Data analysis

Ben Graham

MA930, University of Warwick

October 13, 2015

Ch4: Joint and Marginal Distributions

Def 4.1.1 An n-dimensional random vector is a function $X = (X_i)_{i=1}^n$ from sample space S to $\mathbb{R}^n$.

- $n = 2$; roll two dice
- `rnorm(10)`
- `rbinom(10,5,0.5)`

Discrete case: Joint p.m.f. $f_X(x_1, \ldots, x_n)$:

$$\mathbb{P}[X \in A] = \sum_{x=(x_1,\ldots,x_n)\in A} f_X(x)$$

Continuous case: Joint p.d.f. $f_X(x_1, \ldots, x_n)$:

$$\mathbb{P}[X \in A] = \int_{x=(x_1,\ldots,x_n)\in A} f_X(x)dx_1 \ldots dx_n$$

Marginal distributions

Discrete case: pmf $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$.

$$f_X(x) = \sum_{y : f_{X,Y}(x,y) > 0} f_{X,Y}(x, y)$$

Continuous case: pdf $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$.

$$f_X(x) = \int f_{X,Y}(x, y) dy$$

Example

Discrete

- $X, Y \in \{1, \ldots, 6\}$ independent dice rolls
- $Z = X + Y$
- p.m.f. $f_{X,Y}$
- p.m.f. $f_{X,Z}$

Example

Continuous

- $X, Y \sim N(0,1)$ i.i.d.r.v
- $Z = \rho X + \sqrt{1-\rho^2}\, Y \sim N(0,1)$
- $f_{X,Y}(x,y) = f_X(x) f_Y(y)$
- $f_{X,Z}(x,z) = f_X(x) f_{Z|X}(z \mid x)$ $[(Z \mid X) \sim N(\rho X, 1-\rho^2)]$ bivariate normal distribution

4.2 Conditional Distributions and Independence

Random variables X and Y are independent if

- for all x, y, $\{X < x\}$ and $\{Y < y\}$ are independent
- $f_{X,Y}(x, y) = f_X(x) f_Y(y)$ [continuous p.d.f.s or discrete p.m.f.s]
- $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$ [characteristic functions]

Examples

- $X, Y \sim Bernoulli(1/2)$
- $X, Y \sim N(0, 1)$

Independence $\rightarrow$Covariance=0

- $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$

Covariance=0 $\not\rightarrow$independence

- $X \sim N(0, 1)$
- $Y \in \{-1, +1\}$ independent of X
- Cov(X,XY)=0

Sums of Normal distributions

Example 4.3.4.

- $X \sim N(\mu_X, \sigma_X^2)$
- $Y \sim N(\mu_Y, \sigma_Y^2)$
- X, Y independent
- Then $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

Random sample

- $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$
- Then $\sum_i X_i \sim N(n\mu, n\sigma^2)$
- $$\bar{X} = \sum_i \frac{X_i}{n} \sim N(\mu, \sigma^2/n)$$
- $$Z = \sum_i \frac{X_i - \mu}{\sigma\sqrt{n}} \sim N(0, 1)$$

Sums of Poissons

Theorem 4.3.2

- $X \sim$Poisson(θ)
- $Y \sim$Poisson(λ)
- X, Y independent.
- Then $X + Y \sim \text{Poisson}(\theta + \lambda)$

Ex 4.4.1 Conversely

- $Y \sim$Poisson(λ)
- $X \mid Y \sim$Bin(Y, p)
- Then $X \sim$Poisson(λp)

Random Samples

- $X_1, \ldots, X_n \sim$Poisson(λ) i.i.d.r.v.
- $\sum_i X_i \sim$Poisson$(n\lambda) \approx N(n\lambda, n\lambda)$
- $\bar{X} \sim\approx N(\lambda, \lambda/n)$

Covariance and correlation

Def 4.5.1 Covariance:

$$Cov(X, Y) = E[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}[XY] - \mathbb{E}X\mathbb{E}Y$$

Def 4.5.2 Correlation

$$\rho_{XY} = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

Thm 4.5.6 If X and Y are r.v. and $a, b \in \mathbb{R}$, then

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$$

Special case — X, Y independent.

Def 4.5.10 Bivariate distribution

- $\mu_X, \mu_Y \in \mathbb{R}$
- $\sigma_x, \sigma_Y > 0$
- pdf

$$\begin{aligned} f_{X,Y}(x,y) &= (2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2})^{-1} \\ &\times \exp\left(\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_x}{\sigma_X}\right)^2\right.\right. \\ &\left.\left. -2\rho\left(\frac{x-\mu_x}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)+\left(\frac{Y-\mu_Y}{\sigma_Y}\right)^2\right)\right) \end{aligned}$$

- Or pdf

$$f(x) = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}}\exp\left(-(x-\mu)\Sigma^{-1}(x-\mu)\right)$$

$$k = 2,\ x \in \mathbb{R}^k,\ \mu = \begin{pmatrix}\mu_x \\ \mu_Y\end{pmatrix},\ \Sigma = \begin{pmatrix}\sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{pmatrix}$$

Def 4.6.2 Multinomial distribution

- m trials / repeat events
- n possible outcomes, probabilities $p_1, \ldots, p_n$ sum to one.
- Let $x_i \in \{0, 1, \ldots, n\}$ count the number of type-i outcomes
- joint pmf

$$f(x_1, \ldots, x_n) = \frac{m!}{x_1! \ldots x_n!} p_1^{x_1} \ldots p_n^{x_n} \qquad \text{where } \sum_i x_i = m$$

- Negative correlations $Cov(X_i, X_j) = -mp_i p_j \ (i \neq j)$
- Modeling tables of categorical variables.

Ch5 Random samples

- $X_1, \ldots, X_n$ is called a random sample of size n from population $f(x)$ if they are independent, identically distributed random variables (i.i.d.r.v.) with marginal distribution function $f(x)$.
- Think of them as being a random sample from a population that is much larger than n. The probability distribution represents the true distribution of values in the larger population.
- Condorcet's Jury Principle:
  `plot(sapply(seq(0,1,0.01),function(p)pbinom(6,12,p)))`
- Getting representative samples can be hard, i.e. telephone surveys.
  - Does everyone have a landline?
  - Does everyone choose to talk to strangers phoning you up during dinner?
  - Are people shy about expressing some preferences?
  - http://www.bbc.co.uk/news/uk-politics-33228669

Parameters

- Parameter θ controlling the distribution $f(x \mid \theta)$
- Frequentist statistics:
  - θ is fixed but unknown
  - Choose $\hat{\theta} = \hat{\theta}(data)$ to estimate θ.
- Bayesian:
  - Joint distribution $f(\theta)f(x \mid \theta)$
  - $f(\theta)$ is the prior distribution
- Example.
  - Exponential $f(x \mid \theta) = \theta \exp(-\theta x)$.
  - Joint distribution $f(x \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) = \theta^n \exp(-\theta \sum_i x_i)$
  - N.B. $f(x \mid \theta)$only depends on the x_i via their sum.

Statistics

Def 5.2.1

- Let $X_1, \ldots, X_n \sim f(x \mid \theta)$
- Let $T(x_1, \ldots, x_n)$ denote some function of the data, i.e. $T : \mathbb{R}^n \to \mathbb{R}$.
- T is a statistic
  - $T(x_1, \ldots, x_n) = x_1$
  - $T(x_1, \ldots, x_n) = \max_i x_i$
  - mean, median, etc.
- θ,$\mathbb{E}X$, $\mathrm{Var}(X)$, etc, are not statistics.
- The probability distribution of T is called the sampling distribution of T.

Common statistics

- $X_1, \ldots, X_n$ i.i.d.r.v.
- Sample mean

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \sum_{i=1}^{n} X_i$$

- Sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{1}{n-1} \left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right]$$

- If the mean $\mathbb{E}X_i$ and variance $\mathrm{Var}(X_i)$ exist, then these are *unbiased* estimates.

*Def 5.4.1 Order Statistics

- The order statistics of a random sample $X_1, \ldots, X_n$ are the values placed into increasing order: $X_{(1)}, \ldots, X_{(n)}$.
  - $X_{(1)} = \min_i X_i$
  - $X_{(2)}$ =second smallest
  - ...
  - $X_{(n)} = \max_i X_i$
- The sample range is $R = X_{(n)} - X_{(1)}$
- The sample median is

$$M = \begin{cases} X_{((n+1)/2)} & n \text{ odd} \\ \frac{1}{2}X_{(n/2)} + \frac{1}{2}X_{(n/2+1)} & n \text{ even} \end{cases}$$

The median may give a better sense of what is typical than the mean.

Quantiles

- For $p \in [0,1]$, the p quantile is (R, method 7 of 9)

$$(1-\gamma)x_{(j)}+\gamma x_{(j+1)}, \quad (n-1)p < j \leq (n-1)p+1, \quad \gamma = np+1-p-j$$

- Trimmed/truncated mean
  - $p \in [0,1]$
  - Remove the smallest and biggest np items from the sample
  - Take the mean of what is left
  - i.e. LIBOR, 18 banks, top and bottom 4 removed
  - Cauchy location parameter

Theorem 5.4.4 Distribution of the order statistics

- Sample of size n with c.d.f. F_X and pdf f_X
- Binomial distribution:

$$F_{X_{(j)}}(x) = \sum_{k=j}^{n} \binom{n}{k} [F_X(x)]^k [1 - F_X(x)]^{n-k}$$

- Differentiate:

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f_X(x) [F_X(x)]^{j-1} [1 - F_X(x)]^{n-j}.$$

More sample mean

- Characteristic function

$$\phi_{\bar{X}}(t) = [\phi_X(t/n)]^n \approx \left[1 + \frac{it\mathbb{E}X}{n} + o(t/n)\right]^n$$

- Thm 5.3.1: If $X_i \sim N(\mu, \sigma^2)$, then
  - $\bar{X}$ and S^2 are independent
  - $\bar{X} \sim N(\mu, \sigma^2/n)$.
  - $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$
- Proof: For simplicity, assume $\mu = 0$ and $\sigma^2 = 1$.

Ingredients for the proof

- If $e_1, \ldots, e_n$ form an orthonormal basis for $\mathbb{R}^n$. Then $(e_i \cdot X)_{i=1}^n$ are i.i.d.r.v $N(0,1)$.
- Let $e_1 = (\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})$
- The $\Gamma(\alpha, \beta)$ distribution is defined

$$f(x) = C(\alpha, \beta) x^{\alpha - 1} \exp(-\beta x).$$

- The χ^2_k distribution is the $\Gamma(k/2, 1/2)$ distribution.
- If $X \sim N(0,1)$, then $X^2 \sim \chi^2_1$
- The sum of k independent χ^2_1 r.v. is χ^2_k.

Derived distributions

- If $Z \sim N(0,1)$ and $V \sim \chi^2_k$ then $Z/\sqrt{V/k} \sim t_k$ (Student's t)
- $$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$
- If $U \sim \chi^2_k$ and $V \sim \chi^2_\ell$ are independent, then $\frac{U/k}{V/\ell} \sim F_{k,\ell}$
- Suppose $X_1, \ldots, X_m \sim N(\mu_X, \sigma^2_X)$ and $Y_1, \ldots, Y_n \sim N(\mu_Y, \sigma^2_Y)$. Then

$$\frac{S^2_X/\sigma^2_X}{S^2_Y/\sigma^2_Y} \sim F_{m-1,n-1}$$

- These distributions are also used for linear regression/ANOVA.