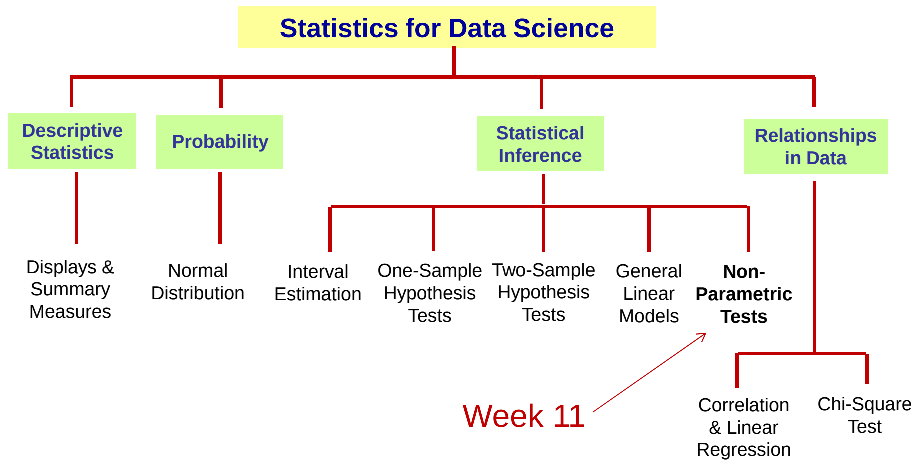
MATH 4044
Statistics for Data Science

Non-Parametric Tests

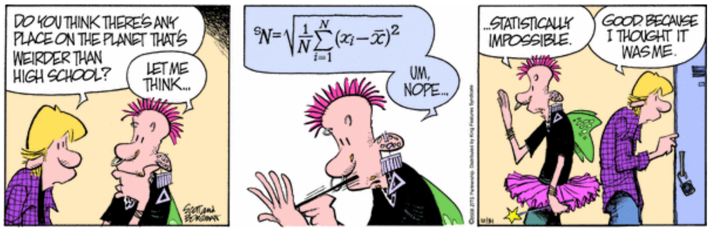
Course outline



Field, A & Miles, J, *Discovering Statistics Using SAS*, Chapter 15, 15.1-15.5.

Topics to be covered

- Non-Parametric Statistics
  - Wilcoxon rank-sum test and Mann-Whitney U test
  - Wilcoxon signed-rank test
  - Kruskal-Wallis test
  - Spearman correlation coefficient



Non-parametric procedures

- Analysis of ranks rather than the actual data.
- Two independent samples:
  - Wilcoxon rank-sum or Mann-Whitney test
  - Populations are not Normal *OR* outliers are present
- One-sample or matched pairs problems:
  - Wilcoxon signed-rank test
  - Population is not Normal *OR* outliers are present
- More than two independent samples:
  - Kruskall-Wallis test
  - Populations are not Normal *OR* variances are not approx equal.

Non-parametric vs parametric procedures

- When all the assumptions of the ‘classical’ procedures are met, it is disadvantageous to apply non-parametric methods:
  - ☐ The researcher will not take full advantage of data.
  - ☐ Non-parametric methods will waste part of the sample information.

Example: Page-to-screen adaptations

- The first three seasons of HBO's Game of Thrones covered about 1,980 pages of the book series, with more than 26.5 total hours on screen.
- According to data from *The Guardian*, however, the show actually spends less time per page than many other recent book-to-TV adaptations.
- Minutes spent per page in recent book to TV or film adaptations were recorded.
- Is there a difference between film and TV adaptations?

Example: Page-to-screen adaptations

Film adaptations

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.85672	Pr < W	0.0019
Kolmogorov-Smirnov	D	0.148862	Pr > D	0.1416
Cramer-von Mises	W-Sq	0.150769	Pr > W-Sq	0.0222
Anderson-Darling	A-Sq	1.069198	Pr > A-Sq	0.0072

The MEANS Procedure

Analysis Variable : MPP						
Type	**N Obs**	**N**	**Mean**	**Std Dev**	**Minimum**	**Maximum**
Film	26	26	0.469	0.282	0.087	1.371
TV	25	25	2.703	5.641	0.264	27.429

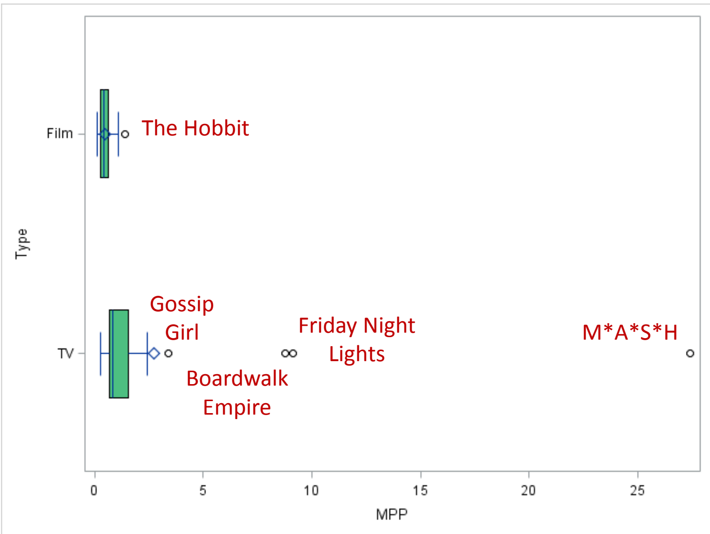
TV adaptations

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.436152	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.371794	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.099771	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	5.466564	Pr > A-Sq	<0.0050

Neither distribution is Normal.

A two-sample t-test is not appropriate.

Example: Page-to-screen adaptations



The Wilcoxon rank-sum & Mann-Whitney U test

- Non-parametric alternative to a two-sample *t*-test for independent samples.
- Assumptions:
  - No assumptions are made about the shape of the population distributions.
- Hypotheses:

  H_0: $\text{median}_1 = \text{median}_2$ (Distributions are similar)
  H_1: $\text{median}_1 \neq \text{median}_2$ (Distributions are different)

Basis for Wilcoxon rank-sum and Mann-Whitney

- Observations are ranked from lowest to highest, ignoring the sample to which they belong.
- If there is no difference, we would expect to find a similar number of high and low ranks in each sample.
  - Summed ranks for both samples should be about the same.
- If there is a difference, we would expect the data from one sample to be concentrated at one end of the scale, and the data from the other sample at the other end.
  - Summed rank for one sample would be higher.

Example: Page-to-screen adaptations

Type	MPP	Potential rank	Actual rank
Film	0.09	1	1
Film	0.18	2	2
Film	0.21	3	3
Film	0.25	4	4.5
Film	0.25	5	4.5
Film	0.26	6	7
Film	0.26	7	7
TV	0.26	8	7
Film	0.28	9	9
Film	0.31	10	10
Film	0.32	11	11
TV	0.33	12	12.5
TV	0.33	13	12.5
Film	0.35	14	14
Film	0.36	15	15

[First 15 observations]

In case of ties, a rank that is the average of potential ranks for these scores.

Sum of ranks for TV = 876
Sum of ranks for Film = 450

Test statistic for Wilcoxon is based on these (smaller value if $n_1 = n_2$)

Procedure for Mann-Whitney U test is very similar.

Example: Page-to-screen adaptations

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable MPP Classified by Variable Type

Type	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Film	26	450.0	676.0	53.071077	17.307692
TV	25	876.0	650.0	53.071077	35.040000

Average scores were used for ties.

Wilcoxon Two-Sample Test	
Statistic (S)	876.0000
Normal Approximation	
Z	4.2490
One-Sided Pr > Z	<.0001
Two-Sided Pr > |Z|	<.0001
t Approximation	
One-Sided Pr > Z	<.0001
Two-Sided Pr > |Z|	<.0001
Exact Test	
One-Sided Pr >= S	<.0001
Two-Sided Pr >= |S - Mean|	<.0001

Z includes a continuity correction of 0.5.

Test statistic W (SAS calls it S)

Since its value is greater than the expected value under the null hypothesis, SAS displays the right-sided P-values.

z-score = (Test statistic – Mean)/SE

Example: Page-to-screen adaptations

The MEANS Procedure

Analysis Variable : MPP		
Type	N Obs	Median
Film	26	0.422
TV	25	0.846

- Minutes per page for TV adaptations (*Mdn* = 0.846) were significantly higher than for film adaptations (*Mdn* = 0.422), W_S = 876, *z* = 4.2584, P-value < 0.0001.
- There is a significant difference between the underlying distributions of minutes per page for TV and film adaptations.

Example: SAS code

```
proc npar1way data=work.adaptations Wilcoxon correct=no;
      class Type;
      var MPP;
      exact Wilcoxon;
      run;
```

Add this option to not include the continuity correction

If you have small sample sizes, you might choose to include an `exact` statement.

It causes an exact P-value to be calculated. Do not use this statement if you have large sample sizes, because the processing time can become excessive.

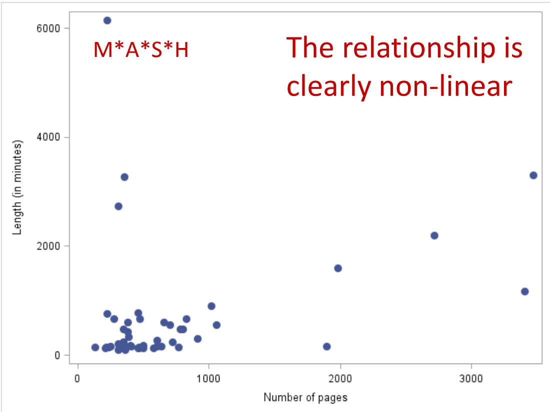
Spearman's correlation coefficient

- The Spearman's rank correlation coefficient measures the strength of curved relationships between two quantitative variables that are strictly increasing or decreasing.
  - Also used when outliers are present.
- It is denoted by r_s or ρ (rho) and calculated by first ranking the data for each quantitative variable and then applying the linear correlation coefficient formula.
- A non-parametric alternative to Pearson's correlation coefficient.

Example: Page-to-screen adaptations

Spearman Correlation Coefficients, N = 51 Prob > |r| under H0: Rho=0		
	Pages	Length
Pages	1.00000	0.31280 0.0254
Length	0.31280 0.0254	1.00000

$H_0: \rho = 0$
$H_1: \rho \neq 0$
$\alpha = 0.05$



There is a statistically significant positive non-linear association between length of the adaptation and the number of pages.

Example: Standardised tests

- A data file contains 200 observations from a sample of high school students with demographic information about the students, such as their gender and socio-economic status.
- It also contains a number of scores on standardized tests, including tests of reading and writing.
- Is there a statistically significant difference between reading and writing scores?

The Wilcoxon signed-rank test

- Used when two sets of measurements are to be compared, but these measurements come from the same subjects (paired samples).
- Non-parametric alternative to a paired t-test.
- Assumptions:
  - There are no assumptions about the distribution shape.
- Hypotheses:

  H_0: median difference = 0
  H_1: median difference $\neq 0$ (or $<$, or $>$)

Basis for Wilcoxon signed-rank test

- Differences between measurements are calculated and ranked.
- The sign of the difference is assigned to each rank, excluding the ties.
- If the null hypothesis is true we would expect approx equal number of + and - signs.
- If either + or - signs predominate, there is evidence that the null hypothesis is false.

Basis for Wilcoxon signed-rank test

- Two sums of ranks are obtained:
  - The sum of the ranks of the difference scores which were positive and the sum of the ranks of the difference scores which were negative.
  - The test statistic, T, is the smaller of these two sums for a non-directional test.
  - For a directional test, the test statistic is the sum which you predicted would be smaller.
- Difference scores of zero are usually discarded from the analysis (prior to ranking), which may bias the test against the null hypothesis.

Example: Standardised tests

The UNIVARIATE Procedure
Variable: difference

Tests for Location: Mu0=0				
Test		Statistic	p Value	
Student's t	t	-0.86731	Pr > |t|	0.3868
Sign	M	-4.5	Pr >= |M|	0.5565
Signed Rank	S	-658.5	Pr >= |S|	0.3677

Test statistic

H_0: median difference = 0
H_1: median difference ≠ 0
$\alpha = 0.05$

Since the P-value = 0.3677 > 0.05, H_0 can't be rejected.

The results suggest that there is no statistically significant difference between reading and writing scores.

Example: SAS code

In SAS there is no direct way to run a non-parametric paired comparison.

We first need to compute a variable that represents the difference of the paired values using a DATA step, and then use PROC UNIVARIATE.

```
data work.hsb2_differences;
  set work.hsb2;
  difference = read - write;
run;

ods select testsforlocation;

proc univariate data = work.hsb2_differences;
  var difference;
run;
```

Restrict the output to the section called 'Tests for Location'

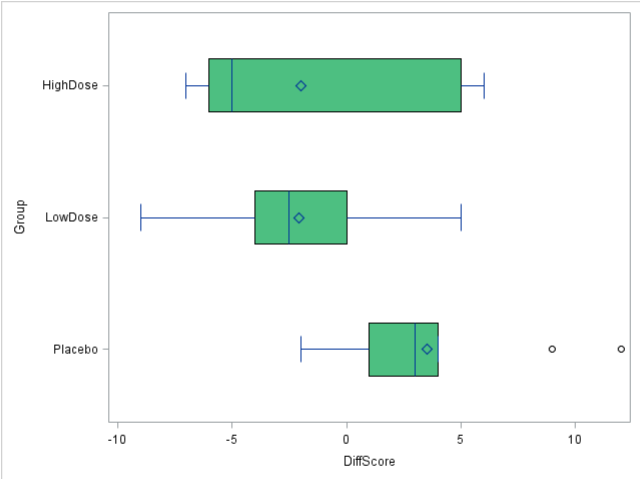
Example: Standardised tests

- If we believed the differences between reading and writing scores were not ordinal but could merely be classified as positive and negative, then we may want to consider a sign test instead of sign rank test.
- Note that the SAS output gives us the results for both the Wilcoxon signed rank test and the sign test without having to use any options.
- Using the sign test, we again conclude that there is no statistically significant difference between reading and writing scores (P-value = 0.5565).

Example: Vitamin C and cold prevention

- In an experiment, thirty volunteers were randomly assigned to take either a placebo, a low dose of vitamin C or a high dose of vitamin C.
- The difference in the number of days of cold symptoms (one year to the next) was recorded.
- Is there a relationship between the amount of vitamin C taken and the change in the number of days that individuals show cold symptoms?

Example: Vitamin C and cold prevention



What can we say about the shape of the distributions?

Example: Vitamin C and cold prevention

High dose

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.761094	Pr < W	0.0049
Kolmogorov-Smirnov	D	0.267396	Pr > D	0.0418
Cramer-von Mises	W-Sq	0.182004	Pr > W-Sq	0.0072
Anderson-Darling	A-Sq	1.077038	Pr > A-Sq	<0.0050

Placebo

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.894913	Pr < W	0.1925
Kolmogorov-Smirnov	D	0.251973	Pr > D	0.0717
Cramer-von Mises	W-Sq	0.106993	Pr > W-Sq	0.0810
Anderson-Darling	A-Sq	0.558998	Pr > A-Sq	0.1125

Low dose

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.96919	Pr < W	0.8832
Kolmogorov-Smirnov	D	0.190193	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.045133	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.247349	Pr > A-Sq	>0.2500

Differences for the low dose and placebo groups can be assumed to be Normal.

Differences for the high dose group are significantly non-Normal.

A parametric ANOVA procedure is not appropriate

The Kruskal-Wallis test

- There are more than two independent samples.
- Population distributions are *not Normal* or *variances are not equal*.
- The hypotheses are:

  H_0: All medians are equal

  H_1: Not all medians are equal
- The idea:
  - Rank all the responses from all groups together and then apply a procedure similar to one-way ANOVA to the ranks rather than the original observations.

Example: Vitamin C and cold prevention

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable DiffScore Classified by Variable Group

Group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Placebo	10	214.50	155.0	22.616239	21.450
LowDose	10	126.00	155.0	22.616239	12.600
HighDose	10	124.50	155.0	22.616239	12.450

Average scores were used for ties.

Kruskal-Wallis Test	
Chi-Square	6.9229
DF	2
Pr > Chi-Square	0.0314

H_0: All medians are equal

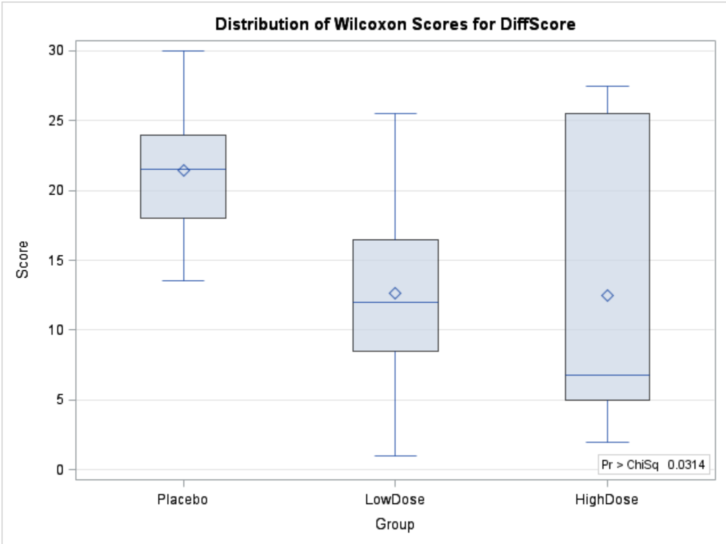
H_1: Not all medians are equal

$\alpha = 0.05$

Test statistic *H* (labelled chi-square because of its distribution)

Since the P-value = 0.031 < 0.05, H_0 is rejected.

Example: Vitamin C and cold prevention



Example: Vitamin C and cold prevention

- The results of the analysis indicate that there is a significant difference in the medians (H = 6.92 has chi-square distribution with 2 df, *P-value* = 0.031).
- Because the overall test is significant, pairwise comparisons among the three groups should be completed.
  - For comparisons with placebo, two Wilcoxon tests are performed with $\alpha = 0.05/2 = 0.025$.

Example: Vitamin C and cold prevention

Wilcoxon Scores (Rank Sums) for Variable DiffScore Classified by Variable Group					
Group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Placebo	10	139.0	105.0	13.148944	13.90
LowDose	10	71.0	105.0	13.148944	7.10
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	139.0000
Normal Approximation	
Z	2.5477
One-Sided Pr > Z	0.0054
Two-Sided Pr > |Z|	0.0108
t Approximation	
One-Sided Pr > Z	0.0098
Two-Sided Pr > |Z|	0.0197
Z includes a continuity correction of 0.5.	

Placebo vs **low** dose of vitamin C

Since the P-value is less than 0.025, the difference between placebo and a low dose of vitamin C is statistically significant.

Example: Vitamin C and cold prevention

Wilcoxon Scores (Rank Sums) for Variable DiffScore Classified by Variable Group					
Group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Placebo	10	130.50	105.0	13.148944	13.050
HighDose	10	79.50	105.0	13.148944	7.950
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	130.5000
Normal Approximation	
Z	1.9013
One-Sided Pr > Z	0.0286
Two-Sided Pr > |Z|	0.0573
t Approximation	
One-Sided Pr > Z	0.0363
Two-Sided Pr > |Z|	0.0725
Z includes a continuity correction of 0.5.	

Placebo vs **high** dose of vitamin C

Since the P-value is higher than 0.025, the difference between placebo and a high dose of vitamin C is not statistically significant.

The effect we have observed seems to be due to low doses only.

Example: SAS code

```
proc npar1way data=work.vitaminc wilcoxon;
   class Group;
   var DiffScore;
   run;

data work.low_placebo;
   set work.vitaminc;
   if _N_ > 20 then delete;
   run;

proc npar1way data=work.low_placebo wilcoxon;
   class Group;
   var DiffScore;
   run;

data work.high_placebo;
   set work.vitaminc;
   if _N_ > 10 and _N_ < 21 then delete;
   run;

proc npar1way data=work.high_placebo wilcoxon;
   class Group;
   var DiffScore;
   run;
```

Kruskal-Wallis test

Wilcoxon test to compare low dose to placebo

Wilcoxon test to compare high dose to placebo

Converting data values into ranks

- When the data is not Normally distributed, one option is to replace the data values with ranks, and then perform parametric tests on ranks.
- For example, for the page-to-screen adaptations data, we could have done the following:

```
proc rank data=work.adaptations out=rank_mpp;
   var MPP;
   ranks Rank_of_MPP;
   run;

proc ttest data=work.rank_mpp;
   class Type;
   var Rank_of_MPP;
   run;
```

If you do not include a `ranks` statement, variables listed on the `var` statement will be replaced with their ranks.