Polling (Basics)

What does it mean if the poll says a candidate wins with 29% - 31% probability, with 96% confidence?

Does it mean the real portion of votes the candidate wins is within 29% - 31% with 96% probability?

Ans: no.
Model of polling

180m votes for C

120m votes for D
Polling

- Pick one person: (random experiment)
  With 0.4 probability, vote for D
  With 0.6 probability, vote for C

- $\Omega = ?$

- $\Omega = \{C, D\}$

- Construct a random variable by mapping D to 0 and C to 1
  \[ Z = \begin{cases} 0, & \text{with probability 0.4} \\ 1, & \text{with probability 0.6} \end{cases} \]

Language

- Bernoulli ($p$)
  \[ Z = \begin{cases} 0, & \text{with probability } 1 - p \\ 1, & \text{with probability } p \end{cases} \]

  $E(Z) = p$
  $Var(Z) = p(1-p)$

  Can also model a coin toss, winning of a game (any experiment with 2 outcomes)
Bernoulli process

Coin toss 1, coin toss 2, coin toss 3, coin toss 4 ......

\[ Z_1 \quad Z_2 \quad Z_3 \quad Z_4 \]

\[ Z_i = \begin{cases} 
0, & \text{with probability } 1 - p \\
1, & \text{with probability } p 
\end{cases} \]

\[ P(Z_1 = 1, Z_2 = 0, Z_3 = 1, Z_4 = 1) = ? \]

\[ p(1-p)p \cdot p = p^3(1-p) \]

i.i.d. : independent and identically distributed

Polling

Voter 1, voter 2, voter 3, voter 4 ......

\[ Z_1 \quad Z_2 \quad Z_3 \quad Z_4 \]

- Are \( Z_i \) i.i.d. Bernoulli?

- Sampling without replacement

\[ P(Z_1 = 1, Z_2 = 0) = \frac{180m \times 120m}{300m \times (300m - 1)} = 0.4 \times (0.6 + 2 \times 10^{-9}) \approx 0.4 \times 0.6 \]

(approximately i.i.d. Bernoulli)
Question 1

- Out of 10 voters sampled, what’s the probability that 5 vote for C?

- $Z_i$ are iid Bernoulli($p$)

- $Z_i = \begin{cases} 0, & \text{with probability } 1 - p \\ 1, & \text{with probability } p \end{cases}$

```
1 1 1 1 0 0 0 0 0 \\
0 1 1 0 0 1 1 0 0 \\
......
```

What’s the probability of such a sequence occurring?
How many such sequences are there?

More generally, out of $n$ voters sampled, what’s the probability that $k$ vote for C?
Binomial \((n,p)\)

- \(X = \sum_{i=1}^{n} Z_i\)
- \(Z_i\) are iid Bernoulli \((p)\)
- \(Z_i = \begin{cases} 0, & \text{with probability } 1 - p \\ 1, & \text{with probability } p \end{cases}\)
- \(P(k \text{ vote for } C)\) is the same as \(P(X = k)\)
  \[P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}\]

Why is it called Binomial?

Binomial \((n,p)\)

\[X = \sum_{i=1}^{n} Z_i\]
\[P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}\]

What’s \(E(X)\) and \(Var(X)\)?

\[E(X) = \sum_{i=1}^{n} E(Z_i) = np\]
\[Var(X) = \sum_{i=1}^{n} Var(Z_i) = np(1-p) \text{ (only true for i.i.d)}\]
Question 2

- What’s the probability that the first vote for C comes from the 5th voter?

- More generally, what’s the probability that the first vote for C comes from the kth voter?

- \[ P(Z_1 = 0, Z_2 = 0, \ldots, Z_k = 1) = (1-p)^{k-1}p \]

- We define a r.v. \( L \), \( P(L=k) = P(Z_1 = 0, Z_2 = 0, \ldots, Z_k = 1) \)

Geometric (p)

- We define a r.v. \( L \), \( P(L=k) = P(Z_1 = 0, Z_2 = 0, \ldots, Z_k = 1) = (1-p)^{k-1}p \)

Why is it called Geometric?

- \( E(L) = \frac{1}{p} \)

- \( \text{Var}(L) = \frac{1-p}{p^2} \)
Memoryless property

- $P(L > 3) = ?$
  
  What should the first digit be? 0 or 1?
  What should the second digit be? 0 or 1?
  What should the third digit be? 0 or 1?
  0 0 0
  What’s the probability of this sequence occurring?
  
  $$(1-p)^3$$

- $P(L > n) = ?$
  
  $$(1-p)^n$$

- $P(L > k+n | L > n) = ?$
  
  $$\frac{(1-p)^{k+n}}{(1-p)^n}$$
  
  $$= (1-p)^k = P(L>k)$$
Memoryless property

\[
P(L > k+n | L > n) = P(L > k)
\]
Question 3

- What’s the probability that the 3rd vote for C comes from the 10th voter?
- More generally, what’s the probability that the rth vote for C comes from the nth voter?

Negative Binomial distribution

- \( P(S_r = n) = \binom{n - 1}{r - 1} p^r (1 - p)^{n-r} \)
- Relationship with Geometric (p)

\[
0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 1
\]

\[
S_r = L_1 + L_2 + \ldots + L_r
\]
Question 1’

- Out of 1000 voters sampled, what’s the probability that 5 vote for C?

- Binomial distribution

\[ P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \]

\[ P(X = 5) = \binom{1000}{5} 0.001^5 0.999995 \]

Precision problem: calculator gives you 0
Voting for C is a rare event
Poisson approximation (of binomial)

\[ P(X = 5) = \binom{1000}{5} 0.001^5 0.999^{995} \]

- Approximated by Poisson distribution with mean
  \[ \lambda = np = 1000 \times 0.001 = 1 \]

\[ p_Y(k) = \frac{\lambda^k e^{-\lambda}}{k!} \]

\[ P(X=5) = p_Y(5) = \frac{1^5 e^{-1}}{5!} = 0.00307 \]

\[ E(Y) = \text{Var}(Y) = \lambda \]

---

Poisson (\( \lambda \))

\[ p_Y(k) = \frac{\lambda^k e^{-\lambda}}{k!} \]

\[ E(Y) = \text{Var}(Y) = \lambda \]

Siméon Denis Poisson
1781 – 1840
French mathematician, geometer and physicist