STK121 - Chapter 18

Learning outcomes:

⇒ Know what the difference in approach between parametric and nonparametric tests are.
⇒ Be able to understand, explain and execute the hypothesis test about a single population median.
⇒ Be able to calculate a p-value by hand and by using EXCEL's BINOM.DIST function.
⇒ Be able to calculate a p-value for larger sample sizes, where the normal distribution approximation of the binomial distribution is used in hand calculations and by using EXCEL's NORM.S.DIST and NORM.DIST functions.
⇒ Know what the requirements for the sample size is in order to do a normal approximation.
⇒ Be able to apply the correction factors when approximating the discrete binomial distribution with the continuous normal distribution

2.4 - Non-Parametric Methods

Introduction:

- Most of the statistical methods referred to as parametric require the use of interval- or ratio-scaled data.
- Nonparametric methods are often the only way to analyse categorical (nominal or ordinal) data and draw statistical conclusions.
- Nonparametric methods require no assumptions about the population probability distributions
- Nonparametric methods are often called distribution-free methods.
- Whenever the data are quantitative, we will transform the data into categorical data in order to conduct the nonparametric test.

Previous Chapters	This chapter
⇒ All previous testing methods regarding the unknown population parameters μ, σ and P were conducted using sample data from the population to calculate the sample parameters which is used as estimates. ⇒ The assumption most often made is the one regarding the population`s distribution form being known and Normal. ⇒ Tests like these are referred to as **parametric tests** since the **test procedures are performed with known parameters** and **distributions characteristics**. ⇒ **Data** is usually **quantitative**	⇒ In this chapter testing procedures and measures will be discussed which can be used **without having** any **knowledge** regarding the **form and distribution of the population from which the data is sampled.** ⇒ Tests like these are referred to as **nonparametric or distribution-free tests**. ⇒ **Data** can be **quantitative or qualitative (ordinal or categorical).** ⇒ The **median** is rather **used** instead of the average, since it is not influenced by outliers in the dataset and the **ranks of the data** instead of the actual measurements **are used.**

Links which will also be helpful:

- https://www.youtube.com/watch?v=xA0QcbNxENs
- https://youtu.be/pWEWHKnwg_0

There are four Non-Parametric methods

1. Sign test (2.4.1)
2. Mann-Whitney-Wilcoxon test (2.4.2)
3. Kruskal-Wallis Test (2.4.3)
4. Spearman rank-correlation coefficient (2.4.4)

2.4.1 – Sign Test

<u>Learning outcomes:</u>

⇒ Be able to understand and explain the hypothesis test about a population median.
⇒ Be able to calculate a p-value by using EXCEL's BINOM.DIST function.
⇒ Be able to calculate a p-value for larger sample sizes, where the normal distribution approximation of the binomial distribution is used and EXCEL's NORM.S.DIST and NORM.DIST functions

The sign Test

⇒ The sign test is a versatile method for hypothesis testing that uses the binomial distribution with **p = 0.50** as the sampling distribution.
⇒ We present two applications of the sign test:

1. **A hypothesis test about a population median**
2. **A matched-sample test about the difference between two populations**

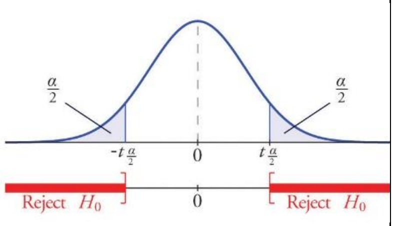
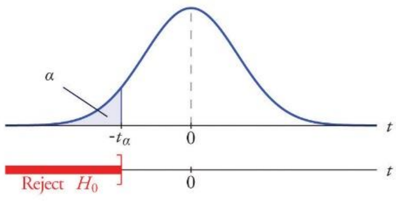
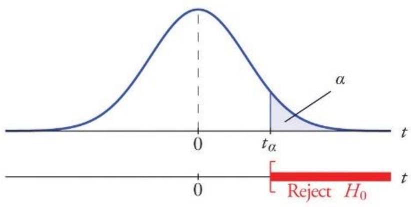
A hypothesis test about a population median

We can apply the sign test by:

+	Using a **plus sign** whenever the data in the sample are above the hypothesized value of the **median**
−	Using a **minus sign** whenever the data in the sample are below the hypothesized value of the **median**
DISREGARD	**Discarding** any data **exactly equal** to the **hypothesized median & your sample size becomes one observation less**

❖ We get two sample cases – small and large
❖ Binomial distribution = sample test small
❖ The sample size = number of trials
❖ There are two outcomes possible per trial:
   1. A plus sign
   2. Minus sign
❖ The trials are independent.
❖ We let **P = probability of a plus sign.**
❖ If the population median is in fact a particular value, p should equal 0.50

	Small sample case	Large sample case
Notes	• Binomial distribution	• Normal distribution approximation of the binomial distribution to compute the p-value
Distribution form	• Whenever n $\leq$ 20.	• Approximately normal for **n > 20**
Hypothesis	• $H0$: p = 0.50 (The population median equals the value assumed) • Ha: $p \neq$ 0.50 (The population median is different than the value assumed)	Number of Plus Signs when $\Rightarrow$ H0: **p = 0.50**
Test statistic	• The number of plus signs is our test statistic	
Rejection rule	• Assuming H0 is true, the sampling distribution for the test statistic is a binomial distribution with p =0.50 • H0 is rejected if the p-value < level of significance, alpha.	
Mean		• **μ = 0.50n**
Standard deviation		• σ **= 0.25**n

Test	Two - sided	Left sided	Right-sided
Ha:	**Median = value**	**Median $\geq$ value**	**Median $\leq$ Value**
Ho:	**Median $\neq$ value**	**Median < value**	**Median > value**
			

My own steps:

1) Identify whether if it's a large (binomial) or small (approximate normal distribution) case test - n $\leq$ 20 or n > 20 or $n \geq 20$
2) Count the plus and minus signs
3) Hypothesis test for POPULATION MEDIAN
4) State Test statistic
5) State Decision rule
6) Determine P-value
7) Rejection rule
8) Conclusion

Calculation for probabilities:

$$P(X = X) = \binom{n}{x} P^x (1-p)^{n-x}$$

Example 1: Small-Sample Case

Lawler's Grocery Store made the decision to carry Cape May Potato Chips based on the manufacturer's estimate that the **median sales should be $450** per week on a per-store basis. Lawler's has been carrying the potato chips for three months. Data showing one-week sales at 10 randomly selected Lawler's stores.

We want to test the statement that the median sales per week at all stores is $450

One week's sale at 10 Lawler grocery stores:

n	Store number	Weekly sales ($)	Sign
1	**56**	**485**	+
2	**19**	**562**	+
3	**36**	**415**	-
4	**128**	**860**	+
5	**12**	**426**	-
6	**63**	**474**	+
7	**39**	**662**	+
8	**84**	**380**	-
9	**102**	**515**	+
10	**44**	**721**	+

Step 1: Determine if it's a large or small test

Small sample test - n ≤ 20

⇒ n = (7+3) 10

Step 2: Count signs

- ⇒ The population median is $450 the 50% of the signs should be positive and 50% of the signs must be minus
- ⇒ Plus sign = 7
- ⇒ Minus sign = 3
- ⇒ n = 10
- ⇒ p = 0.50 and 1 – p = 0.50 (Binomial)

Step 3: Hypothesis test for POPULATION MEDIAN

- ⇒ Ho: Median = 450
- ⇒ Ha: Median $\neq$ 450

Hypothesis test can be converted into binomial distribution's probability p

- ⇒ Ho: Median = 0.50
- ⇒ Ha: Median $\neq$ 0.50

Step 4: Test statistic

- ⇒ Number of plus signs
- ⇒ Test stat = 7

Step 5: Decision rule

- ⇒ Reject Ho if p-value $\leq \alpha$
- ⇒ Do not reject Ho is P-value > α

Step 6: Determine P-value

- ⇒ Use Binomial table with n = 10, p =0.50 to get the P-value
- ⇒ P -value = the probability to get a value of 7 or more plus signs
- ⇒ Because the observed number of plus signs is 7, we begin by computing the probability of obtaining 7 or more plus signs.
- ⇒ The probability of **7, 8, 9, or 10** plus signs is: **0.1172 + 0.0439 + 0.0098 + 0.0010** = 0.1719
- ⇒ P-value = 2(0.1719) = 0.3438 – because it's a two-tailed test
- ⇒ P-value = 0.3438 > 0.05

Step 7: Rejection rule

- ⇒ Do not reject Ho is P-value > α
- ⇒ P-value = 0.3438 > 0.05
- ⇒ Do not reject Ho

Step 8: Conclusion

The statement that the median sales per week per store is $450 cannot be rejected. There is insufficient evidence in the sample to reject the assumption that the median weekly sales are $450.

Number of plus signs	Probability calculation	Probability
0	$P(X=0)=\binom{10}{0}(0.50)^{0}(1-0.50)^{10-0}$	**0.0010**
1	$P(X=1)=\binom{10}{1}(0.50)^{1}(1-0.50)^{10-1}$	**0.0098**
2	$P(X=2)=\binom{10}{2}(0.50)^{2}(1-0.50)^{10-2}$	**0.0439**
3	$P(X=3)=\binom{10}{3}(0.50)^{3}(1-0.50)^{10-3}$	**0.1172**
4	$P(X=4)=\binom{10}{4}(0.50)^{4}(1-0.50)^{10-4}$	**0.2051**
5	$P(X=5)=\binom{10}{5}(0.50)^{5}(1-0.50)^{10-5}$	**0.2461**
6	$P(X=6)=\binom{10}{6}(0.50)^{6}(1-0.50)^{10-6}$	**0.2051**
7	$P(X=7)=\binom{10}{7}(0.50)^{7}(1-0.50)^{10-7}$	**0.1172**
8	$P(X=8)=\binom{10}{8}(0.50)^{8}(1-0.50)^{10-8}$	**0.0439**
9	$P(X=9)=\binom{10}{9}(0.50)^{9}(1-0.50)^{10-9}$	**0.0098**
10	$P(X=10)=\binom{10}{10}(0.50)^{10}(1-0.50)^{10-10}$	**0.0010**

Example 2: Large sample size

- ⇒ H0: p = 0.50
- ⇒ Mean: μ = 0.50n
- ⇒ Standard Deviation: $\sigma = 0.25n$
- ⇒ Distribution Form: Approximately normal for n > 20
- ⇒ Identify whether if it's a large (binomial) or small (approximate normal distribution) case test - n ≤ 20 or n > 20 or $n \geq 20$

Step 1: Determine if it's a large or small test

- ⇒ The media sales must be more than \$450

Step 2: Count signs

- ⇒ Number of plus signs
- ⇒ Test stat = 7

Step 3: Hypothesis test for POPULATION MEDIAN

- ⇒ H0: Median ≤ $450 — H0: Median ≤ 0.50
- ⇒ Ha: Median > $450 — Ha: Median > 0.50

Step 4: Test statistic

- ⇒ Number of plus signs
- ⇒ Test stat = 7

Step 5: Decision rule

- ⇒ Reject Ho if p-value ≤ α = 0.05
- ⇒ Do not reject Ho is P-value > α = 0.05

Step 6: Determine P-value

- ⇒ P-value = probability to get a value of 7 or more plus signs
- ⇒ P-value = **0.1172 + 0.0439 + 0.0098 + 0.0010** = 0.1719
- ⇒ P-value = 0.1719 > 0.05
- ⇒ H0 cannot be rejected

Step 7: Rejection rule

- ⇒ P-value = 0.1719 > 0.05
- ⇒ H0 cannot be rejected

Step 8: Conclusion

- ⇒ Median sales are **not more** than $450

Example 3: Large sample size

Step 1: Determine if it's a large or small test

- ⇒ The media sales must be less than $450

Step 2: Count signs

- ⇒ Number of plus signs
- ⇒ Test stat = 7

Step 3: Hypothesis test for POPULATION MEDIAN

- ⇒ H0: Median ≥ $450 — H0: Median ≥ 0.50
- ⇒ Ha: Median < $450 — Ha: Median < 0.50

Step 4: Test statistic

- ⇒ Number of plus signs
- ⇒ Test stat = 7

Step 5: Decision rule

- ⇒ Reject Ho if p-value ≤ α = 0.05
- ⇒ Do not reject Ho is P-value > α = 0.05

Step 6: Determine P-value

- ⇒ P-value = probability to get a value of 7 or fewer plus signs
- ⇒ P-value = 1 - **0.0439** + **0.0098** + **0.0010** = 0.9453
- ⇒ P-value = 0.9243 > 0.05
- ⇒ H0 cannot be rejected

Step 7: Rejection rule

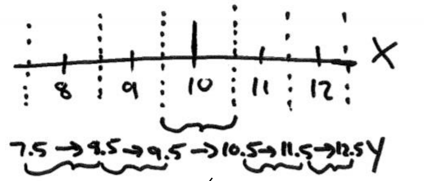
- ⇒ P-value = 0.9243 > 0.05
- ⇒ H0 cannot be rejected

Step 8: Conclusion

- ⇒ Median sales are NOT less than $450

How to determine Binomial probabilities

- ⇒ Determining binomial probabilities for large samples is time consuming
- ⇒ For large sample (20+)
- ⇒ The NORMAL APPROXIMATION of the Binomial can be used with
- ⇒ $\mu = np = 0.50(\text{n})$
- ⇒ $\sigma^2 = np(1-n) = 0.25\text{n}$
- ⇒ $\sigma^2 = \sqrt{np(1-p)} = \sqrt{0.25n}$
- ⇒ Binomial = discrete distribution
- ⇒ Normal = continuous distribution
- ⇒ Continuity correction factor is needed for conversion
- ⇒ Continuity correction factor = $P(X = 10) = P(9.50 \leq Y \leq 10.50)$
- ⇒ $10 - 0.50 = 9.50$ $\quad$ $10+0.50 = 10.50$



Example 4:

A year ago, the median price of a new home was $236000. It is suspected that the economic downturn affected prices negatively. Real estate firms use sample data to see if the current median price of a new home is LESS than a year ago. A sample size of 61 recent new homes sold yielded 22 homes that sold for more than $236000. 38 homes that sold for less than $236000 and 1 sold for $236000. Use alpha = 0.05

Step 1: Determine if it's a large or small test

- ⇒ After deleting the home that sold for $236000 the test will be performed on 60 homes

Step 2: Count signs

- ⇒ 22 plus signs (+)
- ⇒ 38 minus signs (-)

Step 3: Hypothesis test for POPULATION MEDIAN

⇒ H0: Median ≥ $236000	H0: Median ≥ 0.50
⇒ Ha: Median < $236000	Ha: Median < 0.50

Step 4: Determine the median and standard deviation

⇒ $\mu = np$ = 0.50(n) = 0.50(60) = 30 plus (+) signs

⇒ $\sigma^2 = np(1-n)$ = 0.25n = 0.25(60) = 15

⇒ $\sigma^2 = \sqrt{np(1-p)} = \sqrt{0.25n} = \sqrt{15}$

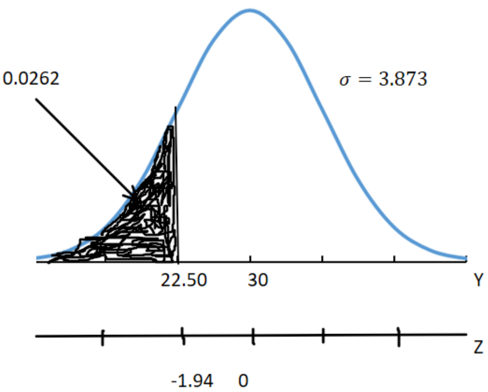
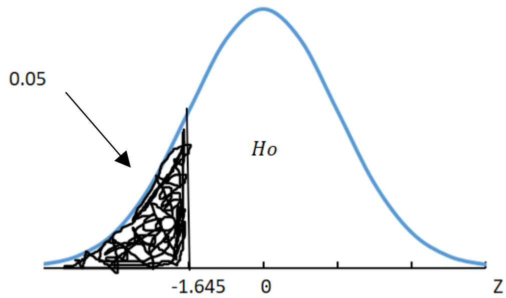
⇒ S = 3.8729

Step 5: Decision rule

⇒ Reject Ho if p-value ≤ α = 0.05

⇒ Do not reject Ho is P-value > α = 0.05

Step 6: Determine P-value or critical value

P – value	Critical Value
⇒ X = the number of plus signs ⇒ Use the normal distribution to approximate the binomial probability P(X < 22). ⇒ The binomial probability of 22 is computed by the normal probability interval 21.5 to 22.5 ⇒ $P(21.5 < X < 22.5)$ ⇒ Z = (x – μ)/s ⇒ $Convert\ Y\ to\ Z$ ⇒ $P\left(\frac{21.50-30}{3.873}\right) < z < P\left(\frac{22.5-30}{3.873}\right)$ ⇒ $P(-2.19 < z < 1.94)$ ⇒ P-value = -1.93648 – (- 2.19468) = 0.0262 ⇒ P-value = 0.0262 	⇒ Z = (x – μ)/s ⇒ = (22.5 – 30) / 3.873 = - 1.94 ⇒ The test statistic = -1.94 ⇒ Critical value = -1.645 ⇒ Reject Ho if Z < - 1.645 ⇒ (- 1.94 < - 1.645) ⇒ Reject Ho 

Step 7: Rejection rule

P - Value	Critical Value
⇒ Using 0.05 level of significance ⇒ Reject H0 if p-value < 0.05 ⇒ 0.0262 < 0.05 ⇒ Reject H0	⇒ The test statistic = -1.94 ⇒ Critical value = -1.645 ⇒ Reject Ho if Z < - 1.645 ⇒ (- 1.94 < - 1.645) ⇒ Reject Ho

Step 8: Conclusion

⇒ Median price of current new homes is less than the $236,000 median price a year ago.

YouTube videos:

1. https://www.youtube.com/watch?v=1bkuW0EJGKs
2. https://www.youtube.com/watch?v=eqZUI18V_dg
3. https://www.youtube.com/watch?v=Oiu9ymGuCVA
4. https://www.youtube.com/watch?v=heobUSjs72c