Lecture 3:

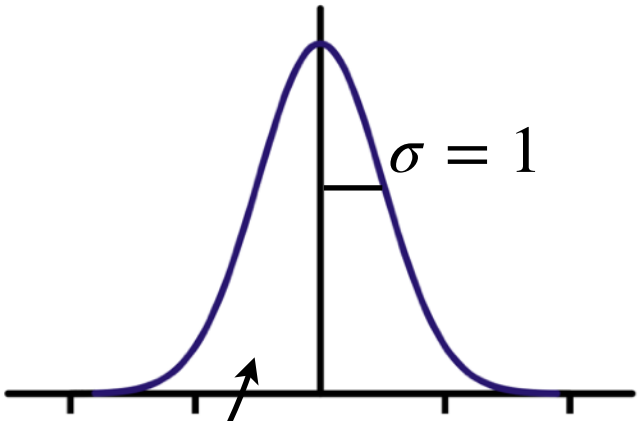
Testing Models

- Chi-Squared
- “Scientific Method”
- Student’s t
- Correlation test
- Non-parametric tests

Testing Models

Consider:

$$\chi^2 \equiv \sum_{i=1}^{n} g_i^2$$



where g_i are samples drawn from a normal (*i.e.* Gaussian) distribution of unit variance

Then the distribution of this quantity defines a χ^2 ("chi-squared") distribution with $\mathbf{n}$ *degrees of freedom*

effective number of independent samples contributing to the variance

The χ^2 probability density function for $\mathbf{n}$ degrees of freedom has the form:

$$P(\chi^2, n) = \frac{(\chi^2)^{\frac{n}{2}-1} e^{-\frac{\chi^2}{2}}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}$$

Where $\Gamma(k) = (k-1)!$ if k = positive integer **or** $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$ for any real value z

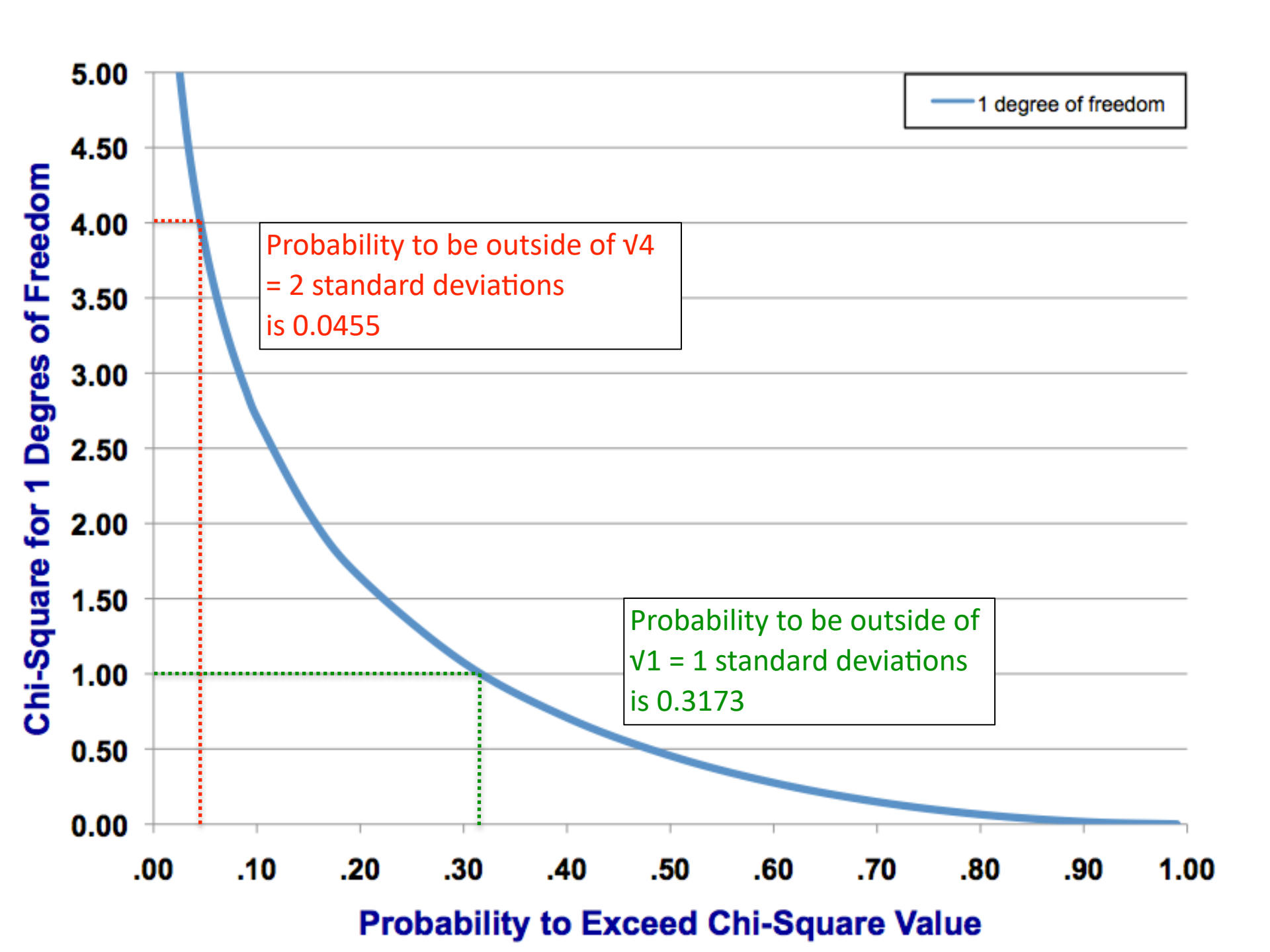
Note that: $P(\chi^2,2) = \frac{1}{2}e^{-\frac{\chi^2}{2}}$

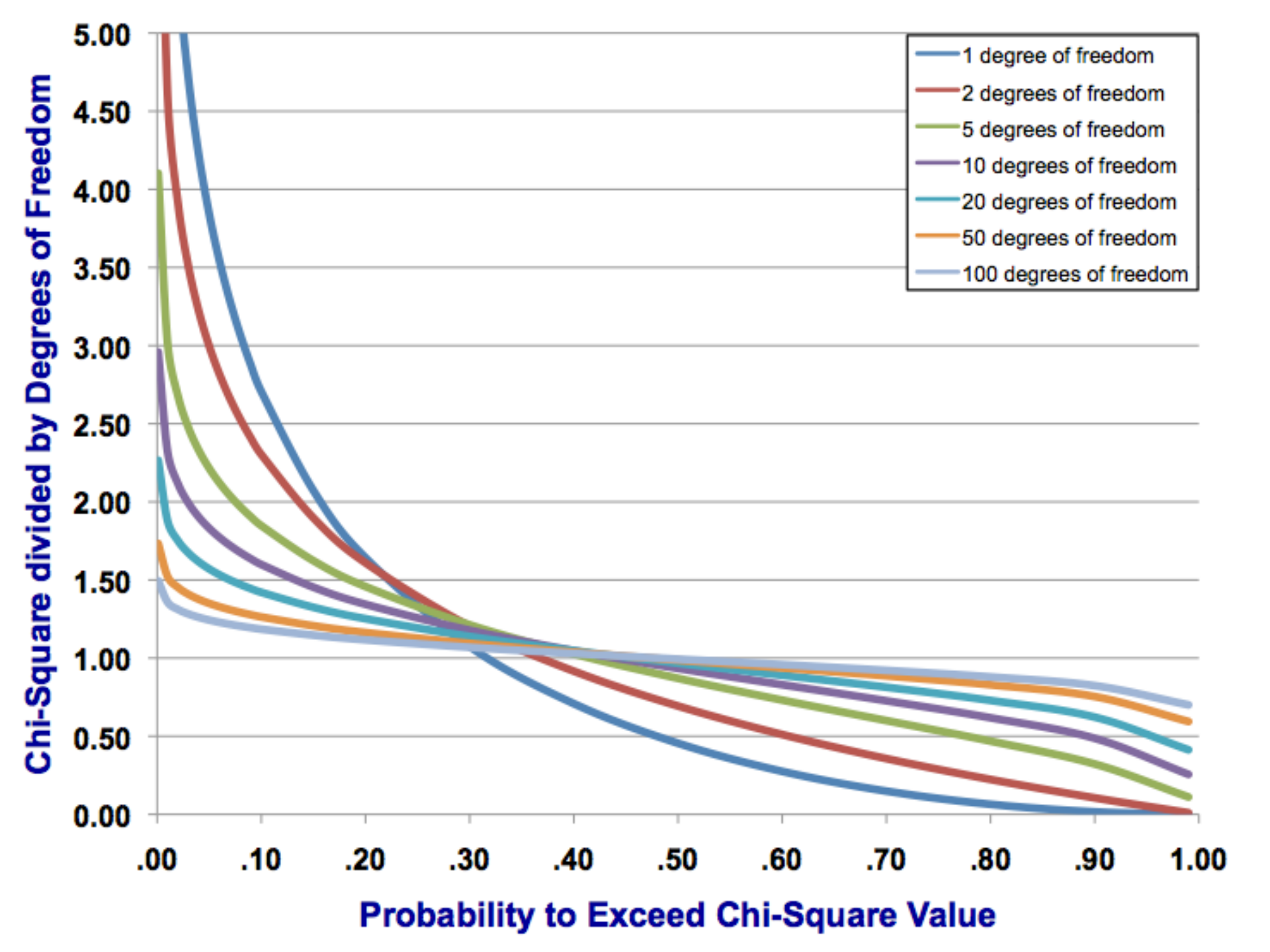
(will come back to this)

And the integral probability is given by:

$$P(>\chi^2, n) = 1 - \frac{\gamma\left(\frac{n}{2}, \frac{\chi^2}{2}\right)}{\Gamma(\frac{n}{2})}$$

where $\gamma(z, \alpha) = \int_0^\alpha x^{z-1}e^{-x}dx$


5.00
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
Chi-Square for 1 Degres of Freedom
1 degree of freedom
Probability to be outside of √4
= 2 standard deviations
is 0.0455
Probability to be outside of
√1 = 1 standard deviations
is 0.3173
.00
.10
.20
.30
.40
.50
.60
.70
.80
.90
1.00
Probability to Exceed Chi-Square Value


Chi-Square divided by Degrees of Freedom
5.00
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
.00
.10
.20
.30
.40
.50
.60
.70
.80
.90
1.00
Probability to Exceed Chi-Square Value
1 degree of freedom
2 degrees of freedom
5 degrees of freedom
10 degrees of freedom
20 degrees of freedom
50 degrees of freedom
100 degrees of freedom

Pearson's χ^2 Test

So, for example, if we have a model, **m**, involving **k** free parameters (determined by a fit to the data) that seeks to predict the values, **x**, of **n** data points, each with *normally distributed uncertainties*, we can construct the sum:

$$S \equiv \sum_{i=1}^{n} \frac{(x_i - m_i)^2}{\sigma_{m_i}^2}$$

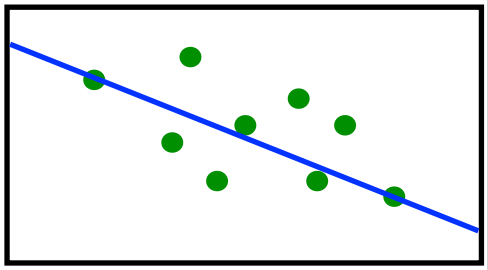
normalises things to give a Gaussian distribution with unit variance

for binned data (Poisson statistics)

$$\sigma_{m_i}^2 \cong m_i$$

S will then be distributed as a χ^2 distribution with **n-k** degrees of freedom, and can thus be used as a statistic to determine how well the model matches the data.

For example, imagine fitting a straight line (2 parameters: slope and intercept) to a set of data. You can always force the line to go through 2 of the data points exactly, so only n-2 of the data points will contribute to the variance around the model



S will then be distributed as a χ^2 distribution with **n-k** degrees of freedom, and can thus be used as a statistic to determine how well the model matches the data.

"If my model is correct, how often would a randomly drawn sample of data yield a value of χ^2 at least as large as this?"

Determining the best values for the model parameters by choosing them so as to minimise χ^2 is called the "**Method of Least Squares**."

Note that, if you vary one of the model parameters from its best fit value until χ^2 increases by 1, this therefore represents the change in the model parameter associated with 1 unit of variance in the fit quality (*i.e.* the "1σ uncertainty" in the model parameter).

Example 1:

Say we have n measurements of some quantity, with each measurement having a different Gaussian uncertainty. What is the best estimate for the mean value of this quantity?

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma_i^2}$$

← Let's find the value of μ that minimises this

→ $$\frac{\partial \chi^2}{\partial \mu} = -2\sum_{i=1}^{n} \frac{(x_i - \mu)}{\sigma_i^2} = -2\sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} + 2\mu \sum_{i=1}^{n} \frac{1}{\sigma_i^2} = 0$$

→ $$\mu_{best} = \frac{\sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \quad where \quad w_i \equiv \frac{1}{\sigma_i^2}$$

Example 2:

A newly commissioned underground neutrino detector sees a rate of internal radioactive contamination decreasing as a function of time. Measurements of the number of such events observed are taken on 10 consecutive days. Determine the best fit mean decay time in order to determine the source of the contamination.

decay probability:

$$P(t) = \frac{1}{t_o} e^{-\frac{t}{t_o}}$$

t_o = mean decay lifetime

Centre of Time Bin (days)	Measured # counts
0.5	115
1.5	76
2.5	60
3.5	66
4.5	56
5.5	35
6.5	25
7.5	32
8.5	22
9.5	25

Table of model predictions for different values of t0

m = (Ntot/(1-exp(-10/t0)))*[exp(-(t-0.5)/t0) - exp(-(t+0.5)/t0)] ~ (Ntot/(1-exp(-10/t0)))*(1/t0)*exp(-t/t0)

normalised to total number of events observed in 10 days

often approximate integral over bin with the average

t0:	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	8.5	9	9.5	10	10.5	11	11.5	12
	151	135	123	114.4	107.3	101.6	96.9						80.3	78.[illegible]					03838
	108	101	96.1	91.62	87.88	84.72	82						71.8	70.[illegible]					14747
	77.3	76.2	74.8	73.37	71.95	70.64	69.4						64.3	63.[illegible]					88526
	55.4	57.3	58.3	58.75	58.91	58.89	58.8						57.5	57.[illegible]					88166
	39.7	43.1	45.4	47.04	48.23	49.1	49.8						51.5	51.[illegible]					79617
	28.4	32.4	35.3	37.67	39.49	40.94	42.1	43.1	43.9	44.6	45.1	45.6	46.1	46.4	46.8	47	47.3	47.5	47.731552
	20.4	24.3	27.5	30.16	32.33	34.13	35.6	36.9	38	39	39.8	40.6	41.2	41.8	42.3	42.8	43.2	43.6	43.915148
	14.6	18.3	21.4	24.15	26.47	28.46	30.2	31.7	33	34.1	35.1	36.1	36.9	37.6	38.3	38.9	39.4	39.9	40.403886
	10.5	13.7	16.7	19.34	21.67	23.73	25.5	27.2	28.6	29.9	31	32.1	33	33.9	34.6	35.3	36	36.6	37.17337
	7.49	10.3	13	15.49	17.74	19.78	21.6	23.3	24.8	26.1	27.4	28.5	29.5	30.5	31.3	32.1	32.9	33.6	34.201151

Histogram of measured times: bin width =1

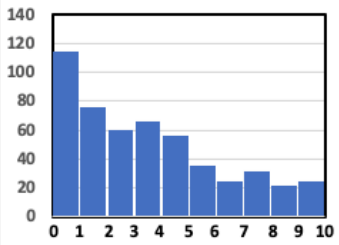


Table of ((n-m)^2)/m for different values of t0

t0:	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	8							?
	8.38	2.96	0.57	0.003	0.547	1.764	3.38	5.22	7.17	9.17	11.2								901
	9.4	6.38	4.2	2.664	1.606	0.897	0.44	0.17	0.04	0	0.03								745
	3.86	3.46	2.94	2.435	1.985	1.601	1.28	1.02	0.8	0.63	0.49								399
	2.04	1.32	1.					3	0.99	1.05	1.12								175
	6.72	3.89	2.					5	0.57	0.51	0.46								491
	1.52	0.22						2	1.8	2.05	2.28								175
	1.05	0.02	0.					6	4.47	5.02	5.52								389
	20.8	10.3	5					0	0.03	0.13	0.28								983
	12.7	4.98	1.					3	1.52	2.07	2.62								432
	40.9	20.9	11.1	5.846	2.969	1.376	0.53	0.13	0	0.05	0.21								399
Chi-squared sum:	107	54.5	29.4	17.54	12.54	11.34	12.3	14.5	17.4	20.7	24.2								305

Variance for Poisson, and then an assumption that Poisson ~ Gaussian (when is this ok?)

How many degrees of freedom?

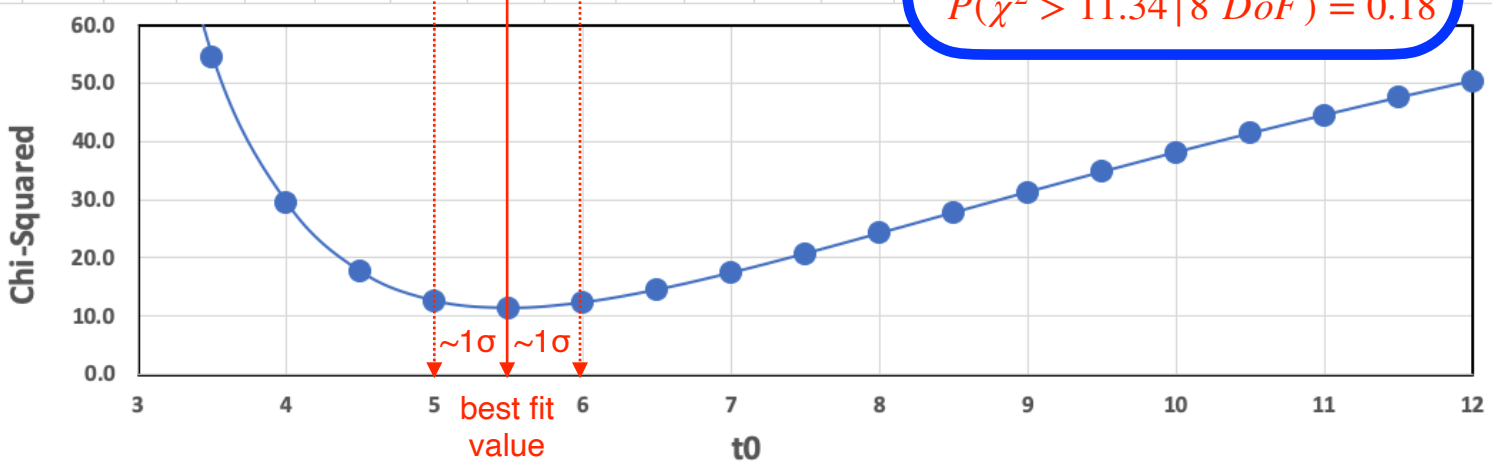
- 10 independent data points
- fit parameter t_o
- but normalisation is also based on observed data (for single bin, variance would be zero)

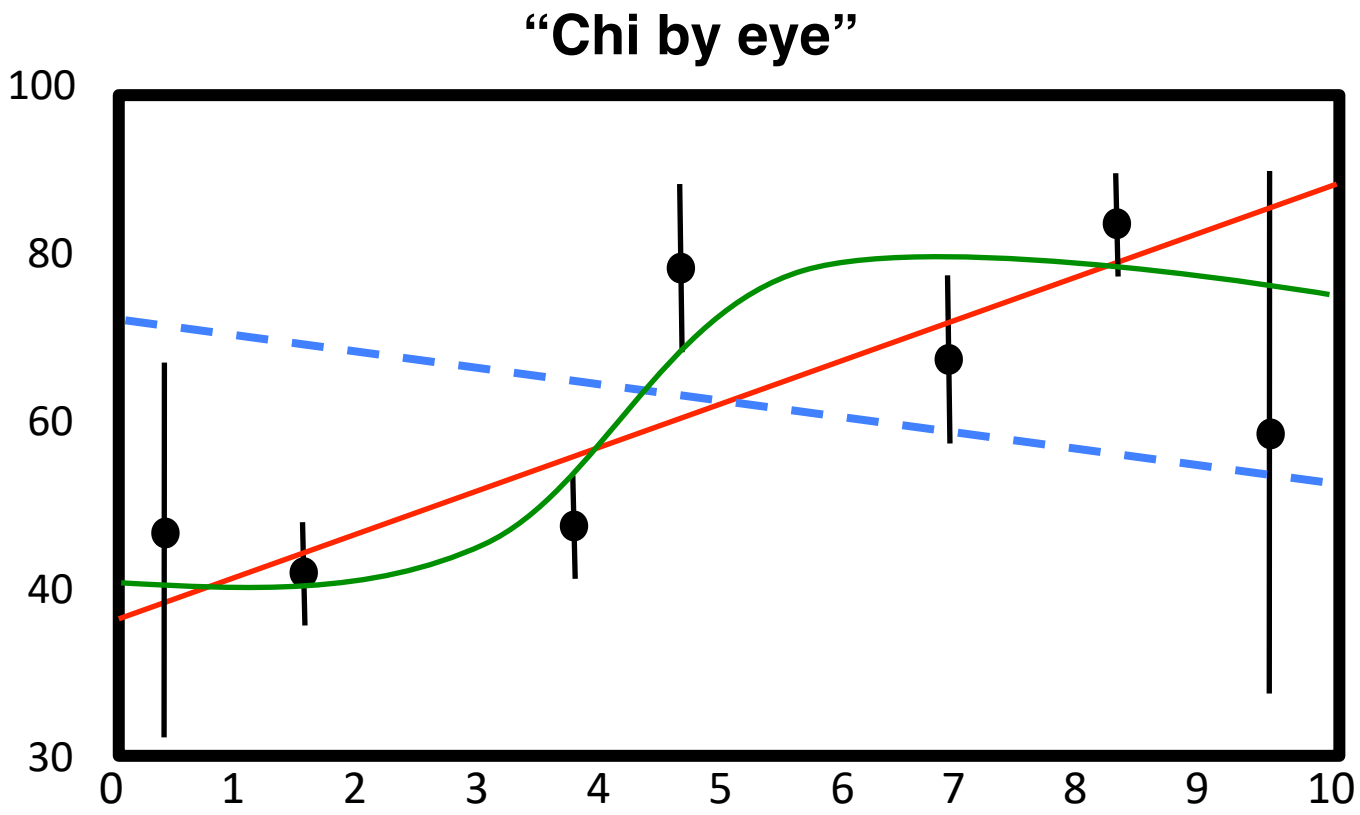
DoF = 10 - 2 = 8

How good is the fit?

$$P(\chi^2 > 11.34 \,|\, 8\ DoF) = 0.18$$

^{222}Rn mean lifetime = 5.51 days

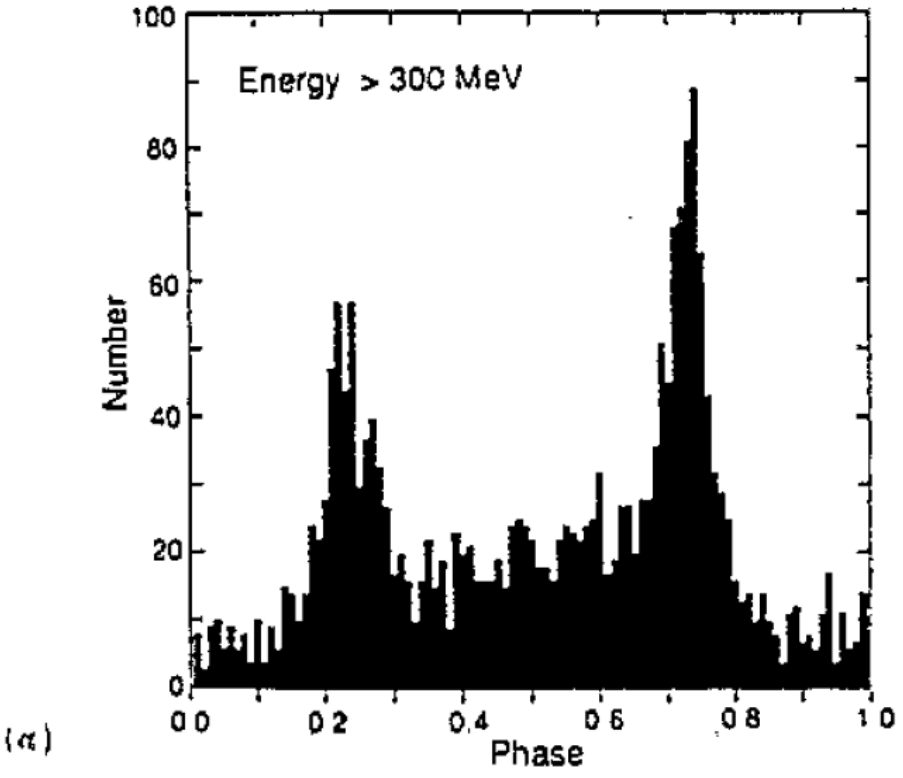




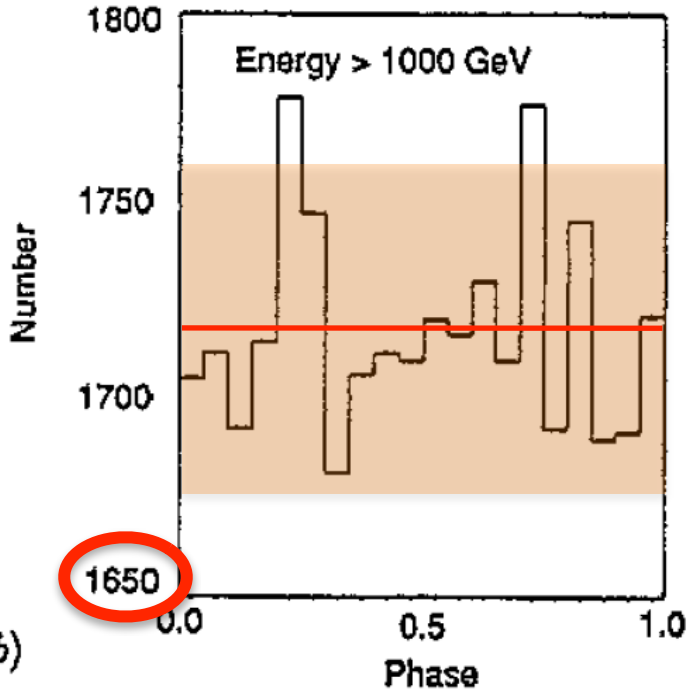
$$\chi^2 \sim (0.3)^2 + (0.1)^2 + (1.2)^2 + (1.7)^2 + (0.3)^2 + (0.8)^2 + (0.8)^2 = 5.8$$

Degrees of Freedom = 7 – 2 = 5

NOTE: This doesn't tell you which model is correct,
but it can tell you which models don't fit well!



Quoted chance probability based on correlation analysis: **6 x 10^{-4}**



x
1705
1712
1693
1715
1778
1756
1681
1707
1712
1710
1721
1717
1731
1710
1777
1693
1747
1690
1692
1722

Figure 1. (*a*) Light curve for Geminga obtained with EGRET (*b*) The VHE γ-ray light curve of Geminga plotted at the GeV γ-ray phase, as derived from the COS-B ephemeris.

Avg (m) = 1718.45

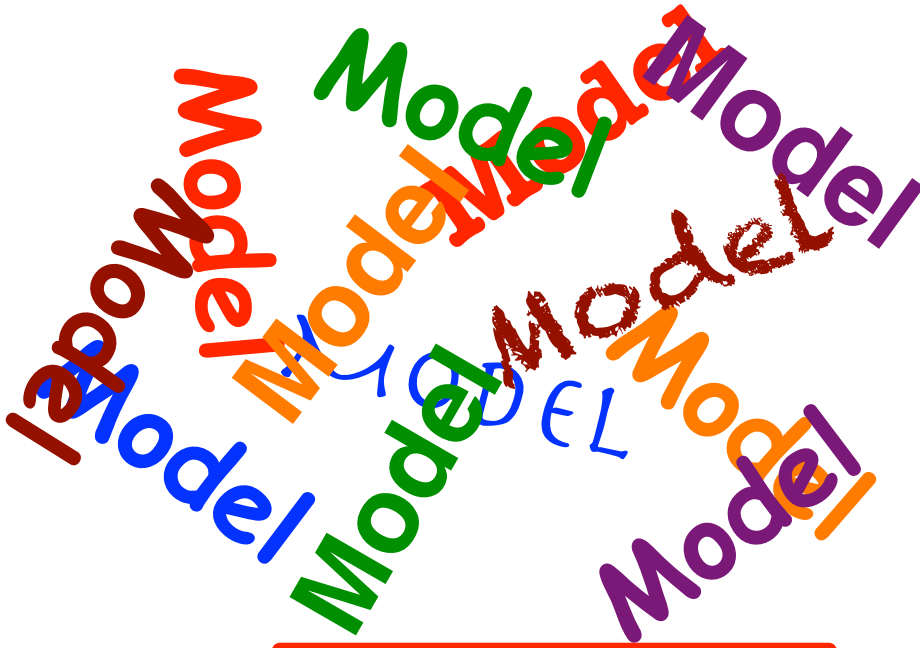
$$\chi^2 = \sum_{i=1}^{20} \frac{(x_i - 1718.45)^2}{1718.45} = 8.22$$

Wuant 'em Effect

DoF = 20 – 1 = 19 **(98.4% chance of getting something larger)**

Scientific Method:

ORDER!!



Simplest and most predictive

Test for reproducible predictions to disprove

Rejected with high confidence

Model

Next simplest & most predictive

Not rejected with high confidence

Model

Test for reproducible

Rejected with high confidence

Model

A theory is judged not on what it can explain, but on what it can reproducibly predict!

We don't prove models correct; we reject those models that are wrong!

Don't state that data are "consistent" with a given model, but rather that they are **"not inconsistent."**

More ways to test models…

Student's t

(Often misinterpreted as referring to being from or for "a student," rather than the fact that the name of the author happens to be "Student")

(Except this was actually a pseudonym used by William Sealy Gosset in his 1908 paper, who was couching himself as "a student"!)

Recall that the rms deviation in the estimated mean from a set of n samples is given by :

$$\sigma_m = \frac{\sigma}{\sqrt{n}}$$

← rms of the full distribution.

But what if we don't know σ *a priori* and all we have are the sampled estimators?

$$t \equiv \frac{\bar{x} - \mu}{\left(s/\sqrt{n}\right)}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

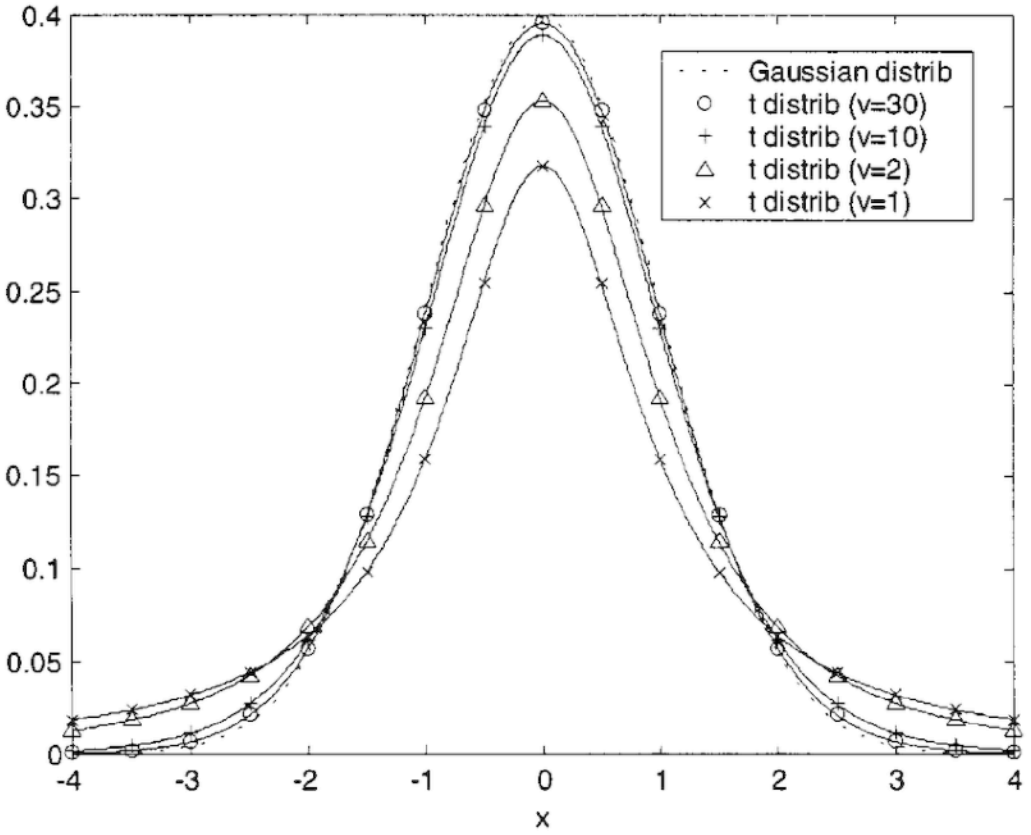
Want to find the distribution of t

$$t \equiv \frac{\bar{x} - \mu}{\left(s/\sqrt{n} \right)}$$

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

ν = # degrees of freedom

As you would expect, this approaches the shape of a Gaussian distribution as the sample size grows:



Pearson Correlation Coefficient

A test of linear correlation between two sets of data

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_i x_i^2 - n\bar{x}^2} \sqrt{\sum_i y_i^2 - n\bar{y}^2}}$$

This is just the covariance normalised by the sample rms deviations.

The value of this quantity runs from 1 (completely correlated) to -1 (completely anti-correlated), with zero indicating no correlation.

The statistics provides a *relative* measure of linear correlation but, in general, the probability distribution for r will depend on the distributions of x and y.

IF x and y are uncorrelated and each drawn from a normal distribution (such that, jointly, they can be described by a 2-D Gaussian), then:

$$\sigma_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

DoF for 2 free parameters in linear fit

From which it is possible to define a t statistic for r:

$$t_r = r\sqrt{\frac{n - 2}{1 - r^2}}$$

Spearman Rank-Order Correlation Coefficient

A non-parametric test of correlation between two sets of data (*i.e.* linearity is not assumed)

Define R_i as the 'rank' of x_i (i.e. the numerical position in an ordered list of the n data points from lowest to highest x value).

Define S_i as the 'rank' of y_i (i.e. the numerical position in an ordered list of the n data points from lowest to highest y value).

Note: it's possible to have identical ranks if a data set contains multiple identical values! In which case you should ascribe to each of these an 'average' rank value

Then define the rank coefficient as:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{R_i - \bar{R}}{s_R}\right)\left(\frac{S_i - \bar{S}}{s_S}\right) = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2}\sqrt{\sum_i (S_i - \bar{S})^2}}$$

Similarly, the probability distribution can be approximated by the t statistic:

$$t_r = r\sqrt{\frac{n-2}{1-r^2}}$$

Generally pretty good and no longer depends on the actual distributions of x & y

Kolmogorov-Smirnov (and the like)

A non-parametric test of distributions

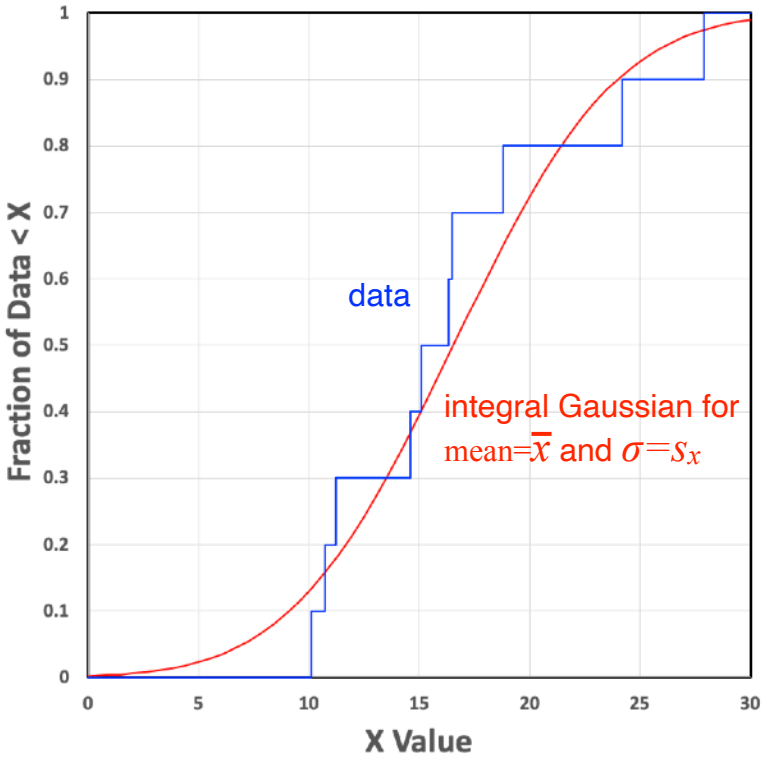
Plot the cumulative fraction of events less than or equal to a particular value of x as a function of x, along with the cumulative distribution for some model:

data point	x value	fraction ≤ x
1	10.1	0.1
2	10.7	0.2
3	11.2	0.3
4	14.6	0.4
5	15.1	0.5
6	16.3	0.6
7	16.5	0.7
8	18.8	0.8
9	24.2	0.9
10	27.9	1.0

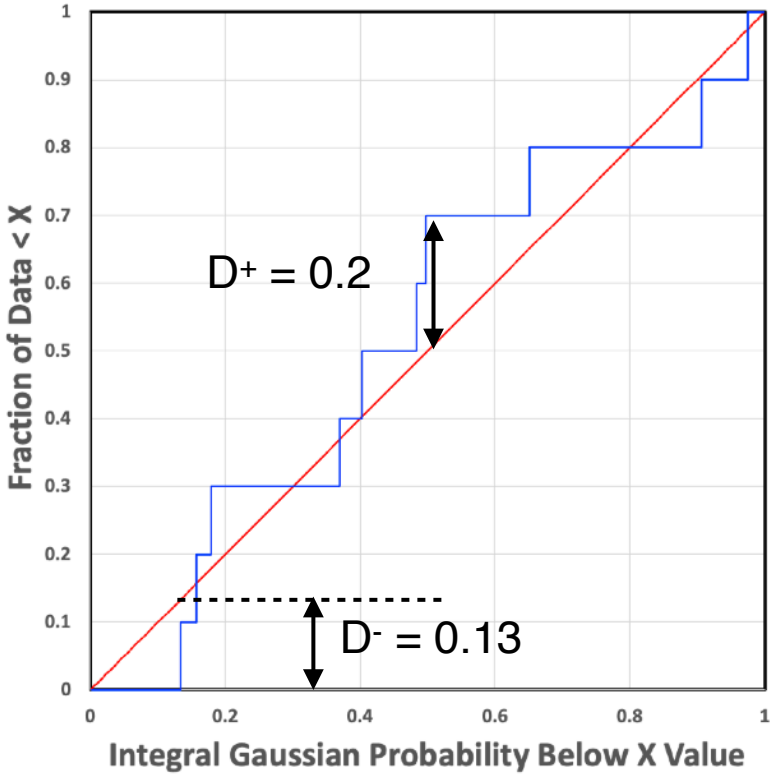
$\bar{x} = 16.54$

$s_x = 5.8$

For example: Is this data Normally distributed?



Equivalently:



A more clearly defined span on the x-axis with a visually simple model expectation that is independent of the test distribution

There are several statistics that can be used to assess the level of agreement:

K-S statistics (Clustering)

D^+ = maximum positive deviation from the model line

D^- = maximum negative deviation from the model line

$D = \max(D^+, D^-)$

$V = D^+ + D^-$ (Kuiper test)

Cramer-von Mises (Variance)

$$W^2 = \sum_{i=1}^{n} \left(y_i - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n}$$

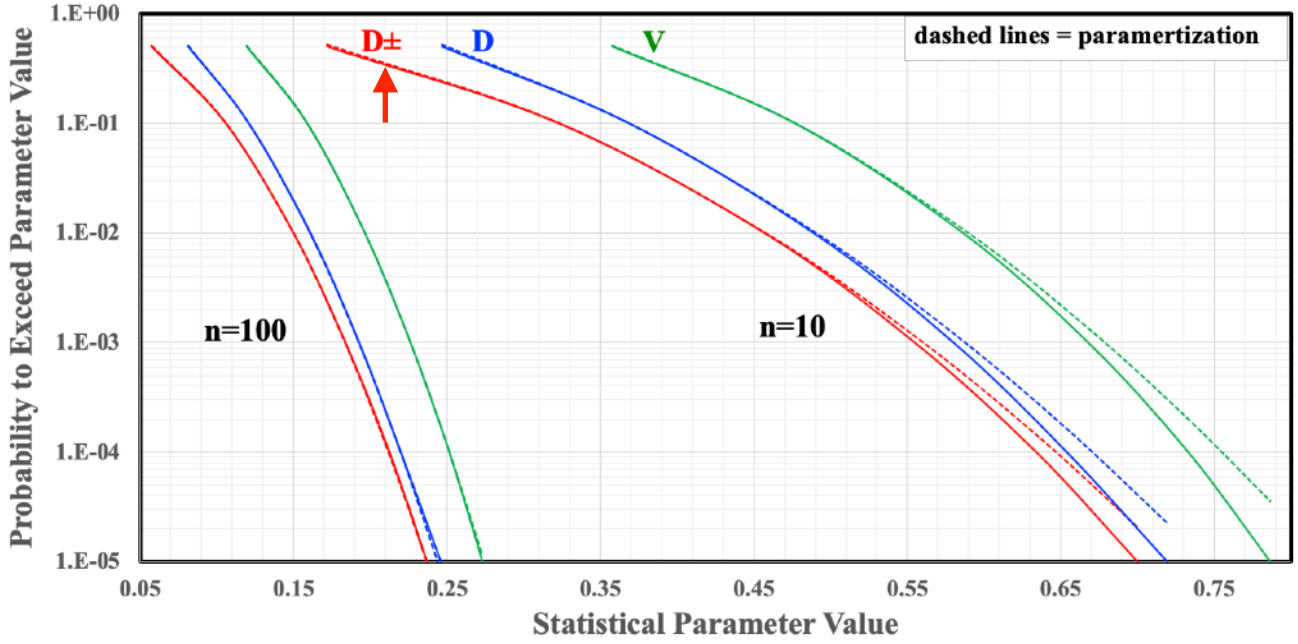
$$U^2 = W^2 - n\left(\bar{y} - \frac{1}{2} \right)^2$$

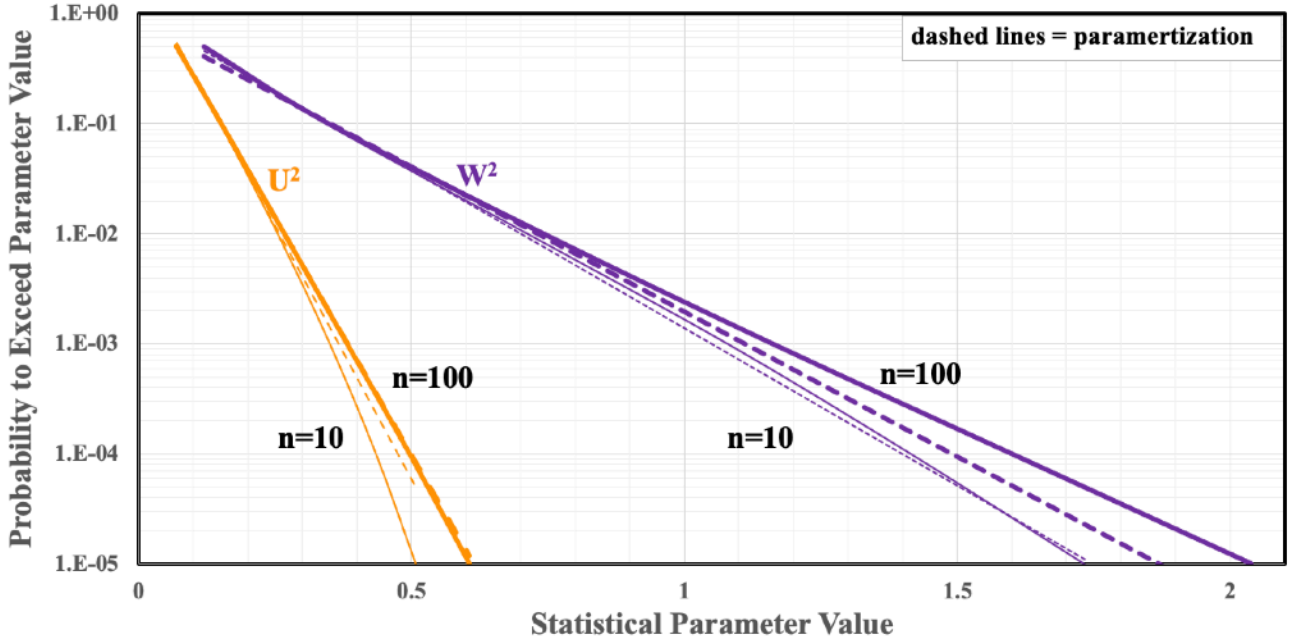
In general, the probability distributions for these statistics need to be determined by Monte Carlo calculations. However, for continuous variables tested against a well-defined model distribution under the null hypothesis, tables and approximate parameterisations exist to obtain p-values:

Test Statistic (T)	Modified Test Statistic (T*)	"High Tail" Approximate Parameterisation for P(T* > z)
D^+	$D^+(\sqrt{n}+0{\cdot}12+0{\cdot}11/\sqrt{n})$	$\exp(-2z^2)$
D^-	$D^-(\sqrt{n}+0{\cdot}12+0{\cdot}11/\sqrt{n})$	$\exp(-2z^2)$
D	$D(\sqrt{n}+0{\cdot}12+0{\cdot}11/\sqrt{n})$	$2\exp(-2z^2)$
V	$V(\sqrt{n}+0{\cdot}155+0{\cdot}24/\sqrt{n})$	$(8z^2-2)\exp(-2z^2)$
W^2	$(W^2-0{\cdot}4/n+0{\cdot}6/n^2)(1{\cdot}0+1{\cdot}0/n)$	$0{\cdot}05\exp(2{\cdot}79-6z)$
U^2	$(U^2-0{\cdot}1/n+0{\cdot}1/n^2)(1{\cdot}0+0{\cdot}8/n)$	$2\exp(-2z^2\pi^2)$

trial factor for choosing best!

M.A. Stephens, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 32, No. 1. (1970), pp. 115-122


Probability to Exceed Parameter Value
1.E+00
1.E-01
1.E-02
1.E-03
1.E-04
1.E-05
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
Statistical Parameter Value
D±
D
V
dashed lines = paramertization
n=100
n=10


Probability to Exceed Parameter Value
1.E+00
1.E-01
1.E-02
1.E-03
1.E-04
1.E-05
0
0.5
1
1.5
2
Statistical Parameter Value
dashed lines = paramertization
U²
W²
n=100
n=10
n=100
n=10