Homework #1 - SOLUTIONS

Due: Tuesday, September 28 @ 5pm [27*points*]

Problem 1: [5*points*]

Consider the following figure from Kim, Chung, & Shin. "Higher levels of serum triglyceride and dietary carbohydrate intake are associated with smaller LDL particle size in healthy Korean women." *Nutr Res Pract.* 6(2), 2012. (http://www.ncbi.nlm.nih.gov/pubmed/22586500)

- Is this figure a histogram? Justify your answer. [2*points*]

No. Histograms are used to show distributions of data grouped into bins or intervals. This graph is more comprable to a bar graph - showing each individual as a bar.

- Describe in words what this figure shows and give your interpretation of it. [1*point*]

This figure shows the LDL phenotype A and B for different subjects.

- How would you improve this figure? Sketch an improved version. [2*points*]

This figure could be simplified to a bar chart for 1) LDL phenotype B (mean 248 with error bars to incorporate all 13 samples) and LDL phenotype A (mean 269 with error bars to incorporate data from all 44 samples).

Alternatively, a boxplot or dotplot (or histogram for each group) could be used to better represent this data. A barplot is misleading when there is only one sample per bar.

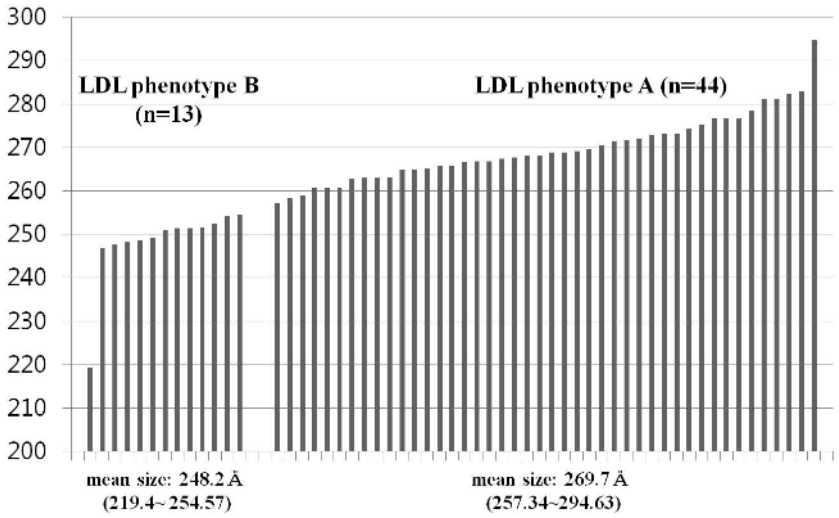


Fig. 1. Distribution of low-density lipoprotein (LDL) particle size in all study subjects (LDL phenotypes A and B). *LDL phenotype A group* (mean size: 269.7 Å, n = 44), subjects with buoyant-mode profiles [peak LDL particle diameter ≥ 264 Å] including intermediate LDL subclass pattern [256 Å ≤ peak LDL particle diameter ≤ 263 Å]; *LDL phenotype B group* (mean size: 248.2 Å, n = 13), subjects with dense-mode profiles [peak LDL particle diameter ≤ 255 Å]

Problem 2: [12*points*]

Consider the following set of measurements of some variable `x`

52	16	180	1	199	8	3	23	156	63
808	25	5	554	85	1	64	52	7	192

Using a handheld calculator, compute: [1*pointeach*]

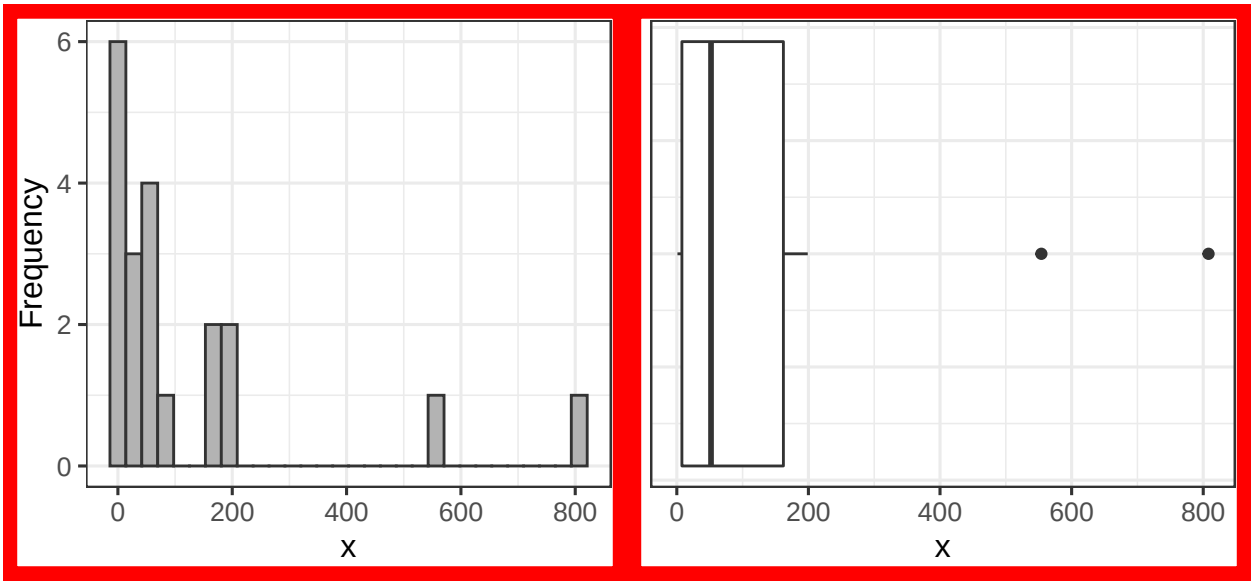
- The mean of `x` = 125
- The median of `x` = 52
- The sample standard deviation of `x` = 206

Note: significant digits were used in this answer because it doesn't make sense to have a mean that is more precise than your data, but points are not taken off for disregarding significant digits.

Sketch: [2*pointseach*]

- A rough histogram of `x`
- A rough boxplot of `x`

Take a point off for axes unlabeled or inappropriately labeled



- Describe the shape of the distribution in words. [2*points*]

The distribution of the data is heavily skewed to the right

Suppose we added two additional observations to `x`, both of which were exactly equal to the mean of `x`.

TA's: please grade this based on the solution given above, even if the original mean is wrong. i.e., answers that are consistent are correct; inconsistence is incorrect.

[1*pointeach*]

- What would the new mean be? = 125
- What would the new median be? = 58
- What would the new sample SD be? = 196

Problem 3: [3*points*]

Suppose `x` is a *sample of body temperatures in Fahrenheit* from patients admitted to the ER in the past month. Let `y` be that same set of measurements, but converted to Celsius.

- Write down the equation for obtaining `y` as a function of `x`. [1*point*]

'y = (5/9)(x - 32)'

Note that this is a **linear transformation**, i.e. 'y' has the form 'mx + b'

- If $\bar{x}$ is the mean of the Fahrenheit measurements, what would the mean of the Celsius measurements $\bar{y}$ be, in terms of $\bar{x}$? [1*point*]

$\bar{y} = (5/9)(\bar{x} - 32)$

The mean of the Celsius-converted values 'y' is the same as taking the mean of the Fahrenheit values 'x' and convert it after the fact. In general, the mean of linearly transformed data is the same as the linear transformation of the mean. Note, however, that this only holds for linear transformations; other types of transformations do not have this property.

- If s_x is the sample SD of the Fahrenheit measurements, what would the sample SD of the Celsius measurements s_y be, in terms of s_x? [1*point*]

$s_y = (5/9)(s_x)$

Note that the standard deviation of the linearly transformed data is NOT the same as a linear transformation of the SD! SD changes accordingly with multiplying by a constant, but does NOT change with adding or subtracting a constant.

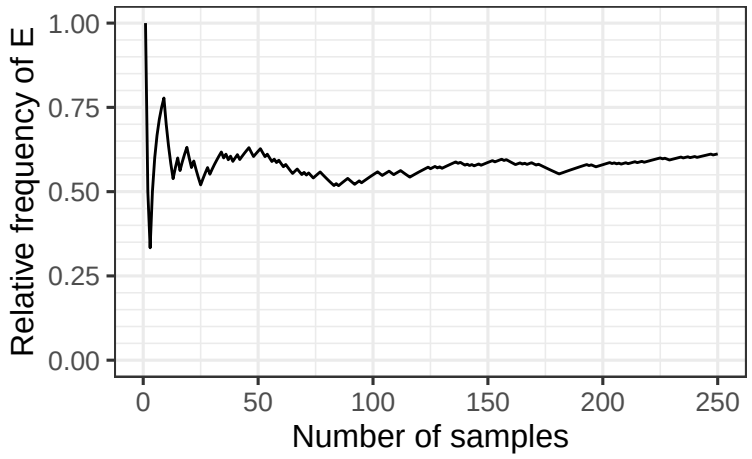
Problem 4: [3*points*]

Consider the following graph showing the relative frequency of an event plotted against the number of independent samples.

- What is the probability of event E? (i.e., `Pr{E}`)? [2*points*] How do you know this is true? [1*point*]

PrE seems to be about 0.60 (answers between 0.65 and 0.55 are correct).

The probability of an event is approximately equal to its relative frequency over large sample size N



Problem 5: [4*points*]

Consider the following histogram of 50 samples:

- Sketch the corresponding **relative frequency histogram** [2*points*]

Sketch should have the same distribution as the original histogram, but the scale of the y axis should be from 0 to 0.4 (divide the height of each bin by the total number of samples = 50)

- Identify the probability that a chosen value is **less than 4** (i.e., `Pr{ Value < 4 }`) [2*points*]

Prvalue < 4 = area of bars in **relative frequency histogram** with values less than 4 (not equal to 4!!) = 0.66. I.e. $(6/50) + (19/50) + (8/50) = 33/50 = 0.66$

This is one difference between a histogram and a bar chart. In a bar chart, you have specific x axis values (i.e. 2, 3, 4, 5) and a count of observations that hit each of these values. However, in a histogram, we are plotting continuous data that is grouped into "bins" of equal size. Your histogram might look different depending on the number/size of bins. In this case, there is a bin that covers values between 3.75-4.25 (approx). I would consider anything in that bar that covers the value 4 as equal to or greater than 4. And this is the power of using density curves instead of relative frequency histograms, because you can use calculus to find the area under the curve at a particular point (i.e. 4 or even 3.9).