# Introduction to Data Analysis

Lecture 1

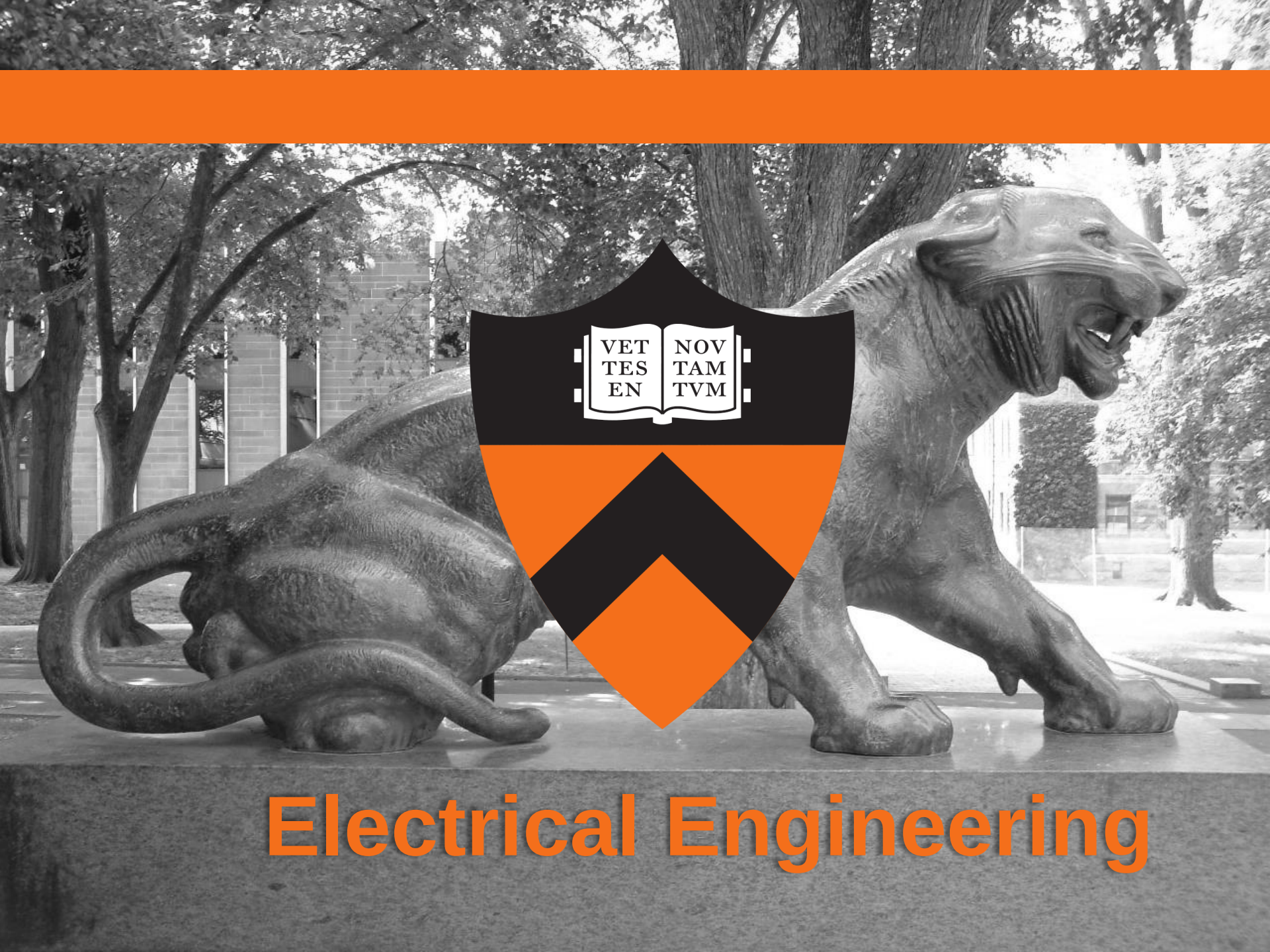
Thomas S. Woodson

# Exercice d' échauffement:

- Here is a list of exam grades. The highest grade you can get on the exam is 200.  How did the students do?
- Did the Master's or PhD students do better?
- Vous trouverez ci-dessous une liste des notes que les étudiants ont reçues lors de l' épreuve finale . La note la plus élevée que vous pouvez obtenir à l'examen est de 200.Comment les étudiants ont-ils réussi?
- Les étudiants de maîtrise ou de doctorat ont-ils fait mieux?

| Grades | Status |
|---|---|
| 140 | PHD |
| 130 | PHD |
| 170 | PHD |
| 100 | PHD |
| 160 | PHD |
| 180 | PHD |
| 120 | PHD |
| 200 | PHD |
| 160 | M |
| 140 | M |
| 140 | M |
| 160 | M |
| 150 | M |
| 190 | M |
| 160 | M |
| 180 | M |
| 120 | M |
| 200 | M |
| 170 | M |
| 150 | M |

ABOUT ME




# Electrical Engineering


PH.D
Public
Policy


# Technology and Society

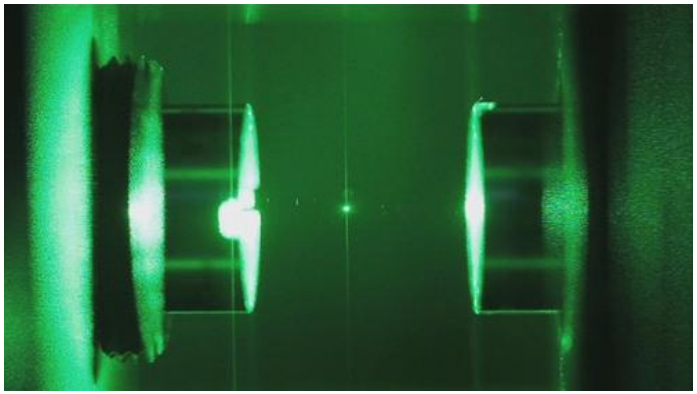
# About Stony Brook

# Department of Technology and Society

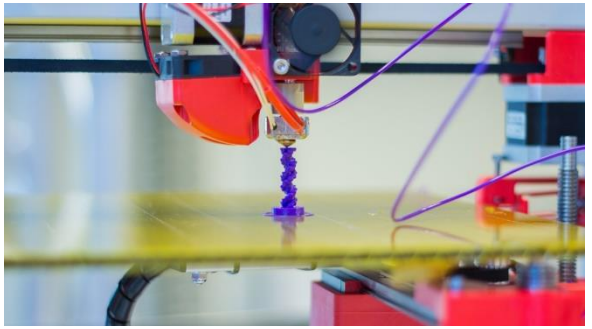
- College of Engineering and Applied Sciences
- Bachelors
  - Technological Systems Management
- Master's degree
  - Technological Systems Management
- PhD
  - Technology, Policy and Innovation Policy
- Collège d'ingénieurs et de sciences appliquées
- Les bacheliers
  - Gestion des systèmes technologiques
- Une maîtrise
  - Gestion des systèmes technologiques
- Doctorat
  - Technologie, politique et politique d'innovation

Wolfie the Seawolf

NANOTECHNOLOGY



3D PRINTING




ENGINEERING EDUCATION

STEM DIVERSITY

VILLAGE CHURCH

Education is Liberation

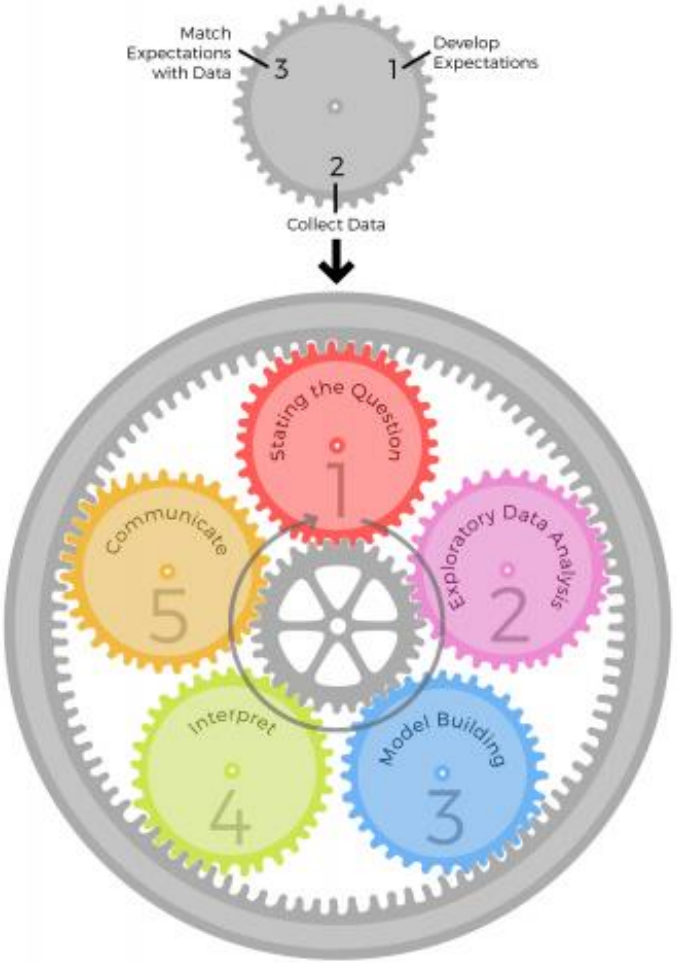
# Qu'est-ce que l'analyse de données (DA)?

Data analysis is sorting and characterizing data to better understand underlying phenomena and predict future events.

L'analyse des données consiste à trier et à caractériser les données pour mieux comprendre les phénomènes sous-jacents et prédire les événements futurs

# Misconceptions about DA

- Data analysis is a not a linear process. But rather it is an iterative non-linear process.
- L'analyse des données n'est pas un processus linéaire. Mais il s'agit plutôt d'un processus itératif non linéaire




source: The Art of Data Science

# How to start an analysis

Stating and refining the question.

- Asking the wrong question get the wrong answer.

Collecting and **Exploring and the data**
**Choosing** and building formal models to analyse the data
Interpreting the results
Drawing conclusion and Communicating the results

- Énoncer et affiner la question.
  - Poser la mauvaise question donne la mauvaise réponse.
- La collecte et l' **exploration et les données**
- **Choisir** et construire des modèles formels pour analyser les données
- Interpréter les résultats
- Conclusion et communication des résultats

# What are issues with data?

# Quelles sont les méthodes pour collecter des données?

- Non-response
- Missing values
- Coded incorrectly
- Data entry errors
- Non-réponse
- Valeurs manquantes
- Codé incorrectement
- Erreurs de saisie de données

# Collect Data/
# **Collecter des données**

What are methods to collect data?

- Survey
- Interview
- Measurements from tools
- Census
- Sondage
- Entrevue
- Mesures à partir d'outils
- Recensement

# What is R?
# **Qu'est ce que R?**

- Statistical computing and graphics language
- Predecessor program, S, was developed by Bell Labs in the 1970s
- Open Source
- Has wide range of applications
- Very powerful and expandable
- Informatique statistique et langage graphique
- Programme prédécesseur, S, a été développé par Bell Labs dans les années 1970
- Open source
- A une large gamme d'applications
- Très puissant et extensible

# Challenges with R **Défis avec R**

- No user support call center
- Big learning curve, but very powerful
- Need some programming skills
- If you don’t practice R you’ll forget
- Pas de centre d'appel utilisateur
- Grande courbe d'apprentissage, mais très puissante
- Besoin de quelques compétences en programmation
- Si vous ne pratiquez pas R, vous oublierez

# Installing R and R Studio

https://www.r-project.org/

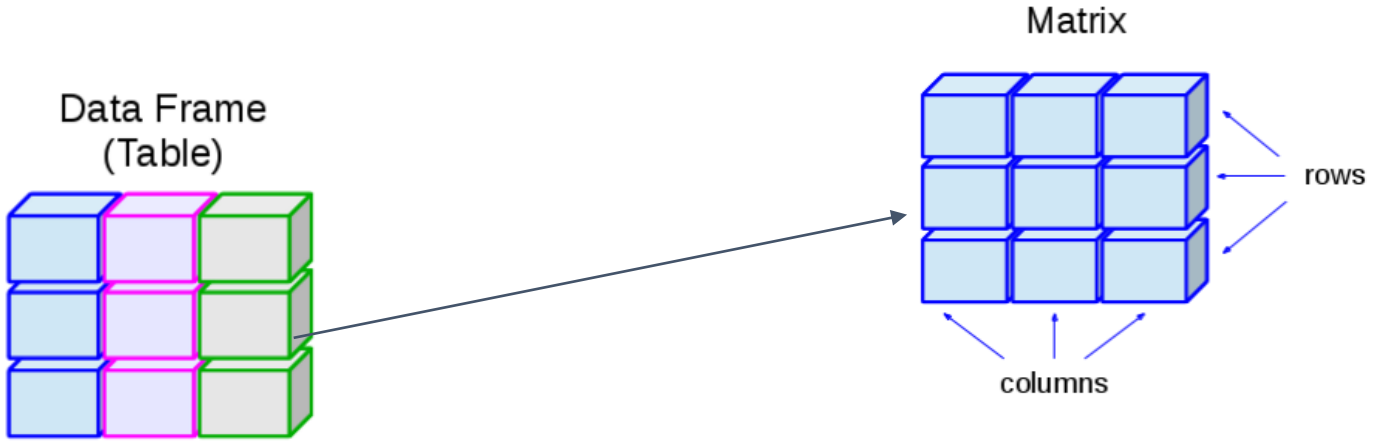
https://www.rstudio.com/

# R Practice

# Practice

- Create a sequence (2,4,6…20)
- Créer une séquence (2,4,6… 20). Appelle la sequence « seq1 »

# Structure of data(Matrix)




$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

$a_{jk}$= jth term (observation) on measurement of kth variable

Rows→Observations
Columns → Variable

**Rangées → Observations**
**Colonnes → Variable**

# Descriptive Statistics

What are descriptive statistics? Why do we do it?

**Que sont les statistiques descriptives? Pourquoi le faisons-nous?**

# Descriptive Statistics

## Mean

$$A = \frac{1}{n}\sum_{i=1}^{n} a_i$$

## Median

“numerical value separating the upper and lower halves of a data vector”

«Valeur numérique séparant les moitiés supérieure et inférieure d'un vecteur de données»

## Mode

“most common value”

«Valeur la plus commune»

# Inferential Statistics

What are inferential statistics?

**Quelles sont les statistiques inférentielles?**

What are the differences?

**Quelles sont les différences?**

# Descriptive/Inferential

Descriptive Statistics : Outlining the nature of the data

Inferential Statistics : Drawing the relationship between sample and population.

Statistiques descriptives: décrivant la nature des données

Statistiques inférentielles: établir la relation entre l'échantillon et la population.

# Variance

Measure of the “spread” of a variable

**Mesure de la «propagation» d'une variable**

$$Var(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$$

Variance of sample

Variance de l'échantillon

# Co-Variance

$$cov(x, y) = \frac{\sum(x - \bar{x})(y - \bar{y})}{n}$$

Co-variance measures the linear association between variables.
The unit of the variables can make a big difference in value

La co-variance mesure l'association linéaire entre les variables.
L'unité des variables peut faire une grande différence de valeur

# Covariance Matrix

| | SO2 | temp | manu | popul | wind | precip | predays |
|---|---|---|---|---|---|---|---|
| SO2 | 550.947561 | -73.560671 | 8527.7201 | 6711.9945 | 3.1753049 | 15.0017988 | 229.92988 |
| temp | -73.560671 | 52.239878 | -773.9713 | -262.3496 | -3.6113537 | 32.8629884 | -82.42616 |
| manu | 8527.720122 | -773.971341 | 317502.8902 | 311718.8140 | 191.5481098 | -215.0199024 | 1968.95976 |
| popul | 6711.994512 | -262.349634 | 311718.8140 | 335371.8939 | 175.9300610 | -178.0528902 | 645.98598 |
| wind | 3.175305 | -3.611354 | 191.5481 | 175.9301 | 2.0410244 | -0.2185311 | 6.21439 |
| precip | 15.001799 | 32.862988 | -215.0199 | -178.0529 | -0.2185311 | 138.5693840 | 154.79290 |
| predays | 229.929878 | -82.426159 | 1968.9598 | 645.9860 | 6.2143902 | 154.7929024 | 702.59024 |

# Correlation coefficient

- Varies from -1 to 1
- Sign indicates direction of association
- Is affected by outliers.
- Varie de -1 à 1
- Le signe indique la direction de l'association
- Est affecté par des valeurs aberrantes.
-

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}} * \sqrt{s_{kk}}} \qquad r_{ik} = \frac{\sum_{j=1}^{n}(x_{ji} - \overline{x_i})(x_{jk} - \overline{x_k})}{\sqrt{\sum_{i=1}^{n}(x_{ji} - \overline{x_i})^2} * \sqrt{\sum_{i=1}^{n}(x_{jk} - \overline{x_k})^2}}$$

# Correlation Matrix

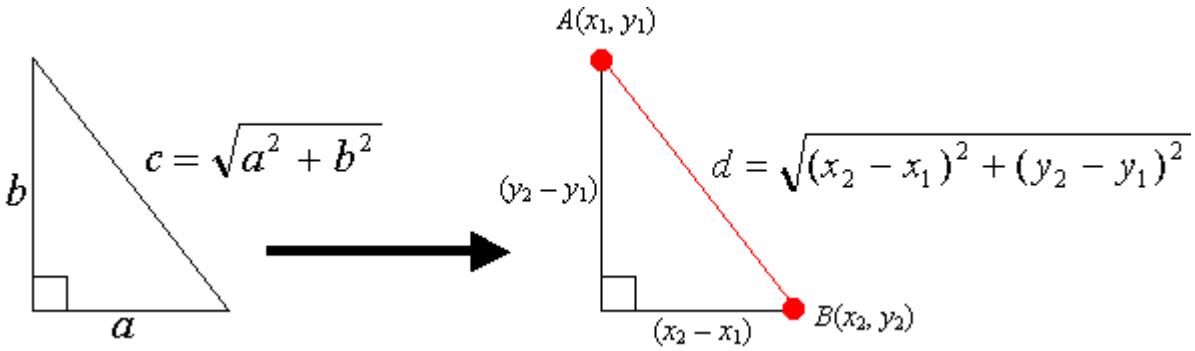
| | SO2 | temp | manu | popul | wind | precip | predays |
|---|---|---|---|---|---|---|---|
| SO2 | 1.00000000 | -0.43360020 | 0.64476873 | 0.49377958 | 0.09469045 | 0.05429434 | 0.36956363 |
| temp | -0.43360020 | 1.00000000 | -0.19004216 | -0.06267813 | -0.34973963 | 0.38625342 | -0.43024212 |
| manu | 0.64476873 | -0.19004216 | 1.00000000 | 0.95526935 | 0.23794683 | -0.03241688 | 0.13182930 |
| popul | 0.49377958 | -0.06267813 | 0.95526935 | 1.00000000 | 0.21264375 | -0.02611873 | 0.04208319 |
| wind | 0.09469045 | -0.34973963 | 0.23794683 | 0.21264375 | 1.00000000 | -0.01299438 | 0.16410559 |
| precip | 0.05429434 | 0.38625342 | -0.03241688 | -0.02611873 | -0.01299438 | 1.00000000 | 0.49609671 |
| predays | 0.36956363 | -0.43024212 | 0.13182930 | 0.04208319 | 0.16410559 | 0.49609671 | 1.00000000 |

# In Class Practice

- Find the mean of the rows
- Find the means of the columns
- Trouver la moyenne des lignes
- Trouver le moyen des colonnes

| | |
|---|---|
| 42 | 4 |
| 52 | 5 |
| 48 | 4 |
| 58 | 3 |

# Distance



# Statistical Distance

- **Statistical distance** quantifies the distance between two statistical objects, which can be t**wo random variables**, or **two probability distributions** or **samples**, or the distance can be between an individual **sample point** and **a population** or a wider sample of points.
- **La distance statistique** quantifie la distance entre deux objets statistiques, qui peuvent être **deux variables aléatoires** , ou **deux distributions de probabilité** ou **échantillons** , ou la distance peut être comprise entre un **point d'échantillonnage** individuel et **une population** ou un échantillon de points plus large.

$$d(Q,P) = \sqrt{\left(\frac{x_1 - y_1}{\sqrt{s_{11}}}\right)^2 + \left(\frac{x_2 - y_2}{\sqrt{s_{22}}}\right)^2} \qquad d(O,P) = \sqrt{\left(\frac{x_1}{\sqrt{s_{11}}}\right)^2 + \left(\frac{x_2}{\sqrt{s_{22}}}\right)^2}$$

Distance from Q to P