Nominal Data

Greg C Elvers

1

Parametric Statistics

- The inferential statistics that we have discussed, such as t and ANOVA, are *parametric statistics*
- A parametric statistic is a statistic that makes certain assumptions about how the data are distributed
  - Typically, they assume that the data are distributed normally

2

Nonparametric Statistics

- *Nonparametric statistics* do not make assumptions about the underlying distribution of the data
- Thus, nonparametric statistics are useful when the data are not normally distributed
- Because nominally scaled variables cannot be normally distributed, nonparametric statistics should be used with them

3

Parametric vs Nonparametric Tests

- When you have a choice, you should use parametric statistics because they have greater statistical power than the corresponding nonparametric tests
  - That is, parametric statistics are more likely to correctly reject H_0 than nonparametric statistics

4

Binomial Test

- The *binomial test* is a type of nonparametric statistic
- The binomial test is used when the DV is nominal, and it has only two categories or classes
- It is used to answer the question:
  - In a sample, is the proportion of observations in one category different than a given proportion?

Binomial Test

- A researcher wants to know if the proportion of ailurophiles in a group of 20 librarians is greater than that found in the general population, .40
  - There are 9 ailurophiles in the group of 20 librarians

Binomial Test

- Write H_0 and H_1:
  - H_0: $P \leq .40$
  - H_1: $P > .40$
- Is the hypothesis one-tailed or two-tailed?
  - Directional, one-tailed
- Determine the statistical test
  - The librarians can either be or not be ailurophiles, thus we have a dichotomous, nominally scaled variable
  - Use the binomial test

Binomial Test

- Determine the critical value from a table of critical binomial values
  - Find the column that corresponds to the p value (in this case .40)
  - Find the row that corresponds to the sample size (N = 20) and α (.05)
  - The critical value is 13

Binomial Test

- If the observed number of ailurophiles (9) is greater than or equal to the critical value (13), you can reject H_0
  - We fail to reject H_0; there is insufficient evidence to conclude that the percentage of librarians who are ailurophiles is probably greater than that of the general population

Normal Approximation to the Binomial Test

- When the sample size is greater than or equal to 50, then a normal approximation (i.e. a z-test) can be used in place of the binomial test
- When the product of the sample size (N), p, and 1 - p is greater than or equal to 9, then the normal approximation can be use

Normal Approximation to the Binomial Test

- The normal approximation to the binomial test is defined as:

$$z = \frac{x - NP}{\sqrt{NP(1-P)}}$$

- x = number of observations in the category
- N = sample size
- P = probability in question

Normal Approximation to the Binomial Test

- A researcher wants to know if the proportion of ailurophiles in a group of 100 librarians is greater than that found in the general population, .40
  - There are 43 ailurophiles in the group of 100 librarians

Normal Approximation to the Binomial Test

- Write H_0 and H_1:
  - H_0: $P \leq .40$
  - H_1: $P > .40$
- Is the hypothesis one-tailed or two-tailed?
  - Directional, one-tailed
- Determine the statistical test
  - The librarians can either be or not be ailurophiles, thus we have a dichotomous, nominally scaled variable
  - Use the z test, because $n \geq 50$

13

Normal Approximation to the Binomial Test

- Calculate the z-score

$$z = \frac{x - NP}{\sqrt{NP(1-P)}}$$
$$= \frac{43 - 100 \cdot .40}{\sqrt{100 \cdot .40 \cdot (1 - .40)}}$$
$$= \frac{3}{4.899}$$
$$= 0.612$$

14

Normal Approximation to the Binomial Test

- Determine the critical value from a table of area under the normal curve
  - Find the z-score that corresponds to an area of .05 above the z-score
  - That value is 1.65
- Compare the calculated z-score to the critical z-score
  - If $|z_{calculated}| \geq z_{critical}$, then reject H_0
  - $0.612 < 1.65$; fail to reject H_0

15

χ^2 -- One Variable

- When you have nominal data that has more than two categories, the binomial test is not appropriate
- The χ^2 (chi squared) test is appropriate in such instances
- The χ^2 test answers the following question:
  - Is the observed number of items in each category different from a theoretically expected number of observations in the categories?

16

χ^2 -- One Variable

- At a recent GRE test, each of 28 students took one of 5 subject tests
- Was there an equal number of test takers for each test?

Test	Psych	Math	Bio	Lit	Engin
Obs.	12	2	4	6	4
Exp.	5.6	5.6	5.6	5.6	5.6

17

χ^2 -- One Variable

- Write H_0 and H_1:
  - H_0: $\Sigma(O - E)^2 = 0$
  - H_1: $\Sigma(O - E)^2 \neq 0$
    - O = observed frequencies
    - E = expected frequencies
- Specify α
  - $\alpha = .05$
- Calculate the χ^2 statistic
- $\chi^2=\Sigma[(O_i-E_i)^2/E_i]$

18

χ^2 Calculations

	Psy	Math	Bio	Lit	Engin
O_i	12	2	4	6	4
E_i	5.6	5.6	5.6	5.6	5.6
O_i-E_i	6.4	-3.6	-1.6	.4	-1.6
$(O_i-E_i)^2$	40.96	12.96	2.56	1.6	2.56
$(O_i-E_i)^2/E_i$	7.31	2.31	0.46	0.29	0.46

19

χ^2 Calculations

- $\chi^2=\Sigma[(O_i-E_i)^2/E_i]$
- $\chi^2=7.31+2.31+0.46+0.29+0.46=10.83$
- Calculate the degrees of freedom:
  - df = number of groups - 1 = 5 - 1 = 4
- Determine the critical value from a table of critical χ^2 values
  - df = 4, $\alpha = .05$
  - Critical $\chi^2_{\alpha=.05}(4) = 9.488$

20

χ^2 Decision

- If the observed / calculated value of χ^2 is greater than or equal to the critical value of χ^2, then you can reject H_0 that there is no difference between the observed and expected frequencies
  - Because the observed χ^2 = 10.83 is larger than the critical χ^2 =9.488, we can reject H_0 that the observed and expected frequencies are the same

21

χ^2 Test of Independence

- χ^2 can also be used to determine if two variables are independent of each other
  - E.g., is being an ailurophile independent of whether you are male or female?
- Write H_0 and H_1:
  - H_0: $\Sigma\Sigma(O - E)^2 = 0$
  - H_1: $\Sigma\Sigma(O - E)^2 \neq 0$
- Specify α
  - α=.05

22

χ^2 Test of Independence

- The procedure for answering such questions is virtually identical to the one variable χ^2 procedure, except that we have no theoretical basis for the expected frequencies
  - The expected frequencies are derived from the data

23

χ^2 Test of Independence

	Male	Female	Total
Ailurophile	24	37	61
Non-ailurophile	12	7	19
Total	36	44	80

- The expected frequencies are given by the formula to the right:

$$E_{ij} = \frac{r_i c_j}{T}$$

E_{ij} = expected frequency for cell at row i and column j

r_i = total for row i

c_j = total for column j

T = total number of observations

24

χ^2 Test of Independence

	Male	Female	Total
Ailurophile	O_{11}=24 E_{11}=(61*36)/80=27.45	O_{12}=37 E_{12}=(61*44)/80=33.55	r_1=61
Non-ailurophile	O_{21}=12 E_{21}=(19*36)/80=8.55	O_{22}=7 E_{22}=(19*44)/80=10.45	r_2=19
Total	c_1=36	c_2=44	T=80

χ^2 Test of Independence

- Calculate the observed value of χ^2

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

$$= \frac{(24-27.45)^2}{27.45}+\frac{(37-33.55)^2}{33.55}+\frac{(12-8.55)^2}{8.55}+\frac{(7-10.45)^2}{10.45}$$

$$= 0.434 + 0.355 + 1.392 + 1.139$$

$$= 3.319$$

χ^2 Test of Independence

- First, determine the degrees of freedom:
  - df = (r - 1) * (c - 1)
  - In this example, the number of rows (r) is 2, and the number of columns (c) is 2, so the degrees of freedom are (2 - 1) * (2 - 1) = 1
  - Determine the critical value of χ^2 from a table of critical χ^2 values
  - Critical $\chi^2_{\alpha=.05}(1)$=3.841

χ^2 Test of Independence

- Make the decision
  - If the observed /calculated value of χ^2 is greater than or equal to the critical value of χ^2, then you can reject H_0 that the expected and observed frequencies are equal
  - If this example, the observed χ^2 = 3.319 is not greater than or equal to the critical χ^2 = 3.841, so we fail to reject H_0

Requirements for the Use of χ^2

- Even though χ^2 makes no assumptions about the underlying distribution, it does make some assumptions that needs to be met prior to use
  - Assumption of independence
  - Frequencies must be used, not percentages
  - Sufficiently large sample size

Assumption of Independence

- Each observation must be unique; that is an individual cannot be contained in more than one category, or counted in one category more than once
- When this assumption is violated, the probability of making a Type-I error is greatly enhanced

Frequencies

- The data must correspond to frequencies in the categories; percentages are not appropriate as data

Sufficient Sample Size

- Different people have different recommendation about how large the sample should be, and what the minimum expected frequency in each cell should be
- Good, Grover, and Mitchell (1977) suggest that the expected frequencies can be as low as 0.33 without increasing the likelihood of making a Type-I error
- Small samples reduce power