Noise, Outliers and Missing Data

J.D. Raffa

October 6, 2017

Outline

Data Generation

- Data generation: How the data is collected and processed before it gets to you.

- We think like this

- But the data is generated from this:

Noise

- All non-systematic components to a statistical model.
  - Random noise
  - Effects of unmodelled variables
  - Contamination

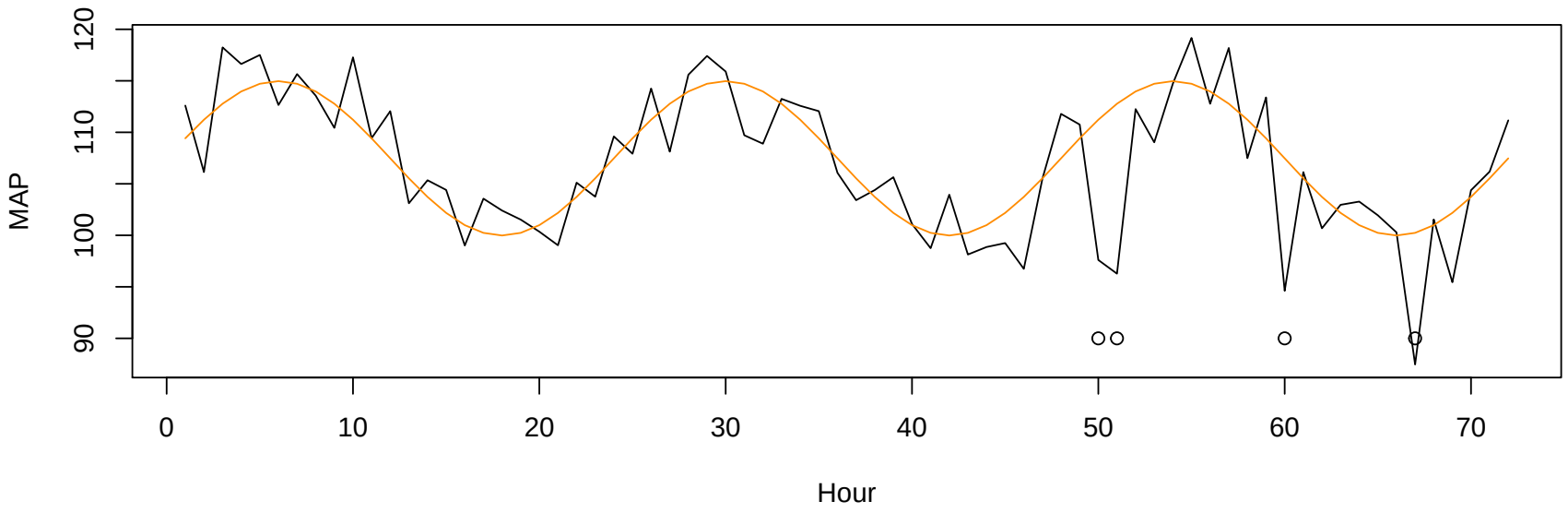


Figure 1: Systematic and Noise Part of Data

Outliers

- Non-technical definition:

Outliers

- Non-technical definition:
  - any data point which seems to be far away from the rest of the data points.

Outliers

- Non-technical definition:
  - any data point which seems to be far away from the rest of the data points.
- Can be a result from the regular distribution of the data, contamination or another part of the data generating process.

Outliers in the ICU

- Most ICU patients are in the ICU because they are quite ill.
- These may be among the more extreme conditions humans experience.
- Extreme values are often of most interest
  - e.g., APACHE often uses "the worst" value.
- Patients are heavily monitored, with multiple high frequency measurements.

Outliers in the ICU

- Most ICU patients are in the ICU because they are quite ill.
- These may be among the more extreme conditions humans experience.
- Extreme values are often of most interest
  - e.g., APACHE often uses "the worst" value.
- Patients are heavily monitored, with multiple high frequency measurements.

- Outliers may be more common here than in other situations in medicine.

Tukey's Rule

- Non-technical: plot the data as a boxplot, any points outside the whiskers, consider outliers.
  - Beware: multiple different boxplots formats.

Tukey's Rule

- Non-technical: plot the data as a boxplot, any points outside the whiskers, consider outliers.
  - Beware: multiple different boxplots formats.

- Technical:
  1. Calculate $IQR = Q3 - Q1$
  2. Setup a "Fence" 1.5 $\times$ IQR outside the IQR.
  3. Data outside interval, $[IQR - 1.5 * IQR, IQR + 1.5 * IQR]$, should be considered outliers.

- Practical considerations:

1. Under normality: This definitions translates to mean $\pm$ $2.7 \times SD$. This will occur about 0.3% of the time in each tail, our about 6 times out of 1000 total.
2. In symmetric distributions with heavier tails, this can occur more often.
3. For non-symmetric distributions, all bets are off.
4. Method is generally robust to the *number* and *magnitude* of the outliers.

Other Approaches: For Univariate Assessment

- *Z-scores*:
  - Calculate z-scores for each data point, exclude any greater than some threshold.
  - Doesn't work well for asymmetric data.
- *Formal Tests*
- *Anomaly Testing*
- Using Clinical Ranges
  - Some multiple of the normal clinical ranges.
  - Physical constraints: e.g., negative heart rate.
  - Needs clinical expertise.
  - Often these are already used for lab results.

Examples

1. eICU
   - BUN: 1270484 lab results in eICU: Normal Range 7-20 mg/dL

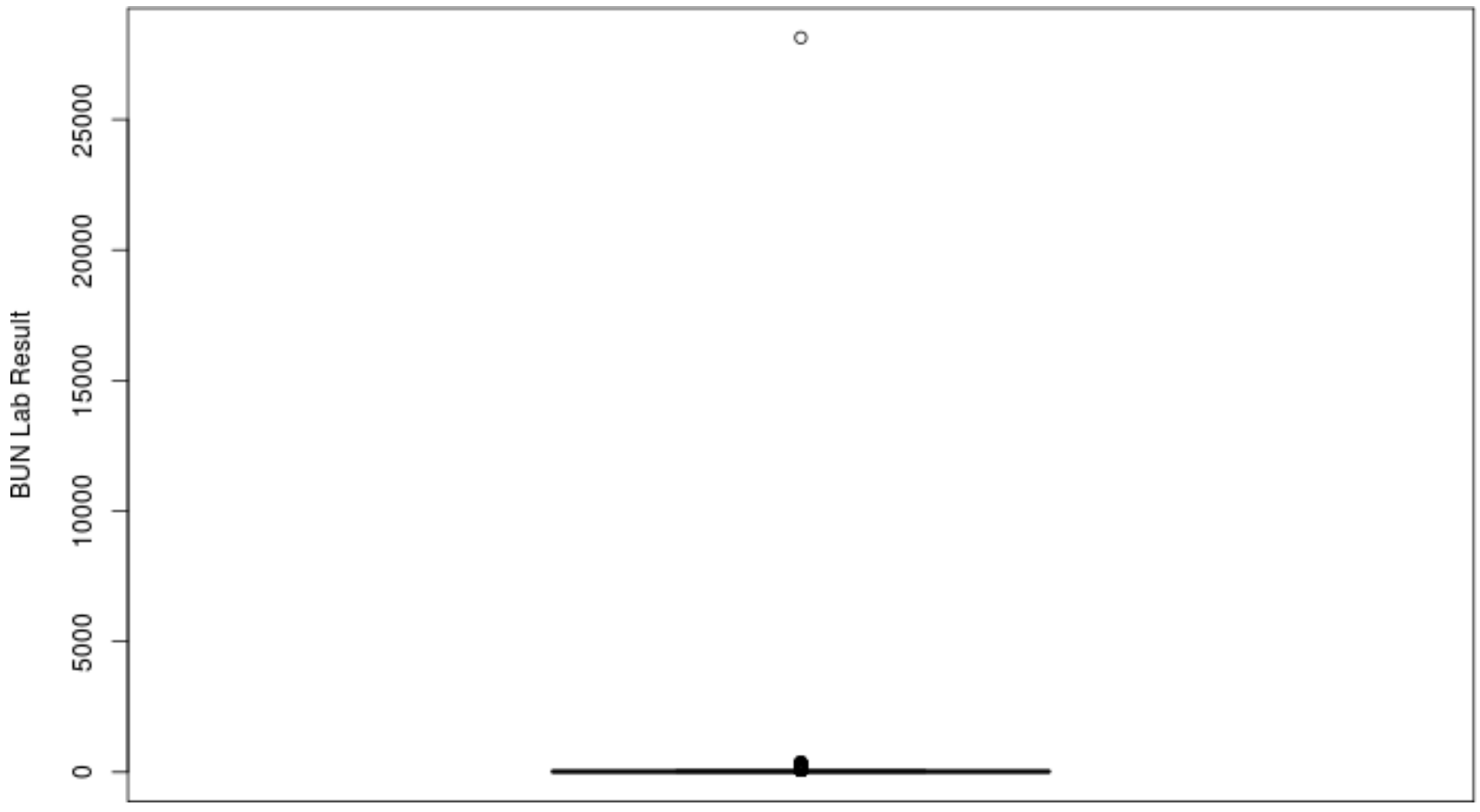
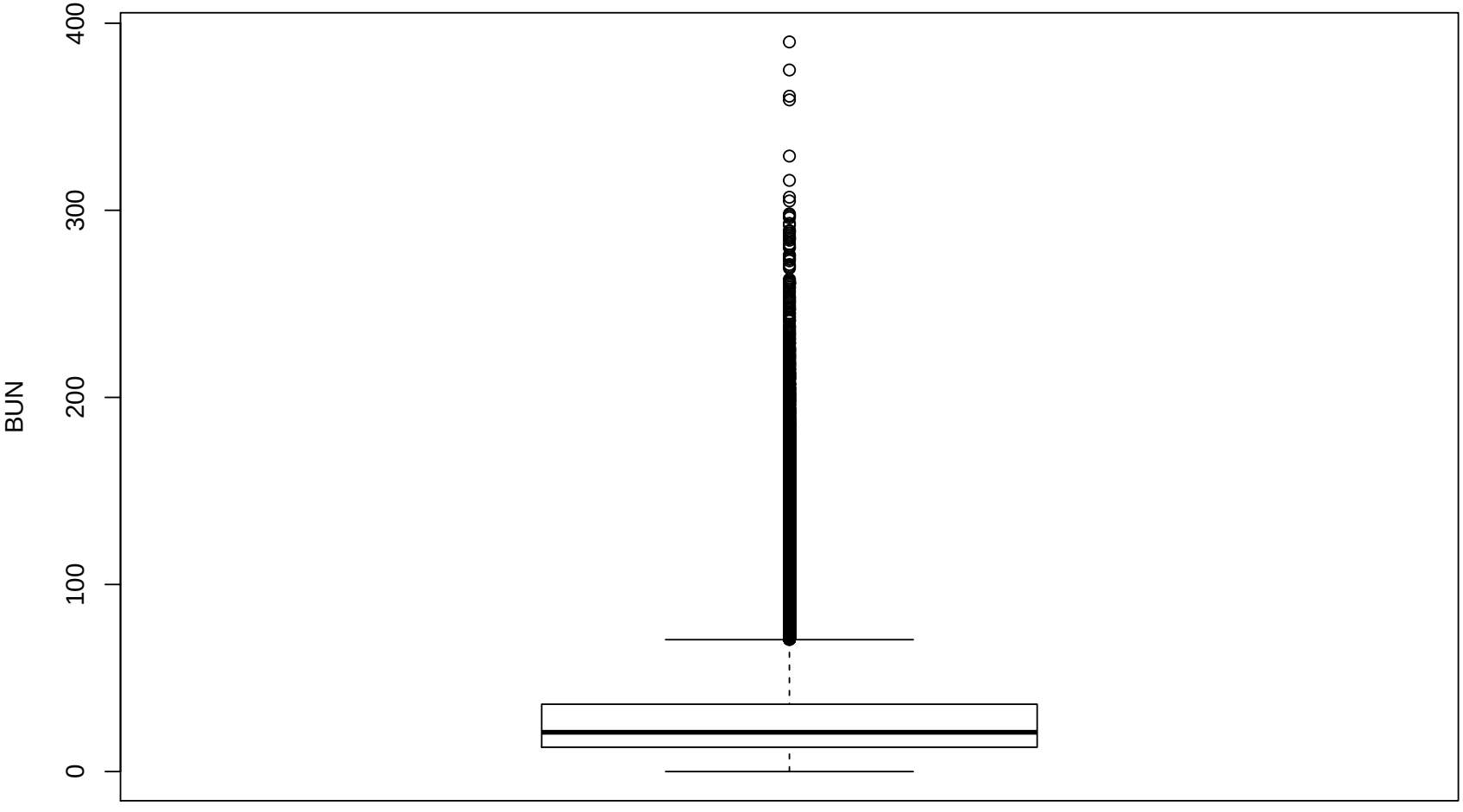


Table 1: Extreme Outlier BUN

patientunitstayid	labresultoffset	labname	labresult	labmeasurenamesystem	labresultrevisedoffset
268728	-160	BUN	31	mg/dL	871
268728	454	BUN	29	mg/dL	684
268728	909	BUN	30	mg/dL	961
268728	2184	BUN	32	mg/dL	2220
268728	3586	BUN	30	mg/dL	3668
268728	4364	BUN	29	mg/dL	6938
268728	5039	BUN	26	mg/dL	5200
268728	5747	BUN	27	mg/dL	6191
268728	6478	BUN	26	mg/dL	6944
268728	7201	BUN	28131	mg/dL	8286
268728	7925	BUN	27	mg/dL	8285

Let's ignore that one point:

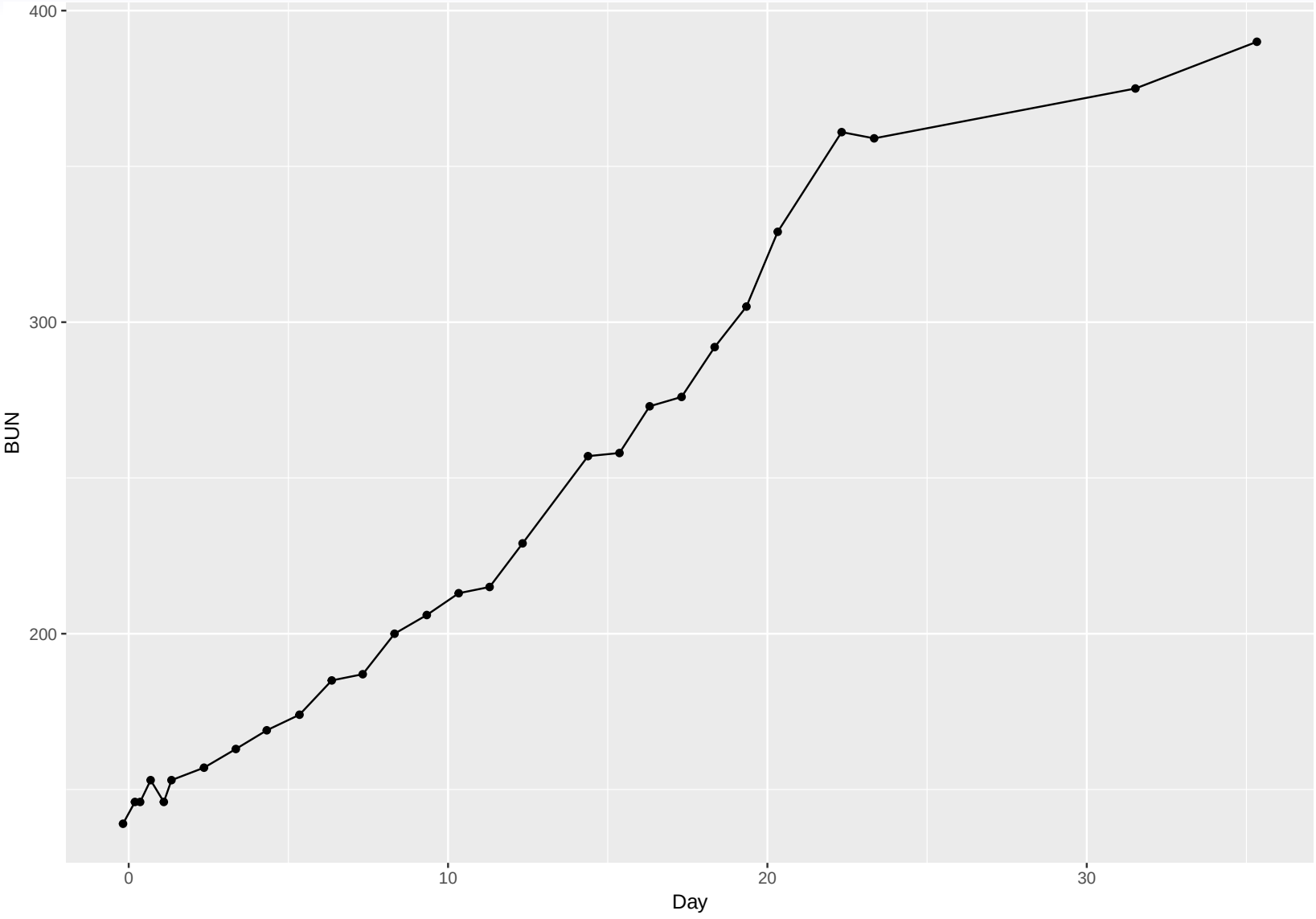


Still have 9.31 percent of data considered Tukey outliers.

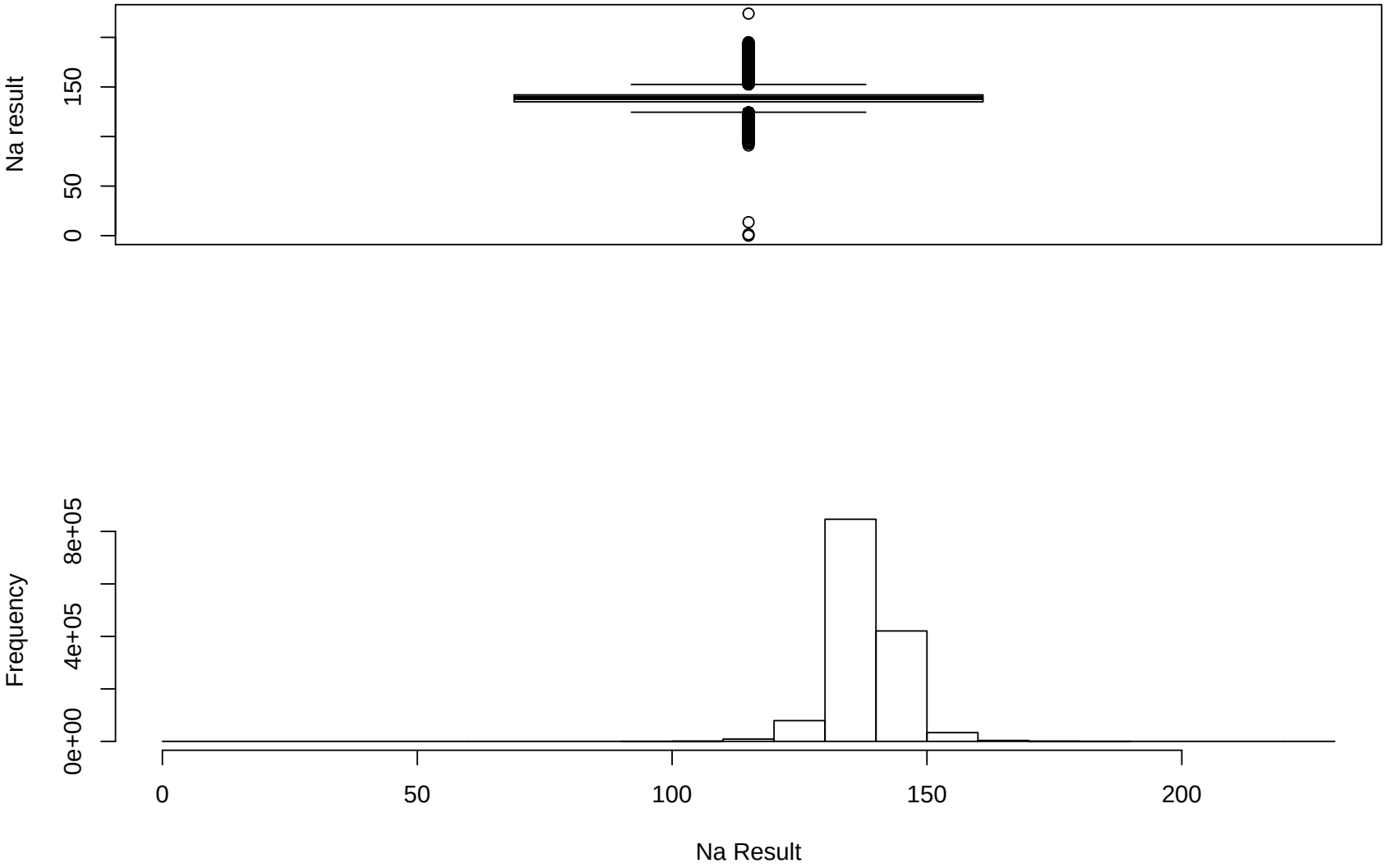
What about the second highest value?

Table 2: Extreme Outlier #2 BUN

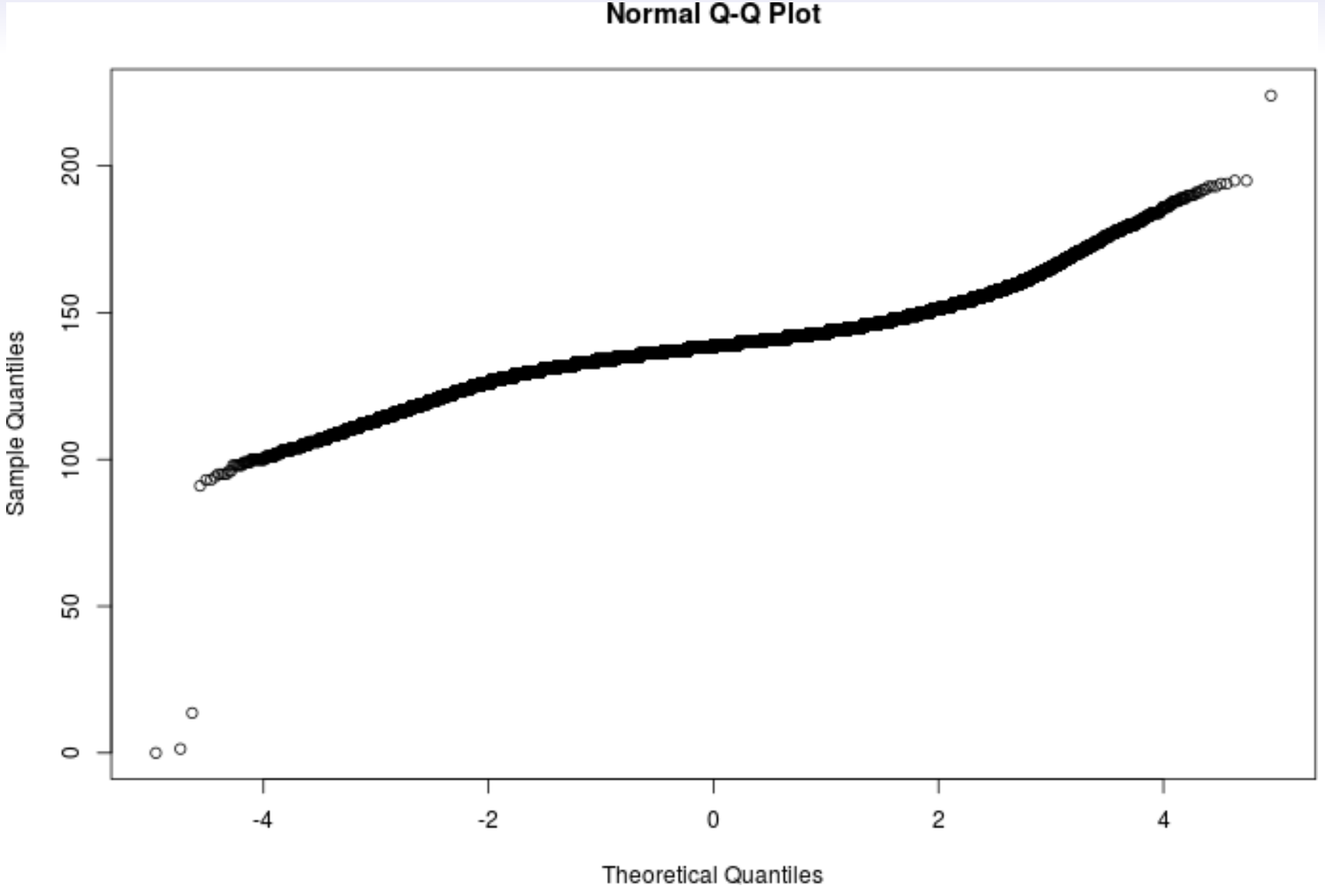
patientunitstayid	labresultoffset	labname	labresult	labmeasurenamesystem	labresultrevisedoffset
3097437	20710	BUN	257	mg/dL	20687
3097437	22130	BUN	258	mg/dL	22112
3097437	23490	BUN	273	mg/dL	23477
3097437	24930	BUN	276	mg/dL	24917
3097437	26422	BUN	292	mg/dL	26417
3097437	27855	BUN	305	mg/dL	27827
3097437	29265	BUN	329	mg/dL	29252
3097437	32147	BUN	361	mg/dL	32147
3097437	33613	BUN	359	mg/dL	33587
3097437	45394	BUN	375	mg/dL	45387
3097437	50872	BUN	390	mg/dL	50867


400
300
200
BUN
0
10
20
30
Day

- Sodium (1393205 tests) Normal Range 135-145 mEq/L

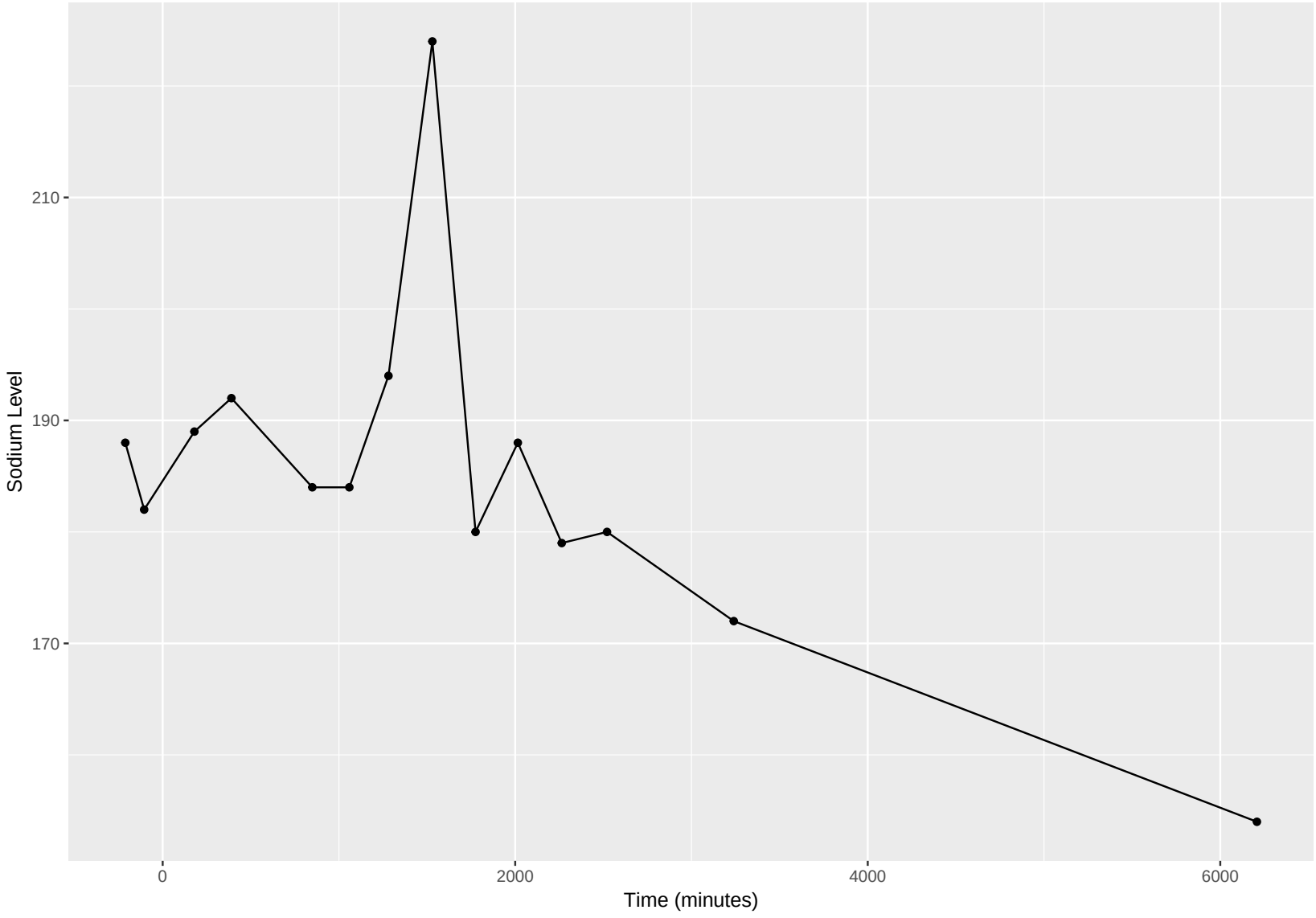


So $\sim$ 6% of data is considered an outlier by this definition. Things aren't really that bad though.



So. . . maybe a half dozen or so points? Let's look at a few.

Patient with Sodium over 200:



Patient with Sodium Close to Zero:

patientunitstayid	labresultoffset	labname	labresult	labmeasurenamesystem	labresultrevisedoffset
1068196	40	sodium	136.00	mmol/L	45
1068196	354	sodium	137.00	mmol/L	410
1068196	564	sodium	1.37	mmol/L	627
1068196	845	sodium	139.00	mmol/L	945
1068196	1889	sodium	139.00	mmol/L	1954
<..snip..>					

Patient with Sodium at Zero:

patientunitstayid	labresultoffset	labname	labresult	labmeasurenamesystem	labresultrevisedoffset
1556807	-214	sodium	139	mmol/L	-214
1556807	907	sodium	140	mmol/L	907
1556807	2244	sodium	138	mmol/L	2244
1556807	3772	sodium	137	mmol/L	3772
1556807	4763	sodium	0	mmol/L	4763
1556807	5113	sodium	139	mmol/L	5113
1556807	5148	sodium	138	mmol/L	5148
1556807	6489	sodium	137	mmol/L	6489
1556807	8121	sodium	138	mmol/L	8121
1556807	9381	sodium	135	mmol/L	9381

My General Philosophy

- I hate throwing away data.
- I will generally not delete data points, unless it's clear that there is an error.
- I won't "edit" points, to correct what I deem is the "correct value".

My General Philosophy

- I hate throwing away data.
- I will generally not delete data points, unless it's clear that there is an error.
- I won't "edit" points, to correct what I deem is the "correct value".

So. . . .what can we do?

Depends on the goal and the specifics.

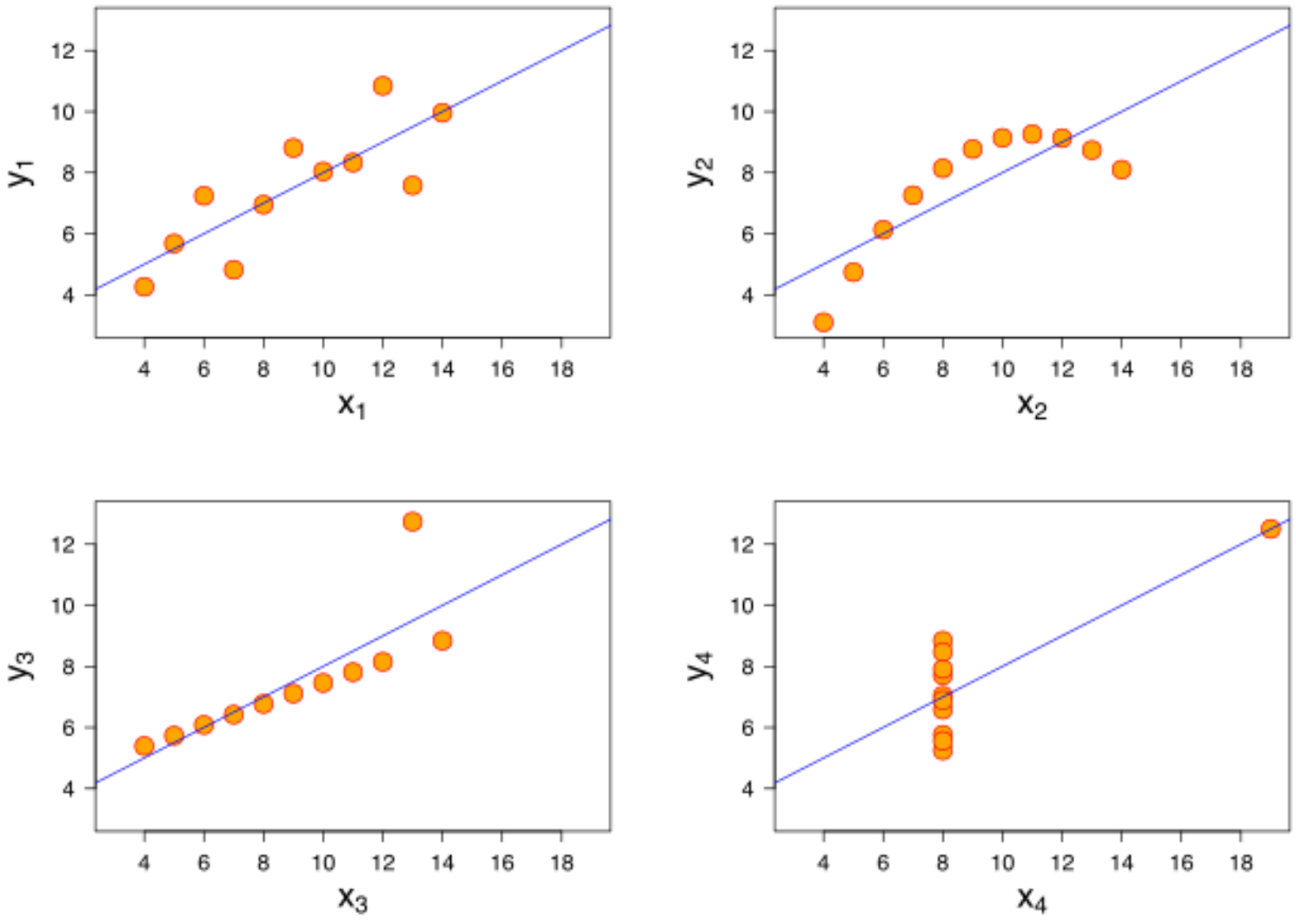
- Parametric models, tests and estimators can be sensitive to outliers.
  - e.g., average BUN: 28.33 w/ data large datapoint, and 28.31 w/o.
- Non-parametric tests and estimators are often a viable alternative:
  - sample median $\leftrightarrow$ sample mean
  - sample MAD/IQR: $\leftrightarrow$ sample SD
  - Mann-Whitney Test $\leftrightarrow$ t-test
  - ...
- Generally a good idea to do the non-parametric and parametric approaches side-by-side.

Depends on the goal and the specifics.

- Parametric models, tests and estimators can be sensitive to outliers.
  - e.g., average BUN: 28.33 w/ data large datapoint, and 28.31 w/o.
- Non-parametric tests and estimators are often a viable alternative:
  - sample median $\leftrightarrow$ sample mean
  - sample MAD/IQR: $\leftrightarrow$ sample SD
  - Mann-Whitney Test $\leftrightarrow$ t-test
  - ...
- Generally a good idea to do the non-parametric and parametric approaches side-by-side.

- The interpretation sometimes is not comparable.

That was the easier case. The harder case is regression methods.



We will return to this in a few weeks.

Some options:

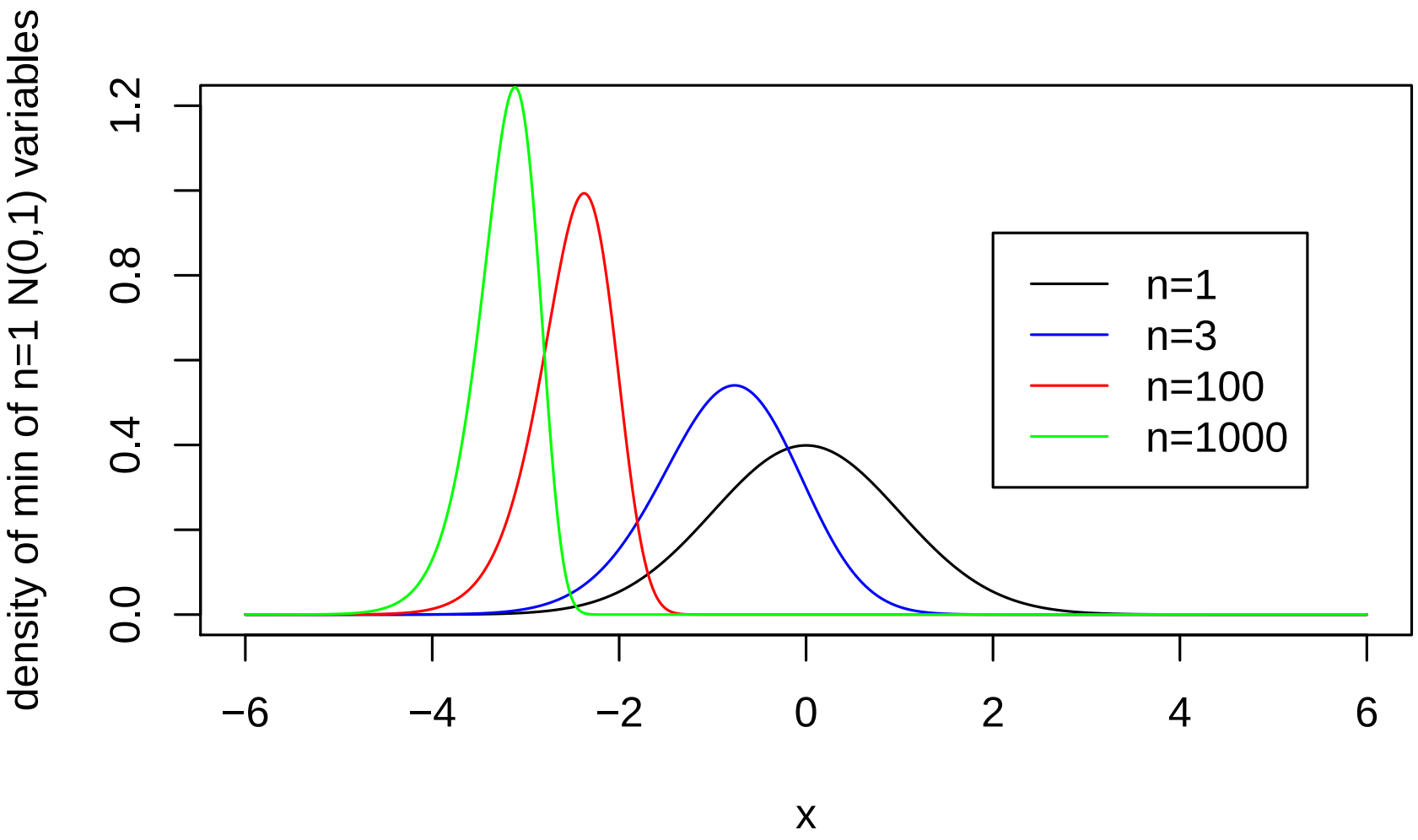
1. Deletion: remove the data point OR replace with a missing value.
2. Transformations of outcome or covariates:

   - Log, sqrt, etc
   - Map to categories: [0-4), [4-7), >=7

3. Alternative Methods or Assumptions
4. Winsorization: Determine what an outlying point is a replace it the most extreme allowable point.
5. Robust Methods: e.g., M-estimators.

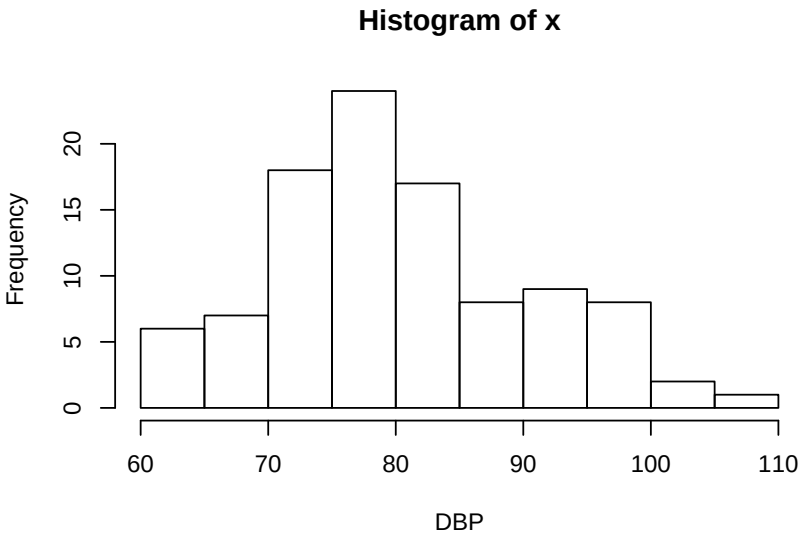
Other considerations

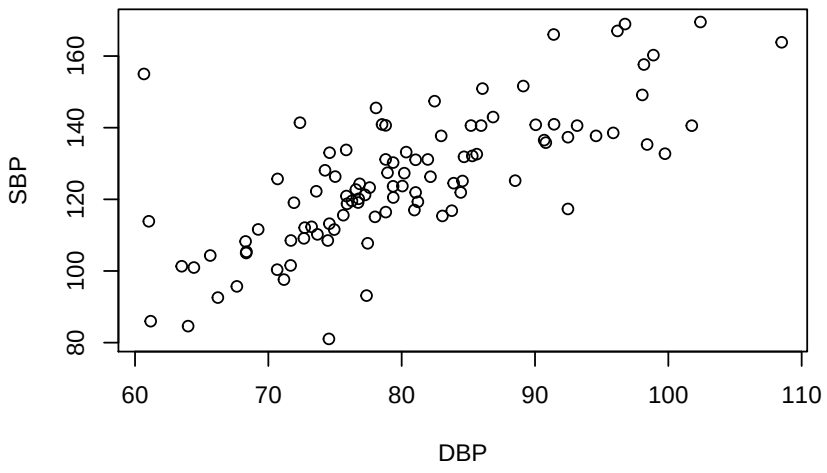
- Be careful when using extreme (min or max) values:

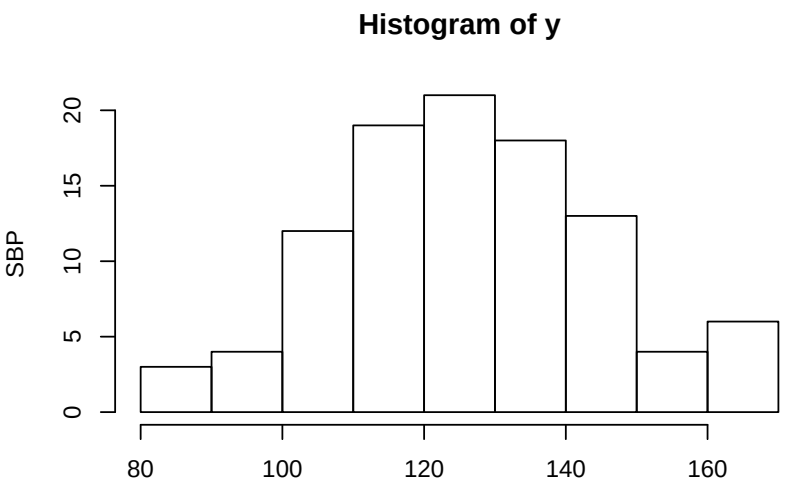


- Consider testing also with second most extreme value.

- In multivariate settings, there's another wrinkle:
  - A variable may not be a outlier when two variables are looked at in the univariate setting, but when examined together, they often can be considered outliers.







Take Home Messages: Noise and Outliers

- Noise and Outliers exist – get used to it.
- Formal criteria exist, but use some common sense and plot!
- Avoid deletion unless absolutely necessary.
  - Outliers are often the most interesting points!
- Handle carefully, check for alternatives, transform if necessary.
- Pay special attention to multivariate relationships between variables and outliers in that context too.

Missing Data

- Missing data: when data you would like is not present in the dataset:
  - Data has been partially or fully destroyed.
  - Data was not provided by the subject.
  - When the subject was lost to follow-up ....

Missing Data

- Missing data: when data you would like is not present in the dataset:
  - Data has been partially or fully destroyed.
  - Data was not provided by the subject.
  - When the subject was lost to follow-up ....
- These are classic examples. In this course more likely to have:
  - A measurement was not measured when you would like it to be measured.
  - When documentation (notes, diagnostic codes) are incomplete.
  - When a type of data is not recorded at that hospital, or wasn't recorded during a time period.
- Can be very complex!

Example:

- Collaborator:
  - lactate: wants 3 tests in the first 24 hours for inclusion in the study.
  - You show a lot of people are missing 1 or more of these 3 lactate tests.
  - Ask: Who gets lactate tests?
    - Likely answer: those patients we are worried about. . . .

Not enough time to get into this course, but some things you should be aware of:

- When fitting most models: any patient with an outcome *or* any covariate that is missing will be dropped.
  - You actual n is the number with complete data.
  - Depending on the models you fit, you may be fitting on different samples.

Suggestions:

- Try to loosen up criteria which create missing data.
- Try to figure out why data is missing.
- Try to run analyses on those with complete data, and those with partially missing data.
- Use multiple imputation (sometimes), carefully.
    - Multiple times for a reason!

Workshop (Optional for extra credit)

1. Download this presentation, and the resulting RMarkdown file.
2. Pick one vital sign and one type of laboratory test in eICU or MIMIC. Choose one related to your project, if possible.
3. Determine the Tukey fences for each. How many data points (and %) would be considered outliers by this definition?
4. Conduct a full investigation of up to three potential outliers on each, and conclude what you would do with them.
5. For the same two variables, determine how many patients have no results for this particular test. Pick two patients for each, and try to determine why the patient may have no data. You may not be able to.

Support all your answers with code, and output. Post code and output on GitHub, and e-mail `hst953hw@mit.edu` the repo's URL.

Please do NOT post data on GitHub.

If you choose to hand this in for credit, please do so by November 23rd, 2017. I would recommend completing this *AFTER* the EDA Problem Set is due. If you have problems, please come to office hours, or arrange an appointment with an instructor.