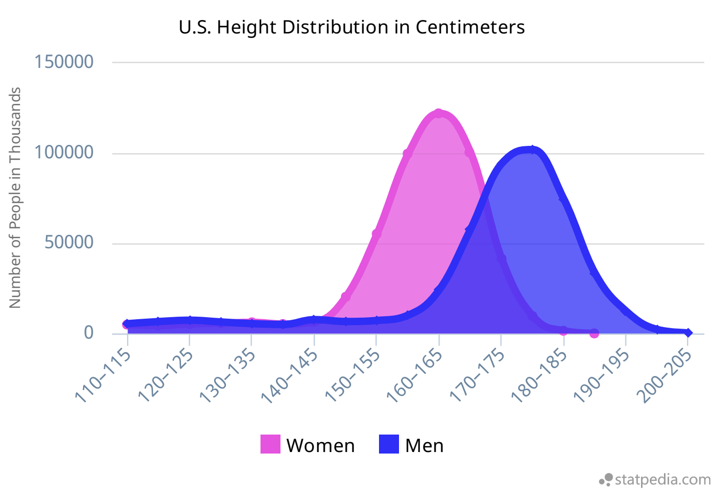
Modeling dependence between variables

DSE 210

Multiple random variables

We've seen many ways to model how a single variable, e.g. height, is distributed in a population.



What if we have more variables, e.g. weight as well?

Dependence

Example: For a person chosen at random from a population, take

$$H = \text{height}$$
$$W = \text{weight}$$

We could treat them as independent, e.g.

- Fit a Gaussian G_1 to the heights
- Fit a Gaussian G_2 to the weights

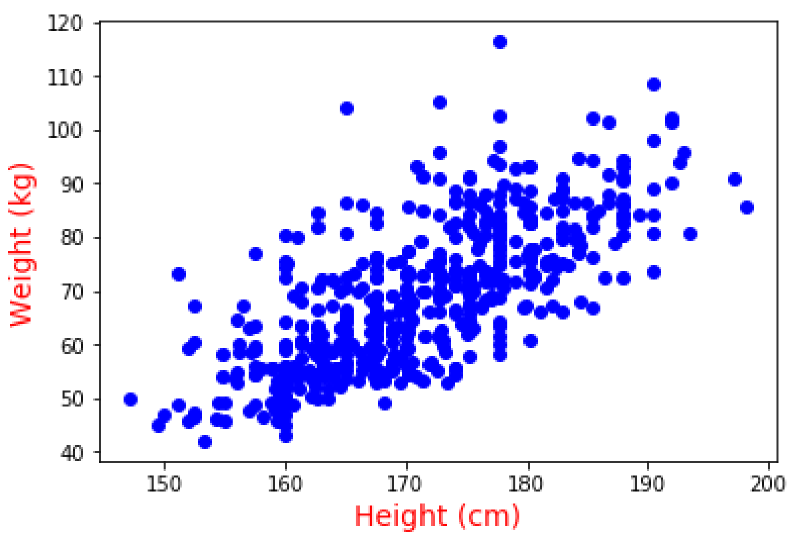
Independence would mean

$$\Pr(H = h, W = w) \ = \ \Pr(H = h) \Pr(W = w).$$

This is not a good model. Why?

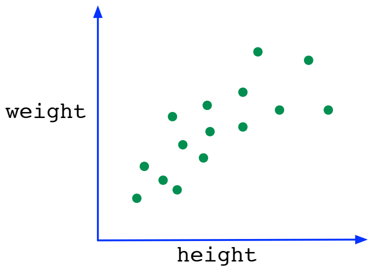
Correlation

Height and weight are **positively correlated**.



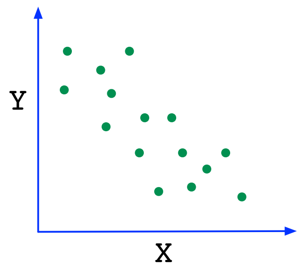
Based on body measurements of 507 people at `https://ww2.amstat.org/publications/jse/datasets/body.txt`

Types of correlation

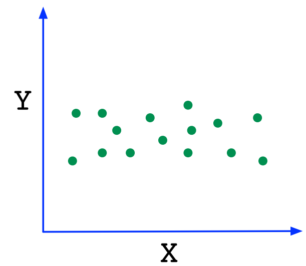


H, W **positively correlated**
This also implies

$$\mathbb{E}[HW] > \mathbb{E}[H]\,\mathbb{E}[W]$$



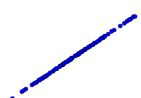
X, Y **negatively correlated**
$\mathbb{E}[XY] < \mathbb{E}[X]\,\mathbb{E}[Y]$



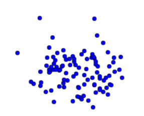
X, Y **uncorrelated**
$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$

Correlation coefficient: pictures

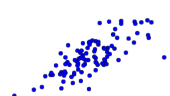
$r = 1$



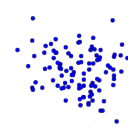
$r = 0$



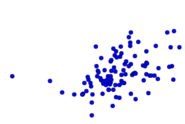
$r = 0.75$



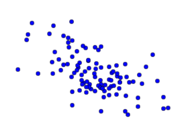
$r = -0.25$



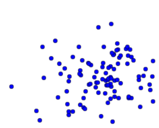
$r = 0.5$



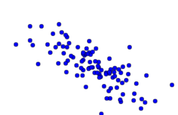
$r = -0.5$



$r = 0.25$



$r = -0.75$



Covariance and correlation

- **Covariance**

$$\begin{aligned}\mathrm{cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]\end{aligned}$$

  Maximized when $X = Y$, in which case it is $\mathrm{var}(X)$.
  In general, it is at most $\mathrm{std}(X)\mathrm{std}(Y)$.

- **Correlation**

$$\mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\mathrm{std}(X)\mathrm{std}(Y)}$$

  This is always in the range $[-1, 1]$.

Example 1

Find $\mathrm{cov}(X, Y)$ and $\mathrm{corr}(X, Y)$

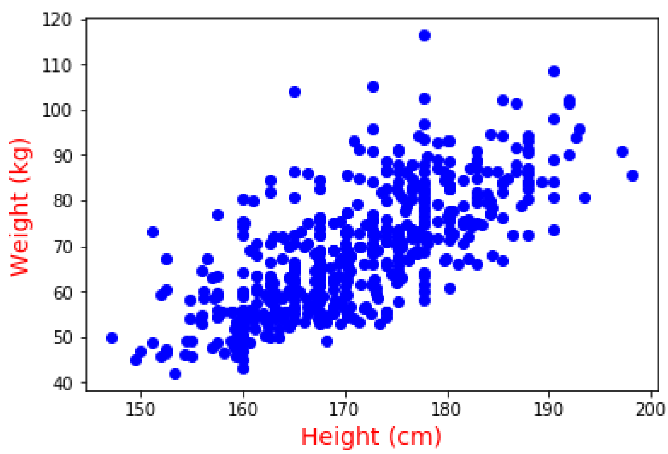
x	y	$\Pr(x, y)$
-1	-1	$1/3$
-1	1	$1/6$
1	-1	$1/3$
1	1	$1/6$

Example 2

Find $\text{cov}(X, Y)$ and $\text{corr}(X, Y)$

x	y	$\Pr(x, y)$
-1	-10	$1/6$
-1	10	$1/3$
1	-10	$1/3$
1	10	$1/6$

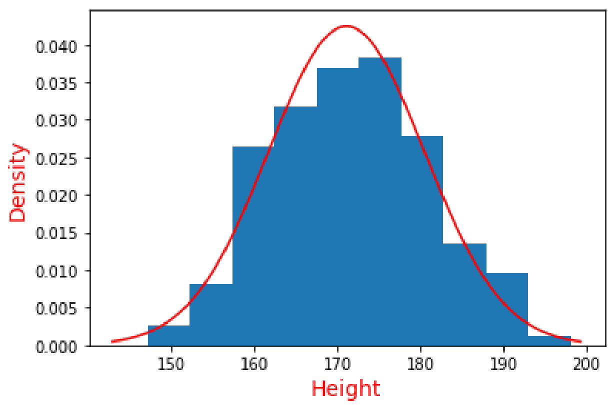
Height and weight again

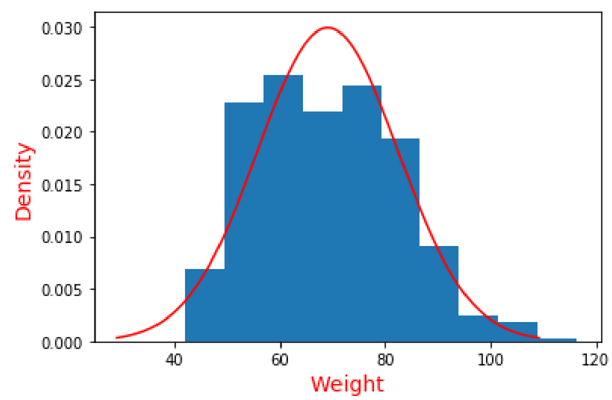


- Height (cm): $\mathbb{E}(H) = 171.1$, $\text{std}(H) = 9.4$
- Weight (kg): $\mathbb{E}(W) = 69.1$, $\text{std}(W) = 13.3$
- $\mathbb{E}(HW) = 11924.0$ while $\mathbb{E}(H)\mathbb{E}(W) = 11834.2$
- $\text{cov}(H, W) = 89.9$ and $\text{corr}(H, W) = 0.72$

A distribution over two variables?

We want a distribution over two variables: $(X_1, X_2) = (\text{height}, \text{weight})$



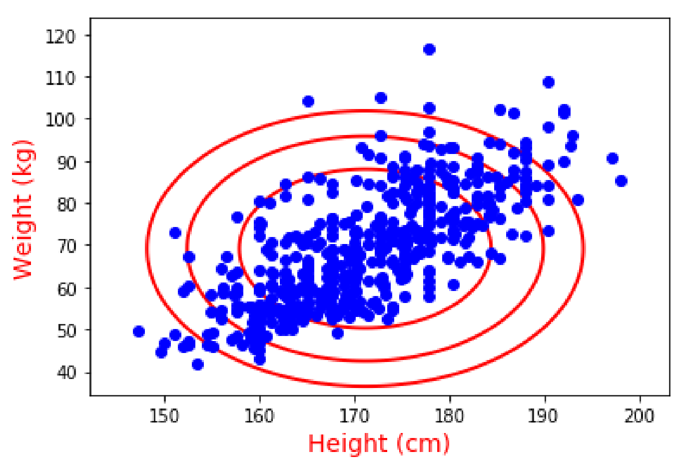


- Mean $\mu_1 = 171.1$
- Standard dev $\sigma_1 = 9.4$

- Mean $\mu_1 = 69.1$
- Standard dev $\sigma_1 = 13.3$

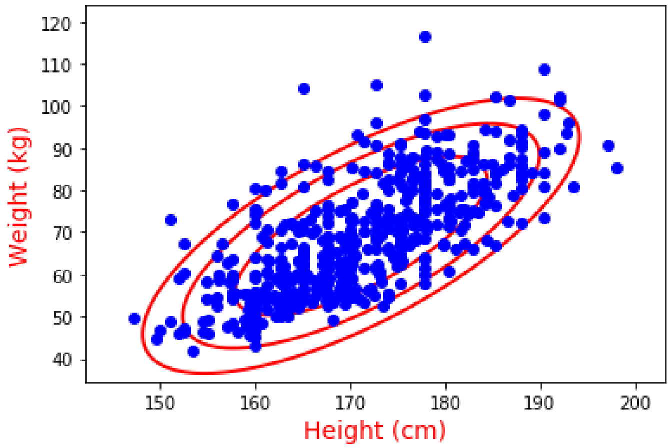
Independent variables

One possibility: Treat the two variables as independent and fit a Gaussian to each.



What is the resulting density over (x_1, x_2)?

The bivariate Gaussian



Model the data by a bivariate Gaussian, parametrized by:

$$\text{mean } \mu = \begin{pmatrix} 171.1 \\ 69.1 \end{pmatrix} \text{ and covariance matrix } \Sigma = \begin{pmatrix} 88.4 & 89.9 \\ 89.9 & 176.9 \end{pmatrix}$$

The bivariate (2-d) Gaussian

A distribution over $(x_1, x_2) \in \mathbb{R}^2$, parametrized by:

- **Mean** $(\mu_1, \mu_2) \in \mathbb{R}^2$, where $\mu_1 = \mathbb{E}(X_1)$ and $\mu_2 = \mathbb{E}(X_2)$
- **Covariance matrix** $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ where $\left\{ \begin{array}{c} \Sigma_{11} = \text{var}(X_1) \\ \Sigma_{22} = \text{var}(X_2) \\ \Sigma_{12} = \Sigma_{21} = \text{cov}(X_1, X_2) \end{array} \right\}$

Density is highest at the mean, falls
off in ellipsoidal contours.

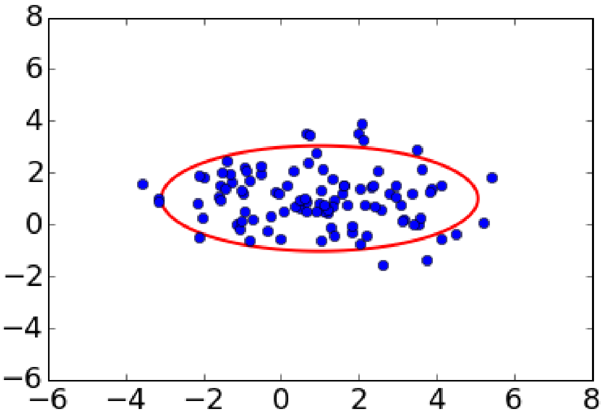
Density of the bivariate Gaussian

- **Mean** $(\mu_1, \mu_2) \in \mathbb{R}^2$, where $\mu_1 = \mathbb{E}(X_1)$ and $\mu_2 = \mathbb{E}(X_2)$
- **Covariance matrix** $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$

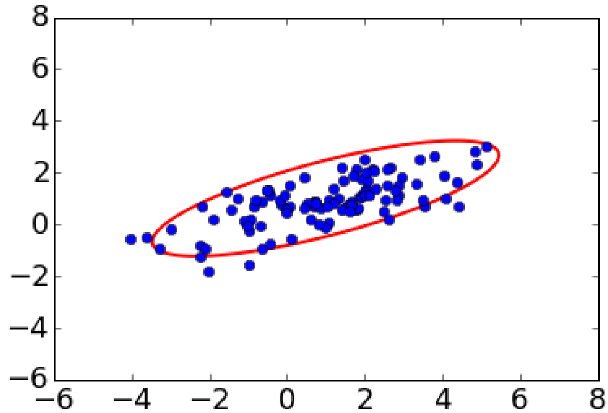
Density $p(x_1, x_2) = \dfrac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\dfrac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \right)$

Bivariate Gaussian: examples

In either case, the mean is $(1, 1)$.

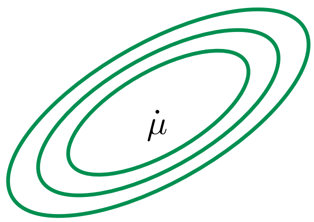


$$\Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 4 & 1.5 \\ 1.5 & 1 \end{bmatrix}$$

The multivariate Gaussian



$N(\mu, \Sigma)$: Gaussian in $\mathbb{R}^d$

- mean: $\mu \in \mathbb{R}^d$
- covariance: $d \times d$ matrix Σ

Generates points $X = (X_1, X_2, \ldots, X_d)$.

- μ is the vector of coordinate-wise means:

$$\mu_1 = \mathbb{E}X_1,\ \mu_2 = \mathbb{E}X_2, \ldots,\ \mu_d = \mathbb{E}X_d.$$

- Σ is a matrix containing all pairwise covariances:

$$\begin{aligned} \Sigma_{ij} = \Sigma_{ji} &= \mathrm{cov}(X_i, X_j) \quad \text{if } i \neq j \\ \Sigma_{ii} &= \mathrm{var}(X_i) \end{aligned}$$

Density $\quad p(x) = \dfrac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\dfrac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$

Special case: independent features

Suppose the X_i are independent, and $\mathrm{var}(X_i) = \sigma_i^2$.

What is the covariance matrix Σ, and what is its inverse Σ^{-1}?

Diagonal Gaussian

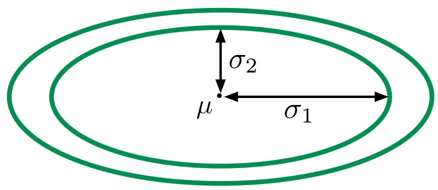
Diagonal Gaussian: the X_i are independent, with variances σ_i^2. Thus

$$\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2) \quad \text{(off-diagonal elements zero)}$$

Each X_i is an independent one-dimensional Gaussian $N(\mu_i, \sigma_i^2)$:

$$\Pr(x) \;=\; \Pr(x_1)\Pr(x_2)\cdots\Pr(x_d) \;=\; \frac{1}{(2\pi)^{d/2}\sigma_1\cdots\sigma_d} \exp\left(-\sum_{i=1}^{d} \frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right)$$

Contours of equal density are **axis-aligned ellipsoids** centered at μ:



How to fit a Gaussian to data

Fit a Gaussian to data points $x^{(1)}, \ldots, x^{(m)} \in \mathbb{R}^d$.

- Empirical mean

$$\mu \;=\; \frac{1}{m}\left(x^{(1)} + \cdots + x^{(m)}\right)$$

- Empirical covariance matrix has i, j entry:

$$\Sigma_{ij} \;=\; \left(\frac{1}{m}\sum_{k=1}^{m} x_i^{(k)} x_j^{(k)}\right) - \mu_i \mu_j$$