

Descriptive Statistics

Once an experiment is carried out and the results are measured, the researcher has to decide whether the results of the treatments are different. This would be easy if the results were perfectly consistent:

Corn heights for Treatment 1 (in): 32, 32, 32, 32, 32, 32, 32, 32
Corn heights for Treatment 2: (in) 36, 36, 36, 36, 36, 36, 36, 36
Obviously treatment 2 results in taller corn.

Unfortunately, real life results are not so simple:

Corn heights for Treatment 1 (in): 28, 32, 36, 39, 25, 30, 34, 32
Corn heights for Treatment 2 (in): 33, 32, 38, 36, 40, 39, 34, 36
Differences are not obvious, so we need statistics.

Statistics describe a sample and use Latin symbols.

Parameters describe a population and use Greek symbols.

Statistics are used when individual characteristics are variable. Note that you have to measure several individuals to know how variable they are, in other words, you need replication.

Some physical properties are very consistent, that is, they have low variability. An example might be the speed at which steel balls fall in a vacuum - the biggest source of variability is likely to be the accuracy of the timing device.

How many sheets of paper do you need to measure to know the average length of the sheets of paper in a ream? How many do you need to measure to know whether the lengths of sheets in 2 reams are the same or different?

Biological properties, on the other hand, tend to have a fairly high variability due to the many variations in genetics and environment even within a single species.

How many class participant heights do you need to measure to know the average height of people in a classroom? How many do you need to measure to know whether the average heights of people on the left and right sides of the room are the same or different?

Most populations and samples follow a normal or Gaussian distribution that looks like a bell-shaped curve.

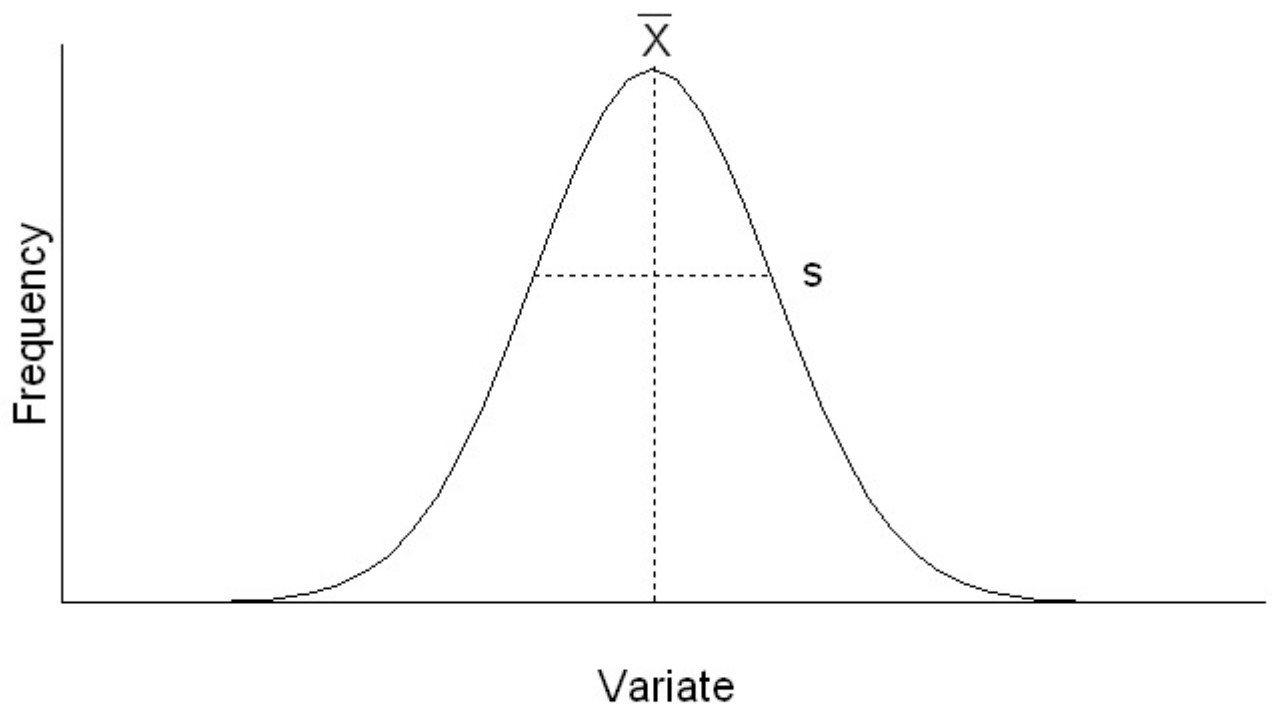
Normal Distribution

Describing a normal distribution

The distribution is described by:

Mean: central value

Population $\mu = \frac{\sum X}{N}$ Sample $\bar{X} = \frac{\sum X}{n}$



Variance: measure of the scatter about the central value

Population $\sigma^2 = \frac{\sum (X - \bar{X})^2}{N}$ Sample $s^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$

Sums of squares: squares of deviations from the mean, numerator of the variance formula

$SS = \sum (X - \bar{X})^2$ _____

Degrees of freedom: number of independently calculated differences from the mean. If

the mean is calculated from the sample, the value of the last observation can be calculated from the mean and the values of all the other observations. Thus there are only $n - 1$ independent observations.

$$df = n - 1$$

Standard deviation: measure of the scatter of the individuals about the central value. Units are the same as for the original data.

$$\text{Population } \sigma = \sqrt{\sigma^2} \quad \text{Sample } s = \sqrt{s^2}$$

A higher variance or standard deviation describes a wider curve.

Coefficient of Variation: ratio of the size of the scatter (standard deviation) to the size of the mean, expressed as percent. Allows comparisons of relative variability of different populations, for example, whether the weights of elephants are more variable than the weights of mice.

$$CV = \frac{s}{\bar{X}} * 100\%$$

The statistics calculated so far describe samples and populations, but do not test for differences between samples and populations. For such tests the distributions of sample means are needed.

Example: Speed (-2 minutes) in seconds of two calculating machines for computing sums of squares

Replication	Machine A			Machine B			A - B	Rep. or Pair Totals
	Time (sec)	Deviation from Mean	Dev. ²	Time (sec)	Deviation from Mean	Dev. ²		
1	30*	8	64	14	0	0	16	44
2	21	-1	1	21*	7	49	0	42
3	22*	0	0	5	-9	81	17	27
4	22	0	0	13*	-1	1	9	35
5	19	-3	9	14*	0	0	5	33
6	29*	7	49	17	3	9	12	46
7	17	-5	25	8*	-6	36	9	25
8	14*	-8	64	16	2	4	-2	30
9	23*	1	1	8	-6	36	15	31
10	23	1	1	24*	10	100	-1	47
Total	$\Sigma X=220$		$\Sigma x^2=214$	$\Sigma X=140$		$\Sigma x^2=316$	$\Sigma=80$	

where $x = X - \bar{X}$

$$\text{Sums of Squares} = SS = \Sigma x^2 = \Sigma (X - \bar{X})^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

$$\text{Mean} = \bar{X} = \frac{\Sigma X}{n}$$

$$\text{Variance} = s^2 = \frac{\Sigma (X - \bar{X})^2}{n - 1}$$

$$\text{Standard Deviation} = s = \sqrt{s^2}$$

$$\text{Coefficient of Variation} = CV = \frac{s}{\bar{X}} \times 100\%$$

$$\text{Standard Error} = s_{\bar{X}} = \sqrt{\frac{s^2}{n}}$$

$$\text{Confidence Limits} = CL = \bar{X} \pm t s_{\bar{X}}$$

$$df = n - 1 \quad t_{.05} \text{ from table}$$

Distribution of Sample Means

Taking many samples from a population and calculating the mean for each sample gives a new distribution that has the same mean as the population distribution but is narrower. The width of the distribution depends on the sizes of the samples taken. The means of large samples are less likely to be in the tails than are the means of small samples, so the distribution becomes narrower as the samples get larger.

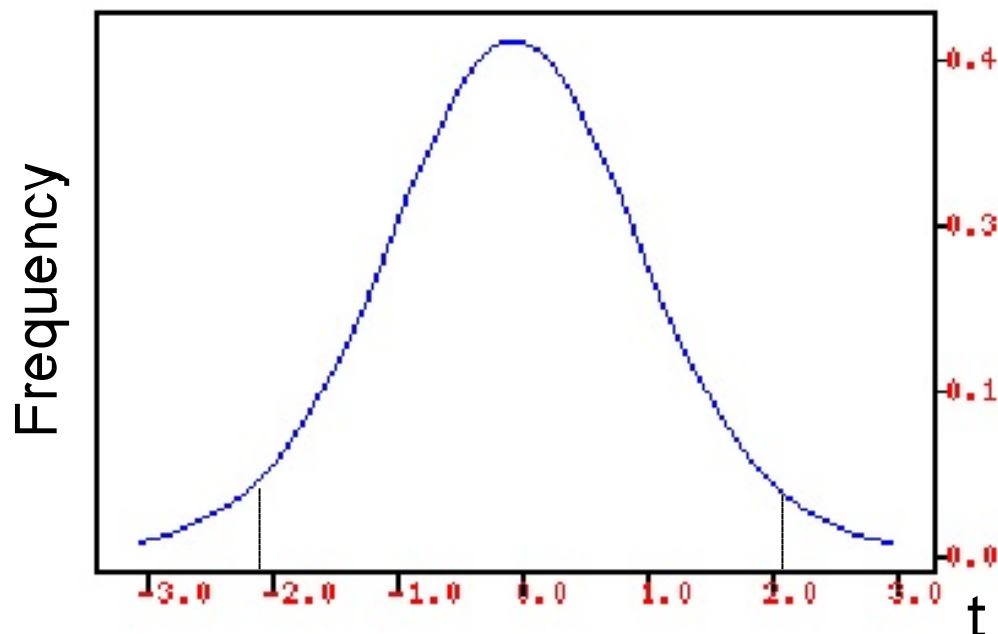
For the distribution of sample means

$$\text{Mean} = \bar{X}$$

$$\text{Standard Deviation} = \text{Standard Error} = \text{SEM} = s_{\bar{x}} = \sqrt{\frac{s^2}{n}}$$

“Student” calculated the distribution of the deviations of the sample means from the population mean relative to the standard error. This t distribution gives the probability of finding a difference between a sample mean and true mean greater than any chosen t. The distribution depends on the sample size or df.

Student t Distribution with 18 df



$$t = \frac{\bar{X} - \mu}{s_{\bar{x}}} \quad \text{where } s_{\bar{x}} = \sqrt{\frac{s^2}{n}}$$

Cutoff points are shown on the graph for the probability of finding a larger t is less than 5%. The points occur at $t_{.05, 18} = 2.101$

The probability of finding a t that is in either one of the two tails, ie a t larger than a cutoff point, is given in the t table.

Distribution of t

Degrees of Freedom	Probability of Obtaining a Value as Large or Larger					
	0.400	0.200	0.100	0.050	0.010	0.001
1	1.376	3.078	6.314	12.706	63.657	
2	1.061	1.886	2.920	4.303	9.925	31.598
3	0.978	1.638	2.353	3.182	5.841	12.941
4	0.941	1.533	2.132	2.776	4.604	8.610
5	0.920	1.476	2.015	2.571	4.032	6.859
6	0.906	1.440	1.943	2.447	3.707	5.959
7	0.895	1.415	1.895	2.365	3.499	5.405
8	0.889	1.397	1.860	2.306	3.355	5.041
9	0.883	1.383	1.833	2.262	3.250	4.781
10	0.978	1.372	1.812	2.228	3.169	4.587
11	0.876	1.363	1.796	2.201	3.106	4.437
12	0.873	1.356	1.782	2.179	3.055	4.318
13	0.870	1.350	1.771	2.160	3.012	4.221
14	0.868	1.345	1.761	2.145	2.977	4.140
15	0.866	1.341	1.753	2.131	2.947	4.073
16	0.865	1.337	1.746	2.120	2.921	4.015
17	0.863	1.333	1.740	2.110	2.898	3.965
18	0.862	1.330	1.734	2.101	2.878	3.922
19	0.861	1.328	1.729	2.093	2.861	3.883
20	0.860	1.325	1.725	2.086	2.845	3.850

t values and probabilities can be looked up on the web at <http://members.aol.com/johnp71/pdfs.html>

The probabilities in the t distribution provide the means of testing for differences between samples.

Confidence Limits: for a given probability, the limits of the range within which the true mean lies.

Determine the confidence limits for μ for calculating machine A, given:

$$\Sigma X = 220 \quad n = 10 \quad \Sigma X^2 = 214 \quad s^2 = \frac{\Sigma X^2}{n - 1} = \frac{214}{9} = 23.77$$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{220}{10} = 22$$

$$s_{\bar{X}} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{23.77}{10}} = 1.54$$

$$CL = 22 \pm ts_{\bar{X}}$$

For P = 0.95

$$t_{(.05,9)} = 2.262$$

$$\begin{aligned} CL &= 22 \pm (2.262)(1.54) \\ &= 22 \pm 3.5 \\ &= 18.5, 25.5 \end{aligned}$$

For P = 0.99

$$t_{(.01,9)} = 3.250$$

$$\begin{aligned} CL &= 22 \pm (3.250)(1.54) \\ &= 22 \pm 5.005 \\ &= 17.0, 27.0 \end{aligned}$$

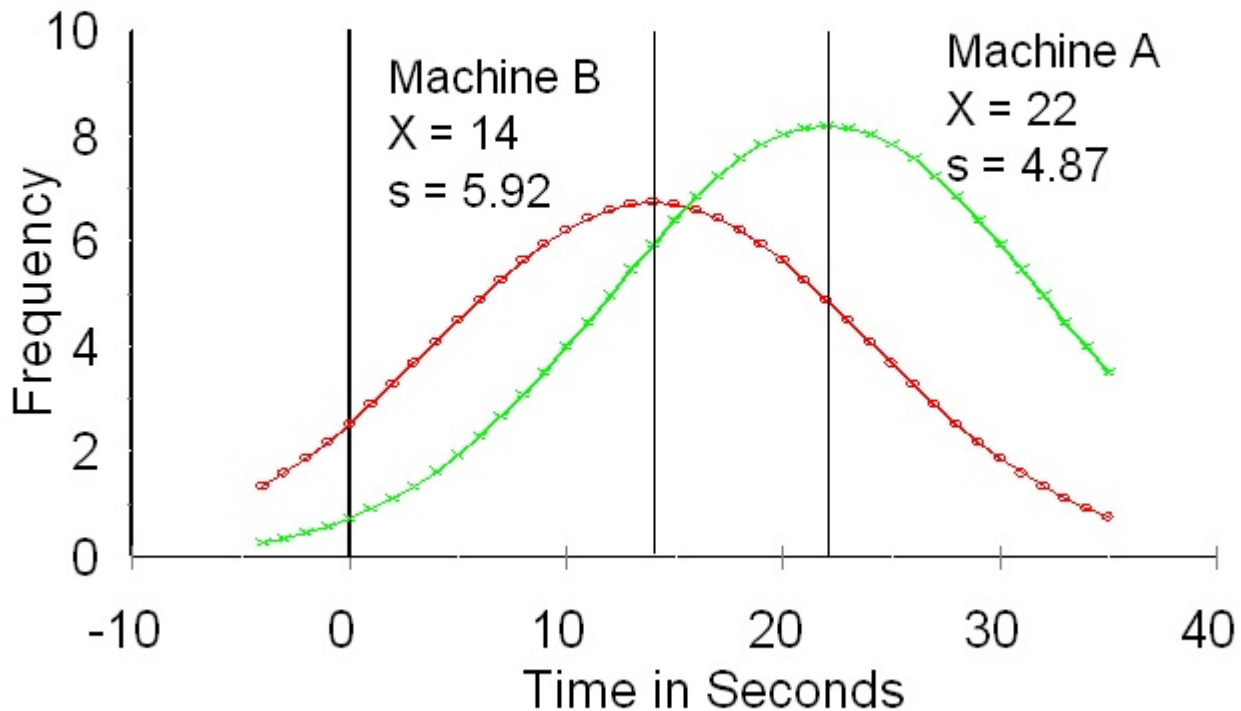


With a confidence of 95%, we can say the true mean, μ , is in the range 18.5 to 25.5. There is 1 chance in 20, or 5 in 100, that the true mean for machine A lies outside this range.

Null Hypothesis: a statement that there is no difference between the parameters involved

Under the null hypothesis, the mean of machine A is equal to the mean of machine B. If the null hypothesis is true, the true mean of machine B will lie in the range 18.5 to 25.5.

Normal Distributions for Time to Complete Calculations



If the true mean for machine B lies within the expected range, we accept the null hypothesis that there is no difference.

If the range in which the true mean for machine B lies does not overlap the range for machine A, the true mean of B lies outside the range and there is a significant difference between the two machines, i.e. we have 95% confidence that the means are different. The null hypothesis is rejected.

Note that the null hypothesis is essential for defining the probability and range in which the second mean is expected to lie if there is no difference. This makes it possible to test for differences with a known confidence level.

If the null hypothesis is rejected, the alternative hypothesis is accepted that the mean of A is not the same as the mean of B. Possible forms are:

- $A \neq B$
- $A > B$
- $A < B$

Rounding Data

Record to a number place that is 1/4 of the standard deviation per unit.

If $s = 6.96$ kg/experimental unit

$$6.96/4 = 1.74$$

Since the first number place in 1.74 is in the one's position, record data to the nearest kg.

If $s = 2.5$ kg/unit

$$2.5/4 = 0.625$$

Since the first number place in 0.625 is in the tenth's position, record data to the nearest tenth of a kg or 1 decimal place.

Rounding Means

Round means to a number place that is 1/4 of the standard error of the mean.

If $s = 6.96$ with 5 reps

$$s_{\bar{x}} = \frac{6.96}{\sqrt{5}} = 3.11$$

$$3.11/4 = 0.8$$

Round means to the nearest tenth or to 1 decimal place.

t Test

The t test compares the difference between two means to the standard error of the means.

As shown before, the t test to compare the difference between the sample mean and the population mean is:

$$t = \frac{\bar{X} - \mu}{s_{\bar{x}}} \quad \text{where} \quad s_{\bar{x}} = \sqrt{\frac{s^2}{n}}$$

The t test can be used to directly test the difference between two means. The t test to compare the difference between means from two different treatments is:

$$t = \frac{\bar{X}_a - \bar{X}_b}{s_d}$$

where $s_d = \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$ is the standard error of a difference

Degrees of Freedom (df) for t

1. If samples are from two populations, the df is the sum of the df for the two populations.
2. If pairs of values or replicated comparisons are being compared, the df is the number of pairs - 1.

For the two calculating machines

$$s_d = \sqrt{\left(\frac{s_A^2}{n} + \frac{s_B^2}{n}\right)} = \sqrt{\left(\frac{23.77}{10} + \frac{35.11}{10}\right)} = 2.426$$

$$t = \frac{\bar{X}_A - \bar{X}_B}{s_d} = \frac{22 - 14}{2.426} = 3.298$$

with $9 + 9 = 18$ df

Compare to the table value of $t_{.05, 18} = 2.101$. Since $3.298 > 2.101$, the two means are different with 95% confidence.

Comparing the calculated t statistic to the table value at 1%, what do you conclude?

Abbreviations in Statistics	Sample		Population
	Preferred symbol	Acceptable symbol	
Arithmetic mean	\bar{X}		μ
Chi-square	χ^2		
Correlation coefficient	r		
Coefficient of multiple determination	R^2		
Coefficient of simple determination	r^2		
Coefficient of variation	CV		
Degrees of freedom	df	DF	
Least significant difference	LSD		
Multiple correlation coefficient	R		
Not significant	NS		
Probability of type I error	α		
Probability of type II error	β		
Regression coefficient	b		β
Sample size	n		N
Standard error of mean	SE	$s_{\bar{x}}$	$\sigma_{\bar{x}}$
Standard deviation of sample	SD	s	σ
Student's t	t		
Variance	s^2		σ^2
Variance ratio	F		

The symbols *, **, and *** are used to show significance at the $P = 0.05$, 0.01 , and 0.001 levels, respectively. Significance at other levels is designated by a supplemental note.

From: Publications Handbook and Style Manual. 1998. Amer. Soc. Agron. Inc., Crop Sci. Soc. Amer. Inc., Soil Sci. Soc. Amer., Madison, WI.