

Applied Statistics

Hypothesis Testing



Troels C. Petersen (NBI)



"Statistics is merely a quantisation of common sense"

Hypothesis terminology

$H_0 = \text{Null Hypothesis:}$

Definition: The initial / simplest hypothesis.

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

$H_1 = \text{Alternative Hypothesis:}$

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

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

$\alpha = \text{Significance level:}$

Definition: Probability to **accept H_0** , even if it is **false**.

Example: Concluding no signal, even if there; deciding electron when pion.

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

$\beta = \text{Significance level:}$

Definition: Probability to **reject H_0** , even if it is **true**.

Example: Concluding signal, even if not there; deciding pion when electron.

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?

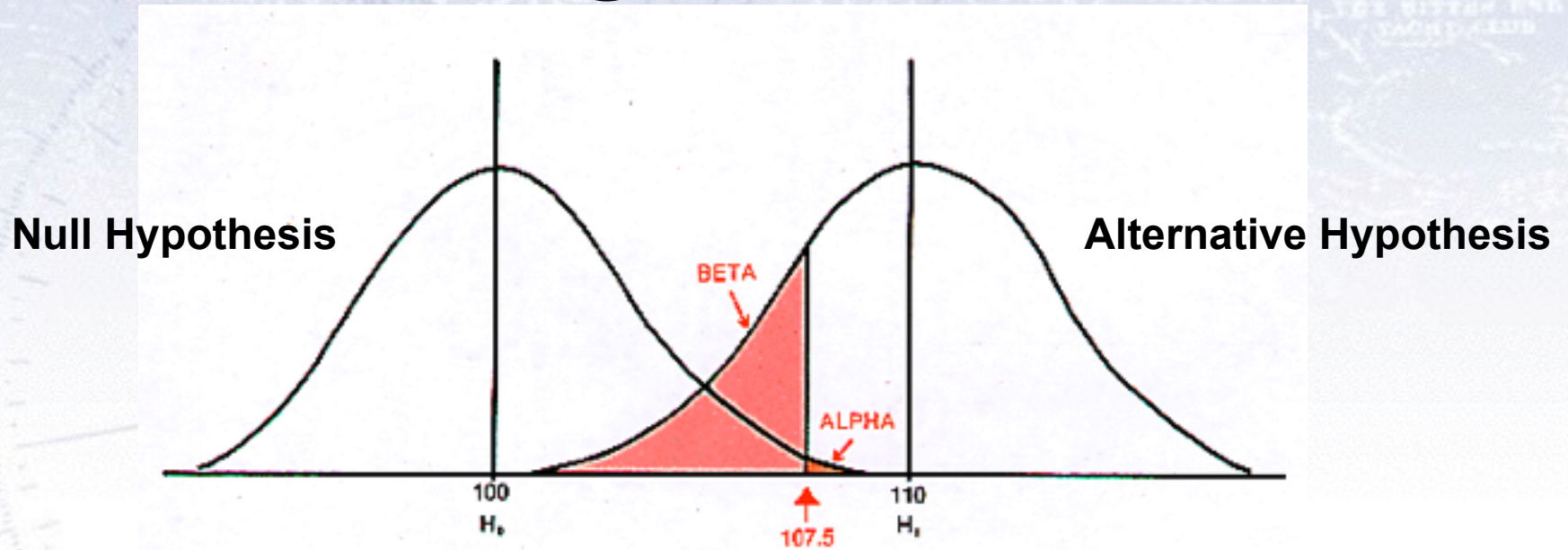
That is the basic question in hypothesis testing.

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).

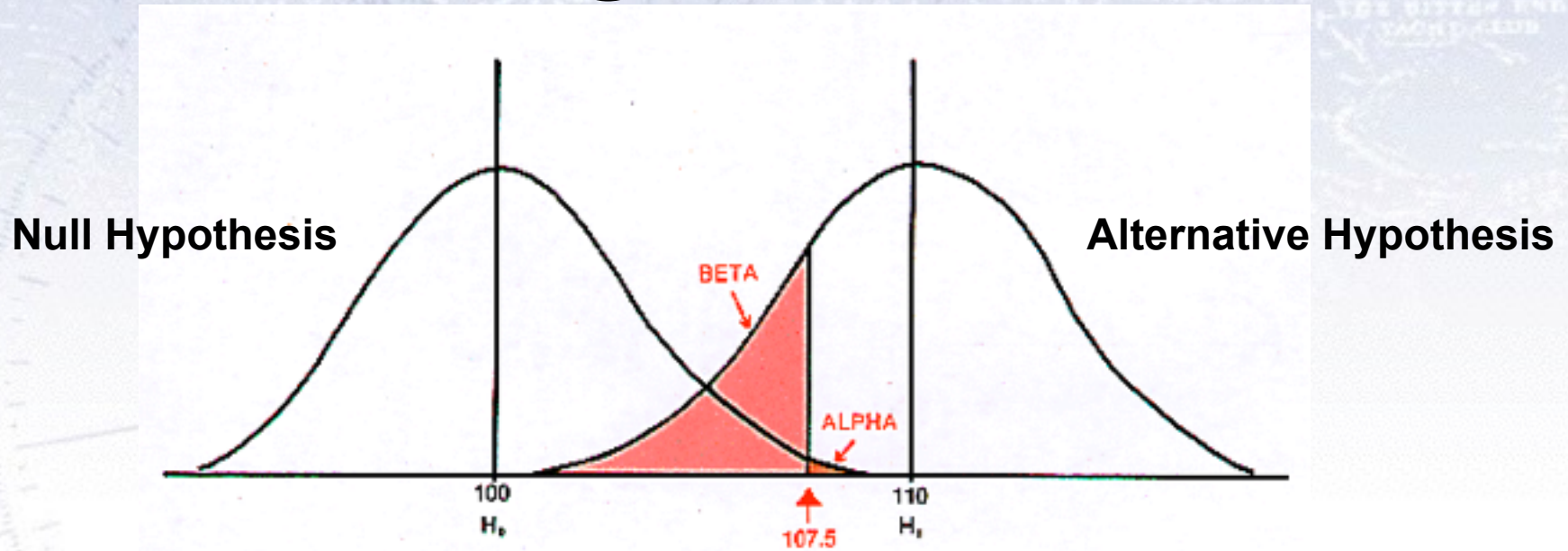
| | | REALITY | |
|--------------------------|--------------------|--------------------------|--------------------------|
| | | Null is True | Null is False |
| STATISTICAL DECISION: | Do Not Reject Null | $1 - \alpha$ Correct | β Type II error |
| | Reject Null | α Type I error | $1 - \beta$ Correct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

Taking decisions



| | | 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 |

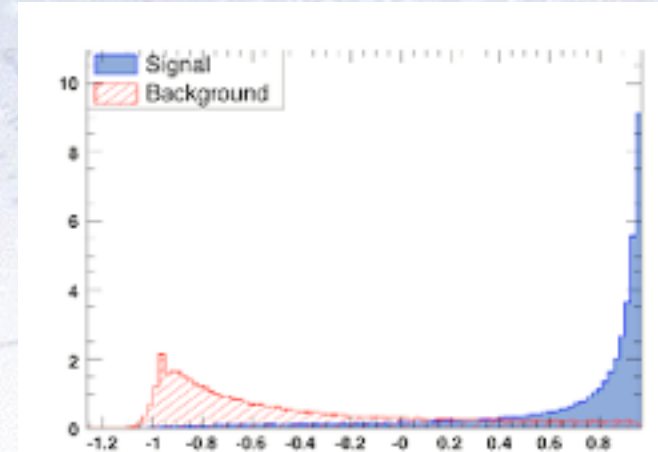
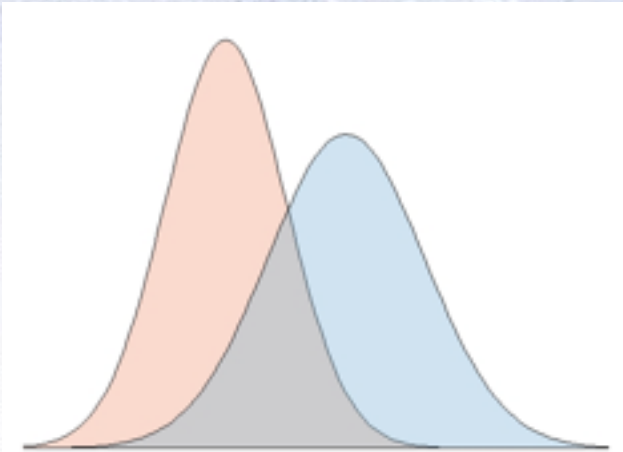
Taking decisions



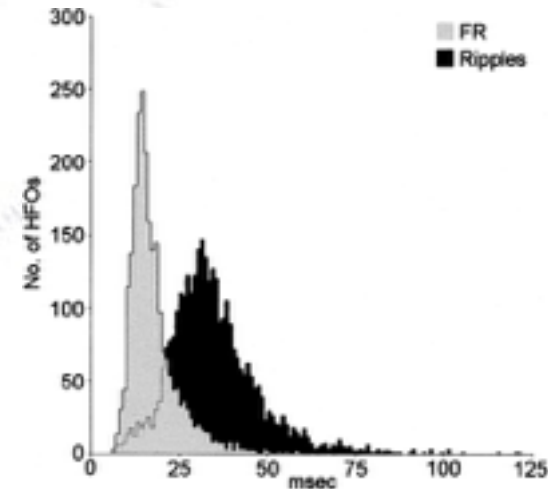
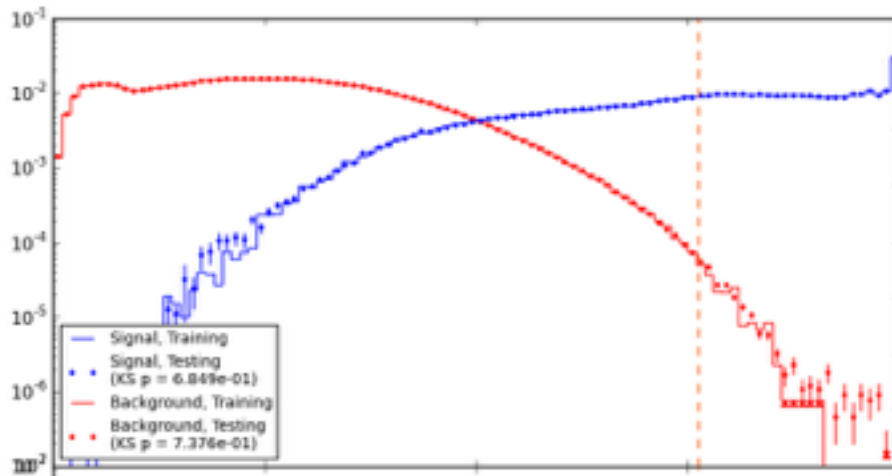
The purpose of a **test** is to yield (calculable/predictable) distributions for the **Null** and **Alternative** hypotheses, which are *as separated from each other as possible*, to minimise α and β .

The likelihood ratio test is in general the best such test.

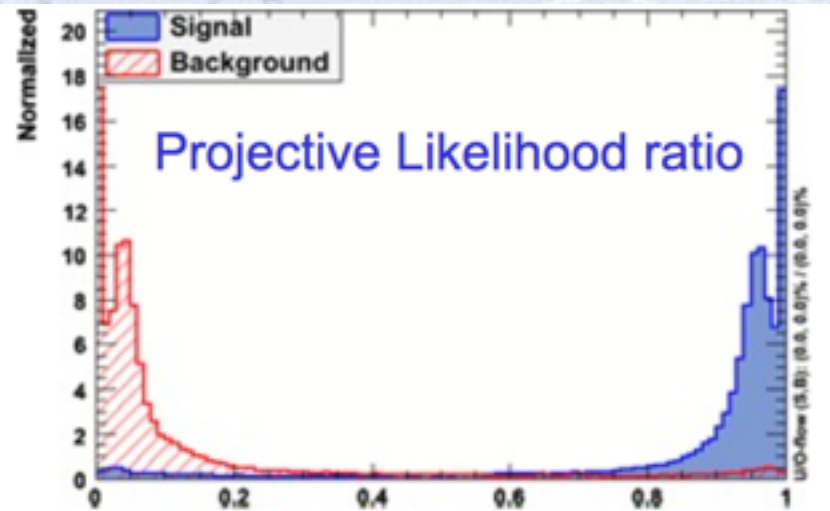
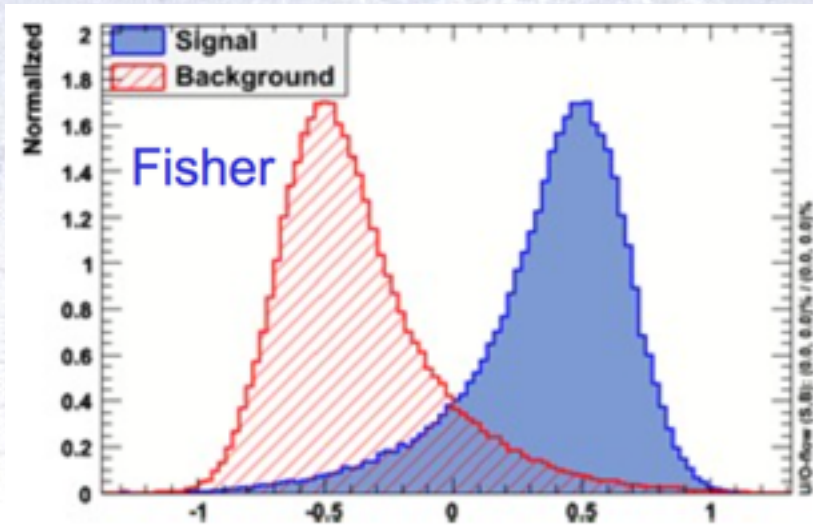
Measuring separation



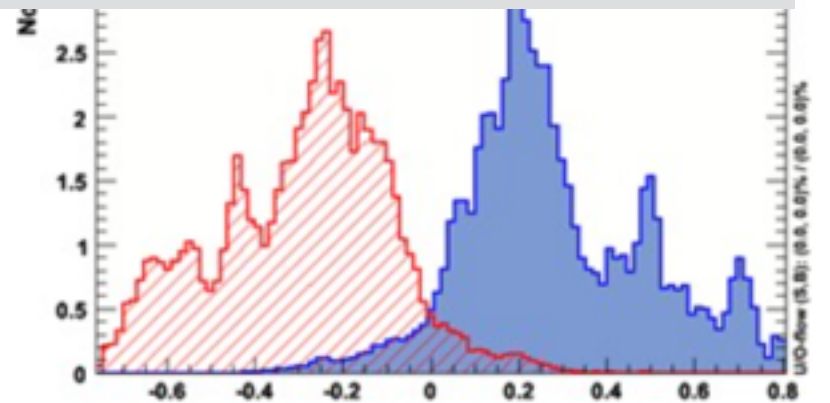
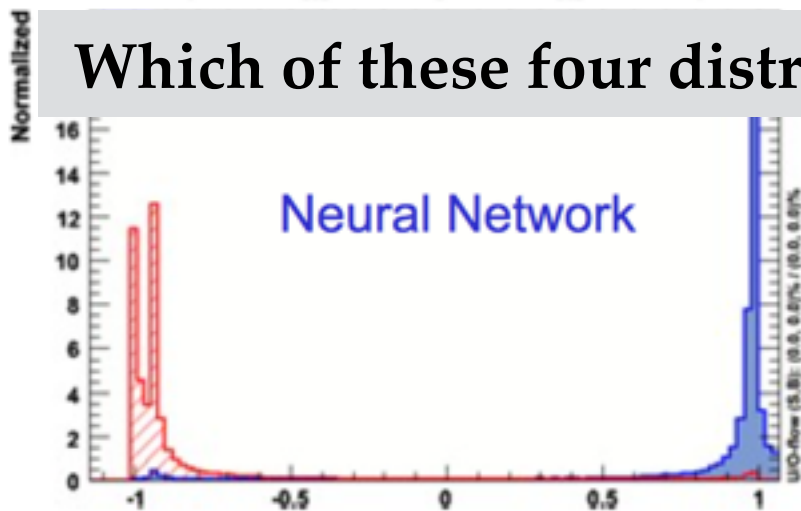
Which of these three distributions are most separated?
How do you “measure” this?



Measuring separation



Which of these four distributions are most separated?



ROC-curves

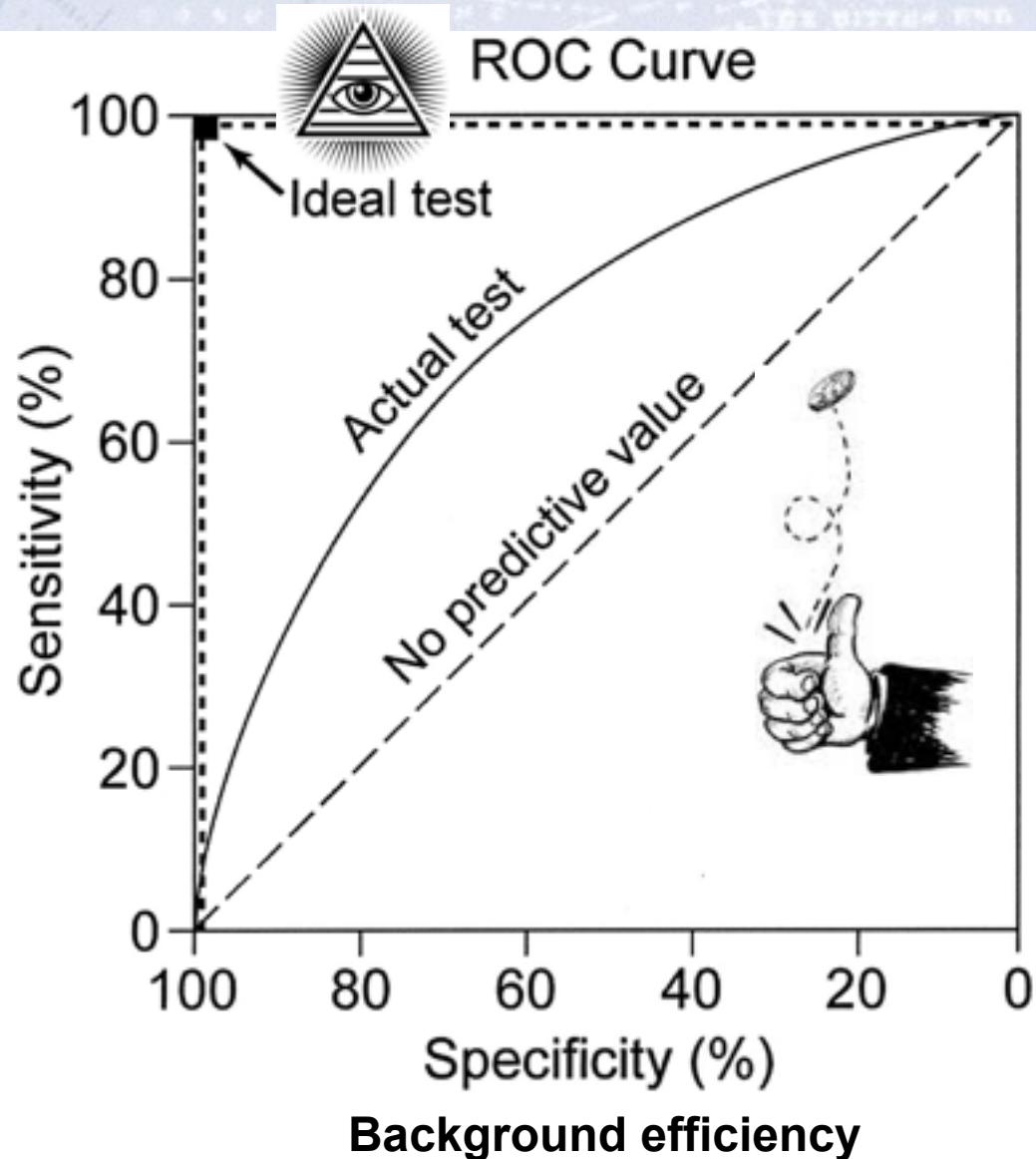
The **Receiver Operating Characteristic** or just ROC-curve is a graphical plot of the sensitivity, or true positive rate, vs. false positive rate.

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!

Signal efficiency



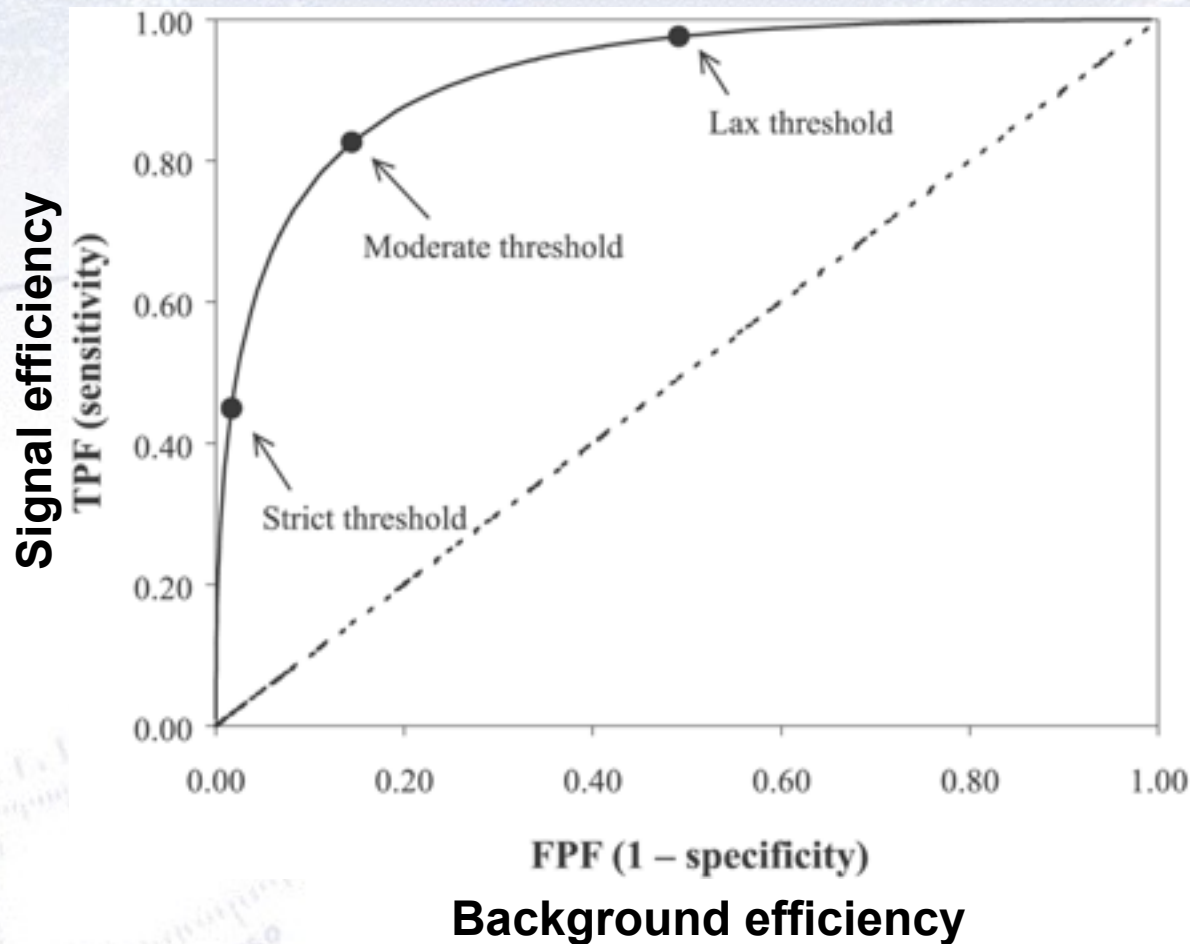
ROC-curves

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!



Useful ROC metrics

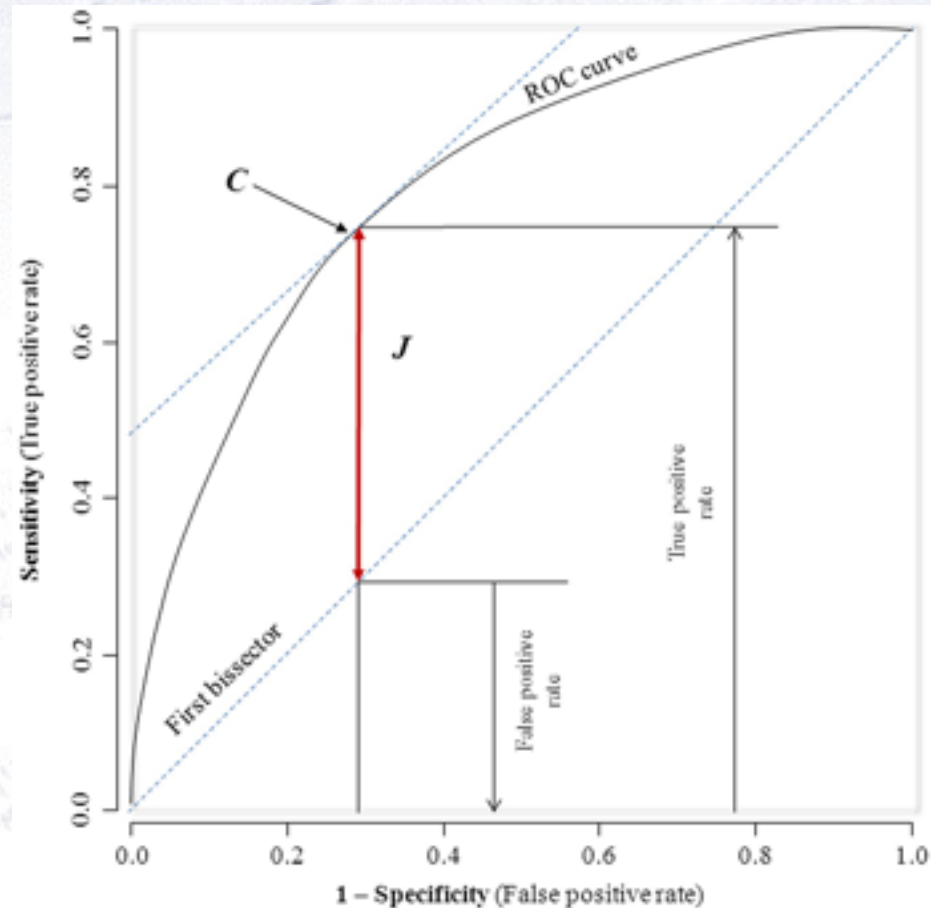
The performance of a test statistic is described fully by the ROC curve itself!

To summarise performance in one single number (i.e. easy to compare!), one used Area Under ROC curve.

Alternatively, people use:

- Signal eff. for a given background eff.
- Background eff. for a given signal eff.
- Youden's index (J), defined as shown in the figure.

The optimal selection **depends entirely on your analysis at hand!**

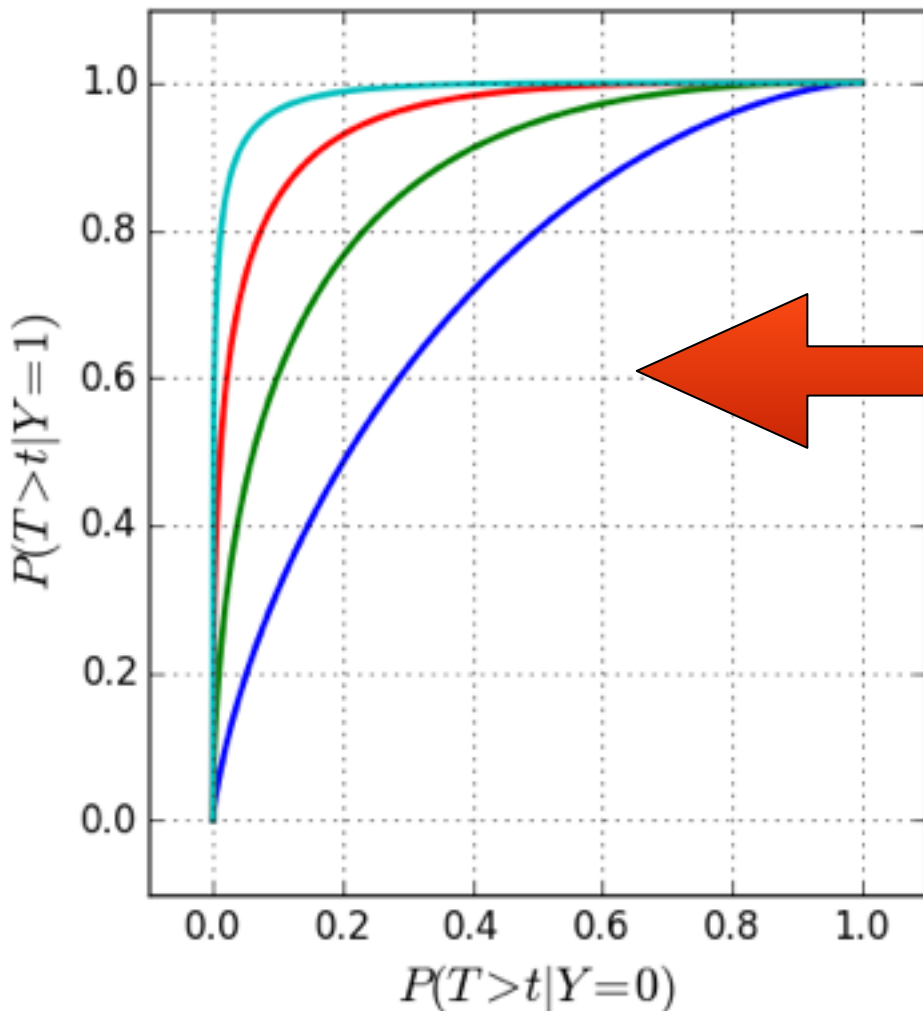




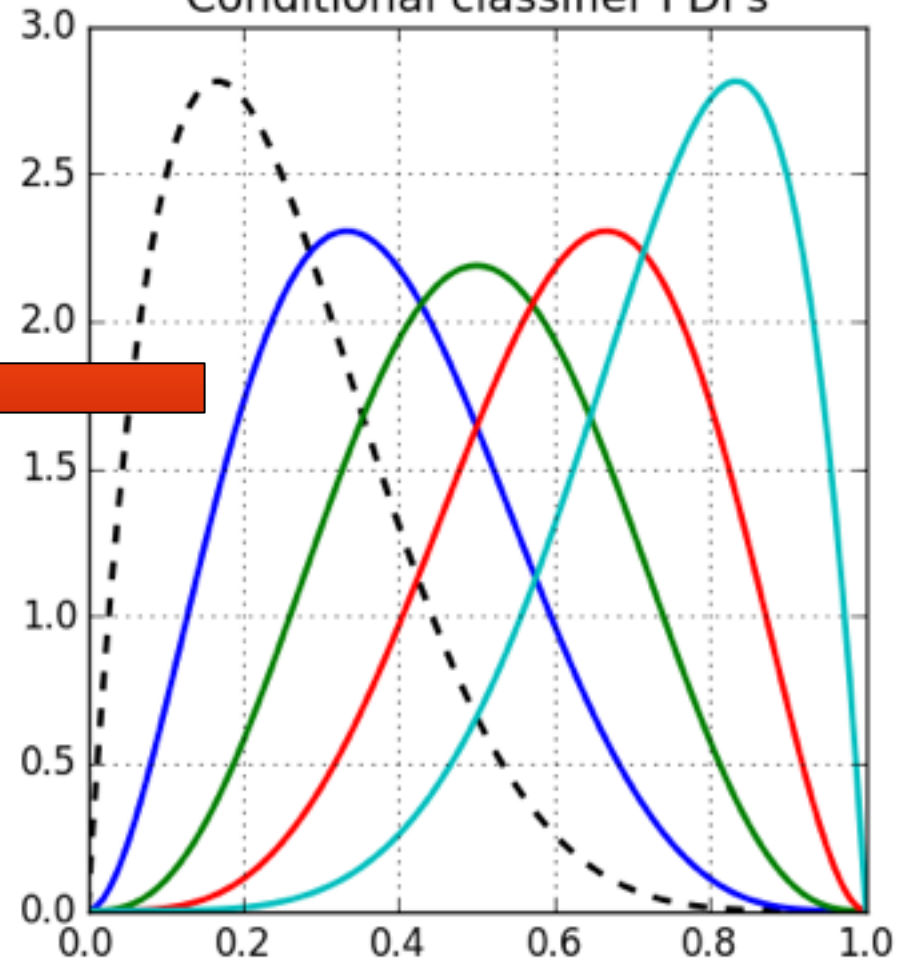
Example of ROC curves in use

Simple case

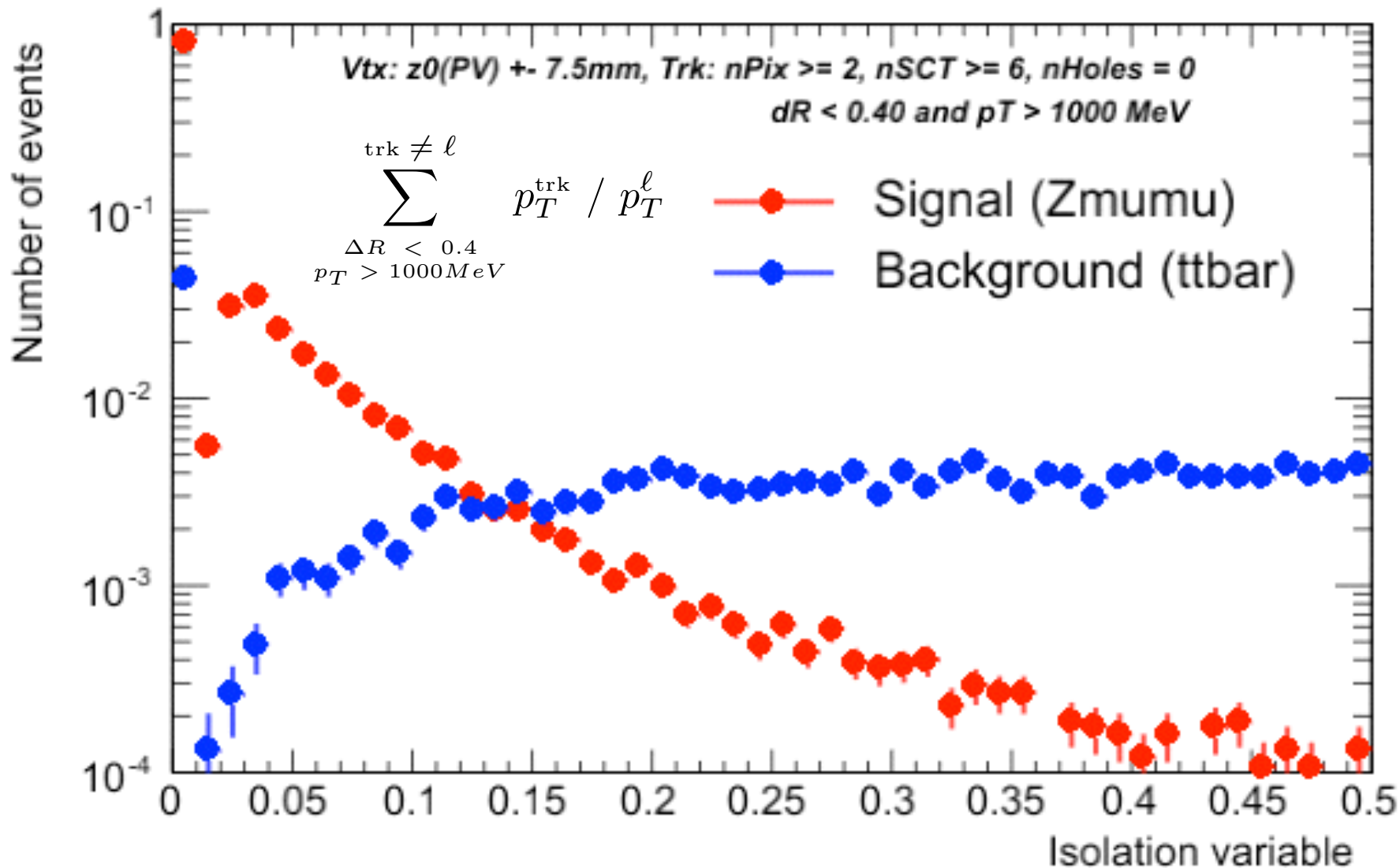
ROC curves



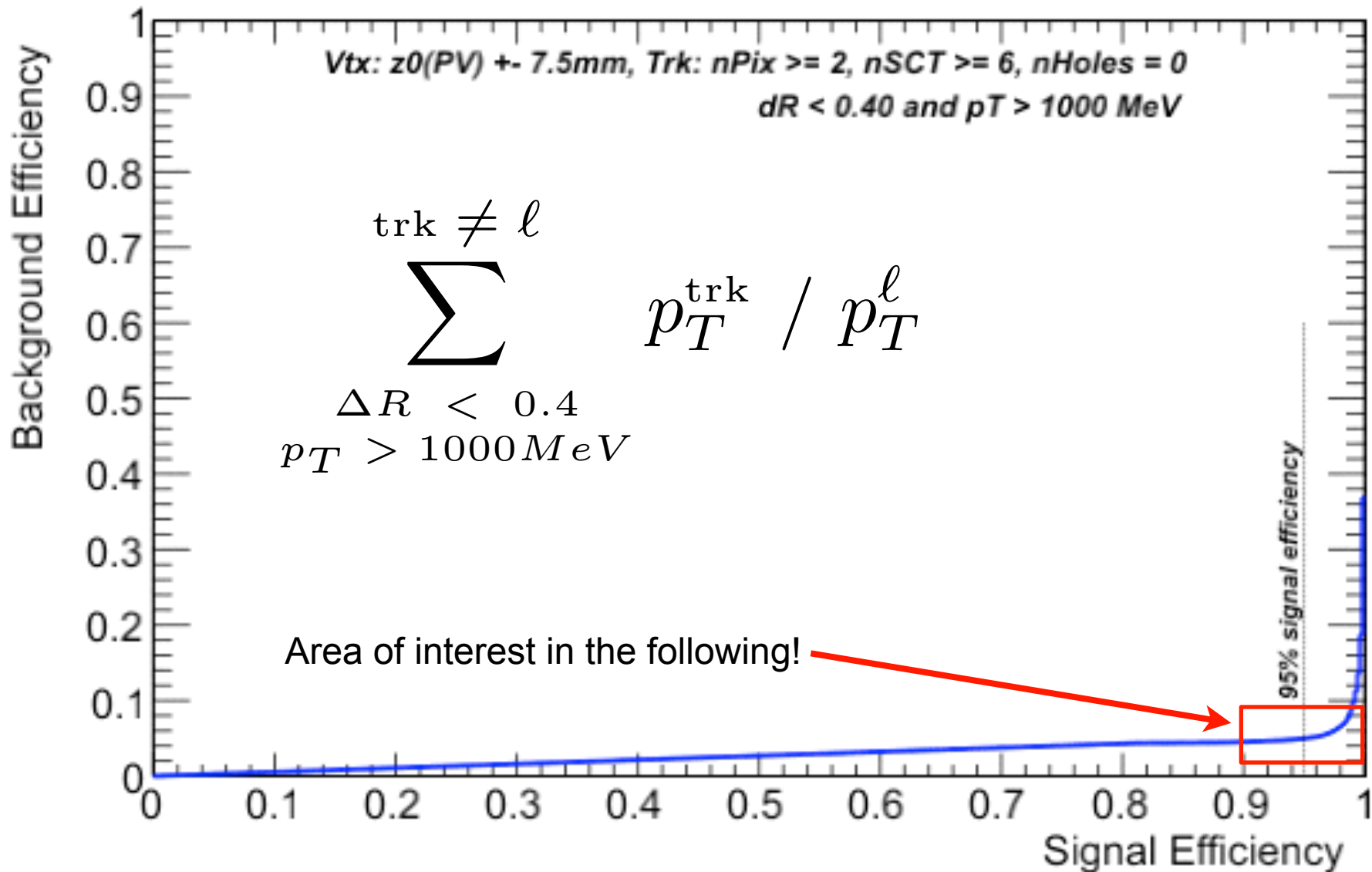
Conditional classifier PDFs



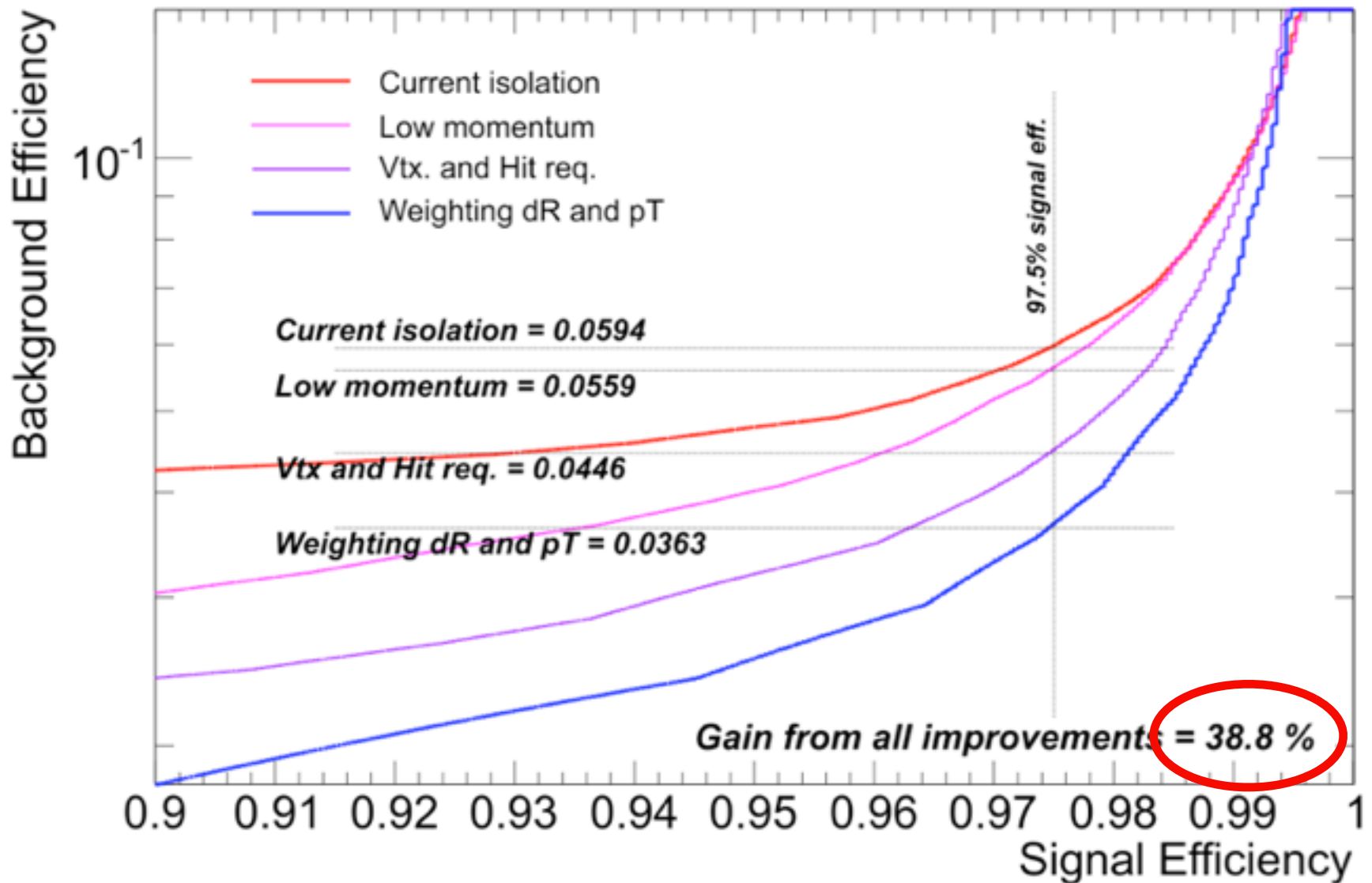
Basic steps - distributions

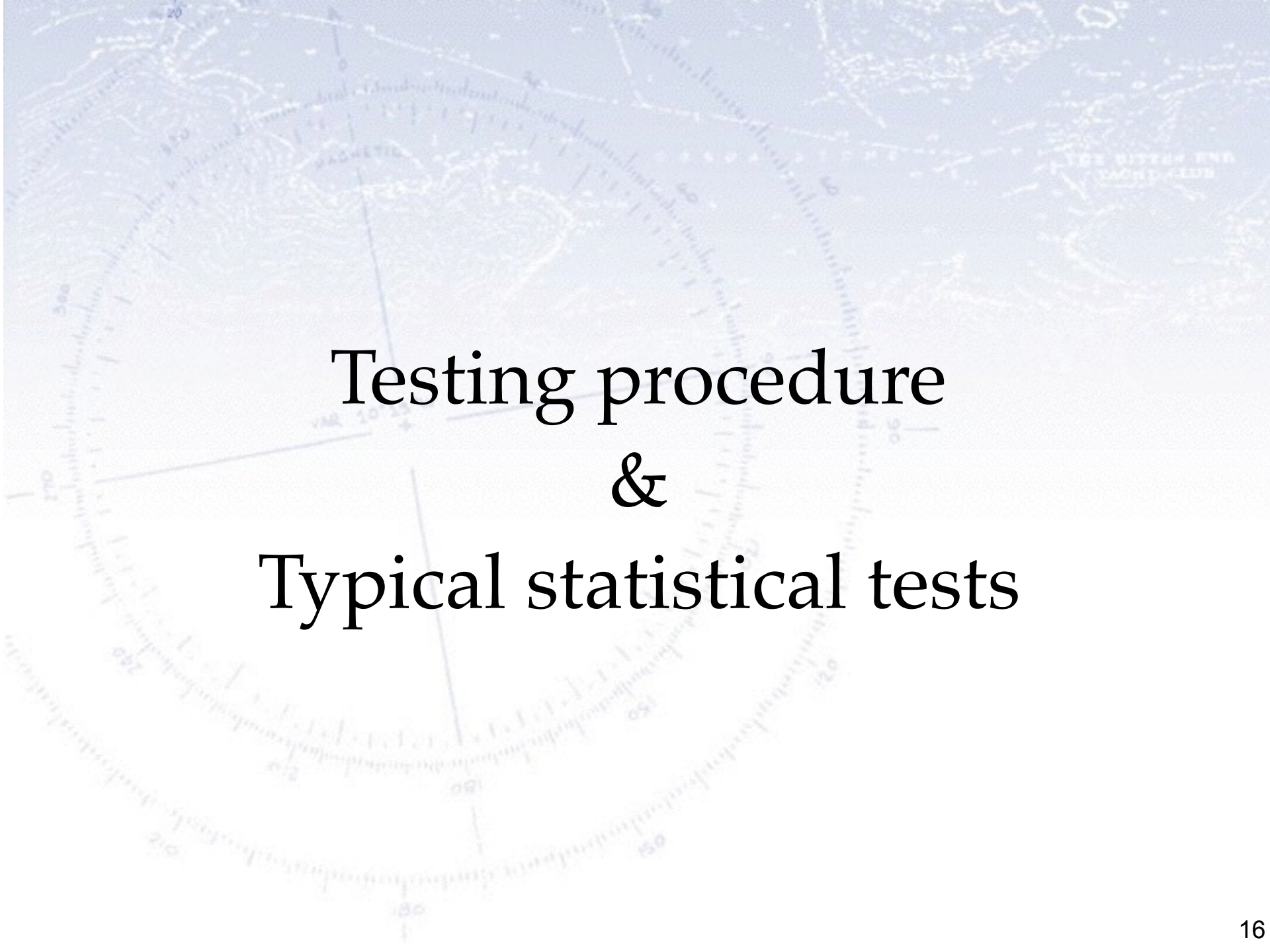


Basic steps - ROC curves



Overall improvement

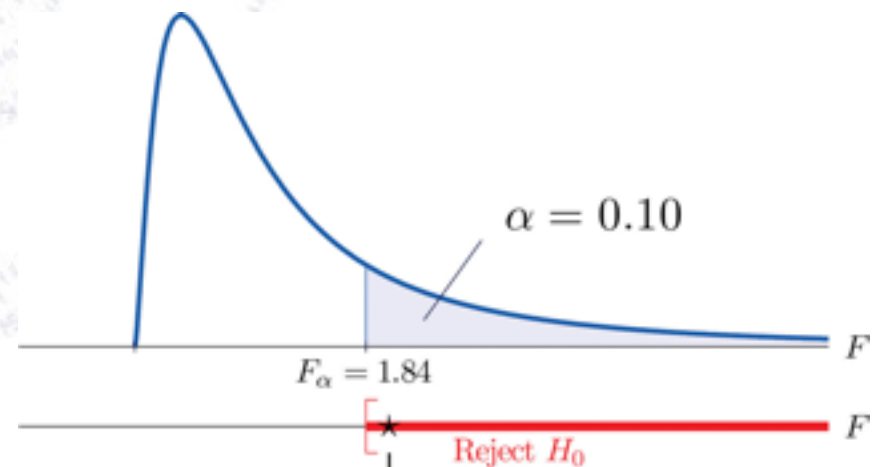
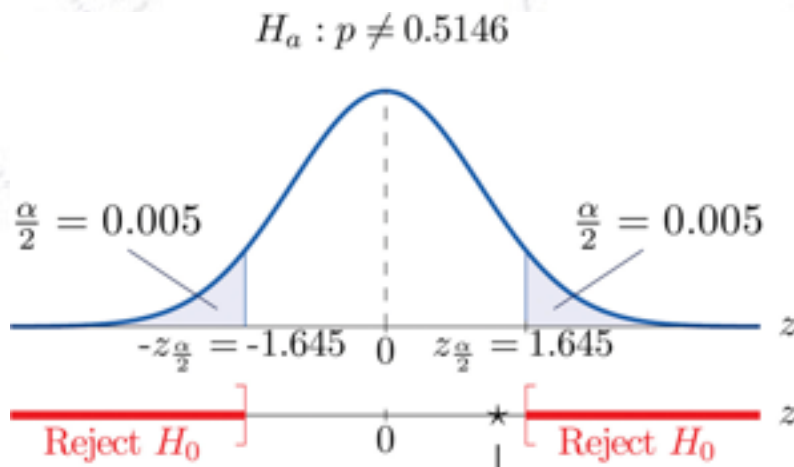




Testing procedure & Typical statistical tests

Testing procedure

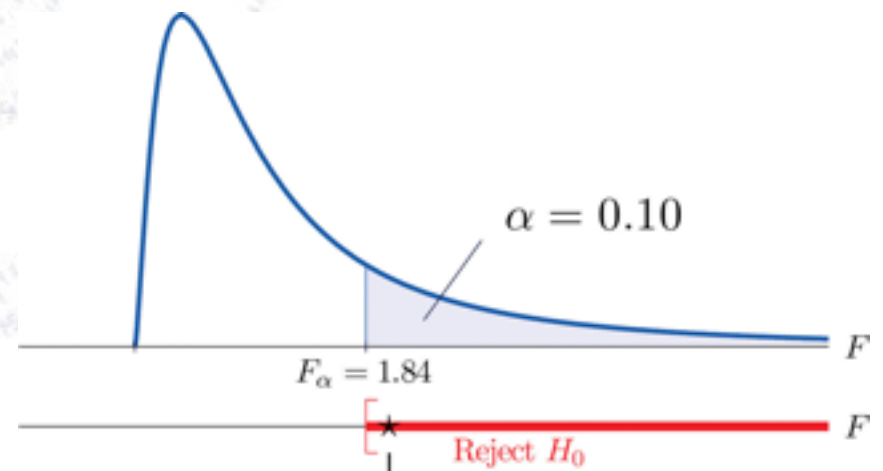
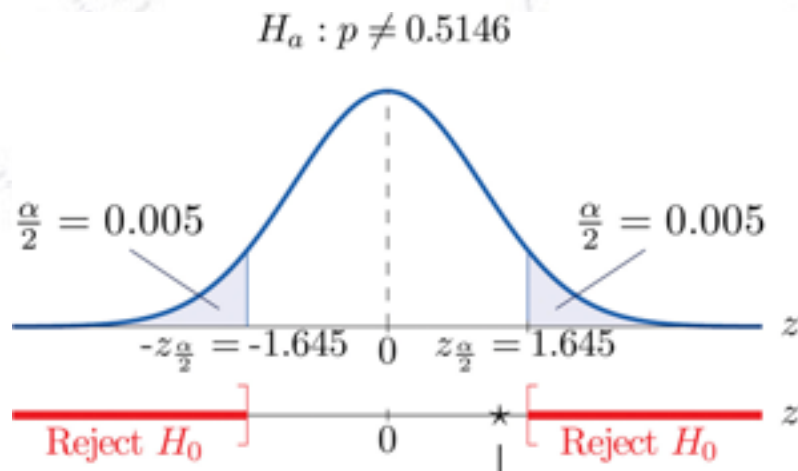
1. Consider an **initial (null) hypothesis**, of which the truth is unknown.
2. State null and **alternative hypothesis**.
3. Consider statistical **assumptions** (independence, distributions, etc.)
4. Decide for appropriate test and state relevant **test statistic**.
5. **Derive the test statistic** 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 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 test statistic** (e.g., t -test, F -test, etc.)
6. **Select a significance level α** (e.g., 0.05, 0.10, etc.)
7. **Compute the test statistic** (e.g., t -statistic, F -statistic, etc.)
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.**

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.

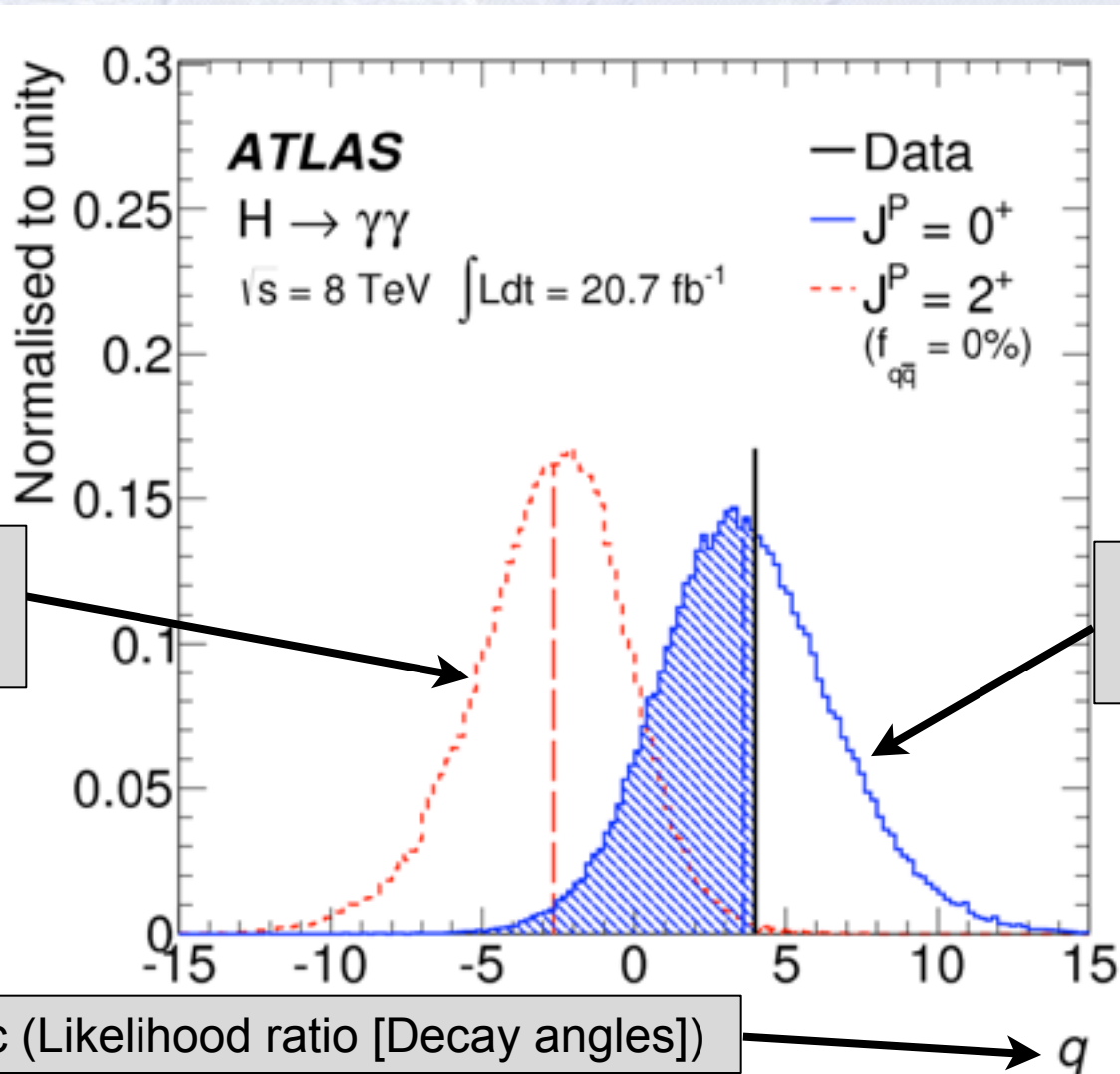
You can **accept** a hypothesis, but this does not exclude accepting other hypothesis.

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?):



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)...

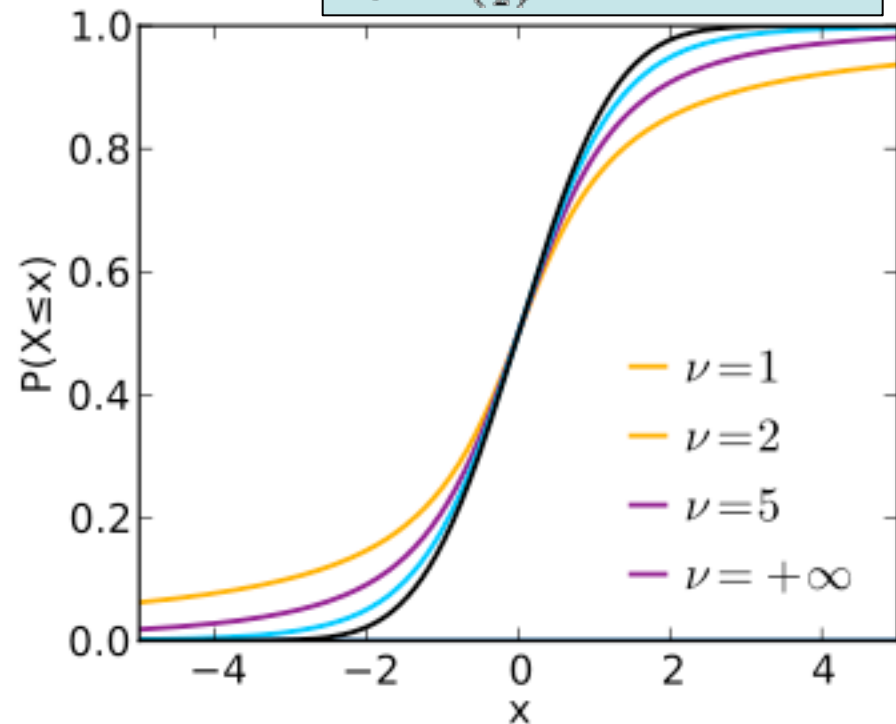
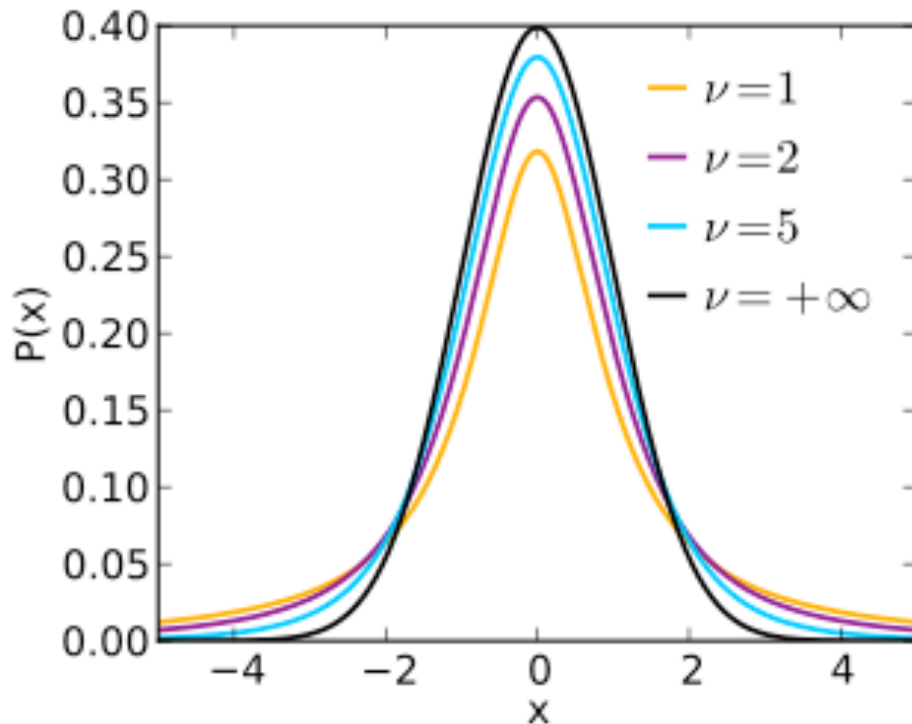
Common statistical tests

- **One-sample test** compares sample (e.g. mean) to known value:
Example: Comparing sample to known constant ($\mu_{\text{exp}} = 2.91 \pm 0.01$ 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 ($\mu_{\text{exp}} = 4.1 \pm 0.6$ vs. $\mu_{\text{control}} = 0.7 \pm 0.4$).
$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma(\bar{x}_1)^2 + \sigma(\bar{x}_2)^2}}$$
- **Paired test** compares paired member difference (to control important variables).
Example: Testing environment influence on twins to control genetic bias ($\mu_{\text{diff}} = 0.81 \pm 0.29$ vs. 0).
- **Chi-squared test** evaluates adequacy of model compared to data.
Example: Model fitted to (possibly binned) data, yielding p-value = $\text{Prob}(\chi^2 = 45.9, N_{\text{dof}} = 36) = 0.125$
- **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.

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 ($\mu_{\text{exp}} = 2.91 \pm 0.01$ 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 ($\mu_{\text{exp}} = 4.1 \pm 0.6$ vs. $\mu_{\text{control}} = 0.7 \pm 0.4$).

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma(\bar{x}_1)^2 + \sigma(\bar{x}_2)^2}}$$

- **Paired test** compares paired member difference (to control important variables).
Example: Testing environment influence on twins to control genetic bias ($\mu_{\text{diff}} = 0.81 \pm 0.29$ vs. 0).

Things to consider:

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

Rule-of-thumb: If $N > 30$ 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.

Two-Tailed Versus One-Tailed Hypothesis Tests

Figure A:
Two-Tailed Test

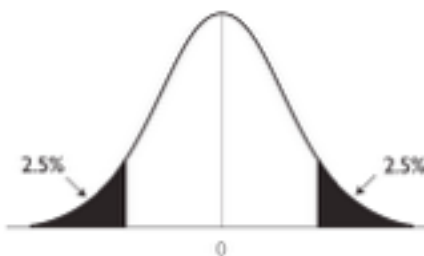
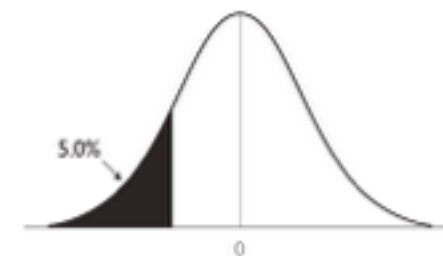


Figure B:
One-Tailed Test
(Left-Tailed Test)



Chi-squared test

Without any further introduction...

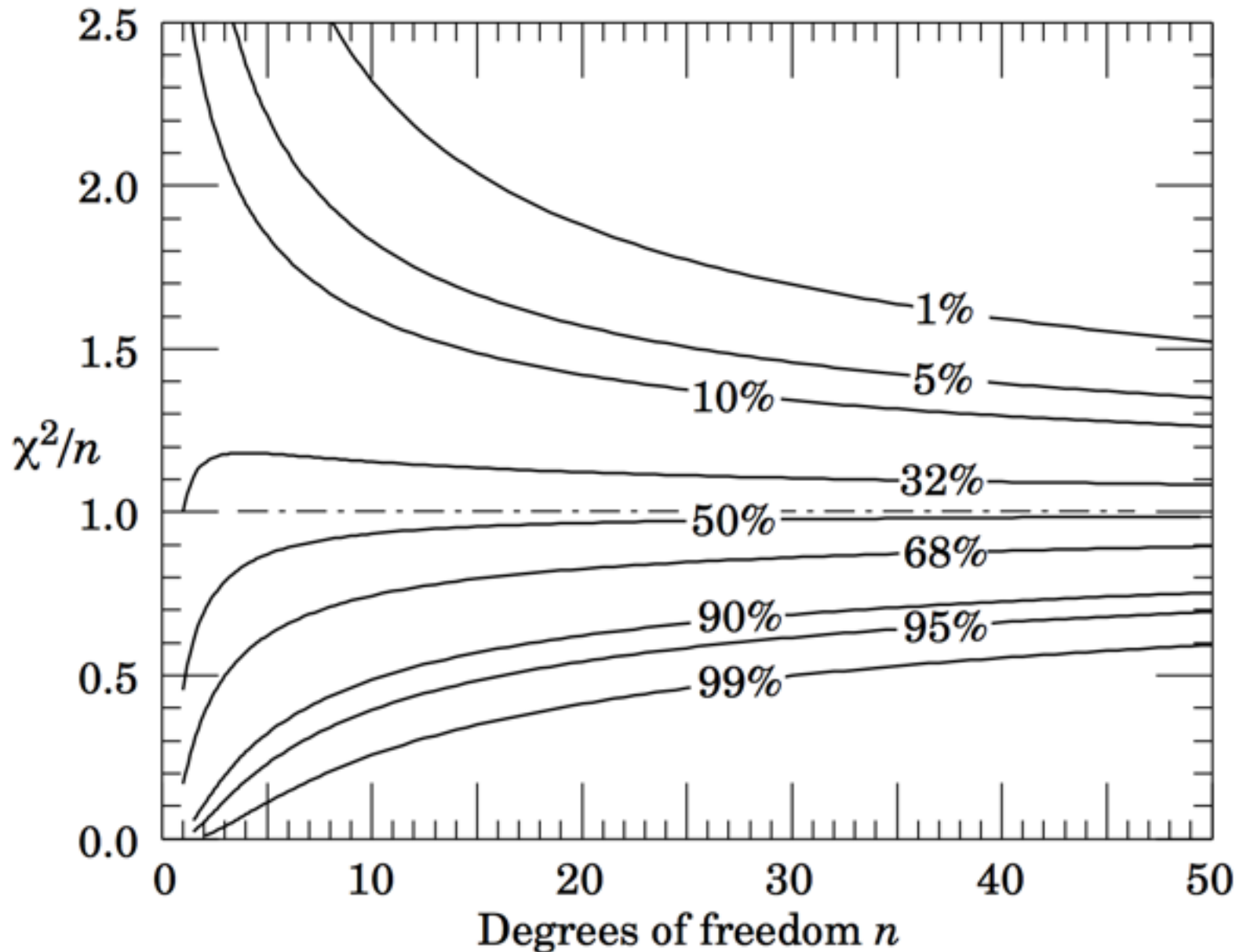
$$\chi^2(\bar{\theta}) = \sum_{i=1}^N \frac{(y_i - \lambda(x_i; \bar{\theta}))^2}{\sigma_i^2}$$

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

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

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

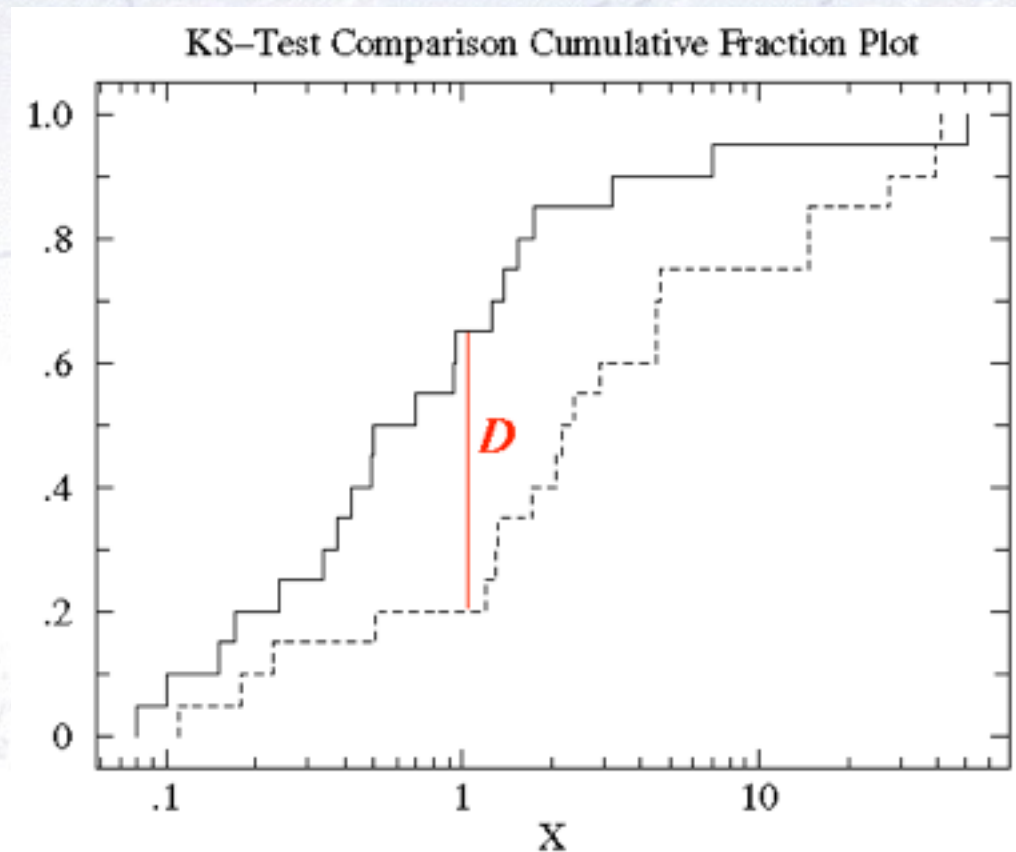
Chi-squared test



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

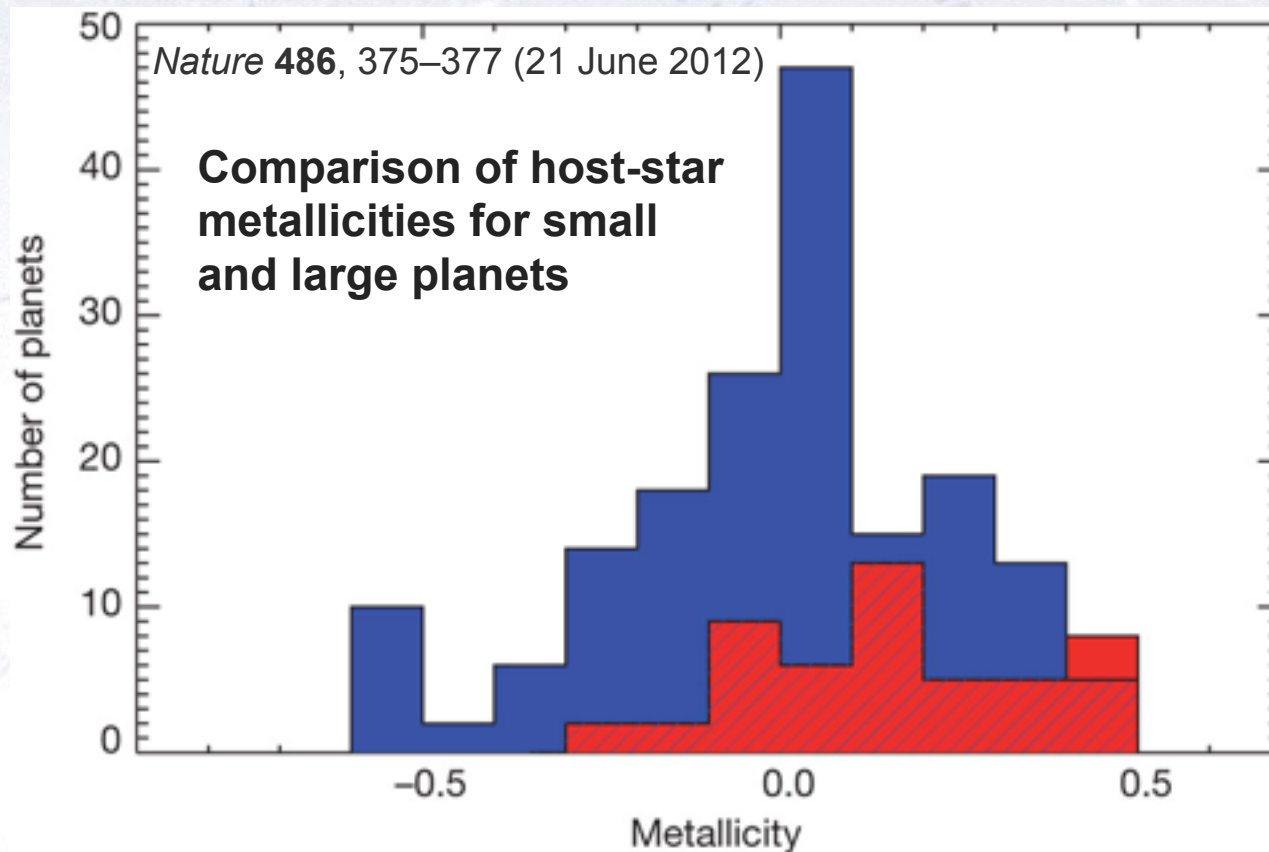


The Kolmogorov test measures the maximal distance between the integrals of two distributions and gives a probability of being from the same distribution.

Kolmogorov-Smirnov test

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

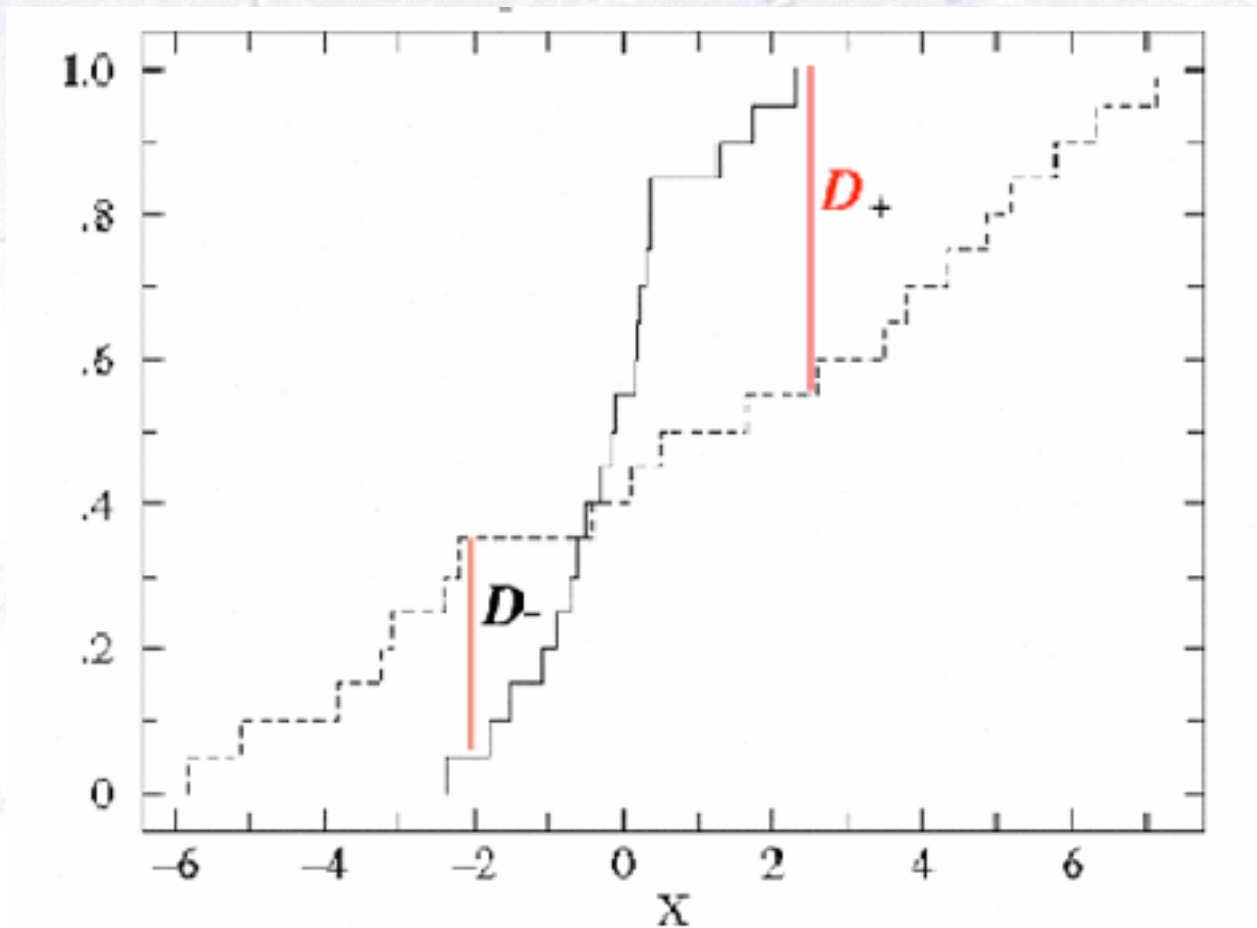
Example: Compatibility between function and sample or between two samples, yielding p-value = 0.87



“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]

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.
Example: Comparing sample to known constant ($\mu_{\text{exp}} = 2.91 \pm 0.01$ vs. $c = 3.00$).
$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$
- **Two-sample test** compares two samples (e.g. means).
Example: Comparing sample to control ($\mu_{\text{exp}} = 4.1 \pm 0.6$ vs. $\mu_{\text{ctrl}} = 3.7 \pm 0.4$).
$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$
- **Paired test** compares paired member difference (to control important variables).
Example: Testing environment influence on twins to control genetic bias ($\mu_{\text{diff}} = 0.81 \pm 0.29$ vs. 0).
- **Chi squared test** evaluates adequacy of model compared to data.
Example: Model fitted to (possibly binned) data, yielding p-value = $\text{Prob}(\chi^2 = 45.9, N_{\text{dof}} = 36) = 0.125$
- **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.

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:

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

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

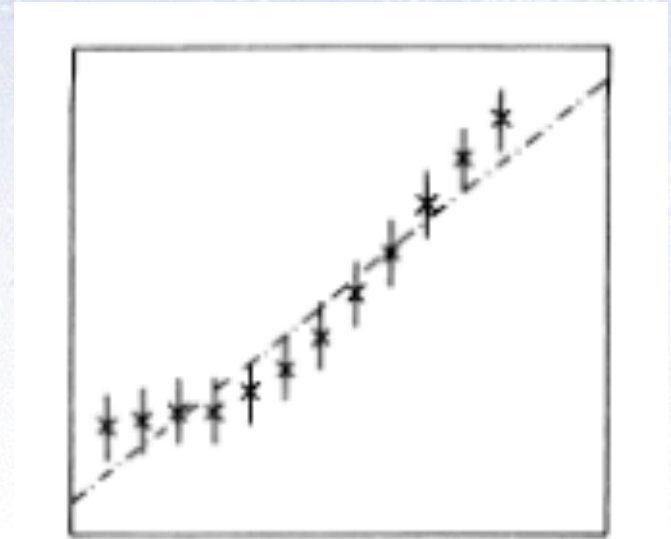


Fig. 8.3. A straight line through twelve data points.

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

Note: The WW runs test requires $N > 10-15$ for the output to be approx. Gaussian! 32

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 categorical 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 | <i>Row total</i> |
|---------------------|------------|--------------|------------------|
| DiETING | 1 | 9 | <i>10</i> |
| Non-dieting | 11 | 3 | <i>14</i> |
| <i>Column total</i> | <i>12</i> | <i>12</i> | <i>24</i> |

Is there a correlation between dieting and gender?

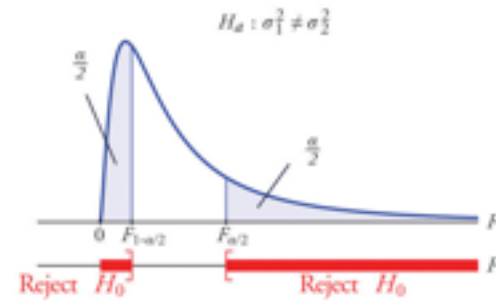
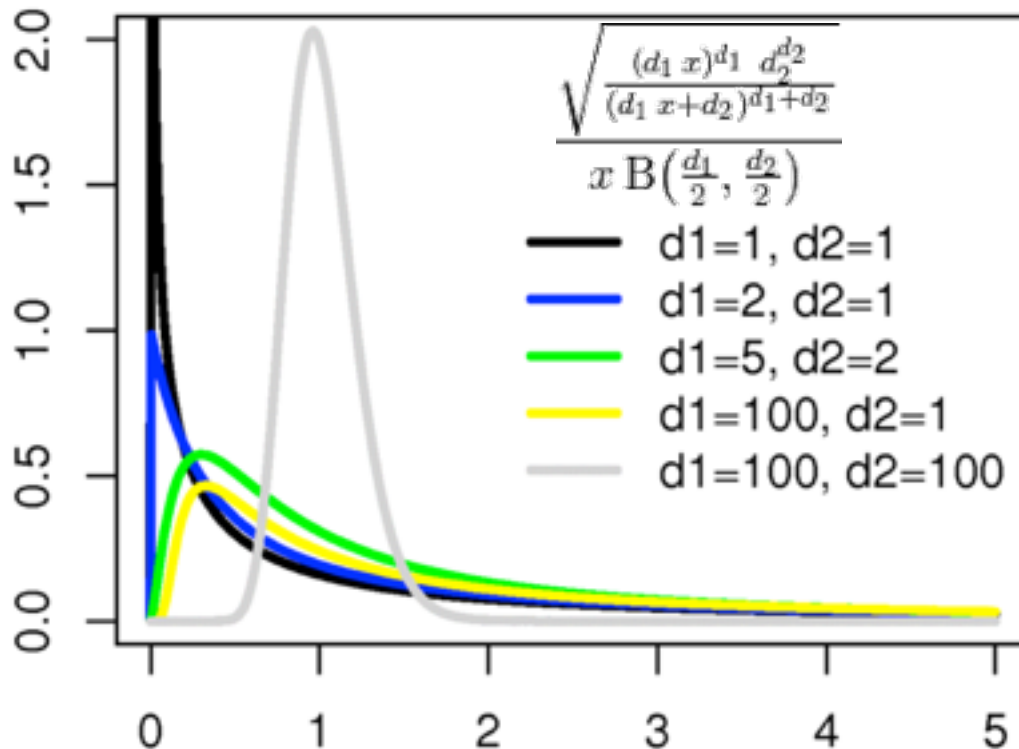
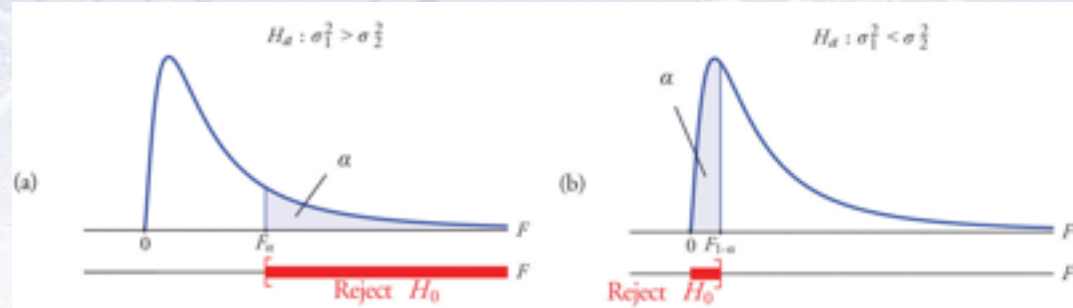
The Chi-square test is not optimal, as there are (several) entries, that are very low (< 5), but Fisher's exact test gives the answer:

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

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.

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 |
| $B_s \rightarrow \mu\mu$ | 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 |