Biostatistics VHM 8010 Course
Fall 2024, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to home assignment I

The solution presents one way in which the descriptive analysis could be done, but other ways are possible as well. It is a bit more detailed than required for a 100% mark, by including both gender groups for the calculations in the last question and throughout by covering multiple ways the questions could be answered.

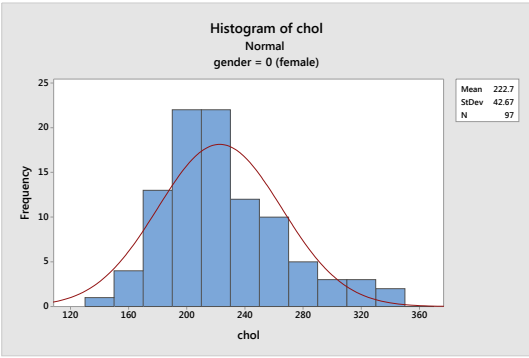
1. Descriptive analysis

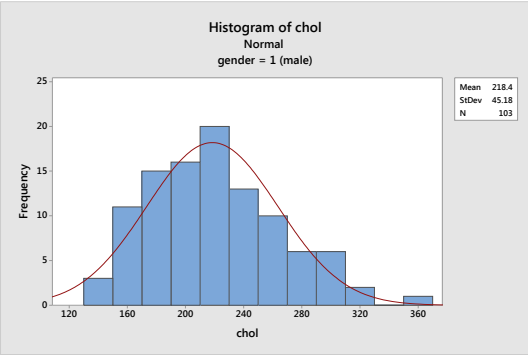
The data may be considered as a simple random sample from the population of study participants, representing the population of Framingham at that time. The focus is on the (total) cholesterol level, a quantitative variable measured in mg/dL. (Note that this study was carried out long before the more recent distinction between "good" and "bad" cholesterol.) We carry out separate analyses for the women and men. The table below gives the most useful descriptive statistics for the cholesterol variable; the list of descriptive statistics should as a minimum include the mean, median and standard deviation.

Descriptive statistics for cholesterol values in women and men:

gender	no. obs.	mean	5-number summary	std.dev.	skewness
women	93	222.7	134 – 191.5 – 216 – 246 – 340	42.67	0.764
men	107	218.5	133 – 182 – 212 – 249 – 357	45.18	0.504

The descriptive statistics were computed in Minitab; other statistical software (e.g. Stata) could give slightly different values for the percentiles (in the 5-number summary). With the moderately large sample size, the preferred graphical display of the continuous distributions is a histogram, which may be overlaid a normal distribution curve to show the agreement with the normal distribution. The stemplot gives about the same information but in a more clumsy layout, and a boxplot displays just the descriptive statistics of the 5-number summary and suspected outliers. The histograms have the same x-axis and binning (11-12 bins) for the two gender groups, in fair agreement with the $\sqrt{n}$ recommendation (slide 1L−11).





We summarise the interpretations of the descriptive statistics and displays in the table below. Note that the Graphical Summary in Minitab includes both an extensive list of descriptive statistics and a histogram (as well as a box plot); it is therefore a natural choice for a combined graphical and descriptive display.

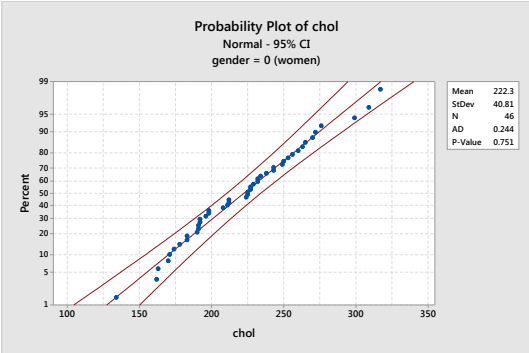
characteristic	women	men
center	around $216 - 223$ (median and mean)	around $212 - 218$ (median and mean)
	thus very slightly larger for the women than men	
spread	$s = 43$ and IQR$= 54.5$	$s = 45$ and IQR$= 67$
	thus std.dev. around 45 in both groups, but somewhat larger IQR for men	
symmetry	for both genders, somewhat right-skewed (both box and tails)	
outliers	two suspected outliers $(334, 340)$	one suspected outlier (357)

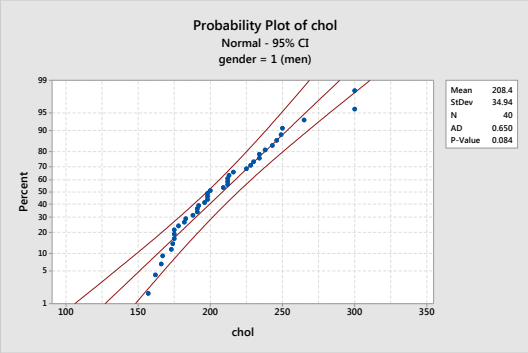
The two suspected outliers among the women are close to other values in the distribution and hence do not appear to be real outliers. The one suspected outlier among the men is substantially (40 units) off the other values and may be a real outlier. This may however also be due to unusually low upper-tail values among the other men, so it is difficult to say this for sure. One would probably need to (i) check this particular observation, and (ii) compare with the normal range in a larger sample of men.

The overall impression of the two distributions is that they are pretty similar, both somewhat (or moderately) right-skewed and with similar center and spread. It is perhaps interesting and a bit surprising to note that a normality test gives clear evidence against a normal distribution only for the distribution for the women. It is probably due to its more pronounced skewness, but that does not change our general assessment and comparison of the two distributions.

2. Assessment of normal distribution

After restricting the data to persons in the age group 35-44 years (both inclusive), we are left with cholesterol values for 46 women and 40 men. The standard tools to assess whether it is reasonable to assume a variable to be normally distributed are the normal probability plot and a normality test. The figures below show this plot and gives a P-value for the Anderson-Darling normality test for the two gender groups.





The plot for the women looks reasonably straight and has all points inside the confidence bands; in addition, with $P = 0.75$ the A-D normality test shows no evidence in the data (whatsoever) against a normal distribution. The plot for the men is more curved with a couple of points outside the bands (one at either end), and the P-value of the A-D normality is 0.084. This is quite close to the usual significance level of 0.05, and in fact other normality tests could give significance for these data (e.g., the Ryan-Joiner test in Minitab's Normality Test menu gives 0.043, and the Shapiro-Wilk test in Stata gives 0.046). Inspection of the distribution using descriptive statistics shows that the distribution is quite right-skewed (e.g., with a skewness of 0.865), and it therefore seems most reasonable to describe this distribution as (somewhat) non-normal, due to in particular its right-skewness.

3. Comparison with population data

The proportion of high (≥ 240) cholesterol values can be estimated with or without assuming a normal distribution. Without assuming a normal distribution, we can simply compute the proportion of high cholesterol values in the sample. The calculation based on the normal distribution is the same as we used in Supplementary Exercise 1.127, except that we replace the true population mean and standard deviation by the estimates from our sample. We should however avoid (or be very cautious) to use such a calculation if the data are not described well by a normal distribution.

Women aged 35-44 years
Our sample mean is somewhat larger than the population mean ($222 > 214$). We concluded above that the sample is described well by a normal distribution, so we may calculate (for $X \sim \mathrm{N}(222.33, 40.81)$):

$$\mathrm{P}(X \geq 240) = \mathrm{P}\left(\frac{X - 222.33}{40.81} \geq \frac{240 - 222.33}{40.81}\right) = \mathrm{P}(Z \geq 0.4330) = 0.3325 \approx 0.333,$$

or using Table A of IPS: $\mathrm{P}(Z \geq 0.43) = 1 - 0.6664 = 0.3336 \approx 0.334$. The z-score could also be obtained directly: $z = (240 - 222.33)/40.81 = 0.4330$. Perhaps not too surprisingly, considering the sample distribution's higher mean, the estimated proportion of high cholesterol values is quite a bit higher than in the population. The conclusion remains the same if we use the sample proportion of high cholesterol values: it equals $15/46 = 0.326$. Apparently, the cholesterol levels of the women in this age group of the Framingham study were very high compared to population standards. This could e.g. be due to differences in the demographics of the populations.

Men aged 35-44 years
The sample mean is clearly lower than the population mean (208 and 227, respectively). We concluded above that the sample was not described well by a normal distribution, so it would seem most natural to use the sample proportion of high cholesterol values: $7/40 = 0.175 \approx 0.175$. This value is clearly off the 0.339 of the population, not too surprisingly considering the differences in means. (For reference, the calculation in the normal distribution would give: $\mathrm{P}(X \geq 240) = \mathrm{P}(Z \geq (240 - 208.38)/34.94) = \mathrm{P}(Z \geq 0.905) = 0.183$.) So also for the men there is a clear discrepancy between the sample from the Framingham study and the population values for men aged 35-44 years. The direction is however the opposite of that for the women.

4. Calculation of standard deviation

It would be possible to determine a standard deviation (σ) to make $P(X \geq 240) = 0.231$ for $X \sim \mathrm{N}(214, \sigma)$, for the women, by successive trial-and-error. For any tentative value of σ, the proportion above 240 would be computed as above, and the value of σ could changed gradually to get closer and closer to the target value (0.231). We can also obtain a formula for σ by noting that the right-tail proportion 0.231 corresponds to a z-score of $z = 0.7356$ (or $z = 0.74$ using Table A), and therefore

$$0.7356 = z = \frac{240 - 214}{\sigma} \quad \Rightarrow \quad \sigma = \frac{240 - 214}{0.7356} = 35.35.$$

That is, for the women aged 35-44 a normal distribution $\mathrm{N}(214, 35.35)$ would correspond to a proportion of high cholesterol values of 0.231. The same approach for the men gives $z = 0.4152$ and $\sigma = (240 - 227)/0.4152 = 31.31$. That is, for the men aged 35-44 a normal distribution $\mathrm{N}(227, 31.31)$ would correspond to a proportion of high cholesterol values of 0.339.

Estimating the standard deviation based on the sample mean and the proportion of values above a certain threshold most importantly has the limitation of being based on a normal distribution assumption. If the data do not correspond well to a normal distribution, a standard deviation computed

this way could be seriously off. The calculation also relies on an accurate estimate of the proportion of values above a threshold, and this will be easier to obtain for a threshold that is relatively central in the distribution (as is the case here in both gender groups). The normality assumption might also be more critical if the proportion was for values in a far tail. Finally, even for a "perfect" normal distribution one might expect the direct calculation of the standard deviation to be more precise, because it uses more information in the data than simply whether values are above the threshold or not. One way to explore this hypothesis would be by a simulation study, but the details are well beyond the home assignment.