Statistical tests and hypothesis testing

Alexander Khanov

PHYS6260: Experimental Methods is HEP
Oklahoma State University

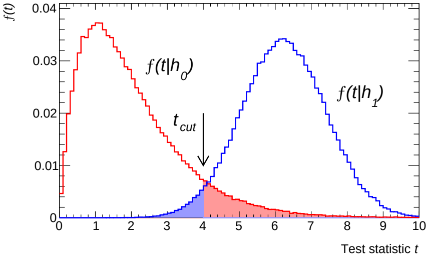
September 29, 2023

Statistics

- Statistic (singular) t: any function defined on a set of data $\mathbf{x} = \{x_1, \ldots, x_N\}$.
- Examples of statistics:
  - sample mean $\langle x \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i$
  - sample variance $\sigma^2 = \left\langle x^2 \right\rangle - \langle x \rangle^2$
  - likelihood $L(\mathbf{p}) = f(\mathbf{x}, \mathbf{p})$
- Since measurements $\mathbf{x}$ are random variables, t is also a random variable with its own p.d.f. $f(t)$

Hypothesis testing

- A typical problem: does the data reveal something interesting?
  - Null or background-only hypothesis h_0: not really (e.g. there is nothing but the Standard Model particles/processes)
  - Alternative or signal+background hypothesis h_1: indeed (e.g. there are some supersymetric particles flying around)
- Test statistic t: a statistic that can be used to test hypotheses
- Let $f(t|h0)$ and $f(t|h1)$ be the p.d.f. of statistic t under hypotheses h_0 and h_1, respectively
- Introduce the t cut value, $t_{\rm cut}$, to discriminate between the two hypotheses



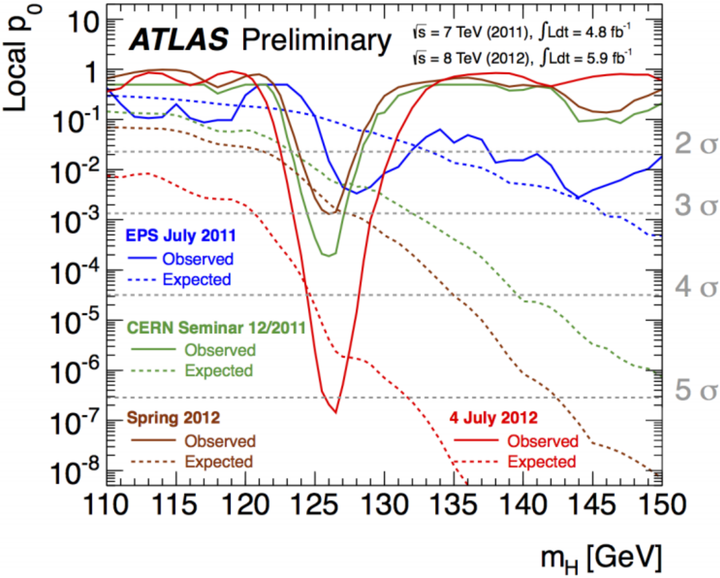
Hypothesis testing (2)

- $\alpha = \int_{t_{\text{cut}}}^{\infty} f(t|h0)\,dt$ is the probability to accept h_0 while it is true
  - α is called significance level of the test
- $\beta = \int_{-\infty}^{t_{\text{cut}}} f(t|h1)\,dt$ is the probability to reject h_1 while it is true
  - $1-\beta$ is called power of the test
- Neyman-Pearson lemma: the test that achieves the highest power for a given significance level is the likelihood ratio $t = \frac{L(\mathbf{x}|h_0)}{L(\mathbf{x}|h_1)}$
  - in practice it is more convenient to work with $\left(1+\frac{L(\mathbf{x}|h_0)}{L(\mathbf{x}|h_1)}\right)^{-1}$ since it is restricted to $(0,1)$

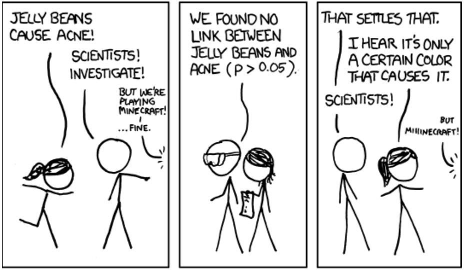
p-value

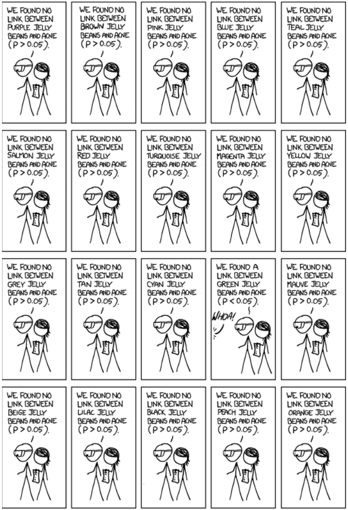
- If the value of test statistic t observed in data is t_{obs}, then
  $P = \int_{t_{\mathrm{obs}}}^{\infty} f(t|h0)\,dt$ is called p-value
- p-value is not significance level (which is a predefined number unrelated to data)
- p-value is not the probability that h_0 is true
  - frequentist: p-value is calculated for a particular hypothesis (h_0), discussing the probability of h_0 being true doesn't make sense
  - bayesian: the probability for h_0 to be true for a given data set is $P(h_0|D)$ while p-value is $P(D|h_0)$ (roughly speaking)
- In general, low p-value doesn't tell anything about the null hypothesis
- The concept of p-value is hated by many people
  - Nature 506 (2014) 150: P values, the "gold standard" of statistical validity, are not as reliable as many scientists assume
  - Nature 519 (2015) 9: Psychology journal bans P values

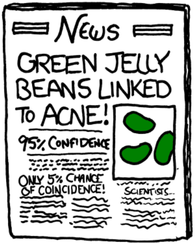
Higgs p-value



Look-elsewhere effect in a nut shell







https://xkcd.com/882

Look-elsewhere effect

- If one is performing multiple tests then a p-value of $1/n$ is likely to occur after n tests
- The local p-value: the probability for the background to fluctuate as much as the observed maximum excess
- The global p-value: the probability for the excess anywhere in the specific parameter range (e.g. mass range)
  - both are quoted e.g. for searches for new particles with unknown mass
- arXiv:2007.13821 [physics]: The look-elsewhere effect from a unified Bayesian and frequentist perspective

Bayes factor

- One alternative to p-value
- Consider two hypotheses h_0 and h_1 with prior probabilities $P(h_0)$ and $P(h_1) = 1 - P(h_0)$
- According to Bayes formula,

$$\frac{P(h_1|D)}{P(h_0|D)} = \frac{P(D|h_1)P(h_1)}{P(D|h_0)P(h_0)}$$

- The Bayes factor $B_{10} = \dfrac{P(D|h_1)}{P(D|h_0)}$
- Interpretation:
  - $B_{10} = 1 - 3$: irrelevant
  - $B_{10} = 3 - 20$: positive evidence
  - $B_{10} = 20 - 150$: strong evidence
  - $B_{10} > 150$: very strong evidence

J. Am. Stat. Assoc. 90 (1995) 773

Pearson's χ^2 statistic

- Suppose we have N measurements $\mathbf{n} = n_1, \ldots, n_N$ which are Poisson distributed random variables, and N predicted values $\boldsymbol{\nu} = \nu_1, \ldots, \nu_N$ which depend on n parameters $\mathbf{p} = p_1, \ldots, p_n$
- Pearson's χ^2 statistic is defined as

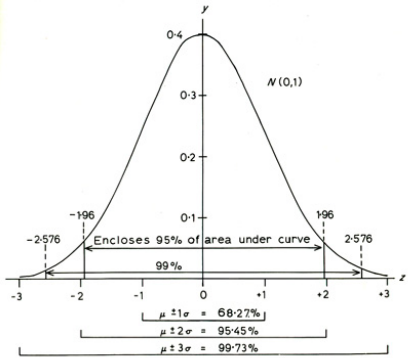
$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i}$$

- If n_i are large, Pearson's χ^2 statistic follows χ^2 p.d.f. $f(\chi^2, d)$, with $d = N - n$ degrees of freedom
- p-value for Pearson's χ^2 statistic:

$$\mathcal{P} = \int_{\chi^2_{obs}}^{\infty} f(\chi^2, d)\, d\chi^2$$

Confidence intervals

- If the parameter estimator P is believed to be Gaussian distributed, one can just quote its mean value $\langle P \rangle$ and the standard deviation $\Delta P = \sqrt{\sigma_P^2}$ as the result of the measurement
- Confidence interval: range of parameter values $p_{min} < p < p_{max}$ such that the probability that the true value of the parameter p_{true} lies within the range is a predefined number α called confidence level



Confidence intervals (2)

- What to do in the case when the parameter estimator distribution is not Gaussian, or there are physical boundaries on possible values of p?
- The procedure:
  - consider a test of the hypothesis that the parameter's true value is p
  - exclude all values of p where the hypothesis would be rejected at a significance level α (in other words, where the p-value is less than α)
  - the remaining values of p constitute the confidence interval at confidence level α
- Which test to use for this procedure?
  - a popular choice is the likelihood ratio
  - the confidence intervals obtained in this way are known as Feldman and Cousins