Chapter 4 Expectations of functions of RVs

Example: Toss a die and win X^2 if X is # of spots.

$$E(X^2) = \sum_x x^2 p_X(x) = 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + \cdots + 6^2 \cdot \frac{1}{6}$$

$$= 15\,1/6$$

Note $E(X^2) \neq (E(X))^2$

$$(E(X))^2 = \left(1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6}\right)^2$$

$$= (3.5)^2 = 12\,1/4$$

In general: $E(g(X)) = \sum_x g(x)\, p_X(x)$

or

$= \int g(x)\, f_X(x)\, dx$

Works for function of random vectors:

Let $Y = g(X_1, X_2, \ldots, X_k)$

with pmf $p(x_1, x_2, \ldots, x_k)$

then $E(Y) = E(g(X_1, \ldots, X_k)) = \sum_{x's} g(x_1, \ldots, x_k)\, p(x_1, \ldots, x_k)$

or $$E(Y) = \int \left[\int \cdots \left[g(x_1, \ldots x_k) f(x_1, \ldots x_k) \, dx_k \cdots dx_2 \right] dx_1 \right]$$

if the integral (sum) with $|g|$ converges.

---

p. 124 Corollary A

If X & Y are independent, then

$$E(g(X)\,h(Y)) = E(g(X)) \times E(h(Y))$$

provided ↑ and ↑ exist.

Special case: If X & Y are independent

$$E(XY) = E(X)E(Y)$$

if $\uparrow$ and $\uparrow$ exist.

Beware: 1) not true in general

2) converse not true

Linear combinations of R.V.'s

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_nX_n$$

then $E(Y) = a + b_1E(X_1) + b_2E(X_2) + \cdots + b_nE(X_n)$

if $\uparrow$ $\cdots$ exist.

Example A (p. 126)

What is $E(Y)$ if $Y \sim Bin(n, p)$?

$$E(Y) = \sum_y \binom{n}{y} y\, p^y (1-p)^{n-y}$$

Easier: Use fact that $Y = X_1 + \cdots + X_n$ where X_i's are indep. $Bin(1, p)$ and $E(X_i) = p$

So $E(Y) = \sum_{i=1}^{n} E(X_i) = np$

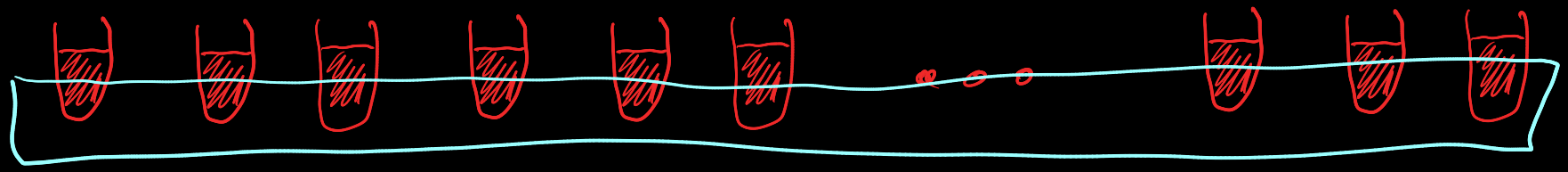
Example C p. 128

Group testing

- n blood samples tested for rare disease

- If you test each individually will need n tests.

Alternative:



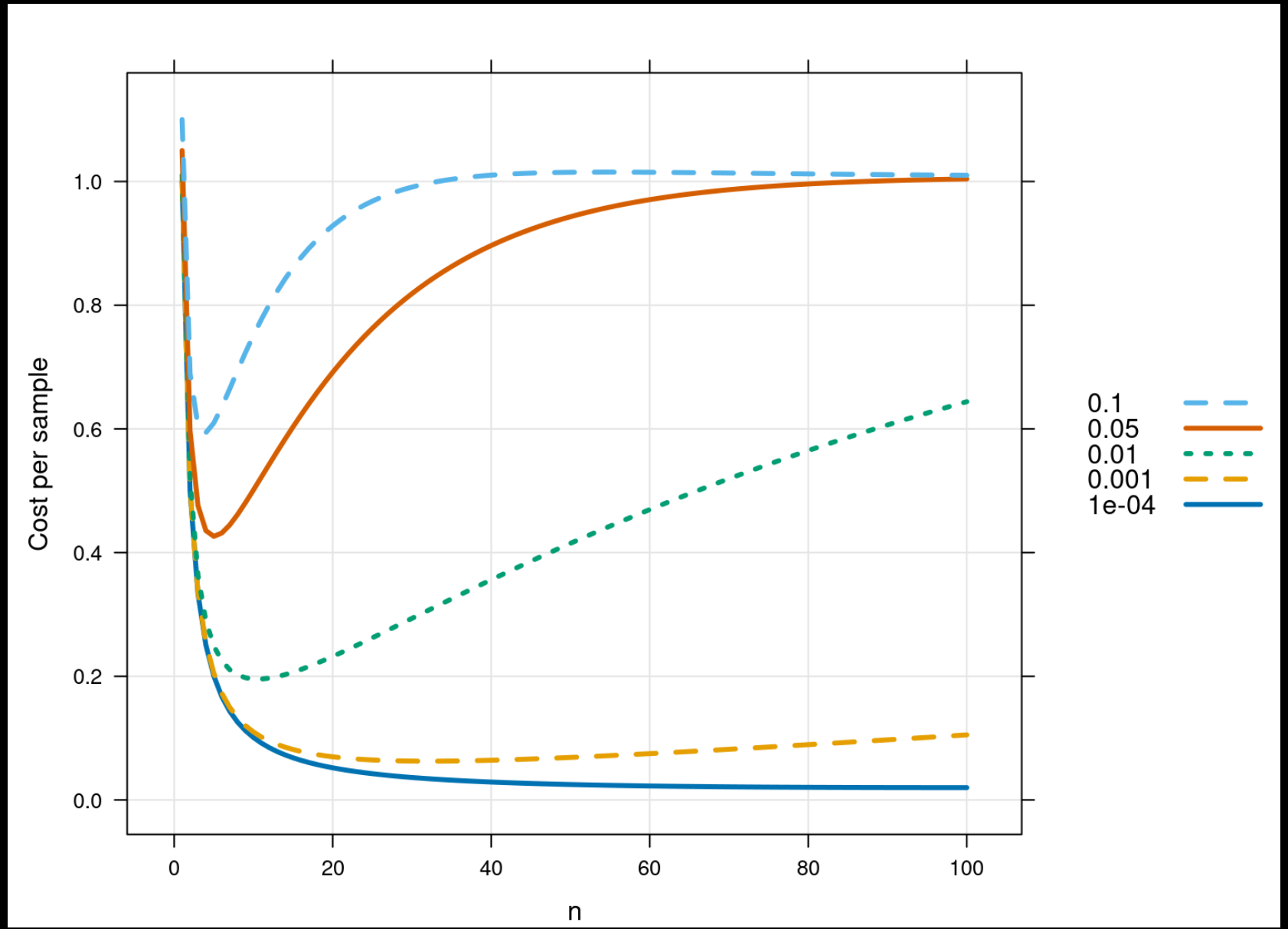
take 1/2 of each sample and combine:

Then test.

If negative - done - all okay
If positive - test n samples.

Let p = probability of a positive.

$$E(\text{Tests}) = 1 \times \underbrace{(1-p)^n}_{\text{pr all negative}} + (n+1)\underbrace{\left(1-(1-p)^n\right)}_{\text{prob. at least 1 positive}}$$


Cost per sample
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
n
0.1
0.05
0.01
0.001
1e-04

```r
df <- expand.grid(n = 1:100, p = c(.0001,.001, .01, .05, .1))
head(df)
dim(df)
df <- within(df,
             {
               Etests <- 1* (1-p)^n + (n + 1) * (1 - (1 - p)^n)
               Cost_per_sample <- Etests/n
             }
             )

library(latticeExtra)
trellis.par.set(superpose.line = list(lwd=3, lty = 1:3))
xyplot(Cost_per_sample ~ n, df,
       groups = p, type = 'l',
       ylab = "Cost per sample",
       auto.key = list(reverse.rows = T)) +
  layer_(panel.grid(h=-1, v = -1))
```

Example D: Illustrates how

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

does not require independence

DNA Sequences: formed from 4 letters A C T G

Of random & each letter with = prob.

ATCAATCGAGT ... TAA

- Suppose length = N
- and each letter has $p = 1/4$

How many ATGC do you expect?

Let I_n = event ATGC starts at position n

$$P(I_n) = \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} = \frac{1}{256}$$

$$= E(I_n)$$

$$\text{if } I_n = \begin{cases} 1 & \text{if ATGC starts at position } n \\ 0 & \text{otherwise} \end{cases}$$

$$E(\#\text{of sequences}) = E\left(\sum_{n=1}^{N-3} I_n\right) = \sum_{n=1}^{N-3} E(I_n)$$

$$= (N-3) \times \frac{1}{256}$$

Variance and Standard Deviation

Mean = "location parameter" ← one of many e.g. median, min, max

Next we need a "spread" parameter

Variance: Average squared distance from the mean

$$Var(X) = E\left[(X-\mu_X)^2\right]$$ if E exists. in squared units

$$SD(X) = \sqrt{Var(X)}$$ in original units

Facts about variance.

$$Var(X) = E(X^2) - (E(X))^2$$

also a useful way to calculate $Var(X)$

Proof $E\left[(X-\mu_X)^2\right]$

$= E(X^2 - 2\mu_X X + \mu_X^2)$

$= E(X^2) - 2\mu_X E(X) + \mu_X^2$

$= E(X^2) - 2E(X)E(X) + (E(X))^2$

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

Corollary: $E(X^2) = E(X)^2$ iff $Var(X) = 0$

Fact: $Var(X) = 0$ iff X is a constant
i.e. $P(X = c) = 1$

Fact: If $Var(X) < \infty$ (i.e. exists)
and $Y = a + bX$
then $Var(Y) = b^2 Var(X)$

Chebyshev's Inequality: Prop re spread

Let X have mean μ and variance σ^2

Then for any $t > 0$:

$$P(|X-\mu| > t) \leq \sigma^2/t^2$$

Proof: Use Markov's inequality on $Y = (X-\mu)^2$

Different form: If $\sigma > 0$, let $t = k\sigma$ for $k > 0$. Then

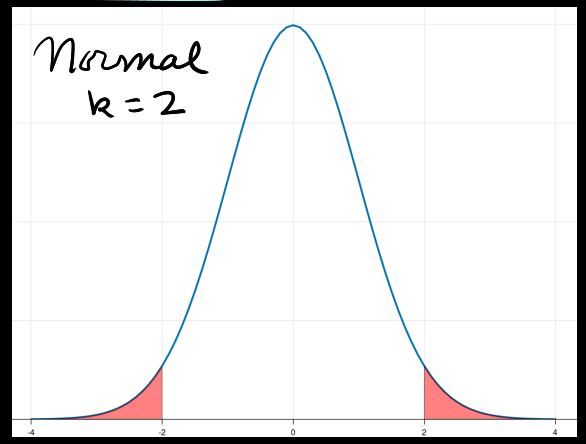
$$P(|X-\mu| > k\sigma) \leq \frac{\sigma^2}{k^2\sigma^2} = 1/k^2$$

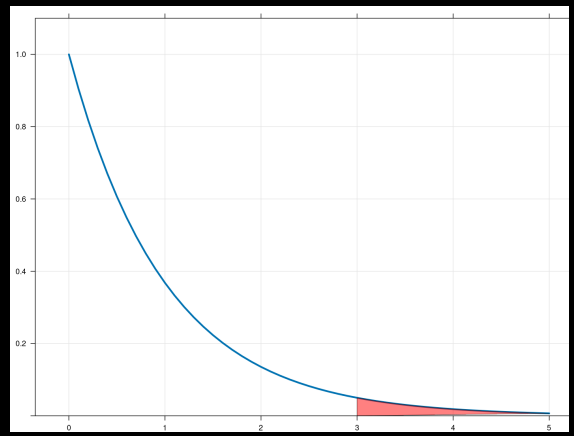
For $k = 1$ this says $P(|X-\mu| \geq \sigma) \leq 1/1^2 = 1$!?

Only useful for $k > 1$; e.g. $P(|X-\mu| \geq 2\sigma) \leq \frac{1}{2^2} = \frac{1}{4}$

Let's compare with some actual values.

	$k=2$	$k=3$	$k=4$
Chebyshev	$\leq 1/4 = 0.25$	$\leq 1/9 = 0.1111\ldots$	$\leq \frac{1}{16} = 0.0625$
Actuals			
Normal	0.0455	0.0027	0.0000633
Exponential	0.0498	0.0183	0.00674
Poisson	0.0803	0.0196	0.00366





Fact (Corollary A, p. 134)

$$\mathrm{Var}(X) = 0 \iff P(X = \mu) = 1$$

$$\iff P(X \text{ is constant}) = 1$$

Proof: Example of Chebyshev's Theorem.

$$P(|X - \mu| > \varepsilon) \leq \sigma^2/\varepsilon^2 = 0$$

for every $\varepsilon > 0$.

So $P(|X - \mu| = 0) = 1$

i.e. $P(X = \mu) = 1$

Investment Portfolios (p. 134)

Two investments: R_1: $E(R_1) = \mu_1 = 0.10$

$\sigma_1 = 0.075$

R_2: $E(R_2) = \mu_2 = 0.03$

$\sigma_2 = 0$

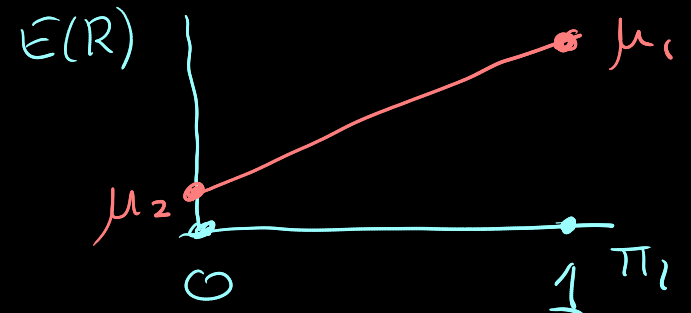
Same cost so investor can choose to invest π_1 in R_1 and $\pi_2 = (1 - \pi_1)$ in R_2

Return is a R.V.:

$$R = \pi_1 R_1 + (1 - \pi_1) R_2$$

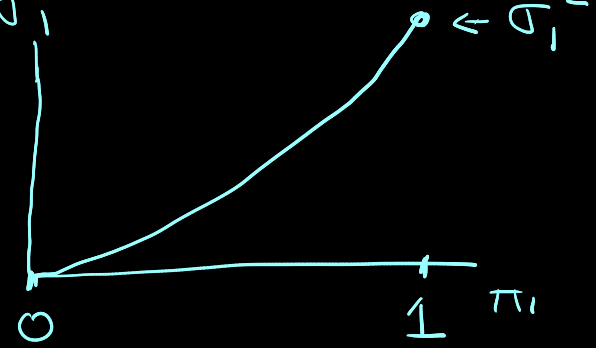
$$E(R) = \pi_1 \mu_1 + (1 - \pi_1) \mu_2$$

Expected return

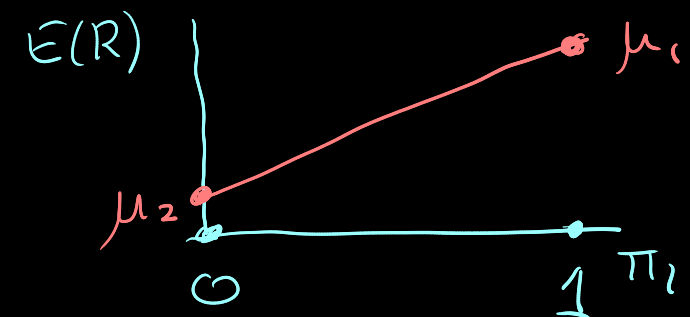


$$Var(R) = \pi_1^2 \sigma_1^2$$

Risk

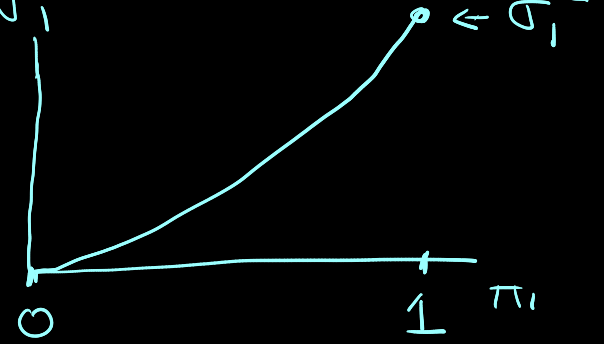


$$E(R) = \pi_1 \mu_1 + (1 - \pi_1)\mu_2$$



$$Var(R) = \pi_1^2 \sigma_1^2$$

Risk



Need to balance expected return against risk.

Notes:
- When $\sigma_2 > 0$ we will also need to consider covariance
- Variance is not the whole story as far as risk is concerned.

Read 4.2.1 Measurement error

Suppose we want to measure some "fixed" but "unknown" quantity.

Call it "x_0"

Measurement $X = x_0 + \beta + \varepsilon$

- x_0: fixed unknown
- β: systematic error - same each time
- ε: random error varies with each measurement.

β = "bias"

ε = random component: $\mathrm{Var}(\varepsilon) = \sigma^2$

$E(\varepsilon) = 0$

Then $Var(X) = Var(\underbrace{x_0 + \beta}_{\text{constant}} + \varepsilon)$

$= Var(\varepsilon) = \sigma^2$

But $E(Error^2) = MSE$

$= E[(X - x_0)^2]$

$= E[(\beta + \varepsilon)^2]$

$= E[\beta^2 + 2\beta\varepsilon + \varepsilon^2]$

$= \beta^2 + 2\beta \underbrace{E(\varepsilon)}_{=0} + E(\varepsilon^2)$

$= Var(X) + (E(X))^2 = Var(X) = \sigma^2$

"MSE" $= \beta^2 + \sigma^2 = Bias^2 + \sigma^2$

4.3 Covariance → Correlation

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

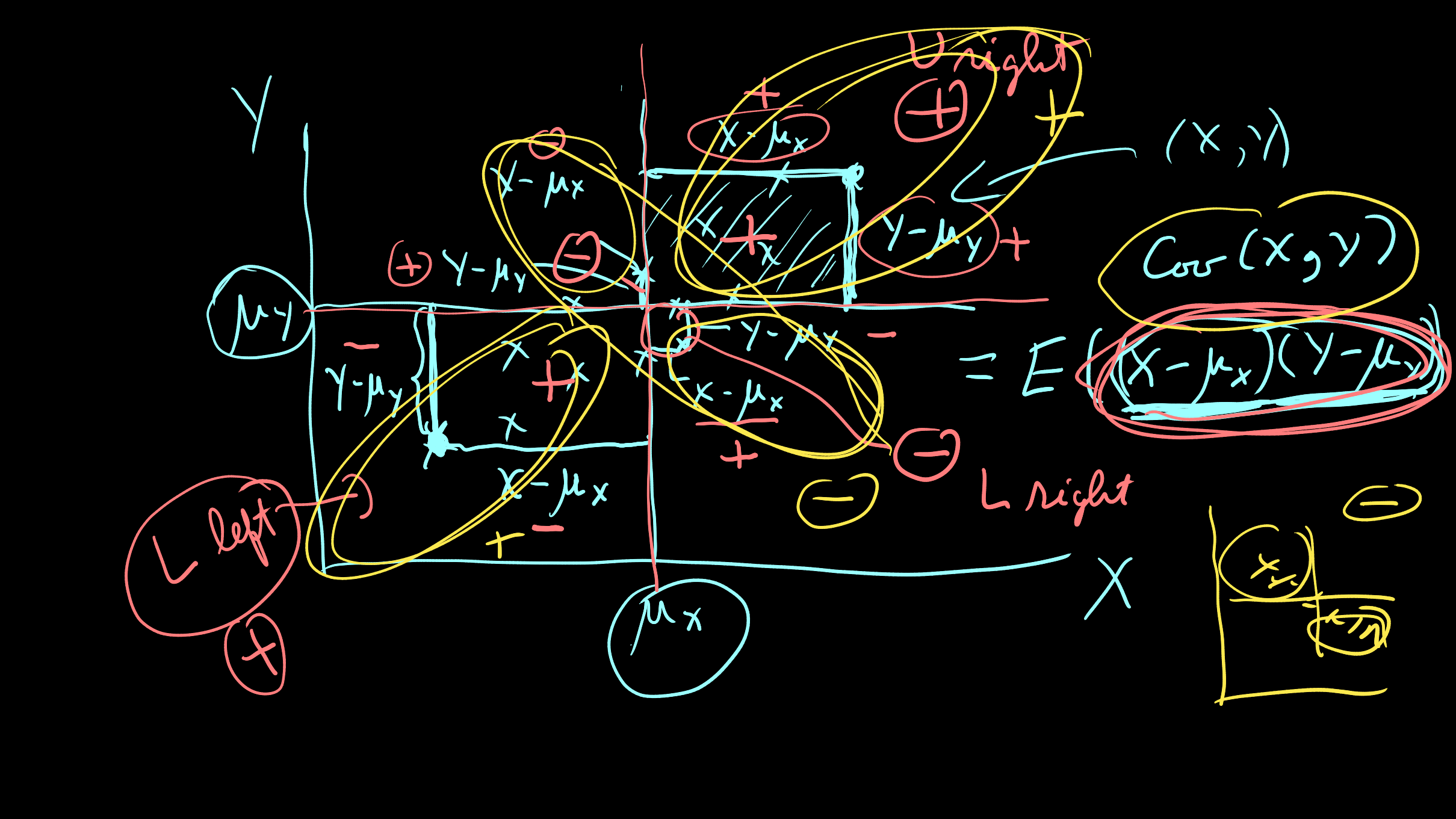
"with" — if E exists. "times" is correct here

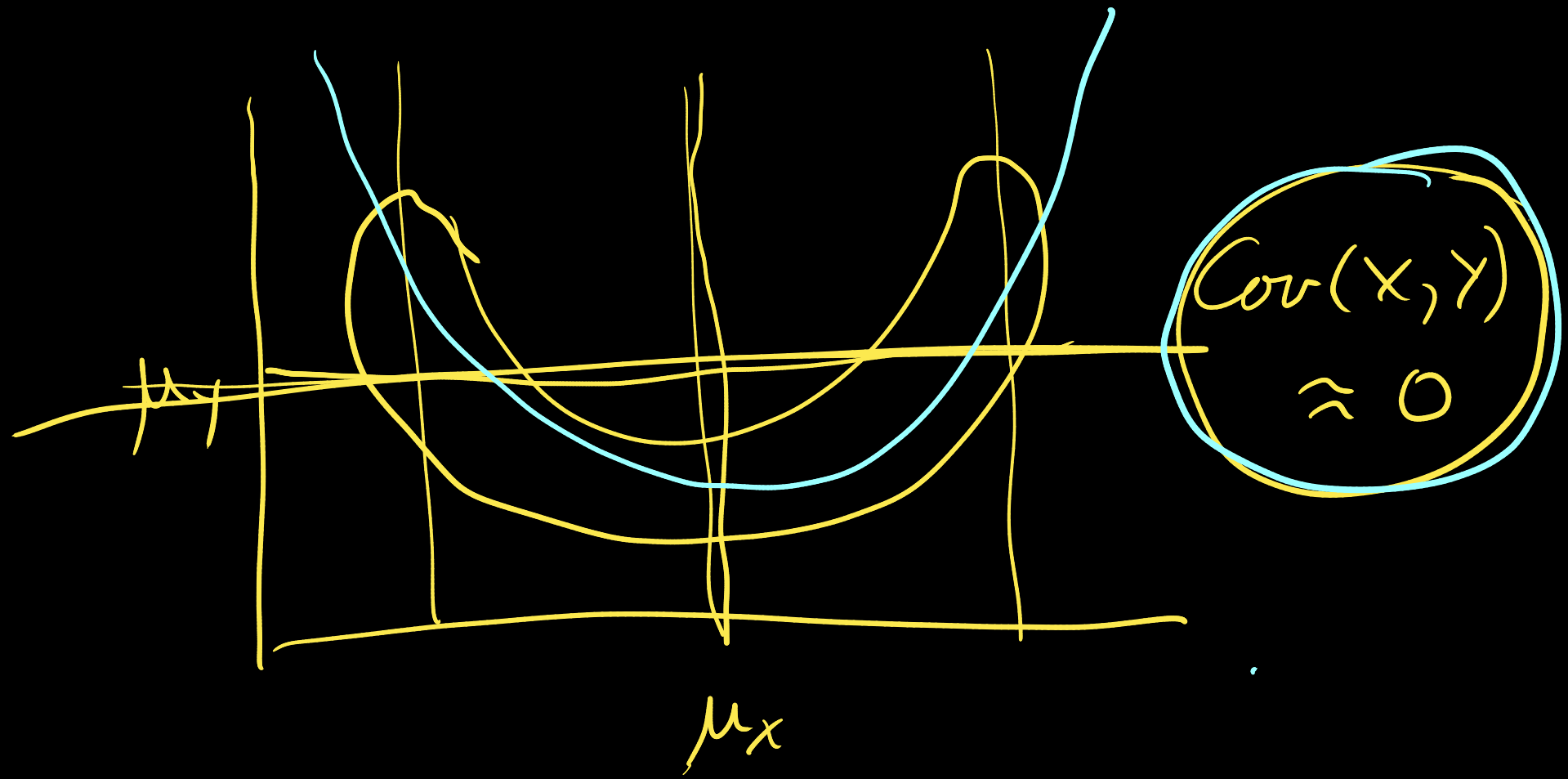
Fact: $\mathrm{Cov}(X, Y) = E\left[XY - \mu_X Y - \mu_Y X + \mu_X \mu_X\right]$

$$= E(XY) - \mu_X \mu_Y - \mu_Y \mu_X + \mu_X \mu_Y$$

$$= E(XY) - E(X)E(Y)$$

Never say Covariance of X ~~times~~ Y BUT always X "with" Y


Y
X
μ_Y
μ_X
$Y-\mu_Y$
$X-\mu_X$
U right
L left
L right
(X, Y)
$= E[(X-\mu_X)(Y-\mu_Y)]$
Cov(X, Y)


Cov(X,Y)
≈ 0
μ_Y
μ_X

(X, Y) joint Normal.

$X \perp Y$ $\iff$ $Cov(X, Y) = 0$

iff X & Y are independent

Note: $Cov(X, X) \stackrel{?}{=} E((X-\mu_X)(X-\mu_X))$
$= Var(X)$

Facts

$Cov(a+X, b+Y) = Cov(X, Y)$ (0?)

a, b constants

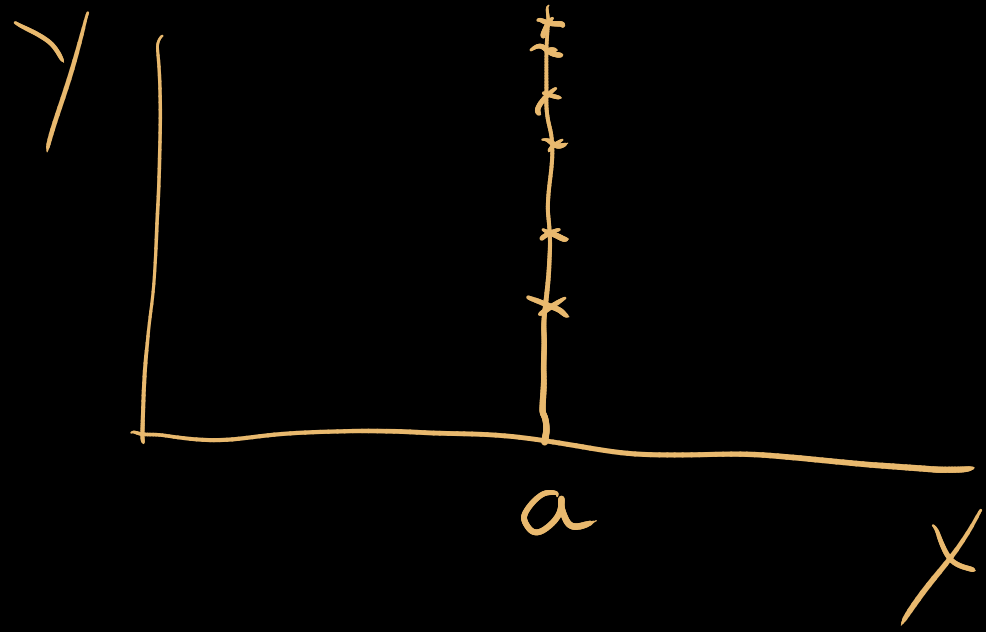
$Cov(aX, bY) = ab\, Cov(X, Y)$

$Cov(X, Y+Z)$
$= Cov(X, Y) + Cov(X, Z)$

$Cov(X+W, Y) = Cov(X, Y) + Cov(W, Y)$

$$Cov(a, Y) = a\,\underline{Cov(\quad, Y)}$$

not defined



$$M_X = a$$

$$E(a) = a$$

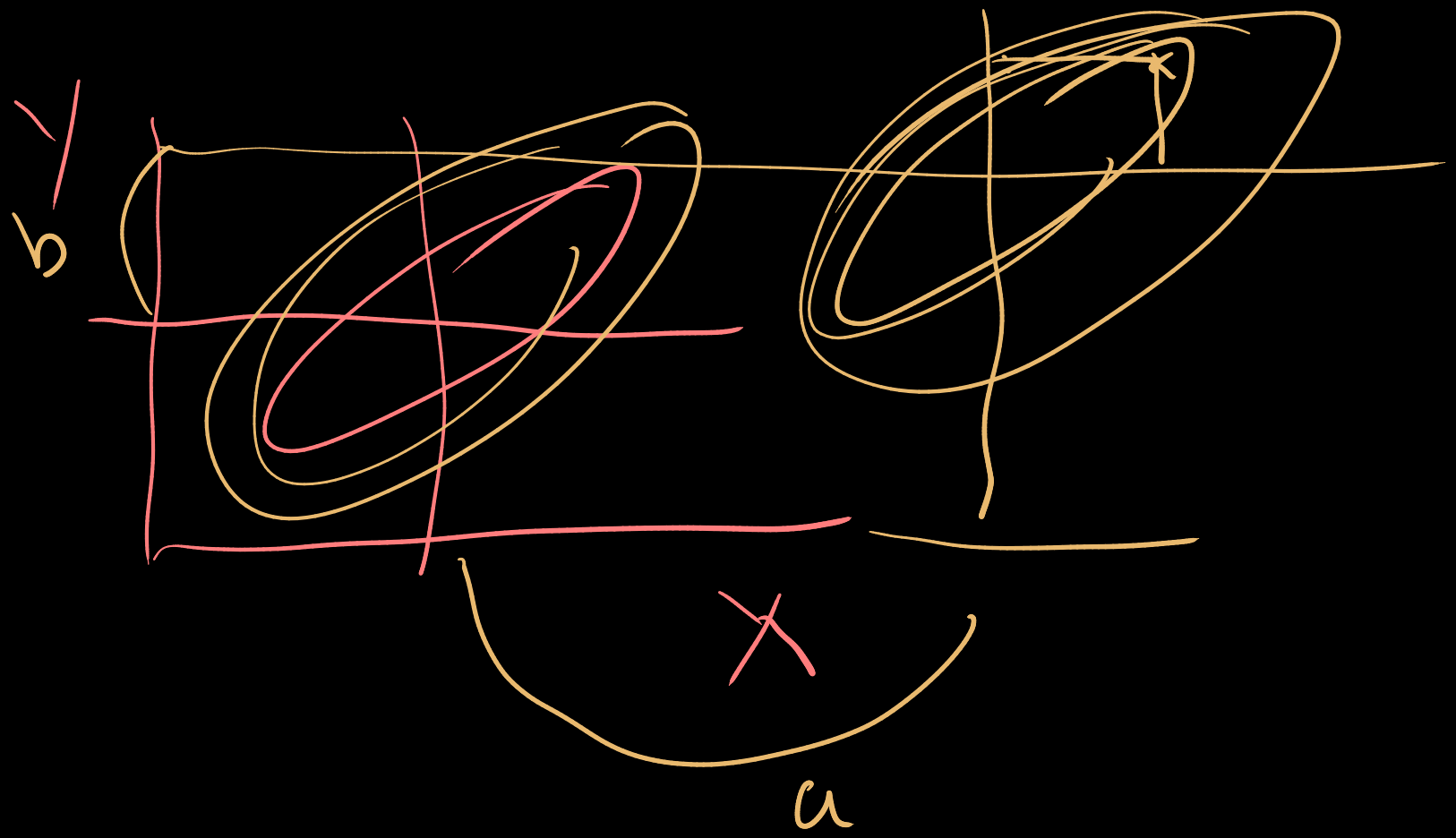
$$Cov(X, Y)$$

with

$$Cov(X, X) = Var(X)$$

$$\boxed{E((X-\mu_X)(Y-\mu_Y)) = 0}$$

$$\underline{(X-\mu_X)} \to 0$$


b
Y
X
a

$$\mathrm{Cov}(a + b_1 X_1 + b_2 X_2, Y)$$

$$= \underbrace{\mathrm{Cov}(a, Y)}_{0} + b_1 \mathrm{Cov}(X_1, Y) + b_2 \mathrm{Cov}(X_2, Y)$$

linear in left argument

" " right "

bilinear

$$\mathrm{Cov}(aW+bX,\ cY+dZ)$$

with

$$= ac\,\mathrm{Cov}(W,Y) + ad\,\mathrm{Cov}(W,Z)$$
$$+ bc\,\mathrm{Cov}(X,Y) + bd\,\mathrm{Cov}(X,Z)$$

Summary:

Linearity of Expectation:

If $Y = a + b_1 X_1 + \cdots + b_n X_n$

then

$$E(Y) = a + b_1 E(X_1) + \cdots + b_n E(X_n)$$

In other words

If $Y = a + \sum_{i=1}^{n} X_i$

then $E(Y) = a + \sum_{i=1}^{n} E(X_i)$

Bilinearity of Covariance

$$\mathrm{Cov}\left(X, a + \sum_{i=1}^{n} b_i Y_i\right)$$

$$= \sum_{i=1}^{n} b_i \,\mathrm{Cov}(X, Y_i)$$

$$\mathrm{Cov}\left(a + \sum_{i=1}^{n} b_i X_i,\ Y\right)$$

$$= \sum_{i=1}^{n} b_i \,\mathrm{Cov}(X_i, Y)$$

$\mathrm{Cov}(a, Y) = 0$

$$\mathrm{Cov}\left(a + \sum_{i=1}^{n} b_i X_i \,,\, c + \sum_{j=1}^{m} d_j Y_j\right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} b_i d_j \, \mathrm{Cov}(X_{(i)}, Y_{(j)})$$

this should look like matrix multiplication

$$= \underbrace{\begin{bmatrix} b_1 & b_2 & \cdots & b_n \end{bmatrix}}_{n \text{ rows}} \times$$

$$\underbrace{\begin{bmatrix} Cov(X_1,Y_1) & Cov(X_1,Y_2) & \cdots & Cov(X_1,Y_m) \\ Cov(X_2,Y_1) & Cov(X_2,Y_2) & & \vdots \\ \vdots & \vdots & & \\ Cov(X_n,Y_1) & Cov(X_n,Y_2) & \cdots & Cov(X_n,Y_m) \end{bmatrix}}_{m \text{ cols.}} \underbrace{\begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_m \end{bmatrix}}_{n \times 1}$$

$$(1 \times n) \times (n \times m) \times (m \times 1) = [\ ]_{1 \times 1}$$

vector · matrix · vector · scalar

Matrix formula:

conformable

Let $\underset{\sim}{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$; $C = \left[Cov(X_i, Y_j) \right]$ $n \times m$ matrix; $\underset{\sim}{d} = \begin{bmatrix} d_1 \\ \vdots \\ d_m \end{bmatrix}$

$Cov\left(\underset{\sim}{b}^T \underset{\sim}{X}, \underset{\sim}{d}^T \underset{\sim}{Y} \right) = \underset{\sim}{b}^T C \underset{\sim}{d}$

$\underset{\sim}{b} \cdot \underset{\sim}{X}$

$b_1 X_1 + b_2 X_2 + \cdots + b_m X_m$

$(b_1 \cdots b_m) \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$

$\underset{\sim}{b}$ and $\underset{\sim}{d}$ could be matrices

$\mathrm{Cov}(\underset{\sim}{b}\cdot\underset{\sim}{X}, \underset{\sim}{d}\cdot\underset{\sim}{Y})$

$\underset{\sim}{b}^T\underset{\sim}{c} = \underset{\sim}{b}\cdot\underset{\sim}{c}$

$$\mathrm{Cov}(B\underset{\sim}{X}, D\underset{\sim}{Y}) = B\,\mathrm{Cov}(\underset{\sim}{X},\underset{\sim}{Y})\,D^T$$

Fact:

If X & Y are independent

then $\mathrm{Cov}(X, Y) = 0$

Proof: Recall: if X, Y are independent then $E(g(X) \cdot h(Y))$

$= E(g(X))\, E(h(Y))$

Let $g(x) = X - \mu_X$, $h(Y) = Y - \mu_Y$

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

$$= E(X - \mu_X) \times E(Y - \mu_Y)$$

$$= [\mu_X - \mu_X] \times [\mu_Y - \mu_Y]$$

$$= 0 \times 0$$

$$= 0$$

Important to remember:

If (X, Y) are bivariate normal

then X and Y are independent

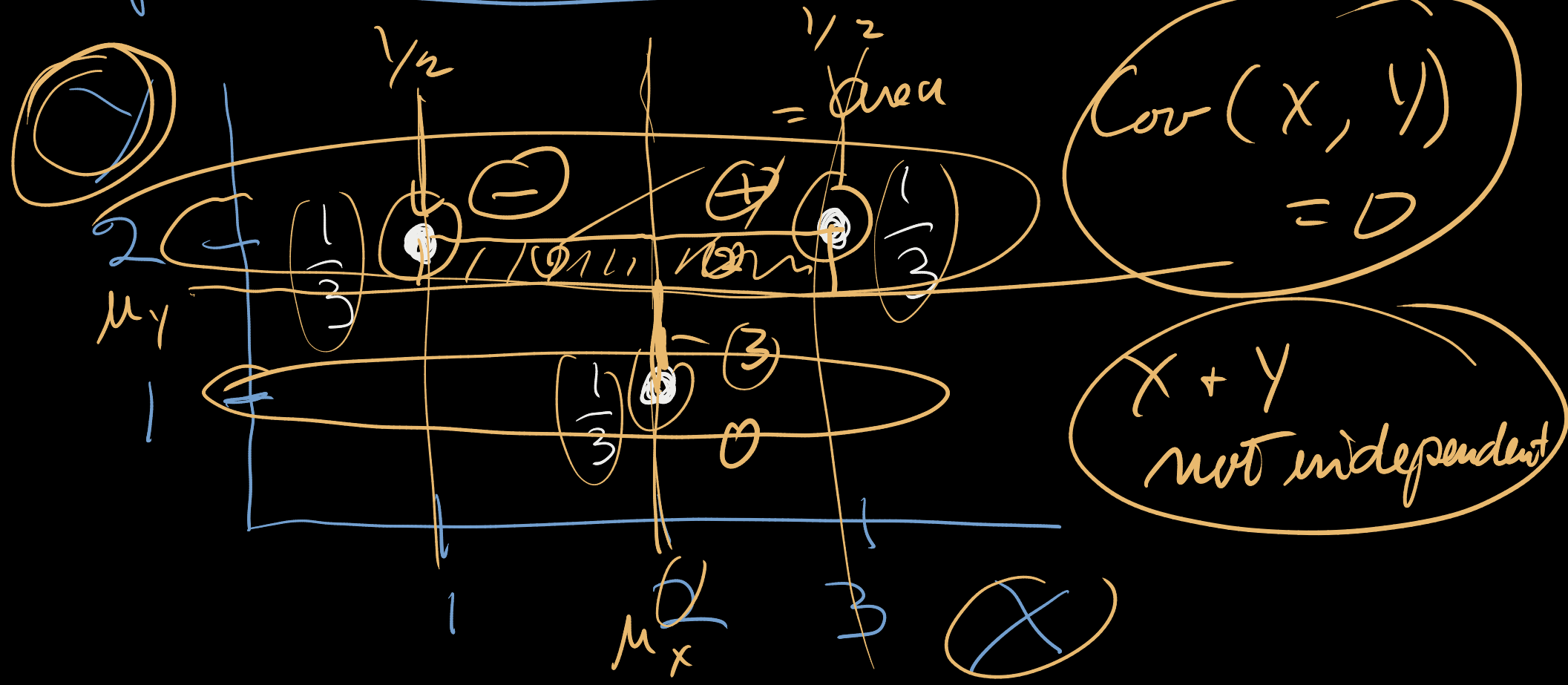
iff $\mathrm{Cov}(X, Y) = 0$

In general: X, Y independent implies $\mathrm{Cov}(X, Y) = 0$

BUT $\mathrm{Cov}(X, Y) = 0$ DOES NOT IMPLY INDEPENDENCE

Simple example

counter



$Cov(X, Y) = 0$

$X + Y$ not independent

IMPORTANT RESULT

If $X_1, \ldots, X_n$ are R.V.s with $\operatorname{Cov}(X_i, X_j) = 0$, $i \neq j$

then

$$\operatorname{Var}(X_1 + \cdots + X_n) = \operatorname{Var}(X_1) + \cdots + \operatorname{Var}(X_n)$$

Corollary: If $X_1, \ldots, X_n$ are independent, then

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}(X_i)$$

NOTE: NOT TRUE IN GENERAL

Recall:

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

whether X_i's are independent or not

Example B (p. 140)

$X \sim \text{Binomial}(n, p)$

What is $\text{Var}(X)$?

$X = X_1 + \cdots + X_n$

where X_i's are

independent Bernoulli(p)

$E(X_i) = p \cdot 1 + (1-p) \cdot 0 = p$

$E(X_i^2) = p \cdot 1^2 + (1-p) \cdot 0^2 =$

$$Var(X_i) = E(X^2) - (E(X))^2$$

$$= p - p^2$$

$$= p(1-p)$$

$$= pq \text{ where } q = 1-p$$

$$\text{So } Var(X) = \sum_{i=1}^{n} (Var(X_i)) = npq$$

Correlation:

"Pure" measure of covariance

X = height in cm

Y = weight in kg

then $\mathrm{Cov}(X, Y)$ is in "cm × kg"

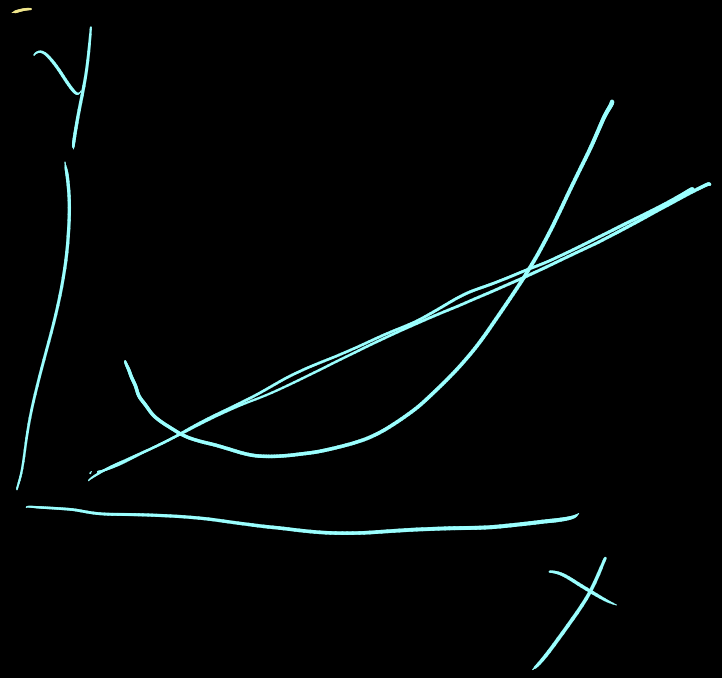
If we change X to $X' = X/2.54$

so X' is in inches, then

$$\mathrm{Cov}(X', Y) = \mathrm{Cov}(X, Y)/2.54$$

in units of "in $\times$ kg"

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{SD(X) \times SD(Y)}$$



$$= \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \times \text{Var}(Y)}}$$

$$= \rho_{XY} = \rho$$

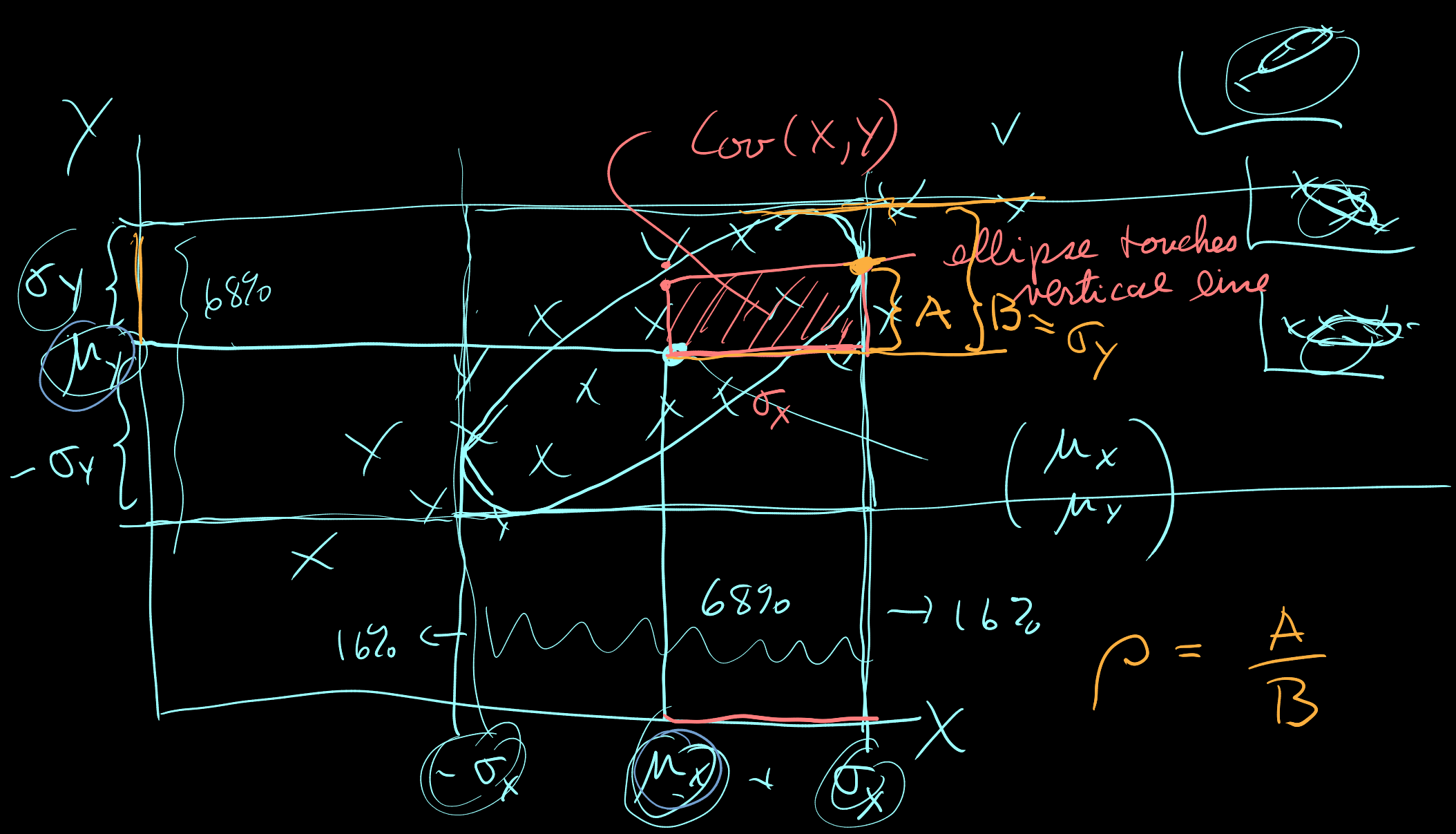
Theorem: $|\rho| \leq 1$ i.e $-1 \leq \rho \leq 1$

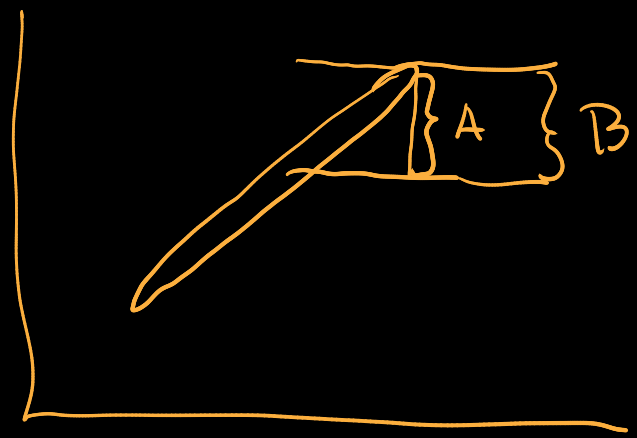
and $\rho = \pm 1$ iff $P(Y = a + bX) = 1$ for some $b \neq 0$

See proof p. 143 = Cauchy-Schwarz Theorem

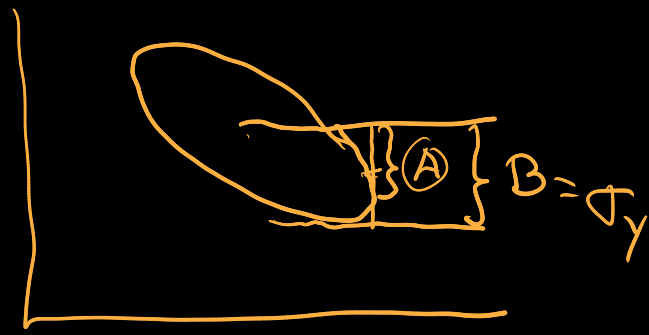
Fact: ρ_{XY} stays the same if X and/or Y are rescaled or translated.

e.g. inches to feet, $X' = X/12$; $^\circ C \to ^\circ F$, $X' = \frac{9}{5}X + 32$

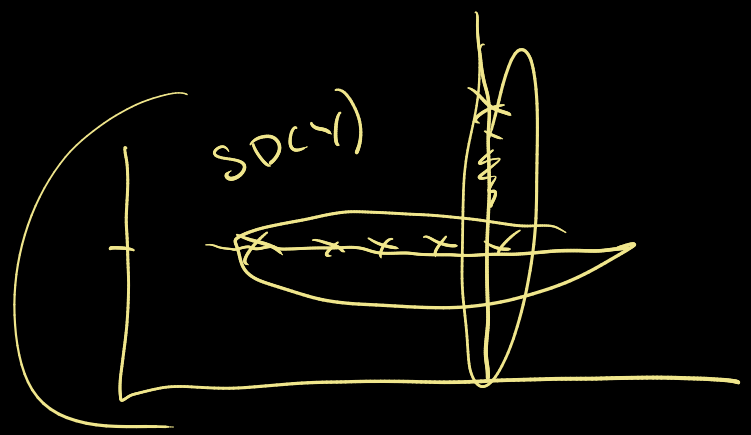

Y
Cov(X,Y)
V
ellipse touches
vertical line
A
B ≈ σY
σY
68%
μY
−σY
σX
(μX
μY)
16%
68%
16%
X
−σX
μX
+
σX
ρ = A/B



$$\rho = \frac{A}{B} \approx .9$$



$$\rho = \frac{A}{B} = -.75$$



$$\rho_{XY} = \frac{Cov(X,Y)}{SD(X)\,SD(Y)} = 0$$

$$= ?$$

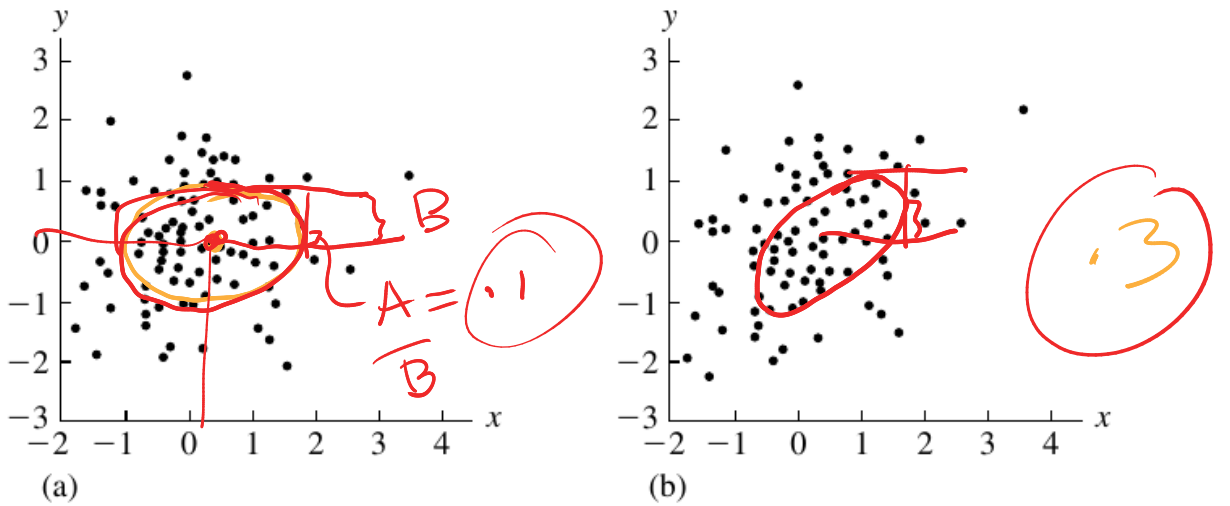


FIGURE 4.7 Scatterplots of 100 independent pairs of bivariate normal random variables, (a) $\rho = 0$, (b) $\rho = .3$, (c) $\rho = .6$, (d) $\rho = .9$.

$-1 \le \rho \le 1$

$Y = a + bX$

$b > 0$

$\rho = 1$

$\rho = -1$

Conditional Expectation & Prediction

$$E(Y \mid X = x) = \sum_y y \, p_{Y|X}(y \mid x) \qquad p_Y(y)$$

$$E(h(Y) \mid X = x) = \int h(y) f_{Y|X}(y \mid x)\, dx$$

Bivariate Normal

See book for formula w/o matrices

k = dimension $\quad p = 2 \quad k/2 = 1$

$$\underset{\sim}{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$$

$$f(x,y) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \times \sqrt{|\Sigma|}$$

$$\exp\left\{ -\frac{1}{2} \left[\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \right]^T \Sigma^{-1} \left[\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \right] \right\}$$

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_y^2(1-\rho^2)}} \exp\left\{-\frac{1}{2}\frac{\left[(y-\mu_y)-\rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)\right]^2}{\sigma_y^2(1-\rho^2)}\right\}$$

$$Y|X=x \sim N\left(\underbrace{\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)}_{E(Y|X)}, \underbrace{\sigma_y^2(1-\rho^2)}_{Var(Y|X)}\right)$$

Var(Y)

$$0 \le 1-\rho^2 \le 1$$

Var(Y|X) =

No info $\rho = 0$ then $1-\rho^2 = 1$ and $\sigma_2^Y(1-\rho^2) = \sigma_2^Y$

Perfect Pred $\rho = 1$ $\quad 1-\rho^2 = 0$ $\quad \sigma_2^Y(1-\rho^2) = 0$

$\rho = 1/2$ $\quad 1-\rho^2 = 3/4$ $\quad \sigma_2^Y(1-\rho^2) = \frac{3}{4}\sigma_2^2$

$\rho = .9$ $\quad 1-\rho^2 = 0.19$ $\quad \sigma_2^Y \cdot 0.19$

Theorem A : Mean = Mean Conditional Mean

marginal $E_Y(Y) = E_X\left[E_{Y|X}(Y|X)\right]$

Proof: $E(Y|X) = \sum_y y\, P_{Y|X}(y|x)$

$E(Y|X)$

$E(Y|X=x)$

$= \sum_y y \frac{P(x,y)}{P_X(x)}$

$E_X(E(Y|X)) = \sum_x E(Y|X=x)\, P_X(x)$

$= \sum_x \left[\sum_y y \frac{P(x,y)}{P_X(x)}\right] P_X(x)$

Let X = spots rolling die

$P(X=3)$

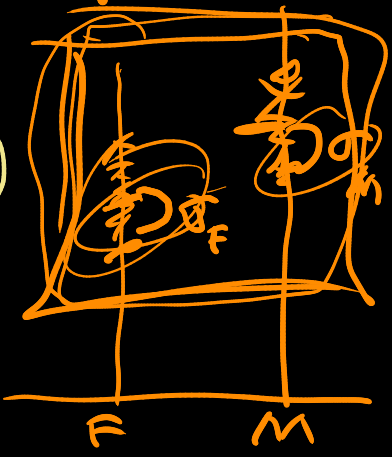
(Y, X)

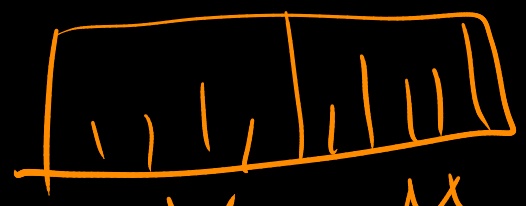
Y

$Y|X$

$X|Y$

X





w Δx

$E(Y)$ = cue height

$E(Y \mid G=f) = \mu_f$

$E(Y \mid G=m) = \mu_m$

$$= \sum_{x,y} y\, p(x,y)$$

$$= \sum_{y} y\, p_Y(y)$$

$$= E(Y)$$

$$E(Y) = E(Y \mid G=f)\, P(G=f) + E(Y \mid G=m)\, P(G=m)$$

Theorem B Var = Var of Cond. Mean + Mean of Cond. Variance

$$\mathrm{Var}(Y) = \mathrm{Var}(E(Y \mid X)) + E(\mathrm{Var}(Y \mid X))$$

Proof: Cond'l variance

$$\mathrm{Var}(Y \mid X=x) = E(Y^2 \mid X=x) - [E(Y \mid X=x)]^2$$

$$E(\mathrm{Var}(Y|X)) = \sum_{x} \mathrm{Var}(Y|X=x)\, p_X(x)$$

$$= \left(\sum_{x} E(Y^2|X=x)\, p_X(x) - \sum_{x} \left(E(Y|X=x)\right)^2 p_X(x) \right)$$

$$= E(E(Y^2|X)) - E\left(\left[E(Y|X)\right]^2\right)$$

$$\mathrm{Var}(E(Y|X)) = E\left(\{E(Y|X)\}^2\right) - \{E(E(Y|X))\}^2$$

$$\mathrm{Var}(Y) = E(Y^2) - (E(Y))^2$$

$$= E_X\left(E(Y^2|X)\right) - \{E(E(Y|X))\}^2$$

$$= E(E(Y^2|X)) - E\{[E(Y|X)]^2\} + E\{[E(Y|X)]^2\} - \{E(E(Y|X))\}^2$$

$$= E[\mathrm{Var}(Y|X)] + \mathrm{Var}[E(Y|X)]$$

Prediction

If you know X,
what's your best guess of Y?

"Best?" - Many criteria
- Depends on what kind of variable Y is.
- Cost of error functions.
- etc. etc.

Start easy - much more in future courses.

minimize MSE$_{\text{error}}$ = $E[(Y - \text{guess})^2]$
can guess

- We have a joint distribution for

$$(X, Y)$$

- We observe X and want a function $h(x)$ to make

$$E[(Y - h(X))^2]$$

as small as possible.

Answer: $h(x) = E(Y|X)$

Proof:

Step 1: Predicting Y with a single value "c".

$$MSE = E[(Y-c)^2]$$

$$= Var(Y-c) + \{E(Y-c)\}^2$$

$$= \underbrace{Var(Y)}_{\text{fixed}} + \underbrace{\{E(Y)-c\}^2}_{\text{minimized by letting } c = E(Y)}$$

Step 2 :

$$E\left[(Y-h(X))^2\right]$$

mean of cond'l mean

$$= E_X\left\{E\left[(Y-h(X))^2 \mid X\right]\right\}$$

$$= \sum_x E_{Y|X=x}\left[(Y-h(x))^2\right] p_X(x)$$

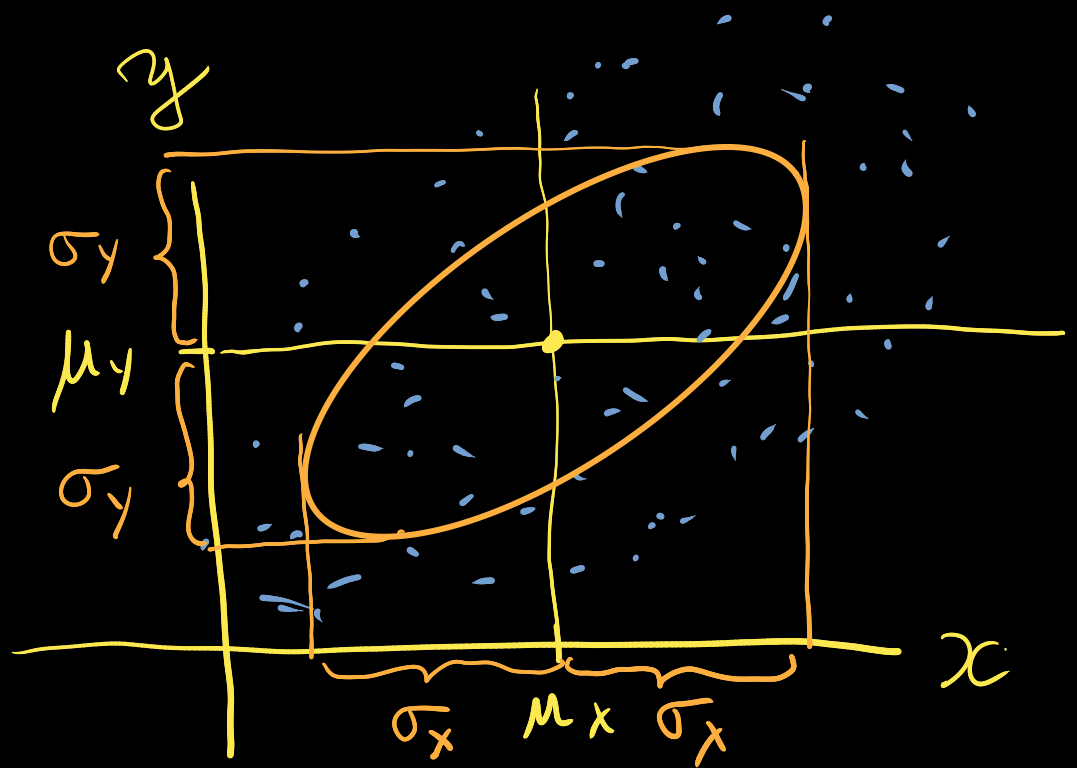
for each x this expectation is minimized by taking $h(x) = E(Y|X=x)$

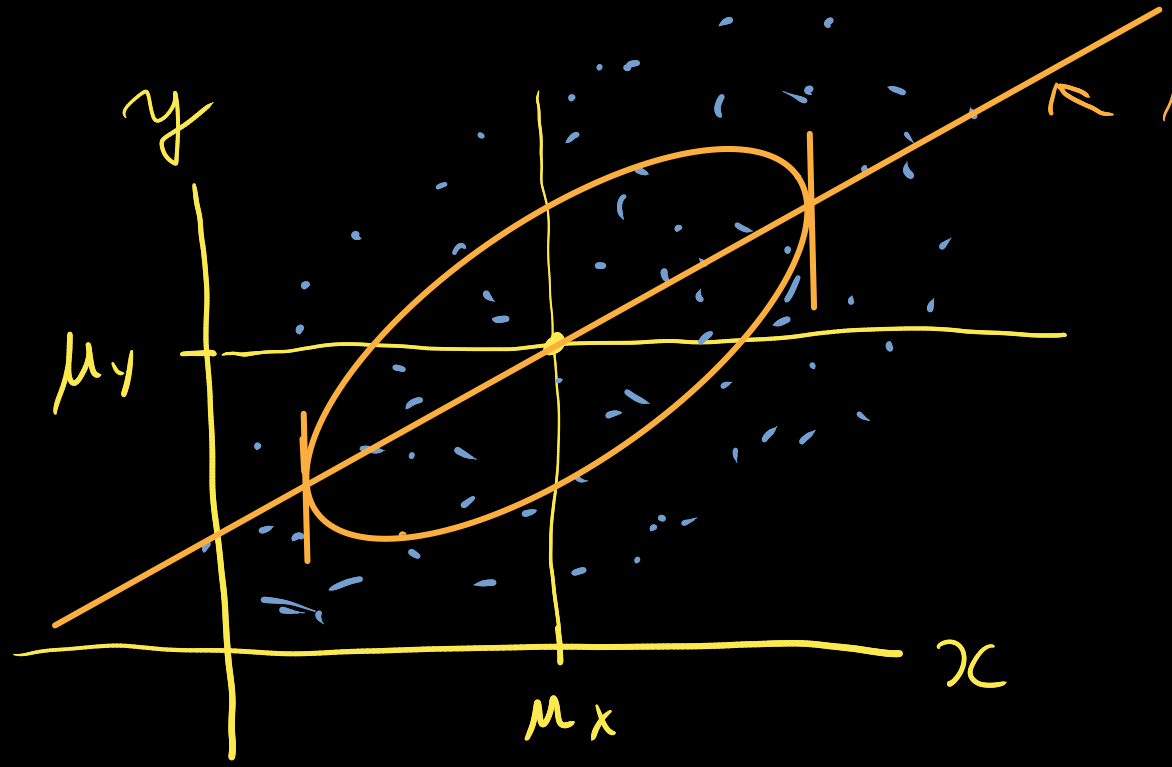
So $h(X) = E(Y|X)$ minimizes MSE

Special Case : Bivariate normal

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2\left(\underbrace{\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}}_{\text{mean}}, \underbrace{\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}}_{\text{variance-covariance matrix}} \right)$$

or just the "variance matrix"





← regression line : $E(Y|X=x)$

Fact 1 :

$$\text{Slope} = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$$

$$= \frac{\rho_{XY}\,\sigma_X\,\sigma_Y}{\sigma_X^2}$$

$$= \rho_{XY}\,\frac{\sigma_Y}{\sigma_X}$$

Fact 2 :

Goes through (μ_x, μ_y)

So $E(Y|X=x) = a + bx$

Fact 1: $b = \frac{\sigma_{xy}}{\sigma_x^2}$

From Fact 2: $\mu_y = a + b\mu_x$

$= a + \frac{\sigma_{xy}}{\sigma_x^2}\mu_x$

So $a = \mu_y - \frac{\sigma_{xy}}{\sigma_x^2}\mu_x$

Easier to remember :. Equation of regression line

(1) $(\hat{Y} - \mu_Y) = \frac{\sigma_{XY}}{\sigma_X^2}(X - \mu_X)$

(2) $(\hat{Y} - \mu_Y) = \frac{\rho_{XY} \times \sigma_Y}{\sigma_X}(X - \mu_X)$

(3) $\frac{(\hat{Y} - \mu_Y)}{\sigma_Y} = \rho_{XY}\left(\frac{X - \mu_X}{\sigma_X}\right)$

(4) $\hat{Z}_Y = 0 + \rho_{XY} Z_X$