Lecture 14:
Non-Parametric Tests

Announcements

1. Assignment #4 is due next class (Thursday March 7^{th})
2. I will discuss the Group Projects on Thursday
3. Assignment #5 will be assigned after spring break

Non-Parametric Tests

- Normality Assumption: Don’t need to have normal distributions
- Data Type Assumption: Data can be measured on any scale
- Variance Assumption: No strict assumptions about the nature of the data
- BUT are more difficult for large samples and you loose information when data converted to ordinal or nominal

When to use Non-Parametric Tests

- When it is clear your dataset does not have a normal distribution
- Relatively normal distribution contains outliers

Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✔	✔	✔
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

Non Parametric Tests

- Hypothesis Testing
  - Parametric
    - Z Test
    - t-Test
    - ANOVA
  - Nonparametric
    - Wilcoxon Rank Sum Test
    - Kruskal-Wallis H-Test
    - Goodness of Fit
      Chi-square

Many More Tests Exist!

Wilcoxon Rank Sum Test

(also known as the Mann-Whitney U test)

- Use when only ordinal data is available or when ratio/interval data is not normally distributed

$$Z_W = \frac{W_i - \overline{W}_i}{s_w}$$

- Very similar to the t-test except that the test is difference in mean rank

$$\overline{W}_i = n_i\left(\frac{n_1 + n_2 + 1}{2}\right)$$

- Calculate the sum of ranks (W_i) for one of the samples and compare to the mean of ranks $\overline{W}_i$

$$s_w = \sqrt{n_1 n_2\left(\frac{n_1 + n_2 + 1}{12}\right)}$$

- Calculate the standard deviation of the ranks

Measure of height between girls and boys

Male Height	Rank	Female Height	Rank
67		62	
69		68	
67		63	
64		64	
70			
Sum		Sum	

$$\overline{W_i} = n_i\left(\frac{n_1 + n_2 + 1}{2}\right)$$

$$s_w = \sqrt{n_1 n_2\left(\frac{n_1 + n_2 + 1}{12}\right)}$$

$$Z_W = \frac{W_i - \overline{W_i}}{s_w}$$

n_1 = number of observations in first sample
n_2 = number of observations in second sample

W_i and n_i can be either the first or the second sample

Measure of height between girls and boys

Male Height	Rank	Female Height	Rank
67	5.5	62	1
69	8	68	7
67	5.5	63	2
64	3.5	64	3.5
70	9		
Sum	31.5	Sum	13.5

$$\overline{W_i} = n_i\left(\frac{n_1 + n_2 + 1}{2}\right)$$

25

$$s_w = \sqrt{n_1 n_2\left(\frac{n_1 + n_2 + 1}{12}\right)}$$

4.08

$$Z_W = \frac{W_i - \overline{W_i}}{s_w}$$

1.59

$n_1 = 5$
$n_2 = 4$

$n_i = 5$
$W_i = 31.5$

Table 5-2 Proportions of the Normal Curve above the Absolute Value of Z

First digit and first decimal of Z	Second decimal of Z									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	**.0250**	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010

Assumptions and Limitations

- Independence Assumption: Two independent random samples
- Distributions: Both population distributions have the same shape
- Data Type Assumption: Variables measured on an ordinal scale

$$Z_W = \frac{W_i - \overline{W}_i}{s_w}$$

$$\frac{\text{Variability between groups}}{\text{Variability within groups}}$$

Let's open up R…

Kruskal-Wallis Test

- Non-parametric test to determine if there is a statistically significant difference between 3 or more samples

- If there is no statistically significant difference in the ranks, then we should expect:

$$\frac{R_1}{n_1} = \frac{R_2}{n_2} = \frac{R_3}{n_3} = ... = \frac{R_i}{n_i}$$

$$H = \frac{12}{n(n+1)}\left(\sum_{i=1}^{k} \frac{{R_i}^2}{n_i}\right) - 3(n+1)$$

- R_i = sum of ranks of each sample, *i*
- *n* = total number of observations
- n_i = number obs. in each sample, *I*
- *k* = number of samples

Kruskal-Wallis Test

Surface Temperatures

Lot	Rank	Cement	Rank	Grass	Rank
30.7		25.7		27	
29.6		27.6		28.3	
26.8		26.6		28.9	
33.4		28		28.3	
31.4		27.3		27.9	
Sum		**Sum**		**Sum**	

$$H = \frac{12}{n(n+1)}\left(\sum_{i=1}^{k} \frac{{R_i}^2}{n_i}\right) - 3(n+1)$$

- R_i = sum of ranks of each sample, i
- n = total number of observations
- n_i = number obs. in each sample, I
- k = number of samples

Kruskal-Wallis Test

Surface Temperatures

Lot	Rank	Cement	Rank	Grass	Rank
30.7	13	25.7	1	27	4
29.6	12	27.6	6	28.3	9.5
26.8	3	26.6	2	28.9	11
33.4	15	28	8	28.3	9.5
31.4	14	27.3	5	27.9	7
Sum	57	**Sum**	22	**Sum**	41

$$H = \frac{12}{n(n+1)}\left(\sum_{i=1}^{k}\frac{{R_i}^2}{n_i}\right) - 3(n+1)$$

- R_i = sum of ranks of each sample, *i*
- *n* = total number of observations
- n_i = number obs. in each sample, *I*
- *k* = number of samples

$$H = \frac{12}{240}\left(\frac{57^2}{5} + \frac{22^2}{5} + \frac{41^2}{5}\right) - 48 = 6.14$$

47.06

df = k -1 = 2
k = number of samples

$$H = \frac{12}{n(n+1)}\left(\sum_{i=1}^{k}\frac{{R_i}^2}{n_i}\right) - 3(n+1)$$

$$H = 6.14$$

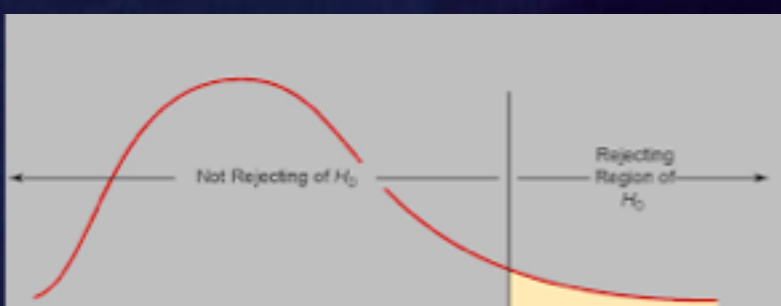

Table 6-6 Critical Values of χ^2

Degrees of freedom	Area to the right of critical value .10	.05	.025	.01
1	2.706	3.841	5.024	6.635
2	4.605	5.991	7.378	9.210
3	6.251	7.815	9.348	11.345
4	7.779	9.488	11.143	13.277
5	9.236	11.070	12.833	15.086
6	10.645	12.592	14.449	16.812
7	12.017	14.067	16.013	18.475
8	13.362	15.507	17.535	20.090
9	14.684	16.919	19.023	21.666
10	15.987	18.307	20.483	23.209
11	17.275	19.675	21.920	24.725
12	18.549	21.026	23.337	26.217
13	19.812	22.362	24.736	27.688
14	21.064	23.685	26.119	29.141
15	22.307	24.996	27.488	30.578
16	23.542	26.296	28.845	32.000
17	24.769	27.587	30.191	33.409
18	25.989	28.869	31.526	34.805
19	27.204	30.144	32.852	36.191
20	28.412	31.410	34.170	37.566
21	29.615	32.671	35.479	38.932
22	30.813	33.924	36.781	40.289
23	32.007	35.172	38.076	41.638
24	33.196	36.415	39.364	42.980
25	34.382	37.652	40.646	44.314
26	35.563	38.885	41.923	45.642
27	36.741	40.113	43.195	46.963
28	37.916	41.337	44.461	48.278
29	39.087	42.557	45.722	49.588
30	40.256	43.773	46.979	50.892
40	51.805	55.758	59.342	63.691
50	63.167	67.505	71.420	76.154
60	74.397	79.082	83.298	88.379
70	85.527	90.531	95.023	100.425
80	96.578	101.879	106.629	112.329
90	107.565	113.145	118.136	124.116
100	118.498	124.342	129.561	135.807

Assumptions and Limitations

- Independence Assumption: Three or more independent random samples

- Distribution Assumption: Each population has an underlying continuous distribution of values

- Data Type Assumption: Variables measured on an ordinal scale

$$H = \frac{12}{n(n+1)}\left(\sum_{i=1}^{k}\frac{{R_i}^2}{n_i}\right) - 3(n+1)$$

Let’s open up R...

Lecture 14:
Non-Parametric Tests

www.terragalleria.com