WORKSHEET — STATISTICAL ANALYSIS OF DATA

Two groups of Biology 101 students achieved the following scores on the third test:
Group 1: 94, 94, 91, 88, 85, 85, 81, 79, 79, 77, 75, 70, 65, 56, 56, 53
Group 2: 67, 64, 63, 62, 60, 54, 54, 49, 47, 36, 33, 32, 30, 26, 12
From these scores, the **range**, the difference between maximum and minimum scores, and the **mean**, the average, can be calculated (Range = $X_{max} - X_{min}$, and Mean = $\overline{X} = (\Sigma X_i)/n$) [note that Σ means sum--add them up], so

	Group 1	Group 2
Range	94 − 53 = 41	67 − 12 = 55
ΣX_i	1228	689
n	16	15
$\overline{X}$	1228÷16 = 76.75	689÷15 = 45.93

Two means and ranges might be close together, causing one to assume the two samples were similar. However, in one sample, the scores could be more closely grouped with only a few at the extremes, and in another sample, the scores could be more widely spread. Thus, it is useful to calculate the **variance**, and from that, the **standard deviation** as indicators of how spread out or clustered the data are (dispersion). If you are interested in the variance of only those individuals in your sample, $s^2 = \Sigma(X_i - \overline{X})^2/n$ for this finite group of data. However, if this group is a representative sampling of a larger population, then it has been found that $^2 = \Sigma(X_i - \overline{X})^2/(n-1)$ is a better estimate of the population variance. Thus, for our two groups of students:

GROUP 1

X_i	$X_i - \overline{X}$	$(X_i - \overline{X})^2$
94	17.25	297.5625
94	17.25	297.5625
91	14.25	203.0625
88	11.25	126.5625
85	8.25	68.0625
85	8.25	68.0625
81	4.25	18.0625
79	2.25	5.0625
79	2.25	5.0625
77	0.25	0.0625
75	−1.75	3.0625
70	−6.75	45.5625
65	−11.75	138.0625
56	−20.75	430.5625
56	−20.75	430.5625
53	−23.75	564.0625

n = 16 Σ = 2701
$\overline{X}$ = 76.75 $s^2 = 2701 \div 16 = 168.81$
$s = \sqrt{168.81} = 12.99$
OR: $^2 = 2701 \div 15 = 180.07$
$= \sqrt{180.07} = 13.42$

GROUP 2

X_i	$X_i - \overline{X}$	$(X_i - \overline{X})^2$
67	21.07	443.80
64	18.07	326.40
63	17.07	291.27
62	16.07	258.14
60	14.07	197.87
54	8.07	65.07
54	8.07	65.07
49	3.07	9.40
47	1.07	1.14
36	−9.93	98.67
33	−12.93	167.27
32	−13.93	194.14
30	−15.93	253.87
26	−19.93	397.34
12	−33.93	1151.47

n = 15 Σ = 3920.92
$\overline{X}$ = 45.93 $s^2 = 3920.92 \div 15 = 261.39$
$s = \sqrt{261.39} = 16.17$
$^2 = 3920.92 \div 14 = 280.06$
$= \sqrt{280.06} = 16.74$

In a "normal" distribution, a graph of scores on the x-axis versus number of people who got that score on the y-axis should give a bell-shaped curve with the mean in the center. 68% of the people should be within one standard deviation unit to either side of the mean and 95% should be within two units of the mean.

Luckily, most scientific calculators and PCs will perform these calculations from entered data and a few simple commands.

When comparing data like these, we could make one of several hypotheses. Either we could say that there is no difference between the two groups (the **null hypothesis--H_o**) or that there is a significant difference of some sort. For plants on two given sites A and B, hypothesis 1

(H_1) might be that density on site A is greater than on site B, while hypothesis 2 (H_2) might be that density on site B is greater than A. Comparing the means and variances alone can be misleading, so we must do further statistical testing to determine whether we can say (with, usually, a 95% confidence level) that, for the null hypothesis, there is no difference between these two samples or for H_1 and H_2 combined, that there is a difference between them. The test used to compare the means of two groups is called the **t-test**:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{\Sigma(X - \overline{X})^2_1 + \Sigma(X - \overline{X})^2_2}{n_1 + n_2 - 2} \times \left[\dfrac{1}{n_1} + \dfrac{1}{n_2}\right]}}$$

Thus, for our two groups of students:

$$t = \frac{76.75 - 45.93}{\sqrt{\dfrac{2701 + 3920.92}{16 + 15 - 2} \times \left[\dfrac{1}{16} + \dfrac{1}{15}\right]}} = \frac{30.82}{\sqrt{\dfrac{6621.92}{29} \times (0.0625 + 0.0667)}}$$

$$= \frac{30.82}{\sqrt{29.49}} = \frac{30.82}{5.43} = 5.68$$

This value is compared to a table of t-scores for a number of possible degrees of freedom [DF] (See copy of table). For one sample, DF = n − 1, so for two samples, DF = $n_1 + n_2 - 2$ (in this case, 16 + 15 − 2 = 29). Since we want to be 95% certain (0.95) we must look at the t-value for 1.00 − 0.95 = 0.05. As an alternative hypothesis, let us say there is a difference between the two groups--either #1 is higher or lower than #2. Since we have two choices, we need to look at the value of t for 0.05 for the "two-tailed" test, which is 2.045. Our calculated value might be positive or negative, depending on if we compared #1 to #2 or #2 to #1, so if the absolute value of t is greater than or equal to the table value, we can accept the alternative hypothesis, and if it is less than the table value, we can reject the alternative and accept the null hypothesis. For our two groups of students, our observed value of 5.68 is greater than the table value of 2.045, thus we can accept the alternative hypothesis and reject the null hypothesis--there is a difference between our two groups.

PROBLEM:

A group of ecology students was studying and comparing two different woodland areas by taking 20 sample plots in each area. In each woods the combined number of sugar plus red maple trees in each of the 20 plots was found to be:
Woods A: 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 2, 0, 1, 0, 0, 0
Woods B: 0, 1, 0, 2, 0, 2, 0, 4, 0, 3, 0, 1, 0, 0, 0, 8, 0, 1, 0, 4
Calculate the range, mean, s, and for the number of maples per plot for each of these two areas. Use a t-test to tell if the number of maples per plot for the two areas is significantly different.

The **density** of a species is the number of organisms of that species per unit area/volume. The **dispersion** of a species refers to the spatial distribution of the individuals. Generally, these are arranged in one of three ways: In **uniform** dispersion, individuals are evenly scattered throughout the habitat. In **random** dispersion, individuals are scattered throughout the habitat, "at random". In **clumped** dispersion, individuals are clustered together in some locations, while other locations have few/none.

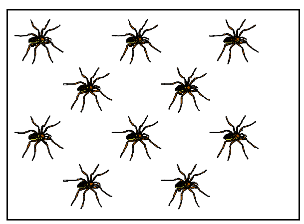

Figure 1. Uniform

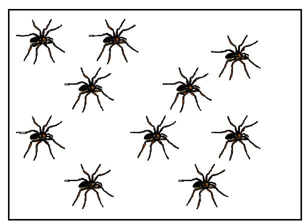

Figure 2. Random

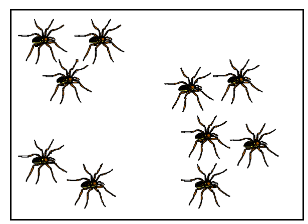

Figure 3. Clumped

Statistically, it is possible to analyze whether a species is randomly or uniformly dispersed or clumped. Consider the following example:

plot #	# of ladybugs	plot #	# of ladybugs	plot #	# of ladybugs
1	0	13	5	25	0
2	4	14	4	26	3
3	2	15	2	27	4
4	3	16	3	28	5
5	0	17	4	29	3
6	4	18	0	30	1
7	2	19	4	31	4
8	4	20	3	32	3
9	0	21	4	33	2
10	3	22	4	34	4
11	3	23	3	35	3
12	4	24	4	36	1

From these data, the following table can be constructed:

# of ladybugs per plot (X)		observed # of plants with that many individuals (O)	total # of individuals	
X	X^2	O	(X)(O)	
0	0	5	0 × 5 =	0
1	1	2	1 × 2 =	2
2	4	4	2 × 4 =	8
3	9	10	3 × 10 =	30
4	16	13	4 × 13 =	52
5	25	2	5 × 2 =	10
		total # of plots (n) = 36	Σ =	102

mean # of individ./plot (m) = 102/36 = 2.833

For each X (# individ./plot) something called the **poisson relative frequency** can be calculated by the following formula: prf = $(e^{-m})(m^x/x!)$. In this equation, e is the base of natural logarithms (= 2.7183) so $e^{-m} = e^{-2.833} = 0.0588$ (note, most scientific calculators have e^x functions--what this is really saying is $2.7183^{-2.833} = 0.0588$). We can then calculate the following values:

X	m^X		X!	prf
0	2.833^0 =	1.000	0! = 1	(0.0588)(1.000/ 1) = 0.0588
1	2.833^1 =	2.833	1! = 1	(0.0588)(2.833/ 1) = 0.1666
2	2.833^2 =	8.028	2! = 2 × 1 = 2	(0.0588)(8.023/ 2) = 0.2359
3	2.833^3 =	22.745	3! = 3 × 2 × 1 = 6	(0.0588)(22.745/ 6) = 0.2230
4	2.833^4 =	64.445	4! = 4 × 3 × 2 × 1 = 24	(0.0588)(64.445/ 24) = 0.1579
5	2.833^5 =	182.595	5! = 5 × 4 × 3 × 2 × 1 = 120	(0.0588)(182.595/120) = 0.0895

By multiplying the total number of plots (n) by the prf, we can predict how many plots we should expect to have each of the X number of species out of the total plots, so for

X	and prf,	then, (prf)(n) = expected # of plots (E).
0	0.0588	× 36 = 2.117
1	0.1666	× 36 = 5.976
2	0.2359	× 36 = 8.492
3	0.2230	× 36 = 8.028
4	0.1579	× 36 = 5.684
5	0.0895	× 36 = 3.222

This means that if the species is randomly distributed, most of the plots should have about the same average number of individuals with only a few plots having extremely high or low numbers--like a bell-shaped curve. We can, then, compare the observed (O) and expected (E) values to see how closely the observed data fit the expected data--is the species really randomly distributed? This can be done by means of a statistical analysis called the **chi-square (^{2}) test** where $^2 = \Sigma[(O - E)^2/E]$. Thus, we can add two more columns to our chart as follows:

X	O	E	(O − E)	$(O - E)^2/E$
0	5	2.117	2.883	3.926
1	2	5.976	−3.976	2.645
2	4	8.492	−4.492	2.376
3	10	8.028	1.972	0.484
4	13	5.684	7.316	9.417
5	2	3.222	−1.222	0.463
			$^2 = \Sigma[(O - E)^2/E]$ =	19.311

This number is then compared to a table of ^{2} values. Because we have six different choices for X, degrees of freedom (DF) = 6 − 1 = 5. Once again, we want to test the data at the 95% level, so we need to look up the value for 0.05. The value corresponding to 0.05 and 5 DF is 11.070. Since $^2_{calc}$ of 19.311 is greater than this, we must reject the null hypothesis that dispersion is random for this group of individuals (that there is no difference between observed and expected values). Thus, the ^{2} test supports an alternate hypothesis that there is a difference between O and E. In order to figure out whether the organisms are uniformly dispersed or clumped, a few more calculations must be done to compare the variance (s^2) and the mean (m). Another formula for s^2 which is easier to use here is: $s^2 = [\Sigma X^2 \cdot O - n(m)^2]/DF$, where (in this example) n = 36 and m = 2.833,thus $n(m)^2 = 289$, so:

X	X^2	O	X^2O
0	0	5	0
1	1	2	2
2	4	4	16
3	9	10	90
4	16	13	208
5	25	2	50
			Σ = 366

then (366 − 289)/5 = 15.4

For any samples, if $m < s^2$ (or $s^2 > m$), then the distribution is clumped and if $m > s^2$, then the distribution is uniform. In this case, m = 2.833 and $s^2 = 15.4$, which is larger, thus the dispersion pattern is clumped.

PROBLEM:

A group of ecology students was working in an oak-hickory forest sampling plots for various herbs, grasses, etc. In 20 plots, the following numbers of goldenrod plants were counted: 4, 0, 0, 1, 1, 7, 3, 11, 2, 2, 0, 7, 13, 1, 0, 0, 6, 0, 2, 7. The null hypothesis would be that the goldenrod plants are randomly distributed. Complete the following table:

X	X^2	X!	O	X·O	m^x	$prf=e^{-m}\cdot m^x/X!$	$E=prf\cdot n$	O−E	$(O-E)^2/E$	$X^2\cdot O$
0										
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
			n=	Σ=					$X^2=\Sigma=$	Σ=
				÷n=m=						$s^2=$
				$e^{-m}=$						
				$n(m)^2=$						

df= ___ X^2 from table = ___

Can we accept the null hypothesis? ___ because ___

If not, is the distribution uniform or clumped? ___ because ___

	^{2} Value		Critical Values of t	
Degrees of Freedom	at P = .05 level		one-tailed @ .05	two-tailed @ .05
1	3.841		6.314	12.706
2	5.991		2.920	4.303
3	7.815		2.353	3.182
4	9.488		2.132	2.776
5	11.070		2.015	2.571
6	12.592		1.943	2.447
7	14.067		1.895	2.365
8	15.507		1.860	2.306
9	16.919		1.833	2.262
10	18.307		1.812	2.228
11	19.675		1.796	2.201
12	21.026		1.782	2.179
13	22.362		1.771	2.160
14	23.685		1.761	2.145
15	24.996		1.753	2.131
16	26.296		1.746	2.120
17	27.587		1.740	2.110
18	28.869		1.734	2.101
19	30.144		1.729	2.093
20	31.410		1.725	2.086
21	32.671		1.721	2.080
22	33.924		1.717	2.074
23	35.172		1.714	2.069
24	36.415		1.711	2.064
25	37.652		1.708	2.060
26	38.885		1.706	2.056
27	40.113		1.703	2.052
28	41.337		1.701	2.048
29	42.557		1.699	2.045
30	43.773		1.697	2.042
31			1.696	2.040
32			1.694	2.037
33			1.692	2.035
34			1.691	2.032
35			1.690	2.030
36			1.688	2.028
37			1.687	2.026
38			1.686	2.024
39			1.685	2.023
40			1.684	2.021
60			1.671	2.000

ANSWERS:

of Sugar and Red Maples in 20 Sample Plots

Woods A			Woods B		
X	$X - \overline{X}$	$(X - \overline{X})^2$	X	$X - \overline{X}$	$(X - \overline{X})^2$
2	1.6	2.56	8	6.70	44.89
1	0.6	0.36	4	2.70	7.29
1	0.6	0.36	4	2.70	7.29
1	0.6	0.36	3	1.70	2.89
1	0.6	0.36	2	0.70	0.49
1	0.6	0.36	2	0.70	0.49
1	0.6	0.36	1	−0.30	0.09
0	−0.4	0.16	1	−0.30	0.09
0	−0.4	0.16	1	−0.30	0.09
0	−0.4	0.16	0	−1.30	1.69
0	−0.4	0.16	0	−1.30	1.69
0	−0.4	0.16	0	−1.30	1.69
0	−0.4	0.16	0	−1.30	1.69
0	−0.4	0.16	0	−1.30	1.69
0	−0.4	0.16	0	−1.30	1.69
0	−0.4	0.16	0	−1.30	1.69
0	−0.4	0.16	0	−1.30	1.69
0	−0.4	0.16	0	−1.30	1.69
0	−0.4	0.16	0	−1.30	1.69
0	−0.4	0.16	0	−1.30	1.69
n=20		Σ=6.80	n=20		Σ=82.20
range: 2−0=2		s^2=0.34	range: 8−0=8		s^2=4.11
		s=0.58			s=2.03
ΣX=8		^{2}=0.36	ΣX=26		^{2}=4.33
$\overline{X}$=0.40		=0.60	$\overline{X}$=1.30		=2.08

$$t = \frac{0.40 - 1.30}{\sqrt{\frac{6.80 + 82.20}{20 + 20 - 2} \times \left[\frac{1}{20} + \frac{1}{20}\right]}} = \frac{-0.90}{\sqrt{\frac{89.00}{38} \times \frac{1}{10}}} = \frac{-0.90}{\sqrt{0.234}} = \frac{-0.90}{0.484} = -1.8597$$

DF = 20 + 20 − 2 = 38

for two-tailed test @ 0.05: t_{30} = 2.042 and t_{40} = 2.021

$$\frac{2.042 - 2.021}{10} \times 8 = 0.0168 \textit{ and } 2.042 - 0.017 = 2.025 = t_{38}$$

Since 1.86 < 2.025, we can accept the null hypothesis--there is no difference between the number of maples per plot in these two areas.

ANSWERS:
Goldenrod

X	X^2	X!	O	X·O	m^x	prf=$e^{-m} \cdot m^x$/X!	E=prf·n	O−E	$(O-E)^2/E$	$X^2 \cdot O$
0	0	1	6	0	1.000	0.0351	0.702	5.298	39.38	0
1	1	1	3	3	3.350	0.1175	2.350	0.659	0.18	3
2	4	2	3	6	11.222	0.1969	3.938	−0.938	0.22	12
3	9	6	1	3	37.595	0.2198	4.396	−3.396	2.62	9
4	16	24	1	4	125.944	0.1841	3.682	−2.682	1.95	16
5	25	120	0	0	421.914	0.1231	2.462	−2.462	2.46	0
6	36	720	1	6	1413.412	0.0689	1.378	−0.378	0.10	36
7	49	5040	3	21	4734.931	0.0330	0.660	2.340	8.30	147
8	64	40320	0	0	15862.019	0.0138	0.276	−0.276	0.28	0
9	81	362880	0	0	53137.762	0.0051	0.102	−0.102	0.10	0
10	100	3628800	0	0	178011.500	0.0017	0.034	−0.034	0.03	0
11	121	39916800	1	11	596338.540	0.0005	0.010	0.990	98.01	121
12	144	4.790016E8	0	0	1997734.100	0.0001	0.002	−0.002	0.002	0
13	169	6.227020E9	1	13	6692409.300	0.00004	0.0008	1.000	1250	169
			n=20	Σ=67					X^2=Σ=1404.23	Σ=513
			÷n=m=67÷20=3.35						s^2=	
			e^{-m}=0.0350843						(513−224.45)÷13	
			$n(m)^2$=224.45						=22.196	

df= 14 − 1 = 13 X^2 from table = 22.362

Can we accept the null hypothesis? no because $X^2_{tab} < X^2_{obs}$

If not, is the distribution uniform or clumped? clumped because $s^2 > m$