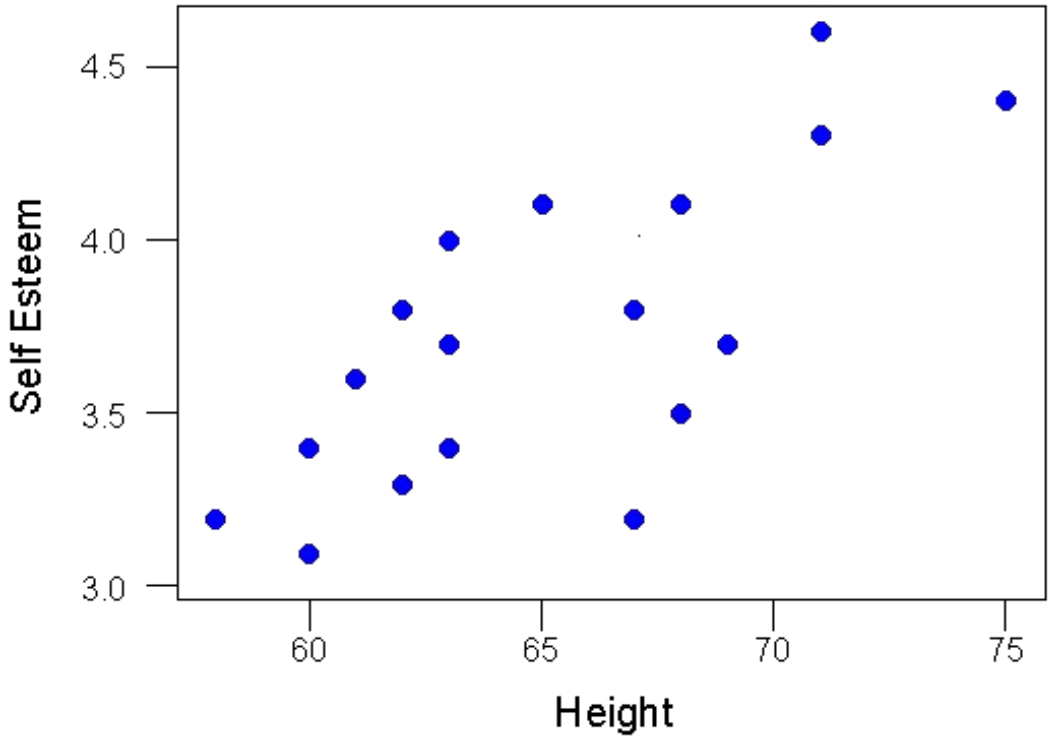
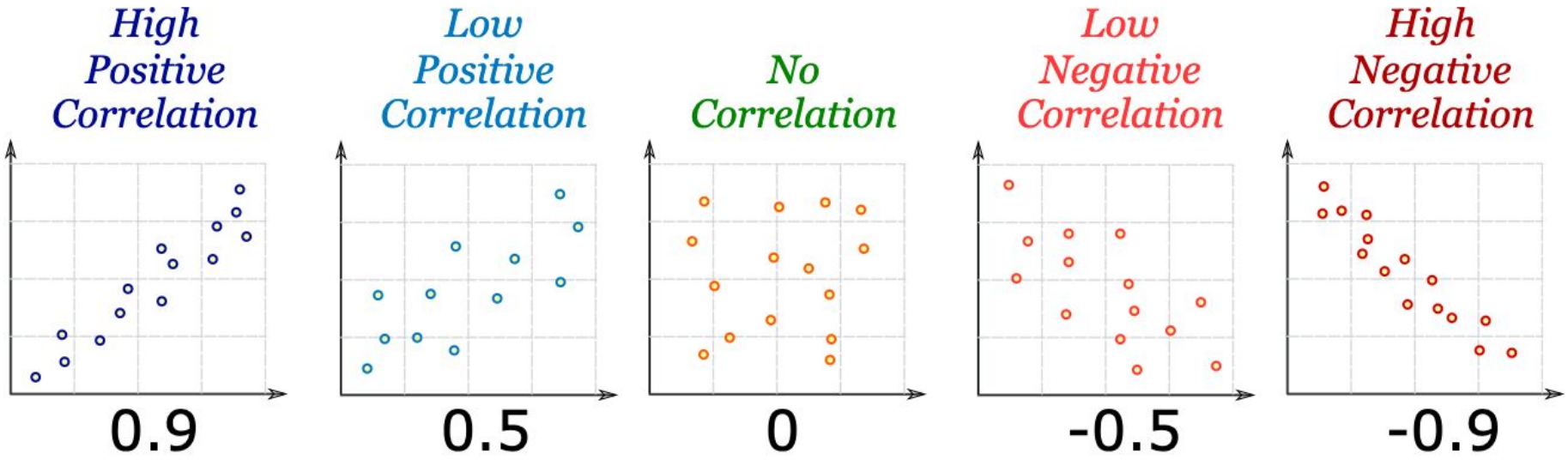
Correlation and p-values



Slides: Ben Cowley

We want some metric to quantify how much two variables go up and down together...



Pearson's correlation, *r*

- developed by Karl Pearson
  - founded world's first statistics department at University College London, 1911
  - also developed p-value, principal component analysis, and first introduction of the histogram

Pearson's correlation, *r*

- developed by Karl Pearson
  - founded world's first statistics department at University College London, 1911
  - also developed p-value, principal component analysis, and first introduction of the histogram

  - terrible human being:
    - work was inspired by eugenics (create "improved" humans through genetic selectivity) and founded "Annals of Eugenics"
    - unfortunately, lots of statistical ideas arose from this field
    - morphed into "biometrics"

Pearson's correlation, *r*

- consider two variables, X and Y

Pearson's correlation, *r*

- consider two variables, X and Y
- let μ_X be the mean of X and μ_Y be the mean of Y

Pearson's correlation, *r*

- consider two variables, X and Y
- let μ_X be the mean of X and μ_Y be the mean of Y
- let σ_X^2 be the variance of X and σ_Y^2 be the variance of Y

Pearson's correlation, *r*

- consider two variables, X and Y
- let μ_X be the mean of X and μ_Y be the mean of Y
- let σ_X^2 be the variance of X and σ_Y^2 be the variance of Y
- then correlation $\rho = \dfrac{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}}$

GUESS THE CORRELATION

NEW GAME

TWO PLAYERS

SCORE BOARD

ABOUT

SETTINGS

Write a function that computes the correlation

input: X, Y (vectors)
output: *r* (correlation)

script1.py

How do we know if a correlation is significant?

How do we know if a correlation is significant?

given true $r = 0.1$, what is estimated r for 10, 100, 1000 samples?

How do we know if a correlation is significant?

given true *r* = *0.1*, what is estimated *r* for 10, 100, 1000 samples?

idea:

- shuffle data
- compute *r* of the shuffled data
- do this many times
- compare the measured *r* to the shuffled *r*'s.
- how many *r_shuffs* have greater magnitude that *r_actual*?

How do we know if a correlation is significant?

given true *r* = *0.1*, what is estimated *r* for 10, 100, 1000 samples?

idea:

- shuffle data
- compute *r* of the shuffled data
- do this many times
- compare the measured *r* to the shuffled *r*'s.
- how many *r_shuffs* have greater magnitude that *r_actual*?

→ p-value: probability that *r_actual* is far from null distribution (*r_shuffs*)

Write a function that computes the correlation and p-value

input: X, Y (vectors)
output: *r* (correlation), p-value

script2.py