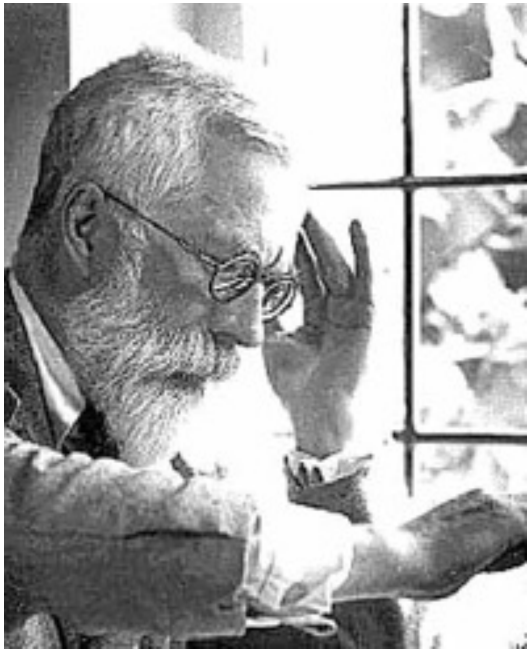
ASTR509 - 8

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

ASTR509 © Jasper Wall Fall term 2013

We're back ....

Remembering that we were dealing with correlation, and PCA

There are many pitfalls!

- Both Bayesian and classical tests for correlation generally depend upon the bi- or multi-variate Gaussian model, with correlation coefficient(s) ρ. In the classical case we are trying to reject the null hypothesis, that ρ is zero.

- The only escape from the bi-/multi-variate Gaussian is using the non-parametric version of the classical tests.

- Partial correlation is a painful exercise in subscripts – it is used by many, and there is a large literature. It is easy enough to carry out, following the prescriptions and using the conventions of the classical (Fisher) test for correlation significance.

- Principal Component Analysis is really the way to search a data-set objectively to determine what depends on what - it is the ultimate correlation searcher.

- We did a couple of examples in detail to show how it works. It is easy enough to implement, is hands-off in terms of deciding what to test against what, and it leaves all the difficult stuff to the physical interpretation at the end. Or so we thought until we tried the assignment?

We now return to look more carefully at the whole area of hypothesis testing.

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 Neyman-Pearson Method

1. 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.

2. Specify **a priori** the **significance level α.** Choose a test which (a) approximate 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.

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

4. 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.

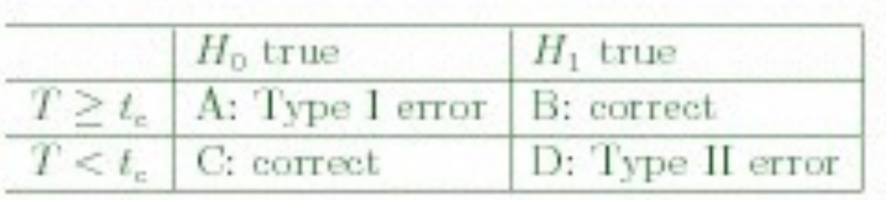
This is the only use we make of H_1; the testing procedure can only convince us to accept H_1 if it is the sole alternative to H_0. **The procedure of elimination serves to reject H_0, not prove H_1. Beware -- it is human nature to think that your H_1 is the only possible alternative to H_0.**

Classical Testing: the Devil is STILL in the Detail

If we don't have a well-defined H_1, then a threshold for action makes no sense (Fisher's objection).

Maybe calculate p - notice that it is uncomfortably small – think about other possibilities? This is OK as long as we don't ascribe a probability to our rejection of H_0.

Suppose we have a critical value of our test statistics, say t_c. Assume that the chance of exceeding t_c under H_0 is α. We then compute our test statistic T. compare with t_c, reject H_0 (accept H_1) if $T > t_c$. Here are the possibilities:



	H_0 true	H_1 true
$T \geq t_c$	A: Type I error	B: correct
$T < t_c$	C: correct	D: Type II error

Note probabilities of A and C add up to 1.0, and so do probabilities B and D

BUT THERE IS NO RELATIONSHIP ALONG THE ROWS, i.e. there is no relationship between Type I error rate and Type II error rate.

The probability of B being occupied is the POWER; tradoff is between POWER and TYPE I ERROR RATE.

Tests for Means and Variances - I

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 $\boldsymbol{\mu_x = \mu_y}$.
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 $\boldsymbol{\mu_x - \mu_y = 0}$, (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 **σ_x = σ_y**, the data are **X_i** (I = 1 … N) and **Y_i** (I = 1 … M) and the test statistic is

$$\mathcal{F} = \frac{\sum_i (X_i - \overline{X})^2/(N-1)}{\sum_i (Y_i - \overline{Y})^2/(M-1)}.$$

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.

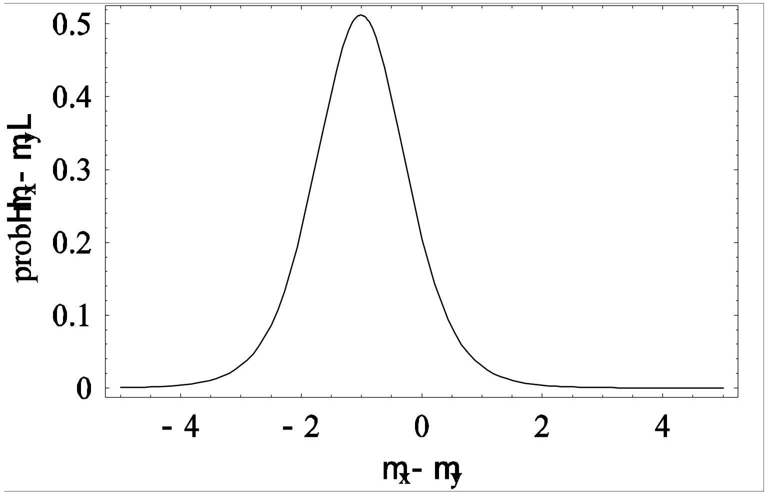
Tests for Means and Variances - Example 1

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 - 4

The Behrens - Fisher Test

Relaxing the assumption of equal variances may be important. The distribution of the difference in means without this assumption is called the **Behrens-Fisher distribution.** It is a rare example of a Bayesian analysis having no classical analogue; there is no classical test for the case of possibly unequal variances.

The analytical form of the Behrens-Fisher distribution is complicated and involves a numerical integration anyway, so we may as well resort to a computer right away to calculate it from Bayes' Theorem. Assume our two std Gaussians in x and y, now with their very own $\boldsymbol{\sigma}_{\mathbf{x}}$ and $\boldsymbol{\sigma}_{\mathbf{y}}$. The joint posterior distribution (using the Jeffreys prior on the σ) is

$$\mathrm{prob}(\mu_x, \mu_y, \sigma_x, \sigma_y) \propto \frac{1}{\sigma_x^{n+1}} \exp\left[-\frac{\sum_i (x_i - \mu_x)^2}{2\sigma_x^2}\right] \times \frac{1}{\sigma_y^{n+1}} \exp\left[-\frac{\sum_i (y_i - \mu_y)^2}{2\sigma_y^2}\right]$$

We have to do a multidimensional integration to get rid of the two **nuisance parameters** ($\boldsymbol{\sigma}_{\mathbf{x}}$ and $\boldsymbol{\sigma}_{\mathbf{y}}$) and to ensure that the resulting joint distribution **prob($\boldsymbol{\mu}_{\mathbf{x}}$, $\boldsymbol{\mu}_{\mathbf{y}}$)** is normalized.

Given the joint distribution of $\boldsymbol{\mu}_{\mathbf{x}}$ and $\boldsymbol{\mu}_{\mathbf{y}}$, we would like the distribution of **($\boldsymbol{\mu}_{\mathbf{y}}$ – $\boldsymbol{\mu}_{\mathbf{x}}$).**

By changing variables,

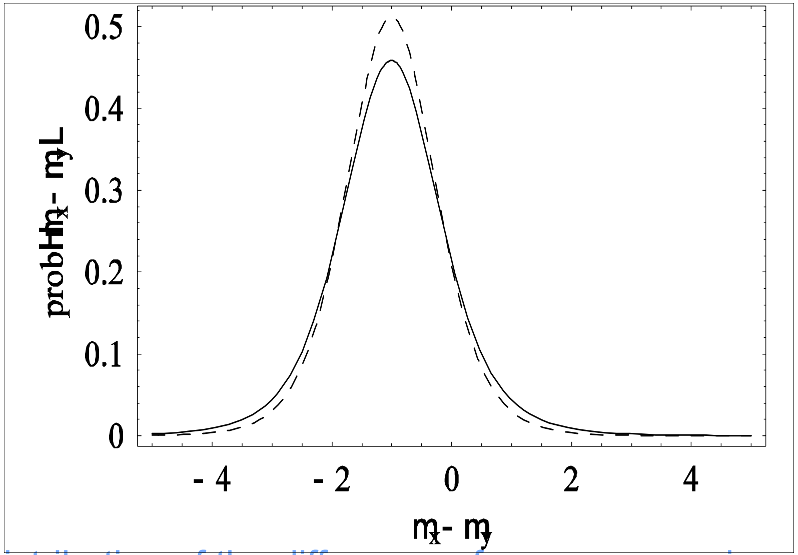
$$\mathrm{prob}(u = \mu_y - \mu_x) = \int_{\infty}^{\infty} \mathrm{prob}(v, v + u)\, dv.$$

(Another integration!).

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.



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

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

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-Gaussian Parametric Testing - 1

- Often we have **little information** about the distributions from which our data are drawn, yet we need to test whether they are the same or not.
- There is only **one way** in which two unknown distributions can be the same, but a multitude in which they may differ.

=> **classical hypothesis tests**, which assume the distributions are the same.

- Some information about distributions => **Bayesian methods**.
- The trick is to use a **multi-parameter generalization** of a familiar distribution, with extra parameters to allow distortions in the shape.
- Eventually can **marginalize out** these 'nuisance parameters', integrating over our prior assumptions about their magnitude.

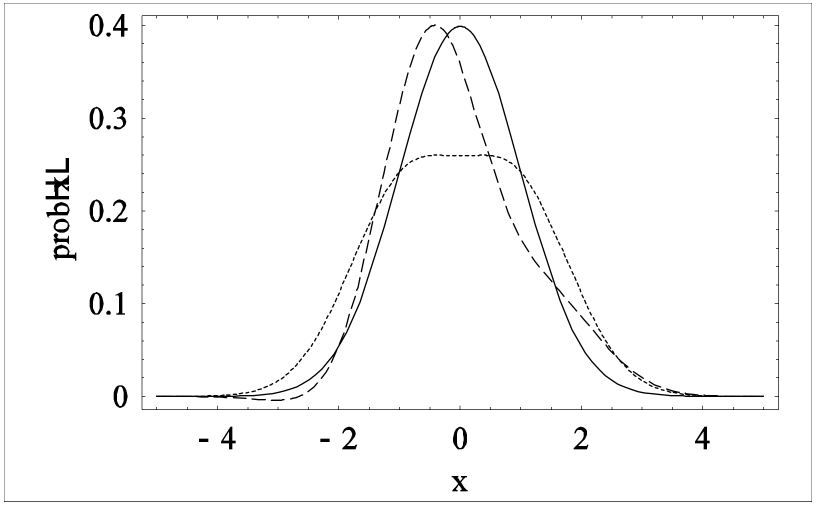
- Most common example: the **Gram-Charlier series** in which the **H**'s are the **Hermite polynomials**. The coefficients **a**$_i$ are the **free parameters** we need. (Because of properties of Hermite polynomials, these coefficients are also related to the moments of the distribution we are trying to create.)

$$\exp(-\frac{x^2}{2\sigma^2})\,(1+\sum_i a_i H_i(x))$$

-The effect of these extra terms is to **broaden and skew a Gaussian**, and so for some data a few-term Gram-Charlier series may give a **useful basis for a parametric analysis.** Priors on the coefficients are set by judgment.

Non-Gaussian Parametric Testing - 2

Example – various Gram-Charlier distributions resulting from using just two terms. The solid curve is a pure Gaussian.



There are two **variants on the Gram-Charlier** series.

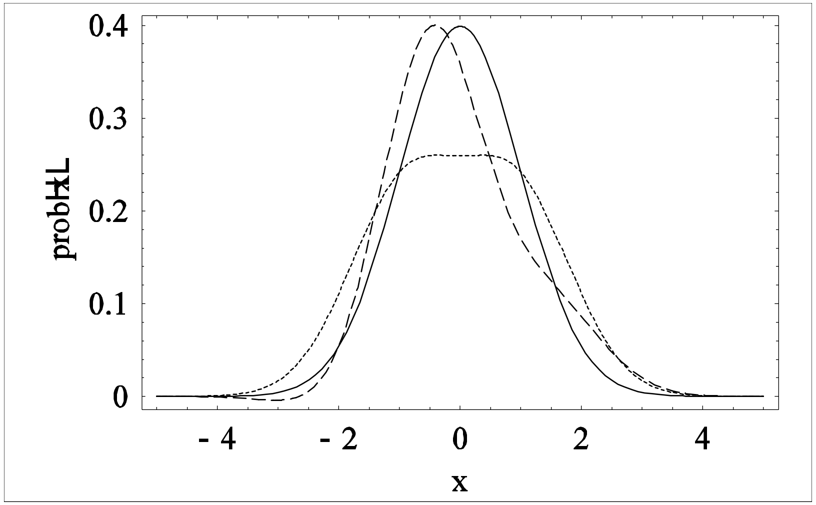
1. For a distribution allied to the exponential **exp(-x/a),** a **Laguerre series** will function in the same way, the distorting functions being the **Laguerre polynomials**.

2. The **Gamma series** is based on the distribution **x^{α} $(1-x)^{\beta}$**, defined on the interval from **0 to 1**; the distorting functions are the **Jacobi polynomials**.

See computer algebra packages (e.g. MATHEMATICA) for support of such special functions.

Non-Gaussian Parametric Testing - 3

This approach clarifies the workings of non-parametric tests.



Suppose we fix on a two-term Gram-Charlier expansion as a realistic representation of our data; the versatility is demonstrated in the Fig.

For **data set 1**, we then get the posterior **prob(μ_1,σ_1,a_1^1,a_2^1),** and similarly for data set 2.

If we ask the apparently innocuous question **"are these data drawn from different distributions?"** we see that there are many possibilities (in fact, 2^4) of the form, for instance, $\mu_1 > \mu_2$ and $\sigma_1 < \sigma_2$ and $a_1^1 > a_1^2$ and $a_2^1 < a_2^2$.

Working through these possibilities could be **'quite tedious'.**

A different question might be **"are these distributions at different locations, regardless of their widths?",** in which case we could marginalize out the **σ's** and **a_2's**; the location, in a Gram-Charlier expansion, is **a simple combination of μ and a_1.**

Which Model is Better?

Suggests the simple question: **“are these data drawn from the same distribution?"**

Here comes the **‘Bayes Factor’** or **‘Weight of Evidence’.**

Suppose we try to describe all of the data **X_i, Y_i** with just one distribution **G**. This distribution may have parameters so call this hypothesis by **(G, θ).**

Alternatively (and by hypothesis exhaustively) we may use **(G_x, θ_x)** for the data **X_i** and **(G_y, θ_y)** for the data **Y_i**. This hypothesis is **(G_x,θ_x,G_y,θ_y).** Note we need **prior probabilities** for our two options, **G** or **G_xG_y**. Bayes' theorem then tells us

$$\mathrm{prob}(G,\theta \mid X,Y) = \frac{\mathrm{prob}(X,Y \mid G,\theta)\mathrm{prob}(G,\theta)}{(\int \mathrm{prob}(G,\theta \mid X,Y)d\theta + \int \mathrm{prob}(G_x,\theta_x \mid X)d\theta_x \int \mathrm{prob}(G_y,\theta_y \mid Y)d\theta_y)}$$

in which the second term of the denominator arises because our alternative to **(G, θ)** is that the data are described as the product of two distinct distributions. The **odds** on the distinct distributions are

$$\frac{\int \mathrm{prob}(G_x,\theta_x \mid X)d\theta_x \int \mathrm{prob}(G_y,\theta_y \mid Y)d\theta_y}{\int \mathrm{prob}(G,\theta \mid X,Y)d\theta},$$

and this ratio is closely related to the **Bayes Factor**. To work out these odds we integrate the likelihood functions, weighted by the priors, over the range of parameters of the distributions.

Which Model is Better? - Example

We have the following two data sets:
X_i = -0.16, 0.12, 0.44, 0.60, 0.70, 0.87, 0.88, 1.44, 1.74, 2.79
Y_i = 0.89, 0.99, 1.29, 1.73, 1.96, 2.35, 2.51, 2.79, 3.17, 3.76.

The **means differ by about one std dev**. We consider two *a priori* equally likely hypotheses. One is that **all 20 data are drawn from the same Gaussian**. The other is that **they are drawn from different Gaussians.**

In the first case, the likelihood function is

$$\frac{1}{(\sqrt{2\pi}\sigma)^{20}} \exp\left[-\frac{\sum_i (X_i-\mu)^2 + \sum (Y_i-\mu)^2}{2\sigma^2}\right]$$

and we take the prior on **σ** to be **1/σ**. We also assume a **uniform prior** for the **μ**'s.

In the second case, the likelihood function is

$$\frac{1}{(\sqrt{2\pi}\sigma_x)^{10}} \exp\left[-\frac{\sum_i (X_i-\mu_x)^2}{2\sigma_x^2}\right] \frac{1}{(\sqrt{2\pi}\sigma_y)^{10}} \exp\left[-\frac{\sum_i (Y_i-\mu_y)^2}{2\sigma_y^2}\right]$$

and the prior is $1/\sigma_x\sigma_y$. Integrating over the range of the **μ**'s and **σ**'s, the odds on the data being drawn from different Gaussians are about 40 to 1 - **a good bet.**

PCA