BAU
BAHÇEŞEHİR ÜNİVERSİTESİ

BME 3005
BIOSTATISTICS

Lecture 9: Kruskal Wallis, Wilcoxon Signed-Rank Test, Friedman Test

Burcu Tunç Çamlıbel

BAU
BAHÇEŞEHİR ÜNİVERSİTESİ

Chapter 10
Alternatives to Analysis of Variance and the t test Based on Ranks

Summary of Some Statistical Methods to Test Hypotheses

Scale of measurement	Type of experiment: Two treatment groups consisting of different individuals	Three or more treatment groups consisting of different individuals	Before and after a single treatment in the same individuals	Multiple treatments in the same individuals	Association between two variables
Interval (and drawn from normally distributed populations*)	Unpaired *t* test (Chapter 4)	Analysis of variance (Chapter 3)	Paired *t* test (Chapter 9)	Repeated-measures analysis of variance (Chapter 9)	Linear regression, Pearson product-moment correlation, or Bland-Altman analysis (Chapter 8)
Nominal	Chi-square analysis-of-contingency table (Chapter 5)	Chi-square analysis-of-contingency table (Chapter 5)	McNemar's test (Chapter 9)	Cochrane Q†	Relative rank or odds ratio (Chapter 5)
Ordinal‡	Mann-Whitney rank-sum test (Chapter 10)	Kruskal-Wallis statistic (Chapter 10)	Wilcoxon signed-rank test (Chapter 10)	Friedman statistic (Chapter 10)	Spearman rank correlation (Chapter 8)
Survival time	Log-rank test or Gehan's test (Chapter 11)				

*If the assumption of normally distributed populations is not met, rank the observations and use the methods for data measured on an ordinal scale.

†Not covered in this text.

‡Or interval data that are not necessarily normally distributed.

TWO DIFFERENT SAMPLES: THE MANN-WHITNEY RANK-SUM TEST

The procedure for testing the hypothesis that a treatment had no effect with this statistic is:

- *Rank all observations according to their magnitude, a rank of 1 being assigned to the smallest observation. Tied observations should be assigned the same rank, equal to the average of the ranks they would have been assigned had there been no tie (i.e., using the same procedure as in computing the Spearman rank correlation coefficient in Chapter 8).*
- *Compute T, the sum of the ranks in the smaller sample. (If both samples are the same size, you can compute T from either one.)*
- *Compare the resulting value of T with the distribution of all possible rank sums for experiments with samples of the same size to see whether the pattern of rankings is compatible with the hypothesis that the treatment had no effect. (use table 10.3 page 210 if sample size is smaller than 8) or calcute z_T (next slide) (use table 4.3)*

TWO DIFFERENT SAMPLES: THE MANN-WHITNEY RANK-SUM TEST

$$\mu_T = \frac{n_S(n_S + n_B + 1)}{2}$$

and standard deviation

$$\sigma_T = \sqrt{\frac{n_S n_B (n_S + n_B + 1)}{12}}$$

$$z_T = \frac{|T - \mu_T| - \frac{1}{2}}{\sigma_T}$$

n_S is the size of the smaller sample.*

compare this statistic with the critical values of the normal distribution that define the, say 5%, most extreme possible values.

z_T can also be compared with the t distribution with an infinite number of degrees of freedom Table 4-1) because it is equals to the normal distribution.

Example page 211: USE OF A CANNABIS-BASED MEDICINE IN PAINFUL DIABETIC NEUROPATHY

USE OF A CANNABIS-BASED MEDICINE IN PAINFUL DIABETIC NEUROPATHY

TABLE 10-4. Diabetic Neuropathy Pain among People Treated with a Placebo and a Cannabis Medicinal

Placebo		Cannabis Medicinal	
Observation	Rank	Observation	Rank
13	16	90	50
8	6.5	10	9.5
46	39	45	38
61	44	70	45.5
28	31.5	13	16
7	4	27	30
93	51	11	11
10	9.5	70	45.5
7	4	14	19
100	53	15	20
4	1.5	13	16
16	21	75	47
23	27	50	40
33	35	30	34
18	22	80	48
51	41	40	37
26	29	29	33
19	23.5	13	16
20	25.5	9	8
54	42	7	4
19	23.5	20	25.5
37	36	85	49
13	16	55	43
8	6.5	94	52
28	31.5		
25	28		$T = 737$
4	1.5		
12	12.5		
12	12.5		

USE OF A CANNABIS-BASED MEDICINE IN PAINFUL DIABETIC NEUROPATHY

The cannabis medicinal group is the smaller sample, so we compute the test statistic T by summing all the ranks in that group, yielding $T = 737$. The cannabis group has $n_S = 24$ people in it and the larger placebo group, $n_B = 29$, so the mean value of T for all studies of this size is

$$\mu_T = \frac{n_S(n_S + n_B + 1)}{2} = \frac{24(24 + 29 + 1)}{2} = 648$$

and the standard deviation is

$$\sigma_T = \sqrt{\frac{n_S n_B(n_S + n_B + 1)}{2}} = \sqrt{\frac{24 \cdot 29(24 + 29 + 1)}{12}} = 55.96$$

So

$$z_T = \frac{|T - \mu_T| - \frac{1}{2}}{\sigma_T} = \frac{|737 - 648| - \frac{1}{2}}{55.96} = 1.581$$

This value is smaller than 1.960, the value of z that defines the most extreme 5% of the normal distribution (from Table 4-1). Hence, this study does not provide substantial evidence that the cannabis medicinal was any more or less effective than placebo in controlling pain associated with diabetic neuropathy.

EACH SUBJECT OBSERVED BEFORE AND AFTER ONE TREATMENT: THE WILCOXON SIGNED-RANK TEST

The procedure for comparing the observed effects of a treatment in a single group of experimental subjects before and after administering a treatment:

- *Compute the change in the variable of interest in each experimental subject.*
- *Rank all the differences according to their magnitude without regard for sign. (Zero differences should be dropped from the analysis with a corresponding reduction of sample size. Tied ranks should be assigned the average of the ranks that would be assigned to the tied ranks if they were not tied.)*
- *Apply the sign of each difference to its rank.*
- *Add all the signed ranks to obtain the test statistic W.**
- *Compare the observed value of W with the distribution of possible values that would occur if the treatment had no effect, and reject this hypothesis if W is "big."*

$$Z_W = \frac{|W| - \frac{1}{2}}{\sqrt{[n(n+1)(2n+1)]/6}}$$

in which *n* equals the number of experimental subjects.

See page 217: Example «Cigarette Smoking and Platelet Function»

EACH SUBJECT OBSERVED BEFORE AND AFTER ONE TREATMENT: THE WILCOXON SIGNED-RANK TEST: Example

TABLE 10-5. Effect of a Potential Diuretic on Six People

Person	Daily Urine Production (mL/d)			Rank* of Difference	Signed Rank of Difference
	Before Drug	After Drug	Difference		
1	1600	1490	−110	5	−5
2	1850	1300	−550	6	−6
3	1300	1400	+100	4	+4
4	1500	1410	−90	3	−3
5	1400	1350	−50	2	−2
6	1010	1000	−10	1	−1
					W = −13

*1 = smallest magnitude; 6 = largest magnitude.

EACH SUBJECT OBSERVED BEFORE AND AFTER ONE TREATMENT: THE WILCOXON SIGNED-RANK TEST

TABLE 10-7. Critical Values (Two-Tailed) of Wilcoxon *W*

n	Critical Value	*P*
5	15	.062
6	21	.032
	19	.062
7	28	.016
	24	.046
8	32	.024
	28	.054

Data from Table A-11 of Mosteller F, Rourke R. *Sturdy Statistics: Nonparametrics and Order Statistics*. Reading, MA: Addison-Wesley; 1973.

Table 10-7 presents the values of *W* that come closest to defining the most extreme 5% and 1% of all possible values for experiments with up to 8 subjects.

EACH SUBJECT OBSERVED BEFORE AND AFTER ONE TREATMENT: THE WILCOXON SIGNED-RANK TEST: Example

TABLE 10-8. Maximum Percentage Platelet Aggregation before and after Smoking One Cigarette

Person	Before Smoking	After Smoking	Difference	Rank of Difference	Signed Rank of Difference
1	25	27	2	2	2
2	25	29	4	3.5	3.5
3	27	37	10	6	6
4	44	56	12	7	7
5	30	46	16	10	10
6	67	82	15	8.5	8.5
7	53	57	4	3.5	3.5
8	53	80	27	11	11
9	52	61	9	5	5
10	60	59	−1	1	−1
11	28	43	15	8.5	8.5
					$W = 64$

Calculate zw (because you have 11 samples), and then check your z value from t-table with infinity degrees of feedom.

EXPERIMENTS WITH THREE OR MORE GROUPS WHEN EACH GROUP CONTAINS DIFFERENT INDIVIDUALS: THE KRUSKAL-WALLIS TEST

The Kruskal-Wallis test is a direct generalization of the Mann-Whitney rank-sum test. The procedure for analyzing an experiment in which different groups of experimental subjects receive each treatment is:

- *Rank each observation without regard for treatment group, beginning with a rank of 1 for the smallest observation. (Ties are treated in the same way as the other rank tests.*)*
- *Compute the Kruskal-Wallis test statistic H to obtain a normalized measure of how much the average ranks within each treatment group deviate from the average rank of all the observations.*
- *Compare H with χ2 distribution with 1 less degree of freedom than the number of treatment groups. If H exceeds the critical value that defines a "big" H, reject the null hypothesis that the treatment has no effect.*

- *See page 218-222 (important)*
- *When you find a statistical difference within the treated groups, follow with multiple comparison tests. (In this case: Mann-Whitney Rank sum test with beforroni/Holm correction)*

EXPERIMENTS WITH THREE OR MORE GROUPS WHEN EACH GROUP CONTAINS DIFFERENT INDIVIDUALS: THE KRUSKAL-WALLIS TEST

$$\bar{R} = \frac{1+2+3+\cdots+N}{N} = \frac{N+1}{2} \qquad \bar{R}_1 = R_1/n_1,\ \bar{R}_2 = R_2/n_2,\ \text{and} \qquad \bar{R}_3 = R_3/n_3$$

$$D = n_1(\bar{R}_1 - \bar{R})^2 + n_2(\bar{R}_2 - \bar{R})^2 + n_3(\bar{R}_3 - \bar{R})^2$$

$$H = \frac{D}{N(N+1)/12} = \frac{12}{N(N+1)}\sum n_t(\bar{R}_t - \bar{R})^2$$

with $\nu = k - 1$ degrees of freedom

The summation denoted with Σ is over all the treatment groups, regardless of how many treatment groups there are.

Use chi-square distribution.

EXPERIMENTS WITH THREE OR MORE GROUPS WHEN EACH GROUP CONTAINS DIFFERENT INDIVIDUALS: THE KRUSKAL-WALLIS TEST: Example

Prenatal Marijuana Exposure and Child Behavior

Lidush Goldschmidt and colleagues† designed a prospective observational study to track children whose mothers used marijuana during pregnancy.

They interviewed women who came to a prenatal clinic and attempted to recruit all women who used two or more joints of marijuana per month during the first trimester of pregnancy and a random selection of other pregnant women who did not smoke marijuana.

They kept in touch with these women, then evaluated temperament and behavioral characteristics of the children when they were 10 years old.

One of the assessments used to address attention deficit disorder and hyperactivity was the Swanson, Noland, and Pelham (SNAP) checklist, which is a questionnaire completed by mothers.

EXPERIMENTS WITH THREE OR MORE GROUPS WHEN EACH GROUP CONTAINS DIFFERENT INDIVIDUALS: THE KRUSKAL-WALLIS TEST: Example

Prenatal Marijuana Exposure and Child Behavior

TABLE 10-9. Average Number of Joints Per Day (AJD)

	AJD = 0 $n_1 = 13$		$0 < AJD \leq 0.89$ $n_1 = 9$		AJD > 0.89 $n_1 = 9$	
	SNAP Score	Rank	SNAP Score	Rank	SNAP Score	Rank
	7.79	4	8.84	12	8.65	11
	9.16	17	9.92	24	10.70	31
	7.34	2	7.20	1	10.24	28
	10.28	29	9.25	20	8.62	10
	9.12	15	9.45	21	9.94	25
	9.24	19	9.14	16	10.55	30
	8.40	7	9.99	26	10.13	27
	8.60	9	9.21	18	9.78	23
	8.04	5	9.06	14	9.01	13
	8.45	8				
	9.51	22				
	8.15	6				
	7.69	3				
Sum of ranks, R_t		146		152		198
Mean rank, $R_t = R_t/n_t$		11.23		16.89		22.00

EXPERIMENTS WITH THREE OR MORE GROUPS WHEN EACH GROUP CONTAINS DIFFERENT INDIVIDUALS: THE KRUSKAL-WALLIS TEST: Example

Prenatal Marijuana Exposure and Child Behavior

$$\bar{R} = \frac{1+2+3+\cdots+31}{31} = \frac{N+1}{2} = \frac{31+1}{2} = 16$$

Therefore, the weighted sum of squared deviations between the average ranks observed in each treatment group and the average of all ranks is

$$D = 13(11.23-16)^2 + 9(16.89-16)^2 + 9(22.00-16)^2$$
$$= 13(-4.77)^2 + 9(0.89)^2 + 9(6.00)^2 = 626.92$$

and, so,

$$H = \frac{D}{N(N+1)/12} = \frac{626.92}{31(31+1)/12} = 7.58$$

This value exceeds 5.991, the value that defines the largest 5% of the χ^2 distribution with $\nu = k-1 = 3-1 = 2$ degrees of freedom (from Tables 5-7). Therefore, we conclude that at least one of these three groups differed in hyperactivity and attention deficit ($P < .05$).

EXPERIMENTS WITH THREE OR MORE GROUPS WHEN EACH GROUP CONTAINS DIFFERENT INDIVIDUALS: THE KRUSKAL-WALLIS TEST: Example

Prenatal Marijuana Exposure and Child Behavior

Nonparametric Multiple Comparisons

TABLE 10-10. Mann-Whitney Tests for Multiple Comparisons for Marijuana Exposure and Child Behavior

None vs. Low				None vs. High				Low vs. High			
SNAP Score	Rank	SNAP Score	Rank	SNAP Score	Rank	SNAP Score	Rank	SNAP Score	Rank	SNAP Score	Rank
None		Low		None		High		Low		High	
7.79	4	8.64	10	7.79	3	8.65	10	8.64	3	8.65	4
9.16	14	9.92	20	9.16	13	10.70	22	9.92	12	10.70	18
7.34	2	7.20	1	7.34	1	10.24	19	7.20	1	10.24	16
10.28	22	9.25	17	10.28	20	8.62	9	9.25	9	8.62	2
9.12	12	9.45	18	9.12	12	9.94	17	9.45	10	9.94	13
9.24	16	9.14	13	9.24	14	10.55	21	9.14	7	10.55	17
8.40	7	9.99	21	8.40	6	10.13	18	9.99	14	10.13	15
8.60	9	9.21	15	8.60	8	9.78	16	9.21	8	9.78	11
8.04	5	9.06	11	8.04	4	9.01	11	9.06	6	9.01	5
8.45	8			8.45	7						
9.51	19			9.51	15						
8.15	6			8.15	5						
7.69	3			7.69	2						
$n_B = 13$		$n_S = 9$	$T = 126$	$n_B = 13$		$n_S = 9$	$T = 143$	$n_B = 9$	$T = 70$	$n_S = 9$	

$$\mu_T = \frac{9(9+13+1)}{2} = 103.5$$

$$\sigma_T = \sqrt{\frac{9 \cdot 13(9+13+1)}{12}} = 14.97$$

$$z_T = \frac{|126 - 103.5| - \frac{1}{2}}{14.97} = 1.470$$

$.20 < P < .10$

$$\mu_T = \frac{9(9+13+1)}{2} = 103.5$$

$$\sigma_T = \sqrt{\frac{9 \cdot 13(9+13+1)}{12}} = 14.97$$

$$z_T = \frac{|143 - 103.5| - \frac{1}{2}}{14.97} = 2.605$$

$P < .01$

$$\mu_T = \frac{9(9+9+1)}{2} = 85.5$$

$$\sigma_T = \sqrt{\frac{9 \cdot 9(9+9+1)}{12}} = 11.32$$

$$z_T = \frac{|70 - 85.5| - \frac{1}{2}}{11.32} = 1.325$$

$.20 < P < .10$

Summary of Some Statistical Methods to Test Hypotheses

Scale of measurement	Type of experiment: Two treatment groups consisting of different individuals	Three or more treatment groups consisting of different individuals	Before and after a single treatment in the same individuals	Multiple treatments in the same individuals	Association between two variables
Interval (and drawn from normally distributed populations*)	Unpaired *t* test (Chapter 4)	Analysis of variance (Chapter 3)	Paired *t* test (Chapter 9)	Repeated-measures analysis of variance (Chapter 9)	Linear regression, Pearson product-moment correlation, or Bland-Altman analysis (Chapter 8)
Nominal	Chi-square analysis-of-contingency table (Chapter 5)	Chi-square analysis-of-contingency table (Chapter 5)	McNemar's test (Chapter 9)	Cochrane Q†	Relative rank or odds ratio (Chapter 5)
Ordinal‡	Mann-Whitney rank-sum test (Chapter 10)	Kruskal-Wallis statistic (Chapter 10)	Wilcoxon signed-rank test (Chapter 10)	Friedman statistic (Chapter 10)	Spearman rank correlation (Chapter 8)
Survival time	Log-rank test or Gehan's test (Chapter 11)				

*If the assumption of normally distributed populations is not met, rank the observations and use the methods for data measured on an ordinal scale.

†Not covered in this text.

‡Or interval data that are not necessarily normally distributed.

EXPERIMENTS IN WHICH EACH SUBJECT RECEIVES MORE THAN ONE TREATMENT: THE FRIEDMAN TEST

The procedure for using the Friedman statistic to analyze experiments in which the same individuals receive several treatments is as follows:

- Rank each observation within each experimental subject, assigning 1 to the smallest response. (Treat ties as before.)

- Compute the sum of the ranks observed in all subjects for each treatment.

- Compute the Friedman test statistic $\chi r2$ as a measure of how much the observed rank sums differ from those that would be expected if the treatments had no effect.

- Compare the resulting value of the Friedman statistic with the χ 2 distribution. *if the experiment involves large enough samples or with the exact distribution of* χr 2 *in Table 10-14 if the sample is small.*

EXPERIMENTS IN WHICH EACH SUBJECT RECEIVES MORE THAN ONE TREATMENT: THE FRIEDMAN TEST

TABLE 10-14. Critical Values for Friedman χ_r^2

$k = 3$ Treatments			$k = 4$ Treatments		
n	χ_r^2	P	n	χ_r^2	P
3	6.00	.028	2	6.00	.042
4	6.50	.042	3	7.00	.054
	8.00	.005		8.20	.017
5	5.20	.093	4	7.50	.054
	6.40	.039		9.30	.011
	8.40	.008	5	7.80	.049
6	5.33	.072		9.96	.009
	6.33	.052	6	7.60	.043
	9.00	.008		10.20	.010
7	6.00	.051	7	7.63	.051
	8.86	.008		10.37	.009
8	6.25	.047	8	7.65	.049
	9.00	.010		10.35	.010
9	6.22	.048			
	8.67	.010			
10	6.20	.046			
	8.60	.012			
11	6.54	.043			
	8.91	.011			
12	6.17	.050			
	8.67	.011			
13	6.00	.050			
	8.67	.012			
14	6.14	.049			
	9.00	.010			
15	6.40	.047			
	8.93	.010			

Data from Owen DB. *Handbook of Statistical Tables*. US Department of Energy. Reading, MA: Addison-Wesley; 1962.

EXPERIMENTS IN WHICH EACH SUBJECT RECEIVES MORE THAN ONE TREATMENT: THE FRIEDMAN TEST

- See page 224-225, example «Anti-asthmatic Drugs and Endotoxin»

Multiple Comparisons after the Friedman Test (Important)

Just as we could use the Mann-Whitney test with a Holm- Sidak (or Bonferroni or Holm) correction for multiple comparisons following a Kruskal-Wallis analysis of variance on ranks, we can use the Wilcoxon signed rank tests with a Holm-Sidak (or Bonferroni or Holm) correction for multiple comparisons following a significant Friedman repeated measures analysis of variance on ranks.

$$\chi_r^2 = \frac{S}{nk(k+1)/12} = \frac{12\sum[R_t - n(k+1)/2]^2}{nk(k+1)}$$

$$= \frac{12}{nk(k+1)}\sum R_t^2 - 3n(k+1)$$

with $\nu = k - 1$ degrees of freedom

EXPERIMENTS IN WHICH EACH SUBJECT RECEIVES MORE THAN ONE TREATMENT: THE FRIEDMAN TEST

- See page 224-225, example «Anti-asthmatic Drugs and Endotoxin»

TABLE 10-15. Forced Expiratory Volume at 1 Second before and after Bronchial Challenge with Endotoxin and Salbutamol

	FEV_1 (L)					
	No Drug (Baseline)		One Hour after Endotoxin		Two Hours after Endotoxin and Salbutamol	
Person (Subject)	Units	Rank	Units	Rank	Units	Rank
1	3.7	2	3.4	1	4.0	3
2	4.0	2	3.7	1	4.4	3
3	3.0	2	2.8	1	3.2	3
4	3.2	2	2.9	1	3.4	3
Rank sums for each group		8		4		12

EXPERIMENTS IN WHICH EACH SUBJECT RECEIVES MORE THAN ONE TREATMENT: THE FRIEDMAN TEST

- See page 224-225, example «Anti-asthmatic Drugs and Endotoxin»

Table 10-15 shows how the three treatments rank in terms of FEV_1 for each of the four people in the study. The last row gives the sums of the ranks for each treatment. Since the possible ranks are 1, 2, and 3, the average rank is $(1 + 2 + 3)/3 = 2$. Since there are four people, if the treatments had no effect, these rank sums should all be about $4(2) = 8$. Hence, our measure of the difference between this expectation and the observed data is

$$S = (8-8)^2 + (4-8)^2 + (12-8)^2$$
$$= (0)^2 + (4)^2 + (4)^2 = 32$$

We convert S into χ_r^2 by dividing by $nk(k+1)/12 = 4(3)(3+1)/12 = 4$ to obtain $\chi_r^2 = 32/4 = 8.0$. Table 10-14 shows that for an experiment with $k = 3$ treatments and $n = 4$ experimental subjects there is only a $P = .042$ chance of obtaining a value of χ_r^2 as big or bigger than 8 by chance if the treatments have no effect. Therefore, we can report that endotoxin and salbutamol alter FEV_1 ($P = .042$).

Multiple Comparisons after the Friedman Test (Important)

We can use the Wilcoxon signed rank tests with a Holm-Sidak (or Bonferroni or Holm) correction for multiple comparisons following a significant Friedman repeated measures analysis of variance on ranks.