Statistics for Linguists
Hale
January 27, 2010

Measurement

There is a distinction between an experiment and an observational study.

observational study A corpus study is typically observational; the linguist doesn't decide what people say, but merely records what is said. With this data, sometimes an association can be established e.g. repetition disfluencies tend to precede heavier constituents.

(controlled) experiment The experimenter (randomly) assigns subjects to levels of some 'treatment' e.g. auditory or written stimuli. Then he or she measures some response — the 'dependent' variable. Typical dependent variables include number of mistakes on a test, reaction times, acceptability judgments etc.

Different kinds of data support different kinds of inference.

nominal Named properties which have no meaningful order on a scale of any type are called 'nominal'. Examples: What language is being observed? What dialect? Which word or variant, i.e. *going to* or *gonna*? What is the gender of the person being observed? There is no average fruit.

ordinal Orderable properties are called 'ordinal'. They aren't observed on a measurable scale, but this kind of property is transitive so that if $a < b$ and $b < c$ then also $a < c$. Example: Zipf's rank frequency of word-attestations, rating scales like Strongly Agree, Agree, Disagree, Strongly Disagree. The order of finishers in a cross-country race is ordinal. It doesn't matter how many seconds ahead of the next runner over the finish line you were — they are still one rank behind you.

interval Properties measured on a scale that does not have a true zero value are called 'interval'. On an interval scale, the magnitude of differences of adjacent observations can be determined (unlike the adjacent items on an ordinal scale), but because the zero value on the scale is arbitrary the scale cannot be interpreted in any absolute sense. Examples: temperature (Fahrenheit or Centigrade scales), magnitude estimates such as "This soup is ten times hotter than that other soup."

ratio When measuring on a scale that does have an absolute zero value, the data are called 'ratio'. They get this name because ratios of these measurements are meaningful. For instance, a vowel that is 100 msec long is twice as long as a 50 msec vowel, and 200 msec is twice 100 msec. Contrast this with temperature where 80 degrees Fahrenheit is not twice as hot as 40 degrees. Examples: Acoustic measures like sound frequency, durations such as reaction time, distances e.g. centimeters.

Cherokee voice onset times

```
> vot <- read.table("cherokeeVOT.txt", header = T)
> head(vot)
```

```
  VOT year Consonant
1  67 1971         k
2 127 1971         k
3  79 1971         k
4 150 1971         k
5  53 1971         k
6  65 1971         k
```

```
> vot$Consonant == "k"
```

```
 [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE
[13] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[25]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
[37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
> attach(vot)
> mythresholds <- c(0, 70, 100, 200)
> coarse <- cut(VOT, mythresholds)
> head(coarse, n = 10)
```

```
 [1] (0,70]    (100,200] (70,100]  (100,200] (0,70]    (0,70]    (70,100]
 [8] (100,200] (100,200] (100,200]
Levels: (0,70] (70,100] (100,200]
```

A **factor** is a kind of indexing vector with names. The factor “coarse” divides up VOTs into classes. The half-open interval $(70, 100]$ contains 100 but not 3. Lets give them more humane names.

```
> bins <- cut(VOT, mythresholds, c("short", "unsure", "long"))
```

In this data, “year” is meant as a cross-classifying **factor**, not a measurement. Turn this column of the data frame into a factor using assignment like this:

```
> year <- factor(year)
```

It’s easy to ask which observations occupy particular categories

```
> Consonant[bins == "short" & year == 2001]
```

```
[1] k t t t t t t t
Levels: k t
```

```
> Consonant[bins == "unsure" & year == 2001]
```

```
 [1] k k k k k k k k t t t t t
Levels: k t
```

```
> Consonant[bins == "long" & year == 2001]
```

```
[1] k k k k t
Levels: k t
```

The `table` function summarizes our measurements according to the new factor “bins.”

```
> table(bins)
```

```
bins
 short unsure   long
    11     15     18
```

Check: if our factorization into short, unsure & long was exhaustive then the sum should add up to the total number of observations.

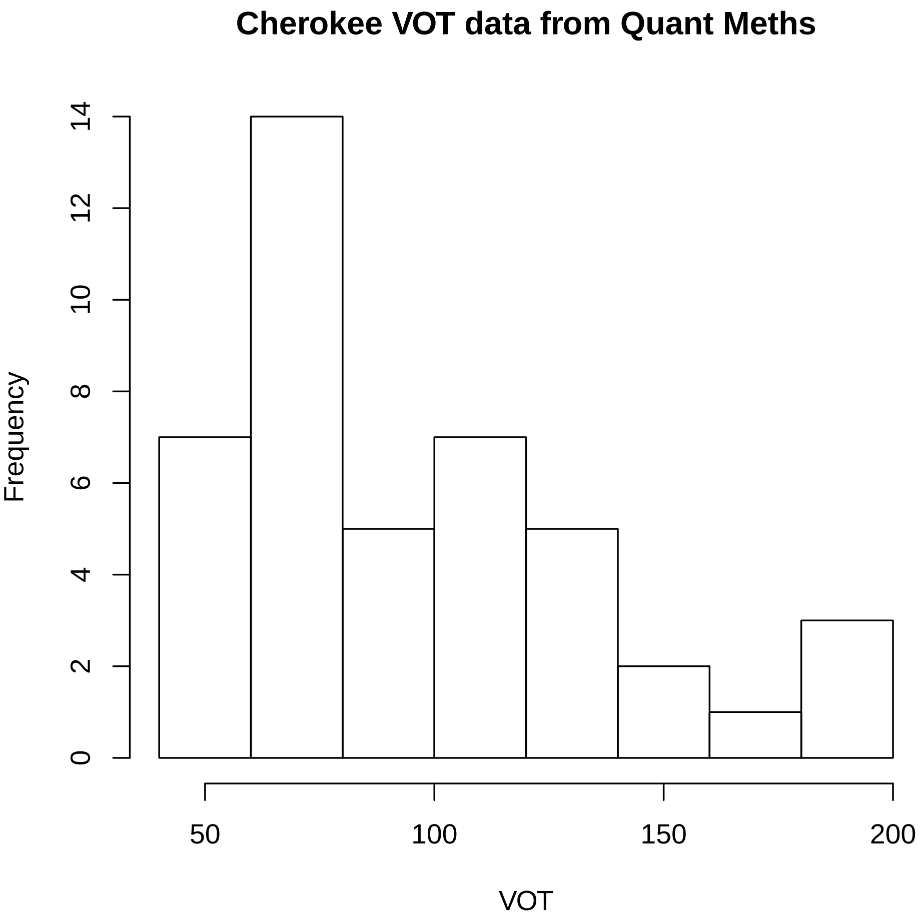
```
> 11 + 15 + 18 == length(VOT)
```

```
[1] TRUE
```

Saying that eleven of the observations were “short” means that $\frac{11}{44}$ ths of the observations meet the criterion of being 70ms or less. In other words, 25% of the measurements satisfy this definition of being short VOTs. The **relative frequency** of shortness is 0.25.

A **histogram** visually summarizes relative frequencies.

```
> hist(VOT, main = "Cherokee VOT data from Quant Meths")
```



By default, R chooses class intervals to ensure that the area of each vertical bar is proportional to the number it represents. Setting `freq=F` yields a histogram whose vertical axis is a proportion, rather than a count as in Figure 1.7 on page 16 of Johnson.

What if we wanted to ask more specific questions? Considering just the 2001 observations,

```
> ohone <- subset(vot, year == 2001)
```

what fraction have VOTs lasting 97ms or less? The answer[1] to questions like is called the **empirical cumulative distribution function**.

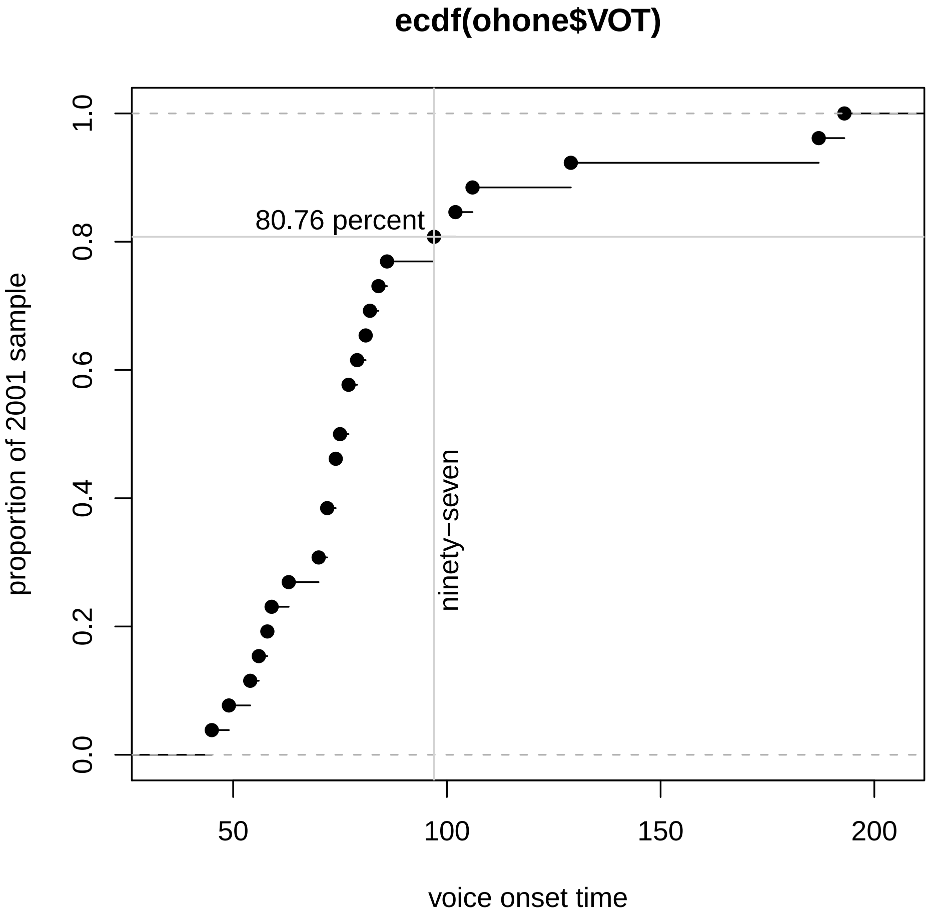
```
> cumulative <- ecdf(ohone$VOT)
> cumulative(97)

[1] 0.8076923
```

[1] You can plot your own empirical CDF by sorting the data points like this: `n <- length(VOT); plot(sort(vot$VOT),(1:n)/n,type="s",ylim=c(0,1))`

The **relative cumulative frequency curve** graphically summarizes all possible arguments to the `cumulative` function we just created.

```
> plot(cumulative, xlab = "voice onset time", ylab = "proportion of 2001 sample")
> abline(v = 97, col = "lightgray")
> text(101, 0.35, "ninety-seven", srt = 90)
> abline(h = cumulative(97), col = "lightgray")
> text(75, 0.83, "80.76 percent")
```



The VOT levels at which 25%, 50% and 75% of the empirical CDF are accounted-for are known as the **first quartile**, **median** and **third quartile** respectively. By definition, half the observations are above the median and half are below.

```
> quantile(ohone$VOT)

    0%    25%    50%    75%   100% 
 45.00  64.75  76.00  85.50 193.00 
```

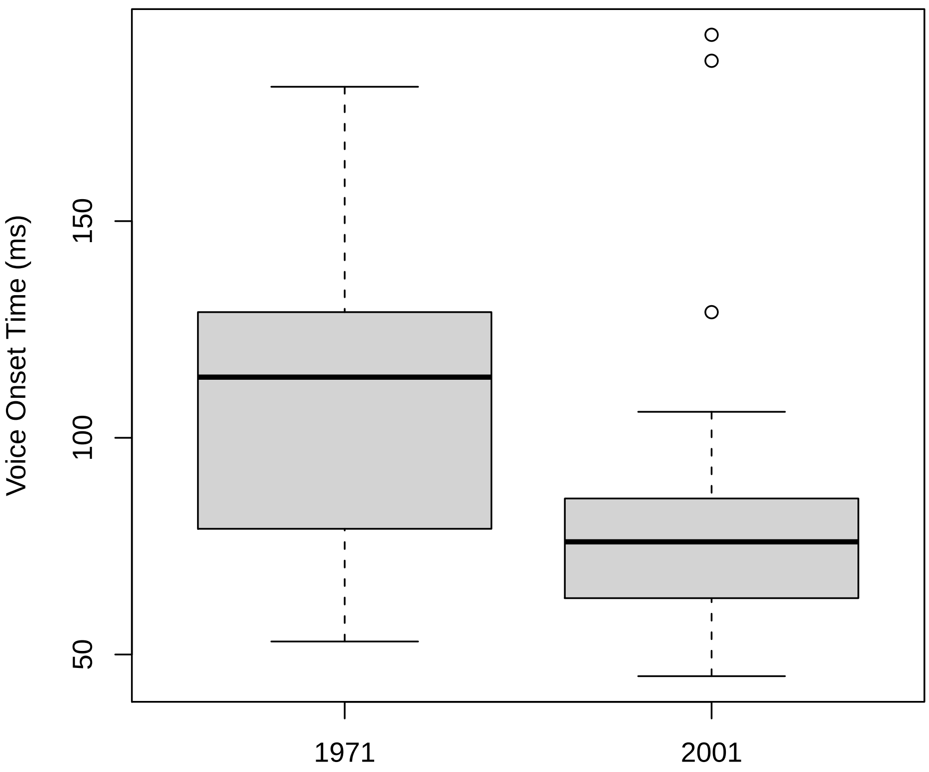
The quartiles help one to get a general handle on a batch of data. When combined with the **minimum** and **maximum** they are known as the Five Number Summary. The R function `summary` gives you the average or **mean** as well.

```
> summary(vot$VOT[vot$year == "2001"])

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  45.00   64.75   76.00   84.65   85.50  193.00 
```

These same five numbers are shown in a **boxplot**.

Figure 3.1 from page 71 of Johnson



The box spans the interquartile range, while the whiskers extend up to the maximum and minimum values – or 1.5 times the size of the box, whichever is smaller. In the case of this VOT data, there are a couple observations that exceed this threshold. These extreme values are rendered as circles. Johnson deems the most extreme two in the 2001 data “outliers” on page 16. They are inconsistent with the expectation that this data is “normally-distributed” — more on that later!

Homework

- Read chapter 1 of Johnson, as well as Vasishth/Broe.