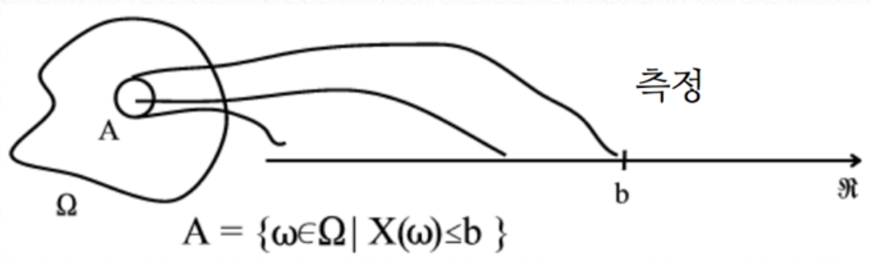
1

확률변수_기대값

1. 확률변수

1) 정의



(정의) 확률실험 표본공간 S가 정의역(입력, 함수의 x), 실수(real number)가 공역(출력, 함수의 $y = f(x)$)인 측정 함수

- (기호) 알파벳 : W, X, Y, Z
- $X(w) = x,\ w \subseteq S$

확률변수는 데이터의 변수와 동일하다.

2) 종류

(1)이산형과 연속형

- 이산형 discrete : 가질 수 있는 값이 유한 (예) 온도, 고통사고 건수, 성별, 직업 종류
- 연속형 continous : 가질 수 있는 값이 무한, 어떤 작은 구간 내에도 값이 발생 (예) 몸무게, 소득, 수능성적

(2)양적 vs. 질적

- 양적 qualitative, 측정형 metric : 숫자로 표현할 수 있는 변수

1) 비율 ratio : 배수의 개념이 성립 (예) 소득, 몸무게

2) 구간 interval : 배수 개념 성립 않음 (예) 온도

- 질적 qualitative, 범주형 non-metric : 개체를 분류하기 위하여

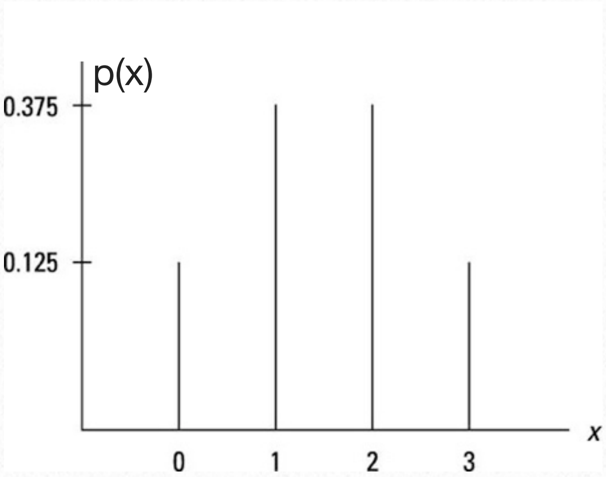
1) 순서형 ordinal : 순서가 있는 분류 (예) 알파벳 성적, 소득 수준 상중하

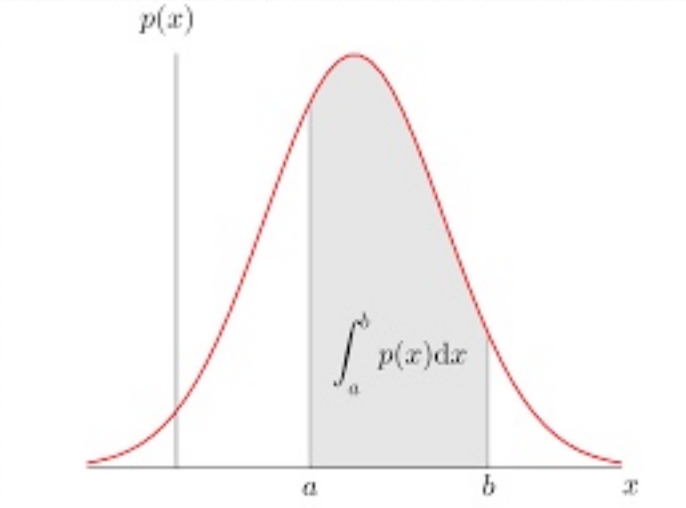
2) 명목형 nominal : 순서 없는 분류 (예) 성별, 직업 종류

2. 확률밀도함수 prob. density/mass fn.

1) (정의) $P(X = x), p(x), f(x)$

- 확률변수의 값이 정의역, 각 값에 대응하는 확률 값을 공역으로 하는 규칙
- 이산형의 확률은 막대 높이, 연속형의 확률은 면적 (그러므로 x의 한 값에서 확률은 0이다)
- 규칙은 함수, 표, 그래프로 표현 - x-축의 확률변수 값, y-축은 대응하는 확률 값





2) 확률공리

1) $p(x) \geq 0,$ for all x

2) $\sum p(x) = 1, \int f(x)dx = 1$

3) 활용

모수의 추정에 사용되는 통계량의 샘플링분포 sampling distribution을 알아야 구간 추정값을 계산하고 통계적 가설을 검정할 수 있다.

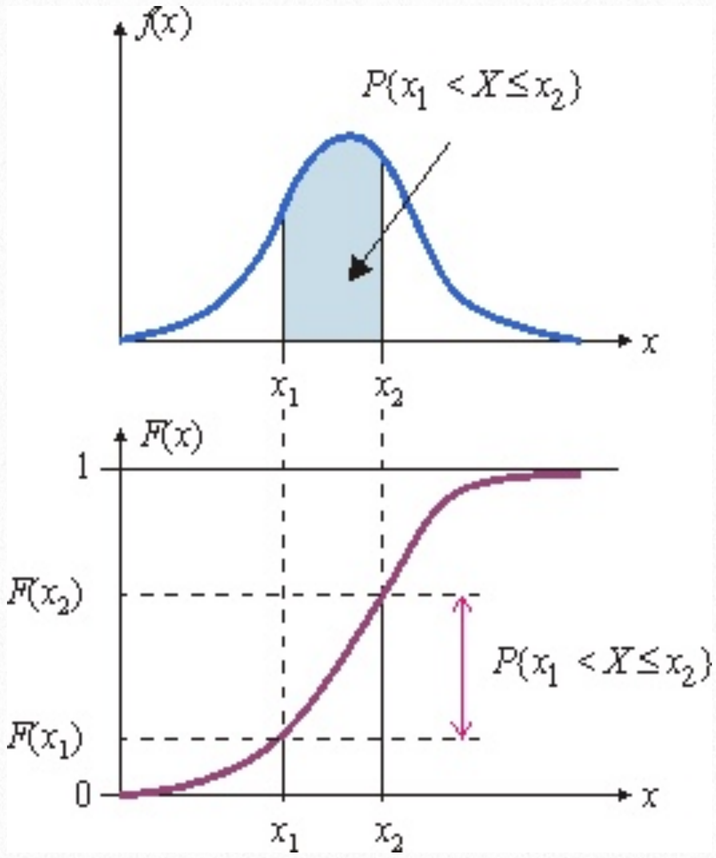
통계의 추론은 신뢰수준 95%, 유의수준(오류 가능성 확률) 5% 등으로 확률에 의해 표현된다.

3. 누적확률밀도함수 cumulative PDF

1) 정의 definition

확률변수 X의 정의역의 가장 작은 값부터 임의의 값 x 까지 (x값을 포함) 확률 값을 누적시킨 함수

(기호) $F(x) = P(X \leq x) = \sum_{-\infty}^{x} p(x) = \int_{-\infty}^{x} f(x)dx$



2) 성질

1) $F(x)$는 비감소 함수 : $x_1 < x_2 => F(x_1) \leq F(x_2)$

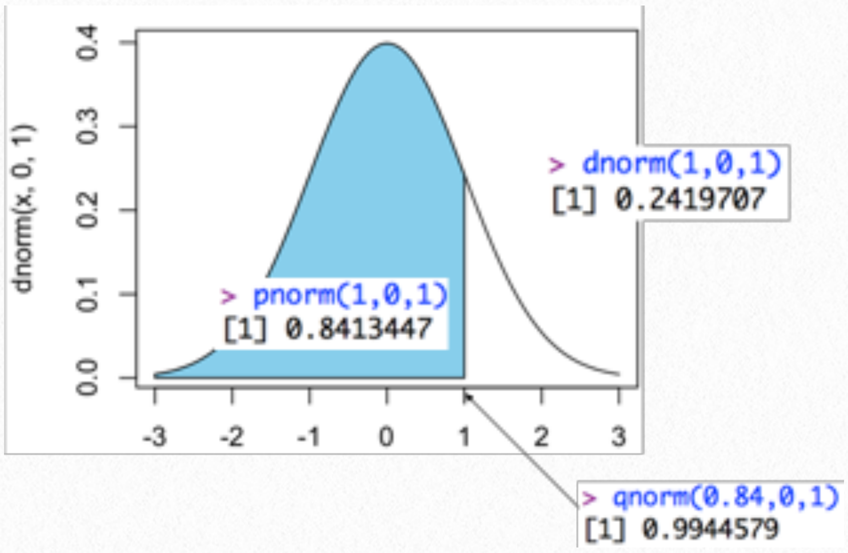
2) $F(-\infty) = 0, F(\infty) = 1$

3) (이산형) $P(x) = F(x) - F(x-)$

4) (연속형) $F'(x) = f(x)$

4. 엑셀 & R 함수

1) in R



함수	기능
d*(x, 모수)	확률밀도함수 확률 값, f(x)
p*(x, p, 모수)	분포함수 값, F(x)
q*(p, 모수)	역분포함수 값, F−1(p)
r*(n, 모수)	분포함수 따르는 데이터 n개 랜덤하게 생성

이항분포 *binom(x=성공회수,시험회수=n, 성공확률=p)

포아송분포 *pois(x=관심 사건 수,평균=λ)

음이항분포 *nbinom(x=실패회수,size=성공회수,p=성공확률)

초기하분포 hyper(관심개수M,모집단수N-M,표본개수n)

카이제곱 *chisq(x, df=자유도)

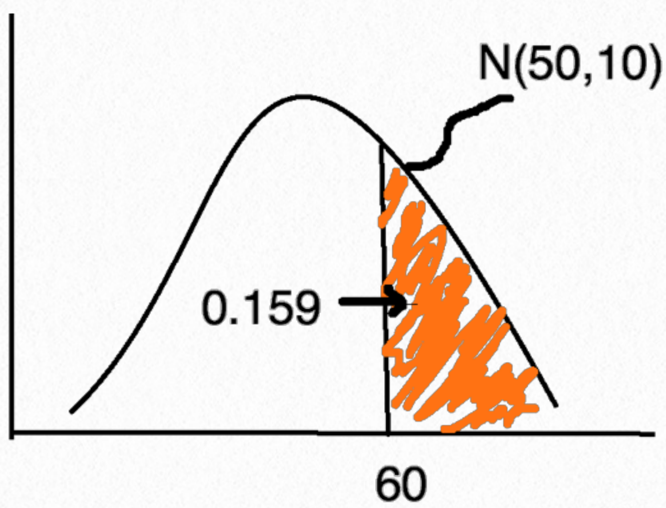
감마분포 *gamma(x, 형상모수=α, 크기모수=1/β)

정규분포 *norm(x,평균=μ,표준편차=sd)

2) 엑셀

http://www.excelfunctions.net/Excel-Statistical-Functions.html

Distribution	PDF/CDF	Inverse	Right Tailed	Test
Beta	BETA.DIST	BETA.INV		
Binomial	BINOM.DIST	BINOM.INV		
Chi Square	CHISQ.DIST	CHISQ.INV	CHISQ.DIST.RT CHISQ.INV.RT	CHISQ.TEST
Exponential	EXPON.DIST			
F	F.DIST	F.INV	F.DIST.RT F.INV.RT	F.TEST
Gamma	GAMMA.DIST	GAMMA.INV		
Hypergeometric	HYPGEOM.DIST			
Lognormal	LOGNORM.DIST	LOGNORM.INV		
Negative Binomial	NEGBINOM.DIST			
Normal	NORM.DIST	NORM.INV		
Poisson	POISSON.DIST			
Standard Normal	NORM.S.DIST	NORM.S.INV		Z.TEST
Student's T	T.DIST	T.INV	T.DIST.RT T.DIST.2T T.INV.2T	T.TEST
Weibull	WEIBULL.DIST			



=1−NORM.DIST(60,50,10,1)

D	E	F
0.159		

=NORM.INV(1−0.159,50,10)

D	E	F
0.159		
60		

```
> 1-pnorm(60,50,10)
[1] 0.1586553
> qnorm(1-0.1586533,50,10)
[1] 60.00008
```

5. 기대값 expected value

확률변수 값을 무한 관측했을 때 평균 개념으로 기대되는 값

1) 정의

$$E(X) = \sum_x p(x)x,\ E(X) = \int xf(x)dx$$

◉ (데이터 평균) $p(x) = \dfrac{1}{n}$

◉ 함수($g(x)$)의 기대값 :

$$E(g(x)) = \sum_x g(x)p(x) = \int g(x)f(x)dx$$

◉ 분산 variance : $g(x) = (X - E(X))^2$

$V(X) = E(X - E(X))^2$ (간편식) $= E(X^2) - E(X)^2$

◉ 표준편차 standard deviation-분산 양의 제곱근

$SD(X) = \sqrt{V(X)}$

주사위 2개를 던졌을 때 첫 주사위 눈금과 두번째 주사위 눈금의 차이에 대한 확률밀도함수를 구하시오. 그리고 기대값과 분산을 구하시오.

(St. Petersburg Paradox) 주머니에 $1이 있고 동전을 던져 앞면이 나타나면 주머니 돈이 2배가 된다. 동전 던지기는 뒷면이 한 번 나타나면 종료되며 주머니의 돈은 상금으로 가져간다. 그러므로 게임참가자가 동전을 던져 뒷면이 첫 번째 나오면 $1, 두번째 나오면 $2, 세 번째는 $4… 받는다. 참가비가 얼마이면 게임에 참여할 것인가? 상금을 확률변수 X라 하자.

(1) E(X) 구하시오.

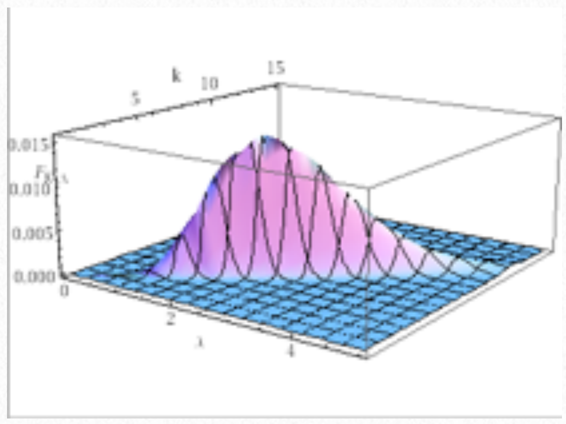
(2) 상금이 $65이상일 확률을 구하시오.

6. 결합확률밀도함수 Joint PDF

1) 개념

◉ 2개 확률변수를 동시에 고려함 - 주유소에서 시간당 주유하기 위하여 방문하는 차량 대수(이산형 확률변수)와 매출액(연속형 확률변수), 매출 경유량(연속형)과 휘발유량(연속형)

◉ 일반적으로 동일 형태의 확률변수의 결합

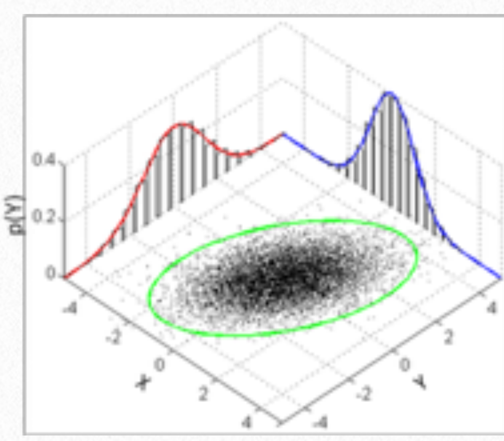


2) 정의

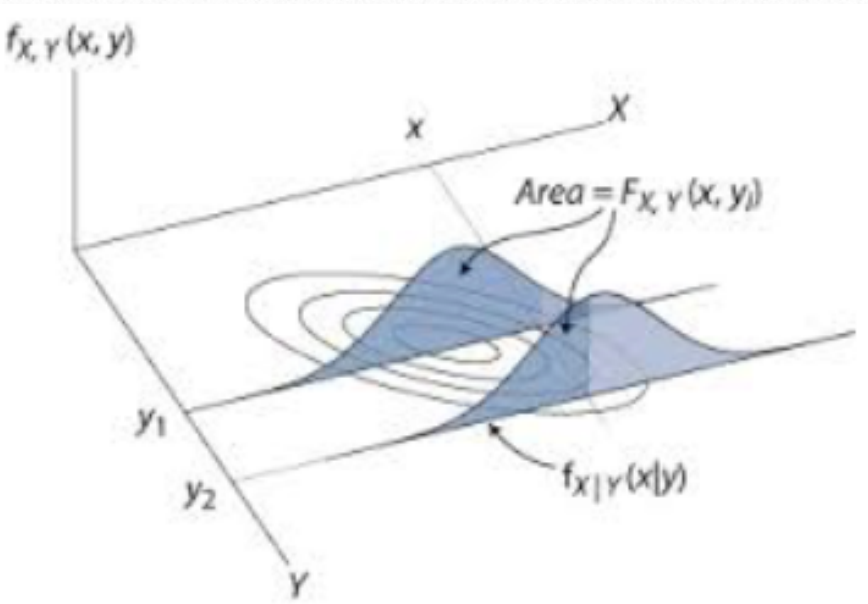
결합 PDF : $P(X = x, Y = y) = P(x, y) = f(x, y)$

누적 CDF : $F(x, y) = P(X \le x, Y \le y)$

주변 PDF : $p(x) = \sum_y p(x, y),\ f(x) = \int f(x, y)dy$



조건부 PDF : $p(x\,|\,y) = \dfrac{p(x,y)}{p(y)}, f(x\,|\,y) = \dfrac{f(x,y)}{f(y)}$



3) 독립 independence

$f(x,y) = f(x)f(y)$ <=> 이변량 확률변수 (X, Y)는 서로 독립이다.

4) 기대값

◉ (a, b)는 상수, (X, Y)는 이변량 변수라고 하자.

$E(aX + bY) = aE(X) + bE(Y)$

$V(aX \pm bY) = a^2V(X) + b^2V(Y) \pm 2COV(X,Y)$

만약 (X, Y) 서로 독립이면, $COV(X,Y) = 0$

◉ a_i 는 상수, X_i는 다변량 변수 ($i = 1,2,...,p$)

$$E(\sum_{i=1}^{p} a_iX_i) = \sum_{i=1}^{p} a_iE(X_i)$$

$$V(\sum_{i=1}^{p} a_iX_i) = \sum_{i=1}^{p} a_i^2V(X_i) + 2\sum_{i>j} COV(X_i, X_j)$$

만약 X_i가 확률표본(서로 독립, 동일분포)이면

$$V(\sum_{p}^{i=1} a_iX_i) = \sum_{p}^{i=1} a_i^2V(X_i)$$

7. 공분산과 상관계수

두 확률변수 간 선형관계 정도를 측정 (한 확률변수의 값이 증가하면 다른 확률변수의 값이 직선의 관계 속에서 변하는 정도)

(정의) $COV(X,Y) = E(X - E(X))(Y - E(Y))$

(간편식) $COV(X,Y) = E(XY) - E(X)E(Y)$

만약 (X, Y) 서로 독립이면 $COV(X,Y) = 0$

(예제) 이변량 확률변수의 확률밀도함수는 $p(x,y) = 1/3, (x,y) = (-1,0), (0,1), (1,0)$ 이다.

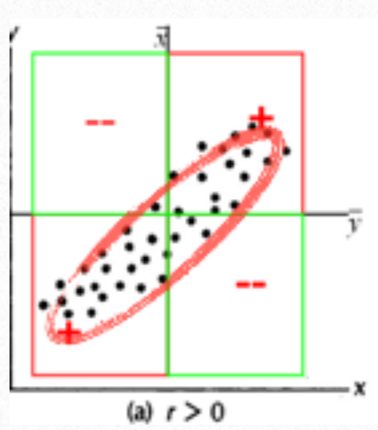
(1) 공분산은 0이다.

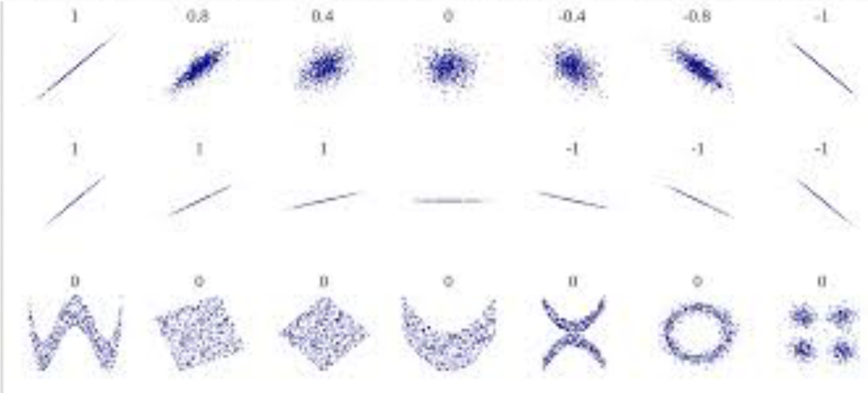
(2) 주변 확률밀도함수 $p(x) = 1/3, x = -1, 0, 1,$ $p(Y=0) = 2/3, P(Y=1)1/3$ 이므로 서로 독립이 아니다. 그러므로 공분산이 0이라고 서로 독립은 아닐 수 있다.

상관계수 $\rho = Corr(X,Y) = \dfrac{COV(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$

공분산 단위의 표준화를 위하여 각 변수의 표준편차로 공분산을 나눈 값으로 $(-1, 1)$ 을 갖는다.

두 변수의 직선 관계 정도에 대한 척도





해석

- 상관계수는 두 확률변수의 직선 관계 정도에 대한 척도이다. (linearly functioned)
- 상관계수는 -1과 1사이 값이며, 1=완벽한(모든 점들이 직선 위에 있음) 양(한 변수 관측값이 증가하면 다른 변수의 관측값이 증가) 상관관계, -1=완벽한 음(한 변수의 관측값이 증가하면 다른 변수 관측값은 감소)
- 상관계수 0은 직선의 관계가 없음을 의미함 (예제 플롯 3행의 경우 4차 함수, 마름모, 사각형, 원 등의 함수 관계는 상관계수가 0이다.
- 상관계수가 크면 관측값이 직선에 가까움 - 타원의 폭이 좁고 길이가 길수록 상관계수 ± 1 에 가까움
- 실험실 자료와 같이 연구자가 자료 수집을 control 할 수 있는 경우는 0.9 (매우 유의), 0.8(유의), 0.7(little 유의) 하다고 한다.
- 설문 조사의 리커드 척도와 같이 변수가 가질 수 있는 값이 한정된 경우 (1-5점, 물론 여러 문항을 합쳐 평균을 이용하는 경우에는 다소 문제가 해결되지만) 상관 계수는 매우 낮다. 그러므로 이런 경우는 비모수 상관 계수를 구하는 것을 권한다. Spearman 순위 상관 계수, Kendall's Tau는 비모수 상관 계수 분석 방법이다.

피어슨 상관계수 추론

(1) 통계적 가설

1) 귀무가설 : 두 변수는 서로 독립이다. $\rho = 0$, 두 변수는 직선의 상관관계가 존재하지 않는다.

2) 대립가설 : 두 변수 간에는 직선의 상관관계가 존재한다. $\rho \neq 0$

2) 검정통계량

$$TS = \frac{r}{\sqrt{(1-r)^2/n-2}} \sim t(n-2)$$

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})}^2}$$

8. 산점도 활용 - 회귀분석

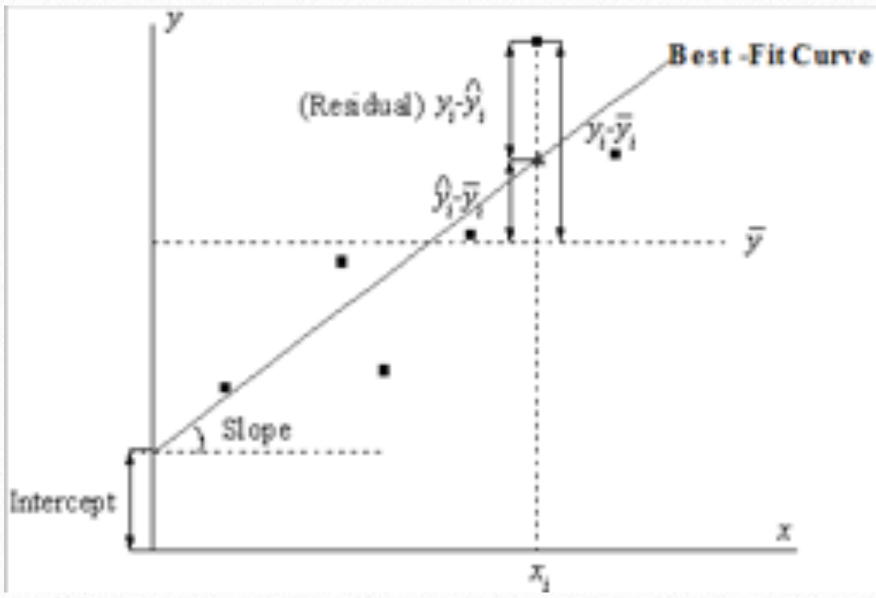
1) 산점도

- 두 확률변수의 함수관계에 대한 시각적 표현
- 함수관계 중 가장 활용성이 높은 것은 직선이다 => $Y = a + bX$

2) 회귀모형 $Y_i = a + bX_i + e_i$

- Y를 종속변수(결과), X를 설명변수(독립변수, 원인)라 한다.
- 종속변수 관측값(y_i)은 패턴(직선, 모형)에 의해 설명되는 부분($a + bX_i$)과 설명되지 않는 오차(e_i)항 부분으로 나뉜다.
- 오차항은 평균 0이고 표준편차가 인 정규분포를 따른다. $e_i \sim N(0, \sigma^2)$
- 회귀모형에서 모수는 (절편=a, 기울기=b)이다. 물론 표준편차 σ 도 모른다.

3)최소자승추정법 OLS ordinary Least Square



◉ 관측 데이터에 가장 적합한(Best fit curve) 직선을 구하고, 이 직선에 점들이 얼마나 가까이 놓여 있는 가를 판단하여 직선관계 유의성 여부를 검정

- $\min_{a,b} Q = \sum_i e_i^2 = \sum_i (y_i - a - bx_i)^2$
- $\hat{\beta} = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$ ($\hat{\beta} = \dfrac{S_{xy}}{S_{xx}}$)
- $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$

회귀계수 b 가설검정

◉ 귀무가설 : $H_0 : b = 0$ ((설명변수 유의하지 않음) (설명변수 X의 유의성 검정 할 때)

◉ 대립가설 $H_a : b \neq 0$

◉ 검정통계량 test statistic :

$$ts = \frac{\hat{b} - 0}{s(\hat{b})} \sim t(n-2),\ s(\hat{b}) = \sqrt{\frac{MSE}{Sxx}}$$

회귀계수와 상관계수 관계

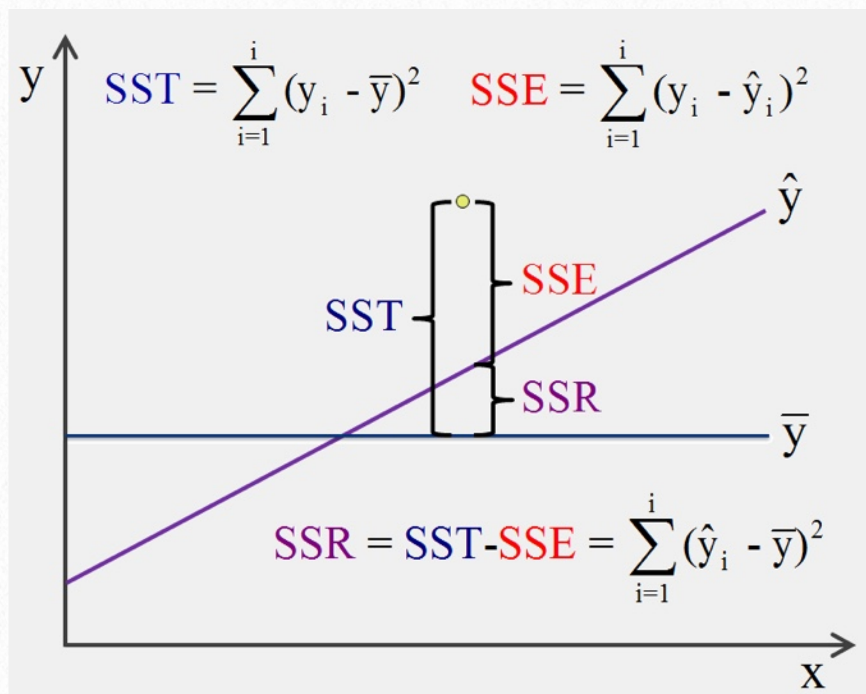
◉ 부호가 동일하며 유의성 검정도 t-검정으로 동일하고 유의확률도 같다.

◉ 즉 상관관계(직선 관계)가 유의하면 회귀(직선)모형도 유의하다.

4) 분산분석표 ANOVA table

귀무가설 : 설정한 $Y = a + bX$가 유의하지 않음 <=> (설명변수 유의하지 않음)

대립가설 : 설정한 $Y = a + bX$가 유의



변동	자승합	자유도	평균자승합	기대평균	F-통계량
모형변동	SSR	1	MSR=SSR/1	$\sigma^2 + \beta^2 \sum (x_i - \bar{x})^2$	F=MSR/MSE ~F(1,n−2)
오차변동	SSE	n−2	MSE=SSE/(n−2)	$E(MSE) = \sigma^2$	
총변동	SST	n−1			

5) 결정계수 Determination Coefficient

(계산) $R^2 = \dfrac{SSR}{SST}$; 총변동 중 모형이 설명하는 비율

- ◉ 결정계수 값의 최대 1이고 최소 0이다. 90% 이상이어야 종속변수를 충분히 설명, 80% 이상이면 보통
- ◉ 단순(설명변수 1개) 회귀분석에서는 상관계수의 제곱이 결정계수이다.

고정비용과 한계비용 BASEBALL.csv : Wins-승리게임 수, Payroll - 구단지불 총년봉(단위: 백만불)

$Payroll = a + b * Win$

a : 고정비용 fixed cost

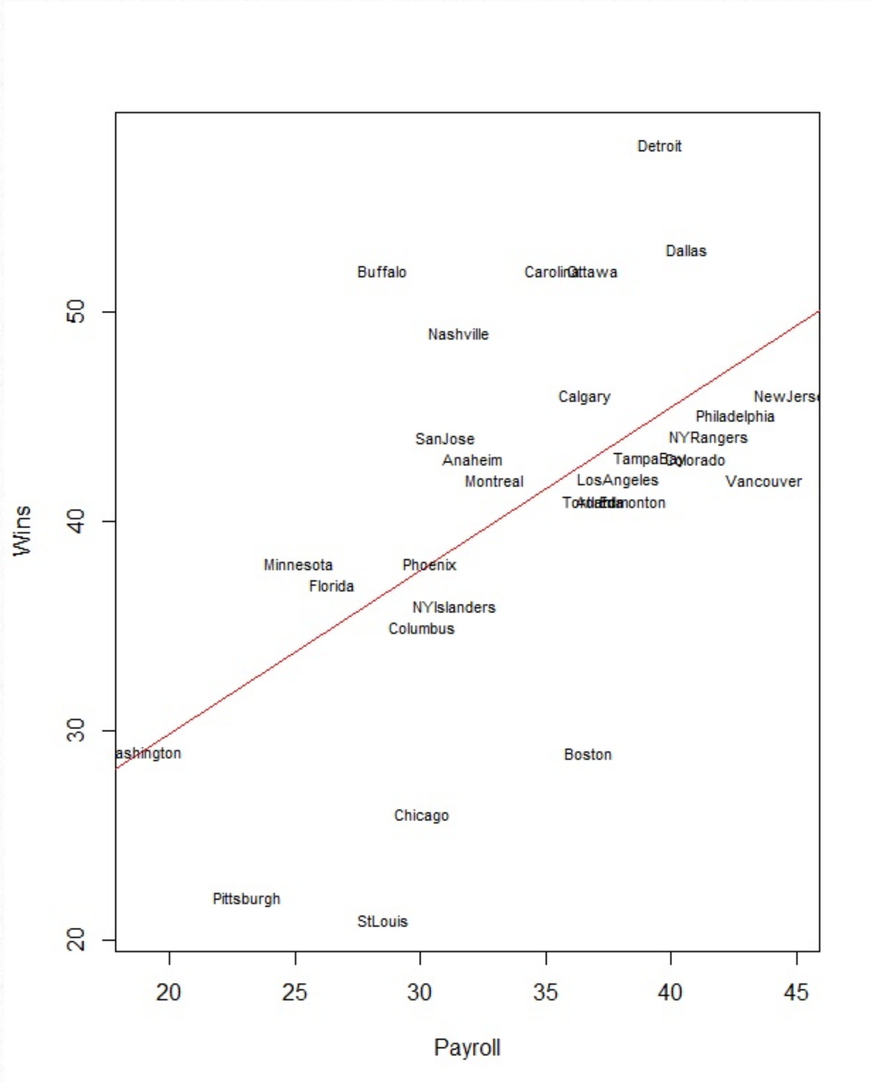
b : 한 게임을 이기기 위하여 구단이 지불해야 하는 한계비용 ❐baseball.csv

```
baseball=read.csv("http://wolfpack.hnu.ac.kr/2015_Fall/D4BE/baseball.csv")
names(baseball)
attach(baseball)
summary(lm(Wins~Payroll))
plot(Payroll,Wins,type="n")
abline(lm(Wins~Payroll),col="red")
text(Payroll,Wins,Team,cex=0.7)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.2961     7.7063   1.855  0.07414 .
Payroll       0.7782     0.2209   3.523  0.00149 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

Residual standard error: 7.59 on 28 degrees of freec
Multiple R-squared:  0.3071,    Adjusted R-squared:
F-statistic: 12.41 on 1 and 28 DF,  p-value: 0.00148
```

$Payroll = 14.29 + 0.78 * Win$ - 한 게임을 이기지 못해도 Win=0, 14.29(천사백만불) 지급해야 한다. (고정비용=1.43천만불) 그리고 한 게임 이기기 위하여 0.78백만불 선수 연봉으로 지급 : 한계비용



시스템 & 기업-특성 위험 RIM.csv RIM(블랙베리 제조사) 회사 수익률, NASDAK 수익률 ❐[RIM.csv](RIM.csv)

http://wolfpack.hnu.ac.kr/2015_Fall/D4BE/RIM.csv

- ◉ 기울기 해석 : - 종합주가 수익률 1% 증가하면 RIM 수익률은 1% 이상 증가하므로 RIM 수익률은 유동성이 높다. 즉, 다른 기업 주가에 비해 위험도가 높다.
- ◉ 결정계수 : RIM 수익률에 대한 시장 시스템이 설명하는 비율, 결정계수가 낮으면 해당 기업의 수익률은 시장 시스템이 결정하는 것이 아니라 기업의 특성이 더 중요한 결정요인이다.