Kerala School of Mathematics
Course in Statistics for Scientists

Descriptive Statistics

T.Krishnan

Strand Life Sciences, Bangalore

Statistical Summaries

- may be single numerical summaries of a batch, such as average or, may be more complex tables and graphs
- reference to a given batch of data rather than to more general population
- closely related field is exploratory data analysis
- both exploratory and descriptive methods may lead to formulate laws or test hypotheses, but focus is on data at hand numbers of arrests by sex in 1985 for selected crimes
- male-female differences in patterns of arrests?

Male-Female Crime Rates Data

CRIME	MALES	FEMALES	CRIME	MALES	FEMALES	CRIME	MALES	FEMALES
Murder	12904	1815	Rape	28865	303	robbery	105401	8639
Assault	211228	32926	burglary	326959	26753	larceny	744423	334053
Auto	97835	10093	arson	13129	2003	battery	416735	75937
Forgery	46286	23181	fraud	151773	111825	embezzle	5624	3184
Vandal	181600	20192	weapons	134210	10970	vice	29584	67592
Sex	74602	6108	drugs	562754	90038	gambling	21995	3879
Family	35553	5086	dui	1208416	157131	drunk	726214	70573
Disorderly	435198	99252	vagrancy	24592	3001	runaway	53808	72473

Possible Problems with the Data:

- Know Your Batch
- Does not cover the gamut of crimes
- curfew and loitering violations, for example, left out
- May be omissions, maybe false arrests
- Definitions may vary between states
- Statistics may be falsified for political purposes

Summary Statistics

- more male than female arrests?
- Some may have been arrested more than once
- Summary Table

Statistic	MALES	FEMALES
No. of Cases	24	24
Minimum	5624	303
Maximum	1208416	334053
Sum	5649688	1237007
Arithmetic Mean	235404	51542
Standard Deviation	305947	74220

Some Questions:

- How about the average (mean) number of arrests for a crime?
- Does the mean make any sense to you as a summary statistic?
- Does standard deviation (s.d.) make any sense?
- Let us examine the shape of these numbers

Stem–Leaf Plots

- stem-and-leaf plot is like a tally
- Most significant digit is stem
- trailing digits leaves
- data are positively skewed toward larger numbers for both males and females

Stem–Leaf Plot for Males

```
Minimum      :     5624
  Lower Hinge :    29224.5
  Median      :   101618
  Upper Hinge :   371847
  Maximum     :  1208416
                  0 H 011222234579
                  1 M 0358
                  2   1
                  3 H 2
                  4   13
                  5   6
                  6
                  7   24
  * * * Outside Values * * *
                 12   0
```


Stem–Leaf Plot for Females

```
Minimum      :     303.000
Lower Hinge :    4482.500
Median      :   21686.500
Upper Hinge :   74205.000
Maximum     :  334053.000
                0 H 00000000011
                0 M 2223
                0
                0 H 6777
                0   99
                1   1
                1
                1   5
* * * Outside Values * * *
                3   3
```


T.Krishnan Strand Life Sciences, Bangalore Kerala School of MathematicsCourse in Statistics for Scientis

Median

- For skewed data mean gets pulled from centre toward the extreme
- A statistic that is not as sensitive to extreme values is the median.
- median is the value above which half the data fall
- if you sort data, the median is the middle value or average of two middle values
- for males the median is 101,618, and for females, 21,686
- Both are considerably smaller than means and more typical of the majority of number
- This is why the median is often used for representing skewed data, such as incomes, populations, or reaction times
- In this case it would be hard to characterize what median would represent

Sorting

- means, standard deviations, and medians as descriptive statistics in most cases.
- useful summary quantities when the observations represent values of a single variable
- in this example they are less appropriate, even if easily computable
- there are better ways to reveal patterns in these data
- let us look at sorting as a way of uncovering structure
- Sorting is one of the most basic and powerful data analysis techniques
- **stem-and-leaf plot, for example, is a sorted display**

Sorting by Frequency

- can sort on any numerical or character variable
- it depends on goal
- We began with a question:
- Are there male-female differences in patterns of arrests
- Sort male and female arrests separately
- we will get a list of crimes in order of decreasing frequency within sex
- connect similar crimes with lines
- number of crossings indicate differences in ranks

Sorted Data

MALES	FEMALES
dui	larceny
larceny	dui
drunk	fraud
drugs	disorderly
disorderly	drugs
battery	battery
burglary	runaway
assault	drunk
vandal	vice
fraud	assault
weapons	burglary
robbery	forgery
auto	vandal
sex	weapons
runaway	auto
forgery	robbery
family	sex
vice	family
rape	gambling
vagrancy	embezzle
gambling	vagrancy
arson	arson
murder	murder
embezzle	rape

Standardizing

- ranking is influenced by prevalence
- most frequent crimes occur at the top of the list in both groups
- comparisons within crimes are obscured by this influence
- men committed almost 100 times as many rapes as women, for example, yet rape is near the bottom of both lists
- if we are interested in contrasting the sexes on patterns of crime we must standardize the data

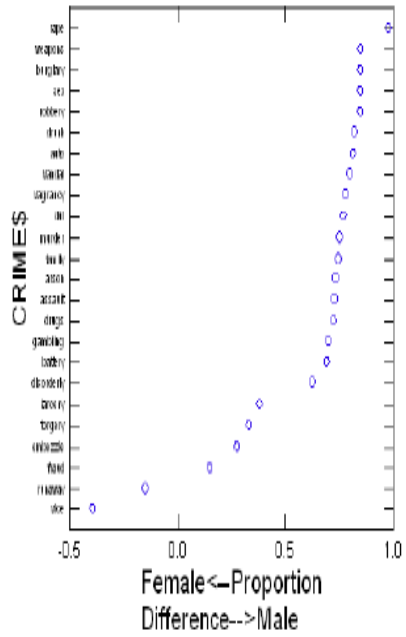
Methods of Standardization

- usually produced by subtracting means and then dividing by standard deviations
- another method is simply to divide by row or column totals
- for the crime data, we will divide by totals within rows (each crime)
- Doing so gives us the proportion of each arresting crime committed by men or women
- The total of these two proportions will thus be 1
- a contrast between men and women on this standardized value should reveal variations in arrest patterns within crime type
- By subtracting the female proportion from the male, we will highlight primarily male crimes with positive values and female crimes with negative

Plot of Differences

- Next sort these differences and plot them in a simple graph
- can see clear contrasts between males and females in arrest patterns
- predominantly aggressive crimes appear at the top of the list
- rape now appears where it belongs an aggressive rather than sexual crime
- a few crimes dominated by females are at the bottom

Gender Difference Plot



Appropriate Descriptors

- not all descriptors are appropriate
- mean, s.d. suitable for normal
- not appropriate for
  - Skewed
  - with outliers
  - mixtures

What Do You Want to Describe

ask what you want to describe:

- location
- spread
- asymmetry
- fatness of tail

Variable Types

- unordered categories
- ordered categories
- counts
- measurements?

Robust Descriptors

- descriptive statistic is called robust if the calculations are insensitive to violations of the assumption of normality
- robust measures include
  - median
  - quartiles
  - frequency counts
  - percentages
- if there are extreme observations
  - Trim (one-sided two-sided)
  - Winsorize

Graph Before Descriptors

a graph of data is useful to see if:

- symmetric or skewed
- there are outliers
- there is heterogeneity (there are groups) in which case separate descriptors for each group
- the tails are fat

Standard Descriptors

- number of observations (N)
- minimum
- maximum
- arithmetic mean (AM)
- geometric mean
- harmonic mean
- sum
- standard deviation
- variance
- coefficient of variation (CV)
- range
- median
- skewness
- kurtosis

Multivariate Descriptors

- Covariances
- Correlations

Geometric and Harmonic Means

- geometric mean is used (for positive numbers) for quantities of multiplicative in nature, e.g., population growth rate, interest rates
- Harmonic mean is used for averaging quantities like speed, number per rupee

Measures of Spread

- variance
- standard deviation
- range
- inter-quartile range (IQR) or mid-range

Correlations

- Pearson product moment correlation measures linear association of two quantitative variables
- other measures Spearman's rank correlation, Kendall's tau

Other Types of Data

- rank-order data
- unordered data
- binary data