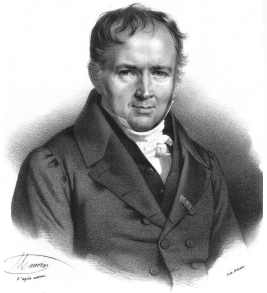
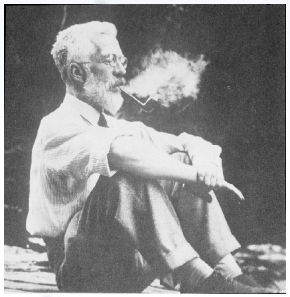
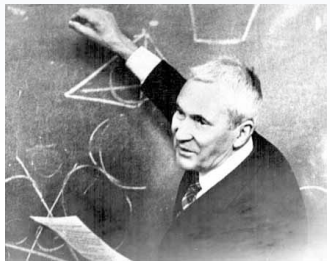
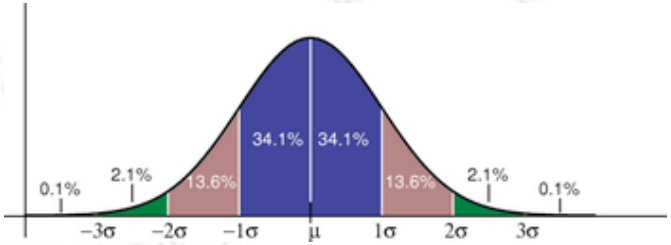
# Applied Statistics

## On p-values



Troels C. Petersen (NBI)




*"Statistics is merely a quantisation of common sense"*

# p-values

A p-value is:

*"the probability of the observed or something more extreme, assuming that the (null) hypothesis model is correct!".*

# p-values

A p-value is:

*"the probability of the observed or something more extreme, assuming that the (null) hypothesis model is correct!"*.

Why this definition?

# p-values

A p-value is:

*"the probability of the observed or something more extreme, assuming that the (null) hypothesis model is correct!"*.

Why this definition?

We want to quantify the degree to which a model matches the data. This can for the ChiSquare be done, since we know the expected ChiSquare distribution, *but only if we have the correct hypothesis that matches the data*.

Since data fluctuates, we should NOT expect a fixed p-value, and occasionally the fluctuations are so significant that we get a low p-value. What is the chance of this or something more extreme? Well… the p-value.

**Note that p-values are not only the result of a ChiSquare test, but ALL tests!**

# p-values

A p-value is:

*"the probability of the observed or something more extreme, assuming that the (null) hypothesis model is correct!"*.

Why this definition?

We want to quantify the degree to which a model matches the data. This can for the ChiSquare be done, since we know the expected ChiSquare distribution, *but only if we have the correct hypothesis that matches the data*.

Since data fluctuates, we should NOT expect a fixed p-value, and occasionally the fluctuations are so significant that we get a low p-value. What is the chance of this or something more extreme? Well… the p-value.

**Note that p-values are not only the result of a ChiSquare test, but ALL tests!**

What p-values are not:

"p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone" and that "a p-value, or statistical significance, does not measure the size of an effect or the importance of a result" [American Statistical Association, 2016]

# Testing p-value calculation

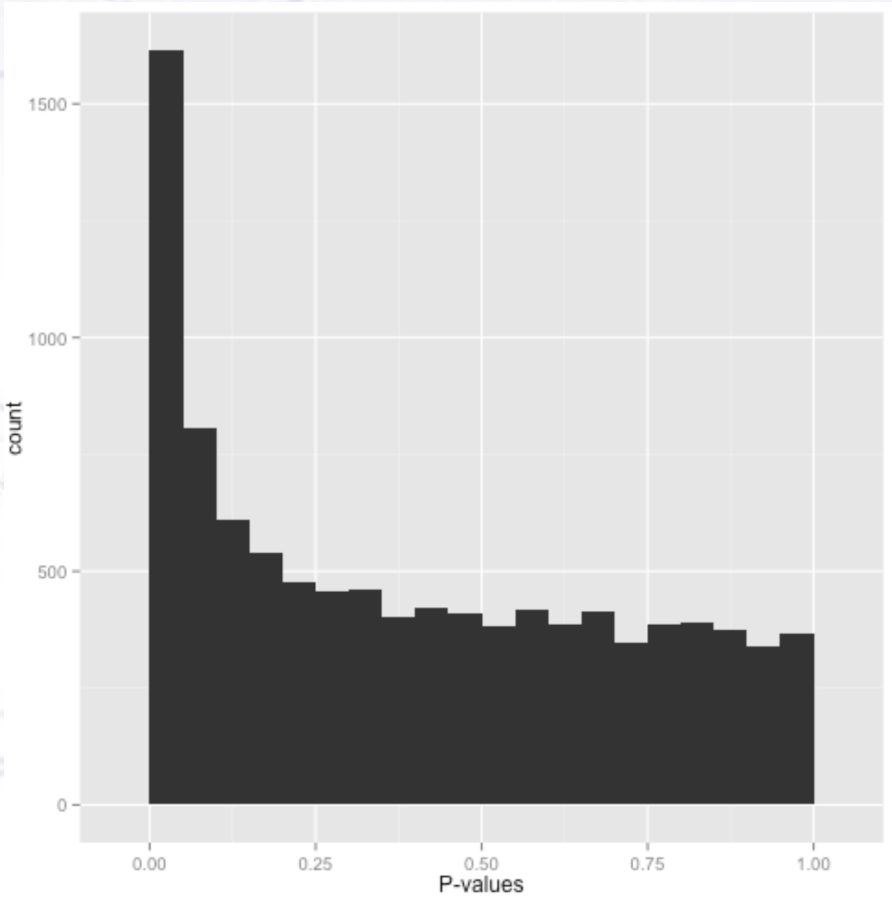
# Peak towards 0

The “null hypothesis case” is **by construction flat**. It may be what you expect or aim for (see research case), and is in all cases a good cross check of the hypothesis test you are doing/repeating.

The next case is possibly a good case.

Most of your tests come out uniformly distributed, but some seem to suggest the alternative hypothesis.

If that is of use of not is up to you, but this is a typical pattern to see.



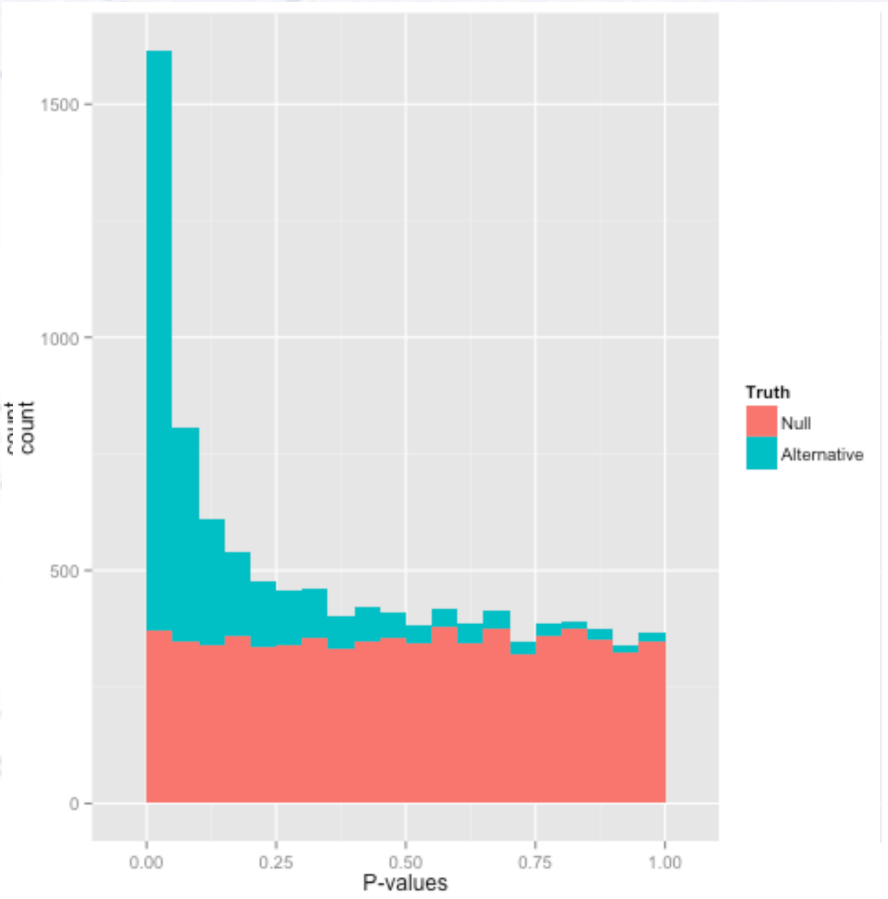
# Peak towards 0

The "null hypothesis case" is **by construction flat**. It may be what you expect or aim for (see research case), and is in all cases a good cross check of the hypothesis test you are doing/repeating.

The next case is possibly a good case.

Most of your tests come out uniformly distributed, but some seem to suggest the alternative hypothesis.

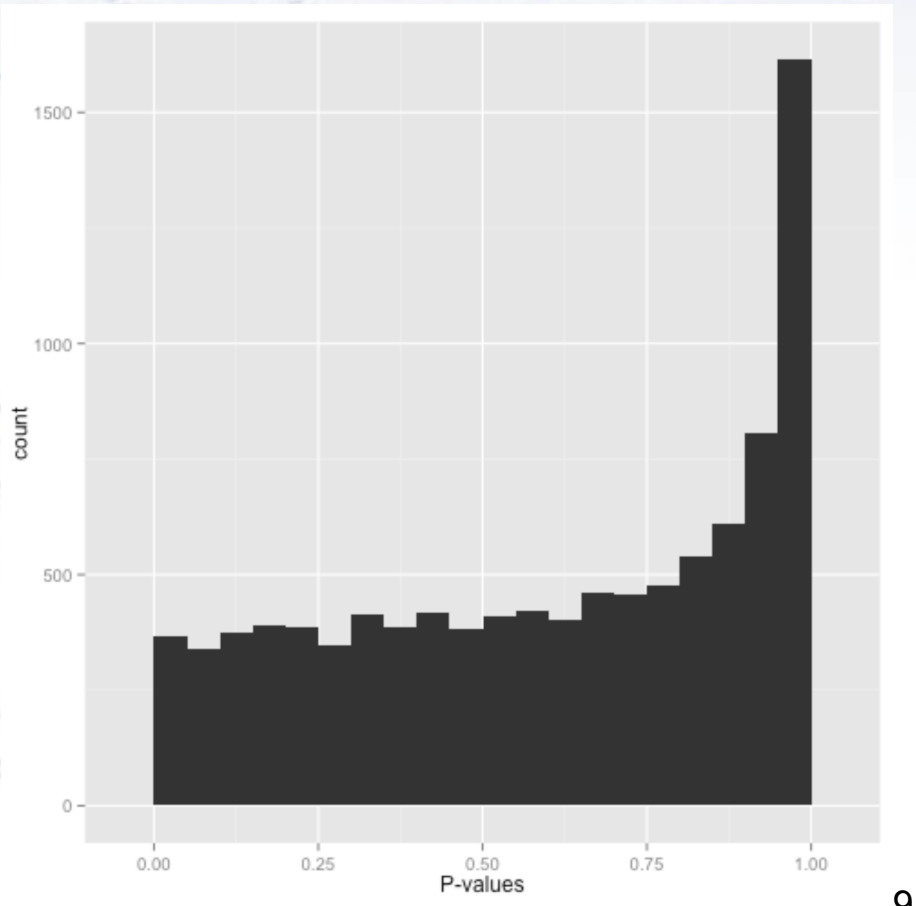
If that is of use of not is up to you, but this is a typical pattern to see.



# Peak towards 1

If the distribution has a peak towards 1 (conservative case), then your test is somehow “too loose”. It simply gives too high p-values in general, suggesting that the null hypothesis is good in more cases, than it should.

This could typically be due to overestimated uncertainties, if you happen to use the ChiSquare.



# Bimodal (peaks at 0 and 1)

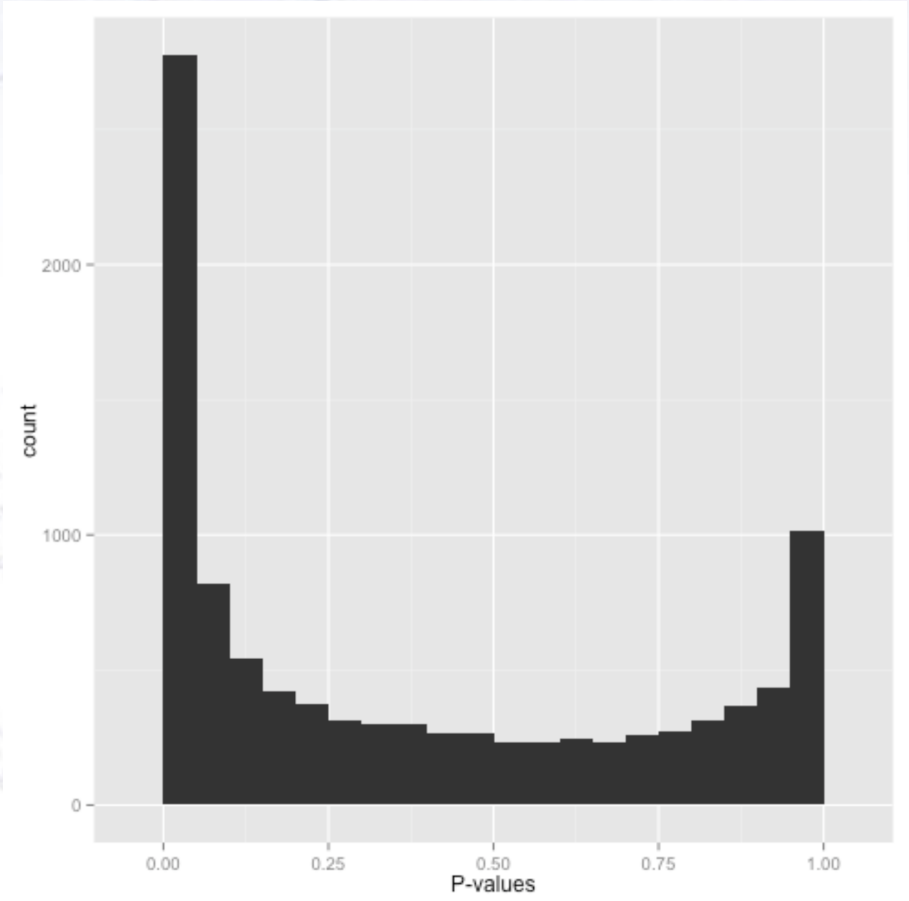
If the distribution is bimodal, i.e. has a peak at both high and low values, that indicates that the test is probably not well calibrated, i.e. there are variations that you might not know of or model well.

An example could be the ChiSquare, where you use fixed uncertainties. But if these in reality vary, the p-values will be either too large or too small.

It might also have to do with testing a one-sided vs. a two-sided hypothesis.

Or you might have a pathological case, which always yields one (high/low) value.

… or just a bug :-)



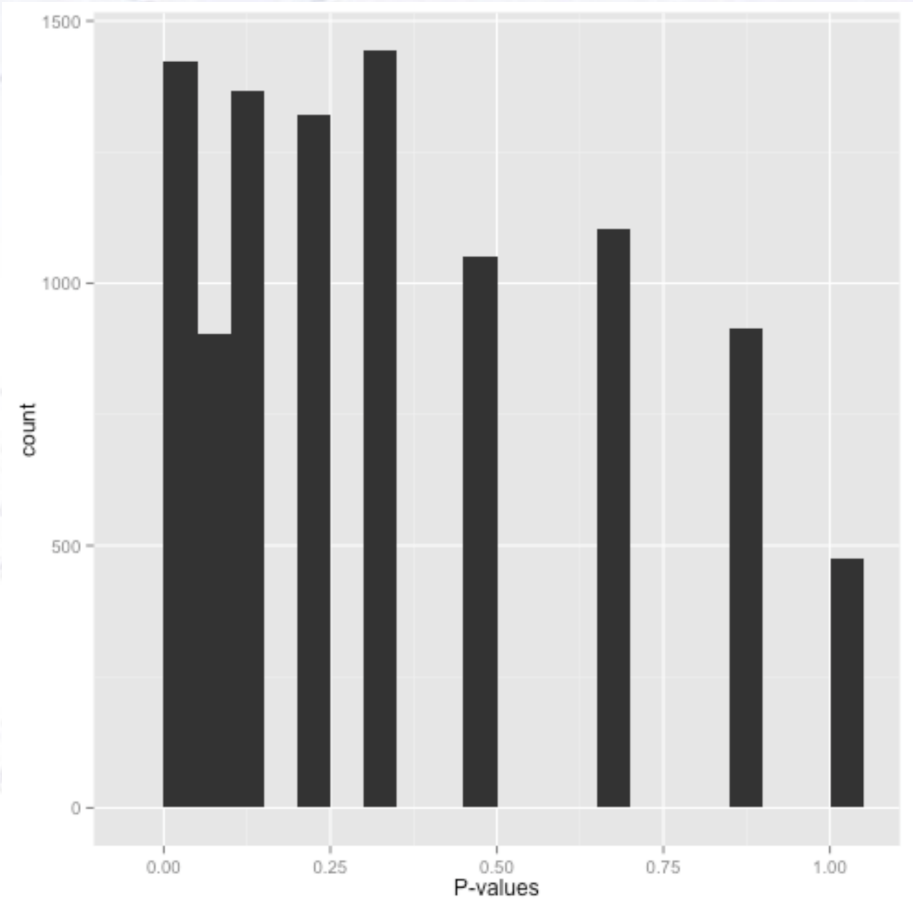
# Sparse

Sometimes the p-values are sparse - i.e. only a few values are represented, even if you have statistics for filling out the whole thing.

This happens, when the number of outcomes of a test is limited!

A typical case is the comparison of two low(ish) statistics histograms with the Kolmogorov- Smirnoff test.
The difference (D) might only be able to take very few values (say 0-9), and though the test gives continuous results, it can still only give 10 distinct values.

PS. To test the number of unique entries in an array, use the "set" command!

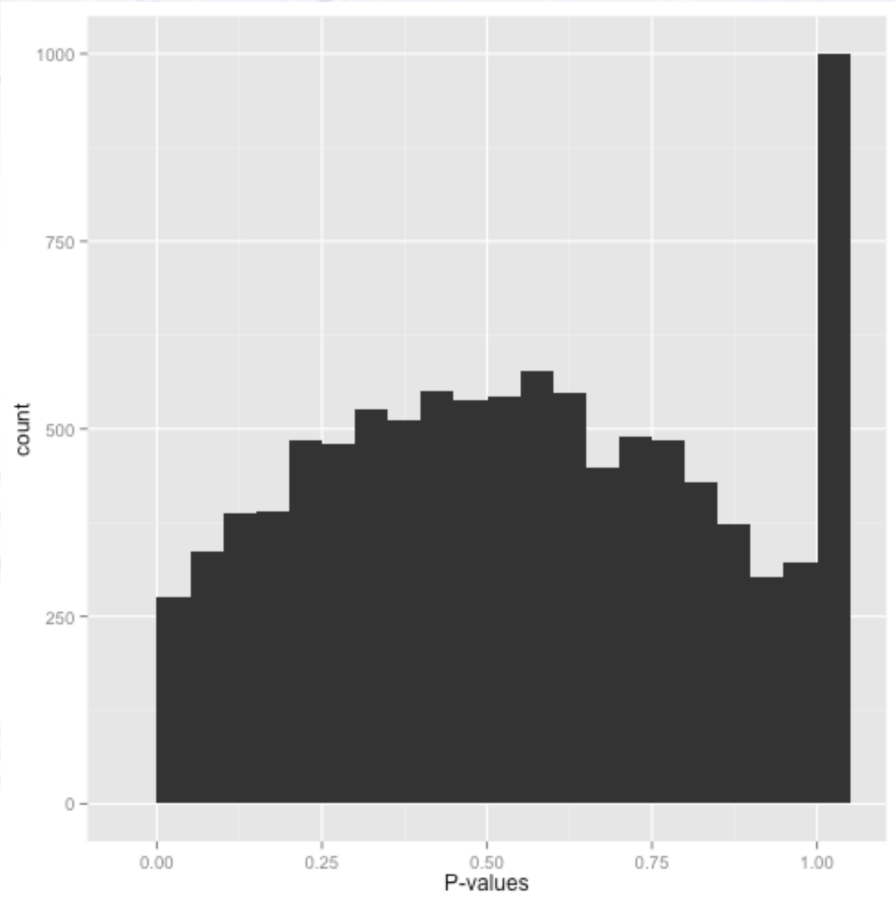


# Weird

If you have something "weird" that curves the opposite way of what you would expect (see previous slides), and/or has strange peaks in it, then there is surely something wrong with your hypothesis test!!!

Stop what you're doing and go back to check what went wrong.

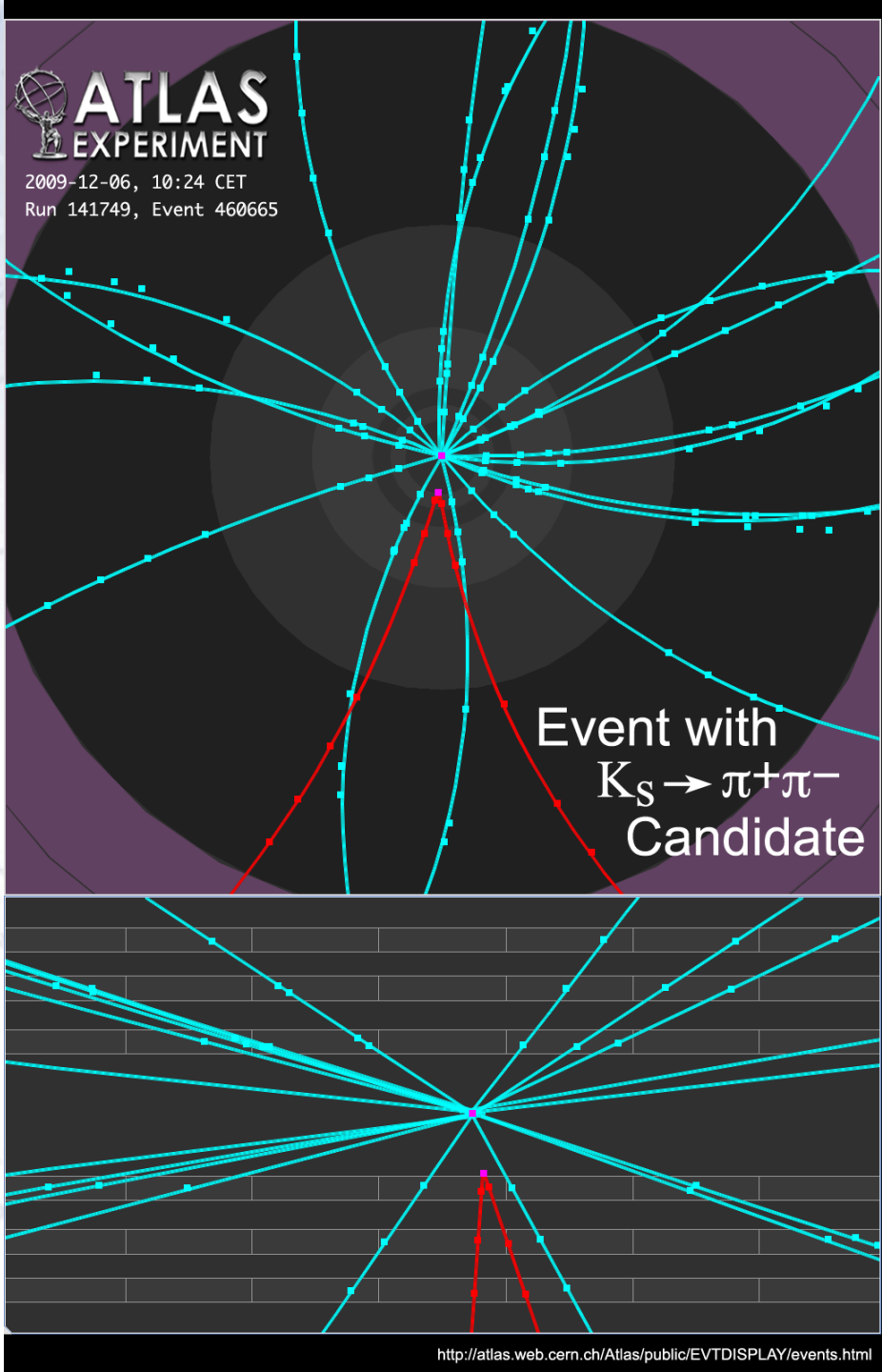
And praise yourself for plotting the p-values, so that you didn't continue far with them!



# Research example

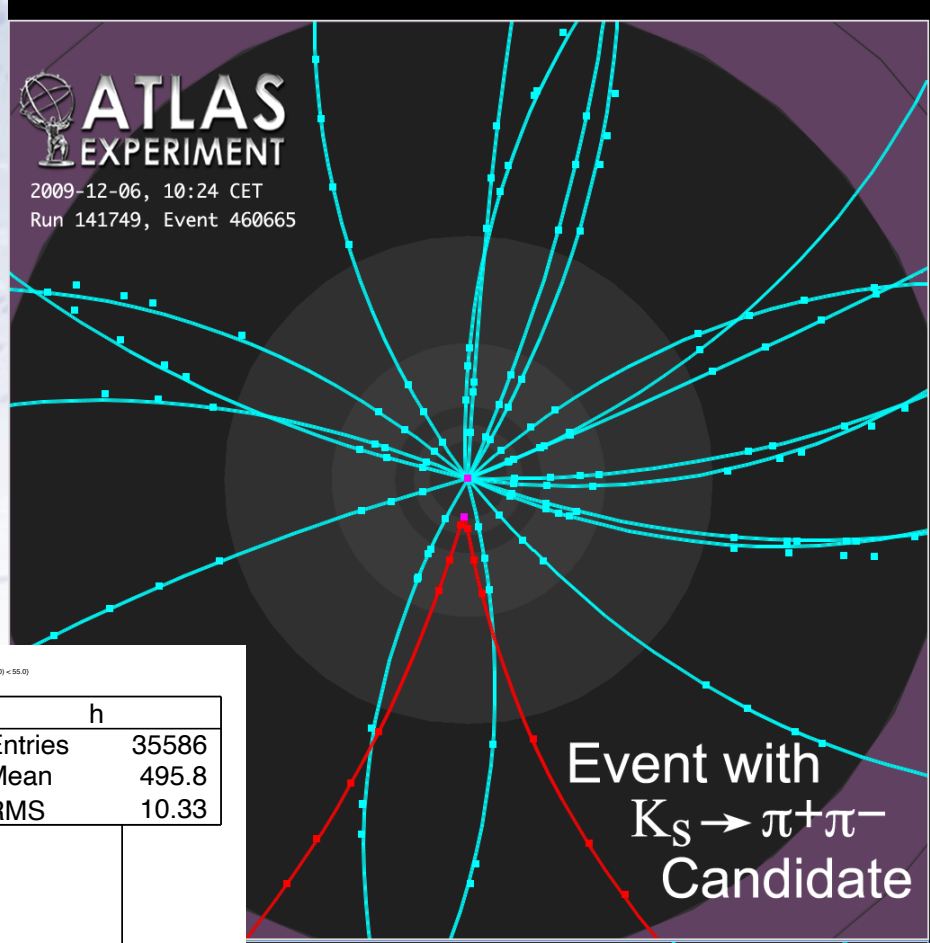
# An example…

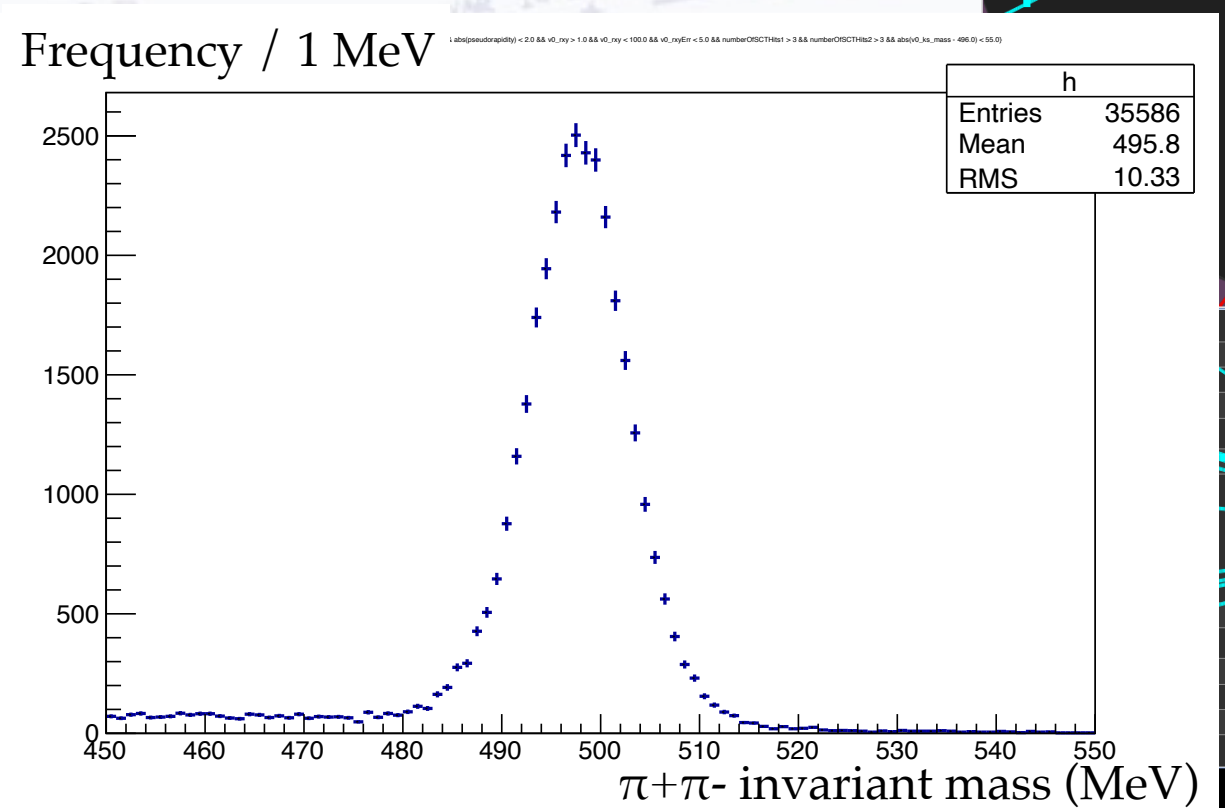
In particle physics, the $K^0$-short particle flies a little bit before it decays to two pions. Their tracks are measured (with full covariant uncertainty), and are then paired in a ChiSquare fit to check for a common vertex.

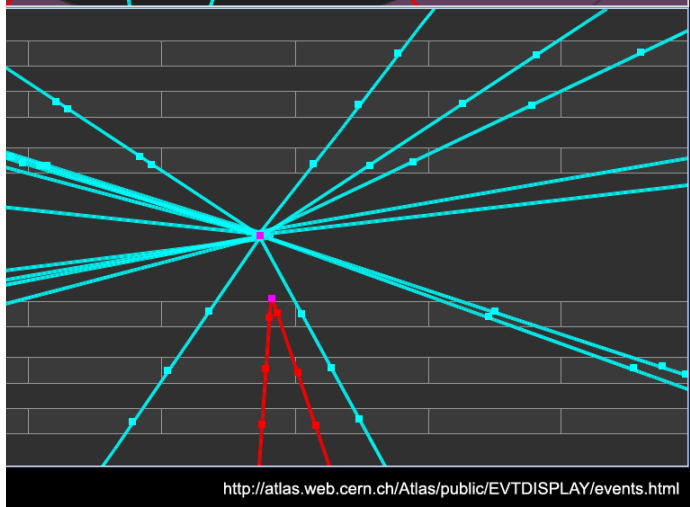


# An example…

In particle physics, the $K^0$-short particle flies a little bit before it decays to two pions. Their tracks are measured (with full covariant uncertainty), and are then paired in a ChiSquare fit to check for a common vertex.



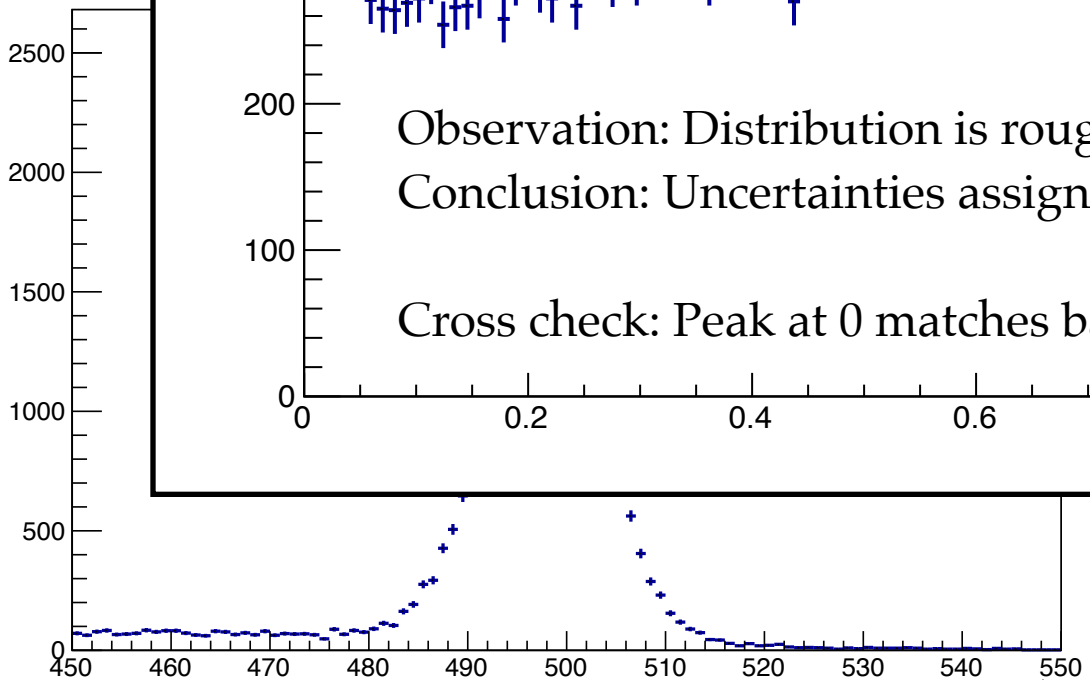



# An example…

In pa
flies a
pions
full c
paire
comm

Frequen







Observation: Distribution is roughly flat.

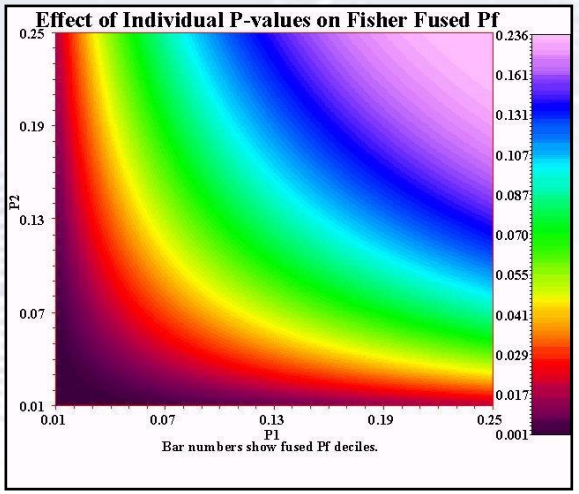
Conclusion: Uncertainties assigned are correct.

Cross check: Peak at 0 matches background.

# Combining p-values

# Multiple p-values

Your data might warrant repeated (independent) hypothesis testing of the same hypothesis, yielding multiple p-values.

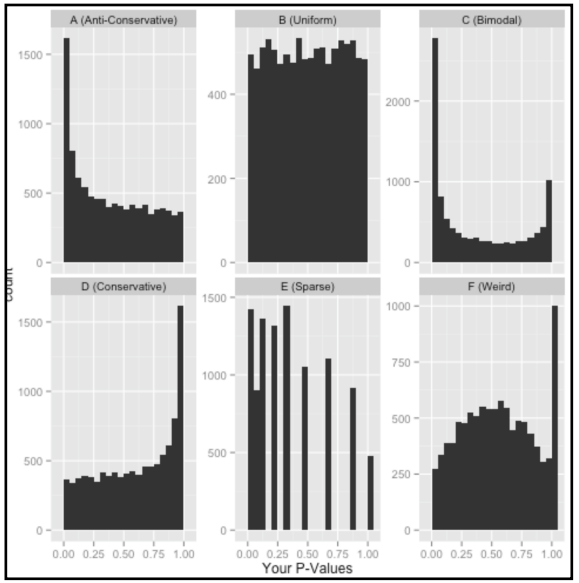



<u>There are two possible things to do:</u>

1) Combine the p-values (Fisher's method)
2) Plot the p-values (if numerous, i.e. > 100)

The **first** allows you to combine the results into one combined p-value (see next slide).

The **second** gives you a way to inspect the behaviour of your tests, and to immediately diagnose some potential problems.
Inspired by "VARIANCE EXPLAINED: How to interpret a p-value histogram", six examples were discussed (The Good, The Bad, and The Ugly!).



# Fisher's method

If you have two or more p-values from **independent hypothesis tests** ($p_i$), then these may be combined into one single p-value, by producing the following sum:

$$\chi^2_{2k} \sim \sum_{i=1}^{k} \ln(p_i)$$

When the null hypothesis are all true and all $p_i$ are independent, then this new Chi2(2k) **follows a ChiSquare distribution with 2k degrees of freedom**.

When the p-values tend to be small, the test statistic Chi2(2k) above will be large, which suggests that the null hypotheses are not true for every test.

Note that if the tests are NOT independent (typically positive correlation), then this test gives a too small p-value (anti-conservative).

# Conclusions

Whichever hypothesis test you're conducting, it is always **very healthy** to test that it gives a flat distribution on the null hypothesis case.

If it doesn't, then something needs to be understood. It might be range from a subtle statistical point to an obvious bug, but it needs to be investigated.

If it does yield a flat distribution, then that is an excellent control plot to include in your argument/slides/thesis/paper, or at least to refer to.

Possibly think ahead of time, how you can produce such a cross check.

And the plot from ATLAS on the right is the result of many iterations!