Statistical Data Analysis 2024/25
Lecture Week 6

London Postgraduate Lectures on Particle Physics

University of London MSc/MSci course PH4515

Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Course web page via RHUL moodle (PH4515) and also

www.pp.rhul.ac.uk/~cowan/stat_course.html

Statistical Data Analysis
Lecture 6-1

- p-values
- Definition
- Important properties
- Relationship to hypothesis test

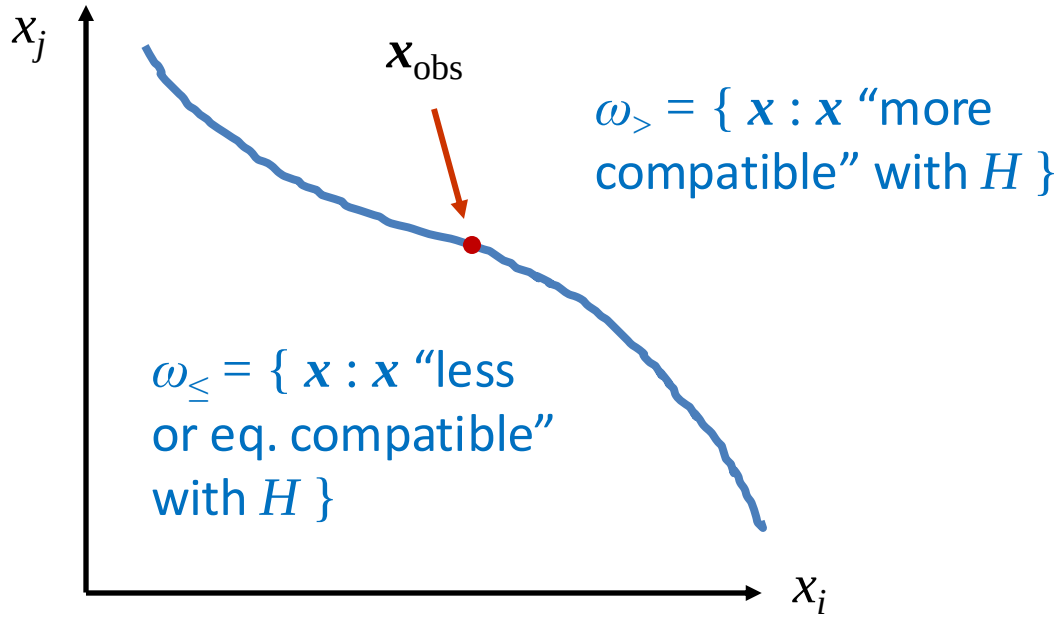
Testing significance / goodness-of-fit

Suppose hypothesis H predicts pdf $f(\boldsymbol{x}|H)$ for a set of observations $\boldsymbol{x} = (x_1,\ldots x_n)$.

We observe a single point in this space: $\boldsymbol{x}_{\text{obs}}$.

How can we quantify the level of compatibility between the data and the predictions of H?

Decide what part of the data space represents equal or less compatibility with H than does the point $\boldsymbol{x}_{\text{obs}}$. (Not unique!)



p-values

Express level of compatibility between data and hypothesis (sometimes 'goodness-of-fit') by giving the p-value for H:

$$p = P(\mathbf{x} \in \omega_{\leq}(\mathbf{x}_{\text{obs}})|H)$$

= probability, under assumption of H, to observe data with equal or lesser compatibility with H relative to the data we got.

= probability, under assumption of H, to observe data as discrepant with H as the data we got or more so.

Basic idea: if there is only a very small probability to find data with even worse (or equal) compatibility, then H is "disfavoured by the data".

If the p-value is below a user-defined threshold α (e.g. 0.05) then H is rejected (equivalent to hypothesis test as discussed previously).

p-value of H is not $P(H)$

The p-value of H is not the probability that H is true!

In frequentist statistics we don't talk about $P(H)$ (unless H represents a repeatable observation).

If we do define $P(H)$, e.g., in Bayesian statistics as a degree of belief, then we need to use Bayes' theorem to obtain

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

where $\pi(H)$ is the prior probability for H.

For now stick with the frequentist approach;
result is p-value, regrettably easy to misinterpret as $P(H)$.

Compatibility with H

What does it mean for a region of data space to be less compatible with the predictions of H?

It must mean that that region of data space is more compatible with some relevant alternative H'.

So although the definition of the p-value does not need to refer explicitly to an alternative, this enters implicitly through its role in determining the partitioning of the data space into more and less-or-equally compatible regions.

As in the case of hypothesis tests, there may be more than one relevant alternative.

Example of p-value: exponential decay time

A nuclear sample contains two radioactive isotopes with mean lifetimes $\tau = 0.2$ s and $\tau = 1.0$ s.

For either isotope we expect the decay time to follow $f(t|\tau) = \frac{1}{\tau} e^{-t/\tau}$

A nucleus is observed to decay after a time $t_{\text{obs}} = 0.6$ s.

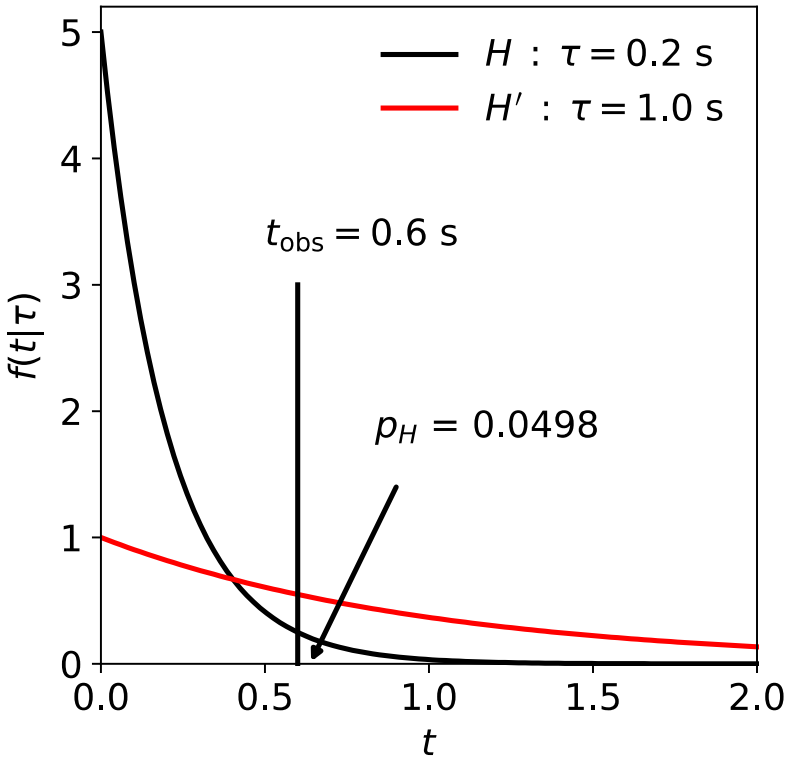
The p-value of the hypothesis H that the nucleus is of the type with $\tau = 0.2$ s is

$$p_H = P(t \geq t_{\text{obs}} | \tau = 0.2\,\text{s}) = 0.0498$$

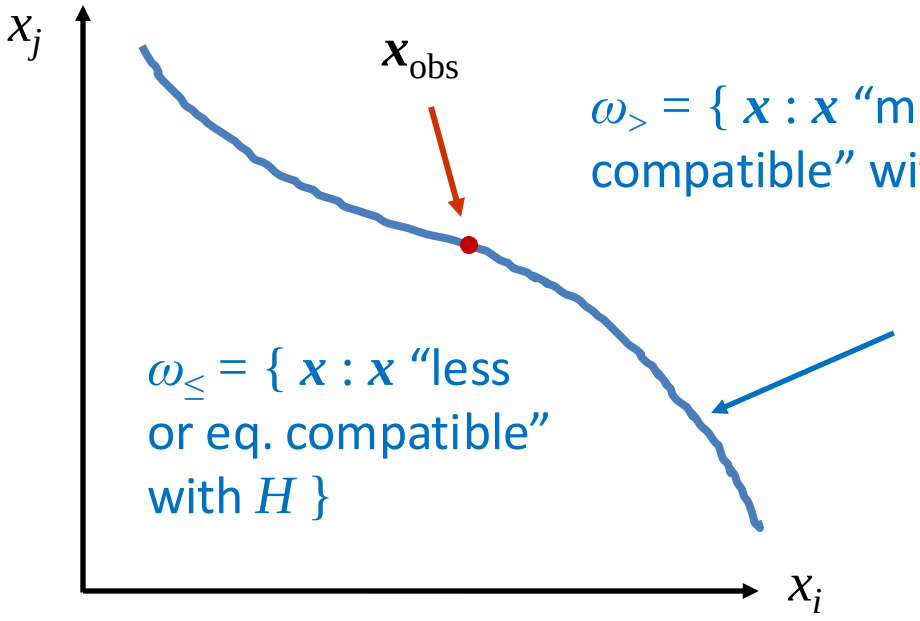
Here we take $t \geq t_{\text{obs}}$ as being less compatible with $\tau = 0.2$ s , because greater t is more characteristic of $\tau = 1.0$ s.

If the relevant alternative had been $\tau = 0.1$ s, then one would define the p-value as

$$p_H = P(t \leq t_{\text{obs}} | \tau = 0.2\,\text{s}) = 0.9502$$



p-value from test statistic



If e.g. we define the region of less or eq. compatibility to be $t(\boldsymbol{x}) \geq t_{\rm obs}$ then the p-value of H is

$$p_H = \int_{t_{\rm obs}}^{\infty} f(t|H)\, dt = \int_{\{\mathbf{x}\,:\,t(\mathbf{x})\geq t_{\rm obs}\}} f(\mathbf{x}|H)\, d\mathbf{x}$$

Distribution of the p-value

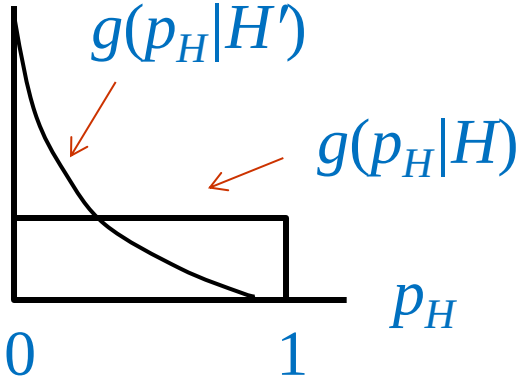
The p-value is a function of the data, and is thus itself a random variable with a given distribution. Suppose the p-value of H is found from a test statistic $t(\boldsymbol{x})$ as

$$p_H = \int_t^\infty f(t'|H)dt'$$

The pdf of p_H under assumption of H is

$$g(p_H|H) = \frac{f(t|H)}{|\partial p_H/\partial t|} = \frac{f(t|H)}{f(t|H)} = 1 \quad (0 \le p_H \le 1)$$

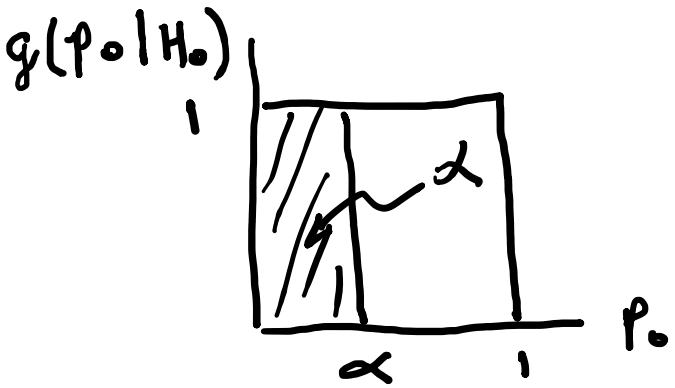
In general for continuous data, under assumption of H, p_H ~ Uniform[0,1] and is concentrated toward zero for some (broad) class of alternatives.



Using a p-value to define test of H_0

So the probability to find the p-value of H_0, p_0, less than α is

$$P(p_0 \leq \alpha | H_0) = \alpha$$



We started by defining critical region in the original data space ($\boldsymbol{x}$), then reformulated this in terms of a scalar test statistic $t(\boldsymbol{x})$.

We can take this one step further and define the critical region of a test of H_0 with size α as the set of data space where $p_0 \leq \alpha$.

Formally the p-value relates only to H_0, but the resulting test will have a given power with respect to a given alternative H_1, and the test statistic $t(\boldsymbol{x})$ used to obtain the p-value can be designed to achieve this, e.g., a likelihood ratio $t(\boldsymbol{x}) = P(\boldsymbol{x}|H_1)/P(\boldsymbol{x}|H_0)$.

Statistical Data Analysis
Lecture 6-2

- More examples *of* p-values
  - Coin
  - Poisson counting experiment
- Equivalent Gaussian significance

p-value example: testing whether a coin is 'fair'

Probability to observe n heads in N coin tosses is binomial:

$$P(n; p, N) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

Hypothesis H: the coin is fair (p = 0.5).

Suppose we toss the coin $N = 20$ times and get $n = 17$ heads.

Region of data space with equal or lesser compatibility with H relative to $n = 17$ is: $n = 17, 18, 19, 20, 0, 1, 2, 3$. Adding up the probabilities for these values gives:

$$P(n = 0, 1, 2, 3, 17, 18, 19, \text{ or } 20) = 0.0026 \; .$$

i.e. $p = 0.0026$ is the probability of obtaining such a bizarre result (or more so) 'by chance', under the assumption of H.

p-value example for coin (2)

Note that the region of equal or lesser compatibility seems "obvious" but could be different.

For example, suppose the person tossing the coin works for the "Mostly-Heads-Trick-Coin Company", then maybe $\omega_{\leq} = \{17,18,19,20\}$, and $p_{\text{fair}} = 0.0013$.

Note as well the clear distinction between the p-value of a fair coin and the probability (degree of belief) that the coin is fair:

Suppose you get the coin as change at a cafe. You then flip the coin 20 times and get 17 heads:

p-value $p_{\text{fair}} = 0.0026$,
P(fair) = probably still close to 1, depending on prior π(fair).

Suppose a representative of the MHTC Co. proposes a betting game in which they win money from you if there is an excess of heads. The result is 17 heads out of 20. P(fair) = low.

The Poisson counting experiment

Suppose we do a counting experiment and observe n events.

Events could be from *signal* process or from *background* – we only count the total number.

Poisson model:

$$P(n|s,b) = \frac{(s+b)^n}{n!}e^{-(s+b)}$$

s = mean (i.e., expected) # of signal events

b = mean # of background events

Goal is to make inference about s, e.g.,

test $s = 0$ (rejecting $H_0 \approx$ "discovery of signal process")

test all non-zero s (values not rejected = confidence interval)

In both cases need to ask what is relevant alternative hypothesis.

Poisson counting experiment: discovery p-value

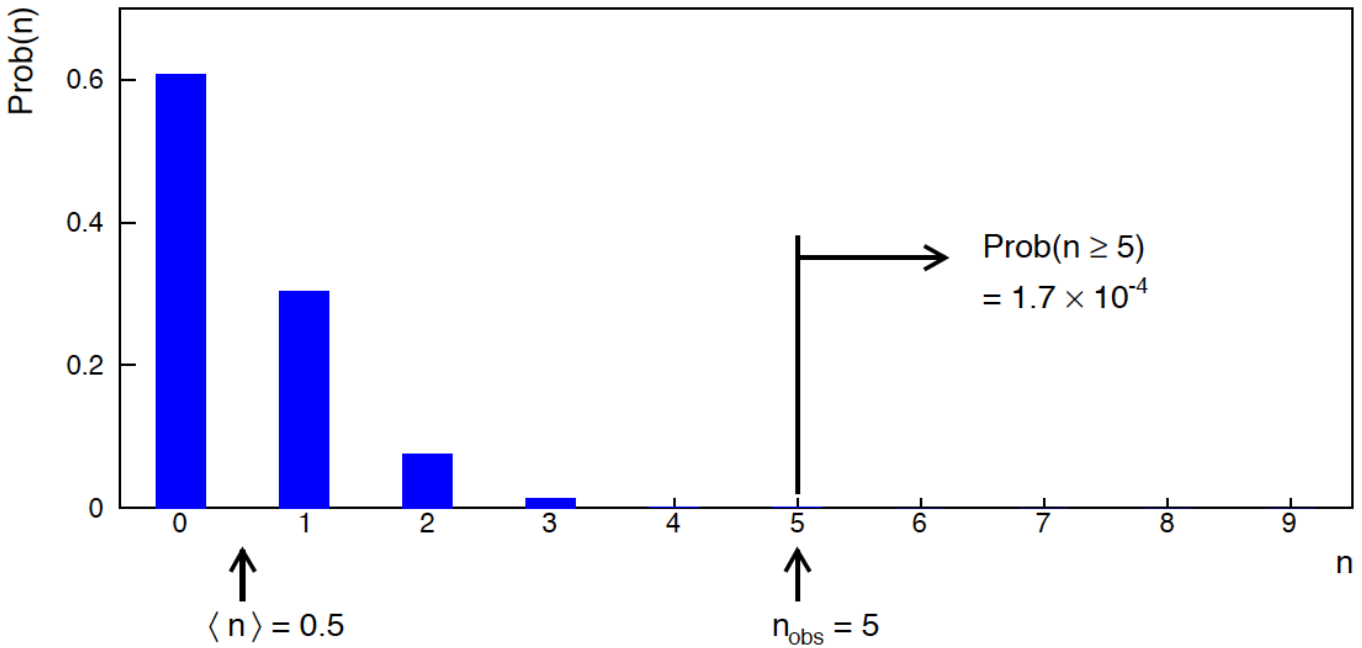
Suppose $b = 0.5$ (known), and we observe $n_{\text{obs}} = 5$.

Should we claim evidence for a new discovery?

Give p-value for hypothesis $s = 0$:

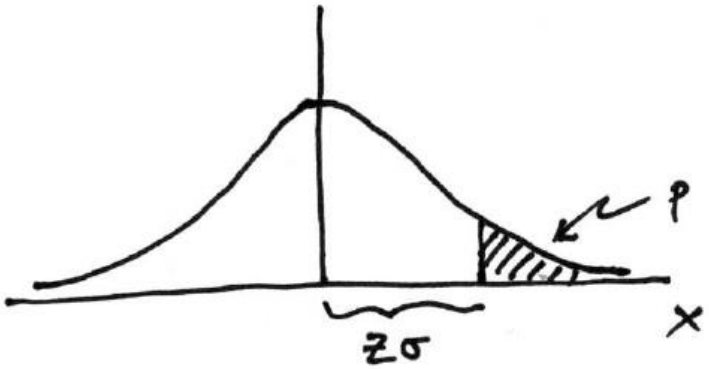
$$p\text{-value} = P(n \geq 5; b = 0.5, s = 0)$$

$$= 1.7 \times 10^{-4} \neq P(s = 0)!$$



Significance from p-value

Often define significance Z as the number of standard deviations that a Gaussian variable would fluctuate in one direction to give the same p-value.



$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1 - \Phi(Z)$$

$$Z = \Phi^{-1}(1-p)$$

```
in ROOT:
p = 1 - TMath::Freq(Z)
Z = TMath::NormQuantile(1-p)
```

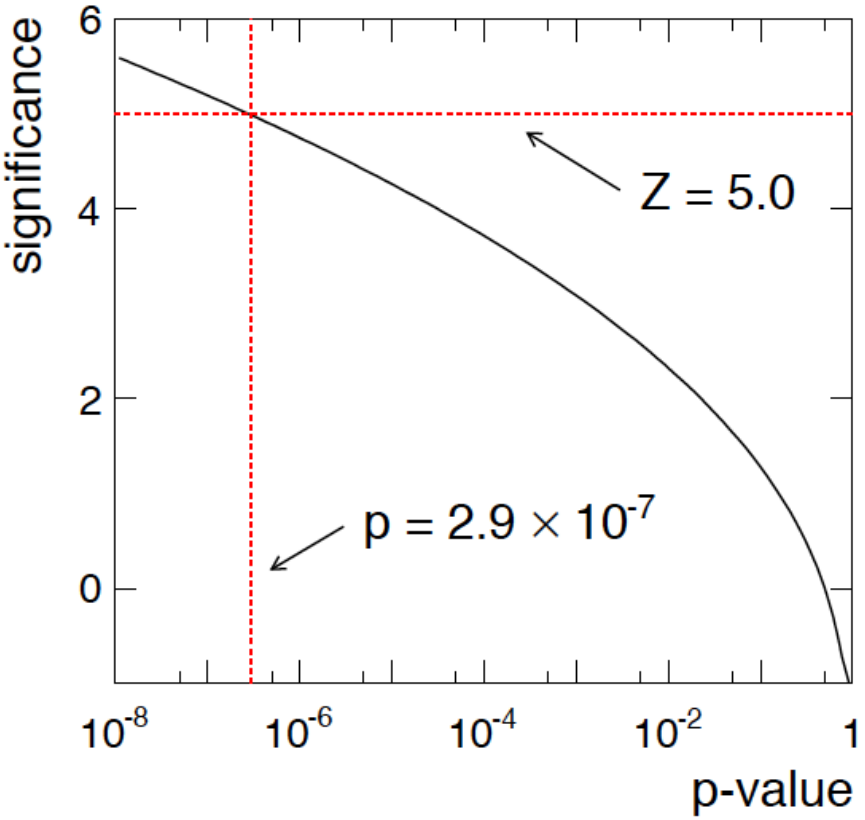
```
in python (scipy.stats):
p = 1 - norm.cdf(Z) = norm.sf(Z)
Z = norm.ppf(1-p)
```

Result Z is a "number of sigmas". Note this does not mean that the original data was Gaussian distributed.

Poisson counting experiment: discovery significance

Equivalent significance for $p = 1.7 \times 10^{-4}$: $Z = \Phi^{-1}(1 - p) = 3.6$

Often claim discovery if $Z > 5$ ($p < 2.9 \times 10^{-7}$, i.e., a "5-sigma effect")



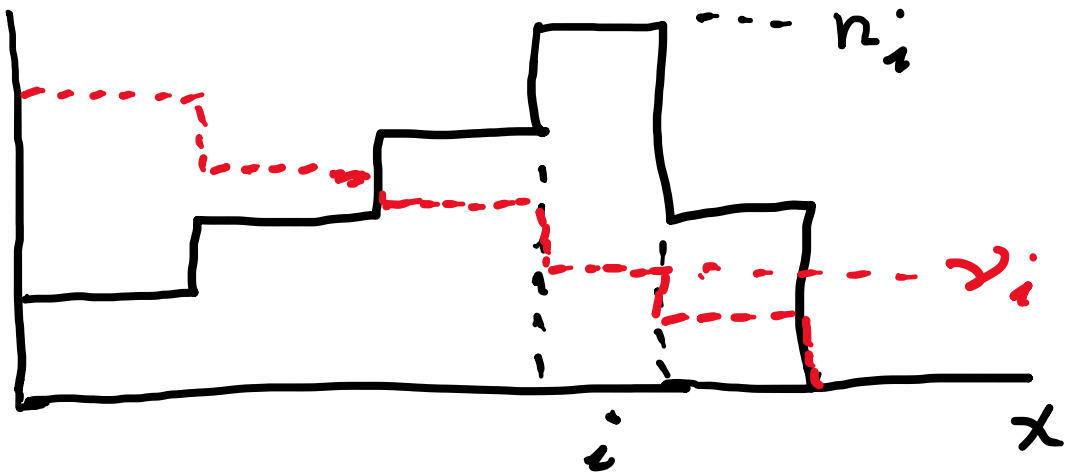
In fact this tradition should be revisited: p-value intended to quantify probability of a signal-like fluctuation assuming background only; not intended to cover, e.g., hidden systematics, plausibility signal model, compatibility of data with signal, "look-elsewhere effect" (~multiple testing), etc.

Statistical Data Analysis
Lecture 6-3

- Test based on histogram
- Pearson's chi-squared

Test using histogram of data

Suppose the data are a histogram $\boldsymbol{n} = (n_1,\ldots,n_N)$ of values and a hypothesis predicts mean values $\boldsymbol{\nu} = E[\boldsymbol{n}] = (\nu_1,\ldots,\nu_N)$.



Modeling the data

Consider e.g. the following hypotheses:

independent, treat as continuous $n_i \sim \text{Gauss}(\nu_i, \sigma_i)$

$$p(\mathbf{n}|\boldsymbol{\nu}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(n_i-\nu_i)^2/2\sigma_i^2}$$

independent $n_i \sim \text{Poisson}(\nu_i)$

$$P(\mathbf{n}|\boldsymbol{\nu}) = \prod_{i=1}^{N} \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}$$

$\boldsymbol{n} \sim \text{Multinomial}(n_{\text{tot}}, \boldsymbol{p})$, $\quad n_{\text{tot}} = \Sigma_i n_i$, $\quad \boldsymbol{p} = \boldsymbol{\nu} / n_{\text{tot}}$

$$P(\mathbf{n}|\boldsymbol{\nu}) = \frac{n_{\text{tot}}!}{n_1! \cdots n_N!} p_1^{n_1} \cdots p_N^{n_N}$$

Pearson's χ^2 statistic

We can take as the test statistic

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\sigma_i^2} \,, \text{ where } \sigma_i^2 = V[n_i] \,.$$

(Pearson's χ^2 statistic)

χ^2 = sum of squares of the deviations of the ith measurement from the ith predicted mean, using σ_i as the 'yardstick' for the comparison.

$\chi^2 \geq \chi^2_{\text{obs}}$ defines the region of "equal or lesser compatibility" for purposes of computing a p-value.

$$p = \int_{\chi^2_{\text{obs}}}^{\infty} f(\chi^2)\, d\chi^2$$

need this pdf

Distribution of Pearson's χ^2 statistic

If $n_i \sim \text{Gauss}(\mu_i, \sigma_i^2)$, then Pearson's χ^2 will follow the chi-square pdf (here write $\chi^2 = z$) for N degrees of freedom:

$$f_{\chi^2}(z; N) = \frac{1}{2^{N/2}\Gamma(N/2)} z^{N/2-1} e^{-z/2}$$

For proof using characteristic functions (Fourier transforms) see e.g. SDA Sec. 10.2.

If the $n_i \sim \text{Poisson}(\mu_i)$ then $V[n_i] = \nu_i$ and so $\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i}$.

If $\nu_i >> 1$ (in practice OK for $\nu_i >$ half dozen) then the Poisson dist. becomes Gaussian (see SDA Sec. 10.2) and therefore Pearson's χ^2 statistic here as well follows the chi-square pdf.

This is called the "large-sample" or "asymptotic" limit.

Pearson's χ^2 with multinomial data

If $n_{\text{tot}} = \sum_i^N n_i$ is fixed, then we can model the histogram using

$\boldsymbol{n} \sim$ Multinomial($\boldsymbol{p}$, n_{tot}) with $p_i = \nu_i / n_{\text{tot}}$.

In this case we can take Pearson's χ^2 statistic to be

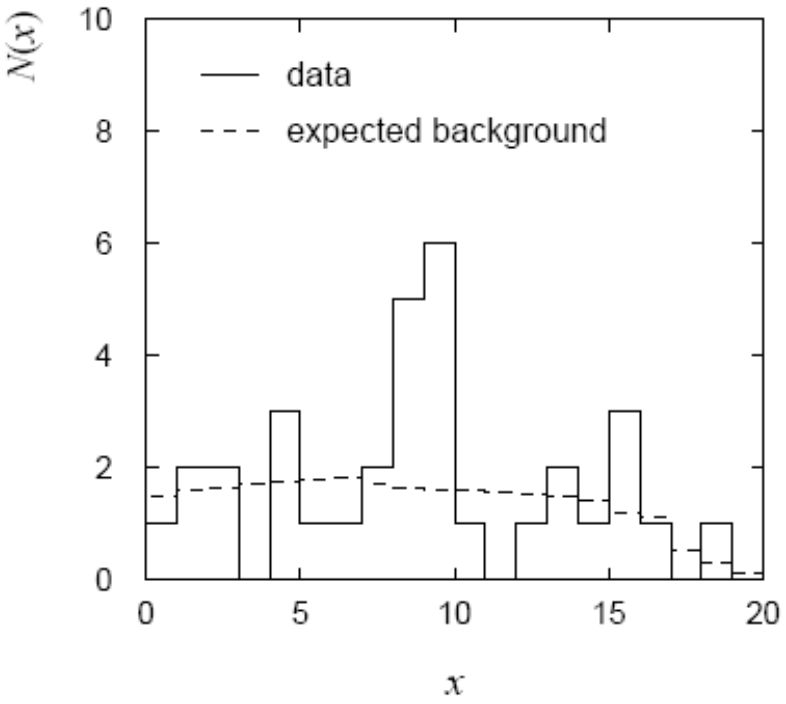
$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - p_i n_{\text{tot}})^2}{p_i n_{\text{tot}}}$$

Note here the denominator is not the variance $V[n_i] = n_{\text{tot}}\, p_i\, (1–p_i)$, and also since the $n_i \sim$ multinomial they are not independent.

But with this definition, if all $p_i\, n_{\text{tot}} >> 1$ (the "large sample limit") one can show the statistic will follow the chi-square pdf for N–1 degrees of freedom.

Example of a $\not{p}$ test

Suppose we have the data below (solid) and prediction (dashed) of a "background" hypothesis, model n_i ~Poisson(ν_i).



← This gives

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i} = 29.8$$

for $N = 20$ dof.

Now need to find p-value, but... many bins have few (or no) entries, so here we do not expect $\not{p}$ to follow the chi-square pdf.

Using MC to find distribution of χ^2 statistic

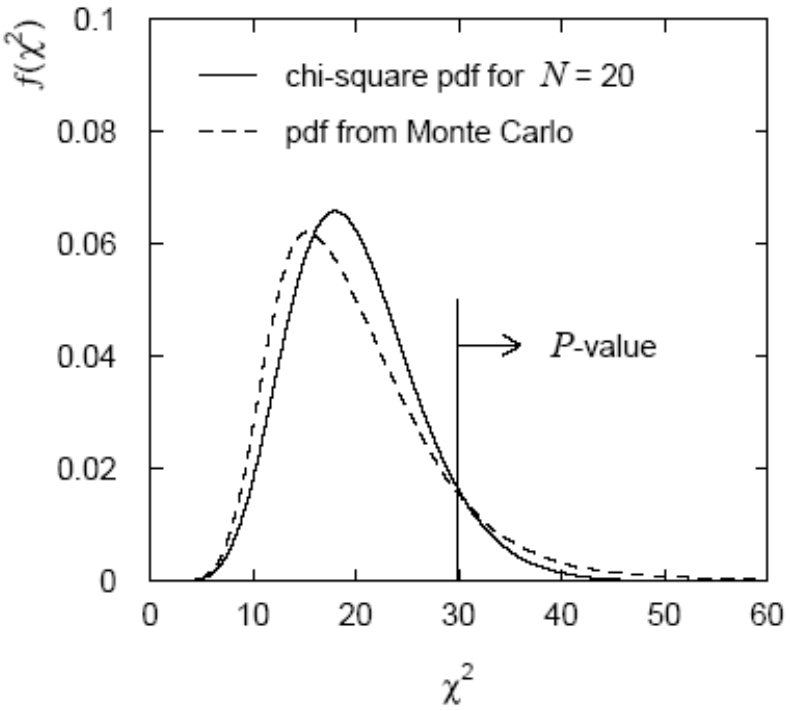
If the distribution of the χ^2 statistic is not expected to be well approximated by the asymptotic chi-square distribution, we can still use it but need some other way to find its pdf.

To find its sampling distribution, simulate the data with a Monte Carlo program, i.e., generate $n_i \sim \text{Poisson}(\nu_i)$ for $i = 1,\ldots,N$

Here data sample simulated 10^6 times. The fraction of times we find $\chi^2 > 29.8$ gives the p-value:

$p = 0.11$

If we had used the chi-square pdf we would find $p = 0.073$.



The 'χ^2 per degree of freedom'

Recall that for the chi-square pdf for N degrees of freedom,

$$E[z] = N\,, \quad V[z] = 2N\,.$$

This makes sense: if the hypothesized v_i are right, the rms deviation of n_i from v_i is σ_i, so each term in the sum contributes ~1.

One often sees χ^2/N reported as a measure of goodness-of-fit. But... better to give χ^2 and N separately. Consider, e.g.,

$$\chi^2 = 15,\ N = 10 \rightarrow p - \text{value} = 0.13\,,$$
$$\chi^2 = 150,\ N = 100 \rightarrow p - \text{value} = 9.0 \times 10^{-4}\,.$$

i.e. for N large, even a χ^2 per dof only a bit greater than one can imply a small p-value, i.e., poor goodness-of-fit.

Statistical Data Analysis
Lecture 6-4

- Introduction to (frequentist) parameter estimation
- The method of Maximum Likelihood
- MLE for exponential distribution

Parameter estimation

The parameters of a pdf are any constants that characterize it,

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}$$

r.v. parameter

i.e., θ indexes a set of hypotheses.

Suppose we have a sample of observed values: $\boldsymbol{x} = (x_1, \ldots, x_n)$

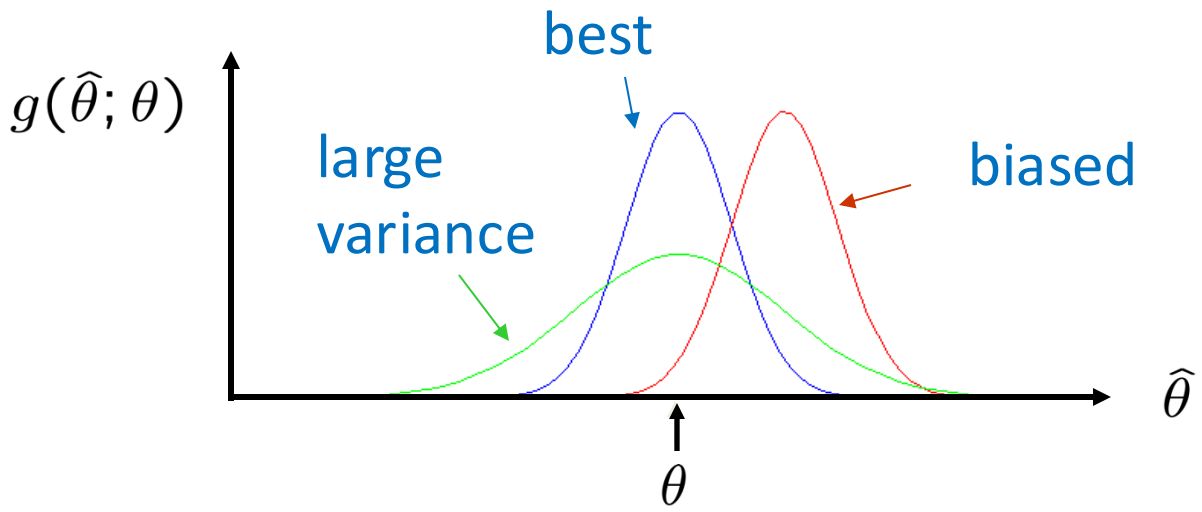
We want to find some function of the data to estimate the parameter(s):

$\widehat{\theta}(\vec{x})$ ← estimator written with a hat

Sometimes we say 'estimator' for the function of $x_1, \ldots, x_n$; 'estimate' for the value of the estimator with a particular data set.

Properties of estimators

If we were to repeat the entire measurement, the estimates from each would follow a pdf:



We want small (or zero) bias (systematic error): $b = E[\hat{\theta}] - \theta$

→ average of repeated measurements should tend to true value.

And we want a small variance (statistical error): $V[\hat{\theta}]$

→ small bias & variance are in general conflicting criteria

An estimator for the mean (expectation value)

Parameter: $\mu = E[x] = \langle x \rangle = \int_{-\infty}^{\infty} x f(x)\, dx$

Suppose we have a sample of n independent values $x_1,\ldots,x_n$.

Estimator: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \equiv \overline{x}$ ('sample mean')

We find: $b = E[\hat{\mu}] - \mu = 0$

$$V[\hat{\mu}] = \frac{\sigma^2}{n} \qquad \left(\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}} \right)$$

An estimator for the variance

Parameter: $\sigma^2 = V[x] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx$

Estimator: $\widehat{\sigma^2} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \equiv s^2$ ('sample variance')

We find:

$b = E[\widehat{\sigma^2}] - \sigma^2 = 0$ (factor of *n*-1 makes this so)

$V[\widehat{\sigma^2}] = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\mu_2\right)$, where

$\mu_k = \int (x - \mu)^k f(x)\, dx$

The likelihood function

Suppose the entire result of an experiment (set of measurements) is a collection of numbers $\boldsymbol{x}$, and suppose the joint pdf for the data $\boldsymbol{x}$ is a function that depends on a set of parameters $\boldsymbol{\theta}$:

$$P(\mathbf{x}|\boldsymbol{\theta})$$

Now evaluate this function with the data obtained and regard it as a function of the parameter(s). This is the likelihood function:

$$L(\boldsymbol{\theta}) = P(\mathbf{x}|\boldsymbol{\theta})$$

($\boldsymbol{x}$ constant)

The likelihood function for i.i.d.*. data

* i.i.d. = independent and identically distributed

Consider n independent observations of x: x_1, ..., x_n, where x follows $f(x; \theta)$. The joint pdf for the whole data sample is:

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

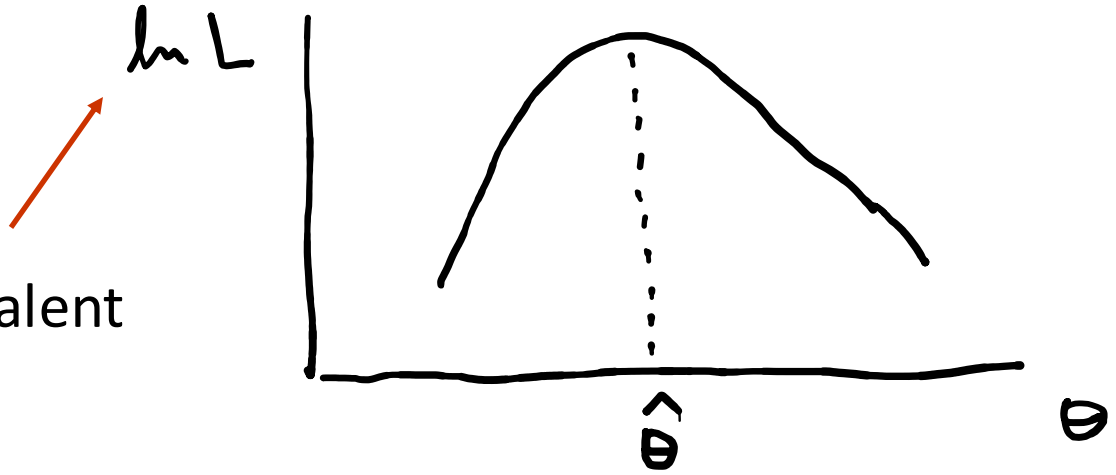
In this case the likelihood function is

$$L(\vec{\theta}) = \prod_{i=1}^{n} f(x_i; \vec{\theta}) \qquad (x_i \text{ constant})$$

Maximum Likelihood Estimators (MLEs)

We *define* the maximum likelihood estimators or MLEs to be the parameter values for which the likelihood is maximum.

Maximizing L equivalent to maximizing $\log L$



Could have multiple maxima (take highest).

MLEs not guaranteed to have any 'optimal' properties, (but in practice they're very good).

MLE example: parameter of exponential pdf

Consider exponential pdf, $f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$

and suppose we have i.i.d. data, $t_1, \ldots, t_n$

The likelihood function is $L(\tau) = \prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau}$

The value of τ for which $L(\tau)$ is maximum also gives the maximum value of its logarithm (the log-likelihood function):

$$\ln L(\tau) = \sum_{i=1}^{n} \ln f(t_i;\tau) = \sum_{i=1}^{n} \left(\ln \frac{1}{\tau} - \frac{t_i}{\tau} \right)$$

MLE example: parameter of exponential pdf (2)

Find its maximum by setting

$$\frac{\partial \ln L}{\partial \tau} = \sum_{i=1}^{n} \left(-\frac{1}{\tau} + \frac{t_i}{\tau^2} \right) = 0$$

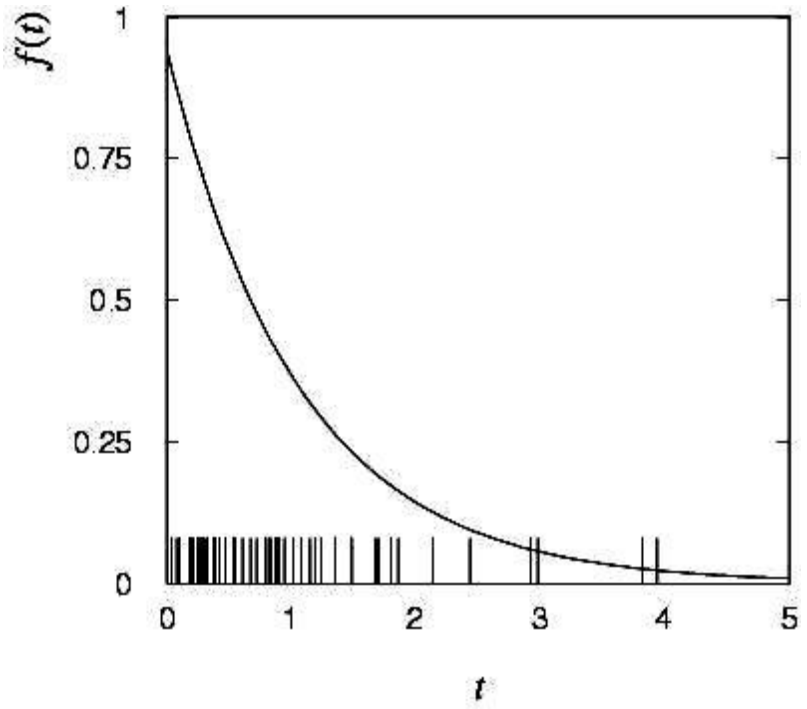
$$\rightarrow \quad \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$$

Monte Carlo test:

generate 50 values using $\tau = 1$:

We find the ML estimate:

$$\hat{\tau} = 1.062$$



MLE example: parameter of exponential pdf (3)

For the exponential distribution one has for mean, variance:

$$E[t] = \int_0^\infty t \frac{1}{\tau} e^{-t/\tau}\, dt = \tau$$

$$V[t] = \int_0^\infty (t-\tau)^2 \frac{1}{\tau} e^{-t/\tau}\, dt = \tau^2$$

For the MLE $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$ we therefore find

$$E[\hat{\tau}] = E\left[\frac{1}{n}\sum_{i=1}^{n} t_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[t_i] = \tau \quad \longrightarrow \quad b = E[\hat{\tau}] - \tau = 0$$

$$V[\hat{\tau}] = V\left[\frac{1}{n}\sum_{i=1}^{n} t_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V[t_i] = \frac{\tau^2}{n} \quad \longrightarrow \quad \sigma_{\hat{\tau}} = \frac{\tau}{\sqrt{n}}$$

Extra slides

Software for Machine Learning

We will practice ML with the Python package scikit-learn

`scikit-learn.org` ← software, docs, example code

scikit-learn built on NumPy, SciPy and matplotlib, so you need

```
import scipy as sp
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
```

and then you import the needed classifier(s), e.g.,

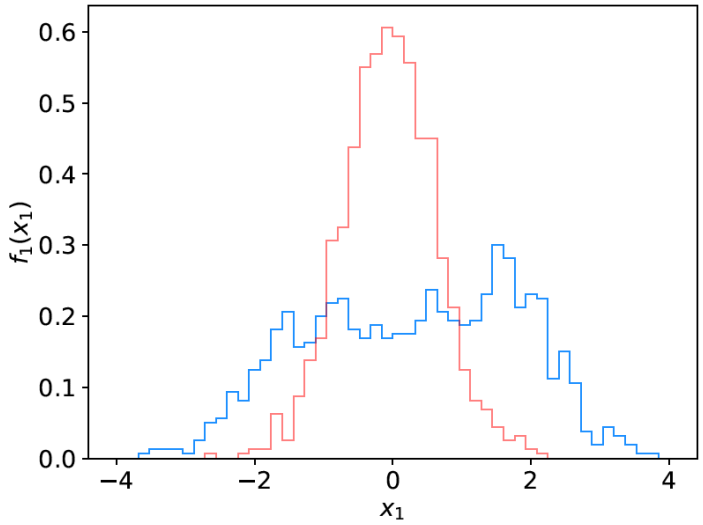
```
from sklearn.neural_network import MLPClassifier
```

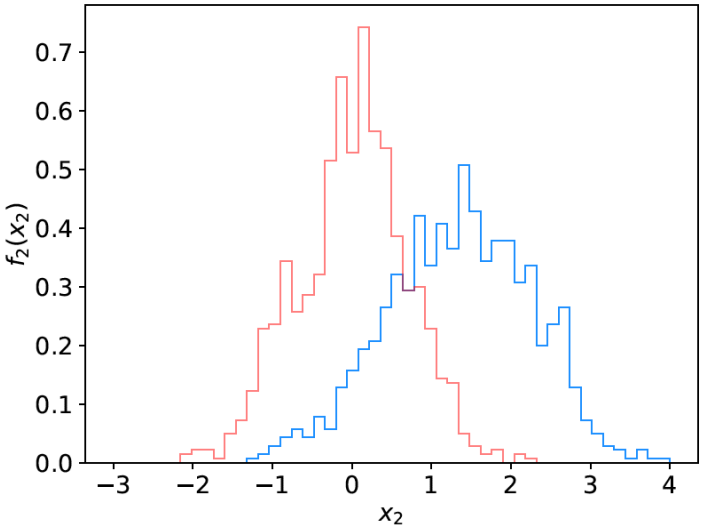
For a list of the various classifiers in scikit-learn see the docs on scikit-learn.org, also a very useful sample program:

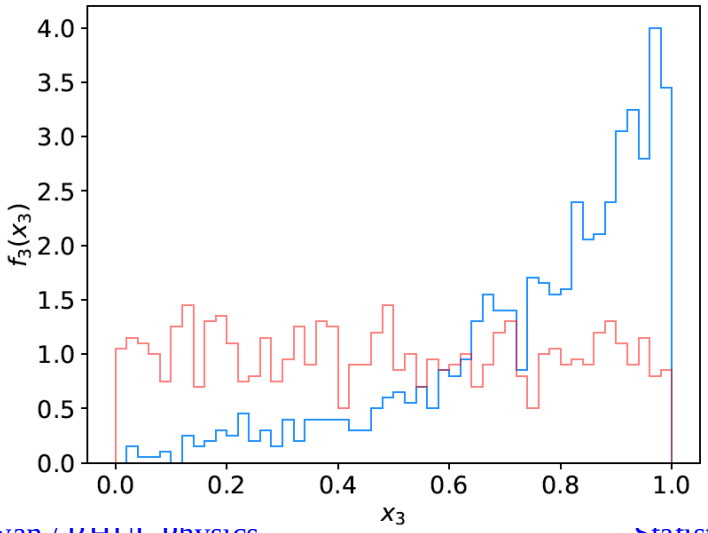
`http://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html`

Example: the data

We will do an example with data corresponding to events of two types: signal ($y = 1$, blue) and background ($y = 0$, red).







Each event is characterised by 3 quantities: $\boldsymbol{x} = (x_1, x_2, x_3)$.

Components are correlated.

Suppose we have 1000 events each of signal and background.

Reading in the data

scikit-learn wants the data in the form of numpy arrays:

```
#  read the data in from files,
# assign target values 1 for signal, 0 for background
sigData = np.loadtxt('signal.txt')
nSig = sigData.shape[0]
sigTargets = np.ones(nSig)
bkgData = np.loadtxt('background.txt')
nBkg = bkgData.shape[0]
bkgTargets = np.zeros(nBkg)

# concatenate arrays into data X and targets y
# split into two parts:  use one for training, the other for testing
X = np.concatenate((sigData,bkgData),0)
y = np.concatenate((sigTargets, bkgTargets))

# split data into training and testing samples
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.5,
                                                    random_state=1)
```

Create, train, evaluate the classifier

Create an instance of the MLP (multilayer perceptron) class and "train", i.e., adjust the values of the weights to minimise the loss function.

Here we request 3 hidden layers with 10 nodes each:

```
# create classifier object and train
clf = MLPClassifier(hidden_layer_sizes=(10,10,10), activation='tanh',
                    max_iter=2000, random_state=6)
clf.fit(X_train, y_train)
```

Use test data to see what fraction of events are correctly classified (default takes threshold of 0.5 for decision function)

```
# evaluate its accuracy (= 1 – error rate) using the test data
y_pred = clf.predict(X_test)
print(metrics.accuracy_score(y_test, y_pred))
```

Evaluating the decision function

So now for any point (x_1, x, x_3) in the feature space, we can evaluate the decision:

```
#  Test evaluation of decision function for a specific point in feature space
xt = np.array([0.37, 2.46, 0.42]).reshape((1,-1))
#t = clf.decision_function(xt)[0]          # not available for MLP
t = clf.predict_proba(xt)[0, 1]           # for MLP use this instead
```

Usually we have an array of points in $\boldsymbol{x}$-space, so we can get an array of probabilities:

```
t = clf.predict_proba(X_test)[:, 1]       # returns prob to be of type y=1
```

Can get this separately for the signal and background events and make histograms (see sample code).

Note for most other classifiers, the decision function is called decision_function – use this instead of predict_proba.

On defining a p-value

Earlier it was argued that the region of "equal or lesser compatibility" with H had greater compatibility with the predictions of some alternative hypothesis.

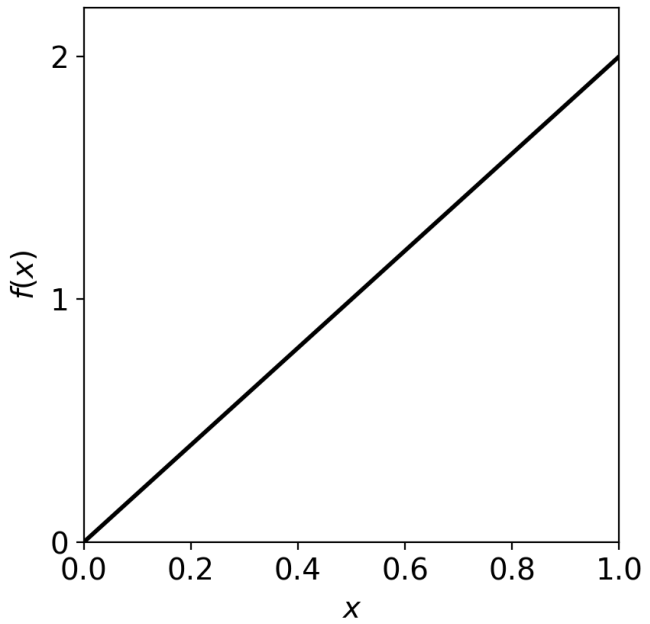
But shouldn't it be possible to identify such a region by using the pdf $f(x|H)$?

In general, no.

Consider cubic crystal grains produced by a process H that have a size distribution

$$f(x|H) = 2x, \quad 0 < x \leq 1$$

Observe grain of uncertain origin,
measure x,
find p-value of H.



If we observe a value x_{obs}, naively we could regard $x \leq x_{\text{obs}}$ as constituting equal or less agreement with the predictions of $f(x|H)$.

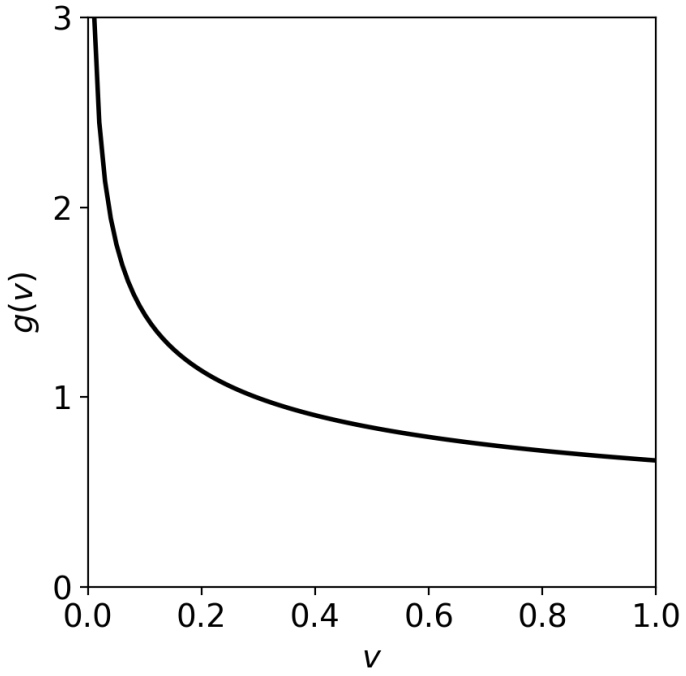
On defining a p-value (2)

But suppose we took the volume $v = x^3$ of the cube to represent its size. The volume distribution is

$$g(v|H) = f(x(v)|H) \left| \frac{dx}{dv} \right|$$

$$= \frac{2}{3} v^{-1/3}$$

$$0 < v \leq 1$$



So now it appears that smaller sizes are more compatible with H.

Conclusion: deciding what region of data space constitutes greater or lesser compatibility with *H* cannot be done by looking at the data distribution alone; it requires that one consider an alternative *H*'.

nature International weekly journal of science

NATURE | RESEARCH HIGHLIGHTS: SOCIAL SELECTION

Psychology journal bans *P* values

Test for reliability of results 'too easy to pass', say editors.

Chris Woolston

26 February 2015 | Clarified: 09 March 2015

A controversial statistical test has finally met its end, at least in one journal. Earlier this month, the editors of *Basic and Applied Social Psychology* (*BASP*) announced that the journal would no longer publish papers containing *P* values because the statistics were too often used to support lower-quality research [1].

https://imgs.xkcd.com/comics/significant.png