Introduction to Engineering I
Lecture 9

Some Statistical Ideas
  Population, Sample
  Measures of Central Tendency and Variation
  Curve Fitting by Linear Regression

The Normal Distribution
  Calculating Probabilities

Studio Problems
Presentations @ 4:30
Populations and Samples

• Population: ALL members of a specific group, or ALL events of a particular description
  • Example: ALL the students in SCUPI
  • Example: ALL the values of the box volume

• A measurable property of a population is a parameter
  • Example: The average height of all SCUPI students

• Parameters are often difficult or costly to obtain when the population contains many elements
Populations and Samples

• Sample: A SUBSET of the population
  • Example: Ten students in the class selected at random
  • Example: Every third team measurement of the box volume

• A measurable property of a sample is a statistic
  • Example: The average height of the ten selected students

• We use the statistics (which characterize the sample) to approximate the parameters (which characterize the population)

• Random sampling is a way of choosing a sample from a population
# Box Volume Measurements* (cm³)

<table>
<thead>
<tr>
<th></th>
<th>Unit</th>
<th>Value</th>
<th></th>
<th>Unit</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MR</td>
<td>250</td>
<td>18</td>
<td>Lux</td>
<td>236.25</td>
</tr>
<tr>
<td>2</td>
<td>MR tape</td>
<td>220</td>
<td>19</td>
<td>Dalton</td>
<td>240.63</td>
</tr>
<tr>
<td>3</td>
<td>KQ</td>
<td>250</td>
<td>20</td>
<td>Barn</td>
<td>238.14</td>
</tr>
<tr>
<td>4</td>
<td>KQ tape</td>
<td>230</td>
<td>21</td>
<td>Hertz</td>
<td>237.852</td>
</tr>
<tr>
<td>5</td>
<td>Mole</td>
<td>227.052</td>
<td>22</td>
<td>Byte</td>
<td>231.34</td>
</tr>
<tr>
<td>6</td>
<td>Volt</td>
<td>231.02</td>
<td>23</td>
<td>Diopter</td>
<td>242.55</td>
</tr>
<tr>
<td>7</td>
<td>Watt</td>
<td>238.4375</td>
<td>24</td>
<td>Kilogram</td>
<td>250.25</td>
</tr>
<tr>
<td>8</td>
<td>Octave</td>
<td>260</td>
<td>25</td>
<td>Farad</td>
<td>230.47</td>
</tr>
<tr>
<td>9</td>
<td>Siemens</td>
<td>240.9825</td>
<td>26</td>
<td>Becquerel</td>
<td>235.03125</td>
</tr>
<tr>
<td>10</td>
<td>Tesla</td>
<td>236.25</td>
<td>27</td>
<td>Lumen</td>
<td>240.41</td>
</tr>
<tr>
<td>11</td>
<td>Sievert</td>
<td>238.14</td>
<td>28</td>
<td>Gray</td>
<td>244.82</td>
</tr>
<tr>
<td>12</td>
<td>Pascal</td>
<td>242.25</td>
<td>29</td>
<td>Kelvin</td>
<td>242.55</td>
</tr>
<tr>
<td>13</td>
<td>Meter</td>
<td>235.62</td>
<td>30</td>
<td>Coulomb</td>
<td>233.541</td>
</tr>
<tr>
<td>14</td>
<td>Semitone</td>
<td>240.625</td>
<td>31</td>
<td>Hectare</td>
<td>234.74</td>
</tr>
<tr>
<td>15</td>
<td>Radians</td>
<td>264.49</td>
<td>32</td>
<td>Candela</td>
<td>242.55</td>
</tr>
<tr>
<td>16</td>
<td>Parsec</td>
<td>240.625</td>
<td>33</td>
<td>Ampere</td>
<td>256.032</td>
</tr>
<tr>
<td>17</td>
<td>Ohm</td>
<td>236.25</td>
<td>34</td>
<td>Henry</td>
<td>253.44</td>
</tr>
</tbody>
</table>

*Data from 2015 class. Some values adjusted for educational purposes
Histogram of Volume
## Sample Every 4\textsuperscript{th} Measurement

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MR</td>
<td>250</td>
<td>18</td>
<td>Lux</td>
<td>236.25</td>
</tr>
<tr>
<td>2</td>
<td>MR tape</td>
<td>220</td>
<td>19</td>
<td>Dalton</td>
<td>240.63</td>
</tr>
<tr>
<td>3</td>
<td>KQ</td>
<td>250</td>
<td>20</td>
<td>Barn</td>
<td>238.14</td>
</tr>
<tr>
<td>4</td>
<td>KQ tape</td>
<td>230</td>
<td>21</td>
<td>Hertz</td>
<td>237.852</td>
</tr>
<tr>
<td>5</td>
<td>Mole</td>
<td>227.052</td>
<td>22</td>
<td>Byte</td>
<td>231.34</td>
</tr>
<tr>
<td>6</td>
<td>Volt</td>
<td>231.02</td>
<td>23</td>
<td>Diopter</td>
<td>242.55</td>
</tr>
<tr>
<td>7</td>
<td>Watt</td>
<td>238.4375</td>
<td>24</td>
<td>Kilogram</td>
<td>250.25</td>
</tr>
<tr>
<td>8</td>
<td>Octave</td>
<td>260</td>
<td>25</td>
<td>Farad</td>
<td>230.47</td>
</tr>
<tr>
<td>9</td>
<td>Siemens</td>
<td>240.9825</td>
<td>26</td>
<td>Becquerel</td>
<td>235.03125</td>
</tr>
<tr>
<td>10</td>
<td>Tesla</td>
<td>236.25</td>
<td>27</td>
<td>Lumen</td>
<td>240.41</td>
</tr>
<tr>
<td>11</td>
<td>Sievert</td>
<td>238.14</td>
<td>28</td>
<td>Gray</td>
<td>244.82</td>
</tr>
<tr>
<td>12</td>
<td>Pascal</td>
<td>242.25</td>
<td>29</td>
<td>Kelvin</td>
<td>242.55</td>
</tr>
<tr>
<td>13</td>
<td>Meter</td>
<td>235.62</td>
<td>30</td>
<td>Coulomb</td>
<td>233.541</td>
</tr>
<tr>
<td>14</td>
<td>Semitone</td>
<td>240.625</td>
<td>31</td>
<td>Hectare</td>
<td>234.74</td>
</tr>
<tr>
<td>15</td>
<td>Radians</td>
<td>264.49</td>
<td>32</td>
<td>Candela</td>
<td>242.55</td>
</tr>
<tr>
<td>16</td>
<td>Parsec</td>
<td>240.625</td>
<td>33</td>
<td>Ampere</td>
<td>256.032</td>
</tr>
<tr>
<td>17</td>
<td>Ohm</td>
<td>236.25</td>
<td>34</td>
<td>Henry</td>
<td>253.44</td>
</tr>
</tbody>
</table>

---

**SCUPI**

Sichuan University - Pittsburgh Institute
Histogram of Volume Samples
Measures of Central Tendency

• Population Mean

\[ \mu = \frac{1}{N} \sum_{i} x_i \]

- N is the number of values in the population
- \(x_i\) is the \(i^{th}\) value
- Summation is over \(i = 1\) to \(i = N\) (entire population)

• Sample Mean

\[ \bar{x} = \frac{1}{N} \sum_{i} x_i \]

- N is the number of values in the sample
Measures of Central Tendency

• Mean
  • Also called the average
  • Most commonly used measure
  • Highly sensitive to outliers - single data points can overwhelm the mean

• Median
  • Half of the values are above and below the median
  • Less sensitive to outliers than the mean

• Mode
  • Most frequently occurring value
  • Might not be unique (can have multiple modes)
Population Mean:
\[ \mu = 240.4 \text{ cm}^3 \]
Median = 239.4 cm³

Sample Mean:
\[ \bar{x} = 242.2 \text{ cm}^3 \]
Median = 241.0 cm³
Measures of Variation

- Population Variance
  \[ \sigma^2 = \frac{1}{N} \sum_{i} (x_i - \mu)^2 \]
  where \( N \) is the number of values in the population.

- Sample Variance
  \[ s^2 = \frac{1}{N - 1} \sum_{i} (x_i - \bar{x})^2 \]
  where \( N \) is the number of values in the sample.

- Standard Deviation
  \[ \sigma = \sqrt{\sigma^2}, \quad s = \sqrt{s^2} \]
Population Standard Deviation:
\( \sigma = 9.2 \text{ cm}^3 \)

Sample Standard Deviation:
\( s = 8.3 \text{ cm}^3 \)

\( s \) is an estimate of \( \sigma \)
Linear Regression

Idea:
Find the straight line $y' = mx + b$ that best fits the data
$y_i$ is the measured value at $x_i$

$y'(x)$ is an estimate of the data found by considering all points

$(y_i - y')$ is called the residual; it can be positive or negative
Linear Regression Method

• Want to find the line $y' = mx + b$ that is "closest" (best fit) to the data

• Simple approach: minimize the sum of the residuals
  • Choose $m$ and $b$ such that $\sum (y_i - y')$ is a minimum
  • Problem: since $y_i - y'$ can be positive or negative, even a poorly fit line can have a small or zero sum of residuals

• Proper approach: minimize $\sum (y_i - y')^2$
  • Square of residuals is always positive
  • Minimizing the sum of squared residuals results in best fit line
Linear Regression Formulas

\[ m = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2} \]

\[ b = \frac{\sum x_i - m(\sum x_i)}{n} \]

Matlab will do this for you!
Normal Distribution*

\[ f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2} \]

*Also called the Gaussian Distribution or bell-shaped curve
Histogram of Box Volume

\[ \mu = 240.4 \text{ cm}^3 \]
\[ \sigma = 9.2 \text{ cm}^3 \]
Calculating Probabilities from the Normal Distribution

- Transform data into z-scale (units of standard deviation)
Calculating Probabilities from the Normal Distribution

• Probability of \( z < z_0 \) = area under the curve from \(-\infty \) to \( z_0 \)

• See examples in Chapter 11
• Need table of areas (Appendix C) – on website
Useful Matlab Functions

mean, std
median, mode, max, min
hist, histfit
plot, bar, figure
hold on, hold off
fit
corrcotef