# NMAI059 – Probability and Statistics 1

Mykhaylo Tyomkyn

## Lecture 13 - Hypothesis testing.

**Definition 1 (Simple/Composite hypothesis)** *If $|\Theta_0| = 1$, that is, when $H_0$ is a single distribution, we say that the null-hypothesis is* simple. *Similarly if $|\Theta_1| = 1$ we say that the alternative is simple. A hypothesis that is not simple is called* composite. *If both $H_0$ and $H_1$ are simple, we refer to this as* simple binary hypothesis testing.

Note that in the latter setting $\alpha$ and $\beta$ are simply the probabilities of "making the wrong choice" under $H_0$ and $H_1$ respectively. In particular, $\beta$ can be viewed as a single number, rather than a function.

**Example 1** *Consider $H_0 \colon X \sim \mathcal{N}(0,1)$ and $H_1 \colon X \sim \mathcal{N}(1,1)$. Fix a confidence level $1-\alpha$. An intuitively sensible decision rule would be to take the test statistic*

$$h(X_1, \dots, X_n) = S_n = \sum_{i=1}^{n} X_i,$$

*and the rejection region $W = [\xi, \infty)$, for some $\xi = \xi(\alpha, n)$. That is, we opt for $H_1$ if and only if*

$$\sum_{i=1}^{n} X_i \geq \xi.$$

*To determine $\xi$ we recall that, by the convolution formula, $S_n \sim \mathcal{N}(0,n)$, and then use the normal tables to determine $\xi$ such that*[1]

$$\mathbb{P}_{H_0}(S_n \geq \xi) = \alpha.$$

*After this, we use the normal tables again in order to compute*

$$\beta = \mathbb{P}_{H_1}(S_n < \xi).$$

**Example 2** *Consider $H_0 \sim \mathcal{N}(0,1)$ and $H_1 \sim \mathcal{N}(0,4)$. Here a reasonable approach would be to take*

$$h(X_1, \dots, X_n) = \sum_{i=1}^{n} X_i^2,$$

*and $W = [\xi', \infty)$ for some $\xi' = \xi'(\alpha, n)$.*

It turns out that we can use the likelihood functions, defined when dealing with estimators, to design good hypothesis tests.

[1] Exercise: do this for $\alpha = 0.05$. Be careful: how does the answer scale with $n$?

**Definition 2 (Likelihood ratio, likelihood-ratio test)** *Suppose $H_0$ and $H_1$ are simple and that the underlying distributions $\mathbb{P}_{\vartheta_0}$ and $\mathbb{P}_{\vartheta_1}$ both have pmf's or pdf's. The* likelihood ratio *is a function $L\colon \mathbb{R}^n \to \mathbb{R}$,*

$$L(x) = \frac{L(x, \vartheta_1)}{L(x, \vartheta_0)},$$

*where $L(x, \vartheta)$ is the likelihood function. A* likelihood-ratio test *rejects $H_0$ if and only if $L(X) \geq \xi$ for $\xi = \xi(\alpha, n)$.*

**Example 3** *Consider again Example 1. For $x = (x_1, \ldots, x_n)$ we have*

$$L(x) = \frac{(1/\sqrt{2\pi})^n \exp\left(-\sum_{i=1}^n (x_i - 1)^2/2\right)}{(1/\sqrt{2\pi})^n \exp\left(-\sum_{i=1}^n x_i^2/2\right)},$$

*which is at least $\xi$ if and only if $x_1 + \cdots + x_n \geq \xi'$ for an appropriately chosen $\xi' = \xi'(\alpha, n)$.*

**Exercise 1** *Verify that the test described in Example 2 is also a likelihood-ratio test.*

It turns out that in many scenarios a likelihood-ratio test is optimal.

**Definition 3 (Uniformly most powerful test (UMP))** *A* uniformly most powerful test (UMP) *for simple binary hypothesis testing is a test that has the greatest power $1 - \beta$ among all tests with a given $\alpha$.*

**Theorem 1 (Neyman-Pearson lemma)** *Suppose $H_0$ and $H_1$ are simple and that the underlying distributions $\mathbb{P}_{\vartheta_0}$ and $\mathbb{P}_{\vartheta_1}$ both have pmf's or pdf's. Then for any $\alpha \in (0, 1)$ there exists an (essentially) unique UMP among tests of significance level $\alpha$, and it is a likelihood-ratio test.*

Let us now move on to studying tests with composite hypotheses.

How would we test if a coin is fair ($H_0\colon X \sim Bern(1/2)$, $H_1\colon X \sim Bern(p), p \neq 1/2$)? One natural method would be to use the test statistic $S_n = \sum_{i=1}^n X_i$ and reject $H_0$ if and only if $|S_n - n/2| > \xi$ for some $\xi = \xi(\alpha, n)$. Let us say, we have $n = 500$ and $\alpha = 0.05$. Then, using CLT and the normal tables, we obtain

$$\mathbb{P}_{p=1/2}(|S - 500| \leq 31) \approx 0.95,$$

so we take $\xi = 31$. But what instead of the coin we had a die; how would we test, whether it is fair?

More broadly, let us consider the 'generalized die', which has $k$ possible outcomes $a_1 \ldots, a_k$. Our null-hypothesis $H_0$ is that $\mathbb{P}(a_i) = p_i$ for all $1 \leq i \leq k$ (where $p_1 + \cdots + p_k = 1$), while $H_1$ is "not $H_0$". Given a sample of size $n$, let $N_i$ be the number of observed occurrences of outcome $a_i$ among $X_1, \ldots, X_n$ (the "histogram" of our sample). The *chi-square goodness of fit test* uses the test statistic

$$T = \sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i},$$

to test $H_0$ against $H_1$. The null-hypothesis $H_0$ is rejected if and only if $T > \xi$, where $\xi = \xi(\alpha, n)$ is chosen to satisfy

$$\mathbb{P}_{H_0}(T > \xi) = \alpha.$$

How to find $\xi$ is described below.

**Lemma 1** *Given independent $\mathcal{N}(0,1)$-variables $X_1, \ldots, X_n$, the random variable $Y = X_1^2 + \cdots + X_n^2$ satisfies $Y \sim \Gamma(1/2, n/2)$. That is, it has the pdf*

$$f_Y(y) = \frac{y^{n/2-1} e^{-y/2}}{\Gamma(n/2) 2^{n/2}}.$$

**Proof** (sketch). First, by considering the cdf and differentiating it, one shows that for $X \sim \mathcal{N}(0,1)$ we have $X^2 \sim \Gamma(1/2, 1/2)$ (note that $\Gamma(1/2) = \sqrt{\pi}$). This proves the lemma statement for $n = 1$. By the convolution formula for the Gamma distribution,

$$\Gamma(\lambda, r) * \Gamma(\lambda, s) = \Gamma(\lambda, r + s),$$

and the general statement follows. □

**Definition 4** *The* chi-square distribution with $m$ degrees of freedom*, denoted $\chi^2_m$, is the distribution* $\Gamma(1/2, m/2)$.

Its values, for various $m$, are well-documented in tables.

**Fact 1** *For large $n$, under assumption of $H_0$ above, the test statistic $T$ has, approximately, the distribution $\chi^2_{k-1}$.*

**Example 4** *Let us illustrate the above with a standard 6-sided die (so we have $k = 6$), and let $\alpha = 0.05$. Suppose that we have thrown it $n = 600$ times and obtained the following histogram*

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $N_i$ | 92 | 120 | 88 | 98 | 95 | 107 |

*We calculate*

$$T(x_1, \ldots, x_{600}) = \sum_{i=1}^{6} \frac{(N_i - 100)^2}{100} = \frac{1}{100}(8^2 + 20^2 + 12^2 + 2^2 + 5^2 + 7^2) = 6.86.$$

*Let $F$ be the cdf of $\chi^5_2$. From the tables we find that $F^{-1}(0.95) = 11.1 > 6.86$, so here we would accept $H_0$. Moreover, also from the tables we can see that*

$$\mathbb{P}_{H_0}(T \geq 6.86) = 1 - F(6.86) = 0.23.$$

*So, informally, under $H_0$ we would obtain a result 'at least as extreme as this" 23% of the time. We say that 0.23 is the* $p$-value *of the given experiment and sample.*

Be careful: one possible misuse of statistical analysis is the so-called *p-hacking*, that is, conducting the experiment without fixing $\alpha$ in advance, and then declaring $\alpha$ to be the $p$-value of the sample.

Let us now consider another non-parametric example, the *Permutation test*. Suppose we are testing a new medication, and have two groups of participants - the first group receive the drug, the second receive the placebo. We want to find out if the drug has any effect on the recovery time. Formally, we have two samples $X_1, \dots, X_n \sim \mathbb{P}$ and $Y_1, \dots, Y_m \sim \mathbb{P}'$, and the null-hypothesis is $H_0 \colon \mathbb{P} = \mathbb{P}'$ and $H_1 \colon$ "not $H_0$". For simplicity, assume that $\mathbb{P}$ is discrete [2]. Note that we are not making any further assumptions about the distributions. Moreover, $\mathbb{P}$ and $\mathbb{P}'$ are unknown to us.

Let us choose a test statistic, for example

$$T(X_1, \dots, X_n, Y_1, \dots, Y_m) = |\bar{X}_n - \bar{Y}_m|,$$

and note that, under $H_0$ this value should be small, so we are aiming for a rejection region $W = [w, \infty)$ with an appropriately chosen $w$.

Observe that,by the mutual independence, given a realization of the sample $(x_1, \dots, x_n, y_1, \dots, y_m)$, for any permutation $\sigma$ of its values, we have

$$\mathbb{P}(X_1 = \sigma(x_1), \dots, X_n = \sigma(x_n), Y_1 = \sigma(y_1), \dots, Y_m = \sigma(y_m)) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n, Y_1 = y_1, \dots, Y_m = y_m).$$

Now we compute $T(\sigma(x_1), \dots, \sigma(x_n), \sigma(y_1), \dots, \sigma(y_m))$ each of $(n+m)!$ possible permutations $\sigma$, and arrange the resulting values in descending order:

$$t_1 \geq t_2 \geq \dots \geq t_{(m+n)!}.$$

So, we can pick $a = \lfloor \alpha \cdot (m+n)! \rfloor$ and $w = t_a$. We have

$$\mathbb{P}(H_0 \text{ falsely rejected} \mid (X_1, \dots, Y_m) \textit{ is a permutation of } (x_1, \dots, y_m)) = \frac{a}{(m+n)!}.$$

Since this holds for all possible instances of $(x_1, \dots, y_m)$, by the law of total probability we obtain

$$\mathbb{P}(H_0 \text{ falsely rejected}) = \frac{a}{(m+n)!} \approx \alpha.$$

**Example 5** *Let $\alpha = 0.05$, $n = 3$, $m = 4$, $(x_1, x_2, x_3) = (23, 33, 40)$ and $(y_1, y_2, y_3, y_4) = (19, 22, 25, 26)$. Then $t = T(x_1, \dots, y_4) = 9$, and, having compared the $\binom{7}{3} = 35$ values for all possible permutations [3] we obtain 9 is the third-largest. This would give the p-value of $3/35 \approx 8.6\%$, and since this is larger than $\alpha = 5\%$, we accept $H_0$.*

[2] the method also works for continuous measures, as they can be approximated by discrete ones.

[3] There are $7! = 5040$ permutations in total, but the value of any two will coincide if between them we only permute the first 3 and the last 4 entries