The End of Chi-Squareds and a New Era in Goodness of Fit Tests

Examples:

The End of Bayesianism
The End of Physics
Let us end chi-squareds too!!!
History and modern developments in Goodness of Fit tests
Statistics in HEP

• Signs of growing interest: 4 workshops in the last 3 yrs
  - 2001 CERN
  - 2001 Fermilab
  - 2002 Durham
  - 2003 SLAC

• Bayesian vs Frequentist: Return of the Old Controversy

• No controversy => complimentary

• Today’s talk is entirely frequentist
• Plenty of statistical papers in physics journals/web
• Two magic words to look out for
  – **New**
    • We propose to use a certain method described in the statistics literature, perhaps with some straightforward modifications, for analysis of physics data. To our knowledge, no one used this method in physics analysis before. Example: Feldman & Cousins
    • We cooked up something and we never cared to read the statistics literature. The method is new because none of our collaborators ever heard about it. Example: …I am trying to be nice today
  – **Obvious**
    • integration of likelihood for extraction of upper limits => implicitly Bayesian with uniform prior
    • using likelihood value at the max likelihood estimator of the parameter to judge fit quality => basically, nonsense; usually says little about goodness of fit (yes, you can use it as a cross-check – just don’t call it “goodness of fit”)
• Two highly subjective and very extreme principles:
  – No physicist can invent a truly new statistical method
  – If a paper written by a physicist about statistical methods fails to quote statistical literature, there is a substantial probability that this paper is garbage
Outline

• Hypothesis tests: basic definitions
• What is goodness of fit (GOF) test?
• History of binned GOF tests
• Unbinned 1D GOF tests
• Unbinned multivariate GOF tests (PHYSTAT 2003 talks)
• What’s next…
Notation

- \( f(x \mid \theta) \)  probability density function (PDF) under null hypothesis
- \( f_n(x) \)  empirical probability density function estimated from data
- \( F(x \mid \theta) \)  cumulative density function (CDF) under null hypothesis
- \( F_n(x) \)  empirical cumulative density function estimated from data
- \( L(\theta \mid x) \equiv f(x \mid \theta) \)  definition of likelihood (nothing Bayesian about it)
- \( \alpha_I \)  Type I error  \( \alpha_{II} \)  Type II error
Hypothesis test

• Test $T$ of $H_0$ vs $H_1$ on observable space $X$

• Accept $H_0$ if $x \in A$; reject $H_0$ if $x \in R = A^C$

• Confidence Level = $1 - \alpha_I = \int f(x \mid H_0) dx$

  $x \in A$

• Power $\beta = 1 - \alpha_{II} = \int f(x \mid H_1) dx$

  $x \in R$

• Asymptotic properties and properties for finite samples

• Optimize power at fixed CL
  – Pitman relative efficiency of tests T1 and T2

  \[ \varepsilon(\alpha_{II}^0) = \frac{n_2(\alpha_{II} \leq \alpha_{II}^0)}{n_1(\alpha_{II} \leq \alpha_{II}^0)} \] at fixed $\alpha_I$

  – Bahadur relative efficiency of tests T1 and T2

  \[ \varepsilon(\alpha_I^0) = \frac{n_2(\alpha_I \leq \alpha_I^0)}{n_1(\alpha_I \leq \alpha_I^0)} \] at fixed $\alpha_{II}$
Ideal test

- Uniformly: \( \forall T \neq T_0 : \beta(T) \leq \beta(T_0) \) at the same instance of \( H_1 \)
- Most: \( \forall h_0 \in H_0 \) and \( \forall h_1 \in H_1 : \alpha_i(h_0) \leq \beta(h_1) \)
- Powerful: \( P(\text{reject } H_0 \text{ if } H_0 \text{ is true}) \leq P(\text{reject } H_0 \text{ if } H_1 \text{ is true}) \)

- Neyman-Pearson Lemma
  - for \( X \sim f(x \mid \theta) \) test \( H_0 : \theta = \theta_0 \) vs \( H_1 : \theta = \theta_1 \)
  - likelihood ratio test \( f(x \mid \theta_0) / f(x \mid \theta_1) > C \) is UMP

- True only for the simple hypothesis
  - if testing \( H_0 : \theta = \theta_0 \) vs \( H_1 : \theta \neq \theta_0 \)
  - LRT \( f(x \mid \theta_0) / f(x \mid \hat{\theta}) > C \) not necessarily UMP
  - however, nice asymptotic property \( 2 \log L(\hat{\theta} \mid x) - 2 \log L(\theta_0 \mid x) \sim \chi^2_p \)

- Usually, it is not clear how to find UMP test \( \theta = \{\theta_1, \theta_2, \ldots, \theta_p\} \)
Locally powerful tests

- Often need a two-sided test \( H_0 : \theta = \theta_0 \) vs \( H_1 : \theta \neq \theta_0 \)
- Taylor expansion

\[
\log L(\theta \mid x) = \log L(\theta_0 \mid x) + (\theta - \theta_0) \frac{d}{d\theta} \left( \log L(\theta \mid x) \right)_{\theta = \theta_0} + \frac{1}{2} (\theta - \theta_0)^2 \frac{d^2}{d\theta^2} \left( \log L(\theta \mid x) \right)_{\theta = \theta_0} + \ldots
\]

- Score \( U_i(\theta) = \frac{\partial}{\partial \theta_i} \log L(\theta \mid x) \)
- Information matrix \( I_{ij}(\theta) = -E_\theta \left[ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log L(\theta \mid x) \right] \)
- Wald, 1943 \( W = (\hat{\theta} - \theta_0)^T I(\hat{\theta})(\hat{\theta} - \theta_0) \sim \chi^2_p \)
- Rao, 1948 \( R = U(\theta_0)^T I(\theta_0)^{-1} U(\theta_0) \sim \chi^2_p \)
Quiz

\[ X = \left\{ x_i = i/10; \quad i = 0,1,\ldots,10 \right\} \quad 0 \leq x \leq 1 \]

Is this sample drawn from a uniform distribution on \([0,1]\)?
What is Goodness of Fit?

• Suppose…
  \[ X = \left\{ x_i = i/10; \quad i = 0,1,\ldots,10 \right\} \quad 0 \leq x \leq 1 \]

• Is this distribution uniform?
  – alternative = presence of peaks in the data \implies YES
  – alternative = highly structured (equidistant) data \implies NO
    • example: measure elapsed time between two events in a Geiger counter and plot intervals sequentially on a straight line

• What will GOF tests tell us?
  – binned chi-squared \implies YES
  – Kolmogorov-Smirnov \implies YES
  – distance to nearest neighbor \implies NO
    • rejects equidistant data
  – Anderson-Darling \implies NO
    • …but an entirely different reason: sensitivity to tails
Formulation of the problem

• **Test**

\[ H_0: \text{ data obey } f(x | \theta = \theta_0, \tau) \]
\[ H_1: \text{ anything else} \]

• Have to assume that we know more about H1 than “anything else”. Otherwise, there is no criterion for choosing one GOF test over another.

• Design a GOF test in any way you like…

• …but we need to know when it will work and when it won’t => test the technique on several alternatives

• **There is no ultimate GOF test**
History of $\chi^2$ tests

• Pearson, 1900

$$\sum_{i=1}^{M} \frac{(n_i - np_i)^2}{np_i} \sim \chi^2_{M-1} \quad n \to \infty$$

$$p_i = \int f(x \mid \theta) \, dx$$

$$\theta = \{\theta_1, \theta_2, \ldots, \theta_p\}$$

• Fisher, 1924
  – if $\theta$ is estimated from chi-squared minimization, then $\chi^2_{M-1-p}$

• If estimated by other means, distribution may be different

• Chernoff & Lehmann, 1954
  – if $\theta$ is estimated by likelihood maximization

$$\chi^2_{M-1-p} \leq \chi^2_{M-1-p} + \sum_{k=1}^{p} \lambda_k(\theta) \chi^2_1 \leq \chi^2_{M-1} \quad 0 \leq \lambda_k(\theta) < 1$$
Alternatives

- Log likelihood ratio
  \[ \sum_{i=1}^{M} n_i \log \left( \frac{n_i}{np_i} \right) \]  
  \( \lambda = 0 \)

- Log likelihood ratio modified
  \[ \sum_{i=1}^{M} np_i \log \left( \frac{n_i}{np_i} \right) \]  
  \( \lambda = -1 \)

- Neyman modified
  \[ \sum_{i=1}^{M} \frac{(n_i - np_i)^2}{n_i} \]  
  \( \lambda = -2 \)

- Freeman-Tukey
  \[ \sum_{i=1}^{M} \left( \sqrt{n_i} - \sqrt{np_i} \right)^2 \]  
  \( \lambda = -\frac{1}{2} \)

Cressie & Read, 1984:

\[ \frac{2}{\lambda(\lambda + 1)} \sum_{i=1}^{M} n_i \left[ \left( \frac{n_i}{np_i} \right)^{\lambda} - 1 \right] \]

all asymptotically \( \sim \chi^2_{M-1} \) (if params known in advance)
Cressie & Read’s conclusions

• moments converge to asymptotic $\chi^2$ moments most fast for $0.3 \leq \lambda \leq 2.7$
• when no knowledge of alternative available, use $0 \leq \lambda \leq 1.5$
• when alternative is peaked, use $\lambda = 1$
• when alternative is dipped, use $\lambda = 0$
• use $\lambda = \frac{2}{3}$ as an “excellent compromise” for $np_i \geq 1$ and $n \geq 10$

among popular chi-squared statistics, this conclusion favors the original Pearson test $\lambda = 1$ in the sense of Pitman efficiency
General quadratic forms

- GOF measure = \( V^T Q V \sim \sum_{n \to \infty} \chi^2 \) \( \theta = \{\theta_1, \theta_2, \ldots, \theta_p\} \)

\[ V_i = \frac{(n_i - np_i)}{\sqrt{np_i}} \quad i = 1, \ldots, M \]

- For Pearson test \( Q = I_{M \times M} \)

- Rao & Robson, 1974

\[ Q_{M \times M} = I_{M \times M} + B_{M \times p} \left( J_{p \times p} - B_{M \times p}^T B_{M \times p} \right)^{-1} B_{M \times p}^T \]

\[ J_{ij} = -E \left( \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log L(\theta | x) \right) \quad B_{ij} = \frac{1}{\sqrt{p_i(\theta)}} \frac{\partial p_i(\theta)}{\partial \theta_j} \]

\[ V^T (\hat{\theta}) Q(\hat{\theta}) V(\hat{\theta}) \sim \chi^2_{M-1} \]
Who cares about binned tests when you can use an unbinned one?

- Unbinned tests are believed to be more powerful, even for large \(n\)

• Univariate tests:
  - general
    - Kolmogorov-Smirnov
    - Cramer-von Mises family
      - Cramer-von Mises test
      - Anderson-Darling
      - Watson
  - specialized
    - uniformity
    - exponentiality
    - normality

Well studied and documented, e.g., book by D’Agostino & Stephens

- Multivariate tests?
Power of univariate EDF tests

as per M. Stephens, co-author of “GOF techniques”

• “EDF statistics are usually much more powerful than the Pearson chi-square statistics”
• KS is most well-known but often less powerful than CvM and AD
• Watson statistic is powerful for detection of clustering of F-values at one point
• AD is similar to CvM but is more sensitive to the tails
Neyman smooth tests

• The idea goes back to Neyman, 1937
  – define alternative to f(x) as $g(x) = C(\theta) \exp \left[ \sum_{i=1}^{K} \theta_i h_i(x) \right] f(x)$
  – now test $H_0: \vec{\theta} = 0$ vs $H_1: \vec{\theta} \neq 0$

• A set of orthonormal functions $h_i(x)$ appropriate for this null hypothesis:
  – uniform ⇒ Legendre polynomials
  – normal ⇒ Hermite-Chebyshev polynomials
  – exponential ⇒ Laguerre polynomials
  – Poisson ⇒ Poisson-Charlier polynomials
  – etc

• GOF statistic derived from Rao statistic: $\sum_{i=1}^{K} \left[ \sum_{j=1}^{n} h_i(x_j) \right]^2$
  – more complicated with nuisance parameters

• Most difficult question: how to choose K?
  – in practice usually choose K=2,3,4
  – Teresa Ledwina, 1994-1996: Data driven smooth tests
    • choose max number of dimensions M large enough
    • define parameter subsets \( \Omega_K = \{ \theta_i = 0; \ i = K + 1, K + 2, \ldots, M \} \)
    • likelihood under smooth alternative
      \[
      \log L(\theta) = \sum_{j=1}^{n} \log[g(x_j | \theta)]
      \]
    • information number \( L_K = \sup_{\theta \in \Omega_K} \log L(\theta) - \frac{1}{2} K \log n \)
    • Schwartz’ Bayesian information criterion: choose K that gives maximal information number
  – This method generally improves powers of Neyman smooth tests for a broad range of alternatives, in some cases dramatically
Multivariate fits

• Typical CLEO/BaBar/Belle analysis:
  – estimate parameters from Max Likelihood fit \( L_0 = L(\hat{\theta} \mid x_{\text{DATA}}) \)
  – generate toy MC assuming \( \theta_0 = \hat{\theta} \)
  – derive GOF from \( L_0 \) and distribution of \( L(\hat{\theta} \mid x_{\text{TOY}}) \)

• What is wrong with this procedure?
  – often does not say anything useful about consistency of model and data
  – …understandably so because distribution of likelihood values is not parameter-independent

Remember chi-squared tests?

\[
\sum_{i=1}^{M} \frac{(n_i - np_i)^2}{np_i} \sim \chi^2_{M-1-p} \quad \text{parameter-free}
\]

Location family \( f(x - \mu) \)

Scale family \( \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right) \)

\[
-2 \log L = 2n \log \sigma - 2 \sum \log \left[ f\left(\frac{x_i}{\sigma}\right) \right]
\]
• A simple example: 
  \[ f(t \mid \tau) = \frac{1}{\tau} \exp \left( -\frac{t}{\tau} \right) \]
  \[ \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i \quad \Rightarrow \quad -2 \log L(\hat{\tau}) = 2n(1 + \log \hat{\tau}) \]

• Toy MC generated with \( \tau = \hat{\tau} \) will give a perfect GOF value

• Not surprisingly: since \( \rho(L(\hat{\tau}), \hat{\tau}) = 100\% \) it is not possible to extract new information from distribution of likelihood.

Correlation between GOF statistic and parameter estimator must be small!!!
What other multivariate methods are there?

- **Kolmogorov-Smirnov**
  \[ F_n(x^{(1)}, x^{(2)}, \ldots, x^{(m)}) = \frac{n(x_i^{(1)} < x^{(1)}, x_i^{(2)} < x^{(2)}, \ldots, x_i^{(m)} < x^{(m)})}{n} \]
  - Saunders & Laud, 1980
  - Justel, Pena & Zamar, 1996
  - Powers??

- **Cramer-von Mises**
  \[ \int_X \left[ F_n(\bar{x}) - F(\bar{x}) \right]^2 \psi(\bar{x}) f(\bar{x}) d\bar{x} \]
  - Powers??

- Problem: empirical CDF is not distribution-free => enhanced sensitivity to some parts of the observable space, reduced sensitivity to others

- My gut feeling: don’t use empirical CDF for multivariate GOF

- Tests of multivariate normality (bivariate mostly)
  - Mardia’s, Shapiro-Wilk’s, bivariate Neyman etc
  - relatively well studied
GOF talks at PHYSTAT 2003

- **Zech**
  - comparison of 2 multivariate samples using “energy test”
- **Yabsley**
  - applied Zech’s “energy test” to 2D and 1D fits in a Belle analysis
- **Kinoshita**
  - applied von Mises test of uniformity on a circle to 1D data
- **Raja**
  - likelihood ratio test using density estimated from data and density specified by the model
- **Bonvicini**
  - "Generalized chi square from extended likelihood moments" (transparencies not available)
- **Narsky**
  - multivariate GOF using distances to nearest neighbors
- **Friedman (discussion panel)**
  - GOF based on decision trees
Two-sample comparison using energy function

- Aslan & Zech,
  - hep-ex/0203010, “A new class of binning-free, multivariate goodness-of-fit tests: the energy tests”
  - math.PR/0309164, “A NEW TEST FOR THE MULTIVARIATE TWO-SAMPLE PROBLEM BASED ON THE CONCEPT OF MINIMUM ENERGY”

\[ \phi = \int dx \int dy [f(x) - f_0(x)] [f(y) - f_0(y)] R(x, y) \]

- N events in sample X and M events in sample Y

\[ \phi = \frac{1}{N^2} \sum_{i \neq j} R(x_i, x_j) - \frac{2}{NM} \sum_{i,j} R(x_i, y_j) + \frac{1}{M^2} \sum_{i \neq j} R(y_i, y_j) \]

- include or not include variability within MC sample? (I vote “yes”)
- use Euclidian distance for R: \( R(x, y) = R(|x - y|) \)
- if MC generation is expensive, use bootstrap
• Several candidates for $R(r)$:

\[
R_{\text{pow}}(r) = \begin{cases} 
\frac{1}{r^\kappa} & \text{for } r > d_{\text{min}} \\
\frac{1}{d_{\text{min}}^\kappa} & \text{for } r \leq d_{\text{min}} 
\end{cases}
\]

\[
R_{\log}(r) = \begin{cases} 
-\ln r & \text{for } r > d_{\text{min}} \\
-\ln d_{\text{min}} & \text{for } r \leq d_{\text{min}} 
\end{cases}
\]

\[
R_G(r) = \exp \left(-\frac{r^2}{(2s^2)}\right)
\]

• Of course, only $1/r$ comes from electrostatic energy; others have purely statistical motivation
• Aslan & Zech did quite a good job in comparison of test powers
• Their favorite choice is $R(x, y) = -\log |x - y|$
• **1D test of uniformity**

• **Bivariate Normality**

\[
 f_0 = \frac{1}{2\pi} \exp \left( -\frac{x^2 + y^2}{2} \right)
\]
• More tests in 2D and 4D (Zech’s talk at PHYSTAT 2003)

Power at 5% significance, 2d:

Power at 5% significance, 4d:

R: Friedman-Rafsky
N: Nearest Neighbor
Phi: energy

• Excellent performance of the “energy” test up to 4D!!!
No physicist can say anything truly new about statistics

- Research: 0.5 hr of my time plus Google engine
  - squared distance between samples $X$ and $Y$

$$
\Delta^2(\Pi_1, \Pi_2) = \int_{S^2} d^2(x, y) f(x)g(y)\lambda(dx)\lambda(dy) - V_d(X) - V_d(Y)
$$

- variability of $X$ with respect to measure of dissimilarity $d(x,x')$

$$
V_d(X) = \frac{1}{2} \int_{S^2} d^2(x, x') f(x)f(x')\lambda(dx)\lambda(dx').
$$

- Jensen difference and Rao’s quadratic entropy (Rao, 1982)

- Remember Zech’s formula?

$$
\phi = \int dx \int dy [f(x) - f_0(x)][f(y) - f_0(y)]R(x, y)
$$
• Difference between physics and math:

A new class of binning-free, multivariate goodness-of-fit tests: the energy tests

Comparison of two multivariate samples using a logarithmic measure of point-to-point dissimilarity

• Is Zech’s method new? Of course, it is!!! Cuadras & Fortiana never attempted to use $\log|x-y|$ as measure of dissimilarity

• Aslan & Zech’s paper was certainly useful because Cuadras & Fortiana did not study the power properties of this method extensively

• Lesson – use Google for everything
Two-sample comparison using nearest neighbor counts

• Friedman & Steppel, 1974; Schilling, 1986; Cuzick & Edwards, 1990
  – mix two samples together and count nearest neighbors that belong to the same sample
  \[ X(N) = \{x_1, x_2, \ldots, x_N\} \quad Y(M) = \{y_1, y_2, \ldots, y_M\} \]
  \[ Z(N + M) = \{z_i = x_i, 1 \leq i \leq N; \quad z_{N+i} = y_i, 1 \leq i \leq M\} \]
  – GOF statistic based on
  \[ T_K = \frac{1}{(N + M)K} \sum_{i=1}^{N+M} \sum_{k=1}^{K} I_i(k) \]
  \[ I_i(k) = \begin{cases} 
  1, & \text{if } k\text{-th NN of point } i \text{ belongs to the same sample} \\
  0, & \text{otherwise} 
\end{cases} \]

• Friedman & Rafsky, 1979
  – minimal spanning tree (straight lines, no loops, min length)
  – GOF statistic = number of connections between samples
Raja, “The End of Bayesianism”

• **Claims that**
  – “With unbinned data, currently, the fitted parameters are obtained but no measure of goodness of fit is available.” – Paper submitted to Elsevier
  – “Prior to my paper, the problem of goodness of fit in unbinned likelihoods was an unsolved one.” – Private Communications

• **This work lacks study of power functions, comparison with other unbinned GOF methods and realistic examples from HEP analysis. Hopefully, will be extended in the future.**

• **Motivated his method through Bayesian approach while, in my opinion, there is nothing Bayesian about it.**
• Basic idea:

Probability Density Estimator

\[ f_n(x) = \frac{1}{n} \sum_{i=1}^{n} K(x, x_i) \]

and then compare \( f_n(x) \) and \( f(x) \)

• Bickel & Rosenblatt, 1973; Bowman, 1989; Gourieroux & Tenreiro, 1996

\[ L_2 = \int_{\mathbb{R}} \left[ f_n(x) - E_0 f_n(x) \right]^2 \psi(x) dx \]

• Bowman, 1989: "Density based tests for goodness of fit"

use \( \int_{\mathbb{R}} [f_n(x) - f(x)]^2 dx \)

beware of bias: \( E_{H_0} [f_n(x)] \neq f(x) \)

use \( \int_{\mathbb{R}} [f_n(x) - E_0 f_n(x)]^2 dx \)

\[ K(x_i, x) = \frac{1}{\sqrt{2\pi \sigma h}} \exp \left[ -\frac{(x - x_i)^2}{2\sigma^2 h^2} \right]; \quad h = h(f, n) \]
• Bowman compared performance of the new statistic with
  – Vasicek, CvM, AD and Shapiro-Wilk tests of normality
  – Watson test of Cramer-von Mises distribution
• Conclusion: performs better for some alternatives, worse
  for others

So is Raja’s method new? Of course, it is!!!

Previous authors: Integrated Squared Error
\[ \int_X \left[ f_n(x) - E_0 f_n(x) \right]^2 dx \]

Raja: Likelihood Ratio Test
\[ -2 \sum_{i=1}^{n} \log \left[ \frac{f(x_i)}{f_n(x_i)} \right] \]
Distance to nearest neighbor

• **Review by Dixon (Iowa State)**

• **Clark & Evans, 1954, 1979**
  – GOF statistic = average distance between nearest neighbors within a population
  – Used to test 2D populations of various plants for uniformity
  – Later generalized this approach to an arbitrary number of dimensions cited 1000+ on Web of Science

• **Diggle, 1979**
  – Build an entire distribution of ordered distances to nearest neighbors and apply KS or CvM test to this distribution
  – Not a true p.d.f., of course, because NN distances are obviously interdependent
• Ripley, 1976-1977 (cited 600+ on WoS)
  – \( \lambda K(t) = E\{n \text{ points within distance } t \text{ of an arbitrary point of the process}\} \)
  – \( \hat{\lambda} = V / N \) - average intensity of the process
  – for a uniform process on a plane \( K(t) = \pi t^2 \)
  – plot expected and observed \( K \) vs \( t \) and estimate GOF from maximal distance between the two \( K \)'s
  – Seems to be a popular method in ecology

• Bickel & Breiman, 1983; Schilling, 1983
  – GOF based on distributions of \( Z_i = \exp[-nf(x_i)V(x_i)] \)
    = Poisson probability of observing one point in a sphere \( V \) centered at this point
  – form a full distribution of ordered \( Z_i \)'s and compare with the expected one
  – studied performance for multivariate normal densities

• SLEUTH at D0: search for new physics
  – instead of spheres, use Voronoi regions (Voronoi region = region of space that is closer to this point than to any other point)
Test of Uniformity of Positions of Redwood seedlings at CL>95%

Left. Fig. 10. The positions of 62 Redwood seedlings.

Right. Fig. 11. The solid curves are the plots of $\hat{K}$ for the Redwood data and of $K$ for the Poisson process (the parabolic curve). The lower and upper pairs of dashed curves are the envelopes of the plots of $\hat{K}$ for 99 simulations of the Poisson process and 20 samples of Strauss' model.

Test of Uniformity of Distances between centers of Spanish towns at CL=97%

Ilya Narsky
Caltech November 2003
• Use a bivariate distribution of maximal vs minimal distance to K nearest neighbors
• Ideal for detection of well-localized irregularities, e.g., unexpected peaks in the data
• Powers for bivariate normals and uniform pdf
  – compared only with KS
• Example: $B \rightarrow K^{(*)}l^+l^-$ How consistent are the data with background pdf?
How to choose number of nearest neighbors?

- Schilling:

Since the best \( k \) depends critically on the size of the clusters which are present, it is doubtful that a general answer to this problem is obtainable without turning to adaptive procedures. This is perhaps the most fundamental difficulty with cluster analysis in general.
GOF based on decision tree

• As sketched by Friedman at the Panel Discussion
  – my free interpretation (because transparencies not available)
• Standard decision tree and CART algorithm
  – \( n(t) \) events with coords \( x_i \) and features \( d_i \)
    in a subset \( t \) of the decision tree \( T \)
  – \( \bar{d}(t) = \frac{1}{n(t)} \sum_{x_i \in t} d_i \)
  – error per subset \( e(t) = \frac{1}{n(t)} \sum_{x_i \in t} (d_i - \bar{d}(t))^2 \)
  – error per tree \( e(T) = \sum_{t \in T} e(t) \)
  – the optimal set of tree splits is the one that minimizes overall error \( e(T) \)
  – hence, \( e(T) \) can be a measure of GOF
Transform or not transform?

- Generally, all multivariate methods described above are applicable to “raw” distributions
- But if you want a GOF test to be equally sensitive to all regions of the observable space, need to perform this test in the flattened space (aka unit hypercube)

- Transformation to uniformity, of course, is not unique
  - transformation to m-dimensional unit cube using marginal CDF’s
    \[
    u_1 = F(x_1) \\
    u_i = F(x_i \mid x_1, x_2, \ldots, x_{i-1}) \quad 2 \leq i \leq m
    \]
  - point-to-point distances not invariant under relabeling of components, rotation etc
Summary

• Physicists are smart – they can independently re-invent statistical methods invented decades ago…
• …and doing so they attract attention of the community to application of statistical methods in HEP practice.
• Joy of rediscovery aside, we need to admit that we are not as statistically advanced as other communities (e.g., ecology, medical research, finance etc). A better idea might be to study the statistics literature instead of re-inventing it.
• It would be very useful to study powers of the mentioned GOF tests on realistic HEP problems.