Exploratory Data Analysis

STOR 565

Andrew Nobel

August, 2021

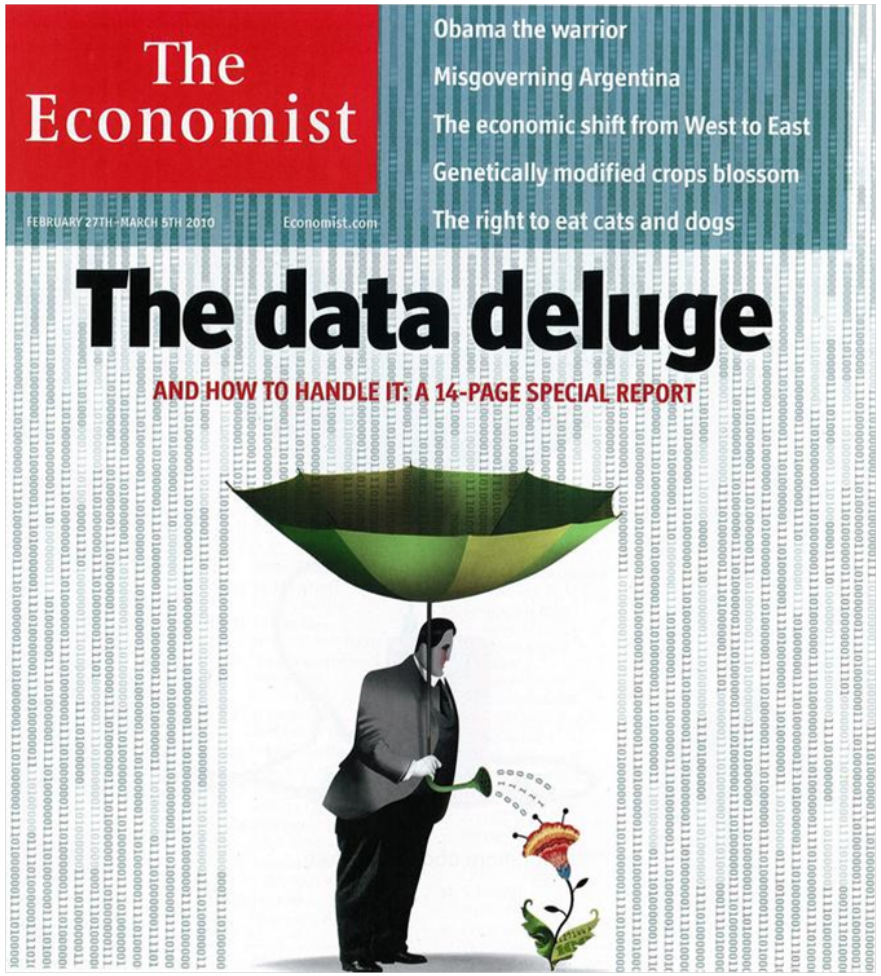
Data: Definitions from the O.E.D.

General: Facts and statistics collected together for reference or analysis.

Philosophy: Things known or assumed as facts, making the basis of reasoning or calculation.

Computing: The quantities, characters, or symbols on which operations are performed by a computer, being stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.

Big Data: Cover of The Economist, 2010



Big Data: Enron Email Graph

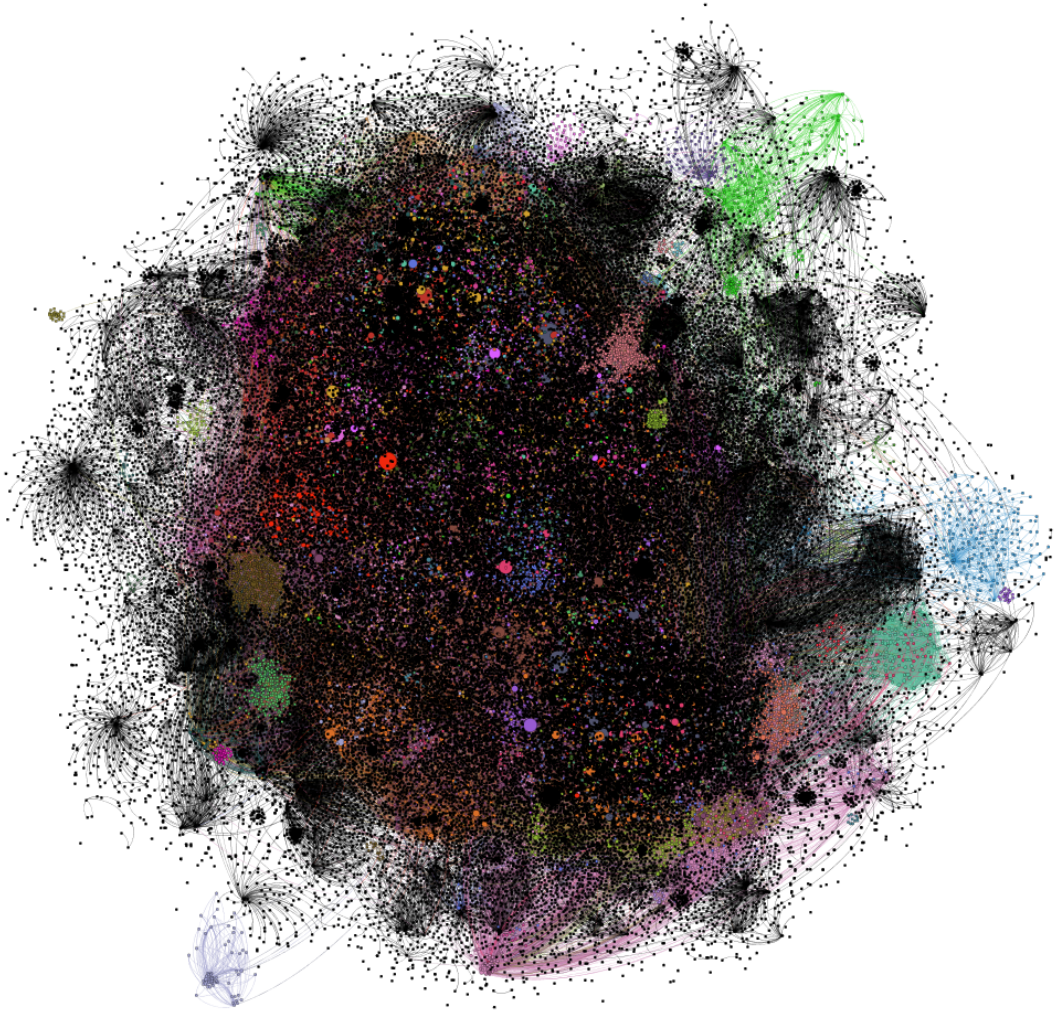


Figure: Graph of three million emails between 80,705 people. Vertices are individuals. Edges are colored according to number of emails exchanged.

Datasets

Data: Obtained by measuring a common set of quantities, usually numerical or categorical, across a set of related objects or individuals

- Objects under study are called **samples**
- Measured quantities referred to as **features** or **variables**
- In supervised problems samples accompanied by **label** or **response**

Example: Administer 75 question survey about eating and hygiene habits to 200 individuals, 75 with Type II diabetes and 125 without

Example: Measure the expression of 20,000 genes in 350 breast tumors that have been assigned to one of 3 disease subtypes.

Samples and Features: Different Regimes

Classical: Number of samples larger than number of variables ($n > p$)

- Low dimension high sample size
- Measurements made manually, e.g., field trials, drug testing

Modern: Number of features larger than number of features ($p > n$)

- High dimension low sample size
- Measurements from high-throughput technologies, e.g., genomics

Representing Data in Matrix Form

Data Matrix: A data set with

- n (labeled) samples
- p numerical features

is described by a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ (and a response vector $\mathbf{y} \in \mathbb{R}^n$).

Note

- i'th row of $\mathbf{X}$ contains measurements from the i'th sample
- j'th column of $\mathbf{X}$ contains measurements of the j'th feature
- i'th entry of $\mathbf{y}$ is the response for the i'th sample

Example: Fisher's Iris Data

- $n = 150$ samples consisting of 50 irises from each of three different species: setosa, versicolor, and virginica
- $p = 4$ features: length and width of sepals and petals

Analyzed by R. A. Fisher "The use of multiple measurements in taxonomic problems", *Annals of Eugenics*, 1936

Iris Data

Data: Matrix with $n = 150$ rows and $p = 4$ columns; 150 dimensional feature vector containing species designation

Species	Sepal Length	Sepal Width	Petal Length	Petal Width
Setosa	5.1	3.5	1.4	0.2
Setosa	4.6	3.4	1.4	0.3
Versicolor	5.0	2.0	3.5	1.0
Virginica	7.2	3.6	6.1	2.5
...	...	...	...	...

Fisher's Iris Data: Scatterplots

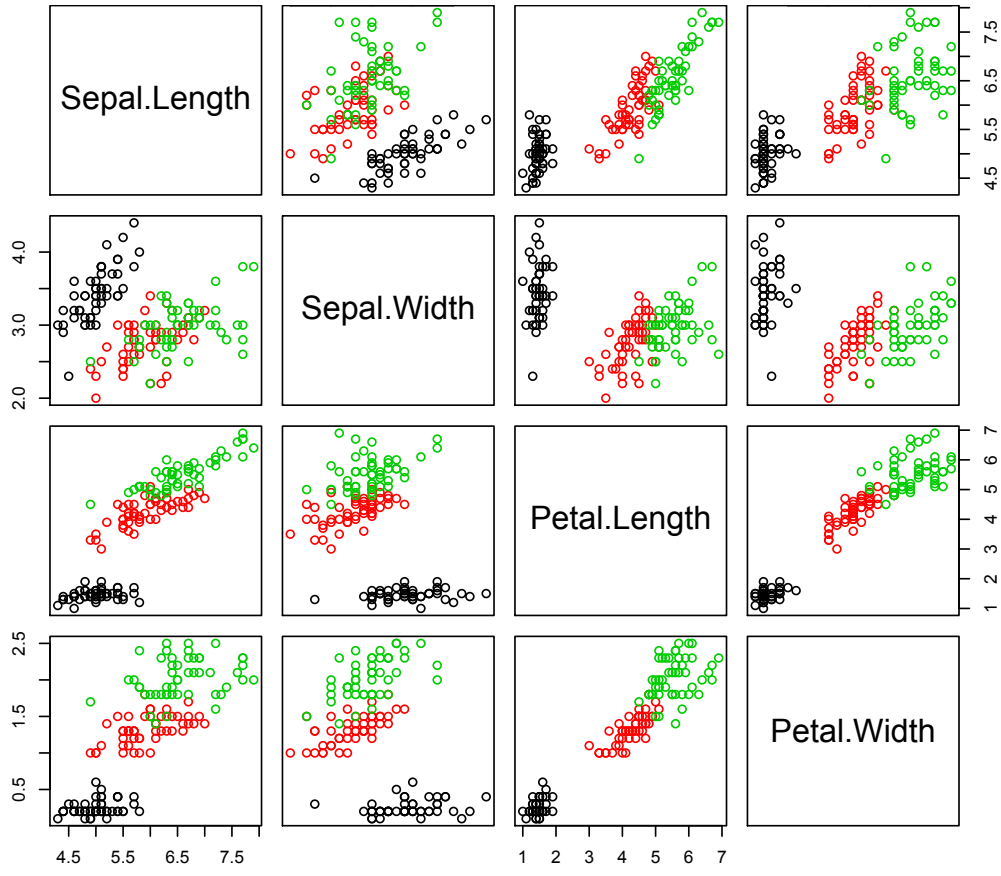


Figure: Pairwise scatterplot of Iris measurements based on four measured features. Colors indicate species: *Setosa*, *Virginica*, *Versicolor*.

Fisher's Iris Data: Principal Component Analysis

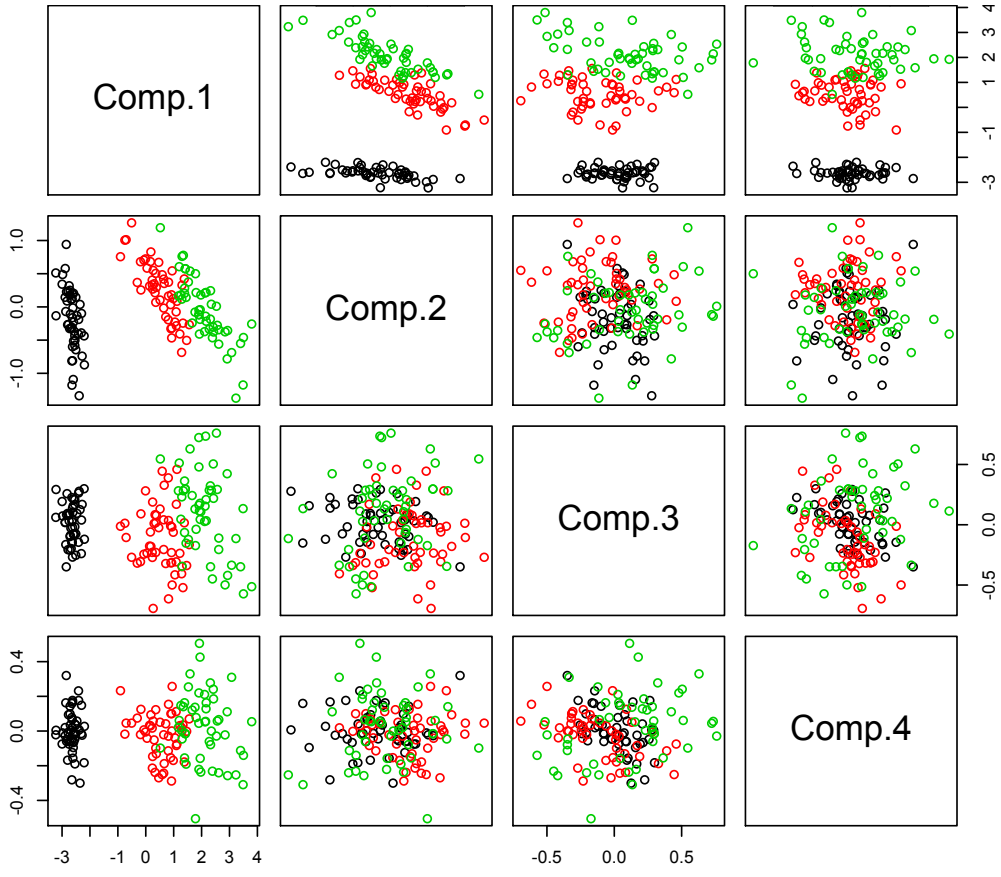
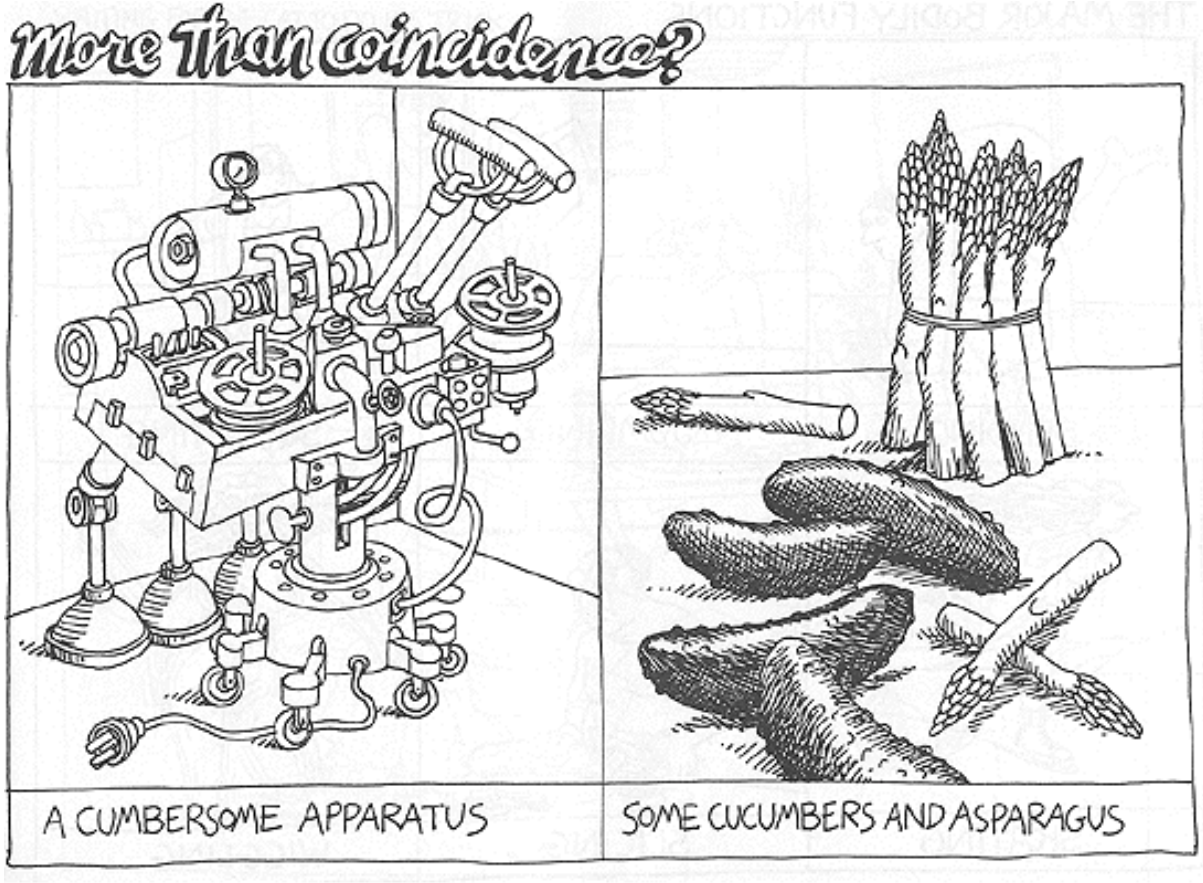


Figure: Pairwise scatterplot of of Iris measurements based on four principal component directions. Colors indicate species: *Setosa*, *Virginica*, *Versicolor*.

Finding Patterns in Data



Drawing by B. Kliban

Exploratory Data Analysis

EDA: First look at a data set, typically in the form of a matrix of numbers

- Data visualization
- Identifying patterns or regularities of interest
- Hypothesis generation

Exploratory Data Analysis: Preliminaries

- Identifying and addressing outlying samples, variables, or entries
- Imputing missing values
- Transforming data values using logarithm or other functions
- Normalization: removing systematic differences between samples
- Checking distributional assumptions

Overview: Univariate Data Analysis

Univariate Sample: Sample $x = x_1, \ldots, x_n$ with $x_i \in \mathbb{R}$

- Sample mean $m(x) = \overline{x} = n^{-1} \sum_{i=1}^{n} x_i$
- Sample variance $s^2(x) = n^{-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$
- Sample standard deviation $s(x)$

Standardized sample: Replace x_i with $\tilde{x}_i = (x_i - \overline{x})/s(x)$

- Standardization ensures $m(\tilde{x}) = 0$ and $s(\tilde{x}) = 1$

Univariate Data, cont.

Rank based statistics

- Order statistics $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$
- αth percentile $= x_{(r)}$, where r is the integer closest to $n\,(\alpha/100) + 1/2$
- Special cases: first (25%), median (50%), and third (75%) quartiles

Visualizing Univariate Distributions

Histogram or density estimate based on $\{x_1, \ldots, x_n\}$. Note: need to specify bin size (for histogram) or bandwidth (for density)

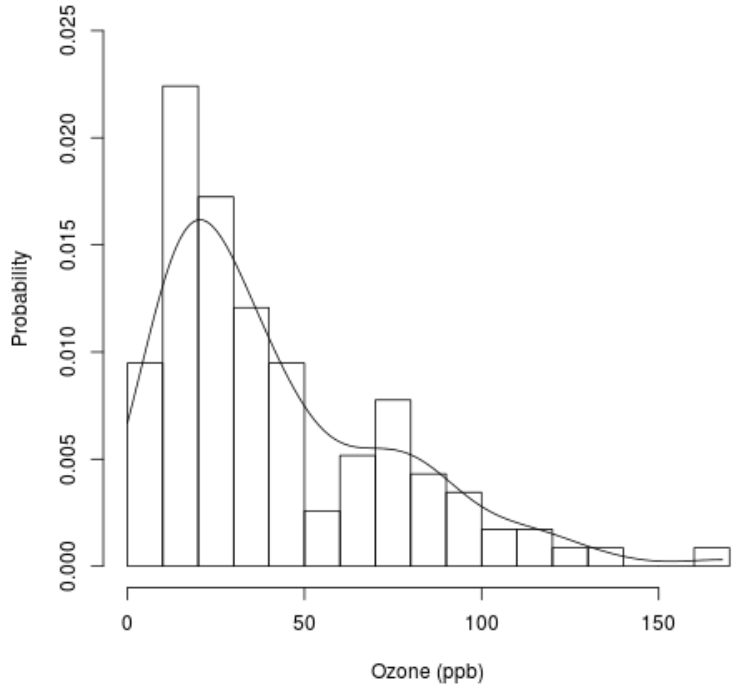


Figure: https://chemicalstatistician.wordpress.com

Visualizing Univariate Distributions

Empirical cumulative distribution function (CDF) of sample x. For each $t \in \mathbb{R}$

$$F_n(t) \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(x_i \leq t) \;=\; \frac{\#\,\text{data points } x_i \leq t}{n}$$

- "staircase shape" with jumps of size $1/n$ at each data point
- can recover dataset x from $F_n(t)$ apart from order

Normal QQ-plots

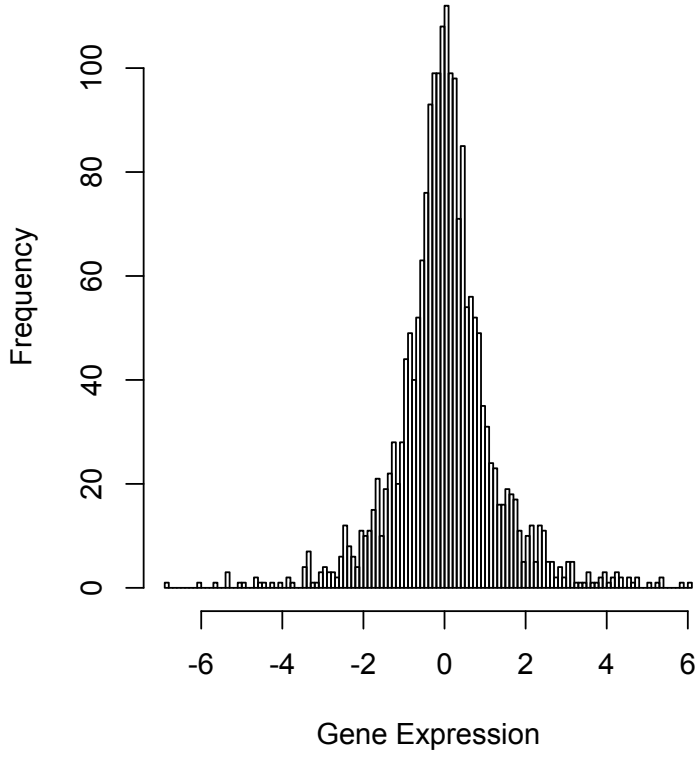
Recall: Normal QQ plot of dataset $x = x_1, \ldots, x_n$

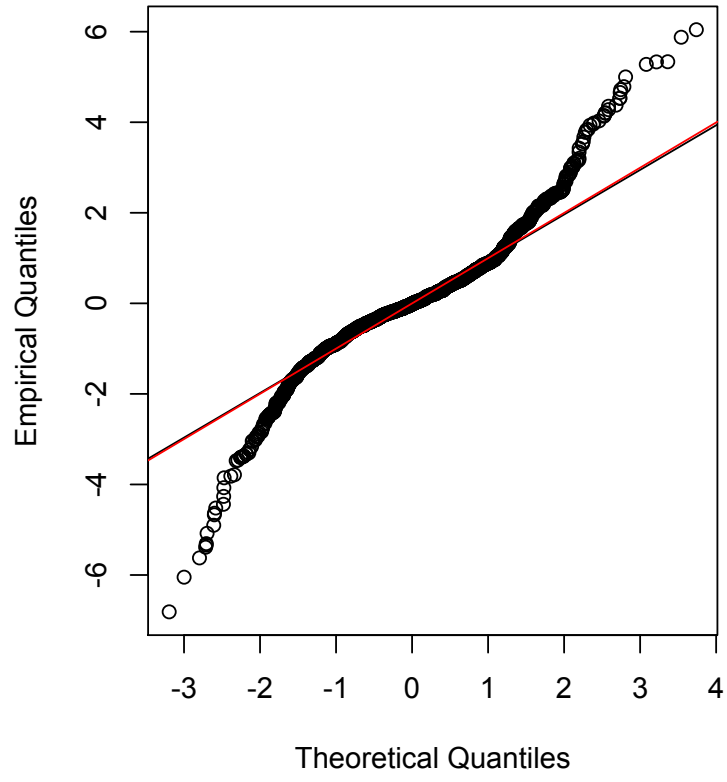
- x-axis is theoretical quantiles of standard normal CDF $\Phi(x)$
- Reference line $y = x$ represents ideal (normal) data
- QQ plot shows x_i versus $y_i = F_n^{-1}(\Phi(x_i))$

Interpretation

- If QQ plot *steeper* than reference line, then empirical distribution has heavier tails (more dispersed) than normal
- If QQ plot *shallower* than reference line, then empirical distribution has lighter tails (less dispersed) than normal

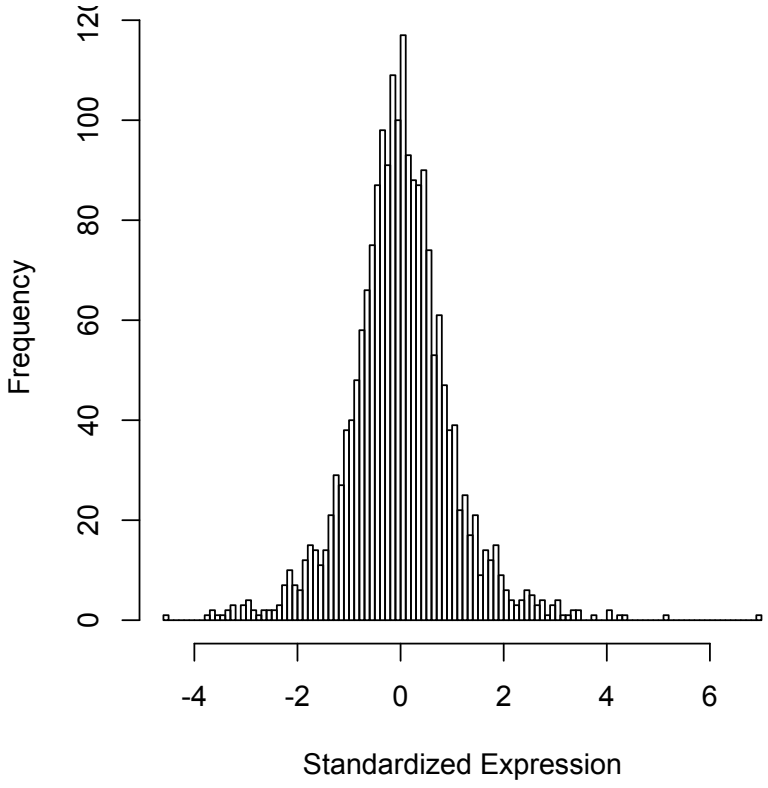
Histogram and QQ-plot, Gene Expression

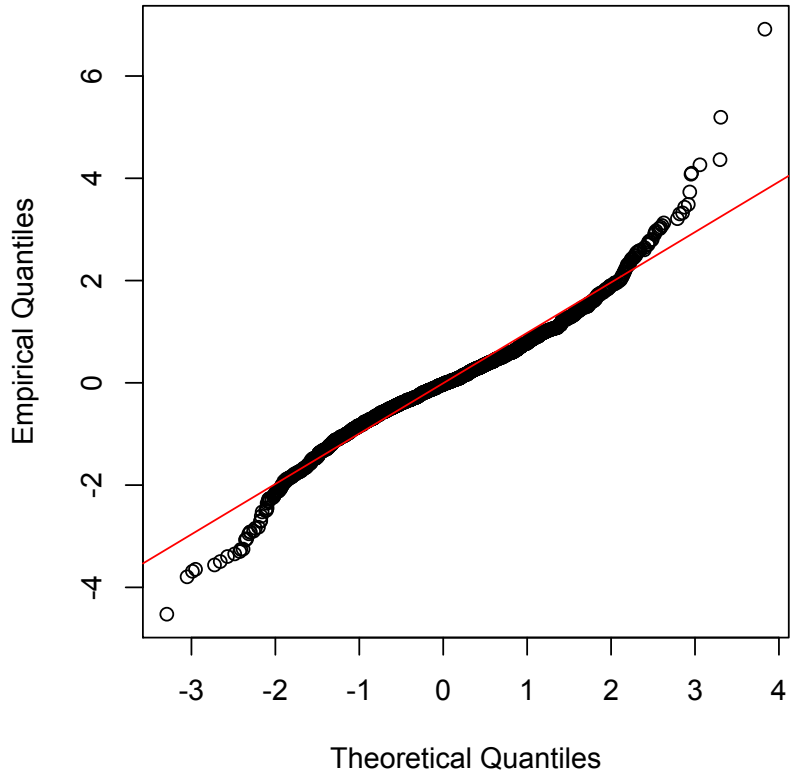




Note: Each figure based on 2000 measurements in first row of data matrix

Histogram and QQ-plot, Gene Expression after Standardization





Bivariate Data

Bivariate Sample: $(x, y) = (x_1, y_1), \ldots, (x_n, y_n)$ with $(x_i, y_i) \in \mathbb{R}^2$

- Univariate statistics $m(x), s^2(x)$ and $m(y), s^2(y)$.

- Sample covariance of x and y

$$s(x, y) \;=\; n^{-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \;=\; n^{-1} \sum_{i=1}^{n} x_i \, y_i - \overline{x} \, \overline{y}$$

- Sample correlation of x and y

$$r(x, y) \;=\; \frac{s(x, y)}{s(x) \, s(y)} \;\in\; [-1, 1]$$

Visualizing Bivariate Data: Scatter Plots

Recall: The *scatter plot* of $(x, y) = (x_1, y_1), \ldots, (x_n, y_n)$ is just the two-dimensional plot of the points (x_i, y_i). Typical uses

- Identifying outliers, looking for associations between x and y
- Identifying linear or nonlinear relationship between x and y

Utility of scatter plots derives from their flexibility, for example, one can

- Compare two samples or two features
- Compare (mean, median) or (mean, SD) across featuires
- Compare the SDs of features under two experimental conditions

The Regression Line

Given bivariate data (x, y) the *sample regression line of* y *on* x is the line $\ell^*(x)$ minimizing

$$\text{MSE}(\ell) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \ell(x_i))^2$$

over all linear functions $\ell(x) = ax + b$.

Fact: Sample regression line ℓ^* is given by

$$\ell^*(x) = m(y) + \frac{s(x,y)}{s^2(x)}[x - m(x)]$$

Moreover $\text{MSE}(\ell^*) = s^2(y)[1 - r^2(x,y)]$. Thus $r^2(x,y)$ tells us how much benefit there is in looking at x when predicting value of y.

High Dimensional Data: Empirical Covariance Matrices

Given: $n \times p$ data matrix $\mathbf{X}$ with

- rows/samples $\mathbf{x}_{i\cdot} = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$, with means $\overline{\mathbf{x}}_{i\cdot}$
- cols/features $\mathbf{x}_{\cdot k} = (x_{1k}, \ldots, x_{nk})^t$, $k = 1, \ldots, p$, with means $\overline{\mathbf{x}}_{\cdot k}$

Empirical covariance matrices

$$\text{Samples} \quad (\Sigma_{\mathsf{s}})_{ij} = s(\mathbf{x}_{i\cdot}, \mathbf{x}_{j\cdot}) = \frac{1}{p} \sum_{r=1}^{p} x_{ir} x_{jr} - \overline{\mathbf{x}}_{i\cdot}\, \overline{\mathbf{x}}_{j\cdot} \quad 1 \le i, j \le n$$

$$\text{Features} \quad (\Sigma_{\mathsf{f}})_{kl} = s(\mathbf{x}_{\cdot k}, \mathbf{x}_{\cdot l}) = \frac{1}{n} \sum_{r=1}^{n} x_{rk} x_{rl} - \overline{\mathbf{x}}_{\cdot k}\, \overline{\mathbf{x}}_{\cdot l} \quad 1 \le k, l \le p$$

Note: Matrix Σ_{s} is $n \times n$ and Σ_{f} is $p \times p$. Both are symmetric

Empirical Correlation Matrices

Definition: Empirical correlation matrices

$$\text{Samples} \quad (R_{\mathsf{s}})_{ij} = \frac{s(\mathbf{x}_{i\cdot}, \mathbf{x}_{j\cdot})}{s(\mathbf{x}_{i\cdot})\, s(\mathbf{x}_{j\cdot})} \quad 1 \leq i, j \leq n$$

$$\text{Features} \quad (R_{\mathsf{f}})_{kl} = \frac{s(\mathbf{x}_{\cdot k}, \mathbf{x}_{\cdot l})}{s(\mathbf{x}_{\cdot k})\, s(\mathbf{x}_{\cdot l})} \quad 1 \leq k, l \leq p$$