



Basic applied techniques

Choose the right stat to make better decisions

Statistical techniques help you better use resources and give credibility to your ideas

The information age has changed the way many of us do our jobs. In the 1980s, we focused on collecting data. By the early 1990s, most organizations were swimming in it. Today, we must analyze this wealth of data and turn it into useful information. Spreadsheets are a start, but they only go so far. Using a statistical package allows you to go beyond basic row-and-column math and simple summaries for better decision-making. In this paper, we will discuss how statistical techniques help you get more useful information out of your data.

Add value to decision-making

Using statistics can help analyze data to make better informed decisions because:

- **statistics are useful** – they summarize data or graphically display relationships and quickly identify unusual points;
- **statistics are easy to understand** – they are easy to read and interpret with straightforward formats, tables and charts; and
- **statistics add value to decision-making** – they allow you to quantify relationships (instead of relying on hunches) and go beyond information spreadsheets and databases provide.

By understanding basic statistical techniques, you can learn more about your data, discover interesting relationships, better use scarce resources and give credibility to your ideas.

Measure what you want to analyze

Data is a measurement of some occurrence or event. By determining your data type, you can select a suitable statistical procedure. On the flip side, to analyze data a certain way, you must collect it a certain way.

For example, if you want to measure work experience, you have a few choices. You can ask a respondent for a specific number of years (two years or five years). Or you can ask each respondent for a category (less than two years, two to five years). If your supervisor wants a graph showing how many employees are at each experience level, you could present information in a bar chart. You would choose a bar chart because it displays data collected in categories (as in Figure 1). However, if your data reflected a specific number of years for each respondent, you would probably choose another type of chart or graph.

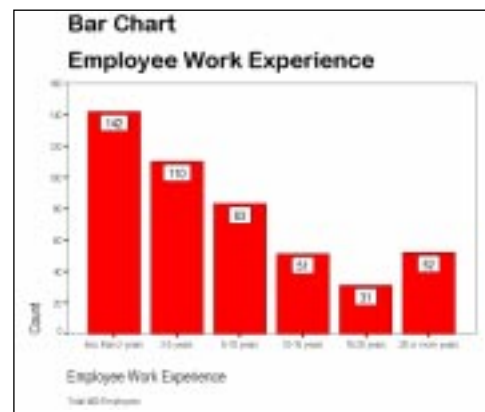


Figure 1. Bar chart displaying categorical measurement of work experience data.

Frequency distribution gives you the big picture on your data

Define levels of measurement

Levels of measurement explain how data are measured: categorically or continuously. Categorical data are measured in a group (less than two years, two to five years). Continuous data are measured with specific numbers (two years or five years). In addition, categorical data can have two distinct variables: nominal and ordinal. Nominal variables are collected in categories of group identifiers (department, business unit). Ordinal variables have implied scales (satisfaction levels, education level).

There are also two types of continuous variables: interval and ratio. Interval variables have a numeric scale without a true zero (temperature, achievement test scores). Ratio variables have a measurable scale and a true zero designating the absence of what is measured (age, net worth). With both interval and ratio variables, a one-unit increase anywhere on the scale represents the same change in quantity. Regardless of whether variables are nominal, ordinal, interval or ratio, they can help you learn more about your data.

Categorical variables

Nominal: categories with group identifiers (*gender, region, country*)

Ordinal: categories with implied scales (*military rank, satisfaction levels*)

Continuous variables

Interval: variables in a numeric scale without a true zero (*temperature, test scores*)

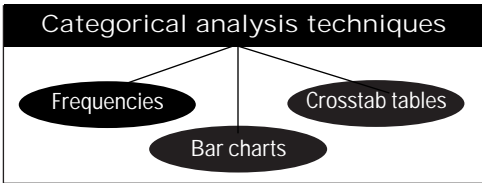
Ratio: variables in a measurable scale with a true zero (*annual income, vacation*)

When to use statistical techniques

Now that we defined levels of measurement, variable types and how they affect data analysis, let's discuss when and how to use specific statistical techniques.

Techniques for categorical data

Frequencies. A frequency distribution is useful with categorical data because it summarizes all possible responses to a question, giving you the big picture. Frequencies are also used to tabulate answers to survey questions and check for data entry errors.



For example in Figure 2, if you were curious about overall customer satisfaction with a purchased product, you could request a frequency of the "I would purchase product again" variable. This table tells you the range of re-

I would purchase product again		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Strongly Agree	10	13.5%	14.7%	14.7%
	Agree	13	17.6%	19.1%	33.8%
	Disagree	33	44.6%	48.5%	82.4%
	Strongly Disagree	12	16.2%	17.6%	100.0%
Total		68	91.9%	100.0%	
Missing	Don't Know	3	4.1%		
	Did Not Answer	3	4.1%		
	Total	6	8.1%		
Total		74	100.0%		

Figure 2. Example for frequency distribution, ordinal variable.

sponses and alerts you to the most popular response. It also allows you to draw conclusions from cumulative percentages: only 34 percent or a third of the customers who responded agreed they would purchase the product again. This is quantitative information that may support a decision to investigate why.

Customers who did not answer the question are in a separate category stats packages record called 'Missing Values.' Not accounting for missing values will cause misleading results, which is one reason stats packages are superior to spreadsheet and database packages. Stats packages are also beneficial because they allow you to produce frequencies with a few mouse clicks. Spreadsheets make you work harder for similar results.

A bar chart offers a range of answers while identifying the most frequent response

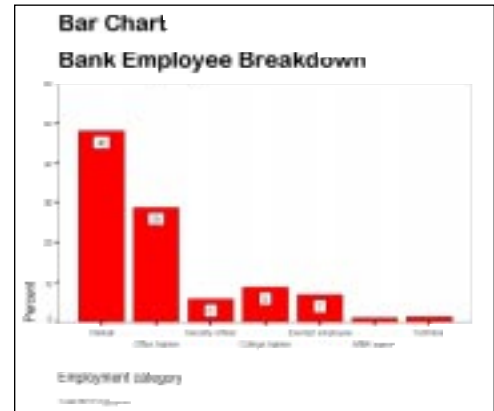


Figure 3. Example of bar chart to show overview of responses.

Bar charts. A graphic overview of a categorical variable shows easy-to-understand information. Like frequency distribution, a bar chart offers a range of answers while identifying the most frequent response. In Figure 3, the “Clerical” category is distinguished as the largest group, or 48 percent of all employees. There is an 8-to-1 ratio of clerical types to security officers. And unlike spreadsheets that make you reserve an area to receive chart results, stats packages can create frequency tables and bar charts with minimal work (one package lets you produce a frequency table and bar chart simultaneously by checking a box).

Crosstabulation tables. With crosstabulation tables, or crosstabs, you can look at the frequencies of two variables at once. Crosstabs help you determine whether variables are related, and if so, can measure the strength of that relationship. And like frequencies, you can use crosstabs to examine for data entry errors. The number of distinct categories in each variable determine the table’s size, with cells created at the intersection of each row-and-column combination.

In Figure 4, each cell gives you the actual counts and row percentages of satisfied and dissatisfied customers and whether they are willing to recommend the paper. You may want to see if a customer’s satisfaction level is a good indicator of whether the customer would

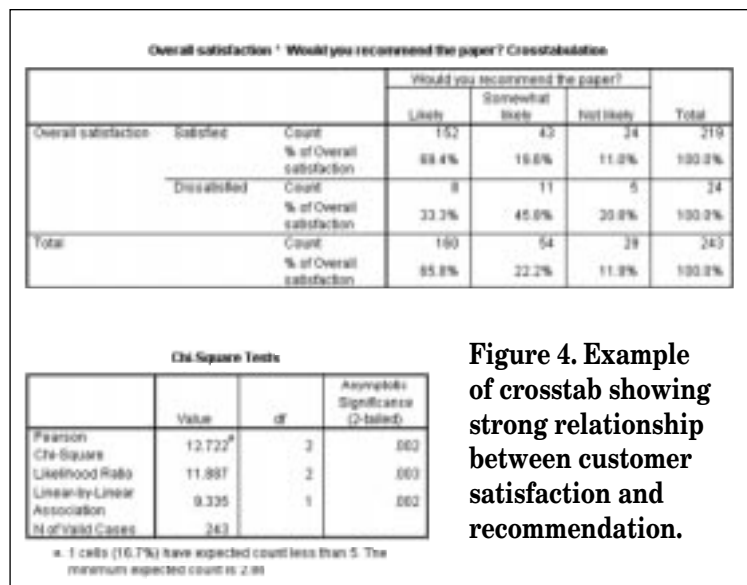


Figure 4. Example of crosstab showing strong relationship between customer satisfaction and recommendation.

recommend the paper. In this case, you are assuming customer satisfaction and recommendation are related.

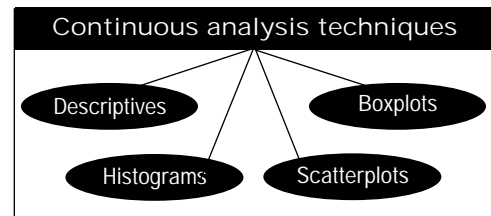
You can test this hypothesis with one of several significance tests to determine if a relationship exists. The significance tells us how probable the relationship is of occurring in a population. Usually significance is measured at .05 (five percent chance) or .01 (one percent chance), the level the relationship occurs by chance. If a significance value is .25, there is a 75 percent chance of obtaining the same results in the population. This indicates no relationship exists between the variables.

For example, in the Pearson Chi-square also in Figure 4, there is a .002 significance or a 0.2 percent level. This indicates a very small probability customer satisfaction and recommendation are not related (the smaller the significance number, the stronger the relationship). Therefore, because there is a strong relationship between satisfaction and customer recommendation, you can be confident a satisfied customer will recommend the paper.

Descriptives summarize information, helping you quickly learn more about your data as a whole

Techniques for continuous data

Descriptives. Descriptive statistics give you a quick overview of continuous variables by calculating the average, spread, minimum, maximum and number of cases for each variable. These statistics work like frequencies for categorical data and can help identify data entry errors. They are especially useful when seeing a dataset for the first time. For example, you may want to know the average age, education level, beginning salary, current salary and work experience of employees in your organization. Descriptives summarize this information, helping you quickly learn more about your data as a whole.



For example, in Figure 5 the education level values range from 8 to 21 years, with an average of 13.5 years. The standard deviation of 2.88 years gives you an idea how the data is spread around this average. Both high school (12 years) and college graduates (16 years) are included in three years either way of the 13.5-year average, so the results are not surprising.

Histogram. In our earlier discussion of levels of measurement, we considered the work experience of bank employees when measured in categories of less than two years, two to five years, etc. But instead of looking at categories, you may want to see work experience measured on a continuous scale (a specific number of years). Used with continuous data, a histogram provides an overview of distributed data values.

Basically, it works like a bar chart does with categorical data.

You can see the mean work experience of bank employees is

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Age of employee	474	23.00	64.50	37.1861	11.7872
Educational level	474	8	21	13.49	2.88
Beginning salary	474	3600	31992	6806.43	3148.26
Current salary	474	6300	54000	13767.83	6830.26
Work experience	474	.00	39.67	7.9886	8.7154
Valid N (listwise)	474				

Figure 5. Example of descriptive statistics for continuous variables.

Used with continuous data, a histogram provides an overview of how data values are distributed

eight years from the histogram in Figure 6. You can also see many employees have five or fewer years experience and several employees have 30 or more years experience. If distribution were normal, as represented by the superimposed curve, you would expect to see a bunching around the eight-year average and a symmetrical distribution on either side of this average. From this histogram, you could interpret most employees have six or less years work experience and the mean years of work experience is eight. Overall, the workers' population has fewer years of experience.

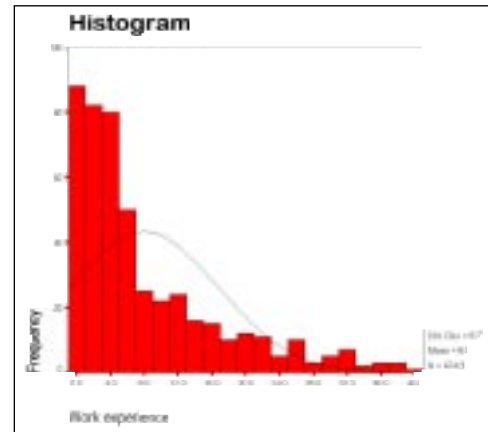


Figure 6. Example of histogram to graphically show distribution of continuous variables.

Boxplot. Again, rather than getting your data overview in table form, you can draw similar conclusions from a graph. A boxplot is a good method for identifying unusual values and gaining insight into the pattern of a majority of values. It shows more than a spreadsheet, displaying the minimum, maximum, range, average and outliers, or extreme cases.

The boxplot also shows data distribution. The whiskers show the upper and lower range of data, up to 1½ times the box length from the mean. The box shows the range of data between the 25 and 75 percentiles. Dots show outliers.

From the boxplot in Figure 7, you can see the median income – the heavy black line – does not vary much across the four groups. The median is about \$28,000, but the range of income among women who read the *Ledger Star* is larger than for those who read the *Virginian Pilot*.

This is apparent from the boxes' size. While most *Pilot* women readers make \$20,000 to \$32,000, *Star* women readers earn more, \$21,000 to \$43,000. With only a spreadsheet, you may have missed this. As a manager armed with this significant information, you could better target your advertising.

Scatterplot. Rather than limiting your investigation to individual variables, you may want to examine whether two continuous variables are related. Plotting the intersecting points on axes can quickly isolate trouble spots and identify groupings.

Scatterplots draw attention to unusual points that may affect averages. These graphs also provide a good overview of the

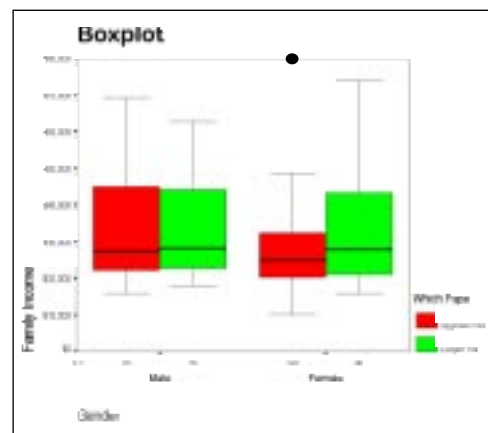


Figure 7. Example of boxplot showing newspaper readership by gender and family incomes.

Scatterplots
give you
a good
overview
of the
relationship
between
variables

relationship between variables. In Figure 8, you see a large cluster of readers (of both papers) who earn between \$15,000 and \$40,000 a year. While readers range in age from their 20s to 60s, the largest concentration of readers in their mid-20s to mid-30s. As a circulation manager knowing this, you could more precisely target your efforts.

Scatterplots also have fit lines with valuable information on differences between groups. For example, looking at the fit lines in this scatterplot we see *Star* readers have lower incomes and are younger than *Pilot* readers, who have higher incomes and are older. But when looking at the total line, we see the total population is more like the *Pilot's* population.

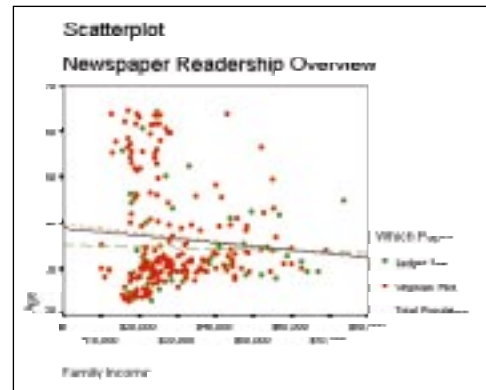


Figure 8. Example of scatterplot showing overview of newspaper readership by gender and family incomes.

Conclusion

In this paper, we discussed basic statistics and how you can use them for more effective decision-making. We considered different levels of measurement and how they affect the statistical techniques you choose to use. We then examined, in more detail, some basic statistical techniques: frequencies, bar charts, crosstabs, descriptives, histograms, boxplots and scatterplots. As demonstrated, these statistics are not difficult to understand. Using a statistical software package that offers these techniques allows you to go beyond what spreadsheets and databases packages provide. And statistics can add significant value to your everyday business decisions. With that knowledge, why not let statistical software help you work smarter?

About SPSS

SPSS Inc. is a multinational software products company that delivers statistical product and service solutions for survey research, marketing and sales analysis, quality improvement, scientific research, government reporting and education. Primary product lines include: SPSS for a variety of business solutions, SYSTAT and BMDP for scientific analysis, and QI Analyst for manufacturing and quality improvement applications. More than 2 million people worldwide use SPSS products.

Chicago-based SPSS has sales and support offices and distributors worldwide. In 1995, SPSS completed the best year in its 28-year history with total revenues of \$63 million.

SPSS software operates on most models of all major computers. It is widely used on personal computers running Microsoft® Windows® and Windows 95. Versions for the Power Macintosh® and many UNIX® platforms are also available. In addition, many products are offered in Catalan, French, German, Italian, Japanese, Spanish and traditional Chinese.

Contacting SPSS

To place an order or to get more information, call your nearest SPSS office or visit our World Wide Web site at <http://www.spss.com>

SPSS Inc. United States and Canada	+1.312.329.2400 Toll-free: 1.800.543.2185	SPSS Italia srl	+39.51.252573
SPSS Federal Systems (U.S.)	+1.703.527.6777	SPSS Japan Inc.	+81.3.5474.0341
SPSS Argentina srl.	+541.816.4086	SPSS Korea	+82.2.552.9415
SPSS Asia Pacific Pte. Ltd.	+65.3922.738	SPSS Latin America	+1.312.494.3226
SPSS Australasia Pty. Ltd.	+61.2.9954.5660 Toll-free: 1800.024.836	SPSS Malaysia Sdn Bhd	+60.3.704.5877
SPSS Belgium	+32.162.389.82	SPSS Mexico Sa de CV	+52.5.575.3091
SPSS Benelux	+31.183.636711	SPSS Middle East and Southeast Asia	+971.4.525536
SPSS Central and Eastern Europe	+44.(0)1483.719200	SPSS Scandinavia AB	+46.8.102610
SPSS East Mediterranean and Africa	+972.9.526700	SPSS Schweiz AG	+41.1.201.0930
SPSS France SARL	+33.1.4699.9670	SPSS Singapore Pte.	+65.2991238
SPSS Germany	+49.89.4890740	SPSS Taiwan Corp.	+886.2.5771100
SPSS Hellas SA	+30.1.7251925	SPSS UK Ltd.	+44.1483.719200
SPSS Hispanoportuguesa S.L.	+34.1.447.37.00		
SPSS Ireland	+353.1.66.13788		
SPSS Israel Ltd.	+972.9.526700		

SPSS is a registered trademark and the other SPSS products named are trademarks of SPSS Inc. All other names are trademarks of their respective owners.