M339G Predictive Analytics

In-Class Work #1

Some Mathematical Statistics.

1.1. The Statistical Set-Up.

Definition 1.1. *A* **random sample** *of size n from a distribution D is a random vector $(X_1, X_2, \ldots, X_n)$ where the $X_i, i = 1, \ldots, n$ are independent and identically distributed (i.i.d.) with distribution D.*

For the purposes of encapsulating a specific feature, estimation or hypothesis testing, we usually do not consider individual data. Instead, we apply suitable functions to the random sample thus distilling the information from the full data set for our purposes. We end up with the following definition:

Definition 1.2. *A* **statistic** *is a function of the random sample that does not depend on any unknown parameters.*

Given a model, one central task in statistics is to work out the distributions of specific statistics which are referred to as **sampling distributions**.

Depending on the use for which we construct the statistic, it can have different names. For example, if we use it to estimate a population parameter, we call it an **estimator**.

Example 1.3. *Let $(X_1, X_2, \ldots, X_n)$ be a random sample. Then the* **sample mean**

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is a natural estimator for the population mean.

The rationale behind using $\overline{X}$ to estimate the population mean is that we are *matching* the theoretical mean to the empirical mean. In general, this method of constructing estimators is called the **method of moments**.

Remark 1.4. *Another perspective on the sample mean is that it is the "center" of the data points in the sample. In fact, the sample mean minimizes the sum of squared deviations from itself, i.e.,*

$$\overline{X} = \arg\min_{a \in \mathbb{R}} \sum_{i=1}^{n} (X_i - a)^2.$$

Example 1.5. *Let $(X_1, X_2, \ldots, X_n)$ be a random sample. Then the* **sample variance**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2,$$

where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$*, is a natural estimator for the population variance.*

Question 1.6. *The above expression almost looks like the variance formula, but not quite. Why do we use* $n-1$ *instead of* n *in the denominator?*

Definition 1.7. *The* **bias** *of an estimator* $\hat{\theta}$ *for a parameter* θ *is defined as*

$$bias(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta.$$

An estimator is called **unbiased** *if its bias is zero, i.e., if* $\mathbb{E}[\hat{\theta}] = \theta$*.*

Importantly, for an estimator to be unbiased, the above equality must be true for all admissible values of the parameter θ.

Example 1.8. *Let* $(X_1, X_2, \ldots, X_n)$ *be a random sample from a distribution with mean* μ *and variance* σ^2*. Then the sample variance* S^2 *defined above is an unbiased estimator for the population variance* σ^2*. To see this, we compute*

$$\begin{aligned}
\mathbb{E}[S^2] &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}\left((X_i - \mu) - (\overline{X} - \mu)\right)^2\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2\right] \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\mathbb{E}[(X_i - \mu)^2] - n\mathbb{E}[(\overline{X} - \mu)^2]\right) \\
&= \frac{1}{n-1}\left(n\sigma^2 - n\frac{\sigma^2}{n}\right) \\
&= \sigma^2.
\end{aligned}$$

Remark 1.9. *The bias of the estimator speaks to the accuracy of the "recipe" we are using. If an estimator is unbiased, that means that we are not making systematic errors in our estimation procedure. Of course, while accuracy is of primary concern, we also care about precision.*

Definition 1.10. *The* **mean squared error** *(MSE) of an estimator* $\hat{\theta}$ *for a parameter* θ *is defined as*

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

Question 1.11. *Show that*

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + \Big(bias(\hat{\theta})\Big)^2.$$