Dealing with lack of Normality - Theory & Practice

http://www.pelagicos.net/classes_biometry_fa16.htm

Assessing Normality: 4 Methods

First: plot your data using a histogram

Note whether bin size affects number of modes

Second: Use P-P (and Q-Q) plot to compare the observed and the expected normal distribution

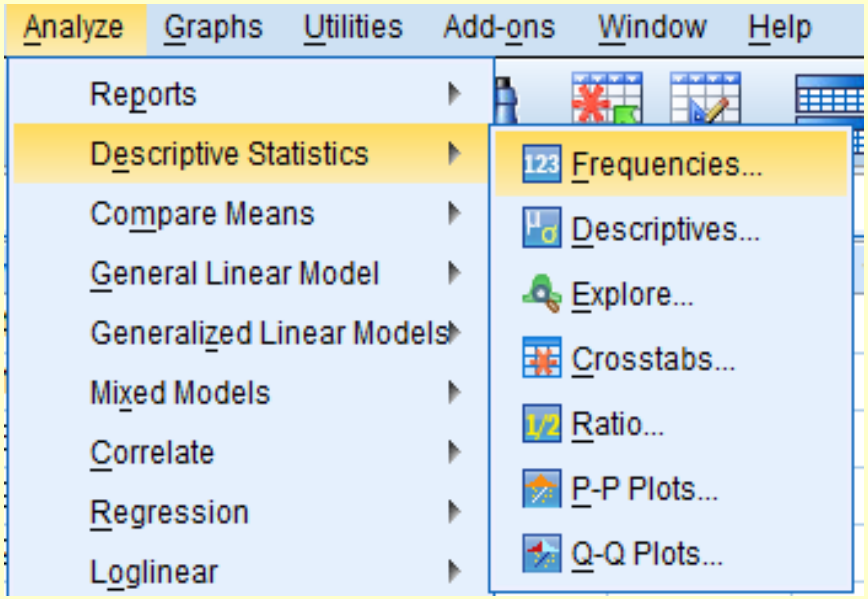
Third: Quantify skewness / kurtosis (point estimate and SE) using descriptive stats in SPSS

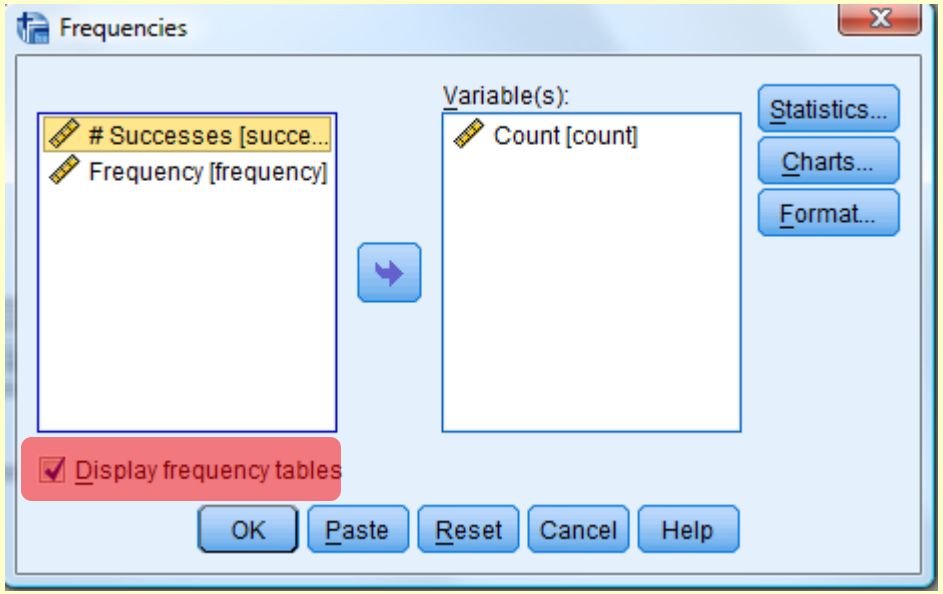
Compare observed parameter estimates (95% C.I.) to expected parameter (normal distribution) 0

Fourth: Perform S-W and K-S statistical tests

PSA-1: Calculate Frequencies

Converting count data into frequencies





NOTE:

Make sure you select "display frequency tables"

Table of Frequencies

Count

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	7	.7	.7	.7
	1.00	40	4.0	4.0	4.7
	2.00	121	12.1	12.1	16.8
	3.00	215	21.5	21.5	38.3
	4.00	251	25.1	25.1	63.4
	5.00	201	20.1	20.1	83.5
	6.00	111	11.1	11.1	94.6
	7.00	42	4.2	4.2	98.8
	8.00	10	1.0	1.0	99.8
	9.00	2	.2	.2	100.0
	Total	1000	100.0	100.0	

What is the mode?

4

What is the median?

4

PSA-2: SPSS Mode Calculation

n25

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-2.42	1	.1	4.0	4.0
	-1.79	1	.1	4.0	8.0
	-1.38	1	.1	4.0	12.0
	-1.24	1	.1	4.0	16.0
	-1.18	1	.1	4.0	20.0
	-1.06	1	.1	4.0	24.0
	-1.02	1	.1	4.0	28.0
	-.95	1	.1	4.0	32.0
	-.89	1	.1	4.0	36.0
	-.73	1	.1	4.0	40.0
	-.63	1	.1	4.0	44.0
	-.38	1	.1	4.0	48.0
	-.17	1	.1	4.0	52.0
	.41	1	.1	4.0	56.0
	.45	1	.1	4.0	60.0
	.57	1	.1	4.0	64.0
	.69	1	.1	4.0	68.0
	.70	1	.1	4.0	72.0
	1.41	1	.1	4.0	76.0
	1.42	1	.1	4.0	80.0
	1.52	1	.1	4.0	84.0
	1.54	1	.1	4.0	88.0
	1.55	1	.1	4.0	92.0
	1.59	1	.1	4.0	96.0
	1.79	1	.1	4.0	100.0
	Total	25	2.5	100.0	
Missing	System	975	97.5		
Total		1000	100.0		

Data from random normal variable ~ N ($\mu = 0$, $\sigma = 1$).

Sample size (n) = 25

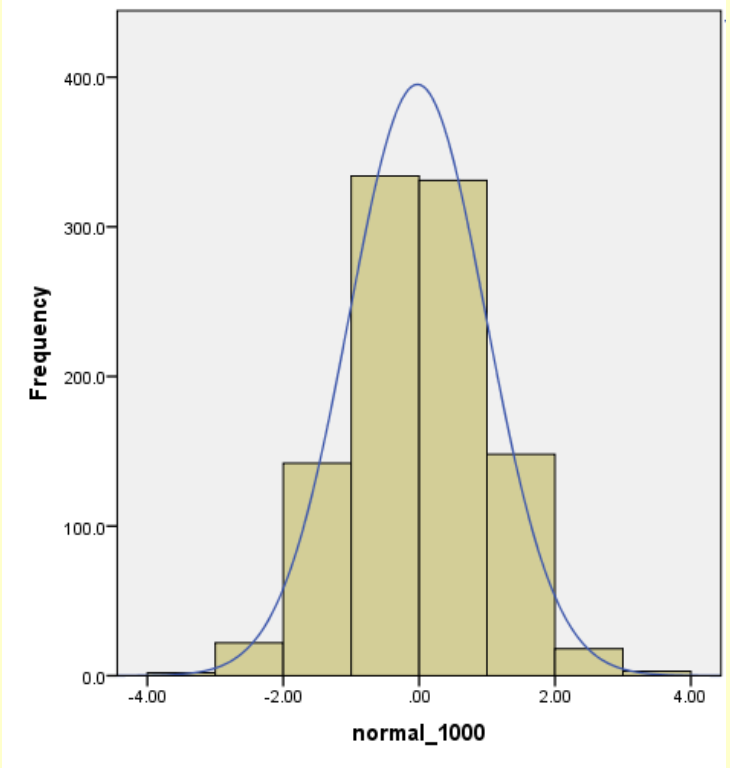
Statistics

n25

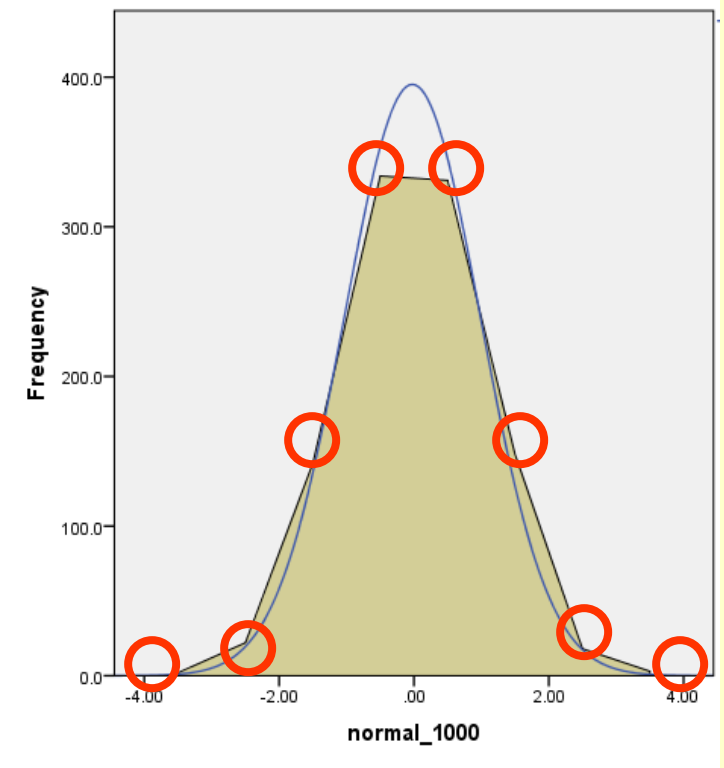
N	Valid	25
	Missing	975
Mean		-.0074
Std. Error of Mean		.24943
Median		-.1697
Mode		-2.42[a]
Std. Deviation		1.24716
Skewness		-.055
Std. Error of Skewness		.464
Kurtosis		-1.229
Std. Error of Kurtosis		.902
Sum		-.19

a. Multiple modes exist. The smallest value is shown

PSA – 3: Histograms vs Line Graphs

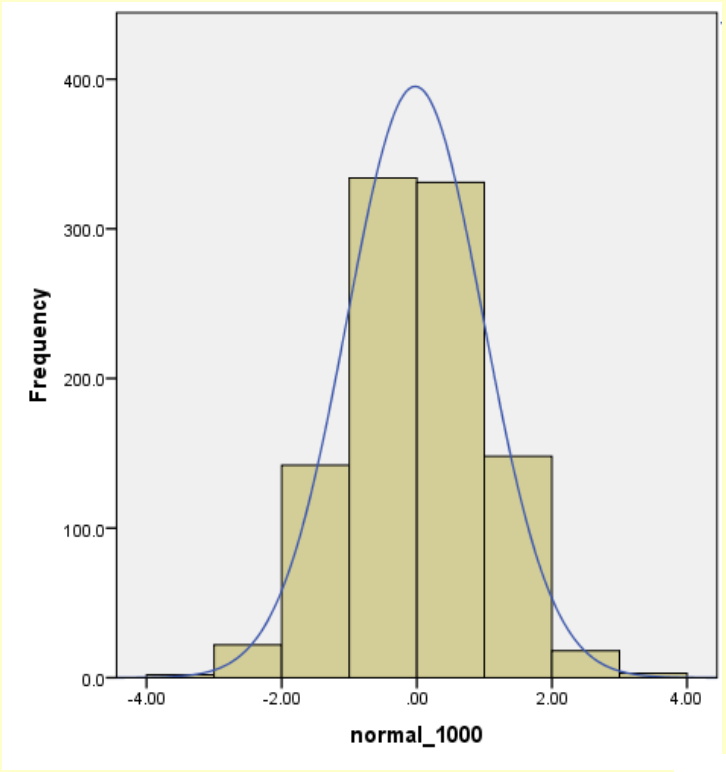


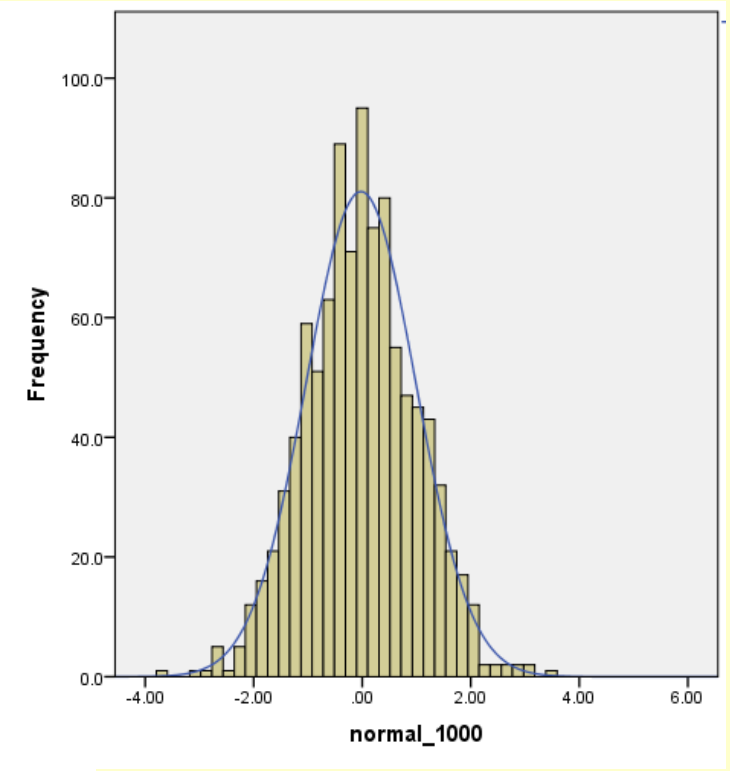
Bin Size = 1



Bin Size = 1

PSA – 4: Frequency Histograms





Mean = -.0234
Std. Dev. = 1.00944
N = 1,000

Bin Size = 1

Bin Size = 0.1

Homework #3

Open the file (BIOL4090_Hw3_data.xls) and use this information from a random normal variable (with a mean of 0 and a S.D. of 1) for the following exercise.

Analyze the values in the "small sample size" sheet. You may use SPSS or another program to calculate the following parameters. Paste the results in the table below.

Homework #3

Tabular Descriptive Statistics

Statistics

n25

N	Valid	25
	Missing	975
Mean		-.0074
Std. Error of Mean		.24943
Median		-.1697
Mode		-2.42[a]
Std. Deviation		1.24716
Skewness		-.055
Std. Error of Skewness		.464
Kurtosis		-1.229
Std. Error of Kurtosis		.902
Sum		-.19

a. Multiple modes exist. The smallest value is shown

Statistics

n50

N	Valid	50
	Missing	950
Mean		.1209
Std. Error of Mean		.17241
Median		.2147
Mode		-2.42[a]
Std. Deviation		1.21910
Skewness		-.047
Std. Error of Skewness		.337
Kurtosis		-.628
Std. Error of Kurtosis		.662
Sum		6.04

a. Multiple modes exist. The smallest value is shown

Statistics

n250

N	Valid	250
	Missing	750
Mean		.0480
Std. Error of Mean		.06702
Median		.0475
Mode		-2.64[a]
Std. Deviation		1.05971
Skewness		.060
Std. Error of Skewness		.154
Kurtosis		-.108
Std. Error of Kurtosis		.307
Sum		12.00

a. Multiple modes exist. The smallest value is shown

Statistics

n500

N	Valid	500
	Missing	500
Mean		-.0199
Std. Error of Mean		.04600
Median		.0001
Mode		-2.85[a]
Std. Deviation		1.02852
Skewness		.082
Std. Error of Skewness		.109
Kurtosis		.008
Std. Error of Kurtosis		.218
Sum		-9.96

a. Multiple modes exist. The smallest value is shown

Homework #3
Statistics Depend on Sample Size

Sample Size (n)	Mean	Median	Mode	S.D.	S.E. (show your work)
25	-0.0074	-0.1697	-2.42	1.24716	= 1.24716 / sqrt(25) = 0.24943
50	0.1209	0.2147	-2.42	1.21910	= 1.21910 / sqrt(50) = 0.17241
250	0.0480	0.0475	-2.64	1.05971	= 1.05971 / sqrt(250) = 0.06702
500	-0.0199	0.0001	-2.85	1.02852	= 1.02852 / sqrt(500) = 0.04600

Homework #3

Sample Size (n)	S.E. (show your work)	95% CI (show your work)
25	= 1.24716 / sqrt(25) = 0.24943	= mean +/- (0.24943 * 1.96) = -0.0074 +/- 0.48888 = from -0.49629 to 0.481487
50	= 1.21910 / sqrt(50) = 0.17241	= mean +/- (0.17241 * 1.96) = 0.1209 +/- 0.33791 = from -0.21702 to 0.45881
250	= 1.05971 / sqrt(250) = 0.06702	= 0.0480 +/- (0.06702 * 1.96) = 0.0480 +/- 0.13136 = from -0.08336 to 0.17936
500	= 1.02852 / sqrt(500) = 0.04600	= -0.0199 +/- (0.04600 * 1.96) = -0.0199 +/- 0.09016 = from -0.01100 to 0.07025

Homework #3

Finally, explain your results:

How well do the means of the four small samples (n = 25, 50, 250, 500) estimate the mean of the large population sample (n = 1000)?

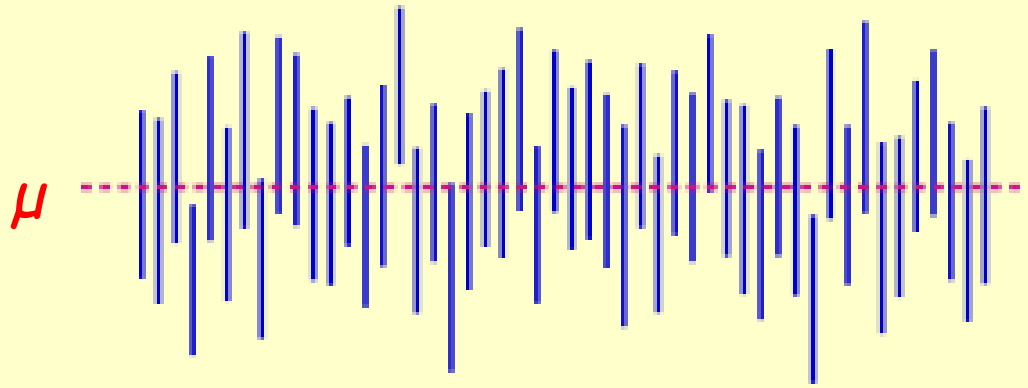
Very well: All the means are very close to the real "mu", since all the 95% CIs overlap "0"

Sample Size (n)	Mean	S.D.	S.E.
25	-0.0074	1.24716	0.24943
50	0.1209	1.21910	0.17241
250	0.048	1.05971	0.06702
500	-0.0199	1.02852	0.04600

95% Confidence Intervals

Estimating the parameter (population mean)=

Lower bound: Point Estimate – (1.96 * SE)
Upper bound: Point Estimate + (1.96 * SE)



By definition: 95% of the C.I, overlap real parameter.

Determine whether the 95% C.I. overlaps μ.

Homework #4

Open the file BIOL4090_Hw4_data.xls with SPSS and use these observations, drawn from three random variable distributions (derived from theoretical distributions with mean = 10 and variance =10) for the following exercise.

Make sure variables are "numeric" and measure is "scale".

Create a frequency table of each dataset of 100 data points each and use this information to fill in the table below.

Homework #4

Statistics

		distribution_1	distribution_2	distribution_3
N	Valid	100	100	100
	Missing	0	0	0
Mean		9.724613	9.720000	9.705097
Median		9.644506	10.000000	9.761769
Mode		2.6396[a]	7.0000	1.5995[a]
Std. Deviation		2.9433543	3.0486957	3.0334299
Skewness		.625	.180	-.082
Std. Error of Skewness		.241	.241	.241
Kurtosis		.843	.362	-.095
Std. Error of Kurtosis		.478	.478	.478
Percentiles	5	5.207715	5.000000	4.936689
	95	15.799824	14.950000	15.167130

a. Multiple modes exist. The smallest value is shown

Homework #4

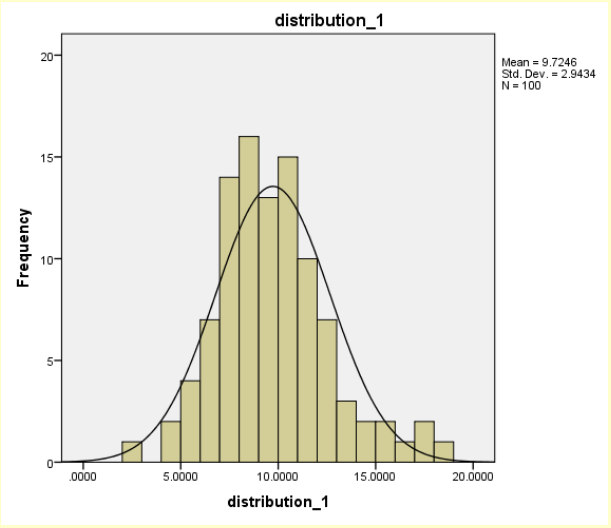
Tests of Normality

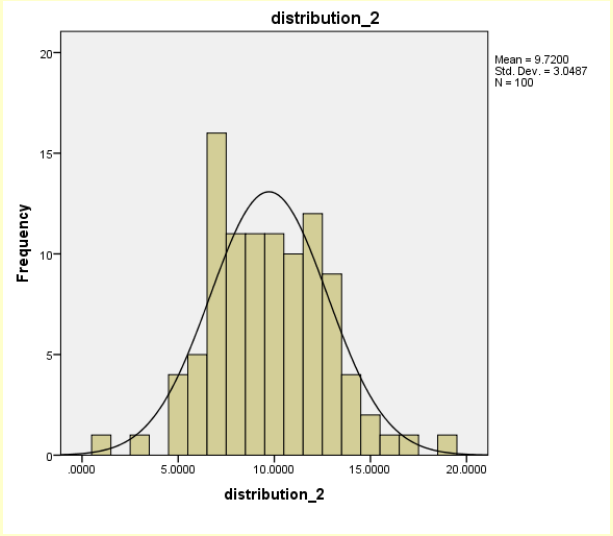
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
distribution_1	.079	100	.126	.970	100	.022
distribution_2	.094	100	.031	.981	100	.154
distribution_3	.035	100	.200*	.995	100	.968

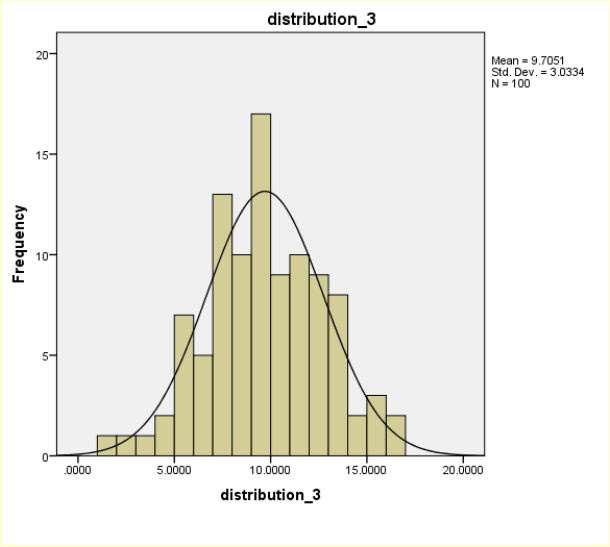
a. Lilliefors Significance Correction
*. This is a lower bound of the true significance.

Graphical Inspection of Data

Consider three random variable distributions (theoretical distributions, mean = 10, variance =10).







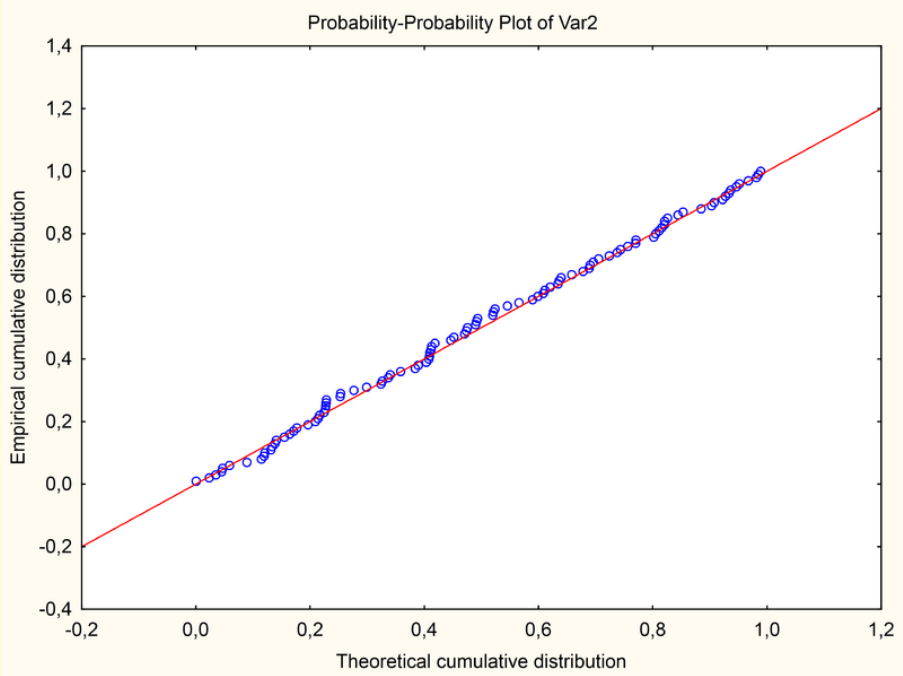
Statistics

		distribution_1	distribution_2	distribution_3
N	Valid	100	100	100
	Missing	0	0	0
Mean		9.724613	9.720000	9.705097
Median		9.644506	10.000000	9.761769

P-P Plots

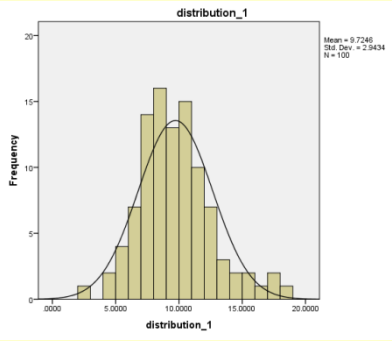
"Visual" tests are more intuitively appealing but are based on a subjective assessment and do not accept or reject the null hypothesis.

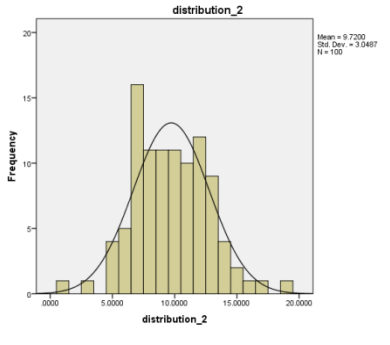
P-P plot:
For normally distributed data this plot should lie on a 45° line between (0, 0) and (1, 1).

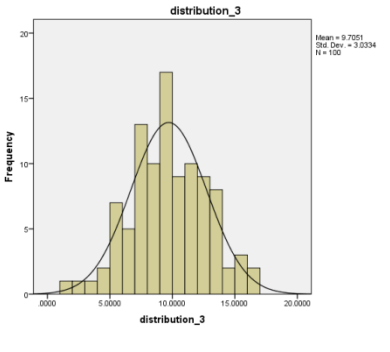


Diagnostics

Consider three random variable distributions (theoretical distributions, mean = 10, variance = 10).







Statistics

		distribution_1	distribution_2	distribution_3
N	Valid	100	100	100
	Missing	0	0	0
Mean		9.724613	9.720000	9.705097
Median		9.644506	10.000000	9.761769
Mode		2.6396[a]	7.0000	1.5995[a]
Std. Deviation		2.9433543	3.0486957	3.0334299
Skewness		.625	.180	-.082
Std. Error of Skewness		.241	.241	.241
Kurtosis		.843	.362	-.095
Std. Error of Kurtosis		.478	.478	.478

Check rule of thumb:

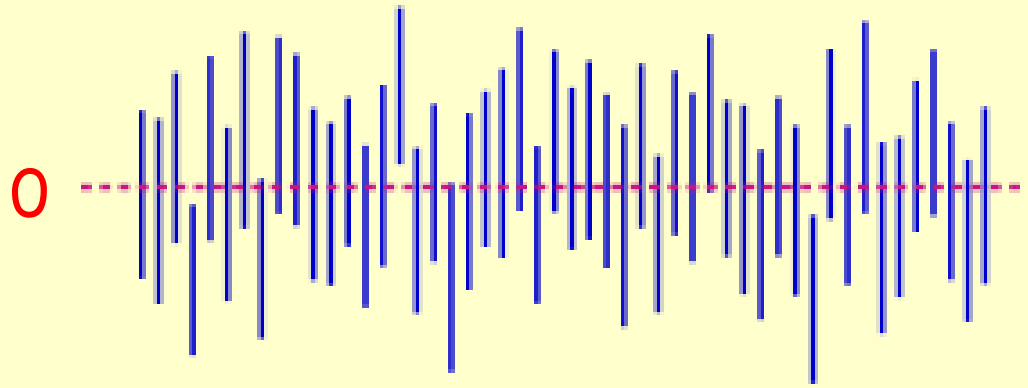
$-1 < \text{skew} < 1$

$-1 < \text{kurtosis} < +1$

95% Confidence Intervals

Formulation for Kurtosis and Skewness =

Lower bound: Point Estimate – (1.96 * SE)
Upper bound: Point Estimate + (1.96 * SE)



By definition: 95% of the C.I, overlap real parameter.

Determine whether the 95% C.I. overlaps 0.

Example - Diagnostics

skewness

Distribution	Skew	SE	lower CI	upper CI
1	0.625	0.241	0.153	1.097
2	0.180	0.241	-0.292	0.652
3	-0.082	0.241	-0.554	0.390

kurtosis

Distribution	Kurtosis	SE	lower CI	upper CI
1	0.843	0.478	-0.094	1.780
2	0.362	0.478	-0.575	1.299
3	-0.095	0.478	-1.032	0.842

Descriptive Tests

The **Kolmogorov–Smirnov (K – S)** is a nonparametric test for continuous probability distributions.

Used to compare a sample distribution (observed) with a reference probability distribution (theoretical).

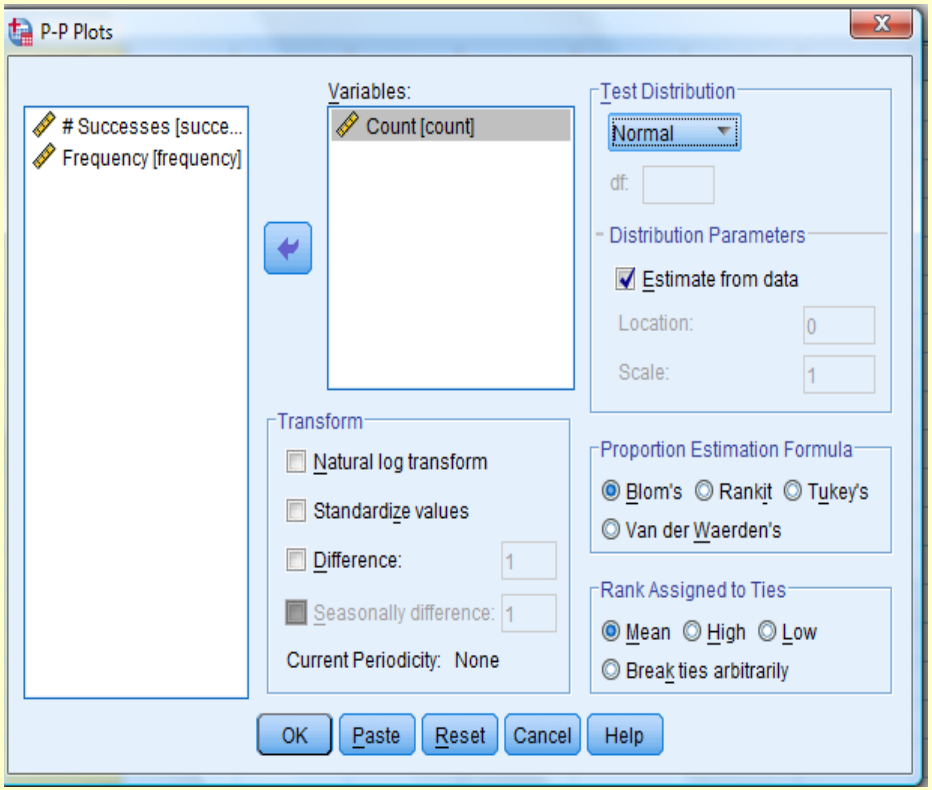
The K-S statistic quantifies the distance between the empirical distribution (sample) and a reference distribution (with the same mean and variance).

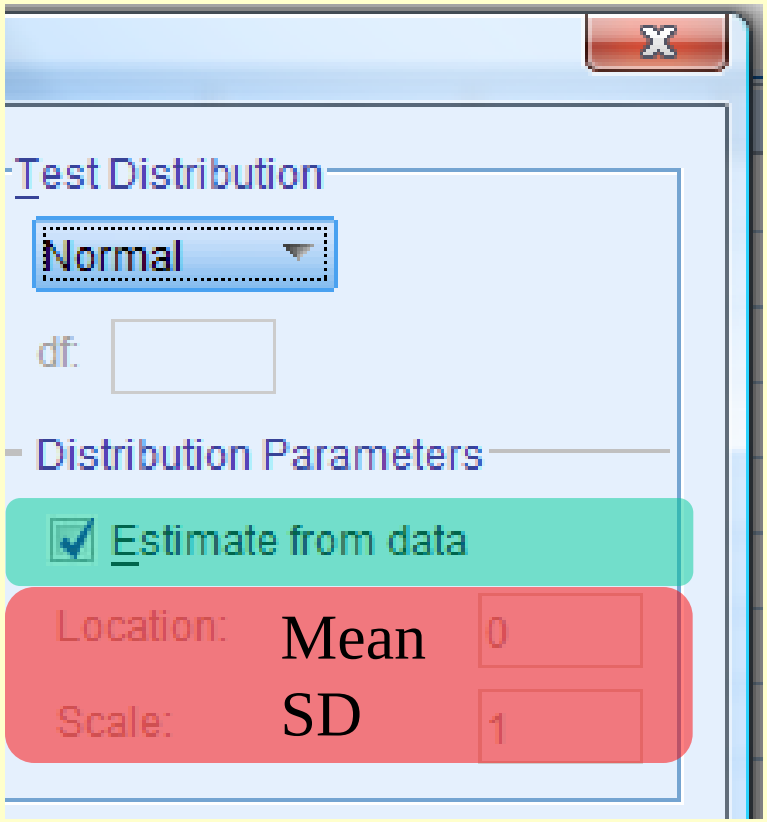
The null hypothesis of this test states that the sample is drawn from the reference distribution.

Normality Tests

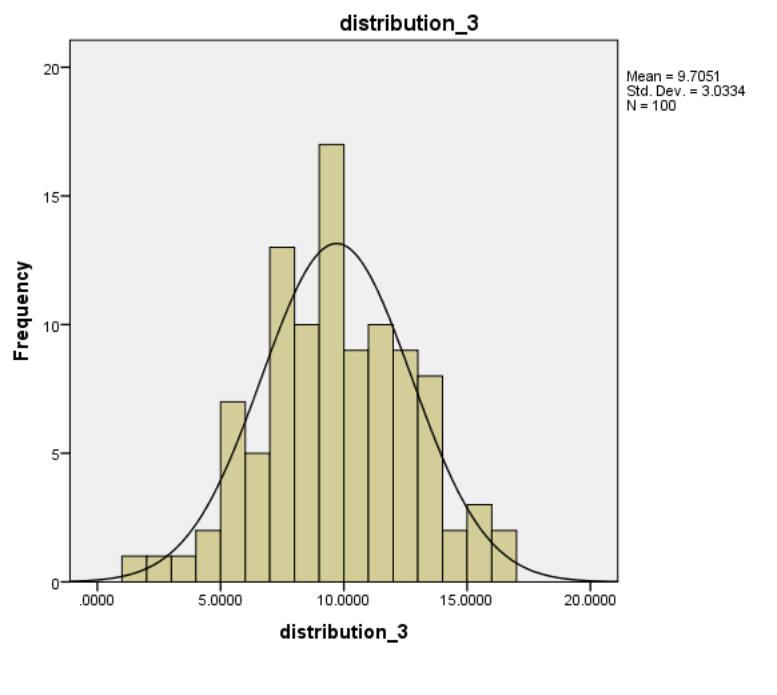
Select Parameters describing the Normal

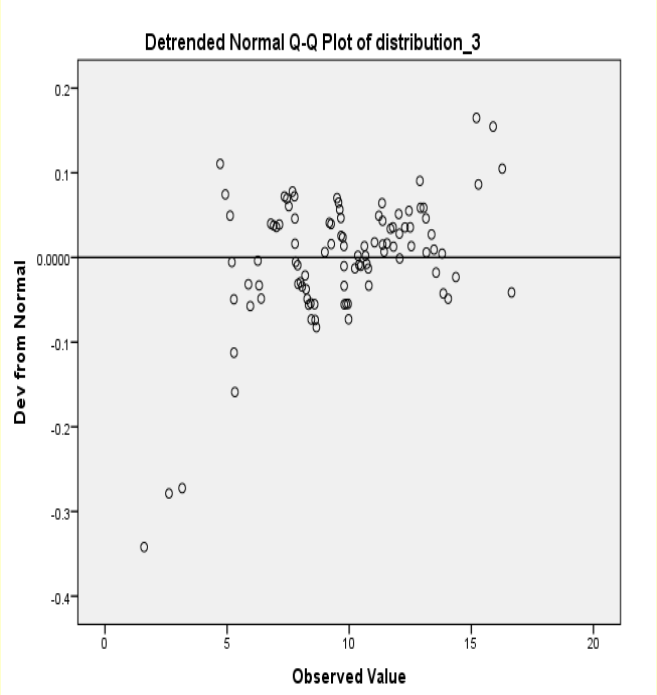
(estimated from the data / determined by user)





The Practice – Distribution 3





Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
distribution_3	.035	100	.200*	.995	100	.968

Not Significant

Not Significant

Outcome – Distribution 3

Visual Inspection: Unimodal, Bell-shaped

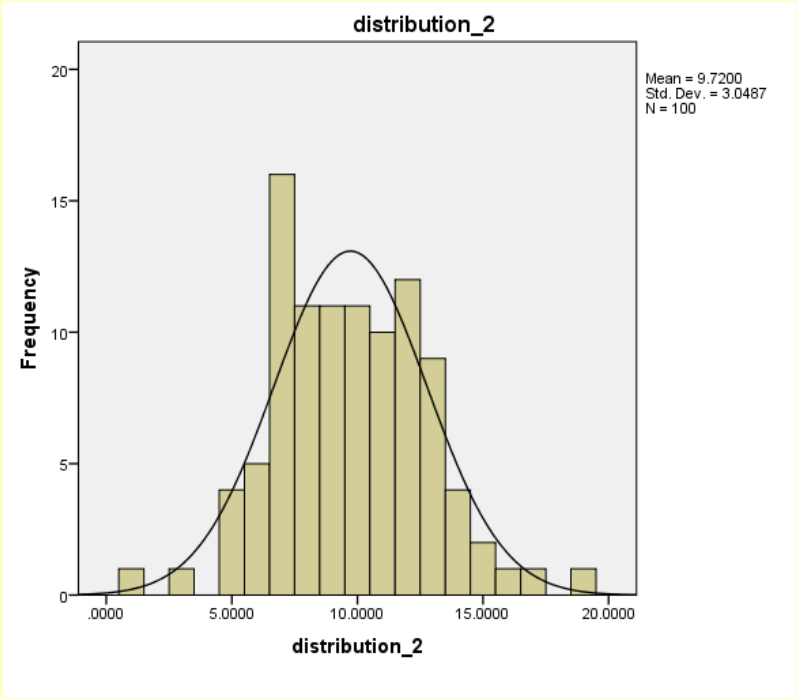
Skewness / Kurtosis C.I.: Not Significant

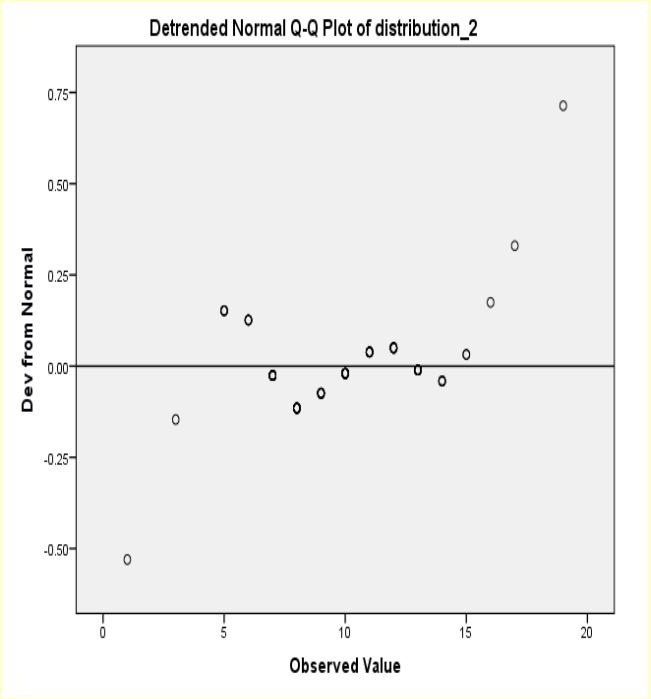
K.S. test.: Not Significant

S.W. test.: Not Significant

Result: normal distribution

The Practice – Distribution 2





Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
distribution_2	.094	100	.031	.981	100	.154

Significant

Not Significant

Outcome – Distribution 2

Visual Inspection: Unimodal, Bell-shaped

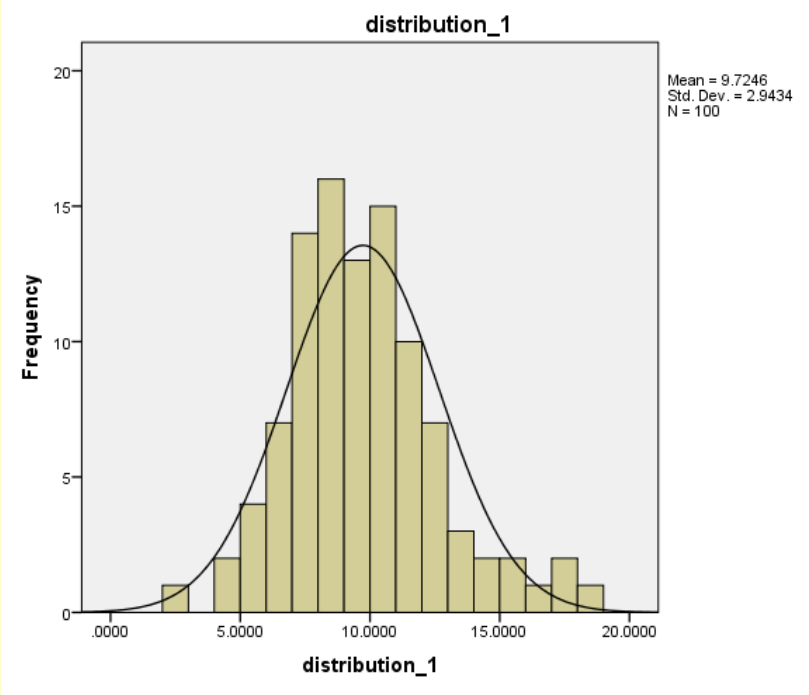
Skewness / Kurtosis C.I.: Not Significant

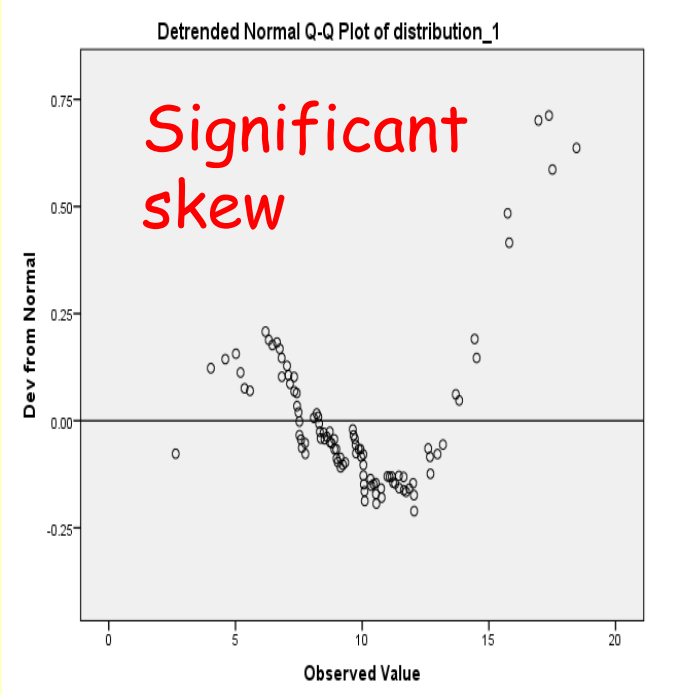
K.S. test.: Significant

S.W. test.: Not Significant

Result: Poisson distribution (lambda = 10)

The Practice – Distribution 1





Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
distribution_1	.079	100	.126	.970	100	.022

Not Significant Significant

Outcome – Distribution 1

Visual Inspection: Unimodal, Bell-shaped

Skewness / Kurtosis C.I.: Skewness Significant

K.S. test.: Not Significant

S.W. test.: Significant

Result: log-normal distribution

Outcome – Recommendations

If one or both descriptors (Skewness / Kurtosis) significant, transform the data

If both tests (KS and SW) agree – No problem

What if both tests (KS and SW) do not agree?

SW or KS ?

How do the KS and SW tests differ?

- KS: less power, tests difference of cumulative frequency distributions

- SW: more power, tests slope of the p-p plot line

Recommendation:

If there are discrepancies in the significance of the results, select test with the highest power (SW)

Remember: Power is the probability of detecting a false null hypothesis (correctly rejecting the Ho)

Take-home Lessons

Data transformations are one of the most difficult issues in parametric statistics:

- Conflicting advice: transform or not

- Conflicting results: various normality tests

Recommendation:

Select one approach that provides multiple evidence and come up with criteria before starting analysis

Be as strict as you wish: one or more criteria

But, if a test significant… cannot back-track

Take-home Lessons

- Parametric tests are more powerful, but are based on assumption of normally distributed data

- Determine normality criteria and undertake data transformations, if needed

- If you are unsure, data transformations can always be attempted to compare the same test results, using transformed and un-transformed data

- Test normality before / after data transformations

- If transformations do not work…

  use non-parametric tests