PS6 Lecture 2

Daniel Lim

University of California, Los Angeles

4/2/2014

Agenda

- Histogram
- Measures of spread
- Boxplots
- Subsetting data

Histograms

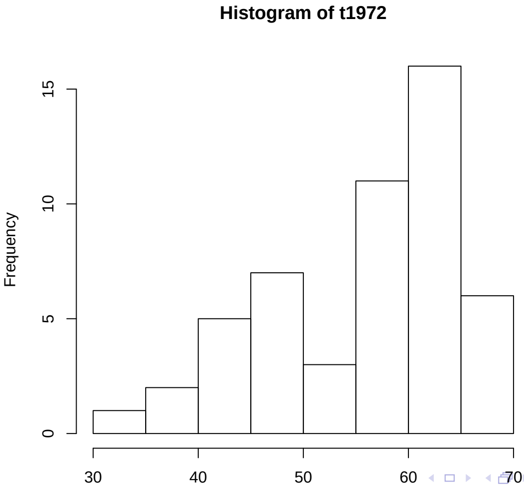
- Previously, we considered measures of central tendency. . . but only as numbers.
- Graphing the data contextualizes and helps us better understand what's going on.
- We'll use a **histogram** – the first of several basic graphs that we'll learn in this course – to visualize the turnout data.

Histograms

To create. . .

1. divide up the *range* of the data into several *bins*
2. count how many observations of the data fit into each bin
3. draw a bar for each bin with height corresponding to the number of observations in that bin.

```
> hist(t1972)
```



Histograms

A few observations we can immediately make.

- there are 8 bins at intervals of 5
- there are 2 bins, at 45-50 and 60-65, where there are *local* peaks - we say that these data are **bimodal**
- the *actual* mode is *probably* somewhere in the range 60-65[1].

[1] This may or may not be true depending on the actual values, and rounding.

Histograms

Add lines for the mean and median to the histogram using the 'abline' command:

1. save the mean and median to variables.
2. draw the histogram
3. call 'abline' using the 'v' argument (vertical).

Histograms

The previous steps, in R code:

```
> myMean = mean(t1972)
> myMedian = median(t1972)
> hist(t1972)
> abline(v = myMean, col='red', lwd=3, lty=1)
> abline(v = myMedian, col='blue', lwd=4, lty=2)
```

Histograms

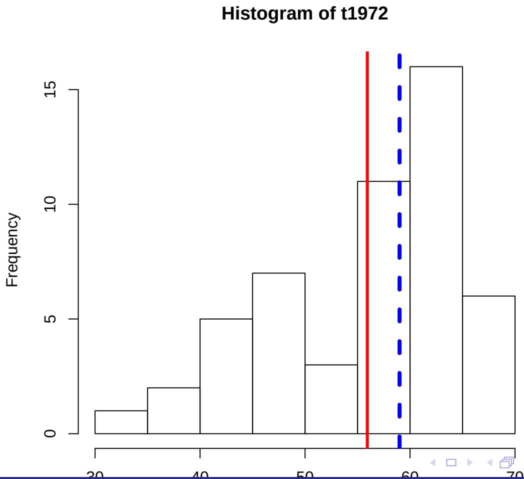
Notice the additional arguments used in the ‘abline’ commands from the previous slide:

‘col’ (color) changes the color of the line.

‘lwd’ (line width) set line thickness

‘lty’ (line type) sets various dot patterns

Histograms



Histograms

Summary statistics can be misleading when taken out of context.

- *For these data*, neither the mean nor median describe what we'd *intuitively* think of as "the center of the data"
- We discovered this discrepancy by examining the summary statistics in the context of a histogram

Main takeaway: Always visualize data – do not just stop with summary statistics.

Measures of spread

More summary statistics: **measures of spread** show how “spread out” the data are.

- Range
- Standard deviation/variance
- Percentiles/Interquartile Range

Range

Definition: the minimum and maximum values taken by the data.

You can find the range manually, using the ‘min’ and ‘max’ commands in R.

```
> min(t1972)
[1] 30.8
```

```
> max(t1972)
[1] 68.8
```

Alternatively, use the ‘range’ command, which does both.

```
> range(t1972)
[1] 30.8 68.8
```

An aside on vectors

- Note that ‘min’ and ‘max’ return single values (scalars) while ‘range’ returns 2 values (a vector).
- In R, any collection of values of length greater than 1 is referred to as a **vector**[2].
- ‘t1972’ is a vector. So is the output from calling ‘range.’

```
> myRange = range(t1972)
> is.vector(myRange)

[1] TRUE
```

[2] This is true for many programming languages, not to mention in physics and mathematics in general.

An aside on vectors

Remember how we used matrix notation to retrieve rows, columns and specific elements from the turnout data? We do the same for vectors.

1st element of 'myRange'

```
> myRange[1]
[1] 30.8
```

20th element of 't1972'

```
> t1972[20]
[1] 61.1
```

An aside on vectors

A few other thoughts before returning to spread.

- No commas for vector elements since they are one-dimensional.
- ‘c’ (the concatenate command) creates vectors
- data.frames (like ‘turnout’) are collections of vectors

Standard Deviation & Variance

Standard deviation and variance are measures of the distance between individual observations and their mean.

Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Variance: $\sigma^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n}$

S.D.: $\sigma = \sqrt{\sigma^2}$

Standard Deviation & Variance

Look at the summands (i.e. individual terms) in the numerator of variance:

$$(x_i - \bar{x})^2$$

You can *think* about S.D. as the (weighted) average distance of an observation from the mean of the data.

Standard Deviation & Variance

Why is the summand squared?

- penalizes observations for being further away from the center.
- consequence: variance is NOT on the same scale as the original variable.
- S.D. $= \sqrt{Var}$ so IS on the same scale as the original variable.

Standard Deviation & Variance

Finding SD and Variance in R is simple. Let's do so for the 1972 turnout data.

```
> var(t1972)
[1] 83.23074
> sd(t1972)
[1] 9.123088
```

Visualizing SD and Var

Let's visualize the range and S.D. using histograms.

```
> myMean = mean(t1972)
> hist(t1972)
> abline(v = myMean, col='black', lwd=3, lty=1)
```

Visualizing SD and Var

We'll first define a few values we are interested in. Let's say, 1 and 2 SDs to the left and the right of the mean[3].

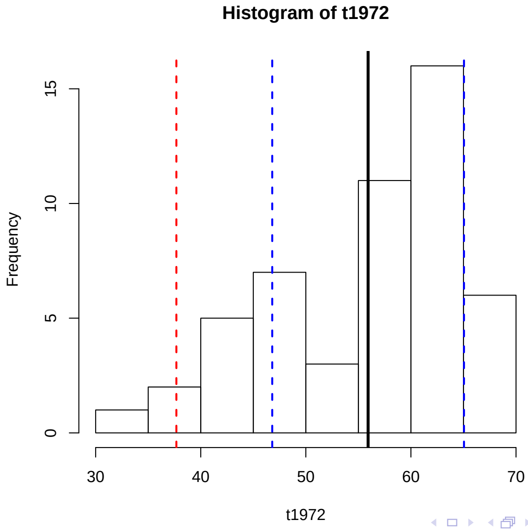
```
> mySd = sd(t1972)
> vals = myMean + c(-2, -1, 1, 2) * mySd
```

[3]Another note on syntax: when you do any basic operation (e.g. add, substract, multiply) between a vector (the term created by the 'c' command) and a scalar (e.g. 'mySd'), you get a new vector where the operation is performed between each element of the original vector and the scalar.

Visualizing SD and Var

Finally, add these new lines to the original histogram to demarcate points 1 and 2 S.D. to the left and right of the mean.

```
> abline(v = vals[1], col='red', lwd=2, lty=2)
> abline(v = vals[2], col='blue', lwd=2, lty=2)
> abline(v = vals[3], col='blue', lwd=2, lty=2)
> abline(v = vals[4], col='red', lwd=2, lty=2)
```


Histogram of t1972
Frequency
0
5
10
15
30
40
50
60
70
t1972

Visualizing SD and Var

What do we notice?

- Most data fall within 2 SDs of the mean (i.e. between the 2 red lines)
- The mean (solid black line) is not a good measure of what we'd intuitively think of as the center(s) of this data
- The data tapers off more steeply on the right hand side.
- The last 2 points have to do with the bimodality of the data. We'll explore this issue later.

Percentiles/Interquartile Range

Definition: the **Xth percentile** is the value under which X percent of observations lie.

- To find percentiles, sort the data and just count until you get to the desired percentile.
- e.g. consider some data ranging from 0 to 10: 0, 1,. . . 10.
  - The 20^{th} percentile is 2, the 75^{th} is 7.5, etc.

Percentiles/Interquartile Range

Certain percentiles have special names.

25 : first quartile
50 : median
75 : 3rd quartile

20 : first quintile
40 : second quintile
60 : third quintile
80 : fourth quintile

Percentiles/Interquartile Range

The interquartile range (IQR) is the distance between the 3rd and 1st quartiles.

$$IQR = Q_3 - Q_1$$

It can be a better measure of spread than range and SD when there are a lot of extreme values (outliers).

Percentiles/Interquartile Range

In R, use the quantile command to retrieve specific percentiles.

1st quantile for the 1972 turnout data.

```
> quantile(t1972, 0.25)

  25%
48.35
```

You can also pass a vector of percentiles that you want to find. E.g. find *all* quartiles.

```
> quantile(t1972, c(0.25, 0.5, 0.75))

  25%   50%   75%
48.35 59.00 62.35
```

Boxplots

To easily visualize quartiles/IQR, use the **boxplot**

- also called a box and whisker plot
- easily see the 3 quartiles and the range of *most* of the data
- has slightly different implementations depending on who's defining it
- hard to describe in words... so lets just see what it looks like

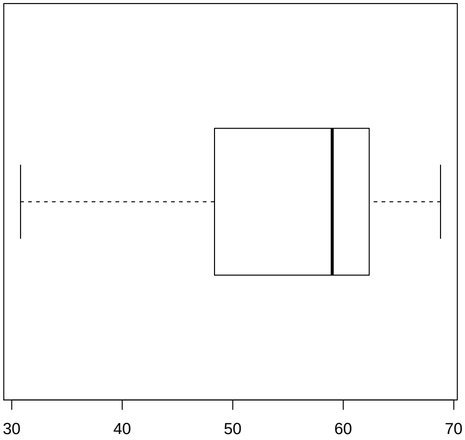
Boxplots

Boxplot syntax is very similar to histogram syntax.

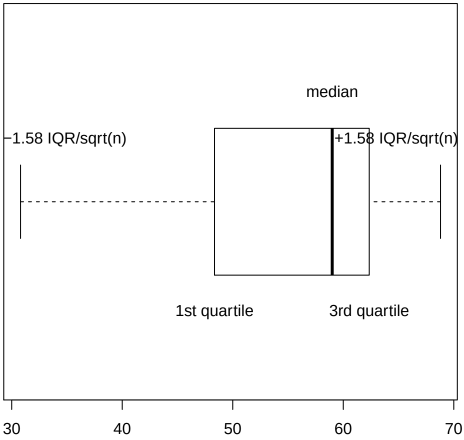
```
> boxplot(t1972, horizontal=T)
```

The 'horizontal = T' argument makes the plot horizontal, rather than the default vertical.

Boxplots



Boxplots

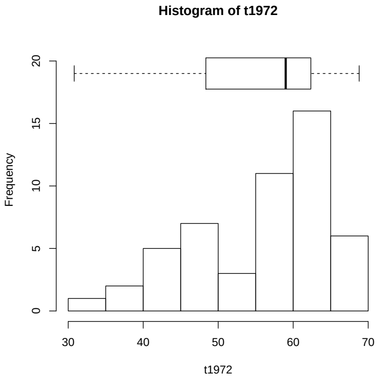

median
-1.58 IQR/sqrt(n)
+1.58 IQR/sqrt(n)
1st quartile
3rd quartile
30
40
50
60
70

Boxplots

Lines coming out the sides of the box are called whiskers

- In R, the default length of the whiskers is $\pm 1.58\times$ IQR$/\sqrt{n}$, where n is the total number of observations.
- Other sources suggest min/max, $1.5\times$IQR, $2\times$SD, etc. The specific implementation is not that important.
- What is important is the idea that *most* observations fall within the whiskers. Anything outside can be considered an **outlier** (i.e. unsually extreme value).

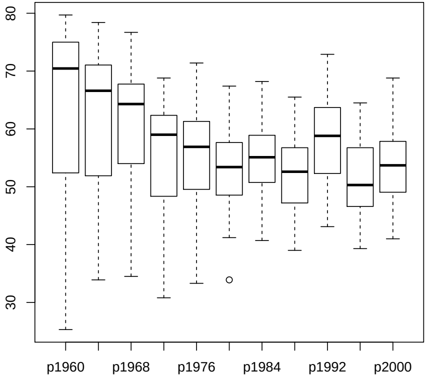
Boxplots



Comparing the histogram and the boxplot. . .

- 1 advantage: boxplot lets us easily see quartiles
- are there any other advantages?

Boxplots



This is called a **side-by-side** boxplot.

Useful for presenting...

- data from different **clusters**
- changes over time (AKA **longitudinal data**)

Boxplots

How was the last plot created?

```
> boxplot(turnout[ , 2:12])
```

- 'X:Y' creates a vector of integers from X to Y.
- 'turnout[, 2:12]' tells R, "create a new data.frame using only columns 2 to 12 of the 'turnout' data.frame."
- Passing a data.frame to 'boxplot' creates a boxplot for each column. Here, each column is one election in 1960-2000.

Boxplots

What should we notice?

- Median turnout across the states has steadily decreased since 1960, settling to $\approx 55\%$, with a spike in 1992.
- There is an outlier in 1980, indicated by the open circle
- The spread of the data has decreased somewhat since 1960.
- We do not see bimodality in the data as with the histograms – a tradeoff between the two types of plots.

Subsetting data

Motivation: So far we've been observing the bimodality of the 1972 turnout data but haven't dug into what's causing it.

Further, because our data are bimodal, our measures of centrality don't quite describe what we intuitively think of as 'central.'

Let's mitigate the bimodality issue by **subsetting**.

Bimodality

Bimodality/multiple peaks are often caused by some underlying factor that separates subgroups of the data.

Examples:

- sex
- race & ethnicity
- government type
- location

For our data (state-by-state turnout) one such factor is region – specifically, whether a state is located in the South.

```
> str(turnout)

'data.frame': 51 obs. of  13 variables:
 $ state     : chr  "AL" "AK" "AZ" "AR" ...
 $ p1960     : num  30.8 43.7 52.4 40.9 65.8 69.2 76.1 72.3 NaN 48.6 ...
 $ p1964     : num  35.9 44 54.8 50.6 65.4 68 70.7 68.9 38.7 51.2 ...
 $ p1968     : num  52.7 50 49.9 54.2 61.6 64.8 68.8 68.3 34.5 53.1 ...
 $ p1972     : num  43.4 48.3 48.1 48.1 59.9 60.1 66.3 62.3 30.8 49.3 ..
 $ p1976     : num  47.2 48.3 48.6 52.2 51.3 60.4 62.4 58.4 33.3 51.5 ..
 $ p1980     : num  47.5 41.2 47.3 52.6 48.5 56.1 59.4 56 33.9 50.6 ...
 $ p1984     : num  49.9 59.4 46.1 51.8 49.6 55.1 61 55.6 43.1 48.3 ...
 $ p1988     : num  46 55.7 46.1 47.3 47.1 56.2 58.3 50.2 40.9 44.7 ...
 $ p1992     : num  54.8 63.8 52.9 53.6 49.4 60.8 64.4 55.6 48.7 51 ...
 $ p1996     : num  47.7 56.9 45.4 47.5 43.3 53.1 56.4 49.6 42.7 48 ...
 $ p2000     : num  50 64.4 42.1 47.8 44 56.8 58.3 56.3 49 50.6 ...
 $ deepsouth: num  1 0 0 1 0 0 0 0 0 1 ...
```

We want the last column – ‘deepsouth’.

Bimodality

- ‘deepsouth,’only takes values of 0 and 1. This type of variable is called an **indicator** because it indicates whether a condition is true (1) or not (0).
- We want to separate out the turnout data by whether the corresponding state is in the South or not; i.e. whether deepsouth $= 1$ or 0.
- There are many ways to do this. For now, we'll use the ‘subset’ command.

Subsetting

First, lets retrieve the 'deepsouth' column from 'turnout'.

```
> deepsouth = turnout$deepsouth
```

Next, take the subset of the data that ARE in the South, and take another subset that are NOT in the South

```
> t1972.South = subset(t1972, deepsouth == 1)
> t1972.Other = subset(t1972, deepsouth == 0)
```

Subsetting

subset(*arg1*, *arg2*)

- arg 1: what data we are subsetting[4]
- arg 2: **condition** to subset on.
  - 'X == Y' is like saying "return TRUE when X is equal to Y; else, return FALSE"
  - Since 'deepsouth' is a vector, 'deepsouth == 1' gets us a vector of TRUEs and FALSEs
  - Enter 'deepsouth == 1' into the console to see how this works.

[4] Here we are subsetting a vector, but the function will take data.frames as well. Try subsetting on 'turnout' to see what happens

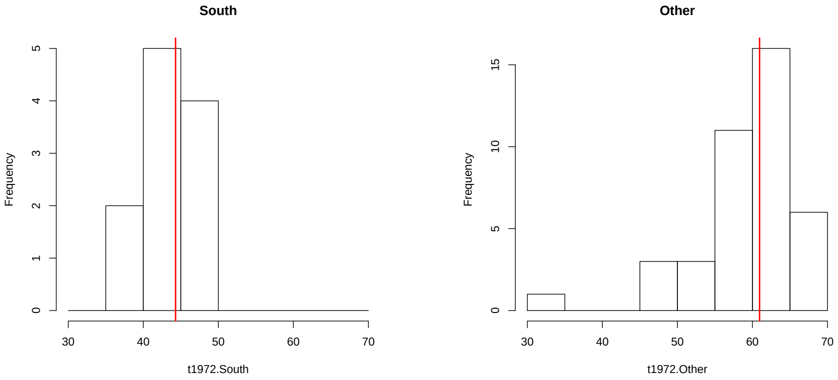
Subsetting

From subsetting, we now have 2 vectors of turnout data: one for the South, and another for the remaining states

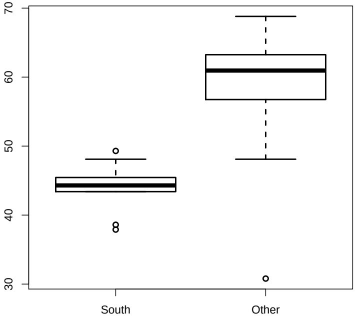
```
> t1972.South

 [1] 43.4 48.1 49.3 37.9 44.3 45.0 43.4 38.6 43.6 45.4 45.5

> t1972.Other

 [1] 48.3 48.1 59.9 60.1 66.3 62.3 30.8 50.4 63.2 62.7 60.8
[12] 63.3 59.0 48.4 61.1 50.3 62.0 59.5 68.4 57.5 67.7 56.0
[23] 50.9 64.2 60.0 57.6 56.6 67.9 57.5 56.9 61.7 56.1 62.0
[34] 68.8 68.5 61.1 63.8 62.4 62.0 63.6
```

Now let's compare them graphically.

Subsetting


South
Frequency
0
1
2
3
4
5
30
40
50
60
70
t1972.South
Other
Frequency
0
5
10
15
30
40
50
60
70
t1972.Other

Subsetting



It's pretty clear that there is something that sets southern states apart from the rest of the Union in the 1972 elections.