Stat 88 lec 24

warm up 2:00-2:10

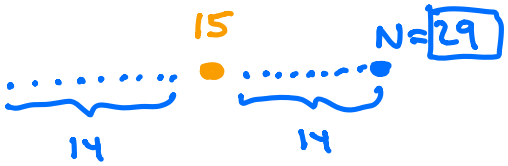
N consecutive positive integers, 1, 2, 3, ..., N are in a hat, with N unknown.

You randomly sample one number from the hat and get 15. Estimate N

Lets call the blue dots to the left of the gold dot a "gap".

By symmetry we would expect an equal sized gap on the other side of the gold dot.

So $14 + 1 + 14 = 29$ is a reasonable estimate of N.



Last time — Q4 – next Friday Chap 8, 9, 10

Sec 10.4 Tests for the equality of two proportions

We have two independent populations of 0, 1s and wish to test that the proportion of 1s is the same.

X ~ p_x : 0 1 — draw n w/ replacement

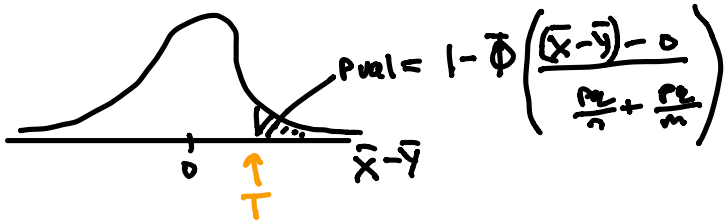
Y ~ p_y : 0 1 — draw m w/ replacement

$X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(p_x)$ $\quad Y_1, Y_2, \dots Y_m \overset{iid}{\sim} \text{Bernoulli}(p_y)$

$H_0: p_x = p_y = p$ level 5% test.

$H_1: p_x > p_y$

$T = \bar{X} - \bar{Y}$ is our T.S. and we reject the null if Pval < .05



Sec 11.1 Bias and Variance

Some estimators for a parameter, θ, are better than others. A good estimator is one with a small mean square error.

$$mse_\theta(T) = E_\theta\left((T-\theta)^2\right) = B_\theta^2(T) + Var_\theta(T)$$

(θ: parameter; T: estimator)

where $B_\theta(T) = E(T) - \theta$ is the bias,

$Var_\theta(T) = E_\theta\left((T - E(T))^2\right)$ is the variance.

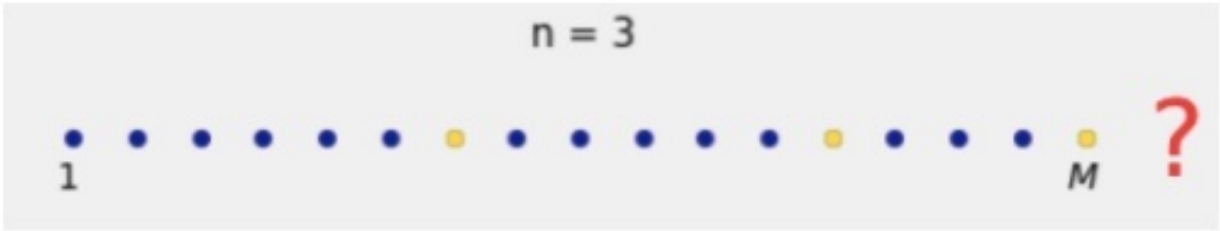
Today

(1) Sec 11.2 The German Tank Problem

(1) Sec 11.2 The German Tank Problem

The Allies during WW II needed to estimate how many Tanks N the Germans had produced. The idea was to model the observed serial numbers as random draws from $1, 2, \ldots, N$ and then estimate N.

Suppose they capture 3 tanks with serial numbers 6, 10, 16.



Estimate N in your breakout room.

average gap size $= \frac{\text{total \# blue dots}}{\text{\# gaps}} = \frac{M-3}{3}$

$= \frac{16-3}{3} = \frac{13}{3}$

By symmetry there should be a gap of size $\frac{13}{3}$ to the right of M $\Rightarrow N \approx M + \frac{M-3}{3} = \frac{4M}{3} - 1 = \frac{4}{3}\cdot 16 - 1$

$= \boxed{20.33}$

we will see the estimator $\frac{4}{3}M - 1$ is unbiased.

So we will now assume, as the Allies did, that the serial numbers of the observed tanks are random variables $X_1, X_2, \ldots, X_n$ drawn uniformly at random *without* replacement from $\{1, 2, 3, \ldots, N\}$. That is, we have a simple random sample of size n from the population $\{1, 2, 3, \ldots, N\}$, and we have to estimate N.

We will compare several estimators

T_1

By symmetry

$$E(X_i) = \frac{N+1}{2}$$

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

This is a linear function of N so can find an unbiased estimator

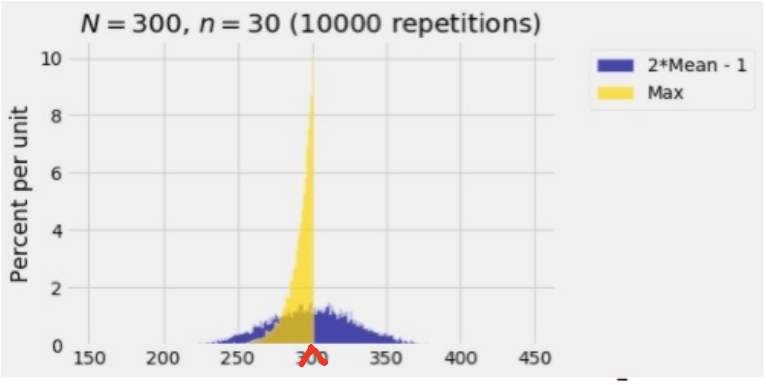
$$N = 2E(\bar{X}) - 1$$

let $T_1 = 2\bar{X} - 1$

Check $E(T_1) = 2E(\bar{X}) - 1 = N$ ✓

T_2

let $T_2 = \max(X_1, \ldots, X_n)$

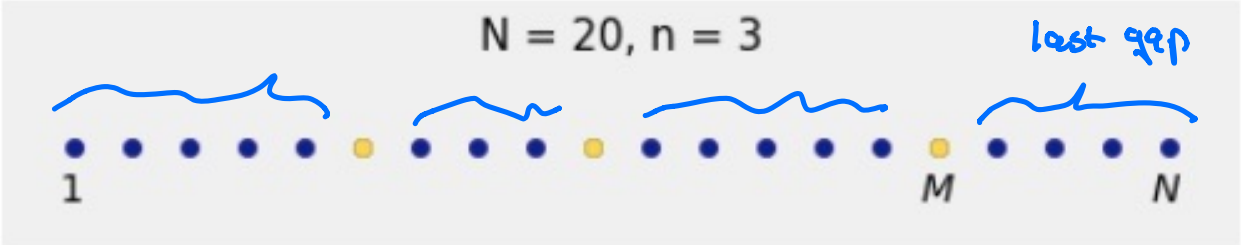


T_2 is clearly biased.

Compute bias of T_2:

$$B_N(T_2) = E(M) - N$$

M



by symmetry, the expected length of all the gaps is the same.

$N - E(M)$ is the expected length of the last gap,

poll $\quad N - E(M)$ is ? $\quad \frac{20-3}{4}$ ← 3 gold, ← 4 gaps

$= E(M) - N$

so $B_N(T_2) = -\frac{(20-3)}{4}$

more generally,

$B_N(T_2) = ?$ $\quad -\frac{(N-n)}{n+1}$

breakout room

write $E(M)$ as a function of N

$$E(M) - N = -\frac{(N-n)}{n+1} \Rightarrow E(M) = N - \frac{(N-n)}{n+1}$$

or $$E(M) = \frac{(N-n)}{n+1} \cdot n + n = \frac{n}{n+1}(N-1) + n$$

$$= \boxed{\frac{n}{n+1}(N+1)}$$

T_3 (the "augmented maximum")

Since $E(M)$ is a linear function of N we can make a new unbiased estimator by solving for N.

$$E(M) = \frac{n}{n+1}(N+1)$$

$$\frac{n+1}{n}E(M) - 1 = N \Rightarrow N = \frac{n+1}{n}E(M) - 1$$

So $T_3 = \frac{n+1}{n}M - 1$ is an unbiased estimator

$$E(T_3) = \frac{n+1}{n}E(M) - 1 = N \checkmark$$

In the case where $M = 16$, $n = 3$

$$T_3 = \frac{4}{3}\cdot 16 - 1 = 20.33$$

How does $Var(T_3)$ and $Var(T_2)$ compare?

poll

$$T_3 = \frac{n+1}{n}T_2 - 1$$

$$Var(T_3) = \left(\frac{n+1}{n}\right)^2 Var(T_2)$$

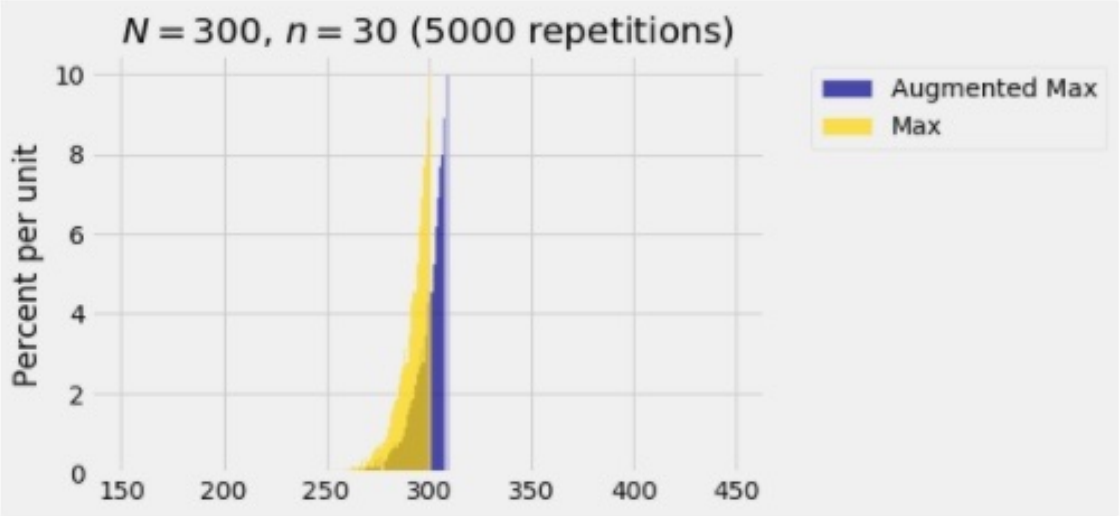
So $Var(T_3) > Var(T_2)$ but almost equal for large n

How does $B_N(T_3)$ and $B_N(T_2)$ compare?

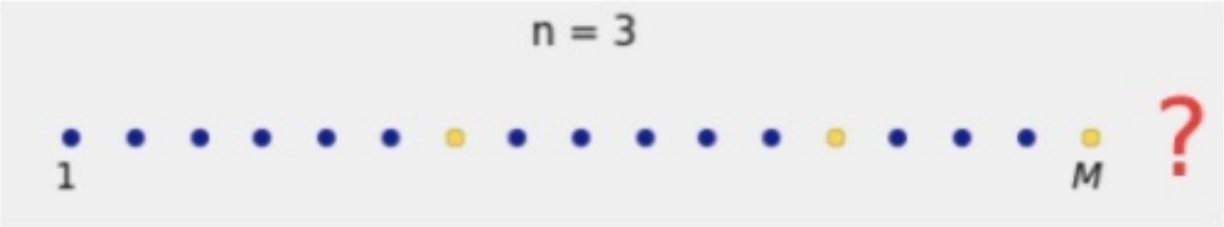
$B_N(T_3) = 0$ $\quad$ $B_N(T_2) = -\frac{(N-n)}{n+1}$

large n, T_3 is a better estimator

```
Average of Augmented Maxes: 300.18587333333335
SD of of Augmented Maxes: 8.947086216787126
Average of Maxes: 291.4702
SD of Maxes: 8.65847053237464
```



Another way to think of the "augmented max"



estimated gap length ? $\frac{M-n}{n}$

$N \approx$? $M + \frac{M-n}{n} = \frac{n+1}{n} M - 1$

which is what we derived earlier,

In summary, we can have many different estimators for a parameter. In this lecture T_1 was unbiased but had a large variance, T_2 was biased but had a smaller variance. T_3 was unbiased and had a bigger variance (but for large n $Var(T_2) \approx Var(T_3)$,

the estimator with the smallest $MSE = bias^2 + var$ is best,