Lecture 2 Data Presentation

BIO210 Biostatistics

Xi Chen

Fall, 2023

School of Life Sciences
Southern University of Science and Technology

南方科技大学生命科学学院
SUSTech · SCHOOL OF
LIFE SCIENCES

Data Presentation

- Types of numerical data
- Tables and graphs

Types of numerical data

- **Nominal data** (categorical, unordered)
  - types of films (sci-fi, thriller, horror ...)
  - gender (male/female), status of a switch (on/off)
  - blood types (A/B/AB/O or Rh+/Rh-)
- **Ordinal data** (categorical, ordered)
  - Game/Music/App rating
  - Customer satisfactory survey
- **Discrete data** (quantitative, countable)
  - Number of email/messages received per day
  - Number of cars passing a traffic light per hour
- **Continuous data** (quantitative, not countable)
  - Time, height, weight $etc.$

What are the lengths of human genes in base pairs?

2540, 15166, 67, 1555, 137, 1527, 839, 6518, 6166, 44428, 1554, 3811, 754, 283, 549, 32388, 103, 1079, 1478, 10194, 67, 101817, 1800, 384, 7195, 157539, 362, 994, 2805, 2016, 103, 241725, 7139, 371, 1043, 1542, 88, 681, 206, 680, 546, 423, 994, 2811, 52741, 103, 1078, 31318, 113, 16833, 2466, 34308, 1316, 7976, 8832, 1057, 2705, 11142, 3513, 800, 4151, 20653, 15106, 5135, 9383, 6889, 2085, 1210, 13402, 153, 1196, 35998, 1083, 9647, 1080, 3339, 34543, 7013, 7039, 94, 89, 82, 986, 6499, 24056, 3084, 2813, 15159, 2804, 4276, 1557, 19976, 5197, 11593, 17219, 60, 3080, 13106, 64, 1108, 4140, 4034, 14142, 58, 9088, 1765, 366, 13624, 2526, 5384, 292, 1522, 3349, 2446, 1659, $\cdots$

Tables: Frequency distributions

Frequency distributions:

- a set of classes along with the numerical counts that correspond to each one.

The Premier League

Team	# of championships
Arsenal	3
Blackburn Rovers	1
Chelsea	5
Leicester City	1
Liverpool	1
Manchester City	6
Manchester United	13

Human Transcription Factors

Family	# of TFs
Zinc figner (C2H2)	868
Homeobox	247
Helix-loop-helix	107
bZip	54
Forkhead	51
STAT	7

Tables: Frequency distributions

Quantitative data:

- Break down the range of the values into non-overlapping intervals
- Trade-off: number of intervals vs information details
- Interval width: equal (but not always)

Lengths of human genes

Length (bp)	# of genes
1-500	14,065
501-1,000	6,603
1,001-5,000	11,867
5,001-50,000	18,567
50,001-100,000	4,485
100,001-2,473,539	5,030

# of published single cells/year

Year	# of single cells
2009	253
2010	421
2011	109
2012	1,978
2013	1,366
2014	23,707
2015	97,793
2016	461,341
2017	1,095,421
2018	6,514,313
2019	16,998,165
2020	20,729,842
2021	19,778,6518

Tables: Relative frequency

- In fraction (0.1) or percentage (10%)
- Comparison
- Unequal sizes

ABO blood groups in different places (Peng, 1991)

Absolute frequency

	A	B	O	AB
Beijing	1,032	1,268	1,195	376
Hubei	20,176	15,429	20,810	5,411
Guangdong	8,856	9,115	15,282	2133

Relative frequency

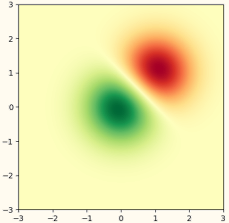
	A	B	O	AB
Beijing	26.66	32.76	30.87	9.71
Hubei	32.63	24.96	33.66	8.75
Guangdong	25.03	25.76	43.19	6.03

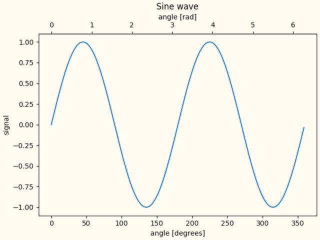
Tables: Cumulative relative frequency

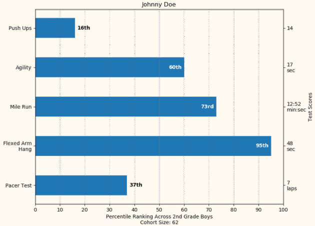
# of single cells published in 2021

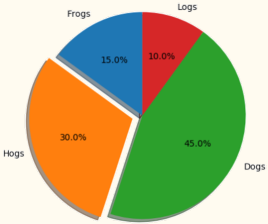
Month	Relative frequency (%)	Cumulative relative frequency
January	4.05	4.05
February	11.26	15.31
March	11.03	26.34
April	5.39	31.73
May	9.67	41.4
June	4.35	45.75
July	13.47	59.22
August	5.95	65.17
September	0.44	65.61
October	19.2	84.81
November	13.94	98.75
December	1.25	100

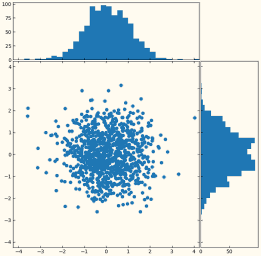
Graphs

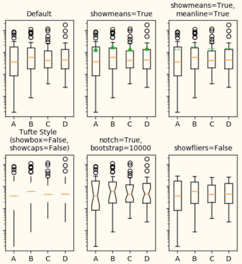



Sine wave
angle [rad]
signal
angle [degrees]

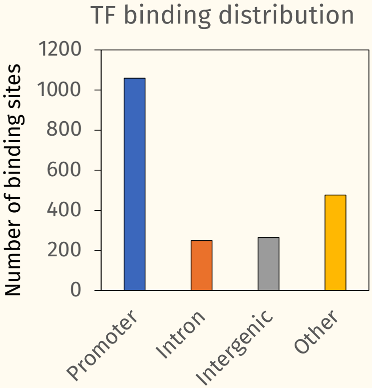

Johnny Doe
Push Ups
16th
14
Agility
60th
17 sec
Mile Run
73rd
12:52 min:sec
Flexed Arm Hang
95th
48 sec
Pacer Test
37th
7 laps
Test Scores
Percentile Ranking Across 2nd Grade Boys
Cohort Size: 62

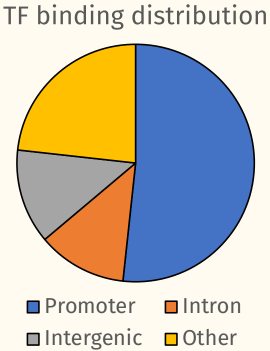

Frogs
Logs
Hogs
Dogs
15.0%
10.0%
30.0%
45.0%




Default
showmeans=True
showmeans=True, meanline=True
Tufte Style (showbox=False, showcaps=False)
notch=True, bootstrap=10000
showfliers=False

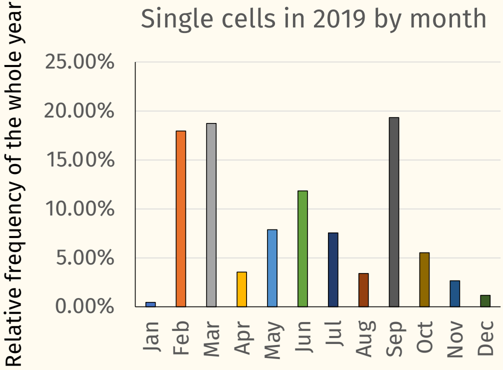
Graphs


TF binding distribution
Number of binding sites
1200
1000
800
600
400
200
0
Promoter
Intron
Intergenic
Other

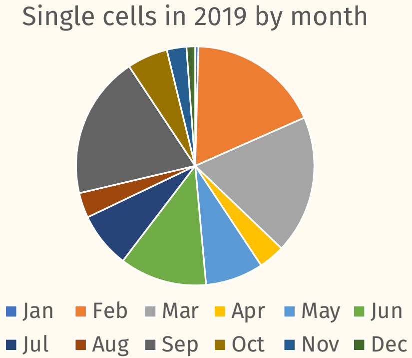

TF binding distribution
Promoter
Intron
Intergenic
Other

Graph

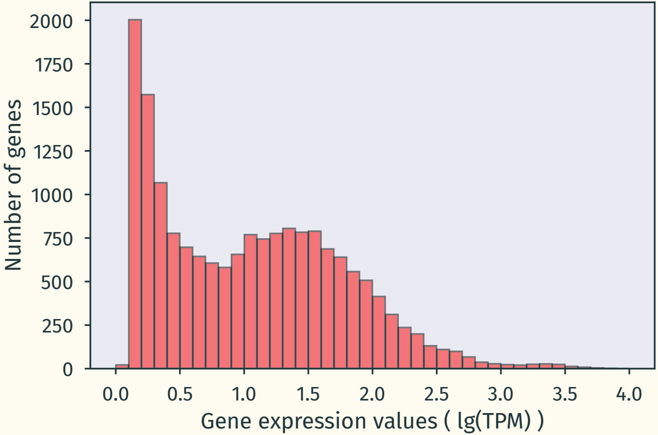
Bar chart



Pie chart



Histogram



40 intervals:

$[0,\ 0.1)$

$[0.1,\ 0.2)$

$[0.2,\ 0.3)$

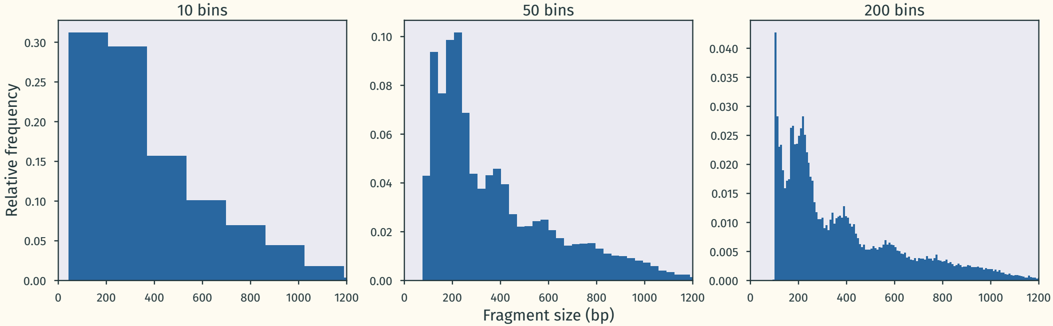
$\vdots$

$[3.8,\ 3.9)$

$[3.9,\ 4.0)$

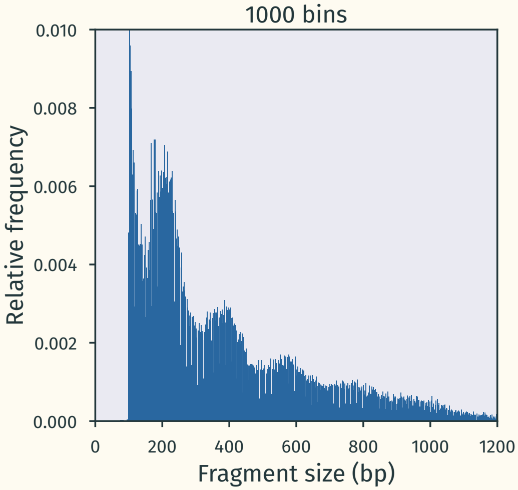
Histogram

Different number of intervals:



Histogram

Too many intervals:



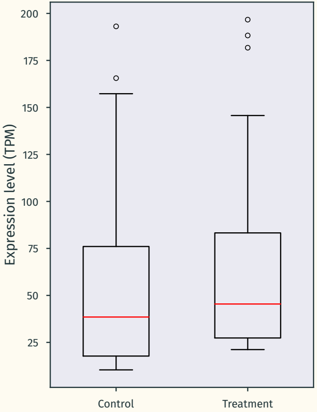
Percentile (quantile)

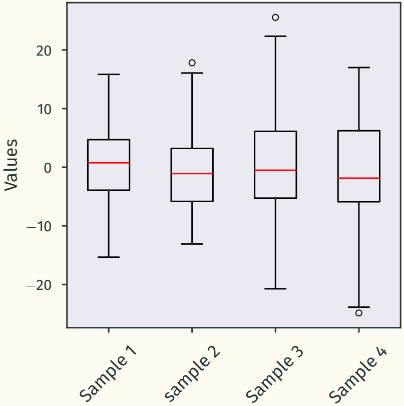
- Declarative definition: the $k-$th percentile of a data set is the value that divides the data, such that $k\%$ of the data points are smaller or equal to $(\leqslant)$ that value.
- Imperative definition: to find the $k-$th percentile of a data set with size n, perform the following steps:

1) sort the data from smallest to the largest
2) If $nk/100$ is an integer, the $k-$th percentile of the data is the average of the $(nk/100)$th and $(nk/100+1)$th largest observations
3) If $nk/100$ is NOT an integer, the $k-$th percentile of the data is the $(j+1)$th largest observation, where j is the largest integer that is less than $nk/100$.

Practice: What are the 25th and 50th percentiles of the first 10 prime numbers?

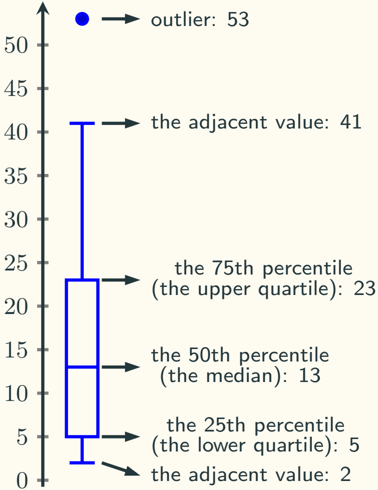
Box plot


200
175
150
125
100
75
50
25
Expression level (TPM)
Control
Treatment


20
10
0
−10
−20
Values
Sample 1
sample 2
Sample 3
Sample 4

The box plot anatomy

Draw a box plot of the following data ($n = 11$): [2, 3, 5, 7, 11, 13, 17, 19, 23, 41, 53]

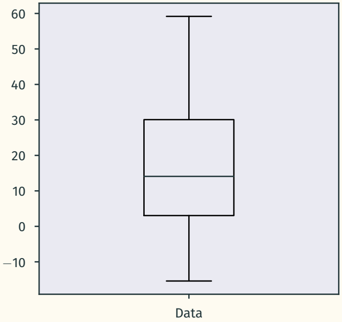


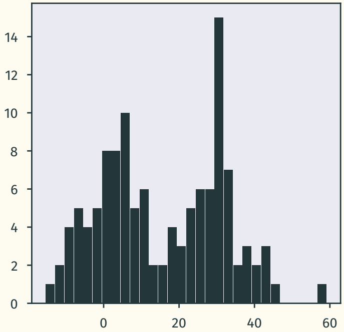
To make a boxplot, find the following key points:

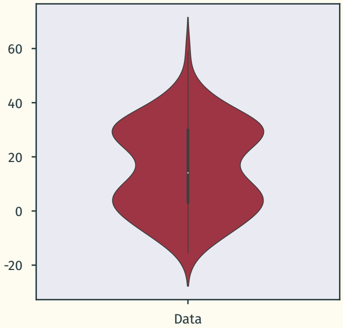
- The 25th percentile (the lower quartile): $11 \times 25/100 = 2.75$, so the lower quartile is the 3rd value: 5
- The 50th percentile (the median): $11 \times 50/100 = 5.5$, so the 50th percentile is the 6th value: 13
- The 75th percentile (the upper quartile): $11 \times 75/100 = 8.25$, so the upper quartile is the 9th largest value: 23
- The interquartile range (IQR): this is the difference between the 75th and 25th quartiles, which is $23 - 5 = 18$
- The adjacent values: these are the most extreme values that are between the lower quartile - 1.5 x IQR and the upper quartile + 1.5 x IQR. The lower quartile - 1.5 x IQR is 5 - 1.5 x 18 = -22, and the upper quartile + 1.5 x IQR is 23 + 1.5 x 18 = 50. Therefore, the most extreme values of the data that are within the range of [-22, 50] are 2 and 41
- Whiskers: draw extended lines (called whiskers) to the adjacent values
- Outliers: mark any values that are outside [-22, 50] with small circles, in this case, is 53

Graphs

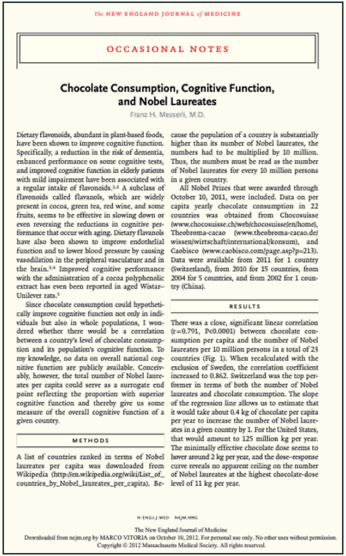
Box plot vs Violin plot







Scatter plot



The NEW ENGLAND JOURNAL of MEDICINE

OCCASIONAL NOTES

Chocolate Consumption, Cognitive Function, and Nobel Laureates

Franz H. Messerli, M.D.

Dietary flavonoids, abundant in plant-based foods, have been shown to improve cognitive function. Specifically, a reduction in the risk of dementia, enhanced performance on some cognitive tests, and improved cognitive function in elderly patients with mild impairment have been associated with a regular intake of flavonoids.[1,2] A subclass of flavonoids called flavanols, which are widely present in cocoa, green tea, red wine, and some fruits, seems to be effective in slowing down or even reversing the reductions in cognitive performance that occur with aging. Dietary flavanols have also been shown to improve endothelial function and to lower blood pressure by causing vasodilation in the peripheral vasculature and in the brain.[3,4] Improved cognitive performance with the administration of a cocoa polyphenolic extract has even been reported in aged Wistar–Unilever rats.[5]

Since chocolate consumption could hypothetically improve cognitive function not only in individuals but also in whole populations, I wondered whether there would be a correlation between a country's level of chocolate consumption and its population's cognitive function. To my knowledge, no data on overall national cognitive function are publicly available. Conceivably, however, the total number of Nobel laureates per capita could serve as a surrogate end point reflecting the proportion with superior cognitive function and thereby give us some measure of the overall cognitive function of a given country.

METHODS

A list of countries ranked in terms of Nobel laureates per capita was downloaded from Wikipedia (http://en.wikipedia.org/wiki/List_of_countries_by_Nobel_laureates_per_capita). Because the population of a country is substantially higher than its number of Nobel laureates, the numbers had to be multiplied by 10 million. Thus, the numbers must be read as the number of Nobel laureates for every 10 million persons in a given country.

All Nobel Prizes that were awarded through October 10, 2011, were included. Data on per capita yearly chocolate consumption in 22 countries was obtained from Chocosuisse (www.chocosuisse.ch/web/chocosuisse/en/home), Theobroma-cacao (www.theobroma-cacao.de/wissen/wirtschaft/international/konsum), and Caobisco (www.caobisco.com/page.asp?p=213). Data were available from 2011 for 1 country (Switzerland), from 2010 for 15 countries, from 2004 for 5 countries, and from 2002 for 1 country (China).

RESULTS

There was a close, significant linear correlation ($r=0.791$, $P<0.0001$) between chocolate consumption per capita and the number of Nobel laureates per 10 million persons in a total of 23 countries (Fig. 1). When recalculated with the exclusion of Sweden, the correlation coefficient increased to 0.862. Switzerland was the top performer in terms of both the number of Nobel laureates and chocolate consumption. The slope of the regression line allows us to estimate that it would take about 0.4 kg of chocolate per capita per year to increase the number of Nobel laureates in a given country by 1. For the United States, that would amount to 125 million kg per year. The minimally effective chocolate dose seems to hover around 2 kg per year, and the dose–response curve reveals no apparent ceiling on the number of Nobel laureates at the highest chocolate-dose level of 11 kg per year.

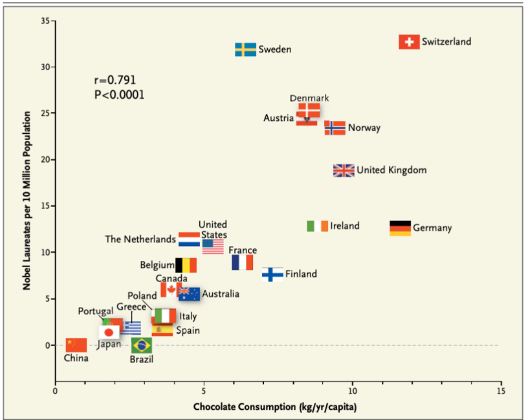


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Line graph

Time series data: