APPENDIX: STATISTICAL TOOLS

I. Notes on random sampling

Why do you need to sample randomly?

In order to measure some value on a population of organisms, you usually cannot measure all organisms, so you sample a subset of the population. In order for this subset to be representative of the population as a whole, it cannot be selected subjectively from the population. That is, there should not be any bias about which particular organisms are selected to measure. The way in which we avoid bias is by taking random samples. This means that each individual to be sampled is chosen randomly (or by chance) from the population. Most statistical tests that you use assume that you have sampled randomly.

Consider an example of how nonrandom sampling could affect your results. Suppose you are testing the hypothesis that plants growing in acidic soils have lower growth rates than those growing in neutral soils. You grow plants from seed, then transfer the seedlings to the two soil types. You subconsciously choose more vigorous seedlings to transfer to the neutral soil, and less vigorous seedlings (perhaps those with some herbivore damage) to transfer to the acidic soils. If you observe a difference in growth rates between these two groups, you cannot be sure that it was due to soil differences. If you had randomly chosen which seedling would be transplanted to the two soil types, you could be more certain that any difference in growth was due to the soil.

Ways to sample randomly

Random sampling can be difficult in field biology. In many other situations, you can assign the individuals numbers, then use a random numbers table to decide which treatment each individual would receive. You sometimes must be creative in devising ways to sample randomly, or at least avoid bias.

Example:

If you need to randomly sample a leaf or branch from a tree, stand with your back to the tree and reach back, sampling the first leaf or branch you happen to touch. Another option is to pick a branch, then choose which leaf you will sample by looking up a number in a random numbers table (close your eyes and put your finger someplace on the page). If the random number is 4, choose the fourth leaf from the branch tip. You can experiment with other methods. The important thing is to not look at the leaves, and don't choose one because it has few or many galls.

II. Random Numbers Table

Enter the table at random (e.g., point to it with a pencil). Proceed horizontally or vertically from that point to obtain as many digits or numbers as needed.

16363	27995	41125	66592	12860	51577	71692	16628	67324	83419	39240	07412
66121	30863	34081	04607	75502	04707	68653	43338	87116	46480	36041	74787
96575	53826	41704	15194	56279	49971	42649	70175	95444	76761	35816	55837
52166	96367	47481	18862	45623	93595	10650	12053	38724	78957	29886	86154
91028	65112	12180	46755	32955	96324	79516	63432	10341	94262	39179	69020
27709	28594	25839	87636	39442	80248	74996	19481	66351	06377	97824	37041
91136	85882	24857	51299	10838	73647	61668	27778	37789	53166	58454	41833
71188	66829	04281	27206	70331	17899	51875	77253	86078	12234	92391	02238
86762	71131	59628	80300	57614	93257	55946	89789	18238	25252	70653	37393
83339	77436	38773	50081	29229	23649	15426	31504	25341	74610	19022	91666
24947	44623	42554	58705	35822	02055	94723	96090	70086	31284	24361	19557
28186	41642	92387	39381	99737	20272	47305	58768	72531	46704	14361	65260
32884	49283	67787	94116	52161	38414	84319	50978	01803	19969	56564	88407
79466	24667	41425	91616	65372	86449	24668	54459	95823	12610	13266	22570
11775	52778	19652	06992	01513	88362	98785	27116	61859	53933	31861	23445
66248	79998	13670	64290	15231	54748	29354	94888	73377	74669	69746	84505
80039	39369	59470	26716	46462	52922	78911	19953	93147	13792	12761	35869
99366	43213	90418	61337	14260	56876	74992	67722	91075	91137	11354	64755
33913	96112	24338	95294	01216	02949	66586	74661	94809	14762	78894	75333
44129	59181	43233	33124	37576	47110	20722	73824	79877	46516	75488	82273
04478	84929	68978	22693	92885	76772	94407	22729	81553	46129	11806	20192
38821	64814	47522	86915	49195	28851	71306	84207	94648	17014	38834	92108
97443	01973	61484	62615	41741	99773	12110	15931	09120	33874	84738	08786
85704	98251	29175	76784	94321	18842	22817	35496	66678	75891	76086	97310
42480	07655	40183	25256	09522	84493	63682	72724	87341	57773	15782	68323
64858	69010	31483	37740	46141	29821	53313	62641	33150	50417	43686	87116
96176	99746	95971	49622	82590	35237	21789	67766	12829	14555	41262	15937
28438	37287	68455	18433	94419	79448	71742	25031	09557	79627	78776	51468
28608	29116	12344	46025	36451	75226	92497	58712	21333	69243	30157	38766
88959	48172	24219	21986	59749	96376	54314	34511	11377	15936	14015	79795
71608	45585	48448	86156	44586	34124	68122	95516	51596	61392	86385	13587
54399	71428	35942	93594	52565	75225	80233	92176	35457	71737	62239	51060
59755	72465	86327	39716	45208	36755	75772	74799	21169	31143	22754	78596
17961	16698	69213	93388	89813	58122	18803	39039	97264	24619	97375	38838
16298	96671	57299	16061	69740	56814	63324	35363	54903	90979	60792	83056
44946	17269	17264	62822	26350	24863	53662	72367	72757	16294	44210	73288

III. Tests of Independence using the G-statistic:

A two-way test of independence is commonly used in ecology for situations in which we want to test whether two different characteristics or conditions occur independently of one another. (An alternative, the Chi Square test, is more commonly used, but G is quite simple to calculate and more closely follows an actual chi-square distribution. An advantage of the chi square test is that students must calculate expected frequencies, so they see for each cell in the table how closely the observed value matches the expected.)

The calculations for a 2-way test of independence follow:

1) $a = \Sigma$ (f ln f) for cell frequencies
2) $b = \Sigma$ (f ln f) for row and column totals
3) c = n ln n
4) G = 2 (a - b+c)
5) Compare G with the critical value of X^2. In a 2 x 2 table, there is (2-1)(2-1) =1 degree of freedom. Use $\alpha = 0.05$. If G < critical value, accept H_o; if $G \geq$ critical value, reject H_o.

Example:
An ecologist wishes to know whether leaves that have sawfly galls are more susceptible to herbivory by leaf-chewing insects than are leaves without galls. The ecologist recorded presence or absence of leaf chewing damage on 50 leaves with galls and 50 leaves without galls. Here are the results (hypothetical data):

Galls on leaf	chewing damage on leaf		
	yes	no	Total
yes	31	19	50
no	22	28	50
Total	53	47	n = 100

To calculate G:
1) a = 31 ln 31 + 22 ln 22 + 19 ln 19 + 28 ln 28 = 323.7
2) b = 50 ln 50 + 50 ln 50 + 53 ln 53 +47 ln 47 = 782.6
3) c = 100 ln 100 = 460.5
4) G = 2(323.7 - 782.6 +460.5) = 3.24
5) In the table of critical values of the chi square distribution, the critical value of G for 1 d.f. and $\alpha = 0.05$ is 3.841.

The ecologist accepts the null hypothesis that chewing damage occurs independently of gall presence on willow leaves at this site and concludes that the observed differences were small enough to have resulted from chance.

IV. Testing a Poisson Distribution

For ecological events that occur relatively rarely and independently of other events of the same type, the frequency of events should follow a Poisson distribution. For example, adult female sawflies may oviposit on only a small fraction of the leaves of a willow tree. We could test whether sawfly galls occur independently of whether there are other galls on the same leaf by comparing the frequency distribution of galls to a Poisson distribution.

H_o: The number of galls per leaf on willows follows a Poisson distribution.
H_a: the number of galls per leaf on willows does not follow a Poisson distribution.

General test procedure: Compare observed frequency data with expected frequencies (Sokal, R. R., and F. J. Rohlf. 1981. Biometry. W.H. Freeman, San Francisco, section 5.3)

	Frequency (# trials)		
#events/trial	observed (f)	expected ($\hat{f}$)	deviation from expected ($f-\hat{f}$)
0	f_0	$\frac{n}{e^{\bar{X}}}$	
1	f_1	$\frac{n}{e^{\bar{X}}}(\bar{X})$	
2	f_2	$\frac{n}{e^{\bar{X}}} \bullet \frac{\bar{X}^2}{2}$	
3	f_3	$\frac{n}{e^{\bar{X}}} \bullet \frac{\bar{X}^3}{6}$	
4	f_4	$\frac{n}{e^{\bar{X}}} \bullet \frac{\bar{X}^4}{24}$	
5	f_5	$\frac{n}{e^{\bar{X}}} \bullet \frac{\bar{X}^5}{120}$	
6	f_6	$\frac{n}{e^{\bar{X}}} \bullet \frac{\bar{X}^6}{720}$	
7+	f_7	$\frac{n}{e^{\bar{X}}} \bullet \frac{\bar{X}^7}{5040}$	
total	$N=\Sigma f=$		

N is the sample size (number of leaves sampled), $\overline{X}$ (average number of galls per leaf) = total number of galls/total number of leaves. For calculations of expected frequencies, see example data, below.

Example (from real class data, 1998): From a sample of 201 leaves, we found 140 galls. So $\overline{X} = 0.70$.

	Frequency (# trials)		
# galls/leaf	observed (f) (= # leaves)	expected ($\hat{f}$)	deviation from expected (f- $\hat{f}$)
0	145	$\frac{n}{e^{\overline{X}}} = \frac{201}{e^{0.7}} = 99.8$	+
1	22	$\frac{n}{e^{\overline{X}}}(\overline{X}) = \frac{201}{e^{0.7}}(0.7) = 69.9$	-
2	12	$\frac{n}{e^{\overline{X}}} \bullet \frac{\overline{X}^2}{2} = 24.5$	-
3	8	$\frac{n}{e^{\overline{X}}} \bullet \frac{\overline{X}^3}{6} = 5.7$	+
4	7	$\frac{n}{e^{\overline{X}}} \bullet \frac{\overline{X}^4}{24} = 1.0$	+
5	4	$\frac{n}{e^{\overline{X}}} \bullet \frac{\overline{X}^5}{120} = 0.1$	+
6	0	$\frac{n}{e^{\overline{X}}} \bullet \frac{\overline{X}^6}{720} = 0$	
7+	3	$\frac{n}{e^{\overline{X}}} \bullet \frac{\overline{X}^7}{5040} = 0$	+
Total	n=Σf = 201		

Note that it is best to lump classes with expected frequencies less than 5. In this example, we would combine leaves with 3 or more galls, giving the table on the following page:

	Frequency (# trials)		
# galls/leaf	observed (f) (= # leaves)	expected ($\hat{f}$)	deviation from expected ($f-\hat{f}$)
0	145	$\frac{n}{e^{\bar{X}}} = \frac{201}{e^{0.7}} = 99.8$	+
1	22	$\frac{n}{e^{\bar{X}}}(\bar{X}) = \frac{201}{e^{0.7}}(0.7) = 69.9$	-
2	12	$\frac{n}{e^{\bar{X}}} \bullet \frac{\bar{X}^2}{2} = 24.5$	-
3+	22	$\frac{n}{e^{\bar{X}}} \bullet \frac{\bar{X}^3}{6} = 5.7$	+
Total	n=Σf = 201		

The data do not fit the Poisson distribution very well. Since the number of leaves with just one gall is much smaller than expected (and the number with 3 or more galls greater), we can conclude that female sawflies do not avoid ovipositing on leaves that already have 1 gall.