Print

Summary: Exploratory Data Analysis

To go to the video where you learned a task or concept, select a link.

Introduction to Descriptive Statistics

Statistics gives us a framework for describing variability in a process or system and for learning about potential sources of variability.

Descriptive statistical methods are used to describe and summarize process characteristics using numerical summaries and graphical displays.

You can use descriptive methods to understand the spread and shape of your data, characterize any central tendency of your data, and see unusual data values or patterns in your data.

Types of Data

There are three modeling types used when analyzing data: nominal, ordinal, and continuous. These modeling types are used to guide you to the type of statistical method that makes the most sense, given the type of data and the number of variables you are analyzing.

Histograms

A histogram is a picture of the distribution of the data. A distribution is the pattern that is formed by your data. You can use the histogram to understand three characteristics of the distribution: the centering (or location), the spread, and the shape. Histograms are also useful for identifying unusual patterns in the data.

Measures of Central Tendency and Location

The most commonly used measures of centering are the mean and the median. If you have data that are highly skewed, the median can be a more representative measure of the central tendency of the distribution.

The median divides the data into two parts, but we can divide the data into more than two parts. For example, we can divide the data into four equal parts, or quarters. The values that separate the quarters are called *quartiles*.

The quartiles are used to create a graph called a *box plot*.

Measures of Spread: Range and Interquartile Range

Although it's important to understand the central tendency of a distribution, understanding the spread or dispersion of a variable can be just as important.

The most common measures of spread are the range, interquartile range, sample variance, and sample standard deviation.

A box plot is an efficient graph of range and interquartile range.

Measures of Spread: Variance and Standard Deviation

The sample variance, s^2, is a measure of the variability of your data around the mean.

The variance is the average of the squared differences between each observation and the sample mean.

Instead of using the sample variance, it can be easier to interpret the sample standard deviation, *s*. This is calculated by taking the square root of the sample variance. As a result, the sample standard deviation is in the same unit of measure as the variable of interest. This makes the standard deviation much easier to interpret.

Visualizing Continuous Data

When you have two continuous variables, a scatterplot shows you the relationship between these two variables. When you have many variables that you want to explore at the same time, you can create a matrix of scatterplots.

Most of the data that you collect will have a time dimension. You might be interested in examining the performance of the process over time. In this situation, you use a *run chart*, also known as a *line graph*.

With a line graph, the variable of interest is plotted on the Y axis and the time-ordered variable is plotted on the X axis.

Describing Categorical Data

For categorical variables, the descriptive measures are largely based on frequencies.

You can use linked bar charts and histograms to explore potential relationships between variables.

When you want to look at many variables at one time, you can use tabular summaries of the data.

A contingency table is an efficient way to summarize all this information in one table. In this contingency table, you see the counts (or N), the column percent, and the row percent. You also see cumulative totals for both variables, the percent for each cell out of the total, and the percent for each level of a variable out of the total.

You can graphically describe the relationship between two categorical variables in several ways.

Side-by-side and stacked bar charts are efficient graphs, as well as a mosaic plot, to show the percent of observations that meet a given criterion. You can add counts or percents to all these graphs to help interpret what the graph is telling you and to better communicate the story in the data.

Samples and Populations

When you analyze data, you usually assume that values in a data set are a sample drawn from a larger population. A *population* is the complete set of observations, or the entire group of objects, that you are interested in understanding.

You are ultimately interested in understanding the population, but you typically can't measure every member of the population. Instead, you select a sample of observations from the population and use this sample to estimate characteristics of the population.

Because you're using sample statistics to estimate population parameters, the sample should be representative of the population. That is, the characteristics of the sample should be similar to the characteristics of the population. Using sample data that are not representative of the population leads you to draw incorrect conclusions about the population characteristics.

Simple random sampling can help ensure that the sample is representative of the population. With simple random sampling, every member of the population has an equal chance of being selected for the sample. A result of simple random sampling is that the sample statistics are good estimates of the population parameters.

Understanding the Normal Distribution

For continuous data, a histogram is the best way to see the distribution of sample data. When you graph sample data using a histogram, you are usually trying to understand or estimate the shape of the distribution of the underlying population. That is, you want to understand the underlying *continuous probability distribution*.

The normal distribution is one of the most common reference distributions in statistics, with many useful mathematical properties. One property of the normal distribution is that the shape depends entirely on two independent values: the mean (μ) and the standard deviation (σ). The population mean (μ) is the center of the distribution, and the standard deviation (σ) measures the spread or dispersion of the distribution. The larger the standard deviation, the wider the distribution.

Approximately 68% (or 68.27%) of the area under of the normal curve falls within +/- 1 standard deviation of the mean. This means that approximately 68% of the data values for a variable that is normally distributed will fall within +/- 1 standard deviation of the mean. Approximately 95% (or 95.48%) of the area under the normal curve falls within +/- 2 standard deviations of the mean. And 99.73% of the area under the normal curve falls within +/- 3 standard deviations of the mean. It is very unlikely that you will observe a value that falls more than 3 standard deviations from the mean, unless something unusual has happened.

Because the normal distribution has many useful mathematical properties, statistical procedures often assume the normal distribution.

Checking for Normality

How can you tell whether your data are approximately normally distributed?

A histogram is the best way to see the distribution of sample data. But histograms can be a bit misleading when it comes to understanding the underlying distribution. Even if your data are known to come from a normal distribution, if you have a small sample size, the data might not appear to be normal.

An alternative is to use a normal quantile plot. A normal quantile plot is also called a *QQ plot* or a *normal probability plot*. A normal quantile plot displays your data, which have been sorted in ascending order, plotted against the percentiles of the standard normal distribution.

If the distribution is approximately normal, the points in the normal quantile plot fall in a straight diagonal line, with no obvious nonlinear patterns.

The Central Limit Theorem

The *central limit theorem* makes the normal distribution particularly useful for statistical analyses.

When you analyze data, you are analyzing one sample of data from a population. If you had selected a different sample, your computed sample statistics would be slightly different. The central limit theorem enables you to understand the behavior of different random samples drawn from the same population.

Suppose you have a variable, **X**, that has a distribution with population mean μ and population standard deviation σ. You draw many samples of size n from this distribution. The central limit theorem tells you that the distribution of the ***means of these samples*** becomes more normal as the sample size increases. This distribution of sample means is centered at true mean, μ, and has population standard deviation, $\sigma/\sqrt{n}$, that becomes smaller as the sample size increases. The estimate of the standard deviation of sample means, $s/\sqrt{n}$, is called the *standard error of the mean*, or simply, the *standard error*.

The central limit theorem explains why we often use sample averages in statistics.

- We get more precise estimates of population behavior by taking samples. Larger samples give more precise estimates than smaller samples.
- The standard deviation of the sample mean is always less than the standard deviation of the raw data (by a factor of the square root of the sample size).
- Sample averages are approximately normally distributed, even if the underlying distribution is not normal, and as the sample size increases, this distribution becomes more normal.

Introduction to Exploratory Data Analysis

The process of using numerical summaries and visualizations to explore your data and identify potential relationships between variables is called *exploratory data analysis*, or *EDA*.

Exploratory data analysis is an investigative process in which you use summary statistics and graphical tools to get to know your data and understand what you can learn from it.

With EDA, you can

- find anomalies in your data, such as outliers or unusual observations,
- uncover patterns in your data,
- understand potential relationships between variables, and generate interesting questions or hypotheses that you'll test using more formal statistical methods.

Exploratory data analysis is like detective work. You're searching for clues and insights that can lead to the identification of potential root causes of the problem that you are trying to solve. You explore one variable at a time, then two variables at a time, and then many variables at a time.

Exploring Continuous Data: Enhanced Tools

Graphical tools for exploring continuous data can be enhanced by

- adding colors and different markers to help you see the patterns in your data
- using column switching to enable you to easily explore many variables without having to re-create your analysis each time
- using a row legend.

Pareto Plots

Bar charts are an efficient tool for visualizing the frequencies of categorical data. To make it easier to identify the largest defect categories, you can sort the data in descending order of the frequency of occurrence.

A *Pareto plot* is a sorted bar chart that also displays a curve for the cumulative frequency or cumulative percent. This curve helps you identify the top few issues that account for the majority of the problems.

The Pareto plot, or Pareto chart, is based on the *Pareto principle,* also known as the *80-20 rule*. This rule states that approximately 80% of the consequences result from 20% of the causes.

Sometimes the biggest problem isn't necessarily the problem that should be addressed first. You need to consider the cost of the problem and the importance to the business and the customer. Also, you should keep in mind that some problems might be much easier to tackle than others.

As with all analyses, your knowledge of the product, the process, and the business environment, along with what the data say, should guide your decision making.

Packed Bar Charts and Data Filtering

Sorted bar charts and Pareto plots are useful when you want to explore categorical variables with many levels. Pareto plots were developed to make it easier to focus on the biggest issues, or signals, through the noise of all the smaller issues.

A more modern and flexible alternative to Pareto plots is *packed bar charts*.

When you change the structure and add labels, the packed bar chart conveys essentially the same information as a Pareto plot. But a packed bar chart is more compact and efficient than a Pareto plot, and there is less unused white space in the graph.

You can use a data filter to stratify a graph or an analysis by the values of other variables in the data set.

Tree Maps and Mosaic Plots

An alternative for graphing two categorical variables with many levels is a *tree map*. In a tree map, the rectangles can be sized by the magnitude of one variable, and they can be colored by the values of another variable.

Using Trellis Plots and Overlay Variables

Visual tools for exploratory data analysis are particularly useful when you have complex or multidimensional data. You might have many observations, many variables, many categorical variables with many levels, many measurements taken over time, or data with a geographic dimension.

You can explore time-ordered data using run charts. When you have many variables, and many levels of a categorical variable, you can efficiently graph these data by overlaying variables and using trellis plots.

You can use an overlay variable to plot lines or curves for levels of a categorical variable on the same graph.

In a trellis plot, you create a matrix of graphs, where each graph shows a subset of the data. You can use trellis plots anytime that you want to create a series of plots for multivariate data—that is, when you want to graph many variables at one time.

Bubble Plots and Heat Maps

Bubble plots are effective for visualizing multivariate data when motion or animation can help communicate the message in your data. A bubble plot is like a scatterplot, but with extra dimensions.

An alternative to run charts, trellis plots, and bubble plots for visualizing multivariate data that is time ordered is a *heat map*. In a heat map, the values are represented as colors.

Visualizing Geographic and Spatial Data

Many data sets include geographic or spatial information.

- You might have ZIP codes, country codes, or city names.
- You might have latitudes and longitudes.
- Or you might have data that can be plotted using x and y coordinates. For example, the coordinates might map to positions within parts or objects.

When you have geographic data, the most effective graphical tool is a *geographic map*. Plotting your data on a map can often reveal geographic patterns that would otherwise be difficult to identify or understand.

When you have the names of the countries (and ISO codes), a geographic map can be created using shape files.

Shape files map the names of countries to latitudes and longitudes, enabling you to easily create geographic maps. Shape files can also be used to create custom maps.

Introduction to Communicating with Data

In *data presentation*, your goal is to effectively communicate what you have learned from your data.

In data presentation, you need to evaluate the effectiveness of your visualization, define your target audience, and customize your graphs to make them more effective.

Creating Effective Visualizations

With an effective visualization, the results of your analysis can be communicated in a clear and concise manner. A poor visualization can confuse your audience. Or your audience might misinterpret your intended message.

A visualization is "a graphical representation of your data." A visualization might consist of one chart or graph, or it might be more complex. For example, you might have a visualization that includes many small graphs. We use the terms *visualization*, *chart*, and *graph* interchangeably.

Evaluating the Effectiveness of a Visualization

What is the practical question we are attempting to answer with this graph? The question should be well defined, and it should be aligned with your research question. The question should also be interesting or compelling. That is, there should be a good reason for creating the graph to begin with.

Do you have the right data? The data that you compile should be aligned with the practical question that you are addressing. That is, the data should enable you to answer your question. If you can't answer the question with your data, then you don't have the right data.

Designing an Effective Visualization

Ultimately, the effectiveness of your visualization in communicating your message is determined by your audience. It's easy to get excited when you create a beautiful or fancy visualization, but the ***best*** graph is often the one that is the easiest for your audience to understand.

If your audience can't see the message, then the visualization isn't effective.

Communicating Visually with Animation

Adding animation to your graphs can be effective when motion or animation helps communicate the message in your data. For example, you can create a geographic map with an animated filter, click **Play**, and see how the map changes over time.

Are animated graphs effective in answering your question? This depends largely on your audience: who your audience is, what they know, and what you want them to know. It also depends on the communication channel you will use to share the visualization.

Designing for Your Audience

Visualizations are developed to answer a question and communicate a message. On the receiving end of this message is your audience.

The audience might be you. When you're doing exploratory data analysis, you use graphical displays to help you get to know the data, to look for patterns or relationship between variables, or to make discoveries.

When the audience is someone else, visualizations are used for data presentation to share a discovery, to inform, or to persuade.

The best visualization depends largely on your audience. If you are creating the visualization for your own use, then the graph can be simple, with minimal labeling or customizations. The default settings used by the software might be okay.

However, if your visualization is intended for someone else, then it should be designed to effectively communicate your message to your target audience. You might change the default settings, apply customizations, or add annotations to enhance the graph and your message.

Understanding Your Target Audience

To ensure that you develop the best visualization for your target audience, you need to ask three questions: Who? What? How?

Perhaps the most important of these questions is ***Who?*** Your visualization should be designed to communicate a specific message to your audience. You need to know who your audience is, and who it is not.

Understanding who your audience is can help you determine the amount of detail to include and the language to use. The same graph might not work for everyone. In other words, you might need to customize your message for different audiences.

Designing Visualizations: The Do's

You might want to use customizations that make it easier for the reader to interpret the graph, and you might want to change the color scales.

When you're selecting a color scale, it's important to keep in mind accessibility. Some people have a hard time distinguishing between certain colors. It's also important to consider what your graph will look like if it's printed without color ink.

It is likely that you will need to present your findings or conclusions to others – to your manager, your project sponsor, or other stakeholders. You might develop a few PowerPoint slides to help organize your story, but your visual displays of data are the jewel and centerpiece of your presentation.

If you are giving a live presentation, you might be tempted to use static graphs or images. Depending on your audience, you might want to use interactive visualizations instead (with the static graphs as a backup).

Designing Visualizations: The Don'ts

Don't clutter a graph with ink that doesn't provide information.

Don't use pie charts if you can help it, and don't use 3-D charts to plot data that have only one dimension.

Don't use a secondary Y axis to plot two variables with different value ranges on the same graph unless you provide annotations explaining the axes.

Don't try to say everything in one graph. You might want to use multiple graphs if you have different messages that you want to communicate.

Introduction to Saving and Sharing Results

If you need to share your analysis results, you need to consider who your audience is, what you want your audience to know or do with the results, and how your audience will see or receive the results.

There are two aspects to the question of ***How?***: which communication channel you will use, and which form or format you will use to share the results.

There are many possible communication channels. You might send your results in a report or slide deck, write a blog or post results on a website, share your results in a recorded video or a live webinar, or provide a live, face-to-face presentation.

Your results can be shared in a variety of formats that align closely with your communication channel.

Saving and Sharing Results in JMP

Why are you saving and sharing your results? Are you saving your results for your own use, to save your work, or to make it easy to repeat your analysis later? Or are you collaborating with others who need to be able to access your data, and your analysis results, in JMP?

The four most popular formats for saving and sharing results directly in JMP are scripts, journals, projects, and dashboards.

Saving and Sharing Results outside of JMP

If you want to share results with people who don't have JMP, or if you don't want them to access your data or your results directly in JMP, you can still share your results in an interactive format that can be accessed outside of JMP.

You can share interactive web reports, and you can also publish your results to JMP Public.

If you cannot make your data accessible to others, or if you don't need to use an interactive format, then you can share your analysis results as static output.

If you want to share results, and movement or motion helps tell your story, then you can save your results as an animated GIF file. You can easily share these files with others. All they need is a browser or an application that displays animated GIFs, such as PowerPoint.

Deciding Which Format to Use

With so many available formats for saving and sharing your work, how do you decide which to use?

This is based, in part, on whether you can share your data, and whether you want others to access the data in JMP.

You also need to consider who your audience is and what you want them to know (or do) with the results.

Of the methods you have been introduced to in this lesson, the most popular and perhaps easiest to use are saving scripts to data tables, saving static output, saving journals, and publishing reports to JMP Public.

Data Tables Essentials

For most statistical analyses, your data set must have a basic structure, consisting of rows and columns. The observations are stored in rows, and the variables are in columns.

An observation is the information about the basic unit of interest. The observational unit, for example, might be a customer, a batch, a part, or a transaction. A variable is a characteristic that can be measured or recorded for an observation.

In any project, a majority of your time might be spent in collecting, compiling, and preparing your data.

Common Data Quality Issues

Common data quality issues include incorrect formatting, incomplete data, missing data, and dirty or messy data.

Identifying Issues in the Data Table

When identifying potential data quality issues, you start by looking at the data table to make sure it is properly formatted. Examine the number of variables and observations. Examine the modeling type of each variable and determine whether it needs to be changed. You might also need to create new variables.

Identifying Issues One Variable at a Time

After scanning the data table for potential issues, you look at the variables one at a time. You use summary statistics and graphical summaries to get familiar with the data. At the same time, you look for potential data quality issues.

For continuous variables, you're interested in summary statistics, such as the mean, the standard deviation, the minimum and maximum values, and the number of missing observations. You're also interested in the shape of the distribution.

Is the distribution more or less symmetric? Is it skewed? Are there clusters of data or severe outliers? Are there values that aren't physically possible?

For categorical data, you're interested in the number of categories (or levels), the number of observations in each category, and the number of missing observations.

Restructuring Data for Analysis

For most statistical analyses, your data set must have a data grid with the observations stored in rows and the variables stored in columns.

Your data might be stored in a wide, or *split*, data table format.

In order to analyze these data, you need to stack the data so that the labels appear in one column and the data values are in a separate column. This is called a *tall*, or *stacked*, format.

There are other ways that you might want to reshape your data for analysis.

For example, you might want to create a new data table from a subset of your data, sort your data, or transpose the rows and columns in your data table.

Combining Data

What if your data are not stored in the same file, or if you need to add new data to an existing data table?

When two files have the same basic structure and the same variables, you can concatenate the two data tables. You can either create a new data table with the information from both tables, or you can add the new data to the bottom of your original table.

You can also join two tables that have a common variable by matching the rows.

Deriving New Variables

Sometimes, the information that you need is in your data table, but you have to do a bit of work to get to it.

For example, you might create a formula to determine the scrap rate, by dividing the number of parts scrapped by the batch size. Scrap rate is a derived variable, because it was computed using data in the data table.

You can create variables using IF-THEN statements or by using data transformations.

Formulas can also be used to extract information out of a column in the data table.

Working with Dates

Most statistical packages store dates as the number of time units since some reference time. In JMP, dates are stored as the number of seconds since midnight, January 1, 1904.

If you work with data that have variables for dates or times, you should use the correct date or time format for these columns.

Columns with date and time formats are stored as numeric continuous data in JMP. This enables you to calculate elapsed times and extract information from the date columns.