UNIVERSITY OF LINCOLN
SCHOOL OF MATHEMATICS AND PHYSICS

MTH1005 PROBABILITY AND STATISTICS TUTORIAL WEEK 7-8

Review

The joint probability density function $f_{XY}(x,y)$ of a continuous random variable (X,Y) is a function whose integral in the set of possible values for (X, Y) is the rectangle $D = (x,y): a \le x \le b, u \le y \le w$. gives the likelihood of the subset in the sample space of (X,Y):

$$P\{a \le X \le b; u \le Y \le w\} = \int_a^b \int_u^w f_{XY}(x,y)dxdy$$

To be a valid probability density function of a random variable, we must have

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)dxdy = 1 \tag{0.1}$$

MEASURES OF CENTRAL TENDENCY

- **Mean.** For a discrete data set $\{x_i \in X, i = 1,\dots N\}$, the arithmetic mean, also called the mathematical expectation or average, is the central value of the numbers (x_i) in the set: specifically, the sum of the values divided by the number of values (N)

$$\mu = \frac{1}{N}\sum_{i=0}^{N} x_i.$$

  The **expectation value or mean** of a continuous random variable Y by

$$E[Y] = \mu_Y = \int_{-\infty}^{\infty} y f_Y(y) dy$$

- **Median.** The median is the value separating the higher half from the lower half of a data sample. For a discrete data set, it corresponds to the "middle" value for a set containing an odd number of data or the average of the two "middle values for a set with an even number of data. The basic advantage of the median in describing data compared to the mean is that it is not skewed so much by extremely large or small values, and so it may give a better idea of a "typical" value.

- **Mode.** The mode of a set of data values is the value that appears most often. It is the value x at which its probability mass function takes its maximum value. In other words, it is the value that is most likely to be sampled.

- **Quantiles.** They are cut points dividing the range of a probability distribution into continuous intervals with equal probabilities or dividing the observations in a sample in the same way. The 4-quantiles are called **quartiles, Q**. The difference between upper (Q_3) and lower (Q_1) quartiles is also called the interquartile range, $\mathbf{IQR = Q_3 - Q_1}$ The n-quartile of a continuous variable X can be calculated from the cumulative function as

$$F_Y(Q_n) = \int_0^{Q_n} f(x) dx = \frac{n}{4}$$

- The **variance of the random variable X** is defined by

$$\sigma_X^2 = Var(X) = E[(X - \mu)^2]$$

  and it can be also written as

$$\sigma_X^2 = Var(X) = E[X^2] - (E[X])^2.$$

  with the second central moment of a continuous random variable given by

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) dx$$

  It is the expected squared distance of a value from the center of the distribution.

- **The Chebyshev inequality.** This powerful relation guarantees that, for a wide class of probability distributions, no more than a certain fraction of values can be more than a certain distance from the mean. It is defined as

$$P\big(|T - \mathrm{E}[T]| \geq k\big) \leq \frac{\sigma_T^2}{k^2}$$

where ET is the expectation value, σ_T^2 is the variance of the random variable T. The relation tell us that no more than $1/k^2$ of the distribution's values can be more than k standard deviations away from the mean.

PROBABILITY DISTRIBUTIONS

The probability density function for a **uniform distribution** is defined as

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

and the probability between $\alpha < X < \beta$ is given by

$$P(\alpha < X < \beta) = \int_\alpha^\beta f_X(x)dx = \frac{\beta - \alpha}{b - a}.$$

Question 1

A random variable Z has a probability density function

$$f_Z(z) = \begin{cases} 0.1e^{-0.1z} & z \geq 0\,, \\ 0 & z < 0 \end{cases}$$

- find the mean of Z
- find the median of Z
- find the interquartile distance of Z

Question 2

The mean and variance of a random variable T with probability density function

$$f_T(t) = \begin{cases} \frac{2}{15} - \frac{2t}{225} & 0 \leq \mathrm{t} \leq 15\,, \\ 0 & \text{otherwise} \end{cases}$$

are $E[T] = 5$ and $\sigma_T^2 = 12.5$, respectively. Use the Chebyshev inequality

$$P\big(|T - \mathrm{E}[T]| \geq k\big) \leq \frac{\sigma_T^2}{k^2}$$

to estimate the probability that - $2.5 \leq T \leq 7.5$ - $0 \leq T \leq 10$ - compare these estimates to the real probabilities.

Question 3

Buses arrive at a bus stop at 15-minute intervals starting at 7am(!). The buses are always perfectly on time, so arrive at 7:00, 7:15, 7:30, 7:45 ...
If a passenger arrives at the stop at a time that is uniformly distributed between 7 and 7:30, find the probability that she waits

- less than 5 minutes for a bus;
- at least 12 minutes for a bus.

Question 4

Suppose the joint pdf of (X, Y) is given by

$$f(x,y) = \begin{cases} \frac{6}{5}(x+y^2) & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

- Verify that this is a legitimate pdf.
- The probability $P(0 \le x \le \frac{1}{4}, 0 \le y \le \frac{1}{4})$.