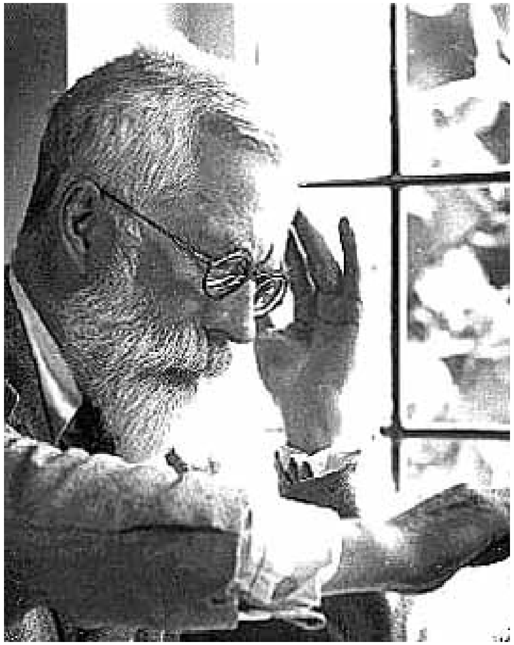
Bertinoro 4 (JVW)

Hypothesis testing

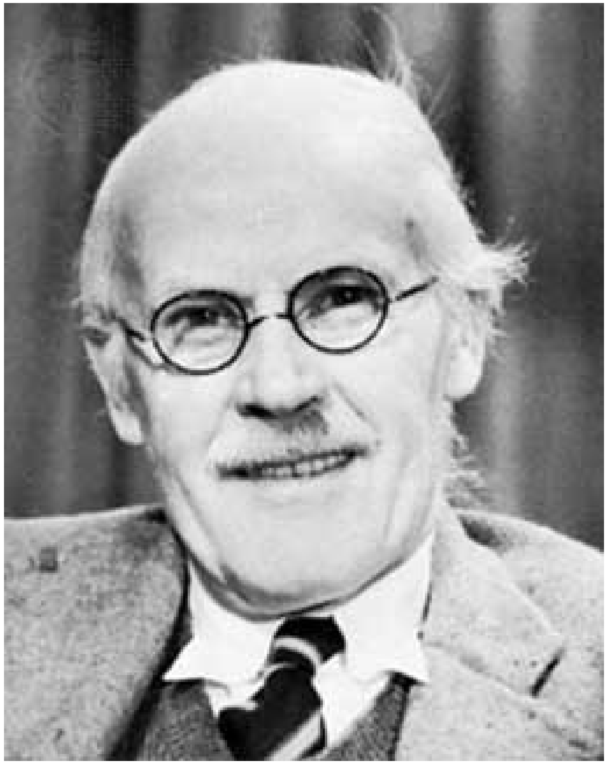
Fisher, Sir Ronald Aylmer, 1890 - 1966.



'.. if he had stuck to the ropes he would have made a first class mathematician, but he would not.'

- his tutor at Cambridge

Data modelling - The Bayesian Way



Sir Harold Jeffreys, 1891-1989
Fellow, St John's College, Cambridge 1914-89

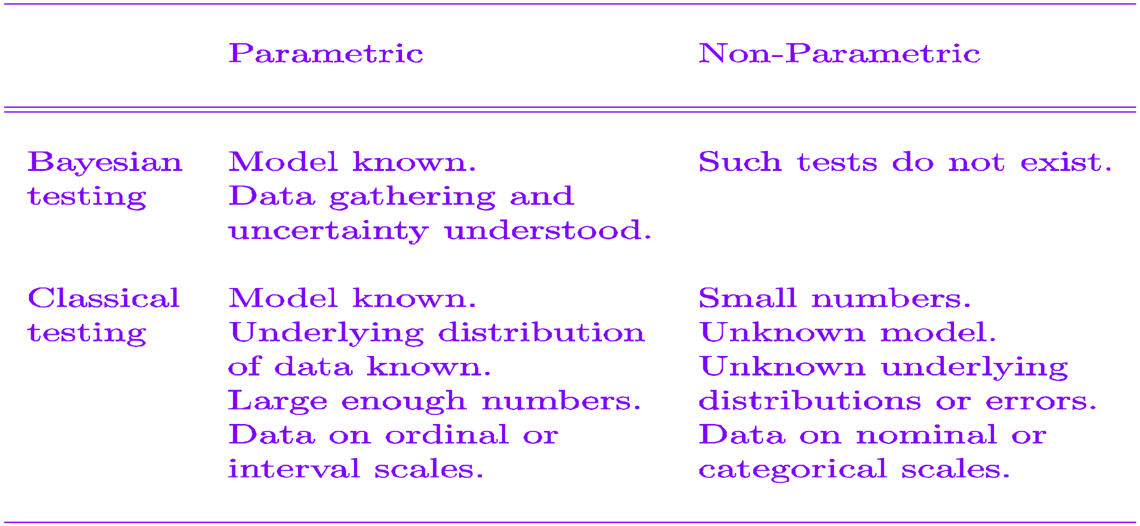
Plumian Professor of Astronomy, after many years of teaching maths.
The first to claim a liquid core for the earth.
>400 papers on celestial mechanics, fluid dynamics, meteorology, geophysics, probability plus these books:

The Earth: Its Origin, History and Physical Constitution (1924),
Theory of Probability (1939)
Methods of Mathematical Physics (1946)

'Fisher and Jeffreys first took serious notice of each another in 1933. About all they knew of each other's work was that it was founded on a flawed notion of probability.'

Bayes/frequentist/parametric/non-parametric?

The essential divide:



	Parametric	Non-Parametric
Bayesian testing	Model known. Data gathering and uncertainty understood.	Such tests do not exist.
Classical testing	Model known. Underlying distribution of data known. Large enough numbers. Data on ordinal or interval scales.	Small numbers. Unknown model. Unknown underlying distributions or errors. Data on nominal or categorical scales.

- non-parametric Bayesian tests do not exist (more or less).

- If we understand the data so that we can model its collection process, then

GO BAYES.

Rejection; elimination

There are situations when classical methods are essential:

1. If we are comparing data with a model and we have **very few of these data**;

or

If we have **poorly defined distributions** or outliers,

then we do not have an adequate model for our data. Moreover we'll have to call on **non-parametric** methods.

2. Classical methods are **widely used**. We therefore need to understand results quoted to us in these terms.

The classical tests involve us in **'rejecting the null hypothesis',** i.e. **rejecting** rather than accepting a hypothesis, at some level of significance.

This null hypothesis may not be one in which we have the slightest interest.

A process of elimination.

A classical test works with probability distributions of a statistic while the Bayesian method deals with probability distributions of a hypothesis –one in which we may be very interested.

Classical testing – the Method

Set up two possible and exclusive hypotheses, each with an associated **terminal action**:

- **H_0, the null hypothesis or hypothesis of no effect, usually formulated to be rejected**

- **H_1, an alternative, or research hypothesis.**

Specify **a priori** the **significance level α.**

Choose a test which (a) approximates the conditions and (b) finds what is needed; obtain the **sampling distribution** and the **region of rejection**, whose area is **a fraction α** of the total area in the sampling distribution.

Run the test; **reject H_0 if the test yields a value of the statistic whose probability of occurrence under H_0 is < α.**

Carry out the **terminal action**.

Classical testing – the Devil is in the Detail

There is no such thing as an inconclusive hypothesis test.

Type I error : H_0 is in fact true; the probability is the probability of rejecting H_0, i.e. α.

Type II error : H_0 is false; the probability is the probability **β** of the failure to reject a false H_0. **β is not related to α in any direct or obvious way.**

The power of a test is the probability of rejecting a false H_1, or (1- β).

The **sampling distribution** is the p.d. or pdf of the test statistic. **The probability of any value of the test statistic occurring in the region of rejection is less than α.**

But **where the region of rejection lies within the sampling distribution depends on H_1**. If **H_1 indicates direction**, then there is a **single** region of rejection and the test is **one-tailed**; if no direction is indicated, the region of rejection is comprised of the two ends of the distribution and we are dealing with a **two-tailed** test.

This is the only use we make of H_1; the testing procedure can only convince us to accept H_1 if it is the sole alternative to H_0. **The procedure of elimination serves to reject H_0, not prove H_1. Beware -- it is human nature to think that your H_1 is the only possible alternative to H_0.**

Correlation Testing - Bayesian

Uses Bayes' Theorem to extract the probability distribution for **ρ** from the likelihood of the data and suitable priors.

We want to know about **ρ** independently of any inference about the means and variances, we have to integrate these ‘nuisance variables’ out of the full posterior probability **prob(**$\rho,\sigma_x,\sigma_y,\mu_x,\mu_y$ **| data).**

For the bivariate Gaussian model, the result is given by Jeffreys (1961) as

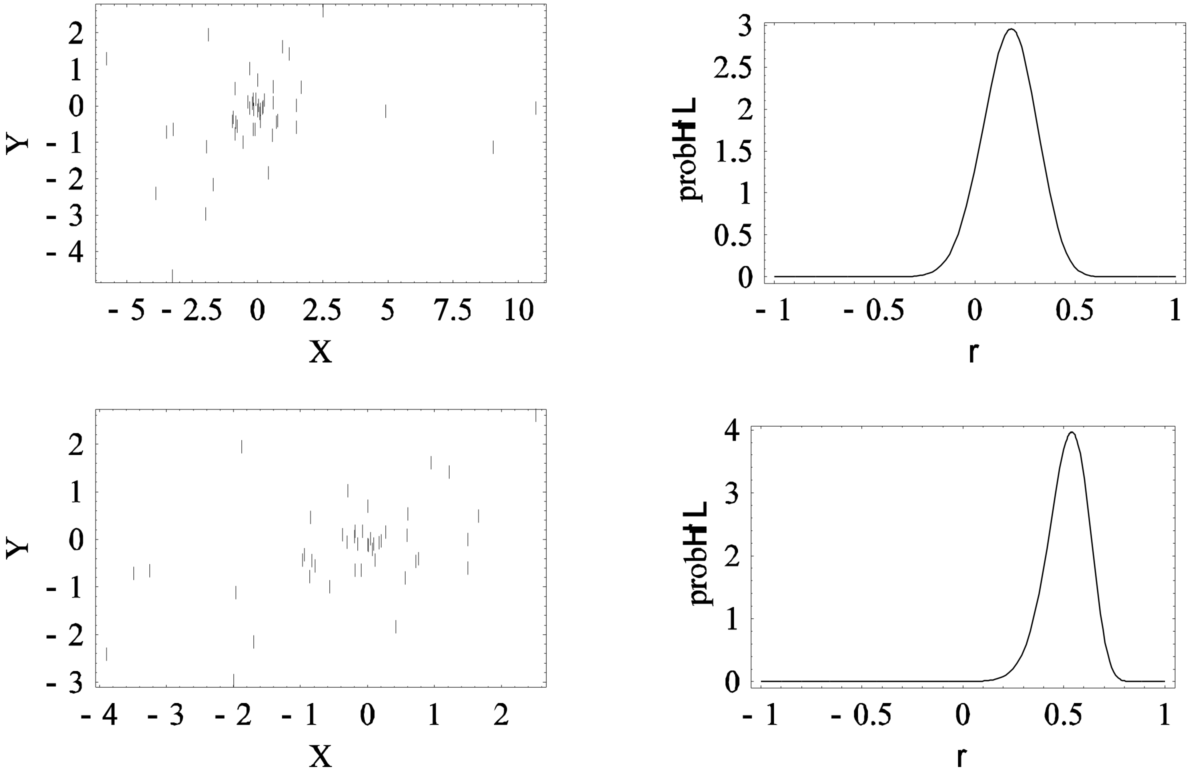
$$\mathrm{prob}((\rho \mid \ \mathrm{data}) \propto \frac{(1-\rho^2)^{\frac{n-1}{2}}}{(1-\rho r)^{n-\frac{3}{2}}}(1+\frac{1}{n-1/2}\frac{1+r\rho}{8}+\ldots)$$

The Bayesian test for correlation is thus simple: compute **r** from the **(X_i,Y_i),** and calculate prob(**ρ**) for the range of interest.

That’s it – and we have, as always with Bayes, what we really want to know.

Bayesian correlation testing – example 1

Generate **50 samples from a bivariate Gaussian** using **true correlation coefficient of 0.5 and add some outliers**, not accounted for by assuming a Gaussian.
Top panels: rotten result! Then remove the outliers **> 4σ. Better!**



50 X_i,Y_i chosen at random from a bivariate Gaussian with ρ = 0.5, outliers added. The Jeffreys probability distribution of correlation coefficient is shown, peaking at around 0.2 for the upper panel. The data have been restricted to ±4σ in the lower panel; the distribution now peaks at 0.54.

Correlation testing - classical approach

ρ now taken to be a fixed quantity -
=> probability of the **data**, given **ρ** (+ hypothesis of bivariate Gaussian).
The result (Fisher 1944) is:

$$\mathrm{prob}(r \mid \rho, H) \propto \frac{(1-\rho^2)^{(n-1)/2}(1-r^2)^{(n-4)/2}}{(1-\rho r)^{n-3/2}}\left(1+\frac{1}{n-1/2}\frac{1+r\rho}{8}+\ldots\right)$$

The standard parametric test is to attempt to reject the null hypothesis that **ρ** = 0:
1. Compute **r**. Note **-1 < r < 1**; **r=0** for no correlation, and the standard deviation in r is

$$\sigma_r = \frac{(1-r^2)}{\sqrt{N-1}}$$

2. Compute the probability, under this hypothesis, of **r** being this big or bigger.
If this probability is 'very small' we may conclude that the null hypothesis is unlikely.
3. To test the significance of a non-zero value for **r**, compute

$$\frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

which obeys the probability distribution of the **'Students' t statistic** with
N-2 degrees of freedom. (The transformation simply allows us to use tables of t.)
4. Consult the **table of critical values** for **t**; if **t** exceeds that corresponding to a critical value of the probability (two-tailed test), then the hypothesis that the variables are unrelated can be rejected at the specified level of significance. This level of significance (say 1%, or 5%) is the **maximum probability** which we are willing to risk in deciding to reject the null hypothesis (no correlation) **when it is in fact true**.

Correlation testing - classical approach 2

☻ This approach probably has not answered the question!

☻ We embark on this sort of investigation when it is apparent that the data contain correlations; we merely want some justification by knowing `how much'.

☻ The inclusion in the test of values of **unobserved values of r** is problematic**.**

☻ The test is widely used, and is formally powerful. **But**
- the data must be on continuous scales
- the relation between them must be linear. (How would we know this?)
- the data must be drawn from Normally-distributed populations. (How …..?)
- they must be free from restrictions in variability or groupings.

☻ There are parametric tests that help: the F-test for non-linearity and the Correlation Ratio test which gets around non-linearity.

☺ However, to circumvent the problems it is far better to go to **non-parametric tests**. These permit additional tests on data which are not numerically defined (binned data, or ranked data), so that in some instances **they may be the only alternative.**

Correlation Testing – Classical, Non-Parametric

The best known non-parametric test for correlation:

1. For the N data pairs of **(X_i,Y_i)**, make rank tables of **X_i** and **Y_i** such that **(XR_i,YR_i)** pairs represent the ranks for the **i**th pair, **$1 < XR_i < N$, $1 < YR_i < N$**.
2. Compute the **Spearman Rank Correlation Coefficient**:

$$r_s = 1 - 6\frac{\sum^{N}(XR_i - YR_i)^2}{N^3 - N}$$

3. The range is **$0 < r_s < 1$**; a high value indicates significant correlation. To find how significant, refer the computed **r_s** to the **table of critical values of r_s** applicable for **$4 \leq N \leq 30$**. If **r_s** exceeds an appropriate critical value, the hypothesis that the variables are unrelated is **rejected** at that level of significance.
4. If **N** exceeds 30, compute

$$t_r = r_s\sqrt{\frac{N-2}{1-r_s^2}},$$

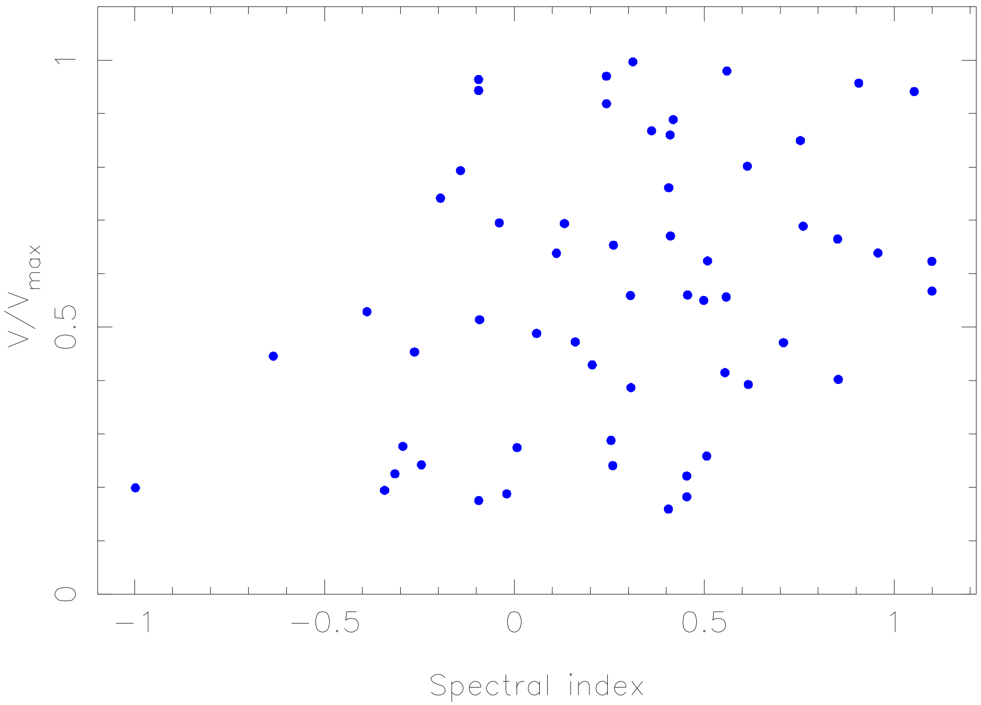
a statistic whose distribution for large **N** asymptotically approaches that of the **t** statistic with **N-2** degrees of freedom. The significance of **t_r** may be found from the **t-distribution table**, and this represents the associated probability under the hypothesis that the variables are unrelated.

In comparison with **r**, **r_s** has an efficiency of 91%.

Moral – if in doubt, go non-parametric.

Correlation – Classical, Non-Parametric: Example

A `correlation' at the notorious **2σ** level is shown. Here, **r_s = 0.28**, **N = 55**, and the hypothesis that the variables are unrelated is rejected at the **5% level of significance**. Here we have no idea of the underlying distributions; nor are we clear about the nature of the axes. The assumption of a bivariate Gaussian distribution would be crazy, especially in view of a uniformly-filled Universe producing a **V/V_{max}** statistic **uniformly** distributed between **0 and 1** (Schmidt 1968).



V/V_{max} as a function of high-frequency spectral index for a sample of radio quasars selected from the Parkes 2.7-GHz survey.

Correlation testing - comments

- The **non-parametric tests** circumvent some of the issues involved in the non-Bayesian approach, but they have no bearing on the fundamental issue – **what was the real question**?

2. But as ever, the Bayesian approach, strong in answering the real question, forces reliance on a **model**.

3. In practice there is little difference between the **Fisher test** and results from **Jeffreys distribution**. We can show this with some random Gaussian data with a correlation of zero. In the standard way, we can use the **r-distribution** to find the probability of **r** being as large, or larger, than we observe, on the hypothesis that **ρ=0**. If this probability is small, the test is hinting at the possibility that the correlation is actually positive. Therefore we compare with the probability, from the Jeffreys distribution, that **ρ** is positive. If the probability from Fisher's **r-distribution** is small we expect the probability from **ρ** to be large; and in fact we can see, either from simulations or from the algebraic form of the distributions, that **the sum of these two probabilities is always → 1**.

Interpreting the standard Fisher test (illegally!) to be telling us the chance that **ρ** is positive, actually works very well!

Tests for Means and Variances -1

Normally-distributed parent populations: the **"Student's" t test** (comparison of means) and the **F test** (comparison of variances).

Let's have n data $\mathbf{X_i}$ drawn from a Gaussian of mean $\boldsymbol{\mu_x}$, and **m** other data $\mathbf{Y_i}$, drawn from a Gaussian of identical variance $\boldsymbol{\sigma^2}$ but a different mean $\boldsymbol{\mu_y}$.

The Bayesian method: calculate the **joint posterior distribution** assuming a prior, integrating over the 'nuisance' parameter $\boldsymbol{\sigma}$, to get the joint prob($\boldsymbol{\mu_x}$, $\boldsymbol{\mu_y}$). From this we can calculate the probability distribution of ($\boldsymbol{\mu_x}$ - $\boldsymbol{\mu_y}$). The result depends on the data via a quantity

$$t' = \frac{(\mu_x - \mu_y) - (\overline{X} - \overline{Y})}{s\sqrt{m^{-1} + n^{-1}}}, \quad \text{where} \quad s^2 = \frac{nS_x + mS_y}{\nu}$$

$$\text{and} \quad S_x = \sum (X_i - \overline{X})^2/n, \;\; S_y = \ldots, \nu = n + m - 2.$$

The distribution for **t'** is

$$\text{prob}(t') = \frac{\Gamma[\frac{\nu+1}{2}]}{\sqrt{\pi\nu}\,\Gamma[\frac{\nu}{2}]} \left(1 + \frac{t'^2}{\nu}\right)^{-(\nu+1)/2} .$$

By this route we do not really hypothesis-test. We regard the **data as fixed** and ($\boldsymbol{\mu_x}$ - $\boldsymbol{\mu_y}$) as the variable, simply computing the probability of any difference in the means. We might work out the **range of differences** which are, say, 90% probable, or carry the distribution of mean difference on into a later probabilistic calculation.

Tests for Means and Variances -2

Classical approach: We do not treat the **μ**'s as random variables. Instead we guess that the difference in the averages **(<X> - <Y>)** will be the statistic we need; and we calculate its distribution on the **null hypothesis** that $\mu_x = \mu_y$.
We find that

$$t = \frac{\overline{X} - \overline{Y}}{s\sqrt{m^{-1} + n^{-1}}}$$

follows a **t-distribution** with **v** degrees of freedom.

This is the basis of a classical hypothesis test, the **Student's t test for means**. Assuming that $\mu_x - \mu_y = 0$, (the null hypothesis) we calculate **t**. If it (or some greater value) is very unlikely (see a **t-table**), we think that the null hypothesis is ruled out.

The **t-statistic** is heavy with history and reflects an era when analytical calculations were essential. The penalty is **total reliance on the Gaussian**. However, with cheap computing power -

we may expect to be able to follow the basic Bayesian approach.

Tests for Means and Variances -3

By analogous calculations, we can arrive at the **F test** for variances. Again, **Gaussian distributions** are assumed.

The null hypothesis is $\boldsymbol{\sigma_x = \sigma_y}$, the data are $\mathbf{X_i}$ (I = 1 … N) and $\mathbf{Y_i}$ (I = 1 … M) and the test statistic is

$$\mathcal{F} = \frac{\sum_i (X_i - \overline{X})/(N-1)}{\sum_i (Y_i - \overline{Y})/(M-1)}.$$

This follows a **F-distribution with N-1 and M-1 degrees of freedom** (F table).

The testing procedure is the same as for Student's t.

This statistic will be particularly sensitive to the Gaussian assumption.

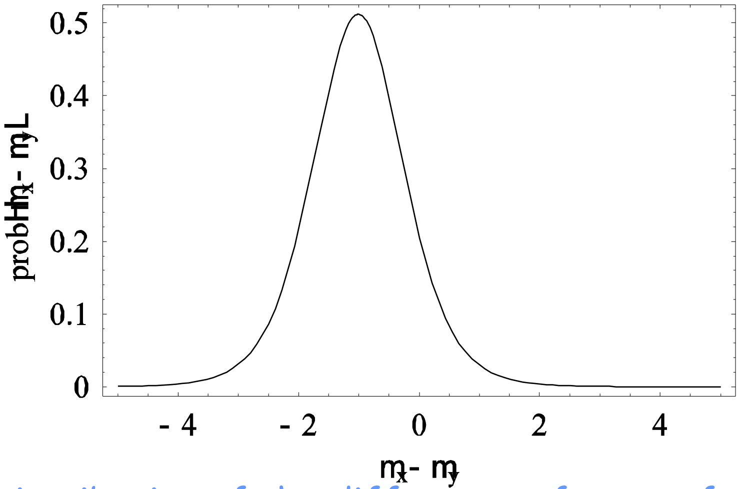
Tests for Means and Variances - Example 1

Take two small sets of data, from Gaussian distributions of equal variance:
-1.22, -1.17, 0.93, -0.58, -1.14 (mean -0.64), and
1.03, -1.59, -0.41, 0.71, 2.10 (mean 0.37), pooled **std dev of 1.2**.
The standard **t-statistic = 1.12**.

If we do a **two-tailed test** (not caring whether one mean is larger than another), we find a **30% chance** that these data would arise if the means were the same.

The **one-tailed test** (testing whether one mean is larger) gives **16%.**

Bayesian point of view? We can calculate the **distribution of (μ_x - μ_y)** for the same data. In the Fig we can see clearly that one mean is smaller; the odds on this being so are about 10 to 1, as can be calculated by integrating the posterior distribution of the difference of means.



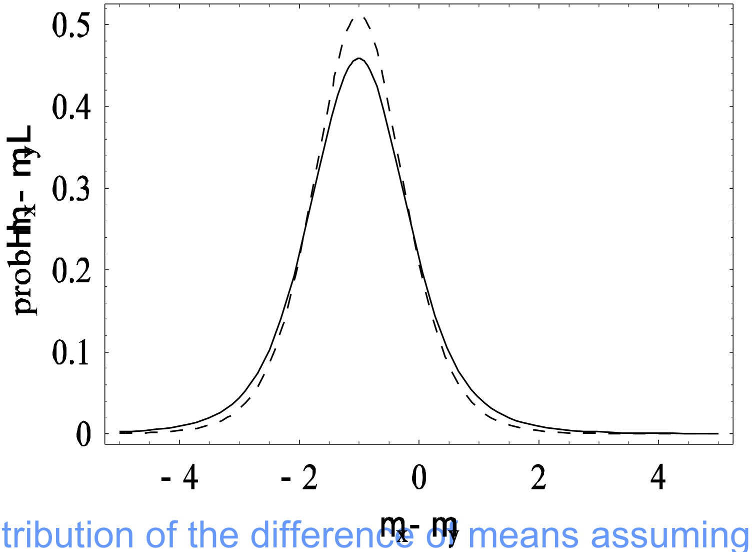
Distribution of the difference of means for the example data.

Tests for Means and Variances – Example 2

Consider the same example data as before, relaxing the assumption that the variances are equal. (The sample standard deviations are 0.9 and 1.4, not significantly different, according to the F-test.)

We see from the Fig that the distributions of **μ_y - μ_x** are similar to the **t-distribution**, although as we might expect the distribution is **a little wider** if we do not assume that the variances are equal.

Thus although we cannot tell (classically) that the variances differ, **we will obtain somewhat different results** by not assuming that they are the same.



Distribution of the difference of means assuming equal variances (dashed) and without this assumption (solid)

This general sort of Bayesian test can be followed for any distribution – as long as we know what it is, and can do the integrations.

Non-parametric tests (classical!)

Why? (1) fewer assumptions about the data - if the underlying distribution is unknown, there is no alternative. (2) they work for small sample sizes, (3) they cope with non-numerical data; and (4) they can treat samples from several populations.

'No distribution is assumed'? Don't be silly. What is assumed?

Counting probabilities!

Example: the **chi-square test**. The number of items in bin **i** is $\mathbf{N_i}$, and we expect $\mathbf{E_i}$. For smallish numbers, **Poisson statistics** tells us that the variance is also $\mathbf{E_i}$. So $\mathbf{(N_i - E_i)^2/E_i}$ should be roughly a **squared Gaussian variable**, of unit variance.

Example: the **runs test -** is just using the assumption that each successive observation is equally likely to be 'up' or 'down', so a **binomial distribution** applies.

The assumptions underlying non-parametric tests are **weaker, and so more general,** than the for parametric tests.

The main counter-argument concerns binning - **binning is bad**; it loses information and therefore **loses efficiency**. The **power** of non-parametric tests may be somewhat less, but typically no more than 10% less than their parametric equivalents.

Chi-square test (Pearson 1900)

If we have **observational data which can be binned**, and a model/hypothesis which predicts **the population of each bin**,

Then the chi-square statistic describes the **goodness-of-fit** of the data to the model.

With the **observed** numbers in each of **k** bins as **O**$_i$, and the **expected** values from the model as **E**$_i$, then this statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

The null hypothesis **H**$_0$ is that the number of objects falling in each category is **E**$_i$; the **chi-square procedure tests** whether the **O**$_i$ are sufficiently close to **E**$_i$ to be likely to have occurred under **H**$_0$. The sampling distribution under **H**$_0$ of the statistic **χ** follows the **chi-square distribution**

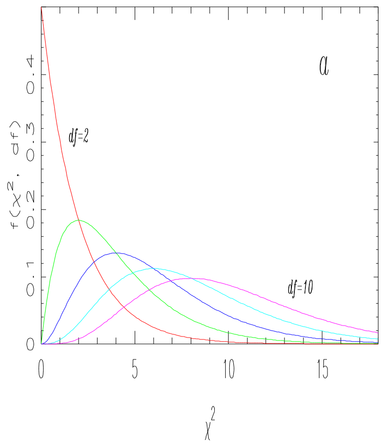
$$f(x) = \frac{2^{-\nu/2}}{\Gamma[\nu/2]} x^{\nu/2-1} e^{-x/2}$$

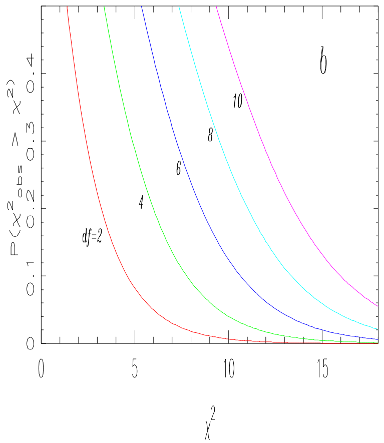
(for x > 0) with **v = (k-1)** degrees of freedom. (One degree of freedom is lost because of the constraint that **Σ**$_i$ **O**$_i$ **= Σ**$_i$ **E**$_i$.) This is the distribution function of the random variable $Y^2 = Z_1^2 + Z_2^2 + \ldots + Z_\nu^2$ where the **Z**$_i$ are independent random variables of standard Normal distribution.

A **chi-square table** presents critical values; if **χ²** exceeds these values, **H**$_0$ is **rejected** at that level of significance.

Chi-square test – 2: The Good News

- **Common** – known, accepted.
- **Additive** – pull in different data sets, bin sizes, etc
- **The contribution** to χ^2 from **each bin** can be examined to look for regions of good/bad fit.
- **Easily** computed.
- **Mean** = no. of deg of freedom; **variance** = 2 x no. of deg of freedom
- **=> Rule of thumb**: if **χ^2 ~ no. of bins**, accept H_0; if **> 2 x (no. of bins)**, reject.
- **Free model-fitting!**





Chi-square test – 3: The Bad News

- The **data must be binned** to apply the test, and the bin populations must reach a certain size because it is obvious that instability results as **$E_i \rightarrow 0$.**

=> Another rule-of-thumb : **> 80% of the bins must have E_i > 5.**
Bins may have to be combined.

- However, the **binning of data** in general, and certainly the binning of bins, results in **loss of efficiency and information,** resolution in particular.

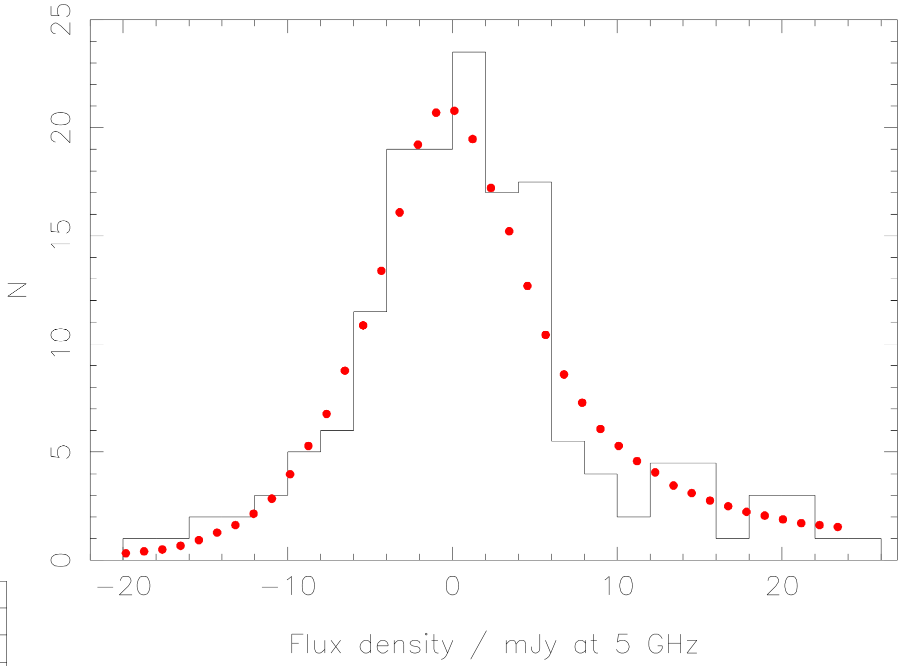
3. Small samples **cannot** be treated.

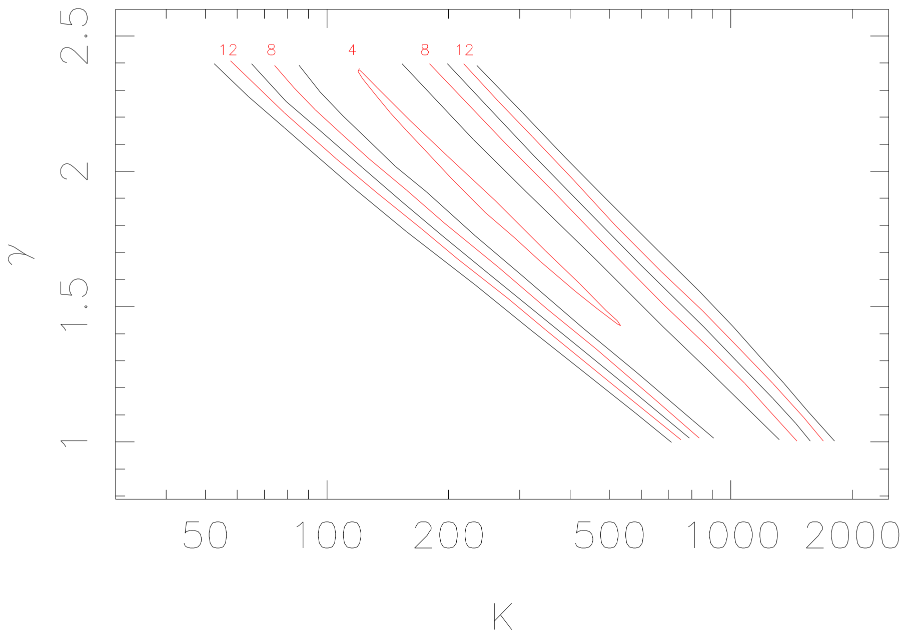
- The **chi-square test** cannot tell **direction**; it is a **'two-tailed'** test; it can only tell whether the differences between sample and prediction exceed those reasonably expected on the basis of statistical fluctuations due to the finite sample size.

There must be something better…..

Chi-square test – 4: Example

Chi-square testing/modelling: the object of the experiment was to estimate the surface-density count (the **N(S)** relation) of faint radio sources at 5 GHz, assuming a power-law **$N(>S) = KS^{-(\gamma-1)}$**, **γ** and **K** to be determined from the distribution of background deflections, the **P(D) method**. The histogram of measured deflections is shown right.





The dotted red curve above represents the optimum model from minimizing χ^2. Contours of χ^2 in the **γ - K plane** are shown left.

With the best-fit model, **χ^2** = 4 for 7 bins, 2 parameters; thus dof = 4. **Right on.**

Chi-square test 5: Two (or k) independent samples

H_0 is that the **k** samples are from the same population.

1. Each sample is binned in the same r bins (a **k x r contingency table**).
2. Compute

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}, \quad \text{with } E_{ij} \text{ the expectation values, from } \quad E_{ij} = \frac{\sum_{j=1}^{k} O_{ij} \cdot \sum_{i=1}^{r} O_{ij}}{\sum_{i=1}^{r}\sum_{j=1}^{k} O_{ij}}.$$

Under $\mathbf{H_0}$ this is distributed as χ^2, with **(r-1)(k-1)** degrees of freedom.

There is a modification of this test for the case of the **N**-object **2 x 2** contingency table:

Sample =	1	2
Category = 1	A	C
= 2	B	D

$$\chi^2 = \frac{N(|AD - BC| - N/2)^2}{(A+B)(C+D)(A+C)(B+D)}$$

which has **dof = 1**.

The usual chi-square caveat applies – cell numbers should stay above 5. If they don't, combine adjacent cells, or abandon ship. And if there are only **2 x 2** cells, the total **N** must exceed 30; if not, use the **Fisher Exact Probability test**. **For data which are not on a numerical scale, this test is probably it.**

A positive: The **k-sample chi-square test** may be used to test a directional alternative to H0; H1 can be that the two groups differ in some predicted sense.

OK, OK, so what's the Fisher Exact Probability Test?

For two independent small samples with discrete binary data, i.e. mutually exclusive bins

Sample =	1	2
Category = 1	A	C
= 2	B	D

H_0: the assignment of ‘scores’ is random

Compute

$$p = \frac{(A+B)!(C+D)!(A+C)!(B+D)!}{N!A!B!C!D!}$$

This is the probability that the total of **N** scores could be as they are **when the two samples are in fact identical.** But the test asks : what is the probability of occurrence of the observed outcome **or one more extreme under H_0**?

Thus we must compute and add the probabilities of the more extreme cases until **both** samples have a zero in one of their boxes. Then

$$p_{tot} = p_1 + p_2 + p_3 + \ldots.$$

Computation can be ‘tedious’; but it’s the best test to use for small samples, and **if N < 20 it is on its own.**

Kolmogorov–Smirnov (K-S) testing

1. Calculate **S_e(x),** the **predicted** cumulative (integral) frequency distribution under **H_0**

2. Compute **S_o(x),** the **observed** cumulative distribution, the sum of all observations to each **x** divided by the sum of all **N** observations.

3. Find $$D = \max \mid S_e(x) - S_o(x) \mid$$

4. Consult the known sampling distribution for **D** under **H_0**, as given in a **K-S table**, to determine the fate of **H_0.** If **D** exceeds a critical value at the appropriate **N**, then **H_0** is rejected at that level of significance.

Thus as for the chi-square test, the sampling distribution indicates whether a divergence of the observed magnitude is **'reasonable' if the difference between observations and prediction is due solely to statistical fluctuations.**

Advantages: (1) no binning
(2) small samples
(3) greater power for intermediate samples
(4) with modification, can be directional

Disadvantages: (1) continuous functions needed, numerical scale
(2) no model fitting side benefit, no minimization of K-S possible.

K-S testing, two samples

1. Calculate **S_m(x),** the cumulative (integral) frequency for sample 1 (**m** members) and **S_n(y),** the cumulative distribution for sample 2 (**n** members).

2. Find $$D = \max \mid S_m(x) - S_n(y) \mid$$

3. Consult the known sampling distribution for **D** under **H_0**, as given in a **K-S two-sample table**, to determine the fate of **H_0.** Now there are tables for both one- and two-tailed tests. If **D** exceeds a **critical value** at the appropriate **N**, then **H_0** is rejected at that level of significance.

If you run off the end of the tables with big samples, **approximations** work:

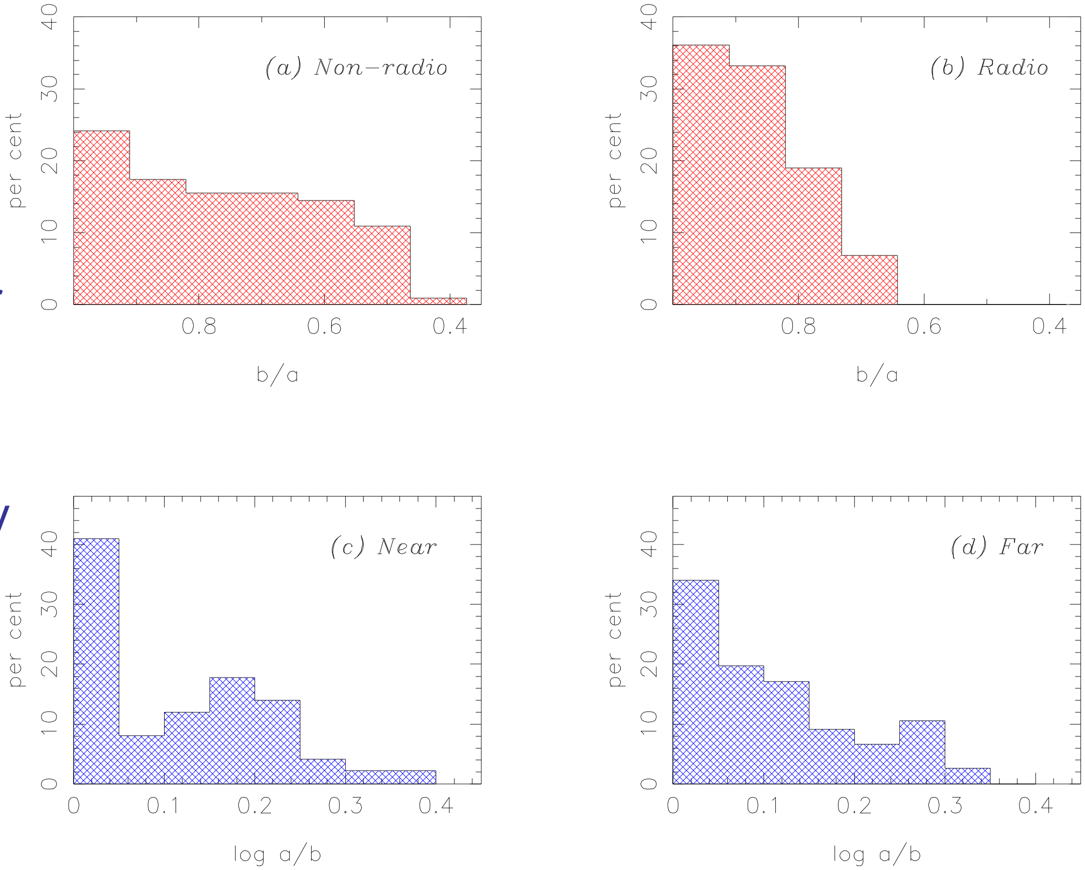
- For the **two-tailed test**, a simple table for the usual levels of significance is given.
- For large samples, **one-tailed test**, compute $$\chi^2 = 4D^2 \frac{mn}{m+n},$$

which has a ~ **chi-square sampling distribution with 2 dof**. Then use a **chi-square table** to determine the fate of **H_0.**

A powerful test: efficiency **always exceeds chi-square**, and just exceeds that of the **Mann-Whitney U test** for very small samples. For **larger samples, U-test is preferred.**

K-S testing, two samples - Example

Kolmogorov-Smirnov tests on subsamples of ellipticals from the Disney-Wall (1977) sample of bright ellipticals.



Upper panels -
distribution functions in **b/a**, minor to major axis, for (a) the 102 undetected and (b) the 30 radio-detected ellipticals in the sample. The Kolmogorov-Smirnov two-sample test rejects $\mathbf{H_0}$, that the subsamples are drawn from the same population, at a significance level of **< 1%.**

Lower panels –
distribution functions in **log a/b** for (c) the 51 ellipticals closer than 30 Mpc, (d) 76 bright ellipticals in the sample more distant than this. The Kolmogorov-Smirnov test indicates no significant difference between these latter subsamples.

Runs test of randomness

So simple - **form a binary (1 - 0) statistic from each sample datum**, e.g. heads-tails, or the sign of the residuals about a mean or a best-fit line. It is to test $\mathbf{H_0}$ that this new statistic is random; successive observations are independent. **Are there too few runs?**

Determine **m**, the number of heads or 1's; **n**, the number of tails or 0's, **N=n+m**; and find **r, the number of runs**.

Look up the level of significance from the tabled probabilities for one or two-tailed test – depending on $\mathbf{H_1}$, which can specify (as the **research hypothesis**) how the non-randomness might occur. (In general we are concerned simply with the **one-tail test**, asking whether or not the number of runs is **too few**, the issue being independence of data in a sequence.)

For **m** 'heads' and **n** 'tails' with **N** data, the expectation value of number of runs is

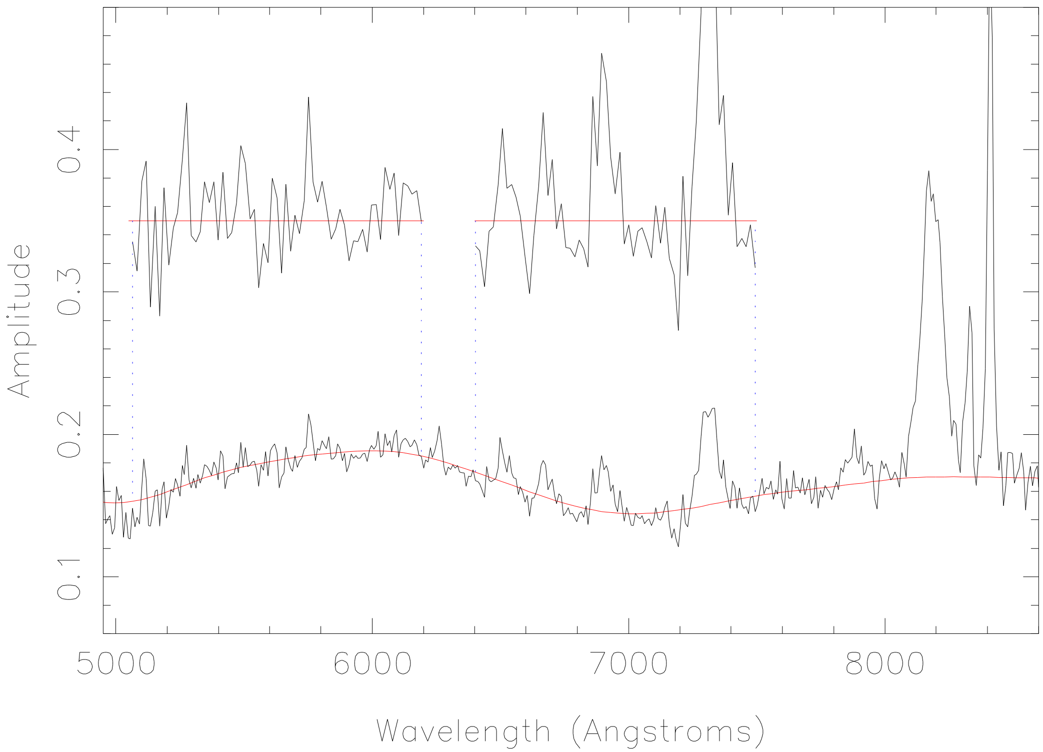
$$\mu_r = \frac{2mn}{m+n+1}, \text{ with } \sigma_r = \sqrt{\frac{2nm(2nm-N)}{N^2(N-1)}}.$$

…becoming **asymptotically Gaussian**, so that the Gaussian distribution or its integral erf can be used by forming

$$z = \frac{r-\mu_r}{\sigma_r}$$

and **consulting tables for the Normal distribution**. This is the procedure **when the numbers exceed 20** and toddle off the end of runs table.

Runs test of randomness - Example



A spectrum of the quasar **3C207**, taken with the 4.2-m William Herschel Telescope. Red curve: baseline fitted by **Fourier minimum-component technique**. The regions considered for runs test are shown in the separated sections, baseline-subtracted and magnified by 3. These carefully-selected regions of the spectrum are examined with the runs test. **Left region - concordance**, 36 positive deflections, 29 negative, **31 runs vs an expectation of 32.1** runs, **z = -0.28**. **Right region** – in the **Hydrogen Balmer-line series**, and several members are present in emission; **rejection** of randomness at about 4σ: 31 positives, 32 negatives, **16 runs against an expectation of 31.5**, **z = -3.94**. Broad emission lines yield contiguous regions decreasing the number of runs.

Wilcoxon-Mann-Whitney U test for two samples

There are two samples, **A** (**m** members) and **B** (**n** members); $\mathbf{H_0}$ is that **A** and **B** are from the same distribution or have the same parent population, while $\mathbf{H_1}$ may be one of three possibilities: **A** stochastically larger than **B**; **B** stochastically larger than **A**; or **A and B differ in some other way**, perhaps in **scatter or skewness**.

The first two hypotheses are directional, resulting in **one-tailed tests**; the third is not, resulting in a **two-tailed test**. To proceed,

1. Decide on $\mathbf{H_1}$ and the significance level **α,**
2. **Rank in ascending order** the combined sample **A+B**, preserving the **A** or **B** identity of each member.
3. (Depending on choice of $\mathbf{H_1}$) **Sum** the number of **A-**rankings to get $\mathbf{U_A}$, or *vv*, the **B-**rankings to get $\mathbf{U_B}$. Tied observations are assigned the average of the tied ranks. Note that if **N=m+n**,

$$U_A + U_B = \frac{N(N+1)}{2}$$

so that only one summation is necessary to determine both.

Look up the result in the table calculated from the sampling distribution (**pdf of U**). The table presents probabilities for **U > observed**, and for **U < observed**. For samples >10, the sampling distribution for **U tends to Normal** with mean $\boldsymbol{\mu_A = m(N+1)/2}$ and variance $\boldsymbol{\sigma_A^2 = mn(N+1)/12}$. Significance can be assessed from the Normal distribution, by calculating

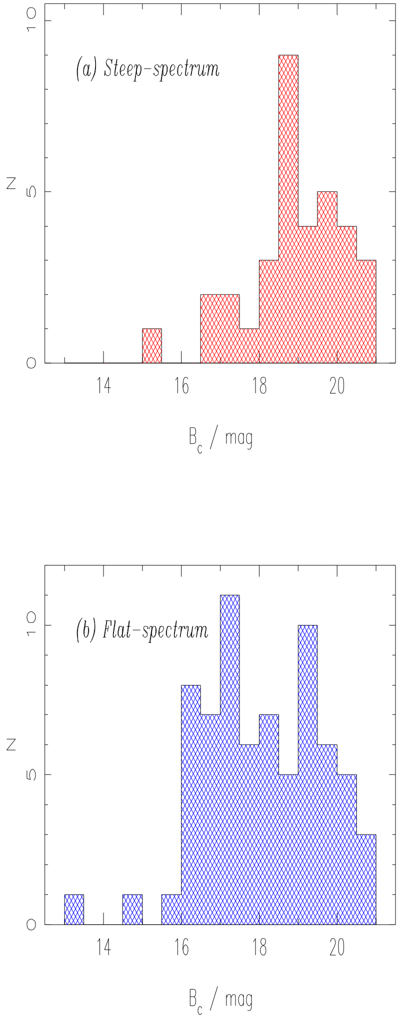
$$z = \frac{U_A \pm 0.5 - \mu_A}{\sigma_A}$$

where +0.5 corresponds to considering probabilities of **U ≤** that observed (lower-tail), and -0.5 for **U ≥** that observed (upper-tail). If the two-tailed test is required, simply double the probabilities.

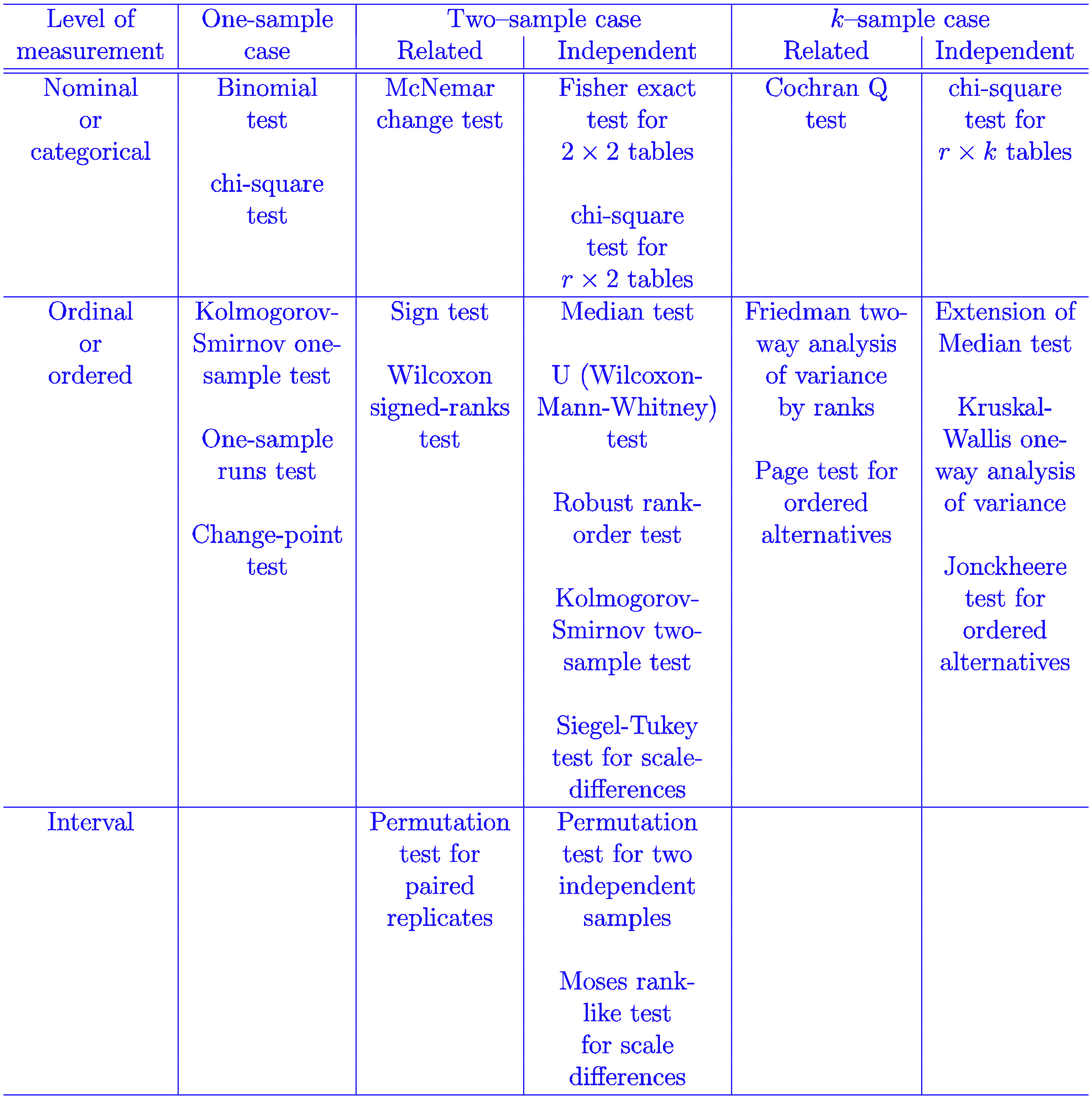
U test - Example

Magnitude distributions for flat and steep (radio) spectrum quasars from a complete sample of quasars in the Parkes 2.7-GHz Survey. **H_1** is that the flat-spectrum quasars extend to significantly lower (brighter) magnitudes than do the steep-spectrum quasars, a claim made earlier by several observers. The eye agrees with **H_1**, and so does the result from the **U test**, in which we found **U = 719, z = 2.69**, rejecting **H_0** in favour of **H_1** at the

0.004 level of significance.



Non-parametric tests for comparison of samples



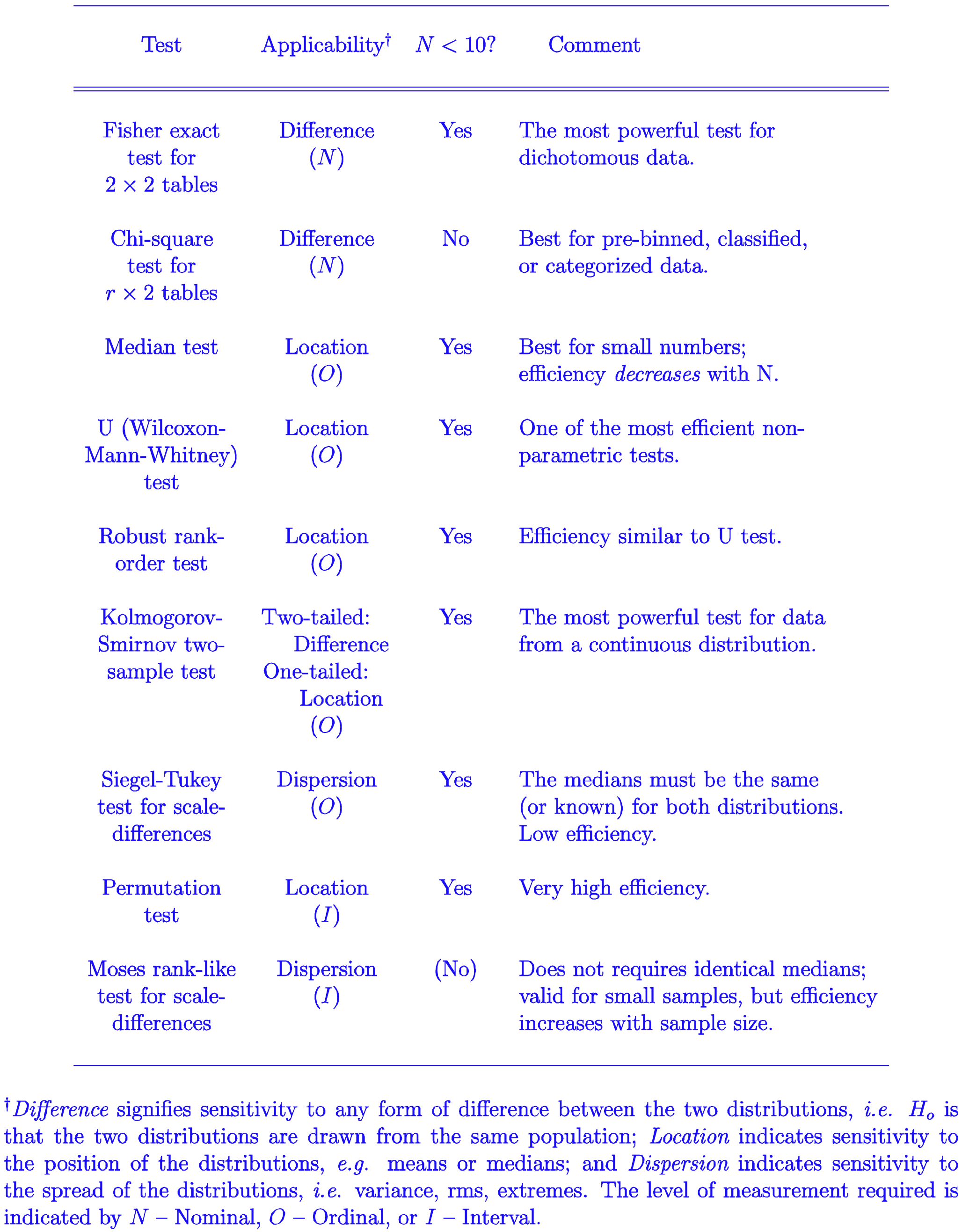
Level of measurement	One-sample case	Two–sample case		k–sample case	
		Related	Independent	Related	Independent
Nominal or categorical	Binomial test chi-square test	McNemar change test	Fisher exact test for 2×2 tables chi-square test for $r \times 2$ tables	Cochran Q test	chi-square test for $r \times k$ tables
Ordinal or ordered	Kolmogorov-Smirnov one-sample test One-sample runs test Change-point test	Sign test Wilcoxon signed-ranks test	Median test U (Wilcoxon-Mann-Whitney) test Robust rank-order test Kolmogorov-Smirnov two-sample test Siegel-Tukey test for scale-differences	Friedman two-way analysis of variance by ranks Page test for ordered alternatives	Extension of Median test Kruskal-Wallis one-way analysis of variance Jonckheere test for ordered alternatives
Interval		Permutation test for paired replicates	Permutation test for two independent samples Moses rank-like test for scale differences		

Single-sample non-parametric tests

Test	Applicability[†]	$N < 10$?	Comment
Binomial test	Goodness-of-fit (N)	Yes	Appropriate for two-category (dichotomous) data; do *not* dichotomize continuous data.
Chi-square test	Goodness-of-fit (N)	No	For testing categorized, pre-binned, or classified data; choose categories with expected frequencies 6 – 10.
Kolmogorov-Smirnov one-sample test	Goodness-of-fit (O)	Yes	The most powerful test for data from a continuous distribution; may always be more efficient than chi-square test.
One-sample runs test	Randomness of event sequences (O)	Yes	Does not estimate differences between groups.
Change-point test	Change in the distribution of an event sequence (O)	Yes	Robust with regard to changes in distributional form; efficient.

[†] *Goodness-of-fit* indicates general testing for any type of difference, *i.e.* H_o is that the distribution is drawn from the specified population. The level of measurement required is indicated by N – Nominal, O – Ordinal, or I – Interval.

Two-sample non-parametric tests



Test	Applicability†	$N < 10$?	Comment
Fisher exact test for 2×2 tables	Difference (N)	Yes	The most powerful test for dichotomous data.
Chi-square test for $r \times 2$ tables	Difference (N)	No	Best for pre-binned, classified, or categorized data.
Median test	Location (O)	Yes	Best for small numbers; efficiency *decreases* with N.
U (Wilcoxon-Mann-Whitney) test	Location (O)	Yes	One of the most efficient non-parametric tests.
Robust rank-order test	Location (O)	Yes	Efficiency similar to U test.
Kolmogorov-Smirnov two-sample test	Two-tailed: Difference One-tailed: Location (O)	Yes	The most powerful test for data from a continuous distribution.
Siegel-Tukey test for scale-differences	Dispersion (O)	Yes	The medians must be the same (or known) for both distributions. Low efficiency.
Permutation test	Location (I)	Yes	Very high efficiency.
Moses rank-like test for scale-differences	Dispersion (I)	(No)	Does not requires identical medians; valid for small samples, but efficiency increases with sample size.

† *Difference* signifies sensitivity to any form of difference between the two distributions, *i.e.* H_o is that the two distributions are drawn from the same population; *Location* indicates sensitivity to the position of the distributions, *e.g.* means or medians; and *Dispersion* indicates sensitivity to the spread of the distributions, *i.e.* variance, rms, extremes. The level of measurement required is indicated by N – Nominal, O – Ordinal, or I – Interval.

End Bertinoro 4 (JVW)