Lecture 5 outline

Statistical tests (part II)

1. **Testing goodness-of-fit, P-values**
2. **The significance of an observed signal**
3. **Pearson's χ^2 test**

General concepts of parameter estimation

1. **Samples, estimators, bias**
2. **Estimators for mean, variance, covariance**

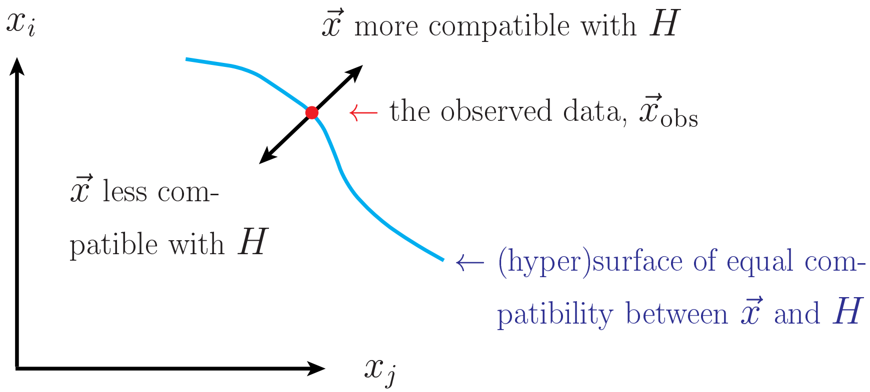
Testing goodness-of-fit

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 $\vec{x}_{\rm obs}$ (or $t(\vec{x}_{\rm obs})$)

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.

An example of a goodness-of-fit test

Probability to observe $n_{\rm h}$ heads in N coin tosses is:

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

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

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

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

Region of t-space with equal or lesser compatibility:

$t \geq 7$

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

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.

The significance of an observed signal

Suppose we observe n events; these can consist of:

$n_{\rm b}$ events from known processes (background)

$n_{\rm s}$ events from new processes (signal)

If $n_{\rm b}, n_{\rm s}$ are Poisson r.v.s with means $\nu_{\rm b}, \nu_{\rm s}$, $\Rightarrow n = n_{\rm s} + n_{\rm b}$ is also Poisson, mean $\nu = \nu_{\rm s} + \nu_{\rm b}$ (cf. SDA Chapter 10):

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

Suppose $\nu_{\rm b} = 0.5$ and we observe $n_{\rm obs} = 5$.
Should we claim evidence for a new discovery?

Hypothesis H: $\nu_{\rm s} = 0$, i.e. only background present.

$$P\text{-value} = P(n \geq n_{\rm obs})$$

$$= \sum_{n=n_{\rm obs}}^{\infty} P(n; \nu_{\rm s} = 0, \nu_{\rm b})$$

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

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

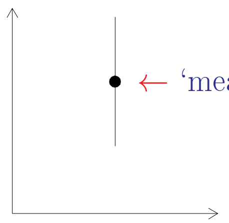
$(\neq P(\nu_{\rm s} = 0)!)$

Pitfalls

A misleading (but often used) representation …

estimate for ν is $n_{\rm obs} = 5$,

estimated standard deviation of n is $\sqrt{n} = 2.2$,



$\leftarrow$ 'measured signal' $= n_{\rm obs} - \nu_{\rm b} = 4.5 \pm 2.2$

i.e. about 2σ from 0

What we want: probability for Poisson variable of mean $\nu_{\rm b} = 0.5$ to give 5 or more. (Answer: 1.7×10^{-4})

What the picture implies: probability for variable of mean 4.5, $\sigma = 2.2$ to give 0 or less. (Answer for Gaussian: 0.021)

$\rightarrow$ not a problem if $\nu \gg 1$, i.e. n Gaussian

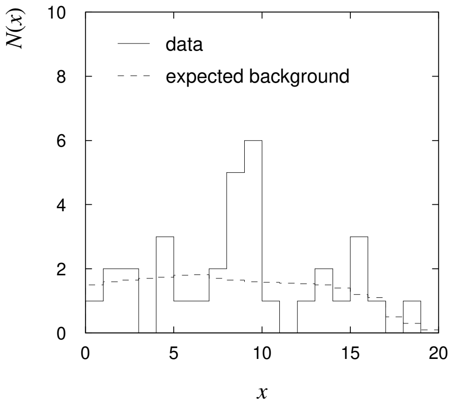
Another pitfall: In practice $\nu_{\rm b}$ has a systematic uncertainty.
Suppose e.g. $\nu_{\rm b} = 0.8$,

$$P(n \geq 5; \nu_{\rm b} = 0.8, \nu_{\rm s} = 0) = 1.4 \times 10^{-3}$$

$\Rightarrow$ report range of P-values for a reasonable variation of $\nu_{\rm b}$.
(No well established convention.)

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???

Pearson's χ^2 test

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

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

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, $E[z] = N$,

→ often give χ^2/N as measure of level of agreement

Better to give χ^2, N separately ...

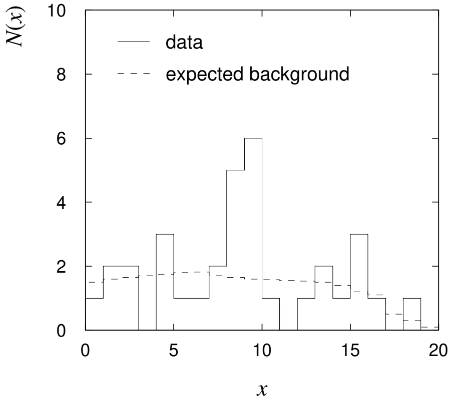
$\chi^2 = 15$, $N = 10$ → P-value $= 0.13$
$\chi^2 = 150$, $N = 100$ → P-value $= 9.0 \times 10^{-4}$

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



$\leftarrow$ This gives

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

$$= 29.8 \text{ for } N = 20 \text{ dof.}$$

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

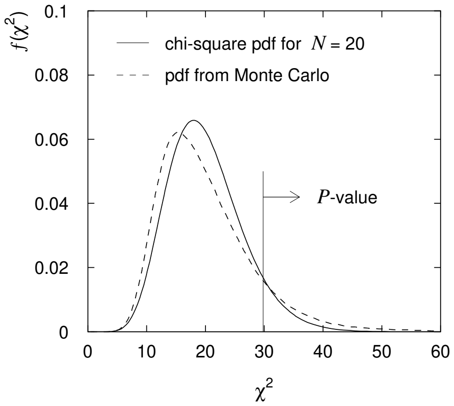
$\rightarrow$ here χ^2 will not follow chi-square pdf.

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

$P = 0.073$

Program for generating Poisson random numbers

```
      program TEST_RNPSSN

c   Test program for CERNLIB routine RNPSSN (V136) for generating
c   Poisson distributed numbers.

      implicit        NONE

c   Needed for HBOOK routines

      integer         hsize
      parameter       (hsize = 100000)
      integer         hmemor (hsize)
      common  /pawc/  hmemor

c   Local variables

      character*80    outfile
      integer         i, icycle, ierror, istat, lun, n
      real            nu

c   Initialize HBOOK, open histogram file, book histograms.

      call HLIMIT (hsize)
      lun = 20
      outfile = 'test_rnpssn.his'
      call HROPEN (lun, 'histog', outfile, 'N', 1024, istat)
      call HBOOK1 (1, 'Poisson n', 100, -0.5, 99.5, 0.)

c   Generate 10000 values and enter into histogram.

      write (*, *) 'enter Poisson mean nu'
      read (*, *) nu
      do i = 1, 10000
        call RNPSSN (nu, n, ierror)
        call HF1 (1, FLOAT(n), 1.)
      end do

c   Store histogram and close.

      call HROUT (0, icycle, ' ')
      call HREND ('histog')

      stop
      END
```

Parameter estimation: general concepts

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

$\rightarrow$ sample of size n

Equivalently, single observation of an n-dimensional vector:

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

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.

Estimators

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: $\lim_{n \to \infty} \hat{\theta} = \theta$

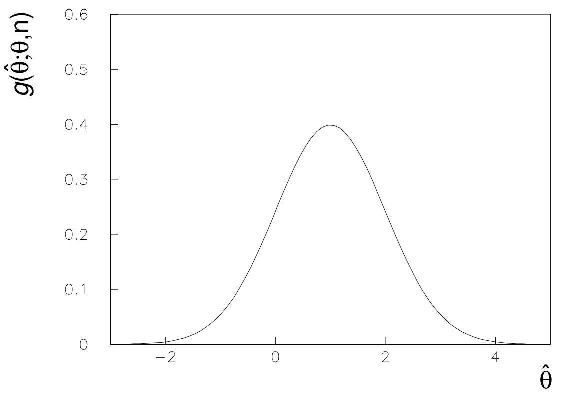
i.e. as size of sample increases, estimate converges to true value:

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

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 $\hat{\theta}_{\mathrm{obs}}$.

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

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

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

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

Sometimes consider mean squared error:

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

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

→ 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.

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

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

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

→ $\overline{x}$ 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

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

Estimator for the variance

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

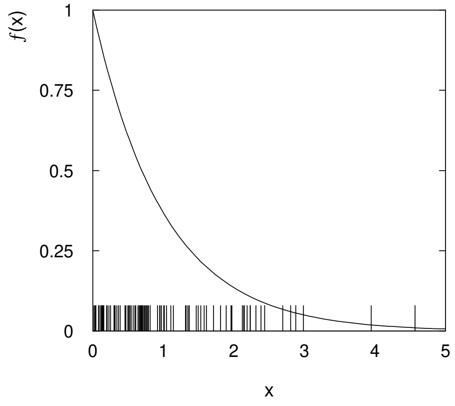
$$V[s^2] = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\mu_2^2\right)$$

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

The μ_k can be estimated using

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

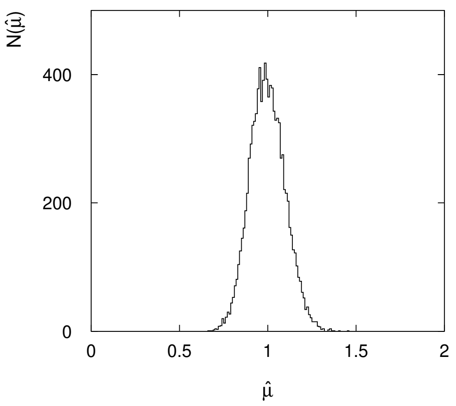
Example of estimator for mean



Data sample of $n = 100$ values from MC with $\mu = 1, \ \sigma^2 = 1.$

$$\hat{\mu} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 1.073$$

Now repeat the experiment 10^4 times with $n = 100$ values each, enter the sample mean for each experiment into histogram:



$\overline{\hat{\mu}} = 0.9981$ ($\hat{\mu}$ unbiased)

Sample standard deviation of $\hat{\mu}$ values $= 0.0995$

$$\approx \frac{\sigma}{\sqrt{n}}$$

N.B. pdf of $\hat{\mu}$ approximately Gaussian (Central Limit Theorem).

Estimator for covariance and correlation coefficient

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

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

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

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.)

Lecture 5 summary

Statistical tests (part II)

1. **Testing goodness-of-fit:** P-value is the probability to get data as inconsistent with the hypothesis (or more so) as is the data that we actually obtained.
2. **The significance of an observed signal:** A minefield. The literature is full of 10^{-4} effects that turned out to be fluctuations.
3. **Pearson's χ^2 test:** Probably most widely used test statistic. For small data samples, doesn't follow chi-square pdf. (Still OK, get pdf from MC.)

General concepts of parameter estimation

1. **Estimators:** No golden rule on how to construct an estimator, pick one according to its properties (consistency, bias, variance).
2. **Estimators for mean, variance, covariance:** Here not derived from any deeper principle, but their properties turn out to be (almost) optimal.