CHAPTER 14

SAMPLING FROM POPULATIONS

Up to this point, we have mainly been interested in providing some description of one or more data sets such as test scores or heights. Generally speaking, if all you are interested in doing is describing a set of data, then we would refer to the analysis of that data as being **DESCRIPTIVE STATISTICS**. For example, in a school, the principal circulates a short survey to 5th and 6th graders concerning their attitudes about certain activities and has a need to summarize the data. Of major priority is what these children think. Perhaps the children's views will impact on what the principal decides to do in the future at this school.

However, in many situations, the primary focus is not on what that particular group thinks (on whom you gather the data). For example, during election years, pollsters survey opinions of small groups of potential voters about their preferences for one candidate or another, or for one side of an issue or another. In these cases, the real concern is not in what this particular group feels but rather how their opinions reflect the views of the larger population of potential voters. In these cases, what the sample feels is taken as a good indication of what the larger population feels. But, how well the sample reflects what the larger population feels hinges to a large extent on how representative the sample is of the entire population. Therefore, the present Chapter examines the concept of sampling and some terminology related to that, and describes several ways in which samples can be taken, some good and one bad. **I CANNOT OVEREMPHASIZE HOW IMPORTANT THE PROCESS OF SAMPLING IS TO THE PROCESS OF MAKING REASONABLE INFERENCES FROM SAMPLE DATA TO THE LARGER POPULATION**. If one "messes up" on the selection of the sample, then the entire project or study may be highly suspect. While there is not much room in this text to go into sampling in much detail, the topic of sampling is very important.

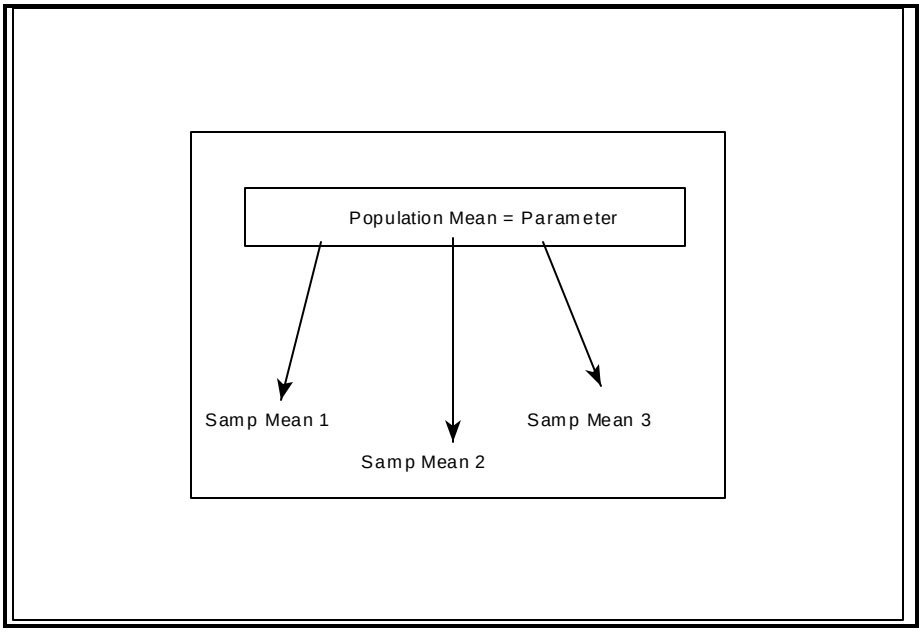
Populations and Samples

First we need to clarify the distinction between populations and samples. The **POPULATION** represents the entire collection of objects or people to which you want to generalize your findings or results. Some possible examples of populations could be: all tenth grade students in Pennsylvania, all General Motors cars, or all Star Kist tuna fish cans produced during 1990. On the other hand, some possible samples drawn from these populations could be: tenth grade students in Altoona, 1989 Oldsmobile Cierras, and Star Kist tuna cans in the Weis food markets in State College, Pennsylvania. **SAMPLES** are subsets of the population. Samples, obviously, could be relatively small or relatively large. For example, the 1989 Oldsmobile Cierras in the parking lot at one local department store in one particular city (which may only be 3 or 4 cars) would be a relatively small sample (out of the population) whereas all the 1989 Oldsmobile Cierras in the state of New York

would represent a rather large sample out of the total population.

When we talk about populations, it may be of interest to consider calculating some numerical value such as the mean or standard deviation on the variable of interest within the population. For example, the actual mean number of miles on all 1989 Oldsmobile Cierras would be an example of finding **THE** numerical value of the characteristic in the overall population. Such a value is called a parameter. A **PARAMETER** is a value calculated on the variable of interest based on every element in the entire population. However, if we take a small sample of cars and calculate the mean number of miles on them as a good estimate of the parameter, then we are dealing with a statistic. A **STATISTIC** is a value based on a sample that is used to estimate what the corresponding parameter is in the population. One way to keep this distinction clear is to think that: **P** is for **POPULATION** and for **PARAMETER**, and **S** is for **SAMPLE** and **STATISTIC**. Note that the use of the term decriptive statistics used above seems to differ somewhat from the current definition of statistic in that descriptive statistics are normally referred to in situations where there is no real interest in generalizing the results. While I realize this small "conflict" in the use of the terminology, that is the way you are likely to see it in the literature.

However, the real chain of events looks as follows when you use some statistic from a sample to estimate a paramter of some population. See below.



As an example, we may be interested in the mean in the population which, recall from Chapter 2, has mu (μ) as its symbol and is pronounced mew. This would be our population parameter. However, if we are interested in what the parameter is, the only realistic way in which we can estimate it is to take one or more samples and to use the

mean(s) of the sample(s) which are usually referred to as X bar's. These would be our statistics. Thus, whereas we are primarily interested in the parameters, we generally have no other choice than to deal with statistics as estimates of the parameter values.

If we use samples and the statistics from those samples to estimate the corresponding parameters, then it should be obvious that the statistic would reflect the parameter fairly well if the sample is representative (and of decent size probably too) of the total population. Thus, the method used to select the sample becomes crucial in the process of making reasonably accurate inferences. Unfortunately, there are variety of methods by which samples can be drawn; some are much better than others. Let's explore several ways that samples can be selected starting with a poor example and moving to better ones.

Accidental or Convenience Sampling

At the bottom of the heap is accidental or convenience sampling. Accidental or convenience sampling is just that; sampling that is done because some collection of objects or people is "handy" to you. For example, what if you are interested in estimating the average height of 11th grade students in the State of Oregon. Being a sports buff, you generally attend the local high school basketball games at your son's school. With the permission of the coach, you quickly measure the heights of the 6 players who happen to be in the 11th grade. You use this sample mean as your "best" estimate of the heights of all 11th graders in the state. What a mistake you will make doing it this way, by convenience, since basketball players tend to be taller than the average student. While it is possible that samples gathered by some accidental or convenient plan could "accidentally" be representative of the entire population, it is doubtful that many would be since there is no systematic plan for the sampling that has been used to make sure the samples are representative. That is the problem: accidental and convenience samples have no plan. Unfortunately, too many samples are collected this way. We all have seen the "on the corner" interviews of people for the evening news broadcast. You know of course that this is not a good way to sample since, usually, they interview many people and then select one or two cases from each side of the story, even if one side of the story occurs rarely in the population. You should be observant when reading reports to make sure that the samples were not gathered in this way and if so, you need to be very conservative in the interpretation of their reported results!

Random Sampling

A second way to sample, and a much better method, would be to take a random sample. **RANDOM SAMPLES** are ones where each person or object you select has an equal chance of being included in the sample. For example, let's say that the population of interest is all car dealerships in Alabama. First, one would have to identify each and every one of these dealerships. Then, one would need to essentially put each one (via a slip of paper) into a hat and then draw out (for example) 20 without seeing which ones you are selecting. In this way, each dealership has essentially the same "luck of the draw" to be

included in the sample. Technically, there are two variations of this plan. Sampling **WITH REPLACEMENT** would be the case if, after each draw, you put that dealership back into the pot. This would maintain the same initial chance of being selected on each and every draw. For example, assume that there were 50 dealerships in the state. On the first draw, the chance is 1 out of 50 of being selected. After being selected, you put that dealership back into the hat and draw the second one. The chance of being selected on the second draw is still 1 out of 50. While it is not very likely that the first one drawn will be drawn a second time, there still remains that chance. Here lies the "rub" in a sampling with replacement plan. An alternative to this would be the more practical approach of sampling **WITHOUT REPLACEMENT** where, after the first draw, you do not replace the dealership back into the pot. In this way, a particular dealership could be included only once. For most practical purposes, selecting a sample without replacement is a very close cousin of true random sampling with replacement and, in general, will provide a good way of selecting a representative sample. The important thing to keep in mind with random sampling with or without replacement is that it is chance that is determining who is included, and not the one who is selecting the sample. If it were entirely up to the data collector, then his or her biases could enter into the plan and produce a sample that is not as representative of the total population as it should be, and may possibly "lean" in a direction that is consistent with what the sampler wants to find.

While random sampling is a very good long run strategy for selecting representative samples, it is not necssarily the best in the short run or, especially if the samples are small. For example, what if you wanted to take a representative sample of 30 faculty members at a college on which to do a survey and you select them using a random process. Is there any guarantee that you will get any full professors (assuming that there are some!)? Not necessarily. It might be the case that in your sampling, not one of the full professors is selected in that the luck of the draw was "against them" in this particular sampling. In the long run, there would be full professors sampled if you take enough samples. What if you wanted to also make sure that some of this sample were female faculty members. Would random sampling necessarily guarantee that? Again, no. In any one random sampling, one may not obtain all the constituent elements that would make for a representative sample.

Stratified Random Sampling

If there are occasional problems with random sampling in that some important population elements are not selected, how can we improve upon simple random sampling so that we are better assured of obtaining representation of all the important elements of the population? The answer is to stratify. **STRATIFICATION** means to subdivide the population into important segments prior to sampling. For example, in the college sampling mentioned above, what if we made a breakdown at the local college in terms of faculty rank and gender. Look at the table below.

Faculty Rank

Inst. Asst. Prof. Assoc. Prof. Prof Total

Male	12	15	20	15	62
Female	6	9	10	6	31
Total	18	24	30	21	93

If you wanted to guarantee inclusion of some at each faculty rank and some from both sexes, then you will have to subdivide the population of 93 faculty members into rank and gender categories before you actually select the sample. In this particular situation, there are some males and females at each faculty rank. Thus, if you randomly selected some from each of the 8 cells (faculty rank by sex), then you would be guaranteed of meeting your predetermined requirement of having some from each rank and from each sex. Thus, stratifying and then randomly sampling from within each stratum will deliberately improve the odds of making sure that your sample is representative of the total population, in the short run. This assumes, of course, that you do in fact understand what the important elements are in the population before sampling so that you can actually stratify the population!

Stratified Random Proportional Sampling

As a final method (and there are many more), which is a variation on the stratified random sampling process, one could also take into account the relative frequency of each category in the total population. For example, in the table above, note that there are about twice as many males in this population as females. Also note that there are more associate professors and fewer instructors. Therefore, one could select according to the approximate proportion that exists in the population. As an illustration, if you wanted a sample of about 30 faculty, then you would sample at random about 20 males and 10 females with the largest number of males coming from the associate professor category (maybe 5 or 6) and fewest from the instructor category (maybe 3 or 4). For females, you would perhaps take 3 or 4 from the associate professor category and only 1 or 2 each from the instructor and full professor categories. Thus, not only would you be guaranteeing that representation will be there from each category but also providing representation at the approximate rate at which it appears in the overall population. A good stratified random proportional sampling plan can be very accurate in representing the population.

As nice as it seems though, stratified random proportional samples are difficult to implement and can be very costly. This is the major drawback to using this technique more often. For example, in nationwide polls, one may want to stratify on several factors such as age, sex, and region. To properly implement such a sampling plan, one will have to travel to several different regions of the country and to make sure that different ages and sexes are included in the samples when one arrives at each region. Such efforts can be very time consuming and expensive. However, one of the major reasons why the major polling companies (like Gallup and Neilsen) are quite accurate is because of the time and energy

expended in developing and implementing the sampling plans. It surely is not because of the size of the samples since most of these polls include less than 1000 persons to represent perhaps the entire nation. Thus, the key to representativeness is the process of sampling, not so much the size of the sample (though size does play some role).

Biased Sampling

As a last note on sampling, one needs to understand that even the best sampling plans can go awry. More generally however, plans that go awry are those that were not very good from the start such as convenience sampling. For example, what if you were interested in the average grade point average of college students at a large university and were unable to convince the registrar to turn over that information to you. Your alternative may be to take several samples of students and, in each, ask students to indicate (confidentially of course!) their grade point average. But, in selecting your several samples, you unfortunately obtain 3 classes that happen to be in "remedial math", although you are not aware of this fact. What impact will this have on your estimation of the grade point average in the entire college? Since students who tend to be in remedial courses are the ones who are having difficulty in school, it is very likely that their grade point averages will be low. Thus, in each of these samples, the average of the grade point averages will be low compared to the overall student population average. This is what we call a **BIASED SAMPLING METHOD** in that systematically, most if not all of the samples will underestimate or overestimate the population mean or the parameter of interest. We would have run into the same problem if the samples we were able to gather had all come from advanced classes in the sciences. In either case, we can seriously "goof" if we use our sample data to estimate the population parameter if the method of sampling is biased. It is not the fact that one particular sample may produce a statistic (sample mean for example) that is much lower or higher than the parameter (population mean), that will happen just by chance. But, it all the samples are too high, or all the sample are too low, then the method of sampling is biased. Specific samples are not biased, but the method of sampling can be.

Final Note on Sampling

So, what is the moral of this "sampling" story? It is simple! If a method is used that will not reasonably insure that the sample is representative of the population to which you want to generalize, then all the care and time and effort that goes into the actual data collection and analysis can be wasted. Bad samples make for a bad study and bad studies or investigations provide data that are difficult to interpret, at best. A word to the wise is to make sure that if you are the one who has to select the sample, then make sure that your sample speaks well for the population to which you want to generalize. Also keep this in mind when reading articles where sampling has been done and inferences are

made to the larger population. Sure, any sample will generalize to some population but will it be the correct population? If there is reasonable doubt in the reported study about the quality of the sampling methods, then take the interpreted results with a large grain of caution!

Practice Problems

1. Distinguish between populations and samples, and parameters and statistics.

2. Describe how you could obtain a representative sample of electricians if the population is all electricians in the state of Virginia.

3. What are the differences among accidental, random, stratified random, and stratified random proportional sampling methods?

4. What is a biased sampling method and how can it be avoided?

CHAPTER 15

SPECIAL STATISTICAL DISTRIBUTIONS

Chapter 14 discussed sampling and its role in the inferential process. The next Chapter will introduce the concept of sampling error and factors related to the amount of sampling error, particularly as it relates to sampling error of means. However, for the moment, we need to introduce (and in one case refresh your memory) several distributions that play key roles in inferential statistics. Our discussion will not exhaust all the distributions that can be used but, will concentrate on the most popular. For purposes here, it will be sufficient to examine the normal, t, chi square and F distributions and describe what each of these look like. In addition, the process of finding baseline values for several different percentile ranks will also be illustrated. We will use each of these distributions later when discussing different methods of making inferences about different parameters of interest such as the population mean or the population variance.

Normal Distributions

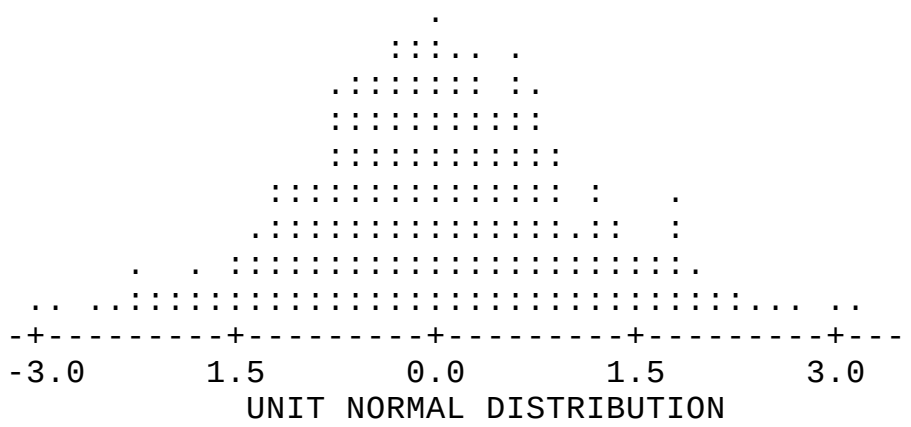
The first distribution should not be any stranger to you because it is the normal distribution. Chapter 6 gave a rather long presentation of the normal distribution and its characteristics. The normal distribution is widely used in descriptive statistics since the model of a normal distribution resembles many sets of real data, especially large data sets. In addition, the normal distribution forms the basis for many inferential procedures and that is why it is important to discuss it first.

As a simulation, I randomly generated 1000 values that came from a normal distribution where the mean is 0 and the standard deviation is 1. There is a command in Minitab called **RANDOM** that easily allows one to do this. This sounds like a typical z score distribution from descriptive statistics and is called the **UNIT NORMAL DISTRIBUTION**. See the following use of the Minitab command. Also note that the **RANDOM** command specifies how many observations Minitab is to generate and the column number where the data are to be placed. There is also a subcommand that specifies the type of distribution from which to sample, in this case, **NORMAL** with mean = **0** and standard deviation = **1**.

Note that the simulated distribution is not exactly smooth (which the real normal distribution would be) but the descriptive statistics come close to what you would expect with a mean close to 0 and a standard deviation close to 1. To refresh your memory, recall that in normal distributions, you can mark off about 3 standard deviation units on either side of the mean and there is little remaining area in a normal distribution (although technically, it goes on forever) after reaching z values of -3 or 3.

MTB > **rand 1000 c1;** **<--- Minitab command line**
SUBC> **norm 0 1.** **<--- Subcommand line**

Now look at a simple dotplot of the data.



DESCRIPTIVE STATISTICS ON GENERATED UNIT NORMAL DISTRIBUTION

	N	MEAN	MEDIAN	STDEV
UnitNorm	1000	0.0199	0.0095	0.9857

What if you wanted to find the percentile ranks for various baseline values in a unit normal distribution; ie, where the mean is 0 and the standard deviation is 1? We could use a normal curve table that shows areas in various segments of the normal distribution. Remember, approximately 34% of the total area in the normal distribution is from the mean out to either a z of -1 or 1, approximately 14% is from a z of -1 to -2, or 1 to 2, and approximately 2% remains from a z of -2 to -3, or 2 to 3. The z score and percentile rank scales would look as follows.

Area		2	14	34	34	14	2
	|)))))|)))))|)))))|)))))|)))))|)))))|						
z score	-3	-2	-1	0	+1	+2	+3
PR	<1	2	16	50	84	98	>99

As long as you are working with nice whole number z scores, then finding the approximate percentile rank is relatively easy. However, if you want to find the percentile rank for a z score of, say -2.5 or +1.5, then all we could do is to "guesstimate" from the chart above. For example, for a z of -2.5, the percentile rank is somwhere between <1 and 2, but what is it exactly? Or, for 1.5, it is between 84 and 98 but, again, what is it exactly? Using an area under the normal curve table (shown in Chapter 6) can help you solve this but Minitab will also do this for you easily. Recall I mentioned that a command cdf would find percentile ranks given various baseline values. Assume that you want to find the

percentile ranks for the z values of -3, -2.5, -2, etc. up to 3. If we set these values into a column, we can find the percentile ranks using the cdf command. See the following.

z Score	Area(%)
-3.0000	.13
-2.5000	.62
-2.0000	2.28
-1.5000	6.68
-1.0000	15.87
-0.5000	30.85
0.0000	50.00
0.5000	69.15
1.0000	84.13
1.5000	93.32
2.0000	97.72
2.5000	99.38
3.0000	99.87

Note that the percentile ranks are listed as decimals and you would have to move the decimal place to the right two places. For example, the z score of -2.5 has less than 1 percent (actually .62 of 1 percent) whereas the z score of 1.5 has a percentile rank of about 93.

It is also possible to find the baseline score value for a particular percentile rank value. For example, what if you want to know the z values that correspond to the percentile rank values of 5, 10, etc. up to 95. Again, Minintab can help with the command invcdf. See the following data table.

PR Value	z	PR Value	z
5.00	-1.6449	55.00	0.1257
10.00	-1.2816	60.00	0.2533
15.00	-1.0364	65.00	0.3853
20.00	-0.8416	70.00	0.5244
25 00	-0.6745	75.00	0.6745
30.00	-0.5244	80.00	0.8416
35.00	-0.3853	85.00	1.0364
40.00	-0.2533	90.00	1.2816
45.00	-0.1257	95.00	1.6449
50.00	0.0000		

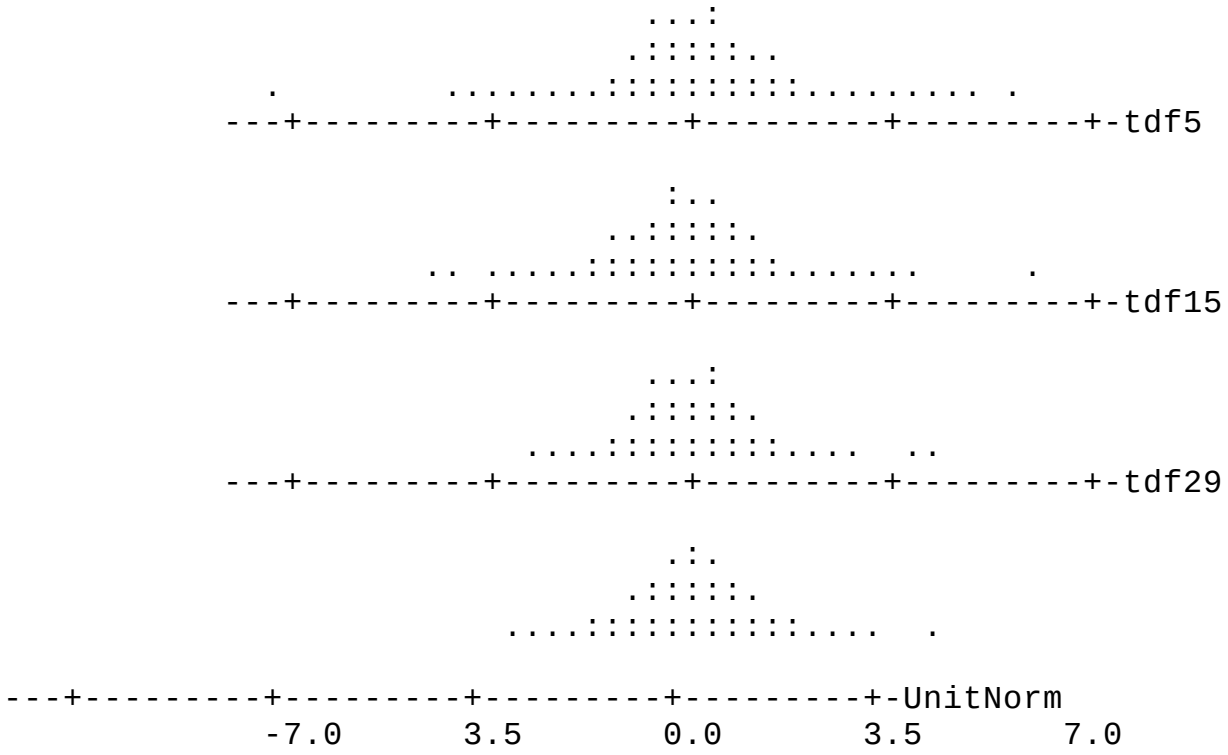
For example, the 30th percentile rank would be a z score of approximately -.52, whereas the 80th percentile rank would be a z score of about .84.

This very brief review of the normal distribution was mainly designed to bring in the term, unit normal distribution. Unit normal distributions look like normal distributions using the z score baseline as the score values. For more information on normal distributions, refer back to Chapter 6. We now turn our attention to a new statistical distribution: t.

t Distributions

A second distribution that is popular in inferential statistical work is called the t distribution (small t, not cap T mentioned as a position measure in Chapter 5). The determining factor that drives the looks of t distributions is a concept called degrees of freedom. Recall from descriptive statistics when you calculate an estimate of the population standard deviation, you divide by n - 1, which is one less than the sample size. Degrees of freedom in many cases is this value of n - 1. So, if n = 34, then df = 33. Look at several t distribution simulations below. Degrees of freedom values used to generate the t distributions were 5, 15, and 29. I also simulated the unit normal distribution.

Notice that all of the t distributions look symmetrical and uni-modal and resemble the normal distribution which is the last of the 4 distributions. Each distribution seems to center around 0 and the widths are about the same although the lower degrees of freedom t distributions seem to be a little wider.



DESCRIPTIVE STATISTICS ON t AND UNIT NORMAL DISTRIBUTIONS

	MEAN	STDEV

tdf5	0.0290	1.2383
tdf15	-0.0471	1.1077
tdf29	0.0099	0.9827
UnitNorm	0.0259	0.9964

The descriptive statistics show that the means are all close to 0 and the standard deviations are close to 1. Actually, what happens is that the standard deviations of the t distributions approach being closer to 1 as the degrees of freedom values increase. You can notice in the t graphs above that the variability seems to get narrower as you move down through the distributions. In comparison, t distributions and unit normal distributions look very much the same in that they are all uni-modal and symmetrical around 0. The only real difference is that t distributions tend to be a little wider, particularly with small degrees of freedom, and hence have somewhat larger standard deviations than 1. However, even these variability differences evaporate if the sample sizes are relatively large (that is, df values are large).

What if you wanted to compare what the baseline t values would be (as opposed to z values in a unit normal distribution) for one specific percentile rank in several t distributions that varied in their degrees of freedom? Again, using the invcdf command in Minitab, we can easily do that. See the following.

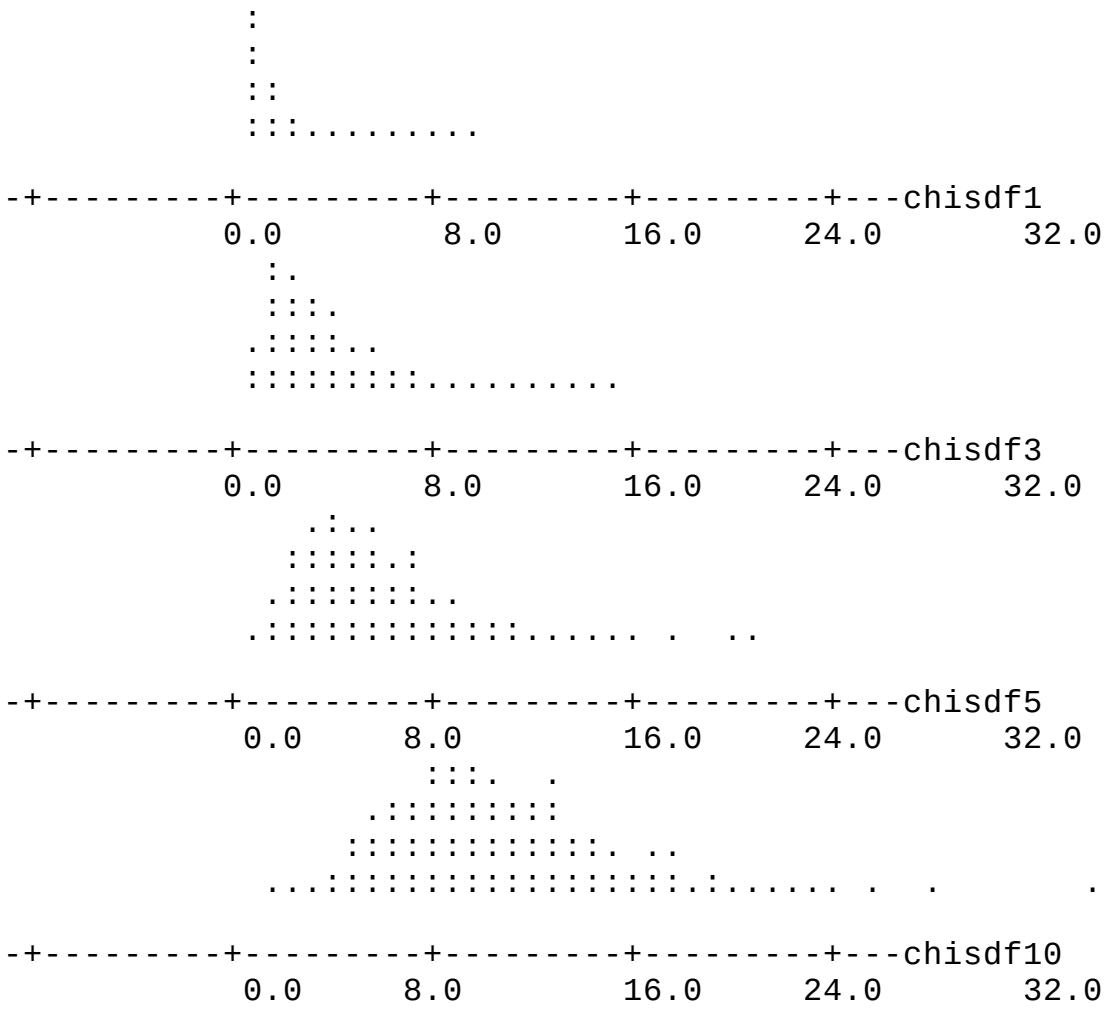
As one example, focus on the percentile rank of 97.5.

tdf	PR	Baseline value
5	97.5	t = 2.5706
15	97.5	t = 2.1315
29	97.5	t = 2.0453
100	97.5	t = 1.9840
UnitNorm	97.5	z = 1.9600

Notice that in t distributions, you have to go out from the mean to the right of 0 further to arrive at the point where there is 97.5 percent of the distribution below. Another way to think of this is to say that you would have to go further from the mean to capture 47.5 percent of the area in the t distribution, compared to the unit normal distribution. But, as you can see from the table above, increases in sample sizes (degrees of freedom) rapidly diminish this impact. For all intents and purposes, baseline values in t distributions and z scores in unit normal distributions are very similar unless the degrees of freedom (sample sizes) are very small. Despite the general similarity of t distributions to unit normal distributions, later we will use t values rather than z scores from the unit normal distribution since using t values will produce (in one inferential application) a model for the data that produces better results. More on that later.

Chi Square Distributions

The third distribution to be discussed is called the chi square distribution (kie like pie). Again, like the t distribution, chi square distributions are determined by degrees of freedom. However, degrees of freedom in many chi square applications are not necessarily related to sample size and how you arrive at them will have to await later discussion. Just as before, Minitab can generate examples of these. See the following simulated chi square distributions.



Unlike the unit normal distribution and t distributions, chi square distributions tend to be positively skewed; in fact, with small degrees of freedom values like 1 or 3, quite radically positively skewed. With larger degrees of freedom values however, the distribution seems to spread out more along the right side of the baseline and it becomes much more similar to the shape of the normal distribution. Look at the descriptive statistics below on the data simulated for these 4 distributions.

	MEAN	STDEV	MIN	MAX
chisdf1	0.9746	1.3701	0.0000	9.0117

chisdf3	3.0698	2.5938	0.0015	14.7810
chisdf5	5.0763	3.1080	0.2014	19.8849
chisdf10	9.8490	4.2690	0.9550	33.3920

Notice that the means of these chi square distributions appear to be very similar to the degrees of freedom for each chi square distribution. In fact, the long run theoretical values for the means of chi square distributions are equal to the degrees of freedom. Hence, if df = 5, then the mean of the theoretical chi square distribution is 5. Also note that the variabilities, as reflected by the standard deviations, get considerably larger as the degrees of freedom values increase. This should have been obvious from the dotplots above in that the larger df valued chi square distributions got wider and wider. If the mean value gets larger and larger, then the distributions have to spread out much further to accomodate the larger mean. However, note that the minimum values are all close to 0. Thus, while the distributions clearly stretch out to the right, they seem to have an anchor at 0 on the left side. Seems like the "rubber band" principle, right? So, in capsule form, what do chi square distributions look like?

1. The shapes of chi square distributions go from being radically positively skewed when the df value is small to being more symmetrical (looking more like a normal distribution) when the df value is large.

2. The means of chi square distributions are equal to the degrees of freedom values.

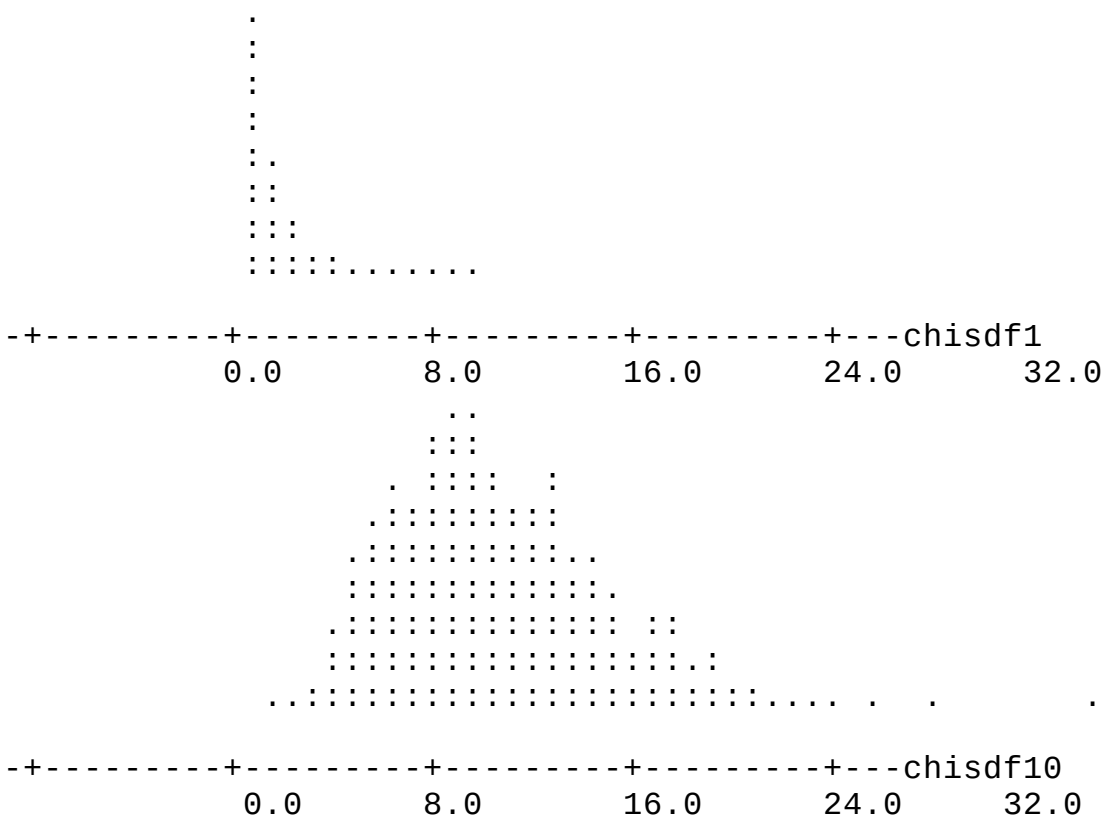
3. The standard deviations of chi square distributions get larger and larger as the df value increases.

4. And finally, note that all chi square distributions have positive values on the baseline; in fact, the point furthermost to the left is 0. No negative values are possible in chi square distributions!

What if we wanted to compare the baseline values of chi square at a given percentile rank in distributions with different degrees of freedom? To make this easier to see, look at the two dotplots on the next page for chi square distributions with 1 and 10 degrees of freedom respectively.

Looking quickly at the first dotplot, it appears that the frequencies basically run out a little past a baseline value of 8. Obviously then, a value of 8 along the baseline in a chi square distribution with 1 degree of freedom will have a very high percentile rank: 90 +, since it is at the very top of the distribution. However, in the distribution with 10 degrees of freedom, a value of 8 does not even seem to move you half way into the frequency distribution. Therefore, since there is more of the distributon above a value of 8, the value of 8 must have a lower percentile rank, perhaps even less than 50. After the dotplots, look at the table that compares the percentile ranks for chi square distributions with 1 and 10

degrees of freedom.

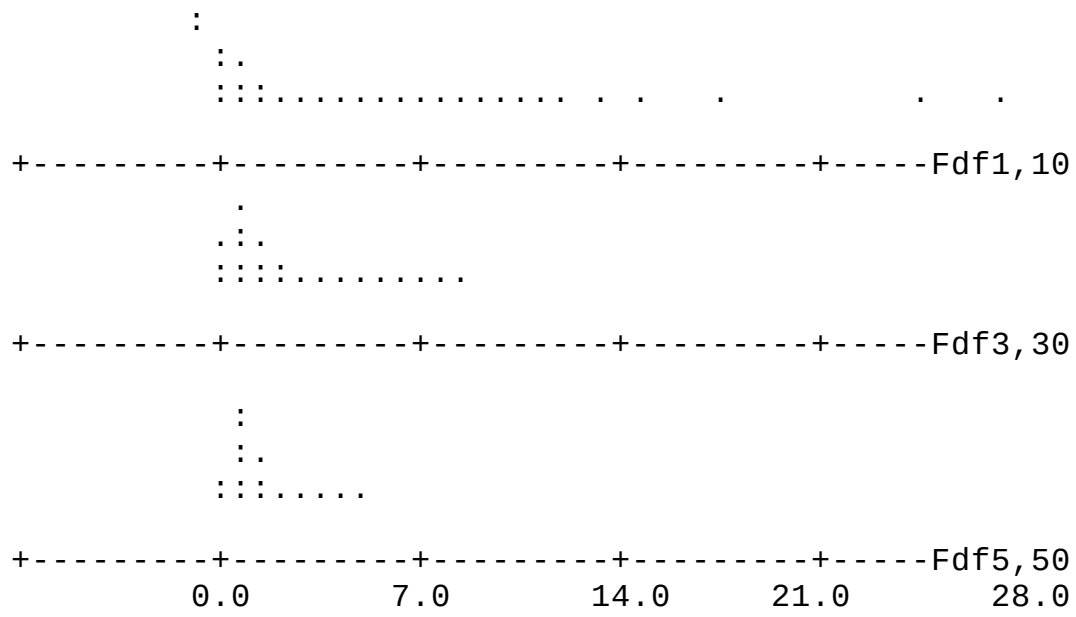


Chis 1 df		Chis 10 df	
PR	Value	PR	Value
50.00	0.4549	50.00	9.3418
75.00	1.3233	75.00	12.5489
95.00	3.8415	95.00	18.3070
99.00	6.6349	99.00	23.2093

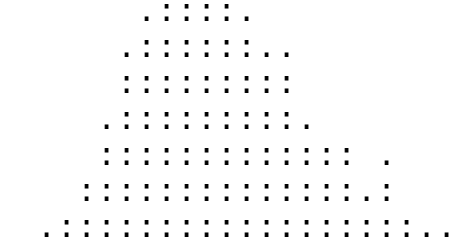
Note, for example, that you arrive at the 95th percentile rank at a baseline value of 3.84 when df = 1, but you must go all the way out to 18.3 to accumulate 95 percent of the area when df = 10. The fact that the same percentile rank can and will yield radically different baseline values depending on df will become important later.

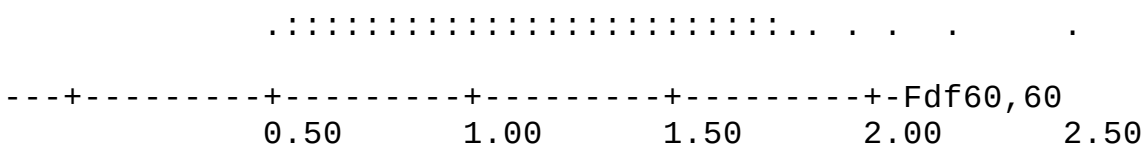
F Distributions

The last distribution to be examined is called the F distribution, named after Sir R. A. Fisher. One thing that is different about F distributions compared to t and chi square distributions is the fact that two degrees of freedom values are needed. For example, F distributions can have 2 and 10, or 3 and 30 degrees of freedom, as just two possibilities. Look at the following simulations for several F distributions.



Note that these 3 F distributions (1 and 10, 3 and 30, 5 and 50) are radically positively skewed, with the far left end being 0. Again, F values cannot be negative. Most situations where F distributions are used, the distributions are decidedly positively skewed (similar to low df chi square distributions), though not always. For example, when the two degrees of freedom values are rather large, the F distribution can look more like a symmetrical distribution that looks similar to the normal distribution. See the following.





In looking at the descriptive statistics on the generated F distributions, note that the means of F distributions all appear to be close to 1. This is a general characteristic of F distributions. But, note that the variabilities, as reflected by the standard deviations, can be quite different. For example (and note that F values cannot be negative, 0 is the minimum), with 1 and 10 degrees of freedom, the range of baseline F values goes from about 0 up to nearly 28 (the standard deviation is about 2.5). However, when df = 60 and 60, the range is severely restricted going from about .5 up to about 2.5 (with a standard deviation of about .3). Thus, while the center points in F distributions tend to be about the same (a value of 1 which is the approximate mean), deviations around 1 can put you in a different relative position (percentile rank value).

	MEAN	STDEV	MIN	MAX
Fdf1,10	1.3797	2.4039	0.0000	27.5451
Fdf3,30	1.0689	1.0209	0.0077	8.4245
Fdf5,50	1.0132	0.6876	0.0196	5.1056
Fdf60,60	1.0265	0.2722	0.4744	2.4466

To see how percentile ranks can be impacted by the degrees of freedom of the F distribution, look at the comparison of the percentile ranks across between the F distributions with 3 and 30, and 5 and 50 degrees of freedom.

F 3 and 30		F 5 and 50	
PR	Value	PR	Value
50.00	0.8069	50.00	0.8822
75.00	1.4426	75.00	1.3739
95.00	2.9223	95.00	2.4004
99.00	4.5098	99.00	3.4077

For example, the 99th percentile rank value in an F distribution with 3 and 30 degrees of freedom is about 4.5 whereas that same position in an F distribution with 5 and 50 degrees of freedom is about 3.4. Again, the fact that the same percentile rank will have different baseline F values will be important later when we use the F distribution for some inferential statistical applications.

Statistical Tables

Most statistics or data analysis books will include many distributional tables in the Appendices. For the most part, I have departed from that tradition since, software packages like Minitab will read these F or t values for you using the cdf or invcdf command. Other packages will do this in similar ways. Thus, since you are to be encouraged to use a package to do most of your work, it is easy to look up needed values.

Final Notes on Statistical Distributions

As a final summary, the 4 distributions we have reviewed are the normal, t, chi square and F. Some overall observations are as follows.

1. t distributions look very much like unit normal distributions except for the fact that t distributions tend to be somewhat wider with low df values.

2. Chi square distributions with low degrees of freedom values and most F distributions look similar in that they are seriously positively skewed. However, both chi square and F distributions can look more symmetrical under certain df conditions.

3. Especially for chi square and F distributions, the same percentile ranks will not produce the same baseline chi square and F values.

Useful Minitab Commands

CDF INVCDF RANDOM

Practice Problems

1. Find the baseline z, t, chi square and F values for the following percentile ranks: 3, 20, 85 and 90.

   A. Unit normal
   B. t with 25 degrees of freedom
   C. Chi square distribution with 8 degrees of freedom
   D. F distribution with 2 and 20 degrees of freedom

2. Find the percentile ranks for the following baseline values.

   A. Unit normal z values of -1.6 and 2.2
   B. t values of -2 and 1 with df = 20
   C. Chi square values of 1 and 5 with df = 4
   D. F values of .8 and 2 with df = 2 and 15