

Analysis of Questionnaires and Qualitative Data – Non-parametric Tests



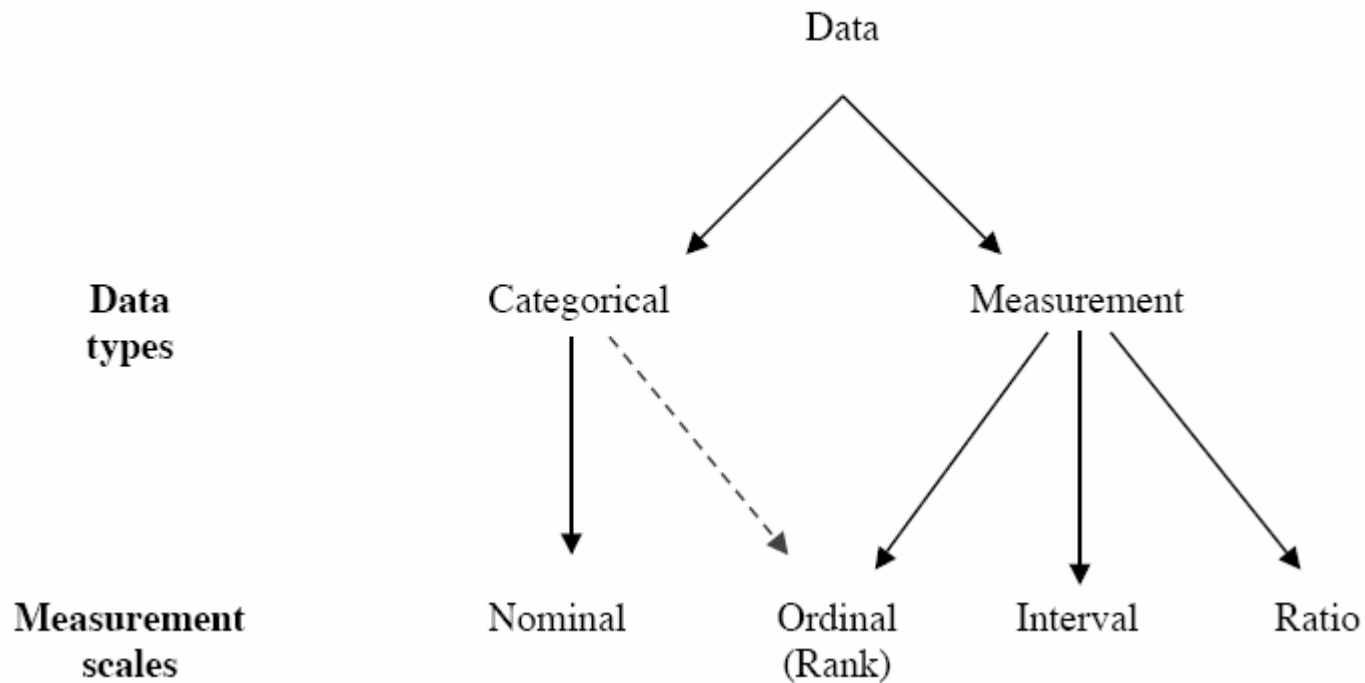
JERZY STEFANOWSKI

Instytut Informatyki
Politechnika Poznańska

Lecture SE 2013, Poznań

Recalling Basics

Measurement Scales and Variable Types



Measurment Scales

- Four scales of measurements commonly used in statistical analysis: nominal, ordinal, interval, and ratio scales
- A **nominal scale** -> there is no relative ordering of the categories,
e.g. sex of a person, colour, trademark,
- **Ordinal** -> place object in a relative ordering,
Many rating scales (e.g. never, rarely, sometimes, often, always)
- **Interval** -> Places objects in order and equal differences in value denote equal differences in what we are measuring
- **Ratio** -> similar interval measurement but also has a 'true zero point' and you can divide values.

Coding values with numbers

Numbers are used to code nominal or ordered values but they are not true numbers!

Only for interval or ratio measurements they are proper number – e.g. you are allowed to perform algebraic operations (+, -, *, /)

Most of our considerations as to statistical data analysis or prediction methods concern → numerical (interval, ratio) data.

- In many domains we collect nominal or ordinal data
- Use of Questionnaires or Survey Studies in SE!
- Also collected in Web applications

nextCOMPANIES

Please rate your level of agreement with the following statement about life-work balance at your job:

	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
My job gives me the flexibility to meet the needs of both my professional and personal life	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

What is your gender?

☐ Female

☐ Male

What is your age?

Types of Variables in Questionnaires

- Yes/No Questions
Any question that has yes or no as a possible response is nominal
- Two or multi-values (options)
 - Gender (Femal vs. Male)
- *Activity* nominal data type of 6 choices of activity in the park:
 - sport,
 - picnic,
 - reading,
 - walk (including with the dog),
 - meditation,
 - jog.

Likert Scales

- A special kind of survey question uses a set of responses that are ordered so that one response is greater (or preferred) than another.
- Generally, this term is used for any question that has about 5 or more possible options.
- An example:
"How would you rate your department admin?"
1=very incompetent, 2=somewhat incompetent,
3=neither competent, 4=somewhat competent, or
5=very competent.

Simple Tools to Analyze Survey Results

Frequency or cross tables

Frequency Distribution of Gender		
Gender	Count	Percentage (%)
Female.....	6	60%
Male.....	4	40%
Total	10	100%

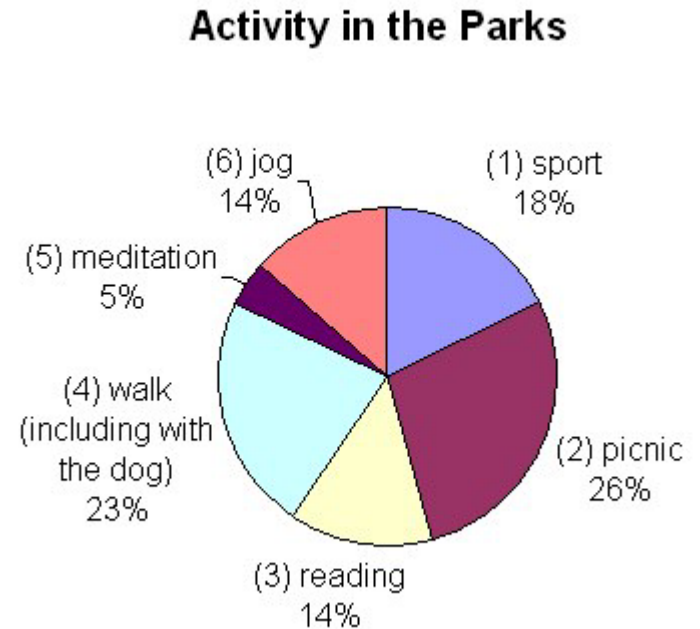
Frequency Distribution of Life-Work Balance at Work		
Life-Work Balance	Count	Percentage (%)
Strongly Agree.....	2	20%
Agree.....	4	40%
Neither Agree nor Disagree....	2	20%
Disagree.....	2	20%
Strongly Disagree.....	0	0
Total	10	100%

Life-Work Balance at Work by Gender			
Life-Work Balance	Female	Male	Total
Strongly Agree.....	0%	50.0% (2)	20.0% (2)
Agree.....	33.3% (2)	50.0% (2)	40.0% (4)
Neither Agree nor Disagree.....	33.3% (2)	0%	20.0% (2)
Disagree.....	33.3% (2)	0%	20.0% (2)
Strongly Disagree.....	0%	0%	0%
Total	100% (6)	100% (4)	100% (10)

Other forms of summaries for question responses

- Think about graphic visualizations

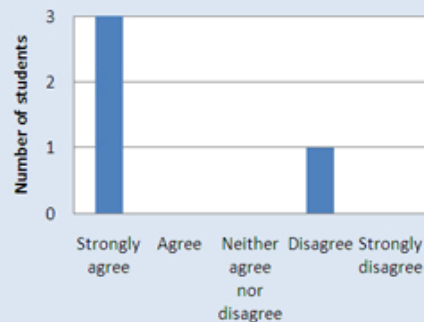
Activity	frequency	Relative frequency
(1) sport	4	18.2%
(2) picnic	6	27.3%
(3) reading	3	13.6%
(4) walk (including with the dog)	5	22.7%
(5) meditation	1	4.5%
(6) jog	3	13.6%
Sum	22	100.0%



Example table

Example chart

Question 1: I like maths lessons	
Response	Number of respondents
Strongly agree	3
Agree	0
Neither agree nor disagree	0
Disagree	1
Strongly disagree	0



Based on four responses

From basic descriptive statistics to statistical tests

- Dependencies of nominal variables
- We want to know whether variable *Playground* has relationship with variable *Satisfaction* , or not
-> use chi2 test

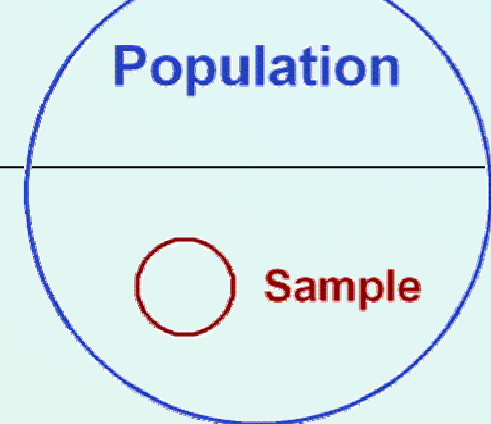
Count	Activity Time in Parks (min)						Grand Total
Activity	15	30	45	60	120	180	
(1) sport	0	1	0	1	1	1	4
(2) picnic	0	1	0	3	1	1	6
(3) reading	0	1	1	0	0	1	3
(4) walk (including with the dog)	0	2	1	0	1	1	5
(5) meditation	0	0	1	0	0	0	1
(6) jog	1	2	0	0	0	0	3
Grand Total	1	7	3	4	3	4	22

Independent Values	Activity Time in Parks (min)						Grand Total
Activity	15	30	45	60	120	180	
(1) sport	0.182	1.273	0.545	0.727	0.545	0.727	4
(2) picnic	0.273	1.909	0.818	1.091	0.818	1.091	6
(3) reading	0.136	0.955	0.409	0.545	0.409	0.545	3
(4) walk (including with the dog)	0.227	1.591	0.682	0.909	0.682	0.909	5
(5) meditation	0.045	0.318	0.136	0.182	0.136	0.182	1
(6) jog	0.136	0.955	0.409	0.545	0.409	0.545	3
Grand Total	1	7	3	4	3	4	22

Chi Square computation	Activity Time in Parks (min)						
Activity	15	30	45	60	120	180	
(1) sport	0.182	0.058	0.545	0.102	0.379	0.102	
(2) picnic	0.273	0.433	0.818	3.341	0.040	0.008	
(3) reading	0.136	0.002	0.854	0.545	0.409	0.379	
(4) walk (including with the dog)	0.227	0.105	0.148	0.909	0.148	0.009	
(5) meditation	0.045	0.318	5.470	0.182	0.136	0.182	
(6) jog	5.470	1.145	0.409	0.545	0.409	0.545	

chi square = 25.012
df = 25

Recall Hypothesis Testing



- State the a priori hypotheses
 - The null hypothesis H_0
 - The alternative hypothesis H_1 (one vs. two tailed tests)
- Make decisions as to significance level α
- Decide on the suitable statistics and calculate its values from the random sample
- Compare it against the value expected according to H_0 (temporary assumed to be true)
- Decide basing on a critical region

Testing mean values (parametric test)

- Assume: X normally distributed $N(\mu_0, \sigma_{\bar{X}})$ cardinality n of the sample, known variance/standard distributions

- Krok 1:** Hypotheses:

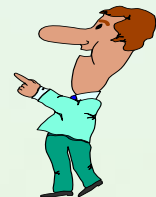
$$H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0 \quad \text{or} \quad H_1 : \mu < \mu_0 \quad H_1 : \mu > \mu_0$$

- Krok 2:** Significance level α .

- Krok 3:** Test statistics (random variable)

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

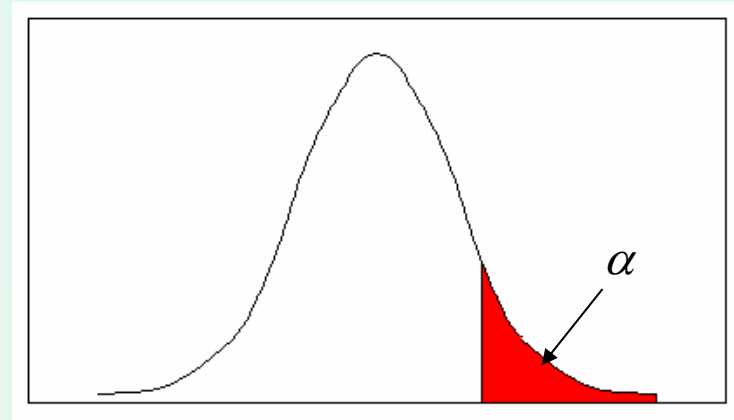
- Krok 4:** Critical values $z(\alpha)$ and rejection regions.
- Krok 5:** Make a final decision
- Z-normal distribution vs. t-Student test?



Critical regions and values (t-student version)

One tailed test

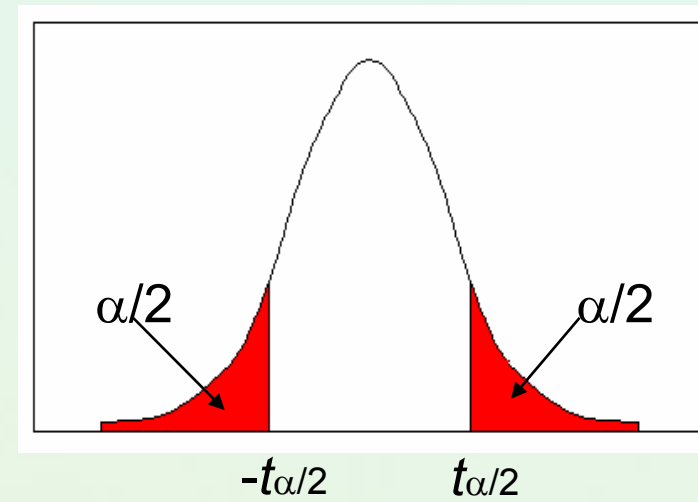
- $H_1 : \mu < \mu_0$
- H_0 reject, is $p \leq \alpha$ ($T \geq t_\alpha$)



t_α

Two tailed test

- $H_1 : \mu \neq \mu_0$



$-t_{\alpha/2}$

$t_{\alpha/2}$

Comparing Means among Two (or More) Independent Populations

- Recall, general “recipe” for hypothesis testing . . .

$$H_0 : \mu_1 = \mu_2$$

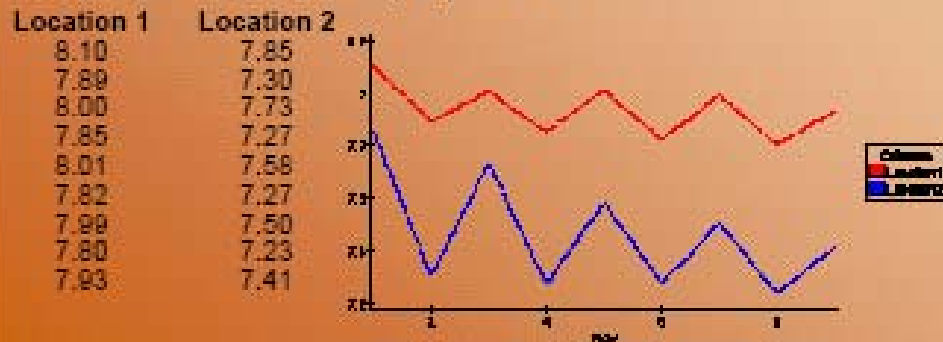
- 1. Start by assuming H_0 true
- 2. Measure values on the sample
- 3. Compare test statistic (distance) to appropriate distribution to get p-value

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{SE}(\bar{x}_1 - \bar{x}_2)} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Two Sample - Examples

Example: Nine observations of surface soil pH were made two different locations. Does the data suggest that the true mean soil pH values differ for the two locations?

Test using $\alpha = 0.05$, and be sure to check any necessary assumptions for the validity of your test.



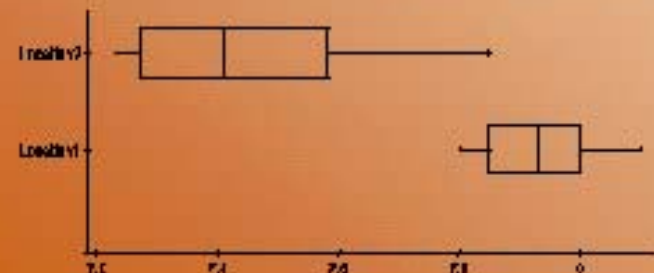
#1 Formulate hypotheses

$$H_0: \mu_1 - \mu_2 = 0$$

(there is no difference between the true mean soil pH of location1 and location2)

$$H_a: \mu_1 - \mu_2 \neq 0$$

(there is a difference between the true mean soil pH of location1 and location2)



#2 Calculate the test statistic

Descriptive Statistics: Location 1, Location 2

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Location 1	9	0	7.9322	0.0335	0.1005	7.8000	7.8350	7.9300	8.0050
Location 2	9	0	7.4600	0.0740	0.2220	7.2300	7.2700	7.4100	7.6550

$$SE_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{0.1005^2}{9} + \frac{0.222^2}{9}} = 0.081$$

$$t_s = \frac{\bar{y}_1 - \bar{y}_2 - 0}{SE_{\bar{y}_1 - \bar{y}_2}} = \frac{7.9322 - 7.460 - 0}{0.081} = 5.827$$

Parametric and non-parametric tests

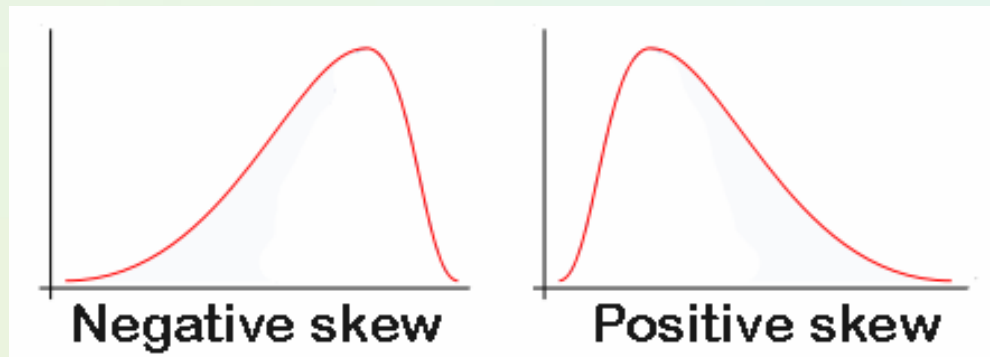
- Parametric statistical tests assume that the data belong to some type of probability distribution. The normal distribution is probably the most common.
 - Moreover homogenous variances and no outliers
- Non-parametric statistical tests are often called distribution free tests since don't make any assumptions about the distribution of data.
- ..., they are more appropriate for questionnaires or qualitative data!

Nonparametric Tests

Make no assumptions about the data's characteristics.

Use if *any* of the three properties below are true:

(a) the data are not normally distributed (e.g. skewed);



(b) the data show inhomogeneity of variance;

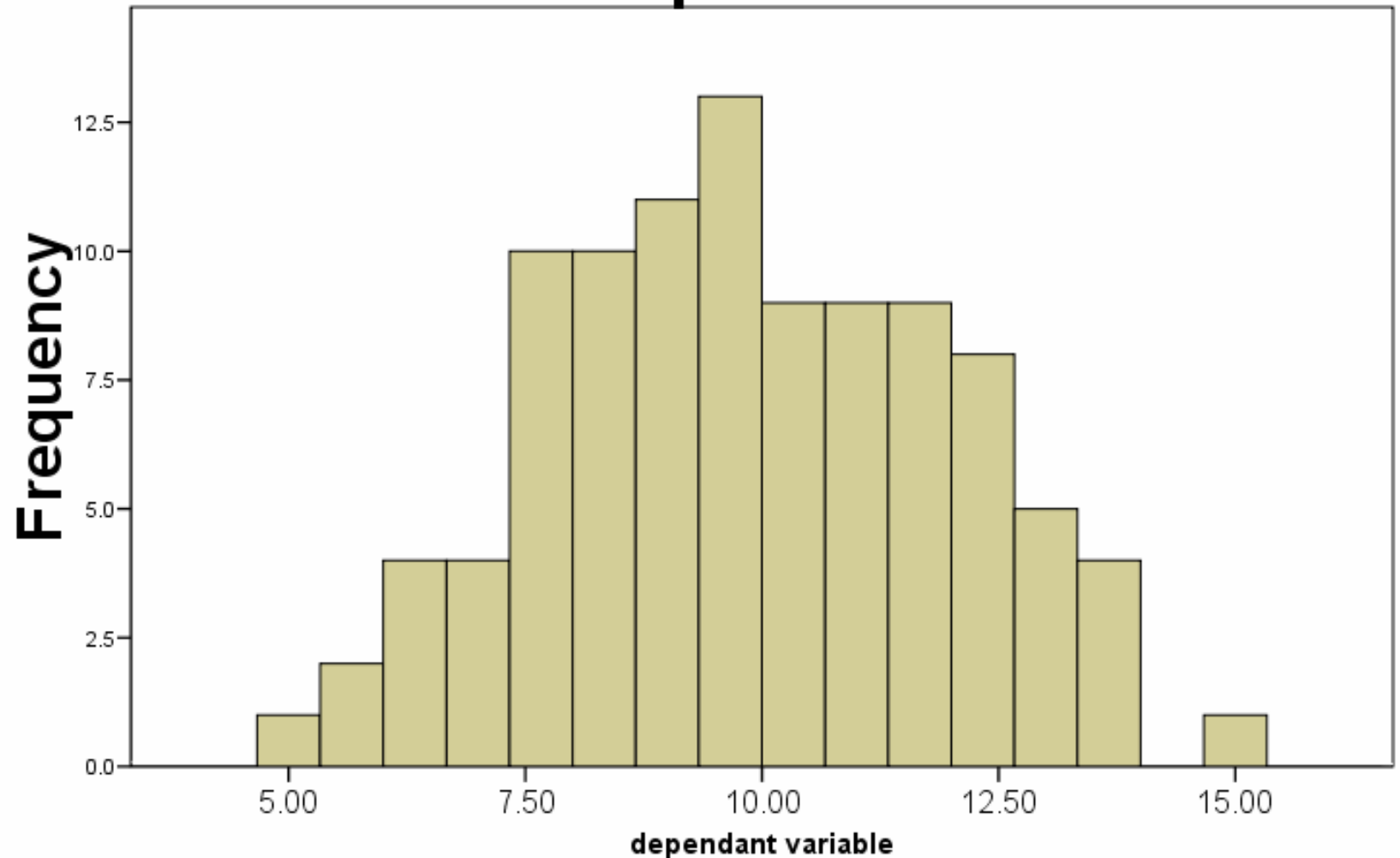
(c) the data are measured on an ordinal scale (ranks).

Assumption 1 - normality

- This can be checked by inspecting a histogram
 - with small samples the histogram is unlikely to ever be exactly bell shaped
- This assumption is only broken if there are large and obvious departures from normality

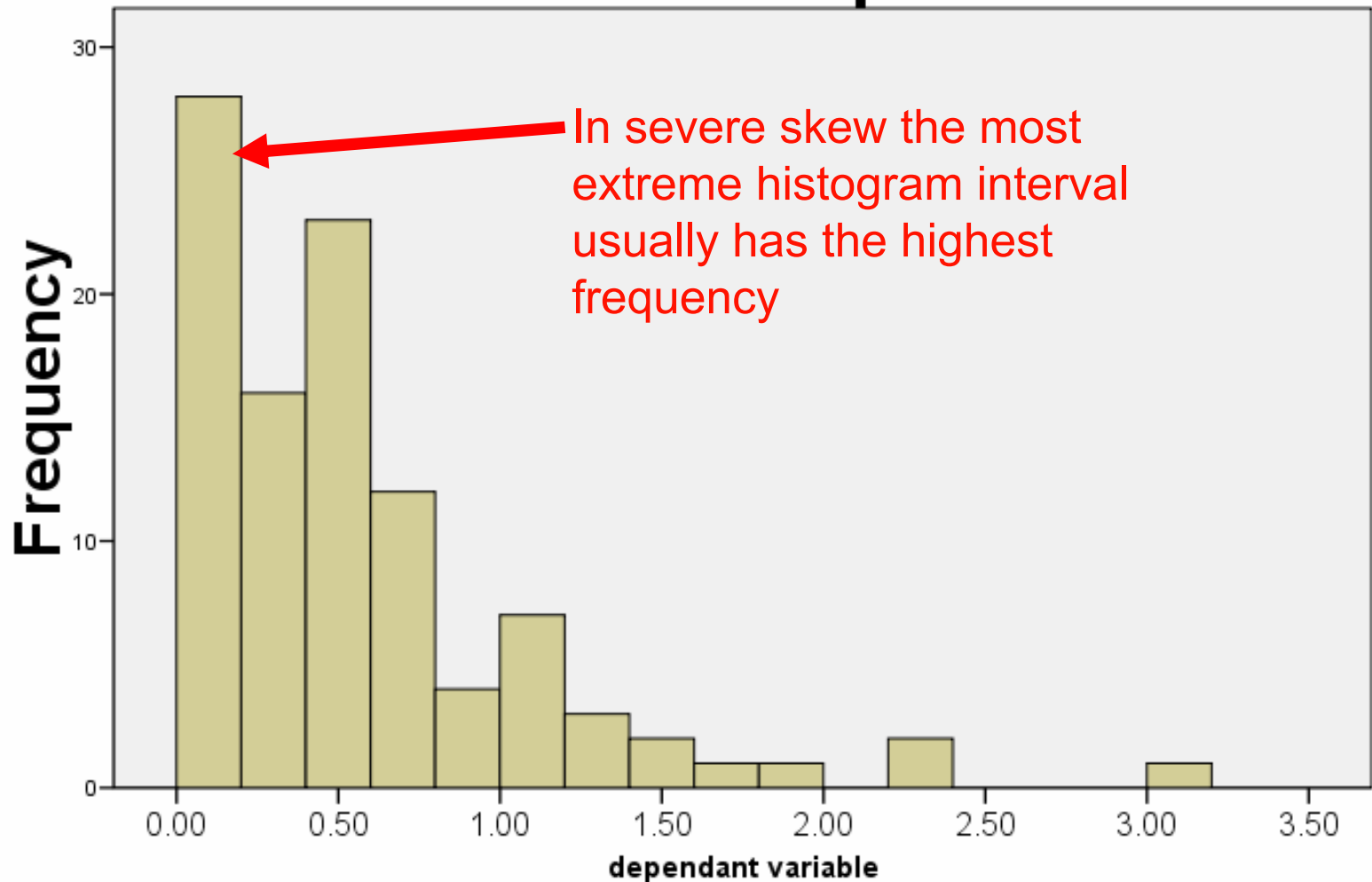
Assumption 1 - normality

normal: use parametric

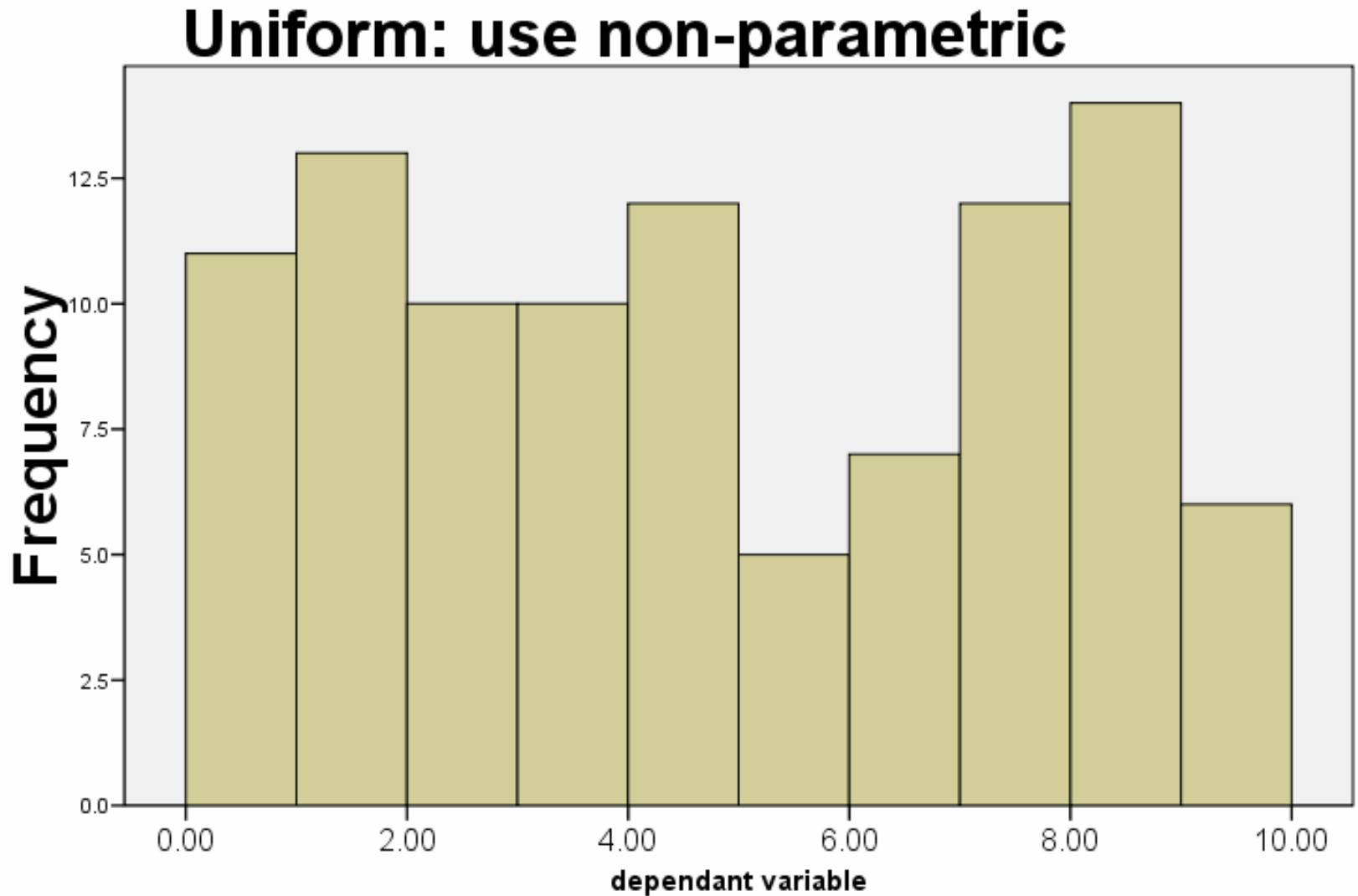


Assumption 1 - normality

Severe skew: use non-parametric

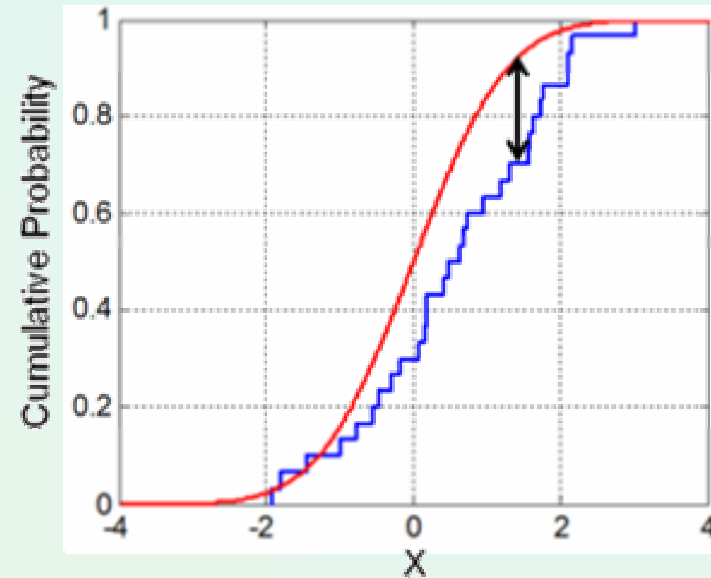


Assumption 1 - normality



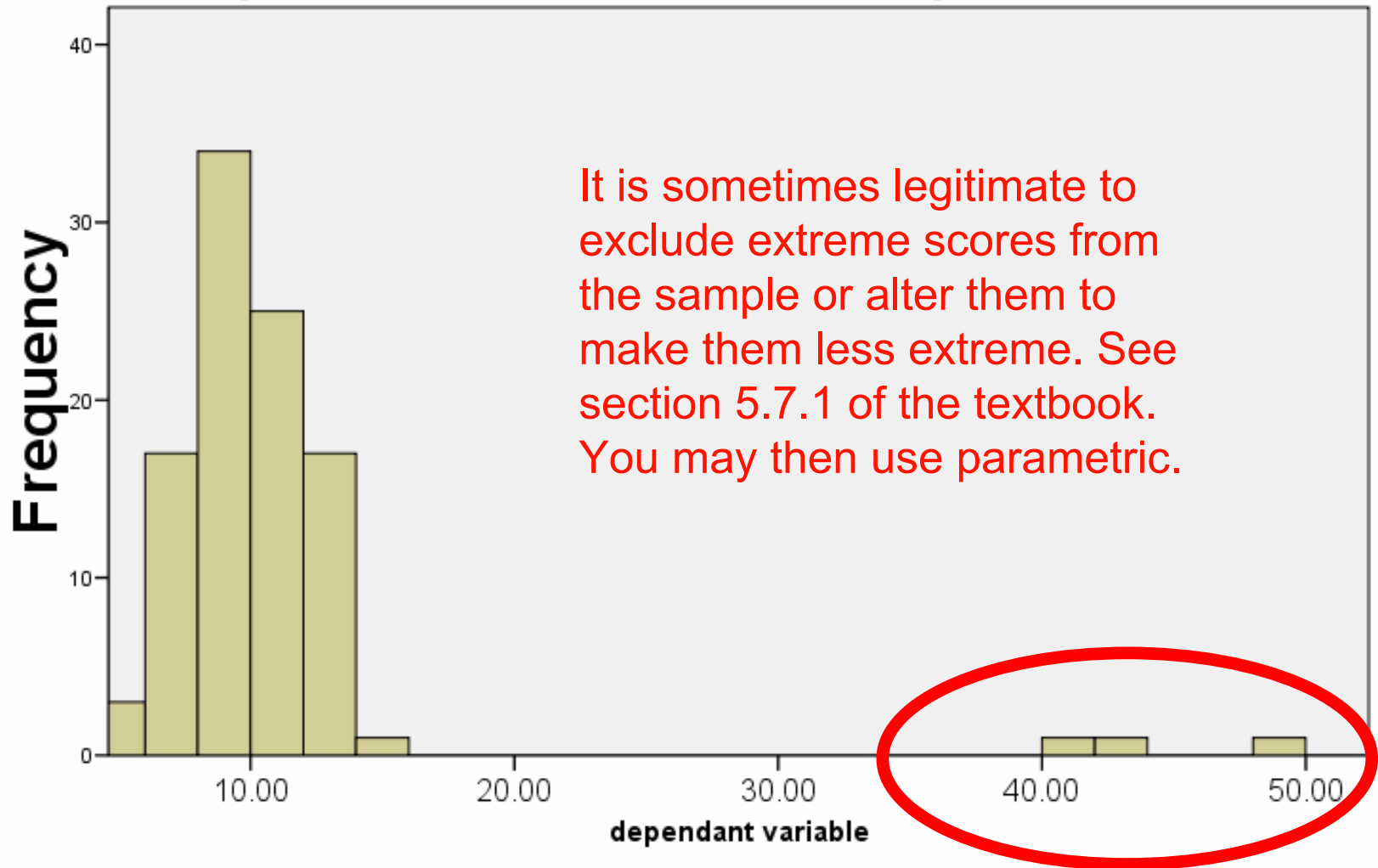
Testing normality more formally

- the **Kolmogorov–Smirnov** test (K–S test) is a nonparametric test for the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution
- In the special case of testing for normality of the distribution, samples are standardized and compared with a standard normal distribution
- More powerfull is the **Shapiro–Wilk test**



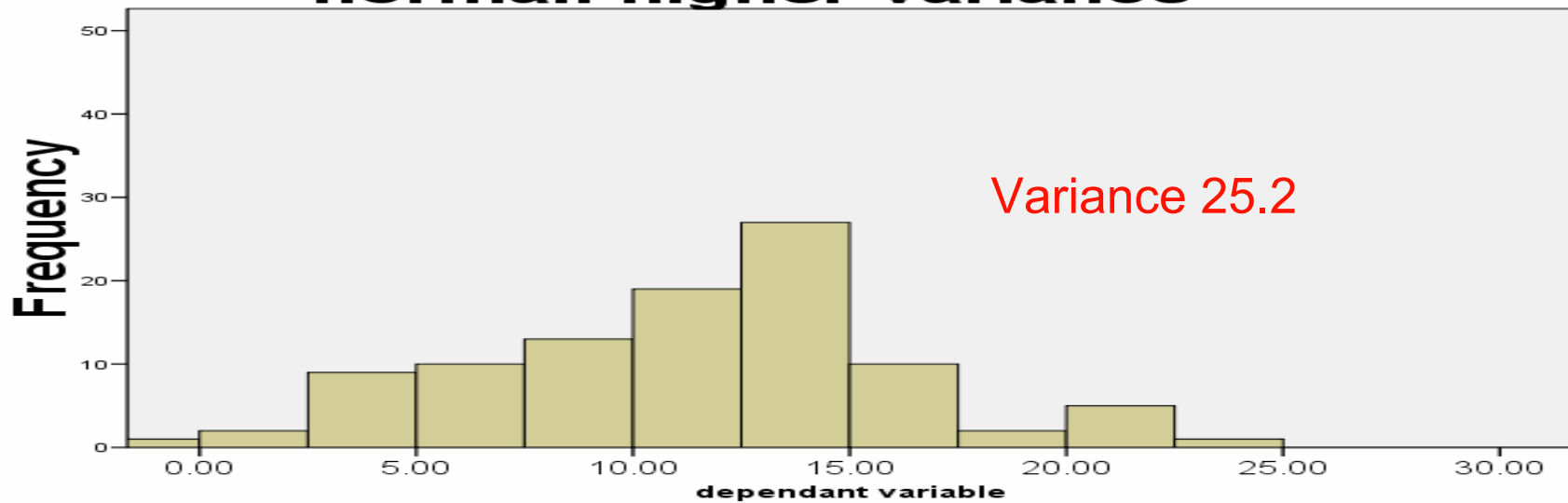
Assumption 3 – no extreme scores

normal plus outliers: use non-parametric

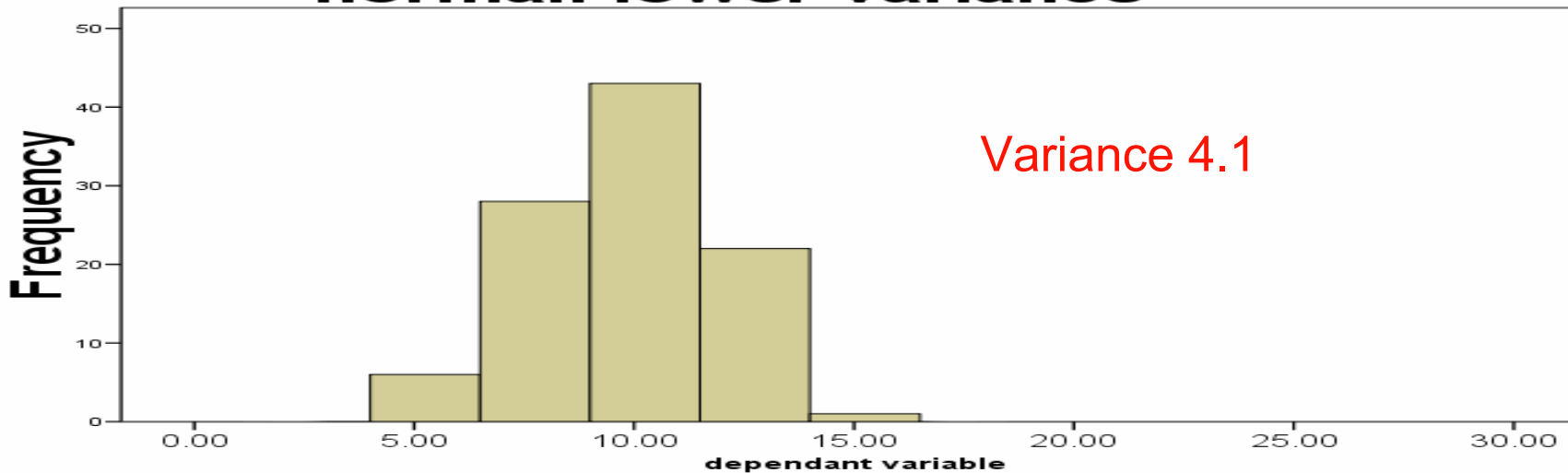


Assumption 4 (independent samples t only) – equal variance

normal: higher variance



normal: lower variance



Examples of parametric tests and their non-parametric equivalents:

Parametric test:

Pearson correlation

(No equivalent test)

Independent-means t-test

Dependent-means t-test

One-way Independent Measures
Analysis of Variance (ANOVA)

One-way Repeated-Measures
ANOVA

Non-parametric counterpart:

Spearman's correlation

Chi-Square test

U-Mann-Whitney test

Wilcoxon test

Kruskal-Wallis test

Friedman's test

WHICH TEST SHOULD I USE?

The type of data that you collect will be important in your final choice of test:

Nominal

Consider a chi-squared test if you are interested in differences in frequency counts using nominal data, for example comparing whether month of birth affects the sport that someone participates in.

Ordinal

If you are interested in the relationship between groups, then use Spearman's correlation.

If you are looking for differences between independent groups, then a Mann-Whitney test may be appropriate.

If the groups are paired, however, then a Wilcoxon Signed rank test is appropriate.

If there are three or more groups then consider a Kruskal-Wallis test.

Wilcoxon's rank sum test for two samples -- the Mann–Whitney U test

Used for

- independent samples when data is not normally distributed
- it is not sure whether the variable follows a normal distribution
- Ordinal scores

Named for

- Frank Wilcoxon in 1945: equal sample sizes
- Henry Berthold Mann (1905-2000), Austrian-born US mathematician and statistician; Donald Ransom Whitney in 1947: arbitrary sample sizes
- Also called the Mann–Whitney U test or Mann–Whitney–Wilcoxon (MWW) test.

THE U MANN-WHITNEY TEST

An alternative to the independent t-test.

Used when data is ordinal and non-parametric.

This test works on ranking the data rather than testing the actual scores (values), and scoring each rank (so the lowest score would be ranked '1', the next lowest '2' and so on) ignoring the group to which each participant belonged.

The principle of the test is that if the groups were equal (coming from the same population), then the sum of the ranks should also be the same.

Mann and Whitney assumptions and formulation

- All the observations from both groups are independent of each other,
- The responses are ordinal (i.e. one can at least say, of any two observations, which is the greater),
- The distributions of both groups are equal under the null hypothesis, so that the probability of an observation from one population (X) exceeding (greater) an observation from the second population (Y) equals the probability of an observation from Y exceeding an observation from X.
That is, there is a symmetry between populations with respect to probability of random drawing of a larger observation.
- Under the alternative hypothesis, the probability of an observation from one population (X) exceeding an observation from the second population (Y) (after exclusion of ties) is not equal to 0.5. The alternative may also be stated in terms of a one-sided test, for example: $P(X > Y) + 0.5 P(X = Y) > 0.5$.

Median Test for Two Independent Samples

- The Median test investigates if the medians of two independent samples are the same.
- The hypothesis under test, H_0 , is that the medians are the same, and this is to be tested against the alternative hypothesis H_1 that they are different.

Revision of how to Rank scores (raw values):

- (a) Lowest score gets rank of “1”; next lowest gets “2”; and so on.
- (b) Two or more scores with the same value are “tied”.
 - (i) Give each tied score the rank it would have had, had it been different from the other scores.
 - (ii) Add the ranks for the tied scores, and divide by the number of tied scores. Each of the ties gets this average rank.
 - (iii) The next score after the set of ties gets the rank it *would* have obtained, had there been no tied scores.

e.g.	raw score:	6	17	34	48
	“original” rank:	1	2	3	4
	raw score:	6	23	23	38
	“original” rank:	1	<u>2</u>	<u>3</u>	4
	“actual” rank:	1	2.5	2.5	4

Rationale of Mann-Whitney U

- Imagine two samples of scores drawn at random from the same population
- The two samples are combined into one larger group and then ranked from lowest to highest
- In this case there should be a similar number of high and low ranked scores in each original group
 - if you sum the ranks in each group, the totals should be about the same (approximately)
 - this is the null hypothesis
- If however, the two samples are from different populations with different medians then most of the scores from one sample will be lower in the ranked list than most of the scores from the other sample
 - the sum of ranks in each group will differ

Calculation procedure

- The test involves the calculation of a statistic, usually called U , whose distribution under the null hypothesis is known.
- In the case of small samples, the distribution is tabulated, but for sample sizes above ~ 20 approximation using the normal distribution is fairly good.
- First, arrange all the observations into a single ranked series. That is, rank all the observations without regard to which sample they are in.

Calculation schema – method one

- For very small samples a direct method is recommended. It is very quick, and gives an insight into the meaning of the U statistic
 - Choose the sample for which the ranks seem to be smaller (The only reason to do this is to make computation easier). Call this "sample 1," and call the other sample "sample 2."
 - For each observation in sample 1, count the number of observations in sample 2 that have a smaller rank (count a half for any that are equal to it). The sum of these counts is U .

Calculation schema – typical method

- Add up the ranks for the observations which came from sample 1. The sum of ranks in sample 2 is now determinate, since the sum of all the ranks equals $N(N + 1)/2$ where N is the total number of observations
- Calculate two statistics U_1 and U_2

$$U_1 = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - R_1$$

$$U_2 = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2 + 1)}{2} - R_2$$

- Choose the smaller U and compare with tabular statistics (for N_1 and N_2 degrees of freedom)

Mann-Whitney test, step-by-step:

Does it make any difference to students' comprehension of statistics whether the lectures are in English or in Polish?

Group 1: statistics lectures in English.

Group 2: statistics lectures in Polish.

DV: lecturer intelligibility ratings by students (0 = "unintelligible", 100 = "highly intelligible").

Ratings - so Mann-Whitney is appropriate.

English group (raw scores)	English group (ranks)	Polish group (raw scores)	Polish group (ranks)
18	17	17	15
15	10.5	13	8
17	15	12	5.5
13	8	16	12.5
11	3.5	10	1.5
16	12.5	15	10.5
10	1.5	11	3.5
17	15	13	8
		12	5.5
Median:	15.5	Median:	13

Step 1:

Rank all the scores together, regardless of group.

Step 2:

Add up the ranks for group 1, to get T_1 . Here, $T_1 = 83$.

Add up the ranks for group 2, to get T_2 . Here, $T_2 = 70$.

Step 3:

N_1 is the number of subjects in group 1; N_2 is the number of subjects in group 2. Here, $N_1 = 8$ and $N_2 = 9$.

Step 4:

Call the larger of these two rank totals T_x . Here, $T_x = 83$.

N_x is the number of subjects in this group. Here, $N_x = 8$.

Step 5:

Find U:

$$U = N1 * N2 + \frac{Nx (Nx + 1)}{2} - Tx$$

In our example,

$$U = 8 * 9 + \frac{8 * (8 + 1)}{2} - 83$$

$$U = 72 + 36 - 83 = 25$$

If there are unequal numbers of subjects - as in the present case - calculate U for *both* rank totals and then use the *smaller* U .

In the present example, for T_1 , $U = 25$, and for T_2 , $U = 47$. Therefore, use 25 as U .

Step 6:

Look up the critical value of U , (e.g. with the table on my website), taking into account N_1 and N_2 . If our obtained U is *equal to or smaller* than the critical value of U , we reject the null hypothesis and conclude that our two groups do differ significantly.

	N_2						
N_1		5	6	7	8	9	10
5		2	3	5	6	7	8
6		3	5	6	8	10	11
7		5	6	8	10	12	14
8		6	8	10	13	15	17
9		7	10	12	15	17	20
10		8	11	14	17	20	23

Here, the critical value of U for $N_1 = 8$ and $N_2 = 9$ is 15. Our obtained U of 25 is *larger* than this, and so we conclude that there is *no* significant difference between our two groups.

Conclusion: ratings of lecturer intelligibility are unaffected by whether the lectures are given in English or in Polish.

Mann-Whitney U: Another example

Step 1:

Rank all the data from both groups in one series, then total each

School A			School B		
Student	Grade	Rank	Student	Grade	Rank
J. S.	B-	9	T. J.	D	4
L. D.	B-	9	M. M.	C+	6.5
H. L.	A+	14	K. S.	C+	6.5
M. J.	D-	3	P. S.	B-	9
T. M.	B+	11	R. M.	E	2
T. S.	A-	12.5	P. W.	C-	5
P. H.	F	1	A. F.	A-	12.5
Median = ; $\sum R_A =$			Median = ; $\sum R_B =$		

Mann-Whitney U: Calculation

Step 2:

Calculate two versions of the U statistic using:

$$U_1 = (n_A \times n_B) + \frac{(n_A + 1) \times n_A}{2} - \sum R_A$$

AND...

$$U_2 = (n_A \times n_B) + \frac{(n_B + 1) \times n_B}{2} - \sum R_B$$

Mann-Whitney U: Calculation

Step 3 *finished*:

Select the smaller of the two U statistics ($U_1 = \dots\dots\dots$; $U_2 = \dots\dots\dots$)

...now consult a table of critical values for the Mann-Whitney test

n	6	7	8	9
0.05	5	8	13	17
0.01	2	4	7	11

Calculated U must be critical
U to conclude a significant difference

Conclusion

Median A Median B

The Wilcoxon signed test

- A non-parametric statistical hypothesis test used when comparing two related samples (paired)
- The test is named for Frank Wilcoxon (1892–1965) who, in a single paper, proposed both it and the rank-sum test for two independent samples (Wilcoxon, 1945).
- Assumptions
 - Data are paired and come from the same population.
 - Each pair is chosen randomly and independent.
 - The data are measured at least on an ordinal scale
- H_0 : median difference between the pairs is zero
- H_1 : median difference is not zero.

The Wilcoxon test – calculation schema

- N pairs of scores (values) x_{1i} and x_{2i} ($i=1, \dots, N$)
- Calculate differences $\text{sgn}(x_{1i} - x_{2i})$
- Exclude pairs with $|x_{1i} - x_{2i}|=0$; N_r – their reduced number
- Order the remaining pairs from smallest absolute difference to largest absolute difference and rank them
- Calculate total sums of ranks for positive and negative differences $T+$ and $T-$
- T statistics is the smaller of these rank sums
- For $N_r < 10$ compare it against tabular values (N_r degrees of freedom)
- If it is smaller than critical tabular value, reject H_0
- For larger N_r – approximation with normal distribution

Rationale of Wilcoxon test

- Some difference scores will be large, others will be small
- Some difference scores will be positive, others negative
- If there is no difference between the two experimental conditions then there will be similar numbers of positive and negative difference scores
- If there is no difference between the two experimental conditions then the numbers and sizes of positive and negative differences will be equal
 - this is the null hypothesis
- If there is a differences between the two experimental conditions then there will either be more positive ranks than negative ones, or the other way around
 - Also, the larger ranks will tend to lie in one direction

Wilcoxon test, step-by-step:

Does background music affect the mood of factory workers?

Eight workers: each tested twice.

Condition A: background music.

Condition B: silence.

DV: workers' mood rating (0 = "extremely miserable", 100 = "euphoric").

Ratings, so use Wilcoxon test.

Worker:	Silence	Music	Difference	Rank
1	15	10	5	4.5
2	12	14	-2	2.5
3	11	11	0	Ignore
4	16	11	5	4.5
5	14	4	10	6
6	13	1	12	7
7	11	12	-1	1
8	8	10	-2	2.5
	Median: 12.5	Median: 10.5		

Step 1:

Find the difference between each pair of scores, keeping track of the sign of the difference.

Step 2:

Rank the differences, *ignoring their sign*. Lowest = 1.

Tied scores dealt with as before.

Ignore zero difference-scores.

Step 3:

Add together the positive-signed ranks. $T^+ = 22$.

Add together the negative-signed ranks. $T^- = 6$.

Step 4:

„T" is the *smaller* sum of ranks; $T = 6$.

Nr is the number of differences, omitting zero differences; $Nr = 8 - 1 = 7$.

Step 5:

Use table to find the critical value of T, for your Nr.

Your obtained T has to be *equal to or smaller* than this critical value, for it to be statistically significant.

	One Tailed Significance levels:		
	0.025	0.01	0.005
	Two Tailed significance levels:		
N	0.05	0.02	0.01
6	0	-	-
7	2	0	-
8	4	2	0
9	6	3	2
10	8	5	3

The critical value of Tcrit (for an N of 7) is 2.
Our obtained T of 6 is *bigger* than this.
Our two conditions are not significantly different.

Conclusion: workers' mood appears to be unaffected by presence or absence of background music.

T Wilcoxon critical values

- See the book F. Cleg „Simple Statistics”.

Tablica S2. Wartości krytyczne T w teście Wilcozona dla wyników połączonych w pary – z uwzględnieniem znaku.
 T musi być równe lub mniejsze od wartości ustalonej jako istotna.

Poziom istotności dla testu jednostronnego					Poziom istotności dla testu jednostronnego				
0,05 0,025 0,01 0,005					0,05 0,025 0,01 0,005				
Poziom istotności dla testu dwustronnego					Poziom istotności dla testu dwustronnego				
N	0,10	0,05	0,02	0,01	N	0,10	0,05	0,02	0,01
5	1	—	—	—	28	130	117	101	92
6	2	1	—	—	29	141	127	111	100
7	4	2	0	—	30	152	137	120	109
8	6	4	2	0	31	163	148	130	118
9	8	6	3	2	32	175	159	141	128
10	11	8	5	3	33	188	171	151	138
11	14	11	7	5	34	201	183	162	149
12	17	14	10	7	35	214	195	174	160
13	21	17	13	10	36	228	208	186	171
14	26	21	16	13	37	242	222	198	183
15	30	25	20	16	38	256	235	211	195
16	36	30	24	19	39	271	250	224	208
17	41	35	28	23	40	287	264	238	221
18	47	40	33	28	41	303	279	252	234
19	54	46	38	32	42	319	295	267	248
20	60	52	43	37	43	336	311	281	262
21	68	59	49	43	44	353	327	297	277
22	75	66	56	49	45	371	343	313	292
23	83	73	62	55	46	389	361	329	307
24	92	81	69	61	47	408	379	345	323
25	101	90	77	68	48	427	397	362	339
26	110	98	85	76	49	446	415	380	356
27	120	107	93	84	50	466	434	398	373

Źródło: F. Wilcoxon, R. A. Wilcox, *Some rapid approximate statistical procedures*, str. 28, tablica 2, New York 1964. Przedruk za zgodą American Cyanamid Company.

Wilcoxon Signed Ranks: Calculation

Step 1:

Rank all the diffs from in one series (ignoring signs), then total each

Athlete	Pre-training OBLA (kph)	Post-training OBLA (kph)	Diff.	Rank	Signed Ranks	
					-	+
J. S.	15.6	16.1	0.5	6		6
L. D.	17.2	17.5	0.3	4.5		4.5
H. L.	17.7	16.7	-1	-7	-7	
M. J.	16.5	16.8	0.3	4.5		4.5
T. M.	15.9	16.0	0.1	1.5		1.5
T. S.	16.7	16.5	-0.2	-3	-3	
P. H.	17.0	17.1	0.1	1.5		1.5

Medians =

\sum Signed Ranks =

Wilcoxon Signed Ranks: Calculation

Step 2:

The smaller of the T values is our test statistic ($T^+ = \dots\dots\dots$; $T^- = \dots\dots\dots$)

...now consult a table of critical values for the Wilcoxon test

n	6	7	8	9
0.05	0	2	3	5

Calculated T must be critical
T to conclude a significant difference

Conclusion

Median A Median B

Nonparametric tests for comparing three or more groups

(a) Kruskal-Wallis test:

Similar to the Mann-Whitney test, except it enables you to compare *three or more* groups rather than just two. *Different* objects /subjects are used for each group.

(b) Friedman's Test:

Similar to the Wilcoxon test, except you can use it with *three or more* conditions.
Each object (subject) does *all* of the experimental conditions.

Friedman Test

- The Friedman test is a non-parametric statistical test developed by the U.S. economist Milton Friedman.
 - Similar to the parametric repeated measures ANOVA, it is used to detect differences in treatments across multiple groups.
1. H_0 : k groups are coming from the same population (or have the same medians)
 2. H_1 : not all groups are identical (or not medians are equal)
- The procedure involves ranking each row (or block) together, then considering the values of ranks by columns.

Calculation schema

- Given data X - a matrix with n rows (the blocks), k columns (the treatments) and a single observation (score/values) at the intersection of each block and treatment, calculate the ranks within each block.
- If there are tied values, assign to each tied value the average of the ranks that would have been assigned without ties.
- Calculate the total sum of ranks for each i -th group R_i ($i=1, \dots, k$)

Friedman Test - Statistics

Assumptions: $n_1 = n_2 = \dots = n_k$

Considered statistics

$$\chi^2 = \frac{12}{n_1 k(k+1)} \sum_{i=1}^k R_i^2 - 3n_1(k+1)$$

*Is asymptotically distributed as χ^2
with $(k-1)$ degrees of freedom*

If $\chi^2 \geq \text{critical } \chi^2$, then reject H_0

Friedman's test, step-by-step:

Effects on worker mood of different types of music:

Five workers. Each is tested three times, once under each of the following conditions (groups):

condition 1: silence.

condition 2: "easy-listening" music.

condition 3: marching-band music.

DV: mood rating ("0" = unhappy, "100" = euphoric).

Ratings - so use a nonparametric test.

NB: to avoid practice and fatigue effects, order of presentation of conditions is varied across subjects.

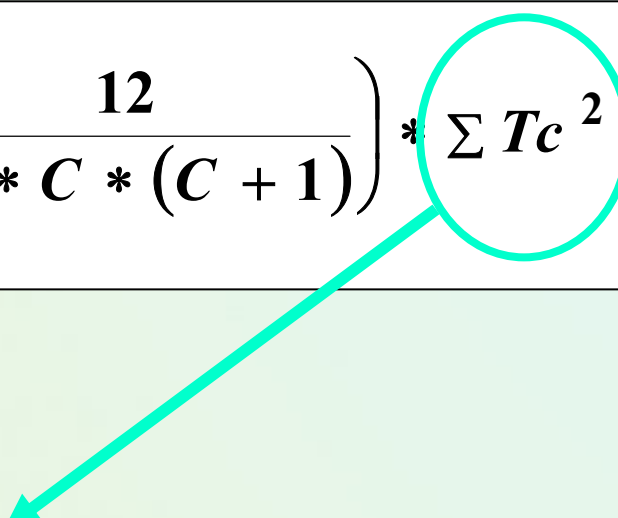
	Silence (raw score)	Silence (ranked score)	Easy (raw score)	Easy (ranked score)	Band (raw score)	Band (ranked score)
Wkr 1:	4	1	5	2	6	3
Wkr 2:	2	1	7	2.5	7	2.5
Wkr 3:	6	1.5	6	1.5	8	3
Wkr 4:	3	1	7	3	5	2
Wkr 5:	3	1	8	2	9	3

Step 1:

Rank *each subject's scores* individually.

Worker 1's scores are 4, 5, 6: these get ranks of 1, 2, 3.

Worker 4's scores are 3, 7, 5: these get ranks of 1, 3, 2 .

$$\chi^2 = \left[\left(\frac{12}{N * C * (C + 1)} \right) * \Sigma T_c^2 \right] - 3 * N * (C + 1)$$


To get ΣT_c^2 :

(a) square each rank total:

$$5.5^2 = 30.25. \quad 11^2 = 121. \quad 13.5^2 = 182.25.$$

(b) Add together these squared totals.

$$30.25 + 121 + 182.25 = 333.5.$$

In our example,

$$\chi r^2 = \left[\left(\frac{12}{N * C * (C + 1)} \right) * \Sigma T_c^2 \right] - 3 * N * (C + 1)$$

$$\chi r^2 = \left[\left(\frac{12}{5 * 3 * 4} \right) * 333.5 \right] - 3 * 5 * 4 = 6.7$$

$$\chi r^2 = 6.7$$

Step 4:

Degrees of freedom = number of conditions minus one. df = 3 - 1 = 2.

Step 5:

Assessing the statistical significance of χ^2 depends on the number of participants and the number of groups.

(a) *Less than 9 participants:*

Should use a special table of critical values.

(b) *9 or more participants:*

Use a Chi-Square table (Typical).

Compare your obtained χ^2 value to the critical value of χ^2 for your d.f.

If your obtained χ^2 is *bigger* than the critical χ^2 value, your conditions are significantly different.

The test only tells you that *some kind* of difference exists; look at the median or mean score for each condition (group) to see where the difference comes from.

Our obtained χ^2 is 6.7.

For 2 d.f., a χ^2 value of 5.99 would occur by chance with a probability of .05.

Our obtained value is *bigger* than 5.99.

Therefore our obtained χ^2 is even *less* likely to occur by chance: $p < .05$.

Conclusion: the conditions (groups) are significantly different. Music *does* affect worker mood.

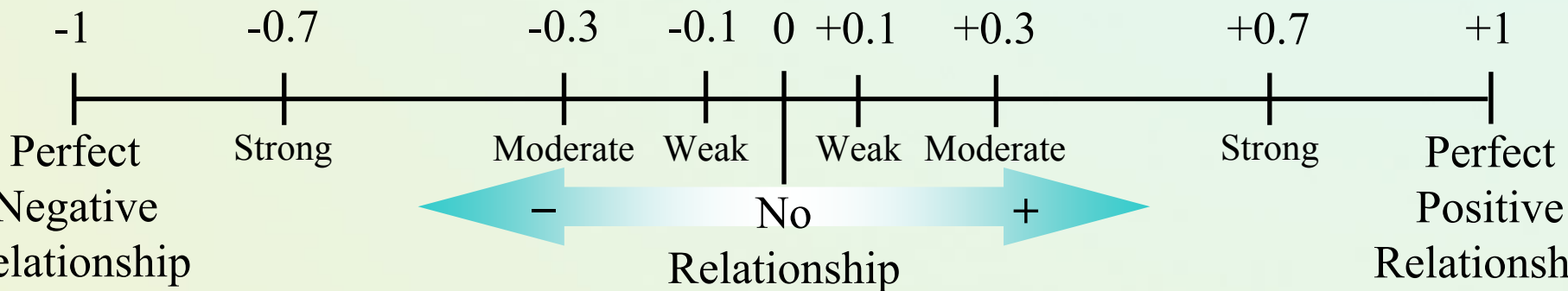
Dependencies and Correlations

- Dependence refers to any statistical relationship between two random variables or two sets of data
- Correlation refers to any of a broad class of statistical relationships involving dependence
- Examples
 - the correlation between the demand for a product and its price;
 - The correlation between electricity demand and weather, etc
- However, statistical dependence or correlation is not sufficient to demonstrate the presence of a causal relationship!

	Nominal	Ordinal	Int/Ratio
Nominal	Clustered bar-graph Chi-squared Phi (ϕ) or Cramer's V	Clustered bar-graph Chi-squared Phi (ϕ) or Cramer's V	Scatterplot, bar chart or error-bar chart Point bi-serial correlation (r_{pb})
Ordinal		Scatterplot or clustered bar chart Spearman's Rho or Kendall's Tau	\Rightarrow β Recode Scatterplot Point bi-serial or Spearman/Kendall
Int/Ratio			Scatterplot Product-moment correlation (r)

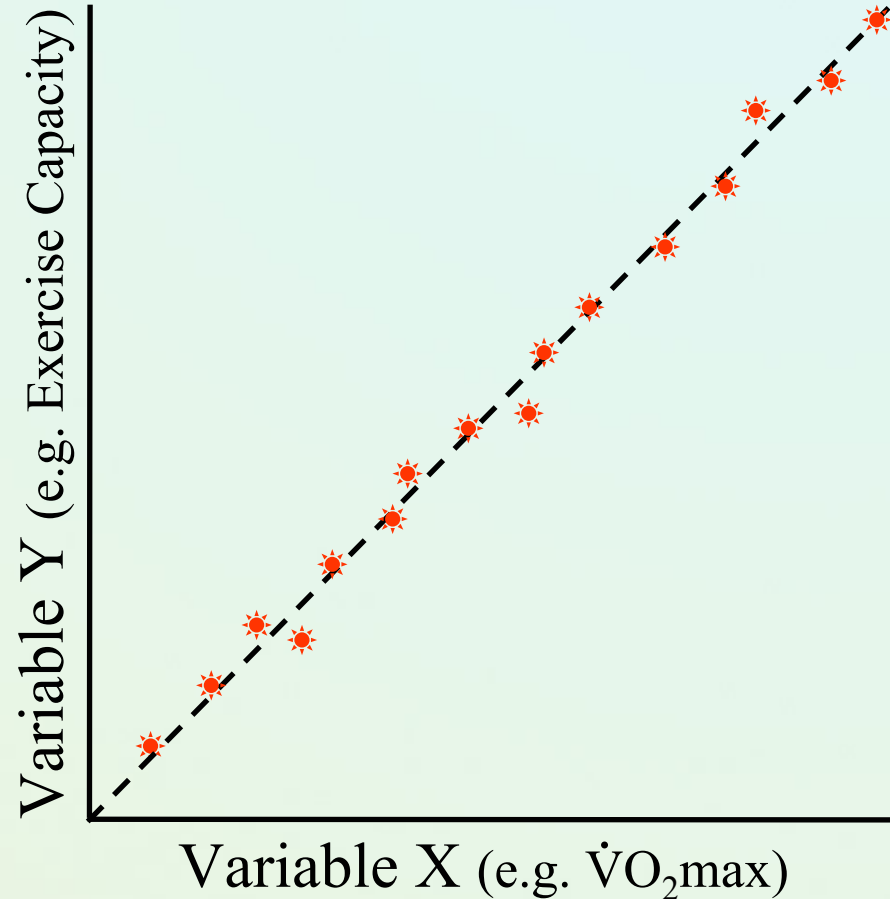
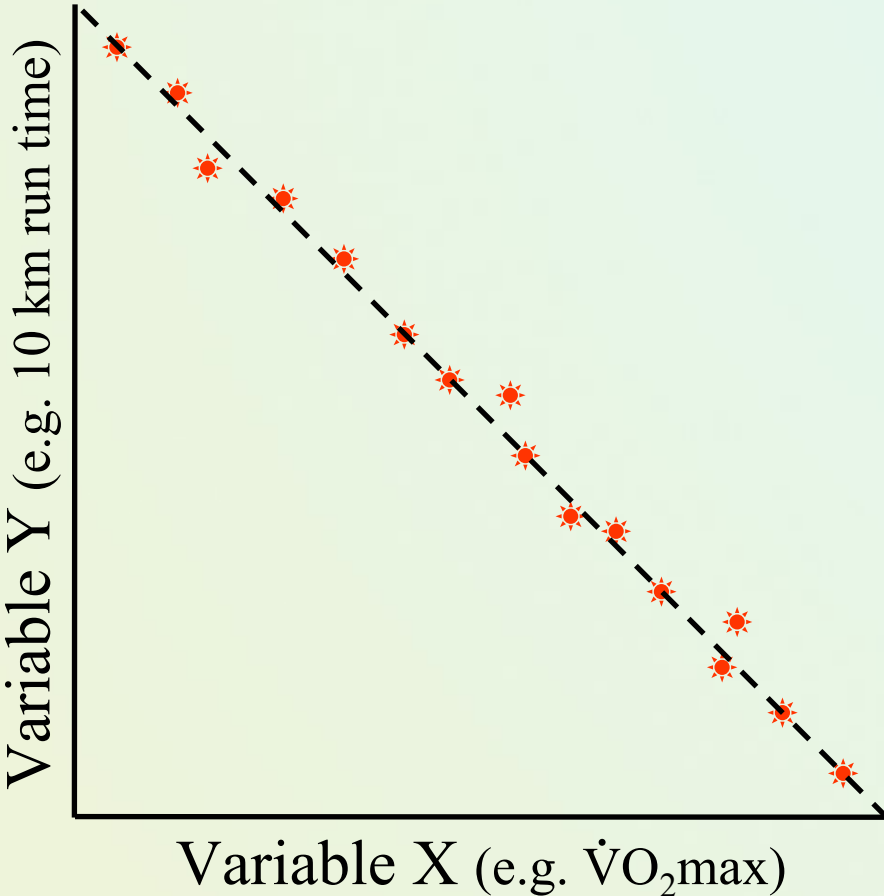
Correlation

- A measure of the relationship (correlation) between interval/ratio variables taken from the same set of subjects
- A ratio which indicates the amount of concomitant variation between two sets of scores
- This ratio is expressed as a correlation coefficient (r)



Correlation Coefficient & Scatterplots

Direction



Pearson correlation

Recall older definition and its assumptions

- interval or ratio level;
- linearly related;
- bivariate normally distributed.

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}},$$

Spearman rank correlation

Use if

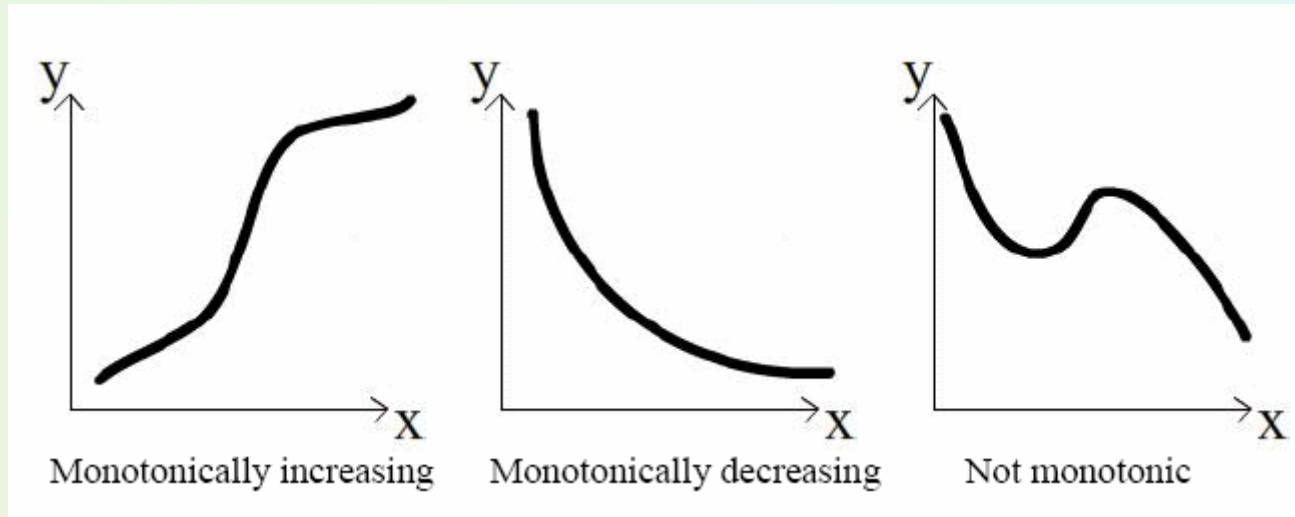
- Pearson assumptions do not hold
- Non-linear monotonic functions
- Ordinal data
- Rank correlation coefficients, such as Spearman's rank correlation coefficient and Kendall's rank correlation coefficient (τ) measure the extent to which, as one variable increases, the other variable tends to increase, without requiring that increase to be represented by a linear relationship. If, as the one variable increases, the other decreases, the rank correlation coefficients will be negative.

Simple illustrations

- The nature of rank correlation, and its difference from linear correlation, consider the following four pairs of numbers (x, y) : $(0, 1)$, $(10, 100)$, $(101, 500)$, $(102, 2000)$.
- an increase in x is always accompanied by an increase in y .
- The perfect rank correlation, and both Spearman's and Kendall's correlation coefficients are 1, whereas in this example Pearson product-moment correlation coefficient is 0.7544,

Monotonic relationships

- We refer to monotonic functions

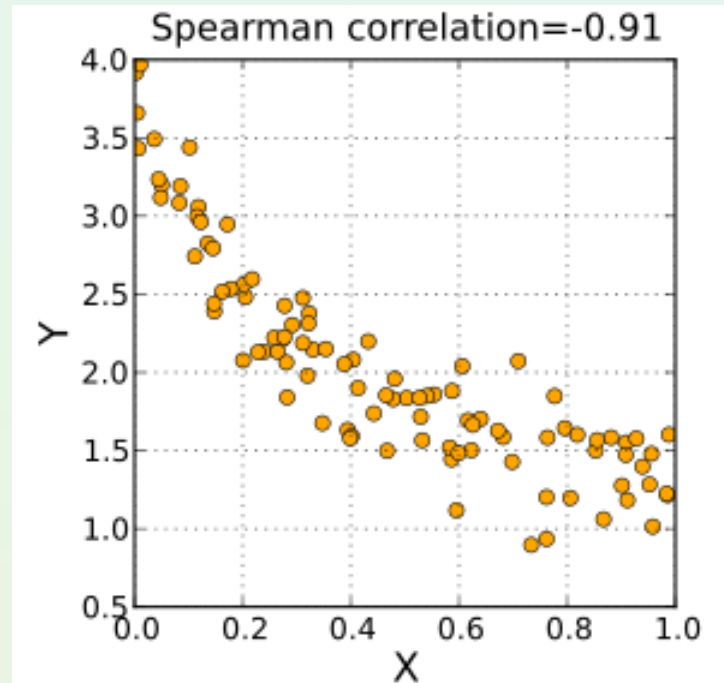
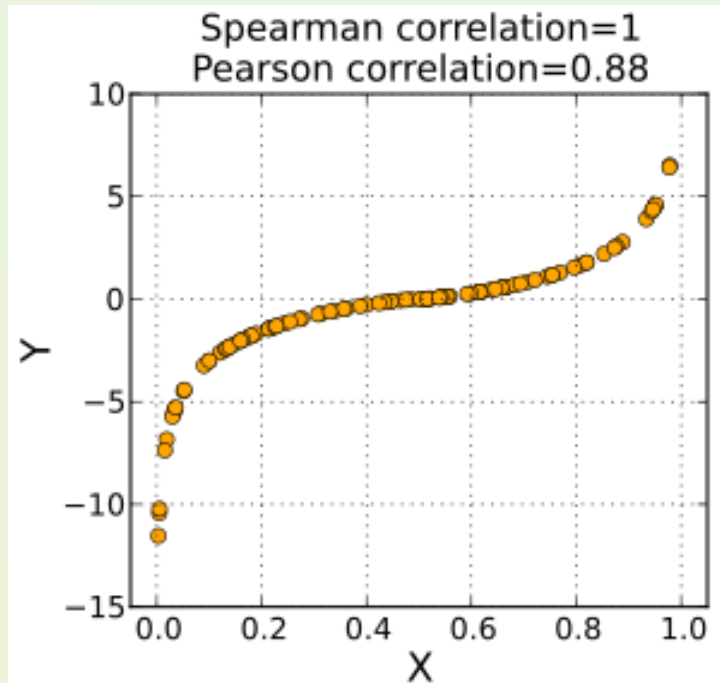


- Spearman's correlation coefficient is a statistical measure of the strength of a monotonic relationship between paired data $[-1; +1]$.

- .00-.19 “very weak”
- .20-.39 “weak”
- .40-.59 “moderate”
- .60-.79 “strong”
- .80-1.0 “very strong”

Differences for Spearman and Pearson correlations

- Non-linear monotonic relationships X and Y
- A perfect monotone increasing relationship implies that for any two pairs of data values X_i, Y_i and X_j, Y_j , that $X_i - X_j$ and $Y_i - Y_j$ always have the same sign.



Spearman correlation – calculation schema

- For a sample of size n , the n raw scores (x,y) are converted to ranks (i.e. rank values of variable x , and then variable y)
- Differences between the ranks of each observation on the two variables are calculated as d_i
- Spearman correlation r_s is defined as

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

Toy examples

- Study evaluations (scores) of 10 students according to 2 experts/teachers.

Student	A	B	C	D	E	F	G	H	I	J
1 exp.	42	27	36	33	24	47	39	52	43	37
2 exp.	39	24	35	29	26	47	44	51	39	32

Calculations

- Rankings

Student	A	B	C	D	E	F	G	H	I	J
1 exp.	42	27	36	33	24	47	39	52	43	37
ranks	7	2	4	3	1	9	6	10	8	5

Student	A	B	C	D	E	F	G	H	I	J
2 exp.	39	24	35	29	26	47	44	51	39	32
ranks	6.5	1	5	3	2	9	8	10	6.5	4

Calculations 2

- Rank differences

Student	A	B	C	D	E	F	G	H	I	J
1 exp.	7	2	4	3	1	9	6	10	8	5
2 exp.	6.5	1	5	3	2	9	8	10	6.5	4
d _i	0.5	1	-1	0	-1	0	-2	0	1.5	1
d _i ²	0.25	1	1	0	1	0	4	0	2.25	1

$$\sum_i d_i^2 = 10.5$$

$$r_s = 1 - \frac{6 \cdot 10.5}{10(10^2 - 1)} = 1 - \frac{63}{10 \cdot 99} = 0.936$$

Pearson's Rank-Order Correlation

X = Alcohol Units	Y = Skill Score	Rank X	Rank Y	D	D ²
15	4	10	1.5	8.5	72
14	6	9	3	6	36
10	4	8	1.5	6.5	42
9	8	7	5.5	1.5	2.3
8	7	5.5	4	1.5	2.3
8	8	5.5	5.5	0	0
7	10	4	8	4	16
6	9	3	7	4	16
4	14	2	10	8	64
2	12	1	9	8	64

Total=

More references

- Complete Business Statistics. Amir D.Aczel, 2008/
- Simple Statistics, Clegg F., 1994.
- and ...other WWW resources on non-parametrics tests



