Statistics ST2334 Topic 5: Probability Densities (part a)

2011/2012 Semester 2

Continuous Random Variables

- Continuous random variable: a random variable that can assume any values within an interval or collection of intervals.
- Examples: heights; length; speed of a car; the amount of alcohol in a person's blood; the tensile strength of a new alloy.
- Probability mass function is not applicable: If X is a continuous random variable, then for any specific real value x, $P(X = x) = 0$.
- Insetad, we characterize the probability distribution of a continuous random variable, X, by $P(a \leq X \leq b)$ for any interval (a, b).

Probability Density Function

- Mathematically, for continuous random variable, X,

$$P(a \leq X \leq b) = \int_a^b f(x)dx.$$

- Since

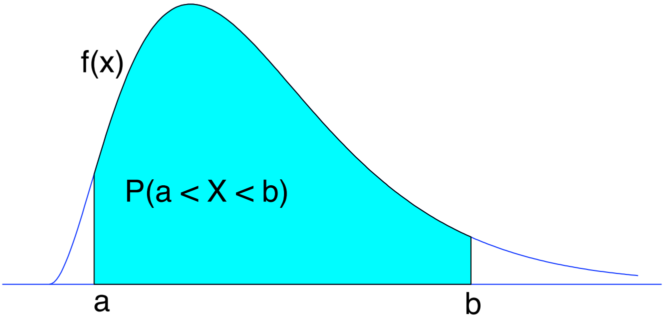
$$P(X = a) = P(X = b) = 0,$$

we also have

$$P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = \int_a^b f(x)dx.$$

Probability Density Function

- $f(x)$ above is called the probability density function (pdf, or density function). It is a continuous function on the interval or collection of intervals, on which X assumes values.



Properties of $f(x)$

- A probability density function $f(x)$ must satisfy the following requirements:
  1. $f(x) \geq 0$ for all x
  2. The total area under the curve is 1. i.e.

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

- Conversely, any function $f(x)$ that satisfies the above two conditions simultaneously can be a density function of a random variable.

Example of Probability Density Function

- Find k so that $f(x)$ is a probability density function:

$$f(x) = \begin{cases} kx^2(1-x), & \text{for } 0 < x < 1 \\ 0, & \text{elsewhere} \end{cases}$$

- Check that $f(x) \geq 0$.
- Now, we need to make sure it integrates to 1.

$$\int_{-\infty}^{\infty} f(x)dx = \int_0^1 kx^2(1-x)dx = k\left[\frac{x^3}{3} - \frac{x^4}{4}\right]_0^1 = \frac{k}{12}$$

so $k = 12$.

Example: Computing Probabilities from pdf

- A random variable has the probability density

$$f(x) = \begin{cases} \frac{1}{2}e^{-\frac{x}{2}}, & \text{for } x > 0 \\ 0, & \text{for } x \leq 0 \end{cases}$$

- Find the probabilities that it will take on a value
  (a) between 3 and 4.5 ;
  (b) greater than 2 .
- Evaluating the necessary integrals, we get
  (a) $\int_3^{4.5} \frac{1}{2}e^{-\frac{x}{2}}\,dx = -e^{-2.25} - \left(-e^{-1.5}\right) \approx 0.118$
  (b) $\int_2^{\infty} \frac{1}{2}e^{-\frac{x}{2}}\,dx = 0 - \left(-e^{-1}\right) \approx 0.368$

CDF for Continuous Random Variables

- $F(x) = P(X \leq x)$ is the Cumulative Distribution Function for random variable X, i.e.

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

- Therefore, we have

$$f(x) = \frac{d}{dx}F(x).$$

- If X has cdf $F(x)$, then

$$P(a \leq X \leq b) = F(b) - F(a)$$

Properties of $F(x)$

- $F(x)$ must satisfy the following conditions:
  1. $F(x)$ is a non-decreasing function of x.
  2. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$
- If any function $F(x)$ satisfies the above two conditions simultaneously, then it can be a cdf of a random variable.
- Note that the two conditions corresponds to the ones for probability density function $f(x)$.

Example continued

- For the above X from the previous example, its cdf is

$$
\begin{aligned}
F(x) &= P(X \leq x) = \int_{-\infty}^{x} f(t)dt \\
&= \begin{cases} \int_0^x \frac{1}{2} e^{-\frac{t}{2}} dt, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0 \end{cases} \\
&= \begin{cases} 1 - e^{-\frac{x}{2}}, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0 \end{cases}
\end{aligned}
$$

Expectation

- Definition: for X has pdf $f(x)$, then its expectation is defined as

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

- and more generally, for any function $g(\cdot)$,

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) \cdot f(x) dx$$

- Compare with discrete random variables.

Example continued

- Again, looking at the previous example, let μ and σ^2 be the mean and variance of X. Then,

$$\begin{aligned}
\mu = E(X) &= \int_{-\infty}^{\infty} xf(x)dx \\
&= \int_{0}^{\infty} x \cdot \frac{1}{2} e^{-\frac{x}{2}} dx \\
&= \left[x \cdot (-e^{-\frac{x}{2}})\right]_0^{\infty} - \int_0^{\infty} (-e^{-\frac{x}{2}})dx \\
&= (0-0) - \left[2e^{\frac{x}{2}}\right]_0^{\infty} \\
&= -(0-2) = 2
\end{aligned}$$

Variance

- Just as before,

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

- and can also be written as

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

Example continued continued

- Taking the second approach with our example,

$$\begin{aligned}
E(X^2) &= \int_{-\infty}^{\infty} x^2 f(x) dx \\
&= \int_{0}^{\infty} x^2 \cdot \frac{1}{2} e^{-\frac{x}{2}} dx \\
&= \left[x^2 \cdot (-e^{-\frac{x}{2}}) \right]_0^{\infty} - \int_0^{\infty} (-e^{-\frac{x}{2}}) \cdot (2x) dx \\
&= 4 \int_0^{\infty} \frac{1}{2} e^{-\frac{x}{2}} dx \\
&= 4(2) = 8 \\
\Rightarrow Var(X) &= E(X^2) - [E(X)]^2 = 8 - 2^2 = 4
\end{aligned}$$

Normal Distribution

- The most important distribution in real application – normal distribution.
- A random variable, X, is normally distributed, if its density function is

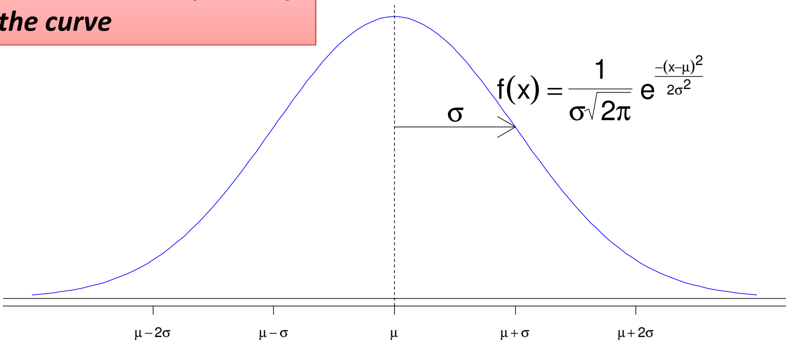
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

for some fixed μ, σ^2.

- The graph of the above density function is bell shaped (refer to the next slide).

Normal Density Function

e = 2.718 and π = 3.142;
μ and σ ($\sigma > 0$) are the parameters that affects the center and spread of the curve



CDF of Normal Distribution

- The cdf of normal distribution is

$$P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

- No closed form for the cdf of normal distribution. Numerical method is required to evaluate $P(X \leq x)$ for some specific x.

Normal Distribution

- If a random variable, X, follows a normal distribution with density function,

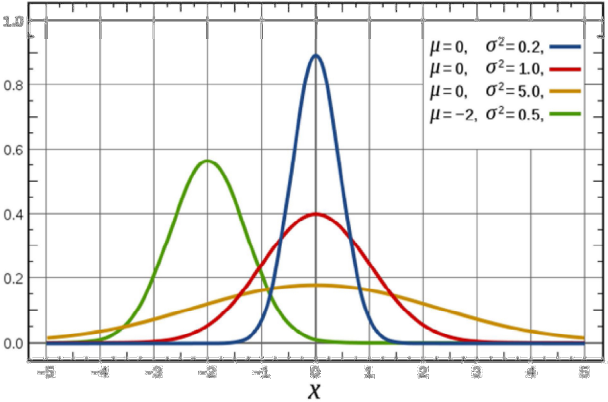
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

then

$$E(X) = \mu \qquad \& \qquad Var(X) = \sigma^2$$

- We use $X \sim N(\mu, \sigma^2)$ to denote a random variable, X, following normal distribution with mean μ , variance σ^2.

Normal Distribution



Standard Normal Distribution

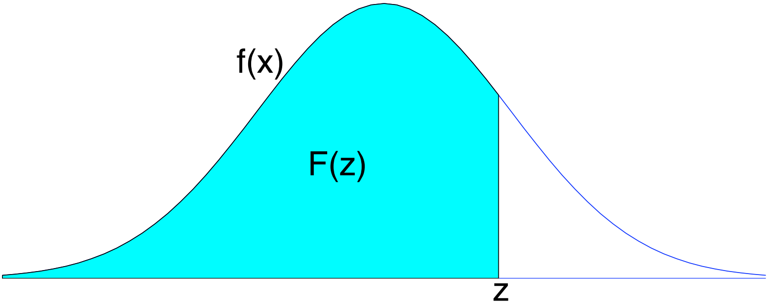
- When $\mu = 0$ & $\sigma^2 = 1$, the resulting normal distribution is called standard normal distribution.
- A random variable that has a standard normal distribution is denoted as Z, whose cdf is given by

$$F(z) = P(Z \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

  This cdf is commonly denoted by $\Phi(z)$.
- $Z \sim N(0,1)$

Standard Normal Distribution

- The value of $\Phi(z)$ is only numerically accessible, and can be read directly from Table 3 at the end of the textbook.



Normal Table

- Note that because the distribution of Z is symmetric about 0, we have
$$P(Z < -z) = P(Z > z)$$
so
$$\Phi(z) = 1 - \Phi(-z)$$
and if $z > 0$
$$P(|Z| > z) = 2\Phi(-z)$$

Examples

- $P(Z > 1.35) =$

- $P(-2.33 < Z \leq -1.12) =$

- $P(|Z| < 2) =$

We can go the other way too

Find z such that

1. $P(Z > z) = 0.05$

   $\Rightarrow \Phi(z) = P(Z < z) = 0.95$

   Thus we have $z \approx 1.645$

2. $P(|Z| \leq z) = 0.5$

   $P(|Z| \leq z) = 1 - (P(Z > z) + P(Z < -z)) = 1 - 2\Phi(-z) = 0.5$

   Thus $\Phi(-z) = 0.25$, we have $z \approx 0.67$

Normalizing Normal Random Variables

- If $X \sim N(\mu, \sigma^2)$, let

$$Z = \frac{X - \mu}{\sigma}$$

  then $Z \sim N(0, 1)$.

- This implies that

$$\begin{aligned} P(X \leq x) &= P(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}) \\ &= P(Z \leq \frac{x - \mu}{\sigma}) \\ &= \Phi(\frac{x - \mu}{\sigma}) \end{aligned}$$

One table is enough!

- When X has the normal distribution with mean μ and standard deviation σ then

$$P(a < X \leq b) = \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})$$

  Similarly,

$$P(X > a) = 1 - \Phi(\frac{a-\mu}{\sigma})$$

  and

$$P(X \leq b) = \Phi(\frac{b-\mu}{\sigma})$$

- We can reference the standard normal table for any normal random variable!

Example: Motorcycle accidents

- Let X be the number of Motorcycle accidents per month in Singapore. Assume approximately

$$X \sim N(7.4,\ 9)$$

Find the probability of at least 5 Motorcycle accidents in any one month

$$\begin{aligned} P(X \geq 5) &= 1 - P(X < 5) \\ &\approx 1 - \Phi\left(\frac{4.5 - 7.4}{\sqrt{9}}\right) \\ &= 1 - \Phi(-0.97) = 1 - 0.1660 = 0.834 \end{aligned}$$

- Why 4.5?

Continuity Correction

- When approximating a discrete random variable by a continuous random variable like the normal distribution, we need to "spread" its values over a continuous scale.
  - It is in an approximation in the interval sense. This makes sense when the interval is large.
  - What about small range, or even say $P(X = 5)$?
- We do this by representing each integer k by the interval from $k - \frac{1}{2}$ to $k + \frac{1}{2}$.
- In the previous example, 5 is represented by 4.5 to 5.5. So to approximate $P(X < 5)$, we exclude 4.5 to 5.5 and look the interval to the left of 4.5.

Example: Height

Suppose the height of a randomly selected individual has mean 170 and standard deviation 5. Find

(a) the probability that the height of the randomly selected individual is between 165 to 180.

(b) the quartiles of the distribution of height.

Normal Approximation to the Binomial Distribution

- For a binomial random variable, X, when n is large, then,

$$X \approx N(np, np(1-p))$$

or

$$P(X \leq x) \approx \Phi(\frac{x - np}{\sqrt{np(1-p)}})$$

This approximation is reasonably accurate when np and n(1-p) are both greater than 15.

Example: Barry Bonds

Barry Bonds has a career on-base percentage of 0.444. Suppose that is an accurate estimate of his probability of getting on base when he's at bat. Also, assume that each at bat is independent.

What is the probability of Barry Bonds getting on base at least 20 times out of 50 at bats?

- $X \sim B(50, 0.444)$, can we use normal approximation?
- Check rule of thumb: np $=$ 50(0.444) $=$ 22.2 $>$15, n(1-p)=27.8$>$15.
- Since $np = 50(0.444) = 22.2, np(1-p) \approx 12.34$, X is can be approximated by $N(22.2, 12.34)$.

Example: Barry Bonds

- Using the normal approximation,

$$\begin{aligned} P(X \geq 20) &\approx 1 - P(X < 20) \\ &= 1 - \Phi(\frac{19.5 - 22.2}{\sqrt{12.34}}) \\ &= 1 - \Phi(-0.77) = 0.7794 \end{aligned}$$

- The exact probability is

$$P(X \geq 20) = \binom{50}{20} p^{20}(1-p)^{30} + \ldots + \binom{50}{50} p^{50}(1-p)^{0} = 0.7781$$

Poisson Approximation vs Normal Approximation

- Possion approximation requires n large and p small. Recall this is a result of taking letting $n \to \infty$ while keeping np fixed.
- Normal approximation as well requires n large. It is a specific case of a very powerful theorem, the central limit theorem, which we will cover in Chapter 6. Here, we think of p as fixed and let $n \to \infty$.