Probability and Mathematical Statistics in Data Science

Lecture 10: 4.2: Waiting Times Section 4.3: Exponential Approximation

Waiting Times

- Say Ali keeps playing roulette, and betting on red each time. The waiting time of a red win is the number of spins until they see a red. Pr(Red) = 18/38

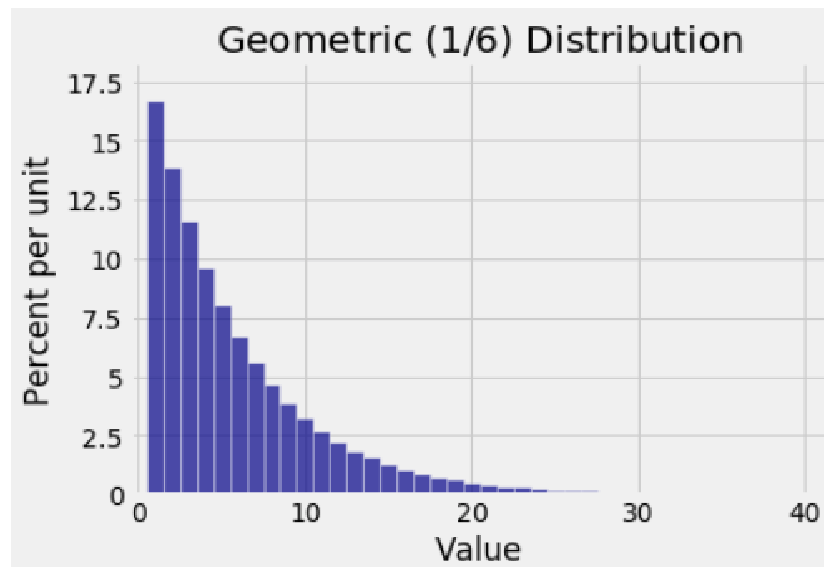
Q. What is the probability that Ali will wait for 4 spins before their first win? (That is, the first time the ball lands in red is the 4^{th} spin or trial)

This is an example of a **geometric distribution**. The number of trial until the **first** success

Geometric distribution

- Recall that T_1 has the **geometric distribution**, denoted $\boldsymbol{T_1 \sim Geom(p)}$ on $\{1, 2, 3, \ldots\}$, when we have k-1 failures, and then first success is on kth trial.
- $f(k) = P(T_1 = k) = P(FFF \ldots FS) =$
- $F(k) = P(T_1 \leq k) = 1 - P(T_1 > k) =$

Roll a die until *first* ace (1 spot):



Waiting time until Rth success

- Say we roll a 8 sided die.
- What is the chance that the first time we roll an eight is on the 11th try?

- What is the chance that it takes us 15 times until the 4th time we roll eight? (That is, the waiting time until the 4th time we roll an eight is 15)

$$= P(\_\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ S)$$

Waiting Times until the Rth Success

- The **negative binomial distribution** is based on an experiment satisfying the following conditions:

1. The experiment consists of a sequence of independent trials.
2. Each trial can result in either a success (S) or a failure (F).
3. The probability of success is constant from trial to trial, so $P(S \text{ on trial } i) = p$ for $i = 1, 2, 3, \ldots$.
4. The experiment continues (trials are performed) until a total of r successes have been observed, where r is a specified positive integer.

Waiting Times until the Rth Success

- The random variable of interest is X = number of failures that precede the *r*th success. Possible values of *X* are 0, 1, 2, ...
- Consider negative binomial with X=7 and r =3.
- P(X=7)? => 10^{th} trial must be a success (S) and there must be 2 *S*'s among 9 trials. Thus

$$nb(7;3,p) = \left\{\binom{9}{2}\cdot p^2(1-p)^7\right\}\cdot p = \binom{9}{2}\cdot p^3(1-p)^7$$

> The pmf of the negative binomial rv X with parameters r = number of S's and $p = P(S)$ is
>
> $$nb(x;r,p) = \binom{x+r-1}{r-1}p^r(1-p)^x \quad x = 0, 1, 2, \ldots$$

Example

- A pediatrician wishes to recruit 5 couples, each of whom is expecting their first child, to participate in a new natural childbirth regimen.
- Let p = P(a randomly selected couple agrees to participate)
- If p=0.2, what is the probability that 15 couples must be asked before 5 are found who agree to participate?
- That is with S = {agrees to participate], what is the probability that 10 *F*'s occur before the fifth *S*?

Example

$$nb(x; r, p) = \binom{x + r - 1}{r - 1} p^r (1 - p)^x$$

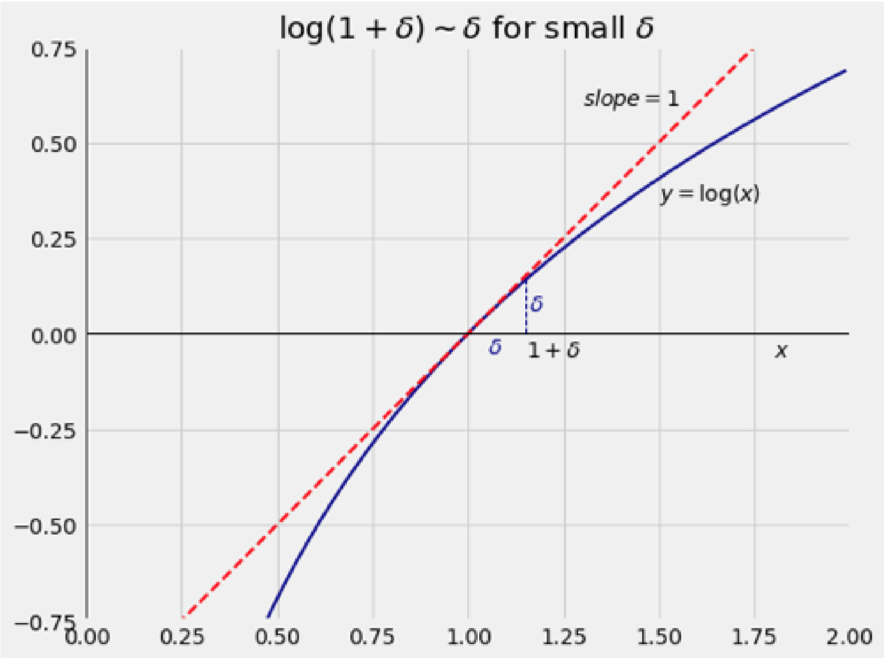
- Substituting r = 5, p = .2, and x = 10, into *nb*(*x*; *r*, *p*) gives

$$nb(10; 5, .2) = \binom{14}{4} (.2)^5 (.8)^{10} = .034$$

- The probability that at most 10 *F*'s are observed (at most 15 couples are asked) is

$$P(X \le 10) = \sum_{x=0}^{10} nb(x; 5, .2) = (.2)^5 \sum_{x=0}^{10} \binom{x + 4}{4} (.8)^x = .164$$

Exponential Approximations (See Text)



- The point is at a distance of δ away from $x = 1$.
- $\log(1+\delta)$ is the height of the blue curve at $x = 1 + \delta$.
- Because δ is small, the tangent line $y = x$ is very close to the curve $y = \log(x)$ at the point $x = 1 + \delta$.
- So the three points $(1, 0)$, $(1+\delta, 0)$, and $(1+\delta, \log(1+\delta))$ essentially form a 45°-90°-45° triangle.
- The two legs of that triangle are equal, so $\log(1+\delta) \approx \delta$.

How to use this approximation (lec 11)

- Result: log(1+x) = log (x) and log(1-x) = -x
- Approximate the value of $x = \left(1 - \frac{3}{100}\right)^{100}$
- $x = \left(1 - \frac{2}{1000}\right)^{5000}$
- $x = (1-p)^n$, for large n and small p

Example

- A book chapter n = 100,000 words and the chance that a word in the chapter has a typo (independently of all other words) is very small : p = 1/1,000,000 = 10^{-6}. Give an approximation of the chance the chapter *doesn't* have a typo. (Note: A typo is a *rare event*)

Bootstraps and probabilities

- Bootstrap sample: sample of size n drawn with replacement from original sample of n individuals

- Suppose one particular individual in the original sample is called Ali. What is the probability that Ali is chosen at least once in the bootstrap sample?

- Use the complement.