Testing Goodness of Fit

Dr. Wolfgang Rolke
University of Puerto Rico – Mayaguez

Table of Content

- Overfitting
- Continuous Data
- EDF Methods
- Kolmogorov–Smirnov
- X^2 vs K–S
- Probability Plots
- Smooth Tests
- Multidimensional Data

The Archetypical Statistics Problem:

- There is a theory
- There is data from an experiment
- Does the data agree with the theory?

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

- 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

Use as measure of deviations

$$X^2 = \sum \frac{(O-E)^2}{E}$$

O: observed counts

E: expected counts

Agreement is bad if X^2 is large

But why $\sum \frac{(O-E)^2}{E}$, why not (say) $\sum \frac{(O-E)^2}{O}$ or $\sum |O-E|$ or $\max\{|O-E|\}$?

	1	2	3	4	5	6
O	187	168	161	147	176	161
E	167	167	167	167	167	167

$$X^2 = \frac{(187-167)^2}{167} + .. + \frac{(161-167)^2}{167} = 5.72$$

Is 5.72 “large”?

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

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)

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

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

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

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

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

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

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

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

Here: mean of $\chi^2(5) = 5$

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

In the derivation of the distribution of X^2 we used the CLT approximation, so this needs a sufficiently large sample size. But how large does it have to be?

Famous answer: $E \geq 5$

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.

Hypothesis Testing Basics

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

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

2: The probability of committing the type II error should be as low as possible (subject to 1)

Historically 1 was achieved either by finding an exact test or having a large enough sample.

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: Mathematical, Physical and Engineering Sciences. **231** (694–706):*

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

$$L(p_1,..,p_k) \sim p_1^{N_1}..p_k^{N_k}$$

$$\Lambda = \frac{L(p_1,..,p_k)}{\max\{L(p_1,..,p_k): n_1+..+n_k=n\}} =$$

$$\frac{L(p_1,..,p_k)}{L(N_1/n,..,N_k/n\}} = \frac{p_1^{N_1}..p_k^{N_k}}{\left(\frac{N_1}{n}\right)^{N_1}..\left(\frac{N_k}{n}\right)^{N_k}} =$$

$$\left(\frac{np_1}{N_1}\right)^{N_1}..\left(\frac{np_k}{N_k}\right)^{N_k}$$

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

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

$$-2\log\Lambda = -2\log\left[\left(\frac{np_1}{N_1}\right)^{N_1}..\left(\frac{np_k}{N_k}\right)^{N_k}\right] =$$

$$2\sum n_i \log\frac{N_i}{np_i} =$$

$$2\sum n_i \log\left(\frac{N_i}{np_i} - 1 + 1\right) =$$

$$2\sum n_i \log\left(\frac{N_i - np_i}{np_i} + 1\right)$$

$\log(x+1) \approx x + x^2/2$ Taylor expansion

$$-2\log\Lambda \approx 2\sum n_i\left(\frac{N_i - np_i}{np_i} + \left(\frac{N_i - np_i}{np_i}\right)^2/2\right) =$$

$$2\sum n_i \frac{N_i - np_i}{np_i} + \sum\left(\frac{N_i - np_i}{np_i}\right)^2 \approx X^2$$

because $N_i \approx np_i$, so $N_i - np_i \approx 0$

so $2\sum n_i \frac{N_i - np_i}{np_i} \approx 0$

And that's not all:

Other standard methods for deriving hypothesis tests:

- Wald → X^2 (without any approximations!)
- Rao → X^2

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.

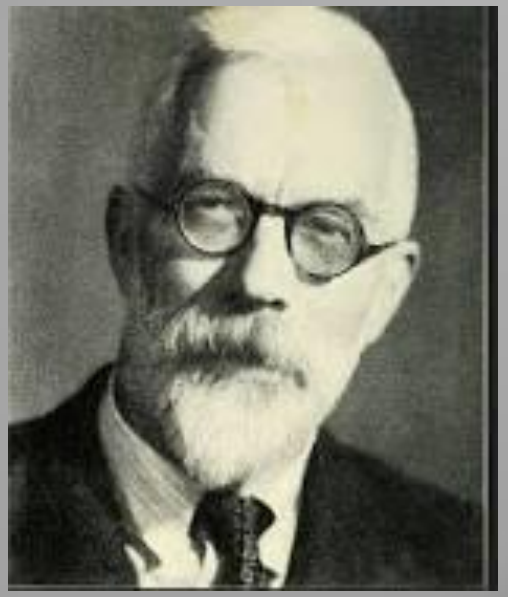
In 1915 Greenwood and Yule publish an analysis of a 2x2 table and note that there is a problem.

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 sample size is used)

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

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"

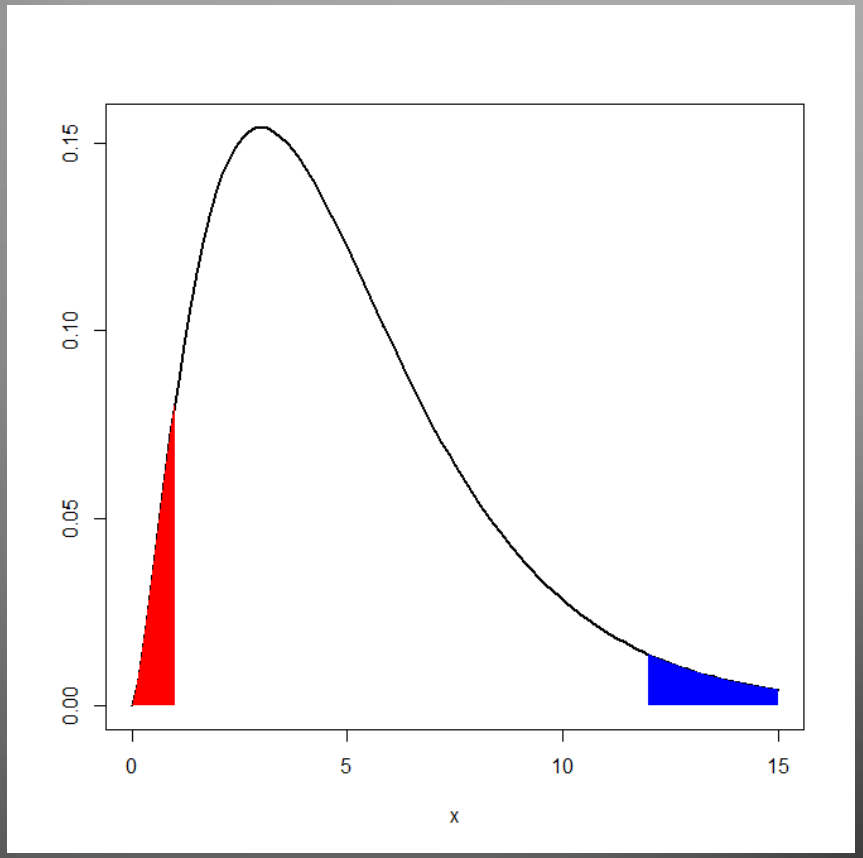
Fisher, R.A. (1936). "Has Mendel's work been rediscovered?" . Annals of Science. 1 (2): 115–137.

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 day 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=T
      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
}
```

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

George Box: All models are wrong, but some are useful

Probability models are theoretical constructs, one can not expect them to be perfect fits in real life (“there is no perfect circle in nature”)

→ how close an agreement between null and data is needed depends on context

→ related to choice of type I error probability α,

5%? 1%?

Overfitting

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

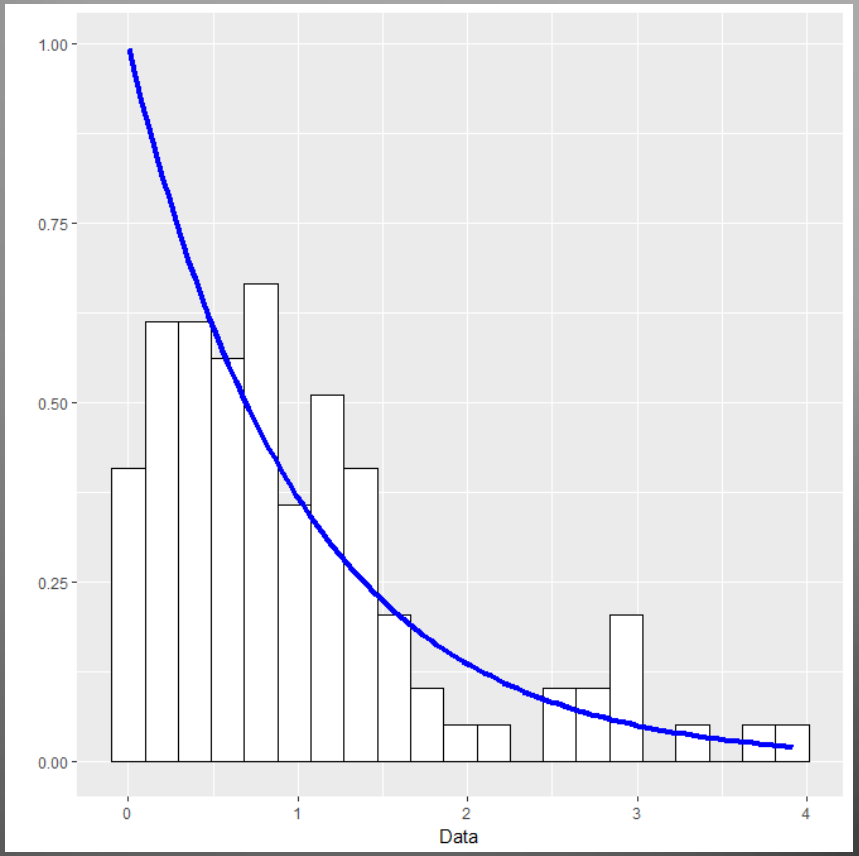
We also should worry about: is our model better than it should be?

- Overfitting!

Exponential Model – Good Fit?

χ^2 (6 bins): p value = 0.111

Good enough!



Continuous Data

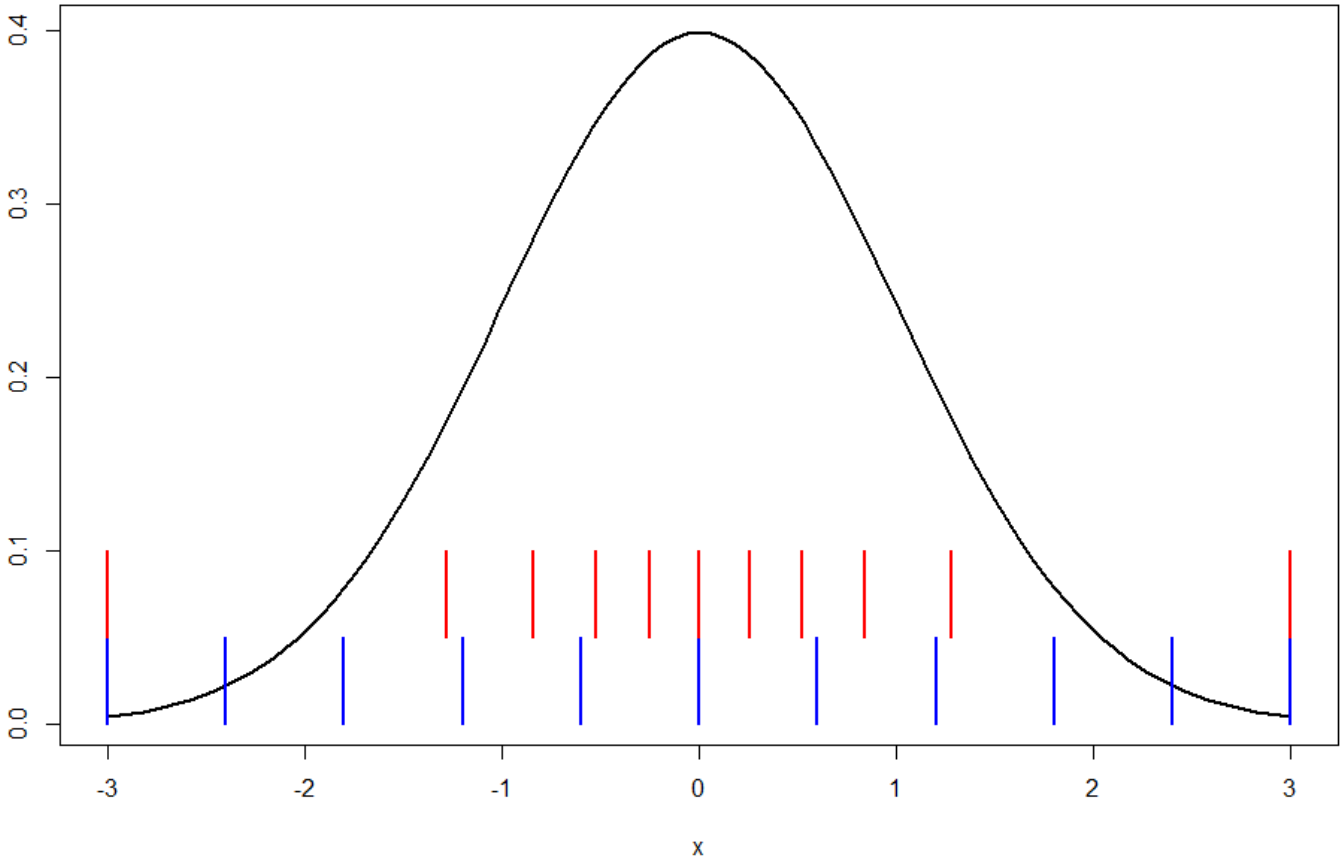
Need to bin the data
In principle any binning is ok

Two Questions:

1) What kind of bins?
2) How many bins?

What kind of bins?

Equi-distant – Equi-probable



Most studies suggest equi-probable is better

One advantage: $E=1/k >> 5$ for all bins, no need to adjust binning

Major advantage: In general leads to tests with higher power

Bins can be found easily as quantiles of F or as quantiles of data

How many bins?

Textbook answer: $k = 2n^{2/5}$

D'Agostini and Stephens (1986) "*Goodness-of-Fit Techiques*"

But: really depends on alternative

Example: $H_0: X \sim U[0,1]$ vs $H_a: X \sim Linear$

Optimal k: k=2!

EDF Methods

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

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

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

so

$$D\{\widehat{F}(x), F(x)\}$$

where D is some distance measure on function space can be used for goodness-of-fit test.

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

One table to rule them all!

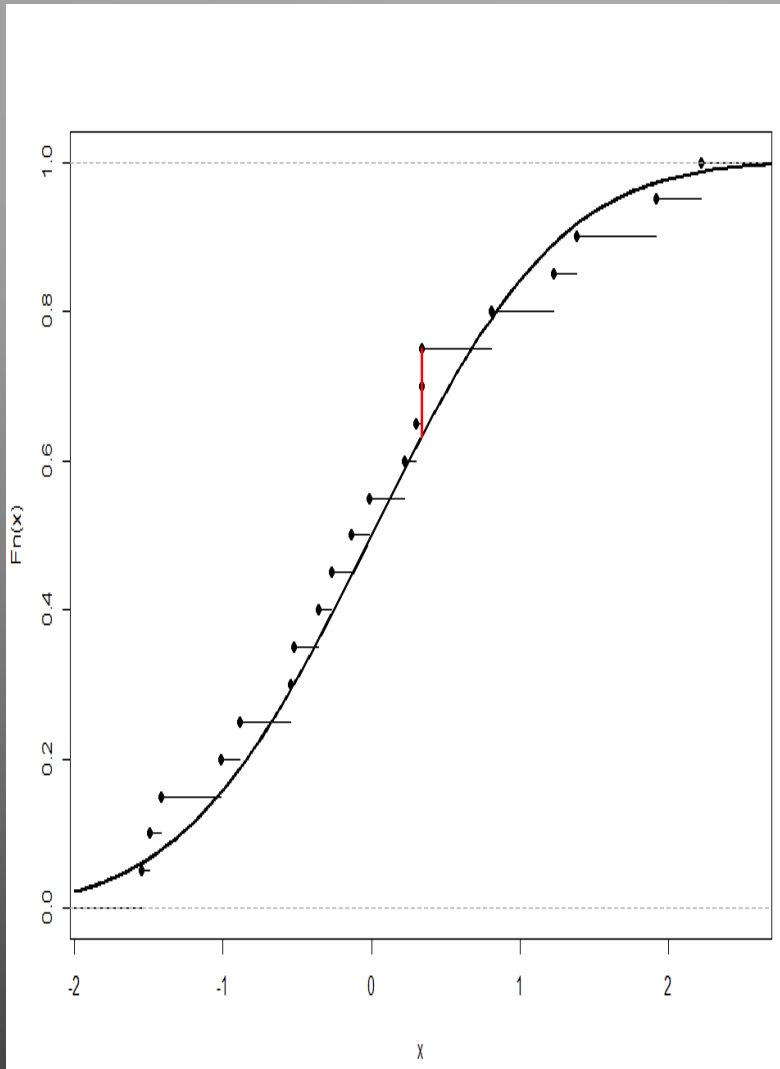
Except this does not work if parameters are estimated from data!

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



Alternatives

Anderson-Darling

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.

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

Cramer-vonMises

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.

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

And more…

Modern theory based on convergence of $\widehat{F}$ to Gaussian process, most importantly Donsker's theorem (Central limit theorem for stochastic processes)

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!

R package KScorrect

Uses maximum likelihood to estimate parameters and Monte Carlo simulation to estimate null distribution

Example:

```
> x<-rexp(1000,1)
> LcKS(x,"pexp")$p.value
[1] 0.3998
```

- `"pnorm"` for normal,
- `"pmixnorm"` for (univariate) normal mixture,
- `"plnorm"` for lognormal (log-normal, log normal),
- `"punif"` for uniform,
- `"plunif"` for loguniform (log-uniform, log uniform),
- `"pexp"` for exponential,
- `"pgamma"` for gamma,
- `"pweibull"` for Weibull.

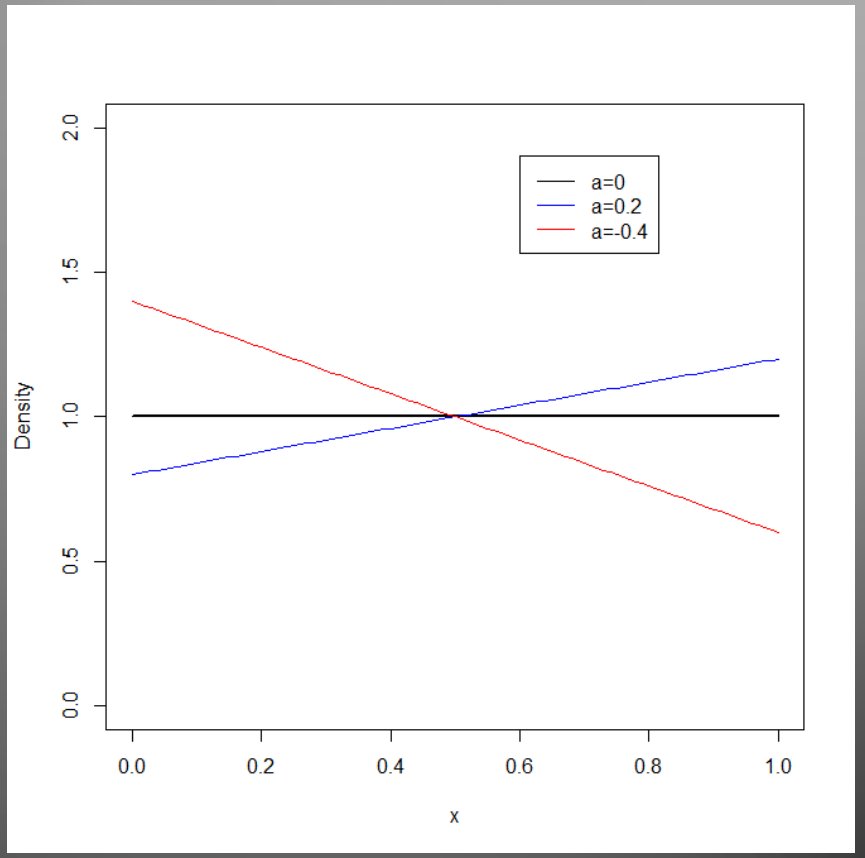
X^2 vs K–S

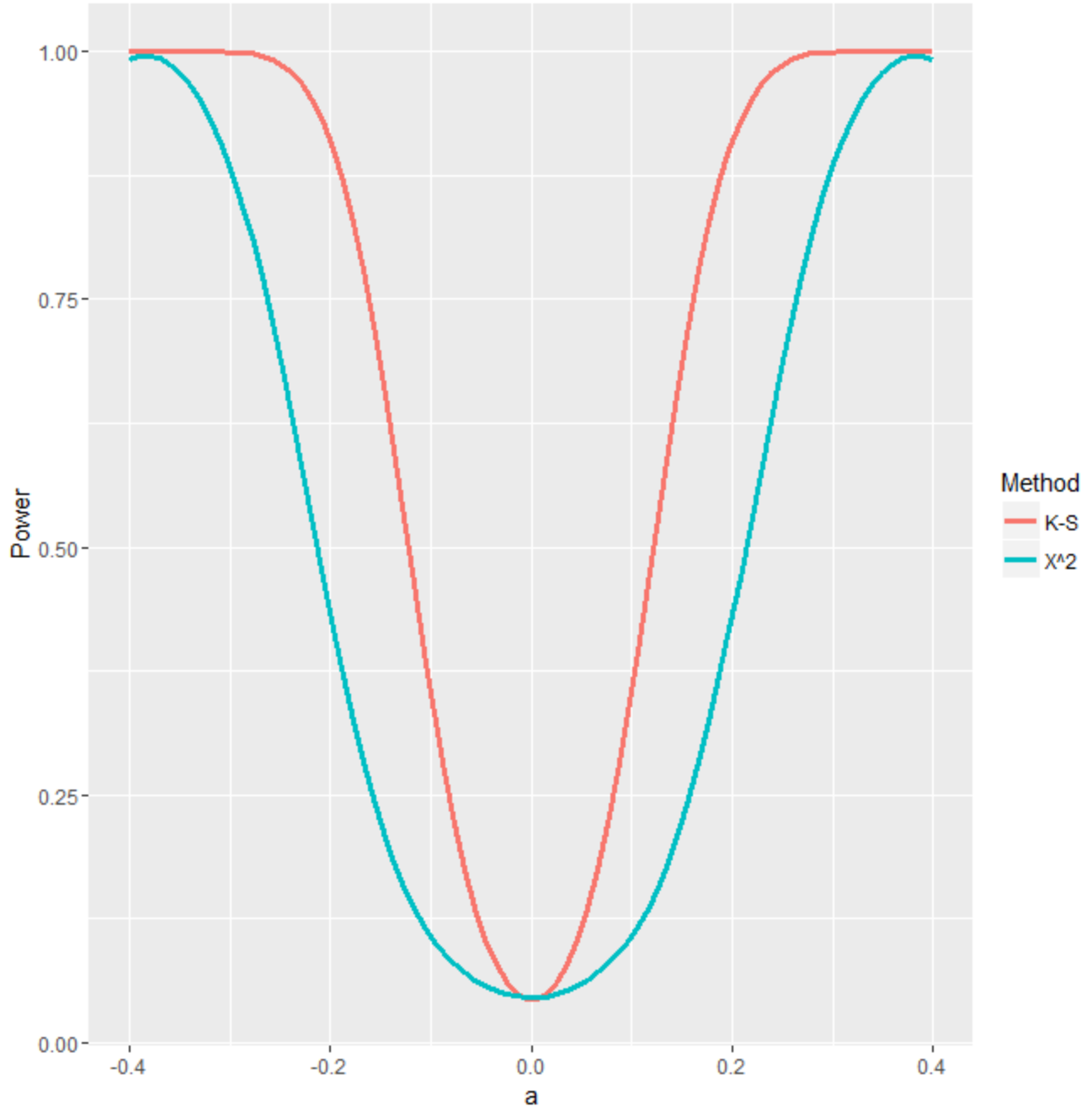
$H_0: F = Uniform[0,1]$

$H_a: F = Linear\ [0,1]$

Sample size: n=1000

X^2: 32 bins
(here Equi-distant = Equi-probable)

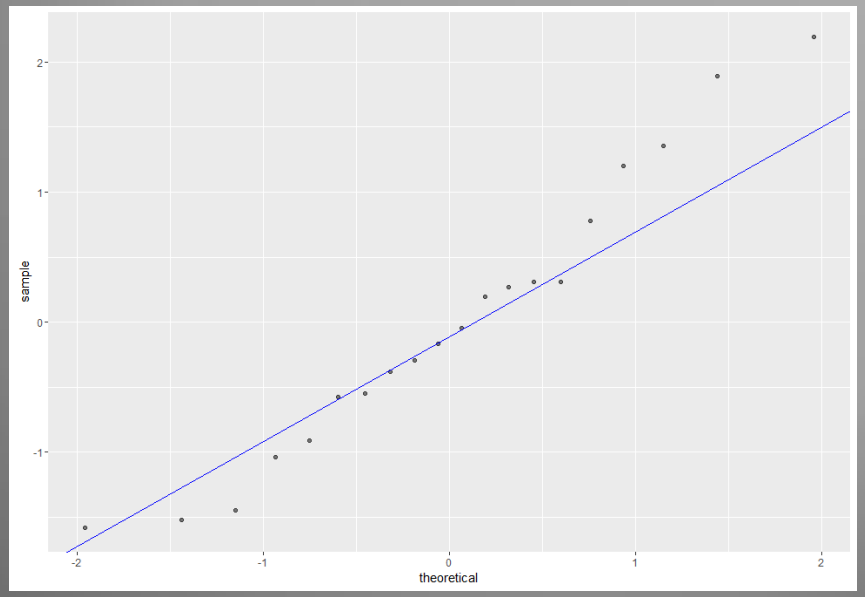



1.00
0.75
0.50
0.25
0.00
Power
-0.4
-0.2
0.0
0.2
0.4
a
Method
K-S
X^2

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

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

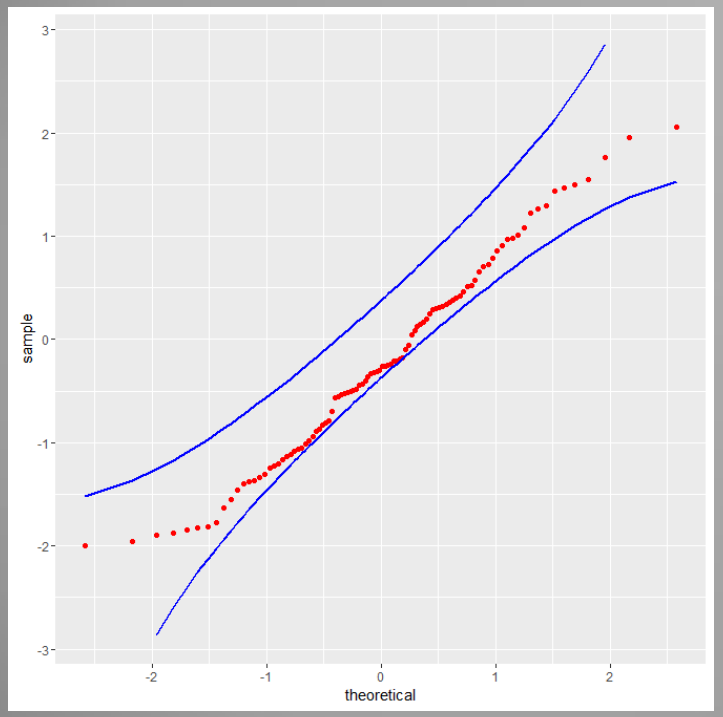
Find pointwise confidence intervals from quantiles of Beta distribution

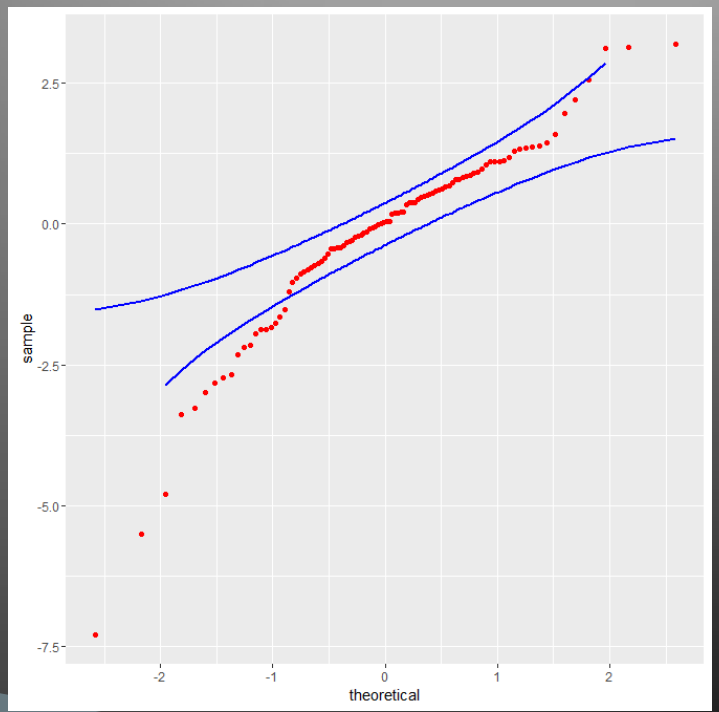
Turn into simultaneous confidence band by adjusting nominal confidence level

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.htm





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!

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.

testing composite hypotheses is possible

Quite unusual: best method for estimating parameters: MoM (method of moments)

Example:

$$X_1,..,X_n \text{ iid } N(\mu,\sigma)$$

$$\mu = E[X_1] \simeq \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\sigma^2 = Var(X_1) = E[X_1^2] - E[X_1]^2$$

$$E[X_1^2] \simeq \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

$$\sigma^2 \simeq \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$$

And many more…

- Tests based on moments
- Tests specific for a distribution (Normal: more than 20 tests)
- A good place to start: “Comparing Distributions”, Olivier Thais, Springer

Multidimensional Data

Richard Bellman : Curse of Dimensionality

Example: H_0: $(X_1,..,X_d) \sim U[0,1]^d$

We want E≥ 5 and we want 10 bins in each dimension. What n do we need?

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

Example: consider box with sides length l:

$$B(l) \quad = \{x \in R^d, 0 < x_i < l\}$$

What is the volume of the box?

$$B(1-\varepsilon) = (1-\varepsilon)^d \to 0 \; as \; d \to \infty$$

$$B(1+\varepsilon) = (1+\varepsilon)^d \to \infty \; as \; d \to \infty$$

Example: $(X_1,..,X_d)$ with $X_i \sim N(0,1)$, independent

$$P(point\ is\ in\ circle\ of\ radius\ 3) =$$
$$P\left(\sqrt{X_1^2+..+X_d^2} < 3\right) =$$
$$P\left(X_1^2+..+X_d^2 < 9\right) =$$
$$P(\chi^2(d) < 9)$$

d=1	99.7%
d=5	89.1%
d=10	46.8%
d=20	1.7%

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

At the very least will require gigantic data sets

Still a wide open problem!

Thanks!