

Solutions to In-Class Problems — Week 14, Fri

Problem 1. In this problem you will check a proof of the following:

Theorem 1.1. Let A_1, A_2, \dots, A_n be independent events, and let T be the number of these events that occur. The probability that none of the events occur is at most $e^{-E[T]}$.

To prove Theorem 1.1, note that

$$T = T_1 + T_2 + \dots + T_n, \tag{1}$$

where T_i is the indicator variable for the event A_i . Also, remember that

$$1 + x \leq e^x \tag{2}$$

for all x and

$$1 + x \approx e^x, \tag{3}$$

for $0 \leq x \leq 1$. Both (2) and (3) follow from the Taylor's expansion of e^x .

(a) Justify each line in the following derivation:

Solution. *Proof.*

$$\begin{aligned} \Pr\{T = 0\} &= \overline{A_1 \cup A_2 \cup \dots \cup A_n} && \text{(def. of } T) \\ &= \Pr\{\overline{A_1} \cap \overline{A_2} \cap \dots \cap \overline{A_n}\} && \text{(De Morgan's law)} \\ &= \prod_{i=1}^n \Pr\{\overline{A_i}\} && \text{(mutual independence of } A_i\text{'s)} \\ &= \prod_{i=1}^n (1 - \Pr\{A_i\}) && \text{(complement rule)} \\ &\leq \prod_{i=1}^n e^{-\Pr\{A_i\}} && \text{(by (2))} \\ &= e^{-\sum_{i=1}^n \Pr\{A_i\}} && \text{(exponent algebra)} \\ &= e^{-\sum_{i=1}^n E[T_i]} && \text{(expectation of indicator variable)} \\ &= e^{-E[T]}. && \text{((1) \& linearity of expectation)} \end{aligned}$$

□



Two special cases of Theorem 1.1 are worth singling out because they come up all the time.

Corollary 1.2. *Suppose an event has probability $1/m$. Then the probability that the event will occur at least once in n independent trials is approximately $1 - 1/e \approx 63\%$. There is a 50% chance the event will occur in $n = \log 2m \approx 0.69m$ trials.*

(b) Prove Corollary 1.2.

Solution. In this case, $\Pr \{A_i\} = 1/m$ for $1 \leq i \leq n$ and

$$E[\text{\# occurrences}] = n \frac{1}{m} = \frac{n}{m}.$$

So by Theorem 1.1,

$$\Pr \{\text{no occurrence}\} \leq e^{-(n/m)},$$

and therefore

$$\Pr \{\text{at least one occurrence}\} \geq 1 - e^{-(n/m)}. \quad (4)$$

In fact, it follows from (3), that the \geq in (4) is an approximate equality.

So if the number, n of trials is m , we have

$$\Pr \{\text{at least one occurrence}\} \approx 1 - e^{-(m/m)} = 1 - \frac{1}{e}.$$

If we want

$$1 - e^{-(n/m)} \approx \Pr \{\text{at least one occurrence}\} \approx \frac{1}{2},$$

then we need

$$e^{-(n/m)} \approx \frac{1}{2},$$

so taking log's we conclude

$$n \approx m \log 2.$$

■

Problem 2. The spider (remember her from the Tutor problem) is expecting guests and wants to catch 500 flies for her dinner. *Exactly* 100 flies pass by her web every hour. Exactly 60 of these flies are quite small and are caught with probability $1/6$ each. Exactly 40 of the flies are big and are caught with probability $3/4$ each. Assume all fly interceptions are mutually independent. Using

this information, the methods from lecture can show that the poor spider has only about 1 chance in 100,000 of catching 500 flies within 10 hours.

Ben Bitdiddle knows he can get the best estimate using the approximations to the binomial distribution developed in Notes 12. He reasons that since 60% of the flies are small and 40% are large, the probability that a random fly will be caught is $0.6(1/6) + 0.4(3/4) = 0.4$, so he will use the approximation for the binomial cumulative distribution function, $F_{1000,0.4}$, to bound the probability that the spider catches at least 500 flies in 10 hours.

As usual, Ben hasn't got it quite right.

(a) According to Ben's reasoning, what is the probability that the spider will catch all 1000 flies that show up during the 10 hours? Show that this is not equal to the actual probability.

Solution. According to Ben, the probability would be $(0.4)^{1000} = (2/5)^{1000}$. But the actual probability is $(1/6)^{600}(3/4)^{400}$, and we don't even need to evaluate these expressions to see that they must have different values. ■

(b) How would you explain to Ben what is wrong with his reasoning?

Solution. Ben's reasoning would be ok if the event that a fly is large is independent of whether the next fly is large. That's not the case here: after the 99th fly in the first hour, we can predict whether the 100th fly will be large or small. The number, R , of flies caught in 10 hours is actually the sum of a random variable with distribution $f_{600,1/6}$ and another variable with distribution $f_{400,3/4}$, and R not only disagrees with Ben's model on the probability that all the flies will be caught, it does not even have a binomial distribution. ■

(c) What would the Markov bound be on the probability that the spider will catch her quota of 500 flies?

Solution. The expected number of flies caught is $600(1/6) + 400(3/4) = 400$, so by Markov, $\Pr\{R \geq 500\} \leq 400/500 = 0.8$. ■

(d) What would the Chebyshev bound be on the probability that the spider will catch her quota of 500 flies?

Solution. The variance is $600(1/6)(5/6) + 400(3/4)(1/4) = 1900/12 \approx 158$, so the Chebyshev bound is

$$\Pr\{R - 400 \geq 100\} \leq \Pr\{|R - 400| \geq 100\} \leq \frac{1900/12}{100^2} = 19/1200 \approx 1/64.$$

■

(e) What would the Chernoff bound be on the probability that the spider will catch her quota of 500 flies? (You can do this without a calculator knowing that $\ln 5/4 \approx 0.223$, $e^3 \approx 20$ and $\sqrt{e} \approx 1.6$.)

Solution.

$$\begin{aligned}
\Pr \{R \geq 500\} &= \Pr \{R \geq (5/4)400\} \\
&\leq \exp(-((5/4) \ln(5/4) - 5/4 + 1)400) \\
&= \exp(-(500 \ln(5/4) - 500 + 400)) \\
&\approx \exp(-(500(0.223) - 100)) \\
&= \exp(-(111.5 - 100)) \\
&= e^{-11.5} \\
&= \frac{\sqrt{e}}{(e^3)^4} \\
&\approx \frac{1.6}{20^4} \\
&= \frac{1}{100,000}.
\end{aligned}$$

■

(f) Ben argues that he made his mistake because the description of the spider's situation is absurd: knowing the *expected* number of flies per hour is one thing, but knowing the *exact* number is far-fetched.

Which of the bounds above will hold if all we know is that the *expected* number of small flies caught per hour is 10 and of large flies is 30?

Solution. We know the expectation is 400 flies in 10 hours, so Markov's bound will hold because it only depends on the expectation and the nonnegativity of the number of flies. In the case of the Chernoff bound, we also need to know that the number of flies is a sum of independent Bernoulli variables; we are no longer given this, so Chernoff does not apply. To apply Chebyshev we need the variance, which we aren't given.

Actually, to apply Chebyshev, all we need is a bound on the variance, and there is one given that the expectation is 400 and R is nonnegative. The maximum possible variance for a nonnegative distribution with mean 400 occurs for the two-valued variable taking values 0 and 800 with equal probability. In this case the variance is 400^2 . But plugging this value into the Chebyshev formula gives a bound greater than 1, which is useless. ■

Problem 3. Let R be the sum of a finite number of mutually independent Bernoulli variables. Let $\mu_R ::= E[R]$, and let σ_R be the deviation of R .

(a) Write formulas in terms of y , μ_R and σ_R for the Markov, Chebyshev, and Chernoff bounds on $\Pr \{R \geq y\}$, where $y \geq \mu_R$. *Hint:* To apply Chebyshev bounds, assume

$$\Pr \{R - \mu_R \geq x\} \approx \frac{\Pr \{|R - \mu_R| \geq x\}}{2} \quad (5)$$

for all $x \geq 0$.

Solution. • The Markov bound is μ_R/y , directly from (7).

- The Chebyshev bound is

$$\begin{aligned} \Pr\{R \geq y\} &= \Pr\{R - \mu_R \geq y - \mu_R\} \\ &\approx \frac{\Pr\{|R - \mu_R| \geq y - \mu_R\}}{2} && \text{(by (5))} \\ &\leq \frac{\sigma_R^2}{2(y - \mu_R)^2}. && (6) \end{aligned}$$

- The Chernoff bound is

$$\begin{aligned} \Pr\{R \geq y\} &= \Pr\left\{R \geq \frac{y}{\mu_R} \mu_R\right\} \\ &\leq \exp\left(-\left(\frac{y}{\mu_R} \ln \frac{y}{\mu_R} - \frac{y}{\mu_R} + 1\right) \mu_R\right) \\ &= \exp\left(-\left(y \ln \frac{y}{\mu_R} - y + \mu_R\right)\right). \end{aligned}$$

■

(b) Compare these bounds when R is a single unbiased Bernoulli variable and $y = 1$.

Solution. In this case $\mu_R = 1/2$ and $\sigma_R^2 = 1/4$, so

- Markov gives $\mu_R/y = 1/2$ which is exactly right,
- the bound from the Chebyshev result (6) is

$$\frac{\sigma_R^2}{2(y - \mu_R)^2} = \frac{1/4}{2(1 - (1/2))^2} = \frac{1}{2},$$

and so is also exactly right,

- the bound from the Chernoff result is

$$\exp(-(\ln(1/(1/2)) - 1 + 1/2)) = e^{1/2 - \ln 2} = \sqrt{e}/2 > 0.83,$$

and so is a large overestimate.

■

A Appendix

Theorem (Markov's Theorem). If R is a nonnegative random variable, then for all $x > 0$

$$\Pr\{R \geq x\} \leq \frac{\mathbb{E}[R]}{x}. \quad (7)$$

Theorem (Chebyshev). Let R be a random variable, and let x be a positive real number. Then

$$\Pr \{|R - \mathbf{E}[R]| \geq x\} \leq \frac{\mathbf{Var}[R]}{x^2}. \quad (8)$$

Theorem (Pairwise Independent Sampling). Let $S_n ::= \sum_{i=1}^n G_i$ where G_1, \dots, G_n are pairwise independent variables with the same mean, μ , and deviation, σ . Then

$$\Pr \left\{ \left| \frac{S_n}{n} - \mu \right| \geq x \right\} \leq \frac{1}{n} \left(\frac{\sigma}{x} \right)^2. \quad (9)$$

Theorem (Chernoff Bound). Let T_1, T_2, \dots, T_N be mutually independent Bernoulli variables, and let $T ::= T_1 + T_2 + \dots + T_N$. Then for all $c \geq 1$,

$$\Pr \{T \geq c \mathbf{E}[T]\} \leq \exp(-(c \ln c - c + 1) \mathbf{E}[T]). \quad (10)$$