IMT573 Problem Set 4: Distributions, Central Limit Theorem

Your name:

Deadline: Thu, Nov 5th 8pm

Instructions

1. Please write clearly. Imagine this is a business report for your boss. She does not care about coding but is very much interested in the results. Can she understand your text? Test it, by making all code invisible and ensuring one can still understand what you have written.

2. Consider outputting your results as inline R chunks in markdown as in this example: `we have `r nrow(data)` observations`.

3. Do not add irrelevant output! Every output you produce must be there for a *good reason*!

1 Compare differently distributed data (40pt)

In this problem set you are comparing the humans in terms of body size (height) and influence. Strictly speaking we do not have data on human influence here, just research paper influence (citations) but it is a good proxy for influence of the individual humans (researchers).

Let's start with the human height data.

1.1 Human Bodies (18pt)

1. (5pt) You'll work about human heights. What kind of measure is this? (nominal, ordered, difference, ratio)? How should it be measured (continuous, discrete, positive...)?

   Hint: read lecture notes https://otoomet.bitbucket.io/machineLearning.pdf/ section 1.1.

2. (3pt) Read the “fatherson.csv” data. It contains two columns, father’s height and son’s height (in cm). Let’s focus on fathers here (variable *fheight*) and ignore the sons. Provide the basic descriptives: how many observations do we have? Do we have any missings? Any unreasonable values?

   Note: These are 19th century people and they are noticeably shorter than us.

3. (5pt) Compute mean, median, standard deviation and range of the heights. Discuss the relationship between these numbers. Is mean larger than median? By how much (in relative terms)? How does standard deviation compare to mean?

4. (5pt) plot a histogram of the data. Add to this histogram mean and median. You can use vertical lines of different color.

   What do you find? Which distribution does the result resemble?

   Hint: To draw a vertical line, check out `geom_vline` geom if you are using ggplot, or `abline` if you do base R plotting.

1.2 Human influence (22pt)

Next, let’s take a look at human influence. We are working with the number of citations of research papers. The data represents research papers from Microsoft Academic Graph (MAG). It contains two columns: paper id and number of times this paper has been cited. We do not need the id and just work with the number of citations.

1. (5pt) What kind of measure is this? What kind of valid figures would you expect to see (continuous, discrete, positive, ...)

2. (3pt) Read the “mag-in-citations.csv” data. Provide the basic descriptives: how many observations do we have? Do we have any missings? Do we have implausible or wrong values? What is the range of the citations?

3. (5pt) Compute mean, median, mode (the most frequent value), standard deviation and range of the number of citations. Discuss the relationship between these numbers. Is mean larger than median? Than mode? By how much (in relative terms)? How does standard deviation compare to mean?

Hint: you can use *modeest* package to estimate the mode as

```
mode <- modeest::mlv(citations, method="mfv").
```

But you can also just make a table of values, you will immediately see which one is the most common.

4. (5pt) plot a histogram of the data. Add to this histogram mean, median, and mode. You can use vertical lines of different color.

   How does the histogram look like? Which distribution does it resemble? Can you get it to be a nice and easy to grasp image?

   Note: you may experiment with log-log scale for the histogram.

5. (4pt) Finally, comment on your findings about human bodies and influence.

2 Explore Central Limit Theorem (40pt)

In this section you will see how does Central Limit Theorem (CLT) work. CLT states that means of random numbers tend to be normally distributed if the sample gets large, and the variance of the mean tends to be $\frac{1}{S}\operatorname{Var} X$ where S is the sample size and X is the random variable, means of which we are analyzing.

CLT, and how variance and mean value change when sample size increases, plays a very important role in computing confidence intervals, something we need later in this course.

You will work with discrete random numbers and observe how means of Bernoullis turns more-and-more into a normal.

Lets create a RV

$$X = \begin{cases} -1 & \text{with probability } 0.5 \\ 1 & \text{with probability } 0.5. \end{cases}$$

One way to sample such values is

```
sample(c(-1, 1), 10, replace=TRUE)  # creates 10 random values

##  [1] -1 -1  1 -1  1  1  1 -1  1 -1
```

1. (4pt) (10pt) Calculate the expected value and variance of this random variable.

   Hint: read Openintro Statistics 3.4 (Random variables), in particular 3.4.2 (Variability). I recommend to use the shortcut formula $\mathrm{Var}\, \mathsf{X} = \mathbb{E}\, \mathsf{X}^2 - (\mathbb{E}\, \mathsf{X})^2$.

2. (1pt) Choose your number of experiments size N. 1000 is a good number.

3. (4pt) Create a vector of N random numbers as explained above. Make a histogram of those. Comment the shape of the histogram.

   Hint: while ggplot is great for plotting data, here you find it easier to work with base-R graphics. You can plot histogram of `x` with

   ```
   hist(x, breaks=30)  # about 30 bins
   ```

4. (4pt) Compute and report mean and variance of the random numbers you created (just use `mean` and `var` functions). Compare these numbers with the theoretical values computed in question 1.

5. (4pt) Now create N *pairs* of random numbers. For each pair, compute its mean. You should have N means. Make histogram of the means. How does this look like?

   Hint: you can do this using loops while creating two random numbers in each iteration. If this seems to simple to you, create a $\mathsf{N} \times 2$ matrix of random numbers, where each row represents one pair. Thereafter your compute means by rows using `rowMeans`.

6. (5pt) Compute and report mean of the pair means, and variance of the means. Compare these numbers with the theoretical values computed in question 1. Remember, as CLT tells, the variance now should be just 1/2 of what (3) suggests as size of the pairs $\mathsf{S} = 2$.

   Hint: Expectation should be 0, variance should be 1/2.

7. (2pt) Now instead of pairs of random variables, repeat this with 5-tuples of random numbers (i.e. 5 random numbers per one observations instead of a pair). Do you spot any noticeable differences in the histogram?

   Hint: expected value should still be 0, variance should be 0.2

8. (2pt) Repeat with 25-tuples...
9. (2pt) ... and with 1000-tuples.
10. (4pt) Comment on the tuple size, and the shape of the histogram.
11. (8pt) Explain why do the distribution becomes to look more and more normal as we take mean of a large sample of individual values.

    In particular, explain what happens when we move from single values $S = 1$ to pairs $S = 2$. Why did two equal peaks turn into a “山”-like histogram?

How much time did you spend on this PS?

3 Extra credit: Pareto Distribution (1 EC point)

Here you basically replicate the previous exercise using Pareto random numbers.[1] Pareto is a popular distribution to describe unequal outcomes, such as human income. It has a single parameter α, often called *shape*. Its pdf is given as

$$f(x) = \alpha x^{-\alpha-1}, \tag{1}$$

its expected value (mean) is

$$\mathbb{E}\, X = \frac{\alpha}{\alpha - 1}, \quad \alpha > 1 \tag{2}$$

and its variance is

$$\operatorname{Var} X = \frac{\alpha}{(\alpha-1)^2(\alpha-2)}, \quad \alpha > 2. \tag{3}$$

Pareto is a peculiar distribution in a sense that it does not have variance if $\alpha \leqslant 2$, and it does not have expected value if $\alpha \leqslant 1$. We initially analyze a nice case with $\alpha = 10$, and later you go for $\alpha = 1.5$ (expectation exists but variance does not) and $\alpha = 0.5$ (neither exists).

Now let’s generate random numbers from this distribution.

1. (0pt) Choose your sample size N. 1000 is a good number.

[1] More precisely, we talk here about Pareto-II or Lomax distribution. This is a shifted version of Pareto-I distribution (see wikipedia for details).

2. (10pt = 0.1EC points) Create a vector of N *pareto(10)* random numbers. Make a histogram of those. Comment the shape of the histogram.

   Hint: you can use `VGAM::rpareto(5, shape=10)` to create 5 such numbers.

3. (10pt) Compute and report mean and variance of the sample you created. Compare these numbers with the theoretical values computed from (2) and (3).

4. (10pt) Now create N *pairs* of random Paretos. For each pair, compute its mean. You should have N means. Make the histogram. How does this look like?

   Hint: while you can do this using loops, it is more useful to create a $\mathsf{N} \times 2$ matrix of random normals, where each row represents one pair. Thereafter your compute means by rows and you have N means.

5. (10pt) Compute and report mean of the pair means, and variance of the means. Compare these numbers with the theoretical values computed from (2) and (3). However, as CLT tells, the variance now should be just 1/2 of what (3) suggests as size of the pairs $\mathsf{S} = 2$.

6. (5pt) Now instead of pairs of random normals, repeat this with 5-tuples of random numbers (i.e. 5 random numbers per one observations). Do you spot any noticeable differences in the histogram?

7. (5pt) Repeat with 25-tuples...

8. (5pt) ... and with 1000-tuples.

9. (10pt) Comment on the tuple size, and the shape of the histogram.

10. (10pt) Now repeat the previous with $\alpha = 1.5$. For simplicity, let's just do the case with 1, 25, and 1000 samples.

    Note: you cannot compute theoretical variance here, so there is nothing to compare to.

11. (10pt) And finally, repeat it with $\alpha = 0.5$. Again, only consider $\mathsf{S} = 1, 25, 1000$.

    Note: you can compute neither the expected value nor the variance, so nothing to compare here.

12. (15pt) Comment the result. What happens when variance does not exist? What happens when expectation does not exist?

Hint: consult Openintro Statistics 5.1.3 (p 172-178).