7. Testing problems - first example

Earlier in the course we discussed the problem of how to test whether a “psychic” can make predictions better than a random guesser. This is a prototype of what are called *testing problems*. We start with this simple example and introduce various general terms and notions in the context of this problem.

Question 172. A “psychic” claims to guess the order of cards in a deck. We shuffle a deck of cards, ask her to guess and count the number of correct guesses, say X.

One hypotheses (we call it the *null hypothesis* and denote it by H_0) is that the psychic is guessing randomly. The *alternate hypothesis* (denoted H_1) is that his/her guesses are better than random guessing (in itself this does not imply existence of psychic powers. It could be that he/she has managed to see some of the cards etc.). Can we decide between the two hypotheses based on X?

What we need is a rule for deciding which hypothesis is true. A rule for deciding between the hypotheses is called a *test*. For example, the following are examples of rules (the only condition is that the rule must depend only on the data at hand).

Example 173. We present three possible rules.

(1) If X is an even number declare that H_1 is true. Else declare that H_1 is false.
(2) If $X \geq 5$, then accept H_1, else reject H_1.
(3) If $X \geq 8$, then accept H_1, else reject H_1.

The first rule does not make much sense as the parity (evenness or oddness) has little to do with either hypothesis. On the other hand, the other two rules make some sense. They rely on the fact that if H_1 is true then we expect X to be larger than if H_0 is true. But the question still remains, should we draw the line at 5 or at 8 or somewhere else?

In testing problems there is only one objective, to avoid the following two possible types of mistakes.

Type-I error: H_0 is true but our rule concludes H_1.

Type-II error: H_1 is true but our rule concludes H_0.

The probability of type-I error is called the *significance level* of the test and usually denote by α. That is, $\alpha = \mathbf{P}_{H_0}\{\text{the test accepts } H_1\}$ where we write $\mathbf{P}_{H_0}$ to mean that the probability is calculated under the assumption that H_0 is true. Similarly one define the *power* of the test as $\beta = \mathbf{P}_{H_1}\{\text{the test accepts } H_1\}$. Note that β is the probability of not making type-II error, and hence we would like it to be close to 1. Given two tests with the same level of significance, the one with higher power is better. Ideally we would like both to be small, but that is not always achievable.

We fix the desired level of significance, usually $\alpha = 0.05$ or 0.1 and only consider tests whose probability of type-I error is at most α. It may seem surprising that we take α to be so small. Indeed the two hypotheses are not treated equally. Usually H_0 is the default option, representing traditional belief and H_1 is a claim that must prove itself. As such, the burden of proof is on H_1.

To use analogy with law, when a person is convicted, there are two hypotheses, one that he is guilty and the other that he is not guilty. According to the maxim “innocent till proved guilty”, one is not required to prove his/her innocence. On the other hand guilt must be proved. Thus the null hypothesis is “not guilty” and the alternative hypothesis is “guilty”.

In our example of card-guessing, assuming random guessing, we have calculated the distribution of X long ago. Let $p_k = \mathbf{P}\{X = k\}$ for $k = 0, 1, \ldots, 52$. Now consider a test of the form "Accept H_1 if $X \geq k_0$ and reject otherwise". Its level of significance is

$$\mathbf{P}_{H_0}\{\text{accept } H_1\} = \mathbf{P}_{H_0}\{X \geq k_0\} = \sum_{i=k_0}^{52} p_i.$$

For $k_0 = 0$, the right side is 1 while for $k_0 = 52$ it is $1/52!$ which is tiny. As we increase k_0 there is a first time where it becomes less than or equal to α. We take that k_0 to be the threshold for cut-off.

In the same example of card-guessing, let $\alpha = 0.01$. Let us also assume that Poisson approximation holds. This means that $p_j \approx e^{-1}/j!$ for each j. Then, we are looking for the smallest k_0 such that $\sum_{j=k_0}^{\infty} e^{-1}/j! \leq 0.01$. For $k_0 = 4$, this sum is about 0.019 while for $k_0 = 5$ this sum is 0.004. Hence, we take $k_0 = 5$. In other words, accept H_1 if $X \geq 5$ and reject if $X < 5$. If we took $\alpha = 0.0001$ we would get $k_0 = 7$ and so on.

Strength of evidence: Rather than merely say that we accepted H_1 or rejected it would be better to say how strong the evidence is in favour of the alternative hypothesis. This is captured by the *p-value*, a central concept of decision making. It is defined as *the probability that data drawn from the null hypothesis would show closer agreement with the alternative hypothesis than the data we have at hand* (read it five times!).

Before we compute it in our example, let us return to the analogy with law. Suppose a man is convicted for murder. Recall that H_0 is that he is not guilty and H_1 is that he is guilty. Suppose his fingerprints were found in the house of the murdered person. Does it prove his guilt? It is some evidence in favour of it, but not necessarily strong. For example, if the convict was a friend of the murdered person, then he might be innocent but have left his fingerprints on his visits to his friend. However if the convict is a total stranger, then one wonders why, if he was innocent, his finger prints were found there. The evidence is stronger for guilt. If bloodstains are found on his shirt, the evidence would be even stronger! In saying this, we are asking ourselves questions like "if he was innocent, how likely is it that his shirt is blood-stained?". That is p-value. Smaller the p-value, stronger the evidence for the alternate hypothesis.

Now we return to our example. Suppose the observed value is $X_{\text{obs}} = 4$. Then the p-value is $\mathbf{P}\{X \geq 4\} = p_4 + \ldots + p_{52} \approx 0.019$. If the observed value was $X_{\text{obs}} = 6$, then the p-value would be $p_6 + \ldots + p_{52} \approx 0.00059$. Note that the computation of p-value does not depend on the level of significance. It just depends on the given hypotheses and the chosen test.

8. Testing for the mean of a normal population

Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. We shall consider the following hypothesis testing problems.

(1) One sided test for the mean. $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$.
(2) Two sided test for the mean. $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.

This kind of problem arises in many situations in comparing the effect of a treatment as follows.

Example 174. Consider a drug claimed to reduce blood pressure. How do we check if it actually does? We take a random sample of n patients, measure their blood pressures $Y_1, \ldots, Y_n$. We administer the drug to each of them and again measure the blood pressures

$Y'_1,\dots,Y'_n$, respectively. Then, the question is whether the mean blood pressure decreases upon giving the treatment. To this effect, we define $X_i = Y_i - Y'_i$ and wish to test the hypothesis that the mean of X_is is strictly positive. If X_i are indeed normally distributed, this is exactly the one-sided test above.

Example 175. The same applies to test the efficacy of a fertilizer to increase yield, a proposed drug to decrease weight, a particular educational method to improve a skill, or a particular course such as the current *probability and statistics course* in increasing subject knowledge. To make a policy decision on such matters, we can conduct an experiment as in the above example.

For example, a bunch of students are tested on probability and statistics and their scores are noted. Then they are subjected to the course for a semester. They are tested again after the course (for the same marks, and at the same level of difficulty) and the scores are again noted. Take differences of the scores before and after, and test whether the mean of these differences is positive (or negative, depending on how you take the difference). This is a one-sided tests for the mean. Note that in these examples, we are taking the null hypothesis to be that there is no effect. In other words, the burden of proof is on the new drug or fertilizer or the instructor of the course.

The test: Now we present the test. We shall use the statistic $\mathcal{T} := \frac{\sqrt{n}(\overline{X}-\mu_0)}{s}$ where $\overline{X}$ and s are the sample mean and sample standard deviation.

(1) In the one-sided test, we accept the alternative hypothesis if $\mathcal{T} > t_{n-1}(\alpha)$.
(2) In the two sided-test, accept the alternative hypothesis if $\mathcal{T} > t_{n-1}(\alpha/2)$ or $\mathcal{T} < -t_{n-1}(\alpha/2)$.

The rationale behind the tests: If $\overline{X}$ is much larger than μ_0 then the greater is the evidence that the true mean μ is greater than μ_0. However, the magnitude depends on the standard deviation and hence we divide by s (if we knew σ we would divide by that). Another way to see that this is reasonable is that $\mathcal{T}$ does not depend on the units in which you measure X_is (whether X_i are measured in meters or centimeters, the value of $\mathcal{T}$ does not change).

The significance level is α: The question is where to draw the threshold. We have seen before that *under the null hypothesis* $\mathcal{T}$ has a t_{n-1} distribution. Recall that this is because, if the null hypothesis is true, then $\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma} \sim N(0,1)$, $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$ and the two are independent. Thus, the given tests have significance level α for the two problems.

Remark 176. Earlier we considered the problem of constructing a $(1-\alpha)$-CI for μ when σ^2 is unknown. The two sided test abovecan be simply stated as follows: Accept the alternative at level α if the corresponding $(1-\alpha)$-CI does not contain μ_0. Conversely, if we had dealt with testing problems first, we could define a confidence interval as the set of all those μ_0 for which the corresponding test rejects the alternative.

Thus, confidence intervals and testing are closely related. This is true in some greater generality. For example, we did not construct confidence interval for μ, but you should do so and check that it is closely related to the one-sided tests above.

9. Testing for the difference between means of two normal populations

Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu_1, \sigma_1^2)$ and let $Y_1, \ldots, Y_m$ be i.i.d. $N(\mu_2, \sigma_2^2)$. We shall consider the following hypothesis testing problems.

(1) One sided test for the difference in means. $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 > \mu_2$.
(2) Two sided test for the mean. $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$.

This kind of problem arises in many situations in comparing two different populations or the effect of two different treatments etc. Actual data sets of such questions can be found in the homework.

Example 177. Suppose a new drug to reduce blood pressure is introduced by a pharmaceutical company. There is already an existing drug in the market which is working reasonably alright. But it is claimed by the company that the new drug is better. How to test this claim?

We take a random sample of $n+m$ patients and break them into two groups of n and of m patients. The first group is administered the new drug while the second group is administered the old drug. Let $X_1, \ldots, X_n$ be the *decrease in blood pressures* in the first group. Let $Y_1, \ldots, Y_m$ be the *decrease* in blood pressures in the second group. The claim is that one average X_is are larger than Y_is.

Note that it does not make sense to subtract $X_i - Y_i$ and reduce to a one sample test as in the previous section (here X_i is a measurement on one person and Y_i is a measurement on a completely different person! Even the number of persons in the two groups may differ). This is an example of a two-sample test as formulated above.

Example 178. The same applies to many studies of comparision. If someone claims that Americans are taller than Indians on average, or if it is claimed that cycling a lot leads to increase in height, or if it is claimed that Chinese have higher IQ than Europeans, or if it is claimed that *Honda Activa* gives better mileage than *Suzuki Access*, etc., etc., the claims can be reduced to the two-sample testing problem as introduced above.

BIG ASSUMPTION: We shall assume that $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (yet unknown). This assumption is not made because it is natural or because it is often observed, but because it leads to mathematical simplification. Without this assumption, no exact level-α test has been found!

The test: Let $\overline{X}, \overline{Y}$ denote the sample means of X and Y and let s_X, s_Y denote the corresponding sample standard deviations. Since σ^2 is the assumed to be the same for both populations, s_X^2 and s_Y^2 can be combined to define

$$S^2 := \frac{(n-1)s_X^2 + (m-1)s_Y^2}{m+n-2}$$

which is a better estimate for σ^2 than just s_X^2 or s_Y^2 (this S^2 is better than simply taking $(s_X^2 + s_Y^2)/2$ because it gives greater weight to the larger sample).

Now define $\mathcal{T} = \sqrt{\frac{1}{n} + \frac{1}{m}} \left(\frac{\overline{X} - \overline{Y}}{S} \right)$. The following tests hav significance level α.

(1) For the one-sided test, accept the alternative if $\mathcal{T} > t_{n+m-2}(\alpha)$.
(2) For the one-sided test, accept the alternative if $\mathcal{T} > t_{n+m-2}(\alpha/2)$ or $\mathcal{T} < -t_{n+m-2}(\alpha/2)$.

The rationale behind the tests: If $\overline{X}$ is much larger than $\overline{Y}$ then the greater is the evidence that the true mean μ_1 is greater than μ_2. But again we need to standardize by dividing this by an estimate of σ, namely S. The resulting statistic $\mathcal{T}$ has a t_{m+n-2} distribution as explained below.

The significance level is α: The question is where to draw the threshold. From the facts we know,

$$\begin{aligned}\overline{X} &\sim N(\mu_1, \sigma_1^2/n),\\ \overline{Y} &\sim N(\mu_2, \sigma_2^2/m),\\ \frac{(n-1)}{\sigma^2} s_X^2 &\sim \chi^2_{n-1},\\ \frac{(m-1)}{\sigma^2} s_Y^2 &\sim \chi^2_{m-1}\end{aligned}$$

and the four random variables are independent. From this, it follows that $(m+n-2)S^2$ has χ^2_{n+m-2} distribution. *Under the null hypothesis* $\frac{1}{\sigma}\sqrt{\frac{1}{n}+\frac{1}{m}}(\overline{X}-\overline{Y})$ has $N(0,1)$ distribution and is independent of S. Taking ratios, we see that $\mathcal{T}$ has t_{m+n-2} distribution (under the null hypothesis).

10. Testing for the mean in absence of normality

Suppose $X_1, \ldots, X_n$ are i.i.d. $\text{Ber}(p)$. Consider the test

$$H_0 : \ p = p_0 \quad \text{versus} \quad H_1 : \ p \neq p_0.$$

One can also consider the one-sided test. Just as in the confidence interval problem, we can give a solution when n is large, using the approximation provided by the central limit theorem. Recall that an approximate $(1-\alpha)$-CI is

$$\left[\overline{X}_n - z_{\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}, \overline{X}_n + z_{\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right].$$

Inverting this confidence interval, we see that a reasonable test is:

Reject the alternative if p_0 belongs to the above CI. That is, accept the alternative if

$$\overline{X}_n - z_{\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}} \leq p_0 \leq \overline{X}_n + z_{\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}$$

This test has (approximately) significance level α.

More generally, if we have data $X_1, \ldots, X_n$ from a population with mean μ and variance σ^2, then consider the test

$$H_0 : \ \mu = \mu_0 \quad \text{versus} \quad H_1 : \ \mu \neq \mu_0.$$

A test with approximate significance level α is given by: Reject the alternative if

$$\overline{X}_n - z_{\alpha/2}\frac{s_n}{\sqrt{n}} \leq \mu_0 \leq \overline{X}_n + z_{\alpha/2}\frac{s_n}{\sqrt{n}}.$$

Just as with confidence intervals, we can find the actual level of significance (if n is not large enough) by simulating data on a computer.