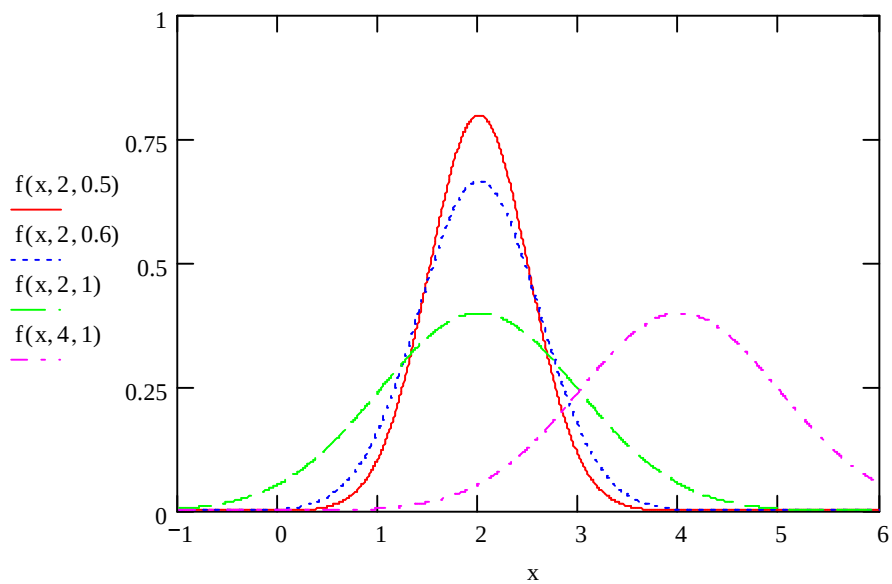
NORMAL (or GAUSSIAN) DISTRIBUTION

The Normal distribution function is a function that is applicable to a wide range of experimental data involving continuous random variables. The equation for the Normal distribution function is:

$$f(x) = \frac{1}{s\sqrt{2p}} e^{\frac{-(x-m)^2}{2s^2}}$$

with m as the population mean and s^2 as the population variance. The following graph shows some examples of the Normal distribution function. The graphs are plotted with MathCAD using the following user-defined function:

$$f(x,\mu,\sigma) := \frac{e^{\frac{-(x-\mu)^2}{2\cdot\sigma^2}}}{\sigma\sqrt{2\pi}}$$



Note the following:

f(x) has all the properties of a probability density function (what were they?)

The function is symmetric about the mean value, m

Changing the standard deviation, s , changes the 'spread' of the distribution

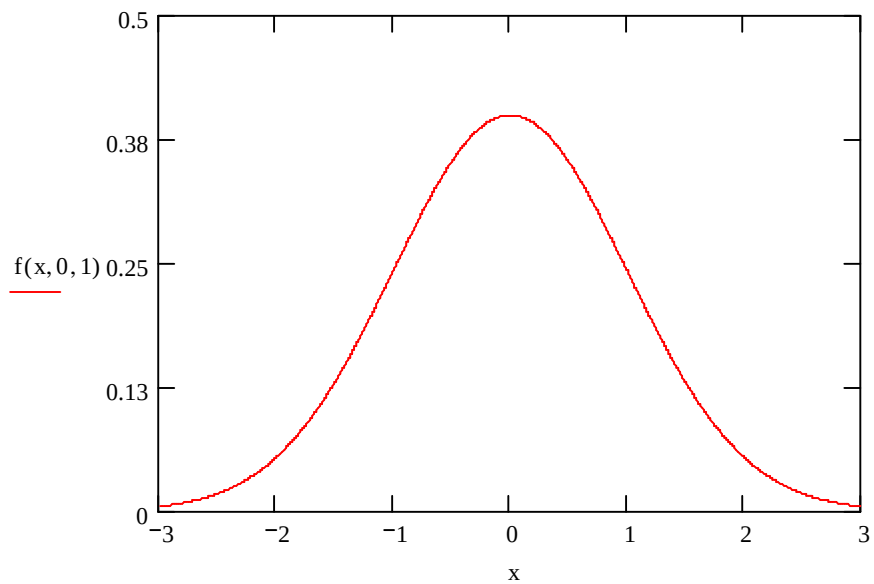
As for all distribution functions, we can calculate the probability of a random variable being in a required range as:

$$P\left(x_1 \le x \le x_2\right) = \int_{x_1}^{x_2} f\left(x\right) dx = \int_{x_1}^{x_2} \frac{1}{s\sqrt{2p}} e^{-\frac{(x-m)^2}{2s^2}} dx$$

Unfortunately, this isn't such a nice integral to evaluate! Since it can't be evaluated analytically, it must be evaluated numerically. We circumvent the problem by defining a new random variable, the *z* variable. This is calculated from the *x* variable as:

$$z = \frac{\left(x - m\right)}{s}$$

This new variable is also normally distributed, with a zero mean and a standard deviation of 1. This distribution is called the *standard normal density function.* This function is plotted below:



By now you should recognize that we calculate the probability of certain events happening by integrating the distribution functions. That is the same as looking up values in the cumulative distribution functions. The cumulative distribution function for the normal distribution is printed in most statistics textbooks, and is issued as a class handout. The following is the table:

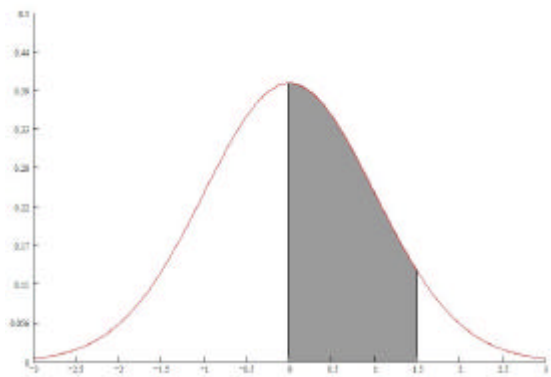
CAUTION – CAUTION – CAUTION – CAUTION – CAUTION – CAUTION

Most texts (and the handout for this course) show the cumulative distribution function, that is, the probability of variable z being <u>less than</u> the tabulated value.

Some text books give tables that show the probability of *z* being in the range 0 to z.

<u>GRAPHICALLY:</u>
The next figure represents the area tabulated in the course handout. It is the 'usual' cumulative normal distribution function and shows $P\left(-\infty < z < z_{TABULATED}\right)$

The following figure represents the area tabulated in one text - Wheeler & Ganji



Since the total area under the curve is 1, and the curve is symmetric, the two tabulated areas are related by:

(Course notes table) = (Wheeler & Ganji table) + 0.5

Cumulative Normal Density Function

$$z = \frac{(x-m)}{s} \approx \frac{(x-\bar{x})}{S}$$

z	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997
4.0	0.99997	0.99997	0.99997	0.99997	0.99997	0.99997	0.99998	0.99998	0.99998	0.99998

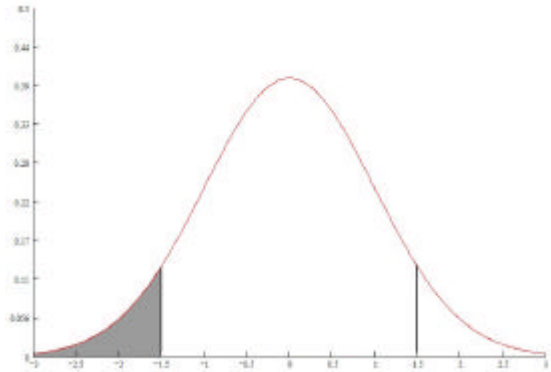
For positive values of *z* find P(x < z) by looking up the value in the table.

For negative *z* we have to use the symmetry of the function and the fact that $P(-\infty \le z \le +\infty) = 1$. We can calculate the probability for negative *z*-values (e.g., for z = -1.5): as:

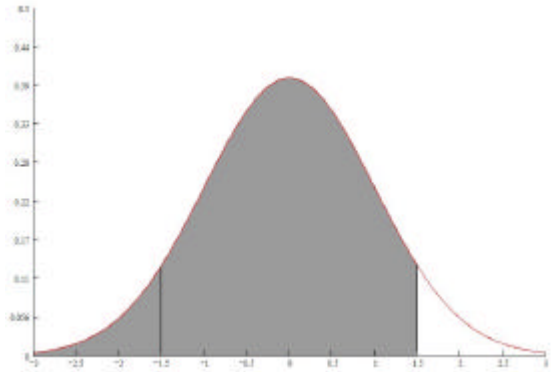
$$P(z \le -1.5) = 1 - P(z \le +1.5)$$

<u>GRAPHICALLY for NEGATIVE Z values:</u>

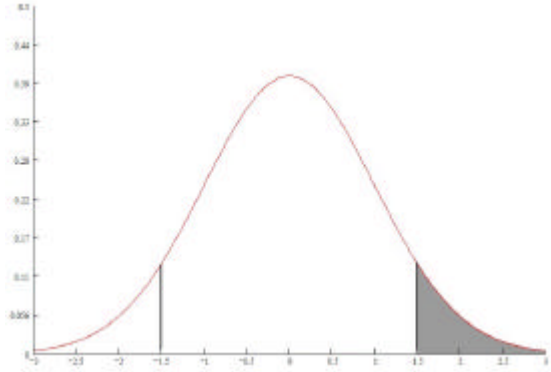
This is the area we want:



This is the tabulated area:



By symmetry, the wanted area is the same as this area:



So we calculate:

(wanted area) = 1- (tabulated area)

EXAMPLE: The average height of midshipman is 69 inches, with a standard deviation of 4 inches. Calculate:

a) The probability that a randomly selected midshipman will be less than 6-ft tall.

b) The probability that a randomly selected midshipman will be more than 6-ft tall.

c) The probability that a randomly selected midshipman will be between 5-ft 8-inches and 6-ft tall.

SOLUTION:
First calculate the z-variables.

$$z(x = 6') = \frac{(x - m)}{s} = \frac{(72 - 69)}{4} = 0.75$$
$$z(x = 5'8'') = \frac{(x - m)}{s} = \frac{(68 - 69)}{4} = -0.25$$

Now look up these z-values in the cumulative distribution function table.

z	P(<z) from table	P(<z) calculated
0.75	0.77337	
+0.25	0.59871	
-0.25		1 - 0.59871 = 0.40129

a) P(shorter than 6-ft) = 0.77337 = 77.337%

b) P(taller than 6-ft) = 1 – P(less than 6-ft) = 1 – 0.77337 = 0.22663 = 22.663%

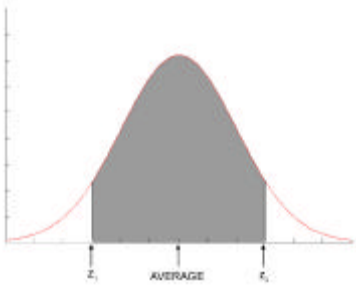
c)

$$\begin{aligned} P(5'8'' \le x \le 6') &= P(x \le 6') - P(x \le 5'8'') \\ &= 0.77337 - 0.40129 \\ &= 0.37208 = 37.208\% \end{aligned}$$

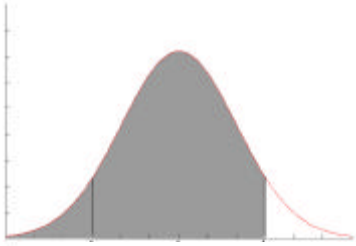
EXAMPLE: On average, automotive tires are $\bar{d} = 19.0$-inches in diameter. The manufacturer can control the quality of manufacture, which affects the standard deviation of the diameter. To reduce wastage he wants at least 99% of tires to be in the range $18.9 \le d \le 19.1$ inches. What is the maximum permissible standard deviation?

SOLUTION: We assume normally distributed tires and note that the required range (18.9 to 19.1 inches) is symmetric about the mean (19.0-inches). Here is a graphical representation of the problem:

Wanted probability = this area = 0.9900



This is the tabulated area. But we can't look it up yet because we don't know either the area or z value.

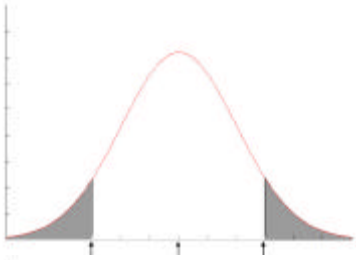


By symmetry and pdf properties,

(wanted) = 1 – (this area)

so

(this combined area) = 0.0100



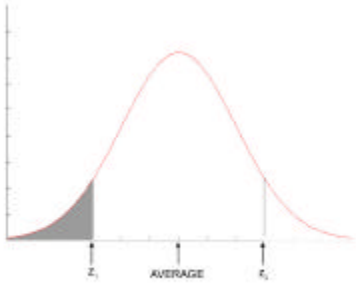
Each tail has half of the above area = 0.0050

and

(wanted) = (tabulated) – (one tail)

hence

(tabulated) = 0.9900 + 0.0050 = 0.9950



So, to solve the problem we look up the z-value for a tabulated value of 0.9950 to get:

For area = 0.99492, z = 2.57
For area = 0.99506, z = 2.58

Hence (linear interpolation), for area = 0.9950 we have z = 2.576

At this stage, we know the mean (19.0-inches), the required z- value and the actual value that correlates to that z-value (19.1-inches). We solve for the standard deviation:

$$z = \frac{(x - m)}{s}$$

$$2.576 = \frac{(19.1 - 19.0)}{s}$$

$$s = 0.0388 - inches$$

So the final answer is that for a rejection of less than 1%, the required standard deviation of the tire diameters must be less than 0.0388 inches.

PARAMETER ESTIMATION

The aim of a statistical experiment is often to estimate the parameters of a population. Once these are known, the distribution function can then be used to make predictions about the population.

Different distribution functions need different parameters. In this section we will develop a method of estimating the mean and standard deviation of a population based on a sample. We will investigate such topics as the determination of sample size, and the rejection of bad data points.

Our best estimate of the population parameter *population mean* is the sample mean. However, we recognize that our sample mean is not the exact population mean. We introduce the concepts of *degree of confidence*, and *level of significance.*

The degree of confidence (also called the confidence level) is the probability that the population mean falls within the specified interval, δ.

$$\text{degree of confidence} = P\left(\bar{x} - d \le m \le \bar{x} + d\right)$$

Thus our estimate of the population mean now takes the form:

$$m = \bar{x} \pm d$$

Rather than giving a degree of confidence, we may also quote a *level of significance,* a. This is the probability of a significant event. The significant event would be that the population mean falls <u>outside</u> the confidence interval. The degree of confidence and level of significance are related by:

$$(\text{degree of confidence}) + (\text{level of significance}) = 100\%$$

$$(\text{degree of confidence}) + \alpha = 100\%$$

When we take a sample, this is just one sample from an (infinitely) large set of possible samples. The *central limit theorem* indicates that the average of all the possible samples is, itself, normally distributed. Thus is, if you repeated the sampling many times, the averages you calculate would, themselves, follow a normal distribution. The mean for this distribution is the population mean – the thing we are trying to find. The standard deviation of this new distribution is related to the population standard distribution, but depends on the number of samples.

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

However, we can only use the normal distribution under the following conditions:

If the original population is normally distributed

If the original population is not normally distributed, but the sample size is large (we normally say the sample size must be more than 30).

STUDENT-t DISTRIBUTION

What do we do if the above criteria are not met? The answer lies in another distribution, the Student-t distribution. The Student-t distribution has many uses, but for this course we only use it to determine confidence intervals. We therefore do not go into full details about this distribution, but focus on our specific needs.

The Student-t statistic is:

$$t = \frac{\left(\bar{x} - m\right)}{S / \sqrt{n}}$$

The Student-t distribution depends on the *number of degrees of freedom* (dof) of the sample. The sample dof is the number of independent measurements minus the minimum number of measurements theoretically necessary to estimate a statistical parameter. For estimates of the mean (our main focus here), the minimum is one measurement. Therefore, the degrees of freedom for an n-sized sample is (n-1).

We use tabulated values of Student's-t. The value depends on the level of significance, α, and the number of degrees of freedom.

The table is issued in class.

Student-t Distribution

	Level of Significance, α						
	10%	5%	2.5%	1%	0.5%	0.05%	1-Sided
dof	20%	10%	5%	2%	1%	0.1%	2-Sided
1	3.07768	6.31375	12.70620	31.82052	63.65674	636.61925	
2	1.88562	2.91999	4.30265	6.96456	9.92484	31.59905	
3	1.63774	2.35336	3.18245	4.54070	5.84091	12.92398	
4	1.53321	2.13185	2.77645	3.74695	4.60409	8.61030	
5	1.47588	2.01505	2.57058	3.36493	4.03214	6.86883	
6	1.43976	1.94318	2.44691	3.14267	3.70743	5.95882	
7	1.41492	1.89458	2.36462	2.99795	3.49948	5.40788	
8	1.39682	1.85955	2.30600	2.89646	3.35539	5.04131	
9	1.38303	1.83311	2.26216	2.82144	3.24984	4.78091	
10	1.37218	1.81246	2.22814	2.76377	3.16927	4.58689	
11	1.36343	1.79588	2.20099	2.71808	3.10581	4.43698	
12	1.35622	1.78229	2.17881	2.68100	3.05454	4.31779	
13	1.35017	1.77093	2.16037	2.65031	3.01228	4.22083	
14	1.34503	1.76131	2.14479	2.62449	2.97684	4.14045	
15	1.34061	1.75305	2.13145	2.60248	2.94671	4.07277	
16	1.33676	1.74588	2.11991	2.58349	2.92078	4.01500	
17	1.33338	1.73961	2.10982	2.56693	2.89823	3.96513	
18	1.33039	1.73406	2.10092	2.55238	2.87844	3.92165	
19	1.32773	1.72913	2.09302	2.53948	2.86093	3.88341	
20	1.32534	1.72472	2.08596	2.52798	2.84534	3.84952	
25	1.31635	1.70814	2.05954	2.48511	2.78744	3.72514	
30	1.31042	1.69726	2.04227	2.45726	2.75000	3.64596	
40	1.30308	1.68385	2.02108	2.42326	2.70446	3.55097	
50	1.29871	1.67591	2.00856	2.40327	2.67779	3.49601	
60	1.29582	1.67065	2.00030	2.39012	2.66028	3.46020	
70	1.29376	1.66691	1.99444	2.38081	2.64790	3.43501	
80	1.29222	1.66412	1.99006	2.37387	2.63869	3.41634	
90	1.29103	1.66196	1.98667	2.36850	2.63157	3.40194	
100	1.29007	1.66023	1.98397	2.36422	2.62589	3.39049	
infinity	1.28157	1.64488	1.96001	2.32642	2.57593	3.29072	

Example: For a 5% level of significance (95% degree of confidence), and 15 measurements, the dof for estimating the mean is (15-1)=14, and Student's-t is 2.14479. Note that for the confidence interval we need to use the 2-sided values of Student's-t.

Also note that when the number of samples (and thus the dof) is very large, Student's-t is exactly the same as the normal distribution z-value.

We now have enough information to estimate the population mean. First a summary:

n	Number of samples
$\bar{x}$	The sample average
S	The sample standard deviation
α	Level of significance. Chosen by you (or the problem statement).
t	Student's-t, looked up in the table for the appropriate α and dof

Combining equations from above:

$$m = \bar{x} \pm d$$

$$t = \frac{(\bar{x} - m)}{S / \sqrt{n}}$$

$$\text{hence } d = \frac{tS}{\sqrt{n}}$$

therefore the confidence interval of the mean is:

$$m = \bar{x} \pm \frac{tS}{\sqrt{n}}$$

Alternatively, the confidence interval of the mean can be quoted as

$$\bar{x} - \frac{tS}{\sqrt{n}} \le m \le \bar{x} + \frac{tS}{\sqrt{n}}$$

THESE ARE GOOD EQUATIONS
TO REMEMBER!

These equations state that, at the chosen degree of confidence (or level of significance), the population mean is in the range given by the sample mean plus/minus an interval.

Remember:

> We are trying to find the population mean, but we can only calculate sample means.
>
> The population mean is a *parameter* of the population. It does not vary.
>
> > Sample means are *statistics.* There is a different sample mean for each chosen sample set
> >
> > > A statistic is an estimate of a parameter

EXAMPLE:
Thirteen measurements of heart rate produced an average value of 66.727 beats per minute, with a standard deviation of 5.368 beats per minute. What is the population (long-term) average heart rate? Compare the results for a 90% and 95% degree of confidence.

SOLUTION:
<u>First, for 90% degree of confidence.</u>
We look up the Student-t values for α = 100 - 90 = 10% with dof=(13-1) = 12 to get:

$$t_{10,12} = 1.78229$$

The population mean is therefore

$$m = \bar{x} \pm \frac{tS}{\sqrt{n}} = 66.727 \pm \frac{1.78229 \times 5.368}{\sqrt{13}} = 66.727 \pm 2.65350 \text{ bpm}$$

Or, with appropriate rounding:

$$m = 66.7 \pm 2.7 \text{ beats per minute}$$

<u>Now we analyze the 95% degree of confidence.</u>
We look up the Student-t values for α = 100 - 95 = 5% with dof=(13-1) = 12 to get:

$$t_{5,12} = 2.17881$$

The population mean is therefore

$$m = \bar{x} \pm \frac{tS}{\sqrt{n}} = 66.727 \pm \frac{2.17881 \times 5.368}{\sqrt{13}} = 66.727 \pm 3.24385 \text{ bpm}$$

Or, with appropriate rounding:

$$m = 66.7 \pm 3.2 \text{ beats per minute}$$

Note that if we want to be more certain (confident) of the result, we have to include a wider interval.

90% confidence gave a 2.7 b.p.m. confidence interval

95% confidence gave a 3.2 b.p.m. confidence interval

EXAMPLE:
What if we had taken 41 measurements, not 13, to get the same sample mean and standard deviations as above? At the 95% degree of confidence, what is the population mean?

SOLUTION:
Look up the Student-t for 5% significance and dof=(41-1)=40 to get t = 2.02108.

$$m = \bar{x} \pm \frac{tS}{\sqrt{n}} = 66.727 \pm \frac{2.02108 \times 5.368}{\sqrt{41}} = 66.727 \pm 1.69435 \text{ bpm}$$

or

$$m = 66.7 \pm 1.7 \text{ beats per minute}$$

By taking more measurements we have increased the accuracy from 3.2 to 1.7 beats per minute.

PROBLEMS

Norm-1:
A high accuracy voltmeter was used to determine that a power source had a voltage of 115V. Readings from another, cheaper voltmeter recorded 60% of its readings more than 0.5V from the mean value. Assume a normal distribution for the error.

a) Calculate the standard deviation of readings for the cheap voltmeter

b) What is the probability that a single reading will be more than 0.75V from the mean value?

Norm-2:
In an automotive manufacturing process, lengths of steel beam are cut to an average length of 12.75 feet with a standard deviation of 0.2 inches. Variations of ±0.3 inches from the average length can be tolerated.

a) Calculate the percentage of beams that will be rejected

b) In order to reduce waste the cutting process will be improved. What is the required standard deviation of cuts if the rejection rate is to be halved?

Norm-3:
The light bulbs used in Dahlgren Hall are known to have a normally distributed life with a mean of 3600 hr and a standard deviation of 160 hr. After how many hours should all light bulbs be simultaneously replaced to ensure that no more than 10% of the bulbs fail?

Norm-4:
Testing of military jeep tires determined the following:

Number of tires tested	10
Average measured lifetime	41,500 miles
with a standard deviation of	5500 miles

a) After what mileage would 10% of tires be expected to have worn out?

b) What percentage of tires will wear out between 50,000 and 60,000 miles?

c) At the 95% level of confidence, what is the mean expected life of a tire?

d) Your platoon is on deployment with 10 jeeps (4 tires per jeep). You carry 5 replacement tires with you. Each jeep averages 200 miles per day. How many days into the deployment should you anticipate needing a supply of more tires?

Hint: Use the 'tire count' to determine a z-value.
Use this value to calculate mileage and then days.