Even more assumptions… but first, a return to normality

Kathleen Torkko

September 18, 2019

Objectives

Revisit assessing normality in small sample sizes

Learn about data transformation to help attain a normal distribution

Learn about homoscedasticity and how to assess it

Learn how to assess linearity

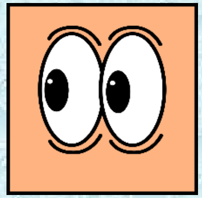
Learn what independence is

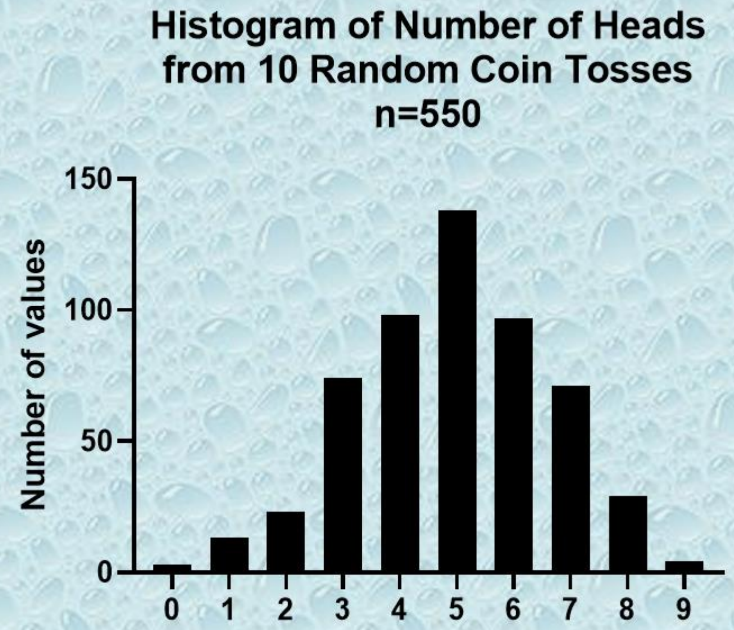
Learn more about outliers

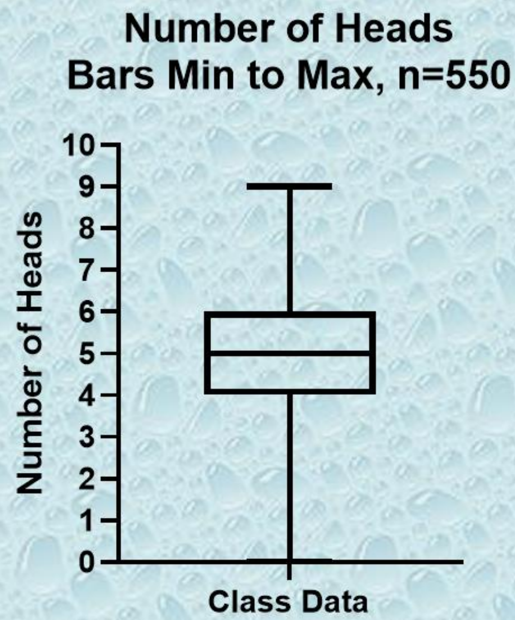
First, let's revisit steps for assessing normality in small data sets

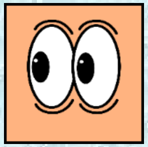
Look at distribution of number of heads per 10 coin tosses

NumberHeadsFrom10CoinTosses2019.xlsx

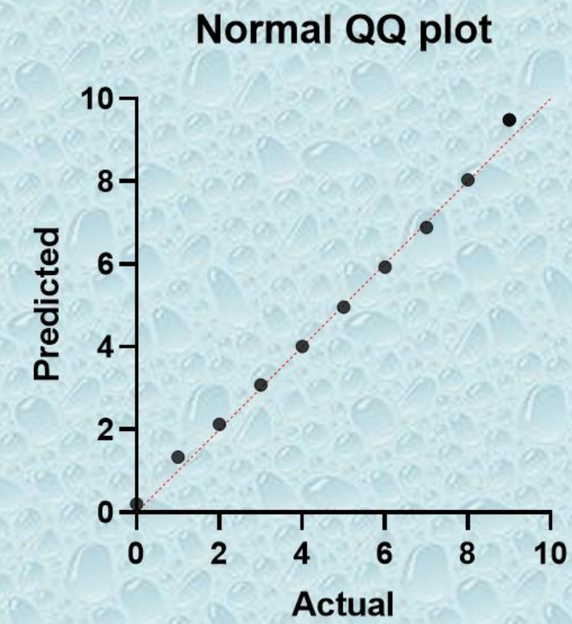








	All Data
N	550
Mean	4.9
Median	5
Skewness	-0.17
~~Anderson-Dar~~	~~No~~
~~D'Agostino~~	~~Yes~~
~~Shapiro-Wilk~~	~~No~~

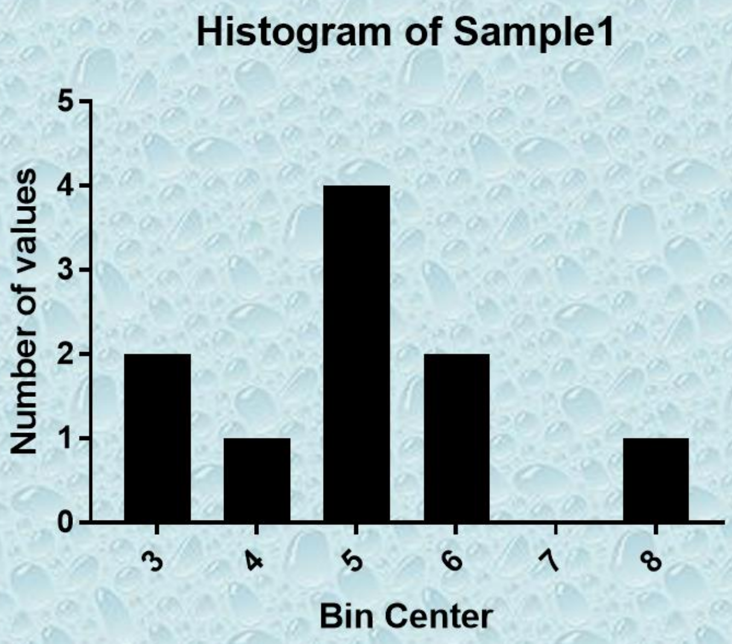


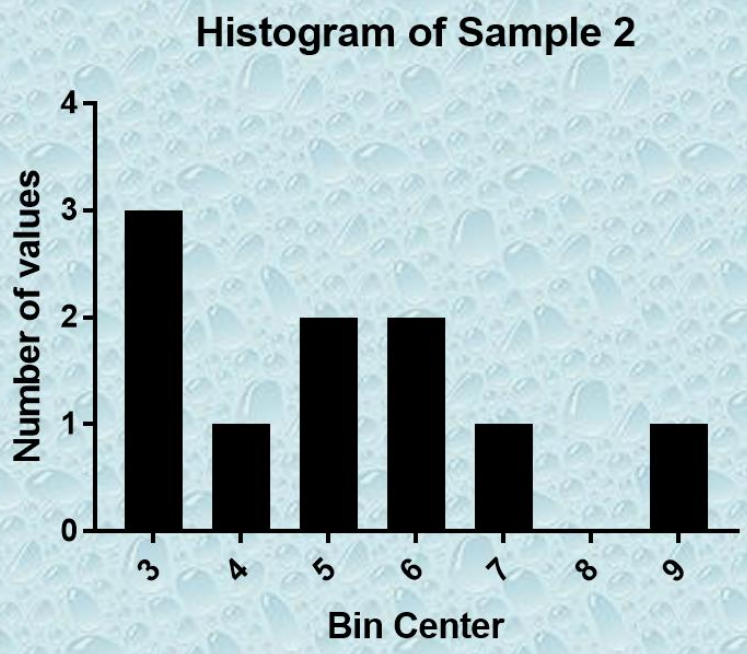
Let’s play with random samples of 10

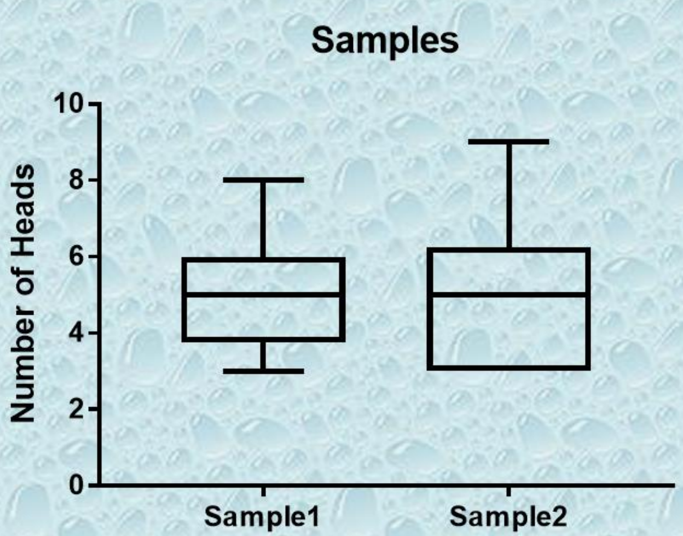
NumberHeadsFrom10CoinTosses2019.xlsx

RandomSample1
RandomSample2
RandomSample3

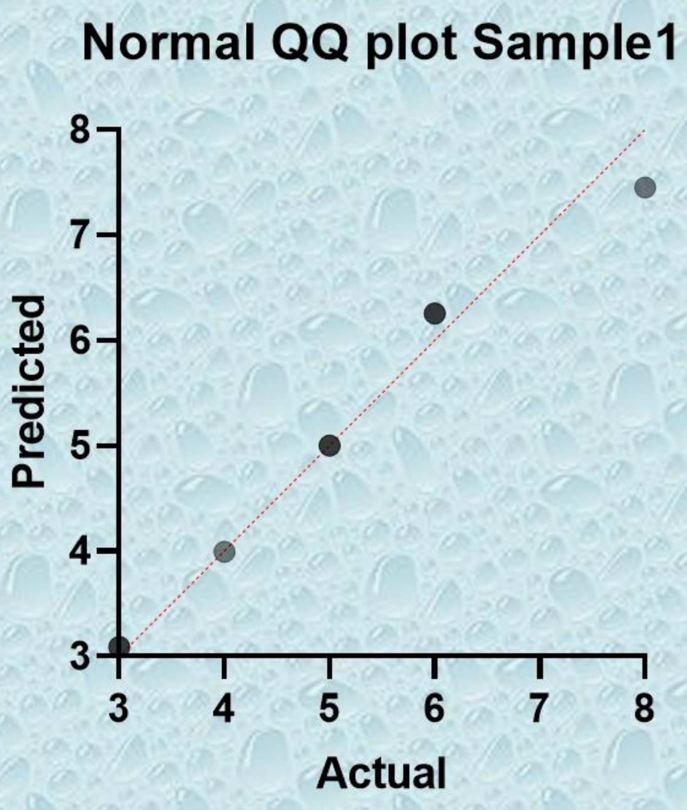
I have done the first 2 and we will do the third in class

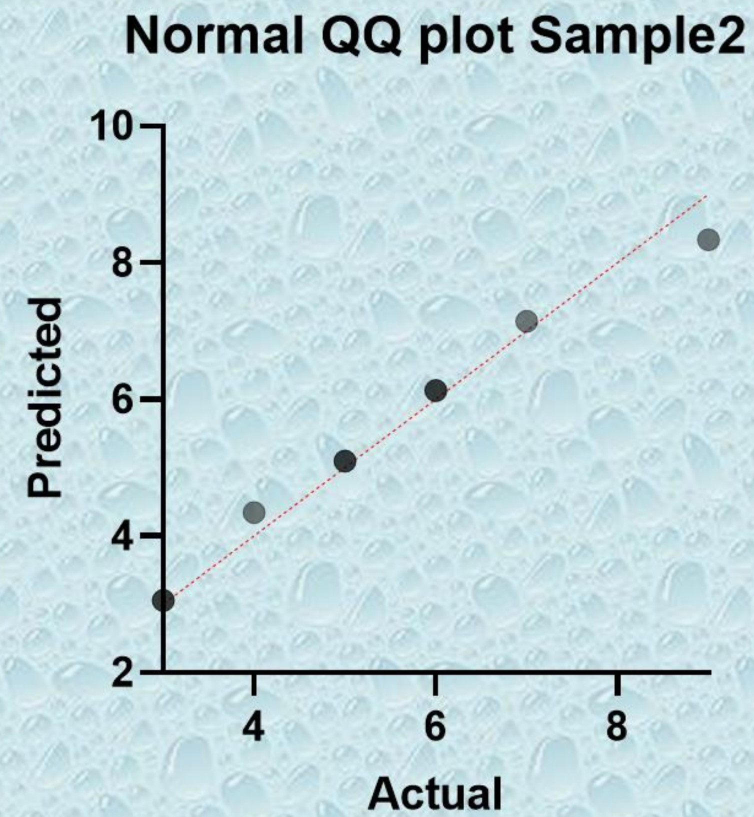






	Sample1	Sample2
N	10	10
Mean	5	5.1
Median	5	5
Skewness	0.503	0.701
~~Anderson-Dar~~	~~Yes~~	~~Yes~~
~~D’Agostino~~	~~Yes~~	~~Yes~~
~~Shapiro-Wilk~~	~~Yes~~	~~Yes~~





Let’s do Sample3

What can be done about non-normality?

1. Transform the data (same operation must be done on all data points for that particular variable)

2. Use a non-parametric test

3. Ignore it (well, not really)

When to transform…maybe…

Don’t transform if:

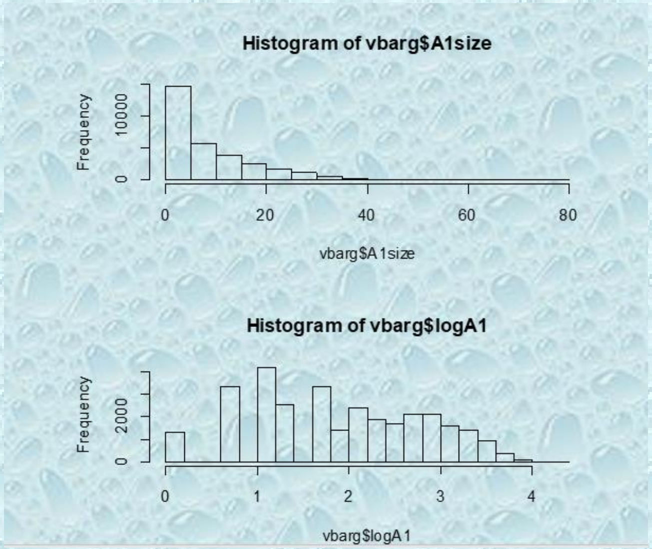
- The deviation from normality is not too extreme
- The sample is >30 and the data are roughly symmetrical
- You are using parametric statistics with known robustness
- The groups you are comparing are similar in distribution and sample size

Do transform when:

- The data are highly skewed
- You need to control unequal variances

Caution: Transforming can make interpretation of the data difficult and results are not always symmetrical

The natural log the data from the top graph is shown on lower graph



Transformations

Logarithmic	1. Data skewed to the right	$\log_{10}(X)$	$\ln(X)$
	2. if values are <1	$\log_{10}(X+1)$	$\ln(X+1)$
Square root	Data are counts skewed to right	SQRT(X+0.05)	
		SQRT(X) + SQRT(X+1)	
Power	Data skewed to left -or-	X to a power	
	SD decreases with increasing X		

Decide to transform before you start the analysis based on the data (don't fish for significance)

Back transforming a mean of a log is not meaningful.

- Report both mean of raw data and mean of transformed data
- Report results from statistical tests on transformed data

State in methods you transformed data for the analyses to meet assumption of normality (and/or homoscedasticity)

What about negative numbers?
A log transformation will not work on negative numbers

add a constant to the data before applying the transformation so that after adding the constant all your data is greater than zero.

x_transformed = log(x + C)

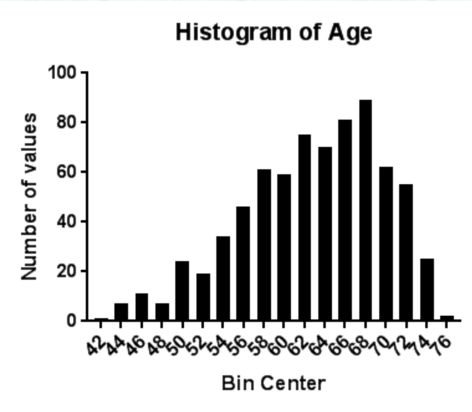
where C is a constant that allows x+C to be greater than zero.

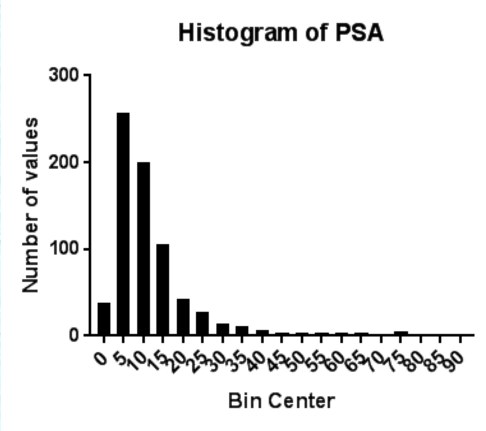
e.g., C = 1 – the smallest x value

Let’s Play…

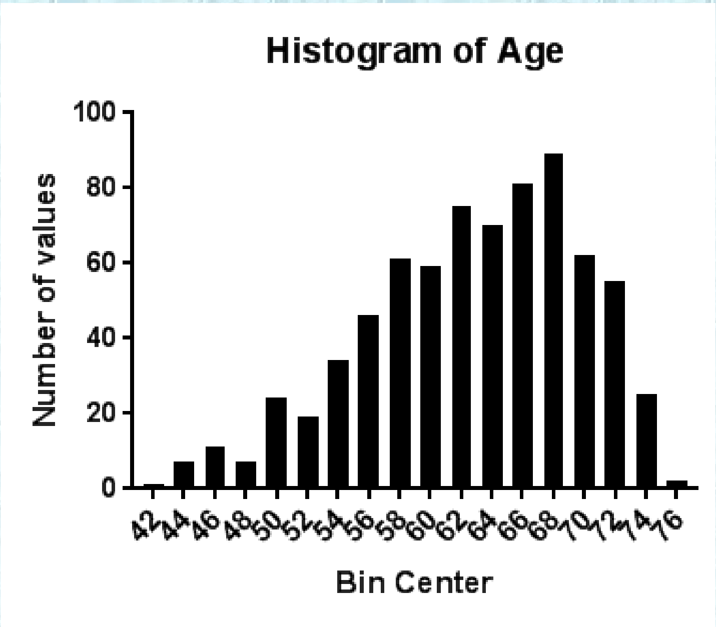
ProstateCancerData2019.xlsx

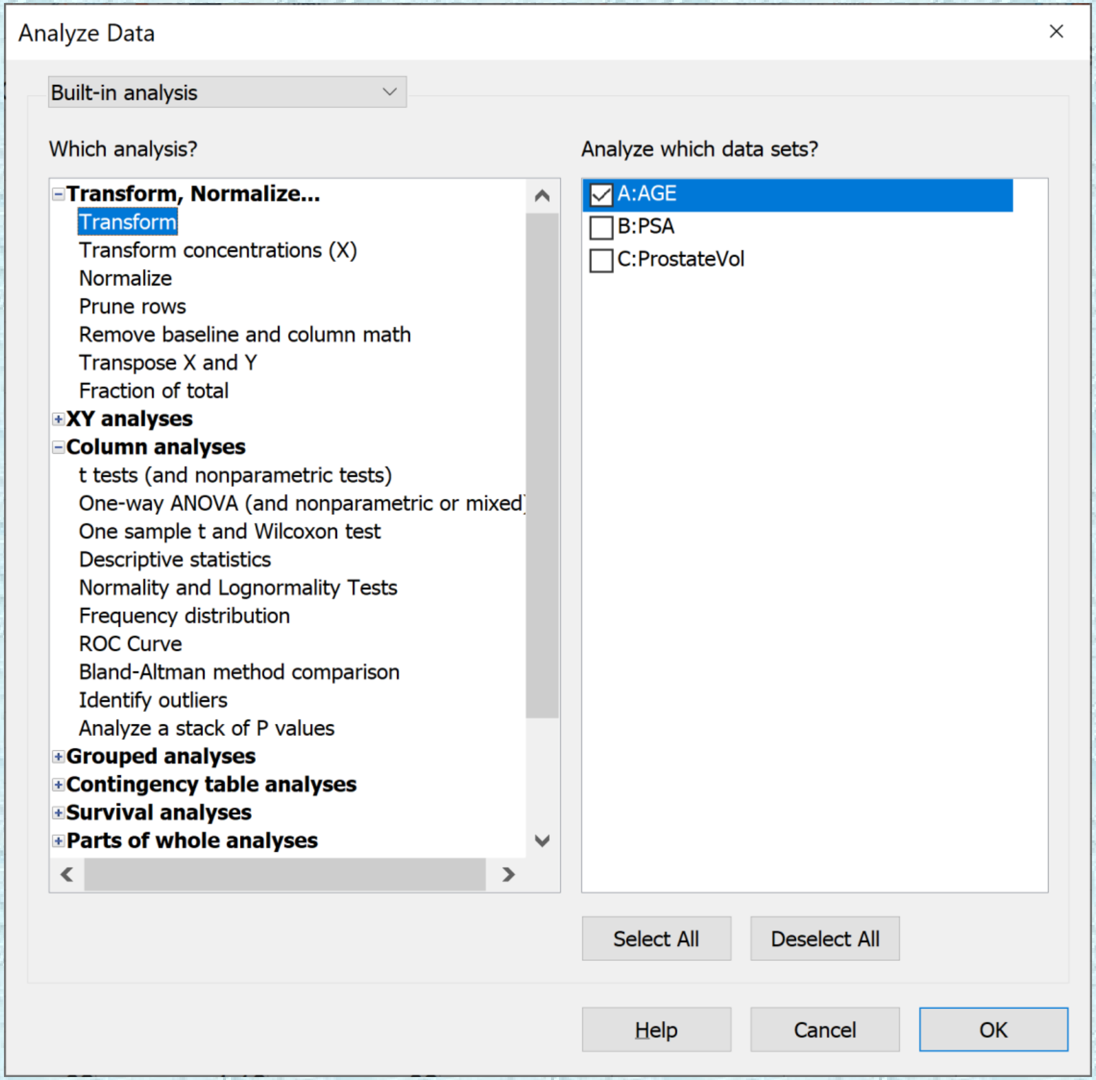
Variables for Age, PSA, Prostate volume
(we will do Prostate volume in class)

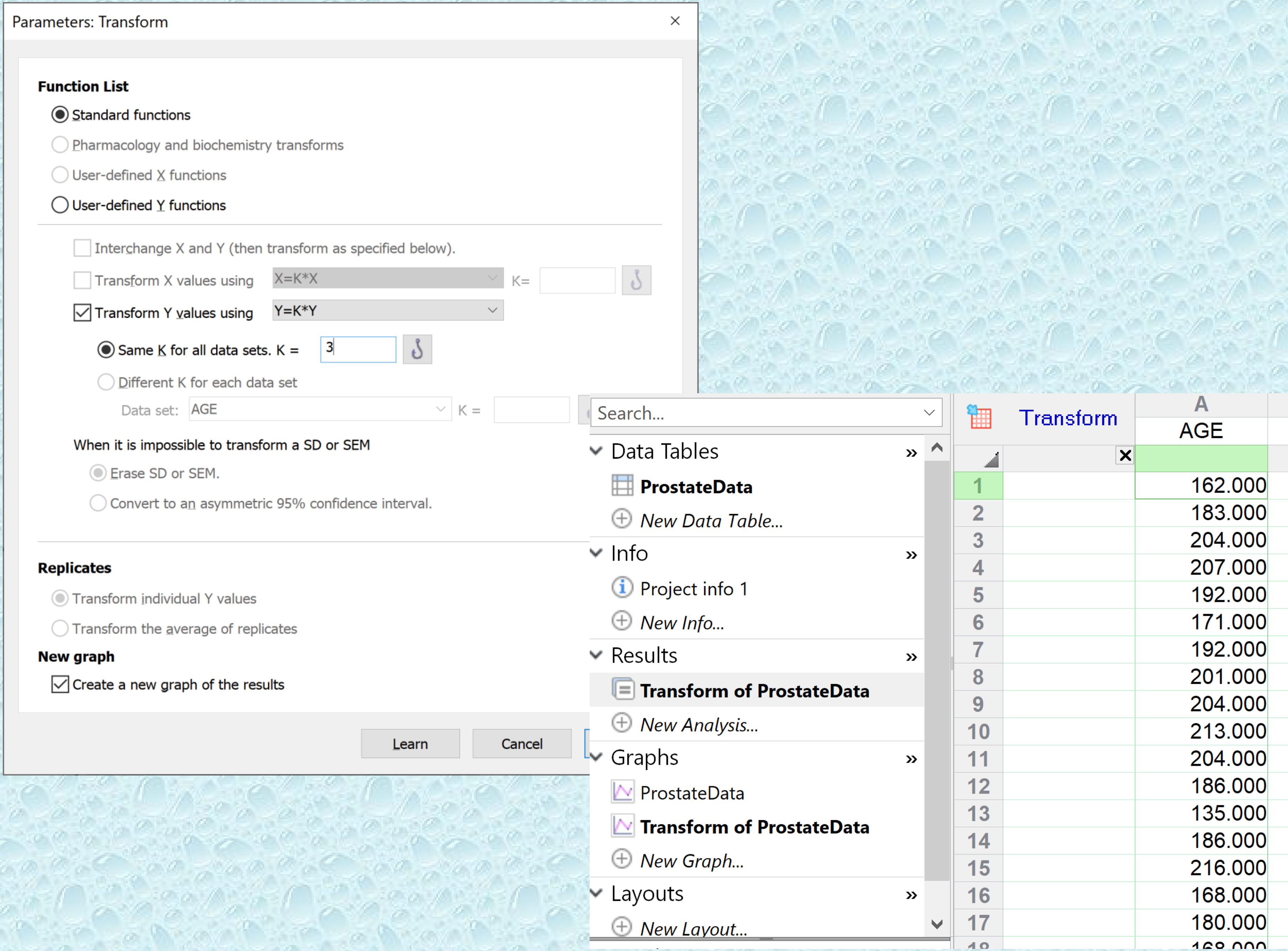


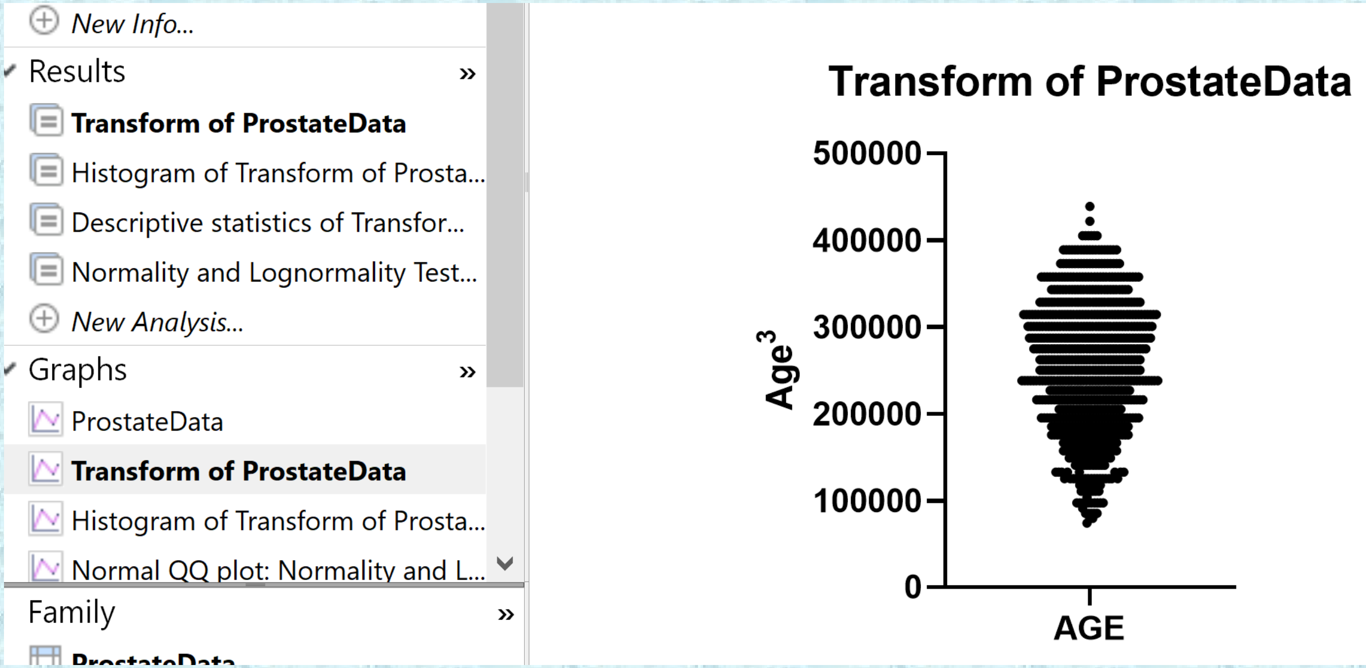


For age with left skew, raise the value to a power.
In this case we cubed it or raised it to 3



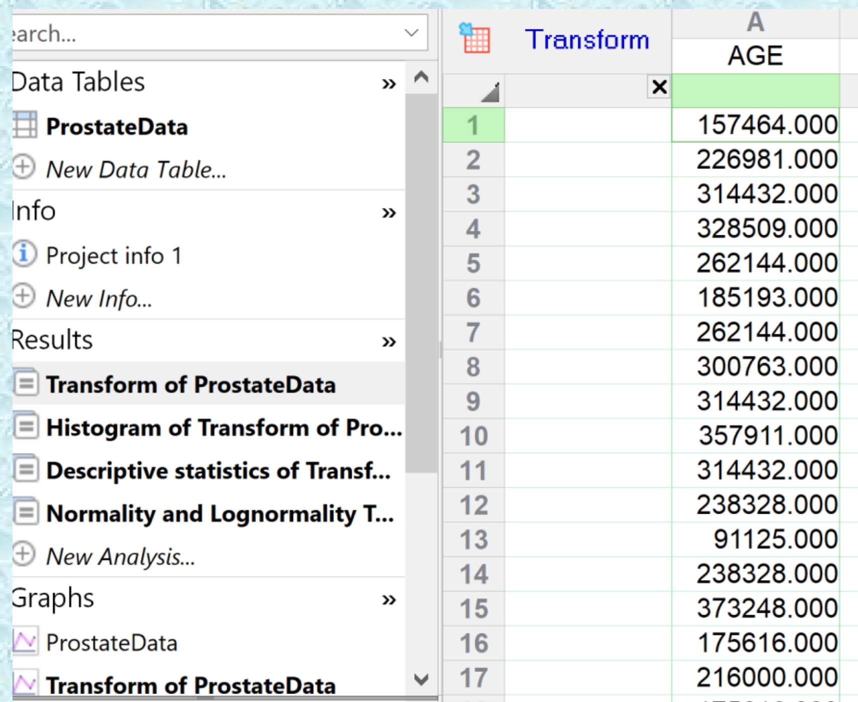


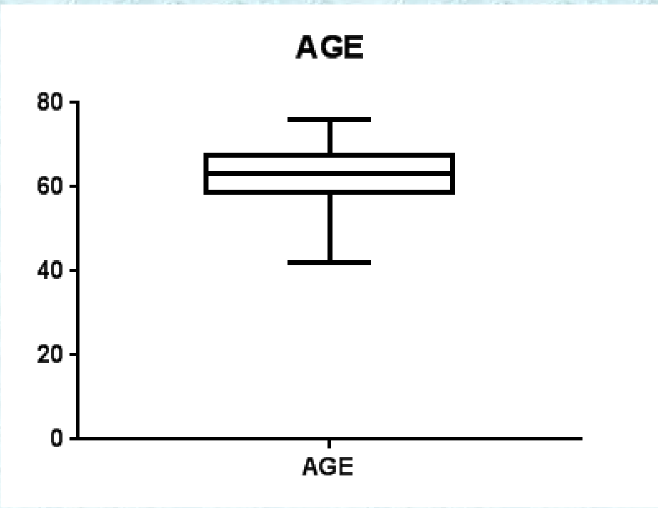

Parameters: Transform
Function List
Standard functions
Pharmacology and biochemistry transforms
User-defined X functions
User-defined Y functions
Interchange X and Y (then transform as specified below).
Transform X values using
X=K*X
K=
Transform Y values using
Y=K*Y
Same K for all data sets. K =
3
Different K for each data set
Data set:
AGE
K =
When it is impossible to transform a SD or SEM
Erase SD or SEM.
Convert to an asymmetric 95% confidence interval.
Replicates
Transform individual Y values
Transform the average of replicates
New graph
Create a new graph of the results
Learn
Cancel
Search...
Data Tables
ProstateData
New Data Table...
Info
Project info 1
New Info...
Results
Transform of ProstateData
New Analysis...
Graphs
ProstateData
Transform of ProstateData
New Graph...
Layouts
New Layout...
Transform
A
AGE
1 162.000
2 183.000
3 204.000
4 207.000
5 192.000
6 171.000
7 192.000
8 201.000
9 204.000
10 213.000
11 204.000
12 186.000
13 135.000
14 186.000
15 216.000
16 168.000
17 180.000

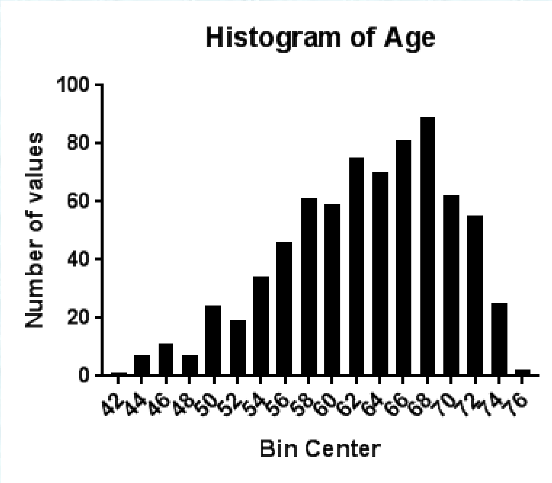


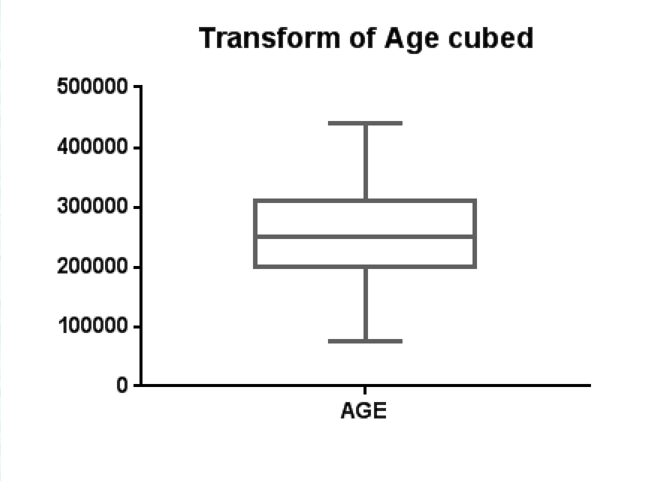
Create the frequency histogram from the transformed data in the Results page

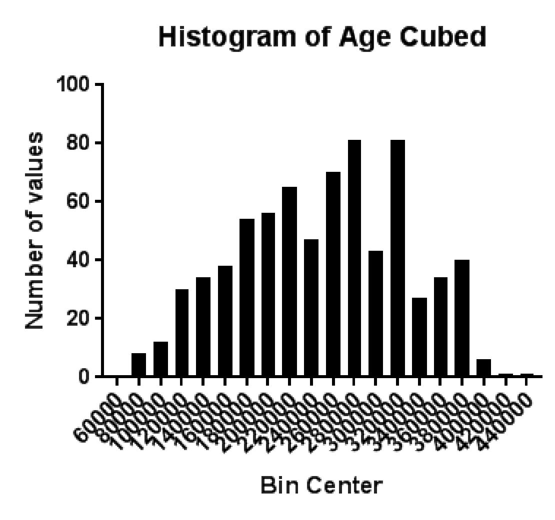
Also use this data to create descriptive and do normality tests




AGE
80
60
40
20
0
AGE


Histogram of Age
Number of values
100
80
60
40
20
0
42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76
Bin Center


Transform of Age cubed
500000
400000
300000
200000
100000
0
AGE


Histogram of Age Cubed
Number of values
100
80
60
40
20
0
Bin Center

		A
		AGE
1	Number of values	728
2		
3	Minimum	42.00
4	25% Percentile	58.00
5	Median	63.00
6	75% Percentile	68.00
7	Maximum	76.00
8	Range	34.00
9		
10	Mean	62.31
11	Std. Deviation	6.871
12	Std. Error of Mean	0.2547
13		
14	Skewness	-0.5250
15	Kurtosis	-0.2938

Transformed Data

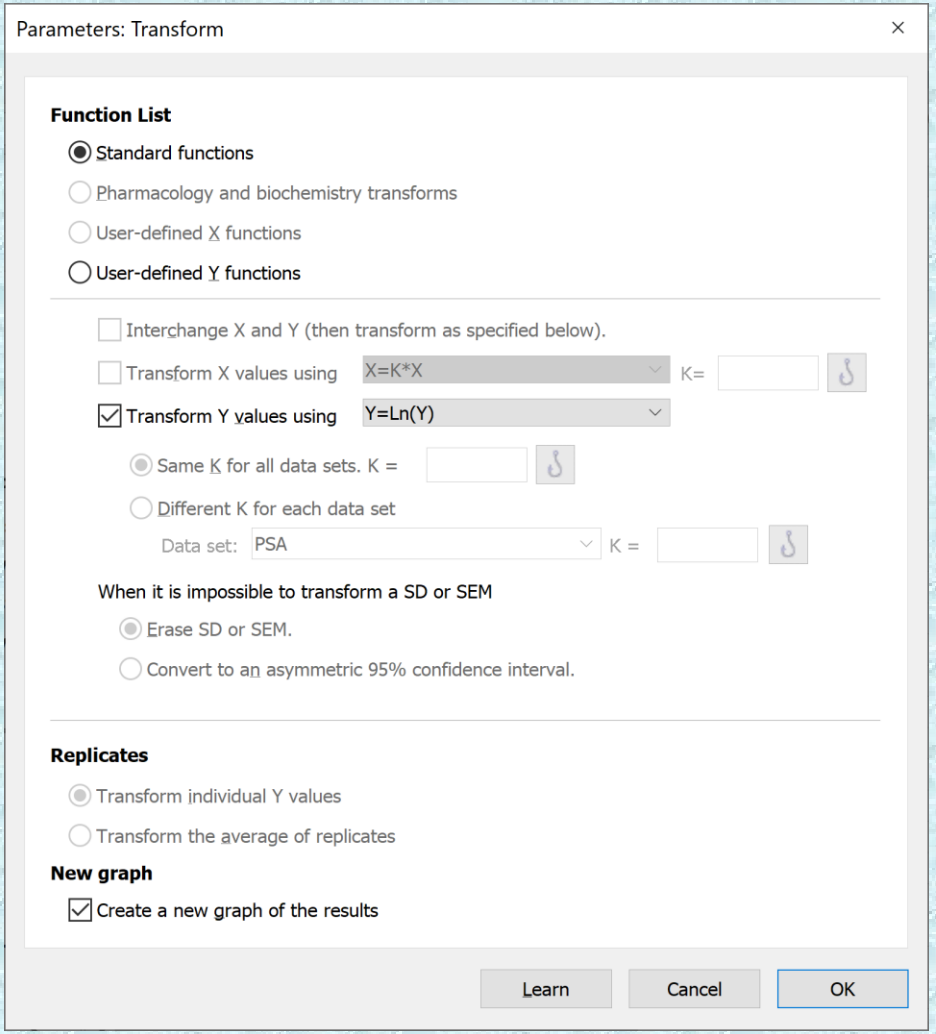
		A
		AGE
1	Number of values	728
2		
3	Minimum	74088
4	25% Percentile	195112
5	Median	250047
6	75% Percentile	314432
7	Maximum	438976
8	Range	364888
9		
10	Mean	250587
11	Std. Deviation	76958
12	Std. Error of Mean	2852
13		
14	Skewness	-0.07325
15	Kurtosis	-0.7591
16		

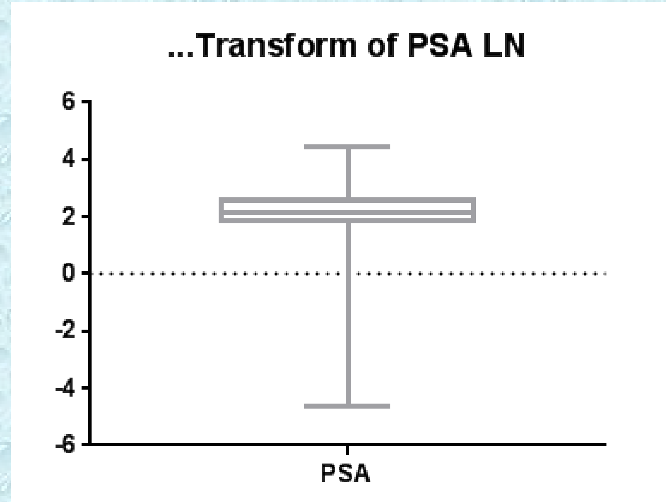
Transformed Data

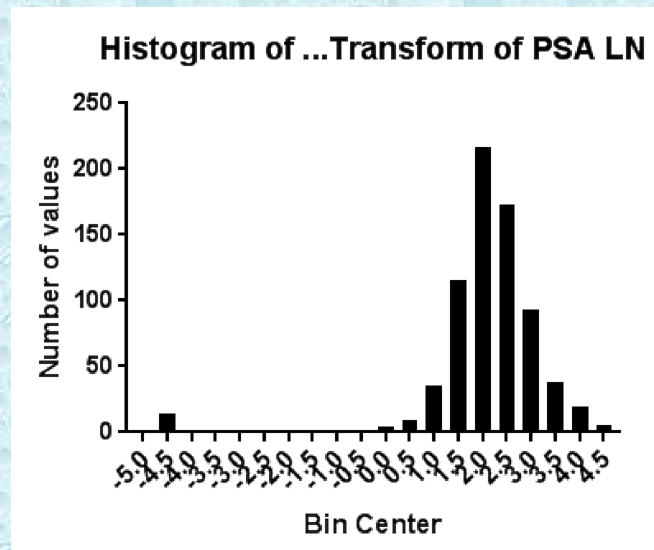
	Normality and Lognormality Tests Tabular results	A AGE
1	**Test for normal distribution**	
2	**Anderson-Darling test**	
3	A2*	5.672
4	P value	<0.0001
5	Passed normality test (alpha=0.05)?	No
6	P value summary	****
7		
8	**D'Agostino & Pearson test**	
9	K2	33.60
10	P value	<0.0001
11	Passed normality test (alpha=0.05)?	No
12	P value summary	****
13		
14	**Shapiro-Wilk test**	
15	W	0.9697
16	P value	<0.0001
17	Passed normality test (alpha=0.05)?	No
18	P value summary	****
19		
20	**Number of values**	728

	Normality and Lognormality Tests Tabular results	A AGE
1	**Test for normal distribution**	
2	**Anderson-Darling test**	
3	A2*	2.682
4	P value	<0.0001
5	Passed normality test (alpha=0.05)?	No
6	P value summary	****
7		
8	**D'Agostino & Pearson test**	
9	K2	49.32
10	P value	<0.0001
11	Passed normality test (alpha=0.05)?	No
12	P value summary	****
13		
14	**Shapiro-Wilk test**	
15	W	0.9851
16	P value	<0.0001
17	Passed normality test (alpha=0.05)?	No
18	P value summary	****
19		
20	**Number of values**	728

For PSA with right skew, take the logarithm. In this case we will use the natural log .





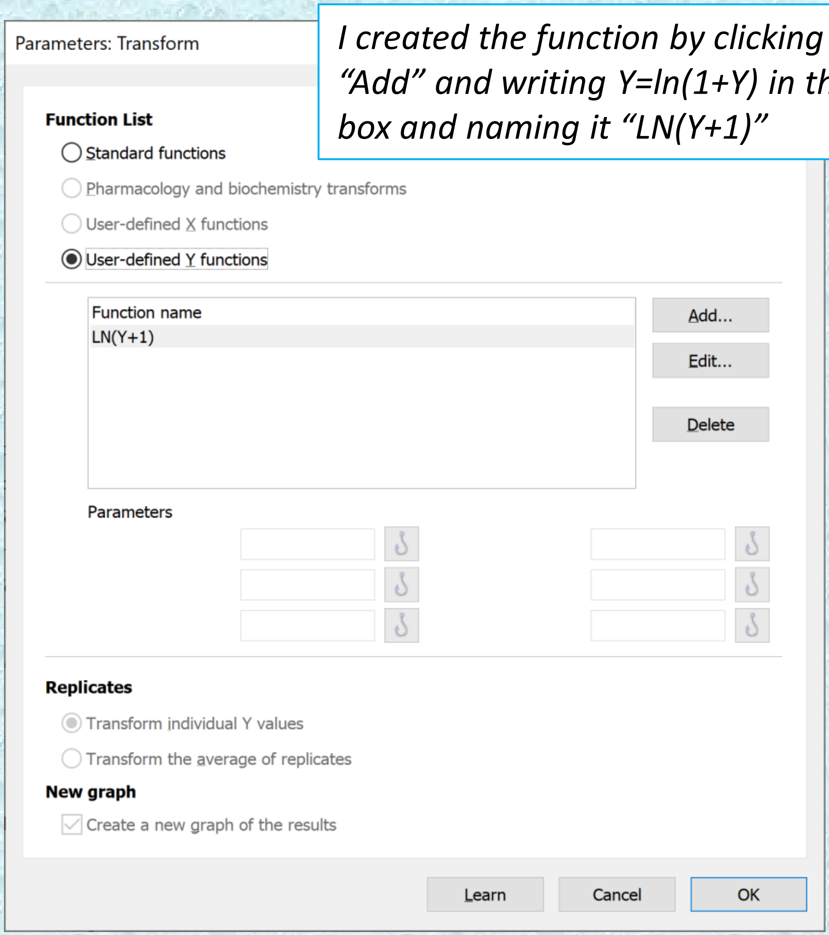


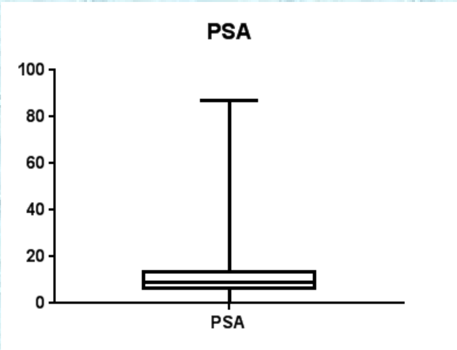
OOPS, the PSA includes values <1.0.

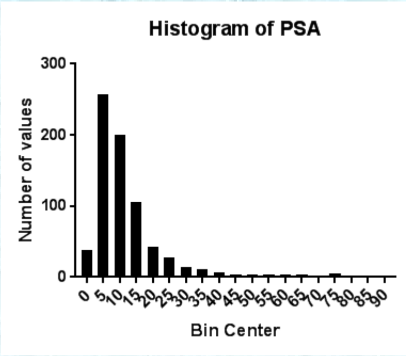
For PSA with right skew, take the logarithm.
In this case we will use the natural log +1.

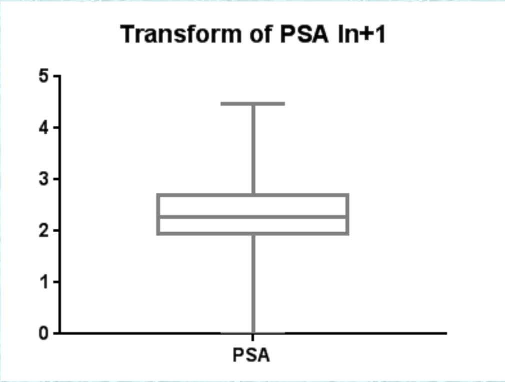
Transformed Y=ln(1+Y)

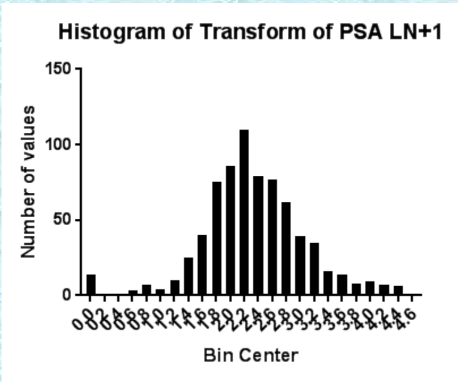
I created the function by clicking "Add" and writing Y=ln(1+Y) in the box and naming it "LN(Y+1)"







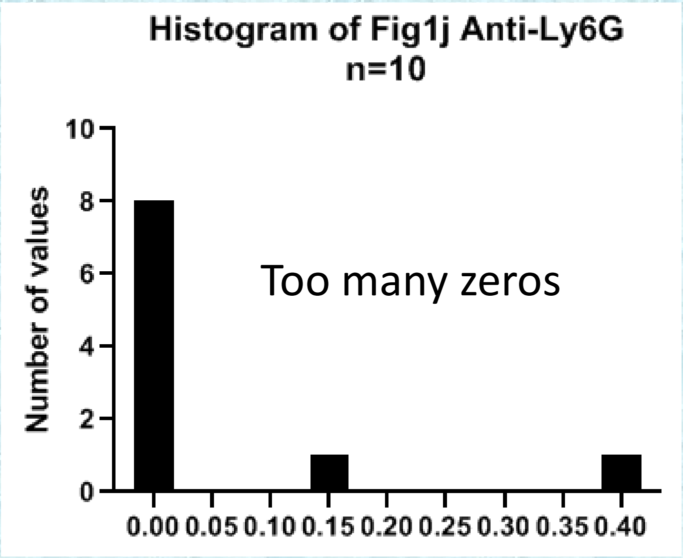


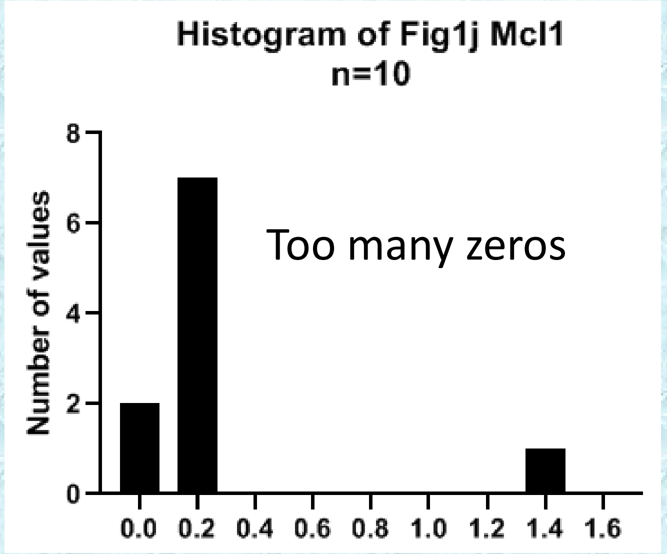


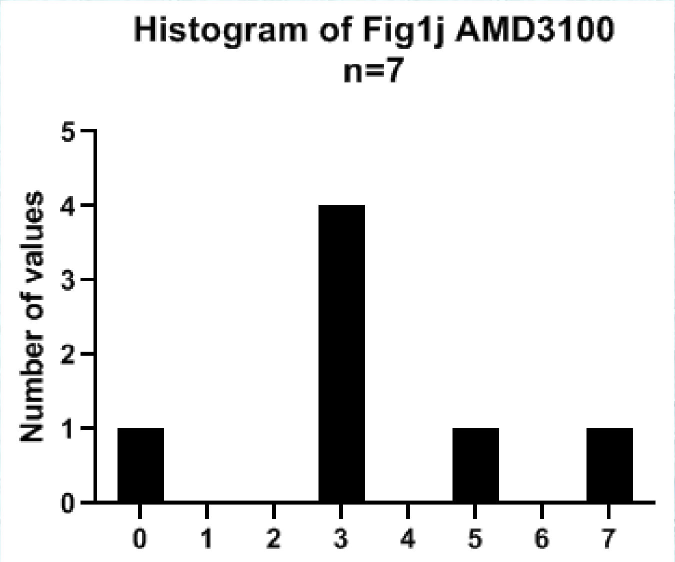
What transformation would you use for prostate volume?

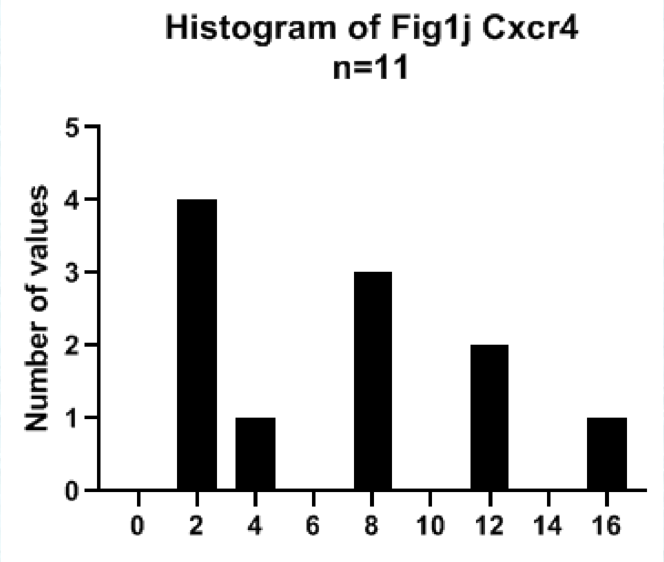
<paste frequency histogram here>

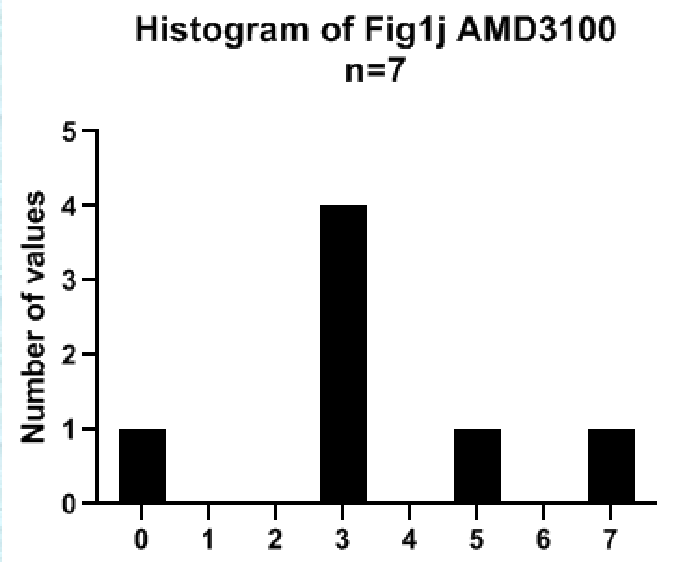
Remember these graphs from Monday? Can transformation help?

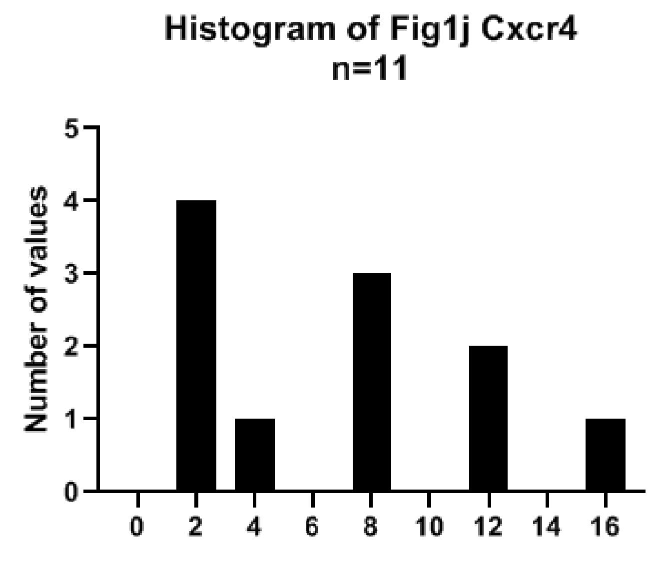


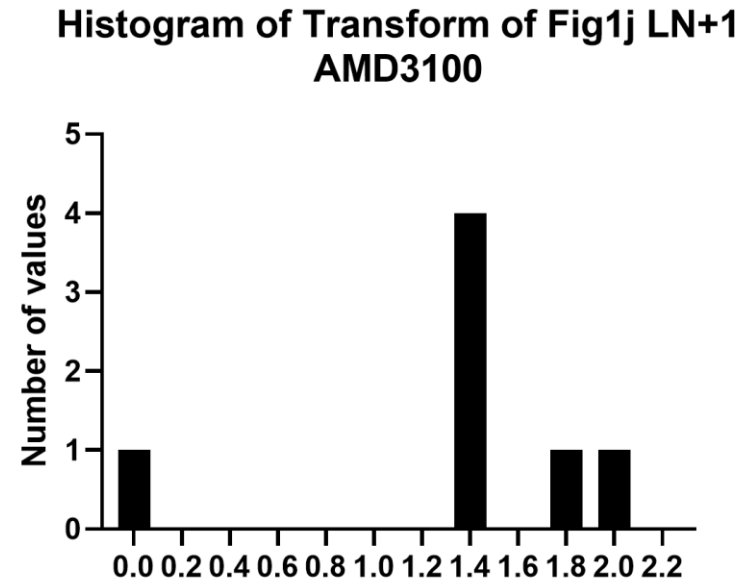


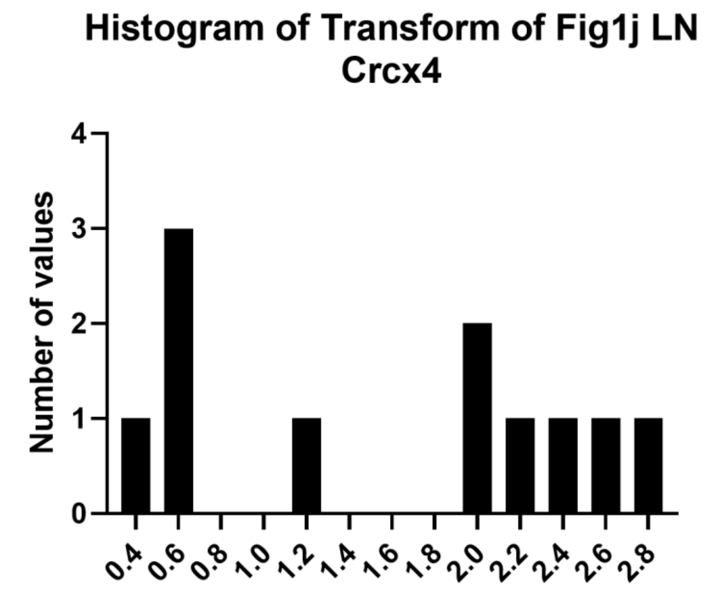





Histogram of Fig1j AMD3100
n=7
Number of values
5
4
3
2
1
0
0
1
2
3
4
5
6
7


Histogram of Fig1j Cxcr4
n=11
Number of values
5
4
3
2
1
0
0
2
4
6
8
10
12
14
16

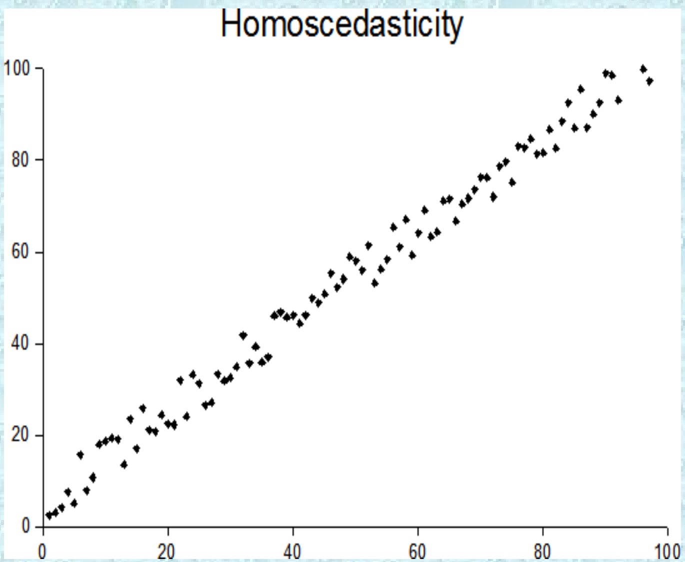

Histogram of Transform of Fig1j LN+1
AMD3100
Number of values
5
4
3
2
1
0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2

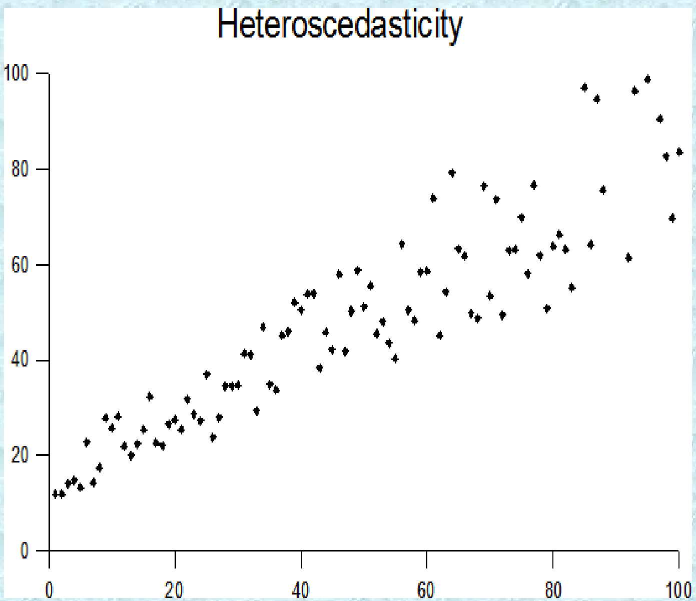

Histogram of Transform of Fig1j LN
Crcx4
Number of values
4
3
2
1
0
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
2.6
2.8

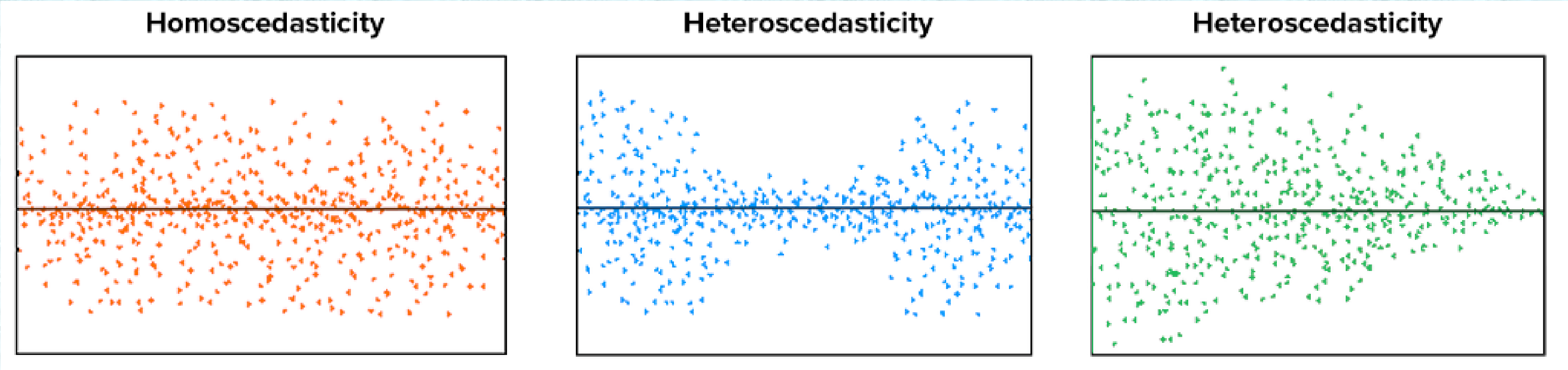
		A	B
		AMD3100	Cxcr4
1	Number of values	7	11
2			
3	Minimum	0.000	1.514
4	25% Percentile	3.358	1.810
5	Median	3.458	7.118
6	75% Percentile	5.293	11.85
7	Maximum	6.609	15.37
8	Range	6.609	13.85
9			
10	Mean	3.654	6.652
11	Std. Deviation	2.041	4.995
12	Std. Error of Mean	0.7715	1.506
13			
14	Skewness	-0.5111	0.4753
15	Kurtosis	1.636	-1.207

		A	A
		AMD3100	Cxcr4
1	Number of values	7	11
2			
3	Minimum	0.000	0.4148
4	25% Percentile	1.472	0.5933
5	Median	1.495	1.963
6	75% Percentile	1.839	2.473
7	Maximum	2.029	2.732
8	Range	2.029	2.317
9			
10	Mean	1.402	1.568
11	Std. Deviation	0.6556	0.9006
12	Std. Error of Mean	0.2478	0.2716
13			
14	Skewness	-2.004	-0.1348
15	Kurtosis	4.836	-1.938

Enough about normality and data transformations, let's move on to the assumption of homoscedasticity





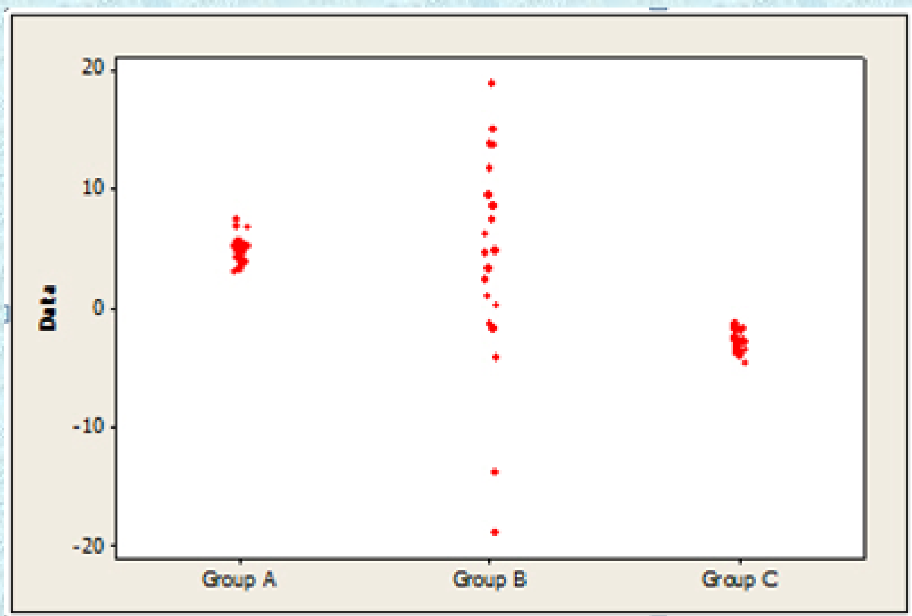


Homoscedasticity = having the same scatter

homo → *homo* - same

scedastic → *skedannýnai* – to scatter

Homoscedasticity, equal variances, homogeneity of variance—they're all saying "same scatter."



Which group(s) demonstrates homoscedasticity?

Which group(s) demonstrates heteroscedasticity?

Heteroscedasticity

Unequal variances are usually only an issue when you are comparing a continuous variable in two or more groups or groups over repeated time points

Independent t-tests
ANOVA
Repeated measures ANOVA

What happens if I ignore heteroscedasticity?

There is an increased probability you will conclude something is different when it really isn't (biased result)

A statistician did a data experiment using 3 populations with **the same mean**:

He generated thousands of random samples of n=10 observations from population A, n=7 from population B, and n=3 from population C (note: unequal sample sizes)

When the three populations were homoscedastic, the one-way ANOVA tests were significant (p<0.05) in about 5% of the simulations

When the standard deviations were different (1.0 for population A, 2.0 for population B, and 3.0 for population C), tests were significant about 18% of the time.

Even though the population means were really all the same, the probability of getting a false positive result was 18%, not the "desired" 5%.

Your statistics prof will be unhappy....

http://www.biostathandbook.com/homoscedasticity.html

How to Assess Heteroscedasticity

Compare the standard deviations of different groups of measurements, to see if they are very different from each other.

Graphing data or residuals

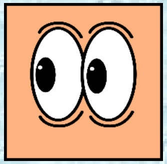
Formal tests

Assessing homoscedasticity

Similar to looking at mean and median to assess normality, you can use the standard deviations of your groups to assess homoscedasticity.

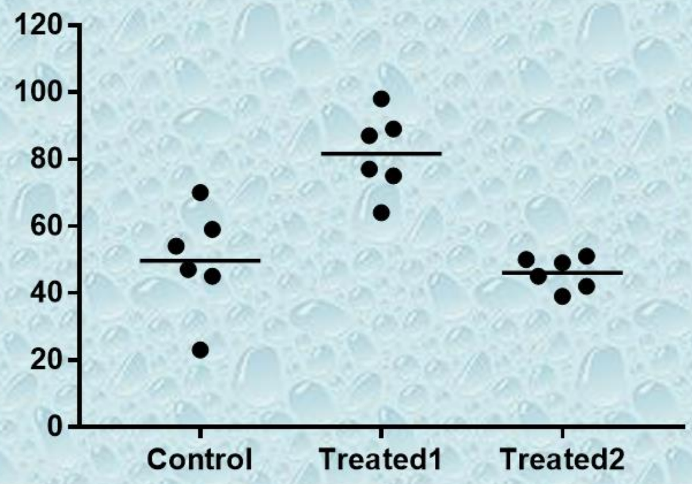
A RULE OF THUMB: If sample sizes are equal, t-tests (and ANOVA) are robust to heterogeneity of variance, provided the **ratio of largest to smallest SD is <2 times**

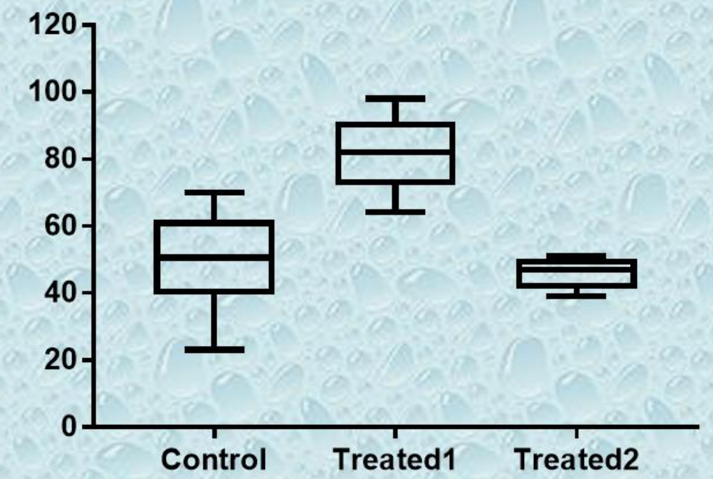
Look at column statistics in Prism



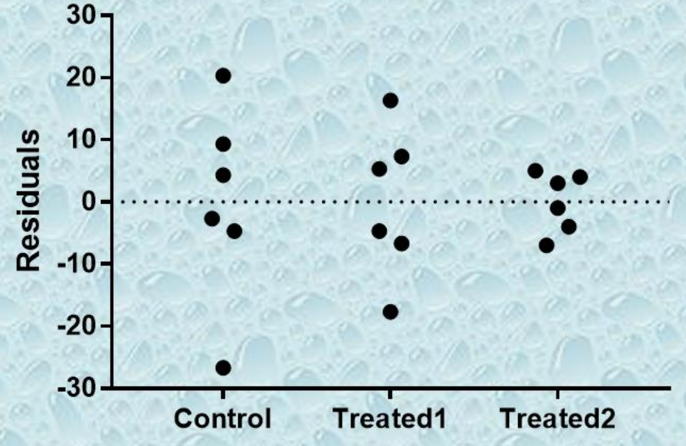
Graphic options

Plot data



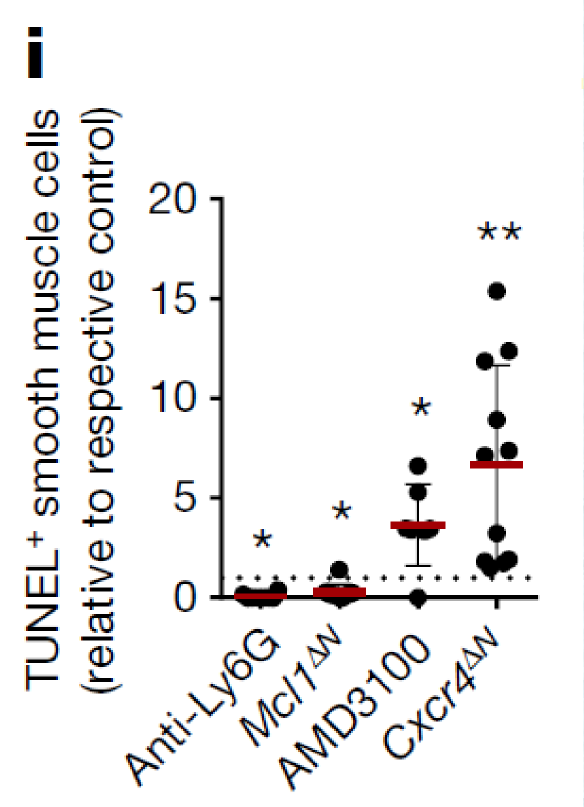


Plot residuals



We will discuss residuals when we talk about t-tests and ANOVA

		A	B	C	D
		Anti-Ly6G	Mcl1	AMD3100	Cxcr4
1	Number of values	10	10	7	11
2					
3	Minimum	0.000	0.000	0.000	1.514
4	25% Percentile	0.000	0.08850	3.358	1.810
5	Median	0.000	0.1590	3.458	7.118
6	75% Percentile	0.04025	0.2530	5.293	11.85
7	Maximum	0.3990	1.403	6.609	15.37
8	Range	0.3990	1.403	6.609	13.85
9					
10	Mean	0.05600	0.2684	3.654	6.652
11	Std. Deviation	0.1307	0.4089	2.041	4.995
12	Std. Error of Mean	0.04133	0.1293	0.7715	1.506
13					
14	Skewness	2.494	2.870	-0.5111	0.4753
15	Kurtosis	6.160	8.688	1.636	-1.207



4.99/0.13 = 38.4 >2

Formal tests for equal variances

Like tests for normality, there are tests to assess homoscedasticity

For the independent t test, GraphPad Prism uses an F test.

For the one-way ANOVA, Prism uses the Brown-Forsythe test and also (if every group has at least five values) the Bartlett's test.

Tests for equal variances have the same issues as tests for normality

- Too much power at large sample sizes
  - Too easy to detect small differences in variances
- Too little power with small sample sizes
  - Too easy to fail to detect differences

So they really have limited usefulness

What to do about heteroscedasticity?

Transform the data

A log transformation will reduce data variation

Use a non-parametric test

These tests still a bit biased if the variation is large between groups

We will talk more about testing for homoscedasticity when we talk about t-tests and ANOVA

The Good News

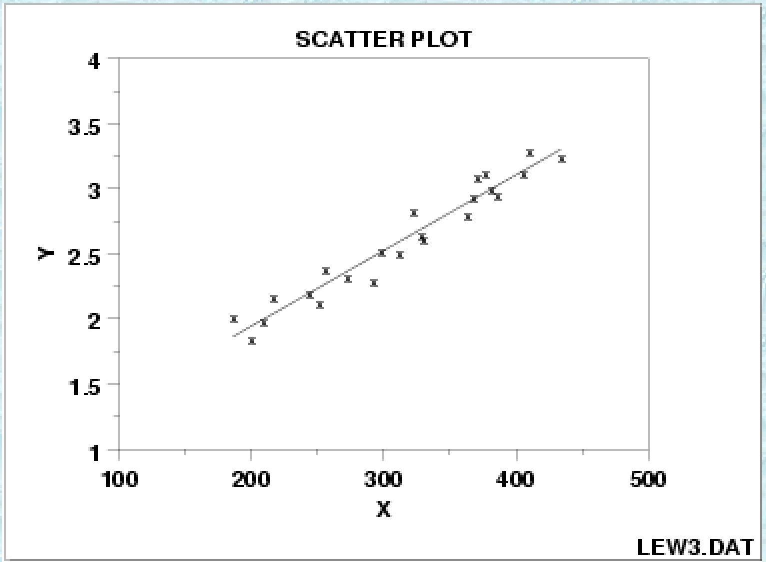
A lot of parametric tests are robust to violating homoscedasticity if data are balanced (same sample size in each group) and the difference in variation is not too large (<2x)

Assumption of linearity

Linearity refers to a linear, or straight line, relationship between two variables

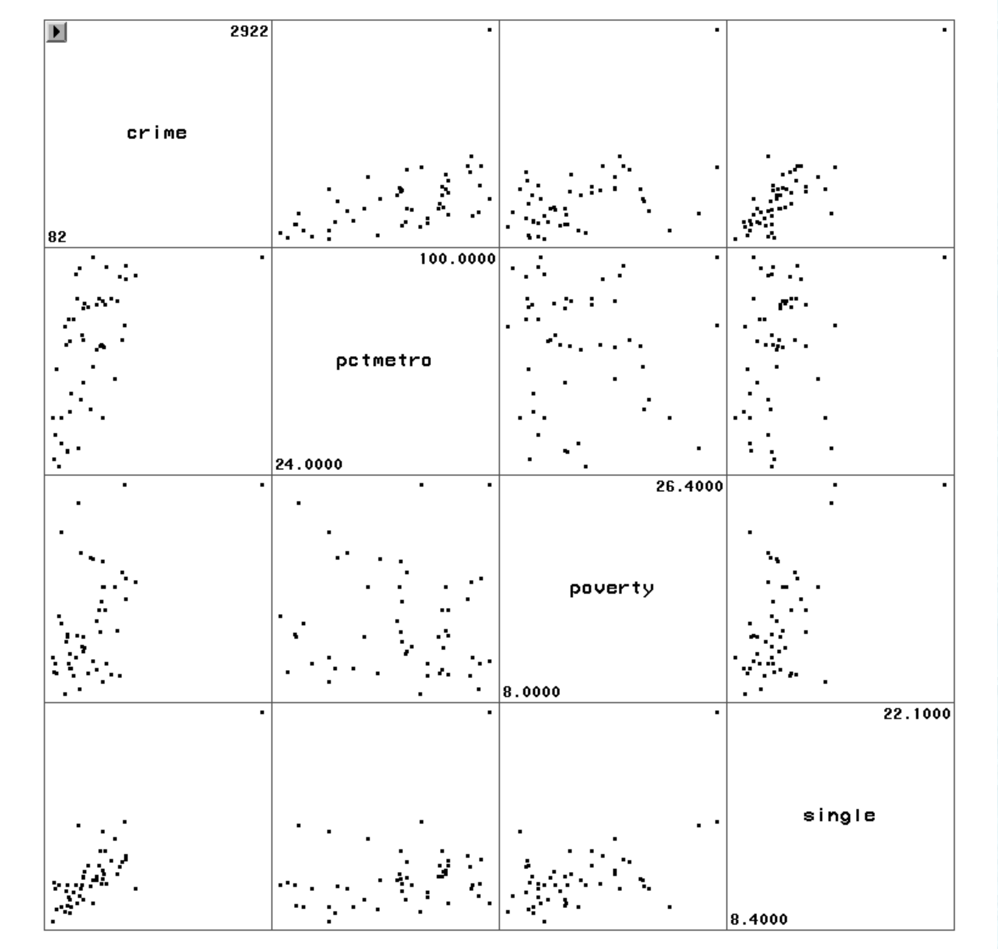
If the assumption of linearity is not met, then results may be inaccurate (biased)

Linearity is typically important in Pearson correlation analyses and regression analyses



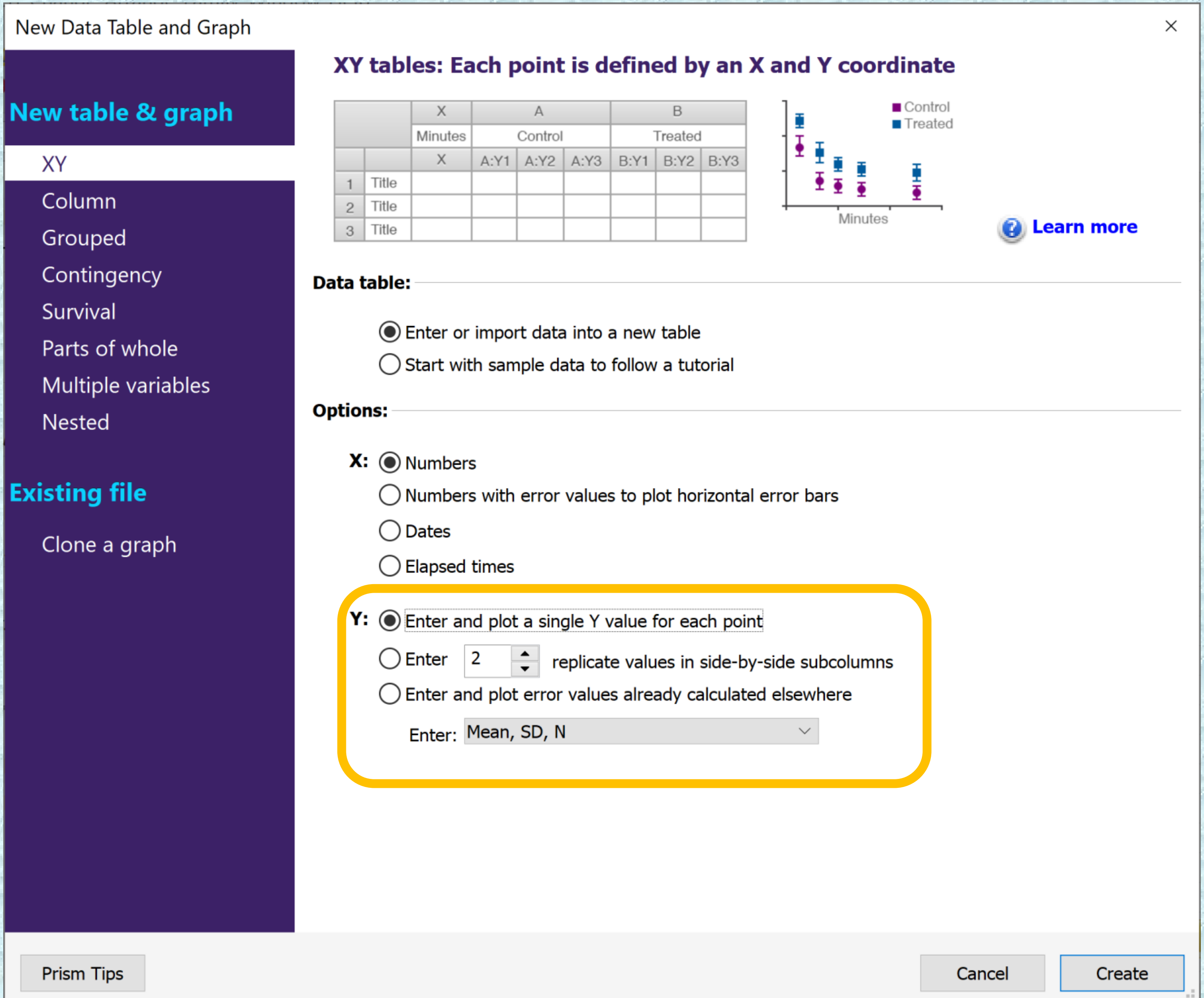
Testing for linearity

The linearity assumption can best be tested with scatter plots of the data or residual plots

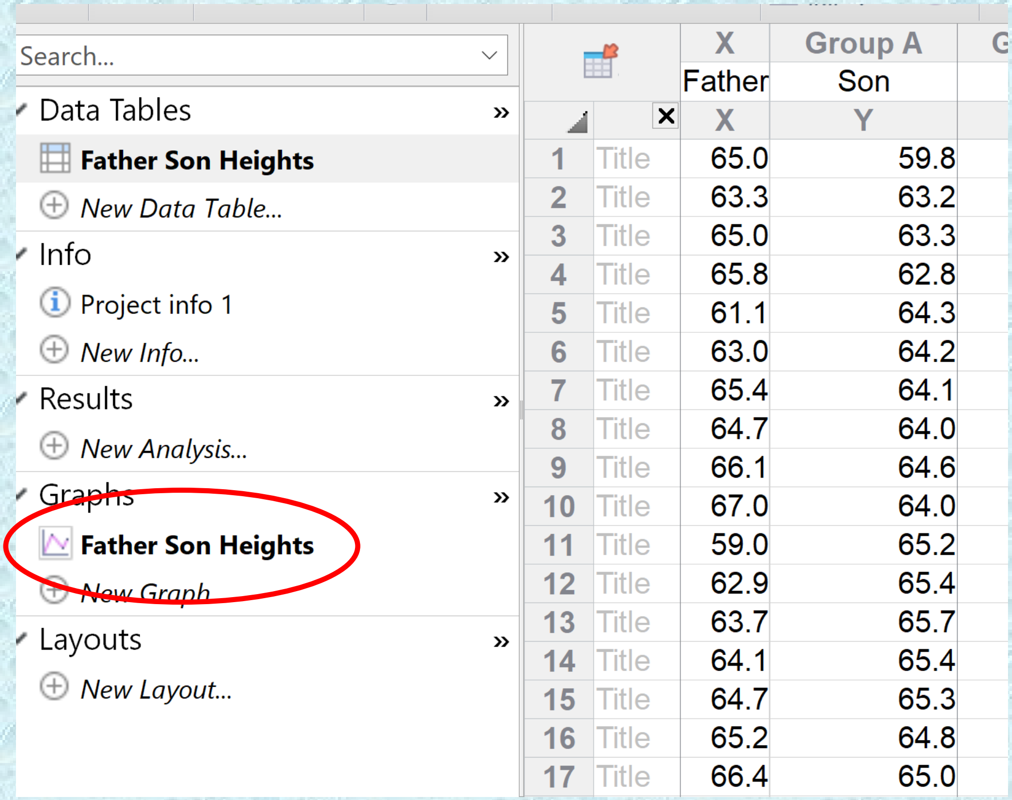


https://stats.idre.ucla.edu/sas/webbooks/reg/chapter2/regressionwith-saschapter-2-regression-diagnostics/

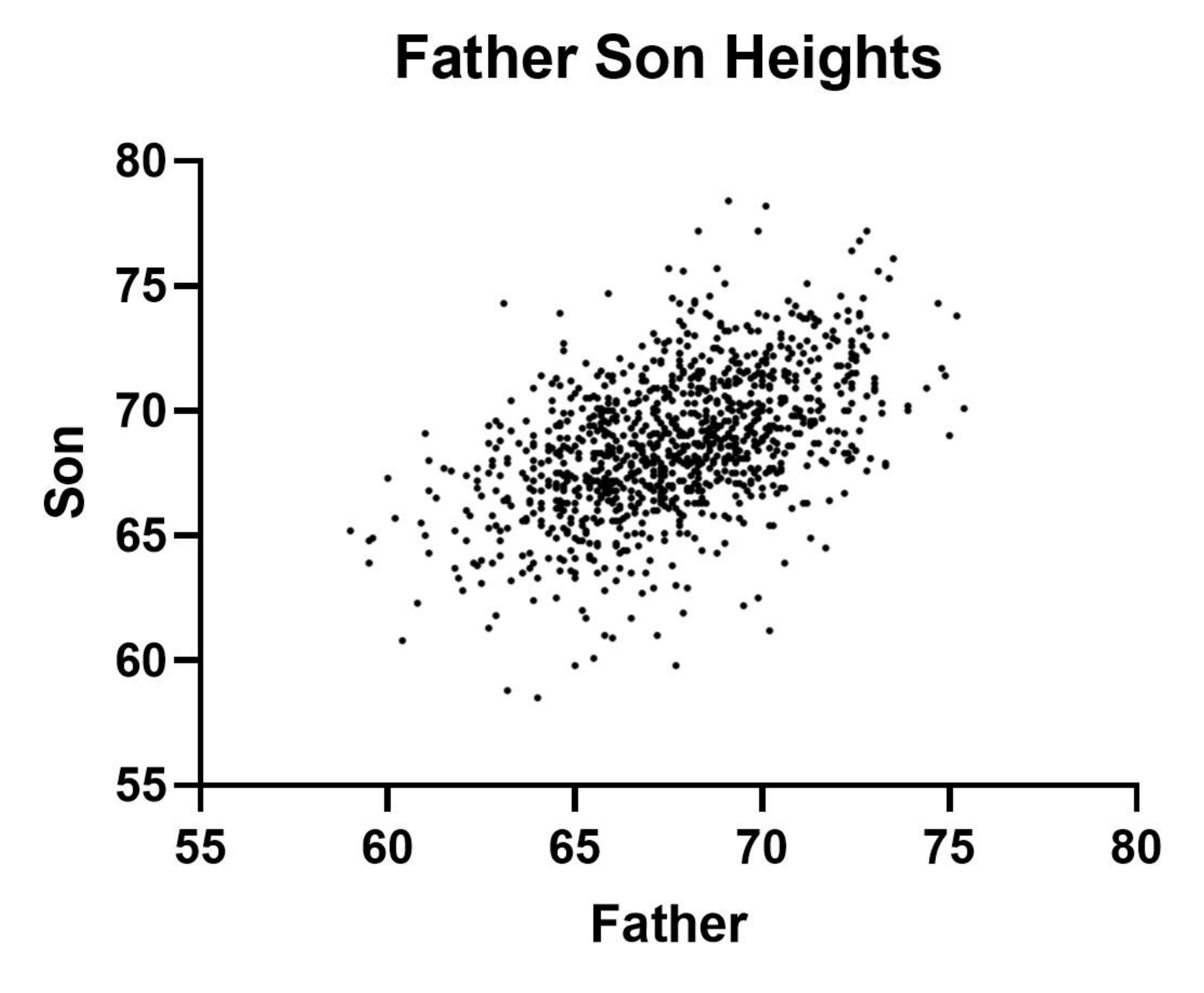
Graphing a scatter plot between two continuous variables



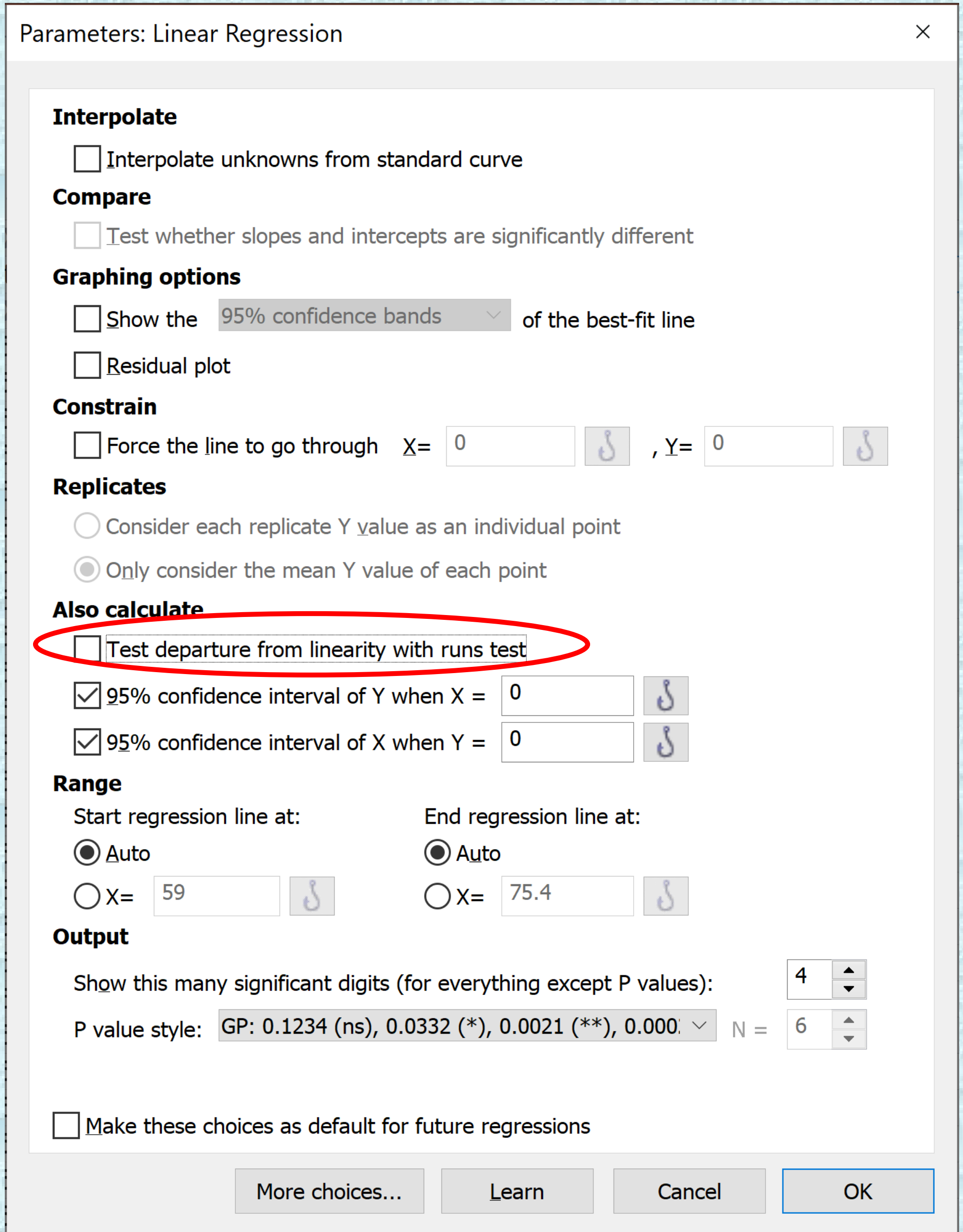
Data structure



Choose graph in “Graphs”


Father Son Heights
80
75
70
65
60
55
Son
55
60
65
70
75
80
Father

Run's test for linearity in linear regression analysis



Testing for Linearity – The Runs Test in Prism

Asks if the line fit by linear regression deviates from the data.

Runs test is impossible for some data sets when several rows have the same X value

The runs test is rarely found in the biomedical literature

We will not rely on it

Remedies if data are not linear?

Data transformation

Non-linear regression

The Assumption of Independence

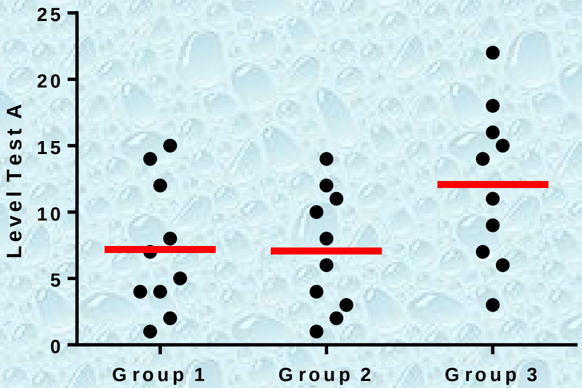
INDEPENDENCE

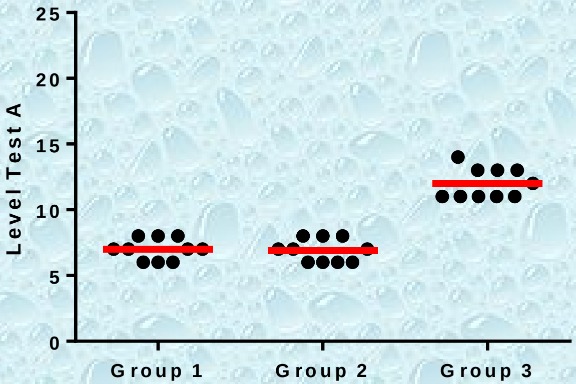
One of the common assumptions of statistical tests

- a study subject is independent from other subject
- measurements within a subject are independent

If two samples are not correlated or one sample does not influence the value of the other sample, then they are probably independent

When observations are not independent, the dependency between them has the potential to decrease data variability within that group and can lead to false positive conclusions





Independence

A function of the study design

The key to avoiding violating the assumption of independence is to make sure your data is independent *while you are collecting it*.

Data can be independent or paired

- Different statistical tests are used for each
  - independent t-test (unpaired t-test)
  - paired t-test

Dependent/paired samples
Samples are correlated in some fashion

pre-test/post-test samples in which a variable is measured before and after an intervention in the same mouse

matched samples in which mice are matched on characteristics such as age and sex

repeated measurements on the same biological samples or in the same mouse

technical replicates (a special case)

Sometimes using independent observations is not always the best approach

- Before and after studies
  - Measuring blood pressure in the same person before and after treatment
  - Each person acts as their own control
    - Reduces variability

Dependent or Independent?

You are doing a pilot study to get information about tumor growth in your new mouse model. You want to measure tumor growth in the flanks of mice over a five week period. You will use calipers to measure the dimensions of the tumors.

A. You measure the tumor size each week for 5 weeks in each of 6 mice.

B. You follow Mouse1 for one week then measure the size of the tumor. You follow Mouse2 for two weeks then measure the size of the tumor. You follow Mouse3 for three weeks … etc. Repeat for Mouse4 and Mouse5. You repeat this experiment 4 more times.

<u>Dependent or Independent?</u>

Tumor growth measurements for A and B

Dependent or Independent?

Samples of a cell line pipetted into three neighboring wells on the same 96-well plate

Samples of a cell line pipetted into one well on three different 96-well plates run on different days

Size of tumors injected unilaterally in two different mice

Size of tumors injected bilaterally in one mouse

Technical Replicates

Testing the same samples under identical conditions

- i.e., same sample in multiple wells of the same 96-well plate

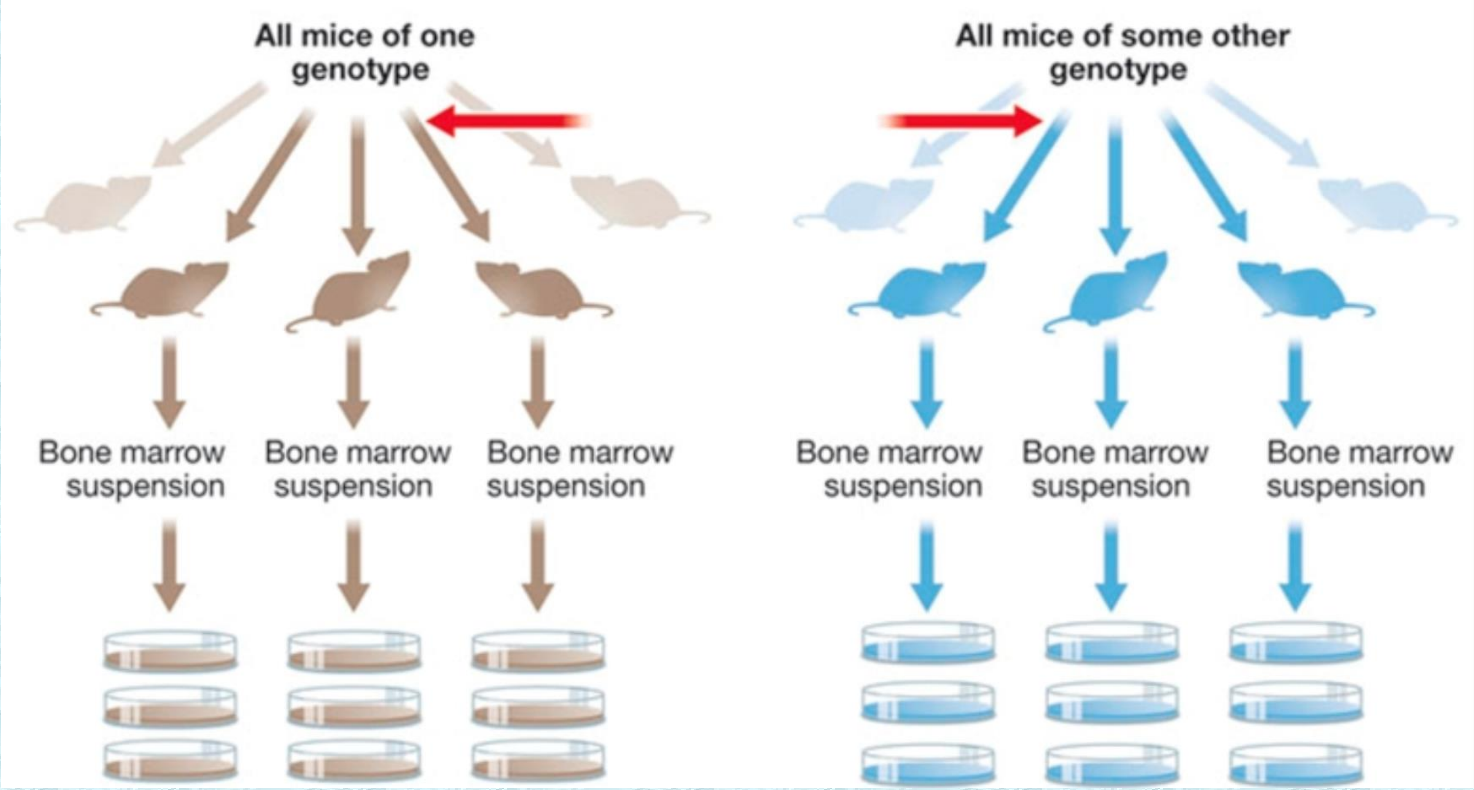
Crucial in laboratory experiments

- Reduce the effect of uncontrolled variation
- To assure that results are reliable and valid

Do not treat technical replicates as independent tests

Average for an n=1

a colony assay using bone marrow cells cultured in soft agar using triplicate plates



Three independent mice were chosen from each genotype, so we can make inferences about all mice of that genotype. Note that in the experiments, *n* = 6 (3 in each group), no matter how many replicate plates are created.

David L Vaux et al. EMBO Rep. 2012;13:291-296

Independent or paired?

Tumor biopsy samples were collected before the study and after 15 plus or minus 7 days of osimertinib treatment in 24 patients

Heart rates were measured in twelve healthy volunteers (six women; mean age: 28 years) who performed a treadmill protocol consisting of: five minutes sitting, five minutes standing, 10 minutes walking at 4 km/h.

Rates of alcoholism were studied in adult identical twins who were separated at birth and grew up in different family environments.

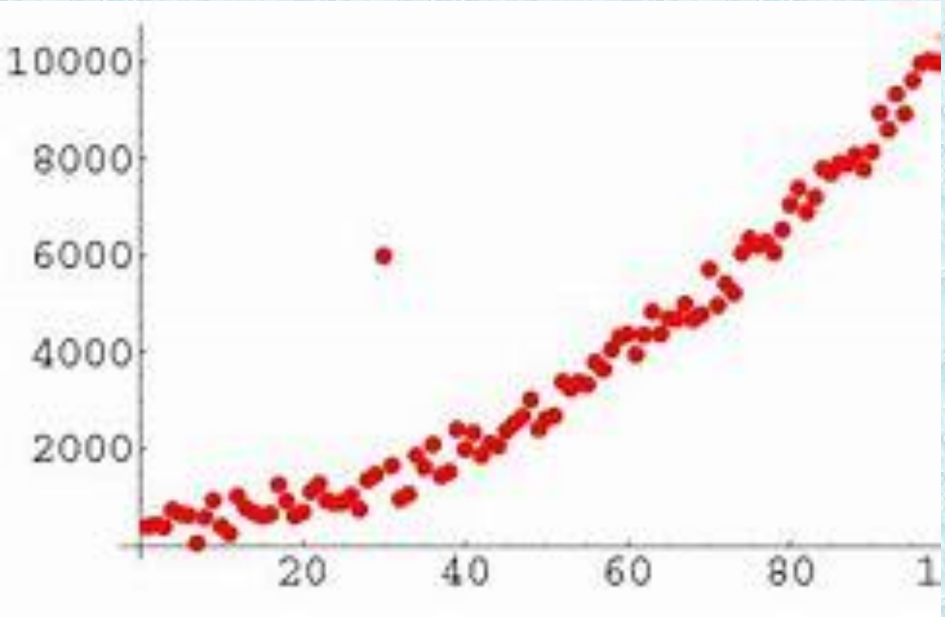
What happens if I ignore independence?

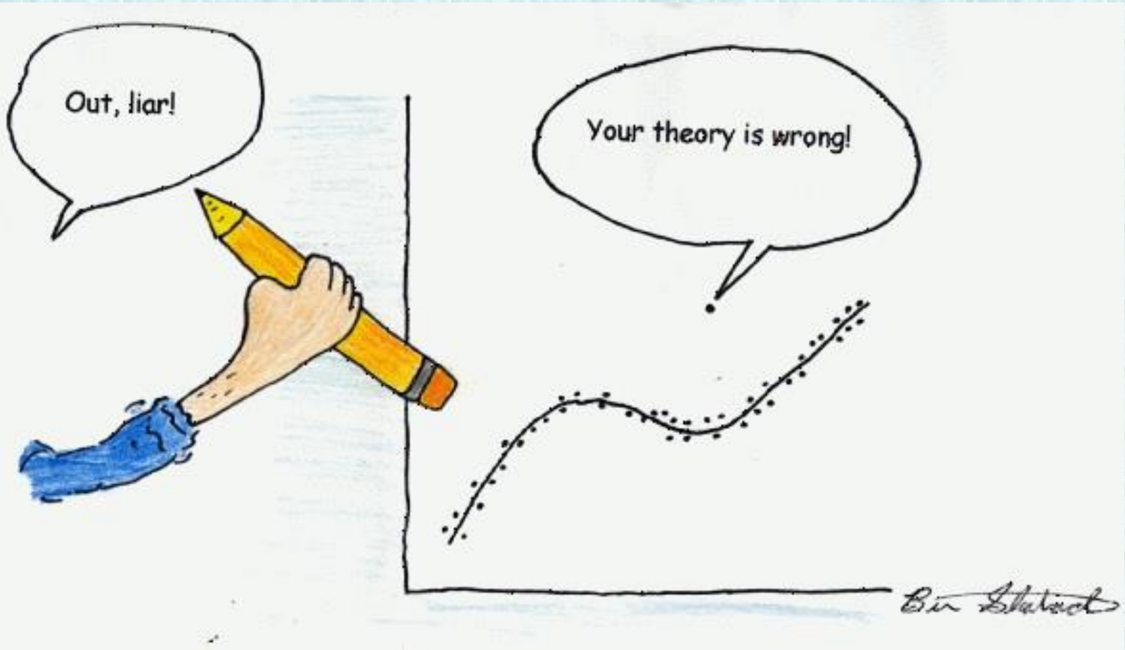
Dependence in data can turn into heavily biased results

Non-independent observations can make your statistical test give too many false positives due to reduced variability.

Your statistics instructor will be angry

Outliers


10000
8000
6000
4000
2000
20
40
60
80


Out, liar!
Your theory is wrong!

Outliers

An outlier is a data point that is outside the range of other values in the dataset

    Can be subjective

May be due to

    biological variability

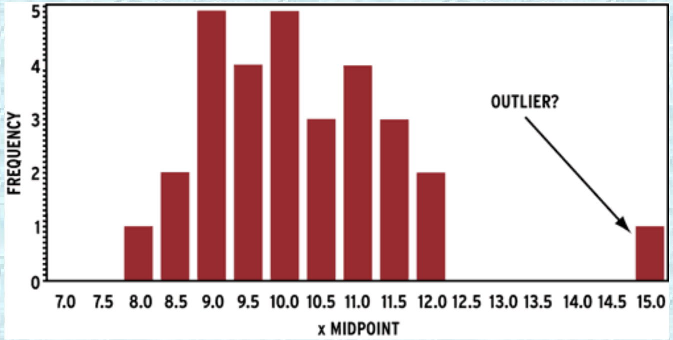
    experimental or measurement error

If they are real data, do not delete

Identifying Outliers

Graph the data and eyeball it

Frequency distribution, scatter, box plots



Two different methods for identifying outliers for a normally distributed dataset:

Interquartile range (IQR) method

Z-score method

Identifying Outliers:
Interquartile Range (IQR) Method
Remember the Tukey whiskers for boxplots

set up a "fence" utilizing Q1 and Q3. Any values that fall outside of this fence are considered outliers.

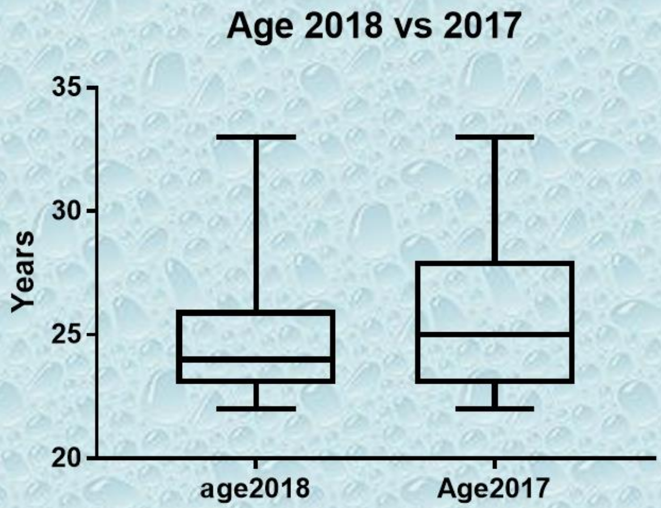
To build this fence take 1.5 times the IQR and then subtract this value from Q1 and add this value to Q3:

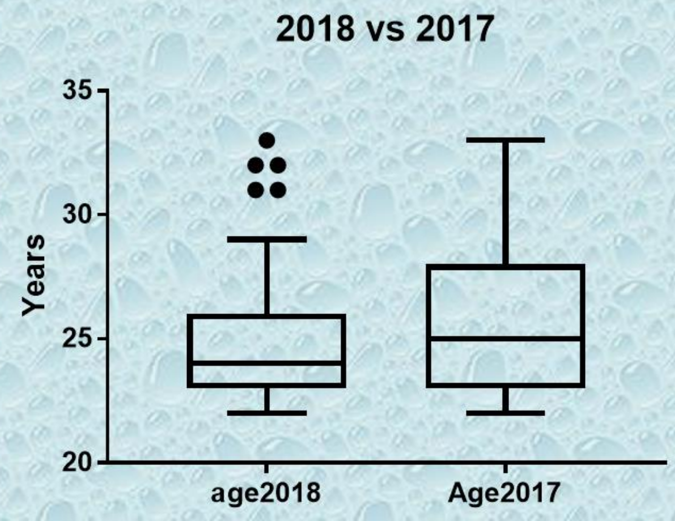
Lower Fence = Q1 – 1.5(IQR); Upper Fence = Q3 + 1.5(IQR)

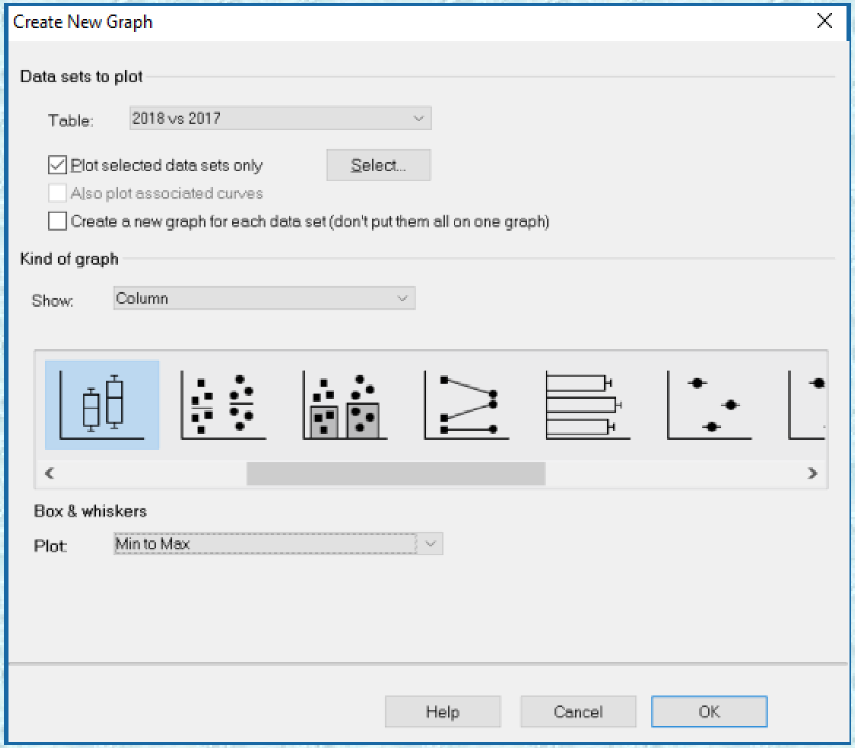
Observations more than 1.5 IQR below Q1 or more than 1.5 IQR above Q3 are considered outliers.

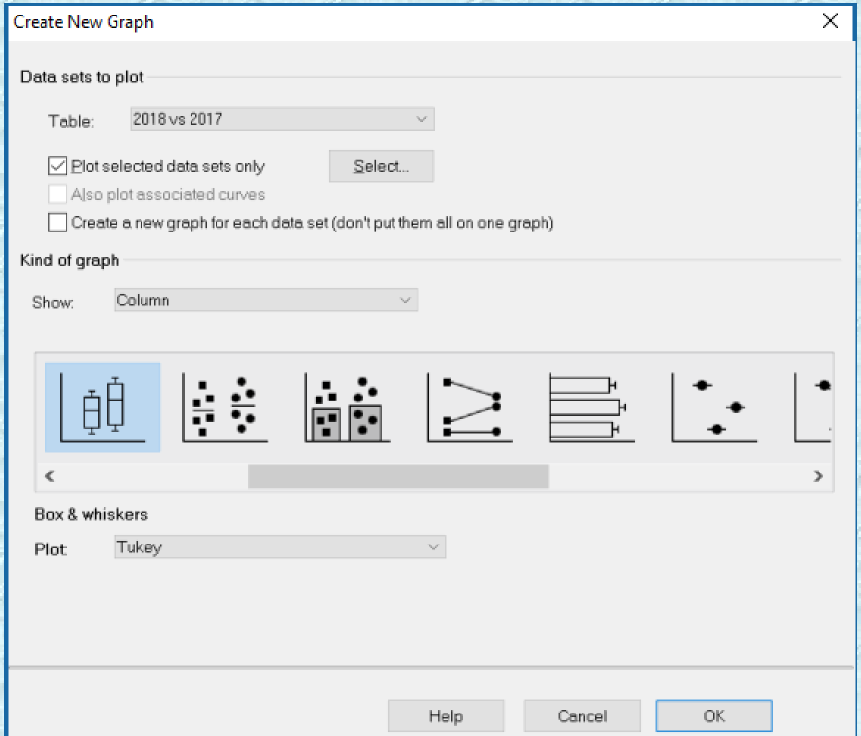
This is the method that Prism uses to identify outliers.

Why 1.5IQR? There is no statistical rationale; it is simply how Tukey decided to do it, and he invented the idea of box-and-whisker plots.


Age 2018 vs 2017
Years
35
30
25
20
age2018
Age2017


2018 vs 2017
Years
35
30
25
20
age2018
Age2017


Create New Graph
Data sets to plot
Table: 2018 vs 2017
Plot selected data sets only
Select...
Also plot associated curves
Create a new graph for each data set (don't put them all on one graph)
Kind of graph
Show: Column
Box & whiskers
Plot: Min to Max
Help
Cancel
OK


Create New Graph
Data sets to plot
Table: 2018 vs 2017
Plot selected data sets only
Select...
Also plot associated curves
Create a new graph for each data set (don't put them all on one graph)
Kind of graph
Show: Column
Box & whiskers
Plot: Tukey
Help
Cancel
OK

Example: Test Scores, Are there "outliers"? IQR Method by hand

A teacher wants to examine students' test scores. Their scores are:
74, 80, 80, 84, 88, 90, 90, 90, 90, 98

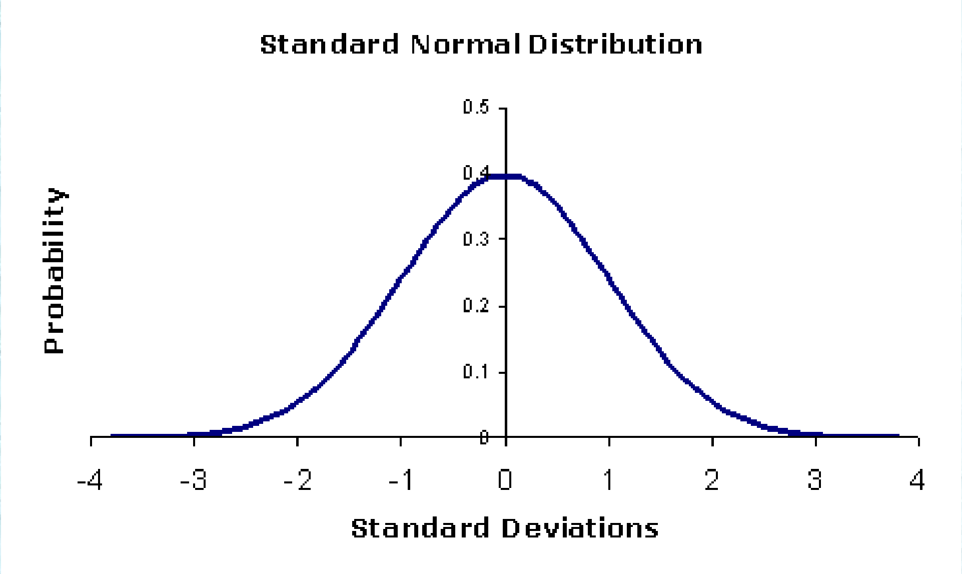
1	2	3	4	5	6	7	8	9	10	Rank

The median is 89 [(88 + 90)/2]
Q1=80
Q2=90
IQR = 90 – 80 = 10
1.5 (IQR) = 1.5 (10) = 15
Our "fences" will be 15 points below Q1 and 15 points above Q3.
Lower fence = 80 – 15 = 65
Upper fence = 90 + 15 = 105
Any scores that are less than 65 or greater than 105 are outliers. In this case, there are no outliers.

But caution, because they statistics say they are outliers, it doesn't tell us the data may be real and evidence of important biological variation

z Score method
Example: Test Scores, Are there "outliers"?

It is unusual for an observation to fall more than 3 standard deviations from the mean. Thus, any observation with a z score less than -3 or greater than +3 could be considered a potential outlier.



But remember, any statistical "outlier" may be real data

Example: Test Scores, Are there "outliers"? "z" Score method

Example: Test Scores
A teacher wants to examine students' test scores. Their scores are: 74, 88, 78, 90, 94, 90, 84, 90, 98, and 80
First, must compute mean and SD. Prism gives us mean (xbar)= 86.6, SD = 7.5

Now we can compute the z score for each student score using the formula "z" = (x-xbar)/s

Student	X	xbar	x-xbar	SD	z score
1	74	86.6	-12.6	7.5	-1.7
2	88	86.6	1.4	7.5	0.2
3	78	86.6	-8.6	7.5	-1.1
4	90	86.6	3.4	7.5	0.5
5	94	86.6	7.4	7.5	1.0
6	90	86.6	3.4	7.5	0.5
7	84	86.6	-2.6	7.5	-0.3
8	90	86.6	3.4	7.5	0.5
9	98	86.6	11.4	7.5	1.5
10	80	86.6	-6.6	7.5	-0.9

Any z scores less than -3 or greater than +3 are considered "outliers". Again, there are no outliers in this distribution.

What to do about an outlier

First, make sure that the data point is not an error

- Measurement error
- Data entry error

If it is real

- Do not automatically delete it despite what formal tests say
  - Could be a rare but important variation in your data

- Data transformation may help
- Use a non-parametric test

Questions?

Assignment – will be given out next week to cover this week and next