

# The first three steps in a descriptive data analysis, with applications in IBM SPSS

Steve Simon

P.Mean Consulting

[www.pmean.com](http://www.pmean.com)

## 2. Why do I offer this webinar for free?

I offer free statistics webinars partly for fun and partly to build up goodwill for my consulting business,

– [www.pmean.com/consult.html](http://www.pmean.com/consult.html).

I also provide a free newsletter about Statistics, The Monthly Mean. To sign up for the newsletter, go to

– [www.pmean.com/news](http://www.pmean.com/news)

## 3. The next free webinar

“What do all these numbers mean? Sensitivity, specificity, and likelihood ratios”

– Wednesday, February 17, 11am-noon, CST.

Abstract: This one hour training class will give you a general introduction to numeric summary measures for diagnostic testing. You will learn how to distinguish between a diagnostic test that is useful for ruling in a diagnosis and one that is useful for ruling out a diagnosis. You will also see an illustration of how prevalence of disease affects the performance of a diagnostic test. Please have a pocket calculator available during this presentation. This class is useful for anyone who reads journal articles that evaluate these tests.

## 4. Abstract

- There are three steps that will help you get started with descriptive data analysis.
  1. Know your count, how much data you have and how much data is missing.
  2. Compute ranges and frequencies for individual variables.
  3. Examine relationships among pairs of variables using crosstabs, boxplots, and scatterplots.

# 5. Objectives

In this class you will learn how to:

- organize a plan for a descriptive data analysis,
- select appropriate summary measures for categorical and continuous data, and
- examine relationships between key variables in your data.

# 6. Sources

Part of the material for this webinar comes from:

- Stats #02: Using SPSS to Describe Your Data
  - [www.childrens-mercy.org/stats/training/hand02.asp](http://www.childrens-mercy.org/stats/training/hand02.asp)
- What is a boxplot? (October 15, 2002)
  - [www.childrensmercy.org/stats/definitions/boxplot.htm](http://www.childrensmercy.org/stats/definitions/boxplot.htm)

## 7. Very bad joke

There are three types of statisticians in the world...

## 8. Very bad joke

There are three types of statisticians in the world...

those who can count,

## 9. Very bad joke

There are three types of statisticians in the world...

those who can count,

and those who can't.

# 10. Pop quiz #1

Categorical data is data that

1. has a large number of possible values
2. has a small number of possible values
3. has missing values
4. has two possible values

# 11. Pop quiz #2

You should compute frequencies for

1. both categorical and continuous data
2. categorical data only
3. continuous data only
4. outcome variables only

## 12. Pop quiz #3

The “box” in a boxplot ranges from:

1. infinity and beyond
2. the minimum value to the maximum value
3. the mean to the standard deviation
4. the 25<sup>th</sup> percentile to the 75<sup>th</sup> percentile

# 13. Pop quiz #4

You should use a scatterplot to examine the relationship between:

1. a categorical variable and a continuous variable.
2. two categorical variables
3. Two continuous variables
4. all of the above

# 14. Categorical data

Data that consist of only small number of values, each corresponding to a specific category value or label. Ask yourself whether you can state out loud all the possible values of your data without taking a breath. If you can, you have a pretty good indication that your data are categorical.

# 15. Categorical data

In a recently published study of breast feeding in pre-term infants, there are a variety of categorical variables:

- Breast feeding status  
(exclusive, partial, and none);
- whether the mother was employed  
(yes, no); and
- the mother's marital status  
(single, married, divorced, widowed).

# 16. Continuous data

Data that consist of a large number of values, with no particular category label attached to any particular data value. Ask yourself if your data can conceptually take on any value inside some interval. If it can, you have a good indication that your data are continuous.

# 17. Know your count

You need to get a feel for how much data you have. This includes the number of subjects in your study; and the number of data values that are missing. When you have a count of the number of subjects in your study, keep that in mind when you examine any statistical procedures. If the total sample size in any of these procedures is less than your count, you may have problems with an undetected missing value.

# 18. Know your count

This seems like a simple thing, but often there are subtle details that you can't ignore. For example, the following table lists the first 10 mothers in the study.

Mother's Medical Record Number

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 224809.00	1	10.0	10.0	10.0
351110.00	1	10.0	10.0	20.0
407858.00	1	10.0	10.0	30.0
414203.00	1	10.0	10.0	40.0
423331.00	2	20.0	20.0	60.0
474640.00	1	10.0	10.0	70.0
511886.00	1	10.0	10.0	80.0
575591.00	1	10.0	10.0	90.0
590890.00	1	10.0	10.0	100.0
Total	10	100.0	100.0	

# 19. Know your count

Look for patterns in the missing values.

- Rows where every single value is missing.
- Variables where every single values is missing.
- Pairs of variables which share missing values.
- Greater number of missing values in variables measured later in time.

## 20. Compute ranges and frequencies

You should know what the maximum and minimum values are for all the important variables in your data set. If any of these are surprising, you should investigate. You should also know how many observations fall into each level of any important categorical variables.

## 21. Compute ranges and frequencies

Our outcome measure, the age when breast feeding was stopped is a continuous variable. Here is a table of statistics for this variable, including the minimum and maximum variables.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Age when bf stopped	82	1.00	34.00	16.5854	10.2415
Valid N (listwise)	82				

## 22. Compute ranges and frequencies

At first glance, the maximum value (34 weeks) seems a bit large (the study followed infants for only 24 weeks after discharge). But when I talked to the nurses involved, they explained that the length of breast feeding included the time the infants were in the hospital.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Age when bf stopped	82	1.00	34.00	16.5854	10.2415
Valid N (listwise)	82				

## 23. Compute ranges and frequencies

Also notice that the sample size for this table (82) is less than the total number of data points. This serves as a reminder that some of the data are missing for the age when breastfeeding was stopped.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Age when bf stopped	82	1.00	34.00	16.5854	10.2415
Valid N (listwise)	82				

## 24. Compute ranges and frequencies

Other tables (not shown) tell us that ranges for

- birth weights: 1 kilogram to 2.4 kilograms
- gestational age: 26 to 36 weeks.

These are reasonable values for a population of pre-term infants. Additional statistics include

- mother's age: 16 to 44 years old.

Again this is a reasonable and reassuring range of values.

## 25. Compute ranges and frequencies

Race/ethnicity is a categorical variable. Here is a table for frequencies for this variable. This table serves as a valuable reminder that this data set does not allow for reasonable comparison between black and white mothers.

Race/ethnicity

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid B	2	2.4	2.4	2.4
W	82	97.6	97.6	100.0
Total	84	100.0	100.0	

## 26. Compute ranges and frequencies

Race/ethnicity is a categorical variable. Here is a table for frequencies for this variable. This table shows that the patient population is almost exclusively white. Not only is this valuable for writing up the description of the patient population in your research paper, it also indicates that any attempt to account for race in later models is probably a waste of time.

Race/ethnicity

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid B	2	2.4	2.4	2.4
W	82	97.6	97.6	100.0
Total	84	100.0	100.0	

## 27. Examine relationships

You should have a general idea of how one variable changes as another one changes.

- For two categorical variables, we can examine this using **crosstabs**.

## 28. Examine relationships

The following is a crosstabulation of feeding type versus delivery type. Notice that I have placed feeding type as the rows of the table.

Feeding type \* Type of delivery Crosstabulation

Count		Type of delivery		Total
		VAG	C/S	
Feeding type	Bottle	21	25	46
	NG Tube	20	18	38
Total		41	43	84

# 29. Examine relationships

Sometimes these tables are easier to interpret with percentages. I selected the row percentages option to get the following table.

Type of delivery \* Feeding type Crosstabulation

			Feeding type		Total
			Control	Treatment	
Type of delivery	C/S	Count	21	20	41
		% within Type of delivery	51.2%	48.8%	100.0%
	VAG	Count	25	18	43
		% within Type of delivery	58.1%	41.9%	100.0%
Total		Count	46	38	84
		% within Type of delivery	54.8%	45.2%	100.0%

# 30. Examine relationships

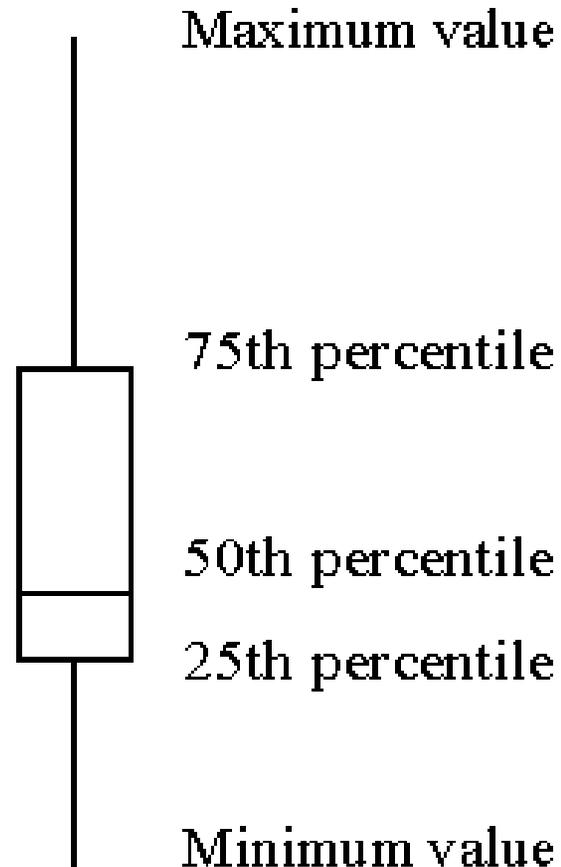
Sometimes these tables are easier to interpret with percentages. I selected the row percentages option to get the following table. We can see that there was a roughly 50-50 change for a C-section birth to find itself in the treatment or control group. In the vaginal births, however, there was a slightly greater tendency to be found in the control group. This is an imbalance which might cause problems with interpretation of the results.

Type of delivery \* Feeding type Crosstabulation

			Feeding type		Total
			Control	Treatment	
Type of delivery	C/S	Count	21	20	41
		% within Type of delivery	51.2%	48.8%	100.0%
	VAG	Count	25	18	43
		% within Type of delivery	58.1%	41.9%	100.0%
Total		Count	46	38	84
		% within Type of delivery	54.8%	45.2%	100.0%

# 31. Examine relationships

The box plot is a graphical display of a five number summary. Sometimes the box plot is also known as a box and whiskers plot.



## 32. Examine relationships

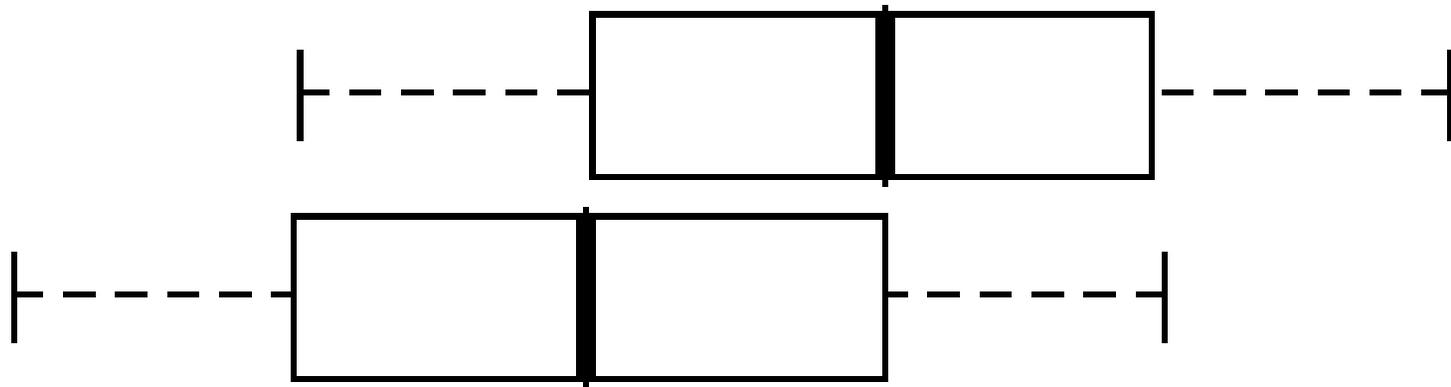
- Here are the four steps you follow to draw a boxplot.
  1. Draw a **box** from the **25th to the 75th** percentile.
  2. Split the box with a **line at the median**.
  3. Draw a thin lines (whisker) from the **75th percentile up to the maximum** value.
  4. Draw another thin line from the **25th percentile down to the minimum** value.

## 33. Examine relationships

The length of the box in a box plot, i.e., the distance between the 25th and 75th percentiles, is known as the interquartile range. You can use this box length to detect outliers. If any whisker is **more than 1.5 times as long as the length of the box**, then we have evidence of **outliers**. A common variation on the box plot is to draw the whisker to the value which is just shy of 1.5 box lengths away, and highlight each individual data point more than 1.5 box lengths away.

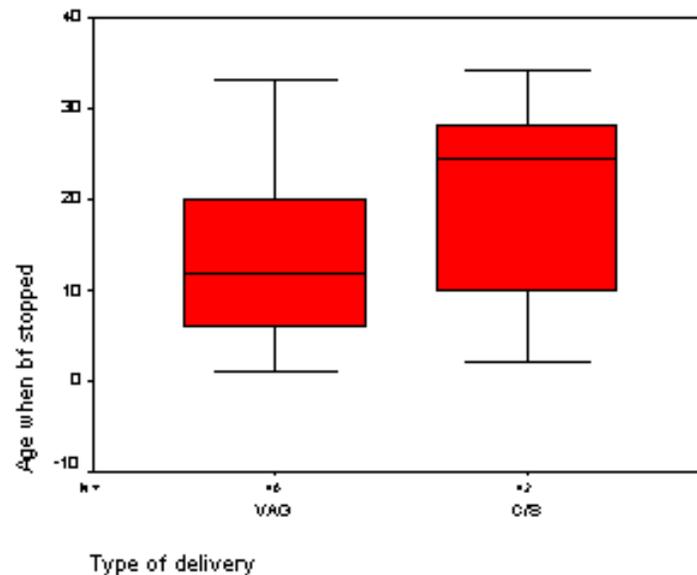
# 34. Examine relationships

The boxplot is useful for comparing the distributions of two different groups. If the median in one box exceeds the end of the box of the other group, that is evidence of a "large" discrepancy between the two groups. What passes as the median for one group would actually be the 25th or 75th percentile of the other group.



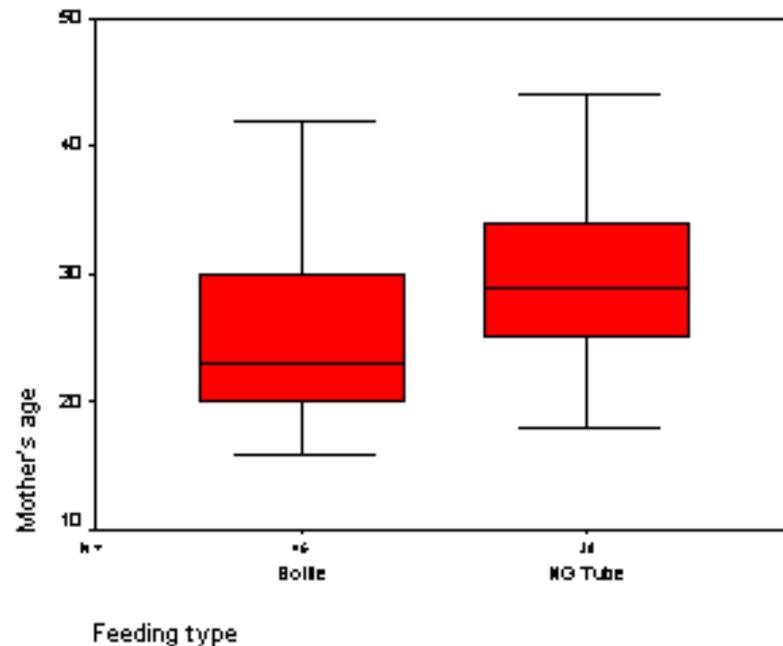
# 35. Examine relationships

Does delivery type also influence duration of breast feeding? The following box plot shows that c-section births tend to have longer durations than vaginal births, a somewhat surprising finding. Because delivery type is related to both feeding type and duration of breast feeding, we should be sure to examine delivery type as a potential confounding variable in any analysis.



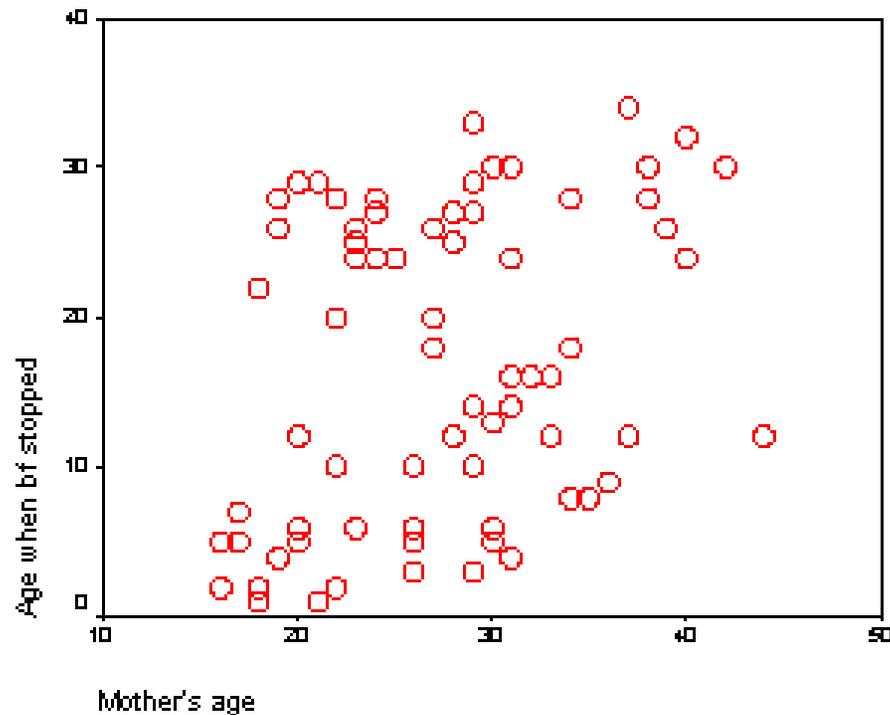
# 36. Examine relationships

Here is a boxplot comparing ages in the two feeding groups. We see that the NG tube group has older mothers than the bottle group. Further statistical analysis shows that the average age is 29 in the NG tube group and 25 in the bottle group, a difference of 4 years.



# 37. Examine relationships

We also should examine the relationship between mother's age and duration of breast feeding. The following scatterplot shows a slight tendency for older mothers to breast feed longer.



# 38. Repeat of pop quiz #1

Categorical data is data that

1. has a large number of possible values
2. has a small number of possible values
3. has missing values
4. has two possible values

## 39. Repeat of pop quiz #2

You should compute frequencies for

1. both categorical and continuous data
2. categorical data only
3. continuous data only
4. outcome variables only

## 40. Repeat of pop quiz #3

The “box” in a boxplot ranges from:

1. infinity and beyond
2. the minimum value to the maximum value
3. the mean to the standard deviation
4. the 25<sup>th</sup> percentile to the 75<sup>th</sup> percentile

# 41. Repeat of pop quiz #4

You should use a scatterplot to examine the relationship between:

1. a categorical variable and a continuous variable.
2. two categorical variables
3. Two continuous variables
4. all of the above