Goodness of Fit Test

Dr. Wolfgang Rolke
University of Puerto Rico – Mayaguez
2nd Pan-European Advanced School on Statistics in High Energy Physics
March 28, 2022

2gdgdgd2nd Pan-European Advanced School on S2tatistics in

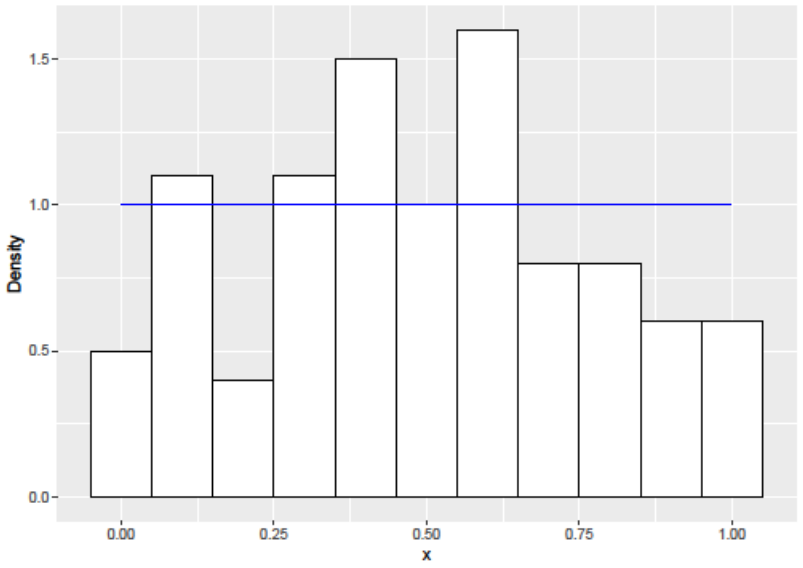
Table of Content

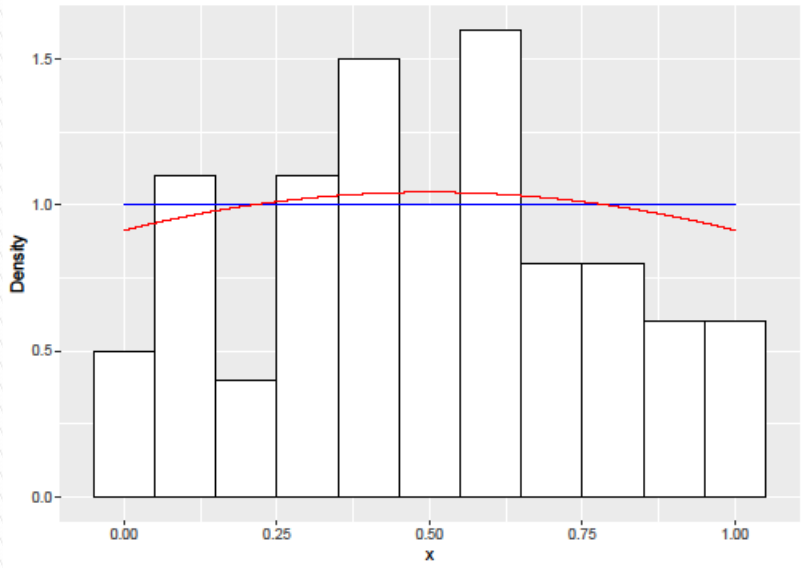
- Problem statement
- Hypothesis testing
- Chi-square test
- Methods based on empirical distribution function
- Other tests
- Power studies
- Running several tests
- Tests for multi-dimensional data
- Special Cases
- Two-Sample Problem (MC vs Sample)
- Conclusions

The Archetypical Statistics Problem:

- We have a probability model (aka density, distribution function)
- We have data from an experiment
- Does the data agree with the probability model?

What's the density?





Good Model?

Or maybe needs more?

General Problem Statement

F: cumulative distribution function (integral of density)

$$H_0: F = F_0$$

Usually more useful:

$$H_0: F \in \mathcal{F}_0$$

$\mathcal{F}_0$ a family of distributions, indexed by parameters. Those need to be estimated from the data.

Hypothesis Testing Basics

- Type I error: reject true null hypothesis
- Type II error: fail to reject false null hypothesis

A: a HT has to have a true type I error probability no higher than the nominal one (α)

B: probability of committing the type II error (β) should be as low as possible (subject to A). Usually we discuss the power of a test $P=1-\beta$, which then should be as high as possible.

Historically A was achieved either by finding an exact test or having a large enough sample.

p value = probability to reject a true null hypothesis when repeating the experiment and observing value of test statistic or something even less likely.

If method works p-value has uniform distribution.

Frequentist vs Bayesian

Not again …

Actually no, GOF equally important to both (everybody has a likelihood)

Maybe more so for Bayesians, no non-parametric methods.

But GOF is frequentist. Bayesian GOF would need prior on space of probability distributions.

"All models are wrong but some are useful"

Box, G. E. P. (1979), "Robustness in the strategy of scientific model building", in Launer, R. L.; Wilkinson, G. N. (eds.), Robustness in Statistics, Academic Press, pp. 201–236.

There is no perfect circle in nature

There is no data set perfectly normally distributed (or exponential, or …)

In frequentist hypothesis testing, if the null hypothesis is wrong, a proper test will always reject it, as long as the sample size is large enough.

But all models are wrong!

So a GOF test will (and should!) always reject the null hypothesis for a sufficiently large sample.

But how bad can a model be if it takes a million events to reject it?

If so, it should be useful! What useful means depends on the context. (for example testing at 5σ vs 3σ levels).

Overfitting

Usual question: is our theory a good enough model for the data?

We also should worry about: is our model better than it should be?

- Overfitting!
- Occam's Razor: *Numquam ponenda est pluralitas sine necessitate* (Quantities should not be multiplied beyond necessity aka keep it simple!)
- Here: the best model is the most basic one that works (aka fits the data)

Simple Example: Is the die fair?

Theory: die is fair ($p_i = {}^1/_6$)

Experiment: roll die 1000 times

If die is fair one would expect 1000*1/6 = 167 1's, 2's and so on

Data:

1	2	3	4	5	6
187	168	161	147	176	161

- Is this a good fit?

Most Famous Answer: Pearson X^2

Sir Karl Pearson (1900),

"*On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling*", Phil. Mag (5) **50**, 157–175

Uses as measure of deviations

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

k: number of classes / categories / bins

O_i : observed counts

E_i : expected counts

Agreement is bad if X^2 is large

	1	2	3	4	5	6
O	187	168	161	147	176	161
E	167.7	167.7	167.7	167.7	167.7	167.7

$$X^2 = \frac{(187 - 167.7)^2}{167.7} + .. + \frac{(161 - 167.7)^2}{167.7} = 5.72$$

Is 5.72 “large”?

If die is fair and rolled 1000 times, how large would X^2 typically be?

Answer: $X^2 \sim \chi^2(k-1)$

Pearson's Reasoning

N_i = frequency of outcome i, $i = 1,..,k$

$(N_1,..,N_k) \sim Multinomial(n,p_1,..,p_k)$

$E[N_i] = np_i, \;\; Var[N_i] = np_i(1-p_i)$

$\frac{N_i - np_i}{\sqrt{np_i(1-p_i)}} \sim_{app} N(0,1)$ by CLT

$\left(\frac{N_i - np_i}{\sqrt{np_i(1-p_i)}} \right)^2 = \frac{(N_i - np_i)^2}{np_i(1-p_i)} \sim_{app} \chi^2(1)$

so maybe $\sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i(1-p_i)} \sim_{app} \chi^2$?

but $N_1 + .. + N_k = n$ fixed (not independent)

$$k = 2 : (N_1, N_2) = (N, n - N)$$

$$X^2 = \sum_{i=1}^{k} \frac{(N_i - E_i)^2}{E_i} =$$

$$\frac{(N - np)^2}{np} + \frac{(n - N - n(1-p))^2}{n(1-p)} =$$

$$\frac{(N - np)^2}{np} + \frac{(n - N - n + np))^2}{n(1-p)} =$$

$$\left(\frac{1}{np} + \frac{1}{n(1-p)} \right) (N - np)^2 =$$

$$\left(\frac{1 - p + p}{np(1-p)} \right) (N - np)^2 =$$

$$\frac{(N - np)^2}{np(1-p)} = \left(\frac{N - np}{\sqrt{np(1-p)}} \right)^2 \sim \chi^2(1)$$

So X^2 has a chi square distribution with k-1 degrees of freedom (k=number of categories/bins)

Here: 95^{th} percentile of $\chi^2(5)$ is 11.07

So our $X^2 = 5.72$ is not unusually large, die is (reasonably) fair.

The derivation of the distribution of X^2 uses several approximations, so this needs a sufficiently large sample size. But how large does it have to be?

Famous answer: $E_i \geq 5$ for all categories.

William G. Cochran The [chi-squared] test of goodness of fit. *Annals of Mathematical Statistics* 1952; 25:315-345.

Seems to have picked 5 for no particular reason. Later research showed this is quite conservative. Test generally works fine if $E_i \geq 5$ for most i and no $E_i < 1$.

Another Derivation of X^2

*Neyman, Jerzy; Pearson, Egon S. (1933). "On the Problem of the Most Efficient Tests of Statistical Hypotheses". Philosophical Transactions of the Royal Society A:. **231** (694–706)*

In a test of a simple vs simple hypotheses likelihood ratio test is most powerful

In the case of a multinomial also leads to X^2!

Samuel S. Wilks: "*The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses",* The Annals of Mathematical Statistics, Vol. 9, No. 1 (Mar., 1938), pp. 60-62

Λ : Likelihood Ratio

$$-2log\Lambda \cong X^2 \sim \chi^2(k-1)$$

The Degree of Freedom Controversy

Not

$H_0: F = Normal(0,1)$ (**simple** hypothesis)

but

$H_0: F = Normal$ (**composite** hypothesis)

Idea: find estimates of parameters, use those.

Any change in test? Pearson said no.

In 1915 Greenwood and Yule publish an analysis of a 2x2 table and note that there is a problem.

In 1922, 1924 and 1926 Sir Karl Fisher published several papers showing that Pearson was wrong.

If m parameters are estimated

$$X^2 \sim \chi^2(k-1-m)$$

The 1922 paper is the first ever to use the term “degrees of freedom”.

In some ways this is an astonishing result: it does not seem to matter how well one can estimate the parameter (aka what the sample size is)

Does it matter what method of estimation is used? Yes, and it has to be minimum chi-square!

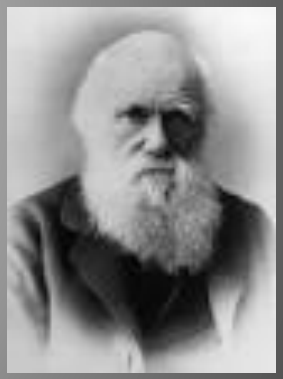
Except these days everyone is using maximum likelihood, and then this result can be wrong

Pearson didn’t acknowledge Fisher was right until 1935!

Mendel-Fisher Controversy

Mendel, J.G. (1866). "Versuche über Pflanzenhybriden", *Verhandlungen des naturforschenden Vereines in Brünn*, Bd. IV für das Jahr, 1865, *Abhandlungen*: 3-47

Discovery of Mendelian inheritance



Immediate impact on Science: ZERO!

Darwin could have used this when he wrote On The Origin of Species.

His cousin Francis Galton (inventor of regression!) could have told him.

Around 1900, Hugo de Vries and Carl Correns first independently repeat some of Mendel's experiments and then rediscover Mendel's writings and laws.

Finally Mendel becomes the "Father of Genetics"

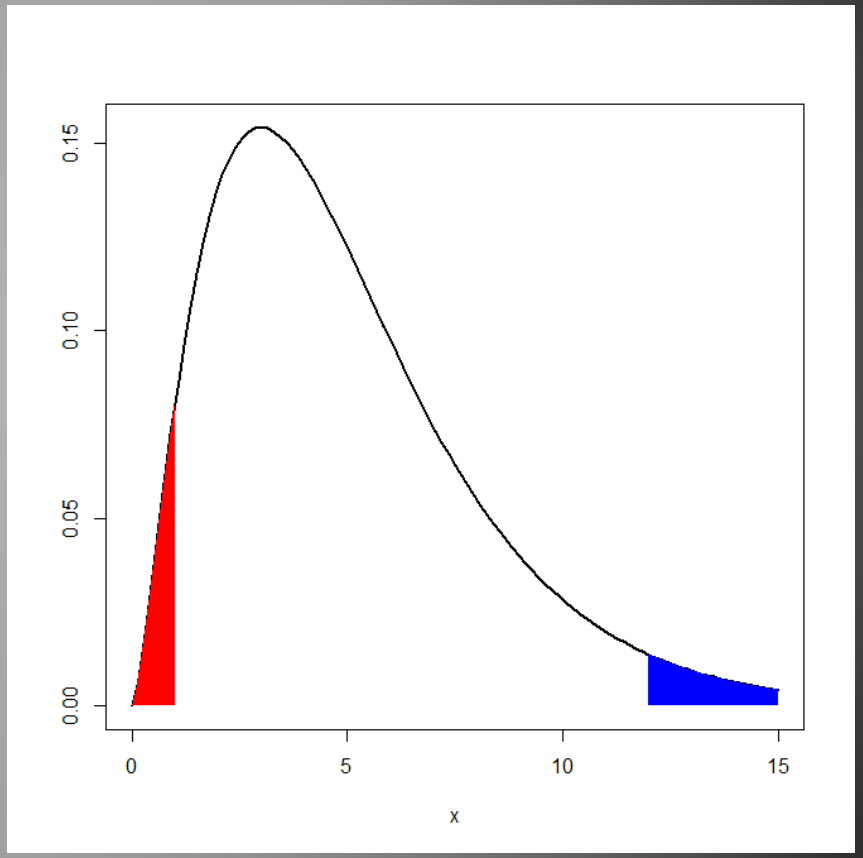
Fisher, R.A. (1936). "Has Mendel's work been rediscovered?" . Annals of Science. 1 (2): 115–137.

Fisher re-analyzed Mendel's data and applied the X^2 test to all of them together. He finds an (almost) perfect agreement. But inheritance is intrinsically random, the agreement should not be that good.

Fisher's words: "to good to be true"

X^2 large (blue area)
→ difference between O and E to large
→ theory doesn't agree with data

X^2 small (red area)
→ difference between O and E to small
→ Cheating!



More than 50 papers published since 1936 have tried to figure out what happened.

For a long time: it was the Gardener!

Another explanation, which seems to have gained momentum in recent years: It was early in the history of experimentation, modern ideas of how to avoid (even unconscious) biases were not yet developed.

Allan Franklin, A. W. F. Edwards, Daniel J. Fairbanks, Daniel L. Hartl and Teddy Seidenfeld. *"Ending the Mendel–Fisher Controversy",* University of Pittsburgh Press, 2008.

Variations on X^2

Cressie-Read	$\frac{1}{n\lambda(\lambda-1)}\sum O\left\{\left(\frac{O}{E}\right)^{\lambda}-1\right\}$
Pearson ($\lambda = 1$)	$\sum \frac{(O-E)^2}{E}$
log likelihood ratio ($\lambda = 0$)	$2\sum O\log(\frac{O}{E})$
Freeman-Tukey ($\lambda = -1/2$)	$4\sum\left[\sqrt{O}-\sqrt{E}\right]^2$
Neyman modified X^2 ($\lambda = -2$)	$\sum \frac{(O-E)^2}{O}$
modified likelihood ratio ($\lambda = -1$)	$2\sum E\log(\frac{E}{O})$

Question used to be: which converges fastest to χ^2?

But these days null distribution can be found most easily using Monte Carlo simulation!

Monte Carlo Simulation

```
function(B=1e4) {
   O<-c(187,168,161,147,176,161)
   E<-rep(1,6)/6*1000
   TS.Data<-rep(0,5)
   TS.Data[1]<-sum( (O-E)^2/E)
   TS.Data[2]<-2*sum(O*log(O/E))
   TS.Data[3]<-4*sum( (sqrt(O)-sqrt(E))^2)
   TS.Data[4]<-sum( (O-E)^2/O)
   TS.Data[5]<-2*sum(E*log(E/O))
   TS.Sim<-matrix(0,B,5)
   for(i in 1:B) {
      O<-table(sample(1:6,size=1000,replace=
      TS.Sim[i,1]<-sum( (O-E)^2/E)
      TS.Sim[i,2]<-2*sum(O*log(O/E))
      TS.Sim[i,3]<-4*sum( (sqrt(O)-sqrt(E))^2)
      TS.Sim[i,4]<-sum( (O-E)^2/O)
      TS.Sim[i,5]<-2*sum(E*log(E/O))
   }
   list(TS.Data,apply(TS.Sim,2,quantile,0.95))
}
```

Method	**Data**	95^{th}
Pearson	5.72	10.95
log likelihood ratio	5.76	10.97
Freeman-Tukey	5.75	10.95
Neyman modified	5.73	11.08
modified likelihood ratio	5.73	11.00

Question today: Which method has highest power?

```
function(B=1e4) {
   crit95<-c(10.95, 10.97, 10.95, 11.08, 11.00)
   E<-rep(1,6)/6*1000
   TS.Sim<-matrix(0,B,5)
   for(i in 1:B) {
      O<-table(sample(1:6,size=1000,replace=T,
         prob=c(1.25,1,1,1,1,1)))
      TS.Sim[i,1]<-sum( (O-E)^2/E)
      TS.Sim[i,2]<-2*sum(O*log(O/E))
      TS.Sim[i,3]<-4*sum( (sqrt(O)-sqrt(E))^2)
      TS.Sim[i,4]<-sum( (O-E)^2/O)
      TS.Sim[i,5]<-2*sum(E*log(E/O))
   }
   power<-rep(0,5)
   for(i in 1:5) power[i]<-        s
sum(TS.Sim[,i]>crit95[i])/B
   power
}
```

Simulated loaded die has a slightly higher probability for a "1".

Method	Power
Pearson	55.47%
log likelihood ratio	53.95%
Freeman-Tukey	53.33%
Neyman modified	50.50%
modified likelihood ratio	52.26%

Continuous Data

Need to bin the data

In principle any binning is ok, as long as expected counts are not to low

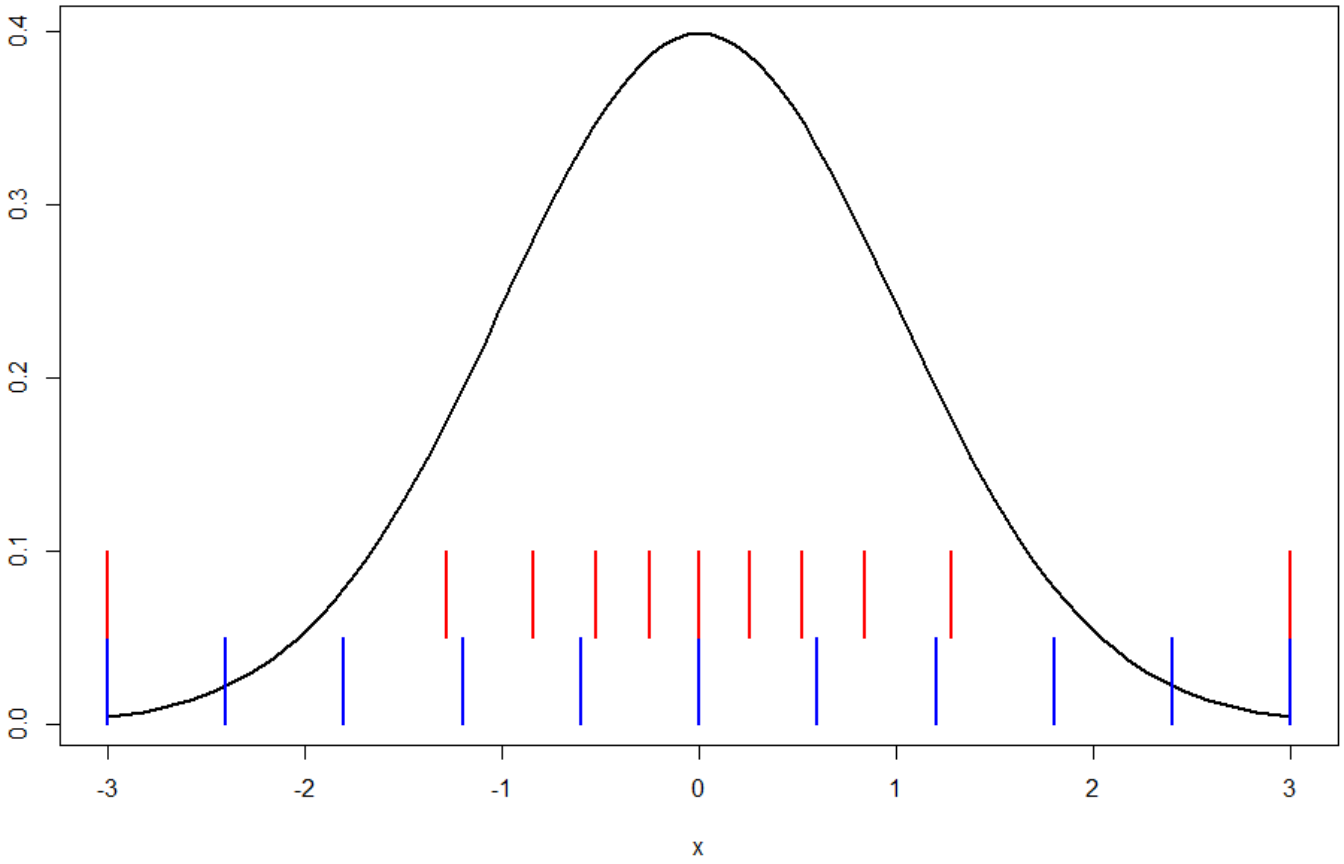
Two obvious questions:

1) What kind of bins?

2) How many bins?

What kind of bins?

Equi-distant vs Equi-probable



Most textbooks suggest equi-probable is better, but this isn't really true.

One advantage: $E=n/k >> 5$ for all bins, no need to adjust binning

Equi-probable bins can be found easily as quantiles of distribution or as quantiles of data

How many bins?

Many textbook answers:

D'Agostini and Stephens $2n^{2/5}$
Sturge's Rule $1 + log_2 n$ (used in a lot of software for histograms)
Mann and Wald $4[\frac{2(n+1)^2}{c^2}]^{1/5}$

And many more

But: really depends on case:
Example: $H_0: X \sim U[0,1]$ vs $H_a: X \sim Linear$
Optimal: k=2!

Formulas above were derived for the purpose of density estimation, but a number of bins that is good for density estimation (aka histogram) need not be good for gof testing.

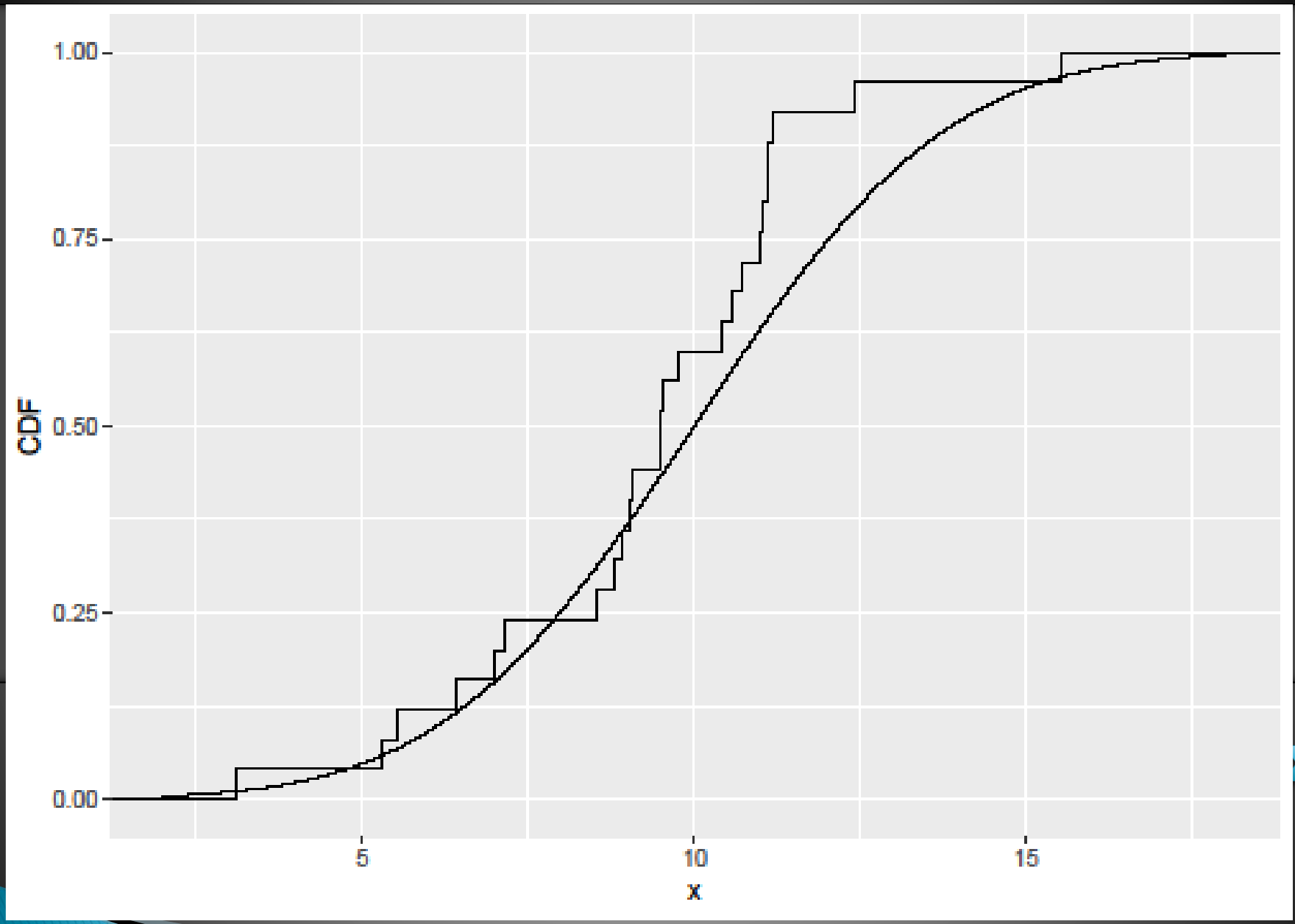
My own studies show that a small number (no more than 10) independently of n is usually best.

EDF Methods

EDF: **E**mpirical **D**istribution **F**unction

$$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i) = \frac{\text{\# of events } \leq x}{n}$$

$\hat{F}(x) \to F(x)$ uniformly (Glivenko–Cantelli lemma)


1.00
0.75
0.50
0.25
0.00
CDF
5
10
15
x

Basic idea for test:

$$\mathrm{TS} = d\{\hat{F}, F\}$$

d: distance measure on function space

Typical Example:

$$\int D\left(\hat{F}(x), F(x)\right) \psi(x) dF(x)$$

Ψ: weight function

Theorem: (*Probability Integral Transform*)
Let X be a continuous random variable with distribution function F, then the random variable $Y = F(X)$ has a uniform (0,1) distribution.

Consequence: test is distribution free, aka does not depend on F.

One table to rule them all!

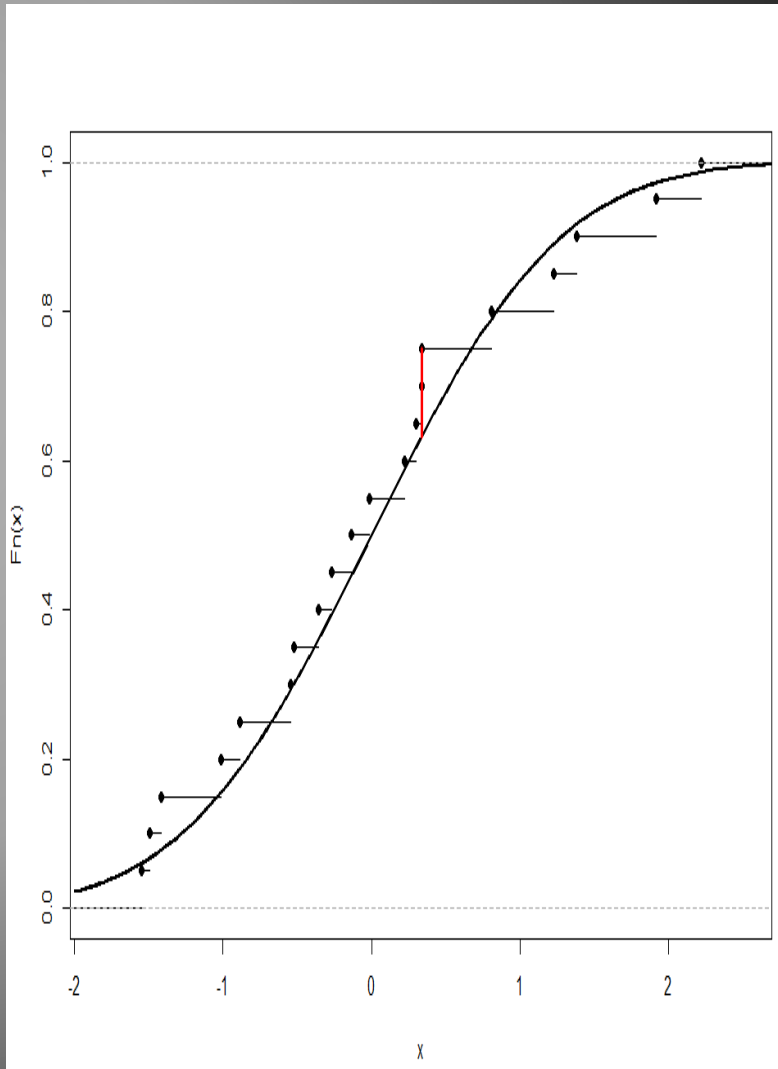
Except this does not work if parameters are estimated from data!

Kolmogorov–Smirnov

$$KS = max\{|\hat{F}(x) - F(x)|; x\} = max\left\{\left|\frac{i}{n} - F(X_{(i)})\right|, \left|F(X_{(i)}) - \frac{i-1}{n}\right|\right\}$$

Kolmogorov A (1933). "Sulla determinazione empirica di una legge di distribuzione". G. Ist. Ital. Attuari. 4: 83–91.

Smirnov N (1948). "Table for estimating the goodness of fit of empirical distributions". Annals of Mathematical Statistics. 19: 279–281

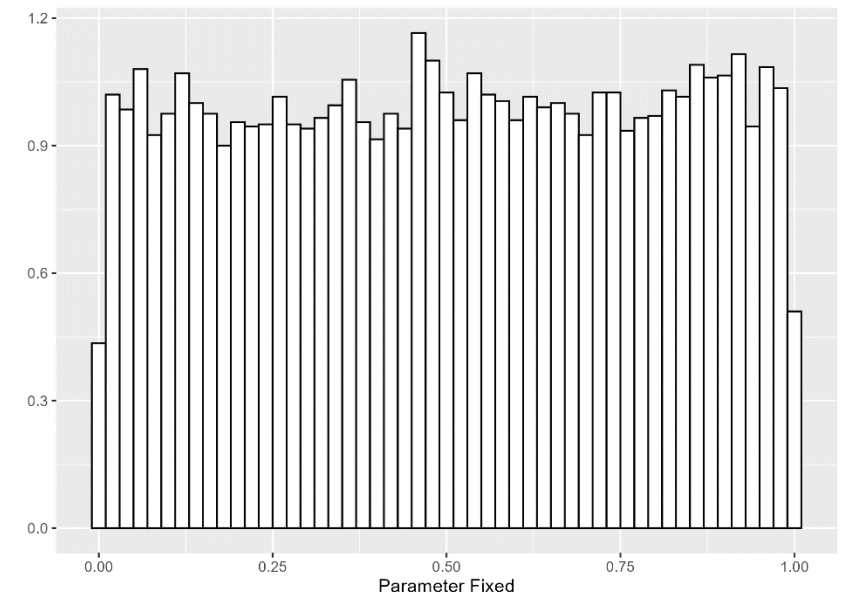


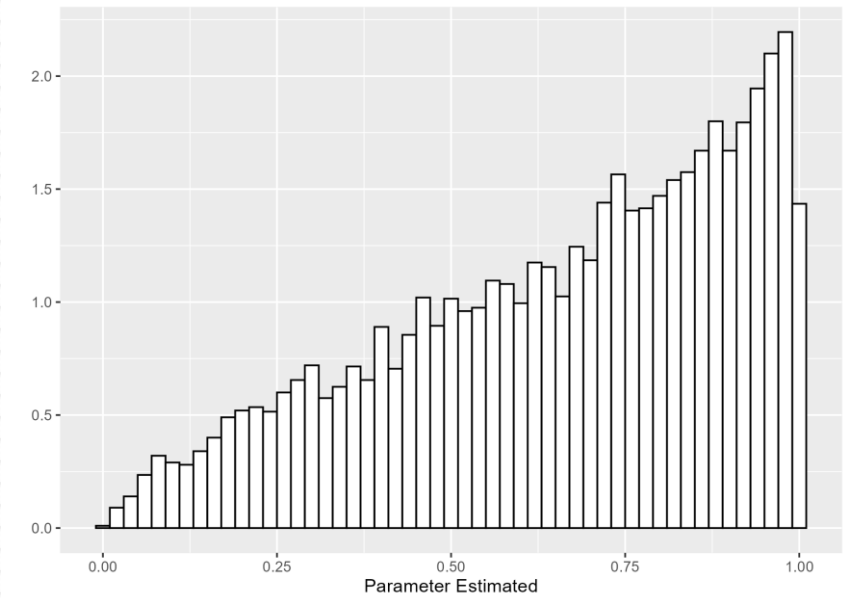
Effect of parameter estimation

- Generate 1000 events from exponential distribution rate 1.0
- KS test whether data comes from an exponential distribution.

- Case 1: rate fixed at 1
- Case 2: rate estimated from data

- Repeat 1000 times, record p values





Test with parameter estimation will be badly under-powered.

Null Distribution via Simulation

- Estimate parameters from data (and you can use any method you like!) $\mapsto \hat{\theta}_D$
- Find test statistic T_D for data, using $F(.|\hat{\theta}_D)$.
- Simulate new data set from $F(.|\hat{\theta}_D)$, find its parameter estimates $\hat{\theta}_1$, and its test statistic T_1 using $F(.|\hat{\theta}_1)$
- Do this (say) 1000 times.
- P-value $= \%\{T_i > T_D\}$ (if large T is bad)
- Parametric bootstrap

More EDF Tests

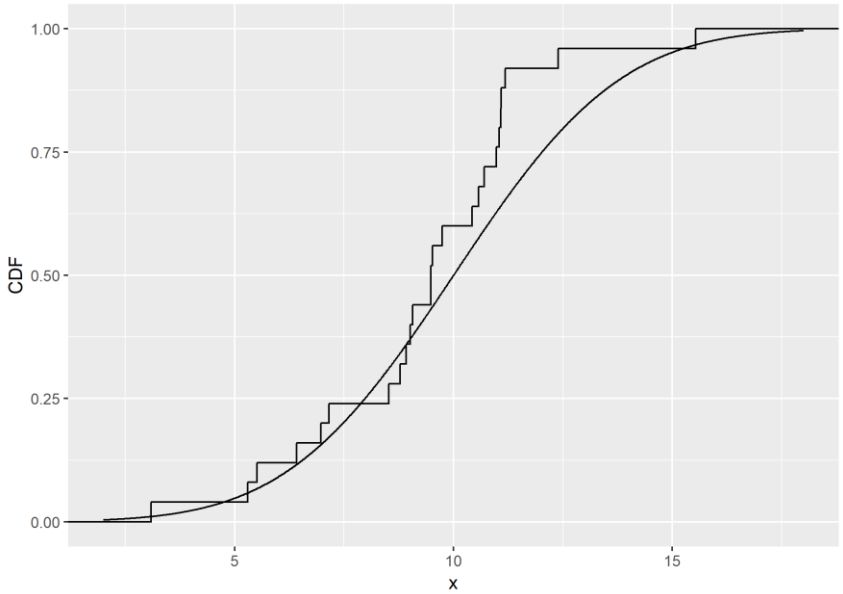
- Obvious issue with KS: uses only one x point (albeit the worst)
- Obvious alternative:

$$CM = \int_{-\infty}^{\infty} \left(\hat{F}(x) - F(x) \right)^2 d\hat{F}(x)$$

Cramer-vonMises Test

Cramér, H. (1928). "On the Composition of Elementary Errors". Scandinavian Actuarial Journal. 1928 (1): 13–74. doi:10.1080/03461238.1928.10416862.

von Mises, R. E. (1928). Wahrscheinlichkeit, Statistik und Wahrheit. Julius Springer.



Note that on the left both functions are 0, and on the right both are 1. So no difference!
Idea of Anderson-Darling: inflate differences there

$$AD = n\int_{-\infty}^{\infty} \frac{(\hat{F}(x) - F(x))^2}{F(x)[1 - F(x)]} dF(x)$$

$$A^2 = -n - \sum_{i=1}^{n} \frac{2i-1}{n} [logF(x_i) + (1 - logF(x_{n+1-i}))]$$

Another Argument

- $n\hat{F}(x) \sim Binomial\big(n, F(x)\big)$
- $Var\big[n\hat{F}(x)\big] = nF(x)\big(1 - F(x)\big)$
- Variance stabilization

$$AD = n\int_{-\infty}^{\infty} \frac{(\hat{F}(x) - F(x))^2}{F(x)[1 - F(x)]} dF(x)$$

Anderson, T. W.; Darling, D. A. (1952). "Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes". Annals of Mathematical Statistics. 23: 193–212.

None of these allows estimation of parameters except in some special cases:

$H_0: X \sim Normal$

Hubert Lilliefors (1967), "*On the Kolmogorov–Smirnov test for normality with mean and variance unknown*", Journal of the American Statistical Association, Vol. 62. pp. 399–402.

$H_0: X \sim Exponential$

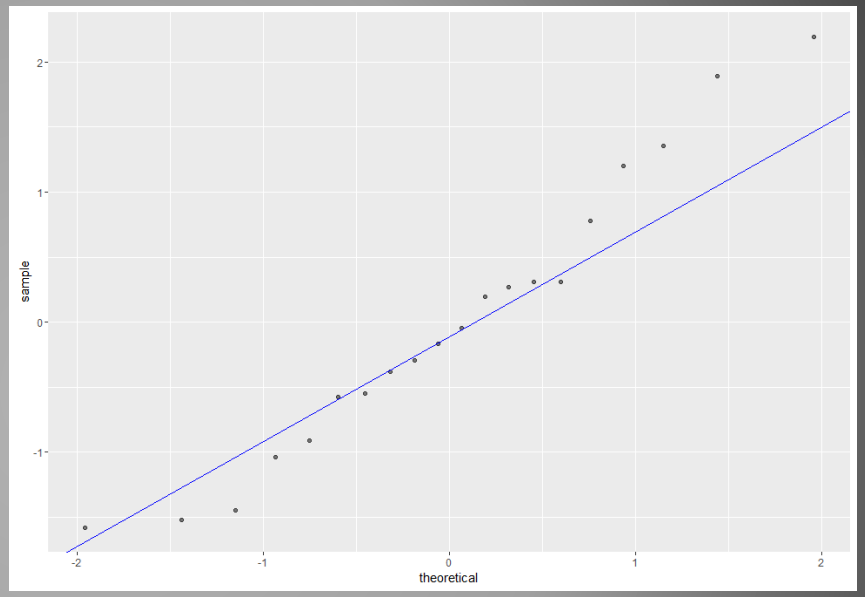
Hubert Lilliefors (1969), "*On the Kolmogorov–Smirnov test for the exponential distribution with mean unknown*", Journal of the American Statistical Association, Vol. 64 . pp. 387–389.

But then again, just find null distribution via Monte Carlo!

Methods based on Probability Plots

Plot quantiles of F vs sample quantiles

If F is correct model, points form a straight line



Turn this into a formal test

Again Probability Integral Transform:

$$X \sim F \;\rightarrow\; F(X) \sim U[0,1]$$

$$(U_1, .., U_n)\ iid\ U[0,1]$$

Order Statistic $U_{(1)} < \ldots < U_{(n)}$

$$U_{(k)} \sim Beta(k, n-k+1)$$

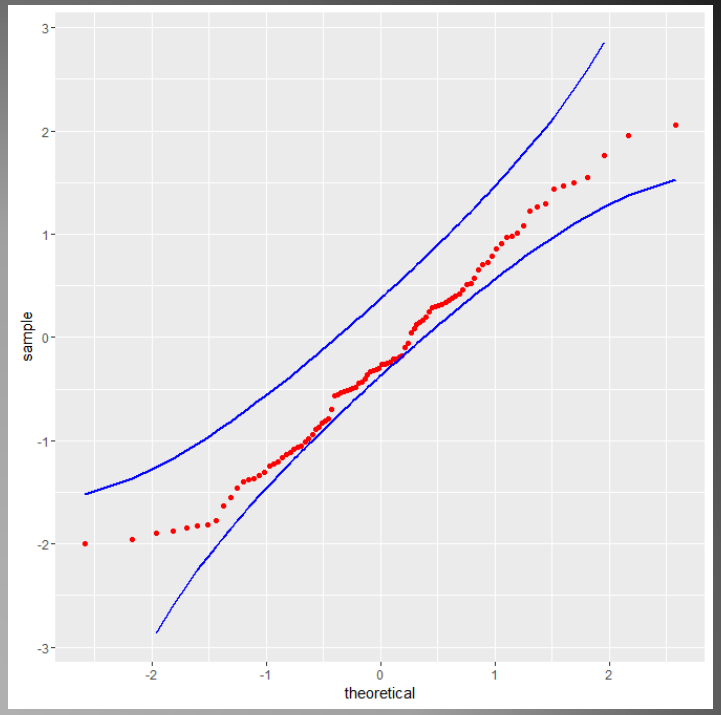
Find pointwise confidence intervals from quantiles of Beta distribution

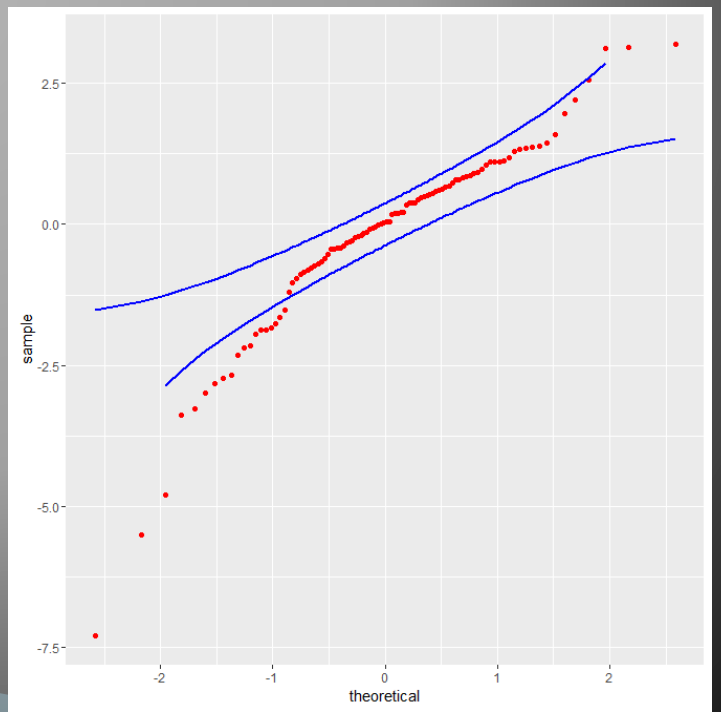
Turn into simultaneous confidence band by adjusting nominal confidence level via MC.

Sivan Aldor-Noima, Lawrence D. Brown, Andreas Buja , Robert A. Stine and Wolfgang Rolke, "*The Power to See: A New Graphical Test of Normality*", The American Statistician (2013), Vol 67/4

Andreas Buja, Wolfgang Rolke "*Calibration for Simultaneity: (Re) Sampling Methods for Simultaneous Inference with Applications to Function Estimation and Functional Data*", Technical Report, Wharton School of Business, Univ. of Pennsylvania

R routines: http://academic.uprm.edu/wrolke/research/publications.html





Smooth Tests

Old idea – goes back to Neyman (1937) – but with some recent improvements.

Basic idea: embed density f in family of densities $\{g_k\}$ indexed by some parameter vector $\Theta = (\theta_1, .., \theta_k)$ which includes true density for some k and such that

H_0: *true density is* $f \leftrightarrow H_0$: $\Theta = \mathbf{0}$

$$g_k(x;\theta,\beta) = C(\theta,\beta)\exp\left\{\sum_{j=1}^{k}\theta_j h_j(x;\beta)\right\} f(x;\beta)$$

$\{h_j\}$ should be orthonormal family of functions, i.e.

$$\int_{-\infty}^{\infty} h_i(x)h_j(x)dx = \delta_{ij}$$

optimal choice of $\{h_j\}$ depends on f, so different tests for different null hypotheses.

Typical choices for $\{h_j\}$:
Legendre Polynomials, Fourier series,
$h_j(x)=\sqrt{2}\cos(j\pi x)$, Haar functions, ….

Basics of the test:

$$U_j = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h_j(X_i)$$

$$T_k = \sum_{j=1}^{k} U_j$$

$$T_k \rightarrow_d \chi_k^2$$

Interesting feature: partial tests $(\theta_1,..,\theta_m) = 0$ for m<k can give insight into HOW null is wrong.

Zhang's Tests

Not so well known, but often have good power.

$$Z_K = \max_{1 \le i \le n} \left((i - \frac{1}{2}) \log \left\{ \frac{i - \frac{1}{2}}{n F_0(X_{(i)})} \right\} + (n - i + \frac{1}{2}) \log \left\{ \frac{n - i + \frac{1}{2}}{n\{1 - F_0(X_{(i)})\}} \right\} \right)$$

$$Z_A = -\sum_{i=1}^{n} \left[\frac{\log F_0(X_{(i)})}{n - i + \frac{1}{2}} + \frac{\log\{1 - F_0(X_{(i)})\}}{i - \frac{1}{2}} \right]$$

$$Z_C = \sum_{i=1}^{n} \left[\log \left\{ \frac{F_0(X_{(i)})^{-1}}{(n - \frac{1}{2})/(i - \frac{3}{4}) - 1} \right\} \right]^2$$

Jin Zhang, "*Powerful Goodness-of-Fit Tests Based on the Likelihood Ratio*", Journal of the Royal Statistical Society. Series B (Statistical Methodology), Vol. 64,No. 2 (2002), pp. 281-294

The distributions of all three test statistics need to be found via MC.

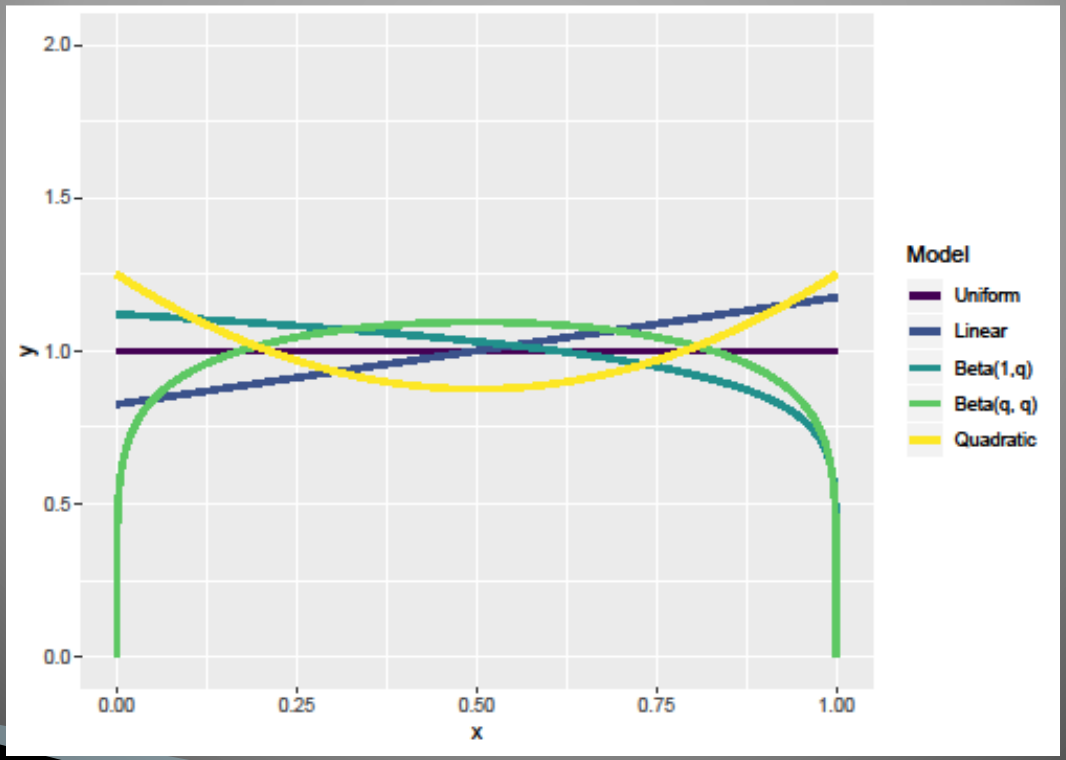
And many more…

- Tests based on moments
- Test based on Wasserstein distance
- Tests specific for a distribution (Normal: more than 30 tests)
- A good place to start: "*Comparing Distributions*", Olivier Thais, Springer

Coffee break

So, how do they do?

$H_0 : F$ =U[0,1] ; n=1000, $\alpha = 0.05$
In all cases highest power ≈80-90%



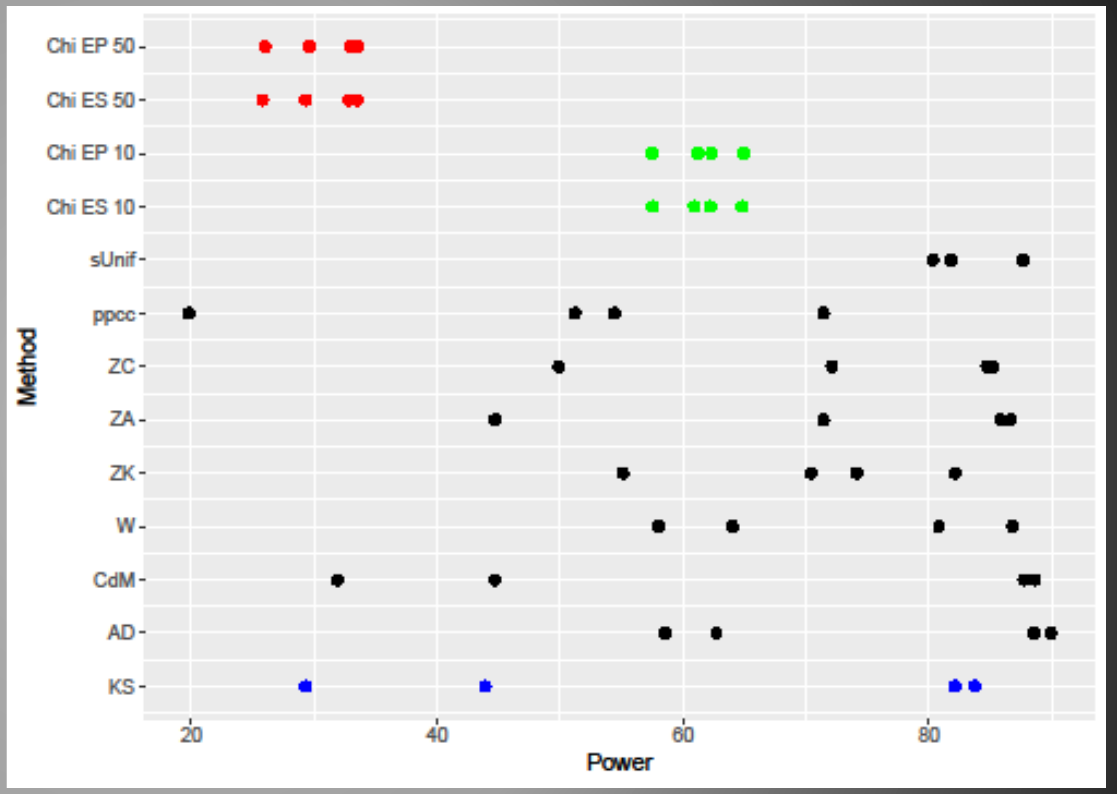
It's a mess!

Any one method might have good power in one case and bad power in another.

Chi-square with large number of bins always bad.

Chis-square with low number of bins better but not great.

KS at least sometimes very bad.



"*Simultaneous Goodness-of-Fit Testing*",
W. Rolke (2020): 21 such studies
(https://arxiv.org/abs/2007.04727).

Most methods sometimes good, sometimes bad.

Chi-square and KS: never very good.

Chi-square with large number of bins (>>10): horrible!

AD and Zhang's Z_C generally quite good.

An obvious Idea:

Do several tests!

If none of them reject the model, it can't be that bad.

But: look-elsewhere-effect

Take a couple of looks effect?

↦ simultaneous inference

Say we perform k tests, each at the α level, and assume model is good. Let T_i be test i rejects null, then:

P(at least one test rejects null) =

$$1 - \text{Prob}(T_i^c; i = 1, .., k)$$

Easy if tests are independent:

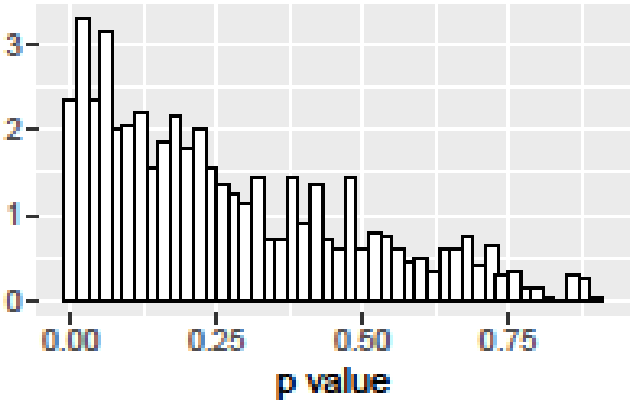
$$1 - \text{Prob}(T_i^c; i = 1, .., k) =$$
$$1 - \prod \text{Prob}(T_i^c) =$$
$$1 - \prod (1 - \alpha) = 1 - (1 - \alpha)^k$$

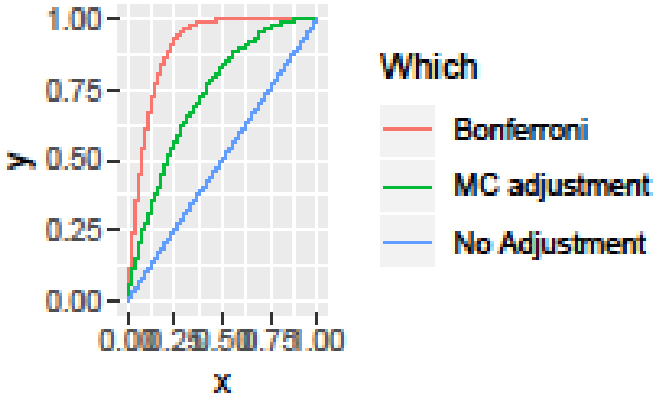
↦ Bonferroni correction

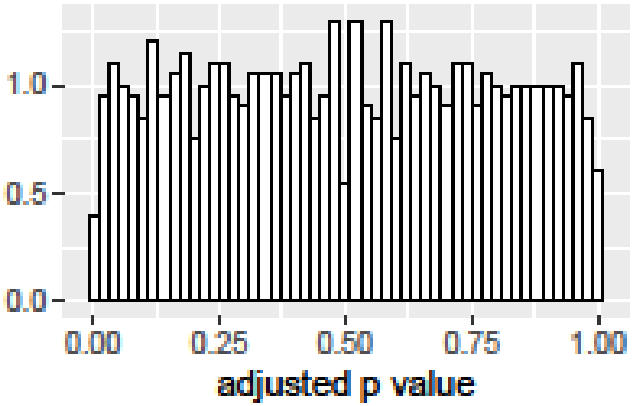
But our tests are not independent, they all use the same data.

We can still find correction using simulation!

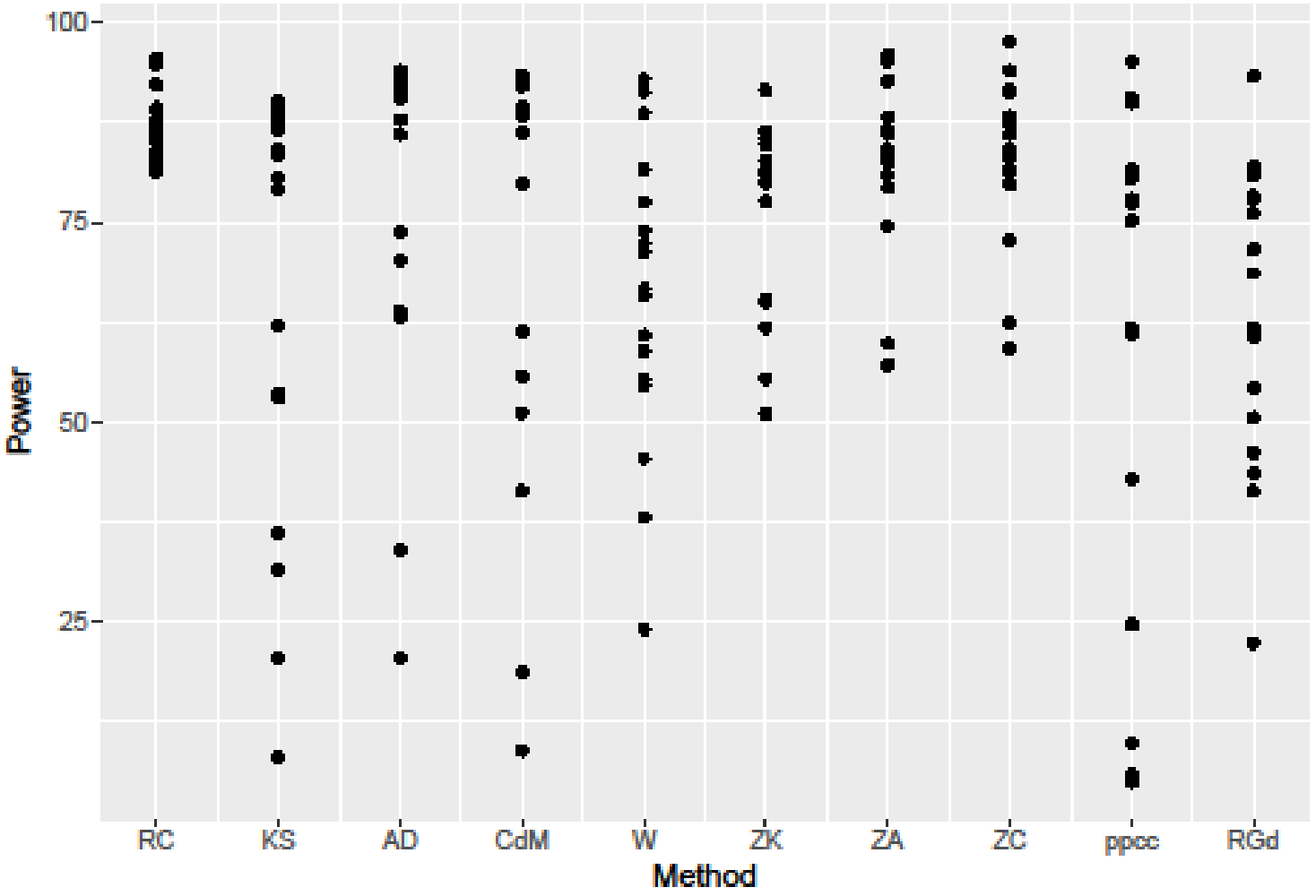
Example: $H_0: X \sim U[0,1]$, use 9 tests:







Results over 21 Studies



How to run this test:

- R package *simgof (*available on CRAN)

```
> library(simgof)
>  x <- rnorm(1000, 100, 20)
>  pnull <- function(x, param) pnorm(x, param[1], param[2])
>  rnull <- function(n, param) rnorm(x, param[1], param[2])
>  qnull <- function(x, param) qnorm(x, param[1], param[2])
>  estimate <- function(x) c(mean(x), sd(x))
>  simgof.test(x, pnull, rnull, qnull, TRUE, estimate)
     RC      KS      AD     CdM       W      ZA      ZK      ZC
0.7572 0.4220 0.6020 0.5450 0.5070 0.8010 0.9110 0.7060
>
```

https://drrolke.shinyapps.io/sgoftest

Simultaneous Goodness-of-Fit Test

Enter all the information required and then hit Go. For a detailed explanation of the app go here

Go

Data is ... Continuous

Sample size is .. fixed

Number of Simulation Runs 10000

Upload file with data Browse... normal.data.txt — Upload complete

Upload file with Routines Browse... normalC.est.txt — Upload complete

General Methods

☑ KS ☑ AD ☑ CdM ☑ W ☑ ZK ☑ ZA ☑ ZC

Normal Distribution

☑ ppcc ☑ SW ☑ JB ☑ Nor

Uniform

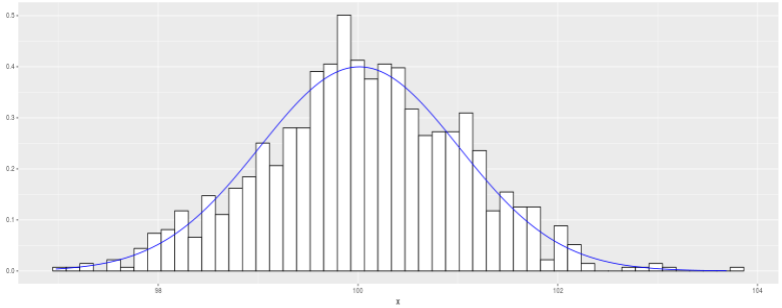
☐ Unif

Exponential

☐ Exp

Chisquare Methods

☐ RGd ☐ Equal Size ☐ Equal Prob



Method	p value
RC	0.6386
KS	0.2932
AD	0.3115
W	0.3504
ppcc	0.3516
ZK	0.3606
CdM	0.3636
SW	0.4106
sNor	0.4158
ZA	0.4628
ZC	0.541
JB	0.9223

Parameter Estimate(s):
100.047
, 0.998

Tests for Multidimensional Data

In principle very useful, but:

Curse of Dimensionality (R. Bellman)

Example: H_0: $(X_1,..,X_d) \sim U[0,1]^d$
We want to do a χ^2 test and we want 10 bins in each dimension. What n do we need to get $E \geq 5$?

d=1: $E = {}^{n}/_{10} \cong 5 \rightarrow n \cong 50$
d=2: $E = {}^{n}/_{10^2} \cong 5 \rightarrow n \cong 500$
d=3: $E = {}^{n}/_{10^3} \cong 5 \rightarrow n \cong 5000$
…
d=10: $E = {}^{n}/_{10^{10}} \cong 5 \rightarrow n \cong 50$ billion

Some other tests not so extreme, but all of them suffer to some degree from the curse.

High-dimensional space is strange!

$X_1, .., X_n \sim N(0,1)$, independent

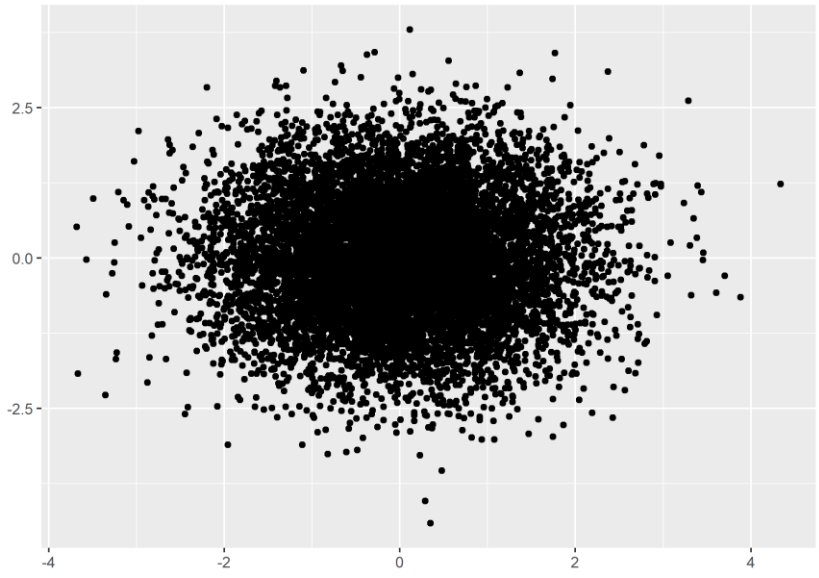
$\boldsymbol{x} = (x_1, .., x_n) \in \mathbb{R}$

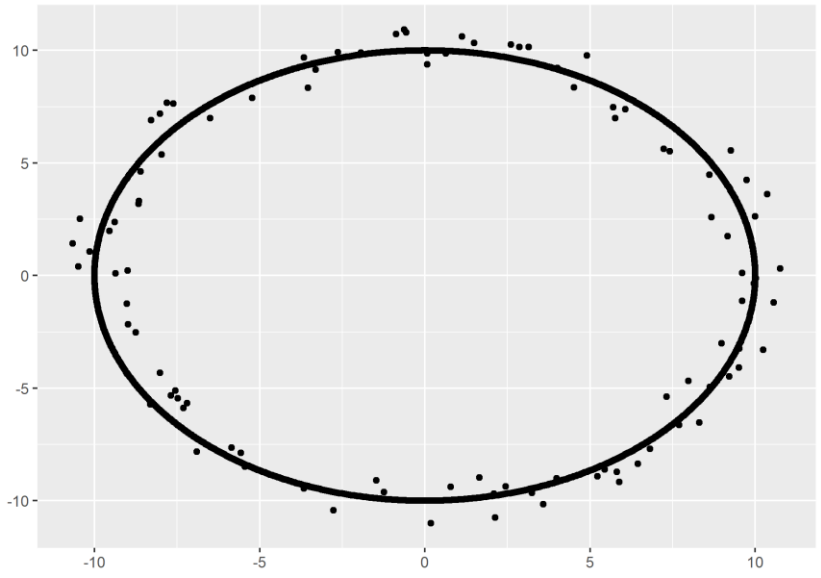
$D = \sqrt{x_1^2 + .. + x_n^2}$ Euclidean distance to origin

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$$

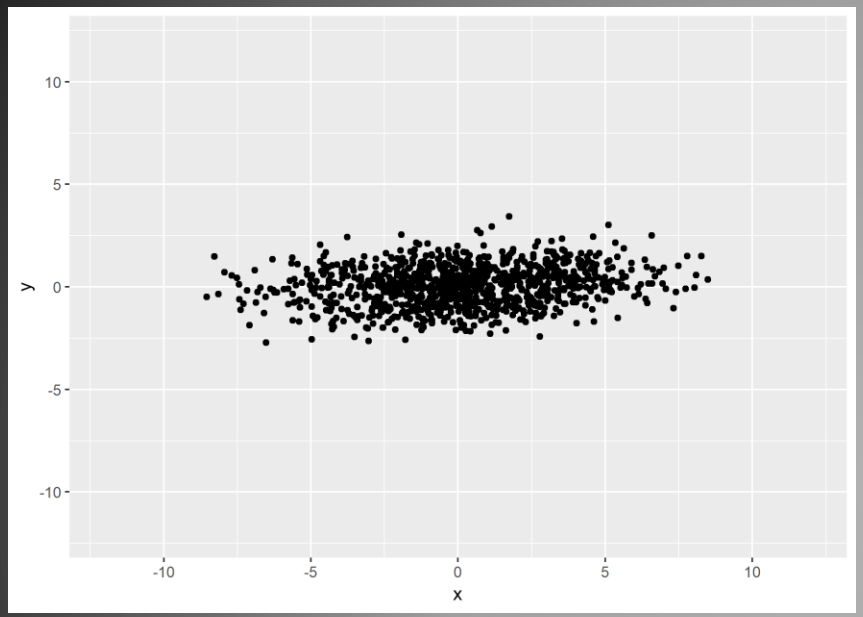
$$E[D] \approx \sqrt{E[D^2]} = \sqrt{n}$$

$$P(|D - \sqrt{n}| > t) \leq 2\exp\{-ct^2\}$$ Hoeffding bound





First: Standardize!



Usual: $\frac{x-\bar{x}}{sd(x)}$

Robust: $\frac{x-\text{Med}(x)}{\text{IQR}(x)}$

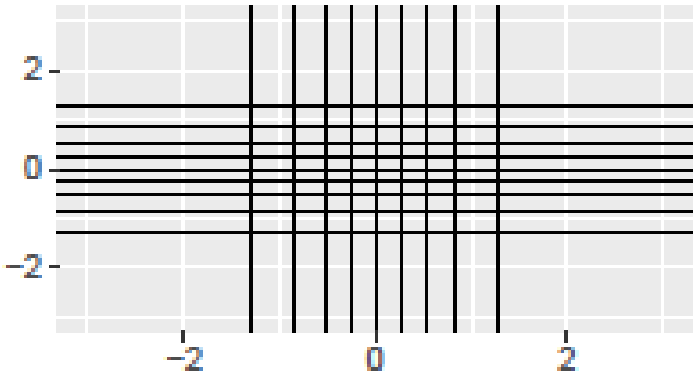
0-1: $\frac{x-\min x}{\max(x)-\min(x)}$

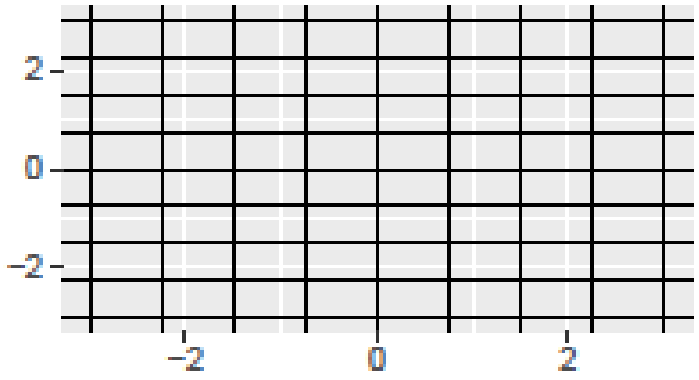
$(\text{IQR} = \text{Inter Quartile Range } = P_{75} - P_{25})$

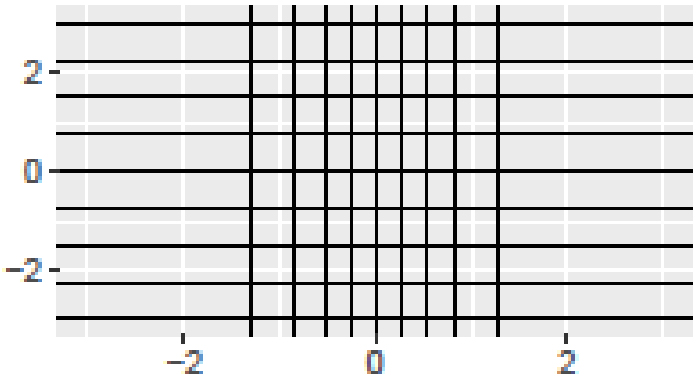
Some methods do this automatically.

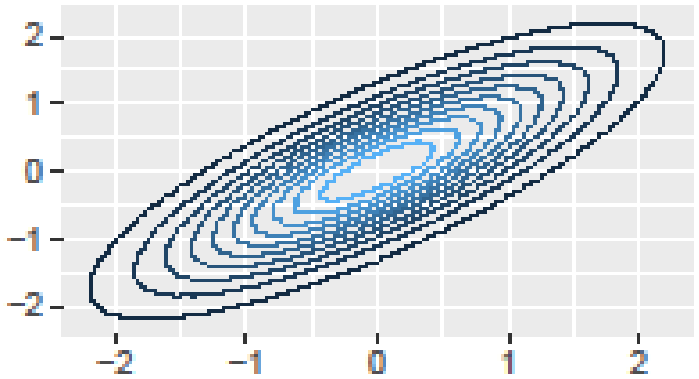
Destroys any analytic null distribution.

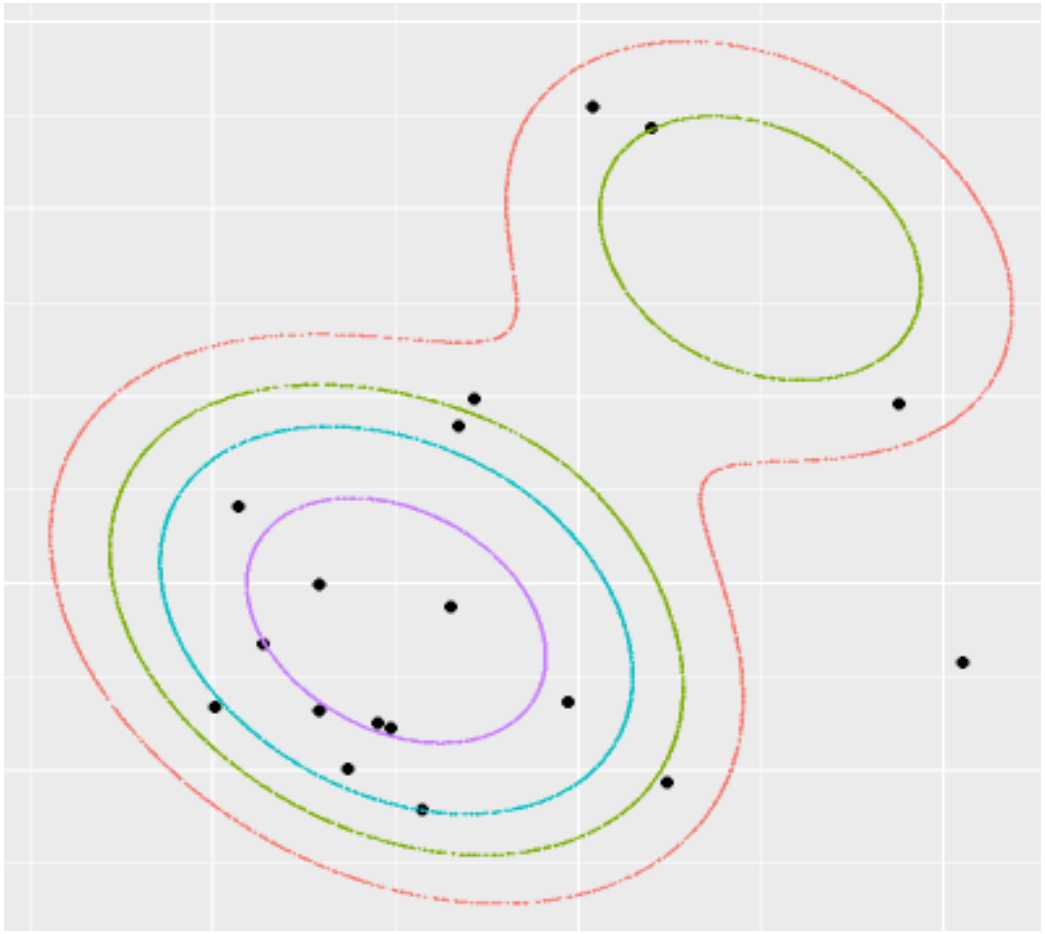
χ^2 Test: How to bin?



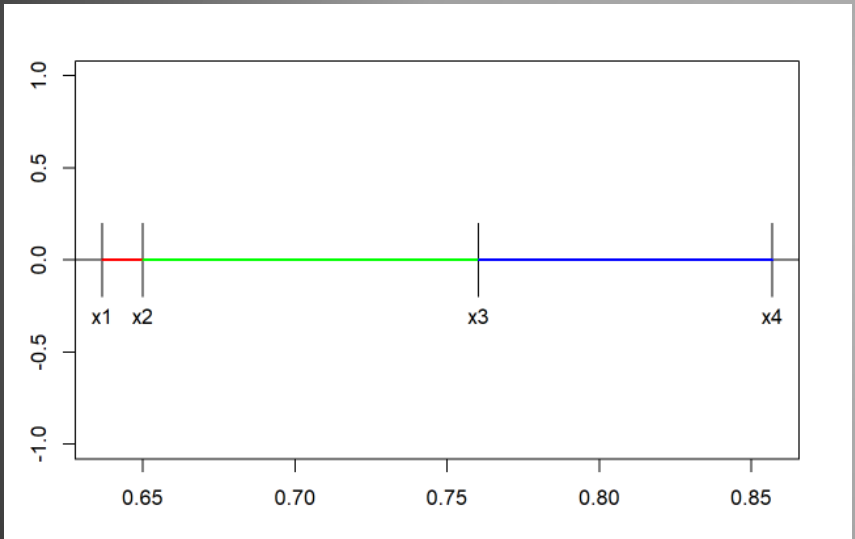






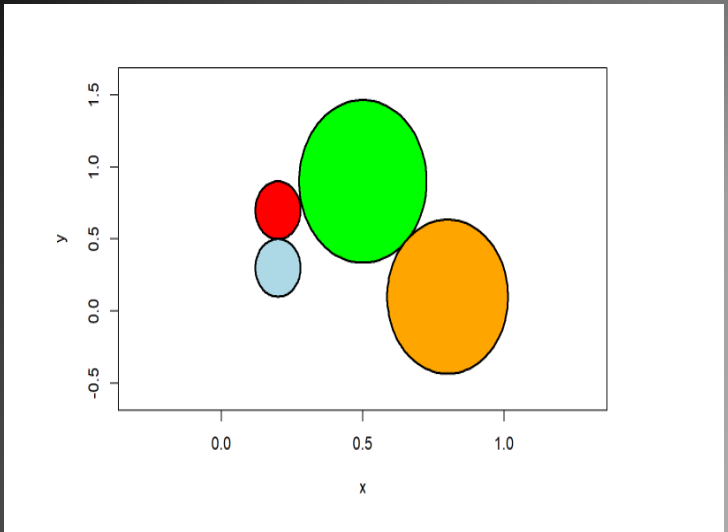


Tests based on Spacings



Under null hypothesis transformed spacings have uniform distributions.

Closely related to nearest-neighbors



Hyperspheres in R^d

- Bickel, P.J., Breiman. L (1983) *Sums of functions of nearest neighbor distances, limit theorems and goodness of fit test*, Ann. Prob. **11**, 185-214.
- Schilling. M (1983), *Goodness of Fit Testing in Rm Based on the Weights Empirical Distribution of Certain Nearest Neighbor Statistics*, Ann. ff Statistics **11**, 1-12.
- Schilling. M (1983), An infinite-dimensional approximation to the nearest neighbor goodness-of-fit tests, Ann. Of Statistics **11**, 13-24
- Hall. P, (1986) *On Powerful Distribution Tests Based on Sample Spacings*, J. of Multivariate Analysis **19**, 201-224.

More Nearest Neighbor

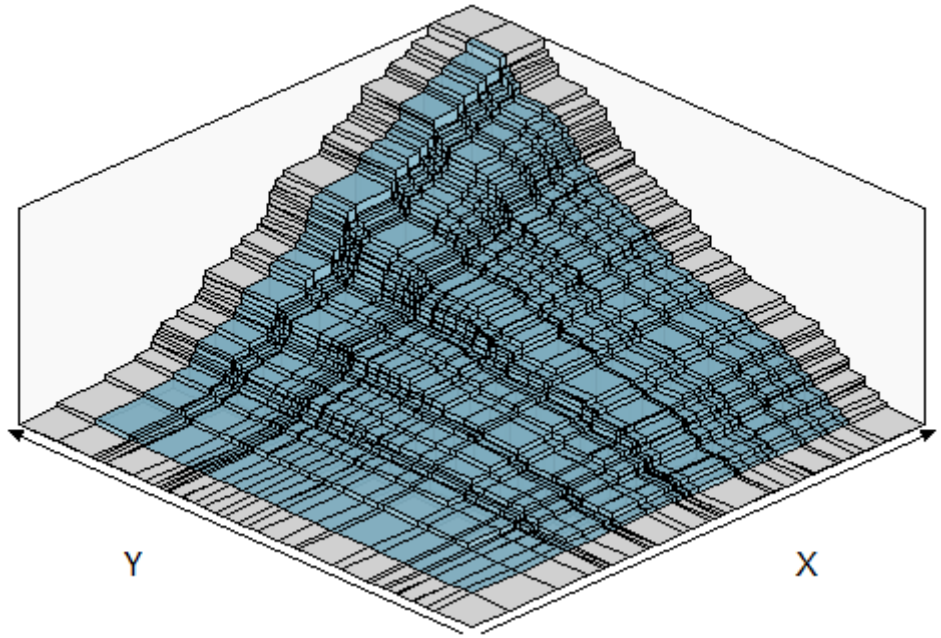
Ilya Narsky (2003), *Estimation of Goodness-of-Fit in Multidimensional Analysis Using Distance to Nearest Neighbor*, arXiv:physics/0306171

Presented at Phystat 2003 – SLAC

Based on Rosenblatt transform and Monte Carlo.

Rosenblatt transform imposes artificial order on variables. In d dimensions there are d! ways to go.

Tests based on EDF



Analytic derivation of null distribution also based on Rosenblatt transform, same issue of order.

These days test statistic can be found directly, but needs a lot of calculations.

KS: Sample size n$\mapsto n^2/4$ function evaluations

Simple Idea: Just look at data points $\mapsto$ fKS

Not as powerful as KS, but not bad either, and much faster.

Anderson-Darling can also be generalized but requires numerical calculation of many integrals.

Literature

- Lopes.RHC, Reid. I and Hobson. PR (2007) *The two-dimensional Kolmogorov-Smirnov test*. Proc. XI Int. Workshop on Advanced Computing and Analysis Techniques in Physics Research April 23-27.

- Fasano, G and Franceschini. A (1987) *A multidimensional version of the Kolmogorov-Smirnov test*, Mon. Not R ast. Soc **225**, 155-170

- Lopes. RHC et al (2008), *Computationally efficient algorithms for the two-dimensional Kolmogorov-Smirnov test*, J. Phys. Conf. Ser, **119**

- Peacock. JA (1983) *Two-dimensional goodness-of-fit testing in astronomy,* Mon. Not. R. Astron. Soc. **202** 615-627

Tests based on Characteristic Functions

Characteristic Function of a Random Vector

$$\phi(t_1,..,t_k) = \mathbb{E}\left[e^{\,i\,t_1X_1+..+t_kX_k}\right]$$

empirical characteristic function

$$\hat{\phi}(t_1,..,t_k) = \frac{1}{n}\sum_{i=1}^{n} e^{\,i\,t_1x_1+..+i\,t_kx_k}$$

$$T_n \approx \left\{\phi(t_1,..,t_k) - \hat{\phi}(t_1,..,t_k)\right\}^2$$

Problem: choice of $t_1,..,t_k$ crucial for power, not obvious

Yanqin Fan, (1997), *Goodness-of-Fit Tests for a Multivariate Distribution by the Empirical Characteristic Function*, Journal of Multivariate Analysis, 62, 36-63

Aslan–Zech Energy tests

Data: $\boldsymbol{x}_1, .., \boldsymbol{x}_n$

Data simulated from F : $\boldsymbol{t}_1, .., \boldsymbol{t}_m$

$$\varphi = \frac{1}{n^2} \sum_{i<j} R(\| \boldsymbol{x}_i - \boldsymbol{x}_j \|) - \frac{1}{nm} \sum_{i,j} R\big(\| \boldsymbol{t}_i - \boldsymbol{x}_j \|\big)$$

R correlation function:

$$R_k(r) = \frac{1}{r^k}$$

$$R_l(r) = -\log r$$

$$R_s(r) = \exp(-r^2/(2s^2)$$

Neyman smooth tests

$$g_k(x;\theta,\beta) = C(\theta,\beta)\exp\left\{\sum_{i,j=1}^{k}\theta_i\theta_j h_i(x;\beta)h_j(x;\beta)\right\}f(x;\beta)$$

In principle easy, but:

Choice of k?

Same basis functions in different dimensions?

For good power basis functions need to "match" F.

General Comments:

GOF tests beyond 2 or 3 dimensions unlikely to be very powerful.

At the very least will require gigantic data sets to get reasonable power.

No easy to use computer programs

Still a wide-open problem!

Special Cases and Closely Related Problems

1) Model Selection
2) Data is truncated
3) Sample size is random
4) Data is discrete
5) Data is binned
6) MC vs Sample

GOF ≠ Model Selection

Note above: no alternative hypothesis H_1

Different problem:

$H_0: F = flat$ vs $H_0: F = linear$

→ model selection

Usually better tests: likelihood ratio test, F tests, BIC etc.

Easy to confuse: all GOF papers do power studies, those need specific alternative.

Say we test

$$H_0: F = flat \text{ vs } H_0: F = linear$$

and we reject the null, but only because we have 1 million events.

Remember, all models are wrong!

Maybe true model is proportional to $x^{1.5}$!

How does choice effect final result?

Maybe better to use both flat and linear and find out

→ sensitivity analysis
→ systematic errors

Truncated Data

Data in High Energy Physics is always truncated to a finite interval.

Care needs to be taken with normalization

(aka $\int_{-\infty}^{\infty} f(x)dx = 1$)

Statisticians (and their methods) usually will assume this is done automatically and at all times.

Random Sample Size

In HEP experiments sample size is not fixed a-priori but is a consequence of the run time

$$n \sim Poisson(\lambda)$$

If n is fixed: $(N_1, .., N_k) \sim Multinomial(n, p_1, .., p_k)$

But if n is Poisson

$$N_i \sim Poisson(\lambda p_i) \; and \; N_1, .., N_k \; independent!$$

(Theory of Marked Poisson processes)

Consequence: $X^2 \sim \chi^2(k - m)$ (not k-m-1)

Not an issue if null distribution is found via MC

Discrete Data

Does my data come from a Poisson distribution?

Chi-square works just as before (but again, don't use to many classes)

EDF based tests (KS, AD, Zhang etc) all still work but require different formulas and p-values have to be found via mini MC.

Binned Data

Data in HEP is often already binned for various reasons, for example detector resolution.

Note: binned data ≠ discrete data

Still need to consider rebinning for chi square tests.

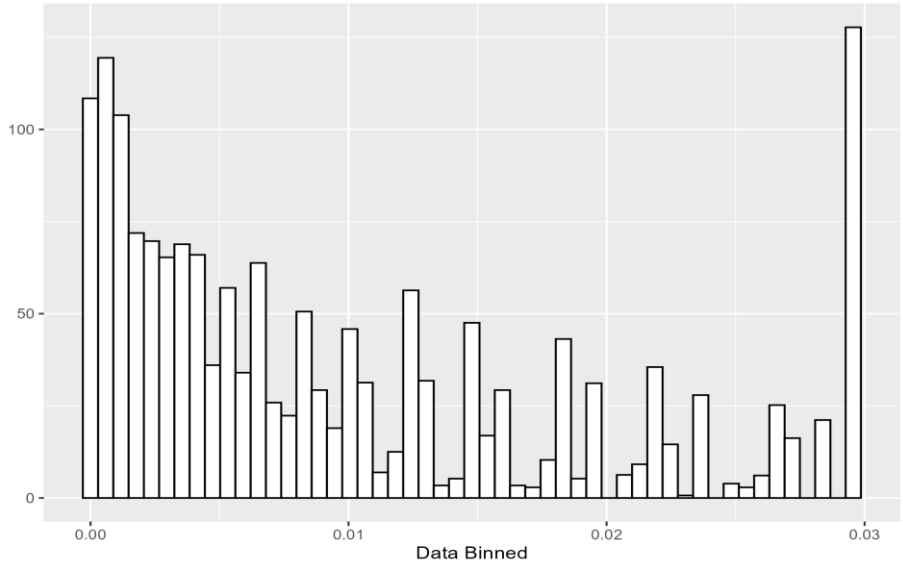
How about tests that require continuous data?

$$KS = max\left\{\left|\frac{i}{n} - F\left(X_{(i)}\right)\right|, \left|F\left(X_{(i)}\right) - \frac{i-1}{n}\right|\right\}$$

But we only know $b_i < X_{(i)} < b_{i+1}$

Obvious answer: $p_i = \frac{b_i + b_{i+1}}{2}$ midpoints, repeat each according to bin counts, run continuous test.

Simulation: generate 1000 Exp(1), bin into 50 equal sized bins, do test as above



Need to use simulation to find p values

Other ways to go:

spread out n_i points in (b_i, b_{i+1}) uniformly.

spread out n_i points in (b_i, b_{i+1}) according to F (can be computer intensive).

Discretize distribution function:

$$p_i = F(b_{i+1}) - F(b_i)$$

Now run discrete version of test (if available)

What's best? Not clear …

MC vs Sample

Today we often don't have F. Note that as long as we can at least calculate values from F (machine learning etc) that's ok. But sometimes we can't even do that.

We can however simulate from F (Monte Carlo)

Question now: does our MC agree with the data, that is, where they generated by the same (unknown) distribution?

↦ Two-Sample Problem

$$H_0: F_x = F_y$$

Note: not $F_x = F_y = F$, so no probability distribution specified, so also no parameter estimation.

→ in many ways an easier problem than GOF.

Classic version: two-sample t test

$$T = \frac{|\bar{X} - \bar{Y}|}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

$$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

$$T \sim t_{n+m-2}$$

Chi-square test

Say we have k classes/bins. x is the counts of x data, y counts of y data, then

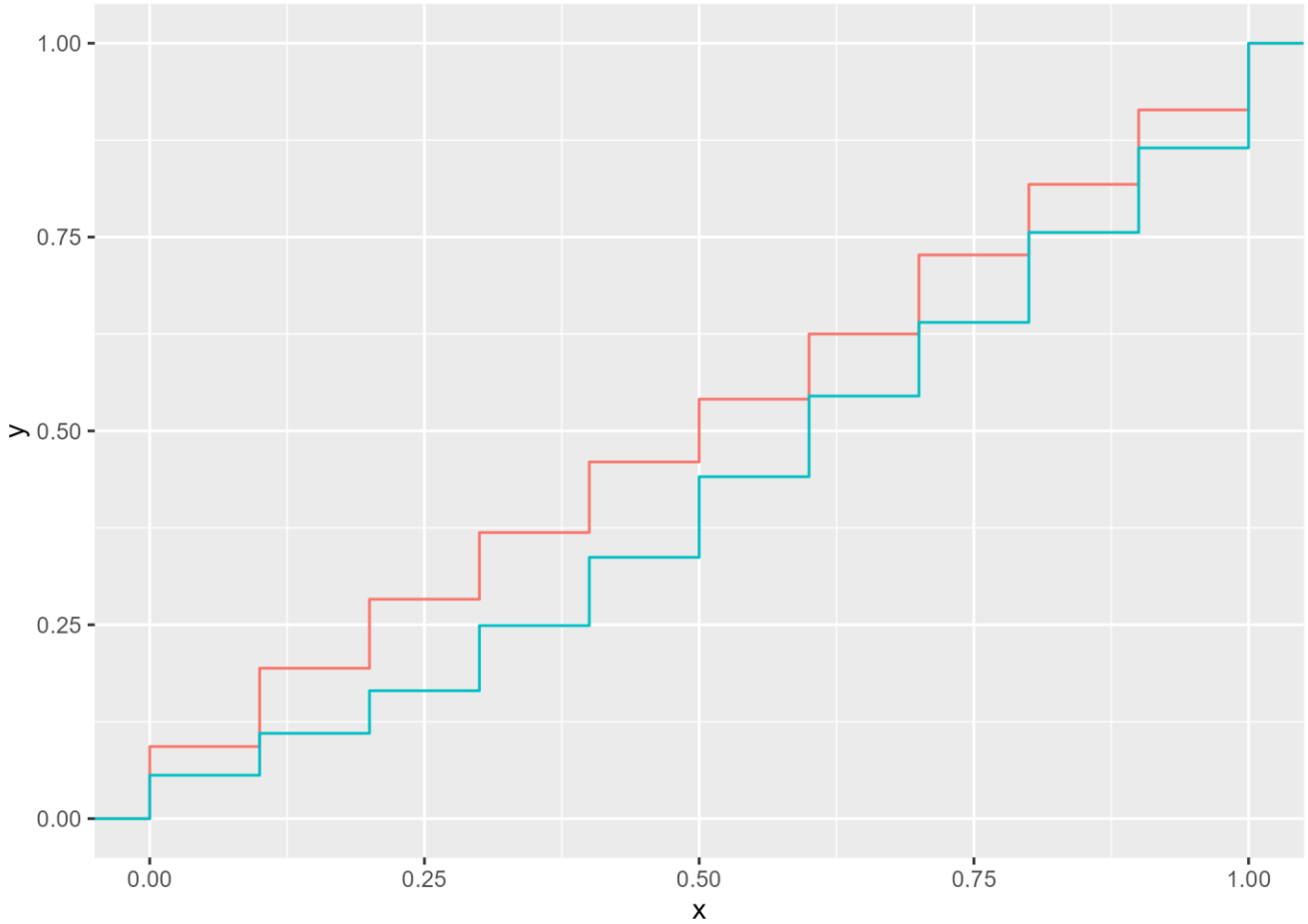
$$n = \sum x_i, m = \sum y_i$$

$$p_i = \frac{x_i + y_i}{n + m}$$

$$X^2 = \sum_{i=1}^{k} \frac{(x_i - np_i)^2}{np_i} + \frac{(y_i - mp_i)^2}{mp_i}$$

$$X^2 \sim \chi^2(k)$$

ECDF based tests



KS and AD tests:

$$KS = \max\{|\frac{1}{n}\sum_{i=1}^{j} x_i - \frac{1}{m}\sum_{i=1}^{j} y_i| : j = 1, .., k\}$$

$$AD = \frac{1}{mn}\sum_{i=1}^{k-1} \frac{(kc_i - mi)^2}{i(k-i)}$$

---

where c_i is the number of x values less or equal to the combined data set.

Similarly formulas for Zhang's tests.

But no analytic formulas for p values for those.

P values via simulation → Permutation tests

x=(0.1, 0.4, 0.8, 0.9, 1.0, 1.1)
y=(0, 0.2, 0.6, 1.0)
xy=(0.1, 0.4, 0.8, 0.9, 1.0, 1.1, 0, 0.2, 0.6, 1.0)
P(xy) =
(0.6, 0.2, 0.8, 1.0, 0.9, 0.4, 0.0, 0.1, 1.0, 1.1)
P(x) = (0.6, 0.2, 0.8, 1.0, 0.9, 0.4)
P(y) = (0.0, 0.1, 1.0, 1.1)

Under the null hypothesis these are just as good a data as original. Generate (say) 1000 such data sets and find values of test statistics. Compare to real data.

Not a new idea: Fisher (1935) "*Design of Experiments*"

How good is this?

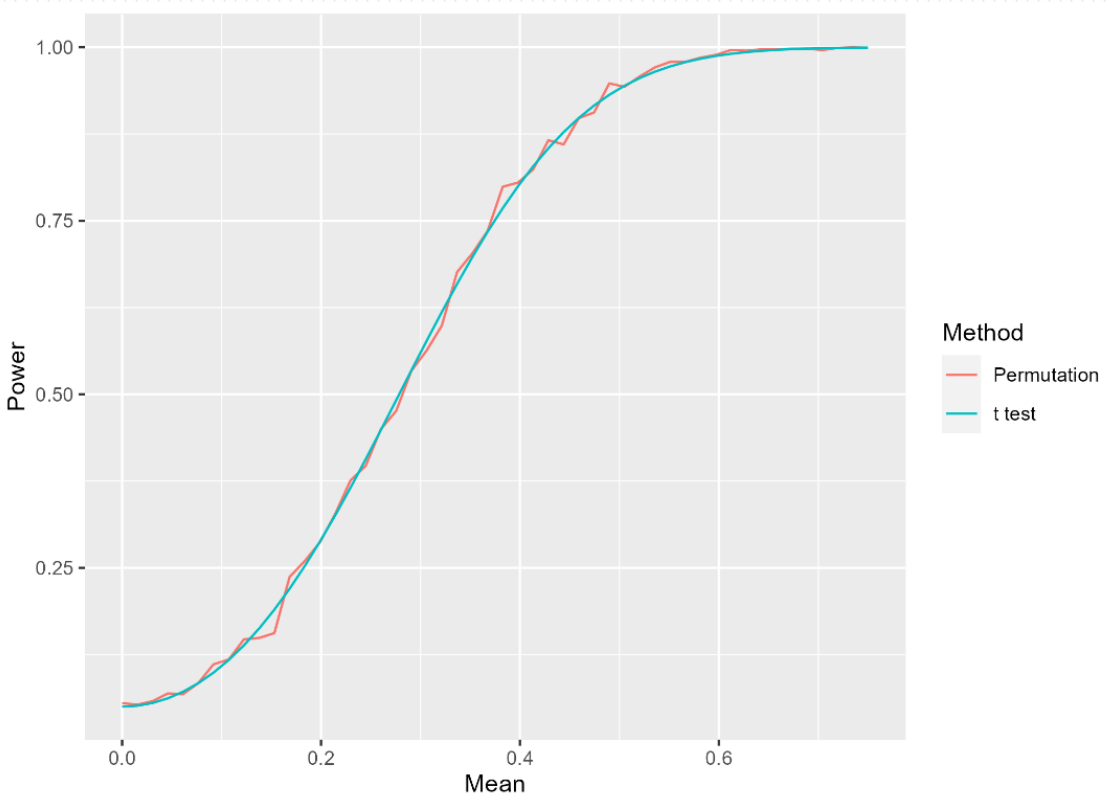
x normal(0,1)

y normal(μ,1)

Sample size 100 each

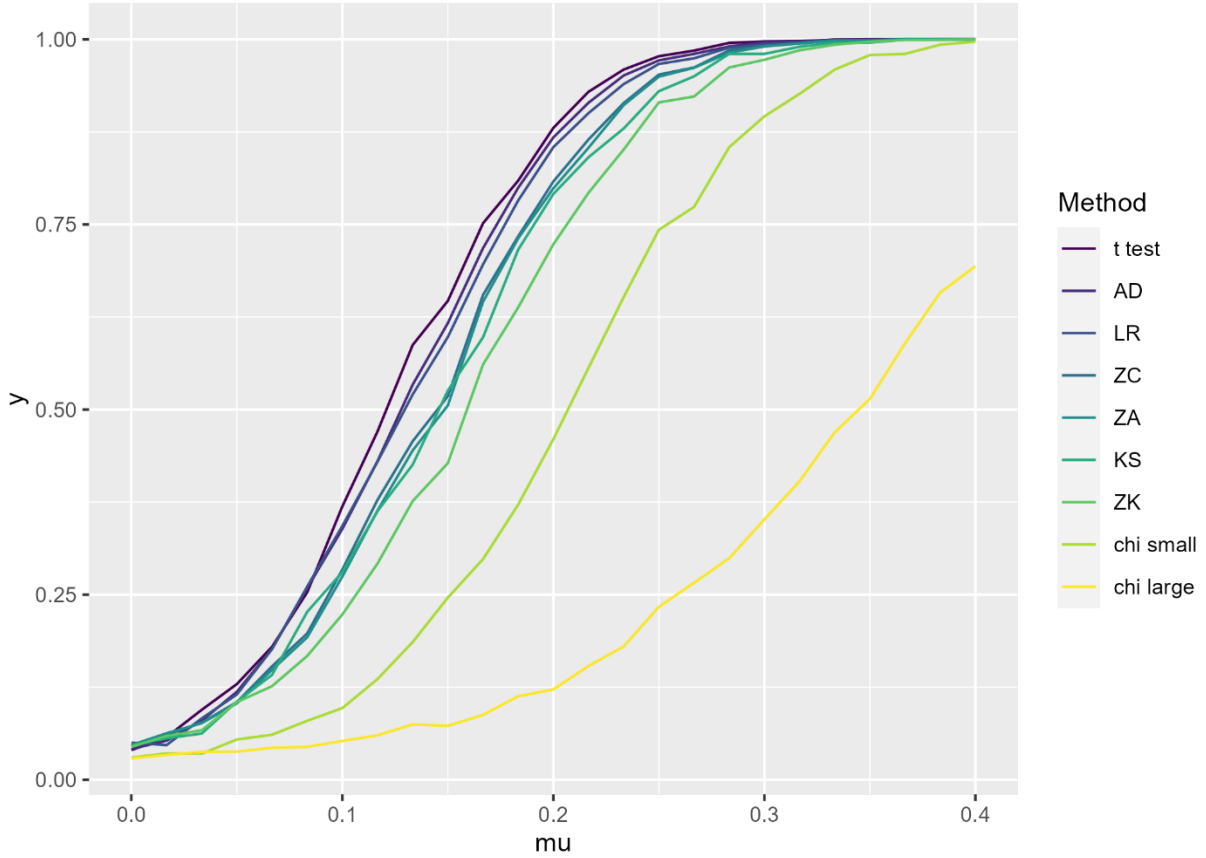
t test is likelihood ratio test, so (near) optimal.

Permutation test just as powerful, without any assumptions!



Application to our two-sample problem

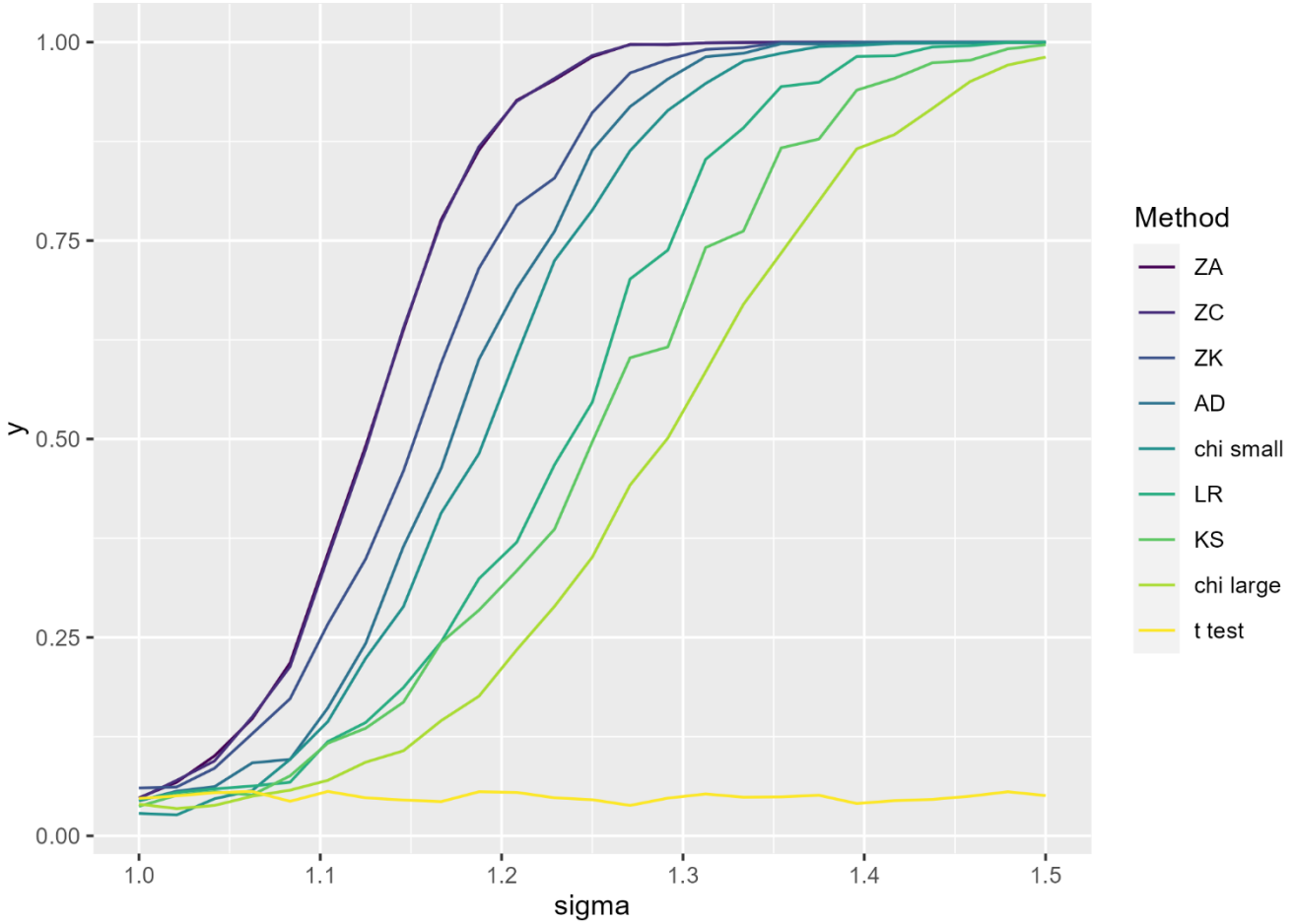
Example (Shift):
Data set 1: 500 events from standard normal
Data set 2: 500 events from normal with other mean.



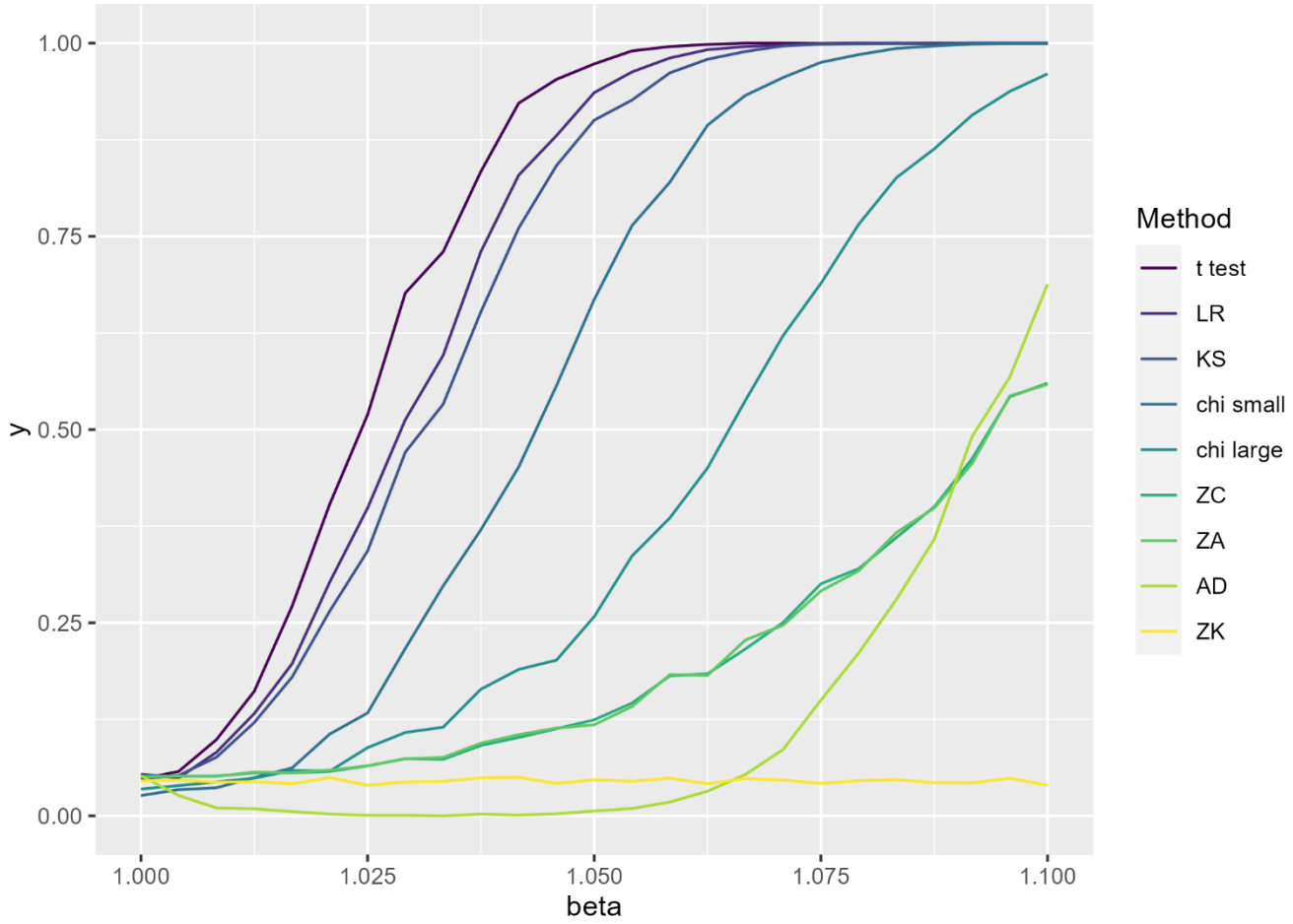
Example (Stretch)

Data set 1: 500 events from standard normal

Data set 2: 500 events from normal with other standard deviation.



Example:
Data set 1: 10000 events from exponential rate 1
Data set 2: 20000 events from normal gamma rate 1, shape
Both data sets are binned into 100 equal probability bins



Relative Sample Sizes

Often when comparing MC with data, it is possible to generate as many MC events as we want, whereas the size of the data set is fixed.

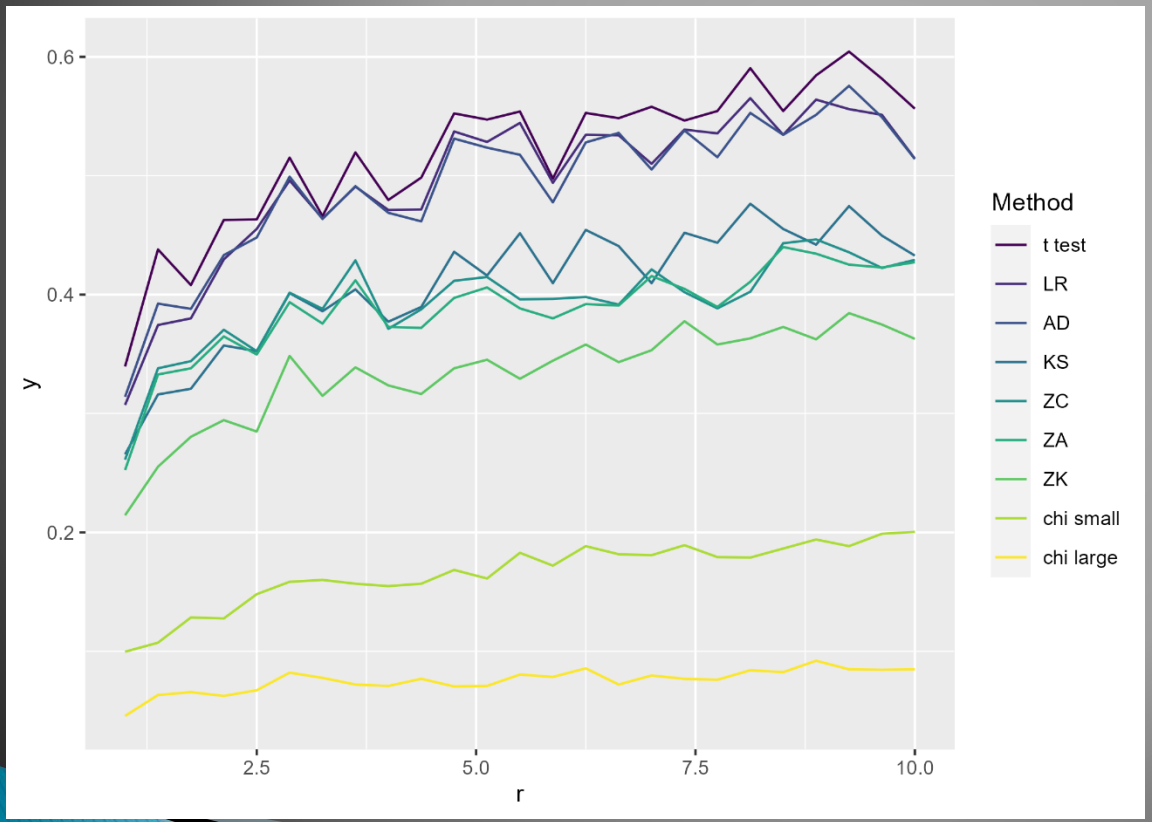
So if the data set has n events, how large should we pick m, the number of MC events?

Example:

Data set 1: 500 events from standard normal

Data set 2: 500r events from normal with mean 0.2

Not much gain in power beyond five-fold



Software for gof

Root:

AD and KS test implemented in the ROT:Math::GoFTest class, uses theoretical distributions for p values.

For binned data the tests are available in the histogram class, as TH1::AndersonDarling and TH1::KolmogorovTest. They assume points at bin centers.

P values can also be found via mini MC.

combine:

KS, AD and Baker-Cousins. For binned data only, parameters are estimated via maximum likelihood and p values found via mini MC.

R: everything …

Conclusions

- GOF testing should be part of (most) statistical analysis.
- Any one test can have low power, so do several.
- Chi-square with large number of bins has very low power.
- Tests for multi-dimensional distributions are not great and likely have low power for much more than two dimensions.
- Related problems such as model selection and two sample require their own methods.

- THANKS!