Absolute frequency of "A"= a = number of apparition of "A"; **Relative frequency of "A"** = a/n*100 or =a/n

Median	**Quartile 0-4** – the value below which, a given percentage of data in a group of data fall	**Mode** - The most frequent value	**Range** – the difference between maximum and minimum value	**Arithmetic mean**
n = odd: $Me = x_{\frac{n+1}{2}}$ n = even: $Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$				$\overline{X} = \frac{1}{n}\sum X$

Sample standard deviation:	**Standard error (ES)**:	**Coefficient of variance (CV)**:	**Population standard deviation**:	where Σ - sum, X individual observations, s – sample standard deviation, $\bar{X}$ - arithmetic mean, n or N number of observations, X_1, … X_N – observation
$S = \sqrt{\frac{\sum(X-\bar{X})^2}{N-1}} = \sqrt{\frac{(X_1-\bar{X})^2+(X_2-\bar{X})^2+\ldots+(X_N-\bar{X})^2}{N-1}}$	$ES = \frac{s}{\sqrt{n}}$	$CV = \frac{s}{\bar{X}}$	$\sigma = \sqrt{\frac{\sum(X-\bar{X})^2}{N}}$	

Probability (A) = $\frac{\text{the number of times that A occurs}}{\text{the total number of trials}}$

P (nonA)=1-P(A)

A, B two events: P(A **or** B) = P(A) + P(B) - P(A and B)

A, B two <u>independent</u> events: P(A **and** B) = P(A) * P(B)

A, B two <u>mutually exclusive</u> events: P(A and B)=0

event B **<u>dependent</u>** on event A:
P(B dependent on A) = P(B|A) = P(A and B) / P(A)

Disease/ Test	B With illness	non(B) Without illness	Total
T Positive test	a (TP)	b (FP)	a+b
non (T) Negative test	c (FN)	d (TN)	c+d
Total	a+c	b+d	n

a – true positive TP	b – false positive FP
c – false negative FN	d – true negative TN
a+b – positive test	c+d – negative test
a+c – with illness	b+d – without illness

<u>Relative risk</u>: $RR = \frac{\textbf{risk of disease when the factor is present}}{\textbf{risk of disease when the factor is absent}} = \frac{P(B|A)}{P(B|\bar{A})} = \frac{P(Disease|Risk\ factor)}{P(Disease|\overline{Risk\ factor})} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$

<u>Positive predictive value</u> - Probability of a positive test to be true (indicate the disease): $PPV = Pr(B\|T) = \frac{TP}{TP+FP} = \frac{a}{a+b}$	**<u>Negative predictive value</u>** - Probability of a negative test to be true (indicate no disease): $NPV = Pr(nonB\|nonT) = \frac{TN}{TN+FN} = \frac{d}{c+d}$
<u>Test sensitivity</u> - Probability of a people with disease to get a positive test: $Se = Pr(T\|B) = \frac{TP}{TP+FN} = \frac{a}{a+c}$	**<u>Test specificity</u>** - Probability of a people without disease to get a negative test: $Sp = Pr(nonT\|nonB) = \frac{TN}{TN+FP} = \frac{d}{b+d}$

A series is **<u>normally distributed</u>** if: arithmetic mean = median = mode (or near equal); Quartile 1, Quartile 3 are simetrical with the mean (or near simetrical); Skewness ≈ 0 (between -1 to 1); Kurtosis ≈ 0 (between -1 to 1); in the interval: mean ± st.dev. there are minimum 68.2% of data; in the interval: mean ±2* st.dev. there are minimum 95.4% of data; in the interval: mean ±3* st.dev. there are minimum 99.7% of data, where st.dev. – standard deviation, mean – arithmetic mean, mean ± st.dev.

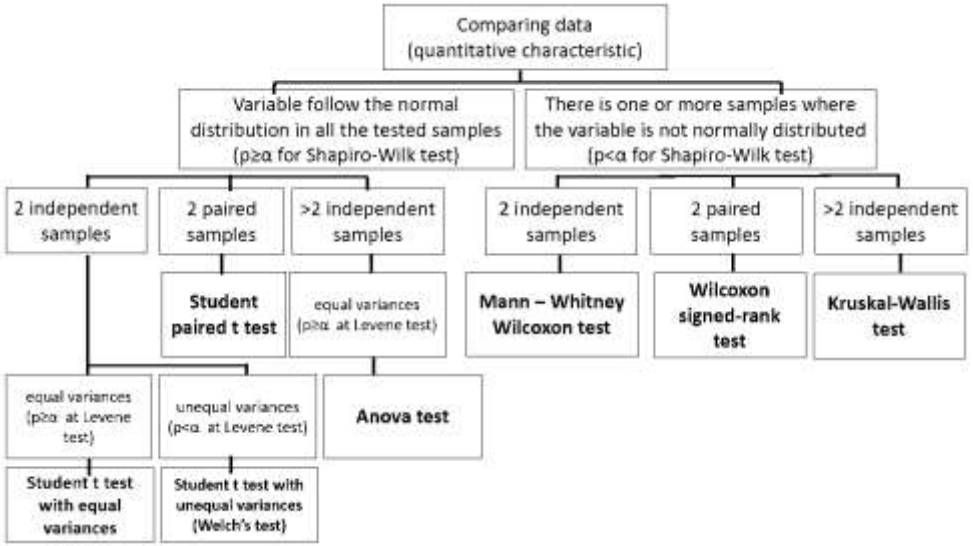
<u>95% confidence interval for the arithmetic mean</u> μ when σ is unknown and sample size n≥30: $[\bar{X} - 1.96\frac{s}{\sqrt{n-1}}, \bar{X} + 1.96\frac{s}{\sqrt{n-1}}]$ or $[\bar{X} - 1.96 \cdot SE; \bar{X} + 1.96 \cdot SE]$, where $\bar{X}$ – the sample arithmetic mean, s – the sample standard deviation, SE – the sample standard error, n – sample size, μ – population arithmetic mean, σ – population standard deviation

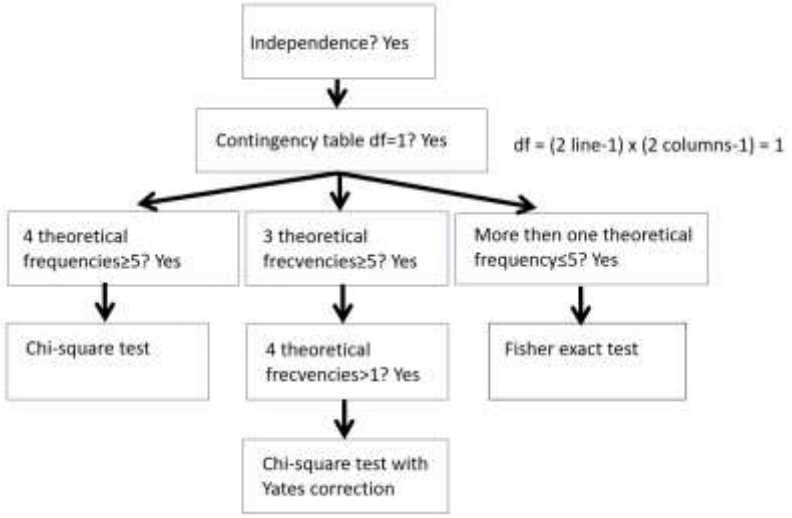
<u>95% confidence interval for the frequency</u> in the population $\left[f - 1.96\sqrt{\frac{f(1-f)}{n}}; f + 1.96\sqrt{\frac{f(1-f)}{n}}\right]$, where f – the sample frequency, f<1, n – sample size

<u>$1-\alpha$ confidence interval for the average</u> of the population μ in the case of small samples n<30 with σ unknown $[\bar{X} - t_{n-1,1-\frac{\alpha}{2}}\frac{s}{\sqrt{n-1}}, \bar{X} + t_{n-1,1-\frac{\alpha}{2}}\frac{s}{\sqrt{n-1}}]$, where $\bar{X}$ - sample arithmetic mean, s – sample standard deviation, n – sample size, $t_{n-1,1-\frac{\alpha}{2}}$ critical t for n-1 degree of freedom (also noted with t_α), $1-\alpha$ level of confidence

Student t-test for independent samples in the case of unequal variances H0 - null hypothesis at the population level, there is no statistical significant difference between group 1 average and group 2 average; **H1 - alternative hypothesis:** at the population level, there is statistical significant difference between group 1 average and group 2 average; **Assumptions:** The observations are independent, two independent samples, unequal variances, normal distributions (n_1 or n_2<30). **Rejection interval** – $(-\infty, -Z_\alpha) \cup (Z_\alpha, \infty)$; **Acceptance interval**: $[-Z_\alpha, Z_\alpha]$;

for α=0.05, Z_α=1.96. **Test parameter** $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$ where $\bar{X}_1$, $\bar{X}_2$- samples arithmetic means, s_1, s_2 – samples standard deviations, n – sample size. If parameter of the test ∈ acceptance interval --> we fail to reject null hypothesis H_0. If parameter of the test ∈ rejection interval --> we reject null hypothesis H_0, accept H_1. If p≥α we fail to reject null hypothesis H_0. If p<α we reject null hypothesis H_0, accept H_1.





If samples are not independent? Mc Nemar test

Chi-square test Null hypothesis H0 - At the population level, there is **no** statistically significant association between 2 variables: risk factor – disease. **Alternative hypothesis H1:** At the population level, there is a statistically significant association between 2 variables: Risk factor – disease. **α=5%**; Rejection region (3.84; +∞); Acceptance area (0; 3.84]

Observed frequency table

	Disease+	Disease-	Total
Risk+	a	b	a+b
Risk-	c	d	c+d
Total	a+c	b+d	n

Theoretical frequency table

	Disease+	Disease-	Total
Risk+	$\frac{(a+c)*(a+b)}{n}$	$\frac{(b+d)*(a+b)}{n}$	a+b
Risk-	$\frac{(a+c)*(c+d)}{n}$	$\frac{(b+d)*(c+d)}{n}$	c+d
Total	a+c	b+d	n

Test parameter: $\chi^2 = \sum_{i=1}^{4} \frac{(f_i^o - f_i^t)^2}{f_i^t}$, where f_i^o observed frequency, f_i^t theoretical frequency. If χ2 belongs to (3.84; +∞) reject H0, we accept H1. If χ2 DOES NOT belong to (3.84; +∞) we are in favor of H0. If p < 0.05 reject H0, we accept H1. If p ≥ 0.05 we are in favor of H0.

Pearson coefficient of correlation – r - Assumption: Normal distribution for both variables, two quantitative variables

Spearman coefficient of correlation – Assumption: two quantitative or ordinal variables. **Regression line**: Y=aX+b

Statistical test for correlation H0 (null hyphotesis): ρ (r at the population level) is not significant statistically different than 0; H1 (alternative hyphotesis): ρ (r at the population level) is significant statistically different than 0; If p<α, we reject the null hypothesis (H0) and accept the alternative hypothesis (H1): the correlation is statistically significant; If p≥α, we fail to reject the null hypothesis (H0): the correlation is not statistically significant