Sampling Distributions

CSU, Chico Math 314

2018-10-01

outline

outline

Review: the sample mean shows up everywhere

We assume data $x_1, \ldots, x_n$ are generated from distribution $G(\theta)$. Given these data, the most likely value of the paramter θ is $\hat{\theta}$ (often some function of $\bar{X}$).

outline

Is $\bar{X}$ good at its job?

How good is $\bar{X}$ at its job? Could we simulate an (not the) answer?
http://onlinestatbook.com/stat_sim/sampling_dist/

Is $\bar{X}$ good at its job?

This is like 101 of us each took a random sample $n = 100$ from the same population and each calculated our own sample mean and sample standard deviation.

```
# 101 random samples each of size 100 from N(0, 1)
X <- matrix(rnorm(100*101), nrow=100)
xbars <- colMeans(X) # 101 sample means
mean(xbars) # mean of the means

## [1] -0.0002633851

sd(xbars) # standard error = standard deviation of means

## [1] 0.09889086
```

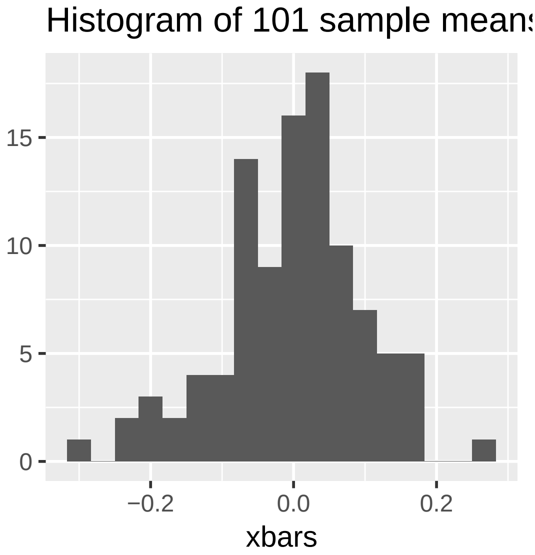
Is $\bar{X}$ good at its job?

What does the distribution of sample means look like?

```
suppressMessages(library(ggplot2))
p <- qplot(xbars, geom="histogram", binwidth=1/30,
           main="Histogram of 101 sample means")
```

Is $\bar{X}$ good at its job?

The sample means seem to on average estimate $\mu = 0$ quite well.



outline

Point Estimators, idea

Since data will vary from sample to sample, $\hat{\theta}$ will vary from sample to sample. But, then we should really be thinking about data theoretically; instead of $x_1, \ldots, x_n$ we should be thinking $X_1, \ldots, X_n$.

Point Estimators, definition

That is to say, data not yet collected are random variables. Hence, the **point estimator** $\bar{X}$ is a random variable.

point estimator

A statistic that estimates a single value, e.g. $\bar{X}, \hat{\sigma}, \hat{p}, \ldots$

Point Estimators are Random Variables

If point estimators are random variables, and random variables have distribution functions, then point estimators have distribution functions. This means that point estimators, follow some functional form and thus have a mean and a variance of their own.

Mean of $\bar{X}$

Assume $X_1, \ldots, X_n \sim_{iid} G$, such that $E(X) = \mu$. Calculate $E(\bar{X})$.

$$
\begin{aligned}
E(\bar{X}) &= E(n^{-1} \sum_{i=1}^{n} X_i) \\
&= n^{-1} \sum_{i=1}^{n} E(X_i) \\
&= n^{-1} n\mu \\
&= \mu.
\end{aligned}
$$

Standard Error

The **standard error** describes the error/uncertainty/inaccuracy of an estimator.

standard error

The standard deviation of a point estimator is called the **standard error**.

Variance of $\bar{X}$

Assume $X_1, \ldots, X_n \sim_{iid} G$, such that $Var(X) = \sigma^2$. Calculate $Var(\bar{X})$.

$$
\begin{aligned}
\sigma^2_{\bar{X}} = Var(\bar{X}) &= Var(n^{-1} \sum_{i=1}^{n} X_i) \\
&= n^{-2} Var(\sum_{i=1}^{n} X_i) \\
&= n^{-2} \sum_{i=1}^{n} Var(X_i) \qquad\qquad \text{(because } X_i \text{ independent)} \\
&= n^{-2} n \sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

Standard Error of $\bar{X}$

Since the standard error is the point estimator analogue to the standard deviation, it holds the same relation to the variance of a point estimator.

standard error of $\bar{X}$

The standard error of $\bar{X}$ is

$$\sigma_{\bar{X}} = \sqrt{\sigma^2_{\bar{X}}} = \frac{\sigma}{\sqrt{n}}.$$

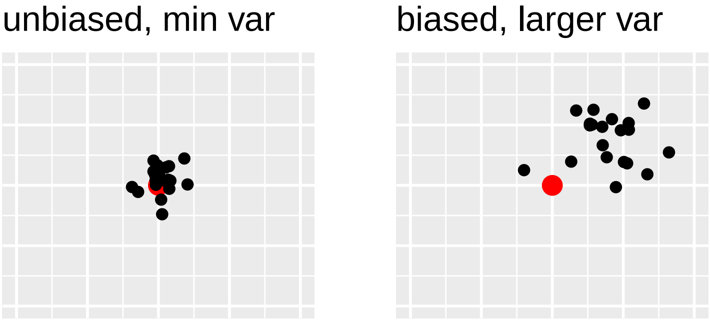
outline

Properties of Point Estimators

We want point estimators to be good at their job. Good, here, can mean many different things, but a common definition requires two criteria

1. un**bias**ed
2. minimum variance.

Summarizing Properties of $\bar{X}$, in picture



Bias

Statistics unfortunately uses the word **bias** in two different ways, sample bias and estimator bias.

(estimator) bias

Bias is defined as the difference between the expected value of an estimator $\hat{\theta}$ and the true parameter value

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

$\bar{X}$ is unbiased

Amongst its good properties, the sample mean is unbiased.
Assume $X_1, \ldots, X_n \sim_{iid} G$, such that $E(X) = \mu$.

$$
\begin{aligned}
Bias(\bar{X}) &= E(\bar{X}) - \mu \\
&= \mu - \mu \\
&= 0.
\end{aligned}
$$

$\bar{X}$ has minimum variance

Amongst its good properties, the sample mean has minimum variance. Assume $X_1, \ldots, X_n \sim_{iid} G$. If $T(\mathbf{X}) = \bar{X}$ and $T'(\mathbf{X})$ is another estimator of $E(X)$, then

$$Var(T(\mathbf{X})) \leq Var(T'(\mathbf{X})).$$

Summarizing Properties of $\bar{X}$, in English

To summarize in English the above properties of $\bar{X}$, we'd say

- on average $\bar{X}$ estimates the thing it's supposed to – $Bias(\bar{X}) = 0$
- $\bar{X}$ estimates the thing it's supposed to with smaller error than any other estimator – minimum variance

outline

Sampling Distribution

While it's nice to know the expected value and variance of an estimator, we would still like to know the distribution function (the functional form) of our estimator. We use the phrase **sampling distribution** to describe the distribution of a point estimator.

sampling distribution

The sampling distribution is the distribution of a point estimator based on a sample of fixed size.

N.B.

any given point estimate can be thought of as a random variable sampled from the sampling distribution.

outline

Central Limit Theorem

The **Central Limit Theorem** describes the sampling distribution of the sample mean.

Central Limit Theorem

Assume $X_1, \ldots, X_n \sim_{iid} G$, such that $E(X) = \mu$ and $Var(X) = \sigma^2 < \infty$. Then $\bar{X}$ is approximately distributed as $N(\mu, \sigma^2/n)$, or

$$\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \dot{\sim} N(0, 1).$$

CLT, notes

The CLT

- requires independent, identically distributed observations with finite variance,
- is an approximation that depends on the sample size n,
  - the approximation improves as $n \to \infty$,
- is population distribution invariant in the limit,
- in practice (finite world) requires a relatively larger sample size when the population distribution is abnormal,
- relies on unknown parameters.

outline

references I

Michael G. Akritas. *Probability and Statistics with R for Engineers and Scientists*. Pearson Education, Inc., 2016.