Non-parametric statistics

Leif E. Peterson

May 18, 2006

Nominal (Categorical) Measurement Scale

Data are often collected to identify the number of subjects that fall into in categories. Examples are number of new patients with breast cancer, lung cancer, or no cancer; low-intermediate-high risk, or diagnostic category A, B, C, etc.

One Sample Case - Binomial Test

- There are many populations for which there are two classes.

- EX: male/female, literate/illiterate, member/nonmember, married/single.

- All events will fall into one or the other of the two discrete categories.

- For any population with two classes, if the proportion of subject in one class is P, the proportion in the other is $1 - P$, $Q = 1 - P$.

Binomial test

- Probability of obtaining x objects in one category and $N - x$ in the other is is the *binomial probability*

$$p(x) = \binom{N}{x} P^x Q^{N-x} \tag{1}$$

where

$$\binom{N}{x} = \frac{N!}{x!(N-x)!} \tag{2}$$

EX: roll a die five times ($N = 5$). What is the probability of rolling a two sixes?

$$
\begin{aligned}
p(x) &= \binom{N}{x} P^x Q^{N-x} \\
&= \frac{5!}{2!3!} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3 \\
&= 0.16
\end{aligned}
\tag{3}
$$

Binomial test

EX: In a study of stress, 18 college students were taught how to tie a knot using two methods. Half of the subjects selected randomly from the 18 were taught method A first while half learned method B first. Later that same day (at midnight) subjects were asked to tie the knot. It was hypothesized that stress would result in regression, that is, subjects would use the first method they learned to tie the knot. Subjects were categorized according to whether they used the first method they learned or the second method to tie the knot.

Binomial Test

Null Hypothesis: $H_o : p_1 = p_2 = 0.5$ That is, there is no difference between the probability of using the first-learned method under stress (p_1) and the probability of using the second method under stress (p_2). Any difference is expected to follow $H_a : p_1 > p_2$.

Results: 16 subjects used the first method and 2 used the second method. $N = 18$ and $x = 2$.

$$
\begin{aligned}
p(x \leq 2) &= p(0) + p(1) + p(2) \\
&= \frac{18!}{0!18!}\left(\frac{1}{2}\right)^{0}\left(\frac{1}{2}\right)^{18} = 0.0000038147 \\
&= \frac{18!}{1!17!}\left(\frac{1}{2}\right)^{1}\left(\frac{1}{2}\right)^{17} = 0.0000686646 \\
&= \frac{18!}{2!16!}\left(\frac{1}{2}\right)^{2}\left(\frac{1}{2}\right)^{16} = 0.0005836487 \\
&= p(0) + p(1) + p(2) = 0.000004 + 0.00007 + 0.0006
\end{aligned} \tag{4}
$$

Thus, the decision is to reject H_o in favor of the alternative H_a, that is $p_1 > p_2$ such that people under stress revert to the first learned of two methods.

Kolmogorov-Smirnov One-Sample test

Let $F_0(X)$ be the theoretical cumulative distribution under H_o. That is, for any value of X, the value $F_0(X)$is the proportion of cases expected to have scores equal to or less than X. Let $S_N(X)$ be the observed cumulative distribution for the data. According to the null hypothesis, it is expected that the sample has been drawn from the theoretical distribution, so that values of $S_N(X)$ are fairly close to $F_0(X)$. The Kolmogorov-Smirnov test focuses on the largest of the differences $D = \max|F_0(X) - S_N(X)|$

Kolmogorov-Smirnov

EX: Researcher believes fair-skinned people have a hierarchy of skin color preferences. A picture is taken of each of 10 subjects, and 5 copies are made with skin shades ranking from 1 (whitish) to 5(dark tan). Each subject is offered a choice among the 5 copies of his/her own photograph. If skin shade is unimportant, photographs of each rank should be equally often. If skin shade is important (hypothesized), then subjects should favor one of the extreme ranks.

Kolmogorov-Smirnov

	Rank of photo chosen (5 is the most t				
	1	2	3	4	5
f	0	1	0	5	4
$F_0(X)$	1/5	2/5	3/5	4/5	5/5
$S_{10}(X)$	0/10	1/10	1/10	6/10	10/10
$D = \|F_0(X) - S_N(X)\|$	2/10	3/10	**5/10**	2/10	0

Lookup of the critical value for $N = 10$ and $D \geq 0.5$ has an associated probability under H_o of $p < 0.01$. Thus, we reject the null hypothesis in favor of the alternative that subjects show significant preference among skin colors.

Kolmogorov-Smirnov

Kolmogorov-Smirnov test is more powerful than χ^2 test for very small samples. As a comparison, in order to perform a χ^2 test on the skin color data, ranks 1 and 2 were joined so that 1 count was in light ranks (1 and 2), and 9 counts were in tan ranks (3,4, and 5). The calculated χ^2=3.75 (uncorrected for continuity) and d.f. = k-1 which is 1 since there are two groups. Thus $0.05 < p < 0.10$. Note that the Kolmogorov-Smirnov p-values was 0.01, whereas the χ^2 p-value was greater than 0.05.

One-sample Runs Test

Consider the sequence of plus and minus scores: $++---+$ $----++-+$ Sequence starts with a run of 2+, followed by a run of 3-, 1+, 4-, 2+, 1- and 1+. The total number of runs in a sequence is an indication of whether or not the sample is random or not. If very few runs occur, a time trend or clustering due to lack of independence is suggested. If many runs occurs, then systematic short-period cyclical fluctuations seem to be influencing the scores.

One-sample Runs Test

EX: Sequence of gender among 30M ($n_1 = 30$) and 20F ($n_2 = 20$) in the queue at a theater box office. Null hypothesis is that order of males and females is random.

M F M F MMM FF M F M F M F MMMM F M F M F MM FFF M F M F M F MM F MM F MMMM F M F MM

The number of runs $r = 35$.

One-sample Runs Test

To determine whether $r \geq 35$ might have readily occurred under the null hypothesis, we calculate

$$z = \frac{r - \left(\frac{2n_1n_2}{n_1+n_2} + 1\right)}{\sqrt{\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}} = \frac{35 - \left(\frac{2(30)(20)}{30+20} + 1\right)}{\sqrt{\frac{2(30)(20)(2(30)(20)-30-20)}{(30+20)^2(30+20-1)}}} = 2.98 \tag{5}$$

Since $p(z \geq 2.98)$=0.0028, we reject the null in favor of the alternative hypothesis that the order of males and females in the queue was not random.

Two Related Samples

Related samples occur when subjects are matched and each is assigned to a different treatments, or when subjects are used as their own controls. When a subject is used as his/her own control he/she is exposed to both treatments at different times. Analysis of results for two related samples is usually addressed with the t-test assuming normally and independently distributed data, and that data are at least on an interval scale. There may be cases, however, when response for the pair are the same or different, or each subject response form the pair is positive or negative.

McNemar Test for the Significance of Changes

		After	
		-	+
Before	+	A	B
	-	C	D

McNemar Test

Since only cells A and D contain subjects who changed between first and second response, the expectation under the null hypothesis is that $\frac{1}{2}(A+D)$ subjects changed one way and $\frac{1}{2}(A+D)$ changed the other. In other words $\frac{1}{2}(A+D)$ is the expected cell frequency in cells A and D.

McNemar Test

Recall the chi-square function

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(A - \frac{A+D}{2})^2}{\frac{A+D}{2}} + \frac{(D - \frac{A+D}{2})^2}{\frac{A+D}{2}} \\ &= \frac{(A - D)^2}{A + D}\end{aligned} \tag{6}$$

McNemar Test

EX: A child psychiatrist is interested in determining whether or not children who are new in nursery school usually initiate personal contacts with adults rather than with other children. The hypothesis is that with increasing experience, children will initiate social contacts with children rather than with adults. He observes 25 children on their first day and categorizes their first initiation of of social contact. After a month, each child is observed again and the same categorizations are made.

McNemar Test

		Initiation at day 30	
		Child	Adult
Initiation on first day	Adult	A	B
	Child	C	D

McNemar Test

		Initiation at day 30	
		Child	Adult
Initiation on first day	Adult	14	4
	Child	3	4

$$\begin{aligned}\chi^2 &= \frac{(|A-D|-1)^2}{A+D} \\ &= \frac{(|14-4|-1)^2}{14+4} \\ &= \frac{9^2}{18} \\ &= 4.5\end{aligned} \tag{7}$$

which is a one-tailed test with 1 d.f.

Sign Test for Related Samples

An experimenter randomly selects 100 adults from a community and asks them whether *less* or *more* punitive action against juvenile delinquents should be taken. She then shows a movie about juvenile delinquents, and hypothesizes that the movie will change the opinion of how the members of the community about how severely juvenile delinquents should be published. After the movie, she asks the same question.

		After film	
		Less	More
Before film	More	59	7
	Less	8	26

Sign Test for Related Samples

Overall, 85 people changes their opinions, and the null hypothesis applies only to these individuals. If the film had no effects, we would expect half to change from more to less and half to change form less to more (42.5 subjects to show both changes equally).

$$z = \frac{x - \mu_x}{\sigma_x} = \frac{x - \frac{1}{2}N}{\frac{1}{2}\sqrt{N}} \tag{8}$$

where N is the number of subjects showing a change in either direction.

$$z = \frac{(x \pm 0.5) - \frac{1}{2}N}{\frac{1}{2}\sqrt{N}} \tag{9}$$

where $x + 0.5$ is used when $x < 1/2N$ and $x - 0.5$ is used when $x > 1/2N$

Substituting in, we have

$$\begin{aligned} z &= \frac{x - \mu_x}{\sigma_x} \\ &= \frac{(59 - 0.5)\frac{1}{2}85}{\frac{1}{2}\sqrt{85}} \\ &= 3.47(2 - tailed) \end{aligned} \tag{10}$$

Wilcoxon Matched Pairs for Related Samples

The sign test only considered the direction of change. The Wilcoxon test considers direction and magnitude of change, so a more powerful test can be made.

Pair	Social perceptiveness of twin in nursery school	Social perceptiveness of twin at home	d	Rank of d	Rank of less frequent sign
A	82	63	19	7	
B	69	42	27	8	
C	73	74	-1	-1	1
D	43	37	6	4	
E	58	51	7	5	
F	56	43	13	6	
G	76	80	-4	-3	4
H	65	82	3	2	
					T=4

Wilcoxon Matched Pairs for Related Samples

$$z = \frac{T - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24}}} = \frac{4 - \frac{(8)(9)}{4}}{\sqrt{\frac{(8)(9)(17)}{24}}} = -1.96 \qquad (11)$$

which is a two-tailed test (P=0.05).

Mann-Whitney Test for Independent Samples

Society	Anxiety rating for child in society w/o oral explanations	Rank	Anxiety rating for child in society with oral explanations	Rank
1	6	1.5	6	1.5
2	7	5	8	9.5
3	7	5	8	9.5
4	7	5	10	16
5	7	5	10	16
6	7	5	10	16
7	8	9.5	11	20.5
8	8	9.5	11	20.5
9	9	12	12	24.5
10	10	16	12	24.5
11	10	16	12	24.5
12	10	16	12	24.5
13	10	16	13	29.5
14	12	24.5	13	29.5
15	12	24.5	13	29.5
16	13	29.5	14	33
17	$n_1 = 16$	$R_1 = 200.0$	14	33
18			14	33
19			15	36
20			15	36
21			15	36
22			16	38
23			17	39
			$n_2 = 23$	$R_2 = 580.0$

Mann-Whitney Test for Independent Samples

$$\begin{aligned} U &= n_1 n_2 + \frac{n_1(n_1+1)}{2} - \min(R) \\ &= (16)(23) + \frac{(16)(17)}{2} - 200 \\ &= 304 \end{aligned} \tag{12}$$

$$
\begin{aligned}
z &= \frac{U - \frac{n_1 n_2}{4}}{\sqrt{\frac{(n_1)(n_2)(n_1+n_2+1)}{12}}} \\
&= \frac{304 - \frac{(16)(23)}{4}}{\sqrt{\frac{(16)(23)(16+23+1)}{12}}} \\
&= 3.43
\end{aligned} \tag{13}
$$

Mann-Whitney Test		
Test var:	**Anxiety**	
Grouping var:	**Society**	
	1	**2**
Sample Size	16	23
R	200.00	580.00
U1,U2	304.00	64.00
Min(R1,R2)	200.00	
U based on Min(R1,R2)	304.00	
Mean U	184.00	
s.d.(U)	35.02	
Z-score	3.43	
Chi-square	11.74	
P-value (Chi-square test)	0.0006	
P-value(Permutation)	0.0010	
Avg. perm. chi-square	0.9957	

k-Independent Samples: Kruskal-Wallis ANOVA by Ranks

Authoritarianism scores for three groups of educators

Teaching oriented teachers	Administration-oriented teachers	Administrators
96	82	115
128	124	149
83	132	166
61	135	147
101	109	

Kruskal-Wallis ANOVA by Ranks

Ranked scores for three groups of educators

Teaching oriented teachers	Administration oriented teachers	Administrators
4	2	7
9	8	13
3	10	14
1	11	12
5	6	
$R_1 = 22$	$R_1 = 37$	$R_3 = 46$

Kruskal-Wallis ANOVA by Ranks

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1)$$
$$= \frac{12}{14(14+1)} \left[\frac{(22)^2}{5} + \frac{(37)^2}{5} + \frac{(46)^2}{5} + \right] - 3(14+1)$$
$$= 6.4 \tag{14}$$

Which is looked up in Table O of Siegel ($p < 0.049$)

Permutation Test for Spearman Rank Correlation

Because correlation involves the comparison of two ordered configurations of data, it is possible to perform *empirical testing* by first calculating the "observed" correlation, and then during a number of *iterations* permuting the data within each vector and calculating correlation from the permuted data. The advantage of permutation tests is that they are a permutation form of an exact test which can account for both the sample size of input data as well as the number of multiple tests (correlation coefficients) being determined.

Permutation Test for Spearman Rank Correlation

Consider stacking of the pair of random variates **x** and **y** being tested into a single vector **a,** with sample sizes n_1 and $n_{2,}$and$N = n_1 + n_{2,}$respectively. The Spearman rank correlation between vectors **x** and **y** is

$$R = 1 - \left(\frac{6T}{N^3 - N}\right)$$

where

$$T = \sum_{i=1}^{n_1} \left(r(a_i) - r(a_{n_1+i})\right)^2.$$

Permutation Test for Spearman Rank Correlation

Permutation Test for Spearman Rank Correlation For each permutation-based test, the Spearman rank correlation coefficient is first determined, call this R. Next, m=1/α*+1 iterations are performed where α*=1-(1-α)$^{1/\#tests}$ is the Bonferroni correction to the Type 1 error level (α=0.05) for multiple tests.

Permutation Test for Spearman Rank Correlation (cont'd)

During each bth iteration (b=1,2,. . . ,B) data within each of the two vectors being correlated are permuted and then a new value of Spearman rank correlation is calculated, call this $R^{(b)}$. After B iterations, the exact p-value is

$$Prob = \frac{\#\{b : R^{(b)} > R\}}{B} \tag{15}$$

Reference

Siegel, S. Nonparametric Statistics.
McGraw-Hill, New York(NY), 1956.