Statistics for the Sciences

Analyzing Frequencies

Xuemao Zhang
East Stroudsburg University

January 18, 2025

Outline

- One-way Frequency Table
  - Multinomial Experiment
  - Single variable goodness-of-fit tests
- Chi-Square Contingency Test
- Lab 1
- Lab 2

Multinomial Experiment

Sometimes samplings results in measurements that are qualitative or categorical rather than quantitiative. The following characteristics, which define a multinomial experiment:

- The experiment consists of **n identical trials**. (binomial)
- Each trial results in **one of k** categories.
- The probability that the outcome falls into a particular category i on a single trial is p_i and **remains constant** from trial to trial. The sum of all k probabilities,

$$p_1 + p_2 + \cdots + p_k = 1.$$

- The trials are **independent**.
- We are interested in the number of outcomes in each category, $O_1, O_2, \ldots, O_k$ with $O_1 + O_2 + \cdots + O_k = n$.

Multinomial Experiment

- Example: A researcher wants to know the distribution of different blood types (A, B, AB, and O) in a sample of 200 individuals. The data collected can be summarized in a one-way table.

```
##   Blood_Type Frequency
## 1          A        80
## 2          B        50
## 3         AB        30
## 4          O        40
```

Multinomial Experiment

- In the multinomial experiment, we make inferences about all the probabilities, $p_1, p_2, p_3, \ldots, p_k$.
- It can be shown that the expected number of outcomes resulting in category i is

$$\boxed{E(O_i) = np_i, \quad i = 1, 2, \ldots, k.}$$

- Suppose that we hypothesize values for $p_1, p_2, \ldots, p_k$ and calculate the expected value for each cell. Certainly if our hypothesis is true, the cell counts ni should not deviate greatly from their expected values np_i for $i = 1, 2, \ldots, k$. Hence, it would seem intuitively reasonable to use a test statistic involving the k deviations,

$$O_i - E(O_i) = O_i - np_i, \quad i = 1, 2, \ldots, k.$$

One-way Chi-Square Test

In 1900 Karl Pearson proposed the following test statistic

$$X^2 = \sum_{i=1}^{k} \frac{[O_i - E(O_i)]^2}{E(O_i)} = \sum_{i=1}^{k} \frac{[O_i - np_i]^2}{np_i}.$$

It can be shown that when n is large, X^2 has an approximate chi-square (χ^2) probability distribution.

One-way Chi-Square Test

- Note that for each category the Pearson statistic computes **(observed-expected)^2/expected** (noting that we assume H_0 true and under this assumption, the expected number in category i is $np_i^{(0)}$) and sums over all categories.

- When H_0 is true, the differences observed-expected for all cells will be small, but large when H_0 is false. We reject H_0 only if X^2 is **large**.

One-way Chi-Square Test

- Sample size requirement: Experience has shown that the cell counts n_i should not be too small if the χ^2 distribution is to provide an adequate approximation to the distribution of X^2. As a rule of thumb, we will require that **all expected cell counts are at least five**, although Cochran (1952) has noted that this value can be as low as one for some situations.

- Determine df: The **principle** to determine the df is: *the appropriate number of degrees of freedom will equal the number of cells, k, less 1 df for each independent linear restriction placed on the cell probabilities.*

One-way Chi-Square Test

- Example: We test $H_0 : P(A) = 34\%, P(B) = 9\%, P(AB) = 4\%, P(O) = 53\%$

```
##
##  Chi-squared test for given probabilities
##
## data:  observed_counts
## X-squared = 160.6, df = 3, p-value < 2.2e-16
```

Contingency Table

- Analysis of categorical data is based on counts, proportions or percentages of data that fall into the various categories defined by the variables.
- Suppose a population is partitioned into rc categories, determined by r levels of variable 1 and c levels of variable 2. The population proportion for level i of variable 1 and level j of variable 2 is p_{ij}. This information can be displayed in the following $r \times c$ table:

Two-Way Table of Proportions

	Column				Marginals
row	1	2	...	c	
1	p_{11}	p_{12}	...	p_{1c}	$p_{1\cdot}$
2	p_{21}	p_{22}	...	p_{2c}	$p_{2\cdot}$
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
r	p_{r1}	p_{r2}	...	p_{rc}	$p_{r\cdot}$
Marginals	$p_{\cdot 1}$	$p_{\cdot 2}$	...	$p_{\cdot c}$	1

Contingency Table

- Data summary:

Two-Way Table of Counts

row	Column 1	2	...	c	Marginals
1	O_{11}	O_{12}	...	O_{1c}	$R_{1\cdot}$
2	O_{21}	O_{22}	...	O_{2c}	$R_{2\cdot}$
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
r	O_{r1}	O_{r2}	...	O_{rc}	$R_{r\cdot}$
Marginals	$C_{\cdot 1}$	$C_{\cdot 2}$	...	$C_{\cdot c}$	n

- We want to test

H_0 : row and column variables are independent

H_a : row and column variables are not independent.

Chi-Square Contingency Test

- To do so, we select a random sample of size n from the population. Suppose the table of observed frequencies is

	Column				Totals
row	1	2	...	c	
1	O_{11}	O_{12}	...	O_{1c}	$R_{1\cdot}$
2	O_{21}	O_{22}	...	O_{2c}	$R_{2\cdot}$
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
r	O_{r1}	O_{r2}	...	O_{rc}	$R_{r\cdot}$
Totals	$C_{\cdot 1}$	$C_{\cdot 2}$	...	$C_{\cdot c}$	n

- It can be shown that under H_0 the expected cell frequency for the ij cell is given by

$$E_{ij} = \frac{\text{row i total} \times \text{column j total}}{\text{sample size}}$$
$$= \frac{R_{i\cdot} C_{\cdot j}}{n} = n\hat{p}_{i\cdot}\hat{p}_{\cdot j},$$

where $\hat{p}_{i\cdot} = R_{i\cdot}/n$ and $\hat{p}_{\cdot j} = C_{\cdot j}/n$.

Chi-Square Contingency Test

To measure the deviations of the observed frequencies from the expected frequencies under the assumption of independence, we construct the Pearson χ^2 statistic

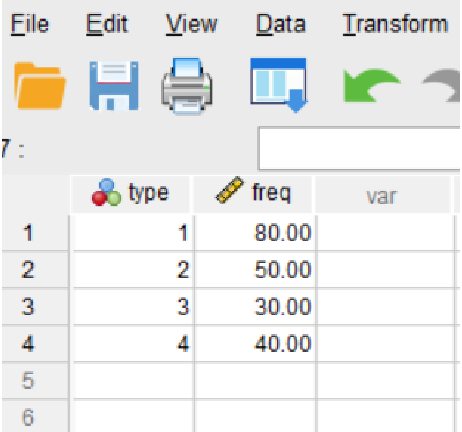
$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

If H_0 is true, X^2 has (approximately) a $\chi^2_{(\mathbf{r-1})(\mathbf{c-1})}$ distribution.

(Note that for the approximation to be valid, we require that $E_{ij} \geq 5$).

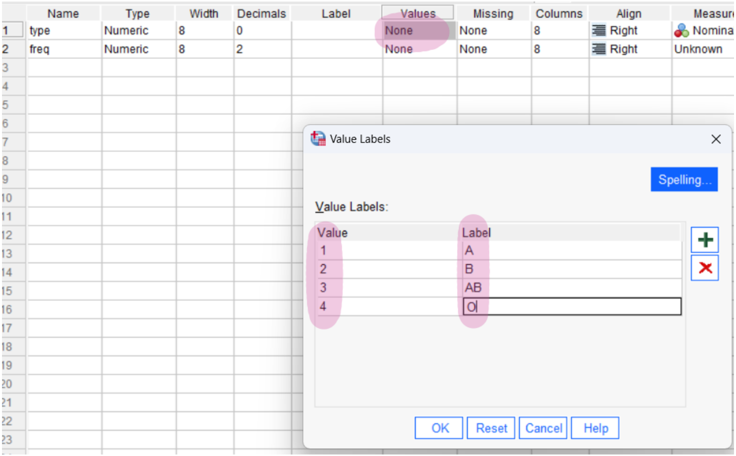
Lab 1

- Enter data manually



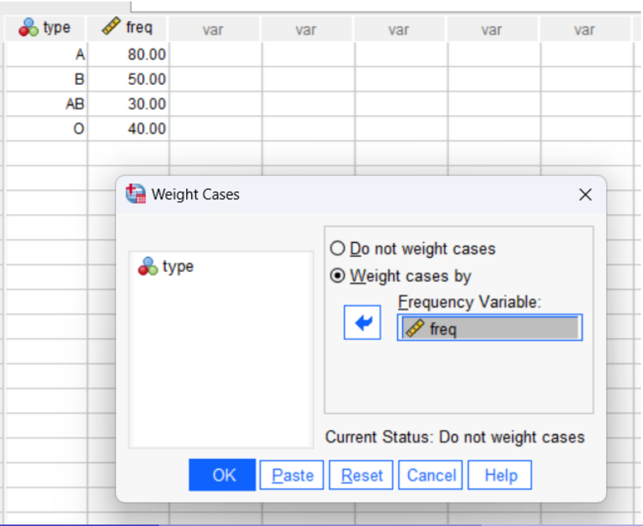
Lab 1

- Go to `Variable View`, value labels

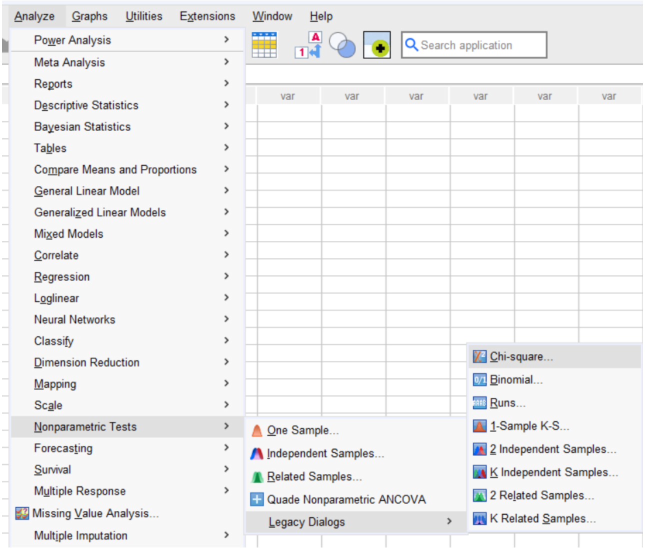


Lab 1

- Note that the data is **NOT** raw data, it is a summarized frequency table.
-
  - So we need to weight the cases
  - Click on `Data` → `Weight Cases ...`

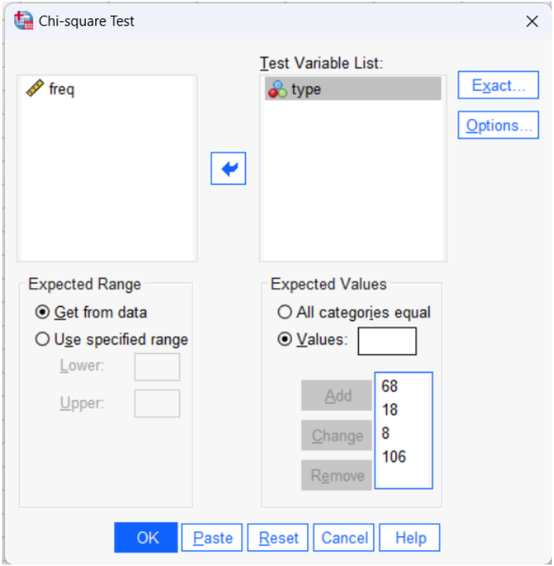


Lab 1



Analyze
Graphs
Utilities
Extensions
Window
Help
Search application
var
var
var
var
var
var
Power Analysis
Meta Analysis
Reports
Descriptive Statistics
Bayesian Statistics
Tables
Compare Means and Proportions
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Mapping
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
One Sample...
Independent Samples...
Related Samples...
Quade Nonparametric ANCOVA
Legacy Dialogs
Chi-square...
Binomial...
Runs...
1-Sample K-S...
2 Independent Samples...
K Independent Samples...
2 Related Samples...
K Related Samples...

Lab 1


Chi-square Test
freq
Test Variable List:
type
Exact...
Options...
Expected Range
Get from data
Use specified range
Lower:
Upper:
Expected Values
All categories equal
Values:
Add
Change
Remove
68
18
8
106
OK
Paste
Reset
Cancel
Help

Lab 1

NPar Tests

Chi-Square Test

Frequencies

type

	Observed N	Expected N	Residual
1 A	80	68.0	12.0
2 B	50	18.0	32.0
3 AB	30	8.0	22.0
4 O	40	106.0	-66.0
Total	200		

Test Statistics

	type
Chi-Square	160.601[a]
df	3
Asymp. Sig.	<.001

a. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected

Lab 2

- `teng.csv`: Teng et al. (2020) analyzed the results of a survey of domestic cat owners in Australia. The survey focused on factors (e.g. cat demographics, owner attitudes and demographics, etc.) that might affect the prevalence of overweight and obese cats. They related nearly 1400 survey responses of owner-assessed body condition score [BCS with five categories: very underweight (1), somewhat underweight (2), ideal (3), chubby/overweight (4), and fat/obese (5)] to a range of categorical predictors with a multivariate multinomial GLM. We will use one aspect of their data to construct a contingency table relating the BCS, reduced to three categories (1&2, 3, 4&5) to cats' begging behavior (four categories: never, sometimes, often, always).

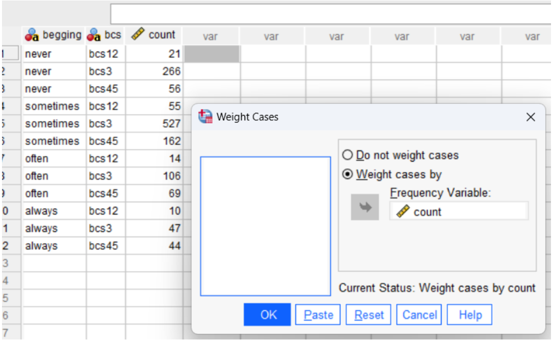
Lab 2

- Data

```
##      begging   bcs count
## 1      never bcs12    21
## 2      never  bcs3   266
## 3      never bcs45    56
## 4  sometimes bcs12    55
## 5  sometimes  bcs3   527
## 6  sometimes bcs45   162
## 7      often bcs12    14
## 8      often  bcs3   106
## 9      often bcs45    69
## 10    always bcs12    10
## 11    always  bcs3    47
## 12    always bcs45    44
```

Lab 2

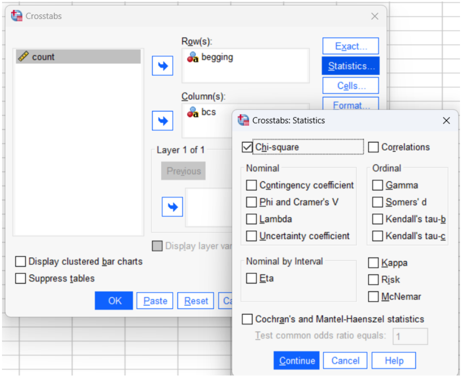
- Like in Lab 1, weight cases



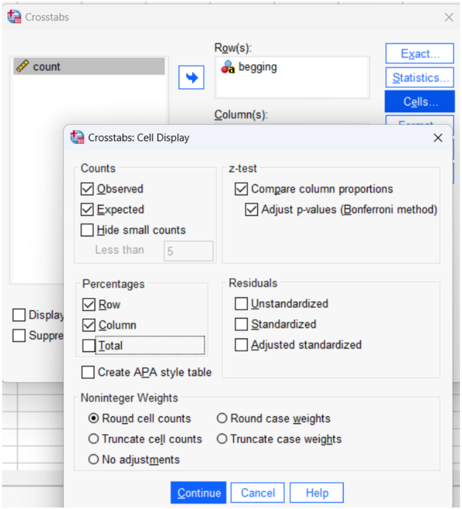
	begging	bcs	count
1	never	bcs12	21
2	never	bcs3	266
3	never	bcs45	56
4	sometimes	bcs12	55
5	sometimes	bcs3	527
6	sometimes	bcs45	162
7	often	bcs12	14
8	often	bcs3	106
9	often	bcs45	69
10	always	bcs12	10
11	always	bcs3	47
12	always	bcs45	44

Lab 2

- Click on `Analyze` → `Descriptive Statistics` → `Crosstabs ...`, and then



Lab 2


Crosstabs
count
Row(s):
begging
Column(s):
Exact...
Statistics...
Cells...
Display
Suppre
Crosstabs: Cell Display
Counts
Observed
Expected
Hide small counts
Less than
5
z-test
Compare column proportions
Adjust p-values (Bonferroni method)
Percentages
Row
Column
Total
Residuals
Unstandardized
Standardized
Adjusted standardized
Create APA style table
Noninteger Weights
Round cell counts
Round case weights
Truncate cell counts
Truncate case weights
No adjustments
Continue
Cancel
Help

Lab 2

- The extremelly small p-value shows that the two categorical variables are **NOT** independent.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	55.928[a]	6	<.001
Likelihood Ratio	53.239	6	<.001
N of Valid Cases	1377		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 7.33.

License