Stats Notes

Ben Sheffield

April 2017

Statistics 3

0.1 Contingency Tables

0.1.1 Contingency Tables

So far we have looked at the frequency a single event happens, but sometimes your interested in the frequencies of two criteria happening.

Example

Trains from 3 stations leave late or on-time, test at 5% is their is any association.

	On Time	Late
A	26	14
B	30	10
C	44	26

1. Hypothesis

   H_0: No association
   H_1: Association

2. Calculate the expected and χ^2 values

O_i	E_i	$\frac{(O_i - E_i)^2}{E_i}$
26	$\frac{80}{3}$	$\frac{1}{60}$
30	$\frac{80}{3}$	$\frac{5}{12}$
44	$\frac{140}{3}$	$\frac{1}{60}$
14	$\frac{40}{3}$	0.152
10	$\frac{40}{3}$	$\frac{1}{30}$
16	$\frac{70}{3}$	0.305

3. Find the statistic χ^2

$$\sum \frac{(O_i - E_i)^2}{E_i} = \chi^2 = 1.757$$

4. Find the critical value

$$v = (3-1) \times (2-1) = 2$$
$$\chi^2_2(0.95) = 5.991$$

5. Conclusion

   $1.757 < 5.991$ so accept H_0, evidence suggests stations and lateness are not associated.

0.1.2 Yates Correction

For contingency tables that are 2x2 we must use yate's correction since it produces a more accurate result.

$$\chi^2 = \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

Example

Test the results of the following drug trial at a 5% confidence interval.

	Cold	No cold
Drug	32	66
Placebo	45	55

1. Hypothesis

   H_0: No association
   H_1: Association

2. Calculate the expected and χ^2 values

O_i	E_i	$\frac{(\lvert O_i - E_i\rvert - 0.5)^2}{E_i}$
34	36.5	0.633
45	39.5	0.633
66	60.5	0.413
55	60.5	0.413

3. Find the statistic χ^2

$$\sum \frac{(|O_i - E_i| - 0.5)^2}{E_i} = \chi^2 = 2.09$$

4. Find the critical value

$$v = (2-1) * (2-1) = 1$$
$$\chi^2_1(0.95) = 3.841$$

5. Conclusion

   $2.09 < 3.841$ so accept H_0, no evidence to show drug is effective.

0.2 Distribution free methods

Distribution free test's are tests that do not require the data to follow a particular distribution. Their are four Distribution free methods you need to know:

Test	Test for	When to use
The sign test	Median	Cant use Wilcoxon (data is not symmetrical)
Wilcoxon signed-rank test	Median/mean	Cant use z or t test
Mann-Whitney U test	Two samples have identical populations	Their are two samples
Kruskal-Wallis test	More than two samples from identical populations	Their are more than two samples

0.2.1 The Sign Test

The sign test works by checking if their is a significant difference in the median value of the samples by comparing each value pair. Unlike the Wilcoxon test the distribution does not have to be symmetrical.

Process

1. Hypothesis.
2. Find the sign of the difference ignoring any pairs that are equal.
3. Count the number of positive and negative signs (these are the test statistics)
4. Find the value of $P(X < min(a, b)|X \sim B(n, \frac{1}{2}))$ where a and b are the test statistics and n is the number of pairs (minus equal pairs)
5. Compare this value with the significant level, if its less than the significant level reject H_0

1-way Example

Example

Results from a 2005 sample of cocaine usage: 9, 26, 17, 18, 21, 16, 19, 13, 15. In 2000 the average was 14, test at the 10% significance level if cocaine usage has risen.

1. Hypothesis

   H_0: Population median $\eta = 14$
   H_1: Population median $\eta > 14$

2. Signs

$x - 14$	-5	12	3	4	7	2	5	-1	1
Signs	$-$	$+$	$+$	$+$	$+$	$+$	$+$	$-$	$+$

3. Test Statistic

$$2^-/7^+$$

4. Critical value

$$X \sim B(9, \frac{1}{2})$$
$$P(X \leq 2) = 0.0898$$

5. Conclusion

   $0.0898 < 0.10$ so reject H_0, significant evidence at 10% level to suggest median cocaine usage has increased since 2000.

2-way Example

Example

Test is their is a difference in the aerosols average effectiveness at a 5% significance level.

1. Hypothesis

H_0: Population median difference $\eta_d = 0$

H_1: Population median difference $\eta_d \neq 0$

2. Signs

Patient	1	2	3	4	5	6	7	8	9
Aerosol A	28	22	10	40	18	52	49	40	34
Aerosol B	24	16	5	17	23	57	30	16	14
Sign	+	+	+	+	−	−	+	+	+

3. Test Statistic

$$2^-/7^+$$

4. Critical value

$$X \sim B(9, \frac{1}{2})$$

$$P(X \leq 2) = 0.0898$$

5. Conclusion

$0.0898 > 0.05$ so accept H_0, significant evidence at 5% level to suggest their is no difference in the average effectiveness of aerosols.

0.2.2 Wilcoxon Signed-rank Test

Wilcoxon signed-rank test is also a distribution free test, for when we can't use a z or t test. Its similar to the sign test except we rank the differences $ignoringthesigns, smallestfirst$, then add the ranks with the same signs.

Process

1. Hypothesis
2. Rank the absolutes of the differences.
3. Give the ranks the same sign as the difference.
4. Sum the positive then the negative ranks and pick the smallest value.
5. Find the critical value in the table.
6. If the statistic is less than the critical value reject H_0.

Example

Test if the distributions of mock and a-level results are the same.

1. Hypothesis

   H_0: Population median difference $\eta_d = 0$
   H_1: Population median difference $\eta_d > 0$

2. Find the difference, rank and signed rank

Candidate	Mock	A level	Difference	Rank	Signed rank
1	40	45	5	7	7
2	65	68	3	4	4
3	53	47	-6	9.5	-9.5
4	79	75	-4	6	-6
5	87	88	1	1	1
6	70	88	18	13	13
7	80	77	-3	4	-4
8	63	69	6	9.5	9.5
9	51	60	9	12	12
10	82	88	6	9.5	9.5
11	27	30	3	4	4
12	71	73	2	2	2
13	29	35	6	9.5	9.5

3. Test statistics

$$T_+ = \sum +\text{Rank} = 71.5$$
$$T_- = \sum -\text{Rank} = 19.5$$
$$T = 19.5$$

4. Critical value

$$\text{Critical value} = 21$$

5. Conclusion

   $19.5 < 21$ so reject H_0, evidence students did better in their A-levels.

0.2.3 Mann-Whitney U-Test

Mann-Whitney U-Test is another non-parametric test that is used to test if two sample were taken from the same population. It is used when we can do a t-test because the data is not normal.

Process

1. Hypothesis
2. Rank the data (as if it was just a single set) and calculate the sum of the ranks of each set.
3. Calculate the statistic for each set using $U = T - \frac{n(n+1)}{2}$ where T is the sum of the ranks and n is the sample size. The test statistic is the minimum of the two U values.
4. Find the critical value in the table.
5. If the statistic is less than the critical value reject H_0.

Example

Nine random plants from two sides of a valley where weighed, carry out a Mann-Whitney U test at 5% level of significance.

East Side	27.1	40.3	15.7	36.4	16.3	15.3	32.0	15.7	27.5
West Side	11.7	14.7	19.1	22.0	6.7	14.1	20.1	24.4	15.4

1. Hypothesis

   H_0: Samples taken from identical populations
   H_1: Samples not taken from identical populations

2. Rank the data

East Side	14	18	7.5	17	9	5	16	7.5	15
West side	2	4	10	12	1	3	11	13	6

$$T_E = 14 + 18 + ... + 15 = 109$$
$$T_W = 2 + 5 + ... + 6 = 62$$

3. Test Statistic

$$U_E = 109 - \frac{9 \times (9+1)}{2} = 64$$
$$U_W = 62 - \frac{9 \times (9+1)}{2} = 17$$
$$U = min(64, 17) = 17$$

4. Critical value

$$\text{Critical value} = 18$$

5. Conclusion

   $17 < 18$ so reject H_0, the population average weight differs across each side.

0.2.4 Kruskal-Wallis Test

Kruskal-Wallis test is a non-parametric version of an ANOVA test. The test determins if their is a difference between some samples. We use this test instead of an ANOVA test if the data is not normally distributed.

Process

1. Hypothesis
2. Rank the data (as if it was a single set) and calculate the sum of the ranks of each set.
3. Calculate the test statistic using: $H = \frac{12}{N(N+1)} \times \sum \frac{T_i^2}{n_i} - 3(N+1)$ where T_i is the sum of all ranks in a set of size n_i and N is the sum of all sample sizes.
4. Find the degrees of freedom, which is the number of samples minus one.
5. Use the χ^2 distribution to find the critical value.
6. Compare the test statistic and critical value, if statistic is larger than critical value reject H_0.

Example

Carry out a Kruskal-Wallis test to see if their is a difference between the fish prices from three different markets at 5% significance level.

A	B	C
220.3	190.1	228.7
226.3	209.7	231.3
227.3	223.4	249.6
228.1	224.2	260.7
242.4	226.4	289.7

1. Hypothesis

   H_0: Samples from the same population
   H_1: Samples from the different populations

2. Rank the data

A	B	C
3	1	10
6	2	11
8	4	13
9	5	14
12	7	15

3. Test statistic

$$T_A = 3 + 6 + ... + 12 = 38$$
$$T_B = 1 + 2 + ... + 7 = 19$$
$$T_C = 10 + 11 + ... + 15 = 63$$

$$\sum \frac{T_i}{n_i} = \frac{38^2}{5} + \frac{19^2}{5} + \frac{63^2}{5}$$
$$= 1154.8$$

$$H = \frac{12}{15(15+1)} \times 1154.8 - 3(15+1)$$
$$= 9.74$$

4. Degrees of freedom

$$v = 3 - 1$$
$$= 2$$

5. Critical value

$$\chi^2_2(5\%) = 5.991$$

6. Conclusion

$9.74 > 5.991$ so reject H_0, samples are not from identical populations

0.3 Correlation

0.3.1 Spearman's Rank Correlation Coefficent

The product moment correlation coefficient works when you have something to measure. Suppose you had an order, for example an order of preference in different blends of tea, this is when we use Spearman's rank correlation coefficient.

To calculate the Spearman's rank correlation coefficient we use the following:

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

Where d_i is the difference between rank x_i and rank y_i and n, the number of pairs.

Example

Two judges judge a competition, comment on how well they agree

Competitor	A	B	C	D	E	F	G	H	I	J
Judge 1	7.8	6.6	7.3	7.4	8.4	6.5	8.9	8.5	6.7	7.7
Judge 2	8.1	6.8	8.2	7.5	8.0	6.7	8.5	8.3	6.6	7.8

1. Rank the scores and find the difference

Competitor	A	B	C	D	E	F	G	H	I	J
x_i	4	9	7	6	3	10	1	2	8	5
y_i	4	8	3	7	5	9	1	2	10	6
d_i	0	1	4	-1	-2	1	0	0	-2	-1

2. Calculate to coefficient

$$\begin{aligned} r_s &= 1 - \frac{6 \times (0^2 + 2^2 + 4^2 + ...)}{10 \times (10^2 - 1)} \\ &= 1 - \frac{168}{990} \\ &= 0.830 \end{aligned}$$

3. Conclusion

   Fairly strong correlation between the judges, suggesting judges have similar criteria and standards.

0.3.2 Testing the correction coefficient

We may be interested in conducting a test to show their is no correlation between two random variables. First we find the correlation coefficient, then get the critical value from the appropriate table.

Example

Test to see if their is no correlation between English and maths marks at the 5% significance level.

Student	A	B	C	D	E	F	G	H
English	25	18	32	27	21	35	28	30
Mathematics	16	11	20	17	15	26	32	20

1. Hypothesis

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

2. Calculate the product moment correlation coefficient

$$\sum x_i = 216$$
$$\sum y_i = 157$$
$$\sum x_i^2 = 6052$$
$$\sum y_i^2 = 3391$$
$$\sum x_i y_i = 4418$$
$$S_{xx} = 6052 - \frac{216^2}{8} = 220$$
$$S_{yy} = 3391 - \frac{157^2}{8} = 309.875$$
$$S_{xy} = 4418 - \frac{216 \times 157}{8} = 179$$
$$r = \frac{179}{\sqrt{220 \times 309.875}} = 0.686$$

3. Critical value (from table)

$$\text{critical value} = \pm 0.7067$$

4. Conclusion

   $-0.7067 < 0.686 < 0.7067$ so value is not in critical region, accept H_0, evidence to suggest product moment correlation coefficient is zero.