Lecture 1

1. **Probability**
2. **Random variables, probability densities, etc.**
3. **Brief catalogue of probability densities**
4. **The Monte Carlo method**

Lecture 2

1. **Statistical tests**
2. **Fisher discriminants, neural networks, etc.**
3. **Goodness-of-fit tests**
4. **The significance of a signal**
5. **Introduction to parameter estimation**

Lecture 3

1. **The method of maximum likelihood (ML)**
2. **Variance of ML estimators**
3. **The method of least squares (LS)**
4. **Interval estimation, setting limits**

Suppose the result of a measurement is $\vec{x} = (x_1, \ldots, x_n)$

e.g. events from e^+e^- collisions; for each event measure

$\quad$ $x_1 =$ number of charged particles produced

$\quad$ $x_2 =$ mean $p_\perp$ of particles

$\quad$ $x_3 =$ number of 'jets' (according to some algorithm)

$\quad$ $x_4 = \ldots$

$\vec{x}$ follows some joint pdf in an n-dimensional space, which depends on the type of event produced, i.e. $e^+e^- \to q\bar{q}$, $e^+e^- \to WW$, etc.

That is, the joint pdf $f(\vec{x})$ is specified by a certain

<p align="center">HYPOTHESIS</p>

i.e. predicted probability densities $f(\vec{x}|H_0)$, $f(\vec{x}|H_1)$, etc.

(Note sloppy but traditional notation: usually H_0, H_1, $\ldots$ not r.v.s.)

Simple hypothesis: $f(\vec{x})$ completely specified,

Composite hypothesis: form of $f(\vec{x}; \theta)$ given, parameter θ unknown.

Usually awkward to work with multidimensional $\vec{x}$,

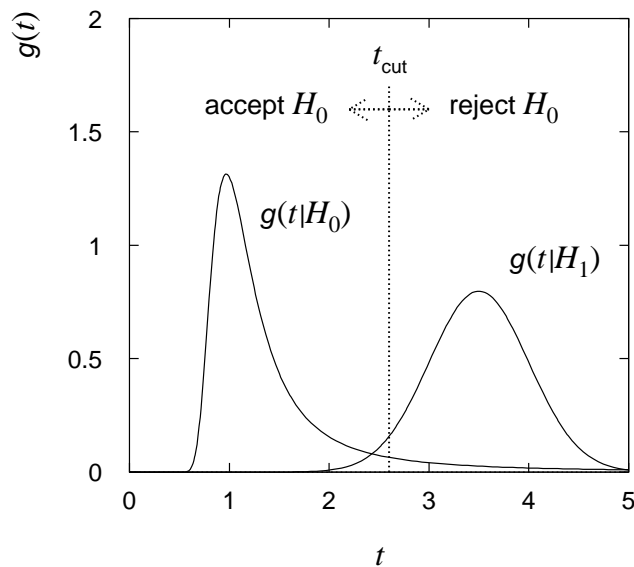$\Rightarrow$ construct test statistic of lower dimension (e.g. scalar), $t(\vec{x})$:

$\quad$ compactify data,

$\quad$ try not to lose ability to discriminate bewteen hypotheses.

The statistic t then has pdfs $g(t|H_0)$, $g(t|H_1)$, $\ldots$

Consider a test statistic t following $g(t|H_0)$, $g(t|H_1)$, ...

Define a critical region where t is not likely to occur if H_0 is true,

e.g. for the case above, $t \geq t_{\mathrm{cut}}$.

If observed value t_{obs} is in critical region, reject H_0, otherwise 'accept'.

Probability to reject H_0 if it is true (error of 1st kind):

$$\alpha = \int_{t_{\mathrm{cut}}}^{\infty} g(t|H_0)\, dt \qquad (\text{significance level})$$

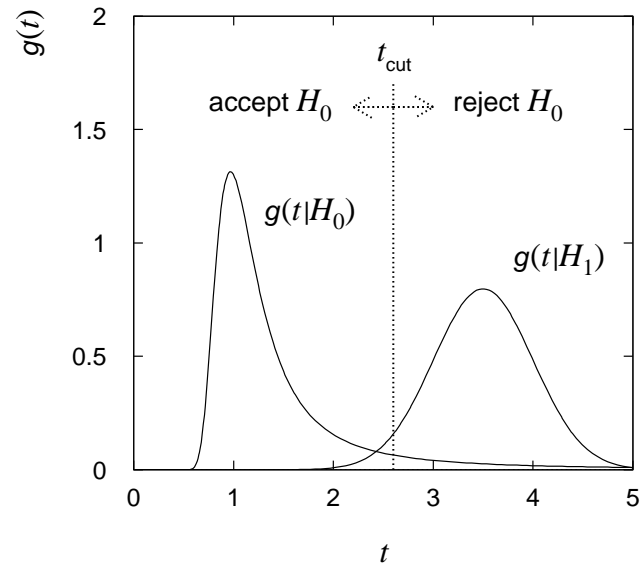Probability to accept H_0 if H_1 is true (error of 2nd kind):

$$\beta = \int_{-\infty}^{t_{\mathrm{cut}}} g(t|H_1)\, dt \qquad (1 - \beta = \text{power})$$

An example with particle selection

Suppose we obtain n energy loss measurements for a particle in a drift chamber, construct $t =$ truncated mean of the measurements, and suppose we know the particles are either electrons or pions:

$H_0 =$ electron (signal)

$H_1 =$ pion (background)

Select electrons by requiring $t < t_{\mathrm{cut}}$. The selection efficiencies are:

$$\varepsilon_{\mathrm{e}} = \int_{-\infty}^{t_{\mathrm{cut}}} g(t|\mathrm{e})dt = 1 - \alpha$$

$$\varepsilon_{\pi} = \int_{-\infty}^{t_{\mathrm{cut}}} g(t|\pi)dt = \beta$$

Loose cut: most e accepted, lots of π background

Tight cut: low signal efficiency, pure sample

Fractions of e, π may be unknown; t follows

$$f(t; a_{\mathrm{e}}) = a_{\mathrm{e}}g(t|\mathrm{e}) + (1 - a_{\mathrm{e}})g(t|\pi)$$

$\rightarrow$ estimate a_{e} (for now assume a_{e}, $a_{\pi} = 1 - a_{\mathrm{e}}$ known)

Purity of selected sample

For a measured value t, what is the probability to be e/π?

$$h(\mathrm{e}|t) = \frac{a_\mathrm{e}\, g(t|\mathrm{e})}{a_\mathrm{e}\, g(t|\mathrm{e}) \,+\, a_\pi\, g(t|\pi)}$$

(Bayes' theorem)

$$h(\pi|t) = \frac{a_\pi\, g(t|\pi)}{a_\mathrm{e}\, g(t|\mathrm{e}) \,+\, a_\pi\, g(t|\pi)}$$

Bayesian: degree of belief that this particle is e or π

Frequentist: fraction of particles at given t which are e/π

$\rightarrow$ here both approaches make sense

Often want purity of selected sample:

$$p_\mathrm{e} = P(\,\mathrm{e}\,|t < t_\mathrm{cut})$$

$$= \frac{\text{number of electrons with } t < t_\mathrm{cut}}{\text{number of all particles with } t < t_\mathrm{cut}}$$

$$= \frac{\int_{-\infty}^{t_\mathrm{cut}} a_\mathrm{e} g(t|\mathrm{e}) dt}{\int_{-\infty}^{t_\mathrm{cut}} (a_\mathrm{e} g(t|\mathrm{e}) + (1-a_\mathrm{e}) g(t|\pi)) dt}$$

$$= \frac{\int_{-\infty}^{t_\mathrm{cut}} h(\mathrm{e}|t)\, f(t)\, dt}{\int_{-\infty}^{t_\mathrm{cut}} f(t)\, dt}$$

$$= \text{electron probability averaged over interval } (-\infty, t_\mathrm{cut}]$$

The Neyman–Pearson lemma

Consider a multidimensional test statistic $\vec{t} = (t_1, \ldots, t_m)$;
hypotheses H_0 ('signal') and H_1 ('background').
What is the optimal choice of the critical region (i.e. cuts)?

The Neyman–Pearson lemma states: to get the highest purity for
a given efficiency, (i.e. highest power for a given significance level),
choose the acceptance region such that

$$\frac{g(\vec{t}|H_0)}{g(\vec{t}|H_1)} > c,$$

where $c = $ constant which determines the efficiency.

(For a proof see Brandt Chapter 8.) Value of c left open; choose
this depending on what efficiency you want.

Equivalently, the optimal scalar test statistic is

$$r = \frac{g(\vec{t}|H_0)}{g(\vec{t}|H_1)},$$

called the likelihood ratio for simple hypotheses H_0 and H_1.
Requiring $r > c$ gives maximum purity for a given efficiency.
N.B. any monotonic function of r is just as good.

Constructing a test statistic

Example: $H_0 = e^+e^- \rightarrow WW \rightarrow$ hadrons (usually four jets)

$$H_1 = e^+e^- \rightarrow q\overline{q} \rightarrow \text{hadrons} \quad \text{(usually two jets)}$$

For each event measure $\vec{x} = (x_1, \ldots, x_n)$.

According to Neyman–Pearson, to select WWs we should cut on

$$t(\vec{x}) = \frac{f(\vec{x}|H_0)}{f(\vec{x}|H_1)},$$

but we need to know $f(\vec{x}|H_0)$ and $f(\vec{x}|H_1)$.

In practice, get these from Monte Carlo event generator:

Generate events, for each, obtain $\vec{x}$ and enter into n-dimensional histogram. If e.g. M bins per component, total number of cells in $\vec{x}$-space $= M^n$

Approximate $f(\vec{x}|H)$ by probability to be in corresponding cell, i.e. determine M^n parameters. But n is potentially large!

$\Rightarrow$ prohibitively large number of cells to populate with MC data.

Compromise solution:

Make Ansatz for form of $t(\vec{x})$ with fewer parameters; determine the parameters (e.g. using MC) to give best discrimination between H_0 and H_1.

Linear test statistic

Ansatz:
$$t(\vec{x}) = \sum_{i=1}^{n} a_i x_i = \vec{a}^T \vec{x}$$

A choice of $\vec{a}$ gives certain pdfs $g(t|H_0)$, $g(t|H_1)$.

Choose the a_i to maximize 'separation' between $g(t|H_0)$, $g(t|H_1)$.

$\longrightarrow$ Must define 'separation'.

We have the expectation values and covariances,

$$(\mu_k)_i = \int x_i f(\vec{x}|H_k)\, d\vec{x},$$

$$(V_k)_{ij} = \int (x - \mu_k)_i\, (x - \mu_k)_j\, f(\vec{x}|H_k)\, d\vec{x},$$

$$k = 0, 1 \quad \text{(hypothesis)},$$

$$i, j = 1, \dots, n \quad \text{(component of } \vec{x}).$$

Similarly for mean and variance of $t(\vec{x})$,

$$\tau_k = \int t(\vec{x}) f(\vec{x}|H_k)\, d\vec{x} = \vec{a}^T \vec{\mu}_k,$$

$$\Sigma_k^2 = \int (t(\vec{x}) - \tau_k)^2\, f(\vec{x}|H_k)\, d\vec{x} = \vec{a}^T V_k\, \vec{a}.$$

We should require:

large $|\tau_0 - \tau_1|$,

small Σ_0^2, Σ_1^2 (pdfs tightly concentrated about their means).

Fisher defines as a measure of separation

$$J(\vec{a}) = \frac{(\tau_0 - \tau_1)^2}{\Sigma_0^2 + \Sigma_1^2}.$$

The numerator of $J(\vec{a})$ is

$$(\tau_0 - \tau_1)^2 = \sum_{i,j=1}^{n} a_i a_j (\mu_0 - \mu_1)_i (\mu_0 - \mu_1)_j$$

$$= \sum_{i,j=1}^{n} a_i a_j B_{ij} = \vec{a}^T B \vec{a}.$$

The denominator is

$$\Sigma_0^2 + \Sigma_1^2 = \sum_{i,j=1}^{n} a_i a_j (V_0 + V_1)_{ij} = \vec{a}^T W \vec{a}.$$

This gives $\quad J(\vec{a}) = \dfrac{\vec{a}^T B \vec{a}}{\vec{a}^T W \vec{a}} = \dfrac{\text{separation between classes}}{\text{separation within classes}}$

Set $\quad \dfrac{\partial J}{\partial a_i} = 0 \quad \Rightarrow \quad \vec{a} \propto W^{-1}(\vec{\mu}_0 - \vec{\mu}_1)$

This defines Fisher's linear discriminant function,

determined up to a scale factor for $\vec{a}$.

R.A. Fisher, *Ann. Eugen.* 7 (1936) 179.

Neural networks (1)

Used in neurobiology, pattern recognition, financial forecasting ...

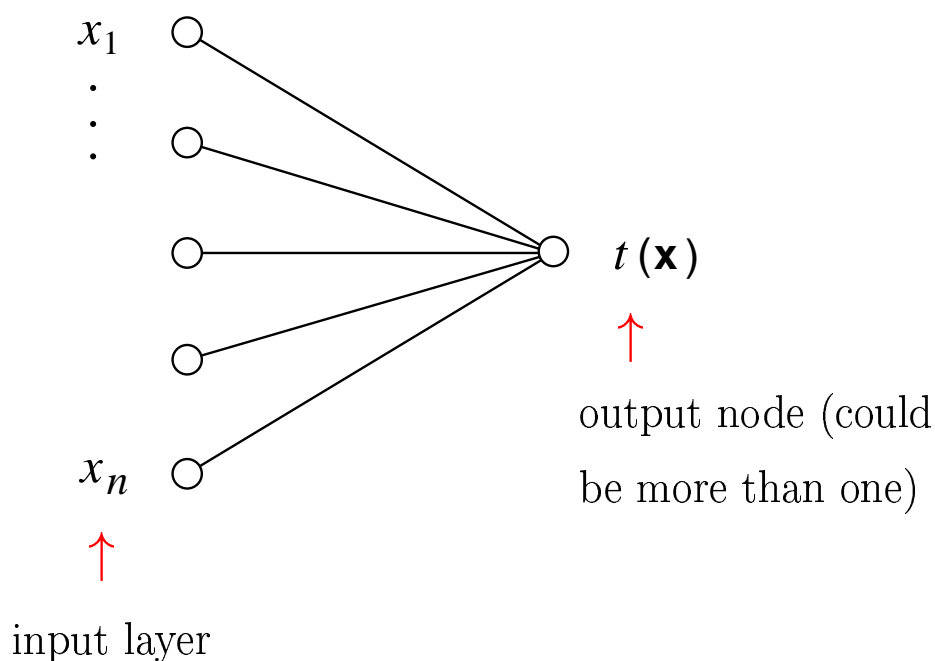here, neural nets are just a type of test statistic.

Suppose we take $t(\vec{x})$ to have the form

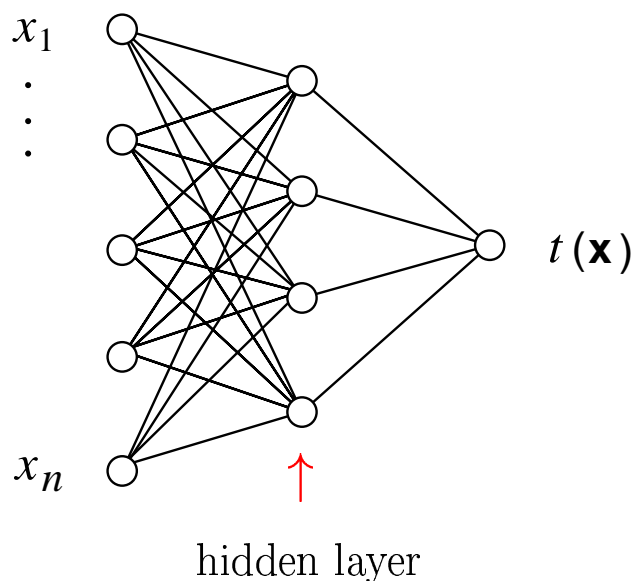$$t(\vec{x}) = s\left(a_0 + \sum_{i=1}^{n} a_i x_i\right)$$

where $s(u) = (1 + e^{-u})^{-1}$ (the 'activation function')

This is the single-layer perceptron.

$s(\cdot)$ is monotic $\Rightarrow$ equivalent to linear $t(\vec{x})$.

x_1 ○
.
.
.
○

○ ────────── ○ $t(\mathbf{x})$

○ ↑

x_n ○ output node (could
be more than one)
↑
input layer

Generalize this to the multilayer perceptron:

x_1

x_n

$t(\mathbf{x})$

hidden layer

The output is defined by $t(\vec{x}) = s\left(a_0 + \sum\limits_{i=1}^{m} a_i h_i(\vec{x})\right)$,

where the h_i are functions of the nodes in the previous layer,

$$h_i(\vec{x}) = s\left(w_{i0} + \sum\limits_{j=1}^{n} w_{ij} x_j\right).$$

$a_i,\ w_{ij}\ =$ weights (connection strengths)

Easy to generalize to arbitrary number of layers.

Feed-forward net: values of a node depend only on earlier layers,

usually only on previous layer $\rightarrow$ 'network architecture'

More nodes $\rightarrow$ neural net gets closer to optimal $t(\vec{x})$,

but more parameters need to be determined.

Parameters usually determined by minimizing an error function,

$$\mathcal{E} = E_0[(t - t^{(0)})^2] + E_1[(t - t^{(1)})^2],$$

where $t^{(0)}$, $t^{(1)}$ are target values, e.g. 0 and 1 for logistic sigmoid, cf. least squares principle with Fisher discriminant.

In practice, replace expectation values by averages of training data from Monte Carlo. (Adjusting parameters = network 'learning'.)

In general this can be tricky; fortunately, programs like JETNET do it for you, e.g. with 'error back-propogation'.

For more information see

L. Lönnblad et al., *Comput. Phys. Commun.* 70 (1992) 167;

C. Peterson, et al., *Comput. Phys. Commun.* **81** (1994) 185;

C.M. Bishop, *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford (1995);

John Hertz, et al., *Introduction to the Theory of Neural Computation*, Addison-Wesley, New York (1991);

B. Müller et al., *Neural Networks: an Introduction*, 2nd edition, Springer, Berlin (1995).
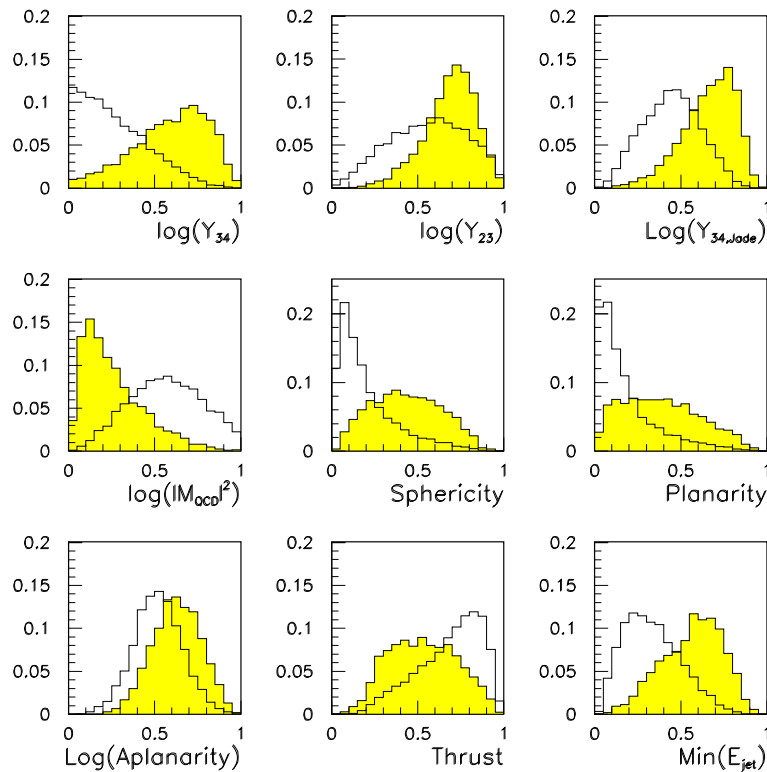
Neural networks (4)

An example with WW event selection
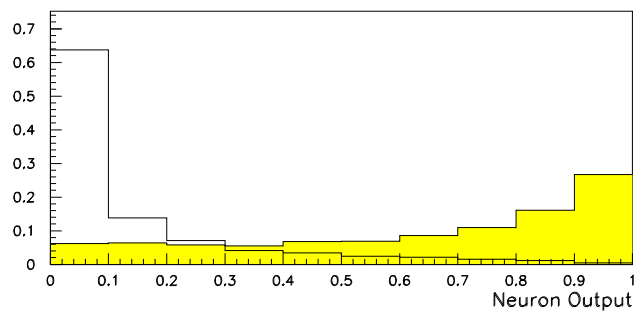
(Garrido, Juste and Martinez, ALEPH 96-144)

The input variables:

Shaded histograms: WW (signal)

Open histograms: $q\overline{q}$ (background)

The neural network output:

Choosing the input variables

Why not use all of the available input variables?

Fewer inputs → fewer parameters to be adjusted,

→ parameters better determined for finite training data.

Some inputs may be highly correlated → drop all but one.

Some inputs may contain little or no discriminating power between the hypotheses → drop them.

NN exploits higher moments of joint pdf $f(\vec{x}|H)$,
but these may not be well modeled in training data.

→ better to have simpler $t(\vec{x})$ where you can 'understand what it's doing'.

Recall that the purpose of the statistical test is usually to select objects for further study; e.g. select WW events, then measure their properties (e.g. particle multiplicity).

⇒ avoid input variables that are correlated with the properties of the selected objects which you want to study. (Not always easy; correlations may not be well known.)
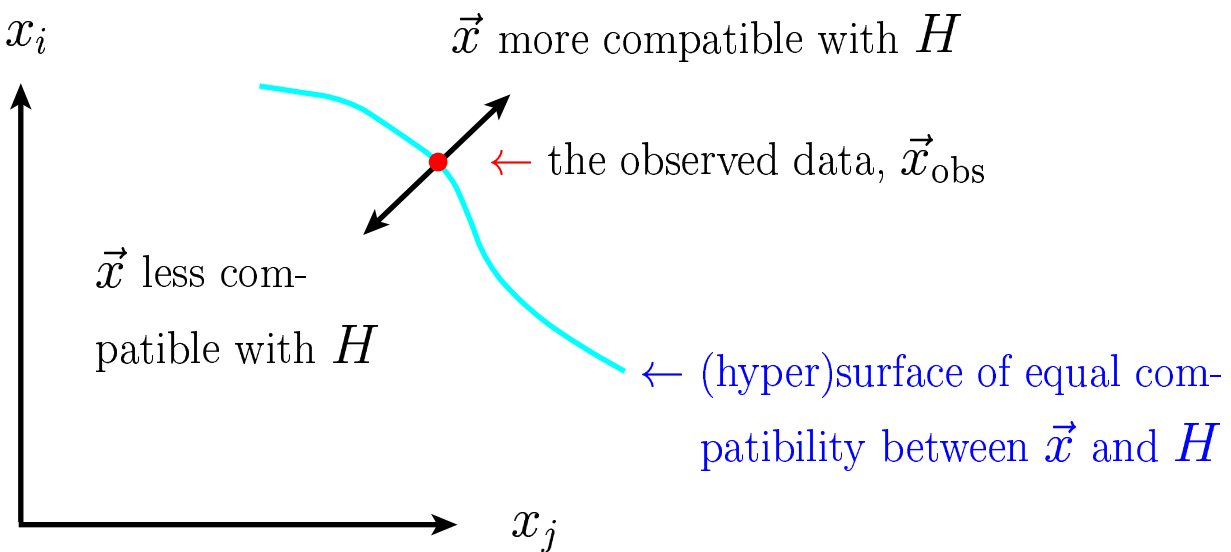
Suppose hypothesis H predicts $f(\vec{x}|H)$ for some vector
of data $\vec{x} = (x_1, \ldots, x_n)$.

We observe a single point in $\vec{x}$-space: $\vec{x}_{\mathrm{obs}}$.

What can we say about the validity of H in light of the data?

$\rightarrow$ Decide what part of $\vec{x}$-space represents less compatibility with
    H than does the observed point $\vec{x}_{\mathrm{obs}}$. (Not unique!)

Usually construct test statistic $t(\vec{x})$ whose value reflects
level compatibility between $\vec{x}$ and H, e.g.

low $t \rightarrow$ data more compatible with H;

high $t \rightarrow$ data less compatible with H.

Since pdf $f(\vec{x}|H)$ known, the pdf $g(t|H)$ can be determined.

P-values

Express 'goodness-of-fit' by giving the P-value (also called observed significance level or confidence level):

$P =$ probability to observe data $\vec{x}$ (or $t(\vec{x})$) having equal or lesser compatibility with H as $\vec{x}_{\text{obs}}$ (or $t(\vec{x}_{\text{obs}})$)

This is not the 'probability' that H is true!

In classical statistics we never talk about $P(H)$.
In Bayesian statistics, treat H as a random variable;
use Bayes' theorem (here symbolically) to obtain

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

where $\pi(H)$ is the prior probability for H; normalize by integrating (or summing) over all possible hypotheses. For now stick with classical approach, i.e. our final answer is the P-value.

N.B. No alternative hypotheses mentioned.

N.B. P-value is a random variable. Previously considered significance level was a constant, specified before the test.

If H true, then (for continuous $\vec{x}$) P is uniform in $[0, 1]$.
If H not true, then pdf of P is (usually) peaked closer to 0.

Probability to observe n_h heads in N coin tosses is:

$$f(n_\mathrm{h}; p_\mathrm{h}, N) = \frac{N!}{n_\mathrm{h}!(N - n_\mathrm{h})!} p_\mathrm{h}^{n_\mathrm{h}} (1 - p_\mathrm{h})^{N - n_\mathrm{h}}$$

Hypothesis H: the coin is fair $(p_\mathrm{h} = p_\mathrm{t} = 0.5)$

Take as goodness-of-fit statistic $t = \left| n_\mathrm{h} - \frac{N}{2} \right|$.

We toss the coin $N = 20$ times and get 17 heads, i.e. $t_\mathrm{obs} = 7$.

Region of t-space with equal or lesser compatibility:

$t \geq 7$

P-value $= P(n_\mathrm{h} = 0, 1, 2, 3, 17, 18, 19$ or $20) = 0.0026$

So does this mean H is false? P-value does not answer this question; it only gives the probability of obtaining such a level of discrepancy (or higher) with H as that observed.

P-value = probability of obtaining such a bizarre result 'by chance'.

A philosophical objection (but not a real problem):

Could have defined experiment to end after at least 3 heads and tails; in ours this happened to occur after 20 tosses. In such an experiment, the P-value is 0.00072!

Pragmatist's solution: 'repetition of experiment' taken to mean repetition with same number of trials per experiment.

<u>The significance of an observed signal</u>

Suppose we observe n events; these can consist of:

n_b events from known processes (background)

n_s events from new processes (signal)

If n_b, n_s are Poisson r.v.s with means $\nu_b, \nu_s, \Rightarrow n = n_s + n_b$
is also Poisson, mean $\nu = \nu_s + \nu_b$ (cf. SDA Chapter 10):

$$P(n; \nu_s, \nu_b) = \frac{(\nu_s + \nu_b)^n}{n!} e^{-(\nu_s + \nu_b)}$$

Suppose $\nu_b = 0.5$ and we observe $n_{obs} = 5$.

Should we claim evidence for a new discovery?

Hypothesis H: $\nu_s = 0$, i.e. only background present.

P-value $= P(n \geq n_{obs})$
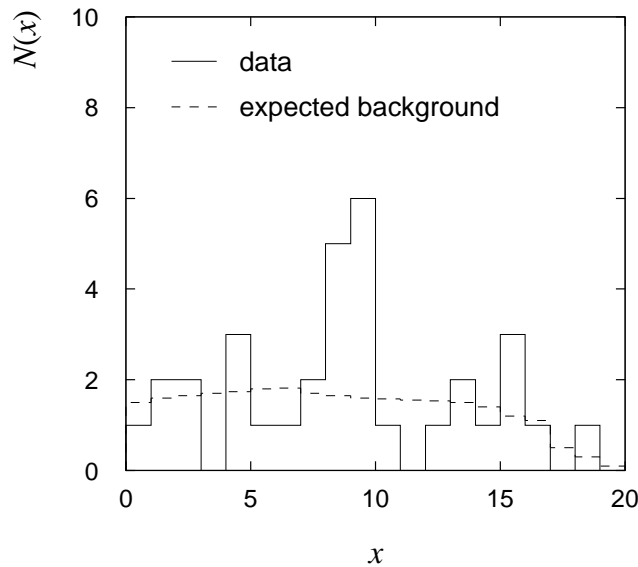
$$= \sum_{n=n_{obs}}^{\infty} P(n; \nu_s = 0, \nu_b)$$

$$= 1 - \sum_{n=0}^{n_{obs}-1} \frac{\nu_b^n}{n!} e^{-\nu_b}$$

$$= 1.7 \times 10^{-4}$$

$$(\neq P(\nu_s = 0)!)$$

The significance of a peak

Suppose in addition to counting events, we measure x for each.

← Histogram of observed and expected data. Each bin is a Poisson variable.

In the 2 bins with peak, 11 entries found, $\nu_b = 3.2$,

$$P(n \geq 11; \nu_b = 3.2; \nu_s = 0) = 5.0 \times 10^{-4}$$

But... did we know where to look for the peak?

→ give $P(n \geq 11)$ in any 2 adjacent bins.

Is the observed width consistent with the expected x resolution?

→ take x window several times expected resolution

How many bins × distributions have we looked at?

→ look at a thousand of them, you'll find a 10^{-3} effect.

Did we adjust the cuts to 'enhance' the peak?

→ freeze cuts, repeat analysis with new data.

How about the bins to the sides of the peak ... (too low!)

Should we publish???

<u>Pearson's χ^2 test</u>

Test statistic for comparing observed data $\vec{n} = (n_1, \ldots, n_N)$
to predicted expectation values $\vec{\nu} = (\nu_1, \ldots, \nu_N)$:

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

If n_i are independent Poisson r.v.s with means ν_i,
and all ν_i not too small (rule of thumb: all $\nu_i \geq 5$),
then χ^2 will follow the chi-square pdf for N dof.
The observed χ^2 then gives a P-value:

$$P = \int_{\chi^2}^{\infty} f(z; N) \, dz$$

where $f(z; N)$ is the chi-square pdf for N degrees of freedom.

Recall for chi-square pdf, $E[z] = N$,

$\rightarrow$ often give χ^2/N as measure of level of agreement

Better to give χ^2, N separately …
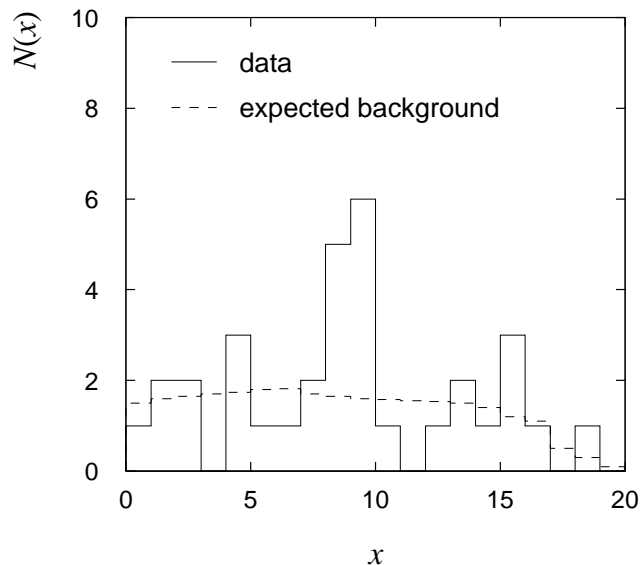
$\chi^2 = 15$, $N = 10 \rightarrow P$-value $= 0.13$
$\chi^2 = 150$, $N = 100 \rightarrow P$-value $= 9.0 \times 10^{-4}$

If $n_{\text{tot}} = \sum_{i=1}^{N} n_i$ is fixed, n_i are binomial, $p_i = \nu_i/n_{\text{tot}}$,

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

will follow chi-square for $N - 1$ dof (all $p_i n_{\text{tot}} >> 1$).

Example of χ^2 test

← This gives

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

$= 29.8$ for $N = 20$ dof.

But... many bins have few (or no) entries,

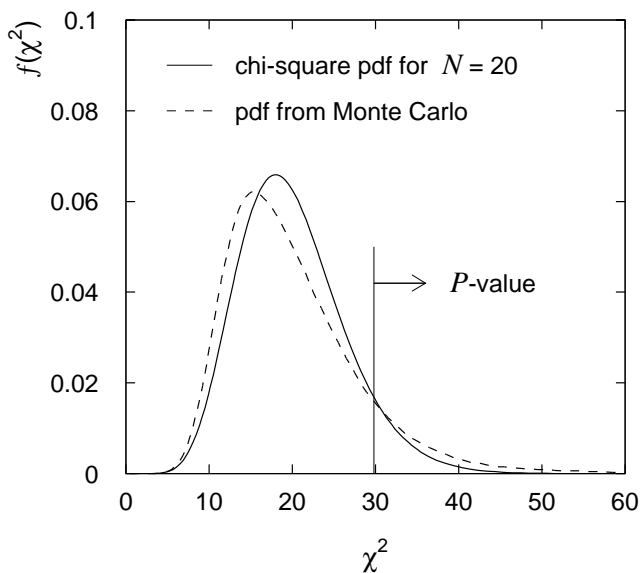$\rightarrow$ here χ^2 will not follow chi-square pdf.

Pearson's χ^2 still usable as a test statistic, but

to compute P-value first get $f(\chi^2)$ from Monte Carlo:

Generate n_i from Poisson, mean ν_i, $i = 1, \ldots, N$,

compute χ^2, record in histogram,

repeat experiment many times (here 10^6).

Using pdf from MC gives
$$P = 0.11$$

Chi-square pdf would give
$$P = 0.073$$

Parameter estimation: general concepts

Consider n independent observations of an r.v. x,

$$\rightarrow \text{ sample of size } n$$

Equivalently, single observation of an n-dimensional vector:

$$\vec{x} = (x_1, \ldots, x_n)$$

The x_i are independent $\Rightarrow$ joint pdf for the sample is

$$f_{\text{sample}}(\vec{x}) = f(x_1)f(x_2) \cdots f(x_n)$$

Task: given a data sample, infer properties of $f(x)$.

$\rightarrow$ construct functions of the data to estimate various
properties of $f(x)$ (mean, variance, ...)

Often, form of $f(x)$ hypothesized, value of parameter(s) unknown

$\rightarrow$ given form of $f(x; \theta)$ and data sample, estimate θ

Statistic = function of the data

Estimator = statistic used to estimate some property of a pdf

notation: estimator for θ is $\hat{\theta}$ (hat means estimator)

Estimate = an observed value of an estimator (often: $\hat{\theta}_{\text{obs}}$)

N.B. $\hat{\theta}(\vec{x})$ is a function of a (vector) random variable,

$\Rightarrow$ it is itself a random variable, characterized by a pdf $g(\hat{\theta})$
with an expectation value (mean), variance, etc.

How do we construct an estimator $\hat{\theta}(\vec{x})$?

> There is no golden rule on how
>
> to construct an estimator.

Construct estimators to statisfy (in general conflicting) criteria.

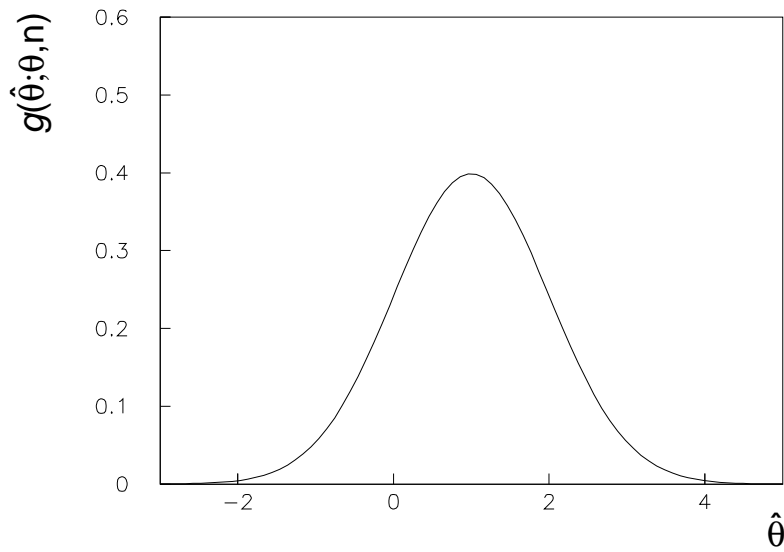As a start, require consistency: $\lim\limits_{n \to \infty} \hat{\theta} = \theta$

i.e. as size of sample increases, estimate converges to true value:

$$\text{for any } \epsilon > 0, \ \lim\limits_{n \to \infty} P(|\hat{\theta} - \theta| > \epsilon) = 0.$$

N.B. convergence in the sense of probability, i.e. no guaranty that any particular $\hat{\theta}_{\text{obs}}$ will be within any given distance of θ.

Properties of estimators

Consider the pdf of $\hat{\theta}$ for a fixed sample size n:

N.B. $g(\hat{\theta}; \theta, n)$ depends on true (unknown!) parameter θ.

We don't know θ, just a single value $\hat{\theta}_{\text{obs}}$.

Properties of $g(\hat{\theta}; \theta, n)$:

variance $V[\hat{\theta}] = \sigma_{\hat{\theta}}^2$.   ($\sigma_{\hat{\theta}} = $ 'statistical error')

bias $b = E[\hat{\theta}] - \theta$ ('systematic error', depends on n)

For many estimators we will have $\sigma_{\hat{\theta}} \propto \dfrac{1}{\sqrt{n}}, \quad b \propto \dfrac{1}{n}$.

Sometimes consider mean squared error:

$$\text{MSE} = V[\hat{\theta}] + b^2$$

In general, there is a trade-off between bias and variance,

$\rightarrow$ often require minimum variance among estimators with 0 bias.

Estimator for the mean (expectation value)

Consider n measurements of r.v. x, $x_1, \ldots, x_n$, we want an estimator for $\mu = E[x]$. Try arithmetic mean of the x_i:

$$\hat{\mu} = \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \text{(the sample mean)}$$

If $V[x]$ finite, $\overline{x}$ is a consistent estimator for μ, i.e.

$$\text{for any } \epsilon > 0, \ \lim_{n\to\infty} P\left(\left|\frac{1}{n} \sum_{i=1}^{n} x_i - \mu\right| \geq \epsilon\right) = 0 \ .$$

This is the Weak Law of Large Numbers. Compute expectation value:

$$E[\overline{x}] = E\left[\frac{1}{n} \sum_{i=1}^{n} x_i\right] = \frac{1}{n} \sum_{i=1}^{n} E[x_i] = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu$$

$\to \overline{x}$ is an unbiased estimator for μ. Compute variance:

$$V[\overline{x}] = E[\overline{x}^2] - (E[\overline{x}])^2 = E\left[\left(\frac{1}{n} \sum_{i=1}^{n} x_i\right)\left(\frac{1}{n} \sum_{j=1}^{n} x_j\right)\right] - \mu^2$$

$$= \frac{1}{n^2} \sum_{i,j=1}^{n} E[x_i x_j] - \mu^2$$

$$= \frac{1}{n^2}\left[(n^2 - n)\mu^2 + n(\mu^2 + \sigma^2)\right] - \mu^2 = \frac{\sigma^2}{n}$$

where σ^2 is the variance of x, and we used

$$E[x_i x_j] = \mu^2 \text{ for } i \neq j \text{ and } E[x_i^2] = \mu^2 + \sigma^2 \ .$$

Suppose mean μ and variance $V[x] = \sigma^2$ both unknown.

Estimate σ^2 with the sample variance:

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

Factor of $1/(n-1)$ included so that $E[s^2] = \sigma^2$ (i.e. no bias).

If $\mu = E[x]$ is known a priori,

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 = \overline{x^2} - \mu^2$$

is an unbiased estimator for σ^2.

Computing the variance of s^2 (long calculation!) gives

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

where μ_k is kth central moment (e.g. $\mu_2 = \sigma^2$).

The μ_k can be estimated using

$$m_k = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^k$$

To estimate the covariance $V_{xy} = \text{cov}[x, y]$, use

$$\widehat{V}_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \frac{n}{n-1} (\overline{xy} - \overline{x}\,\overline{y})$$

which is unbiased.

For the correlation coefficient $\rho = \dfrac{V_{xy}}{\sigma_x \sigma_y}$, use

$$r = \frac{\widehat{V}_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\left(\sum_{j=1}^{n}(x_j - \overline{x})^2 \cdot \sum_{k=1}^{n}(y_k - \overline{y})^2\right)^{1/2}}$$

$$= \frac{\overline{xy} - \overline{x}\,\overline{y}}{\sqrt{(\overline{x^2} - \overline{x}^2)(\overline{y^2} - \overline{y}^2)}} \ .$$

r has a bias which goes to zero as $n \to \infty$.

In general, pdf $g(r; \rho, n)$ is complicated; for Gaussian x, y,

$$E[r] = \rho - \frac{\rho(1 - \rho^2)}{2n} + O(n^{-2})$$

$$V[r] = \frac{1}{n}(1 - \rho^2)^2 + O(n^{-2})$$

(cf. R.J. Muirhead, *Aspects of Multivariate Statistical Theory*, Wiley, New York, 1982.)

- **Statistical tests:** test whether data stand in agreement with predicted probabilities, i.e., hypotheses. Critical region, significance level, power, (related to efficiency, purity).

- **Fisher discriminants, neural networks, etc.:** reduce data vector $\vec{x}$ to a single (or few) component function $t(\vec{x})$. Compactify data while retaining ability to discriminate between hypotheses.

- **Goodness-of-fit tests:** quantify level of agreement between data and hypothesis with P-value.

- **The significance of a signal:** often give P-value of hypothesis that only background present.

- **Introduction to parameter estimation:** try to minimize bias, variance. Simple estimators for mean, variance, covariance.