Section for Applied Statistics and Data Analysis

TA: Cong Mu

Office Hour: Wednesday 10:00AM - 12:00PM

September 6, 2019

Overview

1. Some Statistics
   - Bivariate Normal Distribution
   - Gamma Distribution
   - Chi-squared Distribution

2. Some Programming
   - Introduction to R
   - Initial Example of Data Analysis

- **Multivariate Normal Distribution**

$$f_{\boldsymbol{X}}(x_1, \cdots, x_k) = \frac{\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}}$$

- **Bivariate Normal Distribution**

$$f_{X,Y}(x, y) = \frac{\exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X \sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right)}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}$$

Marginal Distribution of Bivariate Normal

Claim

The marginal distributions of bivariate normal $\mathcal{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ are normal with

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \frac{1}{\sigma_X\sqrt{2\pi}} \exp\left[-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right]$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \frac{1}{\sigma_Y\sqrt{2\pi}} \exp\left[-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}\right]$$

Conditional Distribution of Bivariate Normal

Claim

The conditional distributions of bivariate normal $\mathcal{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ are normal with

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{1}{\sigma_{X|Y}\sqrt{2\pi}} \exp\left[-\frac{(x-\mu_{X|Y})^2}{2\sigma_{X|Y}^2}\right]$$

where

$$\mu_{X|Y} = \mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y-\mu_Y) \qquad \text{and} \qquad \sigma_{X|Y}^2 = (1-\rho^2)\sigma_X^2$$

Gamma Distribution

- **Gamma Function**

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx, \quad \Gamma(n) = (n-1)!, \quad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

- **Gamma Distribution:** $X \sim Gamma(\alpha, \lambda)$ where $\alpha > 0, \lambda > 0$

$$f(x) = \frac{\lambda e^{-\lambda x} (\lambda x)^{\alpha-1}}{\Gamma(\alpha)} \qquad x \geqslant 0$$

with

$$\mathbb{E}(X) = \frac{\alpha}{\lambda} \qquad \text{and} \qquad Var(X) = \frac{\alpha}{\lambda^2}$$

Chi-squared Distribution

- **Definition**
  $X \sim \chi^2_n$ if $X = \sum_{i=1}^n Z_i^2$ where $Z_1, \cdots, Z_n$ are independent, standard normal random variables
- **Chi-squared Distribution**

$$f(x) = \frac{e^{-x/2} x^{n/2-1}}{2^{n/2} \Gamma(n/2)}$$

with

$$\mathbb{E}(X) = n \qquad \text{and} \qquad Var(X) = 2n$$

Normal Distribution v.s. Chi-squared Distribution

Claim

If Z is standard normal, then Z^2 has a chi-squared distribution with 1 degree of freedom, i.e.

$$Z \sim \mathcal{N}(0,1) \qquad \Longrightarrow \qquad Z^2 \sim \chi_1^2$$

Proof Hint

$$\mathbb{P}(Z^2 < y) = \mathbb{P}(-\sqrt{y} < Z < \sqrt{y})$$

Gamma Distribution v.s. Chi-squared Distribution

- **Gamma Distribution:** $X \sim \mathrm{Gamma}(\alpha, \lambda)$ where $\alpha > 0, \lambda > 0$

$$f(x) = \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} \qquad x \geqslant 0$$

- **Chi-squared Distribution:** $X \sim \chi^2_n$

$$f(x) = \frac{e^{-x/2} x^{n/2-1}}{2^{n/2}\Gamma(n/2)}$$

- Chi-squared distribution is a special case of the gamma distribution with

$$\alpha = \frac{n}{2} \qquad \text{and} \qquad \lambda = \frac{1}{2}$$

- **R:** https://www.r-project.org/
- **RStudio:** https://www.rstudio.com/
- **Blackboard:** R videos, R worksheets
- **More in the future**

- **Linear Models with R**
  By Julian Faraway, 2004. Available in JHU Library.
- **R Markdown**
  https://rmarkdown.rstudio.com/

Data

- **Background**
  The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix.
- **Variables**
  - number of times pregnant
  - plasma glucose concentration at 2 hours in an oral glucose tolerance test
  - diastolic blood pressure (mmHg)
  - triceps skin fold thickness (mm)
  - 2-hour serum insulin (mu U/ml)
  - body mass index (weight in kg/(height in m2))
  - diabetes pedigree function
  - age (years)
  - a test whether the patient showed signs of diabetes (coded zero if negative, one if positive)

Summary

- **Formulate the Problem**
  - Background
  - Objective
  - What the client wants
  - Turn the problem into statistical terms
- **Learn Your Data**
  - Outliers
  - Missing value
  - Re-code: quantitative v.s. categorical
  - Visualization: **ggplot2**
    https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf

Thanks for listening!