

Statistical Analysis

Chapter 11, pages 229-269
Appendix 5 – Table of Critical Values t and χ^2
Appendix 8 – Terms and Formulas
Appendix 13 – Test of Effects of Using Parametric Statistics on a Very Non-normal Population
Appendix 14 – Introduction to Resampling Methods
Appendix 16 – Instructions on using DSTPLAN and PC SIZE: CONSULTANT
Appendix 19 – Instructions on using NCSS Probability Calculator to Calculate P values

Statistical Analysis

Objective: Trainees will be able to:

- Understand the importance of random sampling
- Use graphs to explore the nature of monitoring data sets.
- Construct confidence intervals around estimates of population parameters.
- Use significance tests to test for differences between years in means and proportions:
 - For independent samples.
 - For dependent (paired) samples.

- Understand the assumptions regarding parametric statistics.
- Know about the use of nonparametric statistics, including resampling, and when their use is appropriate.
- Graph the results of data analysis.
- Understand a computer printout from a statistical program.
- Interpret the results of monitoring.
- Know about some statistical software programs.

Importance of Random Sampling



Can you select a representative sample?

See if you can do better than a random sample. Pick 5 numbers for the table below that best represent the population of 100 numbers.

4	4243	2792	675	7648	2423	55452	34	1897	39
30486	56712	65	354	25	1306	1626	63	46	293
896	2968	18	1142	2550	1497	4792	30	85	2219
26255	73	93	969	16229	159	102	1944	276	10
272	60	3256	672	41921	16	693	922	1140	1096
110	15	1667	51	67	133	46	68	34	1376
3229	1992	253	4147	3229	84	761	13	698	84
61029	815	1625	76	94	529	21	52	395	889
129	6849	1946	8	87	300	260	29	46	727
409	715	17	23	19	83	570	64	812	167

Submit | **Print** | To see the results without selecting members, just click **submit**.

Responsible party: Paul_Gatzert@usgs.gov
 Thanks to USGS Patuxent Wildlife Research Center for hosting this page for the Science Staff.
[Privacy Statement](#) | [Disclaimer](#) | [FOIA](#) | [Accessibility](#)
 U.S. Department of the Interior |  | 
<http://www.pwrc.usgs.gov/furl1/>
[Contact Us](#)

Can you select a representative sample?

True Mean	True Mean
3,725.32	3,725.32
Everyone's Mean	Random Mean
6,000.58	3,416.36
Difference	Difference
2,275.26	-308.76

Percent in quartiles of population
 Each cell should have 25%, but many people select too many values in the top quartile.

	Bottom 25%	Next 25%	Next 25%	Top 25%
Judgement Samples	20%	17%	21%	42%
Random Samples	27%	26%	23%	23%

182 People were closer to the true mean than a random sample.
 176 People were not as close to the true mean as a random sample.

The population follows a lognormal distribution, with a few large values and many small ones.

Responsible party: Paul_Gatzert@usgs.gov
 Thanks to USGS Patuxent Wildlife Research Center for hosting this page for the Science Staff.
[Privacy Statement](#) | [Disclaimer](#) | [FOIA](#) | [Accessibility](#)
 U.S. Department of the Interior |  | 
<http://www.pwrc.usgs.gov/RepresentativeSample.cfm>
[Contact Us](#)

How Did the Digest Get it So Wrong?

The magazine committed two major errors:

1. The sampled population was vastly different from the target population.
 - Only people with higher than average incomes could subscribe to magazines, own cars, or have telephones in 1936.
 - Those are the people who were most likely to vote Republican.
2. They sampled this population by sending out questionnaires and asking people to respond.
 - This is a volunteer sample: most of those who took the time to fill out and return the questionnaires had strong feelings against Roosevelt.
 - Most of those who supported Roosevelt didn't bother to respond.

The Gallup Poll

- George Gallup was riding high after the 1936 election.
- Not only did he correctly predict that Roosevelt would win, he correctly predicted the results of the Literary Digest poll within 1 percent! (He said the Digest would be incorrect by 18 points, when they were actually incorrect by 19 points).
- Both based on a sample of only 3,000 people (compared to the Digest's sample of 2.4 million).

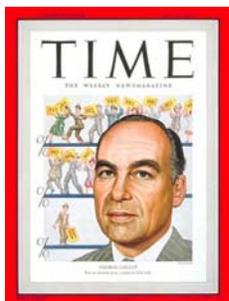
Gallup's Methods

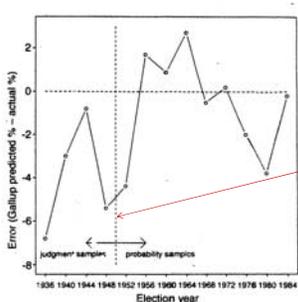
- Until 1956 Gallup and other pollsters used a technique called quota sampling to draw the sample used to predict the winner of the election.
 - Quota sampling involves dividing the population to be sampled into groups expected to vote differently and taking a sample of a specified size (the quota) for each group.
 - For example, quotas would be determined for middle-class urban women, lower-class rural men, etc.
 - Interviewers were then sent out to interview people in these groups and meet their quotas for each group.

Gallup's Methods

- Interviewers, however, were not required to take a random sample from each of these groups; rather, they sampled in any way that was convenient (this type of sampling is called a convenience sample).
 - Thus, they might stand on a street corner and poll any businessman that might happen to walk by until they met their quota of businessmen. Or, they might skip the houses of any lower class rural men who had mean-looking dogs.
 - The result of this type of sampling is a sample that is less representative than a random sample.

Gallup and the 1948 Election



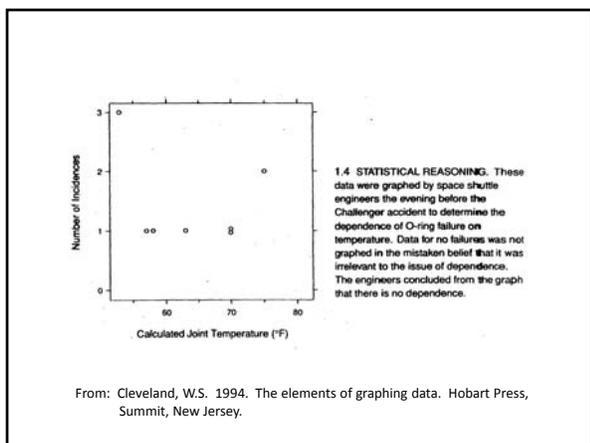


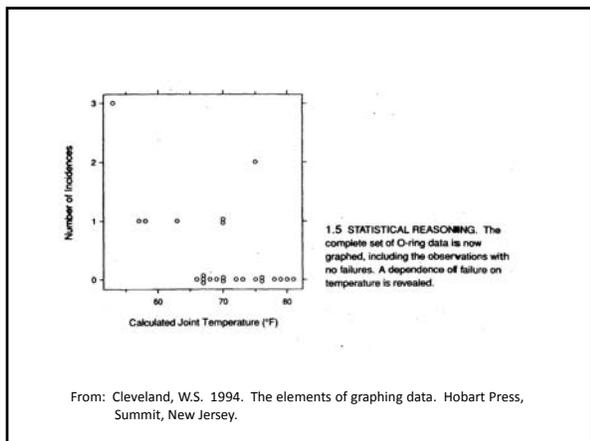
Although this graph shows that Gallup started using probability (= random) sampling shortly after 1948, other sources suggest he didn't start until just before the 1956 election.

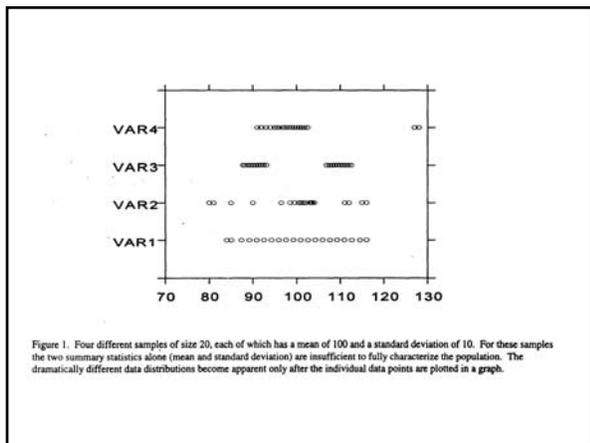
FIG. 1. Presidential election winning candidate estimation errors for Gallup polls, 1948-1984. Until 1948, quota sampling was used in samples on the order of 3000 respondents. The modern probability-based samples aim for 1500 respondents.

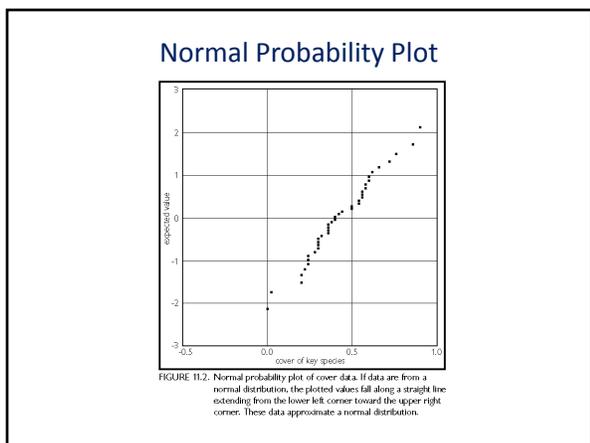


Data Exploration









Normal Probability Plot

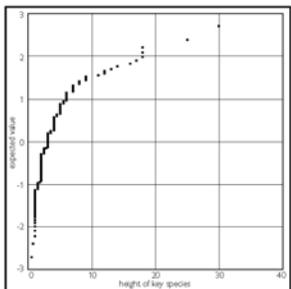


FIGURE 11.3. Normal probability plot of plant heights. If data are from a normal distribution, the plotted values fall along a straight line extending from the lower left corner toward the upper right corner. These data are not from a normal distribution.

Histogram

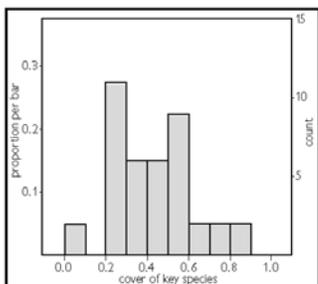


FIGURE 11.4. Histogram of cover data, with 10 bars chosen (2 of the bars contain no values). Notice individual data points cannot be distinguished.

Histogram

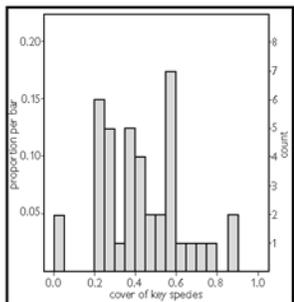


FIGURE 11.5. Histogram of same cover values used to create Figure 11.4, but with 20 bars instead of 10 (6 of the 20 bars contain no values).

Dit Plot

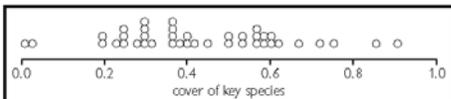
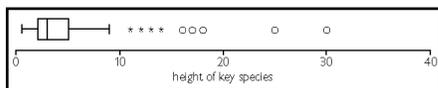
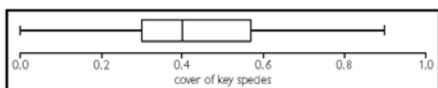
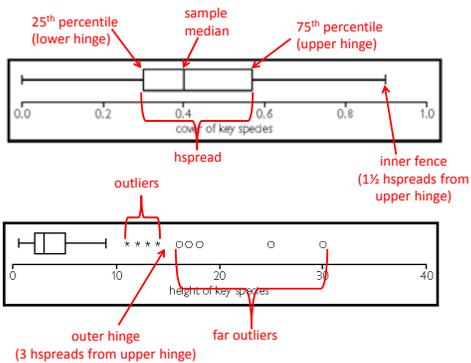


FIGURE 11.6. Dit plot of cover data used to create Figures 11.4 and 11.5. Note that each data point can be distinguished.

Box Plots



Box Plots



Dit Plot + Box Plot

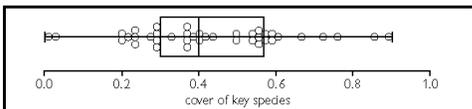
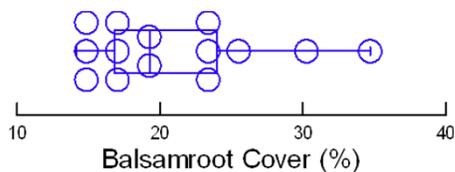


FIGURE 11.9. Overlay of symmetric dit plot on a box plot of cover data. The dits (actual data values) show the underlying data distribution of the box plot.

Combination Box and Symmetrical Dot Density Plot of Line Intercept Cover



Box Plot

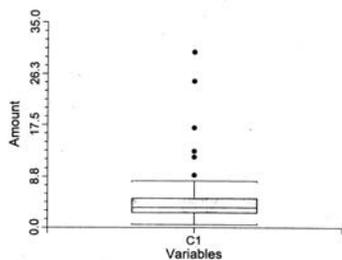
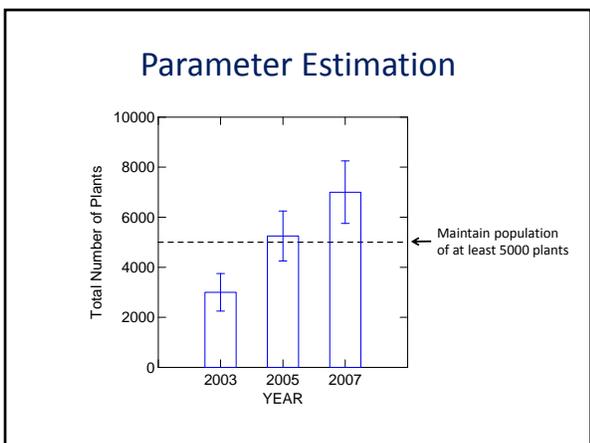


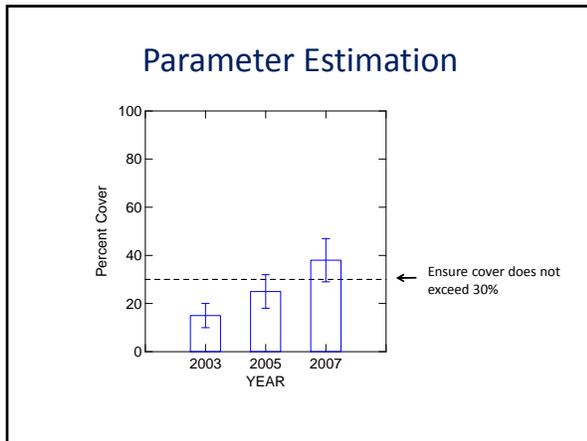
Figure 8.2. A box plot created by the program NCSS Jr. of the same plant height data as used in Figures 8 and 8.1. The box and whiskers are the same as described for the program SYSTAT, but NCSS Jr. does not distinguish between near and far outliers.

Data Analysis

Two Basic Types of Analysis

- **Parameter estimation**
 - For target/threshold management objectives
- **Significance testing**
 - For change/trend management objectives





Significance Testing

Null and Alternative Hypotheses:

H_0 : There has been no change in the parameter of interest

H_A : There has been a change in the parameter of interest

Example

- We estimate the density of rare plant species X in two separate years.
- Each year we take a new random sample of forty 0.25 x 5.0m quadrats and count number of plants in each quadrat.
 - Year 1: Mean (\bar{x}) = 6 plants/quadrat
 - Year 2: Mean (\bar{x}) = 4 plants/quadrat
- We want to determine whether this change is statistically significant or simply due to random variation in the population of all possible quadrats.

P Value

- To do this we must quantify the difference between these two sample means with a *test statistic*.
- When the test statistic is sufficiently large we reject the null hypothesis of no difference between population means and conclude there is in fact a difference.

P Value

- But we must specify how large this test statistic must be for us to reject the null hypothesis.
- To do this we specify a critical or threshold significance level, or P value.
- The P value is the probability of obtaining a value of the test statistic as large or larger than the one computed from the data when in reality there is no difference between the two populations.

Relationship between Test Statistic and P Value

Threshold P (P_{thresh}) = α = false-change error rate in sampling objective

Calculated P (P_{calc}) = actual false-change error rate calculated from sample data

As test statistic \uparrow calculated P \downarrow

This relationship holds true for every test statistic (e.g., t , F , X^2)

P_{thresh} = False Change Error Rate

- In developing a sampling objective we've already specified the threshold P value.
- Now we just need to conduct the statistical test and obtain the calculated P value.
- If $P_{\text{calc}} < P_{\text{thresh}}$ there is a statistically significant difference.
- If $P_{\text{calc}} > P_{\text{thresh}}$ there is no statistically significant difference.

Example (cont'd)

Threshold P = 0.20

This is the same as the false-change error rate in our sampling objective.

It means we are willing to accept a 20% chance of concluding that a change has taken place when it actually hasn't.

Calculated P = 0.125

There is only a 12.5% chance of committing a false-change error based on a statistical analysis of our data.

We therefore reject H_0 .

We conclude that a change has taken place.

Threshold P = 0.20

This is the same as the false-change error rate in our sampling objective.

It means we are willing to accept a 20% chance of concluding that a change has taken place when it actually hasn't.

Calculated P = 0.25

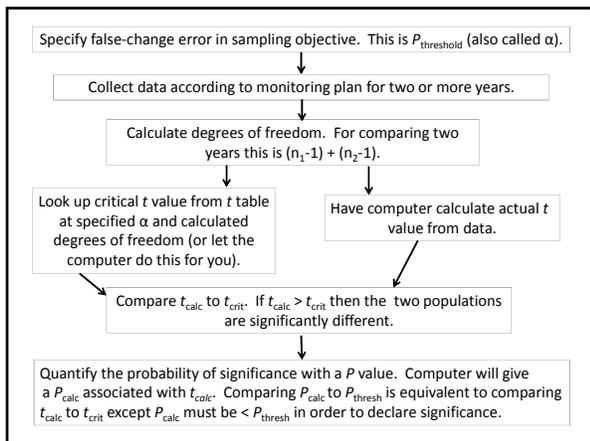
Now there is a 25% chance of committing a false-change error based on a statistical analysis of our data.

In other words, we will be wrong 25% of the time in concluding that a change has taken place based on these data.

We therefore fail to reject H_0 .

Our calculated false-change error rate is greater than the threshold false-change error rate we have said we are willing to accept.

We therefore conclude that no change has taken place.



Exercise 1

Testing for a Difference in the
Number of Flowers/Plant in Two
Populations

Significance Tests to Test for the
Difference between Means or
Proportions of Two or More
Independent Samples

Independent Sample *t* Test (for two samples)

$$t = \frac{\text{difference of sample means}}{\text{standard error of difference of sample means}}$$

As *t* gets larger the *P* value gets smaller.

The *P* value is the probability of obtaining a *t* value as large or larger than the one observed when in reality no change has actually taken place.

Example

- $P_{\text{thresh}} = 0.10$
- Calculated *t* value = 3.21
- Sample size (*n*) = 25 in each of two years
- Degrees of freedom (*v*) = (*n*-1) + (*n*-1) = 48

Critical Values of the *t* distribution

<i>v</i>	$\alpha(2)$ $\alpha(1)$	0.50 0.25	0.20 0.10	0.10 0.05	0.05 0.025	0.02 0.01	0.01 0.005	0.005 0.0025	0.002 0.0010	0.001 0.0005
26		0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27		0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28		0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29		0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30		0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
31		0.682	1.309	1.696	2.040	2.453	2.744	3.022	3.375	3.633
32		0.682	1.309	1.694	2.037	2.449	2.738	3.015	3.365	3.622
33		0.682	1.308	1.692	2.035	2.445	2.733	3.008	3.356	3.611
34		0.682	1.307	1.691	2.032	2.441	2.728	3.002	3.348	3.601
35		0.682	1.306	1.690	2.030	2.438	2.724	2.996	3.340	3.591
36		0.681	1.306	1.688	2.028	2.434	2.719	2.990	3.333	3.582
37		0.681	1.305	1.687	2.026	2.431	2.715	2.985	3.326	3.574
38		0.681	1.304	1.686	2.024	2.429	2.712	2.980	3.319	3.566
39		0.681	1.304	1.685	2.023	2.426	2.708	2.976	3.313	3.558
40		0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
41		0.681	1.303	1.683	2.020	2.421	2.701	2.967	3.301	3.544
42		0.680	1.302	1.682	2.018	2.418	2.698	2.963	3.296	3.538
43		0.680	1.302	1.681	2.017	2.416	2.695	2.959	3.291	3.532
44		0.680	1.301	1.680	2.015	2.414	2.692	2.956	3.286	3.526
45		0.680	1.301	1.679	2.014	2.412	2.690	2.952	3.281	3.520
46		0.680	1.300	1.679	2.013	2.410	2.687	2.949	3.277	3.515
47		0.680	1.300	1.678	2.012	2.408	2.685	2.946	3.273	3.510
48		0.680	1.299	1.677	2.011	2.407	2.682	2.943	3.269	3.505
49		0.680	1.299	1.677	2.010	2.405	2.680	2.940	3.265	3.500
50		0.679	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496

Assessing Site Reclamation/Restoration

Treatment
Area (t)

Reference
Area (r)

Traditional approach:

$$H_0: \text{Cover}_t = \text{Cover}_r$$

$$H_A: \text{Cover}_t \neq \text{Cover}_r$$

Problems with traditional approach:

1. A low power monitoring design would fail to reject H_0 even when populations are quite different—this results in ability to declare success when in fact the treatment has been unsuccessful.
2. A high power monitoring design might reject null hypothesis even when populations are sufficiently similar to conclude success.
3. This approach doesn't explicitly state what level of difference between Cover_t and Cover_r is acceptable.

Assessing Site Reclamation/Restoration

Treatment
Area (t)

Reference
Area (r)

A better approach:

$$H_0: \text{Cover}_t \leq 0.8\text{Cover}_r$$

$$H_A: \text{Cover}_t > 0.8\text{Cover}_r$$

1. This explicitly states what the success criterion is (cover of treatment must be greater than 80% of cover of reference area).
2. Now only a high power design will allow rejection of H_0 and the ability to declare success.
3. Analysis would use a one-tailed test (t test if data are continuous, X^2 test for binomial data).

This is a **one-tailed bioequivalence test**. See Manly, Statistics for Environmental Science and Management, 2001, revised 2nd edition, 2008. Also called **non-inferiority test**.

Site Restoration Along Tuscarora Pipeline

1. Company wanted to use traditional approach and to treat treatment plots and reference plots as two independent samples (i.e., compare the mean of all treatment plots to the mean of all reference plots).
2. This made little sense given the tremendous variation in vegetation types along the pipeline.
3. We had them change the design to use a **one-tailed bioequivalence test** and to treat each treatment and control plot as a pair. A one-tailed paired-sample statistical test would then be used for analysis.

Exercise 2:

Understanding the relationship
between the t statistic and P value
from an independent sample t -test

Analysis of Variance (ANOVA) (for 3 or more independent samples)

$$F = \frac{S_{\text{between}}^2}{S_{\text{within}}^2}$$

S_{between}^2 = between-groups variance: population
variance estimated from sample means

S_{within}^2 = within-groups variance: population
variance estimated as average of sample
variances

ANOVA

- A significant F statistic tells you only that one or more of your means is different from the others. It doesn't tell you which specific means are different from one another.
- Various procedures have been developed to accomplish "mean separation" following an ANOVA.
 - These involve applying various "corrections" to control for the overall experiment-wise error rate (Threshold P value).
 - Recent papers have questioned the use of these correction procedures.

Bonferroni Correction



- One of the most common corrections applied following an ANOVA is the Bonferroni correction.
 - Following a significant ANOVA, a computer program will perform t tests on all pairs of samples.
 - The P_{thresh} (false-change error rate) for the ANOVA is divided by the number of t tests to derive a corrected P_{thresh} to apply to the results of each t test.

How the Bonferroni Correction Works

- For example, if the original $P_{\text{thresh}} = 0.05$ and the population is sampled in 3 years:
 - There are 3 t tests: (1) Year 1 vs. Year 2, (2) Year 1 vs. Year 3, and (3) Year 2 vs. Year 3
 - The original P_{thresh} of 0.05 is corrected by dividing by 3: $0.05/3 = 0.0167$
 - Now P_{calc} must be < 0.0167 in order to declare any of the differences significant.

What's Wrong with Applying the Bonferroni Correction?

- It punishes you for collecting additional information!
- Let's say you collect data in only two years (Year 1 and Year 3) and decide to use a $P_{\text{thresh}} = 0.05$. The mean of year 1 is 7 plants/quadrat and the mean of year 2 is 3 plants/quadrat.
 - You only need one t test so your P_{thresh} is still 0.05 (i.e., no correction is needed).
 - The test yields a $P_{\text{calc}} = 0.031$, so you conclude a change has taken place between years.

- Now suppose we also sampled in Year 2 and came up with a mean of 4 plants/quadrat. So we have the following values:
 Year 1 = 7 plants/quadrat
 Year 2 = 4 plants/quadrat
 Year 3 = 3 plants/quadrat
- The P_{calc} for the Year 1 vs. Year 3 comparison is still 0.031 but because we've now done 3 sets of comparisons we must now compare this to a $P_{thresh} = 0.0167$ and conclude there is no significant difference between Year 1 and Year 3!

Another Example of Problems with the Bonferroni Correction

Table 1. Results of a hypothetical experiment testing the effects of grazing on ten plant species. Response refers to the difference between grazed (G) and ungrazed (U) plots. * indicates significance at $p < 0.05$.

Plant Species	Response	Significance
Grass # 1	G < U	0.45
Grass # 2	G > U	0.67
Grass # 3	G < U	0.93
Grass # 4	G < U	0.25
Grass # 5	G > U	0.53
Forb # 1	G < U	0.04*
Forb # 2	G < U	0.02*
Forb # 3	G < U	0.03*
Forb # 4	G < U	0.01*
Forb # 5	G < U	0.02*

Excerpted from:
 Moran, M. D. 2003. Arguments for rejecting the sequential Bonferroni in Ecological studies. *Oikos* 100: 403-405.

Bonferroni and Other Post-Hoc Corrections—Don't Use Them

- Because of the problems with the Bonferroni and other corrections we recommend not using them.
- Because comparisons between years are essentially planned (as opposed to post-hoc, unplanned comparisons), you are justified in using the uncorrected P_{calc} value.
- This also makes the use of the ANOVA to test differences between 3 or more years unnecessary—simply use separate t tests (or—depending on the type of data—other two-sample tests such as the chi-square test).

ANOVA Printout from a Statistical Program

DENSDATA.DMD Tuesday, March 26, 1996 1:01:14 PM

	1989	1991	1993
1	31	26	24
2	25	20	18
3	6	1	1
4	9	4	2
5	24	19	17
6	18	13	11
7	30	25	23
8	25	20	18
9	0	0	1
10	8	3	1
11	42	37	36
12	32	27	25
13	21	16	14
14	17	12	10
15	14	9	7
16	8	3	1
17	27	22	20
18	29	24	22
19	34	29	27
20	37	32	30
21	25	20	18
22	34	29	27
23	21	16	14
24	18	13	11
25	10	5	3
26	26	21	19
27	11	6	4
28	17	12	10
29	21	16	14
30	23	18	16

StatSoft for Windows Tuesday, March 26, 1996 12:42:09 PM

***** Statistics Report *****

	1989	1991	1993
Sample size (N)	30	30	30
Sum missing	0	0	0
Minimum	0.0000	0.0000	1.0000
Maximum	42.0000	37.0000	35.0000
Std deviation	10.1361	9.8067	9.5390
Variance	102.7402	96.1713	90.9931
Std error	1.6806	1.7904	1.7416
C.V.	47.2178	58.9581	64.4529
Mean	21.6667	16.6333	14.8000
Sum	649.0000	499.0000	444.0000
Sum squares	16924.0000	11089.0000	9210.0000
Median	22.0000	17.0000	15.0000
Mode	23.0667	17.7333	15.4000
Kurtosis	-0.4873	-0.7240	-0.7721
Coeff kurtosis	2.5127	2.2760	2.2279
Skewness	-0.1376	0.0060	0.0932
Coeff skewness	-0.0488	0.0020	0.0466
Percentiles:			
10	8.0000	3.0000	1.0000
25	13.2500	8.2500	4.2500
50	22.0000	17.0000	15.0000
75	29.2500	24.2500	22.2500
90	34.0000	29.0000	27.0000
Quartiles:			
First quartile:	13.2500	8.2500	4.2500
Second quartile:	22.0000	17.0000	15.0000
Third quartile:	29.2500	24.2500	22.2500
95.00% Confidence Interval:			
Lower limit	17.6918	12.9715	11.2381
Upper limit	25.2515	20.2952	18.3619

***** The End *****

Expected Values Under the Null Hypothesis

- The null hypothesis of no change is the same as saying the true proportion is the same in both years.
- If the true proportion is the same, then the values obtained in each year can be considered estimates of the same proportion.
- Thus we can use the total frequencies in the right hand column of the contingency table to determine the expected values under the null.

	2000	2004	Totals
Present	123 (0.31)	157 (0.39)	280 (0.35)
Absent	277 (0.69)	243 (0.61)	520 (0.65)
Totals	400 (1.00)	400 (1.00)	800 (1.00)

Expected Values Under the Null Hypothesis

- Thus, in both 2000 and 2004, 0.35 x 400 quadrats, or 140 quadrats (or points) would be expected to contain the species, and 0.65 x 400 quadrats, or 260 quadrats (or points) would be expected to not contain the species.

	2000	2004	Totals
Present	140	140	280
Absent	260	260	520
Totals	400	400	800

Chi-Square Test

$$X^2 = \sum \frac{(O - E)^2}{E}$$

- Where: X^2 = the chi - square statistic
 Σ = summation symbol
 O = number observed
 E = number expected

Chi-Square Test

Applying this formula to our example we get:

$$X^2 = \frac{(123-140)^2}{140} + \frac{(277-260)^2}{260} + \frac{(157-140)^2}{140} + \frac{(243-260)^2}{260}$$

$$= 2.06 + 1.11 + 2.06 + 1.11$$

$$= 6.34$$

We then compare the X^2 value of 6.34 to a table of critical values of the chi-square statistic (Appendix 5) to see if this value is sufficiently large to be significant.

Degrees of freedom: $v = (r-1)(c-1)$

Where: r = number of rows in contingency table
 c = number of columns in contingency table

Don't Use the Yates Correction

- Some authors (e.g., Zar 1996) advocate the Yates correction for a 2 x 2 contingency table:

$$X^2 = \sum \frac{(|O - E| - \frac{1}{2})^2}{E}$$

- Other authors (e.g., Steel and Torrie 1980, Sokal and Rohlf 1981) say it's overly conservative and recommend against it.
- Simulations by Dan Salzer show it's not needed.
- Most experts now agree it should not be used.

SYSTAT Printout of a chi-square analysis.

Counts

PRESENCE\$(rows) by YEAR\$(columns)

	2001	2004	Total
Absent	166	144	310
Present	134	156	290
Total	300	300	600

Table of Counts and Percents

PRESENCE\$(rows) by YEAR\$(columns)

	2001	2004	Total
Absent	166(27.667%)	144(24.000%)	310(51.667%)
Present	134(22.333%)	156(26.000%)	290(48.333%)
Total	300(50.000%)	300(50.000%)	600(100.0%)

Chi-Square Tests of Association for PRESENCE\$ and YEAR\$

Test Statistic	Value	df	p-Value
Pearson Chi-Square	3.230	1,000	0.072
Yates Corrected Chi-Square	2.943	1,000	0.086

Exercise 3

Understanding the relationship between the χ^2 statistic and P value from a chi-square test

SYSTAT Printout of Exercise 3 Example 1.

Counts

PRESENCE\$(rows) by YEAR\$(columns)

	2001	2004	Total
Absent	166	144	310
Present	134	156	290
Total	300	300	600

Table of Counts and Percents

PRESENCE\$(rows) by YEAR\$(columns)

	2001	2004	Total
Absent	166(27.667%)	144(24.000%)	310(51.667%)
Present	134(22.333%)	156(26.000%)	290(48.333%)
Total	300(50.000%)	300(50.000%)	600(100.0%)

Chi-Square Tests of Association for PRESENCE\$ and YEAR\$

Test Statistic	Value	df	p-Value
Pearson Chi-Square	3.230	1.000	0.072
Yates Corrected Chi-Square	2.945	1.000	0.088

Larger contingency tables for more than two years

- Although you can use larger contingency tables to accommodate more than 2 years (e.g., 2 x 3 table for 3 years, 2 x 4 table for 4 years, etc.), this suffers from the same problem as the ANOVA: a significant result will tell you only that one or more of the years is different—it won't tell you which specific years are different from the others.
- For this reason we recommend you conduct pairwise chi-square tests (with no Bonferroni or other correction).

**Permanent Quadrats, Transects,
and Points ¹**

**The Use of Paired-Sample
Significance Tests**

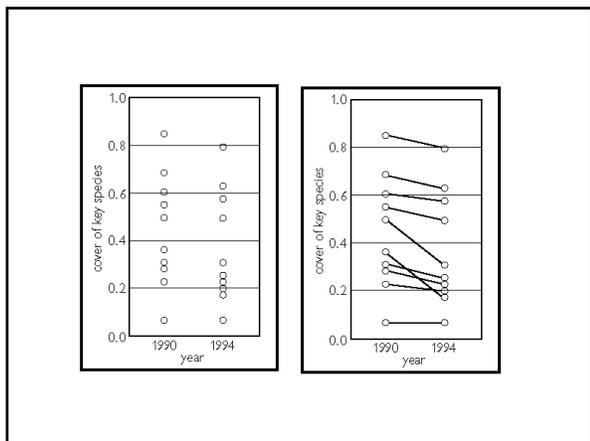
¹This doesn't apply to points when the points are the sampling units

Independent vs. Paired Samples

- Independent samples are ones in which different sets of sampling units are selected randomly (or systematically with random starts) in each year of measurement.
- Now consider the case in which sampling units are randomly selected only in the first year.
 - The sampling units are permanently marked.
 - Same sampling units measured in subsequent year.
 - The two samples are no longer independent—they are dependent or paired.

Paired t test: Use it when you can

- Paired tests, such as the paired t test, are often much more powerful than independent sample tests in detecting change.
- The next slide illustrates why this is so.



Paired Sample *t* Test

- An independent sample *t* test run on these two samples yields a calculated *t* value of 0.617 and a *P* value of 0.55—not significant.
- An independent sample *t* test is not appropriately applied to these paired data.
 - Even if we could conduct this test, we wouldn't want to.
 - The paired *t* test ignores the between-transect variability and looks at the *differences* between the 1990 and 1994 values for each of the transects.

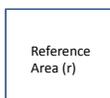
transect number	cover in 1990	cover in 1994	difference between 1990 and 1994
1	0.22	0.20	-0.02
2	0.32	0.26	-0.06
3	0.06	0.06	0.00
4	0.86	0.80	-0.06
5	0.62	0.58	-0.04
6	0.54	0.50	-0.04
7	0.50	0.32	-0.18
8	0.28	0.24	-0.04
9	0.36	0.18	-0.18
10	0.68	0.64	-0.04
			mean difference -0.07
			standard error 0.02

Paired sample *t* test results:
 Calculated *t* value: 3.34
 Calculated *P* value: 0.009

Paired Sample *t* Test

- The higher the degree of positive correlation between the pairs of sampling units, the more powerful the test.
- Can measure this with a correlation coefficient.
- Closer the coefficient is to 1.0 (perfect correlation), the more powerful the test.
- Can apply a paired sample *t* test even when there is no correlation—but won't be as powerful as an independent sample *t* test.

Assessing Site Reclamation/Restoration



Traditional approach:

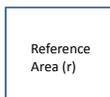
$$H_0: \text{Cover}_t = \text{Cover}_r$$

$$H_A: \text{Cover}_t \neq \text{Cover}_r$$

Problems with traditional approach:

1. A low power monitoring design would fail to reject H_0 even when populations are quite different—this results in inability to declare success when in fact the treatment has been unsuccessful.
2. A high power monitoring design might reject null hypothesis even when populations are sufficiently similar to conclude success.
3. This approach doesn't explicitly state what level of difference between Cover_t and Cover_r is acceptable.

Assessing Site Reclamation/Restoration



A better approach:

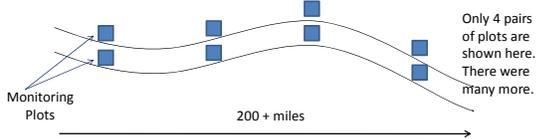
$$H_0: \text{Cover}_t \leq 0.8\text{Cover}_r$$

$$H_A: \text{Cover}_t > 0.8\text{Cover}_r$$

1. This explicitly states what the success criterion is (cover of treatment must be greater than 80% of cover of reference area).
2. Now only a high power design will allow rejection of H_0 and the ability to declare success.
3. Analysis would use a one-tailed test (*t* test if data are continuous, χ^2 test for binomial data).

This is a **one-tailed bioequivalence test**. See Manly, *Statistics for Environmental Science and Management*, 2001, revised 2nd edition, 2008.

Site Restoration Along Tuscarora Pipeline



1. Company wanted to use traditional approach and to treat treatment plots and reference plots as two independent samples (i.e., compare the mean of all treatment plots to the mean of all reference plots).
2. This made little sense given the tremendous variation in vegetation types along the pipeline.
3. We had them change the design to use a one-tailed bioequivalence test and to treat each treatment and control plot as a pair. A **one-tailed paired-sample statistical test** would then be used for analysis.

Repeated Measures ANOVA

- You can compare 3 or more years using a repeated measures ANOVA.
- Just as with the independent sample ANOVA, a significant result tells you only that 1 or more years is different from the others.
- We recommend you use paired sample *t* tests to compare pairs of years.
- No Bonferroni correction is necessary (unlike guidance in MMPP TR).

Paired Sample Testing for Proportions: McNemar's Test

- Can be used for frequency data when quadrats are paired.
- Theoretically could be used for paired points, but this is impractical.
- Uses a 2 x 2 contingency table like the chi square test for two independent samples but the data entered into the McNemar table differs from the independent chi-square table.

Example

100 quadrats in macroplot in each of 2 years
 $P_{\text{thresh}} = 0.10$
 Year 1: 60 quadrats have Species X in them
 Year 2: 50 quadrats have Species X in them

	Year 1	Year 2	Totals
Present	60	50	110
Absent	40	50	90
Totals	100	100	200

Permanent quadrats
 McNemar's $\chi^2 = 8.100$
 $P_{\text{calc}} = 0.0044$

Temporary quadrats
 $\chi^2 = 2.020$
 $P_{\text{calc}} = 0.155$

		Year 1	
		Present	Absent
Year 2	Present	50	0
	Absent	10	40

Formula for McNemar's Test

		Year 1	
		Present	Absent
Year 2	Present	50	0
	Absent	10	40

$$X^2_{mcn} = \frac{(AP - PA)^2}{AP + PA}$$

Applying the Finite Population Correction Factor to the Results of a Statistical Test

- If you've sampled more than 5% of a population you should apply the FPC to the results of a significance test.
- The procedure for applying the FPC depends on the nature of the test statistic.

Applying the FPC to Tests that Use the t Statistic

t statistic from a t test (either an independent sample or a paired t test) is 1.645 and in each of 2 years you sampled $n = 26$ out of a total of $N = 100$ possible quadrats.

$$t' = \frac{t}{\sqrt{1 - (n/N)}}$$

$$t' = \frac{1.645}{\sqrt{1 - (26/100)}} = 1.912$$

You then need to look up the P value for the appropriate degrees of freedom in a t table or use the program NCCS Probability Calculator.

Applying the FPC to Tests that Use the Chi-square Statistic

X^2 statistic from either a chi-square or McNemar's test is 2.706 and in each of 2 years you sampled $n = 77$ out of a total of $N = 300$ possible quadrats.

$$X'^2 = \frac{X^2}{1 - (n/N)}$$

$$X'^2 = \frac{2.706}{1 - (77/300)} = 3.640$$

You then need to look up the P value for the appropriate degrees of freedom in a X^2 table or use the program NCCS Probability Calculator.

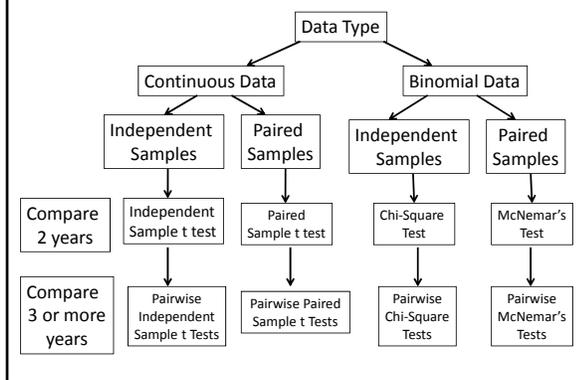
Applying the FPC to Tests that Use the F Statistic

- See MMPP on pp 250-251 for the formula.
- Not shown here because you will probably use pair-wise t tests in lieu of an ANOVA.

Exercise 4

Adjust the Chi-Square Statistic Using
the Finite Population Correction
Factor

Flow Chart of Statistical Tests



Exercise 5

What Statistical Procedure Would
You Use?

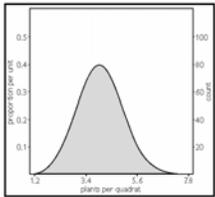
Assumptions Regarding the Statistics Discussed So Far

Parametric Statistics

- Except for the chi-square and McNemar's tests, all of the statistics we've discussed so far are *parametric* statistics.
- Parametric statistics get their name because they are used to estimate population parameters such as means and totals.
- *t* tests, both independent and paired, ANOVA, and confidence intervals calculated using a *t* table are all parametric statistics.

The use of parametric statistics requires that several assumptions be met, at least approximately (no monitoring data will meet these assumptions exactly):

1. That the population being sampled follows a normal distribution.
 - This assumption holds both for the calculation of confidence intervals and for the use of *t* tests and ANOVAs. (For paired *t* tests the *differences* between sampling units should come from a population that follows a normal distribution.)



2. That the sampling units are drawn from populations in which the variances are the same even if the means change from the first year of measurement to the next.

- This assumption, called homogeneity of variances, applies to *significance* tests to detect changes in means.
- This assumption is rarely met by natural resource monitoring data—the variance always tends to increase as the mean increases.

3. That the sampling units are drawn in some random manner from the population.

- This assumption hold both for the calculation of confidence intervals and for significance tests.
- This assumption must also be met when *nonparametric* statistics are used.

How to Check for Normality and Homogeneity of Variance

- Although there are tests of normality, it is often most effective to look at a graphical analysis of your data—we showed you the normal probability plot earlier.
- Several tests are available to determine if the variances of two or more samples are equal, but none of these is very reliable.
 - The most well known, Bartlett’s test, is not recommended because it is very sensitive to departures from normality.
 - Zar recommends no test be used because ANOVA (and *t* test) are robust to departures from this assumption.

What do I do if my data don’t meet the assumptions of normality and homogeneity of variances?

1. **Nothing.** Few if any real data comes from a population that’s normal, or even quasi normal (Koch and Link 1970).

- The only consequences of failure to meet the normality assumption is some distortion of theoretical risk levels and a reduction in the efficiency of estimation.
- These problems are far less serious than the failure to meet the assumption of randomness.
- Both *t* tests and ANOVAs are robust to moderate departures from both assumptions.

2. Increase your sample size. According to Mattson (1981) a sample size of at least 100 sampling units will ensure against problems resulting from severe departures from normality.

- This is conservative.
- Less severe departures from normality will not require as large a sample.
- We'll talk more about this shortly.

3. Transform your data. Data are often converted to another scale prior to analysis in order to more closely meet the assumption of normality and homogeneity of variances.

- Covered in many statistical text books.
- Their utility for vegetation monitoring is limited because of several problems:
 - The most common transformations are seldom helpful.
 - Estimating means, variances, and confidence intervals in their transformed scale leads to biased estimates when data are transformed back to original scale.
 - May be difficult to understand or apply the results of statistical analyses expressed in the transformed scale.

4. Use nonparametric statistics. If you are greatly concerned whether your data meet the assumptions you can use this class of statistics that do not require these assumptions.

- These still require that data be collected randomly.
- Nonparametric statistics require other assumptions that are often not discussed.
- They aren't as powerful as parametric statistics when the assumptions of normality and homogeneity of variances are approximately met.
- We'll talk more about this shortly.

5. Use statistical analyses based on resampling.

Resampling methods (also called computer-intensive methods) are becoming more and more popular with ecologists and other scientists.

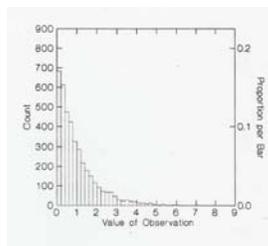
- These methods can be used to calculate confidence intervals and to conduct significance testing.
- Like nonparametric statistics, these methods often don't require the assumptions of normality or homogeneity of variances.
- They are often more powerful than traditional nonparametric statistics.
- Will cover in more detail.

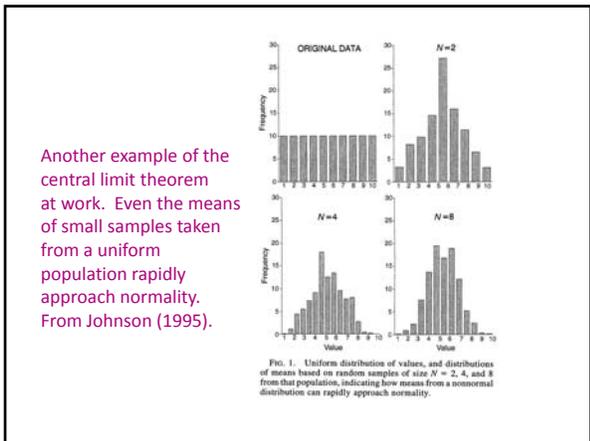
When should I worry about using parametric statistics?

- If you follow the guidance in this class you probably don't have to worry about using parametric statistics.
- We've already stated that both *t* tests and ANOVAs are robust to moderate depatures from the assumptions of normality and homogeneity of variances.
 - There is a *t* test that can be used that allows for differences in variances (see pooled vs. separate variance *t* test paper in Statistics section).
- MMPP (page 253) gives some rules of thumb.

A Small Experiment

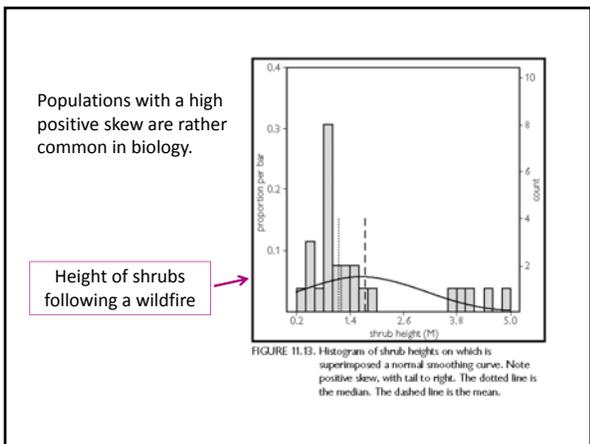
- I created a simulated population of 4000 observations.
- The population follows an exponential distribution.
- Pop mean = 0.995
- Pop SD = 0.962
- Note large # of small values and long tail to right caused by a small # of large values (one value is 8, more than 8 times the SD)





Nonparametric Statistics

- Chi-square and McNemar’s tests are nonparametric.
- Parametric statistics estimate means and standard deviations, and/or conduct significance tests on means with formulas employing standard deviations.
 - CI’s around means and population totals are calculated using the standard deviation (converted to a standard error by dividing by the sample size) and therefore involve parametric statistics.
 - t tests and ANOVAs are parametric procedures.



Nonparametric procedures conduct tests on the ranks of the observations, not on the values of the observations themselves.

This is equivalent to "throwing away" information.

Height (m)	Rank
0.35	1
0.40	2
0.50	3
0.55	4
0.75	5
0.90	6
1.00	7
1.10	8
1.30	9
4.50	10
5.10	11

TABLE 11.1 Sample of 11 shrub. Heights ranked from smallest to largest.

Why not use nonparametric statistics all the time?

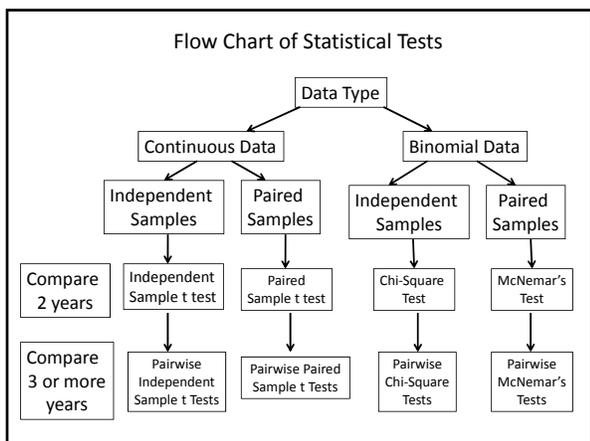
- When necessary assumptions approximated parametric statistics are more powerful than nonparametric analogues.
- Other than the use of resampling techniques (discussed next) there is no nonparametric method available to calculate confidence intervals around means and totals, the two parameters often of the most interest in monitoring.
- Nonparametric statistics have their own set of assumptions that may be problematic.

Table 2 -- Matrix of statistical significance tests.

Parametric and nonparametric significance tests corresponding to type of data and purpose of test. Note that for frequency (present-absent) data only nonparametric tests are available.

Purpose of Test	Parametric Test	Nonparametric Test
Testing for change between two years; samples independent; not frequency data	Independent sample <i>t</i> test	Mann-Whitney <i>U</i> test
Testing for change between two years; samples paired; not frequency data	Paired <i>t</i> test	Wilcoxin's signed rank test
Testing for change between two years; samples independent; frequency data		Chi-square test (2 x 2 contingency table)
Testing for change between two years; samples paired; frequency data		McNemar's test
Testing for change between three or more years; samples independent; not frequency data	Analysis of variance	Kruskal-Wallis test
Testing for change between three or more years; same samples measured each year; not frequency data	Repeated measures analysis of variance	Friedman's test
Testing for change between three or more years; samples independent; frequency data		Chi-square test (2 x 3 contingency table)

Except for the Chi-square and McNemar's tests I don't use any of these non-parametric tests. If I have concerns about the appropriateness of a parametric test I'll use a randomization test.



Statistical Analysis Based on Resampling

What is resampling?

Example 1: What is the probability of throwing 3 heads in a row?

This is mathematically easy to figure out:

$$P = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{8} = 0.125$$

Let's do this empirically using resampling.

```

Start execution.
Where heads.ata
"What are the odds of throwing three heads in a row with a fair coin?"
MAXSIZE DEFAULT 5000      Sets default to allow for 5000 samples
REPEAT 5000                5000 samples
GENERATE 3 1,2 a           Creates 3 random numbers between 1 and 2 and stores in a
SUM a c                    Sums the 3 numbers in a
SCORE c z                  Keeps track of the sum in c and stores in z
END
COUNT z =6 k             Counts the number of times the sum of the three numbers
                           was 6 (if the number 2 is heads, then a sum of 6 represents
                           3 heads in a row)
DIVIDE k 5000 kk           Converts to a proportion and stores in kk
PRINT kk                  Prints the proportion stored in kk

KK = 0.1198                Empirical probability

Successful execution. (2.6 seconds)

-----

Start execution.
Where heads.ata
"What are the odds of throwing three heads in a row with a fair coin?"
MAXSIZE DEFAULT 5000
REPEAT 5000
GENERATE 3 1,2 a
SUM a c
SCORE c z
END
COUNT z =6 k
DIVIDE k 5000 kk
PRINT kk

KK = 0.1306

Successful execution. (2.5 seconds)
  
```

15000 samples

```

Start execution.
'Here heads sta
'What are the odds of throwing three heads in a row with a fair coin?
'
MAXSIZE DEFAULT 15000
REPEAT 15000
  GENERATE 3 1,2 a
  SUM a c
  SCORE c z
END
COUNT z =6 k
DIVIDE k 15000 kk
PRINT kk

KK = 0.1238
Successful execution (7.2 seconds)

```

15000 samples

```

Start execution.
'Here heads sta
'What are the odds of throwing three heads in a row with a fair coin?
'
MAXSIZE DEFAULT 15000
REPEAT 15000
  GENERATE 3 1,2 a
  SUM a c
  SCORE c z
END
COUNT z =6 k
DIVIDE k 15000 kk
PRINT kk

KK = 0.12313
Successful execution (6.8 seconds)

```

Example 2. What is the probability in a class of 25 people, at least two would have the same birthday?

The calculations are *much* more complex than those needed to determine the probability of throwing 3 heads in a row.

But if we do this empirically though resampling, the concept is relatively easy to understand.

5000 samples

```

Two runs of 5000 samples using RESAMPLING STATS to determine the
probability that 2 or more people in a class of 23 have the same birthday.

Start execution.
'birthday sta
'What are the odds that two or more
'persons in a group of 23 people will
'have a birthday on the same day?
'
MAXSIZE DEFAULT 5000
REPEAT 5000
  GENERATE 23 1,31 a
  MULTIPLE a == 2 j
  SCORE j z
END
COUNT z == 1 k
DIVIDE k 5000 kk
PRINT kk

KK = 0.5712
Successful execution (4.3 seconds)

```

5000 samples

```

Start execution.
'birthday sta
'What are the odds that two or more
'persons in a group of 23 people will
'have a birthday on the same day?
'
MAXSIZE DEFAULT 5000
REPEAT 5000
  GENERATE 23 1,31 a
  MULTIPLE a == 2 j
  SCORE j z
END
COUNT z == 1 k
DIVIDE k 5000 kk
PRINT kk

KK = 0.567
Successful execution (3.7 seconds)

```

- This resampling was conducted using the standalone version of Resampling Stats.
- The standalone version is no longer available, but has been replaced with an add-in program for Excel.

- ### Bootstrapping
- Bootstrapping is sampling *with* replacement.
 - For example, let's say we put the following numbers into a hat: 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10
 - We then take a number out of the hat, record it, and put in back in the hat.
 - Thus a possible sample of size 10 could be: 1, 1, 2, 4, 4, 7, 8, 8, 10, 10.
 - Bootstrapping is often used to calculate confidence intervals around an estimate.

Bootstrapping

Plant Heights

25	Mean = 8.23
4	Median = 10.27
30	
4.5	
4	
1.75	
2	
4	
2.5	
4.5	

Parametric 95% confidence interval: 0.874 to 15.576

Bootstrap Confidence Interval:

Take a large number of samples (e.g., 10,000) with replacement and use the distribution of these to determine the values at which the two tails of the distribution correspond to the 2.5 and 97.5 percentiles.

For this example: 95% confidence interval: 3.175 to 15.175

In real life you'd often want to collect more than 10 sampling units, depending on the distribution of your data. A sample of 40 works well even for data corresponding to an exponential distribution (for CI's around a mean).

Randomization Tests

- Randomization tests (also called permutation tests) involve sampling *without* replacement.
- They are used in lieu of parametric or standard nonparametric tests to test whether means, medians, variances, etc. differ between populations.
- There is no assumption of normality or homogeneity of variances.

Randomization Tests

Eastern horned lizard gut contents (in mg of dry biomass)

Size class 1 (adult males and yearling females)
n=24

256	0	6	34	0	332
209	44	0	13	32	0
0	49	0	0	0	31
0	117	75	90	205	0

Size class 2 (adult females)
n=21

0	0	843	311	19	432
89	163	0	232	142	
0	286	158	179	100	
0	3	443	179	0	

Difference between means of the two groups = -108.6

Randomly reallocate 24 of the sample weights to "size class 1" and 21 to "size class 2" (this is sampling *without* replacement) and calculate the difference in means between the two size classes.

Repeat this a large number of times (e.g., 5000 times), keeping track of the difference in means each time.

Compare the original mean difference to the set of randomization mean differences.

If the original mean difference looks like a typical value from the distribution of randomization mean differences then conclusion is no difference between the two size classes.

If, however, the original mean difference is unusually large then the difference is unlikely to have arisen if the null hypothesis were true. Conclusion is then that there is a difference between the two size classes.

Results from statistical tests:

t test : $P = 0.027$
Mann-Whitney non-parametric test: $P = 0.080$
Randomization test (5000 randomizations): $P = 0.018$

Graphing the Results of Data Analysis

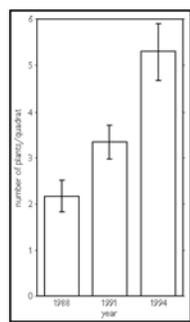


FIGURE 11.14. Bar chart of mean number of plants of the key species per 0.5m x 4.0m quadrat. Error bars are 90% confidence intervals. In each year n = 100.

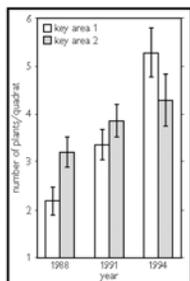
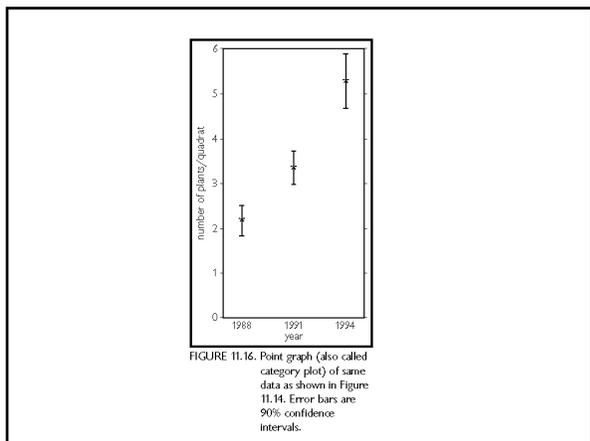
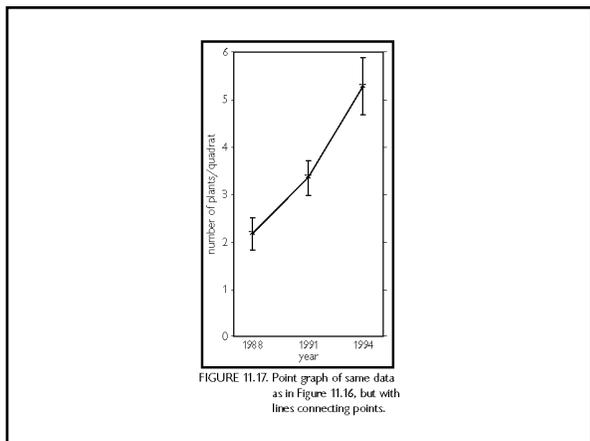
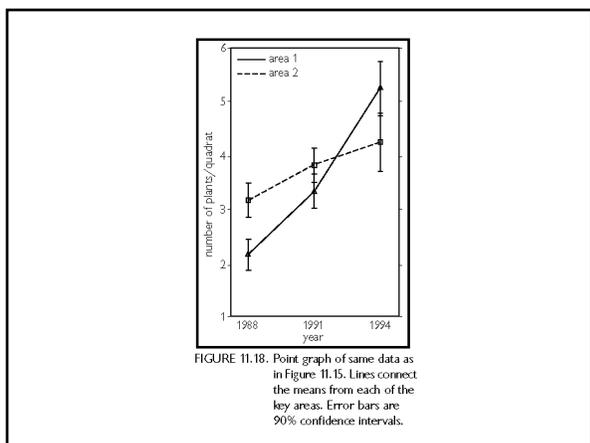
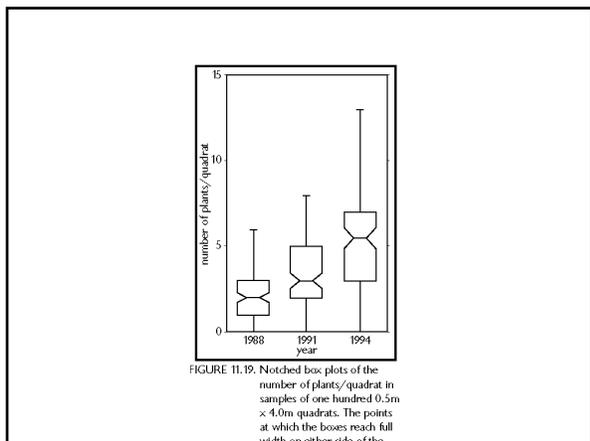


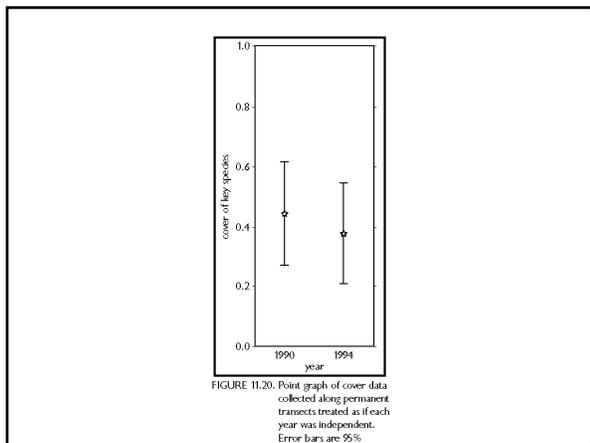
FIGURE 11.15. Side-by-side bar chart of mean number of plants of the key species per 0.5m x 4.0m quadrat, at key area 1 and key area 2. Error bars are 90% confidence intervals. All bars represent n = 100.

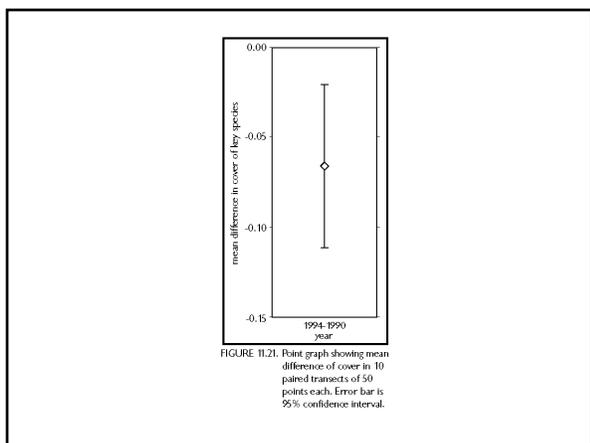






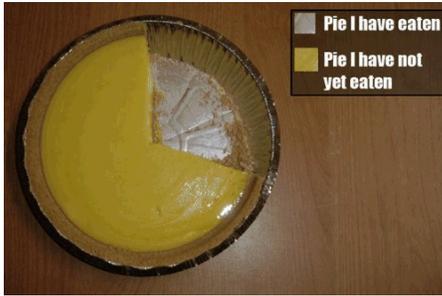






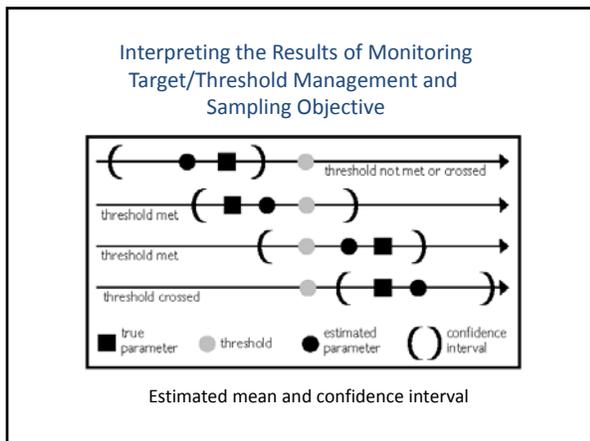
Pie Charts: Don't Use Them!

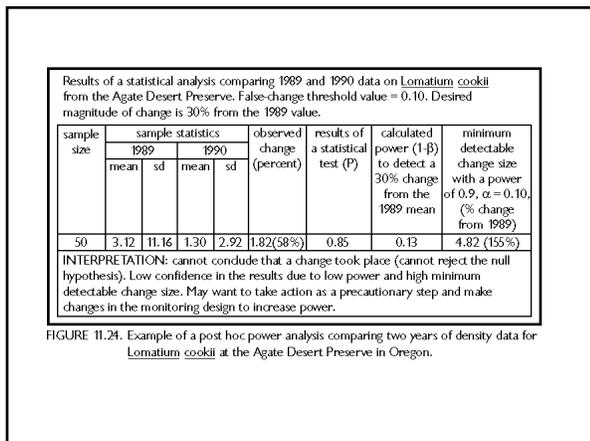
- Edward Tufte: Given their low data-density and failure to order numbers along a visual dimension, pie charts should never be used.
- Pie charts do not allow you to illustrate uncertainty (i.e., error bars).
- There is, however, one (and only one) pie chart that meets with Tufte's approval:



The only acceptable pie chart according to Edward Tufte

Interpreting the Results of Monitoring





Exercise 6

Interpreting the Results of Significance Tests

