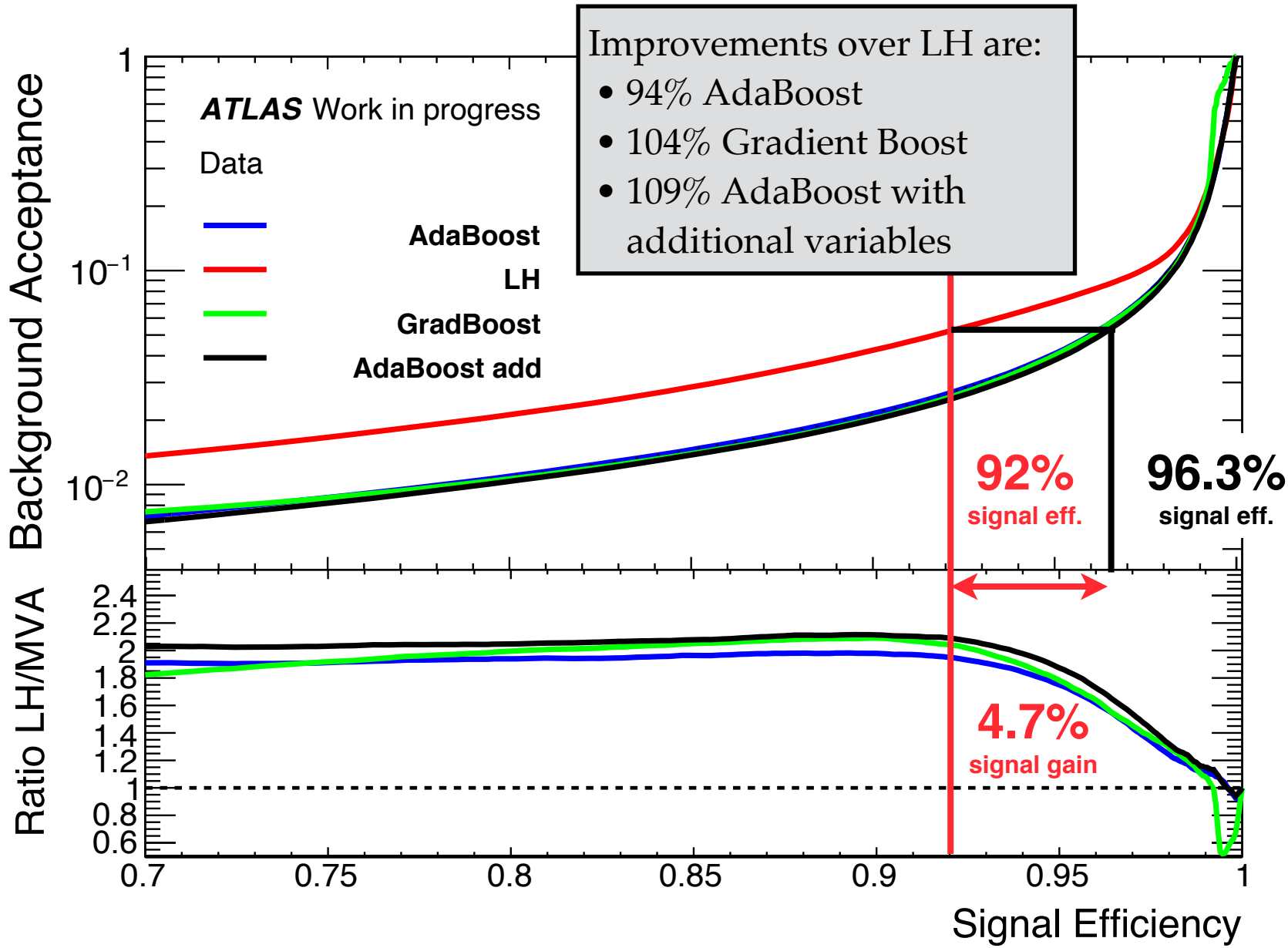
Combined result



Improvements over LH are:

- 94% AdaBoost
- 104% Gradient Boost
- 109% AdaBoost with additional variables

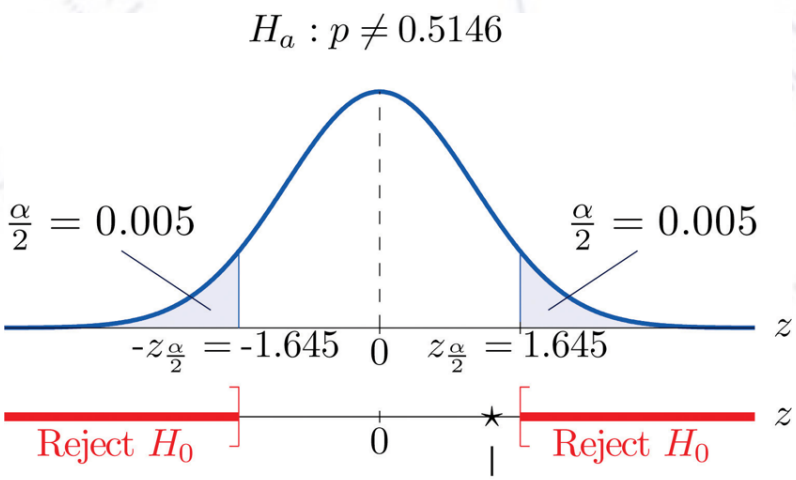
Combined result

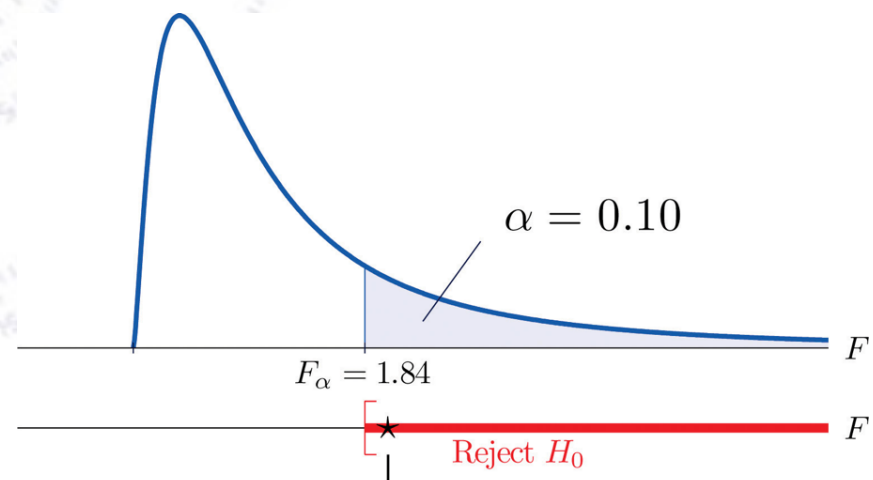


Testing procedure
&
Typical statistical tests

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** 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**.

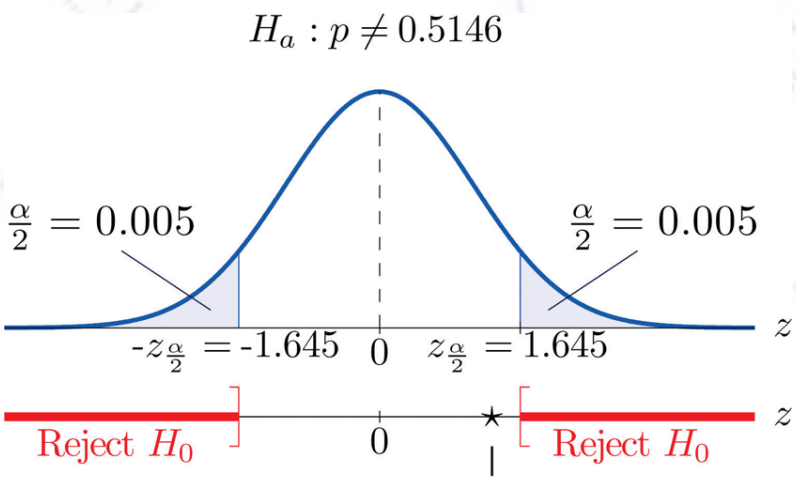


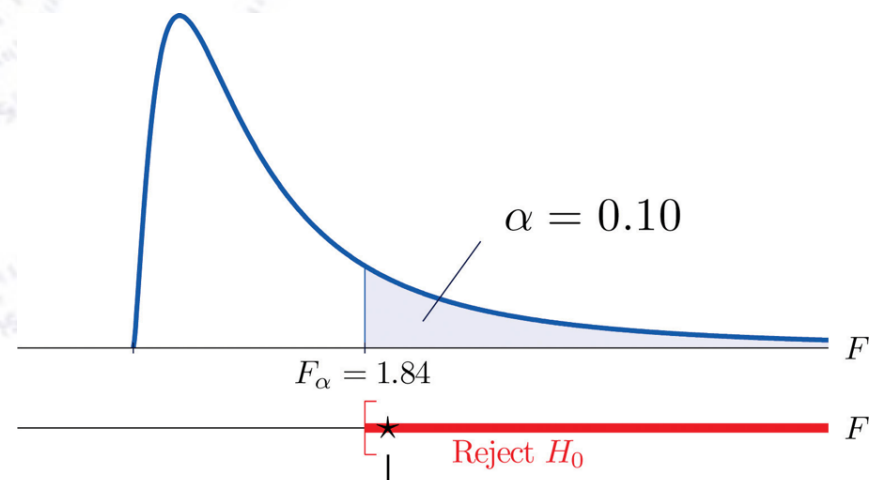


Testing procedure

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 standa… …nt's t, etc.)
6. **Select a s…** … which null
   hypothes… …(physics)).
7. Compute… …**statistic** t.
8. From t **ca…** …(**p-value**).
9. **Reject null hypothesis for alternative if p-value is below significance 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**.

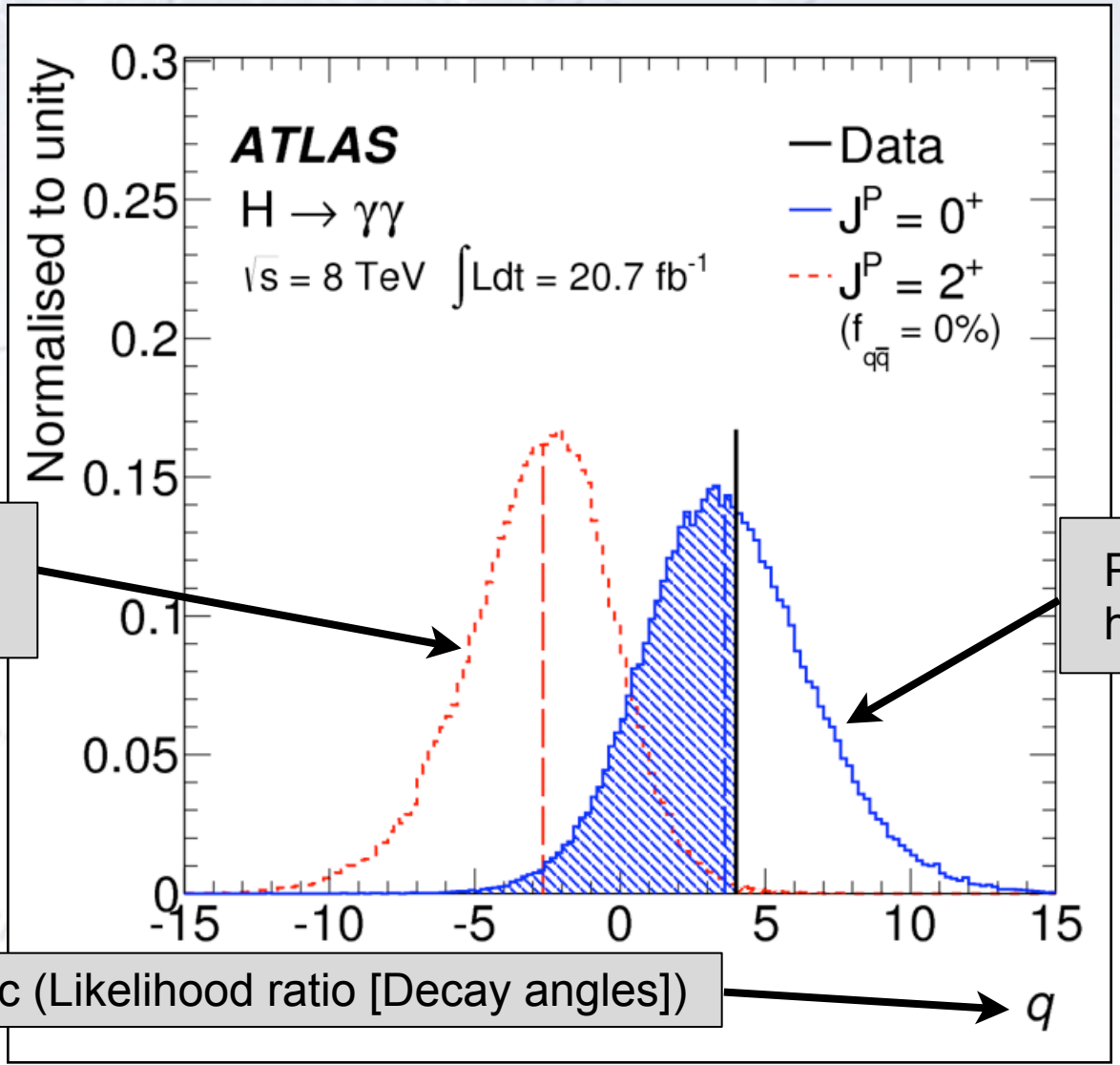
You can **accept** a hypothesis, but this does not exclude accepting other hypothesis.

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.

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])

Neyman-Pearson Lemma

Consider a **likelihood ratio** between the null and the alternative model:

$$D = -2 \ln \frac{\text{likelihood for null model}}{\text{likelihood for alternative model}}$$

The Neyman-Pearson lemma (loosely) states, that this is the most powerful test there is.

In reality, the problem is that it is not always easy to write up a likelihood for complex situations!

However, there are many tests derived from the likelihood...

Likelihood ratio problem

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

$$D = -2 \ln \frac{\text{likelihood for null model}}{\text{likelihood for alternative model}}$$

…it turns out that determining the exact 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 statistical tests

- **One-sample test** compares sample (e.g. mean) to known value: $$z = \frac{\bar{x} - \mu_0}{\sigma(\bar{x})}$$
  Example: Comparing sample to known constant ($\mu_{exp} = 2.91 \pm 0.01$ vs. $c = 2.99$).
- **Two-sample test** compares two samples (e.g. means). $$z = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\sigma(\bar{x_1})^2 + \sigma(\bar{x_2})^2}}$$
  Example: Comparing sample to control ($\mu_{exp} = 4.1 \pm 0.6$ 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
- **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!

The figure of merit is typically the **Power of a Test***, defined as $(1 - \beta)$, 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).

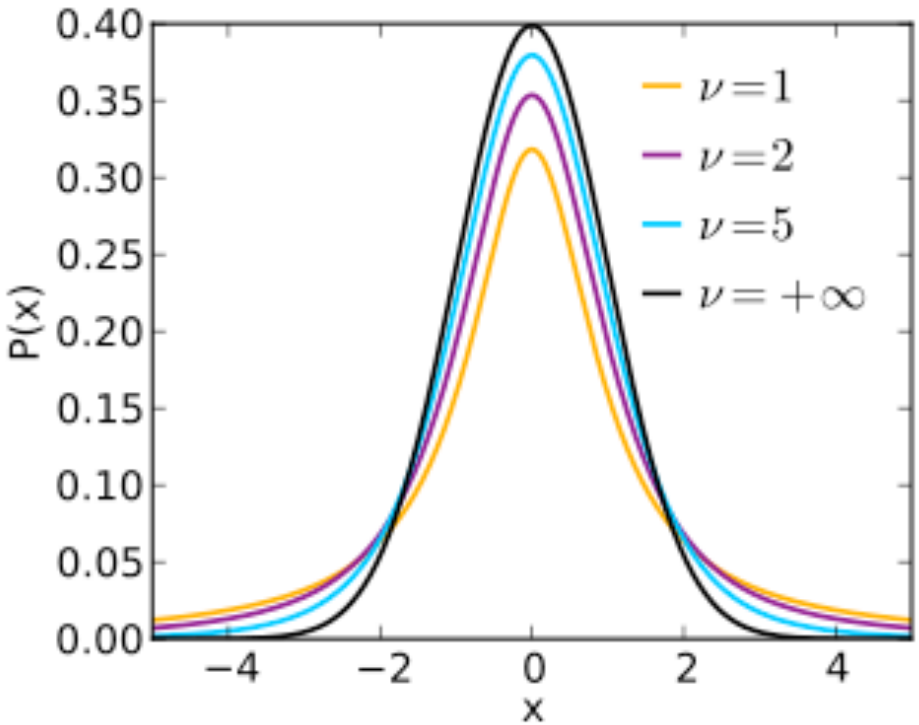
In medical science, it is typically important to have a powerful test (i.e. low β), while in criminal science it is a low type I error rate (i.e. low α), convicting innocents.
In the end, choosing a test comes down to **experience, importance of power, ease of use**, and even standards in the field of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

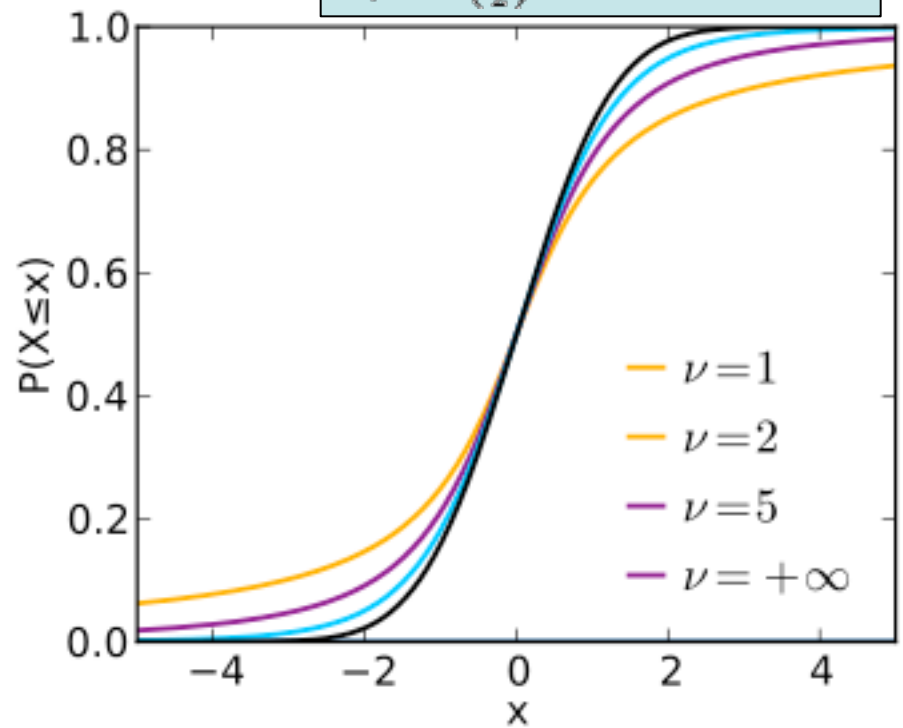
* Power of a test is often termed sensitivity in biostatistics.

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:

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

  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).

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

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

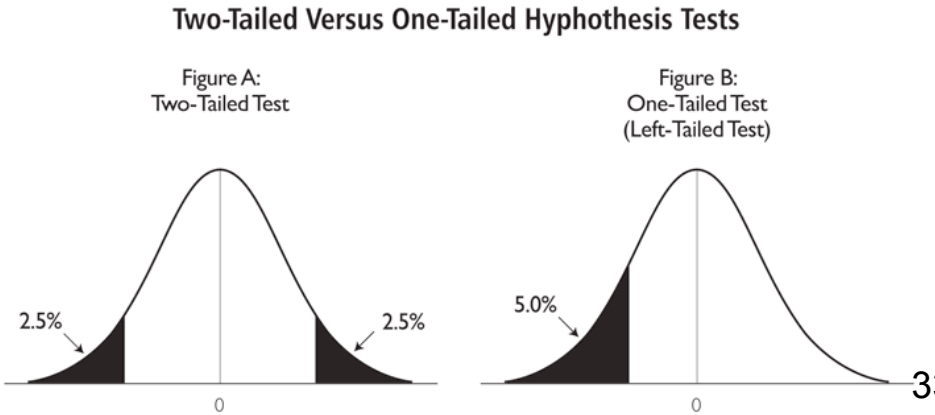
<u>Things to consider:</u>

- 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...

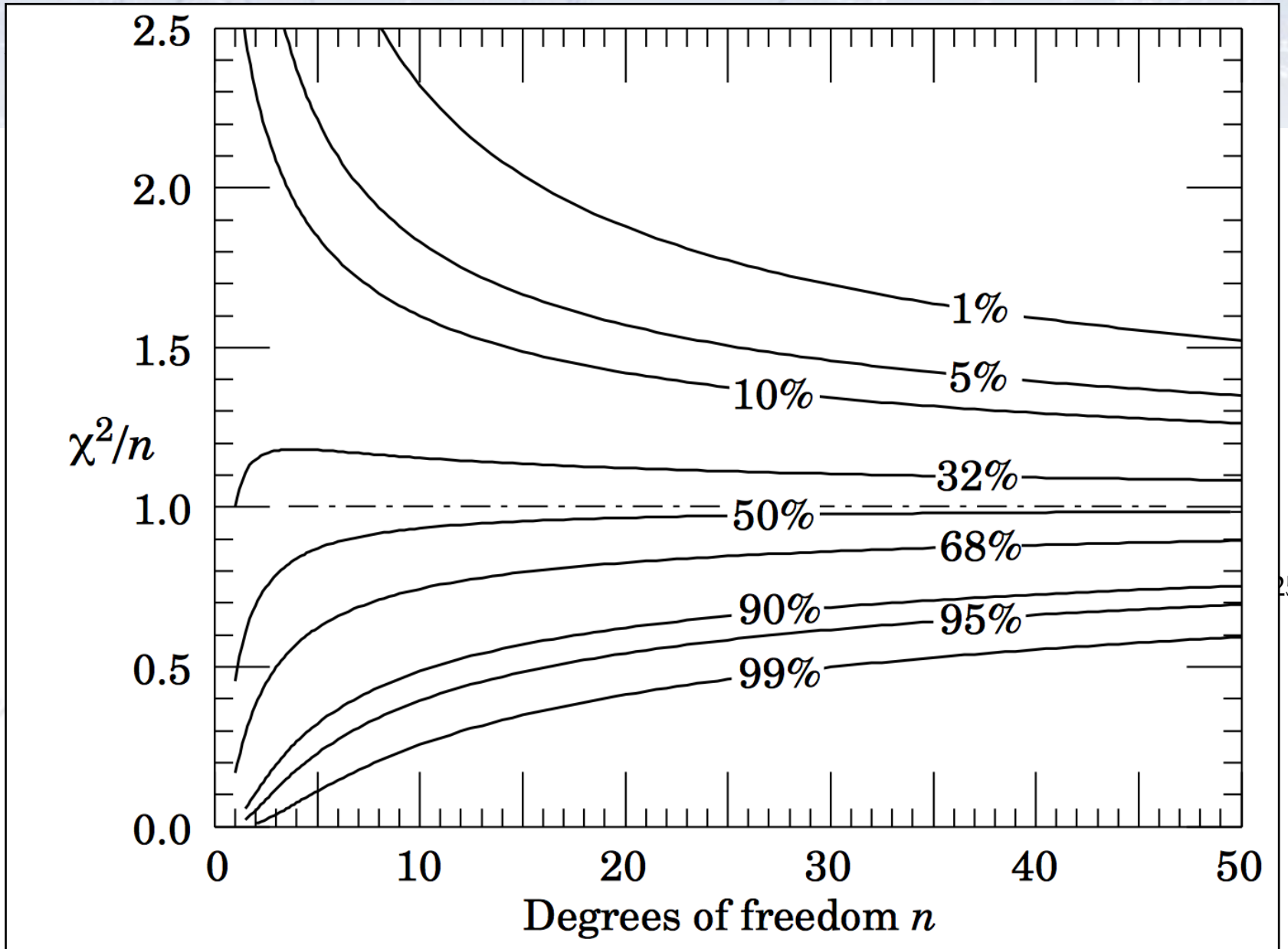
$$\chi^2(\bar{\theta}) = \sum_{i=1}^{N} \frac{(y_i - \lambda(x_i; \bar{\theta}))^2}{\sigma_i^2}$$

- **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...

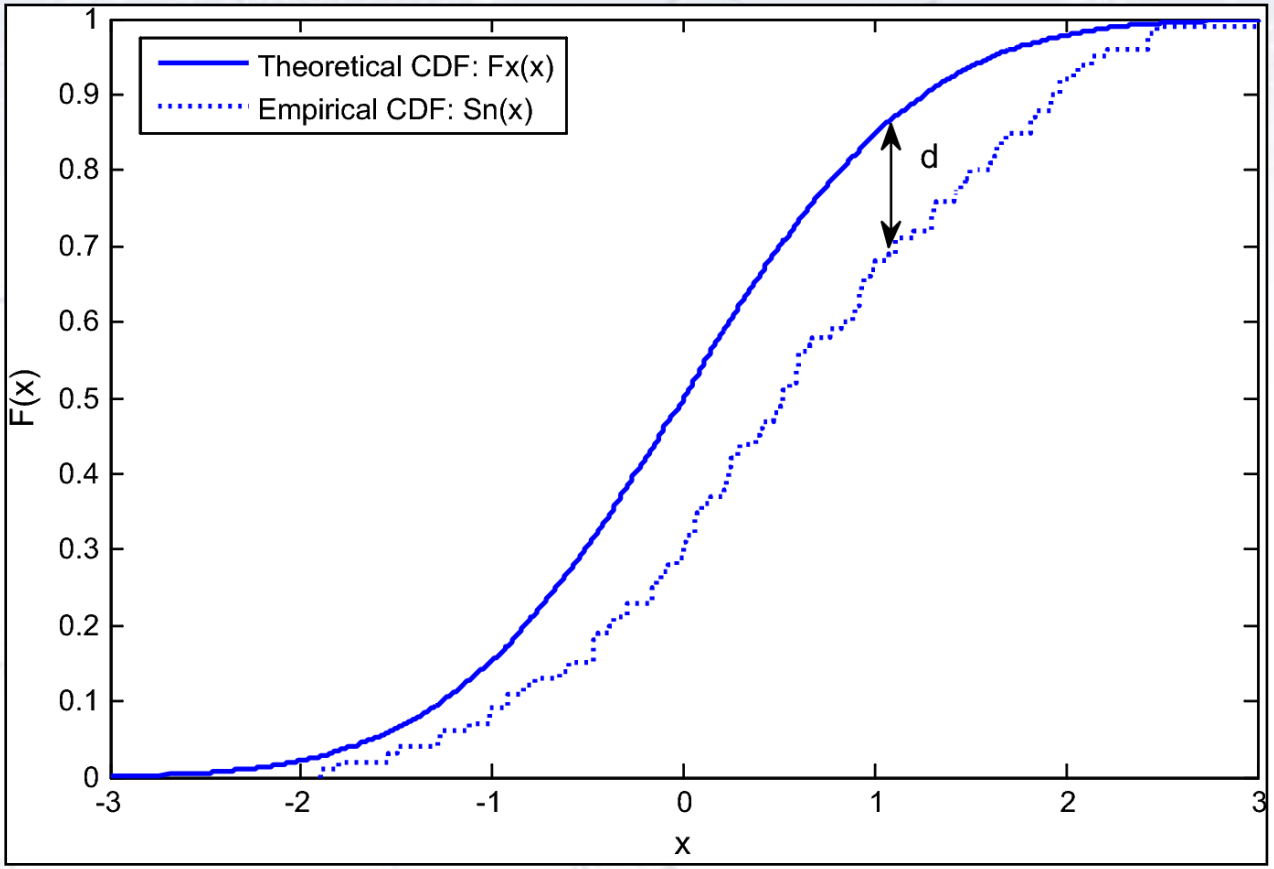
Chi-squared test



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

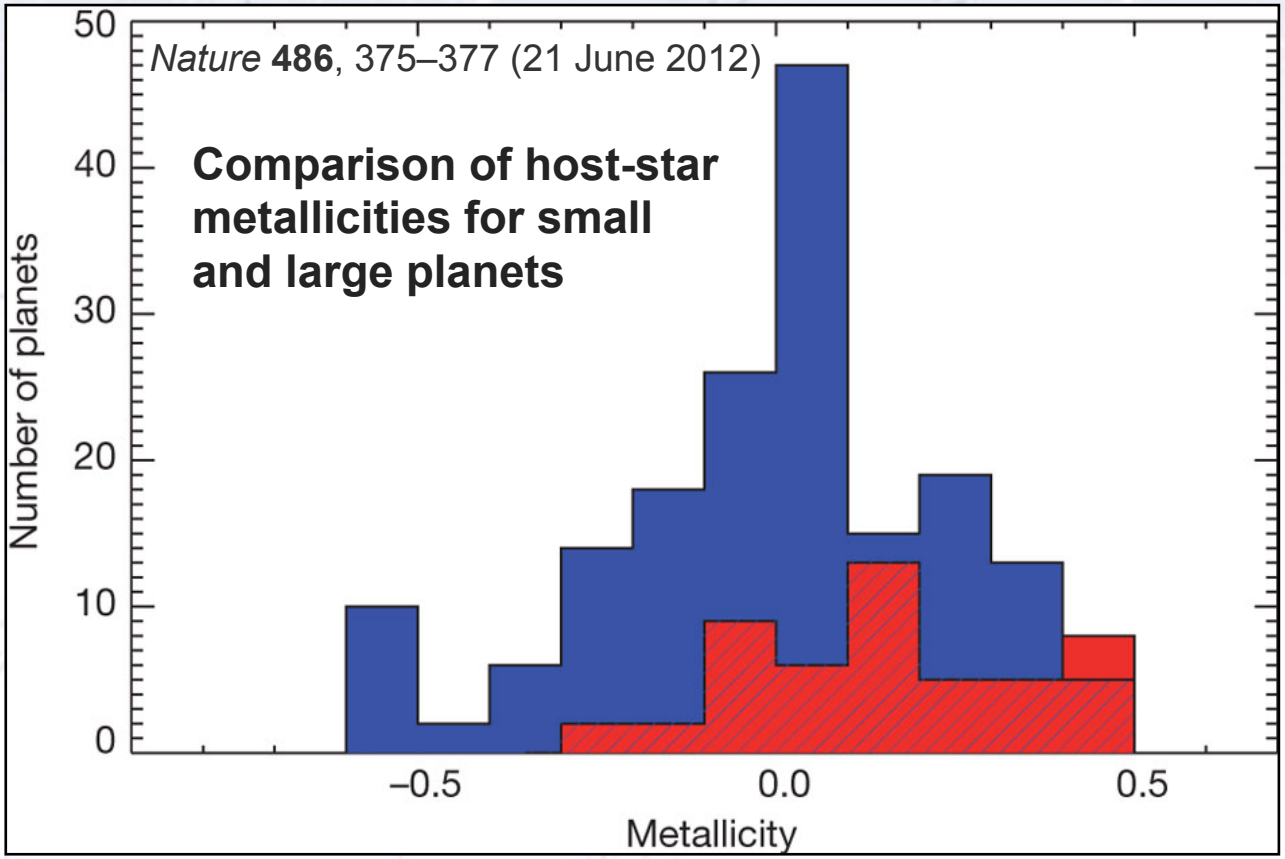


The Kolmogorov test measures the maximal distance between the integrals of two distributions and gives a probability of being from the same distribution.

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

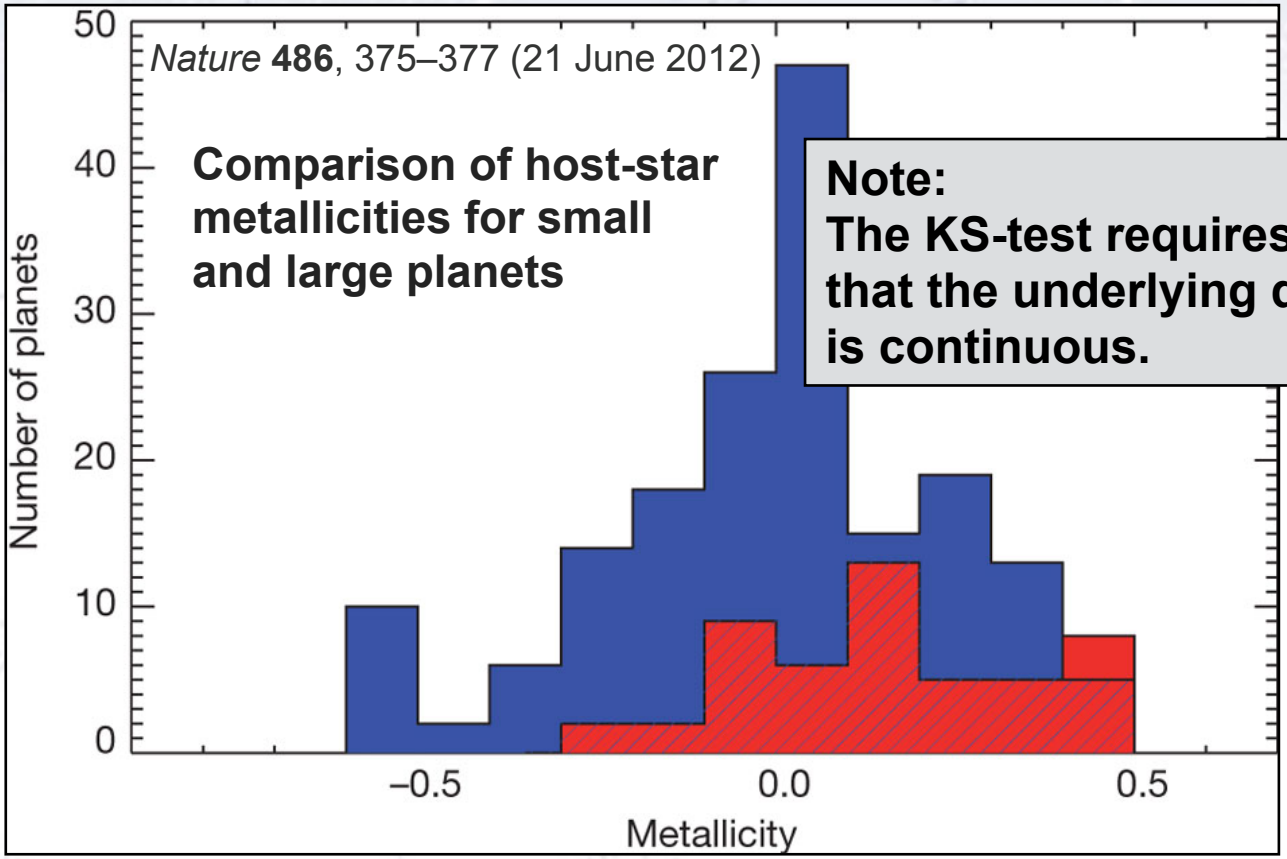


"A Kolmogorov–Smirnov test shows that the probability that the two distributions are not drawn randomly from the same parent population is greater than 99.96%; that is, the two distributions differ by more than 3.5σ". [Quote from figure caption]

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

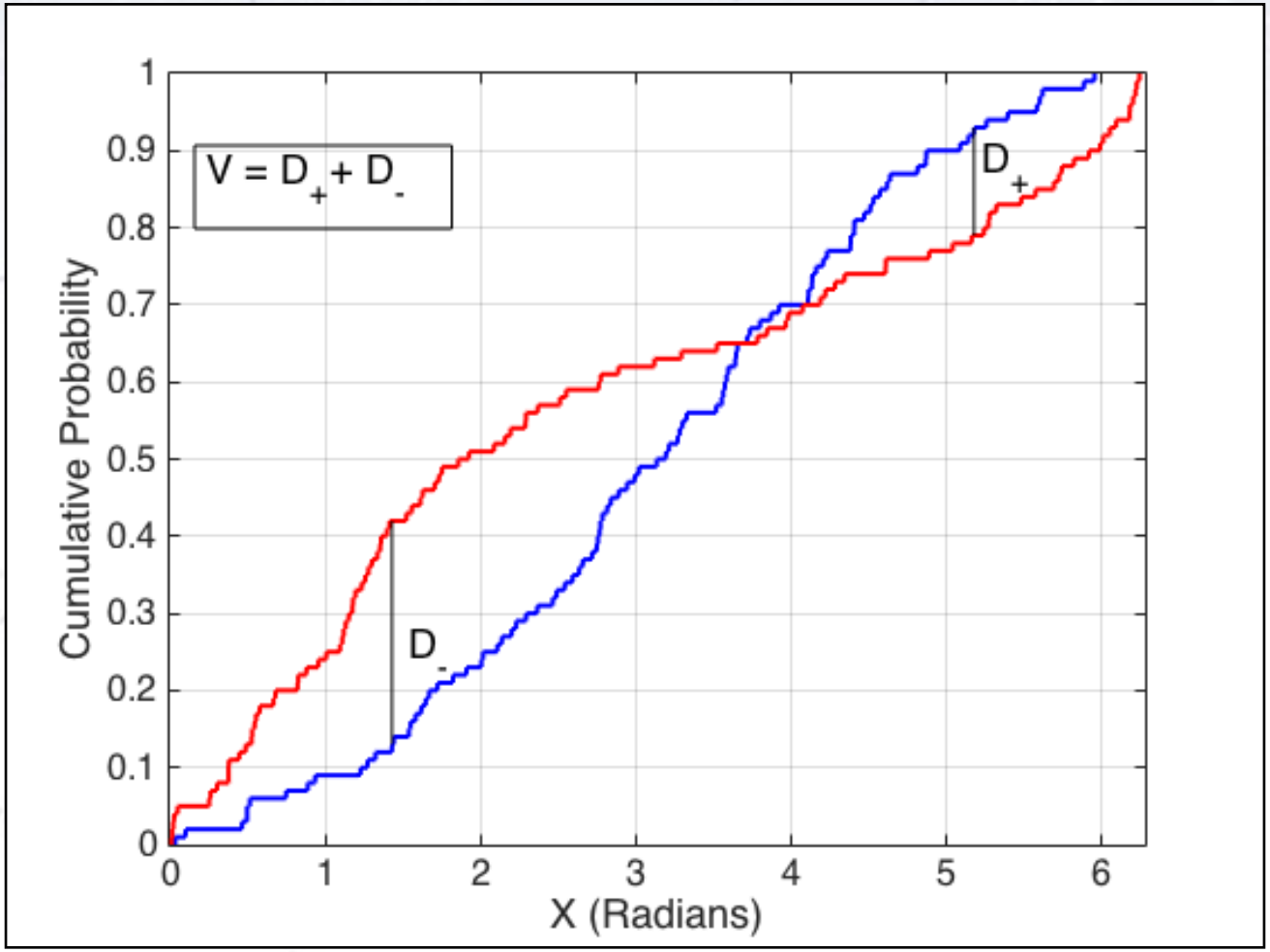


Note:
The KS-test requires/assumes, that the underlying distribution is continuous.

"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: $t = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n}$

  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). $t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

  Example: Comparing sample to control ($\mu_{exp} = 4.1 \pm 0.6$ vs. $\mu_{cntrl} = 3.7 \pm 0.4$).

- **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.

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:

++++−−−+++−−++++++−−−−

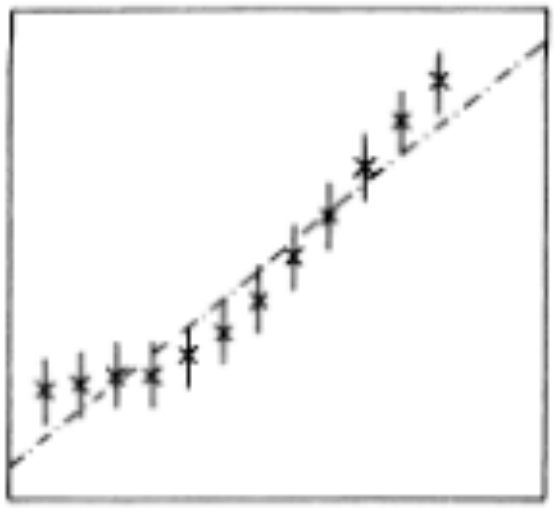


Fig. 8.3. A straight line through twelve data points.

If random, the mean and variance is known:

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

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

N = 12, N_+ = 6, N_- = 6
μ = 7, σ = 1.76
(7-3)/1.65 = 2.4 σ (~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

$$p = \frac{\binom{A+C}{A}\binom{B+D}{B}}{\binom{N}{A+B}} = \frac{(A+B)!\,(C+D)!\,(A+C)!\,(B+D)!}{A!\,B!\,C!\,D!\,N!}$$

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?

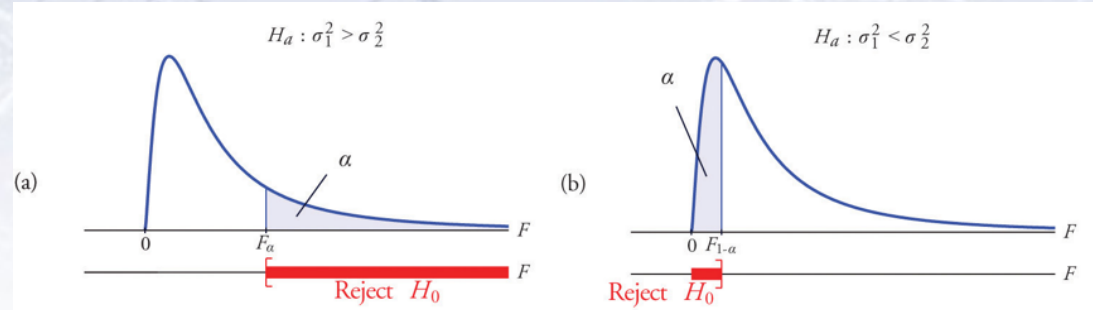
The Chi-square test is not optimal, as there are (several) entries, that are very low (< 5), but Fisher's exact test gives the answer:

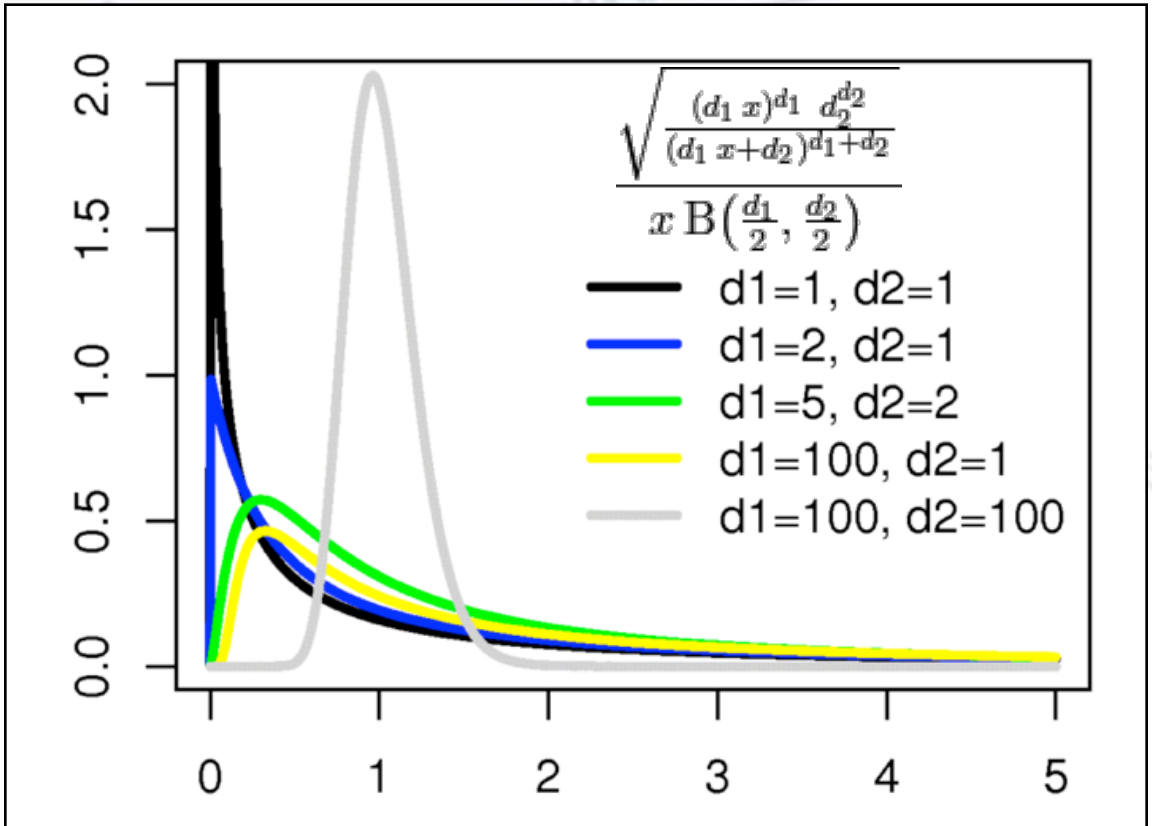
$$p = \binom{10}{1}\binom{14}{11} / \binom{24}{12} = \frac{10!\ 14!\ 12!\ 12!}{1!\ 9!\ 11!\ 3!\ 24!} \simeq 0.00135$$

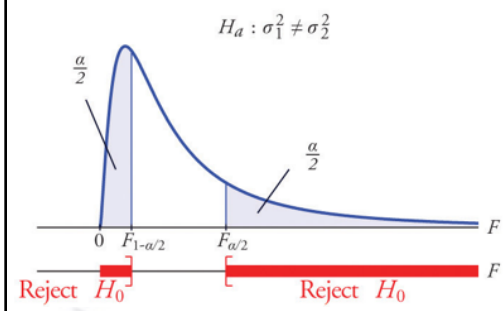
F-test

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.

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
4th 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

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
4th 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!