STAT 3743 • PROBABILITY & STATISTICS • FALL 2010 • KERNS

R Lab #1

Name: ANSWER KEY

Note: the questions are randomly generated so these may (not) exactly match those on your paper. The answers below are for *these* and if you have trouble seeing the connection between these and those, ask me.

Directions: for each of the following variables,

1. Construct several visual displays (appropriate for the data type). Sketch by hand on a piece of paper.
2. Comment on what you have learned about the data (I'm thinking CUSS).
3. Comment on which displays capture those features, and which ones do not.

Here are the variables:

- `Formaldehyde$carb` (see `?Formaldehyde`)
- `InsectSprays$count`
- `cars$speed`
- `chickwts$weight`
- `infert$education`

(Selected) Solutions:

For variable `Formaldehyde$carb`: We read in the help file that the data are from a chemical experiment to prepare a standard curve for the determination of formaldehyde by the addition of chromatropic acid and concentrated sulphuric acid and the reading of the resulting purple color on a spectrophotometer. The variable `carb` is thus a quantitative, continuous variable, Carbohydrate (ml). We can even take a look at these data directly, since the sample size is so small.

```
> Formaldehyde
```

```
  carb optden
1  0.1  0.086
2  0.3  0.269
3  0.5  0.446
4  0.6  0.538
5  0.7  0.626
6  0.9  0.782
```

The types of plots appropriate for quantitative data include histograms, strip charts, stemplots, boxplots, ..., the list continues. Which one is appropriate for these data? Let's look at a stemplot. Rather than typing that whole big name every time, let's save the data in a shorter variable name.

```
> v <- Formaldehyde$carb
```

Now on to the stemplot.

```
> library(aplpack)
> stem.leaf(v)

1 | 2: represents 1.2
 leaf unit: 0.1
            n: 6
   1    0* | 1
   2     t | 3
  (1)    f | 5
   3     s | 67
   1    0. | 9
```

From this we see that the sample median Carbohydrate level is approximately 0.5 (in fact, we can read it from the original data to be 0.55). The range of the data (typically a bad measure, but not so bad for this small data set) is 0.1 to 0.9, and it's hard to see any reliable shape.

A good way to get several descriptive statistics on a variable like this is with the `summary` command.

```
> summary(v)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1000  0.3500  0.5500  0.5167  0.6750  0.9000
```

There we get two measures of center and the quartiles.

Now let's take a look at another visual display. This is a small data set, so let's try a stripchart.

```
> stripchart(v)
```

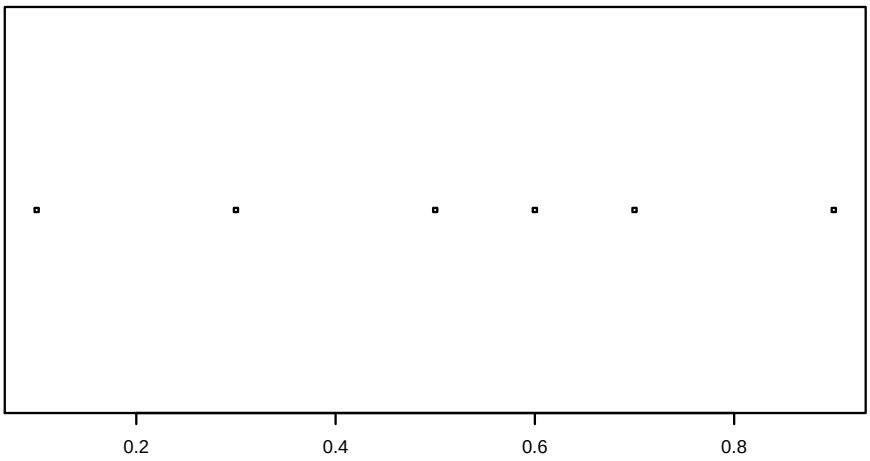

Figure 1: Stripchart of `Formaldehyde$carb`

This gives us a good picture of the data; they spread from just above zero to almost 0.10, and they are nearly uniformly distributed throughout that range. The center looks to be around 0.55. The plot gives us our first tangible idea about shape: these data are not peaked in the middle with heavy tails, rather, they are flat all over with thin tails. We would suspect that a distribution like this would be platykurtic. We're here, so let's check.

```
> library(e1071)
> kurtosis(v)

[1] -1.622692
```

Yes, it is just as we suspected. By the way, the data look ever so slightly right-skewed. Let's check that as well.

```
> skewness(v)  # the e1071 package is already loaded

[1] -0.1388600
```

We are batting one-thousand, so far. There do not seem to be any unusual features to this data set, which should not be a surprise given the data set's size.

Now that we've seen a couple visual displays, let's try another (a boxplot).

```
> boxplot(v, horizontal = TRUE)
```

The boxplot is shown below. It shows a center of around 0.55 (located at the median), the IQR is approximately $0.7 - 0.3 = 0.4$, and there are no extreme values indicated. The boxplot hints at an ever-so-slight right skewness. We should try to bear in mind, however,

that its a little bit silly to draw a boxplot for only six observations. We really don't need a summary display; we would be just as well to look at the individual data values with a stripchart.

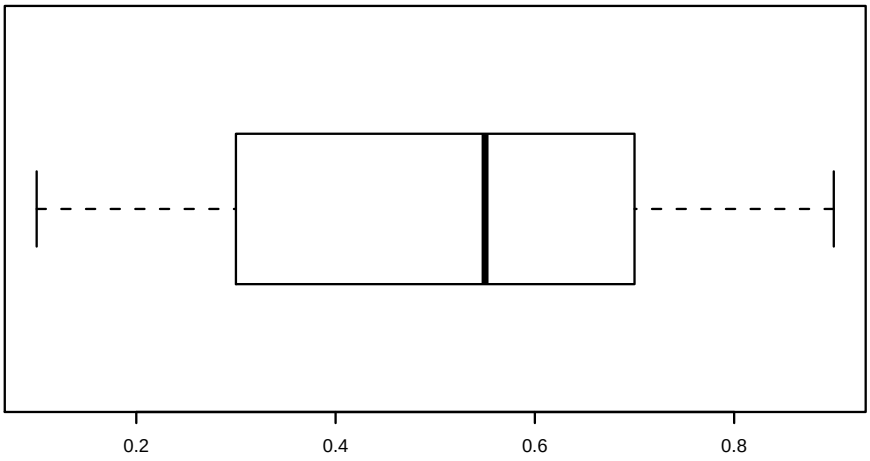

Figure 2: Boxplot of `Formaldehyde$carb`

This data set is a good example of what to do when we have a small data set. Boxplots (and histograms, as well) serve to summarize a data set, and are not so helpful for small data. We can estimate things like center and spread, but it is difficult to really go to the bank with those estimates, because we really do not have a lot of information from the population. The same remarks apply to judgements of shape: they are tentative, at best.

For these data, we would say that the boxplot is not so useful, but the stripchart and stemplot do a pretty good job. If I had to pick a best one, I would say the strip chart.

For variable `InsectSprays$count`**:** We read in the help file that the data are counts of insects in agricultural experimental units treated with different insecticides. Let's save ourselves some typing right now.

```
> v <- InsectSprays$count
```

The sample size is

```
> length(v)

[1] 72

> v
```

```
 [1] 10  7 20 14 14 12 10 23 17 20 14 13 11 17 21 11 16 14 17 17 19 21  7 13  0
[26]  1  7  2  3  1  2  1  3  0  1  4  3  5 12  6  4  3  5  5  5  5  2  4  3  5
[51]  3  5  3  6  1  1  3  2  6  4 11  9 15 22 15 16 13 10 26 26 24 13
```

Here we go; now we have a reasonably sized data set. These are counts, so we are looking at quantitative discrete data. Let's try a stemplot.

Now on to the stemplot.

```
> library(aplpack)
> stem.leaf(v)

1 | 2: represents 12
 leaf unit: 1
            n: 72
   8    0* | 00111111
  20     t | 222233333333
  31     f | 44445555555
  37     s | 666777
  (1)   0. | 9
  34    1* | 000111
  28     t | 223333
  22     f | 444455
  16     s | 667777
  10    1. | 9
   9    2* | 0011
   5     t | 23
   3     f | 4
   2     s | 66
```

Nice. What a lot of information! The median is shown to be 9 insects, the counts range from 0 to 26, and the shape is clearly right skewed. There are no extreme values identified in the stemplot, but there look to be two "humps" on either side of the median, and there may be another hump up higher, around 20. But this makes sense: we already knew that the data were about different groups of insects treated with different insecticides, so what we may be seeing is different performance (or kill rates) for the different insecticides. See how the depths drop off in an unbalanced way on either side of the median? Such behavior is common with skewed data.

Let's try something else, maybe a boxplot.

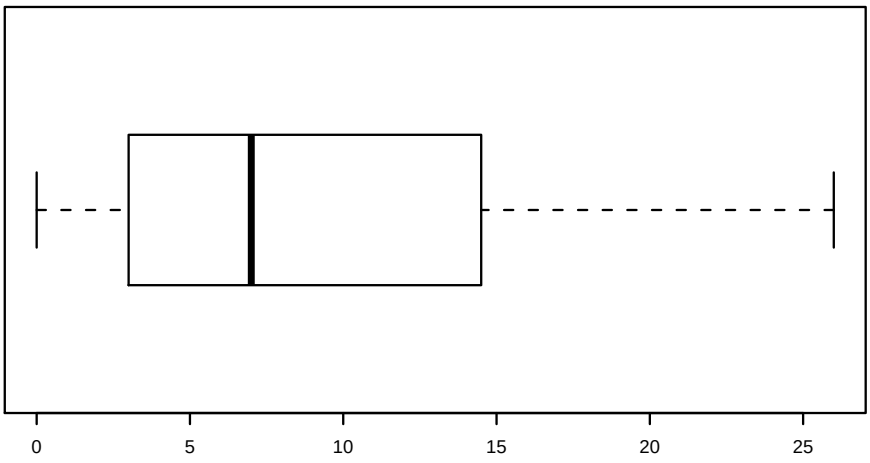

Figure 3: Boxplot of `InsectSprays$count`

```
> boxplot(v, horizontal = TRUE)
```

Well, the boxplot shows the center as before, we can read a measure of spread (approximately the IQR) to be around 12, and the longer whisker on the right side is suggestive of the right skewness. Note, however, that we lost all information about the multiple humps. Boxplots are blind to such things. As a redeeming feature we can also see that there are no extreme values in this data set, just like the stemplot. The boxplot has definitely summarized the data, perhaps too much in this case.

We haven't tried a histogram, yet, let's look at one of those.

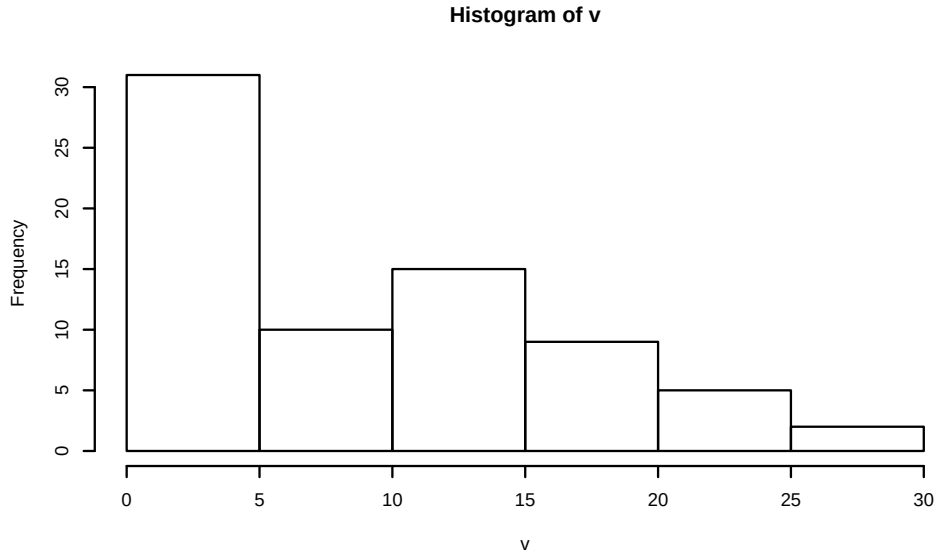

Figure 4: Boxplot of `InsectSprays$count`

```
> hist(v)
```

The default histogram is OK. We see the right skewness, and there is even a hint of maybe multiple humps. We can't tell the actual highest value, though; for all we know, the maximum value could even have been 30! All in all, the default histogram is so-so in its performance with these data. Another thing we could try is to fiddle with the number of bins. Different bins means different histograms. Here is a histogram with (approximately) 15 bins.

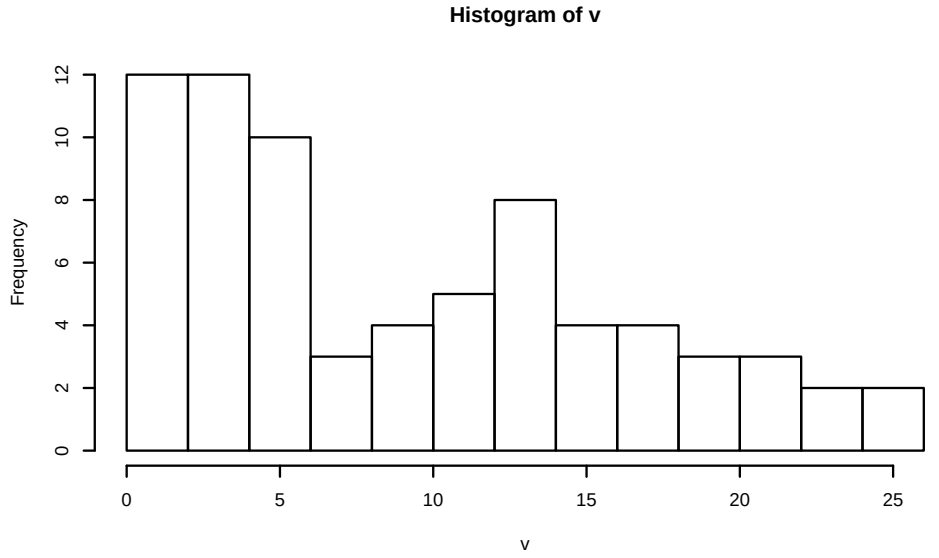

Figure 5: Boxplot of `InsectSprays$count`

```
> hist(v, breaks = 15)
```

This is definitely better. Now we can tell that the max is 26, and the two humps are more prominent. For these data, however, even this histogram does not hold a candle to the stemplot. Stemplot is a winner.

For variable `infert$education`**:** The help file says that the data come from a matched case-control study dating from before the availability of conditional logistic regression. The education variable is shown to be a categorical (qualitative) variable, it is a factor, and the levels are intuitively ordered (but note they are not represented internally as an ordered factor). We can see the first part of the data with a command like this:

```
> v <- infert$education

> v[1:5]

[1] 0-5yrs  0-5yrs  0-5yrs  0-5yrs  6-11yrs
Levels: 0-5yrs 6-11yrs 12+ yrs
```

The sample size is 248, which is the largest we have yet seen. We may quickly make a frequency table like this:

```
> table(v)

v
 0-5yrs 6-11yrs 12+ yrs
     12     120     116
```

There are three categories; the majority of observations are in the second category while there are relatively few observations in the first category. Let's take a look at some visual displays of the data.

```
> barplot(table(v))
```

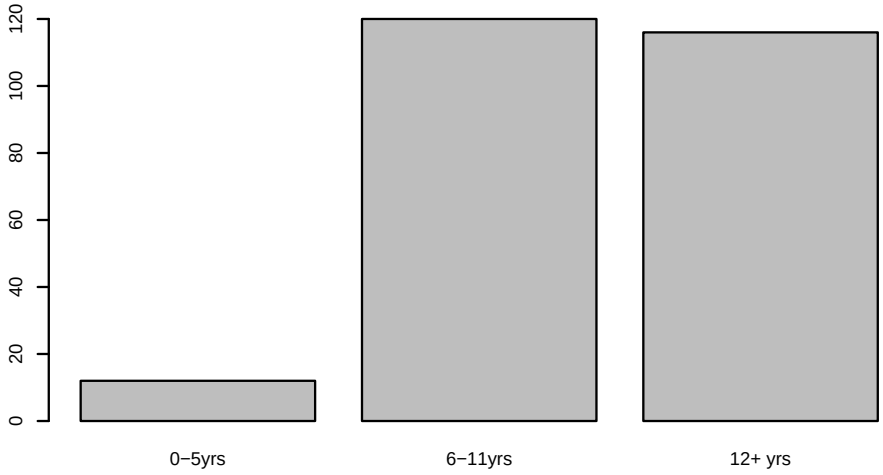

Figure 6: Bar graph of `infert$education`

```
> dotchart(table(v))
```

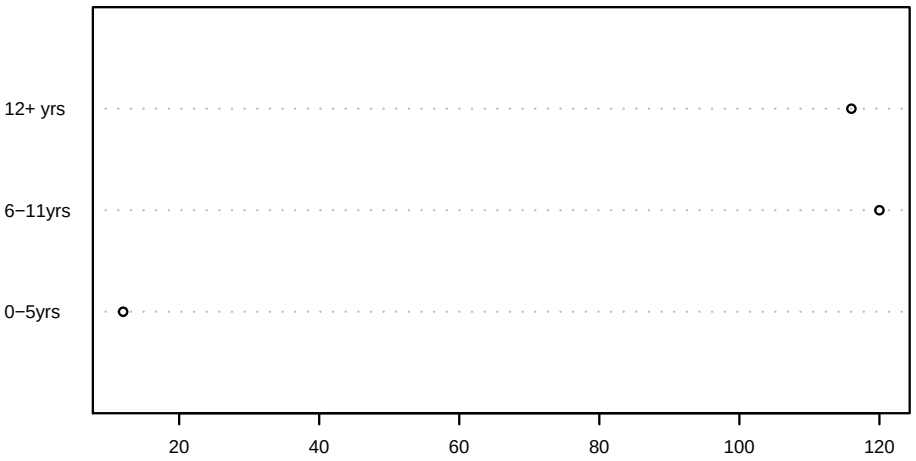

Figure 7: Cleveland dot chart of `infert$education`

```
> library(qcc)
> pareto.chart(table(v))
```

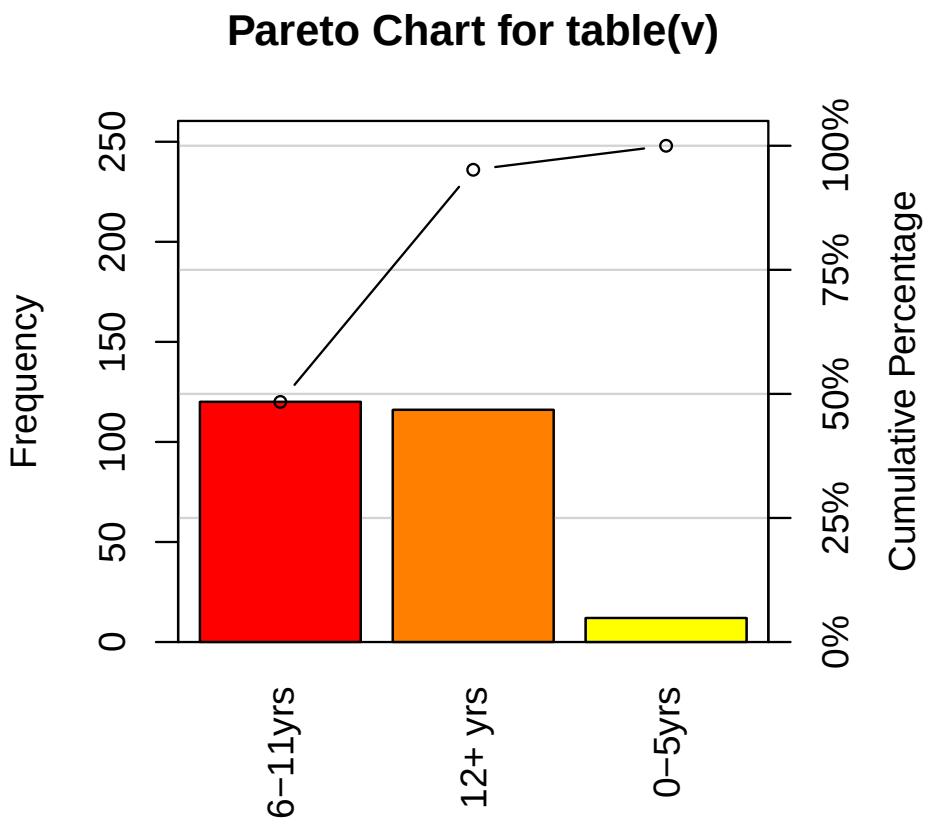

Figure 8: Pareto diagram of `infert$education`

In all honesty, all three graphs convey essentially the same information for these data. There are only three categories, and the relationship between them is reasonably clear. There are no clear winners. It is a good thing when multiple visual displays suggest the same information, or better put, it is BAD when radically different messages are conveyed depending on the display used. We should count our lucky stars in this case and rest for a moment, rest with the knowledge that the next data set we encounter will likely not be so simple.