Why 'N' How seminar series - Introduction to Frequentist Statistics

Hari M. Bharadwaj

Research Affiliate, MGH Neurology

December 3, 2009

Outline

1. Detection and Hypothesis testing
   - Basic problem
   - Some distributions
   - Adult human height - 'Alien' example
2. Important test statistics
   - One sample t-test
   - 2 sample *t*-test
   - Paired *t*-test
   - Multiple Comparisons
3. The General Linear Model (GLM) perspective
   - 2 sample *t*-test
   - One-way ANOVA
   - Correlation
   - ANCOVA
   - Other considerations and references

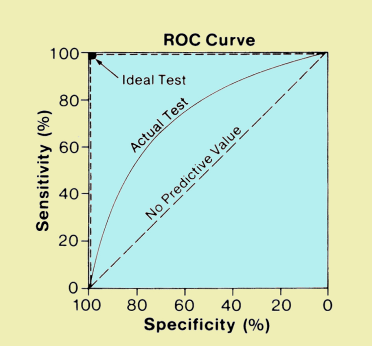
2 Alternatives discrimination problem

$\mathcal{H}_1$: **There is an 'effect'**
$\mathcal{H}_0$: **There is no 'effect'**

- Example: $\mathcal{H}_1$: Average IQ of ASD subjects $<$ TD subjects
  $\mathcal{H}_0$: Average IQ of ASD subjects $=$ TD subjects
- Given data we wish to probabilistically test out the hypotheses, example: given the IQ data of 10 ASD and 10 TD subjects
- Two approaches possible - Bayesian approach: More formal and comprehensive
  Frequentist approach - Easier and widely used but interpretation questionable

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 $p(data|\mathcal{H}_0)$?
- **Bayesian** - Given the data we have, what is the probability that $\mathcal{H}_0$ is true i.e $p(\mathcal{H}_0|data)$? 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

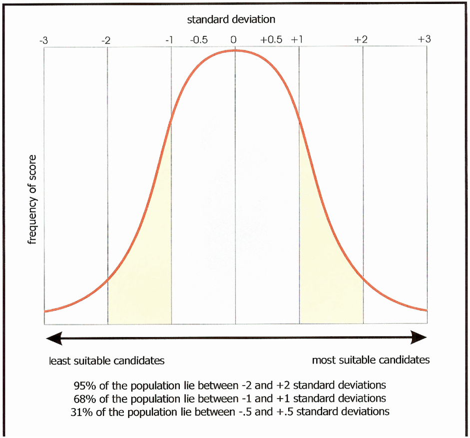


Figure: $p(\chi) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(\chi-\mu)^2}{2\sigma^2}}$, Normal distributions are good models of most real life data where clustering around the average happens, example: Adult human height

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 $\mu = 170$cm and $\sigma = 10$ 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?

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 $\mathcal{H}_1$)

χ^2 distribution

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

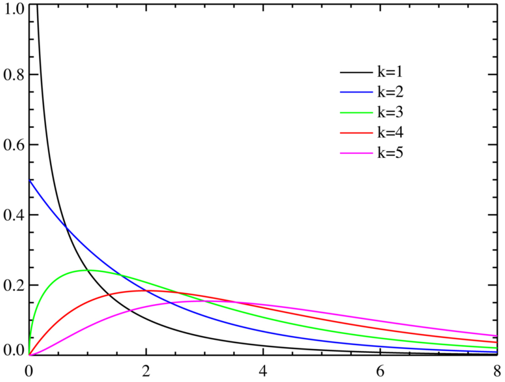


Figure: Sum of squares **Indpendent and Identically distributed** normal variables with mean 0 and variance 1

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 χ^2 variable with k degrees of freedom
- Ratio of a **zero mean** normal variable and square root of a χ^2 variable (with k df)is a **t** variable with k degrees of freedom

$$t = \frac{\chi}{\sqrt{S/k}} \tag{1}$$

- Ratio of two **independent** χ^2 variables is an **F** variable

$$F = \frac{S_1/k_1}{S_2/k_2} \tag{2}$$

F has degrees of freedom k_1 and k_2

t distribution

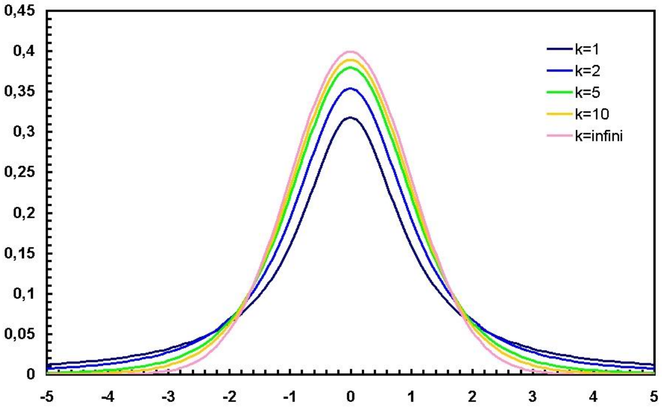


Figure: Ratio of zero mean normal and square root of a χ^2 distibution

F distribution

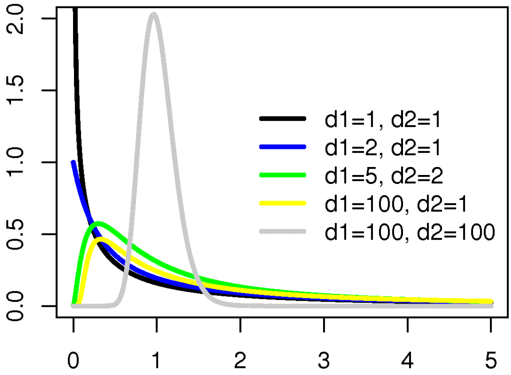


Figure: Ratio of 2 χ^2 distributions

Where we are...

One sample t-test

- Testing for the average of a normal **population** to have a certain mean μ_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

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_k}{k} \tag{3}$$

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

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

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

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

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

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

- $\mathcal{H}_1$ 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

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

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

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

- 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 ⇒ 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 guerentee 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 $\mathcal{H}_1$

Multiple Comparisons

- $\mathcal{H}_1$: **A** is an alien
  $\mathcal{H}_0$: **A** is a human being
- Given: Adult human height is normally distributed with $\mu = 170$cm and $\sigma = 10$ 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, **A** is an alien
- What if we have 100 subjects and you want to test if any of them is an alien?
- We cannot do a t-test for each subject being an alien, because under $\mathcal{H}_0$ about 20 (0.05 $\times$ 100) will be inferred as not humans

MCP - Bonferroni Correction

- Do a hypothesis test on the whole cohort instead of individual subjects
- $\mathcal{H}_1$: There are aliens in this group of 100
  $\mathcal{H}_0$:They are all humans
- Given: Adult human height is normally distributed with $\mu = 170$cm and $\sigma = 10$ cm
- A particular subject's height is **independent** of other subjects' heights
- Set our individual subject t-test threshold p value to be $\frac{0.05}{n}$, where n = 100 here
- This will ensure that just by chance (under $\mathcal{H}_0$), the probability that any 1 out of all the 100 will be misclassified as alien is still 0.05 $\longrightarrow$ Called bonferroni correction

Where we are...

2 group t-test as a linear regression

- $y_i = \mu + \beta x_{group} + \epsilon_i$
- Example: A given ASD IQ $= \mu_{population} + \mu_{asd} + \epsilon_{measurement}$
- Fitting IQ data to groups is like forming a linear regression between IQ (independent variable) and Group (The explanatory variable)
- Asking is IQs are different for the 2 groups is like asking if the slope of the regression between y (IQ) and x (group in this case) is not 0
- This generalises to multiple group tests (F-tests) or ANOVA
- This perspective also gives us ways to assess significance of covariance

GLM with 1 factor (group) → One-Way ANOVA with 2 levels

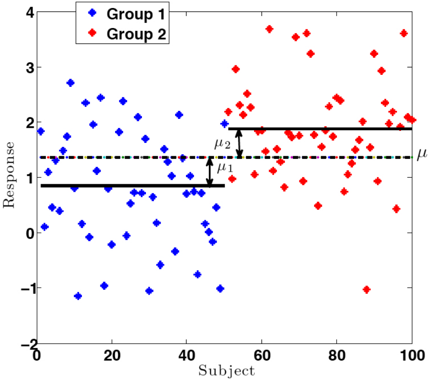


Figure: The data y is explained by 1 factor, namely 'group' $x = 0$ or 1 denoting ASD or TD for example. Does regressing y as a linear function of x help explain the variance in y better than when not modeled as a function of x?

F test instead of t test

- Is $y = \beta x + \mu + \epsilon$ a better model than $y = \mu + \epsilon$?
- Fit both models and see if gives one smaller errors ϵ than the other
- Let S_1 be the sum of squared errors after fitting model 1 and S_2 the sum of squared errors after fitting model 2 (null model): Is S_1 significantly smaller than S_2? (Use out **recipe**)
- Let $F = \frac{(S_2 - S_1)/(k_2 - k_1)}{S_1/k_1}$, under $\mathcal{H}_0$, both numerator and denominator are χ^2 distributed with $k_2 - k_1$ and k_1 degrees of freedom respectively
- $F = \frac{ExplainedVariance}{UnexplainedVariance} = \frac{VarianceOfGroupMeans}{MeanOfWithinGroupVariances}$
- Under $\mathcal{H}_0$, F is F-distributed with k_1, k_2 degrees of freedom
- The p-values should be the same as from a 2 sample t-test, but we wont know which group's mean is higher

One way 2 level ANOVA example

- 2 Group IQs: TDs = 87, 110, 93, 99, 75, 102, 90, 83, 100, 70
  ASDs = 77, 81, 64,100, 84, 72, 69, 90, 68, 70
- $t = 2.5138$, $df = 18$, $p = 0.0217 < 0.05$
- $F = 6.3192$, $df = 1, 18$, $p = 0.0217 < 0.05$
- Lets try this out in MATLAB
- Note that for a 2 group analysis, $F_{anova} = t^2$

One way ANOVA with 3 levels

- With GLMs, we need not restrict to 2 levels:
  TDs = 87, 110, 93, 99, 75, 102, 90, 83, 100, 70
  ASDs = 77, 81, 64,100, 84, 72, 69, 90, 68, 70
  Schizophrenia grp = 88, 87, 98, 120, 111, 97, 89, 132, 114, 126
- Same question asked: Is there any dependence on group (when #groups > 2)?
- Here IQ (y) is a modeled as a linear function of group (x) which can take 3 values
- $F(2, 17) = 11.05$, $p = 0.0003$
- Degrees of freedom is the rank of the linear subspace of $\mathcal{R}^n$ spanned by each explanatory variate: The number of independent variables in the set

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 (ρ): 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 ρ_0 is the actual population correlation

Linear and non-linear predictability

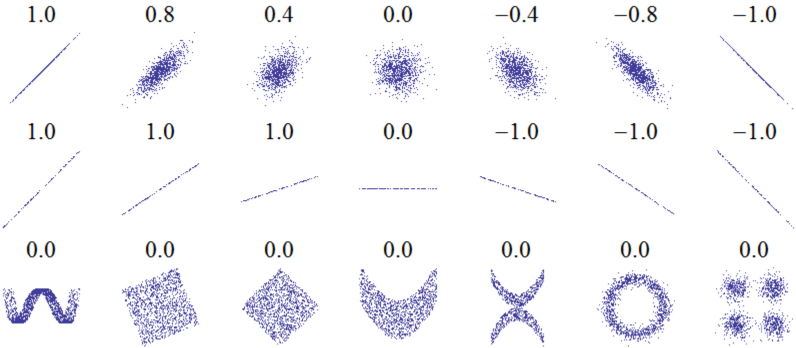


Figure: Correlation just says y is linearly predictable from x. Lower correlation $\Rightarrow$ Higher prediction error. Perfect dependence could result in zero correlation if the dependence is non-linear.

Outliers and bad models

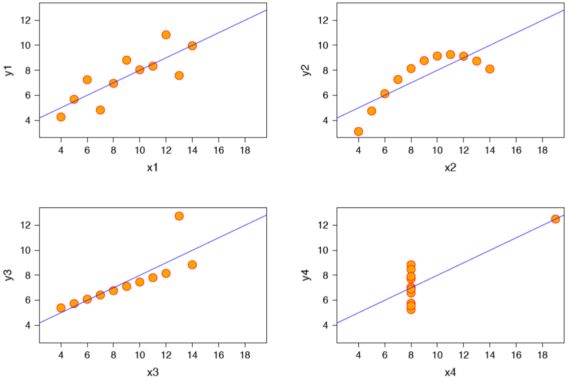


Figure: All the 4 cases have the exact same correlation coefficient of about 0.8. One should plot and look at the curves. A log-linear model might fit better.

Multiple way ANOVA with uninteresting covariate (ANCOVA)

- IQ(td) = 87, 110, 93, 99, 75, 102, 90, 83, 100, 70
  AGE(td) = 9,15,9,10,10,12,8,10,11,7
  IQ(asd) = 77, 81, 64,100, 84, 72, 69, 90, 68, 70
  AGE(asd) = 8,9,9,13,9,8,8,9,7,8
- Linear model for IQ (y) as a function of group(x_1) and age(x_2)
- Are the IQs different between groups given that they were not all 1 age?
- **Group and Age should be uncorrelated:** x_1 **and** x_2 **uncorrelated**
- Is $y = \beta_1 x_1 + \beta_2 x_2 + \mu + \epsilon$ better significantly than $y = \beta_2 x_2 + \mu + \epsilon$
- Again F-test: $F(1, 17) = 3.88$, $p = 0.0653 \Rightarrow$ Not significant at 0.05!

Summary of GLM/ANOVA

- ANOVA generalizes t-test to more than 2 groups

Summary of GLM/ANOVA

- ANOVA generalizes t-test to more than 2 groups
- Correlations, ANOVA, ANCOVA are all instances of the GLM

Summary of GLM/ANOVA

- ANOVA generalizes t-test to more than 2 groups
- Correlations, ANOVA, ANCOVA are all instances of the GLM
- Formulating $\mathcal{H}_0$ and $\mathcal{H}_1$ as competing linear models is the key $\rightarrow$ F-tests

Summary of GLM/ANOVA

- ANOVA generalizes t-test to more than 2 groups
- Correlations, ANOVA, ANCOVA are all instances of the GLM
- Formulating $\mathcal{H}_0$ and $\mathcal{H}_1$ as competing linear models is the key $\longrightarrow$ F-tests
- Assumptions in GLM:

Summary of GLM/ANOVA

- ANOVA generalizes t-test to more than 2 groups
- Correlations, ANOVA, ANCOVA are all instances of the GLM
- Formulating $\mathcal{H}_0$ and $\mathcal{H}_1$ as competing linear models is the key $\longrightarrow$ F-tests
- Assumptions in GLM:
  1. All the data are normally distributed

Summary of GLM/ANOVA

- ANOVA generalizes t-test to more than 2 groups
- Correlations, ANOVA, ANCOVA are all instances of the GLM
- Formulating $\mathcal{H}_0$ and $\mathcal{H}_1$ as competing linear models is the key $\rightarrow$ F-tests
- Assumptions in GLM:
  1. All the data are normally distributed
  2. The explained variable's conditional mean is a linear function of the explanatory variable

Summary of GLM/ANOVA

- ANOVA generalizes t-test to more than 2 groups
- Correlations, ANOVA, ANCOVA are all instances of the GLM
- Formulating $\mathcal{H}_0$ and $\mathcal{H}_1$ as competing linear models is the key $\rightarrow$ F-tests
- Assumptions in GLM:
  1. All the data are normally distributed
  2. The explained variable's conditional mean is a linear function of the explanatory variable
  3. Successive measurements are independent

Summary of GLM/ANOVA

- ANOVA generalizes t-test to more than 2 groups
- Correlations, ANOVA, ANCOVA are all instances of the GLM
- Formulating $\mathcal{H}_0$ and $\mathcal{H}_1$ as competing linear models is the key $\rightarrow$ F-tests
- Assumptions in GLM:
  1. All the data are normally distributed
  2. The explained variable's conditional mean is a linear function of the explanatory variable
  3. Successive measurements are independent
  4. There is one common error variance (example: IQ in TD and ASD have the same variance)

Summary of GLM/ANOVA

- ANOVA generalizes t-test to more than 2 groups
- Correlations, ANOVA, ANCOVA are all instances of the GLM
- Formulating $\mathcal{H}_0$ and $\mathcal{H}_1$ as competing linear models is the key $\rightarrow$ F-tests
- Assumptions in GLM:
  1. All the data are normally distributed
  2. The explained variable's conditional mean is a linear function of the explanatory variable
  3. Successive measurements are independent
  4. There is one common error variance (example: IQ in TD and ASD have the same variance)
  5. The explanatory variables should be uncorrelated in ANCOVA

Summary of GLM/ANOVA

- ANOVA generalizes t-test to more than 2 groups
- Correlations, ANOVA, ANCOVA are all instances of the GLM
- Formulating $\mathcal{H}_0$ and $\mathcal{H}_1$ as competing linear models is the key $\rightarrow$ F-tests
- Assumptions in GLM:
    1. All the data are normally distributed
    2. The explained variable's conditional mean is a linear function of the explanatory variable
    3. Successive measurements are independent
    4. There is one common error variance (example: IQ in TD and ASD have the same variance)
    5. The explanatory variables should be uncorrelated in ANCOVA
    6. One should look out for outliers and bad modeling: Always plot y versus x and look at it

Other considerations/ Future discussions

- Non-parametric tests: Bootstrapping/permutations

Other considerations/ Future discussions

- Non-parametric tests: Bootstrapping/permutations
- **Main effects** which **'interact'**: The level of one explanatory variable affects the effect of another explanatory variable → Example: Exercise increases sleep hours only in heavy people, i.e Exercise predicts sleep hours only when weight is above a certain value, exercise interacts with weight in predicting sleep hours

Other considerations/ Future discussions

- Non-parametric tests: Bootstrapping/permutations
- **Main effects** which **'interact'**: The level of one explanatory variable affects the effect of another explanatory variable → Example: Exercise increases sleep hours only in heavy people, i.e Exercise predicts sleep hours only when weight is above a certain value, exercise interacts with weight in predicting sleep hours
- Greenhouse-Geisser correction could be applied in non-spherical variance situations, but interpretation of results might be difficult

Other considerations/ Future discussions

- Non-parametric tests: Bootstrapping/permutations
- **Main effects** which **'interact'**: The level of one explanatory variable affects the effect of another explanatory variable → Example: Exercise increases sleep hours only in heavy people, i.e Exercise predicts sleep hours only when weight is above a certain value, exercise interacts with weight in predicting sleep hours
- Greenhouse-Geisser correction could be applied in non-spherical variance situations, but interpretation of results might be difficult
- Doing an ANCOVA to eliminate significant differences in the covariate is just wrong! ⇒ If ASDs and TDs differ significantly in AGE, one can **NOT** use an ANCOVA and ask the question as to whether the IQs would be same or not if the AGES were the same

Other considerations/ Future discussions

- Non-parametric tests: Bootstrapping/permutations
- **Main effects** which **'interact'**: The level of one explanatory variable affects the effect of another explanatory variable → Example: Exercise increases sleep hours only in heavy people, i.e Exercise predicts sleep hours only when weight is above a certain value, exercise interacts with weight in predicting sleep hours
- Greenhouse-Geisser correction could be applied in non-spherical variance situations, but interpretation of results might be difficult
- Doing an ANCOVA to eliminate significant differences in the covariate is just wrong! ⇒ If ASDs and TDs differ significantly in AGE, one can **NOT** use an ANCOVA and ask the question as to whether the IQs would be same or not if the AGES were the same
- Correcting for multiple comparisons when the tests are not independent: Permutations and Random Field theory

Other considerations/ Future discussions

- Non-parametric tests: Bootstrapping/permutations
- **Main effects** which **'interact'**: The level of one explanatory variable affects the effect of another explanatory variable → Example: Exercise increases sleep hours only in heavy people, i.e Exercise predicts sleep hours only when weight is above a certain value, exercise interacts with weight in predicting sleep hours
- Greenhouse-Geisser correction could be applied in non-spherical variance situations, but interpretation of results might be difficult
- Doing an ANCOVA to eliminate significant differences in the covariate is just wrong! ⇒ If ASDs and TDs differ significantly in AGE, one can **NOT** use an ANCOVA and ask the question as to whether the IQs would be same or not if the AGES were the same
- Correcting for multiple comparisons when the tests are not independent: Permutations and Random Field theory
- Block effects, efficient design and design matrices

References

1. BJ Winer, DR Brown, KM Michels (1971): Statistical principles in experimental design. McGraw-Hill.
2. 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.