Statistics

Descriptive statistics:

Statistics involving describing data. Involves summarizing and organizing data so they can be easily understood.

Inference statistics:

Complex set of procedures to draw conclusions over large populations with sample data.

Data

Numeric: wind speed, time duration, discrete etc.

Categorical: Car types, Binary, ordinal (ordered).

*------------------- 5/7/2024 Up to Page 5*

Exploratory Data Analysis

Data Structure:

* Rectangular (rows are records & columns are variable or features) and
* Non-Rectangular (spatial or graph)

In statistics we use mostly rectangular data:

**Estimates of Location**

An estimate of where most of the data is located (i.e., its central tendency)

|  |  |  |  |
| --- | --- | --- | --- |
| Key Term | Definition | Formula | Usage |
| Mean | Sum of all values/ number of values |  | average |
| Weighted Mean | Sum of all values times a weight / sum of weights |  | Some variables are intrinsically more variable than other and high variable observations are given lower weight. Ex: when taking average from multiple sensors giving lower weight for sensors that giving less accurate readings. |
| Median | The value such that one-half of data lies above and below |  | While calculating average household income in a city where bill gate lives the mean gives diff value where median gives right value no matter who is rich or not. |
| Percentile | The value such that P percentage of data lies below |  |  |
| Weighted Median | The value such that one half of the weighted sum lies above and below the sorted data. |  |  |
| Trimmed Mean | The average of all values after removing fixed number of extreme values | P smallest and largest values omitted | A trimmed means eliminate the influence of extreme values, EX: International diving the top score & bottom score from five judges are dropped and the final score is the average of scores from 3 remaining judges. This makes it difficult for a single judge to manipulate the scores. |

Note: Trimmed mean, Median and weighted median are robust to outliers.

Outlier: is a any value which is very distant from other values in data set and cause skewness.

*------------------- 5/8/2024 Up to Page 11*

[Refer Estimates of Location in Python Notebooks]

*------------------- 5/9/2024 Up to Page 12*

**Estimates of Variability**

Measures whether the data values are tightly clustered or spread out.

At the heart of the statistics lies variability:

* Measuring it
* Reducing it
* Distinguishing random from real variability
* Identifying the various sources of real variability
* Make decision out of it in presence.

|  |  |  |  |
| --- | --- | --- | --- |
| Key Terms | Why | Formula | Formula Explanation |
| Deviation | To calculate how individual item deviate from mean, useful to calculate Standard deviation | Mean – Individual Value | Simple distance calculation formula |
| Variance | How each data is varied from the mean |  | Its not express in square root as it already quantifies the spread of data in squared units. |
| Standard Deviation  (Population) means whole dataset | Square root of variance |  | Just calculating the mean of deviations (distance form mean), squaring to avoid negatives to cancel with positive. Square root to get back squares to their original unit. |
| Standard Deviation (sample) | Same as population but n-1  Dividing by 𝑛−1 instead of n results in a slightly larger value for the standard deviation, which better reflects the variability in the population from which the sample was drawn |  | This adjustment is known as Bessel's correction. The rationale behind this correction is to provide an unbiased estimate of the population standard deviation. |
| Mean Absolute Deviation (Manhattan Norm, l1 – norm) | Mean of absolute values of the deviation from the mean |  | SD emphasize the large deviation, but MAD won’t because its taking absolute values |
| Median Absolute Deviation from Median | Robustness to outliers |  | Median of absolute values of the deviation from the median |
| Percentile | A percentile tells you where a certain value falls within a dataset when arranged in ascending order. |  | Simple percentage calculation  For instance, if your score is at the 70th percentile, it means you've scored as well as or better than 70% of the other people in the dataset. |
| Interquartile range | The difference between 75th percentile and 25th percentile |  | This range represents where the bulk of the data lies, excluding the extremes. |

*------------------- 5/10/2024 Up to Page 14 (Additional notes pending)*

*------------------- 5/11/2024 Up to Page 14 (Additional notes updated with why and formula explanation)*

**Degrees of freedom and n or n – 1?**

whether you divide by n or n – 1. It is based on the premise that you want to make estimates about a population, based on a sample. If you use the intuitive denominator of n in the variance formula, you will underestimate the true value of the variance and the standard deviation in the population. This is referred to as a biased estimate. However, if you divide by n – 1 instead of n, the variance becomes an unbiased estimate.

To fully explain why using n leads to a biased estimate involves the notion of degrees of freedom, which considers the number of constraints in computing an estimate. In this case, there are n – 1 degrees of freedom since there is one constraint: the standard deviation depends on calculating the sample mean. For most problems, data scientists do not need to worry about degrees of freedom.

Notes:

* Even for normal distribution the calculation of SD, MAD, Median AD are different.
* The percentile is essentially the same as a quantile, with quantiles indexed by fractions (so the .8 quantile is the same as the 80th percentile).

*------------------- 5/12/2024 Up to Page 17*

[Refer Estimates of Variability in Python Notebooks]

*------------------- 5/13/2024 Up to Page 18*

**Explore Data Distribution**

|  |  |  |
| --- | --- | --- |
| Key Term | Usage | visual |
| Box Plot | * Helps to identify the quartiles, Interquartile, median, find outliers | Box Plot - Simply explained - DATAtab |
| Frequency Table | * Able to find how frequent or the count of category or intervals in a dataset. * Helps to summarize data. * Further helps to plot Bar and pie charts. | Frequency Table: How to Make & Examples - Statistics By Jim |
| Histogram | Visual representation of data distribution, the bar height says how frequence of the data occur.  Uses bins to represent frequency of observations. | Histograms Unveiled: Analyzing Numeric Distributions |
| Density Plot | A smoothed version of the histogram, often based on a kernel density estimate.  estimate the probability density function (PDF) of the underlying distribution, providing a smoothed representation of the data distribution.  When dataset is large and continuous data. | Density – from Data to Viz |

*------------------- 5/14/2024 Up to Page 19*

[Refer Estimates of Distribution in Python Notebooks]

*------------------- 5/15/2024 Up to Page 26*

**Exploring Binary and Categorical data**

|  |  |  |
| --- | --- | --- |
| Key Term | Definition | Example |
| Mode | The most commonly occurring category or value in a data set. | In most part of US, the mode of religious preference would be Christian |
| Expected value | When categories can be associated with numerical values, this gives an average value based on category’s probability of occurrence. |  |
| Bar chart | The proportion of each category plotted as bars |  |
| Pie chart | The proportion of each category plotted as wedges in a pie. |  |

*------------------- 5/16/2024 Up to Page 27*

[Refer Exploration of Binary and Categorical Data in Python Notebooks]

*------------------- 5/17/2024 Up to Page 28*

Expected Value

A marketer for a new cloud technology, for example, offers two levels of service, one priced at $300/month and another at $50/month. The marketer offers free webinars to generate leads, and the firm figures that 5% of the attendees will sign up for the $300 service, 15% will sign up for the $50 service, and 80% will not sign up for anything. This data can be summed up, for financial purposes, in a single “expected value,”

The expected value is calculated as follows:

1. Multiply each outcome by its probability of occurrence.

2. Sum these values

In the cloud service example, the expected value of a webinar attendee is thus $22.50 per month, calculated as follows:

EV = (0 .05) (300) + (0 .15) (50) + (0 .80) (0) = 22.5

*------------------- 5/18/2024 Up to Page 29*

**Probability**

For our purposes the probability of an event will happen is the proportion of the event will occur if the situation could be repeated over and over again, infinitely.

**Correlation**

In Exploratory Data Analysis the correlation will be performed among features or between feature and target variables.

If the highest value of X goes with highest value of Y then correlation of X and Y is positively correlated and also for vice versa.

If the highest value of X goes with lowest value of X then correlation of X and Y is negatively correlated and also for vice versa

*------------------- 5/19/2024 Up to Page 30*

**Correlation key terms**

*Correlation coefficient:* a metric to which extend the numerical variables are associated ranges from -1 to +1.

*Correlation matrix*: a matrix in which variables are both in rows and columns and correlation between them as values.

*Scatter plot:* a plot which x-axis is one plot and y axis is one plot.

**Vector Sum of Products**

First, let's clarify what the "vector sum of products" means:

Given two vectors:

Vector 𝐴

A = [1, 2, 3]

Vector 𝐵

B = [4, 5, 6]

The sum of products (or dot product) of these vectors is calculated as:

1⋅4 + 2⋅5 + 3⋅6

Breaking it down:

1⋅4=4

2⋅5=10=10

3⋅6=18

Adding these up:

4+10+18=32

Shuffling and Recalculation

The next part talks about shuffling one of the vectors and recalculating the sum of products. Let's shuffle vector 𝐵

Assume we shuffle vector 𝐵

B to [6, 4, 5]. Now, we recalculate the sum of products:

1⋅6+2⋅4+3⋅5

Breaking it down:

1⋅6=6

2⋅4=8

3⋅5=15

Adding these up:

6+8+15=29

In this case, the new sum of products is 29, which is less than 32.

Why the Original Sum is the Highest

The text states that "the vector sum of products will never be higher than 32" when you shuffle the elements. This is because the original ordering of the vectors (1, 2, 3 with 4, 5, 6) maximizes the sum of products due to the way the values are paired. This property is a result of the vectors being sorted in the same order. When both vectors are sorted in increasing order, the sum of their products is maximized.

Using the Sum as a Metric

The sum of products (32 in the original case) can be used as a metric to compare against random shufflings. By shuffling one of the vectors multiple times and calculating the sum of products each time, you can generate a distribution of sums. This relates to a resampling-based estimate.

Resampling-Based Estimate

In statistics, resampling methods involve repeatedly drawing samples from a data set and calculating a statistic (in this case, the sum of products) for each sample. By comparing the observed statistic (32) to the distribution of statistics from the resampled data, you can determine how likely or unusual the observed value is.

In summary, the vector sum of products is maximized when both vectors are in the same order. Shuffling and recalculating this sum multiple times allows for a comparison of the observed value against a distribution of potential values, which can be used in various statistical analyses, such as hypothesis testing or confidence interval estimation.

To compute **Pearson’s correlation coefficient**, we multiply deviations from the mean for variable 1 times those for variable 2, and divide by the product of the standard deviations



*------------------- 5/20/2024 Up to Page 31*

Variables can have an association that is not linear, in which case the correlation coefficient may not be a useful metric. The relationship between tax rates and revenue raised is an example: as tax rates increase from zero, the revenue raised also increases. However, once tax rates reach a high level and approach 100%, tax avoidance increa‐ ses and tax revenue actually declines.



*------------------- 5/21/2024 Up to Page 32*

NumPy

To work with N Dimensional array

Create arrays -> np.array([List]) | np.zeros() |np.linspace(), np.logspace()

Examining -> .dtype, .ndim, .shape, .size

*------------------- 5/22/2024 [External]*

[Refer Correlation in Python Notebooks]

*------------------- 5/23/2024 Up to Page 32*

[Refer Correlation in Python Notebooks]

*------------------- 5/24/2024 Up to Page 33*

[Refer Scatter plot in Python Notebooks]

*------------------- 5/25/2024 Up to Page 34*

Scatterplot

The standard way to visualize the relationship between two measured data variables is with a scatterplot.

The correlation coefficient measures the extent to which two paired variables (e.g., height and weight for individuals) are associated with one another.

When high values of v1 go with high values of v2, v1 and v2 are positively associated.

• When high values of v1 go with low values of v2, v1 and v2 are negatively associated.

• The correlation coefficient is a standardized metric, so that it always ranges from –1 (perfect negative correlation) to +1 (perfect positive correlation).

• A correlation coefficient of zero indicates no correlation, but be aware that ran‐ dom arrangements of data will produce both positive and negative values for the correlation coefficient just by chance

*------------------- 5/26/2024 Up to Page 35*

**Explore two or more variables.**

**Contingency table**

A tally of counts between two or more categorical variables.

**Hexagonal binning**

A plot of two numeric variables with the records binned into hexagons.

**Contour plot.**

A plot showing the density of two numeric variables like a topographical map.

**Violin plot**

Like a boxplot but showing the density estimate.

*------------------- 5/27/2024 Up to Page 36*

**Hexagonal Binning and Contours (Plotting Numeric Versus Numeric Data)**

Scatter Plot is good for less amount of dataset. But what if the dataset size is 100s of 1000s of Millions?  
Plotting scatter plot will be too dense.

The records will be plotted by grouped bins with two variables and the color coded will be gradient color indicating the number of records in that bin.

*------------------- 5/28/2024 Up to Page 37*

In Contours chart

It is contours overlaid on scatterplot to show relationships between two variables. It is a topographical map to two variables. Each contour band represents a density of points.

Heat map, hexagonal binning and contours chart are similar story telling plots for two-dimensional density.

*------------------- 5/29/2024 Up to Page 38*

**Two Categorical Variables**

A useful way to summarize two categorical variables is a contingency table or Pivot table.

Contingency tables can look only at counts, or they can also include column and total percentages. Pivot tables in Excel are perhaps the most common tool used to create contingency tables

*------------------- 5/30/2024 Up to Page 40*

**Categorical and Numeric Data**

Boxplots are a simple way to visually compare the distributions of a numeric variable grouped according to a categorical variable.

*------------------- 5/31/2024 Up to Page 41*

Violin Plot

Violin plot is an enhancement of box plot, where its show the density and y – axis. The density is mirrored and flipped over and resulting shape is filled in creating an image resembling a VIOLIN.

It shows the nuances in the distribution that aren’t perceptible in box plot. On the other hand the boxplot clearly shows the outliers.

*------------------- 6/1/2024 Up to Page 42*

Visualizing Multiple Variables

The type of chart used to compare two variables can readily extended to more variables through the notion of conditioning.

**EDA Summary:**

Exploratory data analysis (EDA), pioneered by John Tukey

The key idea of EDA is that the first and most important step in any project based on data is to look at the data. By summarizing and visualizing the data, you can gain valuable intuition and understanding of the project.

*------------------- 6/3/2024 Up to Page 46*

Data and Sampling Distributions

Sampling is necessary even though we have big data!

In traditional statistics, there assumptions made on population but in current statistics there is no need and sample is used to train models and use it against the big data.



*------------------- 6/4/2024 Up to Page 47*

**Random Sampling and Sample Bias**

Random sampling is a process in which each available member of the population being sampled has an equal chance of being chosen for the sample at each draw. The sample that results is called a *simple random sample*. Sampling can be done with replacement, in which observations are put back in the population after each draw for possible future reselection. Or it can be done without replacement, in which case observations, once selected, are unavailable for future draws.

Data quality often matters more than data quantity when making an estimate or a model based on a sample. Data quality in data science involves completeness, consistency of format, cleanliness, and accuracy of individual data points. Statistics adds the notion of representativeness.

*------------------- 6/5/2024 Up to Page 48*

**Key Terms for Random Sampling**

*Sample:*

A subset from a larger dataset.

*Population:*

The larger data set or idea of a data set.

*N (n):*

The size of the population (sample).

*Random sampling:*

Drawing elements into a sample at random.

*Stratified sampling:*

Dividing the population into strata and randomly sampling from each strata.

*Stratum (pl., strata):*

A homogeneous subgroup of a population with common characteristics.

*Simple random sample:*

The sample that results from random sampling without stratifying the population.

*Bias:*

Systematic error.

*Sample Bias:*

A sample that misrepresents the population.

*------------------- 6/6/2024 Up to Page 49*

Self-Selection Sampling Bias

The reviews of restaurants, hotels, cafés, and so on that you read on social media sites like Yelp are prone to bias because the people submitting them are not randomly selected. rather, they themselves have taken the initiative to write. This leads to self-selection bias— the people motivated to write reviews may have had poor experiences, may have an association with the establishment, or may simply be a different type of person from those who do not write reviews. Note that while self-selection samples can be unreliable indicators of the true state of affairs, they may be more reliable in simply com‐ paring one establishment to a similar one; the same self-selection bias might apply to each.

**Bias**

Statistical bias refers to measurement or sampling errors that are systematic and pro‐ duced by the measurement or sampling process.

An important distinction should be made between errors due to random chance and errors due to bias. Consider the physical process of a gun shooting at a target. It will not hit the absolute center of the target every time, or even much at all. An unbiased process will produce error, but it is random and does not tend strongly in any direction (see Figure 2-2). The results shown in Figure 2-3 show a biased process—there is still random error in both the x and y direction, but there is also a bias. Shots tend to fall in the upper-right quadrant.



A diagram of an object with blue dots

Description automatically generated

Bias comes in different forms, and may be observable or invisible. When a result does suggest bias (e.g., by reference to a benchmark or actual values), it is often an indica‐ tor that a statistical or machine learning model has been mis specified, or an important variable left out.

*------------------- 6/7/2024 Up to Page 50*

**Random Selection**

To avoid the problem of sample bias that led the Literary Digest to predict Landon over Roosevelt, George Gallup (shown in Figure 2-4) opted for more scientifically chosen methods to achieve a sample that was representative of the US voting elector‐ ate. There are now a variety of methods to achieve representativeness, but at the heart of all of them lies random sampling.

*------------------- 6/8/2024 Up to Page 51*

**How to do Random sampling using strata**

First, we need to define who a customer is. We might select all customer records where purchase amount > 0. Do we include all past customers? Do we include refunds? Internal test purchases? Resellers? Both billing agent and customer? Next, we need to specify a sampling procedure. It might be “select 100 customers at random.” Where a sampling from a flow is involved (e.g., real-time customer transactions or web visitors), timing considerations may be important (e.g., a web visitor at 10 a.m. on a weekday may be different from a web visitor at 10 p.m. on a weekend). In stratified sampling, the population is divided up into strata, and random samples are taken from each stratum. Political pollsters might seek to learn the electoral preferences of whites, blacks, and Hispanics. A simple random sample taken from the population would yield too few blacks and Hispanics, so those strata could be over‐ weighted in stratified sampling to yield equivalent sample sizes.

*------------------- 6/9/2024 Up to Page 52*

**Size Versus Quality: When Does Size Matter?**

In the era of big data, it is sometimes surprising that smaller is better. Time and effort spent on random sampling not only reduces bias but also allows greater attention to data exploration and data quality. For example, missing data and outliers may contain useful information. It might be prohibitively expensive to track down missing values or evaluate outliers in millions of records, but doing so in a sample of several thou‐ sand records may be feasible. Data plotting and manual inspection bog down if there is too much data.

*------------------- 6/10/2024 Up to Page 52*

**So when are massive amounts of data needed?**

Consider the search queries received by Google, where columns are terms, rows are individual search queries, and cell values are either 0 or 1, depending on whether a query contains a term. The goal is to determine the best predicted search destination for a given query. There are over 150,000 words in the English language, and Google processes over one trillion queries per year. This yields a huge matrix, the vast majority of whose entries are “0.” This is a true big data problem—only when such enormous quantities of data are accumulated can effective search results be returned for most queries. And the more data accumulates, the better the results.

Consider the search phrase “Ricky Ricardo and Little Red Riding Hood.” In the early days of the internet, this query would probably have returned results on the band‐ leader Ricky Ricardo, the television show I Love Lucy in which that character appeared, and the children’s story Little Red Riding Hood. Both of those individual items would have had many searches to refer to, but the combination would have had very few. Later, now that trillions of search queries have been accumulated, this search query returns the exact I Love Lucy episode in which Ricky narrates, in dramatic fashion, the Little Red Riding Hood story to his infant son in a comic mix of English and Spanish.

Keep in mind that the number of actual pertinent records—ones in which this exact search query, or something very similar, appears (together with information on what link people ultimately clicked on)—might need only be in the thousands to be effective. However, many trillions of data points are needed to obtain these pertinent records (and random sampling, of course, will not help)

*------------------- 6/11/2024 Up to Page 53*

Sample Mean Versus Population Mean

The symbol x (pronounced “x-bar”) is used to represent the mean of a sample from a population, whereas μ is used to represent the mean of a population. Why make the distinction?

*------------------- 6/12/2024 Up to Page 53*

Information about samples is observed, and information about large populations is often inferred from smaller samples. Statisticians like to keep the two things separate in the symbology

**Key Ideas**

• Even in the era of big data, random sampling remains an important arrow in the data scientist’s quiver.

• Bias occurs when measurements or observations are systematically in error because they are not representative of the full population.

• Data quality is often more important than data quantity, and random sampling can reduce bias and facilitate quality improvement that would otherwise be prohibitively expensive.

*------------------- 6/13/2024 Up to Page 53*

**Selection Bias**

Selection bias refers to the practice of selectively choosing data—consciously or unconsciously—in a way that leads to a conclusion that is misleading or ephemeral.

Data snooping:

Extensive hunting through data in search of something interesting.

Vast search effect:

Bias or non-reproducibility resulting from repeated data modeling, or modeling data with large numbers of predictor variables

The difference between a phenomenon that you verify when you test a hypothesis using an experiment and a phenomenon that you discover by perusing available data can be illuminated with the following thought experiment

Since repeated review of large data sets is a key value proposition in data science, selection bias is something to worry about. A form of selection bias of particular con‐ cern to data scientists is what John Elder (founder of Elder Research, a respected data mining consultancy) calls the vast search effect. If you repeatedly run different models and ask different questions with a large data set, you are bound to find something interesting. But is the result you found truly something interesting, or is it the chance outlier? We can guard against this by using a holdout set, and sometimes more than one hold‐ out set, against which to validate performance. Elder also advocates the use of what he calls target shuffling (a permutation test, in essence) to test the validity of predic‐ tive associations that a data mining model suggests

*------------------- 6/14/2024 Up to Page 54*

Regression to the Mean

Regression to the mean refers to a phenomenon involving successive measurements on a given variable.

Sports fans are familiar with the “rookie of the year, sophomore slump” phenomenon. Among the athletes who begin their career in a given season (the rookie class), there is always one who performs better than all the rest. Generally, this “rookie of the year” does not do as well in his second year. Why not? In nearly all major sports, at least those played with a ball or puck, there are two elements that play a role in overall performance:

• Skill

• Luck

Regression to the mean is a consequence of a particular form of selection bias. When we select the rookie with the best performance, skill and good luck are probably con‐ tributing. In his next season, the skill will still be there, but very often the luck will not be, so his performance will decline—it will regress.

*------------------- 6/15/2024 Up to Page 55*

**Galton’s study that identified the phenomenon of regression to the mean**

******

*------------------- 6/16/2024 Up to Page 56*