This note is to clarify a point raised earlier by email. This email was stating that R does not produce the normal probability plot taught in lecture. This statement is true, but very unimportant, as shown in this note. **You can safely ignore both the statement (from the email) and this note!** If you are curious, this document is for you.

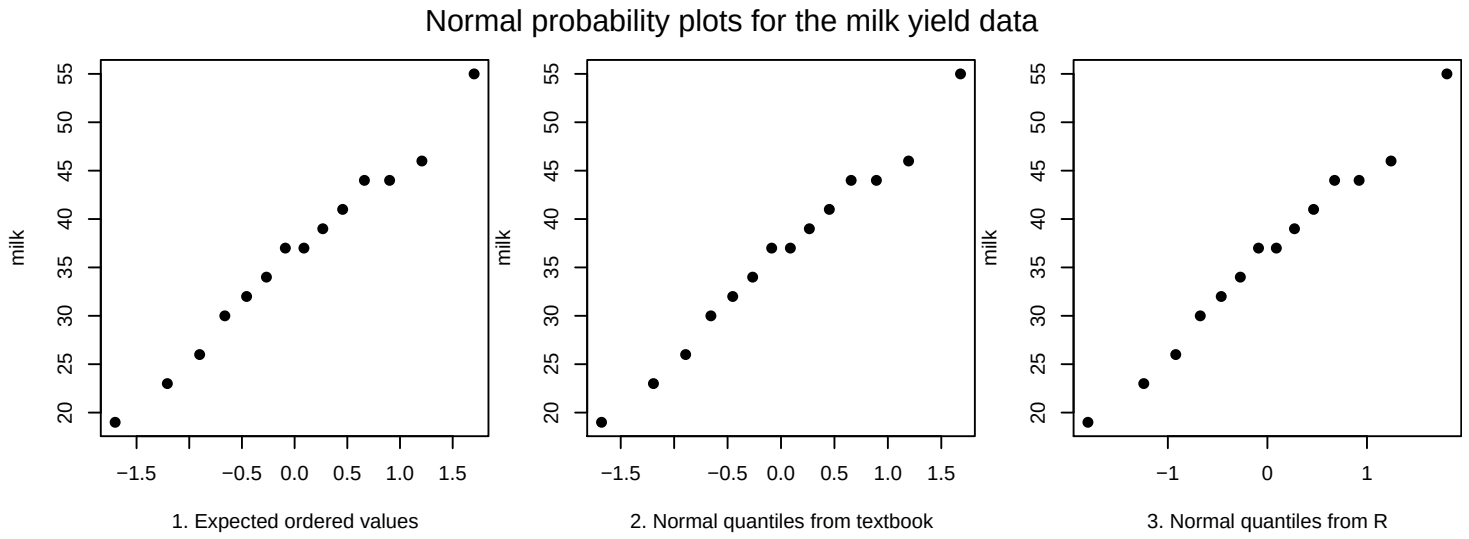
There are 3 different plots

All normal probability plots plot the ordered observations on the y-axis $(y_1, \ldots, y_n)$ against some "normal scores" $x_1, \ldots, x_n$ on the x-axis. What may differ between plots are the normal scores on the horizontal axis.

1. Normal scores taught in class, as well as in the textbook: x_1 is the expected value of the lowest observation in a random sample of size n from the standard normal $\mathcal{N}(0, 1)$. Then x_2 is the expected value of the second lowest observation, etc. With a sample size $n = 14$, we get that the expected smallest value is $x_1 = -1.70$, the expected second smallest value is $x_2 = -1.21$, etc., and the expected largest value is $x_{14} = 1.70$.

2. Normal scores as described in a footnote in the text (on p. 136): These are quantiles from $\mathcal{N}(0, 1)$. More specifically, x_1 is the $0.666/(n + 1/3)$ quantile, x_i is the $(i - 1/3)/(n + 1/3)$, etc. Although the textbook says 1. and 2. are the same, they are not really! The figure below shows that they are very close. That's why the textbook does not bother mentioning the difference. For instance, with a sample size of $n = 14$ we get here $x_1 = -1.69$, $x_2 = -1.19$ and $x_{14} = 1.69$.

3. Finally, R uses quantiles too, but different than in 2. In R plots, x_1 is the $0.5/n$ quantile, x_i is the $(i - 0.5)/n$ quantile, and x_n is the $1 - 0.5/n$ quantile. For this reason, R labels the x-axis with "Normal quantiles" instead of n-scores. With $n = 14$ we now get $x_1 = -1.80$, $x_2 = -1.24$ and $x_{14} = 1.80$.



The bottom line is this:

- We can't tell the difference by eye!
- If the points follow more or less a straight line, then we can assume that the data come from a normal distribution.