Asymptotic Statistics

Lecture 4

Recap

Multivariate normal distribution and CLT

Definition.
A random vector X in $\mathbb{R}^k$ is said to have a multivariate normal distribution with parameters μ and Σ, if X is distributed as $\mu + LZ$, for $Z = (Z_1, \ldots, Z_k)$ a vector of independent, standard normal variables and L such that $\Sigma = LL^T$. Notation: $X \sim N_k(\mu, \Sigma)$.

Theorem.
Let $Y_1, Y_2, \ldots$ be i.i.d. random vectors in $\mathbb{R}^k$ with finite mean μ and covariance matrix Σ. Then

$$\sqrt{n}(\bar{Y}_n - \mu) \rightsquigarrow N_k(0, \Sigma)$$

as $n \to \infty$.

Quadratic forms

Definition.
The chisquare distribution with k degrees of freedom is the distribution of $Z_1^2 + \ldots + Z_k^2$, for $Z_1, \ldots, Z_k$ independent, standard normal random variables. Notation: χ_k^2.

Lemma.
If $X \sim N_k(0, \Sigma)$, then $\|X\|^2 \stackrel{d}{=} \sum_{i=1}^k \lambda_i Z_i^2$, where $\lambda_1, \ldots, \lambda_k$ are the eigenvalues of Σ and $Z_1, \ldots, Z_k$ independent, standard normal random variables.

Analysis of the chi square test

Statistical application: chi square test - 1

Nonparametric goodness of fit test:

Have i.i.d. sample $D_1, \ldots, D_n$ in some interval $I \subset \mathbb{R}$. To test whether the sample comes from some given distribution, divide I into k bins I_i and define

$$X_{n,i} = \#\{\text{data points } D_j \text{ that fall in bin } i\}.$$

Note that $X = (X_{n,1}, \ldots, X_{n,k})$ has a multinomial distribution with parameters n and $p = (p_1, \ldots, p_k)$, where $p_i = P(D_1 \in I_i)$.

If the true distribution of the D_i is μ, then it should hold that $p_i = a_i = \mu(I_i)$. For given μ, the numbers a_i are known. Hence, we can test whether μ is the true distribution by testing whether p equals a.

Statistical application: chi square test - 2

Let $X_n = (X_{n,1}, \dots, X_{n,k})$ be multinomial with parameters n and $p = (p_1, \dots, p_k)$. For given $a = (a_1, \dots, a_k)$, we consider the problem of testing the hypothesis $H_0 : p = a$.

The chi square test uses as test statistic the Pearson statistic

$$C_n^2 = \sum_{i=1}^{k} \frac{(X_{n,i} - na_i)^2}{na_i}.$$

Idea: this test statistic compares for every bin the observed number of observations in that bin with the number that is expected under H_0. The null is rejected for large values of C_n^2.

Statistical application: chi square test - 3

Theorem.

Let $X_n = (X_{n,1}, \ldots, X_{n,k})$ be multinomial with parameters n and $a = (a_1, \ldots, a_k) > 0$. Then $C_n^2 \rightsquigarrow \chi^2_{k-1}$.

Hence, we can use the χ^2_{k-1}-quantiles to construct a test with (approximately) a desired level α.

Delta method

Delta method - idea

Suppose we have an estimator T_n for a real variable θ and we know that $\sqrt{n}(T_n - \theta) \rightsquigarrow T$. Then for a smooth function φ, what can we say about $\varphi(T_n)$ as an estimator for $\varphi(\theta)$?

By Slutsky and the continuous mapping theorem, $\varphi(T_n)$ is consistent for $\varphi(\theta)$ (check!).

Moreover, by a Taylor approximation,

$$\sqrt{n}(\varphi(T_n) - \varphi(\theta)) \approx \sqrt{n}\varphi'(\theta)(T_n - \theta) \rightsquigarrow \varphi'(\theta)T.$$

The delta method makes this precise and multivariate.

Delta method - theorem

Theorem.
Let T_n and T be random vectors in $\mathbb{R}^k$. Suppose that for numbers $r_n \to \infty$, $r_n(T_n - \theta) \rightsquigarrow T$. Suppose that $\varphi : \mathbb{R}^k \to \mathbb{R}^m$ is a (measurable) map that is differentiable at θ, with derivative $\varphi'_\theta : \mathbb{R}^k \to \mathbb{R}^m$. Then

$$r_n(\varphi(T_n) - \varphi(\theta)) - \varphi'_\theta(r_n(T_n - \theta)) \xrightarrow{P} 0$$

and

$$r_n(\varphi(T_n) - \varphi(\theta)) \rightsquigarrow \varphi'_\theta(T).$$

→ exm. 3.2