Tools for Physicists: Statistics

Confidence intervals, hypothesis tests

Wolfgang Gradl

Institut für Kernphysik

Summer semester 2022

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Confidence intervals

In 2006: $M_{\text{top}} = 174.3 \pm 5.1\,\text{GeV}/c^2$

What does this mean?

- 68% of top quarks have masses between 169.2 and 179.4 GeV/c^2
  WRONG: all top quarks have same mass!

In 2006: $M_{\text{top}} = 174.3 \pm 5.1\,\text{GeV}/c^2$

What does this mean?

- 68% of top quarks have masses between 169.2 and 179.4 GeV/c^2
  WRONG: all top quarks have same mass!

- The probability of M_{top} being in the range 169.2 – 179.4 GeV/c^2 is 68%
  WRONG: M_{top} is what it is, it is either in or outside this range. P is 0 or 1.

In 2006: $M_{\text{top}} = 174.3 \pm 5.1\,\text{GeV}/c^2$

What does this mean?

- 68% of top quarks have masses between 169.2 and 179.4 GeV/c^2
  WRONG: all top quarks have same mass!

- The probability of M_{top} being in the range 169.2 – 179.4 GeV/c^2 is 68%
  WRONG: M_{top} is what it is, it is either in or outside this range. P is 0 or 1.

- M_{top} has been measured to be 174.3 GeV/c^2 using a technique which has a 68% probability of being within 5.1 GeV/c^2 of the true result
  RIGHT

In 2006: $M_{\text{top}} = 174.3 \pm 5.1\,\text{GeV}/c^2$

What does this mean?

- 68% of top quarks have masses between 169.2 and 179.4 GeV/c^2
  WRONG: all top quarks have same mass!

- The probability of M_{top} being in the range 169.2 – 179.4 GeV/c^2 is 68%
  WRONG: M_{top} is what it is, it is either in or outside this range. P is 0 or 1.

- M_{top} has been measured to be 174.3 GeV/c^2 using a technique which has a 68% probability of being within 5.1 GeV/c^2 of the true result
  RIGHT
  if we repeated the measurement many times, we would obtain many different intervals; they would bracket the true M_{top} in 68% of all cases

Point estimates, limits

Often reported: point estimate and its standard deviation, $\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}$.

In some situations, an interval is reported instead, e.g. when
p.d.f. of the estimator is non-Gaussian, or
there are physical boundaries on the possible values of the parameter

Goals:

- communicate as objectively as possible the result of the experiment
- provide an interval that is constructed to cover the true value of the parameter with a specified probability
- provide information needed to draw conclusions about the parameter or to make a particular decision
- draw conclusions about parameter that incorporate stated prior beliefs

With sufficiently large data sample, point estimate and standard deviation essentially satisfy all these goals.

Choices, choices!

We can choose:

- The confidence level
  two-sided confidence intervals: typically 68%, corresponding to $\pm 1\sigma$
  upper (or lower) limits: frequently 90%, but 95% not uncommon …
- Whether to quote an upper limit or a two-sided confidence interval
- What sort of two-sided limit
  central (i.e. symmetric), shortest, …

Important: document what you are doing!

Constrained parameters

Measure a mass

$$M_X = -2 \pm 5\,\text{GeV}$$

or even

$$M_X = -5 \pm 2\,\text{GeV}$$

'M_X lies between -7 and -3' with 68% confidence

???

Counting experiment

Expect 2.8 background events

See 0 events; so, 90% CL upper limit is 2.3 events

so, signal < -0.5 events

???

What's happened?

Two views:

Nothing has gone wrong

(Up to) 10% of our 90% CL statements can be wrong; this is just one of them

Publish this, to avoid bias!

Everything wrong!

There are physical constraints (masses are non-negative, so are cross sections!)

No way to input this into the statistical apparatus

We will not publish results that are manifestly wrong

This is broken and needs fixing

What should be done with ‘unphysical’ results?

Best, but mostly not possible: publish full likelihood (or log-likelihood) function. This allows optimal combination of results, but is rarely done.

Preferred solution: publish both solutions,
i.e. the ‘raw’, maybe nonsensical two-sided confidence interval,
and one-sided C.I. taking extra constraints into account

May have to fight against (internal and external) referees who insist that publishing a two-sided confidence interval is equivalent to claiming “observation”

Estimation of confidence intervals

Typically, use **fit** to determine event yields or parameters of a distribution

Least square fit (for binned datasets) or maximum likelihood fits (can also deal with unbinned data)

Error definition, for one degree of freedom:

LSQ : 1σ confidence interval from $S = S_{\text{min}} + 1$

ML : 1σ confidence interval from $\log \mathcal{L} = \log \mathcal{L}_{\text{max}} - \frac{1}{2}$

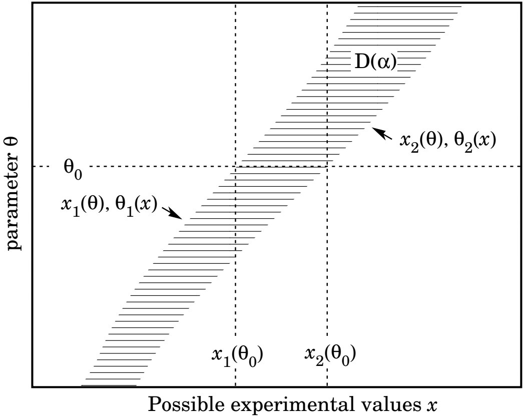
$n\sigma$ conf. intervals from $2\Delta \log \mathcal{L} = n^2$

See today's practical part what happens for joint confidence region for ν parameters

Construction of frequentist confidence intervals

Neyman construction of ‘confidence belts’:

for a given value of parameter θ, find interval of possible measured values x such that $[x_1, x_2]$ is a *CL* confidence interval:



D(α)
$x_2(\theta)$, $\theta_2(x)$
θ_0
$x_1(\theta)$, $\theta_1(x)$
parameter θ
$x_1(\theta_0)$
$x_2(\theta_0)$
Possible experimental values x

Bayesian credible intervals

Bayesian approach: report full posterior p.d.f.

If a range is desired: integrate posterior p.d.f. $p(\theta|x)$

$$1 - \alpha = \int_{\theta_{\mathrm{lo}}}^{\theta_{\mathrm{up}}} p(\theta|x) \mathrm{d}\theta$$

e.g. $1 - \alpha = 0.9$: “90% credible interval”

Several choices possible to construct $[\theta_{\mathrm{lo}}, \theta_{\mathrm{up}}]$:

- $[-\infty; \theta_{\mathrm{lo}}]$ and $[\theta_{\mathrm{up}}; \infty]$ both correspond to probability $\alpha/2$
- Symmetric interval around maximum value of p, corresponding to probability $1 - \alpha$
- $p(\theta|x)$ higher than any θ not belonging to the set
- …

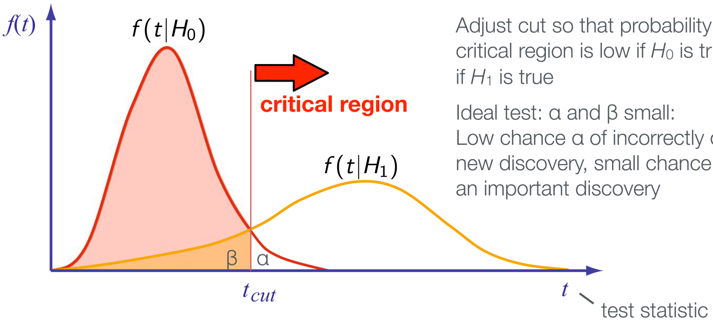
Hypothesis tests

Hypotheses and tests

- Hypothesis test
  - Goal: draw conclusions from the data
  - Statement about validity of a model
  - Decide which of two competing models is more consistent with data
- Simple hypothesis: no free parameters
  - Examples: particle is a π; data follow Poissonian with mean 5
- Composite hypothesis: contains free parameters
- Null hypothesis H_0 and alternative hypothesis H_1
  - H_0 often the background-only hypothesis
    (e.g. Standard Model only; no additional resonance; ...)
  - H_1 often signal or signal+background hypothesis
- Question: can H_0 be rejected by data?
- Test statistic t: (scalar) variable that is a function of the data alone, that can be used to test hypothesis

Critical region

Reject null hypothesis if value of t lies in critical region: $t > t_{\text{cut}}$



Adjust cut so that probability to be in critical region is low if H_0 is true and high if H_1 is true

Ideal test: α and β small:
Low chance α of incorrectly claiming a new discovery, small chance β of missing an important discovery

Probability for H_0 to be rejected while H_0 is true:

$$\int_{t_{\text{cut}}}^{\infty} f(t|H_0)\mathrm{d}t = \alpha$$

α: "size" or significance level of test

Probability for H_1 to be rejected even though it is true:

$$\int_{-\infty}^{t_{\text{cut}}} f(t|H_1)\mathrm{d}t = \beta$$

$1-\beta$: power of the test

Type I and Type II errors

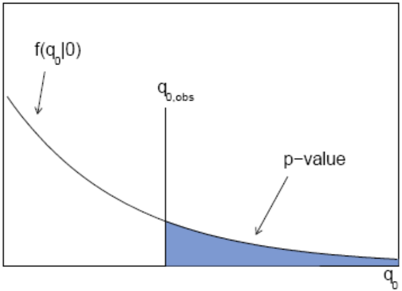
Statistics jargon, getting more and more common also in HEP

Type I error: Probability of rejecting null hypothesis H_0 when it is actually true
also known as false discovery rate

Type II error: Probability to fail to reject null hypothesis H_0 while it is actually false
also known as false exclusion rate

p-value

p-value: probability to observe data set that is as consistent or worse with null hypothesis as the actual observation



test statistic: q_0

pdf for q_0 under H_0: $f(q_0|0)$

critical region: large values of q_0

$q_{0,\mathrm{obs}}$: observed value in data

$$p_0 = \int_{q_{0,\mathrm{obs}}}^{\infty} f(q_0|0)\mathrm{d}q_0$$

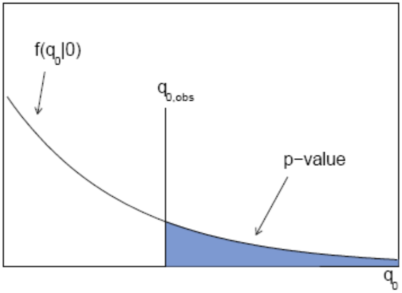
pdf for q_0 under H_0 frequently needs to be estimated with simulation

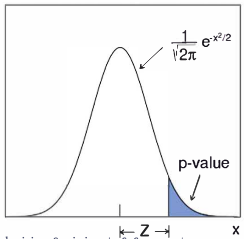
p-value is a random variable (contrast: significance level α fixed before measurement).

if $p_0 < \alpha$: reject H_0

$1 - p_0$: confidence level of test

p-value and significance





(a)

if $p_0 < \alpha$, then reject null hypothesis

Frequent convention in HEP:

for discovery, require $p < 2.87 \times 10^{-7}$

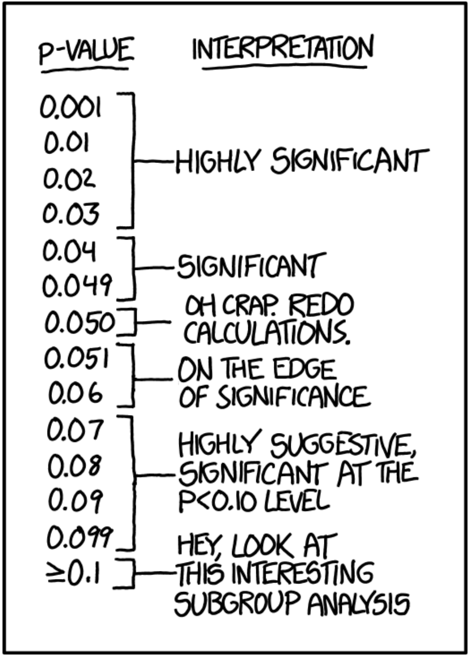
for exclusion, require $p < 0.05$

translate p-value to significance Z via Standard Normal pdf

$$p_0 = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1 - \Phi(Z)$$

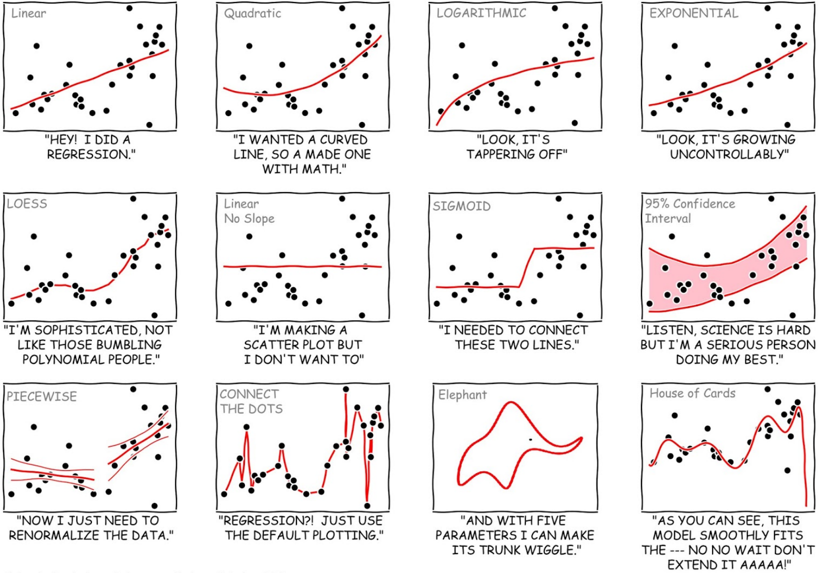
$$Z = \Phi^{-1}(1 - p_0)$$

Significance of 5 (1.64) s.d. corresponds to $p = 2.87 \times 10^{-7} (0.05)$


P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

how can we objectively tell which model fits better?

CURVE-FITTING METHODS AND THE MESSAGES THEY SEND



Linear
"HEY! I DID A REGRESSION."
Quadratic
"I WANTED A CURVED LINE, SO A MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPPERING OFF"
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
Linear No Slope
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO"
SIGMOID
"I NEEDED TO CONNECT THESE TWO LINES."
95% Confidence Interval
"LISTEN, SCIENCE IS HARD BUT I'M A SERIOUS PERSON DOING MY BEST."
PIECEWISE
"NOW I JUST NEED TO RENORMALIZE THE DATA."
CONNECT THE DOTS
"REGRESSION?! JUST USE THE DEFAULT PLOTTING."
Elephant
"AND WITH FIVE PARAMETERS I CAN MAKE ITS TRUNK WIGGLE."
House of Cards
"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE --- NO NO WAIT DON'T EXTEND IT AAAAA!"

by Douglas Higinbotham in Python inspired by https://xkcd.com/2048

Least squares: Goodness-of-fit

Minimum value of S in the least squares method is a measure of agreement between model and data:

$$S_{\text{min}} = \sum_{i=1}^{n} \left(\frac{y_i - f(x_i; \widehat{\vec{\theta}})}{\sigma_i} \right)^2$$

Large value of S_{min}: can reject model.

If model is correct, then S_{min} for repeated experiments follows a χ^2 distribution with n_{df} degrees of freedom:

$$f(t; n_{\text{df}}) = \frac{t^{n_{\text{df}}/2-1}}{2^{n_{\text{df}}/2}\Gamma\left(\frac{n_{\text{df}}}{2}\right)} e^{-t/2}, \quad t = \chi^2_{\text{min}}$$

with $n_{\text{df}} = n - m =$ number of data points $-$ number of fit parameters

Least squares: Goodness-of-fit

Expectation value of χ^2 distribution is n_{df}
➡ $\chi^2 \approx n_{\mathrm{df}}$ indicates good fit

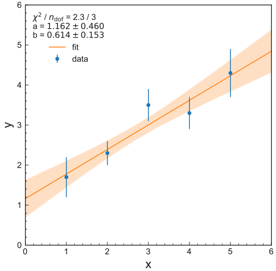
Consistency of a model with data is quantified with the p-value:

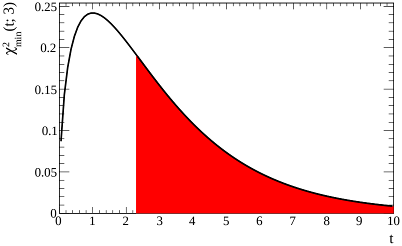
$$p = \int_{S_{\mathrm{min}}}^{+\infty} f(t; n_{\mathrm{df}})\mathrm{d}t$$

p-value: probability to get a χ^2_{min} at least as high as the observed one, if the model is correct.

p-value is **not** the probability that the model is correct!

p-value for the straight line fit example

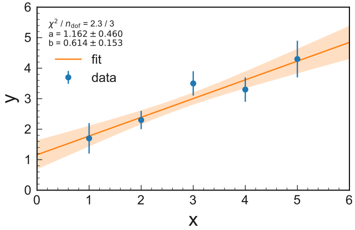




$S_{\mathrm{min}} = 2.29557$, $n_{\mathrm{df}} = 3$

p-value: $\mathrm{prob}(S_{\mathrm{min}}, n_{\mathrm{df}}) = 0.51337011$

p-value for the straight line fit example

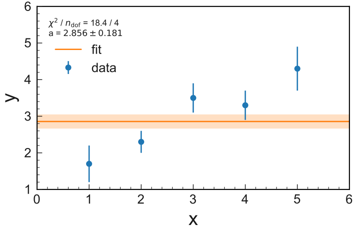


$$S_{\mathrm{min}} = 2.29557, \quad n_{\mathrm{df}} = 3$$

$$p\text{-value} = 0.5134$$

$$\hat{\theta}_0 = 1.16 \pm 0.46$$

$$\hat{\theta}_1 = 0.614 \pm 0.153$$



$$S_{\mathrm{min}} = 18.3964, \quad n_{\mathrm{df}} = 4$$

$$p\text{-value} = 0.00103$$

$$\hat{\theta}_0 = 2.856 \pm 0.181$$

Stat. uncertainty on fit parameter does not tell us whether model is correct

Side remark: quoting χ^2 and ndf

Always remember to quote χ^2 and n_{df} separately,
instead of just the 'reduced χ^2/n_{df} — there *is* a difference!

$$\text{prob}(15, 10) = 0.132$$

$$\text{prob}(1500, 1000) = 1.05 \times 10^{-22}$$

Goodness of fit for unbinned ML fits

In the case of unbinned ML fit, can bin data and model prediction into histogram and then perform χ^2 test

Consider the likelihood ratio

$$\lambda = \frac{\mathcal{L}(\vec{n}|\vec{\nu})}{\mathcal{L}(\vec{n}|\vec{n})}, \qquad \vec{\nu} = \vec{\nu}(\vec{\theta})$$

For multinomially (“M”, n_{tot} fixed) and Poisson distributed data (“P”), one obtains for k bins

$$\lambda_M = \prod_i^k \left(\frac{\nu_i}{n_i}\right)^{n_i}, \qquad \lambda_P = e^{n_{\text{tot}} - \nu_{\text{tot}}} \prod_i^k \left(\frac{\nu_i}{n_i}\right)^{n_i}$$

Now consider test statistic

$$t \equiv -2 \log \lambda$$

Goodness of fit for unbinned ML fits

For multinomially distributed data, in the large sample limit

$$t_M = -2\log\lambda_M = 2\sum_{i=1}^{k} n_i \log\frac{n_i}{\hat{\nu}_i}$$

follows χ^2 distribution for $k - m - 1$ degrees of freedom.

For Poisson distributed data,

$$t_P = -2\log\lambda_P = 2\sum_{i=1}^{k}\left(n_i \log\frac{n_i}{\hat{\nu}_i} + \hat{\nu}_i - n_i\right)$$

follows χ^2 distribution for $k - m$ degrees of freedom.

Profile likelihood ratio: hypothesis tests with nuisance parameters

Base significance test on the profile likelihood

$$\lambda(\mu) = \frac{\mathcal{L}(\mu, \hat{\hat{\theta}})}{\mathcal{L}(\hat{\mu}, \hat{\theta})} = \frac{\text{maximised } \mathcal{L} \text{ for specified } \mu}{\text{globally maximised } \mathcal{L}}$$

Likelihood ratio of point hypotheses gives optimum test (Neyman-Pearson lemma).

Composite hypothesis: parameter μ is only fixed under H_0, but not under H_1.

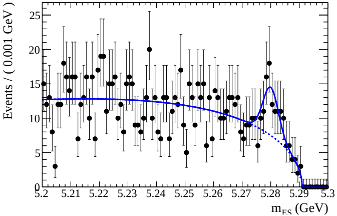
Wilks' theorem:

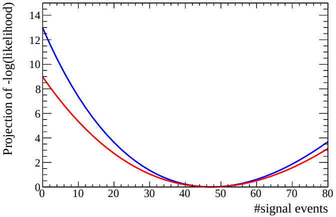
$$q_0 = -2 \log \lambda$$

asymptotically approaches chi-square distribution for k degrees of freedom, where k is the difference in dimensionality of H_1 and H_0

Profile likelihood ratio

Example: *B* mass fit from last time; 40 signal events, 1000 background events





3 parameters in the fit: signal and background yields, shape parameter for background

$$\hat{n}_{\text{sig}} = 47 \pm 12$$

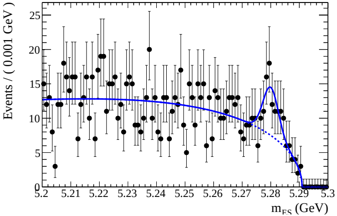
$$\hat{n}_{\text{bkg}} = 992 \pm 33$$

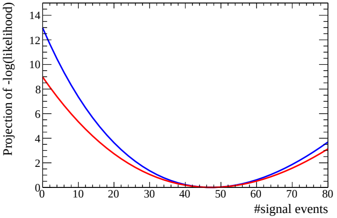
scan of $\mathcal{L}(n_{\text{sig}}, \hat{\theta})$ with nuisance parameters fixed to values from global minimum

profile likelihood: $\mathcal{L}(n_{\text{sig}}; \hat{\hat{\theta}})$

Profile likelihood ratio

Example: *B* mass fit from last time; 40 signal events, 1000 background events





3 parameters in the fit: signal and background yields, shape parameter for background

$$\hat{n}_{\text{sig}} = 47 \pm 12$$

$$\hat{n}_{\text{bkg}} = 992 \pm 33$$

From scan of profile likelihood:

$$2\Delta \log \mathcal{L} = 17.94$$

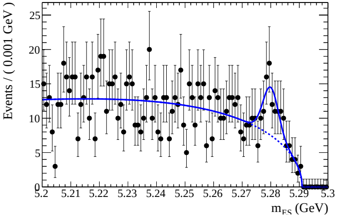
And therefore p-value for H_0: $1.139\,27 \times 10^{-5}$, or significance for $n_{\text{sig}} \neq 0$

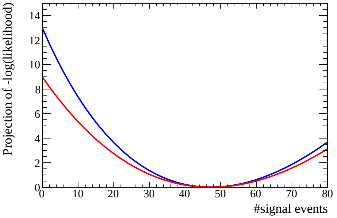
$$Z = \sqrt{2\Delta \log \mathcal{L}} = 4.2\sigma$$

(one degree of freedom!)

Profile likelihood ratio

Example: B mass fit from last time; 40 signal events, 1000 background events





3 parameters in the fit: signal and background yields, shape parameter for background

$$\hat{n}_{\text{sig}} = 47 \pm 12$$
$$\hat{n}_{\text{bkg}} = 992 \pm 33$$

now leave also mean and width of signal peak free in fit: two additional nuisance parameters (that cannot really be determined when $n_{\text{sig}} = 0$).

p-value = 0.0697557

$Z = 1.48\ \sigma$

Look-elsewhere effect

A Swedish study in 1992 tried to determine whether or not power lines caused some kind of poor health effects. The researchers surveyed everyone living within 300 meters of high-voltage power lines over a 25-year period and looked for statistically significant increases in rates of over 800 ailments. The study found that the incidence of childhood leukemia was four times higher among those that lived closest to the power lines, and it spurred calls to action by the Swedish government. The problem with the conclusion, however, was that they failed to compensate for the look-elsewhere effect; in any collection of 800 random samples, it is likely that at least one will be at least 3 standard deviations above the expected value, by chance alone. Subsequent studies failed to show any links between power lines and childhood leukemia, neither in causation nor even in correlation.

`https://en.wikipedia.org/wiki/Look-elsewhere_effect`

Look-elsewhere effect

In general, a p-value of $1/n$ is likely to occur after n tests.

Solution: apply ‘trials penalty’, or ‘trials factor’, *i.e.* make threshold more stringent for large n.

Not entirely trivial to choose trials factor: need to count effective number of ‘independent’ regions. Suppose you look at a range of invariant masses large compared to the mass resolution, then $N \sim \Delta M/\sigma_M$.

See e.g. Gross & Vitells, arXiv:1005.1891 [physics.data-an] for a recipe

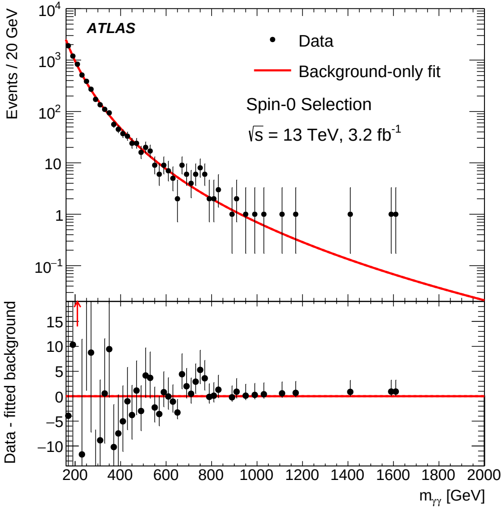
Look-elsewhere effect

Can make substantial change to claimed significance:

for example ATLAS observation of an enhancement around 750 GeV in $\gamma\gamma$ invariant mass:

Local significance 3.9σ, corresponding to a p-value of $p = 9.6 \times 10^{-5}$,
i.e. roughly 1:10000

Global significance only 2.1σ, corresponding to a p-value of $p = 0.0357$,
i.e. roughly 1:28



ATLAS, JHEP 09 (2016) 001

(Final) digression: p-value debate

In many fields (esp. social sciences, psychology, etc.), significant means $p < 0.05$

Relatively weak statistical standard, but often not realised as such!

We've seen that getting $p < 0.05$ isn't that rare, especially if you run many experiments!

May be a contributing factor to the 'reproducibility crisis'
and may be exacerbated by p-value hacking

5σ for discovery in particle physics?

5σ corresponds to p-value of 2.87×10^{-7} (one-sided test)

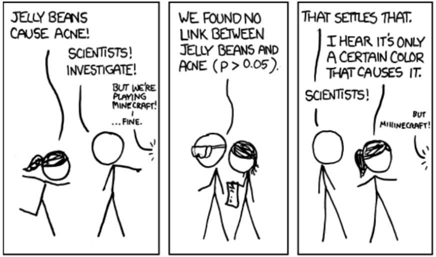
- History: many cases where 3σ and 4σ effects have disappeared with more data
- Look-elsewhere effect
- Systematics: often difficult to quantify / estimate
- Subconscious Bayes factor:
  - physicists tend to (subconsciously) assess Bayesian probabilities $p(H_1|\text{data})$ and $p(H_0|\text{data})$
  - If H_1 involves something very unexpected (e.g. superluminal neutrinos), then prior probability for H_0 is much larger than for H_1
  - Extraordinary claims require extraordinary evidence

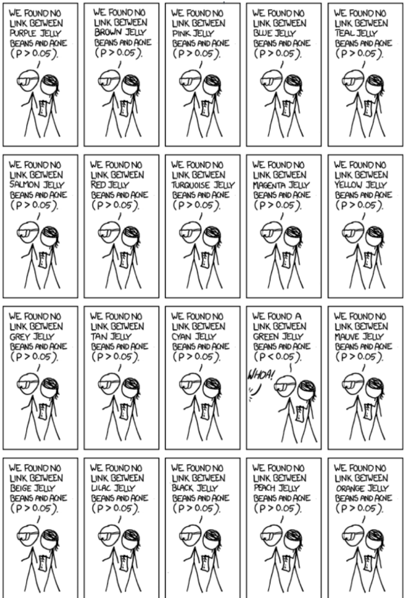
May be unreasonable to have single criterion for all experiments

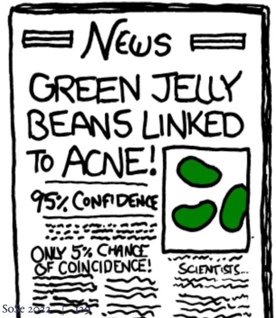
Louis Lyons, Statistical issues in searches for new physics, arXiv:1409.1903

p-value hacking

http://xkcd.com/822


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MINECRAFT!


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...