STA130, Winter 2017: Midterm

(100 minutes; 6 questions; 5 pages; total points = 66)

[SOLUTIONS]

1. [4] In three or four complete English sentences, explain what a P-value <u>is</u>, and what it is <u>used for</u>.

Solution. *You could write, for example: "A P-value is used when we want to test whether a certain effect is real or is just luck. The P-value is the probability that we would have obtained a result as extreme as, or more extreme than, the observed result, assuming the null hypothesis that there was no real effect. If the P-value is sufficiently small (e.g. less than 0.05), then we reject the null hypothesis and conclude that there was a real effect; otherwise, we do not reject the null hypothesis."*

2. Suppose Y is a random quantity having normal probabilities, with mean 20 and variance 16.

(a) [3] Compute $P(Y < 26)$. [Hint: don't forget the standard normal probability table included at the end of this test.]

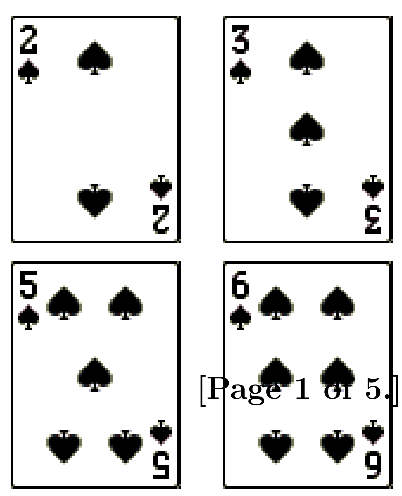
Solution. *Here* $sd(Y) = \sqrt{Var(Y)} = \sqrt{16} = 4$, *so* $Z = (Y - 20)/4$ *has the standard normal distribution, so* $P(Y < 26) = P((Y - 20)/4 < (26 - 20)/4) = P(Z < 1.5) \doteq 0.9332$ *from the table.*

(b) [3] Compute $P(Y < 18)$.

Solution. *Again* $Z = (Y - 20)/4$ *has the standard normal distribution, so* $P(Y < 18) = P((Y - 20)/4 < (18 - 20)/4) = P(Z < -0.5) = P(Z > 0.5) = 1 - P(Z < 0.5) \doteq 1 - 0.6915 \doteq 0.3085$ *from the table.*

(c) [3] Provide an estimate of $P(Y > 100)$. [Hint: if a value is too large for the standard normal table, then what can you conclude?]

Solution. *Again* $Z = (Y - 20)/4$ *has the standard normal distribution, so* $P(Y > 100) = P((Y - 20)/4 > (100 - 20)/4) = P(Z > 20) = 1 - P(Z < 20)$. *Now, the table only goes as high as* $P(Z < 3.09)$. *Since this probability is 0.9990, therefore* $P(Z < 20)$ *must be even <u>more</u> than this, so that* $P(Y > 100)$ *is even <u>less</u> than* $1 - 0.9990 = 0.001$. *(Actually it is <u>much</u> less than that.)*



3. Suppose 50 people each have their own deck of cards. Each person picks either the 2 or 3 or 5 or 6 of Spades, with probability 1/4 each. Let X_1 be the card chosen by the <u>first</u> person. Let T be the <u>sum</u> of all 50 cards, and let $Y = T/50$ be the <u>average</u> of all 50 cards.

(a) [4] For X_1, compute the expected value $E(X_1)$ and the variance $Var(X_1)$.

Solution. $E(X_1) = \sum_x x\, P(X_1 = x) = 2(1/4) + 3(1/4) + 5(1/4) + 6(1/4) = 16/4 = 4$. *Then* $Var(X_1) = \sum_x (x-4)^2\, P(X_1 = x) = (2-4)^2(1/4) + (3-4)^2(1/4) + (5-4)^2(1/4) + (6-4)^2(1/4) = 10/4 = 2.5$.

(b) [**4**] For T, compute the expected value $E(T)$ and the variance $Var(T)$.

Solution. *Here* $T = X_1 + X_2 + \ldots + X_{50}$, *where each* X_i *is the card chosen by the* i^{th} *person. So,* $E(X_i) = 4$ *and* $Var(X_i) = 2.5$ *as above. Then* $E(T) = 4 * 50 = 200$, *and* $Var(T) = 2.5 * 50 = 125$.

(c) [**4**] For Y, compute the expected value $E(Y)$ and the variance $Var(Y)$.

Solution. *Since* $A = T/50$, *therefore* $E(A) = E(T)/50 = 200/50 = 4$, *and* $Var(A) = Var(T)/50^2 = 125/2500 = 0.05$.

4. The champion New England Patriots football team won 14 of the 16 games they played during this year's regular NFL season. Suppose we wish to test the null hypothesis that their games were all just random luck with probability 1/2 of winning each game, versus the alternative hypothesis that they had probability <u>more</u> than 1/2 of winning each game.

(a) [**3**] Under this null hypothesis, what would be the mean and variance and sd for the number of games a team would win (out of 16 games total)?

Solution. *Here* $n = 16$ *and* $p = 1/2$, *so mean* $= np = 16(1/2) = 8$, *and variance* $= np(1-p) = 16(1/2)(1/2) = 4$, *and sd* $= \sqrt{4} = 2$.

(b) [**3**] Under this null hypothesis, what would be the probability of a team winning all 16 games (out of 16 games total)?

Solution. *The probability of winning all 16 games is* $(1/2) * (1/2) * \ldots * (1/2) = 1/2^{16} = 1/65536 \doteq 0.000015$.

(c) [**3**] Under this null hypothesis, what would be the probability of a team winning exactly 15 games (out of 16 games total)? [Hint: How many sequences are there corresponding to 15 W and 1 L?]

Solution. *The one "L" could be in any of 16 positions. So, the probability of winning exactly 15 games is* $16 * (1/2)^{15} * (1/2)^{16-15} = 16(1/2)^{16} = 16/65536 = 1/4096 \doteq 0.00024$.

(d) [**3**] Under this null hypothesis, what would be the probability of a team winning exactly 14 games (out of 16 games total)? [Hint: You may use the fact that $\binom{16}{14} = 120$.]

Solution. *According to the binomial probability formula, the probability of winning exactly 14 games out of 16 is* $\binom{16}{14}(1/2)^{14}(1-(1/2))^{16-14} = 120(1/2)^{16} = 120/65536 = 15/8192 \doteq 0.00183$.

(e) [**3**] Using all of the above information, what is the P-value for this hypothesis test?

Solution. *The P-value here is the probability, under the null hypothesis, of winning 14 or more games out of 16, i.e. of winning 14 or 15 or 16 games. From the above, this is equal to the sum* $0.000015 + 0.00024 + 0.00183 \doteq 0.00208$.

(f) **[2]** What can we conclude from this P-value? (State your conclusion clearly, using complete English sentences.)

Solution. *This P-value is much smaller than 0.05, so we can reject the null hypothesis. Thus, we conclude that during this year's NFL regular season, the Patriots had probability more than 1/2 of winning each game.*

5. A recent study[1] was reported with such headlines[2] as "Patients treated by female doctors less likely to die, study shows". The study examined 415,559 elderly patients who saw female doctors, of which 10.82% of them died within 30 days. It also examined 1,200,296 elderly patients who saw male doctors, of which 11.49% of them died within 30 days.

(a) **[1]** Based on the above, what (approximately) is the actual number of patients who saw female doctors and then died within 30 days?

Solution. *This number is 10.82% of 415,559, i.e. it equals approximately* $415559 * 10.82/100 = 44963.5$*, or about 44,964.*

(b) **[1]** Based on the above, what (approximately) is the actual number of patients who saw male doctors and then died within 30 days?

Solution. *This number is 11.49% of 1,200,296, i.e. it equals approximately* $1200296 * 11.49/100 = 137914$*, i.e. about 137,914.*

(c) **[3]** Under the null hypothesis that 10.9% of all patients who see female doctors will die within 30 days, compute the mean and variance and standard deviation for the number of patients out of $415,559$ who would die within 30 days of seeing a female doctor.

Solution. *Under the null hypothesis that* $p = 10.9\% = 0.109$*, the number of patients out of* $n = 415,559$ *who would die within 30 days has mean* $np = 415559 * 0.109 \doteq 45296$*, and variance* $np(1-p) = 415559 * 0.109 * (1 - 0.109) \doteq 40359$*, hence sd* $\sqrt{40359} \doteq 200.9$.

(d) **[4]** Compute (with explanation) a P-value for the null hypothesis that 10.9% of all patients who see female doctors will die within 30 days, versus the alternative hypothesis that it's less than 10.9%. What can we conclude from this?

Solution. *This P-value is the probability, assuming the null hypothesis* $p = 0.109$*, that if* $n = 415,559$ *patients saw a female doctor, then 44,964 or less of them would die. This is approximately the probability that a normal random quantity with mean 45296 and sd 200.9 would equal 44964 or less. That is the same as the probability that a standard normal would equal* $(44964 - 45296)/200.9$ *or less , i.e. equal about* -1.65 *or less. Using the table, this is* $P(Z < -1.65) = P(Z > 1.65) = 1 - P(Z < 1.65) \doteq 1 - 0.9505 = 0.0495$*. This is slightly less*

[1] http://jamanetwork.com/journals/jamainternalmedicine/article-abstract/2593255
[2] http://www.medicalnewstoday.com/articles/314912.php

than 0.05. So, according to standard scientific practice, we reject the null hypothesis, and conclude that the true percentage is less than 10.9%.

6. Consider again the study described in Question 5.

(a) [**2**] Write down the general formula, in terms of the unknown true fractions p_1 and p_2, for the sd of the difference $\hat{p}_2 - \hat{p}_1$ between the observed fraction of deaths within 30 days after seeing a female doctor, minus the observed fraction after seeing a male doctor.

Solution. *We know from class that the general formula is* $\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2} = \sqrt{p_1(1-p_1)/1200296 + p_2(1-p_2)/415559}$, *where* p_1 *is the true fraction for male doctors, and* p_2 *is the true fraction for female doctors.*

(b) [**2**] Estimate the sd in part (a) using the bold option.

Solution. *For the bold option, we replace* p_1 *by* $\hat{p}_1 = 11.49\% = 0.1149$, *and replace* p_2 *by* $\hat{p}_2 = 10.82\% = 0.1082$, *to obtain the sd estimate*
$\sqrt{0.1149 * (1 - 0.1149)/1200296 + 0.1082 * (1 - 0.1082)/415559} \doteq 0.00056$.

(c) [**2**] Estimate the sd in part (a) using the conservative option. How is it different?

Solution. *For the conservative option, we replace* p_1 *by* $1/2$, *and replace* p_2 *by* $1/2$, *to obtain the sd estimate* $\sqrt{1/1200296 + 1/415559}\,/2 \doteq 0.00090$. *In this case the conservative option is quite a bit larger than the bold option, because the values* $\hat{p}_1$ *and* $\hat{p}_2$ *are very far from 1/2.*

(d) [**4**] Using the bold option, compute (with explanation) an estimate of the P-value for the null hypothesis that the same fraction of patients will die within 30 days whether they see a female or a male doctor, versus the alternative hypothesis that the fractions are different (either larger or smaller, i.e. two-sided).

Solution. *Under the null hypothesis and bold option, the difference between the fraction of deaths with a female doctor, minus the fraction with a male doctor, is approximately normal with mean 0 and sd 0.00056 as above. The P-value is the probability that this random quantity would be as extreme or more extreme than the observed difference of* $|10.82\% - 11.49\%| = |0.1082 - 0.1149| = 0.0067$. *So, the P-value is the probability that a normal with mean 0 and sd 0.00056 is more than 0.0067 or less than* -0.0067. *That equals the probability that a standard normal is more than* $0.0067/0.00056 \doteq 11.96$ *or less than* -11.96. *Since* $11.96 > 3.09$, *all we can conclude from the table is that this probability is less than* $(1 - 0.9990) + (1 - 0.9990) = 0.002$. *(Actually it is much less than that.)*

(e) [**2**] State your final conclusion from part (d) in a clear, complete English sentence.

Solution. *The P-value from part (d) is much less than 0.05, so we can reject the null hypothesis, and conclude that the percentage of patients who will die after visiting a female or male doctor is not the same, i.e. it is different.*

(f) **[3]** Has this study convinced you that seeing a female doctor makes people live longer? Why or why not? (There is no completely right or wrong answer here; just explain your opinion with some reasons.)

Solution. *You could be convinced, since the P-value is so small. Or, you could raise various objections, e.g. that perhaps the patients were not assigned randomly, and the patients who chose to see a female doctor may have been healthier or younger, or more likely to live longer for some other reason besides their doctor's gender. The important thing is that you express your opinion clearly, in good English, with valid reasons.*

[END OF EXAMINATION: total points = 66]