# Lecture 2: Data Summary and Visualization

Ye Tian

Department of Statistics, Columbia University
Calculus-based Introduction to Statistics (S1201)

July 6, 2022

# Review: basic concepts

- A **sample** represents a subset of the cases and is often a small fraction of the population.
- A (summary) (sample) **statistic** indicates some characteristics of the **dataset**
- Usually what we are really interested in is some **parameters** of the **population**
- We **estimate/infer** this information from the samples

# Review: descriptive statistics

- **Sample mean**: $\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$
- **Sample median**:
  - First rank the samples: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ (we call them as **order statistics**)
  - Then

$$x_{\text{median}} = \begin{cases} x_{((n+1)/2)}, & \text{if } n \text{ is odd} \\ \frac{1}{2}[x_{(n/2)} + x_{((n/2)+1)}], & \text{if } n \text{ is even} \end{cases}$$

- **Sample percentiles/quantiles**: $k$th percentile or $k\%$ quantile is the value where $k\%$ of the values are below.
- **Sample quartiles** (not **quantiles**)
  - $25\%$ quantile (25th percentile, lower forth) is also called **the 1st quartile** (Q1)
  - $50\%$ quantile (50th percentile) is also called **the 2nd quartile** (Q2), i.e. the **median**
  - $75\%$ quantile (75th percentile, upper forth) is also called **the 3rd quartile** (Q3)

# Review: descriptive statistics

- **Five number summary**: min, lower forth, median, upper forth, max
- **Range** = maximum − minimum
- **Interquartile range** (IQR) = Q3 − Q1
- **Sample variance** is roughly the average squared deviation from the mean.

  $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

  where $\bar{x}$ is the sample mean
- **Sample standard deviation** is the square root of the sample variance, i.e. $s$ here.
- Trick of **interpolation** when calculating percentiles/quantiles
- Sample median is more "**robust**" than sample mean (w.r.t. extreme values in the data)

# Today's goal

- Know popular data visualization methods (new) and descriptive statistics (review)
  - Stem-and-leaf plot, histogram, box-plot ...
  - Descriptive statistics: mean, median, quartiles, variance, standard deviation
- Can do simple exploratory analysis by using visualization tools and summary statistics

# Data visualization (continuous data)

# Why data visualization?

"The greatest value of a picture is when it forces us to notice what we never expected to see" — John Tukey (1915-2000, American mathematician and statistician)

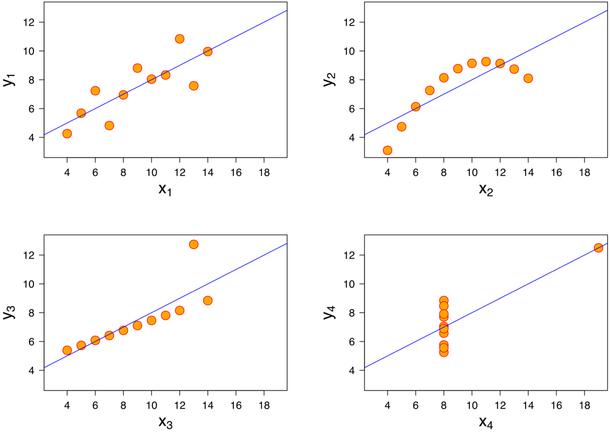
- Processing time in brain: Pictures $<$ Numbers $\ll$ Words
- Road signs (from DMV website)
  (1): Right lane ends - stay to the left
  (2): Merging traffic entering from right
  (3): School crossing
  (4): No U-turn

(1) (2) (3) (4)

# Why data visualization?

Sometimes summary statistics don't deliver enough information.
(especially when we have $\geq 2$ variables)

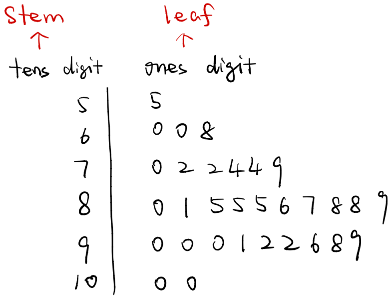
- Anscombe's quartet:




- All 4 figures: sample mean $\bar{x} = 9$, $\bar{y} = 7.5$
  sample variance $s_x^2 = 11$, $s_y^2 = 4.125$

# Stem-and-leaf plot[1]

The grades of final exams of this course last summer: 55, 60, 60, 68, 70, 72, 74, 74, 79, 80, 81, 85, 85, 85, 86, 87, 88, 88, 89, 90, 90, 90, 91, 92, 92, 96, 98, 99, 100, 100

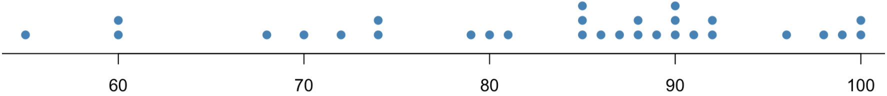



1. Select one or more leading digits for the stem values. The trailing digits become the leaves
2. List possible stem values in a vertical column
3. Record the leaf for each observation beside the corresponding stem value
4. Indicate the units for stems and leaves someplace in the display

[1] will not appear in exams

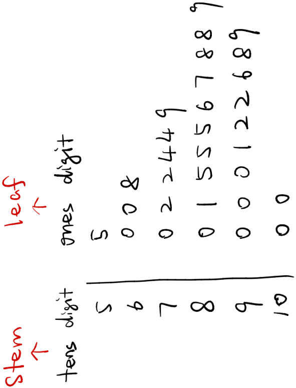
# Dot plot[2]

55, 60, 60, 68, 70, 72, 74, 74, 79, 80, 81, 85, 85, 85, 86, 87, 88, 88, 89, 90, 90, 90, 91, 92, 92, 96, 98, 99, 100, 100




- Compare the 90-degree rotation of stem-and-leaf plot (see the right) with the dot plot.

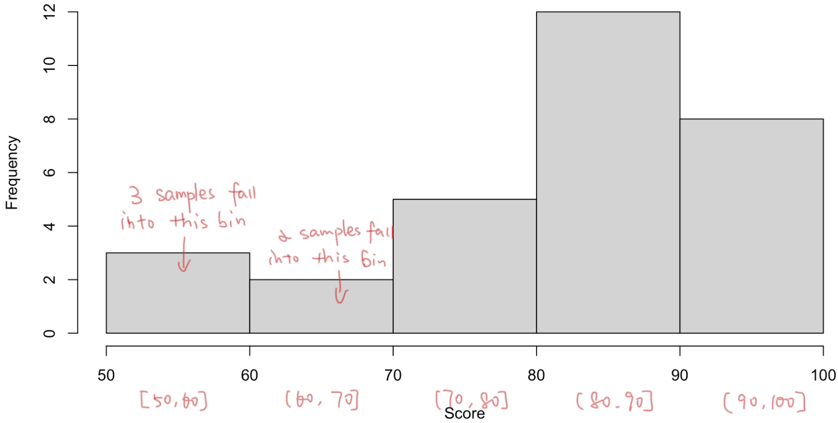
- Next let's see a more fancy version of these two plots — Histogram.



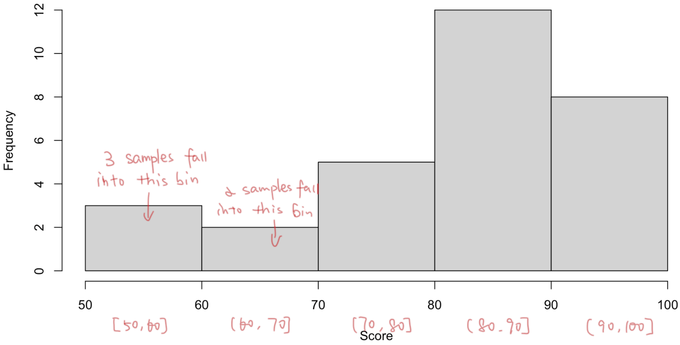

[2]will not appear in HWs or exams

# Histogram

55, 60, 60, 68, 70, 72, 74, 74, 79, 80, 81, 85, 85, 85, 86, 87, 88, 88, 89, 90, 90, 90, 91, 92, 92, 96, 98, 99, 100, 100

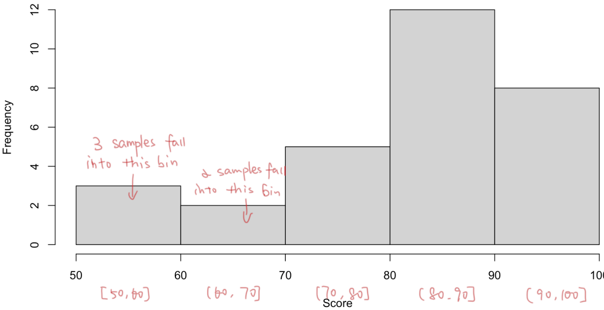


# Histogram




- The intervals are usually left open & right closed. (except the first one)
- Frequency is the number of samples falling into each bin
- The height of each bar indicates the frequency

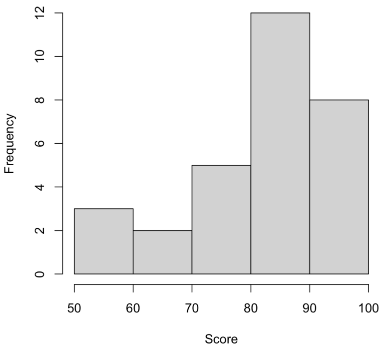
# Histogram

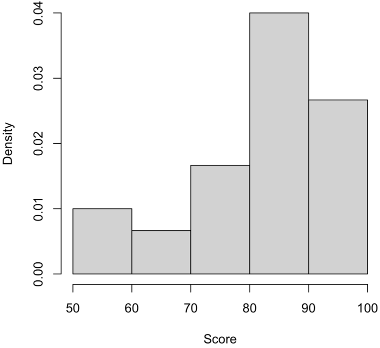



3 steps to draw a histogram:

- **Discretizing:** Divide the range of the data into bins or classes of equal width.
- **Counting:** Determine the frequency (counts) or relative frequency (frequency/sample size) for each bin.
- **Drawing:** Above each class interval, draw a rectangle whose height is the corresponding frequency or relative frequency.

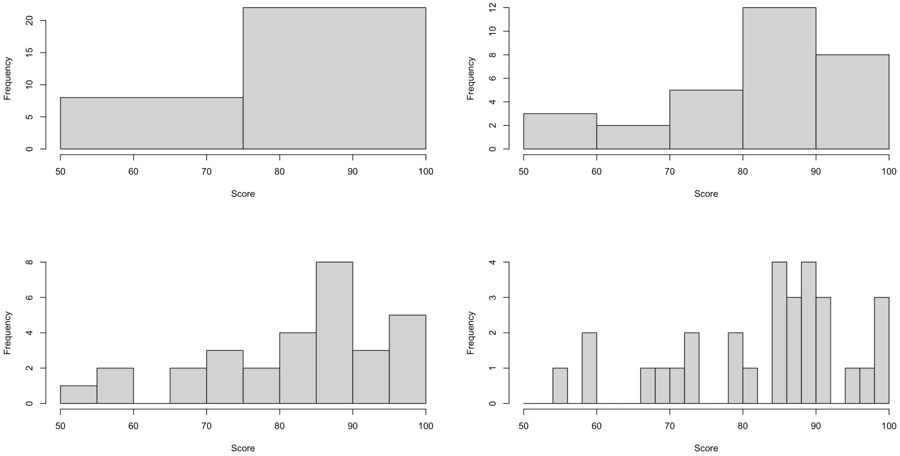
# Two types of histograms




- Relative frequency = frequency/sample size
- Here sample size = 30
- In the **density histogram**, relative frequencies are represented by **area**, NOT by height
- So the height of each block = $\frac{\text{relative frequency}}{\text{bin width}}$
- What's the sum of area of bins in the density histogram?

# Let's try different bin widths

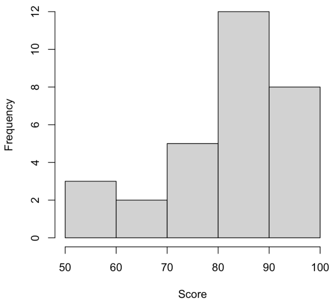


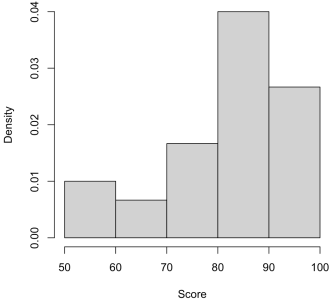

- Which one(s) are appropriate?
- It depends on what story we want to tell...

# From the frequency table to histogram

55, 60, 60, 68, 70, 72, 74, 74, 79, 80, 81, 85, 85, 85, 86, 87, 88, 88, 89, 90, 90, 90, 91, 92, 92, 96, 98, 99, 100, 100

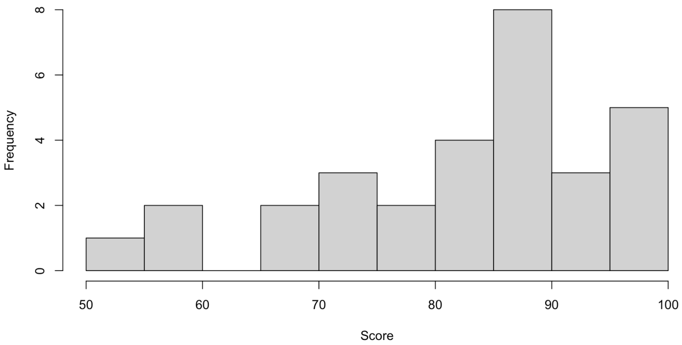
| Intervals | Count | Density |
|---|---|---|
| [50, 60] | 3 | 3/300 |
| (60, 70] | 2 | 2/300 |
| (70, 80] | 5 | 5/300 |
| (80, 90] | 12 | 12/300 |
| (90, 100] | 8 | 8/300 |
| | 30 | 1 |



# Information provided by histogram

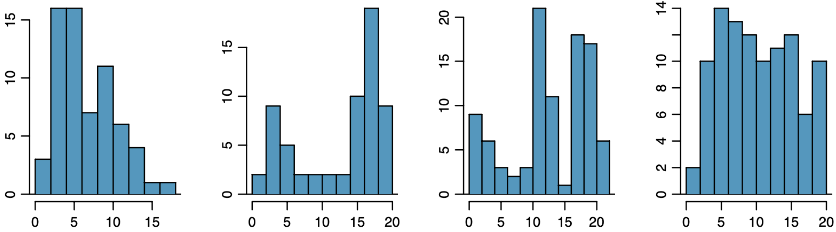
55, 60, 60, 68, 70, 72, 74, 74, 79, 80, 81, 85, 85, 85, 86, 87, 88, 88, 89, 90, 90, 90, 91, 92, 92, 96, 98, 99, 100, 100




- The data density
- The shape of data "distribution"
- Potential "outliners" (a data value that is far away from the **bulk** of data)
- The chosen bin width can alter the story that a histogram is telling.

# Shape of data distribution: modality

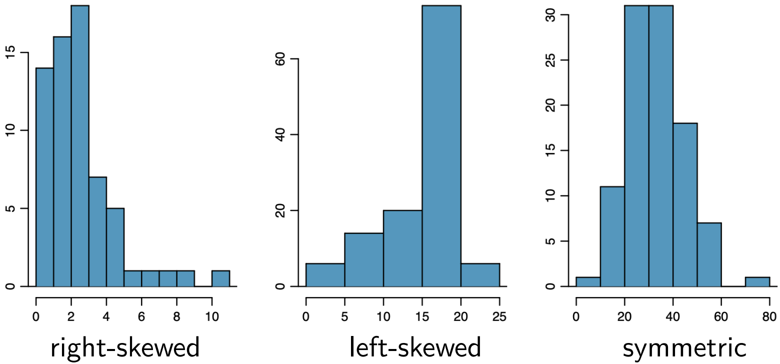
Does the histogram have a single prominent peak (**unimodal**), two prominent peaks (**bimodal**), more than two peaks (**multimodal**), or no apparent peaks (**uniform**)?




- In order to determine modality, step back and imagine a smooth curve over the histogram
- Can the bin width affect the modality?
- The modality is a **population-level** characteristic, which is **deterministic**
- Different choices of bin width or insufficient sample size can only affect our observation and conjecture, but not the modality of the distribution

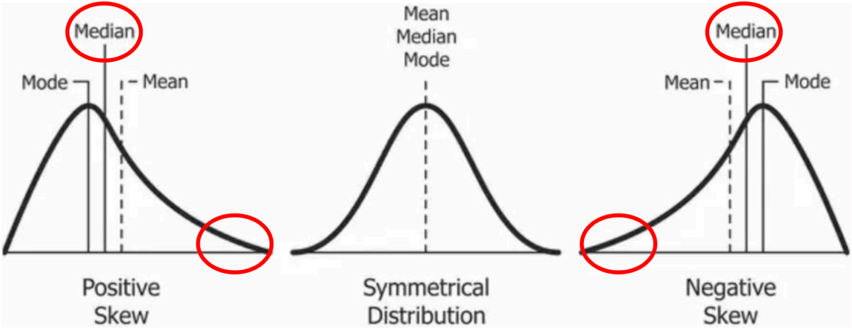
# Shape of data distribution: skewness

Left-skewed, right-skewed and symmetric **distributions**




- We usually only talk about skewness of **unimodal** distributions.
- How to memorize this? Left/right indicates the position of the **long tail**. i.e. *Histograms are said to be skewed to the side of the long tail*
- Again, skewness of distribution is deterministic, while skewness of a histogram can depend on the histogram (even for the same data)

# Shape of data distribution: skewness




(right-skewed) (symmetric) (left-skewed)

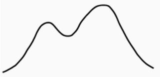
- Right-skewed: mode $<$ median $<$ mean
- Left-skewed: mode $>$ median $>$ mean
- Symmetric: mode $=$ median $=$ mean

# Commonly observed shapes of distributions

- modality

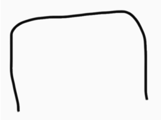
unimodal

bimodal



multimodal

uniform

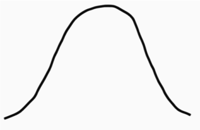


- skewness

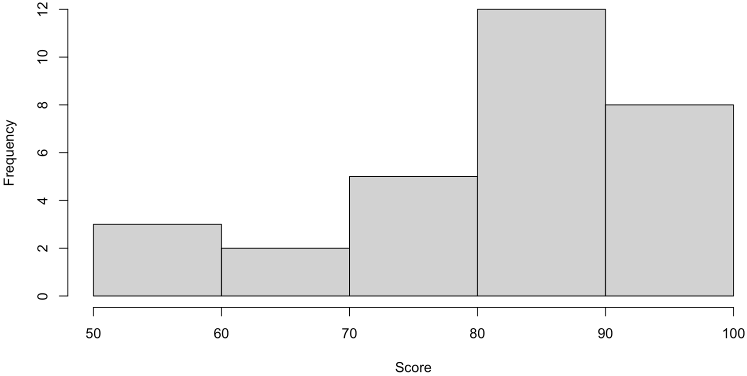
right skew

left skew

symmetric



# Example: final grades (revisited)




- Is the histogram unimodal?
- Is it skewed? If so, left-skewed or right-skewed?

# Example: NYC income statistics

NYC per household income per year (based on data of 2016-2020):

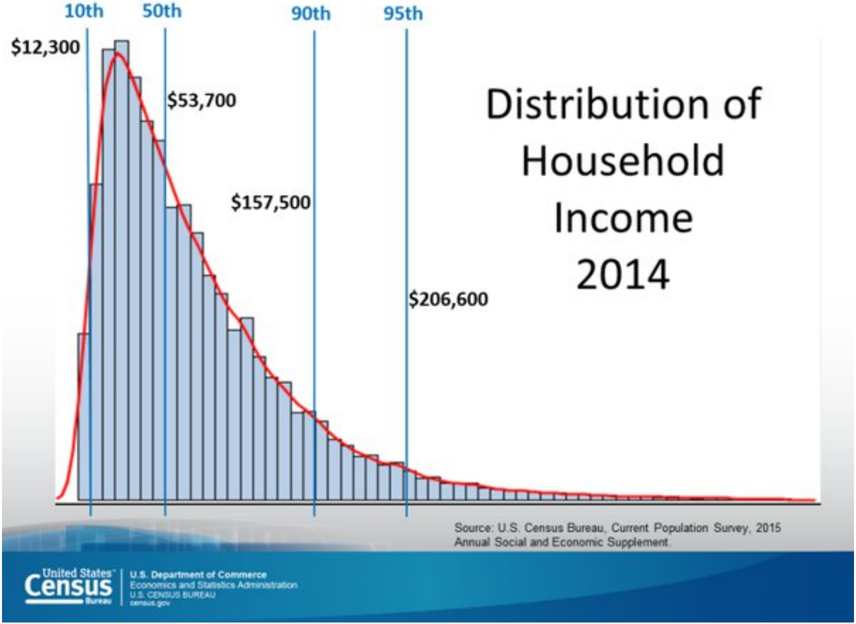
| | |
|---|---|
| **Average Household Income** | $107,000 |
| **Median Household Income** | $67,046 |

Is the household income skewed? How?

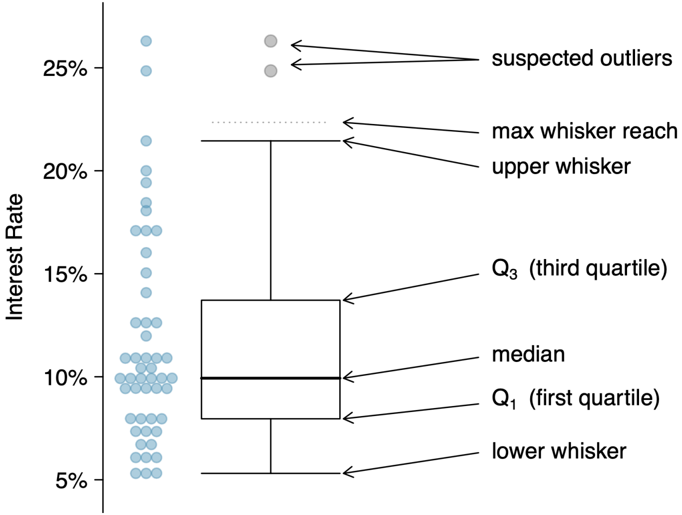
"The $238 Million Penthouse, and the Hedge Fund Billionaire Who May Rarely Live There" --New York Times [Read the story here]

More data available [here].

# Example: US household income



# Box plot



Interest Rate
25%
20%
15%
10%
5%
suspected outliers
max whisker reach
upper whisker
Q3 (third quartile)
median
Q1 (first quartile)
lower whisker

# How to draw a box plot

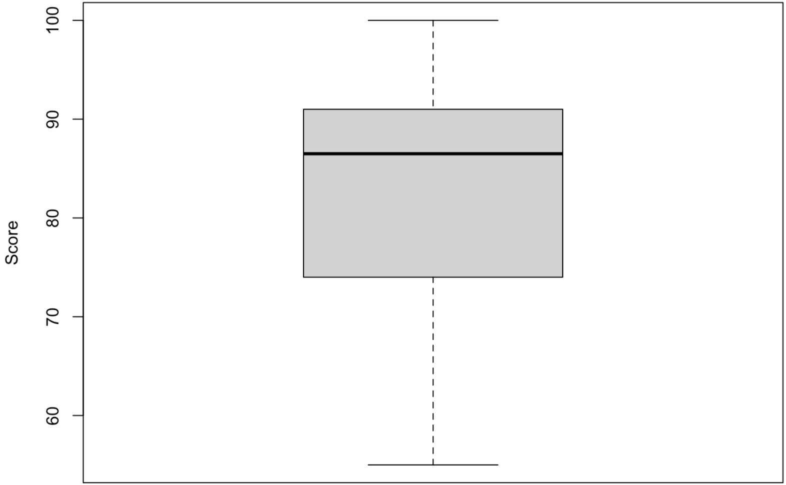

- Calculate Q1, median and Q3 $\rightarrow$ which gives us the location of the box
- Calculate the location of **whisker**
  - Max upper whisker reach $=$ Q3 $+ 1.5 \times$ IQR (not drawn)
  - Max lower whisker reach $=$ Q1 $- 1.5 \times$ IQR (not drawn)
  - **Upper whisker** extends to the maximum data value $\leq$ Max upper whisker reach
  - **Lower whisker** extends to the minimum data value $\geq$ Max lower whisker reach
- Samples outside beyond the maximum whisker reach are **potential outliners** $\rightarrow$ should be drawn separately in the box plot

# Example: final grades (revisited)

55, 60, 60, 68, 70, 72, 74, 74, 79, 80, 81, 85, 85, 85, 86, 87, 88, 88, 89, 90, 90, 90, 91, 92, 92, 96, 98, 99, 100, 100

- Calculate Q1, median and Q3 $\rightarrow$
  - Q1 $= 74$, median $= 86.5$, Q3 $= 90.2$, IQR $=$ Q3 $-$ Q1 $= 16.2$
- Calculate the location of **whisker**
  - Max upper whisker reach $=$ Q3 $+$ 1.5 $\times$ IQR $=$ 114.5
  - Max lower whisker reach $=$ Q1 $-$ 1.5 $\times$ IQR $=$ 49.7
  - Upper whisker: 100
  - Lower whisker: 55
- No potential outliners

# Example: final grades (revisited)



# Data visualization (discrete data)

# Frequency tables

Categories of passengers on Titanic: first class, second class, third class and crew

| Class | Count |
|---|---|
| First | 325 |
| Second | 285 |
| Third | 706 |
| Crew | 885 |

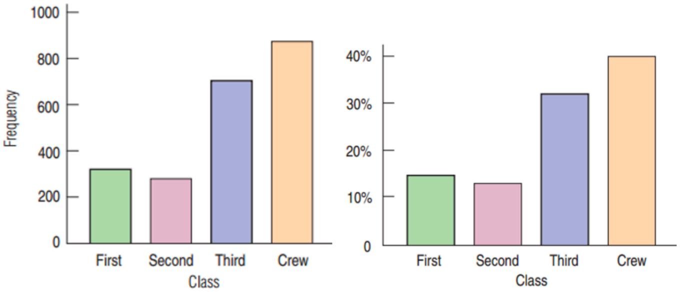
frequency table

| Class | % |
|---|---|
| First | 14.77 |
| Second | 12.95 |
| Third | 32.08 |
| Crew | 40.21 |

relative frequency table

- A **frequency table** is a table whose first column displays each distinct outcome and second column displays that outcome's frequency.
- A **relative frequency table** is a table whose first column displays each distinct outcome and second column displays that outcome's relative frequency.

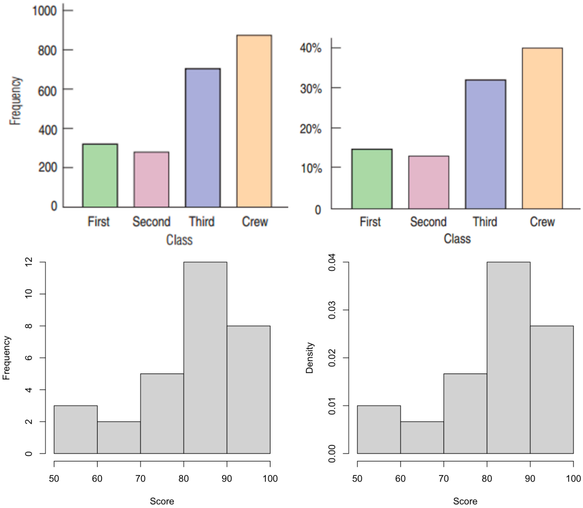
# Bar charts




- A **bar chart** displays the frequency or relative frequency of each category.
- Good for general audience.
- All bars must have **the same width** (The Area Principle[3])
- There should be some spaces between the bars.
- Bar orders can be arbitrary (usually according to the background info or by first occurrence of the category in the dataset)

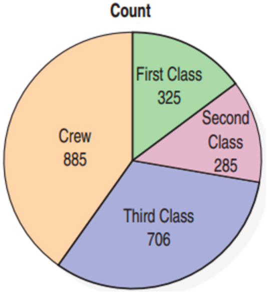
[3]The area occupied by a part of the graph should correspond to the magnitude of the value it represents

# Bar charts v.s. Histograms



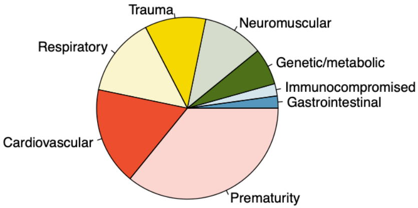

- Data type: Bar charts — categorical data
  Histograms — continuous data
- x-axis: Bar charts — categories, arbitrary order
  Histograms — numbers, must be ordered

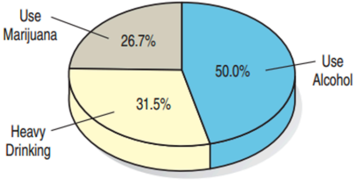
# Pie charts

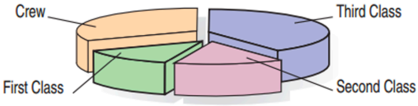



- A **pie chart** presents each category as a slice of a circle so that each slice has a size that is proportion to the whole in each category.
- Also good for a general audience.
- Help to display the fraction of the whole that each category represents.
- It is difficult to decipher details in a pie chart.
- Compared to pie charts, we **prefer bar charts**.

# Some other charts that suck...




Can you tell what problems these charts have?

# Reading list (optional)

(May overlap with previous reading lists)

- "Probability and Statistics for Engineering and the Sciences" (9th edition):
  - Chapter 1.2 and 1.4
- "OpenIntro statistics" (4th edition, free online, download `[here]`):
  - Chapter 2.1.1-2.1.6
  - Chapter 2.2.1 and 2.2.5

**Many thanks to**

- Chengliang Tang
- Anthony Donoghue
- Joyce Robbins
- Yang Feng
- Owen Ward
- Wenda Zhou
- And all my teachers in the past 25 years