2 Point estimation

$$X_1, X_2, X_3, \ldots, X_{n-1}, X_n \overset{\text{iid}}{\sim} f_\theta.$$

Definition 2.1. An **estimator** is a rule that takes in data and outputs a best guess at some quantity of interest. Formally, an estimator could be any transformation of the observations:

$$\hat{\theta}_n = \hat{\theta}(X_1, X_2, \ldots, X_n).$$

So, in go data, out pops a guess. The notation is deliberate. Our generic notation for the **estimand** (the true but unknown quantity we seek to estimate) will be θ, so $\hat{\theta}$ will be our generic notation for an estimator that targets that quantity. The subscript-n reminds us what sample size our estimate is based on, which will be an important consideration.

Example 2.1. Let us come up with an estimator for θ in the Bernoulli setting from Example 1.1. Here, θ is a probability representing the likelihood of an event occurring. Our dataset of binary variables $X_1, X_2, \ldots, X_n$ is essentially a running tally of how often we actually observed the event of interest (e.g. the coin coming up heads), and so a reasonable estimator might be the proportion of successes that we observed:

$$\hat{\theta}_n = \underset{\text{proportion}}{\text{sample}} = \frac{\#(X_i = 1)}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Example 2.2. We get to observe $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, b)$ and want to come up with an estimator for the right end point b. A reasonable choice is

$$\hat{b}_n = \max(X_1, \ldots, X_n).$$

Indeed, we know that $0 \leq X_1, \ldots, X_n \leq b$ since the range of $\text{Unif}(0, b)$ is $[0, b]$. So b has to be larger than each of $X_1, \ldots, X_n$. In particular, it must be larger than $\max(X_1, \ldots, X_n)$. So our best guess for the right endpoint is this maximum.

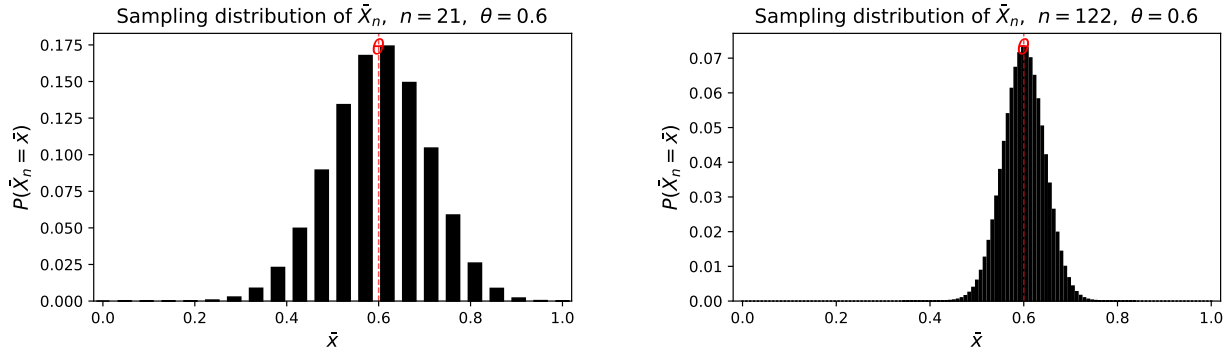
We will soon discuss a systematic way to come up with estimators, so we don't have to apply heuristic reasoning each time.

The main idea of *classical* statistics is that, since the data are random variables and the estimator is a function of the data, the estimator is a random variable as well. So it has its own distribution which we call the **sampling distribution** of the estimator. For the estimator $\bar{X}_n$ from Example 2.1, we can work out its sampling distribution explicitly. We know that $X_i \sim \text{Ber}(\theta)$ and $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Since $\sum_{i=1}^{n} X_i \sim \text{Binom}(n, \theta)$, $\hat{\theta}_n$ is just a "rescaled" binomial with:

$$\text{Range}(\hat{\theta}_n) = \left\{\tfrac{0}{n} = 0, \tfrac{1}{n}, \tfrac{2}{n}, \ldots, \tfrac{n-1}{n}, \tfrac{n}{n} = 1\right\}$$

$$P\left(\hat{\theta}_n = \tfrac{k}{n}\right) = \binom{n}{k}\theta^k(1-\theta)^{n-k}.$$

Figure 28 shows this sampling distribution for two values of n.



(a) $n = 21$ (b) $n = 122$

Figure 28: Sampling distribution of $\hat{\theta}_n = \bar{X}_n$ when $X_i \overset{\text{iid}}{\sim}$ Bernoulli(0.6). As the sample size n increases, the distribution concentrates around $\theta = 0.6$.

We see that, as we collect more data, the sampling distribution concentrates more closely about the true parameter θ. This is the desired behavior: the more data we have, the more confident we should be in our estimate of θ. We will now quantify this good behavior more precisely. The question is, what makes an estimator good?

Remark 2.1. For the Bernoulli example, we were able to work out the sampling distribution of the estimator $\bar{X}_n$ explicitly. This is not always possible. However, whenever our estimator is given by a sample average $\bar{X}_n$, we can use the CLT to at least approximately compute the sampling distribution: when n is large enough, we know $\bar{X}_n$ is approximately Gaussian.

2.1 Measuring the quality of an estimator

Two ways to estimate the quality of an estimator are via its *variance* and its *bias*.

- (**variance**) The variance $\text{var}(\hat{\theta}_n)$ tells us whether the sampling distribution is widely spread out or tightly concentrated. In layman's terms, is $\hat{\theta}_n$ stable or volatile?
- (**bias**) The bias is defined below.

> **Definition 2.2.** The **bias** of $\hat{\theta}_n$ is
>
> $$\text{bias}(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n] - \theta.$$
>
> If $\text{bias}(\hat{\theta}_n) = 0$, i.e. $\mathbb{E}[\hat{\theta}_n] = \theta$, we say the estimator is **unbiased**.

If $\hat{\theta}_n$ is unbiased, then its sampling distribution is centered on the true value θ. In layman's terms, an unbiased estimator is correct *on average*.

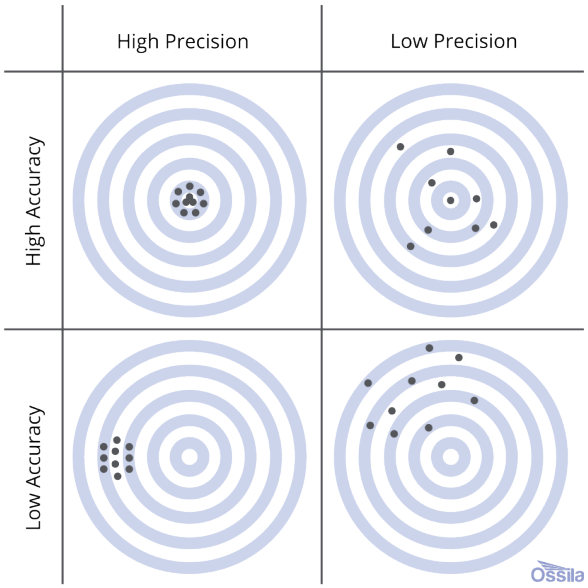


Figure 29: A low variance estimator is like high-precision dart throws at a target: they are clustered in one spot, though not necessarily the bullseye. A low bias estimator is like high-accuracy dart throws: they are centered at the bullseye, though they may be spread out rather than clustered.

Example 2.3. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Ber}(\theta)$. Consider three estimators: $\hat{\theta}_n = \frac{1}{2}$, $\hat{\theta}_n = X_1$, $\hat{\theta}_n = \bar{X}_n$. The first one throws out all the data, and just blindly guesses that $\theta = 1/2$. The second one only looks at X_1. The third one averages all the data.

$\hat{\theta}_n = \frac{1}{2}$:

$$\text{bias}\left(\tfrac{1}{2}\right) = \tfrac{1}{2} - \theta. \quad \text{Nails it when } \theta = \tfrac{1}{2}. \text{ Bad when } \theta \text{ is far from } \tfrac{1}{2}.$$
$$\text{var}\left(\tfrac{1}{2}\right) = 0. \quad \text{Good!}$$

$\hat{\theta}_n = X_1$:

$$\text{bias}(X_1) = \mathbb{E}[X_1] - \theta = \theta - \theta = 0. \quad \text{unbiased}$$
$$\text{var}(X_1) = \theta(1-\theta). \quad \text{When } \theta = 0 \text{ or } 1, \text{ this is good; worse when } \theta \text{ is in between.}$$

$\hat{\theta}_n = \bar{X}_n$:

$$\text{bias}(\bar{X}_n) = \mathbb{E}[\bar{X}_n] - \theta = \theta - \theta = 0. \quad \text{unbiased}$$
$$\text{var}(\bar{X}_n) = \text{var}(X_1)/n = \frac{\theta(1-\theta)}{n}. \quad \text{Good estimator for } n \text{ large, regardless of } \theta.$$

(Recall: if $\mathbb{E}X_1 = \mu$, $\text{var}(X_1) = \sigma^2$, then $\mathbb{E}\bar{X}_n = \mu$, $\text{var}(\bar{X}_n) = \sigma^2/n$.)

Since $\theta(1-\theta) \le \frac{1}{4}$ for all θ (with maximum at $\theta = \frac{1}{2}$), we have $\text{var}(\bar{X}_n) \le \frac{1}{4n}$ in the worst case.

Notice that each of the three estimators in the above example is good in at least one department: $1/2$ has zero variance, and X_1 has zero bias. But only $\bar{X}_n$ has both zero bias and (as $n \to \infty$) vanishingly small variance. Ideally we would have an error metric that combines both variance and bias to produce a single total score of how good an estimator is. The *mean-squared error* gives precisely this.

Definition 2.3. The **mean-squared error** (MSE) of an estimator is

$$\mathrm{MSE}(\hat{\theta}_n) = \mathbb{E}\left[(\hat{\theta}_n - \theta)^2\right].$$

Theorem 2.1 (Bias–variance decomposition)**.** $\mathrm{MSE}(\hat{\theta}_n) = \mathrm{bias}(\hat{\theta}_n)^2 + \mathrm{var}(\hat{\theta}_n)$.

Proof.

$$\begin{aligned}
\mathbb{E}\left[(\hat{\theta}_n - \theta)^2\right] &= \mathbb{E}\left[\left((\hat{\theta}_n - \mathbb{E}\hat{\theta}_n) + (\mathbb{E}\hat{\theta}_n - \theta)\right)^2\right] \\
&= \mathbb{E}\left[(\hat{\theta}_n - \mathbb{E}\hat{\theta}_n)^2 + 2(\hat{\theta}_n - \mathbb{E}\hat{\theta}_n)(\mathbb{E}\hat{\theta}_n - \theta) + (\mathbb{E}\hat{\theta}_n - \theta)^2\right] \\
&= \underbrace{\mathbb{E}\left[(\hat{\theta}_n - \mathbb{E}\hat{\theta}_n)^2\right]}_{\mathrm{var}(\hat{\theta}_n)} + 2(\mathbb{E}\hat{\theta}_n - \theta)\underbrace{\mathbb{E}\left[\hat{\theta}_n - \mathbb{E}\hat{\theta}_n\right]}_{=0} + \underbrace{(\mathbb{E}\hat{\theta}_n - \theta)^2}_{\mathrm{bias}(\hat{\theta}_n)^2} \\
&= \mathrm{var}(\hat{\theta}_n) + \mathrm{bias}(\hat{\theta}_n)^2.
\end{aligned}$$

□

Bias–variance tradeoff It is hard to make both bias and variance small; in practice we trade off between the two. Sometimes we are willing to tolerate a little bias for a reduction in variance that makes the overall MSE smaller. This opens up a whole kettle of worms involving topics like **shrinkage** and **regularization**. Fun stuff for another day!

Definition 2.4 (Consistency)**.** If $\mathrm{MSE}(\hat{\theta}_n) \to 0$ as $n \to \infty$, then we say that an estimator is **consistent**. In order to be consistent, an estimator's bias and variance must both go to zero as the sample size increases. If the estimator is already unbiased, then consistency is guaranteed if the variance goes to 0.

When it comes to point estimation, this is the ultimate sanity check. If we had *infinite* data, would our estimator recover the truth? If the answer is no, then what are we even doing? An estimator should *at least* be consistent. But if the answer is yes, this is not the slam dunk it may seem like. Consistency is an *asymptotic* property. It tells us what happens if we have infinite data. But we never actually have infinite data. We have a finite set of observations. If an estimator is consistent but the rate of convergence is very slow, then that somewhat curbs our enthusiasm for it.

Example 2.4. Back to the example: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Ber}(\theta)$ with $\hat{\theta}_n = \frac{1}{2}$, $\hat{\theta}_n = X_1$, $\hat{\theta}_n = \bar{X}_n$.

$$\text{MSE}\left(\tfrac{1}{2}\right) = \left(\tfrac{1}{2} - \theta\right)^2 + 0. \quad \text{Not going to 0 as } n \to \infty, \text{ so } \textbf{not} \text{ consistent.}$$

$$\text{MSE}(X_1) = 0^2 + \theta(1-\theta). \quad \text{Not going to 0 as } n \to \infty, \text{ so } \textbf{not} \text{ consistent.}$$

$$\text{MSE}(\bar{X}_n) = 0^2 + \frac{\theta(1-\theta)}{n} \to 0 \text{ as } n \to \infty. \quad \checkmark \text{ So it } \textbf{is} \text{ consistent.}$$