Solution to Homework 2

Question 1

(a). The test statistic \( T = \frac{\mu_1 - \mu_0}{\sqrt{s_1^2/n_1 + s_0^2/n_0}} \). For the data we have, we got \( T = -3.095 < -1.96 \). Then the null hypothesis is rejected. DPA measurement is associated with osteoporosis status. The 95% confidence interval is \((-0.19, -0.04)\).

(b). We have at least 95% confidence to claim that there is association.

(c). The p-value is \( P(|T| \geq 3.095) = 0.002 \). i.e., If there is no association between the DPA and osteoporosis, there is only a probability of less than 0.002 to get the test statistic value as extreme as what we got from our data.

(d). The mean difference in DPA between \( osteo = 1 \) and \( osteo = 0 \) groups is estimated to be \(-0.114\). The 95% level confidence interval of this mean difference is in \((-0.19, -0.04)\). The DPA for \( osteo = 1 \) group is significantly lower than that in the \( osteo = 0 \) group with p-value 0.002.

Question 2

The sample correlation coefficient is 0.073. It is 0.384 in the \( osteo = 0 \) group, but only \(-0.0003\) in the \( osteo = 1 \) group. The linear correlation between DPA and age is minor. In fact, there are some association in the \( osteo = 0 \) group. But DPA and age are almost uncorrelated in the \( osteo = 1 \) group.

Question 3
Answer (c).
(a) and (b) are talking about the linear correlation. Independence implies uncorrelated, but not true versa; (d) is right if it is ‘equal’.

Question 4

(a). \( \Pr(\text{user} \mid \text{positive}) = 25\% \)

\[
\Pr(\text{user} \mid +) = \frac{\Pr(\text{user} \cap +)}{\Pr(+)}
= \frac{\Pr(\text{user}) \Pr(+ \mid \text{user})}{\Pr(\text{user}) \Pr(+ \mid \text{user}) + \Pr(\text{nonuser}) \Pr(+ \mid \text{nonuser})}
= \frac{1}{1 + \frac{\Pr(\text{nonuser}) \Pr(+ \mid \text{nonuser})}{\Pr(\text{user}) \Pr(+ \mid \text{user})}}
= \frac{1}{1 + \frac{95\% \times 15\%}{5\% \times 95\%}}
= 25\% \tag{1}
\]

(b). \( \Pr(\text{nonuser} \mid \text{negative}) = 99.7\% \)

(c). This test does not do very well in picking true users among those with positive test results. If we look at the formula in (a), it is because \( \frac{\Pr(\text{nonuser}) \Pr(+ \mid \text{nonuser})}{\Pr(\text{user}) \Pr(+ \mid \text{user})} \) is large. Since \( \frac{\Pr(\text{nonuser})}{\Pr(\text{user})} \) are the percentage in the population and is fixed. We cannot do much on this. So if we can increase \( \Pr(\text{−} \mid \text{nonuser}) \) or \( \Pr(+ \mid \text{user}) \), then \( \Pr(\text{user} \mid +) \) will increase.

Question 5

Answer (a).
The population mean is a fixed number anyway. The confidence level of CI is saying the percentage that the CI’s covering the mean in repeated experiment.

Question 6

Answer (e).
Type I error is the probability of rejecting the null when the null hypothesis is true.