Problem 1.2
The lowess curve in Figure 1.2 has a flat spot between ages 25 and 35. Discuss how one might use bootstrap replications like those in Figure 1.3 to suggest whether the flat spot is genuine or just a statistical artifact.

Solution I think that the flat spot is a statistical artifact. There are several why I think that. First, if you look at Figure 1.3, we can see that the result of 25 bootstrap lowess replications between the ages of 25-35 show a large fluctuation. Actually, if you look at the original data, you can see that the variance of distribution of data is large. Second, lowess curve is not smooth in the 25-35 year old section.
Therefore, no matter how much bootstrap is repeated, I think flat spot is not accurate because the variance of data is large and the distribution is not smooth.

Problem 1.3
Suppose that there were no differences between AML and ALL patients for any gene, so that t in (1.6) exactly followed a student-t distribution with 70 degrees of freedom in all 7,128 cases. About how big might you expect the largest observed t value to be?
Hint : $\frac{1}{7128} = 0.00014$.

Solution For each $t_i \overset{i.i.d}{\sim} t(70) \quad i = 1, \ldots, 7128$ Let F be cdf of the t-distribution with 70 degrees of freedom. F is strictly increasing function, so F has inverse function denoted F^{-1}. Then, we know that $F(t_i) \overset{d}{\equiv} U(0,1)$ for all i. Let $t_{(i)}$ be order statistics. Define $U_i \sim U(0,1)$ $i = 1, \ldots, 7,128$ and $U_{(i)}$ be order statistics.
Then, $F(t_{(n)}) \overset{d}{\equiv} U_{(n)}$ and $U_{(n)} \sim Beta(n,1)$ with n = 7,128. Since the mean of $Beta(n,1)$ is $\frac{n}{n+1}$, then $E(F(t_{(n)}) = E(U_{(n)}) = \frac{n}{n+1}$. Thus, we can expect that $F(t_{(7,128)}) \approx \frac{7,128}{7,129} \Leftrightarrow t_{(7,128)} \approx F^{-1}(\frac{7,128}{7,129})$.
Therefore, $\underset{1 \le i \le 7,128}{max} t_i = t_{(7,128)} \approx t_{0.00014}(70) = 3.826058$ (the 0.00014 quantile for t distribution with degree of freedom 70)

$\therefore$ We can expect that the largest observed t value is 3.826058

Problem 2.3
Page 14 presents two definitions of frequentism, one in terms of probabilistic accuracy and one in terms of an infinite sequence of future trials. Give a heuristic argument relating the two.

Solution Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ indicate n independent draws from a probability distribution F and $x = (x_1, x_2, \ldots, x_n)$ be observed data. First definition of frequentist inference is that the accuracy of an observed estimate $\hat{\theta} = t(x)$ is the probabilistic accuracy of $\hat{\Theta} = t(X)$ as an estimator of $\theta = E_F(X)$. This means, if we estimate $\mu = E_F(\hat{\Theta})$, we use observed data $\hat{\theta} = t(x)$ and the accuracy of $\hat{\theta}$ is measured using the Mean Squared Error method which is decomposing $bias(\hat{\theta})^2 + var(\hat{\theta})$.
Second definition is that frequentism is defined with respect to an infinite sequence of future trials. This means, if we observe data infinitely from a distribution F, $\hat{\theta} = t(x)$ should be the same as $\hat{\Theta} = t(X)$

Problem 2.4
Suppose that in (2.15) we plugged in $\hat{\alpha}$ to get an approximate 95% normal theory hypothesis test for $H_0 : \theta = 0$. How would it compare with the student-t hypothesis test?

Solution If we know σ and under H_0, $\hat{\theta} = \bar{X_1} - \bar{X_2}$ follows exactly $N(0, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2}))$. Thus, the 95% confidence interval for $\theta = \mu_2 - \mu_1$ is $\bar{x_2} - \bar{x_1} \pm z_{0.025} * \sigma * (\frac{1}{n_1} + \frac{1}{n_2})^{1/2}$. This means,

$$Pr_{H_0}\{\bar{x_2} - \bar{x_1} - z_{0.025} * \sigma * (\frac{1}{n_1} + \frac{1}{n_2})^{1/2} \leq \theta \leq \bar{x_2} - \bar{x_1} + z_{0.025} * \sigma * (\frac{1}{n_1} + \frac{1}{n_2})^{1/2}\} = 0.95$$

If we plug in $\hat{\sigma}$ instead of σ,

$$Pr_{H_0}\{\bar{x_2} - \bar{x_1} - z_{0.025} * \hat{\sigma} * (\frac{1}{n_1} + \frac{1}{n_2})^{1/2} \leq \theta \leq \bar{x_2} - \bar{x_1} + z_{0.025} * \hat{\sigma} * (\frac{1}{n_1} + \frac{1}{n_2})^{1/2}\} \approx 0.95$$

which is not exactly the same 0.95
But if we use a different statistic $t = \frac{\bar{x2} - \bar{x_1}}{\hat{sd}}$ where $\hat{sd} = \hat{\sigma}(\frac{1}{n_1} + \frac{1}{n_2})^{1/2}$, then t follows exactly t-distribution with $n_1 + n_2 - 2$ degree of freedom under H_0. So we can calculate exactly the 95% confidence interval for $\theta = \mu_2 - \mu_1$. That is $\bar{x_2} - \bar{x_1} \pm t_{0.025}(n_1 + n_2 - 2) * \hat{\sigma} * (\frac{1}{n_1} + \frac{1}{n_2})^{1/2}$. Similarly,

$$Pr_{H_0}\{\bar{x_2} - \bar{x_1} - t_{0.025}(n_1 + n_2 - 2) * \hat{\sigma} * (\frac{1}{n_1} + \frac{1}{n_2})^{1/2} \leq \theta \leq \bar{x_2} - \bar{x_1} + t_{0.025}(n_1 + n_2 - 2) * \hat{\sigma} * (\frac{1}{n_1} + \frac{1}{n_2})^{1/2}\} = 0.95$$

Therefore, if we don't know the true value of σ, it's more accurate to estimate $\theta = \mu_2 - \mu_1$ using t-statistics than first method.