Statistics

Week 3: Sampling Distributions (Chapter 5)

ESD, SUTD

Term 5, 2017

Outline

1. Sampling distributions
2. Estimators
   - German tank problem

Random variable

Motivating example: before an election, we wish to estimate the population proportion p of people who plan to vote for a particular candidate. We can survey 100 random voters and estimate p.

Each time we run such a survey, the sequence of responses, and the resulting estimate, will generally be different.

Therefore, the data points we are getting are essentially acting like random variables, and can be treated (modeled) as such. In this case, each response can be modeled as a Bernoulli random variable with parameter p; moreover, the responses are independent.

The statistic we are interested in (the estimate for p) is also a random variable.

A *sampling distribution* is the probability distribution of a given statistic based on a random sample.

Sample mean

Consider a population with mean μ and variance σ^2. Let $X_1, X_2, \ldots, X_n$ be a random sample drawn from the population; note that X_i are independent and identically distributed (iid).

Notation: upper case letters (such as X_1, X_2) denote random variables; lower case letters (such as x_1, x_2) denote data values, which are the observed values of the random variables. However, when the context is clear, we will sometimes abuse the notation and use lower case letters to represent both.

How does the sample mean, $\bar{X}$, behave?

$\mathsf{E}(\bar{X}) =$

$\mathsf{Var}(\bar{X}) =$

Moral: larger n means better approximation to μ.

Sample mean – example

Suppose we have 2 electronic scales, whose readings are iid with mean = actual weight of object, and variance = σ^2.

Then it is actually more precise to weigh the same object on both scales, and take the average reading, than it is to just weigh it once.

We can demonstrate this effect using a 'toy' scale and some simple numbers...

Central limit theorem

Some special cases:

- If $X_i \sim N(\mu, \sigma^2)$, then $\bar{X} \sim N(\mu, \sigma^2/n)$.
- If X_i are Bernoulli with parameter p, then $P(\bar{X} = x/n) = P\left(\sum_{i=1}^n X_i = x\right) = \binom{n}{x} p^x (1-p)^{n-x}$ (binomial).

What about X_i drawn from an arbitrary distribution?

Central limit theorem (CLT)

Let $X_1, X_2, \ldots, X_n$ be iid random variables drawn from an arbitrary distribution with finite mean μ and variance σ^2. Then as $n \to \infty$,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to N(0, 1).$$

Excel demo: average of n loaded dice.

CLT; binomial

As a rule of thumb, the central limit theorem is used when the sample size is at least 30.

You should also recall the more elementary, but useful result:

Normal approximation to binomial

Suppose X is a Bin(n, p) random variable. For large n,

$$\frac{X - np}{\sqrt{np(1-p)}} \approx N(0, 1).$$

In practice, do not forget the continuity correction.

Exercise: how is the above a special case of the CLT?

Outline

1 Sampling distributions

2 Estimators
- German tank problem

Point estimator

Let x_i be random samples drawn from a population with an unknown parameter θ. A point *estimator* $\hat{\theta}$ is a statistic computed from x_i, and is used to estimate θ.

Example: if θ is the population mean μ, then we can use $\hat{\theta} = \bar{x}$.

The **bias** of an estimator is defined as $\mathrm{E}(\hat{\theta}) - \theta$. If the bias is 0, then the estimator is called *unbiased*.

Note: do not confuse 'bias' here with selection bias in sampling, which occurs when randomization is not achieved.

Exercise: is $\bar{x}$ an unbiased estimator of μ?

Unbiased estimator

How do we estimate σ^2 when only a sample of size n is taken?

We could try the unbiased estimator

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2,$$

the problem is that we do not know μ, and only know $\bar{x}$!

Exercise: check that the above estimator is unbiased, using $\mathrm{Var}(X) = \mathrm{E}((X-\mu)^2)$.

What about

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2?$$

Intuitively, a sample may miss some extreme values, so this estimator seems too small.

Sample variance

We can modify the factor in front of the $\sum$ to make the estimator unbiased.

Tricky calculation

$$\mathsf{E}\left[\sum_{i=1}^{n}(x_i-\mu)^2\right] = \mathsf{E}\left[\sum_{i=1}^{n}\Big((x_i-\bar{x})^2+(\bar{x}-\mu)^2+2(x_i-\bar{x})(\bar{x}-\mu)\Big)\right]$$

$$\Rightarrow\ n\,\sigma^2 = \mathsf{E}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right] + n\,\mathsf{Var}(\bar{x}) + 0$$

$$\Rightarrow\ (n-1)\,\sigma^2 = \mathsf{E}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right].$$

Therefore

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2,$$

is an unbiased estimator of σ^2, namely $\mathsf{E}(s^2)=\sigma^2$.

German tank problem

During World War II, the Allies had the serial numbers of some captured/destroyed German tanks and wanted to estimate the total number of German tanks.

They correctly guessed that the serial numbers ran from 1 to N, where N was the total number of tanks produced. How can one estimate N?

Month of production	Intelligence estimate	Statistical estimate	Actual (from post-war captured documents)
June 1940	1000	169	122
June 1941	1550	244	271
August 1942	1550	327	342

Exercise

Goal: estimate the maximum value N of a discrete uniform random variable taking values from $\{1, 2, \ldots, N\}$, given a sample of size n drawn without replacement.

Let the sample be $\{x_1, x_2, \ldots, x_n\}$. Some possible estimators for N are:

- $\bar{x} + 3s$
- $2\bar{x} - 1$
- $x_{(n)}$
- $x_{(n)} + x_{(1)} - 1$
- $\dfrac{n+1}{n} x_{(n)} - 1$

The *Excel* sheet provides 500 simulated samples, each of size $n = 5$, from a population of $N = 250$. Evaluate the estimators and compare the bias.

Solution

(1) This estimator is inspired by the normal distribution. Biased.

(2) $\mathsf{E}(\bar{x}) = \frac{1}{5}\big(\mathsf{E}(x_1) + \cdots + \mathsf{E}(x_5)\big) = \frac{N+1}{2}$. Hence $\mathsf{E}(2\bar{x} - 1) = N$ (unbiased). Note however that this may be less than $x_{(n)}$.

(3) $x_{(n)} \leq N$ always, so it is a biased estimator.

(4) This estimator is based on the last one, as well as symmetry. Unbiased.

(5) It turns out that this is unbiased, and has the *least variance* among all unbiased estimators.

Proof sketch of unbiasedness: the probability that the maximum, $x_{(n)}$, equals m is given by

$$P(x_{(n)} = m) = \frac{\binom{m-1}{n-1}}{\binom{N}{n}}.$$

Multiply this by m and sum:

$$\begin{aligned}
\mathsf{E}(x_{(n)}) &= \sum_{m=n}^{N} m\, P(x_{(n)} = m) \\
&= \sum_{m=n}^{N} m \frac{\binom{m-1}{n-1}}{\binom{N}{n}} = \frac{n}{\binom{N}{n}} \sum_{m=n}^{N} \binom{m}{n} \\
&= \frac{n}{\binom{N}{n}} \binom{N+1}{n+1} = \frac{n(N+1)}{n+1}.
\end{aligned}$$

Rearranging gives the required result.