Lecture 24

24.1 Goodness-of-fit test.

Suppose that we observe an i.i.d. sample $X_1, \ldots, X_n$ of random variables that can take a finite number of values $B_1, \ldots, B_r$ with some unknown to us probabilities $p_1, \ldots, p_r$. Suppose that we have a theory (or a guess) that these probabilities are equal to some particular $p_1^\circ, \ldots, p_r^\circ$ and we want to test it. This means that we want to test the hypotheses

$$\begin{cases} H_1: & p_i = p_i^\circ \text{ for all } i = 1, \ldots, r, \\ H_2: & \text{otherwise, i.e. for some } i, p_i \neq p_i^\circ. \end{cases}$$

If the first hypothesis is true than the main result from previous lecture tells us that we have the following convergence in distribution:

$$T = \sum_{i=1}^{r} \frac{(\nu_i - np_i^\circ)^2}{np_i^\circ} \to \chi^2_{r-1}$$

where $\nu_i = \#\{X_j : X_j = B_i\}$. On the other hand, if H_2 holds then for some index i, $p_i \neq p_i^\circ$ and the statistics T will behave very differently. If p_i is the true probability $\mathbb{P}(X_1 = B_i)$ then by CLT (see previous lecture)

$$\frac{\nu_i - np_i}{\sqrt{np_i}} \to N(0, 1 - p_i).$$

If we write

$$\frac{\nu_i - np_i^\circ}{\sqrt{np_i^\circ}} = \frac{\nu_i - np_i + n(p_i - p_i^\circ)}{\sqrt{np_i^\circ}} = \frac{\nu_i - np_i}{\sqrt{np_i}} + \sqrt{n}\frac{p_i - p_i^\circ}{\sqrt{p_i^\circ}}$$

then the first term converges to $N(0, 1 - p_i)$ but the second term converges to plus or minus ∞ since $p_i \neq p_i^\circ$. Therefore,

$$\frac{(\nu_i - np_i^\circ)^2}{np_i^\circ} \to +\infty$$

which, obviously, implies that $T \to +\infty$. Therefore, as sample size n increases the distribution of T under hypothesis H_1 will approach χ^2_{r-1} distribution and under hypothesis H_2 it will shift to $+\infty$, as shown in figure 24.1.

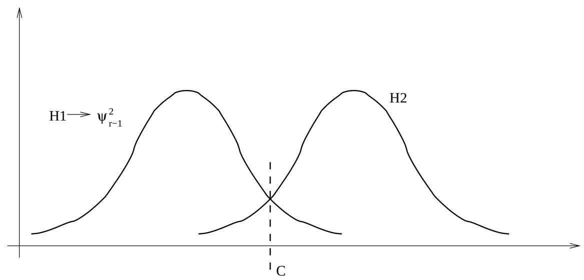

Figure 24.1: Distribution of T under H_1 and H_2.

Therefore, the following test looks very natural

$$\delta = \begin{cases} H_1: & T \le c \\ H_2: & T > c, \end{cases}$$

i.e. we suspect that the first hypothesis H_1 fails if T becomes unusually large. We can decide what is "unusually large" or how to choose the threshold c by fixing the error of type 1 to be equal to the level of significance α :

$$\alpha = \mathbb{P}_1(\delta \neq H_1) = \mathbb{P}_1(T > c) \approx \chi^2_{r-1}(c, \infty)$$

since under the first hypothesis the distribution of T can be approximated by χ^2_{r-1} distribution. Therefore, we find c from the table of χ^2_{r-1} distribution such that $\alpha = \chi^2_{r-1}(c, \infty)$. This test is called the χ^2 goodness-of-fit test.

□

Example. Suppose that we have a sample of 189 observations that can take three values A, B and C with some unknown probabilities p_1, p_2 and p_3 and the counts are given by

A	B	C	$Total$
58	64	67	189

We want to test the hypothesis H_1 that this distribution is uniform, i.e. $p_1 = p_2 = p_3 = 1/3$. Suppose that level of significance is chosen to be $\alpha = 0.05$. Then the threshold c in the χ^2 test

$$\delta = \begin{cases} H_1: & T \le c \\ H_2: & T > c \end{cases}$$

can be found from the condition that

$$\chi^2_{3-1=2}(c, \infty) = 0.05$$

and from the table of χ^2_2 distribution with two degrees of freedom we find that $c = 5.9$. In our case

$$T = \frac{(58 - 189/3)^2}{189/3} + \frac{(64 - 189/3)^2}{189/3} + \frac{(67 - 189/3)^2}{189/3} = 0.666 < 5.9$$

which means that we accept H_1 at the level of significance 0.05.

24.2 Goodness-of-fit for continuous distribution.

A similar approach can be used to test a hypothesis that the distribution of the data is equal to some particular distribution, in the case when observations do not necessarily take a finite number of fixed values as was the case in the last section. Let $X_1, \ldots, X_n$ be the sample from unknown distribution $\mathbb{P}$ and consider the following hypotheses:

$$\begin{cases} H_1 : & \mathbb{P} = \mathbb{P}_0 \\ H_2 : & \mathbb{P} \neq \mathbb{P}_0 \end{cases}$$

for some particular $\mathbb{P}_0$. To use the result from previous lecture we will discretize the set of possible values of Xs by splitting it into a finite number of intervals $I_1, \ldots, I_r$ as shown in figure 24.2. If the first hypothesis H_1 holds then the probability that X comes from the jth interval is equal to

$$\mathbb{P}(X \in I_j) = \mathbb{P}_0(X \in I_j) = p_j^\circ.$$

and instead of testing H_1 vs. H_2 we will consider the following weaker hypotheses

$$\begin{cases} H_1' : & \mathbb{P}(X \in I_j) = p_j^\circ \text{ for all } j \leq r \\ H_2' : & \text{otherwise} \end{cases}$$

Asking whether H_1' holds is, of course, a weaker question that asking if H_1 holds, because H_1 implies H_1' but not the other way around. There are many distributions different from $\mathbb{P}$ that have the same probabilities of the intervals $I_1, \ldots, I_r$ as $\mathbb{P}$. Later on in the course we will look at other way to test the hypothesis H_1 in a more consistent way (Kolmogorov-Smirnov test) but for now we will use the χ^2 convergence result from previous lecture and test the derivative hypothesis H_1'. Of course, we are back to the case of categorical data from previous section and we can simply use the χ^2 goodness-of-fit test above.

The rule of thumb about how to split into subintervals $I_1, \ldots, I_r$ is to have the expected count in each subinterval

$$np_i^\circ = n\mathbb{P}_0(X \in I_i) \geq 5$$

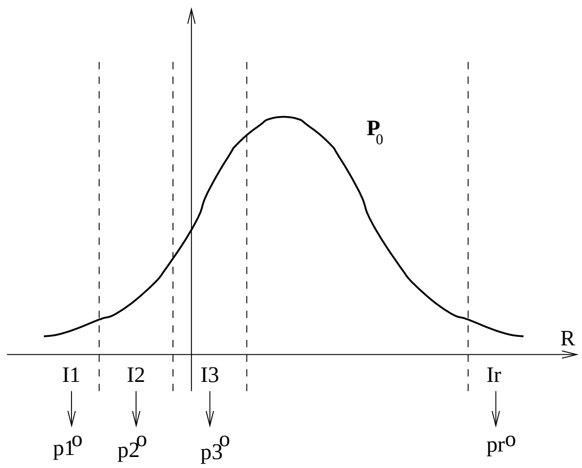

Figure 24.2: Discretizing continuous distribution.

at least 5. For example, we can split into intervals of equal probabilities $p_i^\circ = 1/r$ and choose their number r so that

$$np_i^\circ = \frac{n}{r} \geq 5.$$

Example. (textbook, p. 539) We want to test the following hypotheses:

$$\begin{cases} H_1 : & \mathbb{P} = N(3.912, 0.25) \\ H_2 : & \text{otherwise} \end{cases}$$

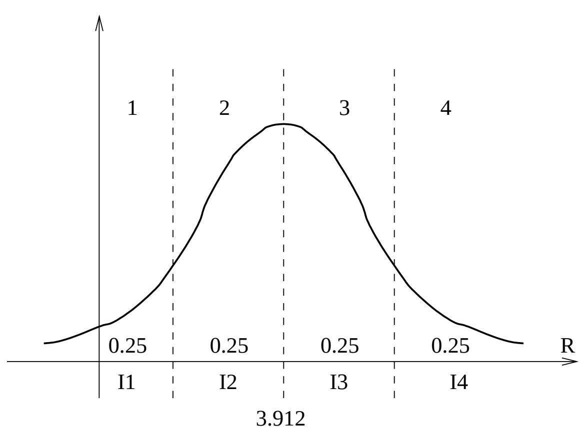

Figure 24.3: Total of 4 Sub-intervals.

We are given $n = 23$ observations and using the rule of thumb we will split into r equal probability intervals so that

$$\frac{n}{r} = \frac{23}{r} \geq 5 \Rightarrow r = 4.$$

Therefore, we split into 4 intervals of probability 0.25 each. It is easy to find the endpoints of these intervals for the distribution $N(3.912, 0.25)$ which we will skip and simply say that the counts of the observations in these intervals are...