MATHEMATICAL STATISTICS, Lessons 8-10.
TESTING STATISTICAL HYPOTHESES

Parametric tests

Let $(\Omega, \mathcal{A}, \mathcal{P})$ be a dominated, identifiable, parametric statistical space, $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$. We want to decide between: $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, where $\Theta_0 \cap \Theta_1 = \emptyset$, $\Theta_0 \cup \Theta_1 = \Theta$. In case of $|\Theta_0| = 1$ we have a *simple zero-hypothesis*, otherwise it is *composite*. The same for the alternative hypothesis.

Our decision is based on the $\mathbf{X} = (X_1, \dots, X_n) \sim \mathbb{P}_\theta$ i.i.d. sample from the sample space $\mathcal{X}$. The decision process is as follows.

1. We calculate an appropriate statistic $T(\mathbf{X})$ (whose distribution under H_0, or on the boundary of H_0, is known).
2. We divide the sample space into *acceptance region* $\mathcal{X}_a$ and *rejection (or critical) region* $\mathcal{X}_c$, where $\mathcal{X}_a \cap \mathcal{X}_c = \emptyset$ and $\mathcal{X}_a \cup \mathcal{X}_c = \mathcal{X}$. Usually, $\mathcal{X}_c = \{\mathbf{x} : T(\mathbf{x}) \geq c_\alpha\}$ or $\mathcal{X}_c = \{\mathbf{x} : |T(\mathbf{x})| \geq c_\alpha\}$ where c_α is a quantile (percentile) value of T, and α is the significance (size) of the test.
3. If $\mathbf{X} \in \mathcal{X}_a$, then accept, otherwise reject H_0 with significance α.

Definition: The significance (size) of the test defined by $\mathcal{X}_c$ is

$$\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\mathbf{X} \in \mathcal{X}_c)$$

(supremum of the so-called Type I. errors).
Definition: The *power* of the test defined by $\mathcal{X}_c$ is

$$\gamma(\theta) = \mathbb{P}_\theta(\mathbf{X} \in \mathcal{X}_c) = 1 - \mathbb{P}_\theta(\mathbf{X} \in \mathcal{X}_a) = 1 - \beta(\theta), \quad \text{for} \quad \theta \in \Theta_1$$

(1 minus the Type II. error $\beta(\theta)$ for the alternative $\theta \in \Theta_1$).
Definition: The test defined by $\mathcal{X}_c$ is *uniformly most powerful (UMP) test* of significance α, if among the tests with significance at most α, its power is the largest possible, for any alternative. That is, $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\mathbf{X} \in \mathcal{X}_c) = \alpha$, and for any other $\mathcal{X}'_c$ with $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\mathbf{X} \in \mathcal{X}'_c) \leq \alpha$, the following also holds:

$$\mathbb{P}_\theta(\mathbf{X} \in \mathcal{X}_c) \geq \mathbb{P}_\theta(\mathbf{X} \in \mathcal{X}'_c), \quad \forall \theta \in \Theta_1.$$

It is easier to formulate the problem with the help of the test function Ψ, which is the characteristic function of $\mathbf{X} \in \mathcal{X}_c$, and

$$\alpha = \sup_{\theta \in \Theta_0} \mathbb{E}_\theta \Psi(\mathbf{X})$$

whereas

$$\gamma(\theta) = \mathbb{E}_\theta \Psi(\mathbf{X}) \quad \text{for} \quad \theta \in \Theta_1.$$

More generally, we consider randomized tests defined with the following test function.
Definition: The test function $\Psi(\mathbf{X})$ is the probability of rejecting H_0 based on the observation $\mathbf{X}$. It is 1 if $\mathbf{X} \in \mathcal{X}_c$, 0 if $\mathbf{X} \in \mathcal{X}_a$, and may be $p \in (0, 1)$ if $\mathbf{X} \in \mathcal{X}_r$ (randomization region), when we cannot decide immediately (see discrete cases).
Neyman–Pearson Theorem (on the existence of a UMP test). In a parametric statistical space let $L_\theta(\mathbf{X})$ be the likelihood function based on the i.i.d. sample $\mathbf{X} = (X_1, \dots, X_n)$. Then for the simple alternative

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1$$

for any $0 < \alpha < 1$ there exists a UMP test of size α, and it is defined uniquely (with probability 1) in the following way:

$$\psi(\mathbf{X}) = 0 \quad \text{if} \quad \frac{L_{\theta_1}(\mathbf{X})}{L_{\theta_0}(\mathbf{X})} < c; \qquad \psi(\mathbf{X}) = p \quad \text{if} \quad \frac{L_{\theta_1}(\mathbf{X})}{L_{\theta_0}(\mathbf{X})} = c; \qquad \psi(\mathbf{X}) = 1 \quad \text{if} \quad \frac{L_{\theta_1}(\mathbf{X})}{L_{\theta_0}(\mathbf{X})} > c,$$

where $p \in [0, 1)$ and $c > 0$ are appropriate constants depending on α.
Remarks: It is easy to see that the construction of the Neyman–Pearson theorem provides an *unbiased test*:

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta \Psi(\mathbf{X}) \le \inf_{\theta \in \Theta_1} \mathbb{E}_\theta \Psi(\mathbf{X}).$$

This theorem can be extended to composite hypotheses, when the likelihood function depends monotonously on a sufficient statistic T. In these cases the above inequalities can be reformulated in terms of T (with other constants).
Definition: The family $\{\mathbb{P}_\theta : \theta \in \Theta\}$ of probability distributions has *monotone likelihood ratio* if for $\theta < \theta'$, $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ are not identical (identifiability), and there exists a statistic T such that $\frac{L_{\theta'}(\mathbf{X})}{L_\theta(\mathbf{X})}$ is a non-decreasing function of $T(\mathbf{X})$.
Theorem: Let the family $\{\mathbb{P}_\theta : \theta \in \Theta\}$ have monotone likelihood ratio. Then for the (one-tail) alternative

$$H_0 : \theta \le \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0$$

with any $0 < \alpha < 1$ there exists a UMP test of significance α. Its test function is the following:

$$\psi(\mathbf{X}) = 0 \quad \text{if} \quad T(\mathbf{X}) < c; \qquad \psi(\mathbf{X}) = p \quad \text{if} \quad T(\mathbf{X}) = c; \qquad \psi(\mathbf{X}) = 1 \quad \text{if} \quad T(\mathbf{X}) > c,$$

where T is that of the definition, and the constants $c > 0$, $p \in [0, 1)$ can be chosen such that $\mathbb{E}_{\theta_0}(\psi(\mathbf{X})) = \alpha$ be satisfied. Further, $\mathbb{E}_\theta(\psi(\mathbf{X}))$ is strictly increasing in θ until it possibly becomes 1.

For the most important parametric tests, see the formula sheet enclosed for z, t, F, and $Welch$-tests (one- and two-sample, one- and two tail situations; further independent sample and paired sample tests).

Likelihood ratio test. Applicable when Θ_0 is a low-dimensional manifold of Θ and our sample is from an absolutely continuous distribution.

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1$$

where $\Theta_0 \cap \Theta_1 = \emptyset$, $\Theta_0 \cup \Theta_1 = \Theta$, and with the notation $\dim(\Theta_0) = r$, $\dim(\Theta) = k$, $r < k$ is satisfied. The test statistic based on an n-element sample is the following:

$$\lambda_n(\mathbf{X}) = \frac{\sup_{\theta \in \Theta_0} L_\theta(\mathbf{X})}{\sup_{\theta \in \Theta} L_\theta(\mathbf{X})}.$$

Note that $\lambda_n(\mathbf{X})$ does not depend on θ and takes on values between 0 and 1. Here $\mathcal{X}_c = \{\lambda_n \le c\}$, where c depends on the significance α; however, we can find it, only if the distribution of λ_n under H_0 is known. Otherwise, we use the fact, that under certain regularity conditions,

$$-2 \ln \lambda_n(\mathbf{X}) \to \chi^2(k - r)$$

as $n \to \infty$, under H_0. Therefore, $\mathcal{X}_c = \{-2 \ln \lambda_n \ge c'\}$, where c' is the upper α-point of the $\chi^2(k - r)$ distribution. For instance, the one-sample, two-tail z-test is such.

Non-parametric tests. H_0 applies not to the parameter. Basic idea: find the limit distribution under the zero hypothesis of a conveniently constructed statistic, where this limiting distribution does not depend on the parameters of the underlying so-called population distribution.

- *Pearson's χ^2-test.* Test statistic:

$$\chi^2 = \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i}$$

where O_i's are the Observed and E_i's are the Expected frequencies. If $n = \sum_{i=1}^r O_i = \sum_{i=1}^r E_i \to \infty$, then under the zero hypothesis it asymptotically follows $\chi^2(df)$-distribution, where $df = r - 1 - e$ (e is the number of estimated parameters). So the critical region, corresponding to a test with significance α, is the upper α-point of the $\chi^2(df)$-distribution ($1-\alpha$ quantile value, see Table).

1. χ^2-test for **goodness of fit**:

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

2. χ^2-test for **homogeneity**:

$$\chi^2 = nm \sum_{i=1}^{r} \frac{(\frac{\nu_i}{n} - \frac{\mu_i}{m})^2}{\nu_i + \mu_i}$$

3. χ^2-test for **independence**:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(\nu_{ij} - n\frac{\nu_{i\cdot}}{n}\frac{\nu_{\cdot j}}{n})^2}{n\frac{\nu_{i\cdot}}{n}\frac{\nu_{\cdot j}}{n}} = n\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(\nu_{ij} - \frac{\nu_{i\cdot}\nu_{\cdot j}}{n})^2}{\nu_{i\cdot}\nu_{\cdot j}}$$

with $df = rs - 1 - [(r-1) + (s-1)] = (r-1)(s-1)$.

- *Kolmogorov–Smirnov-test*

1. **One-sample K–S-test (for goodness of fit)**:

$$D_n = \sup_{x\in\mathbb{R}} |F_n^*(x) - F(x)|.$$

Under the zero hypothesis (the ordered sample is from the continuous F- distribution)

$$\lim_{n\to\infty} \mathbb{P}(\sqrt{n}D_n < z) = K(z), \quad \forall z \in \mathbb{R},$$

that is the test statistic $\sqrt{n}D_n$ asymptotically follows Kolmogorov-distribution.

2. **Two-sample K–S-test (for homogeneity)**:

$$D_{n,m} = \sup_{x\in\mathbb{R}} |F_n^*(x) - G_m^*(x)|.$$

Under the zero hypothesis (the two ordered samples are from the same continuous distribution), the test statistic $\sqrt{\frac{nm}{n+m}}D_{n,m}$ asymptotically follows Kolmogorov-distribution. So the critical region corresponding to a test with significance α is the upper α-point of the Kolmogorov-distribution ($1-\alpha$ quantile value, see Table). Observe that the above suprema are, in fact, maxima of finitely many terms, and the distributions can be transformed into the $\mathcal{U}(0,1)$ distribution.