

Topic 14: Nonparametric Methods (ST & D Chapter 24)

Introduction

- *Parametric statistics* deal with the **estimation of parameters** (e.g., means, variances) and **testing hypotheses for continuous normally distributed variables**.
- *Multivariate statistics* deal with analyses involving several variables at the same time and in techniques to classify experimental units based on multiple variables
- *Nonparametric statistics* do not relate to specific parameters (the broad definition) or maintain their distributional properties irrespective of the underlying distribution of the data (**distribution-free methods**). That is, nonparametric statistics compare distributions rather than parameters.
- Nonparametric statistics are **less restrictive** in terms of the assumptions compared to parametric techniques. Although some **assumptions**, for example, **samples are random and independent**, are **still required**.
- Nonparametric statistics generally involve tests based on **ranked data**, i.e. data that can be put in the order, and/or **categorical data**.
- *Nonparametric statistics* are generally **less powerful (sensitive)** than parametric statistics. That is, type II errors (false null hypothesis is accepted) are more likely. Less power to detect true differences (H_a is true)

14.2. Advantages of using nonparametric techniques

- They are appropriate when only weak assumptions can be made about the distribution.
- They can be used with categorical data when no adequate scale of measurement is available.
- For data that can only be ranked, nonparametric test using ranked data may be the only alternative.
- Relatively quick and easy to apply and to learn since they involve counts, ranks and signs.

14.3 The χ^2 test of goodness of fit (ST&D Chapter 20, 21)

- Comparison of the **observed frequency** of occurrence of classes with that **predicted** by a **theoretical model**.
- N classes with *observed frequencies* O_1, O_2, \dots, O_n ; corresponding *expected frequencies* E_1, E_2, \dots, E_n . (E = expected value when the hypothesis is true = $n \cdot$ hypothesized population proportion).

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

- The statistics has a distribution that is distributed \approx as χ^2 with **$n - 1 - p$ df**, where **p** is the number of parameters estimated to calculate the expected frequencies: **$p=0$** for external hypothesis as a genetic segregation 3:1; **$p=2$** for a normal distribution with mean and variance estimated from the sample).
- Then H_0 is rejected at the α level of significance if

$$X^2 = \sum \frac{(O - E)^2}{E} \geq \chi^2_{1-\alpha, n-1}$$

$$H_0: O_1 = E_1, \dots, O_n = E_n$$

$$H_1: O_i \neq E_i \text{ for at least one } i.$$

- **Restrictions:** 1) **$n > 50$**
 2) **No $E = 0$**
 3) **$E < 5$ less than 20% of classes**
- If the last two restrictions are not satisfied, sometimes the problem can be solved by merging some classes
- An adjusted χ^2 (Yate's correction for continuity) can be used when the criterion has a single degree of freedom (2 classes or 2 by 2 table) in order to make the distribution of X^2 more close to a χ^2 distribution

$$\bullet \text{ adjusted } X^2 = \sum \frac{(|O - E| - .5)^2}{E}$$

14.3.1. Example of hand calculation for a one-way classification

Tests of hypotheses

- Test of hypothesis using the χ^2 criterion can be exemplified by tests of 1:1 sex ratio or 3:1 segregation test of dominance in F₂ generation.

Example 14.1 (ST & D p 488)

Suppose a certain F₁ generation of a *Drosophila* species with 35 males and 46 female. Test the hypothesis of a 1:1 sex ratio. (H₀: p = q (with q = 1-p))

Sex	Observed (O)	Expected (E = p*n)	Deviation (O-E)	(O-E) ²	(O-E) ² /E
Male	35	40.5	-5.5	30.25	0.747
Female	46	40.5	5.5	30.25	0.747
Sum	81	81	0	60.5	1.494

- The X^2 value is 1.494
- The **df** = no. of classes - 1 - p = 1
- p = 0 because this is an external hypothesis and no parameter was estimated to calculate the expected values
- From χ^2 Table A.5 (ST&D), the probability for 1.49 with 1 df is between 0.1 and 0.25.
- Therefore, we don't reject the null hypothesis that the sex ratio is 1:1.

Example of **SAS** calculation for a one-way classification

Test the hypothesis of a **9:3:3:1** ratio, normal dihybrid segregation for the data of F₂ progeny of a barley cross (ST&D p500).

Phenotype	Observed	Expected
1 (Green/Six-row)	1178	(9/16) * 1898
2 (Green/Two-row)	291	(3/16) * 1898
3 (Chlorina/Six-row)	273	(3/16) * 1898
4 (Chlorina/Two-row)	156	(1/16) * 1898
Total	1898	1898

```
data f2;
  input pheno count @@;
  cards;
1 1178 2 291 3 273 4 156
proc freq;
  weight count;
  tables pheno / testp = (0.5625, 0.1875, 0.1875, 0.0625);
run; quit;
```

PROC FREQ produces one-way to n-way frequency and contingency tables. One-way frequency tables: computes statistics to test for specified proportions. Contingency tables: computes stats for the relationships between 2 classification var.

WEIGHT: required when the input data are in **count form**. **WEIGHT** names the variable **Count**, which provides the frequency of each class.

TABLES statement: **Pheno** specifies a table where the rows are pheno. In 2-way TABLES A*B specifies a table with A rows and B columns.

PHENO	Frequency	Percent	Test Percent	Cumulative Frequency	Cumulative Percent
1	1178	62.1	56.3	1178	62.1
2	291	15.3	18.8	1469	77.4
3	273	14.4	18.8	1742	91.8
4	156	8.2	6.3	1898	100.0

Chi-Square Test for Specified Proportions

Statistic = 54.313 DF = 3 Prob = 0.001

The number of **df** is one less than the number of classes. We conclude that the data don't follow the ratio of 9:3:3:1 with a $P=0.001$.

14.3.2. Contingency tables

- Two-way tables called **contingency tables** are useful to test if two classification criteria are independent (**test of independence**) and if two samples belong to the same population in relation to one-classification criteria (**test of homogeneity**).
- **Example homogeneity:** Population 1 and Population 2, data is frequency of dark and clear butterflies of species A. Is it OK to merge the two data sets?
- **Example independence:** Ecologists samples 100 trees and register soil (serpentine vs. non-serpentine) and pubescence of leaves (Pub. Vs. smooth).
- If two events are independent, the probability of their occurring together can be computed as the product of their separate probabilities.

- The expected frequency can be obtained by
$$E = \frac{(\text{row total})(\text{column total})}{\text{grand total}}$$
- $\chi^2 df = (n^\circ \text{ of rows}-1)*(n^\circ \text{ of columns}-1)$

Example independence: color and spots in butterflies using a 2 by 2 contingency table. Color: dark or clear, and presence of absence of a black spot in the wings

	No-spot	spot	Total
Dark	1178 >(E ₁ =1123)	291 (E ₂ =346)	1469 (77%)
Clear	273 (E ₃ =328)	156 >(E ₄ =101)	429 (23%)
Total	1451 (78%)	447 (22%)	1898

$$E_1 = 1469 * 1451 / 1898 = 1123$$

```
data f2;
  input c1 $ c2 $ pheno count;
cards;
dark no_spot 1 1178
dark spot 2 291
clear no_spot 3 273
clear spot 4 156
;
proc freq;
  weight count;
  tables c1*c2 / chisq nopercnt nocum norow nocol;
run; quit;
```

CHISQ option requests chi-square statistics for assessing association.

To simplify the output:

NOPERCENT suppresses display of the percentage in cross-tabulation tables

NOCUM suppresses display of cumulative frequencies and cumulative percentages in one-way frequency tables and in list format

NOROW suppresses display of the row percentage for each cell

NOCOL suppresses display of the column percentage for each cell

TABLE OF C1 BY C2

C1	C2		Total
Frequency	no_spot	spot	
1dark	1178	291	1469
2clear	273	156	429
Total	1451	447	1898

STATISTICS FOR TABLE OF C1 BY C2			
Statistic	DF	Value	Prob
Chi-Square	1	50.538	0.001
Continuity Adj. Chi-Square	1	49.623	0.001
Fisher's Exact Test (Left)			1.000
			(Right) 4.60E-12
			(2-Tail) 7.11E-12

Since $P=0.001$, we reject the null hypothesis that two characters are independent. The dark color seems to be associated with the absence of spots

Example 2: A plant ecologist samples 100 trees of a rare species from a 400 square mile area. He records for each tree whether or not it is rooted in serpentine soils and whether its leaves are pubescent or smooth.

Soil	Pubescent	Smooth
Serpentine	12	22
Non serpentine	16	50

Statistic	DF	Value	Prob
Chi-Square	1	1.360	0.244
Continuity Adj. Chi-Square	1	0.867	0.352
Fisher's Exact Test (Left)			0.176
			(Right) 0.918
			(2-Tail) 0.251

Leaf type is independent of the soil type in which the tree is rooted.

14.8. Spearman's coefficient of rank correlation

- The **parametric correlation** coefficient is applicable to **the bivariate normal distribution**.
- If the two variables do not have a bivariate normal distribution, **Spearman's** coefficient can be used to test for the significance of association using a coefficient of **rank correlation**.

- **This coefficient can be used to correlate**

- Order of emergence in a sample of insects with a ranking in size
- Ranking in flower size of roses with order of branching.
- **Rankings of judges**

True order	1	2	3	4	5
Judge 1	1	3	2	5	4

- **The procedure is:**

- 1) Rank the observations for each variable
- 2) Obtain the **differences in ranks** for the paired observations. Let d_i = the difference between the ranks of pair i
- 3) Calculate r_s as follow

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

Where d_i is the difference for the i th pair, and n is the number of pairs. The estimated r_s is compared with the critical value in the following table for samples of 10 or fewer pairs.

Table of **two-tailed** significance levels of the **Spearman Rank Correlation r_s** for $n < 11$. For $n > 10$ use tables for regular Pearson correlation coefficient.

Sample size n	Significance level	
	5%	1%
5	1.000	none
6	0.886	1.00
7	0.786	0.929
8	0.738	0.857
9	0.683	0.817
10	0.648	0.781

If r_s is >10 then Student's distribution with $n-2$ df test the following statistics:

$$t = r_s \sqrt{\frac{n-2}{1-r_s^2}}$$

is used to

Example 14.5 11.2.1 in ST&D p 290. Spearman's rank correlation. Tube length (T) and limb length (L) of flowers of a Nicotiana cross.

Coefficient of rank correlation for data from p290 ex11.2.1

```
data digest;
input t l @@;
cards;
49 27 44 24 32 12 42 22 32 13 53 29 36 14 39 20 37 16
45 21 41 22 48 25 39 18 40 20 34 15 37 20 35 13
;
proc corr; * Pearson's correlation coefficients;
var t l;
proc corr Spearman; *Spearman's correlation coefficients;
var t l;
proc freq;
table t*l / noprint measures cl;
run; quit;
```

```
Spearman Correlation Coefficients, N = 17
Prob > |r| under H0: Rho=0
          t          l
t      1.00000      0.96180
          <.0001
l      0.96180      1.00000
          <.0001
```

PROC FREQ

- The **noprint** option in the TABLES statement suppresses display of the crosstabulation tables but allows display of the requested statistics.
- The option **measures** in table statement enables SAS to print out correlations including **Spearman's** coefficient of rank correlation
- The **CL** option in the TABLES statement, computes asymptotic **confidence limits** ($\alpha=0.05$ default)

Statistic	Value	ASE	95% confidence bounds	
Pearson Correlation	0.9538	0.0202	0.9143	0.9934
Spearman Correlation	0.9618	0.0203	0.9220	1.0000

ASE is the asymptotic standard error

14.4. One sample tests

14.4.1 The Kolmogorov-Smirnov test

- For a single sample of data, the **Kolmogorov-Smirnov** test is used to test whether or not the data is consistent with a specified distribution function.
- Useful nonparametric test for **goodness of fit** for continuous distributions.
- The Shapiro-Wilk's test is used to test for normality for small number of samples and **Kormogorov-Smirnov statistic (D)** is used for large samples.

14.5. Two sample tests

14.5.1 The **sign test** for two **paired** samples

- The sign test is designed to test a hypothesis about the location of a population distribution. The test needs at least 7 pairs (more useful with >20 pairs).
- It is most often used to test the hypothesis about a **population median**, and often involves the use of **matched pairs**.
- The signed numbers serve to test the null hypothesis that each difference has a median of zero (pluses and minuses occur with the same chances). The values, n1 and n2 are the numbers of pluses and minuses.
- This test can be used in place of the one sample t-test when the normality assumption fails. It is less powerful than Wilcoxon signed ranks test (se next), but does not assume that the population probability distribution is symmetric.

$$X^2 = \frac{(n1 - n2)^2}{n1 + n2}, \text{ with } 1 \text{ df}$$

data Score;

```
input Student $ PreTest PostTest @@;
ScoreChange = PostTest - PreTest;
datalines;
```

Capalleti	94	91	Dubose	51	65	Mcbane	75	78	Mullen	89	82
Engles	95	97	Grant	63	75	Nguyen	79	76	Patel	71	77
Krupski	80	75	Lundsford	92	55	Si	75	70	Tanaka	87	73;

```
ods select BasicMeasures TestsForLocation; * ODS=Output Delivery System;
```

```
proc univariate data=Score;
```

```
var ScoreChange;
run; quit;
```

Test	-Statistic-	-----p Value-----
Student's t	t -0.80079	Pr > t 0.4402
Sign	M -1	Pr >= M 0.7744
Signed Rank	S -8.5	Pr >= S 0.5278

14.5.2 Wilcoxon's signed-rank test for paired treatments

- The **Wilcoxon signed-rank** test is an improvement on the sign test in terms of detecting real differences with paired treatments. In the cases where we cannot measure the variable on the absolute scale, but only on ordinal scale, ranking method is very useful.

1. Rank the differences
2. Assign the signs
3. Obtain T_+ and T_- (sum of positive ranks and negative ranks respectively). Choose smaller one and call it T .
4. Compare T with the critical value.

- The test of location for two independent samples of equal size was developed by Wilcoxon. Mann and Whitney extended the test to deal with unequal samples. The test is called *Wilcoxon-Mann-Whitney* two sample test.

14.5.3 The Kolmogorov-Smirnov test for two independent samples

- The null hypothesis is that the two independent samples come from an identical distribution

- 1) Rank all observations in ascending order.
- 2) Determine the sample distribution functions $F_n(Y_1)$ and $F_n(Y_2)$.
- 3) Compute $|F_n(Y_1) - F_n(Y_2)|$ at each Y value.
- 4) Find the value D given by
 $D = \max |F_n(Y_1) - F_n(Y_2)|$, the maximum being taken over all values of Y is compared with a critical value in Tables A.22A (balanced design) and A.22B (unbalanced design).

Computation for a Kolmogorov-Smirnov two-sample test

1. Rank all observations together
2. Determine the sample cumulative distribution function $F_1(Y_1)$ and $F_2(Y_2)$
3. Compute $|F_1(Y_1) - F_2(Y_2)|$ at each of the values
4. Find the maximum difference **D** and compare it with the critical value in Tables A22A (for $n_1=n_2$) or A22B (for $n_1 \neq n_2$) in order to draw a conclusion

Y_1	$F_1(Y_1)$	Y_2	$F_2(Y_2)$	$ F_1(Y_1) - F_2(Y_2) $
53.2	1/7			$ 1/7 - 0 = 1/7$
53.6	2/7			$ 2/7 - 0 = 2/7$
54.4	3/7			$ 3/7 - 0 = 3/7$
56.2	4/7			$ 4/7 - 0 = 4/7$
56.4	5/7			$ 5/7 - 0 = 5/7$
57.8	6/7			$ 6/7 - 0 = \mathbf{6/7 D}$
		58.7	1/6	$ 6/7 - 1/6 = 29/42$
		59.2	2/6	$ 6/7 - 2/6 = 22/42$
		59.8	3/6	$ 6/7 - 3/6 = 15/42$
61.9	7/7			$ 7/7 - 3/6 = 1/2$
		62.5	4/6	$ 7/7 - 4/6 = 1/3$
		63.1	5/6	$ 7/7 - 5/6 = 1/6$
		64.2	6/6	$ 7/7 - 6/6 = 0$

In this case the maximum difference **D** = $6/7 = 0.857$

The critical value in Table A22B for $\alpha = 0.01$ is $5/6 = 0.83$

Since **D** > critical value, we reject H_0

We conclude that the samples belong to different populations

14.5.4. The Wilcoxon-Mann-Whitney for two independent samples

This tests the hypothesis that two data sets have the same location parameter (you can interpret this as the same median). Assume the data sets have size n_1 and n_2 , where $n_1 < n_2$.

- 1) Rank all observations from both observations together from smallest to largest.
- 2) Add the ranks for the smaller sample. Call this T.
- 3) Compute $T' = n_1(n_1 + n_2 + 1) - T$. T' is the value you would get by adding the ranks if the observations are ranked from the largest to the smallest.
- 4) Compare the smaller of T and T' with Table A.18 in ST&D.

In the sheep - cow data (ST&D p96) here are the steps:

1)

Rank	1	2	3	4	5	6	7	8	9	10	11	12	13
Set	S	S	S	S	S	S	C	C	C	S	C	C	C
Value	53.2	53.6	54.4	56.2	56.4	57.8	58.7	59.2	59.8	61.9	62.5	63.1	64.2

2) C is the smaller set, with $n_1 = 6$. $T = 7+8+9+11+12+13 = 60$.

3) $T' = 6(6 + 7 + 1) - 60 = 1+2+3+5+6+7 = 24$.

4) From Table A.18, $24 < 27$ so reject ($P < 0.05$), $24 = 24$ so don't reject ($P = 0.01$). Note that if the data of C8 and S5 were swapped, for example, the null hypothesis would not be rejected at the 0.05 level.

14.6 More than two sample tests (Non parametric CRD)

14.6.1 **Kruskal-Wallis** k-sample test

- The Kruskal-Wallis test is a nonparametric test used to compare three or more samples.
- It is used to test the null hypothesis that all populations have identical distribution functions against the alternative hypothesis that at least two of the samples differ only with respect to location (median), if at all.
- It is the *analogue to the F-test* used in analysis of variance.
- It is a logical extension of the Wilcoxon-Mann-Whitney Test. So, for $k=2$ it is identical to Wilcoxon-Mann-Whitney test.

14.7 Use of SAS for non parametric statistics PROC NPAR1WAY

PROC NPAR1WAY is a nonparametric procedure for testing that the distribution of a variable has the same location parameter across different groups.

In the case of empirical distribution function tests, that the distribution is the same across the different groups.

PROC NPAR1WAY, computes simple linear rank statistics like

- **Kruskal-Wallis** test
- **Wilcoxon-Mann-Whitney** (W-M-W) test for two samples
- **Kolmogorov-Smirnov** that are based on the empirical distribution of the sample.

Here it is applied to the **Coefficients of digestibility of dry matter, feed corn silage, in percentages. (ST&D p96, 579).**

```
data digest;
* nonparametric tests for data from ST&D p96 *;
do animal = 'sheep', 'steers';
  do rep= 1 to 7;
    input dm @@;
    output;
  end;
end;
cards;
57.8 56.2 61.9 54.4 53.6 56.4 53.2
64.2 58.7 63.1 62.5 59.8 59.2 .
;
proc npar1way anova edf wilcoxon data=digest;
  class animal;
  var dm;
run; quit;
```

- **PROC NPAR1WAY** performs nonparametric tests for location and scale differences across a one-way classification.
- **PROC NPAR1WAY** performs tests for location and scale differences based on the following scores of a response variable: **Wilcoxon**, median, **Kruskal-Wallis**, and others.
- **PROC NPAR1WAY** also provides empirical distribution function (EDF) statistics, which test whether the distribution of a variable is the same across different groups. These include the **Kolmogorov-Smirnov** test

The NPAR1WAY Procedure

Parametric Analysis of Variance for Variable dm
Classified by Variable animal

animal	N	Mean
sheep	7	56.214286
steer	6	61.250000

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Among	1	81.927198	81.927198	11.1834	0.0065
Within	11	80.583571	7.325779		

Parametric ANOVA suggests rejection of null hypothesis under the assumption of normality

Wilcoxon Scores (Rank Sums) for Variable DM
Classified by Variable ANIMAL

ANIMAL	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
sheep	7	31.0	49.0	7.0	4.4285714
steer	6	60.0	42.0	7.0	10.0000000

Wilcoxon 2-Sample Test (Normal Approximation)

(with Continuity Correction of .5)

S = 60.0000 Z = 2.50000 Prob > |Z| = 0.0124

T-Test Approx. Significance = **0.0279**

Output of Wilcoxon-Mann-Whitney test. Sum of ranks within each group is produced

Kruskal-Wallis Test (Chi-Square Approximation)

CHISQ = 6.6122 DF = 1 Prob > CHISQ = 0.0101

Kolmogorov-Smirnov Test for Variable dm

Classified by Variable animal

animal	N	EDF at Maximum	Deviation from Mean at Maximum
sheep	7	0.857143	1.046671
steer	6	0.000000	-1.130534
Total	13	0.461538	

Maximum Deviation Occurred at Observation 1

Value of dm at Maximum = 57.80

Kolmogorov-Smirnov Two-Sample Test (Asymptotic)

KS 0.427302 D 0.857143
KSa 1.540658 Pr > KSa 0.0174