Solutions for review problems

Modules 1-6

Stat 251

updated 2021

The exam will also cover definitions from the first assignment (A1), observational study design, experimental design, understanding graphs, and summary statistics (sample mean, etc.).

(1) Describe the distributions of the following histograms:
   (a) Nile River flow: approximately symmetric (bell curve) and is unimodal with approximate center $90010^8 m^3$
   (b) Old Faithful eruption duration: bimodal with center about 3.5 minutes
   (c) Passenger air miles flown: skewed to the right, is unimodal with approximate center 10,000 miles
   (d) CO_2 levels: skewed to the left, is unimodal with approximate center -0.05 units (cannot remember units of measurement for this one)

(2) Use the Empirical Rule to describe the distribution of the Nile River flow. The mean flow of the Nile is 919.35 and standard deviation is 169.23 (in $10^8 m^3$).

$$68\% \text{ observations are within } \overline{x} \pm 1s = 919.35 \pm 169.23 = 750.12, 1088.58$$

$$95\% \text{ observations are within } \overline{x} \pm 2s = 919.35 \pm 2(169.23) = 580.89, 1257.81$$

$$99.7\% \text{ observations are within } \overline{x} \pm 3s = 919.35 \pm 3(169.23) = 411.66, 1427.04$$

(3) Suppose we have the following 7 ozone measurements from New York, NY (1973). The units of measurement is in ppb (parts per billion). ozone=`{41,36,12,18,28,23,19}`. Calculate the following:
   (a) Mean: $\overline{X} = \frac{41+36+12+18+28+23+19}{7} = 25.2857143\ ppb$
   (b) Variance: $s^2 = \frac{(41-25.29)^2+(36-25.29)^2+(12-25.29)^2+(18-25.29)^2+(28-25.29)^2+(23-25.29)^2+(19-25.29)^2}{7-1} = \frac{643.4285714}{6} = 107.2380952\ ppb^2$
   (c) Standard deviation: $s = \sqrt{107.24} = 10.3555828\ ppb$
   (d) The boxplot shows no sign of outliers (outliers on a boxplot are indicated by small circles that are outside the boundaries of the "whiskers"), so we can assume that there are most likely no outliers.

(4) A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments, and others a diverse mixture of housing structures. Identify the sampling methods described below, and comment on whether or not you think they would be effective in this setting. SRS=simple random sample, StRS=stratified random sample, 1cluster=1-stage cluster, 2cluster=2-stage cluster, SyRS=systematic sample
   (a) Randomly sample 50 households from the city: SRS; could be good for use, may not be representative if the sample does not capture certain neighborhoods (it really depends on the population of interest).

(b) Divide the city into neighborhoods, and randomly sample 20 households from each neighborhood: StRS; is better for representation of all neighborhoods (if that is ideal), could be somewhat costly to sample from every neighborhood.
(c) Divide the city into neighborhoods, randomly sample 10 neighborhoods, and sample all households from those neighborhoods: 1-stage cluster; could be good for use, may not be representative if the sample does not capture certain neighborhoods, is more cost effective than StRS
(d) Divide the city into neighborhoods, randomly sample 10 neighborhoods, and then randomly sample 20 households from those neighborhoods: 2-stage cluster; could be good for use, may not be representative if the sample does not capture certain neighborhoods, is more cost effective than StRS and 1-stage cluster
(e) Sample the 200 households closest to the city council offices: convenience; not a random sample, therefore it is a bad sample. Convenience sample=lazy sample, NEVER use

(5) A study is designed to test the effect of light level on exam performance of students. The researcher believes that light levels might have different effects on males and females, so wants to make sure both are equally represented in each treatment. The treatments are fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps).

(a) Response variable: exam performance of students
(b) Explanatory variable: lighting levels, 3 levels: fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps)
(c) Blocking variable: gender, 2 levels: M,F

(6) You would like to conduct an experiment in class to see if your classmates prefer the taste of regular Coke or Diet Coke. Briefly outline a design for this study: A paired design where the subject is given both sodas in random order and then tells researcher their preference.

(7) Real estate ads suggest that 64% of homes for sale have garages, 21% have swimming pools, and 17% have both. Find the following probabilities:

(a) No garage:
$P(garage') = 1 - P(garage) = 1 - .64 = 0.36$
(b) No pool:
$P(pool') = 1 - P(pool) = 1 - .21 = 0.79$
(c) Pool or garage:
$P(pool \cup garage) = P(pool) + P(garage) - P(pool \cap garage) = 0.21 + 0.64 - 0.17 = 0.68$
(d) Pool but no garage (matrix value): $P(pool \cap garage') = 0.04$
(e) Neither a pool nor a garage (make matrix and value is in matrix):
$P(garage' \cap pool') = 0.32$ OR $= 1 - P(pool \cup garage) = 1 - 0.68 = 0.32$
(f) Are having a pool and a garage independent? Show work
$?P(pool \cap garage) = P(pool)P(garage)? \Rightarrow 0.17? =?(0.64)(0.21) \Rightarrow 0.17 \neq 0.1344 \therefore$ having a pool and a garage are not independent (or they are dependent)
(g) Are having a pool and a garage mutually exclusive? Explain. Since the intersection between having a house with both a pool and a garage exists, they cannot be mutually exclusive (disjoint).

	$P(pool)$	$P(pool')$	
$P(garage)$	0.17	0.47	0.64
$P(garage')$	0.04	0.32	0.36
	0.21	0.79	1

(8) A metal fabricating plant currently has five major pieces under contract, each with a deadline for completion. Let X be the number of pieces completed by their deadlines. Using the following pmf (probability mass function - the table given) to find:

(a) What is the probability of 5 pieces completed by deadline? Use complement rule: $P(X = 5) = 1 - [P(0) + P(1) + P(2) + P(3) + P(4)] = 1 - 0.90 = 0.10$

(b) What is the probability that at least 3 are completed by deadline? $P(X \geq 3) = P(3)+P(4)+P(5) = 0.25 + 0.35 + 0.10 = 0.70$

(c) What is the probability that no more than 3 are completed by deadline? $P(X \leq 3) = P(3) + P(2) + P(1) + P(0) = 0.25 + 0.15 + 0.10 + 0.05 = 0.55$

(d) Calculate EX, VX, and SDX $EX = \sum xp(x) = 1(0.10)+2(0.15)+3(0.25)+4(0.35)+5(0.10) = 3.05$
$VX = \sum (x - EX)^2 p(x) = (0 - 3.05)^2(0.05) + (1 - 3.05)^2(0.10) + (2 - 3.05)^2(0.15)$
$+(3 - 3.05)^2(0.25) + (4 - 3.05)^2(0.35) + (5 - 3.05)^2(0.1) = 1.7475$
$SDX = \sqrt{VX} = \sqrt{1.7475} = 1.3219304$

X	0	1	2	3	4	5
$p(x)$	0.05	0.10	0.15	0.25	0.35	0.10

(9) A 2010 Pew Research poll asked 1,306 Americans "From what you've read and heard, is there solid evidence that the average temperature on earth has been getting warmer over the past few decades, or not?". The table below shows the distribution of responses by party and ideology, with relative frequencies.

(a) Are believing that the earth is warming and being a liberal Democrat mutually exclusive?
Since the intersection between belief that the earth is warming and being a liberal Democrat exists, they are not mutually exclusive

(b) What is the probability that a randomly chosen respondent believes the earth is warming or is a liberal Democrat?
$P(Yes \cup LD) = P(Yes) + P(LD) - P(Yes \cap LD) = 0.6 + 0.2 - 0.18 = 0.62$

(c) Does it appear that whether or not a respondent believes the earth is warming is independent of their party and ideology? Show work
If A and B are independent, then $P(A \cap B) = P(A)P(B)$. Is $P(Yes \cap MSD) = P(Yes)P(MSD)$? $P(Yes \cap MSD) = 0.25$ and $P(Yes)P(MSD) = (0.6)(0.34) = 0.204$ $\therefore$ $P(Yes \cap MSD) \neq P(Yes)P(MSD)$ and they are not independent

Party/ideology	Yes	No	Don't know/refuse	Total
Conservative republican	0.11	0.20	0.02	0.33
Mod/Lib republican	0.06	0.06	0.01	0.13
Mod/Cons democrat	0.25	0.07	0.02	0.34
Liberal democrat	0.18	0.01	0.01	0.20
Total	0.60	0.34	0.06	1

(10) Dr. Peter Venkman wanted to do a test on ESP. He randomly selected his volunteers and they were shown one card of 4 different ones, one card at a time (blank side facing the subject) and were told to guess what shape they thought was on the back side of the card. The test was done for a total of 10 cards per subject. (a) What is the name of the probability distribution for this? What are the parameters of this distribution?
This is the binomial distribution with parameters $n = 10$ and $p = 0.25$.
The probability is 0.25 because by random chance, you have a 1 in 4 chance of guessing correctly (b) What is the probability that a random subject can get exactly one card correct?
$P(X = 1) = \binom{10}{1}(0.25^1)(0.75^{10-1}) = 0.1877117$ (c) What is the probability that they will get at least 8

cards correct?
$P(X \geq 8) = P(X = 8) + P(X = 9) + P(X = 10) : P(X = 10) = \binom{10}{10}(0.25^10)(0.75^{10-10}) = 9.5367432 \times 10^{-7}$
$P(X = 9) = \binom{10}{9}(0.25^9)(0.75^{10-9}) = 2.8610229 \times 10^{-5}$
$P(X = 8) = \binom{10}{8}(0.25^8)(0.75^{10-8}) = 3.862381 \times 10^{-4}$
$P(X \geq 8) = 9.5367432 \times 10^{-7} + 2.8610229 \times 10^{-5} + 3.862381 \times 10^{-4} = 4.15802 \times 10^{-4}$
(d) Suppose that a subject actually has some ESP. Change the probability of success p to 0.5. Now calculate the probability of getting exactly 6 cards correct.
$P(X = 6) = \binom{10}{6}(0.5^6)(0.5^{10-6}) = 0.016222$ (e) Calculate EX, VX and SDX, using both probabilities (so 2 sets of EX, VX, and SDX). $EX = np = 10(.25) = 2.5$ $VX = npq = 10(.25)(.75) = 1.875$ $SDX = \sqrt{npq} = 1.3693064$ $EX = np = 10(.5) = 5$ $VX = npq = 10(.5)(.5) = 2.5$ $SDX = \sqrt{npq} = 1.5811388$

(11) The number of calls coming per minute into a hotels reservation center is Poisson random variable with mean 3. (a) Find the probability that no calls come in a given 1 minute period.
$P(X = 0) = \frac{e^{-3}3^0}{0!} = 0.0497871$ (b) What is the probability that at least 1 call comes in a given 1 minute period?
$P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 1 - \frac{e^{-3}3^0}{0!} = 0.9502129$ (c) Calculate EX, VX and SDX
$EX = \mu = 3$, $VX = \mu = 3$, $SDX = \sqrt{\mu} = 1.7320508$

(12) Find the probability of the following z-scores: (a) $P(Z < 1.89) = 0.970621$ (b) $P(Z > -0.5) = 1 - P(Z < -.5) = 0.6914625$ (c) Find the z-score that represents the top 9% $z_{top9\%} = z_{bottom91\%} = 1.340755$ (d) $P(-1 < Z < 0.87) = P(Z < 0.87) - P(Z < -1) = 0.8078498 - 0.1586553 = 0.6491945$

(13) Suppose the diameter at breast height (in.) of maple trees is normally distributed with mean 8.8 and standard deviation 2.8. $X \sim N(8.8, 2.8)$ and use $z = \frac{X-\mu}{\sigma}$ (a) What is the probability that the diameter of a randomly selected tree will be at least 10 in.? $P(X \geq 10) = P(Z \geq \frac{10-8.8}{2.8}) = P(Z \geq 0.43) = 0.3341$
(b) What is the probability that the diameter of a randomly selected tree will exceed 20 in.? $P(X \geq 20) = P(Z \geq \frac{20-8.8}{2.8}) = P(Z \geq 4) = 0$
(c) What is the probability that the diameter of a randomly selected tree will be between 5 and 10 in.? $P(5 < X < 10) = P\left(\frac{5-8.8}{2.8} < Z < \frac{10-8.8}{2.8}\right) = P(-1.36 < Z < 0.43) = P(Z < 0.43) - P(Z < -1.36)$ $= 0.6659 - 0.0874 = 0.5785$
(d) What is the probability that the diameter of a randomly selected tree will be less than 6 in.? $P(X < 6) = P(Z < \frac{6-8.8}{2.8}) = P(Z < -1) = 0.1587$
(e) How wide are the widest 2% of trees? $z_{top2\%} = z_{.98} = 2.05$ now solve for x where $x = z\sigma + \mu$ $x = (2.05)(2.8) + 8.8 = 14.55$