

Week 1: Descriptive Statistics and Probability

Lecture Handout

Introductory Terms

- Unit: the object on which a construct of interest is defined
- Population: the universe of units over which a construct of interest is defined
- Sample: the subset of the units in the population for which we observe data
- Data: a collection of measures of a construct of interest for sampled units
- Variable: a measure of a specific construct of interest for all sampled units
- Dependent Variable: a variable with an outcome that the data scientist is trying to understand
 - This is the **measured** variable, normally visualized on the y-axis of a graph
- Independent Variable: a variable with an outcome that the data scientist takes as given that relates to the dependent variable
 - This is the **changed** variable, normally visualized on the x-axis of a graph
- Measurement: the multi-stage process of translating a construct of interest into a variable
 - Stages of measurement:
 - 1) specifying the information to be collected
 - 2) sampling units to collect this information
 - 3) recording the information available
- Reliability: Do we get a similar measurement with repetition?
 - **Empirical** question
- Validity: Does the measure accurately capture the construct of interest?
 - **Theoretical** question
 - Forms of validity:
 - Face validity
 - Content validity
 - Construct validity

Central Tendency

Statistical Symbol or Equation	Explanation
$\bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n} = \frac{1}{n} \sum_{i=1}^n y_i$	Sample Mean: the summation of every sample unit (y _i) divided by the total number of sample units (n)
$\bar{y} = \frac{w_1 y_1 + w_2 y_2 + \dots + w_n y_n}{w_1 + w_2 + \dots + w_n} = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}$	Weighted Sample Mean: adjusts the sample mean according to each sample unit's corresponding magnitude (w _i)
∀	Universal quantifier, means “for all instances”

Concept Overview

- Central Tendency (def): refers to the information about a typical value of a variable within a sample
- Most intuitive measure: the sample mean
 - (Def) the summation of every sample unit (y_i) divided by the total number of sample units
 - Or, the weighted sample mean: adjusts the sample mean according to each sample unit's corresponding magnitude
 - ex) college course grading: in a given mathematics course, the grading is as follows:
 - midterm=30%
 - final=60%
 - participation=10%
 - Weighted sample mean for the course grade:
$$(.30)(grade_M) + (.60)(grade_F) + (.10)(grade_P) = G_C$$
- Outlier: a small number of observations that take on values that are substantially different than the rest of the sample
 - Two reasons why outliers are important:
 - 1) could be evidence of a data error
 - 2) can make mean unrepresentative of typical outcome
- Skew: whether the non-erroneous outliers tend to be larger or smaller than the sample mean
 - Mean < median = left-skewed distribution (small outliers)
 - Mean > median = right-skewed distribution (big outliers)
- Ordered Sample: sample such that if $i < j$, then sample unit $y_i \leq y_j$
- Sample median: observation that falls in the middle of an ordered sample
 - When n is odd: median = $y_{(n+1)/2}$
 - When n is even: median = $(y_{n/2} + y_{1+(n/2)})/2$
- Mode: most common value
 - Variables are said to be bimodal if the most common values aren't near each other

Concept Application in R

Mean, Median, and Mode

```
#mean and median
mean(state$Murder.Rate)
median(state$Murder.Rate)
weighted.mean(state$Murder.Rate,
              w=state$Population)
weighted.median(state$Murder.Rate,
               w=state$Population)

#creating function for mode
getmode <- function(v){
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
```

Variability

Statistical Symbol or Equation	Explanation
$s = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}}$	Standard deviation: most common measure of variability of a sample where the deviation of a sample unit y_i from the sample mean is $y_i - \bar{y}$
$mad = \frac{\sum_{i=1}^n y_i - \bar{y} }{n}$	Mean absolute deviation: considers the absolute value of deviation within a sample

Concept Overview

- Variability (def): a measure of the divergence of units within a sample from the sample mean and from each other (i.e. how “spread out” the dataset is)
 - The standard deviation is the most common measure of variability with 2 standard deviations being a generally accepted boundary for variable distribution (i.e. 95% of observations fall between $(\bar{y}-s, \bar{y}+s)$)
- Range: the spread between the minimum and maximum values in a sample
 - Interquartile range: the spread between the 25th and 75th percentiles

Concept Application in R

Variability and Standard Deviation

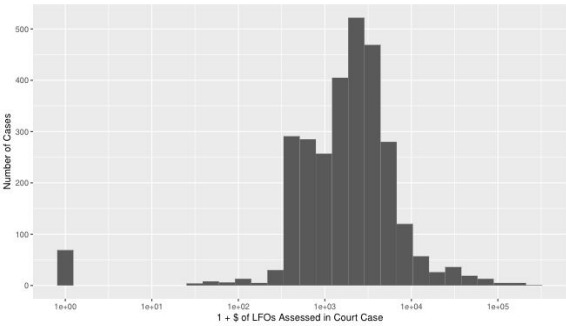
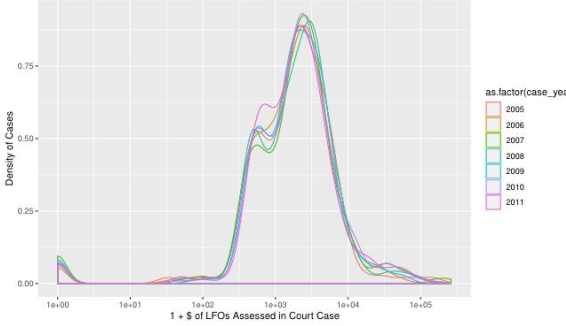
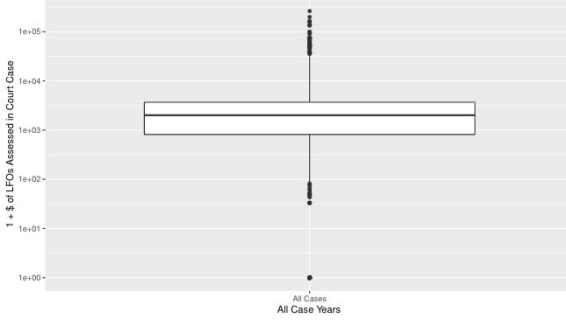
```
var(state$Murder.Rate)
sd(state$Murder.Rate)
mad(state$Murder.Rate)
install.packages("matrixStats")
library(matrixStats)
weightedVar(state$Murder.Rate, w=state$Population)
```

Range

```
max(state$Murder.Rate)-min(state$Murder.Rate)
pcts <- quantile(state$Murder.Rate,
                  p=c(0.05, .25, .5, .75, .95))
print(pcts)
pcts[4]-pcts[2]
install.packages("Hmisc")
library(Hmisc)
wtd.quantile(state$Murder.Rate,
              w=state$Population,
              p=c(.05, .25, .5, .75, .95))
```

How to Communicate Descriptive Statistics

Concept Overview

Communication/Visualization Method	Explanation
<p>Legal Financial Obligations (LFOs) in Alabama Court Cases</p> <pre>===== Statistic N Mean St. Dev. Min Pctl(25) Pctl(75) Max ----- LFOs Paid to Date 2,926 655.354 1,543.663 0 0 785.8 40,328 LFOs Assessed 2,926 4,231.619 11,830.360 0 813.3 3,680 262,245 Year of Case 2,926 2,007.978 1.981 2,005 2,006 2,010 2,011 =====</pre>	Descriptive Statistics Table, using Stargazer (example code shown below)
 <p>A histogram showing the distribution of LFOs assessed in court cases. The x-axis is labeled '1 + \$ of LFOs Assessed in Court Case' and uses a logarithmic scale from 1e+00 to 1e+05. The y-axis is labeled 'Number of Cases' and ranges from 0 to 500. The distribution is right-skewed, with a peak around 1e+03.</p>	Histogram: <code>geom_histogram()</code> (with <code>ggplot()</code>)
 <p>A kernel density plot showing the distribution of LFOs assessed in court cases by year. The x-axis is labeled '1 + \$ of LFOs Assessed in Court Case' and uses a logarithmic scale from 1e+00 to 1e+05. The y-axis is labeled 'Density of Cases' and ranges from 0.00 to 0.75. The plot shows multiple overlapping density curves for the years 2005 through 2011, with a legend on the right.</p>	Kernel Density Plot: <code>geom_density()</code> (with <code>ggplot()</code>)
 <p>A boxplot showing the distribution of LFOs assessed in court cases. The y-axis is labeled '1 + \$ of LFOs Assessed in Court Case' and uses a logarithmic scale from 1e+00 to 1e+05. The x-axis is labeled 'All Cases' and 'All Case Years'. The boxplot shows the median, quartiles, and outliers.</p>	Boxplot: <code>geom_boxplot()</code> (with <code>ggplot()</code>)

Concept Application in R - Descriptive Statistic Table with Stargazer

Code:

```
library(stargazer)
alacourt <- read.csv("~/Dropbox/DATA 201/Data/AlabamaCourt.csv")
head(alacourt)
alacourt_lfos <- subset(alacourt, select=
                        c(amountpaid, amountdue, case_year, race))
stargazer(alacourt_lfos, type = "text")
#add a title
stargazer(alacourt_lfos, type = "text",
          title="Descriptive Stats Table")
#adding median (or any descriptive stat) to the table
stargazer(alacourt_lfos, type = "text",
          title = "Descriptive Stats Table",
          summary.stat=c("n", "mean", "sd", "min",
                        "p25", "median", "p75", "max"))

#specify number of decimal digits
stargazer(alacourt_lfos, type = "text", digits=2)
#specify subset of variables
stargazer(alacourt_lfos[c("amountdue", "amountpaid")], type = "text")
stargazer(subset(alacourt_lfos, alacourt_lfos$case_year == 2011), type = "text")
#labeling variables
stargazer(alacourt_lfos, type="text",
          title="LFOs in Alabama Court Cases",
          covariate.labels=c("LFOs Paid to Date",
                            "LFOs Assessed",
                            "Year of Case"))
```

Probability

Statistical Symbol or Equation	Explanation
$\{ \}$ vs. $()$, $[]$	$\{ \}$ = discrete sample space $()$ = continuous sample space, excluding boundaries $[]$ = continuous sample space, including boundaries
\cup vs. \cap	\cup = union ex) $A \cup B$ = the simple events that occur in EITHER A and B \cap = intersection ex) $A \cap B$ = the simple events that occur in BOTH A and B

	- <u>Property</u> : $p(A \cup B) = p(A) + p(B) - p(A \cap B)$
A^c	Complement: the set of simple events in sample space S that are not included in A - <u>Property</u> : $p(A^c) = 1 - p(A)$
\emptyset	Null set: set that does not contain any values ex) A and B are disjoint events (no simple events are contained in both A and B) when $A \cap B = \emptyset$
\subset	Subset of an event ex) $A \subset B$ denotes that all of the simple events that are contained in A are also contained in B - <u>Property</u> : $A \subset B \rightarrow p(A) \leq p(B)$
$p(B A) = \frac{p(A \cap B)}{p(A)}$	Conditional Probability
$p(A \cap B) = p(A)p(B)$	Independent event

Concept Overview

- Sample Space: the set of all outcomes of an experiment
 - Discrete sample space: contains a finite number of outcomes
 - Continuous sample space: contains an infinite number of outcomes
- Events: all possible subsets of a sample space S
 - Simple event: corresponds to a single outcome
 - Compound event: corresponds to at least two simple events
 - ex) sample space $S = \{1, 2, 3, 4\}$
 - Simple event: $\{1\}$
 - Compound event: $\{2, 3\}$
- Function
- Disjoint Events: given datasets A and B, there are no simple events contained in both A and B
- Subset: given datasets A and B, all simple events of A are also contained in B
- Conditional Probability: the probability that an event will occur given that another event has already occurred
 - Denoted by $p(B | A)$
- Independence: two events are independent if one event's occurrence does not affect the probability of the other from occurring
 - If independent, then:
 - $p(A \cap B) = p(A)p(B)$
 - $p(B | A) = p(B)$
 - $p(A | B) = p(A)$