Statistical Treatment and Analysis of the Data

Alberto Annovi
PhD School in Experimental Physics - University of Siena

Material courtesy of Paolo Francavilla
Mistakes are mine

Hypothesis testing

Hypothesis testing

From a set of observables x we want to check if they support a specific hypothesis for our model H0 WRT another defined model H1

Models: H0: the patient does not have COVID, H1: the patient does have COVID

In the perfect world we like to build a statistics that have T(x) such that T(x)

T(x)=0 if H0 is true

T(x)=1 if H1 is true.

T(x)=1 => The test is significant; T(x)=0 => The test is not significant;

In general

$P(T(x)=1 \mid H0) = \alpha$ (False positive) - test size - to be chosen to be $<<1$

$P(T(x)=0 \mid H1) = \beta$ (False negative)

$1-\beta$ = Power(T) is the power of the test. We want the Power to be the biggest possible for a given test size

Hypothesis testing

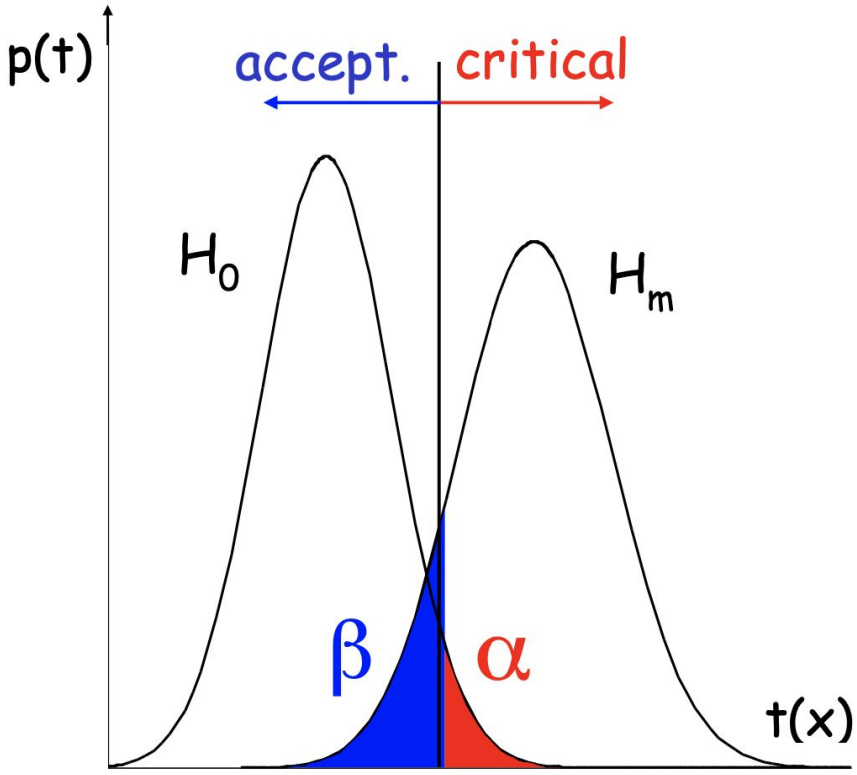
Usually, when starting from a set of observables x_1-x_n, one builds a statistics t(x), such that

p(s|H0) != p(s|H1)

Once the value of **α** (false positive rate) is fixed, one can define a T(t(x)) = bool(t in Critical region), where the critical region is defined such that Prob(t in Critical region | H0) = **α**

Typical values of **α**=0.05 or smaller

β is a function of **α**



What if H1 depends on a parameter

If x under H are described by p(x|H) which are not depending on free parameters, the hypothesis is **SIMPLE**.

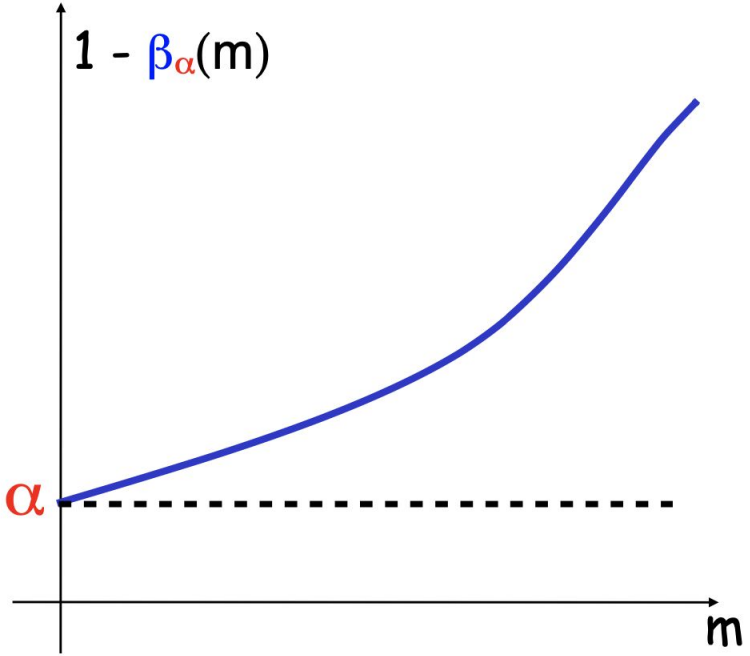
If there is a dependence on the parameter, the hypothesis is **COMPLEX**.

It can frequently happen that we want to test the hypothesis m=m0, VS m>m0.

In this case, H0 is simple, H1 is complex.

What about α and β ?

α is always decided a priori, while β is a function of m.



Properties of a test

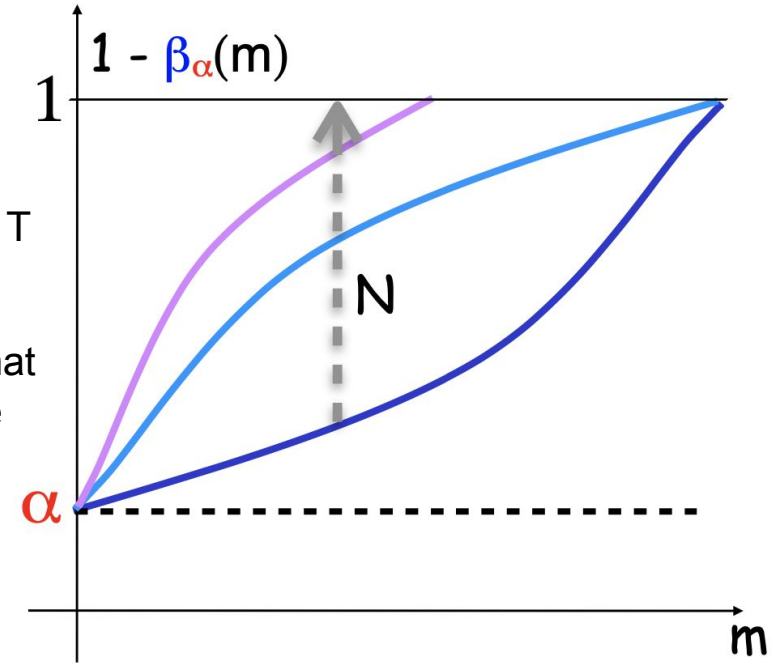
Unbiasedness: $\forall\, m$: $\text{power}_T(m) \geq \alpha$
(very desirable)

Consistency: $\forall\, m$: lim (N->inf) $\text{power}_T(m) = 1$

Maximum power (MP): for simple hypothesis, the test T for which we get the maximum of the power

Uniformly most powerful (UMP): if exist a test T such that $\forall\, m\ \text{power}_T(m) > \text{power}_{T'}(m)\ \forall\, T'$ T is the preferred one

Local most powerful (LMP):
T such that $\forall\, m\ \text{power}_T(m) > \text{power}_{T'}(m)\ \forall\, T'$
for m close to the value in H0.



Simple Hypotheses: Neyman-Pearson test

If H0 and H1 are simple hypotheses, the NP theorem demonstrates that the statistics

$s=p(x|H0)/p(x|H1)$ and the critical region C: $s<c_\alpha$ give the MP test.

NOTE: if x follows H1 and not H0, we could expect $p(x|H0)<p(x|H1)$ => small values of s => it make sense to reject H0 for small values of s.

In general, one has to calculate s, and $p(s|H0)$ to find the value of c_α

=> IF you have 2 simple hypotheses, NP test is the one to be used.

Simple Hypotheses: Neyman-Pearson test

$$\int_{w_\alpha} f_N(\mathbf{X}|\theta_0)d\mathbf{X} = \alpha$$

$$1-\beta = \int_{w_\alpha} f_N(\mathbf{X}|\theta_1)d\mathbf{X}.$$

$$1-\beta = \int_{w_\alpha} \frac{f_N(\mathbf{X}|\theta_1)}{f_N(\mathbf{X}|\theta_0)} f_N(\mathbf{X}|\theta_0)d\mathbf{X}$$

$$= E_{w_\alpha}\left(\frac{f_N(\mathbf{X}|\theta_1)}{f_N(\mathbf{X}|\theta_0)}\,\middle|\,\theta=\theta_0\right).$$

$$\ell_N(\mathbf{X},\theta_0,\theta_1) \equiv \frac{f_N(\mathbf{X}|\theta_1)}{f_N(\mathbf{X}|\theta_0)} \geq c_\alpha\,,$$

$$\begin{array}{llllll} \text{if} & \ell_N(\mathbf{X},\theta_0,\theta_1) > c_\alpha & \text{choose} & H_1: & f_N(\mathbf{X}|\theta_1) \\ \text{if} & \ell_N(\mathbf{X},\theta_0,\theta_1) \leq c_\alpha & \text{choose} & H_0: & f_N(\mathbf{X}|\theta_0)\,. \end{array}$$

Composite Hypotheses: LR test

If H0 and H1 are complex hypotheses but P(x| H0)= p(x,m=m0), P(x| H1)= p(x,m>m1),

one can use $\lambda = -2\log(p(x|H0)/\sup(p(x|H1)))$

and a critical region C: $\lambda > c_\alpha$

NOTE: if x follows H1 and not H0, we could expect p(x|H0)<sup p(x|H1) => large values of λ => it make sense to reject H0 for large values of λ.

In general, one has to calculate λ, and p(λ|H0) to find the value of c_α

BUT: we know that asymptotically λ is distributed like a chi2 => extremely useful to get a fast asymptotic estimate of the q_α

How many degrees of freedom? If general, one can extend the use of λ even for composite H0, and the degrees of freedom is the difference in the number of free parameters in the 2 hypotheses.

When LR is the UMP

In general, one can demonstrate that if

$$p(x;\mu) = \prod_{i=1,N} F(x) \cdot G(\mu) \cdot \exp\left(A(x)B(\mu)\right)$$

the LR test is the UMP test. In this case, LR is a function of $t(x) = \sum_{i=1,N} A(x_i)$ so the test can be done directly on t(x).

NOTE: this is the exponential family, and t is the sufficient statistics!

LMP test

The LMP test is important when a fast decision must be taken, even in the presence of small deviations from H0.

We are just interested in getting the MP test very close to H0.

$$t = \frac{d \log L(x; m))}{dm} \Big|_{m=m_0}$$

it can be demonstrated that the statistics based on the Fisher score calculated at m=m0, and an appropriate critical region, $t>q_\alpha$ or $t<q_\alpha$ gives the LMP

NOTE1: in general, there are 2 possible tests depending of which is the H1 we are considering (m>m0 or m<m0).

Under H0, the distributions of t are asymptotically gaussians with mean value 0, and variance = the Fisher information, if it exists.

Under H1 will be again a gaussian, but shifted and, usually, with a slight bigger variance.

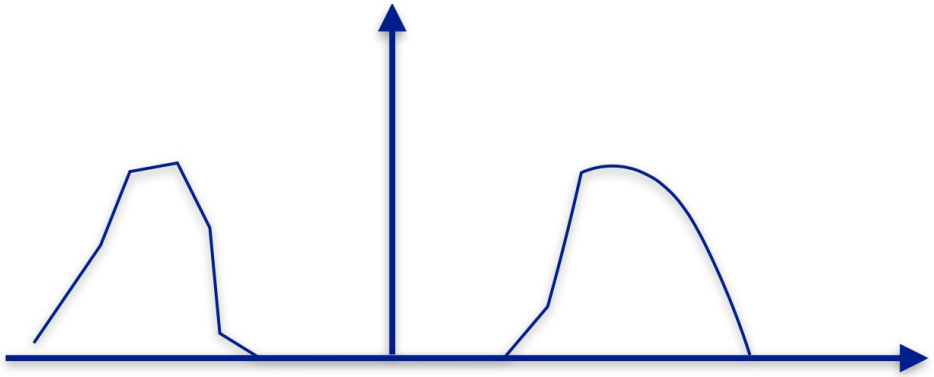
What if ?

We assume that H0 = gaussian with mean value μ=0 and H1 gaussian with μ=μ1?

$$\lambda = 2/(2\sigma^2)\sum x^2 - (x-\mu_1)^2 = 1/\sigma^2 \sum (2\mu_1 x - \mu_1^2) \simeq \sum x \simeq \bar{x}$$

The test is just done by looking at the value of the average. Let's say that we measured an average ~ 0 => We cannot reject H0.

BUT by plotting our data we got:



Goodness of fit

The need for a GOF

It is clear from the example above that testing H0 VS H1 is not the only way in which we would like to reject H0.

There must be another way of proceeding in which H0 is defined (the gaussian with $\mu=0$ of the previous example), while we want to be as open as possible to any kind of alternative!

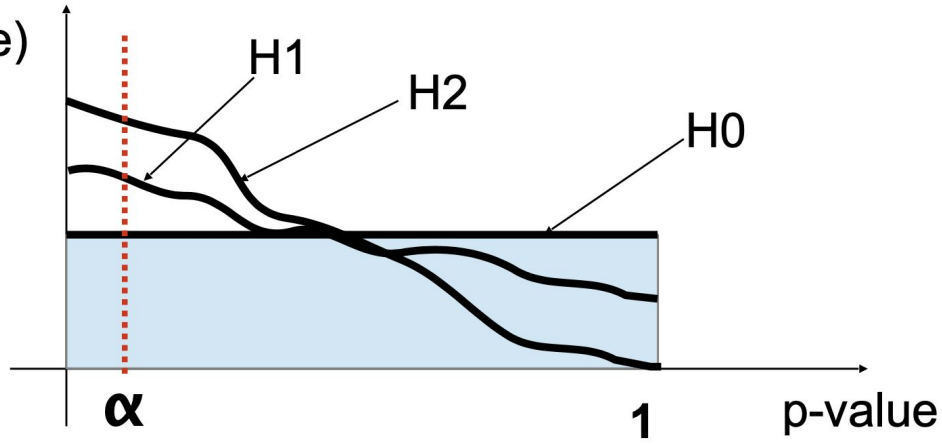
We can still define a statistics T(x) such that PROB(T(x)=1 | H0) =α, but the concept of POWER is completely lost!

We do not have anymore UMP criteria to guide us in the selection of T

In fact, any T is ok, as soon as it is testinting some specific feature of H0.

p-value

To quantify the goodness of our model, we usually have 2 options:



- Define a statistics T such that T=0 or 1, with Prob(T=1|H0)=α as done for the HT
  - if T=1 we reject H0, with a confidence of 1-α=0.95
- Define a statistics named p-value which gives a measure of the goodness of fit.

  Definition: p-value is a statistics such that under H0 p(p-value) = U(0,1)

  Properties: Unbiased: for some of the alternatives hypothesis WRT H0, the distribution of the p-value should move towards smaller values of the p-value

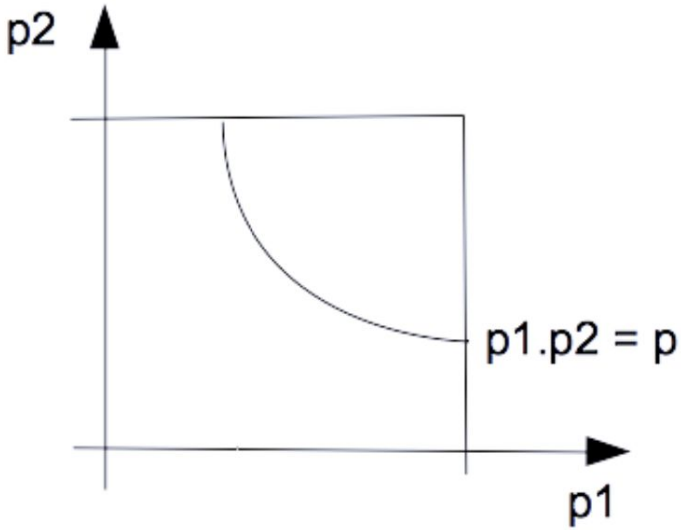
Compilation of p-values

If we have more independent p-values (i.e. p1 and p2) and we want to combine them in a new p-value, there are more ways of doing it.

BUT the new p-value must still be distributed like U(0,1) under H0.

SO IT CANNOT BE p=p1*p2

p is not uniformly distributed.



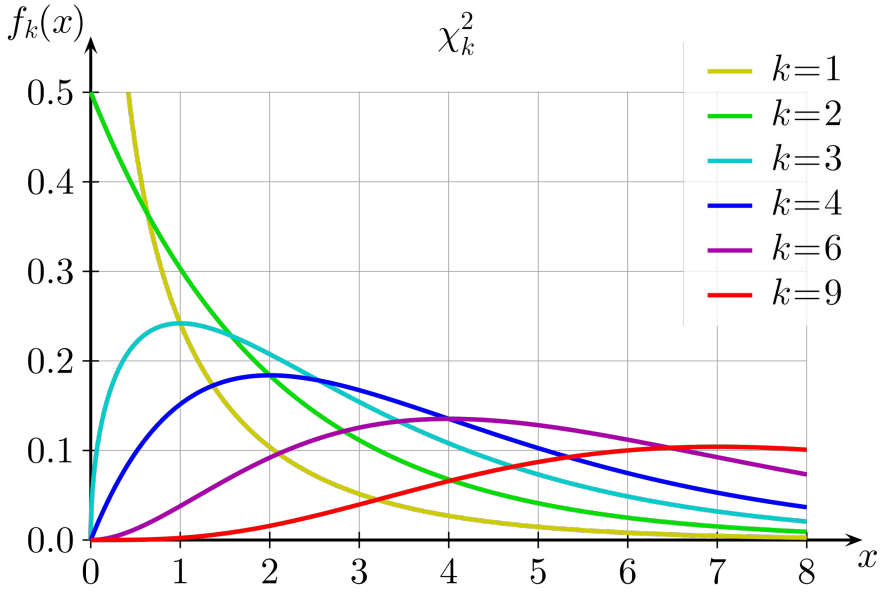
In general, one can use:

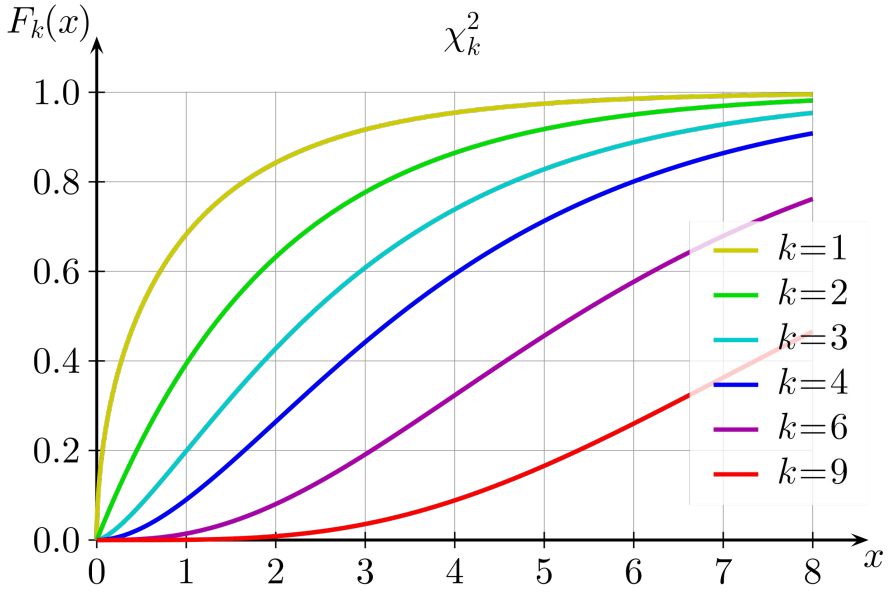
$$q = -2\sum_i \log p_i = -2\log\prod_i p_i$$

Which is distributed like a chi2 with 2N degrees of freedom. The combined p-value is the chi2 percentile

chi2 distributions

From wikipedia





-2 Log p-value distribution

$$q = -2\sum_i \log p_i = -2\log\prod_i p_i$$

$$\int_{x_a}^{x_b} p_x(x)dx = \int_{y(x_a)}^{y(x_b)} p_y(y)dy$$

$$p_y(y) = \frac{p_x(x(y))}{|dy/dx|}$$

$$Y = g(X)$$
$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}\left(g^{-1}(y)\right)\right|$$

y(x) = -2 log x

x is the p-value so U(0,1)

$p_y(y)$?

$p_x(x(y)) = 1$ (p_x is constant)

$|dy/dx| = 2/x$

$p_y(y) = x/2$

replace x with $x(y) = e^{-y/2}$

$p_y(y) = e^{-y/2}/2$ which is the chi2 with 2 d.o.f.

Komlogorov - Smirnov

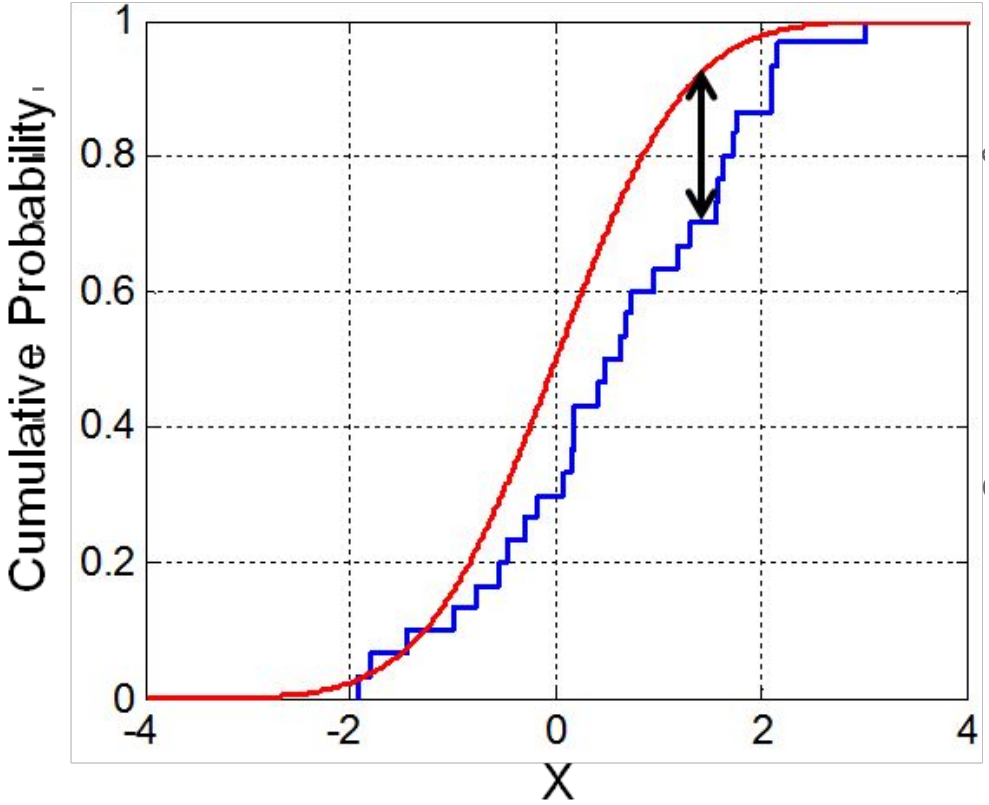
- 1 D problem with no free parameter
- We want to check if x_1-x_n are distributed like p(x|H0).
- F(x) is the cumulant for p(x|H0)
- $D_N \equiv \max_i |S_i - F(x_i)|$ where $S_i = \sum_{j:x_j<x_i} \frac{1}{N}$
-
- The distribution of D_N does not depend on P, so one can calculate the critical region a-priori

$$D_N > q_\alpha$$

can be used to compare two experimental distributions

$$D = \sqrt{(N_1 N_2/(N_1+N_2))}\max|S1_i - S2_i|$$

Komlogorov - Smirnov



$n \backslash \alpha$	0.001	0.01	0.02	0.05	0.1	0.15	0.2
1		0.99500	0.99000	0.97500	0.95000	0.92500	0.90000
2	0.97764	0.92930	0.90000	0.84189	0.77639	0.72614	0.68377
3	0.92063	0.82900	0.78456	0.70760	0.63604	0.59582	0.56481
4	0.85046	0.73421	0.68887	0.62394	0.56522	0.52476	0.49265
5	0.78137	0.66855	0.62718	0.56327	0.50945	0.47439	0.44697
6	0.72479	0.61660	0.57741	0.51926	0.46799	0.43526	0.41035
7	0.67930	0.57580	0.53844	0.48343	0.43607	0.40497	0.38145
8	0.64098	0.54180	0.50654	0.45427	0.40962	0.38062	0.35828
9	0.60846	0.51330	0.47960	0.43001	0.38746	0.36006	0.33907
10	0.58042	0.48895	0.45662	0.40925	0.36866	0.34250	0.32257
11	0.55588	0.46770	0.43670	0.39122	0.35242	0.32734	0.30826
12	0.53422	0.44905	0.41918	0.37543	0.33815	0.31408	0.29573
13	0.51490	0.43246	0.40362	0.36143	0.32548	0.30233	0.28466
14	0.49753	0.41760	0.38970	0.34890	0.31417	0.29181	0.27477
15	0.48182	0.40420	0.37713	0.33760	0.30397	0.28233	0.26585
16	0.46750	0.39200	0.36571	0.32733	0.29471	0.27372	0.25774
17	0.45440	0.38085	0.35528	0.31796	0.28627	0.26587	0.25035
18	0.44234	0.37063	0.34569	0.30936	0.27851	0.25867	0.24356
19	0.43119	0.36116	0.33685	0.30142	0.27135	0.25202	0.23731
20	0.42085	0.35240	0.32866	0.29407	0.26473	0.24587	0.23152
25	0.37843	0.31656	0.30349	0.26404	0.23767	0.22074	0.20786
30	0.34672	0.28988	0.27704	0.24170	0.21756	0.20207	0.19029
35	0.32187	0.26898	0.25649	0.22424	0.20184	0.18748	0.17655
40	0.30169	0.25188	0.23993	0.21017	0.18939	0.17610	0.16601
45	0.28482	0.23780	0.22621	0.19842	0.17881	0.16626	0.15673
50	0.27051	0.22585	0.21460	0.18845	0.16982	0.15790	0.14886
OVER 50	**1.94947** / $\sqrt{n}$	1.62762 / $\sqrt{n}$	1.51743 / $\sqrt{n}$	1.35810 / $\sqrt{n}$	1.22385 / $\sqrt{n}$	1.13795 / $\sqrt{n}$	1.07275 / $\sqrt{n}$

Critical values of Kolmogorov-Smirnov test variable D
as function of sample size n and significance level α

GOF with histograms

Let's assume we have a model p(x,μ) and we group the observables x_1-x_n in k bins.

If n is a stochastic variable distributed like a poissonian, the number of observation in each bin is distributed like a poissonian, and they are independent from each other.

The mean value of each poissonian in each bin i is $f_i(\mu)$ and depends on the model p(x,μ).

In this case we have a natural alternative in to our model, in which the different f_i are free parameters.

We can inherit a statistics from the HT: λ

In this case,

$$\lambda(x) = 2\log \frac{\sup \prod_i \mathrm{Pois}(f'_i, k_i)}{\sup_\mu \prod_i \mathrm{Pois}(f_i(\mu), k_i)}$$

GOF - Gaussian limit

If the statistics in each bin is large enough, the poissonians are approaching gaussian distributions.

In this case $$\lambda(x) = \sum_i \frac{[y_i - f_i(\hat{\mu})]^2}{\sigma_i^2}$$

Chi2 test!

Asymptotically distributed like a chi2 with nbins-dim(μ) DOF