p-values for Model Validation

Frederik Beaujean (MPI for Physics)
Allen Caldwell (MPI for Physics)
Daniel Kollár (CERN)
Kevin Kröninger (University of Göttingen)

PHYSSTAT 2011
Outline

• A Bayesian interpretation of $p$-values
• Common statistics – common pitfalls
• Runs statistic
Example problem

Suppose:

- $N$ measurements (bins/data points) with uncertainty
- Standard Model (SM) predicts quadratic background
- New physics (NP) predicts signal peak (more than one NP model)

Is Standard model enough to explain data?
Example problem

Fit function

\[ f(x | \bar{\lambda}) = A + Bx + Cx^2 + \frac{D}{\sqrt{2\pi\sigma^2}} \exp\left( -\frac{(x - \mu)^2}{\sigma^2} \right) \]

- **I**: quadratic
- **II**: constant + Gaussian
- **III**: linear + Gaussian
- **IV**: quadratic + Gaussian
**Goodness of Fit: standard approach**

**Requirement:**
- Assume a model $M$ with parameters $\lambda$

**Test statistic:**
- Any scalar function of data $T(D)$
- Interpret: large $T(D)$ = discrepancy between $M$ and $D$

**Example:**
- Probability of the data
  \[ P(D|\lambda) \propto \prod \exp \left\{ -\frac{(y_i - f(x_i|\lambda))^2}{2\sigma_i^2} \right\} = \exp \left\{ -\frac{\chi^2}{2} \right\} \]
- Familiar choice
  \[ T(D) = \chi^2(D) \]
- Extension: discrepancy variable $T(D|\lambda)$. Fitting procedure important!
• Definition:

\[ p \equiv P(T > T(D) | M) \]

• Assuming \( M \) and before data is taken: \( p \) uniform in \([0,1]\)

• Confidence level \( \alpha \):

\[ p < 1 - \alpha \Rightarrow \text{reject model} \]
Reasoning behind p-values

- Need prior knowledge about alternatives
- Good model: flat p-value
  \[ P(p|M_0) = 1 \]
- Bad model: peak at \( p=0 \), sharply falling
  \[ P(p|M_i) \approx c_i e^{-c_i p}, \quad c_i \gg 1 \]
Degree-of-belief from p-value

- Similar prior for all models: \( P(M_i) \approx P(M_j) \)
- Bayes Theorem: 
  \[
  P(M_0|p) \approx \frac{P(p|M_0)}{\sum_{i=0}^{K} P(p|M_i)}
  \]

Small \( p \):

\[
P(M_0|p \approx 0) \approx \frac{1}{1 + \sum_{i=1}^{K} c_i} \ll 1
\]

Large \( p \):

\[
P(M_0|p \approx 1) \approx 1
\]

Bayes Theorem gives justification to p-values
Comparison study

Goal: calculate p-value distribution for common statistics

- 10000 experiments
- Sample $N$ data points from Model IV with fixed parameters
- Plot the distribution of the p-value for the statistics after fitting

Beaujean, Caldwell, Kollár, Kröninger
http://de.arxiv.org/abs/1011.1674
Test Statistics: Poisson

Pearson

$$\chi^2_P = \sum_i \frac{(n_i - \nu_i)^2}{\nu_i}$$

- $n_i$ observed events
- $\nu_i = \nu_i(\bar{\lambda}, M)$ expected events

Neyman

$$\chi^2_N = \sum_i \frac{(n_i - \nu_i)^2}{n_i}$$

- Uncertainty if $n_i = 0$? Ignore bin or set uncertainty $= 1$
- Asymptotically (i.e. infinite data, in each bin: $n_i \gg 1$) know distribution of $\chi^2_P, \chi^2_N$. 
Pearson vs. Neyman

- Worrisome peak for Neyman in model III and IV (true)
- Pearson good approximation

**Uncertainties only within a model!**
Gaussian linear regression

\[ \chi^2(\vec{\lambda}, M) = \sum_{i=1}^{N} \left( \frac{f(x_i|\vec{\lambda}, M) - y_i}{\sigma_i^2} \right)^2 = \sum_{i=1}^{N} z_i^2 \]

Least squares constraint, find \( \vec{\lambda}^* \) at global minimum:

\[ \nabla \chi^2 \equiv \frac{\partial \chi^2}{\partial \lambda_j} = 0 \quad j = 1 \ldots n \]

Predictions depend on parameters:

\( f(x_i|\vec{\lambda}^*, M) \) **linear** in \( \vec{\lambda}^* \) \( \Rightarrow \nabla \chi^2 = 0 \) **linear** in \( z_i \) \( \Rightarrow P(\chi^2|N - n \text{ DoF}) \)

Example: \( f(x|\vec{\lambda}) = A + Bx + Cx^2 + \frac{D}{\sqrt{2\pi} \sigma^2} \exp \left( -\frac{(x - \mu)^2}{\sigma^2} \right) \) **nonlinear**!

In real life, usually \( P(\chi^2|\vec{\lambda}^*, N, n) \neq P(\chi^2|N - n \text{ DoF}) \)
Multimodality

Two issues:
1) Find wrong mode within ranges
2) Global mode outside of ranges

- Physics motivates small parameter range: e.g. $C>0$, $\sigma>0.2$ ..., but global mode possibly in larger range
- Gradient based optimization (MINUIT/MIGRAD): need good starting point
- Clever user guess (difficult) or output from Monte Carlo sampler (preferred), e.g. Markov chain [mpp.mpg.de/bat/]

Posterior of model III for particular data set and small range, flat priors
Results: $p$-value distribution for $\chi^2$

Fitting procedure and parameter ranges affect distribution
Local vs Global Minimum

- Use $P(\chi^2|N - n \text{ DoF})$ to turn $\chi^2$ into $p$-value
- Small range: missing global minimum in some case, bias toward $p=0$
- True model, global minimum, but still distribution not flat. → Nonlinear problem

Constraining parameter range = prior belief
Different prior → different $p$-value distribution
• Most statistics disrespect order of data, information wasted

• Human brain good for simple problems

\[ \chi^2 = 32.1 \Rightarrow p = 0.16 \]

**Example:**

• Series of \( N=25 \) datapoints

• Each Gaussian with mean = 0 and variance = 1

\[ \Rightarrow \text{Can we combine information about order and magnitude of deviation?} \]

Beaujean, Caldwell
http://arxiv.org/abs/1005.32
**Proposal:**

- Split ordered data into runs

- Each success run has a weight
  \[ \chi^2_{run} \]

- Test statistic: largest weight of any success run
  \[ T \equiv \max \{ \chi^2_{run} \} \]

- \( p \)-value becomes
  \[ p_{run} \equiv P(T > T_{obs}) \]

- Similarly for failure runs

\[ \chi^2_{run} \]

\[ 23.3 \] \( \text{success} \)

\[ 0.8 \]  \( \text{failure} \)

\[ 3.2 \]
**Gaussian case:**

- Distribution of $T$ exactly calculated for any $N$ (non-parametric)
- Source code available via email

\[
\chi^2_{\text{run}} = \frac{(23.3 - 0.8)^2}{3} = 23.3
\]

\[
p_{\text{run}}(23.3) = 5.3 \times 10^{-4}
\]

\[
\chi^2_{\text{run}} = \frac{(0.9 - 8.3)^2}{3} = 9
\]

\[
p_{\text{run}}(9) = 1.3 \times 10^{-1}
\]

$N = 25$

PDF($T_{\text{obs}}$)
Runs distribution

**Small range, MCMC**

**Good model:**

- a) fitting bias towards $p=1$
- b) Success and failure similar

**Bad model:**

- a) Success and failure different
- b) Bias towards $p=0$
- c) Missed a peak: failures OK

Use nonparametric $P(T|N)$
Runs: Joint distribution

- Good model: symmetric around $p_{\text{failure}} = p_{\text{success}}$
- Clear separation between the two models
Conclusions

- $p$-values useful (even from Bayesian perspective) for goodness-of-fit
- Fitting can make big difference
- Choice of statistic crucial
- Beware: distributions usually approximate, keep uncertainty on $p$-value in mind

FINIS