BIO 682
Nonparametric Statistics
Spring 2010

Steve Shuster

http://www4.nau.edu/shustercourses/

Lecture 3

Homework!

Today: 20 January 2010

Problem Set #1: Due 27 January 2010

January 2010

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

February 2010

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28						

Problem Set #2: Due 3 February 2010

Another Example:

1. The debate in 1970s between ecologists who thought community structure was more influenced by stochastic processes than by competition.

1. *Competitionists* - structure differences *are* due to competition – thus accepted a higher α (and a lower β).

2. *Stochastitians* - wanted to avoid accepting causality too readily: perjoratively known as "β-maximizers“ (low α; high β)

Things to Remember

1. To figure power it is necessary to:

a. Define an alternative hypothesis

b. in the above case, H_1: $p = 2q$

2. Then a frequency distribution is generated for observations for this result.

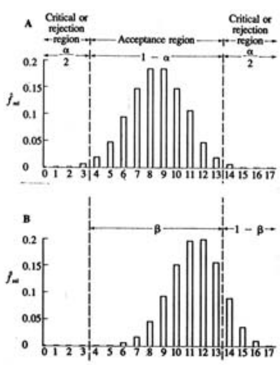
Defining β

3. The area that falls within the critical region is β.

a. Remember that β is *the probability of making a Type II error*.

b. The area that falls outside is (1 - β);

c. This is **statistical power**



Note That:

1. How much of the distribution for H_1 overlaps that of H_o (87%)

a. Thus, for this sample size and for the parameters of H_1, the two hypotheses are *virtually identical*.

3. Note too, that in this case, decreasing α (increasing acceptance region)

also increases β.

a. defeats the purpose of maximizing POWER

Scales of Measurement

A. Nominal Data

1. The weakest level of information.

a. Uses numbers or symbols to classify individuals, objects, etc.

b. Also known as *classificatory scale*.

Nominal Data

2. Formal properties

a. All cases are equal within class, mutually exclusive with other classes.

Nominal Data: Examples

a. Telephone numbers (at least within a geographical area), ssn#.

b. Difference in sign (+, -).

c. It is possible to have totals in each category.

Nominal Data

1. Types of tests:

a. Goodness of fit.

b. Measures of association.

c. Indices of diversity

1. Shannon-Weiner (Shannon-Weaver) Index of diversity.

2. For examining the distribution of observations among categories.

Shannon-Weiner Index of Diversity

$$\mathbf{H} = -\sum_{i=1}^{k} p_i \log p_i$$

where: k = the number of categories;

p_i = proportion of sample in category i,

$p_i = (f_i/n_i)$

f_i = number of cases in category i

n_i = sample size of category i

N = total cases

An Easier Method Is,

$$\mathbf{H} = [\mathbf{N} \log \mathbf{N} - \sum f_i \log f_i\,]/\mathbf{N}$$

where: f_i = number of cases in category i

N = total cases

This eliminates necessity for calculating p_i

An Estimate of Eveness

$$J = H / H_{max}$$

where H_{max} is the maximum diversity possible.

a. Perhaps a better estimator because the magnitude of H is affected by:

1. the number of categories
2. the distribution of data.

Other Problems

1. The SW index is *un-standardized*.

a. Makes its use somewhat suspect unless standardized across analyses.

2. Confidence limits are not clearly defined.

a. Therefore, it is difficult to make comparisons across situations.

The S-W Equation

$$H = \frac{n \log n - \sum f_i \log f_i}{n}$$

Note that here n = N.

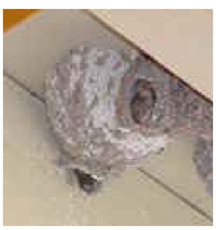
Nests in Different Locations

Sample 1	
Vines	5
Eaves	5
Branches	5
Cavities	5

$$H = \frac{n \log n - \sum f_i \log f_i}{n}$$

$$[20 \log 20 - (5 \log 5 + 5 \log 5 + 5 \log 5 + 5 \log 5)]/20$$
$$= [26.0206 - (3.4949 + 3.4949 + 3.4949 + 3.4949)]/20$$
$$= 12.0410/20 = 0.602$$
$$H_{max} = \log 4 = 0.602$$
$$J = \frac{0.602}{0.602} = 1.000$$

Nests in Different Locations



Sample 2	
Vines	1
Eaves	1
Branches	1
Cavities	17

$$H = \frac{n \log n - \sum f_i \log f_i}{n}$$

$$[20 \log 20 - (1 \log 1 + 1 \log 1 + 1 \log 1 + 17 \log 17)]/20$$
$$= [26.0206 - (0 + 0 + 0 + 20.9176)]/20$$
$$= 5.1030/20 = 0.255$$
$$H_{max} = \log 4 = 0.602$$
$$J = \frac{0.255}{0.602} = 0.424$$

Nests in Different Locations

Sample 3	
Vines	2
Eaves	2
Branches	2
Cavities	34

$$H = \frac{n \log n - \sum f_i \log f_i}{n}$$

$$[40 \log 40 - (2 \log 2 + 2 \log 2 + 2 \log 2 + 34 \log 34)]/40$$
$$= [64.0824 - (0.6021 + 0.6021 + 0.6021 + 52.0703)]/40$$
$$= 10.2058/40 = 0.255$$
$$H_{max} = \log 4 = 0.602$$
$$J = \frac{0.255}{0.602} = 0.424$$

Example 4.3 Indices of diversity for nominal scale data. The nesting sites of sparrows.

Category (i)	Observed Frequencies (f_i)		
	Sample 1	Sample 2	Sample 3
Vines	5	1	2
Eaves	5	1	2
Branches	5	1	2
Cavities	5	17	34
$H = \frac{n \log n - \sum f_i \log f_i}{n}$	$[20 \log 20 - (5 \log 5 + 5 \log 5 + 5 \log 5 + 5 \log 5)]/20 = [26.0206 - (3.4949 + 3.4949 + 3.4949 + 3.4949)]/20 = 12.0410/20 = 0.602$; $H_{max} = \log 4 = 0.602$; $J = \frac{0.602}{0.602} = 1.000$	$[20 \log 20 - (1 \log 1 + 1 \log 1 + 1 \log 1 + 17 \log 17)]/20 = [26.0206 - (0 + 0 + 0 + 20.9176)]/20 = 5.1030/20 = 0.255$; $H_{max} = \log 4 = 0.602$; $J = \frac{0.255}{0.602} = 0.424$	$[40 \log 40 - (2 \log 2 + 2 \log 2 + 2 \log 2 + 34 \log 34)]/40 = [64.0824 - (0.6021 + 0.6021 + 0.6021 + 52.0703)]/40 = 10.2058/40 = 0.255$; $H_{max} = \log 4 = 0.602$; $J = \frac{0.255}{0.602} = 0.424$

1. Note that "diversity" arises as categories become more evenly filled.
2. Differences in sample size do not affect J.

Solution to a Problem

1. Because the SW index is *un-standardized*.
   a. You need to make multiple estimates of the index at each location; the more the better.
   b. Then, calculate the mean and SD of those values.
   c. Estimate 95% CI for comparisions with other locations
   d. Or, compare values of J using ANOVA.

Ordinal Data

1. Data that contains information, not only on category, but also on *relationship* to other categories.

Ordinal Data

2. Formal properties

a. This scale incorporates equivalence *within* class, but also order *between* classes.

Ordinal Data: Examples

1. Military ranks

a.Private, Corporal, Sergeant, Lieutenant

2. Grades

a.Excellent, Very Good, Good, Fair, Do Not Fund.

3. Numerical ranking of any kind.

Ordinal Data: Parameters

1. Since nonparametric statistics make no assumptions about mean or variance.

a. The appropriate indices of central tendency, dispersion are the *median*, and the *range*.

Ordinal Data: Parameters

1.Types of tests:

a. Tests are based on the *median*.

b. Calculated as:

1. $X_{[(n+1)/2]}$ if N is odd.
a. Thus, if n = 7, median is 4.

2. $\{X_{(n/2)}+X_{[(n/2) + 1]}\}/2$ if N is even.
a. Thus, if n = 8, median is 4.5.

Ordinal Data: Parameters

1. There are no assumptions about the shape of the data distribution.
2. But there are hypotheses about *order*.
3. Since there are no estimates of variance, this information is given by the *range*.

a. Calculated as:
1. $X_n - X_1$

b. Thus, if data ranges from 1.2g to 3.4g; range = (3.4g - 1.2g) = 2.2g.

Ranking of Data

1. Recall that information useful for nonparameteric tests is often encoded in *ranks*.
2. This is fine, as long as ranks are *not tied*.
3. If ties occur, it can indicate that the scale of measurement is *not fine enough*.
4. This can cause difficulties because tied ranks can prevent accurate discrimination among groups.

Tied Ranks

1.Most operations permit it unless it becomes excessive.

2.Ties are handled by calculating the sum of all tied ranks and dividing by their number.

$$R_i = \Sigma\, r_j / n_j$$

Where R_i = the i-th assigned rank
r_j = the actual rank of the j-th value
n_j = the number of cases with value j.

Tied Ranks

a. For example.

data:	1	2	3	3	3	6	7	7	9	10
num. rank:	1	2	3	4	5	6	7	8	9	10
assgn. rank:	1	2	4	4	4	6	7.5	7.5	9	10

$R_4 = (3+4+5)/3 = 4$

$R_{7.5} = (7+8)/2 = 7.5$

Interval Data

1. Data that contains information on:

a. Category

b. Order

c. Distance between points

Interval Data

2. Formal properties:

a. Includes all of the above properties (equivalence, order).

b. Adds amount.

1. Important to assume that amount is *standardized*.

2. Many questionnaires are not.

Interval Data: Examples

1. Celsius, Fahrenheit, Kelvin scales for temperature.

2. "agree, somewhat disagree, agree, strongly agree."

a. Doesn't always work; distances are not always equivalent.

b. These are scales without a true zero (its value is arbitrary).

Interval Data

1.Types of tests:

a. This is the first quantitative scale of measurement.

b. The distribution of cases *can* be normal.

1. If so, assumptions of parametric tests *are* met.

2. Parametric tests *are* recommended.

However,

1. If data do not meet assumptions for parametric tests,

a. It is possible to use nonparametric tests, although some information may be lost.

Ratio Data

1. Data that contains all the above information plus:

a. Have a *true zero* point of origin.

2. Formal properties:

a. All of those above

b. A constant ratio between two scales of measurement.

Ratio Data: Examples

speed (distance/time);

pressure (force/area),

1. Types of tests:

a. Nonparametric tests are not really appropriate.

b. However, ratio (and percentage) data can be non-normally distributed.

c. Most data of this sort meets parametric assumptions or can be transformed to do so.

Tests Using Nominal Scale Data

One sample cases:

1. These are tests that consider hypotheses about a single sample of data.

a. Usually a goodness of fit test.

b. To determine whether a sample fits a theoretical distribution, or distribution with pre-specified characteristics.

Questions Include

a. Is there a difference in location (central tendency) between sample and population?

b. Is there a difference between observed and expected frequency?

c. Is there a difference between observed and expected proportions?

d. Is the sample drawn from a population with a specified distribution (normal or uniform)?

All of these questions

1. Can also be addressed using parametric statistics, often a t-test.

a. However you may not want to/be able to use this test because:

1. Assumptions required by parametric tests are violated.

2. Data are not in a form required by parametric tests.

Useful One Sample Tests

1. Binomial test

2. Goodness of fit tests

a. Chi-square

b. G-test: one sample tests, heterogeneity tests

Binomial Test

Many situations exist in which data are arranged as 2 mutually exclusive classes.

1. Called a "dichotomous population"
2. Examples:
a. male, female
b. married, single (mated, unmated)
c. marked, unmarked
d. success, failure

Binomial Test

2. The method for expressing the frequency of these types is familiar:

a. Assuming 2 types, and expressing these as probabilities of occurrence:

1. $P_{[x=0]} = p$; say .4 ("uninfected")
2. $P_{[x=1]} = q$ where $q = (1 - p)$; the portion of the sample that is not p;

3. Thus (1 - .4) = .6 ("infected')

Binomial Test

b. If the population is large, the probability of obtaining an uninfected individual depends on the number of times we draw:

$$(p + q)^N = 1$$

Where N = the number of draws from the population.

Binomial Test

a. If we let $N = 5$, then by expanding the equation, we obtain the probabilities of obtaining different combinations of infected and uninfected individuals :

$$p^5 + 5p^4q + 10p^3q^2 + 10p^2q^3 + 5pq^4 + q^5 = 1$$

Binomial Expansion

$$p^5 + 5p^4q + 10p^3q^2 + 10p^2q^3 + 5pq^4 + q^5 = 1$$

p^5 = all 5 uninfected

$5p^4q$ = 4 uninfected, 1 infected

$10p^3q^2$ = 3 uninfected, 2 infected

$10p^2q^3$ = 2 uninfected, 3 infected

$5pq^4$ = 1 uninfected, 4 infected

q^5 = all infected

Alternatively,

1. We can figure out the exact probability:

a. Let k = the sum of counts of one class.
b. Let N = total number of opportunities to choose.

c. The number of objects in k and in N-k is given by the equation for a binomial distribution.

The Binomial Equation

$$P[k] = \binom{N}{k} p^k q^{N-k}$$

Where,

$$\binom{N}{k} = \frac{N!}{k!(N-k)!}$$