Statistical Considerations in the Study of Obesity

John A. Dawson

Post-Doc, Office of Energetics / SSG

April 11, 2014

Overview of Topics

- Basic terminology and probability
- Randomization
- Bias, confounding, blinding and blocking
- Simple tests for some very general situations
- Am I Normal? What if I’m not?
- Fixed vs. random effects
- General Considerations as a Statistical Reviewer
- Analysis of pre-post experiments
- Analysis of cross-over experiments
- Analysis of time-to-event data

Basic Terminology and Probability

"Somehow it seems to fill my head with ideas–
only I don't exactly know what they are!"

- Alice, *Through the Looking Glass*

Some Basic Terminology

- Mean
- Median
- Mode
- Quantile
- Variance
- Standard deviation
- *Appending ‘sample’ to the front of any of the above*
- Standard error (of the mean)
- Coefficient of variation
- Precision
- Scatterplot
- Histogram
- Box (and Whisker) plot
- Normal (Gaussian)
- Standard Normal

More Basic Terminology

- Null hypothesis (H_0)
- Alternative hypothesis (H_A or H_1)
- p-value
- Type I error
- Type II error
- Type III error
- α (in this context)
- β (in this context)
- Power
- Loss function
- Independent variable
- Dependent variable
- ‘Almost surely’ (a.s.) or ‘with probability 1’ (w.p. 1)
- IID

Even More Basic Terminology

- Categorical variable
- Ordinal variable
- Discrete variable
- Continuous variable
- Randomization
- Observational trial
- Randomized Controlled Trial (RCT)
- Blinded RCT
- Double-blinded RCT
- Triple-blinded RCT
- Multiple comparisons adjustment
- Parametric test
- Nonparametric test
- Permutation test

Some Simple Probability

- Pr(not A) = 1 – Pr(A)
- Pr(A and B) = Pr(A given B)*Pr(B)
- Odds (for) event A = Pr(A) / (1 – Pr(A))
- Odds (against) event A = (1 – Pr(A)) / Pr(A)

- How many ways are there …
  - … to order four items?
  - … to pick three numbered balls from a bag of six …
    - … if order doesn’t matter?
    - … if order does matter?

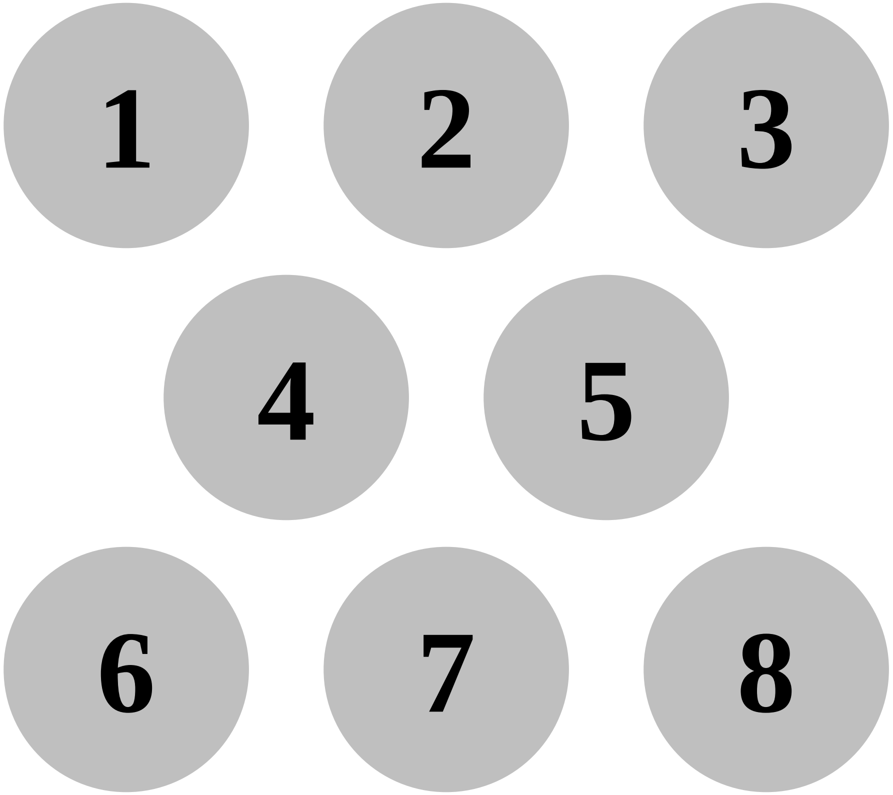
- What is Pr(X), where X is …
  - … rolling doubles with two dice?
  - … getting a flush of diamonds from a five-card draw?
  - … getting any flush from a five-card draw?

Randomization

"To consult a statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination."

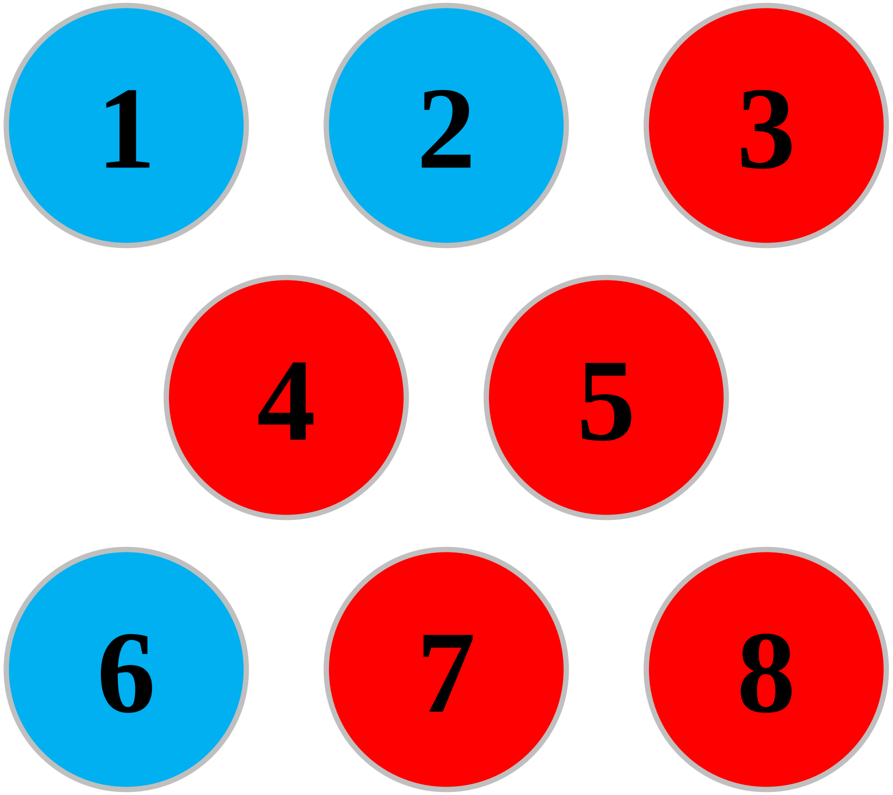
- R. A. Fisher, *Presidential Address to the First Indian Statistical Congress in 1938*

- 8 subjects to be randomized to two treatments **A** and **B**
- One outcome will be measured once for each subject
- Q: Is the average outcome the same for **A** and **B**?



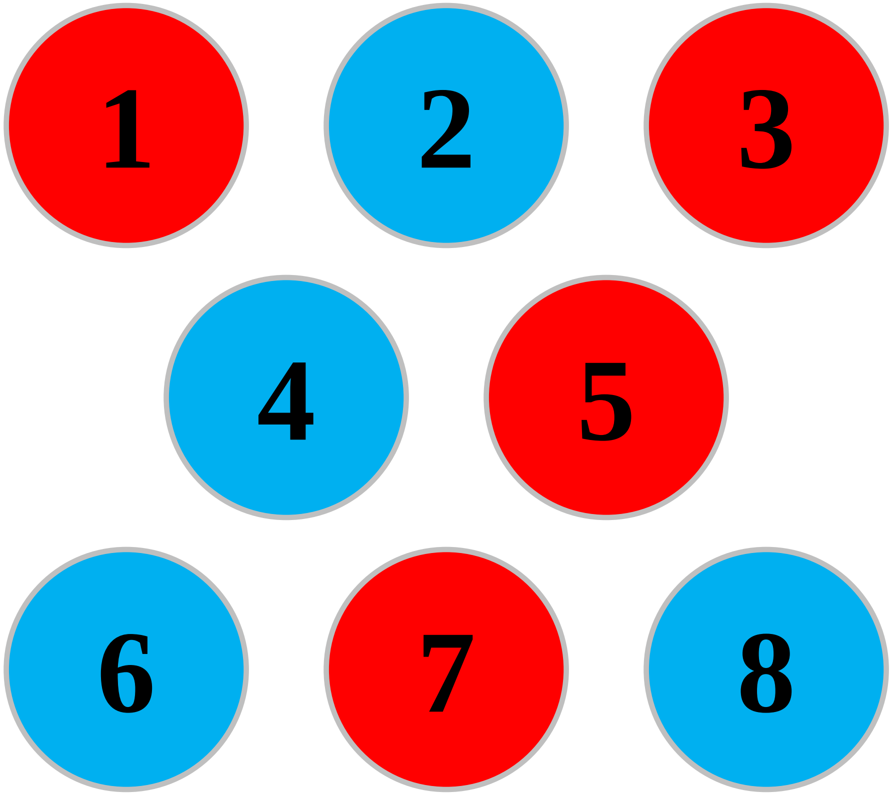
Simple Randomization

- Each subject independently has a fixed chance to be **A**
- Strengths: Simple, independence, largest space, tests
- Weaknesses: Bad (unbalanced) things can happen



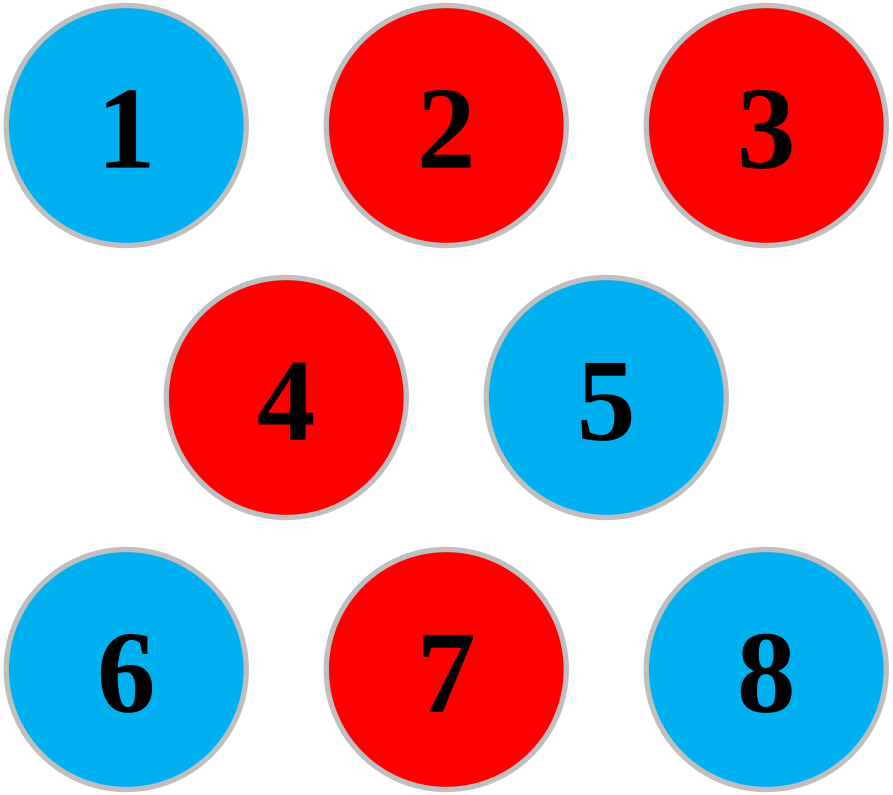
ABABABAB or BABABABA

- Assign to one treatment, then the other, then the first …
- Strengths: Very simple, equal allocation to A and B
- Weaknesses: Completely dependent assignments, space=2



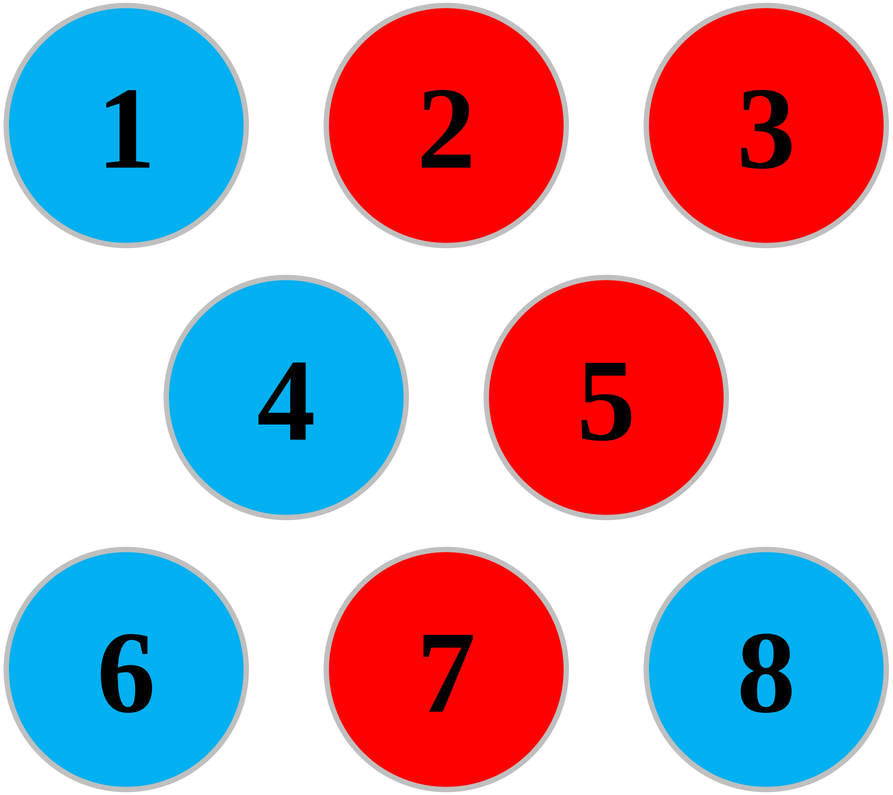
[AAAABBBB]

- Randomly choose four to be assigned to **A** (Urn scheme)
- Strengths: Largest space with equal allocation to **A** and **B**
- Weaknesses: Weak overall dependency



[AB][AB][AB][AB]

- Assign to one treatment, then the other; repeat
- Strengths: Equal allocation to **A** and **B**, balanced over time
- Weaknesses: Local dependency, smallest p-value?



Practicality, Robustness and Wishful Thinking

- (Almost) all of these problems lessen in severity as sample size increases, sometimes vanishingly
- Throwing more sample at a confounded design will not fix it
- ‘Haphazard’ is not the same thing as ‘random’
  - Last name, area code, zip code, day of week, whim

- How robust is the statistical test being employed?
- How will violations of assumptions manifest themselves?

- Hope is not an experimental design feature, it is a bug
- An underpowered study is not worth doing
  - Exception: Gathering pilot data

Bias, Confounding, Blinding and Blocking

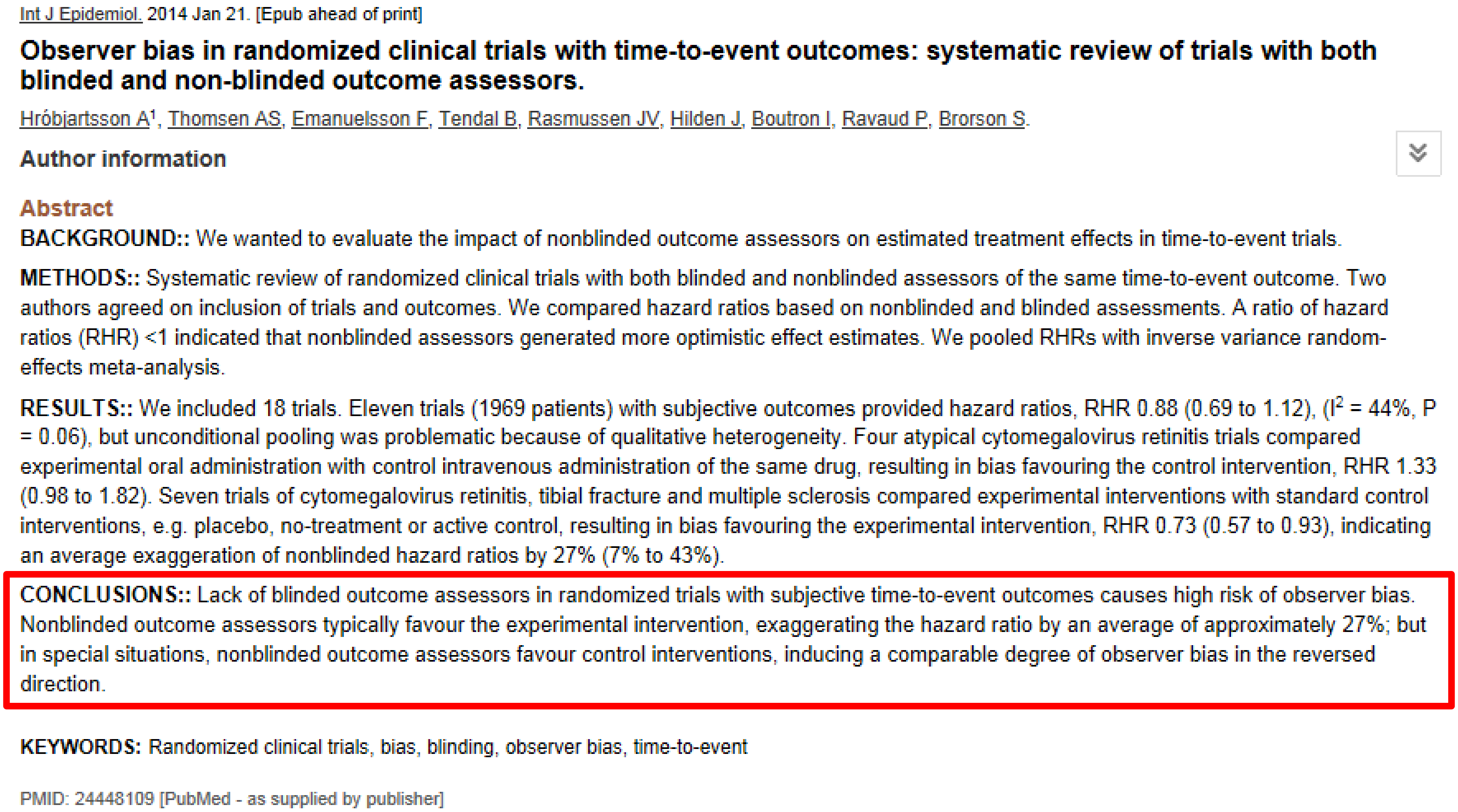
When you wish upon a star;
makes no difference who you are;
when you wish upon a star
it biases your actions in favor of the desired outcome.
- @HardSciFiMovies, *Twitter*

Bias

- Bias has a strictly statistical meaning related to expectations of estimators, but that’s not what we’re focused on here
- Bias: The prior belief for or against *something*
  - The efficacy of a treatment compared to SoC
  - The association between two variables
  - That eating breakfast promotes weight loss
  - That sugar is toxic (No! No! *Added* sugar!)

- Experimental designs ideally should be free of bias from:
  - Investigators
  - Participants / Subjects
  - Providers of funding
- Solution: Randomize in a *blinded* manner

Bias in Action



Int J Epidemiol. 2014 Jan 21. [Epub ahead of print]

Observer bias in randomized clinical trials with time-to-event outcomes: systematic review of trials with both blinded and non-blinded outcome assessors.

Hróbjartsson A[1], Thomsen AS, Emanuelsson F, Tendal B, Rasmussen JV, Hilden J, Boutron I, Ravaud P, Brorson S.

Author information

Abstract

BACKGROUND:: We wanted to evaluate the impact of nonblinded outcome assessors on estimated treatment effects in time-to-event trials.

METHODS:: Systematic review of randomized clinical trials with both blinded and nonblinded assessors of the same time-to-event outcome. Two authors agreed on inclusion of trials and outcomes. We compared hazard ratios based on nonblinded and blinded assessments. A ratio of hazard ratios (RHR) <1 indicated that nonblinded assessors generated more optimistic effect estimates. We pooled RHRs with inverse variance random-effects meta-analysis.

RESULTS:: We included 18 trials. Eleven trials (1969 patients) with subjective outcomes provided hazard ratios, RHR 0.88 (0.69 to 1.12), (I^2 = 44%, P = 0.06), but unconditional pooling was problematic because of qualitative heterogeneity. Four atypical cytomegalovirus retinitis trials compared experimental oral administration with control intravenous administration of the same drug, resulting in bias favouring the control intervention, RHR 1.33 (0.98 to 1.82). Seven trials of cytomegalovirus retinitis, tibial fracture and multiple sclerosis compared experimental interventions with standard control interventions, e.g. placebo, no-treatment or active control, resulting in bias favouring the experimental intervention, RHR 0.73 (0.57 to 0.93), indicating an average exaggeration of nonblinded hazard ratios by 27% (7% to 43%).

CONCLUSIONS:: Lack of blinded outcome assessors in randomized trials with subjective time-to-event outcomes causes high risk of observer bias. Nonblinded outcome assessors typically favour the experimental intervention, exaggerating the hazard ratio by an average of approximately 27%; but in special situations, nonblinded outcome assessors favour control interventions, inducing a comparable degree of observer bias in the reversed direction.

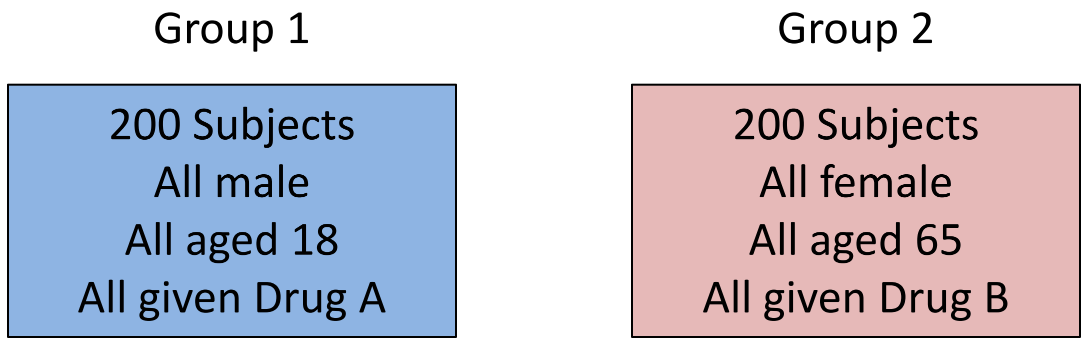
KEYWORDS: Randomized clinical trials, bias, blinding, observer bias, time-to-event

PMID: 24448109 [PubMed - as supplied by publisher]

http://www.ncbi.nlm.nih.gov/pubmed/24448109

Confounding

- When the effects of two or more covariates cannot be teased apart because of the experimental design, we say they are *confounded*
- Quite very bad example:



Group 1

200 Subjects
All male
All aged 18
All given Drug A

Group 2

200 Subjects
All female
All aged 65
All given Drug B

- Confounding is generally avoided through randomization
- However, sometimes the groups *really* need to be balanced with respect to one or more covariates
- Solution: Blocking (not to be confused with stratification)

Blinding and Blocking

- Blind: Being unaware of the treatment assignment
- Many levels of blindness as we’ve already touched on

- Blocking: Enforcing balance through local dependencies in treatment assignment
  - We’ve seen a few examples of this already
  - Blockings can correspond to factor levels (day of week)
  - Blocks need not correspond to a variable at all
- Blocking adds complexity and reduces degrees of freedom
- Important: Marginal assignment probabilities are equal

- Sometimes blinding is not completely practical
- Blocking can provide opportunities for unblinding

Practical Blinding

- Patients: Blind to assignment when possible
  - Similarly-designed placebos
  - Sham surgery or brain implants
  - Deception before debriefing
- Investigators: Blind to assignment when possible
  - At minimum, blind up until revelation of assignment
  - Investigators control ‘who comes in the door when’
- Funders: At minimum no role in design implementation

- Options for blinded randomization:
  - Gather/screen all subjects at once, then randomize
  - Substitute colors or noninformative labels
  - Physically implement the randomization (3rd party)

Practical Blocking

- Why block?
  - Control for day effects
  - Control for technician effects
  - Enforce balance over one or many covariates
  - Account for waves or seasonality
- The last assignment in a block can be determined a.s.
  - So?
    - If single-blinded, clinician knows what’s coming
    - If blocking structure known, patients can game system – less of an issue in obesity studies

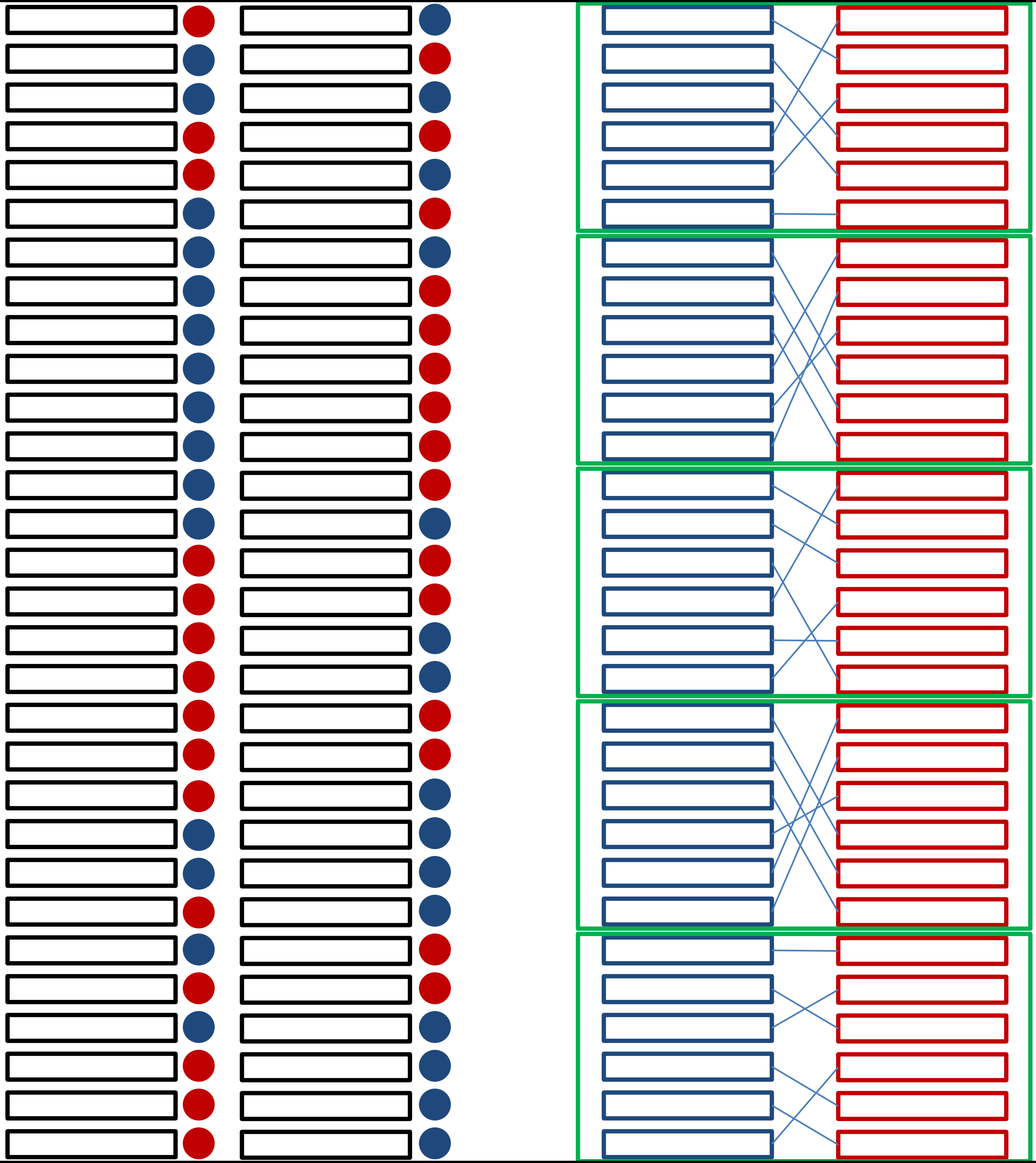
- One solution is to use randomly determined block sizes
- Having others implement a complex randomization protocol can be fraught with peril

Sequentially Numbered Opaque Sealed Envelope (SNOSE)

- The horrors of clever acronyms

- Sequentially Numbered: By external block or overall
- Opaque: So you can’t see inside
- Sealed: So that you can’t look inside beforehand
- Envelope: Something physical that can be opened with the subject and can be set up ahead of time

- The numbering / labeling on the outside helps to maintain multiple blocks as well as record keeping



Step 1:
List the 60 sorted subject IDs in the black boxes, from the top down in the left and then the right columns, sorting by longest recorded duration of limbo dancing (longer to shorter).

Step 2:
Copy subject IDs with a blue dot next to them into the blue boxes, in order, top down, from the left and then the right columns. Repeat for the IDs with a red dot.

Step 3:
Matched pairs are given by lines connecting boxes. Feed the red pair-mate 60% of the kale consumed *ad libitum* by the blue pair-mate subject.

Setup Complete!

Some Simple Tests for Certain General Situations

"Essentially, all models are wrong, some are useful."

- G.E.P. Box, *Empirical Model Building and Response Surfaces*

Building a 95% CI

- Sample estimate θ* of some quantity θ
- Standard error of the sample estimate (SE, f(x) of SS)

- θ* ± 1.96 × SE

- Why 1.96? Upper 2.5 percentile of a std. Normal
- Other upper 2.5 percentiles can sometimes apply instead

- Generally robust against distributional violations
- Does assume an unbounded domain (real line)
  - When bounded, transform to get unbounded
  - Do 95% CI on that scale and transform back
- Non-parametric equivalent: Bootstrap

One sample t test

- Normally distributed with mean μ
- Sample size n and variance σ^2
- H_0: $\mu = 0$ or some fixed δ

- $(\bar{X}-\delta)/s \sim t(n-1)$

- Assumes independence of the sample

- Generally robust against distributional violations
- This is equivalent to building a 95% CI via the t and seeing whether or not 0 falls in the CI
- Non-parametric equivalent: Bootstrap for mean, one sample Wilcoxon signed rank test
- Extension with covariates: OLS, focused on intercept

Pop Quiz!

- Two groups, Normally distributed with means μ_1 and μ_2
- Sample sizes n_1 and n_2; common variance σ^2
- One sample t-tests say:
    - μ_1 is significantly different from 0 ($p < 0.05$)
    - μ_2 is not significantly different from 0 ($p > 0.05$)

- Can we conclude $\mu_1 \neq \mu_2$?

Two sample t test

- Two groups, Normally distributed with means μ_1 and μ_2
- Sample sizes n_1 and n_2; common variance σ^2
- H_0: $\mu_1 = \mu_2$

- $(\bar{X}_1 - \bar{X}_2)/s_{pooled} \sim t(n_1 + n_2 - 2)$

- Welch-Satterthwaite adjustment for unequal variances
- Assumes independence within and across groups
- Paired t test is just one-sample t test on the change scores

- Generally robust against distributional violations
- Non-parametric equivalent: Two sample Wilcoxon signed rank test
- Extension with covariates: OLS, focused on 2-level Group

F test

- K groups, Normally distributed with means μ_k
- Sample sizes n_k; common variance σ^2
- H_0: $\mu_1 = \ldots = \mu_K$

- “variance explained by the group” / “unexplained variance”
- $F(K, [\Sigma\, n_k] – K)$

- Generally robust against distributional violations
- Generally robust against assumption of common variance
- Not generally robust against stupidity

- Non-parametric equivalent: Kruskal-Wallis
- Extension with covariates: OLS, focused on K-level Group

Correlations

- Two variables X and Y with estimated correlation ρ* (SS=n)
- H_0: ρ = 0 or some fixed delta

- Let P’ be the correlation observed from n pairs (X, Y) where X and Y have true correlation ρ
- Define $f_Z(r) = 0.5 * \log((1+r)/(1-r))$
- $f_Z(P') \sim N(f_Z(\rho), 1/(n-3))$ for sufficiently large n
- Test the null or build CI for $f_Z(\rho)$ as usual, transform back for a proper (asymmetric) CI for ρ

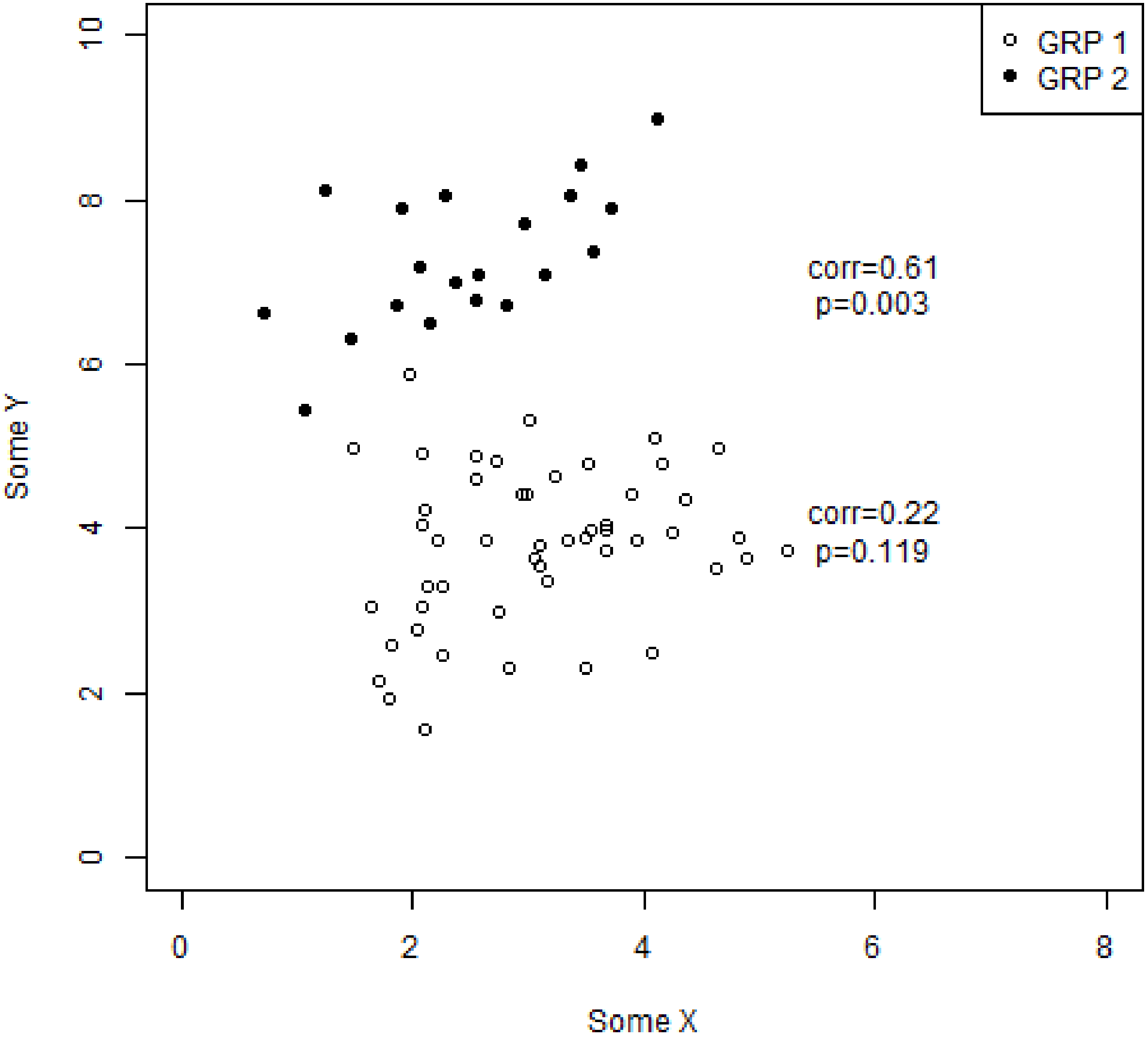
- Assumes that the sample pairs are independent
- Spearman’s can be used as well (Pearson’s on the ranks)

Pop Quiz!

- Two groups, look at corr(X, Y)

- Sample sizes n_1 and n_2
- Fisher Zs say:
  - Claim $\rho_1 \neq 0$
  - !Claim $\rho_2 \neq 0$

- Conclude $\rho_1 \neq \rho_2$?



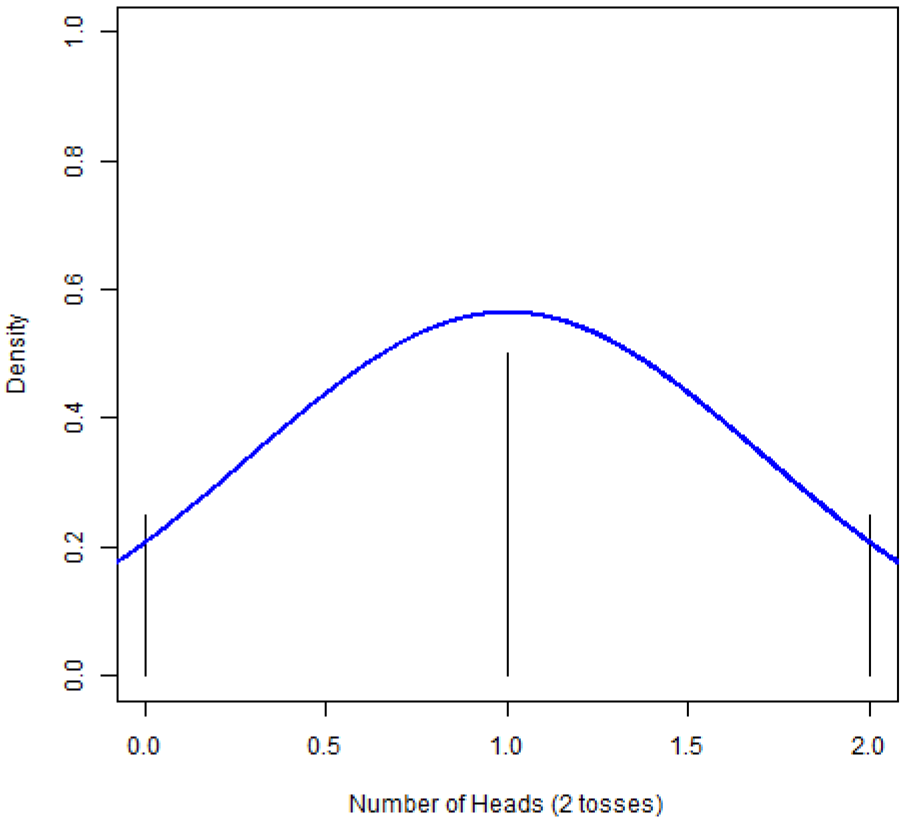
Binomial test of one proportion

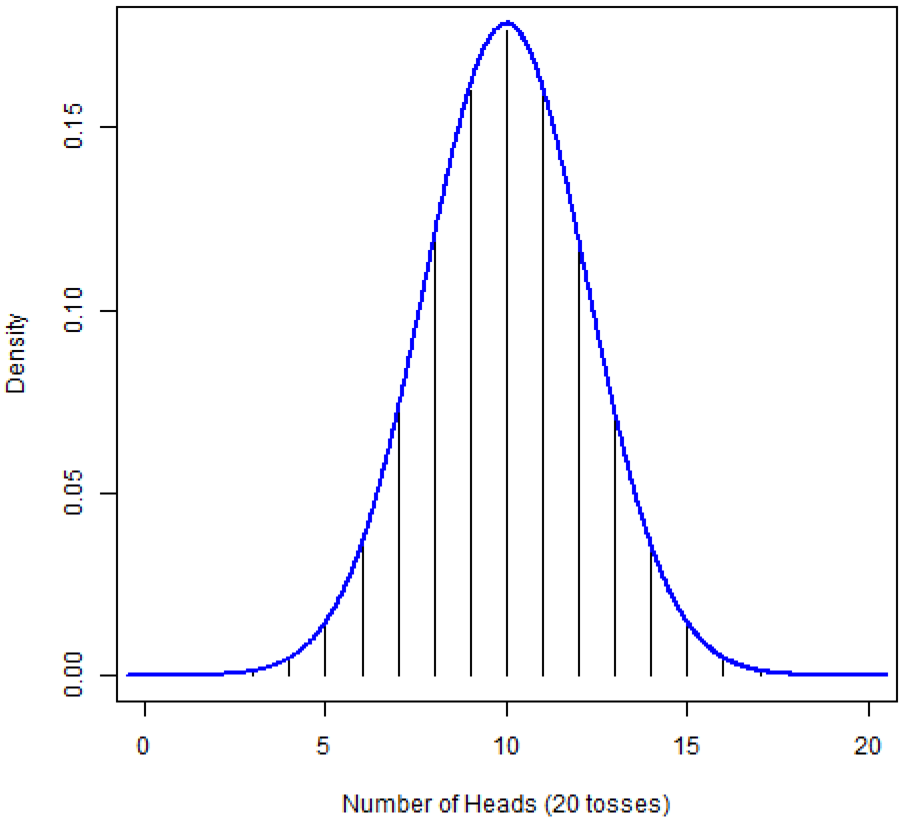
- Series of n trials with yes/no or 1/0 outcomes
- What is the probability of success p?
- H_0: p = some value

- Let X = (# successes). X ~ binom(n, p)
- X/n has expectation p, can build CI for p exactly
- If you can build a CI, you can do a point equality test
- Or: X/n is approximately N(p, p(1-p)/n)
    - Use p* in variance; or, p(1-p) bounded above by 0.25
    - Approximation not valid for small n

- Assumes that p is fixed
- Assumes that the trials are independent

Binomial -> Normal





Binomial test of two proportions

- Series of n_k trials with yes/no or 1/0 outcomes in K=2 groups
- Each group has probability of success p_k
- H_0: $p_1 = p_2$

- Let X_k = (# successes in group k). $X_k \sim \text{binom}(n_k, p_k)$
- $X_1/n_1 – X_2/n_2$ has expectation $p_1 – p_2$
- Exact calculations get tricky here, so:
  - Normal approximations
  - Chi-square tests of association generally favored instead
    - Or the exact equivalent, Fisher's exact test

- Assumes that the p_k are fixed
- Assumes that the trials in each group are independent
- Extension with covariates: Logistic regression

A chi-square test

- Counts across two factors each with 2 or more levels
- Is there an association between the factors?
- H_0: No association

3	12
7	28

10	40
25	25

A chi-square test

- Counts across two factors each with 2 or more levels
- Is there an association between the factors?
- H_0: No association

O

3	12	15
7	28	35
10	40	50

10	40	50
25	25	50
35	65	100

E

$\frac{10 * 15}{50} = 3$	$\frac{40 * 15}{50} = 12$
$\frac{10 * 35}{50} = 7$	$\frac{40 * 35}{50} = 28$

17.5	32.5
17.5	32.5

A chi-square test

- Counts across two factors each with 2 or more levels
- Is there an association between the factors?
- H_0: No association

O

3	12	15
7	28	35
10	40	50

E

$\frac{10*15}{50} = 3$	$\frac{40*15}{50} = 12$
$\frac{10*35}{50} = 7$	$\frac{40*35}{50} = 28$

$$\sum \frac{(O-E)^2}{E} = 0 + 0 + 0 + 0$$

$$= 0 \text{ on } (2-1)*(2-1) = 1 \text{ d.f.}$$

(p = 1)

A chi-square test

- Counts across two factors each with 2 or more levels
- Is there an association between the factors?
- H_0: No association

O

10	40	50
25	25	50
35	65	100

$$\sum \frac{(O-E)^2}{E} = 3.21 + 1.73 + 3.21 + 1.73$$

$$= 9.89 \text{ on 1 d.f. } (p = 0.0017)$$

E

17.5	32.5
17.5	32.5

Fisher’s exact test

- Counts across two factors each with 2 or more levels
- Is there an association between the factors?
- H_0: No association (OR = 1)

X=x	50-X	n=50
35-X	65-(50-X)	50
K=35	65	N=100

$$\sum_{\substack{\text{"}X\text{ as or more extreme than }x\text{"} \\ X\le x\text{ or }X\ge(\min(K,n)-x)}} \frac{\binom{K}{X}\binom{N-K}{n-X}}{\binom{N}{n}}$$

p = 0.003

Yates’ continuity correction

- Counts across two factors each with 2 or more levels
- Is there an association between the factors?
- H_0: No association

O

10	40	50
25	25	50
35	65	100

$$\sum \frac{(|O-E|-0.5)^2}{E} = 8.62$$

on 1 d.f. (p = 0.003)

E

17.5	32.5
17.5	32.5

Linear regression

- Continuous outcome Y
- Predictors / covariates: X_1, X_2, …, X_m
- Sample size n, indexed by j
- "Y is a linear function of the predictors"
- $Y_j = \alpha_0 + \beta_1 * X_{1j} + \ldots + \beta_m * X_{mj} + \varepsilon_j$

- Which βs are different from 0?
- $\varepsilon_j \sim$ i.i.d. $N(0, \sigma^2)$

- Generally robust against distributional violations
- Generally robust against assumption of equal variance across various values of $\alpha_0 + \mathbf{X\beta}$
- Non-parametric equivalent: Permutation test overlay
- We'll have lots more on this in the next section

Logistic regression

- Binary (1 or 0) outcome Y
- Predictors / covariates: X_1, X_2, …, X_m
- Sample size n, indexed by j
- Think: “The logged odds of Y is a linear function of the predictors”
- Log odds(Y_j) = $\alpha_0 + \beta_1 * X_{1j} + \ldots + \beta_m * X_{mj} + \varepsilon_j$

- Which βs are different from 0?
- Actually running a maximum likelihood routine

- Falls into a general class: General Linear Models (GLMs)
  - Poisson regression
  - Gamma regression
  - Many more

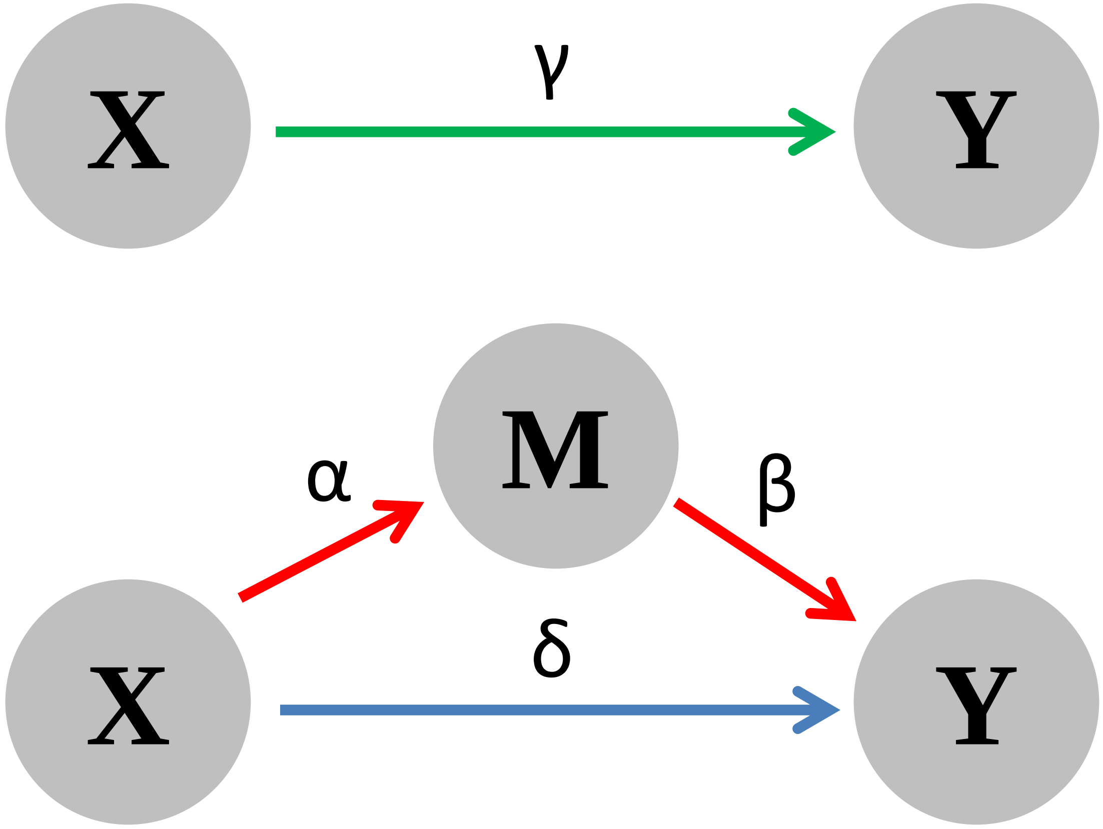
Pop Quiz!

- From *The Guardian* on March 5, 2014:
- “So the people we think of as protein-loons were always eating other stuff beside [sic] it. They are still going to live longer than you. In a longitudinal population study I've been doing, I have demonstrated that just knowing how to pronounce ‘quinoa’ extends your life expectancy by 12 years.”

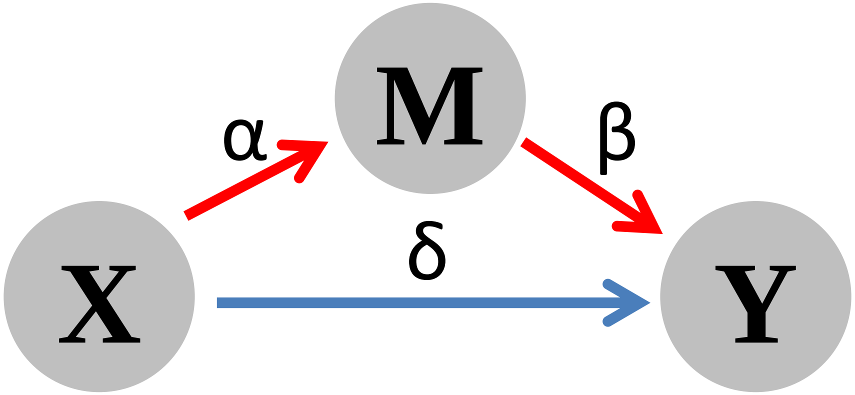
- Has your life expectancy just been extended by 12 years?

Mediation

- "Correlation does not imply causation"
- Mediational analysis attempts to quantify evidence consistent with causality



Mediation: Caveats



- The SE for αβ is not Normal, so bootstrap for it
- Mediational analysis is made *under the assumption* of a causal mediator (and *only* the given mediator(s)) !
- Thus it should really not be used as a test for causality

- Can be useful for probabilistically ruling out potential mediators

Some other situations

- We'll cover the following in greater detail later:

- Pre-post designs (OLS, MEMs or GLMs)

- Cross-over designs (OLS, MEMs or GLMs)

- Time-to-event outcomes (censored outcomes)

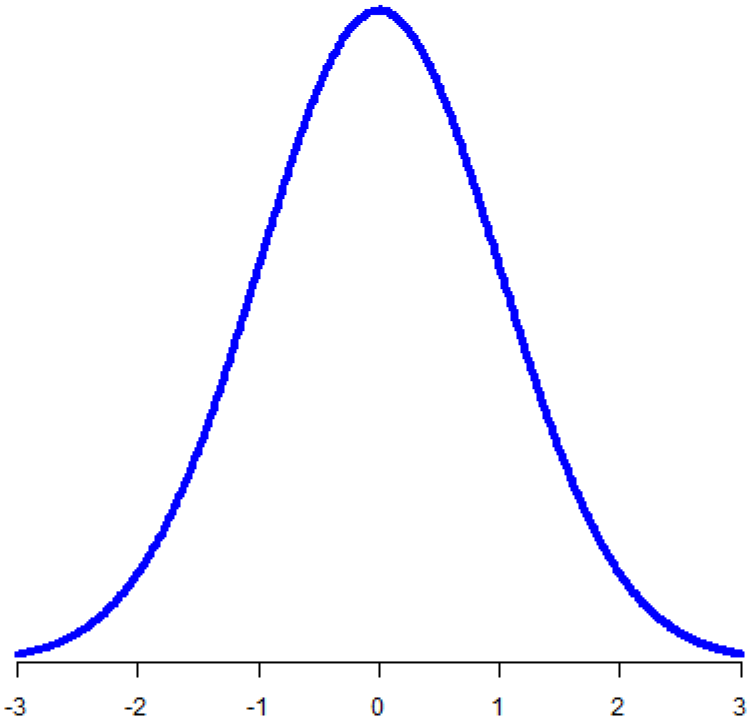
Am I Normal?
What If I’m Not?

“Standard Deviation Not Enough For Perverted Statistician.”

- Fake teaser, *The Onion*

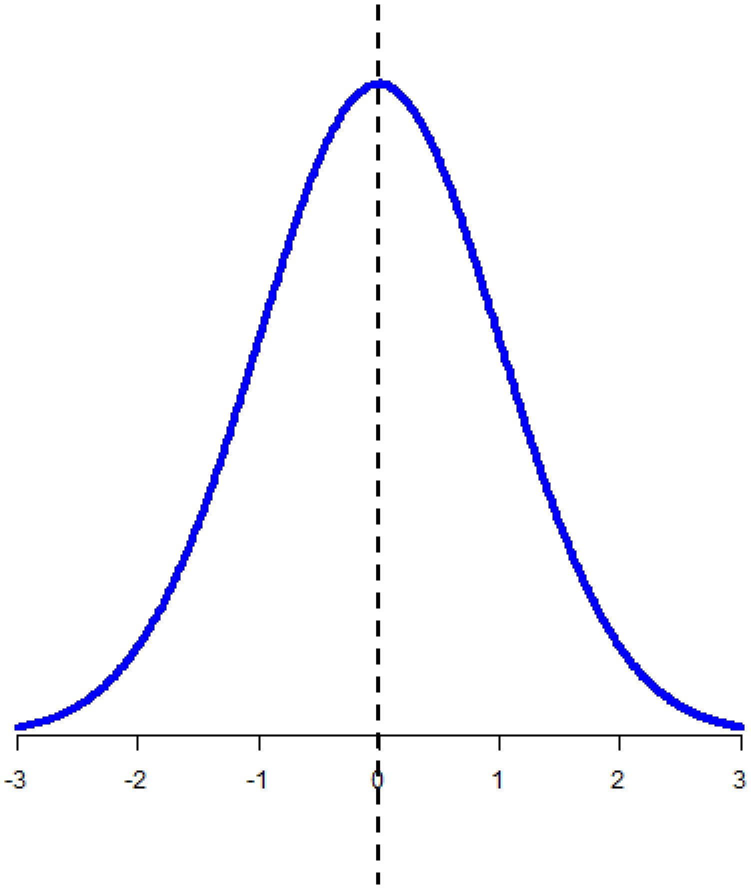
The Normal (Gaussian) Distribution

- What are the main features?
- **Unimodal**
- Normal error is almost always assumed to have mean (and mode) of 0
- Symmetry above and below the mean – no ‘skewness’
- ‘Bell-shaped’ – Kurtosis measures ‘peakedness’



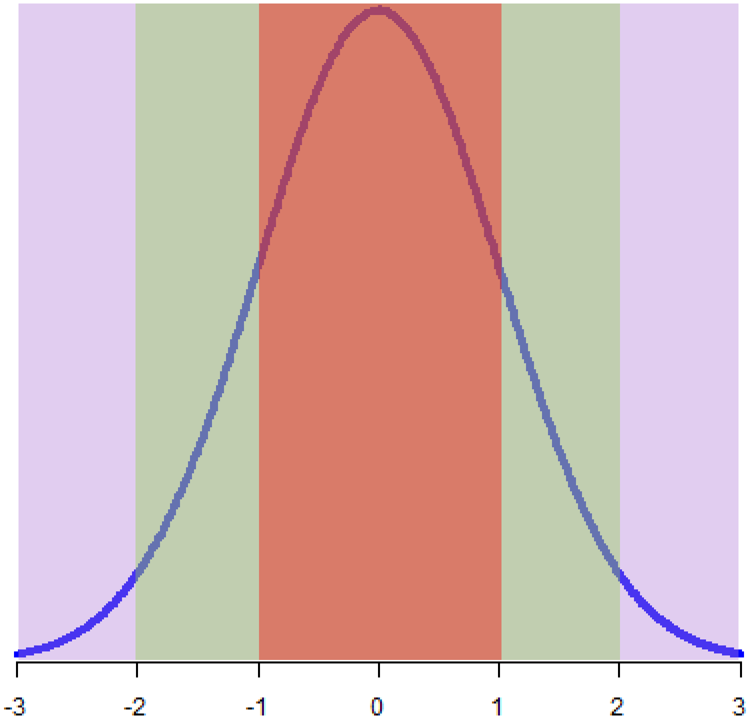
The Normal (Gaussian) Distribution

- What are the main features?
- Unimodal
- **Normal error is almost always assumed to have mean (and mode) of 0**
- **Symmetry above and below the mean – no ‘skewness’**
- ‘Bell-shaped’ – Kurtosis measures ‘peakedness’



The Normal (Gaussian) Distribution

- What are the main features?
- Unimodal
- Normal error is almost always assumed to have mean (and mode) of 0
- Symmetry above and below the mean – no 'skewness'
- **'Bell-shaped' – Kurtosis measures 'peakedness'**



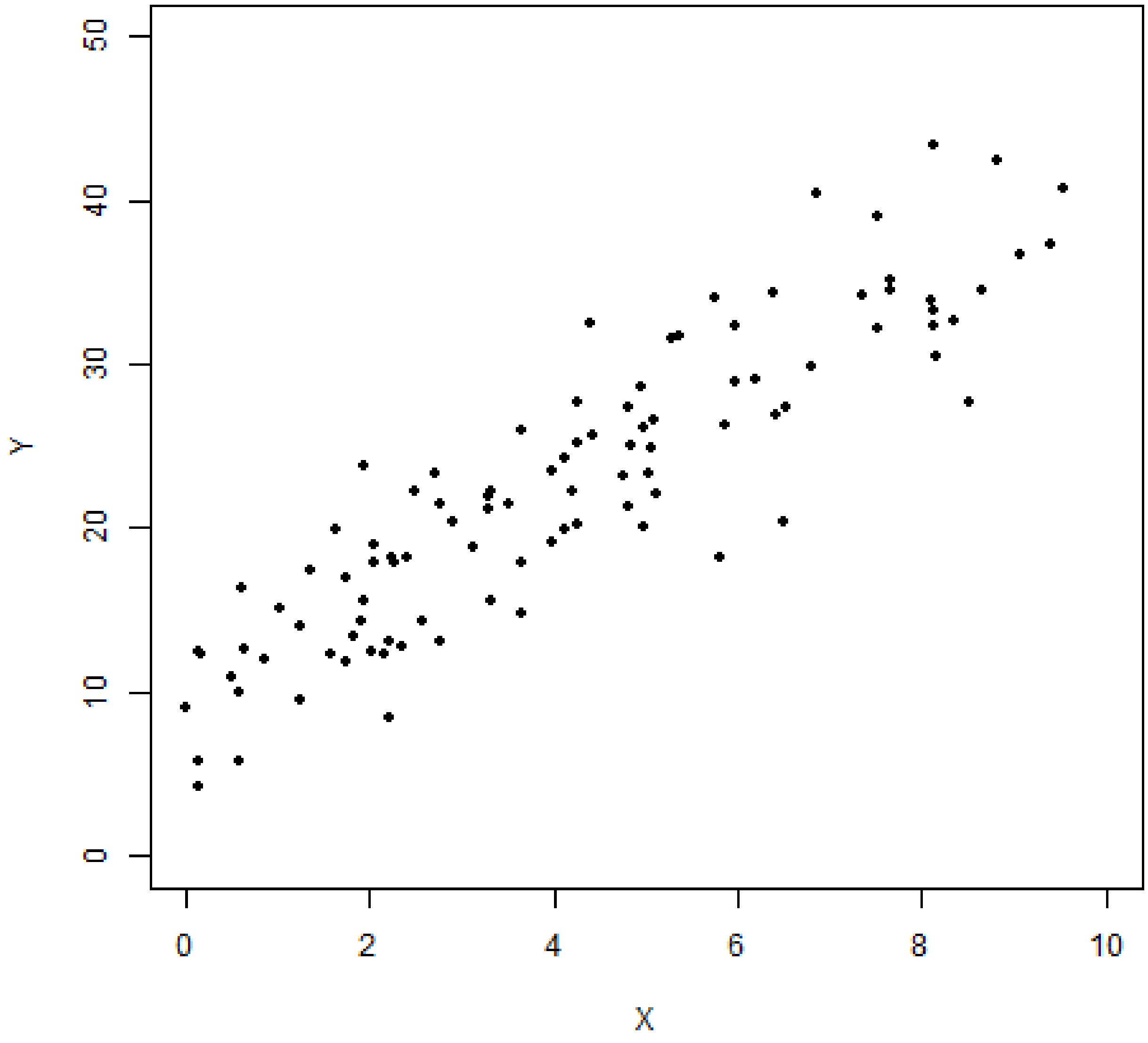
68.29% 95.45% 99.73%

Normality Will Be Violated

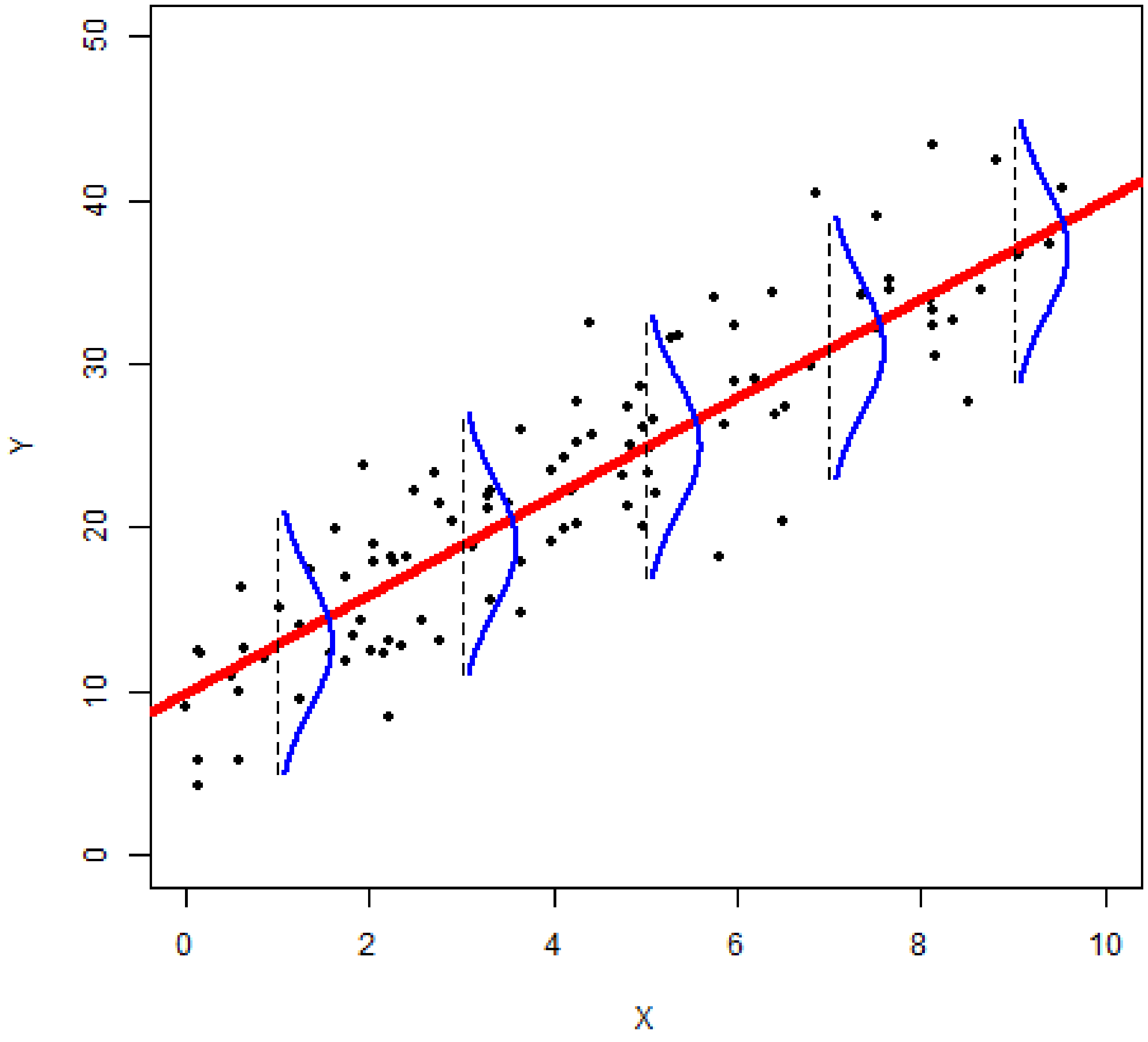
- Linear regression assumes independent, zero-centered, homoscedastic and Gaussian-distributed error
- It’s not a question of ***whether*** these assumptions are violated, but rather ***to what degree***
- Formal tests for goodness of fit exist
- Start with informal tests and diagnostics:
  - ***LOOK AT THE DATA***
  - Ask: How many violations?
  - Ask: How severe are the individual violations?

- It may be that the outcome is not well modeled by a linear combination of the predictors
  - Use some other model
  - But often a Box-Cox transformation can work wonders

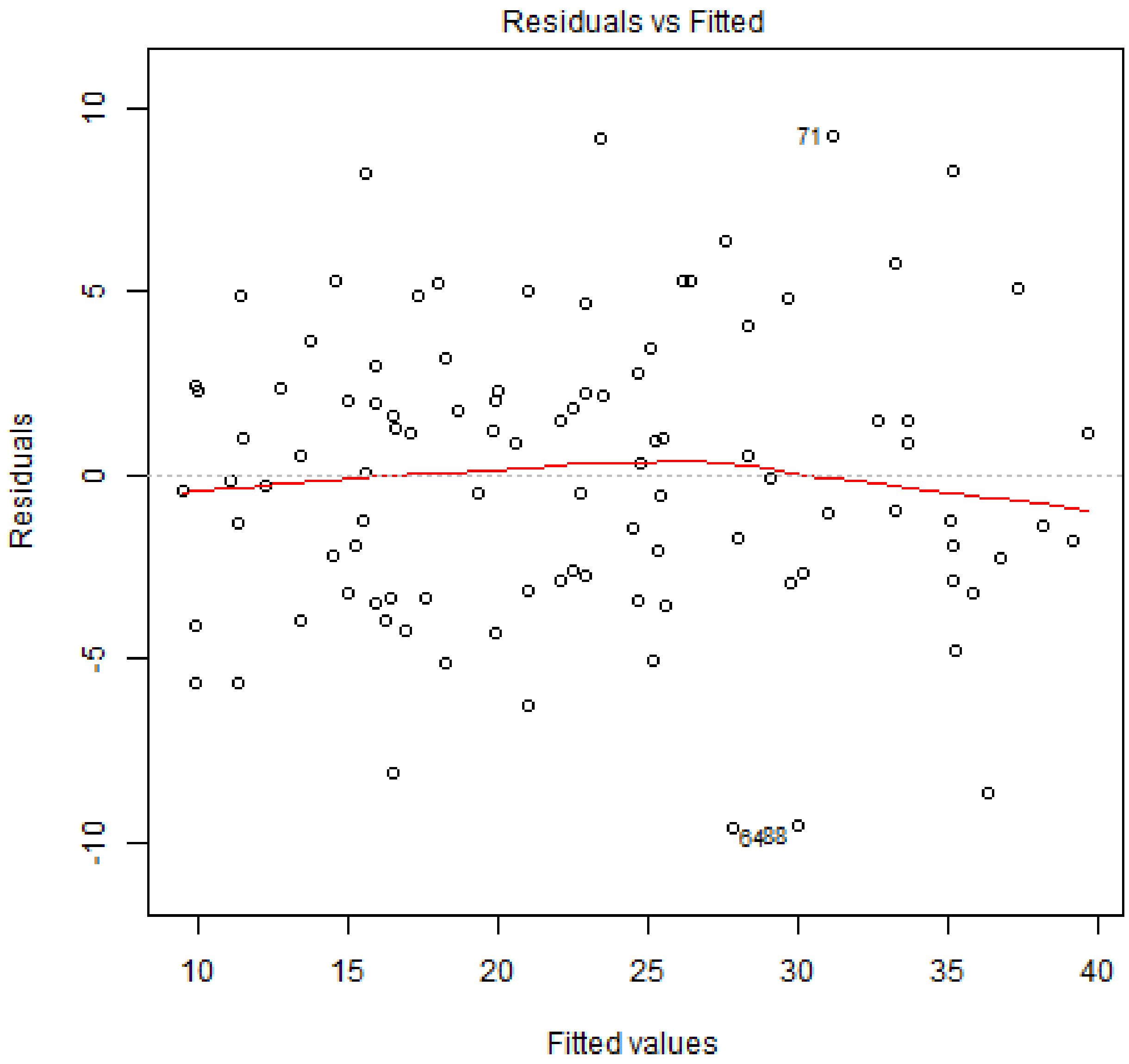
Example (1)



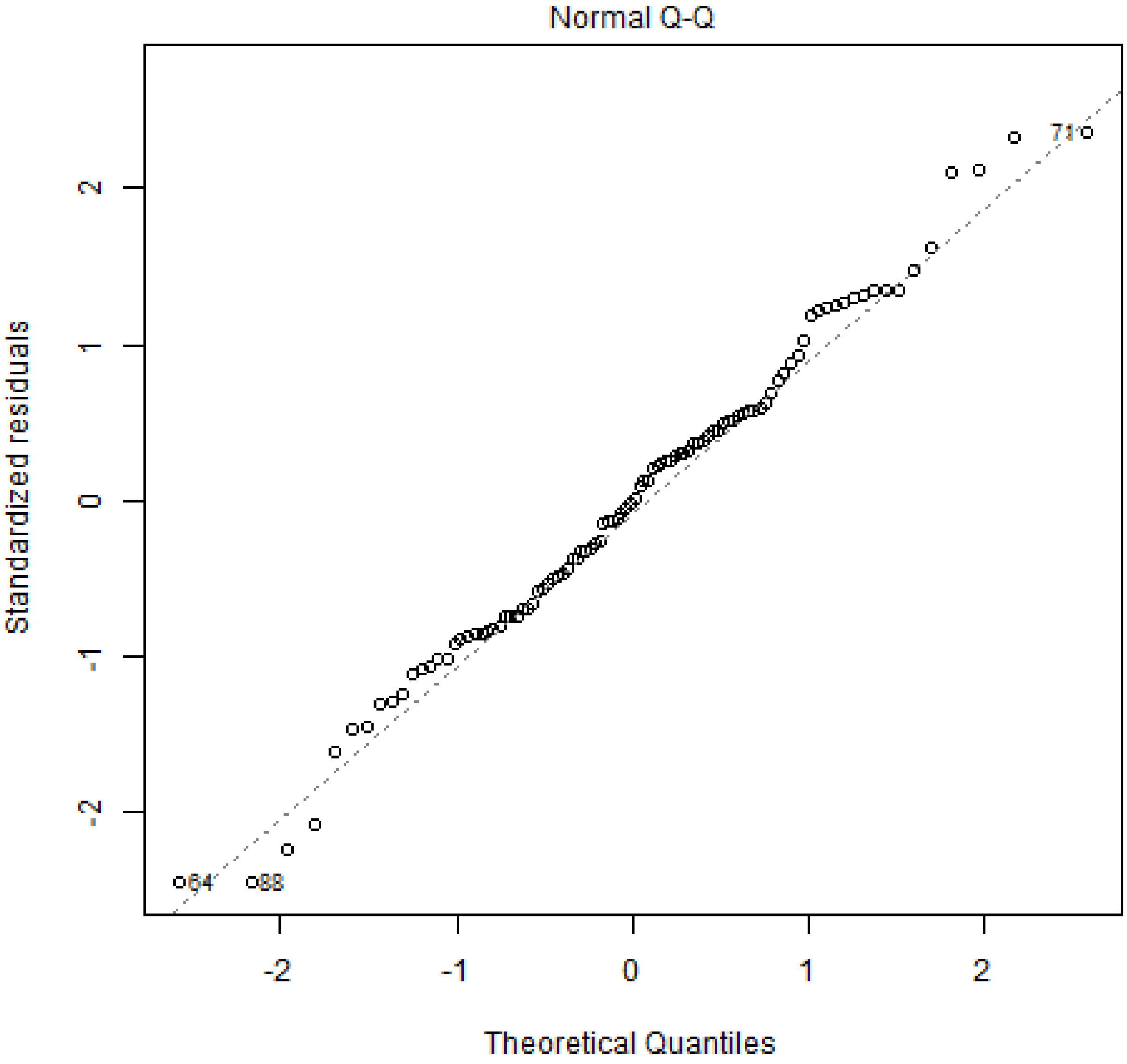
Example (1)



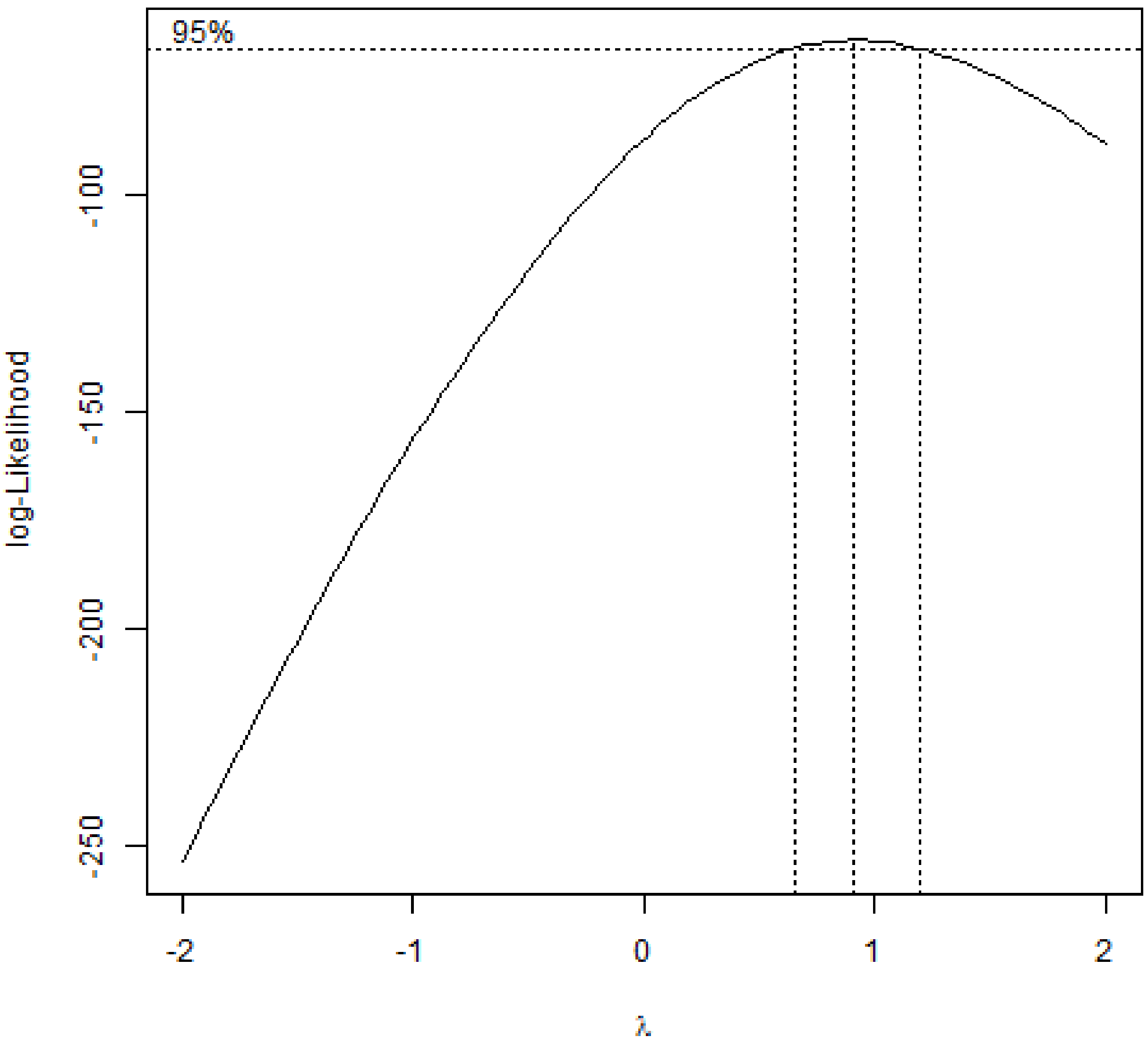
Example (1)


Residuals vs Fitted
71
64
88
Residuals
Fitted values
10
5
0
-5
-10
10
15
20
25
30
35
40

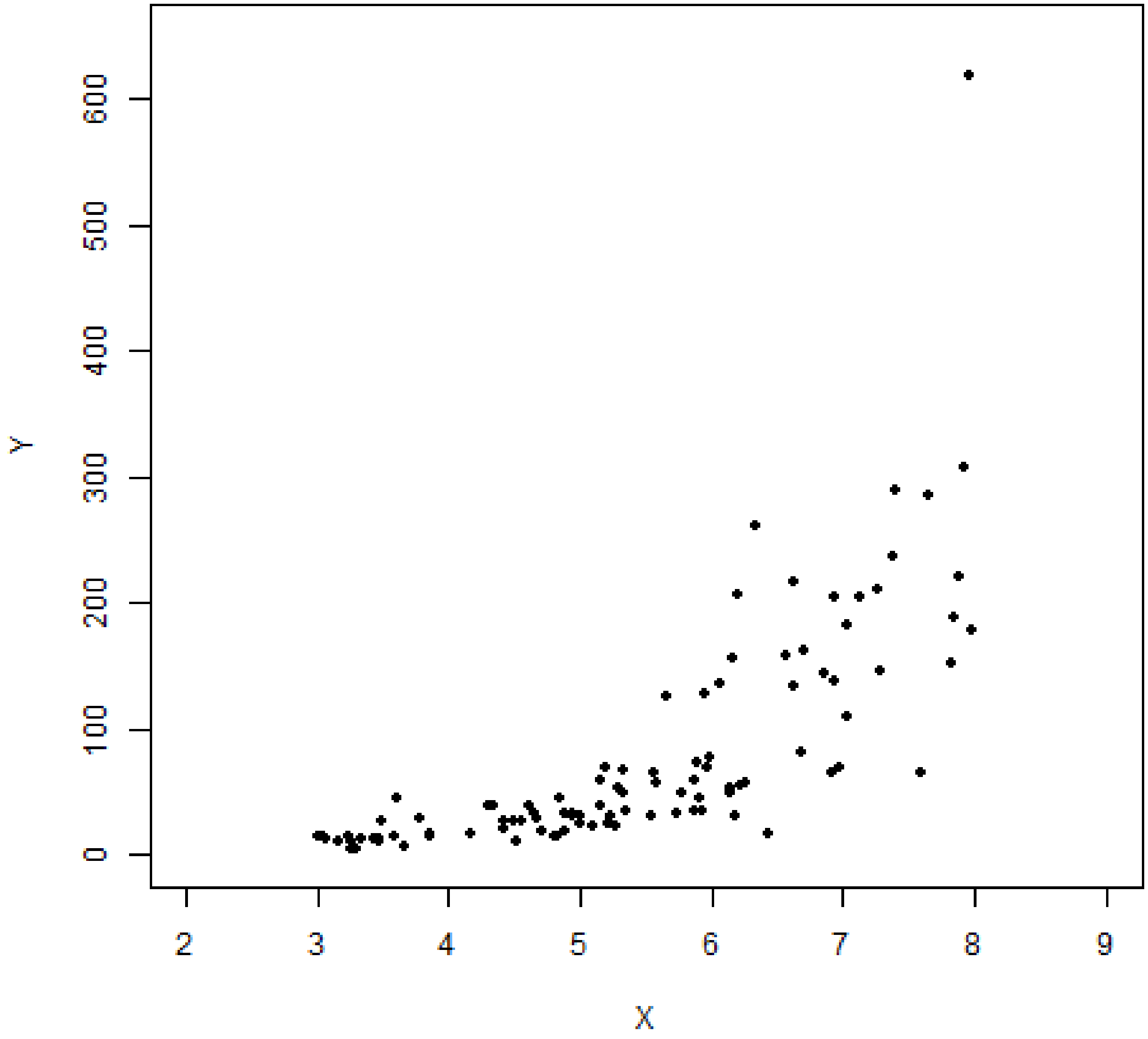
Example (1)


Normal Q-Q
Standardized residuals
Theoretical Quantiles
-2
-1
0
1
2
-2
-1
0
1
2
64
88
71

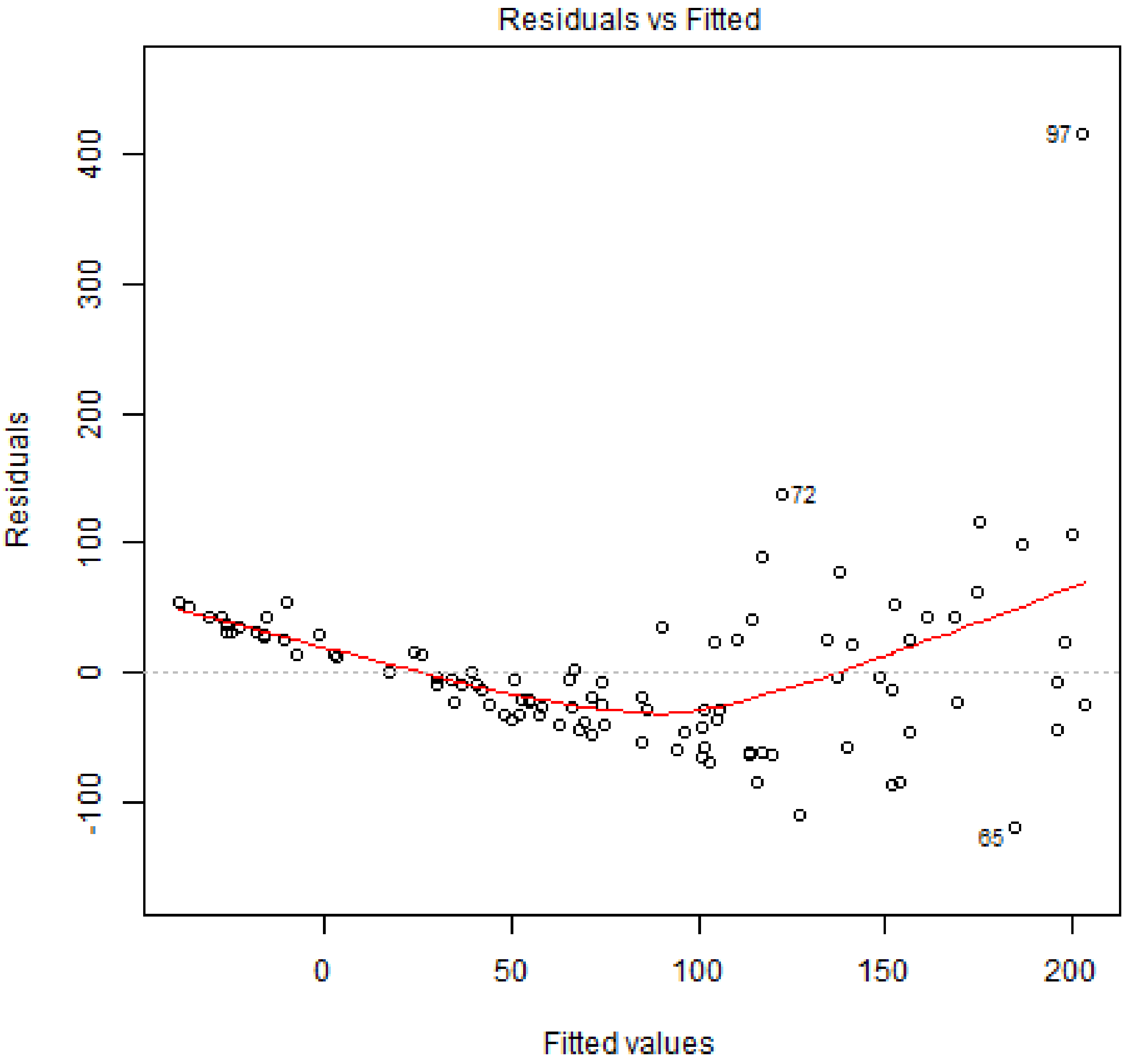
Example (1)



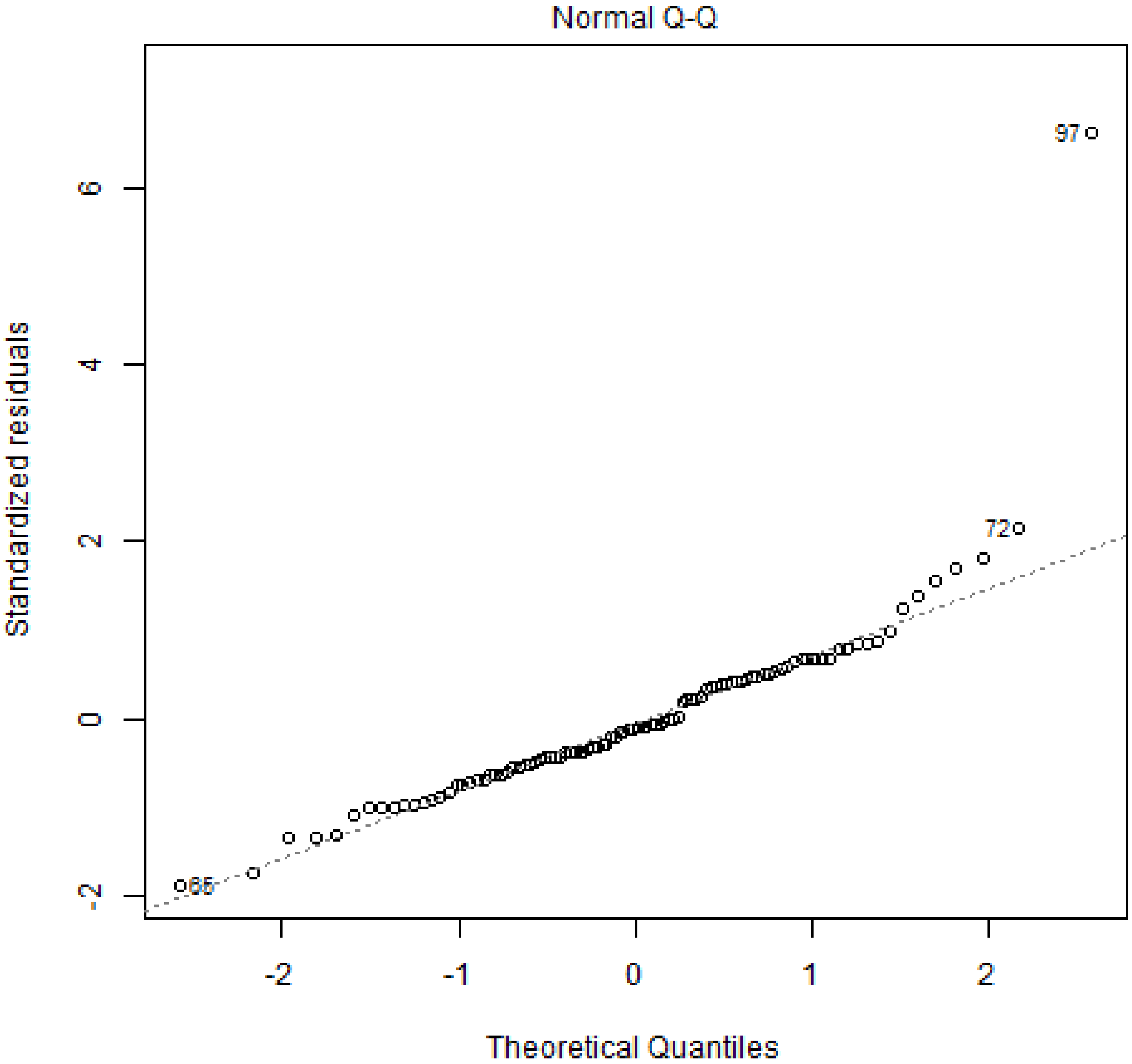
Example (2)



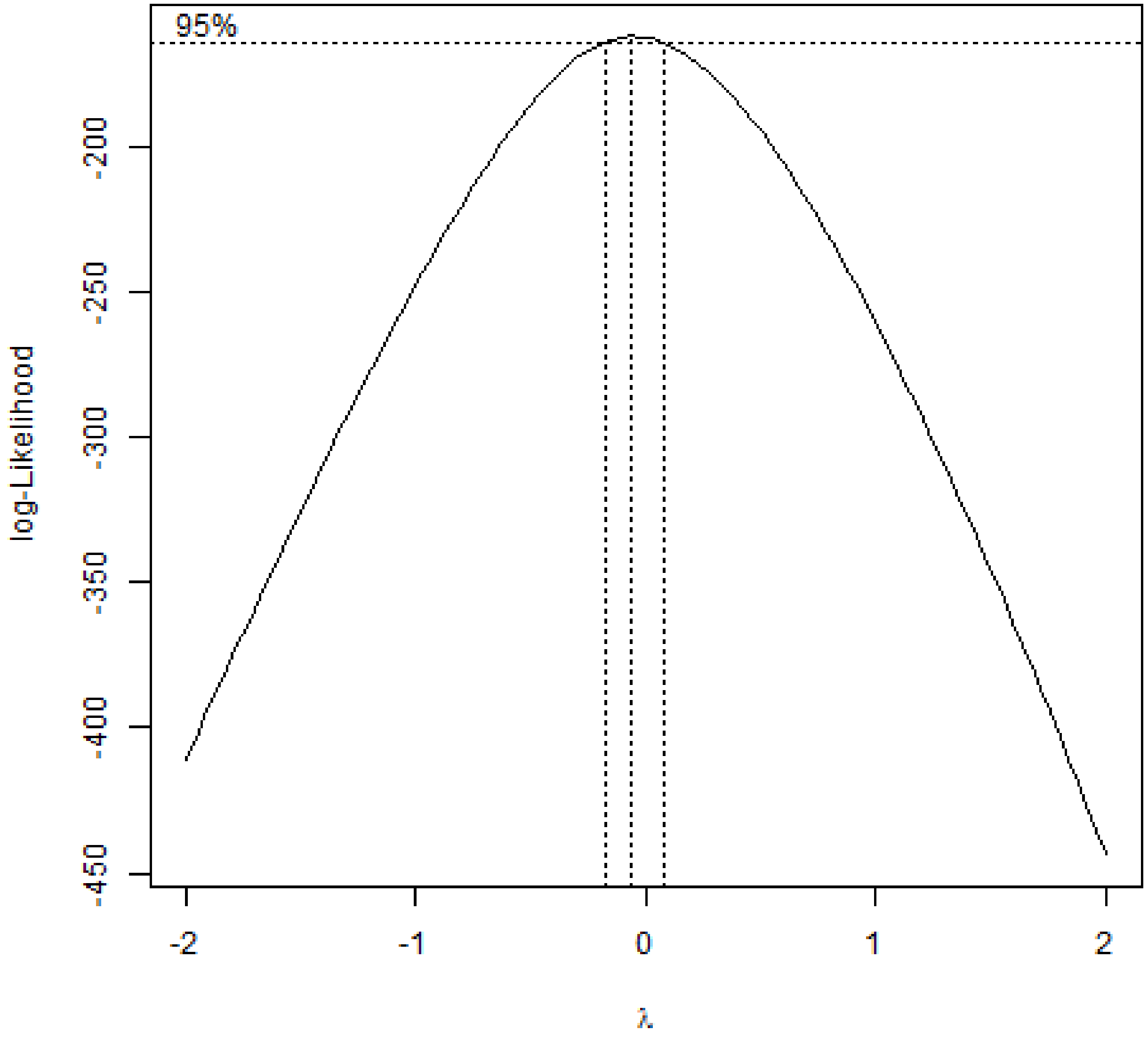
Example (2)


Residuals vs Fitted
Residuals
400
300
200
100
0
-100
Fitted values
0
50
100
150
200
97
72
65

Example (2)


Normal Q-Q
Standardized residuals
Theoretical Quantiles
-2
0
2
4
6
-2
-1
0
1
2
97
72
65

Example (2)



Fixed vs. Random Effects

“Measure twice, cut once.”

- *Unknown*

Pop Quiz!

- Compare the relative strengths of the following:

  - One measurement per subject, 100 subjects

  - Two measurements per subject, 50 subjects
  - 50 measurements per subject, 2 subjects

  - One measurement per subject, 1000 subjects
  - Ten measurements per subject, 100 subjects

  - One reported average of ten measurements per subject, 1000 subjects

Fixed effects

- Taking a clipping from each of 20 banana trees (index j)
- Testing four kinds of fertilizer (A B C D)
- Outcome is height of the new tree after one year

- $Y_j = \beta_A + \beta_B I_{Bj} + \beta_C I_{Cj} + \beta_D I_{Dj} + \varepsilon_j$

- The fertilizer factor is represented by a ‘fixed effect’
- The β values are assumed to be fixed

- Thus OLS can be referred to as a ‘fixed effects’ model

Multiple clippings

- Taking clippings from each of 20 banana trees (index j)
  - Five clippings from each tree (index k)
- Testing four kinds of fertilizer (A B C D)
- Outcome is height of the new tree after one year

- Does it make sense to use the following model?
  - $Y_{jk} = \beta_A + \beta_B I_{Bjk} + \beta_C I_{Cjk} + \beta_D I_{Djk} + \varepsilon_{jk}$

- What about this one?
  - $Y_{jk} = \beta_{A,k=1} + \beta_B I_{Bjk} + \beta_C I_{Cjk} + \beta_D I_{Djk} + \beta_2 I_{k=2} + \dots \beta_5 I_{k=5} + \varepsilon_{jk}$

Random effects

- Taking clippings from each of 20 banana trees (index j)
  - Five clippings from each tree (index k)
- Testing four kinds of fertilizer (A B C D)
- Outcome is height of the new tree after one year

- Need to add a *random effect* for origin of clipping
- $Y_{jk} = \beta_A + \beta_B I_{Bjk} + \beta_C I_{Cjk} + \beta_D I_{Djk} + \delta_j + \varepsilon_{jk}$

- $\varepsilon_{jk} \sim$ i.i.d. $N(0, \sigma^2)$
- $\delta_j \sim$ i.i.d. $N(0, \sigma_0^2)$

- This is a ‘random effects’ or ‘mixed effects’ model
- Think: “Do I care about the specific levels of this factor?”

General Considerations as a Statistical Reviewer

“Lies, damned lies, and statistics.”

- Attributed to Benjamin Disraeli, popularized by Mark Twain, *Chapters from My Autobiography*

Multiple Comparisons

- Problem: As the number of tests with a fixed Type I error rate increases, so does the probability of at least one false discovery
- Pr(At least one FD) = $1 – (1-\alpha)^m$

- Why should you care?
- Why will your statistical reviewer care?

- Bonferroni: $1 – (1-\alpha/m)^m \approx \alpha$

- “Nominal significance”
- “Secondary analyses”
- “Post-hoc tests”
- “Suggestive findings”

False Discovery Rate

- Problem: The number of tests is huge
- FDR says: We found k things to be significant at an FDR rate of (say) 5%. Therefore we expect (0.95*k) of those k to be true discoveries

- Why should you care?
- Why will your statistical reviewer care?

- Question of power but also of philosophy

- “(Storey) q-value”
- Often, these are Bayesian methodologies

Reproducibility

- Problem: “We used PROC MIXED”

- Why should you care?
- Why will your statistical reviewer care?

- Understanding what you did, if it’s valid
- Proliferation of non-reproducible research
- Making data and code available to the public

- List procedures, methods and software with version #s
- “We set an arbitrary but fixed seed”

Data Cleaning / Processing

- Problem: In theory there are data sets that do not require some sort of pre-processing. If you find one, let me know
- Executive decisions will need to be made regarding artifacts, potential outliers, recording errors, variable calibration

- Why should you care?
- Why will your statistical reviewer care?

- “Statistical degrees of freedom” – Andrew Gelman

- “Tertiles” or “fifths”
- Inconsistent sample sizes or “outliers”
- Hazily justified / arbitrary segmentations
- Data-driven segmentations

Let me count the ways

- “Fried food consumption, genetic risk, and body mass index: gene-diet interaction analysis in three US cohort studies”
- http://www.bmj.com/content/348/bmj.g1610

- “We also estimated the differences in BMI per increment of 10 risk alleles stratified by three categories of fried food consumption. An interaction between the genetic risk score and consumption on BMI was tested by including an interaction term in the models. Potential confounders considered in multivariable models were age (continuous), physical activity (in fifths), television watching (0-1, 2-5, 6-20, 21-40, >40 hours/week), smoking (never, past, current), alcohol intake (0, 0.1-4.9, 5.0-9.9, 10-14.9, ≥15 g/day), intake of sugar sweetened beverages (<1 serving/month, 1-4 servings/month, 2-6 servings/week, ≥1 servings/day), alternative healthy eating index (in fifths), *trans*-fat intake (in fifths), Western-diet pattern score (in fifths), and total energy intake (in fifths).”

Protein example

- “Low Protein Intake Is Associated with a Major Reduction in IGF-1, Cancer, and Overall Mortality in the 65 and Younger but Not Older Population”
- Headline result: HR for all-cause mortality, ‘high protein’
  - 50-65: HR 1.74 (1.02 – 2.97)
  - 66+: HR 0.72 (0.55 – 0.94)
- http://www.cell.com/cell-metabolism/abstract/S1550-4131%2814%2900062-X

- “Cox proportional hazard models were used to estimate the association between intake of calories from protein on subsequent all-cause, CVD, cancer, and diabetes mortality, with the latter three run using competing-risks structures. Next we tested the interaction between age and protein consumption on the association with mortality. Based on these results, we categorized subjects into two age groups (50-65 years and 66+ years), which were used in the remainder of the analyses.”

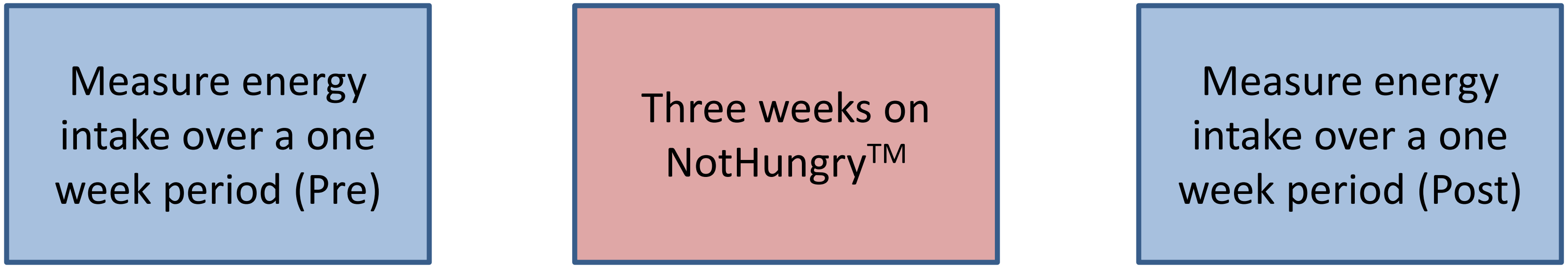
Analysis of Pre-Post Experiments

"I'm not the same man I was yesterday, am I?"

- Sanzo, *Saiyuki*

RCT Proposal

- Testing an appetite suppression drug NotHungry^{TM} for use in obese men to combat obesity
- N subjects
- Want to see if energy intake between post- and pre-intervention is different from 0
- What's wrong here?

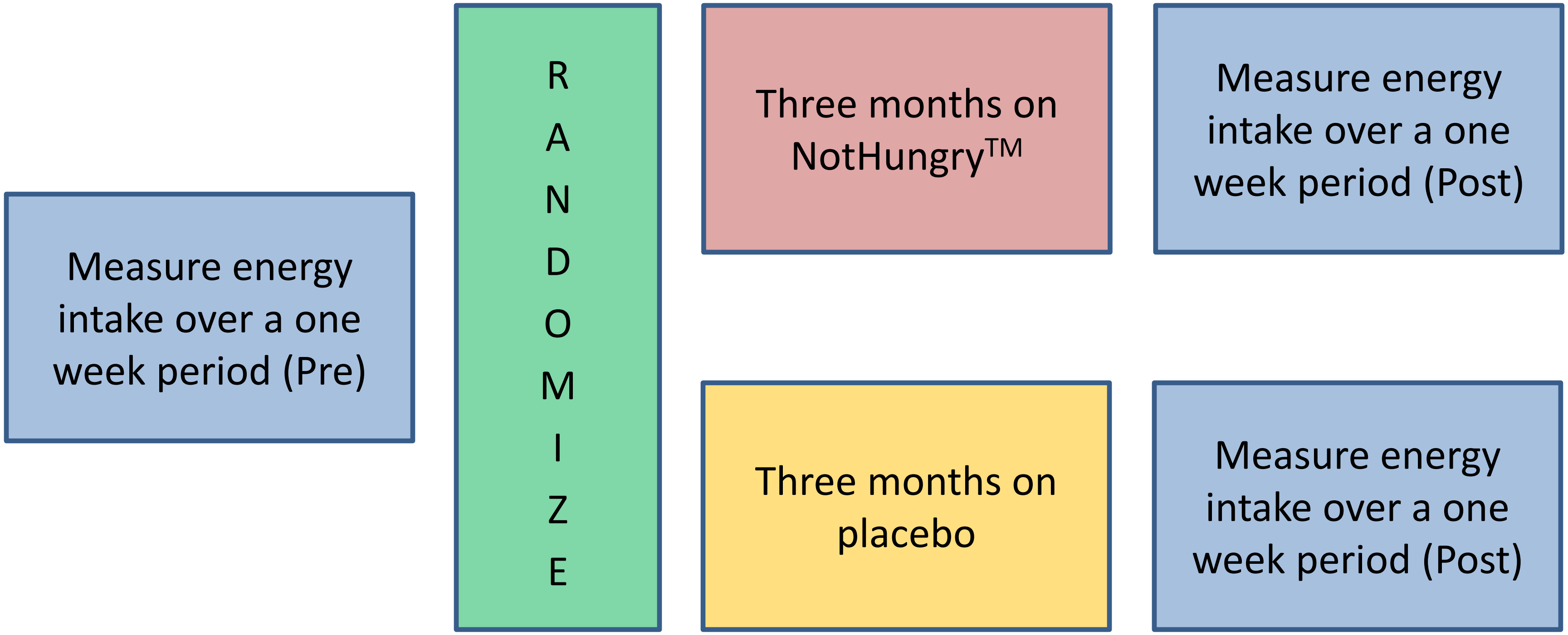

Measure energy intake over a one week period (Pre)

Three weeks on NotHungry^{TM}

Measure energy intake over a one week period (Post)

One stern talking-to later …

- Testing an appetite suppression drug NotHungry^{TM} in obese men against a placebo
- N subjects, randomly allocated to drug or placebo
- Want to see if the change in energy intake between post- and pre- intervention is different between the two arms



Measure energy intake over a one week period (Pre)

RANDOMIZE

Three months on NotHungry^{TM}

Measure energy intake over a one week period (Post)

Three months on placebo

Measure energy intake over a one week period (Post)

Two ways to think about this

- Repeated measures with two time points
  - Need a random effect for subject
  - Immediately extends to multiple time points
  - Potential for abuse by ‘completer’ analysts
  - What’s the effective sample size?

- Two group comparison on the change scores
  - Only one outcome so no random effects
  - Should still include the baseline as a covariate
  - Other covariates can be incorporated easily as well
  - What’s the effective sample size?

- Two related topics: Intent-to-treat and SS / Power calcs

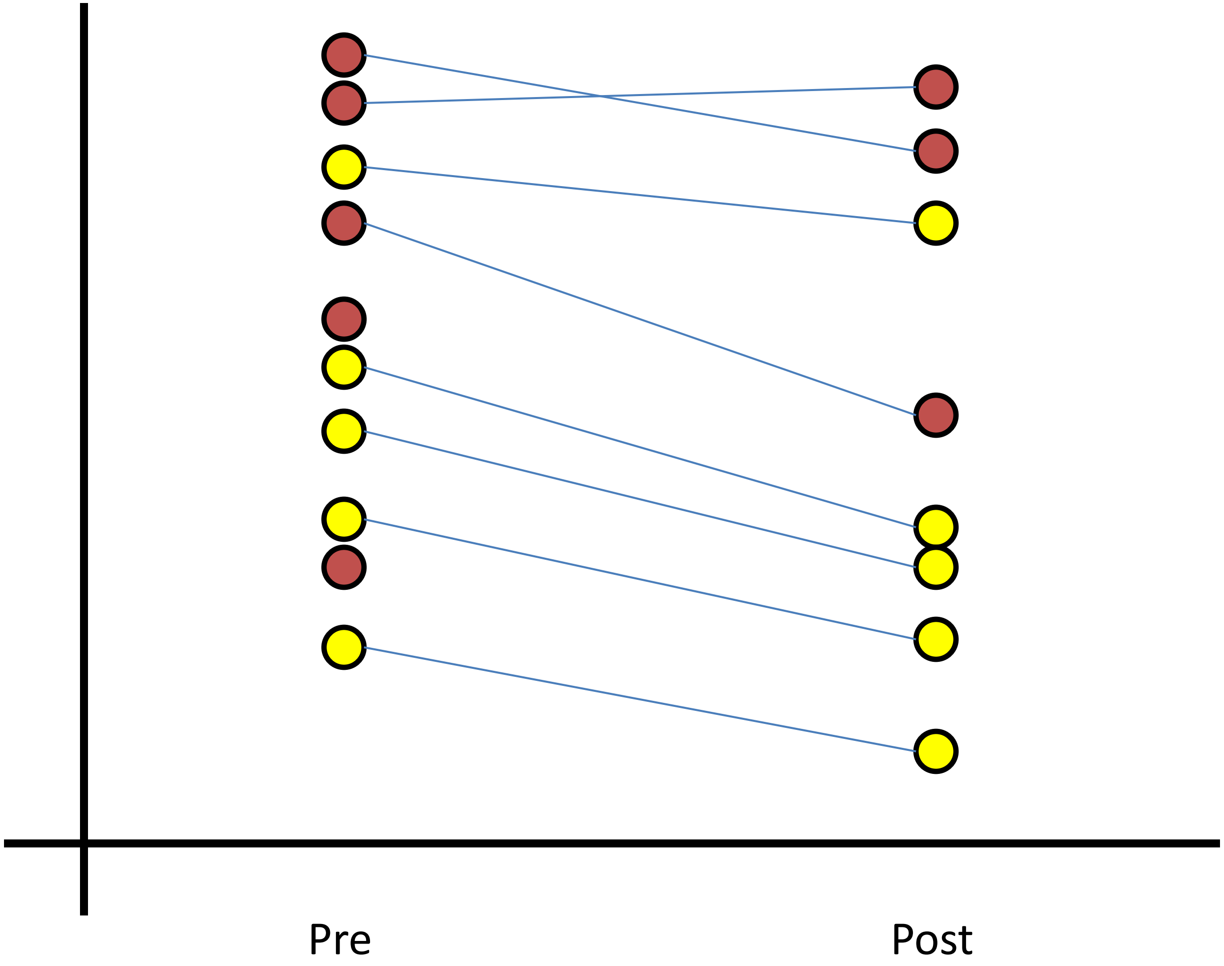
Intent-to-treat

- Related issue not specific to pre-post designs
- Generally subjects are coded according to the group to which they were randomized
- This is called an *intent-to-treat* analysis

- Subjects may not be in compliance
  - Stop taking the drug because of side effects
  - Off-label use of a medication
- Taking this into account is an *as-treated* analysis

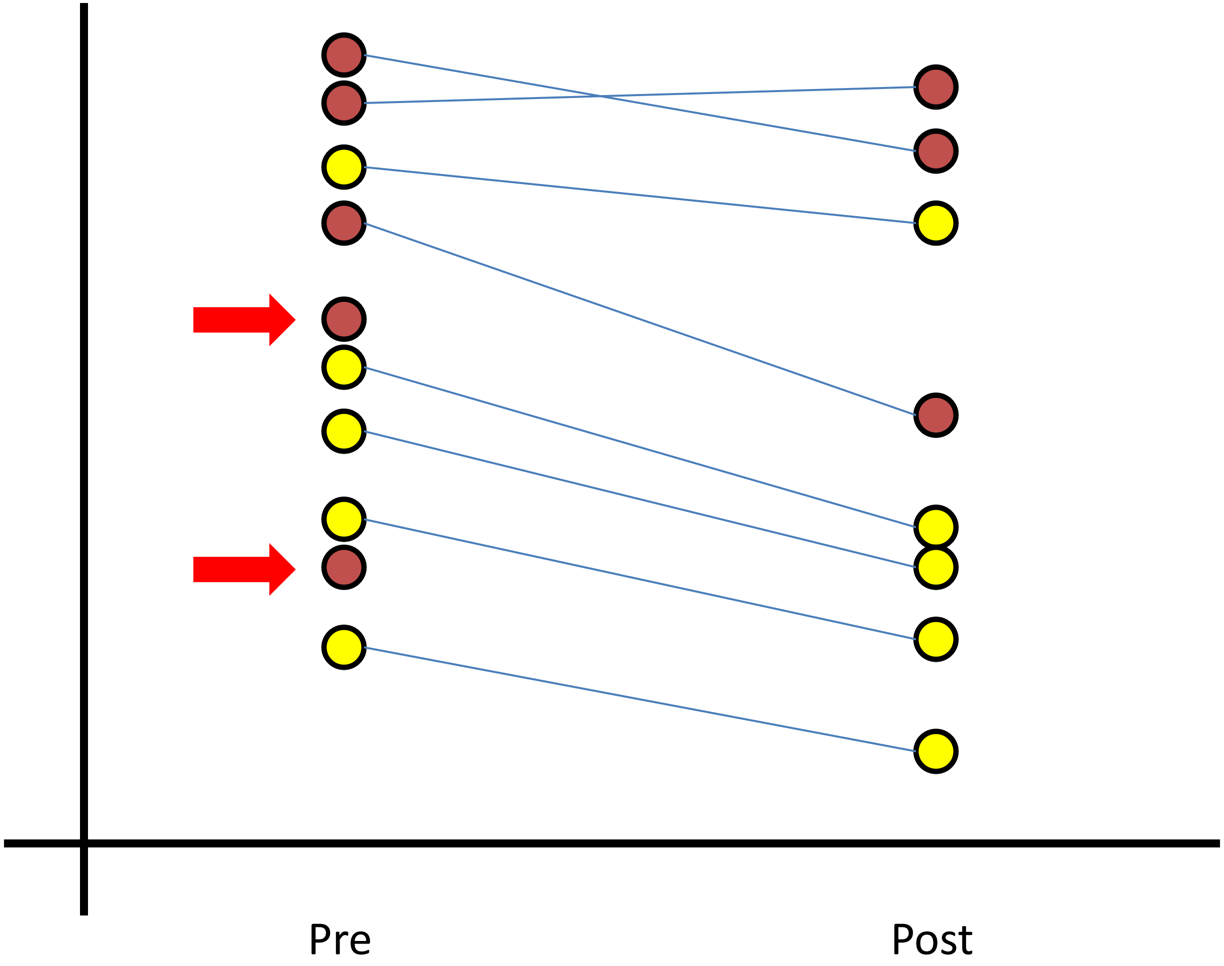
- What are the relative merits?
- How do these change interpretation?
- What do governmental regulatory agencies say?

Example Pre-Post Data

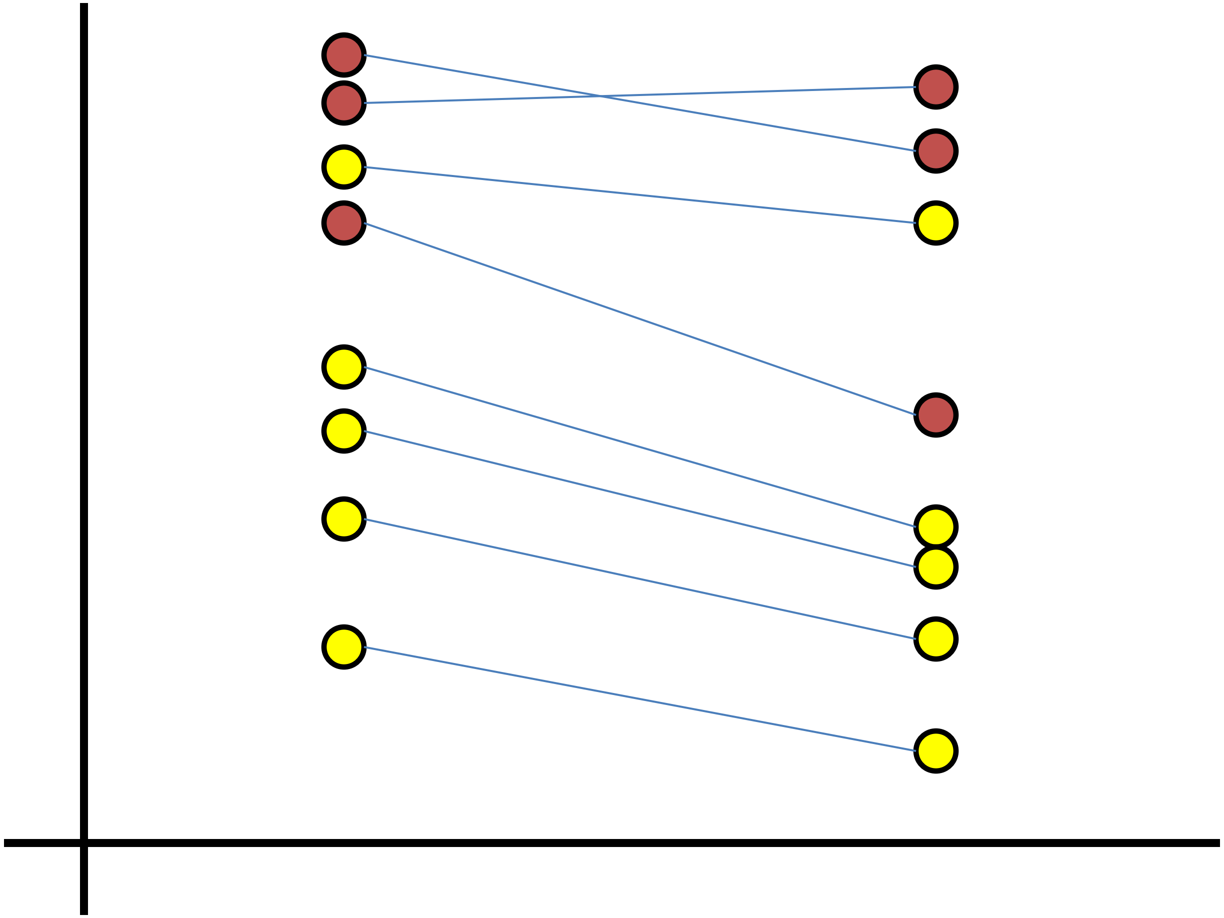


Missing Data

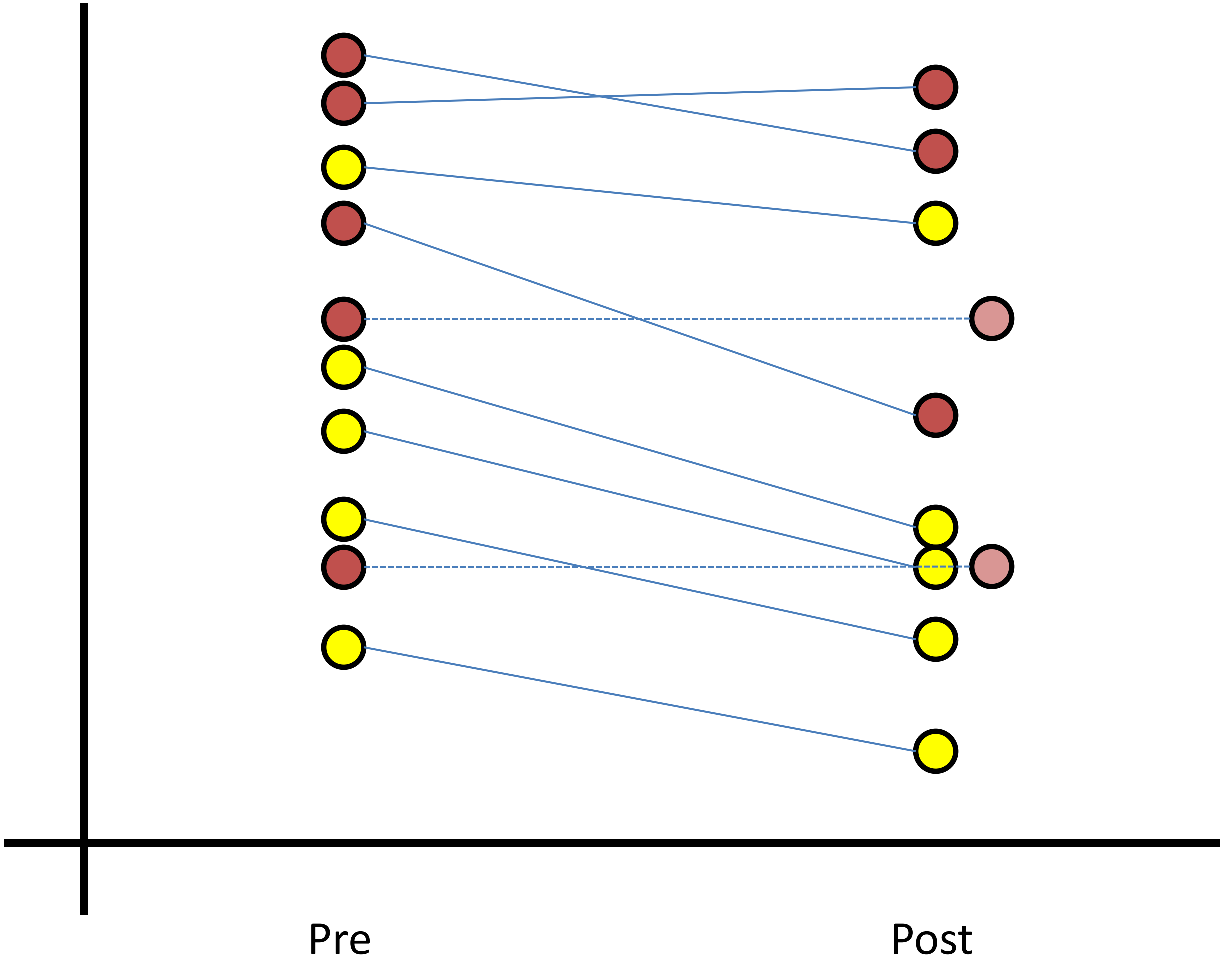
- What problems arise from the incomplete follow-up here?
- What can be done? What should be done?



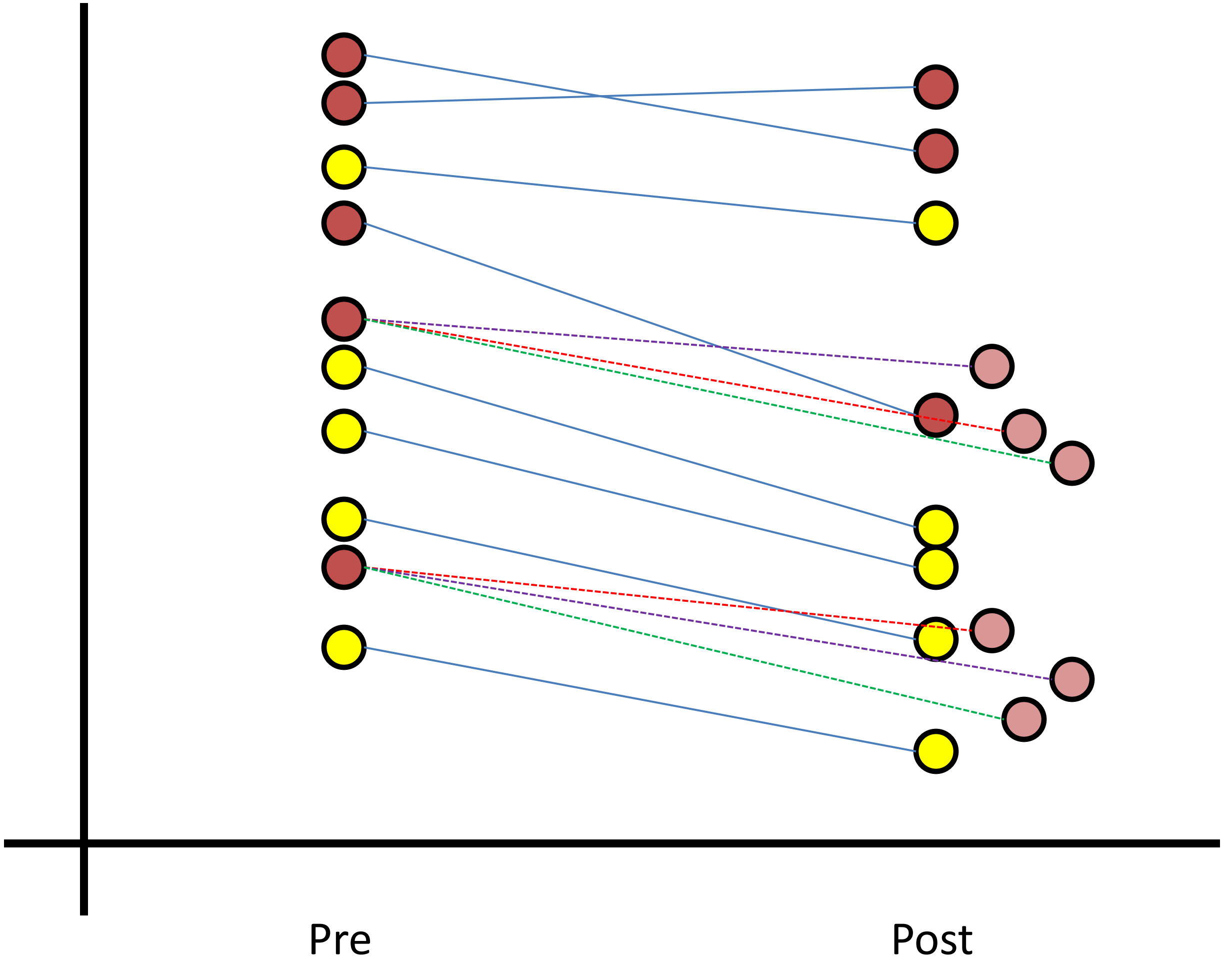
Completers Only



LOCF



Imputation



Sample Size Determination

$$\frac{\text{Available \$}}{\text{\$ per subject}}$$

Sample Size / Power Determination

- Two of the following get you the third:
  - Power
  - Sample size (per group)
  - Effect size

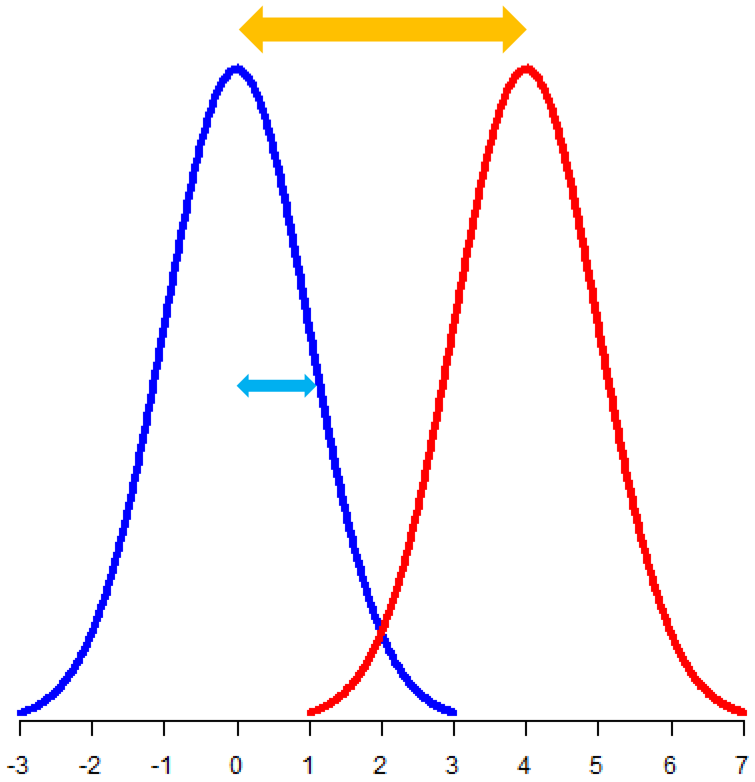
- Power is usually constrained if not a priori fixed (~80%)
- Sample size is often bounded above (available resources)
- Effect size is often based on (educated) guesswork

- 'Quick & dirty' approximations based on t-test
  - Pre-defined software functions for exact calculations
- More tailored power analyses can be done by simulation

What’s the Effect Size?

- ‘Signal-to-noise’ ratio
- $| \mu_1 - \mu_2 | / \sigma$
- Unitless
- In this example the true effect size is 4

- Estimated with sample plug-ins
- Cohen’s d uses the equal-variance t-test estimate of the pooled variance
- Hedge’s g uses the estimate from Welch’s t-test



Take Home Q&D Relation

$$\frac{2 * \left|Z_{\alpha/2} - Z_{1-\beta}\right|^2}{d^2} \leq n$$

Take Home Q&D Relation

Normal lower quantile

Type II error rate
Power = 1-β

$$\frac{2 * \left|Z_{\alpha/2} - Z_{1-\beta}\right|^2}{d^2} \leq n$$

Type I error rate

Effect size

Per group sample size

$Z_{0.05/2} =$ -1.96

$Z_{0.005/2} =$ -2.81

$Z_{0.8} = 0.84$

$Z_{0.9} = 1.28$

Pop Quiz!

- You've done the calculations for your two-arm trial
- Under restrictions / assumptions of
  - Normality in each group
  - Equal variances in each group
  - Two-tailed α of 0.05
  - 90% desired power
  - Effect size of 1

- The formula spits out 21.01 subjects per group
- An exact calculation yields 22.02 subjects per group

- For how many subjects should you budget?

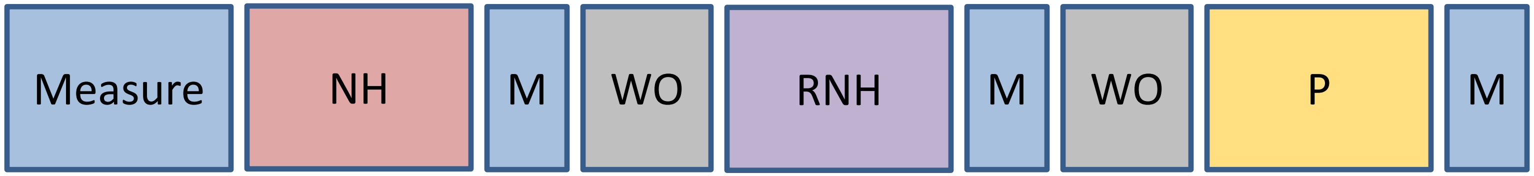
Analysis of Cross-over Experiments

“I owe my solitude to other people.”

- Attributed to Alan Watts

Many Treatments, One Study

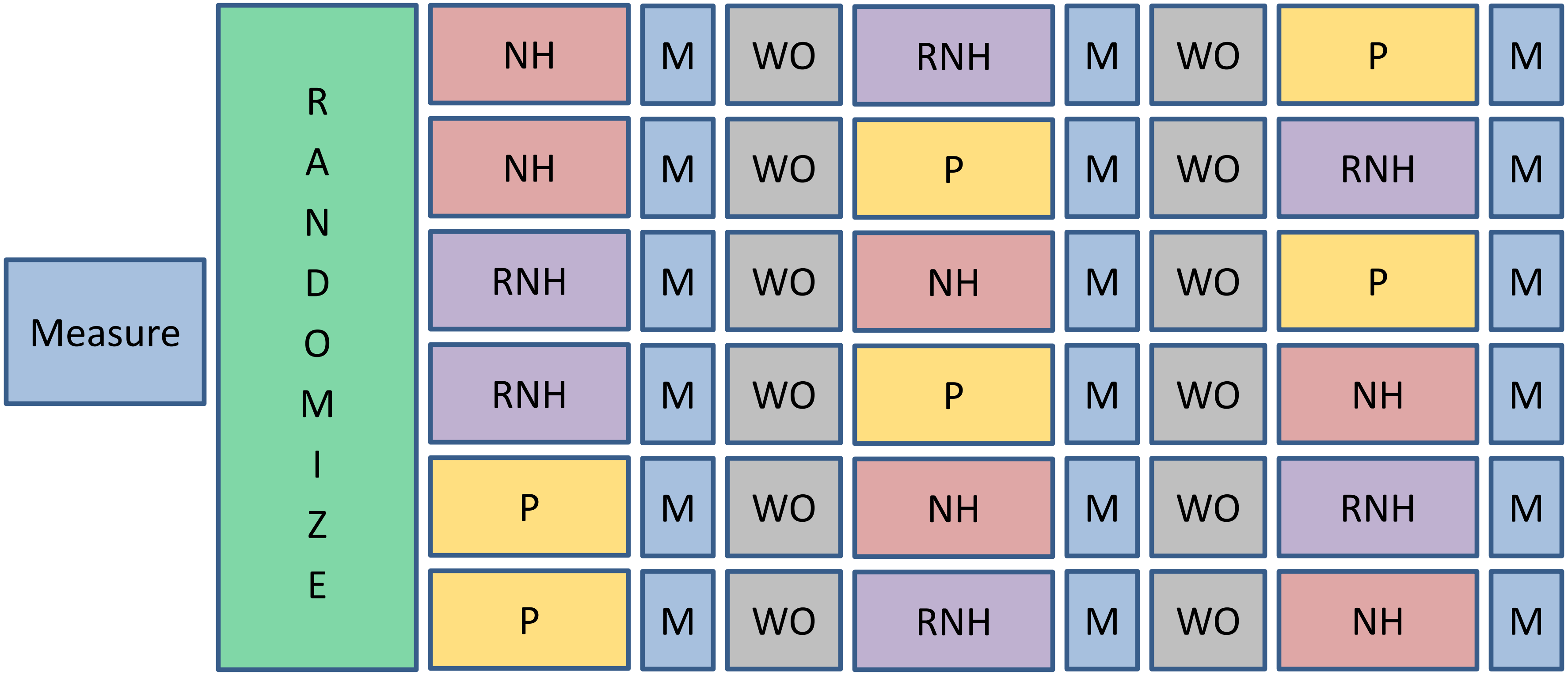
- Test the efficacy of a new appetite suppression drug ReallyNotHungry$^{\text{TM}}$ in obese men
- N subjects in a cross-over design
- “Every subject serves as his own control”
- Wash out period between treatments
- What’s wrong here?



Measure NH M WO RNH M WO P M

A Full Cross-over Design

- Test the efficacy of a new appetite suppression drug ReallyNotHungry^{TM} in obese men
- N subjects in a cross-over design
- “Every subject serves as his own control”



Measure	RANDOMIZE	NH	M	WO	RNH	M	WO	P	M
		NH	M	WO	P	M	WO	RNH	M
		RNH	M	WO	NH	M	WO	P	M
		RNH	M	WO	P	M	WO	NH	M
		P	M	WO	NH	M	WO	RNH	M
		P	M	WO	RNH	M	WO	NH	M

Analysis

- K treatments, all given to each of N subjects
- $\mathrm{EI} = \alpha + \beta_{\mathrm{treat}} + \gamma_{\mathrm{treat,\ prev}} + \delta_{\mathrm{subj}} + \varepsilon$

One grand mean

K of these

K*(K-1) of these

N of these ~ $N(0, \sigma_0^2)$

N*K of these ~ $N(0, \sigma^2)$

- N should be a multiple of K!, for balance
- From here, analyze as any other mixed model
- Can cut down model size by using only K cross-over terms γ
- Or perhaps none; is there complete wash out?

Analysis of Time-to-event data

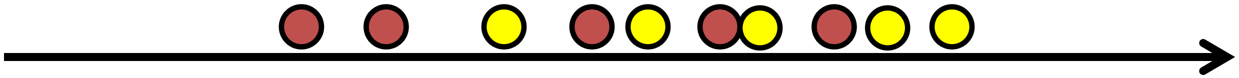
"Ask her to wait a moment – I am almost done."

- C. F. Gauss, *as recorded in Men of Mathematics*

Why are time-to-event data special?

- Given the data below, how might we analyze them?
  - t-test
  - Count models (Poisson, negative binomial)

- What if there are missing data?
  - We can impute much of this problem away

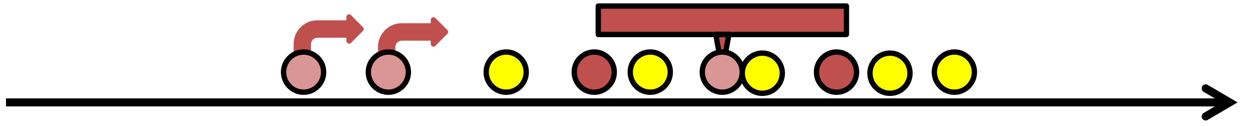


Why are time-to-event data special?

- Given the data below, how might we analyze them?
  - t-test
  - Count models (Poisson, negative binomial)

- What if there are missing data?
  - We can impute much of this problem away

- What if, rather than being missing, some observations are *censored* (usually *right-censored* or *interval-censored*)?



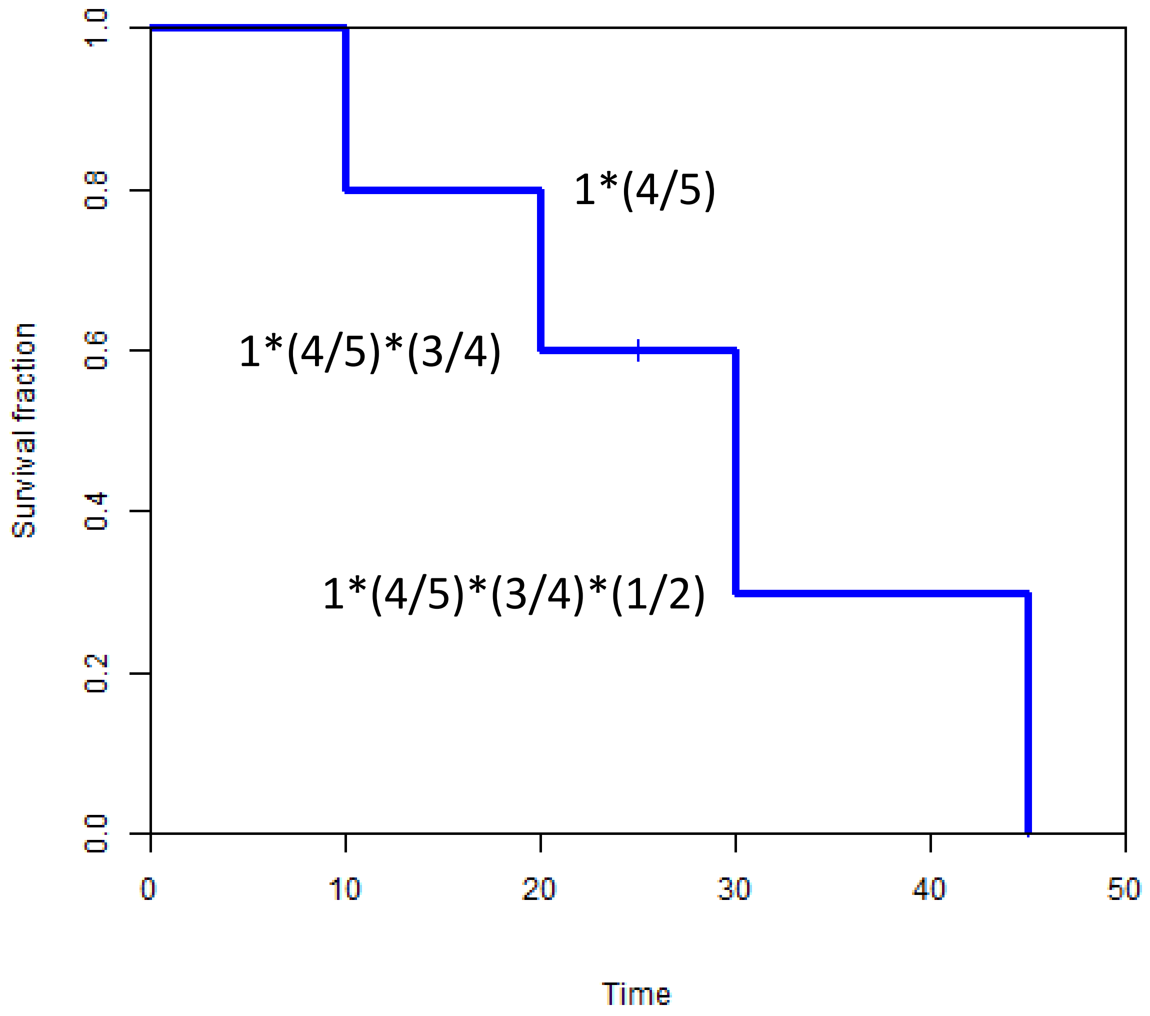
Why are time-to-event data special?

- Examples of time-to-event data:
  - Time to death (Survival)
  - Time to recurrence
  - Time to first myocardial infarction
  - Time to first alcoholic drink
  - Time to first marriage
  - … total reproduction?
  - … number of food pellets eaten?

- Partial information
- Need to incorporate that information into the analysis

Kaplan-Meier Plot

- Empirical, non-parametric estimate of the survival curve
- Example: [10, 20, 25+, 30, 45]

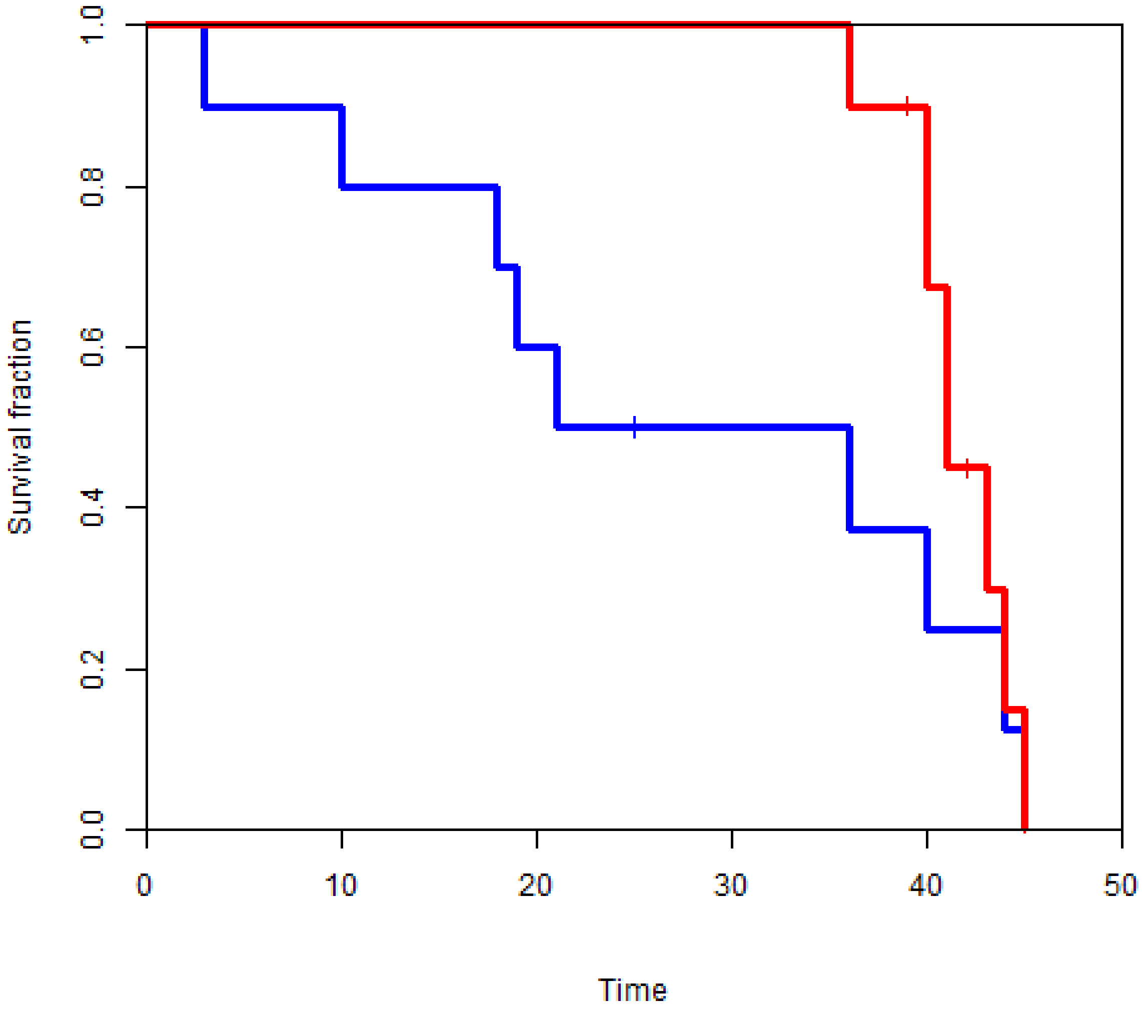


Non-Parametric Survival Tests

- Non-parametric comparison of KM survival curves

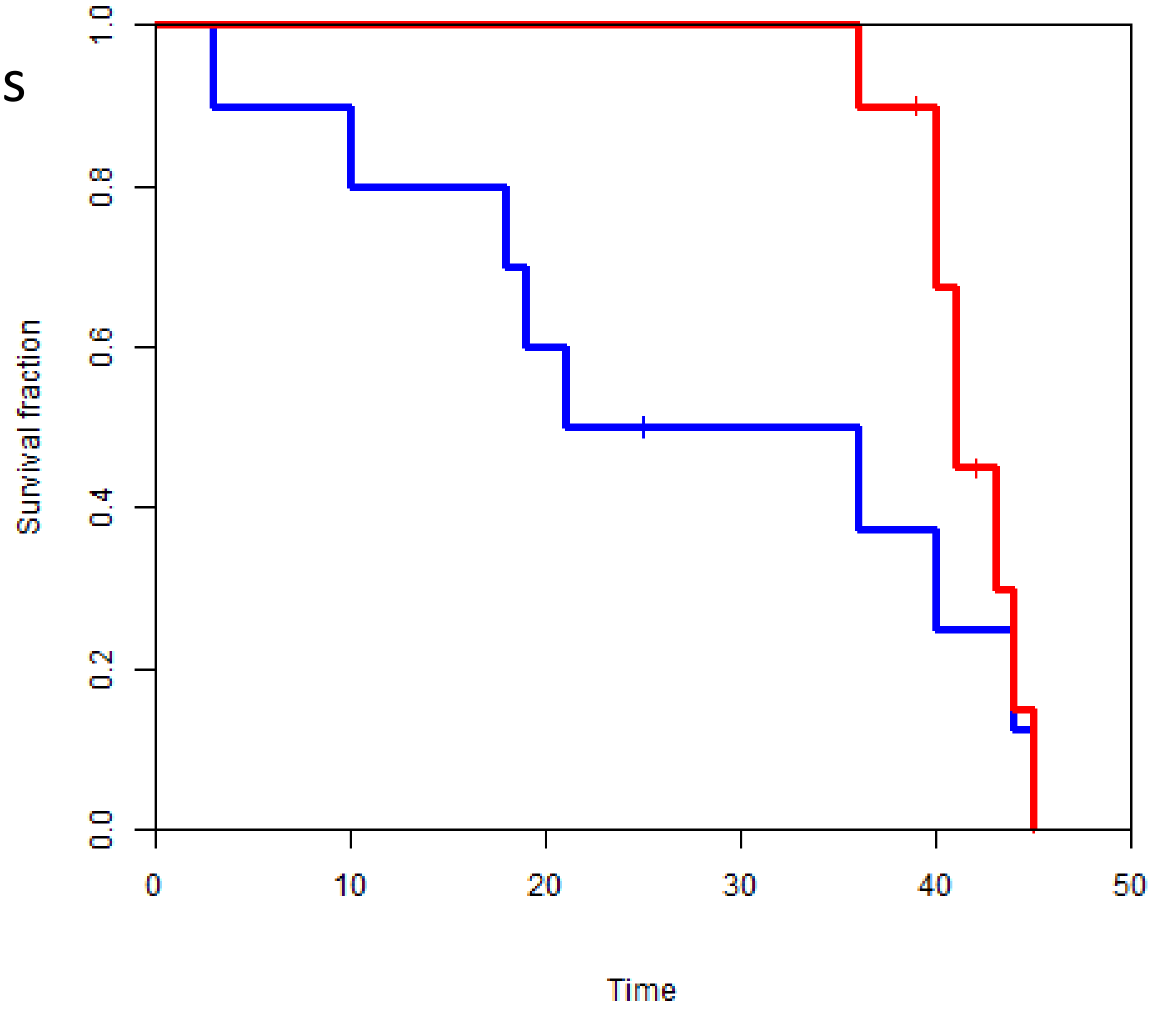
- Log Rank Test
- Mantel-Haenszel
- $\chi^2 = 2.11$ on 1 df
- $p = 0.146$

- Gehan-Wilcoxon
- Weighting
- $\chi^2 = 4.28$ on 1 df
- $p = 0.039$



Semi-Parametric Survival Tests

- Assuming proportional hazards (PH) model
- Can add covariates
- Cox PH
- β_{grp} = -0.63
- exp(β) = 0.53
- Z = -1.26
- p = 0.207



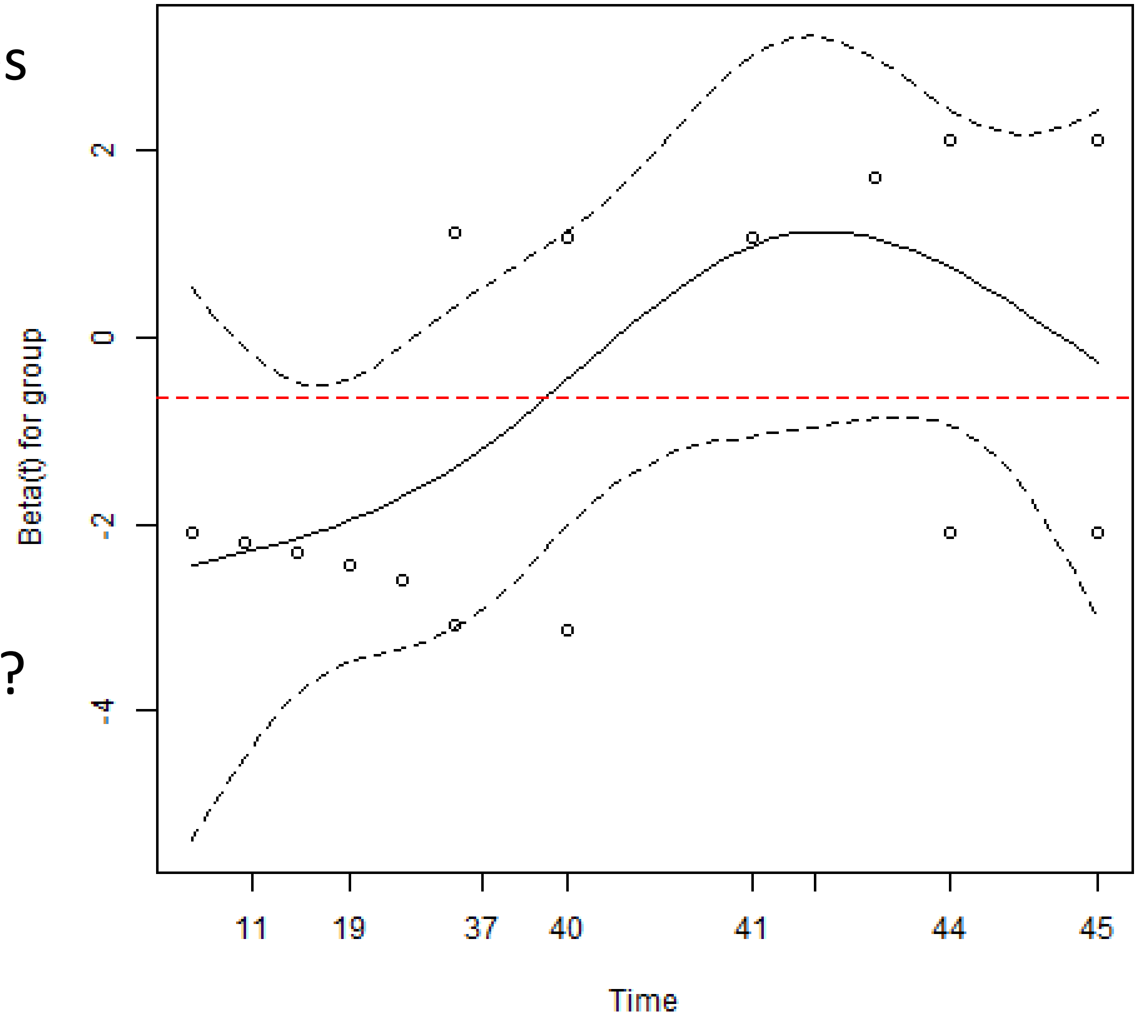
Semi-Parametric Survival Tests

- Assuming proportional hazards (PH) model

- Can add covariates

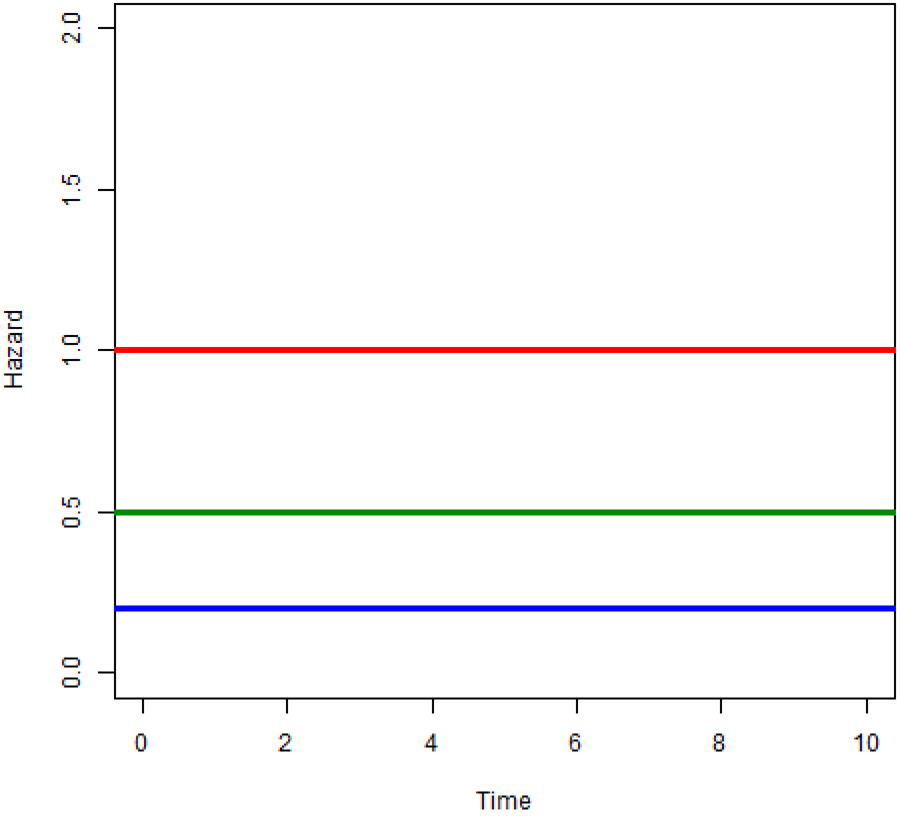
- Cox PH
- β_{grp} = -0.63
- exp(β) = 0.53
- Z = -1.26
- p = 0.207

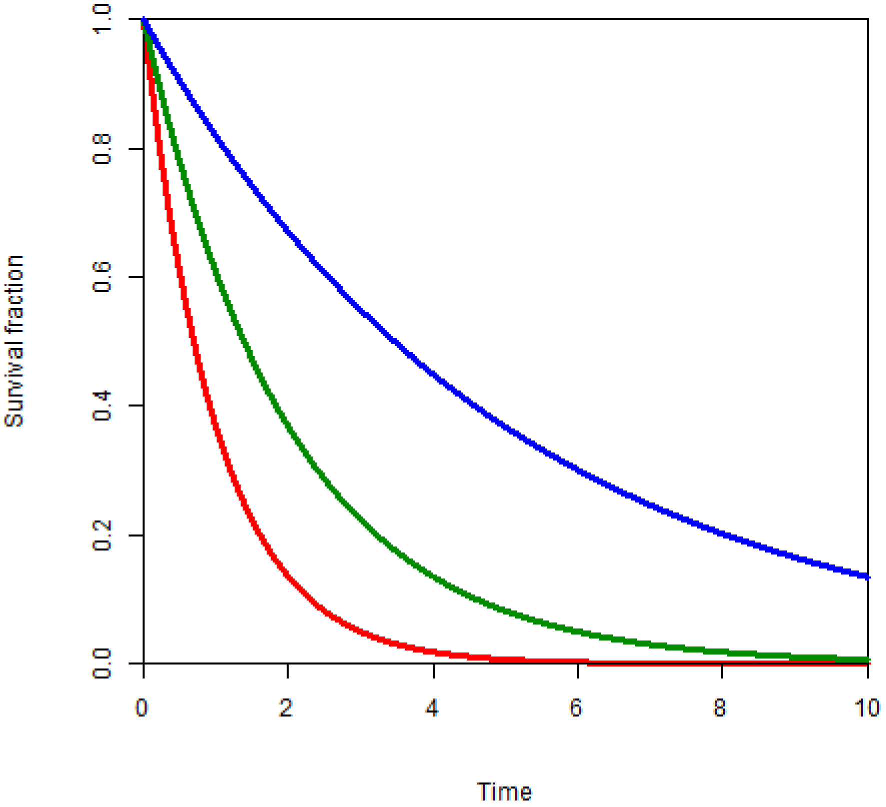
- PH not evidenced?
- GOF $\chi^2 = 4.01$
- p = 0.045



Parametric Survival Tests

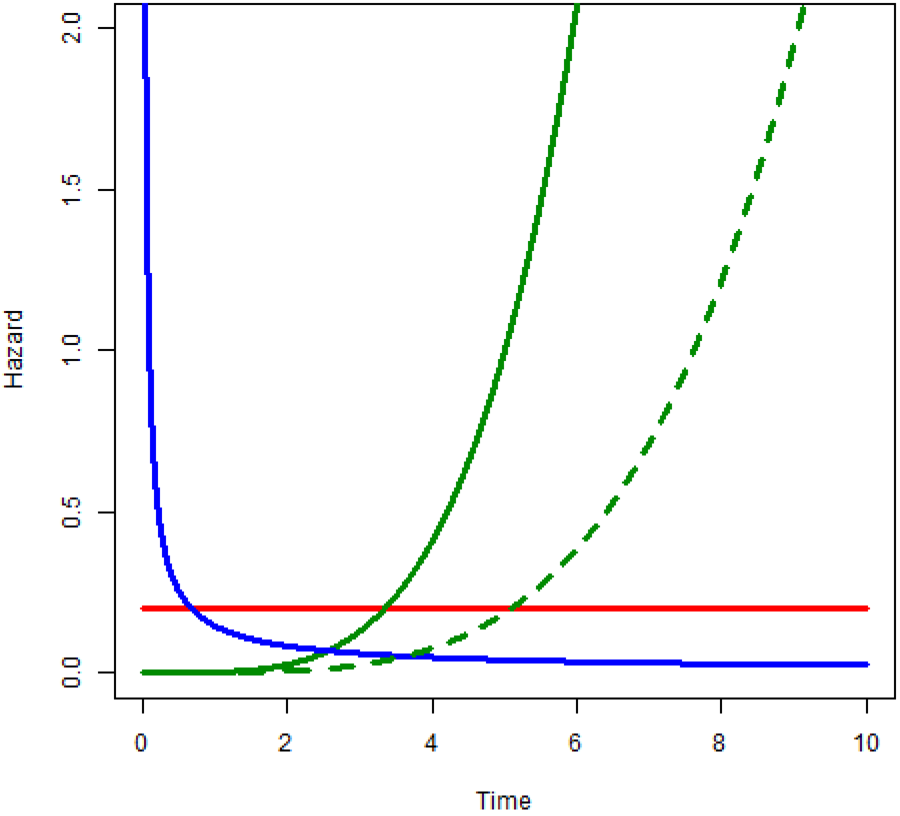
- Exponential
- One parameter family (intensity)
- ‘Light bulb’ survival
- Analysis: Set intensity = **Xβ**

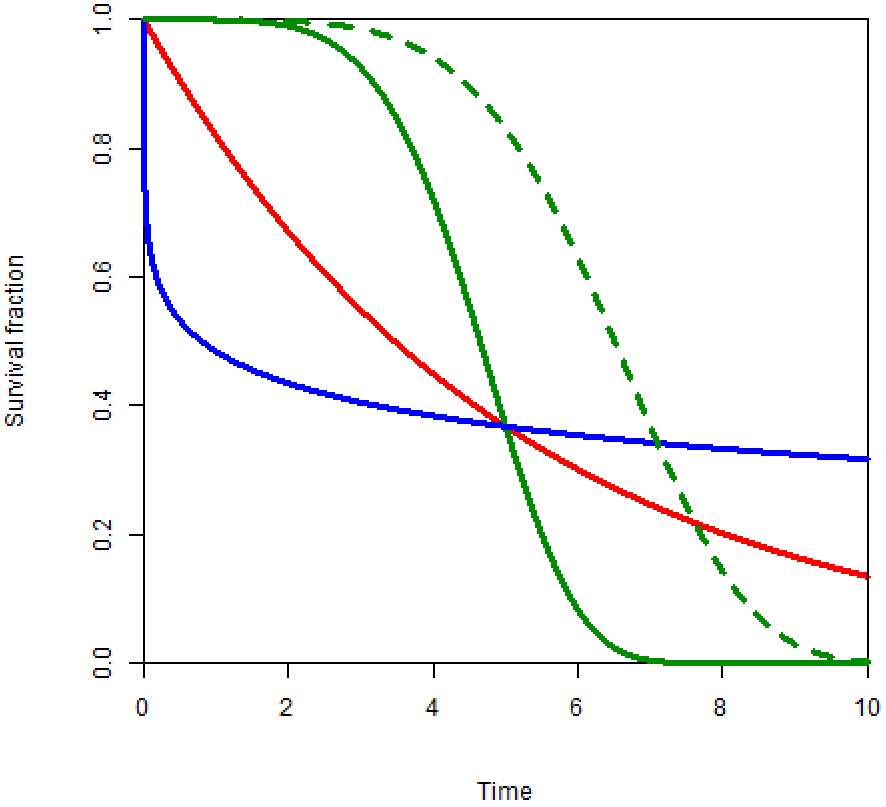




Parametric Survival Tests

- Weibull
- Two parameter family (shape and scale)
- Solid/Dashed: Scale = 5/7; Red/Green/Blue Shape = 1/5/0.2
- Analysis: Fix shape, set scale = **Xβ**



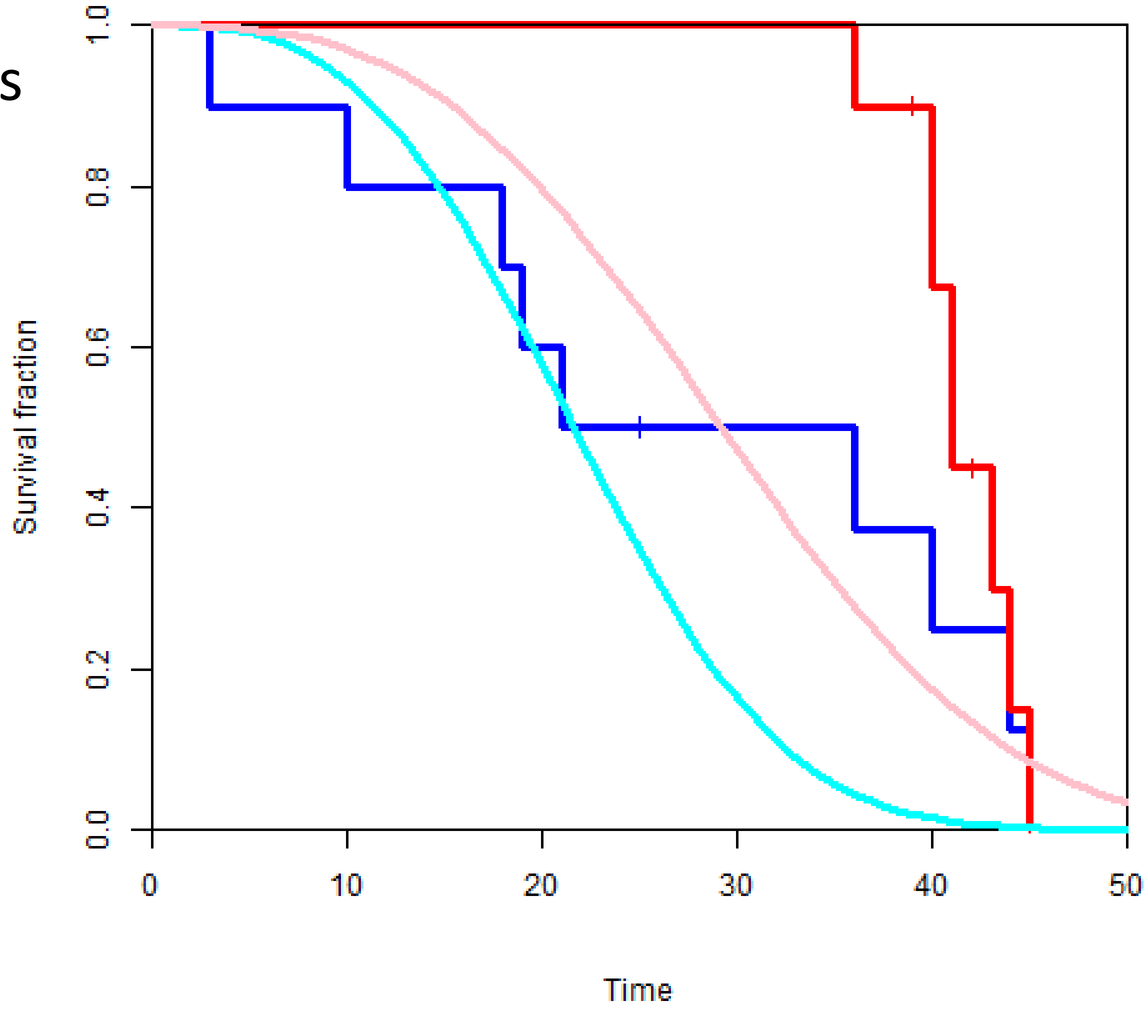


Parametric Survival Tests

- Assuming some parametric form for the hazard

- Can add covariates
- Common families:
  - Weibull
  - Lognormal
  - Gompertz

- Weibull fit:
- p = 0.082



Section Summary

- Account for censoring in the data
- Most people will use NP or SP methods, even if the underlying assumptions are not evidenced
- Parametric models must be rigorously checked
    - And then justified

- First step: Look at the (KM) survival curves

- Any time-to-event analysis that does not present KM estimates of the survival curves is not to be trusted. Period.

Thank You

Questions before the discussion?