Chapter 4 and beyond: Learning about distributions from finite data

Part 3, Fri 15 July

Jess Kunke

MATH/STAT 394: Probability I (Summer 2022 A-term)

Outline

Outline

Final exam logistics

- Next Wed July 20th, 1.5 hours starting at 9am
- In person in CMU 230 (our classroom)
- Please bring your own pens, pencils, and/or erasers
- No calculators, books, phones, computers, etc.
- I will provide scratch/extra paper for you
- You can bring a one-sided 8.5x11 (or smaller) sheet of paper with handwritten notes
- The exam is not explicitly cumulative (e.g. no urns with different color balls and no Bayes, and the heavy lifting will be on the new concepts, but you still need to know the basics about probability when dealing with coin flips, sampling, etc)

Simplifying and showing your work

- When computing the integrals, plug in the numbers and simplify as much as you can; fully simplifying fractions is not necessary
- Show your steps, define your notation/variables
- Not all problems will require the same amount of steps/work, but there should always be more you can show than simply the answer

Exam tips

During the exam

- Skim through the exam to see what the problems probably involve
- Make sure you give some time to each problem; I can only give partial credit for work that you show

Preparation:

- Practice problems are posted; they do not cover everything but they are relevant practice for the exam
- Make sure to also review HW3, HW4, and lecture notes/examples
- I highly encourage modifying the homework and lecture problems and trying to solve them; great learning tool as well as study method

Final exam topics

Chapter 3:

- discrete and continuous RVs
- pmf vs pdf
- definition of cdf in general; properties of the cdf
- of the pmf, pdf, and cdf, which functions are probabilities?
- calculating cdf from pmf vs pdf, and vice versa
- computing probabilities of single values and of intervals (including when the left or right endpoint is +/-infty)

Final exam topics

Chapter 3:

- discrete and continuous RVs
- pmf vs pdf
- definition of cdf in general; properties of the cdf
- of the pmf, pdf, and cdf, which functions are probabilities?
- calculating cdf from pmf vs pdf, and vice versa
- computing probabilities of single values and of intervals (including when the left or right endpoint is +/-infty)
- verifying that a function is a pdf; finding the normalization constant (C)
- be able to take derivatives/integrals of polynomials ($a + bx + cx^2 + ...$), rational powers ($x^{a/b}$) and exponentials (e^{ax})
- expectation (definition, linearity, expectation of an indicator RV, expectation of a transformation $g(X)$ of a RV X, expectation of a product of independent RVs)
- variance (definition, alternative formulation as $\mathrm{E}[X^2] - \mathrm{E}[X]^2$, properties, variance of a sum of independent RVs)

Final exam topics

Distributions:

- Bernoulli, binomial, uniform, exponential, normal (standard and general), Poisson (covered today)

Not covered:

- Median, quantiles, geometric, hypergeometric
- Lecture on Mon July 18th (additional topics)

Final exam topics

Chapter 4 and beyond:

- deriving the pmf/pdf/cdf of $Y = g(X)$ if you know the distribution of X (discrete vs continuous, invertible or not, cdf method)
- mean and variance of the sample mean
- Markov's and Chebyshev's inequalities (when/how to apply them, what their limitations are)
- Weak Law of Large Numbers and its relationship to Chebyshev's inequality
- computing probabilities from the normal distribution using its symmetry and a table
- standardizing variables
- transforming a standard normal into a normal with mean mu and s.d. sigma, and vice versa
- computing the probability that a variable is within some interval or within so many standard deviations
- using the normal approximation for the binomial
- today's material (continuity correction, confidence intervals, Poisson approximation)

Outline

Continuity correction

- Consider that you want to simply approximate $P(S_n = k)$ for some fixed k. What is the problem you'll run into?

Continuity correction

- Consider that you want to simply approximate $P(S_n = k)$ for some fixed k. What is the problem you'll run into?
- The CLT (normal approximation) will not be useful as is, since you would get something like

$$P(S_n = k) = P\left(a_{n,k} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq a_{n,k}\right) \approx \Phi(a_{n,k}) - \Phi(a_{n,k}) = 0$$

  for $a_{n,k} = \frac{k-np}{\sqrt{np(1-p)}}$.
- To circumvent the issue, note that, since S_n only takes integer values,

$$P(S_n = k) = P(k - 1/2 \leq S_n \leq k + 1/2)$$

  and we can rather apply the CLT on the right hand side to have something meaningful

Continuity correction

The method for continuity correction:

- If k_1, k_2 are integers, since $S_n \sim \text{Bin}(n, p)$ can only take integer values, then
$$P(k_1 \leq S_n \leq k_2) = P(k_1 - 1/2 \leq S_n \leq k_2 + 1/2)$$
- Applying the CLT on the second interval,
$$P(k_1 - 1/2 \leq S_n \leq k_2 + 1/2) \approx \Phi\left(\frac{k_2 + 1/2 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k_1 - 1/2 - np}{\sqrt{np(1-p)}}\right)$$
- This correction gives better approximations
- Sometimes this correction is necessary or valuable, and sometimes it is negligible

Continuity correction

Example

We roll a pair of fair six-sided dice 10,000 times. Estimate the probability that the number of times we get snake eyes (two ones) is between 280 and 300 using the continuity correction.

From theory to practice

The thought process:

1. Collect some data
   - e.g. repeat the flips of a coin
2. Make some assumptions about the data
   - e.g. the flips are independent and identically distributed
3. Consider a theoretical model for your experiment
   - e.g. the flips are Ber(p) RVs with some unknown p
4. Use the theory to assess what the unknown parameter p could be given your data
   - e.g. use the CLT to assess how far the empirical frequency of success is from the unknown probability of success p

$\rightarrow$ Here we will develop the last part, i.e., how to assess what could be the true probability of success p.

Confidence intervals: motivation

We return to a question we've been exploring over the past few lectures:

- Suppose we have a biased coin and we want to know $p = P(\text{getting a tail})$
- How can we know if our observed frequency of tails $\hat{p} = \frac{S_n}{n} = \frac{X_1 + \ldots + X_n}{n}$ is a good estimate of p?
- We want to estimate for some $\varepsilon > 0$ fixed, the prob. $P(|\hat{p} - p| \leq \varepsilon)$

Confidence Intervals

Let us reformulate $P(|\hat{p} - p| \leq \varepsilon)$ in terms of the central limit theorem:

Example

How many times should you flip a coin with unknown prob. of success p such that the estimate $\hat{p} = \frac{S_n}{n}$ is within 0.05 of the true p with prob. at least 0.99?

Confidence intervals

Definition

Let $X \sim \text{Ber}(p)$ and $\hat{p}$ be an estimator of p.
A **confidence interval at level α** of p is an interval of the form

$$[\hat{p} - \varepsilon, \hat{p} + \varepsilon] \quad \text{s.t.} \quad P(p \in [\hat{p} - \varepsilon, \hat{p} + \varepsilon]) \geq \alpha$$

which is equivalent to $P(|\hat{p} - p| \leq \varepsilon) \geq \alpha$.

Question: what is the random variable here? Where is the randomness?

Example

We repeat a trial 1000 times and observe 450 successes.
Find a 95% confidence interval for the true success probability p.

Sampling without replacement

- In the previous polling, by considering a binomial approximation, we are considering a sampling with replacement (see lecture 12)
- 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 work?

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 $X \sim \text{Hypergeo}(N, N_A, n)$
- Consider $N \to +\infty$ and $N_A \to +\infty$ such that $N/N_A = p$ remains constant
- Then $P(X = k) \to \binom{n}{k} p^k (1-p)^k$, 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
- We saw this with the DNA vs scrabble tiles example before

[1] See backup slides

Outline

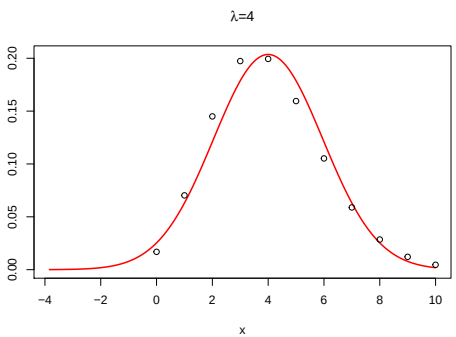
Poisson approximation: motivation

- 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., $p \ll 1$?

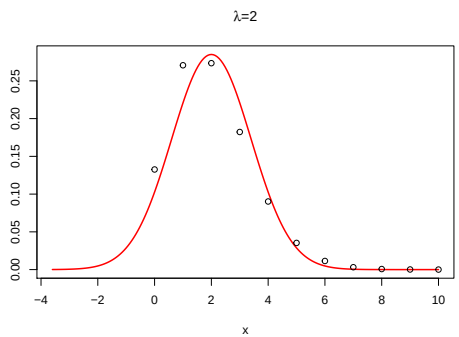
Demo: normal approximation poor for small p

Remember the rule of thumb: normal approx. is fine if $np(1-p) > 10$

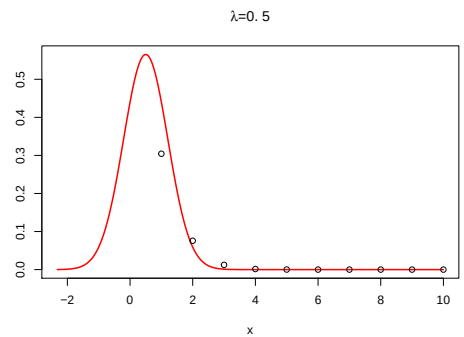
- Here $p = 0.04, 0.02, 0.005$
- Last time normal approx. worked in an example with $p = 1/36 = 0.02\overline{7}$, but $n = 10^4$ in that example; here $n = 100$
- With small enough p, the normal distribution has nonneglible probability on negative numbers whereas binomial RVs are strictly positive



(a) $\lambda = 4$



(b) $\lambda = 2$



(c) $\lambda = 0.5$

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

Approximation for very rare events

Motivation

- As for the normal approximation,
  we would like to understand what happens for very large n
- To model that the probability of success remains small,
  we consider a probability of success $p = \lambda/n$ for some λ
- Namely, we'll consider binomials of the form $S_n = \text{Bin}(n, \lambda/n)$
- Interpretation: for any n, $\text{E}[S_n] = \lambda$, which means that the probability of success is so small that even as n increases the expected number of successes does not increase

Poisson approximation of the binomial

Lemma

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

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

Proof:

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

where we used that $\lim_{n \to +\infty} (1 + x/n)^n = e^x$ (see additional slides for proof of this fact)

Poisson distribution

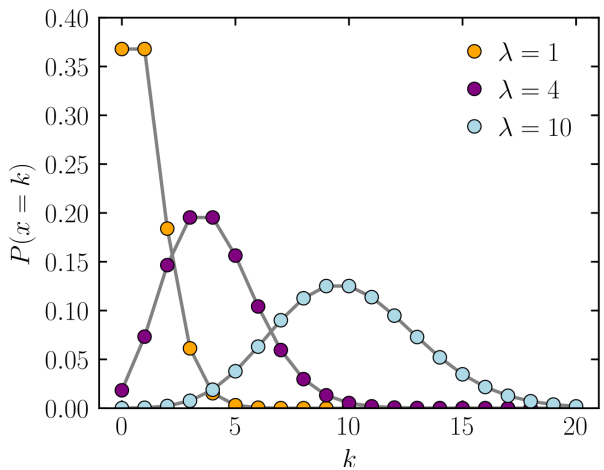
The distribution we obtained is known as a Poisson distribution:

Definition

A RV is a **Poisson** RV with rate parameter $\lambda > 0$ if it has a pmf

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

We denote it $X \sim \text{Poiss}(\lambda)$, and we have $\text{E}[X] = \lambda, \text{Var}(X) = \lambda$.



pmf of $X \sim \text{Poiss}(\lambda)$

Source: Wikipedia

Poisson approximation of the binomial

So the previous lemma can be interpreted as a Poisson approximation of the binomial:

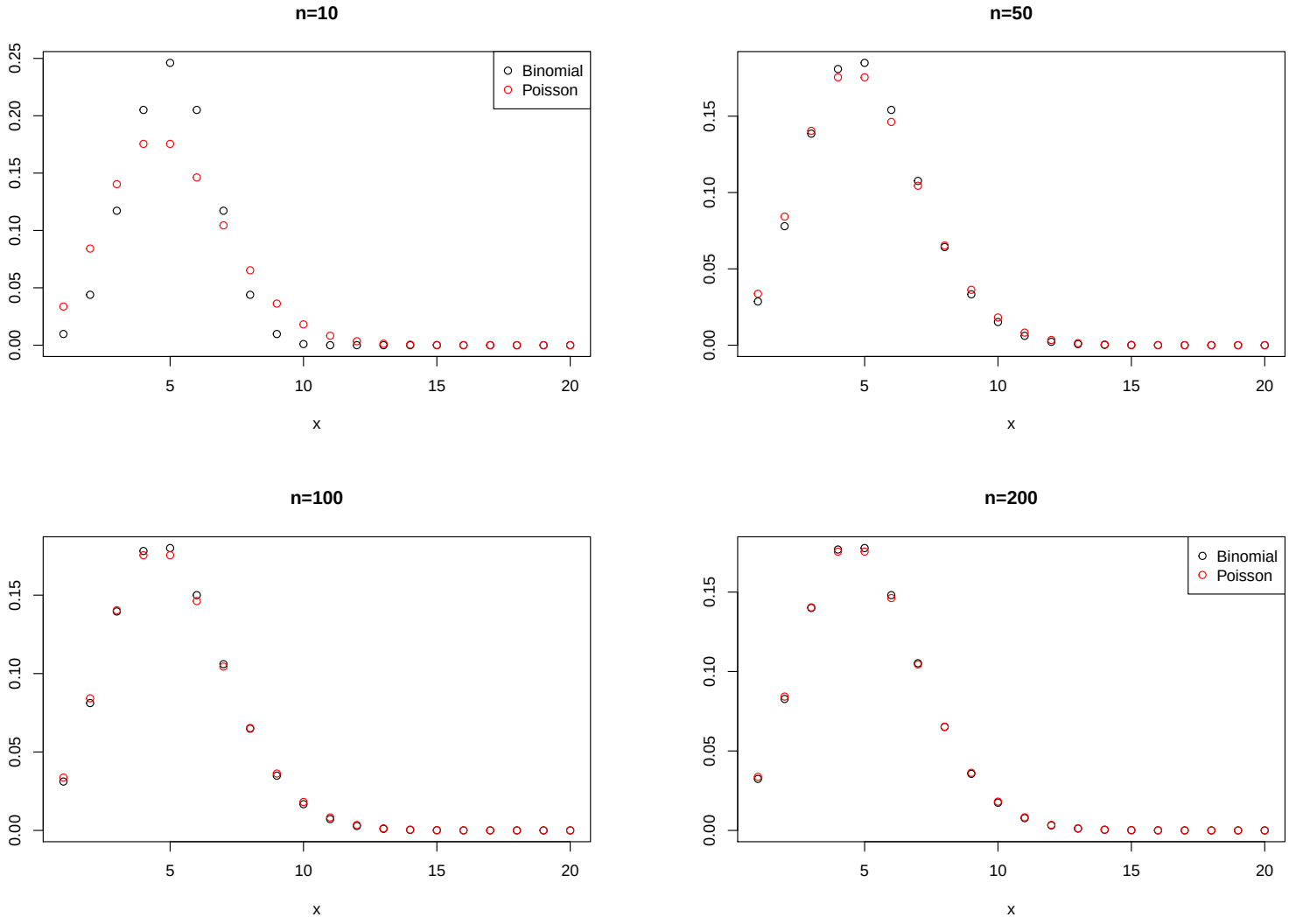
Lemma
Let $\lambda > 0$, consider $S_n \sim Bin(n, \lambda/n)$ for $n > \lambda$. Then

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

Interpretation

If S_n counts the number of successes of n independent trials and the mean $\mathrm{E}[S_n] = \lambda$ 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

Great, but what if n is finite?

Lemma

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

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

Poisson approximation of the binomial

Great, but what if n is finite?

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

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

Interpretation

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

Poisson approximation of the binomial

Example

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

Poisson dist. as a model for rare events

Motivation

- Beyond being used as an approximation, the Poisson distribution can be used directly to model rare events
- The Poisson distribution gives the probability of a given number of events (a nonnegative integer) occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event

Lemma (Poisson modeling of rare events)

If a RV X counts occurrences of rare events that are not strongly dependent on each other, then X is approximately distributed Poiss(λ) for $\lambda = E[X]$.

Poisson distribution

Example

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

Outline

Computing the maximum of $p(1-p)$

Computing the maximum of $p(1-p)$

- Let $f(p) = p(1-p)$
- Then $f'(p) = 1 - 2p$
- Therefore for $p \in [0, 1/2]$, $f'(p) \geq 0$ and for $p \in [1/2, 1]$, $f'(p) \leq 0$
- The maximum of f is then reached on $p = 1/2$, which gives

$$\max_{p \in [0,1]} f(p) = f(1/2) = 1/4$$

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 $X \sim \text{Hypergeo}(N, N_A, n)$

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

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 $P(X = k) \to \binom{n}{k}p^k(1-p)^k$, i.e., X tends to have a binomial distribution

Reminder

Lemma

The Taylor approximation of the logarithm is for any $|x| < 1$,

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots$$

- For any $|t| < 1$,
$$\frac{1}{1+t} = 1 - t + t^2 - t^3 + \dots$$
- Since $\ln(1+x) = \int_0^x \frac{1}{1+t} dt$, we get that for any $|x| < 1$,
$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots$$

Lemma

We have for any $x \in \mathbb{R}$,

$$\lim_{s \to +\infty} (1 + x/s)^s = e^x$$

For any $s \in \mathbb{R}$, $s > 0$, $(1 + x/s)^s = \exp(s \ln(1 + x/s))$. Using the Taylor expansion of ln around 1, we have that $\ln(1 + x/s) = x/s + o(x/s)$ for x small enough, where o is a continuous function such that $\lim_{y \to 0} o(y)/y = 0$. Therefore we get

$$\lim_{s \to 0}(1+x/s)^s = \lim_{s \to +\infty} \exp(s(x/s+o(x/s))) = \exp(x) \lim_{y \to 0} \exp(o(yx)/y)) = \exp(x),$$

by definition of o and continuity of exp for the last equation.