

7 Sampling and Statistical Inference

7.1 Populations and Samples

In several sections of your introduction so far, you have read about *parametric* statistics and *sample* statistics. In this section we examine these concepts – populations and samples – in more depth.

When the data in which we are interested represent everything on which we are focusing, we call the data set a *population*. For example, we could discuss how the students in the QM213 Business & Economic Statistics I course in the School of Business and Management in the Fall semester of 2029 at Norwich University do in their first quiz and in their final exam. These data could constitute the entire population. If we consider the data the population, then the mean, standard deviation, median, mode and any other statistic we compute using these data are all *parametric* values. They are not *estimates* of anything; they are exact representations of the population attributes for that particular, specific group. The group doesn't represent anything; it isn't intended to be anything other than itself.

Similarly, if we summarize the sports statistics for a particular team for a specific period, the data are not a sample of anything; they are the entire population of data for that period. “The Norwich Paint-Drying Team scored an average of 32 points per round in the North American Paint-Drying Tourney in 2029” isn't a description of a sample: it's (presumably) the absolute, incontrovertible truth; it's a parametric statistic.

But what if we consider the QM213 data as part of a larger study? What if we are actually interested in studying the relationship between the score on the first quiz in QM213 and the score on the final exam in QM213 classes in general? The data we just collected could actually be going into a collection spanning the years from, say 2010 through 2029; in that case, the group of interest is not only the particular class and the particular quiz results: the group of interest is *all possible groups of QM213 students* and their first quiz and final exam results. In this case, the data for the Fall 2029 QM213 class's first quiz and final exam results are both *samples* from the larger, theoretical *population* of all possible such values.

As stated above, does the population include students in QM213 courses before 2010 and after 2029? Does the population for which the Fall 2029 results have been collected include Spring QM213 classes? Does the population include other statistics classes in the School of Business and Management such as QM 370 *Quantitative Methods in Marketing & Finance*? Does the population include results from the first quiz and final exams for other statistics courses at Norwich University such as MA232 *Elementary Statistics*? Does it include results for statistics courses at other universities? For that matter, is the population we are studying all possible courses that have a first quiz and a final exam?

The critical concept here is that there is nothing absolute about a set of numbers that tells us instantly whether they are a sample or a population; there's no convenient little flag sticking up to indicate that status. More seriously, the decision on whether to view a group of data as a sample or a population is not based on the data: the decision is based on *the way the data are collected* and *how they are being* used by the analysts.

7.2 Sample Statistics and Parameters

One of the most important concepts in statistics is the idea of representative samples. A sample is representative when the information from the sample can be used to guess at the values of the population from which it was drawn. We say that we can infer the parametric value of a statistic from the value of the sample statistic.

A researcher could claim that the Fall 2010 Norwich University QM213 quiz and final scores were samples from the global population of all statistics courses given anywhere at any time. There would be a number of assumptions in such a claim. Think about some of the claims the researcher could be making by asserting that the sample in question was from the population of all students taking any statistics course (this is only a partial list):

- The QM213 class in Fall 2029 is similar to all other QM213 classes;
- The QM213 course is similar to all other statistics courses the School of Business and Management;
- The statistics courses in the School of Business and Management' to all other statistics courses at Norwich University;
- Statistics courses at Norwich University are similar to all other statistics courses on planet Earth;
- Statistics courses on planet Earth are similar to all other statistics courses in the known universe.

None of these assumptions is obligatory; what the researcher decides to claim about the nature of the population from which the Fall 2029 QM213 first quiz and final exam results determines what assumptions are being made.

Depending on what the population is assumed to be, the researcher will be able to try to infer attributes of that population based on the particular sample; whether those inferences will be accepted by other statisticians is a question of how carefully the researcher thinks about the sampling process.

In ordinary life, we are faced with data that are represented to be from populations defined according to the preferences of the people reporting the statistics. For example, a newspaper article may report that 23% of the college students in Mare Imbrium have been unable to buy a hovercraft in the first year after their graduation. The writer goes on to discuss the general problems of college graduates system wide, including those on Earth, on the Lunar Colonies, and on the Jovian Satellite Colonies. But how do we know that the Mare Imbrium students are in fact from the population defined for the entire Solar System? Is there any evidence presented to suggest that the Mare Imbrium students are in fact a representative sample? What about the fact that the proportion of college graduates who buy a hovercraft on Mars has reached 91%? Or that Earth graduates have a paltry 3% ownership of these much-desired vehicles in their first year after graduation?

Whenever you read statistics, especially in the popular press or in media prepared by people with a strong interest in convincing you of a particular point of view, you should investigate in depth just how the data were collected and on what basis they can be considered representative of the population for which they are claimed to be samples.

INSTANT TEST P 2

Examine published reports that include statistical information. Notice carefully where the authors are discussing *populations* and where they are discussing *samples*. Think about what you would have to do to extend or narrow the definitions to change the populations to larger or smaller groups. Explain your reasoning as if to a fellow student.

7.3 Greek Letters for Parametric Statistics

You will also have noticed in previous sections that parametric statistics are customarily symbolized using lowercase Greek letters. For reference, Figure 7-1 shows the Greek alphabet with names and Roman equivalents. Notice that sigma, the equivalent of our s , has two lowercase versions, σ and ς . The latter is rarely used in mathematics; it is the form that is used in Greek writing only for a sigma that is at the end of a word.

Figure 7-1. Greek letters for parametric statistics.

A α Alpha/a	B β Beta/b	Γ γ Gamma/g	Δ δ Delta/d
Ε ε Epsilon/ě	Ζ ζ Zeta/z	Η η Eta/ē	Θ θ Theta/th
Ι ι Iota/i	Κ κ Kappa/k	Λ λ Lambda/l	Μ μ Mu/m
Ν ν Nu/n	Ξ ξ Xi/x	Ο ο omicron/ō	Π π pi/p
Ρ ρ Rho/r	Σ σ ς Sigma/s	Τ τ Tau/t	Υ υ Upsilon/u
Φ φ Phi/ph	Χ χ chi/ch	Ψ ψ psi/ps	Ω ω omega/ō

7.4 Random Sampling from a Population

What makes a sample representative of a particular population?

Should we inspect the data and pick the ones we think look like what we believe the population to be? Bad idea, don't you think? How would we ever separate the effects of our own preconceptions from the reality of the situation? With a pick-and-choose approach to sampling, we could claim anything we wanted to and pretend to provide a statistical justification for our claims.

For example, suppose a (shall we say) naïve researcher, Arthur Schlemiel,⁷⁴ has a preconceived notion that the score on the first quiz in the QM213 class for Fall 2029 was strongly related to the score on the final exam in that class. Figure 7-2 shows the original data.

Schlemiel could cheerfully (and wrongly) select only the students whose Quiz #1 scores and Final Exam scores were similar; for example, either both low, both middling, or both high. Schlemiel might compute the ratio of the Final Exam score to the Quiz #1 score (F/Q) and pick only the students with a F/Q ratio of, say, 85% to 115%. The students whose data are in bold italics and are shown in the central box in Figure 7-3.

Figure 7-2. Original data on quiz & final scores.

Student	Quiz #1	Final Exam
A	96%	100%
B	61%	68%
C	80%	98%
D	57%	99%
E	90%	82%
F	55%	92%
G	57%	90%
H	98%	82%
I	74%	84%
J	98%	97%
K	66%	75%
L	75%	90%
M	51%	66%
N	74%	60%
O	91%	100%
P	93%	64%
Q	98%	100%

Figure 7-3. Biased sampling using F/Q ratio.

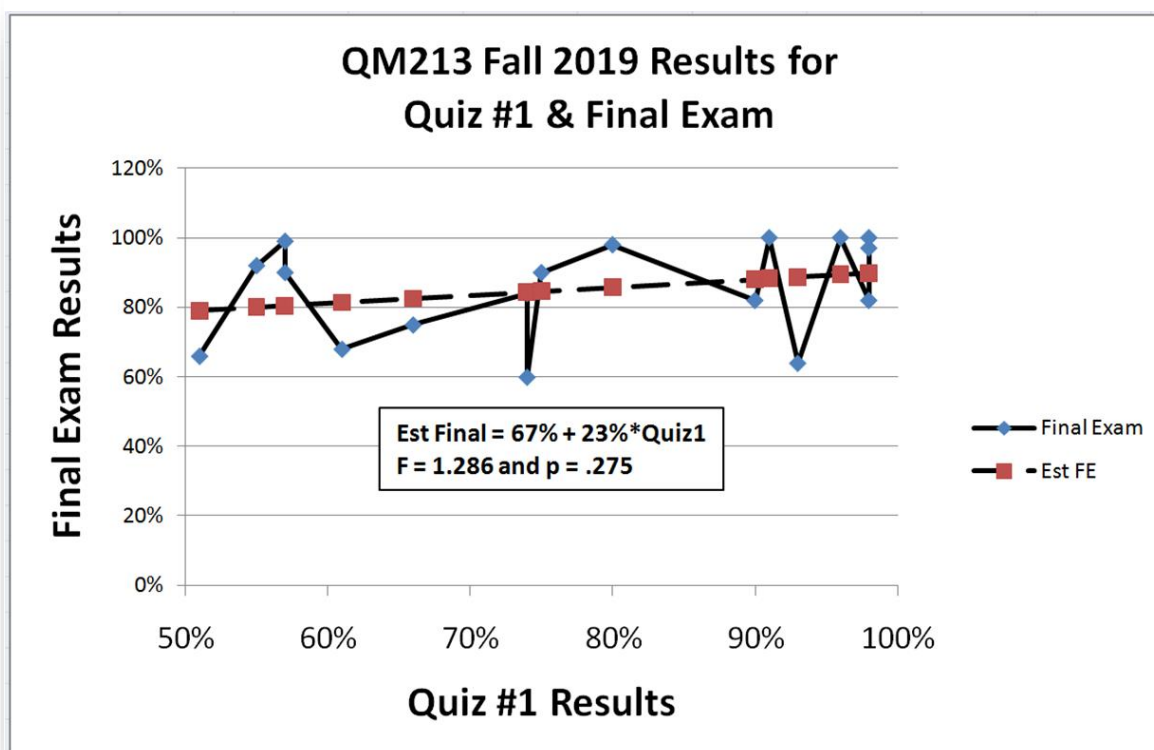
Student	Quiz #1	Final Exam	F/Q
P	93%	64%	69%
N	74%	60%	81%
H	98%	82%	84%
<i>E</i>	<i>90%</i>	<i>82%</i>	<i>91%</i>
<i>J</i>	<i>98%</i>	<i>97%</i>	<i>99%</i>
<i>Q</i>	<i>98%</i>	<i>100%</i>	<i>102%</i>
<i>A</i>	<i>96%</i>	<i>100%</i>	<i>104%</i>
<i>O</i>	<i>91%</i>	<i>100%</i>	<i>110%</i>
<i>B</i>	<i>61%</i>	<i>68%</i>	<i>111%</i>
<i>I</i>	<i>74%</i>	<i>84%</i>	<i>114%</i>
<i>K</i>	<i>66%</i>	<i>75%</i>	<i>114%</i>
L	75%	90%	120%
C	80%	98%	123%
M	51%	66%	129%
G	57%	90%	158%
F	55%	92%	167%
D	57%	99%	174%

Schlemiel would leave out students P, N, H, L, C, M, G, F, and D because their results don't match his preconceptions. We call this a *biased* sample. Simply at a gut level, would you trust anything Schlemiel then has to say about the relationship between Quiz #1 results and Final Exam results in QM213 – or in any other population he chose to define for these data?

⁷⁴ A *schlemiel* is a Yiddish term for a dolt.

Let's pursue this example just a bit farther. You will learn later that an *analysis of variance (ANOVA) with regression* could be used to try to predict the Final Exam score based on the collected data about the Quiz #1 results. Without going into detail, the calculations of a regression equation and the ANOVA for the original data are shown in brief in the box in the chart in Figure 7-4. The results don't support the view that there is much of a relation between the first quiz result and the final exam result. The regression equation, such as it is, has a slope (b) of about 23%, implying that the Final Exam score rises by 23% of the Quiz 1 score. But another way of looking at this weak relationship is to examine the coefficient of determination, r^2 (not shown in the figure) which reflects how much of the variability in one variable can be explained by knowing the other. In this case, r^2 turns out to be about 8%; i.e., only 8% of the variability of the Final Exam score can be explained by the Quiz 1 score; all the rest of the scatter has other, unknown sources.

Figure 7-4. ANOVA with regression for unbiased sample.



INSTANT TEST P 7-5

For the diagram above, explain the meaning of the dashed line with the red squares versus the meaning of the blue solid line with the blue diamonds.

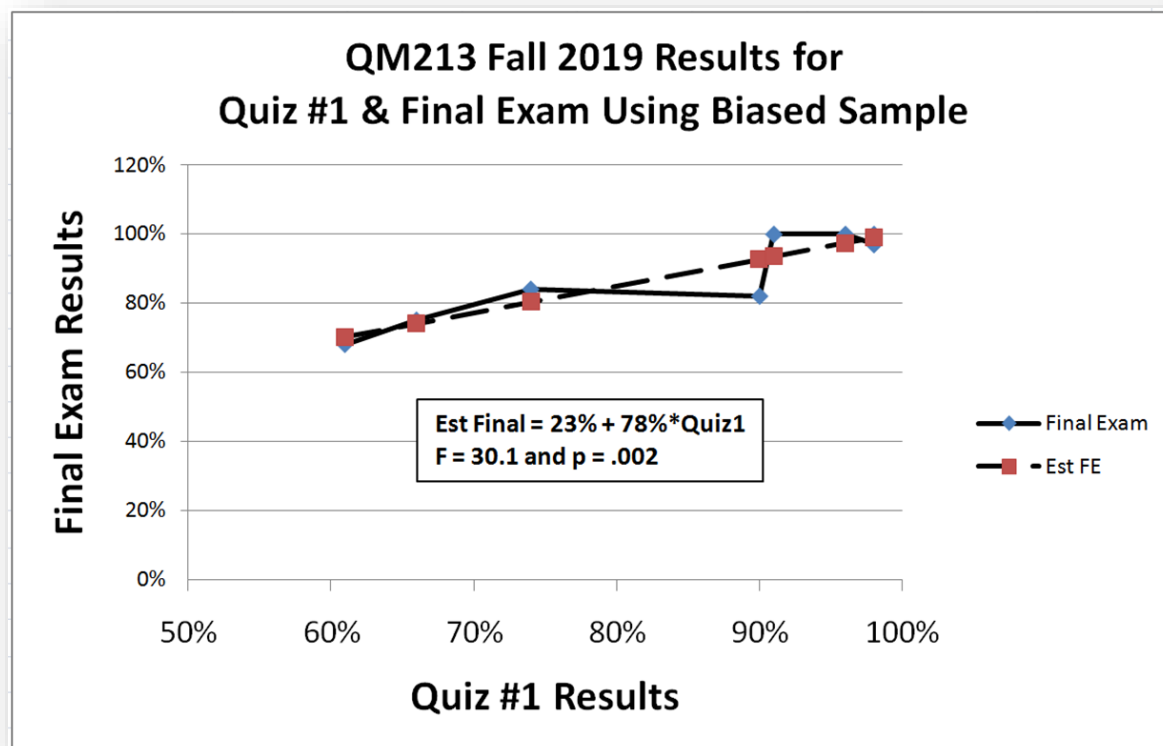
Explain why it's OK to use a line graph instead of histograms in the diagram.

Most people would *not* interchange the axes in this graph; they would not put "Final Exam Results" on the abscissa and "Quiz #1 Results" on the ordinate. Why not?

In contrast, the faulty (fraudulent) analysis using Schlemiel's biased sample, shown in Figure 7-5, produces a misleading result exactly in line with Schlemiel's preconceptions: what a surprise! This time the (fake) regression equation has a slope of 78%; the bogus coefficient of determination is a much higher 83%.

So unless someone examines his raw data and catches his deception, Schlemiel can publish his rubbish and

Figure 7-5. Schlemiel's analysis based on biased sample.



distort the literature about the relationship between initial quizzes and final scores. The practical effects, were these lies to become popularly known, might be to discourage students who do poorly in their first quiz in QM213 (the horror!). In fact, the unbiased data do not support such a pessimistic view.

As you can see, defining representative, unbiased samples is critical to honest use of data. The subject of *publication bias*, which is an observed tendency for editors of scientific journals to discourage publication of negative or neutral results, has serious consequences for the ability of researchers to aggregate results from independent studies using the techniques of *meta-analysis* which you will study later in the course.

7.5 Selecting Random Values for an Unbiased Sample

The solution to getting a *representative, unbiased sample* is *random sampling*. In the simplest terms, a random sampling gives every member of a *defined population* an equal chance of being included in the sample being created. In intuitive terms, there are no special rules for selecting the members of the sample – every member of the population might be picked for study.

Looking at the converse, we can say that the population corresponding to a sample is every element which could have been picked according to the definition of the desired population.

Figuring out what the population is for a given sample is not necessarily easy. As a simple example, suppose we want to study the structural resistance to earthquakes of reinforced concrete used in all buildings built in the last decade that use concrete around the world. We take random samples and everything's OK, right? Well no, not necessarily. For one thing, unless we think of it, the sample won't include concrete from buildings that have collapsed in previous earthquakes! So although the population seems at first to be "all buildings on the planet built in the last decade that use reinforced concrete" it's more correctly described as "all buildings on the planet built in the last decade that use reinforced concrete but have not collapsed yet." Don't you think that the sample might be considered biased? After all, the measurements may exclude the buildings that were built in the last decade using shoddy materials such as concrete with much too much sand and too little cement or "reinforced" concrete without metal reinforcement rods.⁷⁵ The information would give a biased view of how strong the concrete actually has been in the last decade.

The easiest way of understanding random sampling is to see it done. Figure 7-6 shows the beginning and end of a long list of 20,000 observations gathered about a total of 1,000 stock brokers, their consumption of alcohol in the hour before a trade (Y/N) and the occurrence of one or more errors in the particular trade (Y/N). How would one select a random sample of 1,000 observations from this list?

One approach (not the only one) is to assign a random number to each observation; in EXCEL, that's easy: the function `=RAND()` generates a number between 0 and 1 (inclusive) in a uniform distribution. Note that this function takes no argument – the parentheses have nothing between them.

Another EXCEL function is the `=RANDBETWEEN(bottom, top)` function which generates a uniform distribution of numbers between the limits (inclusive).

The key to these applications for *randomizing data* is that the generated numbers are in uniform distributions, so any number can appear anywhere in the list with equal probability.

Figure 7-6. Start and end of list of 20,000 observations about 1,000 data brokers, their alcohol consumption, and their errors.

		hour before trade	trade
1	123	N	N
2	190	N	N
3	437	N	N
4	614	N	Y
5	197	N	N
6	657	Y	Y
7	286	N	N
8	739	Y	N
9	980	N	N
10	206	Y	N
19,995	881	N	N
19,996	847	Y	Y
19,997	404	N	N
19,998	896	Y	N
19,999	502	N	N
20,000	457	N	N

⁷⁵ (Associated Press 2011)

We then sort the entire list by the random numbers and pick the first elements in sorted list as our random sample of the desired size. Figure 7-7 shows the results of this process. Note that the middle extract shows the data around the desired limit of 1,000 entries. We have but to select the first 1,000 entries in the sorted list to have a random sample from the entire 20,000 of the original data.

Figure 7-7. Original data with random numbers assigned and used to sort the data.

Sequence #	Random #	Obs #	Broker	Alcohol consumed 1 hour before trade	Errors in trade
1	0.00181	10,600	140	N	Y
2	0.00461	14,956	245	N	N
3	0.02459	13,649	608	N	N
4	0.03492	19,998	896	Y	N
5	0.04155	8,326	994	N	Y
998	0.1030	6,943	491	N	Y
999	0.1032	18,171	151	N	Y
1,000	0.1053	2,160	986	Y	N
1,001	0.1053	7,217	467	N	Y
1,002	0.1075	19,235	959	N	N
1,003	0.1084	16,737	471	Y	Y
19,995	0.99136	1,799	639	Y	Y
19,996	0.99325	1,505	36	N	N
19,997	0.99401	1,599	467	Y	N
19,998	0.99429	2,348	856	N	N
19,999	0.99446	8,069	17	N	Y
20,000	0.99903	6,317	812	N	N

Looking at this procedure step by step,

- We start by assigning a random number to each of the observations using the =RAND() function in EXCEL.
- Once we've created the list of 20,000 random numbers, we fix them (stop them from changing) by copying the entire list and pasting it as values into the first cell of the list, freezing the numbers as values instead of functions..
- Finally, we sort the list using the random numbers, as shown in Figure 7-7.
- We select the first 1,000 rows and that's it: a random sample of 1,000 records from the original 20,000.
- There is no human judgement (and potential bias) is involved in the choice. This method is easy to apply to any data set and always works.

INSTANT TEST P 7-8

Create a list of 20 random values from 5,000 to 10,000 using =INT(NORM.INV(RAND(),5000,10000)). Copy the list and Paste Special into another column using the Values option to freeze the data. Now apply random numbers using =RAND() and then copy/paste-special next to your 100 values. Sort the two columns by the RAND() values and practice selecting random samples of size 20 from the list.

7.6 More about Probability and Randomness

You've heard about and perhaps even studied probability in other courses; in this course, we've already introduced some basic ideas about probability in §5.3. For now, it suffices to establish probability as a measure of what we expect in the long run: what happens on average when we look at repeated actions in a defined situation.

In §7.4 on random sampling, we used the `=RAND()` function of EXCEL. EXCEL's **HELP** function describes **RAND()** as follows: "Returns an evenly distributed random real number greater than or equal to 0 and less than 1." The phrase *evenly distributed* refers to what mathematicians and statisticians refer to as the uniform probability distribution (§5.4). For the **RAND()** function, we can assert that the frequency of occurrence of numbers 0, 0.1, 0.2... 0.8, and 0.9 are all equal (to 10% of the observations) on average over the long run. But we can also assert that the occurrence of numbers 0, 0.01, 0.02, 0.03... 0.96, 0.97, 0.98 and 0.99 are also equal (to 1% of the observations) on average over the long run. The generated numbers are random precisely because there is equal frequency of all the numbers regardless of precision (within the limits of calculation of EXCEL).

Because the numbers are generated by mathematical processes that actually generate exactly the same sequence of numbers if they start from any given starting point, we call the `=RAND()` function a *pseudo-random number generator*. In other words, the output looks random even though the sequence is theoretically repeatable.

But wouldn't a generator that produced the sequence 0.1, 0.2, 0.3, 0.4... and so on in perpetuity, in the same order, produce equal frequencies of numbers of whatever precision we chose? Yes, but they wouldn't be considered random. Randomness applies also to the *sequence* of the data. There must be no predictability of which number follows any given number in a random sequence. So the frequency of, say, the digit 1 followed by the digit 2 or digit 1 followed by digit 3 must be equal, and so on.

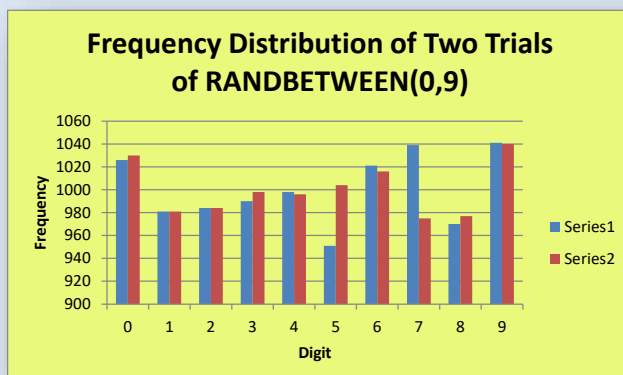
Another function that creates pseudo-random numbers in EXCEL is `=RANDBETWEEN(bottom,top)`. For example, `=RANDBETWEEN(0,9)` might produce the sequence 5 1 0 8 1 5 2 9 7 3 in one usage and 1 8 5 4 6 3 2 9 0 7 in the next.

Later we will study methods of applying statistical inference to this kind of analysis: we will be using Goodness-of-fit tests to evaluate the likelihood that a set of frequency observations deviate from expectation by random sampling alone.

INSTANT TEST P 7-9

Use Excel to create a list of 10,000 pseudo-random whole numbers between 0 and 9. Generate a frequency distribution showing the frequency of each digit in the list. Then do it again and compare the results. Are they what you expected?

Here are the results of such an exercise carried out by the author:



7.7 Random Number Generators

Some mathematical functions have been written that appear to create random number sequences. A simple one is the series of digits in the decimal portion of the number π ; another one is a simple procedure involving taking the fractional part of a starting number, using it as the exponent of a calculation, and taking the fractional part of the result as the next “random” number in the series – and then starting over using this new “random” number as the exponent for the next step. All these methods are called *iterative pseudo-random number generators* because they produce sequences that superficially look random but which in fact are perfectly repeatable and predictable if you know the rule (the *algorithm*) and the starting point (the *seed value*).

Another problem with iterative pseudo-random number generators is the precision of the calculations: eventually, it is possible for a generated number to be created that has already occurred in the sequence. The moment that happens, the sequence enters a loop. For example, if we considered the reduction to absurdity of having an iterative pseudo-random number generator that truncated its calculations at a single decimal digit, then the only numbers it could generate would be 0, .1, .2, .3... and .9. Suppose the sequence it generates were .4, .2, .5, .6 and then .4 again: the system would enter the endless loop of .4, .2, .5, .6, .4, .2, .5, .6, .4, .2, .5, .6 and so on. Not very random, eh?

The point of raising these issues here is not to make you experts on random number generators: it's to make you think about the concept of randomness and the fundamentals of probability. For now, it's enough to have you thinking about probability as the *expectation of observations in a random process*; that is, as average frequencies of occurrence for processes that have no obvious pattern.

7.8 Probabilities in Tossing Coins

A classic example used in introducing probabilities is the tossing of coins. A coin is tossed and lands with either one side (heads) or the other (tails) facing up. We deliberately ignore the rare cases where the coin lands on its edge and stays that way. We say that the probability of heads is $\frac{1}{2}$ and the probability of tails is $\frac{1}{2}$.

In general, the sum of the probabilities of all possible results in a defined system is always 1. The probability of impossible events is 0. So the probability of heads and the probability of tails in a single coin-toss is 1. The probability of 2 heads plus the probability of 1 head and 1 tail plus the probability of 2 tails in tossing a coin twice (or tossing two coins at the same time) is 1. We could write this latter assertion as

$$P\{H,H\} + P\{H,T\} + P\{T,H\} + P\{T,T\} = 1$$

7.9 Probabilities in Statistical Inference

In the ANOVA tables you have seen in previous sections and on the regression charts there were figures labeled p . These refer to the probability that there is no relationship among the data analyzed (that idea is called the *null hypothesis*); it is a measure of how likely we are to see results as deviant from the most likely expected result or more deviant by pure luck – by chance alone. As you will see in the discussion of hypothesis testing, looking at how likely our observed results are as a function of chance variations is a core idea for modern statistics.

INSTANT TEST P 7-10

Explain to yourself or to a buddy why the probability of getting two heads on top if you toss two coins is $\frac{1}{4}$. Then explain why the probability of getting one head and one tail on top is $\frac{1}{2}$ instead of $\frac{1}{4}$.

7.10 The Central Limit Theorem in Practice

What happens when we sample from a population?

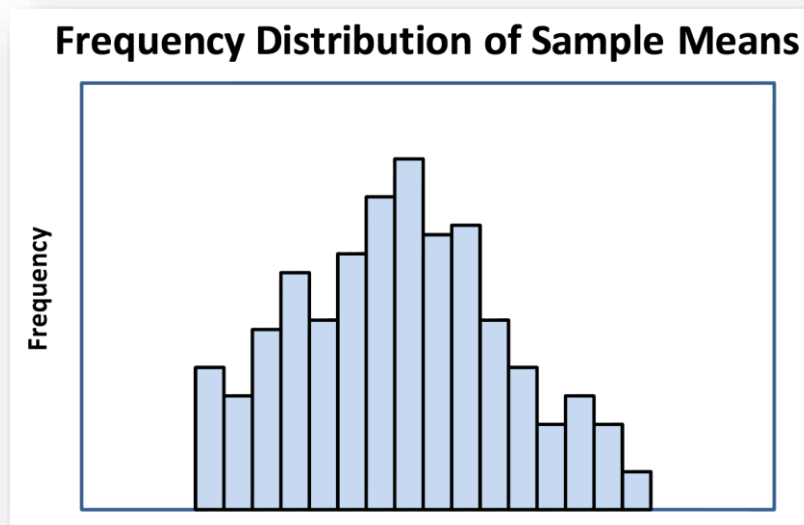
Does the sample reflect the characteristics of the population? Yes, but not in the sense of matching it exactly. Suppose we have a population of 4,821 widgets produced from assembly line #3 at the Urganian Corporation plant in Olympus Mons on June 14, 2219. The parametric mean length of the widgets is determined to be exactly 342 mm by measuring every single widget; the parametric standard deviation of the length is exactly 0.716 mm.

But now we take a sample of 100 widgets from the batch of 4,821 and discover that the sample mean of the lengths is 344 mm and the standard deviation is 0.922 mm. Then we take another sample of 100 widgets and – horrors – it doesn't match the population either: the mean length is 341 mm and the standard deviation is 0.855 mm.

There's nothing wrong here. We are seeing a demonstration of sampling variability and of the Central Limit Theorem. The interesting and important aspect of sampling is that, according to the Central Limit Theorem, the more samples we take, the closer the overall average of the sample statistics approaches the parametric value.

As we accumulate data from dozens of samples, we can build a frequency distribution showing how often different means occur in the samples; Figure 7-8 shows what such a graph might look like.

Figure 7-8. Means from dozens of samples.

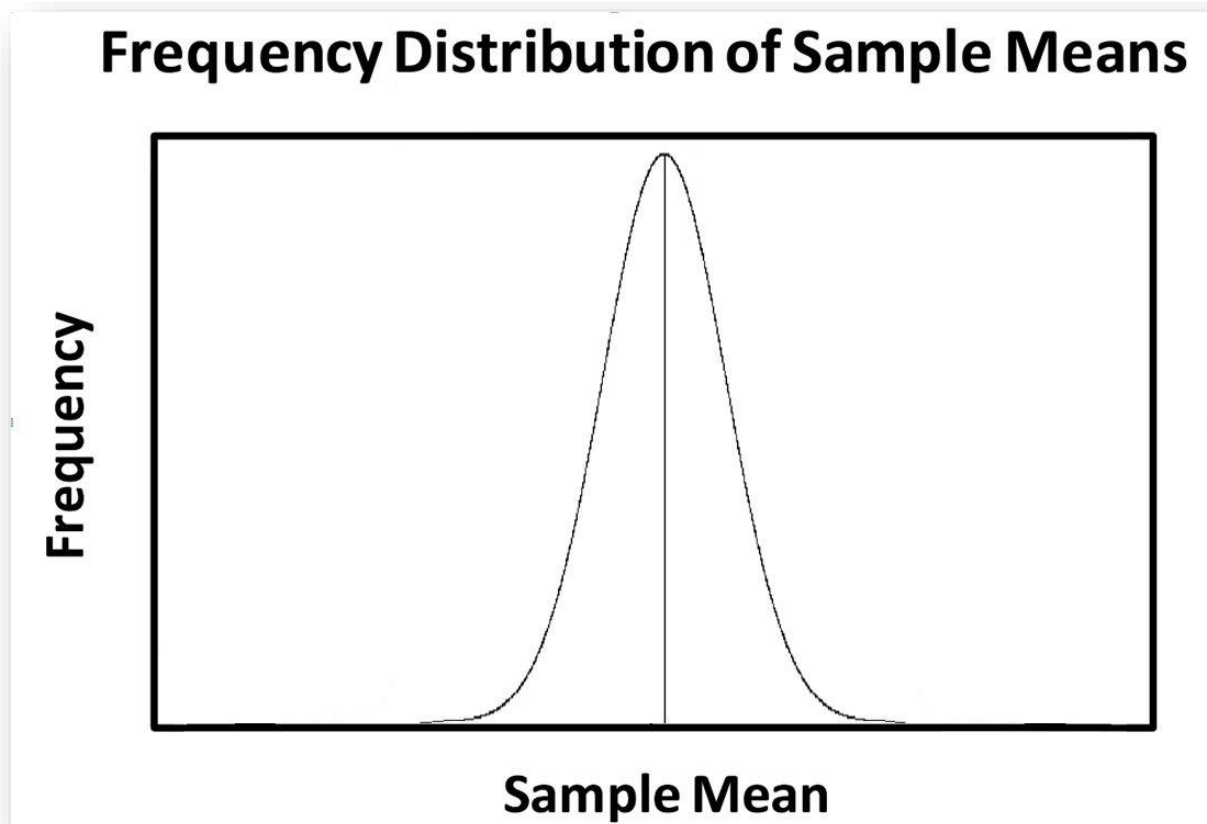


As the number of samples grows, what we find is that

- The distribution curve becomes more and more symmetrical;
- The curve gets smoother-looking;
- The mean of the distribution (\bar{Y}) approaches the parametric mean (μ) more and more closely;
- The variance (and therefore standard deviation) of the distribution gets smaller and the curve gets tighter around the mean.

Figure 7-9 shows the results of these tendencies as the number of samples grows into the hundreds. The image is inserted full width on the page because careful examination will reveal that the curve is actually a step function corresponding to the hundreds of samples. If there were thousands of samples, the curve would look smooth at this scale and would be even narrower around the mean.

Figure 7-9. Sampling distribution with hundreds of samples.



The effect of the Central Limit Theorem is stronger as the size of the individual samples rises; samples of size 100 show a faster approach to the kind of distribution shown in than samples of size 10.

The distribution shown in Figure 7-9 is a Normal distribution. The Central Limit Theorem can be stated in intuitive terms as follows:

As sample size increases, the means of random samples drawn from a population of *any* underlying frequency distribution will approach a Normal distribution *with its mean corresponding to the parametric mean of the source distribution*.

The Central Limit Theorem is enormously important in applied statistics. It means that even if an underlying phenomenon doesn't show the attributes of the Normal distribution, the means of samples will be normally distributed. Since so much of modern statistics assumes a Normal distribution for the underlying variability of the phenomena being analyzed, the Central Limit Theorem means that we can circumvent non-normality by working with samples of observations – groups of data – instead of with individual observations.

7.11 The Expected Value

The Central Limit Theorem also brings to light another concept of great importance: the *expected value* of a statistic. The expected value is the average of a statistic computed over an infinite number of samples.

For example, the *expected value of the sample mean* is the *parametric mean*, μ . We say that the observed sample mean is a *point estimator* of the parametric mean – a single value that indicates what the parameter may be.

Statisticians have also shown that the expected value of the variance of the sample means ($\sigma^2_{\bar{Y}}$) is the ratio of the parametric variance (σ^2) to the sample size (n) and thus the expected standard deviation of the sample means ($\sigma_{\bar{Y}}$) is expected to be the parametric standard deviation (σ) divided by the square root of the sample size (\sqrt{n}):

$$\sigma^2_{\bar{Y}} = \frac{\sigma^2}{n} \qquad \sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

The standard deviation of the mean is called the *standard error of the mean*. In general, the standard deviation of any statistic is known as the *standard error* of that statistic.

Another interesting application of the Central Limit Theorem is that, in the absence of any further information, whatever we encounter is most likely to be *average*. So for example, suppose we are working on a very important project such as writing a statistics textbook and the phone rings; the caller-ID shows an unknown caller. In the absence of further information, the call is most likely to be of *average importance*.⁷⁶

Therefore, one can reasonably defer answering the call and allow it to go to voice-mail for later processing.

Similarly, if one's activity is less than average in importance (for instance, watching the Olympic Paint Drying Championships on holovision), then one can reasonably answer an unknown caller.

Statistics in action!

7.12 More About the Normal Distribution

In this text, the *Normal distribution* has a capital N for *Normal* to be sure that no one thinks that there is anything abnormal about non- Normal distributions! The Normal distribution plays an important role in statistics because many ways of describing data and their relationships depend on a Normal error distribution. For example, in the linear regression that you will study later, the error term ϵ in the equation

$$Y_{ij} = a + bX_i + \epsilon_{ij}$$

represents a Normally-distributed error with a mean of zero and a variance defined by what is called the *Residual MS* (*residual mean square*) in the ANOVA table. In other words, the linear model defines a best-fit line for Y , the expected or predicted dependent variable, as a function of the Y -intercept (a , the value of Y when X , the independent variable, is zero) plus the product of the slope b times the value of X , plus the random (unexplained) error ϵ .⁷⁷

Most of the descriptive statistics and statistical tests we use routinely are called *parametric statistics* because they assume a Normal distribution for the error term or unexplained variance. ANOVA, ANOVA with regression, t-tests, product-moment correlation coefficients (all to be studied in detail later in this course) uniformly assume a Normal error distribution. There are other assumptions too, which we will also discuss later; one important concept is that the mean and the variance of a statistic are supposed to be *independent*. That is, for

⁷⁶ ...and average duration and average volume and average origin and average annoyance-value and average....

⁷⁷ As explained in the Preface, most textbooks *never* make forward references. However, this textbook often makes such references so that when students come to study the details of a technique, they will have encountered it – and some modest amount of information about how it is used – long before they have to learn the details.

parametric statistical analysis like t-tests and ANOVA, we assume that small observed values have the same variability as large observed values for the same variable.

A counter-example is the relation between, say, size and weight. No one would expect the variance of the weights of tiny model sailboats weighing a kilogram or so to be the same as the variance of the weights of battleships weighing tens of thousands of tons.

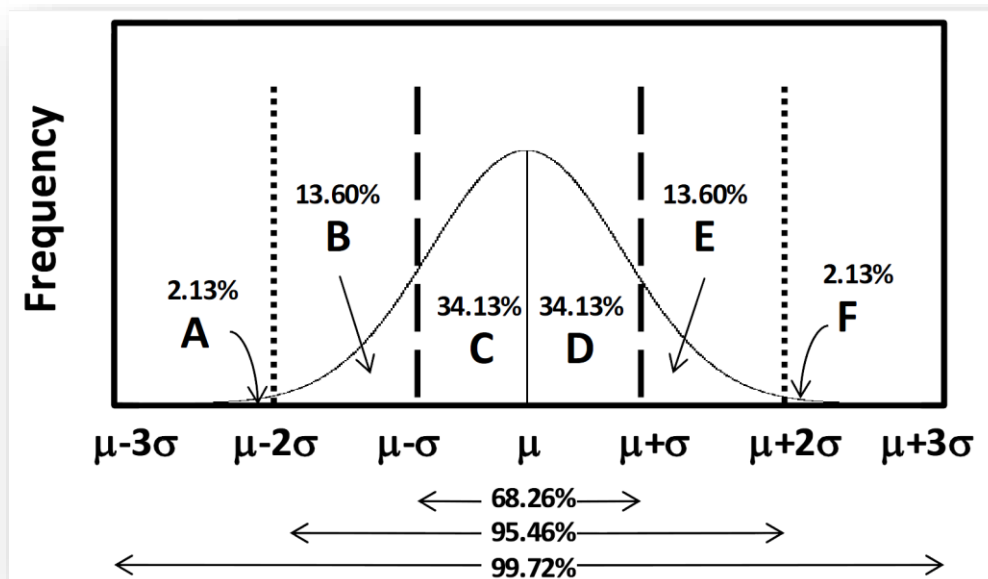
When these assumptions are not justified, we have to use *non-parametric statistics*. Examples include using the median and the mode instead of the mean and the range and other non-parametric measures of variability instead of the variance or standard deviation. Non-parametric tests can include comparisons of ranks instead of measured variables. Examples of such tests include

- Kruskal-Wallis Test for comparisons of central tendency of ranked data
- Friedman's Method for Randomized Blocks for comparisons of data that don't fit the Normal curve
- Mann-Whitney U-Test for comparing percentages
- Kolmogorov-Smirnov Two-Sample Test for comparing frequency distributions
- Wilcoxon's Signed Ranks Test for Two Groups of Paired Data
- Spearman's Coefficient of Rank Correlation.⁷⁸

We use the Normal distribution so much that some attributes become familiar simply by force of repetition.⁷⁹

Figure 7-10 shows some of the most often used characteristics of the Normal distribution: the areas under the curve as a function of distance from the mean (μ) measured in standard deviations (σ).

Figure 7-10. Characteristics of the Normal distribution.



⁷⁸ The names of these tests are included in case you desperately need to analyze rank data or other non-Normal data; you can look up how to perform the tests in your statistical package or in standard textbooks. Eventually, this text will expand to include such methods for upper-year courses.

⁷⁹ Don't memorize this stuff – just learn it by using it!

In Figure 7-10, the letters and numbers indicate the following relationships:

- The probability that an observation picked at random from a Normally-distributed variable with defined mean μ and standard deviation σ will be smaller than three standard deviations (often spoken of as *less than three sigmas*) away from the mean is about 2.13%. This area corresponds to the section marked A in the figure. We can write this statement conventionally as

$$P\{Y \leq \mu - 3\sigma\} \approx 2.13\%$$

- In other words, the probability of encountering a normal variate that is at or less than 3 sigmas below the parametric mean is 2.13% or roughly 1 in 47 tries.
- Because the Normal distribution is perfectly symmetric, area F is identical to area A; that is,

$$P\{Y \geq \mu + 3\sigma\} \approx 2.13\%$$

- Section B in the figure represents the 13.6% probability that an observation picked at random from a Normally distributed variable will lie between 1 and 2 sigmas below the mean. That is,

$$P\{\mu - 2\sigma \leq Y \leq \mu - \sigma\} \approx 13.6\%$$

- Section E corresponds to section B on the other side of the mean, so we can also write

$$P\{\mu + \sigma \leq Y \leq \mu + 2\sigma\} \approx 13.6\%$$

- Areas C and D correspond to the chance of picking a variate at random which lies within one sigma of the mean. Together, C and D add up to 68.2% of the total area under the curve (shown on the arrows below the abscissa), implying that more than 2/3 of all values picked at random from a Normally distributed population will lie between $\mu - \sigma$ and $\mu + \sigma$.

$$P\{\mu - \sigma \leq Y \leq \mu + \sigma\} > 67\%$$

- Similarly, the figure illustrates the rule of thumb that about 95% (95.46%) of randomly-chosen members of a Normally-distributed population will lie between $\mu - 2\sigma$ and $\mu + 2\sigma$ (that is, within two sigmas of the mean).

$$P\{\mu - 2\sigma \leq Y \leq \mu + 2\sigma\} > 95\%$$

- More than 99% (99.72%) of the population lies between $\mu - 3\sigma$ and $\mu + 3\sigma$ (within three sigmas of the mean).

$$P\{\mu - 2\sigma \leq Y \leq \mu + 2\sigma\} > 99\%$$

In ordinary life, we can make use of these approximate values when evaluating statements about the Normality or non-Normality of specific observations if we know something about the underlying population distribution and have some confidence that the error distribution is Normally distributed.

For example,

- Imagine that we know that the average height of Lunar Colony men of 19 years of age in a study is 172.1 cm with standard deviation of 2.8 cm.
- We can assert from these data that about 2/3 of the 19-year-old males in the Lunar Colony have a height between $172.1 - 2.8$ cm and $172.1 + 2.8$ cm.
- That's about 66% of the 19-year-old boys between 169.3 cm and 174.9 cm.
- For you non-metric folks, that's about 6' 3.6" and 6' 6.1" with a mean of 6' 4".
- Similarly, about 99% would be within three sigmas, which would be 163.7 cm and 180.5 cm (6' 1.1" and 6' 8.6").

- So if someone said of a young man, “Wow, he’s really exceptionally short” because he was only 6’ 2” high, we could identify the statement as inaccurate – unless the speaker’s definition of “exceptionally” were unusually liberal and included men within the range of 99% of the Lunar Colony population of that age.

7.13 Statistical Inference: Interval Estimation

Knowing that random samples vary according to particular patterns – for instance, the means of samples approach a Normal distribution – means that we can *estimate the parametric value* based on a *sample value*.

For example, if we sample randomly from the Lunar Colony boys of 19 earth years and measure the heights of 25 of them, the mean of that sample should tell us something about the mean of the population. Using the Central Limit Theorem, we assert that *our best estimate of the parametric mean*, in the absence of any other information, *is the sample mean*.

However, common sense tells us that the sample mean of any one sample may differ from the parametric mean; intuitively, it doesn’t seem reasonable to expect that a random sample would magically be exactly centered on the population mean. Therefore, we compute an *interval estimate* for the parametric mean using our knowledge of the sample mean and of the variability and pattern of distribution of such means.

An *interval estimate* for any statistic is a range with lower and upper *confidence limits*. Typical $(1 - \alpha)$ confidence limits for any interval estimate of a parametric value are called the $(1 - \alpha)$ confidence limits. For example, we often refer to the 95% *confidence limits* of a statistic, where $\alpha = 0.05$. Another common choice is the 99% confidence limits of a statistic, where $\alpha = 0.01$.

These intervals are interpreted as follows:

- The *probability of being correct* in asserting that the $(1 - \alpha)$ confidence limits include the value of the parametric statistic is $(1 - \alpha)$.
- The *probability of being wrong* in asserting that the $(1 - \alpha)$ confidence limits include the value of the parametric statistic is α .

Here are some examples of interval estimates for a variety of made-up statistics and different ways of interpreting them:

- The sample mean cost of a trans-Jovian flight in 2219 is 1,452 credits; the 95% confidence limits are 1167 and 1736 credits. There is a 95% chance of being correct in guessing that the mean cost lies between 1167 and 1736 credits. There is therefore a 5% chance of being wrong in that assertion.
- A sample of Martian fornselling beans has a mean growth potential of 182% per month; the 90% confidence limits are 140% to 224%. There is only a 10% chance of being wrong in claiming that the growth potential is between 140% and 224% per month.
- A study of Norwich University students’ usage of fornselling chips consumed per month in 2219 showed an average of 3.8 kg per student with 80% confidence limits of 2.3 to 5.3 kg. We would be right 80% of the time that we repeat this kind of sampling and computation of the confidence limits for the consumption figures. We’d be wrong in 20% of the estimates based on samples of that size.
- The Gorgonian factor calculated for a sample of 3,491 Sutellian customers indicated an average time to immobility upon exposure to the Gorgonian advertisements of 2 hours 12 minutes with 99% confidence limits of 1 hour 54 minutes through 2 hours 36 minutes. Our chance of being wrong in using this procedure to guess at the correct time to immobility is only 1 time out of a hundred. That is, if we were to repeatedly take random samples of size 3,491 Sutellians exposed to the same ads, in 99 out of a hundred experiments, the computed interval estimates would correctly include the parametric mean time to immobility.
- The mean variance for the sales resulting from exposure to a modified mind-control regimen projected through the holographic broadcasting networks was 3,622 with 95% confidence limits of

1,865 and 5,957. Using this confidence-interval calculation procedure, we have a 95% probability of really including the true parametric sales figure in our computed interval.

Students may have noticed that in *no case* above were the confidence intervals interpreted as follows: “*The probability that the parametric statistic is between the lower and upper $(1 - \alpha)$ confidence limits is $(1 - \alpha)$.*” All of the interpretations were in terms of the chance of *being right (or wrong)* in asserting that *the limits include* the parametric value, *not* the probability that the parameter is between the limits. The parameter is fixed for a population; it is the estimates that vary around it. Sokal and Rohlf explained this subtle point as follows in their classic textbook:

“We must guard against a common mistake in expressing the meaning of the confidence limits of a statistic. When we have set lower and upper limits ... to a statistic, we imply that the probability of this interval covering the mean is 0.95, or, expressed in another way, that on the average 95 out of 100 confidence intervals similarly obtained would cover the mean. We cannot state that there is a probability of 0.95 that the true mean is contained within any particular observed confidence limits, although this may seem to be saying the same thing. The latter statement is incorrect because the true mean is a parameter; hence it is a fixed value and it is therefore either inside the interval or outside it.

It cannot be inside a particular interval 95% of the time. It is important, therefore, to learn the correct statement and meaning of confidence limits.”⁸⁰

Notice that most confidence limits are symmetrical, but that the mind-control variance was not; there is no guarantee that confidence limits will be symmetrical. The exact calculations for the upper and lower limits depend on the nature of the variation of the sample statistics. For example, most statistics are normally distributed, but percentages in the low (less than 10%) and high (greater than 90%) ranges are typically not normally distributed because there is greater uncertainty (and therefore variability) about the extremes than about the central portion of the distribution. We will learn more about how to handle such non-Normal measurement scales in the introduction to *data transforms* later in the text.

7.14 Population Mean Estimated Using Parametric Standard Deviation

One of the most common calculations in statistics is the estimation of the confidence limits for a mean. The exact calculations depend on whether we know the parametric standard deviation or not.

When we know the population standard deviation (for example, if the population has long been studied and the variance of the statistic in question is established from many repeated measurements and no longer considered to be an estimate) then we can use it directly in the calculation of the confidence limits.

In our introduction to the Normal distribution, we learned that for a Normally distributed variable with mean μ and standard deviation σ , 95% of the values lie between -1.96σ and $+1.96\sigma$ from the mean, μ . That is,

$$P = \left\{ -1.96 \leq \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} \leq +1.96 \right\} = 0.95$$

Recalling that for samples of size n , sample means \bar{Y} are normally distributed around the parametric mean with standard error of the mean

$$\sigma_{\bar{Y}} = \sigma / \sqrt{n}$$

we can thus write the calculation for the 95% confidence limits to the mean as

$$P \{ \bar{Y} - (1.96\sigma / \sqrt{n}) \leq \mu \leq \bar{Y} + (1.96\sigma / \sqrt{n}) \} = 0.95$$

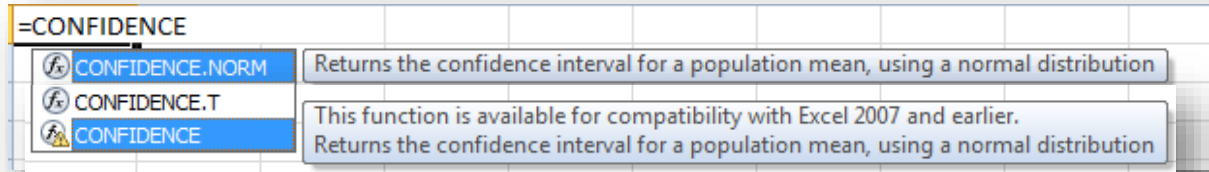
Or more generally, where z_{α} is the critical z-score corresponding to the probability α to the left of its value, then correspondance

⁸⁰ (Sokal and Rohlf, *Biometry: The Principles and Practice of Statistics in Biological Research* 1981) p 144.

$$P\{\bar{Y} - z_{\alpha} s_{\bar{Y}} \leq \mu \leq \bar{Y} + z_{\alpha} s_{\bar{Y}}\} = 1 - \alpha$$

The composite image in Figure 7-11 shows the two EXCEL 2010 functions that perform this calculation automatically:

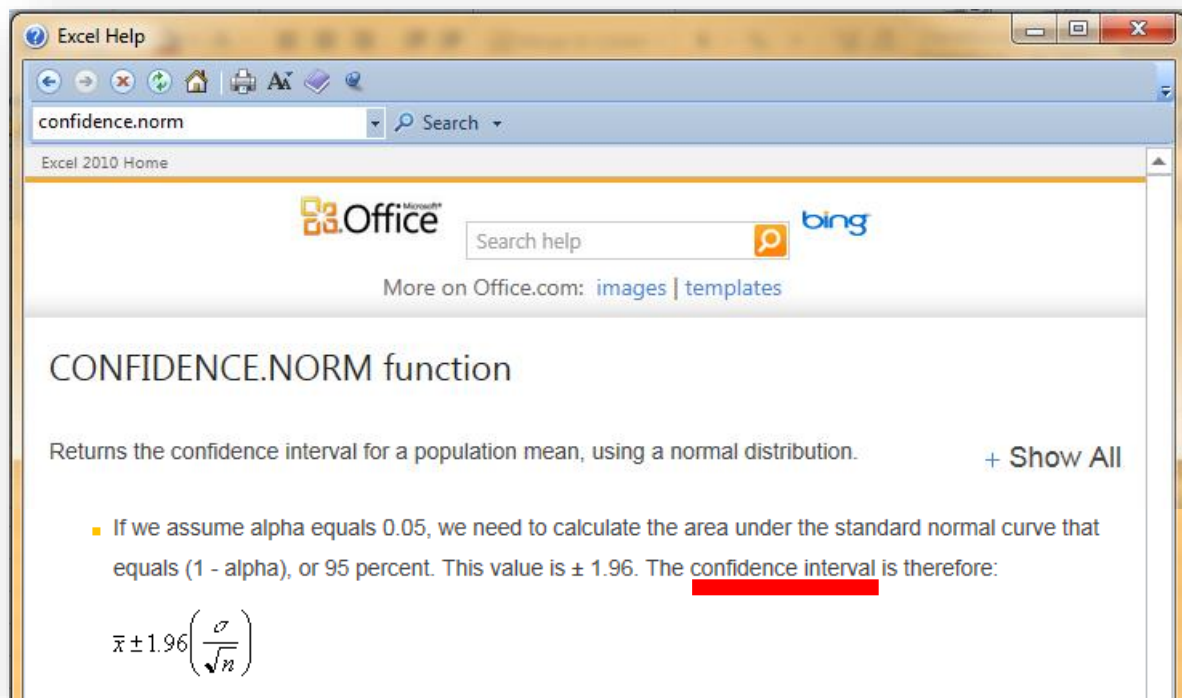
Unfortunately, the **HELP** text in EXCEL 2010 for these functions has a mistake, as highlighted by the red [Figure 7-11. Calculating confidence limits for the mean when the *parametric* standard deviation is known.](#)



underline in the composite image below (Figure 7-12).

This function computes *half* of the confidence interval, not the confidence interval or the confidence limits.

[Figure 7-12. HELP text with error.](#)



The formula shown at the lower left of Figure 7-12 defines the upper and lower confidence *limits*, not the confidence *interval*. The \pm symbol implicitly defines (using the $-$) the *lower confidence limit* (sometimes denoted L_1) and (using the $+$) the *upper confidence limit*, sometimes denoted L_2 . The confidence *interval* is $L_2 - L_1 = 2 * 1.96(\sigma/\sqrt{n})$.

Using the function name and our Y variable,

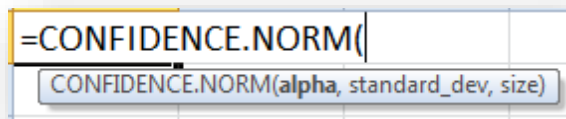
$L_1 = \bar{Y} - \text{CONFIDENCE.NORM}(\text{parms})$ and

$L_2 = \bar{Y} + \text{CONFIDENCE.NORM}(\text{parms})$.

Figure 7-13 shows the parameters (*parms*) for **=CONFIDENCE.NORM**.

- The parameter **alpha** is the complement of the confidence level; thus for 95% confidence, $\alpha = 0.05$.
- The parameter **standard_dev** is the parametric standard deviation.
- The parameter **size** is the sample size n .

For example, if we know from Figure 7-13. Parameters for computing confidence limits to the mean given the parametric standard deviation.



Barsoomian roncinal weights is 23 kg and we acquire a sample of 12 roncinals whose mean weight is 745 kg, we can easily calculate that the lower and upper confidence 95% confidence limits are as shown in the composite image of Figure 7-14 .

Figure 7-14. Calculating lower and upper confidence limits for the parametric mean given the parametric standard deviation.

	A	B		A	B
1	Sample mean:	745	1	Sample mean:	745
2	Confidence level:	95%	2	Confidence level:	0.95
3	Parametric std dev:	23	3	Parametric std dev:	23
4	Sample size:	12	4	Sample size:	12
5	Function output:	13.013	5	Function output:	=CONFIDENCE.NORM((1-B2),B3,B4)
6	Lower confidence limit:	732.0	6	Lower confidence limit:	=+\$B\$1-\$B\$5
7	Upper confidence limit:	758.0	7	Upper confidence limit:	=+\$B\$1+\$B\$5

INSTANT TEST P 7-19

Duplicate the calculations shown above in your own spreadsheet but don't use any \$ signs in the formulas so you can propagate the formulas sideways. Create 6 columns of data with confidence levels 80%, 85%, 90%, 95%, 99% and 99.9%. Graph the lower and upper confidence limits against the confidence level. Discuss your findings in the discussion group on NUoodle for this week.

7.15 Estimating Parametric Mean Using the Sample Standard Deviation

What happens if we don't know the *parametric* standard deviation (which is the same as saying we don't know the parametric variance)?

We use the *sample* standard deviation, s to compute an *estimate* of the standard error of the mean

$$s_{\bar{Y}} = s/\sqrt{n}$$

where n is the sample size and s is the standard deviation of the sample. Then the statistic

$$(\bar{Y} - \mu)/s_{\bar{Y}}$$

is distributed as a *Student's-t distribution* with $n - 1$ degrees of freedom.

These deviates will be more variable than those computed using a single parametric value for the standard error because sometimes the sample s will be smaller than the parametric σ and sometimes it will be larger. It follows that the frequency distribution for the ratio

$$(\bar{Y} - \mu)/s_{\bar{Y}}$$

must be different from the distribution of

$$(\bar{Y} - \mu)/\sigma_{\bar{Y}}$$

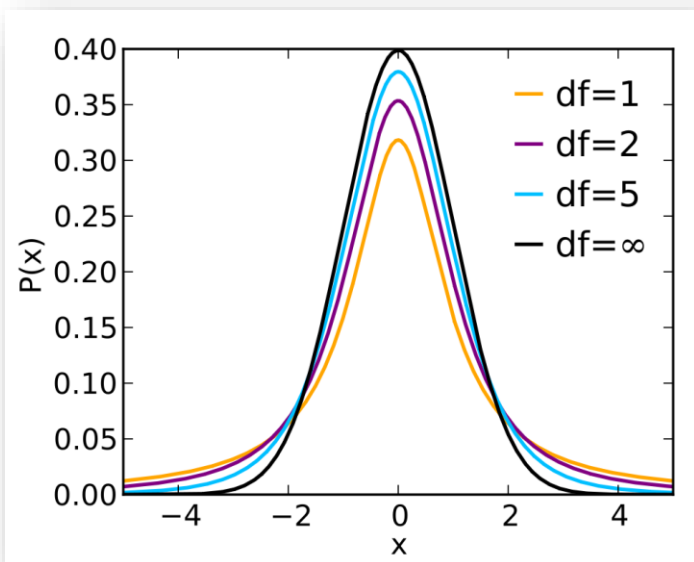
that we studied earlier: it will be broader because of the variations in the denominator.

In fact the distribution of

$$(\bar{Y} - \mu)/s_{\bar{Y}}$$

is called *Student's-t distribution* and was published in 1908 by the famous English mathematician William Sealy Gosset (1876-1937) who published extensively in statistics under the pseudonym *Student*. The distribution is actually a family of curves defined by the sample size: the *degrees of freedom* of each distribution is one less than the sample size on which the sample standard deviation is computed. Note that when $df = \infty$, Student's-t distribution is the Normal distribution. Figure 7-15 illustrates this relationship between Student's t and the Normal distribution.⁸¹

Figure 7-15. Family of Student's-t distributions.



⁸¹ Image used in compliance with Creative Commons Attribution 3.0 license from <
http://upload.wikimedia.org/wikipedia/commons/thumb/4/41/Student_t_pdf.svg/1000px-Student_t_pdf.svg.png > or <
<http://tinyurl.com/9abxyu> >.

7.16 Degrees of Freedom Vary in Statistical Applications

We use *degrees of freedom* (df) extensively in our work in statistics. One interpretation is that if we have n data in a sample, calculating the *sum of the values* fixes $(n - 1)$ of the values; i.e., knowing the sum, we don't need to know the last of the values, since it can be computed as the sum minus the sum of the other $(n - 1)$ data. Thus only $(n - 1)$ of the data are free to vary – hence the degrees of freedom are $(n - 1)$. However, the exact computation of the degrees of freedom for a statistic is particular to each type of statistic.

As mentioned in the previous section, Student's-t distribution approaches the Normal distribution more and more closely as the degrees of freedom rise; indeed the Normal distribution *is* Student's-t distribution with infinite degrees of freedom. Think of the approach of Student's-t distribution to the Normal distribution with increasing sample size as another example of the Central Limit Theorem.

7.17 Notation for Critical Values

The *critical value* of Student's-t distribution with $n - 1$ degrees of freedom below which α of the distribution lies is written as

$$t_{\alpha[n-1]}.$$

In more general use, the degrees of freedom are represented by the letter ν (*nu*), the Greek equivalent of n . Thus you might see this form:

$$t_{\alpha[\nu]}$$

to represent the critical value of Student's t for ν degrees of freedom which has a probability of α of having values that large or smaller. On the probability distribution, $t_{\alpha[\nu]}$ thus demarcates the portion α of the curve to the left and the portion $(1 - \alpha)$ to the right of that value. The square brackets are a typical way of separating the degrees of freedom from the critical probability.

7.18 Two-Tailed Distributions

Because the Normal distribution and the Student's-t distribution are symmetric around the mean, they are called *two-tailed* probability distributions.

In practice, we express the confidence limits based on a sample mean \bar{Y} with unknown parametric standard deviation as follows:

$$P\{\bar{Y} - t_{\alpha/2 [n-1]} s_{\bar{Y}} \leq \mu \leq \bar{Y} + t_{\alpha/2 [n-1]} s_{\bar{Y}}\} = 1 - \alpha$$

Notice that α represents the *total of the area* below and above the $(1 - \alpha)$ confidence limits. Each *tail* of the distribution represents a probability of $\alpha/2$.

So to compute the $(1 - \alpha)$ confidence limits of a population mean given the sample mean \bar{Y} and that sample's standard deviation s and sample size n , we

- (1) Compute the standard error of the mean as

$$s_{\bar{Y}} = s/\sqrt{n}$$

- (2) Locate the absolute value (e.g., $|-3| = 3 = |+3|$) of the t-statistic corresponding to a *left tail* of probability $\alpha/2$; that is,

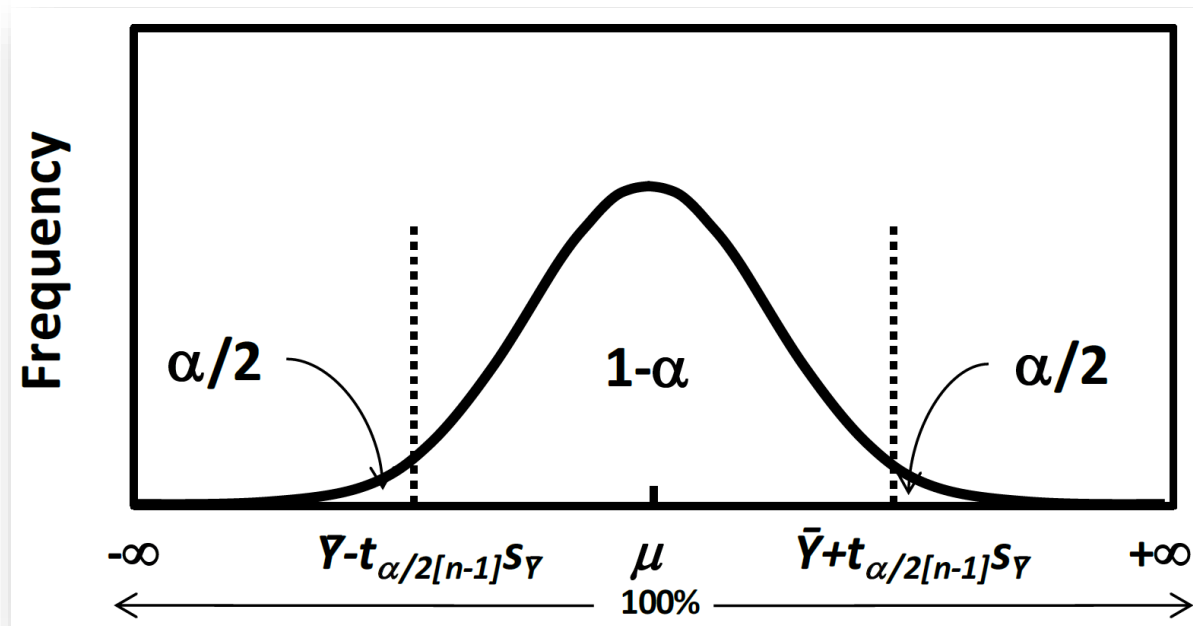
$$|t_{\alpha/2 [n-1]}|$$

- (3) Compute the lower and upper confidence limits as

$$L_1 = \bar{Y} - |t_{\alpha/2 [n-1]}| s_{\bar{Y}} \quad \text{and} \quad L_2 = \bar{Y} + |t_{\alpha/2 [n-1]}| s_{\bar{Y}}$$

Figure 7-16 shows the two-tailed probabilities for distributions like the Normal and the Student's-t.

Figure 7-16. Confidence limits for the mean using Student's-t distribution and two-tailed probabilities.



7.19 EXCEL CONFIDENCE.T Function

The EXCEL 2010 function, `=CONFIDENCE.T`, that computes half the confidence interval for a parametric mean based on a sample mean \bar{Y} and its observed sample standard deviation s . The function is illustrated in the center of Figure 7-11 and is highlighted below in Figure 7-17.

Figure 7-17. Function for confidence limits of a mean knowing the *sample* standard deviation.

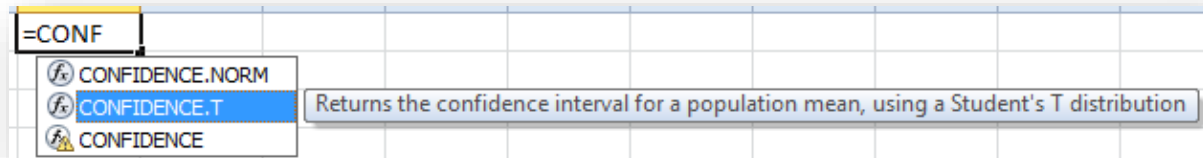


Figure 7-18. Confidence limits for the mean based on the sample standard deviation.



	A	B		A	B
1	Sample mean:	745	1	Sample mean:	745
2	Confidence level:	95%	2	Confidence level:	0.95
3	Sample std dev:	23	3	Sample std dev:	23
4	Sample size:	12	4	Sample size:	12
5	Function output:	14.614	5	Function output:	<code>=CONFIDENCE.T((1-B2),B3,B4)</code>
6	Lower confidence limit:	730.4	6	Lower confidence limit:	<code>=+\$B\$1-\$B\$5</code>
7	Upper confidence limit:	759.6	7	Upper confidence limit:	<code>=+\$B\$1+\$B\$5</code>

`=CONFIDENCE.T` works much like `=CONFIDENCE.NORM`. Figure 7-18 shows the calculations and formulae. Comparing this result with Figure 7-14, we note that the confidence limits are noticeably wider apart. Where the limits based on the parametric standard deviation were 732.0 and 758.0, the limits using the sample standard deviation are 730.4 and 759.6. The latter confidence interval is roughly 112% of the former confidence interval.

7.20 Beware the Definition of α in Inverse Probability Functions

There is another function that can be useful in many other applications, including computing confidence limits for other kinds of statistics than the mean. Many applications can use the `=T.INV` (or its older equivalent, `=TINV`) and the `=T.INV.2T` functions to compute critical values for $t_{\alpha/2}$.

Figure 7-19. Same word, different meanings.

=T.INV	
 T.INV	Returns the left-tailed inverse of the Student's t-distribution
 T.INV.2T	Returns the two-tailed inverse of the Student's t-distribution

However, a critical *issue* when computing critical *values* is that these functions – and other functions you will encounter in your explorations of EXCEL and of other statistical packages – have different, contradictory definitions of the **probability** parameter! Figure 7-19 shows the syntax of the EXCEL 2010 functions; notice that the parameter **probability** occurs in both.

=T.INV(=T.INV.2T(
T.INV(probability, deg_freedom)	T.INV.2T(probability, deg_freedom)

- The `=T.INV` function yields a critical value using the left tail of the probability distribution, which means that if we enter `=T.INV(.025, df)`, the function yields the *left-tail* critical value $t_{0.025|v}$ (a negative number) which produces 0.025 on the *left* of the critical value and 0.975 to the right of the critical value. Because of the symmetry of the Student's-t distribution, that also means that the 0.025 of the distribution lies to the *right* of $|t_{0.025|v}|$ and 0.975 of the distribution lies to the *right* of that value.
 - For example, we can calculate `=T.INV(0.025, 100) = -1.98397`. We write this as $t_{0.025|100} = -1.98397$ or as $t_{0.975|100} = +1.98397$
 - Thus the probability parameter in this function generates a critical value corresponding to a *one-tailed* probability for the *left-tail* critical value, a *negative* number.
- Now consider `=T.INV.2T`, which, as the `.2T` indicates, uses a *two-tailed* probability – and in addition, computes the *right-tail* critical value, a positive number.
 - For example, we can calculate `=T.INV.2T(0.05, 100) = 1.98397`. Exactly as above, we write this as $t_{0.025|100} = 1.98397$. Notice that we have to describe it as cutting off a right tail with 0.025, not the 0.05 entered into the function!
 - So as you can see, the `=T.INV.2T` function automatically computes a critical value that defines the left tail and the right tail for the critical value as having *half the probability* entered in the **probability** parameter.

Remember this example: **you must verify whether a statistical function in any statistical package computes left-tail or right-tail critical values using one-tailed or two-tailed probabilities.** Don't get upset or curse the programmers for inconsistency: just check to make sure you are computing what you *need*, not what you *hope*.

If you want to check your understanding of a new function you haven't used before, you may be able to check your understanding using known results for known parameters to ensure that you are not mistaken in your understanding of what the new parameters mean for the new function.

7.21 Interval Estimate for Any Normally Distributed Statistic

Even more generally, we can extend the applicability of the t-distribution and its use in computing interval estimates of a parametric value to any normally distributed statistic.

Suppose we find a statistical research article that discusses the distribution of a new statistic, say, the “delta” coefficient (δ for parameters, d for samples). Imagine that this (made-up) statistic is important in evaluating the reliability of warp cores on starships. Extensive research confirms that the δ coefficient is indeed normally distributed. Starship engineers need to estimate the confidence limits for the parametric value δ given a sample’s d statistic because any value smaller than 3 or greater than 4 may lead to a faster-than-light engine implosion. Galaxyfleet insists on a chance of implosion of less than 5%.

The principle is that any *normally distributed* statistic (in this imaginary example, d) – whose standard error is s_d with ν (for example, $\nu = n - 1$ degrees of freedom for a sample size of n) will fit the same pattern as what we have seen in computing confidence intervals for means using the Student’s-t distribution:

$$P\{d - |t_{\alpha/2[\nu]}| s_d \leq \delta \leq d + |t_{\alpha/2[\nu]}| s_d\} = 1 - \alpha$$

That is, interpreting this algebraic formulation,

- The probability P
- That we will be correct
- In asserting that the lower and upper computed $(1 - \alpha)$ confidence limits for δ
- With ν degrees of freedom
- Include the parametric statistic δ
- Is $(1 - \alpha)$.

Example:

- An engineer finds that in a sample of 100 warp-core signatures, the sample d statistic is 3.48 and the standard error of d (s_d) with $\nu = 99$ degrees of freedom is 0.081.
- The two-tailed function $=T.INV.2T(.05,96)$ for $\alpha = 0.05$ (i.e., $\alpha/2$ in each tail) and $\nu = 100$ in EXCEL gives us the critical upper-tail value $t_{0.05 [96]} = 1.984216952$ which in turn lets us compute the differential $t_{\alpha/2[\nu]} * s_d = 1.984216952 * 0.081 = 0.160721573$ or ~ 0.161 . So the limits of our interval estimate for δ are 3.48 ± 0.161 or 3.319 and 3.641.
- Thus we have a 95% chance of being correct in asserting that the interval 3.319 to 3.641 based on our sample values of d includes the parametric delta-coefficient δ .
- The confidence limits are within the range of acceptability, so the engineer concludes that the chances of a warp-core implosion today are less than 5%.

This example shows one of the ways that confidence limits can be used in quality control.

7.22 Population Proportion Based on Sample Proportion

In a study of the value of warning labels on pharmaceutical products, the BigPharma Association of the Greater Solar System looked at a sample of 2,000 sentients out of the total population of about 397,452,000 known to have been legally prescribed drugs with such warning labels and counted how many had bothered to read the labels. They found that the proportion p_{sample} of readers-of-labels was 22.5%. What was the 95% confidence interval for the parametric proportion $p_{\text{population}}$ of readers-of-labels?⁸²

Statisticians have shown that repeated measures of p_{sample} with sample size n are distributed as a Normal distribution with the mean $p_{\text{population}}$ (as expected under the Central Limit Theorem) and parametric variance

$$\sigma_p^2 = p(1 - p)/n$$

provided that

- The population size N is infinite *or* that
- That the ratio of the sample size n to the population size N meets the condition

$$n/N \leq 0.05$$

- And that $np > 5$ and $n(1 - p) > 5$

In other words, provided that the sample size n is less than 5% of the total sample size N , the approximation for the parametric variance of proportions works fine.

It follows that the parametric *standard error* of the sample proportion, σ_p is defined as

$$\sigma_p = \sqrt{\frac{p_{\text{sample}}(1 - p_{\text{sample}})}{n}}$$

Once we know how to compute the standard error of the proportion and we know that the sample proportions are Normally distributed, we can compute the interval estimate for the population proportion using same principles described in §7.21 above.

Figure 7-20. Confidence limits for a proportion p .

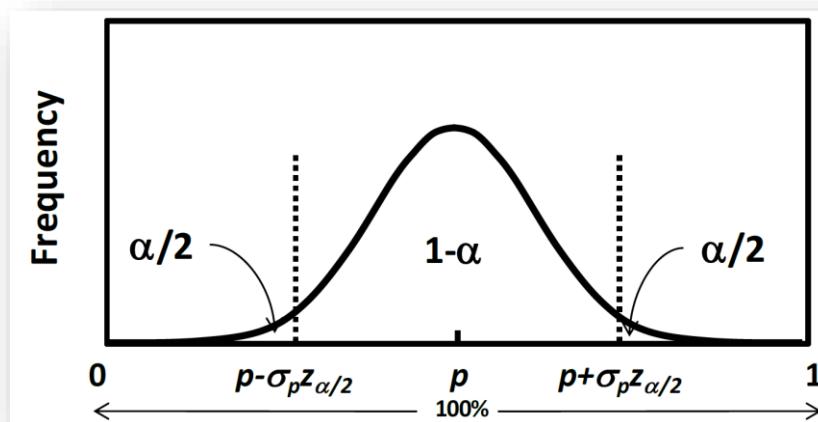


Figure 7-20 shows the meaning of the confidence limits graphically.

⁸² We don't use the Greek version of p (π) because it is too strongly associated with the ratio of the circumference to the radius of a circle in plane geometry. However, note that the capital version, Π , is frequently used to mean "product" in the same way that capital sigma, Σ , is conventionally used to mean "sum."

Figure 7-21 provides an example in EXCEL for computing *confidence limits for the proportion* of sentiments who bothered to read the warning labels on pharmaceuticals in a study on planet Phlrn'thx in the Galaxyfleet year 2712. Column D in the spreadsheet displays the formulas used in Column B.⁸³

Figure 7-21. Demonstration of computing confidence limits for proportion.

	A	B	C	D
1	DATA INPUT		Check	Formulas Displayed for Example Only
2	total population N	397,452,000		
3	sample size n	2,000		
4	observed proportion p	0.225		
5	desired confidence level 1- α	0.95		
6				
7	PRELIMINARY CHECKS			
8	$n/N \leq 5?$	5.03E-06	OK	=+B3/B2 and in Check column =IF(B8<=5,"OK","NO")
9	$np > 5?$	450	OK	=+B3*B4 and in Check column =IF(B9>=5,"OK","NO")
10	$n(1-p) > 5?$	1,550	OK	=+B3*(1-B4) and in Check column =IF(B10>=5,"OK","NO")
11				
12	CALCULATIONS			
13	α	0.05		=1-B5
14	$\alpha/2$	0.025		=B13/2
15	σ	0.009337425		=SQRT(B4*(1-B4)/B3)
16	critical z	1.9600		=ABS(NORM.S.INV(0.025))
17	half the confidence interval	0.0183		=(B\$15*B\$16)
18	lower 95% confidence limit	0.2067		=B\$4-B\$17
19	upper 95% confidence limit	0.2433		=B\$4+B17
20				
21	POST-CALCULATION CHECKS			
22	proportion below lower limit	0.025	OK	=NORM.DIST(+B\$17,B\$4,B\$15,1) and in Check column =IF(B21=B\$14,"OK","NO")
23	proportion above lower limit	0.025	OK	=1-NORM.DIST(+B\$18,B\$4,B\$15,1) and in Check column =IF(B22=B\$14,"OK","NO")

The **CHECKS** sections verify that the results make sense. It's always worth checking your formulas the first time you use them by using information you know. In this case, the formulas in the **PRELIMINARY CHECKS** compute the guidelines and verify that they are within specification. **POST-CALCULATION CHECKS** go backward from the computed confidence limits to verify that the proportion of the curve below the lower confidence limit and above the upper confidence limit match $\alpha/2$. As it should be, the proportions are equal to $\alpha/2$; they also add up to 1.000, as they must.

The method discussed above works as an approximation that is acceptable for proportions that are not very close to zero or to one – something verified by the $np > 5$ and $n(1 - p) > 5$ conditions in the **PRELIMINARY CHECKS** section. Later in the course, you will learn about other methods of determining confidence limits for proportions that don't fit these assumptions.

⁸³ If you ever need to display the formulas and the results in the same sheet, you use F2 to enter EDIT mode, copy the formula, type an apostrophe in the target cell and paste the formula into place right after the apostrophe. It will then be a string rather than an active formula. For example, Cell D3 actually contains '=+B2/B1' and but it does not display the leading apostrophe.

7.23 Conditional Formatting

In Figure 7-21, you may have noticed the green boxes with *OK* in them next to the checks. These boxes are formatted with **CONDITIONAL FORMATTING** in EXCEL 2010. Figure 7-22 shows the drop-down menu for **CONDITIONAL FORMATTING**.

Conditional formatting determines the appearance of a cell according to a wide range of possible conditions or rules. There are a great many options in **CONDITIONAL FORMATTING**, but Figure 7-23 demonstrates how a simple formula can warn a user visually that something is wrong.

In this example, the check-result cells (C8 through C10, C21 and C22) are initially tinted light green by default with dark green letters. If the contents (defined by an **=IF** statement) are “NO” then the box turns light red and the letters are deep red.

Defining appropriate conditional formatting, especially for a spreadsheet that you plan to use often or that you are putting into production for other people to use, can instantly warn a user of an error. Although the colors are helpful, even a color-blind user can see an appropriate message (e.g., “NO” or “BAD” or “ERROR”) to signal something wrong.

The EXCEL 2010 **HELP** function has a number of useful articles about *conditional formatting* that you can access by entering that term in the search box.

Figure 7-22. Accessing Conditional Formatting for an existing rule.

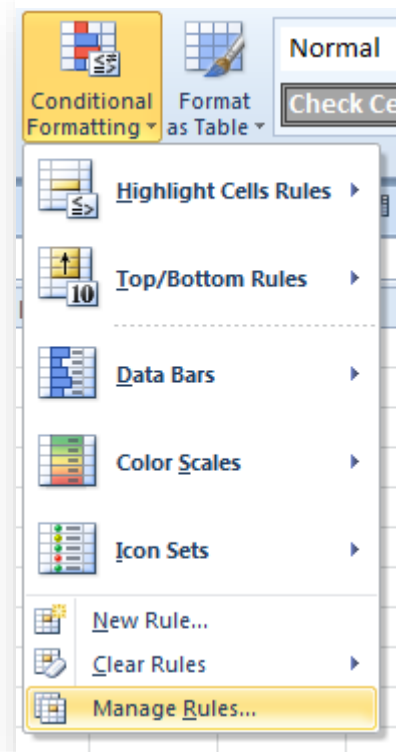
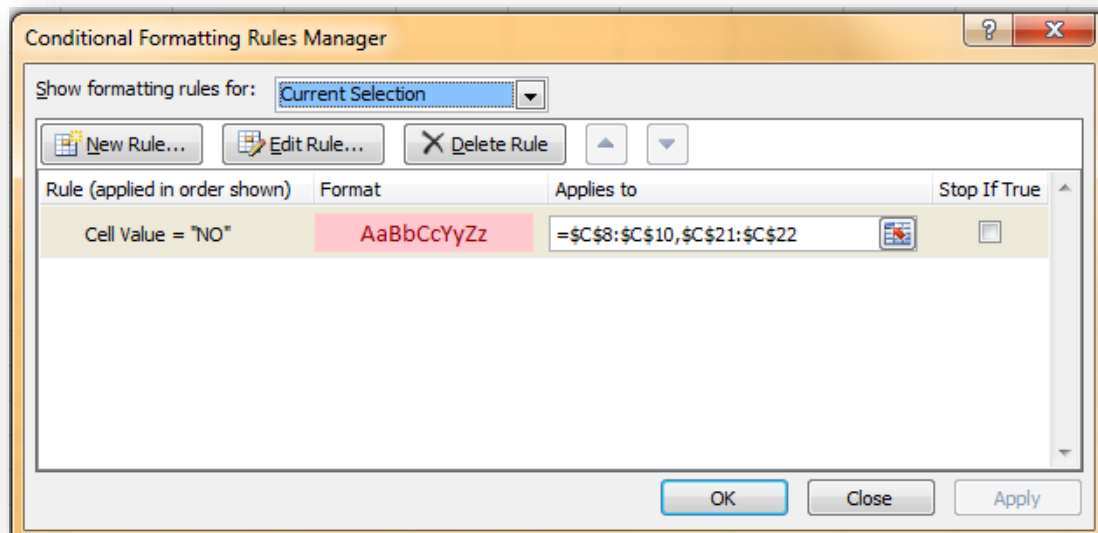


Figure 7-23. Definition of rule for Check cells in example.



7.24 Confidence Limits for Population Variance and Population Standard Deviation Based on Sample Variability

Erewham Natural Foods Corporation are sampling grandiloquent beetle carapaces to make a medicinal grandiloquent beetle carapace extract (GBCE™) highly popular among outworld populations such as the Drazeeli and the Q'ornopiads. The company is deeply concerned about the quality control in its Io plant because deviation from its contractual obligation can result in executive decapitation (on Drazeel) and slow conversion into compost (on Q'ornopia). For the Erewham plant to pass ISO (Interplanetary Standards Organization) 9000 standards, it must monitor the standard deviation of its 1000 gram bottles and start investigating the production line when the standard deviation of the production exceeds a parametric standard deviation of 2 grams.

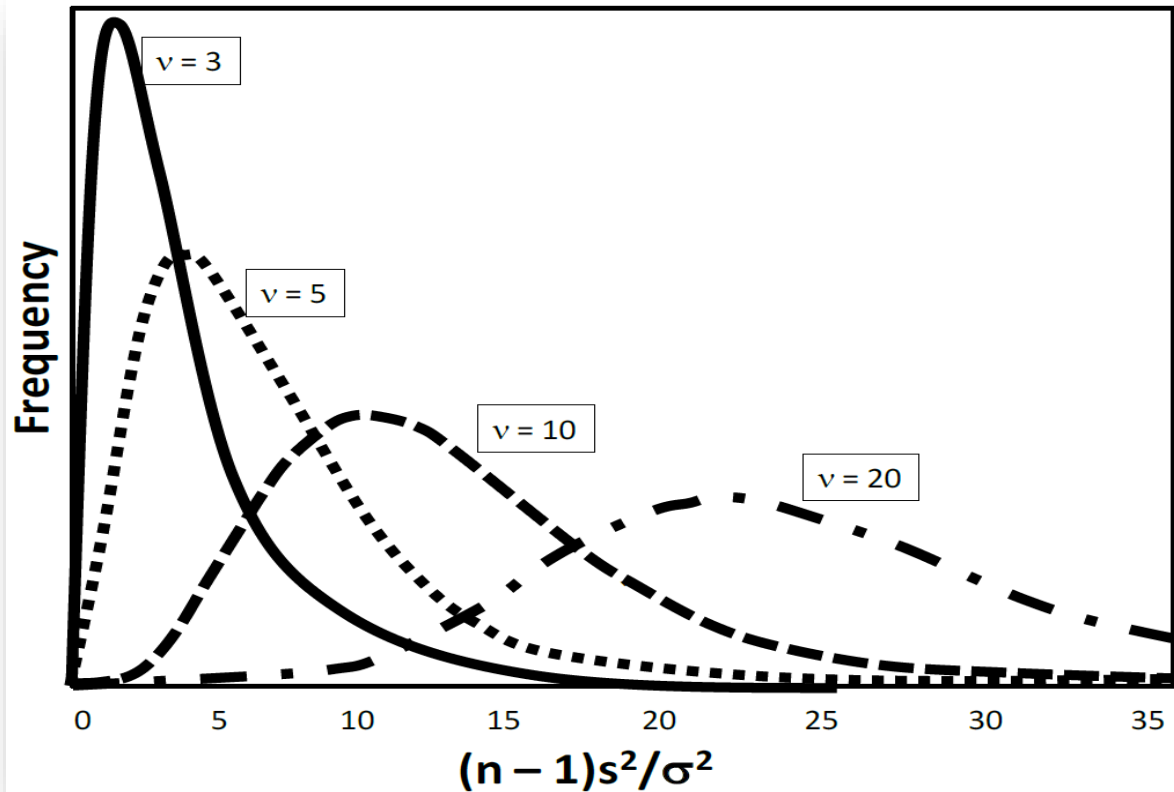
- For extra care in maintaining its corporate reputation, every day the plant managers compute 99% confidence limits for the standard deviation
- Using a random sample of 25 bottles of GBCE™ selected from output using random numbers.
- On a particular day in October 2281, the sample comes back with a standard deviation of 0.8 gm.
- What are the 99% confidence limits for the standard deviation that day?

Variances and standard deviations are *not* normally distributed. Instead, the quantity

$$\chi = (n - 1)s^2/\sigma^2 = \nu s^2/\sigma^2$$

(where n is the sample size, s^2 is the sample variance, and σ^2 is the parametric variance) is distributed according to a theoretical distribution called the *chi-square* (χ^2) with $\nu = (n - 1)$ degrees of freedom. Several of these distributions are shown in Figure 7-24 with different degrees of freedom.

Figure 7-24. Chi-square distributions with different degrees of freedom.

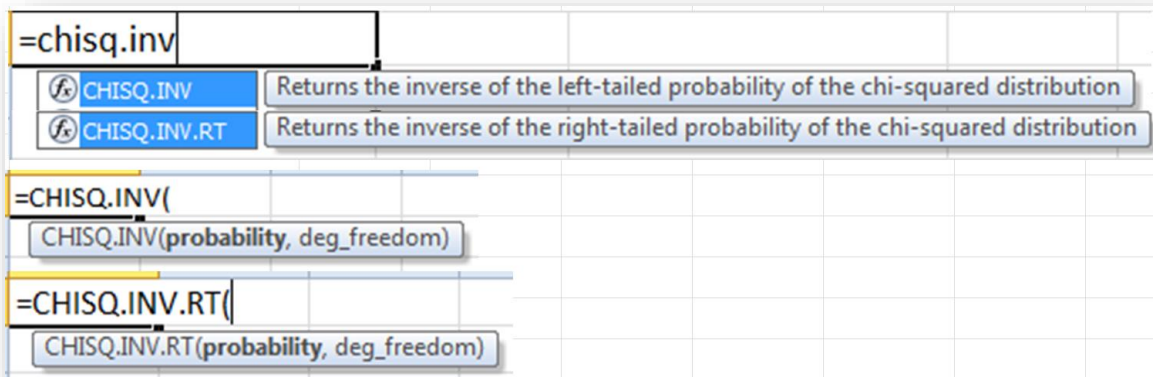


By convention, we use the notation $\chi^2_{\alpha[v]}$ to designate the critical value of the χ^2 distribution with v degrees of freedom for which the probability of sampling a chi-square variable x greater than critical value is α ; i.e., by definition

$$P\{x \geq \chi^2_{\alpha[v]}\} = \alpha$$

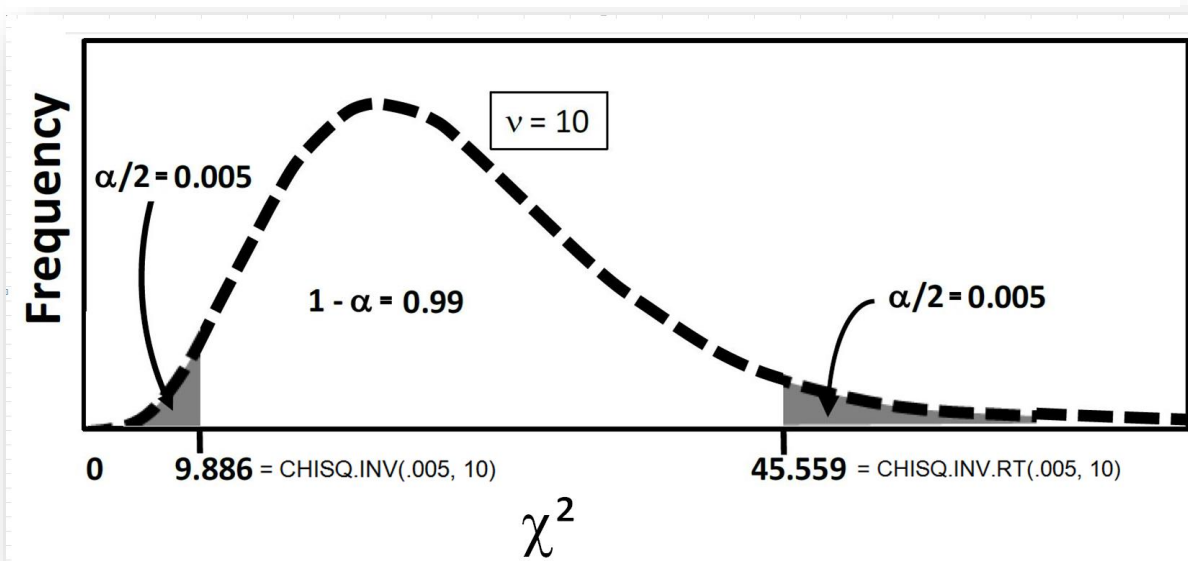
There are two chi-square inverse functions – one for each tail – in EXCEL 2010 that can provide the critical values needed for computation of confidence limits to a variance. Figure 7-25 shows a compound screenshot of these variables in EXCEL:

Figure 7-25. Excel 2010 popups for chi-square inverse functions.



- The EXCEL function =CHISQ.INV(x, deg_freedom) gives the *left-tail* critical value $x = \chi^2_{\alpha[v]}$ of the χ^2 distribution with $df = v$. For example, the critical value $\chi^2_{.005[10]} = \text{CHISQ.INV}(.005, 10) = 9.886$
- The EXCEL function =CHISQ.INV.RT(x, deg_freedom) generates the right-tail critical value corresponding to the value x ; thus $= \text{CHISQ.INV.RT}(.005, 10) = 45.559$
- Thus the $(1 - \alpha) = 0.99$ confidence limits imply $\alpha/2 = 0.005$ in each of the left and right tails of the distribution, and the $(1 - \alpha) = 0.99$ in the area in the middle, as shown in Figure 7-26.

Figure 7-26. Critical values of the chi-square distribution for computation of confidence limits.



To compute confidence limits to the parametric variance based on a sample variance, we need to expand our earlier definition of the critical value.⁸⁴ We know that for the $(1 - \alpha)$ confidence limits, we can start with

$$P\{ \text{CHISQ.INV}(\alpha/2, v) \leq x \leq \text{CHISQ.INV.RT}(\alpha/2, v) \} = 1 - \alpha$$

Substituting the meaning of x ,

$$P\{ \text{CHISQ.INV}(\alpha/2, v) \leq v s^2 / \sigma^2 \leq \text{CHISQ.INV.RT}(\alpha/2, v) \} = 1 - \alpha$$

$$P\{ \text{CHISQ.INV}(\alpha/2, v) / v s^2 \leq 1 / \sigma^2 \leq \text{CHISQ.INV.RT}(\alpha/2, v) / v s^2 \} = 1 - \alpha$$

Therefore⁸⁵

$$P\{ v s^2 / \text{CHISQ.INV}(\alpha/2, v) \geq \sigma^2 \geq v s^2 / \text{CHISQ.INV.RT}(\alpha/2, v) \} = 1 - \alpha$$

Or, putting the expression back in the normal order,

$$P\{ v s^2 / \text{CHISQ.INV.RT}(\alpha/2, v) \leq \sigma^2 \leq v s^2 / \text{CHISQ.INV}(\alpha/2, v) \} = 1 - \alpha$$

This is the general form for the $(1 - \alpha)$ confidence limits to the variance given a sample variance, with the lower limit (L_1) using the *right*-tail critical value (because it's larger and so the quotient is smaller) and the upper limit (L_2) using the *left*-tail critical value (because it's smaller and so the quotient is bigger).

Thus the confidence limits for the parametric variance based on the sample variance are computed as shown below using the functions in EXCEL 2010:

$$L_1 = v s^2 / \text{CHISQ.INV.RT}(\alpha/2, v)$$

$$L_2 = v s^2 / \text{CHISQ.INV}(\alpha/2, v)$$

The confidence limits for the *standard deviation* are the *square roots of the limits of the variance*.⁸⁶

INSTANT TEST P 7-31

A security engineer is trying to establish confidence limits for the variance and of the standard deviation for the number of input-output (I/O) operations per minute on a particular network storage unit under normal load.

The sample size is 10,000 and the observed sample standard deviation is 26.8.

Demonstrate that the lower and upper 95% confidence limits for the variance are 699 and 739; the lower and upper 95% confidence limits for the standard deviation are 26.4 and 27.2. At the 95% level of confidence, why is an observed variance of 600 *unlikely*?

⁸⁴ You are not expected to be able to derive these formulas from memory. They are presented to help all students understand the logic behind the computations and to support those students with an interest in the mathematical underpinnings of statistics. Courses in statistics offered in mathematics departments in universities usually include derivations of this kind for all computational formulas.

⁸⁵ Note the reversal of direction: if $2 < x < 3$ then $1/2 > 1/x > 1/3$.

⁸⁶ Neither a variance nor a standard deviation may be negative.

To illustrate these computations, we can return to the question of quality control of GBCETTM at the Erewham Company production line introduced at the start of this section. The Erewham data produce the following values using some simple EXCEL calculations shown in Figure 7-27.

Figure 7-27. Excel 2010 calculations of confidence limits for the variance and the standard deviation based on sample data.

	A	B	C	D	E	F
1	Data & Calculations			Symbol	Value	Formula
2	Sample size			n	25	
3	Degrees of freedom			v	24	
4	Standard deviation of sample			s	0.8	
5	Variance of sample			s ²	0.64	=E4^2
6	Numerator for limits			(n-1)s ²	15.36	=E3*E5
7	Confidence level			(1 - α)	0.990	
8	Left-tail probability			α/2	0.005	=(1-E7)/2
9	Right-tail probability			1 - α/2	0.995	=1-E8
10	Chi-square for lower limit				9.886	=CHISQ.INV(E8,E3)
11	Chi-square for upper limit				45.559	=CHISQ.INV.RT(E8,E3)
12	Confidence limits					
13	For Variance					
14			Lower		0.337	=E6/E11
15			Upper		1.554	=E6/E10
16	For standard deviation					
17			Lower		0.581	=SQRT(E14)
18			Upper		1.246	=SQRT(E15)
19						
20	CHECKING:					
21	Left tail for left-tail critical value				0.005	=CHISQ.DIST(E10,E3,1)
22	Right tail for right-tail critical value				0.005	=CHISQ.DIST.RT(E11,E3)

The 99% confidence limits for the standard deviation are 0.581 to 1.246 grams⁸⁷ but the maximum acceptable standard deviation with 99% confidence is much higher, at 2 grams. There's no reason to worry about decapitation or conversion into compost today!

Quality-control (QC) charts typically mark the upper and lower limits to the statistics being monitored and graph the periodic measures; in this case, $L_2(\sigma)$ of 1.246 grams would be below the maximum allowable standard deviation of 2 grams. Production goes on and the Solar System will continue to benefit from grandiloquent beetle carapace extract (GBCETTM).

⁸⁷ A quick note about asymmetric confidence limits: the midpoint of 1.246 and 0.581 is 0.932, which is *not* the observed standard deviation of 0.8; the same asymmetry affects the variance limits, where the midpoint of the limits is 0.946, also not the observed variance of 0.64. The reason is that the underlying probability distribution (χ^2) is asymmetrical. Don't expect all confidence limits to be symmetrical around the sample value!