

Warmup 1:00-1:10

Stat 134 lec 15

1. The population of a small town is 1000 with 50% democrats. We wish to know what is a more accurate assessment 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

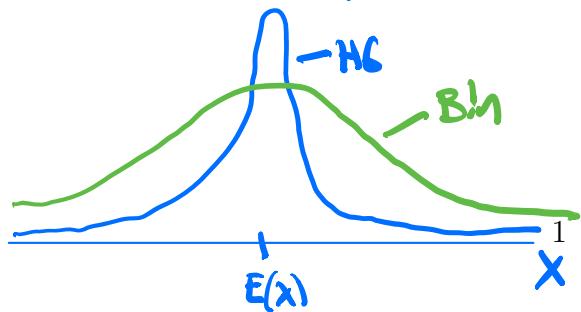
c same accuracy with or without replacement

d not enough info to answer the question

let $X = \# \text{ democrats in town}$

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

$$\text{Var}(\text{Hg}(N, b, n)) < \text{Var}(\text{Bin}(n, \frac{b}{N})) \text{ for } n \geq 1$$



Intuitively you could
if your sample size is
1000 you could get 0%
democrats or 100%
democrats when draw w/
replacement and you would
get 50% when draw w/o
replacement

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 <- 10000
12 vec_percentages <- replicate(B,fun())
13 print(sd(vec_percentages))
14 }
```

replacement *without replacement*

SD = [1] 15.77621 ← replacement

SP = [1] 15.24316 ← without replacement. (Smaller)

Picture

do this 10,000 times and make a list of values >

500 Dem
500 Rep

draw 10 with or without replacement

calculate % dem in your sample

Then take SD of the list.

Announcements:

- Midterm 1: Friday October 9
- review sheets and practice test on website soon
- in class review next wednesday

Last time sec 3.6 identically distributed
 Variance of sum of dependent i.d. indicators:

$$X = I_1 + \dots + I_n$$

$$P_i = E(I_i)$$

$$P_{12} = E(I_{12})$$

I_1, I_2

$$E(X) = nP_i$$

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

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

$$\text{Var}(X) = \underbrace{nP_i}_{E(X^2)} + n(n-1)P_i^2 - \underbrace{(nP_i)^2}_{E(X)^2} = nP_i - nP_i^2$$

$$= nP_i(1-P_i)$$

Today

① student responses from concept test

② sec 3.6 Hypergeometric dist.

③ sec 3.4 geometric distribution

④ sec 3.4 Negative Binomial distribution

(1)

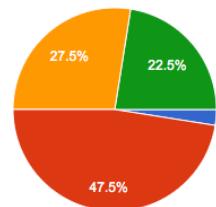
1. A fair die is rolled 14 times. Let X be the number of faces that appear exactly twice. Which of the following expressions appear in the calculation of $Var(X)$

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



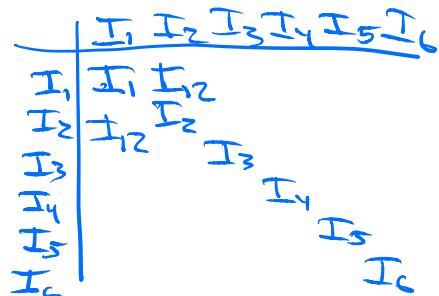
- a
- b
- c
- d

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 14×13 , should be 6×5 (36 - 6 diagonal) for $E(X^2)$

② Sec 3.6 Hypergeometric Distribution

$$E(X) = np_1$$

$$\text{Var}(X) = np_1 + n(n-1)p_{12} - \frac{(np_1)^2}{E(X^2)}$$

A deck of cards has G aces.

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

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

a) Find $E(X)$

$$X = I_1 + I_2 + \dots + I_5$$

$$E(X) = 5 \left(\frac{4}{52} \right)$$

$$= \boxed{n \left(\frac{G}{N} \right)}$$

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

$$P_1 = \frac{6}{N}$$

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

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

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

$$\text{Var}(X) = np_1 + n(n-1)p_{12} - (np_1)^2$$

let $X \sim HG(N, 6, n)$

'identically distributed'

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

From above

$$Var(X) = \frac{nP_1 + n(n-1)P_{12} - (nP_1)^2}{E(X^2)}$$

where

$$P_1 = \frac{6}{N}$$

$$P_{12} = \frac{6}{N} \frac{6-1}{N-1}$$



A more useful formula for $Var(X)$:

Suppose $n=N$ then

constant.

$$\text{then } X = I_1 + \dots + I_N = G$$

$$\text{so } Var(X) = 0$$

$$\text{so } NP_1 + N(N-1)P_{12} - (NP_1)^2 = 0$$

$$\Rightarrow P_{12} = \frac{NP_1(NP_1 - 1)}{N(N-1)}$$

Note that

$$NP_1 = N \cdot \frac{6}{N} = 6$$

This is another way to write

$$\frac{6 \cdot 6-1}{N \cdot N-1}$$

Plug this into



$$\text{Var}(x) = np_i + n(n-1) \frac{np_i(Np_i - 1)}{N(N-1)} - (np_i)^2$$

$$\begin{aligned} &= np_i \left[1 + \frac{(n-1)(Np_i - 1)}{N-1} - np_i \right] \\ &= \frac{np_i}{N-1} \left[(N-1) + (n-1)(Np_i - 1) - np_i(N-1) \right] \\ &\quad \text{“} \\ &\quad N-n-Np_i+np_i \\ &\quad (N-n)(1-p) \end{aligned}$$

$$\boxed{\text{Var}(x) = np_i(1-p_i) \frac{N-n}{N-1}}$$

“ correction factor ≤ 1

Compare with $\boxed{\text{Var}(x) = np_i(1-p_i)}$ for $X \sim \text{Bin}(n, p_i)$

So $X \sim \text{Bin}(n, G, n)$

$$\boxed{\begin{aligned} E(x) &\rightarrow nG/n \\ \text{Var}(x) &\rightarrow nG/n(1-G/n) \left(\frac{N-n}{N-1} \right) \end{aligned}}$$

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

$\equiv X = \# \text{ of coin tosses until the first head}$

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

Find $P(X \geq k)$

$$P(X \geq k+1) + P(X \geq k+2) + \dots$$

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

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

Recall:

$$\begin{aligned} E(X) &= P(X \geq 1) + P(X \geq 2) + P(X \geq 3) + \dots && \text{Tail Sum Formula} \\ &= P(X \geq 0) + P(X \geq 1) + \dots = \sum_{k=0}^{\infty} P(X \geq k) \end{aligned}$$

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

$$\begin{aligned} E(X) &= \sum_{k=0}^{\infty} P(X \geq k) = \sum_{k=0}^{\infty} q^k = 1 + q + \dots \\ &\quad = \frac{1}{1-q} = \boxed{\frac{1}{p}} \end{aligned}$$

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

See appendix to notes

Warning:

Some books define $\text{Geom}(p)$ on $\{0, 1, 2, \dots\}$ as
 $Y = \# \text{ failures until } 1^{\text{st}} \text{ success}$
ex $P(Y=4) = qqqq p$
" $P(X=5)$

$$Y = X - 1$$

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

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

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

generalization of $\text{Geom}(p)$

Sum of
 r indep $\text{Geom}(p)$,
on $\{1, 2, \dots\}$

ex $r=3$

$\underbrace{q q q}_w \underbrace{p}_r \underbrace{q q p}_w \underbrace{p}_r$ # trials until 3rd success

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

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

$\underbrace{\quad}_{k-1} \begin{matrix} r-1 \\ p \end{matrix} \text{ in } k-1 \text{ slots}$

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

$T_r = w_1 + \dots + w_r$ where $w_1, \dots, w_r \stackrel{iid}{\sim} \text{Geom}(p)$

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

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

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.

$X = \# \text{ boxes needed to get all } n \text{ different coupons.}$

$$\text{Ex } n=3 \quad X = X_1 + X_2 + X_3$$



$$x_1 \quad x_2 \quad x_3$$

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

$$\left. \begin{array}{l} X_1 \sim \text{Geom}\left(\frac{1}{3}\right) \\ X_2 \sim \text{Geom}\left(\frac{2}{3}\right) \\ X_3 \sim \text{Geom}\left(\frac{1}{3}\right) \end{array} \right\} \text{Indp}$$

b) What is $E(X)$

$$E(X_1) = \frac{1}{P_1}$$

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

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

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

$$\text{Var}(X_1) = \frac{9}{P_1^2}$$

$$\text{Var}(X) = \frac{10}{\left(\frac{1}{3}\right)^2} + \frac{1}{\left(\frac{2}{3}\right)^2} + \frac{\frac{2}{3}}{\left(\frac{1}{3}\right)^2}$$

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

Soln for n coupons:

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

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

$$\vdots$$

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

$X = X_1 + \dots + X_n$ sum of Indep Geom with diff. P.

$$E(X) = E(X_1) + E(X_2) + E(X_3) + \dots + E(X_n)$$

$$= \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + \dots + \frac{n}{1}$$

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

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

Appendix

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

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

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

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

$$\frac{d}{dq} \left[\sum_{k=0}^{\infty} k(k-1)q^{k-2} \right] = \frac{2}{(1-q)^3} - \frac{2}{p^3}$$

$$\begin{aligned} \text{Var}(x) &= E(x^2) - E(x)^2 \\ &= E(x^2) - E(x) + E(x) - E(x)^2 \\ &= E(x(x-1)) + E(x) - E(x)^2 \end{aligned}$$

$$\begin{aligned} E(x(x-1)) &= \sum_{k=1}^{\infty} k(k-1)p(x=k) \\ E(x) &= \sum_{x \in X} x p(x=x) \quad \frac{k}{q^{k-1}} \quad \frac{1}{p} \\ &= qp \sum_{k=1}^{\infty} k(k-1)q^{k-2} = qp \sum_{k=0}^{\infty} k(k-1)q^{k-2} \\ &= \frac{2q}{p^2} \quad \frac{2}{p^3} \quad (\text{See above}) \end{aligned}$$

$$\text{so } \text{Var}(x) = \frac{2q}{p^2} + \frac{1}{p} + \frac{1}{p^2} = \boxed{\frac{q}{p^2}}$$

