

Applied Biostatistics

Raymond R. Balise Ph.D.

Stanford University

Department of Health Research and Policy



SPCTRM



Stanford / Packard Center for
Translational Research in Medicine

The Love Story!

- A daft biochemist invented a love potion and I talked about her quest to do science. Watch it online by following the links to my May 2006 Biostatistics 101 talk here:
clinicaltrials.stanford.edu/education/index.html
- I talked about general issues for statistical analyses in that lecture. This time I want to talk about the nuts and bolts of data analysis.
- I will talk about going from collection through publication on a couple of projects that I have been involved with.

What is data?

- All pieces of information that you collect and calculate as part of a study are data. Every person's response to a questionnaire is called a data point.
- There are two fundamentally different types of data: character and numeric.
 - Numeric data is always ... numeric. Information that *you could want to do math on* is numeric data.
 - Character data is alphanumeric. It includes the obvious things like names and addresses, but it also includes *numbers that you should not do math on*.
- Some systems make finer distinctions and let you set data so they are forced to be factors.

What is data coding?

- A question such as, “What is your current age in years?” is going to generate numeric data.
- A question such as, “At what age did you first contract a sexually transmitted disease?” is going to generate numeric data

But you are going to need to allow for the possibility that somebody has never contracted a sexually transmitted disease.

... and you always need to allow for people who never knew or do not remember information or who may be dishonest in their answers.

What is data coding? (2)

- When you have a question that generates numeric data and your subject's response is not a “real number” you can code a bogus value.
 - “Not applicable” can be coded as age -1.
 - “Do not know” can be coded as -2.
- The better way to deal with this problem is to use the value “NULL.”
 - Some data systems allow for different types of NULL.
 - Null values make your job easier when you try to do math on the values.

What is data coding? (3)

- Questions that generate alphanumeric data are always complex compared to numeric data.
- “Where were you born?” can be coded as a string of letters from a fill-in-the-blank question or coded as letters or numbers from multiple choice format.
 - Do not use null in fill-in-the-blanks.

What is a Data Dictionary?

- A data dictionary should provide a list of the following things:
 - The variables' names
 - A conceptual description of each variable
 - If the data is stored in character, numeric or some other form
 - The type of data
 - Permitted values
 - Codes for missing or impossible values
 - The source of the information: question(naire) number
- The dictionary may also include:
 - The position of the variables on the page
 - Logic checks that should be applied

Using Computers

- You can force hard error and range checking. Do you really want to do this?
- Print out results, ideally in a different format, immediately after the contact and review them with subjects.
- Collecting with paper and then doing electronic entry is a drag.....
- Back up, back up, back up.
 - Incremental, full and offsite!

Getting Data into a Computer

- What program should I use to gather and store data?
 - Excel is tolerable for toy data sets. Its security and auditing tools are primitive at best.
 - You can force it to only accept certain values.
 - There are good integrated data collection and storage systems.
 - For single investigator projects consider FileMaker Pro.
 - Access has another option but you are a slave to Microsoft's development patterns (patches and upgrades).

Using Excel

- If you don't have the ability to get access to a database and/or database programmer, you can fake it (sort of) with Excel.
 - Think of Excel as a tool of last resort, not a good idea...

Structuring Data in Excel

- Use column headings with no spaces or punctuation (especially leading or trailing spaces).
- Make only one record (row) for each observation on a subject.
- Do not put extra text in a cell.
- Put only one type of information in each column.
- Put only one type of information in each cell.
- Use a standard unit in each column.

Excel Bad and Good

One record for
each observation
remove extra text

| | A | B | |
|---|------|---------------|--|
| 1 | dude | Blood Presure | |
| 2 | | 1 s:120 | |
| 3 | | 1 d:80 | |
| 4 | | 2 s:160 | |
| 5 | | 2 d:85 | |
| 6 | | | |

| | A | B | C | D |
|---|------|-----------------|---------------|---|
| 1 | dude | sysBloodPresure | dBloodPresure | |
| 2 | 1 | 120 | 80 | |
| 3 | 2 | 160 | 85 | |
| 4 | | | | |

Use Validation

- The data menu has a validation option that lets you do things like date validation and enable pick lists for Excel cells.

| | A | |
|---|-----------------|--|
| 1 | Biopsy type | |
| 2 | P -- biopsy 3mo | |
| 3 | E -- biopsy 2mo | |
| 4 | P -- biopsy 1yr | |
| 5 | | |

Split this into 2 columns, use only months and force only legal values.

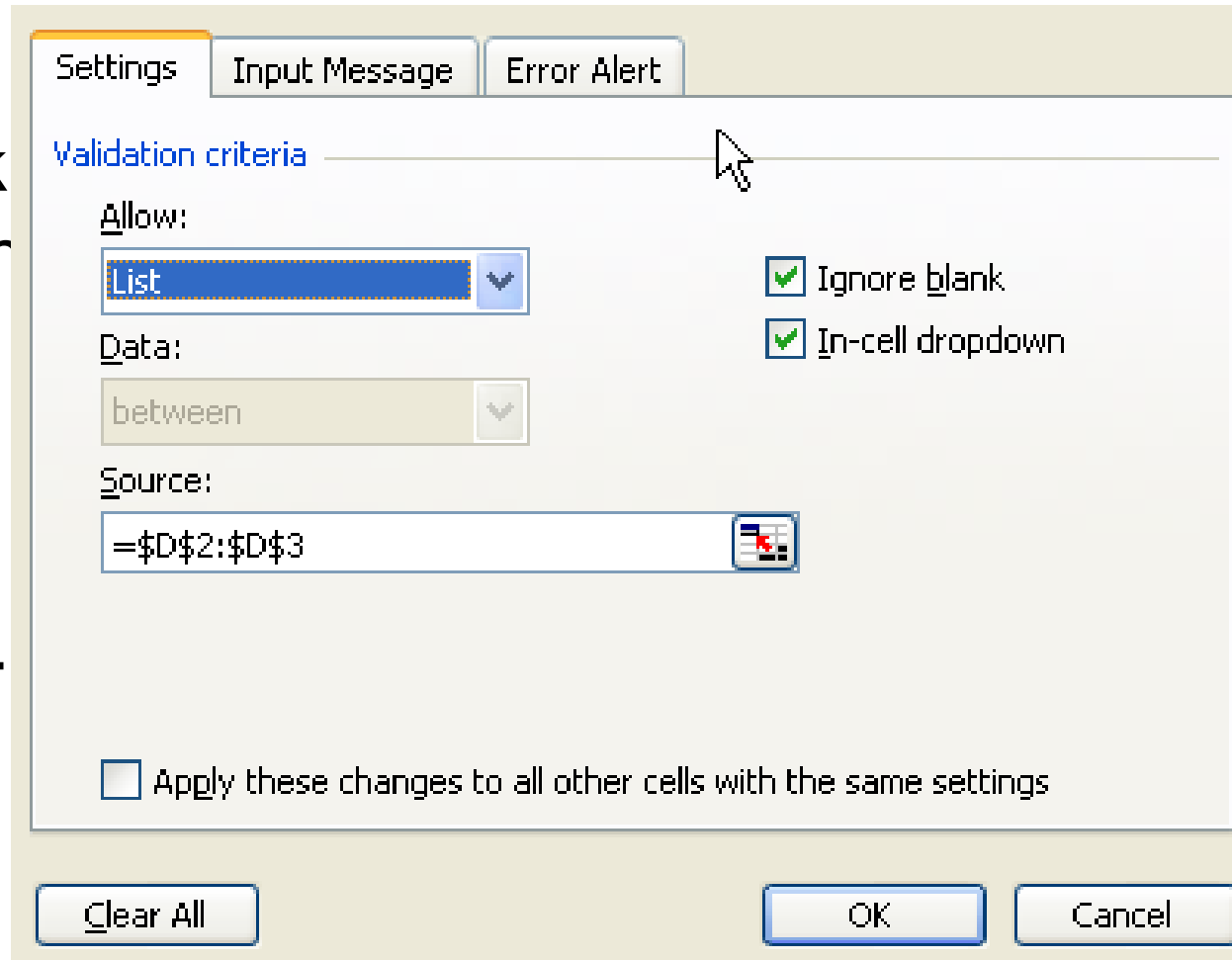
Making a Pick List

- Ideally you want to have a pick list and have a set of choices stored on another tab of the workbook but Microsoft is not that smart. Instead you have to put the choices on the same page (usually off to the right).

| | A | B | C | D | |
|---|-------------|-----------------|---|-----------|--|
| 1 | Biopsy type | biopsyAgeMonths | | | |
| 2 | | 3 | | Planned | |
| 3 | Planned | 2 | | Emergency | |
| 4 | Emergency | 12 | | | |

Use Validation for a Pick List

- Type the legal values for a pick list into a column of the spreadsheet
- Use validation... from the data menu.



Dates in Excel

- You can force Excel to only accept dates in a cell. Click on the column where the dates will be stored then use the validation option on the data menu.
- If you inherit dates typed by someone who was not taught about validation:
 - Check your dates by changing the format.
 - Check your dates by concatenating leading and trailing characters (to find blanks).

| | A | B | C | D | E |
|---|------------|---|---------------|-------------|-------------------------|
| 1 | Original | | Validation on | Reformatted | =x & CellNameHere & "x" |
| 2 | 6/24/1967 | | 6/24/1967 | 24-Jun | x24647x |
| 3 | 6/24/2206 | | | 24-Jun | x111940x |
| 4 | 60/24/1967 | | | 60/24/1967 | x60/24/1967x |
| 5 | 06/241967 | | | 06/241967 | x06/241967x |
| 6 | 1/18/1967 | | 1/18/1967 | 18-Jan | x24490x |
| 7 | 1/18/1967 | | | 1/18/1967 | x 1/18/1967x |

Notice dates are actually the number of days since 1900 (in Windows).

Settings Input Message Error Alert

Validation criteria

Allow: Date Ignore blank

Data: between

Start date: 1/1/1967

End date: 12/31/1967

Apply these changes to all other cells with the same settings

Clear All OK Cancel

Number Alignment Font Border Patterns Protection

Category: Date

Sample: Reformatted

Type:

- *3/14/2001
- *Wednesday, March 14, 2001
- 3/14
- 3/14/01
- 03/14/01
- 14-Mar
- 14-Mar-01

Locale (location): English (United States)

Date formats display date and time serial numbers as date values. Except for items that have an asterisk (*), applied formats do not switch date orders with the operating system.

OK Cancel

Data Containers vs. Data Analysis

- Database programs have, at best, very primitive tools for analysis.
- They can generate automatic reports for compliance and safety monitoring but for hardcore statistics you will need to work with a data analysis system.
 - SAS can talk directly to databases with ODBC
 - R, S-plus and SPSS typically work through an export

Types of Databases

- Any serious project will use a relational database.
 - The relational model says to store related bits of information together and avoid duplication at all costs.
 - The process of setting up relational tables is called normalization.
 - Put only 1 thing into each variable
 - You create tables which are linked by a unique identifier number. This is ideal in situations where people need to be blinded or where confidentiality is critical.
 - Confidential contact data goes in one place, appointments go into another table, blood in another, tissue another, etc..

Normalization

- In a normalized database there is always a variable (or set of variables) that can be used to link the tables, but other than that, the same information is not repeated in a table or across tables.
- This buys you lots of space and speed!

| ID | Sex | DAD ID | Address | Cancer | DX Date | Stage | Grade | Treat Date | H Name | H Address | Res |
|------|-----|--------|--------------|----------|---------|-------|-------|------------|--------|------------|-----|
| 1001 | M | -1 | 1313 Mocking | Lung | 1967 | 2 | 2a | Jan | VA | 234 Main | 10 |
| 1002 | M | 1 | 1721 Cedar | Prostate | 1994 | 2 | 3a | 15-Jun | Stan | 1 Pasture | 2 |
| 1002 | M | 1 | 1721 Cedar | Prostate | 1994 | 1 | 3a | 30-Jun | Stan | 1 Pasture | 2 |
| 1002 | M | 1 | 1721 Cedar | Prostate | 1994 | 1 | 3a | 15-Jul | Stan | 1 Pasture | 3 |
| 1002 | M | 1 | 1721 Cedar | Prostate | 1994 | 1 | 3a | 30-Jul | Stan | 1 Pasture | 4 |
| 1002 | M | 1 | 1721 Cedar | Colon | 1996 | 3 | 4 | July | Cal | 10 Medical | 10 |

A Normalized Database

ID is a key joining the master and cancer tables.

| Master Table |
|--------------|
| ID |
| 1001 |
| 1002 |
| 1003 |
| 1004 |

ID is a key joining the master and dude tables.

| Cancer Table | | | | | |
|--------------|---------|----------|----------|-------|-------|
| ID | DX Date | Site | Icd9Code | Stage | Grade |
| 1001 | 1967 | Lung | 12.9 | 2 | 2a |
| 1002 | 1994 | Prostate | 1096 | 2 | 3a |
| 1002 | 1996 | Colon | 1685.2 | 3 | 4 |

| Dude Table | | | | | | | | |
|------------|--------|-------|------------------|------|-------|-----|-------|-----|
| ID | DAD ID | Name | Address | City | State | Zip | Phone | SSN |
| 1001 | -1 | Bob | 1313 Mockingbird | | | | | |
| 1002 | 1 | JoBob | 1721 Cedar | | | | | |

ID and dxDate make a joint key.

| Response Table | | | | |
|----------------|---------|--------|------------|-----|
| ID | DX Date | H Name | Treat Date | Res |
| 1001 | 1967 | VA | Jan | 10 |
| 1002 | 1994 | Stan | 15-Jun | 2 |
| 1002 | 1994 | Stan | 30-Jun | 2 |
| 1002 | 1994 | Stan | 15-Jul | 3 |
| 1002 | 1994 | Stan | 30-Jul | 4 |
| 1002 | 1996 | Cal | July | 10 |

hName is a key relating the treatment and hospital tables.

| H Name | H Address | City | State | Zip | Phone |
|--------|---------------|------|-------|-----|-------|
| VA | 234 West Main | | | | |
| Stan | 1 Pasture | | | | |
| Cal | 10 Medical | | | | |

Language of Databases

- Any database will use a language called SQL (pronounced S–Q–L or sequel).
 - Asking for information is called querying the database.
 - QBE – query by example is one fairly friendly way to get data out.
 - The ‘real’ way to get data out is to write SQL commands.
 - SQL is deceptively easy to learn to use.
 - Be careful because it is hard to know when you are writing perfectly formed but illogical queries.
 - Tables are related by joins.
 - Inner, outer, left, right, one to one, one to many, many to many

Univariate Descriptive Stats

- Do descriptive statistics often!
 - Once you have code written, just rerun it.
 - Look at the descriptive statistics on ALL of your variables EVERY WEEK during data collection. If you have too many to look at, you have too many.
 - Make someone accountable for spotting problems on each and every variable.
- Look at graphical displays of your data often!

Make Table 1 on Day 1

- Every paper has a descriptive statistics table on the matching, explanatory and outcome variables. Build that table and rerun the code OFTEN.
 - Be sure to have a secondary table 1 that covers all your variables with measures of central tendency and dispersion.

Make Graphics of ALL Your Variables

- Do dot plots for counts on categorical variables.
- For your continuous variables make histogram and box plots (and Q or QQ plots). Look for outliers and departures from the expected shape of the distributions (is the data well described by a bell shaped curve?).

Dude, where's my data?

- Make a flowchart showing when people drop out of the study.
 - Update this constantly.
- If you think you know when people drop out, verify each variable at each time point. Check the data vs. who should be there and who should be there vs. the data.
 - This is in addition to logic checks for patterns of missingness (generally, men don't have labor and delivery records).

Missing Data Summary Table

Lipid table

Who has Lipid data that does not agree with enrollment

| Obs | ID # | BV) In study at baseline | at8w) In study at 8 weeks | at6m) In study at 6 months | at12m) In study at 12 months | Has baseline lipid | Has 8 week lipid | Has 6 month lipid | Has 12 month lipid |
|-----|------|--------------------------|---------------------------|----------------------------|------------------------------|--------------------|------------------|-------------------|--------------------|
| 1 | A001 | 1) Yes | 1) Yes | 0) No | 0) No | MISSING | MISSING | MISSING | MISSING |
| 2 | A002 | 1) Yes | 0) No | 0) No | 0) No | MISSING | MISSING | MISSING | MISSING |
| 3 | A003 | 1) Yes | 0) No | 0) No | 0) No | MISSING | MISSING | MISSING | MISSING |
| 4 | A006 | 1) Yes | 1) Yes | 1) Yes | 1) Yes | 0) No | 1) Yes | 1) Yes | 1) Yes |
| 5 | A009 | 1) Yes | 1) Yes | 1) Yes | 1) Yes | 1) Yes | 0) No | 1) Yes | 1) Yes |
| 6 | A029 | 1) Yes | 1) Yes | 1) Yes | 1) Yes | 1) Yes | 0) No | 1) Yes | 1) Yes |
| 7 | A035 | 1) Yes | 0) No | 0) No | 0) No | MISSING | MISSING | MISSING | MISSING |
| 8 | A038 | 1) Yes | 0) No | 0) No | 0) No | MISSING | MISSING | MISSING | MISSING |
| 9 | A056 | 1) Yes | 1) Yes | 0) No | 0) No | MISSING | MISSING | MISSING | MISSING |
| 10 | A061 | 1) Yes | 1) Yes | 0) No | 0) No | MISSING | MISSING | MISSING | MISSING |
| 11 | A073 | 1) Yes | 0) No | 0) No | 0) No | MISSING | MISSING | MISSING | MISSING |
| 12 | A077 | 1) Yes | 1) Yes | 0) No | 0) No | MISSING | MISSING | MISSING | MISSING |
| 13 | A078 | 1) Yes | 1) Yes | 0) No | 0) No | MISSING | MISSING | MISSING | MISSING |
| 14 | A082 | 1) Yes | 1) Yes | 1) Yes | 1) Yes | 1) Yes | 1) Yes | 1) Yes | 0) No |
| 15 | A085 | 1) Yes | 1) Yes | 1) Yes | 1) Yes | 1) Yes | 0) No | 1) Yes | 1) Yes |
| 16 | A089 | 1) Yes | 1) Yes | 0) No | 0) No | MISSING | MISSING | MISSING | MISSING |
| 17 | B001 | 1) Yes | 1) Yes | 1) Yes | 1) Yes | 1) Yes | 1) Yes | 0) No | 1) Yes |

Data Cleanup

- The first rule of cleaning up data is to back up your data before you start. The second rule is to do nothing by hand. Your goal should be to write a program that does **EVERYTHING** for you.
- You want this because:
 - Your data can and most likely will be damaged (corrupted).
 - You want to know who is responsible when things go wrong.
 - Some projects, like clinical trials, have mandatory audit trails.

How to Do Analyses

- There are two main programs used for biostatistical analyses:
 - R (or its expensive brother S-Plus)
 - Extremely popular with academics
 - Better for data exploration
 - SAS
 - Dominates the market place
 - Better for data management
- The most user-friendly package for biostatistics was designed for the social sciences... SPSS.

Case Study

- The good idea: conduct a pilot study of several proteins that are thought to predict preeclampsia. Women with high risk and controls were selected and followed over their pregnancy.
 - Is there a difference in the average values at each of the time points for women who do and do not get preeclampsia?
 - Can you detect who will be sick later in the hope of providing services to protect the mother and developing child?

Theoretical Issues (and English Translations)

1) Correlated factors

- There are homeostatic feedback loops which (if working at all correctly) force limits on the values. Think about measuring an enzyme and its precursor and its metabolite.

2) Pseudo-replication and variance

- The number of data points you have does not equal the number of people studied. So, any guess at the variability in the population will be tricky.

3) Distribution of the predictors will probably not be normal.

- Will the enzymes be clustered around a mean and with an easy to describe range? Can the data be pictured as a bell shaped curve?

Correlated Factors

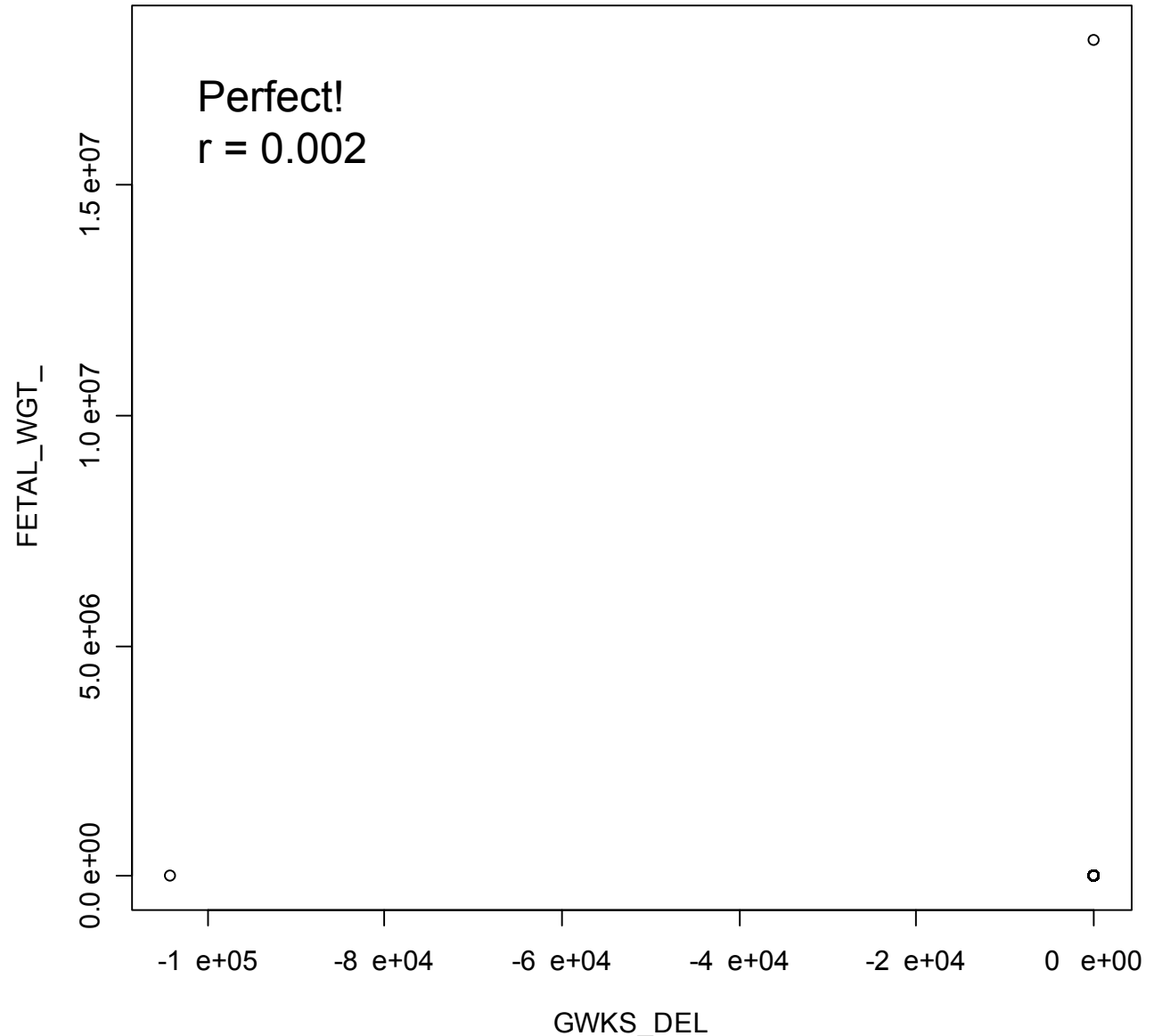
- In the ideal case, you build a model which gives you a probability of disease when it is given a set of values on predictor variables.
 - For every one unit increase in the predictors, your risk goes up or down.
 - You want to build a parsimonious model (only include things that are important). Including the same predictor twice in a model is clearly not a smart idea....

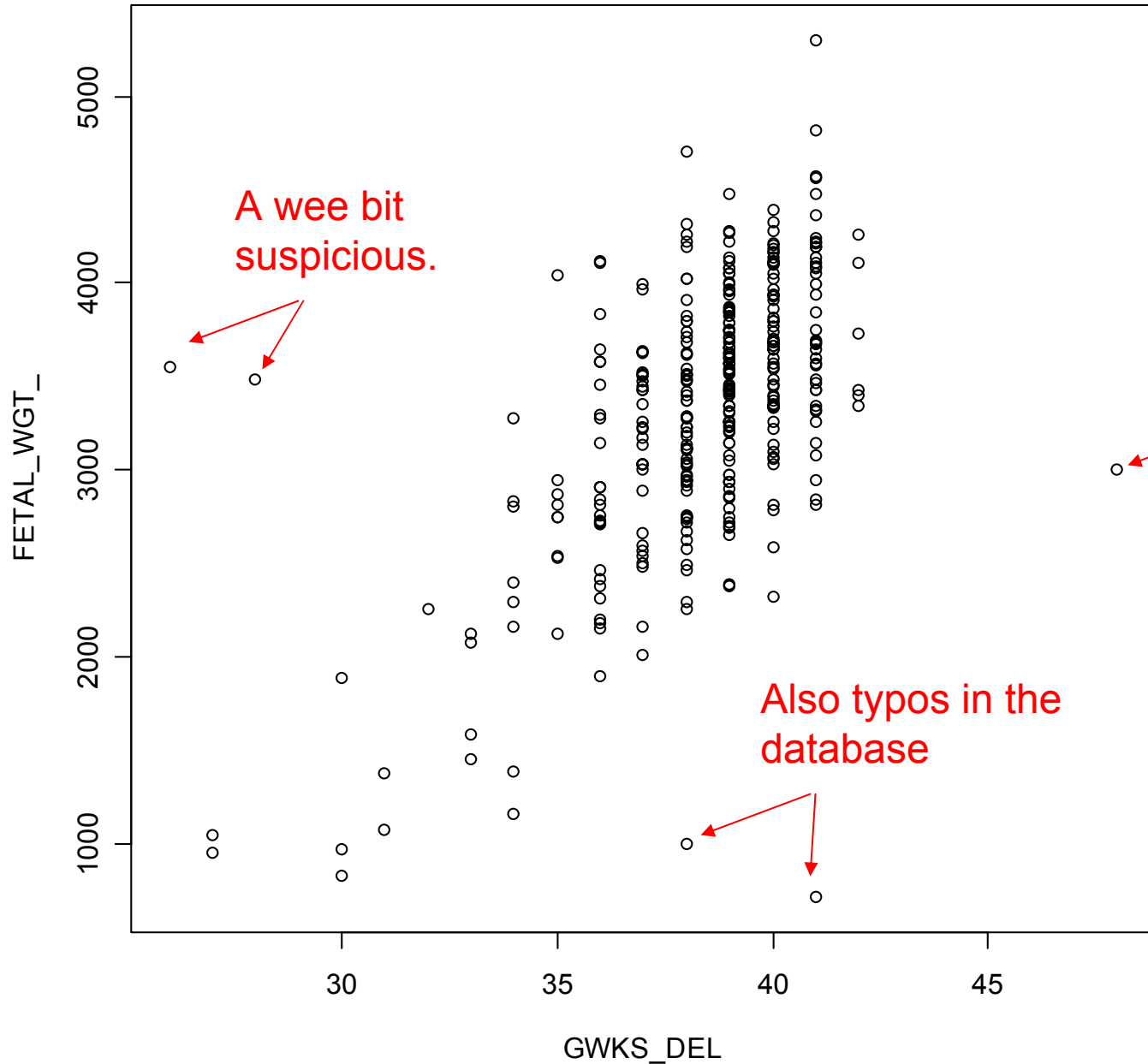
Correlated Factors₍₂₎

- Say you have an excellent predictor of weight at birth, such as gestational age in weeks or gestational age in days. What happens when you include both correlated predictors in a model?
 - When you include two (or more) correlated predictors, either of which predicts that outcome very well, those good predictors may no longer show a strong relation to the outcome!

Correlation between Age and Weight

- Step 1 is always to visualize your data
- Step 2 fix the non-human data





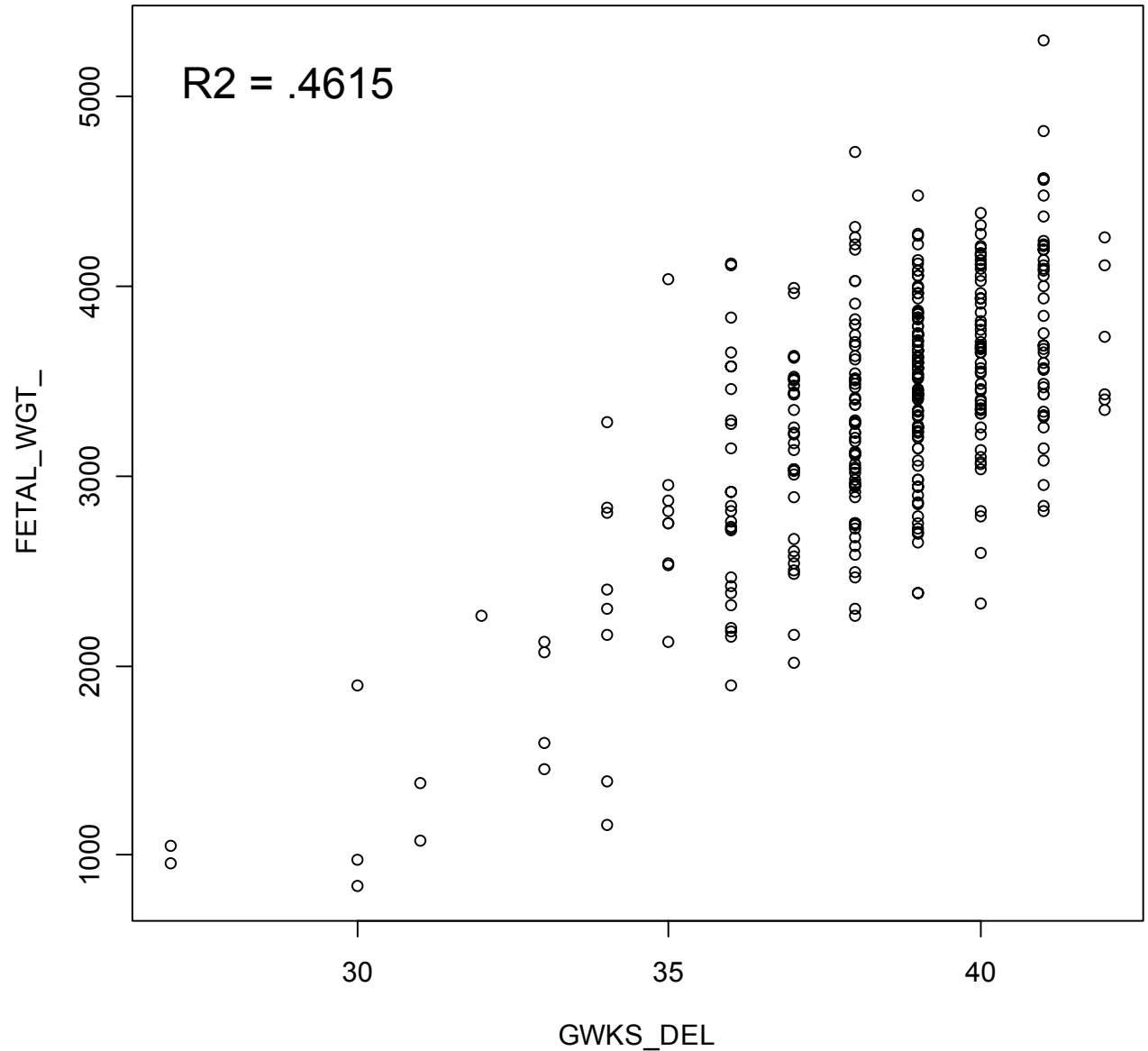
A wee bit suspicious.

Also typos in the database

Think she was grumpy at week 49?

The real data

- There is a clear association. Weight goes up with gestational age in weeks.



A Linear Model

Microscopic p-value
couldn't get these values
by chance if the true values
are 0!

| | Estimate | P-value < |
|-------------|----------|---------------------|
| (Intercept) | -4545.78 | .000000000000000002 |
| GWKS_DEL | 205.16 | .000000000000000002 |

- A baby at 0 weeks gestation weighs in at -4545 grams
- For every 1 unit increase in gestational age (1 week) the average babies weight goes up by 205 grams.

- For a 40 week old baby the model says:

$$\text{weight} = -4545 + 205.16 * 40$$

or on average a 40 week baby should weigh 3661 grams

Correlated Predictors

- What happens when you use two correlated indicators to predict weight? Say your faithful research assistant, Igor, modeled weight with number of days since last menstrual period and also number of weeks...

A Linear Model

| | Estimate | P-Value < |
|-------------|-----------|---------------------|
| (Intercept) | -4557.381 | .000000000000000002 |
| GWKS_DEL | 155.303 | 0.111 |
| days | 7.069 | 0.605 |

The p-value for gestational age is not statistically significant.

- A baby at 0 weeks gestation weighs in at -4557 grams
- For every 1 unit increase in gestational weeks (1 week) the average babies weight goes up by 155 grams.
- For every 1 unit increase in gestational days (1 day) the average babies weight goes up by 7 grams.
- For a 40 week old baby the model says:

$$\text{weight} = -4557 + 155 * 40 + 7 * 280$$

or on average a 40 week baby should weigh 3634 grams

Be VERY Careful!

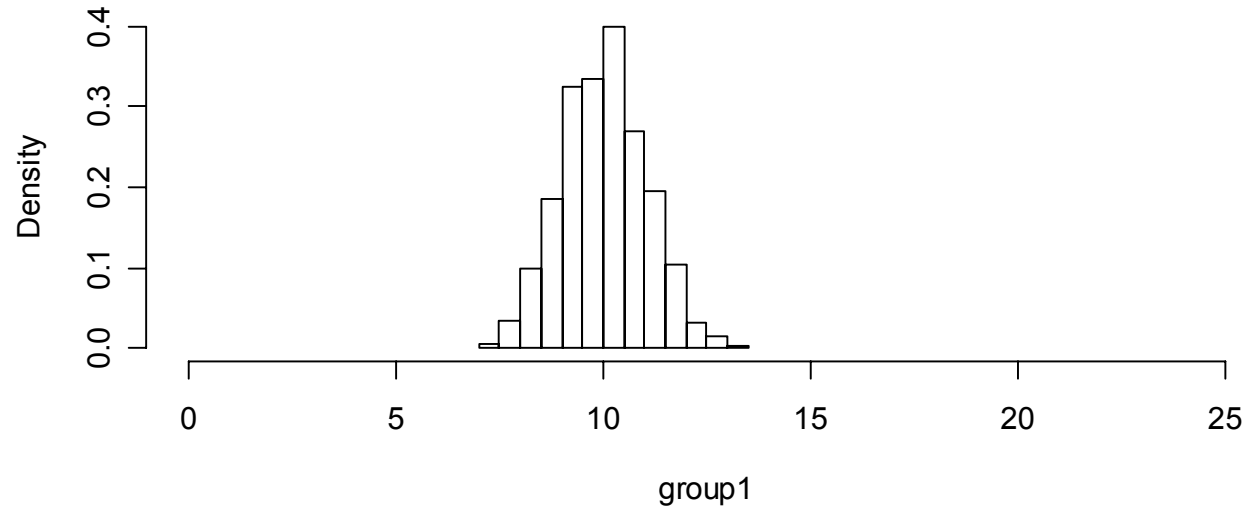
- These correlated predictors are crazy but what about hormones in a homeostatic feedback loop? Can you find each one to be a significant predictor?
 - Explore variance inflation statistics.
 - Look at sequential vs. simultaneous predictors in the ANOVA model.

Pseudo-replication and Variance

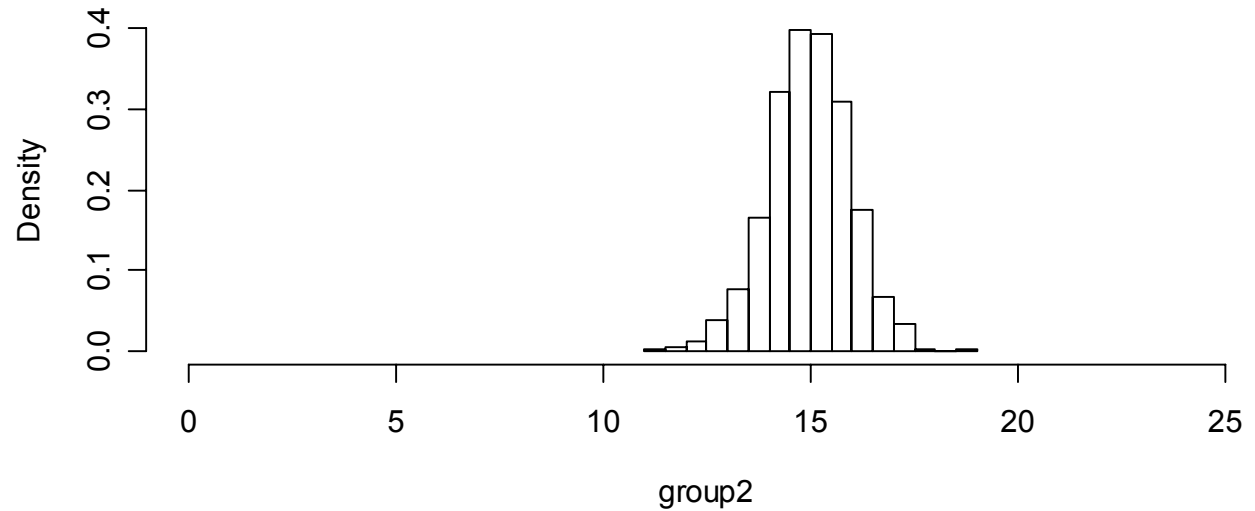
- Most of the statistics you learn about in a basic statistics course and introductory books are parametric.
 - They assume the data can be well described with a couple of numbers, typically the mean and standard deviation.
 - The standard deviation from the population is estimated by sample standard deviation.
 - When you measure the same person over and over you incorrectly specify the population variance!

Parametric means

Histogram of group1

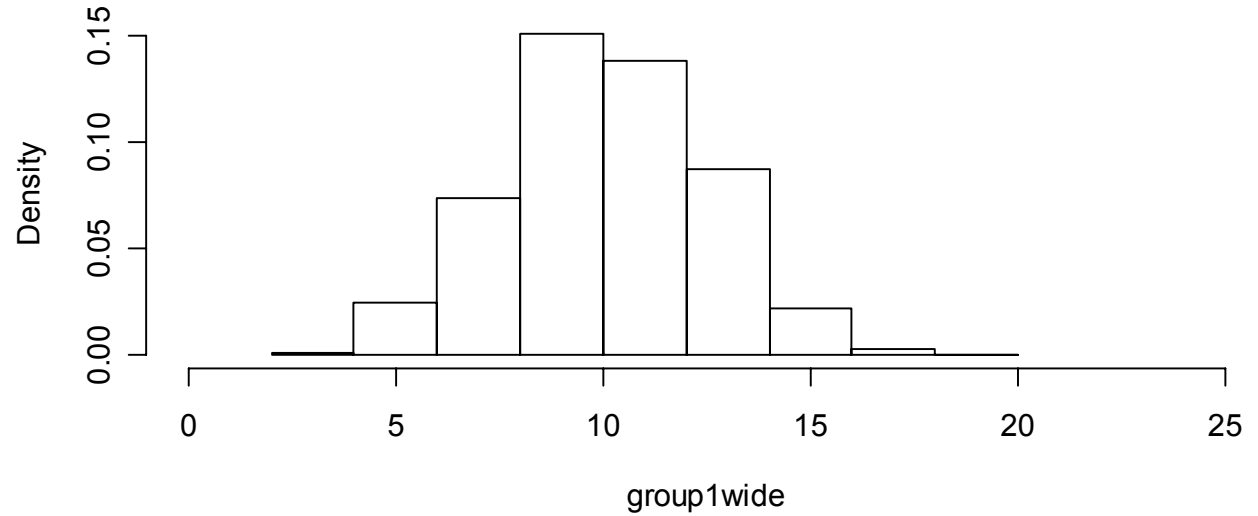


Histogram of group2

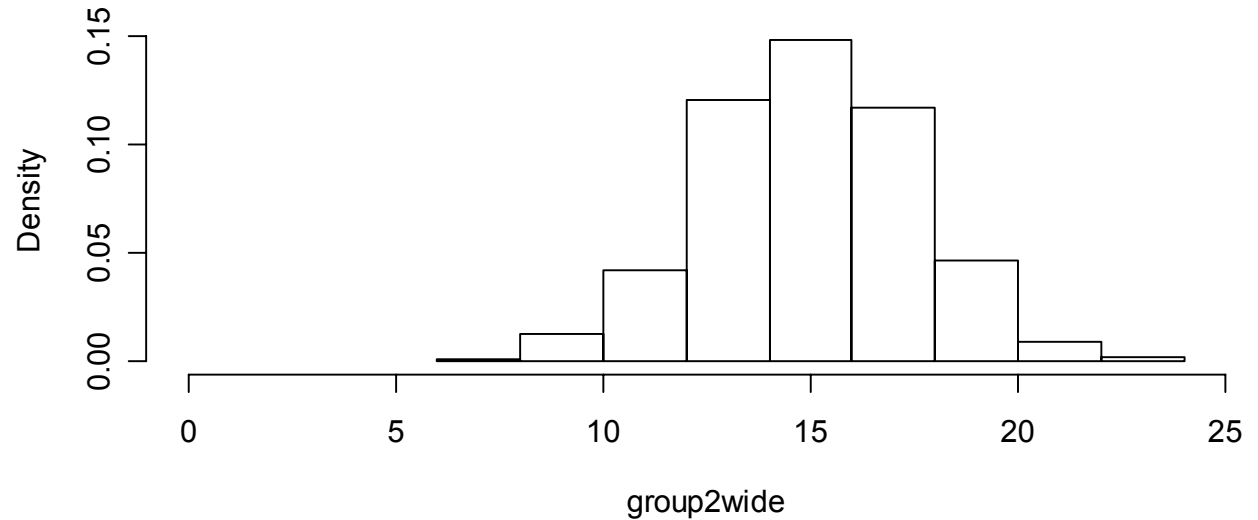


Parametric means

Histogram of group1wide



Histogram of group2wide



Dealing with Repeated Measures

- You can analyze each time point as if it were independent.
 - This means you are losing information in your estimate of the population variance.
- You can do a repeated measures (mixed effects model).
 - This means you are using a statistician to help you get the model specified correctly.

Distribution of the predictors will probably not be normal.

- Many things in nature are well described by a normal distribution. What you want to be extremely careful of is biomarkers that take on a meaningful score of zero and also markers that can take on extreme values.

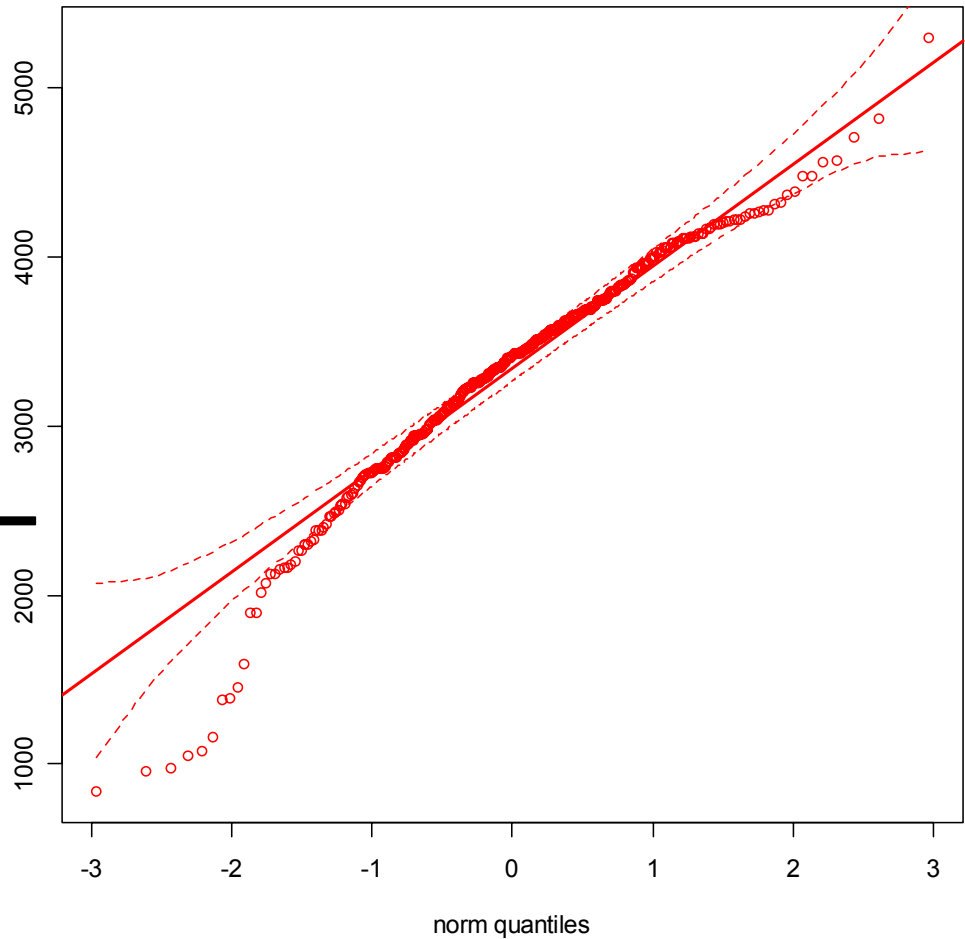
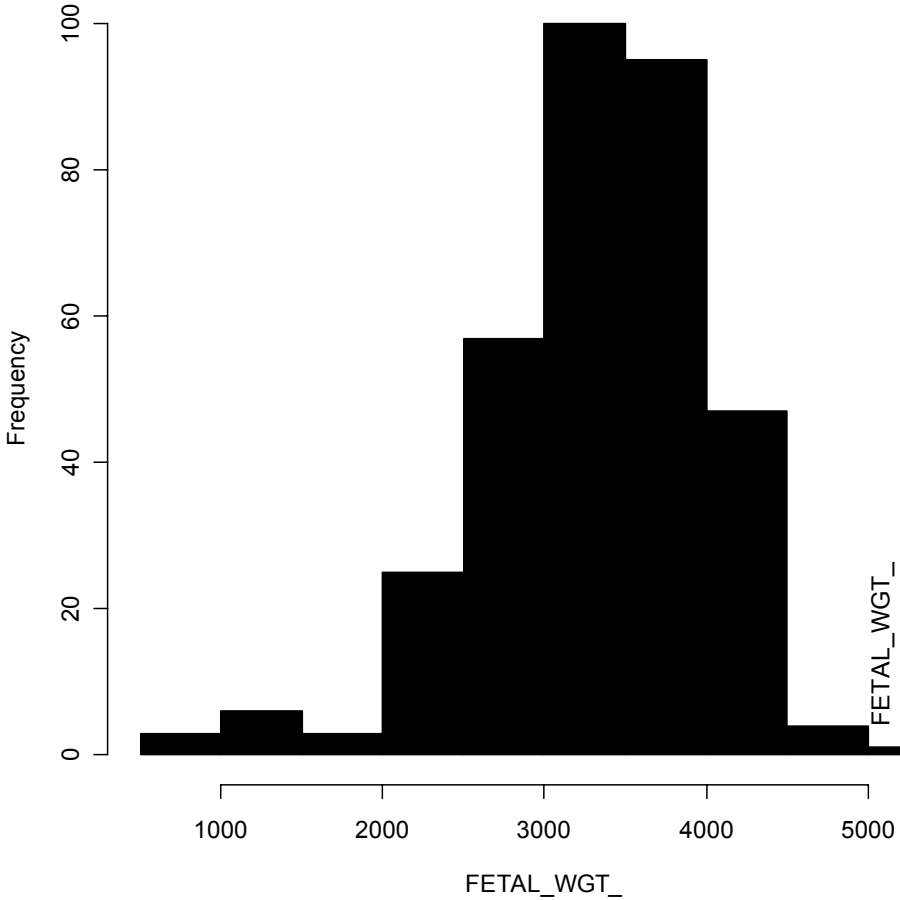
Checking for Normality

- Most people start by doing a histogram and they eyeball what a normal curve would look like.
- Instead, consider doing a graphical or numeric test for normality directly.
 - The graphical method is a QQ plot.
 - There are several statistical tests to see if a variable is normally distributed. Consider Shapiro-Wilk.

QQ plots

Checking for Normality

Histogram of FETAL_WGT_



Transforming

- When in doubt about how to transform, start at one extreme (reciprocal square) and move to the other extreme (square) or use a systematic technique like Box-Cox transforms....
- Be careful if you have negative and positive values.

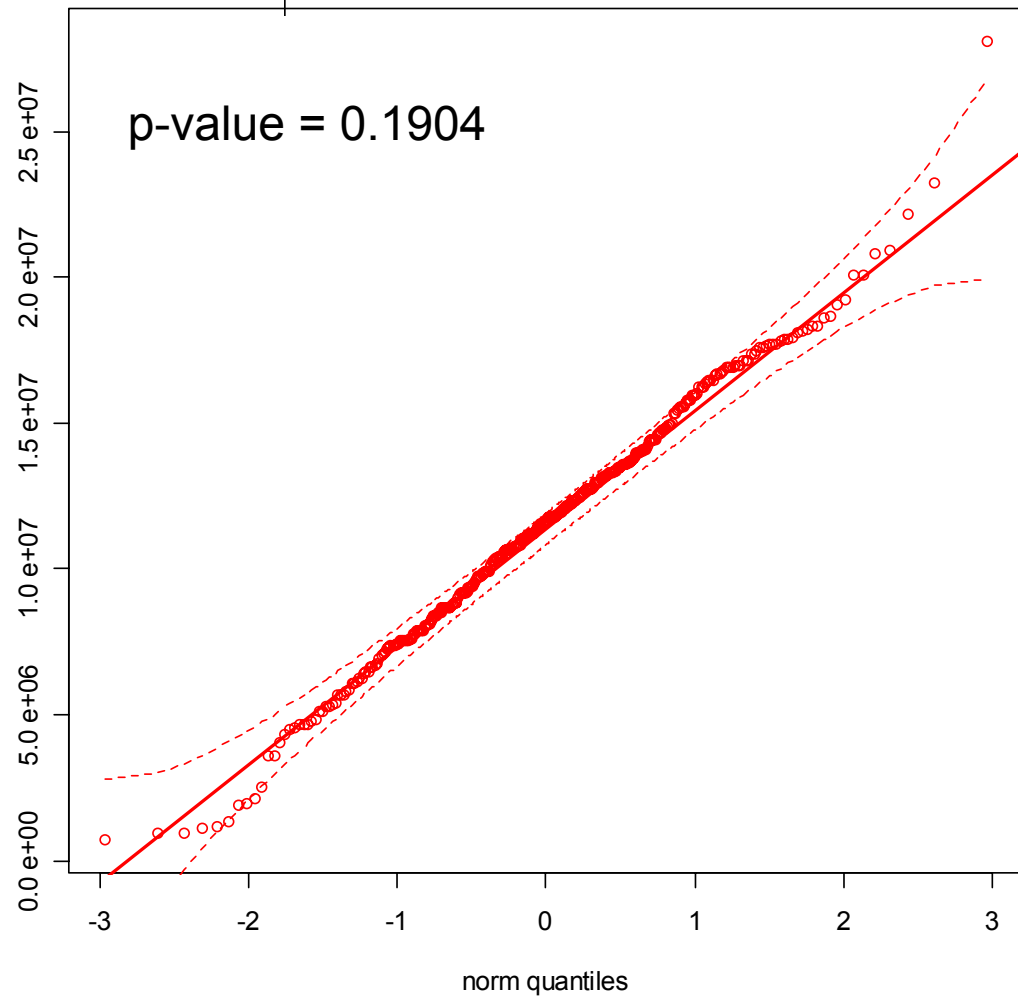
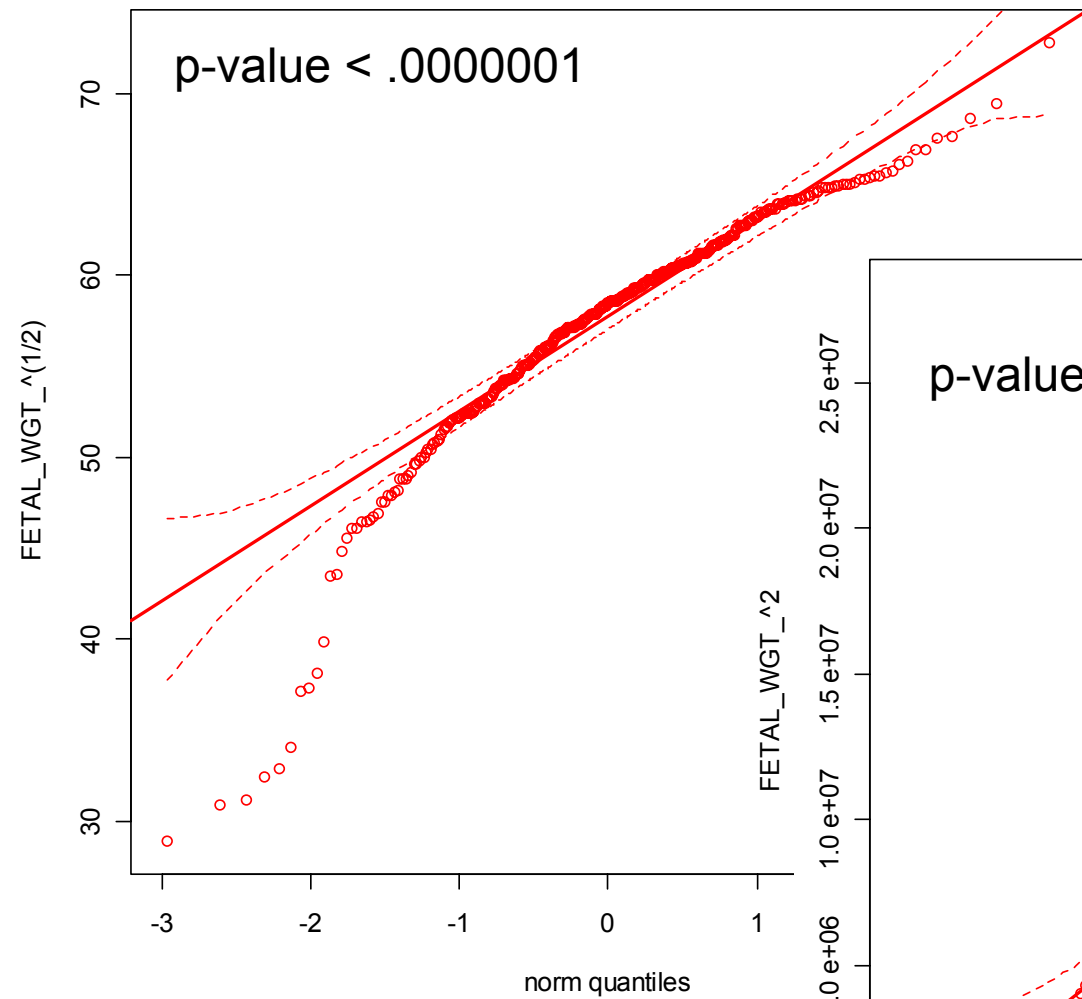
$$y \rightarrow y^P$$

| P | New | Name |
|--------|---------------|-------------------|
| 3 | y^3 | Cube |
| 2 | y^2 | Square |
| 1 | y^1 | original |
| 0 | Log(y) | logarithm |
| $-1/2$ | $-1/\sqrt{y}$ | Reciprocal root |
| -1 | $1/y$ | Reciprocal |
| -2 | $1/y^2$ | Reciprocal square |



compress
lower tail and
expand upper
tail

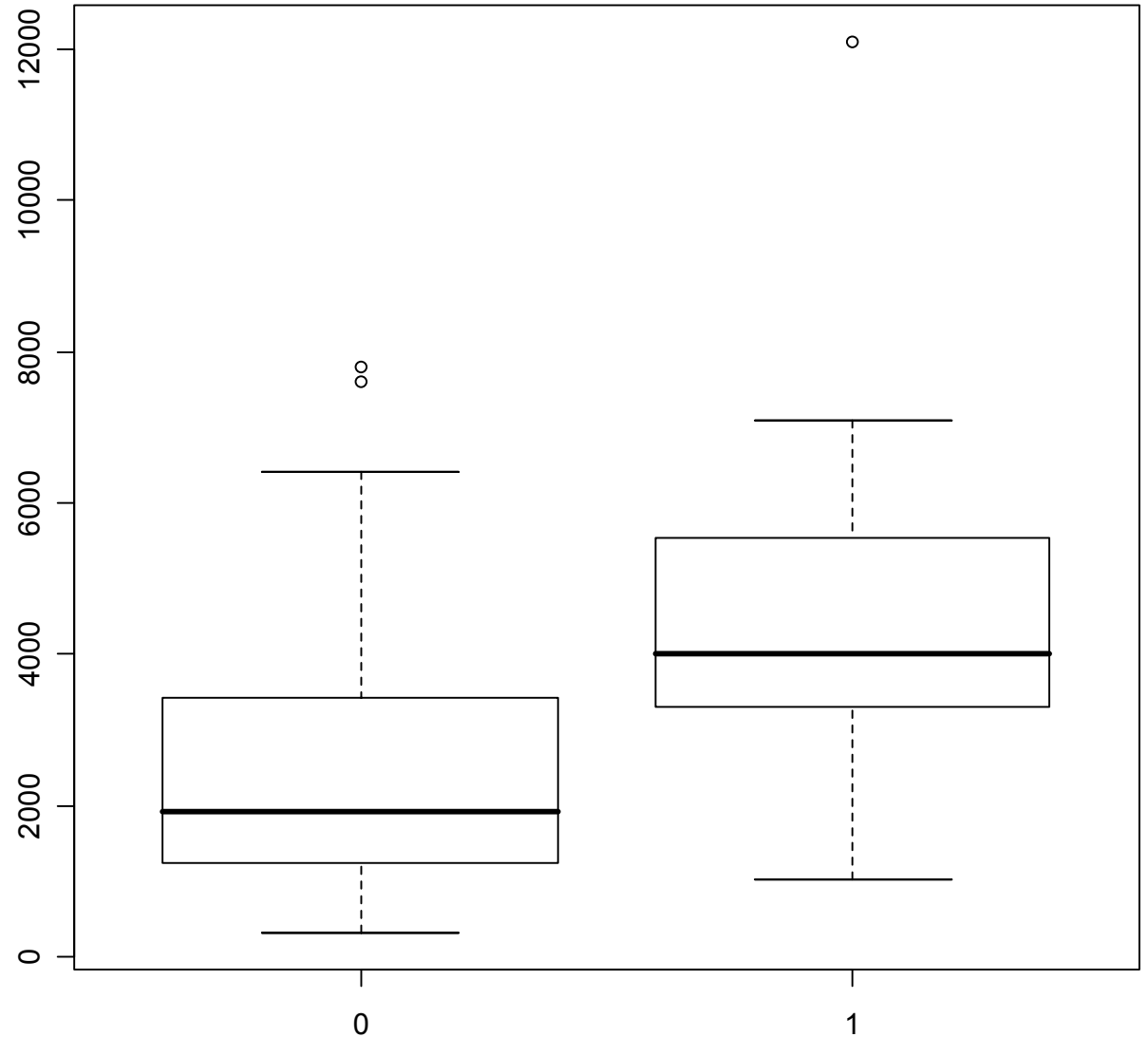
expand lower
tail and
compress
upper tail



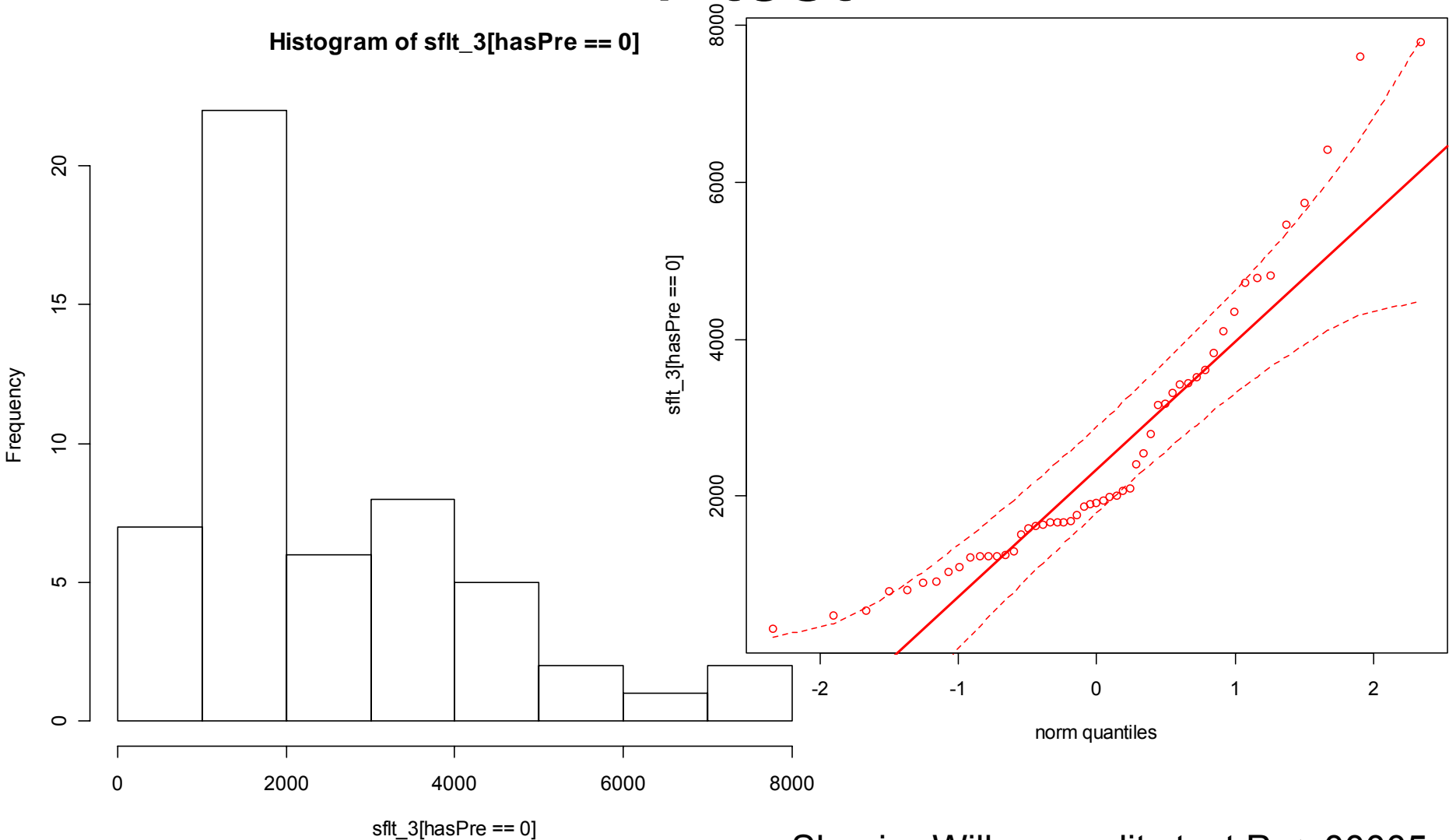
Meanwhile back at the enzymes...

- One of the enzymes Sflit has been shown to be related to preclampsia at time point 4. To examine this, we want to compare the distribution (think histogram when someone says distribution) of the women with and without preclampsia. We expect the center of the distribution to be lower for the women with disease. Take a look before thinking about math...

- It looks like something interesting is going on but look at those extreme values.



Check for Normality Before Doing a T-test

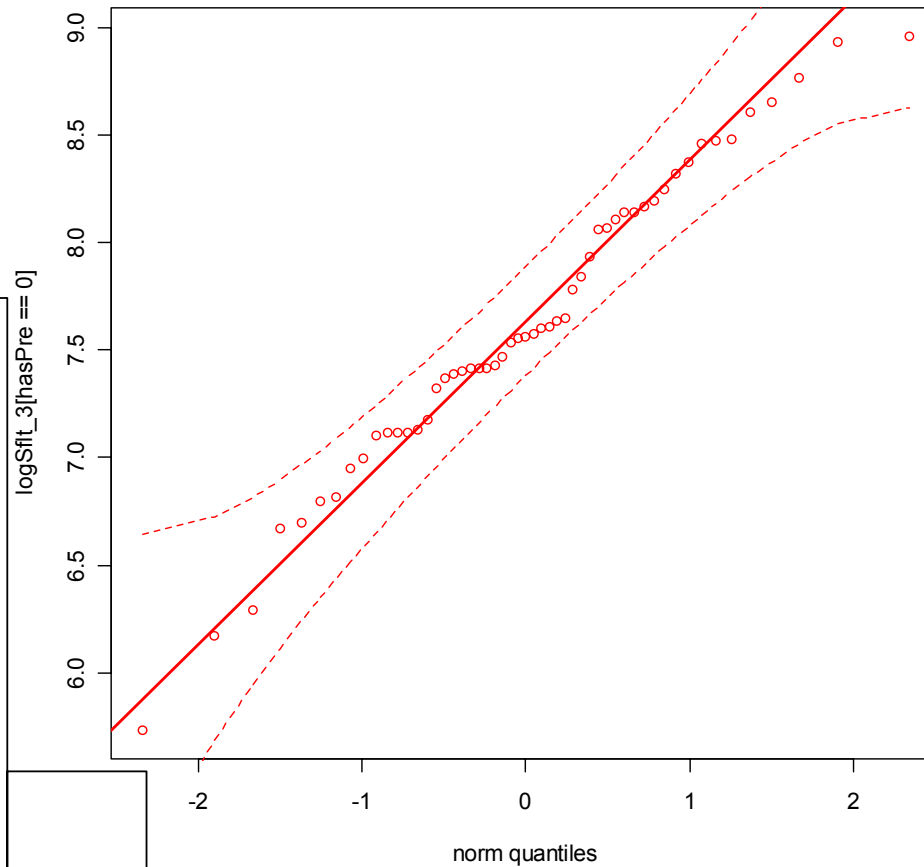
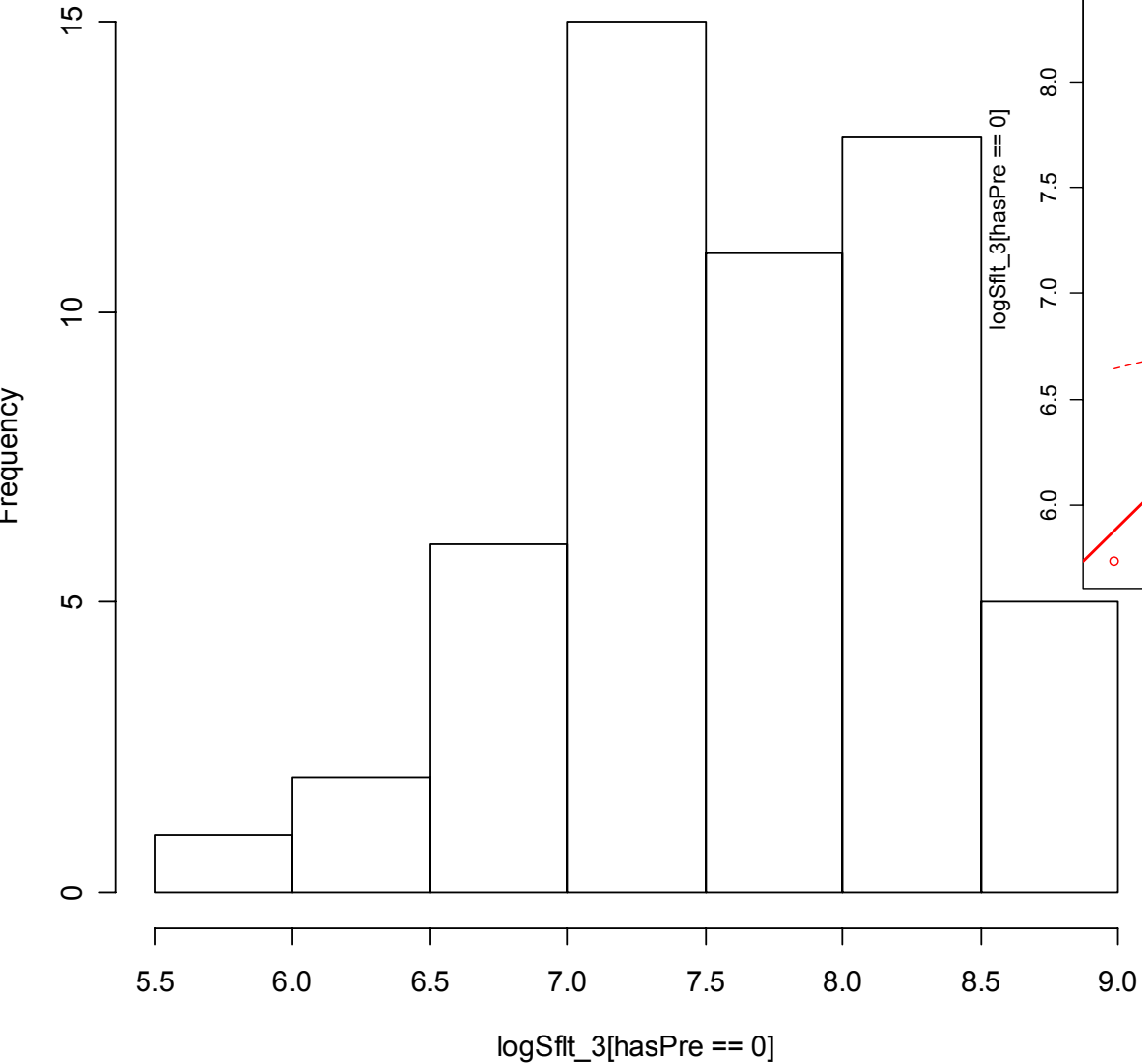


Shapiro-Wilk normality test $P < .00005$

How to Fix Grossly Not Normal Data

- You can either transform the variable or you can do an analysis that does not specify that the distribution has to be normally distributed.
- There is a non-parametric analog to the t-test called the Wilcoxon Rank Sum or Wilcoxon Signed Rank Tests (aka the Mann-Whitney test).
 - They say the shape of the distributions are about the same but one group is shifted to the right or the left.

Histogram of logSflt_3[hasPre == 0]



Shapiro-Wilk normality test $P < .7$

Parametric vs. Nonparametric

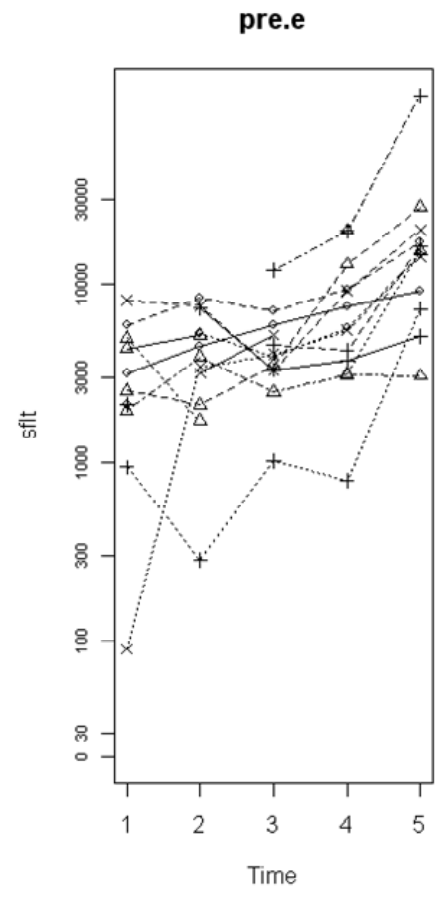
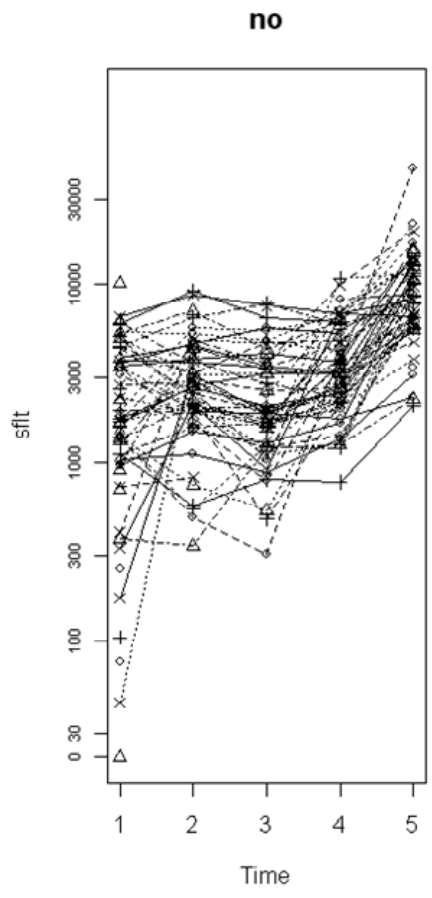
- Ignoring the problem
- $p < 0.02459$
- Fixing it with a log transform
- $p < 0.002769$
- Doing the nonparametric test
- $p < 0.004045$
- Moral of the story... nonparametric tests are powerful when the assumptions of the parametric model are not met.

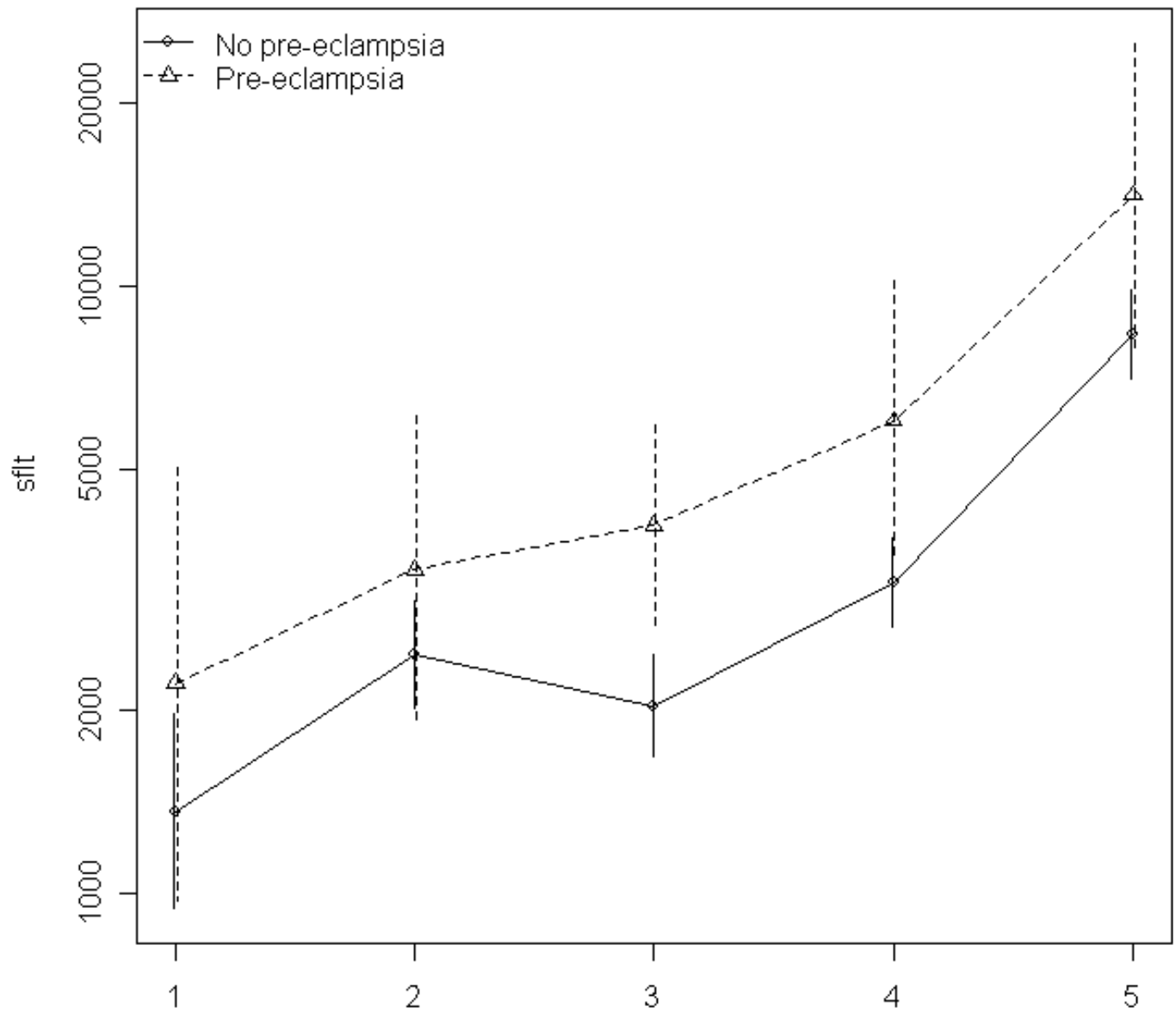
Further Complications

- Most of the women are scored zero on numerous proteins. Transforming the zero data is tricky. So, nonparametric analysis is very appealing...
- What is the nonparametric analog to the repeated measures ANOVA?
 - Welcome to the world of developing new statistics!

What is next...

- Because the relative weakness of the statistical theory for the test, it is best to treat this as a descriptive study...
- The final paper will feature lots of summary graphics showing the trends over time and my job is to discourage inferential tests at each time point for every enzyme.





How to Do This Stuff

- SAS is weak for publication-ready, easy graphics, great for data manipulation, poor GUI (graphical user interface – the point and click stuff)
 - Moderate to easy to learn
- R is great on graphics but mean and nasty for data manipulation and a horrible GUI (but a couple good alternatives for code)
 - HARD to learn
- S-plus is great for graphics, relatively hard to do data manipulation, good GUI (superb integration with Excel)
 - Moderate to hard to learn
- SPSS has good graphics, ok for data manipulation, **GREAT GUI**
 - Easy to learn

Tasks

- Import a dbf file (real data) and xls file (artificially clean)
- make scatter plot of apgar1 and 2
- correlation of apgar1 and 2
- 3d plot of apgar1, apgar2 and moms age
- Make a boxplot of age
- Make side by side boxplots of age by delivery type
- Parametric comparison of age by delivery type
- Nonparametric comparison of age by delivery type