Stat 134 lec 15

- in class Midterm Friday October 11
- review sheets coming today
- in class review next Wednesday.

Last time sec 3.6

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

$X = I_1 + \cdots + I_n$

$p_1 = E(I_1)$

$p_{12} = E(I_1 I_2)$ ← $I_1 I_2$

$E(X) = np_1$

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

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

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

Today

1. student responses from concept test
2. finish sec 3.6 Hypergeometric dist.
3. sec 3.4 geometric distribution
   negative binomial distribution

Monday February 24 2019

(1)

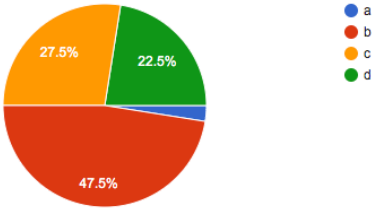
1. A fair die is rolled 14 times. Let X be the number of faces that appear exactly twice. Which of the following expressions appear in the calculation of $Var(X)$

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

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

c more than one of the above

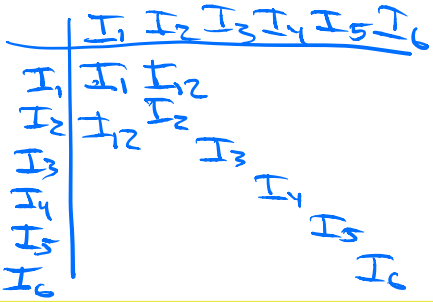
d none of the above



c	a is the nondiagonals and b is the diagonals.

$P_1 = \binom{14}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^{12}$

$P_{12} = \binom{14}{2,2,10}\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2\left(\frac{4}{6}\right)^{10}$



b	Not 14x13, should be 6x5 (36 - 6 diagonal) for E(X^2)

tinyurl.com/october2-pt1

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.

   a with replacement

   b without replacement

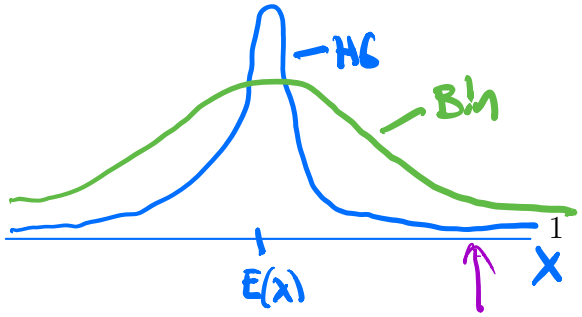
   c same accuracy with or without replacement

   d not enough info to answer the question

let X = # democrats in town

we are comparing variance of $X \sim Bin(10, \frac{500}{1000})$ and $X \sim HG(10, 1000, 500)$

$Var(HG(n,N,G)) < Var(Bin(n, \frac{G}{N}))$ for $n > 1$



R simulation for SD of # democrats in sample.

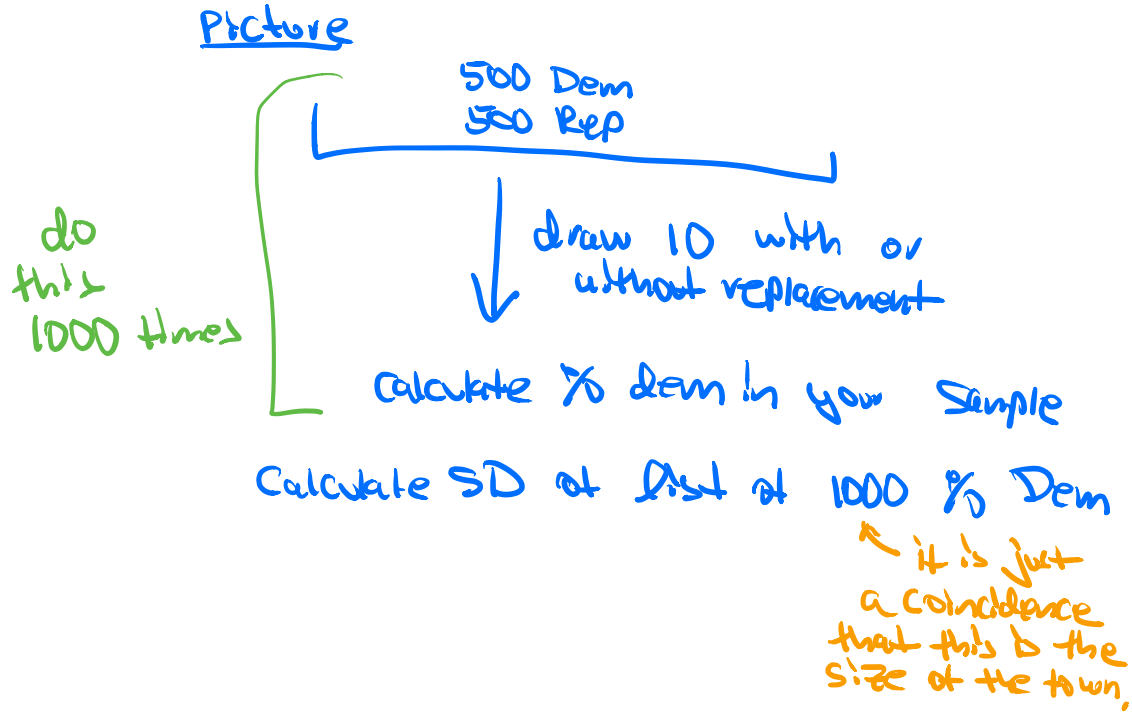
Code | Start Over | Solution

```
1  pop <- 1000
2  sample_size <- 10
3  a <- rep(0,each=pop/2)
4  b <- rep(1,each=pop/2)
5  box <- c(a,b)
6  for(boolean in c(TRUE,FALSE)){
7    fun <- function(){
8      my_sample <- sample(box,size=sample_size,replace=boolean)
9      mean(my_sample)*100
10   }
11  B <- 1000
12   vec_percentages <- replicate(B,fun())
13   print(sd(vec_percentages))
14 }
```

```
[1] 15.77621
[1] 15.24316
```

[1] 15.77621 ← replacement

[1] 15.24316 ← without replacement. (Smaller)



② Sec 3.6 Hypergeometric Distribution

let $X \sim HG(n, N, G)$, $P(X=k) = \dfrac{\binom{G}{k}\binom{N-G}{n-k}}{\binom{N}{n}}$ HG formula

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

last time we saw:

$$Var(X) = \underbrace{nP_1 + n(n-1)P_{12}}_{E(X^2)} - \underbrace{(nP_1)^2}_{E(X)^2} \quad (*)$$

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 $X = I_1 + \cdots + I_N = G$ ← 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 $(*)$

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

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

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

$$= N - n - Np_1 + np_1$$

$$= (N-n)(1-p_1)$$

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

correction factor ≤ 1

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

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

ex X = # p coin tosses until the first head

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

Find $P(X>k)$

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

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

$= q^k p(1 + q + q^2 + \cdots) = q^k p \cdot \frac{1}{1-q} = \boxed{q^k}$

$1-q = p$

Recall:

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

$= P(X>0) + P(X>1) + \cdots = \sum_{k=0}^{\infty} P(X>k)$

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

$$E(X) = \sum_{k=0}^{\infty} P(X>k) = \sum_{k=0}^{\infty} q^k = 1 + q + \cdots$$

$$= \frac{1}{1-q} = \boxed{\frac{1}{p}}$$

To find $\text{Var}(X)$ we need an identity:

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

$\frac{d}{dq}$

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

$\frac{d}{dq}$

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

$$\text{Var}(X) = E(X^2) - E(X)^2$$

$$= E(X^2) - E(X) + E(X) - E(X)^2$$

$$= E(\underbrace{X^2 - X}_{X(X-1)}) + \underbrace{E(X)}_{\frac{1}{p}} - \underbrace{E(X)^2}_{\frac{1}{p^2}}$$

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

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

$P(X=k) = q^{k-1} p$

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

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

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

So $\text{Var}(X) = \underbrace{\frac{2q}{p^2}}_{E(X(X-1))} + \underbrace{\frac{1}{p}}_{E(X)} - \underbrace{\frac{1}{p^2}}_{E(X)^2} = \boxed{\frac{q}{p^2}}$

Warning:

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

Y = # failures until 1st success

i.e. $P(Y=4) = qqqqp$

$\parallel$

$P(X=5)$

$Y = X - 1$

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

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

(4) Negative Binomial Distribution $(\text{NegBin}(r, p))$

generalization of Geom(p)

sum of r indep Geom(p) on $\{1, 2, \ldots\}$

ex $r = 3$

$\underbrace{qqqp}_{w_1}\underbrace{qqp}_{w_2}\underbrace{r}_{w_3}$ # trials until 3rd success

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

T_r = # indep p-trials until r^{th} success

$$P(T_r = k) = \binom{k-1}{r-1} p^{r-1} q^{k-1-(r-1)} p = \binom{k-1}{r-1} p^r q^{k-r}$$

$T_r = w_1 + \cdots + w_r$ where $w_1, \ldots, w_r \overset{iid}{\sim} \text{Geom}(p)$

$$E(T_r) = r E(w_1) = \frac{r}{p}$$

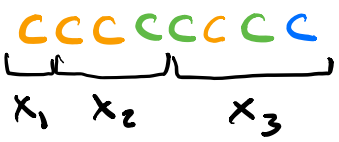
$$\text{Var}(T_r) = r \text{Var}(w_1) = \frac{rq}{p^2}$$

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 $n=3$ $X = X_1 + X_2 + X_3$



a) What is the distribution of X_1, X_2, X_3? Are they independent?

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

b) What is $E(X)$

$E(X_1) = \frac{1}{p_1}$

$E(X) = \frac{1}{3/3} + \frac{1}{2/3} + \frac{1}{1/3}$

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

c) What is $\text{Var}(X)$?

$\text{Var}(X_1) = \frac{q}{p^2}$

$\text{Var}(X) = \frac{0/3}{(3/3)^2} + \frac{1/3}{(2/3)^2} + \frac{2/3}{(1/3)^2}$

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

Soln for n coupons:

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$\left(\frac{n-1}{n}\right)$

$\vdots$

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

$X = X_1 + \cdots + X_n$ sum of indep Geom with diff. P.

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

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

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