Supplementary exercises 7.143 and 7.145 of IPS7e

Data: OC (osteocalcin) measurements (in *mg/ml* blood) for $n = 31$ healthy women between 11 and 32 years. We let $X_1, \ldots, X_{31}$ denote the OC measurements.

Exercise 7.73
We first analyze the data on original scale.

Model: the 31 observations are a simple random sample (i.i.d. sample) from a distribution with mean μ and standard devation σ, both of which are unknown parameters. We could assume a normal distribution $\mathrm{N}(\mu, \sigma)$ right away, but it seems more logical to explore the data first.

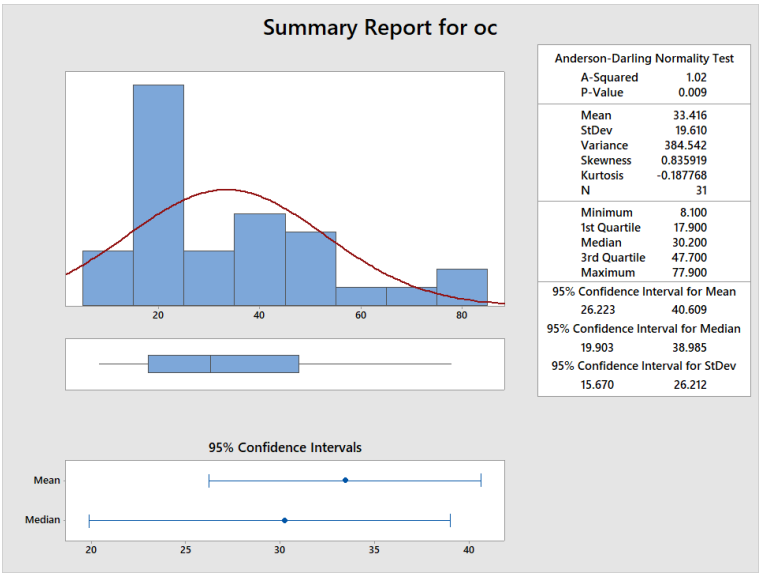
(a) Some descriptive statistics from Minitab (including the stemplot):

Stem-and-Leaf Display: oc

Stem-and-leaf of oc N = 31

2	0	89
3	1	0
10	1	5677799
15	2	00004
15	2	
(3)	3	011
13	3	568
10	4	04
8	4	7
7	5	244
4	5	6
3	6	
3	6	8
2	7	
2	7	67

Leaf Unit = 1



Comments:
The stemplot, histogram and boxplot all show that the data are right-skewed and not approximated well by a normal distribution. The test (Anderson-Darling) for normality is clearly significant. The skewness is 0.84. The distribution appears to be unimodal. There are no obvious outliers, and the $1.5 \times$ IQR rule does not flag any observations as suspected outliers.

(b) In order to use t-procedures to obtain a 95% confidence interval for μ, we need to assume a normal distribution $\mathrm{N}(\mu, \sigma)$ for the OC measurements. We discuss the validity of our analysis below the results.

One-Sample T: oc

Descriptive Statistics

N	Mean	StDev	SE Mean	95% CI for μ
31	33.42	19.61	3.52	(26.22, 40.61)

μ: mean of oc

The confidence interval should be considered *approximate* only, because the observations themselves are clearly not normally distributed (but the sample mean is nevertheless approximately normally distributed, by the central limit theorem (CLT)). It is not quite clear whether this is a situation where the guidelines (7L-3) for use of the t-procedures are met: the skewness is quite

pronounced, and the number of observations is less than 40. If we want exact inference about the population mean, it may be necessary to use more complex non-parametric resampling procedures (mentioned in Session 8, but not part of the VHM 801 course).

Exercise 7.145
For exploration of log-transformation of the OC measurements, it is most natural to work from the original data and carry out the transformation within the statistical software (in Minitab using the `Calc-Calculator` menu), in order to avoid loss of decimals. That is, we calculate $Y_i = \ln(X_i)$, where ln is the natural log transformation. Results with the supplied data for Exercise 7.145 will be slightly different.

Model: the 31 observations (Y_i) are a simple random sample (i.i.d. sample) from a distribution with mean μ_Y and standard devation σ_Y, both of which are unknown parameters.

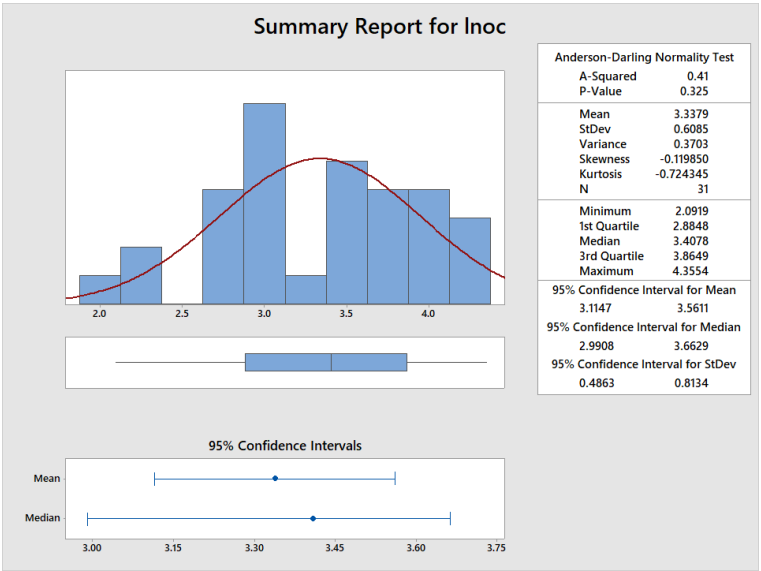
(a) Some descriptive statistics from Minitab (including the stemplot):

Stem-and-Leaf Display: lnoc

Stem-and-leaf of lnoc N = 31

1	2	0
3	2	23
3	2	
4	2	7
11	2	8888999
15	3	0001
15	3	
(5)	3	44455
11	3	667
8	3	89
6	4	000
3	4	233

Leaf Unit = 0.1



Comments:
The distribution of log-transformed OC-values is close to symmetrical (the skewness is -0.12 so the slight left-skewness is of no importance). The stemplot and histogram show a few "gaps" but the normality test does not give any evidence against a normal distribution. There does not appear to be any outlying observations when assessed at this scale either.

(b) As the 95% CI was already given in the Summary Report, we save ourselves the extra display (that was included for the same calculation in 7.143). Because the data show no evidence against being normally distributed, we may consider this interval as exact. There is certainly no problem with using the t-procedure for these data. However, the formulation of the question hints at this now being a confidence interval for the mean OC; it is not, of course: it is 95% CI for the mean logarithmic OC.

(c) The backtransformed mean is $\exp(3.3379) = 28.2$. It is lower than both the estimated mean (33.4) and the estimated median (30.2) for the original data, but it is a valid estimate of the median. If the assumed normal distribution on the log-scale is valid, this estimate is better than the median computed directly from the untransformed data.

The backtransformed endpoints of the 95% CI are: $\exp(3.1147) = 22.5$ and $\exp(3.5611) = 35.2$. This interval is considerably narrower than the 95% CI for the median displayed in the Summary

Report window (its computation will be explained in Session 8). Again, assuming the normal distribution on log-scale to be valid, this is the best interval we can get for the median.

It may seem a nuisance, or a problem, that the transformation provides inference about the median OC instead of the mean OC. Most people find means more intuitive than medians. However, we could argue that for a skewed distribution the median is a more natural estimate of the center, because it does not give equal weight to the two tails, and it is less affected by extreme values (i.e., resistant) than the mean. We didn't observe extreme values on either scale in these data, but for skewed distributions we often encounter extreme values in the longer tail. In recent years publication of results for medians are increasingly being "allowed" and even preferred over means when there is clear justification.