


Statistical Data Analysis: Lecture 7

- 1 Probability, Bayes' theorem, random variables, pdfs
- 2 Functions of r.v.s, expectation values, error propagation
- 3 Catalogue of pdfs
- 4 The Monte Carlo method
- 5 Statistical tests: general concepts
- 6 Test statistics, multivariate methods
-  7 **Significance tests**
- 8 Parameter estimation, maximum likelihood
- 9 More maximum likelihood
- 10 Method of least squares
- 11 Interval estimation, setting limits
- 12 Nuisance parameters, systematic uncertainties
- 13 Examples of Bayesian approach
- 14 tba

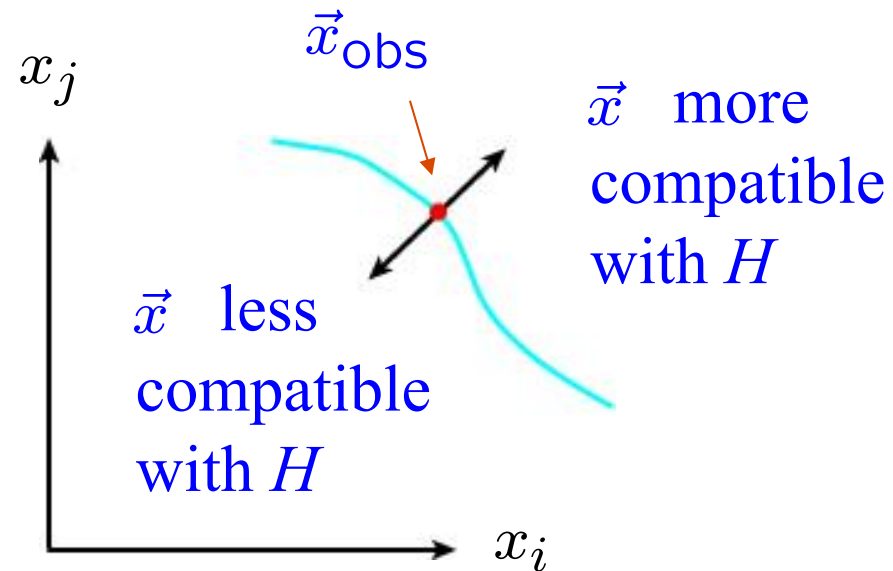
Testing significance / goodness-of-fit

Suppose hypothesis H predicts pdf $f(\vec{x}|H)$ for a set of observations $\vec{x} = (x_1, \dots, x_n)$.

We observe a single point in this space: \vec{x}_{obs}

What can we say about the validity of H in light of the data?

Decide what part of the data space represents less compatibility with H than does the point \vec{x}_{obs} .
(Not unique!)



p-values

Express ‘goodness-of-fit’ by giving the *p*-value for *H*:

p = probability, under assumption of *H*, to observe data with equal or lesser compatibility with *H* relative to the data we got.



This is not the probability that *H* is true!

In frequentist statistics we don’t talk about $P(H)$ (unless *H* represents a repeatable observation). In Bayesian statistics we do; use Bayes’ theorem to obtain

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H) dH}$$

where $\pi(H)$ is the prior probability for *H*.

For now stick with the frequentist approach;
result is *p*-value, regrettably easy to misinterpret as $P(H)$.

p-value example: testing whether a coin is ‘fair’

Probability to observe n heads in N coin tosses is binomial:

$$P(n; p, N) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

Hypothesis H : the coin is fair ($p = 0.5$).

Suppose we toss the coin $N = 20$ times and get $n = 17$ heads.

Region of data space with equal or lesser compatibility with H relative to $n = 17$ is: $n = 17, 18, 19, 20, 0, 1, 2, 3$. Adding up the probabilities for these values gives:

$$P(n = 0, 1, 2, 3, 17, 18, 19, \text{ or } 20) = 0.0026 .$$

i.e. $p = 0.0026$ is the probability of obtaining such a bizarre result (or more so) ‘by chance’, under the assumption of H .

The significance of an observed signal

Suppose we observe n events; these can consist of:

n_b events from known processes (background)

n_s events from a new process (signal)

If n_s, n_b are Poisson r.v.s with means s, b , then $n = n_s + n_b$ is also Poisson, mean = $s + b$:

$$P(n; s, b) = \frac{(s + b)^n}{n!} e^{-(s+b)}$$

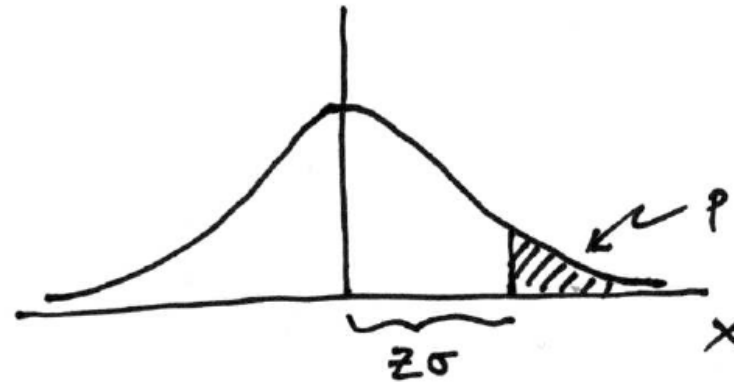
Suppose $b = 0.5$, and we observe $n_{\text{obs}} = 5$. Should we claim evidence for a new discovery?

Give p -value for hypothesis $s = 0$:

$$\begin{aligned} p\text{-value} &= P(n \geq 5; b = 0.5, s = 0) \\ &= 1.7 \times 10^{-4} \neq P(s = 0)! \end{aligned}$$

Significance from p -value

Often define significance Z as the number of standard deviations that a Gaussian variable would fluctuate in one direction to give the same p -value.



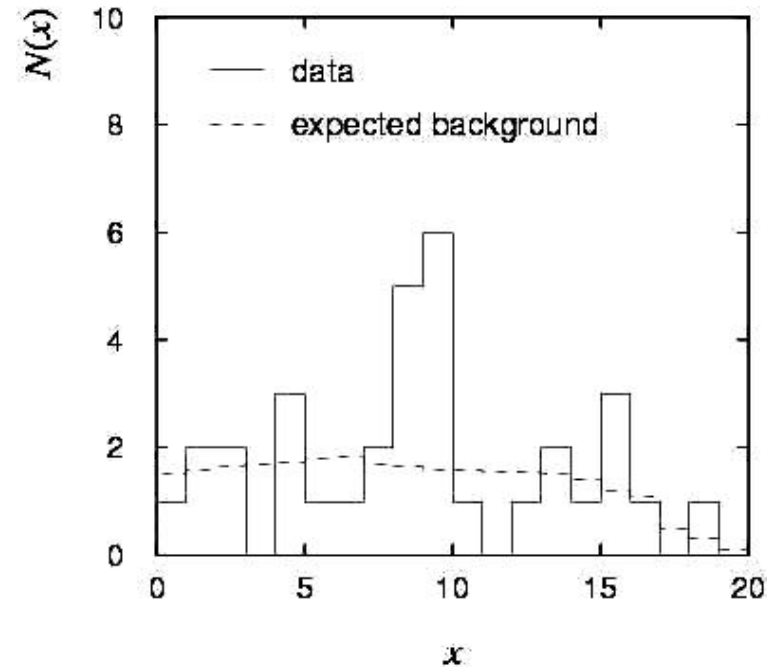
$$p = \int_Z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1 - \Phi(Z) \quad \mathbf{1 - TMath::Freq}$$

$$Z = \Phi^{-1}(1 - p) \quad \mathbf{TMath::NormQuantile}$$

The significance of a peak

Suppose we measure a value x for each event and find:

Each bin (observed) is a Poisson r.v., means are given by dashed lines.



In the two bins with the peak, 11 entries found with $b = 3.2$.
The p -value for the $s = 0$ hypothesis is:

$$P(n \geq 11; b = 3.2, s = 0) = 5.0 \times 10^{-4}$$

The significance of a peak (2)

But... did we know where to look for the peak?

→ give $P(n \geq 11)$ in any 2 adjacent bins

Is the observed width consistent with the expected x resolution?

→ take x window several times the expected resolution

How many bins \times distributions have we looked at?

→ look at a thousand of them, you'll find a 10^{-3} effect

Did we adjust the cuts to 'enhance' the peak?

→ freeze cuts, repeat analysis with new data

How about the bins to the sides of the peak... (too low!)

Should we publish????

When to publish

HEP folklore is to claim discovery when $p = 2.9 \times 10^{-7}$, corresponding to a significance $Z = 5$.

This is very subjective and really should depend on the prior probability of the phenomenon in question, e.g.,

<u>phenomenon</u>	<u>reasonable p-value for discovery</u>
D ⁰ D ⁰ mixing	~ 0.05
Higgs	$\sim 10^{-7}$ (?)
Life on Mars	$\sim 10^{-10}$
Astrology	$\sim 10^{-20}$

One should also consider the degree to which the data are compatible with the new phenomenon, not only the level of disagreement with the null hypothesis; p -value is only first step!

Distribution of the p -value

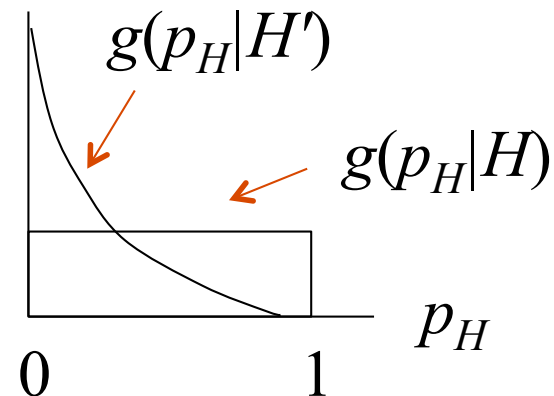
The p -value is a function of the data, and is thus itself a random variable with a given distribution. Suppose the p -value of H is found from a test statistic $t(\mathbf{x})$ as

$$p_H = \int_t^\infty f(t'|H) dt'$$

The pdf of p_H under assumption of H is

$$g(p_H|H) = \frac{f(t|H)}{|\partial p_H / \partial t|} = \frac{f(t|H)}{f(t|H)} = 1 \quad (0 \leq p_H \leq 1)$$

In general for continuous data, under assumption of H , $p_H \sim \text{Uniform}[0,1]$ and is concentrated toward zero for Some (broad) class of alternatives.



Using a p -value to define test of H_0

So the probability to find the p -value of H_0 , p_0 , less than α is

$$P(p_0 \leq \alpha | H_0) = \alpha$$

We started by defining critical region in the original data space (\mathbf{x}), then reformulated this in terms of a scalar test statistic $t(\mathbf{x})$.

We can take this one step further and define the critical region of a test of H_0 with size α as the set of data space where $p_0 \leq \alpha$.

Formally the p -value relates only to H_0 , but the resulting test will have a given power with respect to a given alternative H_1 .

Pearson's χ^2 statistic

Test statistic for comparing observed data $\vec{n} = (n_1, \dots, n_N)$
(n_i independent) to predicted mean values $\vec{\nu} = (\nu_1, \dots, \nu_N)$:

$$\chi^2 = \sum_{i=1}^N \frac{(n_i - \nu_i)^2}{\sigma_i^2}, \text{ where } \sigma_i^2 = V[n_i]. \quad (\text{Pearson's } \chi^2 \text{ statistic})$$

χ^2 = sum of squares of the deviations of the i th measurement from the i th prediction, using σ_i as the 'yardstick' for the comparison.

For $n_i \sim \text{Poisson}(\nu_i)$ we have $V[n_i] = \nu_i$, so this becomes

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

Pearson's χ^2 test

If n_i are Gaussian with mean ν_i and std. dev. σ_i , i.e., $n_i \sim N(\nu_i, \sigma_i^2)$, then Pearson's χ^2 will follow the χ^2 pdf (here for $\chi^2 = z$):

$$f_{\chi^2}(z; N) = \frac{1}{2^{N/2} \Gamma(N/2)} z^{N/2-1} e^{-z/2}$$

If the n_i are Poisson with $\nu_i \gg 1$ (in practice OK for $\nu_i > 5$) then the Poisson dist. becomes Gaussian and therefore Pearson's χ^2 statistic here as well follows the χ^2 pdf.

The χ^2 value obtained from the data then gives the p -value:

$$p = \int_{\chi^2}^{\infty} f_{\chi^2}(z; N) dz .$$

The ‘ χ^2 per degree of freedom’

Recall that for the chi-square pdf for N degrees of freedom,

$$E[z] = N, \quad V[z] = 2N .$$

This makes sense: if the hypothesized v_i are right, the rms deviation of n_i from v_i is σ_i , so each term in the sum contributes ~ 1 .

One often sees χ^2/N reported as a measure of goodness-of-fit. But... better to give χ^2 and N separately. Consider, e.g.,

$$\chi^2 = 15, N = 10 \rightarrow p\text{-value} = 0.13 ,$$

$$\chi^2 = 150, N = 100 \rightarrow p\text{-value} = 9.0 \times 10^{-4} .$$

i.e. for N large, even a χ^2 per dof only a bit greater than one can imply a small p -value, i.e., poor goodness-of-fit.

Pearson's χ^2 with multinomial data

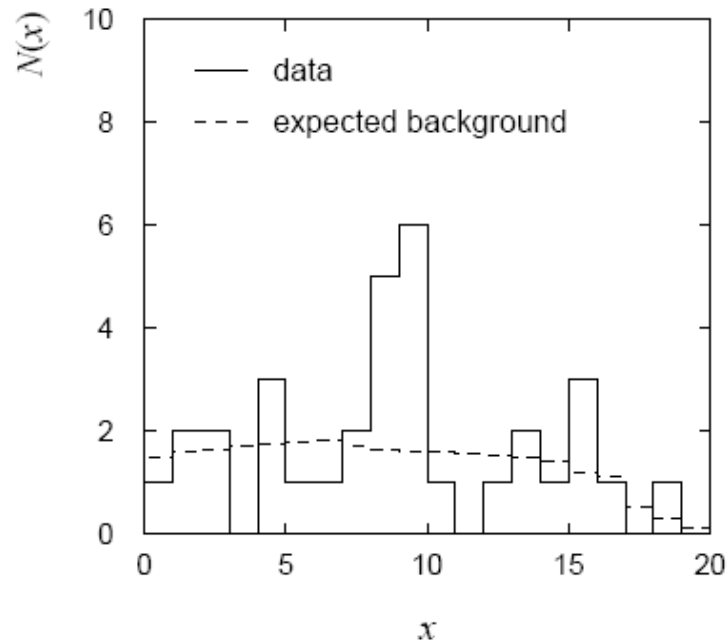
If $n_{\text{tot}} = \sum_{i=1}^N$ is fixed, then we might model $n_i \sim$ binomial
with $p_i = n_i / n_{\text{tot}}$. I.e. $\vec{n} = (n_1, \dots, n_N) \sim$ multinomial.

In this case we can take Pearson's χ^2 statistic to be

$$\chi^2 = \sum_{i=1}^N \frac{(n_i - p_i n_{\text{tot}})^2}{p_i n_{\text{tot}}}$$

If all $p_i n_{\text{tot}} \gg 1$ then this will follow the chi-square pdf for $N-1$ degrees of freedom.

Example of a χ^2 test



← This gives

$$\chi^2 = \sum_{i=1}^N \frac{(n_i - \nu_i)^2}{\nu_i} = 29.8$$

for $N = 20$ dof.

Now need to find p -value, but... many bins have few (or no) entries, so here we do not expect χ^2 to follow the chi-square pdf.

Using MC to find distribution of χ^2 statistic

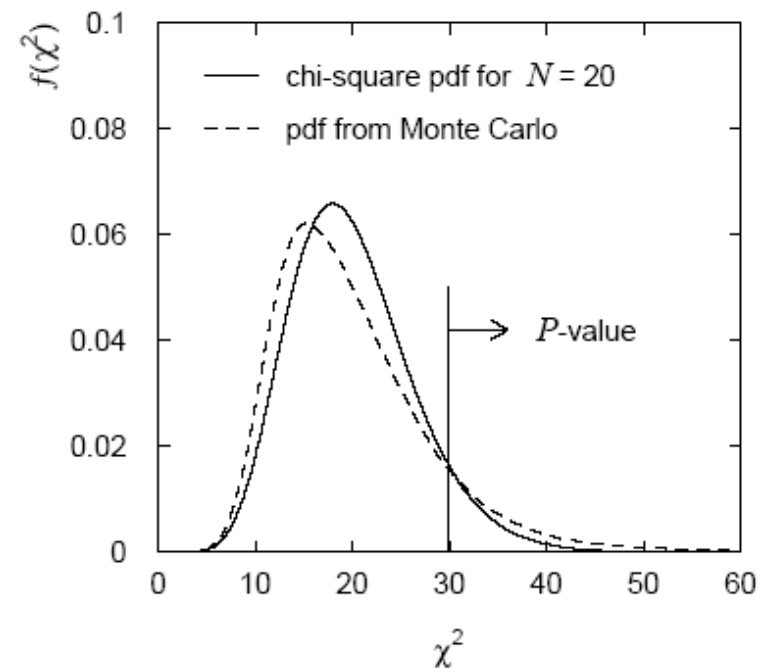
The Pearson χ^2 statistic still reflects the level of agreement between data and prediction, i.e., it is still a ‘valid’ test statistic.

To find its sampling distribution, simulate the data with a Monte Carlo program: $n_i \sim \text{Poisson}(\nu_i)$, $i = 1, N$.

Here data sample simulated 10^6 times. The fraction of times we find $\chi^2 > 29.8$ gives the p -value:

$$p = 0.11$$

If we had used the chi-square pdf we would find $p = 0.073$.



Wrapping up lecture 7

We've had a brief introduction to **significance tests**:

p -value expresses level of agreement between data and hypothesis.

p -value is not the probability of the hypothesis!

p -value can be used to define a critical region, i.e., region of data space where $p < \alpha$.

We saw the widely used χ^2 test:

statistic = sum of (data - prediction)² / variance.

Often $\chi^2 \sim$ chi-square pdf \rightarrow use to get p -value.

(Otherwise may need to use MC.)

Next we'll turn to the second main part of statistics:

parameter estimation