Gov 2001: Problem Set 1

Spring 2025

February 3rd, 2025

Problem Set Instructions:

- This problem set is due on **Feb 12, 11:59 pm** Eastern time. Please upload a PDF of your solutions to **Gradescope**.
- We will accept hand-written solutions but we strongly advise graduate students to typeset your answers in LaTeX.
- Citing your sources is always a good practice in academia. Please list the names of other students / sources / AI you obtained help from on this problem set.

1 Birthday Problem (20 points)

Let's re-consider the birthday problem we discussed during lecture.

Birthday Problem: There are 20 people in a room. Assume each person's birthday is equally likely to be any of the 365 days of the year (no leap babies) and birthdays are independent. What is the probability that at least one pair of people have the same birthday?

1.1 Approach 1 (10 points)

Repeat what we did in class by taking the complement of the event we are interested in. Then, calculate the probability in the birthday problem.

$1 - \left(\binom{365}{20} \cdot 20!\right)/365^{20}$

1.2 Approach 2 (10 points)

Instead of the smarter approach we did in class, can you solve the problem other way around?

Hint: What is the probability of 19 people share the same bday? 18? 17?

Hint: We can add up probabilities of disjoint events.

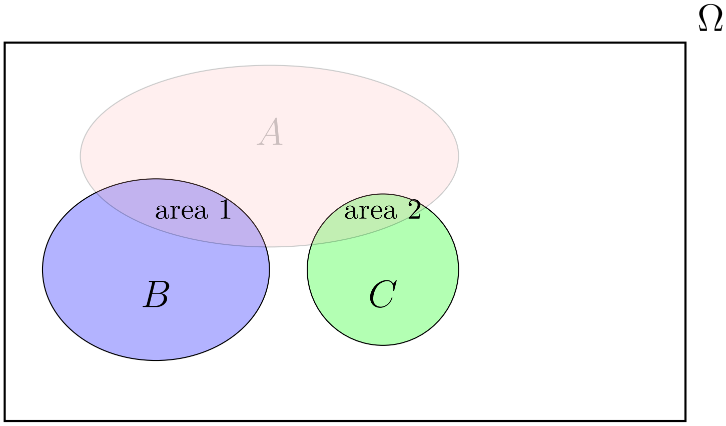
$P(19 \text{ same, } 1 \text{ dif}) = \left(\binom{20}{19} \cdot 365 * 364\right)/365^{20}$,
and $P(18 \text{ same, } 2 \text{ dif}) = \left(\binom{20}{18} \cdot 365 * 364 * 343\right)/365^{20}$

2 Union of Two Events (10 points)

Recall our discussion during class: Even if B and C are disjoint,

$$\mathbb{P}(A \mid B \cup C) \neq \mathbb{P}(A \mid B) + \mathbb{P}(A \mid C).$$

Can you give us an example to illustrate this idea? You can also use visualization to help with your point. For those who are typesetting the pset, `tikz` package is your friend!



$P(A|B \cup C) = \frac{\text{area 1 and 2}}{\text{blue and green}}$
$P(A|B) + P(A|C) = \frac{\text{area 1}}{\text{blue}} + \frac{\text{area 2}}{\text{green}}$

Example (story): Jar A has 4 white balls out of 10 balls, while jar B has 6 out of 15. The probability of drawing a white when randomly drawing from all balls in A and B is $10/25 = 0.4$, while the sum of drawing a white from A and drawing a white from B is $0.4 + 0.4 = 0.8$

3 Peace Between US and UK (10 points)

You and your coauthor are trying to model the probability over how long the peace between the United States and United Kingdom will last, where the sample space is the set of all positive years ($\Omega = \{1, 2, 3, \ldots\}$). You and your coauthor agree that this sample space is countably infinite, but your coauthor insists that you should just model every year as equally likely. Using the axioms of probability, show your coauthor that each year in the sample space cannot be equally likely.

Suppose there exists a probability p. Then we have:

$$\sum_{i=1}^{N} p = 1$$

, where $N \to \infty$. So we have $N \cdot p = 1$, and $p = 0$.

Or we can discuss: if $p > 0$, the sum goes to infinity; if $p = 0$, the sum is 0. Both contradict with the axiom, which implies that the sum should be 1.

4 Survey Study (20 points)

Suppose you are conducting a panel study over two waves, three months apart. In the first wave, you sample without replacement n respondents from your pool of N potential panelists. In the second wave, you take a sample of size m without replacement from the same pool. Our goal will be to obtain the probability that exactly k of the m respondents in the second wave were also in the first wave. We'll do this in a few steps.

You should assume that being a respondent in one wave has no effect on being selected in another wave and that (for now) all those selected participate.

1. (5pts) What are the total number of ways to select m respondents from the pool in the second wave?

   $\binom{N}{m}$

2. (10pts) How many different ways are there to select m respondents in the second wave such that exactly k are from the n selected in the first wave?

   $\binom{n}{k}\binom{N-n}{m-k}$

3. (5pts) Assuming everyone is equally likely to be selected, what is the probability that exactly k of the m respondents in the second wave were also selected in the first wave?

   $\binom{n}{k}\binom{N-n}{m-k}/\binom{N}{m}$

5 Independence of Data (20 points)

1. (5pts) Suppose that, in advance of the 2024 presidential election, you know that Pennsylvania and Georgia are pure toss-ups between the Democratic and Republican candidates and are independent of each other. Let X be the number of states the Republican candidate wins of the two. What are the PMF anf CDF of X?

   Let X be the number of states the Republican candidate wins. Then,

   $$X \sim \text{Binomial}(2, 0.5).$$

   PMF of X: For $x = 0, 1, 2$,

   $$\mathbb{P}(X = x) = \binom{2}{x}(0.5)^x(0.5)^{2-x} = \binom{2}{x}(0.5)^2.$$

   Thus,

   $$\mathbb{P}(X = 0) = \binom{2}{0}(0.5)^2 = 1 \cdot 0.25 = 0.25,$$

   $$\mathbb{P}(X = 1) = \binom{2}{1}(0.5)^2 = 2 \cdot 0.25 = 0.50,$$

   $$\mathbb{P}(X = 2) = \binom{2}{2}(0.5)^2 = 1 \cdot 0.25 = 0.25.$$

CDF of X: Define $F_X(x) = \mathbb{P}(X \leq x)$. Then:

$$F_X(x) = \begin{cases} 0, & x < 0, \\ 0.25, & 0 \leq x < 1, \\ 0.25 + 0.50 = 0.75, & 1 \leq x < 2, \\ 1, & x \geq 2. \end{cases}$$

2. (5pts) Now suppose that you are interested in local elections. Two counties are looking to elect a sheriff, but these elections are not toss-ups - in one, the Republican candidate has a 65% chance of winning, while in the other, the Republican candidate has a 40% chance of winning. Again, the two sheriff's races are independent from each other. Let Y by the number of elections the Republican candidate wins of the two. What are the PMF and CDF of Y?

   Let

   $$Y_1 \sim \text{Bernoulli}(0.65) \quad \text{and} \quad Y_2 \sim \text{Bernoulli}(0.40),$$

   with

   $$Y = Y_1 + Y_2.$$

   PMF of Y:

   - $Y = 0$: Both losses,

     $$\mathbb{P}(Y = 0) = (1 - 0.65)(1 - 0.40) = (0.35)(0.60) = 0.21.$$

   - $Y = 1$: Exactly one win (two cases):

     $$\mathbb{P}(Y = 1) = (0.65)(0.60) + (0.35)(0.40) = 0.39 + 0.14 = 0.53.$$

   - $Y = 2$: Both wins,

     $$\mathbb{P}(Y = 2) = (0.65)(0.40) = 0.26.$$

   CDF of Y: Let $F_Y(y) = \mathbb{P}(Y \leq y)$. Then:

   $$F_Y(y) = \begin{cases} 0, & y < 0, \\ 0.21, & 0 \leq y < 1, \\ 0.21 + 0.53 = 0.74, & 1 \leq y < 2, \\ 1, & y \geq 2. \end{cases}$$

3. (10pts) Finally suppose that you know the joint PMF of X and Y, $P(X = x, Y = y)$, to be as follows:

   Are X and Y independent? Why?

	Y = 0	Y = 1	Y = 2
X = 0	.0525	.15	.065
X = 1	.105	.265	.13
X = 2	.0525	.115	.065

We can check for independence:
X and Y are independent if for every x and y,

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y).$$

Consider the cell $X = 0, Y = 1$:

$$\mathbb{P}(X = 0, Y = 1) = 0.15.$$

The product of the marginals is:

$$\mathbb{P}(X = 0)\mathbb{P}(Y = 1) = 0.25 \times 0.53 = 0.1325.$$

Since $0.15 \neq 0.1325$ (and similar discrepancies occur for other cells), we conclude that X and Y are **not independent**.

6 Stopping a Driver (20 points)

In the United States, roughly 29% of white drivers get stopped by police compared to roughly 42% of non-white drivers. Of white drivers who are stopped by police, 25% have illegal contraband, while 28% of stopped non-white drivers have illegal contraband.[1]

Let C be the event of a driver possessing contraband, W be the event of the driver being white, and S being the event of the driver getting stopped by the police. Suppose that the probability of contraband found among non-stopped drivers is equal across both racial groups.

1. (10 pts) What values of the probability of contraband **among non-stopped drivers** would imply the probability of contraband among whites is higher than contraband among non-whites **in general**?

   - $P(S \mid W) = 0.29$ and $P(S \mid W^c) = 0.42$, where S is the event of being stopped and W is the event that a driver is white.
   - Among stopped drivers, $P(C \mid S, W) = 0.25$ and $P(C \mid S, W^c) = 0.28$, where C is the event of possessing illegal contraband.
   - We assume that the probability of contraband among non-stopped drivers is equal across races. Let
   $$q = P(C \mid S^c, W) = P(C \mid S^c, W^c).$$

[1] These are approximate figures.

The overall probability that a driver of a given race has contraband is obtained by a law of total probability:

$$P(C \mid W) = P(C \mid S, W)P(S \mid W) + P(C \mid S^c, W)P(S^c \mid W),$$

and similarly for non-whites,

$$P(C \mid W^c) = P(C \mid S, W^c)P(S \mid W^c) + P(C \mid S^c, W^c)P(S^c \mid W^c).$$

Plugging in the numbers and writing q for the contraband rate among non-stopped drivers:

$$P(C \mid W) = 0.25(0.29) + q\,(1 - 0.29) = 0.0725 + 0.71\,q,$$
$$P(C \mid W^c) = 0.28(0.42) + q\,(1 - 0.42) = 0.1176 + 0.58\,q.$$

We want to know for what values of q we have

$$P(C \mid W) > P(C \mid W^c).$$

That is,

$$0.0725 + 0.71\,q > 0.1176 + 0.58\,q.$$

Subtract $0.58\,q$ and 0.0725 from both sides:

$$0.13\,q > 0.0451.$$

Thus,

$$q > \frac{0.0451}{0.13} \approx 0.3462.$$

Answer (a): If the probability of contraband among non-stopped drivers exceeds approximately 34.6%, then

$$P(C \mid W) > P(C \mid W^c),$$

i.e. the overall contraband rate among white drivers would be higher than among non-white drivers.

2. (15 pts) Suppose you are asked to find whether there is (and if there is, how much) racial bias in who is stopped by the police. You use the following measure to quantify racial discrimination: $P(S|C, W^c) - P(S|C, W)$. The reasoning for this measure is as follows: if there is no racial bias in police stops, we might expect that $S \perp W \mid C$. This would mean given that if the driver is actually carrying contraband, their race should not update the probability of a police choosing to stop them. To show that this false – that race does update the probability of a stop – we need to simply show that this measure is not equal to zero. In English, it can be interpreted as, "the increase in probability of getting stopped if a contraband carrier is not white". Assuming that the rates of contraband in the population are independent of race, plot and interpret possible bounds for this measure using the information provided in this problem.

Hint: use Bayes' Rule and compute the bounds using R.[2] Hint part 2: let $x = P(C|W^c, S^c) = P(C|W, S^c)$.

We are given the measure:

$$\Delta = P(S \mid C, W^c) - P(S \mid C, W).$$

Under the null hypothesis of no racial bias (i.e. $S \perp W \mid C$) we would expect $\Delta = 0$. In words, Δ represents the increase in the probability of being stopped for a contraband carrier if the driver is non-white.

Using Bayes' rule, we can write:

$$P(S \mid C, W) = \frac{P(C \mid S, W)P(S \mid W)}{P(C \mid W)},$$

and similarly,

$$P(S \mid C, W^c) = \frac{P(C \mid S, W^c)P(S \mid W^c)}{P(C \mid W^c)}.$$

We already computed

$$P(C \mid W) = 0.0725 + 0.71\,q,$$
$$P(C \mid W^c) = 0.1176 + 0.58\,q.$$

Also, we are given:

$$P(C \mid S, W) = 0.25, \quad P(S \mid W) = 0.29,$$

$$P(C \mid S, W^c) = 0.28, \quad P(S \mid W^c) = 0.42.$$

Thus,

$$P(S \mid C, W) = \frac{0.25 \times 0.29}{0.0725 + 0.71\,q} = \frac{0.0725}{0.0725 + 0.71\,q},$$
$$P(S \mid C, W^c) = \frac{0.28 \times 0.42}{0.1176 + 0.58\,q} = \frac{0.1176}{0.1176 + 0.58\,q}.$$

Therefore, the measure of racial bias is:

$$\Delta(q) = \frac{0.1176}{0.1176 + 0.58\,q} - \frac{0.0725}{0.0725 + 0.71\,q}.$$

Plotting and Interpreting the Bounds: Since the true value of $q = P(C \mid S^c)$ is unknown, we can consider how $\Delta(q)$ varies with q. For example, note the following endpoints:

[2] For more on statistical fallacies on estimating racial disparities in policing, see "Administrative Records Mask Racially Biased Policing" (Knox, Dean, Will Lowe, and Jonathan Mummolo).

- When $q = 0$:

$$P(S \mid C, W) = \frac{0.0725}{0.0725} = 1, \quad P(S \mid C, W^c) = \frac{0.1176}{0.1176} = 1,$$

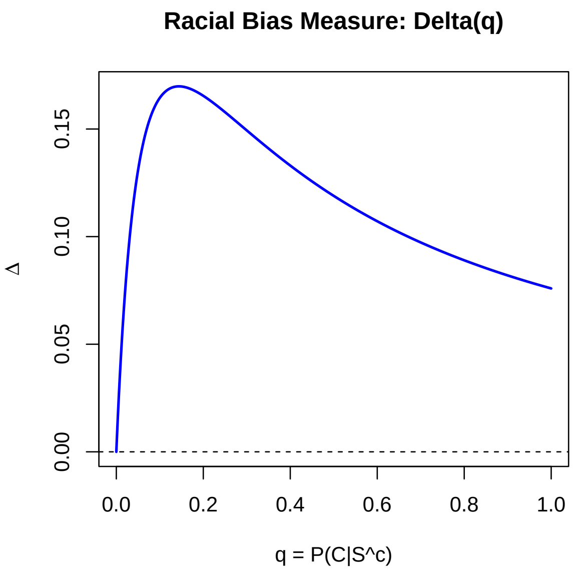
  so $\Delta(0) = 1 - 1 = 0$.
- When q is very large (say $q \to 1$):

$$P(S \mid C, W) \to \frac{0.0725}{0.71\,q} \to 0.10211, \quad P(S \mid C, W^c) \to \frac{0.1176}{0.58\,q} \to 0.20276,$$

  and the difference will not vanish.

Tthe following R code snippet plots $\Delta(q)$ for q between 0 and 1:

```
# R code to plot the measure Delta(q)
q <- seq(0, 1, length.out=500)
Delta <- (0.1176 / (0.1176 + 0.58*q)) - (0.0725 / (0.0725 + 0.71*q))
plot(q, Delta, type="l", lwd=2, col="blue",
     xlab="q = P(C|S^c)", ylab=expression(Delta),
     main="Racial Bias Measure: Delta(q)")
abline(h=0, lty=2)
```



7 Challenge Question (Optional, 10 points)

Prove Vandermonde's Identity:

$$\sum_{k=0}^{r} \binom{m}{k}\binom{n}{r-k} = \binom{m+n}{r}$$

We wish to prove that

$$\sum_{k=0}^{r}\binom{m}{k}\binom{n}{r-k}=\binom{m+n}{r},$$

where $0 \leq r \leq m+n$.

Consider a set X with $m+n$ elements, which we partition into two disjoint subsets:

$$X = A \cup B, \quad \text{with } |A| = m \text{ and } |B| = n.$$

The right-hand side, $\binom{m+n}{r}$, counts the number of ways to choose an r-element subset from X.

Now, any r-element subset of X must contain some elements from A and the remainder from B. Suppose exactly k of the r elements are chosen from A. Then, the remaining $r-k$ elements must be chosen from B. The number of ways to choose these elements is

$$\binom{m}{k} \text{ ways from } A \quad \text{and} \quad \binom{n}{r-k} \text{ ways from } B.$$

For a fixed k, there are

$$\binom{m}{k}\binom{n}{r-k}$$

ways to choose an r-element subset of X with exactly k elements from A.

Since k can vary from 0 (no elements chosen from A) to r (all r elements chosen from A), we sum over all possible values of k. This gives

$$\sum_{k=0}^{r}\binom{m}{k}\binom{n}{r-k},$$

which counts the total number of r-element subsets of X by partitioning according to how many elements come from A.

Since both the right-hand side $\binom{m+n}{r}$ and the sum represent the number of ways to choose r elements from X, we have

$$\sum_{k=0}^{r}\binom{m}{k}\binom{n}{r-k}=\binom{m+n}{r}.$$

□

Second approach:

Recall the Binomial Theorem, which states that for any nonnegative integer p,

$$(1+x)^p = \sum_{j=0}^{p}\binom{p}{j}x^j.$$

Consider the product

$$(1+x)^m(1+x)^n.$$

By the Binomial Theorem, we have:

$$(1+x)^m = \sum_{i=0}^{m} \binom{m}{i} x^i \quad \text{and} \quad (1+x)^n = \sum_{j=0}^{n} \binom{n}{j} x^j.$$

Therefore, their product is

$$(1+x)^m(1+x)^n = \left(\sum_{i=0}^{m} \binom{m}{i} x^i\right)\left(\sum_{j=0}^{n} \binom{n}{j} x^j\right).$$

When we multiply these two sums, the coefficient of x^r is given by summing over all pairs (i, j) such that $i + j = r$:

$$\text{Coefficient of } x^r = \sum_{i+j=r} \binom{m}{i}\binom{n}{j}.$$

If we set $i = k$ and $j = r - k$, the above sum becomes

$$\sum_{k=0}^{r} \binom{m}{k}\binom{n}{r-k}.$$

On the other hand, note that

$$(1+x)^m(1+x)^n = (1+x)^{m+n}.$$

Again by the Binomial Theorem,

$$(1+x)^{m+n} = \sum_{r=0}^{m+n} \binom{m+n}{r} x^r.$$

Hence, the coefficient of x^r in $(1+x)^{m+n}$ is $\binom{m+n}{r}$.

Since the coefficient of x^r must be the same regardless of how we obtain it, we have:

$$\sum_{k=0}^{r} \binom{m}{k}\binom{n}{r-k} = \binom{m+n}{r}.$$

This completes the proof.

□