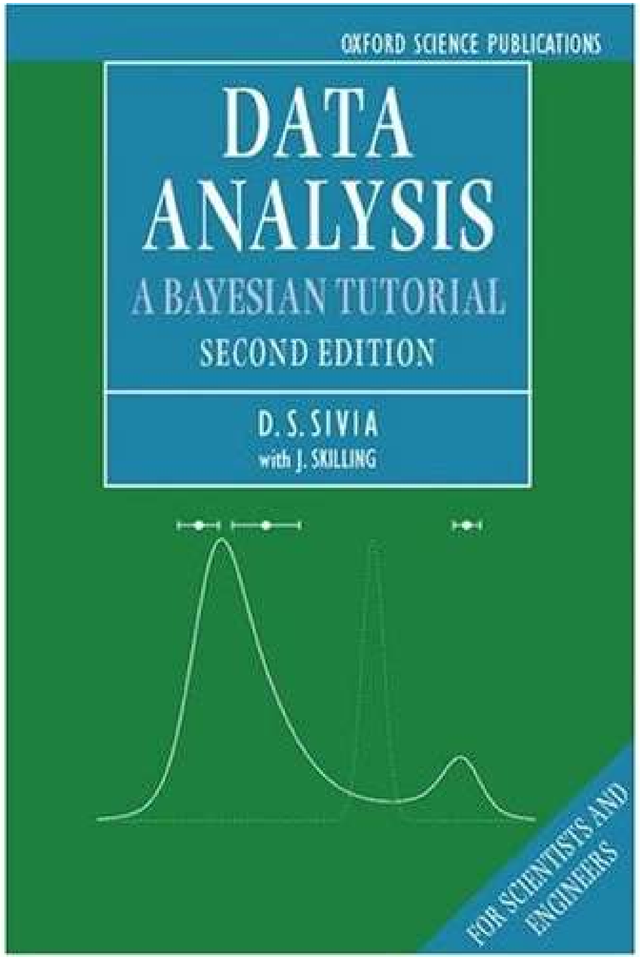
4. Parameter Estimation and Goodness of Fit – part two



Sivia Chapter 3 gives a very clear discussion of least squares fitting within a Bayesian framework.

In particular, contrasts, for Gaussian residuals:

- known σ
- unknown σ $\rightarrow$ Student's t

The principle of maximum likelihood

Frequentist approach:

A parameter is a fixed (but unknown) constant

From actual data we can compute Likelihood,

L = probability of obtaining the observed data, given the value of the parameter θ

The principle of maximum likelihood

Frequentist approach:

A parameter is a fixed (but unknown) constant

From actual data we can compute Likelihood,

L = probability of obtaining the observed data, given the value of the parameter θ

Now define **likelihood function**: (infinite) family of curves generated by regarding L as a function of θ , for data fixed.

Principle of Maximum Likelihood

A good estimator of θ maximises L -

i.e. $\dfrac{\partial L}{\partial \theta} = 0$ and $\dfrac{\partial^2 L}{\partial \theta^2} < 0$

The principle of maximum likelihood

Frequentist approach:

A parameter is a fixed (but unknown) constant

From actual data we can compute Likelihood,

L = probability of obtaining the observed data, given the value of the parameter θ

Now define **likelihood function**: (infinite) family of curves generated by regarding L as a function of θ, for data fixed.

Principle of Maximum Likelihood

A good estimator of θ maximises L -

i.e. $\frac{\partial L}{\partial \theta} = 0$ and $\frac{\partial^2 L}{\partial \theta^2} < 0$

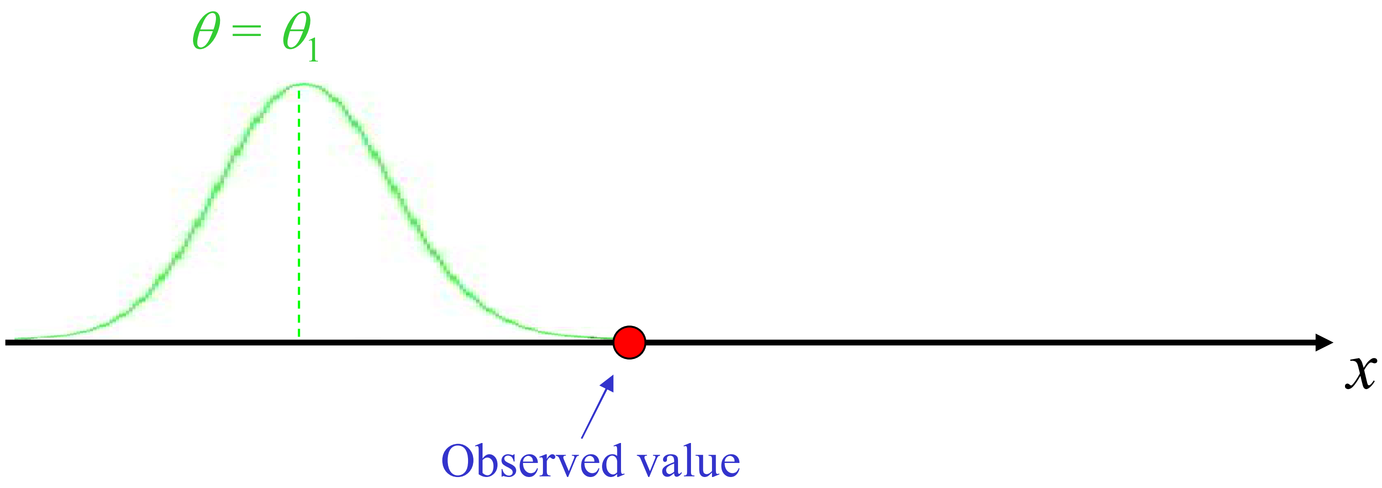
We set the parameter equal to the value that makes the actual data sample we *did* observe – out of all the possible random samples we *could have* observed – the most likely.

Aside: Likelihood function has same definition in Bayesian probability theory, but subtle difference in meaning and interpretation – no need to invoke idea of (infinite) ensemble of different samples.

Principle of Maximum Likelihood

A good estimator of θ maximises L -

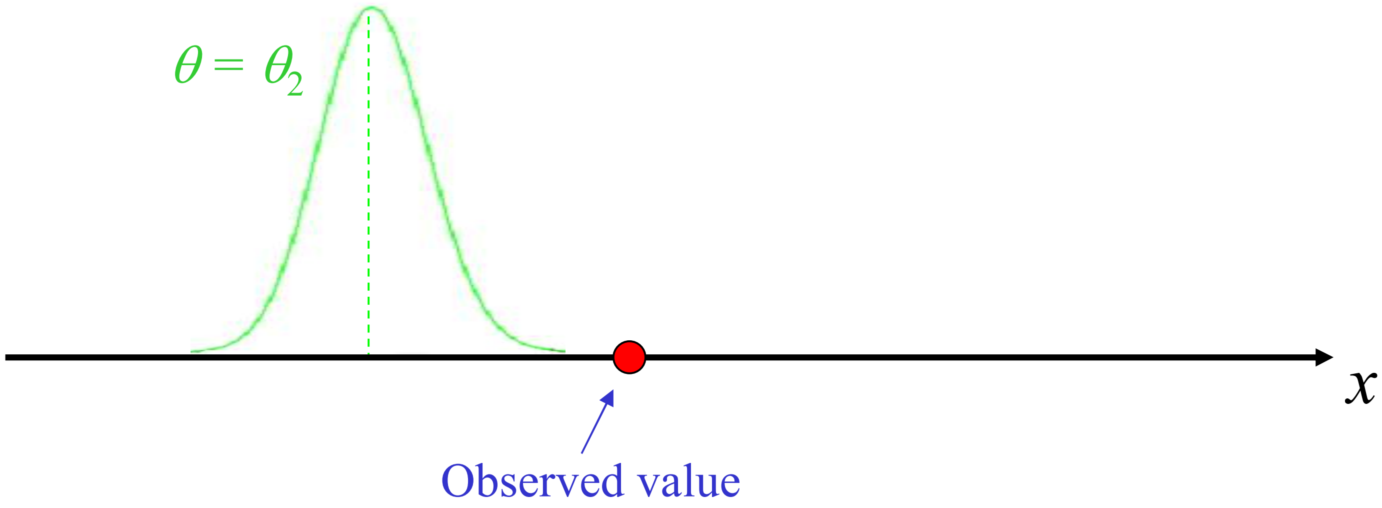
i.e. $\frac{\partial L}{\partial \theta} = 0$ and $\frac{\partial^2 L}{\partial \theta^2} < 0$



Principle of Maximum Likelihood

A good estimator of θ maximises L -

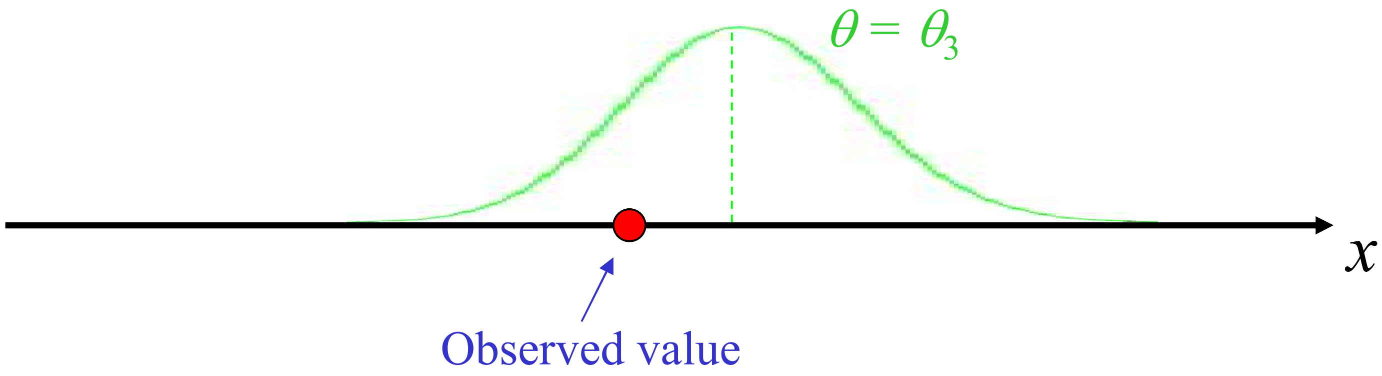
$$\text{i.e.} \quad \frac{\partial L}{\partial \theta} = 0 \quad \text{and} \quad \frac{\partial^2 L}{\partial \theta^2} < 0$$



Principle of Maximum Likelihood

A good estimator of θ maximises L -

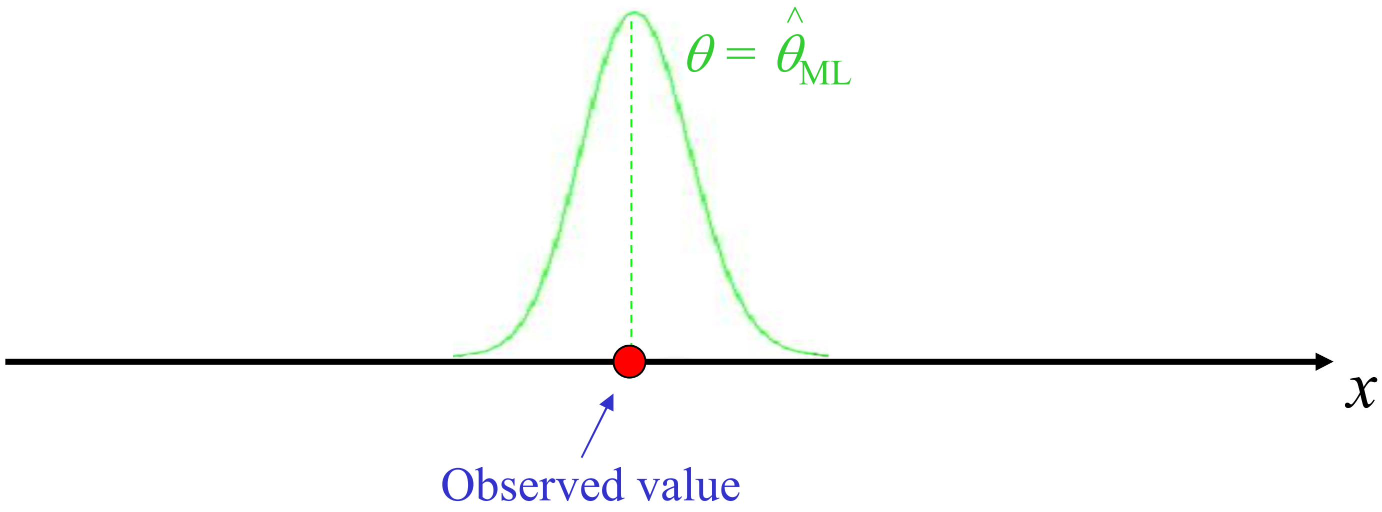
$$\text{i.e.} \quad \frac{\partial L}{\partial \theta} = 0 \quad \text{and} \quad \frac{\partial^2 L}{\partial \theta^2} < 0$$



Principle of Maximum Likelihood

A good estimator of θ maximises L -

i.e. $\frac{\partial L}{\partial \theta} = 0$ and $\frac{\partial^2 L}{\partial \theta^2} < 0$



Least squares as maximum likelihood estimators

To see the maximum likelihood method in action, let's consider again **weighted least squares** for the simple model $y_i \quad = \quad a + bx_i + \epsilon_i$

Suppose the i^{th} residual, $\{\epsilon_i\}$, is assumed to be drawn from some underlying pdf with mean zero and variance σ_i^2, where the variance is allowed to be different for each residual.

Let's assume the pdf is a Gaussian

Least squares as maximum likelihood estimators

To see the maximum likelihood method in action, let's consider again weighted least squares for the simple model $y_i = a + bx_i + \epsilon_i$

Suppose the i^{th} residual, $\{\epsilon_i\}$, is assumed to be drawn from some underlying pdf with mean zero and variance σ_i^2, where the variance is allowed to be different for each residual.

Let's assume the pdf is a Gaussian

Likelihood $$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i}} \exp\left[-\frac{1}{2} \frac{\varepsilon_i^2}{\sigma_i^2} \right]$$

Question 7: How can we justify writing the likelihood as a product?

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2} \frac{\varepsilon_i^2}{\sigma_i^2} \right]$$

A Because the residuals are all equal to each other

B Because the residuals are all Gaussian

C Because the residuals are all positive

D Because the residuals are all independent

Least squares as maximum likelihood estimators

To see the maximum likelihood method in action, let's consider again weighted least squares for the simple model $y_i = a + bx_i + \epsilon_i$

Suppose the i^{th} residual, $\{\epsilon_i\}$, is assumed to be drawn from some underlying pdf with mean zero and variance σ_i^2, where the variance is allowed to be different for each residual.

Let's assume the pdf is a Gaussian

Likelihood
$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\frac{\varepsilon_i^2}{\sigma_i^2}\right]$$

(note: L is a product of 1-D Gaussians because we are assuming the ε_i are independent)

Substitute $\varepsilon_i = y_i - a - bx_i$

$$\Rightarrow \qquad L \;=\; \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2} \frac{\left(y_i - a - bx_i\right)^2}{\sigma_i^2} \right]$$

and the ML estimators of a and b satisfy $\partial L / \partial a = 0$ and $\partial L / \partial b = 0$

Substitute $\varepsilon_i = y_i - a - bx_i$

$$\Rightarrow \qquad L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2} \frac{(y_i - a - bx_i)^2}{\sigma_i^2} \right]$$

and the ML estimators of a and b satisfy $\partial L / \partial a = 0$ and $\partial L / \partial b = 0$

But maximising L is equivalent to maximising $\ell = \ln L$

Here

$$\ell = -\frac{n}{2}\ln(2\pi) - \ln\sum_{i=1}^{n}\sigma_i - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - a - bx_i}{\sigma_i}\right)^2$$

$$= \text{constant} - \frac{1}{2}S$$

This is exactly the same sum of squares we defined earlier

Substitute $\varepsilon_i = y_i - a - bx_i$

$$\Rightarrow \qquad L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i}} \exp\left[-\frac{1}{2} \frac{(y_i - a - bx_i)^2}{\sigma_i^2} \right]$$

and the ML estimators of a and b satisfy $\partial L / \partial a = 0$ and $\partial L / \partial b = 0$

But maximising L is equivalent to maximising $\ell = \ln L$

Here
$$\ell = -\frac{n}{2}\ln(2\pi) - \ln\sum_{i=1}^{n}\sigma_i - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - a - bx_i}{\sigma_i}\right)^2$$

$$= \text{constant} - \frac{1}{2} S$$

This is exactly the same sum of squares we defined earlier

So in this case maximising L is **exactly equivalent** to minimising the sum of squares.

i.e. for Gaussian, independent errors, ML and weighted LS estimators are identical.

Example application

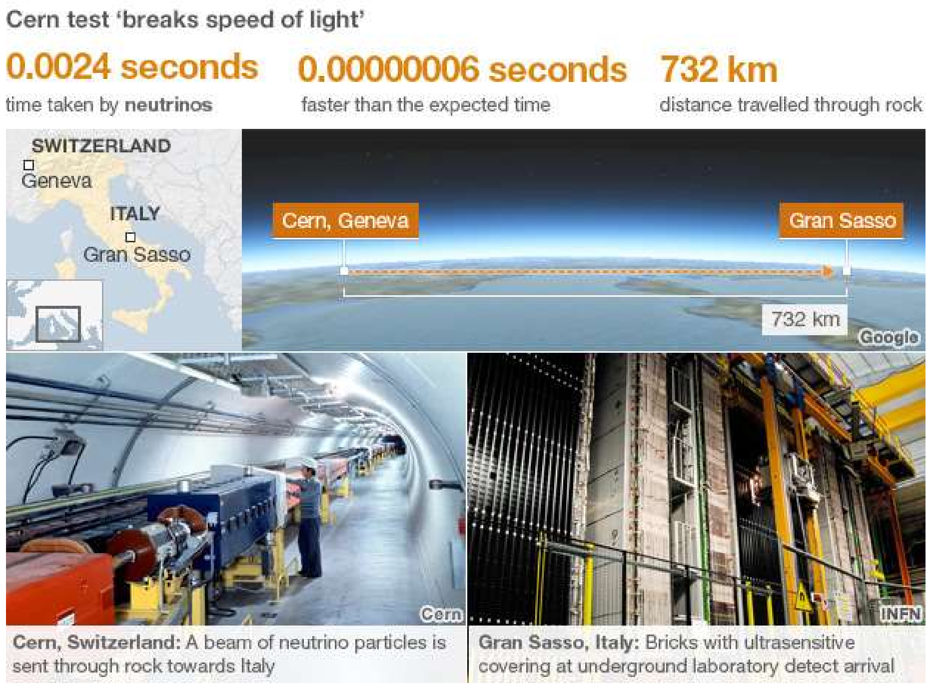
Measurement of the neutrino velocity with the OPERA detector in the CNGS beam*

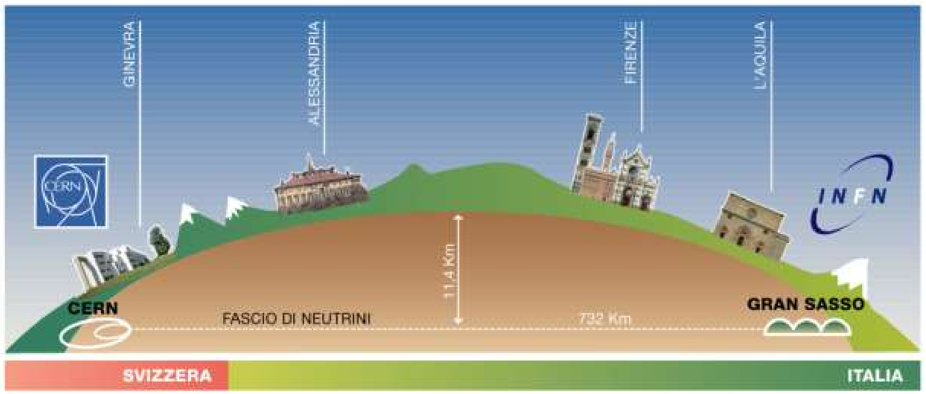
T. Adam[a], N. Agafonova[b], A. Aleksandrov[c,1], O. Altinok[d], P. Alvarez Sanchez[e], A. Anokhina[f], S. Aoki[g], A. Ariga[h], T. Ariga[h], D. Autiero[i†], A. Badertscher[j], A. Ben Dhahbi[h], A. Bertolin[k], C. Bozza[l], T. Brugière[i], R. Brugnera[m,k], F. Brunet[n], G. Brunetti[o,i,2], S. Buontempo[c], B. Carlus[i], F. Cavanna[q], A. Cazes[i], L. Chaussard[i], M. Chernyavsky[r], V. Chiarella[s], A. Chukanov[t], G. Colosimo[u], M. Crespi[u], N. D'Ambrosio[v], G. De Lellis[w,c], M. De Serio[x], Y. Déclais[i], P. del Amo Sanchez[n], F. Di Capua[c], A. Di Crescenzo[w,c], D. Di Ferdinando[p], N. Di Marco[v], S. Dmitrievsky[t], M. Dracos[a], D. Duchesneau[n], S. Dusini[k], J. Ebert[y], I. Efthymiopoulos[e], O. Egorov[z], A. Ereditato[h], L.S. Esposito[j], J. Favier[n], T. Ferber[y], R.A. Fini[x], T. Fukuda[aa], A. Garfagnini[m,k], G. Giacomelli[o,p], M. Giorgini[o,p,3], M. Giovannozzi[e], C. Girerd[i], J. Goldberg[ab], C. Göllnitz[y], D. Golubkov[z], L. Goncharova[r], Y. Gornushkin[t], G. Grella[l], F. Grianti[s,ac], E. Gschwendtner[e], C. Guerin[i], A.M. Guler[d], C. Gustavino[ad], C. Hagner[y], K. Hamada[ae], T. Hara[g], M. Hierholzer[y], A. Hollnagel[y], M. Ieva[x], H. Ishida[aa], K. Ishiguro[ae], K. Jakovcic[af], C. Jollet[a], M. Jones[e], F. Juget[h], M. Kamiscioglu[d], J. Kawada[h], S.H. Kim[ag,4], M. Kimura[aa], E. Kiritsis[ah], N. Kitagawa[ae], B. Klicek[af], J. Knuesel[h], K. Kodama[ai], M. Komatsu[ae], U. Kose[k], I. Kreslo[h], C. Lazzaro[j], J. Lenkeit[y], A. Ljubicic[af], A. Longhin[s], A. Malgin[b], G. Mandrioli[p], J. Marteau[i], T. Matsuo[aa], N. Mauri[s], A. Mazzoni[u], E. Medinaceli[m,k], F. Meisel[h], A. Meregaglia[a], P. Migliozzi[c], S. Mikado[aa], D. Missiaen[e], K. Morishima[ae], U. Moser[h], M.T. Muciaccia[aj,x], N. Naganawa[ae], T. Naka[ae], M. Nakamura[ae], T. Nakano[ae], Y. Nakatsuka[ae], V. Nikitina[f], F. Nitti[ak], S. Ogawa[aa], N. Okateva[r], A. Olchevsky[t], O. Palamara[v], A. Paoloni[s], B.D. Park[ag,5], I.G. Park[ag], A. Pastore[aj,x], L. Patrizii[p], E. Pennacchio[i], H. Pessard[n], C. Pistillo[h], N. Polukhina[r], M. Pozzato[o,p], K. Pretzl[h], F. Pupilli[v], R. Rescigno[l], F. Riguzzi[al], T. Roganova[f], H. Rokujo[g], G. Rosa[am,ad], I. Rostovtseva[z], A. Rubbia[j], A. Russo[c], O. Sato[ae], Y. Sato[an], J. Schuler[a], L. Scotto Lavina[h,6], J. Serrano[e], A. Sheshukov[t], H. Shibuya[aa], G. Shoziyoev[f], S. Simone[aj,x], M. Sioli[o,p], C. Sirignano[v], G. Sirri[p], J.S. Song[ag], M. Spinetti[s], L. Stanco[k], N. Starkov[r], S. Stellacci[l], M. Stipcevic[af], T. Strauss[h], S. Takahashi[g], M. Tenti[o,p,i], F. Terranova[s,ao], I. Tezuka[an], V. Tioukov[c], P. Tolun[d], N.T. Tran[i], S. Tufanli[h], P. Vilain[ap], M. Vladimirov[r], L. Votano[s], J.-L. Vuilleumier[h], G. Wilquet[ap], B. Wonsak[y], J. Wurtz[a], C.S. Yoon[ag], J. Yoshida[ae], Y. Zaitsev[z], S. Zemskova[t], A. Zghiche[n]

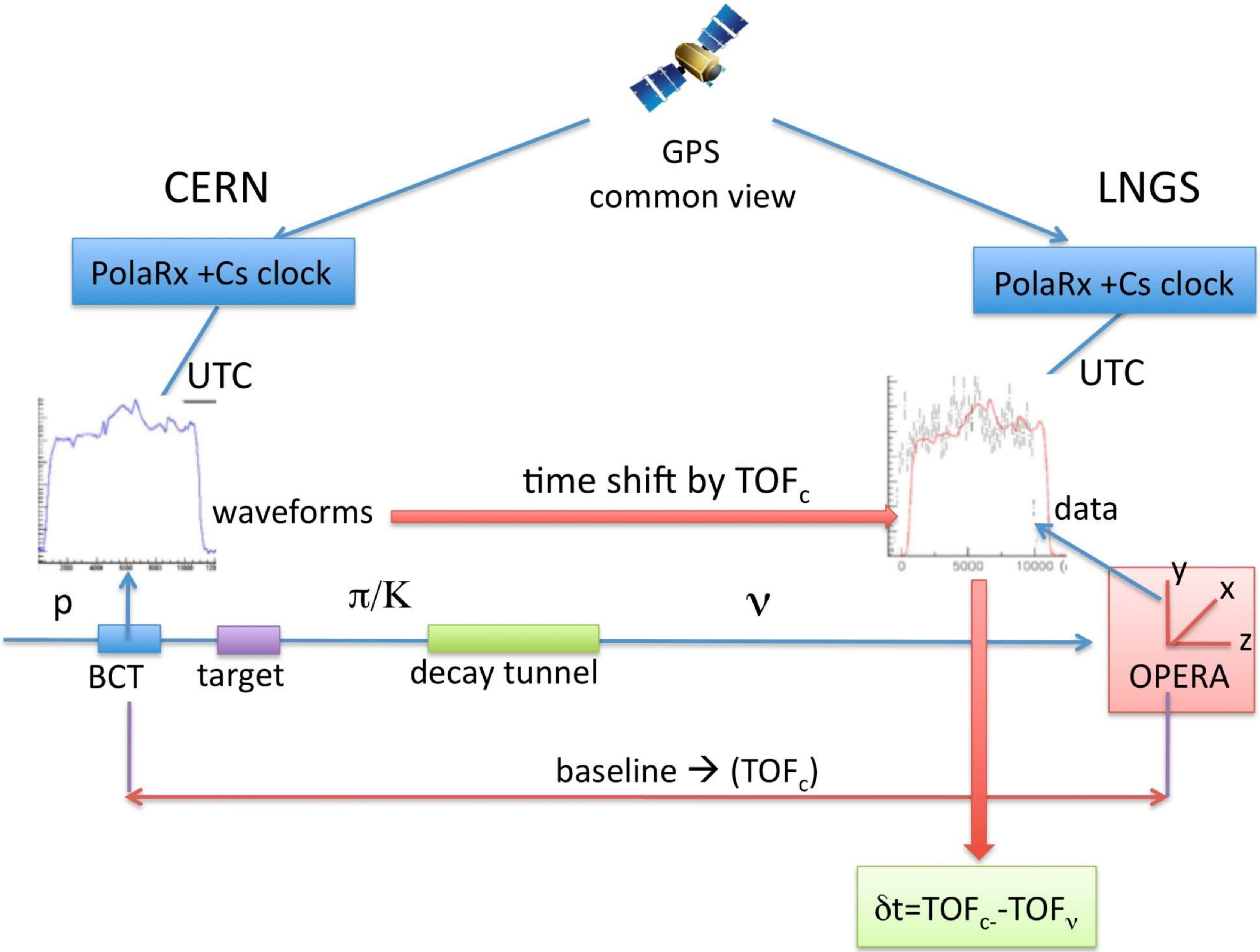
[a] IPHC, Université de Strasbourg, CNRS/IN2P3, F-67037 Strasbourg, France
[b] INR-Institute for Nuclear Research of the Russian Academy of Sciences, RUS-327312 Moscow, Russia
[c] INFN Sezione di Napoli, I-80125 Napoli, Italy
[d] METU-Middle East Technical University, TR-06532 Ankara, Turkey
[e] European Organization for Nuclear Research (CERN), Geneva, Switzerland
[f] (MSU SINP) Lomonosov Moscow State University Skobeltsyn Institute of Nuclear Physics, RUS-119992 Moscow, Russia
[g] Kobe University, J-657-8501 Kobe, Japan
[h] Albert Einstein Center for Fundamental Physics, Laboratory for High Energy Physics (LHEP), University of Bern, CH-3012 Bern, Switzerland
[i] IPNL, Université Claude Bernard Lyon I, CNRS/IN2P3, F-69622 Villeurbanne, France
[j] ETH Zurich, Institute for Particle Physics, CH-8093 Zurich, Switzerland

* *Preprint submitted to the Journal of High Energy Physics (17 November 2011)*
† *Corresponding author* Dario.Autiero@cern.ch

1

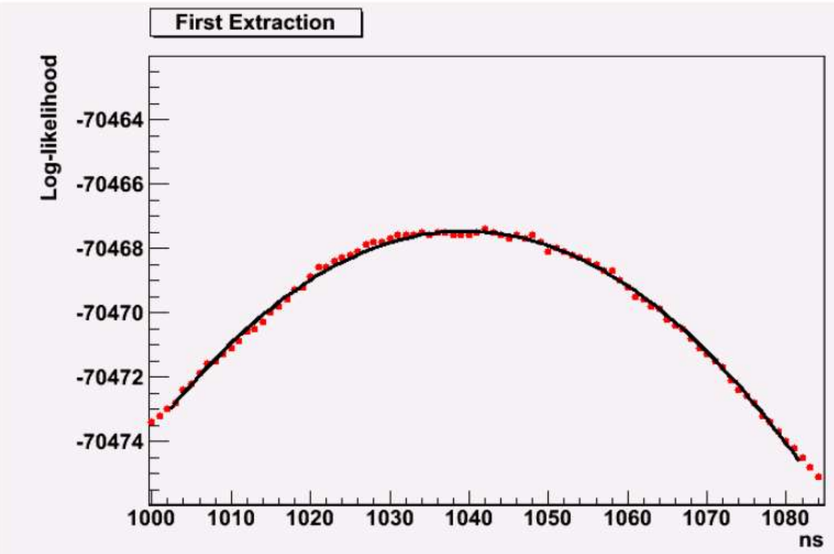


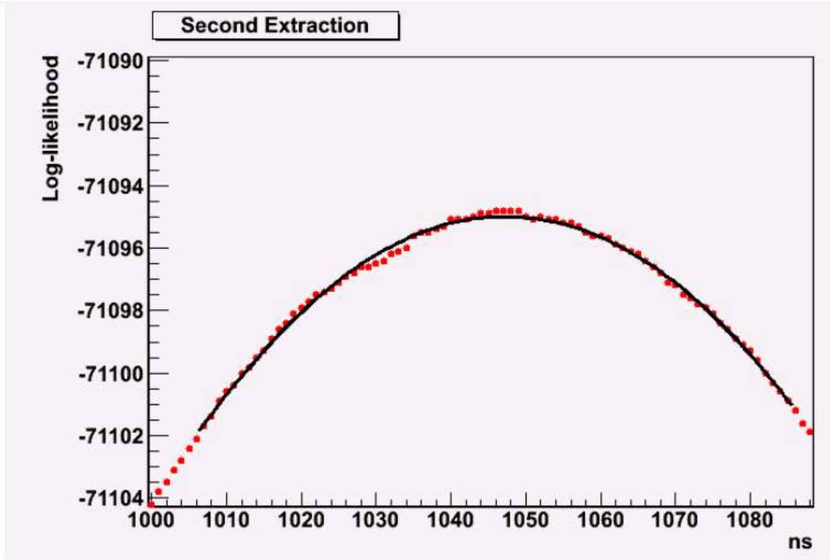



GPS
common view
CERN
LNGS
PolaRx +Cs clock
PolaRx +Cs clock
UTC
UTC
time shift by TOF_c
waveforms
data
0
5000
10000
p
π/K
ν
y
x
z
BCT
target
decay tunnel
OPERA
baseline → (TOF_c)
$\delta t = TOF_{c}\text{-}TOF_{\nu}$

$$L_k(\delta t_k) = \prod_j w_k(t_j + \delta t_k) \qquad k = 1,2 \text{ extractions}$$

Near the maximum the likelihood function can be approximated by a Gaussian whose variance is a measure of the statistical uncertainty on δt. The data used for the maximum likelihood calculation are unbinned and the dependence on δt is computed by making a scan in steps of 1 ns. A parabolic fit is performed on the log-likelihood function for the evaluation of the maximum and of the statistical uncertainty (Fig. 10). As seen in Fig. 11, the PDF representing the time-structure of the proton extraction is not flat but exhibits a series of peaks and valleys, reflecting the features and the inefficiencies of the proton extraction from the PS to the SPS via the Continuous Transfer mechanism [41]. Such structures may well change with time. The way the PDF are built automatically accounts for the beam conditions corresponding to the neutrino interactions detected by OPERA.





Hypothesis testing

In the previous section we have discussed how to estimate parameters of an underlying pdf model from sample data.

We now consider the closely related question:

How good is our pdf model in the first place?

Hypothesis testing

In the previous section we have discussed how to estimate parameters of an underlying pdf model from sample data.

We now consider the closely related question:

How good is our pdf model in the first place?

Simple example.

Null hypothesis: sampled data are drawn from a normal pdf, with mean $\mu_{\rm model}$ and variance σ^2.

We want to **test** this null hypothesis: are our data consistent with it?

Assume (for the moment) that σ^2 is known.

Example

Measured data: $\{x_i : i = 1, \ldots, 10\}$ $\sum_{i=1}^{10} x_i = 47.8$

Null hypothesis: $x \sim N(\mu, \sigma^2)$ with $\mu_{\text{model}} = 4$

Assume: $\sigma = 2$ $\sigma_\mu{}^2 = 0.4$

Under NH, sample mean $\boxed{\bar{x}_{\text{model}} \sim N(4, 2^2/10)}$

Observed sample mean $\bar{x}_{\text{obs}} = 4.78$

We transform to a **standard normal variable**

Under NH: $$Z = \left(\frac{\bar{x}_{\text{obs}} - \bar{x}_{\text{model}}}{\sigma_\mu} \right) \sim N(0,1)$$

From our measured data: $$Z_{\text{obs}} = \frac{4.78 - 4}{\sqrt{0.4}} = 1.233$$

If NH is true, how probable is it that we would obtain a value of Z_{obs} as large as this, or larger?

We call this probability the **p-value**

Question 8: Suppose that X is sampled from a normal distribution with mean $\mu = 5$ and variance $\sigma^2 = 9$.

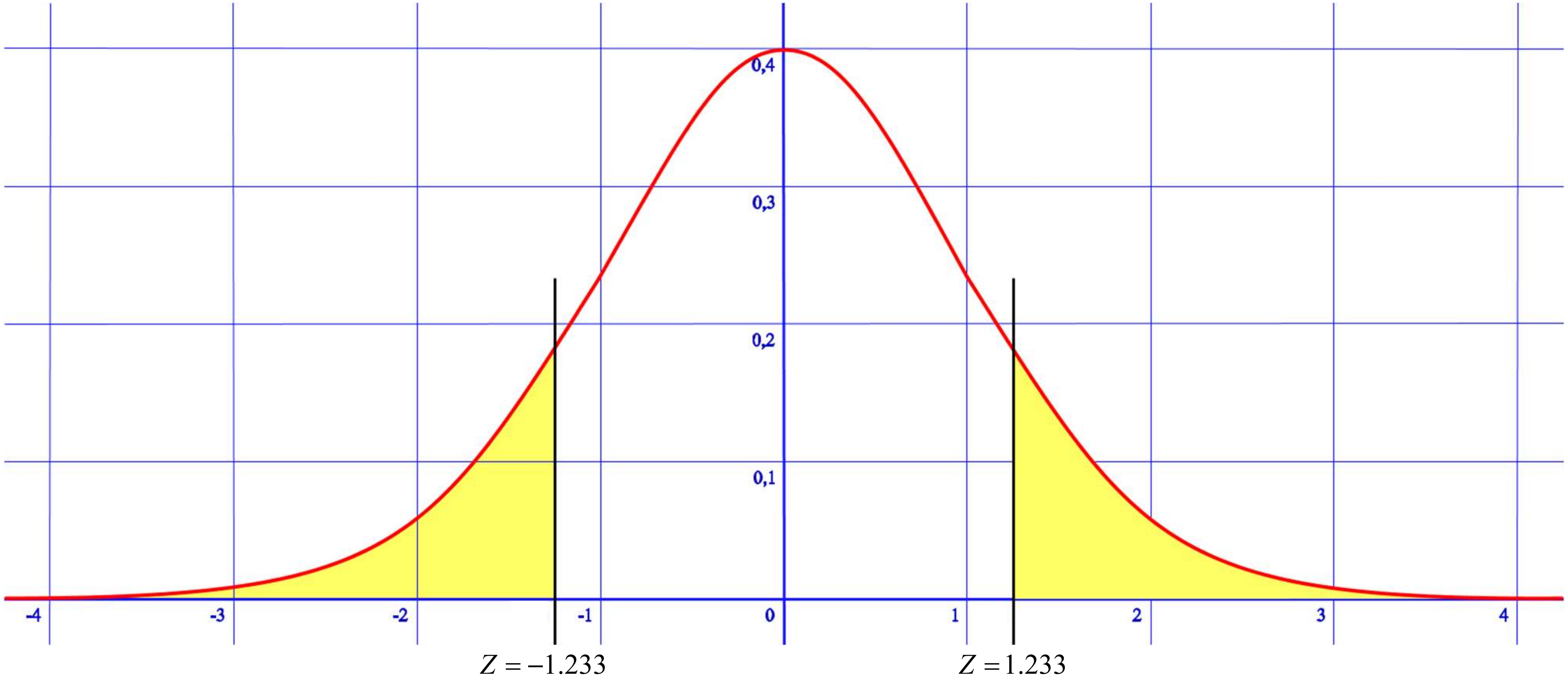
Which of the following is a standard normal variable?

A $Z = \dfrac{X-5}{9}$

B $Z = \dfrac{X-5}{3}$

C $Z = \dfrac{X-3}{5}$

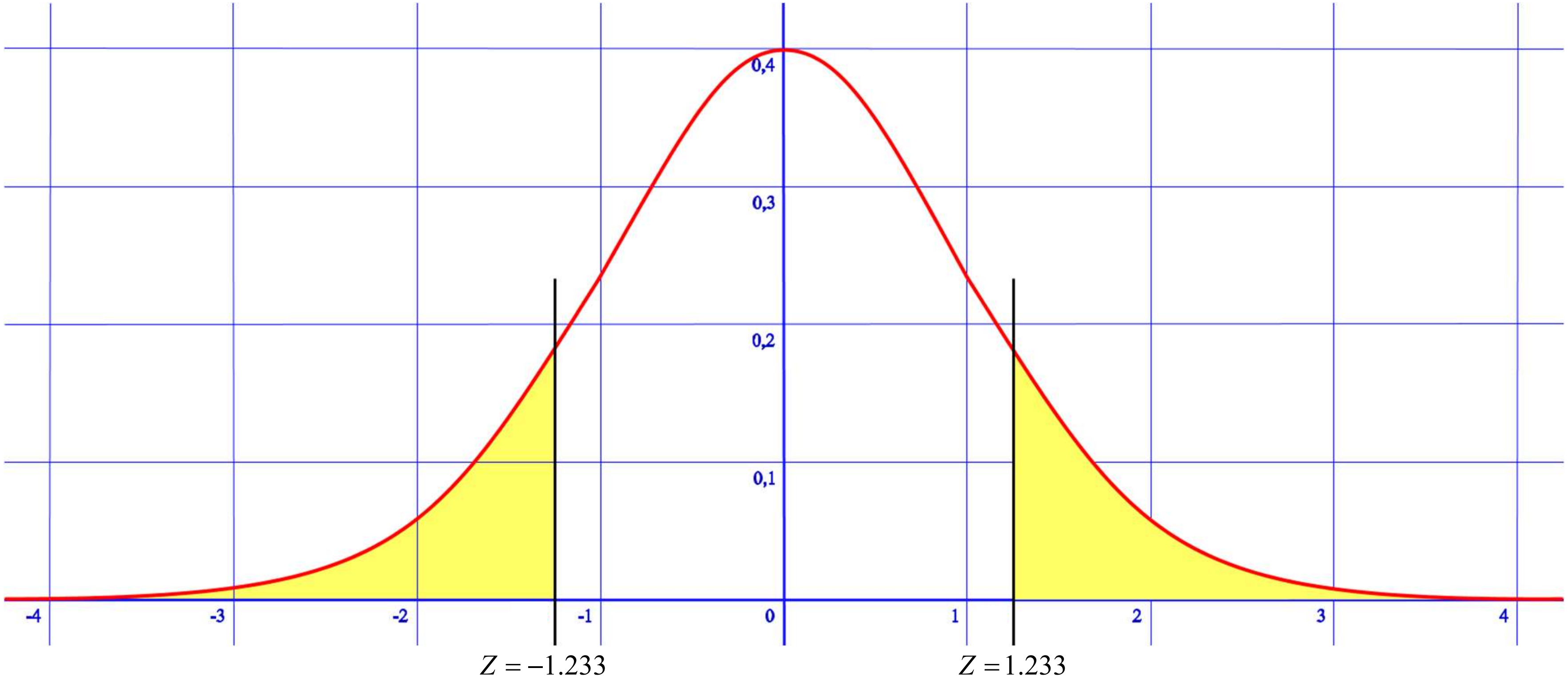
D $Z = \dfrac{X-3}{9}$



$$\text{p-value} \quad = \quad \mathrm{Prob}\left(|Z| \geq |Z_{\mathrm{obs}}|\right) \quad = \quad 1 - \int_{-Z_{\mathrm{obs}}}^{Z_{\mathrm{obs}}} \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2} z^2\right) dz$$

Simple programs to perform this probability integral (and many others) can be found in numerical recipes, or built into e.g. MATLAB or MAPLE.

Java applets also available online at http://statpages.org/pdfs.html (here).



$$\text{p-value} \quad = \quad \text{Prob}\left(|Z| \geq |Z_{\text{obs}}|\right) \quad = \quad 0.2176$$

The ***smaller*** the p-value, the less credible is the null hypothesis.

We can also carry out a *one-tailed* hypothesis test, if appropriate, and for statistics with other sampling distributions.

Question 9: A one-tailed hypothesis test is carried out. Under the NH the test statistic has a uniform distribution $U[0,1]$.

The observed value of the test statistic is 0.8.

The p-value is:

A 0.8

B 0.9

C 0.2

D 0.1

What if we *don't* assume that σ^2 is known?

We can estimate it from our observed data (provided $n \geq 2$)

We form the statistic $t_{\text{obs}} = \left(\dfrac{\bar{x}_{\text{obs}} - \bar{x}_{\text{model}}}{\hat{\sigma}_{\mu}} \right)$

where $\hat{\sigma}_{\mu}{}^2 = \dfrac{1}{n(n-1)} \sum_{i=1}^{n} \left(x_i - \bar{x}_{\text{obs}} \right)^2$

Accounts for the fact that we don't know μ, but must use $\bar{x}_{\text{obs}}$ when we estimate σ_{μ}

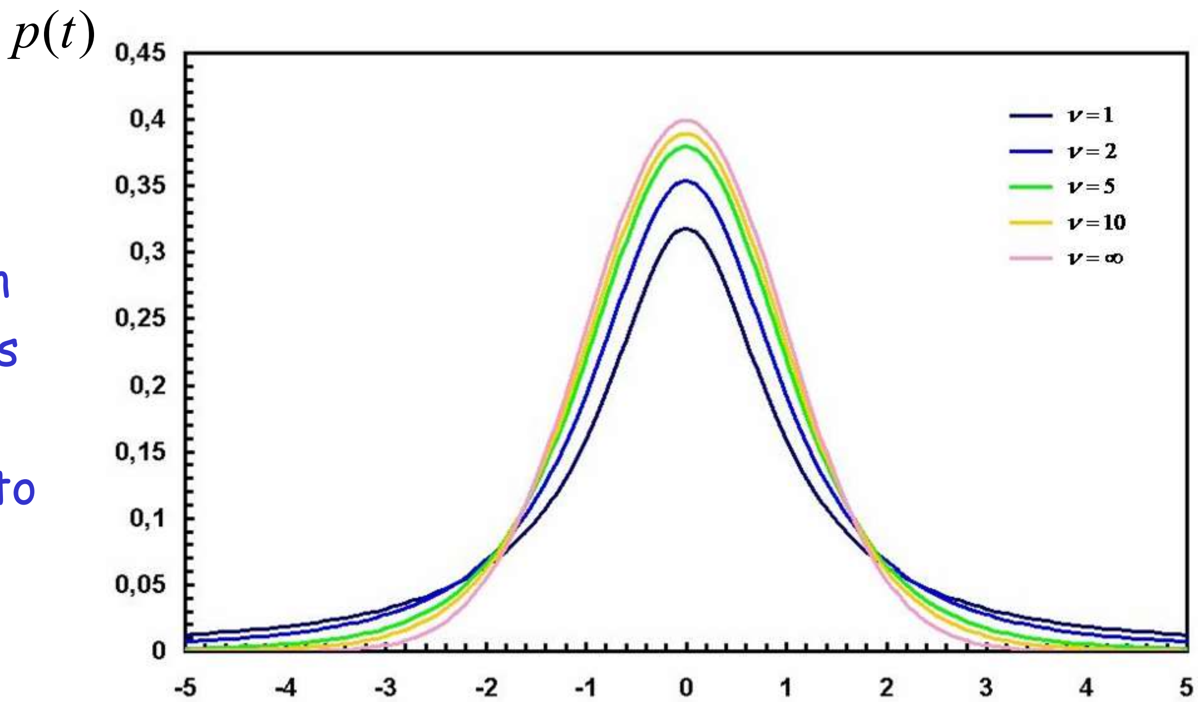
However, now t_{obs} no longer has a normal distribution.

In fact t_{obs} has a pdf known as the **Student's t distribution**

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

where $\nu = n-1$ is the no. degrees of freedom and $\Gamma(\nu) = \int_0^\infty x^{\nu-1}e^{-x}dx$

For small n the Student's t distribution has more extended tails than Z, but as $n \to \infty$ the distribution tends to $N(0,1)$



Question 10: The more extended tails of the students' t distribution mean that, under the null hypothesis

A larger values of the test statistic are more likely

B larger values of the test statistic are less likely

C smaller values of the test statistic are more likely

D smaller values of the test statistic are less likely

Hypothesis tests and decision theory

In a **simple hypothesis test**, we test our **null hypothesis** against a single **alternative hypothesis**.

We choose a **critical region**: set of values of the test statistic for which we choose to reject the **NH** and accept the **AH**.

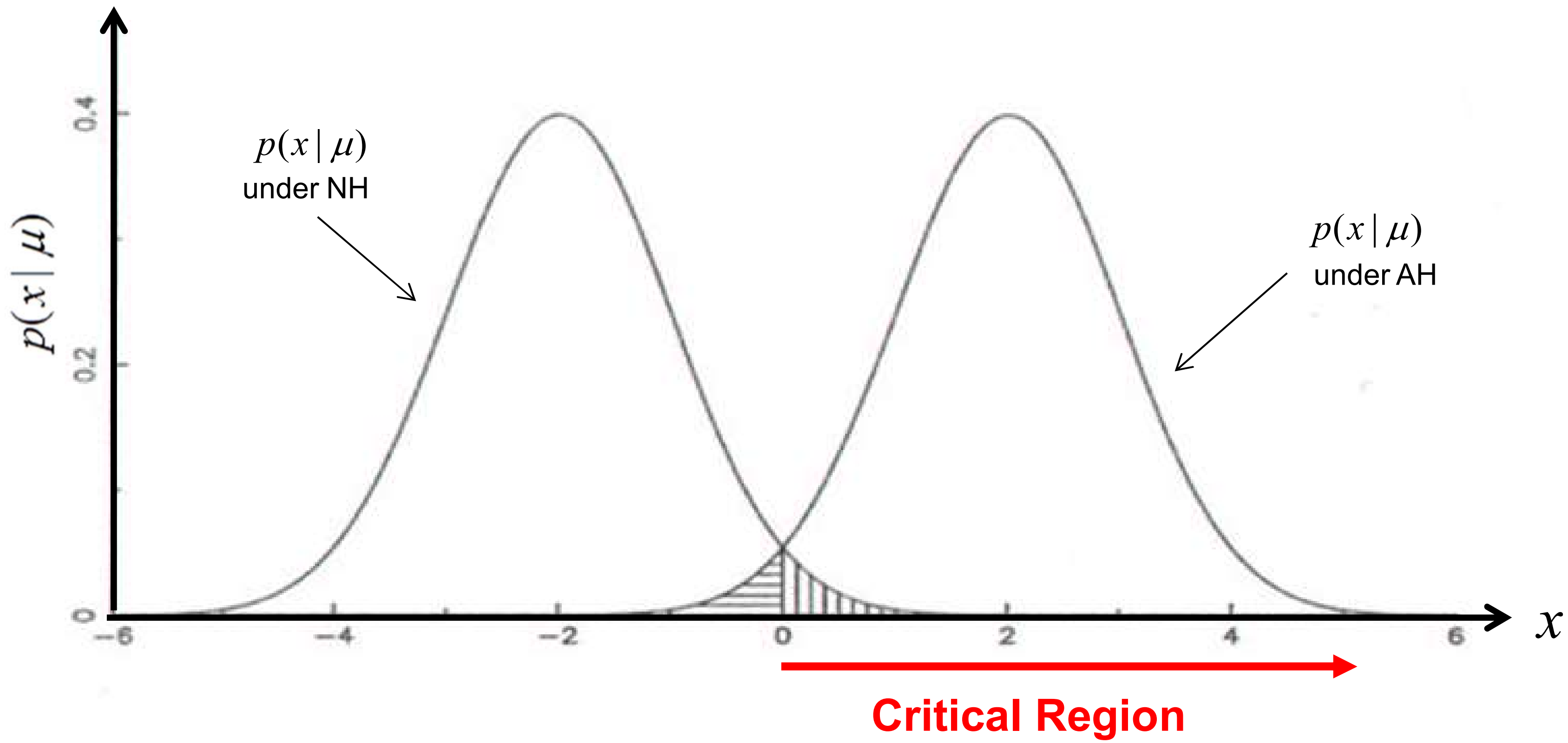
This means we need to consider the distribution of our test statistic under the NH and the AH.

Example

Measured data leads to test statistic x, estimator of μ :

NH: $\mu = -2; \quad x \sim N(-2,1)$ **AH**: $\mu = +2; \quad x \sim N(2,1)$

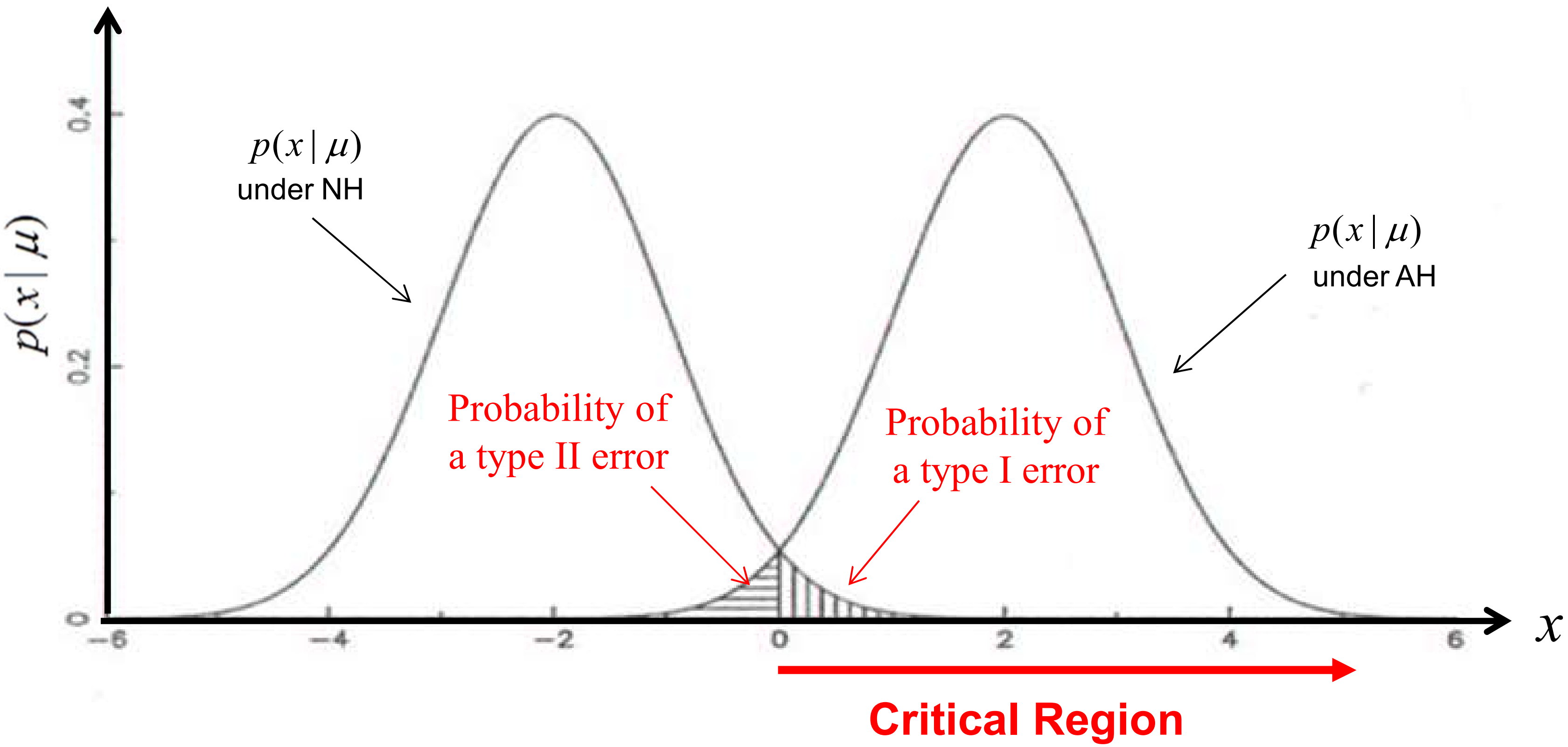
Suppose we choose critical region to be $x > 0$



There are two ways in which we can make an incorrect decision:

Type I error: we reject the NH when it is **TRUE**

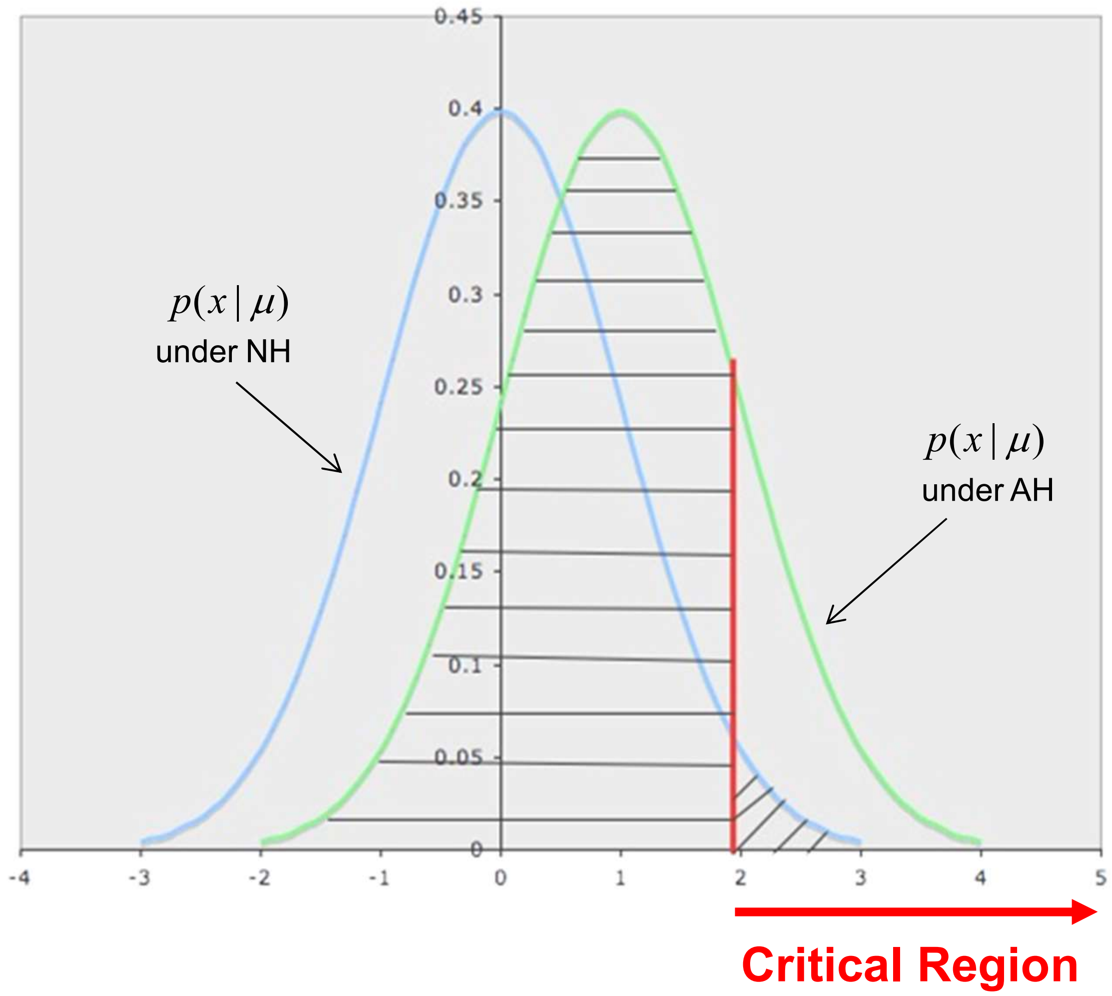
Type II error: we accept the NH when it is **FALSE**



Consequences of incorrect decision will be worse when the sampling distributions under NH and AH have a greater overlap.

This may influence our choice of critical region.

We want to reduce the probability of type I and type II errors – but we can't do *both* at the same time…



Lots of terminology:

Type I error: also known as **false alarm** or **false positive**

Type II error: also known as **false negative** or **"miss"**

Sensitivity = *probability (rate) of obtaining* ***true positive*** *("hit")*

Specificity = *probability (rate) of obtaining* ***true negative***

Sensitivity (or power) = 1 – prob(type II error)

Specificity = 1 – prob(type I error)

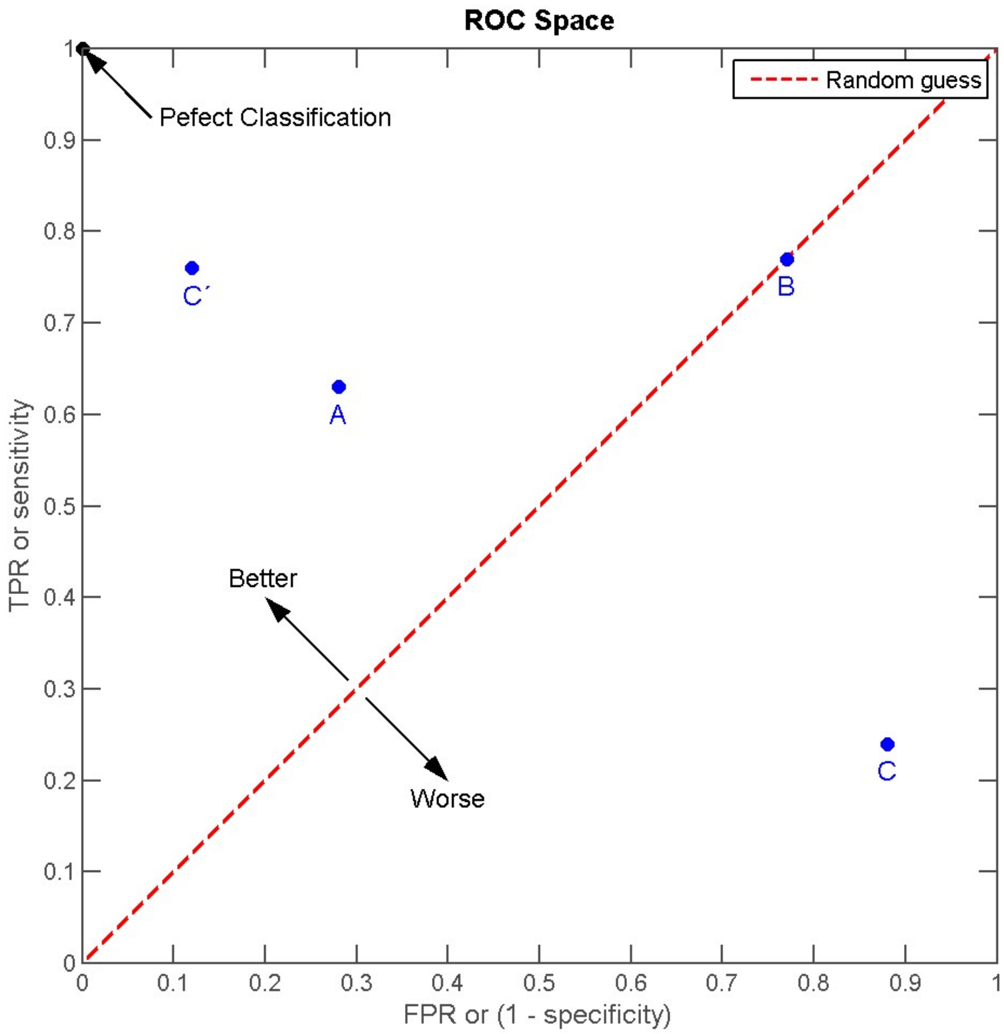
In many fields (particularly medical applications) a **Receiver Operating Chracteristic** (ROC) plot can be used to assess the performance of a simple hypothesis test.

x-axis = (1 – specificity)

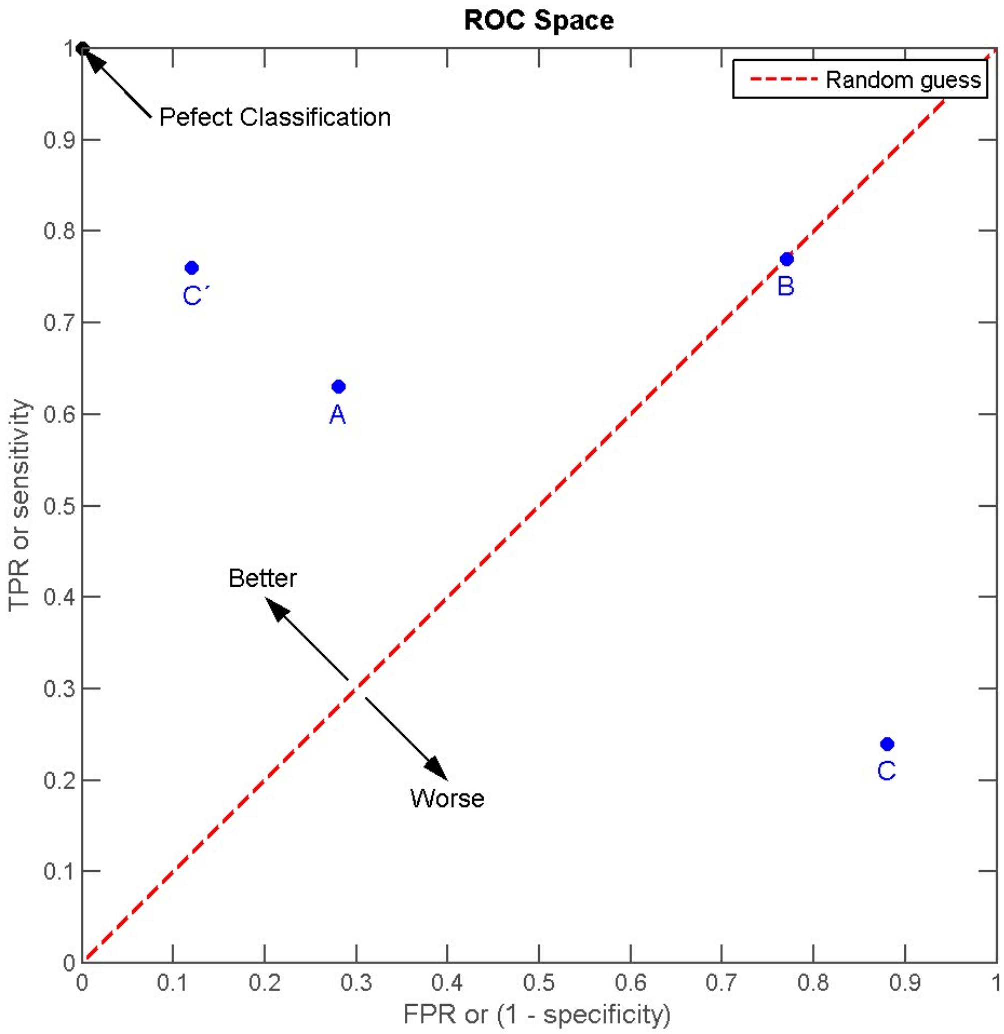
y-axis = sensitivity

Blue dots = results for different hypothesis tests.

Red diagonal = what we'd expect from a random guess alone.

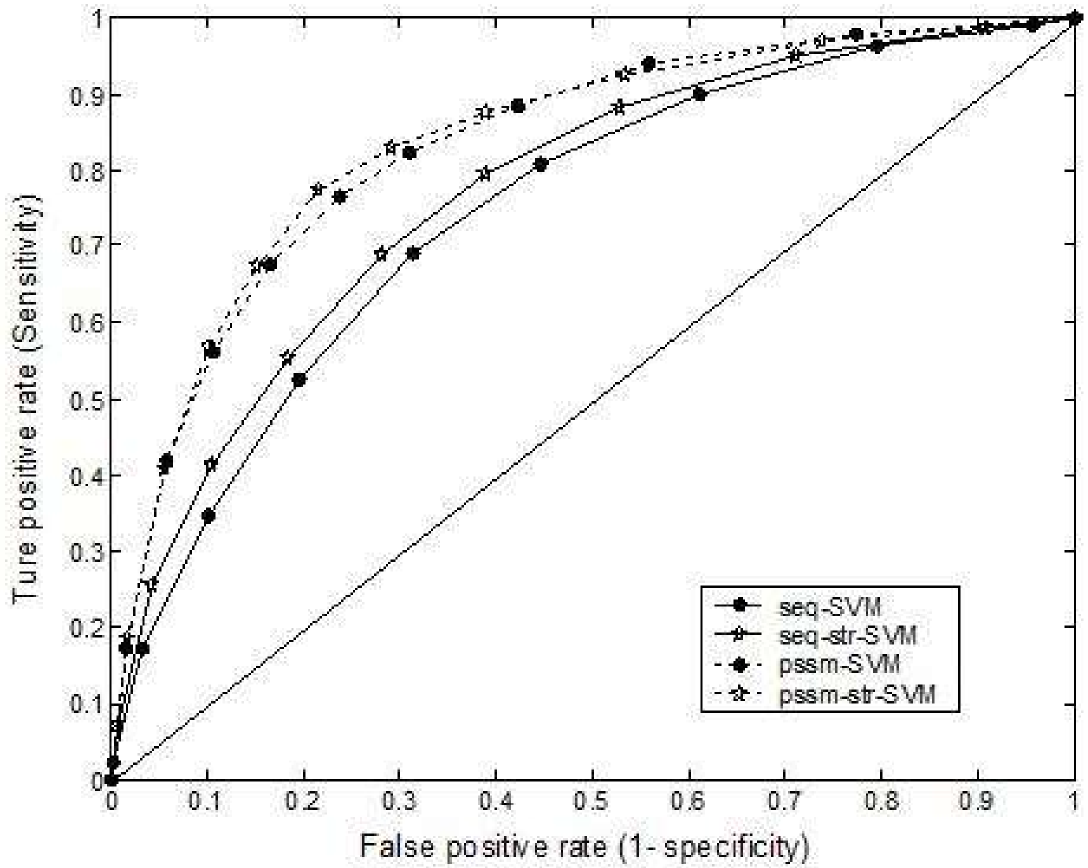


We can generate a ROC curve for a given test by varying the critical region. This can help to optimise our choice of CR – trading off type I and type II error, and getting us as close as possible to a perfect decision / classification.



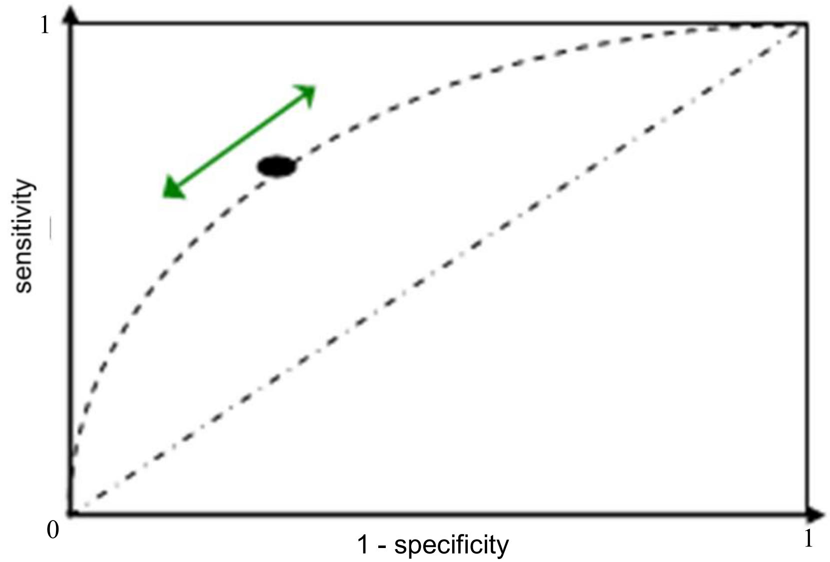
We can generate a ROC curve for a given test by varying the critical region. This can help to optimise our choice of CR – trading off type I and type II error, and getting us as close as possible to a perfect decision / classification.

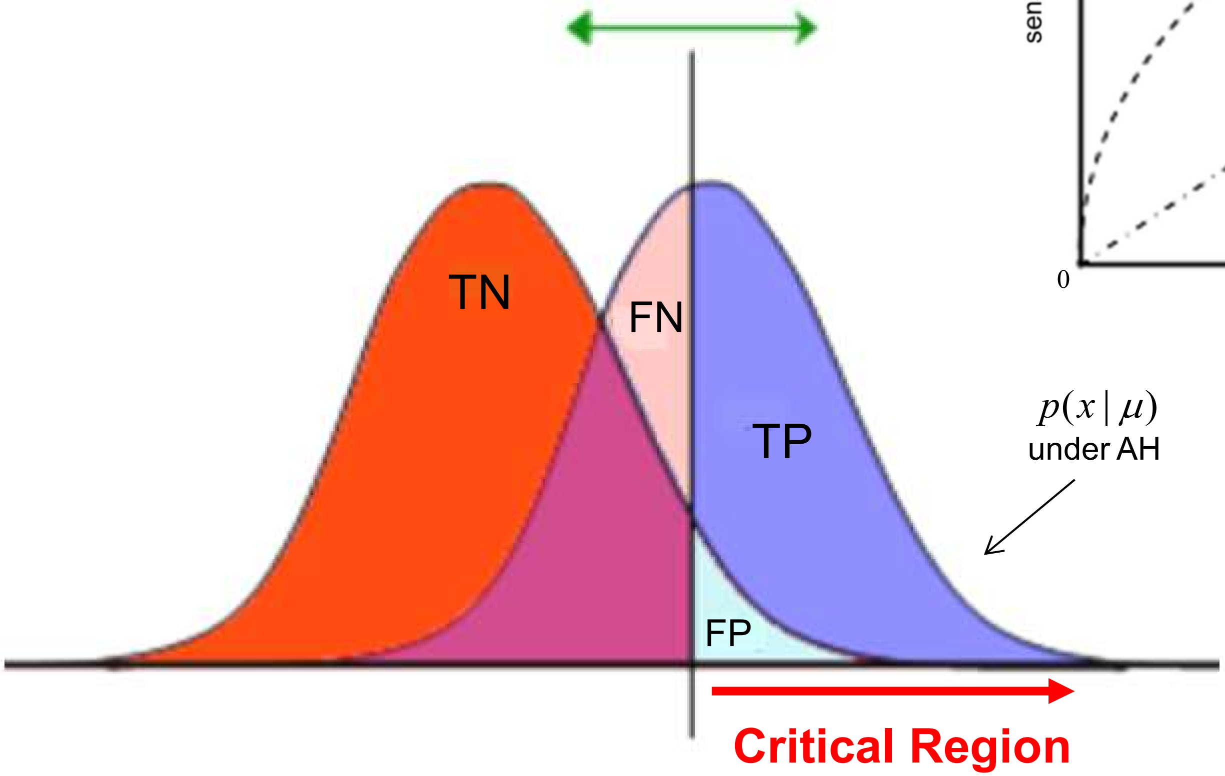
The area under the ROC curve can be used as a measure to compare different tests and choose the best.



ROC curves for predictors of DNA-binding sites.
SUNY Albany Center for Excellence in Cancer Genomics

We can generate a ROC curve for a given test by varying the critical region. This can help to optimise our choice of CR – trading off type I and type II error, and getting us as close as possible to a perfect decision / classification.





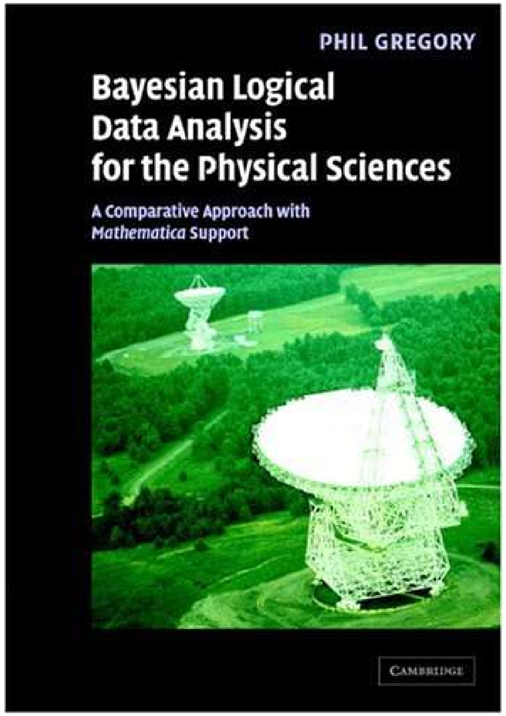
As we move the CR from left to right, we move along the ROC curve (dashed line)

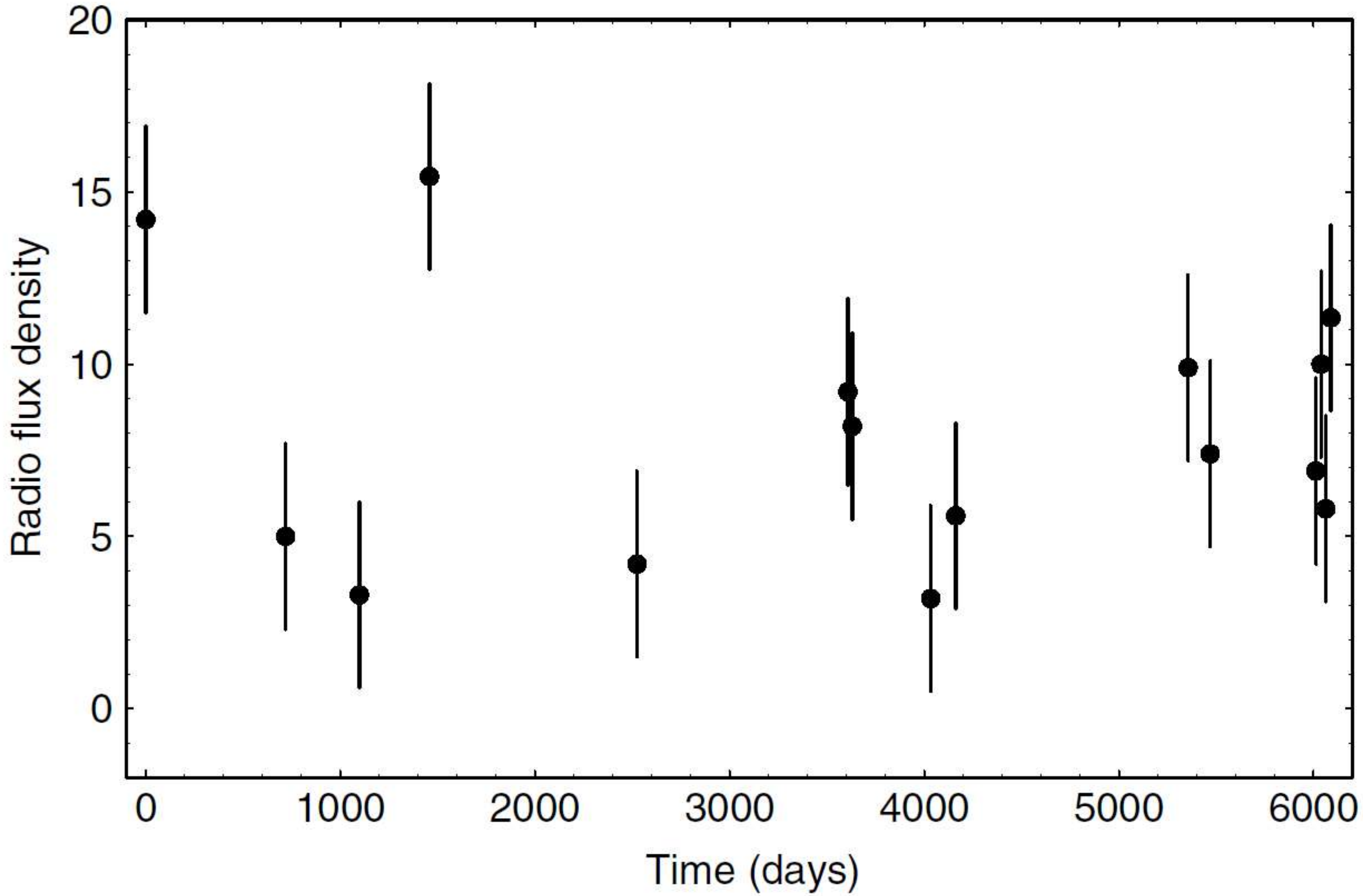
Hypothesis testing

More generally, we now illustrate the frequentist approach to the question of how good is the fit to our model, using the Chi-squared goodness of fit test.

We take an example from Gregory (Chapter 7)

(book focusses mainly on Bayesian probability, but is very good on frequentist approach too)





Model: radio emission from a galaxy is constant in time.

Assume residuals are iid, drawn from $N(0,\sigma)$

Goodness-of-fit Test: the basic ideas

1. Choose as our null hypothesis that the galaxy has an unknown but constant flux density. If we can demonstrate that this hypothesis is absurd at say the 95% confidence level, then this provides indirect evidence that the radio emission is variable. Previous experience with the measurement apparatus indicates that the measurement errors are independently normal with a $\sigma = 2.7$.
2. Select a suitable statistic that (a) can be computed from the measurements, and (b) has a predictable distribution. More precisely, (b) means that we can predict the distribution of values of the statistic that we would expect to obtain from an infinite number of repeats of the above set of radio measurements under identical conditions. We will refer to these as our hypothetical reference set. More specifically, we are predicting a probability distribution for this reference set.

   To refute the null hypothesis, we will need to show that scatter of the individual measurements about the mean is larger than would be expected from measurement errors alone.
3. Evaluate the χ^2 statistic from the measured data. Let's start with the expression for the χ^2 statistic for our data set:

$$\boxed{\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{\sigma^2}}$$

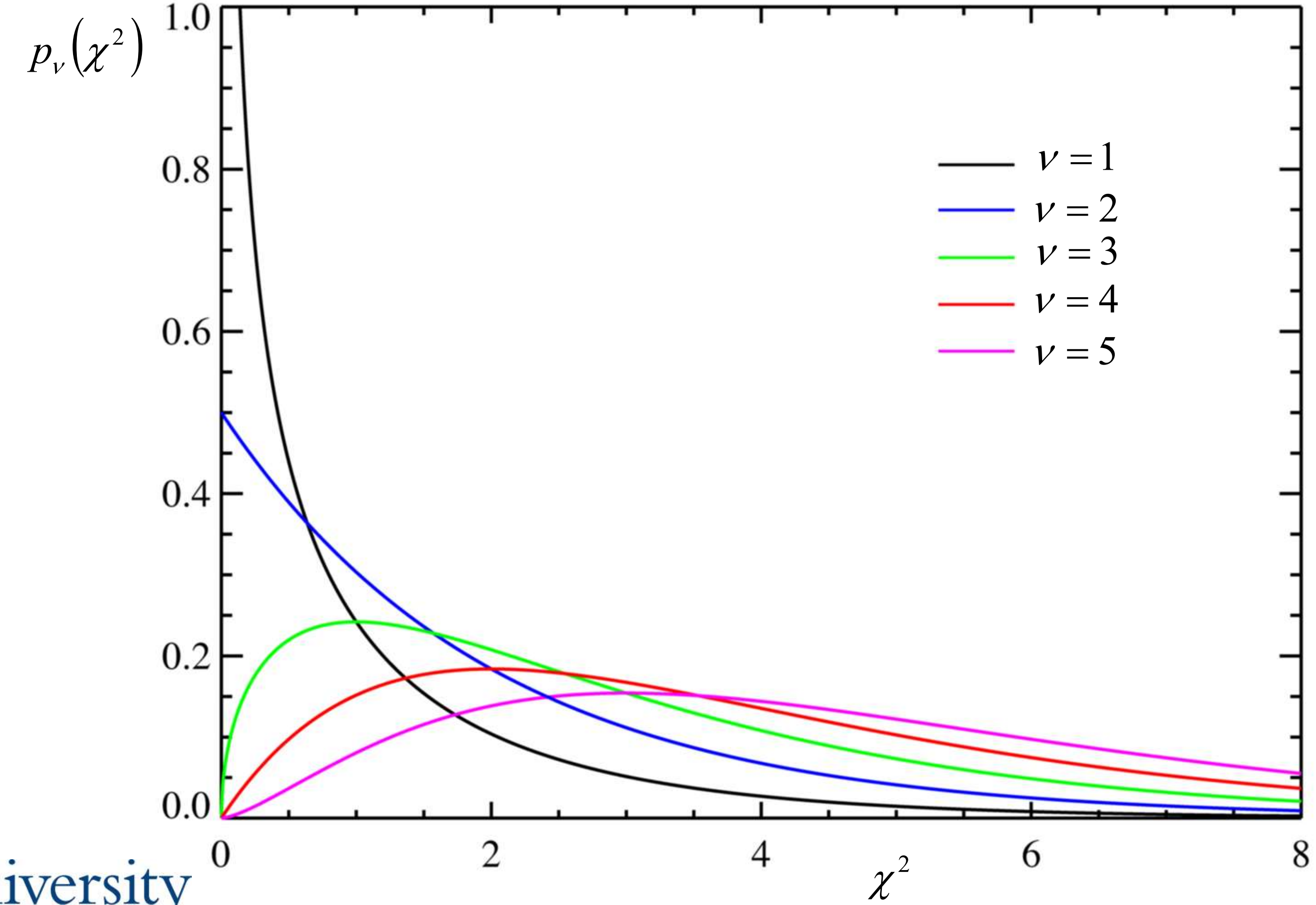
From Gregory, pg. 164

The χ^2 pdf

$$p_\nu\left(\chi^2\right) = p_0 \times \left(\chi^2\right)^{\frac{\nu}{2}-1} e^{-\chi^2/2}$$

Here ν is known as the number of **degrees of freedom** of the pdf.

The mean value of the pdf is ν and the variance is 2ν.

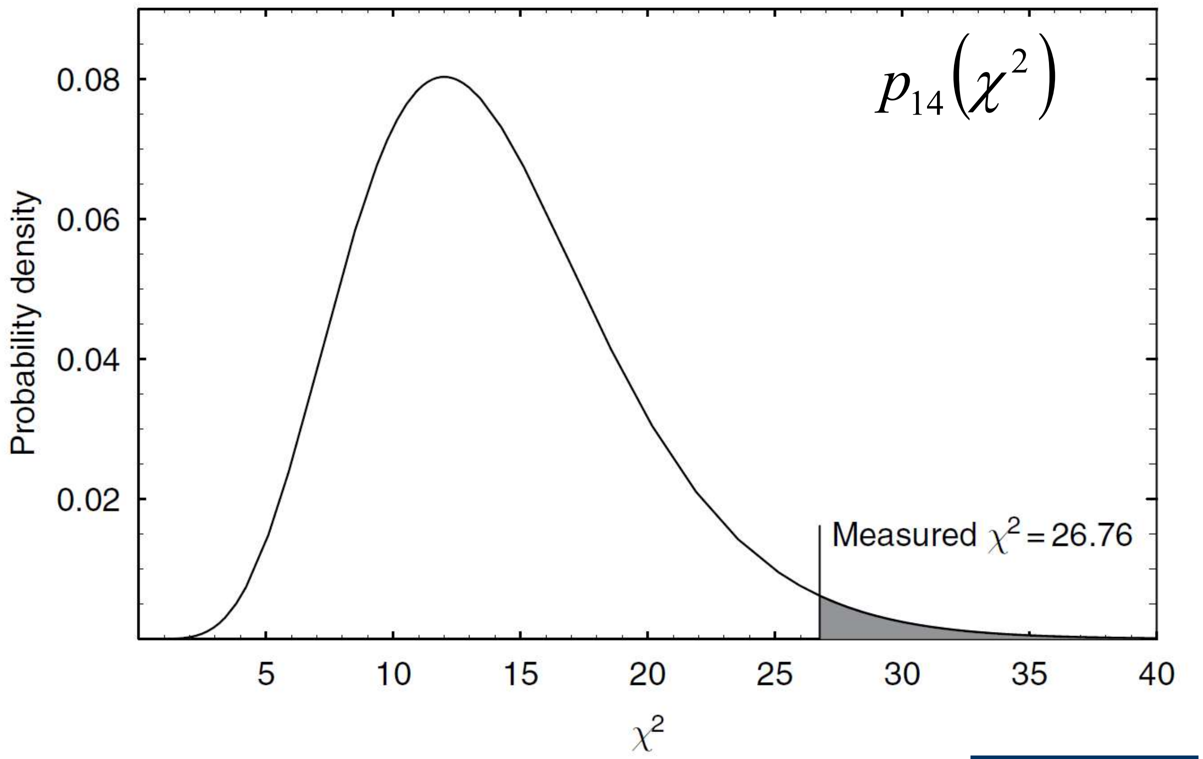


Day Number	Flux Density (mJy)
0.0	14.2
718.0	5.0
1097.0	3.3
1457.1	15.5
2524.1	4.2
3607.7	9.2
3630.1	8.2
4033.1	3.2
4161.3	5.6
5355.9	9.9
5469.1	7.4
6012.4	6.9
6038.3	10.0
6063.2	5.8
6089.3	11.4

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(x_i - 7.98)^2}{2.7^2} = 26.76$$

$n = 15$ data points, but $\nu = 14$ degrees of freedom, because χ^2 statistic involves the *sample mean* and not the true mean.

We subtract one d.o.f. to account for this.



Question 11: Given that the mean and variance of a chi-squared distribution with n degrees of freedom are n and $2n$ respectively, in the Gregory example (with 14 degrees of freedom) estimate the number of sigma by which the value $\chi^2_{\text{obs}} = 26.76$ exceeds the expected value.

A between 0 and 1 sigma

B between 1 and 2 sigma

C between 2 and 3 sigma

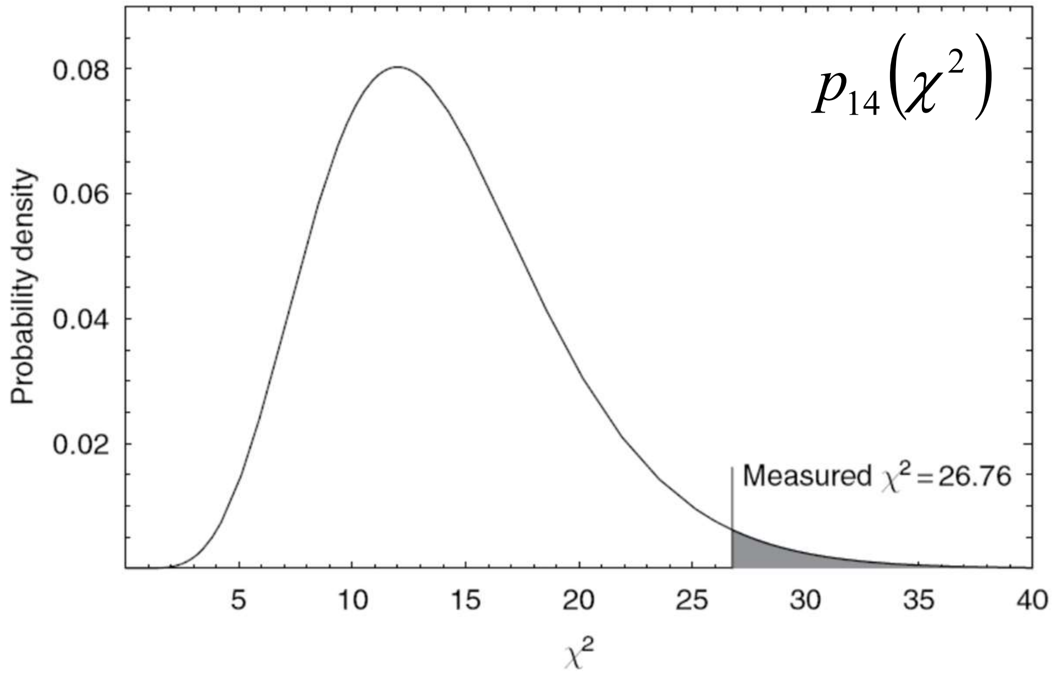
D between 3 and 4 sigma

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(x_i - 7.98)^2}{2.7^2} = 26.76.$$

Day Number	Flux Density (mJy)
0.0	14.2
718.0	5.0
1097.0	3.3
1457.1	15.5
2524.1	4.2
3607.7	9.2
3630.1	8.2
4033.1	3.2
4161.3	5.6
5355.9	9.9
5469.1	7.4
6012.4	6.9
6038.3	10.0
6063.2	5.8
6089.3	11.4

$n = 15$ data points, but $\nu = 14$ degrees of freedom, because χ^2 statistic involves the *sample mean* and not the true mean.

We subtract one d.o.f. to account for this.



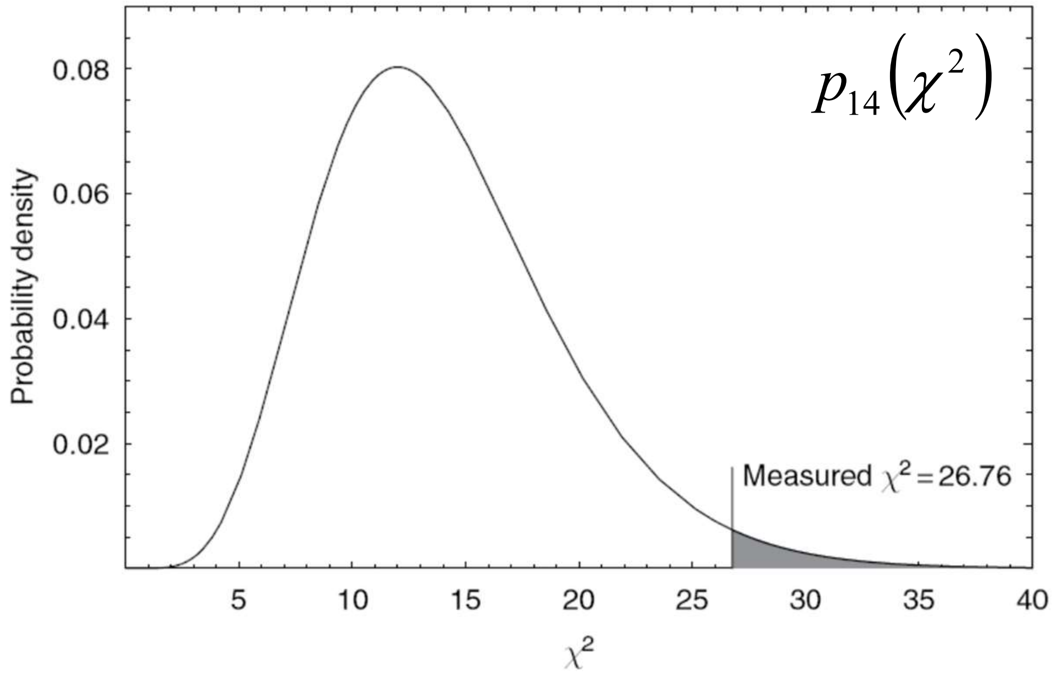
If the null hypothesis is true, how probable is it that we would measure as large, or larger, a value of χ^2 ?

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(x_i - 7.98)^2}{2.7^2} = 26.76.$$

Day Number	Flux Density (mJy)
0.0	14.2
718.0	5.0
1097.0	3.3
1457.1	15.5
2524.1	4.2
3607.7	9.2
3630.1	8.2
4033.1	3.2
4161.3	5.6
5355.9	9.9
5469.1	7.4
6012.4	6.9
6038.3	10.0
6063.2	5.8
6089.3	11.4

$n = 15$ data points, but $\nu = 14$ degrees of freedom, because χ^2 statistic involves the *sample mean* and not the true mean.

We subtract one d.o.f. to account for this.



If the null hypothesis is true, how probable is it that we would measure as large, or larger, a value of χ^2 ?

If the null hypothesis were true, how probable is it that we would measure as large, or larger, a value of χ^2 ?

Recall that we refer to this important quantity as the **p-value**

$$\text{p-value} \;=\; 1 - P\left(\chi^2_{\text{obs}}\right) \;=\; 1 - \int_0^{\chi^2_{\text{obs}}} p_0\, x^{\frac{\nu}{2}-1} \exp\left(-\frac{x}{2}\right) dx \;=\; 0.02$$

If the null hypothesis were true, how probable is it that we would measure as large, or larger, a value of χ^2?

Recall that we refer to this important quantity as the **p-value**

$$\text{p-value} \;=\; 1 - P\left(\chi^2_{\text{obs}}\right) \;=\; 1 - \int_0^{\chi^2_{\text{obs}}} p_0 \, x^{\frac{\nu}{2}-1} \exp\left(-\frac{x}{2}\right) dx \;=\; 0.02$$

What precisely does the p-value mean?

> "If the galaxy flux density really *is* constant, and we repeatedly obtained sets of 15 measurements under the same conditions, then only 2% of the χ^2 values derived from these sets would be expected to be greater than our one actual measured value of 26.76"

From Gregory, pg. 165

If we obtain a very small p-value (e.g. a few percent?) we can interpret this as providing little support for the null hypothesis, which we may then **choose to reject**.

(Ultimately this choice is subjective, but χ^2 may provide objective ammunition for doing so)

If the null hypothesis were true, how probable is it that we would measure as large, or larger, a value of χ^2 ?

Recall that we refer to this important quantity as the **p-value**

$$\boxed{\text{p-value} \;=\; 1 - P\left(\chi^2_{\text{obs}}\right) \;=\; 1 - \int_0^{\chi^2_{\text{obs}}} p_0 \, x^{\frac{\nu}{2}-1} \exp\left(-\frac{x}{2}\right) dx \;=\; 0.02}$$

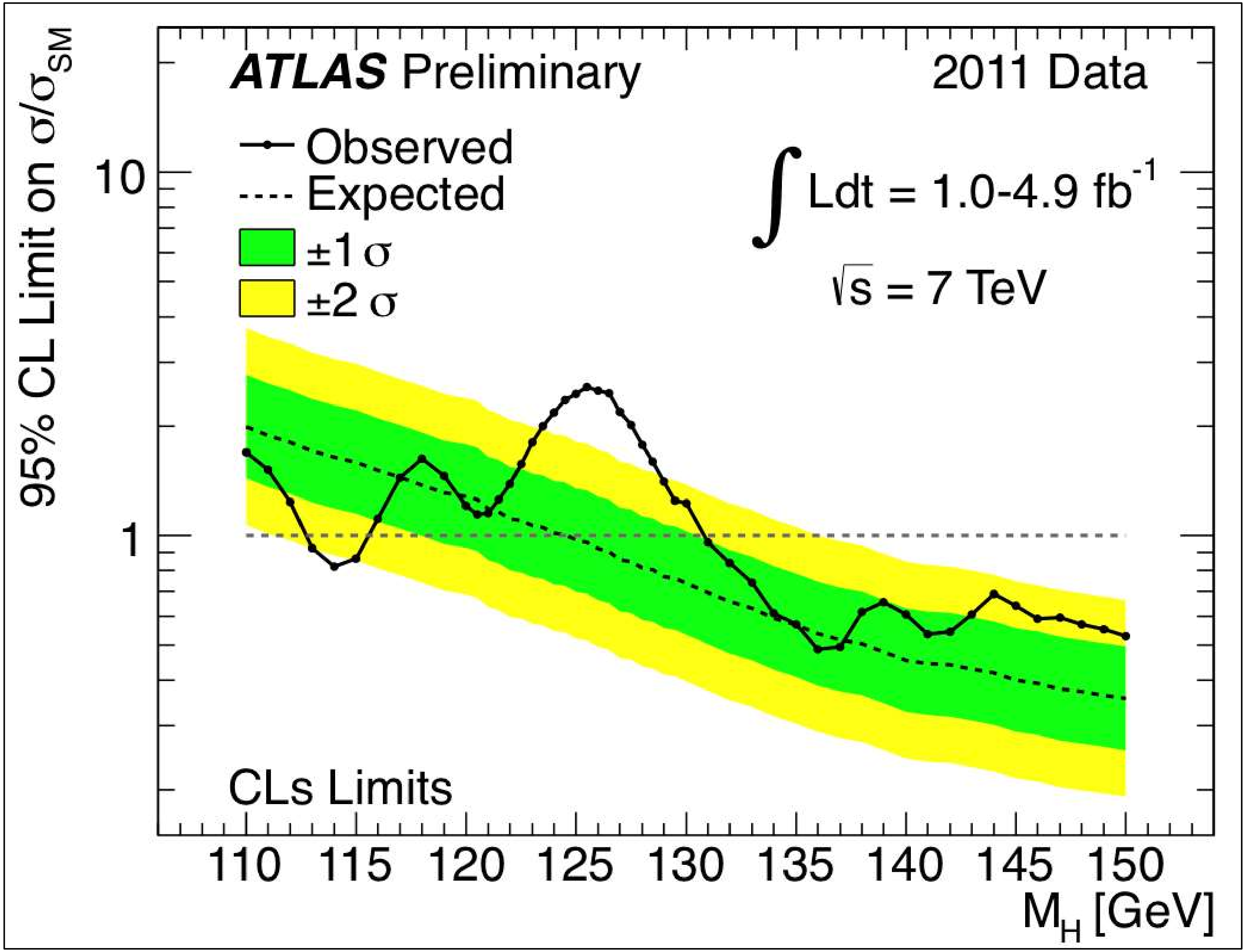
What precisely does the p-value mean?

"If the galaxy flux density really *is* constant, and we repeatedly obtained sets of 15 measurements under the same conditions, then only 2% of the χ^2 values derived from these sets would be expected to be greater than our one actual measured value of 26.76"

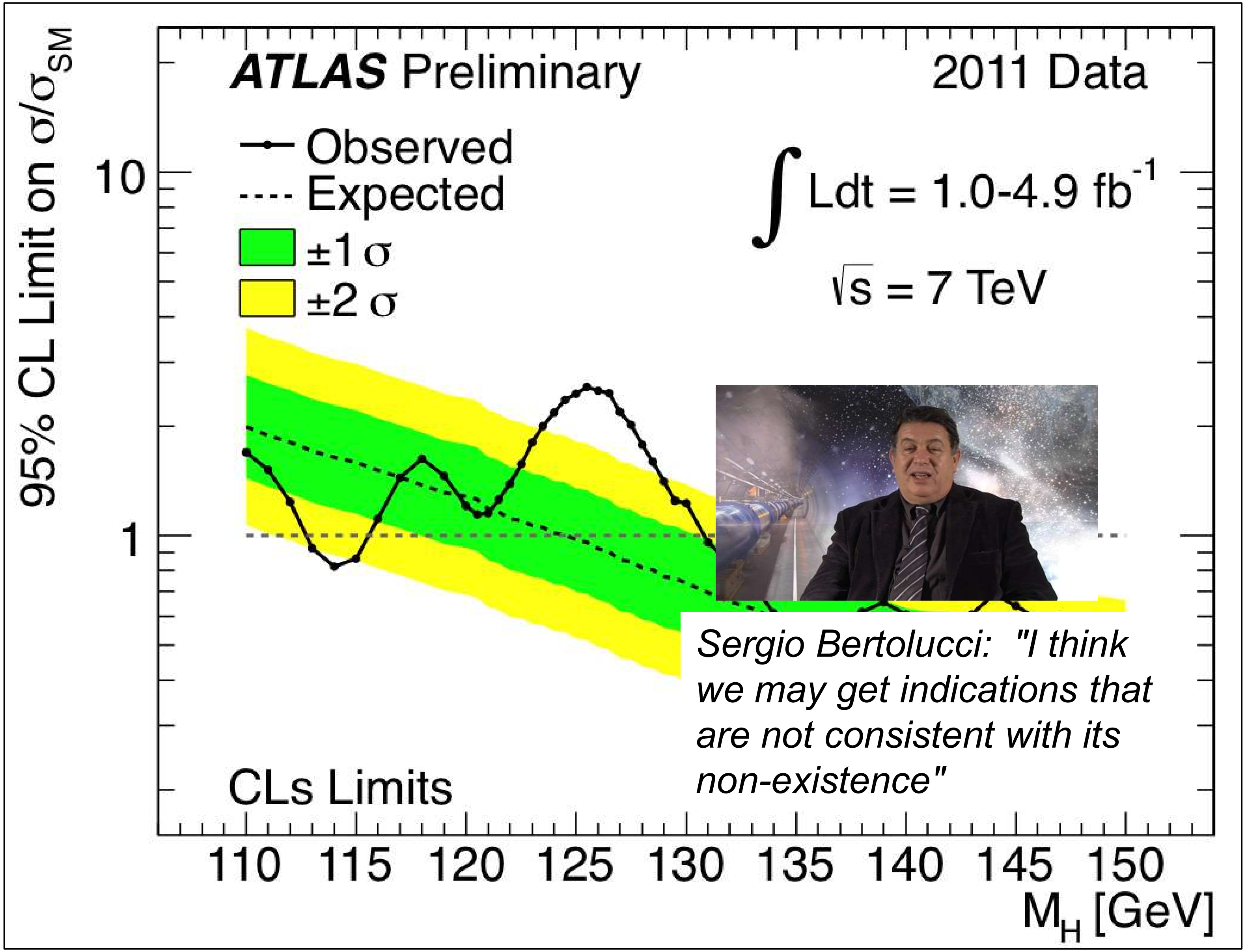
From Gregory, pg. 165

"At this point you may be asking yourself why we should care about a probability involving results never actually obtained"

From Gregory, pg. 166



ATLAS Preliminary
2011 Data
Observed
Expected
±1 σ
±2 σ
∫Ldt = 1.0-4.9 fb⁻¹
√s = 7 TeV
CLs Limits
95% CL Limit on σ/σ_SM
10
1
110
115
120
125
130
135
140
145
150
M_H [GeV]



Sergio Bertolucci: "I think we may get indications that are not consistent with its non-existence"

Nevertheless, p-value based frequentist hypothesis testing remains very common in the literature:

Type of problem	test	References
Line and curve goodness-of-fit	χ^2 test	NR: 15.1-15.6
Difference of means	Student's t	NR: 14.2
Ratio of variances	F test	NR: 14.2
Sample CDF	K-S test Rank sum tests	NR: 14.3, 14.6
Correlated variables?	Sample correlation coefficient	NR: 14.5, 14.6
Discrete RVs	χ^2 test / contingency table	NR: 14.4

Nevertheless, p-value based frequentist hypothesis testing remains very common in the literature:

Type of problem	test	References
Line and curve goodness-of-fit	χ^2 test	NR: 15.1-15.6
Difference of means	Student's t	NR: 14.2
Ratio of variances	F test	NR: 14.2
Sample CDF	K-S test Rank sum tests	NR: 14.3, 14.6
Correlated variables?	Sample correlation coefficient	NR: 14.5, 14.6
Discrete RVs	χ^2 test / contingency table	NR: 14.4

See also supplementary notes on my.SUPA and Moodle

In the Bayesian approach, we can test our model, in the light of our data (e.g. rolling a die) and see how our knowledge of its parameters evolves, for any sample size, considering only the data that we *did* actually observe

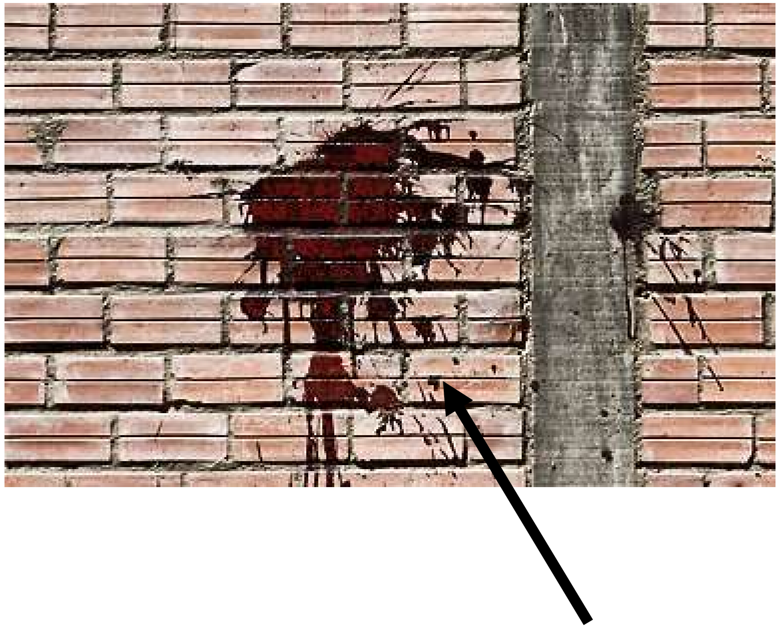
$$\underbrace{p(\text{model}\,|\,\text{data}, I)}_{\text{Posterior}} \;\propto\; \underbrace{p(\text{data}\,|\,\text{model}, I)}_{\text{Likelihood}} \times \underbrace{p(\text{model}\,|\,I)}_{\text{Prior}}$$

What we know now

Influence of our observations

What we knew before

What do we choose as our prior?



Blood on the walls

Good question!

Source of much argument between Bayesians and frequentists

What do we choose as our prior?

Blood on the walls

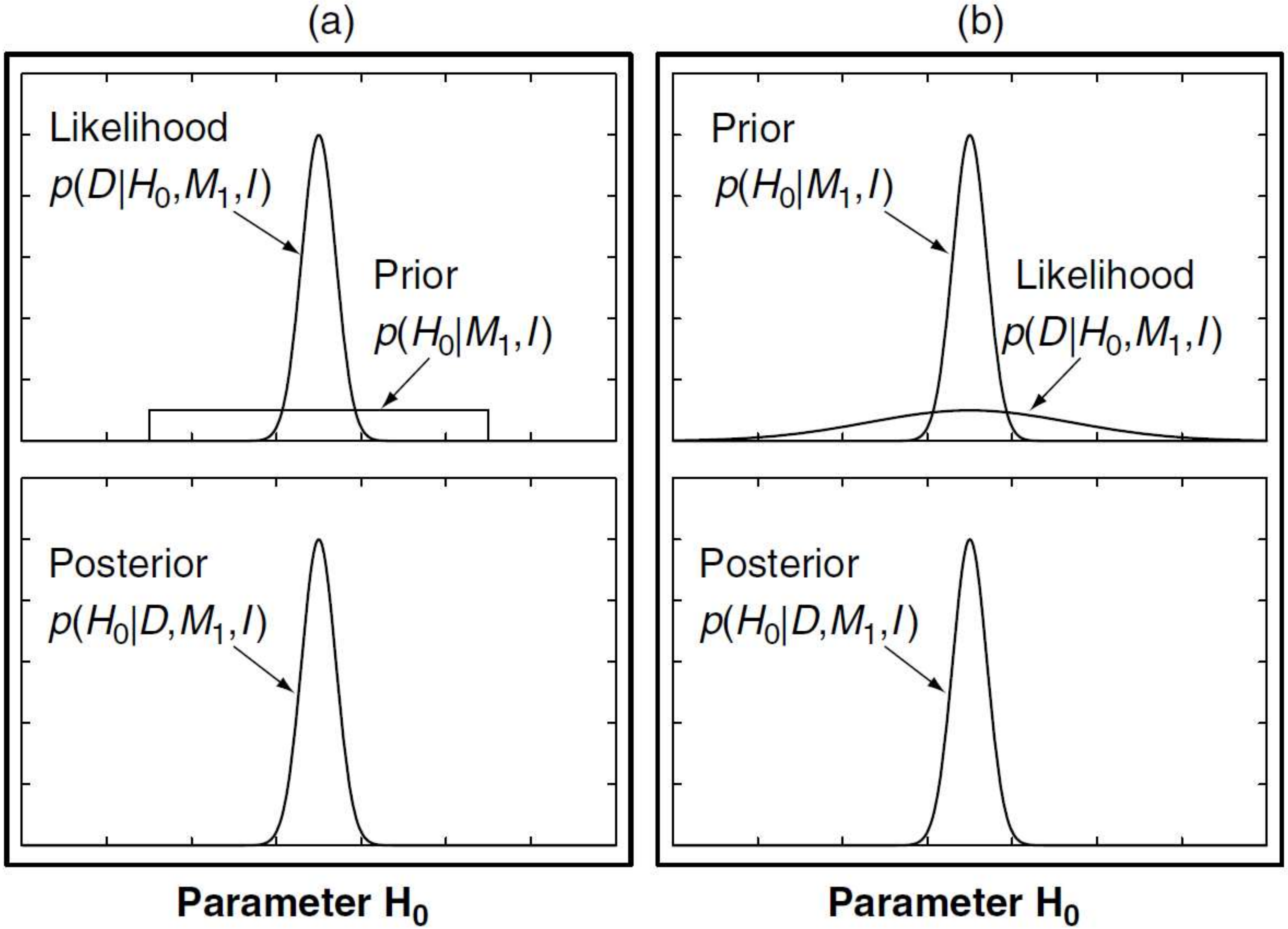
Good question!

Source of much argument between Bayesians and frequentists

If our data are good enough, it shouldn't matter

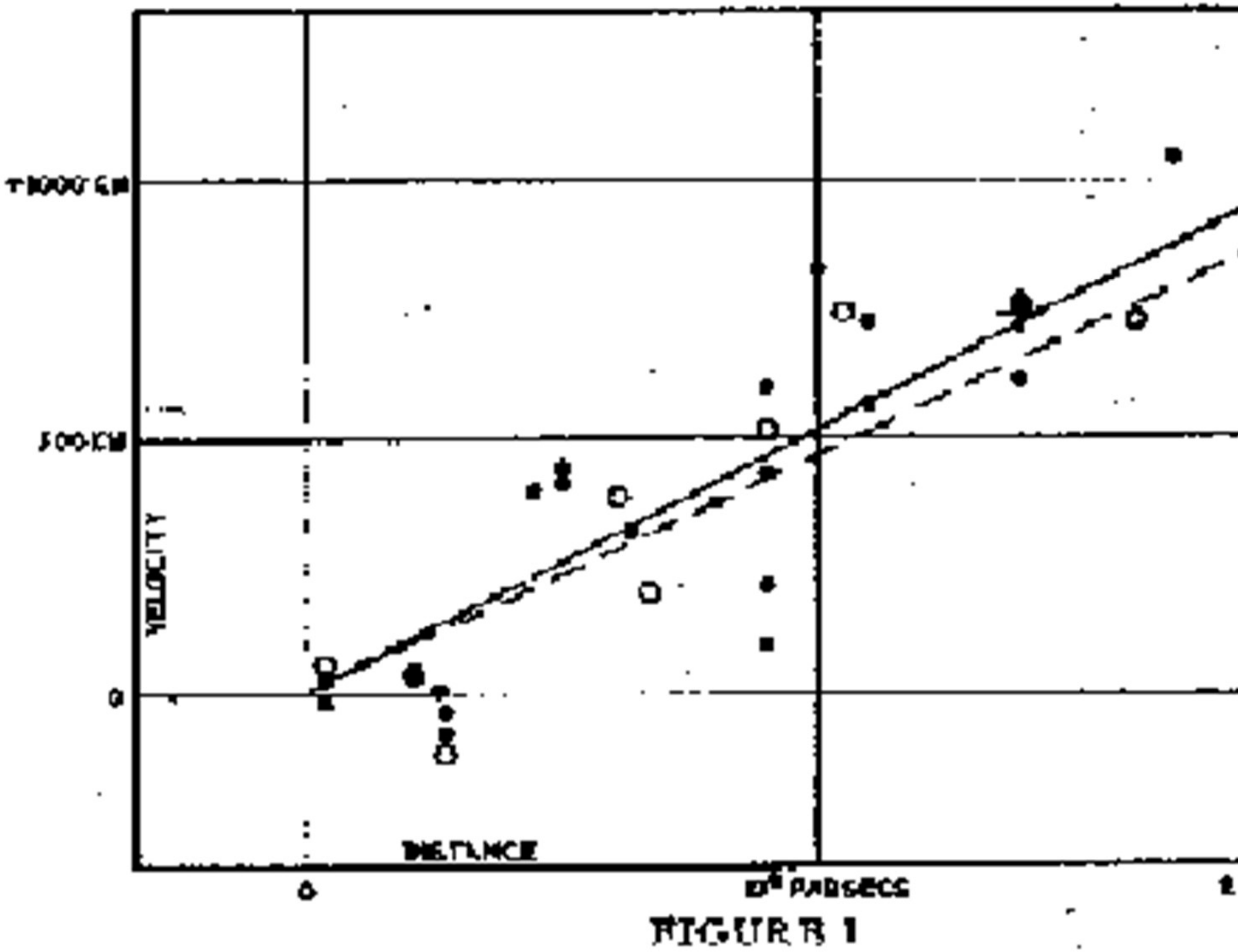
Posterior Likelihood Prior

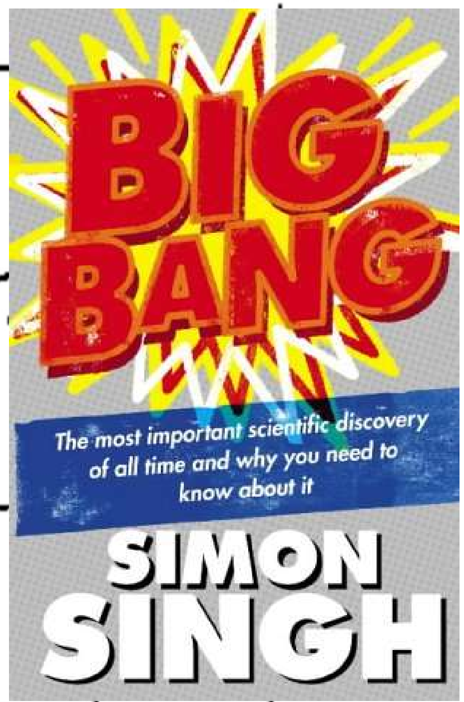
$$\boxed{p(\text{model} \mid \text{data}, I) \propto \underbrace{p(\text{data} \mid \text{model}, I)}_{\textit{Dominates}} \times p(\text{model} \mid I)}$$



From Gregory, pg 8.

Hubble's Law: 1929





Hubble parameter = expansion rate of the Universe
= slope of Hubble's law