FIN 3380

Tutorial 3
Brief Review of Probability Theory

dongliyuan@cuhk.edu.cn
zhaidaojing@cuhk.edu.cn
09/23/2019

Outline

- Discrete random variable
  - probability density function, cumulative distribution function
  - expectation, variance, s.d.
  - conditional expectation, conditional distribution
- Continuous random variable
- Distribution
  - the normal distribution
  - the Chi-Squared distribution
  - the student t distribution

Discrete random variable

- Discrete: Within a range of numbers, discrete variables can take on only certain values.
- Example: dice roll

```
sample(1:6, 1)
```

Outcome	1	2	3	4	5	6
Probability	1/6	1/6	1/6	1/6	1/6	1/6
Cumulative Probability	1/6	2/6	3/6	4/6	5/6	1

Discrete random variable

FIN 3380

- Discrete: Within a range of numbers, discrete variables can take on only certain values.
- Example: dice roll
- Probability density function
  - the list of all possible values of the variable and their probabilities which sum to 1.
- Cumulative probability distribution function:
  - gives the probability that the random variable is less than or equal to a particular value.

```
sample(1:6, 1)
```

```
# generate the vector of probabilities
probability <- rep(1/6, 6)

# generate the vector of cumulative probabilities
cum_probability <- cumsum(probability)
```

Discrete random variable

- Expected value: the long-run average value of its outcomes when the number of repeated trials is large.

Suppose the random variable Y takes on k possible values, $y_1, \dots, y_k$, where y_1 denotes the first value, y_2 denotes the second value, and so forth, and that the probability that Y takes on y_1 is p_1, the probability that Y takes on y_2 is p_2 and so forth. The expected value of Y, $E(Y)$ is defined as

$$E(Y) = y_1 p_1 + y_2 p_2 + \cdots + y_k p_k = \sum_{i=1}^{k} y_i p_i$$

where the notation $\sum_{i=1}^{k} y_i p_i$ means "the sum of y_i p_i for i running from 1 to k". The expected value of Y is also called the mean of Y or the expectation of Y and is denoted by μ_Y.

```r
# compute the sample mean of 10000 dice rolls
mean(sample(1:6,
            10000,
            replace = T))
```

Discrete random variable

- Variance and standard deviation

The variance of the discrete random variable Y, denoted σ_Y^2, is

$$\sigma_Y^2 = \mathrm{Var}(Y) = E\left[(Y - \mu_y)^2\right] = \sum_{i=1}^{k}(y_i - \mu_y)^2 p_i$$

The standard deviation of Y is σ_Y, the square root of the variance. The units of the standard deviation are the same as the units of Y.

- Sample variance: an estimator of the population variance

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Discrete random variable

FIN 3380

- Conditional expectation

$$E(Y|X=x_i)=\sum_{y_i} y_i P(Y=y_j|X=x_i)=\sum_{y_i} y_i \frac{P(X=x_i, Y=y_j)}{P(X=x_i)}=\sum_{y_i} y_i \frac{p_{ij}}{p_i}, j=1,2,\ldots$$

- Conditional distribution

$$F(y \mid x_i)=\sum_{y_j \leq y} P\left(Y=y_j \mid X=x_i\right)=\sum_{y_j \leq y} p_{j|i}.$$

- example

	Dance	Sports	TV	Total
Men	0.04	0.20	0.16	0.40
Women	0.32	0.12	0.16	0.60
Total	0.36	0.32	0.32	1.00

Continuous random variable

- Probability density function and cumulative distribution function

Let $f_Y(y)$ denote the probability density function of Y. The probability that Y falls between a and b where $a < b$ is

$$P(a \leq Y \leq b) = \int_a^b f_Y(y) \mathrm{d}y.$$

We further have that $P(-\infty \leq Y \leq \infty) = 1$ and therefore $\int_{-\infty}^{\infty} f_Y(y) \mathrm{d}y = 1$.

Continuous random variable

- Expected value and variance

As for the discrete case, the expected value of Y is the probability weighted average of its values. Due to continuity, we use integrals instead of sums. The expected value of Y is defined as

$$E(Y) = \mu_Y = \int y f_Y(y) \mathrm{d}y.$$

The variance is the expected value of $(Y - \mu_Y)^2$. We thus have

$$\mathrm{Var}(Y) = \sigma_Y^2 = \int (y - \mu_Y)^2 f_Y(y) \mathrm{d}y.$$

Continuous random variable

- Expected value and variance – implementation in R

Consider the continuous random variable X with probability density function:

$$f_X(x) = \frac{3}{x^4}, x > 1.$$

```
# define functions
f <- function(x) 3 / x^4
g <- function(x) x * f(x)
h <- function(x) x^2 * f(x)

# compute area under the density curve
area <- integrate(f,
                  lower = 1,
                  upper = Inf)$value
area

# compute E(X)
EX <- integrate(g,
                lower = 1,
                upper = Inf)$value
EX

# compute Var(X)
VarX <- integrate(h,
                  lower = 1,
                  upper = Inf)$value - EX^2
VarX
```

Continuous random variable

- Conditional expectation

$$E[X|Y=y] = \sum_x x\, f_{X|Y}(x|y)$$

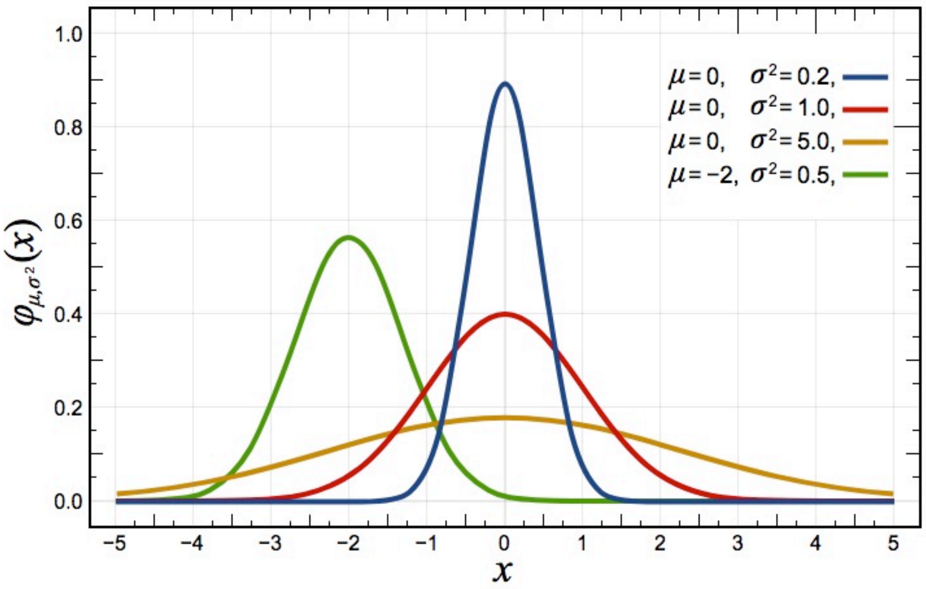
- Conditional distribution

$$F(x \mid y) = \int_{-\infty}^{x} \frac{f(u,y)}{f_Y(y)} du, f(x \mid y) = \frac{f(x,y)}{f_Y(y)}.$$

Distribution

- Normal distribution
  - symmetric and bell-shaped.
  - characterized by its mean μ and standard deviation σ, concisely expressed by $N(\mu,\sigma^2)$.
  - The normal distribution has the PDF:

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

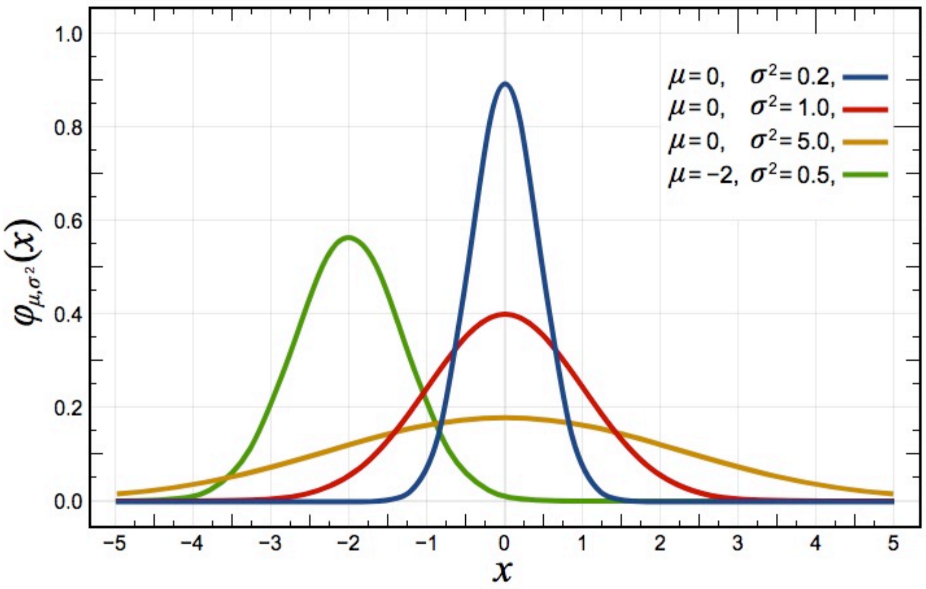


Distribution

- Skewness and kurtosis

$$Skewness = \int (y - \mu_Y)^3 f_Y(y)dy$$

$$Kurtosis = \int (y - \mu_Y)^4 f_Y(y)dy$$



Distribution

- Standardized normal distribution
  - Standard normal distribution *μ=0* and *σ=1*.
  - Standard normal variable is denoted by *Z*.
  - the standard normal PDF is denoted by φ
  - the standard normal CDF is denoted by Φ

$$\phi(c) = \Phi'(c) \ , \ \Phi(c) = P(Z \leq c) \ , \ Z \sim \mathcal{N}(0,1).$$

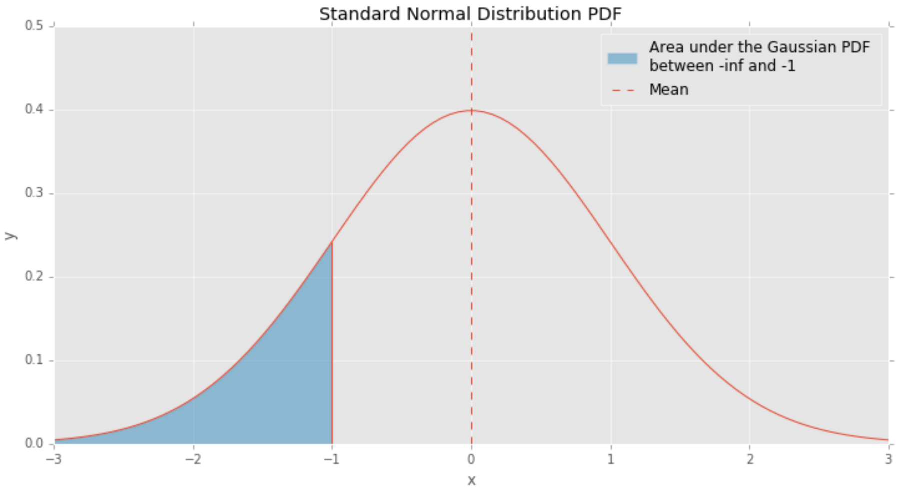
```r
# draw a plot of the N(0,1) PDF
curve(dnorm(x),
      xlim = c(-3.5, 3.5),
      ylab = "Density",
      main = "Standard Normal Density Function")

# compute denstiy at x=-1.96, x=0 and x=1.96
dnorm(x = c(-1.96, 0, 1.96))


# plot the standard normal CDF
curve(pnorm(x),
      xlim = c(-3.5, 3.5),
      ylab = "Density",
      main = "Standard Normal Cumulative Distribution Function")


# compute the probability using pnorm()
pnorm(1.337)
```

Distribution

- Standardized normal distribution
  - Standard normal distribution $\mu=0$ and $\sigma=1$.
  - Standard normal variable is denoted by Z.
  - the standard normal PDF is denoted by φ
  - the standard normal CDF is denoted by Φ

$$\phi(c) = \Phi'(c)\ ,\ \Phi(c) = P(Z \leq c)\ ,\ Z \sim \mathcal{N}(0,1).$$

Q: How to calculate $P(-1.96 \leq Z \leq 1.96)$



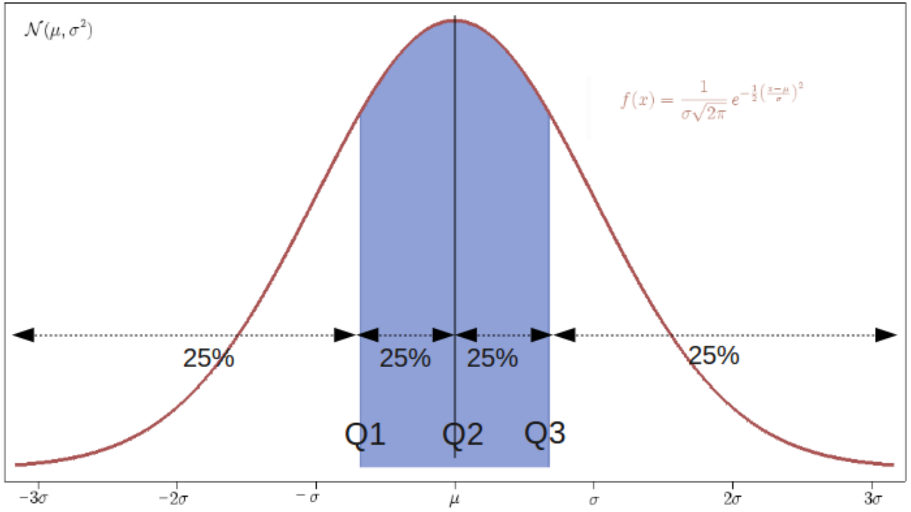
Distribution

- Standardized normal distribution
  - Standard normal distribution $\mu=0$ and $\sigma=1$.
  - Standard normal variable is denoted by Z.
  - the standard normal PDF is denoted by φ
  - the standard normal CDF is denoted by Φ
- Quantile

$$P(X \le Z_{\alpha}) = \alpha$$

$$P(X \le Z_{0.025}) = 0.025$$

Q: How to calculate $P(-1.96 \le Z \le 1.96)$



Distribution

- Standardize normal distribution

Suppose Y is normally distributed with mean μ and variance σ^2:

$$Y \sim \mathcal{N}(\mu, \sigma^2)$$

Then Y is standardized by subtracting its mean and dividing by its standard deviation:

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

Q: If we have a random variable *Y* with *Y~N(5, 25)*,
what is $P(3 \le Y \le 4)$?

Distribution

- Standardize normal distribution

Suppose Y is normally distributed with mean μ and variance σ^2:

$$Y \sim \mathcal{N}(\mu, \sigma^2)$$

Then Y is standardized by subtracting its mean and dividing by its standard deviation:

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

Q: If we have a random variable *Y* with *Y~N(5, 25),*
what is $P(3 \leq Y \leq 4)$?

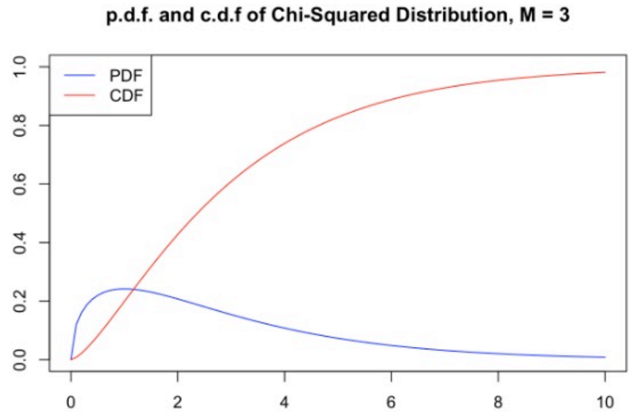
```
pnorm(4, mean = 5, sd = 5) - pnorm(3, mean = 5, sd = 5)
```

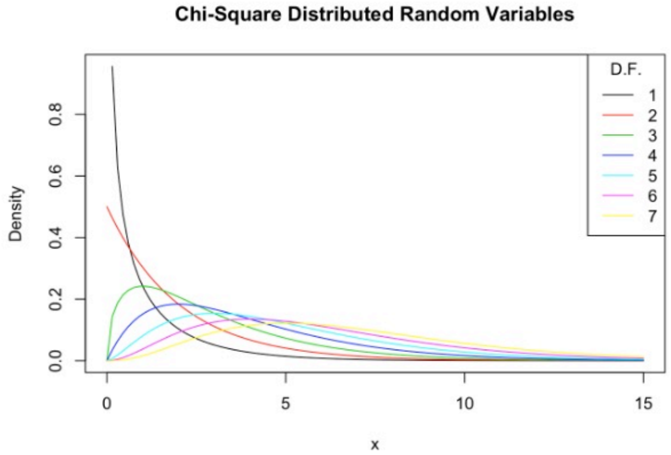
Distribution

FIN 3380

- The Chi-squared distribution
  - The sum of M squared independent standard normal distributed random variables
  - Freedom degrees M

$$Z_1^2 + \cdots + Z_M^2 = \sum_{m=1}^{M} Z_m^2 \sim \chi_M^2 \text{ with } Z_m \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$$





Distribution

- The t-student distribution
  - Z: a standard normal variate
  - W: χ^2_M random variable
  - Z and W are independent
  - Then X follow

$$\frac{Z}{\sqrt{W/M}} =: X \sim t_M$$

- Characteristic
  - symmetric, bell-shaped
  - look similar to a normal distribution, especially when M is large (M is larger than 30).