Statistics in Genetics
and Genetic Association Studies

Jonathan Davis PhD
Instructor Department of Psychiatry
Jonathan.davis@ucdenver.edu
General outline

1) “statistical things I wish I had been taught 20 years ago” (what is a p-value anyway?).

2) Genetic association study design and review of confounding.

3) Some statistical methods commonly used.
Part 1) 5 statistical things I wished I had been taught 20 years ago

  Ewan Birney, European Molecular Biology Laboratory, European Bioinformatics Institute

1) Non-parametric statistics
2) R
3) Multiple testing
4) p-value, effect size, and sample size
5) simple linear models (generalized linear models)

Good course to take.
BIOS 6655; Statistical Methods in Genetic Association Studies
What’s a p-value anyway?

• A p-value represents the strength of evidence against the null hypothesis. We usually choose 0.05 as ‘important’, and if less than 0.05 there is strong evidence to reject the null.

• Also could be considered as a measure of the results as happening as “pure chance alone”. A p-value of 0.002 suggests that the results, if due to chance alone, would happen 2 times in a 1,000 studies.

• Or – the probability that randomness in sampling could have lead to the differences detected

• Proposed by Fisher for use in research in the 20’s.

• Dependent on sample size, effect size and a measure of spread.
How are they calculated

For example; some math for a Z test

\[ Z = \frac{(x - u)}{\left(\frac{\sigma}{\sqrt{n}}\right)} \]

\((x - u)\) = effect size.
\(\sigma\) = standard deviation.
\(n\) = sample size.
\(Z\) = a value, that if more than a critical level (+ - 1.96) \(p < 0.05\).

A good question to ask is “Is the effect size clinically relevant”
Multiple Comparisons

• P-values assume 1 independent test
• If we make many tests p-values are not accurately giving us a measure of the strength of evidence against the null
• For example;
  – if we conduct 100 independent tests we would expect 5 to be positive due to chance alone at an alpha of 0.05
How do we address this?

• Correct via Bonferroni or False Discovery Rate (FDR) or others.

• Or don’t do anything, and let readers judge.

• Bonferroni simply divides a critical alpha by the number of tests. 0.05/number of tests.
  – 100 tests new critical value would be 0.05/100 = 0.0005

• FDR ranks p-values from tests and adjusts for an acceptable amount of false discoveries.
A little more on FDR

• FDR uses an ‘acceptable’ proportion of false positives

• A p-value of 0.05 implies that 5% of all tests will result in false positives. An FDR adjusted p-value (or q-value) of 0.05 implies that 5% of significant tests will result in false positives.
• Assuming a 0.05 false discovery rate as acceptable

• Some R code below

```r
p_values <- c(0.00078, 0.00080, 0.00090, 0.00700, 0.02500, 0.04000, 0.04800, 0.0490, 0.05000)
test <- c("a", "b", "c", "d", "e", "f", "g", "h", "i")
#enter the target FDR
alpha.star <- .05
real.p <- as.data.frame(cbind(test, p_values))
threshold <- alpha.star * (1:length(p_values)) / length(p_values)
cbind(real.p, threshold)
```

Taken with minor tweaks from
• http://www.unc.edu/courses/2007spring/biol/145/001/docs/lectures/Nov12.html
<table>
<thead>
<tr>
<th>test</th>
<th>P-value</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.00078</td>
<td>0.006</td>
</tr>
<tr>
<td>B</td>
<td>0.00080</td>
<td>0.011</td>
</tr>
<tr>
<td>C</td>
<td>0.00090</td>
<td>0.017</td>
</tr>
<tr>
<td>D</td>
<td>0.007</td>
<td>0.022</td>
</tr>
<tr>
<td>E</td>
<td>0.025</td>
<td>0.028</td>
</tr>
<tr>
<td>F</td>
<td>0.04</td>
<td>0.033</td>
</tr>
<tr>
<td>G</td>
<td>0.048</td>
<td>0.039</td>
</tr>
<tr>
<td>H</td>
<td>0.049</td>
<td>0.044</td>
</tr>
<tr>
<td>I</td>
<td>0.04999</td>
<td>0.05</td>
</tr>
</tbody>
</table>

9 tests
Bonferoni = 0.05/9 = 0.006 as critical value. (A, B, C) are significant
FDR of 0.05 (A, B, C, D, E) are significant

What if we did 91 more tests none of which make p <0.05?
Bonferoni = 0.05/100 = 0.0005. None are significant.
With an FDR of 0.05 which are significant?
What about corrections in genome wide association studies?

- Making as many as 1,000,000 comparisons
- Is $0.05 \times 10^{-8}$ appropriate (bonferroni)
- Bonferroni assumes independent tests
- Staged/joint analyses, permutation tests, correcting for LD blocks, effective number of independent tests...others... active area of research.
  - In staged analyses Stage 1 uses a proportion of subjects with a large number of markers, Stage 2 a proportion of markers on remaining subjects.
  - In permutation effective correlations among SNPs is preserved but the phenotype is randomly shuffled 10k times to simulate the null. (computationally prohibitive in large studies).
  - LD blocks and effective number of independent tests, essentially more or less independent SNP clusters.
Reproducibility

• How come studies that meet stringent corrections for multiple comparisons are not reproduced?
• Recall a p-value is only referring to random sampling chance alone.
• Experiment and analysis errors and biases are not included.
• Are these populations true random representative and generalizable samples?
  – Especially with complex heterogeneous conditions
  – Autism? Schizophrenia?
• Take a look at table 1... a joint GWAS analysis of coronary artery disease in two different populations
Table 1 CAD case characteristics, two study populations WTCCC and German

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>WTCCC Study (N = 1926)</th>
<th>German Study (N = 875)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex — no. (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1527 (79.3)</td>
<td>591 (67.5)</td>
</tr>
<tr>
<td>Female</td>
<td>399 (20.7)</td>
<td>284 (32.5)</td>
</tr>
<tr>
<td><strong>Mean age at first event—yr</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td>49.4±7.7</td>
<td>49.0±7.4</td>
</tr>
<tr>
<td>Women</td>
<td>51.0±7.4</td>
<td>52.8±8.2</td>
</tr>
<tr>
<td><strong>Mean age at recruitment—yr</strong></td>
<td>60.1±8.1</td>
<td>58.1±8.7</td>
</tr>
<tr>
<td><strong>First event—no. (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MI</td>
<td>1377 (71.5)</td>
<td>875 (100.0)</td>
</tr>
<tr>
<td>PCI</td>
<td>202 (10.5)</td>
<td></td>
</tr>
<tr>
<td>CABG</td>
<td>347 (18.0)</td>
<td></td>
</tr>
<tr>
<td><strong>Family history of CAD—no. (%)†</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Father affected</td>
<td>570 (29.6)</td>
<td>151 (17.3)</td>
</tr>
<tr>
<td>Mother affected</td>
<td>302 (15.7)</td>
<td>114 (13.0)</td>
</tr>
<tr>
<td>Both parents affected</td>
<td>186 (9.7)</td>
<td>49 (5.6)</td>
</tr>
<tr>
<td>Sibling affected</td>
<td>1464 (76.0)</td>
<td>789 (90.2)</td>
</tr>
<tr>
<td><strong>Cardiovascular risk factor—no. (%)‡</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Former or current smoker</td>
<td>1460 (75.8)</td>
<td>615 (70.3)</td>
</tr>
<tr>
<td>Diabetes mellitus</td>
<td>211 (11.0)</td>
<td>108 (12.3)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>822 (42.7)</td>
<td>757 (86.5)</td>
</tr>
<tr>
<td>Hyperlipidemia</td>
<td>1542 (80.0)</td>
<td>666 (76.1)</td>
</tr>
<tr>
<td>Obesity (body-mass index &gt;30)</td>
<td>448 (23.3)</td>
<td>197 (22.5)</td>
</tr>
<tr>
<td><strong>Receipt of lipid-lowering drugs—no. (%)</strong></td>
<td>1429 (74.2)</td>
<td>575 (65.7)</td>
</tr>
</tbody>
</table>

From Samani et al, NEJM 2007 Aug 2;357(5):443-53
**Table 2**

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Lead SNP</th>
<th>Minor Allele in Controls</th>
<th>Risk Allele</th>
<th>Data</th>
<th>Frequency of Minor Allele</th>
<th>Odds Ratio for Risk Allele (95% CI)</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>rs2943634</td>
<td>A</td>
<td>C</td>
<td>WTCCC</td>
<td>0.30</td>
<td>0.34</td>
<td>1.22 (1.11-1.33)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>German</td>
<td>0.32</td>
<td>0.37</td>
<td>1.20 (1.06-1.35)</td>
</tr>
</tbody>
</table>

Samani et al, NEJM 2007 Aug 2;357(5):443-53
Part 2) Association Study Designs and Confounding

• Correlation -> causation?
  – Never definitively prove causation but we can design our studies to suggest it.
    – Bench Experiment
    – Randomized Clinical Trial
    – Prospective Cohort
    – Case Control
    – Cross-sectional
    – Ecologic
Common study designs

• Prospective Cohort
  – Start with exposure and follow individuals to see development of disease.

• Case Control
  – Start with disease status and compare exposures between unaffected and affected.
  – The important difference is time. In a cohort study we know the exposure preceded disease. In a case control we do not know the order of events.
  – In genetics a common and fair assumption is that the exposure does in fact precede disease.
  – Limited information for public health implications because in case control studies cases are usually oversampled
Cases and Controls

• Selection is important
  – Population based?
  – Super controls?
  – Large phenotypic differences?
  – Complex case phenotypes?
  – Matching?
  – How does selection affect power?
  – What about generalizability?
Common strategies for association studies

- **Candidate gene**
  - Known or suspected functional component

- **Linkage Candidate region**
  - Candidate genes within a region
  - Use a number of markers

- **Genome wide**
Common measure of association (and effect size) in case control studies

• Odds Ratio

<table>
<thead>
<tr>
<th></th>
<th>Disease +</th>
<th>Disease -</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure +</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Exposure -</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td></td>
<td>a+c</td>
<td>b+d</td>
</tr>
</tbody>
</table>

“probability odds ratio” = a/(a+b)/1-(a/a+b)c/c+d/1-(c/c+d)... and some algebra =

OR = a/b/c/d
## Odds Ratio example (MI and hypertension)

<table>
<thead>
<tr>
<th></th>
<th>MI +</th>
<th>MI -</th>
<th>Probability</th>
<th>Probability Odds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyp +</td>
<td>180</td>
<td>9820</td>
<td>180/10000</td>
<td>180/(10000-180) = 180/9820 = 0.018</td>
</tr>
<tr>
<td>Hyp -</td>
<td>30</td>
<td>9970</td>
<td>30/10000</td>
<td>30/(10000 – 30) = 30/9970 = 0.0031</td>
</tr>
</tbody>
</table>

“Probability odds ratio” or just odds ratio = 0.018/0.0031 = OR = 6.09

Szklo Neito  Epidemiology Beyond the Basics ch3

Which is then followed by a $X^2$ with 1fd test

Where $X^2 = \sum \frac{(observed - expected)^2}{expected}$
• OR and $X^2$ methods very similar in genetics
• Exposure, however, has a few more complexities

- Consider an autosomal marker with 2 alleles, 1 and 2

<table>
<thead>
<tr>
<th>Genotype</th>
<th>1,1</th>
<th>1,2</th>
<th>2,2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affected</td>
<td>30</td>
<td>40</td>
<td>50</td>
</tr>
<tr>
<td>Unaffected</td>
<td>30</td>
<td>60</td>
<td>30</td>
</tr>
</tbody>
</table>

$X^2 = \sum_{l=1}^{6} \frac{(O_l - E_l)^2}{E_l}$, $l=1...6$ 2df

How would this table look if we believe allele 1 is dominant? What about recessive, or partial penetrance?
Constructing Exposure, things to consider

• Haplotypes and Linkage Disequilibrium, markers of causal locus
  – Haplotype; a collection of markers along a chromosome.
  – Linkage Disequilibrium; (LD), the deviation of independence of alleles at two genetic loci in a natural breeding population... eg correlation

• Penetrance of disease (dominant, recessive, additive)

• Gene dosage effects with highly polymorphic copy number variants (eg DUF1220)
Confounding review

Genetic exposure of interest → Disease (CAD) → Confounder, Population admixture, Smoking status, Others?
confounding adjustment example with stratification

<table>
<thead>
<tr>
<th></th>
<th>CAD+</th>
<th>CAD-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure +</td>
<td>50</td>
<td>501</td>
</tr>
<tr>
<td>Exposure -</td>
<td>65</td>
<td>384</td>
</tr>
</tbody>
</table>

Exposure+ OR = 0.59

OR less than 1 exposure is ‘protective’
confounding adjustment example with stratification

Smokers

<table>
<thead>
<tr>
<th></th>
<th>CAD+</th>
<th>CAD-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure +</td>
<td>11</td>
<td>40</td>
</tr>
<tr>
<td>Exposure -</td>
<td>49</td>
<td>200</td>
</tr>
</tbody>
</table>

OR = 1.12

Non-Smokers

<table>
<thead>
<tr>
<th></th>
<th>CAD+</th>
<th>CAD-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure +</td>
<td>39</td>
<td>461</td>
</tr>
<tr>
<td>Exposure -</td>
<td>16</td>
<td>184</td>
</tr>
</tbody>
</table>

OR = 0.97
Confounding cont....

• Before adjustment exposure was protective OR=0.59
• After adjustment via stratification no good evidence (ORs 1.1, 0.97)
• What gives?
  – Exposure group had fewer smokers (9.3% vs 55.5%)
Table 1 again

• Does the control group have smoking status?
• Is this SNP related to smoking somehow?
• Was smoking status a covariate?
• Can we conclude this SNP is a risk of CAD independent of smoking?

<table>
<thead>
<tr>
<th>Cardiovascular risk factor — no. (%)</th>
<th>1460 (75.8)</th>
<th>615 (70.3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Former or current smoker</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes mellitus</td>
<td>211 (11.0)</td>
<td>108 (12.3)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>822 (42.7)</td>
<td>757 (86.5)</td>
</tr>
<tr>
<td>Hyperlipidemia</td>
<td>1542 (80.0)</td>
<td>666 (76.1)</td>
</tr>
<tr>
<td>Obesity (body-mass index &gt;30)</td>
<td>448 (23.3)</td>
<td>197 (22.5)</td>
</tr>
<tr>
<td>Receipt of lipid-lowering drugs — no. (%)</td>
<td>1429 (74.2)</td>
<td>575 (65.7)</td>
</tr>
</tbody>
</table>
Population Stratification adjustment

• A popular method uses Principal Components (PCA)
• Dimension reduction technique to find linear combinations of the predictors
• First component is the combination of variables with the largest variance, second is uncorrelated with the first but has the next two variables with the largest variance and so forth.
• Generate an ethnicity covariate to utilize in a multivariate regression analysis.
Part3) Some statistical methods commonly used

- ttest
- ANOVA
- Linear regression
- Logistic regression
- Poison regression
ANOVA

• Extension of a ttest where there are more than 2 groups
• Get an F statistic that signifies one of the groups is different (doesn’t say which). Follow up with ttests
• Assumptions
  – Homoscedasticity (constant variance)
  – Independent observations
  – Distribution of residuals is normal (normality)
ANOVA and t-tests

Mean Plot
with 95% CI

continuous outcome

n=14  n=7  n=11
aa  aA  AA
Some R code and output

# genotype is a 3 level categorical variable “AA, aA, aa”

lec.out <- lm(outcome ~ genotype, data = data_lec)
summary(lec.out)

Coefficients:

|            | Estimate | Std. Error | t value | Pr(>|t|)  |
|------------|----------|------------|---------|-----------|
| (Intercept)| 26.664   | 1.017      | 26.221  | < 2e-16 ***|
| aA         | -3.464   | 1.631      | -2.124  | 0.0423 *  |
| aa         | -11.564  | 1.359      | -8.510  | 2.24e-09 ***|

Residual standard error: 3.373 on 29 degrees of freedom
Multiple R-squared: 0.7258, Adjusted R-squared: 0.7069
F-statistic: 38.39 on 2 and 29 DF, p-value: 7.097e-09
Linear Regression

- Comparing two continuous variables
- Very similar to ANOVA
- \( Y = \beta_0 + \beta_{\text{con1}} + \varepsilon \)
- Assumptions (among others)
  - linearity
  - Homoscedasticity (constant variance)
  - Independent observations
  - Distribution of residuals is normal (normality)
Linear regression plots
DUF1220 CON 1 and head circumference
Linear regression R code and output

```r
> bay.con1 <- lm(focz ~ con1_clade, data= bay)
> summary(bay.con1)

Call: lm(formula = focz ~ con1_clade, data = bay)
Residuals:    Min     1Q   Median     3Q    Max
-2.8211 -1.0444 -0.2028  1.0698  3.8995

Coefficients:    Estimate    Std. Error    t value    Pr(>|t|)
(Intercept)  -0.5673       0.3164      -1.793       0.0837 .
con1_clade   8.4902       1.4441       5.879    2.53e-06 ***

--- Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.685 on 28 degrees of freedom
(7 observations deleted due to missingness) Multiple R-squared: 0.5525, Adjusted R-squared: 0.5365 F-statistic: 34.57 on 1 and 28 DF, p-value: 2.533e-06
```

“For each unit increase in CON1 copy ratio head size increases 8.5 z-scores.” or consider scaling
“For each .1 unit increase in CON1 copy ratio head size increases .85 z-scores”
Generalized linear models

• Linear regression
  – Outcome can fall between +∞

• Logistic
  – Outcome is constrained to be a 0 or a 1

• Poisson (quasi-poisson, and negative binomial)
  – Outcome from 0 to +∞
  – Outcome is typically a rare count, but the distribution is very right tailed
Logistic function (Wikipedia)
Logistic regression

• Linear regression has the ‘identity’ link function, logistic transforms this to bound the outcome to be 1 or 0 through the ‘logit’ link function

• \( p = \) probability of being a case, \( 1-p = \) probability of not being a case

• \( \ln(p/(1-p)) = \beta_0 + \beta_{\text{con}1} + ... \)

• Odds = \( e^{\beta_0 + \beta_{\text{con}1}} \)

• Get Odds Ratios out – simple exponentiated beta value
Some R code and output

```r
bay.con1_logit <- glm(big ~ con1_clade, family = 'binomial', data = bay)
summary(bay.con1_logit)

Call: glm(formula = big ~ con1_clade, family = "binomial", data = bay)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
   -2.0043  -0.4393  -0.2702   0.4618   1.5933

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.34000    0.64350  -2.082  0.03732 *
con1_clade  10.01400    3.67043   2.728  0.00637 **
```

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 . ’ 1

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 34.795 on 29 degrees of freedom Residual deviance: 20.871 on 28
degrees of freedom (7 observations deleted due to missingness) AIC: 24.871 Number of Fisher Scoring iterations: 6

Intercept not so important

OR for CON1 \( \exp(10.01) = 22,247 \)

Yikes?! Why is the OR so big? Try scaling, this is for a unit increase in copy ratio, for a tenth unit increase
OR = \( \exp(1.001) = 2.7 \)

“For each .10 copy ratio increase of CON1 the odds of a big head increase 2.7
Poisson

• Log link function.
• Outcome is right tailed.
• Exponentiated betas give the percent change.
• Critical assumptions.
  – Not dispersed
    • Mean = variance
    • If over dispersed p-values and SEs are substantially directionally biased to be better than truth.
    • Use negative binomial instead
Good Resources

- http://www.ats.ucla.edu/stat/dae/
- http://www.statmethods.net/
- http://cran.r-project.org/