NMAI059 – Probability and Statistics 1

Mykhaylo Tyomkyn

Lecture 10 - Introduction to Statistics. Estimators.

In *Probability theory* we dealt with measures and (distributions of) random variables that were *given*. In *Statistics*, on the other hand, we work with *data* that carries some 'hidden' randomness, and we try to *infer* that randomness and its parameters.

Example 1 *We have a delivery of* 10000 *avocados some of which are rotten inside. We take a sample of* 50 *and check them (checking an avocado destroys it, so we can only afford to sample a small number). We want to come up with some answer regarding the total number of rotten ones. There are several ways of doing so:*

- *An* estimator*: a single number, our 'best guess', as to how many are rotten. For instance, if* 2 *avocados out of our sample of* 50*, that is* 4%*, were bad, we may guess that* 4% *out of the* 10000*, i.e.,* 400 *in total are rotten.*[1]
- *A* confidence interval*: we want to give a range corresponding to a 'degree of certainty', e.g., we want to say "I am* 85% *certain that the number of rotten avocados lies in the interval* $[300, 500]$*." But what does that even mean, i.e., what probability space does* 85% *in that sentence refer to?*
- *A* hypothesis test*. We have agreed with the supplier that we reject the delivery if more than* 5% *of the avocados are rotten. So we choose a number* $c \in \{0, \dots, 50\}$ *and reject the delivery if more than* c *avocados out of the sampled* 50 *are bad. How should we pick* c*? Note that we can make two different kinds of error: we can wrongly reject a good delivery, or wrongly accept a bad one.*

We shall study all three of the above approaches (and more), but first we need to *formalize* the setup. Note first that in the above example we sampled *without* replacement. This may be an accurate description of reality but is computationally more difficult to handle than sampling *with* replacement. So, going forward, we will always sample with replacement. Now, let us introduce the *statistical model.*

Definition 1 (Statistical model) *We are given an event space* $(\Omega, \mathcal{F})$ *and an unknown/hidden (but fixed) probability measure* $\mathbb{P}$ *on* $\mathcal{F}$*. We take a* random sample *represented by iid random variables* $(X_1, \dots, X_n)$ *with* $X_i \sim \mathbb{P}$ *for all* i*. Our goal is to infer* $\mathbb{P}$ *or its parameters, such as mean and variance, from the sample.*

This approach is known as the *Classical statistics.* By contrast, *Bayesian statistics* (not part of this course) makes an a priori assumption about the measure, and adjusts it according to the observations. [2]

[1] This is a very reasonable guess, as we shall see soon.

[2] These two approaches in some sense reflect the two interpretations of Probability, discussed in Lecture 1.

We distinguish between

- Non-parametric models: $\mathbb{P}$ can by 'anything'. More precisely, we have $\mathbb{P} \in \mathcal{P}$, where $\mathcal{P}$ is a (very general) family of probability measures, e.g. $\mathcal{P} = L^1(\Omega, \mathcal{F})$.
- Parametric models: $\mathbb{P} \in \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$. That is, we know in advance what type of measure $\mathbb{P}$ is, but do not know one or multiple parameters.

Here are some examples of parametric models. We use $\mathbb{R}^+ = (0, \infty)$.

Example 2 *With $(\Omega, \mathcal{F})$ defined accordingly, let*

- $\mathbb{P} \in \{\mathbb{P}_\vartheta : \vartheta \in \Theta\} = \{Pois(\lambda) : \lambda \in \mathbb{R}^+\}$. *Here we have $\vartheta = \lambda$ and $\Theta = \mathbb{R}^+$.*
- $\mathbb{P} \in \{\mathbb{P}_\vartheta : \vartheta \in \Theta\} = \{Unif(a, b) : a, b \in \mathbb{R}\}$. *Here we have $\vartheta = (a, b)$ and $\Theta = \mathbb{R}^2$.*
- $\mathbb{P} \in \{\mathbb{P}_\vartheta : \vartheta \in \Theta\} = \{\mathcal{N}(\mu, \nu) : \mu \in \mathbb{R}, \nu \in \mathbb{R}^+\}$. *Here we have $\vartheta = (\mu, \nu)$ and $\Theta = \mathbb{R} \times \mathbb{R}^+$.*

Definition 2 (Statistic) *Any (real) function $T = T(X_1, \ldots, X_n)$ of the random sample is called a* statistic*.*

A statistic is basically a random variable on Ω^n. The choice of a different name is solely in order to emphasize the context.

Example 3 *We have a dataset of heights in a certain homogeneous population of humans. Prior experience tells us that it exhibits a normal distribution*[3] *$X \sim \mathcal{N}(\mu, \nu)$. To infer μ we may use a statistic such as $(X_1 + \cdots + X_n)/n$. To infer ν we use a different statistic (more on this later).*

Definition 3 (Estimator) *If a statistic T is used to estimate a parameter of the model (such as ϑ or some function $g(\vartheta)$), then such a statistic is called an* estimator *of that parameter.* [4] *In this case, given an outcome $(x_1, \ldots, x_n) = (X_1(\omega), \ldots, X_n(\omega))$ of our random sample, the value $T(x_1, \ldots, x_n)$ is called an* estimate *for said parameter.*

Example 4 *We want to infer the bias (probability of coming up 'heads') p of a coin. To this end, we may use the estimator $\hat{p} = k/n$, where k is the number of 'heads' among the n samples. Note that this is essentially the same statistic as in the previous example.*

How to tell if an estimator is 'good'? There are several criteria for this.

Definition 4 (Bias, unbiased, asymptotically unbiased) *The* bias *of an estimator T of $g(\vartheta)$, is the function $bias : \Theta \to \mathbb{R}$ given by*

$$bias_\vartheta(T) = \mathbb{E}_\vartheta(T) - g(\vartheta).$$

[3] We shall often use X here, implying $X \sim \mathbb{P}$.

[4] The estimator of $g(\vartheta)$ is sometimes denoted $\hat{g}(\vartheta)$ or $\bar{g}(\vartheta)$.

Here $\mathbb{E}_\vartheta(T)$ stands for the expectation of T under assumption $\mathbb{P} = \mathbb{P}_\vartheta$. The estimator is unbiased *if $bias_\vartheta(T) = 0$ for all $\vartheta \in \Theta$.* [5] *A family of estimators $(T_n = T_n(X_1, \dots, X_n))_{n=1}^\infty$ is unbiased if each T_n is unbiased. The family is* asymptotically unbiased *if*

$$\lim_{n\to\infty} bias_\vartheta(T_n) = 0$$

for all $\vartheta \in \Theta$.

Example 5 *The statistic $\bar{X}_n = (X_1 + \cdots + X_n)/n$ considered in the previous examples is an unbiased estimator for μ (also in the non-parametric setting). This follows from the linearity of expectation: for any measure $\mathbb{P}$ with a finite mean μ we have (using $X \sim \mathbb{P}$)*

$$\mathbb{E}(\bar{X}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) = E(X) = \mu.$$

Definition 5 (Consistent) *A family of estimators $(T_n = T_n(X_1, \dots, X_n))_{n=1}^\infty$ of a parameter of the model, say $g(\vartheta)$, is* consistent *if for all $\epsilon > 0$ and $\vartheta \in \Theta$:*

$$\lim_{n\to\infty} \mathbb{P}_\vartheta(|T_n - g(\vartheta)| > \epsilon) = 0.$$

Note this is basically convergence in probability. Thus, unsurprisingly, we have

Example 6 *The estimator $\bar{X}_n$ for μ from before is consistent, by WLLN.*

Example 7 (Consistent $\neq$ unbiased) *$\hat{\mu} = X_1$ is an unbiased estimator of μ, but (taking $T_n = X_1$ for all n) it is, in general, not consistent. On the other hand, the family $(X_1 + \cdots + X_n)/(n+1)$ of estimators of μ is consistent but not unbiased. It is asymptotically unbiased, though.*[6]

Now that we have seen a consistent and unbiased estimator for the mean, we may ask to find one for the variance. A naïve guess would be to take the *uncorrected sample variance*

$$\bar{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Surprisingly, there turns out to be a better candidate, namely the *Bessel correction* of the sample variance

$$\bar{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Theorem 1 *$\hat{S}_n^2$ is an unbiased estimator of the variance σ^2. Whereas $\bar{S}_n^2$ is merely asymptotically unbiased.* [7]

[5] The above concepts also make sense in a non-parametric setting. For instance, an estimator T of the mean μ is unbiased if $\mathbb{E}_\mathbb{P}(T) = \mu$ for all $\mathbb{P} \in \mathcal{P}$.

[6] Not a coincidence: under the additional assumption that $Var(T_n)$ is bounded, 'consistent' implies 'asymptotically unbiased'.

[7] It follows that $\hat{S}_n^2$ and $\bar{S}_n^2$ are consistent, by the previous footnote, provided all $\mathbb{P} \in \mathcal{P}$ are in L^4, i.e., each $\mathbb{E}_\mathbb{P}(X^4)$ is finite.

Proof

$$
\begin{aligned}
\bar{S}_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}((X_i - \mu) - (\bar{X}_n - \mu))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\big((X_i - \mu)^2 - 2(X_i - \mu)(\bar{X}_n - \mu) + (\bar{X}_n - \mu)^2\big) \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{2}{n}(\bar{X}_n - \mu)\sum_{i=1}^{n}(X_i - \mu) + \frac{1}{n}\sum_{i=1}^{n}(\bar{X}_n - \mu)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{2}{n}(\bar{X}_n - \mu)\cdot n(\bar{X}_n - \mu) + (\bar{X}_n - \mu)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\bar{X}_n - \mu)^2.
\end{aligned}
$$

Therefore, since $\bar{X}_n$ is an unbiased estimator for μ,

$$
\begin{aligned}
\mathbb{E}(\bar{S}_n^2) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i - \mu)^2 - \mathbb{E}((\bar{X}_n - \mu)^2) = \frac{n}{n}\mathbb{E}((X - \mu)^2) - \mathbb{E}((\bar{X}_n - \mathbb{E}(\bar{X}_n))^2) \\
&= Var(X) - Var(\bar{X}_n) = (1 - \frac{1}{n})\sigma^2,
\end{aligned}
$$

as, by independence,

$$
Var(\bar{X}_n) = Var\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) = \frac{n \cdot Var(X)}{n^2} = \frac{Var(X)}{n}.
$$

Which means

$$
\mathbb{E}(\hat{S}_n^2) = \mathbb{E}\left(\frac{n}{n-1}\bar{S}_n^2\right) = \frac{n}{n-1}\mathbb{E}(\bar{S}_n^2) = \sigma^2,
$$

so $\hat{S}_n^2$ is unbiased. As for $\bar{S}_n^2$, we have

$$
\lim_{n\to\infty}(\mathbb{E}(\bar{S}_n^2) - \sigma^2) = \lim_{n\to\infty}\frac{-\sigma^2}{n} = 0,
$$

so $\bar{S}_n^2$ is asymptotically unbiased. $\square$