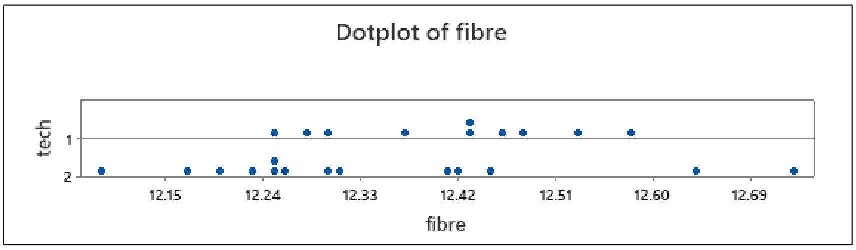
Biostats VHM 801 Course
Fall 2021, Atlantic Veterinary College, PEI
Henrik Stryhn

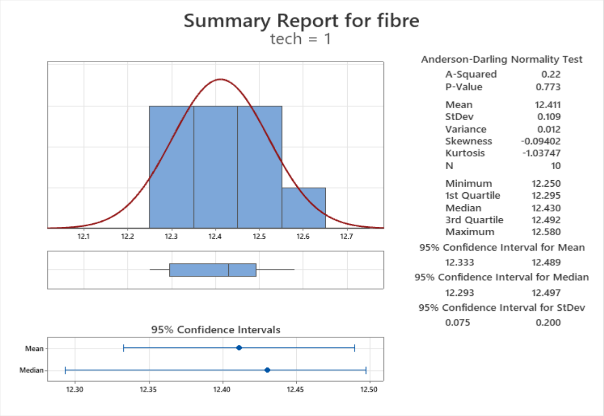
Solution to home assignment 2

The discussion of methods throughout, but in particular for question **4**, is more detailed than expected for a 100% mark. All analyses shown used Minitab, but Stata or other programs should give the same results.

As a start, we display the measurements by the two technicians of the fibre content in soy-bean cake samples, using a dotplot with measurements of the two technicians on the same scale.



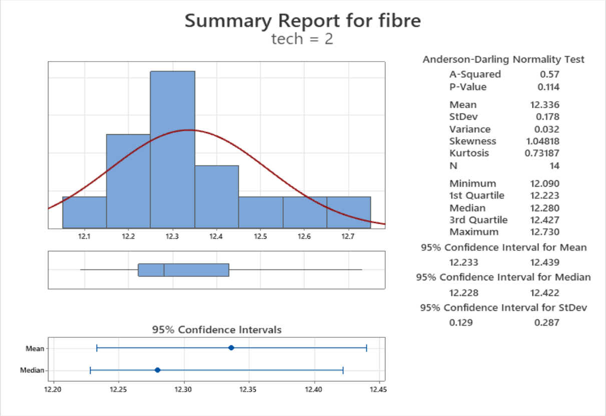
The values of technician 2 are seen to have a larger spread than those of technician 1, and possibly also a lower center of the distribution. We introduce the notation $X_1, \ldots, X_{n_1}$ ($n_1 = 10$) for technician 1, and $Y_1, \ldots, Y_{n_2}$ ($n_2 = 14$) for technician 2. The natural model (set of assumptions) for the measurements of technician 1 is that $X_1, \ldots, X_{n_1}$ are i.i.d. (independent with the same distribution) and normally distributed variables with unknown (population) mean μ_1 and unknown standard deviation σ_1. Similarly, we assume $Y_1, \ldots, Y_{n_2}$ to be i.i.d. and $\sim \mathrm{N}(\mu_2, \sigma_2)$, and finally the two samples are assumed independent of each other. The graphical summaries give estimates of means and standard deviations as well as other useful descriptive statistics.



```
Anderson-Darling Normality Test
   A-Squared          0.22
   P-Value           0.773

   Mean             12.411
   StDev             0.109
   Variance          0.012
   Skewness       -0.09402
   Kurtosis       -1.03747
   N                    10

   Minimum          12.250
   1st Quartile     12.295
   Median           12.430
   3rd Quartile     12.492
   Maximum          12.580
95% Confidence Interval for Mean
   12.333           12.489
95% Confidence Interval for Median
   12.293           12.497
95% Confidence Interval for StDev
   0.075             0.200
```



```
Anderson-Darling Normality Test
   A-Squared          0.57
   P-Value           0.114

   Mean             12.336
   StDev             0.178
   Variance          0.032
   Skewness        1.04818
   Kurtosis        0.73187
   N                    14

   Minimum          12.090
   1st Quartile     12.223
   Median           12.280
   3rd Quartile     12.427
   Maximum          12.730
95% Confidence Interval for Mean
   12.233           12.439
95% Confidence Interval for Median
   12.228           12.422
95% Confidence Interval for StDev
   0.129             0.287
```

Although the sample for technician 2 shows some right-skewness, none of the two distributions are strongly non-normal, as judged by the non-significant P-values (> 0.05) for the A-D normality tests. For later reference, we give the sample means and standard deviations:

$$\bar{X} = 12.4110,\ s_X = 0.1094, \quad \bar{Y} = 12.3364,\ s_Y = 0.1784.$$

1. Confidence intervals for each technician

Based on the above model assumptions, t-procedures lead to 99% confidence intervals for the population means (using $t^* = t_{.995}(9) = 3.250$ and $t^* = t_{.995}(13) = 3.012$ from Table D in IPS):

$$\begin{aligned} \text{99\% CI for } \mu_1: &\quad \bar{X} \pm t_{.995}(9)\, s_X/\sqrt{10} = 12.411 \pm 0.112 = (12.30, 12.52), \\ \text{99\% CI for } \mu_2: &\quad \bar{Y} \pm t_{.995}(13)\, s_Y/\sqrt{14} = 12.336 \pm 0.144 = (12.19, 12.48). \end{aligned}$$

We can interpret these intervals by saying, somewhat imprecisely but still quite acceptably, that we are 99% confident the intervals contain the population means (for the experienced and unexperienced technicians, respectively). In order to make the statements more precise, we will have to refer to repeated sampling by the technicians. For example, in the long run 99% of the confidence intervals obtained by repeated sampling for either technician will contain the respective population mean.

As for the intervals being exact or approximate, this depends on whether we think the data follow a normal distribution well (exact) or we need to rely on the distribution of the sample mean being approximately normal despite the original data are not (approximate). Our initial descriptive exploration of the data showed both samples to be reasonably normal in their distribution, maybe with a small concern for technician 2 (due to the right-skewness). Therefore we would interpret the interval for technician 1 as exact and perhaps the interval for technician as approximate, due to the (small) concern with the normal distribution. Both samples are small ($n < 15$), so if had major concerns with the normal distribution, the validity of the confidence intervals would be questioned (see "Inference for non-normal data" slide in Lecture 6).

2. Tests for each technician

The producer stated the mean fibre content to be 12.36%, and for both of the technicians we can test the hypothesis $\mathrm{H}_0 : \mu = 12.36$ against the alternative $\mathrm{H}_a : \mu \neq 12.36$ using a t-test. We take the alternative hypothesis as two-sided because there is no a priori knowledge that a deviation from the stated mean would go in a particular direction. The stated value 12.36 is included in both confidence intervals above, but this only tells us that $P > 0.01$ and we need a more precise assessment. We could redo the CIs with 95% confidence levels and look whether the value 12.36 is inside those, but we can also use the one-sample t-test.

$$\begin{aligned} \text{technician 1}: &\quad t = \frac{\bar{X} - 12.36}{s_X/\sqrt{10}} = 1.47, \qquad P = 2 \cdot \mathrm{P}(t(9) > 1.47) = 0.17, \\ \text{technician 2}: &\quad t = \frac{\bar{Y} - 12.36}{s_Y/\sqrt{14}} = -0.49, \qquad P = 2 \cdot \mathrm{P}(t(13) > 0.49) = 0.63. \end{aligned}$$

These P-values were obtained by statistical software, but the observed t-values are in any case clearly less extreme than the 97.5% percentiles in the $t(9)$- and $t(13)$-distributions (2.262 and 2.160, respectively). So neither of the t-statistics are significant at the 5% level, and with both P-values even above 0.10 it is fair to say that there is no evidence for either technician against the producer's stated value. Note that there is no contradiction between using the standard significance level of $\alpha = 0.05$ for the test even if the confidence intervals had a confidence level of 99%; tests and CIs can be computed and interpreted independently of each other.

3. Bias among technicicans

The bias (if any), as defined in the question, is the difference between the population means of the two technicians, that is $\mu_1 - \mu_2$. In the previous question, we looked at whether the means were in agreement with the producer's value, and even if they both showed no disagreement, that does not tell us they are both equal to the producer's value. Here we go one step further by comparing the two means directly (without involving the producer value); a bias between the two technicians would show up as a difference in their means. We want to test the hypothesis of no bias, $H_0 : \mu_1 - \mu_2 = 0$, or $H_0 : \mu_1 = \mu_2$. Again without any indication of the direction such a bias would have, we use a two-sided alternative, $H_a : \mu_1 \neq \mu_2$. The two samples are independent so the data invite a two-sample t-test. Two versions exist: with or without an assumption of equal standard deviations. As the data show some difference in estimated standard deviations between the technicians, it seems most natural to work without an assumption of equal standard deviations. Therefore,

$$\begin{aligned} t &= \frac{\bar{X} - \bar{Y}}{\sqrt{(s_X^2/n_1) + (s_Y^2/n_2)}} = \frac{12.4110 - 12.3364}{\sqrt{(0.1094^2/10) + (0.1784^2/14)}} = 1.27, \\ P &= 2 \cdot \mathrm{P}(t(\mathrm{df}) > 1.27) = 0.22 \quad \text{for} \quad \mathrm{df} = 21. \end{aligned}$$

The degrees of freedom were determined using statistical software. With a t-value of 1.27, the actual degrees of freedom are irrelevant for the conclusion that the P-value exceeds 0.05 and hence that the test is clearly non-significant. We conclude that there is no evidence of bias among the technicians. (When assuming equal standard deviations ($\sigma_1 = \sigma_2$), a calculation with the pooled standard deviation of $s = 0.154$ gives $t = 1.17$ with $\mathrm{df} = 10 + 14 - 2 = 22$ and $P = 0.25$, thus a similar P-value and the same conclusion.)

4. Measurements outside expected interval

The expected interval (or range) for a single measurement should, according to our statistical model, give 95% probability for a single measurement to fall inside the interval, and 5% probability of falling outside. This is *not* the same concept or requirement as for a 95% confidence interval for the population mean. A confidence interval states how precisely we have estimated the mean. Single measurements are more variable than estimates of the population mean, and their range will stay the same no matter the number of observations, whereas confidence intervals shrink when n increases. The assumed distribution for a single measurement is the normal distribution $\mathrm{N}(\mu_0, \sigma)$, with $\mu_0 = 12.36$ and $\sigma = 0.10$, and the desired interval is therefore according to the "68 - 95 - 99.7% rule",

$$\mu_0 \pm 2\sigma = 12.36 \pm 2 \cdot 0.10 = (12.16, 12.56).$$

Using the more precise value 1.96 instead of 2 in the formula gives an interval of $(12.164, 12.556)$. Such intervals are sometimes called *prediction intervals*; this term is not used in our textbooks (in this sense).

By going through the measurements of the two technicians one by one, we see that 1 out of 10 measurements are outside the interval for technician 1, and 3 out of 14 measurements for technician 2. Under the assumptions of the model, the measurements should be outside the interval independently of each other and with the same probability (0.05). Therefore we have a binomial setting, and the number of measurements outside the interval should follow a binomial distribution $\mathrm{B}(n_i, p)$, $i = 1, 2$. We want to test the null hypothesis $H_0 : p = 0.05$ against the alternative $H_a : p > 0.05$; we choose here a one-sided alternative because probabilities < 0.05 seem somewhat awkward (why would the technicians have a better agreement with the interval than expected from past laboratory experience?). With the small n's, the appropriate test procedure is an exact test in the binomial distribution. Let $N_i \sim \mathrm{B}(n_i, 0.05)$ denote the number of measurements outside the interval for $i = 1, 2$ (the two technicians), so that our observed counts are: $N_1 = 1$ out of $n_1 = 10$, and $N_2 = 3$ out of $n_2 = 14$.

Then we compute the P-values as,

$$\begin{aligned} \text{technician 1}: \quad P = & \quad \mathrm{P}(N_1 \geq 1) = 1 - \mathrm{P}(N_1 = 0) = 1 - 0.5987 = 0.40 \ \text{ using Table 1 in S}, \\ \text{technician 2}: \quad P = & \quad \mathrm{P}(N_2 \geq 3) = 0.030 \ \text{ using Minitab/Stata.} \end{aligned}$$

The P-values can also be obtained from Minitab's "1 Proportion" menu. For technician 1, the count of measurements outside the expected interval is not larger than would be expected by chance alone. On the other hand, for technician 2 the probability of getting 3 or more out of 14 observations outside the interval is small, and we have some evidence that the measurements do not follow the stated model. We can say that technician 2 has a too high proportion of extreme values, when compared to standard performance in the laboratory. Upon inspection of the values for technician 2 we see that several measurements are inside the interval but close to the endpoints, so the count could easily have been more more extreme.

5. Summary

In the first steps of our analysis, we computed confidence intervals for the two population means, and showed that there was no evidence against any of them being equal to the producer's stated value of 12.36. In the second step, we tested whether the two population means could be equal, and the non-significant P-value did not give us evidence to state a difference between them. In the third step, we computed an expected interval for single observations and compared the measurements of each technician to this interval. It turned out that technician 2 had a too high proportion of measurements outside the interval to be with any reason explained by chance alone. We noted already in the initial examination of the data that the spread for technician 2 was larger, and the estimated standard deviation of 0.1784 is quite a bit above 0.10. Therefore, one explanation of the excess number of measurements outside the prediction interval is that technician 2 has a higher standard deviation than routinely seen at the laboratory. Our previous inspection of the measurements also showed that the distribution for technician 1 complied better with a normal distribution than that of technician 2 which was somewhat right-skewed (although not strongly enough to give evidence against the normal distribution). Taking these findings together, it seems fair to say that the measurements of technician 2 show some disagreement with the established distribution at the laboratory, and this is therefore the most likely the unexperienced technician.
(*Note:* An alternative analytical approach would use confidence intervals and tests for standard deviations; some of this material is in Section 7.3 of IPS, but completely outside the course curriculum.)