# Section #3: Continuous Random Variables

1. **Website Visits**: On average, visitors leave your website after 5 minutes. Assume that the length of stay is exponentially distributed. What is the probability that a user stays more than 10 minutes?

2. **Approximating Normal**: Your website has 100 users and each day each user independently has a 20% chance of logging into your website. Use a normal approximation to estimate the probability that more than 21 users log in.

3. **Continuous Random Variable**: Let $X$ be a continuous random variable with the following probability density function:

$$f_X(x) = \begin{cases} c(e^{x-1} + e^{-x}) & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

   a. Find the value of $c$ that makes $f_X$ a valid probability distribution.

   b. What is $P(X > 0.75)$?

4. **Student Heights**: Adult heights can be considered to be normally distributed, but the distributions are different between men and women.

   a. Adult women have a mean height of 65 inches and a standard deviation of 3.5 inches. What is the probability that a randomly selected adult woman is over 72 inches? What is the probability that a randomly selected woman is between 63 and 65 inches?

   b. Adult men are slightly taller, and the distribution of their heights has a slightly different spread. We know that the average adult man is 70 inches tall, and that 10 percent of adult men are under 65 inches tall. What is the standard deviation of adult men's heights, if they are normally distributed?

   c. Chris Piech is 6'5" (77 inches) tall. What is the probability that, of the six men in a typical section, at least one of them is taller than Chris? What is the probability that there another man is exactly the same height as Chris? What is the probability that another man is approximately his same height (i.e. within +/- half an inch)?

5. **Elections**: We would like to see how we could predict an election between two candidates in France (A and B), given data from 10 polls. For each of the 10 polls, we report below their sample size, how many people said they would vote for candidate A, and how many people said they would vote for candidate B. Not all polls are created equal, so for each poll we also report a value "weight" which represents how accurate we believe the poll was. The data for this problem can be found on the class website in polls.csv:

| Poll | N samples | A votes | B votes | Weight |
|------|-----------|---------|---------|--------|
| 1 | 862 | 548 | 314 | 0.93 |
| 2 | 813 | 542 | 271 | 0.85 |
| 3 | 984 | 682 | 302 | 0.82 |
| 4 | 443 | 236 | 207 | 0.87 |
| 5 | 863 | 497 | 366 | 0.89 |
| 6 | 648 | 331 | 317 | 0.81 |
| 7 | 891 | 552 | 339 | 0.98 |
| 8 | 661 | 479 | 182 | 0.79 |
| 9 | 765 | 609 | 156 | 0.63 |
| 10 | 523 | 405 | 118 | 0.68 |
| **Totals:** | **7453** | **4881** | **2572** | |

a. First, assume that each sample in each poll is an independent experiment of whether or not a random person in France would vote for candidate A (disregard weights).

  - Calculate the probability that a random person in France votes for candidate A.

  - Assume each person votes for candidate A with the probability you've calculated and otherwise votes for candidate B. If the population of France is 64,888,792, what is the probability that candidate A gets more than half of the votes?

b. Nate Silvers at fivethirtyeight pioneered an approach called the "Poll of Polls" to predict elections. For each candidate A or B, we have a random variable $S_A$ or $S_B$ which represents their strength on election night (like ELO scores). The probability that A wins is $P(S_A > S_B)$.

  - Identify the parameters for the random variables $S_A$ and $S_B$. Both $S_A$ and $S_B$ are defined to be normal with the following parameters:

$$S_A \sim \mathcal{N}\left(\mu = \sum_i p_{A_i} \cdot \text{weight}_i, \; \sigma^2\right) \qquad S_B \sim \mathcal{N}\left(\mu = \sum_i p_{B_i} \cdot \text{weight}_i, \; \sigma^2\right)$$

where $p_{A_i}$ is the ratio of A votes to N samples in poll $i$, $p_{B_i}$ is the ratio of B votes to N samples in poll $i$, $\text{weight}_i$ is the weight of poll $i$, $m_i$ is the N samples in poll $i$ and:

$$\sigma = \frac{K}{\sqrt{\sum_i m_i}} \quad \text{s.t. } K = 350; \text{ thus } \sigma = 4.054.$$

  - We will calculate $P(S_A > S_B)$ by simulating 100,000 fake elections. In each fake election, we draw a random sample for the strength of A from $S_A$ and a random sample for the strength of B from $S_B$. If $S_A$ is greater than $S_B$, candidate A wins. What do we expect to see if we simulate so many times? What do we actually see?

c. Which model, the one from (a) or the model from (b) seems more appropriate? Why might that be the case? On election night candidate A wins. Was your prediction from part (b) "correct"?