

Your career today is a

Marketing statistician

Leaders' notes

Do not give to the students

Red text in italics denotes comments and example answers for leaders

Equipment and preparation required for one group (3-5 students) to complete the workshop

- One printed worksheet and one pen or pencil for each student
- A computer loaded with the accompanying Excel spreadsheet
- One calculator (unless Excel can be used instead)



Today you'll be taking on the role of a marketing statistician. You will work through an exercise typical of those undertaken by marketing statisticians in their day-to-day job and you'll see how valuable the use of statistics is in marketing.

What is a marketing statistician?

Statistics is used to assess and quantify the typical level and extent of variation in customers' needs and wants. Statisticians design experiments for new products, conduct focus groups and sample surveys to gather consumer feedback, and perform field experiments in test markets to determine product viability and marketability. Statistics and data mining are also used to analyse sales data and predict future trends.

Market researchers use both government data and their own surveys to answer questions such as:

- Are consumer tastes in television programmes changing?
- What are promising locations for a new retail outlet?
- How much do people understand about our charity and how can we help them to understand more?
- If I launch this new pasta sauce, will anyone buy it?

Statisticians design the elaborate surveys that gather data for both public and private use.

See <http://www.statslife.org.uk/careers> for further details including a profile of a marketing statistician.

Today's objective

An online dating website wants to get media exposure in the run up to Valentine's Day so as to increase their market share. To do this, they decide to run a survey among their members to see what they think of Valentine's Day. They are hoping for some amusing & unusual responses that will allow them to write an article that could be published on-line or in magazines.

The survey takes place eight weeks before Valentine's Day and runs for a week. During that week, if a member logs into their profile, they will immediately be greeted with the following question which they can choose to answer or not.

'Which statement best describes how you feel about Valentine's Day?'

1. Essential to a relationship! I couldn't go out with someone who didn't celebrate it.
2. It's a nice occasion for couples to enjoy.
3. Bit of nonsense really. Doesn't mean much to me.
4. Hate it! It's for fools."

Statements 1 and 2 are positive responses. Statements 3 and 4 are negative responses. The dating site already has information on each member recorded on their profile such as age, relationship goal, job, ethnicity, education and religion.



Your job today is to analyse the data and come up with insights that can be included in an article. The data you will be investigating is real data from a real company who wanted to answer the following two questions:

- How do men and women differ in their attitudes to Valentine's Day?
- What kind of people are the most positive & the most negative about Valentine's Day?

You will have 50 minutes to answer all the following questions.

Spend no more than 25 minutes on questions 1 to 9 so that you have 25 minutes to answer questions 10 to 12. Spend the remaining 10 minutes of your time preparing a 5-minute summary of your work which you will feed back to the rest of the group.

NB Decide in your group who will write notes and who will report back at the end of the session.

1. The dating site has chosen to use a web-based survey. Can you think of two other ways they could have done this survey? Which method do you think is the best way of carrying out this survey?

The purpose of this question is to encourage students to think about the different types of surveys and the strengths & weaknesses of each type. Sample answers include:

- *Telephone survey – can sample from whole population (who have a telephone) to ensure a representative sample, allows you to check the answers respondents give. However, response rates can be poor as most people do not like having their time wasted, is also more expensive and takes longer.*
- *Face to Face survey – can sample from a population of people out on the streets, allows you to ask kinds of questions. But is the most time consuming method and costly method.*
- *Web survey – cheapest and fastest method. Can result in larger samples which allows more detailed analysis. However, sample tends to come from IT-literate population and respondents have to self-complete the questionnaire without any checking as to whether their answers are consistent.*



2. What are the problems with using their existing members to carry out the survey? Can you think of any ways that this might distort the results?

This asks students to think about sampling bias and the different ways this survey could be biased. Obvious points are

- A dating site usually excludes married people
- On-line dating is generally done by people comfortable with IT which tends to exclude older people.
- By definition a dating site includes those who are actively dating. Single people who are not interested in dating will be excluded.
- A dating site might be targeting a certain segment of the population e.g. religious people, older people, people in the military, people living in Scotland, etc or charge a high membership fee which puts off poorer people. In this case, FreeDating is a free dating site aimed at the whole of the UK and has no restrictions other than you have to be over 18. Married people are allowed on the site though their responses have been excluded from the data.

3. Look at the question that was asked in the survey and the four possible answers? Do you think this is a good way of asking the question? Can you think of a better way of asking people what they think of Valentine's Day?

This is about the issues with designing survey questions and linking the type of response (numerical/categorical etc) to the kind of analysis that can be done.

- The respondents have to choose from 4 statements. Issues that can arise when two different respondents interpret the intent behind each statement differently.
- It is possible for one statement to be non-controversial and thus become the default choice for those with no strong opinions.
- There is no middle of the road or neutral statement to choose. This is a deliberate act by the dating site and is a permissible action. However, those who genuinely have no opinion may end up choosing a statement at random or not replying at all.
- An alternative formulation that is often used in market research is to give a single statement such as "I think Valentine's Day is a cynical commercial ploy to exploit weakminded people's notions of romance" and then invite people to choose from a numerical scale how much they agree or disagree with the statement. This can make it easier to write a headline e.g. 25% of men think valentines is for feeble minded people, but respondents are then forced into evaluating something that may not be how they would express themselves.

An option that was considered by the dating site was to use emoticons instead of statements to measure people's emotional perceptions of valentines. After all Valentine's Day is an explicitly emotional occasion that might be better described using images rather than words.



Understanding the Data

In the spreadsheet that you have been given, you will see a worksheet called Data. This contains the responses from 3832 men and 2704 women. This is a lot of data so you will need to spend some time understanding what it means. Questions 4, 5 and 6 will help you do this. The first five rows of the data corresponding to the first five respondents are shown below.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	rpd_id	response	positive_resp	gender	age	ethnicity	status	religion	education	job	politics	want_child	want_serious	Weight	wt_resp
2	a00001	nonsense	0	Female	40-49	White	single	Not religious	oth	fin	strong	no	Yes	0.78	0.00
3	a00002	nice	1	Male	30-39	White	single	Other	alvl	oth	none	maybe	Yes	0.53	0.53
4	a00003	nice	1	Female	40-49	White	single	Not religious	gcse	NA	none	no	Yes	0.78	0.78
5	a00004	nice	1	Female	25-29	White	single	Christian	alvl	sal	none	maybe	Yes	0.94	0.94

Briefly familiarise yourself with each column in the worksheet Data using the descriptions below.

Column	Column name	Explanation of column
A	rpd_id	An ID code for each respondent. Using a code means you do not know the identity of the respondent. This is good market research practice.
B	response	The statement the respondent selected which best describes how they feel about Valentine's Day. 4 options were available: 'vital' = Essential to a relationship! I couldn't go out with someone who didn't celebrate it. 'nice' = It's a nice occasion for couples to enjoy. 'nonsense' = Bit of nonsense really. Doesn't mean much to me. 'hate' = Hate it! It's for fools.
C	positive_resp	If the respondent chose 'vital' or 'nice', they are deemed to be a POSITIVE respondent and marked with a 1. Those choosing 'nonsense' or 'hate' are deemed to have made a NEGATIVE response and are marked with a 0.
D	gender	The gender of the respondent: Male or Female. All respondents are heterosexual.
E	age	The Age of the respondent recorded in age bands: 18-24, 25-29, 30-39, 40-49, 50-59 and 60+. Nobody under 18 is allowed to join the site.
F	ethnicity	The ethnicity of the respondent categorised as: White, Black and Other (which includes Mixed Race, Asian, Chinese and all other ethnicities).
G	status	The marital status of the respondent: Single or sep/div/wdw for Separated/Divorced/Widowed.
H	reli	The religion of the respondent: Agnostic, Atheist, Christian, Not religious, Other, Spiritual.

I	educ	The highest level of education of the respondent: GCSE, ALVL (A level), BTEC, UNI (for university degrees), OTH (for other qualifications including who did not answer)
J	job	The respondents job (see table below for more details on how the jobs are coded)
K	poli	The respondents strength of political views: Strong views (left or right) or None for no strong views.
L	want_child	Whether the respondent wants to have children: Yes, No or Maybe.
M	want_serious	Whether the respondent is looking for a serious relationship: Yes or No.
N	Weight	A quantity (weight) given to the respondent based on Age & Gender. You will learn about this column in question 8.
O	wt_resp	A calculation done for the purposes of the pivot table. You can ignore this column and will learn about pivot tables in question 10.

Jobs are coded as follows:

Code	Explanation	Code	Explanation	Code	Explanation
acc:	Accountancy	fin:	Finance	mgr:	Manager
adm:	Administration	hlth:	Healthcare	mkt:	Marketing
cat:	Catering / Hospitality	hr:	HR/ Training/ Recruitment	NA:	No answer given
cha:	Charity	it:	IT/Computing	oth:	Other
con:	Construction	lei:	Leisure / Tourism	pub:	Public sector worker
cus:	Customer Service	lgl:	Legal	rtl:	Retail
des:	Creative / Designer	log:	Transport/Logistics	sal:	Sales
edu:	Education	man:	Manufacturing	sc:	Social Care
eng:	Engineering	med:	Media	sci:	Scientist

Open the spreadsheet and view the worksheet called Data. In order to become familiar with the information this worksheet contains answer the following questions.



4. What is the gender, job and educational level of the respondent with id number a04522?


Male, manufacturing, A levels

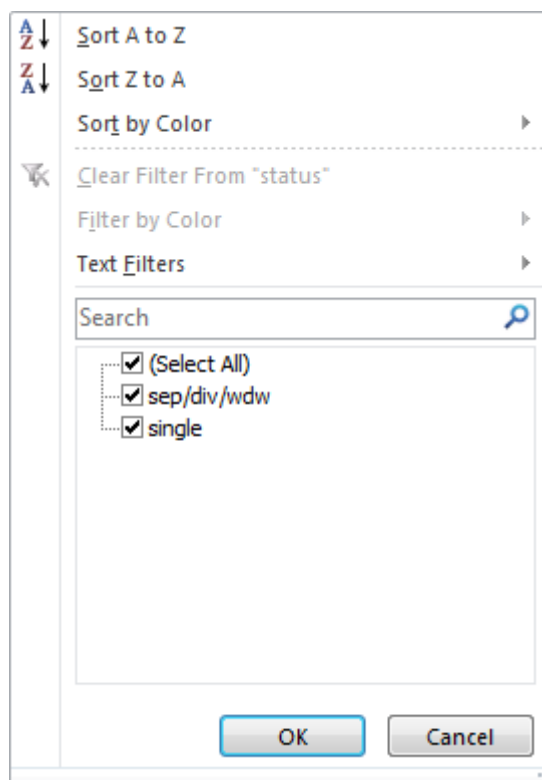
Hint: you will need to use the job codes above to decipher what their job actually is.

5. What codes would you see in the worksheet for a 37 year old, divorced, mixed race, designer?


30-39 sep/div/wdw, other, des

Hint: There might not be anyone fitting this category in this database.

On the right hand side of the header for each column, there is a “data filter” button: 



By clicking on the button a box comes up which allows you to sort or filter the data you see.

Using Status as an example, if you uncheck “(Select All)” and then check “single”, the spreadsheet only lists respondents who are single. When a filter is applied on a column the button changes to look as follows: 

Remember to always remove the filter (so you see all rows again) after answering each question.

6: Using the filter buttons in the Data worksheet, find out how many women under 25 responded to the survey with “hate”=Hate it! It’s for fools ?

8

Hint: make sure you scroll to the top of the Excel page.



The census

The census has collected information about the population every 10 years since 1801 (except in 1941). The latest census in England and Wales took place on 27 March 2011. Census statistics help paint a picture of the nation and how we live. They provide a detailed snapshot of the population and its characteristics, and underpin funding allocation to provide public services. (Source: <http://www.ons.gov.uk/ons/guide-method/census/2011/index.html>)

Table 1: Census as % of adult population

Age Group	Male	Female	Total
18-24	6.0%	5.9%	11.9%
25-29	4.3%	4.4%	8.7%
30-39	8.4%	8.4%	16.8%
40-49	9.2%	9.4%	18.6%
50-59	7.6%	7.8%	15.4%
60+	13.0%	15.5%	28.6%
TOTAL	48.6%	51.4%	100.0%

Table 2: Survey as % of total number of respondents

Age Group	Male	Female	Total
18-24	7.7%	5.7%	13.4%
25-29	9.1%	4.6%	13.7%
30-39	15.8%	8.7%	24.4%
40-49	15.1%	12.1%	27.1%
50-59	8.6%	7.9%	16.4%
60+	2.4%	2.4%	4.9%
TOTAL	58.6%	41.4%	100.0%

In Table 1, you have been given some data from the England and Wales Census. It shows how the population of England and Wales is broken down by Gender & Age.

Table 2 shows how the respondents from our Valentine's Day survey are broken down by Gender & Age.

7. Compare Table 2 with Table 1 and describe how the respondents in our survey differ from the England and Wales census? In what way could these differences cause problems in your analysis?

This introduces students to the census, why it is important and how it can be used to weight responses where responses are uneven. Students should be encouraged to learn about the census but it is not the point of this exercise. Students should be able to spot facts such as

- Men account for 59% of the survey respondents but only 49% of the whole population.
- Over 60s account for 29% of the adult population but only 5% of the survey respondents.

Students are encouraged to think how these discrepancies could be compensated for. An explicit method is demonstrated in question 8. But basically with too many men in the survey, their answers need to be given less weight whilst women need to be given more weight otherwise the survey will not be representative of the population.

In addition, There are clearly too few over 60s in the survey to really be representative. One solution is to exclude the over 60s from the analysis (which can be done in the pivot table) rather than trying to reweight them. It turns out that the final conclusions are not dependent on whether the over 60s are included or not.

We can compensate for such differences by using a process known as “weighting”. If a section of the population is under represented in a survey, then those respondents that fall in that section can be given higher “weights”. Similarly, a section that is over represented can be given lower “weights”. This process helps adjust our analysis for bias which may be due to the sample of people from whom the data are collected.

Table 3: Calculated weights associated with each gender and age category

Age Group	Male	Female	Total
18-24	0.78	1.04	0.89
25-29	0.48	0.94	0.63
30-39	0.53	0.97	0.69
40-49	0.61		0.68
50-59	0.89	0.99	0.94
60+	5.36	6.38	5.87
TOTAL Adults		1.24	1.00

Table 3 contains the weights that would be given for the Gender and Age breakdown. For example, males in the age group 18–24 represent 7.7% of the respondents in the survey but 6% of the adult population according to the census. Therefore we “weight” their results by $\frac{6}{7.7} = 0.78$.

8. Using Table 1 and Table 2, calculate the weights (to 2 decimal places) which should appear in Table 3 in highlighted empty cells.

Females 40-49: $9.4/12.1 = 0.78$

This means each female respondent in their 40s is counted as 0.78 of a person.

Adult Males: $48.6/58.6 = 0.83$

By contrast, on average adult male respondents are counted as 0.83 people.

Students should notice that because 60+ are very underrepresented in the survey, each one is counted as 5.87 people on average and that this could influence the results.

Using the Data worksheet in the Excel spreadsheet, check that subject a00001 has been assigned the correct weight according to your calculations in Table 3.

Yes 0.78 as calculated above



Explanation of Pivot Tables in Excel

This section shows the students how to use pivot tables however if they are running out of time, they can use the graphs already created in the Charts worksheet in the Excel spreadsheet.

Pivot tables in Excel allow you to quickly investigate in a table and graph the percentage of respondents summarised into categories of the other variables collected. A pivot table has been created for you in the 'Pivot' worksheet (another tab) in the Excel spreadsheet. It has been set up to display the percentage of respondents answering positively from the survey. The 'Fields to investigate' are the characteristics recorded in the survey. The first function you will learn is how to move these fields around the pivot table.

This is a blank pivot table:

%Positive	Total
Total	82%

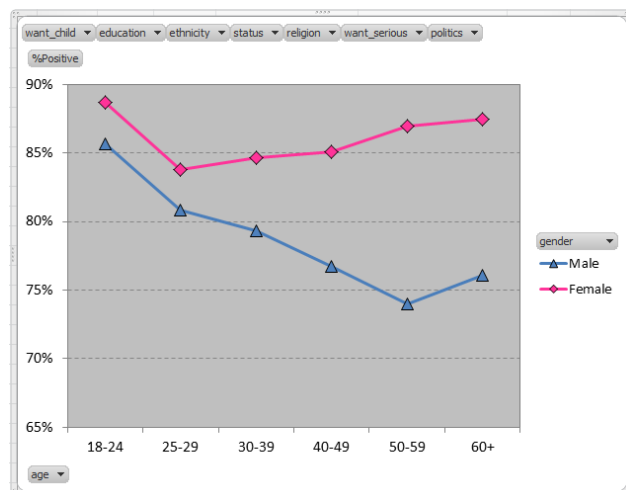
Fields to investigate	
want_child	(All)
education	(All)
ethnicity	(All)
status	(All)
religion	(All)
want_serious	(All)
politics	(All)

The table to the right has been created by dragging and dropping the age field (from the fields to investigate list) to the cell beneath %Positive (for the rows) and the gender field into the cell to the right of the %Positive (for the columns).

This informs us of the percentage of respondents choosing a positive response ("Vital" or "Nice") for each combination of age & gender.

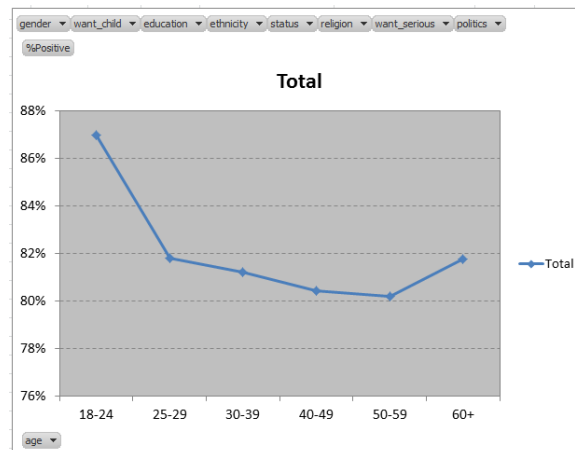
%Positive	gend		
age	Male	Female	Grand Total
18-24	86%	89%	87%
25-29	81%	84%	82%
30-39	79%	85%	81%
40-49	77%	85%	80%
50-59	74%	87%	80%
60+	76%	87%	82%
Grand Total	79%	86%	82%

The table feeds into the following graph which gets automatically updated depending on what fields you put in the pivot table.



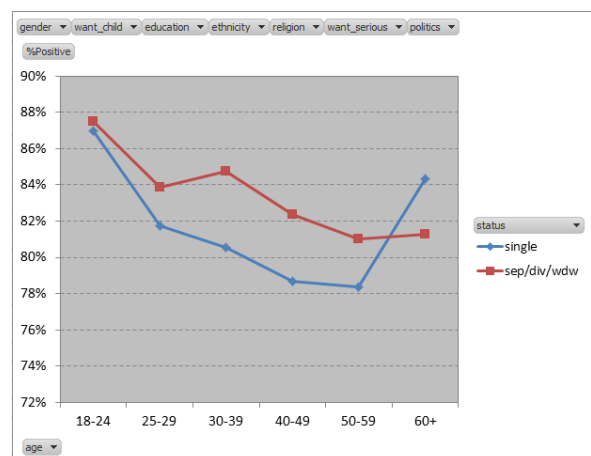
If you remove gender from the table by dragging it from the cell to the right of %Positive and dropping it anywhere in the Fields to investigate list, then the table and graph automatically change to just show Age. You can also remove Age and replace it with any other variable.

%Positive	
age	Total
18-24	87%
25-29	82%
30-39	81%
40-49	80%
50-59	80%
60+	82%
Grand Total	82%



If you now click on status in the fields to investigate list and drop it into the cell to the right of %Positive then the table and graph are updated to show marital status by age. Graph colours and symbols used for markers (+, diamond or square) can be edited in the normal Excel ways.

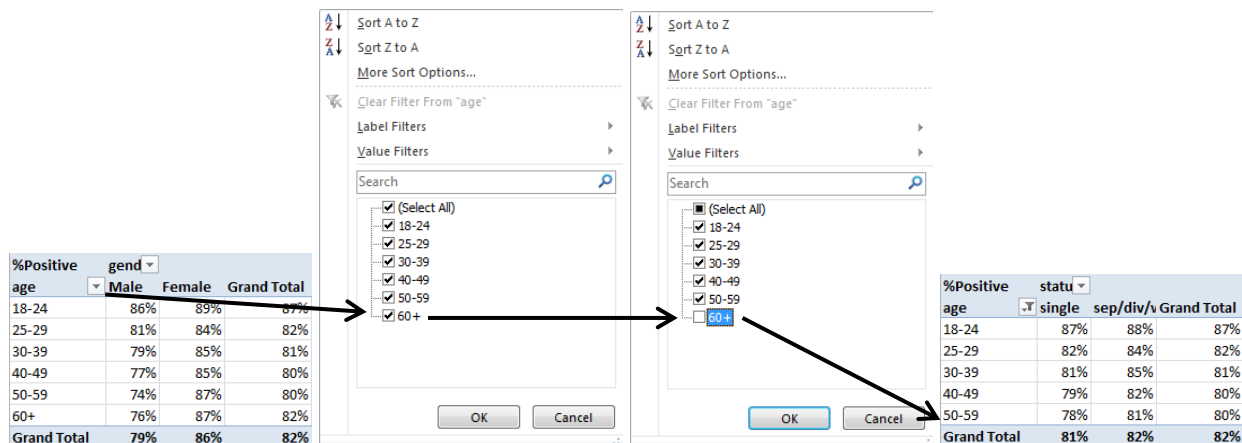
%Positive	status	
age	single	sep/div/v
18-24	87%	88%
25-29	82%	84%
30-39	81%	85%
40-49	79%	82%
50-59	78%	81%
60+	84%	81%
Grand Total	82%	82%



You can also drag gender back into the table and create a three way table (example not shown).

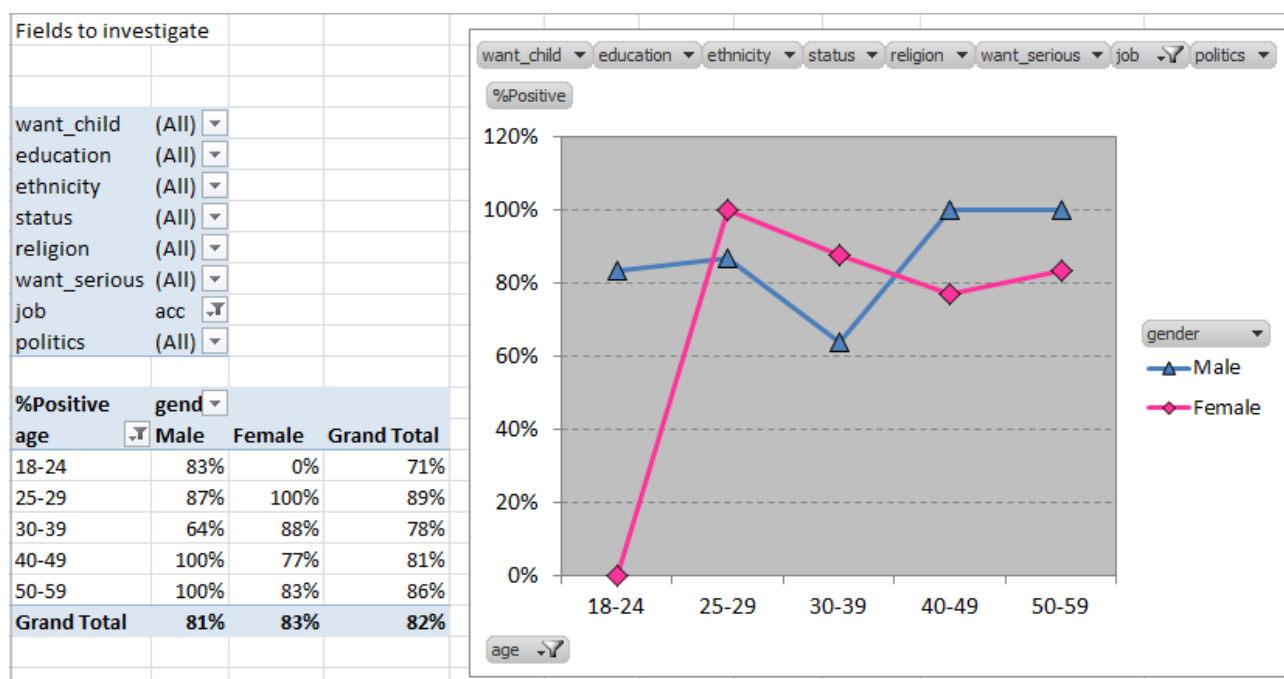
In addition to changing the fields being summarised you can also include and exclude categories within a field. Next to each field name (whether it is in the pivot table or in the fields to investigate list) is a button . If you click on this, it will display a list of categories that exist within your chosen field. Next to each category is a tick mark. Similar to the filter function on the data worksheet, you can deselect a category from the field which means that category will be excluded from the analysis.

Excluding the 60+ category from the analysis is shown below.



Note how the category disappears from the table and the graph. To bring it back, click on the age button again and select the 60+ category.

You can include and exclude multiple categories from any field. For example, the output below shows age (excluding the over 60 category) by gender only for respondents in 'acc' (accountancy). Notice you have to be careful what filters are applied as the graph does not necessarily inform you of the accountancy filter.



To add the job field, select Field List from the Options tab in the PivotTable Tools ribbon and drag job to the Fields to investigate list.

These are merely the simplest functions of a pivot table. There are many other things you can do with pivot tables. If you have time at the end of this session you can use the Help facility in Excel to find out more.

The following questions require you to use the pivot table on the pivot worksheet.

9. Compare various fields (i.e. education, age, status, politics etc.) to see how the % of men making positive responses and the % of women making positive responses are different. List 3 fields where men & women show similar behaviour and 3 fields where they show differing behaviour.

Hint: Make sure Gender is on the right of % positive and change age to be the other fields. If you are running out of time some charts have been created for you in the charts worksheet.

Similar behaviour: Education, religion, politics, want serious

Differing behaviour: Status, ethnicity, age, want child

The pivot table will be preset so that it will be very easy to change variables and interpret the results. The pivot chart facility will be preset as well. Some of the variables give fascinating results, some of which are obvious, others which aren't. Students will be encouraged to consider what constitutes a "significant" difference between categories. Some points to watch out for are:

- Analysis is focused on the % of people choosing a positive statement (option 1 or 2 from the list). The reason why is that very few people chose the 'Hate' statement (option 4).
- The scale on the pivot chart does change with each change of field. Sometimes the scale is narrow e.g. from 70% to 90%, other times it is wide 0% to 100%. This can distort perceptions of the strength of the relationship i.e. from the chart, the relationship looks dramatic but, in fact, the range on the scale is not much. Students are encouraged to look at the table and to decide what they would consider to be a significant difference when writing an article
- Overall men are less positive than women about Valentines. For some fields, this difference is consistent across the categories e.g. want_serious or education, for others the differences vary by category e.g. status, religion. Students should be looking at differences between categories within each gender as well as the differences between categories within each gender as well as the differences between gender within each category
- Analysis of the job field is discouraged. This is because there are a large number of available categories some of which overlap e.g. teacher and public sector.



10. What are the typical characteristics of MALES who are the most positive about Valentine's Day?

Question 10 & 11 involve combining the most negative/positive categories within each field (assuming those categories are significantly different from the others in the first place!) to arrive at a description of the most negative/positive persons. Students can test to see their descriptions are correct by choosing those specific categories within each field in the pivot table and seeing what % of those permutations hold positive views.

Typical characteristics were: Young, religious, single, wanting children, wanting a serious relationship with no strong political views and with less education.

11. What are the typical characteristics of FEMALES who are the most positive about Valentine's Day?

Typical characteristics were: Young or Old (but not middle aged 25-49), religious, separated/divorced/widowed, wanting children, wanting a serious relationship. No strong political views with highest qualification less than university educated.

12. What would your headline for the article be? Which chart should be the highlight of your article?

This encourages students to consider all of their results together and to identify the most salient point of their analysis. With different groups of students, it will be interesting to see how their headlines differ!

Prepare to feedback to the rest of the class a 5-minute summary of what you were tasked with today, what statistical tools you used to solve it and what your conclusions were focusing on the headline you would use in an article.

Credits

Produced by the RSS Careers in Statistics Workshop group with support from the Royal Statistical Society. Published July 2015.

All images from Microsoft Excel, created by the RSS Careers in Statistics Workshop group.

Excel spreadsheet 'Marketing Statistics Data' created by the RSS Careers in Statistics Workshop group.

