Statistics 120
Information Visualisation

Final Grade

- The final grade will be made up as follows:
  - 15% assignments
  - 20% in-class test
  - 65% final exam
- A minimum of 45% must be obtained in the final exam

Contact Details

- Ross Ihaka
- Room 275, Department of Statistics
- Extension 5054
- Office Hours:
  - Friday, Noon – 2pm
  - By arrangement

Midterm Quiz

- Wednesday, 27nd August, 11am – 12noon.
- The test will examine all material covered up to the time of the test. More details will be given closer to the date …
- Both the test and final exam will require answers in essay form.

This Course Aims To Help You To . . .

- Understand how we see.
- Know how this constrains visualisation.
- Recognise good and bad graphs.
- Be familiar with a wide range of graph types.
- Develop computing and graphics skills.

Course Topics

- Introduction.
- Computing for graphics.
- Human vision and perception.
- Displays for counts and proportions, numeric data and time series.
- Graphics for multivariate data.
- Dynamic graphics.

The Place of Computing

- Computing and computer graphics will be of crucial importance in the course.
- The department provides laboratory facilities which you can use.
- The software used in the course is also freely available on the internet.
- CD's containing the software will be available.

Dictionary Definition

- The action or fact of visualising; the power or process of forming a mental picture or vision of something not actually present to the sight; a picture thus formed.
- The action or process of rendering visible.

Operational Definition

- The construction of images which represent important aspects of some situation or process.
- Synonyms for image are:
  - plot
  - graph
  - diagram
  - picture

A Map of Modern Europe

Why Visualise?

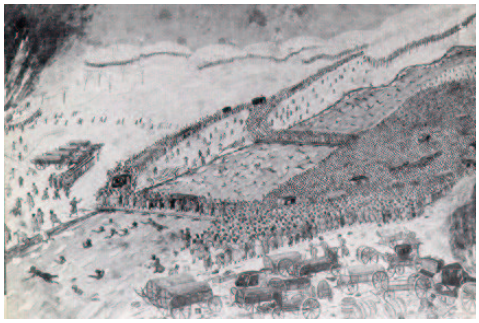
- The human visual cortex is arguably the most powerful computing system we have access to.
- Visualisation allows us to put information into a form which allows us to use the power of this computing system.
- By harnessing some of the capabilities of our visual system we can free other parts of our brains to work on problems.

Napoleon In Russia

How Visualisation Can Be Useful

- **Communication** - visualisation provides a quick way to communicate a very rich message.
- **Discovery** - visualisation provides a way of displaying a large amount of information so we can uncover new facts and relationships.
- **Insight** - visualisation provides a way to obtain better insight into things we already know.

The Bérézina Crossing



Crossing the Bérézina, 26th November. Of 40,000 men, 25,000 were lost.

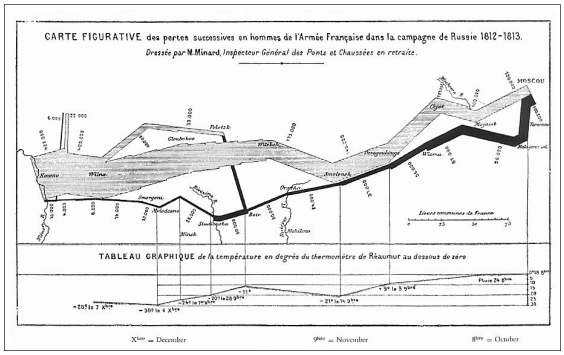
Communication: Napoleon's 1812 Campaign

- In 1812, the French Emperor Napoleon invaded Russia with army of 500,000 men. The French occupied Moscow, but had to retreat through the Russian winter.
- During the retreat, temperatures fell as low as $-30°C$. The bitter weather, together with guerrilla attacks by Russian forces, decimated the French.
- Just 10,000 men returned from Russia.

Minard's Map

- Napoleon's 1812 campaign decimated an entire European generation. Hardly a French family escaped its impact.
- Fifty years later Charles Joseph Minard, a French engineer, created a map which summarised the French experience.
- The map has a brutal eloquence and has been termed "the best graph ever produced."

The Minard Map



Choosing a Strategy

- The results of the games (winning number and winning amount) are publicly available.
- Does this data contain information which will enable us to choose a profitable strategy for this game?
- We will use the results of 254 consecutive games to look for a profitable strategy.

Why the Minard Map Succeeds

- The map describes a complex, multi-dimensional process. It shows the spatial location, marching direction and size of the invading army as well as the temperature during the retreat.
- The complexity is handled by abstracting out the most important data features and presenting them in a simple graphical form.

Winning Numbers and Amounts

(810, $190.0), (156, $120.5), (140, $285.5), (542, $184.0), (507, $384.5),
(972, $324.5), (431, $114.0), (981, $506.5), (865, $290.0), (499, $869.5),
(020, $668.5), (123, $83.0), (356, $188.0), (015, $449.0), (011, $289.5),
(160, $212.0), (507, $466.0), (779, $548.5), (286, $260.0), (268, $300.5),
(698, $556.5), (640, $371.5), (136, $112.5), (854, $254.5), (069, $368.0),
(199, $510.0), (413, $102.0), (192, $206.5), (602, $261.5), (987, $361.0),
(112, $167.5), (245, $187.0), (174, $146.5), (913, $205.0), (828, $348.5),
(539, $283.5), (434, $447.0), (357, $102.5), (178, $219.0), (198, $292.5),
(406, $343.0), (079, $332.5), (034, $532.5), (089, $445.5), (257, $127.0),
(662, $557.5), (524, $203.5), (809, $373.5), (527, $142.0), (257, $230.5),
(008, $482.5), (446, $512.5), (440, $330.0), (781, $273.0), (615, $171.0),
(231, $178.0), (580, $463.5), (987, $476.0), (391, $290.0), (267, $176.0),
(808, $195.0), (258, $159.5), (479, $296.0), (516, $177.5), (964, $406.0),
(742, $182.0), (537, $164.5), (275, $137.0), (112, $191.0), (230, $298.0),
(310, $110.0), (335, $353.0), (238, $192.5), (294, $308.5), (854, $287.0),
(309, $203.5), (026, $377.5), (960, $211.5), (200, $342.0), (604, $259.0),
(841, $231.0), (659, $348.0), (735, $159.0), (105, $130.5), (254, $176.0),
(117, $128.5), (751, $159.0), (781, $290.0), (937, $335.0), (020, $514.0),
(348, $191.0), (653, $304.5), (410, $167.0), (468, $257.0), (077, $640.0),
(921, $142.0), (314, $146.0), (683, $356.0), (000, $96.0), (963, $295.0),

Discovery: Learning About Lottery Strategies

- Forms of lotto are played world-wide and many people have theories about how to make money at the game.
- We will examine a particular lotto game, to see whether it might be possible to play it profitably.
- The game we'll look at is the daily pick-it lottery run by the state of New Jersey in the USA.

How Visualisation Helps

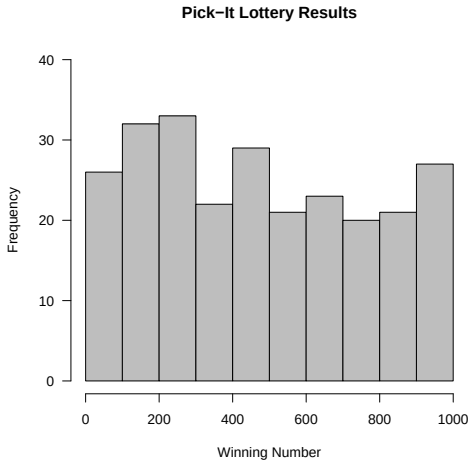
- Humans can really only make sense of three or four numbers at a time.
- By representing the values in a graphical form we make it easier to handle large numbers of values.
- Using graphs should make it possible to learn more about this data.

Playing "Pick-It" Lotto

- Each player selects a three digit number between 000 and 999.
- A winning number is selected by independently picking three digits between 0 and 9 at random.
- All players who hold the winning numbers split the prize money for the game. The size of the prize depends on the number of players who choose the winning numbers.

Choosing Good Numbers

- One approach to making money at "Pick It" is to try to select numbers which are more likely to win.
- Since we have have data on the winning numbers we can look at the distribution of the winning numbers and see whether some ranges of values are more like to produce a winner than others.
- One way to do this is to produce a histogram of the winning numbers.

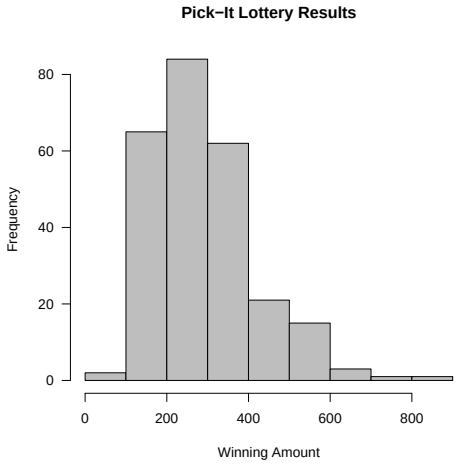


Conclusions

- There is no reason to believe that the winning values are anything other than random.
- This says that we have no reason to believe that any particular value is more likely to win than any other.
- Since we can't choose values which are more likely to win than others, we might instead see if we can influence the amount won.

Analysis

- It looks there tend to be more winners in the region from 100 to 300 than in other regions.
- This suggests that we might be best to choose numbers in this range as they are more likely to be drawn as winners than other values.
- We have to be careful about making this kind of judgement because the winning numbers are chosen randomly and we can expect to sometimes see clusters of winning number.
- To judge the significance of what we see in the histogram we must resort to formal statistical theory.



Statistical Variability

There are 254 values. We would expect the number of values in each cell to be approximately:

$$25.4 = 254 \times \frac{1}{10}.$$

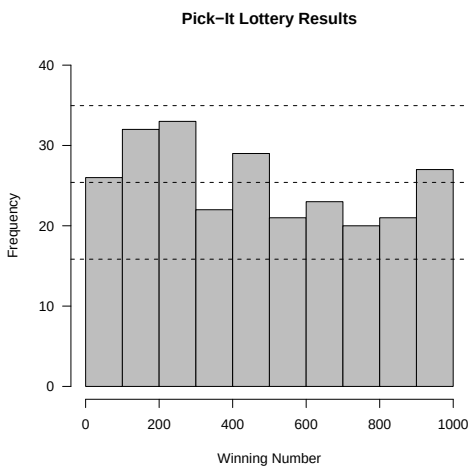
Statistical theory tells us that, if the winning ticket is chosen randomly, the level of variability in each cell is:

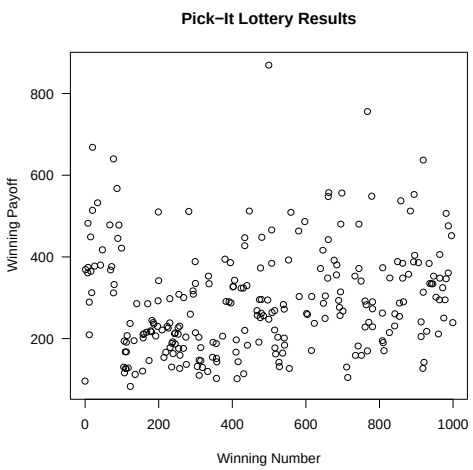
$$4.78 = \sqrt{254 \times \frac{1}{10} \times \frac{9}{10}}.$$

Discrepancies of up to twice this amount can be attributed to "random variability."

Winning Amounts

- The histogram shows that there is a wide range amounts won in the game.
- This suggests that it *might* be possible to choose the numbers which win larger amounts.
- To do thus we need to see if there is some relationship between ticket number and winning amount.
- A scatter plot is the natural way to look for such a relationship.



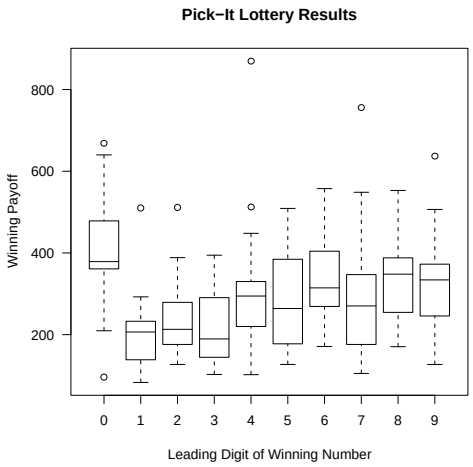


Scatterplot Features

- The winning amounts in a band to the left of the plot appear to generally be higher than those in the rest of the plot.
- We can investigate this further by separating the values into groups according to the first digit of the ticket number and drawing box plots for each group.

Choosing a "Pick-It" Lotto Strategy

- Choose numbers which are less likely to be chosen by other players.
- Then, when you win, you will tend to win more.
- Possible ways to choose:
  - Choose a number with a leading zero.
  - Choose a number with repeated digits
  - Avoid "obvious" numbers.
    E.g. 000, 123, 246, ...



Conclusions

- Tickets with a leading zero digit clearly tend to produce larger winnings.
- It is also apparent that there are some very large and some very small winning amounts.
- It is probably of interest to identify the ticket numbers corresponding to these extremes.