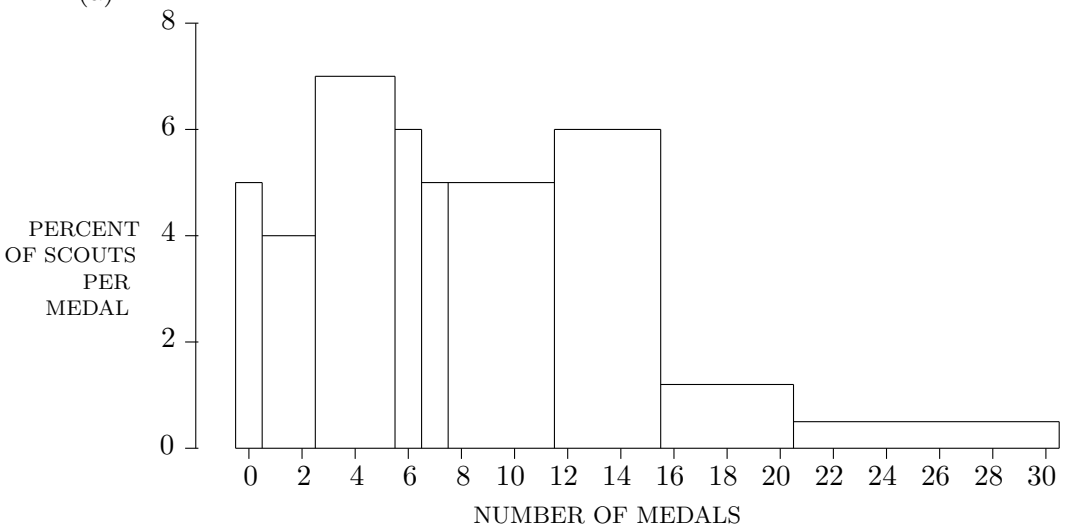
Answers to the Sample Final for Spring 2022

Math 125, Exam Time: May 20, 2022, from 3 to 6 p.m.

1. (a)



(b) i. A.
ii. C.
iii. B.
iv. D.
v. C.
vi. D.
vii. B.

Calculation Example for 3–5 medals.

42 scouts out of 200 total, so 21%. That's the area of the block: 21% of the population, a percentage. The base of the block is three medals. To get the height, divide the are by the base and get 7% per medal. *Be very careful about the units.* The endpoints in a discrete counting case require $\pm\frac{1}{2}$, so it will come out to 2.5 to 5.5 medals.

2. (a) 8.85%.
(b) about 5%.

The calculation for part (a).

The standard units for 700 are $\frac{700-538}{120} = 1.35$.

The area between -1.35 and 1.35 on the normal curve is 82.30%.

Subtract that from 100% and then divide by 2 to get the area of one tail. It is 8.85%.

3. The average of all men in the sample that were 66 inches tall is estimated to be 143 pounds. Work: 66 inches is 4 inches, or $4/3 \approx 1.33$ SDs below average. The estimated weight is below the average weight by $r \times 1.33 = 0.47 \times 1.33 \approx 0.625$ SDs. This is $0.625 \times 30 \approx 19$ pounds. This man is a little lighter than the estimated average of all the men, so he should be a little lighter than the average of all the other men.

4. In a run of one SD of x, the regression line rises

$$r \times \text{SD of } y.$$

The slope is

$$\text{rise/run} = 0.60 \times 20/10 \;=\; 1.2 \;\; \text{final points per midterm point.}$$

The intercept is

$$55 \;-\; 1.2 \times 70 \;=\; -29.$$

[Or solve the equation

$$y = mx + b$$

for b, using the average values of 70 and 55 for x and y.

We get

$$55 = 1.2(70) + b,$$

so

$$b = 55 - 1.2(70) = -29,$$

as before.]

So the equation is

$$\text{predicted final score} \;=\; 1.2 \times \text{midterm score} \;-\; 29\,.$$

5. $\text{weight} \;=\; 6 \times \text{height} \;-\; 240$ pounds.

The work is:

find the slope by taking r times (avg y)/(avg x). That gives 0.40 times 45/3 = 6 pounds per inch.

Now plug in the point of averages to find b: $180 = 6 \times 70 - b$ and $b = -240$.

Write the equation $y = mx + b$ as $y = 6x - 240$.

6. (a) The chances of A are 3/5 or 0.60. The chances of B are 1/2 or 0.50.

   (b) The chances of B are 1/2. The chances of B given A are found to be 3/6, which equals 1/2. Since the chances of B given A are the same as the unconditional chances of B, A and B are determined to be independent.

   Note: the calculation of B given A looks at the six numbers greater than 4 and asks how many are even? That is three of them: 6, 8, 10.

   (c) Finding the chance that a ticket selected at random falls into at least one of the 2 categories: over 4 or even.

      i. No; A and B are not mutually exclusive. Both could happen together, as for example if the 8 was drawn. The addition rule does not apply.

      ii. Look at the opposite event: that neither A or B occurs.
      Since A and B are independent, not A and not B will also be independent (proof omitted).
      So both not happening is the product of each not happening, which is 2/5 times 1/2 = 1/5.
      That's the opposite of what we wanted, so the answer to the problem is 4/5.

      This is a rather complicated way of solving the problem, but it teaches us about the rules of prbability.
      A simpler solution would just be by enumeration. There are 10 tickets. Getting either ticket greater than 4 or an even-numbered ticket is just the list of 5,6,7,8,9,10 with all even numbers that are not on the list so far being added to it. The missing even numbers are the 2 and the 4. The new list is 2,4,5,6,7,8,9,10 and that is 8/10=4/5, in agreement with the earlier result.

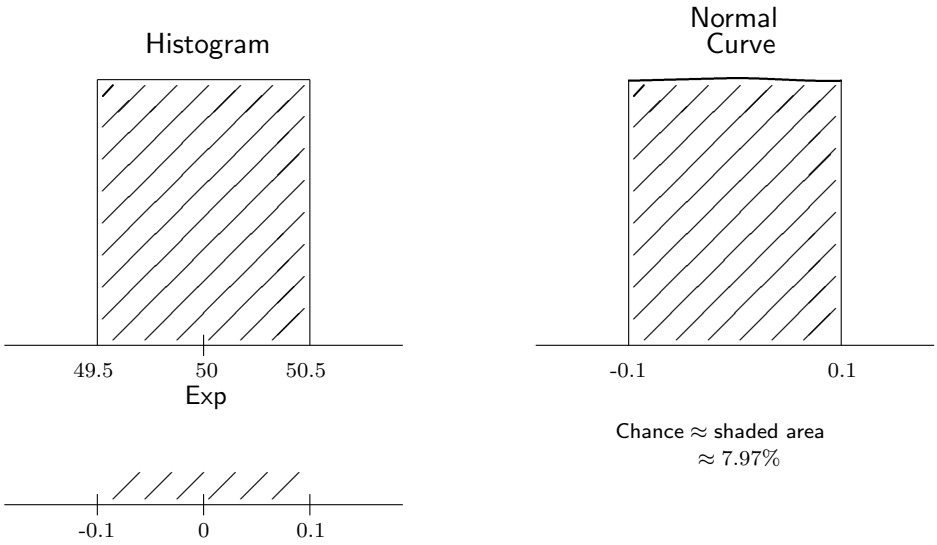
7. The chance is

$$\frac{9!}{3!6!}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^6 \approx 0.130238 \approx 13\%.$$

As an exact fraction, the answer is 109375/839808.

8. The expected number of heads is 50; the SE is 5. Since the SD of a zero-one box with one 0 and one 1 is 0.50, and the SE is the product of the square root of the number of draws and the SE, we get $\sqrt{100} \times 0.50 = 10 \times 0.50 = 5$.

We seek the area of the rectangle over 50, and will need to find the standard units for its endpoints, 49.5 and 50.5. That gives us $\frac{49.5-50}{5} = -0.1$ and $\frac{50.5-50}{5} = 0.1$. (The formula: std. units for the sum of draws = $\frac{\text{endpoint} - \text{EV of the sum}}{\text{SE for the sum}}$.)



Comment: the exact chance is $\frac{100!}{50!\times 50!}(.5)^{50}(.5)^{50} \approx 7.96\%$.

When doing the approximation, the curve doesn't curve much. The area under the normal curve is nearly rectangular, so there's not much error.

9. The box has millions of tickets, one for each 17-year-old in school that year. Tickets are marked 1 for those who knew that Chaucer wrote *The Canterbury Tales*, and 0 for the others. The data are like 6,000 draws from the box, and the number of students in the sample who knew the answer is like the sum of the draws. The fraction of 1's in the box can be estimated from the sample as 0.361. On this basis, the SD of the box is estimated as $\sqrt{0.361 \times 0.639} \approx 0.48$. The SE for the number of students in the sample who know the answer is estimated as $\sqrt{6{,}000} \times 0.48 \approx 37$. The SE for the percentage is $37/6000 \times 100\%$, which is about 0.6 of 1%. The percentage of students in the population who know the answer is estimated as 36.1%, give or take 0.6 of 1% or so. The 95%-confidence interval is 36.1% ± 1.2%.

10. (a) The percentage of Democrats in the sample is

$$\frac{811}{1440} \times 100\% \approx 56.319\%$$

The estimate: about 56.319% of the eligible voters are Democrats. For the standard error, a box model is needed. There is one ticket in the box for each eligible voter in the city, making 240,000 tickets in all. There are 1,440 draws, corresponding to the sample size of 1,440. This problem involves classifying people (Democrat or not) and counting, so each ticket is marked 1 or 0. It is Democrats that are being counted, so the tickets corresponding to Democrats are marked 1, the others are marked 0. There are 1,440 draws made at random from the box. The data are like the draws, and the number of Democrats in the sample is like the sum of the draws. That completes the model.

The fraction of 1's in the box (translation–the fraction of Democrats among the 240,000 eligible voters) is unknown, but can be estimated by 0.56319, the fraction of Democrats in the sample. Similarly, the fraction of 0's in the box is estimated as 0.43681. So the SD of the box is estimated by the bootstrap method as $\sqrt{0.56319 \times 0.43681} \approx 0.4960$. The SE for the percentage of Democrats in the sample is estimated as

$$\frac{0.4960}{\sqrt{1{,}440}} \times 100\% = 1.307\%.$$

In other words, the percentage of Democrats in the sample is likely to be off the percentage of Democrats in the population by 1.307 percentage points or so.

(b) observed (c) estimated from the data as

(d) A 95%-confidence interval for the percentage of Democrats among all 240,000 eligible voters is

$$56.319\% \pm 2 \times 1.307\%.$$

That is the answer. We can be about 95% confident that between 53.7% and 58.9% of the eligible voters in this city are Democrats.

Comment. In this problem, the compostion of the box has to be estimated from the data. You reason backward, from the draws to the box.

11. (a) True: the SE is estimated from the sample data, as on page 416.

(b) False. There is no such thing as a 95%-confidence interval for the *sample average*; you *know* the sample average. It's the population average that you have to worry about (pages 385–386).

(c) True.

(d) False, This confuses the SD with the SE. And it's ridiculous in the first place, because a household must have a whole number of persons (1, or 2, or 3, and so forth). The range 2.16 to 2.44 is impossible for any particular household, let alone 95% of them; although this range is fine for the average of all the households.

(e) False. For instance, if household size followed the normal curve, there would be many households with a negative number of occupants; we're not ready for that. There was a long right tail.

(f) False. See pages 411 and 418–419. Even though household size does not follow the normal curve, you can still use the normal curve to approximate the probability histogram for the sample average.

12. Rolling the die is like drawing at random with replacement 6000 times from a 0–1 box, with 0 = non-threes and 1 = threes. The fraction of 1's in the box is unknown.

Null hypothesis: this fraction equals 1/6.

Alternative: the fraction is bigger than 1/6. (We observe that 963 is somewhat smaller than 1,000, the expected value.)

The number of aces (one-spots) is like the sum of the draws.

$z = 1.26$, $P \approx 11$.

Calculation: The average of the box = 1/6 and the SD of box = $\sqrt{1/6 \times 5/6} = \sqrt{5}/6 \approx 0.372678$.

So the EV sum $= 6{,}000 \times 1/6 = 1000$ and the SE sum $= \sqrt{6{,}000} \times \sqrt{5}/6 = \sqrt{30000}/6 = 173/6 = 28.87$.

Using decimals, the SE for the sum $= \sqrt{6000} \times 0.372678 \approx 28.87$.

Then $z = \dfrac{\text{observed sum - EV sum}}{\text{SE sum}} = \dfrac{963.5 - 1000}{28.87} = 1.26.$

Ignoring the continuity correction would give $(963 - 1000)/28.87 = 1.28$, and hardly any difference in P.

To get P, the area of the right tail, look up 1.25 and subtract that 78.87% from 100% to get 21.13%. Then divide by 2 to get 10.56%.

This looks like chance variation. That large sigificance level P (more than 5%) makes an explanation of the null with chance error reasonable.

It looks like the die is fair.

13. Make a χ^2-test. The null hypothesis says that the probability of each digit is 1/10 (that is, that the random-number generator is fair). The alternative says that the random-number generator is not fair. The expected frequencies (the sums of various zero-one boxes for the ten digits) are found by multiplying the averages of the boxes (always 1/10) by the number of draws—400—to get 40. Or simply observe that for a fair generator, you expect the same number of each digit, 40 each.

    $\chi^2 = 6^2/40 + 1^2/40 + 8^2/40 + 0^2/40 + 14^2/40 + 3^2/40 + 4^2/40 + 6^2/40 + 5^2/40 + 9^2/40 = 11.6$ on 9 degrees of freedom, $P \approx 25\%$, a good fit.

14. The argument is not valid.

    The association between a 96 on the midterm and a 78 on the final does not signify an important connection between these two test scores. It is only a casual relationship.

    Once either the midterm grade or the final grade is specified it becomes the given. From this there is derived an estimate of the other grade that begins with a point of focus that has both grades the same. Since the association between the two grades is not perfect, this prediction is subject to chance variability and must be recalculated using the regression method, bringing the estimate closer to the mean of the predicted score. The strength of the association, represented by the correlation coefficient, will decide how close to the starting point of an equal test score will be the prediction used.

    Once the other test score becomes the given, it tends to be higher than the predicted. The 96 on the midterm leading to a 78 on the final is a result of the midterm being exactly known and the final being subject to uncertainty. If the final score is known to be 92 points, the midterm is unknown and subject to chance variation. Its score must be estimated at less than 92 points using the regression method of Galton.

    It is not a firm pairing of 96 points on the midterm with 78 points on the final. Rather it must be presented as the mapping of 96 points given on midterm to 78 points predicted on the final. The reason that score on the final was only 78 points was not becuase the midterm is generally higher. In fact the average of each is 60 points.

    The regression calculation was misinterpreted in the argument presented. A flawed assumption led to an invalid conclusion.

    Another way to look at the problem is to note that all of the students who got 96 on the midterm got about 78 on the final. If there were no other students who got 78 on the final, it might be reasonable to think that the 78 midterm predicts a 96 final. However, the students who got 96 on the midterm are only some of the students who got 78 on the final. Since there is a larger number with 78 on the final than there is with 96 on the midterm, there are surely others in the 78 group. There is no reason to thnnk that the others got 96 on the midterm. For the small group who got 96 on the midterm, the midterm score was higher than the final score, but for the others: who knows?

    After this argument is set aside, it becomes rather simple to use regression to decide that a 92 on the final must predict a lower score on the midterm.

    This was a classic example of the regression fallacy.