Multivariate distributions

Li Li

Department of Mathematics and Statistics

Coronary heart patient data

- A random sample of N = 200 coronary heart disease patients had their blood pressure (BP) and serum cholesterol (SC) levels measured resulting in the following data summary:

Table: 2×2 table.

		SC		
		Below 240 mg/dL	Above 240 mg/dL	Total
BP	Below 120/80 mm Hg	0.115	0.13	245
	Above120/80 mm Hg	0.41	0.345	0.755
	Total	0.525	0.475	1

Coronary heart patient data

- What is the percentage of patients that have SC below 240 mg/dL and BP below 120/80 mm Hg? 0.115 = 11.5%
- What is the percentage of patients that have SC below 240 mg/dL? 0.525 = 52.5%
- Are the outcome of SC and the outcome of BP independent? (We do not consider sampling variability here)

whether SC and BP are independent is equivalent to whether

$$P(SC < 240 \text{ and } BP < 120/80)$$

$$= P(SC < 240)\, P(BP < 120/80)$$

Since there are not equal, SC and BP are not independent

2×2 table

- Empirical percentage is a estimate of probability.
- Consider joint probability mass function in the following 2×2 table:

Table: 2×2 table.

		X_2		
		a_1	a_2	Total
X_1	b_1	p_{11}	p_{12}	p_{1+}
	b_2	p_{21}	p_{22}	p_{2+}
	Total	p_{+1}	p_{+2}	1

- Interpret p_{11} as $P(X_1 = b_1$ and $X_2 = a_1)$. In short, we write it as $P(X_1 = b_1, X_2 = a_1)$.
- Interpret p_{12} as $P(X_1 = b_1, X_2 = a_2)$.
- Interpret p_{1+} as $P(X_1 = b_1)$.
- Interpret p_{+1} as $P(X_2 = a_1)$.

Coronary heart patient data

- Hypothesis of particular interest: whether X_1 and X_2 are independently distributed.
- It is testing whether $P(X_1 = b_j, X_2 = a_k) = P(X_1 = b_j)P(X_2 = a_k)$ for $j = 1, 2; k = 1, 2$.

Multivariate discrete distribution

- Let $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ be a p-dimensional vector of random variables and $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ be an observation of $\mathbf{X}$.
- If $X_1, X_2, \ldots, X_p$ are discrete, the joint distribution of $\mathbf{X}$ is specified through a **joint probability mass function**, denoted by $f_{\mathbf{X}}(\mathbf{x})$.

$$f_{\mathbf{X}}(\mathbf{x}) = P(X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p),$$

  which is understood as the probability of $X_1, X_2, \ldots, X_p$ taking value $x_1, x_2, \ldots, x_p$ respectively at the same time.

Multivariate continuous distribution

- If $X_1, X_2, \ldots, X_p$ are continuous, the joint distribution of $\mathbf{X}$ is specified through a **joint density function**, denoted by $f_{\mathbf{X}}(\mathbf{x})$ where for any p-dimensional region R,

$$P(\mathbf{X} \in R) = \int \cdots \int_R f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}.$$

For example, when $p = 2$, for $R = [0, 1] \times [-1, 2]$,

$$P(\mathbf{X} \in [0, 1] \times [-1, 2]) = \int_0^1 \int_{-1}^2 f_{\mathbf{X}}(x_1, x_2) dx_2 dx_1$$

Why are multivariate distributions important?

- It is an important tool to understand the statistical relationship between multiple variables, for example, the relationships between the risk of car accident and its various risk factors.
- Concepts like correlation, independence are defined using multivariate distributions.

Marginal distributions

- Consider discrete variables $X_1, X_2, \ldots, X_p$ and its the joint distribution of $\mathbf{X}$ is specified through a **joint probability mass function**, denoted by $f_{\mathbf{X}}(\mathbf{x})$.

$$f_{\mathbf{X}}(\mathbf{x}) = P(X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p).$$

- The distribution of one random variable or a subset of random variables is called a **marginal distribution.**
- The marginal of distribution of X_1 is the also called the marginal distribution of X_1.
- It gives the probabilities of various values of the variables in the subset without reference to the values of the other variables.

Marginal distributions for multivariate discrete distributions example

Consider a case when $p = 2$ and both X_1 and X_2 are binary random variables taking value 1 and 2. The joint probability mass function is given in the following 2×2 table:

Table: 2×2 table.

		X_2	
		a_1	a_2
X_1	b_1	p_{11}	p_{12}
	b_2	p_{21}	p_{22}

Marginal distributions for multivariate discrete distributions example

- X_1 has the marginal distribution

Table: Marginal distribution of X_1.

x_1	b_1	b_2
$f(x_1)$	$p_{11} + p_{12}$	$p_{21} + p_{22}$

- X_2 has the marginal distribution

Table: Marginal distribution of X_2.

x_2	a_1	a_2
$f(x_2)$	$p_{11} + p_{21}$	$p_{12} + p_{22}$

For the Coronary heart patient data example, what are the marginal distributions of BP and SC respectively?

BP: denote '0' as BP < 120/80 mmHg

denote '1' as BP $\geq$ 120/80 mmHg

SC: denote '0' as SC < 240 mg/dl

denote '1' as SC $\geq$ 240 mg/dl

Marginal distribution of BP

BP

x	0	1
f(x)	0.245	0.755

Marginal distribution of SC

SC

y	0	1
f(y)	0.525	0.475

Marginal distributions for multivariate discrete distributions

- Suppose you are given the joint distribution.
- The marginal probability mass function of X_1 is

$$f(x_1) = P(X_1 = x_1) = \sum_{x_2,\ldots,x_p} f_{\mathbf{x}}(x_1, x_2, \ldots, x_p)$$

  for all possible values of X_1.
- The marginal probability mass function of X_2 is

$$f(x_2) = P(X_2 = x_2) = \sum_{x_1, x_3\ldots,x_p} f_{\mathbf{x}}(x_1, x_2, \ldots, x_p)$$

  for all possible values of X_2.

Conditional distributions in the discrete case

- Consider the 2×2 table again.

Table: 2×2 table.

		X_2	
		a_1	a_2
X_1	b_1	p_{11}	p_{12}
	b_2	p_{21}	p_{22}

- X_1 and X_2 have the marginal distributions:

Table: Marginal distribution of X_1 and X_2.

x_1	b_1	b_2
$f(x_1)$	$p_{11} + p_{12}$	$p_{21} + p_{22}$
x_2	a_1	a_2
$f(x_2)$	$p_{11} + p_{21}$	$p_{12} + p_{22}$

Conditional distributions in the discrete case

- The conditional probability $P(X_1 = x_1 | X_2 = x_2)$ is denoted as
$$f(x_1|x_2) = \frac{P(X_1 = x_1, X_2 = x_2)}{P(X_2 = x_2)} = \frac{f_{\mathbf{X}}(x_1, x_2)}{f(x_2)}$$
where $f(x_2)$ is the marginal distribution of X_2.
- Conditional distributions of X_1 given that X_2 takes value b_1 or b_2 are as follows

Table: Conditional distributions of X_1 given $X_2 = a_1$ or a_2.

x_1	b_1	b_2	total
$f(x_1\|x_2 = a_1)$	$p_{11}/(p_{11} + p_{21})$	$p_{21}/(p_{11} + p_{21})$	1
$f(x_1\|x_2 = a_2)$	$p_{12}/(p_{12} + p_{22})$	$p_{22}/(p_{12} + p_{22})$	1

For the Coronary heart patient data example, what are the conditional distributions of BP and SC respectively?

Conditional distributions of BP given SC

BP	0	1
$f(BP \mid SC=0)$	$0.115/0.525$	$0.41/0.525$
$f(BP \mid SC=1)$	$0.13/0.475$	$0.345/0.475$

Independence

- Intuitively, two random variables X and Y are independent if knowing the value of one of them does not change the probabilities for the other one.
- Two random variables are independent if and only if

  $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ for all sets of A and B.

- When both random variables are discrete, they are independent if and only if

  $$P(X = x, Y = y) = P(X = x)P(Y = y) \text{ for all } x \text{ and } y,$$

  i.e. $f_{X,Y}(x, y) = f(x)f(y)$ for all pairs of x and y.

Review (11/5/2019)

Consider a discrete joint distribution table

$f(x,y)$		y: 1	2	3	$f_X(x)$
x	1	0.25	0.25	0	0.5
	2	0	0.25	0.25	0.5
$f_Y(y)$		0.25	0.5	0.25	

$\mu_X = \frac{3}{2}$ $\mu_Y = 2$ $\sigma_X = \frac{1}{2}$ $\sigma_Y = \sqrt{\frac{1}{2}}$

$Cov(X,Y) = \frac{1}{4}$

$\rho = 0.71$

- Independence
- Covariance
- Correlation

Probability $\left(\begin{array}{l} \text{K·M rule} \\ 1)\ 0 \le P(A) \le 1 \\ 2)\ P(S) = 1 \end{array} \right.$ 3) If $A \cap B = \phi$, $\Rightarrow \boxed{\begin{array}{l} P(A \cup B) \\ = P(A) + P(B) \end{array}}$

$\downarrow$

Random variables

(Discrete / Continuous r.v.)

$\downarrow$

Probability mass function $f(x) = P(\underline{X = x})$

Probability density function $P(\underline{a < X < b})$
$f(x)$

$$= \int_a^b f(x)\,dx$$

Example

If X is a discrete r.v. where

$X = \{1, 2, 6, 10\}$

$$P(\underbrace{X < 6}_{A \cup B}) = P(\underbrace{X = 1}_{A}) + P(\underbrace{X = 2}_{B})$$

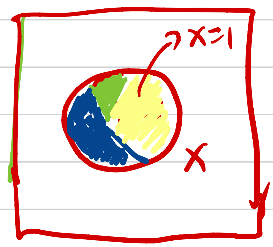
If X is a continuous r.v

$X = \{x \mid -1 < x < 3\}$

$$P(\underbrace{0 < X < 1}_{A \cup B}) = P(\underbrace{0 < X < 0.5}_{A}) + P(\underbrace{0.5 \le X \le 1}_{B})$$

Consider two discrete r.v. jointly

$f(x,y)$ x \ y	1	2	3
1	0.25	0.25	0
2	0	0.25	0.25



1) $0 \le f(x,y) \le 1$

2) $\sum_{x} \sum_{y} f(x,y) = 1$

3) $p(X=1, Y=3) = 0$

$p(X=1) = P(\{X=1, Y=1\} \cup \{X=1, Y=2\} \cup \{X=1, Y=3\})$

$= 0.5$ $\qquad = 0.5$

$p(X=1, Y<3) = f(1,1) + f(1,2)$

$f(x,y)$ / x \ y	y_1	y_2	$\cdots$	y_m
x_1				
x_2				
x_3				
$\vdots$				
x_k				

$f(x,y)$ / x \ y	1	2	3
1	$2c$	$3c$	$4c$
2	$3c$	$4c$	$5c$

$$f(x,y) = c(x+y)$$

$x = 1, 2$

$y = 1, 2, 3$

If X and Y are continuous r.v;

the joint density function

$f(x,y)$ satisfy the following

Conditions;

1) $f(x,y) \geq 0$

2) $$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\,dx\,dy = 1$$

3) $$P((X,Y) \in R)$$

$$= \iint_R f(x,y)\,dx\,dy$$

Bivariate Normal distribution

$$f(x,y) = \frac{1}{\sqrt{2\pi|\Sigma|}} e^{-\frac{1}{2}(x-\mu_x,\, y-\mu_y)^T \Sigma^{-1} \begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}}$$

$-\infty < x < \infty$ Σ : Covariance

$-\infty < y < \infty$ $\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}$: mean vector

---

Marginal distributions

Conditional distributions

$f(x,y)$	y: 1	2	3	$f_X(x)$
x = 1	0.25	0.25	0	0.5
2	0	0.25	0.25	0.5
$f_Y(y)$	0.25	0.5	0.25	1

$$f_Y(y) = P(Y=y)$$

Conditional distributions

Conditional distribution of X given $Y=1$

X	1	2
$f_{X\|Y}(x\|y=1)$	1	0

$$P(A\mid B) = \frac{P(A\cap B)}{P(B)} \qquad f(x\mid y=1) = \frac{P(X=x, Y=1)}{P(Y=1)}$$

$$f(x=1\mid y=1) = \frac{P(X=1, Y=1)}{P(Y=1)} = \frac{0.25}{0.25} = 1$$

Independence of two/multiple r.v.s.

Covariance of two r.v.s

Correlation of two r.v.s.

Independence of X and Y $\Leftrightarrow$ for any sets A, B

$$P(X \in A, Y \in B) = P(X \in A)\, P(Y \in B)$$

Covariance

$$\sigma_{X,Y} = E\left[(X - \mu_X)(Y - \mu_Y)\right]$$

Correlation

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \cdot \sigma_Y} \; ; \quad -1 \le \rho_{X,Y} \le 1$$

Are X an Y independent ?

$f(x,y)$	$y = 1$	2	3	$f_X(x)$
$x = 1$	0.25	0.25	0	0.5
2	0	0.25	0.25	0.5
$f_Y(y)$	0.25	0.5	0.25	1

$\Longrightarrow$ Does $f(x,y) = f_X(x)\, f_Y(y)$ for all possible (x,y)?

$$f(1,1) \stackrel{?}{=} f_X(1) \cdot f_Y(1)$$

$$0.25 \neq 0.5 \cdot 0.25 \Longrightarrow$$ X and Y not independent

$$f(1,2) \stackrel{?}{=} f_X(1)\, f_Y(2)$$

$$\underline{0.25} = 0.5 \cdot 0.5$$

Examples

- Denote X_1 as the outcome for flipping the first coin and X_2 as the outcome for flipping the second coin. Assume these two flips being independent. The independence implies
  - $P(X_1 = H, X_2 = H) = P(X_1 = H)P(X_2 = H)$
  - $P(X_1 = T, X_2 = H) = P(X_1 = T)P(X_2 = H)$
  - $P(X_1 = H, X_2 = T) = P(X_1 = H)P(X_2 = T)$
  - $P(X_1 = T, X_2 = T) = P(X_1 = T)P(X_2 = T)$

How to show that two random variables are dependent?

- In general, if two random variables are dependent if and only if

  $P(X \in A, Y \in B) \neq P(X \in A)P(Y \in B)$ for some sets of A and B.

- If both random variables are discrete and dependent if and only if

  $P(X = x, Y = y) \neq P(X = x)P(Y = y)$ for some x and y.

Coronary heart patient data revisited

For the Coronary heart patient data example, are BP and SC independent? (Note that we do not take sampling variability into account here) skipped

Independence assumption for multiple variables

- Consider multiple random variables $X_1, X_2, \cdots, X_p$ where $p \geq 3$.
- $X_1, X_2, \cdots, X_p$ are mutually independent if and only if any sub-collection of the random variables from $X_1, \ldots, X_p$ are independent from another non-overlapping sub-collection.
- Usually verifying mutual independence for multiple variables is not trivial.
- If $X_1, X_2, \cdots, X_p$ are mutually independent, $P(X_1 = x_1, X_2 = x_2, \cdots, X_p = x_p) = P(X_1 = x_1)P(X_2 = x_2) \cdots P(X_p = x_p)$ for all possible $x_1, x_2, \ldots, x_p$.

Covariance

- Common measure of the relationship between two random variables are covariance and correlation.
- The covariance between two random variables X and Y, denoted as $cov(X, Y)$ or $\sigma_{X,Y}$, is

$$\sigma_{X,Y} = E[(X - \mu_X)(Y - \mu_Y)]$$

  where μ_X and μ_Y are the mean of the marginal distributions of X and Y respectively.
- $E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X \mu_Y$.
- **Expected value** of a function of two **discrete** random variables:

$$E[h(X, Y)] = \sum \sum h(x, y) f(x, y)$$

Example: 2×2 table

- In the 2×2 table,

Table: 2×2 table.

		X_2	
		1	2
X_1	1	p_{11}	p_{12}
	2	p_{21}	p_{22}

$$
\begin{aligned}
E(X_1X_2) &= 1*1*p_{11} + 1*2*p_{12} + 2*1*p_{21} + 2*2*p_{22} \\
&= p_{11} + 2p_{12} + 2p_{21} + 4p_{22} \\
&= 1 + p_{12} + p_{21} + 3p_{22}
\end{aligned}
$$

- The marginal mean of X_1 is $\mu_{X_1} = 1 + p_{21} + p_{22}$.
- The marginal mean of X_2 is $\mu_{X_2} = 1 + p_{12} + p_{22}$.
- Hence covariance of X_1 and X_2 is
  $\sigma_{X_1 X_2} = E(X_1 X_2) - \mu_{X_1}\mu_{X_2} = p_{22}p_{11} - p_{12}p_{21}$.

For the Coronary heart patient data example, find the covariance of BP and SC. Denote the two levels of BP or SC as 0 or 1.

Table: 2 × 2 table.

		SC (0)	(1)	
		Below 240 mg/dL	Above 240 mg/dL	Total
BP	Below 120/80 mm Hg (0)	0.115	0.13	0.245
	Above120/80 mm Hg (1)	0.41	0.345	0.755
Total		0.525	0.475	1

$$\operatorname{cov}(BP, SC) = E(BP \cdot SC) - E(BP) \cdot E(SC)$$

$$= 0.345 - 0.755 \cdot 0.475$$

$$= -0.013$$

$$\operatorname{corr}(BP, SC) = \frac{-0.013}{\sqrt{0.185 \times 0.249}} = -0.06$$ where 0.185 is the variance of BP, 0.249 is the variance of SC.

Correlation

- The correlation measures the direction and strength of the linear relationship between two quantitative variables.
- The correlation between two random variables X and Y, denoted as ρ, is
$$\rho = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$
where σ_X and σ_Y are standard deviations of marginal distribution of X and Y respectively.
- For any two random variable X and Y, $-1 \leq \rho \leq 1$.

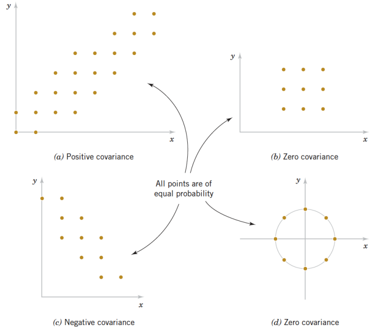


FIGURE 5-12 Joint probability distributions and the sign of covariance between X and Y.

- $\rho > 0 \Leftrightarrow$ (equivalent to) variables have a positive linear relationship.
- $\rho > 0 \Leftrightarrow$ variables have a negative linear relationship.
- The absolute value of ρ is close to 1 $\Leftrightarrow$ the linear association is strong. The absolute value of ρ is close to 0 $\Leftrightarrow$ the linear association is weak.

For the Coronary heart patient data example, find the correlation of BP and SC.

Linear combination of random variables

- **Linear combination**: Given random variables $X_1, X_2, \ldots, X_n$ and constants $c_1, c_2, \ldots, c_n$,

$$Y = c_1 X_1 + c_2 X_2 + \cdots + c_n X_n$$

  is a linear combination of $X_1, X_2, \ldots, X_n$.
- Random sample mean is a direct linear combination of $X_1, X_2, \ldots, X_n$:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{1}{n} X_1 + \frac{1}{n} X_2 + \cdots + \frac{1}{n} X_n.$$

Linear combination of random variables

- The distribution of the linear combination may be difficult to obtain but we can find the find the mean and variance of it.
- **Mean of a linear function**: If
  $Y = c_1X_1 + c_2X_2 + \cdots + c_nX_n$, then

$$E(Y) = c_1E(X_1) + c_2E(X_2) + \cdots + c_nE(X_n)$$

- For example, if $X_1, X_2, \ldots, X_n$ have the same mean 3, then $E(\bar{X}) = \frac{1}{n}E(X_1) + \frac{1}{n}E(X_2) + \cdots + \frac{1}{n}E(X_n) = 3$.

Linear combination of random variables

- Variance of the linear combination

$$\begin{aligned} Var(Y) &= c_1^2 Var(X_1) + c_2^2 Var(X_2) + \cdots + c_n^2 Var(X_n) \\ &+ 2 \sum \sum_{i<j} c_i c_j Cov(X_i, X_j) \end{aligned}$$

  where $Var(\cdot)$ denotes variance and $Cov(\cdot,\cdot)$ denotes covariance.

- If $X_1, X_2, \ldots, X_n$ are **mutually independent**,

$$Var(Y) = c_1^2 Var(X_1) + c_2^2 Var(X_2) + \cdots + c_n^2 Var(X_n)$$

If $Y = 2X_1 + X_2$, $E(X_1) = 1$, $E(X_2) = 2$, $Var(X_1) = 1$, $Var(X_2) = 1$ and $Cov(X_1, X_2) = -0.5$. What are the mean and variance of Y?

$$E(Y) = 2E(X_1) + E(X_2)$$

$$= 2 \cdot 1 + 2 = 4$$

$$Var(Y) = \overset{c_1^2}{4}\, Var(X_1) + \overset{c_2^2}{1} \cdot Var(X_2) + 2 \cdot \overset{c_1 \cdot c_2}{2}\, cov(X_1, X_2)$$

$$= 4 \cdot 1 + 1 + 4 \cdot (-0.5) = 3$$

If $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$ with $E(X_i) = \mu$, $Var(X_i) = \sigma^2$ for $i = 1, \ldots, n$ and $Cov(X_i, X_j) = -0.1\sigma^2$. Find $E(\bar{X})$ and $Var(\bar{X})$.

$$E(\bar{X}) = \frac{1}{n} E(X_1) + \frac{1}{n} E(X_2) + \cdots \frac{1}{n} E(X_n)$$

$$= \frac{1}{n} \cdot [\underbrace{\mu + \mu + \cdots \mu}_{n \text{ copies}}] = \frac{1}{n} \cdot n \cdot \mu = \mu$$

$$Var(\bar{X}) = \frac{1}{n^2} Var(X_1) + \frac{1}{n^2} Var(X_2) + \cdots \frac{1}{n^2} Var(X_n)$$

$$+ 2 \cdot \frac{1}{n} \cdot \frac{1}{n} \cdot \sum_{i<j} \sum Cov(X_i, X_j)$$

$$= \frac{1}{n^2} (\underbrace{\sigma^2 + \sigma^2 + \cdots \sigma^2}_{n \text{ copies}}) + \frac{2}{n^2} \cdot \frac{n(n-1)}{2} \cdot (-0.1\sigma^2)$$

$$= \frac{1}{n} \sigma^2 - 0.1 \frac{n-1}{n} \sigma^2$$

$$= \frac{1.1 - 0.1n}{n} \cdot \sigma^2$$

If $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$ with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ for $i = 1, \ldots, n$. Suppose that $X_1, X_2, \ldots, X_n$ are independent. Find $E(\bar{X})$ and $Var(\bar{X})$.

$E(\bar{x})$ is still μ

$$Var(\bar{x}) = \frac{1}{n^2} Var(x_1) + \frac{1}{n^2} Var(x_2) + \cdots \frac{1}{n^2} Var(x_n)$$

$$= \frac{1}{n^2}[\underbrace{\sigma^2 + \sigma^2 + \cdots \sigma^2}_{n \text{ copies}}] = \frac{\sigma^2}{n}$$

Independence in Central limit theorem

- Central limit theorem (CLT): If $X_1, \ldots, X_n$ are mutually independent random variables having a distribution (not necessarily Normal distribution) with mean μ and variance σ^2 and if $\bar{X}$ is the sample mean, then

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

for large n.

- If $X_1, X_2, \ldots X_n$ are identically and independently distributed from $N(\mu, \sigma^2)$, regardless how large n is, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$