Simple summary statistics

A: Location and scale

Mean

The *mean* (*expected value*) of random variable X, or equivalently of its distribution, is

$$\mathbb{E}(X) = \begin{cases} \displaystyle\sum_x x \Pr(X = x) & \text{if } X \text{ is discrete} \\ \displaystyle\int x \; p(x) \; dx & \text{if } X \text{ is continuous with density } p(x) \end{cases}$$

The *empirical mean* of a set of data points $x_1, \ldots, x_n$:

$$\frac{1}{n} \sum_{i=1}^{n} x_i.$$

What is the relationship between these two definitions?

Median

Two ways of summarizing a set of numbers by a single number.

- The **(empirical) mean**
- The **(empirical) median**: the number in the middle, if you sort them

Find the median of the following sets of numbers:

- $10, -20, 100, 20, 50$
- $50, 100, 60, 90, 20, 10$

How can we define the median of a random variable X?

Mean vs median

In a certain neighborhood, there are 100 houses.

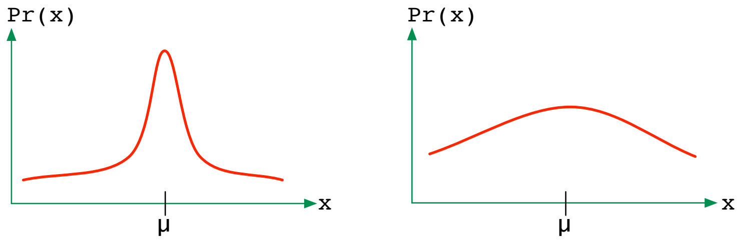
- 10 of the houses cost \$100K
- 60 of the houses cost \$200K
- 29 of the houses cost \$300K
- one house costs \$100M

- What is the mean house cost, roughly?
- What is the median cost?

Variance

We can summarize a random variable X by its mean μ (or median).

Problem: This doesn't capture the **spread** of X.



Possible measure of spread: average distance from the mean, $\mathbb{E}(|X - \mu|)$?

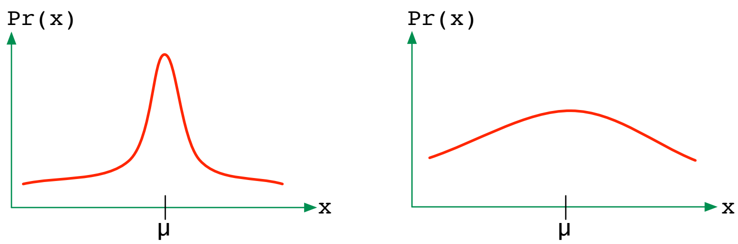
For convenience, take the square instead of the absolute value.

$$\textbf{Variance:} \qquad \text{var}(X) \;=\; \mathbb{E}(X-\mu)^2 \;=\; \mathbb{E}(X^2) - \mu^2,$$

where $\mu = \mathbb{E}(X)$. The variance is always ≥ 0.

Standard deviation

Recall: $\mathrm{var}(X) = \mathbb{E}(X - \mu)^2$, where $\mu = \mathbb{E}(X)$.



The **standard deviation** of X is $\text{std}(X) = \sqrt{\mathrm{var}(X)}$.
It is, *roughly*, the average amount by which X differs from its mean.

Question: How does $\text{std}(X)$ relate to $\mathbb{E}(|X - \mu|)$? Are they equal?

B: Measuring dependence between variables

Independent random variables

Random vars X, Y are **independent** if $\mathrm{Pr}(X = x, Y = y) = \mathrm{Pr}(X = x)\mathrm{Pr}(Y = y)$.

Independent or not? $X, Y \in \{-1, 0, 1\}$, with these probabilities:

		Y		
		-1	0	1
	-1	0.4	0.16	0.24
X	0	0.05	0.02	0.03
	1	0.05	0.02	0.03

Testing independence

Suppose you are given samples (X, Y) from a bivariate distribution:

$$(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2.$$

How would you test whether X and Y are independent?

Dependence

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

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

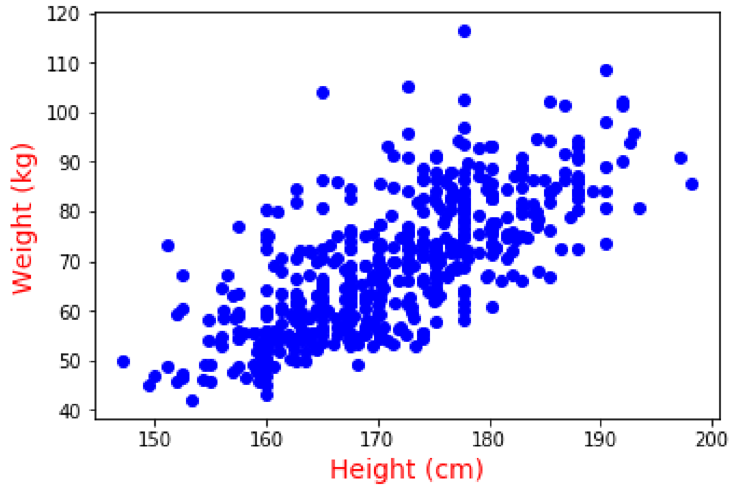
Independence would mean

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

This is unlikely to be true. Why?

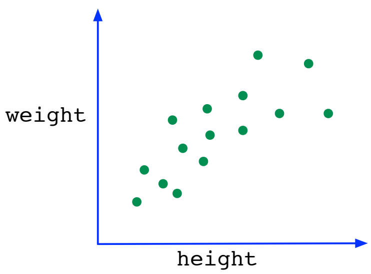
Correlation

Height and weight are **positively correlated**.



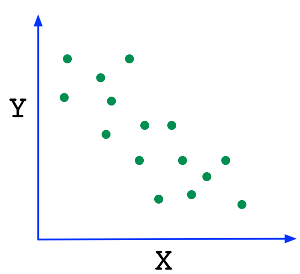
Based on body measurements of 507 people at
`https://ww2.amstat.org/publications/jse/datasets/body.txt`

Types of correlation

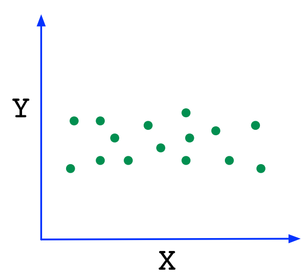


H, W **positively correlated**
This also implies

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



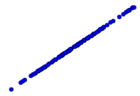
X, Y **negatively correlated**
$\mathbb{E}[XY] < \mathbb{E}[X]\,\mathbb{E}[Y]$



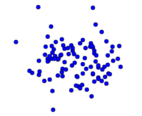
X, Y **uncorrelated**
$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$

Correlation coefficient: pictures

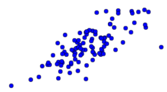
$r = 1$



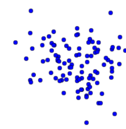
$r = 0$



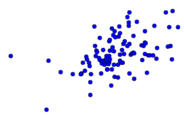
$r = 0.75$



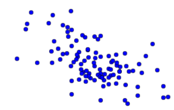
$r = -0.25$



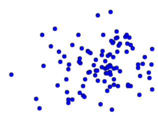
$r = 0.5$



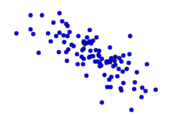
$r = -0.5$



$r = 0.25$



$r = -0.75$



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

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

x	y	$\Pr(x, y)$
1	4	1/4
1	−4	1/4
−1	4	1/8
−1	−4	3/8

Independent $\not\equiv$ uncorrelated

C: Key properties of the mean and variance

Linear functions of a single random variable

- If you double a set of numbers, how are their mean and variance affected?
- If you increase a set of numbers by 1, how much do their mean and variance change?
- Let X be any random variable.
  For some constants a, b, define a new random variable $V = aX + b$.
  Express $\mathbb{E}(V)$ and $\mathrm{var}(V)$ in terms of $\mathbb{E}(X)$ and $\mathrm{var}(X)$.

Linearity of expectation

A powerful and extremely useful property:

Linearity of expectation: For any random variables $X_1, \ldots, X_m$,

$$\mathbb{E}(X_1 + X_2 + \cdots + X_m) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_m).$$

Linearity of variance

We've seen that $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.
Is this also true of variance, i.e., is $\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y)$?

- In general, **no**. Give a counterexample.

- But it is true if X and Y are **independent**.

D: Postscript: Information-theoretic quantities

Entropy and mutual information

How "random" is a distribution?

- Easy but crude: variance.
- Much better: **entropy**.

How "dependent" are two variables?

- Easy but crude: correlation.
- Much better: **mutual information**.

Unfortunately these quantities are hard to empirically estimate.