

School of Languages and Social Sciences

**Quantitative Methods in Linguistics Worksheets**

Tim Grant, Urszula Clark, Gertrud Reershemius, David Pollard, Sarah Hayes, Garry Plappert, Jack Grieve

|  |
| --- |
| These worksheets were created as part of an ESRC-funded project “Integrating quantitative methods into undergraduate programmes in linguistics” undertaken at Aston University from August 2012 to August 2013. The aims were:   * design learning and teaching materials for English Language and German; * integrate these into existing undergraduate modules; * make resources available.   Prior to this project, the English and German curricula at Aston contained little if any QM. The teaching staff were also unskilled. The project trained four lecturing staff, targeted two core 1st year modules and two core 2nd year modules and ran workshops for final year students. All the teaching took place in October and November 2012.  The ethos was that students are interested in their core topic (linguistics) and therefore the route to QM training should be through their linguistics. Teaching was in sequenced, bite-sized sessions within existing tutorials. Where possible, QM was made part of the module assessment.  Training in statistical software was avoided in order to focus on understanding of principles. Supplementary screencasts cover software use for those students who wish to learn more.  http://i.creativecommons.org/l/by-nc-sa/3.0/88x31.png  This document is made available under Creative Commons CC BY-NC-SA 4.0.  16 December, 2013 |

**Contents**

*Overview*

Schema 3

*Year 1 German (LG1011, German Language Past and Present)*

Worksheet 1 (what is a number) 5

Worksheet 2 (tables) 7

Worksheet 3 (tables) 10

Worksheets 4,5,6 (visualisation) 13

*Year 2 German (LG2043, German Language for Business II)*

Worksheet 1 (questionnaire design) 17

Worksheet 2 (descriptive statistics) 20

Worksheet 3 (project instructions) 26

*Year 2 English (LE2012, Variations of English)*

Worksheet 1 (visualisation) 28 (see 13)

Worksheet 2 (descriptive statistics) 29

Worksheet 3 (normalisation, t-test) 35

Worksheet 4 (Chi-square) 43



**Year 1 German**

**LG1011**

**German Language Past and Present**

# Worksheet 1: What is a number?

Here are 3 different types of number.

|  |  |
| --- | --- |
| Nominal (Categorical) | Numbers used as labels, names or categories.  e.g. 1 – Male, 2 – Female |
| Continuous | Numbers that can take any value within a range.  e.g. a person’s height or weight. |
| Ordinal | Numbers arranged in order.  e.g. 1, strongly agree; 2, agree; 3 disagree; 4 strongly disagree |

# Question 1

Inspired by Andy Field (2009)

What type of number fits the descriptions below? Nominal, ordinal, or continuous?

|  |  |
| --- | --- |
| The current number of downloads of One Direction’s single “Live while we’re young” |  |
| The position of “Live while we’re young” in the BBC Top 40 chart |  |
| The number 1 in the street name for the band: “1D” |  |
| The amount of money the group have made so far |  |
| The number of boys in the band |  |
| The phone numbers that fans have given to the band members |  |

# Question 2

This is a screenshot from **amazon.de**. What types of number can you see here?



# Worksheet 2: What is quantitative sociolinguistic data? Part A.

## Example: attitudes towards “Low German”

Low German is a variety spoken in parts of Germany and the Netherlands. This area is shown on the map below.



In 2000, the inhabitants of one village in this area completed a survey that measured their attitudes towards Low German. Some of the results are presented overleaf.

### Further reading on Low German

Matras, Yaron & Reershemius, Gertrud (2003) Low German (East Frisian dialect). Lincom: Munich, pp.1-7

# Task

Work in pairs.

Analyse the two tables below and answer the questions on the following page.

**Table 1: Would you actively support the preservation of Low German?**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Age | Yes | Undecided | No | No answer | Total |
| 15-30 | 7 | 2 |  | 4 | 13 |
| 31-40 | 15 | 5 |  | 7 | 27 |
| 41-50 | 11 | 2 | 2 | 3 | 18 |
| 51-60 | 19 | 2 | 1 | 5 | 27 |
| over 61 | 34 |  |  | 4 | 38 |
| No answer |  | 1 | 1 |  | 2 |
| Total | 86 | 12 | 4 | 23 | 125 |

**Table 2: Which language do you speak with your children?**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Age | Low German | Low German and Standard German | Standard German | No answer | Total |
| 15-30 | 1 | 5 | 2 |  | 8 |
| 31-40 |  | 12 | 9 |  | 21 |
| 41-50 | 2 | 6 | 9 |  | 17 |
| 51-60 | 5 | 6 | 12 |  | 23 |
| over 61 | 20 | 6 | 9 |  | 35 |
| No answer |  |  |  |  | 0 |
| Total | 28 | 35 | 41 | 0 | 104 |

## Questions

1. What different kinds of numbers can you detect in the tables? Are these 3 types of number present:
   1. Nominal/categorical  
      i.e. numbers used as labels or names
   2. Continuous  
      i.e. numbers that can be any value (such as height or weight)
   3. Ordinal  
      i.e. numbers arranged in order

|  |
| --- |
|  |

1. What sort of data could have been the basis of these tables?

|  |
| --- |
|  |

1. Which are the different categories displayed in these tables and why might they be relevant for research into Low German?

|  |
| --- |
|  |

1. If you look a bit closer, you will see that the numbers tell you a story. Write that story in 150 to 200 words.

|  |
| --- |
|  |

# Worksheet 3: What is quantitative sociolinguistic data? Part B.

# Task

Work in pairs.

Analyse the two tables below and answer the questions on the following page.

**Table 1: Would you actively support the preservation of Low German?**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age | Yes | Undecided | No | No answer |
| 15-30 | 54% | 15% |  | 31% |
| 31-40 | 56% | 19% |  | 26% |
| 41-50 | 61% | 11% | 11% | 17% |
| 51-60 | 70% | 7% | 4% | 19% |
| over 61 | 89% |  |  | 11% |
| No answer |  | 50% | 50% |  |
| Total | 69% | 10% | 3% | 18% |

**Table 2: Which language do you speak with your children?**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age | Low German | Low German and Standard German | Standard German | No answer |
| 15-30 | 13% | 63% | 25% |  |
| 31-40 |  | 57% | 43% |  |
| 41-50 | 12% | 35% | 53% |  |
| 51-60 | 22% | 26% | 52% |  |
| over 61 | 57% | 17% | 26% |  |
| No answer |  |  |  |  |
| Total | 27% | 34% | 39% | 0% |

# Questions

1. What is different compared with the first set of tables we looked at (in worksheet 2)?

|  |
| --- |
|  |

1. Write the research story based on this data (up to 300 words).

|  |
| --- |
|  |

1. Explain the difference between this research story and the research story you wrote on the previous worksheet.

|  |
| --- |
|  |

# Picturing Your Data: Overall results

Compare these three types of visualisation.

**Would you actively support the preservation of Low German?**

**Which language do you speak with your children?**

|  |  |
| --- | --- |
| **Pie charts** | |
| **Pros** | **Cons** |
|  |  |

**Would you actively support the preservation of Low German?**

**Which language do you speak with your children?**

|  |  |
| --- | --- |
| **Column charts** | |
| **Pros** | **Cons** |
|  |  |

**Would you actively support the preservation of Low German?**

**Which language do you speak with your children?**

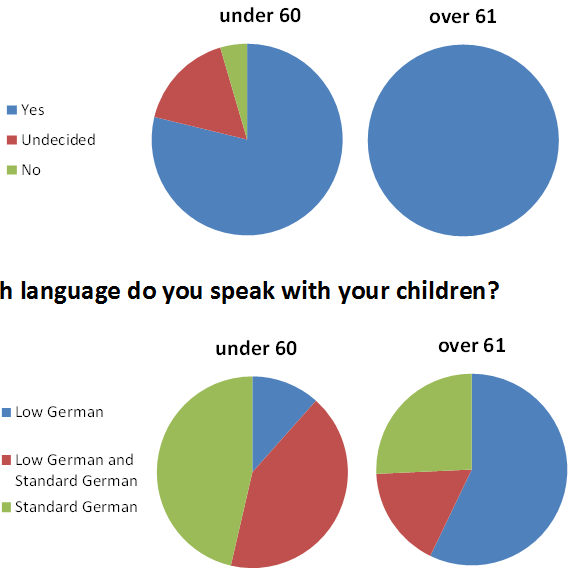
|  |  |
| --- | --- |
| **Stacked column charts** | |
| **Pros** | **Cons** |
|  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **The best visualisation is** |  | **because** |  |

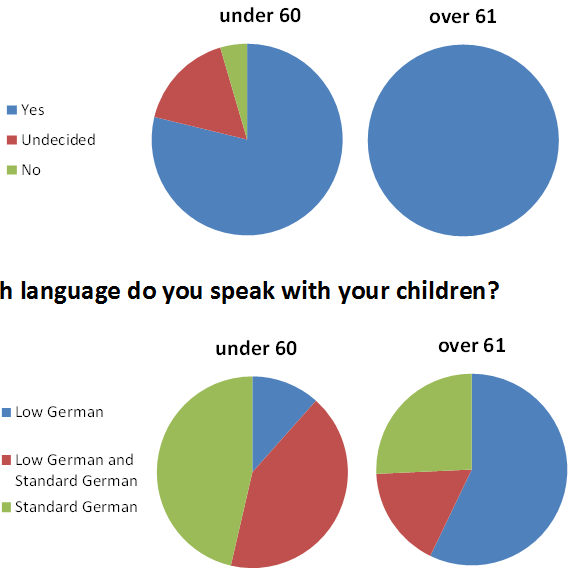
# Picturing Your Data: Under-60s vs Over-60s.

Compare these three types of visualisation.

**Would you actively support the preservation of Low German?**

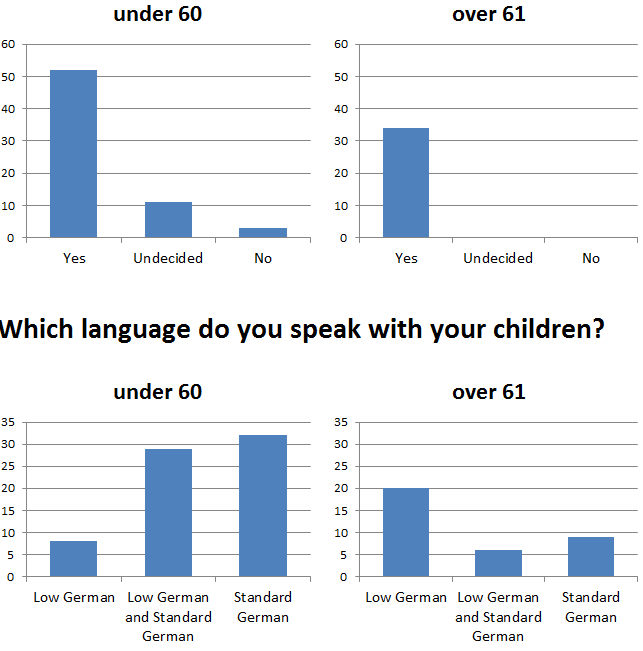


**Which language do you speak with your children?**

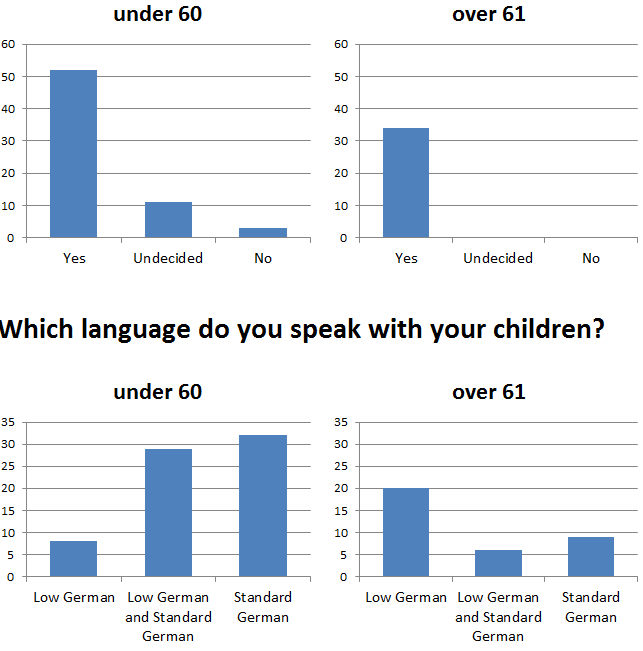


|  |  |
| --- | --- |
| **Pie charts** | |
| **Pros** | **Cons** |
|  |  |

**Would you actively support the preservation of Low German?**



**Which language do you speak with your children?**



|  |  |
| --- | --- |
| **Column charts** | |
| **Pros** | **Cons** |
|  |  |

**Would you actively support the preservation of Low German?**

****

**Which language do you speak with your children?**

****

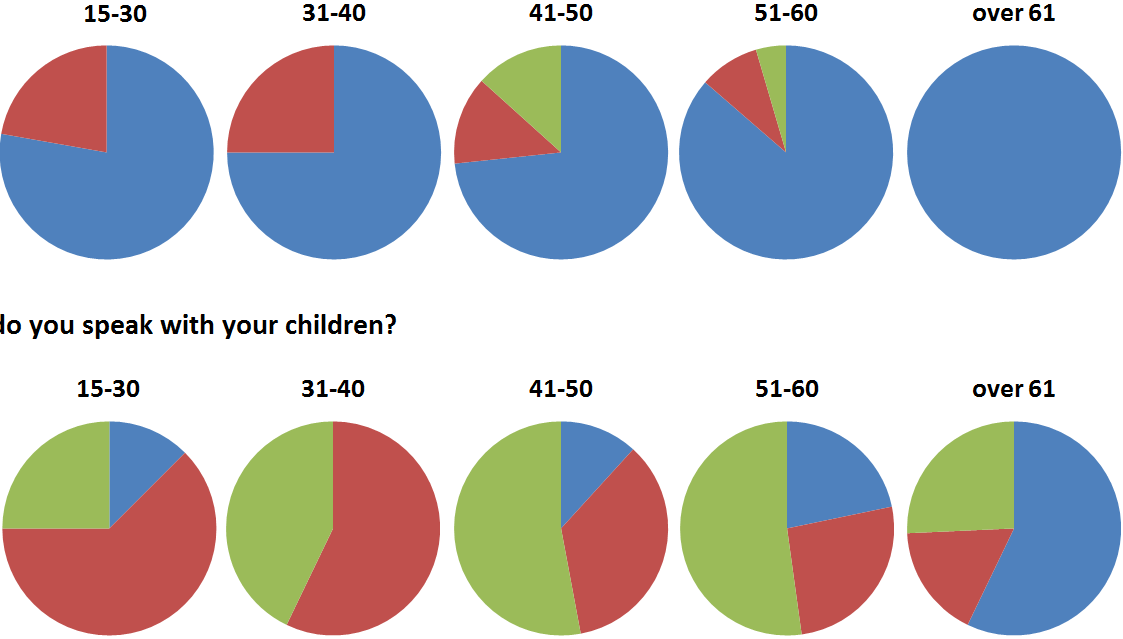
|  |  |
| --- | --- |
| **Stacked column charts** | |
| **Pros** | **Cons** |
|  |  |

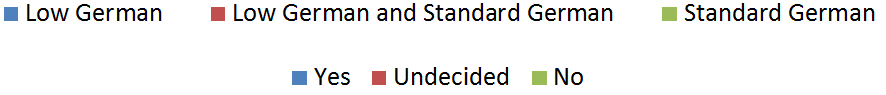
|  |  |  |  |
| --- | --- | --- | --- |
| **The best visualisation is** |  | **because** |  |

# Picturing Your Data: Results by Age Group

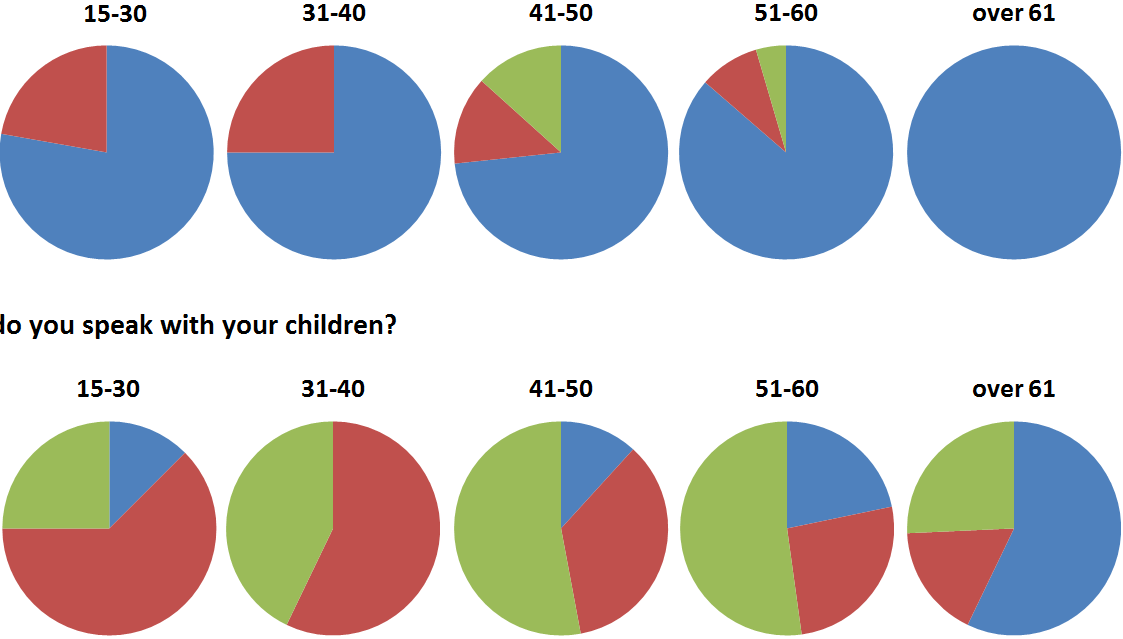
Compare these three types of visualisation.

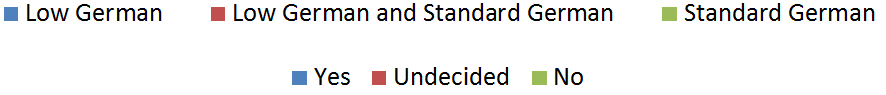
**Would you actively support the preservation of Low German?**





**Which language do you speak with your children?**





|  |  |
| --- | --- |
| **Pie charts** | |
| **Pros** | **Cons** |
|  |  |

**Would you actively support the preservation of Low German?**

**Which language do you speak with your children?**

|  |  |
| --- | --- |
| **Column charts** | |
| **Pros** | **Cons** |
|  |  |

**Would you actively support the preservation of Low German?**

**Which language do you speak with your children?**

|  |  |
| --- | --- |
| **Stacked column charts** | |
| **Pros** | **Cons** |
|  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **The best visualisation is** |  | **because** |  |

**Year 2 German**

**LG2043**

**German Language for Business II**

# Arbeitsblatt 1: Fragebogendesign

## Zunächst einmal …

Das Design des Fragebogens legt fest, was man hinterher mit den Daten tun kann.

Studierende kommen zur Beratung über statistische Fragen oft NACHDEM sie ihre Daten schon erhoben haben. Das ist dann leider eigentlich schon zu spät. Idealerweise solltet Ihr Euch schon vorher darüber im Klaren sein, was man mit den Daten machen will, bevor man sie erhebt, zum Beispiel, welche statistischen Tests man machen möchte.

Es gibt aber eine Reihe von Faustregeln zum Design von Fragebögen, die wir hier mit Euch diskutieren möchten.

Bitte beantworten Sie daher die folgenden Fragen:

### Frage 1

Welche dieser Fragen sind am sinnvollsten? Setzen Sie die Beispiele bitte in eine Rangordnung

|  |  |  |
| --- | --- | --- |
| A | Dein Alter (bitte einkreisen): | unter 15 15-30 31-45 46-60 61+ |
| B | Dein Alter: |  |
| C | Dein Alter (bitte einkreisen): | 0-10 11-20 21-30 31-40 41-50 51-60 61-70 71-80 81+ |

Frage 2

Welche der unten folgenden Fragetechniken würden Statistiker empfehlen, welche eher nicht?

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1. Das Arbeiten mit Zahlen macht mich ängstlich (bitte einkreisen) | | | | | | | | | | | | | | | | | | |
| Ich stimme nicht zu | | | | | Ich weiss nicht | | | | | | | Ich stimme zu | | | | | | |
| 1. Das Arbeiten mit Zahlen macht mich ängstlich (bitte einkreisen) | | | | | | | | | | | | | | | | | | |
| Ich stimme definitiv nicht zu | | | Ich stimme überwiegend nicht zu | | | | Weder das eine noch das andere | | | | Ich stimme überwiegend zu | | | | Ich stimme definitiv zu | | | |
| 1. Das Arbeiten mit Zahlen macht mich ängstlich (bitte einkreisen) | | | | | | | | | | | | | | | | | | |
| Ich stimme definitiv nicht zu | | Ich stimme überwiegend nicht zu | | | Ich habe leichte Vorbehalte | | | | Ich kann gerade eben zustimmen | | | Ich stimme überwiegend zu | | | | | Ich stimme definitiv zu | |
| 1. Das Arbeiten mit Zahlen macht mich ängstlich (bitte einkreisen) | | | | | | | | | | | | | | | | | | |
| Ich stimme **nicht** zu | 0 | | 1 | 2 | | 3 | 4 | 5 | | 6 | 7 | | 8 | 9 | | 10 | | Ich stimme zu |
| Das Arbeiten mit Zahlen macht mich ängstlich (kreuzen Sie bitte auf der Linie an, in welchem Maβe Sie zustimmen): | | | | | | | | | | | | | | | | | | |
| Ich stimme nicht zu |  | | | | | | | | | | | | | | | | | Ich stimme zu |

# Arbeitsblatt 2: Deskriptive Statistik

## Hintergrund

Dieses Arbeitsblatt untersucht Daten, die im Zusammenhang mit einer Konsumentenbefragung erhoben wurden. 79 Personen (35 männlich, 44 weiblich) testeten ein neues alkoholisches Getränk – Knoblauchschnapps – und bewerteten es auf einer Skala von 0 (unbeliebt) bis 10 (sehr beliebt).

## Daten

Dies sind die Daten:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Geschlecht** | **Bewertung** |  |  |  |  |  |  |
| m | 9 | m | 7 | f | 1 | f | 9 |
| m | 6 | m | 3 | f | 5 | f | 5 |
| m | 7 | m | 7 | f | 4 | f | 3 |
| m | 7 | m | 5 | f | 5 | f | 1 |
| m | 7 | m | 6 | f | 3 | f | 7 |
| m | 7 | m | 9 | f | 6 | f | 9 |
| m | 3 | m | 8 | f | 0 | f | 7 |
| m | 2 | m | 4 | f | 8 | f | 6 |
| m | 9 | m | 7 | f | 2 | f | 5 |
| m | 3 | m | 8 | f | 5 | f | 7 |
| m | 1 | m | 6 | f | 5 | f | 10 |
| m | 3 | m | 6 | f | 10 | f | 10 |
| m | 6 | m | 4 | f | 8 | f | 3 |
| m | 10 |  |  | f | 10 | f | 2 |
| m | 8 |  |  | f | 6 | f | 8 |
| m | 7 |  |  | f | 1 | f | 10 |
| m | 8 | f | 0 | f | 4 |  |  |
| m | 7 | f | 10 | f | 7 |  |  |
| m | 6 | f | 9 | f | 4 |  |  |
| m | 10 | f | 6 | f | 8 |  |  |
| m | 8 | f | 8 | f | 5 |  |  |
| m | 3 | f | 5 | f | 5 |  |  |

|  |
| --- |
| JARGON: Der technische Begriff für ‘die Personen, die an der Befragung teilgenommen haben’ ist das Sample. |

## Verschiedene Formen von Mittelwerten

Zunächst einmal errechnen wir die durchschnittliche Bewertung bei Männern und Frauen.

Die Mittelwerte sind sind:

|  |  |  |
| --- | --- | --- |
|  | Männer | Frauen |
| Durchschnitt | 6.20 | 5.73 |
| Medianwert | 7.00 | 5.50 |
| Modus | 7.00 | 5.00 |

### Frage 1

Was verraten uns die Mittelwerte?

|  |
| --- |
|  |

### Frage 2

Wir unterscheiden die folgenden drei Formen von Mittelwerten: Geht es in der Tabelle um den ‚Durchschnitt‘, den ‚Median- oder Zentralwert‘ oder um den ‚Modus‘?

|  |  |
| --- | --- |
| **Beschreibung** | **Art von Mittelwert** |
| Dies ist die üblichste Form von Mittelwert.  Um ihn im Beispiel oben für die an der Befragung teilgenommenen Frauen zu errechnen, gehen wir folgendermaβen vor:   1. Man addiert alle Bewertungen der Frauen (252). 2. Diese Summe teilt man durch die Anzahl der teilgenommenen Frauen (252/44) |  |
| Dieser Wert beschreibt die Ausprägung mit höchster Häufigkeit. |  |
| Wenn man alle Ergebnisse nimmt und sie vom niedrigsten bis zum höchsten sortiert, dann bezeichnet der Medianwert genau die Mitte.  Wenn wir in unserem Beispiel die Ergebnisse für Männer betrachten:   1. Wir sortieren alle Umfrageergebnisse für Männer (1, 2, 3, 3, 3, 3, 3, 4, 4, 5, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 9, 9, 9, 10, 10). 2. Wir suchen die mittlere Position auf dieser Liste (Position 18 von 35). 3. Dann nehmen wir die Zahl, die auf dieser Position steht (7). |  |

### Frage 3

Für welche der unten beschriebenen Situationen nimmt man welche Art von Mittelwert?

|  |  |
| --- | --- |
| **Verwendungssituation** | **Art von Mittelwert** |
| Wenn die Zahlen gleichmäβig verteilt sind, d.h. wenn es nicht zu viele extreme Abweichungen gibt |  |
| Wenn die Zahlen extreme Abweichungen zeigen, zum Beispiel wenn wir die Gehaltsliste einer Firma betrachten, dann sind die Gehälter der Topmanager Abweichungen nach oben hin. |  |
| Wenn man herausfinden will, welche Zahl in der Umfrage von den meisten Beteiligten genannt wurde. Diese Art von Mittelwert wird relativ selten verwendet. |  |

### Frage 4

Welche Art von Mittelwert sollten wir sinnvollerweise für unsere Koblauchschnappsdaten verwenden?

|  |
| --- |
|  |

## Standardabweichungen – wie misst man Verteilung?

Wenn der Durchschnittswert für eine Gruppe höher ist als der einer anderen Gruppe, dann sagt das noch nichts darüber aus, ob es einen entscheidenden Unterschied zwischen den beiden Gruppen gibt.

Hier noch einmal die Durchschnittswerte unserer Befragung:

|  |  |  |
| --- | --- | --- |
|  | Männer | Frauen |
| Durchschnittswert | 6.20 | 5.73 |

Der Durchschnittswert erzählt uns einen Teil der Geschichte, aber noch nicht alles. Was wir herausfinden müssen ist: Befinden sich die meisten Ergebnisse der Begragung in der Nähe des Durchschnittswertes oder sind sie weiter verteilt.

Der Wert, der uns dies zeigt, ist die sogenannte Standardabweichung. Man kann sie mit Hilfe von Programmen wie SPSS oder EXCEL errechnen.

Hier die Standardabweichung für die Knoblauchschnappsumfrage:

|  |  |  |
| --- | --- | --- |
|  | Männer | Frauen |
| Durchschnittswert | 6.20 | 5.73 |
| Standardabweichung | 2.31 | 2.94 |

Frage: Was sagt uns in diesem Fall die Standardabweichung über die Verteilung der Ergebnisse?

Antwort: Sie zeigt uns, dass ungefähr 68% der Ergebnisse für Männer sich 2.31 um den Durchschnittswert ( 6.20) herum befinden. 68% der Antworten fallen also zwischen 3.89 und 8.51.

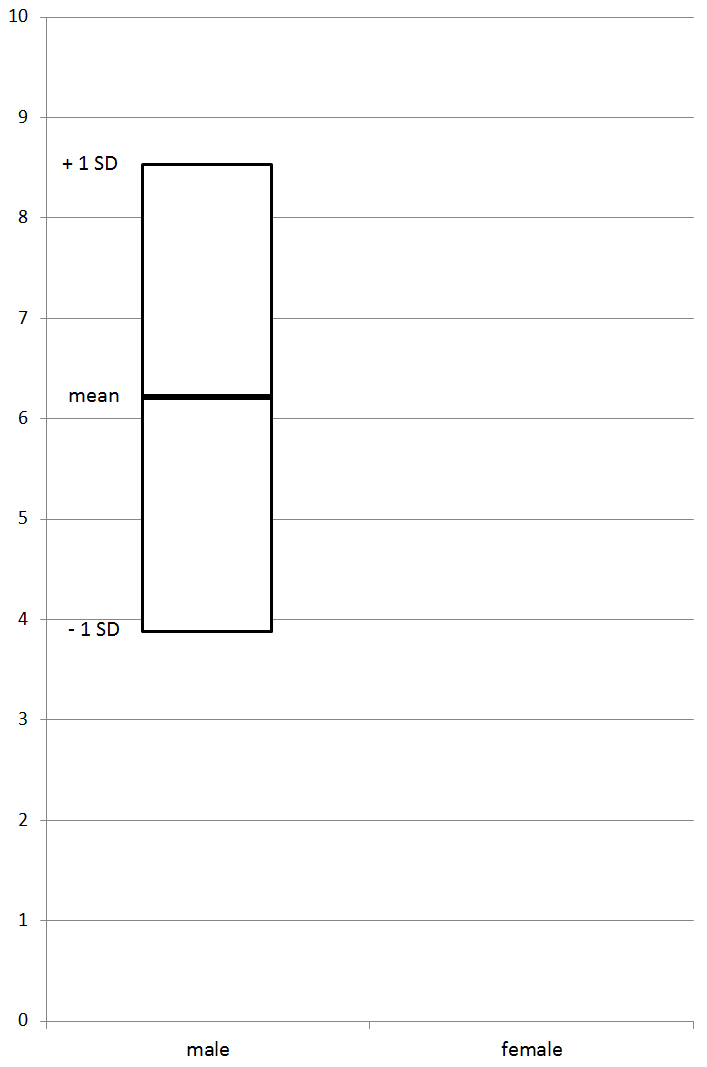
### Frage 5

Benutze den Taschenrechner in deinem Handy um die Lücken zu füllen:

|  |
| --- |
| 68% der Ergebnisse für Männer sind zwischen 3.89 und 8.51.  68% der Ergebnisse für Frauen sind zwischen und . |

### Frage 6

Unten siehst du eine Boxplotgrafik, auf der der Durchschnittswert und die Standardabweichung für die Ergebnisse bei den Männern verzeichnet sind. Vervollständige die Grafik mit den Ergebnissen für die Frauen.



### Frage 7

Sieh dir noch einmal die Boxplotgrafik an, die du gerade vervollständigt hast. Was sagt sie über den Unterschied zwischen Männern und Frauen aus?

|  |
| --- |
|  |

## Ist der Unterschied signifikant oder nicht?

Wenn wir uns die Boxplotgrafik ansehen, dann könnten wir annehmen, dass der Unterschied zwischen den beiden Gruppen nicht signifikant ist.

Um diesen ersten Eindruck zu bestätigen, sollten wir einen statistischen Test mit unseren Daten durchführen , in diesem Fall einen sogenannten **t-Test**.

Wie man einen t-Test durchführt und interpretiert, lernen wir an anderer Stelle.

|  |
| --- |
| JARGON: **Der Begriff signifikante Abweichung ist ein statischer Terminus, d.h. ein statistischer Fachbegriff.**  In einer statistischen Analyse sollte man also den Begriff nur dann verwenden, wenn man einen statistischen Test durchgeführt hat, wie zum Beispiel einen t-Test, einen Chi-Quadrat Test oder einen ANOVA. |

# Worksheet 3: Your Questionnaire

# After this lesson

As part of your market research project you will be expected to

* Gather consumer feedback on your given product
  + using a well-designed questionnaire
  + from at least 30 people, roughly 50% male and 50% female
* Analyse whether there is a significant difference between males and females in their responses to the product EITHER
  + By carrying out a t-test on the data yourself using SPSS  
    OR
  + Sending your data to David Pollard (d.j.pollard1@aston.ac.uk), who will carry out the t-test for you and email you the results
* Note: you will have to analyse the t-test results yourselves.

# In your presentation

If there was a statistically significant difference between the genders, you should include this in your presentation (stating the “p value”).

**Year 2 English**

**LE2021**

**Variations of English**

# Worksheet 1: Visualisation

See worksheets 4 to 6 for module LG1011 on pages 13 to 16

# Worksheet 2: Descriptive Statistics

## Background

This worksheet will look at data collected through a survey. 79 people (35 male, 44 female) listened to an alien speaking “Klingon” and then rated how erotic they found the language on a scale of 0 (not at all) to 10 (extremely).

## Data

These are the data.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Gender** | **Rating** |  |  |  |  |  |  |
| m | 9 | m | 7 | f | 1 | f | 9 |
| m | 6 | m | 3 | f | 5 | f | 5 |
| m | 7 | m | 7 | f | 4 | f | 3 |
| m | 7 | m | 5 | f | 5 | f | 1 |
| m | 7 | m | 6 | f | 3 | f | 7 |
| m | 7 | m | 9 | f | 6 | f | 9 |
| m | 3 | m | 8 | f | 0 | f | 7 |
| m | 2 | m | 4 | f | 8 | f | 6 |
| m | 9 | m | 7 | f | 2 | f | 5 |
| m | 3 | m | 8 | f | 5 | f | 7 |
| m | 1 | m | 6 | f | 5 | f | 10 |
| m | 3 | m | 6 | f | 10 | f | 10 |
| m | 6 | m | 4 | f | 8 | f | 3 |
| m | 10 |  |  | f | 10 | f | 2 |
| m | 8 |  |  | f | 6 | f | 8 |
| m | 7 |  |  | f | 1 | f | 10 |
| m | 8 | f | 0 | f | 4 |  |  |
| m | 7 | f | 10 | f | 7 |  |  |
| m | 6 | f | 9 | f | 4 |  |  |
| m | 10 | f | 6 | f | 8 |  |  |
| m | 8 | f | 8 | f | 5 |  |  |
| m | 3 | f | 5 | f | 5 |  |  |

|  |
| --- |
| JARGON: The technical term for “the people who took part” is **the** **sample**. |

## Averages

A first look at the data could be to calculate the average ratings for males and females.

The averages are:

|  |  |  |
| --- | --- | --- |
|  | Male | Female |
| mean | 6.20 | 5.73 |
| median | 7.00 | 5.50 |
| mode | 7.00 | 5.00 |

### Question 1

What do these averages tell you?

|  |
| --- |
|  |

### Question 2

Here are definitions of 3 types of average. Do they describe the mean, median or mode?

|  |  |
| --- | --- |
| **Description** | **Type of average** |
| The most commonly used type of average.  In the above example, to calculate it for females you:   1. Add up all the female ratings (252) 2. Divide this by the number of females (252/44) |  |
| The score that occurs most frequently. |  |
| The score that appears in the middle position if you sort all the responses from highest to lowest.  In the above example, to calculate it for males you:   1. Sort all the male scores in order (1, 2, 3, 3, 3, 3, 3, 4, 4, 5, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 9, 9, 9, 10, 10) 2. Work out the middle position in this list (item 18 of 35) 3. Take the number that is in that position (7) |  |

|  |
| --- |
| JARGON: The technical term for “type of average” is **measure of central tendency**. |

### Question 3

Below are descriptions of when you use each measure of central tendency. Which do they describe?

|  |  |
| --- | --- |
| **Description of usage** | **Measure of central tendency** |
| When the numbers in the responses are evenly spaced out – that is, there are no great outliers, |  |
| When the numbers in the responses have great outliers – for example, if you have a list of salaries in a company, the salaries of the highest executives may be outliers. |  |
| When you want to know the most common or most “popular” category. More rarely useful than the other two measures. |  |

### Question 4

Which measure of central tendency is best to use with the “Klingon” data?

|  |
| --- |
|  |

## Variance and standard deviation – measures of spread

Knowing that the mean of one group is higher than another does not tell you enough to work out whether there is a difference between the two groups.

Take a look at these averages again.

|  |  |  |
| --- | --- | --- |
|  | Male | female |
| mean | 6.20 | 5.73 |

The average figures tell us part of the story, but not all of it. We need to know: “were most of the scores clustered quite close to the average figure, or were they spread out more widely?” In our case the male scores might be more spread out than the female scores.

The figures that tell us this are the **variance** and the **standard deviation**. You can calculate both using Excel or SPSS. Both are measures of spread but for now, we will concentrate on the standard deviation.

These are the figures for our Klingon data:

|  |  |  |
| --- | --- | --- |
|  | Male | Female |
| Mean | 6.20 | 5.73 |
| Stdev | 2.31 | 2.94 |

Q: What does the standard deviation tell us about the spread of scores?

A: The calculation of the standard deviation shows us that roughly 68% of the males score within 2.31 of the mean (which is 6.20). That is, 68% of the responses fall between 3.89 and 8.51.

|  |
| --- |
| JARGON: Standard deviation tells you how spread out the numbers are. It tells us how close to the mean 68% of the them come. |

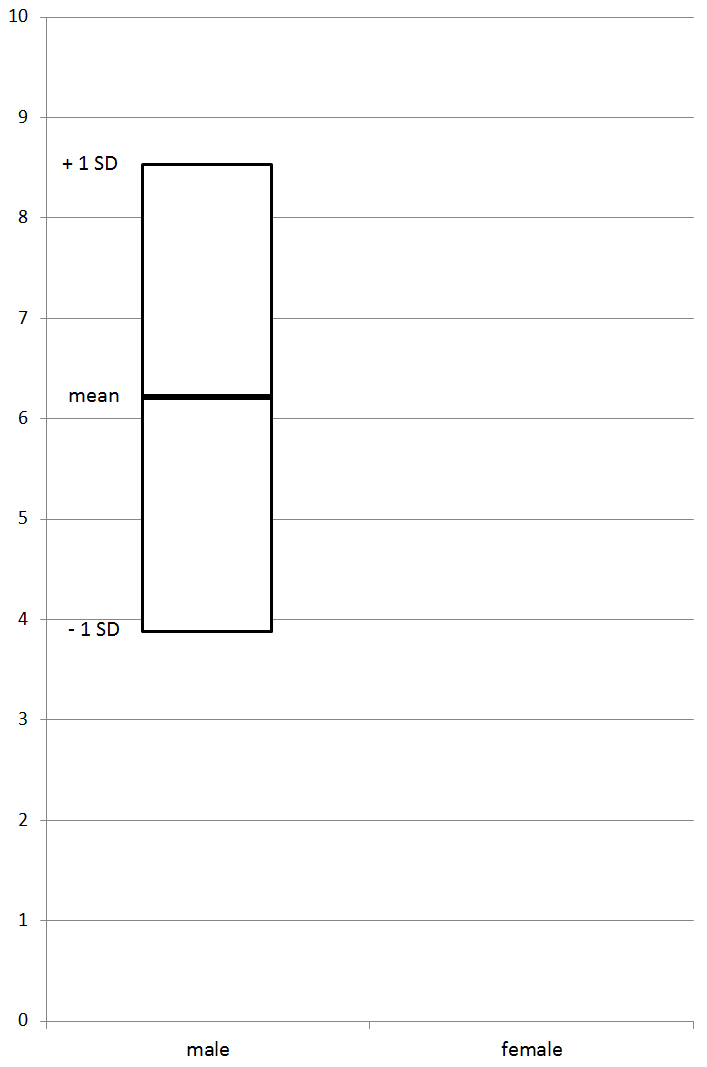
### Question 5

Use the calculator on your phone to fill in the gaps below.

|  |
| --- |
| 68% of the male ratings are between 3.89 and 8.51.  68% of the female ratings are between and . |

### Question 6

Below is a “boxplot” graph showing the mean and standard deviation for the male category of this sample. Draw in the boxplot for the female rankings.



### Question 7

Look at the boxplot graph you have just completed. What does it tell you about the difference in rankings between the males and females?

|  |
| --- |
|  |

## How significant or insignificant is a difference?

Looking at the boxplot graph, it would seem there is little difference between the males and females. We could guess that the difference is not significant.

If you wanted to confirm that there is no significant difference, you would perform a statistical test on these data – in this case, you would use a **t-test**.

How to perform a t-test and interpret the results is a topic for another day.

|  |
| --- |
| JARGON: **Significant difference** is a specific statistical term. To be able to use this term, you need to have carried out a specific statistical test. Examples of such tests are a **t-test**, a **chi-square test** or an **ANOVA** test. |

# Worksheet 3: Comparison of Film Review Corpora: do World Englishes differ?

Table 1 below is a film corpus of film reviews of the top 20 grossing films for 2011. The reviews are taken from the British newspapers *The Times* and *The Guardian*; American Newspapers *Washington Post* and *New York Times*; and Indian papers *The Times of India* and the *Indian Express*.

The first column for each newspaper gives the number of times the perfect tense is used. If there was no review for a film, then the cell is shaded in light green.

The second column gives the word length of each review.

|  |
| --- |
| Note on how we determined the perfect tense?  * 0all lexical (i.e. non-auxiliary) uses of 'have' in any variety * have got' for ownership etc. because its sense is never perfective (and it is more common in BrE than Ame anyway) * have' + NP + Verb(-ing or -ed) which are causatives * all 'have to' forms which are not perfects but a kind of modal auxiliary form |

Table 1: Occurances of perfect tense

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Film title | Times (London) | Word count  count | Guardian | Word count  count | Washington Post | Word count  count | NY Times | Word count  count | Times of India | Word count  count | Indian Express | Word count  count |
| bridesmaids | 2 | 637 | 8 | 991 | 2 | 665 | 1 | 992 |  |  | 0 | 737 |
| captain-america | 2 | 483 | 5 | 908 | 2 | 818 | 4 | 1235 | 0 | 375 | 2 | 571 |
| cars-2 | 1 | 493 | 2 | 470 | 3 | 620 | 5 | 1055 | 0 | 473 | 1 | 458 |
| fast-five | 3 | 989 | 3 | 635 | 7 | 706 | 2 | 905 | 1 | 375 | 2 | 354 |
| hangover | 1 | 624 | 6 | 895 | 3 | 721 | 0 | 784 | 1 | 434 | 0 | 737 |
| harry-potter | 4 | 646 | 2 | 536 | 6 | 894 | 16 | 1745 | 0 | 379 | 6 | 929 |
| kung-fu-panda | 3 | 665 | 0 | 272 | 3 | 521 | 4 | 882 | 1 | 390 | 0 | 513 |
| mission-impossible | 6 | 859 | 0 | 807 | 6 | 768 | 5 | 1082 | 1 | 456 | 2 | 694 |
| pirates | 2 | 410 | 6 | 875 | 5 | 792 | 4 | 969 | 1 | 371 | 1 | 469 |
| planet-of- -apes | 3 | 521 | 4 | 851 | 0 | 654 | 0 | 1180 | 1 | 593 |  |  |
| puss-in-boots | 2 | 646 | 1 | 276 | 3 | 597 | 2 | 756 | 1 | 222 | 1 | 519 |
| rio | 2 | 620 | 1 | 136 | 1 | 560 | 1 | 406 | 0 | 343 | 4 | 401 |
| sherlock | 5 | 642 | 2 | 474 | 2 | 624 | 2 | 659 | 1 | 608 | 1 | 555 |
| smurfs | 2 | 433 | 0 | 195 | 1 | 643 | 1 | 420 | 0 | 176 | 5 | 667 |
| super-8 | 2 | 632 | 2 | 990 | 3 | 796 | 3 | 1190 |  |  | 0 | 495 |
| the-help | 0 | 500 | 3 | 275 | 4 | 835 | 7 | 1308 | 1 | 300 | 9 | 591 |
| thor | 4 | 450 | 1 | 483 | 0 | 644 | 6 | 1181 | 0 | 226 | 1 | 224 |
| transformers | 1 | 584 | 3 | 491 | 4 | 821 | 4 | 1269 | 1 | 451 | 3 | 551 |
| twilight | 4 | 662 | 0 | 338 | 2 | 593 | 6 | 1241 | 0 | 661 | 4 | 519 |
| x-men | 2 | 614 | 0 | 1262 | 4 | 858 | 2 | 938 | 2 | 382 | 0 | 379 |
| total | 51 | 12110 | 49 | 12160 | 61 | 14130 | 75 | 20197 | 12 | 7215 | 42 | 10363 |

## Task 1

Work in pairs to answer these questions with reference to table 1.

1. How did we get this data?

|  |
| --- |
|  |

2. What do the numbers tell us?

|  |
| --- |
|  |

3. In what ways could you analyse this data?

|  |
| --- |
|  |

Table 2: Examples of perfect tense per 1000 words. Reviews with missing data removed across other titles.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Film title | Times (London) | The Guardian | Washington Post | New York Times | Times of India | The Indian Express |
| captain-america | 4.1 | 5.5 | 3.0 | 1.0 | 0.0 | 3.5 |
| cars-2 | 2.0 | 4.3 | 2.4 | 3.2 | 0.0 | 2.2 |
| fast-five | 3.0 | 4.7 | 4.8 | 4.7 | 2.7 | 5.6 |
| hangover | 1.6 | 6.7 | 9.9 | 2.2 | 2.3 | 0.0 |
| harry-potter | 6.2 | 3.7 | 4.2 | 0.0 | 0.0 | 6.5 |
| kung-fu-panda | 4.5 | 0.0 | 6.7 | 9.2 | 2.6 | 0.0 |
| mission-impossible | 7.0 | 0.0 | 5.8 | 4.5 | 2.2 | 2.9 |
| pirates | 4.9 | 6.9 | 7.8 | 4.6 | 2.7 | 2.1 |
| puss-in-boots | 3.1 | 3.6 | 6.3 | 4.1 | 4.5 | 1.9 |
| rio | 3.2 | 7.4 | 0.0 | 0.0 | 0.0 | 10.0 |
| sherlock | 7.8 | 4.2 | 5.0 | 2.6 | 1.6 | 1.8 |
| smurfs | 4.6 | 0.0 | 1.8 | 2.5 | 0.0 | 7.5 |
| the-help | 0.0 | 10.9 | 3.2 | 3.0 | 3.3 | 15.2 |
| thor | 8.9 | 2.1 | 1.6 | 2.4 | 0.0 | 4.5 |
| transformers | 1.7 | 6.1 | 3.8 | 2.5 | 2.2 | 5.4 |
| twilight | 6.0 | 0.0 | 4.8 | 5.4 | 0.0 | 7.7 |
| x-men | 3.3 | 0.0 | 0.0 | 5.1 | 5.2 | 0.0 |
| total | 72.0 | 66.1 | 71.1 | 57.1 | 29.4 | 76.9 |

## Task 2

Work in pairs to answer these questions with reference to tables 1 and 2

1. What is the difference between Table 1 and Table 2?

|  |
| --- |
|  |

2. How has the data been normalised?

|  |
| --- |
|  |

3. Why is normalisation important?

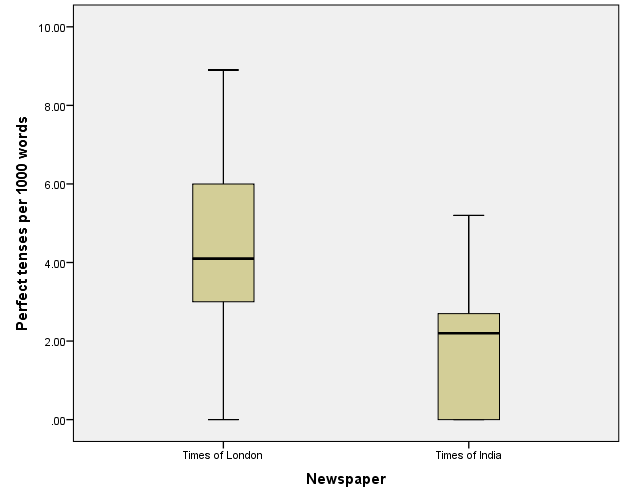
|  |
| --- |
|  |

## Statistical analysis of data from table 2

A comparison was carried out between the Times of India and the Times of London.

**First** a boxplot was created to compare the two data sets graphically.

Figure 1: Boxplot of data drawn from the film review corpus comparing the use of the perfect tense in the Times (of London) and in The Times of India.



In this standard version of the boxplot produced by SPSS the line represents ***the median*** score; the box represents the central 50% of the distribution and contains half of all of the scores for each newspaper; and the ‘whiskers’ or T-lines generally indicate the minimum and maximum figures. As such it is a useful summary of the two distributions. By examining the box plot you should be able to judge whether there is a difference in scores between the two groups.

What do you think?

**Second** some basic descriptive statistics were calculated.

Table 3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Group Statistics** | | | | |
|  | Newspaper | N | Mean | Std. Deviation |
| Perfect tenses per 1000 words | Times of London | 17 | 4.2294 | 2.38243 |
| Times of India | 17 | 1.7235 | 1.70570 |

This table indicates that across the seventeen films the mean number of perfect tenses used in the reviews in the Times of London is higher by more than two and a half occurrences per thousand words than in the Times of India. The ***standard deviation,*** which measures how spread out the distribution is, is only around about two for both groups.

Think about this.

The difference between the two means is about 2.5 but the standard deviation for each group is only about 2. This suggests that there is little overlap in the distributions of the two groups and confirms the picture given by the box plot.

**Third** an independent samples t-test was carried out to find out whether this apparent difference between the two groups is statistically significant. The results you need to interpret the t-test have been provided in the table below.

Table 4: Results from t-test

|  |  |
| --- | --- |
| t value | 3.53 |
| degrees of freedom | 32 |
| Significance or probability value | 0.001 |

In a research paper this result would be written up like this: t(32) = 3.53; p= 0.001

## Task 3

1. What does the result mean? Is the result statistically significant?

|  |
| --- |
|  |

2. Is there a statistically significant difference between British and Indian Englishes in the use of the perfect tense?

|  |
| --- |
|  |

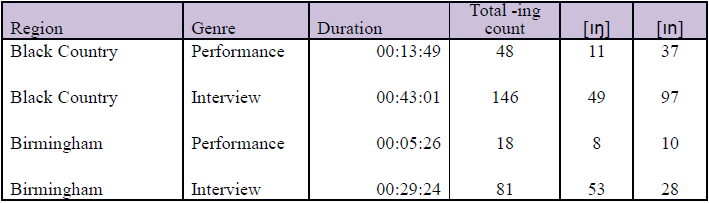
3. Do you think that this statistical result is linguistically important?

|  |
| --- |
|  |

# Worksheet 4: West Midlands ING: How do pronunciations of the progressive ING differ?

The data below shows the counts for -ing and -in in recordings of two live performances and two interviews with performers in Birmingham and the Black Country

Table 1: Raw data



## Task 1

Answer the questions below with reference to table 1.

1. Does the data tell you a story?

|  |
| --- |
|  |

2. What type of variable is ING pronunciation?

|  |
| --- |
|  |

3. Can you identify any problems when it comes to comparing the data?

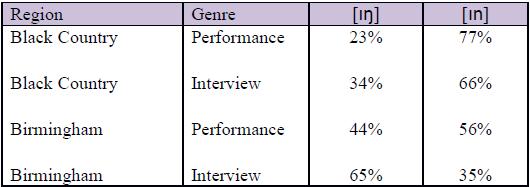
|  |
| --- |
|  |

4. How would you graph this type of data?

|  |
| --- |
|  |

In order to normalise the data we can compute the proportion of each variant by dividing the count for that variant by the total count for both variants.

Table 2: Normalised data



## Task 2

Answer the questions below with reference to table 2.

1. What does the data tell you now?

|  |
| --- |
|  |

2. Are the patterns clearer than in table 1?

|  |
| --- |
|  |

3. Are there any issues with this approach to data normalisation?

|  |
| --- |
|  |

4. How would you graph this type of data?

|  |
| --- |
|  |

## Statistical analysis of the data

In order to test if these differences in the pronunciation of ING across regions and genres are significant you can use a *Chi-square Test for Independence* by following these three steps.

**First** state the hypotheses.

Null hypothesis 1: ING pronunciation in the Black Country is independent of genre (performance vs. interview)

Alternative hypothesis 1: ING pronunciation in the Black Country is dependent on genre (performance vs. interview)

Null hypothesis 2: ING pronunciation in Birmingham is independent of genre (performance vs. interview)

Alternative hypothesis 2: ING pronunciation in Birmingham is dependent on genre (performance vs. interview)

Null hypothesis 3: ING pronunciation in interviews is independent of region (Black Country vs. Birmingham)

Alternative hypothesis 3: ING pronunciation in interviews is dependent on region (Black Country vs. Birmingham)

Null hypothesis 4: ING pronunciation in performances is independent of region (Black Country vs. Birmingham)

Alternative hypothesis 4: ING pronunciation in performances is dependent on region (Black Country vs. Birmingham)

**Second** build contingency tables.

|  |  |
| --- | --- |
| Table 3: Black Country Genre Contingency Table | Table 4: Birmingham Genre Contingency Table |
| Table 5: Performance Region Contingency Table | Table 6: Interview Region Contingency Table |

**Third** calculate Chi-Square (χ2) Test of Independence for each Contingency Table

#### Genre in the Black Country:

χ2 = 1.916, df =1, p = .116

Since p > .05 retain the null hypothesis:

There is not a significant relationship between genre and ING pronunciation in the Black Country

#### Genre in Birmingham:

χ 2 = 2.743, df =1, p = .098

Since p > .05 retain the null hypothesis:

There is not a significant relationship between genre and ING pronunciation in Birmingham

#### Region in Performance Genre:

χ 2 = 2.959 df =1, p = .085

Since p > .05 retain the null hypothesis:

There is not a significant relationship between region and ING pronunciation in the performance genre

#### Region in Interview Genre:

2 = 21.386, df =1, p < .001

Since p < .05 reject the null hypothesis:

There is a significant relationship between region and ING pronunciation in the interview genre, with Birmingham informants using more -ing than the Black Country informants.

## Task 3

Answer the questions below with reference to table 2.

1. What do these results tell you?

|  |
| --- |
|  |

2. Why is it useful to run a Chi-square?

|  |
| --- |
|  |

3. Why did we need to run four Chi-squares?

|  |
| --- |
|  |

4. Do you think this statistical result is linguistically important?

|  |
| --- |
|  |