

video 1: RANDOM SAMPLING AND ESTIMATION BIAS

wbg231

December 2022

1 introduction

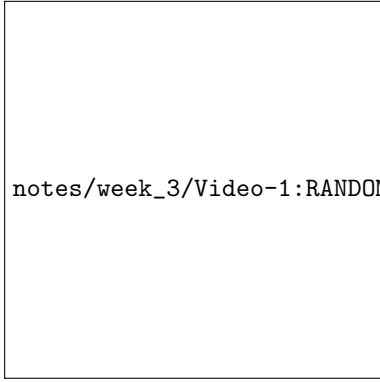
- video link

2 problem set up

- the goal is to estimate a population parameter
- for instance we could want to understand the average weight of rats in NYC. in principle we could catch all rats weigh them and find the average
- so we could chose a subset of the rats find there average weight and use that as an estimate

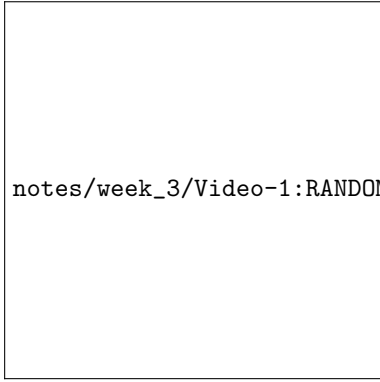
3 estimation a population mean

- suppose we are looking at the heights of a population of $N = 400$ people
- Heights: $h_1..H_N$
- population mean $\mu_{pop} = \frac{1}{N} \sum_{i=1}^N h_i = 175.6$ note that this is a fixed value we want to estimate



notes/week_3/Video-1:RANDOM-SAMPLING-AND-ESTIMATION-BIAS/images/v1_1.jpg

-
- so just keep in mind that our sample average is a random variable our population parameter is a constant
- here is the distribution of the sample mean of size 400



notes/week_3/Video-1:RANDOM-SAMPLING-AND-ESTIMATION-BIAS/images/v2_2.jpg

-
- so know that our sample mean is centered at the population mean and has pretty low variance
- this tells us with high probability our estimate will be close to the population parameter

4 estimating a population proportion

- instead of just getting a population mean we could be interested in a population proportion for instance the proportion of people in NYC with COVID-19 (ie COVID prevalence) call that population proportion $\theta_{pop} = 0.05$ the true constant we want to estimate
- we can try to estimate this with a random sample as well.
- as before our random sample proportion is a random variable we are using to estimate a fixed quantity θ_{pop}

5 random sampling for sample mean

- data $a_1 \dots a_n$ this is our fixed data from the population
- random sample $\tilde{x}_1 \dots \tilde{x}_n$ these are random variables
- we assume that each \tilde{x}_i is selected independently and uniformly at random with replacement (so we could pick someone twice but that will happen with low likely hood in large population)
- independent means it does not matter who we picked before we pick the next person the same way
- uniformly means we have the same likely hood of picking anyone in the population
- samples (each \tilde{x}_i) are IID (independent and identically distributed) random variables with pmf

$$P_{\tilde{x}_i} = P(\tilde{x}_i = a_i) = \frac{1}{n}$$

so note that our random sample is discrete and has equal likely hood of being equal to any individual in our population

- here because our data set is random samples our sample average will be a random viable \tilde{m} such that

$$\tilde{m} = \frac{1}{n} \sum_{i=1}^n \tilde{x}_i$$

recall that N is the number of individuals in our population and n is the number of individuals in our sample

6 random sampling for sample proportion

- data $a_1 \dots a_N$ this is our fixed data from the population, where $a_i = 1$ if a person meets the condition (ie has COVID) or zero otherwise
- random sample $\tilde{x}_1 \dots \tilde{x}_n$ these are random variables
- we assume that each \tilde{x}_i is selected independently and uniformly at random with replacement (so we could pick someone twice but that will happen with low likely hood in large population)
- independent means it does not matter who we picked before we pick the next person the same way
- uniformly means we have the same likely hood of picking anyone in the population

- samples (each \tilde{x}_i) are IID (independent and identically distributed) random variables with pmf

$$P_{\tilde{x}_i} = P(\tilde{x}_i = a_i) = \frac{1}{n}$$

so note that our random sample is discrete and has equal likelihood of being equal to any individual in our population

- here because our data set is random samples our sample proportion will be a random variable and because the data only takes values zero one our sample proportion will be our sample mean \tilde{m} such that

$$\tilde{m} = \frac{1}{n} \sum_{i=1}^n \tilde{x}_i$$

recall that N is the number of individuals in our population and n is the number of individuals in our sample

- so the main point is these problems are the same

7 estimation of population parameters

- note we are taking a frequentist perspective, ie we assume that the parameter of interest is deterministic (this is a choice as opposed to Bayesian inference)
- so we want to think about the probabilistic behavior of the estimator

8 bias

- is the estimator centred at the parameter of interest?

8.1 bias definition

- for random measurements $\tilde{x}_1 \dots \tilde{x}_n$
- and deterministic parameter of interest $\gamma \in \mathbb{R}$
- and estimator as a function of the random samples $h(\tilde{x}_1 \dots \tilde{x}_n)$
- the bias of the estimator is the mean of the error that is

$$\text{bias} = E[h(\tilde{x}_1 \dots \tilde{x}_n) - \gamma]$$

- if $E[h(\tilde{x}_1 \dots \tilde{x}_n)] = \gamma$ the estimator is unbiased
- so this is asking is this thing centered?

8.2 sample mean

- what is the mean of the sample mean? $E[\tilde{m}] = E[\frac{1}{n}\sum_{i=1}^n \tilde{x}_i] = \frac{1}{n}\sum_{i=1}^n E[\tilde{x}_i]$
- the samples are from the population so $E[\tilde{x}_i] = \sum_{j=1}^N a_j P(\tilde{x}_i = a_j) = \frac{1}{n}\sum_{j=1}^n a_j = \frac{1}{n}\sum_{j=1}^N a_j = \mu_{pop}$
- so taking this back over we can get $E[\tilde{m}] = E[\frac{1}{n}\sum_{i=1}^n \tilde{x}_i] = \frac{1}{n}\sum_{i=1}^n E[\tilde{x}_i] = \frac{1}{n}\sum_{i=1}^n \mu_{pop} = \frac{n}{n}\mu_{pop} = \mu_{pop}$
- so the take away is that the sample mean is an unbiased estimator of the population mean

8.3 sample proportion

- recall that the sample proportion is the sample mean ie $\tilde{m} = \frac{1}{n}\sum_{j=1}^n \tilde{x}_j$
- so we can do a similar argument $E[\tilde{x}_j] = \sum_{i=1}^N a_i p_{\tilde{x}_j}(a_i) = \sum_{i=1}^N a_i P(\tilde{x}_j = a_i) = \frac{1}{N}\sum_{i=1}^N a_i = \frac{\text{number of covid cases}}{N}$
- then we can see that $E[\tilde{m}] = E[\frac{1}{n}\sum_{j=1}^n \tilde{x}_j] = \frac{1}{n}\sum_{j=1}^n E[\tilde{x}_j] = \frac{1}{n}\sum_{j=1}^n \frac{\text{number of covid cases}}{N} = \frac{\text{number of covid cases}}{N} = \theta_{pop}$ so the population parameter is also unbiased
- notice that we are assuming that the tests are perfect in this case as well
- so the fact that the estimator is unbiased is why we observe this behavior with the data centred around the true mean

notes/week_3/Video-1:RANDOM-SAMPLING-AND-ESTIMATION-BIAS/images/v1_2.jpg

8.4 sample variance

- we know that the population mean $\mu_{pop} = \frac{1}{N}\sum_{i=1}^N a_i$
- and that the population variance is $\sigma_{pop}^2 = \frac{1}{N}\sum_{i=1}^N (a_i - \mu_{pop})^2$
- that is the distance of each value from the population mean

- so we can define the sample variance as

$$\tilde{v} = \frac{1}{n-1} \sum_{j=1}^n (\tilde{x}_j - \tilde{m})^2$$

(we need to subtract 1 from the denominator to make sure it is unbiased)

- we end up finding that $E[\tilde{v}] = \sigma_{pop}^2$
- the derivation is kinda pain full. so i am not going to write it out