Analysis of Frequencies

Lecture 10

1 Outline

- Introduction to Categorical Data
- Goodness-of-Fit Tests
- Test of Homogeneity
- Test of Independence
- Fisher's Exact Test
- McNemar's Test

2 Categorical Data

2.1 Associations between Categorical Data

Often, we are interested in testing for association between categorical data. These variables can be nominal or ordinal.

> ℹ Note
>
> Suppose we want to test whether there is an association between the type of hospital a patient is admitted to and their diagnosed conditions.

2.2 Contingency Tables

A contingency table is a table of frequencies or counts for each possible combination of the variables.

Example

The hair and eye color of male statistics students were measured. The findings were summarized in the following contingency table.

```
       Eye
Hair    Brown Blue Hazel Green
  Black    32   11    10     3
  Brown    53   50    25    15
  Red      10   10     7     7
  Blond     3   30     5     8
```

2.3 Contingency tables in R

You can use `xtabs()` to produce contingency tables in `R`. The data should be formatted such that the two variables are defined by two columns. If these columns are `x` and `y` from a data frame `df`, the sample code would look like:

```
xtabs(~x+y,data=df)
```

`x` will be assigned as the row variable, `y` will be assigned as the column variable.

2.4 Example

The data set `penguins` preloaded in `R` includes data from penguins at the Palmer Archipelago in Antarctica.

2.4.1 Question

Create a contingency table for this data that summarizes the number of penguins belonging to each species and sex.

2.4.2 Answer

```
xtabs(~species+sex,data=penguins)
```

```
           sex
species     female male
  Adelie        73   73
  Chinstrap     34   34
  Gentoo        58   61
```

2.5 Chi-Squared Distribution

Tests involving categorical data use the chi-squared distribution to approximate the distribution of the test statistics.

2.5.1 Support

The chi-squared distribution is defined for non-negative values $(0, \infty)$.

2.5.2 Parameter

The chi-squared distribution can be defined by the degrees of freedom ν or df.

Like the t-distribution, we only need one value for the degree of freedom.

2.5.3 Relation to Gaussian Distribution

Suppose Z follows the standard Gaussian distribution N(0,1). Then Z^2 follows a chi-square distribution with 1 degree of freedom.

2.5.4 Mean and Variance

The mean of the chi-squared distribution is k, and the variance of the chi-squared distribution is $2k$.

3 Goodness-of-Fit Tests

3.1 Goodness of Fit

Suppose we want to test if the data follows a specified distribution.

ℹ Example

Suppose we want to know if a six-sided die is fair. We would expect to roll the numbers uniformly after a large number of throws. However, there will be variability due to randomness. The goodness-of-fit test will provide information if we have evidence of deviating from the pre-specified uniform distribution.

3.2 Hypothesis Test

The null hypothesis of the goodness-of-fit test is that **the observed data follows the specified distribution**, while the alternative is that it does not follow the specified distribution.

3.3 Test Statistic

Suppose there are k bins separating the data and that the distribution provides an expected number/counts of events E_i for $i = 1, 2, ...k$. If the observed number/counts in the data is O_i, we define the test statistics Q such that

$$Q = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

! Important

Q approximately follows the chi-squared distribution with $k-1$ degrees of freedom, denoted by χ^2_{k-1}.
For the chi-squared distribution assumption to be valid, the expected value of each bin should be greater than 5.

3.4 p-value calculation

The p-values can be calculated using the chi-squared distribution such that:

$$p - value = P(\chi^2_{k-1} \geq Q)$$

3.5 `R` implementation

There are two ways to check for goodness of fit: formal statistical analysis and exploratory data visualization.

3.5.1 Expected Values

For discrete distributions, the expected values can be calculated using the PDF functions (`dpois`, `dbinom`) for discrete distributions, and CDF functions (`pnorm`,`punif`, `pt`).

3.5.2 Formal Statistical Analysis

The `chisq.test` function performs the chi-squared test of goodness of fit. The function needs a vector of the observed variables, `x_observed` and a vector of probabilities `p`.

```
chisq.test(x=x_observed, p=p)
```

❗ Important

The `chisq.test()` function includes a continuity correction in calculating the test statistic and the corresponding p-value. If we want the uncorrected statistic and p-value, we need to specify `correct=FALSE`.

3.5.3 Visualization

Once the expected values are calculated, we can plot these values using `ggplot()`.

3.6 Example

Electronic integrated circuits are produced from thin wafers that are cut from some material. The wafers produced sometimes have tiny flaws on them that make part of the wafer unusable. Suppose we produce 1000 wafers and for each we determine the number of flaws.

3.6.1 Data

Number of Flaws	Observed Frequency
0	10
1	220
2	130
3	80
4	60
5 or more	90

3.6.2 Question

Test whether these data follow a Poisson distribution with $\lambda = 1.44$. Visualize the observed and expected counts to support the results of the test.

3.6.3 Answer

We specify the observed variable.

```r
observed <- c(10,220,130,80,60,90)
observed
```

```
[1]  10 220 130  80  60  90
```

```r
total <- sum(observed)
total
```

```
[1] 590
```

The specified distribution is the Poisson distribution. We can then calculate the expected probabilities using `dpois` and `ppois`.

```r
expected <- c(
dpois(0,lambda=1.44),
dpois(1,lambda=1.44),
dpois(2,lambda=1.44),
dpois(3,lambda=1.44),
dpois(4,lambda=1.44),
1-ppois(4,lambda=1.44)
)
expected
```

```
[1] 0.23692776 0.34117597 0.24564670 0.11791042 0.04244775 0.01589140
```

```r
sum(expected) # must be 1
```

```
[1] 1
```

```
total*expected # there should not be more than 1.5 bins that have less than 5 expected counts
```

```
[1] 139.787378 201.293824 144.931553  69.567145  25.044172   9.375928
```

All expected counts are above 5. We can now use the chi-square approximation for the p-value.

```
chisqtest <- chisq.test(x=observed,p=expected)
chisqtest
```

```

	Chi-squared test for given probabilities

data:  observed
X-squared = 867.42, df = 5, p-value < 2.2e-16
```

The test statistic is 867.4246622 with a p-value < 2.2e-16. We reject the null hypothesis. We have sufficient evidence to conclude that the observed data does not follow the Poisson distribution.

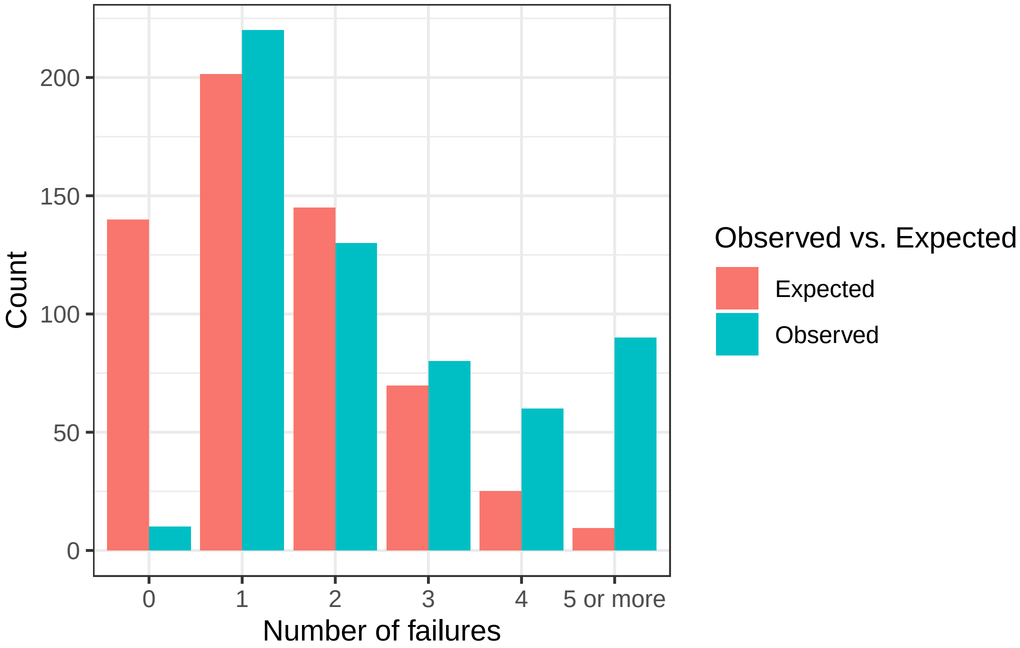
3.6.4 Plot

The function `bind_rows` appends the expected data frame to the observed data frame.

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.2
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
df_observed <- data.frame(bin = c(0,1,2,3,4,"5 or more"),y=observed,type="Observed")
df_expected <- data.frame(bin = c(0,1,2,3,4,"5 or more"),y=total*expected,type="Expected")
df <- bind_rows(df_observed,df_expected)

ggplot(df,aes(x=bin,y=y,group=type,fill=type)) +
  geom_bar(position="dodge",stat="identity") +
  theme_bw() +
  labs(x="Number of failures", y="Count",fill="Observed vs. Expected")
```



3.7 Exercise

In the “nighttime” lottery run by the state of Texas, three numbers are selected from the digits 0 through 9. The frequencies of the first digit selected over a period of almost 30 years (from 1993 to 2023) are shown below for each of the 9,215 days.

3.7.1 Data Set

Digit	Frequency
0	918
1	905

Digit	Frequency
2	908
3	916
4	900
5	911
6	963
7	948
8	937
9	909

3.7.2 Question

Test whether each digit is equally likely to have been selected in the Texas “nighttime” lottery.

3.7.3 Answer

We specify the observed variable.

```
observed <- c(918,905,908,916,900,911,963,948,937,909)
observed
```

```
 [1] 918 905 908 916 900 911 963 948 937 909
```

```
total <- sum(observed)
total
```

```
[1] 9215
```

The specified distribution is the discrete uniform distribution. We can then calculate the uniform probabilities as 1/10 (10 bins).

```
expected <- c(1/10,1/10,1/10,1/10,1/10,1/10,1/10,1/10,1/10,1/10
)
expected
```

```
 [1] 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1
```

```
sum(expected) # must be 1
```

```
[1] 1
```

```
total*expected # there should not be more than 1.5 bins that have less than 5 expected counts
```

```
 [1] 921.5 921.5 921.5 921.5 921.5 921.5 921.5 921.5 921.5 921.5
```

All expected counts are above 5. We can now use the chi-square approximation for the p-value.

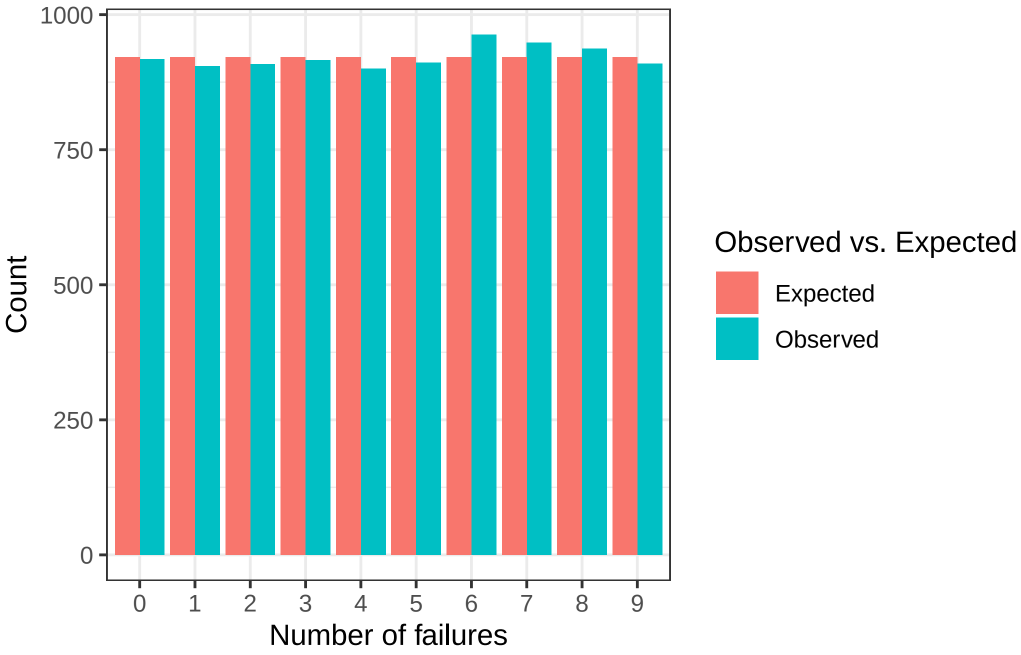
```
chisqtest <- chisq.test(x=observed,p=expected)
chisqtest
```

```


	Chi-squared test for given probabilities

data:  observed
X-squared = 4.2219, df = 9, p-value = 0.8962
```

The test statistic is 4.2219208 with a p-value 0.8962084. We fail to reject the null hypothesis. We have insufficient evidence to conclude that the digits are not equally likely to be chosen for the “nighttime” lottery.

3.7.4 Plot

```
library(tidyverse)
df_observed <- data.frame(bin = 0:9,y=observed,type="Observed")
df_expected <- data.frame(bin = 0:9,y=total*expected,type="Expected")
df <- bind_rows(df_observed,df_expected)

ggplot(df,aes(x=as.factor(bin),y=y,group=type,fill=type)) +
  geom_bar(position="dodge",stat="identity") +
  theme_bw() +
  labs(x="Number of failures", y="Count",fill="Observed vs. Expected")
```


1000
750
500
250
0
Count
0
1
2
3
4
5
6
7
8
9
Number of failures
Observed vs. Expected
Expected
Observed