Random variables

Computational Mathematics and Statistics Camp
University of Chicago
September 2018

1. Sophia who took the Graduate Record Examination (GRE) scored 160 on the Verbal Reasoning section and 157 on the Quantitative Reasoning section. The mean score for Verbal Reasoning section for all test takers was 151 with a standard deviation of 7, and the mean score for the Quantitative Reasoning was 153 with a standard deviation of 7.67. Suppose that both distributions are nearly normal.

   a. Write down the short-hand for these two normal distributions.

   Let X denote scores on the Verbal Reasoning section and Y denote scores on the Quantitative Reasoning section. Then,

$$\text{Verbal} : N(\mu = 151, \sigma = 7)$$
$$\text{Quant} : N(\mu = 153, \sigma = 7.67)$$

   b. What is Sophia's Z-score on the Verbal Reasoning section? On the Quantitative Reasoning section?

$$Z_{VR} = \frac{x - \mu}{\sigma} = \frac{160 - 151}{7} \approx 1.29$$

$$Z_{QR} = \frac{y - \mu}{\sigma} = \frac{157 - 151}{7.67} \approx 0.78$$

   c. What do these Z-scores tell you?

   She scored 1.29 standard deviations above the mean on the Verbal Reasoning section and 0.78 standard deviations above the mean on the Quantitative Reasoning section.

   d. Relative to others, which section did she do better on?

   She did better on the Verbal Reasoning section since her Z score on that section was higher.

   e. Find her percentile scores for the two exams.

   The percentile scores can be calculated as follows:

$$Perc_{VR} = P(Z < 1.29) = 0.90815 \approx 91\%$$

$$Perc_{VR} = P(Z < 0.78) = 0.7823 \approx 78\%$$

   f. What percent of the test takers did better than her on the Verbal Reasoning section? On the Quantitative Reasoning section?

   $100 - 91 = 9\%$ did better than her on the Verbal Reasoning section and $100 - 78 = 22\%$ did better than her on the Quantitative Reasoning section.

   g. Explain why simply comparing raw scores from the two sections could lead to an incorrect conclusion as to which section a student did better on.

   We cannot compare the raw scores since they are on different scales. Her scores will be measured relative to the merits of other students on each exam, so it is helpful to consider the Z score. Comparing her percentile scores is a more appropriate way of determining how well she did compared to others taking the exams.

h. If the distributions of the scores on these exams are not nearly normal, would your answers to parts (b) - (f) change? Explain your reasoning.

Answer to part (b) would not change as Z scores can be calculated for distributions that are not normal. However, we could not answer parts (c)-(f) since we cannot use the Z table to calculate probabilities and percentiles without a normal model.

2. The average daily high temperature in June in LA is 77° F with a standard deviation of 5° F. Suppose that the temperatures in June closely follow a normal distribution.

a. What is the probability of observing an 83° F temperature or higher in LA during a randomly chosen day in June?

Let X denote temperatures, then $X \sim N(\mu = 77, \sigma = 5)$.

$$\begin{aligned}(S > 83) &= P(Z > \frac{83-77}{5}) \\ &= P(Z > 1.2) \\ &= 1 - 0.8849 = 0.1151\end{aligned}$$

b. How cool are the coldest 10% of the days (days with lowest average high temperature)during June in LA?

The z score corresponding to the bottom 10% (or 10^{th} percentile) is -1.28.

$$Z = -1.28 = \frac{x-77}{5} \rightarrow x = -1.28 \times 5 + 77 = 70.6$$

The coldest 10% of days are 70.6^oF or colder.

3. Using the information from the previous problem, we use the following equation to convert ° F (Fahrenheit) to ° C (Celsius):

$$C = (F - 32) \times \frac{5}{9}$$

$$sd_C = sd_F \times \frac{5}{9}$$

a. Write the probability model for the distribution of temperature in ° C in June in LA. The parameters of the distribution in ^oC can be calculated as follows:

$$\begin{aligned}\mu_C &= (\mu_F - 32) \times \frac{5}{9} & \sigma_C &= \sigma_F \times \frac{5}{9} \\ &= (77 - 32) \times \frac{5}{9} & &= 5 \times \frac{5}{9} \\ &= 25^oC & &= 2.78^oC\end{aligned}$$

b. What is the probability of observing a 28° C (which roughly corresponds to 83° F) temperature or higher in June in LA? Calculate using the ° C model from part (a).

Probability of observing a 28.3° C temperature or higher in June in LA can be calculated as follows:

$$Z = \frac{28-25}{2.78} = 1.08$$

$$P(X > 28) = P(Z > 1.08) = 1 - P(Z < 1.08) = 1 - 0.8599 = 0.1401$$

c. Did you get the same answer or different answers in part (b) of this question and part (a) of the previous problem? Are you surprised? Explain.

The answers are very close only because the units were changed. (The only reason they differ at all is because 28 °C is 82.4 °F, not 83 °F.)

4. The National Vaccine Information Center estimates that 90% of Americans have had chickenpox by the time they reach adulthood.

a. Is the use of the binomial distribution appropriate for calculating the probability that exactly 97 out of 100 randomly sampled American adults had chickenpox during childhood.

In order to determine if we can we use the binomial distribution to calculate the probability of finding exactly 97 people out of a random sample of 100 American adults had chickenpox in childhood, we need to check if the binomial conditions are met:

1. Independent trials: In a random sample, whether or not one adult has had chickenpox does not depend on whether or not another one has.
2. Fixed number of trials: $n = 100$.
3. Only two outcomes at each trial: Have or have not had chickenpox.
4. Probability of a success is the same for each trial: p = 0.90.

b. Calculate the probability that exactly 97 out of 100 randomly sampled American adults had chickenpox during childhood.

Let X be the number of people who have had chickenpox in childhood, using a binomial distribution with $n = 100$ and $p = 0.90$:

$$P(X = 97) = \binom{100}{97} \times 0.90^{97} \times 0.10^{3} = 0.0059$$

c. What is the probability that exactly 3 out of a new sample of 100 American adults have **not** had chickenpox in their childhood?

$P(97 \text{ out of } 100 \text{ did have chickenpox}) = P(3 \text{ out of } 100 \text{ did not have chickenpox in childhood}) = 0.0059$

d. What is the probability that at least 1 out of 10 randomly sampled American adults have had chickenpox?

$$P(\text{at least } 1) = P(\text{greater than or equal to } 1)$$

$$\begin{aligned} P(X \geq 1) &= P(X = 1) + P(X = 1) + P(X = 2) + \cdots + P(X = 10) \\ &= 1 - P(X = 0) \\ &= 1 - 0.10^{10} \\ &\approx 1 \end{aligned}$$

e. What is the probability that at most 3 out of 10 randomly sampled American adults have **not** had chickenpox?

$$P(\text{at most 3 did not have chickenpox}) = P(\text{less than or equal to 3 where } p = 0.10)$$

$$\begin{aligned}P(X \leq 3) &= P(X=0)+P(X=1)+P(X=2)+P(X=3)\\ &= \binom{10}{0} \times 0.10^0 \times 0.90^{10} + \binom{10}{1} \times 0.10^1 \times 0.90^9 + \binom{10}{2} \times 0.10^2 \times 0.90^8 + \binom{10}{3} \times 0.10^3 \times 0.90^7\\ &= 0.3487 + 0.3874 + 0.1937 + 0.0574\\ &= 0.9872\end{aligned}$$

5. X is a discrete random variable. It takes the value of the number of days required for a governmental agency to respond to a request for information. X is distributed according to the following PMF:

$$f(x) = e^{-4}\frac{4^x}{x!} \text{ for } X \in \{0, 1, 2...\}$$

a. Given this information, what is the probability of a response from the agency in 3 days or less?

$$P(X \leq 3) \approx .43$$

$$\begin{aligned}&P(X \leq 3) = P(X=0)+P(X=1)+P(X=2)+P(X=3)\\ &P(X=0) = e^{-4}\frac{4^0}{0!} = 0.01831564\\ &P(X=1) = e^{-4}\frac{4^1}{1!} = 0.07326256\\ &P(X=2) = e^{-4}\frac{4^2}{2!} = 0.1465251\\ &P(X=3) = e^{-4}\frac{4^3}{3!} = 0.1953668\\ &P(X \leq 3) = 0.01831564 + 0.07326256 + 0.1465251 + 0.1953668 = 0.4334701\end{aligned}$$

b. What is the probability the agency response takes more than 10 but less than 13 days?

$$P(10 < X < 13) \approx .003$$

$$\begin{aligned}P(10 < X < 13) &= P(X=11) + P(X=12)\\ P(X=11) &= \frac{4^{11}}{11!} = 0.001924537\\ P(X=12) &= \frac{4^{12}}{12!} = 0.0006415123\\ P(10 < X < 13) &= 0.001924537 + 0.0006415123 = 0.002566049\end{aligned}$$

c. What is the probability the agency response takes more than 5 days?

$$P(X > 5) = 1 - P(X \leq 5) \approx 0.22$$

We already know $P(X \leq 3) = .43$ so now we only need to find P(X=4) and P(X=5)

$$\begin{aligned}&P(X \leq 3) = 0.4334701\\ &P(X=4) = \frac{4^4}{4!} = 0.1953668\\ &P(X=5) = \frac{4^5}{5!} = 0.1562935\\ &P(X \leq 5) = 0.4334701 + 0.1953668 + 0.1562935 = 0.7851304\\ &P(X > 5) = 1 - 0.7851304 = 0.2148696\end{aligned}$$

d. Suppose using X you generate a new variable, *Responsive*. *Responsive* equals 1 if an agency responds in 5 days or less and 0 otherwise. What is the expected value of *Responsive*?

Since *Responsive* is dichotomous, $E[Responsive] = P(Responsive = 1) = P(X \leq 5)$. From the previous question we know that $P(5 < X) = 0.2148696$ and $P(X \leq 5) = 0.7851304$. So $E[Responsive] = 0.7851304$.

e. What is the variance of *Responsive*?

Responsive is distributed Bernoulli. $Var(Bernoulli) = p(1-p)$. In this case $p = 0.7851304$. So $Var(Responsive) = (0.7851304 \times (1 - 0.7851304)) = 0.1687007$.

6. Suppose we've developed a model predicting the outcome of the upcoming midterm elections in a state with 4 Congressional districts. In each district there are two candidates, a Republican and a Democrat. We have reason to believe the following PMF describes the distribution of potential election results where $K \in \{0, 1, 2, 3, 4\}$ and is the number of seats won by Republican candidates in the upcoming election.

$$P(K = k|\theta) = \binom{4}{k}\theta^k(1-\theta)^{4-k}$$

Based on polling information, we think the appropriate value for θ is 0.55.

a. What's the expected number of seats Republicans will win in the upcoming election?

The election outcomes are distributed according to a binomial distribution so we find the expected value given our two parameters, n and θ.

$$\begin{aligned} E[Binomial] &= n \times \theta \\ &= 4 \times .55 \\ &= 2.2 \end{aligned}$$

b. Given this PMF, what's the probability that no Republican legislators win in the upcoming election?

$$P(k = 0) \approx 0.04100625$$

$$\begin{aligned} P(k = 0) &= \binom{4}{0}.55^0(1 - .55)^4 \\ &= 1 \times 1 \times 0.04100625 \\ &= 0.04100625 \end{aligned}$$

c. What's the probability that Republican legislators win a majority of the seats in this state?

$$P(2 < k) \approx .39$$

$$\begin{aligned} P(k = 3) &= \binom{4}{3}.55^3(1 - .55)^1 = 4 \times 0.07486875 = 0.299475 \\ P(k = 4) &= \binom{4}{4}.55^4(1 - .55)^0 = 1 \times 0.09150625 = 0.09150625 \\ P(2 < k) &= 0.299475 + 0.09150625 = 0.3909813 \end{aligned}$$

d. A prominent political pundit declares they are certain that Republicans will win a majority of seats in the next election and offers the following bet. If Republicans win a majority of the seats, we must pay the pundit \$15.00. If Republican's fail to win a majority of states, we will win \$20.00. Based on our model, should we take this bet?

Hint: Think of the betting outcomes as a random variable. Find the expected value of this random variable.

We can should take this bet as we will net an expected return of \$6.32 from it.

We can think of *bet* as a variable with the associated values:

$$bet = \begin{cases} -15 & \text{if Republican majority;} \\ 20 & \text{otherwise.} \end{cases}$$

$$\begin{aligned} E[bet] &= E[bet|Republican\ Majority] + E[bet|Otherwise] \\ &= -15 \times P(Republican\ Majority) + 20 \times P(Otherwise) \\ &= -15 * (0.3909813) + 20 * (1 - 0.3909813) \\ &= -5.864719 + 12.18037 \\ &\approx 6.3 \end{aligned}$$

e. Suppose we are offered a second bet with a more complicated structure. In this case we'll receive \$100 if the Republicans win a majority, \$50 if neither party wins a majority and we'll have to pay \$200 if the Democrats win a majority. Should we take this bet?

We should take this bet as well since we will net an expected \$9.18.

We can think of bet_2 as a variable that takes the following values:

$$bet_2 = \begin{cases} 100 & \text{if Republican majority;} \\ 50 & \text{if tie;} \\ -200 & \text{otherwise.} \end{cases}$$

$$\begin{aligned} P(Rep\ Majority) &= 0.3909813 \\ P(tie) &= \binom{4}{2}.55^2(1 - .55)^2 = 6 \times 0.06125625 = 0.3675375 \\ P(Dem\ Majority) &= 1 - P(Rep\ Majority) - P(tie) \\ &= 1 - 0.3909813 - 0.3675375 \\ &= 0.2414812 \\ E[bet_2] &= E[bet_2|Rep\ Majority] + E[bet_2|Tie] + E[bet_2|Otherwise] \\ &= 100 \times P(Republican\ Majority) + 50 \times P(Tie) \\ &\quad + -200 \times P(Republicans\ don't\ win\ majority) \\ &= 100 * 0.3909813 + 50 * 0.3675375 - 200 * 0.2414812 \\ &\approx 9.2 \end{aligned}$$

7. The random variable X has a probability density function:

$$f(x; \lambda) = \frac{\lambda^x exp(-\lambda)}{x!}$$

for $x = 0, 1, 2, \ldots$, (i.e. X has a Poisson distribution with parameter λ). In a lengthy manuscript, it is discovered that only 13.5 percent of the pages contain no typing errors.

If we assume that the number of errors per page is a random variable with a Poisson distribution, find the percentage of pages with exactly one error.

With a Poisson variable, the pdf supplies probability of a given number (count) of events, i.e. $Pr(X = x)$. From the prompt, we know that $Pr(X = 0) = .135$. Using this, we can solve for λ, which is the average rate of "success" in a given interval, (i.e. the average number of times an event occurs in a given interval). Here, the event is an error occurring and the interval is the page.

$$
\begin{aligned}
f(x;\lambda) &= \frac{\lambda^x exp(-\lambda)}{x!} \\
f(x=0;\lambda) = \frac{\lambda^0 exp(-\lambda)}{0!} &= .135 \\
\frac{exp(-\lambda)}{1} &= .135 \\
\frac{1}{exp(\lambda)} &= .135 \\
exp(\lambda) &= \frac{1}{.135} \\
exp(\lambda) &= 7.407 \\
\lambda &= 2.002
\end{aligned}
$$

Now that we know the value of λ, we know that on average, a page contains 2.002 errors. (Note: If we want to change the interval to say, 10 pages, we just adjust λ accordingly. If there are 2.002 errors per page, there are 2.002*10=20.02 errors per 10 pages.) Now we can solve for $Pr(X = 1)$.

$$f(x=1;\lambda) = \frac{2.002^1 exp(-2.002)}{1!} = .27$$

8. A prominent hypothesis in the comparative politics literature is the "resource curse": in countries that depend economically on extraction of natural resources, there is a lower probability of democratic political institutions. Suppose we want to investigate this hypothesis. For each country, we collect data on whether the country is democratic and whether it depends on oil exports. Denote the outcome variable (democracy) for country i as y_i, where $y_i = 1$ for democracies and $y_i = 0$ for non-democracies. Similarly, denote the predictor variable (dependence on oil) as x_i, with $x_i = 1$ for oil-dependent states and $x_i = 0$ for non-oil-dependent states.

   For each country, we model y_i as a Bernoulli random variable with rate π_i (note the subscript), with the following PMF:

$$Pr(y_i = 1 \mid \pi_i) = (\pi)^{y_i}(1-\pi)^{1-y_I}$$

   To incorporate the predictor, we'll let π_i vary as a function of x_i. We'll model this using the inverse logit link function:

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

   Our goal in future courses would be to estimate β_0 and β_1 using maximum likelihood estimation. These parameters will tell us how the probability varies as a function of X. Let's say for now we ran this regression and found that $\hat{\beta_0} = .35$ and $\hat{\beta_1} = -0.9$. Using the estimates of β_0 and β_1, what is the probability that an oil-dependent state is a democracy? What is the probability that a non-oil-dependent state is a democracy?

   We can plug in the estimates to the formula for π to figure out the relevant PMF. For oil-dependent states, $x_i = 1$, so we have

$$
\begin{aligned}
\pi_{x_i=1} &= \frac{\exp(0.35 - 0.9(1))}{1 + \exp(0.35 - 0.9(1)} \\
&= \frac{\exp(-0.55)}{1 + \exp(-0.55)} \\
&\approx 0.37
\end{aligned}
$$

So the PMF for oil-dependent states is $f(y_i \mid x_i = 1) = 0.37^{y_i}(0.63)^{1-y_i}$, so the probability of democracy for oil-dependent states is $f(1 \mid x_i = 1) = 0.37$.

For non-oil-dependent states, $x_i = 0$ so we have

$$\begin{aligned}\pi_{x_i=0} &= \frac{\exp(0.35 - 0.9(0))}{1 + \exp(0.35 - 0.9(0)} \\ &= \frac{\exp(0.35)}{1 + \exp(0.35)} \\ &\approx 0.59\end{aligned}$$

So the PMF for oil-dependent states if $f(y_i \mid x_i = 0) = 0.59^{y_i}(0.41)^{1-y_i}$, so the probability of democracy for non-oil-dependent states is $f(1 \mid x_i = 0) = 0.59$.