Stat 134 Lec 14

Warmup

$X = \#$ of coin tosses until the first head

$$P(X=k) = \underbrace{qq\cdots q}_{k-1}p = q^{k-1}p$$

Find $P(X>k) = P(X=k+1) + P(X=k+2) + \cdots$

$$= q^k p + q^{k+1}p + \cdots$$

$$= q^k \cancel{p}\left(1 + q + q^2 + \cdots\right) = \boxed{q^k.}$$

$$1 + q + q^2 + \cdots = \frac{1}{1-q} = \cancel{\frac{1}{p}}$$

Announcements:

- Midterm 1: Chaps 1–3. Wednesday March 1
- review sheets and practice test on website next week.
- in class review Friday/Monday before test.

Last time sec 3.6

Variance of sum of dependent i.d. indicators: ← identically distributed

$X = I_1 + \dots + I_n$

$p_1 = E(I_1)$

$p_{12} = E(I_{12})$ ← $I_1 I_2$

Note if I_1, I_2 indep

$p_{12} = E(\underbrace{I_1 I_2}_{= I_{12}}) = E(I_1)E(I_2) = p_1^2$

$E(X) = np_1$

$$\text{var}(X) = \underbrace{np_1 + n(n-1)p_{12}}_{E(X^2)} - \underbrace{(np_1)^2}_{E(X)^2}$$

Variance of sum of i.i.d. indicators:

$$\text{var}(X) = \underbrace{np_1 + n(n-1)p_1^2}_{E(X^2)} - \underbrace{(np_1)^2}_{E(X)^2} = np_1 - np_1^2$$

$$= np_1(1-p_1)$$

Stat 134

Monday February 24 2019

1. A fair die is rolled 14 times. Let X be the number of faces that appear exactly twice. Which of the following expressions appear in the calculation of $Var(X)$

   $X = I_1 + I_2 + \cdots + I_6$

   $I_2 = \begin{cases} 1 & \text{if } 2^{nd} \text{ face appears twice} \\ 0 & \text{else.} \end{cases}$

   a $\overset{6}{14} * \overset{5}{13} * \binom{14}{2,2,10}(1/6)^2(1/6)^2(4/6)^{10}$

   b $\binom{14}{2}(1/6)^2(5/6)^{12}$

   c more than one of the above

   d none of the above

No

b	Independent

b	Not 14*13, but 6*5

Today

1. sec 3.6 Hypergeometric dist.
2. sec 3.4 geometric distribution

① Sec 3.6 Hypergeometric Distribution

ex

A deck of cards has G aces.

X = # aces in n cards drawn without replacement from a deck of N cards.

$p = G/N$

above $N=52$
$G=4$
$n=5$

$X = I_1 + \cdots + I_n$

$I_2 = \begin{cases} 1 & \text{if } 2^{nd} \text{ card is ace} \\ 0 & \text{else} \end{cases}$

$E(X) = np_1$

$\mathrm{Var}(X) = \underbrace{np_1 + n(n-1)p_{12}}_{E(X^2)} - \underbrace{(np_1)^2}_{E(X)^2}$

a) Find $E(X) = n \cdot \frac{G}{N}$ ← $p = E(I_1)$

$p_{12} = \frac{G}{N} \cdot \frac{G-1}{N-1}$

b) Find $\mathrm{Var}(X)$

$I_{12} = \begin{cases} 1 & \text{if } 1^{st} \text{ and } 2^{nd} \text{ card is both ace} \\ 0 & \text{else} \end{cases}$

$\mathrm{Var}(X) = \underset{G/N}{\underset{\parallel}{np}} + n(n-1)\underset{\frac{G}{N}\cdot\frac{G-1}{N-1}}{\underset{\parallel}{p_{12}}} - (np)^2$

$(np)^2 = \left(n\frac{G}{N}\right)^2$

let $X \sim HG(n, N, G)$

identically distributed

$X = I_1 + \dots + I_n$ sum of dependent i.d. indicators

From above

$$Var(X) = \underbrace{np_1 + n(n-1)p_{12}}_{E(X^2)} - \underbrace{(np_1)^2}_{E(X)^2} \qquad (*)$$

where

$p_1 = \frac{G}{N}$

$p_{12} = \frac{G}{N} \frac{G-1}{N-1}$

A more useful formula for Var(X):

Suppose $n = N$ then

then $X = I_1 + \dots + I_N = G$ constant.

So $Var(X) = 0$

so $Np_1 + N(N-1)p_{12} - (Np_1)^2 = 0$

$$\Rightarrow p_{12} = \frac{Np_1(Np_1 - 1)}{N(N-1)}$$

Note that $Np_1 = N \cdot \frac{G}{N} = G$

This is another way to write $\frac{G}{N} \cdot \frac{G-1}{N-1}$

Plug this into $(*)$

$$\mathrm{Var}(X) = np_1 + n(n-1)\frac{\cancel{N}p_1(Np_1-1)}{\cancel{N}(N-1)} - (np_1)^2$$

$$= np_1\left[1 + \frac{(n-1)(Np_1-1)}{N-1} - np_1\right]$$

$$= \frac{np_1}{N-1}\left[\underbrace{(N-1) + (n-1)(Np_1-1) - np_1(N-1)}_{\substack{= \\ N-n-Np_1+np_1 \\ = \\ (N-n)(1-p)}}\right]$$

$$\boxed{\mathrm{Var}(X) = np_1(1-p_1)\frac{N-n}{N-1}}$$

correction factor ≤ 1

Compare with $\boxed{\mathrm{Var}(X) = np_1(1-p_1)}$ for $X \sim \mathrm{Bin}(n, p_1)$

So $X \sim HG(n, N, G)$

$E(X) = n\frac{G}{N}$

$\mathrm{Var}(X)$ $n\frac{G}{N}\left(1-\frac{G}{N}\right)\left(\frac{N-n}{N-1}\right)$

tinyurl.com/febr17-2023

Stat 134

Wednesday October 2 2019

1. The population of a small town is 1000 with 50% democrats. We wish to know what is a more accurate assesment of the % of democrats in the town, to randomly sample with or without replacement? The sample size is 10.

$\frac{N-X}{N-X} \approx \frac{N}{N} = 1$ if N is big.

a with replacement

b without replacement

c same accuracy with or without replacement

d not enough info to answer the question

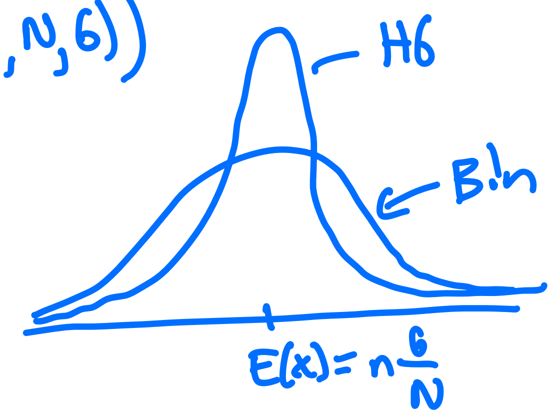
Equivalent to estimate of # democrats in town (500)

X = # democrats in your sample

$X = I_1 + I_2 + \dots I_{10}$ $\quad I_2 = \begin{cases} 1 & \text{if } 2^{nd} \text{ draw is democrat} \\ 0 & \text{else.} \end{cases}$

$Var(Bin(n, \frac{G}{N})) \geq Var(HG(n, N, G))$

So answer .



$X = I_1 + \dots + I_{10}$

(2) Sec 3.4 Geometric distribution (Geom(p)) on $\{1,2,3,\ldots\}$

ex $X = \#$ p coin tosses until the first head

$$P(X=k) = \underbrace{qq\cdots q}_{k-1}p = q^{k-1}p$$

You showed in the warm up that

$$P(X>k) = q^k$$

Recall:

$E(X) = P(X\geq 1) + P(X\geq 2) + P(X\geq 3) + \cdots$ Tail Sum Formula

$= P(X>0) + P(X>2) + \cdots = \sum_{k=0}^{\infty} P(X>k)$

Find $E(X)$ using the tail sum formula

$$E(X) = \sum_{k=0}^{\infty} q^k = \frac{1}{1-q} = \boxed{\frac{1}{p}}$$

see appendix to notes

Fact $\mathrm{Var}(X) = \frac{q}{p^2}$

Warning:

Some books define $Geom(p)$ on $\{0,1,2,\dots\}$ as

$Y = \#$ failures until 1^{st} success

ex $P(Y=4) = qqqqp$
$\quad\quad = P(X=5)$

$Y = X - 1$

$E(Y) = E(X) - 1 = \frac{1}{p} - 1 = \frac{1}{p} - \frac{p}{p} = \boxed{\frac{q}{p}}$

$Var(Y) = Var(X) = \boxed{\frac{q}{p^2}}$

Memoryless property of the Geometric distribution

Definition

$X \sim Geom(p)$ is memoryless if

$P(X > j+k \mid X > j) = P(X > k)$

i.e

The memoryless property says

if you don't get heads in 5 tosses

j

j+k

the chance you don't get a head in 7 tosses is the same as the unconditional probability that you don't get a heads in 2 tosses.

k

$$P(X > j+k \mid X > j) = \frac{P(X > j+k)}{P(X > j)} = \frac{q^{j+k}}{q^{j}} = q^{k} = P(X > k)$$

see appendix for proof

<u>Thm</u> A discrete distribution, X, taking values $1, 2, 3, \ldots$ is memoryless iff it is $\mathrm{Geom}(p)$ where $p = P(\underbrace{X=1}_{\text{event that you get a head in first trial.}})$

ex

Coupon Collector's problem

You have a collection of boxes each containing a coupon. There are n different coupons. Each box is equally likely to contain any coupon independent of the other boxes.

X = # boxes needed to get all n different coupons.

ex $n=3$ $\quad X = X_1 + X_2 + X_3$

$$\underbrace{C}_{X_1}\,\underbrace{C\,C\,C}_{X_2}\,\underbrace{C\,C\,C\,C}_{X_3}$$

a) What is the distribution of $\underset{\text{Geom}\left(\frac{3}{3}\right)}{\overset{}{X_1}}$, $\underset{}{X_2}$, X_3?

$X_1 = \text{Geom}\left(\frac{3}{3}\right)$, $X_2 = \text{Geom}\left(\frac{2}{3}\right)$, $X_3 = \text{Geom}\left(\frac{1}{3}\right)$

Are they independent? ~ indep

b) What is $E(X)$

$$= E(X_1 + X_2 + X_3)$$
$$= E(X_1) + E(X_2) + E(X_3)$$
$$= \frac{1}{3/3} + \frac{1}{2/3} + \frac{1}{1/3} = \boxed{3\left(1 + \frac{1}{2} + \frac{1}{3}\right)}$$

$$\boxed{Var(X) = \frac{q}{p^2}}$$

c) What is $Var(X)$? $= Var(X_1 + X_2 + X_3)$

$$= Var(X_1) + Var(X_2) + Var(X_3)$$

$$\frac{\frac{0}{3}}{\left(\frac{3}{3}\right)^2} + \frac{\frac{1}{3}}{\left(\frac{2}{3}\right)^2} + \frac{\frac{2}{3}}{\left(\frac{1}{3}\right)^2} = \boxed{3\left(\frac{0}{3^2} + \frac{1}{2^2} + \frac{2}{1^2}\right)}$$

<u>soln for n coupons:</u>

X_1 = # boxes to 1^{st} coupon $\sim Geom\left(\frac{n}{n}\right)$

$X_1 + X_2$ = # boxes to 2^{nd} coupon so $X_2 \sim Geom\left(\frac{n-1}{n}\right)$

$\vdots$

$X_1 + \dots + X_n$ = # boxes to n^{th} coupon so $X_n \sim Geom\left(\frac{1}{n}\right)$

$X = X_1 + \dots + X_n$ sum of <u>indep</u> Geom with diff. p.

$$E(X) = E(X_1) + E(X_2) + E(X_3) + \dots + E(X_n)$$
$$= \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + \dots + \frac{n}{1}$$

$$\boxed{E(X) = n\left(1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n}\right)}$$

$$\boxed{Var(X) = n\left(\frac{0}{n^2} + \frac{1}{(n-1)^2} + \dots \frac{n-1}{1^2}\right)}$$

Appendix

Fact $\mathrm{Var}(X) = \dfrac{q}{p^2}$

To find Var(X) we need an identity:

$$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q} \qquad \text{geometric sum}$$

$$\frac{d}{dq}$$

$$\sum_{k=0}^{\infty} k q^{k-1} = \frac{1}{(1-q)^2}$$

$$\frac{d}{dq}$$

$$\boxed{\sum_{k=0}^{\infty} k(k-1) q^{k-2} = \frac{2}{(1-q)^3} = \frac{2}{p^3}}$$

$$\begin{aligned}
\mathrm{Var}(X) &= E(X^2) - E(X)^2 \\
&= E(X^2) - E(X) + E(X) - E(X)^2 \\
&= E(\underbrace{X^2 - X}_{X(X-1)}) + \underbrace{E(X)}_{=\frac{1}{p}} - \underbrace{E(X)^2}_{=\frac{1}{p^2}}
\end{aligned}$$

$$E(X(X-1)) = \sum_{k=1}^{\infty} k(k-1) \underbrace{P(X=k)}_{= q^{k-1}p}$$

$$E(g(X)) = \sum_{x \in X} g(x) P(X=x)$$

$$= qp \sum_{k=1}^{\infty} k(k-1) q^{k-2} = qp \underbrace{\sum_{k=0}^{\infty} k(k-1) q^{k-2}}_{= \frac{2}{p^3}} \quad \text{(see above)}$$

$$= \frac{2q}{p^2}$$

so $\mathrm{Var}(X) = \underbrace{\dfrac{2q}{p^2}}_{E(X(X-1))} + \underbrace{\dfrac{1}{p}}_{E(X)} + \underbrace{\dfrac{1}{p^2}}_{E(X)^2} = \boxed{\dfrac{q}{p^2}}$

Appendix

The defⁿ for memoryless in the discrete case is a little different. It says, given that $X > j$ then the chance $X > j+k$ is the same as the unconditional probability that $X > k$.

Definition

$X \sim Geom(p)$ is memoryless if

$$P(X > j+k \mid X > j) = P(X > k)$$

Thm A discrete distribution, X, taking values $1, 2, 3, \ldots$ is memoryless iff it is $Geom(p)$ where $p = P(X=1)$.

Pf/ We already showed in class that $Geom(p)$ is memoryless.

Next we show that if X is memoryless with values $X = 1, 2, 3, \ldots$ then X is $Geom(p)$ where $p = P(X=1)$.

For positive integers k, j

Suppose $P(X > k+j \mid X > j) = P(X > k)$

$$P(X > k+j \mid X > j) = \frac{P(X > k+j)}{P(X > j)}$$

$\Leftarrow$ $P(X > k+j) = P(X > j)P(X > k)$

It follows from this that

$P(X > j) = P(X > 1)^j$ for $j = 1, 2, \ldots$

This can be shown by induction

base case $j=1$ ✓

assume true for $j-1$

$$P(X > j) = P(X > j-1+1) = P(X > j-1)P(X > 1)$$

$$\underset{P(X>1)^{j-1}}{\overset{\shortparallel}{}}$$

$$= P(X > 1)^j \checkmark$$

let $q = P(X > 1)$

then $P(X > j) = q^j \Rightarrow X \sim \text{Geom}(1-q)$

where $p = 1 - q = P(X = 1)$.

$\square$