Statistics and Machine Learning 1

Lecture 7A: Probabilistic Reasoning

Mark Muldoon
Department of Mathematics, Alan Turing Building
University of Manchester

Week 7

Main theme

In statistical modelling, the more you make things probabilistic, the more you can do.

1. Modelling data probabilistically
   - Joint, conditional & marginal distributions
   - Statistical predictions *are* conditional probabilities
2. Treating the parameters of a statistical model probabilistically
   - Changing one's mind in light of data, *Bayes*-style
   - A simple, but useful, example: estimating a proportion
   - What to try in harder cases:
     - the *Laplace approximation*
     - *Markov-Chain Monte-Carlo* (MCMC, coming next semester)
3. An introduction to *Gaussian Processes*
   - A highly flexible, "nonparametric" family of statistical models.
   - It enables us to do inference on functions

Rules of probability, revisited on the Titanic

If A and B are probabilistic events then:

$P(A)$ and $P(B)$ are the probabilities that A and B happen

$P(A, B)$ is the probability that *both* A and B happen

$P(A|B)$ which is read as “the probability of A given B”, is the probability that A happens, given that B has.

Consider a single passenger, chosen randomly and with uniform likelihood from the dataset about the Titanic that you are working with in coursework 2. If we take our events to be

A : the passenger is male

B : he survived

then we can use R’s `xtabs()` function to draw up a helpful table.

An example: the Titanic dataset, revisited

```
> # Read and attach the data, forcing Survived to be a factor
> titanic.df <- read.csv( "titanic.csv", header=TRUE )
> attach(titanic.df)
> Survived <- factor(Survived, labels=c("no", "yes") )
>
> # Make a useful table
> xtabs( ~ Survived + Sex )
        Sex
Survived female male
     no      81  464
     yes    233  109
```

A randomly selected passenger thus has:

$$
\begin{aligned}
P(\text{male}) &= (464+109)/(81+233+464+109) &&= 573/887 \\
P(\text{survived}) &= (233+109)/(81+233+464+109) &&= 342/887 \\
P(\text{male, survived}) &= 109/(81+233+464+109) &&= 109/887 \\
P(\text{survived} \mid \text{male}) &= 109/(464+109) &&= 109/573
\end{aligned}
$$

Joint and marginal distributions for a table

We can now build a complete table of probabilities of the form $P(\text{fate}, \text{sex})$, which is an especially simple case of a *joint probability distribution*.

	female	male	
drowned	81/887	464/887	545/887
survived	233/887	109/887	342/887
	314/887	573/887	887/887

Probabilities such as $P(\text{survived}) = 342/887$ and $P(\text{male}) = 573/887$ appear as sums across rows or down columns—such sums form the *marginal distributions*.

Conditional distributions

Finally, we can work out *conditional distributions* such as $P(\text{fate} \,|\, \text{male})$, which is

$$P(\text{survived} \,|\, \text{male}) = 109/573 \quad \text{and} \quad P(\text{drowned} \,|\, \text{male}) = 464/573.$$

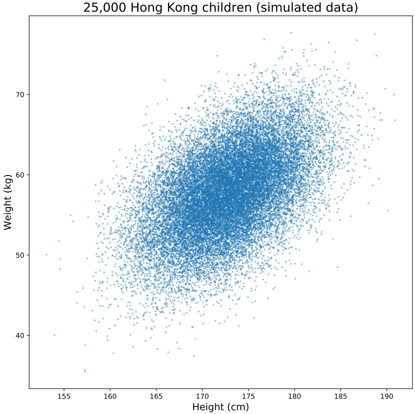
One can compute this from our original table and its marginal distributions using a result from the first lecture,

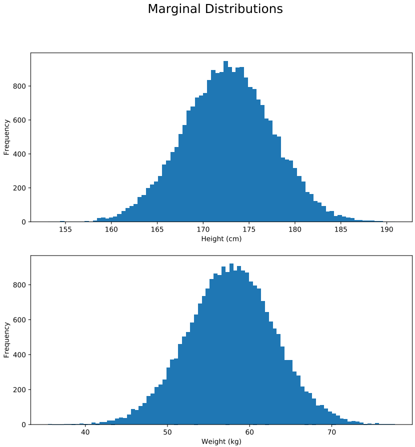
$$P(A \,|\, B) = \frac{P(A, B)}{P(B)},$$

which implies

$$\begin{aligned} P(\text{survived} \,|\, \text{male}) &= P(\text{survived, male})/P(\text{male}) \\ &= (109/887)/(573/887) \\ &= 109/573. \end{aligned}$$

Joint & marginal distributions for continuous variables





The data here are 25,000 simulated measurements of height and weight from a growth survey of children in Hong Kong. I screen-scraped them from a page linked to

http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_020108_HeightsWeights

From point clouds to conditional densities

The (height, weight) data is just a cloud of points, but it would be useful to have the underlying probability density function $f(h, w)$ from which they were drawn. That is, we'd like to be able to compute directly such things as the conditional density

$$f(w \mid h_\star) = \frac{f(h_\star, w)}{f(h_\star)} \text{ where } f(h_\star) = \int f(h_\star, w)\, dw$$

Here $f(h_\star)$ is the value of the marginal density when $h = h_\star$. The integral that defines $f(h_\star)$ is the continuous analogue of the row and column sums we had for the discrete, tabulated Titanic data.

The lectures on EDA suggest two ways to represent these distributions and, for later in the lecture, we'll want a third.

- histogram-based estimators;
- kernel density estimators (KDEs);
- approximation by a multivariate normal.

The histogram as a density estimator

Given a collection $\{x_1, \ldots, x_N\}$ of independent samples of a random variable X, we can approximate its density $f(x)$ with a piecewise-constant function $\hat{f}(x)$ by ...

1. Choosing a set of bin-edges $\{b_0,\, b_1,\, \ldots,\, b_n\}$ for a histogram.
2. Counting the number n_j of the x_k that fall into the j-th histogram bin,

$$n_j = \left|\{x_k \,|\, b_j \le x_k < b_{j+1}\}\right|.$$

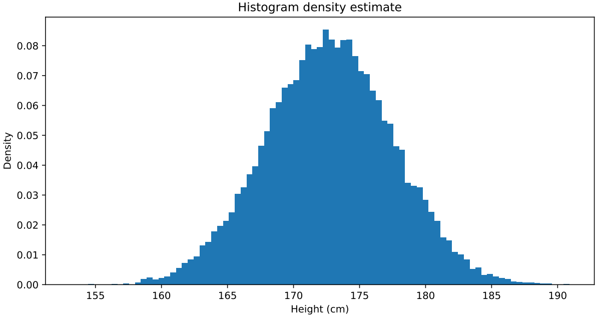
3. Approximating the density $f(h)$ over the interval $b_j \le x \le b_{j+1}$ with the constant value

$$\hat{f}_j = \left(\frac{n_j}{N}\right) \frac{1}{b_{j+1} - b_j}$$

   where N is the size of our original sample.

Histogram estimator for a marginal density: in python

```
# Draw a histogram-based estimate of the
# marginal density of height
hMarginalDensityFig = plt.figure(figsize=[10,5])
plt.hist( hw_df['Height'], bins='fd', density=True )
plt.title( 'Histogram density estimate' )
plt.xlabel('Height (cm)')
plt.ylabel('Density')
plt.show()
```

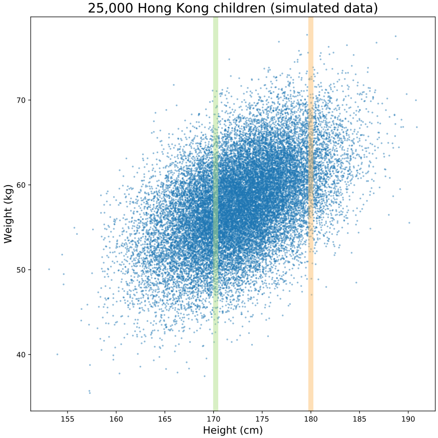


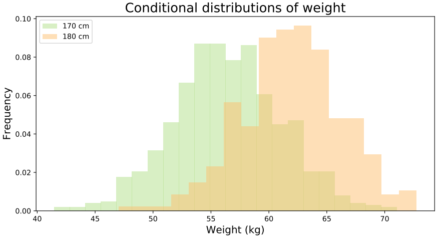
Estimating conditional densities with histograms

In a similar spirit, given a collection $\mathcal{C} = \{(x_1, y_1),\, \ldots,\, (x_N, y_N)\}$ of independently-sampled (X, Y) pairs we can approximate the conditional density $f(y \,|\, x_\star)$ as follows.

1. Choose bin-edges $\{b_0,\, b_1,\, \ldots,\, b_n\}$ for a histogram of the x-values $\{x_1,\, \ldots,\, x_N\}$.
2. Find the bin that contains $x_\star$. That is, find the index $j_\star$ of the bin-edge $b_{j_\star}$ such that
$$b_{j_\star} \leq x_\star < b_{j_\star+1}$$
3. Collect all the $(x,\, y)$ pairs whose x-value falls into the same bin as $x_\star$. That is, find the collection $\mathcal{C}_\star$ of pairs
$$\mathcal{C}_\star = \{(x_k, y_k) \mid b_{j_\star} \leq x_k < b_{j_\star+1}\}\,.$$
The points in $\mathcal{C}_\star$ lie in a vertical strip on the xy-plane.
4. Make a histogram of the y-values that appear in the pairs in $\mathcal{C}_\star$.

Histograms for conditional densities: in pictures

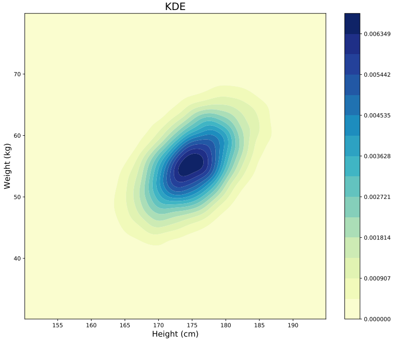

25,000 Hong Kong children (simulated data)

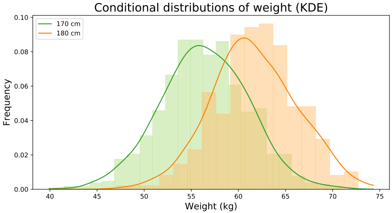

Conditional distributions of weight

At left: the full dataset with strips that correspond to $h_\star = 170$ cm (green) and $h_\star = 180$ cm (orange).
Above: the corresponding histogram estimates of $f(w \mid h_\star)$ are plotted together.

KDEs for the Hong Kong data

The KDE estimator generalises straightforwardly to multivariate data.



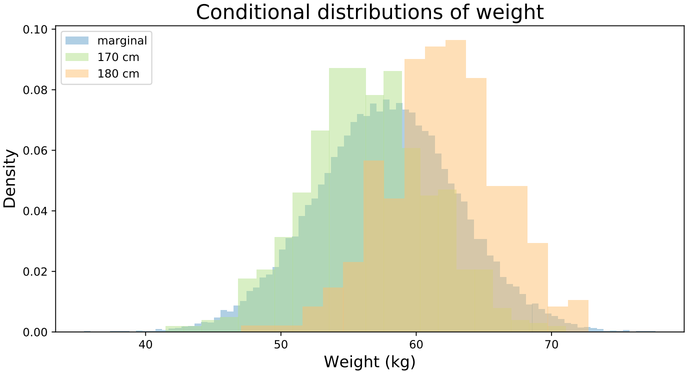


At left: a two-dimensional KDE for the Hong Kong data. It's based on a random subsample of 12,500 points.

At right above: KDE and histogram estimates for the conditional distributions $f(w \mid h_\star)$ with $h_\star = 170$ (green) and $h_\star = 180$ (orange). These conditional KDEs are derived directly from a two-dimensional KDE for the full data set.

Why should we care about conditional distributions?

Conditioning on height tells us something about weight.



Distrib.	Type	Mean (kg)	Std. dev. (kg)
$f(w)$	marginal	57.8	5.30
$f(w \mid h = 170)$	conditional	56.4	4.64
$f(w \mid h = 180)$	conditional	62.0	4.32