* Announcement

① Quiz 2 tomorrow (9/17) ~ Chapter 3

② HW4 due Mon (9/21)

③ Exam-prep section. Every Friday 2-3pm

④ +1 HW/Quiz drop

(HW4. Q1,2,3 are most relevant to today's lecture)

STAT 88: Lecture 9

Contents

Last time

Sec 4.1 Cummulative distribution function (CDF):

The CDF of a random variable X is $F(x) = P(X \leq x)$.

Purpose The CDF is an alternative way to specify a distribution:

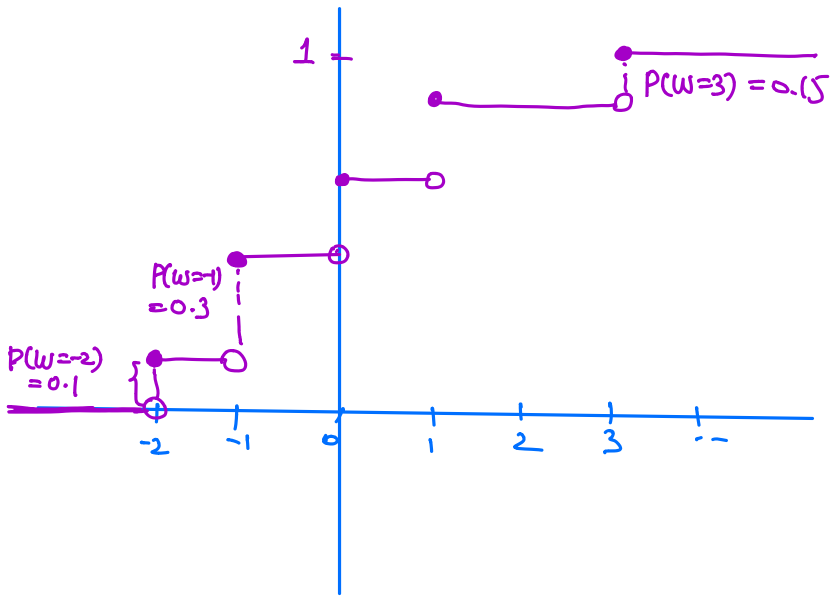
$$P(X = x) = P(X \leq x) - P(X \leq x - 1) = F(x) - F(x - 1).$$

Use Solutions to many problems can be expressed in terms of CDF and Python has built-in CDF function.

Warm up: (Exercise 4.5.2) A random variable W has the distribution shown in the table below. Sketch a graph of the cdf of W.

$P(W \leq w)$

w	-2	-1	0	1	3
$P(W = w)$	0.1	0.3	0.25	0.2	0.15



Computation You can use the stats module of SciPy to calculate CDF.

```
from scipy import stats
import numpy as np
```

x N G n

```
1 - stats.hypergeom.cdf(49, 100, 80, 60)
```

```
0.22097998866696655
```

```
sum(stats.hypergeom.pmf(np.arange(50,61), 100, 80, 60))
```

```
0.22097998866696314
```

$$X \sim HG(100, 80, 60)$$

$$P(X \geq 50) = \sum_{g=50}^{60} \frac{\binom{80}{g}\binom{20}{60-g}}{\binom{100}{60}}$$

Complement rule

$$= 1 - P(X < 50) = 1 - P(X \leq 49) = 1 - F(49)$$

4.2. Waiting Times

Waiting Time to the first success:

Consider a sequence of independent and identically distributed (iid) trials, each of which results in a success or a failure. Let p be the chance of success and q the chance of failure ($q = 1 - p$).

$T_1 = 4$

F F F S

Let $T_1 = \#$ trials until the first success. T_1 follows a distribution called "Geometric" distribution,

$$T_1 \sim \text{Geom}(p).$$

$\overbrace{F\ F\ F\ F\ \cdots\ F}^{k-1 \text{ trials}}\ S$, k trials

What is $P(T_1 = k)$? $P(T_1 = k) = q^{k-1} \cdot p$

($q = 1-p$, chance of failure)

What values does T_1 take? $k = 1, 2, 3, \cdots$

What is the chance it takes at most 5 trials for 1st success?

$P(T_1=1) + P(T_1=2) + \cdots + P(T_1=5) = P(T_1 \le 5)$

$= \sum_{k=1}^{5} q^{k-1} \cdot p$

By def. CDF for Geom(p)?

$P(T_1 \le k)$

$= 1 - P(T_1 > k)$

$= 1 - q^k$

$P(T_1 > k)$

$= P($you need more than k trials to get 1st success$)$

$= P($First k trials are failure$)$

$= q^k$

$\underbrace{F\ F\ F\ \cdots\ F}_{k \text{ trials}}$ $\left(\begin{matrix}F\\S\end{matrix}\right.$

$k+1$ $k+2$

Example: Cards are dealt one by one at random with replacement till the first ace appears. Let X be the number of cards dealt.

$X \sim \text{Geom}\left(\frac{4}{52} = \frac{1}{13}\right)$

(a) Find $P(X = 39)$.

(b) Find $P(X > 20)$.

(a) $P(X = 39) = \left(\frac{12}{13}\right)^{38} * \frac{1}{13}$

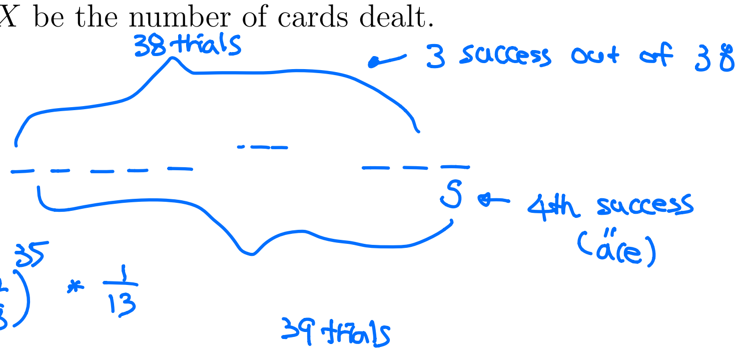
(b) $P(X > 20) = \left(\frac{12}{13}\right)^{20}$

$q = 1 - p$

Waiting time till the rth success: Cards are dealt one by one at random with replacement till the fourth ace appears. Let X be the number of cards dealt.

(a) Find $P(X = 39)$.

(b) Find $P(X > 20)$.



(a) $P(X = 39) = \binom{38}{3} \left(\frac{1}{13}\right)^{3} \left(\frac{12}{13}\right)^{35} * \frac{1}{13}$

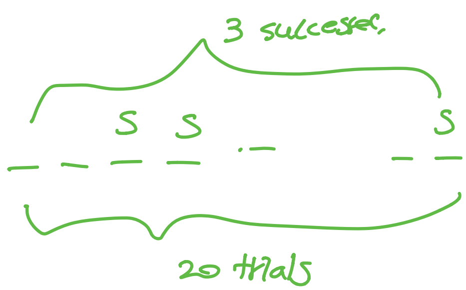
(b) $P(X > 20) =$ P(you need more than 20 trials to see 4 successes)

$=$ P(in the first 20 trials you have at most 3 success)

$= \sum_{k=0}^{3} \binom{20}{k} \left(\frac{1}{13}\right)^{k} \left(\frac{12}{13}\right)^{20-k}$

$1 - P(X \le 20)$

$= 1 - \sum_{k=1}^{20} P(X = k)$



One more variation

Example: (Exercise 4.5.5) Cards are dealt one by one at random without replacement till the fourth ace appears. Let X be the number of cards dealt.

(a) Find $P(X = 39)$.

(b) Find $P(X > 20)$.

Exercise

ace card

(a) $P(X=39) = P(\text{4th success appears in trial 39})$

$= P(\underbrace{\_\_\_ \; \cdots \; \_\_}_{\text{3 successes out of 38 trials}} \underbrace{S}_{\text{4th success}})$

total 39 trials

The sampling is w/o replacement, # 3 ace cards out of 38 sampled cards : HG(52, 4, 38)

cards in a deck; # ace cards; # samples

$$= \frac{\binom{4}{3}\binom{48}{35}}{\binom{52}{38}} * \frac{1}{14}$$

$14 = 52-38$ # remaining cards in trial 39

(b) $P(X>20) = P(\text{you have at most 3 successes in the first 20 trials})$

$$= \sum_{k=0}^{3} \frac{\binom{4}{k}\binom{48}{20-k}}{\binom{52}{20}}$$

$= F(3)$ where F is the CDF of HG(52, 4, 38)

[In Python, this is stats.hypergeom.cdf(3, 52, 4, 38)]

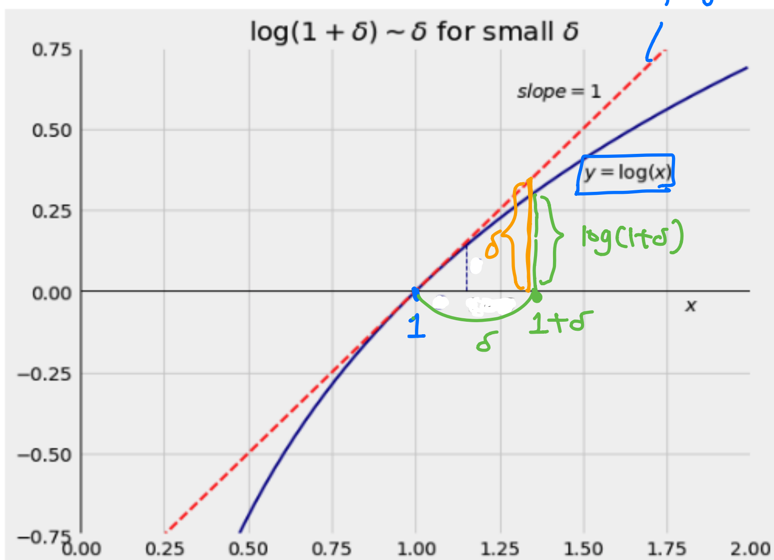
4.3. Exponential Approximations

A useful approximation from Calculus:

$$\log(1+\delta) \approx \delta \quad \text{for small } \delta$$

base e

tangent line of $y = \log x$ at $x = 1$



$f(x) = \log x$ is locally flat at $x = 1$ with slope 1. Since $f'(x) = \frac{1}{x}$, so $f'(1) = 1$. So starting at $x = 1$ if run by δ, you rise by δ. So $\log(1+\delta) \approx \delta$.

$\log(1+\delta) \approx \delta$

Example: Approximate $x = (1 - \frac{3}{100})^{100}$.

Take log on both side

$$\log x = \log\left(1 - \frac{3}{100}\right)^{100} = 100 \cdot \log\left(1 - \frac{3}{100}\right) \approx 100\left(-\frac{3}{100}\right). = -3$$

So $x \approx e^{-3}$.

$\delta = \frac{-3}{100}$

By Exponential Approx.,

$\log(1-\frac{3}{100}) = \log(1+\delta) \approx \delta = \frac{-3}{100}$

$\Downarrow$

$\log(1-\frac{3}{100}) \approx \frac{-3}{100}$

Give exponential approximation for

(a) $x = (1 - \frac{2}{1000})^{5000}$.

(b) $(1-p)^n$ for large n and small p.

(a) $\log x = 5000 \log(1 - \frac{2}{1000})$

Exp. Approx.

$\approx 5000 * (-\frac{2}{1000})$

$= -10$

$\Rightarrow x \approx e^{-10}$

(b) $x = (1-p)^n \Rightarrow \log x = n \log(1-p)$

$\approx n(-p)$

$= -np$

$\Rightarrow x \approx e^{-np}$