MATH242 Lecture Notes

Zhangir Azerbayev

Spring 22

Introduction

Notes from Prof. Zhou Fan's statistics course. These notes are unofficial and all mistakes are almost certainly mine.

Contents

Chapter 1

Probability Review

1.1 Moment Generating Function

A moment-generating function (MGF) is a clothesline on which to hang the moments of a probability distribution. The *nth moment* of a random variable X is EX^n. the MGF is defined as

$$M_X(t) = Ee^{tX}$$

for $t \in \mathbb{R}$. Notice that

$$M_X(t) = \sum_{n=1}^{\infty} \frac{t^n EX^n}{n!}.$$

Proposition 1.1.1. If X and Y are independent, $M_{X+Y}(t) = M_X(t)M_Y(t)$.

So the MGF is useful because it turns convolution into multiplication. Notice that this is like a Laplace transform. We have the following theorem that we won't prove.

Theorem 1.1.2. Let X and Y be two random variables such that, for some $h > 0$ and there is a positive measure set on which $M_X(t)$ and $M_Y(t)$ are finite and $M_X(t) = M_Y(t)$. Then X and Y have the same distribution.

If $Z \sim \mathcal{N}(0, 1)$, then $M_X(t) = e^{t^2/2}$.

1.2 Multivariate Normal Distribution

The multivariate normal distribution is defined by a mean vector $\mu \in \mathbb{R}^k$ and a symmetric covariance matrix $\Sigma \in \mathbb{R}^{k \times k}$.

Definition 1.2.1. $(X_1, \ldots, X_k)$ has a normal distribution if for every linear combination $\alpha_1 X_1 + \cdots + \alpha_k X_k$ is normal, and $EX_i = \mu_i$ and $\text{Cov}(X_i, X_j) = \Sigma_{ij}$.

Chapter 2

Hypothesis Testing

2.1 Sampling Distributions

For data $X_1, \ldots, X_n$ a *statistic* $T(X_1, \ldots, X_n)$ is any real-valued function of the data. For example the sample mean, sample variance (remember sample variance is the unbiased estimator for variance), etc. We care about the *sampling distribution* of the statistic T. A central question of statistics is if we understand the distribution of $X_1, \ldots, X_n$, how can we understand the sampling distribution of T?

The *chi-square distribution with n degrees of freedom*, abbreviated χ^2_n, is the distribution of the statistic $X_1^2 + \cdots + X_n^2$.

For many simple statistics, the sampling distribution is difficult to describe exactly. For example sample mean of uniform or Bernoulli. In this case, we sampling methods or asymptotic methods. Sampling methods do monte carlo simulation on a computer and asymptotic methods apply the $n = \infty$ approximation.

The *central limit theorem* is that for large n, the distribution of $\sqrt{n}\frac{\bar{X}-\mu}{\sigma}$ approaches $\mathcal{N}(0, 1)$.

2.2 The Null Hypothesis

A *hypothesis test* is a binary question about eh distribution of data. Our goal is to figure out whether the data provides sufficiently strong evidence to reject the null hypothesis H_0 in favor of H_1. The hypotheses H_0 and H_1 are treated asymmetricaly.

A test statistic T is a statistic computed from the data such that an extreme value of T provides evidence against H_0 in favor of H_1. The *null distribution* is the distribution of T given that H_0 is true. We divide the possible values of T into a *rejection region* for H_0 and an acceptance region.

Type I error is the probability we wrongly reject H_0, that is $P_{H_0}(T$ belongs to rejection region). We choose the rejection region to ensure that the probability of Type I error is at most α, we call α the significance level. We call the smallest significance level at which we wish to reject the test the *p-value*.

To determine the rejection region for a test, we need to know the null distribution.

2.3 Simple Hypotheses

A hypothesis H_0 or H_1 is *simple* if it completely specifies the distribution of the data. For simple hypotheses, we can simulate the null distribution on a computer.

Type II error is the probability of wrongly accepting the null hypotheses, or $\beta = P_{H_1}(\text{accept } H_0)$. Equivalently we can speak of the *power of the test* $1 - \beta$, which is the probability of rejecting H_0 in favor of H_1 when H_1 is true.

Suppose we have data $X = (X_1, \ldots, X_n)$. Suppose H_0 is that X has distribution $f_0(x)$ and H_1 is that X has distribution $f_1(x)$. An intuitive statistic for testing two simple hypotheses against each other is

$$L(X) = \frac{f_0(X)}{f_1(X)}.$$

This is called the *likelihood ratio statistic* and the test that rejects H_0 for small values of $L(X)$ is called the likelihood ratio test.

Theorem 2.3.1. Neyman-Pearson Lemma Let H_0 and H_1 be simple hypotheses and fix a significance level α. Suppose there is a likelihood ratio test which rejects H_0 when $L(X) < c$ and has Type I error α. Then for any other test with Type I error α its power against H_1 is at most the power of the likelihood ratio test.

2.4 Composite Hypotheses and Pivotal Statistics

Hypotheses that are not *simple* are *composite*. A test with a composite null hypothesis has to have $P(\text{reject } H_0) \leq \alpha$ for all distributions specified by H_0. This makes things very difficult! We have to reason about an infinite family of distributions.

A way we often deal with this is to find a test statistic T whose sampling distribution is the same for every distribution in H_0, such a statistic is called *pivotal*. There is a similar consideration for a composite H_1. We often cannot maximize the power of our test for all possible distributions described by H_1, so we try to strike a balance.

2.5 One-sample t-test

Suppose $X_1, \ldots, X_n$ IID from $\mathcal{N}(\mu, \sigma^2)$ and we wish to test the hypotheses $H_0 : \sigma = 0$ and H_1 also something about μ where μ and σ^2 is unknown. The following statistic, called the *one-sample* t-statistic is pivotal.

$$T = \frac{\sqrt{n}\bar{X}}{S}.$$

If $Z \sim \mathcal{N}(0,1)$ and $U \sim \chi^2_n$ and Z and U are independent, then the distribution of $Z/\sqrt{\frac{1}{n}U}$ is called the t-distribution with n degrees of freedom, denoted t_n. The logic is that the mean and the sample variance are basically independent. This is in fact the distribution of our test statistic T. So we can do one and two sided hypothesis tests.

2.6 Sign Test

We just looked at a parametric test, but what about a non-parametric test? Consider the hypothesis $H_0 : f$ has median and $H_1 : f$ has positive median. Then the t-statistic is no longer pivotal because it relies on a normal assumption. Consider instead the sign statistic

$$S = \sum_{i=1}^{n} \mathbf{1}\{X_i > 0\}.$$

This has a binomial distribution for the null distribution! We can also use the normal approximation of the binomial distribution.

2.7 Testing Multiple Hypotheses

Science is fucked. John P.A Ioannidis "Why Most Published Research Findings Are False". If testing n null hypotheses at level α, all of which are true, then on average I'll falsely reject αn of them. Suppose you're testing if a disease is associated with 1,000,000 different genetic markers.

We will abstract tests merely as returning a p-value. A p-value is really a transformed test statistic. Under H_0, the p-value P is uniform on $[0, 1]$. Therefore one way to approach testing multiple hypotheses is to check whether the distribution of p-values is uniform.

The simplest multiple-testing correction is the *Bonferri method*. When testing n different null hypotheses, perform each at the significance level α/n instead of α. This basically sets P(rejects any null hypothesis) to α. Another way to think about the Bonferri method is this: suppose we test n null hypotheses n_0 of which are true nulls and $n - n_0$ of which are false nulls. The *family-wise error rate (FWER)* is the probability we reject at least one of the n_0 true null hypotheses. A procedure *controls FWER at level* α if it guarantees that FWER$\leq \alpha$. It's obvious that the Bonferri method controls FWER at level α. In fact, the Bonferri method controls FWER at level $\alpha\frac{n_0}{n}$, although in practice we do not know n_0. But if n_0 is small the Bonferri method is quite strict!

Chapter 3

Frequentist Parametric Models

3.1 Parameters

The focus of this chapter is estimating the parameters of a distribution and quantifying our uncertainty in our estimates. A *parametric model* is a family of probability distributions that can be described by a small number of parameters. Examples include the family of normal distributions, the family of Bernoulli distributions, and the family of Gamma distributions.

We will denote a general parametric model by $f(x \mid \theta)$, which depends on k *parameters* $\theta \in \mathbb{R}^k$. For example, for the normal family we have $\theta = (\mu, \sigma^2) \in \mathbb{R}^2$. The set of possible parameters is called *parameter space*.

The fundamental question of this chapter is the following. Suppose we have $X_1, \ldots, X_n$ drawn IID from $f(x|\theta)$, how can we estimate θ and quantify our uncertainty in this estimate?

3.2 Bias, Variance, MSE

Consider the determination of a single parameter $\theta \in \mathbb{R}$. Any estimator $\hat{\theta}$ is a statistic, that is a function $\hat{\theta}(X_1, \ldots, X_n)$ of the observed data. Suppose $X_1, \ldots, X_n$ are IID from $f(x|\theta)$. Consider three ways of determining whether $\hat{\theta}$ is a good estimate:

- The **bias** of $\hat{\theta}$ is $E_\theta[\hat{\theta}] - \theta$.
- The **standard error** of $\hat{\theta}$ is $\sqrt{\mathrm{Var}_\theta \hat{\theta}}$.
- The **mean-squared-error (MSE)** is $E_\theta[(\hat{\theta} - \theta)^2]$.

Notice that for a random variable Y and constant c we have

$$\begin{aligned} E(Y-c)^2 &= E[(Y - EY + EY - c)^2] \\ &= E[(Y-EY)^2] + E[2(Y-EY)(EY-c)] + E[(EY-c)^2] \\ &= \mathrm{Var}Y + 2(EY-c)E[Y-EY] + (EY-c)^2 \\ &= \mathrm{Var}Y + (EY-c)^2. \end{aligned}$$

A consequence of this is that

$$\mathrm{MSE} = \mathrm{variance} + \mathrm{bias}^2.$$

3.3 Method of Moments

If $\theta \in \mathbb{R}$, then a simple possibility is to set the value of θ so that $EX = \frac{1}{n}(X_1 + \cdots + X_n)$.

Generally, if $\theta \in \mathbb{R}^k$, we equate the first k moments of X with their empirical value. We call this the **method of moments estimator**.

3.4 MLE

The **likelihood** of the observed data $X_1, \ldots, X_n$ is

$$L(\theta) = \prod f(X_1 \mid \theta).$$

The **maximum likelihood estimator (MLE)** is the value of θ in the parameter space that maximizes $L(\theta)$.

In practice, we usually work with the log-likelihood

$$l(\theta) = \sum \log f(X_i \mid \theta).$$

3.5 Confidence Intervals

3.5.1 A concrete motivation

Consider a Poisson model with true parameter $\lambda_0 > 0$. We know that the MLE is $\hat{\lambda} = \overline{X}$. We also know that

$$E_{\lambda_0}[\hat{\lambda}] = \lambda_0, \mathrm{Var}_{\lambda_0}[\hat{\lambda}] = \frac{\lambda_0}{n}.$$

So $\hat{\lambda}$ is unbiased with standard error $\sqrt{\lambda_0/n}$.

By the law of large numbers, $\hat{lambda} \to \lambda_0$ in probability as $n \to \infty$. We say this means that $\hat{\lambda}$ is **consistent**. Furthermore, by the central limit theorem we have

$$\sqrt{n}(\hat{\lambda} - \lambda_0) \to \mathcal{N}(0, \lambda_0)$$

in distribution as $n \to \infty$. So for large n the distribution of $\hat{\lambda}$ is approximately $\mathcal{N}(\lambda_0, \lambda_0/n)$. We say $\hat{\lambda}$ is asymptotically normal.

This normal approximation suggests a method for quantifying the uncertainty in λ. For a desired coverage level $1 - \alpha \in (0, 1)$ we can construct a **confidence interval**, that is, a random interval that contains the true parameter λ_0 with probability $1 - \alpha$. To construct such an interval, let $z(\alpha/2)$ be the upper-$\alpha/2$ point of the standard normal distribution. Then asymptotic normality implies

$$\sqrt{\frac{\lambda_0}{n}} z(\alpha/2) \leq \hat{\lambda} - \lambda_0 \leq \sqrt{\frac{\lambda_0}{n}} z(\alpha/2).$$

holds with probability $1 - \alpha$ for large n. Therefore λ_0 belongs to the interval $\hat{\lambda} \pm \sqrt{\hat{\lambda}/n} \cdot z(\alpha/2)$ with probability approximately $1 - \alpha$.

The formal guarantee is that as $n \to \infty$, the probability that the confidence interval covers λ_0 converges to $1 - \alpha$.

3.5.2 Asymptotic normality of the MLE

In the previous part's example, we were able to analyze $\hat{lambda}$ using the LLN and CLT because the estimate is the sample average. However, consistency and asymptotic normality hold for the MLE in general.

Theorem 3.5.1. Let $f(x|\theta)$ be a parametric model, with a single parameter $\theta \in \mathbb{R}$. Let θ_0 be the true parameter and $X_1, \ldots, X_n$ drawn IID from $f(x \mid \theta_0)$. Let $\hat{\theta}$ be the MLE. Under some rather technical conditions, as $n \to \infty$

1. $\hat{\theta}$ is consistent, meaning $\hat{\theta} \to \theta_0$ in probability.
2. $\hat{\theta}$ is asymptotically normal, and $\sqrt{n}(\hat{\theta} - \theta_0) \to \mathcal{N}(0, 1/I(\theta_0))$.

The function $I(\theta)$ has two equivalent forms

$$I(\theta) = \mathrm{Var}_\theta \left[\frac{\partial}{\partial \theta} \log f(X|\theta) \right] = -E_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f(X|\theta) \right].$$

where E_θ and Var_θ are the expectation and variance over $X \sim f(x|\theta)$.

The quantity $I(\theta)$ is called the **Fisher information** and measures how sharp or shallow the maximum in the likelihood function is. Intuitively, we get an estimator with less variance when the maximum of the likelihood function is sharp.

This theorem tells us that the distribution of $\hat{\theta}$ is approximately $\mathcal{N}(\theta_0, \frac{1}{nI(\theta_0)}$. In practice, since we don't know θ_0, we approximate $I(\theta_0)$ by $I(\hat{\theta})$, which we can do because **why can we do this**. So our confidence interval with coverage $1 - \alpha$ is

$$\hat{\theta} \pm \sqrt{\frac{1}{nI(\hat{\theta})}} \cdot z(\alpha/2).$$

3.6 Plug-in Estimators and the Delta Method

Theorem 3.6.1. (The delta method). If a function $g : \mathbb{R} \to \mathbb{R}$ is continuously differentiable at $\theta \in \mathbb{R}$, and if

$$\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}(0, v(\theta))$$

in distribution as $n \to \infty$ for some variance $v(\theta)$, then

$$\sqrt{n}(g(\hat{\theta} - g(\theta)) \to \mathcal{N}(0, g'(\theta)^2 v(\theta)).$$

3.7 Cramer-Rao Bound and Asymptotic Efficiency

Chapter 4

Bayesian Inference

So far, we have modelled the parameter θ is a fixed but unknown value. The crux of Bayesian inference is to model the unknown parameter itself as a random variable. The RV Θ has a probability distribution, $f_\Theta(\theta)$ called the **prior distribution**. The parametric model describing the conditional distribution of X given Θ we now write as $f_{X|\Theta}(x \mid \theta)$, where previously it was $f(x \mid \theta)$.

This defines a joint probability distribution over Θ and X, as

$$f_{X,\Theta}(x, \theta) = f_{X|\Theta}(x|\theta) f_\Theta(\theta).$$

Thus the marginal distribution of X is

$$f_X(x) = \int f_{X,\Theta}(x, \theta) d\theta = \int f_{X|\Theta}(x|\theta) f_\Theta(\theta) d\theta.$$

Bayesian inference is based on the conditional distribution of Θ given X, that is

$$f_{\Theta|X}(\theta|x) = \frac{f_{X,\Theta}(x, \theta)}{f_X(x)} = \frac{f_{X|\Theta}(x|\theta) f_\Theta(\theta)}{f_X(x)}.$$

This is called the **posterior distribution** of Θ. This is summarized as

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

Note that the $1/f_X(x)$ factor does not depend on Θ.

Example 4.0.1. Consider a $X_1, \ldots, X_n \overset{\text{IID}}{\sim} \text{Bernoulli}(p)$ model with a uniform prior. Then we have that

$$f_{P|X}(p|x_1, \ldots, x_n) \propto f_{X|P}(x_1, \ldots, x_n|p) f_P(p) = p^s (1-p)^{n-s}.$$

The only possible choice for the normalization constant is $B(s+1, n-s+1)$, so we know that

$$f_{P|X}(p|x_1, \ldots, x_n) = \frac{1}{B(s+1, n-s+1)} p^s (1-p)^{n-s}.$$

It's very simple to show that if $f_P(p) = \text{Beta}(\alpha, \beta)$, then

$$f_{P|X}(p|x_1, \ldots, x_n) \sim \text{Beta}(s + \alpha + 1, n + \beta - s).$$