Nonparametric Methods

Elias Zintzaras, M.Sc., Ph.D.

Professor in Biomathematics-Biometry
Department of Biomathematics
School of Medicine
University of Thessaly

Institute for Clinical Research and Health Policy Studies
Tufts University School of Medicine
Boston, MA, USA

Theodoros Mprotsis, MSc, PhD
Teacher & Research Fellow
(http://biomath.med.uth.gr)
University of Thessaly
Email: tmprotsis@uth.gr

What does parametric mean?

- Most statistical methods we are familiar with (t-test, ANOVA, linear regression, confidence intervals, etc.) make assumptions about the probability distribution of the population under analysis (normal distribution)
- From this assumption we can develop sampling distributions
- Then from the sampling distributions we can derive sample statistics (mean, standard deviation, etc.)

What does nonparametric mean?

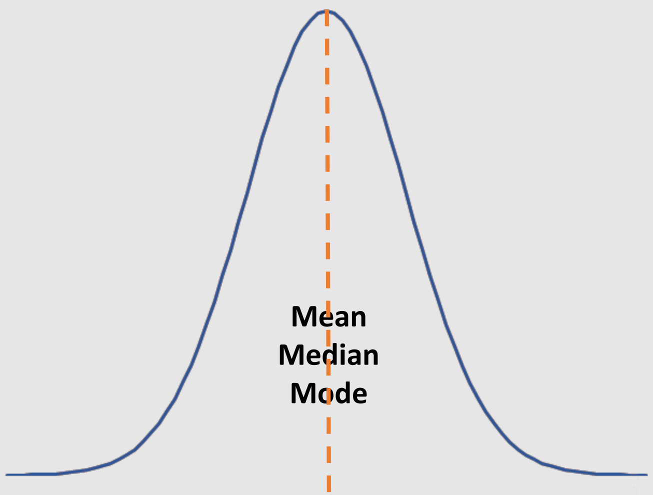
- Nonparametric methods however do not make these assumptions
- While parametric methods mostly require quantitative data, nonparametric methods allow us to work with qualitative data (nominal / ordinal)
- Most of the time, even quantitative data is converted to nominal / ordinal data for use with non-parametric methods; the most common type being ranked observations

When to use?

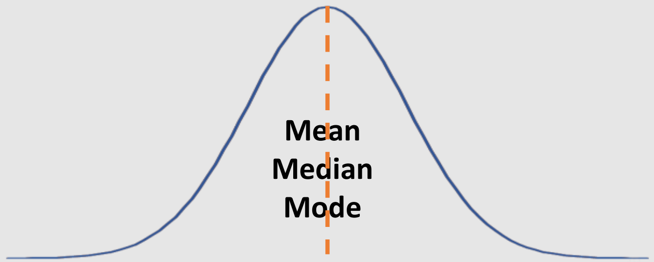
- When the sample size is small,
- the data is not normally distributed,
- variances are not similar,
- the data is ordinal, or
- there are many outliers that cannot be corrected by transformation

Parametric methods: Normality

Symmetric

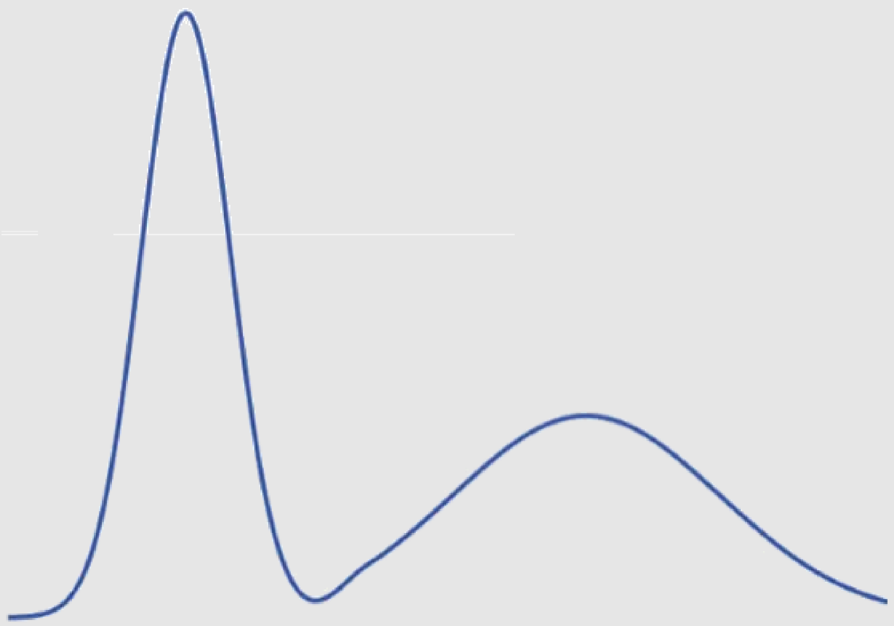


Symmetric

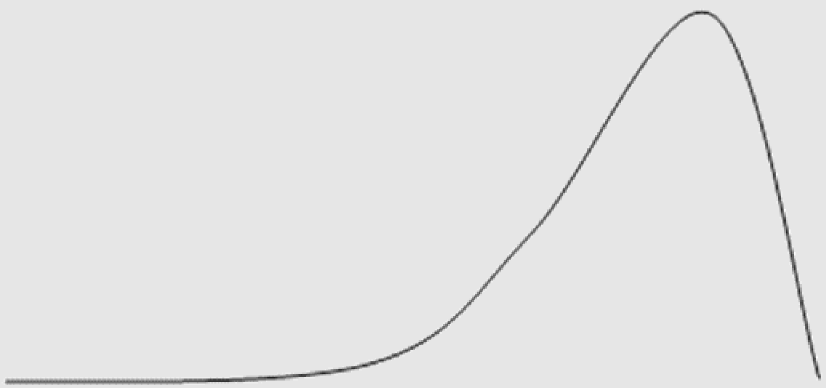


Parametric tests assume the normality of the data

Nonparametric methods: Normality



Bimodal distribution



Skewed Distribution

Nonparametric tests does assume normality of data

Advantages

- They can be applied to **quantitative variables with non-normal distributions**, regardless of the sample size
- They are suitable for **quantitative variables** when the sample size is small, and the distribution is unknown or appears to deviate from normality
- They can be used for **ordinal data**
- They involve simpler numerical calculations

Disadvantages

- When **parametric tests** are applicable, they tend to have **more statistical power** than nonparametric tests
- Nonparametric methods are generally not suited for more **complex statistical analyses**, such as multi-factor analysis of variance with interactions
- Calculating confidence intervals with nonparametric tests can be challenging

Commonly used nonparametric methods

- The **Wilcoxon Signed-Rank Test** is used for paired observations and corresponds to the parametric **paired samples t-test**
- The **Mann-Whitney U Test** (also known as the Mann-Whitney-Wilcoxon test) is used for comparing two independent samples and corresponds to the parametric **independent samples t-test**
- The **Kruskal-Wallis Test** is used for comparing more than two independent samples and corresponds to the parametric **One-Way ANOVA**
- The **Spearman's rank correlation** coefficient measures the degree of correlation between quantitative traits that are not normally distributed or between ordinal traits, and it corresponds to **Pearson's parametric correlation coefficient**

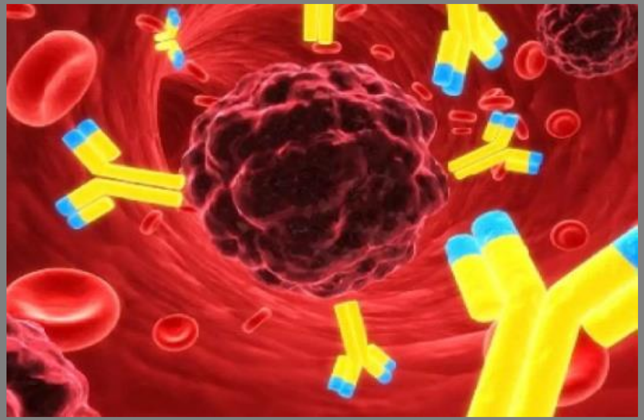
Matching Parametric and Nonparametric Statistical Tests

Parametric methods	Nonparametric methods
Paired samples t-test	**Wilcoxon Signed-Rank test**
Independent samples t-test	**Mann-Whitney U test**
One way ANOVA	**Kruskal-Wallis**
Pearson correlation coefficient	**Spearman's correlation coefficient**

Comparison of Serum M Immunoglobulin Levels in Adults Before and After Cholera Vaccination

Wilcoxon Signed-Rank Test is used for paired observations

(Paired sample t-test)



Background

- Comparing two matched samples using a parametric test would lead us to use a paired samples t-test; before/after
- In this case we would test if the mean difference between the mached pairs is zero, $\bar{d} = 0$
- The assumption is that each population is normally distributed
- Wilcoxon Signed-Rank test allows us to compare two populations that violate the assumption of normality
- The only requirement is that the distribution of differences is symmetrical (median), not necessarily normal distributed

Step by step calculations

a) The differences (A – B) are calculated for each pair of observations.
   Differences may be positive, negative, or equal to zero. Pairs of observations where the difference is zero are **excluded** from further analysis.,

b) The absolute values of the differences A – B are recorded,

c) These absolute values are then ranked by size, with the smallest absolute value assigned **rank 1**, the next smallest assigned **rank 2**, and so on. If two or more absolute differences have the same magnitude, they are assigned the **average of the ranks** they would hold if there were no ties.

d) The original signs (+ or -) of each difference (A - B) are then reapplied to the ranks corresponding to each difference

e) Finally, the ranks with a positive sign (summed as T_+) and the ranks with a negative sign (το απόλυτο summed as T_-) are added separately

Step by step calculations

f) When the absolute values of T_+ and T_- are very close, there is **no** statistically significant difference between groups A and B. The greater the difference between the absolute values of T_+ and T_-, the more likely it is that there is a real difference between groups A and B. Specifically, if $T_+ > T_-$ it suggests that A > B, while if $T_+ < T_-$, it suggests that A < B,

g) Statistical significance is assessed using the table on the right. If the value of T_+ (or T_-) is less than or equal to the lower threshold in the table, or greater than or equal to the upper threshold, the difference between the two groups is statistically significant at the corresponding significance level

Number of pairs without ties	Statistically significant at level 0.05	
	Lower limit	Upper limit
5	–	–
6	0	21
7	2	26
8	3	33
9	5	40
10	8	47
11	10	56
…		

Sampling distribution

$$Mean:\ \mu_{T^{+}} = \frac{n(n+1)}{4}$$

$$Standart\ deviation:\ \sigma_{T^{+}} = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

n = number of pairs excluding ties

Normal approximiation for $n \geq 10$

Test statistic and probability

What is μ_{T^+}?

Is T^+ in the upper tail or lower tail of the sampling distribution relative to μ_{T^+}?

If T^+ test statistic is in upper tail:

$$P\left(Z \geq \frac{\mathrm{T}^+ - \mu_{\mathrm{T}^+}}{\sigma_{\mathrm{T}^+}}\right)$$

If T^+ is in the lower tail:

$$P\left(Z \leq \frac{\mathrm{T}^+ - \mu_{\mathrm{T}^+}}{\sigma_{\mathrm{T}^+}}\right)$$

Apply continuity correction factor to T^+

Then double the probability in the tail to make it two-tailed

Comparison of Serum M Immunoglobulin Levels in Adults Before and After Cholera Vaccination

Serum M immunoglobulin levels were measured in 9 adults before and 15 days after cholera vaccination. The results are presented in columns 2 and 3 of the following table

The objective is to evaluate the difference in immunoglobulin levels measured before and after vaccination

ID	Serum M immunoglobulin levels		Difference (2) – (3)	Absolute values of differences	Ascending Order (Ranking) of Absolute Differences	Ranks of (6) with signs of (4)
	Vaccination					
	Before	After				
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	158	168	−10	10	3 – 3rd order	-3
2	140	168	−28	28	8 – 8th order	-8
3	165	161	+4	4	1 – 1st order, 4 is the lowest value	+1
4	100	120	−20	20	7 – 7th order	−7
5	110	122	−12	12	4.5 - (4th+5th)/2=4.5	−4.5
6	142	134	+ 8	8	2 – 2nd order, 8 is the next greatest value	+2
7	104	116	−12	12	4.5 - (4th order+5th order)/2=4.5	−4.5
8	254	270	−16	16	6 – 6th order	−6
9	180	180	0	(excluded from further analysis)		

T_+ vs T_-

ID	Serum M immunoglobulin levels		Ranks
	Vaccination		
	Before	After	
1	158	168	-3
2	140	168	-8
3	165	161	+1
4	100	120	−7
5	110	122	−4.5
6	142	134	+2
7	104	116	−4.5
8	254	270	−6
9	180	180	0

The positive and negative ranks are summarized, resulting in:

$$T_+ = 1 + 2 = 3$$

$$T_- = (-3) + (-8) + (-7) + (-4.5) + (-4.5) + (-6) = -33$$

Absolute value of T_-

$$|T_-| = |-33| = 33$$

Statistical significance

Number of pairs without ties	Statistically significant at level 0.05	
	Lower limit	Upper limit
5	–	–
6	0	21
7	2	26
8	**3**	**33**
9	5	40
10	8	47
11	10	56
…		

$T_+ = 3 \leq 3$. The value 3 is the lower limit for a sample size of 8 (ties not counted).

Therefore, the difference is statistically significant a the 5% significant level.

Thus, it can be claimed that 15 days after vaccination, there is an increase in serum M immunoglobulins.

Note: If T_+ is less than or equal to the lower limit of the table, then T_- is greater than or equal from the upper limit, and vice versa.

Sampling distribution results

$$Mean: \ \mu_{\mathrm{T}^{+}} = \frac{8(8+1)}{4} = \frac{72}{4} = 18$$

Remember that we had a difference of zero; therefore the number of pairs reduced from 9 to 8.

$$Standart\ deviation: \ \sigma_{\mathrm{T}^{+}} = \sqrt{\frac{8(8+1)(2 \cdot 8+1)}{24}} = 7.14$$

n = number of pairs excluding ties

Normal approximiation for $n \geq 10$

Test statistic, hypothesis result

$$Z = \frac{\mathrm{T}^+ - \mu_{\mathrm{T}^+}}{\sigma_{\mathrm{T}^+}} \qquad Z = \frac{3 - 18}{7.14} \qquad Z = -2.101$$

$$p - value = \mathrm{NORM.S.DIST}(-2.101, \mathrm{TRUE})$$

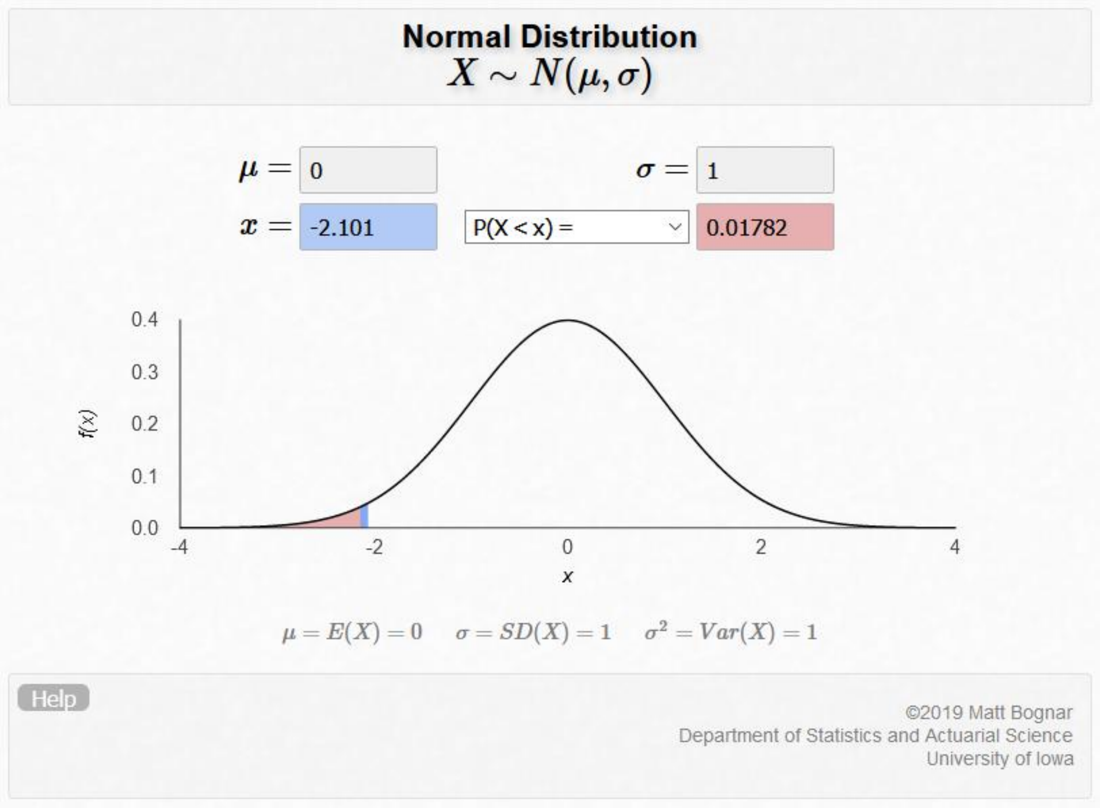
$$p - value\ (\text{lower} - \text{tailed}) = 0.01784595$$

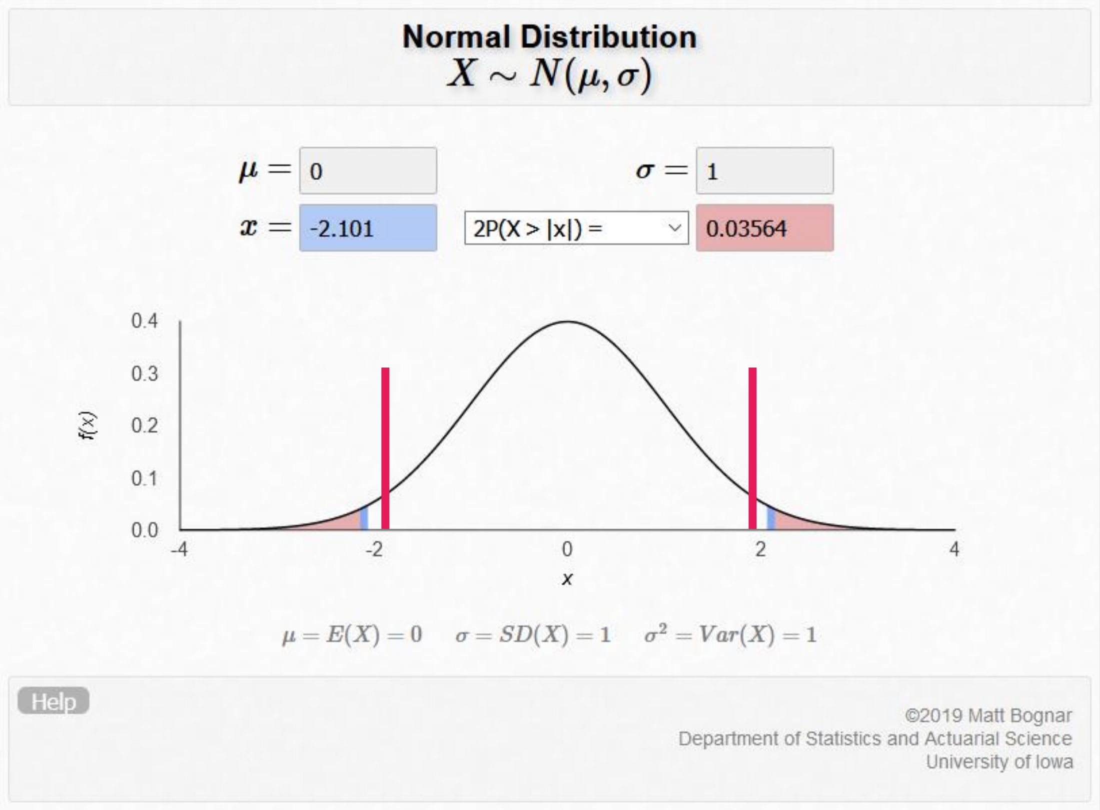
Then double the probability in the tail to make it two-tailed

$$p - value\ (\text{two} - \text{tailed}) = 0.0356919 \qquad \alpha = 0.05$$

FAIL TO REJECT THE H_0: The results provide sufficient evidence to conclude that there is a statistically significant increase in serum immunoglobulin M 15 days after vaccination

Normal distribution visuals





$$p - value \text{ (lower} - \text{tailed)} = 0.01784595$$

$$p - value \text{ (two} - \text{tailed)} = 0.0356919$$

$$z = \pm 1.96, 95\% \text{ interval}$$

https://homepage.divms.uiowa.edu/~mbognar/applets/normal.html

Energy consumption between thin and obese women
Mann-Whitney U test

(Independent sample t-test)

Comparing Two Samples: Parametric and Nonparametric Approaches

- To compare two samples using a parametric test, we would use the t-test for independent samples
- In this case, we would check if the mean values of the two samples are equal, represented as: $\bar{x}_1 - \bar{x}_2 = 0$
- The assumption we make is that each population is normally distributed
- The Wilcoxon Rank Sum / Mann-Whitney test, however, allows us to compare populations without assuming normal distribution
- This is why we use **ranks** of observations instead of actual values

Assumptions

The only assumptions for carrying out a Mann-Whitney test are

- that the two groups must be independent and
- that the dependent variable is ordinal or numerical (continuous)

Note: However, in order to report the difference between groups as medians, the shape of the distributions of the dependent variable by group must be similar (skewness)

Energy consumption between thin and obese women

The energy consumption over 24 hours for a group of thin women and a group of obese women is shown in the table.

Is there a difference in energy consumption between the two groups?

$$median_{thin} = 7.9$$

$$median_{obese} = 9.69$$

Subject ID	Thin	Obese
1	6.13	8.79
2	7.05	9.19
3	7.48	9.21
4	7.48	9.68
5	7.53	9.69
6	7.58	9.97
7	7.90	11.51
8	8.08	11.85
9	8.09	12.79
10	8.11	
11	8.40	
12	10.15	
13	10.88	

Step 1: Calculate the Rank Sums

Combine both groups and assign ranks from smallest to largest. If there are ties, assign each tied value the average rank for those values.

Thin		Obese	
Rank	Energy	Energy	Rank
1	6.13		
2	7.05		
3.5	7.48		
3.5	7.48		
5	7.53		
6	7.58		
7	7.90		
8	8.08		
9	8.09		
10	8.11		
11	8.40		

Thin		Obese	
Rank	Energy	Energy	Rank
		8.79	12
		9.19	13
		9.21	14
		9.68	15
		9.69	16
		9.97	17
18	10.15		
19	10.88		
		11.51	20
		11.85	21
		12.79	22
$R = 103$			$R = 150$

Subject ID	Thin	Obese
1	6.13	8.79
2	7.05	9.19
3	7.48	9.21
4	7.48	9.68
5	7.53	9.69
6	7.58	9.97
7	7.90	11.51
8	8.08	11.85
9	8.09	12.79
10	8.11	
11	8.40	
12	10.15	
13	10.88	

Step 2: Calculate the U values

$$U_{thin} = R_{thin} - \frac{n_{thin}(n_{thin} + 1)}{2}$$

$$U_{obese} = R_{obese} - \frac{n_{obese}(n_{obese} + 1)}{2}$$

$n_{thin} = \text{sample size of the thin group}$
$n_{obese} = sample\ size\ of\ the\ obese\ group$

Step 2: Calculate the U values

$$U_{thin} = 103 - \frac{13(13+1)}{2} = 103 - 91 = 12$$

$$U_{obese} = 150 - \frac{9(9+1)}{2} = 150 - 45 = 105$$

n_{thin} = sample size of the thin group
$n_{obese} = sample\ size\ of\ the\ obese\ group$

Step 3: Determine the Smaller U Value

$$U = min(12, 105) = 12$$

The smaller U value is used for interpretation, so $U = 12$

Step 4: Determine the Critical Value

$n_1 \Rightarrow$ $n_2 \Downarrow$	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3																			
4			0																
5		0	1	2															
6		1	2	3	5														
7		1	3	5	6	8													
8	0	2	3	6	8	10	13												
9	0	2	4	7	9	12	15	17											
10	0	3	5	8	11	14	17	20	23										
11	0	3	6	9	13	16	19	23	26	30									
12	1	3	7	11	14	18	22	26	29	33	37								
13	1	4	8	12	16	20	24	28	33	37	41	45							
14	1	5	9	13	17	22	26	31	36	40	45	50	54						
15	1	5	10	14	19	24	29	34	38	44	49	54	59	64					
16	1	6	11	16	21	26	31	36	42	48	53	59	64	69	74				
17	2	6	12	17	22	28	34	39	45	51	57	63	69	75	80	86			
18	2	6	13	18	24	30	36	42	49	55	61	67	73	80	86	93	99		
19	2	7	13	19	26	31	39	45	52	58	65	72	78	85	92	99	106	112	
20	2	8	14	20	27	34	41	48	55	62	68	76	83	90	98	104	112	119	127

For a Mann-Whitney U test with $n_{thin} = 13, n_{obese} = 9$, we can check a U distribution table or use software to get the critical value

Here, the critical value is **28** for a two-tailed test

https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470776124.app1 (page 8 of 14)

Step 5: Compare U to the Critical Value

Since $U = 12$ is less than 28, we **reject the null hypothesis**. This suggests a statistically significant difference between the Thin and Obese groups

Wilcoxon W

- The **Wilcoxon W** is simply the smallest sum of ranks, in this case 103, but
- **SPSS** uses an approximation of the normal distribution to calculate the Z statistic and
  the resulting p-value (Asymptotic Sig).

Test Statistics[a]

	energy
Mann-Whitney U	12.000
Wilcoxon W	103.000
Z	-3.106
Asymp. Sig. (2-tailed)	.002
Exact Sig. [2*(1-tailed Sig.)]	.001[b]

a. Grouping Variable: group

b. Not corrected for ties.

Interpretation

A Mann-Whitney U test showed that there was a significant difference between the energy consumption for the thin group and the obese group ($U = 12$, $p < 0.05$). The median energy consumption was 7.9 for the thin group compared to 9.69 for the obese group, suggesting that the obese group consumes more energy.

Sampling distribution

$$Mean:\ \mu_U = \frac{n_{thin} n_{obese}}{2}$$

$$Standart\ deviation:\ \sigma_U = \sqrt{\frac{n_{thin} n_{obese}(n_{thin} + n_{obese} + 1)}{12}}$$

n_{thin} = sample size of the thin group
$n_{obese} = sample\ size\ of\ the\ obese\ group$

Normal approximiation for $n \geq 10$

Sampling distribution

$$Mean: \mu_U = \frac{n_{thin} n_{obese}}{2} = \frac{13 \cdot 9}{2} = 58.5$$

$$Standart\ deviation:\ \sigma_U = \sqrt{\frac{13 \cdot 9(13 + 9 + 1)}{12}} = 14.97$$

n_{thin} = sample size of the thin group
$n_{obese} = sample\ size\ of\ the\ obese\ group$

Normal approximiation for $n \geq 10$

Test statistic

$$z = \frac{U - \mu_U}{\sigma_U}$$

Test statistic and result

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{12 - 58.5}{14.97} = \frac{-46.5}{14.97} = -3.106$$

$$p - value = \mathrm{NORM.S.DIST}(-3.106, \mathrm{TRUE})$$

$$p - value\ (\mathrm{lower} - \mathrm{tailed}) = 0.0009475$$

Then double the probability in the tail to make it two-tailed

$$p - value\ (\mathrm{two} - \mathrm{tailed}) = 0.01895 \qquad \alpha = 0.05$$

Reject H_0: The results provide sufficient evidence to conclude that there is a significant difference in energy consumption between thin and obese women.

Headache Activity Reduction in Children: A Comparison of Treatment Methods
Kruskal-Wallis Test

(one-way ANOVA)

Background

- Comparing more than two independent samples using a **parametric test** would lead us to use the **ANOVA procedure**
- In this case we would test if the means of samples are equal: $\overline{\boldsymbol{x}}_{\boldsymbol{1}} = \overline{\boldsymbol{x}}_{\boldsymbol{2}} = \overline{\boldsymbol{x}}_{\boldsymbol{3}} = \cdots \overline{\boldsymbol{x}}_{\boldsymbol{n}}$
- However, the **assumption** for ANOVA is that each population is **normally distributed**
- The **Kruskal-Wallis test** allows us to compare more than two populations that **violate the assumption of the normality**
- We use the **sum of ranks for observations (locations)**

Sampling distribution

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

Iterate through each sample/group

$n_i =$ sample size in sample/group i
$n =$ total sample size
$R_i =$ the sum of the ranks for sample/group i

Test statistic and p-value

Test statistic: $H > \chi^2{}_{\alpha,k-1}$ chi-square

Is the sum of the ranks for each group the same?

p-value: $P(H < \chi^2)$

The Critical Values of the Chi-Square Distribution can be found at http://www.itl.nist.gov/div898/handbook/eda/section3/eda3674.htm

Headache Activity Reduction in Children: A Comparison of Treatment Methods

Eighteen children suffering from migraines underwent three different treatments. One group received **relaxation response and biofeedback** (treatment A), the second group **received relaxation response only** (treatment B), and the last group was **untreated** (treatment C). The response variable was **"reduction in headache activity"** (where a negative value indicates an increase in headache activity, and a value of 100 indicates the absence of headaches). The data are from Fentress (1986).

Treatment A	Treatment B	Treatment C
62	69	50
74	43	-120
86	100	100
74	94	-288
91	100	4
37	98	-76

https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1469-8749.1986.tb03847.x

Headache Activity Reduction in Children: A Comparison of Treatment Methods

The 18 tests are mixed, and the ranks are calculated from lowest to highest

Treatment	Response	Rank
B	100	17
B	100	17
C	100	17
B	98	15
B	94	14
A	91	13
A	86	12
A	74	10.5
A	74	10.5
B	69	9
A	62	8
C	50	7
B	43	6
A	37	5
C	4	4
C	-76	3
C	-120	2
C	-288	1

Sum of ranks

The 18 tests are mixed, and the ranks are calculated from lowest to highest

$$\sum(A) = 59$$

$$\sum(B) = 78$$

$$\sum(C) = 34$$

$$\sum(Ranks) = 171$$

Treatment	Response	Rank
B	100	17
B	100	17
C	100	17
B	98	15
B	94	14
A	91	13
A	86	12
A	74	10.5
A	74	10.5
B	69	9
A	62	8
C	50	7
B	43	6
A	37	5
C	4	4
C	-76	3
C	-120	2
C	-288	1

Test statistic and p-value

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

$$H = \frac{12}{18(18+1)} \left[6\left(\frac{59}{6} - \frac{(18+1)}{2} \right)^2 + 6\left(\frac{78}{6} - \frac{(18+1)}{2} \right)^2 + 6\left(\frac{34}{6} - \frac{(18+1)}{2} \right)^2 \right]$$

$$H = 0.0351(0.6666 + 73.5 + 162.3333)$$

$$H = 0.0351(162.3333) = 5.6959$$

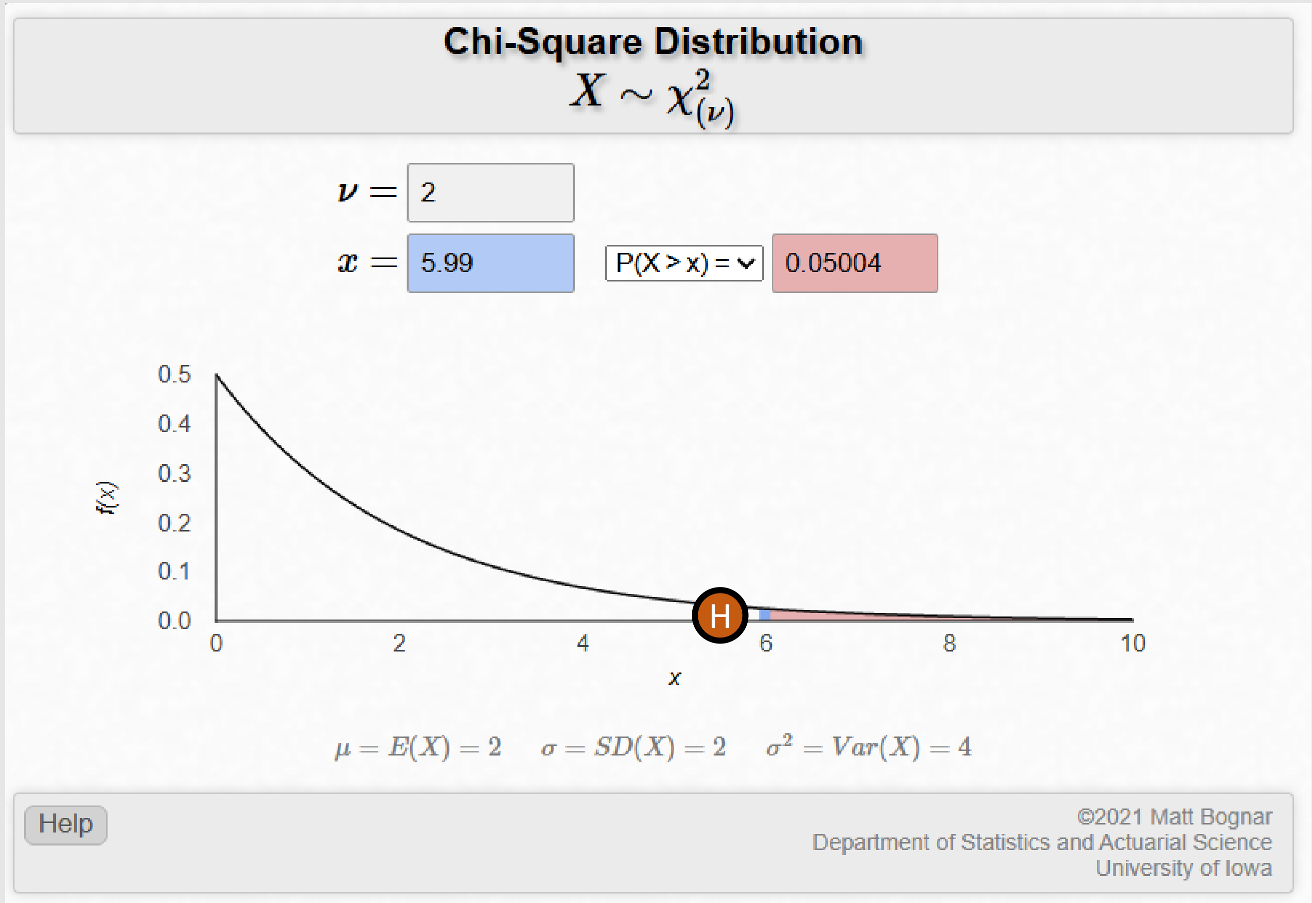
$n_i =$ sample size in sample/group i
$n =$ total sample size
$R_i =$ the sum of the ranks for sample/group i

$\sum(A) = 59$ $\sum(B) = 78$ $\sum(C) = 34$

Test statistic and p-value

Test statistic: $5.6959 < 5.99$

p-value: 0.06



https://homepage.divms.uiowa.edu/~mbognar/applets/chisq.html

http://www.itl.nist.gov/div898/handbook/eda/section3/eda3674.htm

Interpretation

At a significance level of 5%, it follows that the three methods do not differ statistically significantly, as the value of the statistic $H = 5.69$ is smaller than the critical value from the χ^2- distribution with 2 degrees of freedom, which is 5.99 $(P \geq 0.05)$. The p-value calculated using statistical software is 0.06.

If a statistically significant result were observed, we would perform multiple comparisons using the Mann-Whitney test with an adjustment, such as the Bonferroni correction.