ELG 3121 W2005

Hypothesis Testing

Another type of problem faced in statistics is to determine whether a condition or statement about a random situation from among different alternatives is true. This is phrased as deciding between different hypotheses regarding the situation and so is termed hypothesis testing. We shall restrict ourselves to the simplest type of such a problem where we have two hypothesis, labelled H_0 (the null hypothesis) and H_1 (the alternative hypothesis).

Ex① We might observe n samples from a Gaussian population with known variance and unknown mean and be asked to decide between

H_0: The population mean is zero
H_1: The population mean is not zero

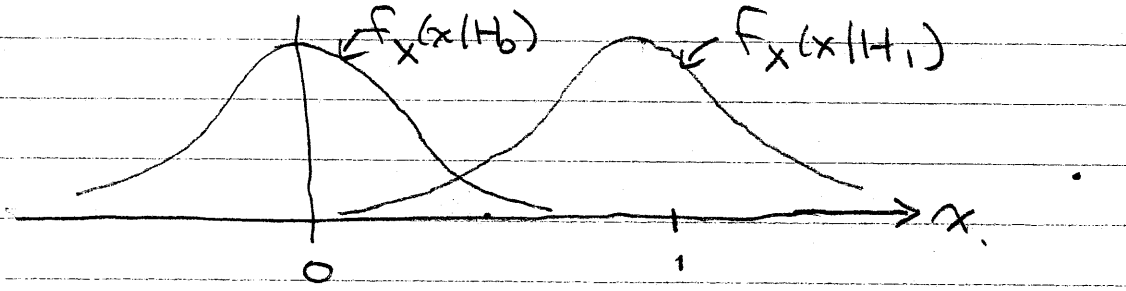
Ex② We might observe the received signal $r(t)$ in a binary digital communication signal on a symbol interval and be asked to decide between

H_0: A "0" symbol was sent
H_1: A "1" symbol was sent

The basis for making a decision is the observed data — if the data is in some set R_0 we will decide in favour of H_0 and otherwise we decide in favour of H_1. This discrimination may be phrased in terms

of a decision rule.

Ex: Suppose we observe a random value whose value is described by a Gaussian random variable X which is Gaussian with variance 1, but the mean may be either 0 or 1. We seek to determine where H_0: the mean = 0 H_1: the mean = 1



To decide between the two hypothesis we use the decision rule: "decide H_0" if $X < 1/2$ & "decide H_1 if $X \geq 1/2$". which we can express as

$$X \underset{\text{accept } H_0}{\overset{\text{accept } H_1}{\gtrless}} \frac{1}{2}$$

We note that with this decision rule, we can accept H_1 when H_0 is valid and vice versa This is typical of all hypothesis testing. With binary hypothesis testing we can make two types of errors:

Type I: accept (decide) H_1 when H_0 is valid

Type II: accept (decide) H_0 when H_1 is valid.

The probability of making a Type I error is termed the level of significance of a test. In our example, the level of significance is $P(X \geq 1/2 \mid \text{mean}=0) = Q(1/2) \simeq .3085$. We could reduce the level of significance of a test (5% is a typical desired value) if we raised the threshold 1/2. But this would raise the probability of making a Type II error. The probability accepting H_1 when H_1 is valid is termed the power of the test.

Hypothesis testing problems don't always allow for all the probabilities of error to be computed. Suppose for example we modified the situation in the above situation where we want to test if the mean is zero or not

$$H_0: X \sim N(0,1) \quad , \quad H_1: X \sim N(\mu, 1) \text{ for some } \mu \neq 0.$$

We might then adopt a "two sided test":

$$|X| \underset{H_0}{\overset{H_1}{\gtrless}} \lambda$$

$P(\text{Type I error}) = P(|X| > \lambda \mid H_0) = 2Q(\lambda)$. We can then set λ to achieve a given level of significance ($\lambda = 1.96$ for 5% level of significance). We cannot find $P(\text{Type II error})$ since we have knowledge of the complete distribution when H_1 is valid. We could find it in terms of μ_1. A plot of $P(\text{Type II error})$ vs μ_1 is termed the operating characteristic curve of the test. A plot of $1 - P(\text{Type II error}) = P(\text{decide } H_1 \mid H_1 \text{ valid})$ vs μ_1 is termed the

power function of the test

An important question hypothesis testing is how to design the best test ~~that~~ makes use of the available data. The answer to this depends on the criteria we adopt to judge the quality of a test and on the type of hypotheses we have. This is a large topic which is unfortunately beyond the scope of this course. . .

It should be noted that when we perform a test and say decide to accept H_0, we should not say H_0 is valid. The correct conclusion we can reach should be stated as
"base on the observation we accept the hypotheses that ... at the ... level of significance".
This recognizes the everpresent possibility of an error in a conclusion. Absolute conclusions are rarely possible in statistics.

<u>The χ^2-goodness of fit test</u>: There is one statistical test that we employ commonly to judge the validity or not of some model for the distribution. In this test we observe N supposed random samples of a random variable X. We hypothesize that the follows some given type of distribution. To determine this we divide the range of observed values some K bins or intervals (usually of some fixed length but not necessarily so). and then count the # of occurrences

of samples in each "bin". Let f_i denote the count in the ith bin. A bar graph of these frequencies vs the midpoint of the bin is termed a histogram plot.

If the hypothesized distribution is valid, we can calculate the probability p_i that a value lies in the ith bin, (this may involve estimating some parameters of the distribution from the data set). Then the expected value for f_i should be Np_i. The probability that the bin counts have a particular set of values is given by the multinomial probability

$$\frac{N!}{f_1!f_2!\cdots f_N!}(p_1)^{f_1}(p_2)^{f_2}\cdots(p_N)^{f_N}$$

The test statistic used in the χ^2-goodness of fit test is

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - Np_i)^2}{Np_i}$$

It is found that for moderately large values of N, this statistic is approximately distributed according to a χ^2 distribution with $k-l-1$ degrees of freedom where l = # of parameters that were estimated to allow p_i to be computed. The approximation is improved if we arrange the bins so that $np_i \geq 5$ (grouping bins if needed)

In the χ^2-goodness of fit test we accept H_0 (that the assumed distribution is valid) if $\chi^2 < \lambda$ for some value λ we determine from the χ^2 distribution

Example 1: A icosahedral (i.e., 20-sided) has two sides marked 1, two sides marked 2,..., and two sides marked 10. We would like to test if the dies is fair or not on the basis of the results of 200 independent throws of the die. The results of the 200 throws are shown in the table below:

Face no.	Observed No. of Samples, f_i	Expected No. of Samples, np_i	$(f_i - np_i)^2$	$\frac{(f_i - np_i)^2}{np_i}$
1	17	20	9	0.45
2	19	20	1	0.05
3	26	20	36	1.80
4	18	20	4	.20
5	16	20	16	.80
6	23	20	9	0.45
7	21	20	1	0.05
8	24	20	16	0.80
9	20	20	0	0.00
10	16	20	16	0.80
	200	200		5.40

If the die is fair, then the probability of each number from 1 to 10 is $2/20 = 0.1$, so we expect that on everage, in 20 independent throws, each number turns up 20 times on average. The question we ask, then, is whether the set of observed f_i is compatible with the null hypothesis that the die is fair...that is that $p_i = 0.1$ for each i. We note that $np_i \geq 5$ in all cases, and apply the χ^2 test. In this example, the number of bins is 10, and the distribution we are testing is fully defined, so the χ^2 statistic is expected to have a χ^2 distribution with 9 degrees of freedom. From the computation in the above table we have that the value of the statistic here is

$$\sum_{i=1}^{10} \frac{(f_i - np_i)^2}{np_i} = 5.40.$$

From the table of the tail probabilities of the χ^2 distribution in the lab sheets, we observe that for 9 degrees of freedom and a 5% level of significance, the threshold value for the test is 16.92. Since the observed value is 5.40 which is less than 16.92, we accept the hypothesis that the die is fair (i.e., that the probability of each number is 1/10) at the 5% level of

significance. Put another way, the results of the throws do not contradict, at the 5% level of significance, the null hypothesis that the die is fair.

It must be pointed out that it would be quite incorrect to state, for example, that we know the die is fair from the above results—accepting a hypothesis that the die is fair is not a claim that the die truly is fair, just as rejecting a hypothesis in a statistical test is not a claim that the hypothesis is truly incorrect. We must be careful in writing conclusions to avoid any such extravagant claims from a statistical test. Statistical tests rarely permit such absolute conclusions.

Example 2: In telephone system engineering, knowledge of the distribution of the load of telephone calls being placed is important when deciding how much capacity for handling calls is to be installed. It is believed that the number of telephone calls placed in a given time interval is accurately described by the Poisson distribution. In this description, the probability that k calls are placed in a given interval is

$$\frac{\mu^k}{k!}e^{-\mu},$$

where μ is a parameter of the distribution (which is the average number of telephone calls that could be expected in the interval). In an experiment, the number of calls placed in a two minute interval in a certain telephone office was monitored for 100 such nonoverlapping intervals with the results reported below. The telephone engineers would like to validate the Poisson model they use for the frequency of telephone calls, and so would like to test the above distribution against the observed data. No assumption is being made about the value of μ.

No. of Calls Placed	Observed No. of 2 min. Intervals
0	1
1	5
2	16
3	17
4	26
5	11
6	9
7	9
8	2
9	1
10	2
11	1
	100

The parameter μ of the model is not known, so it must be estimated. Given the interpretation of μ as the mean, a good estimator of μ would be provided by the sample mean. From the data in the above table, the average number of calls per interval is found to be 4.20 which is our estimate of μ. With this choice of μ, we have then that the model predicts that the probability that there would be k calls initiated in an interval is given by

$$p_k = \frac{(4.2)^k}{k!} e^{-4.2}.$$

Thus for 100 intervals, we would expect $100p_k$ two minute intervals would be found in which k calls were placed on average (for $k = 0, 1, 2, \ldots$). The table below lists the observed frequency and the predicted frequency of k calls in a two minute interval.

No. of Calls Made	Observed No. of 2 min. Intervals	Expected No. of 2 min. Intervals
0	1	1.5
1	5	6.3
2	16	13.2
3	17	18.5
4	26	19.4
5	11	16.3
6	9	11.4
7	9	6.9
8	2	3.6
9	1	1.7
10	2	0.7
11	1	0.4 [≥ 11]
	100	99.9

In order to use the χ^2-goodness of fit test we cannot simply compute the χ^2-statistic and compare it to a threshold determined from the distribution function tables for the χ^2-distribution, since the χ^2-statistic is closely described by the χ^2-distribution only when at least 5 observations are expected in each cell. To achieve this we simply group observations into cells (lumping the case of $k = 0$ and $k = 1$ together as well as the cases of $k = 8$, $k = 9$, $k = 10$ and $k \geq 11$. [Which cells are grouped with which is somewhat arbitrary.] This produces the table below from which we see that the

χ^2-statistic for our grouping into 8 classes has the value 6.257.

Observed No. of Samples, f_i	Expected No. of Samples, np_i	$(f_i - np_i)^2$	$\frac{(f_i - np_i)^2}{np_i}$
6	7.8	3.24	0.415
16	13.2	7.84	0.594
17	18.5	2.25	0.122
26	19.4	43.56	2.245
11	16.3	28.09	1.723
9	11.4	5.76	0.505
9	6.9	4.41	0.639
6	6.3	0.09	0.014
100	99.9		6.257

Since one of the parameters of the model was estimated, the χ^2-statistic we have is supposedly well described by a χ^2-distribution with $8-1-1=6$ degrees of freedom. If we test at the 5% level of significance, the threshold for the test is found from table in Appendix B to be given by 12.59. Hence we find that *the observed data is consistent with a Poisson distribution hypothesis for the number of telephone calls placed in the chosen interval in the particular telephone office at the 5% level of significance.* We have also *estimated* the appropriate parameter for the Poisson distribution to apply is $\mu = 4.2$.

Note that this test is NOT the same as having tested for the observation being described as a Poisson distribution with mean 4.20 (the value 4.20 is assumed fixed before the observations were made and is only exactly the sample mean by a lucky coincidence). If we were making this test, the computation of the χ^2-statistic would be exactly as above, but this statistic would be described by a χ^2-distribution with 7 degrees of freedom, for which the threshold increases to 14.07 at the 5% level of significance. This test is also passed, from which we would be able to state that *the observed data is consistent with a Poisson-distribution-with-mean-4.2 hypothesis for the number of telephone calls placed in the chosen interval in the particular telephone office at the 5% level of significance.* (The difference here between the two tests is that we are testing the value of the mean as well as the distribution shape, while in the first test we are only testing for the shape of the distribution; had the sample mean not been 4.20, the computations of the χ^2-statistic would be different for the two tests.)