PART 2: FREQUENTIST STATISTICAL BUILDING BLOCKS

Matthew Pitkin

GraWIToN School

25 March 2015

University of Glasgow

OVERVIEW

In this part of the course we will discuss

- parameter estimation
- maximum likelihood
- hypothesis testing
- goodness-of-fit

from a Frequentist perspective.

Suppose we observe n different realisations of a variable x (e.g. the height of everyone in the class), which has a pdf $p(x|I)$ (e.g. a normal distribution). This set $\{x_1, \ldots, x_n\}$ is called a **random sample from the population with pdf** $p(x|I)$. The joint pdf of these samples, $g(x_1, \ldots, x_n)$, is known as the **sampling distribution**. If all the elements, x_i, are **independently and identically distributed** (iid) then

$$g(x_1, \ldots, x_n|I) = p(x_1|I)p(x_2|I)\ldots p(x_n|I)$$

The sampling distribution is more commonly encountered as the **likelihood function**, which we will discuss more later, and a *random sample* will be considered to be some observed data set.

We may wish to study a population which has (or is assumed to have) a pdf $p(x|\theta, I)$, where | indicates that the pdf is dependent some (possibly unknown) parameter θ[1].

If we observe a random sample from the population $\{x_1, \ldots, x_n\}$ we might want to try and use these to estimate θ. How can we do that?

[1] In frequentist terms the pdf $p(x|\theta, I)$ would really be expressed as $p(x; \theta)$, where x is the RV dependent on the particular true value of θ, which is therefore not a RV. In the Bayesian view both x and θ have associated pdfs, so our knowledge of θ is just given by its pdf and any particular value range just has an associated plausibility.

A **statistic** is a function of observable RVs that *does not* depend on any unknown parameters. For a random sample drawn from $p(x|\theta, I)$ any function of $\{x_1, \ldots, x_n\}$ that *does not* depend on θ is an example of a statistic.

E.g., if a value x_1 is drawn from a normal distribution, $p(x|\mu, \sigma, I)$, where μ and σ are not known *a priori* then $x_1 - \mu$ is **not** a statistic.

However, in frequentist parameter estimation the idea is to use *statistics* to estimate the unknown parameters of a pdf[2].

[2] In the Bayesian view you would just calculate the pdf of the unknown parameters.

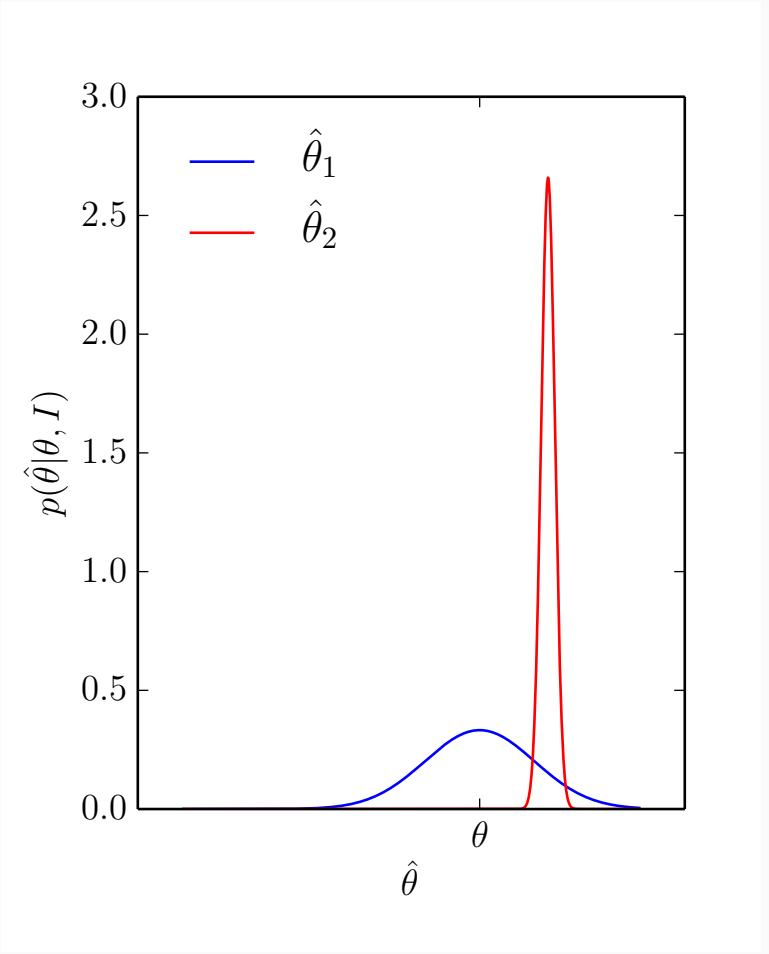
An **estimator** (e.g. $\hat{\theta}$) is a statistic used to estimate the value of a parameter (e.g. θ). $\hat{\theta}$ is not a funciton of θ, but is itself a RV as it is just a function of RVs $\{x_1, \ldots, x_n\}$.

Since x depends on the *true* value of θ then so does pdf of $\hat{\theta}$ and also $p(\hat{\theta}|\theta)$. The distribution $p(\hat{\theta}|\theta)$ can be determined by repeated trials (observations/experiments) giving new **random samples**, and the properties of $p(\hat{\theta}|\theta)$ can be used to determine if $\hat{\theta}$ is a *"good"* estimator of the *true* θ.

The pdfs of two estimators of θ ($\hat{\theta}_1$ and $\hat{\theta}_2$) are shown.

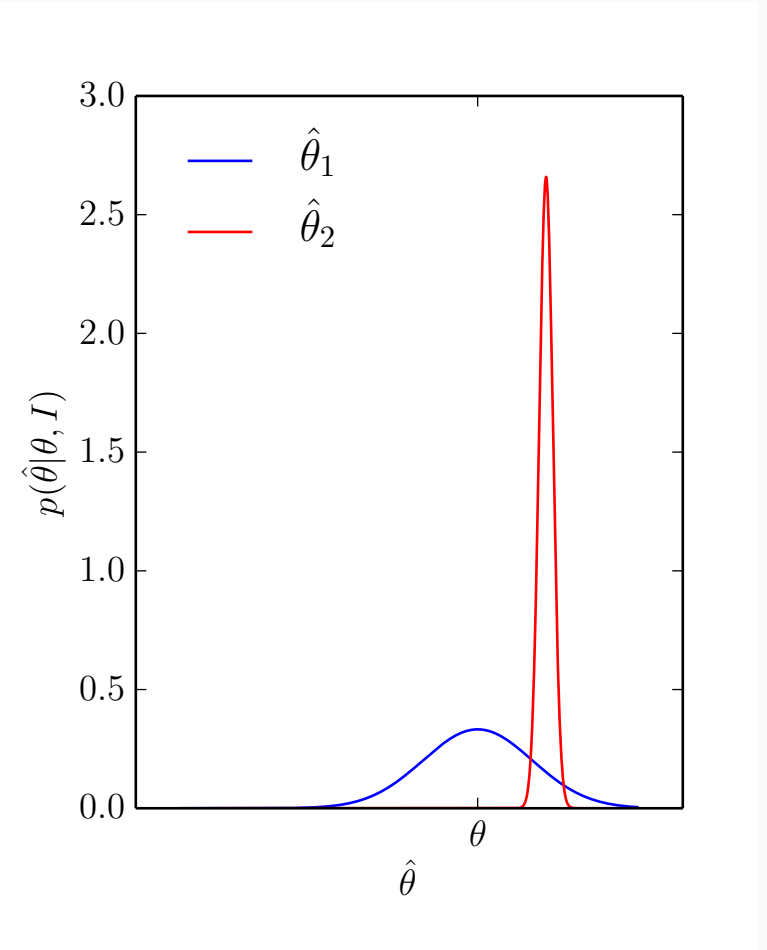
- $p(\hat{\theta}_1|\theta, I)$ is broad and carries a large **statistical error**, but does encompass the true value of θ
- $p(\hat{\theta}_2|\theta, I)$ is narrow, but offset from the θ, and can be said to have large **systematic errors**

It can be difficult to decide which estimator is *"best"* (especially when the true value is unknown).



- $\hat{\theta}_1$ is an **unbiased** estimator
  - repeated observations would average to the true value, e.g. $E[\hat{\theta}_1] = \int \hat{\theta}_1 p(\hat{\theta}_1|\theta, I) \mathrm{d}\hat{\theta}_1 = \theta$.
  - *but* var$[\hat{\theta}_1]$ is large
- $\hat{\theta}_2$ is a **biased** estimator
  - $E[\hat{\theta}_1] = \int \hat{\theta}_2 p(\hat{\theta}_2|\theta, I) \mathrm{d}\hat{\theta}_2 \neq \theta$.
  - *but* var$[\hat{\theta}_2]$ is small

If we could correct for bias $\hat{\theta}_2$ would be a better choice of estimator.



The simplest unbiased estimator is the **sample mean**. If we have a random sample $\{x_1, \ldots, x_n\}$ of length n drawn from pdf $p(x|I)$, which has unknown mean μ and variance σ^2, then

$$\text{sample mean} = \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

This is an unbiased estimator of μ in that $E[\hat{\mu}] = \mu$.

The variance, $\sigma^2_{\hat{\mu}}$, of the sample mean (i.e. the width of the distribution $p(\hat{\mu}|\mu, I)$) is

$$\sigma^2_{\hat{\mu}} = \sigma^2/n,$$

so as the sample size increases the sample mean distribution becomes more concentrated around the true mean (the *law of large numbers*).

If μ is known then an estimator for the variance is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2.$$

and $E[\hat{\sigma}^2] = \sigma^2$ (i.e. it is *unbiased*).

However, if μ is unknown and we instead use $\hat{\mu}$ in its place then

$$E[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2,$$

so it is *biased*. In this case an *unbiased* estimator of the **sample variance** is

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2.$$

Given sampled data $\{(x_i, y_i); i = 1, \ldots, n\}$ we can *estimate* the linear correlation between the variables as

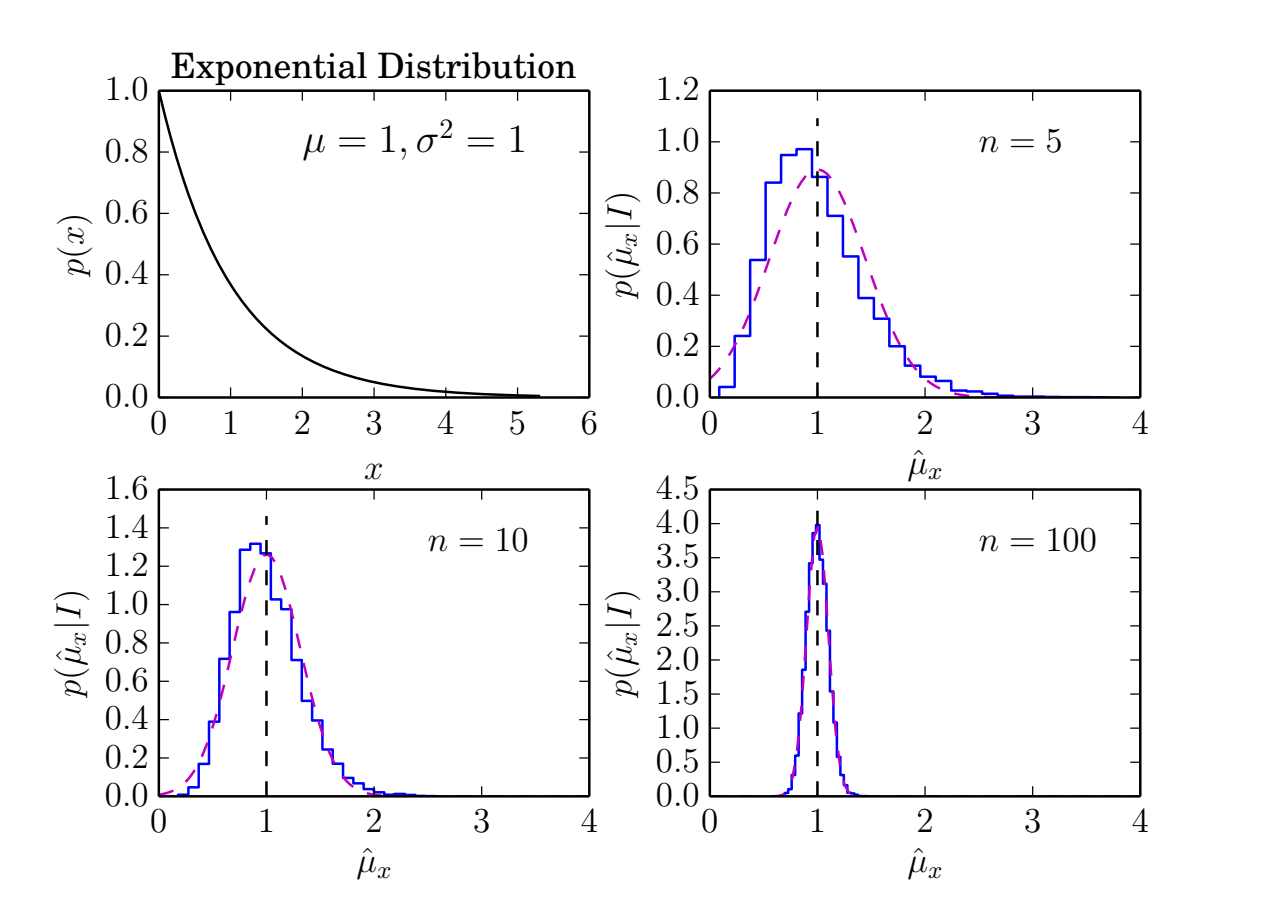
$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overbrace{\hat{\mu}_x}^{\text{sample mean}}}{\hat{\sigma}_x} \right) \left(\frac{y_i - \hat{\mu}_y}{\hat{\sigma}_y} \right)$$

where, e.g.,

$$\hat{\sigma}_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu}_x)^2}.$$

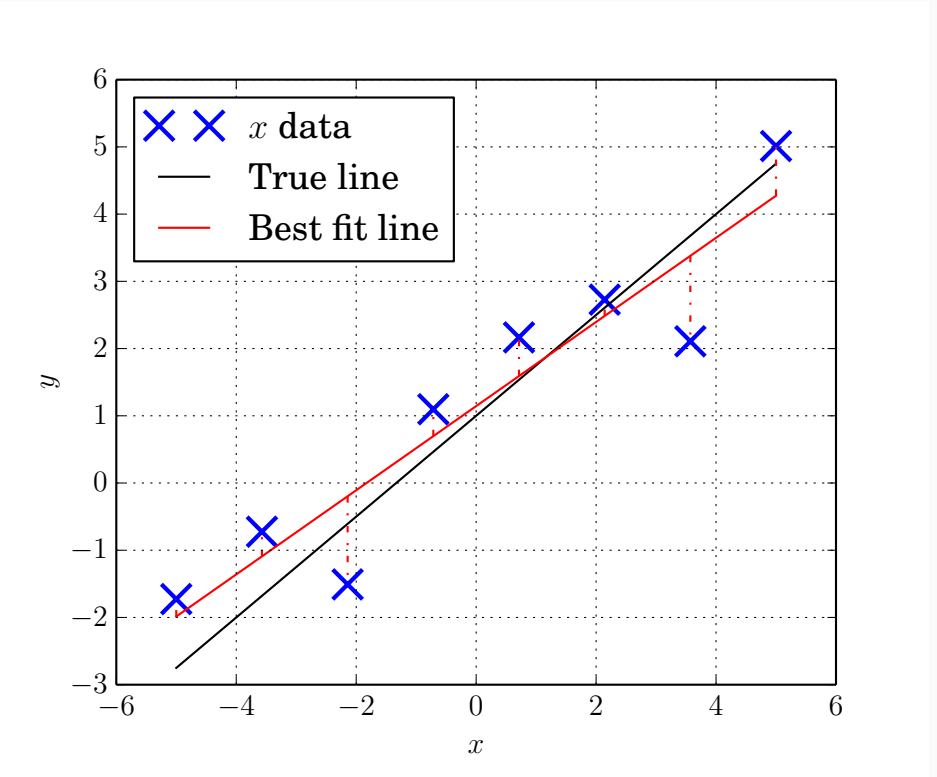
If $p(x, y|I)$ is a bivariate normal distribution then r is an **estimator** is the correlation coefficient ρ.

For *any* pdf with finite variance, σ^2, and mean, μ, central limit theorem states that as $n \to \infty$ the sample mean $\hat{\mu}$ has a normal pdf with mean μ and variance σ^2/n.



The method of **least squares** (LS) is a *standard* (often "black box") method for fitting lines and curves to data.

E.g. if we have some $\{x, y\}$ data least squares provides a way to find the *"best fit"* straight line $y = mx + c$ for it (i.e. estimates of the values of the parameters m and c that minimise the sum of the squared residuals).



Ordinary linear least squares assumes scatter in a plot of $\{x_i, y_i\}$ arises from errors in only one of the two variables. We call the variable with (say y) and without (say x) error the **dependent variable** and **independent variable** respectitvely. For each data point suppose we can write

$$y_i = mx_i + c + \epsilon_i$$

where ϵ_i is known as the **residual** of the i^{th} data point, i.e. the difference between the observed value y_i and the value predicted by the best-fit straight line.

We assume that the $\{\epsilon_i\}$ are an *iid* random sample from some underlying pdf with $\mu = 0$ and variance σ^2.

The **least squares estimators** of m and c minimise the function

$$S = \chi^2(m, c) = \underbrace{\sum_{i=1}^{n} (y_i - (mx_i + c))^2}_{\sum_{i=1}^{n} \epsilon_i^2}$$

so $\hat{m}_{\mathrm{LS}}$ and $\hat{c}_{\mathrm{LS}}$ satisfy

$$\left|\frac{\partial S}{\partial m}\right|_{m=\hat{m}_{\mathrm{LS}}} = 0 \text{ and } \left|\frac{\partial S}{\partial c}\right|_{c=\hat{c}_{\mathrm{LS}}} = 0$$

Solving these equations for the estimators we have:

$$\hat{m}_{\mathrm{LS}} = \frac{n\sum y_i x_i - \sum y_i \sum x_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

and

$$\hat{c}_{\mathrm{LS}} = \frac{\sum y_i \sum x_i^2 - \sum y_i x_i \sum x_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}.$$

We can see these estimators *are* statistics as they just depend on the data and not the m and c parameters, or the (potentially unknown) variance of the residuals σ^2.

In can also be shown that these least squares estimators are **unbiased**, i.e. $E[\hat{m}_{\mathrm{LS}}] = m$ and $E[\hat{c}_{\mathrm{LS}}] = c$.

Assuming a *known* variance on the residuals the variance of the estimators are

$$\begin{aligned}\mathrm{var}[\hat{m}_{\mathrm{LS}}] &= \frac{\sigma^2 n}{n\sum x_i^2 - (\sum x_i)^2} \\ \mathrm{var}[\hat{c}_{\mathrm{LS}}] &= \frac{\sigma^2 \sum x_i^2}{n\sum x_i^2 - (\sum x_i)^2}\end{aligned}$$

In general $\hat{m}_{\mathrm{LS}}$ and $\hat{c}_{\mathrm{LS}}$ will not be statistically independent, so they have a covariance given by

$$\mathrm{cov}[\hat{m}_{\mathrm{LS}}, \hat{c}_{\mathrm{LS}}] = \frac{-\sigma^2 \sum x_i}{n\sum x_i^2 - (\sum x_i)^2}$$

We can see that if $\sum x_i = 0$ (which, provided uniform sampling in x, we could always transform into) the estimators are uncorrelated.

If the residuals $\{\epsilon_i\}$ are drawn from pdfs with $\mu = 0$, but different *known* variances σ_i^2, then

$$S = \chi^2(m, c) = \sum_{i=1}^{n} \left[\frac{y_i - (mx_i + c)}{\sigma_i} \right]^2$$

We can find the **weighted least squares** estimators by again finding the solutions that minimise S, giving

$$\hat{m}_{\mathrm{LS}} = \frac{\sum \frac{1}{\sigma_i^2} \sum \frac{y_i x_i}{\sigma_i^2} - \sum \frac{y_i}{\sigma_i^2} \sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2} \right)^2}$$

and

$$\hat{c}_{\mathrm{LS}} = \frac{\sum \frac{y_i}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \sum \frac{y_i x_i}{\sigma_i^2} \sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2} \right)^2}.$$

This also gives variances and a covariance of

$$\mathrm{var}[\hat{m}_{\mathrm{LS}}] = \frac{\sum \frac{1}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\frac{x_i}{\sigma_i^2}\right)^2},$$

$$\mathrm{var}[\hat{c}_{\mathrm{LS}}] = \frac{\sum \frac{x_i^2}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\frac{x_i}{\sigma_i^2}\right)^2},$$

and

$$\mathrm{cov}[\hat{m}_{\mathrm{LS}}, \hat{c}_{\mathrm{LS}}] = \frac{-\sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\frac{x_i}{\sigma_i^2}\right)^2}$$

If σ_i^2 is constant these all reduce to the unweighted case.

What about errors on *both* variables? The χ^2 function to minimise becomes

$$\chi^2(m, c) = \sum_{i=1}^{n} \frac{(y_i - mx_i - c)^2}{\sigma_{y,i}^2 + m^2\sigma_{x,i}^2}.$$

The equations to minimise this are *non-linear*, so these is no simple analytic solution. They must be solved using numerical methods instead.

We can generalise the linear model, e.g. to an $(M-1)^{\text{th}}$ order polynomial

$$y(x) = a_1 + a_2 x + a_3 x^2 + \ldots + a_M x^{M-1} = \sum_{k=1}^{M} a_k x^{k-1}$$

or even more generally $y(x) = \sum_{k=1}^{M} a_k X_k(x_i)$, where $X(x)$ is some function of x multiplied by a coefficient, a, that we want an estimator for. We therefore have

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{y_i - \sum_{k=1}^{M} a_k X_k(x_i)}{\sigma_i} \right]^2$$

This can be put into matrix form to solve for the a parameters.

For the weighted case we have

$$\mathbf{a} = \underbrace{\begin{bmatrix} a_1 \\ \vdots \\ a_m \end{bmatrix}}_{\text{model parameters}}, \quad \mathbf{b} = \underbrace{\begin{bmatrix} y_1/\sigma_1 \\ \vdots \\ y_n/\sigma_n \end{bmatrix}}_{\text{weighted observations}}, \quad \mathbf{A} = \underbrace{\begin{bmatrix} \frac{X_1(x_1)}{\sigma_1} & \cdots & \frac{X_1(x_M)}{\sigma_1} \\ \vdots & \ddots & \vdots \\ \frac{X_n(x_1)}{\sigma_n} & \cdots & \frac{X_n(x_M)}{\sigma_n} \end{bmatrix}}_{\text{Design matrix } (n \times M)}$$

and the model

$$\mathbf{b} = \mathbf{A}\mathbf{a} + \mathbf{e}, \text{ where } \mathbf{e} = \begin{bmatrix} \epsilon_1/\sigma_1 \\ \vdots \\ \epsilon_n/\sigma_n \end{bmatrix}$$

We solve for the parameter vector $\widehat{\mathbf{a}}_{\mathrm{LS}}$ that minimises

$$\mathbf{S} = \mathbf{e}^{\mathrm{T}} \cdot \mathbf{e} = \sum_{i=1}^{n} e_i^2,$$

which has the solution

$$\widehat{\mathbf{a}}_{\mathrm{LS}} = \underbrace{\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1}}_{M \times M \text{matrix}} \mathbf{A}^{\mathrm{T}} \cdot \mathbf{b}.$$

and

$$\mathrm{cov}[\widehat{\mathbf{a}}_{\mathrm{LS}}] = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1}.$$

Note that inverting $\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)$ can be problematic in case where $\mathbf{A}$ is sparse and/or close to singular.

Other common cases are where the models are non-linear, or the errors on the observations are correlated.

In these cases numerical approaches have to be taken to find the least squares estimators.

In the frequentist approach a parameter is *a fixed (but unknown) constant*.

From actual data we can compute a **likelihood**, L, which is the probability of obtaining the data, given a the value of the parameter θ. Now define the **likelihood function**, $L(\theta)$, as the (infinite) family of curves L as a function of θ for fixed data.

Also, note the **likelihood function**, $L(\theta)$, is the *sampling distribution* $g(x_1, \ldots, x_n|\theta)$, but now with the *random sample* $\{x_1, \ldots, x_n\}$ (the data) being a function of a parameter θ.

The *principle of maximum likelihood* state that a good estimator of θ, $\hat{\theta}_{\mathrm{ML}}$, maximises $L(\theta)$, i.e.

$$\left.\frac{\partial L}{\partial \theta}\right|_{\theta=\hat{\theta}_{\mathrm{ML}}} = 0 \text{ and } \left.\frac{\partial^2 L}{\partial \theta^2}\right|_{\theta=\hat{\theta}_{\mathrm{ML}}} < 0$$

So $\hat{\theta}_{\mathrm{ML}}$ is the value of θ corresponding to the pdf from which it is 'most likely' that the data (random sample) was drawn.

We can see how the weighted least squares estimator falls out of the maximum likelihood method. Let's again consider the model $y_i = mx_i + c + \epsilon_i$, and assume that the i^{th} residual ϵ_i is drawn from a Gaussian (normal) pdf with mean of zero and variance σ_i^2.

We therefore have a *likelihood*

$$L = \prod_{i=1}^{n} p(\epsilon_i | I) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2}\frac{\epsilon_i^2}{\sigma_i^2}\right).$$

[Note: we use the product rule for each of the pdfs $p(\epsilon_i)$ as they are assumed independent.]

Substituting in $\epsilon_i = y_i - mx_i - c$ we have

$$L = (2\pi)^{n/2} \prod_{i=1}^{n} \frac{1}{\sigma_i} \exp\left(-\frac{1}{2}\frac{(y_i - mx_i - c)^2}{\sigma_i^2}\right).$$

The ML estimators of m and c will satisfy $\partial L/\partial m = 0$ and $\partial L/\partial c = 0$.

But, maximising L is equivalent to maximing $\ell = \ln L$, so

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

$$\ell = -\frac{n}{2}\ln(2\pi) - \ln\sum_{i=1}^{n}\sigma_i - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - mx_i - c}{\sigma_i}\right)^2$$
$$= \text{constant} - \frac{1}{2}S,$$

where S is exactly the same sum of squares defined earlier.

So, to find e.g. $\hat{m}_{\mathrm{ML}}$ we would have

$$\left.\frac{\partial L}{\partial m}\right|_{m=\hat{m}_{\mathrm{ML}}} = \left.\frac{\partial \ell}{\partial m}\right|_{m=\hat{m}_{\mathrm{ML}}} = \left.\frac{\partial S}{\partial m}\right|_{m=\hat{m}_{\mathrm{ML}}} = 0.$$

In the case where we have Gaussian, independent errors, the maximum likelihood and least squares estimators are identical.

In general a **simple hypothesis test** is one where we test a **null hypothesis**, H_0, against an alternative hypothesis, H_1. We construct a **test statistic**, t, and based on t we make a decision:

- accept H_0 and reject H_1
- accept H_1 and reject H_0

We must choose the **critical region** for the test statistic, t, as the set of values of t for which we choose to **reject** H_0 and accept H_1. The region in which we accept H_0 is the **acceptance region**.

We can make an incorrect decision in two ways:

- **type I error** - we *reject* the null hypothesis when it is actually **true** (also known as false dismissal)
- **type II error** - we *accept* the null hypothesis when it is actually **false** (also known as a false alarm)

The probability of type I and type II errors are sometimes denoted $P(I)$ and $P(II)$, and $P(II)$ is also known as the false alarm probability (FAP).

A 'good' hypothesis test should have small $P(I)$ *and* $P(II)$, however reducing $P(I)$ (by suitable choice of *critical region*) comes at the cost of increasing $P(II)$. So, often one must try and minimise some combination of $P(I)$ and $P(II)$.

One criterion is the **power** of the hypothesis test

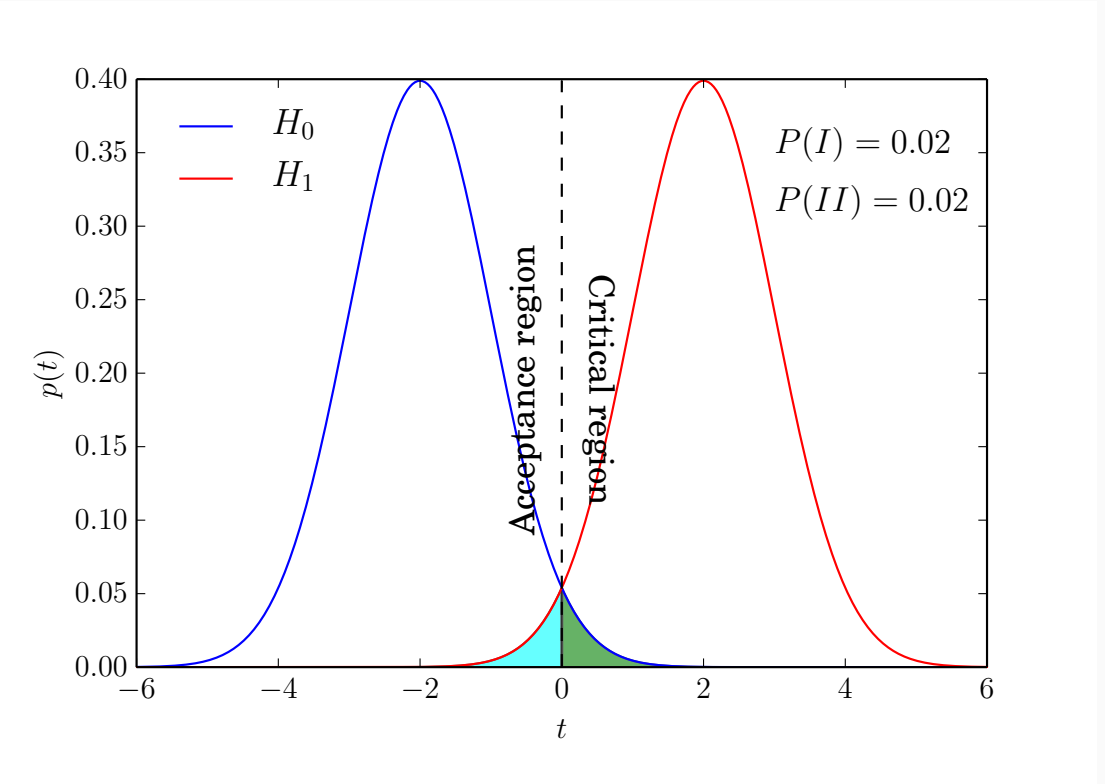
- *the probability of rejecting* H_0 *when it is false*
  power $= 1 - P(II)$

Choosing a critial region that *maximises* the power for a given alternative hypothesis can be a useful way to define a 'good' test.

A variable $x \sim N(\mu, 1)$ and our two hypotheses are:

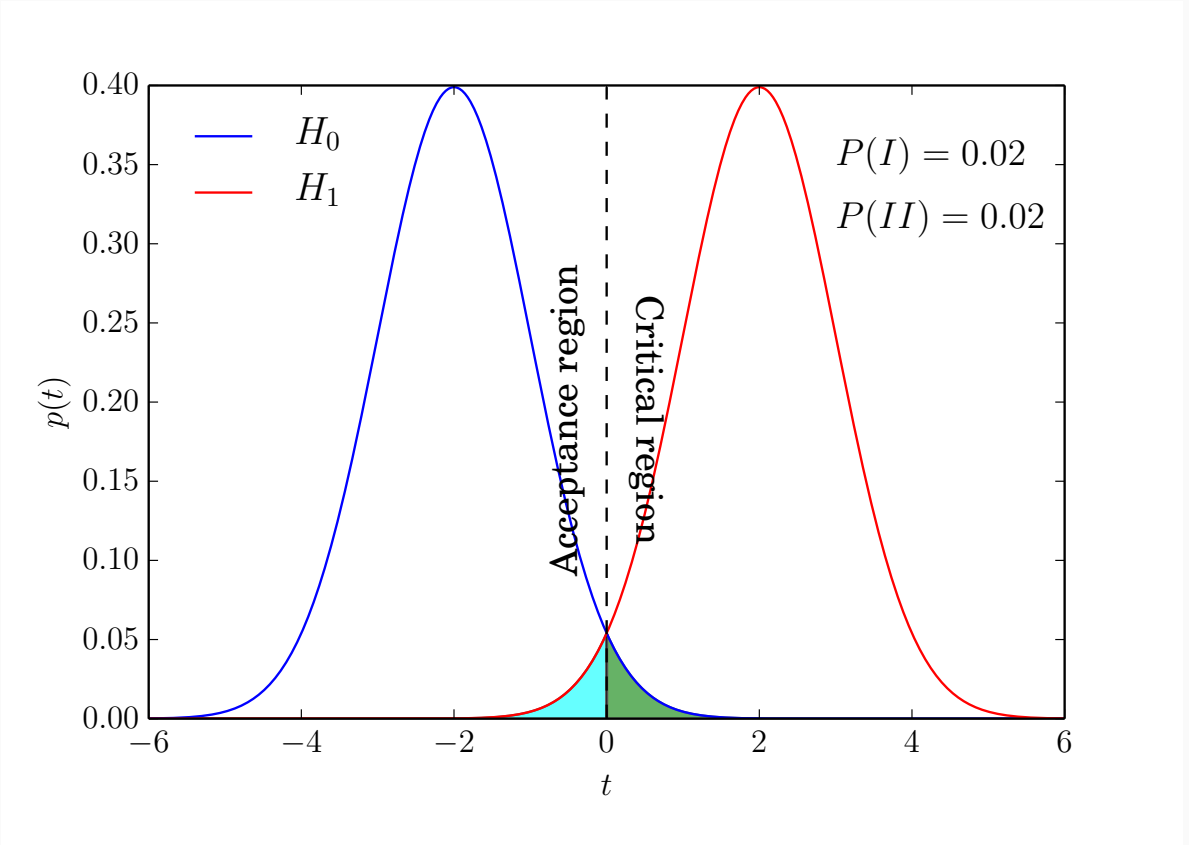
- H_0: $\mu = -2$
- H_1: $\mu = 2$

Our test statistic is simply $t = x$. The figure shows the distributions (assuming an infinite number of trials) for $p(t|H_0)$ and $p(t|H_1)$.



If we define our **critical region** for t (i.e. where we *reject* H_0) as $t > 0$ (so the *acceptance region* is $t \leq 0$) then we have:

- $p(I) = 0.02$
- $p(II) = 0.02$
- power $= 0.98$



The **level of significance** of a hypothesis test is the maximum probability of incurring a *type I* error that we are willing to risk. Commonly adopted levels are 5% or 1%.

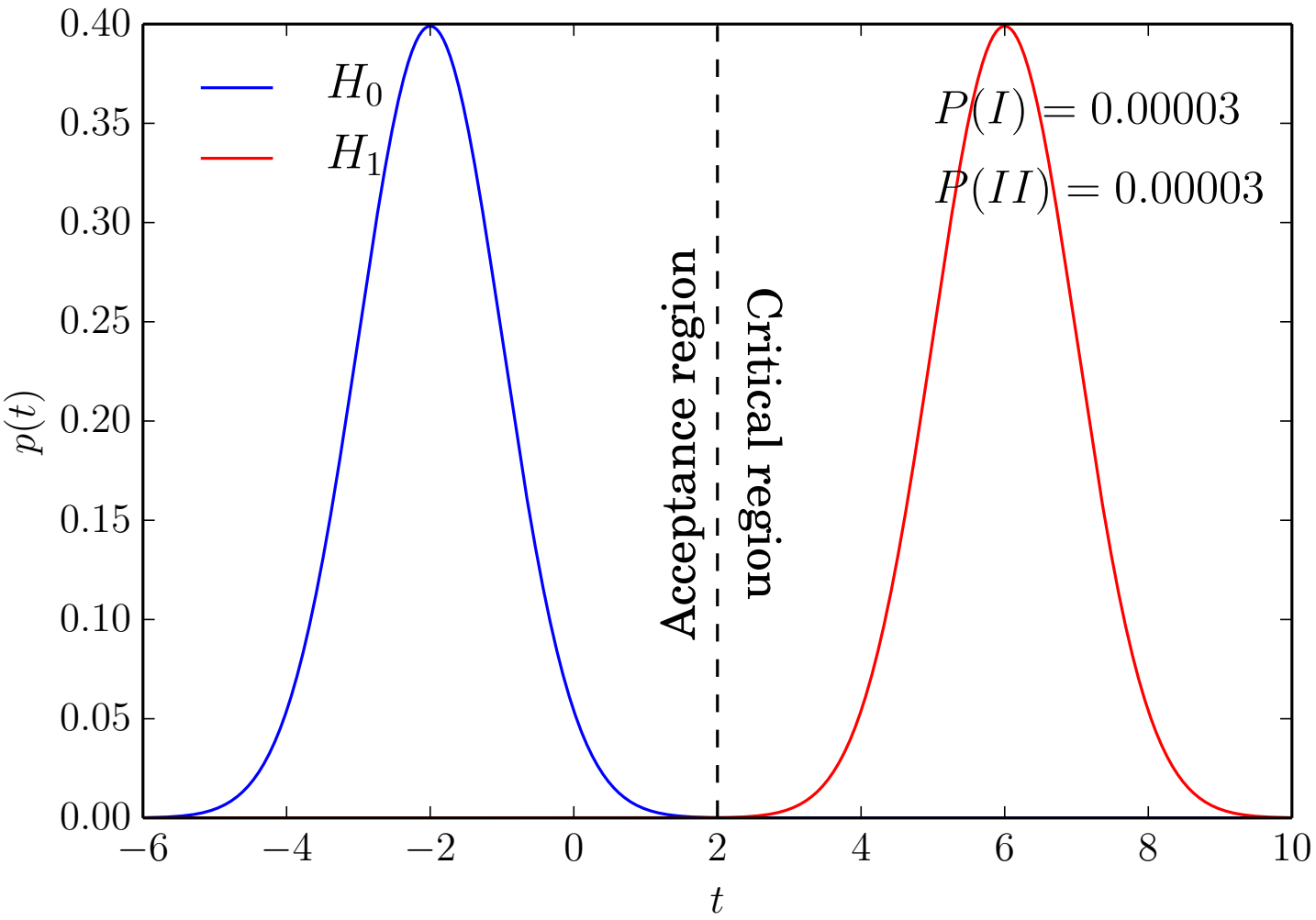
If using 5% then the *critical region* is chosen so that $P(I) \leq 0.05$. If the test statistic *is* within the critical region then we could say:

- the null hypothesis is rejected at the 5% level, or
- our rejecting of the null hypothesis is significant at the 5% level, or
- we are '95% confident' we have made the correct decision in rejecting the null hypothesis

This is saying: *if* the null hypothesis *is* true and we repeat our experiment a large number of times then we expect (by chance) the value of t to lie within the critical region in no more than 5% of them.

The choice of critical region and level of significance to assign is subjective. Ideally if the distributions of t for the two hypotheses have very little overlap (i.e. when signals are quite strong) one can be stringent in setting the critical region so that $P(I)$ is very small, whilst only modestly increasing $P(II)$. But, often that is not the case.

HYPOTHESIS TESTS: SIGNIFICANCE



Null hypothesis: *sampled data are drawn from a normal pdf* $N(\mu, \sigma^2)$ *with with* μ_{model} *and variance* σ^2. We want to **test** this null hypothesis (NH): are our data consistent with it?

Let's take some data, assuming a known variance and mean of their pdf, where:

- measured data: $\{x_i : i = 1, \ldots, 10\}$ and $\sum_{i=1}^{10} x_i = 47.8$, so the observed sample mean $\hat{\mu}_{\text{obs}} = 4.78$.
- null hypothesis is that $x \sim N(\mu_{\text{model}}, \sigma^2)$ where $\mu_{\text{model}} = 4$ and $\sigma = 2$.

Under the null hypothesis the sample mean, $\hat{\mu}_{\text{model}} \sim N(4, 2^2/10)$ (i.e. $\sigma_{\hat{\mu}} = \sigma^2/n = 2^2/10 = 0.4$). [Note: $x \sim N(\mu, \sigma^2)$ means x *is drawn from* N].

We can transform to a *standard normal variable*, so under the NH:

$$Z = \left(\frac{\hat{\mu}_{\text{obs}} - \hat{\mu}_{\text{model}}}{\sigma_{\hat{\mu}}} \right) \sim N(0, 1).$$

From our measured data:

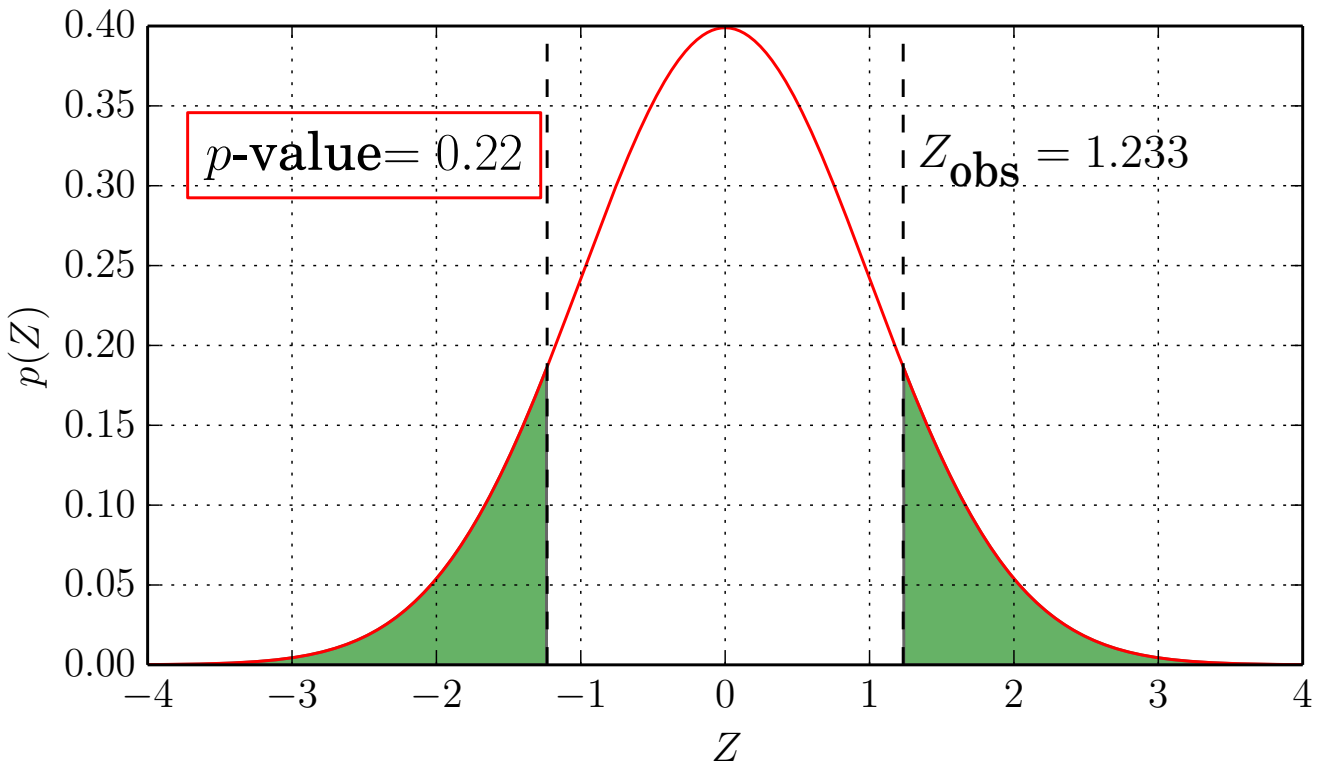
$$Z_{\text{obs}} = \frac{4.78 - 4}{\sqrt{0.4}} = 1.233.$$

So, *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.

SIMPLE HYPOTHESIS TEST EXAMPLE

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



In this case the p-value is 0.2176. The *smaller* the p-value, the less credible is the null hypothesis.

Note: the p-value can be calculated using

$$p\text{-value} = 1 - \mathrm{erf}(Z_{\mathrm{obs}}/\sqrt{2}),$$

where erf is the *error function*[3] $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-y^2} \mathrm{d}y$.

This is a *two-tailed* test, but the *one-tailed* test could be used when appropriate for statistics with other sampling distributions.

[3] The error function is available in e.g Matlab as `erf` and in python in `scipy.special.erf`.

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

Provided $n \geq 2$ we can estimate it from our observed data. We again form the statistic

$$t_{\text{obs}} = \left(\frac{\hat{\mu}_{\text{obs}} - \hat{\mu}_{\text{model}}}{\sigma_{\hat{\mu}}} \right),$$

but now the variance on the sample mean is

$$\sigma^2_{\hat{\mu}} = \frac{1}{n} \underbrace{\frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \hat{\mu}_{\text{obs}})^2}_{\text{sample variance}}$$

However, unlike Z_{obs} previously, t_{obs} no longer has a normal distribution.

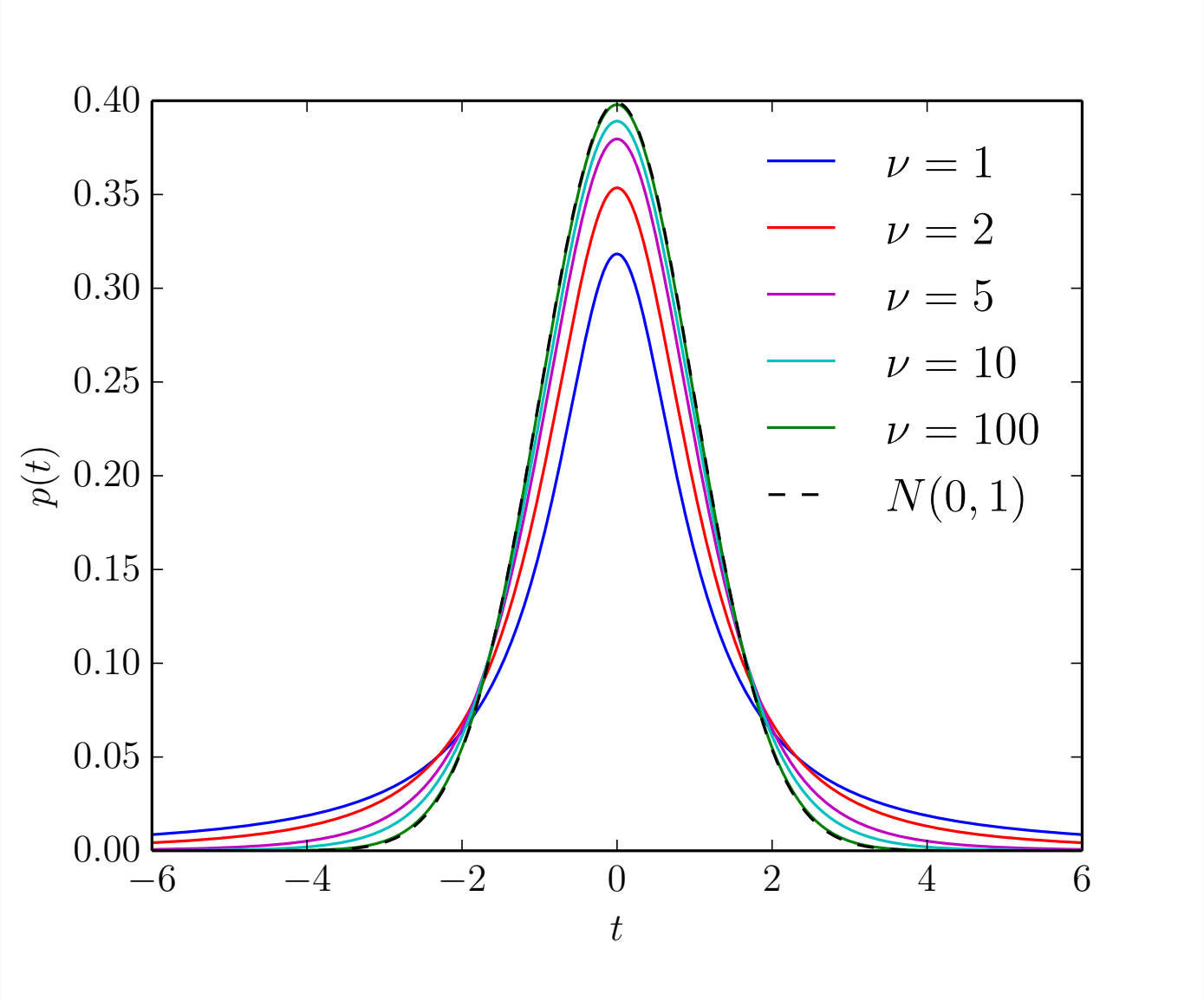
t_{obs} instead has a pdf known as the **Student's t-distribution**

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

where $\nu = n-1$ is the **number of degrees of freedom** and $\Gamma(\nu) = \int_0^\infty x^{\nu-1}e^{-x}\mathrm{d}x$.

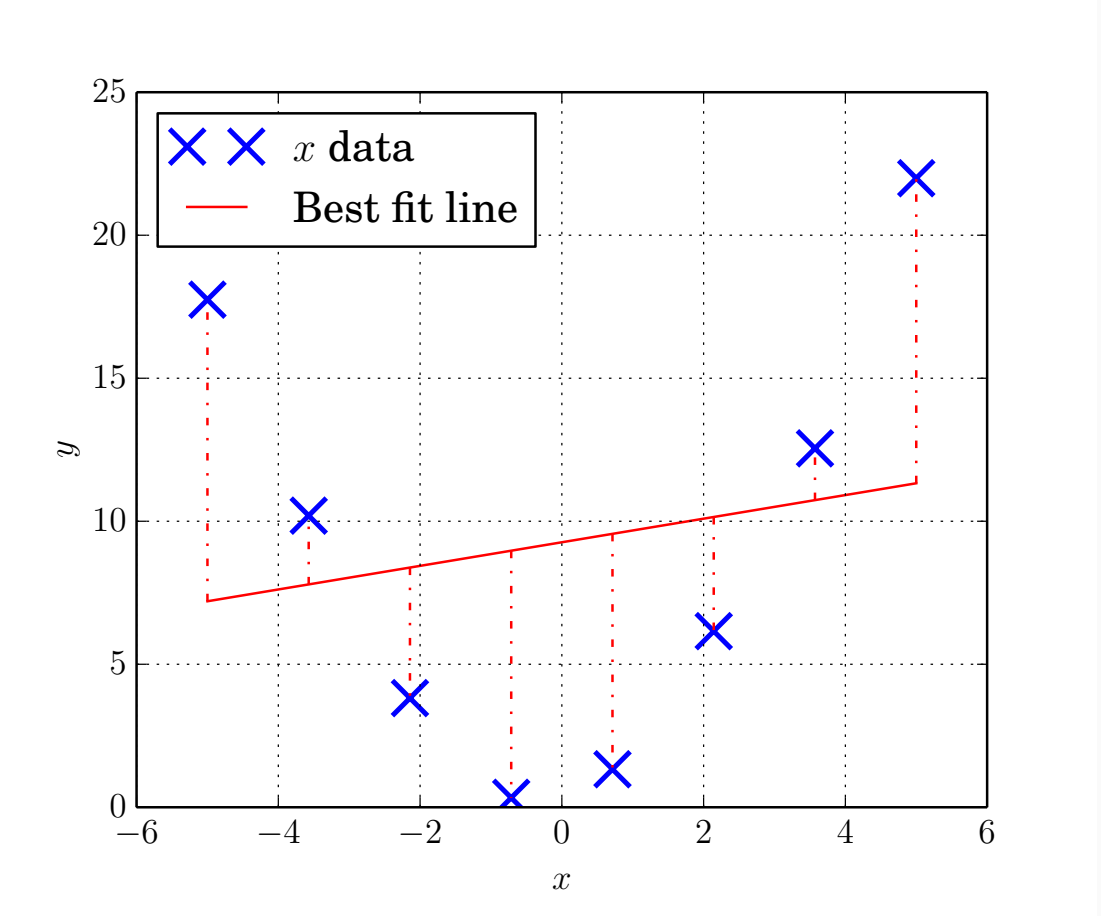
For small n (and therefore ν) the Students's t-distribution has more extended tails than a normal, but as $n \to \infty$ the distribution tends to $N(0,1)$.

SIMPLE HYPOTHESIS TEST EXAMPLE



GOODNESS OF FIT

As well as estimating parameters we might also want to ask how good our model was (e.g. a straight line) in the first place. Given some data we can always obtain a best fit line, but it still might be a very poor fit to the data.



Answering how good our model is is tantamount to asking whether the residuals of the data are actually drawn from their assumed distribution, i.e. a pdf with zero mean and a particular variance σ^2.

Suppose we have some true model m_{true} in the data y. The true residuals are given by

$$\epsilon_i = y_i - m_{\text{true},i},$$

but unless m_{true} is already known these residuals are, in fact, *unknown*. We only have our 'best fit' model (e.g. through least squares) $\hat{m}_{\text{LS}}$, so we *estimate* the residuals as

$$\hat{\epsilon}_i = y_i - \hat{m}_{\text{LS},i}.$$

A goodness of fit test is an example of a simple hypothesis test. The basic ideas of a goodness-of-fit test are:

- choose a **null hypothesis**, where in this case a null hypothesis is defined as the statement being tested in our goodness of fit test[4], for which we can evaluate a confidence level for its validity
- select a suitable statistic that can be computed from the data and has a predictable distribution[5], e.g. a normal distribution.

[4] Often a null hypothesis is the statement that there is 'no effect' present (e.g. the sampled data are consistant with random noise, with a specified distribution, rather than contain a signal).

[5] the distribution we could expect to obtain under an infinite number of repeats of the data.

A commonly used goodness-of-fit test for model fitting is the χ^2 statistic, given by

$$\chi^2 = \sum_{i=1}^{n} \frac{\hat{\epsilon}_i^2}{\sigma_i^2} = \sum_{i=1}^{n} \frac{(y_i - \hat{m}_i)^2}{\sigma_i^2}$$

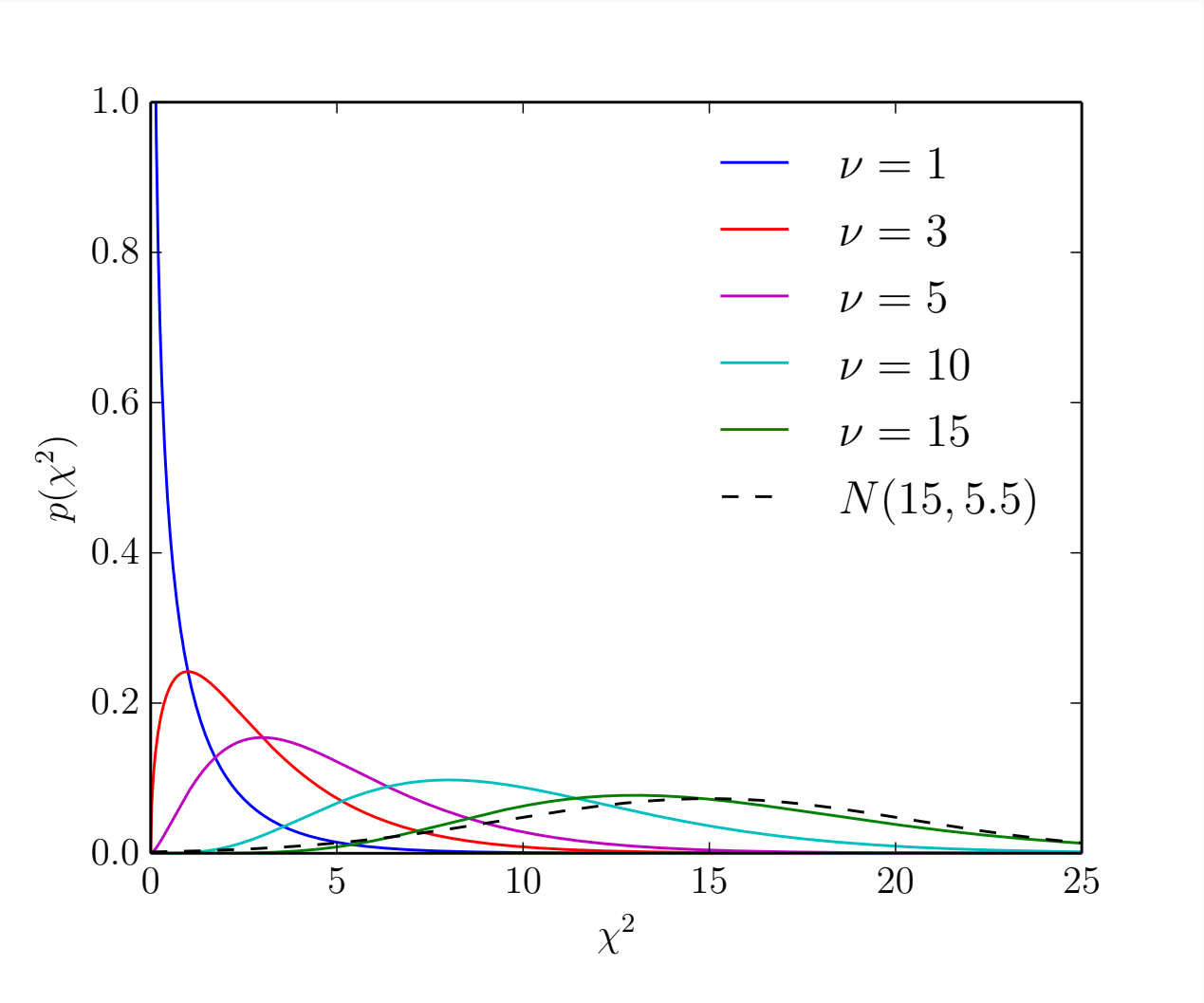
where y is the observed data, $\hat{m}$ is the 'best fit' model (from e.g. the LS fit) and $\hat{\epsilon} = y_i - \hat{m}_i$ are the estimated residuals.

In this case, unless we know σ_i *a priori*, we can say nothing about the goodness of fit of our model. But, if the residuals are distributed as $N(0, \sigma_i^2)$, then the χ^2 statistic has a pdf given by

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

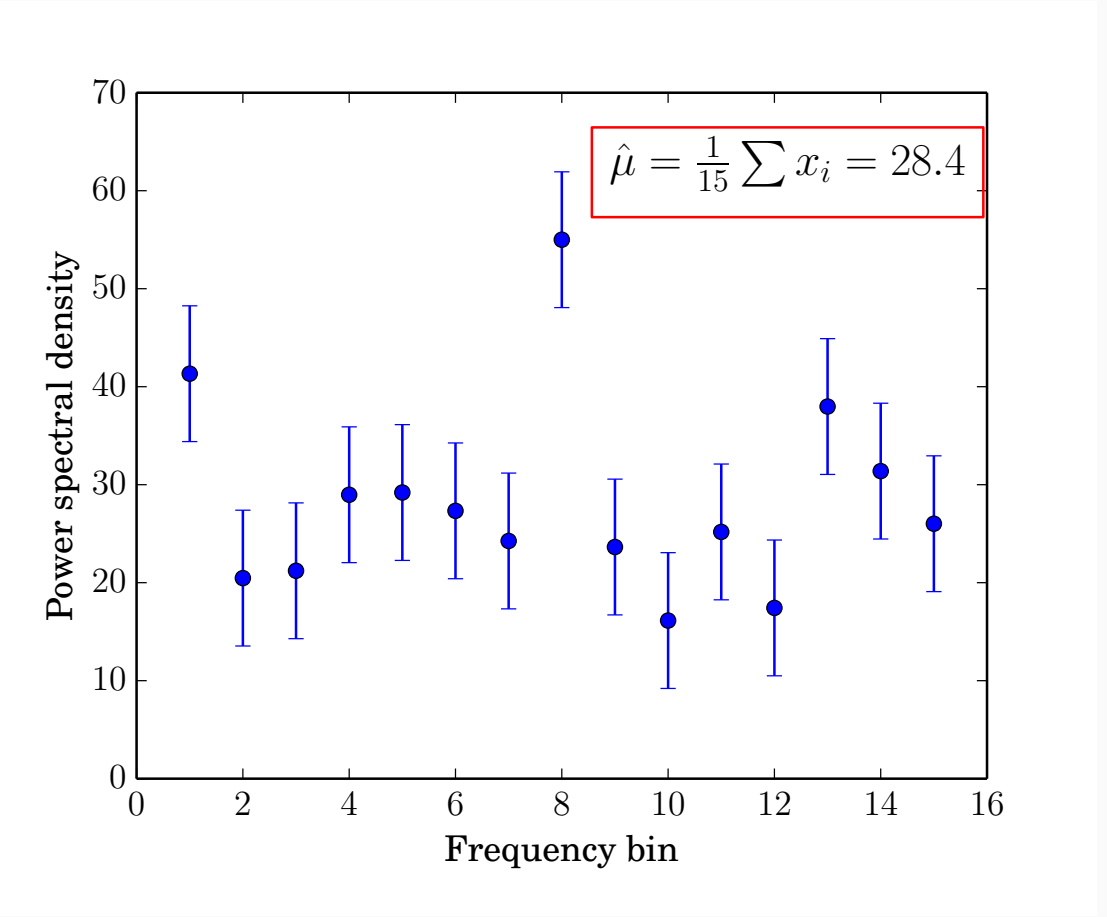
GOODNESS OF FIT: χ^2 TEST

Here ν is the number of **degrees of freedom** of the pdf, and the pdf has a mean of ν and variance of 2ν. As $\nu \to \infty$ the pdf tends to a normal pdf.



GOODNESS OF FIT: χ^2 TEST EXAMPLE

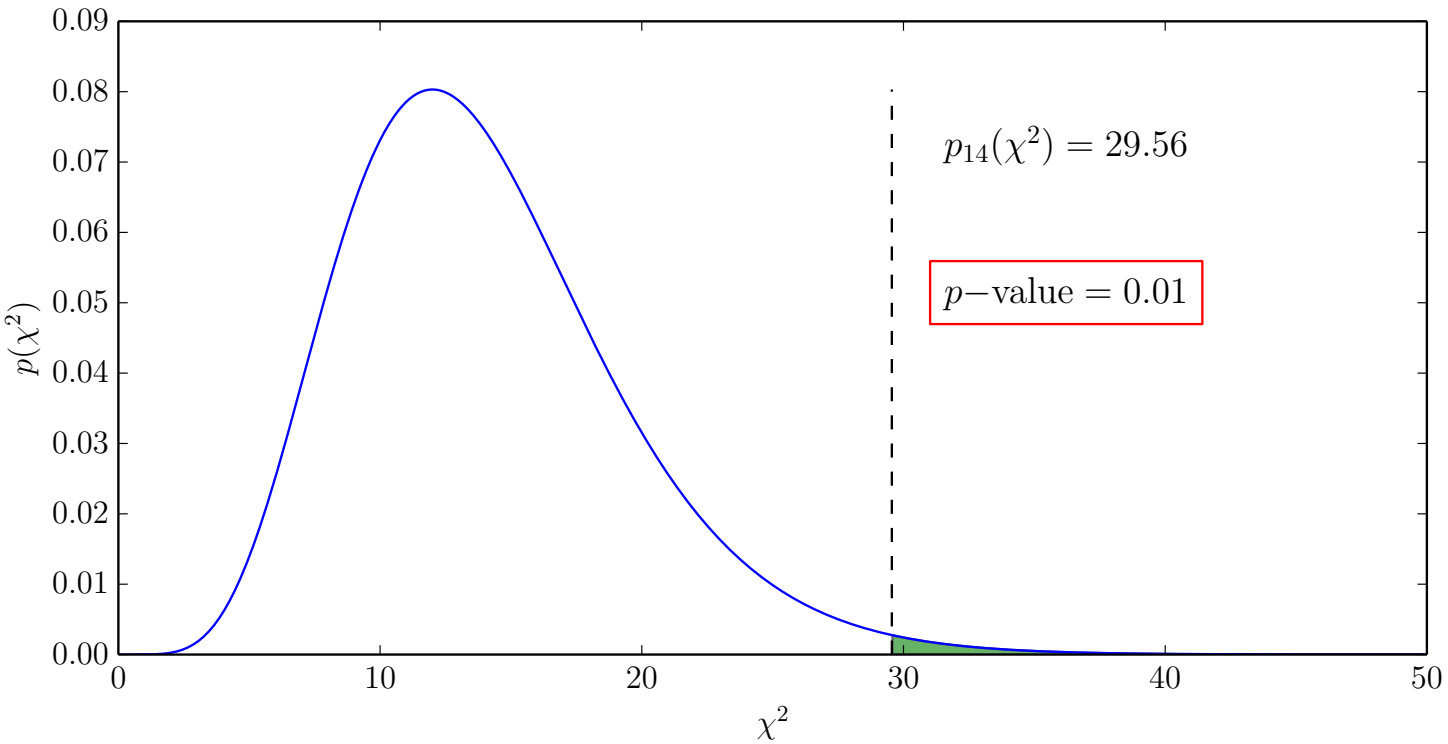
We have some power spectral density data, with a measurement error on each point $\sim N(0, 6.9^2)$. Our **null hypothesis** is: *the spectrum is constant, or flat, over all frequencies (i.e. there are no spectral lines).* We assume the residuals are iid and $\epsilon \sim N(0, \sigma^2)$.



GOODNESS OF FIT: χ^2 TEST EXAMPLE

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \hat{\mu})^2}{\sigma^2} = \sum_{i=1}^{15} \frac{(x_i - 28.4)^2}{6.9^2} = 28.4$$

$$p_{\nu=(n-1)=14}(\chi^2) = p_{14}(28.4) = 29.6.$$



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

We can again calculate a **p-value**, where this time is it given by

$$\begin{aligned} p\text{-value} &= 1 - P(\chi^2_{\text{obs}} \geq \chi^2(\nu = 14)), \\ &= 1 - \int_0^{\chi^2_{\text{obs}}} p_0 x^{\frac{\nu}{2}-1} e^{-x/2} \mathrm{d}x = 0.01, \end{aligned}$$

where $p_0 = 1/(2^{\nu/2}\Gamma(\nu/2))$.

What does this p-value mean?

If the spectrum really is flat, and we repeatedly obtained spectra of the same length under the same conditions, then only 1% of the χ^2 values derived from these sets would be expected to be greater than our one actual measured value of 29.6.

I.e. 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 to reject a hypothesis is subjective, but the χ^2 test can help in the decision.

The **Kolmogorov-Smirnov test** (KS test) is a useful way to test the null hypothesis that a random sample is drawn from a particular underlying pdf[6], $p(x)$, with a known cdf $P(x)$.

If the iid sample $\{x_1, \ldots, x_n\}$ is arranged in ascending order, the sample cdf, $S_n(x)$, is

$$S_n(x) = \begin{cases} 0, & \text{if } x < x_1 \\ \frac{i}{n}, & \text{if } x_i \leq x < x_{i+1}, \text{ for } 1 \leq i \leq n-1 \\ 1, & \text{if } x \geq x_n, \end{cases}$$

then the KS test statistic is defined as: $D_n = \max |P(x) - S_n(x)|$.

[6]The two sample KS test can also be used to compare whether two random samples are drawn from the same underlying pdf
$D_{m,n} = \max |S_m(x) - S_n(x)|$.

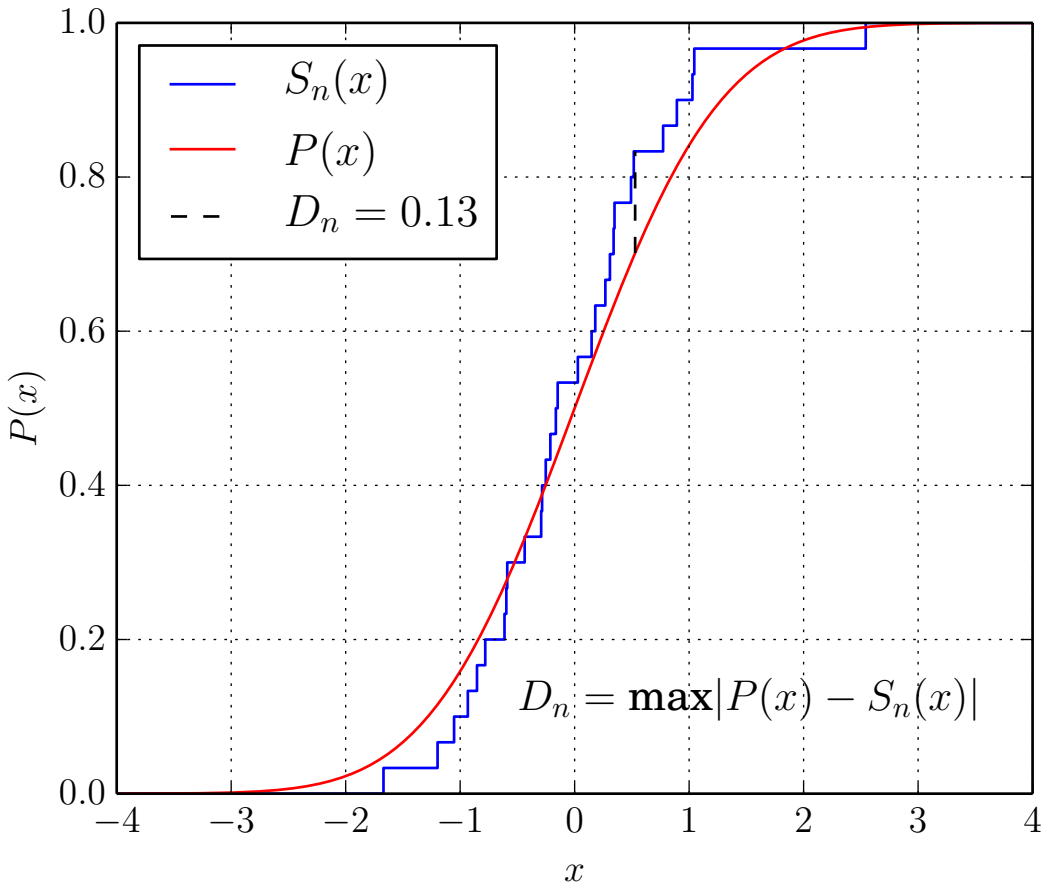
The distribution of D_n under the null hypothesis is independent of the actual form of $P(x)$. Critical values for the Kolmogorov distribution can be found online and associated p-values calculated.

The KS test for the case that the underlying pdf is normal is often called the *Lilliefors test*.

The KS test is an example of a **nonparametric** test as there are minimum assumptions of a parametric form of $p(x)$. However, this means its **power** is often lower than for other parametric tests, i.e. there is a higher chance of false acceptance of the null hypothesis.

GOODNESS OF FIT: KOLMOGOROV-SMIRNOV TEST

We have a random sample of 30 points $x \sim N(0,1)$ compared to a null hypothesis that $p(x) = N(0,1)$.



A simple test of the goodness of fit of two competing hypotheses, H_0 and H_1, defined by parameters θ_0 and θ_1 respectively, is to form the likelihood ratio[7] (generally using the maximum likehood estimators $\hat{\theta}_0$ and $\hat{\theta}_1$)

$$\Lambda(d) = \frac{L(\hat{\theta}_0)}{L(\hat{\theta}_1)}$$

As with the other statistics we need to know the pdf of $\Lambda(d)$ to define critical regions for accessing acceptance/rejection of hypotheses.

[7] Unlike the Bayesian odds ratio discussed in Part 3 this does not take into account any Occam factor

Additional one could use the log-likelihood ratio test:

$$D = -2 \ln \left(\frac{L(\hat{\theta}_0)}{L(\hat{\theta}_1)} \right) = 2 \left(\ln L(\hat{\theta}_1) - \ln L(\hat{\theta}_0) \right)$$

The pdf of the test statistic D is approximately a χ^2-distribution with $\nu = (n_1 - n_0)$ degrees of freedom, where n_1 and n_0 are the number of free parameters for H_0 and H_1 respectively.

This type of statistic is common in gravitational wave data analysis, e.g. the $\mathcal{F}$-statistic [1] used in continuous wave searches.

A simple example is:

- H_0: the data, d, consists of Gaussian noise with known σ^2, but an unknown mean μ
- H_1: the data, consists of Gaussian noise with known σ^2 and a mean of zero.

The best estimate of μ is just the sample mean $\hat{\mu}$, so using a Gaussian likelihood function we have

$$\Lambda = \frac{\cancel{(2\pi\sigma^2)^{-n/2}} \exp\left(-\sum_{i=1}^{n} \frac{(d_i-\hat{\mu})^2}{2\sigma^2}\right)}{\cancel{(2\pi\sigma^2)^{-n/2}} \exp\left(-\sum_{i=1}^{n} \frac{d_i^2}{2\sigma^2}\right)} = \exp\left(-\sum_{i=1}^{n} \frac{\hat{\mu}^2 - 2d_i\hat{\mu}}{2\sigma^2}\right)$$

and

$$D = \frac{\hat{\mu}\sum_{i=1}^{n}(\hat{\mu} - 2d_i)}{\sigma^2} = -\frac{\sum_{i=1}^{n}\hat{\mu}\, d_i}{\sigma^2} = -\frac{\left(\sum_{i=1}^{n} d_i\right)^2}{n\sigma^2}$$

GOODNESS OF FIT: OTHERS TESTS

Other common tests are:

- F-test - test where two random samples have the same variance using the f-statistic defined as the ratio of the variances
- sample correlation coefficient - test whether two variables are statistically independent

In parameter estimation we had a point **estimator** for a parameter i.e. a single number, $\hat{\theta}$, which we associate with the *true* (but unknown) value of the parameter θ. But, we might want to assess the likely *range* of the true values of θ. To do this we can define **confidence intervals**.

If we know (or assume) the pdf of the estimator $p(\hat{\theta})$ (e.g. a normal $N(\hat{\theta}, \sigma^2_{\hat{\theta}})$) then we can define a confidence interval for θ $[\theta_a, \theta_b]$ as

$$X = \text{Prob}(\theta_a \leq \theta \leq \theta_b) = \int_{\theta_a}^{\theta_b} p(\hat{\theta}) \mathrm{d}\hat{\theta}.$$

$$X = \text{Prob}(\theta_a \leq \theta \leq \theta_b) = \int_{\theta_a}^{\theta_b} p(\hat{\theta}) \mathrm{d}\hat{\theta},$$

where, e.g. $X = 0.95$ would give the **95% confidence interval** and $[\theta_a, \theta_b]$ would be the **95% confidence limits**.

Note that θ_a and θ_b are not unique, but you could define them to represent the *shortest interval* or to be symmetric about $\hat{\theta}$.

If $p(\hat{\theta})$ is a normal distribution then

$$\text{Prob}\left(\hat{\theta} - 1.96\sqrt{\sigma^2_{\hat{\theta}}} \leq \theta \leq \hat{\theta} + 1.96\sqrt{\sigma^2_{\hat{\theta}}}\right) = 0.95$$

CONFIDENCE INTERVALS: INTERPRETATION

For a single random sample (experiment) $\hat{\theta}$ is a unique number, so the probability that the true θ lies in a chosen confidence interval is either zero or one.

To interpret confidence intervals one needs to think in terms of repeating the experiment/observation that produced the random sample a large number of times. For each a different value of $\hat{\theta}$ would be produced, and hence also different confidence limits for θ, for the *same* fixed (but unknown) θ.

Thus confidence intervals, for e.g. $X = 0.95\%$, mean that we would expect θ to lie within their range in 95% of a large number of experiments. Thus, we are **95% confident** that θ lies within the interval given from our actual observed value $\hat{\theta}$.

[1] P. Jaranowski, A. Królak, and B. F. Schutz. Data analysis of gravitational-wave signals from spinning neutron stars: The signal and its detection. *Phys. Rev. D*, 58(6):063001, September 1998.