Anna Janicka

Mathematical Statistics 2017/2018
Lecture 8
First draft, to be revised

1. Confidence Intervals

We devoted the last couple of lectures to the study of different methods of estimating the unknown value of a parameter governing the distribution underlying observed data, as well as analyzing the properties of these estimators. We have seen which methods may perform better than others – in terms of bias, MSE, efficiency etc.. We can never assure, however, that even for the best estimation technique and formula, the precise value of the estimator will be equal to the true value of the parameter. First of all, differences arise for purely "technical" reasons. For example, let us assume that we want to calculate the unknown probability of obtaining heads on a coin based on the observed frequency of outcomes (toss results), a natural and efficient way to estimate. Let us assume that we have an odd number of observations in our sample. In such a case, we will *never* obtain the true value of the parameter θ, if it were equal to $\frac{1}{2}$, just because we will never have exactly half of the observations equal to anything (since the number of observations is odd). Second, the discrepancies in the estimated values will also arise due to probabilistic reasons. Continuing the coin tossing example, if in reality the coin is symmetric ($\theta = \frac{1}{2}$), we can obtain a sequence of n tosses consisting of heads only (which happens with probability $\frac{1}{2^n}$). If we calculate an estimate of the unknown probability of heads based on this sample, using any reasonable technique will give us a result which is very far from the true probability of $\frac{1}{2}$. This is because the outcome we have based our considerations on is (highly) unlikely for the given value of the parameter. It is unlikely, but it is still possible!

Interval estimation, where – instead of calculating one precise value for the unknown parameter we provide an interval – allows to overcome both types of inaccuracy, signalled above: coming from sample size constraints and probabilistic variability.

1.1. The Philosophy of Confidence Intervals.

Formally, if $g(\theta)$ is a function of the unknown parameter θ that we wish to estimate (based on a sample $X_1, X_2, \ldots, X_n$), and $\bar{g} = \bar{g}(X_1, X_2, \ldots, X_n)$ and $\underline{g} = \underline{g}(X_1, X_2, \ldots, X_n)$ are statistics, then

Definition 1. *$(\underline{g}, \bar{g})$ is a* **confidence interval** *for $g(\theta)$ with* **confidence level** *$1-\alpha$, if we have that for any θ*

$$\mathbb{P}(g(\theta) \in (\underline{g}, \bar{g})) = \mathbb{P}(\underline{g} < g(\theta) < \bar{g}) \geqslant 1 - \alpha.$$

The value $1-\alpha$ may also be called a confidence coefficient; it describes the probability that the unknown value $g(\theta)$ falls into the random interval $(\underline{g}, \bar{g})$. This probability may be equated to the fraction of the empirical intervals, constructed based on random samples, that will include the true value of $g(\theta)$. Typically, we would like this probability to be large (i.e. equal to 0.9, 0.95, 0.99 etc.). Please note that we may talk about the probability that $g(\theta)$ falls into the confidence interval only if the upper and lower bounds of the interval are random, i.e. given in function al form (for example, as $\underline{g} = \bar{X} - \frac{1}{n}$ and $\bar{g} = \bar{X} + \frac{1}{n}$). Once we substitute specific values (numbers), and obtain a **realization** of a confidence interval, i.e. fixed upper and lower bounds (for example, $\underline{g} = 1$, $\bar{g} = 3$), we are no longer allowed to talk about the probability that the true value of $g(\theta)$ falls into the interval $(1, 3)$. This is because once we substitute the empirical results, we no longer have any randomness left; the true value of $g(\theta)$ either is inside the interval or it isn't. The situation is not random; it is fixed, but it is unknown to the researcher.[1]

[1]Let's say that the true value of θ is 2, and that we construct a confidence interval for θ, and we get a realization (obtained by substituting our data into the formulas) of $(1, 3)$. This interval includes 2. We can't say, however, that 2 falls into the interval $(1, 3)$ with such-and-such probability. 2 is always larger than 1 and smaller than 3, there is no randomness there. On the other hand, if the true value of the parameter θ was 5 and we constructed the same realization of the confidence interval, $(1, 3)$, 5 would not be inside the interval.

Obviously, the exact form of a confidence interval constructed for $g(\theta)$ must depend on the distribution governing the data. Just like in the case of point estimation, in the case of interval estimation we may also think of different methods of obtaining formulas for the upper and lower bounds of the interval. The method which is used most frequently is the so-called **pivotal method**. This technique is based on the observation that it is easiest to find a general estimation rule which works for different values of the parameter θ, if we can find a statistic whose value depends on θ, but whose distribution does not.

We have already seen examples of such pivotal quantities. If $X_1, X_2, \ldots, X_n$ are a sample from a normal distribution with mean μ and variance σ^2, then we know that $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, but $U = \frac{\bar{X}-\mu}{\frac{\sigma^2}{n}} \sim N(0,1)$. U is a pivotal quantity – it depends both on the data and on the value of the unknown parameter(s), while it's distribution does not. Note that in this case if we are going to estimate the value of μ based on U, we would need to assume additionally that σ is known, since we can only have one unknown parameter at a time.

Pivotal quantities are convenient, since they may easily be translated into confidence intervals. If U is a pivotal quantity for parameter θ, then the probability $\mathbb{P}(a \leqslant U \leqslant b)$ does not depend on θ (and can be calculated without knowing θ). Therefore, we can rephrase the confidence interval question and say that when looking for the confidence interval for θ, we will look for values a and b such that

$$\mathbb{P}(a \leqslant U \leqslant b) = 1 - \alpha.$$

Usually, for simplicity, we will be looking for "symmetric" intervals, such that

$$\mathbb{P}(a > U) = \frac{\alpha}{2} = \mathbb{P}(b < U),$$

which means that we will be looking for upper and lower bounds such that the probabilities that the parameter will not fall into the interval because it is too large and too small to fit inside are equal to each other.

The exact formulation of the confidence interval will depend on what we know about the underlying distribution. In the following section, we will provide examples of confidence interval construction for the most common distributions.

1.2. **Confidence Intervals for the Normal Model.**

1.3. **Approximate confidence intervals.**

Again, there would be no randomness there, since 5 is always larger than both 1 and 3. The randomness in confidence interval considerations may appear only when we look at general formulas – statistics, not when we have already substituted data into the formulas.