Applied Statistics

Lecture 11+12

David Barton & Sabine Hauert

Department of Engineering Mathematics

Outline

- Independence testing
- Goodness-of-fit testing

$\Big\}$ Pearson's Chi^2 test

- Joint distribution
- Marginal distribution
- Conditional distribution
- Bivariate normal distribution

OpenIntro Statistics

Chapter 6, particularly §6.3 and §6.4

Independence testing

Pearson's χ^2 test can also be used to test for independence.

Are men and women aged 15–64 equally accident prone?

Data from one hospital with injuries divided by male/female for people aged 15–64 (HASS dataset; 2002)

Injury mechanism	Male	Female
Fall on same level	126	210
Struck — moving object	112	103
Struck — static object	152	179
Cut/tear (sharp)	264	171
Foreign body	99	61
Total	753	724

is the probability of having a particular type of accident independent from gender

H_0: the probabilities are independent

Independence testing

Use the probabilities (calculated assuming male/female have the same probabilities — independence) to calculate the expected values.

Injury mechanism	Total	Probability	E_{male} (753·0.227)	E_{female} (724·0.227)
Fall on same level	126+210= 336	336/1477= 0.227	170.9	164.3
Struck — moving object	215	0.146	109.9	105.7
Struck — static object	331	0.224	168.7	162.2
Cut/tear (sharp)	435	0.295	222.1	213.6
Foreign body	160	0.108	81.3	78.2
Total	1477	1.0	752.9 (rounding error)	724.0

Degrees-of-freedom is given by $(\text{rows} - 1) \times (\text{columns} - 1)$
$= (5-1)(2-1) = 4$ *not the number of categories minus one!*

Calculating things from the data (the probabilities) adds constraints, which reduce the number of degrees-of-freedom

Independence testing

The test statistic is the same as before

observed expected Falls for men Falls for women

$$\sum_i \frac{(O_i - E_i)^2}{E_i} = \frac{(126-170.9)^2}{170.9} + \frac{(210-164.3)^2}{164.3} + \frac{(112-109.9)^2}{109.9}$$
$$+ \frac{(103-105.7)^2}{105.7} + \frac{(152-168.7)^2}{168.7} + \frac{(179-162.2)^2}{162.2}$$
$$+ \frac{(264-222.1)^2}{222.1} + \frac{(171-213.6)^2}{213.6} + \frac{(99-81.3)^2}{81.3}$$
$$+ \frac{(61-78.2)^2}{78.2} = 52.05$$

The critical value $\chi^2_{\underbrace{4}_{\text{dof}}}(0.05) = 9.488 < 52.05$, hence we reject H_0.

There is a statistically significant difference between men and women with regard to the accidents they have!

$$\chi^2 = \frac{(14-15.6)^2}{15.6} + \frac{(16-14.4)^2}{14.4} + \frac{(12-10.4)^2}{10.4} + \frac{(8-9.6)^2}{9.6} = 0.855$$

$\chi^2_{1} = 2.7055$ $(2-1)(2-1) = 1$ dof $\rightarrow$ can not reject H_0

Exercise

Is the probability of improvement independent from the skin cream used?

50 people, who are suffering from a skin rash, are the test set for a new skin ointment, to evaluate whether the new treatment appears effective. 30 are given the usual skin cream and 20 are given the new ointment. The results are as follows.

Treatment		Improved (obs)		Not improved (obs)	
Usual skin cream	$30 \cdot 26/50$ = 15.6	14	$30 \cdot 24/50 = 14.4$	16	30
New ointment	10.4	12	9.6	8	20
		26/50		24/50	50

Investigate whether these results lead you to conclude that the new ointment is more effective that the original skin cream at a 10% significance level (the null hypothesis is that the two treatments are equally effective). H_0: both creams are equally effective. The ability to improve skin is independent from the skin cream used.

Goodness-of-fit testing

One final use of Pearson's χ^2 test is to determine if data is taken from a particular distribution (e.g., normal or Poisson).

Days spent in hospital following an accident for children aged 5–14 years

	0	1	2	3	4	5	
Days in hospital	0–2	3–5	6–10	11–20	21–30	31+	Total
Frequency	463	47	18	12	0	6	546

observed data

Does the data follow a Poisson distribution?

descriptive — relables data in expanded way

Goodness-of-fit testing

What are the expected values if they follow a Poisson distribution

Calculate the mean from the data (all that is needed for Poisson — using a Normal distribution would require the standard deviation as well)

Use the mid-points of the intervals for the calculations. (Actually, will relabel the data since we only need a descriptive model.)

$$\bar{x} = \frac{1}{546}(0 \times 463 + 1 \times 47 + 2 \times 18 + 3 \times 12 + 4 \times 0 + 5 \times 6)$$
$$= 0.273$$

$546 \cdot P(\underset{\text{days}}{0}) = 546 e^{-0.273} \cdot \frac{0.273^0}{0!} = 415.6$ $\quad 0! = 1$

The mean can be used to generate expected values from the Poisson distribution.

Days in hospital	0–2 (0)	3–5 (1)	6–10 (2)	11–20 (3)	21–30 (4)	31+ (5)	Total
Expected	415.6	113.4	15.5	1.4	0.1	0.0	546

merge

< 5

Goodness-of-fit testing

Pearson's χ^2 test is only an approximate test, to work as expected each of the expected values should be greater than 5. Any categories with smaller numbers should be merged together.

The final table is thus (first two categories merged due to small expected numbers)

18+12+0+6

Days in hospital	0–2	3–5	6+	Total
Observed	463	47	36	546
Expected	415.6	113.4	17.0	546

Goodness-of-fit testing

The final calculation is

obs exp

$$\sum_i \frac{(O_i - E_i)^2}{E_i} = \frac{(463 - 415.6)^2}{415.6} + \frac{(47 - 113.4)^2}{113.4} + \frac{(36 - 17.0)^2}{17.0}$$

$$= 65.5$$

Number of degrees of freedom = number of categories (3) minus number of calculated quantities (1 — the mean) minus one = 1; hence use $\chi^2_1(0.05) = 3.841 < 65.5$ and reject H_0.

dof

dof = 3 − 1 − 1 constraint on the mean

categorie always do

Constraints (imposed by calculating expected values from the observed)

- Calculated mean for generating the Poisson distribution

For a normal distribution, you need the mean and variance → subtract 2 from dof

Exercise

$x = (0 \cdot 63 + 28 \cdot 1 + 8 \cdot 2 + 1 \cdot 3)/100 = 0.47$

The number of pages containing 0, 1, 2, 3, . . . misprints in a 100-page magazine were counted, with the results shown below.

$100 \cdot e^{-0.47} \cdot \frac{0.47^1}{1!} = 29.32$

	Number of misprints	0	1	2	$\geqslant 3$	Total
Obs	Number of pages	63	28	8 9	1	100
Expected		$100 \cdot p(0) = 100 e^{-0.47} \frac{0.47^0}{0!} = 62.5$				

$100 - 62.5 - 29.32 = 8.13$

The probability of a misprint is small and the number of pages large, so it seems reasonable that the Poisson distribution would be an appropriate model. Use hypothesis testing to find out if the Poisson distribution is appropriate (at a 5% significance level).

$\chi^2_1 = 3.841$

Note 1: to handle $\geqslant$3 we note that the frequencies must add to 100.
Note 2: the last data column has an expected frequency less than 5.

$$\chi^2 = \frac{(63-62.5)^2}{62.5} + \frac{(28-29.32)^2}{29.32} + \frac{(9-8.13)^2}{8.12} = 0.5$$

χ^2 values

	Significance level		
D.o.F.	5%	1%	0.1%
1	3.841	6.635	10.828
2	5.991	9.210	13.816
3	7.815	11.345	16.266
4	9.488	13.277	18.467
5	11.070	15.086	20.515
6	12.592	16.812	22.458
7	14.067	18.475	24.322
8	15.507	20.090	26.124
9	16.919	21.666	27.877
10	18.307	23.209	29.588

More than one random variable

So far we've considered a single random variable

- either a continuous random variable, or
- a discrete random variable (categories).

What happens when we have more than one random variable?

- Simplest case — independent random variables

$$P(A \text{ and } B) = P(A \cap B) = P(A)P(B)$$

- More difficult case — dependent random variables

$$P(A \text{ and } B) = P(A \cap B) = P(A)P(B|A) = P(A|B)P(B)$$

$\neq P(B)$ $\quad \neq P(A)$

Exam results

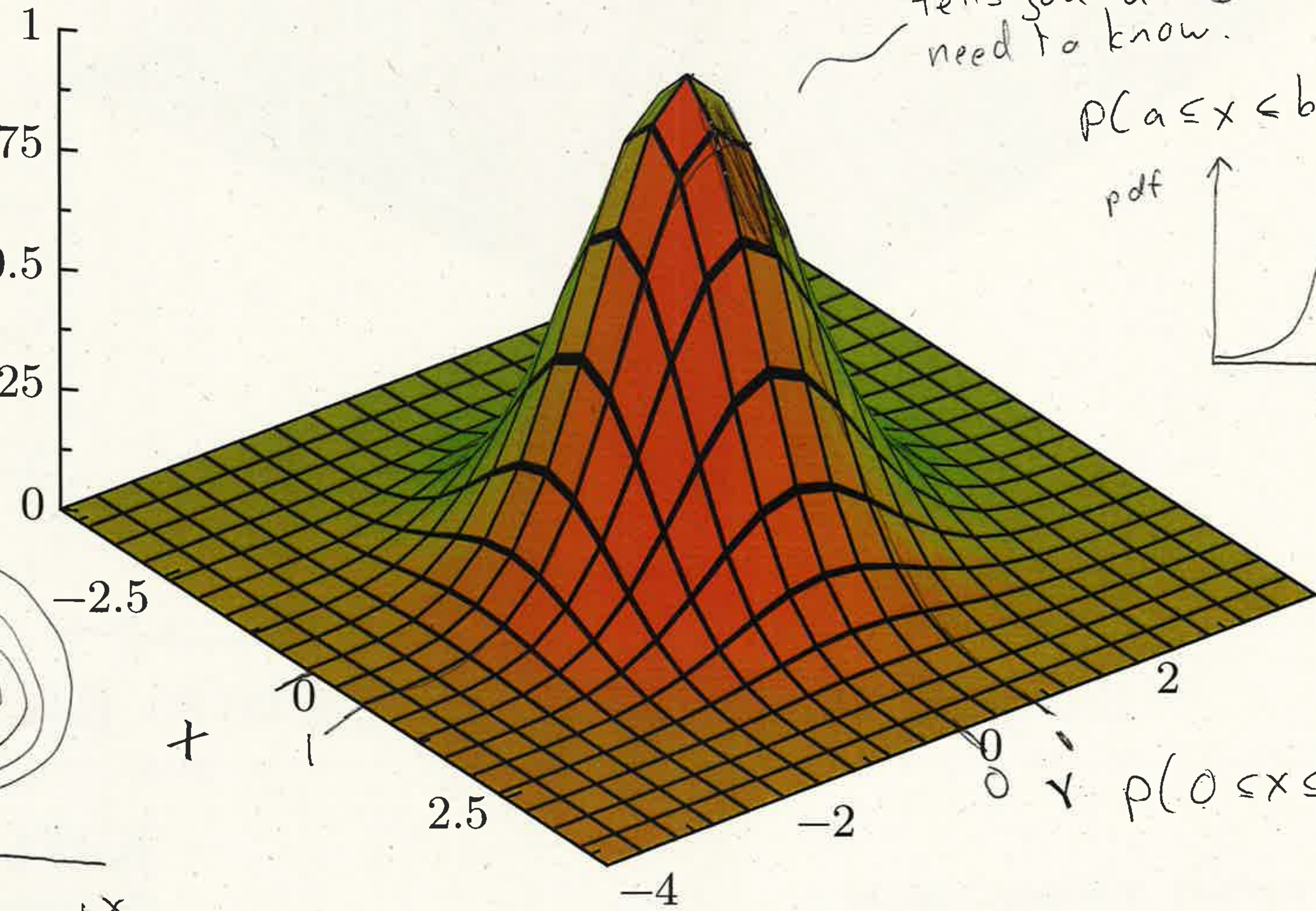
If you get good A-level grades, how likely are you to get good grades in your first year at university?

Questions you might be interested in. . .

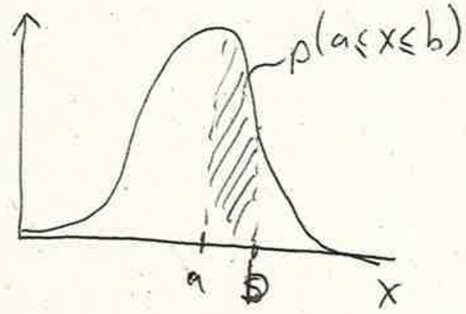
- What is the probability of getting 70+% in the first year?
- If I get A^*AA, what is the probability of getting 70+%?
- What is the probability of getting A^*A^*A at A-level and then failing the first year?

Independent normal — joint distribution

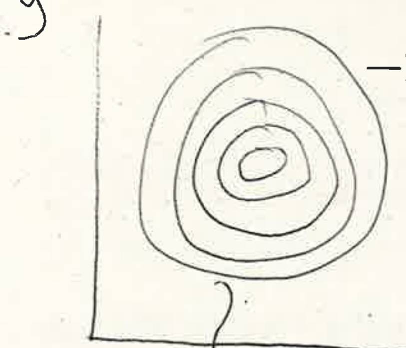
tells you all ~~some~~ the probabilities you need to know.



$p(a \le x \le b) = \int_a^b pdf(x)\,dx$



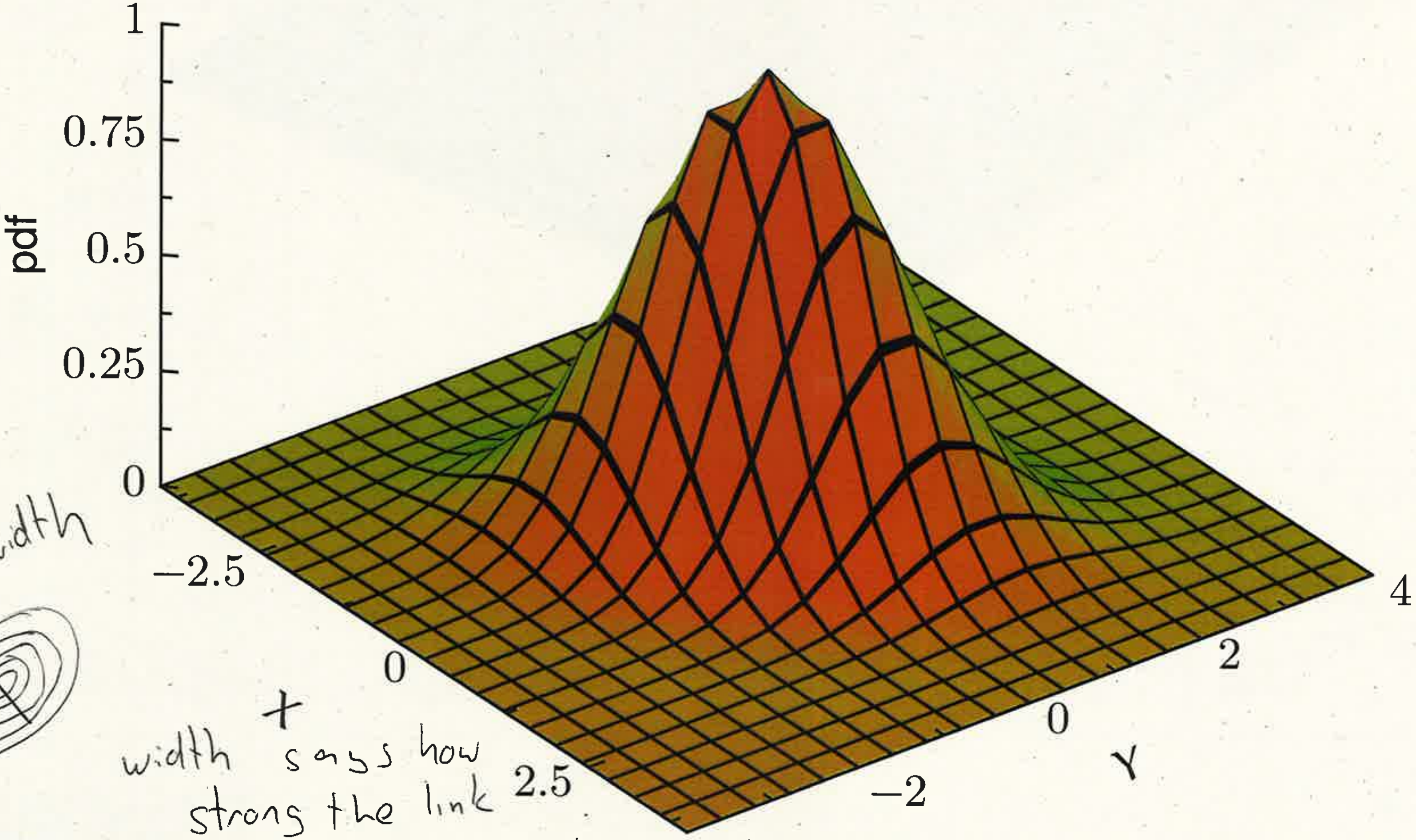
$p(0 \le x \le 1 \text{ and } 0 \le y \le 1)$

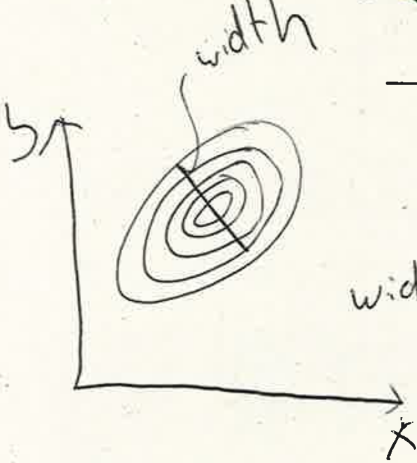


curves of equal probability

Covariant normal — joint distribution

└ dependent ⇒ link between X and Y

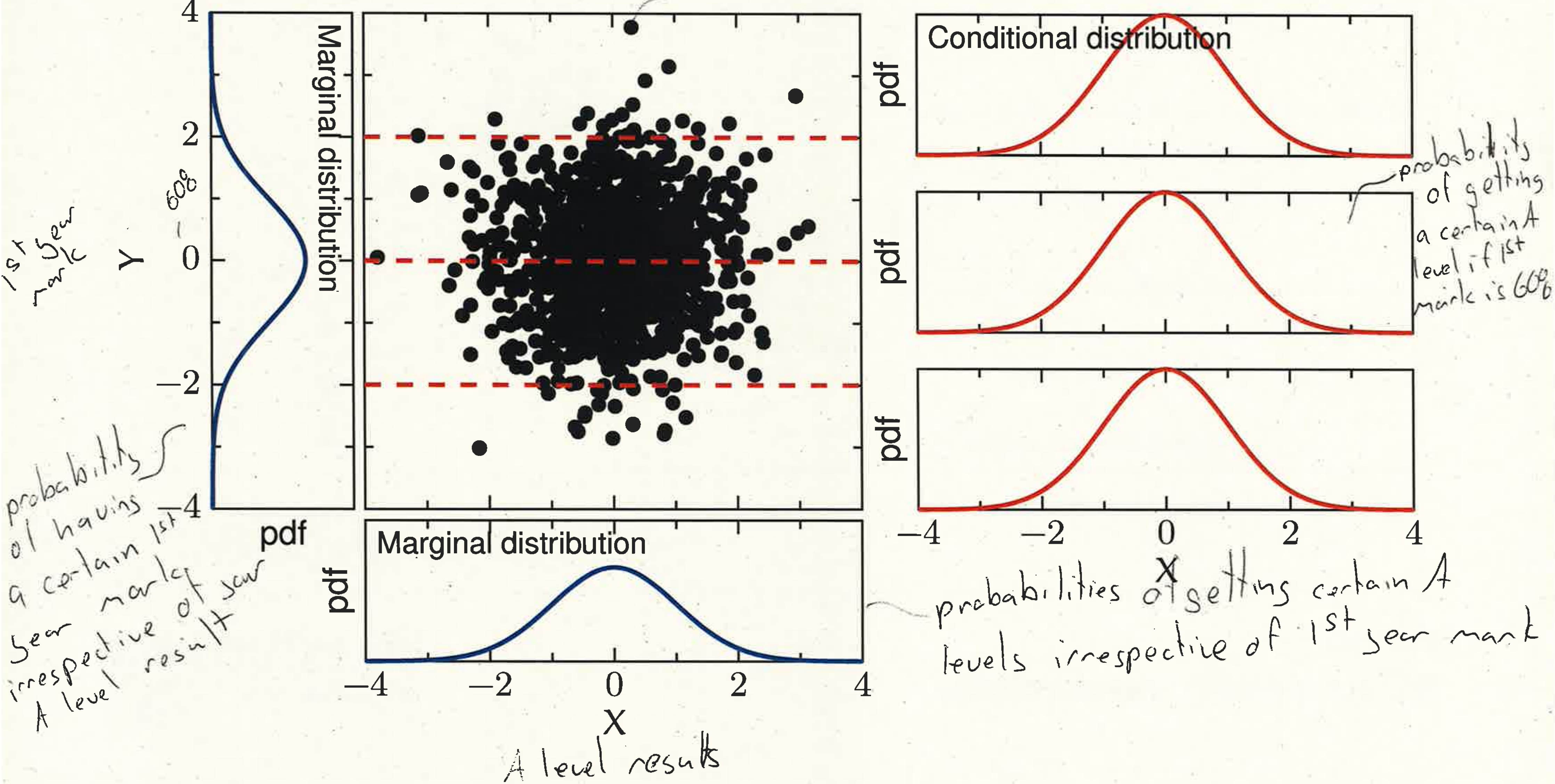




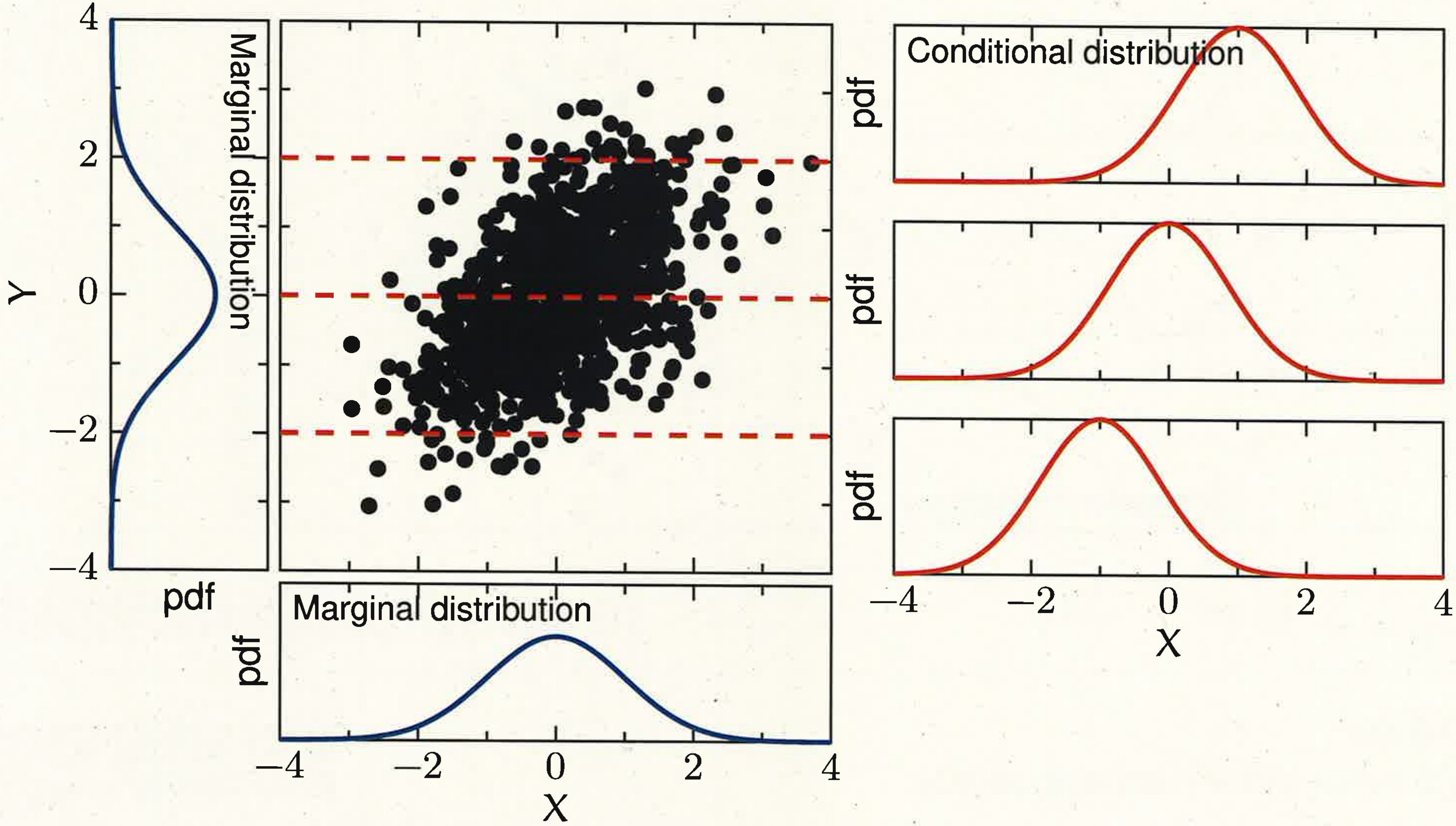
width says how strong the link between x and y is
narrow link ⇒ strong link

Independent normal

data points
Lhistogram gives 3D joint distribution



Covariant normal



Joint distribution

all the info is there but sometimes difficult to gather

The *joint distribution* $p_{X,Y}(x, y)$ is the key distribution from which all others can be derived —

Volume integral to get the probability.

$$P\left((a_X \leqslant X \leqslant b_X) \text{ and } (a_Y \leqslant Y \leqslant b_Y)\right) = \int_{a_X}^{b_X} \int_{a_Y}^{b_Y} p_{X,Y}(x, y)\, dy\, dx$$

If the random variables X and Y are independent, then we have that

$$P\left((a_X \leqslant X \leqslant b_X) \text{ and } (a_Y \leqslant Y \leqslant b_Y)\right) = P(a_X \leqslant X \leqslant b_X) P(a_Y \leqslant Y \leqslant b_Y)$$

which implies that

true if independent

$$p_{X,Y}(x, y) = p_X(x) p_Y(y)$$

but this is only for *independent random variables*.

Marginal distribution

The *marginal distribution* is the distribution of one of the variables, ignoring what the other variable is doing. To find it, integrate over all values of the other variable

joint distribution

$$p_X(x) = \int_y p_{X,Y}(x,y)\,dy$$

all the values of y

This is called *marginalisation*.

This name comes from having two discrete random variables and writing out the probabilities of all possibilities in a table. The sums of the rows/columns (which give the marginal distribution) end up being written in the margins.

Marginal distribution — example

A pedestrian crossing at traffic light controlled junction but ignoring the colour of the lights; do they get hit by a car? [Wikipedia]

	Traffic light colour			
	Red	Amber	Green	Total
Not hit	0.198	0.09	0.14	0.428
Hit	0.002	0.01	0.56	0.572
Total	0.2	0.1	0.7	1

probability of getting hit by a car

probabilities of lights being a certain color

There are two variables and so two marginal distributions. These tells us

1. the probability that someone will be hit (ignoring the colour they crossed on) and
2. the probability that someone will cross on red (ignoring whether they will be hit or not).

Conditional distribution

The *conditional distribution* is the distribution for one variable when the other variable takes a specific value. That is

joint distribution

condition

$$p_X(x|Y = y) = \frac{\overbrace{p_{X,Y}(x, y)}^{\text{joint distribution}}}{p_Y(y)}$$

probability of getting a first A given A*A*A.

marginal distribution

where $p_Y(y)$ is the marginal distribution.

At one level this concept is relatively easy but it's actually not as straightforward as it seems (see the Borel's paradox) — we'll stick with the intuitive definition. . .

Conditional distribution — example

Back to the traffic light example

joint distribution

marginal distribution

	Traffic light colour			
	Red	Amber	Green	Total
Not hit	0.198	0.09	0.14	0.428
Hit	0.002	0.01	0.56	0.572
Total	0.2	0.1	0.7	1

Probability of outcomes while crossing on amber

$$P(\text{Hit}|\text{Amber}) = \frac{P(\text{Hit and Amber})}{P(\text{Amber})} = \frac{0.01}{0.1} = 0.1$$

$$P(\text{Not hit}|\text{Amber}) = \frac{P(\text{Not hit and Amber})}{P(\text{Amber})} = \frac{0.09}{0.1} = 0.9$$

Bivariate normal distribution

For a bivariate normal distribution all you need to specify is the mean μ of each random variable and the *covariance matrix* Σ

Normal distribution is completly defined by mean and variance.

$$(X, Y) \sim N(\mu, \Sigma) = N\left(\underbrace{\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}}_{\text{two means}}, \begin{bmatrix} \sigma_X^2 & \boxed{\sigma_{X,Y}} \\ \boxed{\sigma_{X,Y}} & \sigma_Y^2 \end{bmatrix}\right)$$

two variance values

covariance
↳ how x and y vars together.
0 if independent.
↳ no link.

The *variances* are defined as normal (using expected values)

expected value mean

$$\sigma_X^2 = E\left[(X - \mu_X)^2\right] = \frac{1}{n-1} \sum (x - \bar{x})^2$$

$$\sigma_Y^2 = E\left[(Y - \mu_Y)^2\right]$$

and the *covariance* $\sigma_{X,Y}$ is given by

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

Increase the dimensions of μ and Σ for multi-variate normal distributions.

University of BRISTOL

Exercise

Bivariate normal distribution

The covariance matrix Σ is (like the mean μ and variance σ^2) a deterministic quantity since it is a function of expected values rather than estimated values.

In the previous figures, the covariance matrix was

$$\Sigma = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$$

We will look at estimating the *sample covariance matrix* from data next lecture. The sample covariance matrix is a random variable since it is estimated from random samples.