OVERVIEW

• STATISTICS

PANIK ...THE THEORY AND METHODS OF COLLECTING, ORGANIZING, PRESENTING, ANALYZING, AND INTERPRETING DATA SETS SO AS TO DETERMINE THEIR ESSENTIAL CHARACTERISTICS...

HYPERSTAT ...IN THE BROADEST SENSE, "STATISTICS" REFERS TO A RANGE OF TECHNIQUES AND PROCEDURES FOR ANALYZING DATA, INTERPRETING DATA, DISPLAYING DATA, AND MAKING DECISIONS BASED ON DATA...
• DESCRIPTIVE - SUMMARIZE ESSENTIAL FEATURES OF DATA (CENTRAL TENDENCY, VARIABILITY, DISTRIBUTION)

INFERENTIAL - ESTIMATES, PREDICTIONS, FORECASTS, GENERALIZATIONS

PANIK ...INDUCTIVE STATISTICS, INFERRING SOMETHING ABOUT THE WHOLE FROM THE EXAMINATION OF ONLY ONE PART...
• POPULATION

DANIEL ...LARGEST COLLECTION OF ENTITIES FOR WHICH WE HAVE AN INTEREST AT A PARTICULAR TIME...

HYPERSTAT ...ENTIRE SET OF OBJECTS, OBSERVATIONS, OR SCORES THAT HAVE SOMETHING IN COMMON...EXAMPLE, A POPULATION MIGHT BE DEFINED AS ALL MALES BETWEEN THE AGES OF 15 AND 18...THE DISTRIBUTION OF A POPULATION CAN BE DESCRIBED BY SEVERAL PARAMETERS SUCH AS THE MEAN AND STANDARD DEVIATION...ESTIMATES OF THESE PARAMETERS TAKEN FROM A SAMPLE ARE CALLED STATISTICS...
• CENSUS
  ...COUNT EVERY ENTITY IN THE POPULATION...

• SAMPLE
  ...PART OF A POPULATION...

HYPERSTAT...SUBSET OF A POPULATION...SINCE IT IS USUALLY IMPRACTICAL TO TEST EVERY MEMBER OF A POPULATION, A SAMPLE FROM THE POPULATION IS TYPICALLY THE BEST APPROACH AVAILABLE...
• PARAMETER (POPULATION)

Hyperstat...Numerical quantity measuring some aspect of a population of scores...Example, the mean is a measure of central tendency...Parameters are rarely known and are usually estimated by statistics computed in samples...
• STATISTIC (SAMPLE)

HYPERSTAT ..."STATISTIC" IS DEFINED AS A NUMERICAL QUANTITY (SUCH AS THE MEAN) CALCULATED IN A SAMPLE. SUCH STATISTICS ARE USED TO ESTIMATE PARAMETERS...

HYPERSTAT ..."STATISTICS" SOMETIMES REFERS TO CALCULATED QUANTITIES REGARDLESS OF WHETHER OR NOT THEY ARE FROM A SAMPLE...EXAMPLES, BATTING AVERAGE, "GOVERNMENT STATISTICS"
• QUANTITATIVE (TEMPERATURE)
QUALITATIVE (GENDER)

THE QUALITATIVE-QUANTITATIVE DEBATE

HYPERSTAT...QUALITATIVE VARIABLES ARE SOMETIMES CALLED "CATEGORICAL VARIABLES"...QUANTITATIVE VARIABLES ARE MEASURED ON AN ORDINAL, INTERVAL, OR RATIO SCALE...QUALITATIVE VARIABLES ARE MEASURED ON A NOMINAL SCALE...
DISCRETE (NUMBER OF HOSPITAL ADMITS IN A DAY)
CONTINUOUS (HEIGHT)

SOME VARIABLES (SUCH AS REACTION TIME)
ARE MEASURED ON A CONTINUOUS SCALE...THERE IS AN
INFINITE NUMBER OF POSSIBLE VALUES THESE VARIABLES
CAN TAKE ON...OTHER VARIABLES CAN ONLY TAKE ON A
LIMITED NUMBER OF VALUES AND SUCH VARIABLES ARE
CALLED "DISCRETE" VARIABLES...
LEVELS OF MEASUREMENT

- NOMINAL (GENDER, ETHNICITY, RACE)

... NOMINAL MEASUREMENT CONSISTS OF ASSIGNING ITEMS TO GROUPS OR CATEGORIES
... NO QUANTITATIVE INFORMATION IS CONVEYED AND NO ORDERING OF THE ITEMS IS IMPLIED
... NOMINAL SCALES ARE QUALITATIVE RATHER THAN QUANTITATIVE
... FREQUENCY DISTRIBUTIONS ARE USUALLY USED TO ANALYZE DATA MEASURED ON A NOMINAL SCALE
... MAIN STATISTIC COMPUTED IS THE MODE
... REFERRED TO AS CATEGORICAL OR QUALITATIVE VARIABLES
• ORDINAL (ATTITUDE SCALE: 0, 1, 2, 3, 4, 5)

... ORDERED IN THE SENSE THAT HIGHER NUMBERS REPRESENT HIGHER VALUES
... INTERVALS BETWEEN THE NUMBERS ARE NOT NECESSARILY EQUAL
... NO "TRUE" ZERO POINT FOR ORDINAL SCALES SINCE THE ZERO POINT IS CHOSEN ARBITRARILY
• INTERVAL (TEMPERATURE)

**HYPERSTAT**...

... ONE UNIT ON THE SCALE REPRESENTS THE SAME MAGNITUDE ON THE TRAIT OR CHARACTERISTIC BEING MEASURED ACROSS THE WHOLE RANGE OF THE SCALE

... NO "TRUE" ZERO POINT

• RATIO (HEIGHT)

**HYPERSTAT**...

... RATIO SCALES ARE LIKE INTERVAL SCALES EXCEPT THEY HAVE TRUE ZERO POINTS
RANDOM SAMPLING

• RANDOM SAMPLE
  ...EACH MEMBER OF A POPULATION HAS THE SAME CHANCE OF BEING SELECTED...

• SIMPLE RANDOM SAMPLE
  ...EACH SAMPLE OF SIZE N IS SELECTED SUCH THAT EACH SUCH SAMPLE OF SIZE N HAS THE SAME CHANCE OF BEING SELECTED...
OTHER SAMPLE TYPES

- SYSTEMATIC
  ... Nth ENTITY

- STRATIFIED
  ... RANDOM ENTITIES WITHIN STRATA (GENDER, AGE GROUP, COUNTY)

- CLUSTER
  ... ALL ENTITIES WITHIN RANDOMLY SAMPLED CLUSTERS (COUNTY, CENSUS TRACT, ZIP CODE)

- CONVENIENCE
  ... WHATEVER
CRITICAL THINKING CHAPTER 1

AVERAGE AGE AT DEATH FOR VARIOUS OCCUPATIONS

STUDENTS...MEAN = 20.7

MOST DANGEROUS OCCUPATION = STUDENT

???
DESCRIPTING & COMPARING DATA

• TRIOLA...ONE DATA SET TO ILLUSTRATE CONCEPTS IN CHAPTER

SERUM COTININE LEVELS IN BLOOD (PRODUCED BY NICOTINE)...INDICATOR OF EXPOSURE TO CIGARETTE SMOKE

• FIVE (WELL, THREE + TWO) IMPORTANT CHARACTERISTICS OF DATA...

CENTER VARIATION DISTRIBUTION OUTLIERS TIME
• FREQUENCY DISTRIBUTIONS

DATA VALUES GROUPED INTO INTERVALS

CHOICE OF INTERVALS AFFECTS HOW YOU (OTHERS) SEE (INTERPRET) YOUR DATA...NO ONE RIGHT ANSWER AS TO WHAT INTERVALS ARE APPROPRIATE

USER-DEFINED QUARTILES NATURAL BREAKS STURGES RULE SOFTWARE-DEFINED

TABLES OR GRAPHICS
Using Statcrunch

Histograms
Using Statcrunch

Stem-and-Leaf

Variable: SMOKER
0 : 00002344
0 : 599
1 : 012233
1 : 56777
2 : 011233
2 : 55577899
3 : 1
3 :
4 :
4 : 89
Using Statcrunch

Boxplots

No Fences

With Fences
Using SAS

SERUM COTININE LEVELS IN THREE STUDY GROUPS

<table>
<thead>
<tr>
<th>ETS</th>
<th>NOETS</th>
<th>SMOKE</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>150</td>
<td>250</td>
</tr>
<tr>
<td>150</td>
<td>200</td>
<td>260</td>
</tr>
<tr>
<td>210</td>
<td>300</td>
<td>350</td>
</tr>
<tr>
<td>300</td>
<td>400</td>
<td>450</td>
</tr>
<tr>
<td>450</td>
<td>500</td>
<td>550</td>
</tr>
</tbody>
</table>

- ETS: 34, 2, 1, 1, 0, 2
- NOETS: 2, 1, 1, 0, 0, 0
- SMOKE: 11, 12, 14, 1, 2, 0
• OTHER GRAPHICS

SCATTER PLOTS (2 VARIABLES)

TIME SERIES (TREND LINES)

OGI VE (CUMULATIVE FREQUENCY)

MISCELLANEOUS (USER CREATIVITY !!!)
MEASURES OF CENTER

MEAN

Sum of all values divided by the number of values (very sensitive to extreme values)

MEDIAN

Midpoint - nothing to do with the values themselves (not sensitive to extreme values)

MODE

Most frequently occurring value (possibility of multiple modes, bimodal not uncommon)

Relationship of mean, median, mode (symmetric and skewed distributions)
### Measures of Variation

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range</td>
<td>Difference between minimum and maximum values</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>Measure of variation of values around the mean</td>
</tr>
<tr>
<td>Variance</td>
<td>Average squared deviation of a value from the mean (square of standard deviation) - denominator, $N$ versus $N-1$ and bias</td>
</tr>
<tr>
<td>Coefficient of Variation</td>
<td>Standard deviation expressed as a percentage of the mean</td>
</tr>
</tbody>
</table>
• STANDARD DEVIATION AND 'TYPICAL VALUES'  

EMPIRICAL RULE - IN A BELL-SHAPED DISTRIBUTION...

68% OF VALUES FALL WITHIN 1 STANDARD DEVIATION OF THE MEAN...95% OF VALUES FALL WITHIN 2 STANDARD DEVIATIONS OF THE MEAN...99.7% OF VALUES FALL WITHIN 3 STANDARD DEVIATIONS OF THE MEAN

ACCEPTED CONCEPT OF 'TYPICAL VALUES' --- WITHIN 2 STANDARD DEVIATIONS OF THE MEAN
• CHEBYSHEV'S THEOREM

REGARDLESS OF THE SHAPE OF A DISTRIBUTION, THE PROPORTION OF DATA LYING WITHIN K STANDARD DEVIATIONS OF THE MEAN IS \( AT \ LEAST \ldots 1 - \frac{1}{K^2} \)

AT LEAST 3/4 (75%) OF ALL VALUES LIE WITHIN 2 STANDARD DEVIATIONS OF THE MEAN

AT LEAST 8/9 (89%) OF ALL VALUES LIE WITHIN 3 STANDARD DEVIATIONS OF THE MEAN

• RATIONALE FOR STANDARD DEVIATION (AND VARIANCE)
• RELATIVE STANDING

Z-SCORES STANDARDIZED SCORES USING MEAN AND STANDARD DEVIATION
(UNUSUAL VALUES - JORDAN, LOBO, BOGUES - OUTSIDE OF +/- Z=2)

QUARTILES Q1, Q2, Q3
MEDIAN, MEDIANS OF UPPER AND LOWER HALF INTERQUARTILE RANGE (Q1 AND Q3)
EDA (EXPLORATORY DATA ANALYSIS - JOHN TUKEY, 1977)

EDA IS DETECTIVE WORK (NUMERICAL, COUNTING, GRAPHICAL)

EDA CAN NEVER BE THE WHOLE STORY, BUT NOTHING ELSE CAN SERVE AS THE FOUNDATION STONE, AS THE 1ST STEP

SEARCH FOR PATTERNS, GROUPS, UNEXPECTEDLY POPULAR VALUES, UNUSUAL VALUES, OUTLIERS

TRIOLA 5-NUMBER SUMMARY (MINIMUM, Q1, MEDIAN, Q3, MAXIMUM)

STEM-AND-LEAF PLOTS, BOX PLOTS