

## Lecture 26 : Random Intervals and Confidence Intervals

## 1. The definition of a random interval

Let  $X_1$  and  $X_2$  be random variables defined on the same sample space  $S$  such that  $X_1(s) < X_2(s)$  for all  $s \in S$ . Then  $I = (X_1, X_2)$  is called an (open) random interval. For each  $s \in S$  we obtain an ordinary interval  $I(s) = (X_1(s), X_2(s))$ . Thus we may think of a random interval as an *interval-valued* random variable defined on  $S$ . The point of this lecture is that *confidence intervals are random intervals*.

### Example

Suppose  $X$  is a random variable defined on  $S$  and  $a$  is a positive number then  $I = (X - a, X + a)$  is the random interval with random center  $X$  and (deterministic) width  $2a$ . More generally the random interval  $I = (X - Y, X + Y)$  has random center  $X$  and random width  $2Y$ .

## 2. Probabilities connected with random intervals

Now consider a random interval  $I = (X_1, X_2)$  and a fixed number  $a$ . We want to compute the probability that the random interval  $I$  will contain (or cover) the fixed number  $a$ . But this is just the probability that  $a$  will be between  $X_1$  and  $X_2$  hence we have

$$P(a \in (X_1, X_2)) = P(X_1 < a, a < X_2) = P(X_1 < a, X_2 > a). \quad (1)$$

The probability on the right is the probability that the random variable  $X_1$  will be less than the number  $a$  and the random variable  $X_2$  will be greater than the number  $a$ . This is just a random variable computation of the type we have done many times in the course already. Technically we should think of the formula (1) as the *definition* of the probability that  $a$  will be inside  $I$  but this is a technical point - it is the only reasonable definition.

### Warning

The probability  $P(X_1 < a, X_2 > a)$  in equation (1) is almost never equal to the product probability  $P(X_1 < a) \cdot P(X_2 > a)$  because  $X_1$  and  $X_2$  are almost never independent. For example in the above problem  $X_1 = Z - 1$  and  $X_2 = Z + 1$  so  $X_2 = X_1 + 2$  so  $X_1$  and  $X_2$  are perfectly correlated and in particular not independent. We will conclude with an example of how to compute such probabilities.

## Problem

Suppose that  $Z$  has standard normal distribution. Compute  $P(0 \in (Z - 1, Z + 1))$ .

## Solution

According to the equation (1) we have

$$P(0 \in (Z - 1, Z + 1)) = P(Z - 1 < 0, 0 < Z + 1).$$

But

$$P(Z - 1 < 0, 0 < Z + 1) = P(Z < 1, -1 < Z) = P(-1 < Z < 1) = P(-1 \leq Z \leq 1).$$

By the “handy formula” we have

$$P(-1 \leq Z \leq 1) = 2\Phi(1) - 1 = 2(.8413) - 1 = .6826$$

### 3. In which we go completely random

In the first part of the course we were given a random variable  $X$  and we computed probabilities like  $P(a \leq X \leq b)$ . But we have an equality of events

$$(a \leq X \leq b) = (X \in (a, b))$$

so

$$P(a \leq X \leq b) = P(X \in (a, b)).$$

In other words we were computing the probability that a *random variable* was in an *ordinary interval*. We have just learned how to compute the probability that a fixed *real number* is in a *random interval* e.g.  $P(0 \in (Z - 1, Z + 1))$ . It remains to “go completely random” and learn how to compute the probability that a *random variable* is in a *random interval*. Actually we can already do this. Let’s do an example.

## Problem

Suppose  $Z$  has standard normal distribution. Compute  $P(2Z \in (Z - 1, Z + 1))$ .

## Solution

We have

$$\begin{aligned} P(2Z \in (Z - 1, Z + 1)) &= P(Z - 1 \leq 2Z, 2Z \leq Z + 1) = P(-1 \leq Z, Z \leq 1) \\ &= P(-1 \leq Z \leq 1) = 2\Phi(1) - 1 = .6826. \end{aligned}$$

## Remark

In the above we had to do a little manipulation of inequalities, namely  $Z - 1 \leq 2Z \Leftrightarrow -1 \leq Z$  (subtract  $Z$  from each side or “bring the  $Z$  from the left-hand side to the right-hand side”) and  $2Z \leq Z + 1 \Leftrightarrow Z \leq 1$  (again subtract  $Z$  from each side or bring the  $Z$  from the right-hand side to the left-hand side”).

#### 4. The definition of a confidence (random) interval

Suppose now that  $X_1, X_2, \dots, X_n$  is a random sample from a population whose probability mass function (or probability density function) depends on an unknown parameter  $\theta$ . Let  $\alpha$  be a real number between 0 and 1. Then a  $100(1 - \alpha)\%$  confidence interval for the unknown parameter  $\alpha$  is a *random interval*  $I = (W_1, W_2)$  where  $W_1 = h(X_1, X_2, \dots, X_n)$  and  $W_2 = g(X_1, X_2, \dots, X_n)$  are *statistics* such that

$$P(\theta \in (W_1, W_2)) = 1 - \alpha. \quad (2)$$

If we hadn't given the definition in Equation (1) we wouldn't have been able to make the correct definition in Equation (2). If we have an actual sample  $x_1, x_2, \dots, x_n$  then we plug  $x_1, x_2, \dots, x_n$  into the functions  $h$  and  $g$  to get numbers  $w_1$  and  $w_2$  and an ordinary interval  $(w_1, w_2)$ . This ordinary interval is the *observed value* of the confidence interval  $I = (W_1, W_2)$  on the sample  $x_1, x_2, \dots, x_n$ . This actual interval is also called a confidence interval for  $\theta$ . It is important to keep the difference between the confidence random interval and its observed value on a sample firmly in mind.

The rest of the course will be concerned with finding formulas for confidence intervals in various situations - e.g. a 90% confidence interval for the mean in a normal distribution. In each case we will verify that the equation (2) is satisfied. It is imperative that you all learn how to do these verifications - these will be “good citizen” problems.