**476 Statistics, Spring 2022.**

**Lecturer: Sergey Nadtochiy.**

**Lecture 2. Point estimators: bias, error, mean square error, consistency, asymptotic normality. Properties of some popular estimators. (Sections 8.1–8.4)**

# 1 (Point) estimators and their quality

Point estimator is a statistic of the sample that is meant to approximate the unknown parameter. The parameter is usually denoted $\theta$, and the estimator is denoted $\hat{\theta}$, The point estimators are called this way because each estimator, on each realization of the sample, provides a single number (point) as an approximation of the unknown $\theta$. Note, however, that this approximating number (point) will be different on a different realization of the sample, hence $\hat{\theta}$ **is a random variable**! We will often drop the term "point" and simply say "estimator".

What are the **desirable properties of estimators**? Typically, we want $\hat{\theta}$ to converge to true $\theta$, as the sample size grows. For a finite sample, we also care about the distributional properties of $\hat{\theta}$. This is because we use the abstraction in which the estimation procedure itself can be repeated many times, and in any given experiment we only observe one of many possible outcomes (this is another way of saying that $\hat{\theta}$ is a r.v.).

## 1.1 Finite-sample properties

Draw a graph of the pdf of $\hat{\theta}$. Notice that we want this pdf to be concentrated around the true $\theta$.

**Def 1.** *The bias of $\hat{\theta}$ is*

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

*$\hat{\theta}$ is unbiased if $B(\hat{\theta}) = 0$.*

In general, we can define the error of a point estimator.

**Def 2.** *The error of $\hat{\theta}$ is*

$$|\hat{\theta} - \theta|$$

Note that the error is random. One way to ensure that the error is small is to control the expectation of its square.

**Def 3.** *Mean square error of $\hat{\theta}$ is*

$$MSE(\hat{\theta}) = \mathbb{E}|\hat{\theta} - \theta|^2 = V(\hat{\theta}) + (B(\hat{\theta}))^2.$$

We will typically restrict ourselves to unbiased estimators. Then, the only term in MSE is the variance. We denote

$$V(\hat{\theta}) =: \sigma^2_{\hat{\theta}}$$

However, variance is only one way to control a distr. In general, we would like to be able to say smth about prob.'s of the form

$$\mathbb{P}(|\hat{\theta} - \theta| > b)$$

Recall that, if $\hat{\theta}$ has a known distribution (as in the examples of previous lecture, for normal samples), we can compute the above probability explicitly. However, if the sample does not come from one of the special distributions (such a normal or exponential), or if we do not know this distribution at all, we cannot find the distribution of $\hat{\theta}$ exactly, and we resort to the asymptotic methods that work when the sample size is large.

## 1.2 Large-sample properties

**Def 4.** *A family of estimators $\{\hat{\theta}_n\}$, parameterized by the sample size $n$, is* **consistent** *if*

$$\lim_{n\to\infty} \hat{\theta}_n = \theta.$$

**Rem 1.** *The above convergence is understood* **in the sense of probability***: for any $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(|\hat{\theta}_n - \theta| \geq \varepsilon) = 0.$$

*If $\hat{\theta}_n$ converges with probability one (i.e., for almost every random outcome), it also converges in the sense of probability.*

Recall that the sample mean $\bar{Y}$ is consistent (as follows from LLN), and so is the sample variance $S^2$ (as shown in the previous lecture).

Next, we notice that in practice it is important to know not only that $\hat{\theta}_n$ gets closer to $\theta$ (with high probability) as $n$ gets larger, but also how fast $\hat{\theta}_n$ approaches $\theta$. In particular, as mentioned before, we would like to be able to say smth about prob.'s of the form

$$\mathbb{P}(|\hat{\theta}_n - \theta| > b)$$

Recall again that, for normal samples, we can compute the above prob. explicitly. It turns out that, even if the sample is not normal, $\hat{\theta}$ may be asymptotically normal, which will allow us to find the asymptotic behavior of the above probabilities as $n \to \infty$.

**Def 5.** *A family of estimators $\{\hat{\theta}_n\}$, parameterized by the sample size $n$, is asymptotically normal if*

$$\sqrt{n}(\hat{\theta}_n - \theta) \to N(0, \tilde{\sigma}^2),$$

*as $n \to \infty$, for some constant $\tilde{\sigma}^2 > 0$.*

**Rem 2.** *The above convergence in understood* **in the sense of distribution***: the cdf of $\sqrt{n}(\hat{\theta}_n - \theta)$ converges to the cdf of $N(0, \tilde{\sigma}^2)$ at each point. If $\sqrt{n}(\hat{\theta}_n - \theta)$ converges with probability one (i.e., for almost every random outcome) or in the sense of probability, then it also converges in the sense of distribution.*

**Rem 3.** *If $\{\hat{\theta}_n\}$ is asymptotically normal then it is also consistent. To see why, assume that it is not consistent. Then, for some $\varepsilon > 0$ and with some strictly positive probability, we have $|\hat{\theta}_n - \theta| \geq \varepsilon$, which implies that $|\sqrt{n}(\hat{\theta}_n - \theta)| \geq \sqrt{n}\varepsilon \to \infty$, which is a contradiction (otherwise, the limiting distribution would have a strictly positive mass at infinity, which is not allowed).*

If $\hat{\theta}$ is asymptotically normal, in the sense that

$$\sqrt{n}(\hat{\theta}_n - \theta) \to N(0, \tilde{\sigma}^2),$$

then

$$\frac{\sqrt{n}}{\tilde{\sigma}}(\hat{\theta}_n - \theta) \to N(0, 1),$$

Then, we have for large $n$:

$$\mathbb{P}(|\hat{\theta} - \theta| > b) = 1 - \mathbb{P}(\sqrt{n}|\hat{\theta} - \theta| \leq b\sqrt{n}) = 1 - \mathbb{P}\left(\frac{-b\sqrt{n}}{\tilde{\sigma}} \leq \frac{\sqrt{n}(\hat{\theta} - \theta)}{\tilde{\sigma}} \leq \frac{b\sqrt{n}}{\tilde{\sigma}}\right)$$

$$\approx 1 - \Phi\left(\frac{b\sqrt{n}}{\tilde{\sigma}}\right) + \Phi\left(\frac{-b\sqrt{n}}{\tilde{\sigma}}\right) = 2\Phi\left(\frac{-b\sqrt{n}}{\tilde{\sigma}}\right),$$

where $\Phi$ is the standard normal cdf. Thus, for large $n$,

$$\mathbb{P}(|\hat{\theta} - \theta| > b) \approx 2\Phi(\frac{-b\sqrt{n}}{\tilde{\sigma}}),$$

where the latter can be computed explicitly. Very often, we want the above probability (i.e., the probability of error to be above $b$) to be approximately $0.05$, which corresponds to choosing $b$ so that

$$\Phi(\frac{-b\sqrt{n}}{\tilde{\sigma}}) = 0.05/2 = 0.025.$$

The equation $\Phi(z) = 0.025$ has the approximate solution $z \approx -2$. Thus,

$$\frac{-b\sqrt{n}}{\tilde{\sigma}} = -2, \quad b = \frac{2\tilde{\sigma}}{\sqrt{n}}.$$

The above is often referred to as **"putting a 2-standard deviation bound on the estimation error":**

$$\mathbb{P}\left(|\hat{\theta} - \theta| > \frac{2\tilde{\sigma}}{\sqrt{n}}\right) \approx 0.05,$$

**where $\tilde{\sigma}$ is such that $\frac{\sqrt{n}}{\tilde{\sigma}}(\hat{\theta}_n - \theta)$ is approximately (or exactly) $N(0,1)$.**

# 2 Properties of some popular estimators

Consider an i.i.d. sample $Y_1, \ldots, Y_n$ from a distribution with finite mean $\mu$ and variance $\sigma^2$.

**Thm 1.** *The sample mean $\hat{\mu} = \bar{Y}$ is an unbiased estimator of $\mu$.*

*Proof:*
$\mathbb{E}\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}Y_i = \mu$. ∎

**Thm 2.** *$\hat{\mu} = \bar{Y}$ is asymptotically normal:*

$$\sqrt{n}(\hat{\mu} - \mu) \to N(0, \sigma^2),$$

*Proof:*
The statement follows directly from the central limit theorem (CLT).

**Thm 3.** *(CLT) Consider i.i.d. r.v.'s $Y_1, \ldots, Y_n$, with finite mean $\mu$ and variance $\sigma^2$. Then,*

$$\sqrt{n}(\bar{Y}_n - \mu) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i - \mu)$$

*converges weakly to $N(0, \sigma^2)$ as $n \to \infty$.*

∎

**Rem 4.** *For a normal sample, the 2-st. dev. error bound for $\hat{\mu}$ can be written exactly (not asymptotically), through a standard distr., if we know $\sigma$. In all other cases, we have to rely on asymptotic arguments.*

**Ex 1.** *We need to compare the durability of two brands of tires, measured in the number of miles until a tire is dead (its tread decreases to a critical level). Two samples (one for each brand) of equal sizes $n = n_1 = n_2 = 100$ are collected. (Note that we denote by lowercase letters the actual values of statistics and samples.)*

*The following values of sample means and sample variances were observed:*

$$\bar{y}_1 = 26,400, \quad s_1^2 = 1,440,000,$$

$$\bar{y}_2 = 25,100, \quad s_2^2 = 1,960,000.$$

**Q 1.** *Estimate the difference in means and put a 2-standard deviation bound on the error (asymptotically, in the large-sample regime).*

*Denote by $Y_1^1, \dots, Y_{100}^1$ the durabilities of the 100 randomly selected tires of type 1, and by $Y_1^2, \dots, Y_{100}^2$ the durabilities of the 100 randomly selected tires of type 2.*

*A natural estimator for the difference $\theta = \mu_1 - \mu_2$ is*

$$\hat{\theta} := \bar{Y}_1 - \bar{Y}_2.$$

*Indeed the above estimate is actually the sample mean of the sample $Y_1^1 - Y_1^2, \dots, Y_n^1 - Y_n^2$. Hence, it is unbiased and asymptotically normal, with*

$$\tilde{\sigma}^2 = \sigma_1^2 + \sigma_2^2$$

*Thus, the value of the proposed estimator for $\mu_1 - \mu_2$ is*

$$\bar{y}_1 - \bar{y}_2 = 1300.$$

*To put the 2-standard deviation bound on the error, we find $b$ s.t.*

$$0.05 = \mathbb{P}(|\hat{\theta} - \theta| > b) = \mathbb{P}\left( \frac{\sqrt{100}}{\sqrt{\sigma_1^2 + \sigma_2^2}} |\hat{\theta} - \theta| > b \right) \approx 2\Phi\left( \frac{-b\sqrt{100}}{\sqrt{\sigma_1^2 + \sigma_2^2}} \right),$$

*which gives*

$$b = \frac{2\sqrt{\sigma_1^2 + \sigma_2^2}}{10} \approx \frac{1}{5}\sqrt{1,440,000 + 1,960,000} = 368.8$$

*Note that in the above we replaces the true variances $\sigma_1^2$, $\sigma_2^2$ by their estimates $s_1^2$, $s_2^2$, and that we don't yet have a rigorous statement that justifies that. Nevertheless, such approximation is appropriate, and we will justify it later.*

**Rem 5.** *Note that we haven't discussed the asymptotic normality of the standard estimators for the difference $\mu_1 - \mu_2$ in the cases where the two samples have different sizes. We will come back to this question later.*

Next, consider the sample variance:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

**Thm 4.** *The sample variance $\hat{\sigma}^2 = S^2$ is an unbiased estimator of $\sigma^2$.*

*Proof:*

$$\mathbb{E}\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\sum_{i=1}^{n}\mathbb{E}Y_i^2-2\mathbb{E}\left(\bar{Y}\sum_{i=1}^{n}Y_i\right)+\mathbb{E}\sum_{i=1}^{n}\bar{Y}^2=\sum_{i=1}^{n}\mathbb{E}Y_i^2-n\mathbb{E}\bar{Y}^2$$

$$=n(V(Y_i)+(\mathbb{E}Y_i)^2)-n(V(\bar{Y})+(\mathbb{E}\bar{Y})^2)=n(\sigma^2+\mu^2)-n\sigma^2/n-n\mu^2=(n-1)\sigma^2,$$

where we recalled that $V(\bar{Y})=\sigma^2/n$ and that, for any r.v. $Z$, we have $V(Z)=\mathbb{E}Z^2-(\mathbb{E}Z)^2$. ∎

**Thm 5.** *If $\mathbb{E}Y_i^4<\infty$, then $\hat{\sigma}^2=S^2$ is an asymptotically normal estimator of $\sigma^2$:*

$$\sqrt{n}(\hat{\sigma}^2-\sigma^2)\to N\left(0,\mathbb{E}\left[(Y_i-\mu)^2-\sigma^2\right]^2\right).$$

Assume $Y_1,\ldots,Y_n$ are i.i.d. Bernoulli with success prob. $p$: i.e., $Y_i$ takes value 1 w. prob. $p$ and 0 w. prob. $1-p$. Notice that, in this case, $\mu=\mathbb{E}Y_i=p$ and $\sigma^2=V(Y_i)=p(1-p)^2+(1-p)p^2=p(1-p)$. Then, a natural estimator for $p$ is

$$\hat{p}=\bar{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$$

As we already know the properties of sample mean, we conclude that:

- $\hat{p}=\bar{Y}$ is unbiased
- and asymptotically normal:

$$\sqrt{n}(\hat{p}-p)\to N(0,p(1-p)).$$

**Ex 2.** *A sample of $n=1000$ voters is collected, $560$ of them are in favor of a given candidate.*

**Q 2.** *Estimate the true fraction of all voters who will vote for the candidate, and place a 2-st. dev. bound on the estimation error.*

*As discussed above we use sample mean as the estimator:*

$$\hat{p}=\bar{y}=0.56.$$

*To place a 2-st. dev. bound, we proceed as before:*

$$\mathbb{P}(|\hat{p}-p|>b)\approx 2\Phi(\frac{-b\sqrt{1000}}{\sqrt{p(1-p)}})\approx 0.05,$$

*for*

$$\frac{b\sqrt{1000}}{\sqrt{p(1-p)}}=2,$$

$$b=\frac{2\sqrt{p(1-p)}}{10\sqrt{10}}.$$

*But we do not know $p$. As before, we use replace $p$ with $\hat{p}$ in the above (this is justified further in the lectures):*

$$b\approx\frac{2\sqrt{0.56(1-0.56)}}{10\sqrt{10}}\approx 0.03.$$

Note that, in the above, we replaced the true (unknown) variance of Bernoulli sample $\{Y_i\}_{i=1}^n$ by $\bar{Y}(1-\bar{Y})$. Why did we do that? While we do not provide a completely rigorous justification of this, it is worth noting that, for a Bernoulli sample (and not for other distributions!), we have

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{n-1}\sum_{i=1}^{n} Y_i^2 - \frac{2}{n-1}\bar{Y}\sum_{i=1}^{n} Y_i + \frac{n}{n-1}\bar{Y}^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} Y_i^2 - \frac{2n}{n-1}\bar{Y}^2 + \frac{n}{n-1}\bar{Y}^2 = \frac{1}{n-1}\sum_{i=1}^{n} Y_i - \frac{n}{n-1}\bar{Y}^2$$

$$= \frac{n}{n-1}\bar{Y} - \frac{n}{n-1}\bar{Y}^2 = \frac{n}{n-1}\bar{Y}(1-\bar{Y}) \approx \bar{Y}(1-\bar{Y}).$$