

# Statistics for experimental data analysis

Hari Bharadwaj

Research Assistant, Auditory Neuroscience Laboratory, Boston University

February 7, 2013



# Outline

- 1 Detection and Hypothesis testing
- 2 Some distributions
- 3 The Universal Frequentist Recipe
- 4 Common traditional test statistics
- 5 ANOVA & The General Linear Model (GLM) perspective
  - Some design matrices
  - Contrasts and Interactions
- 6 Non-parametric approaches
- 7 Multiple Comparisons and Topological Inference
- 8 False Discovery rates
- 9 Miscellaneous Issues



## 2 Alternatives discrimination problem

$\mathcal{H}_1$ : **There is an 'effect'**

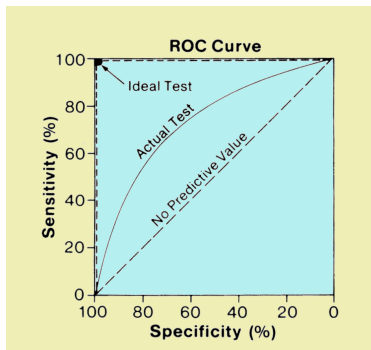
$\mathcal{H}_0$ : **There is no 'effect'**

- Example:  $\mathcal{H}_1$ : Average IQ of Group1 subjects < Group2 subjects  
 $\mathcal{H}_0$ : Average IQ of Group1 subjects = Group2 subjects
- Given data we wish to probabilistically test out the hypotheses
- Frequentist: Is  $p(\text{data}|\mathcal{H}_0) < 0.05$  (or anything else arbitrary) ?
- Bayesian: How do  $p(\mathcal{H}_0|\text{data})$  and  $p(\mathcal{H}_1|\text{data})$  compare?



# Frequentist and Bayesian approaches

- **Frequentist** - When  $\mathcal{H}_0$  is true, what is the probability (p value) that we'll see the data that we have i.e  $p(\text{data}|\mathcal{H}_0)$ ?
- **Bayesian** - Given the data we have, what is the probability that  $\mathcal{H}_0$  is true i.e  $p(\mathcal{H}_0|\text{data})$ ? Which is more likely:  $\mathcal{H}_0$  or  $\mathcal{H}_1$ ?
- ROC curve - Hit (no type II error) probability versus False Alarm (type I error) probability



# Normal distribution

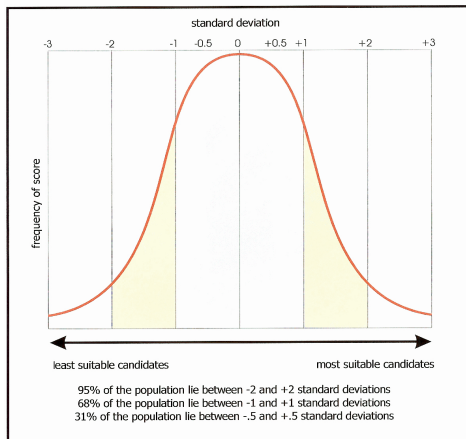


Figure:  $p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ , Normal distributions are good models of most real life data where clustering around the average happens, example: Adult human height

# 'Alien' example

- $\mathcal{H}_1$ : **A** is an alien  
 $\mathcal{H}_0$ : **A** is a human being
- Given: Adult human height is normally distributed with  $\mu = 170\text{cm}$  and  $\sigma = 10\text{ cm}$
- **A** is 195 cm tall (Our data)
- Frequentist: Given  $\mathcal{H}_0$ , the height of **A** is normally distributed
- $p(\chi > \mu + 2\sigma) < 0.05 \Rightarrow$  With  $p < 0.05$ ,  $\mathcal{H}_0$  is false. Is **A** is an alien?
- What if all aliens were shorter than 100cm?



# Frequentist versus Bayesian

Clinical test to screen school children for a certain disease. The test is 96% accurate. That is, if the test is administered on a population of children with disease ( $\mathcal{H}_1$ ), it tests +ve 96% of the time. Similarly if we test a population of children with no disease ( $\mathcal{H}_0$ ), it tests -ve 96% percent of the time.

- Is this a good test?
- If a random school child tests positive:
  - 1 What is the conclusion based on the frequentist approach with a  $p < 0.05$  threshold?
  - 2 What is the probability that he/she actually has the disease?

Bayes Rule:  $p(\mathcal{H}_0|data) \propto p(data|\mathcal{H}_0)p(\mathcal{H}_0)$



# Where we are...

- 1 Detection and Hypothesis testing
- 2 Some distributions
- 3 The Universal Frequentist Recipe**
- 4 Common traditional test statistics
- 5 ANOVA & The General Linear Model (GLM) perspective
  - Some design matrices
  - Contrasts and Interactions
- 6 Non-parametric approaches
- 7 Multiple Comparisons and Topological Inference
- 8 False Discovery rates
- 9 Miscellaneous Issues





# Universal Frequentist Recipe

**ALL** univariate statistical tests entail the following:

- 1 Construct  $\mathcal{H}_0$  and  $\mathcal{H}_1$ , could be competing models



# Universal Frequentist Recipe

**ALL** univariate statistical tests entail the following:

- 1 Construct  $\mathcal{H}_0$  and  $\mathcal{H}_1$ , could be competing models
- 2 Calculate a statistic, a scalar ( $T$ ), that summarizes the effect you are trying to capture (example: difference in mean IQs of 2 groups)



# Universal Frequentist Recipe

**ALL** univariate statistical tests entail the following:

- 1 Construct  $\mathcal{H}_0$  and  $\mathcal{H}_1$ , could be competing models
- 2 Calculate a statistic, a scalar ( $T$ ), that summarizes the effect you are trying to capture (example: difference in mean IQs of 2 groups)
- 3 Determine the distribution of  $T$  when  $\mathcal{H}_0$  is true (Here is where usually many assumptions come in)



# Universal Frequentist Recipe

**ALL** univariate statistical tests entail the following:

- 1 Construct  $\mathcal{H}_0$  and  $\mathcal{H}_1$ , could be competing models
- 2 Calculate a statistic, a scalar ( $T$ ), that summarizes the effect you are trying to capture (example: difference in mean IQs of 2 groups)
- 3 Determine the distribution of  $T$  when  $\mathcal{H}_0$  is true (Here is where usually many assumptions come in)
- 4 If  $p(T|\mathcal{H}_0) < 0.05$  or any other *ad hoc* threshold, reject  $\mathcal{H}_0$  (This doesn't necessarily mean we have evidence for  $\mathcal{H}_1$ )



# Important properties of the normal distribution

- Linear combinations of IID normal variables is a normal variable  $\Rightarrow$   
Average of IID normal variables is normal
- Sum of squares of  $k$  **zero mean** normal normal variables is a  $\chi^2$  variable with  $k$  degrees of freedom
- Ratio of a **zero mean** normal variable and square root of a  $\chi^2$  variable (with  $k$  df) is a **t** variable with  $k$  degrees of freedom

$$t = \frac{\chi}{\sqrt{S/k}} \quad (1)$$

- Ratio of two **independent**  $\chi^2$  variables is an **F** variable

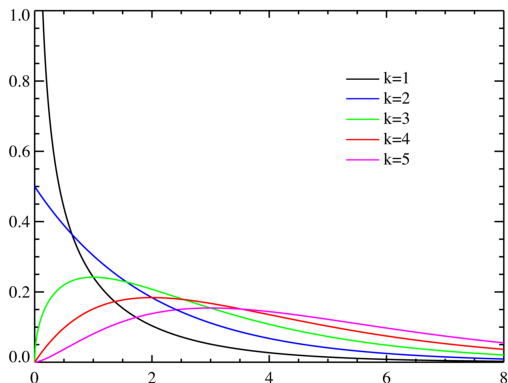
$$F = \frac{S_1/k_1}{S_2/k_2} \quad (2)$$

F has degrees of freedom  $k_1$  and  $k_2$



$\chi^2$  distribution

$$S = x_1^2 + x_2^2 + \cdots + x_k^2$$



**Figure:** Sum of squares **Independent and Identically distributed** normal variables with mean 0 and variance 1



# t distribution

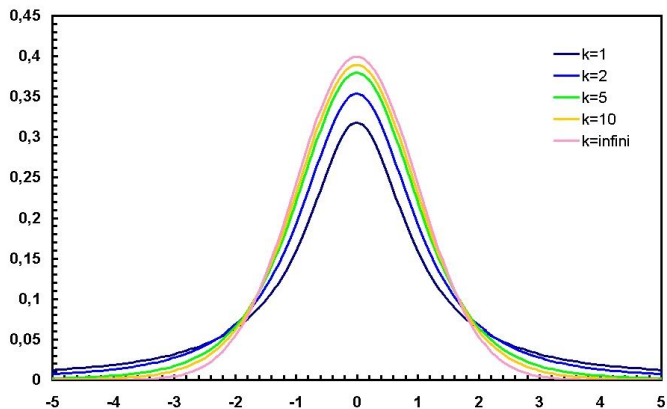


Figure: Ratio of zero mean normal and square root of a  $\chi^2$  distribution



# F distribution

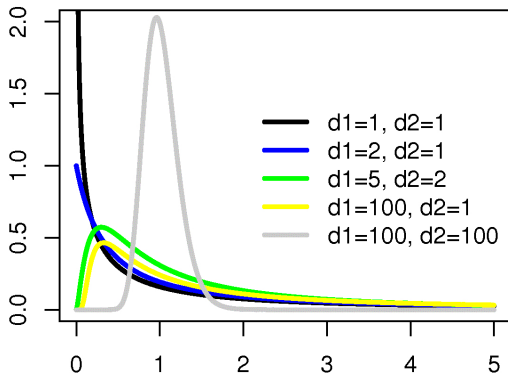


Figure: Ratio of 2  $\chi^2$  distributions





# Where we are...

- 1 Detection and Hypothesis testing
- 2 Some distributions
- 3 The Universal Frequentist Recipe
- 4 Common traditional test statistics**
- 5 ANOVA & The General Linear Model (GLM) perspective
  - Some design matrices
  - Contrasts and Interactions
- 6 Non-parametric approaches
- 7 Multiple Comparisons and Topological Inference
- 8 False Discovery rates
- 9 Miscellaneous Issues



# One sample $t$ -test

- Testing for the average of a normal **population** to have a certain mean  $\mu_0$
- Example: **sample** of 10 subjects  
 $\mathcal{H}_1$ : The average IQ of TDs is different from 100  
 $\mathcal{H}_0$ : The average IQ of TDs is 100
- IQs = 87, 110, 93, 99, 75, 102, 90, 83, 100, 70

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_k}{k} \quad (3)$$

$$S = \frac{1}{k-1} \sum_1^k (x_i - \bar{x})^2 \quad (4)$$

$$t = \frac{\bar{x} - 100}{\sqrt{S/k}} \quad (5)$$

- $t = -2.3$ ,  $p = 0.047 \Rightarrow \mathcal{H}_0$  is rejected



## Two (independent) sample $t$ -test

- Testing for the means of 2 independent **populations** to be equal
- Example: **sample** of 10 subjects in each group (need not be same number)

$\mathcal{H}_1$ : The average IQ of TDs is different from ASDs

$\mathcal{H}_0$ : The average IQ of TDs is same as ASDs

- TDs = 87, 110, 93, 99, 75, 102, 90, 83, 100, 70  
ASDs = 77, 81, 64, 100, 84, 72, 69, 90, 68, 70

$$\bar{x}_{td}, \bar{x}_{asd} = \frac{x_1 + x_2 + \cdots + x_k}{k} \quad (6)$$

$$S_{td}, S_{asd} = \frac{1}{k-1} \sum_{i=1}^k (x_i - \bar{x})^2 \quad (7)$$

$$t = \frac{\bar{x}_{td} - \bar{x}_{asd}}{\sqrt{S_{asd}/k_{asd} + S_{td}/k_{td}}} \quad (8)$$

- $\mathcal{H}_1$  can be **one-sided**: IQ of TDs > IQ of ASD



# Paired $t$ -test

- Testing for changes between conditions in the same block (subject)
- Example: **sample** of 10 TDs at ages 5 and 15  
 $\mathcal{H}_1$ : IQ increases when you grow to 15 from 5  
 $\mathcal{H}_0$ : IQ does not change between ages 5 and 15
- 15 years = 87, 110, 93, 99, 75, 102, 90, 83, 100, 70  
 5 years = 68, 90, 63, 80, 70, 70, 88, 83, 90, 60

$$\bar{x} = \frac{(x_1^{15y} - x_1^{5y}) + (x_2^{15y} - x_2^{5y}) + \dots + (x_k^{15y} - x_k^{5y})}{k} \quad (9)$$

$$S = \frac{1}{k-1} \sum_1^k (x_i^{15y} - x_i^{5y} - \bar{x})^2 \quad (10)$$

$$t = \frac{\bar{x}}{\sqrt{S}} \quad (11)$$

- More 'sensitive' than an unpaired ( $\mathcal{H}_0$ : IQs of 5 and 15 year olds is the same on an average), Unpaired with this data  $\Rightarrow$  block effects!



# Summary of basic tests

- One sample t-test: IID normal data, test for mean being a certain value



# Summary of basic tests

- One sample t-test: IID normal data, test for mean being a certain value
- 2 sample t-test: IID normal data, test for means being equal, If variance unequal, then its called a Behrens-Fisher problem



# Summary of basic tests

- One sample t-test: IID normal data, test for mean being a certain value
- 2 sample t-test: IID normal data, test for means being equal, If variance unequal, then its called a Behrens-Fisher problem
- Paired t-test: IID normal differences, test for paired differences/changes being 0 or not 0.



# Summary of basic tests

- One sample t-test: IID normal data, test for mean being a certain value
- 2 sample t-test: IID normal data, test for means being equal, If variance unequal, then its called a Behrens-Fisher problem
- Paired t-test: IID normal differences, test for paired differences/changes being 0 or not 0.
- Depending on what  $\mathcal{H}_1$  is, the test is one-sided or two sided





# Summary of basic tests

- One sample t-test: IID normal data, test for mean being a certain value
- 2 sample t-test: IID normal data, test for means being equal, If variance unequal, then its called a Behrens-Fisher problem
- Paired t-test: IID normal differences, test for paired differences/changes being 0 or not 0.
- Depending on what  $\mathcal{H}_1$  is, the test is one-sided or two sided
- We know the significance under the null hypothesis  $\Rightarrow$  We dont know the sensitivity, we only guarentee a specificity



# Summary of basic tests

- One sample t-test: IID normal data, test for mean being a certain value
- 2 sample t-test: IID normal data, test for means being equal, If variance unequal, then its called a Behrens-Fisher problem
- Paired t-test: IID normal differences, test for paired differences/changes being 0 or not 0.
- Depending on what  $\mathcal{H}_1$  is, the test is one-sided or two sided
- We know the significance under the null hypothesis  $\Rightarrow$  We dont know the sensitivity, we only guarentee a specificity
- To know the sensitivity, (i.e) the probability that we detect an 'effect' when **there is** an effect, we need to analyze distributions of data under  $\mathcal{H}_1$



# Where we are...

- 1 Detection and Hypothesis testing
- 2 Some distributions
- 3 The Universal Frequentist Recipe
- 4 Common traditional test statistics
- 5 ANOVA & The General Linear Model (GLM) perspective**
  - Some design matrices
  - Contrasts and Interactions
- 6 Non-parametric approaches
- 7 Multiple Comparisons and Topological Inference
- 8 False Discovery rates
- 9 Miscellaneous Issues



# ANOVA basics

- Observed data ( $y$ ) is modeled as coming from a normal population
- Conditional mean of  $y$  is modeled as a linear function of explanatory variables ( $\mathbf{x}$ )
- $\mathcal{H}_1$ :  $y$  depends on all the variables in  $\mathbf{x}$

$$\begin{aligned} E(y|\mathbf{x}) &= \beta\mathbf{x} \\ y &= \beta\mathbf{x} + \epsilon \end{aligned}$$

$\mathcal{H}_0$ :  $y$  depends only on  $\mathbf{x}_0 \subset \mathbf{x}$

$$E(y|\mathbf{x}_0) = \beta\mathbf{x}_0$$

- The 2 models of the data are compared in the least squared sense to generate an **F**-statistic



## Example - continuous factors

A study with  $n$  subjects of different ages and heights (at one time)

- $\mathcal{H}_1$ : Occipital alpha power depends on age and height  
 $\mathcal{H}_0$ : Occipital alpha power depends only on height
- For each subject the alpha power  $y_i$  is measured
- $\mathcal{H}_1 : y_i = (a_i, h_i, 1)(\beta_a, \beta_h, \beta_\mu)^T + \epsilon_i$   
 $\mathcal{H}_0 : y_i = (h_i, 1)(\beta_h, \beta_\mu)^T + \epsilon_i$
- $S_{1,0} = \sum_{i=1}^n \epsilon_i^2$  is the model error for  $\mathcal{H}_1$  and  $\mathcal{H}_0$

$$F_{age} = \frac{(S_0 - S_1)/(k_0 - k_1)}{S_1/k_1}$$

Under  $\mathcal{H}_0$  the ratio has an F-distribution

- If  $p(F > F_{age} | \mathcal{H}_0) < 0.05$ , then  $\mathcal{H}_0$  is rejected and age is said to have a main effect on alpha power

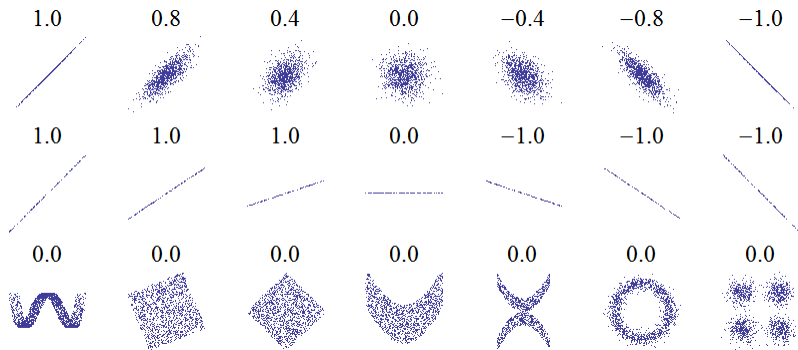


# One way ANOVA with continuous explanatory variable: Correlation

- Does IQ depend on age?
- Same as asking 'Is IQ correlated with AGE'?
- 10 subjects:
- IQ ( $y$ )= 87, 110, 93, 99, 75, 102, 90, 83, 100, 70  
AGE ( $x$ )= 9,15,9,10,10,12,8,10,11,7
- $y = \mu + \beta x + \epsilon$  versus  $y = \mu + \epsilon$  : Is one significantly better than the other
- $F$  - test would give us the answer,  $p = 0.0095$
- Alternate way to test the significance of Correlation ( $\rho$ ): Fisher RA, 1915: When  $x$  and  $y$  are jointly normal,  $0.5 \log \frac{1+\rho}{1-\rho}$  is normally distributed with mean  $0.5 \log \frac{1+\rho_0}{1-\rho_0}$  and variance  $\frac{1}{N-3}$ , where  $\rho_0$  is the actual population correlation



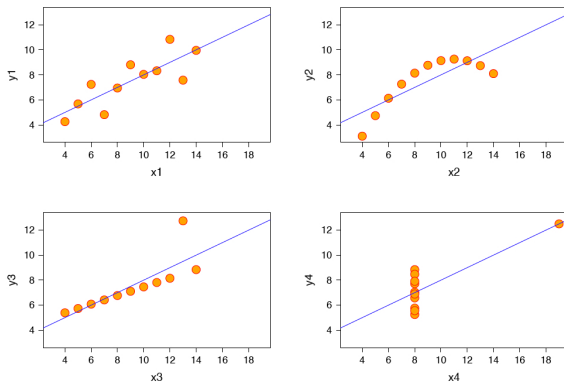
# Linear and non-linear predictability



**Figure:** Correlation just says  $y$  is linearly predictable from  $x$ . Lower correlation  $\Rightarrow$  Higher prediction error. Perfect dependence could result in zero correlation if the dependence is non-linear.



# Outliers and bad models



**Figure:** All the 4 cases have the exact same correlation coefficient of about 0.8. One should plot and look at the curves. A log-linear model might fit better.





## Design Matrix - Model specification

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} a_1 & | & h_1 & | & 1 \\ a_2 & | & h_1 & | & 1 \\ \vdots & & \vdots & & \\ a_n & | & h_n & | & 1 \end{pmatrix} \begin{pmatrix} \beta_a \\ \beta_h \\ \beta_\mu \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$Y = (\mathbf{a}, \mathbf{h}, \mathbf{1})\beta + \epsilon$$

$$Y = \mathbf{X}\beta + \epsilon$$

- $\mathbf{X}$  is called the design matrix
- $Y = (y_1, y_2, \dots, y_n)^T \in \mathcal{R}^n$
- The projection length ( $S_0 - S_1$ ) of  $Y$  onto the subspace spanned by  $\mathbf{a}$  is the variance of  $Y$  that is explained by age (1 degree of freedom)
- The size of the orthogonal projection ( $S_1$ ) is the model error ( $n - 1$  degrees of freedom)
- Thus  $F$  will be small if  $\mathbf{a}$  does not account for the variance in  $Y$  significantly



# Example - Categorical Factors

Study of 20 subjects divided into 2 groups

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ \vdots & \vdots \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{n-1} \\ \epsilon_n \end{pmatrix}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ \vdots & \vdots \\ -1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{n-1} \\ \epsilon_n \end{pmatrix}$$



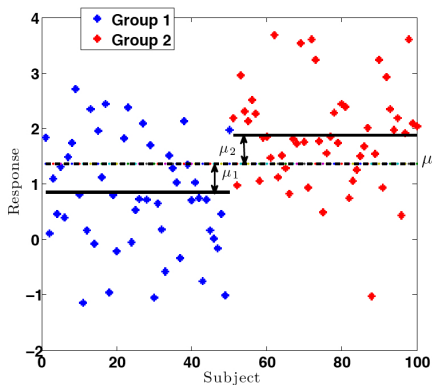
## In general

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \left( \mathbf{G}_1 \quad \mathbf{H}_1 \mid \mathbf{G}_0 \quad \mathbf{H}_0 \right) \begin{pmatrix} \gamma_1 \\ \kappa_1 \\ \gamma_0 \\ \kappa_0 \end{pmatrix} + \boldsymbol{\epsilon}$$

- All design using linear models and assuming a normal distribution with common error covariances are an instance of the above
- $\mathbf{G}_1$  and  $\mathbf{H}_1$  are interesting categorical and continuous factors respectively
- $\mathbf{G}_0$  and  $\mathbf{H}_0$  are uninteresting categorical and continuous factors
- The null model contains only the partition with  $\mathbf{G}_0$  and  $\mathbf{H}_0$
- Experiment design is equivalent to deciding on what the design matrix is



# GLM with 1 factor (group) $\rightarrow$ One-Way ANOVA with 2 levels



**Figure:** The data  $y$  is explained by 1 factor, namely 'group'  $x = 0$  or  $1$  denoting Group1 or Group2 for example. Does regressing  $y$  as a linear function of  $x$  help explain the variance in  $y$  better than when not modeled as a function of  $x$ ?

# F distribution - A reminder

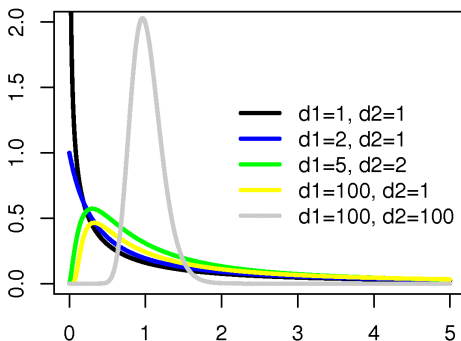
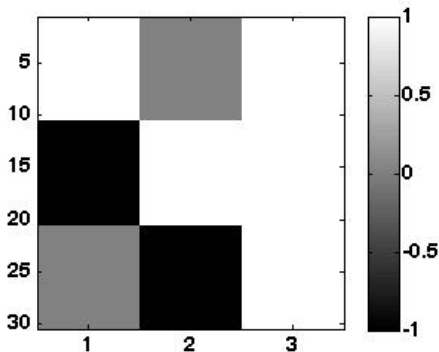


Figure: Distribution of sum of squared mean-zero normal variables



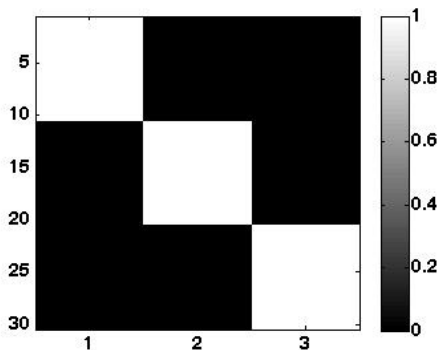
# Design matrix for 1 way 3 level ANOVA with 30 subjects



**Figure:** One way 3 level ANOVA has 3 experimental effects: 2 Group Differences and 1 Overall Mean



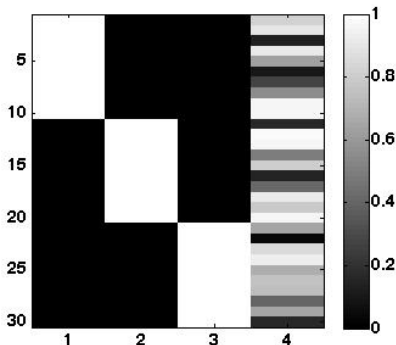
# Design matrix - cell mode



**Figure:** Equivalent design matrix as 2 group differences and 1 overall mean: 3 different group means



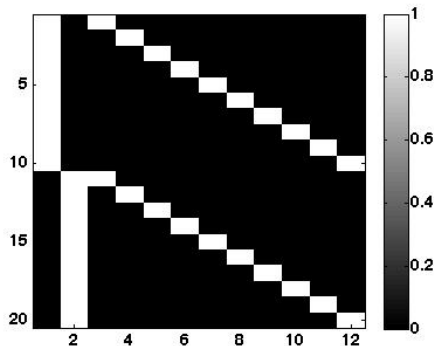
# Design matrix - 2 way ANOVA: 1 categorical and 1 continuous factor



**Figure:** Design matrix for 30 subjects divided into 3 groups of 10 with AGE as a covariate/factor. **What is the NULL model?** - A subset of the full design matrix

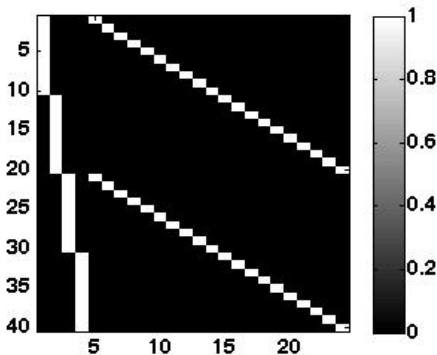


# Design matrix - 1 way repeated measures ANOVA with 2 conditions



**Figure:** 1 way ANOVA with **block (subject)** effects, 20 subjects each measured in 2 conditions: First 2 columns are the cells corresponding to the conditions and then other 10 model effects. **What is the NULL model?**

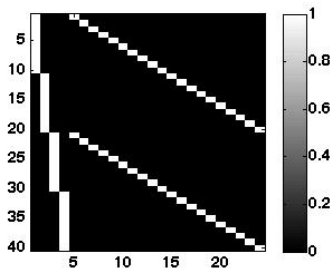
## 2 Groups, 2 Conditions: 2 way ANOVA with interactions



**Figure:** 2 way ANOVA with **block (subject)** effects, 20 subjects each measured in 2 conditions, divided into 2 groups: First 4 columns are the cells corresponding to every condition-group pair: (cond1, grp1), (cond1, grp2), (cond2,grp1) and (cond2,group2). **How do we assess the main effect of group?**



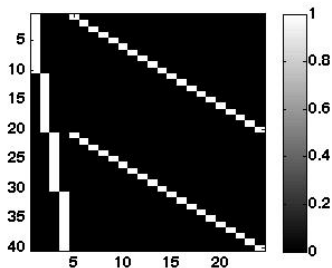
# Contrast for main effect of group



- The cells for the first 4 columns are (cond1, grp1), (cond1, grp2), (cond2,grp1) and (cond2,group2)
- Let  $\mathbf{c} = (1, -1, 1, -1, 0, \dots, 0)^T$  : The data in the subspace of  $\mathbf{Xc}$  represent the variance because of group differences **averaging over conditions**
- Contrast matrix for **main effect of group**



# Contrast for main effect of condition



- The cells for the first 4 columns are (cond1, grp1), (cond1, grp2), (cond2,grp1) and (cond2,group2)
- Let  $\mathbf{c} = (1, 1, -1, -1, 0, \dots, 0)^T$  : The data in the subspace of  $\mathbf{Xc}$  represent the variance because of condition differences **averaging over groups**
- Contrast matrix for **main effect of conditions**

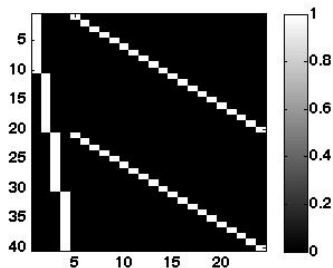


# Interactions

- The **effect** of one factor may depend on the **level** of another factor
- Example: Sleep hours modeled as a function of amount of exercise and weight of a person: Regular exercise increases the amount of sleep more for heavier people than for lighter people
- For our 2 group - 2 condition example: There may be group differences that are condition independent (main effects) but there might be group differences that occur only in one condition (interaction)
- If  $x$  and  $y$  are the factors, an interaction is a dependence on  $xy$



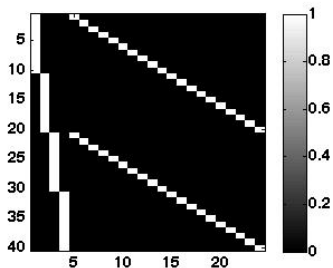
# Contrast for interaction between group and condition



- The cells for the first 4 columns are (cond1, grp1), (cond1, grp2), (cond2,grp1) and (cond2,group2)
- Let  $\mathbf{c} = (1, -1, -1, 1, 0, \dots, 0)^T$  : The data in the subspace of  $\mathbf{Xc}$  represents the interaction **Difference of differences**
- Contrast matrix for **Interaction**



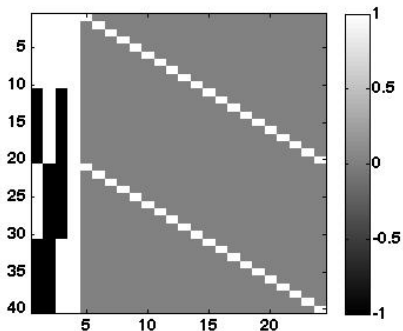
# Contrast for overall mean



- The cells for the first 4 columns are (cond1, grp1), (cond1, grp2), (cond2,grp1) and (cond2,group2)
- Let  $\mathbf{c} = (1, 1, 1, 1, 0, \dots, 0)^T$  : The data in the subspace of  $\mathbf{Xc}$  represents the effects common to each of the cells
- Contrast matrix for **overall mean**



# With experimental effects along the columns of the design matrix



- Column 1 is group difference, column 2 is condition
- Column 3: interaction (Note that this is column1 (dot) column2)
- Column 4: Overall mean





# Where we are...

- 1 Detection and Hypothesis testing
- 2 Some distributions
- 3 The Universal Frequentist Recipe
- 4 Common traditional test statistics
- 5 ANOVA & The General Linear Model (GLM) perspective
  - Some design matrices
  - Contrasts and Interactions
- 6 Non-parametric approaches**
- 7 Multiple Comparisons and Topological Inference
- 8 False Discovery rates
- 9 Miscellaneous Issues



# Universal Frequentist Recipe

**ALL** univariate statistical tests entail the following:

- 1 Construct  $\mathcal{H}_0$  and  $\mathcal{H}_1$ , could be competing models



# Universal Frequentist Recipe

**ALL** univariate statistical tests entail the following:

- 1 Construct  $\mathcal{H}_0$  and  $\mathcal{H}_1$ , could be competing models
- 2 Calculate a statistic, a scalar ( $T$ ), that summarizes the effect you are trying to capture (example: difference in mean IQs of 2 groups)



# Universal Frequentist Recipe

**ALL** univariate statistical tests entail the following:

- 1 Construct  $\mathcal{H}_0$  and  $\mathcal{H}_1$ , could be competing models
- 2 Calculate a statistic, a scalar ( $T$ ), that summarizes the effect you are trying to capture (example: difference in mean IQs of 2 groups)
- 3 Determine the distribution of  $T$  when  $\mathcal{H}_0$  is true (Here is where usually many assumptions come in)



# Universal Frequentist Recipe

**ALL** univariate statistical tests entail the following:

- 1 Construct  $\mathcal{H}_0$  and  $\mathcal{H}_1$ , could be competing models
- 2 Calculate a statistic, a scalar ( $T$ ), that summarizes the effect you are trying to capture (example: difference in mean IQs of 2 groups)
- 3 Determine the distribution of  $T$  when  $\mathcal{H}_0$  is true (Here is where usually many assumptions come in)
- 4 If  $p(T|\mathcal{H}_0) < 0.05$  or any other *ad hoc* threshold, reject  $\mathcal{H}_0$  (This doesn't necessarily mean we have evidence for  $\mathcal{H}_1$ )



# Permutation tests: Example 1

IQ of 2 groups of 10 subjects each: Use our **recipe**

- Let  $T = \text{mean IQ of group 1} - \text{mean IQ of group 2}$
- Under  $\mathcal{H}_0$ , we want to know what the distribution of  $T$  is
- Non-parametric approach: Under  $\mathcal{H}_0$ , group does not have any effect on data  $\Rightarrow$  We can assign group to subjects randomly
- Thus we can get many groupings, here we can have up to  $\binom{20}{10} > 180,000$  permutations where the subjects from the 2 groups are mixed
- For each of these permutations we can get a  $T_{perm} \Rightarrow$  We have a distribution for  $T$  under  $\mathcal{H}_0$
- Is this generalizable to the population or applicable only to the cohort?



# Permutation tests: Example 2

10 subjects who are politicians or have an IQ score less than 80 or both

$\mathcal{H}_1$ : Politicians are more likely to have IQ  $< 80$

$\mathcal{H}_0$ : They are unrelated attributes

- The data is not normally distributed, its categorical
- Let  $x_1 = (0 \ 1 \ 1 \ 0 \ 0 \ 0 \ 1 \ 1 \ 1 \ 1)$  and  $x_2 = (1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 1 \ 1 \ 1 \ 1)$  denote politician or not and IQ less than 80 or not respectively for the 10 subjects
- $\mathbf{d} = \text{norm}(x_1 - x_2)$  is a good measure of the conjunction between the 2 attributes,  $d = \sqrt{2}$  here
- Permute  $x_1$  or  $x_2$  values randomly and get a distribution for  $\mathbf{d}$  and find  $p(d \leq \sqrt{2})$



# Where we are...

- 1 Detection and Hypothesis testing
- 2 Some distributions
- 3 The Universal Frequentist Recipe
- 4 Common traditional test statistics
- 5 ANOVA & The General Linear Model (GLM) perspective
  - Some design matrices
  - Contrasts and Interactions
- 6 Non-parametric approaches
- 7 Multiple Comparisons and Topological Inference**
- 8 False Discovery rates
- 9 Miscellaneous Issues





## MCP

$\mathcal{H}_1$ : Coin is biased

$\mathcal{H}_0$ : Coin is unbiased

- Test: Toss coin 10 times, if Head or Tail shows up 9 or more times, reject  $\mathcal{H}_0$  ( $p(9 \text{ or more heads}/\mathcal{H}_0) \approx 0.02$ )
- This means if we repeat the test 100 times we'll get 9 or more heads only 2 times on an average
- What if we have a million coins and test each of them with this test?
- On an average 20,000 coins will turn head more than 9 times even when non of them are biased  $\Rightarrow$  We have a family wise error which we **must** correct for



# Why should we worry about the MCP

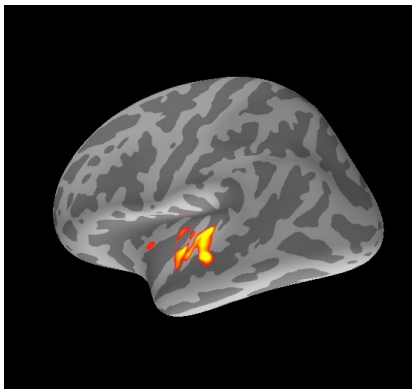


Figure: Of the order of 10,000 sources  $\Rightarrow$  Large number of correlated tests



# Why should we worry about the MCP

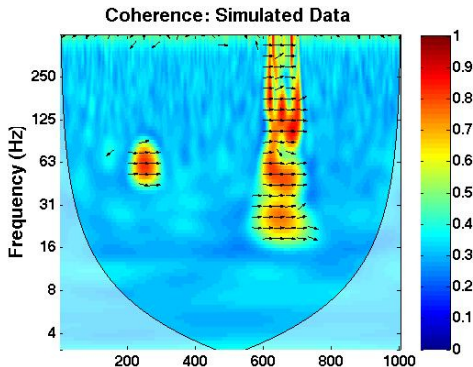


Figure: 1000 time bins  $\times$  50 frequency bins  $\Rightarrow$  Large number of correlated tests

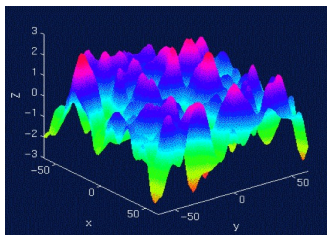


# Multiple testing corrections

- For discrete tests (example: multiple end point drug trials):  
Non-parametric family-wise testing or False Discovery Rate (**FDR**) approaches
- For data sets with an inherent topology (example: Time courses, whole brain signals, time-frequency maps): Random field theory or Non-parametric topological inference tests



# Topological Inference



- We have a topological map of statistics (mean power, t-values, F-values, TF coherence etc.)
- Unlikely excursions under  $\mathcal{H}_0$  of this map should be identified as evidence for  $\mathcal{H}_1$



# How to set thresholds?

Given a statistical map (example t-test at each source)

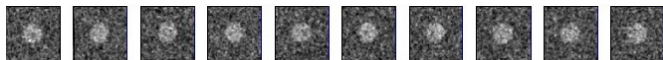


Figure: Simulated: signal+noise

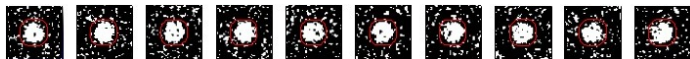


Figure: Thresholding at  $\alpha < 0.05$  at each voxel  $\Rightarrow$  Lots of significant voxels outside of true signal



Figure: Bonferroni thresholding at  $FWE < 0.05$  at each voxel  $\Rightarrow$  Too conservative

# Familywise Error Rate (FWE)

- Each observation is a topological map (TF maps, Scalp power, Whole brain response etc.)
- Example: Wavelet coherence data between STG and rIFG during a Roving paradigm
- 10 ASD and 10 Control Subjects
- No apriori hypothesis about any particular frequency band or time frame:  
 **$\mathcal{H}_0$ : There are no differences in coherence between groups anywhere in the TF plane**
- The hypothesis is not about any TF bin  $\Rightarrow$  inference is also about the whole map
- p-value is a familywise p-value



# RFT versus Bonferroni correction

Example:  $100 \times 100$  images: Bonferroni Correction too conservative when smooth

RFT models error fields (our  $\epsilon$  data) as a Gaussian random fields



Figure: Null field with no spatial covariance  $\rightarrow$  10,000 elements

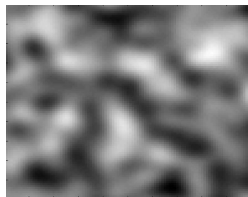


Figure: Smooth null field  $\rightarrow$  100 elements





# Height thresholding

For an *ad hoc* threshold ( $u$ ), we want to find the familywise p-value of each blob that we see. Example: **Under  $\mathcal{H}_0$ , what is the probability that you'll find a peak  $> u$  anywhere** ? Can be answered using RFT or permutations/other non-parametric

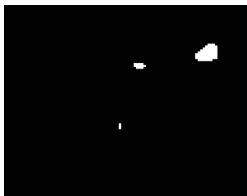


Figure: Thresholded at  $u = 2.5$



Figure: Thresholded at  $u = 3$



# RFT - height threshold (peak level inference)

Assumptions:

- ① Error fields conform to a reasonable lattice approximation of a gaussian RF
- ② The covariance of the error fields is continuous and differentiable (**need not be spatially stationary**)

Once we have established that

- Euler characteristic ( $EC$ ): Property of a map upon thresholding (#blobs - #holes)
- For large thresholds  $p(\text{peak} > u | \mathcal{H}_0) = E(EC | \mathcal{H}_0)$
- $E(EC)$  have for random z-fields, t-fields or F-fields

$$E(EC) = R(4\ln 2)(2\pi)^{-3/2} z e^{(-z^2/2)}$$

- The only data dependent parameter for calculating  $E(EC)$  is the smoothness (FWHM) (specified through  $R$ ) at each voxel or TF element



# RFT - Cluster extent threshold

- Like sharp large peaks, wide plateaus (albeit not so tall) are also unlikely excursions under  $\mathcal{H}_0$
- Inference can be made about the extent of a cluster above an *ad hoc* threshold  $u$
- **Under  $\mathcal{H}_0$ , what is the probability that you'll find a cluster/blob containing more than  $k$  voxels above a threshold  $u$**
- This can also be calculated from from RFT with only the smoothness being specified from the data



# Permutation test

Procedure to determine p-value for height and cluster extent above a threshold  $u$

- 1 Generate a large number  $N$  of random permutations of data (example permutations of 'group' or 'condition')
- 2 The proportion of permutations having a peak  $> u$  **anywhere** is the p-value for the height =  $u$
- 3 The proportion of permutations having clusters  $> u$  containing  $k$  or more voxels **anywhere** gives the cluster extent p-values
- 4 A hybrid measure of 'exceedence mass' ( $m$ ) could be calculated as the mass of the blobs exceeding  $u$ : Sensitive to both sharp tall peaks and flat wide plateaus



# Where we are...

- 1 Detection and Hypothesis testing
- 2 Some distributions
- 3 The Universal Frequentist Recipe
- 4 Common traditional test statistics
- 5 ANOVA & The General Linear Model (GLM) perspective
  - Some design matrices
  - Contrasts and Interactions
- 6 Non-parametric approaches
- 7 Multiple Comparisons and Topological Inference
- 8 **False Discovery rates**
- 9 Miscellaneous Issues



# False Discovery Rate procedure

- New approach to multiple testing
- Instead of controlling for FWE or p-values, control for the the 'False Discovery Rate'
- $FDR = E(\text{proportion of rejected null hypotheses that are falsely rejected})$

When we test for  $m$  null hypotheses of which  $m_0$  are true

	#Accepted	#Rejected	Total
# True	U	V	$m_0$
# False	T	S	$m - m_0$
Total	$m - R$	R	m

$$FDR = q\text{-value} = E\left(\frac{V}{V+S}\right) = E\left(\frac{V}{R}\right)$$



## FDR

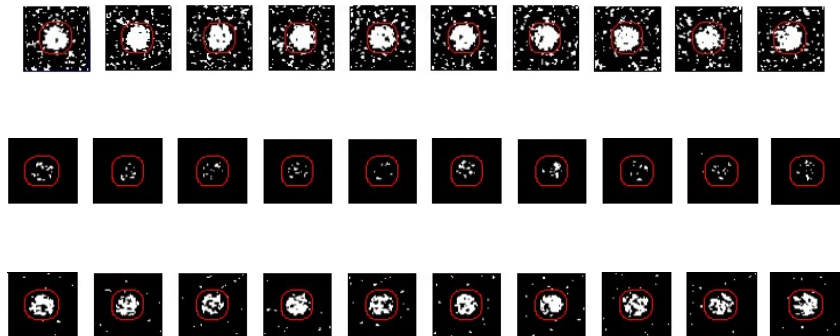


Figure: Setting FDR is not as conservative as bonferroni, we accept some false discoveries anyway



# Where we are...

- 1 Detection and Hypothesis testing
- 2 Some distributions
- 3 The Universal Frequentist Recipe
- 4 Common traditional test statistics
- 5 ANOVA & The General Linear Model (GLM) perspective
  - Some design matrices
  - Contrasts and Interactions
- 6 Non-parametric approaches
- 7 Multiple Comparisons and Topological Inference
- 8 False Discovery rates
- 9 Miscellaneous Issues





# Model misspecification and assumptions

- Non-normality of data: EEG power, Coherence → Transformations can be applied:  $\log(\text{power})$ ,  $\text{arctanh}(\text{coherence})$  or consider only 'differences'
- Estimation bias: Coherence is biased on the number of trials ( $n$ ) i.e.  $E(\text{coherence}) = \text{TrueCoherence} + \frac{1}{2n-2}$  → When comparing groups or conditions with unequal number of trials, corrections have to be applied
- Variance of data different between groups of conditions → Hierarchical models with partitioned errors
- Correlated measurements → Greenhouse - Geisser correction
- Correlated factors in ANOVA (comparing kids with autism to adult controls) → Bad design, sorry!



# Generating surrogate data

- Surrogate data may be generated at times to non-parametrically derive the null distributions of various statistics
- Coherence between 2 channels: Jumble up trials of 1 channel and compute coherence between 2 channels (tricky for event related design)
- 'Empty room' or 'Cap in electrolyte bath' data for MEG and EEG to derive null distributions
- Realistic simulations from the null such as white noise filtered and processed in the same way as the data



# Uncited References

- 1 B.J. Winer, D.R. Brown, K.M. Michels (1971): Statistical principles in experimental design. McGraw-Hill.
- 2 H Bokil, K Purpura, JM Schoffelen, D Thomson, P Mithra (2007): Comparing spectra and coherences for groups of unequal sizes. *J Neuro Methods*, 30(2): 337 - 345.
- 3 K.J. Friston, A.P. Holmes, K.J. Worsley, J.-P. Poline, C.D. Frith, R.S.J. Frackowiak (1995): Statistical Parametric Maps in Functional Imaging: A General Linear Approach. *Human Brain Mapping* 2:189-210.
- 4 Y. Benjamini and Y. Hochberg (1995): Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc Ser B*. 57: 289 - 300.
- 5 Genovese, C.R., Lazar, N.A. & Nichols, T.E. (2002): Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, 15:772-786.
- 6 Nichols T.E. & Holmes A.P. (2001): Nonparametric permutation tests for functional neuroimaging, a primer with examples. *Human Brain Mapping* 15:125.
- 7 Worsley, K.J., Marrett, S., Neelin, P., Vandal, A.C., Friston, K.J. & Evans, A.C. (1996a): A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, 4:58-73.
- 8 Worsley, K.J., Marrett, S., Neelin, P. & Evans, A.C. (1996b): Searching scale space for activation in PET images. *Human Brain Mapping*, 4:74-90.
- 9 Worsley, K.J., Andermann, M., Koulis, T., MacDonald, D. & Evans, A.C. (1999): Detecting changes in nonisotropic images. *Human Brain Mapping*, 8:98-101.

