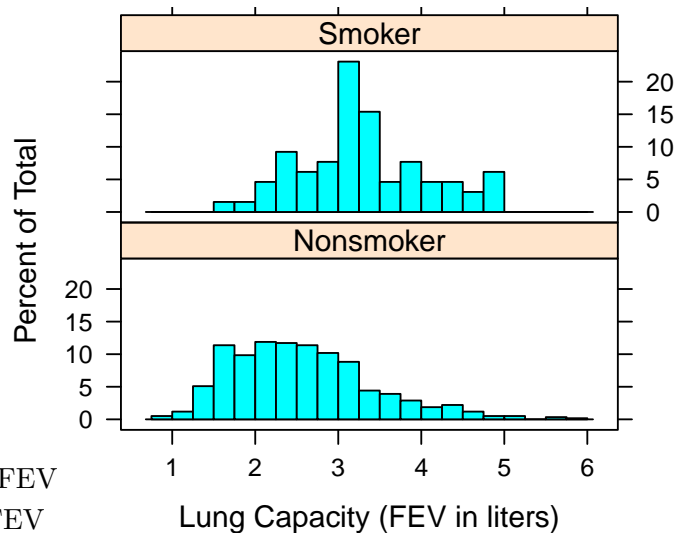


2020 Summer STAT 22000 Midterm Exam July 9, 2020

1. During the exam, you may refer to the textbook, slides, homework, and the solutions and other materials posted on Canvas. You can use R during the exam. However, you cannot use Google or other search engines during the exam. You must do the exam all by yourself. You cannot get assistance from other people.
2. If a question asks you do some calculations, you must **show your work to receive full credit**.
3. Please check Canvas or email regularly during the exam. Yibi might send out corrections or clarifications about exam problems that you don't want to miss.
4. If you are unsure of what a question is asking for, **you may send questions to Yibi by email**.
5. Whenever appropriate, parts of a question will be graded conditionally on how you answered the preceding part(s). For example, even if you get part (a) of a question wrong, you will still get credit for the rest of the question provided your answers to parts (b), (c), etc. are consistent with how you answered part (a).

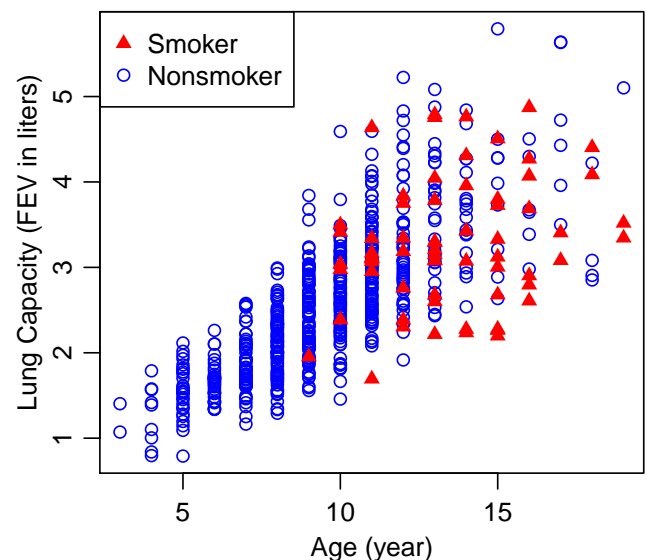
1. We will examine data on lung capacity, as measured by a quantity called forced expiratory volume (to be abbreviated FEV), the amount of air an individual can exhale in the first second of forceful breath (in liters). The following graph displays the distributions of FEV values for 65 smokers and 589 nonsmokers who participated in a research study.



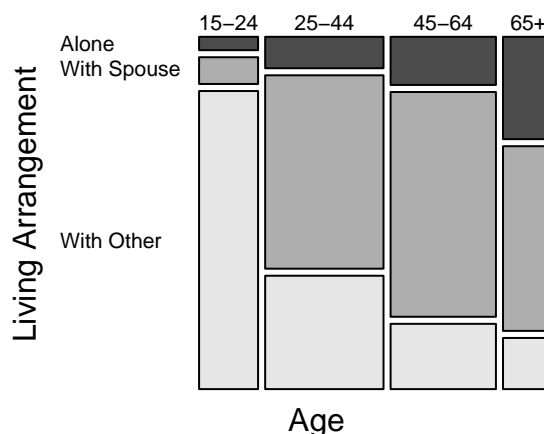
- (1a) (2pts) For the 589 nonsmokers in the study,
- (i) the mean FEV is higher than the median FEV
 - (ii) the mean FEV is lower than the median FEV
 - (iii) the mean FEV is about equal to the median FEV
- Select only ONE answer.**
- (1b) (2pts) For the 65 smokers in the study, the standard deviation (SD) of the FEV values is closest to which of the following numbers **Select only ONE answer.**
- (i) 0.25
 - (ii) 0.75
 - (iii) 1.5
 - (iv) 3
- (1c) (2pts) The Q3 (third quartile = 75th percentile) of the FEV of the 589 nonsmokers in the study is closest to which of the following numbers. **Select only ONE answer.**
- (i) 1
 - (ii) 2
 - (iii) 3
 - (iv) 4.5
- (1d) (2pts) Based on the histograms, which group — smokers or non-smokers — tends to have larger lung capacities (larger FEV)? Or are they about the same? **Select only ONE answer.**
- (i) smokers
 - (ii) nonsmokers
 - (iii) about the same

- (1e) (5pts) Now let's consider one more variable — **age** of the participants. The scatterplot below shows the relationship between the age and lung capacity of smokers and non-smokers.

After accounting for the age of the participants, which group — smokers or non-smokers — tends to have larger lung capacities (larger FEV)? Or are they about the same? Do you get the same conclusion as in the previous part? How do you explain the seemingly contradictory conclusions here and the previous part?



2. As reported by the U.S. Census Bureau in *American's Families and Living Arrangements*, the living arrangements by age of U.S. citizens age 15+ are as shown in the mosaic plot on the right.



- (2a) (2pts) Which of the following groups contains the most number of people? **Select only ONE answer.**
- (i) age 15-24 and live alone
 - (ii) age 25-44 and live with other
 - (iii) age 45-64 and live alone
 - (iv) age 65+ and live with other
- (2b) (2pts) Based on the mosaic plot, which of the following statement is TRUE? **Select only ONE answer.**
- (i) Over half of people in the age group 45-64 live with spouse.
 - (ii) The living arrangement is independent of age for U.S. citizens age 15+.
 - (iii) Older people are more likely to live with others than younger people.
 - (iv) There are more people in the age group 65+ than in the age group 45-64.

3. “Where do you get most of your information about current news events?” This question was asked in the 2008 General Social Survey. Possible answers included television, Internet, and newspapers, as well as other possibilities such as radio, family, and friends. The table one the right summarizes the results by age group.

Age	TV	Internet	Newspapers	Other	Total
18-29	109	92	25	36	262
30-49	272	157	88	63	580
50+	345	59	165	63	632
Total	726	308	278	162	1474

- (3a) (2pts) What percentage of respondents were between 18 and 29 years old and got news primarily from the internet? **Show your calculation.**
- (3b) (4pts) Which of the following statements implies there is an association between the respondents' age and the primary source they get news. **Select all that apply.**
- (i) There was a higher percentage of people getting news primarily from the internet among younger people than among older people.
 - (ii) In the 18-29 age group, there were more people getting news primarily from TV and from the internet than from newspaper and from other sources.
 - (iii) Those who got news primarily from the internet were generally younger than those got news primarily by reading newspaper.
 - (vi) The percentage of people getting news primarily from the internet in the 18-29 age group was higher than the percentage getting news primarily from the internet over all age groups.
 - (v) Among those who got news primarily from the Newspaper, a majority of them were 50 or older.
4. (4pts) Please find the mean, median and standard deviation (SD) for the data below.

3, 2, 8, 4 8

Show your work to receive full credit.

5. Below are the midterm exam scores of 23 introductory statistics students, sorted from low to high:

46, 47, 61, 74, 75, 77, 83, 84, 85, 86, 88, 88, 90, 90, 90, 93, 94, 94, 95, 95, 96, 96, 98

The five-number summary is: Min = 46, Q1 = 80, Median = 88, Q3 = 94, Max = 98.

- (5a) (2pts) What value(s) would be identified as outliers based on the 1.5 IQR rule? **Select only ONE answer.**

(i) none (ii) 46 (iii) 46, 98 (vi) 98 (v) 46, 47 (v) 46, 47, 98

- (5b) (1pt) Where does the lower whisker of the boxplot for the 23 scores above extend to? **Select only ONE answer.**

(i) 46 (ii) 47 (iii) 59 (iv) 61 (v) 67 (vi) 74

6. For a period of 5 years, physicians at McGill University Health Center followed about 5000 adults over the age of 50. The researchers were investigating whether people taking a certain class of antidepressants (SSRIs) might be at greater risk of bone fractures. Their observations are summarized in the table on the right.

Experienced Fracture?	Taking SSRI?		Total
	Yes	No	
Yes	14	244	258
No	123	4627	4750
Total	137	4871	5008

- (6a) (2pts) Which pair of numbers should we examine if we want to know whether people taking SSRI might be at greater risk of bone fractures? **Select only ONE answer.**

(i) 14 vs 244 (ii) 14/5008 vs 244/5008 (iii) 14/137 vs 244/4871
(iv) 14/258 vs 244/258 (v) 258/5008 vs 4750/5008

- (6b) (2pts) Can we conclude that taking SSRI increases the risk of bone fractures? **Select only ONE answer.**

(i) No, because the study was not blinded
(ii) No, because the 5008 subjects were not randomly sampled from the population of interest
(iii) No, because the subjects were not randomized to take SSRI
(iv) Yes, because the sample size 5008 was large
(v) Yes, because the study is a comparative experiment

7. (2pts) Suppose that 35% of the registered voters in a state are registered as Republicans, 40% as Democrats, and 25% as Independents. A newspaper wants to select a sample of 1000 registered voters to predict the outcome of the next election. If they randomly select 350 Republicans, randomly select 400 Democrats, and randomly select 250 Independents, what kind of sampling method is used here?

Select only ONE answer.

(i) simple random sampling (ii) stratified sampling
(iii) clustered sampling (iv) multistage sampling
(v) convenience sampling (vi) voluntary response sampling

8. (4pts) After menopause, some women take supplemental estrogen. There is some concern that if these women also drink alcohol, their estrogen levels will rise too high. Nineteen (19) volunteers who were receiving supplemental estrogen were randomly divided into 2 groups, as were 20 other volunteers not on estrogen. In each case, one group drank an alcoholic beverage, the other a nonalcoholic beverage. An hour later, everyone's estrogen level was checked. Only those on supplemental estrogen who drank alcohol showed a marked increase. For each of the following statements about the study, determine whether it is true or false. No explanation is required.

- (8a) T or F: This study blocked on whether subjects drank alcohol

- (8b) T or F: We cannot make a causal conclusion on the effect of supplemental estrogen on women's estrogen levels as the women were not randomized to take supplemental estrogen
- (8c) T or F: We can generalize the conclusion of the study as the subjects were randomized to drink an alcoholic or a nonalcoholic beverage
- (8d) T or F: This study used stratified sampling
9. (4pts) In some jurisdictions, there are “pretrial conferences,” where the judge confers with the opposing lawyers to settle the case or at least to define the issues before trial. Observational data suggests that pretrial conferences promote settlements and speed up trials, but there were doubts.
- In New Jersey courts, pretrial conferences were mandatory. However, an experiment was done in 7 counties. During a 6-month period, 2954 personal injury cases (mainly automobile accidents) were assigned at random to treatment or control. For the 1495 control cases (group A), pretrial conferences remained mandatory. For the 1459 treatment cases, the conferences were made optional—either lawyer could request one. Among the treatment cases, 701 opted for a pretrial conference (group C), and 758 did not (group B).
- To check whether pretrial conferences promote settlements and speed up trials, which two groups should the investigator compare? **Explain your choice.**
- (i) A v.s. B (ii) B v.s. C (iii) A v.s. (B+C) (iv) (A+C) v.s. B
- (v) Any of the above works because this is a randomized controlled experiment
10. (4pts) A local news agency conducted a survey about unemployment by randomly dialing phone numbers until they had gathered responses from 1000 adults in their state. In the survey, 19% of those who responded said they were not currently employed. In reality, only 6% of the adults in the state were not currently employed. How do you explain the difference in the two percentages?
11. (4pts) Traffic checks on a certain section of highway suggest that 40% of drivers are speeding there. Two vehicles in a row are selected at random at that section of highway. For each of the following statements, determine whether it is TRUE or FALSE and **explain briefly**.
- (11a) The probability that at least one of the two vehicles in a row is speeding is $0.4 + 0.4$
- (11b) The probability that the two vehicles in a row are both speeding is 0.4×0.4
12. The 2018 General Social Survey (GSS) interviewed a national sample of American adults and found that 47% of survey respondents have a pet dog, 25% have a pet cat, and 14% of survey respondents had both a dog and a cat. For all the parts below, please **show your calculation**.
- (12a) (2pts) What percentage of survey respondents have either a pet dog or a pet cat or both?
- (12b) (2pts) What percentage of survey respondents have no pet dog nor pet cat?
- (12c) (3pts) Among the respondents that have no pet cat, what percentage of them have a pet dog?
13. Leah is flying from Boston to Denver with a connection in Chicago. The probability her first flight leaves on time is 0.8. If the flight is on time, the probability that her luggage will make the connecting flight in Chicago and arrive in Denver is 0.9, but if the first flight is delayed, the probability that the luggage will make it is only 0.6.
- (13a) (3pts) What is the probability that her luggage arrives in Denver with her? Show your work.
- (13b) (3pts) Suppose you pick her up at the Denver airport, and her luggage is not there. What is the probability that Leah's first flight was delayed? Show your work.

- (13c) (2pts) Let A be the event that “the first flight leaves on time”, and B be the event that “the luggage makes the connection and arrives in Denver.” Are the two events A and B independent? Explain briefly.
14. You are dealt a hand of 3 cards, one at a time, without replacement. Recall that there are 4 aces and 13 hearts in a deck of 52 poker cards. Three cards are drawn at random WITHOUT replacement from a deck of poker cards.
- (14a) (3pts) Find the probability that you get no aces. **Show your calculation.**
- (14b) (2pts) Find the probability that the third card is an ace given that the first two cards are not aces.
15. The state of Illinois has several state-wide lottery options. One is the *Pick 3* game in which you pick one of the 1000 three-digit numbers between 000 and 999. The lottery selects a three-digit number at random. It takes \$1 to buy a *Pick 3* ticket. You win \$500 if your number is selected and nothing (\$0) otherwise. The distribution of the net profit X from a *Pick 3* ticket is hence:

Net Profit X	−\$1	\$499
Probability	0.999	0.001

For all the parts below, please **show your calculation.**

- (15a) (2pts) What is the expected value of X ?
- (15b) (3pts) What is the standard deviation of X ?
- (15c) (2pts) *Pick 3* lottery draws a three-digit number twice a day, once during the day and once in the evening. The drawings are independent of each other. So gamblers can make two independent bets a day. If a gambler buys two *Pick 3* tickets everyday for a year, one for the midday draw and one for the evening draw, (i.e., he makes $365 \times 2 = 730$ independent bets in total), what is the expected value of the total net profit the gambler can get from the 730 *Pick 3* tickets?
- (15d) (3pts) What is the standard deviation of the total net profit the gambler from the 730 *Pick 3* tickets?
16. The time it takes a driver to react to the brake lights on a decelerating vehicle is critical in avoiding rear-end collisions. The article “*Fast-Rise Brake Lamp as a Collision-Prevention Device*” (Ergonomics, 1993: 391–95) suggests that reaction time for an in-traffic response to a brake signal from standard brake lights can be modeled with a normal distribution having a mean value 1.25 s and an SD of 0.46 s.
- (16a) (3pts) What is the probability that the reaction time is between 1.00 and 1.75 s? If you calculate using R, please show (or write down) the R code you used.
- (16b) (3pts) What is the 99th percentile of the reaction time? That is, find the length of time that is long enough for at least 99% of the drivers to react.
If you calculate using R, please show (or write down) the R code you used.
17. The game of roulette involves spinning a wheel with 38 slots: 18 red, 18 black, and 2 green. A ball is spun onto the wheel and will eventually land in a slot, where each slot has an equal chance of capturing the ball. One popular bet is that it will stop on a red slot; such a bet has an $18/38$ chance of winning. Suppose a gambler bets on red in 6 different spins.
- (17a) (3pts) Find the probability the gambler wins the first 3 spins but loses the next 3 spins.
- (17b) (3pts) Find the probability the gambler wins exactly 3 of the 6 spins.
- (17c) (3pts) Find the probability the gambler wins at least 3 of the 6 spins.

For all 3 parts above, please **show your calculation.**

18. (3pts) A box contains 8 red marbles and 3 green ones. Six draws are made at random without replacement. True or false: the probability that the 3 green marbles are drawn equals

$$\frac{6!}{3!3!} \left(\frac{8}{11}\right)^3 \left(\frac{3}{11}\right)^3$$

Explain briefly.