Anna Janicka

Mathematical Statistics 2021/2022
Lecture 7

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*

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

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 $(\underline{g}, \bar{g})$.[1]. 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 functional 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, data), 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.[2]

[1] The value α, therefore, describes the probability of the unknown value of the parameter falling outside the confidence interval

[2] 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

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 formula (random variable) 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 $U = \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$. U is, therefore, a pivotal quantity – it depends both on the data and on the value of the unknown parameter(s), while its distribution does not. Note that in this case if we are going to estimate the value of μ based on U having a standard normal distribution, 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 $\mathbb{P}(a \leq U \leq b)$ 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

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

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.** We will start our considerations with the construction of confidence intervals for data drawn from a normal distribution, $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$.

1.2.1. *Confidence Interval for the mean μ, when the variance is known.* We are looking for an interval (μ_L, μ_U), such that $P(\mu \in (\mu_L, \mu_U)) = 1 - \alpha$.

As signalled above, in a normal model we have a handy pivotal quantity $U = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$. Let's say that we want to construct a confidence interval for a confidence level $1-\alpha$ which is "symmetric", in the sense that the probability that the true value of the parameter μ is below the lower bound of the interval μ_L, is equal to the probability that the true value of the parameter μ is above the upper bound of the interval μ_U, and amounts to $\frac{\alpha}{2}$:

$$P_\mu(\mu < \mu_L) = P_\mu(\mu > \mu_U) = \frac{\alpha}{2}.$$

This construction will be equivalent to finding some values u_L and u_U, lower and upper bounds for the pivotal quantity, such that

$$P_\mu(U < u_L) = P_\mu(U > u_U) = \frac{\alpha}{2}.$$

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. 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.

Since the standard normal distribution is symmetric around 0, this means that the absolute values of the upper and lower bound must be equal, and $u_U = -u_L = u$. Therefore, we will be looking for a value u such that $P_\mu(u_L \leq U \leq u_U) = P_\mu(|U| \leq u) = 1 - \alpha$.

Since we know that the distribution of U is standard normal, we have that

$$P_\mu(|U| \leq u) = \Phi(u) - \Phi(-u) = \Phi(u) - (1 - \Phi(u)) = 2\Phi(u) - 1 = 1 - \alpha,$$

which means that

$$\Phi(u) = \frac{2 - \alpha}{2} = 1 - \frac{\alpha}{2},$$

which in turn means that u is the quantile of rank $1 - \frac{\alpha}{2}$ of the standard normal distribution. We shall denote such a quantile by $u_{1-\frac{\alpha}{2}}$.

Knowing that

$$1 - \alpha = P_\mu(-u_{1-\frac{\alpha}{2}} \leq U \leq u_{1-\frac{\alpha}{2}}) = P_\mu\left(-u_{1-\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq u_{1-\frac{\alpha}{2}}\right),$$

we can transform the formulae inside the probabiliy in order to get a range of values for μ:

$$P_\mu(-u_{1-\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq u_{1-\frac{\alpha}{2}}) = P_\mu(-u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) =$$

$$P_\mu(\bar{X} - u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}),$$

which leads us to the conclusion that if $\mu_L = \bar{X} - u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$, and $\mu_U = \bar{X} + u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$, then $P_\mu(\mu \in (\mu_L, \mu_U)) = 1 - \alpha$ and (μ_L, μ_U) is the requested confidence interval.

Please note that the confidence interval (μ_L, μ_U) is symmetric around the point estimate $\bar{X}$ of μ. Due to the symmetry of the normal distribution, this confidence interval is also the narrowest possible confidence interval for the requested confidence level. To see this, first note that there are many possible intervals satisfying the property that $P_\mu(\mu \in (m_L, m_U)) = 1 - \alpha$, i.e. covering the requested, fixed fraction $1 - \alpha$ of possible outcomes. One of such intervals is $(-\infty, \bar{X} + u_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}})$. Another one is $(\bar{X} - u_\alpha \cdot \frac{\sigma}{\sqrt{n}}, \infty)$. Each of these two confidence intervals have the unwelcome property that they are very wide (infinitely wide, in fact). This is because in each of these two cases we only "discard" (i.e., do not pick as possible values of our estimator) values on one side of the distribution. Apart from these two infinite intervals, there exist many confidence intervals of finite length, where we include $1 - \alpha$ values and "discard" both some values which are too small and some values which are too large. However, the shape of the density of the standard normal distribution is such that the highest density is at the center, and the further we are from the center, the smaller the value of the density. Therefore, if we want to obtain an interval which covers a fixed fraction of possible outcomes and is the narrowest possible, we must include the most common observations (i.e., all of those around the center) and discard the least frequent observations (i.e., all of those far from the center). Since the distribution is symmetric, we need to discard the same amount of "too low" and "too high" values.

The length of the confidence interval we constructed is equal to $2d$, where $d = u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ is the so-called error of estimation, or estimation precision. Note that for a given value of σ and confidence level $1 - \alpha$, if we want to obtain a given level of precision d, we need to have a sample size at least equal to

$$n \geq \frac{\sigma^2 u^2_{1-\frac{\alpha}{2}}}{d^2}.$$

1.2.2. *Confidence Interval for the mean μ, when the variance is unknown.* We are looking for an interval (μ_L, μ_U), such that $P(\mu \in (\mu_L, \mu_U)) = 1 - \alpha$. This time, however, we will not be able to use the pivotal quantity U (as in the previous section) for our construction, due to the fact that the parameter σ, which is unknown, appears in the formula for U. It is possible to approximate the true value of σ^2 with a value calculated on the base of the data – for example, using the unbiased estimator of the variance, S^2. If we substitute this value into

the formula for U instead of σ^2, however, the variable U no longer has a standard normal distribution. The distribution it has, however, has also been described and studied, and goes by the name of the t-Student distribution (with $n-1$ degrees of freedom, where n is the size of the sample used to calculate the variance).

The t-Student distribution is also symmetric around the mean (which is also equal to 0), and therefore a reasoning analogous to the reasoning conducted in the subsection above leads us to a very similar construction of the confidence interval for μ, the only differences being that the quantiles of the standard normal distribution are substituted by quantiles of the t-Student distribution, and the value of σ^2 is substituted by the value of its estimator S^2. In other words, we have that the confidence interval for μ for a confidence level $1-\alpha$ is equal to

$$\left(\bar{X} - t_{1-\frac{\alpha}{2}}(n-1) \cdot \frac{\sqrt{S^2}}{\sqrt{n}}, \bar{X} + t_{1-\frac{\alpha}{2}}(n-1) \cdot \frac{\sqrt{S^2}}{\sqrt{n}}\right),$$

where $t_{1-\frac{\alpha}{2}}(n-1)$ is a quantile of rank $1-\frac{\alpha}{2}$ of the t-Student distribution with $n-1$ degrees of freedom.

This time, the length of the confidence interval we constructed is equal to $2d$, where $d = t_{1-\frac{\alpha}{2}}(n-1) \cdot \frac{\sqrt{S^2}}{\sqrt{n}}$ is the error of estimation. Note that this time, for a given confidence level $1-\alpha$, the question of finding a sample size which assures a given level of precision d is not that simple, since when the number of observations increases, the estimator of the variance need not decrease (new observations may change it in both directions), and simultaneously, the number of degrees of freedom of the t-Student distribution, and hence the value of the quantile of this distribution, also changes. Finding a suitable sample size is nevertheless possible, and the procedure goes by the name of Stein's two-stage procedure.

The first stage of the procedure requires collecting a sample size of an arbitrarily chosen size n_0. For this sample, the mean $\bar{X}$ and the variance S_0^2 are calculated, and the resulting precision of the estimate is assessed: $d_0 = t_{1-\frac{\alpha}{2}}(n_0-1) \cdot \frac{\sqrt{S_0^2}}{\sqrt{n_0}}$. If this value does not exceed the required precision threshold d, no further actions need to be undertaken and the standard version of the confidence interval is good for this sample. If, however, the length of the interval is larger than required, $d_0 > d$, then in a second stage, an additional sample of n_a observations should be collected, such that

$$n_a + n_0 \geq \frac{S_0^2 \cdot t_{1-\frac{\alpha}{2}}^2(n_0-1)}{d^2}$$

In this case, the confidence interval of the required length is:

$$\left(\bar{X}^* - t_{1-\frac{\alpha}{2}}(n_0-1) \cdot \frac{\sqrt{S_0^2}}{\sqrt{n_a+n_0}}, \bar{X}^* + t_{1-\frac{\alpha}{2}}(n_0-1) \cdot \frac{\sqrt{S_0^2}}{\sqrt{n_a+n_0}}\right),$$

where the mean $\bar{X}^*$ is calculated based on the whole sample size (initial n_0 and additional n_a observations), while the variance S_0^2 is estimated based on the initial sample of n_0 observations (and the corresponding t-Student statistic has n_0-1 degrees of freedom).

What is worth noting is that if the variance of the distribution is unknown, even if the point estimate for the variance happens to be exactly equal to the true variance of the distribution, the length of a confidence interval for a confidence level $1-\alpha$ in this model will be greater than a similar confidence interval constructed assuming that the variance is known (the values of the quantiles of the t-Student distribution have larger absolute values than the corresponding values of the standard normal distribution). This is because not knowing the variance introduces additional uncertainty into the model, which translates to larger variability of possible outcomes (and hence a larger range for similarly probable values). As n tends to infinity, however, the t-Student distribution with n degrees of freedom tends to the standard normal distribution. Therefore, for large sample sizes, the quantiles of the t-Student distribution will be practically the same as the quantiles of the standard normal distribution, meaning that the length of the interval constructed based on a variance calculated from the sample will

be practically equal to the confidence interval calculated using the exact (true) value of the variance.

1.2.3. *Confidence Interval for the variance σ^2.* If our aim is to construct a confidence interval for the variance, we will need to look for a different pivotal quantity (one that does not explicitly use the value of μ). It can be shown that if S^2 is the unbiased estimator of the variance in a normal model, then

$$U = \frac{(n-1)S^2}{\sigma^2}$$

is a random variable dependent on σ, but which has a distribution which does not depend on σ, namely – the so-called chi-squared distribution with $n-1$ degrees of freedom $\chi^2(n-1)$ (again, n is the size of the sample used to calculate the variance).

The chi-squared distribution is concentrated on the positive real numbers only, and therefore is not symmetric around 0 (nor any other point). This affects the simplicity of the confidence interval construction, albeit not the general philosophy: we will want to discard as much "too low" possible values as "too high" possible values, i.e. have

$$\mathbb{P}(u_L > U) = \frac{\alpha}{2} = \mathbb{P}(U > u_H).$$

This time, however, the appropriate quantiles of the distribution will not be symmetric around 0, and will therefore be provided separately. The confidence interval for the variance, for a confidence level of $1-\alpha$, is equal to

$$\left(\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}, \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} \right),$$

where $\chi^2_{\frac{\alpha}{2}}(n-1)$ and $\chi^2_{1-\frac{\alpha}{2}}(n-1)$ are quantiles of rank $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$, respectively, of the chi-squared distribution with $n-1$ degrees of freedom.

1.3. **Approximate confidence intervals.** In many cases, data do not follow a normal distribution. If the sample size is large, however, we have – from the CLT – that the distribution of the sample mean $\bar{X}$ is approximately normal, and thus so is its normalization into a pivotal quantity U (even using an estimator of the variance instead of the true variance). This leads us to the conclusion that in such cases, a confidence interval for the mean constructed as above, using the quantiles of the standard normal distribution instead of the quantiles of the true distribution, will be very close to the "true" confidence interval. This is why in many cases, if the sample sizes are large, approximate confidence intervals for the means are used. Two most commonly used cases are:

$$\left(\bar{X} - u_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{S^2}}{\sqrt{n}}, \bar{X} + u_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{S^2}}{\sqrt{n}} \right),$$

as a confidence interval of the mean of a distribution, if the sample size is large, and

$$\left(\hat{p} - u_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + u_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \right),$$

as a confidence interval for the probability of success for a Bernoulli variable, where $\hat{p}$ is the empirical fraction of successes in the sample. Note that for a single Bernoulli trial, the probability of success is in fact the average of the distribution, so $\hat{p}$ is an estimator of the mean of this distribution and $\sqrt{\hat{p}(1-\hat{p})}$ is an estimator of the variance.