Compact course notes

Stat 241, Winter 2011

Statistics

Professor: M. Zhu
transcribed by: J. Lazovskis
University of Waterloo
April 11, 2011

Contents

1 Model review **2**

2 Point estimation **2**
2.1 General estimations 2

3 Distribution theory **3**
3.1 Mean-squared error 3
3.2 Asymptotic approach 3

4 Hypothesis testing **4**
4.1 Interval estimation 4
4.2 Hypothesis testing 4
4.3 The $2\log(\Lambda)$ transformation 5
4.4 t-tests 5

1 Model review

Model 1.0.1. Bernoulli(p)
· Binary outcome

$$\begin{aligned} f(x) &= p^x(1-p)^{1-x} \\ E(x) &= p \\ \text{Var}(x) &= p(1-p) \end{aligned}$$

Model 1.0.2. Binomial(n, p)
· n independent Bernoulli trials with x successes

$$\begin{aligned} f(x) &= \binom{n}{x} p^x(1-p)^{1-x} \\ E(x) &= np \\ \text{Var}(x) &= np(1-p) \end{aligned}$$

Model 1.0.3. Poisson(λ)
· Limit of binomial model as $n \to \infty$

$$\begin{aligned} f(x) &= \frac{e^{-\lambda}\lambda^x}{x!} \\ E(x) &= \lambda \\ \text{Var}(x) &= \lambda \end{aligned}$$

Model 1.0.4. Exponential(λ)
· Continuous random variable

$$\begin{aligned} f(x) &= \lambda e^{-\lambda x} \\ E(x) &= 1/\lambda \\ \text{Var}(x) &= 1/\lambda^2 \end{aligned}$$

Model 1.0.5. Multinomial($n; p_1, p_2, \ldots, p_k$)
· Generalization of binomial

Constraints: $p_1 + p_2 + \cdots + p_k = 1$
$X_1 + X_2 + \cdots + X_k = n$

Joint probability function:

$$f(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1!x_2!\ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

Model 1.0.6. Normal(μ, σ^2)

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\} \\ E(x) &= \mu \\ \text{Var}(x) &= \sigma^2 \end{aligned}$$

2 Point estimation

2.1 General estimations

Definition 2.1.1. Given $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$, the likelihood function is $L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$.

The log likelihood is $\ell(\theta) = \log(L(\theta))$.

Definition 2.1.2. The maximum likelihood estimation is $\tilde{\theta} = \max_{\text{all } \theta}\{L(\theta)\} = \max_{\text{all } \theta}\{\ell(\theta)\}$.

Definition 2.1.3. Given $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$ and $\theta \in \mathbb{R}^d$, the kth moment is given by

$$E(X^k) = \int_{\text{all } x} x^k f(x;\theta)dx = g_k(\theta) = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

This is termed the method of moments.

Remark 2.1.4. In general, MLE is superior to MOM.

Definition 2.1.5. For the MLE method, the maximum of the function f may be estimated, given an initial guess θ_0, by:

$$\text{Newton's method: } \theta_i = \theta_{i-1} - \frac{\ell'(\theta_{i-1})}{\ell''(\theta_{i-1})} \qquad \text{Fisher scoring: } \theta_i = \theta_{i-1} - \frac{\ell'(\theta_{i-1})}{E(\ell''(\theta_{i-1}))}$$

3 Distribution theory

Definition 3.0.1. The quantity $\tilde{\theta}$ is an estimator for the quantity θ. In essence, both are random variables.

Remark 3.0.2. The distribution of $\tilde{\theta} = g(X_1, X_2, \ldots, X_n)$ can be found through:
1. Simulation (performing the experiment)
2. n-dimensional integration
3. the moment generating function method
4. the asymptotic method (CLT)

3.1 Mean-squared error

Definition 3.1.1. If $E(\tilde{\theta}) = \theta$, then $\tilde{\theta}$ is an unbiased estimator for θ. Otherwise, define the bias to be

$$\text{Bias}(\tilde{\theta}) = E[\tilde{\theta}] - \theta$$

Definition 3.1.2. If $\tilde{\theta}_1$ and $\tilde{\theta}_2$ are both unbiased for θ, then $\tilde{\theta}_1$ is said to be more efficient if $\text{Var}(\tilde{\theta}_1) < V(\tilde{\theta}_2)$.

Definition 3.1.3. The mean-squared error of $\tilde{\theta}$ is defined to be $E[(\tilde{\theta} - \theta)^2]$, abbreviated MSE. It is used to evaluate the distribution of the estimator $\tilde{\theta}$.

Theorem 3.1.4. $\text{MSE}(\tilde{\theta}) = \text{Bias}^2(\tilde{\theta}) + \text{Var}(\tilde{\theta})$

Theorem 3.1.5. [James-Stein]
Let $X \sim N(\theta, \sigma^2 I \in M_{n\times n})$ for $x, \theta \in \mathbb{R}^n$. For $n \geqslant 3$, $\text{MSE}(\tilde{\theta}_{JS}) < \text{MSE}(\tilde{\theta}_{MLE})$, where

$$\tilde{\theta}_{MLE} = x \qquad \text{and} \qquad \tilde{\theta}_{JS} = \left(1 - \frac{n - 2\sigma^2}{||x||^2}\right) x$$

The above is termed biased estimation.

Theorem 3.1.6. For $Y = X_1 + X_2 + \cdots + X_n$ with $X_i \perp X_j$ for all $i \neq j$, $\text{mgf}_y(t) = \prod_{i=1}^{n} \text{mgf}_{x_i}(t)$.

Theorem 3.1.7. For $Y = aX + b$, the moment generating function of Y is $\text{mgf}_y(t) = e^{bt}\text{mgf}_x(at)$.

3.2 Asymptotic approach

Definition 3.2.1. If $\text{mgf}_{x_n}(t) \to \text{mgf}_x(t)$ as $n \to \infty$, then $X_n \xrightarrow{D} X$.
This is termed convergence in distribution.

Definition 3.2.2. If for every $\epsilon > 0$, $P(|x_n - c| > \epsilon) \to 0$ as $n \to \infty$, then $X_n \xrightarrow{P} c$.
This is termed convergence in probability.

Theorem 3.2.3. [Central Limit Theorem]
Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ some distribution. For $E(X_i) = \mu$ and $V(X_i) = \sigma^2 < \infty$ and $S_n = \sum_{i=1}^{n} X_i$,

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{D} N(0,1) \qquad \text{or} \qquad \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} N(0,1) \text{ for } \bar{X}_n = \tfrac{S_n}{n}$$

Remark 3.2.4. If $\text{mgf}_{X_n}(t) \to \text{mgf}_X(t)$ as $n \to \infty$, then $X_n \xrightarrow{D} X$.

Remark 3.2.5. If $F_n(x) \to F(x)$ as $n \to \infty$, then $X_n \xrightarrow{D} X$.
Here F_n is the cumulative distribution function for X_n, and F is the cdf for X.

Theorem 3.2.6. [Chebyshev's inequality]
Let X be a random variable and $\epsilon > 0$ such that $E(X) = \mu$ and $\text{Var}(X) = \sigma^2 < \infty$. Then

$$P(|x - \mu| < \epsilon) \leqslant \frac{\sigma^2}{\epsilon^2}$$

Theorem 3.2.7. [Weak Law of Large Numbers]
Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ some distribution with $E(X_i) = \mu$ and $V(X_i) = \sigma^2 < \infty$ and $\bar{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$. Then

$$\bar{X}_n \xrightarrow{P} \mu$$

Remark 3.2.8. For the special case where $\mu = 0$ and $\sigma^2 = 1$,

CLT: $\sqrt{n}\,\bar{X}_n \xrightarrow{D} N(0,1)$
WLLN: $\bar{X}_n \xrightarrow{P} 0$

4 Hypothesis testing

4.1 Interval estimation

This is the canonical setting:
- Assume that σ^2 is known.
- We know also that $\tilde{\mu}_{MLE} = \tilde{\mu}_{MOM} = \bar{x}$.

We want to find bounds L, U such that $P(L < \mu < U) = 1 - \alpha$ for α small.

Definition 4.1.1. A quantity that depends upon the given data and a single unknown parameter with known distribution (and no other unknowns) is termed a pivotal quantity.

Definition 4.1.2. Wrt the above constants, define the threshold value C_α so that $(\mu - C_\alpha, \mu + C_\alpha) = (L, U)$.

Definition 4.1.3. The interval $(\mu - C_\alpha, \mu + C_\alpha)$ is termed the confidence interval.

Theorem 4.1.4. Let $\tilde{\theta}_n$ be the MLE of θ based on n iid observations. Under certain regularity conditions,

$$\sqrt{nI(\theta)}\,(\tilde{\theta}_n - \theta) \xrightarrow{D} N(0,1) \quad \text{for} \quad I(\theta) = E\left[\frac{d^2}{d\theta^2}\log(f(x_i, \theta))\right]$$

where $I(\theta)$ is the Fisher information.

Remark 4.1.5. In the above, $I(\theta)$ may be replaced with $I(\tilde{\theta}_n)$.

Corollary 4.1.6. Above, $\sqrt{nI(\theta)}\,(\tilde{\theta}_n - \theta)$ is an approximate pivotal quantity.

Corollary 4.1.7. Let X be a random variable with $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$. Then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} N(0,1)$$

4.2 Hypothesis testing

Hypothesis testing involves the evaluation of the null hypothesis H_0 against the alternative hypothesis H_A.

The null hypothesis predicts that there is no relation between observed variables, and the alternative hypothesis either denies the null hypothesis, or predicts a specific relationship between observed variables.

Definition 4.2.1. Define the following terms:

type I error	: reject H_0 when H_0 is true
type II error	: accept H_0 when H_0 is false
significance level	: P(type I error)
power	: $1 - P$(type II error)

Definition 4.2.2. The likelihood ratio test compares the value of a certain parameter under the hypotheses:

$$\Lambda = \frac{L(\theta_A)}{L(\theta_0)}$$

Then the LRT rejects H_0 if $\Lambda > C = C_\alpha$ such that $P(\text{type I error}) = \alpha$.

Theorem 4.2.3. [NEYMAN-PEARSON LEMMA]
Among all tests with a significance level of α, the LRT has the highest power.

The canonical setting used below is $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$.

Remark 4.2.4. The LRT in the canonical case rejects H_0 if $\bar{x} > \mu_0 + C_\alpha \frac{\sigma}{\sqrt{n}}$.
Further, the power in the canonical case is $1 - \Phi\left(C_\alpha - \frac{\mu_A - \mu_0}{\sigma/\sqrt{n}}\right)$ for $\Phi(\cdot)$ the cdf of $N(0,1)$.

Definition 4.2.5. The generalized likelihood ratio test assumes that $\tilde{\mu}_{MLE}$ is the MLE if $\mu \neq \mu_0$.

$$\Lambda = \frac{L(\tilde{\mu}_{MLE})}{L(\mu_0)}$$

Remark 4.2.6. GLRT rejects H_0 is Λ is too large $\iff$ $2\log(\Lambda)$ is too large.

4.3 The $2\log(\Lambda)$ transformation

Theorem 4.3.1. Under regularity conditions, for Λ a GLRT statistic and $df = \dim(H_A) - \dim(H_0)$,

$$2\log(\Lambda) \xrightarrow{D} \chi^2_{(df)}$$

Note that Λ depends upon a sample of size n, and the above holds as $n \to \infty$.

Remark 4.3.2. The above theorem allows us to find C_α using the asymptotic distribution.

Definition 4.3.3. If $Z_1, Z_2, \ldots, Z_n \overset{iid}{\sim} N(0,1)$, then $\sum_{i=1}^{n} Z_i^2 \sim \chi^2_{(n)}$. This is the chi-squared distribution.

Remark 4.3.4. For $N(\mu, \sigma^2)$ with σ^2 known, $2\log(\Lambda) = \left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2_{(1)}$

Definition 4.3.5. The rejection region is the complement of the confidence interval.
Rejection region : $\{\theta \mid 2\ell(\tilde{\theta}_{MLE}) - 2\ell(\theta_0) > C_\alpha\}$
Confidence set : $\{\theta \mid 2\ell(\tilde{\theta}_{MLE}) - 2\ell(\theta_0) \leqslant C_\alpha\}$

The above has demonstrated that $2\log(\Lambda) = 2\ell(\tilde{\theta}_{MLE}) - 2\ell(\theta_0)$ is an approximate pivotal quantity.

4.4 t-tests

Using the t-test for $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, neither μ nor σ^2 are known. In this case σ^2 is termed a nuisance parameter - it is an unknown, but we do not want to say anything about it. It will be replaced by an estimate.

With respect to the hypotheses, $H_0 : \mu = \mu_0$ and $H_A : \mu \neq \mu_0$.

Remark 4.4.1. Replace $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ with $T = \frac{\bar{x}-\mu}{S/\sqrt{n}} \sim t_{(n-1)}$ which is a pivotal quantity for $S^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$.

Definition 4.4.2. Let $U \sim N(0,1)$, $V \sim \chi^2_{(n)}$ with $U \perp V$. Then $\dfrac{U}{\sqrt{V/n}} \sim t_{(n)}$ has the <u>t-distribution</u>.

Theorem 4.4.3. Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Then

1. $\frac{\sum(x_i - \bar{x})^2}{\sigma^2} \sim \chi^2_{(n-1)}$

2. $\bar{x} \perp \sum(x_i - \bar{x})^2$

Corollary 4.4.4. Then S is an unbiased estimator for σ^2, whereas $\tilde{\sigma}^2_{MLE} = \frac{1}{n}\sum(x_i - \bar{x})^2$ is slightly biased.

Definition 4.4.5. The <u>p-value</u> is defined to be $P(\Lambda \geqslant \Lambda_{observed} \,\big|\, H_0)$.