Analiza rozkladu empirycznego

Lucas van der Velde

Task 1: Game of Thrones

(a) Compute the means and the coefficients of variation for season 1 and 6. What can you say about viewership changes in those seasons?

We solve the excercise using the formulas learnt during class. We use the classic measure of mean ($\sum x_i/n$) as there is no need to weight the observations. Then we compute the variance. In order to make it faster, I use the formula = $Var(x) = E(x^2) - E(X)^2 = \frac{\sum x^2}{n} - (\frac{\sum x}{n})^2$. The original data and the solutions are below.

	Season 1		Season 6	
	x	x^2	x	x^2
1	2.22	4.93	7.94	63.04
2	2.2	4.84	7.29	53.14
3	2.44	5.95	7.28	53.00
4	2.45	6.00	7.82	61.15
5	2.58	6.66	7.89	62.25
6	2.44	5.95	6.71	45.02
7	2.4	5.76	7.8	60.84
8	2.72	7.40	7.6	57.76
9	2.66	7.08	7.66	58.68
10	3.04	9.24	8.89	79.03
Sum	25.15	63.81	76.88	593.92
Average	2.52	6.38	7.69	59.39
Variance	0.06		0.29	
sd	0.24		0.54	
Coeff Var	0.1		0.07	

We observe that the number of viewers almost tripled between season 1 and season 6. When we analyzed variation in raw numbers, it would seem like it was larger in the case of season 6; though once we take into account the higher value of the mean in that season the inequality is reversed.

(b) Draw a histogram using 5 equally spaced bins.

Here we need to create first a table with grouped data and then plot the values of x (viewers) against the relative frequencies. (only the table is shown below)

Viewers	Frequency	Relative frequency
2-4	19	0.279
4-6	12	0.176
6-8	27	0.397
8-10	5	0.074
10-13	5	0.074

(c) Compare the mean in the grouped version and in the individual version. Are they the same? Why?

- Mean in individual data: 5.95
- Mean in grouped data (table from previous point): 6.013

Some differences between the two are expected, as in the second we have less information available. In particular, in the last bin, the mid point is 11.5, even though only one observation in the bin is above 11. We can expect that increasing the number of bins should bring the grouped mean result closer to the measure based on individual data.

Task 2: Chinese babies

(a) Compute the mean weight at birth and the standard deviation for boys and girls.

As pointed out during classes, we need to have information on the minimum and maximum weight of babies in order to compute the mean in grouped data. Since this information was not available, we will compute the mean for the 94% central observations, that is ignoring the top and bottom 3%. This procedure is not uncommon, particularly when the presence of outliers can skew the results. (During class, we assumed the minimum weight to be zero and the maximum to be 4.5 and 5 for girls and boys, respectively).

Table 1: Weights for boys

Lower	upper	Dystrybuanta	Frequency	Freq. normalized	Midpoint	Product
2704	2783	0.05	0.02	0.021	2743.5	57.614
2783	2908	0.1	0.05	0.053	2845.5	150.812
2908	3127	0.25	0.15	0.16	3017.5	482.8
3127	3382	0.5	0.25	0.266	3254.5	865.697
3382	3652	0.75	0.25	0.266	3517	935.522
3652	3907	0.9	0.15	0.16	3779.5	604.72
3907	4065	0.95	0.05	0.053	3986	211.258
4065	4171	0.97	0.02	0.021	4118	86.478

Table 2: Weights for girls

Lower	upper	Dystrybuanta	Frequency	Freq. normalized	Midpoint	Product
2627	2703	0.05	0.02	0.021	2665	55.965
2703	2824	0.1	0.05	0.053	2763.5	146.466
2824	3035	0.25	0.15	0.16	2929.5	468.72
3035	3281	0.5	0.25	0.266	3158	840.028
3281	3541	0.75	0.25	0.266	3411	907.326
3541	3787	0.9	0.15	0.16	3664	586.24
3787	3940	0.95	0.05	0.053	3863.5	204.766
3940	4042	0.97	0.02	0.021	3991	83.811

Frequency normalized is the frequency divided by .94, that is the percentage of observations in our trimmed sample. We could interpret it as a conditional mean, where the condition is that the baby's weight belongs to the middle 94% of the observations. The computed means are the sum of values from the product column. (All calculations were performed in excel rounding numbers at three digits.

- Mean(boys) = 3394.901
- Mean(girls) = 3293.322

(b) Knowing that the number of boys is 205506 and the number of girls is 183982, please obtain the overall mean for all babies.

The mean for all babies is a weighted average of the means for boys and girls, where weights represent the relative number of observations. Hence, $w_{boys} = \frac{n_{boys}}{n_{boys}+n_{girls}}$. So we have that

$$\begin{aligned}
\bar{x}_{all} &= \frac{n_{boys} * \bar{x}_{boys} + n_{girls} * \bar{x}_{girls}}{n_{boys} + n_{girls}} \\
&= \frac{205506 * 3394.901 + 183982 * 3293.322}{205506 + 183982} \\
&= 3346.918
\end{aligned}$$

Notice that correcting for the fact that we use just the 94% central observations should not affect the results, as this would be equivalent to replacing n_i with $.94 * n_i$, i (boys, girls). The additional .94 cancels out in the equation.