UCSC AMS 5 **Quiz 3** **Summer Session I, 2016**

NAME: ***Solutions***

- For full credit, ***show your work and explain your reasoning on every question.***
- Please write clearly — if we can't read it, we won't give you credit for it.

1. A study of the relationship between cigarette smoking and blood pressure in adult men collected data from 6235 U.S. men aged 20 - 40, and generated the following statistics:

$$\overline{X} = 24 \quad SD_X = 5.5$$
$$\overline{Y} = 135 \qquad SD_Y = 9 \qquad r = 0.7$$

where X = number of cigarettes per day, and Y = systolic blood pressure, measured in mmHG.

(a) (5 pts) Use this data to find the regression equation that predicts blood pressure from number of cigarettes smoked per day, and find the R.M.S. error of regression (SER).

The regression equation is $\hat{y} = \beta_0 + \beta_1 x$, where $\hat{y}$ is the predicted average y-value for all observations with the same x-value (x). The coefficients β_0, β_1 are found using the formulas

$$\beta_1 = r_{xy} \cdot \frac{SD_y}{SD_x} = 0.7 \cdot \frac{9}{5.5} \approx 1.145 \text{ and } \beta_0 = \overline{y} - \beta_1 \overline{x} \approx 135 - 1.145 \cdot 24 = 107.52,$$

so the regression equation is

$$\hat{y} = 107.52 + 1.145x.$$

This is the same as the formula that we get using the 'regression method':

$$\hat{y} = \overline{y} + r_{xy} \cdot \frac{x - \overline{x}}{SD_x} \cdot SD_y = 135 + \frac{0.7 \cdot (x - 124) \cdot 9}{5.5}.$$

The R.M.S. error of regression is the number

$$SER = \sqrt{1 - r_{xy}^2} \times SD_y = \sqrt{1 - 0.49} \times 9 \approx 6.43.$$

(b) (3 pts) What is the predicted systolic blood pressure for a 30-year old man who smokes 30 cigarettes a day? Include a margin of error (a 'plus/minus' number) with your estimate.

The predicted blood pressure for a 30-year old man who smokes 30 cigarettes/day is the average blood pressure of all men age 20-40 who smoke 30 cigarettes/day, which we estimate using the regression equation:

$$\hat{y}(30) = 107.52 + 1.145 \cdot 30 = 141.87.$$

The margin of error is given by the SER, so that the predicted blood pressure for such a man is

$$141.87 \pm 6.43 \text{ mmHg}.$$

(c) (2 pts) Jane, a 32-year old woman who smokes 20 cigarettes a day, uses this regression equation to predict her systolic blood pressure to be about 130.4. Is this right? Explain.

This is ***not*** right. Jane is a woman, so we can't accurately predict her blood pressure using data about men. (Gender is a *confounding variable* for blood pressure.)

2. A fair coin is tossed three times.

(a) (3 pts) What is the probability that ***all three*** tosses result in *Heads*?

To find the probability of HHH, we rely on the assumption that the outcomes of the different tosses are *independent*, so that we can multiply the probabilities, as below:

$$\begin{aligned} P(HHH) &= P(H \text{ on 1st toss } \textbf{\textit{and}}\ H \text{ on 2nd toss } \textbf{\textit{and}}\ H \text{ on 3rd toss}) \\ &= P(H \text{ on 1st toss}) \times P(H \text{ on 2nd toss}) \times P(H \text{ on 3rd toss}) \\ &= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8} = 12.5\% \end{aligned}$$

(b) (2 pts) What is the probability that ***exactly one*** of the tosses results in *Heads*?

If we observe exactly one H in three tosses, then exactly one of the three sequences HTT, THT or TTH must occur. The probability of each of these is 12.5% (just like HHH), and since they are *mutually exclusive*,[†] we can add the probabilities as below:

$$\begin{aligned} P(\text{exactly one } H \text{ in three tosses}) &= P(\ HTT\ \textbf{\textit{or}}\ THT\ \textbf{\textit{or}}\ TTH) \\ &= P(HTT) + P(THT) + P(TTH) \\ &= 12.5\% + 12.5\% + 12.5\% = 37.5\% \qquad (= 3/8) \end{aligned}$$

3. Four draws are made at random, ***with replacement*** from the box $\boxed{\boxed{1}\boxed{2}\boxed{2}\boxed{3}}$.

(a) (3 pts) What is the probability that $\boxed{2}$ is drawn *at least one time*?

First, the opposite of '*at least one* $\boxed{2}$ *in four draws*' is '***no*** $\boxed{2}$*s in four draws*'.

Second, The probability of ***not*** drawing a $\boxed{2}$ on any one draw is 1/2 because half of the tickets in the box are not $\boxed{2}$s.

Third, the outcomes of the individual draws are *independent* because the draws are done *at random, with replacement*, and this means that

$$\begin{aligned} P(\text{no } \boxed{2}\text{s in four draws}) &= P(\text{no } \boxed{2} \text{ on 1st } \textbf{\textit{and}} \text{ no } \boxed{2} \text{ on 2nd } \textbf{\textit{and}} \text{ no } \boxed{2} \text{ on 3rd } \textbf{\textit{and}} \text{ no } \boxed{2} \text{ on 4th}) \\ &= P(\text{no } \boxed{2} \text{ on 1st}) \times P(\text{no } \boxed{2} \text{ on 2nd}) \times P(\text{no } \boxed{2} \text{ on 3rd}) \times P(\text{no } \boxed{2} \text{ on 4th}) \\ &= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16} \end{aligned}$$

Finally,

$$P(\text{at least one } \boxed{2} \text{ in 4 draws}) = 1 - P(\text{no } \boxed{2}\text{s in four draws}) = 1 - \frac{1}{16} = \frac{15}{16} = 93.75\%.$$

(b) (2 pts) Does your answer change if the draws are made ***without replacement***? If so, what is the new probability? If not, explain why not.

Yes – the answer changes! If we draw four times ***without replacement*** then we draw ***all*** the tickets and we definitely see at least on $\boxed{2}$ (in fact we see two of them). I.e., the probability is 100% that we see at least one $\boxed{2}$ in this case.

[†]If we see HTT then we don't see THT, for example.