Basic Statistics and a Bit of Bootstrap

David S. Rosenberg

CDS, NYU

April 16, 2019

Contents

Bias and Variance

Parameters

- Suppose we have a probability distribution P.
- Often we want to estimate some characteristic of P.
  - e.g. expected value, variance, kurtosis, median, etc...
- These things are called **parameters** of P.
- A **parameter** $\mu = \mu(P)$ is any function of the distribution P.
- Question: Is μ random?
- Answer: Nope. For example if P has density $f(x)$ on R, then mean is

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx,$$

which is just an integral - nothing random.

Statistics and Estimators

- Suppose $\mathcal{D}_n = (x_1, x_2, \ldots, x_n)$ is an i.i.d. sample from P.
- A **statistic** $s = s(\mathcal{D}_n)$ is any function of the data.
- A statistic $\hat{\mu} = \hat{\mu}(\mathcal{D}_n)$ is a **point estimator** of μ if $\hat{\mu} \approx \mu$.
- Question: Are statistics and/or point estimators random?
- Answer: Yes, since we're considering the data to be random.
  - The function $s(\cdot)$ isn't random, but we're plugging in random inputs.

Examples of Statistics

- Mean: $\bar{x}(\mathcal{D}_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$.
- Median: $m(\mathcal{D}_n) = \text{median}(x_1, \ldots, x_n)$
- Sample variance: $\sigma^2(\mathcal{D}_n) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}(\mathcal{D}_n))^2$

Fancier:

- A data histogram is a statistic.
- Empirical distribution function.
- A confidence interval.

Statistics are Random

- Statistics are random, so they have probability distributions.
- The distribution of a statistic is called a **sampling distribution**.
- We often want to know some **parameters** of the sampling distribution.
  - Most commonly the mean and the standard deviation.
- The standard deviation of the sampling distribution is called the **standard error**.
- Question: Is standard error random?
- Answer: Nope. It's a parameter of a distribution.

Bias and Variance for Real-Valued Estimators

- Let $\mu = \mu(P)$ be a real-valued parameter of distribution P.
- Let $\hat{\mu} = \hat{\mu}(\mathcal{D}_n)$ be a point estimator of μ.
- We define the **bias** of $\hat{\mu}$ to be $\text{Bias}(\hat{\mu}) = \mathbb{E}\hat{\mu} - \mu$.
- An estimator is **unbiased** if $\text{Bias}(\hat{\mu}) = \mathbb{E}\hat{\mu} - \mu = 0$.
- We define the **variance** of $\hat{\mu}$ to be $\text{Var}(\hat{\mu}) = \mathbb{E}\hat{\mu}^2 - (\mathbb{E}\hat{\mu})^2$.

Neither bias nor variance depend on a specific sample $\mathcal{D}_n$. We are taking expectation over $\mathcal{D}_n$.

- Why might we care about the bias and variance of an estimator?

Putting “Error Bars” on Estimators

- Why do we even care about estimating variance?
- May want to report a confidence interval for our point estimate, e.g.

$$\hat{\mu} \pm \sqrt{\widehat{\operatorname{Var}(\hat{\mu})}}$$

- Where $\sqrt{\widehat{\operatorname{Var}(\hat{\mu})}}$ is our **estimate of the standard error** of $\hat{\mu}$.

Estimating Variance of an Estimator

- To estimate $\text{Var}(\hat{\mu})$, we can use estimates of $\mathbb{E}\hat{\mu}$ and $\mathbb{E}\hat{\mu}^2$.
- Instead of a single sample $\mathcal{D}_n$ of size n, suppose we had
  - B independent samples of size n: $\mathcal{D}_n^1, \mathcal{D}_n^2, \ldots, \mathcal{D}_n^B$
- Can then estimate

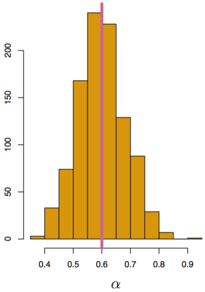
$$\begin{aligned}
\mathbb{E}\hat{\mu} &\approx \frac{1}{B}\sum_{i=1}^{B}\hat{\mu}\left(\mathcal{D}_n^i\right) \\
\mathbb{E}\hat{\mu}^2 &\approx \frac{1}{B}\sum_{i=1}^{B}\left[\hat{\mu}\left(\mathcal{D}_n^i\right)\right]^2
\end{aligned}$$

and

$$\mathrm{Var}\left(\hat{\mu}\right) \approx \frac{1}{B}\sum_{i=1}^{B}\left[\hat{\mu}\left(\mathcal{D}_n^i\right)\right]^2 - \left[\frac{1}{B}\sum_{i=1}^{B}\hat{\mu}\left(\mathcal{D}_n^i\right)\right]^2.$$

Histogram of Estimator

- Want to estimate $\alpha = \alpha(P) \in \mathrm{R}$ for some unknown P, and some complicated α.
- Point estimator $\hat{\alpha} = \hat{\alpha}(\mathcal{D}_{100})$ for samples of size 100.
- How to get error bars on $\hat{\alpha}$?
- Histogram of $\hat{\alpha}$ for 1000 random datasets of size 100 (estimates sampling distribution of $\hat{\alpha}$):



Pink line indicates true value of α. This is Figure 5.10 from *An Introduction to Statistical Learning, with applications in R* (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

Practical Issue

- We typically get only one sample $\mathcal{D}_n$.
- We could divide it into B groups.
- Could use first group as point estimator $\hat{\mu} = \hat{\mu}\left(\mathcal{D}_{n/B}^{(1)}\right)$,
- And use the remaining groups $\mathcal{D}_{n/B}^{(2)}, \ldots, \mathcal{D}_{n/B}^{(B)}$ to get a variance estimate for $\hat{\mu}(\mathcal{D}_{n/B}^{(1)})$
- But then our point estimate only uses a fraction of the data.
- Would be much better if we used all the data: $\hat{\mu} = \hat{\mu}(\mathcal{D}_n)$.
- Can we get the best of both worlds?
  - A good point estimate AND a variance estimate?

The Bootstrap

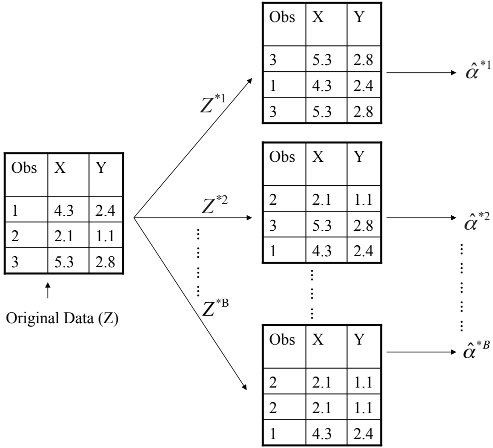
The Bootstrap Sample

- A **bootstrap sample** from $\mathcal{D}_n = (x_1, \ldots, x_n)$ is a sample of size n drawn *with replacement* from $\mathcal{D}_n$.
- In a bootstrap sample, some elements of $\mathcal{D}_n$
  - will show up multiple times, and
  - some won't show up at all.
- Each x_i has a probability of $(1 - 1/n)^n$ of not being selected.
- Recall from analysis that for large n,

$$\left(1 - \frac{1}{n}\right)^n \approx \frac{1}{e} \approx .368.$$

- So we expect ~63.2% of elements of $\mathcal{D}$ will show up at least once.

The Bootstrap Sample



From *An Introduction to Statistical Learning, with applications in R* (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

The Bootstrap Method

Definition

A **bootstrap method** is when you *simulate* having B independent samples from P by taking B bootstrap samples from the sample $\mathcal{D}_n$.

- Given original data $\mathcal{D}_n$, compute B bootstrap samples $D_n^1, \ldots, D_n^B$.
- For each bootstrap sample, compute some function

$$\phi(D_n^1), \ldots, \phi(D_n^B)$$

- Work with these values as though $D_n^1, \ldots, D_n^B$ were i.i.d. P.
- **Amazing fact:** Things often come out very close to what we'd get with independent samples from P.

Independent vs Bootstrap Samples

- Want to estimate $\alpha = \alpha(P)$ for some unknown P and some complicated α.
- Point estimator $\hat{\alpha} = \hat{\alpha}(\mathcal{D}_{100})$ for samples of size 100.
- Histogram of $\hat{\alpha}$ based on
  - 1000 independent samples of size 100, vs
  - 1000 bootstrap samples of size 100

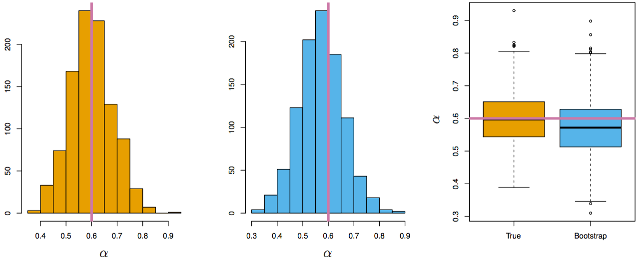


Figure 5.10 from *ISLR* (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

The Bootstrap in Practice

- Suppose we have an estimator $\hat{\mu} = \hat{\mu}(\mathcal{D}_n)$.
- To get error bars, we can compute the "bootstrap variance".
    - Draw B bootstrap samples.
    - Compute sample or empirical variance of $\hat{\mu}(D_n^1), \ldots, \hat{\mu}(D_n^B)$..
- Could report

$$\hat{\mu}(\mathcal{D}_n) \pm \sqrt{\text{Bootstrap Variance}}$$