

Meeting Assumptions: the Normal Distribution

Kathleen Torkko
September 16, 2019

Objectives

Introduce common test assumptions

Identify properties of normal distribution, a common test assumption

Learn about determining if data meet a normal distribution

Understand the difference between parametric and non-parametric tests

Learn what z-scores are and how to use them

The ARRIVE Guidelines Checklist

Animal Research: Reporting In Vivo Experiments

Carol Kilkenney¹, William J Browne², Innes C Cuthill³, Michael Emerson⁴ and Douglas G Altman⁵

¹The National Centre for the Replacement, Refinement and Reduction of Animals in Research, London, UK, ²School of Veterinary Science, University of Bristol, Bristol, UK, ³School of Biological Sciences, University of Bristol, Bristol, UK, ⁴National Heart and Lung Institute, Imperial College London, UK, ⁵Centre for Statistics in Medicine, University of Oxford, Oxford, UK.

Statistical methods	13	<p>a. Provide details of the statistical methods used for each analysis.</p> <p>b. Specify the unit of analysis for each dataset (e.g. single animal, group of animals, single neuron).</p> <p>c. Describe any methods used to assess whether the data met the assumptions of the statistical approach.</p>	<p>a/b. Analysis</p> <p>c. Analysis; also Results, paragraph 1.</p>
---------------------	----	---	---

5. For every figure, are statistical tests justified as appropriate?

Do the data meet the assumptions of the tests (e.g., normal distribution)?

Is there an estimate of variation within each group of data?

Is the variance similar between the groups that are being statistically compared? (Give section/paragraph or page #)

April 2015

Important for
parametric tests on
continuous data

Parametric vs. Non-Parametric Tests

Parametric statistical test -

makes assumptions about the parameters (*i.e.*, μ , σ) of the population distribution from which the data sample is drawn

makes assumptions about the distribution of the data, *i.e.*, the data meet a normal distribution

Parametric tests: t-test, ANOVA, Pearson correlation, etc.

Non-parametric test -

makes no assumptions about parameters or the data distribution

Non-parametric tests: Wilcoxon rank sum, Mann-Whitney U, Kruskal-Wallis

Assumptions in Statistics

Some common assumptions that must be met for some statistical tests
(i.e., parametric tests on continuous data)

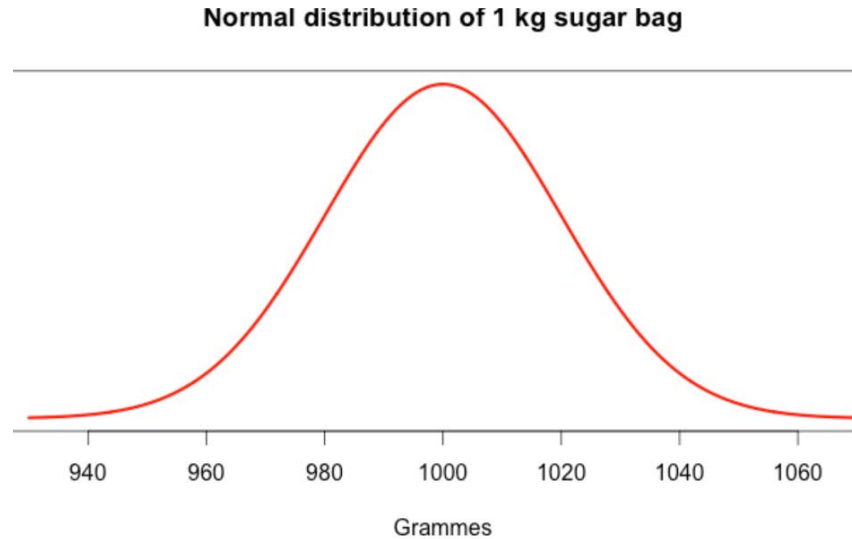
normality (normal distribution or at least symmetric data distribution)

homoscedasticity (data from multiple groups have the same variance)

independence (part of study design)

linearity

Assumptions in Statistics: Normality

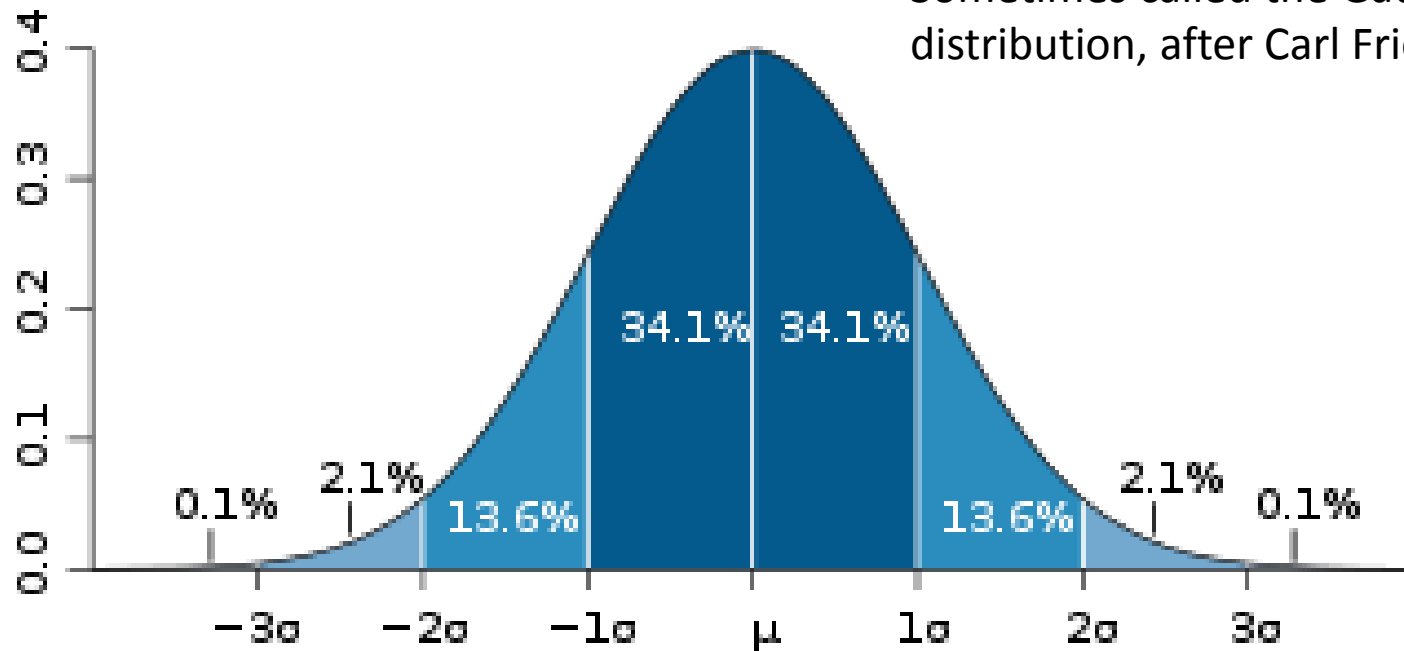


Deviations from normality potentially cause statistical tests to be inaccurate (biased)

The Normal Distribution

A normal distribution curve is a symmetrical, bell-shaped curve defined by the mean and standard deviation

Sometimes called the Gaussian distribution, after Carl Friedrich Gauss



Features of a Normal Distribution

Normal distributions are symmetric around their means

The mean, median, and mode are equal

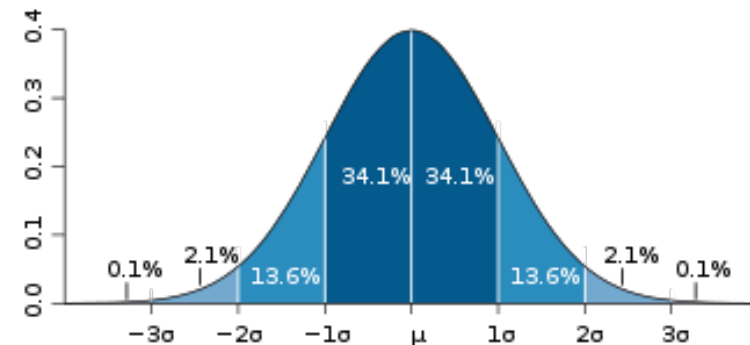
Normal distributions are denser in the center and less dense in the tails.

Normal distributions are defined by two parameters, the mean and the standard deviation.

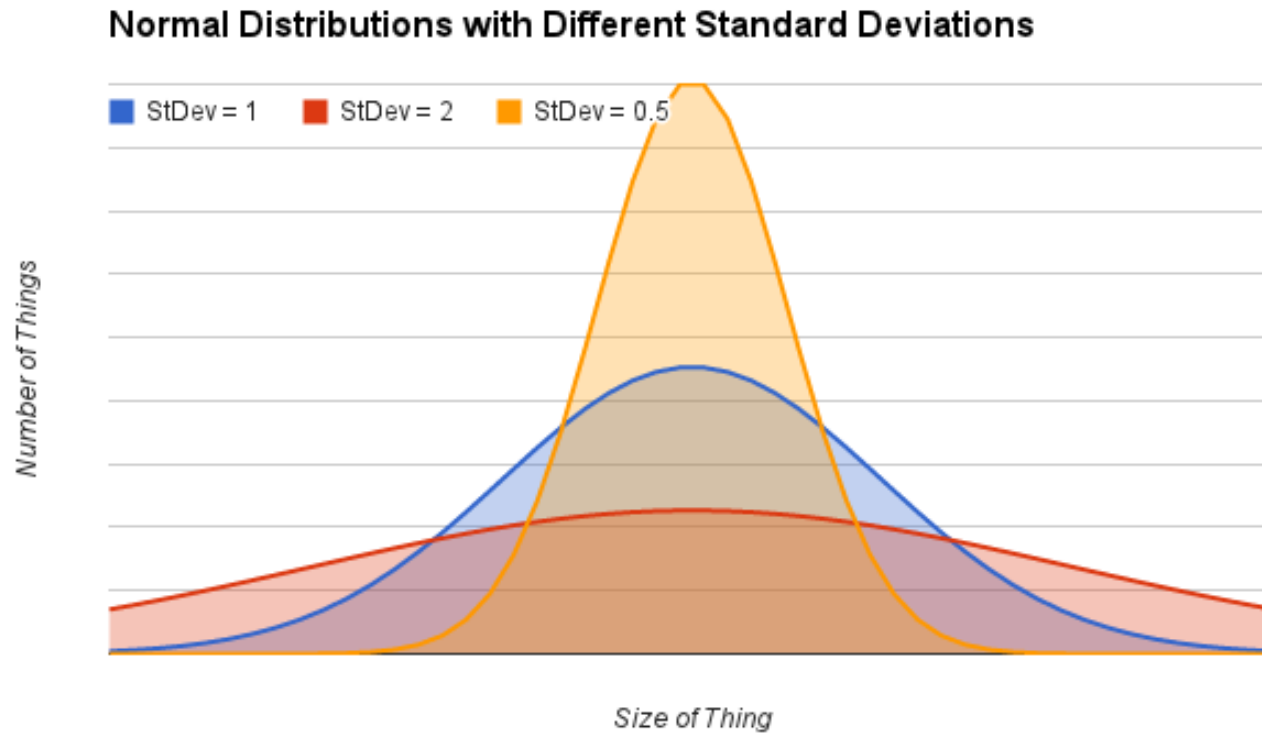
About 68% of the area of a normal distribution is within 1 SD of the mean.

About 95% is within 2 SD of the mean

The area under the normal curve is equal to 1.0

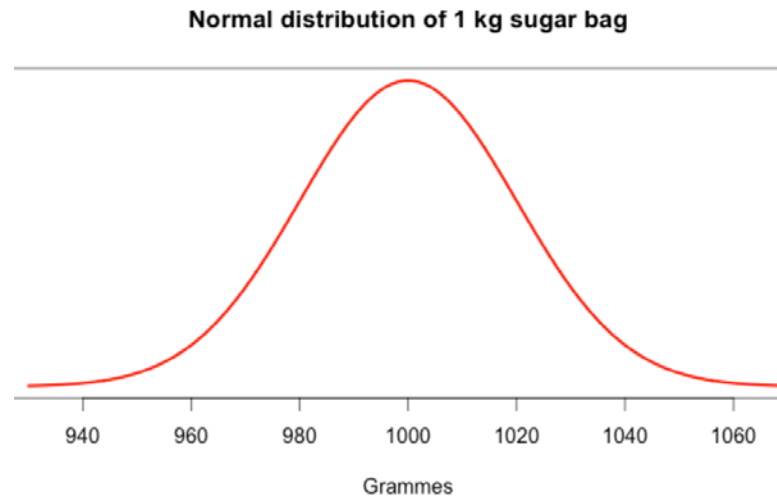


The Bell Curve Doesn't Always Have to be Bell Shaped

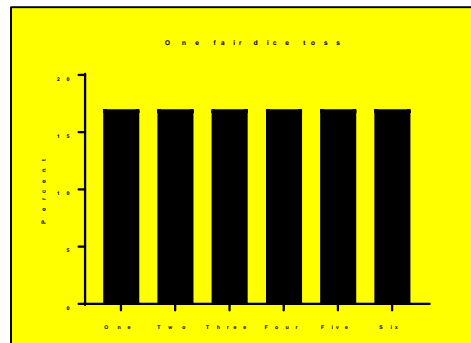


What is the normal distribution really?

It is a *model*, a way of describing the expected distribution of a continuous variable

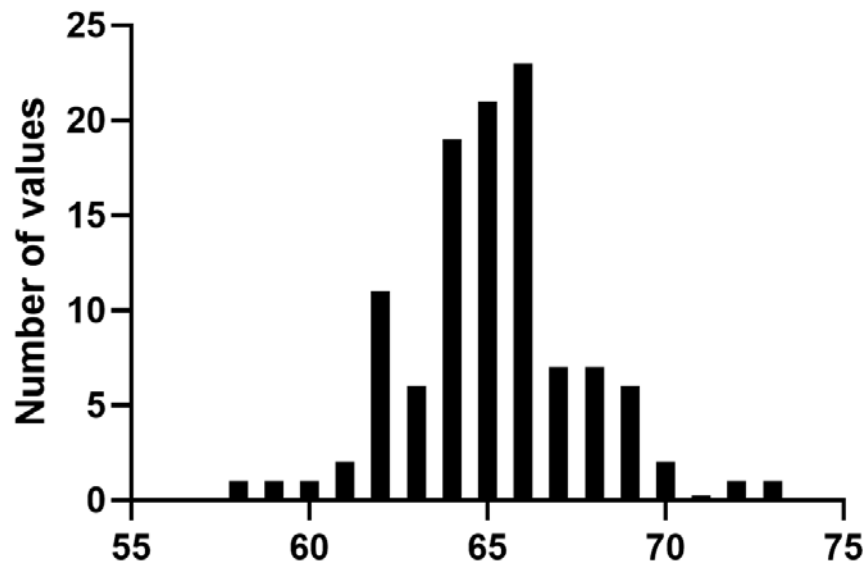


Remember the probability distributions we discussed earlier with discrete variables, our coin and dice toss experiments

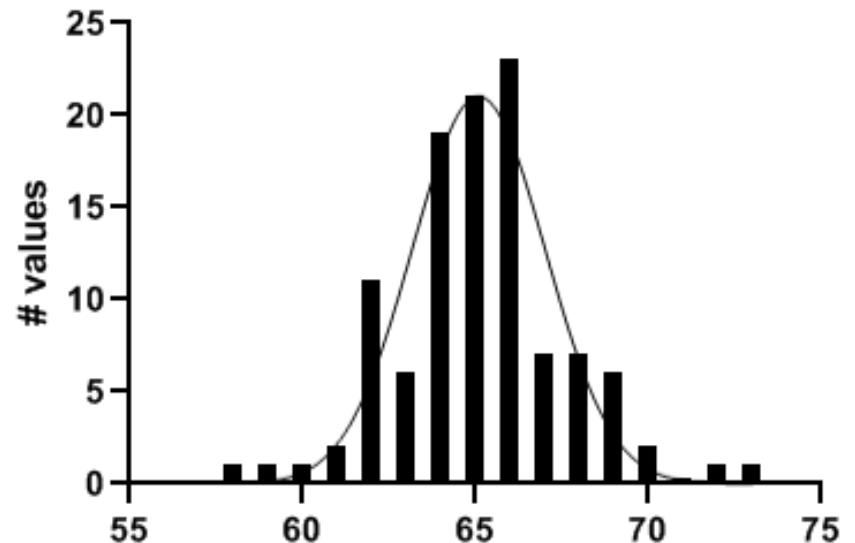


Example: The height of females in this class could be modeled by a normal distribution with a mean of 65.2 inches and a standard deviation of 2.5.

Histogram Spikes
Height Female Students 2017-19



Histogram
Height Female Students 2017-19
with Gaussian Distribution



To get the Gaussian distribution overlay

First create the frequency histogram using choices below

Parameters: Frequency Distribution

Create

☒ Frequency distribution
☐ Cumulative frequency distribution

Tabulate

☒ Number of values
☐ Relative frequency (fractions)
☐ Relative frequency (percentages)

Bin range

Center of first bin: ☒ Auto ☐ 21
Center of last bin: ☒ Auto ☐ 35

Bin width

☒ Choose automatically
☐ Bin width 1
☐ No bins. Tabulate exact cumulative frequency

Replicates

☐ Bin each replicate
☐ Bin only means

New graph

☒ Create a new graph of the results
Graph type: XY graph. Histogram spikes

Learn Cancel OK

Go to page with the histogram, click on “Analyze” and choose “Nonlinear Regression (curve fit)” and click OK

Analyze Data

Data to analyze
Table: Histogram of Height Female Students 2017-19:Frequency distribution

Type of analysis
Which analysis?

Transform, Normalize...
Transform
Transform concentrations (X)
Normalize
Prune rows
Remove baseline and column math
Transpose X and Y
Fraction of total

XY analyses
Nonlinear regression (curve fit)
Linear regression
Fit spline/LOWESS
Smooth, differentiate or integrate curve
Area under curve
Deming (Model II) linear regression
Row means with SD or SEM
Correlation
Interpolate a standard curve

Column analyses
Grouped analyses
Contingency table analyses
Survival analyses
Parts of whole analyses
Multiple variable analyses
Nested analyses

Analyze which data sets?
☒ A: # values

When you analyze tables or graphs with more than one data set, use this space to select which data set(s) to analyze.

Select All Deselect All

Help Cancel OK

Choose “Gaussian” (might be under “Standard Curves”) and hit OK

Parameters: Nonlinear Regression

Model Method Compare Constrain Initial values Range Output Confidence Diagnostics Flag

Choose an equation

- Standard curves to interpolate
- Dose-response - Stimulation
- Dose-response - Inhibition
- Dose-response - Special, X is concentration
- Dose-response - Special, X is log(concentration)
- Binding - Saturation
- Binding - Competitive
- Binding - Kinetics
- Enzyme kinetics - Inhibition
- Enzyme kinetics - Velocity as a function of substrate
- Exponential
- Lines
- Polynomial
- Gaussian
 - Gaussian
 - Sum of two Gaussians
 - Lognormal
 - Cumulative Gaussian -- Percents
 - Cumulative Gaussian -- Fraction
 - Cumulative Gaussian -- Counts
 - Lorentzian
 - Sum of two Lorentzian

Gaussian

Analytical derivatives

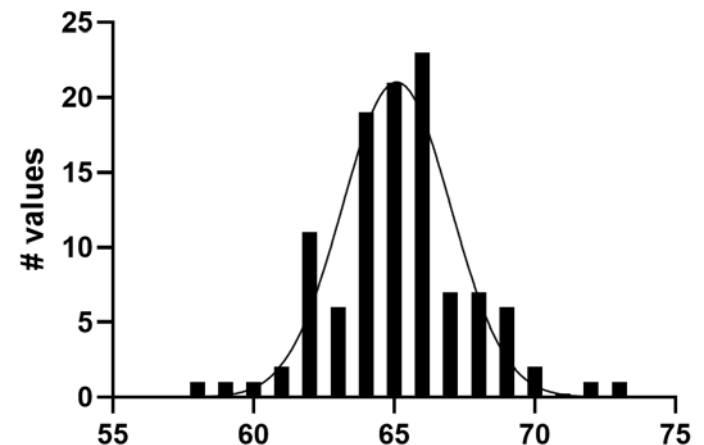
Interpolate

☐ Interpolate unknowns from standard curve. Confidence interval: None

Learn Cancel OK

The Gaussian curve using the mean and SD from your data will be placed on the histogram

Histogram
Height Female Students 2017-19
with Gaussian Distribution



The Normal Distribution: Why do we care?

In probability theory, the normal distribution is a very common and probably the most important *continuous data probability distribution*.

In statistics, normal distributions are often used to represent real variables from random samples whose population distributions are not known.

The normal distribution can be normalized and used to compare individual values to known population distributions (z-scores)

It is important for the central limit theorem

The math behind it is exciting and beautiful for the mathematical statisticians

Why do I care about normality?
What happens if I ignore non-normality?

Type-II errors (false negatives) can increase
Due largely to the fact that the SD is larger
Skewed distributions
Adv in Health Sci Educ (2011) 16:291–296

Your results will be biased
Due to the error of using the wrong test

Your statistics teacher will be upset

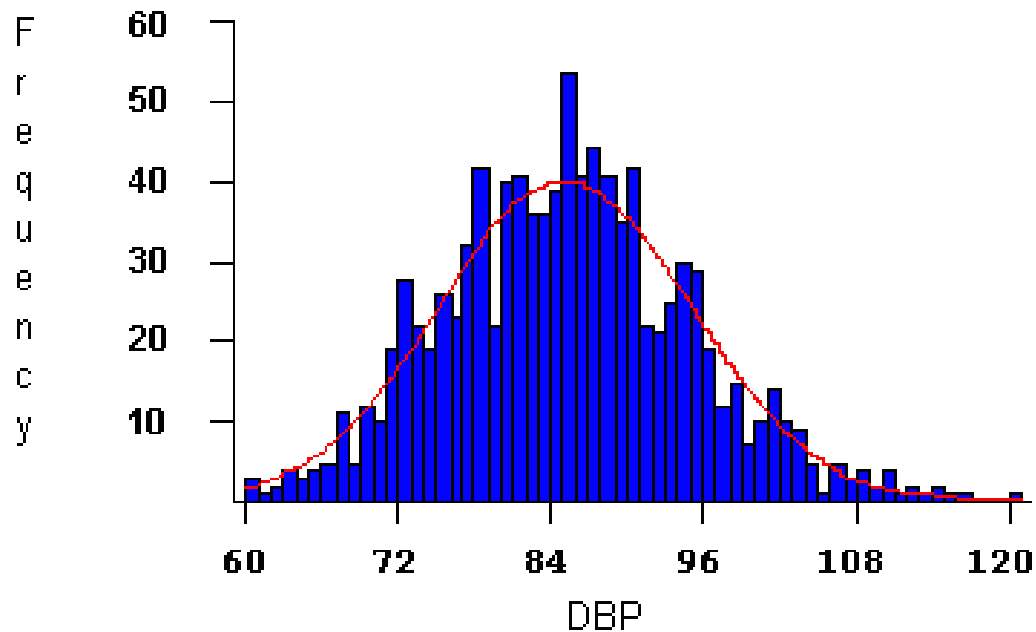
But true normality is elusive....



Example of a real “normal” distribution

Distribution of blood pressure can be approximated as a normal distribution with a mean 85 mm and standard deviation 20 mm.

Below is the histogram of 1,000 observations with a Gaussian curve



How do we determine “normality”?

Numerical methods (mean and median, skewness indices)

For mean and median, if they are similar, it is evidence of a potential normal (symmetrical) distribution

Formal normality tests in Prism

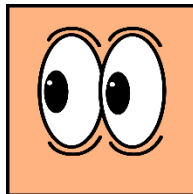
Anderson-Darling

D’Agostino-Pearson

Shapiro-Wilk

Kolmogorov-Smirnov (not recommended)

Graphical methods (frequency histograms, boxplots, Q-Q plots)



Let's practice using data from the class survey: Age and height

StudentHeightFemalesAgesAll2017to19.xlsx

Data structure in a Prism "Column" table

	Group A	Group B	Gr
	Height Females	AgeAll	
1	58	24	
2	59	25	
3	60	24	
4	61	27	
5	61	28	
6	62	23	
7	62	25	
8	62	23	
9	62	25	
10	62	24	
11	62	25	
12	62	29	
13	62	30	
14	62	24	
15	62	24	
16	62	23	
17	63	26	
18	63	23	
19	63	28	
20	63	23	

Choose “Analyze” from the Data Table, then select “Column Statistics” in the pop-up menu

The image shows two overlapping dialog boxes from a software application. The background dialog is titled "Analyze Data" and has a dropdown menu set to "Built-in analysis". It is divided into two main sections: "Which analysis?" and "Analyze which data sets?".

Analyze Data - Which analysis?

- Transform, Normalize...**
 - Transform
 - Transform concentrations (X)
 - Normalize
 - Prune rows
 - Remove baseline and column math
 - Transpose X and Y
 - Fraction of total
- XY analyses**
- Column analyses**
 - t tests (and nonparametric tests)
 - One-way ANOVA (and nonparametric or mixed)
 - One sample t and Wilcoxon test
 - Descriptive statistics** (highlighted)
 - Normality and Lognormality Tests
 - Frequency distribution
 - ROC Curve
 - Bland-Altman method comparison
 - Identify outliers
 - Analyze a stack of P values
- Grouped analyses**
- Contingency table analyses**
- Survival analyses**
- Parts of whole analyses**

Analyze which data sets?

- ☒ A: Height Females
- ☒ B: AgeAll

Buttons at the bottom: Select All, Deselect All, Help, Cancel.

The foreground dialog is titled "Parameters: Descriptive statistics".

Basics

- ☒ Minimum and maximum, range
- ☒ Mean, SD, SEM
- ☒ Quartiles (Median, 25th and 75th percentile)
- ☐ Column sum

Advanced

- ☐ Coefficient of variation
- ☒ Skewness and kurtosis
- ☐ Percentile: 90
- ☐ Geometric mean
- ☐ Harmonic mean
- ☐ Quadratic mean

Confidence intervals

- ☒ CI of the mean
- ☐ CI of harmonic mean
- ☐ CI of geometric mean
- ☐ CI of quadratic mean
- ☐ CI of median

Confidence level: 95%

Subcolumns

- ☒ Average the replicates in each row, and then perform the calculation for each column
- ☐ Perform the calculation for each subcolumn separately
- ☐ Treat all the values in all the subcolumns as one set of data

Output

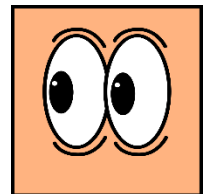
Show this many significant digits: 4

☐ Make these choices be the default for future analyses.

Buttons at the bottom: Learn, Cancel, OK.

Mean and median close or not? Skewed?

Descriptive statistics		A	B	
		Height Females	AgeAll	
1	Number of values	109	166	
2				
3	Minimum	58.00	21.00	
4	25% Percentile	64.00	23.00	
5	Median	65.00	24.00	
6	75% Percentile	66.00	27.00	
7	Maximum	73.00	35.00	
8	Range	15.00	14.00	
9				
10	Mean	65.16	25.31	
11	Std. Deviation	2.461	2.970	
12	Std. Error of Mean	0.2358	0.2305	
13				
14	Lower 95% CI of mean	64.69	24.86	
15	Upper 95% CI of mean	65.62	25.77	
16				
17	Skewness	0.1736	1.171	
18	Kurtosis	1.063	0.8232	
19				

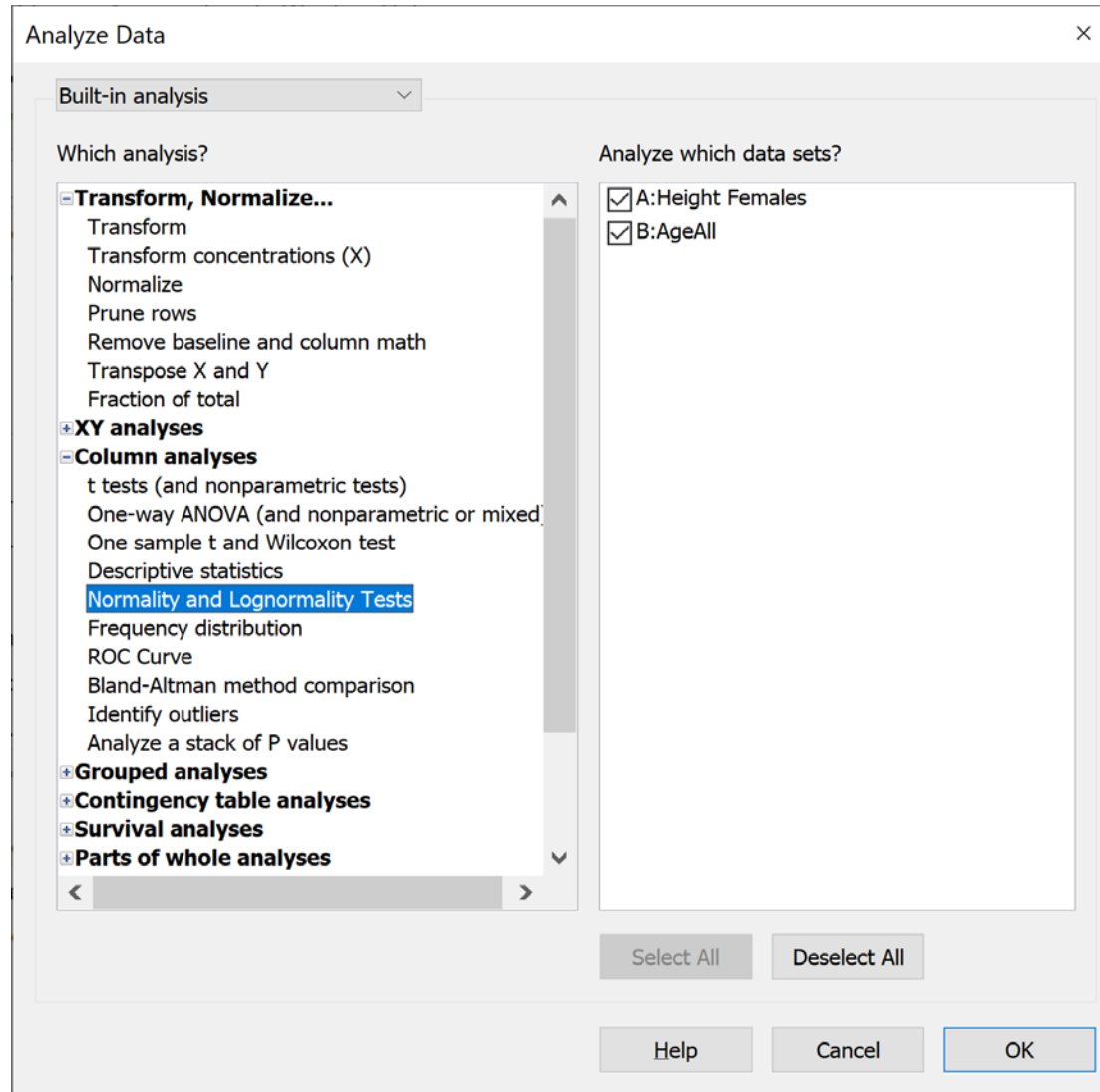


Skewness Index – a reminder

Numerical Example - Skewness

- If skewness = 0, the data are perfectly symmetrical.
- [Bulmer \(1979\)](#) suggests this rule of thumb:
 - If skewness is less than -1 or greater than $+1$, the distribution is **highly skewed**.
 - If skewness is between -1 and $-\frac{1}{2}$ or between $+\frac{1}{2}$ and $+1$, the distribution is **moderately skewed**.
 - If skewness is between $-\frac{1}{2}$ and $+\frac{1}{2}$, the distribution is **approximately symmetric**.

Normality tests in Prism



Which distribution(s) to test?

- ☒ Normal (Gaussian) distribution
- ☐ Lognormal distribution
- ☐ Compute the relative likelihood of sampling from a Gaussian (normal) vs. a lognormal distribution (assuming no other possibilities)

Methods to test distribution(s)

- ☒ Anderson-Darling test
- ☒ D'Agostino-Pearson omnibus normality test
- ☒ Shapiro-Wilk normality test
- ☒ Kolmogorov-Smirnov normality test with Dallal-Wilkinson-Lilliefors P value

Graphing options

- ☒ Create a QQ plot

Subcolumns

- ☒ Average the replicates in each row, and then perform the calculation for each column
- ☐ Perform calculations on each subcolumn separately
- ☐ Treat all the values in all subcolumns as single set of data

Calculations

Significance level (alpha)

Output

Show this many significant digits (for everything except P values):

P value style: GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**), 0.0001 (***) N =

- ☐ Make these choices the default for future analyses.

[Learn](#)

Cancel

OK

D'Agostino-Pearson normality test.

Prism recommends this test above the others

Test uses skewness and kurtosis to quantify how far the distribution is from Gaussian in terms of asymmetry and shape.

Calculates how far each values differs from the value expected with a Gaussian distribution, and computes a single P value from the sum of these discrepancies.

Prism recommended alternatives:

Anderson-Darling test.

Computes p-value by comparing the distribution of your data set against the ideal distribution of a Gaussian distribution.

Shapiro-Wilk normality test.

The Shapiro-Wilk test works very well if every value is unique, it does not work as well when several values are identical.

Then there is the...

Kolmogorov-Smirnov test.

Prism offers this test to be compatible with earlier versions of the program. They do not recommend it.

HOW USEFUL ARE NORMALITY TESTS?

- ☐ ***Not very useful in most situations.***


With small samples (<20), normality tests don't have much power to detect non-Gaussian distributions.

With larger samples (>50), the tests have too much power to detect small deviations from the Gaussian distribution.

But, it doesn't matter so much if data are non-Gaussian, since t-tests and ANOVA are fairly robust to normality violations in certain circumstances

What you would want is a test that tells you whether the deviations from the Gaussian ideal are severe enough to bias statistical methods that assume a Gaussian distribution.

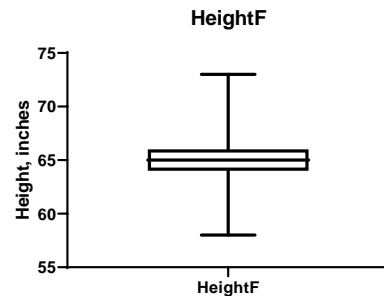
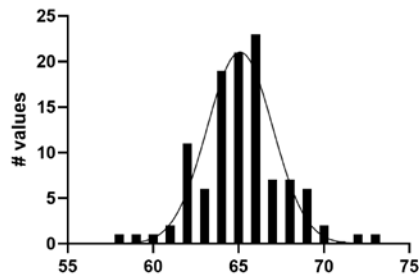
But normality tests don't do this

 Normality and Lognormality Tests Tabular results		A	B	
		Height Females	AgeAll	
1	Test for normal distribution	n=109	n=166	
2	Anderson-Darling test			
3	A2*	1.452	7.134	
4	P value	0.0009	<0.0001	
5	Passed normality test (alpha=0.05)?	No	No	
6	P value summary	***	****	
7				
8	D'Agostino & Pearson test			
9	K2	4.221	30.88	
10	P value	0.1212	<0.0001	
11	Passed normality test (alpha=0.05)?	Yes	No	
12	P value summary	ns	****	
13				
14	Shapiro-Wilk test			
15	W	0.9691	0.8754	
16	P value	0.0123	<0.0001	
17	Passed normality test (alpha=0.05)?	No	No	
18	P value summary	*	****	
19				
20	Kolmogorov-Smirnov test			
21	KS distance	0.1457	0.2130	
22	P value	<0.0001	<0.0001	
23	Passed normality test (alpha=0.05)?	No	No	

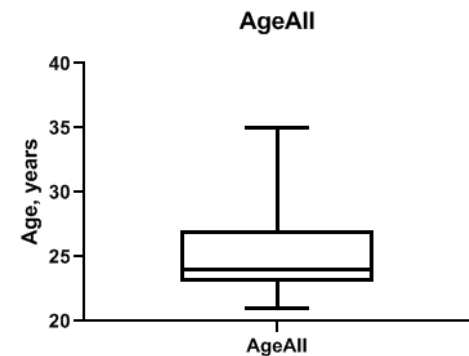
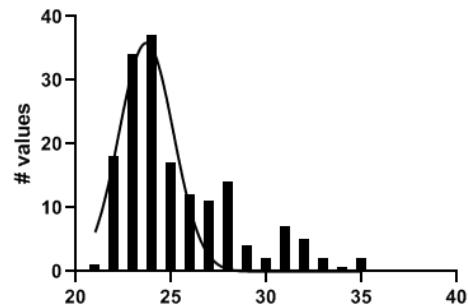
Normality and Lognormality Tests		A	B
Tabular results		Height Females	AgeAll
1	Test for normal distribution		
2	Anderson-Darling test		
3	A2*	1.452	7.134
4	P value	0.0009	<0.0001
5	Passed normality test (alpha=0.05)?	No	No
6	P value summary	***	****
7			
8	D'Agostino & Pearson test		
9	K2	4.221	30.88
10	P value	0.1212	<0.0001
11	Passed normality test (alpha=0.05)?	Yes	No
12	P value summary	ns	****
13			
14	Shapiro-Wilk test		
15	W	0.9691	0.8754
16	P value	0.0123	<0.0001
17	Passed normality test (alpha=0.05)?	No	No
18	P value summary	*	****
19			
20	Kolmogorov-Smirnov test		
21	KS distance	0.1457	0.2130
22	P value	<0.0001	<0.0001
23	Passed normality test (alpha=0.05)?	No	No

Descriptive statistics		A	B
		Height Females	AgeAll
1	Number of values	109	166
2			
3	Minimum	58.00	21.00
4	25% Percentile	64.00	23.00
5	Median	65.00	24.00
6	75% Percentile	66.00	27.00
7	Maximum	73.00	35.00
8	Range	15.00	14.00
9			
10	Mean	65.16	25.31
11	Std. Deviation	2.461	2.970
12	Std. Error of Mean	0.2358	0.2305
13			
14	Lower 95% CI of mean	64.69	24.86
15	Upper 95% CI of mean	65.62	25.77
16			
17	Skewness	0.1736	1.171
18	Kurtosis	1.063	0.8232
19			

**Histogram
Height Female Students 2017-19
with Gaussian Distribution**



**Histogram of Age All
with Gaussian Distribution**



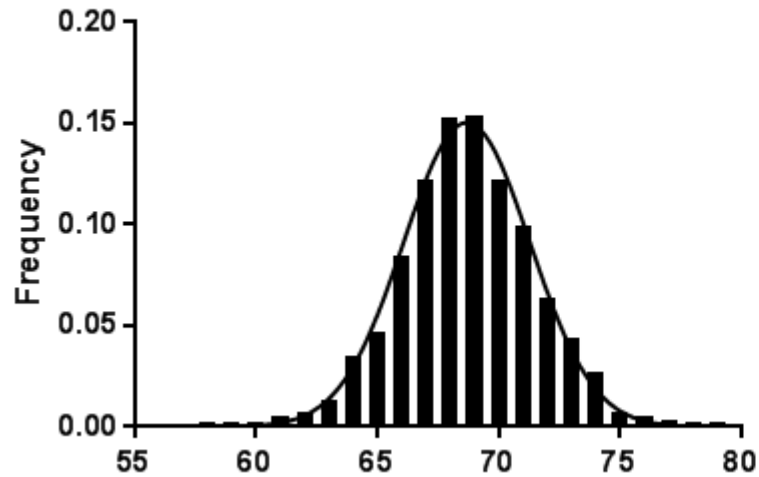
Let me show you how I got the graphs for the age data....

StudentHeightFemalesAgesAll2017to19.xlsx

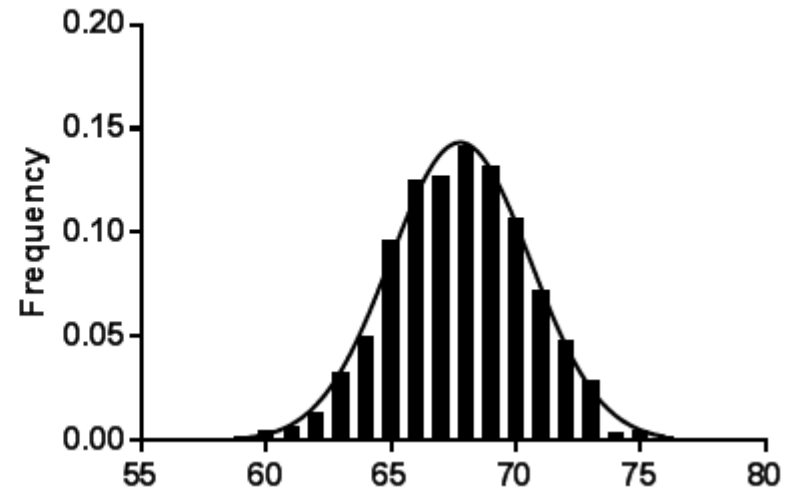
Let's practice eyeballing
 FatherSonsHeights2019.xlsx”
 Height measurements in 1,078 father son pairs

Descriptive statistics		A	B
		Father	Son
1	Number of values	1078	1078
2			
3	Minimum	59.00	58.50
4	25% Percentile	65.80	66.90
5	Median	67.80	68.60
6	75% Percentile	69.60	70.50
7	Maximum	75.40	78.40
8	Range	16.40	19.90
9			
10	Mean	67.69	68.68
11	Std. Deviation	2.746	2.816
12	Std. Error of Mean	0.08363	0.08577
13			
14	Lower 95% CI of mean	67.52	68.52
15	Upper 95% CI of mean	67.85	68.85
16			
17	Skewness	-0.08836	-0.03646
18	Kurtosis	-0.1545	0.5343
19			

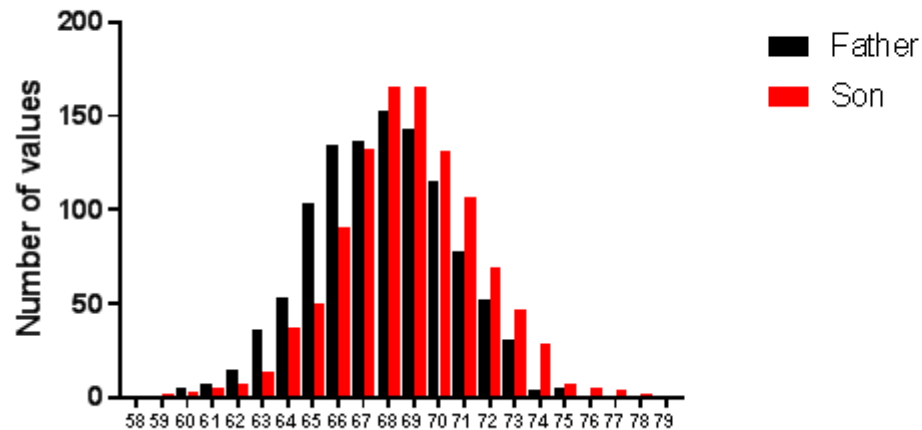
Histogram of Son Height (inches)



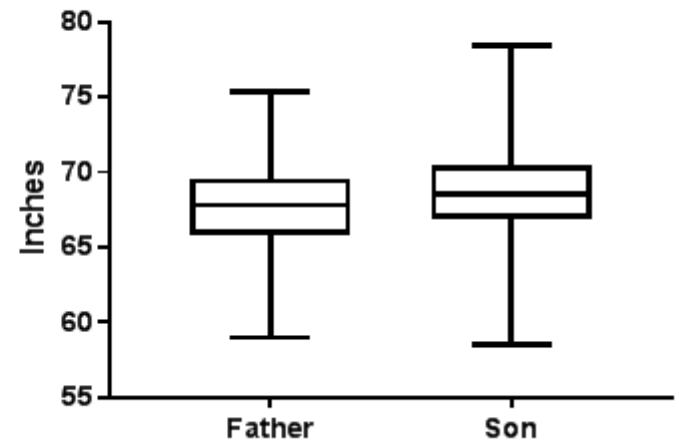
Histogram of Father Height (inches)



Histogram of Heights



Heights Father Son Pairs



What do the normality tests conclude?



Parameters: Normality and Lognormality Tests

Which distribution(s) to test?

☒ Normal (Gaussian) distribution

☐ Lognormal distribution

☐ Compute the relative likelihood of sampling from a Gaussian (normal) vs. a lognormal distribution (assuming no other possibilities)

Methods to test distribution(s)

☒ Anderson-Darling test

☒ D'Agostino-Pearson omnibus normality test

☒ Shapiro-Wilk normality test

☐ Kolmogorov-Smirnov normality test with Dallal-Wilkinson-Lilliefors P value

Graphing options

☒ Create a QQ plot

Subcolumns

☒ Average the replicates in each row, and then perform the calculation for each column

☐ Perform calculations on each subcolumn separately

☐ Treat all the values in all subcolumns as single set of data

Calculations

Significance level (alpha)

Output

Show this many significant digits (for everything except P values):

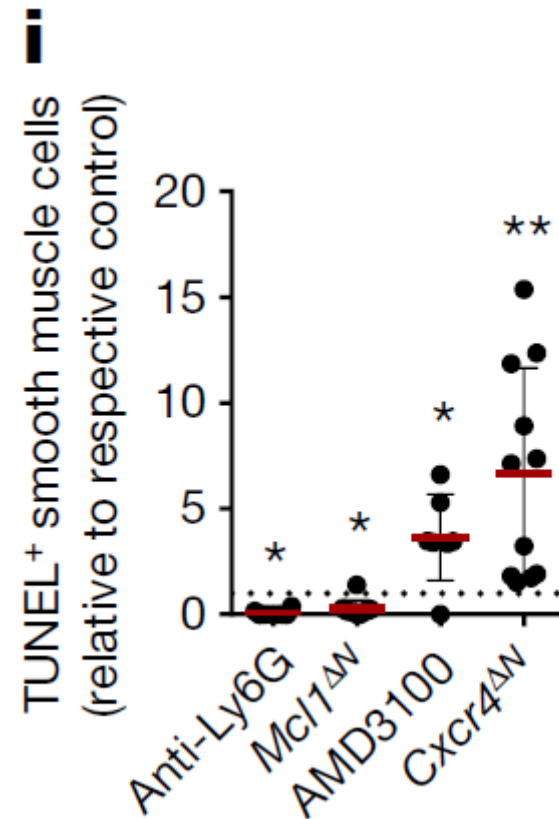
P value style: N =

☐ Make these choices the default for future analyses.

[Learn](#) [Cancel](#) [OK](#)

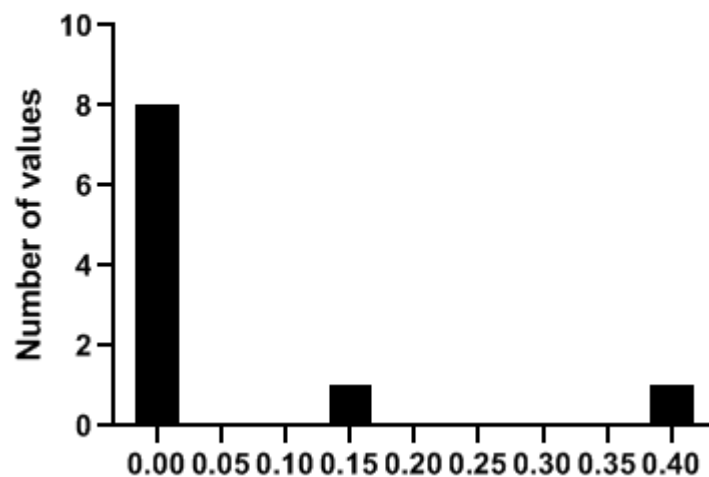
Normality and Lognormality Tests		A	B
Tabular results		Father	Son
1	Test for normal distribution		
2	Anderson-Darling test		
3	A2*	0.4509	0.9033
4	P value	0.2741	0.0212
5	Passed normality test (alpha=0.05)?	Yes	No
6	P value summary	ns	*
7			
8	D'Agostino & Pearson test		
9	K2	2.537	9.021
10	P value	0.2813	0.0110
11	Passed normality test (alpha=0.05)?	Yes	No
12	P value summary	ns	*
13			
14	Shapiro-Wilk test		
15	W	0.9978	0.9964
16	P value	0.1594	0.0130
17	Passed normality test (alpha=0.05)?	Yes	No
18	P value summary	ns	*
19			
20	Number of values	1078	1078

Now, let's look at "SilvestreRoigFig1j.xlsx"
n=7 to 11 for different groups

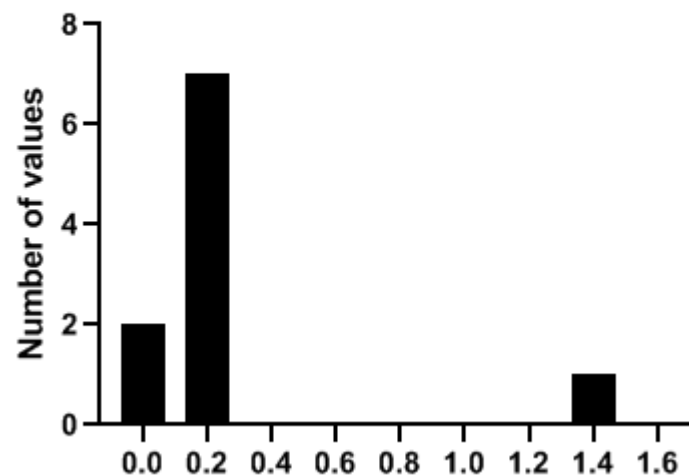


What do the normality tests tell you?

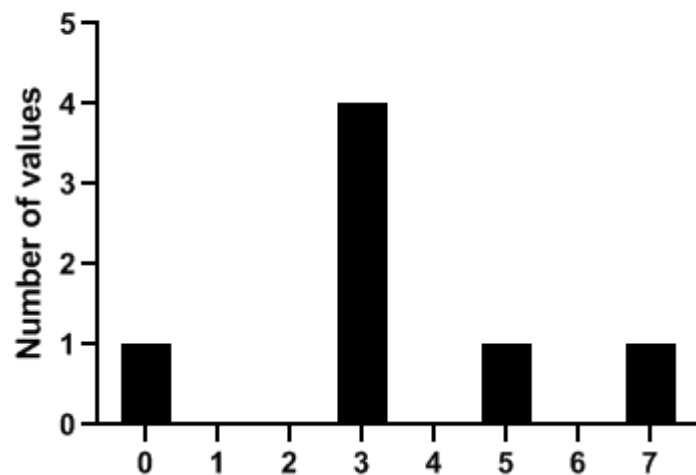
Histogram of Fig1j Anti-Ly6G
n=10



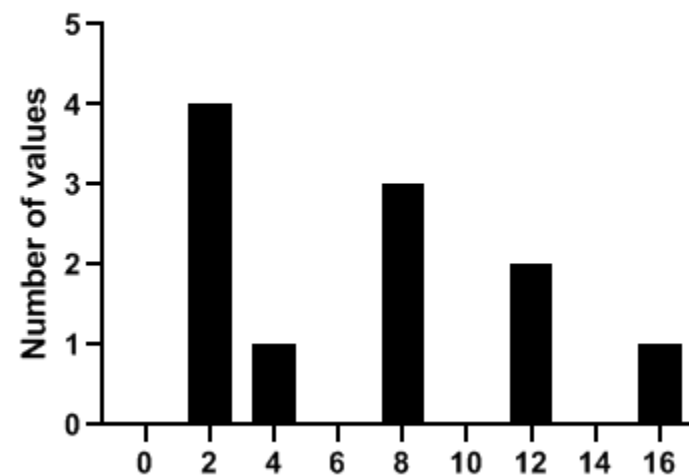
Histogram of Fig1j Mcl1
n=10

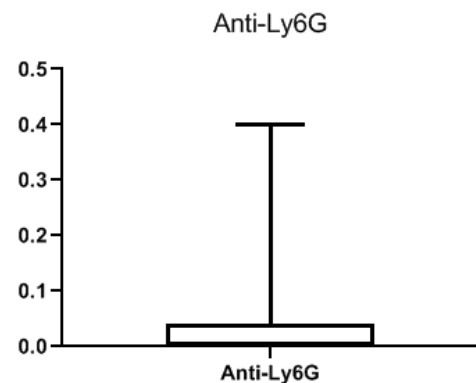


Histogram of Fig1j AMD3100
n=7

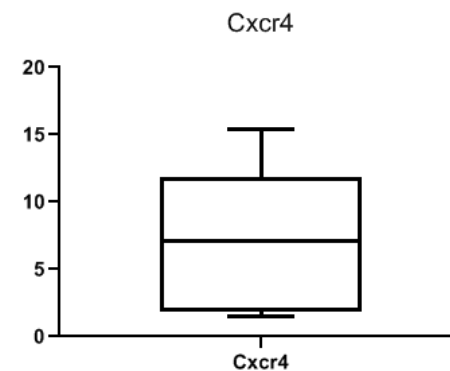
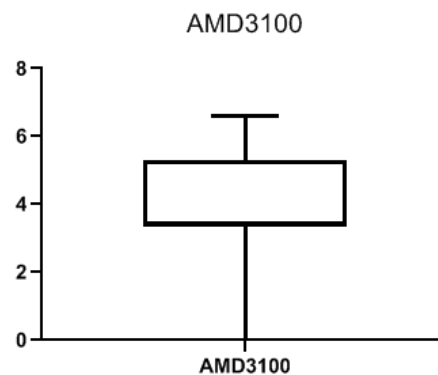
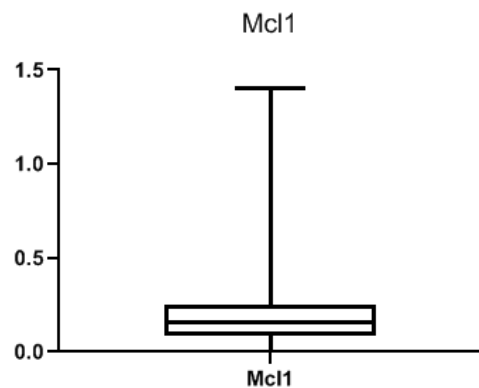


Histogram of Fig1j Cxcr4
n=11





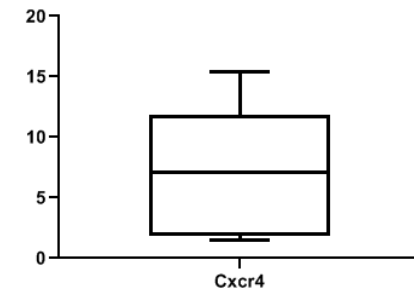
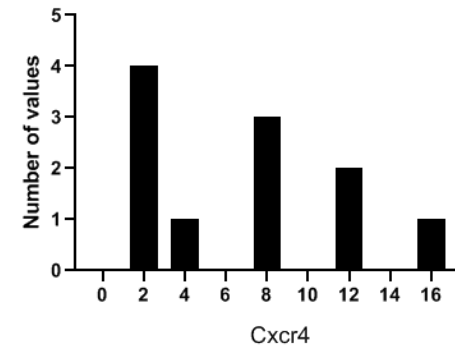
Descriptive statistics		A	B	C	D
		Anti-Ly6G	Mcl1	AMD3100	Cxcr4
1	Number of values	10	10	7	11
2					
3	Minimum	0.000	0.000	0.000	1.514
4	25% Percentile	0.000	0.08850	3.358	1.810
5	Median	0.000	0.1590	3.458	7.118
6	75% Percentile	0.04025	0.2530	5.293	11.85
7	Maximum	0.3990	1.403	6.609	15.37
8	Range	0.3990	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	Lower 95% CI of mean	-0.03750	-0.02411	1.766	3.296
15	Upper 95% CI of mean	0.1495	0.5609	5.542	10.01
16					
17	Skewness	2.494	2.870	-0.5111	0.4753
18	Kurtosis	6.160	8.688	1.636	-1.207



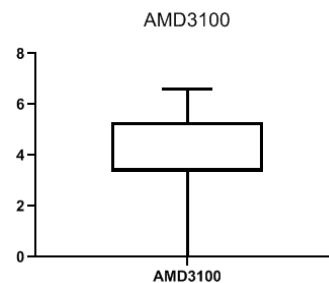
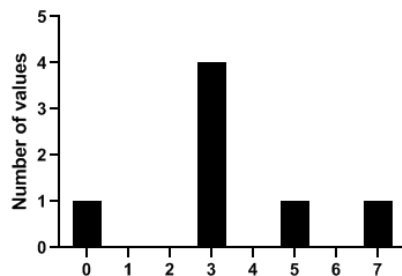


Normality and Lognormality Tests Tabular results		A	B	C	D
		Anti-Ly6G	Mcl1	AMD3100	Cxcr4
1	Test for normal distribution				
2	Anderson-Darling test				
3	A2*	2.341	1.846	N too small	0.5268
4	P value	<0.0001	<0.0001		0.1375
5	Passed normality test (alpha=0.05)?	No	No		Yes
6	P value summary	****	****		ns
7					
8	D'Agostino & Pearson test				
9	K2	18.93	24.88	N too small	1.716
10	P value	<0.0001	<0.0001		0.4239
11	Passed normality test (alpha=0.05)?	No	No		Yes
12	P value summary	****	****		ns
13					
14	Shapiro-Wilk test				
15	W	0.5167	0.5759	0.8893	0.8824
16	P value	<0.0001	<0.0001	0.2709	0.1117
17	Passed normality test (alpha=0.05)?	No	No	Yes	Yes
18	P value summary	****	****	ns	ns
19					
20	Number of values	10	10	7	11

Histogram of Fig1j Cxcr4
n=11

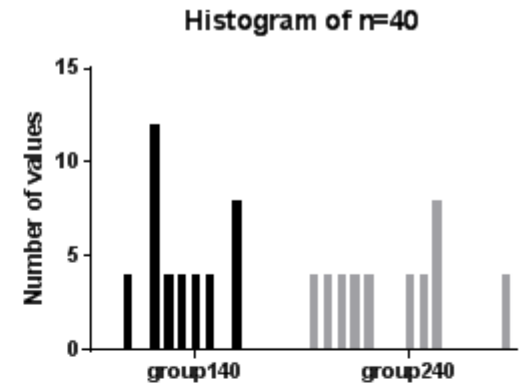
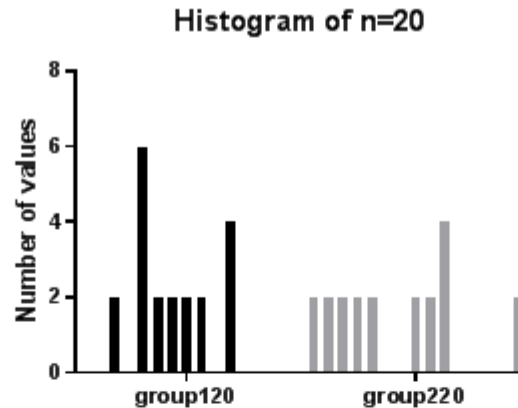


Histogram of Fig1j AMD3100
n=7



Effect of sample size

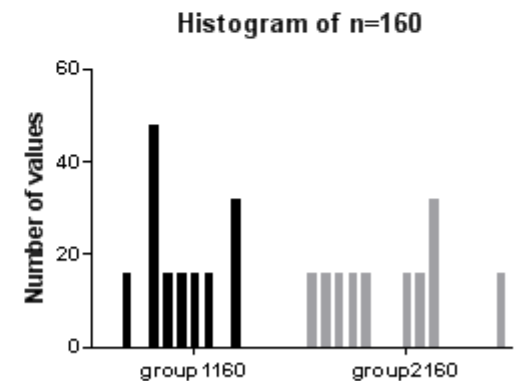
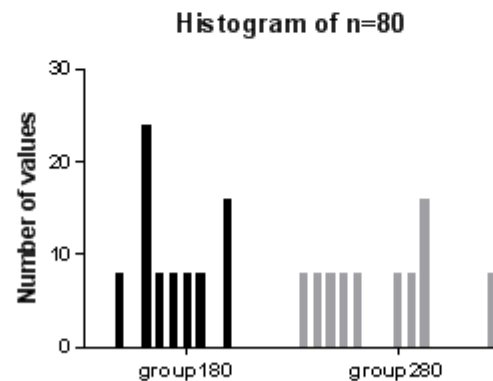
sample size increase the power of the test



D'Agostino & Pearson
Shapiro-Wilk

Normal Normal
Normal Normal

Normal Normal
NO NO



D'Agostino & Pearson
Shapiro-Wilk

Normal NO
NO NO

NO NO
NO NO

What to do?

For sample sizes $n=20+$

If mean and median are similar and the graph looks symmetrical with no outliers, assume you met the assumption of normality,

Particularly if $n=30+$

Can use test for normality if sample size 20-50 for assurance

If mean and median are dissimilar and the data are skewed, try transforming the data

Check mean, median, graphs

If not normal with transformation, use non-parametric test

If there are outliers check to see if the data are in error

If real, try transforming data or use a non-parametric test

For small sample sizes (<20)

Basically the same rules as for larger sample sizes, except

Do not trust normality tests

Add extra scrutiny with Q-Q (next) and boxplots

The truth is many parametric tests are robust to violations of normality

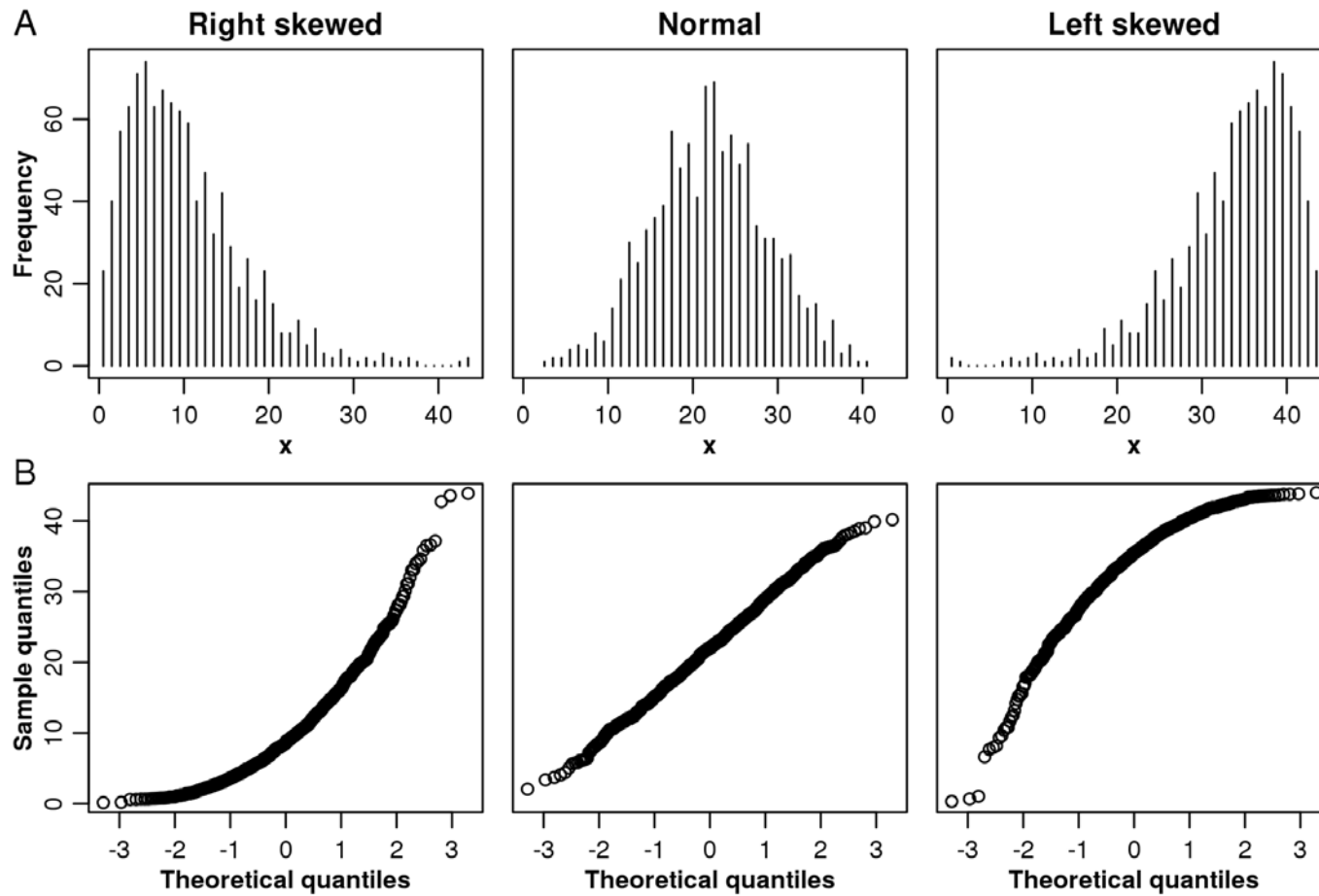
Many times experienced statisticians will just eye-ball the data
If it looks like a duck, it probably is a duck

The Q-Q Plot (Quantile-Quantile plot)

It is a graphical tool to help assess if a set of data could have come from a normal distribution

It is a scatterplot created by plotting two sets of quantiles (one observed from your sample and a theoretical sample)

If both sets of quantiles came from a normal distribution, the points should form a line that's roughly straight.



Parameters: Normality and Lognormality Tests



Which distribution(s) to test?

- ☒ Normal (Gaussian) distribution
- ☐ Lognormal distribution
- ☐ Compute the relative likelihood of sampling from a Gaussian (normal) vs. a lognormal distribution (assuming no other possibilities)

Methods to test distribution(s)

- ☒ Anderson-Darling test
- ☒ D'Agostino-Pearson omnibus normality test
- ☒ Shapiro-Wilk normality test
- ☒ Kolmogorov-Smirnov normality test with Dallal-Wilkinson-Lilliefors P value

Graphing options

- ☒ Create a QQ plot

Subcolumns

- ☒ Average the replicates in each row, and then perform the calculation for each column
- ☐ Perform calculations on each subcolumn separately
- ☐ Treat all the values in all subcolumns as single set of data

Calculations

Significance level (alpha)

Output

Show this many significant digits (for everything except P values):

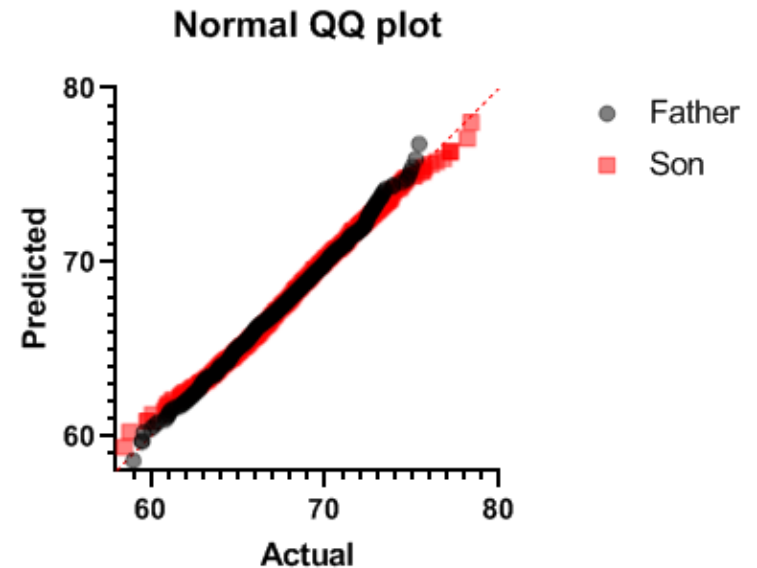
P value style: N =

☐ Make these choices the default for future analyses.

Learn

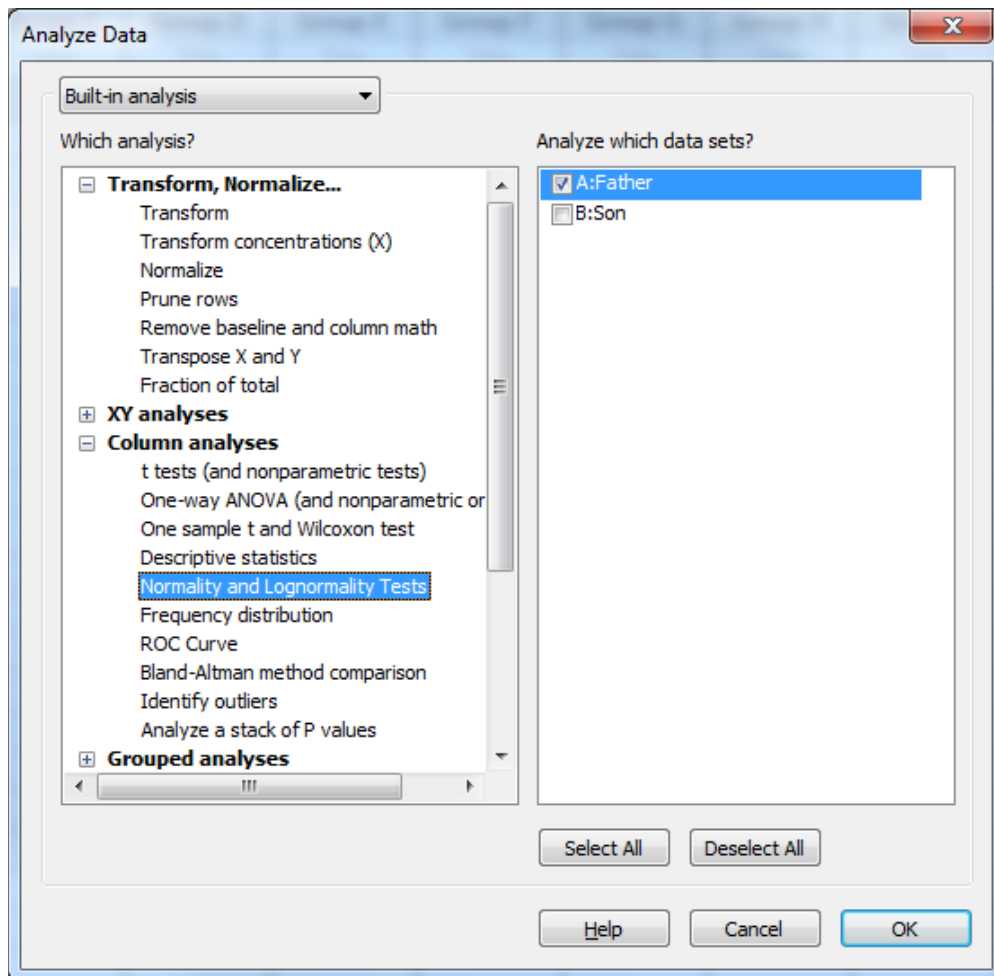
Cancel

OK

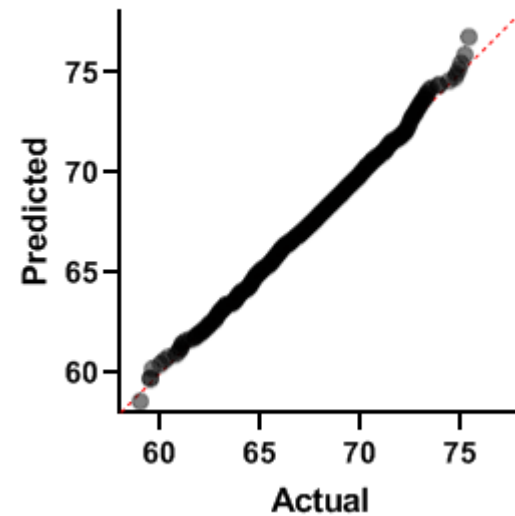


If you want separate plots,
do each group separately

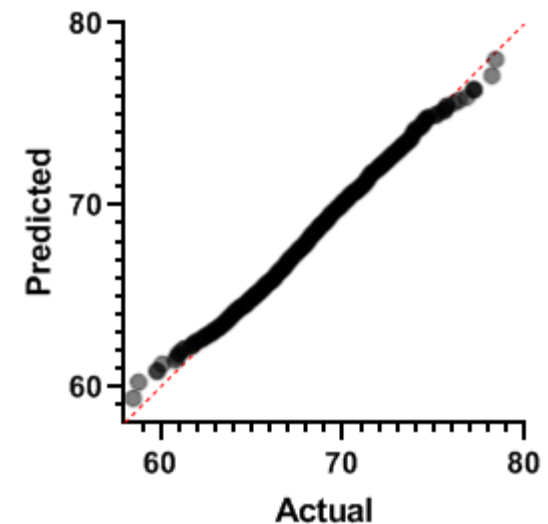
FatherSonHeights2019.xlsx



Normal QQ plot Father



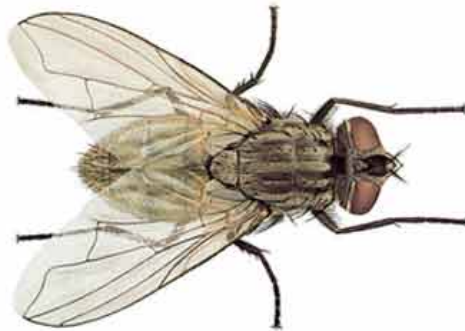
Normal QQ plot Son



Let's play with some data

HouseFly Wing Length 2019.xlsx

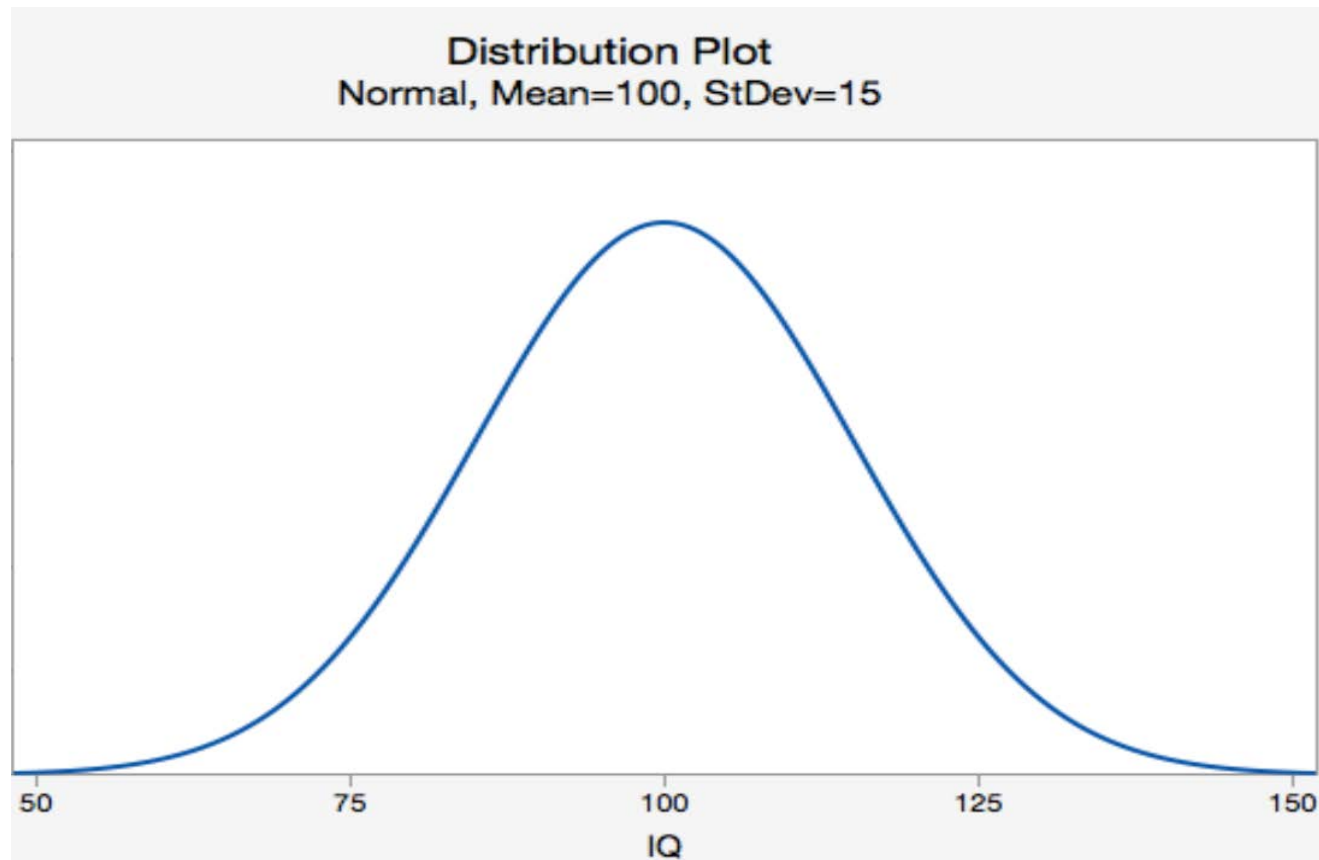
From Q-Q plots which of the two variables might be normal



The standard normal distribution and the z-score

IQ scores

We know the population distribution of IQ scores



The Standard Normal Distribution and z-scores

When a set of data values are normally distributed, the data can be standardized by converting it into a z-score.

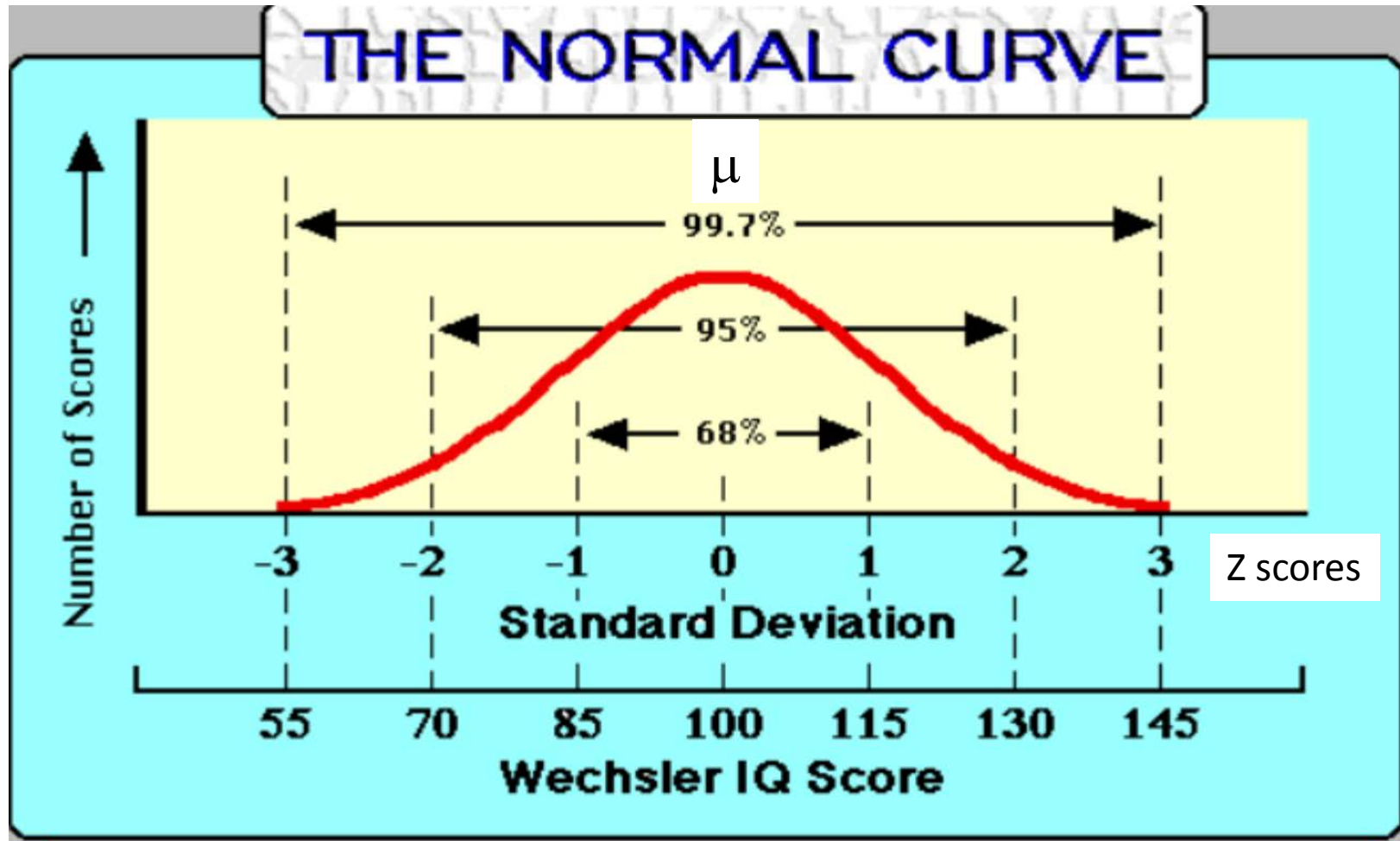
z-scores make it easier to

- compare individual data to standardized data

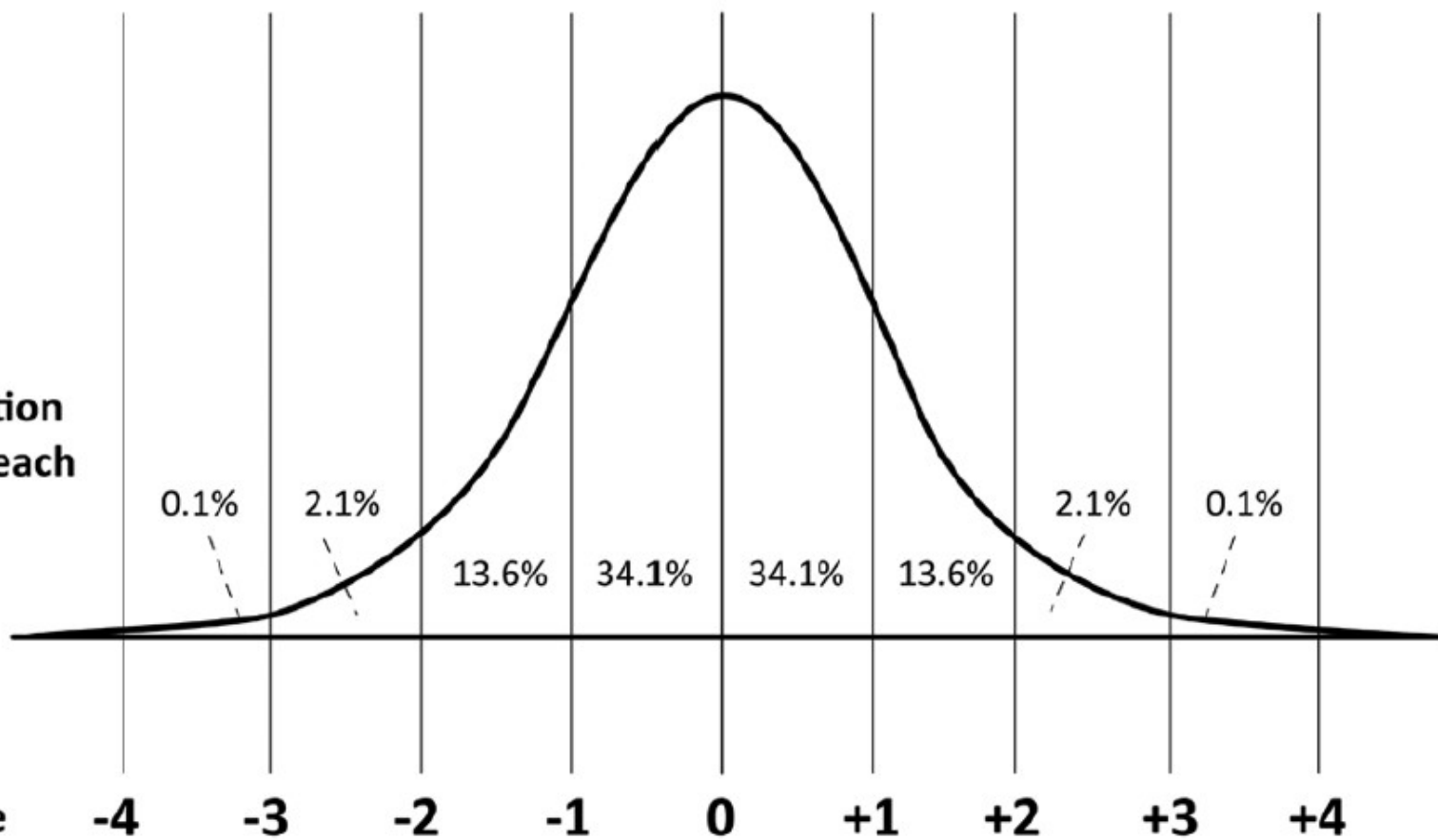
- can also compare data values measured on different scales

The standard normal distribution

The mean of the data in a standard normal distribution is 0 and the standard deviation is 1.



**Proportion
within each
sector**



Z-Score	Centile
-4	0.003%
-3	0.13%
-2	2.3%
-1	15.9%
0	50%
+1	84.1%
+2	97.7%
+3	99.87%
+4	99.997%

The Standard Normal Distribution and z-scores

A standard normal distribution is the set of all z-scores

A z-score reflects how many standard deviations above or below the mean an individual data point is

The z-score is positive if the data value lies above the mean and negative if the data value lies below the mean.

z-score (aka, a standard score)

Indicates how many standard deviations a measurement for an individual is from the population mean.

The **z-score** formula:

$$z = (X - \mu) / \sigma$$

where **z** is the **z-score**,

X is the measured data value for an individual

μ is the *population* mean

σ is the *population* standard deviation.

How do we use z-scores?

The commonest use of z-scores is in the analysis of human nutritional data, especially for children.

- Weight for age
- Height for age
- Weight for height
- Bone mineral density

Z-scores are computed using international reference data intended to reflect human growth patterns under optimal conditions.

Cut-off scores of -2 and -3 are used to identify children suffering from malnutrition.

Mean z-scores are used to evaluate the nutritional state of populations relative to the reference population.

National Center for Health Statistics

Growth Charts

CDC Growth Charts -

Background

Frequently Asked Questions

Clinical Growth Charts +

Individual Growth Charts

Data Tables -

Selected percentiles and LMS Parameters

Selected Z-score values

1977 NCHS Growth Chart Equations

Educational Materials +

Computer Programs

Reports

WHO Growth Charts +

Related Sites

National Health and Nutrition Examination Survey

[CDC](#) > [National Center for Health Statistics](#) > [Growth Charts](#) > [CDC Growth Charts](#) > [Data Tables](#)

Z-score Data Files



The values corresponding to specific z-scores (-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2) are contained in 8 Excel data files representing the 8 different growth curves for infants (weight-for-age; length-for-age, weight-for-recumbent length; head circumference-for-age) and older children (weight-for-stature; weight-for-age; stature-for-age; and BMI-for-age). The file and corresponding chart names are below:

1. [ZWTAGEINF](#) [XLS - 43 KB] [ZWTAGEINF](#) [CSV - 17 KB]
Weight-for-age charts, birth to 36 months, selected weight z-scores in kilograms, by sex and age
2. [ZLENAGEINF](#) [XLS - 42 KB] [ZLENAGEINF](#) [CSV - 17 KB]
Length-for-age charts, birth to 36 months, selected recumbent length z-scores in centimeters, by sex and age
3. [ZWTLENINF](#) [XLS - 60 KB] [ZWTLENINF](#) [CSV - 27 KB]
Weight-for-recumbent length charts, birth to 36 months, selected weight z-scores in kilograms, by sex and recumbent length (in centimeters)
4. [ZHCAGEINF](#) [XLS - 42 KB] [ZHCAGEINF](#) [CSV - 17 KB]
Head circumference-for-age charts, birth to 36 months, selected head circumference z-scores in centimeters, by sex and age
5. [ZWTSTAT](#) [XLS - 48 KB] [ZWTSTAT](#) [CSV - 20 KB]
Weight-for-stature charts, selected weight z-scores in kilograms, by sex and stature (in centimeters)
6. [ZWTAGE](#) [XLS - 188 KB] [ZWTAGE](#) [CSV - 98 KB]
Weight-for-age charts, 2 to 20 years, selected weight z-scores in kilograms, by sex and age
7. [ZSTATAGE](#) [XLS - 188 KB] [ZSTATAGE](#) [CSV - 98 KB]
Stature-for-age charts, 2 to 20 years, selected stature z-scores in centimeters, by sex and age
8. [ZBMIAGE](#) [XLS - 188 KB] [ZBMIAGE](#) [CSV - 98 KB]
BMI-for-age charts, 2 to 20 years, selected BMI (kilograms/meters squared) z-scores, by sex and age

These files contain the z-scores values for the z-scores of -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, and 2 by sex (1=male; 2=female) and half month of age. For example, 1.5 months represents 1.25-1.75 months. The only exception is birth, which represents the point at birth.

An example

Suppose SAT scores among the population of college students are normally distributed with a mean of 500 and a standard deviation of 100. If a student scores a 700, what would her **z**-score be?

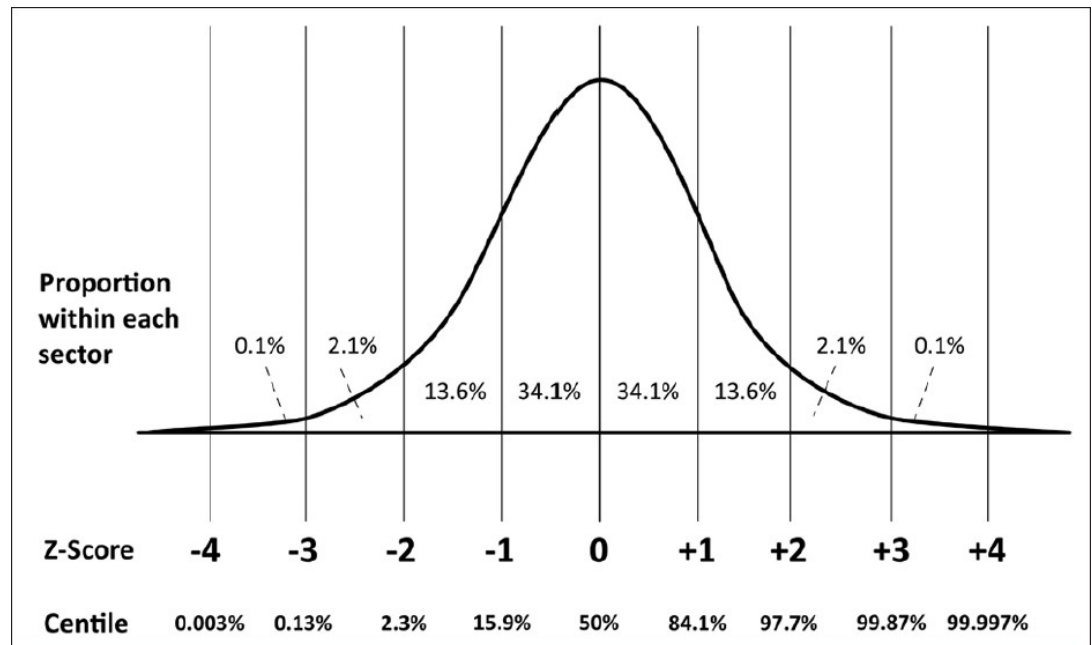
An example

Suppose SAT scores among college students are normally distributed with a mean of 500 and a standard deviation of 100. If a student scores a 700, what would her z -score be?

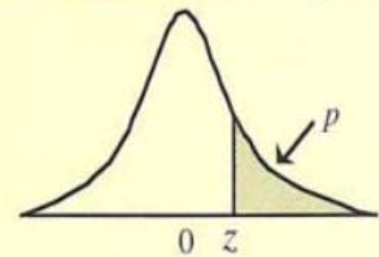
$$Z = (700 - 500) / 100 = 200 / 100 = 2$$

With a z -score of 2, she is 2 SD above the mean

She did better than 97.7% of the population of college students



Finding a p-value
from a z-score,
negative scores



z	Second decimal place of z									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2297	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026

Early Growth and Neurologic Outcomes of Infants with Probable Congenital Zika Virus Syndrome

Antonio Augusto Moura da Silva, et al

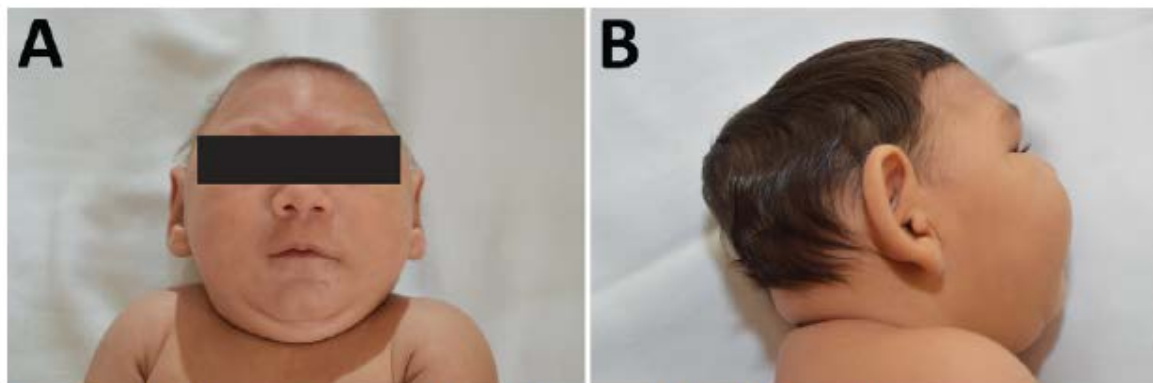


Figure 1. Characteristic phenotype of fetal brain disruption sequence in infants with probable congenital Zika virus syndrome, Sao Luís, Brazil, 2015–2016. A) Craniofacial disproportion and biparietal depression. B) Prominent occiput.

Table. Clinical characteristics of probable congenital Zika virus syndrome in infants from birth to 1–8 months of age, Sao Luis, Brazil, 2015–2016

Characteristic	No. (%)
Rash in mother during pregnancy, n = 46	
First trimester	24 (52.2)
First month	1 (2.2)
Second month	12 (26.1)
Third month	11 (23.9)
Second trimester	10 (21.7)
Fourth month	9 (19.6)
Sixth month	1 (2.2)
No rash	12 (26.1)
Sex, n = 48	
M	25 (52.1)
F	23 (47.9)
Gestational age at birth, n = 47	
Preterm	4 (8.5)
Term	41 (87.2)
Postterm	2 (4.3)
Head circumference z-score at birth,* n = 45	
≥ -2	6 (13.3)
Microcephaly, < -2	10 (22.2)
Severe microcephaly, < -3	29 (64.5)
Birth length z-score,* n = 3	
≥ -2	21 (56.8)
< -2	11 (29.7)
< -3	5 (13.5)
Birthweight z-score,* n = 46	
≥ -2	37 (80.4)
< -2	8 (17.4)
< -3	1 (2.2)

Calculation of z-scores

You need to have a table of reference values showing the mean (average) and standard deviation (SD) for the age, gender, race, skeletal site, and densitometer measurement units. I call this the "expected BMD". The following table gives values from NHANES dataset. Then you use the formula:

$$\text{Z-score} = (\text{Patient's BMD} - \text{expected BMD}) / \text{SD}$$

To calculate BMD if you know the Z-score, use the same equation rearranged:

$$\text{BMD} = \text{expected BMD} + (\text{Z-score} \times \text{SD})$$

Total hip, standardized units, mg/cm ²								
	White Women		White men		Black Women		Black Men	
Age	BMD	SD	BMD	SD	BMD	SD	BMD	SD
25	955	123	1055	146	1040	135	1189	171
35	945	130	1038	144	1017	142	1141	166
45	920	136	1002	140	1034	160	1094	162
55	876	139	990	143	973	175	1072	185
65	809	140	969	157	890	154	1027	168
75	740	129	928	151	838	154	984	173
85	679	135	859	161	723	146	933	194

For example, a white woman aged 55 with BMD of 850 has a Z-score of $(850-876)/139 = -0.18$

A black man aged 55 with BMD of 850 has a Z-score of $(850-1072)/185 = -1.20$

A 65 year old white woman with a Z-score of -2 has a BMD of $809 + (-2 \times 140) = 529$

A 25 year old white woman with a Z-score of -2 has a BMD of $955 + (-2 \times 123) = 832$

An example to discuss in class

At Hogwarts School of Witchcraft and Wizardry, Professor Snape was concerned about grade inflations, and suggested that the school should issue standardized grades (or-z-score), in addition to the regular grades. How might this be used to determine how well Harry Potter did in compared to the rest of the class. Harry was in four classes, each with 20 students. Computed his standardized grade in each class. If he was judged by standardized grades, where did he do best? Where did he do worse?

	<u>Harry's Score</u>	<u>Mean</u>	<u>Std Dev</u>
Care of Magical Creatures	3.80	3.75	.15
Defense Against the Dark Arts	3.60	3.25	.60
Transfiguration	3.10	3.20	.38
Potions	2.50	2.90	.75

Using the data for the transfiguration class, what score would a student need to get to be 2 SD above the mean?