Biostats VHM 801 Course
Fall 2016, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to home assignment I

This solution presents one way in which the descriptive analysis could be done, but other ways are possible as well. The solution is more detailed than required for a 100% mark, by including all the variables when only 4 selected variables were required for the assignment, and by discussing multiple outliers and errors/inconsistencies in the data.

1. Study type and sampling

The study is *observational*, as opposed to an experiment, because no treatments or other interventions were undertaken. The study may be considered a *survey* although the selection of subjects was *not random*, nor was it really a convenience sample, because the sample comprised all women giving birth during a certain period. The difficulty lies in determining what population the subjects may be considered representative for. Three obvious factors to consider are (i) the location of the hospital and the demographics in the population in that area, (ii) the health care services the women had access to during pregnancy, and (iii) the restriction to a certain period of the year. It is probably fair to say that the sample could be considered representative for pregnancies during summer months in parts of the country (as well as similar other countries) with a comparable population and health care system.

2. Descriptive analysis

The variables *age*, *height*, *weight0*, *weight1*, and *weightb* are quantitative and continuous, *sev* and *prevsev* are categorical with four categories, and the variables *r1–r5* and *a1–a5* are all dichotomous (categorical with two categories). The variables *month* and *nkids* are quantitative and discrete, and can be dealt with as either quantitative or categorical; quantitative descriptive statistics may be computed but their distribution should be displayed as a discrete distribution.

The table below gives the most useful descriptive statistics for continuous variables; the list of descriptive statistics should as a minimum include the mean, median and standard deviation. There are no missing values so the number of observations is 174 for all variables. For categorical variables it is more useful to give the probabilities (that is, the proportion of women) for each possible value, as shown in the table below.

Descriptive statistics for continuous variables:

statistic	*age*	*height*	*weight0*	*weight1*	*weightb*	*month*	*nkids*
mean	26.0	1.61	59.5	70.6	3.21	3.81	0.77
minimum	15	1.47	38.2	47.3	1.08	0	0
1st quartile	22	1.55	52.3	62.7	2.84	0	0
median	25	1.60	58.4	70.0	3.24	4	0
3rd quartile	29	1.65	65.9	77.3	3.62	7	1
maximum	42	1.75	95.5	100.0	6.28	9	7
standard deviation	5.5	0.07	10.1	11.2	0.64	3.15	1.24
inter-quartile range	7	0.10	13.6	14.6	0.78	7	1
skewness	0.76	0.14	0.72	0.28	0.61	-0.08	2.37
kurtosis	0.44	-0.63	0.74	-0.31	4.74	-1.60	6.59

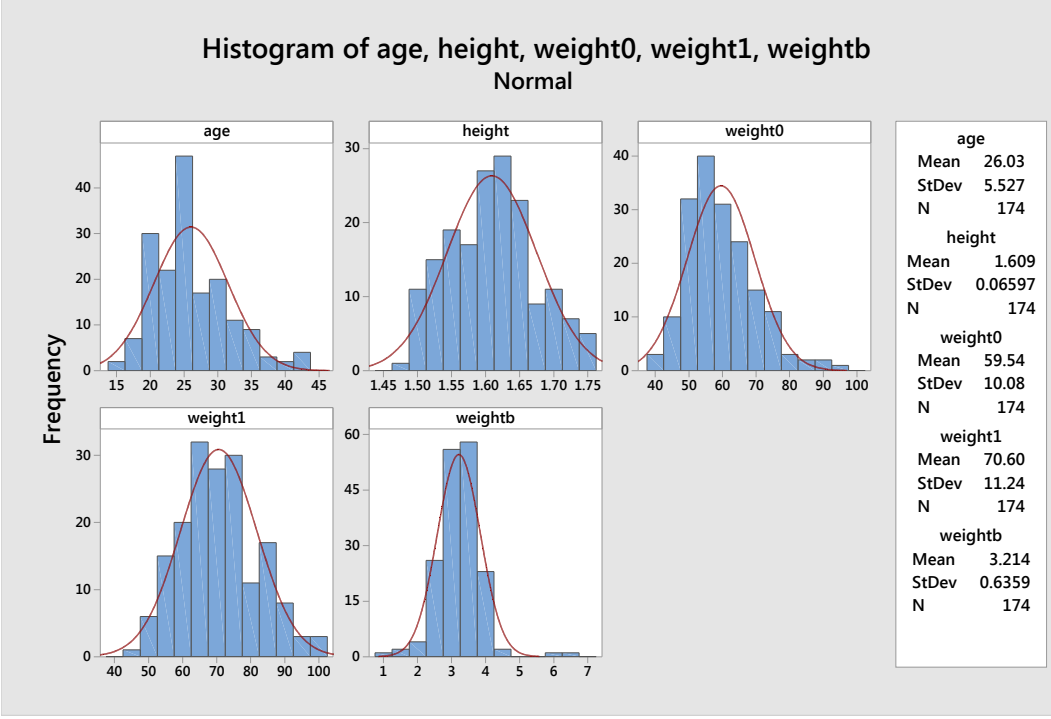
Observed probability distribution for categorical (excl. binary) and discrete variables:

value	*sev*	*prevsev*	*month*	*nkids*
0	.075	.144	.339	.563
1	.443	.466	.011	.282
2	.339	.172	.034	.069
3	.144	.161	.057	.040
4		.057	.063	.017
5			.080	.017
6			.138	.006
7			.144	.006
8			.121	
9			.011	

Observed probability distribution for binary variables:

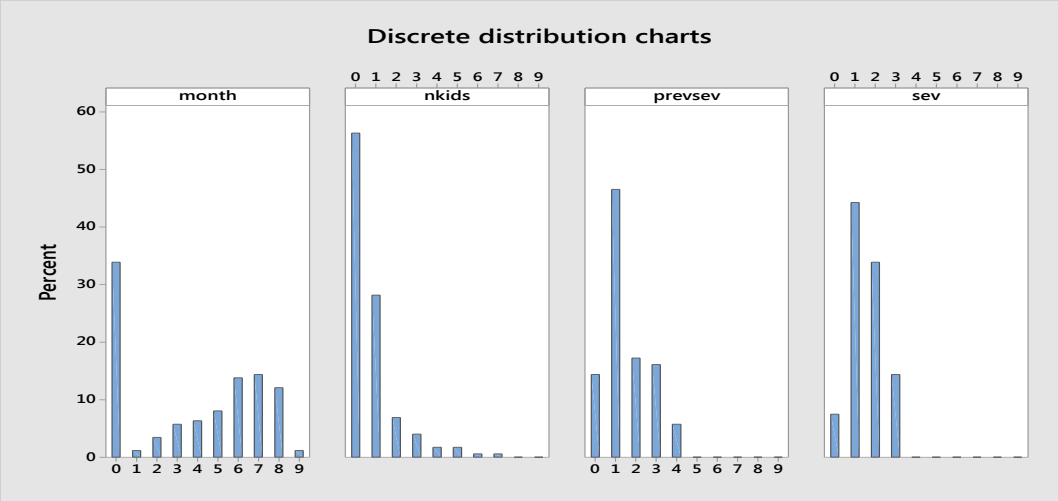
value	*r1*	*r2*	*r3*	*r4*	*r5*	*a1*	*a2*	*a3*	*a4*	*a5*
0	.879	.960	.885	.787	.971	.782	.730	.833	.920	.856
1	.121	.040	.115	.213	.029	.218	.270	.167	.080	.144

With 174 observations, 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 (where of interest). The stemplot gives about the same information but in a more clumsy layout, and a boxplot displays just the descriptive statistics involved and potential outliers.



The default number of bins was quite variable, with most bins (22 and 19) for *weight1* and *age*, respectively. All histograms have been adjusted to have 13 ($\approx \sqrt{174}$) bins. The preferred graphical

display for a categorical variable is a chart with one bar for each value representing its observed proportion; alternatively, a pie chart could be used as well. Bar charts for the categorical and discrete variables are shown below; note that the bars are separated — which distinguishes bar graphs from histograms. (The panel display in Minitab unfortunately assumes the range of values to be the same for all variables.) The binary variables could also be shown with bar or pie charts, but the proportions of the table above will suffice here.



Finally, brief summaries of the distributions based on the computed statistics and graphs:

- *age*: unimodal; centered around 25 years; somewhat right-skewed,
- *height*: unimodal; centered around 1.60 m; narrow distribution with $s = 0.07$ and quite short tails; fairly symmetrical,
- *weight0*: unimodal; centered around 59 kg; fairly large spread ($s = 10$) and a long right tail; right-skewed,
- *weight1*: unimodal; centered around 70 kg; similar spread as *weight0* but considerably more symmetrical,
- *weightb*: unimodal; centered around 3.2 kg; apart from multiple extreme observations in both tails apparently quite symmetrical, but very peaked (due to the extreme observations),
- *month*: bimodal (highest mode at 0 and lower mode at 6-7 months); mode at zero seems to contrast rest of distribution which is left-skewed (see Question 4 for further discussion),
- *nkids*: unimodal with highest proportion for 0 kids, and values ≥ 2 comprise less than 20% of distribution; strongly right-skewed,
- *sev*: highest proportion for category 1 (minor) followed by 2 (troublesome),
- *prevsev*: highest proportion for 1 (n/a) with roughly same proportion for all the remaining categories; surprisingly, about 14% of the values are zero (value not indicated as possible),
- *r1–r5*: highest proportion (21%) of pain relief for $r4$, almost none (3-4%) for *r2* and *r5*,
- *a1–a5*: highest proportions (22% and 27%) of pain aggravation for *a1* and *a2*, respectively.

3. Outlying observations

The boxplots of the continuous variables (not shown) indicate the following potentially outlying observations (marked by asterisks):

- four women of age 42 years (all other women are below 40 years),
- a women of (initial) weight 95.5 kg, below that two women of 89 kg, and all other women below 84 kg,
- one women with final weight 100.0 kg, but several values close to that (e.g. 97.7 kg and 98.6 kg),
- three babies of weight less than 1.67 kg (down to 1.08 kg), and two very large baby weights (5.97 and 6.28 kg), much larger than the third-largest weight (4.49 kg).

Among these observations, only the large baby weights seem plausible as real outliers. It *is* biologically possible to give birth to a baby of this size, but it seems unlikely that a dataset of 174 observations would contain two so extreme values by chance. A possible explanation is that these were twin births.

The high values for age are hardly outlying in any real sense; the data just happened to have three women above 40 years of the same age. Given the right-skewness of the age distribution, it is not unusual to have single large values. The distribution of *weight0* is also right-skewed, and the highest value (for woman 47) may be perfectly normal. It is somewhat suspect though that *weight1* for this woman is identical to *weight0* (the same situation also occurs for woman 88); one should check whether a transcription error has been made.

Among the categorical and discrete variables, no single observation stands out as a potential outlier.

4. Errors and/or inconsistencies

One obvious error has already been noted: 25 women have the code zero for *prevsev*, despite that this code is not among the valid response categories. One could imagine these values to be mistaken for "n/a", but further inspection shows that 6 out of the 24 women had $nkids \geq 1$; it is unclear why these women should be unable to give a score for *prevsev*. Clearly, this variable should be checked before any conclusions are drawn about its values.

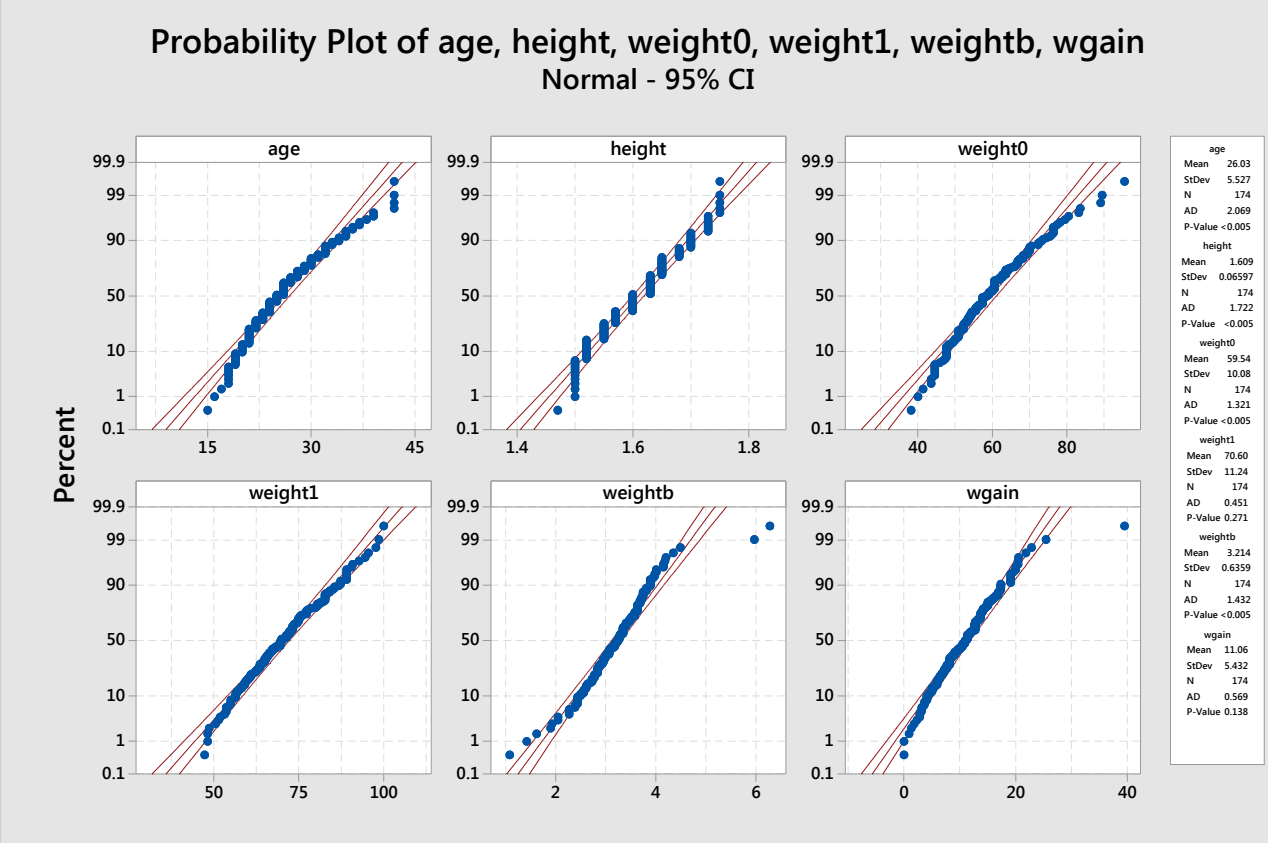
Some further potential errors or inconsistencies:

- *weight0* and *weight1* being identical for the woman (no. 50) with the by far largest value of the former, as discussed above,
- a total of 13 women reported a severity score of 0 (none); two of them reported the month as 1 while the remaining reported the month as 0; the correct coding is probably that women without any backpain should not answer to the question about when the pain started,
- four out of the 81 women who give *prevsev* as "n/a" had $nkids = 1$; it is unclear why these women should not be able to give a score for *prevsev*; perhaps no distinction was made between born and adopted children when enumerating the number of kids,
- several women reported both relief and aggravation of pain by the same activity (questions 2-5); this may be an undesired ambiguity in the responses,
- the response categories for the questions about relief and aggravation of pain do not account for some of the women not experiencing any pain. There should probably have been an "n/a" category, or a rule for how the question should be answered for women not experiencing any pain.

5. Normal distribution for continuous variables

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 figure below (next page) shows these plots

and gives a P-value for the Anderson-Darling test for the 5 continuous variables plus the weight gain (*wgain*) computed as *weight1*$-$*weight0*.



Only two of the six variables have non-significant P-values: *weight1* and *wgain*. For *weight1*, the points are roughly on a straight line with only a few observations beyond the bounds, and it would seem reasonable to assume a normal distribution for this variable. For *wgain*, the points also seem roughly on a straight line, except for one point way out to the right. This woman (no. 163) had the highest weight at end of pregnancy, while her weight at beginning of the pregnancy was just about average. While this not may be impossible, it does look strange and an error could have occurred. Despite the non-significance of the normality test, it hardly makes sense to assume a normal distribution for *wgain* with this extreme observation included. Without this observation, the normal plot obviously looks much better, and the A-D normality test becomes clearly non-significant ($P = 0.52$).

All other variables show $P < 0.005$ and thus strong evidence against a normal distribution. The right-skewness of *age* and *weight0* was already noted. We also noted the extreme values of *weightb* which make this distribution highly peaked; the table of descriptive statistics gave a kurtosis of 4.74. Finally, the points for *height* are roughly on a straight line but show an apparent granularity (multiple points vertically above each other, indicating that the same values are repeated multiple times in the data). Inspection of the data shows that only 12 different values exist for the heights. Apparently the height was recorded in inches (58–69 inches) and then converted to the metric scale. Another problem with the height distribution is that the tails are somewhat too thin for a normal distribution. Altogether it might be acceptable to approximate the height distribution by a normal distribution; in fact, heights are commonly considered as normally distributed.

Note. The data originate from a study carried out at the London Hospital in 1977; further description of the study and its results can be found in Mantle MJ, Greenwood RM & Currey HLF (1977), Backache in pregnancy, *Rheumatology and Rehabilitation* **16**, 95−101.