How To Not Predict An Election

Probability and Statistics for Data Science

Carlos Fernandez-Granda

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Motivation

In probabilistic modeling

- Independence assumptions are unavoidable
- Ignoring some dependencies can be catastrophic

United States presidential election

- Indirect election, citizens of the US cast ballots for *electors* in the Electoral College
- These electors vote for the President and Vice President
- Number of electors per state = members of Congress (Washington D.C. gets 3)
- All electors in a state are assigned to candidate who wins the state (except in Maine and Nebraska)

Cartoon election: 51 states, one elector each

Probabilistic modeling

Result in state i modeled by Bernoulli random variable

$$\tilde{s}_i = \begin{cases} 1 & \text{if Republican candidate wins} \\ 0 & \text{otherwise} \end{cases}$$

Election is determined by the sum

$$\sum_{i=1}^{51} \tilde{s}_i$$

If it is larger than 25, the Republican wins

Joint pmf

To compute the probability

$$\mathrm{P}\left(\text{Republican wins}\right) = \mathrm{P}\left(\sum_{i=1}^{51} \tilde{s}_i > 25\right)$$

we need the joint pmf of $\tilde{s}_1$, $\tilde{s}_2$, ..., $\tilde{s}_{51}$

Number of entries? $2^{51} - 1 \geq 10^{15}$!

We need assumptions!

Independence

If states are independent, what do we need to estimate?

Only 51 marginal pmfs

Plan:

1. Estimate marginal pmf of each state
2. Aggregate them

Toy model

Probability of Republican winning state i depends only on rural turnout $\tilde{r}_i$

$$\mathrm{P}\left(\text{Republican wins state } i \,|\, \text{Rural turnout } = r\right)$$
$$= p_{\tilde{s}_i \,|\, \tilde{r}_i}(1 \,|\, r) := 0.6r + 0.1(1 - r)$$

If $\tilde{r}_i = 0$:

Urban voters dominate $\implies p_{\tilde{s}_i \,|\, \tilde{r}_i}(1 \,|\, r) = 0.1$

If $\tilde{r}_i = 1$:

Rural voters dominate $\implies p_{\tilde{s}_i \,|\, \tilde{r}_i}(1 \,|\, r) = 0.6$

Important

Probability of winning a state is **not** equal to fraction of voters in that state!

Real poll (Pennsylvania 2020)

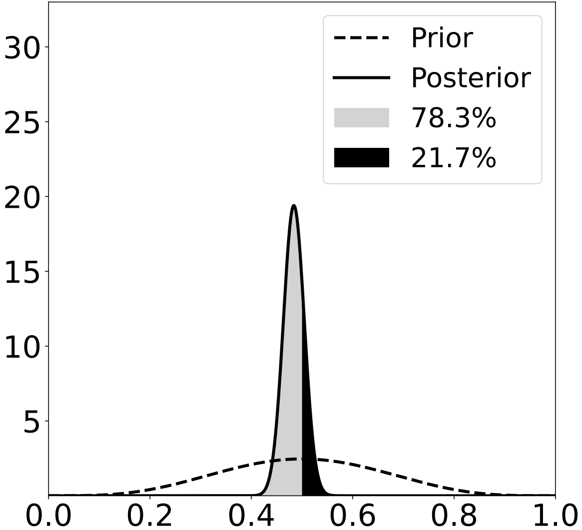
Data: 281 people intend to vote for Trump, 300 for Biden

Fraction of Biden voters: 51.6%

Probability that Biden wins in Pennsylvania?
(if data are truly i.i.d.)

Bayesian parametric model

Posterior pmf of fraction of Trump voters given 581 data points



P(Biden wins) = 78.3% (51.6% = Fraction of Biden voters)

Back to cartoon election

$$\mathrm{P}\left(\text{Republican wins} \,\middle|\, \text{Rural turnout } = r\right)$$
$$= p_{\tilde{s}_i \mid \tilde{r}_i}(1 \mid r) := 0.6r + 0.1(1 - r)$$

Marginal pmf of rural turnout in every state is uniformly distributed between 0 and 1

$$\begin{aligned}
\mathrm{P}\left(\text{Republican wins state } i\right) &= p_{\tilde{s}_i}(1) \\
&= \int_{r=0}^{1} f_{\tilde{r}_i}(r) p_{\tilde{s}_i \mid \tilde{r}_i}(1 \mid r)\, \mathrm{d}r \\
&= \int_{r=0}^{1} (0.5r + 0.1)\, \mathrm{d}r \\
&= 0.35
\end{aligned}$$

Aggregating probabilities

Under our assumptions,

$p_{\tilde{s}_1}(1) = p_{\tilde{s}_2}(1) = \cdots = p_{\tilde{s}_{51}}(1) = 0.35$

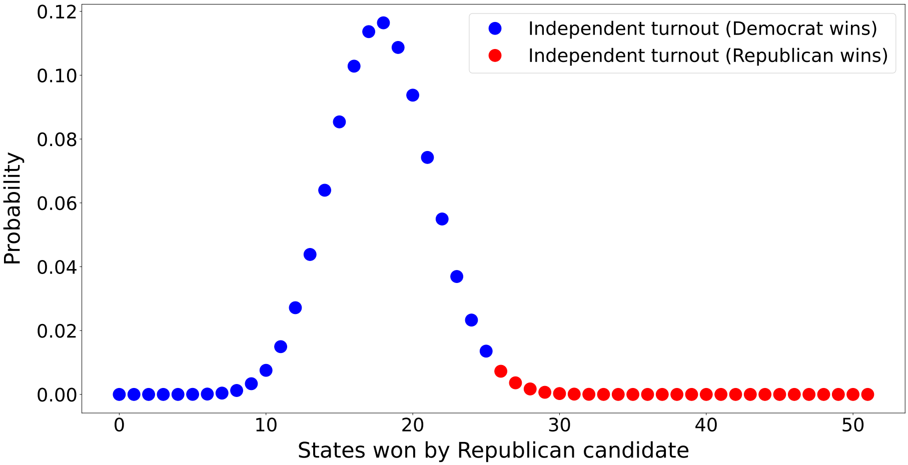
$\tilde{s}_1$, $\tilde{s}_2$, ..., $\tilde{s}_{51}$ are independent

What is the distribution of $\sum_{i=1}^{51} \tilde{s}_i$?

Binomial with parameters $n = 51$ and $\theta = 0.35$

$$\begin{aligned}
\mathrm{P}\left(\text{Republican wins}\right) &= \mathrm{P}\left(\sum_{i=1}^{51} \tilde{s}_i > 25\right) \\
&= \sum_{i=26}^{51} \binom{51}{i} 0.35^i (1 - 0.35)^{51-i} \\
&= 0.014
\end{aligned}$$

Independent turnouts: $\mathrm{P}\left(\text{Republican wins}\right) = 1.4\%$



What could go wrong?

Fictitious ground truth

Rural turnout is highly dependent, it is the same in every state

Probabilistically, modeled as a single random variable $\tilde{r}$

Same marginal distribution as before: Uniform in $[0,1]$

Same conditional probability for each state

$$p_{\tilde{s}_i \mid \tilde{r}}(1 \mid r) := 0.6r + 0.1(1-r)$$

$$\begin{aligned}
\mathrm{P}\left(\text{Republican wins state } i\right) &= p_{\tilde{s}_i}(1) \\
&= \int_{r=0}^{1} f_{\tilde{r}}(r) p_{\tilde{s}_i \mid \tilde{r}}(1 \mid r)\,\mathrm{d}r \\
&= 0.35
\end{aligned}$$

Same as before!

Election results

$$
\begin{aligned}
&\mathrm{P}\left(\text{Republican wins}\right)\\
&= \mathrm{P}\left(\sum_{i=1}^{51} \tilde{s}_i > 25\right)\\
&= \int_{r=0}^{1} f_{\tilde{r}}(r)\mathrm{P}\left(\sum_{i=1}^{51} \tilde{s}_i > 25 \,|\, \tilde{r} = r\right) \mathrm{d}r\\
&= \int_{r=0}^{1} \sum_{i=26}^{51} \binom{51}{i} (0.6r + 0.1)^i (1 - (0.6r + 0.1))^{51-i} \,\mathrm{d}r
\end{aligned}
$$

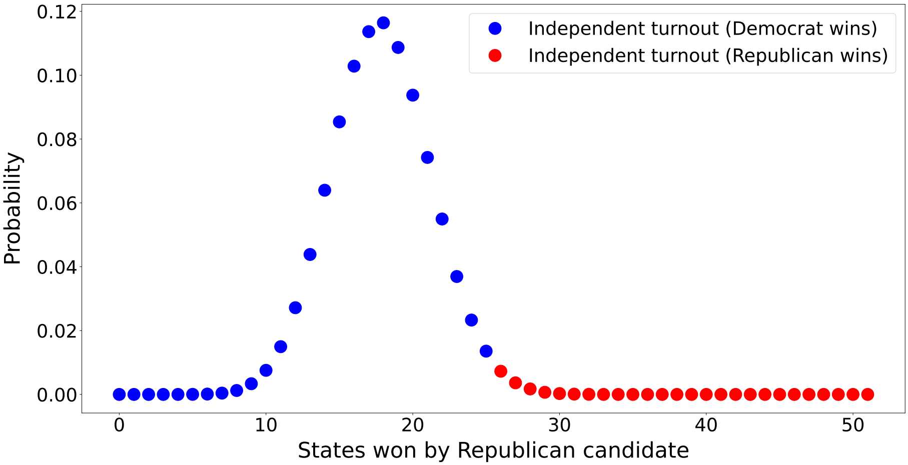
What if we don't know how to compute the integral?

Monte Carlo method

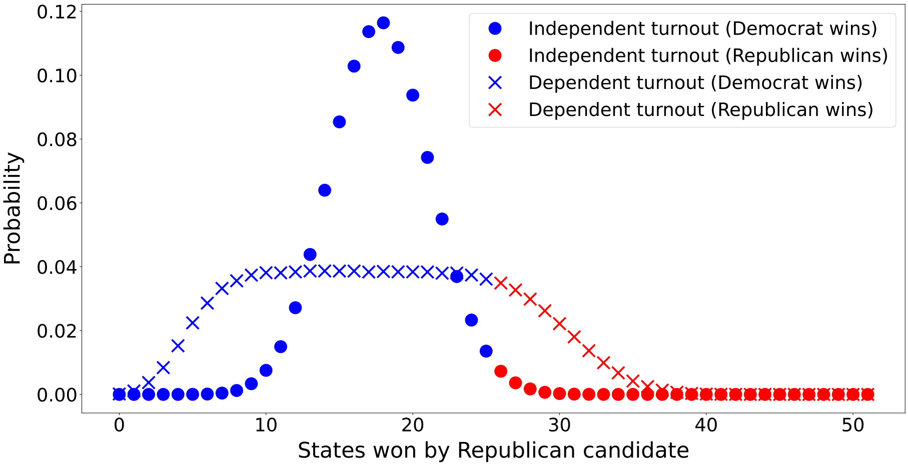
1. Simulate n elections:
   1.1 Generate rural turnout r
   1.2 Generate result in each state conditioned on r
   1.3 Check what candidate wins
2. Probability estimate = fraction of times each candidate wins

Many real election forecasts are more complicated versions of this

Independent turnouts: $\mathrm{P}\,(\text{Republican wins}) = 1.4\%$



Dependent turnouts: $\mathrm{P}\,(\text{Republican wins}) = 20.4\%$



What have we learned?

In probabilistic modeling

- Independence assumptions are unavoidable
- Ignoring some dependencies can be catastrophic