```
*g1g2.sas;
*************************************************************************************;
*Illustrates use of PROC STANDARD to standardize a variable, and;
*The computation of G1, Fisher's skewness, and;
*The computation of G2, kurtosis;
*************************************************************************************;

data EDA; infile 'D:\_Stats\StatData\EDA.dat'; input Y;
proc means mean skewness kurtosis N; var Y; run;
```

The MEANS Procedure

Analysis Variable : Y

Mean	Skewness	Kurtosis	N
72.5104167	0.5255689	0.0323668	96

The estimated population skewness, g_1, is .53, mildly positively skewed. The estimated population kurtosis, g_2, is .03.

```
PROC STANDARD data=eda mean=0 std=1 out=z_scores; run;
proc means mean skewness kurtosis N; var Y; run;
```

The scores are standardized to mean 0, standard deviation 1, and placed in the data set "z_scores." Notice that the mean is now 0, but the skewness and kurtosis have not been changed.

The MEANS Procedure

Analysis Variable : Y

Mean	Skewness	Kurtosis	N
-4.09395E-16	0.5255689	0.0323668	96

Transforming scores to *z* scores changes the mean and the standard deviation but has absolutely no effect on the shape of the distribution.

```
data z34; set z_scores;
Z3=Y**3; Z4=Y**4;
proc means data=z34 noprint; var Z3 Z4; output out=sumZ34 N=N sum=sumZ3 sumZ4; run;
```

This code creates a new data set, "sumZ34" which contains the cubed *z* scores and the *z* scores to the 4th power.

```
data skew; set sumz34; G1=N/(n-1)/(n-2)*sumZ3;
 G2=N*(n+1)/(n-1)/(n-2)/(n-3)*sumZ4 - 3*(n-1)*(n-1)/(n-2)/(n-3);
proc print; run;
```

This code computes g_1 and g_2 using the formulae presented in Skewness, Kurtosis, and the Normal Curve

Obs	_TYPE_	_FREQ_	N	sumZ3	sumZ4	G1	G2
1	0	96	96	48.8889	279.103	0.52557	0.032367

```
*Kurtosis-Uniform.sas;
*************************************************************************************;
options formdlim='-' pageno=min nodate;
TITLE 'One Sample of 500,000 Scores From Uniform(0,1) Distribution'; run;
DATA uniform; DROP N; DO N=1 TO 500000; X=UNIFORM(0);
OUTPUT; END;
PROC MEANS mean std skewness kurtosis; VAR X; run;
```

X=UNIFORM(0); selects one score from a uniform distribution that ranges from 0 to 1. Embedding this random number generator within a “Do Loop” makes SAS sample half a million such scores. Proc Means produced this output:

Analysis Variable : X

Mean	Std Dev	Skewness	Kurtosis
0.4996790	0.2885179	-0.000057309	-1.1983384

If we were to obtain the entire population, the mean would be .5, the standard deviation .2887. skewness 0, and kurtosis -1.2/ Sampling error caused us to get a tiny bit away from those values. Such sampling error can be reduced by increasing the sample size.

```
*Kurtosis-T.sas;
*************************************************************************************;
options formdlim='-' pageno=min nodate;
TITLE 'T ON 9 DF, T COMPUTED ON EACH OF 500,000 SAMPLES';
TITLE2 'EACH WITH 10 SCORES FROM A STANDARD NORMAL POPULATION'; run;
DATA T9; DROP N; DO SAMPLE=1 TO 500000; DO N=1 TO 10; X=NORMAL(0);
OUTPUT; END; END;
PROC MEANS NOPRINT; OUTPUT OUT=TS T=T; VAR X; BY SAMPLE;
PROC MEANS MEAN STD N KURTOSIS; VAR T; run;
```

This code creates half a million samples, each with 10 scores drawn from a standardized normal distribution. For each of those samples the value of Student’s *t* is computed. The resulting distribution of 500,000 values of t is the sampling distribution of Student’s *t* on N – 1 = 9 degrees of freedom. Then some basic descriptive statistics are computed on the sampling distribution.

The MEANS Procedure

Analysis Variable : T

Mean	Std Dev	N	Kurtosis
-0.0024258	1.1349675	500000	1.2223092

Student’s *t* is like the standard normal distribution in that it has a mean of zero, and a skewness of zero, but it has a standard deviation greater than 1 and a kurtosis greater than 1. It has more scores in its tails than would be expected in a normal distribution.

I ran this code a few more times with different sample sizes.

T ON 10 DF, SAMPLING DISTRIBUTION OF 500,000 TS

The MEANS Procedure

Analysis Variable : T

Mean	Std Dev	N	Kurtosis
0.0031798	1.1161173	500000	0.9787776

Increasing the degrees of freedom caused the standard deviation and kurtosis to decrease.

T ON 16 DF, SAMPLING DISTRIBUTION OF 500,000 TS

The MEANS Procedure

Analysis Variable : T

Mean	Std Dev	N	Kurtosis
0.000846732	1.0709861	500000	0.5218626

Increasing the degrees of freedom caused the standard deviation and kurtosis to decrease.

T ON 28 DF, SAMPLING DISTRIBUTION OF 500,000 TS

The MEANS Procedure

Analysis Variable : T

Mean	Std Dev	N	Kurtosis
-2.414558E-6	1.0385425	500000	0.2372472

Increasing the degrees of freedom caused the standard deviation and kurtosis to decrease. If we were to continue to increase the degrees of freedom the standard deviation and the kurtosis of Student's *t* would keep getting closer and closer to those of the standard normal distribution. This is what is meant by "Student's *t* approaches the normal curve as degrees of freedom increase."

Here is Table 1 from the document Skewness, Kurtosis, and the Normal Curve.

Table 1.

Kurtosis for 7 Simple Distributions Also Differing in Variance

X	freq A	freq B	freq C	freq D	freq E	freq F	freq G
05	20	20	20	10	05	03	01
10	00	10	20	20	20	20	20
15	20	20	20	10	05	03	01
Kurtosis	-2.0	-1.75	-1.5	-1.0	0.0	1.33	8.0
Variance	25	20	16.6	12.5	8.3	5.77	2.27
Shoulders	5, 15	5.5, 14.5	5.9. 14.1	6.5, 13.5	7.1, 12.9	7.6, 12.4	8.5, 11.5

Platykurtic Leptokurtic

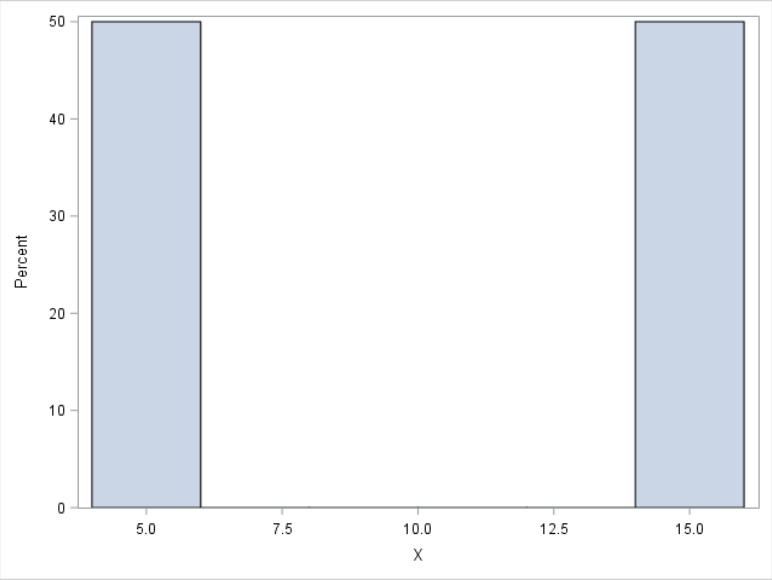
```
*Kurtosis_Beta2.sas;
*Illustrates the computation of population kurtosis;
*Using data from the handout Skewness, Kurtosis, and the Normal Curve;
*************************************************************************;
```

```
options pageno=min nodate formdlim='-' FORMCHAR="|----|+|---+=|-/\<>*";
data A; do s=1 to 20; X=5; output; X=15; output; end;
*SS=1000, SS/N = 25, M = 10;
data ZA; set A; Z=(X-10)/5; Z4A=Z**4;
proc means mean; var Z4A; run;
```

This code creates a distribution of 40 scores, 20 with value 5 and 20 with value 15. This is distribution A from Table 1. Each score is standardized and then raised to the 4th power and then the mean is found for these transformed scores. Karl Pearson (1905) defined a distribution's degree of kurtosis as $\eta = \beta_2 - 3$, where $\beta_2 = \frac{\sum(Y-\mu)^4}{n\sigma^4}$, the expected value of the distribution of *Z* scores which have been raised to the 4th power. β_2 is often referred to as "Pearson's kurtosis," and β_2 - 3 (often symbolized with γ_2) as "kurtosis excess" or "Fisher's kurtosis," even though it was Pearson who defined kurtosis as β_2 - 3. β_2 is, for distribution A,

Kurtosis excess (γ_2) is 1 – 3 = -2, the lowest possible value of kurtosis excess, and that shown in Table 1. Note that it has a perfect U shape, with half the scores at one value and half at another. It has no scores in its tails because it has no tails. All of the scores are at its shoulders (one standard deviation below the mean and one standard deviation above the mean.

Distribution A



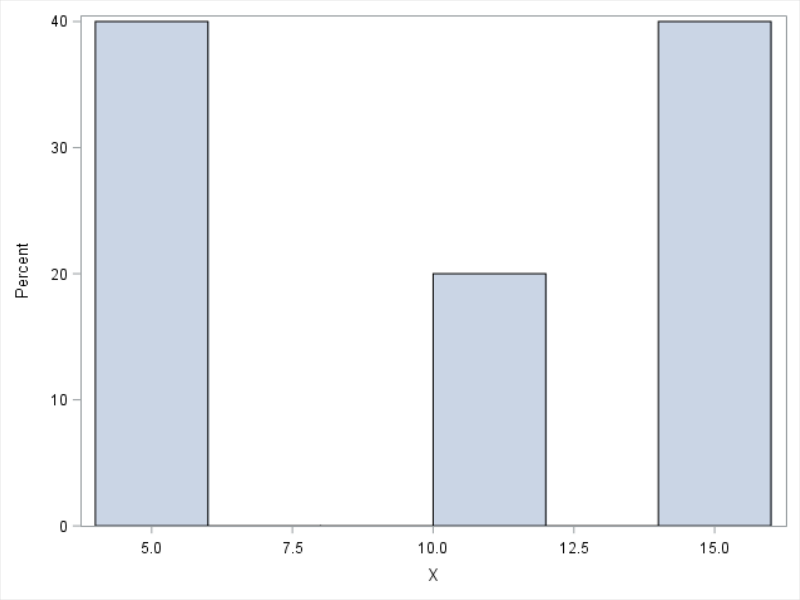
The MEANS Procedure

Analysis Variable : Z4A

	Mean
β_2	1.0000000
γ_2	1 – 3 = -2

Now, watch what happens as I move scores from the shoulders and into the tails and the center:

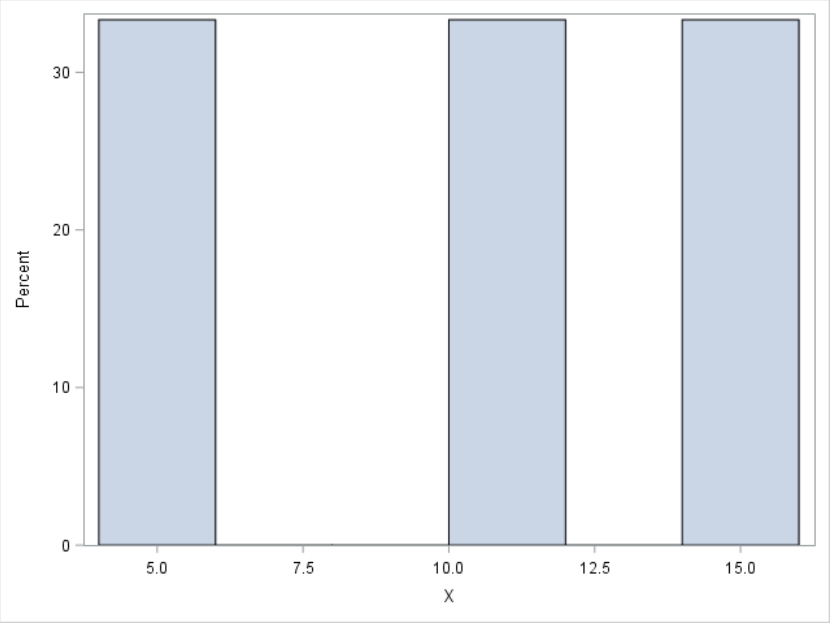
Distribution B



$\beta_2 = 1.2500000$

$\gamma_2 = 1.25 - 3 = -1.75$

Distribution C



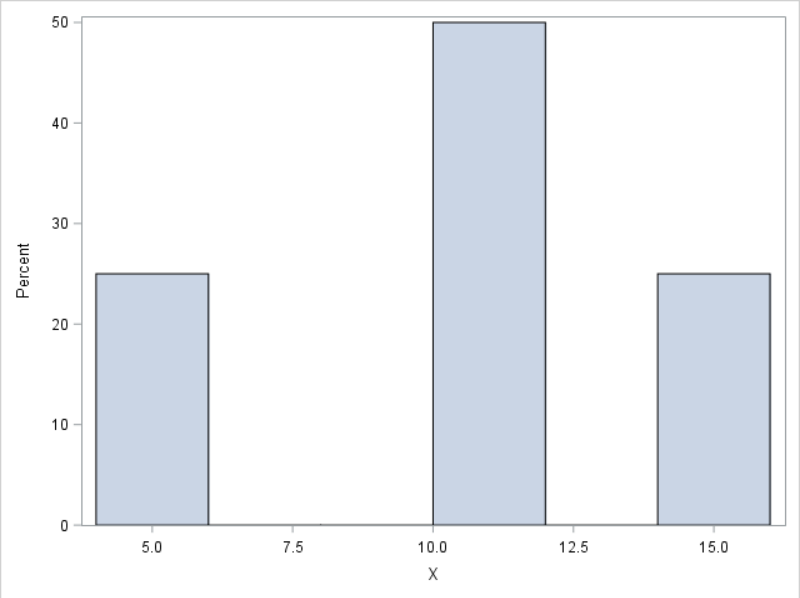
The middle bar should be centered at 10. Sometimes Sgplot messes up.

Mean

$\beta_2 = 1.5000000$

$\gamma_2 = 1.5 - 3 = -1.5$

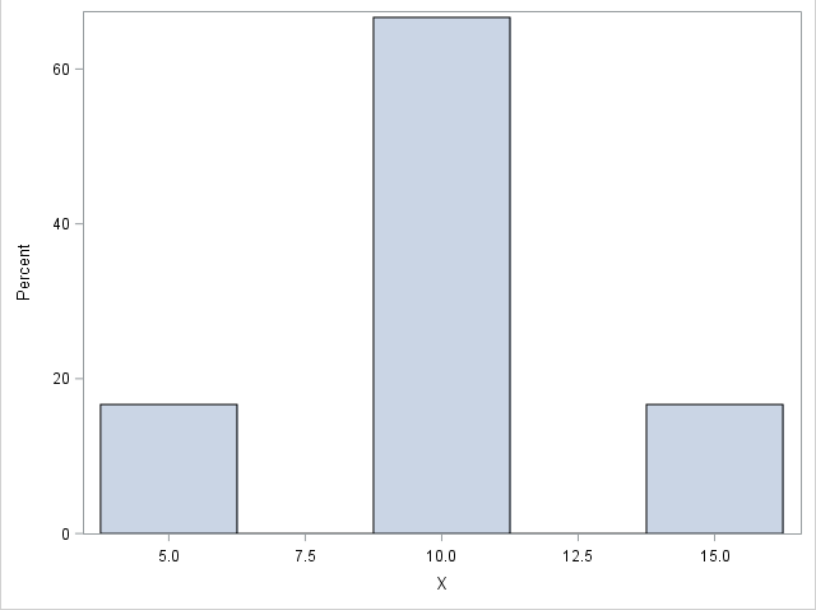
Distribution D



β_2 = 2.0000000

γ_2 = 2 – 3 = -1

Distribution E

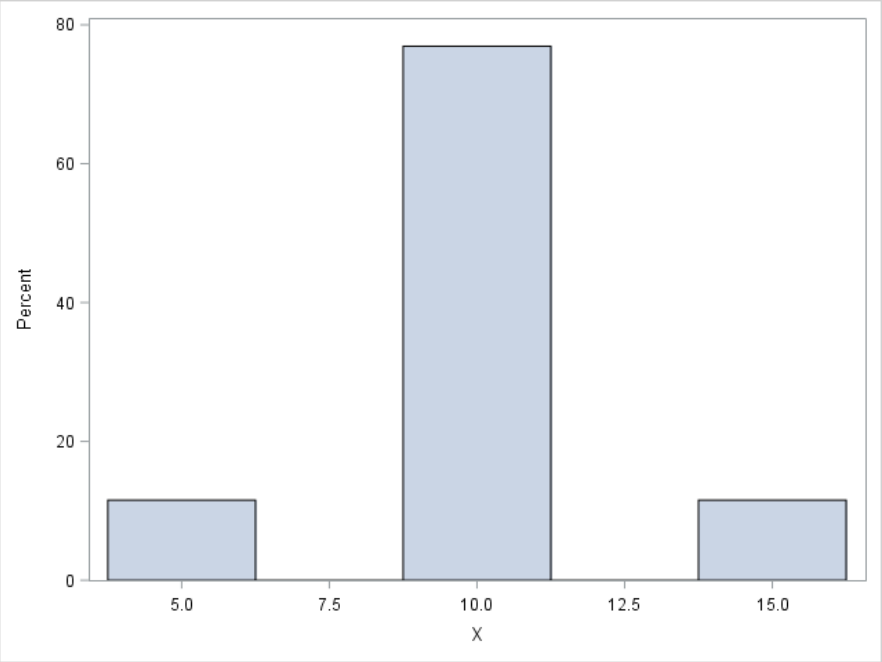


Mean

β_2 = 3.0000000

γ_2 = 3 – 3 = 0

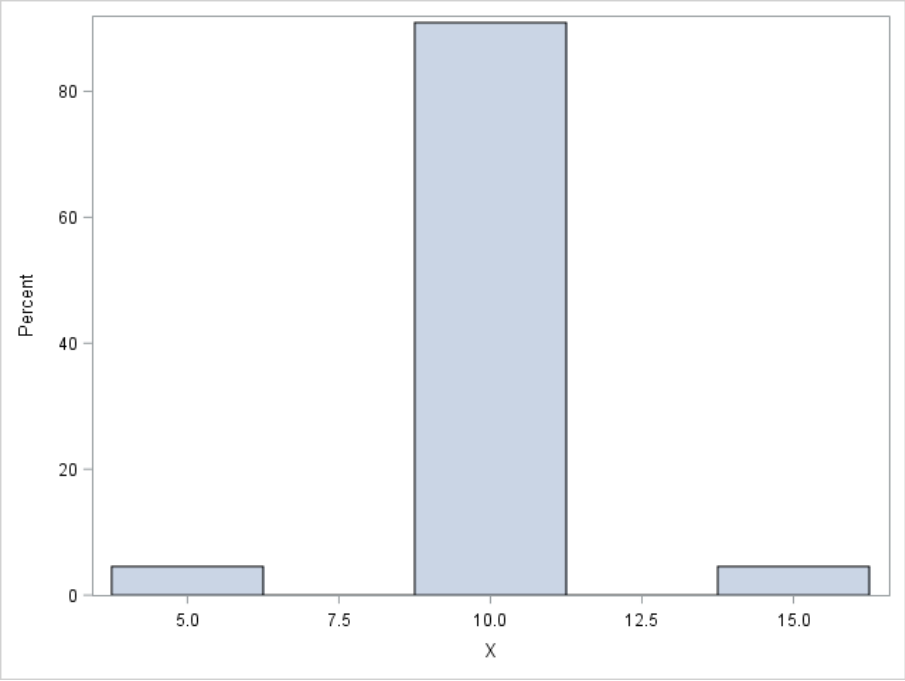
Distribution F



β_2 = 4.3333333

γ_2 = 4.33 – 3 = 1.33

Distribution G



β_2 = 11.0000000

γ_2 = 11 – 3 = 8