Statistics and Expectations

"There are three kinds of lies: lies, damned lies, and statistics."

Benjamin Disraeli (1804-81)
attrib (erroneously?) by Mark Twain
(who was really Samuel Clemens anyway)

Where were we?

- generating random numbers

- pitfalls emphasized – do what you're told!

- random numbers from a frequency distribution: transform method

- random numbers from a frequency distribution: rejection method

- Monte Carlo integration introduced in its most basic form - as a prelude to how to do it right

- Sorting, indexing, ranking, etc.

Statistics - what and why

- enormously developed for the Gaussian distribution in particular.
- classical territory :
  statistics were developed **because the Bayesian approach fell out of favour**
- direct probabilistic inferences were superseded by the indirect,
  going through statistics and intimately linked to hypothesis-testing.
- these alternatives to Bayes methods are subtle and not very obvious.
- here I avoid the math, presenting results and showing the use of statistics.
- I concentrate on conceptual foundations.

Statistics are designed to summarize, reduce or describe data.

The formal definition of a **statistic** is that it is some function of the data alone. For a set of data X_1,X_2 ..... , some examples of **statistics** might be the average, the maximum value, or the average of the cosines.

Statistics are combinations of finite amounts of data.

Statistics what and why, continued ....

The summarizing aspect of statistics – e.g. (a) **location** and (b) **spread** or **scatter**.

(a) Location: various combinations –

Average, denoted by overlining: $\overline{X} = 1/N \sum_{i=1}^{N} X_i$

Median: arrange X_i according to size; renumber. Then $X_{\rm med} = X_j$ where $j = N/2+0.5$, N odd, $X_{\rm med} = 0.5(X_j + X_{j+1})$ where $j = N/2$, N even.

Mode: $X_{\rm mode}$ is the value of x_i occurring most frequently; it is the location of the peak in the histogram of X_i.

(b) Spread: likewise -

Mean deviation: $\overline{\Delta X} = \frac{1}{N} \sum_{i=1}^{N} \mid X_i - X_{\rm med} \mid$.

Mean square deviation: $S^2 = \frac{1}{N} \sum_{i=1}^{N} (X_i - \overline{X})^2$

Root mean square deviation: $\mathrm{rms} = S$

Statistics - what and why, continued, continued ....

We think we are familiar ground; we think we 'know' what "D = 8.3 +/- 0.1 Mpc" means.

- we assume a Gaussian distribution, with faith in the Central Limit theorem.
- then 'knowing' the distribution of the errors allows probabilistic statements.

=> one key aspect of statistics: they are associated with distributions!

=> most useful when they are estimators of the parameters of distributions:

8.3 is an estimate of the parameter μ of some Gaussian

0.1 is an estimate of σ.

Statistics - what and why, continued, continued ....

=> 2nd key aspect of statistics: interpretation is in a classical, not Bayesian framework.

=> Serious difference!

Assuming a true distance D_0, classical analysis tells us that D is (say) Normally distributed around D_0, with a standard deviation of 0.1.

We are to imagine many repetitions of our experiment, each yielding a value of the estimate D which dances around D_0: form a confidence interval (like [8.2, 8.4]) which will also dance around randomly, but will contain D_0 with a probability we can calculate.

=> Thus this approach ***assumes the thing we want to know, and tells us how the data will behave (!!!!!!! Is this what we want?????)***

Statistics - what and why, even more continued ....

- **Bayesian approach circumvents all this!**
- deduces directly the probability distribution of D_0 from the data.
- assumes the data, and tells us the thing we want to know.
- no imagined repetitions of the experiment.
- conceptually clearer than classical methods.

- but these are so well developed and established (particularly for the Gaussian) that we need to know how to handle them.

¤ **Remember that statistics of known usefulness are quite rare.**

¤ In many cases of astronomical interest we may need to derive useful statistics for ourselves.

¤ Maximum Likelihood is by far the easiest method.

¤ So close to a Bayesian method that we may expect to be doing Bayesian, not classical, inference.

Expectation values

- Statistics are properties of the data and only of the data; they summarize, reduce, or describe the data.

- Variables such as **μ** and **σ** of the Poisson and Gaussian distributions define these distribution functions and are **NOT** statistics.

- We may anticipate that our data do follow these or other distributions

- We may therefore wish to relate statistics from the data to parameters describing the distributions.

- This is done through **Expectations** or **Expectation values**, average properties depending on distribution functions. The expectation **E[f(x)]** of some function **f** of a random variable **x**, with distribution function **g**, is defined as

  $$E[f(x)] = \int f(x) g(x) \, dx$$

  i.e. the sum of all possible values of f, weighted by the probability of occurrence.

- We can think of the expectation as being the result of repeating an experiment many times, and averaging the results.

Expectation values, continued ....

- For example, compute an average value of <X>. If we repeat the experiment many times, we will find that the average of <X> will converge to the true mean value, the expectation of the function $f(x)=x$:

$$E[x] = \int x g(x) dx.$$

- Note that the expectation is not to be understood as referring to a very large sample; we can ask for the expectation value of a combination of a finite number of data.

- For example the statistic **S²** should likewise converge to the variance, defined by

$$\begin{aligned} \mathrm{var}[(x-\mu)^2] &= E[(x-\mu)^2] && (1) \\ &= \int (x-\mu)^2 g(x) dx. && (2) \end{aligned}$$

- However, we do have to take some care that the integrals actually exist.

Expectation values - Example

Take a Gaussian. Probability of getting a datum **x** near **μ** is

$$g(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(x-\mu)^2}{2\sigma^2}$$

But what are the parameters **μ** and **σ**? Can easily show

$$E[x] = \int x g(x \mid \mu, \sigma)\, dx = \mu,$$

$$E[(x-\mu)^2] = \int (x-\mu)^2 g(x \mid \mu, \sigma)\, dx = \sigma^2.$$

⇒Expect average **<X>** and mean square deviation **S^2** would be related to **μ**, **σ^2**

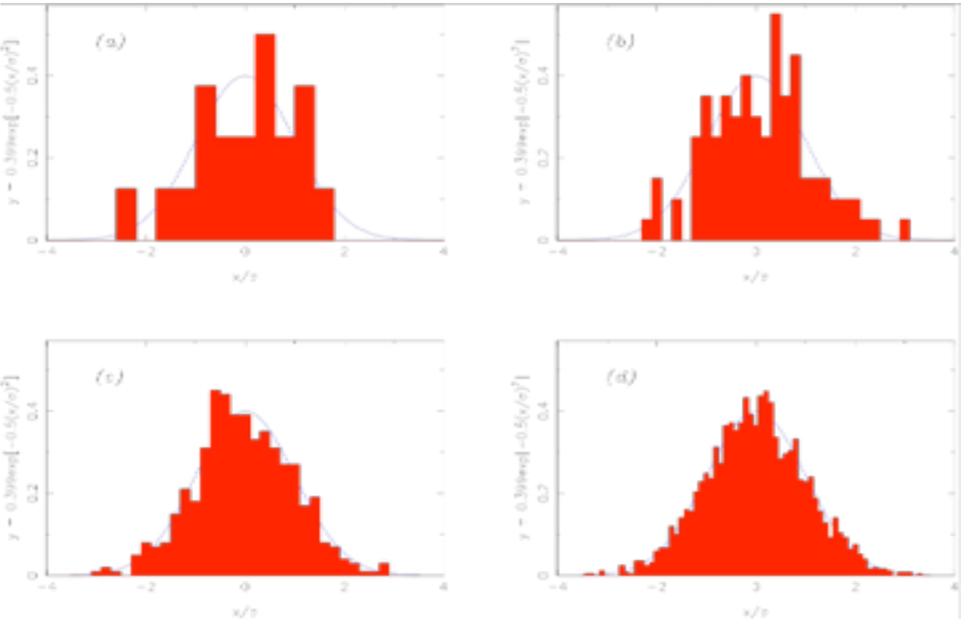
⇒In fact (<X>,S^2) (functions of data alone) converge to (μ, σ^2) when we have plenty of data.

Waht do we want from our statistics?

We have a few of the data **X_i** but we want to know about all of them.

We want their probability or frequency distribution **cheaply (efficiently)** and **accurately (robustly, unbiased).**

Example: draw samples from a population obeying a Gaussian defined by μ= 0, σ= 1. How does size of sample affect estimates?



(a) 20 values, (b) 100 values, (c) 500 values, (d) 2500 values. The average values are 0.003, 0.080, -0.032 and -0.005; the median values 0.121, 0.058, -0.069 and -0.003; and the rms values 0.968, 1.017, 0.986, and 1.001. Solid curves represent Gaussians of unit area and standard deviation.

What do we want from our statistics? cont ....

Thus at least four requirements:

♫They should be **unbiased**, meaning that the expectation value of the statistic turns out to be the true value.

♫ They should be **consistent**, the case if the descriptor for arbitrarily large sample size gives the true answer.

♫ The statistic should obey **closeness**, yielding smallest possible deviation from the truth.

♫ The statistic should be **robust**.

Examples of shaky statistics

Bias: for the Gaussian distribution, **<X>** is an unbiased estimate of the mean **μ**, but the unbiased estimate of the variance **σ^2** is

$$\sigma_s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2$$

which differs from the expectation value of **S^2** by the factor **N/(N-1). σ_s^2**, sometimes called the **sample variance**, is the estimator for the **population variance σ^2**.

(The difference is understandable as follows. The **X_i** of our sample are first used to get **<X>**, an estimate of **μ**, and although this is an unbiased estimate of **μ**, it is the estimate which yields a **minimum** value from the sum of the squares of the deviations of the sample, and thus a low estimate of the variance. The theory provides the appropriate correction factor **N/(N-1)**; of course the difference disappears as **N** becomes large.)

Examples of shaky statistics 2

Thus the **rms** is a **consistent** measure of the standard deviation of a Gaussian distribution in that it gives the right answer for large **N** but it is a **biased** estimator for small **N** unless modified as above.

The Cauchy distribution looks somewhat similar to a Gaussian. But with infinite variance, estimating dispersion via the standard deviation yields massive scatter and little info.

Robustness: consider a symmetric distribution with a few outliers (errors?). As a measure of central location the **median** is far more robust than the **average**.

Also consider salaries – mean vs median.

Error analysis - random or systematic?

The average is a very common statistic; it is what we are doing all the time, for example, in `integrating' on a faint object. The variance:

$$S_m^2 = E[(\frac{1}{N}\sum_{i=1}^{N} X_i - \mu)^2]$$

…and after some fiddling

$$S_m^2 = \frac{\sigma^2}{N} + \frac{1}{N^2}\sum_{i\neq j} E[(X_i - \mu)(X_j - \mu)].$$

The first term expresses generally-held belief : the error on the mean of some data decreases like √N, as the amount of data is increased. This is one of the most important tenets of observational astronomy.

Error analysis - random or systematic? continued..

But apart from infinite variances (e.g. the Cauchy distribution), the $\sqrt{}$N result holds only when the last term is zero. The term contains the covariance, defined as

$$\mathrm{cov}[X_i, X_j] = E[(X_i - \mu_i)(X_j - \mu_j)];$$

it is closely related to the correlation coefficient between x_i and x_j.

In the simplest cases, the data are independent and identically distributed [(probability of x_i and $x_{j)}$ = (probability of x_i) X probability of x_j)].

=>covariance is zero.

This is a condition (probably the likeliest) for the $\sqrt{}$N averaging away of noise.

If it holds, errors are called ‘random’.

If not – ‘systematic’ – but there’s a continuum.

Error propagation

Often the thing we need to know is some function of the measured data. Knowing data error, how do we estimate error in the desired quantity?

If the errors are small, by far the easiest way is to use a Taylor expansion. Measure variables **x,y,z...** with independent errors **δX, δY, δZ...,** and we want some function **f(x,y,z...).** Change in **f** caused by the errors is, to first order

$$\delta F = \frac{\partial f}{\partial x}\,|_{x=X}\,\delta X + \frac{\partial f}{\partial y}\,|_{y=Y}\,\delta Y + \frac{\partial f}{\partial z}\,|_{z=Z}\,\delta Z + ...$$

The variance on a sum is the sum of the variances of the individual terms (because the errors are assumed to be independent) so we get

$$\mathrm{var}[f] = (\frac{\partial f}{\partial x})^2\,|_{x=X}\,\sigma_x^2 + (\frac{\partial f}{\partial y})^2\,|_{y=Y}\,\sigma_y^2 + (\frac{\partial f}{\partial z})^2\,|_{z=Z}\,\sigma_z^2 + ...$$

where the **σ** represent the variances in each of the variables.

Error propagation - Examples

1. This leads to a well-known result for combining measurements: if we have **n** independent estimates, say **X$_j$**, each having an associated error **σ$_j$**, the best combined estimate is the **weighted mean**

$$\overline{X}_w = \sum_{j=1}^{n} w_j \overline{X}_j / \sum_{j=1}^{n} w_j$$

where the weights are given by **$w_j = 1/\sigma_j^2$**, the reciprocals of the sample variances. The **variance of the combined estimate** is

$$\sigma_w^2 = 1/\sum_{j=1}^{n} 1/\sigma_j^2 .$$

2. Suppose **f(x,y)=x/y.** Then the rule gives us immediately

$$\frac{\mathrm{var}[f]}{f^2} = (\frac{\sigma_x}{x})^2 + (\frac{\sigma_y}{y})^2;$$

=> we simply add up the relative errors.

3. If **f(x)=log(x)** then the rule gives

$$\mathrm{var}[f] = (\frac{\sigma_x}{x})^2$$

and the error in the log is just the relative error in the quantity we have measured.

Combining distributions

But we may need to know details of the **probability distribution** of the derived quantity.

The simplest case is a transformation from the measured **x**, with probability distribution **g**, to some derived quantity **f(x)** with probability distribution **h**. Since probability is conserved, we have the requirement that

$$h(f)\,df = g(x)\,dx$$

so that **h** involves the derivative **df/dx.** Beware if **f** is not monotonic!

This technique rapidly becomes difficult to apply for more than one variable.

Combining distributions 2

Results for some useful cases:

1. Suppose we have two identically-distributed independent variables **x** and **y,** both with distribution function **g**. What is the distribution of their sum **z=x+y**? For each **x**, we have to add up the probabilities of the all the numbers **y=z-x** that yield the **z** we are interested in. The probability distribution **h(z)** is therefore

$$h(z) = \int g(z-x)g(x)\,dx$$

where the probabilities are simply multiplied because of the assumption of independence. **h** is the **autocorrelation** of **g**. The result generalizes to the sum of many variables, and is often best calculated using the Fourier transform of the distribution **g**.

This transform is called the **characteristic function**.

Combining distributions 3

2. We often need the distribution of the product or quotient of two variables. Without details, the results are as follows:

For **z=xy**, the distribution of **z** is

$$h(z) = \int \frac{1}{|x|} g(x) g(z/x)\, dx$$

For **z=x/y**, the distribution of **z** is

$$h(z) = \int |x| g(x) g(zx)\, dx.$$

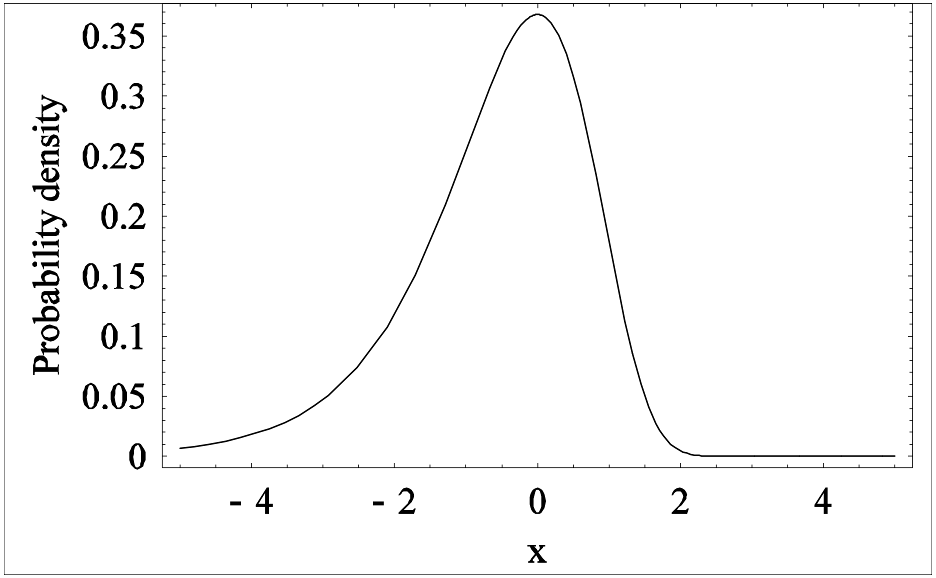
In almost any case of interest, these integrals are too hard to do analytically.

Combining distributions - Example

Suppose we are taking the logarithm of some exponentially-distributed data. Here **g(x)=exp(-x)** for positive **x**, and **f(x)=log(x).** Applying our rule gives

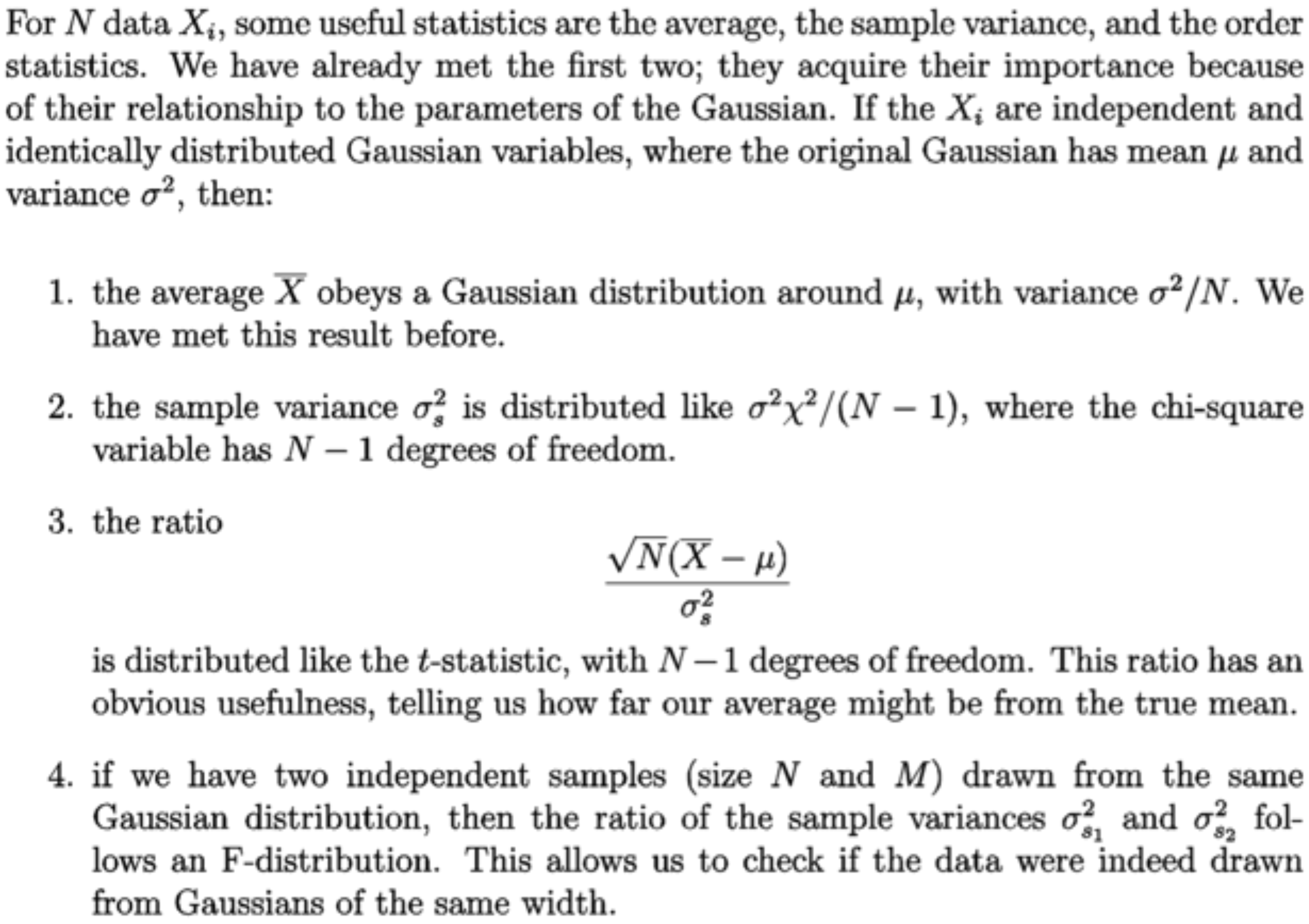
$$h(f) = \exp(-\exp(f))\exp(f)$$

which has a pronounced tail to negative values and is correctly normalized to unity. Our simpler methods would give us **δh = δx/x**, which cannot give a good representation of the asymmetry of **h**. Quoting ``**h ± δh**" is clearly not very informative.



The probability distribution of logarithm of data drawn from an exponential distribution.

Some statistics and their distributions



For N data X_i, some useful statistics are the average, the sample variance, and the order statistics. We have already met the first two; they acquire their importance because of their relationship to the parameters of the Gaussian. If the X_i are independent and identically distributed Gaussian variables, where the original Gaussian has mean μ and variance σ^2, then:

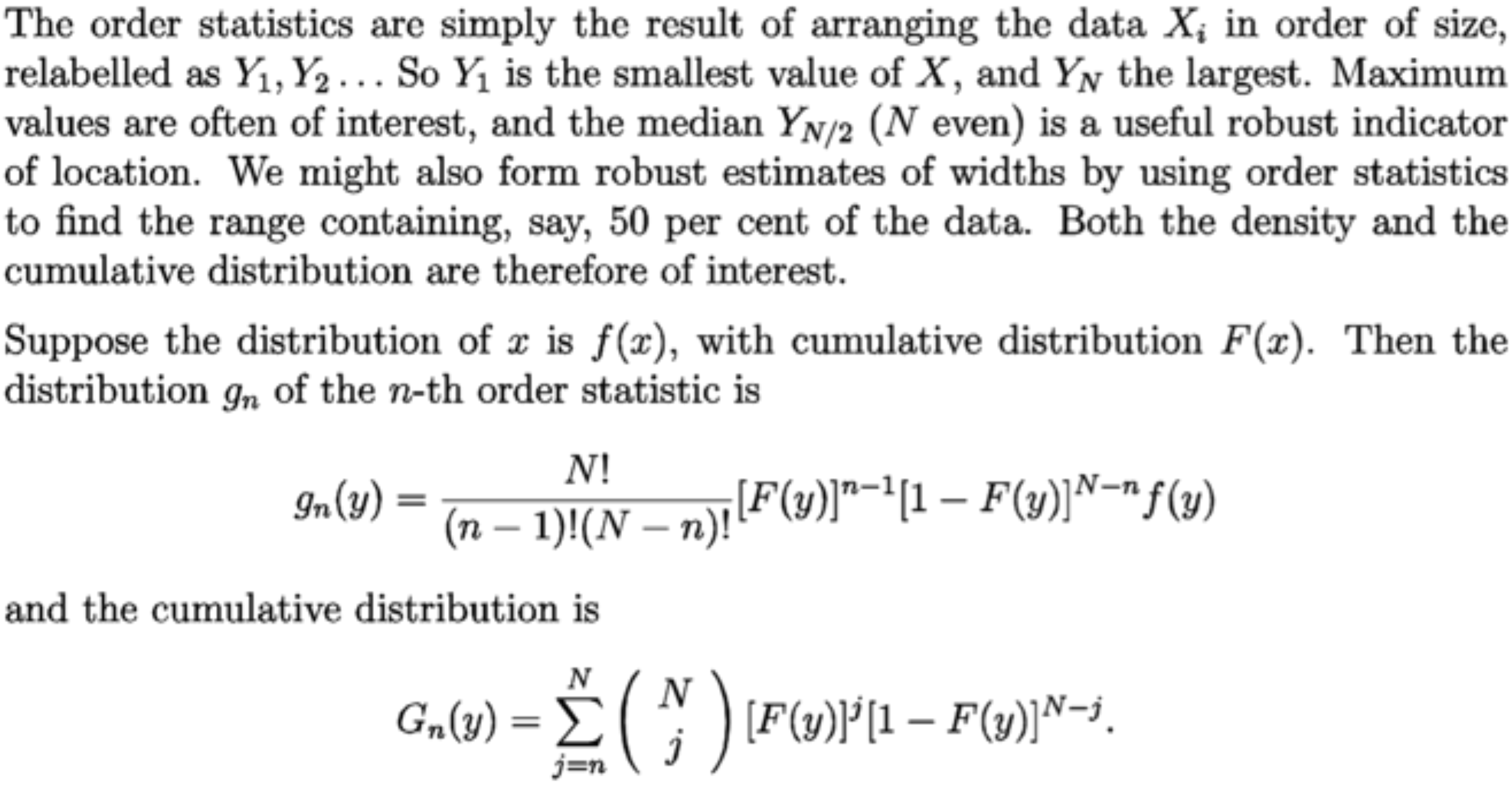
1. the average $\overline{X}$ obeys a Gaussian distribution around μ, with variance σ^2/N. We have met this result before.

2. the sample variance σ_s^2 is distributed like $\sigma^2\chi^2/(N-1)$, where the chi-square variable has $N-1$ degrees of freedom.

3. the ratio
$$\frac{\sqrt{N}(\overline{X}-\mu)}{\sigma_s^2}$$
is distributed like the t-statistic, with $N-1$ degrees of freedom. This ratio has an obvious usefulness, telling us how far our average might be from the true mean.

4. if we have two independent samples (size N and M) drawn from the same Gaussian distribution, then the ratio of the sample variances $\sigma_{s_1}^2$ and $\sigma_{s_2}^2$ follows an F-distribution. This allows us to check if the data were indeed drawn from Gaussians of the same width.

Some statistics and their distributions 2



The order statistics are simply the result of arranging the data X_i in order of size, relabelled as $Y_1, Y_2 \ldots$ So Y_1 is the smallest value of X, and Y_N the largest. Maximum values are often of interest, and the median $Y_{N/2}$ (N even) is a useful robust indicator of location. We might also form robust estimates of widths by using order statistics to find the range containing, say, 50 per cent of the data. Both the density and the cumulative distribution are therefore of interest.

Suppose the distribution of x is $f(x)$, with cumulative distribution $F(x)$. Then the distribution g_n of the n-th order statistic is

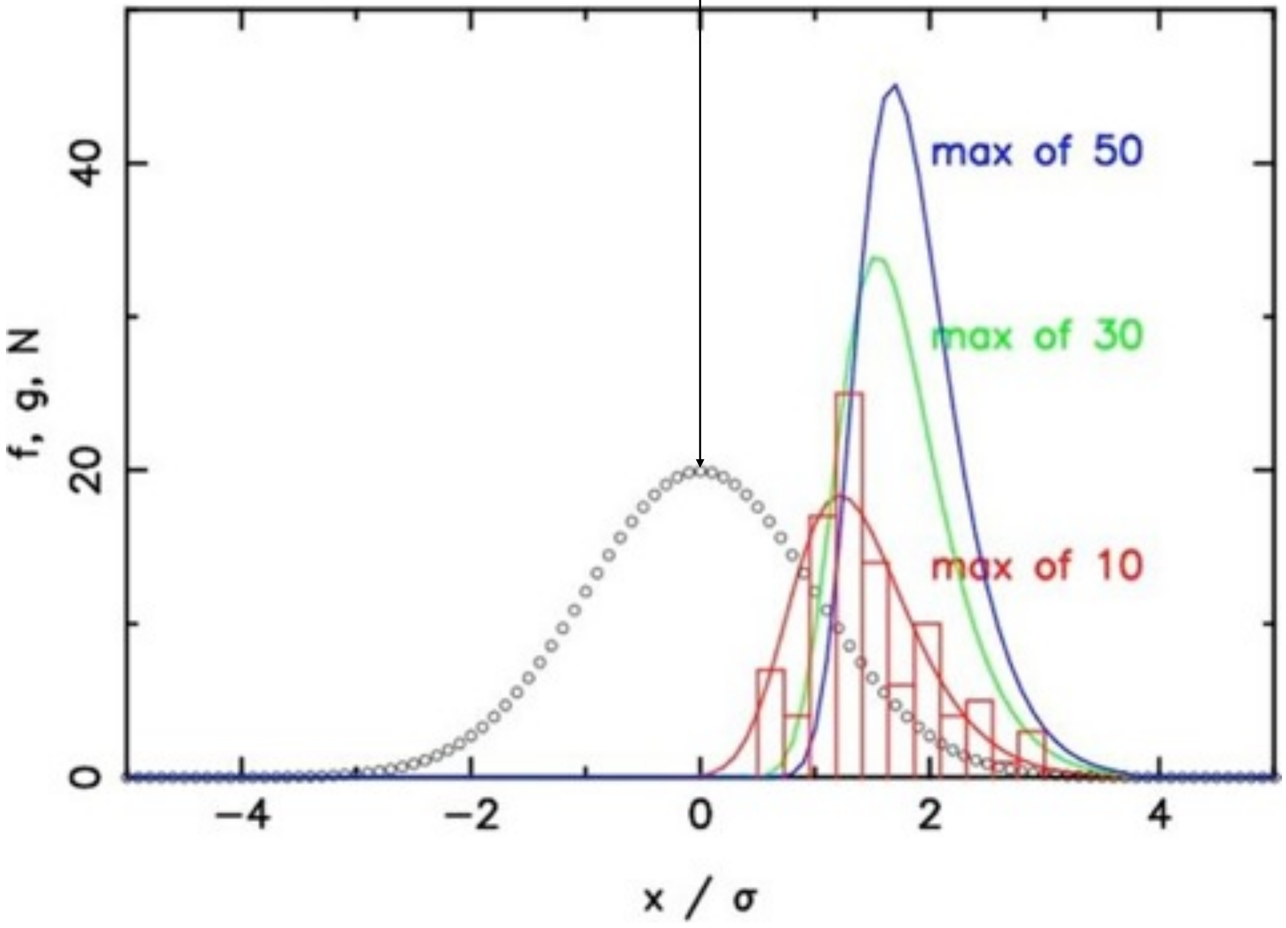
$$g_n(y) = \frac{N!}{(n-1)!(N-n)!}[F(y)]^{n-1}[1-F(y)]^{N-n}f(y)$$

and the cumulative distribution is

$$G_n(y) = \sum_{j=n}^{N} \left(\begin{array}{c} N \\ j \end{array} \right) [F(y)]^j[1-F(y)]^{N-j}.$$

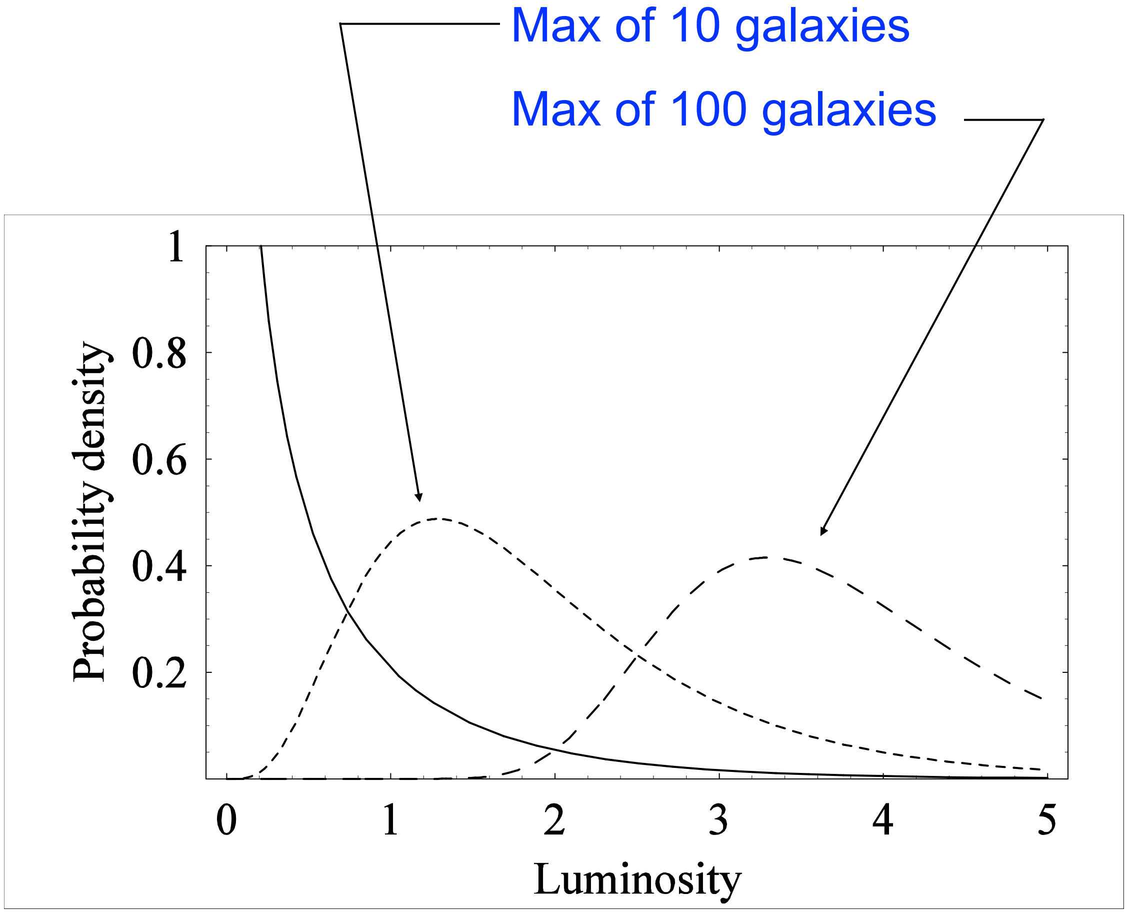
Order statistics - Example 1

Gaussian parent
Select max of 10, 30, 50:



Order statistics - Example 2

Solid curve: A Schechter luminosity function ***$x^{\gamma}exp[-x/x^*]$,*** a useful model for the luminosity function of field galaxies. Take ***$\gamma = 0.5$, $x^* = 1$***.



Max of 10 galaxies

Max of 100 galaxies