**Exploratory Data Analysis**

*Exploratory data analysis*, or *EDA*, is a comparatively new area of

statistics. Classical statistics focused almost exclusively on *inference*, a

sometimes complex set of procedures for drawing conclusions about large populations based on small samples.

**Elements of Structured Data**

Data comes from many sources: sensor measurements, events, text, images, and videos. The *Internet of Things* (IoT) is spewing out streams of information. Much of this data is unstructured: images are a collection of pixels with each pixel containing RGB (red, green, blue) color information. Texts are sequences of words and nonword characters, often organized by sections, subsections, and so on. Clickstreams are sequences of actions by a user interacting with an app or web page. In fact, a major challenge of data science is to harness this torrent of raw data into actionable information. To apply the statistical concepts unstructured raw data must be processed and manipulated into a structured form — as it might emerge from a relational database — or be collected for a study.

**KEY TERMS FOR EDA**

***Mean***

The sum of all values divided by the number of values.

*Synonyms*

Average

***Weighted mean***

The sum of all values times a weight divided by the sum of the weights.

*Synonyms*

weighted average

***Median***

The value such that one-half of the data lies above and below.

*Synonyms*

50th percentile

***Weighted median***

The value such that one-half of the sum of the weights lies above and below the sorted data.

***Trimmed mean***

The average of all values after dropping a fixed number of extreme values.

(A truncated mean or trimmed mean is a statistical measure of central tendency, much like the mean and median. It involves the calculation of the mean after discarding given parts of a probability distribution or sample at the high and low end, and typically discarding an equal amount of both)

*Synonyms*

truncated mean

***Robust***

Not sensitive to extreme values.

*Synonyms*

resistant

***Outlier***

A data value that is very different from most of the data.

*Synonyms*

extreme value

Compute the mean, trimmed mean, and median for the population using R:

mean(state[["Population"]])

mean(state[["Population"]], trim=0.1)

median(state[["Population"]])

Compute the weighted median (R doesn’t have a function for

weighted median, we need to install a package such as matrixStats)

weighted.mean(state[["Murder.Rate"]], w=state[["Population"]])

library("matrixStats")

weightedMedian(state[["Murder.Rate"]], w=state[["Population"]])

**Estimates of Variability**

Location is just one dimension in summarizing a feature. A second dimension, *variability*, also referred to as *dispersion*, measures whether the data values are tightly clustered or spread out. At the heart of statistics lies variability: measuring it, reducing it, distinguishing random from real variability, identifying the various sources of real variability, and making decisions in the presence of it.

**KEY TERMS FOR VARIABILITY M ETRICS**

***Deviations***

The difference between the observed values and the estimate of location.

*Synonyms*

errors, residuals

***Variance***

The sum of squared deviations from the mean divided by *n* – 1 where *n* is the number of data values.

*Synonyms*

mean-squared-error

***Standard deviation***

The square root of the variance.

*Synonyms*

l2-norm, Euclidean norm

***Mean absolute deviation***

The mean of the absolute value of the deviations from the mean.

*Synonyms*

l1-norm, Manhattan norm

***Median absolute deviation from the median***

The median of the absolute value of the deviations from the median.

***Range***

The difference between the largest and the smallest value in a data set.

***Order statistics***

Metrics based on the data values sorted from smallest to biggest.

*Synonyms*

*ranks*

***Percentile***

The value such that *P* percent of the values take on this value or less and (100–P) percent take on

this value or more.

*Synonyms*

quantile

***Interquartile range***

The difference between the 75th percentile and the 25th percentile.

*Synonyms*

IQR

Compute the standard deviation, interquartile range (IQR), and the median absolution deviation

sd(state[["Population"]])

IQR(state[["Population"]])

mad(state[["Population"]])

**Exploring the Data Distribution**

**KEY TERMS FOR EXPLORING THE DISTRIBUTION**

***Boxplot***

A plot introduced by Tukey as a quick way to visualize the distribution of data.

*Synonyms*

Box and whiskers plot

***Frequency table***

A tally of the count of numeric data values that fall into a set of intervals (bins).

***Histogram***

A plot of the frequency table with the bins on the x-axis and the count (or proportion) on the yaxis.

***Density plot***

A smoothed version of the histogram, often based on a *kernal density estimate*.

boxplot(state[["Population"]]/1000000, ylab="Population (millions)")

The top and bottom of the box are the 75th and 25th percentiles, respectively. The median is shown by the horizontal line in the box. The dashed lines, referred to as *whiskers*, extend from the top and bottom to indicate the range for the bulk of the data.

**Frequency Table and Histograms**

A frequency table of a variable divides up the variable range into equally spaced segments, and tells us how many values fall in each segment.

breaks <- seq(from=min(state[["Population"]]),

to=max(state[["Population"]]), length=11)

pop\_freq <- cut(state[["Population"]], breaks=breaks,

right=**TRUE**, include.lowest = **TRUE**)

table(pop\_freq)

A histogram is a way to visualize a frequency table, with bins on the x-axis and data count on the y-axis.

hist(state[["Population"]], breaks=breaks)

**Density Estimates**

Related to the histogram is a density plot, which shows the distribution of data values as a continuous line.

hist(state[["Murder.Rate"]], freq=**FALSE**)

lines(density(state[["Murder.Rate"]]), lwd=3, col="blue")

* Lwd = is an argument which is the line width

**Exploring Binary and Categorical Data**

For categorical data, simple proportions or percentages tell the story of the data.

**KEY TERMS FOR EXPLORING CATEGORICAL DATA**

***Mode***

The most commonly occurring category or value in a data set.

***Expected value***

When the categories can be associated with a numeric value, this gives an average value based on

a category’s probability of occurrence.

***Bar charts***

The frequency or proportion for each category plotted as bars.

***Pie charts***

The frequency or proportion for each category plotted as wedges in a pie.

barplot(as.matrix(dfw)/6, cex.axis=.5)

cex.axis – Specify the size of the tick label numbers/text with a numeric value of length 1