

Essential Statistics

Christian Knigge

- Basic Ideas
- Probability Distributions
- Samples and Populations
- Mean, Median, Mode
- Variance and Covariance
- Least-Squares/Chi-Squared Fitting
- Error Propagation

Basic Ideas

Definitions

- $P(A)$ probability of event A happening
(equivalently: probability of statement A being true)
- $P(A, B)$ probability of events A and B happening
(equivalently: probability of statements A and B both being true)
- $P(A|B)$ probability of event A happening, given that event B has happened
(equivalently: probability of statement A being true, given that B is true)
(note: if A and B are independent, $P(A|B) = P(A)P(B)$)

Fundamental Relation

$$P(A;B) = P(A|B)P(B) = P(B|A)P(A)$$

Bayes Theorem

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Example of Bayes' Theorem

DNA Evidence

- Statement A : DNA matches
- Statement B : Defendant is guilty
- Suppose $P(A) = 10^{-9}$
 - Note that this is not the same as the probability of a false positive – see later
- If world population 5×10^9 , have $P(B) = 2 \times 10^{-10}$ a priori
- Assume that if guilty, then DNA matches, i.e. $P(A|B) = 1$

$$P(B|A) = \frac{P(A|B) P(B)}{P(A)} = \frac{1 \times 2 \times 10^{-10}}{10^{-9}} = 0.2$$

So why is DNA evidence considered to be so strong?

$P(B) = 2 \times 10^{-10}$ assumes arrest was made only on basis of DNA evidence and no other evidence exists! Otherwise $P(B) \gg 2 \times 10^{-10}$ (but difficult to quantify)

Prior knowledge can be important!

Question: the equation above suggests that $P(B|A)$ could be > 1 !
Can this happen? What would it mean? (Well come back to this in a minute)

More Basics: The Negation of A Statement and Its Uses

Definitions:

- $P(A)$: Probability of event happening
- $P(\sim A)$: Probability of event not happening

Fundamental Relation:

$$P(A) + P(\sim A) = 1$$

Example:

- What's the probability $P(\text{catch})$ that a dwarf nova should be found at least once in eruption in N observing epochs (assume epoch spacing \gg outburst recurrence time)?
- Assume dwarf novae spent 30% of their life in eruption (and 70% in quiescence), i.e.

$$P_e = 0.3$$

$$P_q = 1 - P_e = 0.7$$

- Negation of $P(\text{catch})$ is $P(\text{miss})$, the probability of being quiescent in all N epochs

$$P(\text{miss}) = P(\sim \text{catch}) = P_q^N$$

- So probability of catching DN is

$$P(\text{catch}) = 1 - P(\sim \text{catch}) = 1 - P(\text{miss}) = 1 - P_q^N$$

The Example of Bayes' Theorem revisited

Let's return to Bayes' theorem and check if $P(B|A)$ could ever be > 1 . To do this, let's take a closer look at the priors:

- Statement A : DNA matches
- Statement B : Defendant is guilty
- First consider a simple (but sensible) prior for B
 - suppose the defendant is one of N_{pop} people who are equally likely suspects in the absence of DNA evidence
 - $P(B) = 1/N_{pop}$
- Now consider the prior on $P(A)$,
 - DNA will match if either
 - (i) guilty and test worked OK
 - (ii) innocent and test produced a false positive
 - So $P(A) = P(A|B) * P(B) + P(A|\neg B) * P(\neg B)$
- We already know $P(B)$ and $P(\neg B) = 1 - P(B)$.
- So we only need $P(A|B)$ and $P(A|\neg B)$
 - Let's assume the test never gives false negatives, i.e. if guilty, there will always be a match
 - $P(A|B) = 1$ (in practice it only matters that this probability be close to unity)
 - The probability of a false positive is
 - $P(A|\neg B) = P_{fp}$
- We then have
 - $P(A) = 1 \times 1/N_{pop} + [P_{fp} \times (1 - 1/N_{pop})]$
 - For $N_{pop} \gg 1$, this is
 - $P(A) = 1/N_{pop} + P_{fp}$
- Bayes' Theorem said that probability of guilt given DNA match was
 - $P(B|A) = P(A|B) * P(B) / P(A)$
- So we finally have
 - $P(B|A) = (1/N_{pop}) / (1/N_{pop} + P_{fp}) = 1 / (1 + N_{pop} * P_{fp})$
 - Note that this varies exactly between zero and one in the sort of way you'd expect:
 - $P(B|A)$ is small if N_{pop} large
 - Guilt is less likely if there are lots of a priori equally likely suspects
 - $P(B|A)$ is small if P_{fp} large
 - Guilt is less likely if test produces lots of false positives

Probability Distribution Functions

- A continuous random variable x is said to be distributed according to the probability distribution function $p(x)$ if the probability of finding the variable between x and $x+dx$ is $p(x) dx$

- Continuous PDFs are probability density functions (probability per unit "length")
- PDFs should (usually) be normalized:

$$\int_{x_{min}}^{x_{max}} p(x) dx = 1$$

- A discrete random variable n is said to be distributed to the probability distribution function $p(n)$ if the probability of observing exactly n is given by $p(n)$

- Normalization condition for discrete distributions is:

$$\sum_{n_{min}}^{n_{max}} p(n) = 1$$

- It's sometimes useful to approximate discrete distributions $p(n)$ by continuous ones $p(x)$

$$\sum_{n_{min}}^{n_{max}} p(n) \approx \int_{n_{min}}^{n_{max}} p(x) dx = 1$$

- The point is not to simply replace discrete n with continuous x (while keeping functional form), but to replace a "difficult" $p(n)$ with a simpler $p(x)$ (e.g. a Gaussian; see later)
- To recover an estimate of the discrete $p(n)$ from a continuous approximation $p(x)$, can use sth like

$$p(n) \approx \int_{n-0.5}^{n+0.5} p(x) dx$$

Transforming PDFs

- How do we get from one PDF to another? I.e. if $p(x)$ is PDF of x , what is the PDF of $y = f(x)$?

- Key point is that probability in interval x to $x + dx$ must be conserved in the corresponding interval y to $y + dy$

- So we must have

$$p(y) |dy| = p(x) |dx|$$

where $y = y(x)$ and $y + dy = y(x + dx)$

- $p(y)$ and $p(x)$ must therefore be related via

$$p(y) = \left| \frac{dy}{dx} \right|^{-1} p(x)$$

Cumulative Distribution Functions

- For any PDF $p(x)$, can define a cumulative distribution function $P(x)$, which gives the probability that the random variable takes a value less than x

$$P(x) = \int_{x_{\min}}^x p(x') dx'$$

- CDFs have useful properties:
 1. monotonically increasing
 2. always have $P(x_{\min}) = 0$ and $P(x_{\max}) = 1$
 3. the random variable $y = P(x)$ is itself distributed uniformly between 0 and 1:

$$p(y) = \left| \frac{dy}{dx} \right|^{-1} p(x) = \left| \frac{dP(x)}{dx} \right|^{-1} p(x) = p(x)^{-1} p(x) = 1$$

- The last property means we can use CDFs to construct pseudo-random numbers distributed according to any PDF by drawing uniformly distributed numbers between 0 and 1 and finding the corresponding x values from $P(x)$

Mean and Variance of a PDF

- Two key properties of a PDF are

- Mean: a measure of central tendency

$$m = \int_{x_{\min}}^{x_{\max}} x p(x) dx$$

- Variance: a measure of the dispersion around the mean

$$s^2 = \int_{x_{\min}}^{x_{\max}} (x - m)^2 p(x) dx$$

$$s = \sqrt{s^2} = \text{standard deviation}$$

Expectation Values

- If $f(x)$ is a function of the random variable x and the PDF of x is given by $p(x)$, then the expectation value of that function is defined as

$$\langle f \rangle = \int_{x_{m \text{ in}}}^{x_{m \text{ ax}}} f(x)p(x)dx$$

- So the mean is just the expectation value of x

$$1 = \langle x \rangle$$

- And the variance is the expectation value of the square of the deviations from the mean

$$\frac{3}{4}^2 = \langle (x - 1)^2 \rangle$$

– Useful formula:

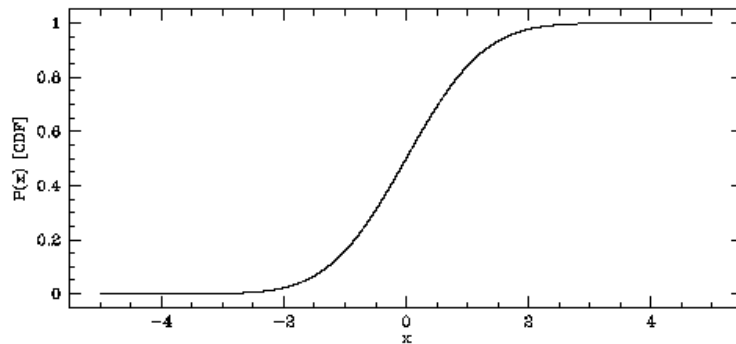
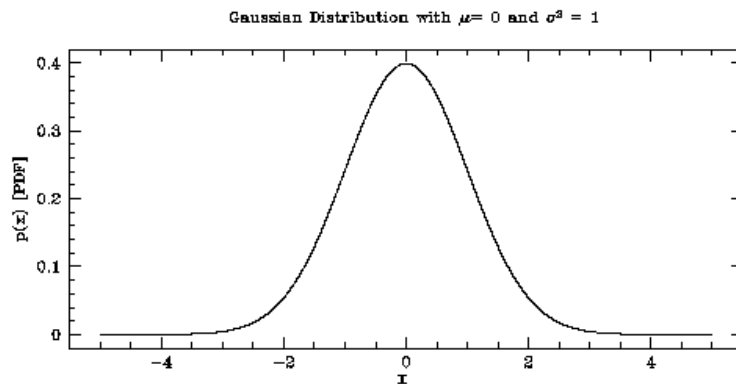
$$\begin{aligned} \frac{3}{4}^2 &= \langle (x - 1)^2 \rangle \\ &= \langle x^2 - 2x + 1 \rangle \\ &= \langle x^2 \rangle - 2\langle x \rangle + 1 \\ &= \langle x^2 \rangle - 2(1) + 1 \\ &= \langle x^2 \rangle - 1 \end{aligned}$$

“Variance = Mean of Squares – Square of Mean”

Examples of PDFs: Gaussian Distribution

- Gaussian (or Normal) Distribution is given by

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$$



“Confidence Intervals”

$$1 \pm n\frac{\sigma}{\sqrt{n}}$$

$n = 1:$	68.3%
$n = 2:$	95.4%
$n = 3:$	99.7%

Common shorthand for Gaussian with mean μ and variance σ^2 :

$$G(\mu, \sigma^2)$$

$$N(\mu, \sigma^2)$$

Standard Form of the Normal Distribution

- The Gaussian Distribution with zero mean and unit variance, $G(0,1)$, is called the "Standard Normal Distribution"
- This is useful because any normal distribution can be transformed into standard form by a change of variables (and vice versa)
 - Suppose you have a random variable x that is distributed according to $G(\mu, \sigma^2)$
 - Then the random variable $z = (x - \mu)/\sigma$ is distributed according to the standard normal distribution, $G(0,1)$.
 - Inverting this, if z is distributed according to standard normal, then $x = \mu + z\sigma$ is distributed according to $G(\mu, \sigma^2)$
 - So no need to store or calculate any Gaussian distributions or look-up tables except for the standard normal one!

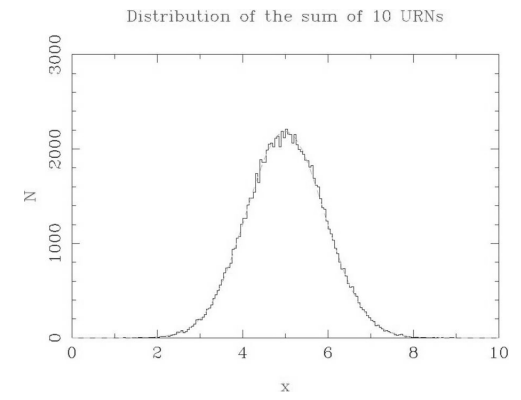
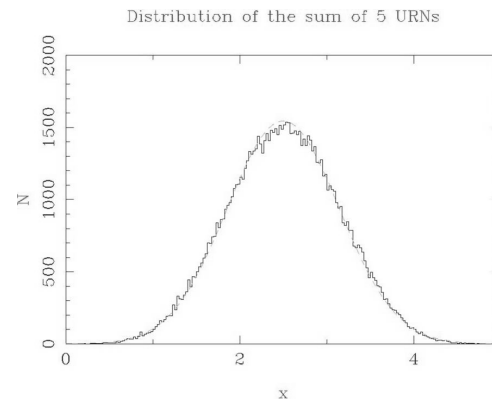
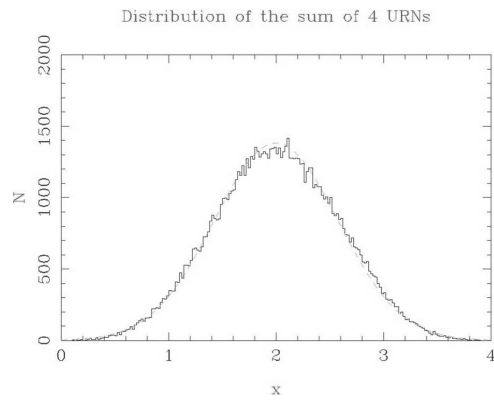
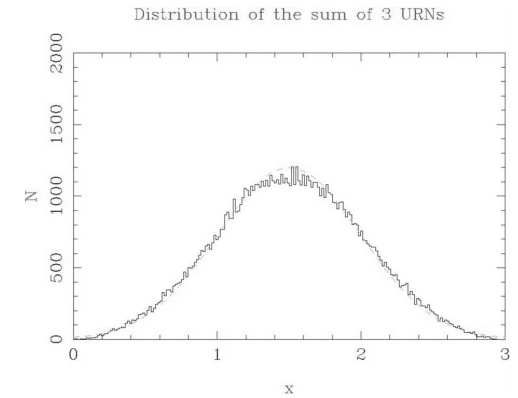
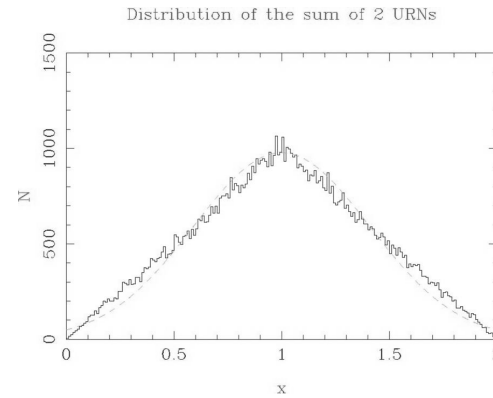
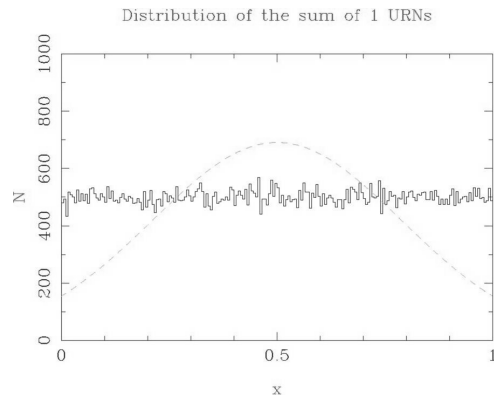
Why is the Gaussian Distribution so Important?

- Central Limit Theorem (aka "Law of Large Numbers")
 - Consider N independent random variables, x_i , drawn from arbitrary parent distributions with means μ_i and variances s_i^2 . The sum, X , of these random variables is itself a random variable, and as $N \rightarrow \infty$, the PDF of X will tend to a Gaussian distribution with mean $N\mu_i$ and variance $N s_i^2$.

$$X = \sum_{i=1}^N x_i \sim G(N\mu_i; N s_i^2)$$

- This result is hugely important! It's the reason Gaussians pop up all over the place.
- For example, observational errors are often due to lots of small, separate effects
 - The CLT says we should then expect the overall errors to have Gaussian distributions!

An Illustration of Central Limit Theorem : Sums of Uniformly Distributed Random Variables



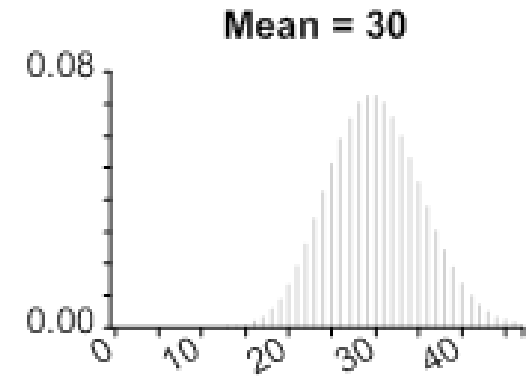
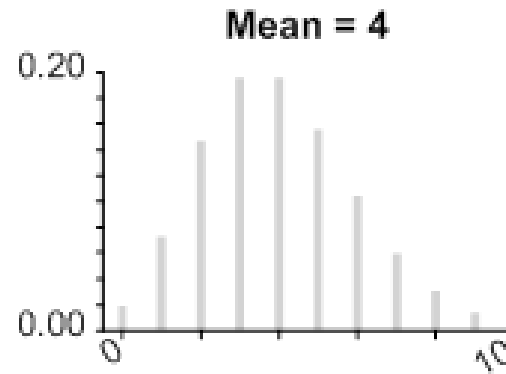
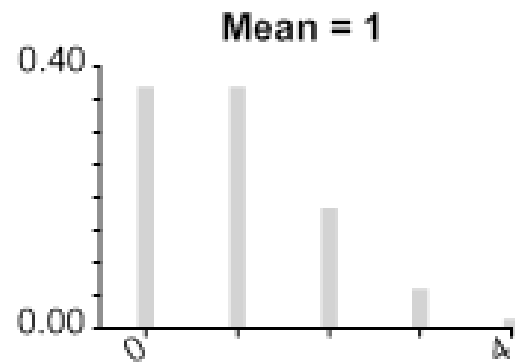
Examples of PDFs: The Poisson Distribution

- The Poisson distribution with mean μ is given by

$$P(n) = \frac{\mu^n e^{-\mu}}{n!}$$

- Describes probability of seeing n "counts" if mean "count rate" is μ
- The Poisson distribution is a discrete PDF
 - n is restricted to integer values
- Very important in astronomy
 - e.g. number of photons hitting a detector should obey Poisson distribution
- Variance of Poisson distribution is $s^2 = \mu$ (and standard deviation $s = \sqrt{\mu}$)
 - Example: observe 100 counts in an image; can immediately estimate that error on this number is roughly 10
 - Beware of the low count limit:
 - Suppose we observe zero counts in some interval; is the error on this zero?
 - NO: the variance is equal to the expected number of counts (not the observed number)
 - Need to be careful: taking $s = \sqrt{N}$ is wrong in the low count limit and can lead to seriously incorrect (biased) estimates!

The Poisson Distribution (contd)



As μ increases, the Poisson distribution tends to the Gaussian distribution $G(\mu, \mu)$

Examples of PDFs: The χ^2 Distribution

- Let z_i be a set of standard normal random variables; then

$$\hat{A}^2 = \sum_{i=1}^N z_i^2$$

is distributed according to a χ^2 distribution with N degrees of freedom

- Don't bother writing the distribution down explicitly, because
 - it's fairly complicated (it's a special case of the class of gamma-distribution and includes a gamma-function)
 - one rarely needs the full distribution, but rather critical values corresponding to some particular significance level, so one can generally just use look-up tables or public software
- Mean of χ^2 distribution is N , variance is $2N$
- Example: you want to test if a set of z_i are consistent with being standard normal random variables
 - Calculate the corresponding value of χ^2
 - Look up critical value χ^2_{crit} defined by $P(\chi^2_{\text{crit}}) = 0.95$ (say); note that P here is CDF, not PDF
 - If $\chi^2 > \chi^2_{\text{crit}}$ the z_i are inconsistent with standard normal at the 5% significance level
 - Hmm, is this right? What if $P(\chi^2) < 0.05$ (i.e. χ^2 is smaller than expected)?
 - Answer depends on context—always think about whether your test should be one-sided or two-sided!
- χ^2 is important because it crops up as the “goodness of fit” statistic we get in least-squares, maximum-likelihood model fitting problems (see later)

Samples and Populations

- In the real world, all we have are finite samples of data drawn from the infinite (or very large) parent populations whose statistical properties are described by PDF, CDF, m , s^2 etc
- Our current definitions of quantities like mean, variance, expectation values are only relevant to the parent populations (and assume known PDF), but how do we estimate such quantities from an observed sample?
- Key idea: each data point in the sample can provide an independent estimate of the expectation value; so the best estimate is just the average of all these estimates

– Sample Mean:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

– Sample Variance:

$$s^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2$$

- The denominator is $(N-1)$ – not N – to remove the bias that would otherwise arise since we're using the sample mean in the sum (which was itself estimated from the data).
- Apart from a factor $N/(N-1)$, the handy formula from before carries over

$$s^2 = \frac{N}{N-1} \left(\overline{x^2} - \bar{x}^2 \right)$$

- This is useful because it's a one-pass algorithm for calculating sample variance

– Sample estimate of expectation value of $f(x)$:

$$\bar{f} = \frac{1}{N} \sum_{i=1}^N f(x_i)$$

- Note that in general

$$\bar{f} \neq f(\bar{x})$$

Other Measures of Central Tendency: Mean, Median and Mode

- Mean

- The most common measure
- Formally: the mean is the centroid of the PDF
- Operationally: the average value of x obtained from N random draws converges to m as $N \rightarrow \infty$

- Median:

- The location which splits the area under the PDF into two equal pieces;
- Can use CDF to define this as

$$P(x = \text{median}) = \frac{1}{2}$$

- Mode:

- The single most likely value
- Corresponds to the peak of the PDF, so we can define this as

$$p(x = \text{mode}) = \text{Max}[p(x)]$$

- Note that this is not necessarily equivalent to the condition $dp/dx = 0$
 - $P(x)$ could have more than one maximum!

Mean, Median & Mode (contd)

- Symmetric distributions:
 - mean = median = mode
- Asymmetric distributions:
 - mean > median > mode (if tail towards large x)

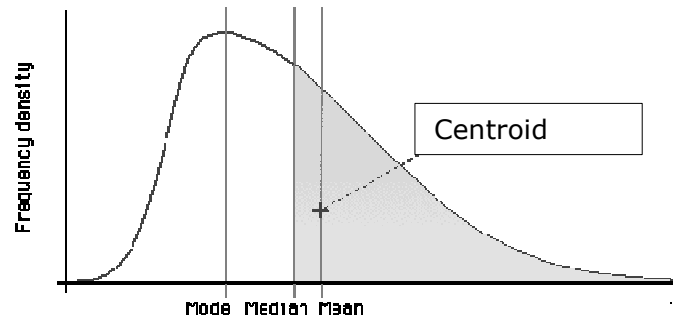


Figure Illustration of mean, median and mode

- Bimodal or Multimodal distributions:
 - mode still gives location of highest peak
 - expect median to lie closer to main peak than mean
- So if we're after the location of the main peak of a distribution, mode and median can be preferable to the mean since both are less sensitive to outliers!
 - Example 1: background estimation in crowded field photography
 - Example 2: cosmic ray rejection by "medianing" of multiple images
- BUT
 - Median and Mode are more difficult to calculate (see below)
 - Can show that errors on median and mode are larger than those on mean

Estimating Median and Mode From Samples

- Median:

- Sort data
- Then median is the value which has an equal number of values greater and less than it; for sorted x and N odd, this means

$$\text{Median}(x) = x_{(N+1)/2}$$

- For N even, the median is the mean of the two middle values

$$\text{Median}(x) = \frac{x_{N/2} + x_{(N/2)+1}}{2}$$

- Mode:

- The mode is the most common value
- May be non-unique (two equal sized peaks in PDF)
- May need to bin & construct histogram if data is non-integer (or integer but no recurring values)
 - But note that binning is subjective!

Error Propagation

- Suppose you have measured two quantities \bar{u} and \bar{v} , where the overline is there to remind us that what we have (usually) really measured is an estimate of the expectation value of the random variables u and v .
- Suppose further that we also have an estimate of the error (standard deviation) s_u and s_v associated with these measured values
- How do we use this information to estimate the expectation value (or perhaps the most probable value?) of the function $f(u, v)$ and the associated error s_f ?
- Strictly speaking, what we should do is work out the PDF of $f(u, v)$ from the PDFs of u and v and take it from there
 - see “Transformation of PDFs” from before, although would still need to generalize to multi-variable case
- However, in practice, one often cheats!

Error Propagation (contd)

- Approximate the expectation value of f by

$$\bar{f}(u;v) \approx f(\bar{u};\bar{v})$$

- Taylor expand $f(u,v)$ about the point $(\bar{u};\bar{v})$

$$f(u;v) \approx \bar{f}(u;v) + f(u;v) - f(\bar{u};\bar{v}) + (u - \bar{u}) \frac{\partial f}{\partial u} + (v - \bar{v}) \frac{\partial f}{\partial v}$$

- Now from the definition of variance, the variance of f is $\sigma_f^2 = \overline{(f - \bar{f})^2}$
- So from the Taylor expansion (and replacing $\sigma_u^2 = \overline{(u - \bar{u})^2}$ and ditto for v)

$$\sigma_f^2 \approx \sigma_u^2 \left(\frac{\partial f}{\partial u} \right)^2 + \sigma_v^2 \left(\frac{\partial f}{\partial v} \right)^2 + 2h(u - \bar{u})(v - \bar{v}) \frac{\partial f}{\partial u} \frac{\partial f}{\partial v}$$

- Finally let $\sigma_{uv}^2 = h(u - \bar{u})(v - \bar{v})$ and allow for possibility of additional variables

$$\sigma_f^2 \approx \sigma_u^2 \left(\frac{\partial f}{\partial u} \right)^2 + \sigma_v^2 \left(\frac{\partial f}{\partial v} \right)^2 + \dots + 2\sigma_{uv}^2 \frac{\partial f}{\partial u} \frac{\partial f}{\partial v} + \dots$$

- This is the best-known form of the error propagation equation

— You should remember this!

Error Propagation (contd)

- The quantity

$$\frac{\sigma_{uv}^2}{4} = h(u; \bar{u})(v; \bar{v})$$

is called the Covariance of u and v

- If u and v are independent, then $p(u, v) = p(u)p(v)$, i.e. u doesn't care about v and vice versa. In that case $\frac{\sigma_{uv}^2}{4} = 0$ and the error propagation equation simplifies to

$$\frac{\sigma_f^2}{4} = \frac{\sigma_u^2}{4} \left(\frac{\partial f}{\partial u}\right)^2 + \frac{\sigma_v^2}{4} \left(\frac{\partial f}{\partial v}\right)^2 + \dots$$

- Example: let $f(u, v) = u + v$; then $\frac{\partial f}{\partial u} = \frac{\partial f}{\partial v} = 1$ and we recover the familiar "errors add in quadrature" rule

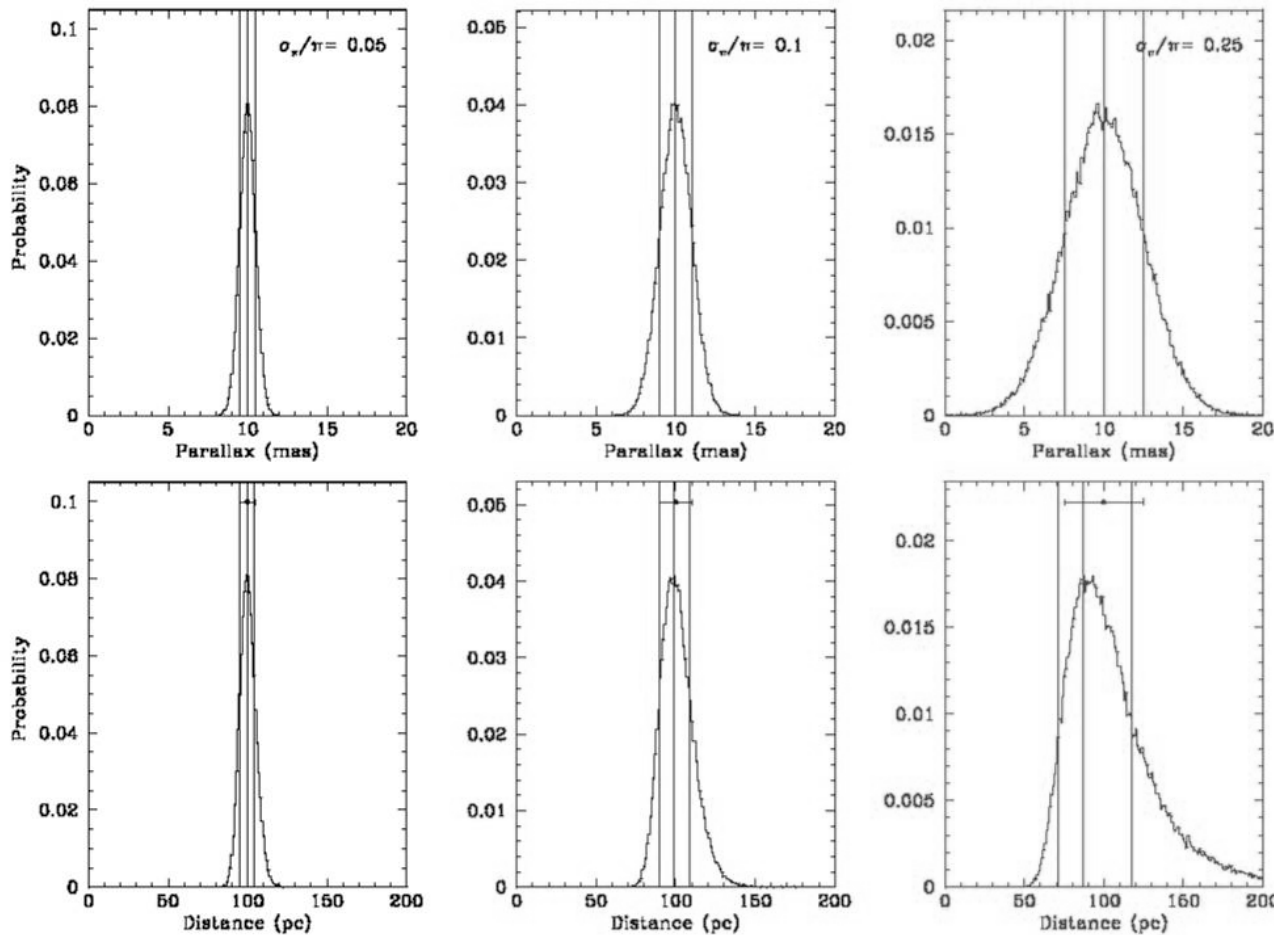
$$\frac{\sigma_f^2}{4} = \frac{\sigma_u^2}{4} + \frac{\sigma_v^2}{4}$$

- If $u, v \dots$ are directly measured data points, the assumption of independence is often OK (but you should always think about it!).
- However, if $u, v \dots$ are parameters that were themselves inferred from a common data set (e.g. by model fitting), then it's quite likely that their covariance will be non-zero!

- Watch out!
- Will come back to this later

Error Propagation Example: Distances from Parallax Measurements

- Relationship between parallax π (in arcsecond) and distance d (in parsecs) is $d = 1/\pi$
- So given a parallax measurement (with error), what's the best distance estimate (and its error)? Assume parallax measurements are normally distributed
- Well, our error propagation rules suggest that we take $\bar{d} = 1/\bar{\pi}$ and $\frac{\sigma_d}{d} = \frac{\sigma_\pi}{\pi}$
- How does that compare to the true PDF of d ?



Vertical Lines: Mean and 1-s intervals of the Gaussian parallax distributions

Vertical Lines: Most probable distance value and the smallest 68% (1-s) confidence interval

Horizontal Error Bar: Distance estimate and associated 1-s error from standard error propagation

Likelihood

- In modelling data D with a model M , the quantity of interest is the probability that the model is correct, given the data

$$P(M | D) = \frac{P(D | M) P(M)}{P(D)}$$

- $P(D)$ is a normalization factor
 - $P(M)$ is the “prior” (or “a priori”) probability of the model
 - $P(D | M)$ is the “likelihood” (the data dependent part of the right hand side)
- $P(M)$ is often almost totally unknown (and disagreement about if and how to treat it at heart of “Bayesian” vs “Frequentist” battles)
 - Focus on the likelihood instead, i.e. maximize the probability that we should have observed the data, given our model

Least Squares

Consider modelling a dataset of N points, y_i , each of gaussian distribution with standard deviation σ_i .

The model \hat{y} is a function of variables \mathbf{x} and model parameters \mathbf{a} .

Then the likelihood

$$P(D|M) = \prod_{i=1}^N \left[\frac{1}{\sqrt{2\pi}\sigma_i} \exp -\frac{(y_i - \hat{y}_i)^2}{2\sigma_i^2} \right].$$

Taking logs,

$$\log P(D|M) = C - \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{2\sigma_i^2}.$$

Maximise likelihood by *minimising*

$$\chi^2 = \sum_{i=1}^N \left(\frac{y_i - \hat{y}_i}{\sigma_i} \right)^2.$$

Minimise by optimising model parameters \mathbf{a} .

Linear Least Squares

Minimum $\chi^2 \Rightarrow \partial\chi^2/\partial a_i = 0$ for all a_k , $k = 1$ to N , thus

$$\sum_{i=1}^N \left(\frac{y_i - \hat{y}_i}{\sigma_i^2} \right) \frac{\partial \hat{y}_i}{\partial a_k} = 0$$

If \hat{y} depends *linearly* on the a_i so that

$$\hat{y}_i = X_j(\mathbf{x}_i) a_j$$

Suffix convention: sum over repeated indices is implied

(suffix convention) then

$$\sum_{i=1}^N \left(\frac{y_i - X_j(\mathbf{x}_i) a_j}{\sigma_i^2} \right) X_k(\mathbf{x}_i) = 0$$

equivalent to the matrix equation $\mathbf{A}\mathbf{a} = \mathbf{b}$ where

$$A_{kj} = \sum_{i=1}^N \frac{X_k(\mathbf{x}_i) X_j(\mathbf{x}_i)}{\sigma_i^2}$$

and

$$b_k = \sum_{i=1}^N \frac{X_k(\mathbf{x}_i) y_i}{\sigma_i^2}$$

These are called the *normal equations*.

Important: Can show that \mathbf{A}^{-1} is equal to the covariance matrix \mathbf{C} , where

$$C_{ij} = \frac{1}{4} \delta_{ij}$$

i.e. diagonal elements of \mathbf{C} are the variances of the parameters, off-diagonal elements are covariances!

Very useful and important for estimating errors on fit parameters and quantities inferred from fit parameters

χ^2 & Goodness-of-Fit

The sum of the squares of N independent gaussian variables follows the χ^2 distribution with N degrees of freedom.

In a linear least-squares fit of M parameters to N points the minimum χ^2 is distributed as χ^2 of $N - M$ degrees of freedom.

Since for a gaussian variable x , $\langle x^2 \rangle = 1$, in linear least-squares we expect:

$$\langle \chi^2 \rangle = N - M$$

$N - M$ degrees of freedom, rather than N , for same reason that we had $N - 1$ (rather than N) in denominator of sample variance estimate.

Example: Fit a parabola to 10 data points, obtain $\chi_{\min}^2 = 18.5$. Is this a good fit?

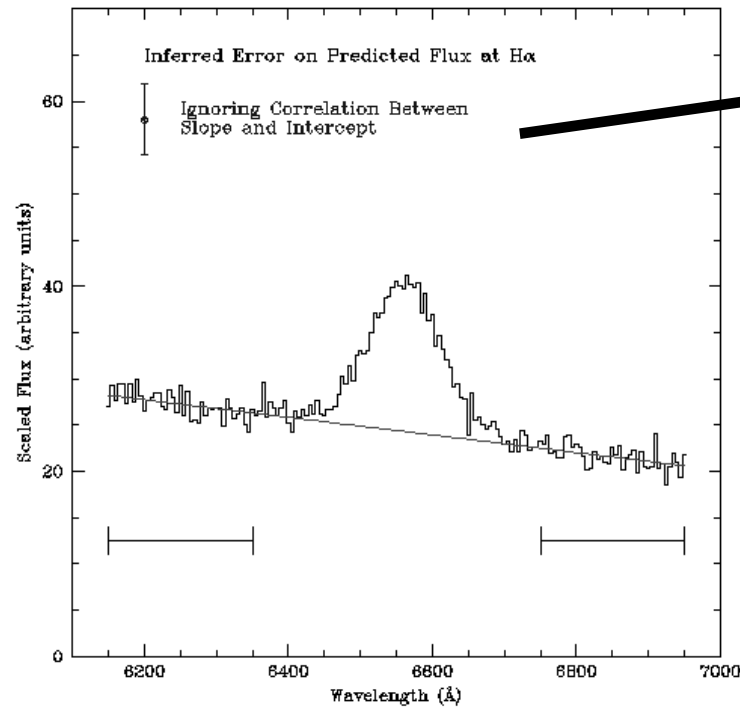
3 coefficients are needed to define a parabola.

Look up *percentiles of χ^2* for 7 d.o.f..

Find $\chi_7^2 > 18.5$ only 1% of the time.

Model Fitting and Error Propagation: The Covariance Strikes Back

- You have obtained a spectrum that contains an emission line (say H α) and now want an estimate of the continuum flux (and its error) right underneath the line peak (at $\lambda = 6563$ Angstroms).
- Fit a straight line $f = m\lambda + b$ to suitable continuum windows, yielding m and b , along with their standard deviations.
- How do you estimate the error on $f_{H\alpha} = f(\lambda = 6563)$?
- Tempting to assume that m and b are independent, in which case error propagation yields $\sigma_{f_{H\alpha}}^2 = \sigma_b^2 + \lambda^2 \sigma_m^2$

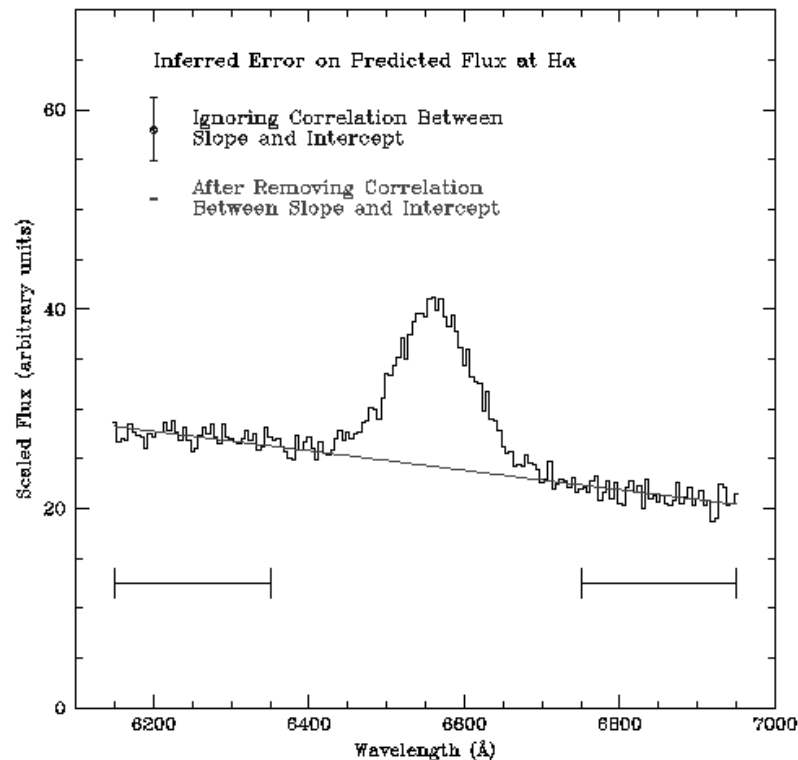


Unfortunately, the resulting error estimate is seriously wrong (overestimated), because m and b are strongly correlated!

Model Fitting and Error Propagation: The Covariance Strikes Back (contd)

So how can we do better?

- 1) We could explicitly account for the correlation by calculating the covariance matrix (i.e. the inverse of the matrix A that appears in the normal equations)
- 2) Easier way: consider the off-diagonal elements of A . If we could arrange things such that these are all zero, then the covariance between slope and intercept would be zero!



Remember that for any linear model $y = \sum_{i=1}^{N_{\text{par}}} a_i f_i(x)$ the elements of A are given by

$$A_{ij} = \sum_{k=1}^{N_{\text{data}}} \frac{f_i(x_k) f_j(x_k)}{\frac{3}{4}^k}$$

Now for a straight-line fit $y = mx + b$, we have

$$f_1 = x \text{ and } f_2 = 1$$

The off-diagonal elements are then simply

$$A_{12} = A_{21} = \sum_{k=1}^{N_{\text{data}}} x_k = \frac{3}{4}^k$$

But (modulo a constant factor) this is just the weighted average of x (or 1 in our example)!!!

So if we shift the x (or 1) values to zero (weighted) mean before the fit, that slope and intercept won't be correlated!