ST2132 Probability Review II

Semester 1 2022/2023

If printing, do DOUBLE-SIDED, each side TWO slides.

Outline

- Prediction, mean square error, random conditional expectation
- Normal and related distributions

Empirical prediction

Niels Bohr: *Prediction is very difficult, especially if it's about the future.*

- What will be the temperature measured at Changi Airport tomorrow at noon? No formula will give you an answer, but you can try any method to predict it, then compare with the measurement.
- In many prediction problems, the quantity of interest is assumed to be random. Its distribution is typically unknown, and has to be guessed, or estimated from relevant data.
- Statistical inference plays a crucial role in prediction, even if the randomness assumption is invalid.

Predicting an RV

A statistician uses a constant c to predict a random variable Y. To quantify success, she uses the **mean square error** (MSE):

$$\mathrm{E}\{(Y-c)^2\}$$

- Show that the MSE is $\mathrm{var}(Y) + \{\mathrm{E}(Y) - c\}^2$.

- Which c minimises MSE, and what is this MSE value?

Prediction with data

Y and X are correlated, and a realisation x has been observed. To predict Y, it is better to choose c that depends on x.

What is the relevant distribution for the MSE?

Show that the MSE is

$$\text{var}[Y|x] + \{\text{E}[Y|x] - c\}^2$$

Deduce the c that minimises MSE, and the minimum MSE. Is the answer intuitive?

Conclusion on prediction (1)

- Without observing X, the best predictor of Y is $\mathrm{E}(Y)$.
  The MSE is __________.
- Given realisation x, the best predictor of Y is $\mathrm{E}[Y|x]$.
  The MSE is __________.

  If we use $\mathrm{E}(Y)$ instead of $\mathrm{E}[Y|x]$, how much larger will our MSE be?
- If success not measured by MSE, the best predictor may not be a conditional expectation.

Conclusion on prediction (2)

- To make actual predictions, $\mathrm{E}[Y|x]$ has to be estimated from data. The simplest form is $\mathrm{E}[Y|x] = ax + b$ where a, b are constants, estimated by simple regression.
- Often x is a vector of realisations from covariates. Then a multiple regression is natural: $\mathrm{E}[Y|x]$ is a linear combination of the covariates; the coefficients estimated from data.
- Our MSE relies on the distribution being right. In practice, better to use the sum of squared errors across all predictions.

Various MSE formulae

We encountered:

$$
\begin{aligned}
\mathrm{E}\{(Y-c)^2\} &= \mathrm{var}(Y) + \{\mathrm{E}(Y) - c\}^2 \\
\mathrm{E}[(Y-c)^2|x] &= \mathrm{var}[Y|x] + \{\mathrm{E}[Y|x] - c\}^2 \\
\mathrm{E}[\{Y-\mathrm{E}(Y)\}^2|x] &= \mathrm{var}[Y|x] + \{\mathrm{E}[Y|x] - \mathrm{E}(Y)\}^2
\end{aligned}
$$

Instead of memorising these equations, understand their purpose and how they relate to the deceptively powerful (*):

$$\mathrm{var}(X) = \mathrm{E}(X^2) - \mu^2$$

Random conditional expectations

A box has N identical tickets: K marked '1' and $N - K$ mark '0', where $0 < K < N$.

X_1, X_2: results of two random draws without replacement.

- $\mathrm{E}[X_2|X_1]$ is a random variable, which takes value $\mathrm{E}[X_2|x_1]$ with probability $\Pr(X_1 = x_1)$. Show that

$$\mathrm{E}(\mathrm{E}[X_2|X_1]) = \mathrm{E}(X_2)$$

- $\mathrm{var}[X_2|X_1]$ is similarly defined. Show that

$$\mathrm{var}(\mathrm{E}[X_2|X_1]) + \mathrm{E}(\mathrm{var}[X_2|X_1]) = \mathrm{var}(X_2)$$

- These formulae are generally valid.

Mean MSE

In practice, Y is predicted from multiple realisations $x_1, \ldots, x_n$.
Suppose for any x, we use $\mathrm{E}[Y|x]$.
The overall success is quantified by the mean MSE:

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{var}[Y|x_i] \approx \underline{\hspace{3cm}}$$

How much smaller is the mean MSE compared to the MSE of $\mathrm{E}(Y)$, as $n \to \infty$?

"$\mathrm{E}[Y|X]$ is the best predictor of Y using X" is best thoughts of as a shorthand: prediction is done using realisations of X, not X itself.

Some symbolic traps

- Sometimes $\mathrm{E}[Y|x]$ is written as

$$\mathrm{E}[Y|X = x]$$

  True or false: $\mathrm{E}[Y|X] = \mathrm{E}[Y|X = X]$?

- For a continuous X, the cumulative density function (CDF) is

$$F(x) = \Pr(X \leq x)$$

  True or false: $F(X) = \Pr(X \leq X)$?

Standard normal distribution

- The standard normal random variable Z has density function

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}, \qquad -\infty < z < \infty$$

$\mathrm{E}(Z) = 0$, $\mathrm{var}(Z) = 1$. What about $\mathrm{E}(Z^2)$, $\mathrm{E}(Z^3)$, etc?

- Express, in terms of f:

$$\Phi(x) = \Pr(Z \leq x) = \underline{\hspace{6cm}}$$

What does the graph of the CDF $\Phi : \mathbb{R} \to (0, 1)$ look like?

Other normal distributions

- Let μ and σ be real numbers, with $\sigma > 0$. Define

  $$Y = \mu + \sigma Z$$

  Calculate $\mathrm{E}(Y)$ and $\mathrm{var}(Y)$. The distribution of Y is denoted $\mathrm{N}(\mu, \sigma^2)$.

- Express the CDF of Y in terms of Φ, then find Y's density.

- Standardising a random variable means to subtract the expectation then divide by the SD. What is the distribution of the standardised Y?

χ^2_1 distribution

- Let $Z \sim \mathrm{N}(0,1)$. $V = Z^2$ has a χ^2 distribution with 1 degree of freedom.
- Write down $\Pr(V \leq v)$ in terms of Φ.
- Show that the density of V is

$$\frac{1}{\sqrt{2\pi}} v^{-1/2} e^{-v/2}, \qquad v > 0$$

Gamma distribution

- Let $\alpha > 0$ and $\lambda > 0$. The Gamma(α, λ) density is

$$\frac{\lambda^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t}, \qquad t \geq 0$$

  where $\Gamma(\alpha)$ is a number that makes the density integrate to 1.
- Can you express $\Gamma(\alpha)$ as an integral that depends only on α?

- Does χ_1^2 have a Gamma distribution?

χ^2_n distribution

- Let $V_1, \ldots, V_n$ be IID χ^2_1.

$$V = \sum_{i=1}^{n} V_i$$

has a χ^2 distribution with n degrees of freedom.

- Calculate $\mathrm{E}(V)$ and $\mathrm{var}(V)$. What can you say about V/n as $n \to \infty$?

- From the density of χ^2_n, can you tell if it has a Gamma distribution?

t distribution

- Let $Z \sim \mathrm{N}(0,1)$ and $V \sim \chi^2_n$ be independent.

  $$t_n = \frac{Z}{\sqrt{V/n}}$$

  has a t distribution with n degrees of freedom.

- The t distribution is symmetric around 0. What can you say about t_n as $n \to \infty$?

F distribution

- Let $V \sim \chi^2_m$ and $W \sim \chi^2_n$ be independent.

$$F_{m,n} = \frac{V/m}{W/n}$$

  has an F distribution with (m, n) degrees of freedom.

- If $X \sim t_n$, what is the distribution of X^2?

- What can you say about $F_{m,n}$ as $n \to \infty$?

IID random variables

Let $X_1, \ldots, X_n$ be IID RVs, with mean $\bar{X}$.

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

is known as the sample variance.

This term is standard, though it seems more reasonable to call the following the sample variance:

$$\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

$\bar{X}$ and S^2 in the normal case

Let $X_1, \ldots, X_n$ be IID $\mathrm{N}(\mu, \sigma^2)$.

(A) $\bar{X}$ and S^2 are independent.

(B)

$$\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

(C)

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

(D)

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Justification (1)

- (B) is true because summing and scaling normal RVs gives a normal RV.
- (D) follows from the rest.
- Rice proves (A) and (C). The best argument is via a change of coordinates, in the general context of multiple regression (see the classic *The Analysis of Variance* by Scheffé.).

  These are beyond our scope, but we can get some intuitive results.

Justification (2)

- For any i, $\bar{X}$ and $X_i - \bar{X}$ are uncorrelated, hence independent by normality.

  In fact, $\bar{X}$ and $(X_1 - \bar{X}, \ldots, X_n - \bar{X})$ are independent, which implies (A).

- From Tutorial 1, we have

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 + n \left(\frac{\bar{X} - \mu}{\sigma} \right)^2$$

  The left side $\sim \chi_n^2$, the rightmost term $\sim \chi_1^2$, which makes (C) plausible.

Conclusions

This lecture finishes the review of probability.

- General concepts: prediction, mean square error, random conditional expectation.
- Distributions: normal, χ^2, t, F, independence of $\bar{X}$ and S^2 under normality.