

5/10/2015

TEAM PLUS

Nick D. wondergem, upadhyay ishant, kichukov veselin, velev dan, draganov dragan, ngabo francky

EDA report

ORIKAMI

Document history

Revisions

|  |  |  |  |
| --- | --- | --- | --- |
| Version | Status | Date | Changes |
| 0.1 | concept | 05-10-2015 | Template created |
| 0.2 | concept |  |  |
| 0.3 | concept |  |  |
| 0.4 | concept |  |  |
| 1.0 | draft |  |  |

Approval

This document needs to be approved by:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Version | Approval date | Name | Function | Paraph |
| 1.0 |  | Theo Cats | Tutor |  |
| 1.0 |  | Bram den Teuling | Product owner |  |

Distribution

This document has been sent to:

|  |  |  |  |
| --- | --- | --- | --- |
| Version | Send date | Name | Function |
| 1.0 | 28-09-2015 | Theo Cats | Tutor |
| 1.0 | 28-09-2015 | Bram den Teuling | Product owner |

Contents

[Introduction 3](#_Toc441594942)

[Descriptive statistics 4](#_Toc441594943)

[Correlations 8](#_Toc441594944)

[t-Test ( dependent and independent ) 9](#_Toc441594945)

[Breakdowns 11](#_Toc441594946)

[Frequency tables 11](#_Toc441594947)

[Appendix A - Data structure 13](#_Toc441594948)

# Introduction

The goal of EDA is to discover patterns in data. John Tukey often likened EDA to detective work. The role of the data analyst is to listen to the data in as many ways as possible until a plausible "story" of the data is apparent, even if such a description would not be borne out in subsequent samples. Finch asserted that "we claim for exploratory investigation no more than that it is an activity directed toward the formation of analogy. The end of it is simply a statement that the data look as if they could reasonably be thought of in such and such a way"

Exploratory data analysis or “EDA” is a critical first step in analyzing the data from an experiment. The main reasons we use EDA:

• detection of mistakes

• checking of assumptions

• preliminary selection of appropriate models

• determining relationships among the explanatory variables, and

• assessing the direction and rough size of relationships between explanatory and outcome variables.

# Descriptive statistics

Description Mean and confidence intervals

The mean is a particularly informative measure of the "central tendency" of the variable if it is reported along with its confidence intervals. The confidence intervals for the mean give us a range of values around the mean where we expect the "true" (population) mean is located.

Purpose

to derive the central tendency of the data

Activities

- find individual artists

- scan database for tweets with target artist/song combination

- create plot (x = day | y = #mentions)

Results

Scanned the first 150.000 documents from mozart.tweets, selected on having an artist and songtitle.

Within these 150.000 documents there were 25640 unique artists identified, this includes some dirty values.

|  |  |  |
| --- | --- | --- |
| Artist | Song | Number of tweets |
| Wings | Aku Pun Tahu | 58 ( in 4 weeks - reply in 2082 seconds ) |
| Chris Brown | Post to be | error on timestamp |
| Teresa Jenee | Remember (Kaytronic dub) | 13 ( in 4 weeks - reply in 1913 seconds) |
| 98 Degrees | Because of You | 40 ( in 4 weeks - reply in 2926 seconds \*) |
| Mark Ronson | Uptown Funk | CursorNotFound: cursor id '113091073817' not valid at server |
| spin doctors | two princes | 158 ( in 4 weeks - reply in 1944 seconds \*) |
| Mr. Probz | Waves | 825 ( in 4 weeks - reply in 3352 seconds ) |

\* multiple queries running

Averages are drawn in the histograms

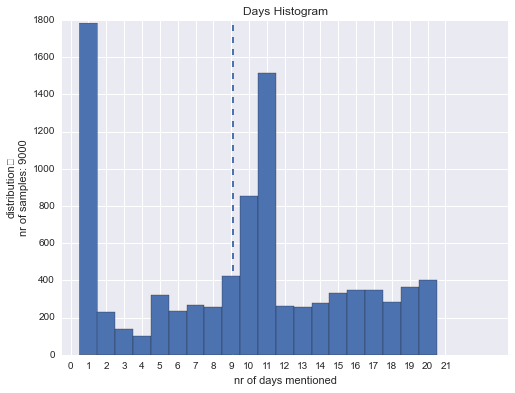
Description Shape of distribution, normality

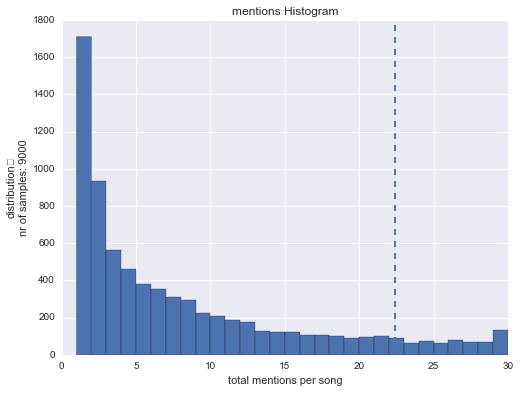
the shape of its distribution, which tells you the frequency of values from different ranges of the variable. More precise information can be obtained by performing one of the tests of normality to determine the probability that the sample came from a normally distributed population of observations (e.g., the so-called Kolmogorov-Smirnov test, or the Shapiro-Wilks' W test. However, none of these tests can entirely substitute for a visual examination of the data using a histogram (i.e., a graph that shows the frequency distribution of a variable).

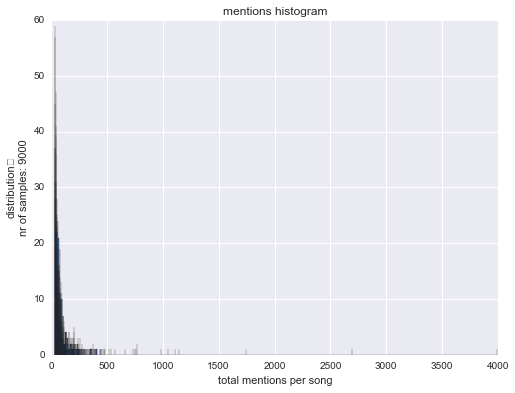
Purpose

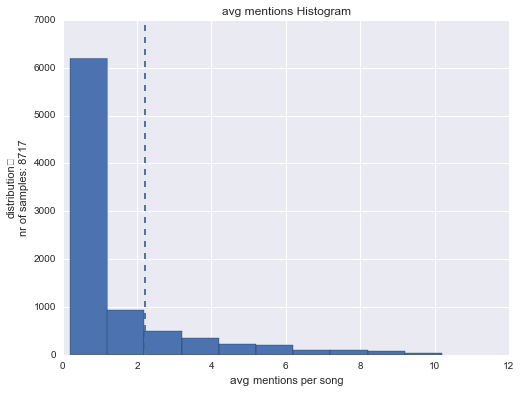
The graph allows you to evaluate the normality of the empirical distribution. It also allows you to examine various aspects of the distribution qualitatively. For example, the distribution could be bimodal (have 2 peaks). This might suggest that the sample is not homogeneous but possibly its elements came from two different populations, each more or less normally distributed. In such cases, in order to understand the nature of the variable in question, you should look for a way to quantitatively identify the two sub-samples.

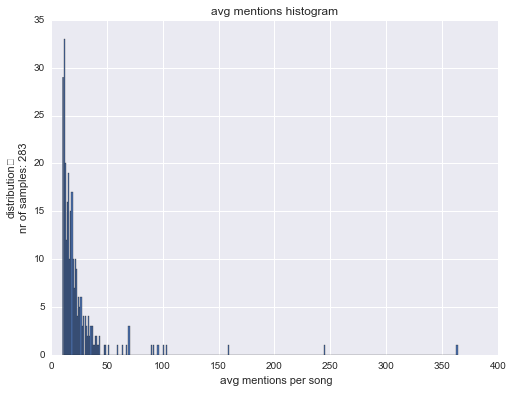
Results











|  |  |  |  |
| --- | --- | --- | --- |
| artist: | song: | Pearsonr: | p-value: |
| ["['James Fortune & FIYA']"] | ['Live Though It'] | -0.21 | 0.6 |
| ["['Wings']"] | ['Aku Pun Tahu'] | -0.19 | 0.62 |
| ["['Hurts']"] | ['Blood, Tears & Gold (Lotus Eaters on My Mind remix by Pantha du Prince)'] | -0.39 | 0.23 |
| ["['Good Charlotte']"] | ['Dance Floor Anthem'] | 0.41 | 0.41 |
| ["['Train']"] | ['Calling All Angels'] | 0.4 | 0.22 |
| ["['S.A.R.S.']"] | ['Neka te kraj mene'] | 0 | 1 |
| ["['Cem Adrian']"] | ['Mutlu Y?llar'] | 0.32 | 0.33 |
| ["['Mumm-Ra']"] | ["She's Got You High"] | 0.35 | 0.36 |
| ["['Erykah Badu']"] | ['Next Lifetime'] | 0.61 | 0.06 |
| ["['Gökhan Türkmen']"] | ['Dön'] | 0.08 | 0.81 |
| ["['Wale']"] | ['The Girls On Drugs'] | 0.87 | 0 |
| ["['Duman']"] | ['Deli'] | -0.08 | 0.83 |
| ["['Haddaway']"] | ['What Is Love'] | 0.79 | 0.01 |
| ["['AC/DC']"] | ['Thunderstruck'] | 0.38 | 0.25 |
| ["['Manika']"] | ['Vegas Party (Jump Smokers Remix)'] | -0.04 | 0.92 |
| ["['Golden Earring']"] | ['Radar Love'] | 0.33 | 0.32 |
| ["['The Whispers']"] | ['Rock Steady'] | 0.23 | 0.5 |
| ["['Myles Kennedy and The Conspirators', 'Slash']"] | ['Anastasia'] | 0.23 | 0.53 |
| ["['Ellie Goulding']"] | ['Love Me Like You Do'] | 0.72 | 0.01 |
| ["['All We Are']"] | ['Keep Me Alive'] | -0.33 | 0.42 |
| ["['Seyyâl Taner']"] | ['Seni Çok Özledim'] | -0.23 | 0.59 |
| ["['Paulina Rubio']"] | ['Dame otro tequila'] | 0 | 1 |
| ["['Wisin & Yandel']"] | ['En la disco bailoteo'] | -0.55 | 0.08 |
| ["['Jagged Edge']"] | ['Love Come Down'] | 0.74 | 0.01 |
| ["['The Glitch Mob']"] | ['Bad Wings'] | 0 | 1 |
| ["['deadmau5']"] | ['Animal Rights'] | -0.07 | 0.89 |
| ["['Madeon']"] | ["You're On (feat. Kyan)"] | 0.48 | 0.13 |
| ["['Cem Adrian']"] | ['Anlad?m'] | 0.28 | 0.49 |
| ["['Rico Love']"] | ['He Got Monry Feat. Future'] | 0.41 | 0.24 |
| ["['Mary Chapin Carpenter']"] | ['Shut Up and Kiss Me'] | -0.17 | 0.69 |
| ["['Müslüm Gürses']"] | ['Affet'] | -0.27 | 0.43 |
| ["['Bullet for My Valentine']"] | ['A Place Where You Belong'] | -0.09 | 0.81 |
| ["['Feeding Like Butterflies']"] | ['Jack'] | 0.36 | 0.31 |

# Correlations

Description Simple Linear Correlation (Pearson r)

Pearson correlation (hereafter called correlation), assumes that the two variables are measured on at least interval scales (see Elementary Concepts), and it determines the extent to which values of the two variables are "proportional" to each other. The value of correlation (i.e., correlation coefficient) does not depend on the specific measurement units used; for example, the correlation between height and weight will be identical regardless of whether inches and pounds, or centimeters and kilograms are used as measurement units. Proportional means linearly related; that is, the correlation is high if it can be "summarized" by a straight line (sloped upwards or downwards). This line is called the regression line or least squares line, because it is determined such that the sum of the squared distances of all the data points from the line is the lowest possible. Note that the concept of

squared distances will have important functional consequences on how the value of the correlation coefficient

reacts to various specific arrangements of data.

Purpose

Correlation is a measure of the relation between two or more variables. The measurement scales used should be at least interval scales, but other correlation coefficients are available to handle other types of data. Correlation coefficients can range from -1.00 to +1.00. The value of -1.00 represents a perfect negative correlation while a value of +1.00 represents a perfect positive correlation. A value of 0.00 represents a lack of correlation.

# t-Test ( dependent and independent )

Description Independent

Purpose, Assumptions. The t-test is the most commonly used method to evaluate the differences in means between two groups. For example, the t-test can be used to test for a difference in test scores between a group of patients who were given a drug and a control group who received a placebo. Theoretically, the t-test can be used even if the sample sizes are very small (e.g., as small as 10; some researchers claim that even smaller n's are possible), as long as the variables are normally distributed within each group and the variation of scores in the two groups is not reliably different (see also Elementary Concepts). As mentioned before, the normality assumption can be evaluated by looking at the distribution of the data (via histograms) or by performing a normality test. The equality of variances assumption can be verified with the F test, or you can use the more robust Levene's test. If these conditions are not met, then you can evaluate the differences in means between two groups using one of the nonparametric alternatives to the t- test (see Nonparametrics and Distribution Fitting).

The p-level reported with a t-test represents the probability of error involved in accepting our research hypothesis about the existence of a difference. Technically speaking, this is the probability of error associated with rejecting the hypothesis of no difference between the two categories of observations (corresponding to the groups) in the population when, in fact, the hypothesis is true. Some researchers suggest that if the difference is in the predicted direction, you can consider only one half (one "tail") of the probability distribution and thus divide the standard p-level reported with a t-test (a "two-tailed" probability) by two. Others, however, suggest that you should always report the standard, two-tailed t-test probability.

Examples

|  |  |
| --- | --- |
| **Question** | **Strategy** |
| Do users located in the US have the same/or similar taste of music, as those located in Europe? | We will first begin with dividing our users in two groups, users located in US and, second group, EU. We will try to find a relation between those groups as we look at the top titles of songs. |
| Do the two groups have the same preference of genre? | Knowing the titles from the above research, hence we will analyze and find a relation of the genre preferences between the two groups. |
| Does the age of the users affect the preferred genre? | At this stage we will split the users in two groups, one from 0-35, second – 36-70. We will then check the most listened genres of those groups. |
| Does the weather/seasons affect the genre? | By splitting the tweets between summer and winter, we will try to find a relation with the genres. We will try define a border between those seasons. |

Description Dependent

Within-group Variation. As explained in Elementary Concepts, the size of a relation between two variables, such as the one measured by a difference in means between two groups, depends to a large extent on the differentiation of values within the group. Depending on how differentiated the values are in each group, a given "raw difference" in group means will indicate either a stronger or weaker relationship between the independent (grouping) and dependent variable. For example, if the mean WCC (White Cell Count) was 102 in males and 104 in females, then this difference of "only" 2 points would be extremely important if all values for males fell within a range of 101 to 103, and all scores for females fell within a range of 103 to 105; for example, we would be able to predict WCC pretty well based on gender. However, if the same difference of 2 was obtained from very differentiated scores (e.g., if their range was 0-200), then we would consider the difference entirely negligible. That is to say, reduction of the within-group variation increases the sensitivity of our test.

Purpose

The t-test for dependent samples helps us to take advantage of one specific type of design in which an important source of within-group variation (or so-called, error) can be easily identified and excluded from the analysis. Specifically, if two groups of observations (that are to be compared) are based on the same sample of subjects who were tested twice (e.g., before and after a treatment), then a considerable part of the within-group variation in both groups of scores can be attributed to the initial individual differences between subjects. Note that, in a sense, this fact is not much different than in cases when the two groups are entirely independent (see t-test for independent samples), where individual differences also contribute to the error variance; but in the case of independent samples, we cannot do anything about it because we cannot identify (or "subtract") the variation due to individual differences in subjects. However, if the same sample was tested twice, then we can easily identify (or "subtract") this variation. Specifically, instead of treating each group separately, and analyzing raw scores, we can look only at the differences between the two measures (e.g., "pre-test" and "post test") in each subject. By subtracting the first score from the second for each subject and then analyzing only those "pure (paired) differences," we will exclude the entire part of the variation in our data set that results from unequal base levels of individual subjects. This is precisely what is being done in the t-test for dependent samples, and, as compared to the t-test for independent samples, it always produces "better" results (i.e., it is always more sensitive).

***Assumptions.*** The theoretical assumptions of the t-test for independent samples also apply to the dependent samples test; that is, the paired differences should be normally distributed. If these assumptions are clearly not met, then one of the nonparametric alternative tests should be used.

Examples

|  |  |
| --- | --- |
| **Question** | **Strategy** |
| Do users located in the US have the same/or similar taste of music, as those located in Europe? | We will first begin with dividing our users in two groups, users located in US and, second group, EU. We will try to find a relation between those groups as we look at the top titles of songs. |
| Do the two groups have the same preference of genre? | Knowing the titles from the above research, hence we will analyze and find a relation of the genre preferences between the two groups. |
| Does the age of the users affect the preferred genre? | At this stage we will split the users in two groups, one from 0-35, second – 36-70. We will then check the most listened genres of those groups. |
| Does the weather/seasons affect the genre? | By splitting the tweets between summer and winter, we will try to find a relation with the genres. We will try define a border between those seasons. |

# Breakdowns

Description

Breakdowns are typically used as an exploratory data analysis technique; the typical question that this technique can help answer is very simple: Are the groups created by the independent variables different regarding the dependent variable? If you are interested in differences concerning the means, then the appropriate test is the breakdowns one-way ANOVA (F test). If you are interested in variation differences, then you should test for homogeneity of variances.

Purpose

The breakdowns analysis calculates descriptive statistics and correlations for dependent variables in each of a number of groups defined by one or more grouping (independent) variables.

# Frequency tables

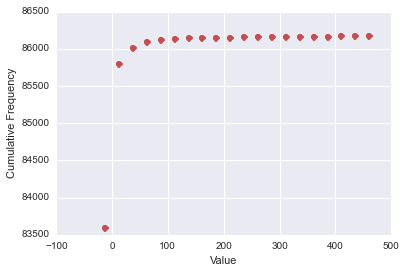
Description

In practically every research project, a first "look" at the data usually includes frequency tables. For example, in survey research, frequency tables can show the number of males and females who participated in the survey, the number of respondents from particular ethnic and racial backgrounds, and so on. Responses on some labeled attitude measurement scales (e.g., interest in watching football) can also be nicely summarized via the frequency table. In medical research, one may tabulate the number of patients displaying specific symptoms; in industrial research one may tabulate the frequency of different causes leading to catastrophic failure of products during stress tests (e.g., which parts are actually responsible for the complete malfunction of television sets under extreme temperatures?). Customarily, if a data set includes any categorical data, then one of the first steps in the data analysis is to compute a frequency table for those categorical variables.

Purpose of the frequencies

Frequency or one-way tables represent the simplest method for analyzing categorical (nominal) data (refer to Elementary Concepts). They are often used as one of the exploratory procedures to review how different categories of values are distributed in the sample. For example, in a survey of spectator interest in different sports, we could summarize the respondents' interest in watching football in a frequency table as follows:

The table above shows the number, proportion, and cumulative proportion of respondents who characterized their interest in watching football as either (1) Always interested, (2) Usually interested, (3) Sometimes interested, or (4) Never interested.



# Appendix A - Data structure

Documents are not required to have the same schema. Number of documents are ever increasing, just there to give an idea of the content

|  |  |
| --- | --- |
| npCountry | 613972 |
| \_id |  |
| value |  |
| npTimeseriesDaily | 460053 |
| \_id |  |
| value |  |
| npTimeseriesYearly | 471628 |
| \_id |  |
| value |  |
| npTotals | 471628 |
| \_id |  |
| value |  |

|  |  |
| --- | --- |
| songs.daily | 1318650 documents |
| \_id |  |
| value |  |
| songs.hourly | 2107493 documents |
| \_id |  |
| value |  |
| songs.weekly | 778540 documents |
| \_id |  |
| value |  |

|  |  |
| --- | --- |
| panelData | 3294 documents |
| \_id |  |
| week |  |
| questions |  |
| sessionID |  |
| year |  |
| identifier |  |
| jobId |  |
| youtube | 0 documents |
| empty | |
| twitter-np | 93230675 documents |
| \_id |  |
| song |  |
| tweet |  |
| users | 1 document |
| \_id |  |
| services |  |
| emails |  |
| username |  |
| createdAt |  |