

MIT OpenCourseWare
<http://ocw.mit.edu>

14.30 Introduction to Statistical Methods in Economics
Spring 2009

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Problem Set #2 - Solutions

14.30 - Intro. to Statistical Methods in Economics

Instructor: Konrad Menzel

Due: Tuesday, February 24, 2009

Question 1

Recall that a random variable X has the binomial distribution if

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x},$$

where n is the number of trials and p is the chance of success. For the following crudely-disguised questions about the binomial distribution, do the following: Define p , n , what the specific “trials” are, what “success” is. Then write down the relevant distribution and answer the specific question.

1. If 25 percent of the balls in a certain box are red, and if 15 balls are selected from the box at random, with replacement, what is the probability that more than four red balls will be obtained?

- Solution to 1: For this problem, we have that $p = .25$, $n = 15$. Each “trial” is the selection of a ball from the box, and each “success” is drawing a red ball.

The probability that more than four red balls will be obtained can be computed two ways. One way would be to compute the direct probability by summing up the probabilities of getting 5, 6, ..., or 15 balls. That’s the long way. I instead will compute it using the complementary event $\{k > 4\}^c = \{k \leq 4\}$ where k is the number of red balls drawn. So, this makes the problem a little simpler as I only have to compute 5 probabilities and then sum over them using the binomial formula:

$$\begin{aligned} P(K = k) &= \binom{n}{k} p^k (1 - p)^{n-k} \\ P(K = 0) &= \binom{15}{0} p^0 (1 - p)^{15-0} = .01336346 \\ P(K = 1) &= \binom{15}{1} p^1 (1 - p)^{15-1} = .06681731 \\ P(K = 2) &= \binom{15}{2} p^2 (1 - p)^{15-2} = .15590705 \\ P(K = 3) &= \binom{15}{3} p^3 (1 - p)^{15-3} = .22519907 \\ P(K = 4) &= \binom{15}{4} p^4 (1 - p)^{15-4} = .22519907 \end{aligned}$$

So, when we sum over the probability of getting $P(\{k \leq 4\}) = .68648594$. Thus, the complementary event's (the complementary event to the complementary event is what we're looking for) probability will be $P(\{k > 4\}) = 1 - P(\{k \leq 4\}) = 1 - .68648594 = .31351406$. So, instead of counting 11 cases, we only had to count 5.

2. (3 pts.) Suppose an economist is organizing a survey of American minimum wage workers, and is interested in understanding how many workers that earn the minimum wage are teenagers.¹ Suppose further that one out of every four minimum wage workers is a teenager. If the economist finds 80 minimum wage workers for his survey, what's the probability that he interviews exactly 14 teenagers? 35 teenagers? What's the probability that he gets at least 5 teenagers in his survey?

- Solution to 2: In this case, we have $p = 0.25$ again and $n = 80$. We again use the binomial formula to obtain the solution to the first question:

$$P(K = 14) = \binom{80}{14} .25^{14} (1 - .25)^{80-14} = 0.0319.$$

The second uses the same formula:

$$P(K = 35) = \binom{80}{35} .25^{35} (1 - .25)^{80-35} = 0.00011704.$$

3. (Bonus Question) A city has 5000 children, including 800 who have not been vaccinated for measles. Sixty-five of the city's children are enrolled in a day care center. Suppose the municipal health department sends a doctor and nurse to the day care center to immunize any child who has not already been vaccinated. Find a formula for the probability that exactly k of the children at the day care center have not been vaccinated. (Hint: This is not exactly a binomial distribution problem.)

- Solution to 3: In previous years, a similar problem was just approximated by the binomial distribution. It turns out that for large n , the binomial distribution gives a good approximation to problems like this one. Further, for large n and large k , the Normal (also known as the Gaussian) distribution also gives a very good approximation. I am going to solve this questions all three ways to give you a sense as to the exact answer.

Since the problem is actually a discrete problem with *sampling without replacement*, the binomial distribution is not exactly correct. If we had *sample with replacement*, then it would work just fine. This problem is an example of the hypergeometric distribution (Wikipedia: Hypergeometric Distribution). The hypergeometric distribution is typically couched in a "defective parts" problem where you have a batch of N parts, m of which are defective. You then sample n different parts without replacement and wish to know the probability of getting

¹In the debate over the minimum wage, one point that is always made is that a lot of workers on the minimum wage are middle-class teenagers. Since most of them are not supporting a family, the harms of an increase in the minimum wage outweigh the benefits. This point never convinces anyone.

exactly k defective units. In our case, $N = 5000$ children, $m = 800$ unvaccinated children, $n = 65$ children at the day care center, and k represents the number of unvaccinated children at the day care center. Without going into the details of its derivation, the hypergeometric distribution takes into account the change in conditional probabilities when you sample without replacement. Intuitively, though, you're trying to figure out how to split the 5000 children into two groups—the sampled and unsampled groups. Then, you're trying to figure out the probability that you have a particular k number of unvaccinated children in the sampled part. From Wikipedia, “The formula can be understood as follows: There are $\binom{N}{n}$ possible samples (without replacement). There are $\binom{m}{k}$ ways to obtain k defective objects and there are $\binom{N-m}{n-k}$ ways to fill out the rest of the sample with non-defective objects.” The formula follows:

$$f(k; N, m, n) = \frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}}$$

which would be, in our case:

$$f(k; 5000, 800, 65) = \frac{\binom{800}{k} \binom{5000-800}{65-k}}{\binom{5000}{65}}.$$

It is extremely interesting to compare how the different approximations perform. Here are two charts that show how well the approximations do for large N ($N = 5000$) and small N ($N = 125$):

N=5000, m=800, n=65				N=125, m=20, n=65			
k	Hypergeometric	Binomial	Normal	k	Hypergeometric	Binomial	Normal
0	0	0	0.0003	0	0	0	0.0003
1	0.0001	0.0001	0.0009	1	0	0.0001	0.0009
2	0.0009	0.0009	0.0025	2	0	0.0009	0.0025
3	0.0036	0.0035	0.006	3	0.0002	0.0036	0.006
4	0.0107	0.0104	0.0132	4	0.0014	0.0107	0.0132
5	0.0248	0.0243	0.0257	5	0.0061	0.0248	0.0257
6	0.0472	0.0467	0.0448	6	0.02	0.0472	0.0448
7	0.0758	0.0755	0.0698	7	0.0503	0.0758	0.0698
8	0.1047	0.1048	0.0969	8	0.0987	0.1047	0.0969
9	0.1263	0.1269	0.1202	9	0.1531	0.1263	0.1202
10	0.1347	0.1356	0.1331	10	0.1886	0.1347	0.1331
11	0.1283	0.1292	0.1316	11	0.1849	0.1283	0.1316
12	0.11	0.1106	0.1162	12	0.144	0.11	0.1162
13	0.0854	0.0856	0.0916	13	0.0886	0.0854	0.0916
14	0.0604	0.0603	0.0644	14	0.0427	0.0604	0.0644
15	0.0391	0.0389	0.0405	15	0.0158	0.0391	0.0405
16	0.0233	0.023	0.0227	16	0.0044	0.0233	0.0227
17	0.0128	0.0125	0.0114	17	0.0009	0.0128	0.0114
18	0.0065	0.0063	0.0051	18	0.0001	0.0065	0.0051
19	0.0031	0.0029	0.002	19	0	0.0031	0.002
20	0.0013	0.0013	0.0007	20	0	0.0013	0.0007
21	0.0005	0.0005	0.0002	21	0	0.0005	0.0002
22	0.0002	0.0002	0.0001	22	0	0.0002	0.0001
23	0.0001	0.0001	0	23	0	0.0001	0
24	0	0	0	24	0	0	0
25	0	0	0	25	0	0	0

So, the real question is, if I didn't care if you had shown your work, could you have just taken the easy route on this problem and used the binomial formula? I guess the answer depends on how many decimals you show! ;)

Question Two

Suppose you flip a weighted coin (probability of heads is p and probability of tails is $q = 1 - p$) n times.

1. What is the probability that you get a particular ordering of k heads and $n - k$ tails?

- Solution to 1: The probability of getting a particular ordering of k heads and $n - k$ tails would be $p^k q^{n-k}$.

2. What is the probability that you get k heads and $n - k$ tails?

- Solution to 2: From part 1, the probability of getting any particular ordering of k heads and $n - k$ tails would be $p^k q^{n-k}$, but we have to take into account the number of permutations that exist for getting k heads and $n - k$ tails. We can think about this as the “ALGEBRA” and “CALCULUS” problem, by simply labeling the heads as “H” and tails as “T” and then figure out how many different words we can write. This is pretty simple: $\frac{n!}{k!(n-k)!}$ which just so happens to be the same as the binomial coefficient, $\binom{n}{k}$. So, it turns out that we just have the binomial formula for this problem:

$$P(k \text{ heads}) = \binom{n}{k} p^k q^{n-k}.$$

3. Let X be the number of heads in n flips. What is the probability density function of X ?

- Solution to 3: Well, I just have to rewrite the answer to 2 slightly differently to change the notation to refer to the random variable, X and its PDF, $f(x; n)$:

$$P(X = x) = f(x; n) = \binom{n}{x} p^x (1 - p)^{n-x}.$$

4. What does this question have to do with the MIT Beaver’s football team question from Problem Set #1? Explain.

- Solution to 4: The MIT Beaver’s football team question was virtually the same problem, with a slight additional generalization, because we had three different outcomes. We could write this as two binomial trials (number of ties, then number of wins in remaining games) or we could write it using a slightly more general version of the binomial theorem which known as the multinomial theorem (Wikipedia: Multinomial Theorem). We can write these coefficients as: $\binom{n}{k_1, \dots, k_m} = \frac{n!}{k_1! \dots k_m!}$ and the distribution that we would use would be the multinomial distribution (Wikipedia: Multinomial Distribution):

$$P(K_1 = k_1, \dots, K_m = k_m; n) = \binom{n}{k_1, \dots, k_m} p_1^{k_1} \dots p_m^{k_m}.$$

Taking a closer look at the solution for the MIT Beavers question from Problem Set #1, it quickly becomes apparent that there are 3 possibilities with varying probabilities: Wins (p_W), Losses (p_L), and Ties (p_T), each with a specific k . Thus, the MIT Beavers question maps into the composition of two binomials, which is just the multinomial theorem.

I must highlight, however, that even if you knew nothing about the binomial theorem, the techniques of the multiplication rule and permutation counting rules was sufficient to answer that question. This is merely to help you learn about more general methods of dealing with these types of questions.

Suppose now that you have a hat with two coins, one weighted as above and one fair. You choose one at random and flip that one n times.

1. Let $Y \equiv$ the number of heads in n flips. What is the PDF of Y ?

- Solution to 1: This question is about a mixture distribution. We have two coins, each of which we use with probability $p = \frac{1}{2}$. So, the PDF will be that we use the fair coin's PDF half of the time and the unfair coin the other half of the time:

$$P(Y = y; n) = \underbrace{\frac{1}{2} \binom{n}{y} 0.5^y (1 - 0.5)^{n-y}}_{\text{Fair Coin}} + \underbrace{\frac{1}{2} \binom{n}{y} p^y (1 - p)^{n-y}}_{\text{Weighted Coin}}.$$

2. What is the probability that you chose the fair coin given that $Y = k$?

- Solution to 2: In order to deduce the answer to this question, we apply Bayes' rule again, but with PDFs. We would like to know $P(\text{Fair} | Y = k)$. We know what the denominator is in Bayes' rule from part 1 (just above):

$$P(Y = k) = P(Y = k | \text{Fair})P(\text{Fair}) + P(Y = k | \text{Weighted})P(\text{Weighted}).$$

We now just need to get the other parts explicitly: $P(\text{Fair}) = \frac{1}{2}$ and $P(Y = k | \text{Fair}) = \binom{n}{k} 0.5^k (1 - 0.5)^{n-k}$, the fair coin's PDF. So, we now use Bayes' Rule:

$$\begin{aligned} P(\text{Fair} | Y = k) &= \frac{P(Y = k | \text{Fair})P(\text{Fair})}{P(Y = k | \text{Fair})P(\text{Fair}) + P(Y = k | \text{Weighted})P(\text{Weighted})} \\ &= \frac{\frac{1}{2} \binom{n}{k} 0.5^k (1 - 0.5)^{n-k}}{\underbrace{\frac{1}{2} \binom{n}{k} 0.5^k (1 - 0.5)^{n-k}}_{\text{Fair Coin}} + \underbrace{\frac{1}{2} \binom{n}{k} p^k (1 - p)^{n-k}}_{\text{Weighted Coin}}} \end{aligned}$$

Question Three

Suppose that two balanced dice are rolled. Determine and sketch the probability distribution of each of the following random variables.

1. Let X denote the absolute value of the difference between the two numbers that appear.

- The PDF for the absolute difference between two dice can take on 11 different values ranging from -5 to $+5$. We could count all of the possibilities, or we could just recognize that due to the symmetry of the dice we can just treat this like the regular two dice problem where the same outcomes are just relabeled. So, it turns out that we can use the same distribution with a shifted support:

<i>Outcome, x</i>	-5	-4	-3	-2	-1	0	$+1$	$+2$	$+3$	$+4$	$+5$
<i>Number of Ways to Obtain</i>	1	2	3	4	5	6	5	4	3	2	1

which, since there are 36 outcomes for the two dice, we have:

$X = x$	-5	-4	-3	-2	-1	0	$+1$	$+2$	$+3$	$+4$	$+5$
$f(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

The sketch of $f(x)$ will be done on the board.

2. Let Y denote the product of the two numbers that appear.

- Solution to 2: The product of two numbers on a dice can take on the following outcomes:

<i>Outcome, y</i>	1	2	3	4	5	6	8	9	10	12
<i>Number of Ways to Obtain</i>	1	2	2	3	2	4	2	1	2	4
$f(y)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{4}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{4}{36}$

<i>Outcome, y</i>	15	16	18	20	24	25	30	36
<i>Number of Ways to Obtain</i>	2	1	2	2	2	1	2	1
$f(y)$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

The sketch of $f(y)$ will be done on the board.

3. Let Z denote the number of even numbers that appear.

- Solution to 3: There are three possible outcomes: 0, 1, or 2. The outcome of 1 even number has twice the probability of 0 or 2. Thus, we have that the PDF has $f(0) = 0.25$, $f(1) = 0.5$, and $f(2) = 0.25$. The sketch of $f(z)$ is straightforward from this information and will be done on the board.

Question Four

Suppose that you have just purchased a new battery for your smoke detector, and the life of the battery is a random variable X , with pdf

$$f_X(x) = ke^{-x/\beta},$$

where $x \in (0, \infty)$. Assume that t and s are real non-negative numbers.

1. Use the properties of a pdf to find the value of k .

- Solution to 1: We know that the PDF must integrate to 1:

$$\begin{aligned}\int_0^{\infty} f_X(x)dx &= \int_0^{\infty} ke^{-x/\beta}dx \\ 1 &= k \int_0^{\infty} e^{-x/\beta}dx \\ \frac{1}{k} &= -\beta [e^{-x/\beta}]_0^{\infty} \\ k &= \frac{1}{-\beta[0-1]} \\ k &= \frac{1}{\beta}\end{aligned}$$

2. Find an expression for $P(X \geq t)$.

- Solution to 2: This expression is straightforward now that we have the integration constant, k :

$$P(X \geq t) = \int_t^{\infty} \frac{1}{\beta}e^{-x/\beta}dx$$

which can be simplified to

$$P(X \geq t) = e^{-t/\beta}$$

3. Find an expression for $P(X \geq t+s|X \geq s)$.

- Solution to 3: We can just use the formula for conditional probabilities:

$$\begin{aligned}P(X \geq t+s|X \geq s) &= \frac{e^{-(t+s)/\beta}}{e^{-s/\beta}} \\ P(X \geq t+s|X \geq s) &= e^{-t/\beta}\end{aligned}$$

4. Suppose that your batteries have lasted s weeks without dying. Based on your above answers, are you more concerned that the battery is about to die that you were when you first put it in?

- Solution to 4: If my batteries have lasted s weeks without dying, based on my answer to part 3, I should be just as worried as I was before, since survival of the battery tells me nothing new about its likelihood of dying. The exponential distribution (which this is) has this very special property, that no matter how long something has lasted, its rate/probability of failure is constant at any given time.

Question Five

Suppose we investigate the pattern of genetic inheritance for the color of an exotic flower which has either blue or red blossoms. Since the flower lives in a close symbiotic relationship with the very shy squirrel monkey which can't be held in captivity, there is no way of doing a controlled laboratory experiment to answer the research questions.

Each flower carries the color genes of both its "father" and its "mother," so its genetic information can be described as a pair of genes (G_M, G_F) as given by the following table:

		"Father"	
		B	R
"Mother"	B	(B,B)	(B,R)
	R	(R,B)	(R,R)

The phenotype corresponding to red blossoms, R, is said to be *dominant* if any flower which contains at least one gene of the R type (e.g. the combination (R,B)) has red blossoms. Either the blue or the red phenotype is dominant, but before having seen a single specimen of the flower, we think that each possibility is equally likely.

- Suppose we know beforehand that the R and the B alleles are equally frequent, i.e. $P_i(B) = P_i(R) = \frac{1}{2}$ for $i = F, M$, and independent across the "parents," i.e. $P_{FM}(G_F, G_M) = P_F(G_F)P_M(G_M)$. If the blue phenotype B is dominant, what is the probability that a given specimen of the flower has red blossoms? What is the probability of red blossoms if the R is dominant?

- Solution to 1: The probability of a flower having red blossoms is the probability of only getting the recessive genotype on both alleles. Thus, we can compute $P_{FM}(R_F, R_M) = P_F(R_F)P_M(R_M) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. If the R is dominant, then we can solve this by relabeling blue as red and taking the complementary event and probabilities which yields $\frac{3}{4}$.

- It takes a lot of effort to find a single specimen of the plant, so all a well-funded two-month expedition by a team of MIT botanists to the Amazon could gather was a sample of 15 flowers. If R is dominant, what is the probability of 9 out of the 15 flowers having red blossoms?

- Solution to 2: This is another basic binomial question. We have a sample of size 15 with $P(\text{Red Blossom}) = \frac{3}{4}$ which yields the following formula:

$$P(9 \text{ Red Blossom}; 15 \text{ Flowers}) = \binom{15}{9} \left(\frac{3}{4}\right)^9 \left(\frac{1}{4}\right)^{15-9}$$

- Our expedition did in fact return with a sample of 9 red and 6 blue blossoms. Given that, what is the likelihood that the red phenotype is dominant?

- Solution to 3: Bayes' rule is our friend here, again. We just need to pull together the pieces. We know that the alleles are equally frequent, which means that the distribution of alleles should not influence our sampling. We should just consider

what we suspect, beforehand the chance of the red versus the blue allele being dominant as equal. Thus, we have $P(\text{Red Dominant}) = \frac{1}{2}$. Then, we have the binomial distribution for the red and blue phenotypes:

$$P(9 \text{ Red Blossoms} | \text{Red is Dominant}; 15 \text{ Flowers}) = \binom{15}{9} \left(\frac{3}{4}\right)^9 \left(\frac{1}{4}\right)^{15-9}$$

$$P(6 \text{ Blue Blossoms} | \text{Blue is Dominant}; 15 \text{ Flowers}) = \binom{15}{6} \left(\frac{3}{4}\right)^6 \left(\frac{1}{4}\right)^{15-6}$$

which gives us the denominator for Bayes' rule and all of the pieces to compute what we are seeking:

$$P(\text{Red Dom} | 9 \text{ Red}) = \frac{P(9 \text{ Red Blossoms} | \text{Red Dominant})P(\text{Red Dominant})}{P(9 \text{ R} | \text{R Dom})P(\text{R Dom}) + P(6 \text{ B} | \text{B Dom})P(\text{B Dom})}$$

$$= \frac{\frac{1}{2} \binom{15}{9} \left(\frac{3}{4}\right)^9 \left(\frac{1}{4}\right)^{15-9}}{\frac{1}{2} \binom{15}{9} \left(\frac{3}{4}\right)^9 \left(\frac{1}{4}\right)^{15-9} + \frac{1}{2} \binom{15}{6} \left(\frac{3}{4}\right)^6 \left(\frac{1}{4}\right)^{15-6}}$$

$$P(\text{Red Dom} | 9 \text{ Red}) = 0.9643$$

This is really interesting, as a difference of only three blossoms is enough to tell you something with 95% significance. However, it should be noted that this is actually a difference of 6 blossoms, since $9 - 6 = 3$ versus $6 - 9 = -3$.

- At the same time, there is a lonely graduate student working in the same area for two entire years on the same research question. The graduate student is totally cut off from the outside world and doesn't know about the other expedition's findings yet, but bases his inference solely on his own sample. What is his posterior probability of R being dominant given that he found N flowers, out of which x have red blossoms? Show that this probability does not depend directly on N , but only the difference between the number x of red blossoms, and the number $N - x$ of blue blossoms.

- Solution to 4: We just use the answer from part 3 to solve this:

$$\begin{aligned}
P(\text{Red Dom} | x \text{ Red}) &= \frac{P(x \text{ Red Blossoms} | \text{Red Dominant})P(\text{Red Dominant})}{P(x \text{ R} | \text{R Dom})P(\text{R Dom}) + P(N - x \text{ B} | \text{B Dom})P(\text{B Dom})} \\
&= \frac{\frac{1}{2} \binom{N}{x} \left(\frac{3}{4}\right)^x \left(\frac{1}{4}\right)^{N-x}}{\frac{1}{2} \binom{N}{x} \left(\frac{3}{4}\right)^x \left(\frac{1}{4}\right)^{N-x} + \frac{1}{2} \binom{N}{N-x} \left(\frac{3}{4}\right)^{N-x} \left(\frac{1}{4}\right)^x} \\
&= \frac{\binom{N}{x} \left(\frac{3}{4}\right)^x \left(\frac{1}{4}\right)^{N-x}}{\binom{N}{x} \left(\frac{3}{4}\right)^x \left(\frac{1}{4}\right)^{N-x} + \binom{N}{N-x} \left(\frac{3}{4}\right)^{N-x} \left(\frac{1}{4}\right)^x} \\
&= \frac{\binom{N}{x}}{\binom{N}{x} + \binom{N}{N-x} \left(\frac{3}{4}\right)^{N-2x} \left(\frac{1}{4}\right)^{2x-N}} \\
&= \frac{1}{1 + 3^{(N-x)-x}}
\end{aligned}$$

which is just a function of the difference between x and $N - x$, or the difference between red and blue blossoms discovered.