Probability Theory

Lecture 2

Mattias Villani

Division of Statistics
Dept. of Computer and Information Science
Linköping University

Overview Lecture 2

- **Conditional distributions**
- **Conditional expectation, conditional variance**
- **Distributions with random parameters and the Bayesian approach**
- **Regression and Prediction**

Conditional distributions

- For events [if $P(B) > 0$]

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- A and B are **independent** if and only if $P(A|B) = P(A)$.
- For **discrete** random variables

$$p_{Y|X=x}(y) = p(Y = y|X = x) = \frac{p_{X,Y}(x,y)}{p_X(x)}$$

$$p_{Y|X=x}(y) = \frac{p_{X,Y}(x,y)}{\sum_y p_{X,Y}(x,y)}.$$

- For **continuous** random variables

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{\int_{-\infty}^{\infty} f_{X,Y}(x,z)dz}$$

Conditional expectation

- **Conditional expectation of Y given $X = x$ is**

$$E(Y|X = x) = \begin{cases} \sum_y y \cdot p_{Y|X=x}(y) & \text{if } Y \text{is discrete} \\ \int_{-\infty}^{\infty} y \cdot f_{Y|X=x}(y)dy & \text{if } Y \text{is continuous} \end{cases}$$

- Note that $h(X) = E(Y|X)$ is a random variable that only depends on X.
- Theorem 2.1. **Law of iterated expectation**.

$$E\left[E(Y|X)\right] = E(Y)$$

- Note that the *inner expectation* $(E(Y|X))$ is with respect to $f_{Y|X}(y)$, while the *outer expectation* is with respect to $f_X(x)$. [Ex. 2.1, Page 33]
- The law of iterated expectation is an "expectation version" of the law of total probability.
- $E(Y|X) = E(Y)$ if X and Y are independent.

Conditional variance

- **Conditional variance of** Y **given** $X = x$ is

$$Var(Y|X = x) = E\left[(Y - E(Y|X = x))^2 \,|X = x\right]$$

- Note that $v(X) = Var(Y|X)$ is a random variable that only depends on X.
- Corollary 2.3.1

$$Var(Y) = E\left[Var(Y|X)\right] + Var\left[E(Y|X)\right]$$

- Note the naive version $Var(Y) = E\left[Var(Y|X)\right]$ misses the uncertainty in Y that comes from not knowing X in $E(Y|X)$.[Ex. 2.1, Page 33]
- See the more general version in Theorem 2.3.

Distributions with random parameters

- $X|\theta \sim f_X(x;\theta)$ and θ is a random variable.
- Example 1:
  - $X|N = n \sim Bin(n,p)$ and $N \sim Po(\lambda)$.
  - If the number of potential bidders in an auction is $N = n$ and each of them bids with probability p, then $X \sim Bin(n,p)$ bids will be placed.
  - The number of potential bidders is uncertain, $N \sim Po(\lambda)$.
  - The marginal distribution for X is $Po(\lambda \cdot p)$[Ex. 3.2, Page 40]

Distributions with random parameters

- $X|\theta \sim f_X(x;\theta)$ and θ is a random variable.
- Example 1:
  - $X|N = n \sim Bin(n,p)$ and $N \sim Po(\lambda)$.
  - If the number of potential bidders in an auction is $N = n$ and each of them bids with probability p, then $X \sim Bin(n,p)$ bids will be placed.
  - The number of potential bidders is uncertain, $N \sim Po(\lambda)$.
  - The marginal distribution for X is $Po(\lambda \cdot p)$[Ex. 3.2, Page 40]
- Example 2:
  - $X|(\sigma^2 = 1/\lambda) \sim N(0, 1/\lambda)$ and $\lambda \sim \Gamma\left(\frac{n}{2}, \frac{2}{n}\right)$, then $X \sim t(n)$.
  - X is daily stock market returns. $X|\lambda \sim N(0, 1/\lambda)$, where $1/\lambda$ is the daily variance.
  - The daily variance varies from day to day according to $\lambda \sim \Gamma\left(\frac{n}{2}, \frac{2}{n}\right)$. Turbulent day: realization of λ is very small.

Bayesian coin tossing

- X_n=number of heads after n tosses.

$$X_n|P = p \sim Bin(n,p)$$

- **Prior distribution**: $P \sim U(0,1)$.
- **Posterior distribution**: $P|(X_n = k) \sim Beta(k+1, n+1-k)$.
- Marginal of X_n

$$X_n \sim U(\{1, 2, ..., n\})$$

- Conditional of X_{n+1} given X_n **and** p

$$P(X_{n+1} = n+1|X_n = n, p) = p$$

- Conditional of X_{n+1} given X_n

$$P(X_{n+1} = n+1|X_n = n) = \frac{n+1}{n+2} \to 1 \text{ as } n \to \infty$$

- Coin flips are no longer independent when p is uncertain and we learn about p from data.

Regression and Prediction

- The **regression function**

$$h(\mathbf{x}) = h(x_1, ..., x_n) = E(Y|X_1 = x_1, ..., X_n = x_n) = E(Y|\mathbf{X} = \mathbf{x})$$

- **Predictor**: $\hat{Y} = d(\mathbf{X})$.
- **Linear predictor** $d(\mathbf{X}) = a_0 + a_1X_1 + ...a_nX_n$.
- Expected **quadratic prediction error**: $E\left[Y - d(\mathbf{X})\right]^2$
- The **best predictor** of Y [minimizes expected quadratic prediction error] is the regression function $E(Y|\mathbf{X} = \mathbf{x})$.
- Best **linear** predictor - **least squares**:

$$\hat{Y} = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(X - \mu_x)$$

- When (X, Y) is jointly normal, $E(Y|X = x)$ is linear. Linear is best of all.