Discrete Random Variable Zoo II

CSE 312 26Su

Lecture 12

Announcements

Quiz 3 grades are out

Midterm on Wednesday during lecture

- Bring your student ID, something to write with
- Bring your own reference sheet + use the one that's given
- See all the details on course website

Where Are We?

Last time:

- Zoo of Discrete RVs

Today:

- Finish Zoo

Scenario: Negative Binomial

Run <u>independent</u> trials with probability $\boldsymbol{p}$. How many trials do you need *until $\boldsymbol{r}$ successes*?

Example

You're playing a carnival game, and there are r little kids nearby who all want a stuffed animal. You can win a single game (and thus win one stuffed animal) with probability p (independently each time) How many times will you need to play the game before every kid gets their toy?

Try it

Run <u>independent</u> trials with probability $\boldsymbol{p}$. How many trials do you need *until* $\boldsymbol{r}$ *successes*?

X is the number of trials till (and including) the r'th success

What is the **support** of X?

What's the **PMF**?

i.e., what is the probability it takes exactly k trials till the r'th success?

Fill out the poll everywhere: pollev.com/jolie312

Try it

Run <u>independent</u> trials with probability $\boldsymbol{p}$. How many trials do you need *until $\boldsymbol{r}$ successes*?

X is the number of trials till (and including) the r'th success

What is the **support** of X? $\Omega_X = \{r, r+1, r+2, \dots\}$

What's the **PMF**?

i.e., what is the probability it takes exactly k trials till the r'th success?

Negative Binomial Analysis

Run <u>independent</u> trials with probability p
X is the number of trials till (and including) the r'th success

What's the **PMF**? Well how would we know $X = k$?

Negative Binomial Analysis

What’s the **PMF**? Well how would we know $X = k$?

Of the first $k-1$ trials, $r-1$ must be successes.
And trial k must be a success.

1. We want **exactly $\boldsymbol{r-1}$ of the first $\boldsymbol{k-1}$ to be successes** – this sounds like a binomial! It’s the $p_Y(r-1)$ where $Y \sim \text{Bin}(k-1, r-1)$:
$\binom{k-1}{r-1}(1-p)^{k-1-(r-1)}p^{r-1} = \binom{k-1}{r-1}(1-p)^{k-r}p^{r-1}$

2. Multiply by p, probability **$\boldsymbol{k}$’th trial is success**

Total: $p_X(k) = \binom{k-1}{r-1}(1-p)^{k-r}p^r$

Negative Binomial Analysis

X is the number of trials till we see r successes

To see r successes:

We do trials until we see success 1.

Then do trials until success 2.

…do trials until success r.

What's the **expectation** and **variance** (hint: linearity)?

How can we write X as a sum of random variables?

Fill out the poll everywhere: pollev.com/jolie312

Negative Binomial Analysis

X is the number of trials till we see r successes

To see r successes:

We do trials until we see success 1.

Then do trials until success 2.

…do trials until success r.

The total number of flips is…the sum of geometric random variables!

Negative Binomial Analysis

Let $Z_1, Z_2, \ldots, Z_r$ be independent copies of $\mathrm{Geo}(p)$

Z_i are called "independent and identically distributed" or "i.i.d.'

Because they are independent...and have identical pmfs.

$X \sim \mathrm{NegBin}(r, p)$ $X = Z_1 + Z_2 + \cdots + Z_r$.

$$\mathbb{E}[X] = \mathbb{E}[Z_1 + Z_2 + \cdots Z_r] = \mathbb{E}[Z_1] + \mathbb{E}[Z_2] + \cdots + \mathbb{E}[Z_r] = r \cdot \frac{1}{p}$$

Negative Binomial Analysis

Let $Z_1, Z_2, \ldots, Z_r$ be independent copies of $\mathtt{Geo}(p)$

$X \sim \mathrm{NegBin}(r, p)$ $X = Z_1 + Z_2 + \cdots + Z_r$.

$\mathrm{Var}(X) = \mathrm{Var}(Z_1 + Z_2 + \cdots + Z_r)$

Up until now we've just used the observation that $X = Z_1 + \cdots + Z_r$.

$= \mathrm{Var}(Z_1) + \mathrm{Var}(Z_2) + \cdots + \mathrm{Var}(Z_r)$ **because the $\boldsymbol{Z_i}$ are independent.**

$= r \cdot \frac{1-p}{p^2}$

Negative Binomial

$X \sim \text{NegBin}(\text{r}, \text{p})$

Parameters: r: the number of successes needed, p the probability of success in a single trial

X is the number of trials needed to get the r^{th} success.

PMF: $p_X(k) = \binom{k-1}{r-1}(1-p)^{k-r}p^r$

CDF: $F_X(k)$ is ugly, don't bother with it.

Expectation: $\mathbb{E}[X] = \frac{r}{p}$

Variance: $\text{Var}(\text{X}) = \frac{\text{r}(1-\text{p})}{p^2}$

Scenario: Hypergeometric

You have an urn with N balls, of which K are purple. You are going to draw n balls out of the urn **without** replacement. How many purple balls do we get in this sample?

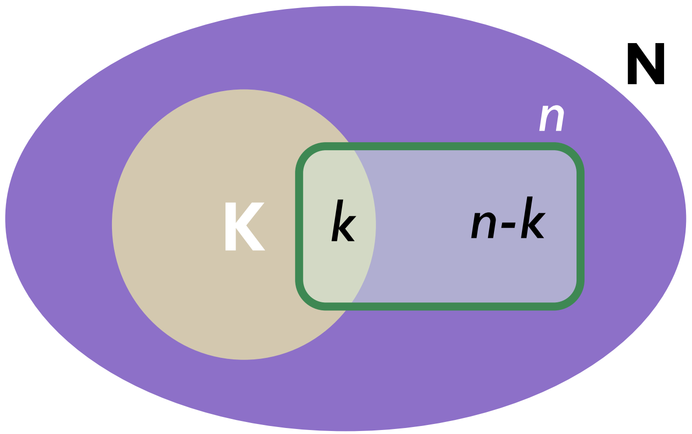
X is the number of purple balls in this sample

Hypergeometric: Analysis (PMF)

You have an urn with N balls, of which K are purple. You are going to draw n balls out of the urn **without** replacement. How many purple balls do we get in this sample?

<u>X is the number of purple balls in this sample</u>

If you draw out $\boldsymbol{n}$ balls, what is the probability you see $\boldsymbol{k}$ purple ones?



Hypergeometric: Analysis (PMF)

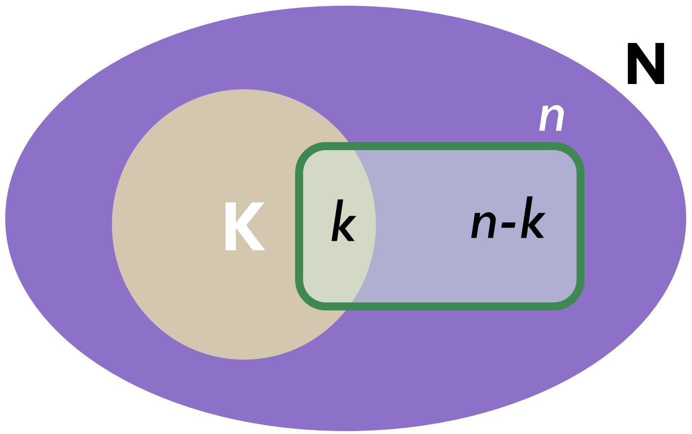
You have an urn with N balls, of which K are purple. You are going to draw n balls out of the urn **without** replacement. How many purple balls do we get in this sample?

X is the number of purple balls in this sample

If you draw out $\boldsymbol{n}$ balls, what is the probability you see $\boldsymbol{k}$ purple ones?

Of the K purple, we draw out k choose which k will be drawn

Of the $N-K$ other balls, we will draw out $n-k$, choose which $N-K-(n-k)$ will be removed.



Sample space all subsets of size n

$$\mathbb{P}(X=k)=\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Hypergeometric: Analysis (Expectation)

X is the number of purple balls in the sample

$X = D_1 + D_2 + \cdots + D_n$

Where D_i is the indicator that draw i is purple.

D_1 is 1 with probability K/N.

What about D_2?

$$\mathbb{P}(D_2 = 1) = \frac{K-1}{N-1} \cdot \frac{K}{N} + \frac{K}{N-1} \cdot \frac{K-N}{N} = \frac{K(K-N+K-1)}{N(N-1)} = \frac{K}{N}$$

In general $\mathbb{P}(D_i = 1) = \frac{K}{N}$

It might feel counterintuitive, but it's true!

Hypergeometric: Analysis

$\mathbb{E}[X]$

$= \mathbb{E}[D_1 + \cdots D_n] = \mathbb{E}[D_1] + \cdots + \mathbb{E}[D_n] = n \cdot \frac{K}{N}$

Can we do the same for variance?

No! The D_i are dependent. Even if they have the same probability.

Hypergeometric Random Variable

$X \sim \text{HypGeo}(N, K, n)$

X is the number of success balls drawn in the sample.

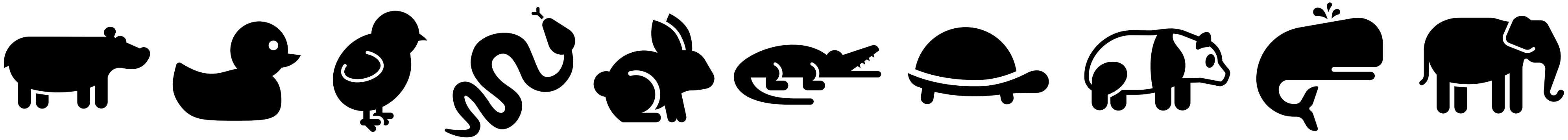
Parameters: A total of N balls in an urn, of which K are successes.
Draw n balls without replacement.

PMF: $p_X(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$

CDF: $\mathbb{E}[X] = \frac{nK}{N}$

Variance: $\text{Var}(X) = n \cdot \frac{K}{N} \cdot \frac{N-K}{N} \cdot \frac{N-n}{N-1}$

Zoo!



$X \sim \mathrm{Unif}(a, b)$

$$p_X(k) = \frac{1}{b - a + 1}$$

$$\mathbb{E}[X] = \frac{a + b}{2}$$

$$\mathrm{Var}(X) = \frac{(b - a)(b - a + 2)}{12}$$

$X \sim \mathrm{Ber}(p)$

$$p_X(0) = 1 - p; \quad p_X(1) = p$$

$$\mathbb{E}[X] = p$$

$$\mathrm{Var}(X) = p(1 - p)$$

$X \sim \mathrm{Bin}(n, p)$

$$p_X(k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

$$\mathbb{E}[X] = np$$

$$\mathrm{Var}(X) = np(1 - p)$$

$X \sim \mathrm{Geo}(p)$

$$p_X(k) = (1 - p)^{k-1} p$$

$$\mathbb{E}[X] = \frac{1}{p}$$

$$\mathrm{Var}(X) = \frac{1 - p}{p^2}$$

$X \sim \mathrm{NegBin}(r, p)$

$$p_X(k) = \binom{k - 1}{r - 1} p^r (1 - p)^{k-r}$$

$$\mathbb{E}[X] = \frac{r}{p}$$

$$\mathrm{Var}(X) = \frac{r(1 - p)}{p^2}$$

$X \sim \mathrm{HypGeo}(N, K, n)$

$$p_X(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

$$\mathbb{E}[X] = n\frac{K}{N}$$

$$\mathrm{Var}(X) = \frac{K(N - K)(N - n)}{N^2(N - 1)}$$

$X \sim \mathrm{Poi}(\lambda)$

$$p_X(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$\mathbb{E}[X] = \lambda$$

$$\mathrm{Var}(X) = \lambda$$

Zoo Takeaways

You can do relatively complicated counting/probability calculations much more quickly than you could week 1!

You can now explain why your problem is a zoo variable and save explanation on homework (and save yourself calculations in the future).

Don't spend extra effort memorizing…but be careful when looking up Wikipedia articles.

- The exact definitions of the parameters can differ (is a geometric random variable the number of failures before the first success, or the total number of trials including the success?)

Discrete Zoo of Random Variables

- **Uniform:** Every integer between a and b are equally likely
  $\text{Un}if(a,b)$
- **Bernoulli:** Whether there is success in one trial
  $\text{Ber}(p)$ is 1 with probability p and 0 otherwise
- **Binomial:** Number of successes in n independent trials
  $\text{Bin}(n,p)$ - n independent trials, probability p of success on each trial
- **Geometric:** Number of trials till first success
  $\text{Geo}(p)$ - probability p of success on each trial
- **Poisson:** Number of successes in a time interval
  $\text{Poi}(\lambda)$ - average number of successes in the time interval
- **Negative Binomial**: Number of trials till r'th success
  $\text{NegBin}(r,p)$ - probability p of success on each trial, want trials till the r'th success
- **Hypergeometric:** Number of successes when drawing a sample
  $\text{HypGeo}(N,K,n)$ - drawing a sample of n items from a set of N with K successes

Halfway Point!

What have we done over the past 4 weeks?

Counting

Combinations, permutations, indistinguishable elements, starts and bars, inclusion-exclusion…

Probability foundations

Events, sample space, axioms of probability, expectation, variance

Conditional probability

Conditioning, independence, Bayes' Rule

Refined our intuition

Especially around Bayes' Rule

What’s next?

Continuous random variables.

- So far our sample spaces have been countable. What happens if we want to choose a random real number?
- How do expectation, variance, conditioning, etc. change in this new context?
- Mostly analogous to discrete cases, but with integrals instead of sums.

Analysis when it’s inconvenient (or impossible) to exactly calculate probabilities.

- Central Limit Theorem (approximating discrete distributions with continuous ones)
- Tail Bounds/Concentration (arguing it’s unlikely that a random variable is far from its expectation)

A first taste of making predictions from data (i.e., a bit of ML)