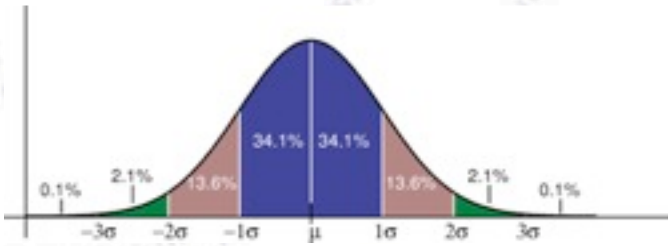
Applied Statistics

Hypothesis Testing






Troels C. Petersen (NBI)



"Statistics is merely a quantization of common sense"

Taking decisions

You are asked to take a decision or give judgement - it is yes-or-no.

Given data - how to do that best?

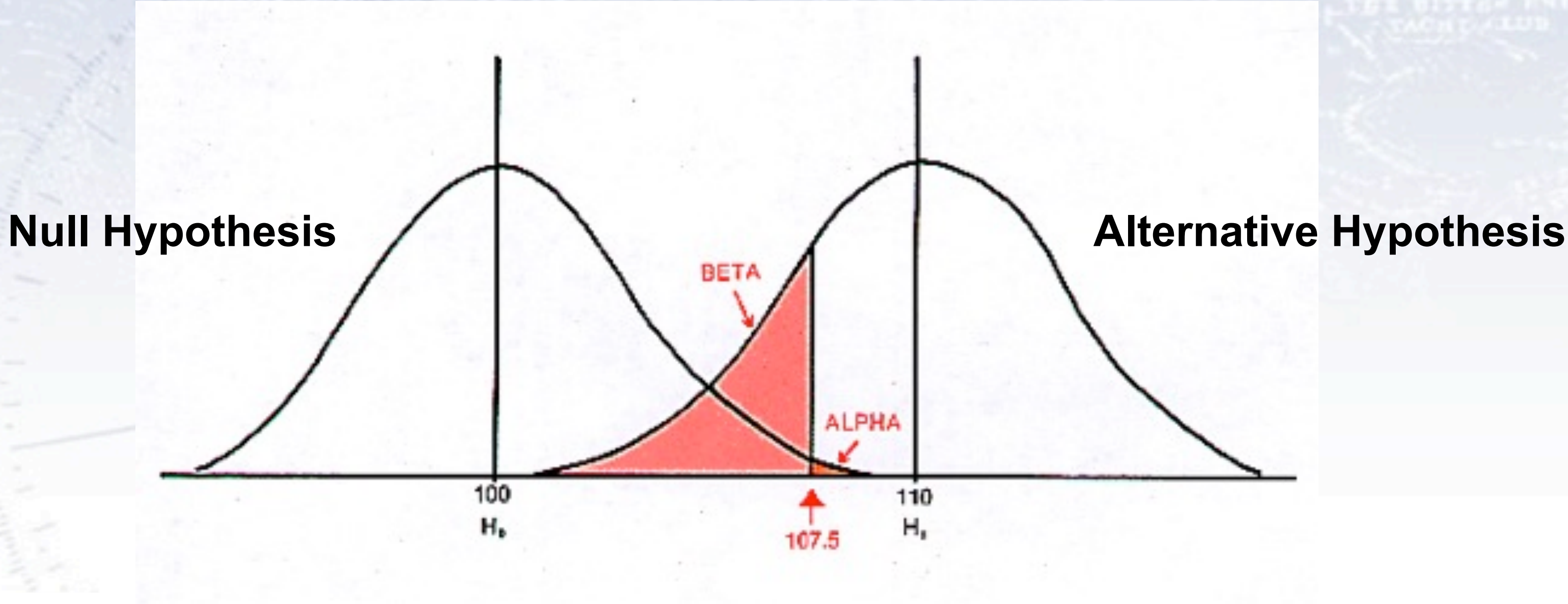
That is the basic question in hypothesis testing.

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



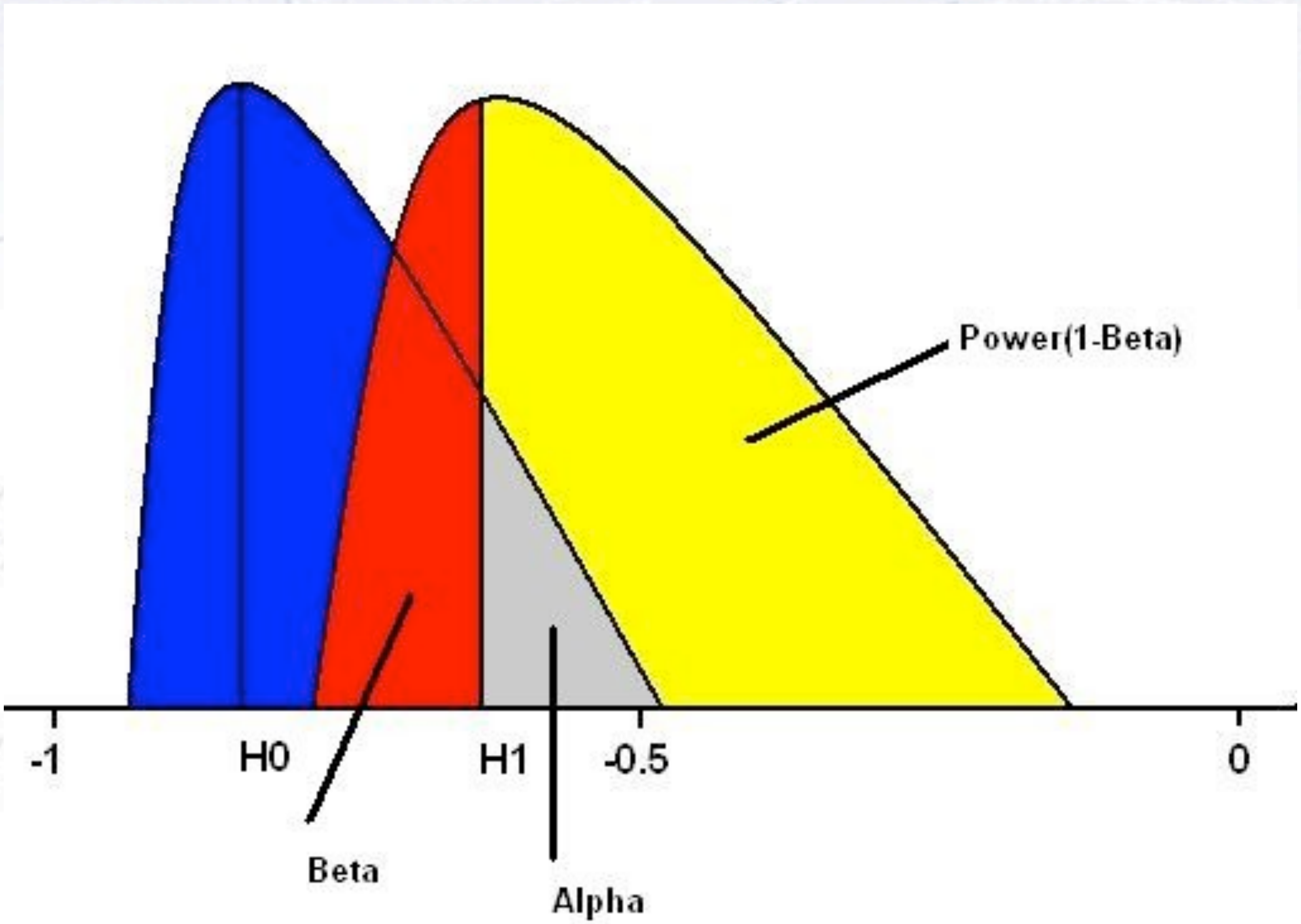
REALITY

STATISTICAL DECISION:

	Null is True	Null is False
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 distributions for the Null and Alternative, which are as separated from each other as possible (to minimize α and β).



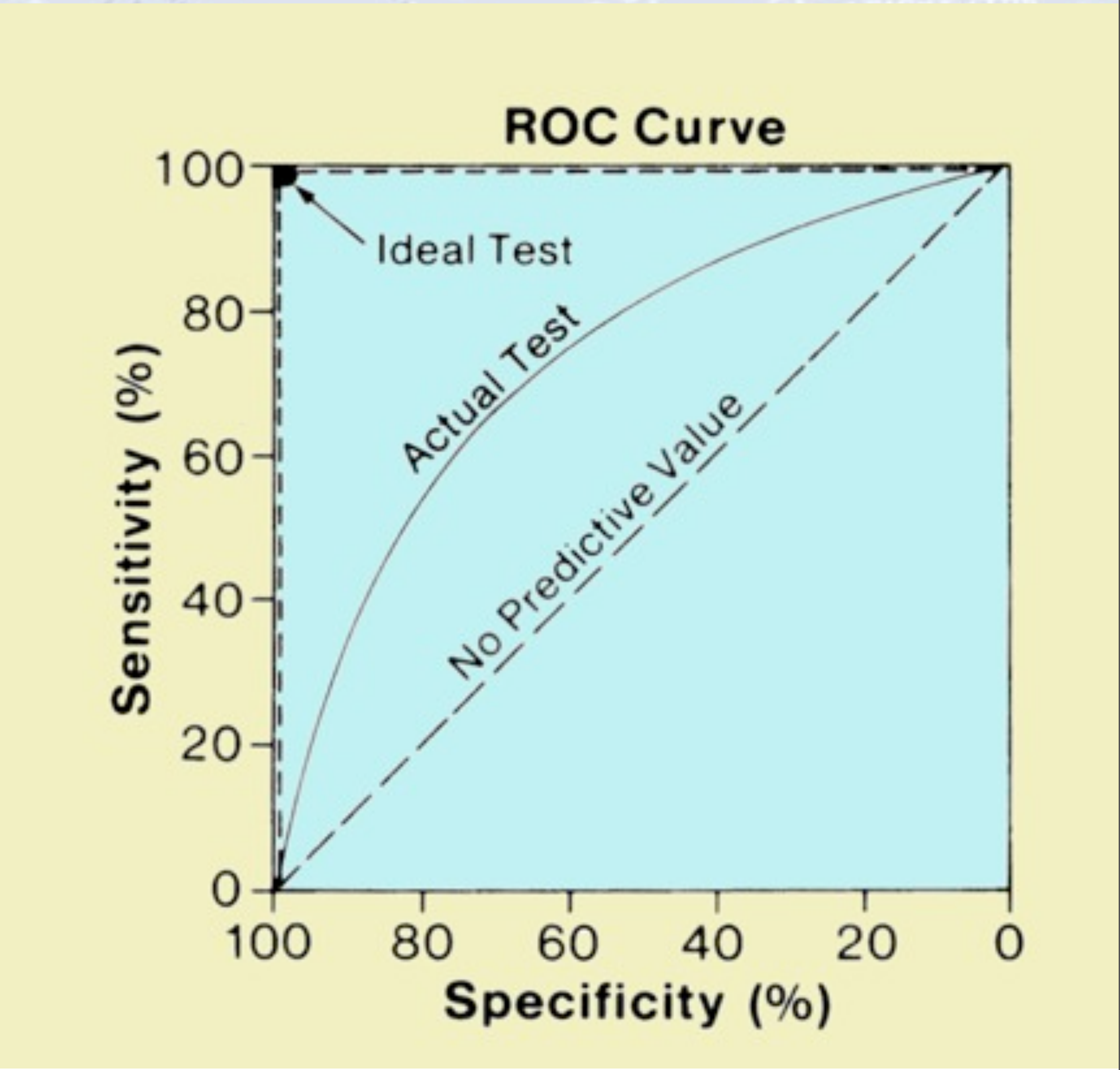
ROC-curves

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.

It can also detect overtraining!



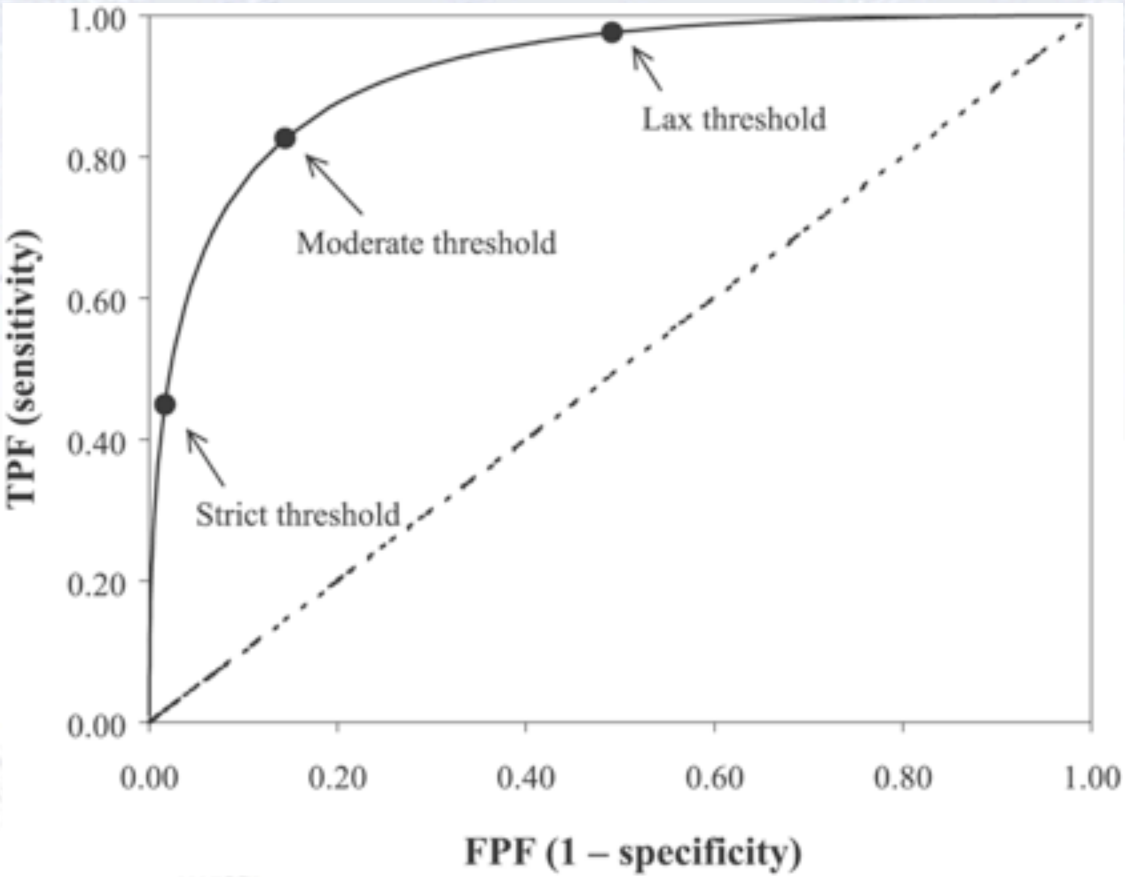
ROC-curves

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.

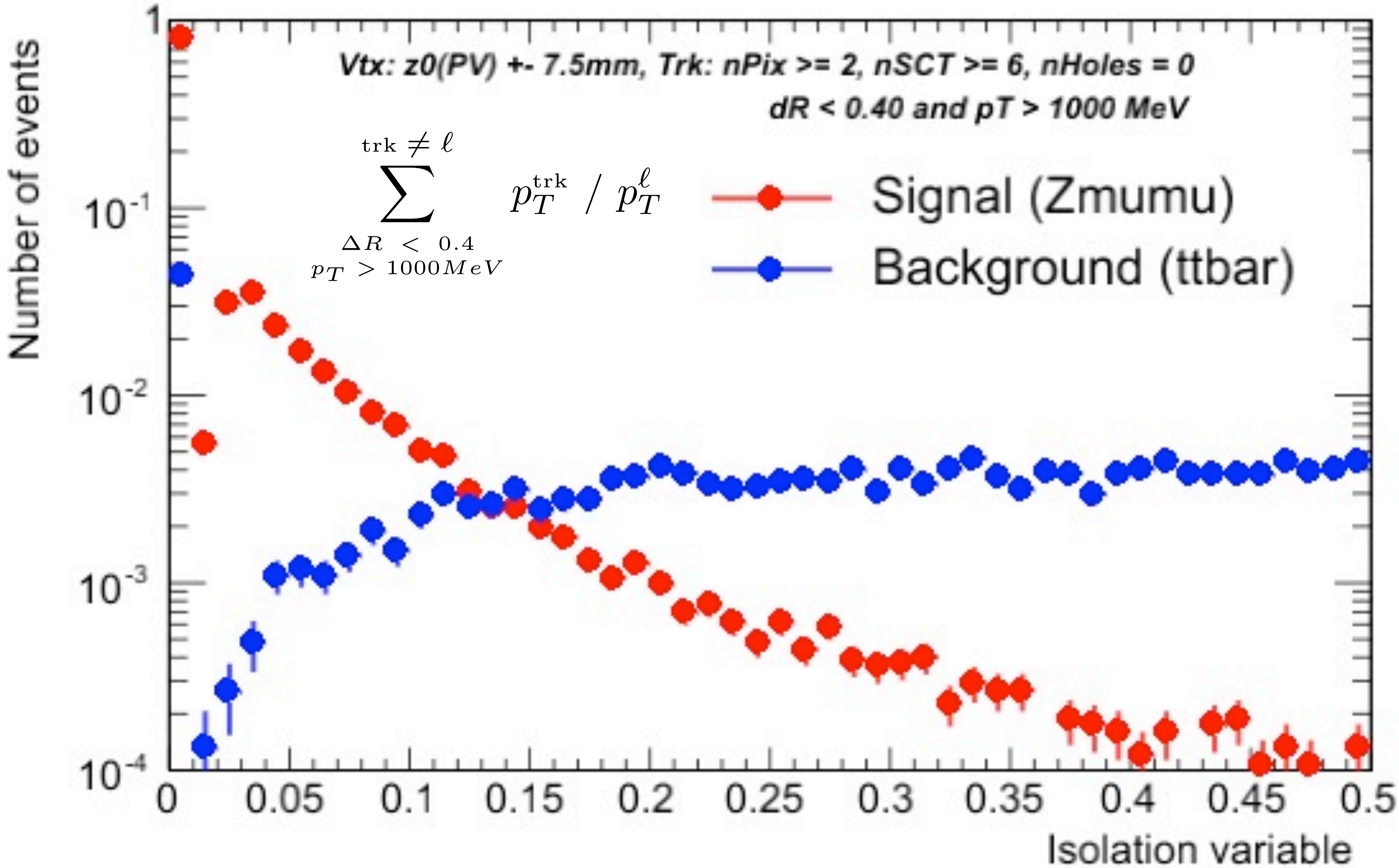
Often, it requires a testing data set to actually see how well a test is performing.

It can also detect overtraining!

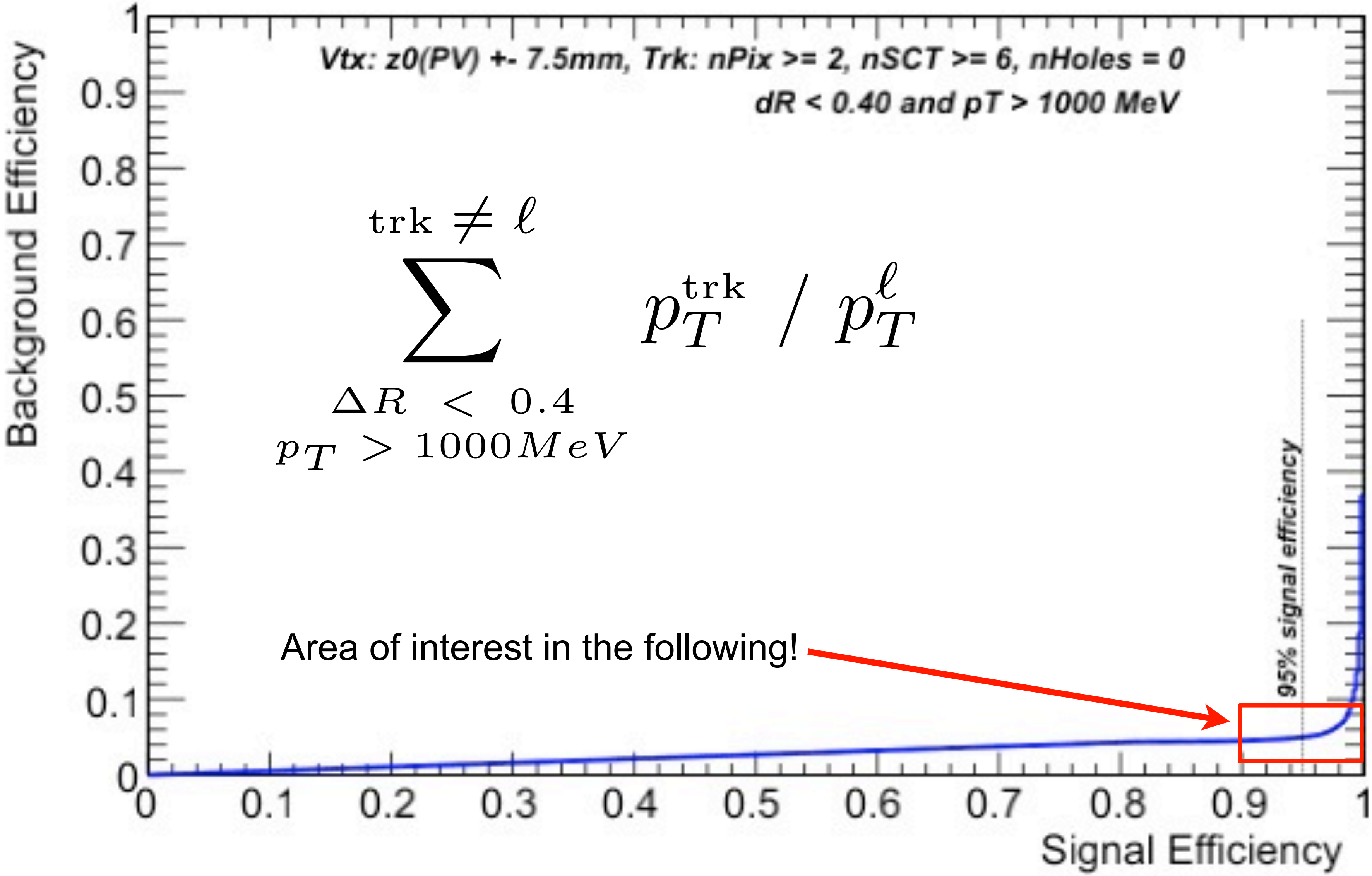


Example of ROC curves in use

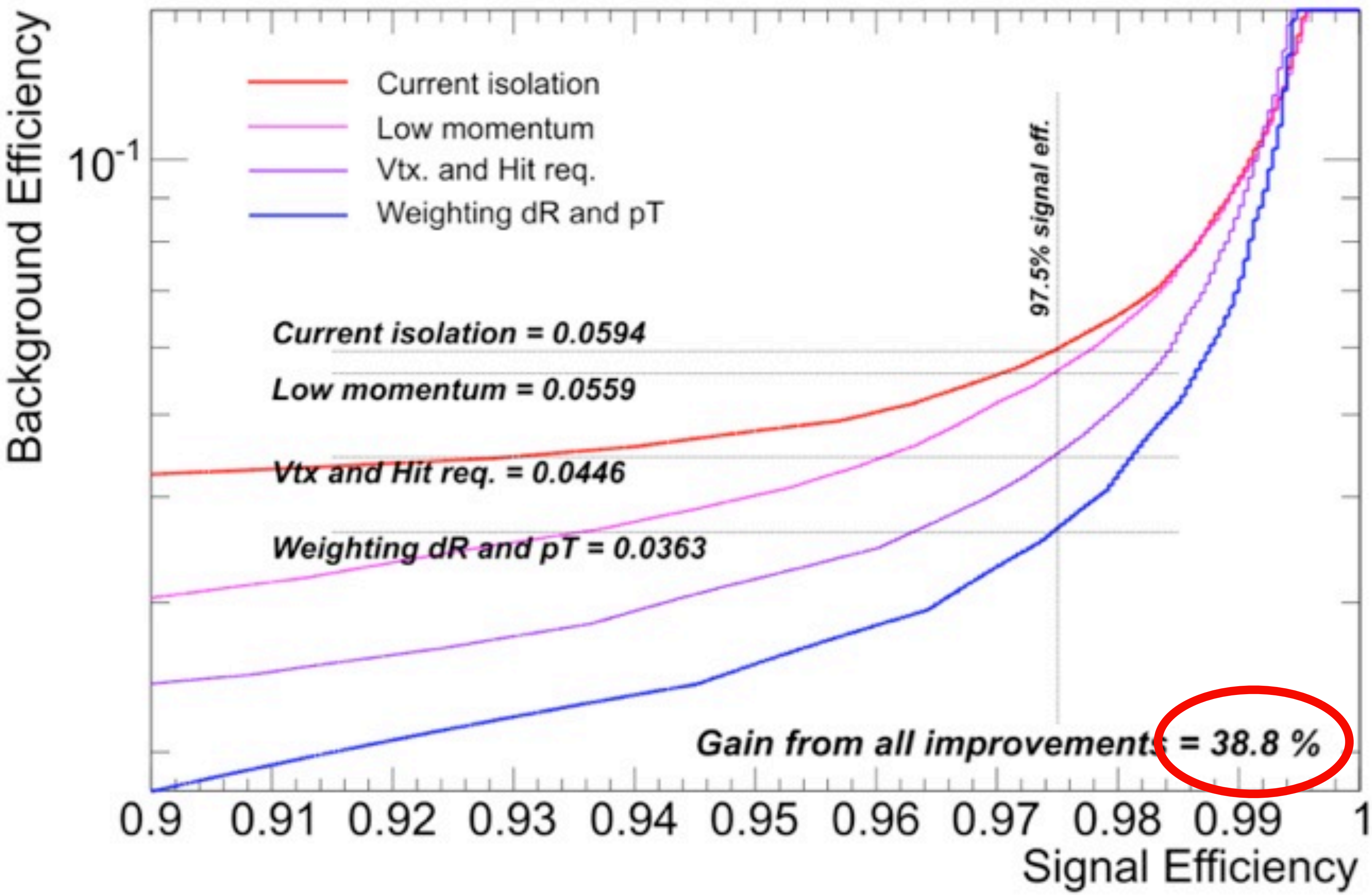
Basic steps - distributions



Basic steps - ROC curves



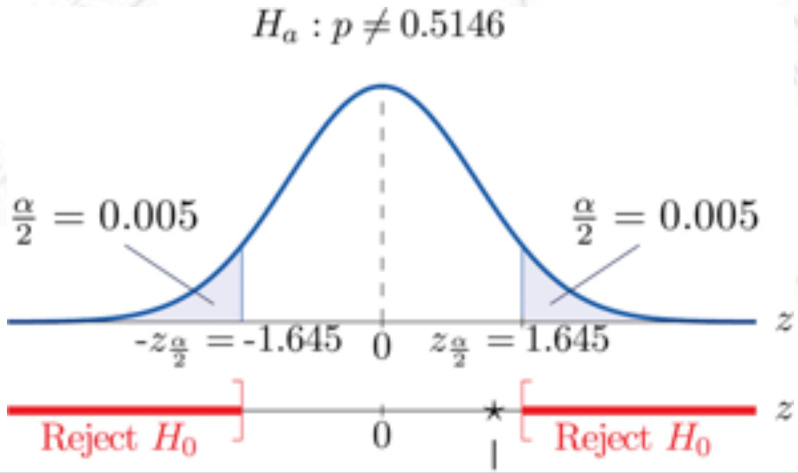
Overall improvement

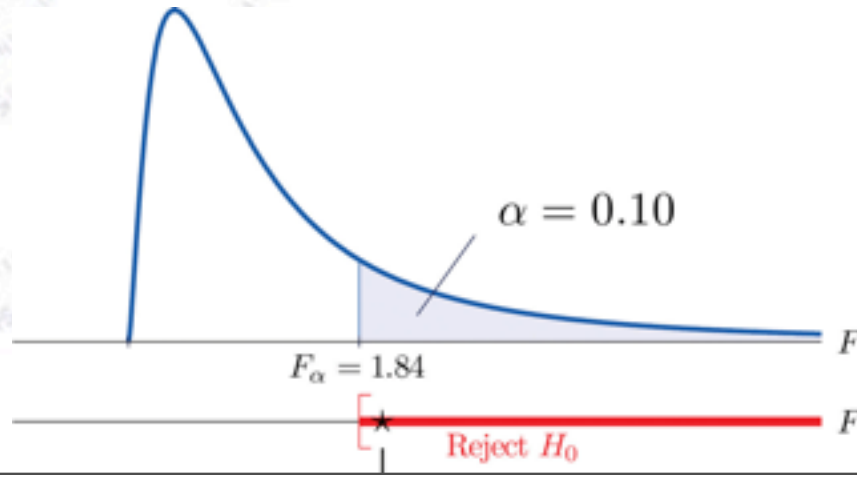


Testing procedure
&
Typical statistical tests

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 observations/data (blinded) **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**.



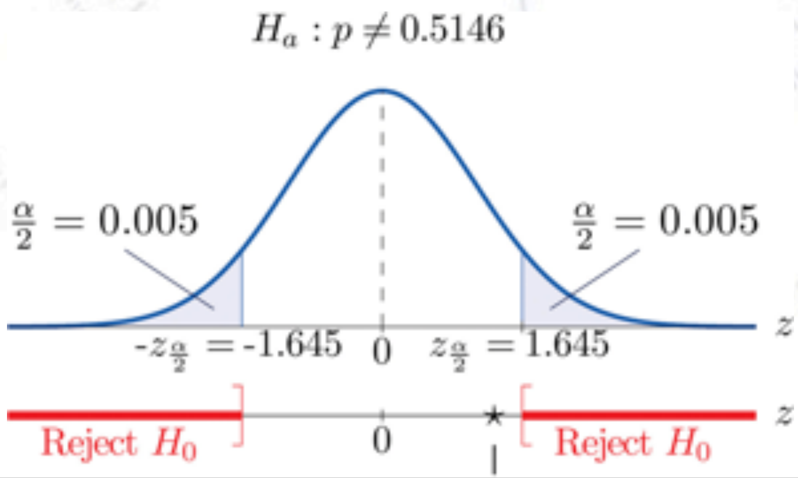


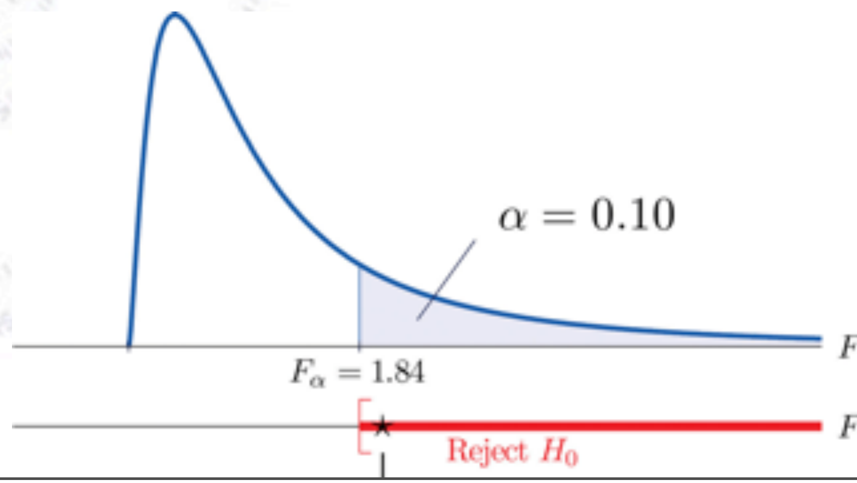
Monday, September 23, 13

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** … othesis.
   In star … nt's t, etc.)
6. **Select** … which null
   hypoth … (physics)).
7. Compu …
8. From *t* … (**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.





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...

Common statistical tests

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

  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). $$z = \frac{(\overline{x}_1 - \overline{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n1} + \frac{\sigma_2^2}{n2}}}$$

  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.

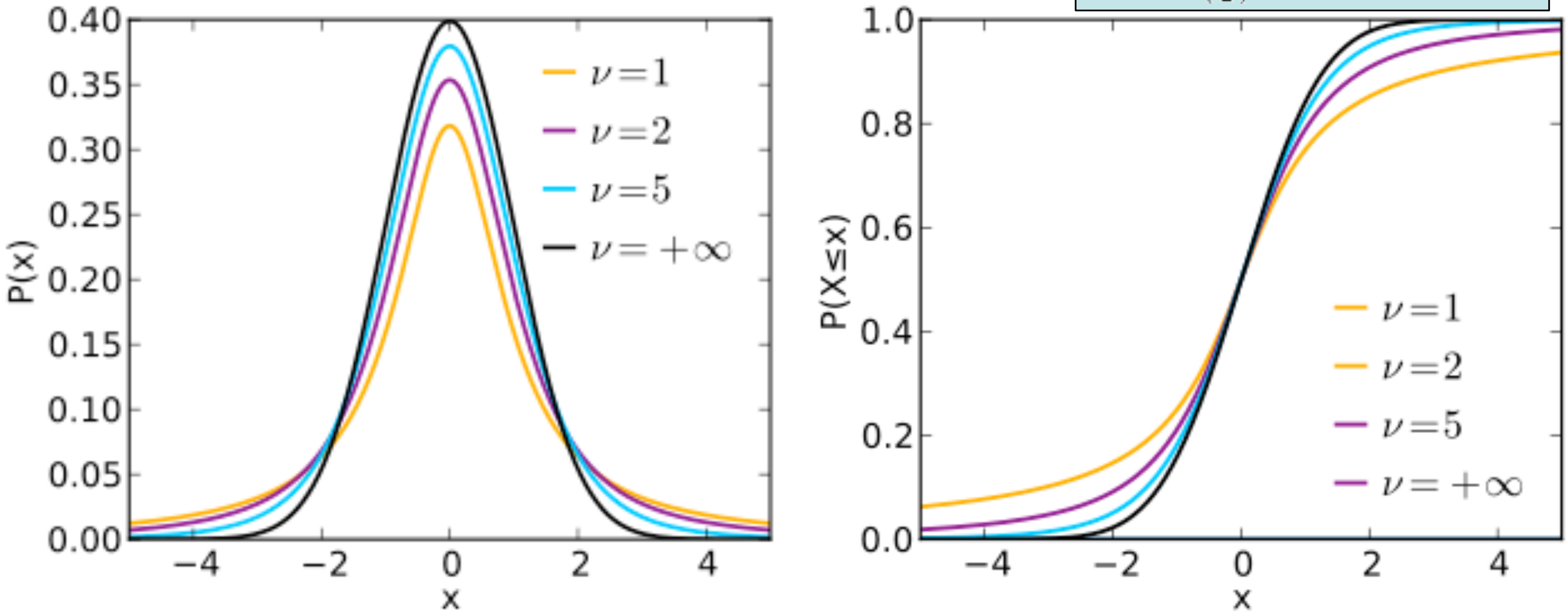
  Example: Compatibily between function and sample or between two samples, yielding p-value = 0.87

- **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: $$z = \frac{\overline{x} - \mu_0}{\sigma}\sqrt{n}$$
  Example: Comparing sample to known constant ($\mu_{exp} = 2.91 \pm 0.01$ vs. $c = 3.00$).

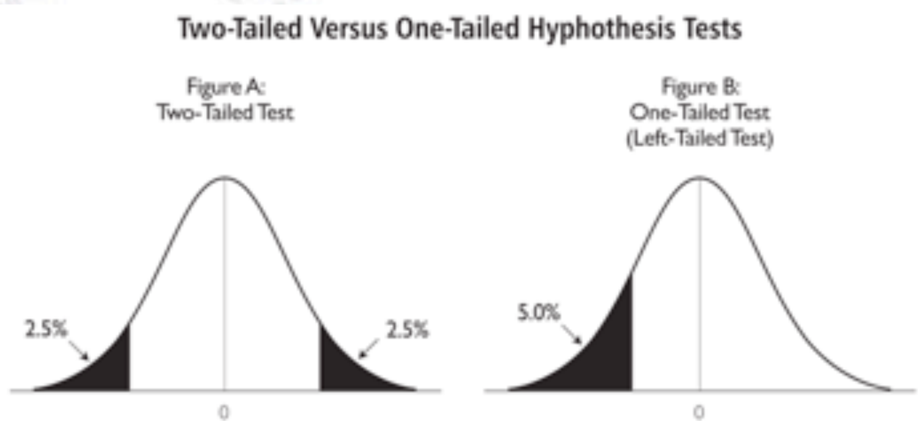
- **Two-sample test** compares two samples (e.g. means). $$z = \frac{(\overline{x}_1 - \overline{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n1} + \frac{\sigma_2^2}{n2}}}$$
  Example: Comparing sample to control ($\mu_{exp} = 4.1 \pm 0.6$ vs. $\mu_{control} = 0.7 \pm 0.4$).

- **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).

Things to consider:

- Variance known (Z-test) vs. Variance unknown (T-test).
  Rule-of-thumb: If $N > 30$ and 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.

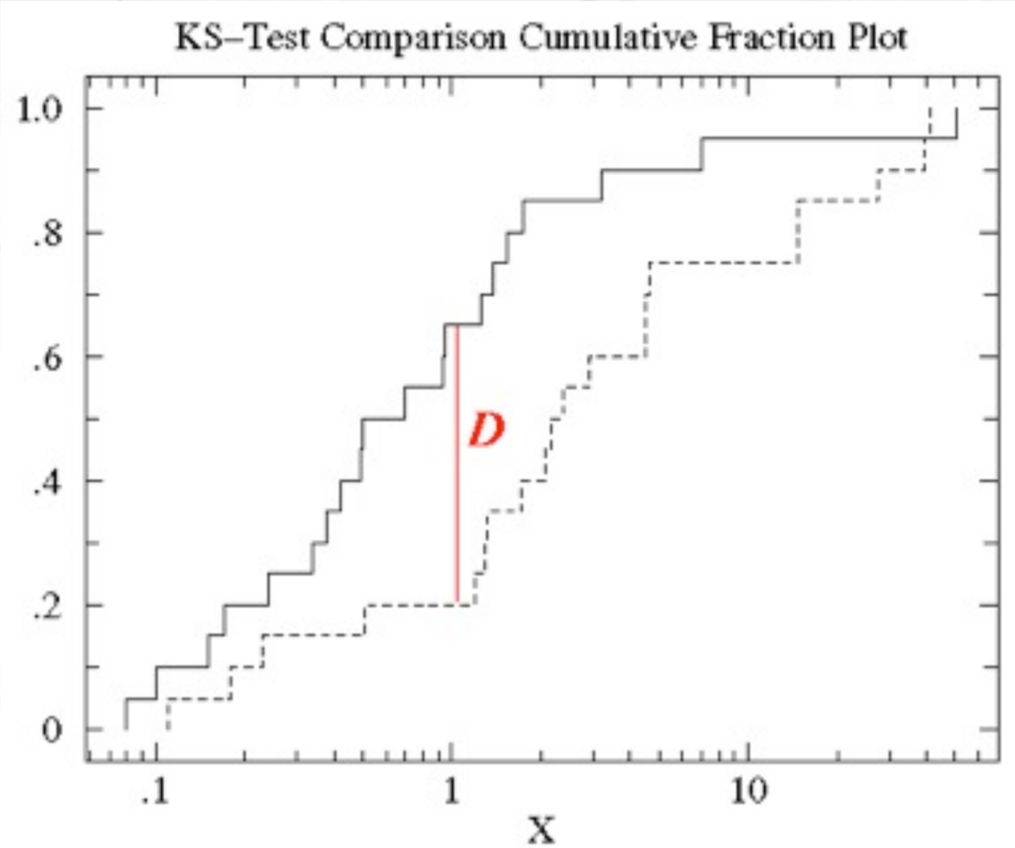
  Example: Model fitted to (possibly binned) data, yielding p-value = Prob($\chi^2 = 45.9$, $N_{dof} = 36$) = 0.125

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

Kolmogorov-Smirnov test

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

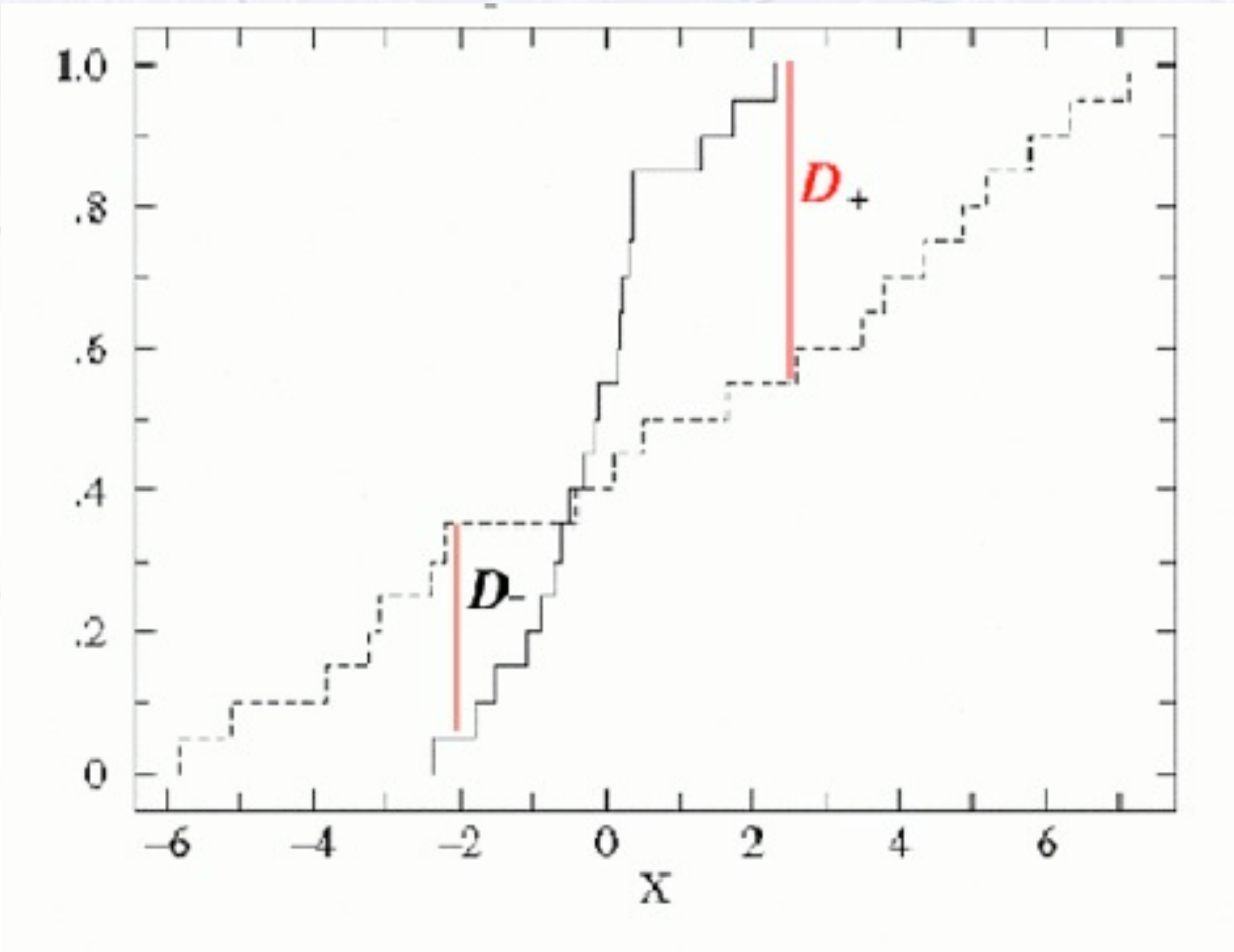
Example: Compatibily between function and sample or between two samples, yielding p-value = 0.87



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

Kuiper test

Is a similar test, but it is more specialized 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: $z = \frac{\overline{x} - \mu_0}{\sigma}\sqrt{n}$

  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). $z = \frac{(\overline{x}_1 - \overline{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n1} + \frac{\sigma_2^2}{n2}}}$

  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: Compatibily 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.

These are the tests you should know about!

Those below are for general education.

Wald-Wolfowitz runs test

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:

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

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

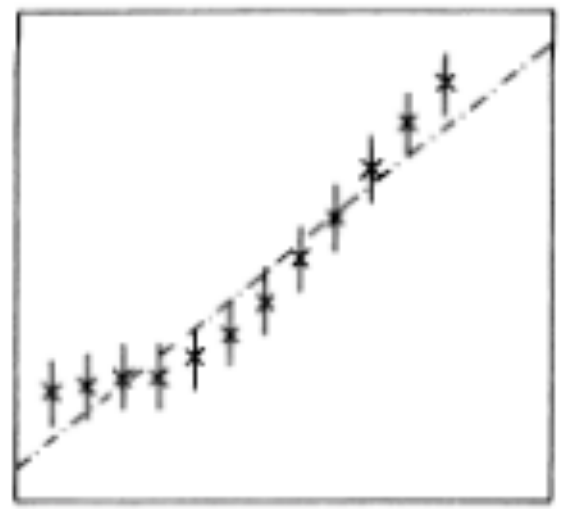


Fig. 8.3. A straight line through twelve data points.

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

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

Simple way to test categorial data (though 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?

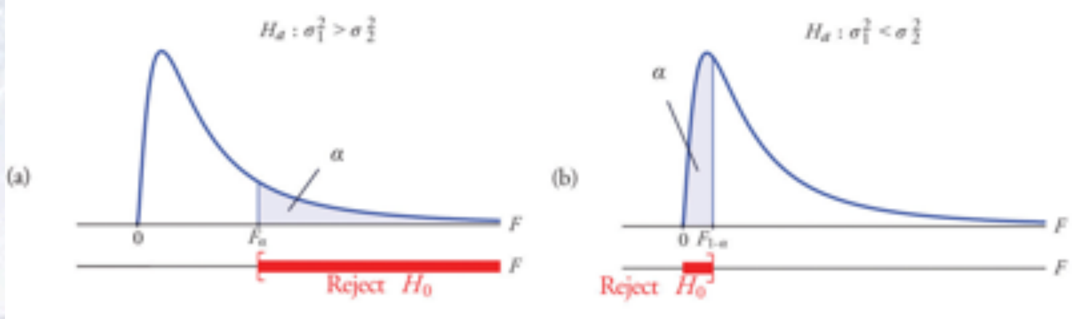
The Chi-square test is not optimal, as there are (several) entries, that are very low (< 5), but Fisher's exact test gives the answer:

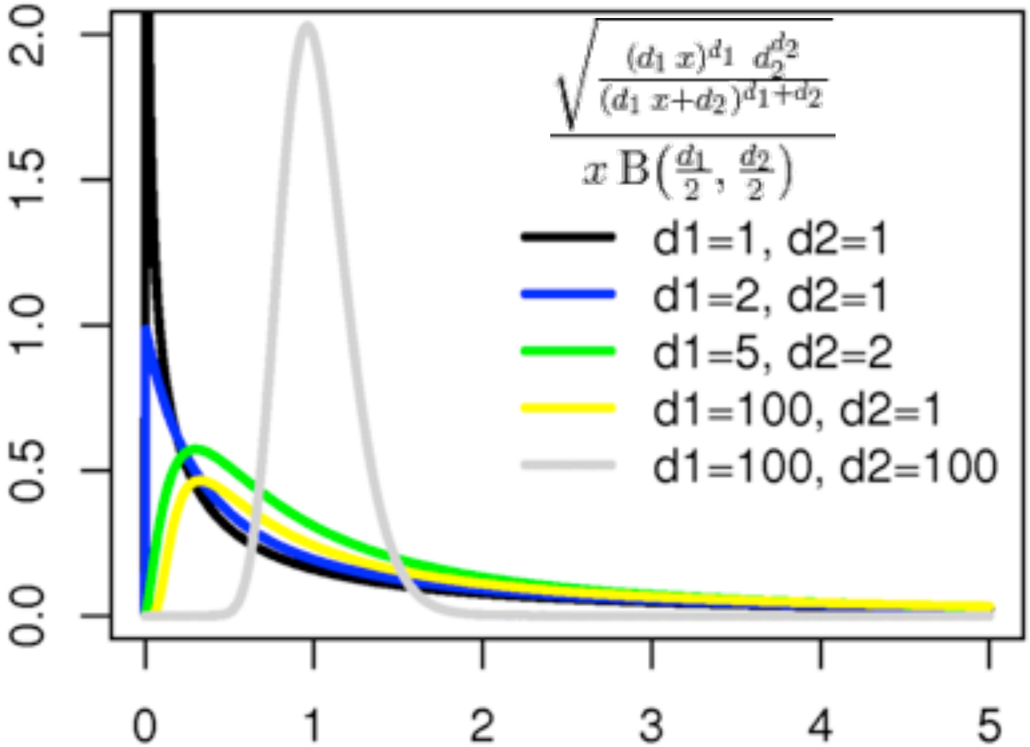
$$p = \binom{10}{1}\binom{14}{11} / \binom{24}{12} = \frac{10!\ 14!\ 12!\ 12!}{1!\ 9!\ 11!\ 3!\ 24!} \simeq 0.00135$$

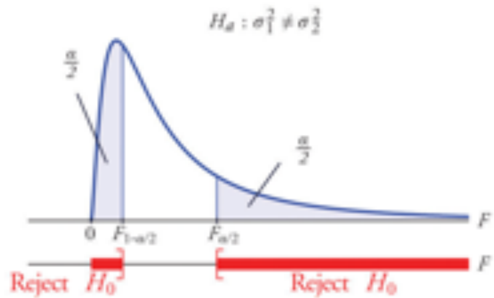
F-test

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.