Transformations and non-parametric tests

Presenter: Steve Baskauf
steve.baskauf@vanderbilt.edu

DiSC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

CodeGraf landing page

- vanderbi.lt/codegraf

Transforming data that aren't normally distributed

DiSC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

Some common transformations

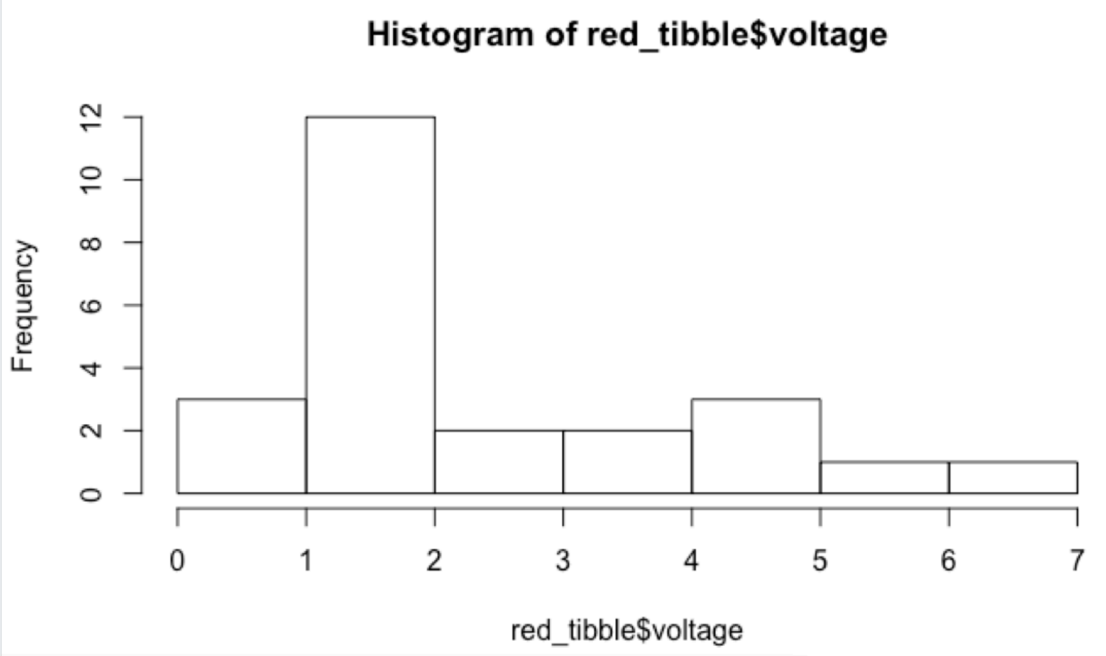
- Data skewed to the right: **`log()`**
  - data without negative values (range: 0-infinity)

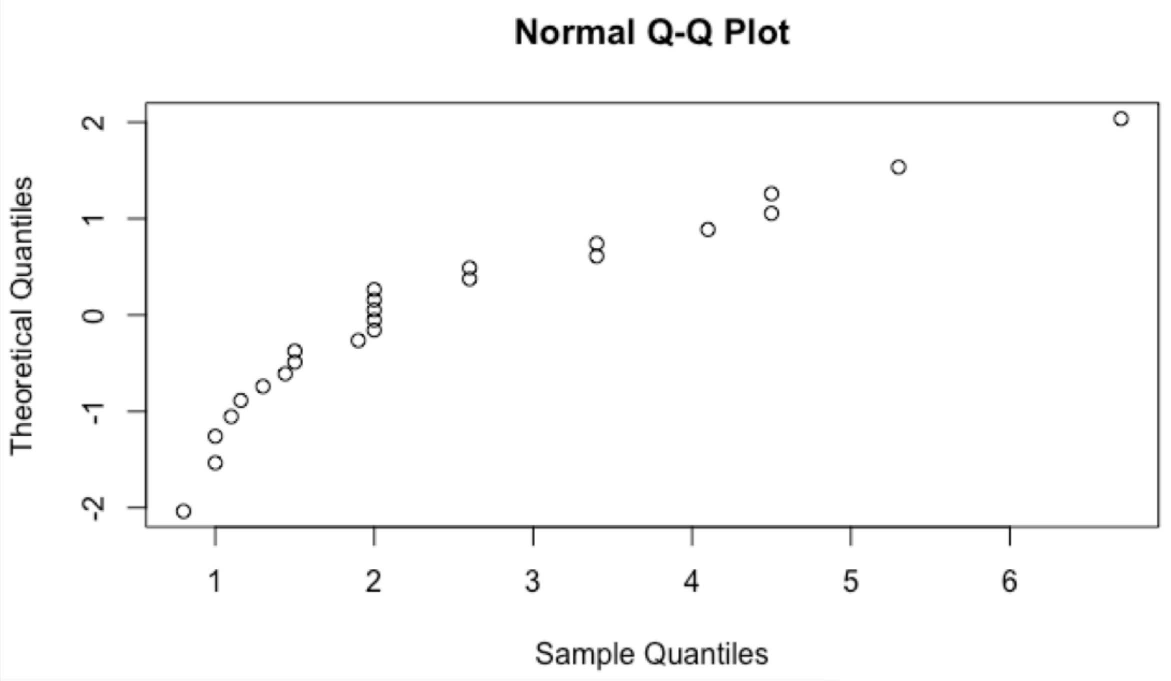
- Counts of things: **`sqrt()`**

- Proportions: **`asin(sqrt())`**
  - But usually you are doing the wrong test and should actually be using a logistic regression.

Reference: http://www.biostathandbook.com/transformation.html

Distribution of red responses

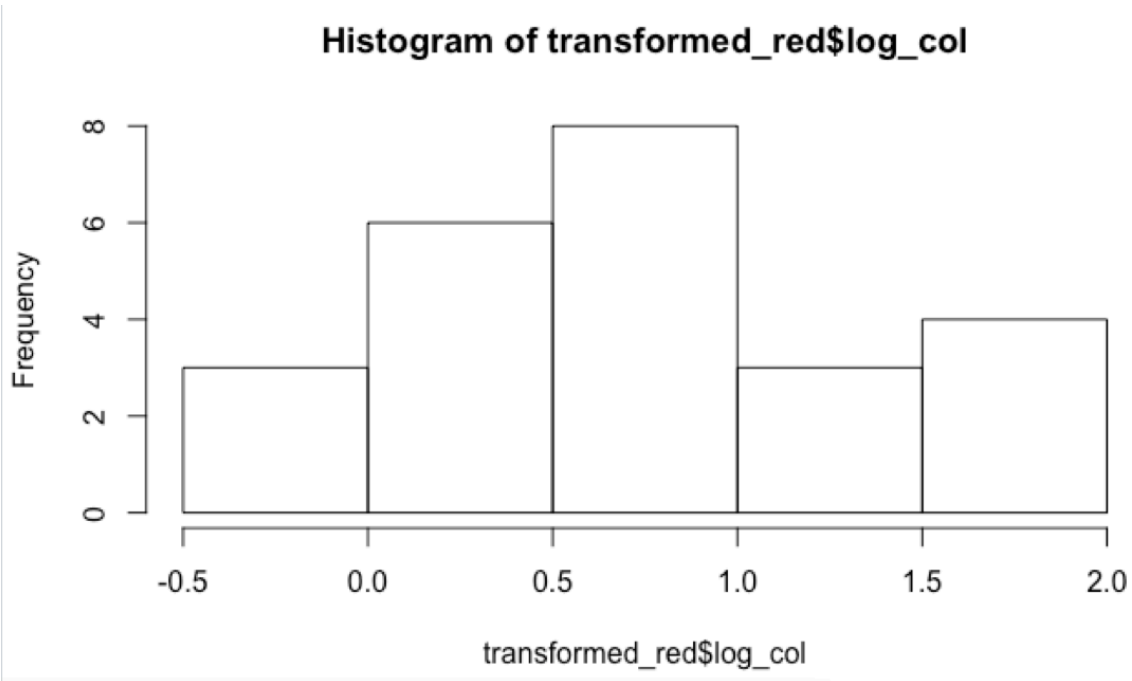


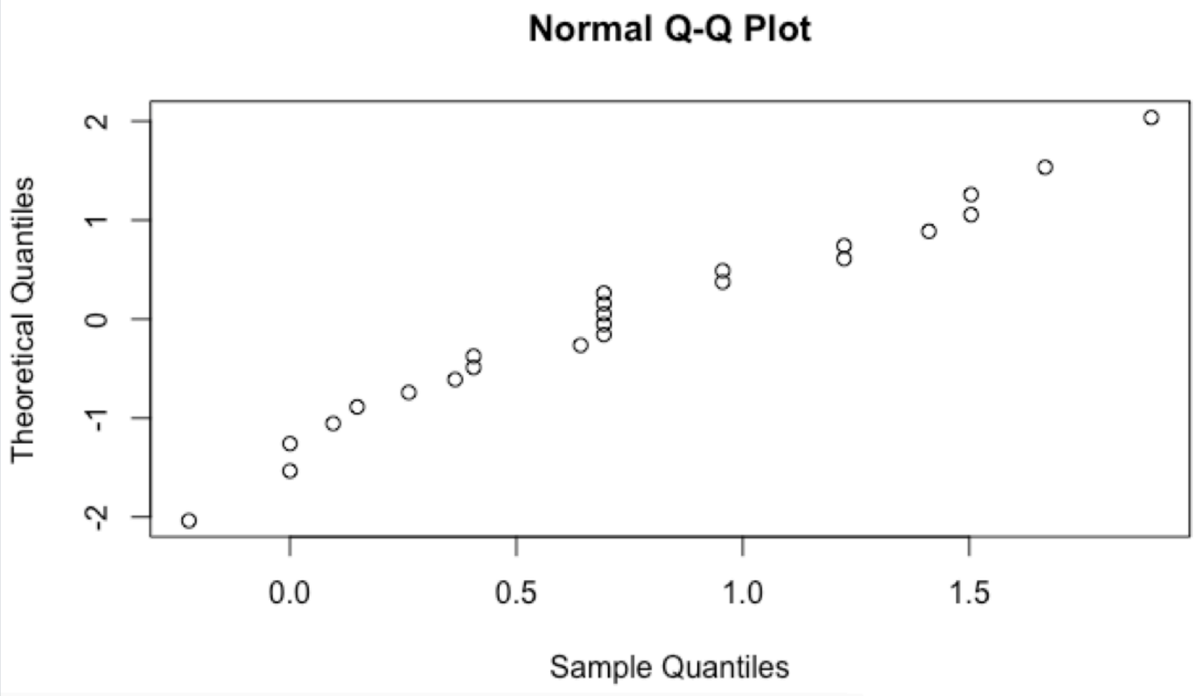


skewed to the right

Shapiro-Wilk normality test
p-value = 0.004284

`log(voltage)` transformation (red)

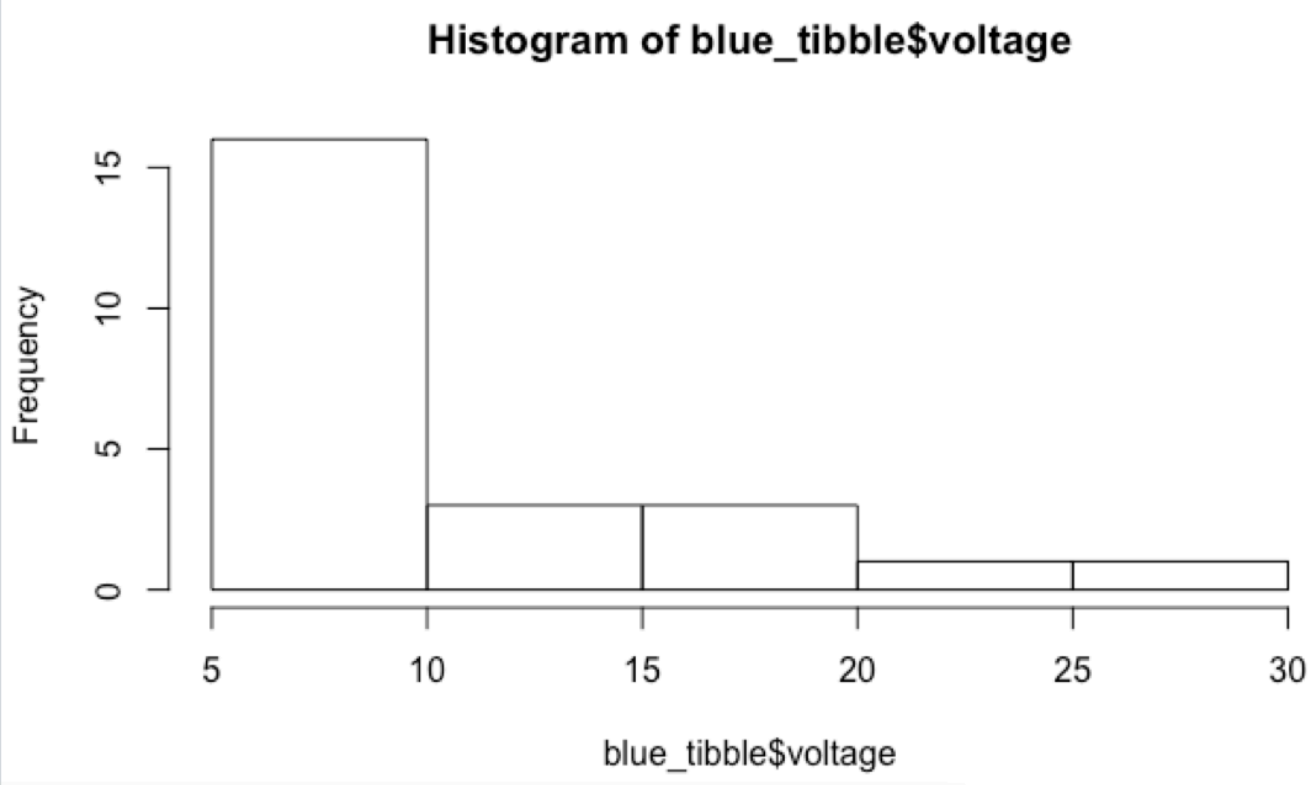


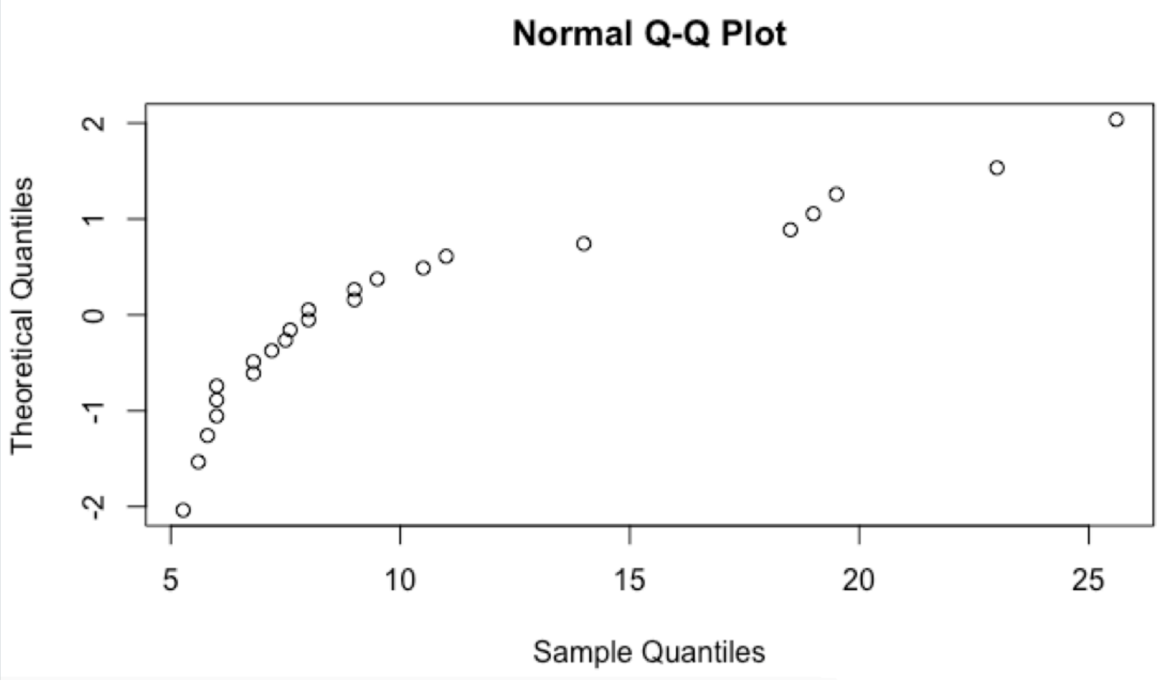


data now normally distributed

Shapiro-Wilk normality test
p-value = 0.4922

Distribution of blue responses

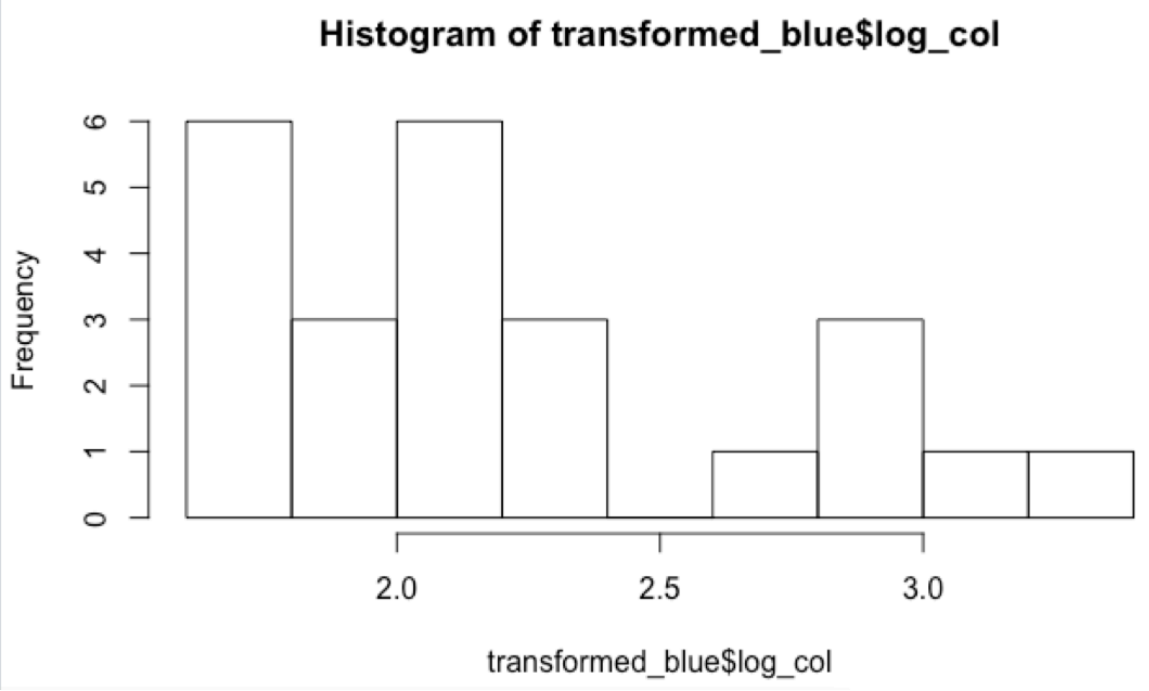


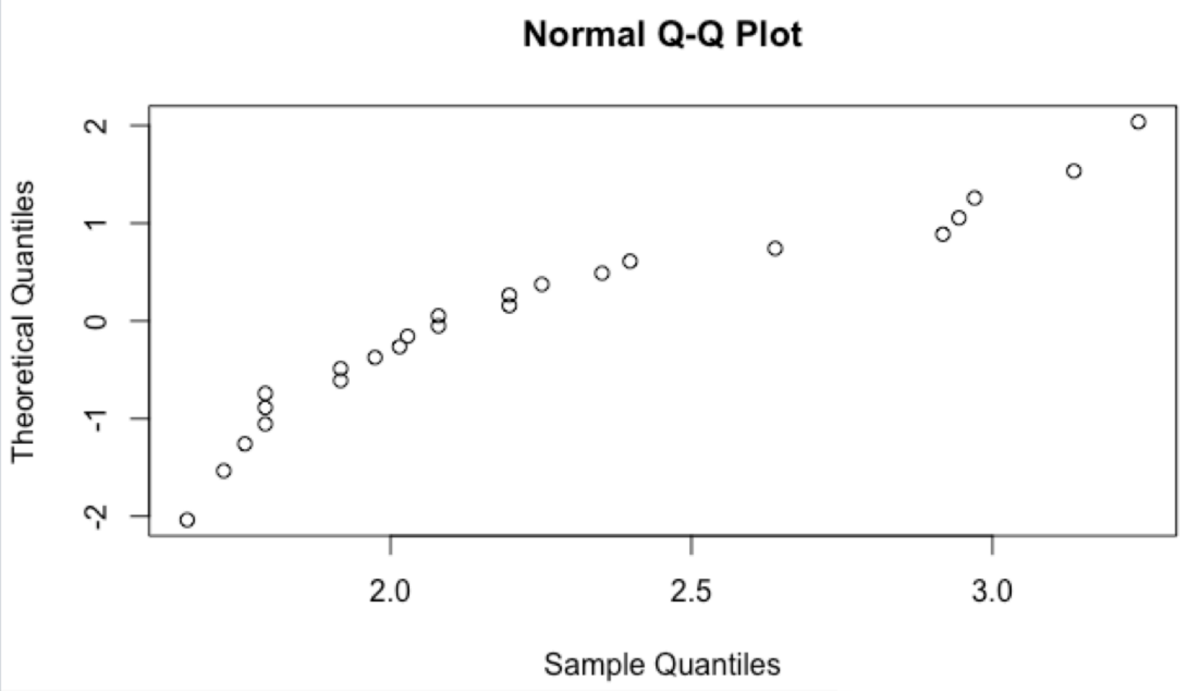


skewed badly to the right

Shapiro-Wilk normality test
p-value = 0.0002075

`log(voltage)` transformation (blue)





data still not so great but probably OK

Shapiro-Wilk normality test
p-value = 0.01005

Transformation may also fix heterogeneous variances

- Bartlett's test before transformation: P = 1.146e-08

- Bartlett's test after log() transformation: P = 0.3763

Test results after transformation

DiSC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

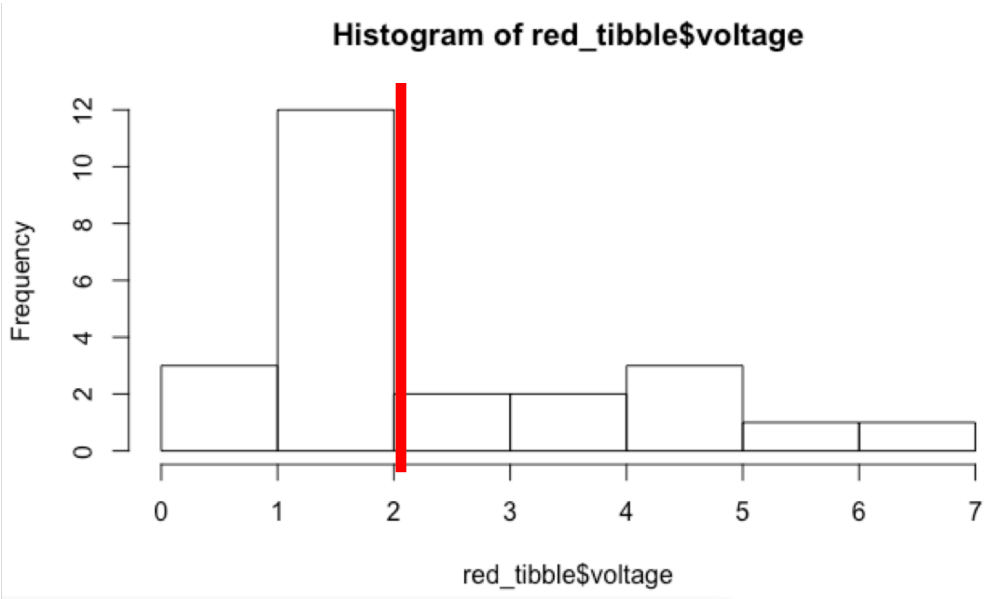
t-test of means after transformation

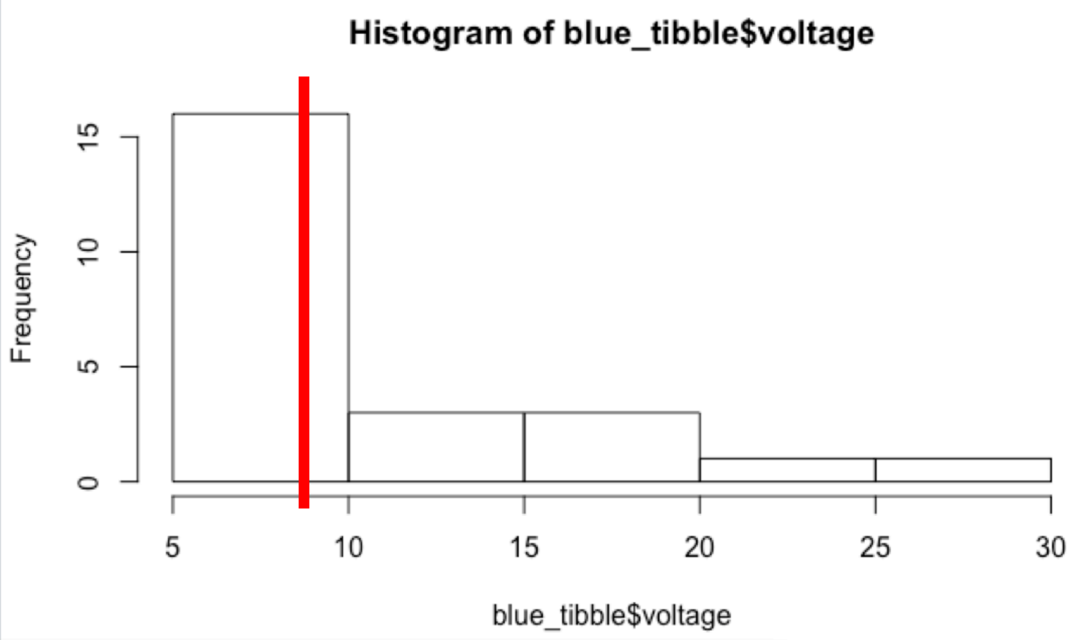
- P = 1.133e-12

- But confidence interval and estimated means are on log scale!
- "back transform" by inverse function:
    - inverse of ln(x) is e^x
    - e^x function in R is **`exp()`**

t-test of means after transformation

- estimated log() means: blue 2.2405239, red 0.7464018
- estimated means: blue 9.398254, red 2.109396





Non-parametric tests

DISC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

Non-parametric alternatives to tests

- If we fail to meet the assumptions of a test, there may be alternatives
- Non-parametric alternative to t-test of means: Wilcoxon-Mann-Whitney (WMW) test
  - a.k.a "rank sum" test
  - a.k.a Mann-Whitney U test

- WMW test tests whether the distributions of two groups are different; those differences won't always be in the means

```
wilcox.test(y ~ x)
```

Why not always use a non-parametric test?

- Parametric test result (t-test of means with log transformation)

  P = 1.133e-12

- Non-parametric test result (Wilcoxon-Mann-Whitney test)

  P = 7.087e-09

- General principle:non-parametric tests have less statistical power than parametric tests
- In this case, both were highly significant, but in borderline cases, it could make a difference.
- If transformation made normal but variances still unequal, use the t-test of means for unequal variances (still a parametric test)

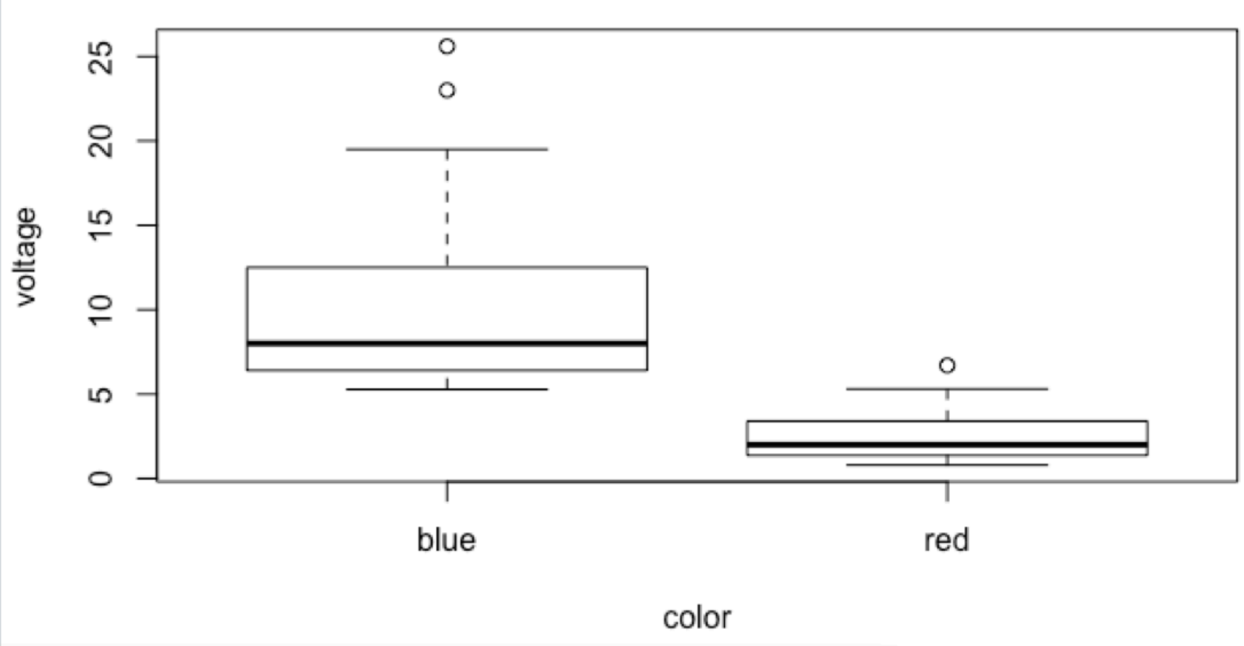
Visualization options for t-test

DISC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

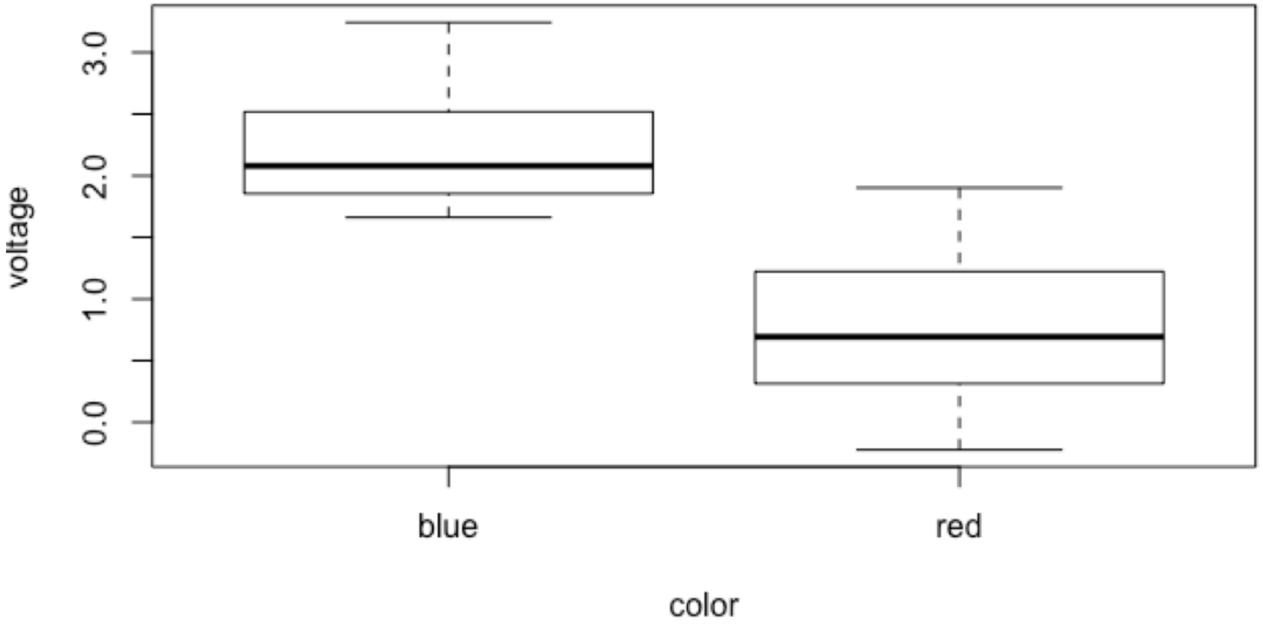
Box and whisker plot

- plot(voltage ~ color, data=erg_factor)
- plot(voltage ~ color, data=transformed_factor)



untransformed

y axis is understandable, but does not reflect the test

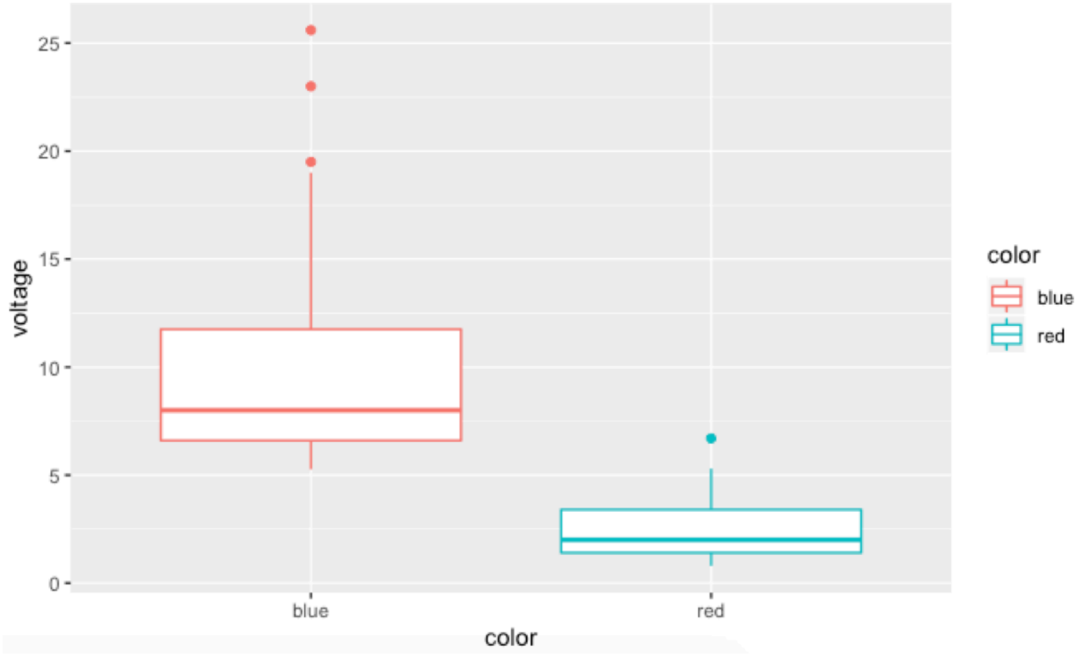


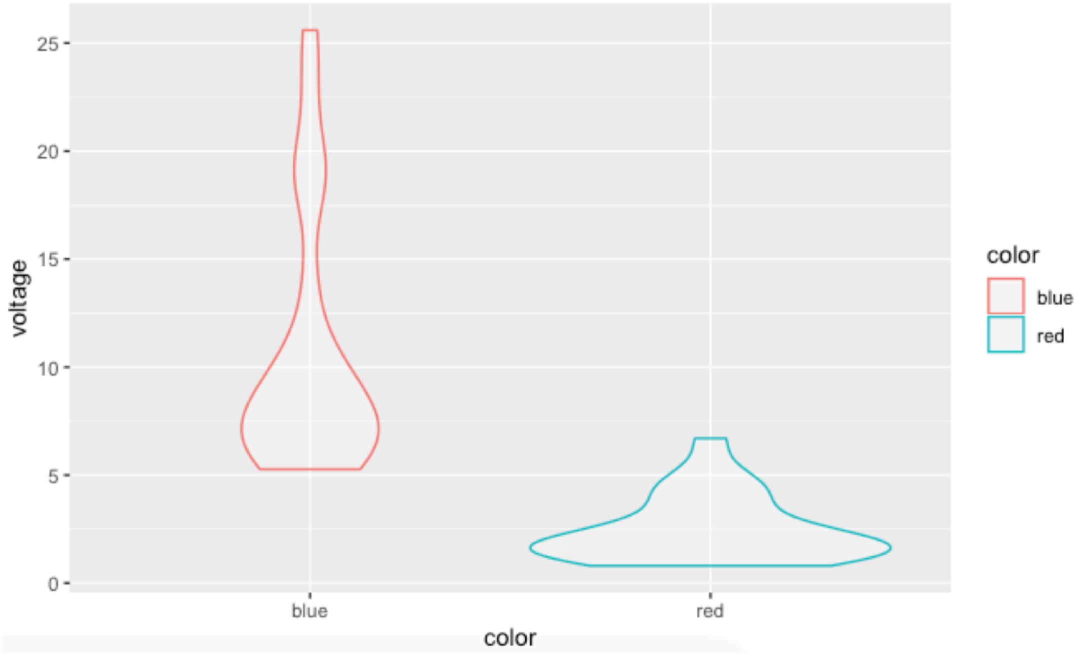
log transformed

y axis is obscure (log), but reflects the actual test test

More sophisticated plots

- The ggplot package provides much more control over the plot parameters.





```
ggplot(data = erg_factor, aes(x=color, y=voltage, color=color)) +
    geom_boxplot()
```

```
ggplot(data = erg_factor, aes(x=color, y=voltage, color=color)) +
    geom_violin(alpha = 0.3) # alpha controls transparency
```