STAT 88: Lecture 34

Contents

<u>Warm up:</u>

N consecutive positive integers, $1, 2, \ldots, N$ are in a hat, with N unknown. You randomly sample one number from the hat and get 15. Estimate N.

Last time

Suppose you have two independent samples, $X_1, X_2, \ldots, X_n$ i.i.d. with mean μ_X and SD σ_X, and $Y_1, Y_2, \ldots, Y_m$ are i.i.d. with mean μ_Y and SD σ_Y. By CLT, $\bar{X} - \bar{Y}$ is approximately distributed as

$$\mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right).$$

An approximate 95% CI for $\mu_X - \mu_Y$ is given by

$$\bar{X} - \bar{Y} \pm 2\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}.$$

If we wish to test $H_0 : \mu_X = \mu_Y$ vs $H_A : \mu_X > \mu_Y$, $T = \bar{X} - \bar{Y}$ is our test statistic. Under H_0, we have

$\mu_X - \mu_Y = 0$

$$T \sim \mathcal{N}\left(0, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right).$$

If t is an observed value of our test statistic, p-value is given by

$$\text{p-val} = P(T \geq t) = P\left(Z \geq \frac{t}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}\right) = 1 - \Phi\left(\frac{t}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}\right).$$

Next, you have two independent populations of 0's and 1's, so $X_1, X_2, \ldots, X_n$ are i.i.d. from Bernoulli(p_X), and $Y_1, Y_2, \ldots, Y_m$ are i.i.d. from Bernoulli(p_Y). By CLT, $\bar{X} - \bar{Y}$ is approximately distributed as

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(p_X - p_Y, \frac{p_X(1-p_X)}{n} + \frac{p_Y(1-p_Y)}{m}\right).$$

An approximate 95% CI for $p_X - p_Y$ is given by

Estimate $p_X \leftarrow \bar{X}$

$p_Y \leftarrow \bar{Y}$

$$\bar{X} - \bar{Y} \pm 2\sqrt{\frac{p_X(1-p_X)}{n} + \frac{p_Y(1-p_Y)}{m}}.$$

If we wish to test $H_0 : p_X = p_Y = p$ vs $H_A : p_X > p_Y$, $T = \bar{X} - \bar{Y}$ is our test statistic. Under H_0, we have

Estimate p by $\frac{n}{n+m}\bar{X} + \frac{m}{n+m}\bar{Y}$

$$T \sim \mathcal{N}\left(0, \frac{p(1-p)}{n} + \frac{p(1-p)}{m}\right).$$

$p_X - p_Y = 0$

If t is an observed value of our test statistic, p-value is given by

$$\text{p-val} = P(T \geq t) = P\left(Z \geq \frac{t}{\sqrt{\frac{p(1-p)}{n} + \frac{p(1-p)}{m}}}\right) = 1 - \Phi\left(\frac{t}{\sqrt{\frac{p(1-p)}{n} + \frac{p(1-p)}{m}}}\right).$$

11.1. Bias and Variance

Bias-Variance Decomposition

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

" Bias^2 + Var "

estimator

$$\mathrm{MSE}_\theta(T) = E_\theta((T-\theta)^2) = B_\theta^2(T) + \mathrm{Var}_\theta(T),$$

parameter

where

$$B_\theta(T) = E_\theta(T) - \theta \text{ and } \mathrm{Var}_\theta(T) = E_\theta((T - E_\theta(T))^2).$$

Bias Variance

11.2. The German Tank Problem

The Allies during WWII 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.

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

Now we will compare several estimators.

T_1: By symmetry, for each i,

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

Since $\bar{X}$ is an unbiased estimator for the pop. mean,

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

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

$$E(\underbrace{2\bar{X} - 1}_{=T_1}) = N.$$

If we define

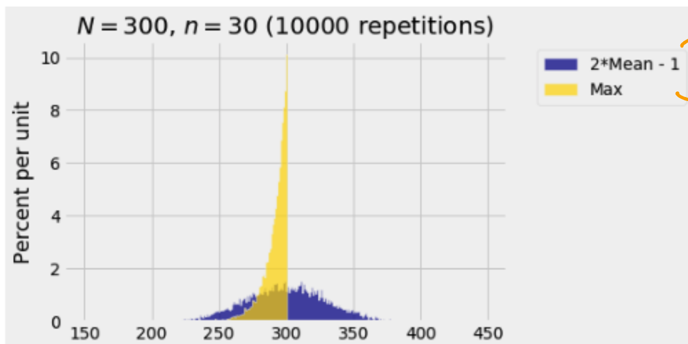
$$T_1 = 2\bar{X} - 1,$$

it is an unbiased estimator of N.

T_2: Another natural estimator is

$$T_2 = \max\{X_1, X_2, \ldots, X_n\},$$

the maximum of the observed numbers.



Unbiased and large var

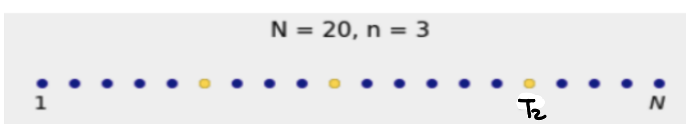
biased and small var

T_2 is clearly biased. We compute the bias of T_2.

The Bias of the Sample Maximum The bias of T_2 is

$$B(T_2) = E(T_2) - N.$$

Imagine a row of N spots for the serial numbers 1 through N, with marks at the spots corresponding to the observed serial numbers.



The $n = 3$ sampled spots create $n + 1 = 4$ blue "gaps" between sampled values: one before the leftmost gold spot, two between successive gold spots, and one after the rightmost gold spot that is at position T_2.

By symmetry, the lengths of all four gaps have the same distribution. Therefore all four gaps have the same expected length.

- The gaps are made up of $N - n = 17$ blue spots;

- Since each of the four gaps has the same expected length, the expected length of a single gap is $\frac{17}{4}$.

More generally,

$$\text{expected length of gap} = \frac{N-n}{n+1}.$$

Note that $N - E(T_2)$ is the expected length of the last gap. So,

$$N - E(T_2) = \frac{N-n}{n+1},$$

and therefore

$$B(T_2) = E(T_2) - N = \frac{-(N-n)}{n+1}.$$

T_3: (the "augmented maximum")

What is $E(T_2)$?

Since $E(T_2)$ is a linear function of N, we can make a new unbiased estimator by solving for N.

$$\begin{aligned} E(T_2) &= \frac{n}{n+1}(N+1) \\ \Rightarrow E\left(\frac{n+1}{n} \cdot T_2 - 1\right) &= N. \end{aligned}$$

If we define

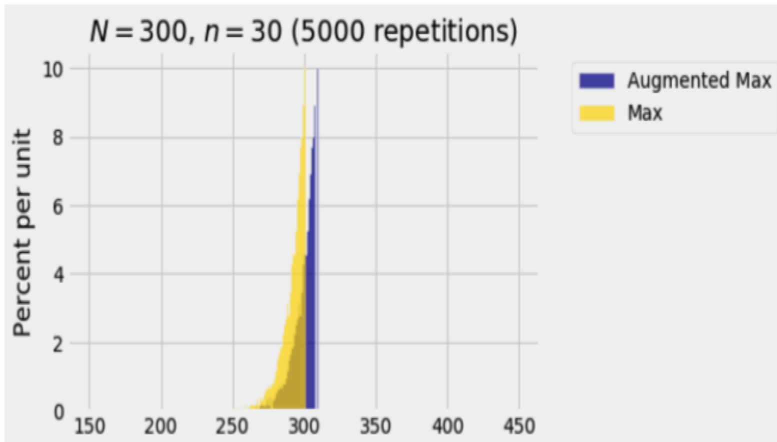
$$T_3 = \frac{n+1}{n} \cdot T_2 - 1,$$

it is an unbiased estimator of N.

- How does $\mathrm{Var}(T_3)$ and $\mathrm{Var}(T_2)$ compare?
- How does $B(T_3)$ and $B(T_2)$ compare?

So for large n, T_3 is better than T_2.

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



— unbiased and small var

— biased and small var

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 $\text{Var}(T_3) \approx \text{Var}(T_2)$. The estimator with the smallest MSE $= \text{Bias}^2 + \text{Var}$ is the best.

Another way to think of the "augmented maximum" If we could see the gap to the right of T_2, we would see N. But we can't. So we can try to do the next best thing, which is to augment T_2 by the estimated size of that gap.

What is the estimated gap length?

Hence we can try to improve upon T_2 using the estimator

$$T_3 =$$