Types of Data, Descriptive Statistics, and Statistical Tests for Nominal Data

Patrick F. Smith, Pharm.D.
University at Buffalo
Buffalo, New York

NONPARAMETRIC STATISTICS

I. DEFINITIONS

A. Parametric statistics
   1. Variable of interest is a measured quantity.
   2. Assumes that the data follow some distribution which can be described by specific parameters
      a. Typically a normal distribution
   3. Example: There are an infinite number of normal distributions, all which can be uniquely defined by a mean and standard deviation (SD).

B. Nonparametric statistics
   1. Variable of interest is not measured quantity. Mean and SD have little meaning.
   2. Does not make any assumptions about the distribution of the data
   3. "Distribution-free" statistics

C. Dependent variable
   1. The variable of interest, the outcome of which is *dependent* on something else

D. Independent variable
   1. The variable that is being tested for an effect on the dependent variable

E. Example
   1. Does high-dose ciprofloxacin lead to seizures?
      a. Seizures = dependent variable
      b. Dose = independent variable

II. PARAMETRIC STATISTICS

A. Developed primarily to deal with categorical data (non-continuous data)
   1. Example: disease vs no disease; dead vs alive

B. Nonparametric statistical tests may be used on continuous data sets.
   1. Removes the requirement to assume a normal distribution
   2. However, it also throws out some information, as continuous data contains information in the way that variables are related.

Some Commonly Used Statistical Tests		
Normal theory-based tests	**Corresponding nonparametric tests**	**Purpose of test**
t test for independent samples	Mann-Whitney *U* test; Wilcoxon rank sum test	Compares two independent samples
Paired t test	Wilcoxon matched pairs signed-rank test	Examines a set of differences
Pearson correlation coefficient	Spearman rank correlation coefficient	Assesses the linear association between two variables
One-way analysis of variance (F test)	Kruskal-Wallis analysis of variance by ranks	Compares three or more groups
Two-way analysis of variance	Friedman two-way analysis of variance	Compares groups classified by two different factors

III. NONPARAMETRIC PROS AND CONS

A. Nonparametric pros
   1. Nonparametric tests make less stringent demands of the data.
      a. For a parametric test to be valid, certain underlying assumptions must be met.
         i. example: For a paired t test, assume that: data are drawn from normal distribution; every observation is independent of each other, and the SDs of the two populations are equal. Data are continuous.
      b. Nonparametric tests do not require these assumptions.
         i. can be used to evaluate data that are not continuous
         ii. no assumptions about distributions, independence, etc.

B. Nonparametric cons
   1. If using for a continuous data set, nonparametric tests throw information inherent in continuous data.
   2. Reduces power to detect a statistical difference
      a. A more conservative approach
   3. Example: For data from a normally distributed population, if the Wilcoxon signed-rank test requires 1000 observations to demonstrate statistical significance, a t test will only require 955.

IV. CONTINGENCY TABLES

A. Contingency tables are used to examine the relationship between subjects' scores on two qualitative or categorical variables.

B. One variable determines the row categories; the other variable defines the column categories.

C. Example: In studying the association between smoking and disease, the row categories in the figure below denote the categories of smoking status while the columns denote the presence or absence of disease.

		A		**B**		
		Disease		Disease		
		Yes	No	Yes	No	
Smoke	Yes	13	37	26%	74%	100%
	No	6	144	4%	96%	100%

V. CHI-SQUARED TEST

A. Commonly used procedure, uses contingency tables

B. Used to evaluate **unpaired samples** (unrelated groups)

C. Often used to evaluate proportions

D. Is there a difference in the proportion of viral infections in patients administered a vaccine? (12/100 vs. 2/100)

E. Assumes nominal data (no ordering between variable groups)

F. Limited when the numbers of subjects in any "cell" is low (rule of thumb, <5)

G. <u>General logic</u>

1. Given two groups (vaccine vs control), the EXPECTED infection rate if the vaccine has no effect would be equal among the two groups. This is the null hypothesis. The chi-squared test compares the EXPECTED frequency of a particular event to the OBSERVED frequency in the population of interest.

H. Formulas

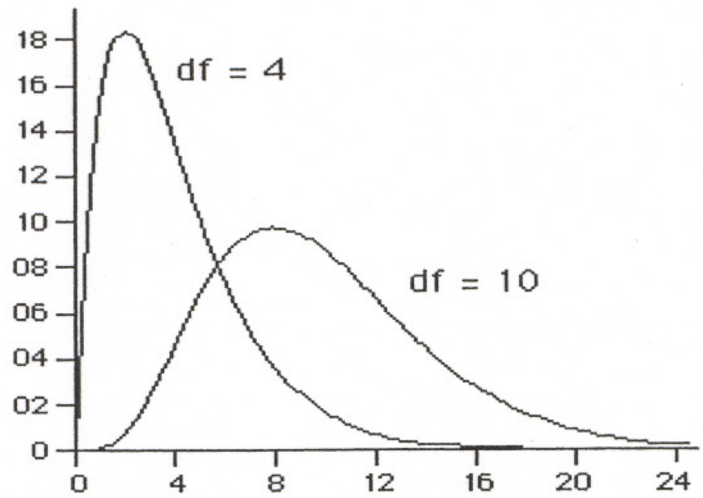
$$X^2 = \sum \frac{(O-E)^2}{E}$$

with df = (r – 1)(c –1)

<u>Expected Frequencies (E) for each cell:</u>

$$E_{ij} = \frac{T_i \times T_j}{N}$$

I. Distribution



Chi-Square distribution

Chi-squared, by strict definition, is not a true nonparametric test. It assumes a distribution that can be described by a single parameter, degrees of freedom.

J. Chi-squared example problems (refer to Example Problem handout)

J. Chi-squared example problems (refer to Example Problem handout)

VI. FISHER'S EXACT TEST

A. Alternative to chi-squared for 2 x 2 contingency tables
   1. Improves accuracy when expected frequencies are small (<5) or sample size is small (n=20)
   2. Calculates exact probabilities

a	b	(a +b)
c	d	(c + d)
(a + c)	(b + d)	N

$$P_{(outcome)} = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{N!\ a!\ b!\ c!\ d!}$$

VII. MCNEMAR'S TEST OF SYMMETRY

A. Chi-squared test requires samples to be independent of each other.

B. McNemar's test is used when samples are related (similar to paired t test).

C. There are often times where measures may be repeated.

D. Example. Does drug X cause insomnia?
   1. Patients may be questioned about insomnia before and after starting the drug.
   2. The researcher asks the question, "Do more patients have insomnia since starting the drug?"

E. Refer to Example Problems handout

VIII. KRUSKAL-WALLIS TEST

A. Compares two independent samples

B. Values of a variable are transformed to ranks.
   1. Tests that there is no shift in the center of the groups (that is, the centers do not differ)

C. If there are only two groups, the procedure reduces to the Mann-Whitney test—the analogue of the unpaired t test.

IX. WILCOXON SIGNED-RANK TEST

A. Nonparametric analogue of the paired t test

B. Compares the rank values of variables pair-by-pair
   1. The sum of the ranks associated with positive and negative differences is computed.
   2. The test statistic is the lesser of the two sums of ranks.

C. Refer to Example Problems handout

J. Chi-squared example problems (refer to Example Problem handout)

VI. FISHER'S EXACT TEST

A. Alternative to chi-squared for 2 x 2 contingency tables
   1. Improves accuracy when expected frequencies are small (<5) or sample size is small (n=20)
   2. Calculates exact probabilities

a	b	(a +b)
c	d	(c + d)
(a + c)	(b + d)	N

$$\mathbf{P}_{(\text{outcome})} = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{N!\ a!\ b!\ c!\ d!}$$

VII. MCNEMAR'S TEST OF SYMMETRY

A. Chi-squared test requires samples to be independent of each other.

B. McNemar's test is used when samples are related (similar to paired t test).

C. There are often times where measures may be repeated.

D. Example. Does drug X cause insomnia?
   1. Patients may be questioned about insomnia before and after starting the drug.
   2. The researcher asks the question, "Do more patients have insomnia since starting the drug?"

E. Refer to Example Problems handout

VIII. KRUSKAL-WALLIS TEST

A. Compares two independent samples

B. Values of a variable are transformed to ranks.
   1. Tests that there is no shift in the center of the groups (that is, the centers do not differ)

C. If there are only two groups, the procedure reduces to the Mann-Whitney test—the analogue of the unpaired t test.

IX. WILCOXON SIGNED-RANK TEST

A. Nonparametric analogue of the paired t test

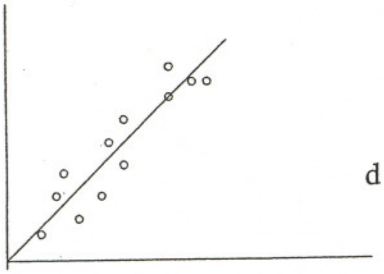
B. Compares the rank values of variables pair-by-pair
   1. The sum of the ranks associated with positive and negative differences is computed.
   2. The test statistic is the lesser of the two sums of ranks.

C. Refer to Example Problems handout

X. SPEARMAN RANK CORRELATION COEFFICIENT

A. Nonparametric analogue of linear regression and the correlation coefficient

Nonparametric analogue of linear regression
and the correlation coefficient (r)



$$r_s = 1 - \frac{6\sum d^2}{n^3 - n}$$

d = difference of ranks at each point

B.

Height	Rank	Weight	Rank	d
31	1	7.7	2	-1
32	2	8.3	3	-1
33	3	7.6	1	2
34	4	9.1	4	0
35	5.5	9.6	5	0.5
35	5.5	9.9	6	-0.5

$\mathbf{R_s} = 6(-1^2 + -1^2 + 2^2 + 0 + 0.5^2 +- 0.5^2)/6^3 - 6) = 0.81$

For statistical significance, can look up critical values from table or obtain from software package.

EXAMPLE PROBLEMS

NONPARAMETRIC STATISTICS

Example Problem 1: Association between tryptophan dietary supplements and eosinophilia-myalgia syndrome (EMS). A number of subjects from a particular area are evaluated; 80 patients with EMS were identified, along with 200 matched controls. Is there a statistically significant association between tryptophan use and EMS ?

- Unrelated groups, categorical (yes/no) data – chi-squared is appropriate

Observed Results:

		EMS	No EMS	Total
Tryptophan use	Yes	**42**	**34**	76
	No	**38**	**166**	204
	Total	80	200	280

(42 of 76 patients taking tryptophan had EMS, compared to 38 of 204 not taking tryptophan)

Expected values if no association exists (null hypothesis):

		EMS	No EMS	Total
Tryptophan use	Yes	**21.7**	**54.3**	76
	No	**58.3**	**145.7**	204
	Total	80	200	280

The rate of EMS in the overall population, assuming no effect, would be 80/280 (28.6%). (.286*76 = 21.7; .286x204 = 58.3). The No EMS cells can then be calculated from subtracting the total (ex: 76 – 21.7 = 54.3).

$$E_{11} = \frac{76x80}{280} \qquad E_{12} = \frac{76x200}{280}$$

$$E_{21} = \frac{204x80}{280} \qquad E_{22} = \frac{204x200}{280}$$

To evaluate significance, one needs a mean and measure of dispersion (ex. – standard deviation, standard error, variance, etc.). The chi-squared test is based on a Poisson distribution, where mean = variance); therefore, the chi-squared test assumes that the variance is equal to the expected mean value.

$$X^2 = \sum \frac{(O-E)^2}{E}$$ **Therefore, in this example:**

$\mathbf{X^2 = (42/21.7)^2/21.7 + (34\text{-}54.3)^2/54.3 + (38\text{-}58.3)^2/58.3 + (166\text{-}145.7)^2/145.7 = 36.4}$

→ Look up the result in a chi-squared table (a 2 x 2 contingency table has 1 degree of freedom). To be significant at the 0.05 level, X^2 must be > 3.84. Since 36.4 >> 3.84, the result is highly significant.

Critical Values for the Chi-Squared Distribution

	Significance Level				
df	0.10	0.05	0.025	0.01	0.005
1	2.7055	3.8415	5.0239	6.6349	7.8794
2	4.6052	5.9915	7.3778	9.2104	10.5965
3	6.2514	7.8147	9.3484	11.3449	12.8381
4	7.7794	9.4877	11.1433	13.2767	14.8602
5	9.2363	11.0705	12.8325	15.0863	16.7496
6	10.6446	12.5916	14.4494	16.8119	18.5475
7	12.017	14.0671	16.0128	18.4753	20.2777
8	13.3616	15.5073	17.5345	20.0902	21.9549
9	14.6837	16.919	19.0228	21.666	23.5893
10	15.9872	18.307	20.4832	23.2093	25.1881
11	17.275	19.6752	21.92	24.725	26.7569
12	18.5493	21.0261	23.3367	26.217	28.2997
13	19.8119	22.362	24.7356	27.6882	29.8193
14	21.0641	23.6848	26.1189	29.1412	31.3194
15	22.3071	24.9958	27.4884	30.578	32.8015
16	23.5418	26.2962	28.8453	31.9999	34.2671
17	24.769	27.5871	30.191	33.4087	35.7184
18	25.9894	28.8693	31.5264	34.8052	37.1564
19	27.2036	30.1435	32.8523	36.1908	38.5821
20	28.412	31.4104	34.1696	37.5663	39.9969
21	29.6151	32.6706	35.4789	38.9322	41.4009
22	30.8133	33.9245	36.7807	40.2894	42.7957
23	32.0069	35.1725	38.0756	41.6383	44.1814
24	33.1962	36.415	39.3641	42.9798	45.5584
25	34.3816	37.6525	40.6465	44.314	46.928
26	35.5632	38.8851	41.9231	45.6416	48.2898
27	36.7412	40.1133	43.1945	46.9628	49.645
28	37.9159	41.3372	44.4608	48.2782	50.9936
29	39.0875	42.5569	45.7223	49.5878	52.3355
30	40.256	43.773	46.9792	50.8922	53.6719
31	41.4217	44.9853	48.2319	52.1914	55.0025
32	42.5847	46.1942	49.4804	53.4857	56.328
33	43.7452	47.3999	50.7251	54.7754	57.6483
34	44.9032	48.6024	51.966	56.0609	58.9637
35	46.0588	49.8018	53.2033	57.342	60.2746
36	47.2122	50.9985	54.4373	58.6192	61.5811
37	48.3634	52.1923	55.668	59.8926	62.8832
38	49.5126	53.3835	56.8955	61.162	64.1812
39	50.6598	54.5722	58.1201	62.4281	65.4753
40	51.805	55.7585	59.3417	63.6908	66.766
41	52.9485	56.9424	60.5606	64.95	68.0526
42	54.0902	58.124	61.7767	66.2063	69.336
43	55.2302	59.3035	62.9903	67.4593	70.6157
44	56.3685	60.4809	64.2014	68.7096	71.8923
45	57.5053	61.6562	65.4101	69.9569	73.166
46	58.6405	62.8296	66.6165	71.2015	74.4367
47	59.7743	64.0011	67.8206	72.4432	75.7039
48	60.9066	65.1708	69.0226	73.6826	76.9689
49	62.0375	66.3387	70.2224	74.9194	78.2306
50	63.1671	67.5048	71.4202	76.1538	79.4898

Eample Problem 2:

A sociological study evaluated the characteristics of marriage by religion; 256 people were surveyed for religion and marital status. The results were as follows:

	Protestant	Catholic	Jewish	None	Other	Total
Never	29	16	8	20	0	73
Married	75	21	11	19	1	127
Divorced	21	6	3	13	0	43
Separated	8	3	1	0	1	13
Total	133	46	23	52	2	256

Is there a relationship between marital status and religion?

SYSTAT – chi-squared output

```
WARNING: More than one-fifth of fitted cells are sparse (frequency < 5).
         Significance tests computed on this table are suspect.
```

Test statistic	Value	df	Prob
Pearson chi-squared	22.718	12.000	0.030

What happened??

Omitting sparse cells: Leave out 'other' and 'separated':

	Protestant	Catholic	Jewish	None	Total
Never	29	16	8	20	73
Married	75	21	11	19	126
Divorced	21	6	3	13	43
Total	125	43	22	52	242

Test statistic	Value	df	Prob
Pearson chi-squared	10.368	6.000	0.110

There is no statistically significant difference between the groups ($p=0.11$)

Example Problem 3: McNemar Test of Symmetry

In November of 1993, the U.S. Congress approved the North American Free Trade Agreement (NAFTA). Let's say that two months before the approval and before the televised debate between Vice President Al Gore and businessman Ross Perot, political pollsters queried a sample of 350 people, asking "Are you for, unsure, or against NAFTA?" Immediately after the debate, the pollsters contacted the same people and asked the question a second time. Here are the results:

BEFORE$ (rows) by AFTER$ (columns)

	for	unsure	against	Total
for	51	22	28	101
unsure	46	18	27	91
against	52	49	57	158
Total	149	89	112	350

Percents of total count
BEFORE$ (rows) by AFTER$ (columns)

	AFTER				
	for	unsure	against	Total	N
for	**14.571**	6.286	8.000	28.857	101
unsure	13.143	**5.143**	7.714	26.000	91
against	**14.857**	14.000	**16.286**	45.143	158
Total	42.571	25.429	32.000	100.000	
N	149	89	112		350

Test statistic	Value	df	Prob
Pearson chi-squared	11.473	4.000	0.022
McNemar Symmetry chi-squared	22.039	3.000	0.000

The McNemar test of symmetry focuses on the counts in the off-diagonal cells (those along the diagonal are not used in the computations). We are investigating the direction of change in opinion. First, how many respondents became more negative about NAFTA?

Among those who initially responded For, 22 (6.29%) are now Unsure and 28 (8%) are now Against. Among those who were Unsure before the debate, 27 (7.71%) answered Against afterwards. The three cells in the upper right contain counts for those who became more unfavorable and comprise 22% (6.29 + 8.00 + 7.71) of the sample. The three cells in the lower left contain counts for people who became more positive about NAFTA (46, 52, and 49) or 42% of the sample.

The null hypothesis for the McNemar test is that the changes in opinion are equal. The chi-squared statistic for this test is 22.039 with 3 df and $p<0.0005$. You reject the null hypothesis. The pro-NAFTA shift in opinion is significantly greater than the anti-NAFTA shift.

Example Problem 4: Wilcoxon Signed-Rank Test

Evaluate the effect of a diuretic in healthy volunteers:

	Daily UOP				
Subject	No drug	+ Drug	Difference	Rank of difference	Signed rank of difference
1	1600	1490	-110	5	-5
2	1850	1300	-550	6	-6
3	1300	1400	+100	4	+4
4	1500	1410	-90	3	-3
5	1400	1350	-50	2	-2
6	1010	1000	-10	1	-1

W = sum of signed ranks = -13

If the drug has no effect, the ranks associated with a positive change should be similar to the ranks associated with a negative change; hence, the sum (W) should = 0.

How large must W be to call this a statistically significant difference? Refer to Critical Values table:

N	Critical Value	P
5	15	.062
6	21	.032
	19	.062
7	28	.016
	24	.046
8	32	.024
	28	.054
9	39	.020
	33	.054
10	45	.02
	39	.048
11	52	.018
	44	.054
12	58	.02
	50	.052
13	65	.022
	57	.048
14	73	.02
	63	.05
15	80	.022
	70	.048

*Due to the nature of discrete possible values of W, p values at traditional breakpoints are usually not possible (ex.: p=0.05).