Poisson Approximation

MATH/STAT 394: Probability I
Summer 2021 A Term

Introduction to Probability
D. Anderson, T.Seppäläinen, B. Valkó

§ 4.4

Aaron Osgood-Zimmerman

Department of Statistics

Logistics

- course evaluation: https://uw.iasystem.org/survey/245028 (*please consider filling this out to improve the course and my teaching for future students!*)
- HW5 due tomorrow, Tuesday, at 11:59am (noon) PST so we can release HW5 solutions for you to review before the final is due
- Final
  - will primarily cover material since the midterm, but you may need leverage knowledge you learned per-midterm
  - will be available after today's lecture
  - will be due Wednesday July 21 at 11:59pm
  - unlimited time allowed during that window
- Last day of lecture will be Q+A (basically extra office hours)
- I have posted a review lecture deck that you can look at beforehand if you like

Outline

Practice solution

Practice

Suppose we interviewed 400 people and 100 of them liked spinach

Find a 90% confidence interval for the true probability that people like spinach assuming that we amy call the same person twice (sampling with replacement)

Solution

- We seek to find $\varepsilon > 0$ s.t.

$$\mathbb{P}(|p - \hat{p}| \leq \varepsilon) \geq 0.9$$

- Using that $\mathbb{P}(|p - \hat{p}| \leq \varepsilon) \geq 2\Phi(2\varepsilon\sqrt{n})$, this amounts to find ε s.t.

$$\begin{aligned} 2\Phi(2\varepsilon\sqrt{n}) - 1 \geq 0.9 &\Leftrightarrow \Phi(2\varepsilon\sqrt{n}) \geq 0.95 \\ &\Leftrightarrow 2\varepsilon\sqrt{n} \geq 1.645 \Leftrightarrow 40\varepsilon \geq 1.645 \Leftrightarrow \varepsilon \geq 0.041 \end{aligned}$$

- Therefore a 90% confidence interval for p is (given that $\hat{p} = 1/4$)

$$[0.25 - 0.041, 0.25 + 0.041] = [0.209, 0291]$$

Polling

Sampling without replacement

- In reality one would not call twice the same person
- In other words the variable would not be a binomial but a hypergeometric distribution
- Would our approximation still works?

Binomial limit of the hypergeometric distribution

Binomial limit of the hypergeometric distribution

- Consider picking n people from a population of size N with N_A people linking spinach and $N - N_A$ people who do not like spinach
- Let X be the number of people you sampled that liked spinach $X \sim \text{Hypergeom}(N, N_A, n)$
- Consider $N \to +\infty$ and $N_A \to +\infty$ such that $N/N_A = p$ remains constant
- Then $\mathbb{P}(X = k) \to \binom{n}{k} p^k (1-p)^k$, i.e., X tends to have a binomial distribution[1]
- So a normal approximation could again be used for polling a large population using sampling without replacement

[1] See backup slides

Recap

Normal Approximation

- A standardized binomial can be approximated by a standard normal dist. for $np(1-p)$ not too small
$$\mathbb{P}\left(a \leq \frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}(S_n)}} \leq b\right) \approx \Phi(b) - \Phi(a)$$
where $S_n \sim \mathrm{Bin}(n, p)$ and Φ is the c.d.f. of $Z \sim \mathcal{N}(0, 1)$

Confidence interval

- For $X \sim \mathrm{Ber}(p)$ and $\hat{p} = S_n/n$ an estimate of p, we saw that
$$\mathbb{P}(|p - \hat{p}| \leq \varepsilon) \geq 2\Phi(2\varepsilon\sqrt{n}) - 1$$
- A confidence interval of level e.g. 95% consists in finding ε s.t.
$$\mathbb{P}(|p - \hat{p}| \leq \varepsilon) \geq 95\%$$
- To do that, compute $z_{95\%}$ s.t. $2\Phi(z_{95\%}) - 1 = 95\%$ (here $z_{95\%} = 1.96$)
- Then an $\varepsilon > 0$ s.t.
$$2\varepsilon\sqrt{n} \geq z_{95\%}$$
gives you a confidence interval!

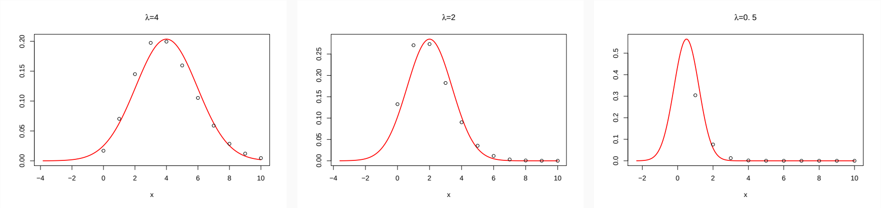
Outline

Poisson approximation

Motivation

- We have seen limits of distribution when p is not too close to 0 or 1
- What happens if the event is extremely rare, i.e., $p \ll 1$?

Bad normal approximation



Bin$(100, \lambda/100)$ and its normal approximation.

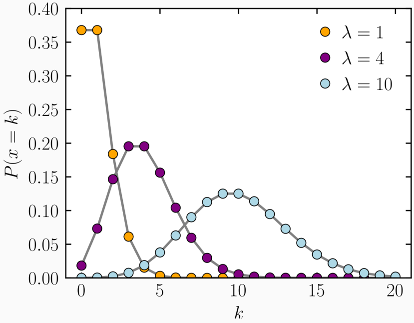
Poisson distribution

Definition

A r.v. has a Poisson dist. with param. $\lambda > 0$ if it has a p.m.f.

$$\mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad \textit{for } k \in \{0, 1, 2, \ldots\}$$

We denote it $X \sim \text{Poisson}(\lambda)$



p.m.f. of $X \sim \text{Poisson}(\lambda)$

Poisson Distribution

Properties

- From the quiz of lecture 19, if $X \sim \text{Poisson}(\lambda)$, then

$$\mathbb{E}[X] = \lambda \quad \text{Var}(X) = \lambda$$

- Typically models rare events

Poisson approximation of the binomial

Lemma

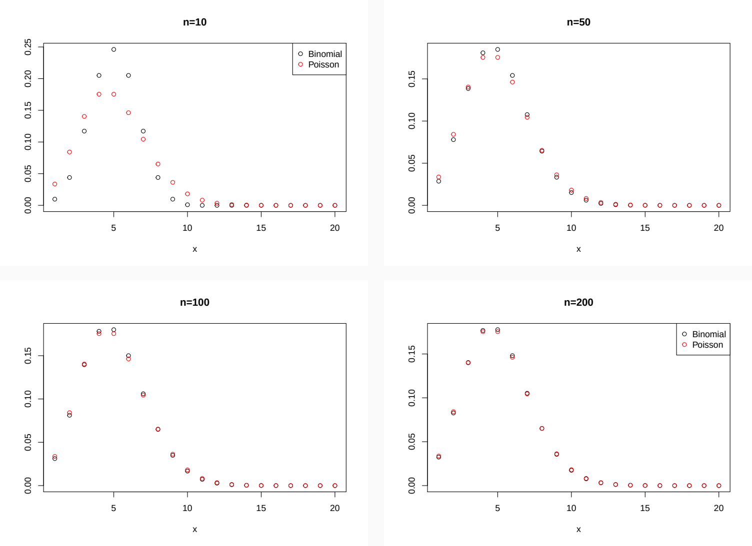
Let $\lambda > 0$*, consider* $S_n \sim \mathrm{Bin}(n, \lambda/n)$ *for* $n > \lambda$*.*

$$\lim_{n \to +\infty} \mathbb{P}(S_n = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

Interpretation

If S_n counts the number of successes of n independent trials and the mean $\mathbb{E}[S_n] = \lambda$ does not change with n, then as $n \to +\infty$. the dist. of S_n approaches the dist. of a Poisson dist.

Poisson approximation to binomial



$$p = \lambda/n,\ \lambda = 5$$

Poisson approximation of the binomial

Lemma
Let $\lambda > 0$, consider $S_n \sim \mathrm{Bin}(n, \lambda/n)$ for $n > \lambda$.

$$\lim_{n\to+\infty} \mathbb{P}(S_n = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

Proof

$$\begin{aligned}
\mathbb{P}(S_n = k) &= \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \frac{n(n-1)\dots(n-k+1)}{k!}\frac{\lambda^k}{n^k}\left(1-\frac{\lambda}{n}\right)^n\frac{1}{(1-\lambda/n)^k} \\
&= \frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n\left[1\cdot\left(1-\frac{1}{n}\right)\cdot\left(1-\frac{2}{n}\right)\dots\left(1-\frac{k-1}{n}\right)\right]\frac{1}{(1-\lambda/n)^k} \\
&\underset{n\to+\infty}{\longrightarrow} \frac{\lambda^k}{k!}e^{-\lambda}\cdot 1\cdot 1
\end{aligned}$$

where we used that $\lim_{n\to+\infty}(1+x/n)^n = e^x$

Poisson approximation of the binomial

Great, but what if n is finite?

Lemma
Let $X \sim \text{Bin}(n, p)$ *and* $Y \sim \text{Poisson}(np)$ *then for any* $A \subset \{0, 1, 2, \ldots\}$,

$$|\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)| \leq np^2$$

Interpretation

- When approximating a binomial by a Poisson r.v. you'll make an error of at most np^2
- So if $np^2 \ll 1$, then the Poisson dist. is a good approx. of the binomial
- So if p is very small (rare events), the Poisson dist. is a good approx. of the binomial

Poisson approximation of the binomial

Exercise

Let $X \sim \text{Bin}(10, 1/10)$. Compare the Poisson and normal approx. of $\mathbb{P}(X \leq 1)$

Solution

- The exact value is

$$\mathbb{P}(X \leq 1) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) = \left(\frac{9}{10}\right)^{10} + 10 \cdot \frac{1}{10}\left(\frac{9}{10}\right)^{9} \approx 0.7631$$

- $np^2 = 10 \cdot (1/10)^2 = 0.1$ so the Poisson approx. is 0.1 close to the exact value
- Here $\mathbb{E}[X] = 10 \cdot 1/10 = 1$ so the Poisson approx. is $Y \sim \text{Poisson}(1)$ which gives

$$\mathbb{P}(Y \leq 1) = \mathbb{P}(Y = 0) + \mathbb{P}(Y = 1) = e^{-1}\frac{1^0}{0!} + e^{-1} \cdot \frac{1^1}{1!} \approx 0.7358$$

- On the other hand a normal approx. would give ($np = 1$, $np(1-p) = 0.9$)

$$\begin{aligned}\mathbb{P}(X \leq 1) = \mathbb{P}(X \leq 3/2) &= \mathbb{P}\left(\frac{X-1}{\sqrt{9/10}} \leq \frac{3/2-1}{\sqrt{9/10}}\right) \\ &\approx \mathbb{P}\left(\frac{X-1}{\sqrt{9/10}} \leq 0.53\right) \approx \Phi(0.53) = 0.7019\end{aligned}$$

  where in the first line, we used that X is discrete and that the normal approx. will be more accurate when looking at the middle of the interval rather than the integer
- Here $np(1-p) = 9/10 \ll 10$, so, the normal approx. is not expected to be great

Poisson dist. as a model for rare events

Motivation

- Beyond being used as an approx., the Poisson dist. can be directly use to model rare events

Lemma (Poisson modeling of rare events)
Assume that a r.v. X counts occurrences of rare events that are not strongly dependent on each other.

Then the dist. of X can be approx. as $X \sim \text{Poisson}(\lambda)$ for $\lambda = \mathbb{E}[X]$

Poisson dist.

Exercise

Suppose a factory experiences on average 3 accidents per month. What is the proba. that there are at most 2 accidents a given month?

Solution

- Accidents are a priori rare events
- Under this assumption (rare events), we model the number of accidents as $X \sim \text{Poisson}(\lambda)$ with $\lambda = \mathbb{E}[X] = 3$
- This gives

$$\mathbb{P}(X \leq 2) = e^{-3}\frac{3^0}{0!} + e^{-3}\frac{3^1}{1!} + e^{-3}\frac{3^2}{2!} \approx 0.423$$

Practice next lecture

Practice

Assume that the prob. of at least one typo in the slides is 0.2.

What is the prob. that you find at least 2 typos?

Outline

Binomial limit of the hypergeometric distribution

- Consider picking n people from a population of size N with N_A people linking spinach and $N - N_A$ people who do not like spinach
- Let X be the number of people you sampled that liked spinach $X \sim \text{Hypergeom}(N, N_A, n)$

$$\mathbb{P}(X = k) = \frac{\binom{N_A}{k}\binom{N-N_A}{n-k}}{\binom{N}{n}} = \frac{\frac{(N_A)_k}{k!}\frac{(N-N_A)_{n-k}}{(n-k)!}}{\frac{(N)_n}{n!}} = \binom{n}{k}\frac{(N_A)_k(N-N_A)_{n-k}}{(N)_n}$$

where $(a)_k = a(a-1)\dots(a-k+1) = a!/(a-k)!$

Then

$$\frac{(N_A)_k(N-N_A)_{n-k}}{(N)_n} = \frac{N_A(N-1)\dots(N_A-k+1)}{N(N-1)\dots(N-k+1)} \cdot \frac{(N-N_A)(N-N_A-1)\dots(N-N_A-n+k+1)}{(N-k)(N-k-1)\dots(N-n+1)}$$

$$= \left(\frac{N_A}{N}\right)^k \prod_{i=1}^{k} \frac{(1-\frac{i-1}{N_A})}{\left(1-\frac{i-1}{N}\right)} \left(1-\frac{N_A}{N}\right)^{n-k} \prod_{i=k+1}^{n} \frac{\left(1-\frac{i-k-1}{N-N_A}\right)}{\left(1-\frac{i-1}{N}\right)}$$

$$\to p^k(1-p)^k$$

- Consider $N \to +\infty$ and $N_A \to +\infty$ such that $N/N_A = p$ remains constant
- Then $\frac{(N_A)_k(N-N_A)_{n-k}}{(N)_n} \to p^k(1-p)^{n-k}$
- Thus $\mathbb{P}(X = k) \to \binom{n}{k}p^k(1-p)^k$, i.e., X tends to have a binomial