

Introduction to Statistics Lecture Notes

ALECK ZHAO

September 13, 2016

This is EN.550.430 Introduction to Statistics, taught by Avanti Athreya.

Contents

1	September 7, 2016	2
1.1	Inferential Statistics	2
1.2	Finite Population Sampling	2
1.3	Simple Random Sampling and Unconscious Statisticians.	3
1.4	Multiple Observation Sampling	4
2	September 12, 2016	5
2.1	Review of Terms	5
2.2	Variance of \bar{X}	5
2.3	Sampling With Replacement	6
2.4	Sampling Without Replacement	6

1 September 7, 2016

1.1 Inferential Statistics

Definition 1.1 (Inferential Statistics). Using information from a sample to draw conclusions about a population.

Example 1.2

Some examples...

1. Sample of newborn birth weights to understand mean birth weights in a country.
2. Sample of respondents to a survey whether or not household has Direct TV. We refer to this as a **dichotomous** case because there are two possible values a response could take - yes or no.

1.2 Finite Population Sampling

Let a population be represented by $\{X_1, X_2, \dots, X_N\}$ where N is the size of the population. We have the **population parameters**:

- **total** $\tau := \sum_{i=1}^N X_i$
- **mean** $\mu := \tau/N$
- **variance** $\sigma^2 := \frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2$

Remark 1.3. None of these are random values, they are deterministic. This distinction will become important.

Proposition 1.4

In the dichotomous case, $\mu = p$ (the population **proportion**) and $\sigma^2 = p(1 - p)$.

Proof. Since $X_i = 1$ if the observation is true and $X_i = 0$ if false, the sum is just the number of true observations, so divided by the total number of observations gives the proportion.

Next, observe that

$$\begin{aligned}
 \sigma^2 &= \frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^N (X_i^2 - 2\mu X_i + \mu^2) \\
 &= \frac{1}{N} \left(\sum_{i=1}^N X_i^2 - 2\mu \sum_{i=1}^N X_i + \sum_{i=1}^N \mu^2 \right) \\
 &= \frac{1}{N} \left(\sum_{i=1}^N X_i^2 - 2\mu(\mu N) + N\mu^2 \right) = \frac{1}{N} \left(\sum_{i=1}^N X_i^2 - \mu^2 \right)
 \end{aligned}$$

In the dichotomous case,

$$\frac{1}{N} \sum_{i=1}^N X_i^2 = p$$

since $X_i^2 = X_i$ because X_i takes on either 0 or 1, and $\mu^2 = p^2$, so

$$\sigma^2 = p - p^2 = p(1 - p).$$

□

1.3 Simple Random Sampling and Unconscious Statisticians...

Draw a sample of n measurements from population: $\{X_1, X_2, \dots, X_n\}$, and there are $\binom{N}{n}$ different possible subsets. Recall that $\{X_1, X_2, \dots, X_N\}$ is the population. Then let $\{\xi_1, \xi_2, \dots, \xi_m\}$ represent the set of all distinct observation values of X_i . Note in particular that $m \leq N$.

Consider a random draw X_i from the population so that $X_i \in \{\xi_1, \xi_2, \dots, \xi_m\}$. Define the distribution

$$P(X_i = \xi_\ell) = \frac{\# \text{ of } \xi_\ell \text{ in population}}{N} = \frac{n_\ell}{N},$$

then we can compute

$$\begin{aligned} E[X_i] &= \sum_{j=1}^m \xi_j P(X_i = \xi_j) \\ &= \sum_{j=1}^m \xi_j \cdot \frac{n_j}{N} = \frac{1}{N} \sum_{j=1}^m \xi_j n_j = \frac{1}{N} \tau = \mu. \end{aligned}$$

The crucial step is the equality $\sum_{j=1}^m \xi_j n_j = \tau$. Think about what ξ_j and n_j mean, and this becomes obvious.

We may think of X_i as a way to estimate μ . It is **unbiased** because $E[X_i] = \mu$, but it may not be a “great” estimate. We formulate a more specific notion of “greatness” of an estimate by considering its variance: the smaller the variance, the better the estimate.

Theorem 1.5 (Law of the Unconscious Statistician)

Assume that g behaves nicely.

- If X is a discrete random variable with PMF $p_X(x)$, then $E[g(X)] = \sum_x g(x)p_X(x)$
- If X is a continuous random variable with PDF $f_X(x)$, then $E[g(X)] = \int_x g(x)f_X(x) dx$

This basically means the distribution of $g(X)$ has the same structure as the distribution of X . Using this theorem, we can derive the variance of X_i from earlier:

$$\begin{aligned} V(X_i) &= E[X_i^2] - (E[X_i])^2 = E[X_i^2] - \mu^2 \\ &= \sum_{j=1}^m \xi_j^2 P(X_i = \xi_j) - \mu^2 \\ &= \frac{1}{N} \sum_{j=1}^m \xi_j^2 n_j - \mu^2 = \frac{1}{N} \sum_{j=1}^m X_i^2 - \mu^2 = \sigma^2 \end{aligned}$$

thus $V(X_i)$ is an unbiased estimator for σ^2 .

1.4 Multiple Observation Sampling

Now we shift from a single draw X_i to a set of multiple draws. Given a sample of size n , the **sample mean** is defined as $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. Now,

$$E[\bar{X}] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{1}{n} n\mu = \mu$$

so \bar{X} is an unbiased estimator for μ as well.

Computing the variance of \bar{X} is slightly complicated by the different ways we can sample the observations. We address the case where we sample with replacement. Then

$$\begin{aligned} V(\bar{X}) &= V\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} V\left(\sum_{i=1}^n X_i\right) \\ &= \frac{1}{n^2} \left(E\left[\left(\sum_{i=1}^n X_i\right)^2\right] - \left(E\left[\sum_{i=1}^n X_i\right]\right)^2 \right) \\ &= \frac{1}{n^2} \left(E\left[\left(\sum_{i=1}^n X_i\right)^2\right] - n^2 \mu^2 \right) \end{aligned}$$

and discussion of this is left to next week.

2 September 12, 2016

2.1 Review of Terms

From last time, we defined a population to consist of a set of observations $\{X_1, X_2, \dots, X_N\}$ where N is the size of the population.

The **population parameters** are defined as

$$\mu = \frac{1}{N} \sum_{i=1}^N X_i$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2$$

We considered a single random draw X_i from the population, and we showed that $E[X_i] = \mu$, thus X_i is an unbiased estimator for μ . We also showed that $V(X_i) = \sigma^2$. Refer to the previous day's notes to confirm these.

Next we considered the random sample $\{X_1, X_2, \dots, X_n\}$ where $n \leq N$ and the **sample mean**

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

We showed that $E[\bar{X}] = \mu$ so that \bar{X} is an unbiased estimator for μ as well. Next, we set out to compute the variance of \bar{X} , which is where we will continue the discussion.

2.2 Variance of \bar{X}

The variance of \bar{X} is

$$\begin{aligned} V(\bar{X}) &= V\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} V\left(\sum_{i=1}^n X_i\right) \\ &= \frac{1}{n^2} \left(E\left[\left(\sum_{i=1}^n X_i\right)^2\right] - \left(E\left[\sum_{i=1}^n X_i\right]\right)^2 \right) \\ &= \frac{1}{n^2} \left(E\left[\left(\sum_{i=1}^n X_i\right)^2\right] - (n\mu)^2 \right) \end{aligned}$$

Example 2.1 ($n = 2$)

When $n = 2$, this expected value is equivalent to

$$\begin{aligned} V(\bar{X}) &= \frac{1}{2^2} (E[(X_1 + X_2)^2] - (2\mu)^2) \\ &= \frac{1}{4} (E[X_1^2 + 2X_1X_2 + X_2^2] - 4\mu^2) \\ &= \frac{1}{4} [(E[X_1^2] - (E[X_1])^2) + 2(E[X_1X_2] - E[X_1]E[X_2]) + (E[X_2^2] - (E[X_2])^2)] \\ &= \frac{1}{4} [V(X_1) + 2cov(X_1, X_2) + V(X_2)] \end{aligned}$$

Note the substitutions for μ^2 , and recall that covariance is $cov(X_1, X_2) = E[X_1X_2] - E[X_1]E[X_2]$.

This example motivates us to write the variance of \bar{X} purely in terms of the covariances of the X_i , since $V(X_i) = \text{cov}(X_i, X_i)$. We may write

$$V(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \text{cov}(X_i, X_j),$$

and this is indeed the correct expression since there is one copy of $\text{cov}(X_i, X_i)$ for each i and two copies of $\text{cov}(X_i, X_j)$ whenever $i \neq j$ since $\text{cov}(X_i, X_j) = \text{cov}(X_j, X_i)$.

Now we must consider two different methods of sampling: with replacement and without replacement.

2.3 Sampling With Replacement

When we sample with replacement, each draw is independent from every other draw because they are all taken from the same population. Since X_i is independent from X_j whenever $i \neq j$, it follows that $\text{cov}(X_i, X_j) = 0$, so all terms in the double summation vanish except where $i = j$, so the sum is exactly

$$V(\bar{X}) = \frac{1}{n^2} \sum_{k=1}^n \text{cov}(X_k, X_k) = \frac{1}{n^2} (n\sigma^2) = \frac{\sigma^2}{n}.$$

Remark 2.2. Note that as $n \rightarrow \infty$, the sample variance $\frac{\sigma^2}{n} \rightarrow 0$, so then $\bar{X} \rightarrow \mu$. This is known as the [Law of Large Numbers](#).

2.4 Sampling Without Replacement

Next we examine sampling without replacement. To do so, we must understand $\text{cov}(X_i, X_j)$, since these comprise all terms of the sum of interest.

Claim 2.3. $\text{cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}$

Proof. This derivation is very, very long and tedious, but it follows the ideas presented in lecture. Please bear with me.

By the definition of covariance, we have

$$\text{cov}(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j] = E[X_i X_j] - \mu^2.$$

To compute $E[X_i X_j]$, we can treat $X_i X_j$ as a random variable whose density function is the joint density, and evaluate the sum

$$\sum_{x,y} (xy) \cdot P(X_i = x, X_j = y).$$

Recall the terminology from the previous lecture, where we define the set of distinct observations $\{\xi_1, \xi_2, \dots, \xi_m\}$ where $m \leq N$. These are just the distinct values of all the X_k in the population. We also have n_k is the number of times ξ_k appears in the population, that is, how many of the X_k take on the value ξ_k . Using this change of variables, we can rewrite the sum as

$$\sum_{k,\ell} \xi_k \xi_\ell P(X_i = \xi_k, X_j = \xi_\ell) = \sum_{k=1}^m \sum_{\ell=1}^m \xi_k \xi_\ell P(X_i = \xi_k, X_j = \xi_\ell).$$

The draws X_i and X_j are not independent because the result of X_i affects X_j . To understand this, imagine a bag with 2 blue and 3 red marbles. If I draw a blue marble first with probability 2/5, the probability I draw

a blue marble on the second try is only $1/4$, so the draws are not independent. To complete the analogy, the colors of the marbles are the distinct ξ_k , and we are drawing from them.

Since the draws are not independent, it makes sense to condition on one of them. We can rewrite

$$P(X_i = \xi_k, X_j = \xi_\ell) = P(X_i = \xi_k | X_j = \xi_\ell)P(X_j = \xi_\ell)$$

where we have conditioned the value of X_i on what X_j turned out to be. Then we can rewrite the double sum as

$$\begin{aligned} & \sum_{k=1}^m \sum_{\ell=1}^m \xi_k \xi_\ell P(X_i = \xi_k | X_j = \xi_\ell) P(X_j = \xi_\ell) \\ &= \sum_{\ell=1}^m \left(\xi_\ell P(X_j = \xi_\ell) \cdot \sum_{k=1}^m \xi_k P(X_i = \xi_k | X_j = \xi_\ell) \right) \end{aligned}$$

where we were allowed to move the $\xi_\ell P(X_j = \xi_\ell)$ outside the inner summation because it does not have any k in it.

Consider the right hand sum,

$$\sum_{k=1}^m \xi_k P(X_i = \xi_k | X_j = \xi_\ell).$$

Note that as k varies, $k \neq \ell$ most of the time, except when $k = \ell$. In the case that $k = \ell$, that particular contribution to the sum is

$$\xi_\ell P(X_i = \xi_\ell | X_j = \xi_\ell).$$

We want the probability that $X_i = \xi_\ell$ given that X_j was already ξ_ℓ . Since one copy of ξ_ℓ was already drawn as X_j , there are $n_\ell - 1$ copies remaining, out of a total of $N - 1$ possibilities, so the probability is $\frac{n_\ell - 1}{N - 1}$, and the contribution is $\xi_\ell \frac{n_\ell - 1}{N - 1}$.

In the other cases, $k \neq \ell$, so that after X_j is determined to be ξ_ℓ , there are still n_k copies of ξ_k remaining, out of a total of $N - 1$, so their contributions are all $\xi_k \frac{n_k}{N - 1}$. Thus, the sum is split as

$$\begin{aligned} \xi_\ell \frac{n_\ell - 1}{N - 1} + \sum_{k \neq \ell, k=1}^m \xi_k \frac{n_k}{N - 1} &= \xi_\ell \frac{n_\ell}{N - 1} - \frac{\xi_\ell}{N - 1} + \sum_{k \neq \ell, k=1}^m \xi_k \frac{n_k}{N - 1} \\ &= -\frac{\xi_\ell}{N - 1} + \sum_{k=1}^m \xi_k \frac{n_k}{N - 1} \end{aligned}$$

where the value $\xi_\ell \frac{n_\ell}{N - 1}$ was incorporated into the summation. Finishing, this is equal to

$$-\frac{\xi_\ell}{N - 1} + \frac{1}{N - 1} \sum_{k=1}^m \xi_k n_k = \frac{\tau - \xi_\ell}{N - 1}.$$

Note that the sum here is equal to τ the total sum of all observations in the population. Think about what ξ_k and n_k mean if this is not clear.

Finally, we compute the summation

$$\begin{aligned} & \sum_{\ell=1}^m \xi_\ell P(X_j = \xi_\ell) \frac{\tau - \xi_\ell}{N - 1} = \sum_{\ell=1}^m \xi_\ell \frac{n_\ell}{N} \frac{\tau - \xi_\ell}{N - 1} = \frac{1}{N(N - 1)} \sum_{\ell=1}^m \xi_\ell n_\ell (\tau - \xi_\ell) \\ &= \frac{1}{N(N - 1)} \sum_{\ell=1}^m \xi_\ell n_\ell (N\mu - \xi_\ell) = \frac{1}{N(N - 1)} \left(N\mu \sum_{\ell=1}^m \xi_\ell n_\ell - \sum_{\ell=1}^m \xi_\ell^2 n_\ell \right) \\ &= \frac{1}{N(N - 1)} \left(N\mu(N\mu) - \sum_{i=1}^N X_i^2 \right) = \frac{N\mu^2}{N - 1} - \frac{1}{N(N - 1)} \sum_{i=1}^N X_i^2 \end{aligned}$$

Notice that summing $\xi_\ell^2 n_\ell$ is essentially taking each unique value in the population, squaring it, then multiplying by the number of times it appears. This is exactly the same as just taking the sum of the square of each value in the population. We know that

$$V(X) = E[X^2] - (E[X])^2 = \frac{1}{N} \sum_{i=1}^N X_i^2 - \mu^2 = \sigma^2$$

is the variance of the entire population, so it follows that

$$\sum_{i=1}^N X_i^2 = N(\mu^2 + \sigma^2).$$

Substituting this in above, we the expression becomes

$$\begin{aligned} \frac{N\mu^2}{N-1} - \frac{1}{N(N-1)} N(\mu^2 + \sigma^2) &= \frac{N\mu^2}{N-1} - \frac{\mu^2 + \sigma^2}{N-1} \\ &= \frac{\mu^2(N-1) - \sigma^2}{N-1} = \mu^2 - \frac{\sigma^2}{N-1} \end{aligned}$$

This is the value of $E[X_i X_j]$. To finally (yes, finally) compute the covariance $cov(X_i, X_j)$, we must subtract off $E[X_i]E[X_j] = \mu^2$, so then

$$\begin{aligned} cov(X_i, X_j) &= E[X_i X_j] - E[X_i]E[X_j] \\ &= \left(\mu^2 - \frac{\sigma^2}{N-1} \right) - \mu^2 \\ &= -\frac{\sigma^2}{N-1}, \end{aligned}$$

as desired. □

Hopefully this clears up Dr. Athreya's lecture, which I noticed had many "typos" and confusing switching of indices.