BACS - HW (Week 2)

Submitting Your Work

IMPORTANT: Please do NOT put your name on your assignment. Only put your Student ID at the top.

Format your solution as a PDF file. Clearly indicate which question and part you are answering; Show your code and comment it appropriately; Show visualizations when appropriate; Mark your answers clearly.

Question 1) Let's have a look at how the mean and median behave.

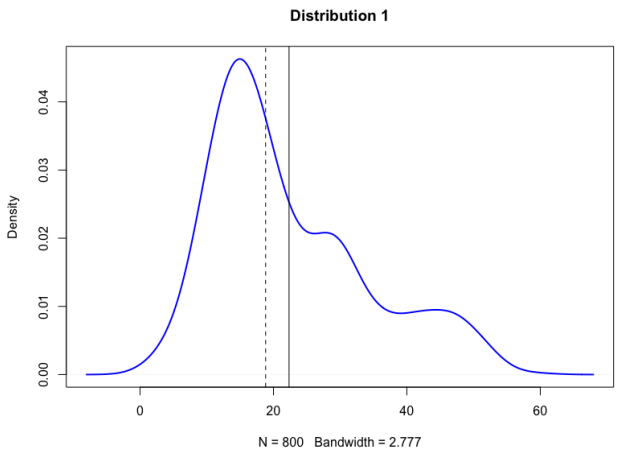
In the box below, we have created a composite distribution by combining three normal distributions, and drawn a density plot. The mean (thick line) and median (dashed line) are drawn as well. Two important things to observe: first, the distribution is positively skewed (tail stretches to the right); second, the mean and median are different!

```
# Three normally distributed data sets
d1 <- rnorm(n=500, mean=15, sd=5)
d2 <- rnorm(n=200, mean=30, sd=5)
d3 <- rnorm(n=100, mean=45, sd=5)

# Combining them into a composite dataset
d123 <- c(d1, d2, d3)

# Let's plot the density function of d123
plot(density(d123), col="blue", lwd=2,
     main = "Distribution 1")

# Add vertical lines showing mean and median
abline(v=mean(d123))
abline(v=median(d123), lty="dashed")
```



Similarly, let's create the following distributions. You can reuse and modify the code above.

(a) Create and visualize a new "**Distribution 2**": a combined dataset (`n=800`) that is *negatively skewed* (tail stretches to the left). Change the mean and standard deviation of d1, d2, and d3 to achieve this new distribution. Compute the mean and median, and draw lines showing the mean (thick line) and median (thin line).

(b) Create a "**Distribution 3**": a single dataset that is *normally distributed* (bell-shaped, symmetric) -- you do not need to combine datasets, just use the `rnorm()` function to create a single large dataset (`n=800`). Show your code, compute the mean and median, and draw lines showing the mean (thick line) and median (thin line).

(c) In general, which measure of central tendency (mean or median) do you think will be *more sensitive* (will change more) to outliers being added to your data?

Question 2) Let's try to get some more insight about what standard deviations are.

a) Create a random dataset (call it 'rdata') that is *normally distributed* with: n=2000, mean=0, sd=1. Draw a density plot and put a solid vertical line on the mean, and *dashed* vertical lines at the 1st, 2nd, and 3rd standard deviations to the *left and right* of the mean. You should have a total of 7 vertical lines (one solid, six dashed).

b) Using the `quantile()` function, which data points correspond to the 1st, 2nd, and 3rd *quartiles* (i.e., 25th, 50th, 75th percentiles)? How many *standard deviations away from the mean* (divide by standard-deviation; keep positive or negative sign) are those points corresponding to the 1st, 2nd, and 3rd quartiles?

c) Now create a new random dataset that is *normally distributed* with: n=2000, mean=35, sd=3.5.
In this distribution, how many *standard deviations away from the mean* (use positive or negative) are those points corresponding to the 1st and 3rd quartiles? Compare your answer to (b)

d) Finally, recall the dataset d123 shown in the description of question 1. In that distribution, *how many standard deviations* away from the mean (use positive or negative) are those data points corresponding to the 1st and 3rd quartiles? Compare your answer to (b)

Question 3) We mentioned in class that there might be some objective ways of determining the bin size of histograms. Take a quick look at the Wikipedia article on Histograms ("Number of bins and width") to see the different ways to calculate bin width (h) and number of bins (k).

Note that, for any dataset *d*, we can calculate number of bins (k) from the bin width (h):

```
k = ceiling((max(d) - min(d))/h)
```

and bin width from number of bins:

```
h = (max(d) - min(d)) / k
```

Now, read this discussion on the Q&A forum called "Cross Validated" about choosing the number of bins

a) From the question on the forum, which formula does *Rob Hyndman's* answer (1st answer) suggest to use for bin widths/number? Also, what does the Wikipedia article say is the benefit of that formula?

b) Given a random normal distribution:

```
rand_data <- rnorm(800, mean=20, sd = 5)
```

Compute the bin widths (*h*) and number of bins (*k*) according to each of the following formula:

i. Sturges' formula

ii. Scott's normal reference rule (uses standard deviation)

iii. Freedman-Diaconis' choice (uses IQR)

c) Repeat part (b) but extend the `rand_data` dataset with some outliers (create a new dataset `out_data`):

```
out_data <- c(rand_data, runif(10, min=40, max=60))
```

From your answers above, in which of the three methods does the bin width (h) change *the least* when outliers are added (i.e., which is least sensitive to outliers), and (briefly) WHY do you think that is?