Week 5. Partial correlation, heatmap, and uniform inequality

Section 6.6.2.

Correlation heatmap

Statistical analysis of stock prices

From investment standpoint, stock return is the most important quantity of interest:

$$\text{return} = \frac{p_t - p_{t-1}}{p_{t-1}}.$$

From a statistical standpoint, it's convenient to define return on the log scale

$$r_t = \ln p_t - \ln p_{t-1}.$$

Justification: (a) easy to do stat analysis on the log scale including autoregression, (b) q-q plot confirms that $\ln p$ is normally distributed, and (c) due to

$$\ln(1+x) \simeq x$$

for small x we have

$$\ln p_t - \ln p_{t-1} = \ln \frac{p_t}{p_{t-1}} = \ln\left(1 + \frac{p_t - p_{t-1}}{p_{t-1}}\right) \simeq \frac{p_t - p_{t-1}}{p_{t-1}} = \text{return},$$

or

$$\text{return} \simeq r_t.$$

How to visualize correlation among large number of variables? Answer: represent the number by color: this is **heatmap.**

For example, assuming that the screen resolution of a computer is 1000 × 1000 pixels one can see the correlation matrix of 1000 variables!

R function `cimcorSP` for the data analysis of 17 major stock price returns

Run `cimcorSP(job=1)` for 17x17 pairwise stock returns correlation

`cimcorSP(job=1.1)` for the high-resolution 5,000 by 5,000 pixels saved in the `png` graphics format

`cimcorSP(job=2)`for correlation heatmap

Partial correlation

Section 4.1.2

When computing correlation, how to exclude common factors? This common factors may lead to false high correlation, like high correlation between revenue of the company and the number of truck drivers (Section 6.7.4).

Example 1 *U and V are independent ($U \perp V$), and*

$$\begin{aligned} X &= U + aZ, \\ Y &= V + aZ \end{aligned}$$

where Z is independent of U and V. The correlation coefficient between X and Y is given by

$$\rho_{XY} = \frac{a^2}{1+a^2}.$$

ρ_{XY} is high when $|a|$ is large due to the presence of the common factor Z.

Sometimes, random variables X and Y are correlated because there is a third (latent, invisible, confounfing) variable Z that contributes to correlation between X and Y. Partial correlation **refines** the correlation by excluding the impact of other variables through conditioning.

Notation:

$$\begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix} \text{ and } \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix}^{-1} = \begin{bmatrix} \rho^{11} & \rho^{12} & \rho^{13} \\ \rho^{12} & \rho^{22} & \rho^{23} \\ \rho^{13} & \rho^{23} & \rho^{33} \end{bmatrix}$$

Definition 2 *The partial correlation coefficient is the **conditional** correlation coefficient between X and Y when other variables are held fixed (conditional correlation coefficient). It can be computed as*

$$\rho_{XY|Z} = -\frac{\rho^{12}}{\sqrt{\rho^{11}\rho^{22}}},$$

where ρ^{jk} denotes the (j,k)th element of the inverse correlation matrix. Alternatively, the partial correlation coefficient may be computed as the correlation between residuals in regression of X on Z and Y on Z.

The following is a continuation of Example 2.

Example 3 *Show that the partial correlation between X and Y is zero regardless of a, that is,*

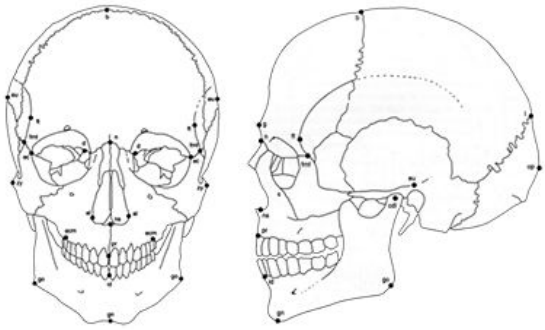
$$\rho_{XY|Z} = 0.$$

Run `cimcorSP(job=3)`for partial correlation heatmap
Using package `pheatmap`
Run `cimcorSP(job=4)`

Goodman osteometric skeleton remain measurements

Statistics for anthropology

https://web.utk.edu/~auerbach/GOLD.htm



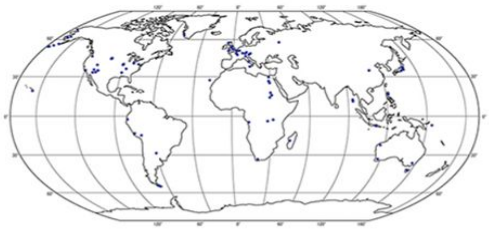
Not secure | web.utk.edu/~auerbach/GOLD.htm

BENJAMIN M. **AUERBACH**, Ph.D.

GOLDMAN OSTEOMETRIC DATA SET

HOME PUBLICATIONS DATA SETS COURSES RESEARCH

PLEASE READ THIS PAGE BEFORE DOWNLOADING THE DATA SET



ABOUT THE DATA

The Goldman Data Set consists of osteometric measurements taken from 1538 human skeletons dating from throughout the Holocene. Links to the data are at the bottom of this page. Links to sample provenience and dating information may be found in links below. Measurements were taken bilaterally from four of the long bones: humerus, radius, femur, and tibia. Three measurements were additionally obtained from the pelvis. Sex and age were estimated as well from pelvic observations (see Auerbach and Ruff, 2004, and Auerbach and Ruff, 2006, for a description of the methods employed).

Dr. Benjamin Auerbach obtained all of these data during three solo research trips beginning in September 2001 and ending in July 2003. These trips were made possible as part of a research fellowship provided by generous funding from the Joanna Jackson Goldman Memorial Prize; the data set is named after this organization. Dr. Auerbach continues to be grateful to the staff and curators at the institutions in the United Kingdom, Italy, Austria, Germany, France, Belgium, the United States of America, and Japan for allowing permission to work with their collections. A list of source institutions may be found below.

These data are now made available to researchers for download. The data are available below as a Microsoft Excel legacy file and as a comma-separated text file. A supporting document describing the osteometric measurements is also provided as a PDF file. Additional provenience information for the samples is currently available by request from Dr. Auerbach.

ALL RESEARCHERS WHO PLAN TO USE THESE DATA ARE WELCOMED AND ENCOURAGED TO CONTACT DR. AUERBACH. ANY USE OF THESE DATA IN PRESENTED OR PUBLISHED RESEARCH CARRIES THE STIPULATION THAT THE SOURCE OF THE DATA BE CITED. ACCEPTABLE CITATIONS FOR THE DATA INCLUDE THE REFERENCE OF THE DATA'S ANALYSIS (AUERBACH & RUFF, 2004 or 2006) AND OF THIS WEB SITE.

ADDITIONAL INFORMATION ABOUT THE DATA

Description of the data set and measurements: PDF file

General dates of the sites sampled: Excel file

1,538 skeleton remains with 43 bone measurements uncovered around the globe: file `Godman.csv`
R function runs: `ostmeas(job=1)` and `ostmeas(job=2)`.

Generalized matrix inverse How to inverse a singular matrix?

$\mathbf{A}$ is a $m \times m$ symmetric nonnegative definite matrix, such as covariance or correlation matrix.

Jordan spectral decomposition

$$\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}'$$

where $\mathbf{\Lambda} = \mathbf{diag}(\lambda_1, ..., \lambda_m)$ is a diagonal matrix with nonnegative eigenvalues on the diagonal. Can we suggest an inverse when some $\lambda_i = 0$? We use notation $\mathbf{A}^+$ for generalized matrix inverse.

Define the generalized matrix inverse for a diagonal matrix as

$$\mathbf{\Lambda}^+ = \begin{cases} 0 \text{ if } \lambda_i = 0 \\ 1/\lambda_i \text{ if } \lambda_i \neq 0 \end{cases}$$

Now define

$$\mathbf{A}^+ = \mathbf{P}\mathbf{\Lambda}^+\mathbf{P}'$$

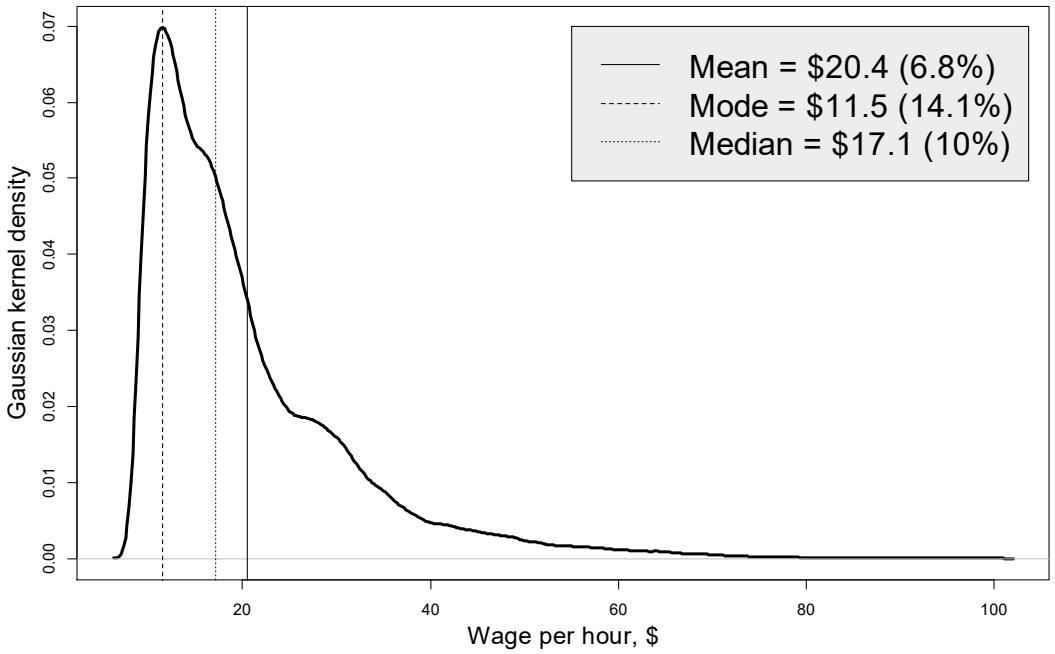
Overview of central tendency measures

Section 2.2

My goal is to plant a seed of doubt in your mind on the wide-spread use of the mean to report the center of the distribution. Reporting a central tendency measure, such as mean, mode, or median, is the most frequent task of data science. How to compare samples using the cumulative distribution function (cdf)?

- What we mean by saying that the prices in one grocery store are higher than in the other?
- What we mean by saying that females are shorter than males?
- What we mean by saying that drug A is better than drug B?

Mean is not a good quantity to make judgment on what is less and what is large.
When we say that females are shorter, how to interpret the sum of heights of females?
Median is better but it reflects the comparison at the middle point of the data.



The Gaussian kernel density for a sample of 234,986 hourly wages in the country.

Example 4 *What central tendency measure to use for a house price in town? Mean, median, or mode?*

Case 1. Mean. The arithmetic average of house prices is a suitable average characteristic for a town clerk who is concerned with the total amount of the property tax to collect from the residents.

Answer. Arithmetic average = Total value of houses in town/number of houses. When reporting the average house price, town officials prefer to use the arithmetic average because they collect the property tax proportional to Total value of houses in town.

Case 2. Median. A real estate agent shows houses to a potential buyer. What is a suitable measure for house price for the buyer, the mean or the median?

Answer. While the mean price makes sense for a town or state official but it is not useful for the buyer who is thinking of the chance of affording the house he/she likes. Instead, the median means that 50% of the houses he/she saw will have price lower than the median and 50% of the houses will have higher price. In this case, the median has a better interpretation from the buyer's perspective.

Case 3. Mode. What central tendency to use to reflect the distribution of house prices in town on the real estate webpage?

Answer. "The most typical house price sold in town is from \$350K to \$400K." The mode.

Conclusion 5 *Mean has an interpretation if and only sum has an interpretation.*

Mode is the most important central tendency measure and yet the most under-appreciated.
How to compute mode?
If `d` is array of data

1. Estimate the density: `densD=density(d)`
2. Estimate the mode: `modeD=mean(densD$x[densD$y==max(densD$y)])`

Example: `D=rchisq(10,df=2);densD=density(D);plot(densD);rug(D);modeD=mean(densD$x[densD$y==max(densD$y)]);segments(modeD,-1,modeD,1000)`

CDF for the uniform sample comparison

Section 5.1

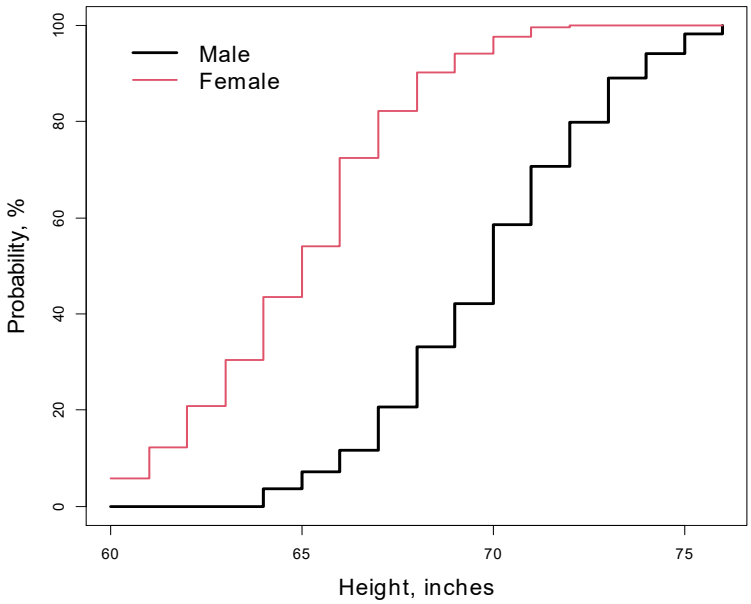
Definition 6 *If X is a random variable (general population) with the pdf (probability density function) $f(x)$, the cdf (cumulative distribution function) is defined as*

$$F(x) = \int_{-\infty}^{x} f(u)du,$$

or the area under the pdf to the left of x. If $\{X_1, X_2, ..., X_n\}$ is a random sample, the empirical cdf (an estimator of the true cdf) is defined as

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \leq x)$$

where 1 *is an indicator function. In words, the empirical cdf is the proportion of data points to the left of x.*



Women are uniformly shorter than men because the proportion of women shorter than x is bigger than proportion of men: $F_{\text{woman}}(x) \geq F_{\text{man}}(x)$ for all x.

Definition 7 *Let X and Y be two random variables with cdfs $F_X(x)$ and $F_Y(x)$. We say that Y is **uniformly** smaller than X, or symbolically, $Y \preceq X$, if $F_Y(x) \geq F_X(x)$ for all x. The same definition applies to empirical cdfs. We say that Y is uniformly smaller than X if the proportion of data Y smaller than x is bigger than the proportion of data X smaller than x for all x.*

It easy to see that $Y \preceq X$ implies $\text{median}_Y < \text{median}_X$, but the reverse is not true.
Run `cdf.dyn(job=2)`
Example. *Salary comparison Vermont versus Connecticut.*
See the R code `salary`