COVARIANCE AND CORRELATION

(Rice P. 129; Lindgren P. 134; Bain and Engelh. P. 179, 188)

If we consider the 2 dimn r. var X, Y,
We can introduce the concept of a MIXED MOMENT (or more usually — a PRODUCT MOMENT):

$$E[(X-a)^r (Y-b)^s]$$ for r, s integers > 0

We shall be mainly concerned with one particular product moment:

$$E[(X-\mu_X)(Y-\mu_Y)]$$

— known as the covariance of X and Y — or the covariance of the Bivariate distribution.

— equivalent form: $E[XY] - (EX)(EY)$

NOTATION: σ_{XY} or $\mathrm{Cov}(X, Y)$

The covariance provides some measure of the extent to which X and Y co vary

NOTICE that if X, Y are indept, then $\mathrm{Cov}(X, Y) = 0$
(but not the converse)

CORRELATION COEFF.

However the covariance is a poor measure of covariation because it is scale-dependent
— e.g. if rv X, Y took values in inches the $\mathrm{Cov}(X, Y)$ would be 144 times what it would be if X, Y were measured in feet.

To remove this defect, the CORR COEFF is introduced:

DEFN: $$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

NOTE

In attempting to get an intuitive understanding of what the corr. coeff measures, it is important to realize that it measures a PARTICULAR KIND OF DEPENDENCE

— but there are other kinds of dependence which are not measured by it.

(We have already pointed out that it is possible to have $\sigma_{XY} = 0$ when X, Y are DEPENDENT)

SOME RESULTS ON ρ

I $\quad |\rho_{XY}| \leq 1$

Proof: Suppose U, V are any random vars

Consider $(U - kV)^2 \longrightarrow$ a nonnegve r.var

(k real const)

$$E[U - kV]^2 = E[U^2] - 2k\,E[UV] + k^2\,E[V^2]$$

and this is ≥ 0

Because this is ≥ 0, we see that we must have

$$B^2 - 4AC \leq 0$$

$$\left[\begin{array}{l} A = E[V^2] \\ B = -2E[UV] \\ C = E[U^2] \end{array}\right.$$

I.E. $\quad [E(UV)]^2 \leq E[U^2]\,E[V^2]$

Now choose $\quad U = X - EX$

$\quad V = Y - EY$

so that $\quad \{E[(X - EX)(Y - EY)]\}^2 \leq \sigma_X^2\,\sigma_Y^2$

I.E. $\quad |Cov(X, Y)| \leq \sigma_X\,\sigma_Y$

and so $\quad |\rho_{XY}| \leq 1$

RESULT 2

if $\rho^2 = 1$,

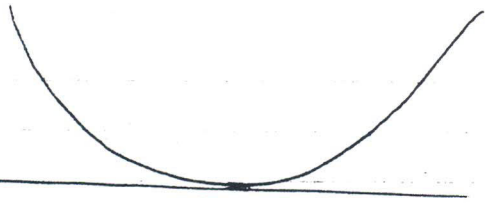
Then X and Y are linearly related with probability 1

(i.e. $\text{Prob}\left[(X - E(X)) = k_0 [Y - EY]\right] = 1$

Proof

$(EUV)^2 = [EU^2][EV^2]$ if $\rho^2 = 1$, then the function $E\left[\underline{U} - k\underline{V}\right]^2$

$\Rightarrow$ $B^2 = 4AC$ has the following appearance:



i.e. $\exists$ some value of k for which

$E\left[\underline{U} - k_0\underline{V}\right]^2 = 0$ $\quad (B^2 - 4AC = 0)$

Now, we chose $\underline{U} = X - EX$

$\underline{V} = Y - EY$

So that $E\underline{U} = E\underline{V} = E\left(\underline{U} - k\underline{V}\right) = 0$

And so $\text{Var}\left[\underline{U} - k_0\underline{V}\right] = 0$

But, arising out of Chebyshev's Ineq, we saw that if $\text{Var}\, X = 0$, then $(X = \mu)$ will prob. 1

In this case

$(X - EX) - k_0 (Y - E(Y)) = 0$ with prob 1

i.e. $X - EX = k_0 (Y - EY)$ with prob 1

i.e. X and Y are linearly related with prob 1

The Linear Relnship

Lets say k_0 — What is k_0 = ? : Just solve the Quadratic Eqn

In fact $k_0 = -\frac{B}{2A}$ $\left(\text{and } B^2 - 4AC = 0\right)$

$$= \frac{2E\left[(X - EX)(Y - EY)\right]}{2\,E(Y - EY)^2}$$

$$= \frac{\sigma_{XY}}{\sigma_Y^2} \quad ; \left(\text{Since } \rho^2 = 1\right) \text{ this } \Rightarrow \quad \frac{\sigma_X \sigma_Y}{\sigma_Y^2} = \frac{\sigma_X}{\sigma_Y}$$

Thus the linear relnship is

$$X - E\bar{X} = \frac{\sigma_{XY}}{\sigma_Y^2}\left(Y - EY\right)$$

$$\left(\text{Since } \rho^2 = 1, \quad \sigma_{XY} = \pm \sigma_X \sigma_Y\right)$$

$$= \pm \frac{\sigma_X}{\sigma_Y}\left(\bar{Y} - E\bar{Y}\right)$$

SLOPE OF THE LINE

If $\rho = 1$, then $\sigma_{XY} > 0$

and the slope of the line $\left(\frac{\sigma_{XY}}{\sigma_Y^2}\right)$ is +Ve

If $\rho = -1$ (due to $\sigma_{XY} < 0$)

The slope of the line will be −Ve.

$\sigma_{XY} = -\sigma_X \sigma_Y$

NOTE: If $Y = AX + B$ $(A, B \text{ consts})$

Then $\rho^2 = 1$

and $\rho = +1$ if $A > 0$

$\rho = -1$ if $A < 0$

$\rho = 0$ DOES NOT $\Rightarrow$ INDEPCE

Counterexample : 3 coin tosses

$\overline{X}$ = # heads ; # runs = $\overline{Y}$

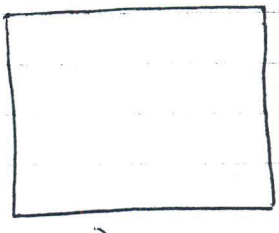
$\overline{X}$ \ $\overline{Y}$	1	2	3	
0	$\frac{1}{8}$	0	0	$\frac{1}{8}$
1	0	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{3}{8}$
2	0	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{3}{8}$
3	$\frac{1}{8}$	0	0	$\frac{1}{8}$
	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	

$$f(x,y) \neq f_{\overline{X}}(x)\, f_{\overline{Y}}(y)$$

$$\mathrm{Cov}(\overline{X},\overline{Y}) = E\left[(\overline{X}-\mu_{\overline{X}})(\overline{Y}-\mu_{\overline{Y}})\right]$$
$$= E(\overline{X}\,\overline{Y}) - \mu_{\overline{X}}\,\mu_{\overline{Y}}$$

$\mu_{\overline{X}} = 3/2$

$\mu_{\overline{Y}} = 2$



$$E(\overline{X}\overline{Y}) = 0(\quad) + 2\left(\tfrac{2}{8}\right) + 3\left(\tfrac{1}{8}\right) + 4\left(\tfrac{2}{8}\right) + 6\left(\tfrac{1}{8}\right) + 3\left(\tfrac{1}{8}\right)$$
$$= \frac{24}{8} = 3$$

Thus $\mathrm{Cov}(\overline{X},\overline{Y}) = 3 - \left(\frac{3}{2}\right)2$

$$= 0$$

NOTE : $\rho = 0 \Rightarrow$ linear covariation is absent

However this does not mean that $\overline{X}$ and $\overline{Y}$ are (in depth) unrelated

There are other kinds of dependence apart from Linear dependence — but these other kinds may not be detected by ρ

OTHER CHARACTERISTICS OF PROB. DISTRIBUTIONS

We have pointed out that the Mean Value provides us with a AVERAGE — or measure of location for the random variable. (CENTRAL VALUE)

However there are other quantities which are also used to indicate the location of a prob distrib.

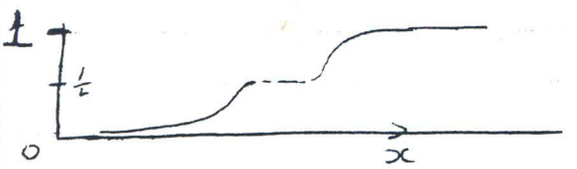
Shall only mention 2

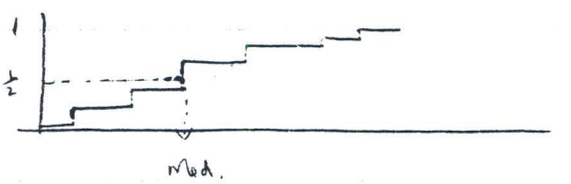
1) Median : That point x for which

$$P[X \leq x] = \frac{1}{2} \qquad F(x) = \frac{1}{2}$$

and $1 - F(x) = \frac{1}{2}$

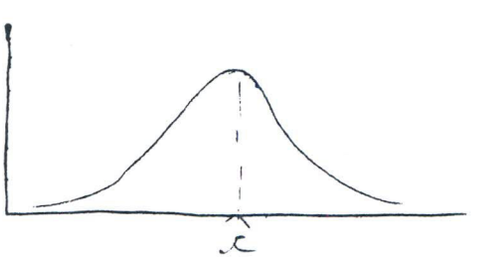
NOTE : If $F(x)$ is never const on an interval, the median is uniquely defined

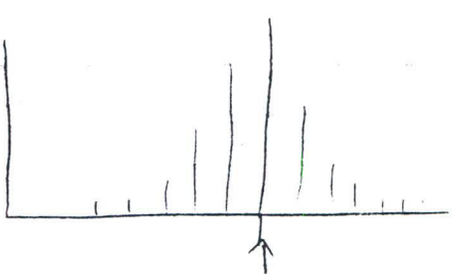




2) Mode

The value of x for which $f(x)$ is maximum (either for contin. or discrete).





OTHER MEASURES OF DISPERSION : QUANTILES

We can define quantiles x_α as follows

$$F(x_\alpha) = \alpha \qquad \text{for } 0 < \alpha < 1$$

Usual α values are: 0.01, 0.02, ---- , 0.99 $\Rightarrow$ PERCENTILES

0.1, 0.2, ---, 0.9 $\Rightarrow$ DECILES

0.25, 0.5, 0.75 $\Rightarrow$ QUARTILES

Dispersion Measures : $P_{90} - P_{10}$ (10-90 PERCentile range)

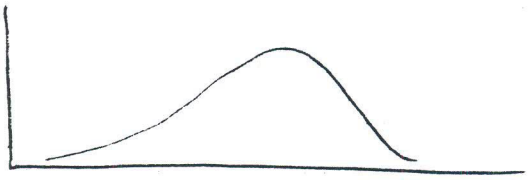
$Q_3 - Q_1$ (Interquartile range)

Some Higher Moments : μ_3 & μ_4. Skewness & Kurtosis

A few words about μ_3 and μ_4

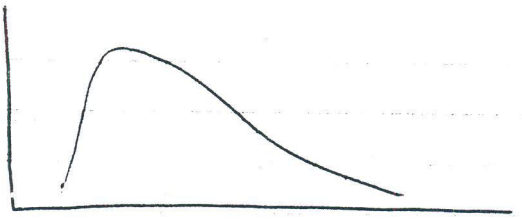
μ_3 is often termed a measure of SKEWNESS (or ASYMMETRY)

It can be shown that symmetric distribs will have $\mu_3 = 0$ and that pdfs such as :



have negative μ_3

and



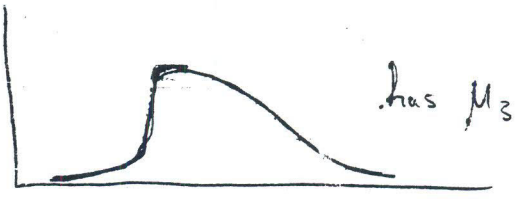
have +ve μ_3

γ_1 → (In fact $\frac{\mu_3}{\sigma^3}$ is termed the Coeff. of SKEWNESS)

Mood & Gr. P. 109-110

However, a knowledge of μ_3 gives little information as to the shape of the distrib. :

e.g.



has $\mu_3 = 0$

μ_4 : This is regarded as a Measure of KURTOSIS — i.e. measures the extent of PEAKEDNESS in the distrib.

Coeff. of KURTOSIS (Index of) λ $= \frac{\mu_4}{\sigma^4}$

$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$ (Excess of Kurtosis)

For Normal Distrib, $\gamma_2 = 0$

THIS REFERS TO PAGE 3

IF $V(X) = 0$, THEN $X = \mu$ WITH PROBABILITY 1

PAGE 7a

Proof:

Using $P[g(X) \geq c] \leq \frac{E[g(X)]}{c}$ for $c > 0$ and $g(x) \geq$

choose $g(x) = (x-\mu)^2$ and $c = t^2$

$$\Rightarrow P[|X-\mu| \geq t] \leq \frac{\sigma^2}{t^2} \qquad (*)$$

Now, Suppose $P(X = \mu) < 1$ WE WILL TRY TO GET A CONTRADICTION

then there must be probability associated with values <u>other</u> than μ, so that for some $\varepsilon > 0$,

$$P[|X-\mu| \geq \varepsilon] > 0 \qquad (**)$$

But $P[|X-\mu| \geq \varepsilon] \leq \frac{\sigma^2}{\varepsilon^2}$ (from * above)

and using the fact that $\sigma^2 = 0$, we have

$$P[|X-\mu| \geq \varepsilon] = 0$$

and this contradicts (**) above.

So, we must have

$P(X = \mu)$ with probability 1.

Variance of Sums of Random Vars

Lindgren P 137 - 140

Let $X_1, X_2, \ldots, X_n$ be n random vars and let $Z = \sum_{i=1}^{n} a_i X_i$ (a_i, real consts)

Then we want an expression for $Var(Z)$

Now $E(Z) = E\left(\sum a_i X_i\right)$

$= \sum_{i=1}^{n} a_i E X_i$ (Props of E)

$V(Z) = E[Z - EZ]^2$

$= E\left[\sum a_i X_i - \sum a_i (E X_i)\right]^2$

$= E\left[\sum_{i=1}^{n} a_i (X_i - E X_i)\right]^2$

$= E\left[\sum_{i=1}^{n} a_i^2 (X_i - E X_i)^2 + \right.$

$\left. + 2 \sum_{i<j} \sum a_i a_j (X_i - E X_i)(X_j - E X_j)\right]$

$= \sum_{i=1}^{n} a_i^2 E(X_i - \mu_{X_i})^2 + 2 \sum_{i<j} \sum a_i a_j E(X_i - E X_i)(X_j - E X_j)$

$= \sum_{i=1}^{n} a_i^2 Var\, X_i + 2 \sum_{i<j} \sum a_i a_j Cov(X_i, X_j)$

Another useful result:

$$Z = \sum_{i=1}^{n} a_i X_i \; ; \; W = \sum_{i=1}^{n} b_i X_i$$

$$Cov(Z, W) = \sum_{i=1}^{n} a_i b_i Var\, X_i + 2 \sum_{i<j} \sum a_i b_j Cov(X_i, X_j)$$

Proof: $Cov(Z, W) = E[(Z - EZ)(W - EW)]$

$= E\left[\left\{\sum a_i (X_i - \mu_i)\right\}\left\{\sum b_j (X_j - \mu_j)\right\}\right]$

$= E\left[\sum_{i=1}^{n} a_i b_i (X_i - \mu_i)^2 + 2 \sum_{i<j} \sum a_i b_j (X_i - \mu_i)(X_j - \mu_j)\right]$

$= \sum_{i=1}^{n} a_i b_i Var(X_i) + 2 \sum_{i<j} \sum a_i b_j Cov(X_i, X_j)$

Special cases

$a_i = 1 \; \forall i$ $\quad \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{i} \sum_{<i} \mathrm{Cov}(X_i, X_j)$

If $\mathrm{Cov}(X_i, X_j) = 0$ for all pairs X_i, X_j.

Then $\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}\, X_i$.

$\mathrm{Var}\left[\sum_{i=1}^{n} (a_i X_i)\right] = \sum_{i=1}^{n} a_i^2 \,\mathrm{Var}\, X_i$

$\mathrm{Cov}(\underline{Z}, \underline{W}) = \sum_{i=1}^{n} a_i b_i \,\mathrm{Var}\, X_i$

Stronger Condn; all X_i indept

CONDITIONAL EXPECTATION

(MEYER 148-150
L. 114-115
BAIN/ENGELHARDT 190-194)

Suppose X, Y are r.v.s with pdf $f(x,y)$

and $f(x|y)$ is Condit. PDF for x for given y

Then:

DEFN We define the COND. EXPECTATION of X given $Y = y$ as

$$E(X|y) = \int_{-\infty}^{\infty} x\, f(x|y)\, dx \qquad \text{CONTIN CASE}$$

and

$$E(X|y) = \sum_{x_i} x_i\, f(x_i|y_j) \qquad \text{DISCRETE CASE}$$

$E(X|y)$ is a fn of $y \longrightarrow$ say $u(y)$. We can then consider $u(Y)$ or $E(X|Y)$

Useful Result: $E_y\left[E_x[X|Y]\right] = E[X]$

Proof:

$$E_x[X|y] = \int_{-\infty}^{\infty} x\, f(x|y)\, dx$$

$$= \int_{-\infty}^{\infty} x\, \frac{f(x,y)}{f_Y(y)}\, dx$$

This is a function of y : $\rightarrow u(y)$ — not a random var.

Now let us consider $u(Y)$, which is what we mean by $E(X|Y)$, — and find $E_y(u(Y))$

$$E_y[u(Y)] = E_y\left[E_x[X|Y]\right] = \int_{-\infty}^{\infty} E(X|y)\, f_Y(y)\, dy$$

$$= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x\, \frac{f(x,y)}{f_Y(y)}\, dx\right] f_Y(y)\, dy \qquad \text{(stet)}$$

Reversing the order of integration (permissible if the expectations exist.)

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\, f(x,y)\, dy\, dx = \int x\, f_X(x)\, dx$$

$$= E(X)$$

Exa Shipments of tomatoes arrive in a shop each day. However the number of tomatoes in the shipment varies and letting N denote the number of tomatoes, we shall assume that the distrib of N is as follows

n	25	26	27	28	29	30	31	32
Prob $(N=n)$	0.05	0.1	0.1	0.2	0.3	0.15	0.05	0.05

The shopkeeper knows from experience that the average proportion of bad tomatoes is 10%.

What is the expected number of bad tomatoes per day?

$\bar{X}$ = # bad toms per day
N = # toms in shipment
We want $E(\bar{X})$

We use $E(\bar{X}) = E\left(E(\bar{X}/N)\right)$

$E(\bar{X}/n) = 0.10\,n$

$E_n\left[E(\bar{X}|N)\right] = E_n[0.10N]$
$= 0.10\,E(N)$

now $E(N) = 28.5$

Thus $E(\bar{X}) = 2.85$

1.25
2.60
2.70
5.60
8.70
4.50
1.55
1.60
28.50

NOTE For any real valued function $g(x,y)$, it can be shown (using the same procedure as above) that

$$E[g(\bar{X},Y)] = E_y\left[E_{x/y}(g(\bar{X},Y)|y)\right]$$
$$= E_x\left[E_{y/x}(g(\bar{X},Y)|x)\right]$$

$$E[g(\bar{X},Y)] = \iint g(x,y)\,f(x,y)\,dy\,dx$$
$$= \iint g(x,y)\frac{f(x,y)}{f_x(x)}\,f_{\bar{X}}(x)\,dy\,dx = \int \underbrace{\left[\int g(x,y)f(y/x)\,dy\right]}_{E[g(\bar{X},Y)|x]} f_{\bar{X}}(x)\,dx$$
$$E_x[\quad]$$

CONDITIONAL VARIANCE

Lindgren P 130

X, Y has joint distrib $f(x,y)$

Defn $\quad \mathrm{Var}(X|y) = E\left[(X - \mu_{X|y})^2 \mid y\right]$ $\qquad \mu_{X|y} = E(X|y)$

This is again a function of y, say $u(y)$

Thus we can consider $u(Y)$, written as $\mathrm{Var}(X|Y)$

We might expect that $E_y[\mathrm{Var}(X|Y)] = \mathrm{Var}(X)$ ——— but NO

to investigate this:

We need the following result:

$$E[(X-a)^2] = \mathrm{Var}\, X + (\mu - a)^2$$

PARALLEL AXIS THEOREM

Now

$$E\left[(X-\mu_X)^2 \mid y\right] = \mathrm{Var}(X|y) + (\mu_{X|y} - \mu_X)^2$$

I.E.

$$\underbrace{\int_{-\infty}^{\infty} (x-\mu_X)^2 \frac{f(x,y)}{f_Y(y)}\, dx}_{g(y) \text{ say}} = \;\ldots\ldots$$

Let us take the expected value of $g(Y)$

$$\Rightarrow \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} (x-\mu_X)^2 \frac{f(x,y)}{f_Y(y)}\, dx\right] f_Y(y)\, dy = E_y \mathrm{Var}(X|Y) + E_y\left[\mu_{X|Y} - \mu_X\right]^2$$

$$\underbrace{\iint (x-\mu_X)^2 f(x,y)\, dx\, dy}_{} = \;\ldots\ldots$$

$$\mathrm{Var}(X) = E_y \mathrm{Var}(X|Y) + \mathrm{Var}\left(\mu_{X|Y}\right)$$

Since $E\left[\mu_{X|Y}\right] = EX = \mu_X$

Exa: Conditional Mean & Variance for Trinomial Multinomial

$$P[X_1 = r_1, \; X_2 = r_2] = \frac{n!}{r_1!\, r_2!\, (n-r_1-r_2)!}\, p_1^{r_1}\, p_2^{r_2} (1-p_1-p_2)^{n-r_1-r_2}$$

We saw earlier that

$$P[X_2 = r_2 \mid X_1 = k] = \frac{(n-k)!}{r_2!\,(n-k-r_2)!} \left(\frac{p_2}{1-p_1}\right)^{r_2} \left(\frac{1-p_1-p_2}{1-p_1}\right)^{n-k-r_2}$$

$$\sim b\left(n-k, \frac{p_2}{1-p_1}\right)$$

Thus it is clear that

$$E[X_2 \mid X_1 = k] = \text{Mean of binomial r.v. with parameters } (n-k) \text{ and } \frac{p_2}{1-p_1}$$

$$\text{i.e.} \quad = (n-k)\frac{p_2}{1-p_1}$$

$$= \frac{n p_2}{1-p_1} - \left(\frac{p_2}{1-p_1}\right) k$$

and

$$\mathrm{Var}[X_2 \mid X_1 = k] = \text{Var. of r.var having } b\left(n-k, \frac{p_2}{1-p_1}\right)$$

$$\text{i.e.} \quad = (n-k)\left(\frac{p_2}{1-p_1}\right)\left(1 - \frac{p_2}{1-p_1}\right)$$

Note: (Terminology)

The Conditional expectation $E(Y \mid x)$ is termed the REGRESSION FUNCTION of Y on X

— and the plot of this function against x is termed a REGRESSION CURVE

When this function is linear, it is referred to as the REGRESSION LINE.

Check:

$$E[E(X_2 \mid X_1)] = \frac{n p_2}{1-p_1} - \left(\frac{p_2}{1-p_1}\right) E(k)$$

$$= \left(\frac{n p_2}{1-p_1}\right) - \frac{p_2}{(1-p_1)}\, n p_1 = \frac{n p_2}{1-p_1}\left[1-p_1\right]$$

$$= n p_2$$