Study guide for first midterm exam - solutions

Joshua Loftus

Practice questions with solutions

1. A certain company sells a supplement for cold relief. To prove the supplement is effective, they commission a lab to do a study. Participants suffering from common cold symptoms are recruited into the study and given the supplement for free. They are asked to record if their symptoms improved after taking the supplement. At the conclusion of the study, about 70% of the participants said their symptoms improved after taking the supplement, another 25% didn't respond to the follow up, and 5% said their symptoms got worse. Do you think the study was effective at proving the supplement relieves symptoms? Why or why not?

The study **does not have a control group**. It is possible the study only demonstrated the placebo effect.

2. Use the definition $\mathbf{Var}(X) = E[(X - E[X])^2]$ to show $\mathbf{Var}(X) = E[X^2] - E[X]^2$.

Expand: $(X - E[X])^2 = X^2 - 2X \cdot E[X] + E[X]^2$. Remember that $E[X]$ is a constant, so $E[2X \cdot E[X]] = E[X] \cdot E[2X] = 2E[X]^2$. Plugging this all in to the definition,

$$\text{Var}(X) = E[X^2 - 2X \cdot E[X] + E[X]^2] = E[X^2 - 2E[X]^2 + E[X]^2] = E[X^2 - E[X]^2] = E[X^2] - E[X]^2$$

3. Fact: If X and Y are independent then $E[XY] = E[X]E[Y]$. Use this fact to show that if X and Y are independent then $\mathbf{Var}(X + Y) = \mathbf{Var}(X) + \mathbf{Var}(Y)$.

Expand:

$$(X + Y - E[X + Y])^2 = (X - E[X] + Y - E[Y])^2 = (X - E[X])^2 + 2(X - E[X])(Y - E[Y]) + (Y - E[Y])^2$$

Taking expectation, the first and last terms give $\text{Var}(X)$ and $\text{Var}(Y)$ respectively. We need to show the expectation of the middle term is zero. Expand the middle term, ignoring the 2 for a moment:

$$(X - E[X])(Y - E[Y]) = XY - XE[Y] - E[X]Y + E[X]E[Y]$$

Taking expectation:

$$E[XY - XE[Y] - E[X]Y + E[X]E[Y]] = E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] = E[XY] - E[X]E[Y] = 0$$

The factor of 2 in front doesn't matter, since $2 \cdot 0 = 0$.

4. Suppose $U \sim \mathbf{Bin}(n, p)$ and $V \sim \mathbf{Bin}(m, p)$ are independent Binomials. What are $E[U + V]$ and $\mathbf{Var}(U + V)$?

$E[U + V] = E[U] + E[V] = np + mp = (n + m)p$, by linearity.

$\text{Var}(U + V) = \text{Var}(U) + \text{Var}(V)$ by linearity because U and V are independent. This equals $np(1 - p) + mp(1 - p) = (n + m)p(1 - p)$

5. According to the Department of Health and Human Services, African American children are 4 times more likely to be admitted to the hospital for asthma compared to non-Hispanic white children. A certain politician concludes from this that asthma must have a genetic explanation related to race, therefore we cannot expect any public health policy to reduce the difference in health outcomes. Leaving aside *ethical* questions, can you find any *statistical* errors in his argument?

There could be many possible **confounding factors**. One example: pollution can make the symptoms of asthma worse, and there are racial discrepancies in exposure to pollution or other environmental factors due to this country's history and ongoing structural racism.

6. A credit card company's algorithm uses *big data* and *machine learning* to detect fraudulent transactions and automatically warn customers that their card may have been stolen. The algorithm is very accurate: $P(\textbf{detection}|\textbf{fraud}) = 0.999$, but it sometimes makes errors $P(\textbf{detection}|\textbf{normal activity}) = 0.01$. You receive a message that the bank has detected fraudulent activity on your account. You wonder: what is the probability that your account has experienced fraudulent activity given that the bank's algorithm says it was detected. Could you calculate this probability, or do you need more information?

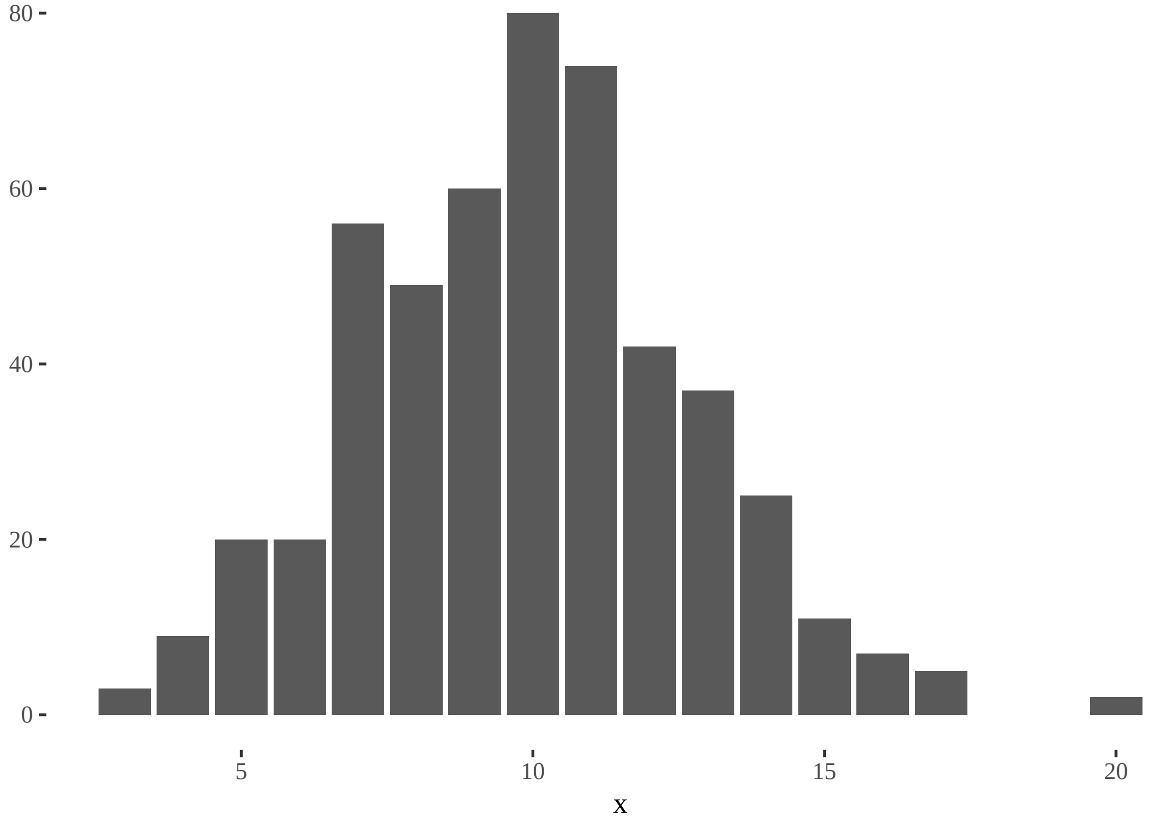
We recognize this as a Bayes' rule type of problem, but there is **not enough information**. We need to know the underlying rates, like $P(\text{fraud})$. If we know this, we could use the law of total probability to compute $P(\text{detection})$, and then we would have all the pieces needed to use Bayes' formula.

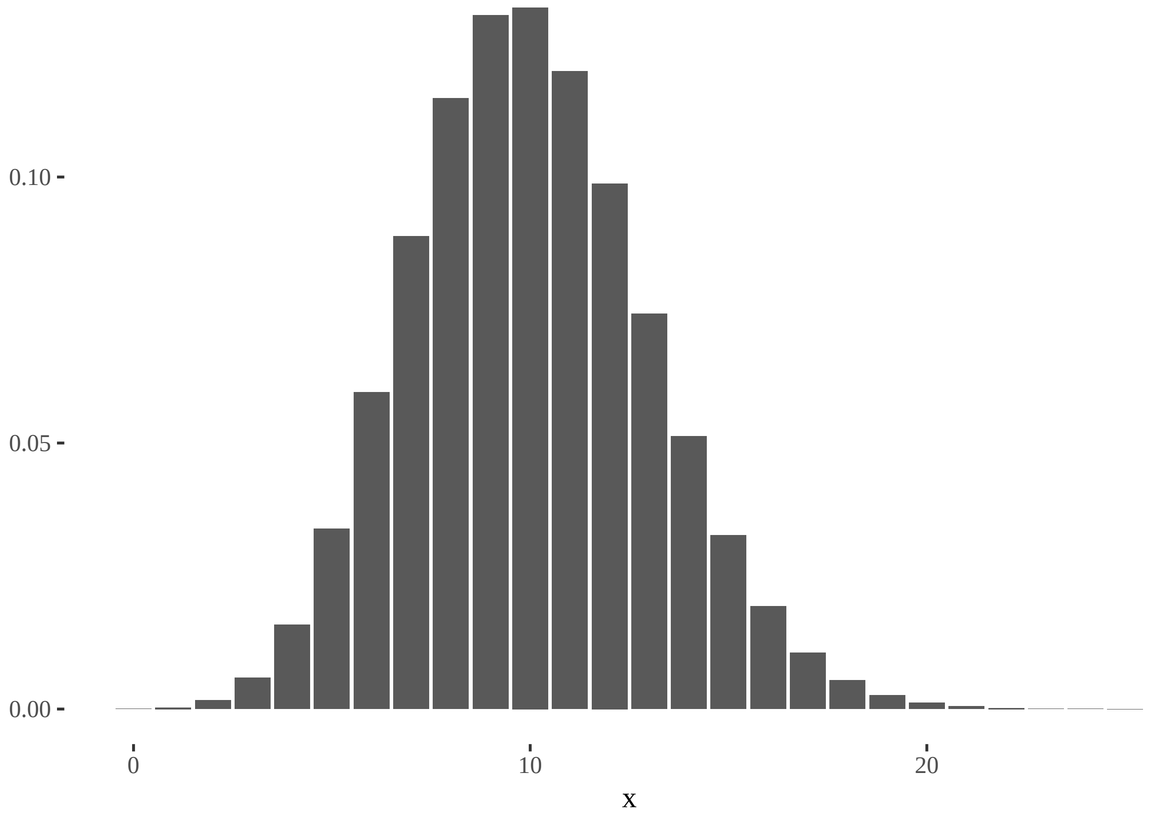
7. You and your friend are sharing a pizza. There are 8 slices, half are cheese and half are pepperoni. As you are eating, each time you take a new slice you pick it randomly. You and your friend each finish half of the pizza. Is Bin$(4, 1/2)$ a good model for the number of pepperoni slices that you eat? Explain.

Binomial is **not a good model** for this scenario because the **trials are not independent**. If we know the first trial was a success, then we know the next trial has a lower probability of success, because this is **sampling without replacement**.

8. One of the following plots is the probability distribution function (pdf) of a Binomial random variable, and one is a histogram of random samples from the same Binomial distribution. Which one is the pdf, first or second? Can you guess the expected value? Do you think $\sigma^2 = 4$ is a good guess for the variance, or too high/low?

The pdf is the second plot. The expected value should be near the center, which appears to be about 10. The bulk of the distribution should be between 10 plus or minus σ, so it looks like $\sigma = 2$ is too low of a guess, but it's not very easy to tell.


80
60
40
20
0
5
10
15
20
x


0.10
0.05
0.00
0
10
20
x

9. A dice game: roll a regular 6-sided die. If it lands on 6, you win. If it lands on 1, you lose. If it lands anywhere in between, you roll the dice a second time, and if the sum of the two rolls is greater than 6 you win. What is the probability that you win *on the second roll*? What is the probability that you win on any roll?

The key here is to use the multiplication rule: $P(E_1 \cap E_2) = P(E_2|E_1)P(E_1)$.

Let D_1 be the outcome of the first die, and D_2 the second. Then the probability of winning on the second roll is:

$$P(2 \leq D_1 \leq 5 \text{ and } D_1 + D_2 > 6) = P(D_1 + D_2 > 6 | 2 \leq D_1 \leq 5) P(2 \leq D_1 \leq 5)$$

By counting, $P(2 \leq D_1 \leq 5) = 4/6 = 2/3$

Conditional on $2 \leq D_1 \leq 5$, there are $4 \cdot 6 = 24$ possibilities for the values of both dice (instead of $6 \cdot 6$, because 1 and 6 are no longer possible for the first die). Out of these 24, we need to count which ones result in a win. We'll do this by listing the 4 possible values of the first die and then which values for the second die would result in a sum greater than 6.

D_1	D_2
2	5, 6
3	4, 5, 6
4	3, 4, 5, 6
5	2, 3, 4, 5, 6

There are $2 + 3 + 4 + 5 = 14$ outcomes where the sum is greater than 6.

$P(D_1 + D_2 > 6 | 2 \leq D_1 \leq 5) = 14/24 = 7/12$

Multiplying these together gives the probability of winning on the second roll as $(7/12)(2/3) = 7/18$.

Winning on the second roll is disjoint from winning on the first, so probability of winning overall is the sum of the two probabilities (addition rule for disjoint events). Winning on the first roll happens with probability $1/6$, so the overall winning probability is $1/6 + 7/18 = 3/18 + 7/18 = 10/18$.