# Statistics notes

## Introductory comments

These notes provide a summary or "cheat sheet" covering some basic statistical recipes and methods. These will be discussed in more detail in the lectures!

*What is the purpose of statistics?* They provide the framework for posing meaningful scientific questions and describing the answers. Some common examples are:

- **Measuring a quantity** *("parameter estimation"):* given some data, what is our best estimate of a particular parameter? What is the uncertainty in our estimate?
- **Searching for correlations:** are two variables we have measured correlated with each other, implying a possible physical connection?
- **Testing a model** *("hypothesis testing"):* given some data and one or more models, are our data consistent with the model(s)? Which model best describes the data?

To do any of these things, we need statistics. Why?

- A clear statistical framework formulates the logic of *what* we are doing and *why*. It allows us to make precise statements.
- A clear statistical framework established in advance helps to prevent *confirmation bias,* where our conclusions are distorted by our preconceived bias about what the result should be.
- A clear statistical framework allows us to quantify the *uncertainty* in any measurement, which should always be stated.

In statistics there are often multiple tests or approaches that have been devised to address the same question (e.g., are two variables correlated? are two samples drawn from the same parent population?) The most appropriate test to use depends on the assumptions underlying the tests (e.g., are the data Gaussian? are they sampled from a parametric model? does the test hold for small sample size?) and to some extent on personal choice. Be cautious if two statistical tests produce conflicting results; it likely shows that the data are not adequate to address the hypothesis.

Broadly speaking there are two schools of statistical thought: the *frequentist* approach works from the probability distribution of particular statistics constructed from the data, and the *Bayesian* approach focuses on the probabilities of the underlying models or hypotheses.

We must always be careful not to use the same dataset for which a hypothesis was proposed to verify that hypothesis (*"a posteriori"* statistics). Testing a hypothesis requires a fresh and unbiased dataset.

The following notes are concerned only with statistical uncertainties arising from random fluctuations. We should never forget the importance of *systematic errors*, which may in fact be dominant in many areas of astronomy.

## Estimating basic descriptive statistics

A *statistic* is a quantity which summarizes or combines our data. A sample $x_i$ of $N$ independent data points may be summarized in some basic ways:

- The mean ("typical value") of the sample:

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

- The standard deviation ("spread" or "scatter") of the sample $\sigma$, closely related to the variance $\sigma^2$:

$$\mathrm{Var}(x) = \sigma^2(x) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2 = \frac{N}{N-1}\left(\overline{x^2} - \overline{x}^2\right) \tag{2}$$

- The error in the estimate of the mean:

$$\sigma(\overline{x}) = \frac{\sigma(x)}{\sqrt{N}} \tag{3}$$

  i.e. the uncertainty in the mean is $\sqrt{N}$ smaller than that in each single measurement. This statement is independent of the probability distribution of $x$.

- The error in the estimate of the variance:

$$\sigma[\mathrm{Var}(x)] = \mathrm{Var}(x)\sqrt{\frac{2}{N-1}} \tag{4}$$

- The error in the median:

$$\sigma[\mathrm{Med}(x)] = 1.25\,\frac{\sigma(x)}{\sqrt{N}} \tag{5}$$

  Note that the last two relations assume a Gaussian probability distribution for $x$.

## Probability distributions

A "probability distribution" is a function $p(x)$ that assigns a probability for each particular value (or range of values) of a random variable $x$. Probability distributions must be *normalized* such that the combined probability of all possible values of $x$ is equal to 1. For example, for a continuous variable:

$$\text{Normalization}: \int_{-\infty}^{\infty} p(x)\,dx = 1 \tag{6}$$

Just as a data sample may be characterized by its mean and variance, a probability distribution may be summarized in a similar manner.

$$\text{Mean} = \mu = \overline{x} = < x > = \int_{-\infty}^{\infty} x\,p(x)\,dx \tag{7}$$

$$\text{Variance} = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2\,p(x)\,dx = < x^2 > - < x >^2 \tag{8}$$

Whenever we quote a measurement of a quantity with an error, we are summarizing the full probability distribution for that parameter. For example, a measurement of the Hubble parameter may be quoted in a paper as $H_0 = 70 \pm 5$ km s$^{-1}$ Mpc$^{-1}$. This statement usually means "there is 68% probability that the value of $H_0$ lies in the range 65 to 75 km s$^{-1}$ Mpc$^{-1}$". Often the stronger implication "The probability distribution for $H_0$ is a Gaussian distribution with mean 70 and standard deviation 5 km s$^{-1}$ Mpc$^{-1}$" will also be intended.

Although it is usually difficult to deduce the full probability distribution of a variable from limited observations, certain types of variable are known to have specific probability distributions:

- The **binomial distribution** applies in problems where there is a random process with 2 possible outcomes with probabilities $p$ and $1-p$, which is repeated $N$ times. Example: tossing a coin. The probability distribution of the number of occurences $n$ of outcome 1 is the binomial distribution:

$$P_{\rm binomial}(n) = \frac{N!}{n!\,(N-n)!}\,p^n(1-p)^{N-n} \tag{9}$$

  The mean and variance of this distribution are $\overline{n} = pN$, $\mathrm{Var}(n) = Np(1-p)$.

- The **Poisson distribution** applies in problems where we are counting something, and describes a discrete random process with some fixed mean. Examples: number of atoms radioactively decaying, photons arriving in a CCD pixel. If $\mu$ is the expected mean number of events, and $n$ is the number observed, the probability distribution of $n$ is the Poisson distribution:

$$P_{\rm Poisson}(n) = \frac{\mu^n \exp\left(-\mu\right)}{n!} \tag{10}$$

  The mean and variance of this distribution are $\overline{n} = \mathrm{Var}(n) = \mu$. Poisson fluctuations are the ultimate limit to any counting experiment. For binned data, if an individual bin contains $n$ events we often place a "Poisson error" in the count in that bin, $n \pm \sqrt{n}$. Note that this assumes that the mean count in the bin is equal to the observed value, $\mu = n$, which is a bad approximation for small $n$ (e.g. $n = 0$). The Poisson error is only an approximation if other processes are also causing numbers to fluctuate – e.g. counting galaxies in a region of space, where galaxy clustering also causes a fluctuation from place-to-place; counts in a CCD pixel, where we must also add read noise.

- The **Gaussian** or **Normal distribution** is important and ubiquitous for two reasons. Firstly, the Gaussian distribution is the "high statistics" limit for both the Poisson and binomial distributions. Secondly, the *central limit theorem* tells us that if we average together $N$ variables *drawn from any probability distribution,* the resulting averages will follow a Gaussian distribution in the limit of high $N$. The Gaussian probability distribution of $x$ for mean $\mu$ and standard deviation $\sigma$ is:

$$P_{\rm Gaussian}(x) = \frac{\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]}{\sigma\sqrt{2\pi}} \tag{11}$$

  The fraction of probability in the ranges $|x-\mu| < (1,2,3)\sigma$ is $(68.27, 95.45, 99.73)\%$. We can use this concept to associate the probability of a statement with an "$N$-$\sigma$ significance".

## Error propagation

Given a function of one or more variables, error propagation allows us to convert measurements and errors of those variables into an error in the function.

- Simplest case: linear function $z = a\,x + b\,y$ of variables $(x, y)$, constants $(a, b)$:

$$\mathrm{Var}(z) = a^2\,\mathrm{Var}(x) + b^2\,\mathrm{Var}(y) \tag{12}$$

- Generalization 1: non-linear function of one variable $z = f(x)$

$$\sigma_z = \left|\frac{\partial f}{\partial x}\right| \sigma_x \tag{13}$$

- Generalization 2: non-linear function of two independent variables $z = f(x, y)$

$$\mathrm{Var}(z) = \left(\frac{\partial f}{\partial x}\right)^2 \mathrm{Var}(x) + \left(\frac{\partial f}{\partial y}\right)^2 \mathrm{Var}(y) \tag{14}$$

  Note: these approximations neglect higher-order terms, effectively assuming $\partial f/\partial x$ is constant over the range $\sigma_x$. They can fail badly in cases such as $z = x/y$ if $\overline{y} \approx 0$!

- Generalization 3: function of two correlated variables

$$\mathrm{Var}(z) = \left(\frac{\partial f}{\partial x}\right)^2 \mathrm{Var}(x) + \left(\frac{\partial f}{\partial y}\right)^2 \mathrm{Var}(y) + 2\frac{\partial f}{\partial x}\frac{\partial f}{\partial y}\mathrm{Cov}(x, y) \tag{15}$$

  in terms of the covariance $\mathrm{Cov}(x, y) =< (x - \overline{x})(y - \overline{y}) >= \rho\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}$, where $\rho$ is the correlation coefficient.

## Optimal combination of data

Say we have $N$ multiple independent estimates $x_i$ of some quantity, with errors $\sigma_i$. What is our best combined estimate? Any linear combination is an unbiased estimate, $y = \sum_{i=1}^{N} w_i x_i$ with weights $w_i$ such that $\sum_{i=1}^{N} w_i = 1$. Error propagation implies that $\mathrm{Var}(y) = \sum_{i=1}^{N} w_i^2\,\mathrm{Var}(x_i)$. The optimal combination gives more weight to the more precise estimates such that

$$w_i = \frac{1/\sigma_i^2}{\sum_{i=1}^{N} 1/\sigma_i^2} \tag{16}$$

("inverse-variance weighting"). The optimal combined estimate is then

$$y = \frac{\sum_{i=1}^{N} x_i/\sigma_i^2}{\sum_{i=1}^{N} 1/\sigma_i^2} \tag{17}$$

and the error in the combined estimate is given by

$$\frac{1}{\mathrm{Var}(y)} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \tag{18}$$

## Searching for correlations

A very frequent task in astronomy is to search for correlations between two variables, $x$ and $y$. What statistical tests are available for quantifying correlation, and assessing its statistical significance?

Mathematically, if the two variables have means $(\mu_x, \mu_y)$, standard deviations $(\sigma_x, \sigma_y)$ and follow a joint probability distribution $P(x, y)$, their *correlation coefficient* $\rho$ is defined by

$$\rho = \frac{< (x - \mu_x)\,(y - \mu_y) >}{\sigma_x\,\sigma_y} = \frac{< xy > - \mu_x\,\mu_y}{\sigma_x\,\sigma_y} \tag{19}$$

where $< xy >= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy\,P(x, y)\,dx\,dy$. If there is no correlation, then $< xy >=< x >< y >= \mu_x \mu_y$ hence $\rho = 0$. If there is complete correlation, then $y = C\,x$ (where $C$ is a positive constant) hence $\rho = 1$. For complete anti-correlation, $y = -C\,x$ hence $\rho = -1$.

Given two sets of $N$ variables $(x_i, y_i)$, we can estimate the correlation coefficient $r$ (sometimes called the "Pearson product-moment correlation coefficient") by calculating:

$$r = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2 \sum_{i=1}^{N}(y_i - \overline{y})^2}} = \frac{\sum_{i=1}^{N} x_i\,y_i - N\,\overline{x}\,\overline{y}}{(N-1)\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}} \tag{20}$$

The value of $r$ lies in the range $-1$ to $+1$. First we must determine *whether or not the observed correlation is statistically significant?* This is particularly important in astronomical applications, where sample sizes can be small. The most common techniques for assessing the statistical significance are described below.

**If** the (anti-)correlation is statistically significant, the size of $r$ indicates its strength. For example, we may refer to $0 \leq |r| \leq 0.3$ as a "weak correlation", $0.3 \leq |r| \leq 0.7$ as a "moderate correlation", and $0.7 \leq |r| \leq 1$ as a "strong correlation".

Sometimes we can assume a model in which the two variables $x$ and $y$ are drawn from a bivariate Gaussian distribution about an underlying linear relation:

$$P(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right\} \tag{21}$$

If we can assume this model, then an estimate of the uncertainty in the measurement of $r$ using the $N$ data pairs is

$$\sigma(r) = \sqrt{\frac{1-r^2}{N-2}} \tag{22}$$

Also assuming this model, we can determine *is this correlation significant* or, what is the probability of obtaining this measured value of $r$ if the true correlation is zero ($\rho = 0$)? This probability can be evaluated by calculating the $t$ statistic

$$t = r\sqrt{\frac{N-2}{1-r^2}} \tag{23}$$

which obeys the *Student's t probability distribution* with $\nu = N - 2$ degrees of freedom. We then need to consult standard tables which list the critical values that $t$ must exceed, as a function of $\nu$, for the hypothesis that the two variables are unrelated to be rejected at a particular level of statistical significance (such as 95% or 99%). If we are satisfied that a correlation is significant,

we may wish to determine the best-fitting linear relation between $x$ and $y$ (using for example the method of least-squares deviation), which is known as the *regression line*. Note that these analyses do not require any knowledge of the errors in the individual data points.

If we do not want to assume that $x$ and $y$ are drawn from a bivariate Gaussian probability distribution, we can use a *non-parametric* correlation test. The most common method is to calculate the *Spearman rank cross-correlation coefficient*

$$r_s = 1 - 6\,\frac{\sum_{i=1}^{N}(X_i - Y_i)^2}{N^3 - N} \tag{24}$$

where $(X_i, Y_i)$ represent the ranks of the $i$th variables in an overall ordering of $x_i$ and $y_i$ such that $1 \leq X_i \leq N$, $1 \leq Y_i \leq N$. Again, we must determine the statistical significance of the correlation. If $N \leq 30$, we consult standard tables which give the critical values $r_s$ must exceed (for a given value of $N$) for the hypothesis that the variables are unrelated to be rejected at a particular level of significance. If $N > 30$, we compute the same $t$ statistic as given above, and compare the value of $t$ to standard tables for $\nu = N - 2$ degrees of freedom.

Some important points to remember when searching for correlations:

- Think about *selection effects:* if some part of the parameter space is unobservable (e.g. falls below a flux threshold in the case of "Malmquist bias") then a false correlation can easily be produced.
- Is the correlation *robust*? – we should be suspicious if a correlation is driven by a small number of outliers (i.e. potentially by systematic errors).
- A correlation does not necessarily imply a cause-and-effect relationship between two variables: a third variable may be involved.

## Comparison of two samples

In astronomy we often want to divide objects into different classes. The *Kolmogorov-Smirnov (KS) test* allows us to determine the probability that two samples are drawn from the same parent population (or conversely, if there is evidence that they are in fact distinct classes). The test is described in basic statistics books; it considers the maximum deviation between the normalized cumulative distributions of the two samples, and converts that deviation into a probability. We note that statistical tests are also available comparing the means or variances of the two populations (Student's $t$ test, $F$ test).

## The $\chi^2$ statistic and its uses

We are often gathering data in order to test a hypothesis or constrain a model. When comparing data and models, we are typically doing one of two things:

- *Hypothesis testing:* we perform a set of $N$ measurements $x_i \pm \sigma_i$, which a theorist says should have values $\mu_i$. How probable is it that a set of measurements such as these would have been obtained if the theory is correct?
- *Parameter estimation:* We have a parameterized model which describes the data, e.g. $y = ax + b$, and we want to determine the best-fitting parameters and errors in these parameters.

To help answer both of these questions we can use the $\chi^2$ *statistic* as a measure of the goodness-of-fit of the data to the model:

$$\chi^2 = \sum_{i=1}^{N} \left( \frac{x_i - \mu_i}{\sigma_i} \right)^2 \tag{25}$$

The intuition behind this statistic is we are penalizing data points according to how many standard deviations they lie away from the model prediction. If the variables $x_i$ result from a counting experiment, the error may be taken as a Poisson error $\sigma_i^2 = \mu_i$.

## Hypothesis testing

We first address the question of whether the data can be considered to be consistent with the model, such that we are testing the null hypothesis that $x_i = \mu_i$. If this probability is low, then the model can be "ruled out" in some sense. If the probability is not low, then the model is "ruled in" and has passed the test. We note here that a more sophisticated discussion of this point would involve Bayesian inference (see below).

If we can assume that the $N$ variables $x_i$ are Gaussian-distributed, then $\chi^2$ has a particular probability distribution:

$$P(\chi^2) \propto \left(\chi^2\right)^{\frac{\nu-2}{2}} \exp\left(-\chi^2/2\right) \tag{26}$$

where $\nu$ is the number of *degrees of freedom*. If the model has no free parameters, then $\nu = N$. If we are fitting a model with $p$ parameters to $N$ data points we can "force the model to exactly agree" with $p$ data points. The number of degrees of freedom is then reduced to $\nu = N - p$.

The mean and variance of this distribution are $\overline{\chi^2} = \nu$, $\mathrm{Var}(\chi^2) = 2\nu$. This mean value makes intuitive sense because we expect that each data point should lie about 1-$\sigma$ from the model and hence contribute on average 1.0 to the $\chi^2$ statistic. Thus if the model is correct, we expect $\chi^2 \sim \nu \pm \sqrt{2\nu}$.

We use the $\chi^2$ distribution to perform the hypothesis test by asking the question: *if the model is correct, what is the probability that this value of $\chi^2$, or a larger one, could arise by chance?* This probability, sometimes called the *p-value*, may be calculated from the $\chi^2$ distribution given the values of $\chi^2$ and $\nu$ using standard libraries. Some examples:

- $(\chi^2 = 37.5, \nu = 30) \rightarrow p = 0.163$
- $(\chi^2 = 52.1, \nu = 30) \rightarrow p = 0.007$

Some notes on the interpretation of the $\chi^2$ and $p$-values:

- If the $p$-value is not very low, then the data are consistent with being drawn from the model.
- If $\chi^2$ is "improbably high" (the $p$-value is very low) – either the model can be rejected, or the errors have been under-estimated.
- If $\chi^2$ is "improbably low" – either the errors have been over-estimated, or there are too many free parameters (for example, the number of degrees of freedom is zero!)
- Since it is usually difficult to reliably determine errors, a model is typically only rejected for very low $p$, e.g. $< 0.01$.
- As a way of summarizing the measurement we often quote a *reduced chi-squared* value $\chi^2/\nu$ – the expected value for a good fit is around 1.0, but the true probability must be described by quoting both $\chi^2$ and $\nu$, rather than just $\chi^2/\nu$ (this is particularly true when $\nu$ is small).
- *If the data points are correlated:* The number of degrees of freedom is unchanged, for all correlation coefficients $< 1$. The equation for $\chi^2$ is modified to

$$\chi^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} y_i \left(C^{-1}\right)_{ij} y_j \tag{27}$$

  where $y_i = x_i - \mu_i$ are the differences between the data and the model, $C_{ij}$ is the covariance matrix of the data, and $C^{-1}$ is the inverse covariance matrix.
- *Comments about the interpretation of the p-value.* Suppose we perform a $\chi^2$ test for a model and find $p = 0.01$. This implies *there is a 1% chance of obtaining a set of measurements at least this discrepant from the model* **assuming the model is true.** We may be tempted to summarize our conclusions by making one of the following statements, **all of which are incorrect**:
  - The probability that the model is true is 1%.
  - The probability that the model is false is 99%.
  - If we reject the model, there is a 1% chance that we would be mistaken.
  - There is a 99% chance that a replicating experiment would yield the same conclusion.

  The reason that these statements cannot be made is that significance tests can only compute the likelihood of a dataset arising from a given model; they cannot assess the probability that the model itself is correct. *Bayesian inference* can be used to assess the probability that a model is true, as discussed below.

A disadvantage of using $\chi^2$ statistics for hypothesis testing that can occur is that the raw data are often binned before comparing with the model. For example, suppose we have a data sample of galaxy luminosities; to determine the best-fitting Schechter function parameters by minimum $\chi^2$ we bin the data into a luminosity function. The binning of data loses information, can cause bias if the bin sizes are large compared to changes in the function, and if the bin numbers are too small then the probability distribution of the variables can become non-Gaussian, invalidating the usual confidence intervals.

## Parameter estimation

A model typically contains free parameters. We often want to determine the most likely values of these parameters, and their error ranges.

We will take an example of a model with two free parameters, $a$ and $b$. The most likely ("best-fitting") values of $a$ and $b$ can be determined by varying these parameters to find the *minimum* $\chi^2$ *statistic* of the data, which we write as $\chi^2_{\rm min}$.

The joint error distribution of $a$ and $b$ can be determined by calculating the values of $\chi^2$ for a grid of $a$ and $b$ values, where the grid spans a parameter range much wider than the eventual errors in $a$ and $b$. Within this 2D grid we can plot contours of constant $\chi^2 = \chi^2_{\rm min} + \Delta\chi^2$. A joint "confidence region" for $a$ and $b$ can be defined by the zone in this parameter space that satisfies $\chi^2 < \chi^2_{\rm min} + \Delta\chi^2$. Assuming the variables are Gaussian-distributed, 68% and 95% confidence regions are enclosed by $\Delta\chi^2 = 2.30$ and 6.17. (These values assume two free parameters, the values for other numbers of parameters are given in standard tables).

We may also wish to quote the 1D probability distribution for parameter $a$, considering all possible values for parameter $b$. This is known as *marginalization* of parameter $b$, and may be carried out using the relation between $\chi^2$ and likelihood that holds for Gaussian variables:

$$\text{Likelihood} \propto \exp\left(-\chi^2/2\right) \tag{28}$$

The process is as follows:

- Convert the $\chi^2$ grid into a probability grid $P_{\rm 2D}(a,b) \propto \exp\left(-\chi^2/2\right)$. Normalize the probabilities such that $\sum_{a,b} P_{\rm 2D}(a,b) = 1$.
- Produce the *marginalized* probability distribution for each individual parameter by summing the probability grid over all the other parameters, for example $P_{\rm 1D}(a) = \sum_b P_{\rm 2D}(a,b)$. Here we are determining the probability distribution of $a$ given all possible values of $b$.
- Use the shape of this 1D probability distribution to quote a confidence range (e.g. 68%) for each parameter.

If the probability distribution of a parameter is Gaussian, the best-fitting value is equal to the mean of the distribution, and the 68% confidence error range is equivalent to the range $\pm 1\sigma$. However, in general the shape of the probability distribution for a parameter may be non-Gaussian or asymmetric. This implies that there is more than one method for quoting our measured parameter value and its error.

- For the measured value of the parameter we may choose to quote the best-fitting (maximum-likelihood) value, or if we have calculated the probability distribution, we may choose to quote the mean value $\overline{a} = \int_{-\infty}^{\infty} a\, P_{\rm 1D}(a)\, da$.
- For the error we may choose to quote the standard deviation $\sigma_a$ of the parameter such that $\sigma_a^2 = \int_{-\infty}^{\infty} (a-\overline{a})^2\, P_{\rm 1D}(a)\, da$. Alternatively we may quote an asymmetric error $a^{+atop}_{-abot}$. In this case we can use the probability distribution to determine $a_{\rm top}$ and $a_{\rm bot}$ such that 68% of the probability lies in the range $a_{\rm bot} < a < a_{\rm top}$.

## Bayesian inference

In the preceding section we calculated the likelihood of some data given a model, $P(\text{data}|\text{model})$. (In this notation $P(A|B)$ means "the probability of $A$ given $B$"). However, most often in science we really want to do the reverse: calculate the likelihood of a model being true given the data: $P(\text{model}|\text{data})$. *Conditional probability* tells us that these two quantities are in general not the same.

As a simple demonstration of conditional probability, consider the following statement: "in the general population the probability of a randomly-selected woman being pregnant is 2%", or $P(\text{pregnant}|\text{woman}) = 0.02$. Clearly it does not follow that $P(\text{woman}|\text{pregnant}) = 0.02$, since of course $P(\text{woman}|\text{pregnant}) = 1$.

To relate conditional probabilities we need to use *Bayes' theorem:*

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \tag{29}$$

Applied to our simple example:

$$P(\text{woman}|\text{pregnant}) = \frac{P(\text{pregnant}|\text{woman})P(\text{woman})}{P(\text{pregnant})} = \frac{0.02 \times 0.5}{0.01} = 1 \tag{30}$$

The application of Bayes' theorem allows us to calculate

$$P(\text{model}|\text{data}) = \frac{P(\text{data}|\text{model})P(\text{model})}{P(\text{data})} \tag{31}$$

In this relation, $P(\text{model}|\text{data})$ is called the *posterior probability* of the model. $P(\text{data}|\text{model})$ is called the *likelihood function* which can be evaluated using the $\chi^2$ method described above. $P(\text{model})$ is the *prior probability* of the model, describing what we knew before the experiment. The denominator $P(\text{data})$ is usually not important as it is absorbed as a constant into the normalization of the posterior probability distribution.

The Bayesian framework makes explicit that a prior probability function is needed when assessing the likelihood of models given data. This could also be considered a disadvantage of the framework, in cases where the prior is subjective.

The evaluation of the model likelihood allows us to perform *model selection.* The *Bayes factor* between two models $M_1$ and $M_2$ given data $D$ is defined by

$$K = \frac{P(M_1|D)}{P(M_2|D)} \tag{32}$$

The size of $K$ quantifies how strongly we can prefer one model to the other (described, for example, by the "Jeffreys scale").

When fitting a model to data we sometimes need to ask *is adding another parameter justified by the data?* For example, what order of polynomial can be determined? Tests are available which answer this question on the basis of the $\chi^2$ statistic and the number of free parameters $p$ in the model. Two common criteria are:

- The *Bayesian information criteria:* minimize $BIC = \chi^2 + p \log N$
- The *Akaike information criteria:* minimize $AIC = \chi^2 + 2p$.

Both of these criteria penalize models with large numbers of parameters.

## More advanced techniques of error estimation

Consider a scatter-plot of two correlated variables $(x, y)$ to which we wish to fit a relation $y = ax + b$ by linear least-squares. It is hard to directly determine the errors in $a$ and $b$ if the individual data points do not have errors (and so we cannot define a likelihood). However, we can use the data itself to estimate these errors with a *bootstrap* approach, using the following procedure:

- Determine the best-fitting $(a, b)$ using linear least-squares.
- Generate a new dataset of $N$ points by sampling the old dataset *with replacement*.
- Repeat the least-squares fit for $(a, b)$, obtaining slightly different values.
- Repeat for many new datasets: the scatter of the best-fitting $a$ and $b$ across the re-samplings defines the errors.

If our dataset can be split into independent subsets (e.g. time series, or by distribution on the sky), then another technique is available for error estimation known as *jack-knife re-sampling*.