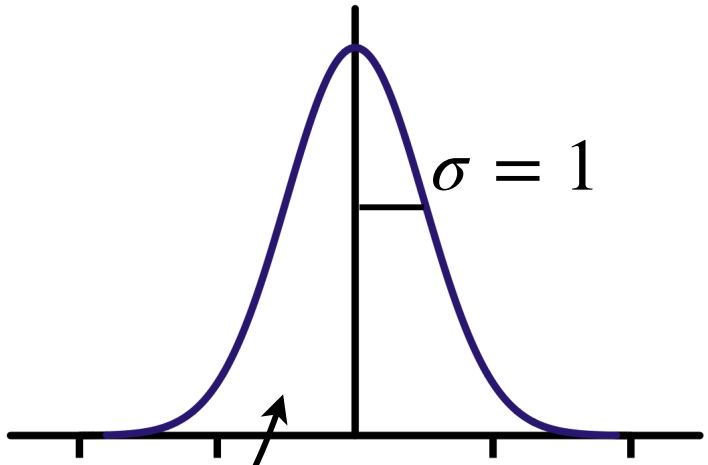
Lecture 5:

- Testing Models: chi-squared
- p-values
- Combined p-values as a Statistic

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

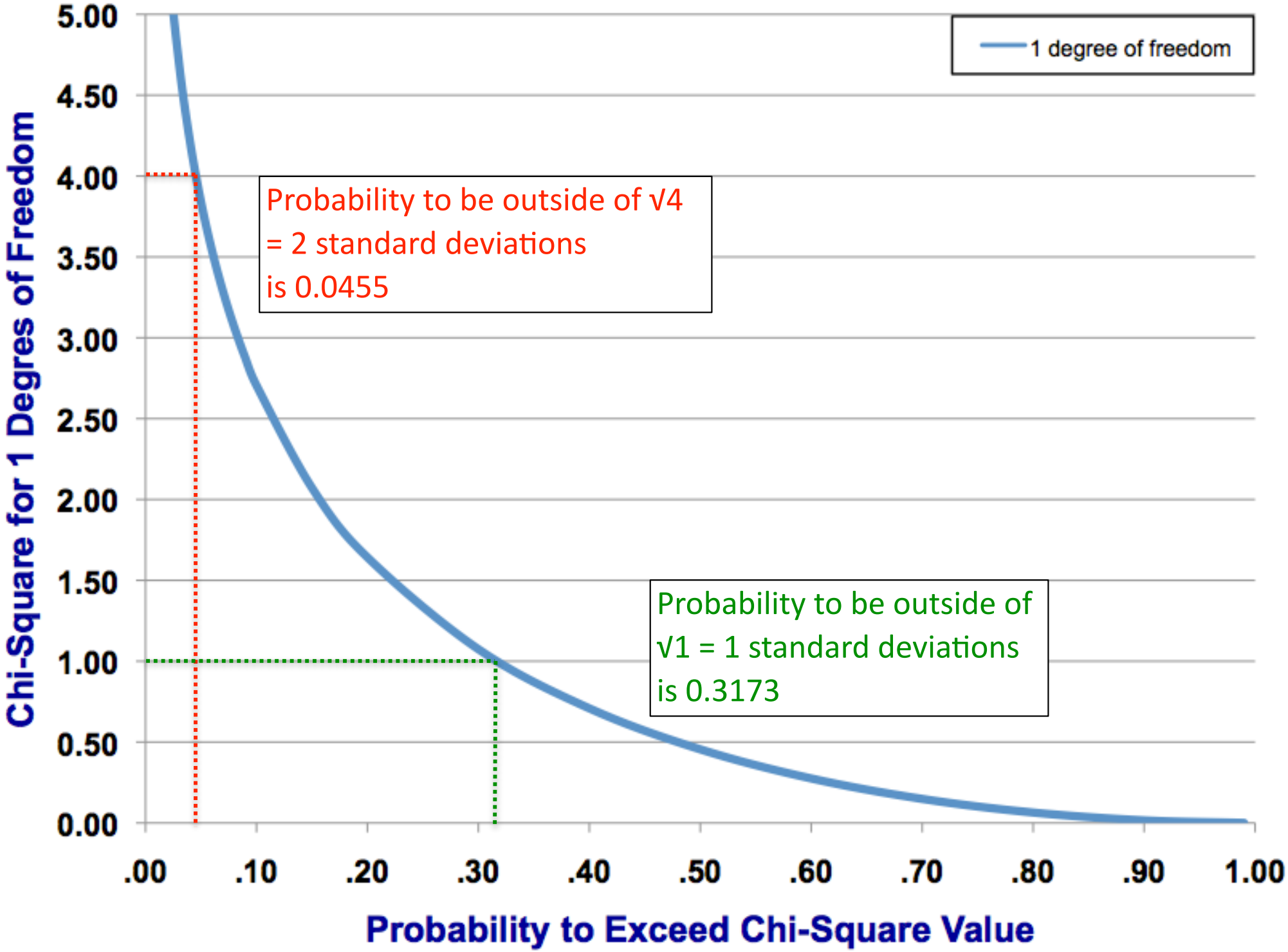
Note that: $P(\chi^2,2) = \frac{1}{2}e^{-\frac{\chi^2}{2}}$

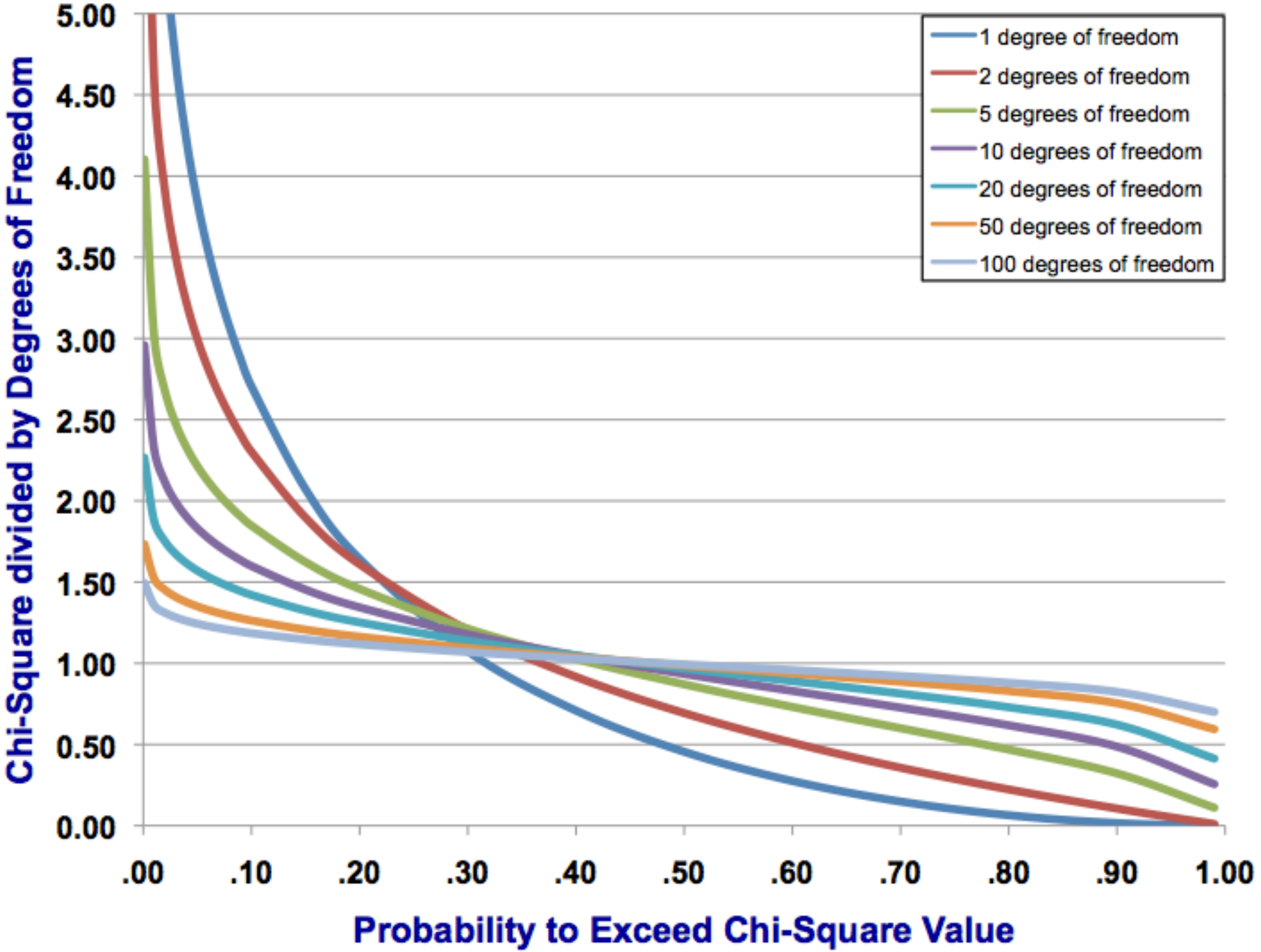
(will come back to this)

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


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
Chi-Square for 1 Degres 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


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:

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

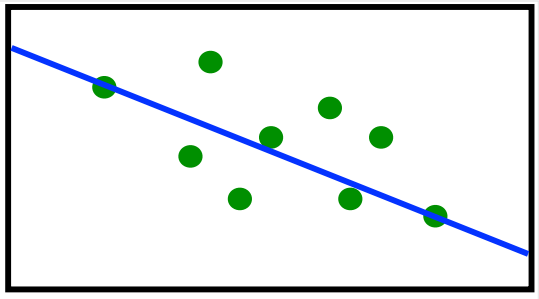
normalised to give a Gaussian distribution with unit variance

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:

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

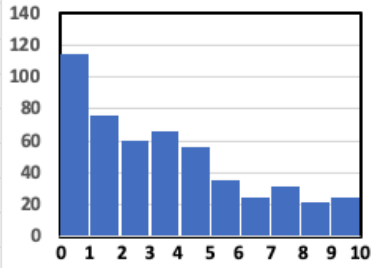


Table of ((n-m)^2)/m for different values of t0

t0:	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	8							2
	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	1.02	0.896	0.854	0.858	0.89	0.93	0.99	1.05	1.12								175
	6.72	3.89	2.48	1.707	1.252	0.969	0.78	0.66	0.57	0.51	0.46								491
	1.52	0.22	0	0.189	0.51	0.861	1.2	1.52	1.8	2.05	2.28								175
	1.05	0.02	0.23	0.883	1.662	2.444	3.18	3.86	4.47	5.02	5.52								389
	20.8	10.3	5.2	2.551	1.156	0.441	0.11	0	0.03	0.13	0.28								983
	12.7	4.98	1.68	0.366	0.005	0.126	0.49	0.98	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								805

How many degrees of freedom?

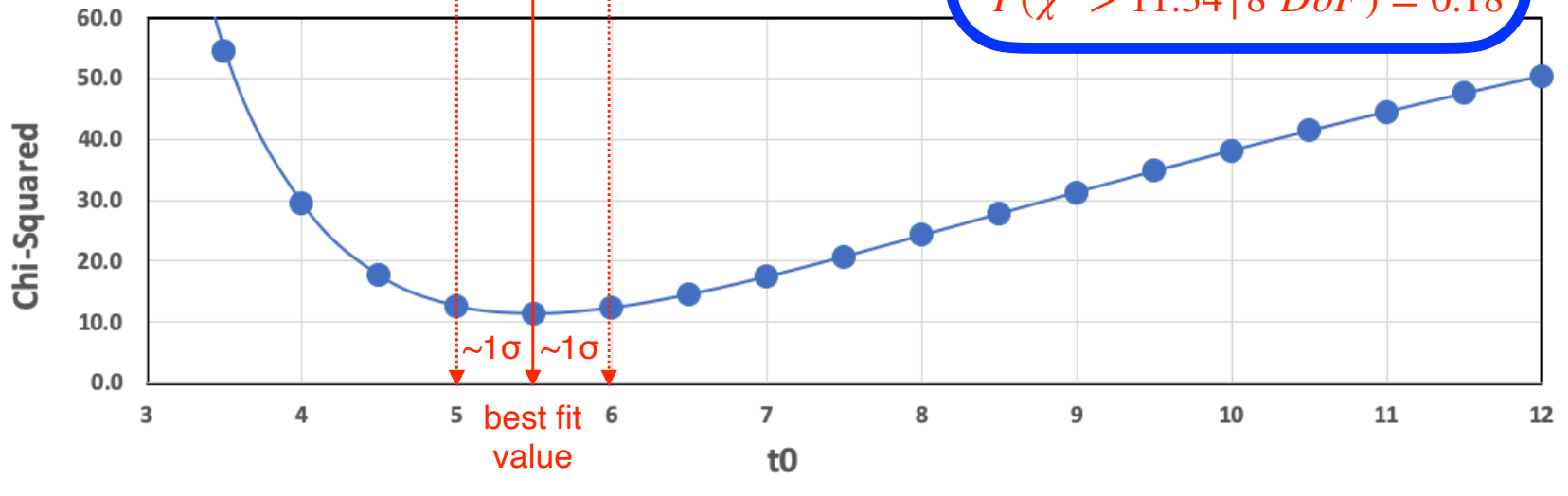
- 10 independent data points
- fit parameter t_o
- but normalisation is also based on observed data (for single bin, variance would be zero)

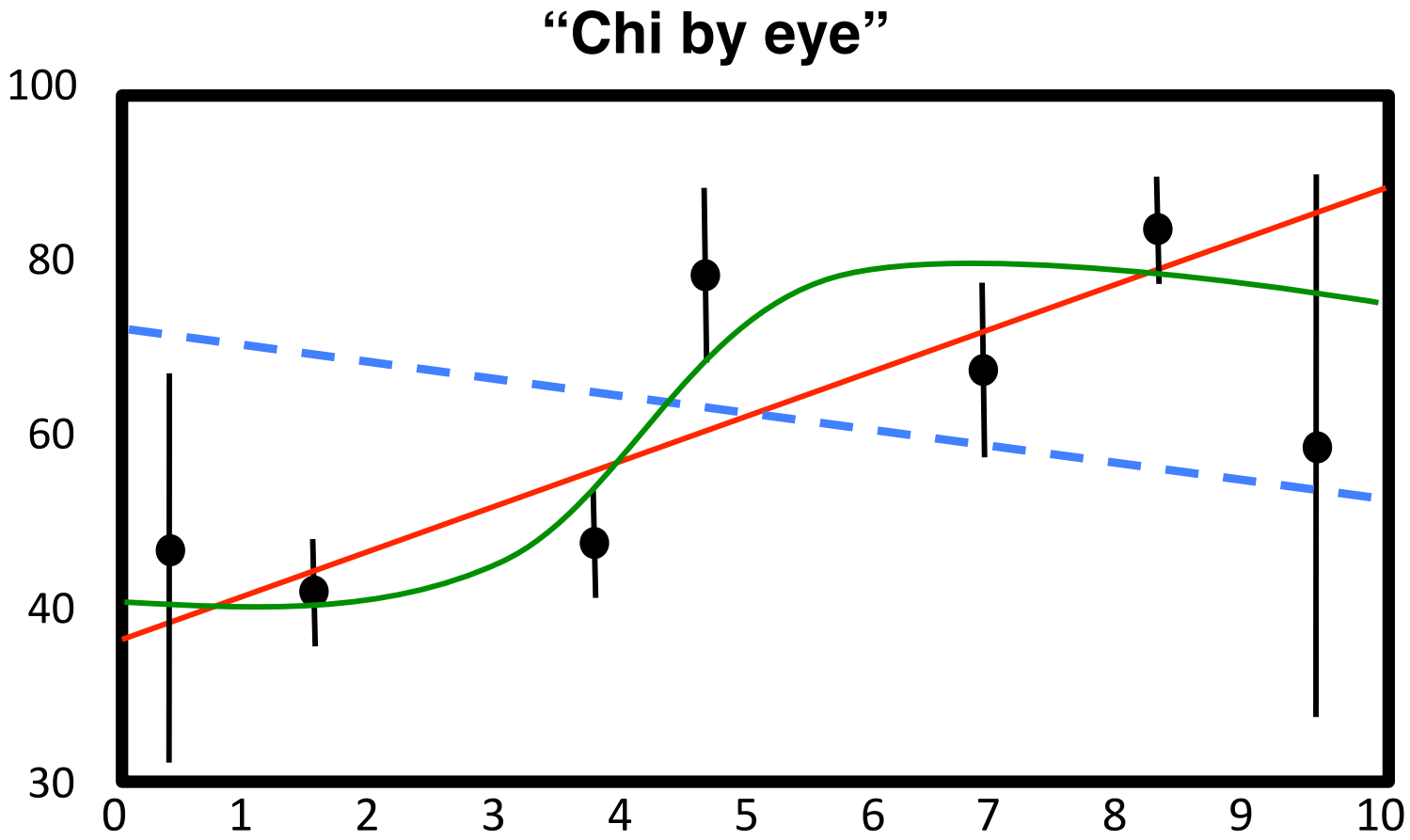
DoF = 10 - 2 = 8

How good is the fit?

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

^{222}Rn mean lifetime = 5.51 days

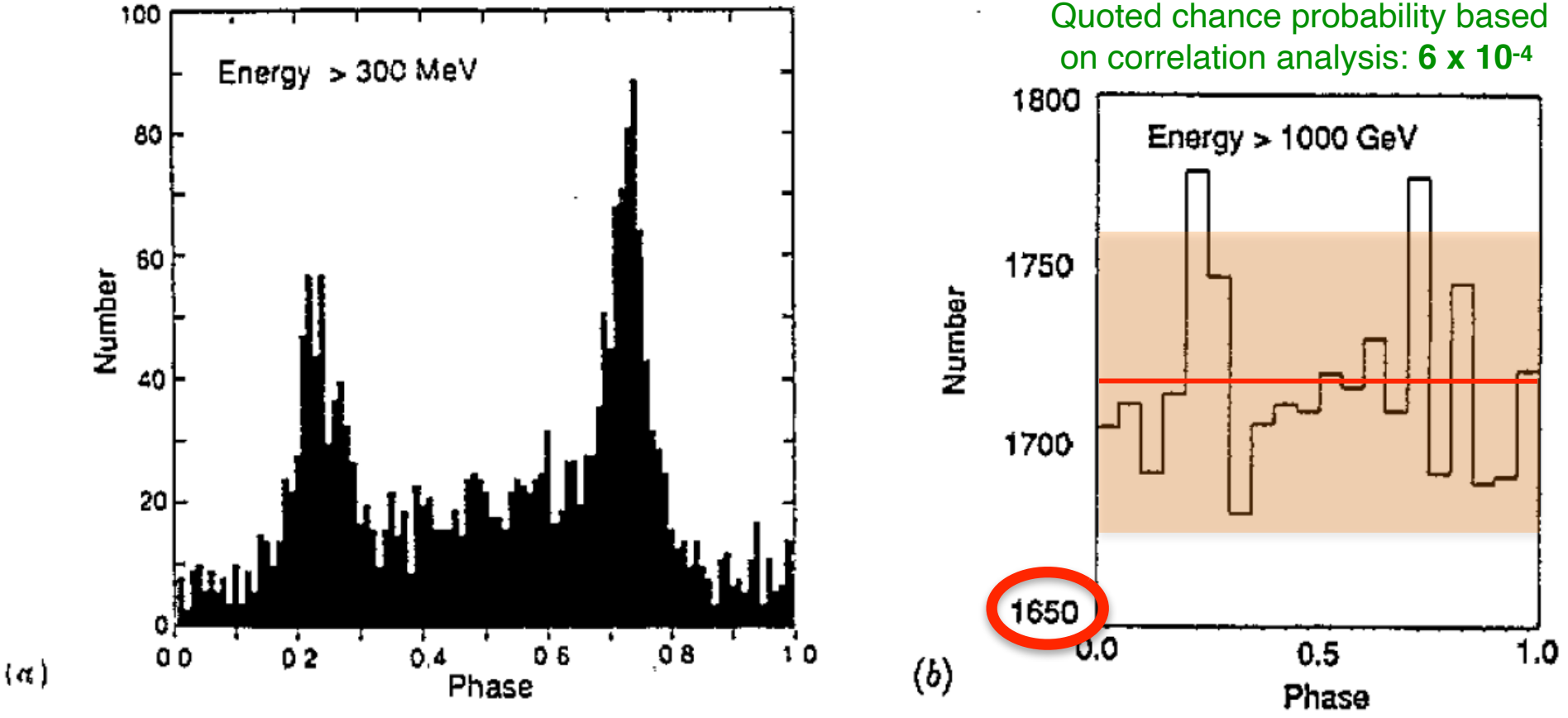




$$\chi^2 \sim (0.3)^2 + (0.1)^2 + (1.2)^2 + (1.7)^2 + (0.3)^2 + (0.8)^2 + (0.8)^2 = 5.8$$

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.

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

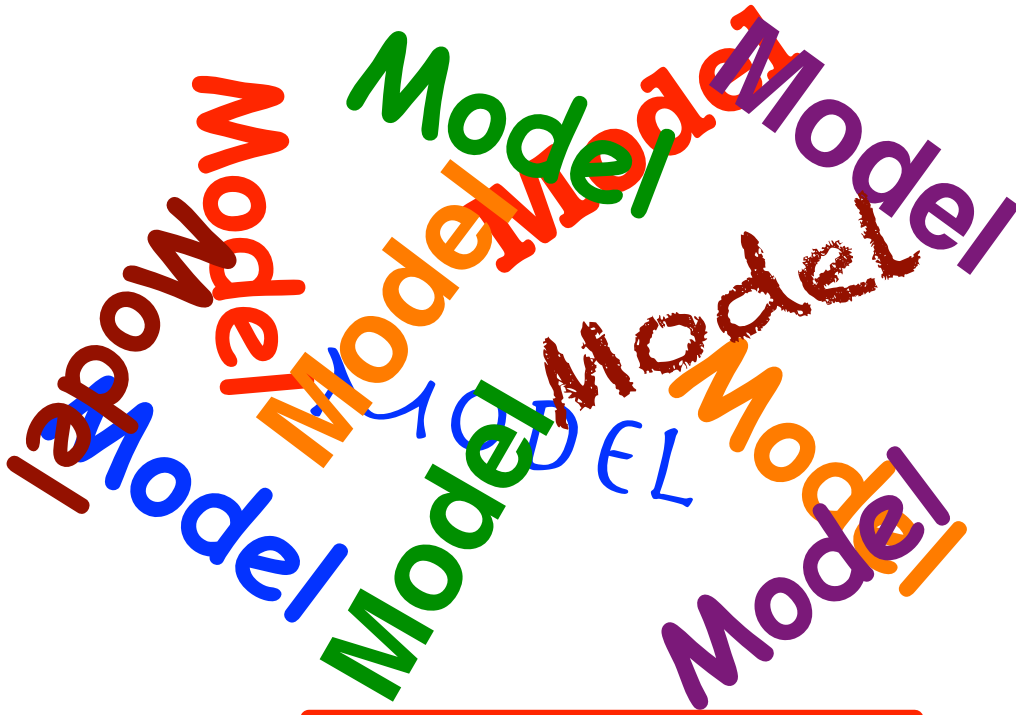
Avg (m) = 1718.45

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

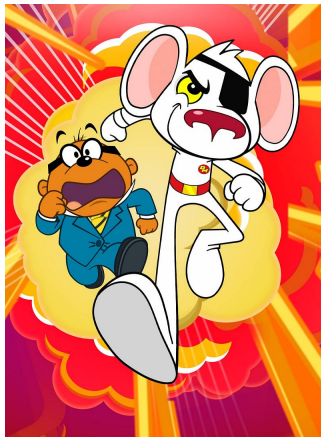
Model

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.”**

A common quantity to compute when testing the null hypothesis:



p-value ("chance probability"):

The probability of obtaining a value of some parameter at least as extreme as that which is observed, assuming the null hypothesis is true.

"How much does this particular data set look like what is expected from the null hypothesis?"

But the p-value is **NOT** the probability of the null hypothesis being true or of the alternative hypothesis being false!

Example 1:

Search for Episodic X-Ray Emission

Over the course of a year, 36000 x-rays are observed to come from a particular astrophysical object. However, on one particular day, 130 events are observed. What is the statistical significance of this observed burst?

$$\langle x \rangle = \frac{36000}{365} = 98.6 \qquad \mu \simeq \langle x \rangle \qquad \sigma = \sqrt{\mu}$$

$$s \cong \frac{(130-98.6)}{\sqrt{98.6}} = 3.16\sigma$$

odds of getting at least this many events by a chance fluctuation from the average rate of emission

$$P = 8 \times 10^{-4}$$

p-value for this test, but need to look at it in the context of all other tests

Is this sufficient to claim the observation of a burst from this object?

Correct question:

What is the chance of seeing at least one burst with an excess at least as large given the number of independent tests I've done ?

Binomial !!

N Bernoulli trials where the chance of each success is **P**

$$\sum_{i=1}^{\infty} \binom{N}{i} P^i (1-P)^{N-i} = 1 - \binom{N}{0} P^0 (1-P)^{N-0}$$

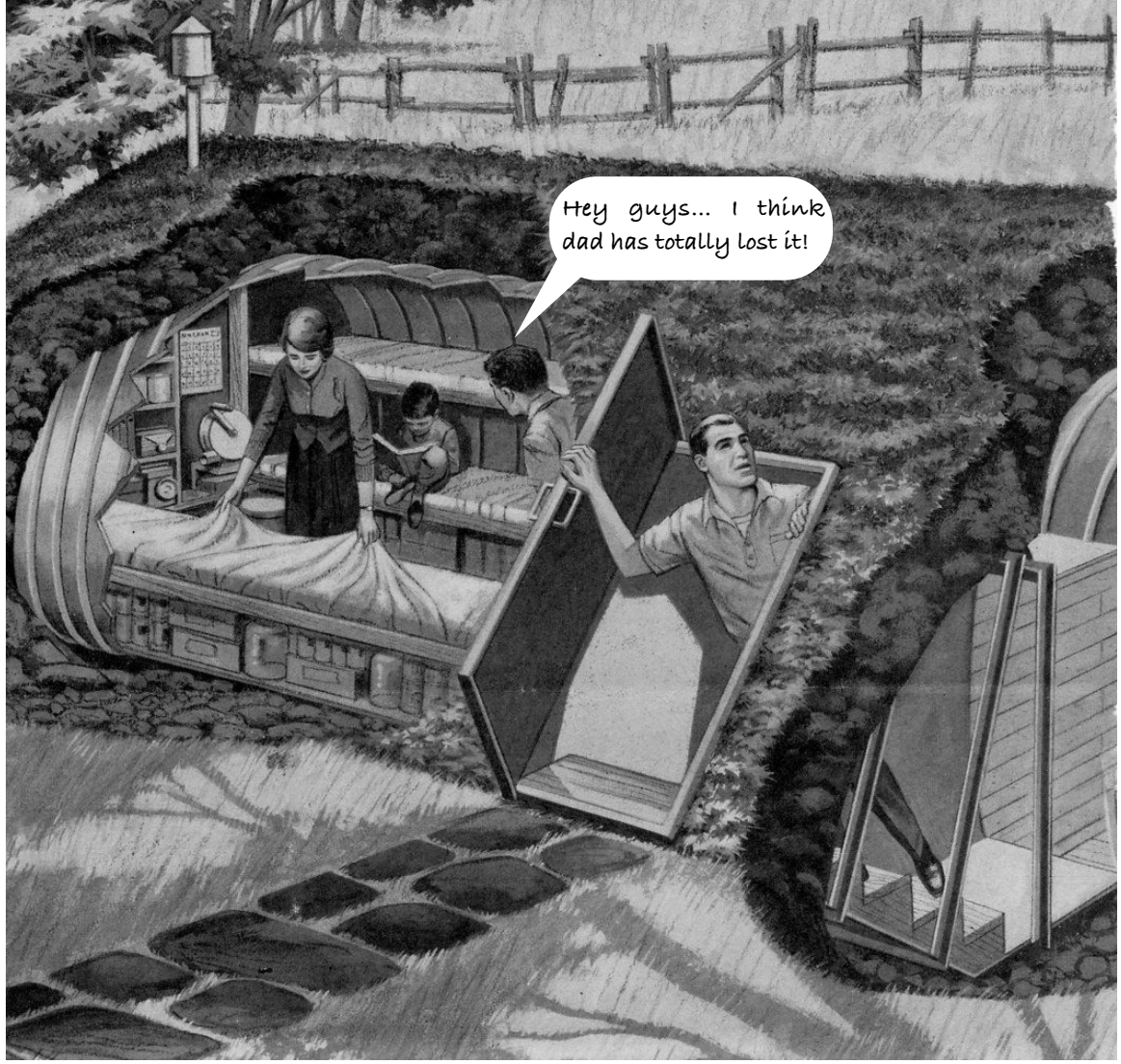
$$\mathrm{P}_{\text{post-trial}} = 1 - (1-P)^N \qquad (\sim NP \quad \text{for} \quad NP << 1)$$

$\mathrm{P} = 8 \times 10^{-4}$, $\mathrm{N} = 365$ → $\mathrm{P}_{\text{post-trial}} = 25\%$

How many timescales were considered? How many objects examined?

Example 2:

During his year of self-isolating, Dave peered out of his bunker on six random occasions and found that it was always dark.



Assuming that the earth goes around the sun, you would expect it to be dark about half the time, averaged over the year. So the chance probability for it to be dark outside on all six occasions is:

$$P(dark\ all\ 6\ times) = (0.5)^6 = 0.0156$$

Importance of prior probabilities (more on this later)

"Gosh, That's pretty small! **Hey everyone, it looks like there's a very good chance that we're no longer going around the sun!!**"

Pragmatism!

Look carefully at context:

Very small p-values, even after careful accounting of trials, confirmed by independent observations, which could be explained by plausible alternative hypotheses…

Reject H0

Combination of p-values

Two identical experiments observe evidence of the brexiton (a particle now outside of the Standard Model that inevitably then decays to a less attractive state). The first experiment assess the odds that their observation is due to chance fluctuations as being 1%, while the second assesses their observation to have a chance probability of 10%. What is the combined chance probability that these two data sets are consistent with the null hypothesis (*i.e.* there is no brexiton)?

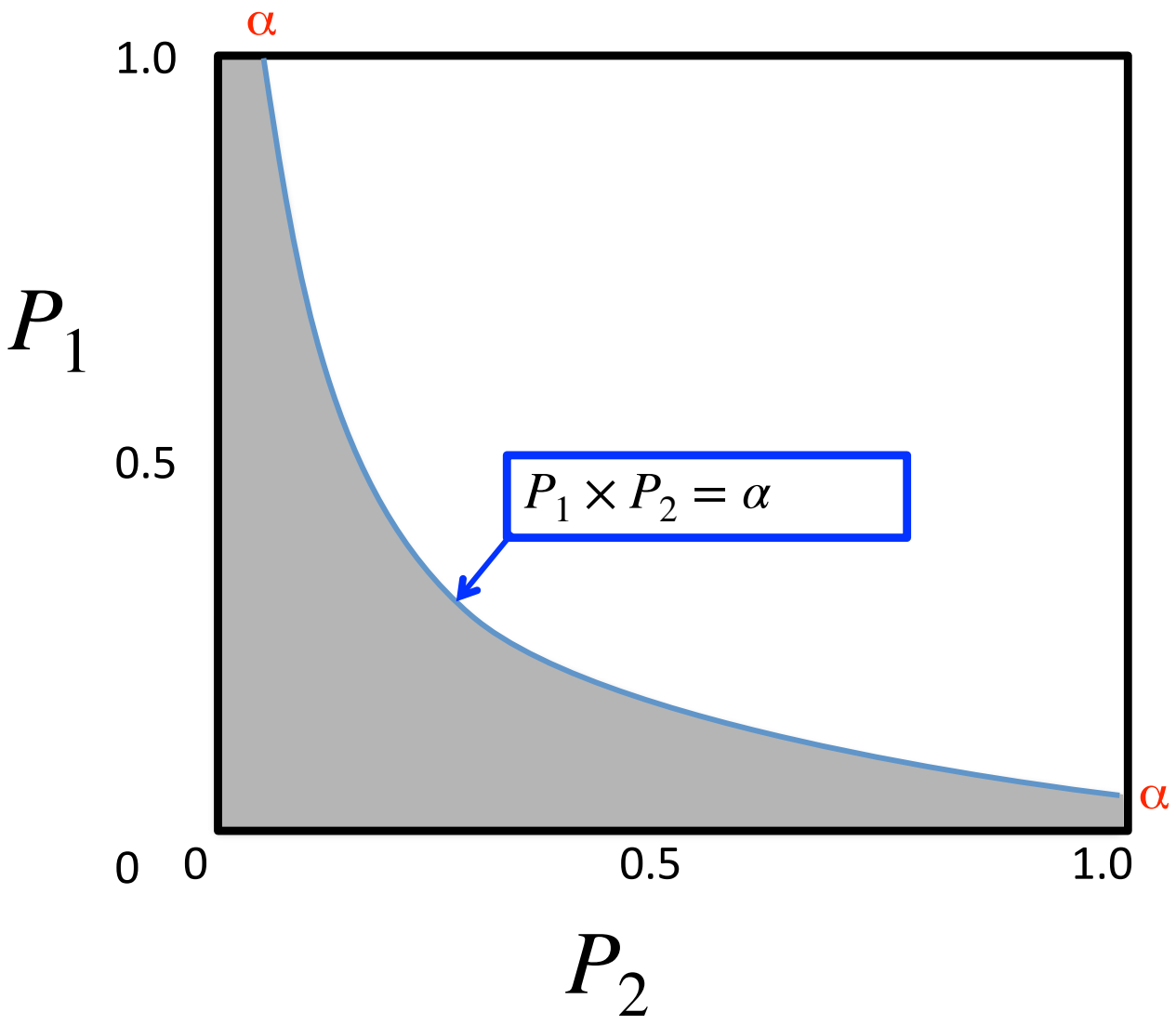
$$P_1 \times P_2 = 0.001 \ ?$$

Need to look at properties of the product:

Define the statistic: $\Gamma \equiv P_1 \times P_2$

What is the chance probability for $\boldsymbol{\Gamma}$

to be at least as small as some value $\boldsymbol{\alpha}$?



Integrated area under the curve:

$$\alpha\,(\,1 - \ln \alpha\,)$$

$$= P(\alpha)$$

i.e. this is the chance that a background fluctuation would yield a value of Γ that is at least as small as α.

So, for the case here: $\alpha = (0.01)(0.1) = 0.001$

$$P(\leq \alpha) = 0.001(1 - \ln(0.001)) = 0.004$$

More generally…

Fisher's Method

$$F \equiv -2\ln\left(\prod_{i=1}^{n} p_i(\leq p_{obs})\right) = \sum_{i=1}^{n}\left(-2\ln p_i(\leq p_{obs})\right) \equiv \sum_{i=1}^{n} f_i$$

$$p_i(\leq p_{obs}) = e^{-\frac{f_i}{2}}$$

or $p_i(> p_{obs}) = 1 - e^{-\frac{f_i}{2}}$

$$p_i^{diff}(x) = \frac{1}{2}e^{-\frac{f_i}{2}}$$

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

so f_i values are distributed like a χ^2 distribution with 2 DoF

and we can express:

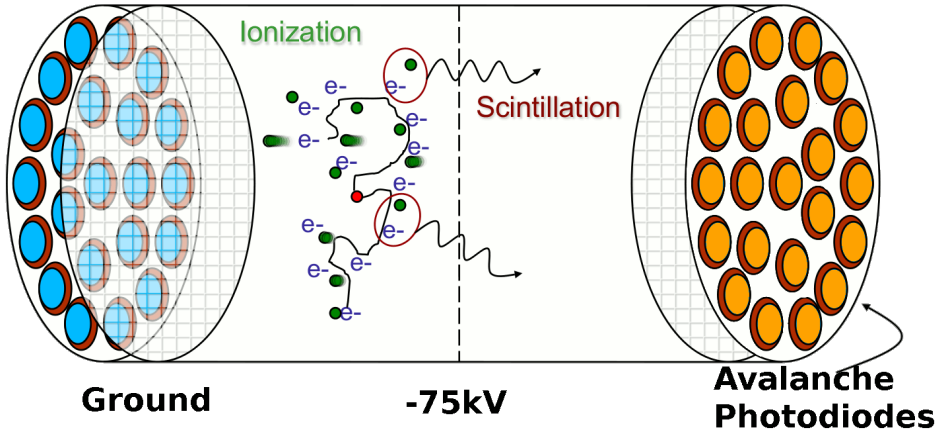
$$\chi^2 = \sum_{i=1}^{k(\equiv 2n)} g_i^2 = \sum_{i=1}^{n} \overbrace{\left(g_{2i-1}^2 + g_{2i}^2\right)}^{\chi_i^2}$$

$$P(\chi^2,2n) = \sum_{i=1}^{n} P(\chi_i^2,2)$$

F is distributed like a χ^2 distribution with 2n DoF

Example:

The EXO experiment searches for evidence of neutrinoless double beta decay, which produces 2 electrons with a total energy that is well defined. The interaction produces scintillation light in the liquid xenon target, and the ionisation tracks of charged particles are also drifted to a readout plane to record the time and position of charges. Backgrounds come from radioactivity in the xenon and, to a greater extent, from the walls of the detector.



Assume that an event is observed and the chance probability for it to be background is assessed using several independent measures:

Event energy estimated from the scintillation light: $P_{scint} = 0.14$

Event energy estimated from the total charge: $P_{charge} = 0.05$

The proximity of the event to the cavity walls: $P_{charge} = 0.32$

The density of charge deposition (event topology): $P_{charge} = 0.53$

What is the overall chance probability (p-value) that this event is background?

$$-2\log(0.14 \times 0.05 \times 0.32 \times 0.53) = 13.43$$

$$P(\chi^2 > 13.43,\ DoF = 2 \times 4) = 0.10$$