Computer Science

CSC196: Analyzing Data

Introduction to Statistics and Data Analysis

Jason Pacheco and Cesim Erten

Outline

- Overview
- Sampling Procedures
- Measures of Location & Variability
- Graphical Diagnostics

Outline

- Overview
- Sampling Procedures
- Measures of Location & Variability
- Graphical Diagnostics

Example: Drug Selection

- Old drug is 80% effective
- New drug is 85% effective, but costs more
- Should we adopt the new drug?

But the 85% finding is based on a set of patients:

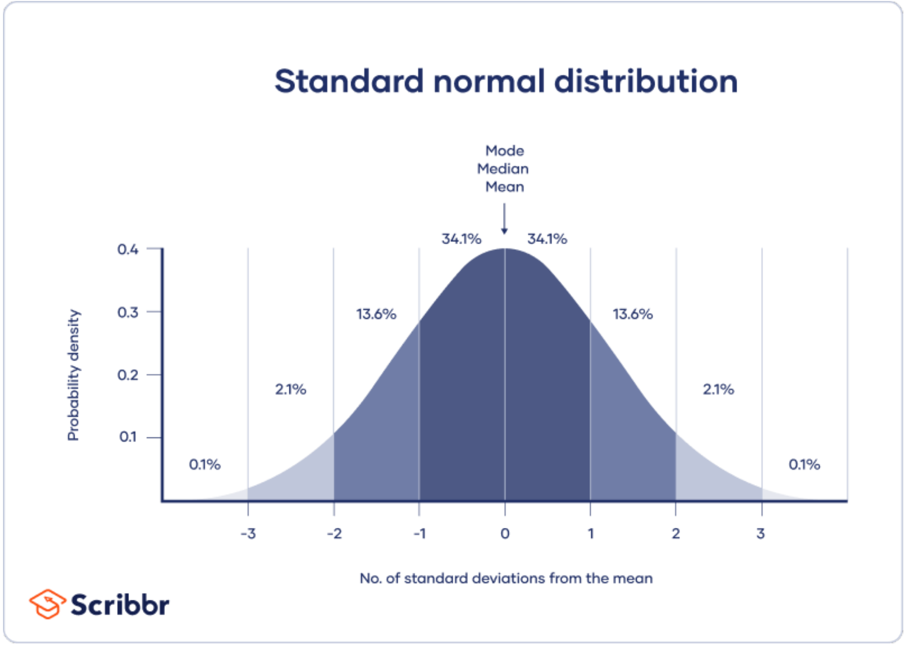
- Perhaps, if we run the trial again we will find that the new drug is only 75% effective…
- Natural variation from trial to trial must be accounted for
- Variation from patient to patient is endemic to the problem
- **Need to analyze sources of variation**

Variability in Scientific Data

If there were no variability in patient-to-patient data,
there would be no need for statisticians

Statisticians:

- Make use of fundamental laws of probability and statistical inference
- Draw conclusions (or inferences)
- Gather information as **samples** or collections of **observations**

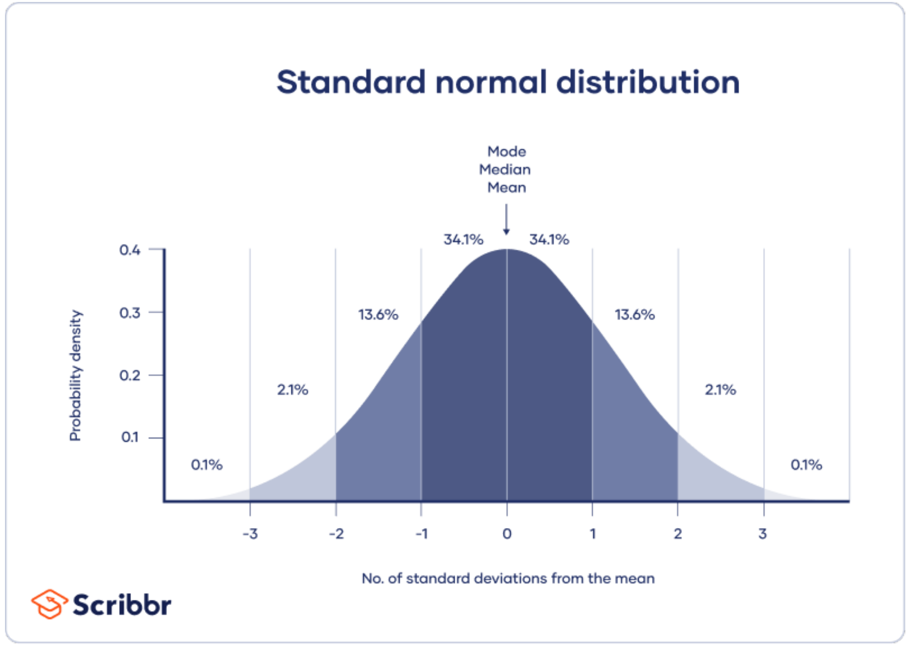


Descriptive Statistics

- Derive a set of single-number statistics from data

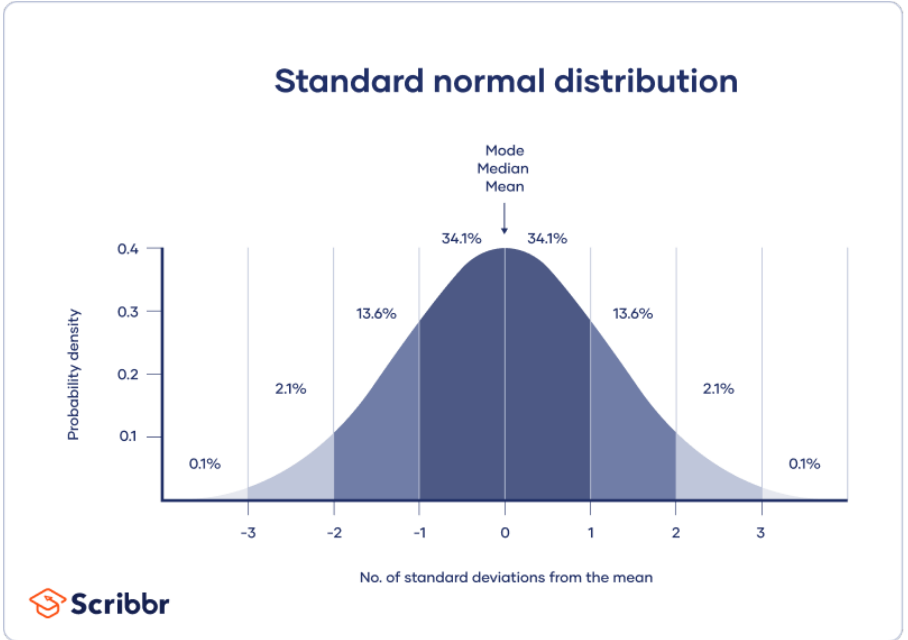
- Explain:
  - Location of the data
  - Variability of the data
  - General nature of the distribution of observations in a sample

- Show *footprint* of the nature of a sample via:
  - Mean
  - Median
  - Standard Deviation



Inferential Statistics

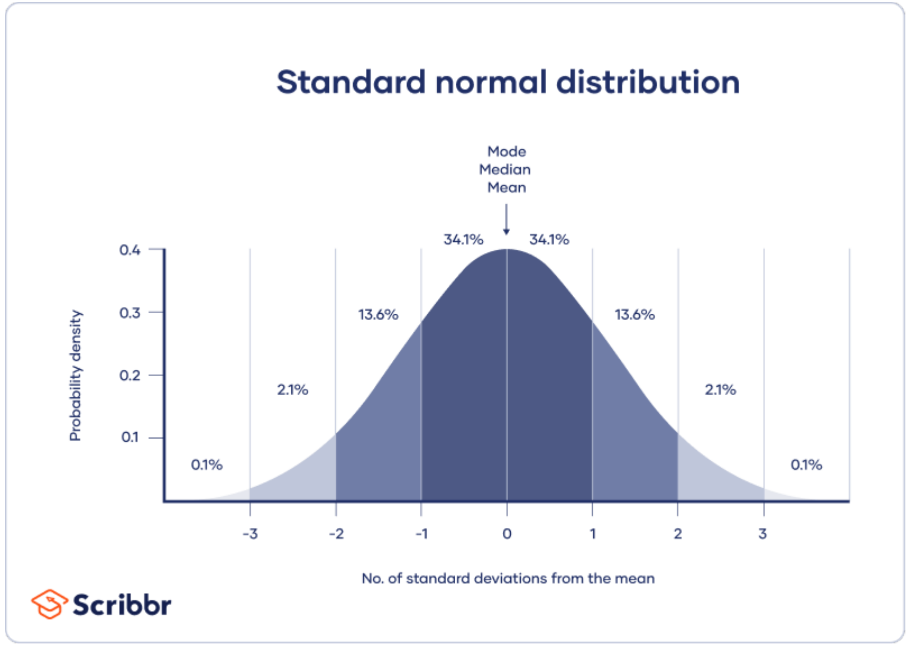
- Large *toolbox* of statistical methods employed by practitioners
- Goal: Make scientific judgements in the face of *uncertainty* and *variation*
- Often used to:
  - Analyze data from a *stochastic (random) process*
  - Determine how to improve process quality
  - Analyze sources of variation



Variability in Scientific Data

It is very important to collect scientific data in a systematic way

- Samples are collected from **populations**
- E.g. population of patients → all adults in a certain age range
- Typically focus on certain characteristics, or **factors**
- Ideally collected via **experimental design**
- Alternative is an **observational study**
- Both lend themselves to *statistical inference*

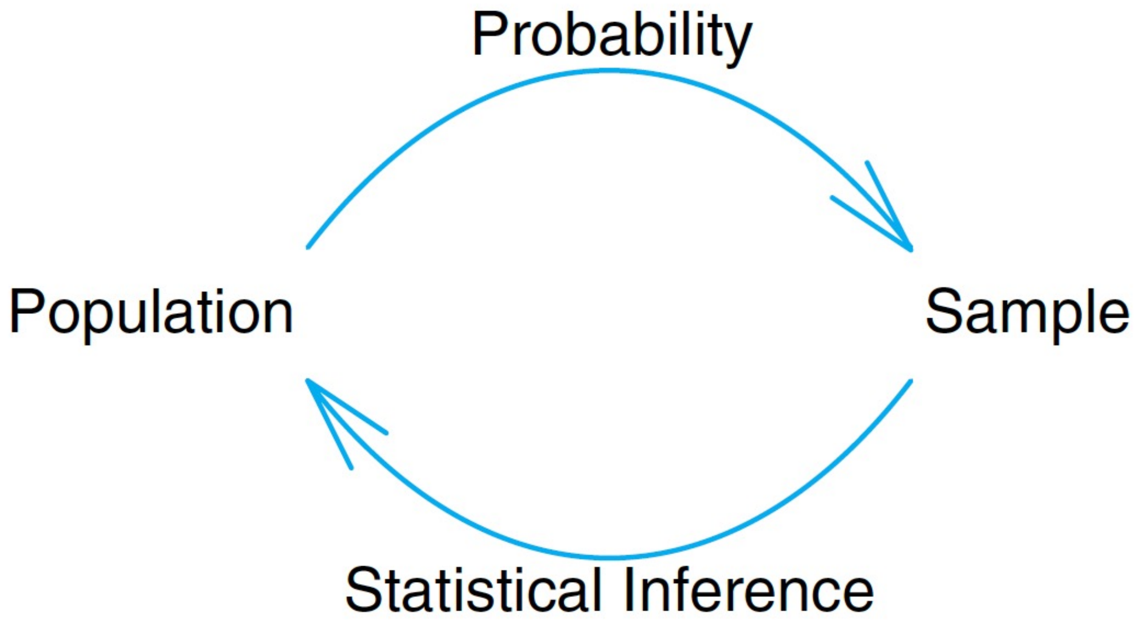


How do probability and statistics work together?

The sample + inferential statistics allows us to draw conclusions about the population

Probability allows us to draw conclusions about characteristics of hypothetical data taken from the population

Nothing can be learned about a population from a sample until the analyst learns the rudiments of uncertainty in that sample



Example: Nitrogen vs. No Nitrogen

Table 1.1: Data Set for Example 1.2

No Nitrogen	Nitrogen
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

<u>Purpose:</u> To determine if nitrogen has effect on stimulating root growth

- Two separate populations
- What conclusions do you draw?
- How can we summarize the data?

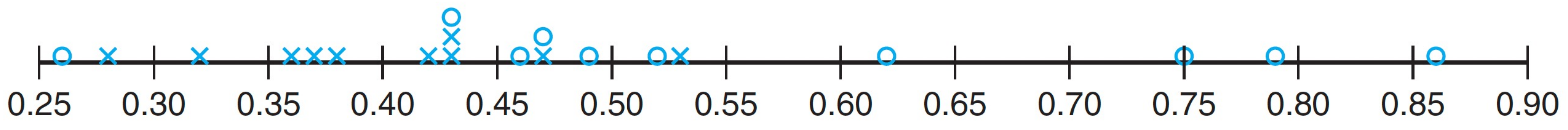


Figure 1.1: A dot plot of stem weight data.

<u>Can make a probability statement</u>

The probability that these data would be observed if there were no effect (e.g. P-value)

Outline

- Overview
- Sampling Procedures
- Measures of Location & Variability
- Graphical Diagnostics

Simple Random Sampling (SRS)

SRS implies that any sample of a specified sample size has the same chance of being selected as any other sample of the same size.

- **Sample size** – Number of elements in the sample
- E.g. we want to collect a sample of political leanings for a state
  - Sample size is 1,000
  - What if all 1,000 are in urban areas
  - Is this a representative sample?
  - Is it a biased sample?
  - Can we use it to draw inferences about the state?

Stratified Random Sampling

- Sampling group can often be divided into nonoverlapping groups that are *homogeneous*
- Homogeneous groups referred to as *strata*
- Perform simple random sampling within each strata
- Ensure no strata is over- or under-represented
- Eg. Sample 500 people from urban areas and 500 people from rural areas

Experimental Design

- Populations defined by a set of **treatments**
- E.g. nitrogen vs. no nitrogen populations
- Often considerable variability within and between groups due to the **experimental unit**
- Standard approach is to assign experimental units to the treatment conditions randomly
- E.g. assign 20 seedlings at random to treatment (nitrogen) group

Example: Corrosion Resistance

Treatment applies coating to surface. Also consider two humidity levels.

- 8 experimental units
- Each assigned randomly to 4 treatment combinations
- Cycles to failure → higher is more corrosion resitant

Table 1.2: Data for Example 1.3

Coating	Humidity	Average Corrosion in Thousands of Cycles to Failure
Uncoated	20%	975
	80%	350
Chemical Corrosion	20%	1750
	80%	1550

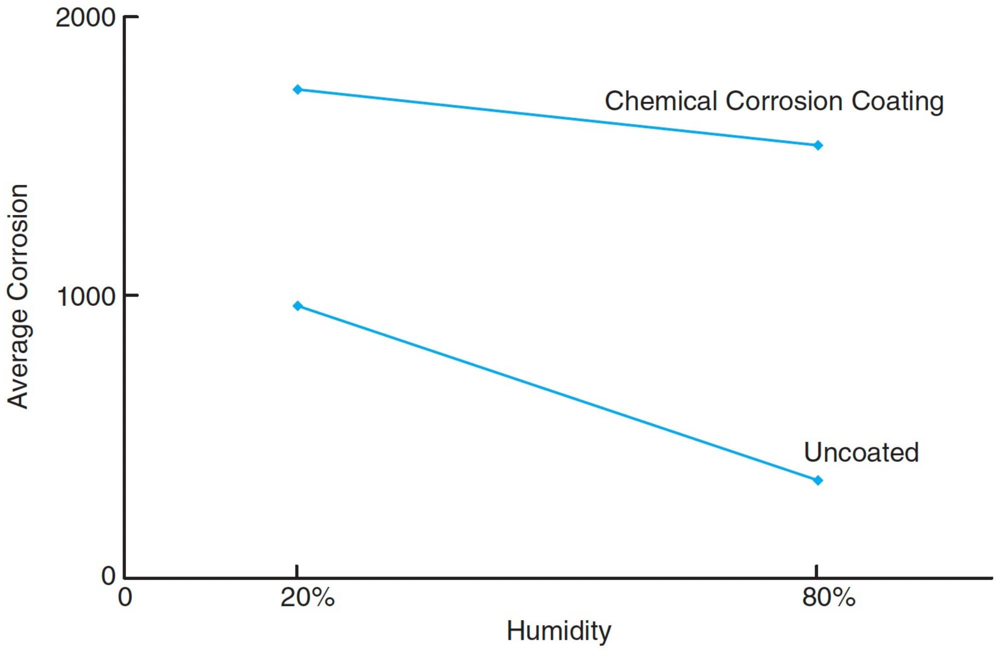


Figure 1.3: Corrosion results for Example 1.3.

Outline

- Overview
- Sampling Procedures
- Measures of Location & Variability
- Graphical Diagnostics

Measures of Location

Sample mean:

Suppose that the observations in a sample are $x_1, x_2, \ldots, x_n$. The **sample mean**, denoted by $\bar{x}$, is

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Measures of Location

Sample median:

Given that the observations in a sample are $x_1, x_2, \ldots, x_n$, arranged in **increasing order** of magnitude, the sample median is

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

Measures of Location

Suppose the data set is the following: 1.7, 2.2, 3.9, 3.11, and 14.7. The sample mean and median are, respectively,

$$\bar{x} = 5.12, \quad \tilde{x} = 3.9.$$

What properties do you observe between these statistics?

Other Measures of Location

Trimmed Mean – Compute mean after “trimming away” largest and smallest set of values,

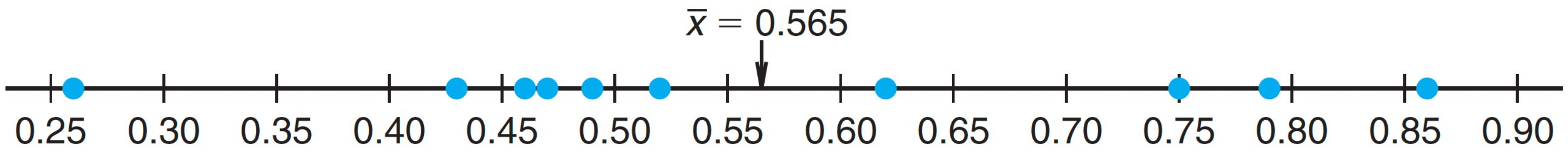


Figure 1.4: Sample mean as a centroid of the with-nitrogen stem weight.

$$\bar{x}_{\text{tr}(10)} = \frac{0.43 + 0.47 + 0.49 + 0.52 + 0.75 + 0.79 + 0.62 + 0.46}{8} = 0.56625.$$

Less sensitive to *outliers* than the sample mean, but more sensitive than the sample median.

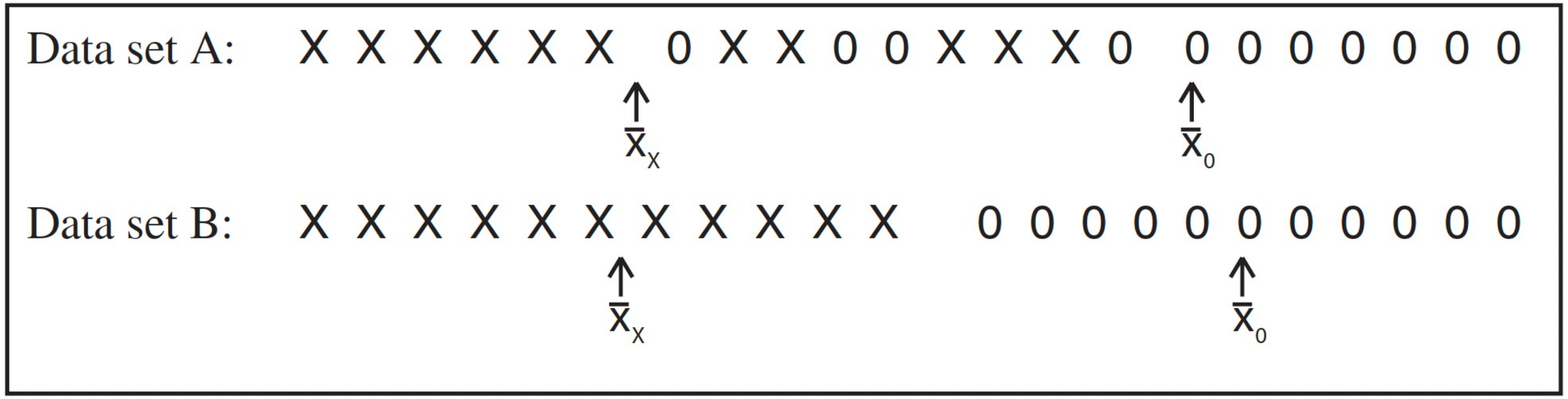
Python Example

```
import numpy as np
from scipy import stats

# Example dataset
data = np.array([1, 2, 2, 3, 4, 30, 4, 4, 5])

# Calculate the standard mean
mean_val = np.mean(data)
print(f"Standard Mean: {mean_val}")

# Calculate the median
median_val = np.median(data)
print(f"Median: {median_val}")

# Calculate the 20% trimmed mean (proportiontocut=0.2)
# This removes the lowest 20% and highest 20% of values
trimmed_mean_val = stats.trim_mean(data, 0.2)
print(f"20% Trimmed Mean: {trimmed_mean_val}")
```

```
Standard Mean: 6.111111111111111
Median: 4.0
20% Trimmed Mean: 3.4285714285714284
```

Measures of Variability

Compare / contrast samples from the following two datasets,



Dataset A exhibits large variability *within* the two groups.

Measures of Variability

Sample range: $X_{max} - X_{min}$

Sample variance / standard deviation:

The **sample variance**, denoted by s^2, is given by

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

← Degrees of Freedom

The **sample standard deviation**, denoted by s, is the positive square root of s^2, that is,

$$s = \sqrt{s^2}.$$

Python Example: Variability

```python
# Example dataset
data = np.array([1, 2, 2, 3, 4, 30, 4, 4, 5])

# Calculate variance & standard deviation
var = np.var(data)
std = np.std(data)

# Calculate the range
data_range = max(data) - min(data)

print(f"Variance: {var}")
print(f"STDEV: {std}")
print(f"The minimum value is: {min(data)}")
print(f"The maximum value is: {max(data)}")
print(f"The statistical range is: {data_range}")
```

```
Variance: 72.76543209876543
STDEV: 8.530265652297437
The minimum value is: 1
The maximum value is: 30
The statistical range is: 29
```

Discrete and Continuous Data

E.g. a chemical engineer is interested in measuring the yield (in %) of a chemical process (continuous data).

E.g. a toxicologist is testing a new drug and the patient either responds or does not (binary data).

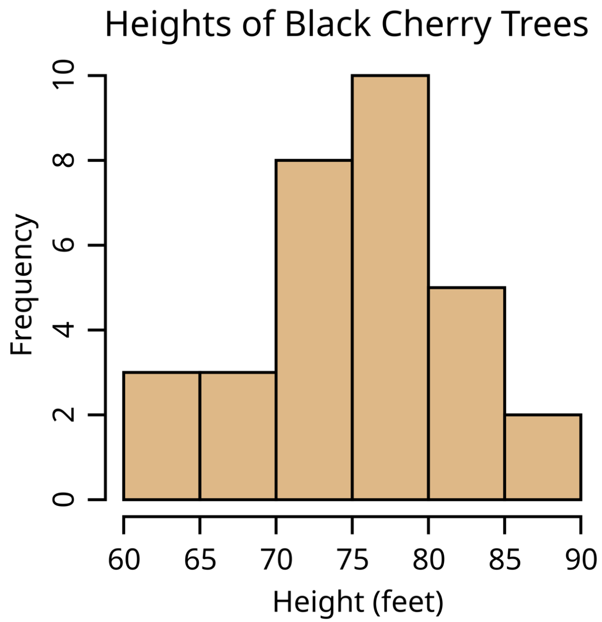
- Oftentimes binary data are reported as continuous ratios (e.g. successes / total)

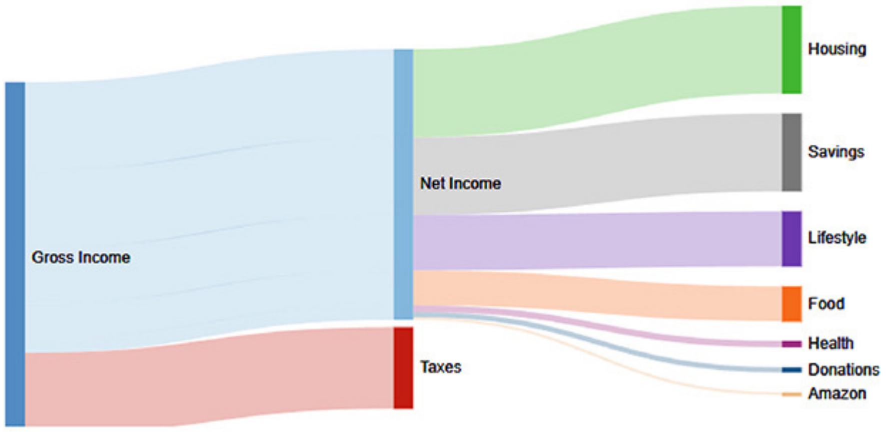
- We will see significant distinctions between continuous / discrete data when we cover probability theory

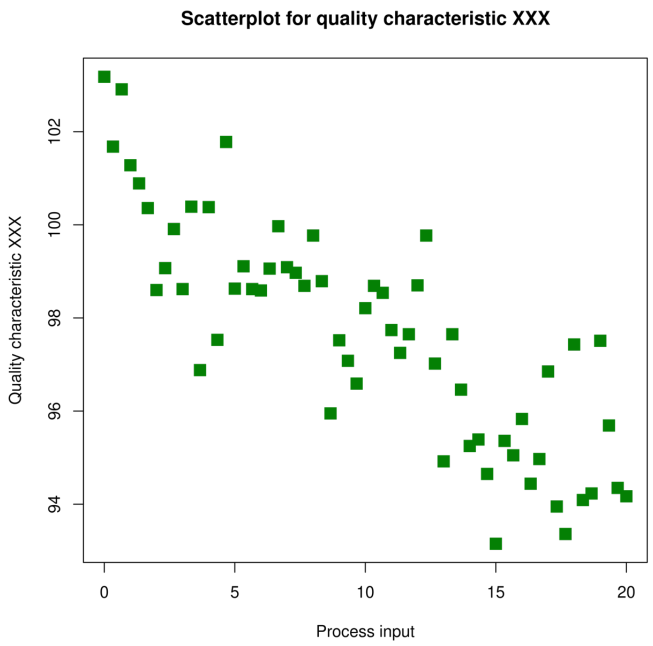
Outline

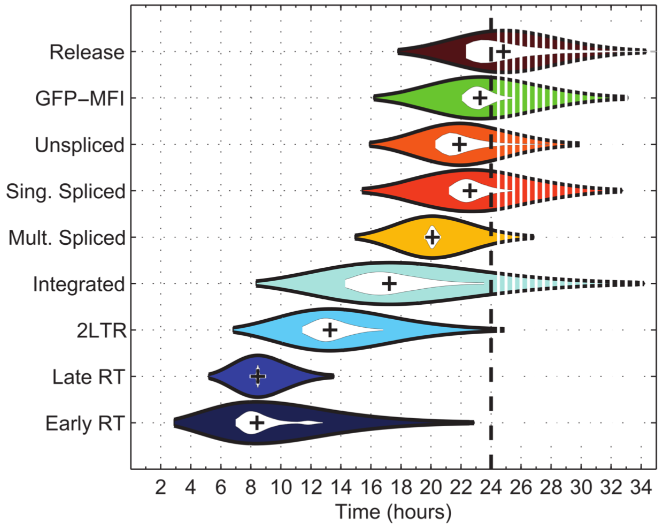
- Overview
- Sampling Procedures
- Measures of Location & Variability
- Graphical Diagnostics

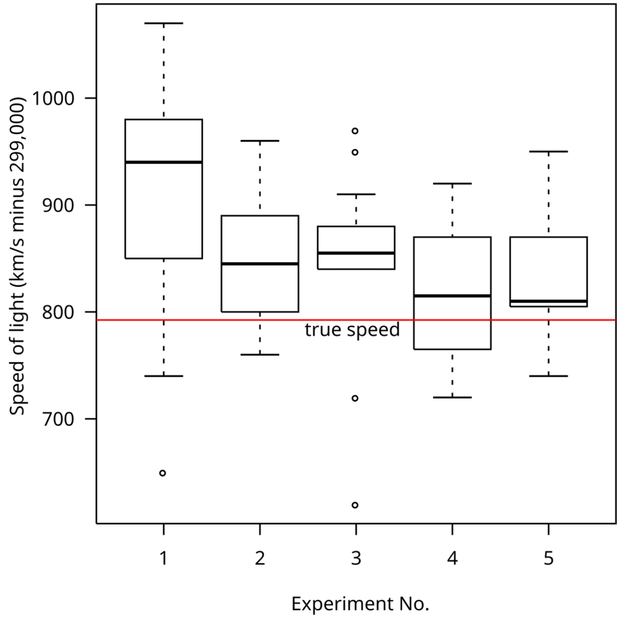
Graphs

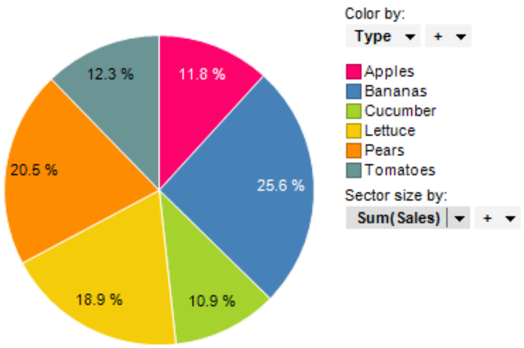

Heights of Black Cherry Trees
Frequency
0 2 4 6 8 10
60 65 70 75 80 85 90
Height (feet)


Gross Income
Net Income
Taxes
Housing
Savings
Lifestyle
Food
Health
Donations
Amazon


Scatterplot for quality characteristic XXX
Quality characteristic XXX
94 96 98 100 102
0 5 10 15 20
Process input


Release
GFP–MFI
Unspliced
Sing. Spliced
Mult. Spliced
Integrated
2LTR
Late RT
Early RT
2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34
Time (hours)


Speed of light (km/s minus 299,000)
700 800 900 1000
true speed
1 2 3 4 5
Experiment No.


12.3 %
11.8 %
25.6 %
10.9 %
18.9 %
20.5 %
Color by:
Type
Apples
Bananas
Cucumber
Lettuce
Pears
Tomatoes
Sector size by:
Sum(Sales)

Graphical Diagnostics: Scatterplot

Visual diagnostics can be helpful in identifying differences between groups.

Table 1.3: Tensile Strength

Cotton Percentage	Tensile Strength
15	7, 7, 9, 8, 10
20	19, 20, 21, 20, 22
25	21, 21, 17, 19, 20
30	8, 7, 8, 9, 10

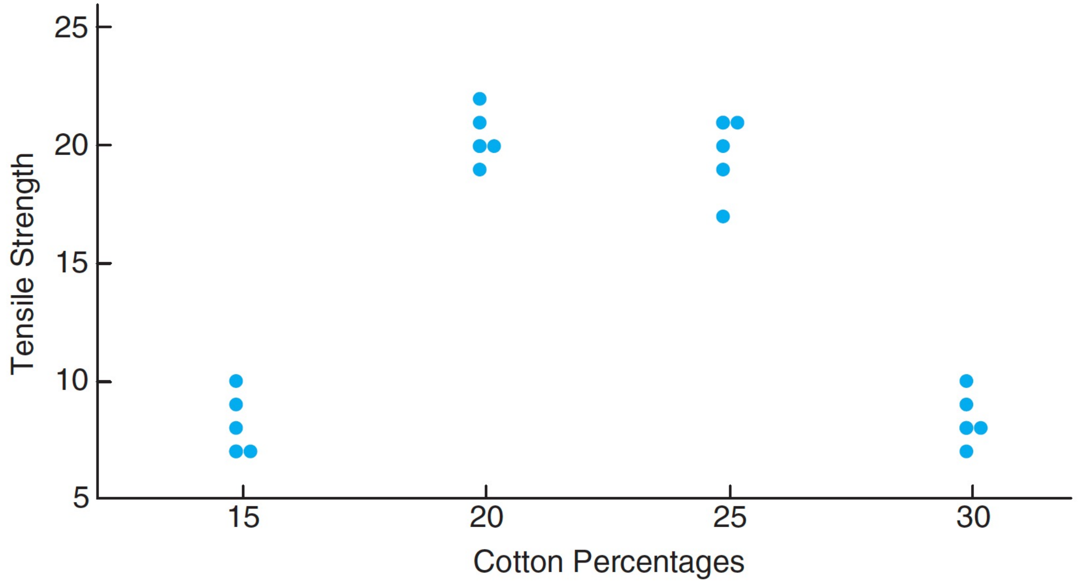


Figure 1.5: Scatter plot of tensile strength and cotton percentages.

Graphical Diagnostics: Histogram

Visual representation of the distribution of values.

Table 1.7: Relative Frequency Distribution of Battery Life

Class Interval	Class Midpoint	Frequency, f	Relative Frequency
1.5–1.9	1.7	2	0.050
2.0–2.4	2.2	1	0.025
2.5–2.9	2.7	4	0.100
3.0–3.4	3.2	15	0.375
3.5–3.9	3.7	10	0.250
4.0–4.4	4.2	5	0.125
4.5–4.9	4.7	3	0.075

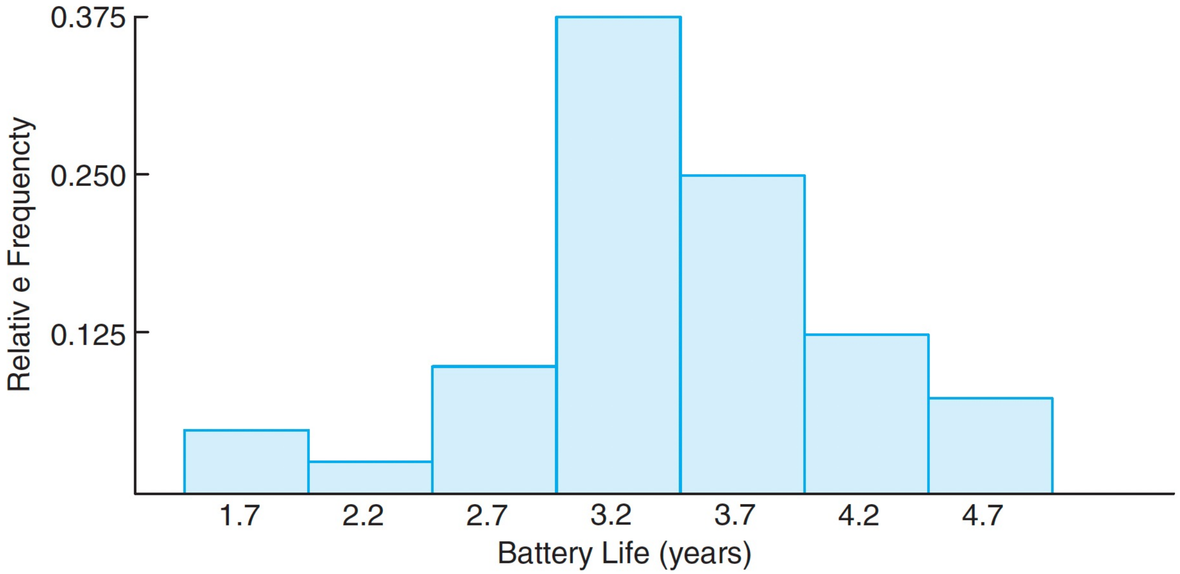


Figure 1.6: Relative frequency histogram.

Graphical Diagnostics: Box Plot

Whiskers indicate quartiles, dots indicate outliers. Note: Outlier determination is implementation-specific.

Table 1.8: Nicotine Data for Example 1.5

1.09	1.92	2.31	1.79	2.28	1.74	1.47	1.97
0.85	1.24	1.58	2.03	1.70	2.17	2.55	2.11
1.86	1.90	1.68	1.51	1.64	0.72	1.69	1.85
1.82	1.79	2.46	1.88	2.08	1.67	1.37	1.93
1.40	1.64	2.09	1.75	1.63	2.37	1.75	1.69

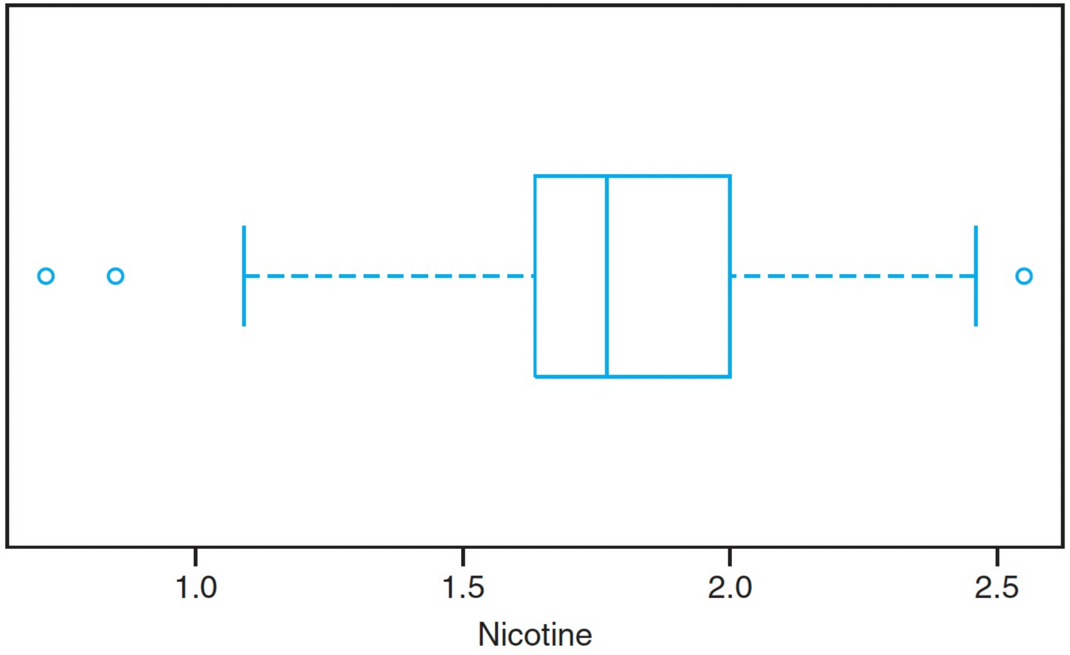
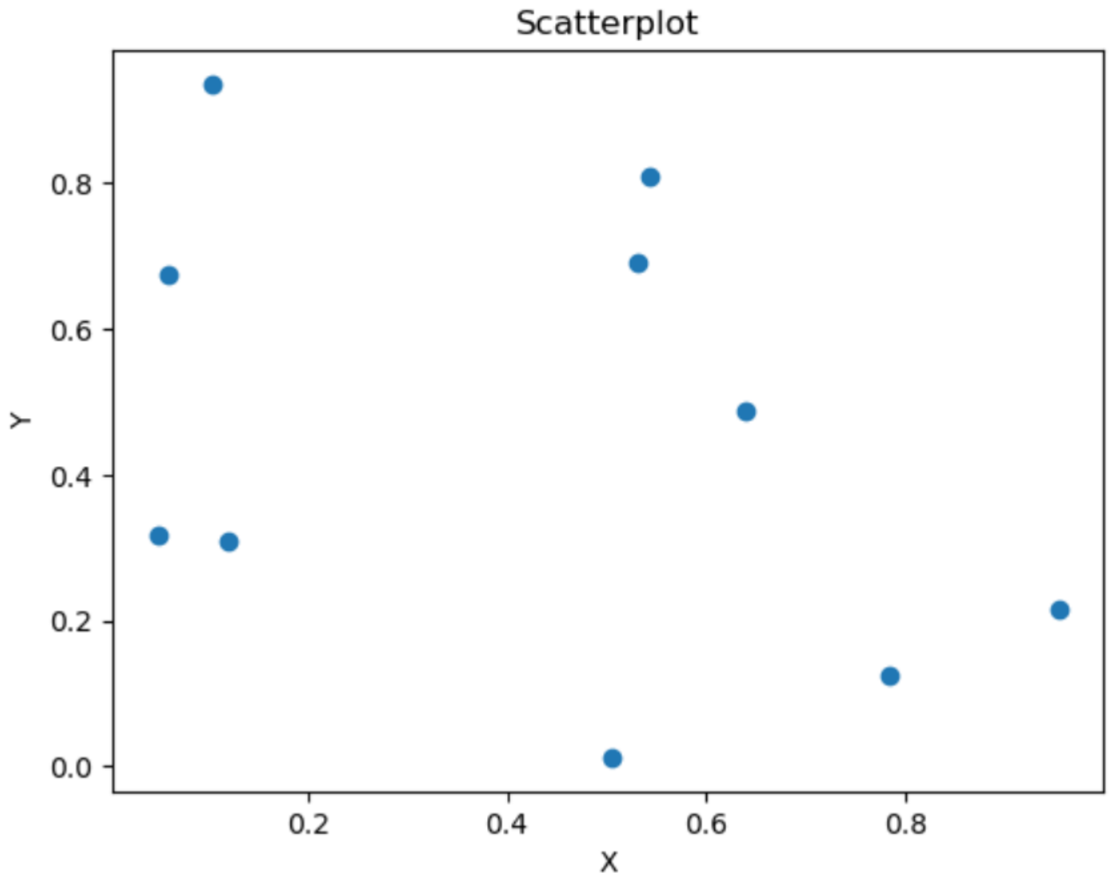


Figure 1.9: Box-and-whisker plot for Example 1.5.

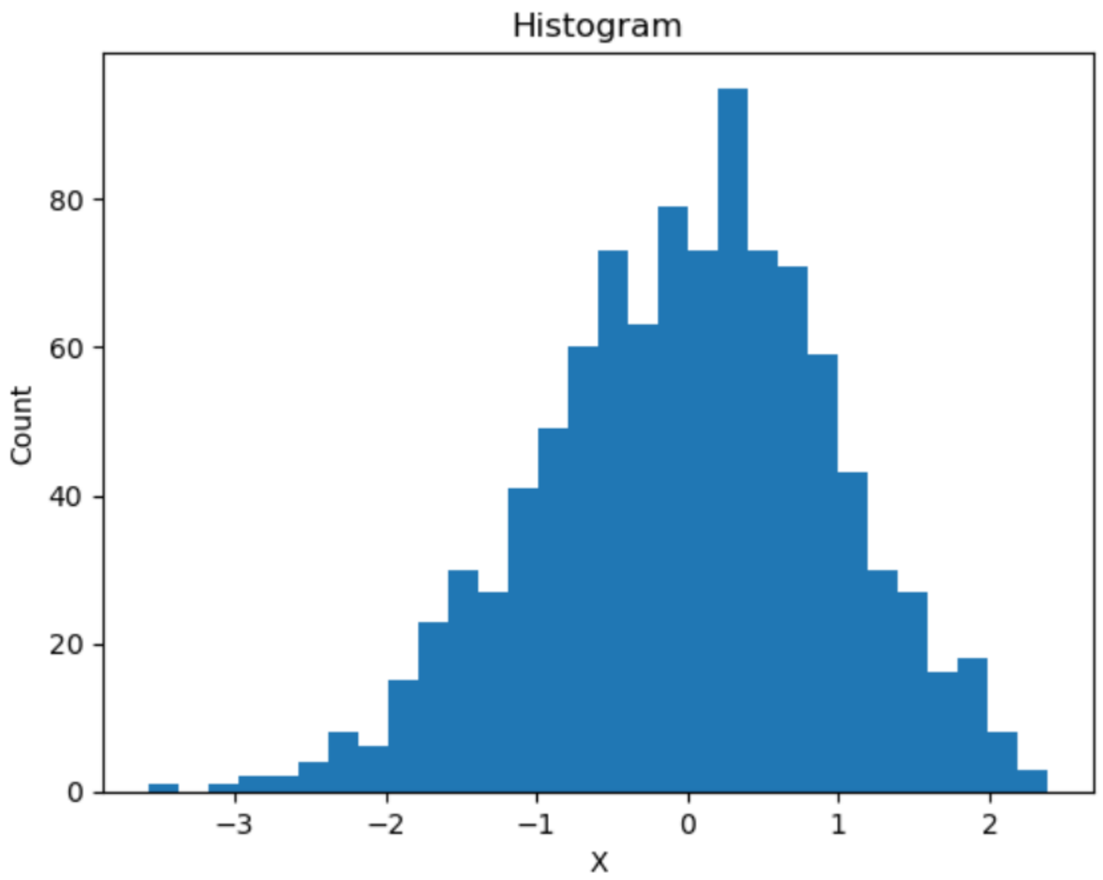
Python Example: Scatterplot

```
import matplotlib.pyplot as plt
import numpy as np

# generate random X / Y coordinates
x = np.random.rand(10)
y = np.random.rand(10)

# scatterplot
plt.scatter(x, y)
plt.xlabel("X")
plt.ylabel("Y")
plt.title("Scatterplot")
plt.show()
```



Python Example: Histogram

```
# Sample from the standard Normal distribution
s = np.random.normal(size=1000)

# Plot histogram
count, bins, ignored = plt.hist(s, 30, density=False)
plt.xlabel("X")
plt.ylabel("Count")
plt.title("Histogram")
plt.show()
```



Python Example: Boxplot

```
# Sample from the standard Normal distribution
s1 = np.random.normal(loc=0, size=1000)
s2 = np.random.normal(loc=10, size=1000)
s3 = np.random.normal(loc=5, size=1000)
s = np.array((s1, s2, s3))

# Boxplot
plt.boxplot(s.T)
plt.xlabel("Group")
plt.ylabel("Value")
plt.title("Boxplot")
plt.show()
```