Statistical tools for nuclear experiments, 3

Karsten Riisager

Aarhus University

March 20, 2013

Contents

Main topics

- Preliminaries
- Fitting
- Testing
- Other topics

Subtopics

- Hypothesis test
- Goodness-of-fit
- Significance level
- Specific tests:
  - χ^2
  - likelihood ratio
  - EDF tests

Hypothesis test

Examples:

- is the detected particle a pion or a kaon ?
- is my level scheme regular or chaotic ?
- does my data support model A or model B ?

The dilemma: exclude correct results vs. include wrong ones
two types of error (probability α and β), need to optimize...

Significance level (α) = probability of excluding correct result
Power ($1-\beta$) = probability of excluding wrong result

Tests should be powerful, consistent (and without bias).

B8.1, **C**4.1, **J**10, **P**36.2

Neyman-Pearson test

A test consists (for our purpose) of two things:

- an algorithm giving a number x
- specification of which region of x is accepted, which rejected

Let our hypothesis give distribution $P_H(x)$ and the alternative hypothesis $P_A(x)$.
The **Neyman-Pearson test** uses the likelihood ratio $P_A(x)/P_H(x)$ to rank the x-regions; low values accepted, high ones rejected.

There is the "best test" for cases without fit parameters ("simple hypotheses").

Likelihood ratio test

If your hypothesis/model has parameters one may often rephrase into a likelihood ratio test using the standard likelihood $L(x;a)$. This is the case when you wish to test that the parameters a belong to a subset Ω_0 of the total parameter-space Ω_{tot}.

The maximum likelihood ratio is:

$$\lambda = \max_{a\in\Omega_0} L(x;a) / \max_{a\in\Omega_{tot}} L(x;a) \ , \quad 0 \le \lambda \le 1$$

Asymptotically $-2\ln\lambda$ has a χ^2-distribution (number of dof = difference in number of parameters in Ω_0 and Ω_{tot}).

Example: Poisson likelihood χ^2_λ, $n_i = y_i$ versus n_i free.
Dof $= N-$ number of parameters in y.

J10.5

Goodness-of-fit

Question: does the model fit my data ?
— a subset of testing, but only one "hypothesis" specified.

Most often no unique best procedure !

Two things to consider:

- Do you really want the answer ??
- If yes, do you care sufficiently to go beyond χ^2 ?

Tests can be combined. . .

The chi-square test

All time favourite: (Pearson's) chi-square test for histograms.

$$X^2 = \sum_{i=1}^{M} \frac{(n_i - Np_i)^2}{Np_i} \;, \quad p_i = \int_i y \mathrm{d}x$$

Neyman modification: $Np_i \rightarrow n_i$ in denominator.
Asymptotically a χ^2-distribution, works better when bins are chosen so that p_i about equal.

"Too nice to be true": # dof = # points − # parameters

Very general method — therefore often not powerful.

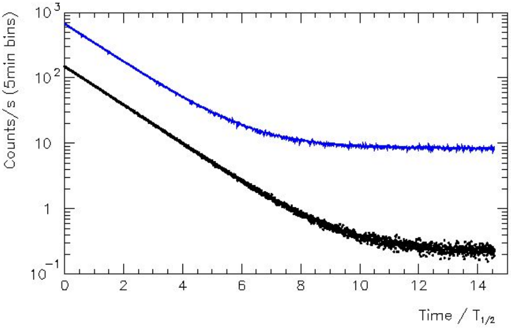
B8.3.1, **C**7.5, **J**11.2

Example: Halflife of ^{64}Cu

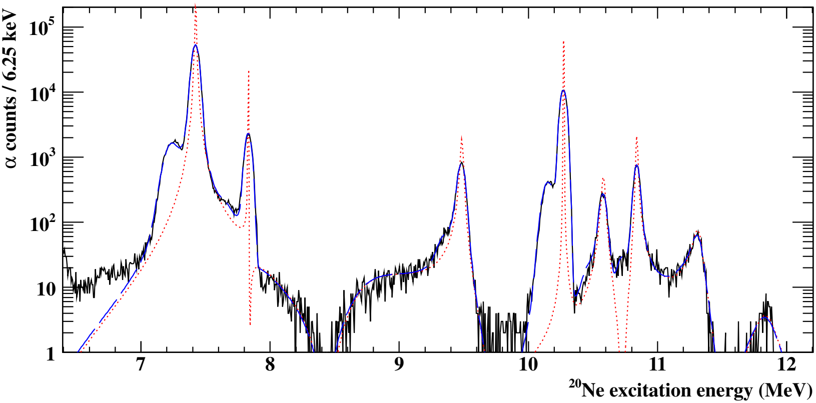
Gate on γ, correct for deadtime

$\chi^2 = 2222$
$\nu = 2228$

Courtesy Hans Fynbo



Example: ^{20}Na $\beta\alpha$ decay



$\chi^2/\nu = 2.21,\ \nu = 800$

Problems with chi-square

A technical problem: is the χ^2 value acceptable ?
Can use that $\sqrt{2\chi^2}$ is, for $\nu > 30$, Gaussian with mean $\sqrt{2\nu - 1}$ and standard deviation 1.
I.e. $\sqrt{2\chi^2} - \sqrt{2\nu - 1}$ gives the number of σ's.

The deeper problems:

- Finite sample $\rightarrow$ when does the asymptotic behaviour set in ?
- Throws away much information
  - insensitive to sign of deviation
  - cannot see systematic trends (correlated deviations)
- Sensitive to "outliers" (deviation squared !)

May combine with complimentary test, e.g. run test.

EDF tests

Compare **Empirical Distribution Function** (EDF) with cumulative distribution.

$$\mathrm{EDF}(x) = (\text{number of data points} \leq x)/N$$

Kolmogorov(-Smirnov), Cramér-von Mises, Anderson-Darling...

Pros	*Cons*
Easy to interpret for parameter-free distributions.	Need Monte-Carlo when parameters are fitted.
Distribution-free.	Mainly for 1-D data.

B8.3.3, **J**11.4

(Illustrated with figure 4 + table 1 from H. Jeppesen et al., Nucl. Phys. A709 (02) 119)

P-values etc

Prove a new effect vs. disprove "null hypothesis" ??

P-value = probability of observed data (or more extreme departures) *if* the null hypothesis is true.

If $P < \alpha$ (significance level), null hypothesis is rejected. Typically use $\alpha = 0.05$. Particle physics want positive 5σ signal.
[Are most published research findings in medicin "false positives" ? J.P.A. Ioannidis, PLoS Medicine vol 2, issue 8 (2005) e124]

P is *not* the probability of data arising by chance.
P is *not* the probability of the null hypothesis, need Bayesian methods/decision theory to decide on models

Example: χ^2 per dof and P-values

As figure 36.2 in **P**, that gives “reduced” χ^2 with corresponding P-values, shows explicitly:

Alway quote χ^2 *and n*, never just χ^2/n !

or use the $\sqrt{2\chi^2}$-rule. . .

Illustrated with figure in W review, **P** p 470:
χ^2 and dof for two sets of W boson mass measurements.

What if χ^2 is bad ?

In principle: reject model, incompatible with data !

In practice: find out where the large χ^2 comes from (plot of residuals).

If from a specific feature, it may not affect physics results (parameter errors etc).

If no obvious cause (and you insist in using the model):

enlarge error bars by $\sqrt{\chi^2/\nu}$

a PDG recipe — always think before use. . .

Other procedures

Monte Carlo simulations !

Many other tests available: Shapiro-Wilk test, run test, ranking tests . . .

Neural networks.

Bayesians employ "Bayes factor" for testing (ratio of posterior probability of two models — related to likelihood ratio).