Jerry Cain
CS109

Section #4
April 27, 2021

# Section #4

## 1 Warmups

### *1.1 Joint Distributions*

1. Given a Normal RV $X \sim N(\mu, \sigma^2)$, how can we compute $P(X \leq x)$ from the standard Normal distribution Z with CDF $\phi$?

2. What is a continuity correction and when should we use it?

3. If we have a joint PMF for discrete random variables $p_{X,Y}(x, y)$, how can we compute the marginal PMF $p_X(x)$?

### *1.2 Independent Random Variables*

1. What distribution does the sum of two independent binomial RVs $X + Y$ have, where $X \sim Bin(n_1, p)$ and $Y \sim Bin(n_2, p)$? Include the parameter(s) in your answer. Why is this the case?

2. What distribution does the sum of two independent Poisson RVs $X + Y$ have, where $X \sim Poi(\lambda_1)$ and $Y \sim Poi(\lambda_2)$? Include the parameter(s) in your answer.

### *1.3 Joint Random Variables Statistics*

**True or False?** The symbol $Cov$ is covariance, and the symbol $\rho$ is Pearson correlation. And $X \perp Y$ is just a fancy way to say that $X$ and $Y$ are independent.

| | |
|---|---|
| $X \perp Y \implies Cov(X, Y) = 0$ | $Var(X + X) = 2Var(X)$ |
| $Cov(X, Y) = 0 \implies X \perp Y$ | $X \sim \mathcal{N}(0, 1) \wedge Y \sim \mathcal{N}(0, 1) \implies \rho(X, Y) = 1$ |
| $Y = X^2 \implies \rho(X, Y) = 1$ | $Y = 3X \implies \rho(X, Y) = 3$ |

## 2 Problems

### 2.1 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 Silver 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 $i$, weight$_i$ is the weight of poll $i$, $m_i$ is the N samples in poll $i$ and:

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

- 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"?

## 2.2   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.