FS - UNIT IV

Hypothesis Testing is a type of statistical analysis in which you put your assumptions about a population parameter to the test. It is used to estimate the relationship between 2 statistical variables.

Let's discuss few examples of statistical hypothesis from real-life -

- A teacher assumes that 60% of his college's students come from lower-middle-class families.
- A doctor believes that 3D (Diet, Dose, and Discipline) is 90% effective for diabetic patients.

Now that you know about hypothesis testing, look at the two types of hypothesis testing in statistics.

Null Hypothesis and Alternate Hypothesis

- The Null Hypothesis is the assumption that the event will not occur. A null hypothesis has no bearing on the study's outcome unless it is rejected.
- H0 is the symbol for it, and it is pronounced H-naught.
- The Alternate Hypothesis is the logical opposite of the null hypothesis. The acceptance of the alternative hypothesis follows the rejection of the null hypothesis. H1 is the symbol for it.
- Let's understand this with an example.
    1. A sanitizer manufacturer claims that its product kills 95 percent of germs on average.
    2. To put this company's claim to the test, create a null and alternate hypothesis.
    3. H0 (Null Hypothesis): Average = 95%.
    4. Alternative Hypothesis (H1): The average is less than 95%.
    5. Another straightforward example to understand this concept is determining whether or not a coin is fair and balanced. The null hypothesis states that the probability of a show of heads is equal to the likelihood of a show of tails. In contrast, the alternate theory states that the probability of a show of heads and tails would be very different.

One-Tailed and Two-Tailed Hypothesis Testing

- The One-Tailed test, also called a directional test, considers a critical region of data that would result in the null hypothesis being rejected if the test sample falls into it, inevitably meaning the acceptance of the alternate hypothesis.
- In a one-tailed test, the critical distribution area is one-sided, meaning the test sample is either greater or lesser than a specific value.
- In two tails, the test sample is checked to be greater or less than a range of values in a Two-Tailed test, implying that the critical distribution area is two-sided.
- If the sample falls within this range, the alternate hypothesis will be accepted, and the null hypothesis will be rejected.

Example:

- Suppose H0: mean = 50 and H1: mean not equal to 50
- According to the H1, the mean can be greater than or less than 50. This is an example of a Two-tailed test.
- In a similar manner, if H0: mean >=50, then H1: mean <50
- Here the mean is less than 50. It is called a One-tailed test.

Type 1 and Type 2 Error

A hypothesis test can result in two types of errors.
Type 1 Error: A Type-I error occurs when sample results reject the null hypothesis despite being true.
Type 2 Error: A Type-II error occurs when the null hypothesis is not rejected when it is false, unlike a Type-I error.

Example:

- Suppose a teacher evaluates the examination paper to decide whether a student passes or fails.
- H0: Student has passed
- H1: Student has failed
- Type I error will be the teacher failing the student [rejects H0] although the student scored the passing marks [H0 was true].
- Type II error will be the case where the teacher passes the student [do not reject H0] although the student did not score the passing marks [H1 is true].

Level of Significance

- The alpha value is a criterion for determining whether a test statistic is statistically significant.
- In a statistical test, Alpha represents an acceptable probability of a Type I error.
- Because alpha is a probability, it can be anywhere between 0 and 1.
- In practice, the most commonly used alpha values are 0.01, 0.05, and 0.1, which represent a 1%, 5%, and 10% chance of a Type I error, respectively (i.e. rejecting the null hypothesis when it is in fact correct).

P-Value

- A p-value is a metric that expresses the likelihood that an observed difference could have occurred by chance.
- As the p-value decreases the statistical significance of the observed difference increases. If the p-value is too low, you reject the null hypothesis.
- Here you have taken an example in which you are trying to test whether the new advertising campaign has increased the product's sales.
- The p-value is the likelihood that the null hypothesis, which states that there is no change in the sales due to the new advertising campaign, is true.
- If the p-value is .30, then there is a 30% chance that there is no increase or decrease in the product's sales.
- If the p-value is 0.03, then there is a 3% probability that there is no increase or decrease in the sales value due to the new advertising campaign.

- As you can see, the lower the p-value, the chances of the alternate hypothesis being true increases, which means that the new advertising campaign causes an increase or decrease in sales.

Interpretation of statistical results

- Clinical research is an indispensable means for the advancement of scientific knowledge and to implement it into the routine clinical practice in order to provide the patients with the best opportunities to recover or improve their health understood as years and quality of life.
- But to accomplish this we are going to need tools to be able to conduct research, describe the biological reality, facilitate the understanding of clinical research and allow manipulation through experiments in order to establish associations between stimuli (drugs, surgical technique, etc.) and interesting results.
- Statistical techniques are mathematical models that require certain knowledge for their interpretation. Without an adequate understanding, the generalization of the study results can be useless or dangerous.
- From an ethical point of view making the effort of trying to understand is essential if we wish to be updated on scientific advances.
- Also, given the huge scientific production available today fuelled by the need for publishing to be promoted professionally, it is essential to know how to interpret statistical results in order to distinguish the important stuff from the unimportant one, develop an analytical spirit, and assess any possible implications to our clinical and research practice.

Association VS. Causal relationships

Association

- When two variables are related, we say that there is association between them. When researchers find a correlation, which can also be called an association, what they are saying is that they found a relationship between two, or more, variables.

Causal

- One variable has a direct influence on the other, this is called a causal relationship.
- Causality can only be determined by reasoning about how the data were collected.
- The data values themselves contain no information that can help you to decide.
- If two variables are causally related, it is possible to conclude that changes to the explanatory variable, X, will have a direct impact on Y.
- If one variable causally affects the other, then adjusting the value of that variable will cause the other to change.
- Obviously, it is much more difficult to prove causation than it is to prove an association.

Statistical Inference Definition

Statistical inference is a method of making decisions about the parameters of a population, based on random sampling. It helps to assess the relationship between the dependent and independent variables. The purpose of statistical inference to estimate the uncertainty or sample to sample variation.

Types of Statistical Inference

- Pearson Correlation
- Bi-variate regression
- Multi-variate regression
- Chi-square statistics and contingency table

Importance of Statistical Inference

Inferential Statistics is important to examine the data properly. To make an accurate conclusion, proper data analysis is important to interpret the research results. It is majorly used in the future prediction for various observations in different fields.

The statistical inference has a wide range of application in different fields, such as:

- Business Analysis
- Artificial Intelligence
- Financial Analysis
- Fraud Detection
- Machine Learning
- Share Market
- Pharmaceutical Sector

An example of statistical inference is given below.

Question: From the shuffled pack of cards, a card is drawn. This trial is repeated for 400 times, and the suits are given below:

Suit	Spade	Clubs	Hearts	Diamonds
No.of times drawn	90	100	120	90

While a card is tried at random, then what is the probability of getting a

1. Diamond cards
2. Black cards
3. Except for spade

Solution:

By statistical inference solution,

Total number of events = 400

i.e.,90+100+120+90=400

(1) The probability of getting diamond cards:

Number of trials in which diamond card is drawn = 90

Therefore, P(diamond card) = 90/400 = 0.225

(2) The probability of getting black cards:

Number of trials in which black card showed up = 90+100 =190

Therefore, P(black card) = 190/400 = 0.475

(3) Except for spade
Number of trials other than spade showed up = 90+100+120 =310
Therefore, P(except spade) = 310/400 = 0.775

Sample Statistic

- A **sample statistic** is any quantity from the sample of a population. A sample is a group of elements chosen from the population.
- The features that describe the population are called the parameters and the properties of the sample data are known as **statistics**. Population and sample both are important parts of statistics.

Example
The following GPA score of 30 High School students. Find the sample mean and standard deviation.
3.1, 2.9, 2.8, 2.9, 3.8, 4.8, 4.2, 3.9, 3.4, 2.5, 4.2, 3.7, 3.3, 2.1, 3.8, 3.0, 3.7, 4.0, 2.7, 3.8, 3.2, 3.5, 3.5, 3.6, 2.2, 3.1, 3.5, 4.0, 2.7, 4.5.
Solution: Using the formulas to compute the sample mean and standard deviation, we get,
Sample mean:
$\bar{x} = \sum x/n = 102.4/30 = 3.41$
Sample standard deviation:
$S = \sqrt{(\sum(x_i - \bar{x})^2/n-1)} = 0.65.$
statistical and practical significance
While **statistical significance** shows that an effect exists in a study, **practical significance** shows that the effect is large enough to be meaningful in the real world. Statistical significance is denoted by *p*-values whereas practical significance is represented by effect sizes.

Statistical Significance

- The hypothesis testing procedure determines whether the sample results that you obtain are likely if you assume the null hypothesis is correct for the population.
- If the results are sufficiently improbable under that assumption, then you can reject the null hypothesis and conclude that an effect exists. In other words, the strength of the evidence in your sample has passed your defined threshold of the significance level (alpha).
- Your results are statistically significant.
- You use p-values to determine statistical significance in hypothesis tests such as t-tests, ANOVA, and regression coefficients among many others.
- Consequently, it might seem logical that p-values and statistical significance relate to importance. However, that is false because conditions other than large effect sizes can produce tiny p-values.
- Hypothesis tests with small effect sizes can produce very low p-values when you have a large sample size and/or the data have low variability.
- Consequently, effect sizes that are trivial in the practical sense can be highly statistically significant.

Practical vs. Statistical Significance

Statistical significance does not necessarily mean that the results are practically significant in a real-world sense of importance.
In this, I'll talk about the differences between practical significance and statistical significance, and how to determine if your results are meaningful in the real world.

Statistical Significance

- The hypothesis testing procedure determines whether the sample results that you obtain are likely if you assume the null hypothesis is correct for the population.
- If the results are sufficiently improbable under that assumption, then you can reject the null hypothesis and conclude that an effect exists.
- In other words, the strength of the evidence in your sample has passed your defined threshold of the significance level (alpha). Your results are statistically significant.
- You use p-values to determine statistical significance in hypothesis tests such as t-tests, ANOVA, and regression coefficients among many others.
- Consequently, it might seem logical that p-values and statistical significance relate to importance. However, that is false because conditions other than large effect sizes can produce tiny p-values.
- Hypothesis tests with small effect sizes can produce very low p-values when you have a large sample size and/or the data have low variability.
- Consequently, effect sizes that are trivial in the practical sense can be highly statistically significant.

Here's how small effect sizes can still produce tiny p-values:

You have a very large sample size.

1. As the sample size increases, the hypothesis test gains greater statistical power to detect small effects.
2. With a large enough sample size, the hypothesis test can detect an effect that is so minuscule that it is meaningless in a practical sense.

The sample variability is very low.

1. When your sample data have low variability, hypothesis tests can produce more precise estimates of the population's effect. This precision allows the test to detect tiny effects.
2. Statistical significance indicates only that you have sufficient evidence to conclude that an effect exists. It is a mathematical definition that does not know anything about the subject area and what constitutes an important effect.

Practical Significance

1. While statistical significance relates to whether an effect exists, practical significance refers to the magnitude of the effect. However, no statistical test can tell you whether the effect is large enough to be important in your field of study.
2. Instead, you need to apply your subject area knowledge and expertise to determine whether the effect is big enough to be meaningful in the real world.
3. Again, this process requires that you use your knowledge of the subject to make this determination.
4. If your study's effect size is greater than this smallest meaningful effect, your results are practically significant.

For example, suppose you are evaluating a training program by comparing the test scores of program participants to those who study on their own. Further, we decide that the difference

between these two groups must be at least five points to represent a practically meaningful effect size. An effect of 4 points or less is too small to care about.

After performing the study, the analysis finds a statistically significant difference between the two groups. Participants in the study program score an average of 3 points higher on a 100-point test. While these results are statistically significant, the 3-point difference is less than our 5-point threshold. Consequently, our study provides evidence that this effect exists, but it is too small to be meaningful in the real world. The time and money that participants spend on the training program are not worth an average improvement of only 3 points.

The Two-Sample *t*-Test

- The two-sample *t*-test (also known as the independent samples *t*-test) is a method used to test whether the unknown population means of two groups are equal or not.
- For the two-sample *t*-test, we need two variables. One variable defines the two groups. The second variable is the measurement of interest.
- We also have an idea, or hypothesis, that the means of the underlying populations for the two groups are different.

Here are a couple of examples:

1. We have students who speak English as their first language and students who do not. All students take a reading test. Our two groups are the native English speakers and the non-native speakers. Our measurements are the test scores. Our idea is that the mean test scores for the underlying populations of native and non-native English speakers are not the same. We want to know if the mean score for the population of native English speakers is different from the people who learned English as a second language.
2. We measure the grams of protein in two different brands of energy bars. Our two groups are the two brands. Our measurement is the grams of protein for each energy bar. Our idea is that the mean grams of protein for the underlying populations for the two brands may be different. We want to know if we have evidence that the mean grams of protein for the two brands of energy bars is different or not.

Wilcoxon Sign Test

The Wilcoxon signed rank test is the non-parametric of the dependent samples t-test. Because the dependent samples t-test analyzes if the average difference of two repeated measures is zero, it requires metric (interval or ratio) and normally distributed data; the Wilcoxon sign test uses ranked or ordinal data; thus, it is a common alternative to the dependent samples t-test when its assumptions are not met.

The Wilcoxon signed rank test relies on the W-statistic. For large samples with n>10 paired observations the W-statistic approximates a normal distribution. The W statistic is a non-parametric test, thus it does not need multivariate normality in the data.

The first step of the Wilcoxon sign test is to calculate the differences of the repeated measurements and to calculate the absolute differences.

Case	X	Y	X-Y	Absolute X-Y
1	1	7	-6	6
2	5	7	-2	2
3	4	10	-6	6
4	5	9	-4	4
5	5	8	-3	3
6	0	3	-3	3
7	5	2	3	3
8	5	2	3	3
9	0	10	-10	10
10	0	10	-10	10
11	5	6	-1	1
12	0	10	-10	10
13	6	10	-4	4
14	6	6	0	0
15	0	10	-10	10
16	0	5	-5	5
17	0	10	-10	10
18	1	8	-7	7
19	2	9	-7	7
20	2	10	-8	8

The next step of the Wilcoxon sign test is to sign each rank. If the original difference < 0 then the rank is multiplied by -1; if the difference is positive the rank stays positive

Case	X	Y	X-Y	Absolute X-Y
14	6	6	0	0
11	5	6	-1	1
2	5	7	-2	2
5	5	8	-3	3
6	0	3	-3	3
7	5	2	3	3
8	5	2	3	3
4	5	9	-4	4
13	6	10	-4	4
16	0	5	-5	5
1	1	7	-6	6
3	4	10	-6	6
18	1	8	-7	7
19	2	9	-7	7
20	2	10	-8	8
10	0	10	-10	10
15	0	10	-10	10
9	0	10	-10	10
12	0	10	-10	10
17	0	10	-10	10

For the Wilcoxon signed rank test we can ignore cases where the difference is zero. For all other cases we assign their relative rank. In case of tied ranks the average rank is calculated. That is if rank 10 and 11 have the same observed differences both are assigned rank 10.5.

The next step of the Wilcoxon sign test is to sign each rank. If the original difference < 0 then the rank is multiplied by -1; if the difference is positive the rank stays positive.

Case	X	Y	X-Y	Absolute Y-X	Rank	With tied ranks	Signed ranks
14	6	6	0	0	-	-	-
11	5	6	-1	1	1	1.0	-1.0
2	5	7	-2	2	2	2.0	-2.0
5	5	8	-3	3	3	4.5	-4.5
6	0	3	-3	3	4	4.5	-4.5
7	5	2	3	3	5	4.5	4.5
8	5	2	3	3	6	4.5	4.5
4	5	9	-4	4	7	7.5	-7.5
13	6	10	-4	4	8	7.5	-7.5
16	0	5	-5	5	9	9.0	-9.0
1	1	7	-6	6	10	10.5	-10.5
3	4	10	-6	6	11	10.5	-10.5
18	1	8	-7	7	12	12.5	-12.5
19	2	9	-7	7	13	12.5	-12.5
20	2	10	-8	8	14	14.0	-14.0
10	0	10	-10	10	15	17.0	-17.0
15	0	10	-10	10	16	17.0	-17.0
9	0	10	-10	10	17	17.0	-17.0
12	0	10	-10	10	18	17.0	-17.0
17	0	10	-10	10	19	17.0	-17.0

The next step is to calculate the W+ and W-.

W+ = 4.5 + 4.5 = 9

W- = 1+ 2 + 4.5 + 4.5 + 7.5 + 7.5 + 9 + 10.5 + 10.5 + 12.5 + 12.5 + 14 + 17+ 17 + 17 + 17 + 17 = 181.

The shortcut to the hypothesis testing of the Wilcoxon signed rank-test is knowing the critical z-value for a 95% confidence interval (or a 5% level of significance) which is z = 1.96 for a two-tailed test and directionality. Whenever a test is based the normal distribution the sample z value needs to be 1.96 or higher to reject the null hypothesis

Mann Whitney U Test (Wilcoxon Rank Sum Test)

A popular nonparametric test to compare outcomes between two independent groups is the Mann Whitney U test. The Mann Whitney U test, sometimes called the Mann Whitney Wilcoxon Test or the Wilcoxon Rank Sum Test, is used to test whether two samples are likely to derive from the same population (i.e., that the two populations have the same shape). Some investigators interpret this test as comparing the medians between the two populations. Recall that the parametric test compares the means (H_0: $\mu_1=\mu_2$) between independent groups.

In contrast, the null and two-sided research hypotheses for the *nonparametric test* are stated as follows:

H_0: The two populations are equal versus

H_1: The two populations are not equal.

This test is often performed as a two-sided test and, thus, the research hypothesis indicates that the populations are not equal as opposed to specifying directionality. A one-sided research hypothesis is used if interest lies in detecting a positive or negative shift in one population as compared to the other. The procedure for the test involves pooling the observations from the two samples into one combined sample, keeping track of which sample each observation comes from, and then ranking lowest to highest from 1 to n_1+n_2, respectively.

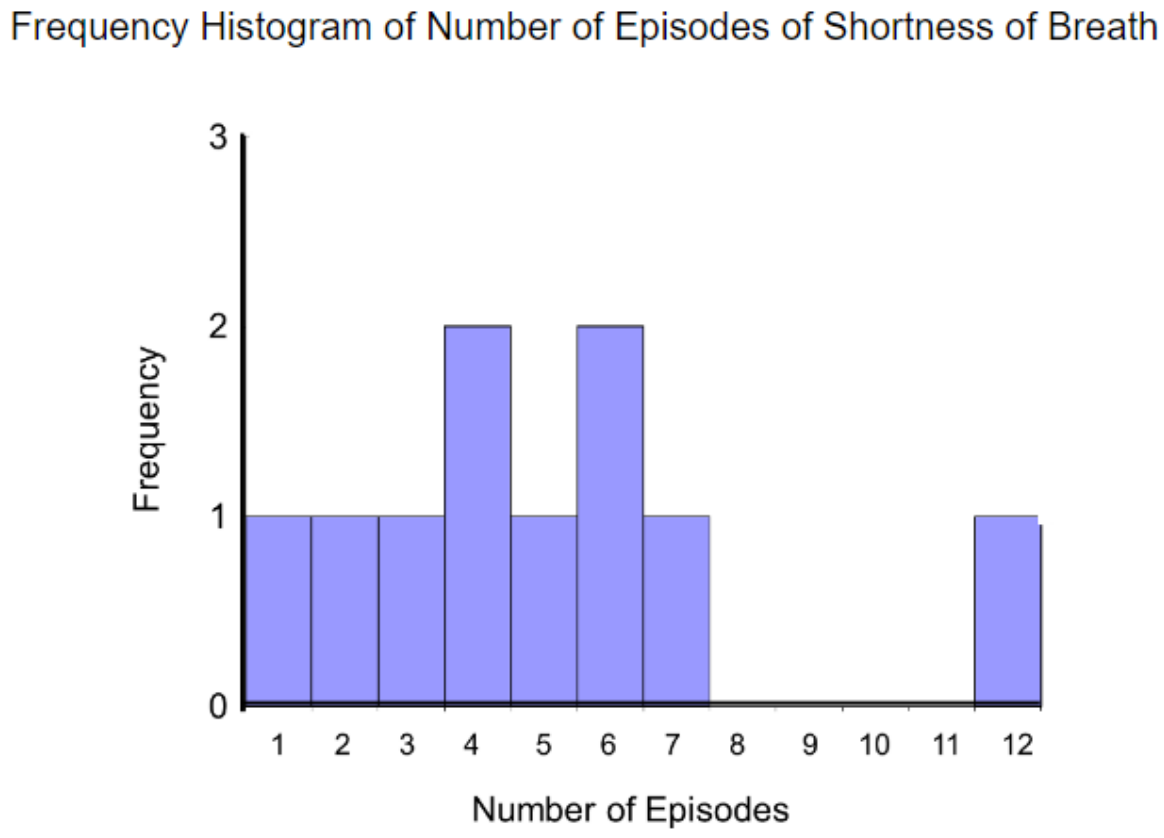
Example:

Consider a Phase II clinical trial designed to investigate the effectiveness of a new drug to reduce symptoms of asthma in children. A total of n=10 participants are randomized to receive either the new drug or a placebo. Participants are asked to record the number of episodes of shortness of breath over a 1 week period following receipt of the assigned treatment. The data are shown below.

Placebo	7	5	6	4	12
New Drug	3	6	4	2	1

Is there a difference in the number of episodes of shortness of breath over a 1 week period in participants receiving the new drug as compared to those receiving the placebo? By inspection, it appears that participants receiving the placebo have more episodes of shortness of breath, but is this statistically significant?

In this example, the outcome is a count and in this sample the data do not follow a normal distribution.



In addition, the sample size is small ($n_1=n_2=5$), so a nonparametric test is appropriate. The hypothesis is given below, and we run the test at the 5% level of significance (i.e., $\alpha=0.05$).

H_0: The two populations are equal versus

H_1: The two populations are not equal.

Note that if the null hypothesis is true (i.e., the two populations are equal), we expect to see similar numbers of episodes of shortness of breath in each of the two treatment groups, and we would expect to see some participants reporting few episodes and some reporting more episodes in each group.

This does not appear to be the case with the observed data. A test of hypothesis is needed to determine whether the observed data is evidence of a statistically significant difference in populations.

The first step is to assign ranks and to do so we order the data from smallest to largest. This is done on the combined or total sample (i.e., pooling the data from the two treatment groups (n=10)), and assigning ranks from 1 to 10, as follows. We also need to keep track of the group assignments in the total sample.

		Total Sample (Ordered Smallest to Largest)		Ranks	
Placebo	New Drug	Placebo	New Drug	Placebo	New Drug
7	3		1		1
5	6		2		2
6	4		3		3
4	2	4	4	4.5	4.5
12	1	5		6	
		6	6	7.5	7.5
		7		9	
		12		10	

Note that the lower ranks (e.g., 1, 2 and 3) are assigned to responses in the new drug group while the higher ranks (e.g., 9, 10) are assigned to responses in the placebo group. Again, the goal of the test is to determine whether the observed data support a difference in the populations of responses.

Recall that in parametric tests (discussed in the modules on hypothesis testing), when comparing means between two groups, we analyzed the difference in the sample means relative to their variability and summarized the sample information in a test statistic.

A similar approach is employed here. Specifically, we produce a test statistic based on the ranks.

First, we sum the ranks in each group. In the placebo group, the sum of the ranks is 37; in the new drug group, the sum of the ranks is 18. Recall that the sum of the ranks will always equal

n(n+1)/2. As a check on our assignment of ranks, we have n(n+1)/2 = 10(11)/2=55 which is equal to 37+18 = 55.

For the test, we call the placebo group 1 and the new drug group 2 (assignment of groups 1 and 2 is arbitrary).

We let R_1 denote the sum of the ranks in group 1 (i.e., R_1=37), and R_2 denote the sum of the ranks in group 2 (i.e., R_2=18). If the null hypothesis is true (i.e., if the two populations are equal), we expect R_1 and R_2 to be similar.

In this example, the lower values (lower ranks) are clustered in the new drug group (group 2), while the higher values (higher ranks) are clustered in the placebo group (group 1).

This is suggestive, but is the observed difference in the sums of the ranks simply due to chance? To answer this we will compute a test statistic to summarize the sample information and look up the corresponding value in a probability distribution.

Test Statistic for the Mann Whitney U Test

The test statistic for the Mann Whitney U Test is denoted **U** and is the ***smaller*** of U_1 and U_2, defined below.

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

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

where R_1 = sum of the ranks for group 1 and R_2 = sum of the ranks for group 2.

For this example,

$$U_1 = 5(5) + \frac{5(6)}{2} - 37 = 3$$

$$U_2 = 5(5) + \frac{5(6)}{2} - 18 = 22$$

Critical Values of the Mann-Whitney U
(Two-Tailed Testing)

n_2	α	n_1																	
		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	.05	--	0	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
	.01	--	0	0	0	0	0	0	0	0	1	1	1	2	2	2	2	3	3
4	.05	--	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14
	.01	--	--	0	0	0	1	1	2	2	3	3	4	5	5	6	6	7	8
5	.05	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
	.01	--	--	0	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6	.05	1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
	.01	--	0	1	2	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7	.05	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
	.01	--	0	1	3	4	6	7	9	10	12	13	15	16	18	19	21	22	24
8	.05	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
	.01	--	1	2	4	6	7	9	11	13	15	17	18	20	22	24	26	28	30
9	.05	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
	.01	0	1	3	5	7	9	11	13	16	18	20	22	24	27	29	31	33	36
10	.05	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
	.01	0	2	4	6	9	11	13	16	18	21	24	26	29	31	34	37	39	42
11	.05	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
	.01	0	2	5	7	10	13	16	18	21	24	27	30	33	36	39	42	45	48
12	.05	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
	.01	1	3	6	9	12	15	18	21	24	27	31	34	37	41	44	47	51	54
13	.05	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
	.01	1	3	7	10	13	17	20	24	27	31	34	38	42	45	49	53	56	60
14	.05	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
	.01	1	4	7	11	15	18	22	26	30	34	38	42	46	50	54	58	63	67
15	.05	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
	.01	2	5	8	12	16	20	24	29	33	37	42	46	51	55	60	64	69	73
16	.05	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
	.01	2	5	9	13	18	22	27	31	36	41	45	50	55	60	65	70	74	79
17	.05	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
	.01	2	6	10	15	19	24	29	34	39	44	49	54	60	65	70	75	81	86
18	.05	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
	.01	2	6	11	16	21	26	31	37	42	47	53	58	64	70	75	81	87	92
19	.05	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
	.01	3	7	12	17	22	28	33	39	45	51	56	63	69	74	81	87	93	99
20	.05	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127
	.01	3	8	13	18	24	30	36	42	48	54	60	67	73	79	86	92	99	105

n_2	α	n_1																	
		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	.05	0	0	1	2	2	3	4	4	5	5	6	7	7	8	9	9	10	11
	.01	--	0	0	0	0	0	1	1	1	2	2	2	3	3	4	4	4	5
4	.05	0	1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	17	18
	.01	--	--	0	1	1	2	3	3	4	5	5	6	7	7	8	9	9	10
5	.05	1	2	4	5	6	8	9	11	12	13	15	16	18	19	20	22	23	25
	.01	--	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
6	.05	2	3	5	7	8	10	12	14	16	17	19	21	23	25	26	28	30	32
	.01	--	1	2	3	4	6	7	8	9	11	12	13	15	16	18	19	20	22
7	.05	2	4	6	8	11	13	15	17	19	21	24	26	28	30	33	35	37	39
	.01	0	1	3	4	6	7	9	11	12	14	16	17	19	21	23	24	26	28
8	.05	3	5	8	10	13	15	18	20	23	26	28	31	33	36	39	41	44	47
	.01	0	2	4	6	7	9	11	13	15	17	20	22	24	26	28	30	32	34
9	.05	4	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48	51	54
	.01	1	3	5	7	9	11	14	16	18	21	23	26	28	31	33	36	38	40
10	.05	4	7	11	14	17	20	24	27	31	34	37	41	44	48	51	55	58	62
	.01	1	3	6	8	11	13	16	19	22	24	27	30	33	36	38	41	44	47
11	.05	5	8	12	16	19	23	27	31	34	38	42	46	50	54	57	61	65	69
	.01	1	4	7	9	12	15	18	22	25	28	31	34	37	41	44	47	50	53
12	.05	5	9	13	17	21	26	30	34	38	42	47	51	55	60	64	68	72	77
	.01	2	5	8	11	14	17	21	24	28	31	35	38	42	46	49	53	56	60
13	.05	6	10	15	19	24	28	33	37	42	47	51	56	61	65	70	75	80	84
	.01	2	5	9	12	16	20	23	27	31	35	39	43	47	51	55	59	63	67
14	.05	7	11	16	21	26	31	36	41	46	51	56	61	66	71	77	82	87	92
	.01	2	6	10	13	17	22	26	30	34	38	43	47	51	56	60	65	69	73
15	.05	7	12	18	23	28	33	39	44	50	55	61	66	72	77	83	88	94	100
	.01	3	7	11	15	19	24	28	33	37	42	47	51	56	61	66	70	75	80
16	.05	8	14	19	25	30	36	42	48	54	60	65	71	77	83	89	95	101	107
	.01	3	7	12	16	21	26	31	36	41	46	51	56	61	66	71	76	82	87
17	.05	9	15	20	26	33	39	45	51	57	64	70	77	83	89	96	102	109	115
	.01	4	8	13	18	23	28	33	38	44	49	55	60	66	71	77	82	88	93
18	.05	9	16	22	28	35	41	48	55	61	68	75	82	88	95	102	109	116	123
	.01	4	9	14	19	24	30	36	41	47	53	59	65	70	76	82	88	94	100
19	.05	10	17	23	30	37	44	51	58	65	72	80	87	94	101	109	116	123	130
	.01	4	9	15	20	26	32	38	44	50	56	63	69	75	82	88	94	101	107
20	.05	11	18	25	32	39	47	54	62	69	77	84	92	100	107	115	123	130	138
	.01	5	10	16	22	28	34	40	47	53	60	67	73	80	87	93	100	107	114

Example:

A new approach to prenatal care is proposed for pregnant women living in a rural community. The new program involves in-home visits during the course of pregnancy in addition to the usual or regularly scheduled visits.

A pilot randomized trial with 15 pregnant women is designed to evaluate whether women who participate in the program deliver healthier babies than women receiving usual care. The outcome is the **APGAR score** measured 5 minutes after birth.

Recall that APGAR scores range from 0 to 10 with scores of 7 or higher considered normal (healthy), 4-6 low and 0-3 critically low. The data are shown below.

Usual Care	8	7	6	2	5	8	7	3
New Program	9	9	7	8	10	9	6	

Is there statistical evidence of a difference in APGAR scores in women receiving the new and enhanced versus usual prenatal care? We run the test using the five-step approach.

- **Step 1.** Set up hypotheses and determine level of significance.

H_0: The two populations are equal versus

H_1: The two populations are not equal. $\alpha = 0.05$

- **Step 2.** Select the appropriate test statistic.

Because APGAR scores are not normally distributed and the samples are small ($n_1=8$ and $n_2=7$), we use the Mann Whitney U test. The test statistic is U, the smaller of

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \text{ and } U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

where R_1 and R_2 are the sums of the ranks in groups 1 and 2, respectively.

- **Step 3.** Set up decision rule.

The appropriate critical value can be found in the table above. To determine the appropriate critical value we need sample sizes ($n_1=8$ and $n_2=7$) and our two-sided level of significance ($\alpha=0.05$). The critical value for this test with $n_1=8$, $n_2=7$ and $\alpha = 0.05$ is 10 and the decision rule is as follows: Reject H_0 if $U \leq 10$.

- **Step 4.** Compute the test statistic.

The first step is to assign ranks of 1 through 15 to the smallest through largest values in the total sample, as follows:

		Total Sample (Ordered Smallest to Largest)		**Ranks**	
Usual Care	**New Program**	**Usual Care**	**New Program**	**Usual Care**	**New Program**
8	9	2		1	
7	8	3		2	
6	7	5		3	
2	8	6	6	4.5	4.5
5	10	7	7	7	7
8	9	7		7	
7	6	8	8	10.5	10.5
3		8	8	10.5	10.5
			9		13.5
			9		13.5
			10		15
				$R_1=45.5$	$R_2=74.5$

Next, we sum the ranks in each group. In the usual care group, the sum of the ranks is R_1=45.5 and in the new program group, the sum of the ranks is R_2=74.5. Recall that the sum of the ranks will always equal n(n+1)/2. As a check on our assignment of ranks, we have n(n+1)/2 = 15(16)/2=120 which is equal to 45.5+74.5 = 120.

We now compute U_1 and U_2, as follows:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 8(7) + \frac{8(9)}{2} - 45.5 = 46.5$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = 8(7) + \frac{7(8)}{2} - 74.5 = 9.5$$

Thus, the test statistic is U=9.5.

- **Step 5.** Conclusion:

We reject H_0 because 9.5 $\leq$ 10. We have statistically significant evidence at α =0.05 to show that the populations of APGAR scores are not equal in women receiving usual prenatal care as compared to the new program of prenatal care.

Contingency Table: Overview

- Contingency tables (also called crosstabs or two-way tables) summarize the relationship between several categorical variables. It is a special type of **frequency distribution table**, where two variables are shown simultaneously.
- For example, a researcher might be investigating the relationship between AIDS and sexual preference. The two variables would be AIDS and SEXUAL PREFERENCE. If the question is "Is there a significant relationship between AIDS and sexual preference?", then a chi-square test could then be run on the table to determine if there is a relationship between the two variables.
- The following contingency table shows exposure to a potential source of foodborne illness (in this case, ice-cream). From the table, you can see that 13 people in a case study ate ice cream; 17 people did not:

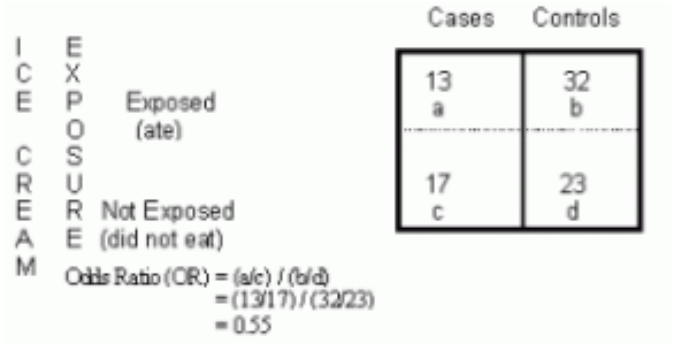


Image: Michigan Dept. of Agriculture

Chi-Square Tests

A chi^2 test can be conducted on contingency tables to test whether or not a relationship exists between variables. These effects are defined as relationships between rows and columns. The chi^2 test:

$$c^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

- Where "O" is the Observed value, "E" is the expected value and "i" is the "ith" position in the table. The sigma (Σ) is the summation symbol.
- The following picture shows what your contingency table might look like with your data, plus the results from running a chi^2 test on your data. A **small chi^2 value** means that there is little relationship between the categorical variables.
- A **large chi^2 value** means that there is a definite correlation between the two variables. As there is strong evidence that sexual orientation is linked to a higher risk of contracting AIDS, it's no surprise that the chi^2 value is high:

AIDS * SEXPREF Crosstabulation

Count

		SEXPREF			
		Males	Females	Both	Total
AIDS	Yes	4	2	3	9
	No	3	16	2	21
Total		7	18	5	30

Chi-Square Tests

	Value	df	Asymp. Sig. (2-tailed)
Pearson Chi-Square	7.657[a]	2	.022
Likelihood Ratio	7.803	2	.020
Linear-by-Linear Association	.062	1	.803
N of Valid Cases	30		

a. 4 cells (66.7%) have expected count less than 5. The minimum expected count is 1.50.

However, the note under the results states that "4 cells (66.7%) have expected count less than 5." Generally, if this is over 25%, the result could be due to chance alone. Therefore, the results from this particular test are **not statistically significant**.

Fisher's Exact Test

Fisher's exact test is a statistical test used to determine if there are non-random associations between two categorical variables.

Let there exist two such variables X and Y, with m and n observed states, respectively. Now form an m×n matrix in which the entries a_(ij) represent the number of observations in which x=i and y=j. Calculate the row and column sums R_i and C_j, respectively, and the total sum

$$N = \sum_i R_i = \sum_j C_j$$

of the matrix. Then calculate the conditional probability of getting the actual matrix given the particular row and column sums, given by

$$P_{\text{cutoff}} = \frac{(R_1!\,R_2!\cdots R_m!)(C_1!\,C_2!\cdots C_n!)}{N!\prod_{i,j} a_{ij}!},$$

which is a multivariate generalization of the hypergeometric probability function. Now find all possible matrices of nonnegative integers consistent with the row and column sums R_i and C_j. For each one, calculate the associated conditional probability using (2), where the sum of these probabilities must be 1.

- To compute the P-value of the test, the tables must then be ordered by some criterion that measures dependence, and those tables that represent equal or greater deviation from independence than the observed table are the ones whose probabilities are added together. There are a variety of criteria that can be used to measure dependence.
- In the 2×2 case, which is the one Fisher looked at when he developed the exact test, either the Pearson chi-square or the difference in proportions (which are equivalent) is typically used. Other measures of association, such as the likelihood-ratio-test, G-squared, or any of the other measures typically used for association in contingency tables, can also be used.
- The test is most commonly applied to 2×2 matrices, and is computationally unwieldy for large m or n. For tables larger than 2×2, the difference in proportion can no longer be used, but the other measures mentioned above remain applicable (and in practice, the Pearson statistic is most often used to order the tables).
- In the case of the 2×2 matrix, the P-value of the test can be simply computed by the sum of all P-values which are <=P_(cutoff).
- For an example application of the 2×2 test, let X be a journal, say either Mathematics Magazine or Science, and let Y be the number of articles on the topics of mathematics and biology appearing in a given issue of one of these journals.
- If Mathematics Magazine has five articles on math and one on biology, and Science has none on math and four on biology, then the relevant matrix would be

	Math. Mag.	Science	
math	5	0	$R_1 = 5$
biology	1	4	$R_2 = 5$
	$C_1 = 6$	$C_2 = 4$	$N = 10.$

Computing P_{cutoff} gives

$$P_{\text{cutoff}} = \frac{5!^2\,6!\,4!}{10!\,(5!\,0!\,1!\,4!)} = 0.0238,$$

and the other possible matrices and their Ps are

$$\begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} P = 0.2381$$

$$\begin{bmatrix} 3 & 2 \\ 3 & 2 \end{bmatrix} P = 0.4762$$

$$\begin{bmatrix} 2 & 3 \\ 4 & 1 \end{bmatrix} P = 0.2381$$

$$\begin{bmatrix} 1 & 4 \\ 5 & 0 \end{bmatrix} P = 0.0238,$$

which indeed sum to 1, as required. The sum of P-values less than or equal to $P_{(cutoff)}=0.0238$ is then 0.0476 which, because it is less than 0.05, is significant. Therefore, in this case, there would be a statistically significant association between the journal and type of article appearing.