Descriptive statistics:
Populations, Samples
Histograms, Quartiles
Two types of reasoning

Logical reasoning

- Physical laws
- Product designs

Types of reasoning

- Population
- Sample

Statistical reasoning: Inference of population properties from a finite sample
Numerical Summaries of Data

• Data are the **numerical observations** of a phenomenon of interest.

• The totality of all observations is a **population**.
  – Population can be infinite (e.g. random variable)
  – It can be very large (e.g. 7 billion humans)

• A (usually small) portion of the population collected for analysis is a random **sample**.

• We want to **use sample** to **infer facts** about populations

• The **inference** is not perfect but **gets better and better** as sample size increases.
Some Definitions

• The random variables $X_1, X_2, \ldots, X_n$ are a random sample of size $n$ if:
  a) The $X_i$ are independent random variables.
  b) Every $X_i$ has the same probability distribution.

• Such $X_1, X_2, \ldots, X_n$ are also called independent and identically distributed (or i. i. d.) random variables.
Ways to describe a sample:
Histogram
load PINT_binding_energy;
dfittool(binding_energy)
Histograms with Unequal Bin Widths

• If the data is tightly clustered in some regions and scattered in others, it is visually helpful to use narrow bin widths in the clustered region and wide bin widths in the scattered areas.

• To approximate the PDF, the rectangle area, not the height, must be proportional to the bin relative frequency.

\[
\text{Rectangle height} = \frac{\text{bin relative frequency}}{\text{bin width}}
\]
Median, Quartiles, Percentiles

• The median \( q_2 \) divides the sample into two equal parts: 50% \((n/2)\) of sample points below \( q_2 \) and 50% \((n/2)\) points above \( q_2 \).

• The three quartiles partition the data into four equally sized counts or segments.
  – 25% of the data is less than \( q_1 \).
  – 50% of the data is less than \( q_2 \), the median.
  – 75% of the data is less than \( q_3 \).

• There are 100 percentiles. \( n\)-th percentile \( p_n \) is defined so that \( n\% \) of the data is less than \( p_n \).
Matlab exercise

• Find the median and lower & upper quartiles of n=100 sample drawn from a continuous uniform distribution in [0,1]

• Do not use built-in Matlab functions for this exercise!

• Hint: use [a,b]=sort(r1); to rank order your sample. The variable a returns r1 sorted in the order of increasing values.

• How to find quartiles from a?
How to find median & quartiles

• % Example: find median and lower quartile of
• % a sample with n=100 drawn from uniform
• r1=rand(100,1);
• [a,b]=sort(r1);
• med=(a(50)+a(51))./2
• sum(r1<med) % verify
• q1=(a(25)+a(26))./2
• sum(r1<q1) % verify
Box-and-Whisker Plot

- A box plot is a graphical display showing Spread, Outliers, Center, and Shape (SOCS).
- It displays the 5-number summary: min, $q_1$, median, $q_3$, and max.

**Figure 6-13** Description of a box plot.
Matlab exercise:

• Generate a sample with $n = 1000$ following standard normal distribution
• Calculate median, first, and third quartiles
• Calculate IQR and find ranges shown below
• Find and count left and right outliers
• Do not use built-in Matlab functions for this!
• Make box and whisker plot: use boxplot
How many right outliers one expects in a sample of \( n=1000 \) following normal distribution?

- % find the third quartile of a standard distribution
  \[ \text{norminv}(0.75) \quad \% \text{ans} = 0.6745 \]
- % Calculate IQR - Inter Quartale Range
  \[ \text{IQR}=2.*\text{norminv}(0.75) \quad \% 1.3490 \]
- % Calculate \( 0.5\times\text{IQR}+1.5\times\text{IQR} \) - the right whisker position
  \[ \text{whisker}=0.5.*\text{IQR}+1.5.*\text{IQR} \quad \% \text{ans} = 2.6980 \]
- % Find the probability to be above the right whisker
  \[ 1-\text{normcdf(whisker)} \quad \% \text{ans} = 0.00349 \]
- % Find number of right outliers in a sample of 1000 points
  \[ 1000.\times(1-\text{normcdf(whisker)}) \quad \% \text{ans} = 3.49 \]
Descriptive statistics:
Sample mean and variance
Linear Functions of Random Variables

• A function of random variables is itself a random variable.

• A function of random variables can be formed by either linear or nonlinear relationships. We start with linear functions.

• Given random variables $X_1, X_2, ..., X_p$ and constants $c_1, c_2, ..., c_p$

\[ Y = c_1X_1 + c_2X_2 + ... + c_pX_p \]  \hspace{1cm} (5-24)

is a linear combination of $X_1, X_2, ..., X_p$. 

Mean & Variance of a Linear Function

\[ Y = c_1X_1 + c_2X_2 + \ldots + c_pX_p \]

\[
E(Y) = c_1E(X_1) + c_2E(X_2) + \ldots + c_pE(X_p) \quad (5-25)
\]

\[
V(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \ldots + c_p^2V(X_p) + 2 \sum_{i<j} c_ic_j \text{cov}(X_iX_j) \quad (5-26)
\]

If \( X_1, X_2, \ldots, X_p \) are independent, then \( \text{cov}(X_iX_j) = 0 \),

\[
V(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \ldots + c_p^2V(X_p) \quad (5-27)
\]
Example 5-31: Error Propagation

A semiconductor product consists of three layers. The variances of the thickness of each layer is 25, 40 and 30 nm\(^2\). What is the variance of the finished product?

Answer:

**IMPORTANT**

\[ X = X_1 + X_2 + X_3 \]

\[ V(X) = \sum_{i=1}^{3} V(X_i) = 25 + 40 + 30 = 95 \text{ nm}^2 \]

\[ SD(X) = \sqrt{95} = 9.7 \text{ nm} \]

If adding SDs one would get \( \sqrt{25 \text{ nm} + \sqrt{40 \text{ nm} + \sqrt{30 \text{ nm}}}} = 16.18 \text{ nm} \)
Some Definitions

• The random variables $X_1, X_2,\ldots,X_n$ are a random sample of size $n$ if:
  a) The $X_i$ are independent random variables.
  b) Every $X_i$ has the same probability distribution.

Such $X_1, X_2,\ldots,X_n$ are also called independent and identically distributed (or i. i. d.) random variables.

• A statistic is any function of the observations in a random sample.

• The probability distribution of a statistic is called a sampling distribution.
Statistic #1: Sample Mean

If the $n$ observations in a random sample are denoted by $x_1, x_2, \ldots, x_n$, the sample mean is

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$  \hspace{1cm} (6-1)
**Important:**

Sample mean $\bar{X}$ is drawn from a random variable

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

$$E(\bar{X}) = \frac{n \cdot E(X_i)}{n} = \frac{h \cdot \mu}{h} = \mu$$

$$V(\bar{X}) = \frac{n \cdot V(X_i)}{n^2} = \frac{h \cdot \sigma^2}{h^2} = \frac{\sigma^2}{h}$$

Standard dev. $(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
Central Limit Theorem

If \( X_1, X_2, \ldots, X_n \) is a random sample of size \( n \) is taken from a population (either finite or infinite) with mean \( \mu \) and finite variance \( \sigma^2 \), and if \( \bar{X} \) is the sample mean, then the limiting form of the distribution of

\[
Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}
\]

for large \( n \), is the standard normal distribution. If \( X_1, X_2, \ldots, X_n \) are themselves normally distributed - for any \( n \).
Sampling Distributions of Sample Means

**Figure 7-1** Distributions of average scores from throwing dice.
Mean = (6+1)/2 = 3.5
Sigma^2 = [(6-1+1)^2 - 1]/12 = 2.92
Sigma = 1.71

**Formulas**

\[
\mu = \frac{b + a}{2}
\]

\[
\sigma_{X}^2 = \frac{(b - a + 1)^2 - 1}{12}
\]

\[
\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}
\]

show Matlab
Example 7-1: Resistors

An electronics company manufactures resistors having a mean resistance of 100 ohms and a standard deviation of 10 ohms. What is the approximate probability that a random sample of $n = 25$ resistors will have an average resistance of less than 95 ohms?
Example 7-1: Resistors

An electronics company manufactures resistors having a mean resistance of 100 ohms and a standard deviation of 10 ohms. What is the approximate probability that a random sample of \( n = 25 \) resistors will have an average resistance of less than 95 ohms?

\[
\mu = 100 \text{ ohms}, \quad \sigma = 10 \text{ ohms}, \quad n = 25
\]

\[
\mu_{\bar{x}} = \mu \cdot \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = \frac{10}{5} = 2 \text{ ohms}
\]

\[
Z_{\bar{x}} = \frac{95 - \mu_{\bar{x}}}{\frac{\sigma}{\sqrt{n}}} = \frac{95 - 100}{2} = -2.5
\]

\[
\Pr(\bar{X} < 95) = \Phi(Z_{\bar{x}}) = \Phi(-2.5) = 0.0062
\]
Example 7-1: Resistors

An electronics company manufactures resistors having a mean resistance of 100 ohms and a standard deviation of 10 ohms. What is the approximate probability that a random sample of \( n = 25 \) resistors will have an average resistance of less than 95 ohms?

Answer:

\[
\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2.0
\]

\[
\Phi\left(\frac{\bar{X} - \mu}{\sigma_{\bar{X}}}\right) = \Phi\left(\frac{95 - 100}{2}\right) = \Phi(-2.5) = 0.0062
\]
Two Populations

We have two independent populations. What is the distribution of the difference of their sample means?

The sampling distribution of $\bar{X}_1 - \bar{X}_2$ has the following mean and variance:

$$
\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2
$$

$$
\sigma_{\bar{X}_1 - \bar{X}_2}^2 = \sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}
$$
Sampling Distribution of a Difference in Sample Means

- If we have two independent populations with means $\mu_1$ and $\mu_2$, and variances $\sigma_1^2$ and $\sigma_2^2$,
- And if $X_{-bar_1}$ and $X_{-bar_2}$ are the sample means of two independent random samples of sizes $n_1$ and $n_2$ from these populations:
- Then the sampling distribution of:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$  \hspace{1cm} (7-4)$$

is approximately standard normal, if the conditions of the central limit theorem apply.
- If the two populations are normal, then the sampling distribution is exactly standard normal.
Example 7-3: Aircraft Engine Life

The effective life of a component used in jet-turbine aircraft engines is a random variable with $\mu_{\text{old}}=5000$ hours and $\sigma_{\text{old}}=40$ hours (old). The engine manufacturer introduces an improvement into the manufacturing process for this component that changes the parameters to $\mu_{\text{new}}=5050$ hours and $\sigma_{\text{new}}=30$ hours (new).

Random samples of 16 components manufactured using “old” process and 25 components using “new” process are chosen.

What is the probability new sample mean is at least 25 hours longer than old?
Example 7-3: Aircraft Engine Life

The effective life of a component used in jet-turbine aircraft engines is a random variable with $\mu_{\text{old}}=5000$ hours and $\sigma_{\text{old}}=40$ hours (old). The engine manufacturer introduces an improvement into the manufacturing process for this component that changes the parameters to $\mu_{\text{new}}=5050$ hours and $\sigma_{\text{new}}=30$ hours (new).

Random samples of 16 components manufactured using “old” process and 25 components using “new” process are chosen.

What is the probability new sample mean is at least 25 hours longer than old?

$$\bar{x}_{\text{old}} = \frac{6010}{\sqrt{16}} = 10 \text{ hrs}$$

$$\bar{x}_{\text{new}} = \frac{6025}{\sqrt{25}} = 7 \text{ hrs}$$

$$\bar{x}_{\text{total}} = \sqrt{\bar{x}_{\text{old}}^2 + \bar{x}_{\text{new}}^2} = \sqrt{100 + 36} \approx 11.7 \text{ hrs}$$

$$\mu_{\text{new}} - \mu_{\text{old}} = 50 \text{ hrs}$$

$$z = \frac{25 - (50)}{11.7} = -2.14$$

$$\text{Prob}(z \geq -2.14) = 0.9840$$
Example 7-3: Aircraft Engine Life

The effective life of a component used in jet-turbine aircraft engines is a normal-distributed random variable with parameters shown (old). The engine manufacturer introduces an improvement into the manufacturing process for this component that changes the parameters mu and sigma as shown (new). Random samples are selected from the “old” process and “new” process as shown.

What is the probability new sample mean is at least 25 hours longer than old?

<table>
<thead>
<tr>
<th>Process</th>
<th>Old (1)</th>
<th>New (2)</th>
<th>Diff (2-1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mu</td>
<td>5,000</td>
<td>5,050</td>
<td>50</td>
</tr>
<tr>
<td>sigma</td>
<td>40</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td>n</td>
<td>16</td>
<td>25</td>
<td></td>
</tr>
</tbody>
</table>

Calculations

\[
\frac{s}{\sqrt{n}} = 10\quad 6\quad 11.7
\]

\[
z = -2.14
\]

\[
P(x_{\text{bar}_2} - x_{\text{bar}_1} > 25) = P(Z > z) = 0.9840
\]
**WHY ARE THERE SLAVES IN THE BIBLE?**

1. Why is there no king in England any longer?
2. Why do I feel dizzy?
3. Why are dogs afraid of fireworks?
4. Why aren’t my arms growing?
5. Why is there no King in America?

---

**WHY DO WHALES JUMP OUT OF THE BIBLE?**

- Why do whales jump out of the Bible?
- Why do whales jump out of the Bible?
- Why do whales jump out of the Bible?
- Why do whales jump out of the Bible?

---

**WHY ARE THERAPISTS BORED?**

- Why are therapists bored?
- Why are therapists bored?
- Why are therapists bored?
- Why are therapists bored?

---

**WHY ARE THERE EMOTIONS?**

- Why are there emotions?
- Why are there emotions?
- Why are there emotions?
- Why are there emotions?

---

**WHY ARE THERE PENGUINS IN RHODESIA?**

- Why are there penguins in Rhodesia?
- Why are there penguins in Rhodesia?
- Why are there penguins in Rhodesia?
- Why are there penguins in Rhodesia?

---

**WHY ARE THERE BUILDINGS INSIDE MILK?**

- Why are there buildings inside milk?
- Why are there buildings inside milk?
- Why are there buildings inside milk?
- Why are there buildings inside milk?

---

**WHY DO I SAY UH?**

- Why do I say uh?
- Why do I say uh?
- Why do I say uh?
- Why do I say uh?

---

**WHY ARE THERAPISTS BORED?**

- Why are therapists bored?
- Why are therapists bored?
- Why are therapists bored?
- Why are therapists bored?

---

**WHY ARE THERAPISTS BORED?**

- Why are therapists bored?
- Why are therapists bored?
- Why are therapists bored?
- Why are therapists bored?