General statistical principles

Descriptive statistics

- Used to describe the nature of your data.

Inductive statistics

- The use of descriptive statistics to make a statement, prediction or decision.

Descriptive statistics are commonly reported but both are needed to interpret results.

Error and uncertainty

Error - difference between your answer and the 'true' one.

Systematic - problem with a method, all errors are of the same size, magnitude and direction. Easy to correct.

Random - based on limits and precision of a measurement. Can be treated statistically.

Blunders - you screw up. Best to just repeat the work.

Probability

A character associated with an event
- its tendency to take place.

To see what we're talking about, let's use an example - a 10 sided die.

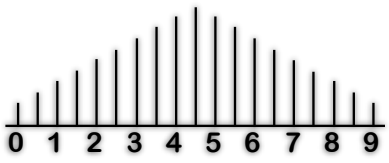
If a person had a die and gave it a series of rolls, what would be the expected result?

One roll of the die

For a single roll, each value is equally likely to come up. What if a person had two dice?

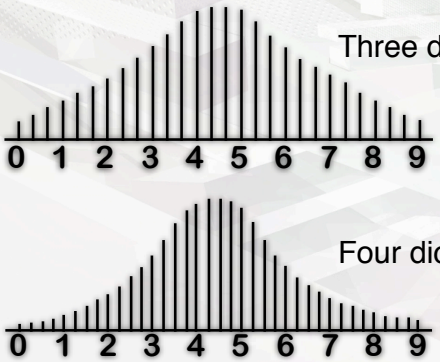
Two dice

Average of two dice - one roll



We can continue this trend, using more dice and a single roll.

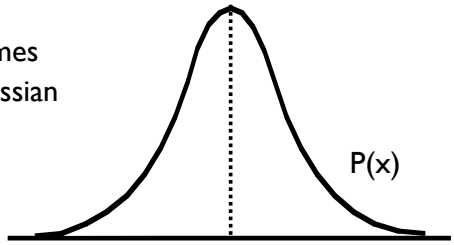
More dice



Even more dice

As the number of dice approaches infinity, the values become continuous.

The curve becomes a normal or Gaussian distribution

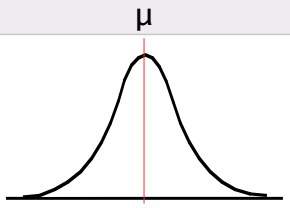


Normal distribution

The distribution can be described by:

$$f_{(x)} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$-\infty < x < +\infty$$



σ - standard deviation
μ - universal mean

An infinite data set is required.

Normal distribution

These terms can be calculated by:

Universal mean $\mu = \sum_{i=1}^{N} \frac{x_i}{N}$

Variance $\sigma^2 = \sum_{i=1}^{N} (x_i - \mu)^2$

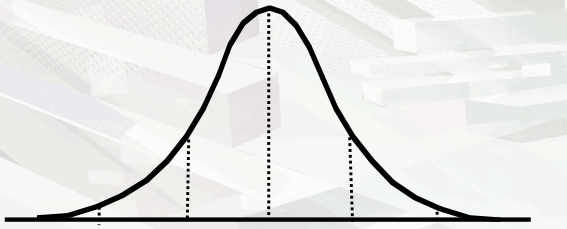
Standard deviation $\sigma = \sqrt{\sigma^2}$

These are for infinite data sets but can be used for data sets where N > 100 and any variation is truly random in nature.

Normal distribution

- Both variance and standard deviation measure the dispersion of data around the universal mean.
- Standard deviation is usually easier to use because it has the same units as the original data - provides more weight to the central data.
- Variance will have units that are the square of the original data. It is a more sensitive measure of dispersion - providing more weight to the outlying data.
- Variances are additive, Standard deviations are not. That why we'll manipulate data via variance no SD.

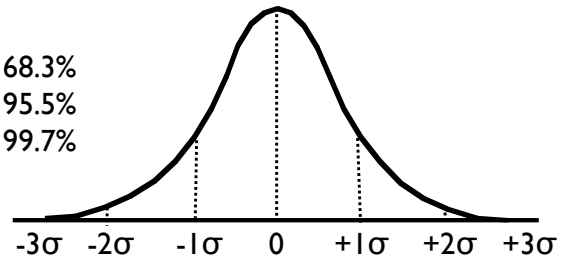
Normal distribution



When determining σ, you are assuming a normal distribution and relating your data to σ units.

Normal distribution areas

± 1σ = 68.3%
± 2σ = 95.5%
± 3σ = 99.7%



The area under any portion of the curve tells you the probability of an event occurring.

Large data sets

We can use the normal distribution curve to predict the likelihood of an event occurring.

This approach is only valid for large data sets and is useful for things like quality control of mass produced products.

In the following examples, we will assume that there is a very large data set and that μ and σ have been tracked.

The reduced variable

Assuming you know μ and σ for a dataset, you can calculate u (the reduced variable) as:

$$u = \frac{(x - \mu)}{\sigma}$$

This is simply converting your test value from your normal units (mg, hours, ...) to standard deviation.

Using the reduced variable

Assuming that your data is normally distributed, you can use u to predict the probability of an event occurring.

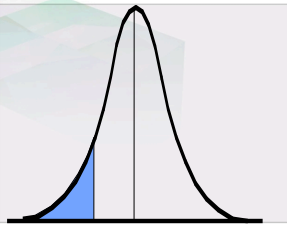
The probability can be found by looking up u on a table - Form A and Form B. Many spreadsheets also allow for calculation of these values.

Which form you use is based on the question being asked.

Form A

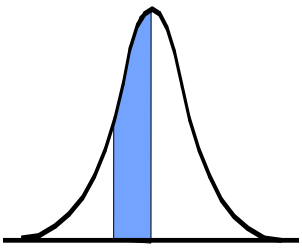
This form will give you the area under the curve from u to infinity.

|u|	area	|u|	area
0.0	0.5000	2.0	0.0227
0.2	0.4207	2.2	0.0139
0.4	0.3446	2.4	0.0082
0.6	0.2743	2.6	0.0047
0.8	0.2119	2.8	0.0026
1.0	0.1587	3.0	1.3×10^{-3}
1.2	0.1151	4.0	3.2×10^{-5}
1.4	0.0808	6.0	9.9×10^{-10}
1.6	0.0548	8.0	6.2×10^{-16}
1.8	0.0359	10.0	7.6×10^{-24}



Form B

This form will give you the area under the curve from 0 to u



|u|	area	|u|	area
0.0	0.0000	1.6	0.445
0.2	0.0793	1.8	0.464
0.4	0.1554	2.0	0.477
0.6	0.2258	2.2	0.486
0.8	0.2881	2.4	0.491
1.0	0.3413	2.6	0.495
1.2	0.3849	2.8	0.497
1.4	0.4192	3.0	0.498

Example

A tire is produced with the following statistics regarding usable milage.

μ = 58000 miles
σ = 10000 miles

What mileage should be guaranteed so that less than 5% of the tires would need to be replaced?

Example

Looking on Form A, we find that an area of 0.05 comes closest to 1.6 σ.

Now use the reduced variable equation using a σ of -1.6 (we want the value to be less than the mean.)

$$-1.6 = (x - 58000) / 10000$$

$$x = 42000 \text{ miles}$$

Using Excel

The same problem can be solved for using a spreadsheet like Excel.

This problem can be solved using the function, NORMINV.

It provides more accuracy because it is not limited to the 'resolution' of the table.

Using Excel

stats.exl

	B	C	D	E
1				
2	NORMINV (probability, mean, Standard deviation)			
3	Returns the inverse of the normal cumulative distribution for			
4	the specified mean and standard deviation.			
5				
6	mean	58000	miles	
7	Stdev	10000	miles	
8	Failure rate	0.05	(5 percent)	
9				
10	NORMINV (0.05, 58000,10000)			
11	41551.4652	miles		
12				

Sheet1 Sheet2 Sheet3

Another example

You install a pH electrode to monitor a process stream. The manufacturer provided you with the follow values regarding electrode life time:

$$\mu = 8000 \text{ hours}$$

$$\sigma = 200 \text{ hours}$$

If you needed to replace the electrode after 7200 hours of use, was the electrode 'bad?.

Another example

$u = (7200 - 8000) / 200 = -4.0\ \sigma$

Looking on form A, we find that the probability at a value of 4 is 3.2×10^{-5}

That means that only 0.0032 % of all electrodes would fail at 7200 hours or earlier. You have a bad electrode.

Using Excel

stats.exl

	A	B	C	D
1	NORMDIST(x,mean,standard,TRUE)			
2	Returns the normal cumulative distribution for the			
3	specified mean and standard deviation.			
4				
5	x is the value for which you want a distribution.			
6				
7	mean	8000	hours	
8	Stdev	200	hours	
9	Failure time	7200	hours	
10				
11	NORMDIST(7200,8000,200,TRUE)			
12	3.1686E-05	0.0031686	percent	
13				

Sheet1 Sheet2 Sheet3

More on Excel functions.

NORMDIST

Will determine the area beyond a specific point on a Gaussian curve.

NORMINV

Inverse of NORMDIST. Used to find the specific X value that will result in the target area.

Both work with 'Form A' values. Form B values can be found by subtracting areas from one (1).

Smaller data sets

$$mean = \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

$$variance = s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

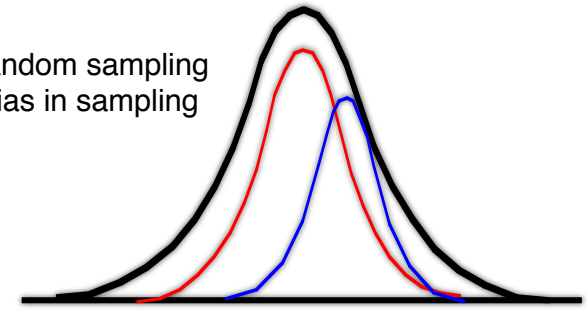
Smaller data sets

When we use smaller data sets, we must be concerned with:

- Are the samples representative of the population? The values must be truly random.
- If we pick a non-random set like all men or women, are the differences significant?

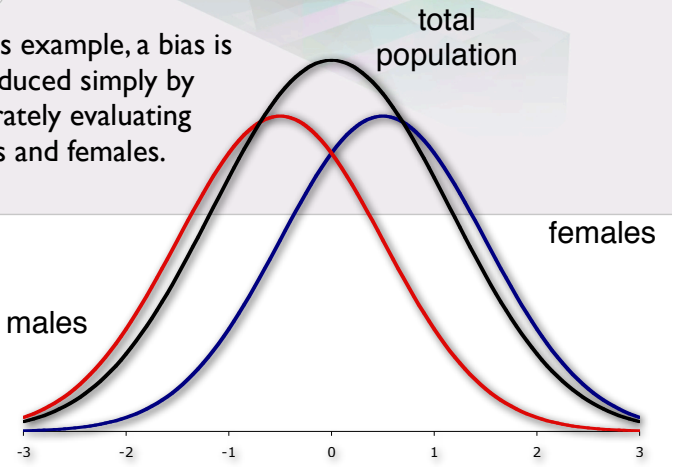
This example shows what can happen if a bias is introduced during sampling.

Red - random sampling
Blue - bias in sampling

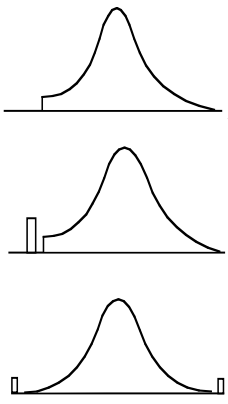


Smaller data sets

In this example, a bias is introduced simply by separately evaluating males and females.



Other types of bias



detection limit bias

nominal value bias

outliers

Smaller data sets

Introducing a bias is not always a bad thing.

The bias could be due to a true difference between the populations

It also could be due to poor sampling of a population.

We need tools to tell the difference.

Univariate tools

For a normal distribution:

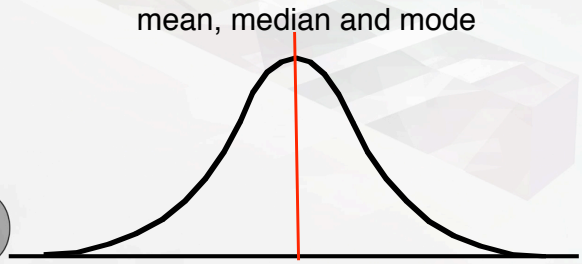
Mean
- numerical average of values

Median
- central tendency
- center value of ranked data
- actual value if N is odd
- average of two center value if N is even.

Mode
- most frequent value

Univariate tools

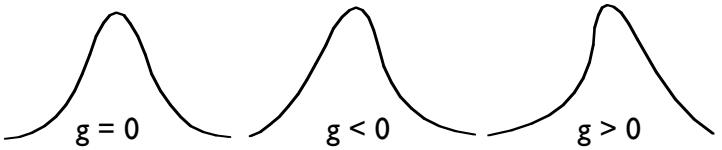
Ideally, all three values should be the same.
If not the same, at least very close



Skewness

This is a test to see if a population is Gaussian.

$$g = \frac{\sum_{i=1}^{N}(x-\bar{x})^3}{N s_x^3}$$

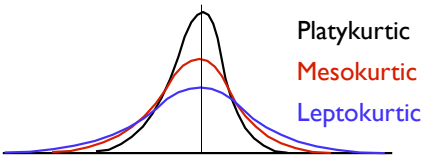


Kurtosis

Closely related to skewness.

While skewness measures a bias in the distribution, kurtosis measures how 'flat' a distribution is.

$$kurtosis = \frac{\sum(x-\bar{x})^4}{N s_x^4} - 3$$



Using Kurtosis or Skewness

Divide the skew (or kurtosis) figure by the standard error of the skew (kurtosis).

If the value is greater than 1.96 or less than -1.96 your data is significantly skewed (kurtotic)

Using Kurtosis or Skewness

$$SE_{Skew} = \sqrt{\frac{6}{N}}$$

$$Z = \frac{g}{SE_{Skew}}$$

$$SE_{Kertosis} = \sqrt{\frac{24}{N}}$$

For small data sets

variance $s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{(n-1)}$

standard deviation $s = \sqrt{s^2}$

degrees of freedom $\varphi = n - 1$

standard deviation of the mean $s_{\bar{x}} = \frac{s_x}{\sqrt{n}}$

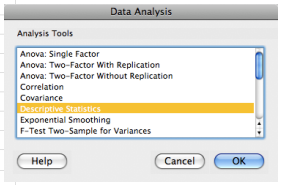
coefficient of variance $cv = \frac{s_x}{\bar{x}}$

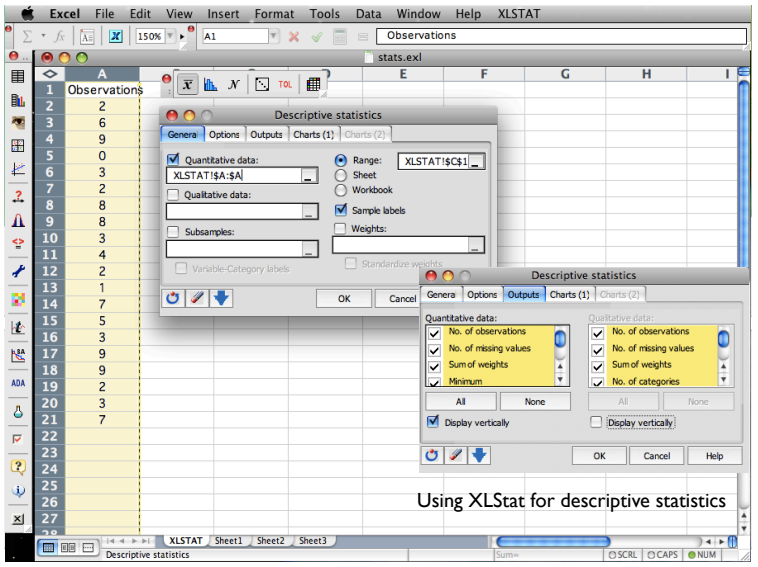
relative standard deviation $RSD = 100\,\frac{s_x}{\bar{x}}$

Using Excel for descriptive statistics

stats.exl

	A	B	C	D	E	F	G
1	Observations	Obtained from analysis add-on					
2	2	*Descriptive statistics*				Excel function	
3	6						
4	9		Mean	4.65		average(range)	
5	0		Median	3.5		median(range)	
6	3		Mode	2		mode(range)	
7	2		Standard Deviation	2.96		stdev(range)	
8	8		Sample Variance	8.77		var(range)	
9	8		Kurtosis	-1.41		kurt(range)	
10	3		Skewness	0.25		skew(range)	
11	4		Range	9		range(range)	
12	2		Minimum	0		min(range)	
13	1		Maximum	9		max(range)	
14	7		Sum	93		sum(range)	
15	5		Count	20		count(range)	
16	3						
17	9						
18	9		STDEV Mean	0.148		stdev/sqrt(20)	
19	2		Coefficient of variance	0.6367		stdev/average	
20	3		Relative STD	63.671		stdev/average * 100	
21	7						



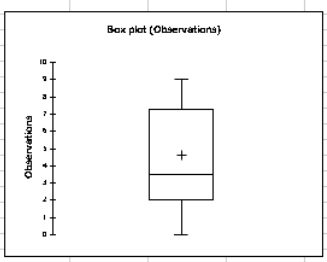


Using XLStat for descriptive statistics

Using XLStat for descriptive statistics

stats.exl

	A	C	D
1	Observations	Descriptive statistics (Quantitative data):	
2	2		
3	6	Statistic	Observations
4	9	No. of observations	20
5	0	No. of missing values	0
6	3	Sum of weights	20
7	2	Minimum	0.000
8	8	Maximum	9.000
9	8	Freq. of minimum	1
10	3	Freq. of maximum	3
11	4	Range	9.000
12	2	1st Quartile	2.000
13	1	Median	3.500
14	7	3rd Quartile	7.250
15	5	Sum	93.000
16	3	Mean	4.650
17	9	Variance (n)	8.328
18	9	Variance (n-1)	8.766
19	2	Standard deviation (n)	2.886
20	3	Standard deviation (n-1)	2.961
21	7	Variation coefficient	0.621
22		Skewness (Pearson)	0.226
23		Skewness (Fisher)	0.245
24		Skewness (Bowley)	0.429
25		Kurtosis (Pearson)	-1.369
26		Kurtosis (Fisher)	-1.412
27		Standard error of the mean	0.662
28		Lower bound on mean (95%)	3.264
29		Upper bound on mean (95%)	6.036
30		Mean absolute deviation	2.615



Degrees of freedom

φ *or df =n −# of parameters*

Example.

If you had 10 measurements, they could be used in pairs to obtain 9 different measurements of the mean.

If you tried to do a 10th, one of the pairs would have already been used - same standard deviation.

Degrees of freedom

When doing a linear regression fit of a line using X,Y data pairs, the model used

(Y = mX + b) results in two parameters.

The degrees of freedom would be N-2 in this case (or model).

So the model used determines the degrees of freedom.

Pooled statistics

In many cases it is necessary to combine results:

- values from separate labs
- data collected on separate days
- a different instrument was used
- a different method of analysis was used

When we combined the data, we refer to this as 'pooling' the data.

Pooled statistics

- We can't simply combine all of the values and calculate the mean and other statistical values.
- There may have been differences with the results obtained.
- There might have been different numbers of samples collected with each set.
- We also would like a way to tell if the results are significantly different.

Pooled statistics

$$s_p = \sqrt{\frac{df_1 s_1^2 + df_1 s_2^2 + \ldots + df_1 s_k^2}{df_1 + df_2 + \ldots + df_k}}$$

The pooled standard deviation is weighted by the degrees of freedom. This accounts for the number of data points and any parameters used in obtaining the results.

Example

Set	s	n	φ	s^2	φs^2
1	1.35	10	9	1.82	16.4
2	2.00	7	6	4.00	24.0
3	2.45	6	5	6.00	30.0
4	1.55	12	11	2.40	26.4

$$s_p = \sqrt{\frac{16.4 + 24.0 + 30.0 + 26.4}{9 + 6 + 5 + 11}} = 1.77$$