Here is annotated output from the assignment at http://core.ecu.edu/psyc/wuenschk/SAS/Monte-SAS.htm

One Sample of 100,000 scores from Normal Distribution

N	100000	Sum Weights	100000
Mean	100.047226	Sum Observations	10004722.6
Std Deviation	15.0165009	Variance	225.495299
Skewness	-0.0022158	Kurtosis	-0.0242482
Uncorrected SS	1023494044	Corrected SS	22549304.4
Coeff Variation	15.0094126	Std Error Mean	0.04748635

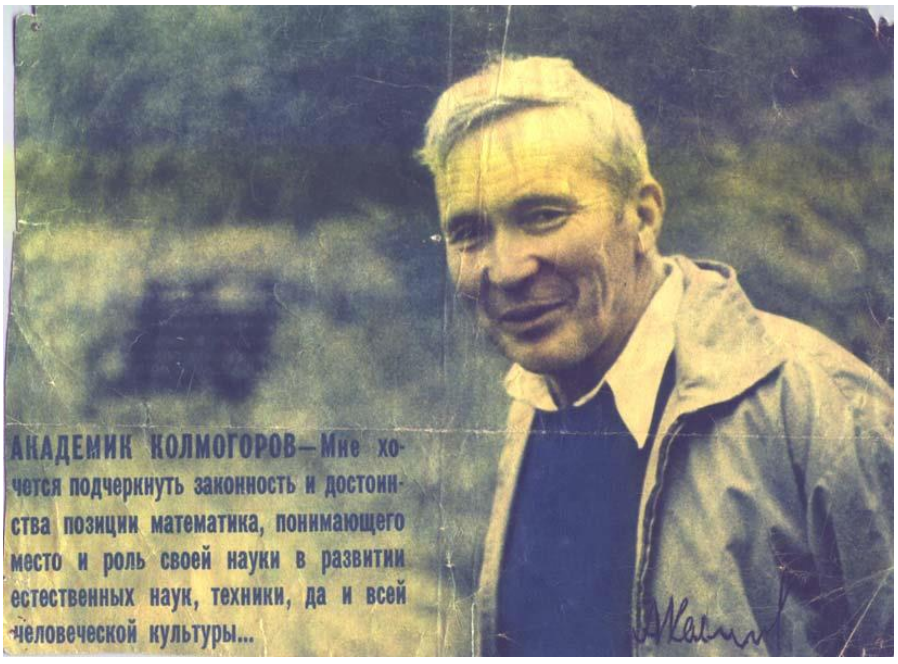
1. Look at the PROC UNIVARIATE output from the sample of 100,000 scores from the normal population. Does it appear that this sample came from a normally distributed population? Refer to the skewness, kurtosis, and the histogram (which may be split between pages if you printed your output, but you can view it intact on your computer screen) in your answer. Looks normal to me. The population μ = 100, σ = 15.

Tests for Normality

Test	--Statistic---		-----p Value------	
Kolmogorov-Smirnov	D	0.001409	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.023279	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.163224	Pr > A-Sq	>0.2500

2. Look at the Kolmogorov-Smirnov statistic. This statistic tests the null hypothesis that the sample was randomly drawn from a normally distributed population. With 100,000 scores in this sample, this statistic should have enormous power -- that is, if the population is not normal, we are almost certain to get a significant result here. Report the *p* value obtained from the Kolmogorov-Smirnov for your sample and interpret that test. We retain the hypothesis that the population is normally distributed.

Historical Note: Without Kolmogorov, we might have lost World War II to the Japanese: "In the 1940's, he created a powerful technique for using probability to make predictions in the face of randomness, on the basis of a series of observations. The technique was applied to a wide range of systems, such as the problem of landing an airplane on an aircraft carrier bobbing in the sea, calculating ahead of time what its likely position would be." From his obituary in the NY Times, linked above.



```
                          Histogram                          #             Boxplot
      157.5+*                                                  9                0
           .*                                                 19                0
           .*                                                 98                0
           .*                                                266                0
           .***                                              617                |
           .*****                                           1345                |
           .*********                                       2425                |
           .*****************                               4374                |
           .************************                          6827                |
           .**********************************              9539             +-----+
           .****************************************       11613             |     |
           .********************************************** 12978             *--+--*
           .********************************************** 13011             |     |
       92.5+****************************************       11658             |     |
           .**********************************              9295             +-----+
           .************************                          6819                |
           .*****************                               4371                |
           .**********                                      2477                |
           .*****                                           1272                |
           .***                                              609                |
           .*                                                251                0
           .*                                                 86                0
           .*                                                 34                0
           .*                                                  4                0
           .*                                                  2                0
           .
       27.5+*                                                  1                *
            ----+----+----+----+----+----+----+----+----+----+---
            * may represent up to 272 counts
```

The program obtains the actual frequency of Type I errors when the nominal alpha is .05 and Student's *t* is used to test the (true) null that the population mean is 100.

Frequency of Type I Errors

The FREQ Procedure

Type1_N9	Frequency	Percent	Frequency	Percent
No	95037	95.04	95037	95.04
Yes	4963	4.96	100000	100.00

Type1_N25	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	94949	94.95	94949	94.95
Yes	5051	5.05	100000	100.00

3. The program obtains the actual frequency of Type I errors when the nominal alpha is .05. What percentage of the samples resulted in Type I errors when *N* = 9? When *N* = 25? Looks fine, actual alpha close to nominal alpha.

Normal Population, Distribution of Sample Means, N = 9

μ = 100, σ = 15;

Obs	mean_ mean9	med_ mean9	std_ mean9	g1_mean9	g2_mean9	min_ mean9	max_ mean9
1	100.007	100.019	5.02217	-0.015113	0.015697	79.0953	123.462

4. The sample mean is an unbiased estimator of the population mean. The mean of your 100,000 sample means from the normal distribution should then be approximately 100 ("Approximately" because we simulated only 100,000 samples, not an uncountably large number of samples.) What mean did you obtain when *N* = 9? When *N* = 25? The mean of sampling distribution is the same as that of population (100) – the sample mean is an unbiased estimator of population mean.

5. Given that you know that the population standard deviation is 15, what should the standard deviation of the sample means (the standard error of the mean) be when *N* = 9? When *N* = 25?

$\sigma_M = 15/\sqrt{9} = 5 \quad \sigma_M = 15/\sqrt{25} = 3$

6. What standard deviations did you actually obtain for these two sampling distributions? Explain how you have demonstrated that the sample mean is a consistent estimator. The observed standard errors are as expected. That the standard error went down as the sample size went up demonstrates that the mean is a consistent estimator.

7. Report the skewness and kurtosis of these two distributions of sample means and describe the shape of the distributions. Skewness and kurtosis are as expected for a normal distribution.

Normal Population, Distribution of Sample Means, N = 25

5. $\sigma_M = 15/\sqrt{25} = 3$

Obs	mean_ mean25	med_ mean25	std_ mean25	g1_mean25	g2_mean25	min_ mean25	max_ mean25
1	99.9998	100.007	3.00676	-0.016430	.009493838	85.1157	113.043

4. The mean is as expected.

5. $\sigma_M = 15/\sqrt{25} = 3$

6. The standard error is as expected. It decreased with increasing sample size, demonstrating that the mean is a consistent estimator.

7. Report the skewness and kurtosis of these two distributions of sample means (that with *N* = 9 and that with *N* = 25) and describe the shape of the distributions. Skewness and kurtosis are as expected for a normal distribution.

Normal Population, Distribution of Sample Variances, N = 9

$\sigma^2 = 225$

Obs	mean_var9	med_var9	std_var9	g1_var9	g2_var9	min_var9	max_var9
1	225.202	206.556	112.649	0.98773	1.40403	5.55318	1085.71

8. Look at the statistics on the distribution of sample variances, *N* = 9. Since you know that the sample variance is an unbiased estimator, you expect a variance of $15^2 = 225$. What did you get? The mean sample variance is as expected for an unbiased estimator.

9. Is this distribution skewed, and if so, how much and in what direction? The variances are positively skewed. The mean is displaced to the right of the median. And $g_1 > 0$ by a considerable amount. Remember that this results in the distribution of Student's *t* being leptokurtic.

Normal Population, Distribution of Sample Variances, N = 25

Obs	mean_var25	med_var25	std_var25	g1_var25	g2_var25	min_var25	max_var25
1	225.066	218.917	65.0678	0.56024	0.44193	46.3876	628.740

10. Compare the skewness of the distribution of sample variances when *N* = 25 to that when *N* = 9. What has happened to the distribution of sample variances as *N* increased? The skewness of the sampling distribution of variances decreases as sample size increases. Remember that this results in the distribution of Student's *t* become less leptokurtic as degrees of freedom increase.

Normal Population, Distribution of Sample Standard Deviations, N = 9

Obs	mean_std9	med_std9	std_std9	g1_std9	g2_std9	min_std9	max_std9
1	14.5457	14.3720	3.69121	0.27037	-0.018511	2.35652	32.9501

11. Look at the skewness of the distributions of the standard deviations. Compare the skewness of the distributions of standard deviations with that observed for the distributions of the variances. Keeping in mind that the standard deviation is just the square root of the variance, draw a conclusion regarding the effect of a square root transformation when applied to a distribution of scores which is positively skewed. The skewness is less than it was with the sample variances. The square root transformation reduces positive skewness. This is important to keep in mind if you have data that are positively skewed and you want to analyze them with a procedure that assumes that they came from a normally distributed population. If you transform them you just might eliminate the skewness problem.

Normal Population, Distribution of Sample Standard Deviations, N = 25

Obs	mean_ std25	med_ std25	std_ std25	g1_std25	g2_std25	min_ std25	max_ std25
1	14.8460	14.7959	2.15929	0.13421	-0.020219	6.81085	25.0747

Normal Population, Distribution of Z Test Statistic, N = 9

Obs	mean_Z9	med_Z9	std_Z9	g1_Z9	g2_Z9	min_Z9	max_Z9
1	.001323939	.003886093	1.00443	-0.015113	0.015697	-4.18093	4.69232

12. Look at the distributions of the *z* statistics. On skewness and kurtosis, for *N* = 9, compare the distribution of sample means with the distribution of *z* statistics. Do the same for *N* = 25. Keeping in mind that the *z* statistics are just a linear transformation of the means, draw a conclusion regarding the effect of linear transformations on the shape of a distribution. Skewness and kurtosis are identical to what they were for the distribution of sample means. The *z* transformation is linear. Accordingly, it has no effect on the shape of the distribution that is transformed. An all too common delusion is that the *z* transformation somehow magically makes the data normal. In fact, it has absolutely no effect on the shape of the distribution.

Normal Population, Distribution of Z Test Statistic, N = 25

Obs	mean_Z25	med_Z25	std_Z25	g1_Z25	g2_Z25	min_Z25	max_Z25
1	-.000078202	.002198043	1.00225	-0.016430	.009493838	-4.96142	4.34782

Normal Population, Distribution of T Test Statistic, N = 9

Obs	mean_T9	med_T9	std_T9	g1_T9	g2_T9	min_T9	max_T9
1	.002035514	.004009438	1.15413	-0.031274	1.45933	-10.5847	8.16035

13. Now look at the *t*-distributions. For *N* = 9, how does the standard deviation and the kurtosis of the *t*-distribution differ from that of the *z*-distribution? What characteristic of the distribution of sample variances causes *t* to differ from *z* in this way (explain your answer)? Notice that the standard deviation and kurtosis of the *t* distribution are greater than they were with the *z* distribution.

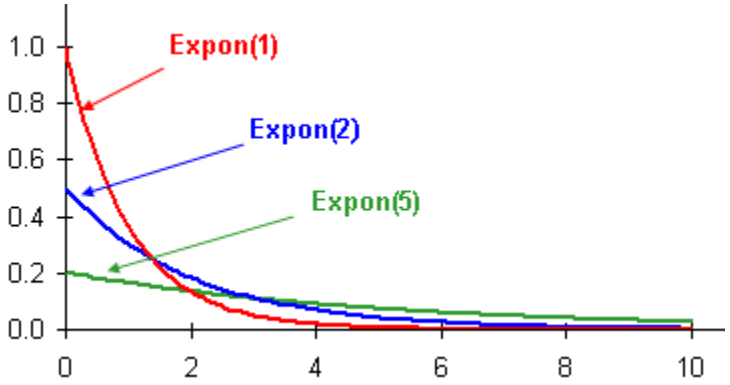
- This is price we pay for estimating the population standard deviation from the sample data. Both of these differences will result in the power of the *t* test being less than the power of the *z* test.
- The standard deviation of Student's *t* distribution is the square root of *df* / (*df* – 2). For 8 *df* that is SQRT(8 / 6) = 1.15.
- The skewness of Student's *t* distribution is 0 if *df* > 3. I ran a separate Monte Carlo, five times, with 1,000,000 samples each, and 4 scores in each sample. The obtained g_1 statistic for the empirical distributions of *t* ranged from -.51 to 1.32. The distributions of means was not skewed.
- The kurtosis of Student's *t* is infinite when *df* ≤ 4. Otherwise it is 6/(*df* – 4), here 6/(8-4) = 1.5.

Normal Population, Distribution of T Test Statistic, N = 25

Obs	mean_T25	med_T25	std_T25	g1_T25	g2_T25	min_T25	max_T25
1	-.000509231	.002211353	1.04591	-0.020967	0.31931	-6.10827	5.44554

14. How did the shape of the *t*-distribution change when we went from *N* = 9 to *N* = 25? Draw a conclusion with respect to how the shape of the *t*-distribution changes with increasing *df*. Notice that the *t* distribution is rapidly approaching the standardized normal distribution as sample size increases.

- The standard deviation of Student's *t* distribution is the square root of *df* / (*df* – 2). For 24 *df* that is SQRT(24 / 22) = 1.044.
- The kurtosis of Student's *t* distribution is 6 / (*df* – 4). For 24 *df* that is 6 / 20 = 0.3.



The exponential distribution is the time between discrete events which occur at an average rate of λ. The mean of the distribution is 1/λ. For example, if you get a robocall on average twice every hour, you expect to have one half hour between calls. The variance is $1/\lambda^2$. The value of λ here is 1 and the output of the random number generator was multiplied by 100. All exponential distributions have a skewness of 2 and a kurtosis of 6.

One Sample of 100,000 scores from Exponential Distribution

Moments

N	100000	Sum Weights	100000
Mean	100.151214	Sum Observations	10015121.4
Std Deviation	99.6933354	Variance	9938.76112
Skewness	1.98688255	Kurtosis	5.86391511
Uncorrected SS	1996892743	Corrected SS	993866173
Coeff Variation	99.5428125	Std Error Mean	0.31525801

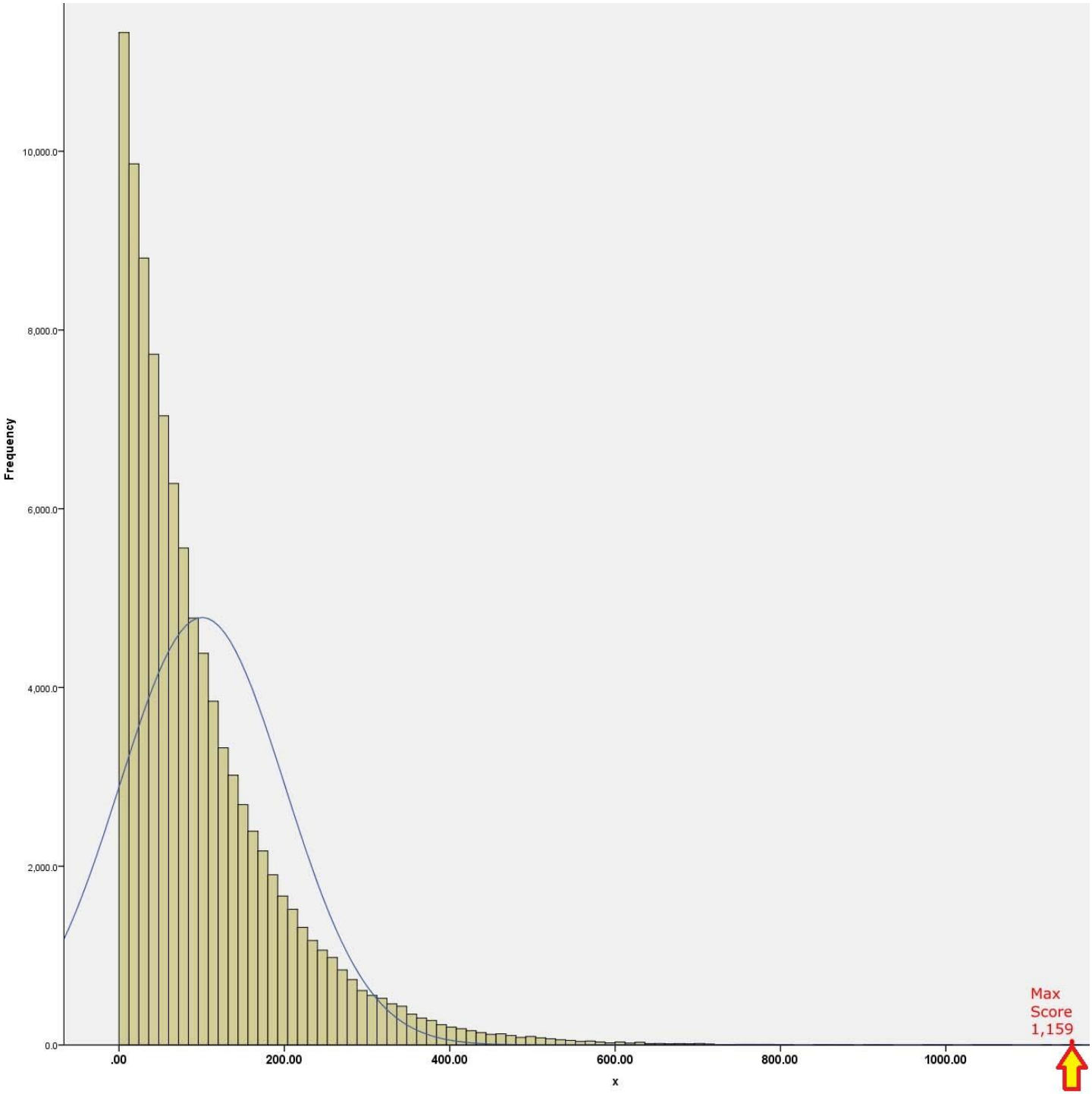
Tests for Normality

Test		--Statistic---	-----p Value------	
Kolmogorov-Smirnov	D	0.157547	Pr > D	<0.0100
Cramer-von Mises	W-Sq	784.0373	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	4599.352	Pr > A-Sq	<0.0050

15. Look at the PROC UNIVARIATE output from the sample of 100,000 scores from the exponential population. Does it appear that this sample came from a normally distributed population? Refer to the skewness, kurtosis, histogram, and Kolmogorov-Smirnov statistic in your answer. This distribution is far from normal, with great positive skewness and kurtosis. We reject the hypothesis that the population is normally distributed. Recall that high values of kurtosis often result from the presence of an unusually large number of outliers. When most of those outliers are in one tail, they also produce skewness.

```
-----------------------------------------------------------------------------------------------
                          Histogram                          #             Boxplot
         1075+*                                              2                *
             .*                                              5                *
             .*                                              5                *
             .*                                              1                *
             .*                                              4                *
             .*                                             12                *
             .*                                             22                *
          725+*                                             34                *
             .*                                             64                *
             .*                                             92                *
             .*                                            175                *
             .*                                            250                *
             .*                                            432                0
             .*                                            756                0
          375+**                                          1169                0
             .***                                         1951                0
             .****                                        3210                |
             .*******                                     5351                |
             .***********                                 8666                |
             .******************                         14805             +--+--+
             .****************************              23928             *-----*
           25+************************************************ 39066      +-----+
              ----+----+----+----+----+----+----+----+----+---
              * may represent up to 814 counts
```

Here is a histogram of another sample of scores drawn from an exponential distribution. Superimposed over it is a normal curve with the same mean and variance.



Mean = 100. Median = 70. Skewness = 2. Kurtosis = 6.

Frequency of Type I Errors

The FREQ Procedure

Type1_N9	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	89614	89.61	89614	89.61
Yes	10386	10.39	100000	100.00

Type1_N25	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	92301	92.30	92301	92.30
Yes	7699	7.70	100000	100.00

16. The program obtains the actual frequency of Type I errors when the nominal alpha is .05. What percentage of the samples resulted in Type I errors when *N* = 9? When *N* = 25? The actual alpha exceeds the nominal alpha. This is the result of violation of the normality assumption.

Exponential Population, Distribution of Sample Means, N = 9

Obs	mean_ mean9	med_ mean9	std_ mean9	g1_mean9	g2_mean9	min_ mean9	max_ mean9
1	99.8042	96.2520	33.3652	0.66758	0.68708	14.1360	343.389

17. Look at the statistics from the distribution of sample means from an exponential population, *N* = 9. How does this sampling distribution differ in shape from that you obtained with sample means drawn from a normally distributed population? The sampling distribution here is clearly not normally distributed, given the positive skewness and kurtosis.

Exponential Population, Distribution of Sample Means, N = 25

Obs	mean_ mean25	med_ mean25	std_ mean25	g1_mean25	g2_mean25	min_ mean25	max_ mean25
1	99.8938	98.4490	20.0326	0.42589	0.28340	33.5751	211.172

18. Look at the statistics from the distribution of sample means from an exponential distribution, *N* = 25. How does its shape differ from that of the sample means from an exponential distribution, *N* = 9? You have empirically verified what theorem discussed in our class and our textbook? As sample size increases, the shape of the distribution of sample means is approaching normality (skewness and kurtosis approaching zero). Now you have seen the Central Limit Theorem in action.

Exponential Population, Distribution of Sample Standard Deviations, N = 9 22

Obs	mean_ std9	med_std9	std_std9	g1_std9	g2_std9	min_std9	max_std9
1	91.4870	84.6019	40.0396	1.18006	2.49561	9.86589	459.719

19. Look at the statistics from the distributions of sample standard deviations. How does the skewness compare with that obtained you sampled from a normal population? Greater skewness than when the population was normal (g_1 was .27 then)

Exponential Population, Distribution of Sample Standard Deviations, N = 25 23

Obs	mean_ std25	med_ std25	std_ std25	g1_std25	g2_std25	min_ std25	max_ std25
1	96.4412	93.2493	26.0547	0.80896	1.27574	27.1228	299.385

Exponential Population, Distribution of Z Test Statistic, N = 9

Obs	mean_Z9	med_Z9	std_Z9	g1_Z9	g2_Z9	min_Z9	max_Z9
1	-.005873714	-0.11244	1.00096	0.66758	0.68708	-2.57592	7.30166

- Again, the shape of the sampling distribution for the test statistic *z* is identical to that of the distribution of sample means.

Exponential Population, Distribution of Z Test Statistic, N = 25

Obs	mean_Z25	med_Z25	std_Z25	g1_Z25	g2_Z25	min_Z25	max_Z25
1	-.005310055	-0.077552	1.00163	0.42589	0.28340	-3.32124	5.55858

- As sample size increases, the shape of the distribution of *z* is approaching normal.

Exponential Population, Distribution of T Test Statistic, N = 9

Obs	mean_T9	med_T9	std_T9	g1_T9	g2_T9	min_T9	max_T9
1	-0.44917	-0.12841	1.56328	-2.08174	10.0540	-23.0804	6.56564

20. Look at the *t*-distributions computed with our samples from an exponential distribution. Are the skewness measures here what you would expect for Student's *t*?

- This is most certainly not Student's *t* distribution.
- The standard deviation should be 1.15.
- The skewness should be 0.
- The kurtosis should be 1.5
- Student's *t* is constructed from samples randomly drawn from a normal distribution, and the parent population here was distinctly non-normal.

Exponential Population, Distribution of T Test Statistic, N = 25

Obs	mean_T25	med_T25	std_T25	g1_T25	g2_T25	min_T25	max_T25
1	-0.22946	-0.081890	1.17882	-0.93726	2.11844	-10.5704	3.46830

- The standard deviation should be 1.044.
- The skewness should be 0.
- The kurtosis should be 0.3.
- The obtained distribution of *t* is not approaching that of Student's *t* as rapidly as the obtained distribution of sample means is approaching normal.

I have extended this Monte Carlo study to sample sizes much larger than 25. Here are the results:

Exponential Population, Distribution of T Test Statistic, N = 49

The FREQ Procedure

Type1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	93311	93.31	93311	93.31
Yes	6689	6.69	100000	100.00

Exponential Population, Distribution of Sample Means, N = 49

Obs	mean_ mean	med_mean	std_mean	g1_mean	g2_mean	min_mean	max_mean
1	100.023	99.4049	14.3116	0.28212	0.14157	52.3589	178.882

Exponential Population, Distribution of T Test Statistic, N = 49

Obs	mean_T	med_T	std_T	g1_T	g2_T	min_T	max_T
1	-0.15054	-0.042487	1.09229	-0.63287	0.90348	-6.84412	3.2939

Notice that the skewness and kurtosis of the distribution of sample means (and thus of the z test statistic as well) is close to normal, but the distribution of the t test statistic is still far from normal).

Exponential Population, Distribution of T Test Statistic, N = 100

The FREQ Procedure

Type1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	93856	93.86	93856	93.86
Yes	6144	6.14	100000	100.00

Exponential Population, Distribution of Sample Means, N = 100

Obs	mean_ mean	med_mean	std_mean	g1_mean	g2_mean	min_mean	max_mean
1	99.9583	99.6616	9.99124	0.19068	0.027422	62.7076	147.530

Exponential Population, Distribution of T Test Statistic, N = 100

Obs	mean_T	med_T	std_T	g1_T	g2_T	min_T	max_T
1	-0.10689	-0.034573	1.04328	-0.41804	0.39616	-6.33850	3.44675

Exponential Population, Distribution of Sample Means, N = 256

Exponential Population, Distribution of T Test Statistic, N = 256

The FREQ Procedure

Type1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	94694	94.69	94694	94.69
Yes	5306	5.31	100000	100.00

Exponential Population, Distribution of Sample Means, N = 256

Obs	mean_ mean	med_mean	std_mean	g1_mean	g2_mean	min_mean	max_mean
1	100.005	99.8551	6.22245	0.12389	.004985574	76.9861	128.270

Exponential Population, Distribution of T Test Statistic, N = 256

Obs	mean_T	med_T	std_T	g1_T	g2_T	min_T	max_T
1	-0.061615	-0.023633	1.01246	-0.25089	0.14387	-5.31619	3.69340

As you can see, things are looking better with a sample size of 256, but the distribution of *t* is still skewed. The actual alpha is not far from the nominal alpha though.

Exponential Population, Distribution of T Test Statistic, N = 900

Type1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	94897	94.90	94897	94.90
Yes	5103	5.10	100000	100.00

Exponential Population, Distribution of Sample Means, N = 900

Obs	mean_ mean	med_mean	std_mean	g1_mean	g2_mean	min_mean	max_mean
1	99.9893	99.9529	3.32984	0.062028	.003032090	86.1940	117.136

Exponential Population, Distribution of T Test Statistic, N = 900

Obs	mean_T	med_T	std_T	g1_T	g2_T	min_T	max_T
1	-0.036543	-0.014106	1.00436	-0.13813	0.044858	-4.73671	4.64696

Exponential Population, Distribution of T Test Statistic, N = 3,025

The FREQ Procedure

Type1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	94970	94.97	94970	94.97
Yes	5030	5.03	100000	100.00

Exponential Population, Distribution of Sample Means, N = 3,025

Obs	mean_mean	med_mean	std_mean	g1_mean	g2_mean	min_mean	max_mean
1	99.9977	99.9840	1.81851	0.046122	0.013240	92.5457	108.106

Exponential Population, Distribution of T Test Statistic, N = 3,025

Obs	mean_T	med_T	std_T	g1_T	g2_T	min_T	max_T
1	-0.019475	-.008697489	1.00152	-0.063785	0.019703	-4.46793	4.13612

Here is the take-home message: When the parent population is very skewed (in this case, $g_1 = 2$), the distribution of sample means (and of the *z* test-statistic) can get close to normal with sample sizes that do not make the distribution of the *t* test-statistic be close to that of Student's *t*.

So, why are simulations like this called "Monte Carlo?"

Monte Carlo is one of the administrative divisions of Monaco. Monaco is the second smallest county in the world – only Vatican City is smaller. It is a monarchy, and most of its residents are wealthy persons who reside there to evade taxes. It is bordered by France and the Mediterranean Sea.

John von Neumann is credited with associating the name "Monte Carlo" with simulations:

> Physicists at Los Alamos Scientific Laboratory were investigating radiation shielding and the distance that neutrons would likely travel through various materials. Despite having most of the necessary data, such as the average distance a neutron would travel in a substance before it collided with an atomic nucleus or how much energy the neutron was likely to give off following a collision, the problem could not be solved with analytical calculations. John von Neumann and Stanislaw Ulam suggested that the problem be solved by modeling the experiment on a computer using chance. Being secret, their work required a code name. Von Neumann chose the name "Monte Carlo". The name is a reference to the Monte Carlo Casino in Monaco where Ulam's uncle would borrow money to gamble.

Such simulations can, however, be traced back to earlier times, including work by William Sealy Gosset. You should already be familiar with Gosset, who developed Student's *t* . His employer (Guiness Breweries) did not allow their employees to publish under their true names (which might reveal trade secrets), so Gosset published under the pseudonym "Student.:

Chevrolet also borrowed the name “Monte Carlo” for one model of their cars, almost certainly because Monte Carlo is well known one of the sites for Formula One Racing – they run this race through the streets of Monte Carlo. You won’t find any Chevrolets running in those races, however. If you want to see a racing Chevrolet, watch NASCAR racing. NASCAR legend Dale Earnhart drove a Monte Carlo. In fact, he was driving a Monte Carlo at the Daytona 500 (2001) when a wreck near the end of his race took his life.