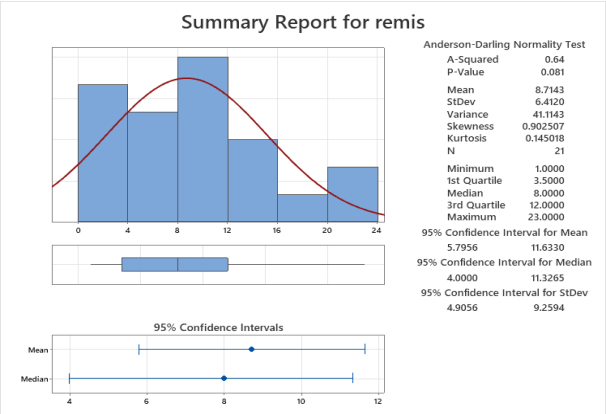
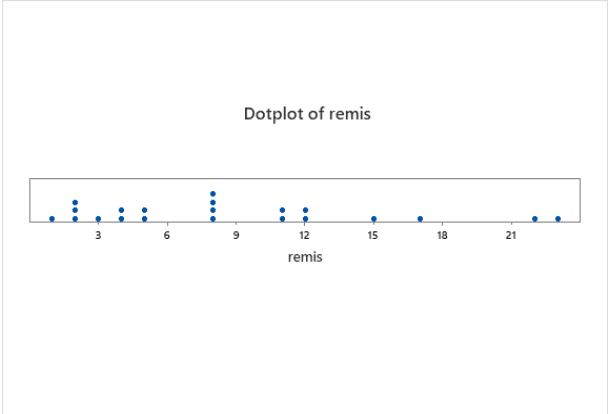
Solution to home assignment II

The solution gives roughly the amount of detail expected for a full score, except for some added explanations of procedures and interpretations, in particular the discussion of transformation and the inclusion of both answers in Question 5 when only one answer was required.

1. Descriptive analysis

For the brief descriptive analysis, we present the graphical summary and a dotplot. The distribution of remission times is unimodal and moderately right-skewed, without any suspected outliers in the boxplot. The median and mean are close, and the box is quite symmetrical around the median, but the two largest values at 22 and 23 days perhaps appear a bit off the rest. The normal probability plot (not shown) is quite curved, and the (A-D) normality test gives $P = 0.081$. It does not provide formal evidence against a normal distribution, but this is due to low power to detect deviations from a normal distribution at a fairly low sample size. The shape of the distribution and the fact that the 95% predicted range based on normality (mean ± 2 standard deviations) includes negative values indicates that the true distribution hardly is normal.





2. Statistical inference for known σ

Let $X_1, \ldots, X_{21}$ denote the 21 remission times. The natural model assumes $X_1, \ldots, X_{21}$ to form an i.i.d. sample (or a simple random sample) from an unknown distribution with unknown mean μ and known standard deviation $\sigma = 6.2$ weeks (assumed valid also for the present trial). It is quite unnecessary to assume a normal distribution for the individual remission times because the inference only relies on the distribution of $\bar{X}$ to be (approximately) normal, and this will follow from the central limit theorem.

With known σ, z-procedures are appropriate. We estimate μ by the sample mean: $\hat{\mu} = \bar{X} = 8.714$. For the 99% confidence interval we use $z^* = z_{.995} = 2.576$, and for testing the hypothesis $\mathrm{H}_0 : \mu = 11.5$ (for treatment B) against a two-sided alternative $\mathrm{H}_a : \mu \neq 11.5$ we use a z-test:

$$\begin{aligned} 99\%\ \text{CI} \quad &: \quad \bar{X} \pm z^* \times \sigma/\sqrt{n} = 8.714 \pm 3.485 = (5.23, 12.20), \\ z\text{-test} \quad &: \quad z = \frac{\bar{X} - 11.5}{\sigma/\sqrt{n}} = -2.06, \quad P = 2 \times \mathrm{P}(Z > 2.06) = 0.039. \end{aligned}$$

The sample mean remission time is 8.71 weeks, and the 99% confidence interval includes the (population) mean value for the treatment B (11.5 weeks). That however only tells us that $P > 0.01$, so we need the extra information from the z-test — it gives statistical significance at $P < 0.05$ with $P = 0.025$. Therefore, the data provide (weak) evidence that the (population) mean remission time for treatment A is not equal to (and hence less than) 11.5 weeks.

3. Statistical inference for unknown σ

If σ must be estimated from the data, we need to assume $X_1, \ldots, X_{21}$ to be i.i.d. from $\mathrm{N}(\mu, \sigma)$. As mentioned, this assumption is somewhat unrealistic, so we need to rely on the robustness of the t-distribution procedures for the results to be reliable. According to the textbook guidelines (see page 9 of Lecture 6), t-dstribution procedures can be used for a sample size between 15 and 40 except in the presence of outliers or strong skewness. According to these guidelines, we should be ok here.

With estimated σ ($\hat{\sigma} = s = 6.412$), t-procedures are appropriate. For the 99% confidence interval we use $t^* = t_{20,.995} = 2.845$, and for testing the hypothesis $\mathrm{H}_0 : \mu = 11.5$ against the two-sided alternative $\mathrm{H}_a : \mu \neq 11.5$ we use a t-test:

$$\begin{aligned} 99\%\ \text{CI} \quad &: \quad \bar{X} \pm t^* \times s/\sqrt{n} = 8.714 \pm 3.981 = (4.73, 12.70), \\ t\text{-test} \quad &: \quad t = \frac{\bar{X} - 11.5}{s/\sqrt{n}} = -1.99, \quad P = 2 \times \mathrm{P}(t_{20} > 1.99) = 0.060. \end{aligned}$$

The margin of error increased and the test statistic became non-significant (even if just above the 0.05 cut-off), because the sample standard deviation (6.41) was larger than the previously used known value (6.2), and because the t-distribution with 20 degrees of freedom is somewhat wider than the standard normal distribution. With a P-value of 0.060, the data provide no formal evidence of different mean remission times for treatments A and B. We still note that the mean remission time was lower for treatment A than for treatment B, and we should use our confidence interval (or a 95% confidence interval if that was considered more appropriate) to indicate the range we are confident the population mean lies in. The CI is mostly below the value for treatment B, but not entirely.

Additional note:
If we were concerned about using t-distribution procedures because of the apparent non-normality of the distribution of remission times, we could consider transforming the values to another scale. Among the two suggested transformations, square-root transformation works better to achieve a more symmetrical distribution; the log-transformed values end up showing some left-skewness. We can use the same procedures as above on square-root scale to get a 99% CI of $(2.074, 3.431)$ and $t = -3.58$ with $P = 0.002$ for testing $\mathrm{H}_0 : \mu = \sqrt{13} = 3.606$ against a two-sided alternative. The value 13 weeks for the median of treatment B was actually a typo in the question, the intended value was 11.5 weeks, and in that case would get $t = -2.68$ and $P = 0.014$. We can and should backtransform the CI to original scale, by squaring both endpoints, giving a 95% CI for the median as $(4.30, 11.77)$. This interval is considerably narrower than the one displayed in the Graphical Summary based on the sign test, reflecting the impact of using stronger assumptions about the data.

4. Treatment C group

Here our comparison is with another sample instead of with known values. The second sample is of similar size ($n = 25$) as our first sample ($n = 21$), therefore we should not ignore the uncertainty on the estimates for treatment C (as we effectively do when considering population values as known). The appropriate model and procedure is now for two independent samples. Without the actual data values for the second sample, it is difficult to know what distributional assumptions would be

reasonable, but the most obvious choice is a normal distribution, as for treatment A. That is, we assume the values, say $Y_1, \ldots, Y_{25}$, to be i.i.d. from $\mathrm{N}(\mu_C, \sigma_C)$. We are told that $\hat{\mu}_C = \bar{Y} = 11.5$ and $\hat{\sigma}_C = s_C = 6.412$. For the two-sample t-distribution procedures there is a choice between working under the extra assumption $\sigma_A = \sigma_C$ or without that assumption. Because the sample standard deviations are the same, it seems tempting to adopt the first assumption, but it turns out the results are almost identical. The explanation is that the standard error for the mean difference is the same for both analyses, so the only difference is the degrees of freedom (which differ slightly).

For the solution, the calculations under the equal variances assumption are shown. The pooled standard deviation estimate is naturally equal to the same value as in both samples, $s_p = 6.412$. The degrees of freedom equals $21+25-2 = 44$. For the 99% confidence interval we use $t^* = t_{44,.995} = 2.692$ (from statistical software), and for testing the hypothesis $\mathrm{H}_0 : \mu_A = \mu_C$ against a two-sided alternative $\mathrm{H}_a : \mu_A \neq \mu_C$ we use a t-test:

$$\begin{aligned} 99\%\ \text{CI} \quad &: \quad \bar{X} - \bar{Y} \pm t^* \times s_p\sqrt{(1/21) + (1/25)} = -2.786 \pm 5.109 = (-7.90, 2.32), \\ t\text{-test} \quad &: \quad t = \frac{\bar{X} - \bar{Y}}{s_p\sqrt{(1/21) + (1/25)}} = -1.47, \quad P = 2 \times \mathrm{P}(t_{44} > 1.47) = 0.15. \end{aligned}$$

The margin of error increased, causing the confidence interval to extend further across zero than the previous interval did across 11.5, and the t-test value decreased to not really being close to significant any more. All of this is as expected because the analysis now incorporates more uncertainty.

Additional note:
If we did the previous analysis on transformed scale, it is perhaps natural to also work on transformed scale here, even if the robustness of t-procedures increases with the larger combined sample size (and the two samples being similar in sample size and spread, both of which can otherwise cause trouble). The results for square-root transformed scale are as follows, again under the equal variances assumption. The pooled standard deviation is $s_p = 1.093$. The 99% CI for the mean difference becomes $(-1.509, 0.233)$, and the t-test to compare the means gives $t = -1.97$ with $P = 0.055$ in $t(44)$ against a two-sided alternative. These results also reflect the larger uncertainty. We cannot backtransform the CI to original scale, but because means on transformed scale correspond to medians on original scale we can think of the test as comparing medians on original scale.

5. Patients with unknown remission time

Both of the outlined approaches lead to problems. The best approach is to use a statistical procedure that allows one to take into account the incompleteness of some observations (beyond the course).

If patients with unknown remission time are excluded (ii), the inference remains valid for the restricted population of patients for which a remission time is available. This population is of less interest than the full population if it only includes remission times less than a certain value (e.g. 26 weeks). However, the reduced sample will be representative for this restricted population and distribution. Patients who drop out of the study early do not contribute much information about the remission time, so the loss is less serious and may not be involve any substantial bias (unless the drop-out is directly related to the remission time). Sample size and power will however be reduced.

If patients with unknown remission time are included with the remission time recorded from the last available information (i), their remission time is systematically too small. This means that the correct distribution of remission times has a longer tail to the right, and thus probably shows stronger right-skewness. Both the mean and standard deviation will be underestimated with the incompletely observed remission times, so the confidence interval will be centered wrongly and have a too small margin of error. It is less clear how the test statistic will be affected (because the bias involves both the mean and the standard deviation), but most likely it will be larger in magnitude and thus more significant.