Forum 6
Inference: Testing Median(s) and Categories
Data Skills for Scientists

Inference

The 3 topics in the second module, ‘Inference’ are:

- Estimating Mean(s);
- Testing Mean(s); and
- Testing Median(s) and Categories.

In this forum we will cover the third topic, Testing Median(s) and Categories.

Key topics for Testing Median(s) and Categories

- Hypothesis testing for categorical data, focusing on the Chi-squared test.
- Hypothesis testing for distributions and medians, focusing on the Mann-Whitney U test.

Chi-square Test

Two-way tables

How can we compare proportions in more than two samples or groups?

We can use **two-way tables of counts** to describe relationships between any two categorical variables.

Example - Research question

"Does age influence whether people consider it rude to recline a seat on a plane?"

Data

843 people were asked their age and their answer to the question "Is it rude to recline a seat on a plane?".

Data

We first load their answers from "flying_etiquette.xlsx",

```
flying <- readxl::read_excel(
  here::here("data", "flying_etiquette.xlsx")
  )
```

Data

```
flying %>%
  head() %>%
  gt()
```

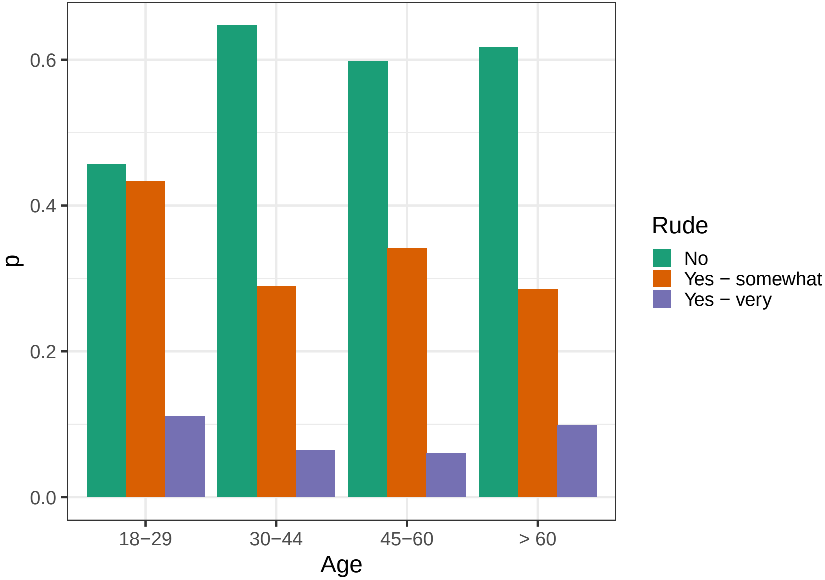
recline	rude	Gender	Age	Education
Always	No	Male	45-60	Bachelor degree
Once in a while	No	Male	45-60	Some college or Associate
Usually	No	Male	45-60	Graduate degree
Never	No	Male	45-60	Bachelor degree
Always	No	Male	45-60	Bachelor degree
Always	No	Male	45-60	Bachelor degree

Visualising the data - Two-way table of counts

```r
mytable <- table(flying$Age,flying$rude)
mytable
```

```
        No Yes - somewhat Yes - very
  > 60  133            61         21
  18-29  78            74         20
  30-44 143            64         15
  45-60 140            80         14
```

Visualising the data - Side-by-side bar chart



Null and alternative hypotheses

H_0 : the variables $<$variable 1$>$ and $<$variable 2$>$ are independent.

H_a : the variables $<$variable 1$>$ and $<$variable 2$>$ are dependent.

What is the null and alternative hypotheses for the flying dataset?

Expected cell counts

Two-way tables sort the data according to two categorical variables.

We aim to test the hypothesis that there is no relationship between these two categorical variables.

To test this hypothesis, we compare actual counts from the sample data with expected counts, given the null hypothesis of no relationship.

The expected count in any cell of a two-way table when H_0 is true is:

$$\text{Expected count} = \frac{\text{Row total} \times \text{Column total}}{N}.$$

where N is the total number of observations.

Example

What are the expected cell counts for the (18-29, No) cell?

Example

We can use R to compute the expected cell counts,

```
mytest <- chisq.test(mytable)
mytest$expected
```

```
             No Yes - somewhat Yes - very
  > 60  125.9905       71.15658   17.85291
  18-29 100.7924       56.92527   14.28233
  30-44 130.0925       73.47331   18.43416
  45-60 137.1246       77.44484   19.43060
```

The Chi-square statistic

To see if the data give convincing evidence against the null hypothesis, we compare the observed counts from our sample with the expected counts assuming H_0 is true.

The test statistic that makes the comparison is the Chi-square statistic:

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{Observed - Expected})^2}{\text{Expected}}.$$

Example

What is the contribution to the Chi-square statistic for the (18-29, No) cell?

Example

We can use R to evaluate the contribution to the Chi-square statistic for each cell,

```
Numerator <- (mytest$observed - mytest$expected)^2
Denominator <- mytest$expected
ChiSquare <- Numerator/Denominator
round(ChiSquare,1)
```

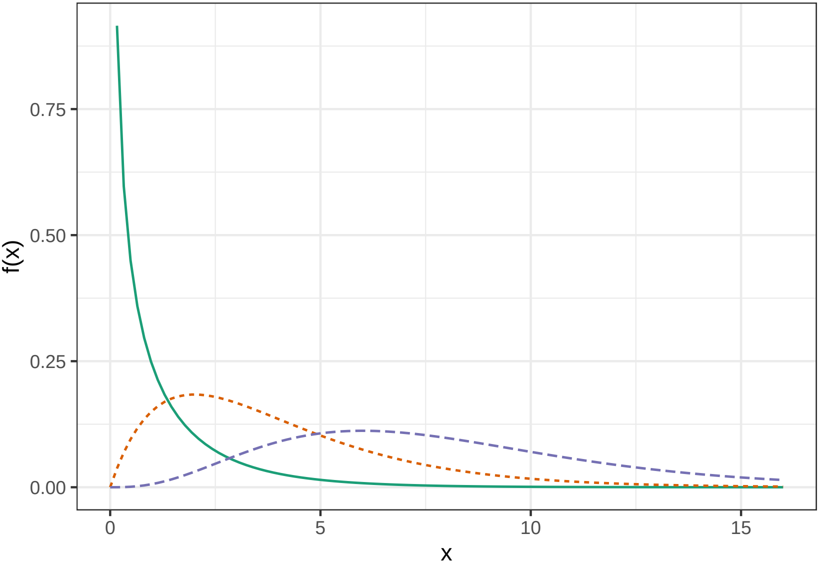
```
        No Yes - somewhat Yes - very
  > 60  0.4           1.4        0.6
  18-29 5.2           5.1        2.3
  30-44 1.3           1.2        0.6
  45-60 0.1           0.1        1.5
```

Calculating the P-value

If the null hypothesis is true, then the test statistic has a χ^2 distribution with $(r-1) \times (c-1)$ degrees of freedom, where r is the number of rows and c is the number of columns.

Chi-square distribution

Green: df = 4; orange: df=8; purple: df = 1



When can we use the Chi-square test?

Since the Chi-square test is an approximate method that becomes more accurate as the counts in the cells of the table get larger, a "rule-of-thumb" is to use the Chi-square test when:

- No more than 20% of expected counts are less than 5,
- All individual expected counts are greater than or equal to 1.

We also require that:

- There are fixed number of independent observations,
- Each observation falls into one of a finite number of complementary and mutually exclusive outcomes.

Chi-square test in R

```r
mytest <- chisq.test(mytable)
mytest
```

```
    Pearson's Chi-squared test

data:  mytable
X-squared = 19.763, df = 6, p-value = 0.003051
```

Interpret the test output

- What is the conclusion for a $\alpha = 0.05$ significance level?
- What is the conclusion for a $\alpha = 0.01$ significance level?

Significance level

If you were performing this study when should you choose your significance level?

- At the start?
- When analysing the data?
- At the end?

Residuals

In cases where we reject the null hypothesis, we can examine the residuals to identify where the association is.

A residual is defined as

$$\text{Residual} = \text{Observed} - \text{Expected}.$$

The standardized residual accounts for how much each cell contributes to the Chi-square statistic and is defined as

$$\text{Standardized Residuals} = \frac{\text{Observed} - \text{Expected}}{\sqrt{\text{Expected}}}.$$

Rule of thumb: any standardized residual larger than 2 or less than -2 indicates that the observed number in the cell is significantly above or below the expected value for that cell.

Example

```r
mytest$stdres
```

```
               No Yes - somewhat Yes - very
  > 60   1.1244821     -1.7054866  0.9011835
  18-29 -3.9548490      3.1011884  1.7709103
  30-44  2.0491989     -1.5742698 -0.9731982
  45-60  0.4490075      0.4176392 -1.5136806
```

Interpret the standardized residuals

Based on the standardized residuals, where is the association (if there is an association)?

Summary of Chi-square test

Hypothesis testing association between two categorical variables

- **Hypotheses:**

$$H_0 : \text{the two variables are independent}$$
$$H_a : \text{the two variables are dependent}$$

- **Test statistic:**

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

- **P-value:** Calculate using a χ^2 distribution with $(r-1) \times (c-1)$ degrees of freedom, where r is the number rows and c is the number of columns.

Testing Distributions and Medians

What if I would like to apply a t-test but my data is not normally distributed?

- If the sample size is sufficiently large (around 30 or above) and does not contain outliers then the central limit theorem ensures that the sample means should be normally distributed (may apply t-tests).
- If the sample size is small or contains outliers, there are other options.

Approaches to non-normal data

- If due to outliers, check if you can remove them.
- Transform the data, e.g. take logs.
- Use other distributions.
- Bootstrap methods.
- Non-parametric methods

Non-parametric tests

One approach is to use the ranking of the numbers rather than their values.

Examples:

- The Mann–Whitney U test (also known as the Wilcoxon rank sum test) is a nonparametric alternative to the two-sample t-test.
- The Wilcoxon signed rank test is a nonparametric alternative to the one-sample t-test and matched pairs t-test.

We will focus on the Mann–Whitney U test.

Example - Research question

Which coffee shop (Dreams versus Regresso) has the shortest waiting time?

Data

A researcher measures the waiting time (seconds) of the queue for each coffee shop 5 times.

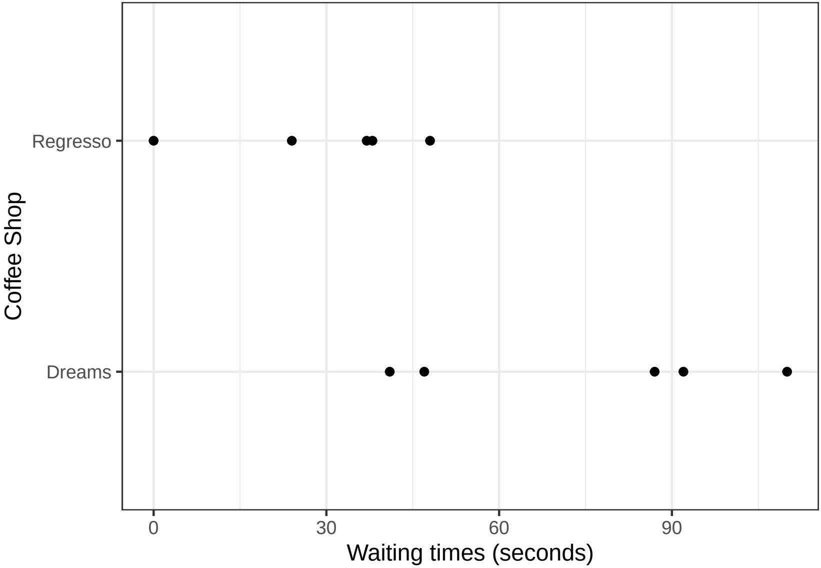
Dreams

110 92 41 87 47

Regresso

0 37 24 48 38

Visualising the data



Visualising the data

Why did we use a dot plot rather a box plot?

Ranking

We will replace their times with their rank:

CoffeeShop	times	rank
Regresso	0	1
Regresso	24	2
Regresso	37	3
Regresso	38	4
Dreams	41	5
Dreams	47	6
Regresso	48	7
Dreams	87	8
Dreams	92	9
Dreams	110	10

The Mann–Whitney U test

1. Order the data and get the rank of each observation.
2. Add up the ranks for the first group R_1.
3. Calculate the test statistic (the U-statistic):

$$U = R_1 - n_1(n_1 + 1)/2,$$

where n_1 is the number of observations in the first group.

- What is the smallest value U can take?
- What is the largest value U can take?

Examples

What is the Mann–Whitney U test statistic for the coffee dataset?

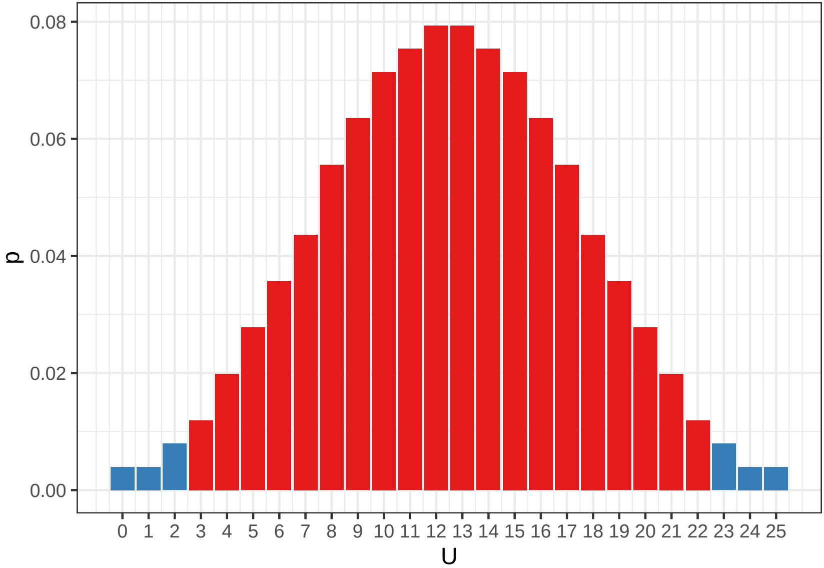
Null and alternative hypotheses

In the Mann–Whitney U test, we have

$$H_0 : \text{no difference between the two groups.}$$
$$H_a : \text{a difference between the two groups.}$$

- What are the null and alternative hypotheses for the coffee dataset?

Calculating the P-value



Calculating the P-value

U	p
0	0.0040
1	0.0040
2	0.0079
23	0.0079
24	0.0040
25	0.0040

The P-value is 0.031746

R output

```r
wilcox.test(Dreams,Regresso)
```

```
	Wilcoxon rank sum exact test

data:  Dreams and Regresso
W = 23, p-value = 0.03175
alternative hypothesis: true location shift is not equal to 0
```

Note

The R function, `wilcox.test` use an exact p-value if the samples contain less than 50 finite values and not ties. Otherwise, a normal approximation is used.

Interpreting the results

- What is the conclusion for a $\alpha = 0.05$ significance level?
- What is the conclusion for a $\alpha = 0.01$ significance level?

What about checking assumptions?

There are no assumptions made about the shape of the population distribution for a Mann-Whitney U test.

What if we had ties?

Usually assign all tied values the **average** of the ranks they occupy.

E.g.,

Observation	0	2	2	4	8
Rank	1	2.5	2.5	4	5

Special case

When **both populations have distributions of the same shape and spread**, the Mann Whitney U test is a test about population medians (rather than population means like in a two sample t-test),

$$H_0 : \text{median}_1 - \text{median}_2 = 0.$$
$$H_a : \text{median}_1 - \text{median}_2 \neq 0.$$

where median_1 is the population median of population 1 and median_2 is the population median of population 2.

Summary of Mann-Whitney U test

Hypothesis testing for a difference in the distributions of two groups.

- **Hypotheses:**

$$H_0 : \text{no difference between the two groups.}$$
$$H_a : \text{a difference between the two groups.}$$

- **Test statistic:**

$$U = R_1 - n_1(n_1 + 1)/2,$$

- **P-value:** Exact calculation when less than 50 finite values and no ties. Otherwise, normal approximation used. We will use `wilcox.test` in R that handles both cases.

Wilcoxon signed rank test

The Wilcoxon signed rank test is a non-parametric alternative to the **one-sample t-test** or **matched pairs t-test**.

Matched pairs Wilcoxon signed rank test

- **Hypotheses:**

$$H_0 : \text{no systematic difference within the pairs.}$$
$$H_a : \text{systematic difference within the pairs.}$$

- **Test statistic:**

1. Compute the differences between each pair of values.
2. Order the differences based on absolute value and rank (pairs with zero difference are excluded and pairs with equal differences share the ranking values).
3. Calculate W_+: the sum of the rankings of all the positive differences.

- **P-value:** Exact calculation when less than 50 finite values and no ties. Otherwise, normal approximation used. We will use `wilcox.test` in R that handles both cases.

One-sample Wilcoxon signed rank test

- **Hypotheses:**

$$H_0 : \text{median} = m_0.$$
$$H_a : \text{median} \neq m_0.$$

- **Test statistic:**

1. Compute the difference between each value and m_0.
2. Order the differences based on absolute value and rank (pairs with zero difference are excluded and pairs with equal differences share the ranking values).
3. Calculate W_+: the sum of the rankings of all the positive differences.

- **P-value:** Exact calculation when less than 50 finite values and no ties. Otherwise, normal approximation used. We will use `wilcox.test` in R that handles both cases.

Summary

- The Mann–Whitney U test (also known as the Wilcoxon rank sum test) is a nonparametric alternative to the two-sample t-test.
- The matched pairs Wilcoxon signed rank test is a non-parametric alternative to the matched pairs t-test.
- The one-sample Wilcoxon signed rank test is a non-parametric alternative to the one sample t-test.

Note: While we have focused on numeric data, these ranked tests also apply to ordinal variables. For example, to analyse survey responses such as Disagree, Neutral, Agree.

Further reading/tools

- Textbook:
  - Chapter 2.6 Data analysis for two-way tables
  - Chapter 9.1 Inference for two-way tables
  - Chapter 9.2 Goodness of fit
  - Chapter 15.1 The Wilcoxon rank sum test
  - Chapter 15.2 The Wilcoxon signed rank test