Non-Parametric Statistics

Tom Bruning

2018-04-16

Non-parametric Statistics

Introduction

Parametric statistics describe procedures that assume that the population from which the sample is drawn follows a known probability distribution

- Parametric statistics require the level of data measurement to be interval or ratio

Parametric procedures were used to form confidence intervals and test hypotheses we have previously studied.

Nonparametric statistics rely on fewer assumptions about the population probability distribution

- Used for nominal or ordinal level data
- Used when the sample size is very small (less than 30) because it is difficult to test if the population is normally distributed
- Used when there is no corresponding parametric procedure for the hypothesis test

All else equal (the same sample size, same level of significance, and so on), a parametric procedure will result in a higher probability of correctly rejecting a null hypothesis

The Sign Test

The sign test is a very versatile nonparametric procedure that can be used to estimate a population median

The sign test assigns a plus sign or a minus sign to each observation in a sample

- the test statistic is the number of plus or minus signs from the sample

Using a One-Tail Sign Test for the Median (Small Sample Case)

The small sample case is used when n $\leq$ 20

Example: United Medical wants to see if the number of steps walked per day by its population of employees is less than 6,000

- Data were collected from a random sample of ten employees

Table 1: Randon Sample of Number of step taken per day

2890	3760	5217	9736	6000
5799	4649	15758	4127	5105

- **Step 1:** Identify the null and alternative hypotheses
  - H_0: Median $\geq$ 6,000 steps
  - H_1: Median $<$ 6,000 steps
    Suppose that $\alpha = 0.10$ is chosen for this test
- **Step 2:** Count the plus and minus signs
  - assign a plus sign to any data value above the hypothesized median (6,000) and a negative sign to any data value below the median
  - If any value happens to fall exactly on the median, remove it from the sample

Table 2: Randon Sample of Number of step taken per day

2890(-)	3760(-)	5217(-)	9736(+)	6000( )
5799(-)	4649(-)	15758(+)	4127(-)	5105(-)

There are two plus signs and seven minus signs

- **Step 3:** Determine the test statistic, S
  - The test statistic for the one-tail sign test, S, is the number of plus signs or minus signs, depending on the alternative hypothesis
  - Because this alternative hypothesis is that the median is less than 6,000 steps per day, the test statistic is the number of minus signs
  - Therefore S = 7
- **Step 4:** Determine the sample size, n
  - The sample size is the total number of plus signs and minus signs
  - here, n = 7 + 2 = 9
- **Step 5:** Determine the p-value
  - If the null hypothesis is true the probability that an observation is assigned a plus sign (above the median of 6,000) equals 0.50

- The same can be said for the probability that the observation is assigned a minus sign
- This sampling distribution can be said to follow the binomial probability distribution with p = 0.50
- Find the binomial probabilities for all possible values of x (the number of minus signs) when p = 0.50 and n = 9
- In our sample of nine observations, seven were below the hypothesized median of 6,000 steps per day.
- The p-value for this hypothesis test represents the probability of observing seven or more observations below the median if the null hypothesis is true

Table 3: Binomial Probability Distribution when p = 0.50 and n = 9

Number of minus signs. x	Probability
0	0.0020
1	0.0176
2	0.0703
3	0.1641
4	0.2461
5	0.2461
6	0.1641
7	0.0703
8	0.0176
9	0.0020

$$p - value = P(x \geq S) \tag{1}$$

$$P(x\ 7) = 0.0703 + 0.0176 + 0.0020 = 0.0899 \tag{2}$$

- **Step 6:** State the conclusion

Because the p-value (0.0899) is less than $\alpha = 0.10$, United Medical can reject the null hypothesis and conclude that the median number of steps walked per day by the population of employees is less than 6,000

- Decision Rule for hypotheses tests using p-values
  p-value $\geq \alpha \rightarrow$ Do not reject H_0
  p-value $< \alpha \rightarrow$ Reject H_0

Using the Sign Test to Determine Preferences

- The sign test can be applied to tests of whether consumers prefer one product or service to another
- In this application, the data values are the brand names consumers prefer when faced with two choices
- Because the data (brand names) are merely labels, the level of measurement is nominal

Example: Brand preference data (n = 14)
Do more people prefer Cheez-It brand?

Table 4: Sample Results for Cheez-Its vs. Cheese Nips

Person	Preference	Sign	Person	Preference	Sign
1	Cheez-It	+	8	No Preference	
2	Cheez-It	+	9	Cheez-It	+
3	Cheez-It	+	10	Cheese Nips	-
4	Cheez-It	+	11	Cheez-It	+
5	Cheese Nips	-	12	Cheez-It	+
6	Cheez-It	+	13	Cheez-It	+
7	Cheez-It	+	14	Cheez-It	+

Step 1: Identify the null and alternative hypotheses - $H_0 : p \leq 0.50$ (a majority of the population does not favor Cheez-It)
- $H_0 : p > 0.50$ (a majority of the population does favor Cheez-It)
Suppose that $alpha = 0.05$ is chosen for this test
Step 2: Count the plus and minus signs

- For this sample, we have 11 plus signs and 2 minus signs
Step 3: Determine the test statistic, S
We want to support that the preferred brand is Cheez-It, those choosing Cheez-Its were assigned a plus sign. So, S = 11
Step 4: Determine the sample size, n
n = the total number of plus signs and minus signs, n = 11 + 2 = 13

Step 5: Determine the p-value The p-value is found using the binomial probabilities for p = 0.50 and n = 13

Table 5: Binomial Probability Distribution when p = 0.50 and n = 13

Number of plus signs. x	Probability	Number of plus signs. x	Probability
0	0.0001	7	0.0295

Number of plus signs. x	Probability	Number of plus signs. x	Probability
1	0.0016	8	0.1571
2	0.0095	9	0.0873
3	0.0349	10	0.0349
4	0.0873	11	0.0095
5	0.1571	12	0.0016
6	0.2095	13	0.0001

p-value $= P(x \geq S) = P(x \geq 11) = 0.0095 + 0.0016 + 0,0001 = 0.0112$

Step 6: State the conclusion

Because the p-value (0.0112) is less than $\alpha = 0.05$, we reject the null hypothesis and conclude that the majority of people prefer Cheez-Its.