MOUNT ROYAL
UNIVERSITY
1910

Mount Royal University

Department of
Mathematics and Computing

COMP 4522

Database II: Advanced Databases

Contents

EXERCISE

NINE

DESCRIPTIVE STATISTICS

Means, medians, modes
Distributions

AN IMPORTANT PART OF DATA MINING is developing a thorough understanding of the data at hand. To do this, one must understand the attribute domains, and must understand how values are distributed within the domain.

An important tool in your "Data Miners Tool Box" is descriptive statistics. The goal of descriptive statistics is to provide values that describe a data set without someone having to examine all of the data. For example it is simpler to say that the class average on an exam was 78% than to list all of the individual scores. Some useful descriptive statistics follow. Functions to calculate all of these statistics are available in Oracle, MySQL, Excel and Python. If your dataset is not too large, Excel can be a valuable tool for calculating and visualizing your descriptive statistics.

Numeric Data

For an attribute with a numeric domain:

Count How many values are there? This is not the same as how many rows are in the table as some values may be NULL.

The number of values, n, present is important for how much confidence you have in the statistic. If your n is small you may not put as much faith into the other statistics that you capture.

Range What is the range of the value? This can be found using the $MAX()$ and $MIN()$ functions.

Mean The arithmetic mean of the attribute values: $\frac{\sum_{i=1}^{n} x_i}{n}$. The mean is a useful statistic, but it can sometimes be misleading if there are any outliers: values that are much bigger or smaller than most. For example if I teach two sections of *COMP 2521 Database I* and both have an average grade of 70% did the students in both sections do equally well? Not necessarily. In one class all of the students could have got 70% while in the other 15 students could get 100% and 15 could get 40%. Both classes have the same average.

A common operation is to "centre" values around a mean. That is to take the each value and subtract the mean from it This will make values that are less than the mean negative and values that are greater than the mean positive.

Median The median is that value that splits the data in half. Half of the values are larger and half are smaller. If the median is lower than the mean then there are more low values and if the median is higher than the mean then there are more high values.

Mode The mode is the most frequently occurring value. With continuous values it may be necessary to round or otherwise "bin" the values before calculating the mode.

For an example of binning, think of temperature data. In Calgary we have temperatures that range from about -35.0 degrees to +35.0 degrees. For some applications we could create bins that span 5 degrees. We could to the nearest 5 and end up with -35, -30, -25, -20, -15, -10, -5, 0, 5, 10, 15, 20, 25, 30, 35.

Standard Deviation The standard deviation measures the variance in the data set. For normally distributed data 2/3 of the values are within +/- one standard deviation of the mean. A small standard deviation means values tend to be close to the mean.

If you find that the mode, median and mean are all the same (or very close) you can assume that the data is *normally distributed*. You can see what that implies herehttps://www.mathsisfun.com/data/standard-normal-distribution.html.

Non-Numeric Data

Obviously, most of the statistics above will not work for non-numeric data, but it is possible to look at the range, the mode and at a frequency distribution for non-numeric values. For example the distribution of letter grades is something I often calculate.

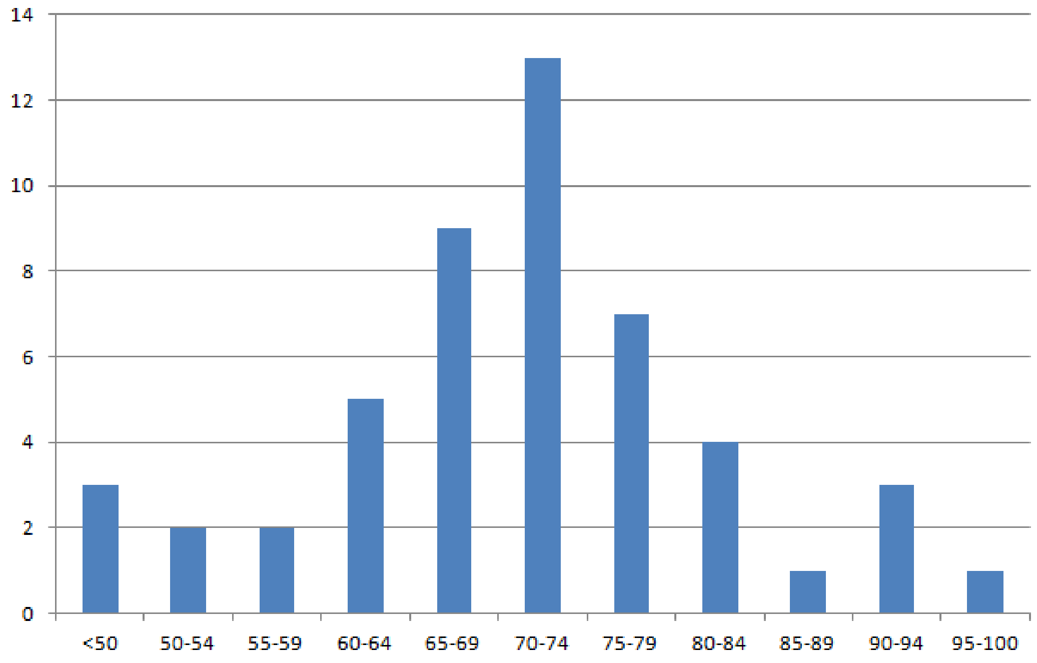
Frequency Distributions

A very useful way to visualize the data is with a frequency distribution histogram. This works for continuous, discrete and even non numeric data.

If there are a large number of possibilities, the idea here is to place the data into *bins*. For example if looking at integer class grades one could create bins of $<$ 50, 50-54, 55-59, 60-64, ... 95-100. Then count how many grades fall into each bin.

If you have data with a smaller number of possibilities (a more restricted domain), like letter grades, you can just use the data as is.

A graph can be made:



Distributions are one of the most useful and interesting things you can do when exploring data. Is it a normal distribution? Is is skewed? Are there outliers?

Relationships Between Attributes

So far we have been looking at only one attribute at a time. If you suspect that two attributes (A and B) may have some kind of relationship you can calculate the correlation coefficient between two data sets.

The correlation coefficient is a number between -1 and 1. If the value is 0, then there is no relationship between the two values.

If the value is 1, then there is a strong positive correlation, that is when the value of A increases, the values of B also increases.

If the value is -1, then there is a strong negative correlation, that is when the value of A increases, the values of B decreases.

When the value is between 0 and 1 it is a judgement call about how strong a correlation exists.

Remember that just because a two variables have a high correlation coefficient it does not necessarily mean there is any relationship between the two, it may simply be coincidence. *The correlation value does not say* *anything about cause*. If A and B have a strong correlation (0.9), that may be interesting, but it does not tell you if A is causing B or if B is causing A.

When you discover a correlation try to find an explanation for the correlation.

Finding a strong correlations is interesting but you always need to be cautious because there are far more spurious correlations than there are valid, meaningful ones. Explore this website https://www.tylervigen.com/discover to see some examples.

In addition to examples where there is a strong correlation but no causal link between the attributes there are also correlations that are valid, but not helpful.

A good example that I have often used is cycling data. I have a database that records bike rides and includes the duration and distance for each trip. There is a very strong correlation, nearly 1, between the trip duration and the distance traveled. While it is a valid correlation and there is an easily explained causal relationship, it is not very interesting.

INFERENTIAL STATISTICS

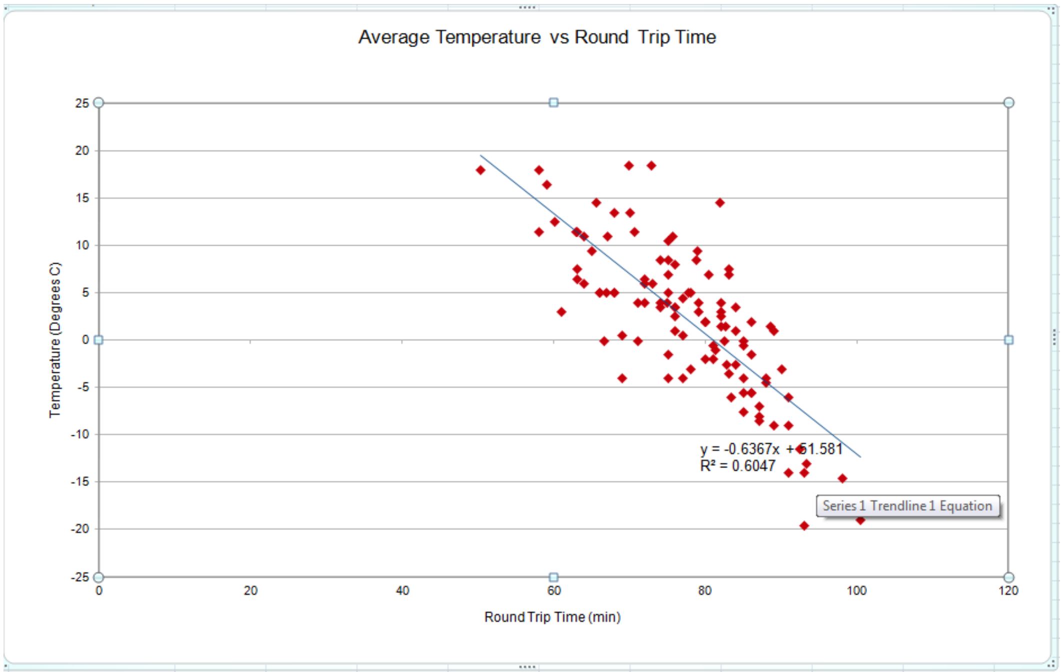
While descriptive statistics describe the data, as it is, inferential statistics are used to make predictions. You use the data that you have to create a formula that can be used to predict values outside of the data you have already collected.

Regression and Making Predictions

Inferential statistics is a huge area, far beyond the scope of this course, but to give a taste we can look at linear regression, the most common type of inferential statistics.

In this case you attempt to find a linear equation, $y = mx + b$ that describes how two attributes are related. It is possible to try to find polynomial, linear equations with many variables, exponentials etc. but these are beyond the scope of this course.

Usually a scatter plot is created graphing one of the attributes on the x-axis and one on the y-axis. Then a line of best fit and be calculated. The Seaborn package in Python and Excel, and Oracle all have built in functions to do these.



To show how this can be done directly in the database, Oracle includes a set of functions that can be used to calculate regressions.

```
/* Using Oracle to perform linear regression */

/* Create a data table and populate it with data. */
DROP TABLE foo;
CREATE TABLE foo ( x INTEGER, y INTEGER);

INSERT INTO foo VALUES (1,3);
INSERT INTO foo VALUES (2,5);
INSERT INTO foo VALUES (3,7);
INSERT INTO foo VALUES (5,10);
INSERT INTO foo VALUES (10,22);
INSERT INTO foo VALUES (11,23);
INSERT INTO foo VALUES (12,25);

/* For a linear equation we need to find the slope, m, and the y-interc

   We also want to estimate how well our line fits the data. The
   R2 coefficient of determination R^2 gives us a number betweeen
   0 and 1. 1 is a perfect fit.

SELECT
   REGR_SLOPE(y, x) SLOPE,
   REGR_INTERCEPT(y, x) INTERCEPT,
   REGR_COUNT(y, x) COUNT,
   REGR_R2(y, x) R2
FROM foo;
```

You then have a linear equation that you can plot or use to make predictions.