1. Can a "psychic" make better predictions than a random guesser?

A psychic claims to have divine visions unavailable to most of us. You are assigned the task of testing her claims. You take a standard deck of cards, shuffle it well and keep it face down on the table. The psychic writes down the list of cards in some order - whatever her vision tells her about how the deck is ordered. Then, you count the *number of correct guesses*. If the number is 1 or 2, perhaps you can dismiss her claims. If it is 45, perhaps you ought to be take her seriously. Again, where to draw the line? This is a prototype of what are called *testing problems*.
Let us first do some standard calculations.

Example 1. A well shuffled deck of cards is placed face down on a table. A 'psychic' writes down her guess of the order of cards in the deck. Finally, the psychic's list is compared with the actual deck. Let X be the number of correct guesses. Find $\mathbf{P}[X = k]$.

Solution: Number the cards from 1 to 52 in some way. Then, $\Omega = \{(i_1, \ldots, i_{52}) : i_j$ are distinct numbers from $1, 2, \ldots, 52\}$ is the set of all permutations of the numbers $1, \ldots, 52$, and $p_\omega = 1/52!$.

Let A_k be the event that the k^{th} card is guessed right. Then, $X = \sum_{k=1}^{52} \mathbf{1}_{A_k}$ and $\mathbf{1}_A$ is the indicator function of the event A. Its range is $\{0, 1, \ldots, 52\}$. Now,

$$\mathbf{P}[X = k] = \mathbf{P}\left[\text{exactly } k \text{ out of } A_1, \ldots, A_{52} \text{ occur}\right] = S_k - \binom{k+1}{k} S_{k+1} + \cdots \pm \binom{52}{k} S_{52},$$

where $S_j = \sum_{1 \le i_1 < \cdots < i_j \le 52} \mathbf{P}(A_{i_1} \cap \cdots \cap A_{i_j})$.

Recall this expression from the 'Inclusion-Exclusion formula' in Module 2 of Note 2. In our case, we can see that

$$S_j = \binom{52}{j} \frac{1}{(52)(52-1)\cdots(52-j+1)} = \frac{1}{j!}.$$

Hence, we get

$$\mathbf{P}[X = k] = \frac{1}{k!} - \binom{k+1}{k} \frac{1}{(k+1)!} + \cdots \pm \binom{52}{k} \frac{1}{52!} = \frac{1}{k!}\left(1 - \frac{1}{2!} + \frac{1}{3!} - \cdots \pm \frac{1}{(52-k)!}\right).$$

Remark 2. If k is small, then the sum inside the bracket is fairly close to e^{-1}, and we get the approximation $\mathbf{P}[X = k] \approx \frac{e^{-1}}{k!}$ (which is valid for small k). So, the number of correct guesses is approximately Poisson distributed with parameter 1. The sum in the series is up to $52 - k$ instead of up to infinity. This is the approximation step. If we fix k and play with n cards (instead of 52), then as $n \to \infty$, the approximation gets better and better and we get $\mathbf{P}[X = k] \to \frac{e^{-1}}{k!}$ as $n \to \infty$. One can quantify the approximation with more work.

2. HYPOTHESIS TESTING - FIRST EXAMPLES

We start with simple examples, and then introduce various general terms and notions.

Question 3. A p-coin ($0 < p < 1$) is tossed n times with outcomes $X_1, \ldots, X_n$. Suppose we are told that the value of p is either $1/2$ or $3/4$. How to decide which, using the data $X = (X_1, \ldots, X_n)$?

Question 4. For example, one can imagine a new drug released by a pharmaceutical company to cure a disease $\mathcal{D}$. The company claims that it is effective in 75% of the cases, while there is a contrary opinion that it is effective in only 50% of the cases.

Question 5. 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 her guesses are better than random guessing (in itself this does not imply existence of psychic powers, it could be that she has managed to see some of the cards, etc.). Can we decide between the two hypotheses based on the data 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 6. We present three possible rules:

(1) If X is an even number declare that H_0 is false. Else, declare that H_0 is true.

(2) If $X \geq 5$, then reject H_0, else accept H_0.

(3) If $X \geq 8$, then reject H_0, else accept H_0.

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_0 is false, 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, 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 denoted by α. So,

$$\alpha = \mathbf{P}_{H_0}\{\text{the test rejects } H_0\},$$

where we write $\mathbf{P}_{H_0}$ to mean that the probability is calculated under the assumption that H_0 is true. Similarly, one defines the *power* of the test as

$$\beta = \mathbf{P}_{H_1}\{\text{the test rejects } H_0\}.$$

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 errors 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 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. Let $p_k = \mathbf{P}\{X = k\}$ for $k = 0, 1, \ldots, 52$. Now, consider a test of the form "Reject H_0 if $X \geq k_0$ and accept otherwise". Its level of significance is

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

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, reject H_0 if $X \geq 5$ and accept H_0 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 (or, rejected) H_1, 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 has 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 $\mathbf{P}\{X \geq 4\} = p_4 + \cdots + p_{52} \approx 0.019$. If the observed value was $X_{\text{obs}} = 6$, then the p-value would be $p_6 + \cdots + p_{52} \approx 0.00059$. Note that the computation of p-value does not depend on the level of significance α. It just depends on the given hypotheses and the chosen test.

3. 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 7. 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 $Y_1', \ldots, Y_n'$, 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 sample mean of X_is is strictly positive. If X_i are indeed normally distributed, this is exactly the one-sided test above.

Example 8. 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* (MSO201A) 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 (at the same level of difficulty with equal total marks) 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 test 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, fertilizer, or the instructor of the MSO201A course!

The test: Now, we present the test. Fix $0 < \alpha < 1$. We shall use the statistic $\mathcal{T} := \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{s_n}$, where $\bar{X}_n$ and s_n are the sample mean and sample standard deviation, respectively.

(1) In the one-sided test, we reject the null hypothesis (H_0) if $\mathcal{T} > t_{n-1}(\alpha)$.

(2) In the two sided-test, we reject the null hypothesis (H_0) if $\mathcal{T} > t_{n-1}(\alpha/2)$ or $\mathcal{T} < -t_{n-1}(\alpha/2)$.

The rationale behind the tests: If $\bar{X}$ 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_n (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 $\frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma} \sim N(0,1)$, $(n-1)s_n^2/\sigma^2 \sim \chi^2_{n-1}$ and the two are independent. Thus, the given tests have significance level α for the two problems.

Remark 9. Earlier, we had considered the problem of constructing a $(1-\alpha)$-CI for μ when σ^2 is unknown. The two sided test above can 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 of μ for the one-sided tests above, but you should do so, and check that they are closely related.

4. TESTING FOR THE MEAN IN ABSENCE OF NORMALITY

Fix $0 < \alpha < 1$. Suppose $X_1, \ldots, X_n$ are i.i.d. 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 (CLT). Recall that an approximate $(1-\alpha)$-CI for p is

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

Inverting this confidence interval, we see that a reasonable test is:
Accept the null if p_0 belongs to the above CI, i.e., accept the null if

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

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: Accept the null if

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

Just as with confidence intervals, we can find the actual level of significance (if n is not large enough) by using simulations.

5. Testing for the difference in means of two normal populations

The mathematical setup is as follows. Let $X_1, \ldots, X_n$ be i.i.d. samples from $N(\mu_1, \sigma_1^2)$ distribution, and $Y_1, \ldots, Y_m$ be i.i.d. samples from $N(\mu_2, \sigma_2^2)$. We assume that X_is are independent of Y_js. Our objective is to test

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_1 : \mu_1 > \mu_2.$$

This is a very common question that arises in comparing two populations. For example, imagine a new fertilizer that claims to give better yields than an existing one. To test this, we conduct an experiment where we divide a large tract of land into $m + n$ equal areas. The new fertilizer is applied to the first n tracts, while the old fertilizer is applied to the next m plots. The yields at the end of the season the first set of tracts are $X_1, \ldots, X_n$ and in the second set of plots are $Y_1, \ldots, Y_m$. Based on this data, we must decide whether the second fertilizer is indeed better as claimed. Note that we have taken the null hypothesis to be "$\mu_1 = \mu_2$", indicating that the burden of proof is on the new fertilizer.

Many similar problems can be considered:

- A pharmaceutical company releases a new drug. Test if it is better than the old one.
- A new method of teaching is claimed to be better than an existing one.
- It is claimed that one set of people have higher IQ than another.

Since comparing things is an ever present obsession of humans, you can add any number of other examples on your own! It is clear that the above testing problem captures all these situations except for one serious issue, the assumption of normality. We will only say that if m and n are both large, then like in the one sample test for the mean, we can use the central limit theorem to obtain approximate level α ($0 < \alpha < 1$) tests.

Now, we return to the mathematical setting. We have $X_1, \ldots, X_n$ to be an i.i.d. sample from $N(\mu_1, \sigma_1^2)$ distribution and $Y_1, \ldots, Y_m$ to be an i.i.d. sample from $N(\mu_2, \sigma_2^2)$ distribution. We assume that X_is are independent of Y_js. Our objective is to test

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_1 : \mu_1 > \mu_2.$$

From previous facts, recall (about the sample mean and the sample variance) that

$$\bar{X} \sim N(\mu_1, \sigma_1^2/n), \quad \frac{(n-1)}{\sigma_1^2} s_X^2 \sim \chi^2_{n-1},$$

$$\bar{Y} \sim N(\mu_2, \sigma_2^2/m), \quad \frac{(m-1)}{\sigma_2^2} s_Y^2 \sim \chi^2_{m-1}.$$

Further, all these four random variables are mutually independent (Why?). Consequently, we see that

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right) \text{ and } \frac{(n-1)}{\sigma_1^2} s_X^2 + \frac{(m-1)}{\sigma_2^2} s_Y^2 \sim \chi^2_{m+n-2}.$$

As with testing for the mean (when the variance was unknown), we want to make a random variable that depends on the data and the difference in means (since we are testing whether $\mu_1 - \mu_2$ is zero, or positive), but not on the unknown variance. Unfortunately, it is known how to do this (you can try manipulating the above random variables!). But, we can do it, under a simplifying assumption.

Assumption: The population variances are equal, i.e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (not necessarily known). In that case, we have

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0, 1),$$

$$\frac{1}{\sigma^2}\{(n-1)s_X^2 + (m-1)s_Y^2\} \sim \chi^2_{m+n-2}.$$

Further, the two random variables are independent. Therefore,

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{n} + \frac{1}{m}}\sqrt{\frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}}} \sim t_{n+m-2}.$$

If the null hypothesis was true, the left hand side yields the following statistic:

$$\mathcal{T}_{m,n} := \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{1}{n} + \frac{1}{m}}\sqrt{\frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}}}.$$

Hence, we can use this to test for $\mu_1 - \mu_2$ (or equivalently, to construct a confidence interval for it).

Testing rule: If $\mathcal{T}_{m,n} > t_{m+n-2}(\alpha)$, then reject the null hypothesis. Else, accept it.

Without the normality assumption: For large (m, n), we can still say that $\mathcal{T}_{m,n}$ has approximately standard normal distribution (under the null hypothesis). Hence, we can use $\mathcal{T}_{m,n}$ to test for $\mu_1 - \mu_2$.