Outline

- Introduction to Statistics and Probability
- FE Reference Handbook
- Works Cited
Statistics 1

- Population (whole group)
- Sample (part of a population)
- Important parameters
  - Arithmetic mean or average
  - Variance
  - Standard deviation
  - Coefficient of variation
  - Geometric mean
  - Median
  - Mode
  - Range
  - root-mean-square (rms)
  - Relative error
  - Linear regression
- Source for text and equations is NCEES, unless otherwise stated.
Statistics 2 (Olia 84)

- Measures of central tendency
  - Mean, Median, and Mode
- Measures of dispersion ("spread or variability")
  - Variance and Standard Deviation
Arithmetic Mean

- If $X_1, X_2, \ldots, X_n$ represent the values of a random sample of $n$ items or observations, the arithmetic mean or average of these items or observations, denoted $\overline{X}$, is given by:

$$\overline{X} = \frac{1}{n} (X_1 + X_2 + \ldots + X_n)$$

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$\overline{X} \to \mu \quad \text{for sufficiently large values of } n$$
Arithmetic Mean Example 1

• 11 sample observations
• 0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1
• Find the arithmetic mean or average of the sample observations
• Solution:
  - \( n = 11 \)
  
  \[
  \overline{X} = \frac{1}{11} (0.5+100+1000.25+345+0.0213+0+45+99+23+11+1)
  \]

  \( \overline{X} = 147.71 \approx 148 \text{ units} \)
Arithmetic Mean Example 1

• Find the arithmetic mean or average of the sample observations using R

• Solution:

```r
samp <- c(0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1)
length(samp) # number of sample observations
mean(samp)
```

```
[1] 147.7065 # units
```
Arithmetic Mean Example 2

- Refer back to Data Import (clipboard) from Lecture Notes 5b (next slides)
```r
library(psych)

# Excerpt of MillCreekcsv (rows 112 - 125) found in 03431000_MillCreek_AntiochTN_revised.csv

"agency_cd","site_no","datetime","02_00060_00003","02_00060_00003_cd"
"USGS","03431000",1954-01-20,2670,"A"
"USGS","03431000",1954-01-21,1120,"A"
"USGS","03431000",1954-01-22,1300,"A"
"USGS","03431000",1954-01-23,360,"A"
"USGS","03431000",1954-01-24,225,"A"
"USGS","03431000",1954-01-25,156,"A"
"USGS","03431000",1954-01-26,127,"A"
"USGS","03431000",1954-01-27,510,"A"
"USGS","03431000",1954-01-28,183,"A"
"USGS","03431000",1954-01-29,131,"A"
"USGS","03431000",1954-01-30,98,"A"
"USGS","03431000",1954-01-31,78,"A"
"USGS","03431000",1954-02-01,66,"A"
"USGS","03431000",1954-02-02,59,"A"
```
Data Import (clipboard) 2

- MillCreekclip <- read.clipboard.csv() # copy the comma-separated table [03431000_MillCreek_AntiochTN_revised.csv opened in an advanced text editor], including the table header [column names]
<table>
<thead>
<tr>
<th>agency_cd</th>
<th>site_no</th>
<th>datetime</th>
<th>X02_00060_00003</th>
<th>X02_00060_00003_cd</th>
</tr>
</thead>
<tbody>
<tr>
<td>USGS</td>
<td>03431000</td>
<td>1954-01-20</td>
<td>2670</td>
<td>A</td>
</tr>
<tr>
<td>USGS</td>
<td>03431000</td>
<td>1954-01-21</td>
<td>1120</td>
<td>A</td>
</tr>
<tr>
<td>USGS</td>
<td>03431000</td>
<td>1954-01-22</td>
<td>1300</td>
<td>A</td>
</tr>
<tr>
<td>USGS</td>
<td>03431000</td>
<td>1954-01-23</td>
<td>360</td>
<td>A</td>
</tr>
<tr>
<td>USGS</td>
<td>03431000</td>
<td>1954-01-24</td>
<td>225</td>
<td>A</td>
</tr>
<tr>
<td>USGS</td>
<td>03431000</td>
<td>1954-01-25</td>
<td>156</td>
<td>A</td>
</tr>
<tr>
<td>USGS</td>
<td>03431000</td>
<td>1954-01-26</td>
<td>127</td>
<td>A</td>
</tr>
<tr>
<td>USGS</td>
<td>03431000</td>
<td>1954-01-27</td>
<td>510</td>
<td>A</td>
</tr>
<tr>
<td>USGS</td>
<td>03431000</td>
<td>1954-01-28</td>
<td>183</td>
<td>A</td>
</tr>
<tr>
<td>USGS</td>
<td>03431000</td>
<td>1954-01-29</td>
<td>131</td>
<td>A</td>
</tr>
<tr>
<td>USGS</td>
<td>03431000</td>
<td>1954-01-30</td>
<td>98</td>
<td>A</td>
</tr>
<tr>
<td>USGS</td>
<td>03431000</td>
<td>1954-01-31</td>
<td>78</td>
<td>A</td>
</tr>
<tr>
<td>USGS</td>
<td>03431000</td>
<td>1954-02-01</td>
<td>66</td>
<td>A</td>
</tr>
<tr>
<td>USGS</td>
<td>03431000</td>
<td>1954-02-02</td>
<td>59</td>
<td>A</td>
</tr>
</tbody>
</table>
Arithmetic Mean Example 2

- What is the mean discharge (cfs) for the given dates (using R)? Note: The mean discharge is what is shown in the table, but for this example we will calculate the mean discharge as if it is not given.

```r
mean(MillCreekclip[, 4]) # 02_00060_00003 or column 4 with all rows

[1] 505.9286 # cfs
```
Population variance

- The variance of the population is the arithmetic mean of the squared deviations from the population mean. If \( \mu \) is the arithmetic mean of a discrete population of size \( N \), the population variance

\[
\sigma^2 = \frac{1}{N} \left[ (X_1 - \mu)^2 + (X_2 - \mu)^2 + \ldots + (X_N - \mu)^2 \right]
\]

\[
\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (X_i - \mu)^2
\]
Population Variance
Example 1

- 17 population observations
- 0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1, 89, 0, 34, 65, 98, 3
- Find the variance of the population observations
- Solution:
  - \( N = 17 \)
  \[
  \mu = \frac{1}{17} \left( 0.5 + 100 + 1000.25 + 345 + 0.0213 + 0 + 45 + 99 + 23 + 11 + 1 + 89 + 0 + 34 + 65 + 98 + 3 \right)
  \]
  \[
  \mu = 112.57 \approx 113 \text{ units}
  \]
Population Variance

Example 2

\[ \sigma^2 = \left( \frac{1}{17} \right) \left[ (0.5 - \mu)^2 + (100 - \mu)^2 + (1000.25 - \mu)^2 + (345 - \mu)^2 + (0.0213 - \mu)^2 + \ldots + (3 - \mu)^2 \right] \]

\[ \mu = 112.57 \approx 113 \ units \]

\[ \sigma^2 = 55851.46 \approx 55851 \ units^2 \]
Population Variance
Example 3

- Find the population variance using R
- Solution:

```
pop <- c(0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1, 89, 0, 34, 65, 98, 3)
length(pop)
source("varpop.R")
varpop(pop)
```

The population variance is 55851 units\(^2\).

[1] 55851.47 # units\(^2\)
Population standard deviation

\[ \sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (X_i - \mu)^2} \]
Population standard deviation Example 1

- 17 population observations
- 0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1, 89, 0, 34, 65, 98, 3
- Find the standard deviation of the population observations
- Solution:
  - \( N = 17 \)
  
  \[
  \sigma^2 = 55851.46 \approx 55852 \ units^2
  \]
  
  \[
  \sigma = 236.32 \approx 236 \ units
  \]
Population standard deviation Example 2

- Find the standard deviation of the population observations using R

- Solution:

```r
pop <- c(0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1, 89, 0, 34, 65, 98, 3)
length(pop)
source("varpop.R")
varpop(pop)
```

The population standard deviation is 236 units.

```
[1] 236.3291 # units
```
Sample variance

\[ s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \]
Sample variance Example 1

- 11 sample observations
- 0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1
- Find the variance of the sample observations
- Solution:
  - \( n = 11 \)
  \[ \bar{X} = 147.71 \approx 148 \text{ units} \]
  \[ s^2 = 90201.31 \approx 90201 \text{ units}^2 \]
Sample variance Example 2

- Find the variance of the sample observations using R

- Solution:

```r
samp <- c(0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1)

length(samp)
var(samp)
[1] 90201.31 # units^2
```
Sample standard deviation

\[ s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} X_i - \bar{X}^2} \]
Sample standard deviation

Example 1

• 11 sample observations
• 0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1
• Find the standard deviation of the sample observations

Solution:

\[ n = 11 \]

\[ s^2 = 90201.31 \approx 90201 \text{ units} \]

\[ s = 300.34 \approx 300 \text{ units} \]
Sample standard deviation
Example 2

- Find the standard deviation of the sample observations using R

- Solution:

```r
samp <- c(0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1)
length(samp)
sd(samp)
[1] 300.3353 # units
```
Sample coefficient of variation

\[ CV = \frac{s}{\bar{X}} \]

\[ s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} X_i - \bar{X}^2} \]

\[ \bar{X} = \frac{1}{n} (X_1 + X_2 + \ldots + X_n) \]

\[ \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \]
Sample coefficient of variation Example 1

- 11 sample observations
- 0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1

- Find the coefficient of variation of the sample observations

- Solution:
  - \( n = 11 \)
  - \( s = 300.34 \)
  - \( \bar{X} = 147.71 \)
  - \( CV = \frac{300.34}{147.71} \approx 2 \)
Sample coefficient of variation Example 2

- Find the sample coefficient of variation using R
- Solution:

```r
samp <- c(0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1)
length(samp)
source("cv.R")
cv(samp)
The sample coefficient of variation is 2.
[1] 2.033325
```
Sample geometric mean

\[ \text{sgm} = \sqrt[n]{X_1 X_2 X_3 X_4 X_N}, x_i > 0 \]
Sample geometric mean
Example 1

• 11 sample observations
• 0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1

• Find the geometric mean of the sample observations

• Solution:

\[
10 \sqrt[10]{0.5 \times 100 \times 1000.25 \times \ldots \times 1} = 14.5119
\]

– N = 10 (reject the value of 0)
• Find the geometric mean of the sample observations using R
• Solution:
samp <- c(0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1)
length(samp)
source("sgm.R")
sgm(samp)
The sample geometric mean is 15 units.
[1] 14.51191 # units
Compare Arithmetic & Geometric Means Example

- Plot a histogram of the sample data along with the Arithmetic & Geometric Means using R
- Solution:

```r
samp <- c(0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1)
# histogram of samp data
# point for both arithmetic and geometric means
meansamp <- mean(samp)
sgmsamp <- sgm(samp)

sampdf <- data.frame(samp) # create a data.frame out of the numeric vector
colnames(sampdf) <- "samp" # change the column name to samp

library(ggplot2)
ggplot(sampdf, aes(x = samp)) + geom_histogram() + geom_point(data = sampdf, shape = 3, size = 3, colour = "green", aes(x = meansamp, y = 1, show_guide = TRUE)) + geom_text(aes(label = "A M", x = meansamp, y = 1), vjust = 2, colour = "red") + geom_point(data = sampdf, shape = 3, size = 3, colour = "yellow", aes(x = sgmsamp, y = 1, show_guide = TRUE)) + geom_text(aes(label = "G M", x = sgmsamp, y = 1), vjust = 2, colour = "blue") + ggtitle("Comparing Arithmetic (A M) and Geometric Means (G M) For Sample Data")
```
Compare Arithmetic & Geometric Means Example
Compare Means 1 (Software)

- **Arithmetic Mean**
  - “The arithmetic mean is a good measure when numbers are of the same order of magnitude – like students scores on a test.

- **Geometric Mean**
  - Geometric mean would be appropriate if the numbers are in different ranges (ballparks) entirely and you do not want one very large number to affect things that much.
• For example, if you have following numbers: 1, 10, 100, 1000, 10000, the average is more than 2200. But a more appropriate “middle” number is 100 in this case. And 100 is the geometric mean here.

• Real world example: 0.98, 8.7, 121, 1400, 9000. From these 5 numbers, arithmetic mean is about 2100. That number hardly means anything. Geometric mean is about 105, which represents more of a central point.

• Interestingly, median is also a good measure in the previous case. But one problem with median is that the largest number (9000) was changed to 10,000 or 10 million, then still the median would be 121, and would not change at all. However, the geometric mean would change (increase) a bit. Arithmetic mean would change TOO much.

• Yet another scenario where geometric mean is more appropriate than arithmetic mean is when the numbers are given as percentage increases or decreases, rather than absolute values. For example, if the housing market rose 40% 1 year, dropped 40% next year, then an appropriate representation of average growth rate can be found by taking a geometric mean of 1.4 and 0.6, which comes up to be about 0.91, that is about 9% drop. The arithmetic mean would have conveyed a 0% average change.”
A histogram is a plot showing the distribution of a set of values. The histogram shows not only the range of values, but how they are distributed. The histogram represents the frequency distribution of the number of occurrences in each class interval.
samp <- c(0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1)

# histogram of samp data using base R graphics
hist(samp, main = "Histogram of Sample Data")
Histogram 3

Histogram of samp

Frequency

samp
Median

- When the discrete data are rearranged in increasing order and \( n \) is odd, the median is the value of the \( \left(\frac{n+1}{2}\right)^{th} \) item.

- When \( n \) is even, the median is the average of the \( \left(\frac{n}{2}\right)^{th} \) and \( \left(\frac{n}{2}+1\right)^{th} \) items.
Median Example 1

- 17 observations
- 0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1, 89, 0, 34, 65, 98, 3
- Find the median
- Solution:
  - n = 17, odd
  - Rearrange in increasing order
  - 0, 0, 0.0213, 0.5, 1, 3, 11, 23, 34, 45, 65, 89, 98, 99, 100, 345, 1000.25
  - median = 34
  \[
  \left(\frac{n+1}{2}\right) = \left(\frac{17+1}{2}\right) = 9\text{th item}
  \]
Median Example 1

- Find the median of the sample observations using R
- Solution:

```r
samp1 <- c(0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1, 89, 0, 34, 65, 98, 3)
length(samp1)
median(samp1)
[1] 34
```
Median Example 2

- 18 observations
- 0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1, 89, 0, 34, 65, 98, 3, 61
- Find the median
- Solution:
  - n = 18, even
  - Rearrange in increasing order
  - 0, 0, 0.0213, 0.5, 1, 3, 11, 23, 34, 45, 61, 65, 89, 98, 99, 100, 345, 1000.25
  - median = average of 34 and 45
  - median = 39.5

\[ \left( \frac{n}{2} \right) = 9\text{ th item} = 34 \]

\[ \left( \frac{n}{2} + 1 \right) = 10\text{ th item} = 45 \]
Median Example 2

• Find the median of the sample observations using R

• Solution:

```r
samp2 <- c(0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1, 89, 0, 34, 65, 98, 3, 61)
length(samp2)
median(samp2)
[1] 39.5
```
Mode

- The mode of a set of data is the value that occurs with greatest frequency.
Mode Example 1

- 17 observations
- 0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1, 89, 0, 34, 65, 98, 3
- Find the mode
- Solution:
  - mode = 0
Mode Example 2

- Find the mode of the sample observations using R
- Solution:

```r
samp <- c(0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1, 89, 0, 34, 65, 98, 3)
library(DescTools)
Mode(samp)
```

```
[1] 0
```
The sample range $R$ is the largest sample value minus the smallest sample value. The sample range $R$ can also be the maximum sample value and the minimum sample value.
Sample range Example 1

- 17 observations
- 0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1, 89, 0, 34, 65, 98, 3
- Find the sample range
- Solution:
- The maximum value is 1000.25 and the minimum value is 0.

\[ R = 1000.25 - 0 = 1000.25 \]
Sample range Example 2

- Find the range of the sample observations using R
- Solution:

```r
samp <- c(0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1, 89, 0, 34, 65, 98, 3)
length(samp)
range(samp)
[1] 0.00 1000.25

source("ranges.R")
ranges(samp)
The sample range is 1000 units.
[1] 1000.25 # units
Test Scores Example 1
(Onwubiko 177)

• Test scores for a class

<table>
<thead>
<tr>
<th>95</th>
<th>81</th>
<th>63</th>
<th>84</th>
<th>91</th>
<th>70</th>
<th>88</th>
<th>97</th>
</tr>
</thead>
<tbody>
<tr>
<td>70</td>
<td>76</td>
<td>80</td>
<td>77</td>
<td>78</td>
<td>91</td>
<td>93</td>
<td>60</td>
</tr>
<tr>
<td>85</td>
<td>79</td>
<td>84</td>
<td>90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Use R to determine the following: a) range 1, b) range 2, c) standard deviation, d) variance, e) mean, f) geometric mean, g) median
Test Scores Example 2

scores <- c(95, 81, 63, 84, 91, 70, 88, 97, 70, 76, 80, 77, 78, 91, 93, 60, 85, 79, 84, 90)

range(scores)
[1] 60 97

ranges(scores)
The sample range is 37 units.
[1] 37 # units

sd(scores)
[1] 10.29767 # units

var(scores)
[1] 106.0421 # units^2
scores <- c(95, 81, 63, 84, 91, 70, 88, 97, 70, 76, 80, 77, 78, 91, 93, 60, 85, 79, 84, 90)

mean(scores)
[1] 81.6 # units

sgm(scores)
The sample geometric mean is 81 units.
[1] 80.94735 # units

median(scores)
[1] 82 # units
Root-mean-square

- Sample root-mean-square value

\[ rms = \sqrt{\frac{1}{n} \sum_{i=1}^{n} X_i^2} \]
Sample Root-mean-square Example 1

- 17 observations
- 0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1, 89, 0, 34, 65, 98, 3
- Find the root-mean-square value
- Solution:

\[ rms = 261.77 \approx 262 \text{ units} \]
Sample Root-mean-square Example 2

- Find the range of the sample observations using R
- Solution:

```r
samp <- c(0.5, 100, 1000.25, 345, 0.0213, 0, 45, 99, 23, 11, 1, 89, 0, 34, 65, 98, 3)
length(samp)
source("rmssam.R")
rmssam(samp)
The sample root-mean-square value is 262 units.

[1] 261.7719 # units
Relative error (Chapra 82)

$$\varepsilon_t = \frac{\text{true value} - \text{approximation}}{\text{true value}} \times 100$$
Relative error Example 1

• What is the relative error between the function $c$ (cfunct) and the tabulated values of $c$ (ctable) for the same time range?
Here is some data for concentration (c) versus time (t) for the photodegradation of aqueous bromine:

<table>
<thead>
<tr>
<th>t, min</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>c, ppm</td>
<td>3.4</td>
<td>2.6</td>
<td>1.6</td>
<td>1.3</td>
<td>1.0</td>
<td>0.5</td>
</tr>
</tbody>
</table>
This data can be described by the following function:

cfunct = \(4.84e^{-0.034t}\)

Use R to calculate cfunct for \(t = 10\) to 60 min.
Relative error Example 4

- Find the relative error using R
- Solution:

```r
ctable <- c(3.4, 2.6, 1.6, 1.3, 1.0, 0.5)
t <- seq(10, 60, by = 10)
cfunct <- 4.84 * exp(-0.034 * t)
source("relerror.R")
relerror(ctable, cfunct)
```

The relative error is -1 6 -9 4 12 -26 %.

Consider the following data for a launcher aiming at a target 100 m downrange:

<table>
<thead>
<tr>
<th>Launcher B (m)</th>
<th>96</th>
<th>107</th>
<th>95</th>
<th>100</th>
<th>108</th>
<th>85</th>
<th>116</th>
<th>102</th>
<th>90</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The error for a launch $i$ at a target distance $d_{\text{target}}$ is defined as the actual distance $d_i$ of the launch minus the target distance:

$$\text{err}_i = d_{\text{target}} - d_i$$

Store the trial results in a single R vector variable and then answer the questions below.
d_i <- c(96, 107, 95, 100, 108, 85, 116, 102, 90, 111)

• a) Write a R vector expression to calculate the error associated with each of the trials.

d_target <- 100 # m
erg_i <- d_target - d_i

[1] 4 -7 5 0 -8 15 -16 -2 10 -11 # m

• b) Write a R vector expression to calculate the average error for each of the trials.

avg_error <- mean(err_i)

[1] -1 # m
c) Write a R vector expression to calculate the average of the square of the error for each of the trials, and then takes the square root of this quantity.

```r
err_i_sq <- (d_target - d_i) ^ 2
avg_err_i_sq <- mean(err_i_sq)
avg_sq_error <- sqrt(avg_err_i_sq)
avg_sq_error
[1] 9.273618 # m
```
• d) Use the R `sd` function to find the standard deviation of the data.

```
sd(d_i)
[1] 9.718253 # m
```
Linear regression 1

- Linear least-squares regression
  
  \[ y = a_0 + a_1 x + e \]

- \( a_0 \) and \( a_1 \) are the y-intercept and the slope coefficients

- \( e \) is the error or residual between the “best fit” line and the observations
  
  \[ e = y - a_0 - a_1 x \]

  - \( e \) is the difference between the true value of \( y \) and the approximate value, \( a_0 + a_1 x \), determined by the “best fit” linear equation
Linear regression 2

\[ a_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \]

\[ n = \text{sample length} \]

\[ y = \frac{1}{n} \sum_{i=1}^{n} y_i \]

\[ x = \frac{1}{n} \sum_{i=1}^{n} x_i \]

\[ a_0 = \bar{y} - a_1 \bar{x} \]
Linear regression 3

- Error estimation

\[ S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 \]

\[ S_t = \sum_{i=1}^{n} (y_i - \bar{y})^2 \]

- \( S_r \) – sum of the squares
- \( S_t \) – square of the residual represented by the square of the difference between the data and the mean
Linear regression 4

- Standard error of the estimate ($S_{y/x}$)

$$S_{y/x} = \sqrt{\frac{S_r}{n-2}}$$

- Coefficient of determination ($r^2$ or $R^2$)

$$r^2 = \frac{S_t - S_r}{S_t}$$
Linear regression 5

- $r^2$ – between 0 and 1
  - 1 (perfect fit)
- $S_r$ – between $S_r = S_t$ and 0
  - 0 (perfect fit)
• Correlation coefficient \((r)\)

\[
r = \sqrt{r^2}
\]

\[
r = \frac{n \sum_{i=1}^{n} (x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \sqrt{n \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}}
\]
Problem Statement: Fit a straight line to the values in the following table

<table>
<thead>
<tr>
<th>v, m/s</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td>F, N</td>
<td>25</td>
<td>70</td>
<td>380</td>
<td>550</td>
<td>610</td>
<td>1220</td>
<td>830</td>
<td>1450</td>
</tr>
</tbody>
</table>

Experimental data for force (N) and velocity (m/s) from a wind tunnel experiment
Solution: In this application, force is the dependent variable \((y)\) and velocity is the independent variable \((x)\). The data can be set up in tabular form and the necessary sums computed in the following table.

Fitting a straight line to the data requires that you perform a linear regression analysis.
### Linear regression Example 1

<table>
<thead>
<tr>
<th>i</th>
<th>$x_i$</th>
<th>$y_i$</th>
<th>$x_i^2$</th>
<th>$x_i y_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>25</td>
<td>100</td>
<td>250</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>70</td>
<td>400</td>
<td>1400</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>380</td>
<td>900</td>
<td>11400</td>
</tr>
<tr>
<td>4</td>
<td>40</td>
<td>550</td>
<td>600</td>
<td>22000</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>610</td>
<td>2500</td>
<td>30500</td>
</tr>
<tr>
<td>6</td>
<td>60</td>
<td>1220</td>
<td>3600</td>
<td>73200</td>
</tr>
<tr>
<td>7</td>
<td>70</td>
<td>830</td>
<td>4900</td>
<td>58100</td>
</tr>
<tr>
<td>8</td>
<td>80</td>
<td>1450</td>
<td>6400</td>
<td>116000</td>
</tr>
<tr>
<td>Σ</td>
<td>360</td>
<td>5,135</td>
<td>20,400</td>
<td>312,850</td>
</tr>
</tbody>
</table>
Linear regression Example 1

- Calculate the mean values

\[ \bar{x} = \frac{360}{8} = 45 \]

\[ \bar{y} = \frac{5135}{8} = 641.875 \]

- Calculate the slope and y-intercept

\[ a_1 = \frac{8(312,850) - 360(5,135)}{8(20,400) - (360)^2} = 19.47024 \]

\[ a_0 = 641.875 - 19.47024(45) = -234.2857 \]
### Linear regression Example

\[
a_i = a_0 + a_1 x_i
\]

<table>
<thead>
<tr>
<th>(i)</th>
<th>(x_i)</th>
<th>(y_i)</th>
<th>(a_0 + a_1 x_i)</th>
<th>((y_i - \bar{y})^2)</th>
<th>((y_i - a_0 - a_1 x_i)^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>25</td>
<td>-39.58</td>
<td>380535</td>
<td>4171</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>70</td>
<td>155.12</td>
<td>327041</td>
<td>7245</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>380</td>
<td>349.82</td>
<td>68579</td>
<td>911</td>
</tr>
<tr>
<td>4</td>
<td>40</td>
<td>550</td>
<td>544.52</td>
<td>8441</td>
<td>30</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>610</td>
<td>739.23</td>
<td>1016</td>
<td>16699</td>
</tr>
<tr>
<td>6</td>
<td>60</td>
<td>1220</td>
<td>933.93</td>
<td>334229</td>
<td>81837</td>
</tr>
<tr>
<td>7</td>
<td>70</td>
<td>830</td>
<td>1128.63</td>
<td>35391</td>
<td>89180</td>
</tr>
<tr>
<td>8</td>
<td>80</td>
<td>1450</td>
<td>1323.33</td>
<td>653066</td>
<td>16044</td>
</tr>
<tr>
<td>(\Sigma)</td>
<td>360</td>
<td>5,135</td>
<td>216,188</td>
<td>1,808,297</td>
<td></td>
</tr>
</tbody>
</table>
Linear regression Example 1

1. Calculate the standard deviation
   
   \[ s_y = \sqrt{\frac{1,808,297}{8-1}} = 508.26 \]

2. Calculate the standard error of the estimate
   
   \[ s_{y/x} = \sqrt{\frac{216,118}{8-2}} = 189.79 \]

3. Calculate the coefficient of determination
   
   \[ r^2 = \frac{1,808,297 - 216,118}{1,808,297} = 0.8805 \]

4. The results show that 88.05% of the original uncertainty has been explained by the linear model defined by the equation (this equation is used to create the straight line to fit the values)
   
   \[ F = 19.47024 \nu - 234.2857 \]
Linear regression Example 1

- Plot the data points along with the line generated by the equation on the previous slide \((F = 19.47024 \times v - 234.2857)\) using R (the process to plot by hand is similar)

- Solution:

  \[
  \begin{align*}
  v & \leftarrow \text{seq}(10, 80, \text{by} = 10) \quad \# \text{Generate x-values (v) for the data points from 10 to 80 by 10} \\
  F & \leftarrow \text{c}(25, 70, 380, 550, 610, 1220, 830, 1450) \quad \# \text{These are the y-values (F) for the data points (from the table)} \\
  \text{vequation} & \leftarrow \text{c}(0:80) \quad \# \text{Generate x-values (vequation) for the linear regression line from 0 to 80 by 1} \\
  \text{Fequation} & \leftarrow 19.47024 \times \text{vequation} - 234.2857 \quad \# \text{Generate y-values (Fequation) for the linear regression line}
  \end{align*}
  \]

  \[
  \text{dfeq} \leftarrow \text{data.frame(vequation, Fequation)}
  \]

  \[
  \text{dfp} \leftarrow \text{data.frame(v, F)}
  \]

  \[
  \text{library(ggplot2)}
  \]

  \[
  \text{ggplot()} + \text{geom_line(data = dfeq, aes(x = vequation, y = Fequation, label = Fequation))} + \\
  \text{geom_text(aes(label = "F = 19.47024 \times v - 234.2857 (linear regression line)", x = 20, y = 1000), vjust = 2) + labs(list(title = "Experimental Data From a Wind Tunnel", x = "Velocity (m/s)", y = "Force (N)"))) + geom_point(data = dfp, size = 3, aes(x = v, y = F, show_guide = TRUE))}
  \]
Experimental Data From a Wind Tunnel

$F = 19.47024 \, v - 234.2857 \text{ (linear regression line)}$
• Perform the linear regression using R

Solution:

v <- seq(10, 80, by = 10)
F <- c(25, 70, 380, 550, 610, 1220, 830, 1450)
lm.F <- lm(F ~ v)
coef(lm.F)
(Intercept) v
-234.28571 19.47024
summary(lm.F)

Call:
lm(formula = F ~ v)

Multiple R-squared: 0.8805

- The results show that 88.05% (multiple $r^2$) of the original uncertainty has been explained by the linear model defined by the equation

$$F = 19.47v - 234.29$$
df <- data.frame(v, F) # create data.frame for use in ggplot2 plot
library(ggplot2)
source("ggplot_smooth_func2.R")

ggplot(data = df, aes(x = v, y = F, label = F)) +
stat_smooth_func(geom = "text", method = "lm", hjust = 0,
parse = TRUE) + geom_smooth(method = "lm", se = FALSE) + geom_point() + labs(list(title = "Experimental Data From a Wind Tunnel", x = "Velocity (m/s)", y = "Force (N)"))

\[ F = 19.47v - 234.29 \]
Linear regression Example 1 (plot)

Experimental Data From a Wind Tunnel

\[ F = -234 + 19.5 \cdot v, \quad r^2 = 0.88 \]
A sample of nine polishing times for bowls and their diameters.

Use the linear regression line to estimate the polishing time of a 10.0 inch diameter bowl.
• Use R to estimate the polishing time of a 10.0 inch diameter bowl.

Solution:
Diameter <- c(7.4, 5.4, 7.5, 14.0, 7.0, 9.0, 12.0, 5.5, 6.0) # inches

Time <- c(16.41, 12.02, 20.21, 32.62, 17.84, 22.82, 29.48, 15.61, 13.25) # min
Linear regression Example 2 (Olia 92)

```r
lm.T <- lm(Time ~ Diameter) # linear model
coeff(lm.T) # coefficients of linear model
lm.T
summary(lm.T) # summary of linear model

df <- data.frame(Diameter, Time) # create data.frame for use in ggplot2 plot
library(ggplot2)
source("ggplot_smooth_func3.R")
ggplot(data = df, aes(x = Diameter, y = Time, label = Time)) + stat_smooth_func(geom = "text", method = "lm", hjust = 0, parse = TRUE) + geom_smooth(method = "lm", se = FALSE) + geom_point() + labs(list(title = "Polishing Times for Bowls and Their Diameters", x = "Diameter (inches)", y = "Time (min)"))

Diameter10in <- 10 # inches
Time10in <- coef(lm.T)[[2]]*Diameter10in + coef(lm.T)[[1]]
Time10in
[1] 24.21794 # minutes
```
Polishing Times for Bowls and Their Diameters

\[ Time = 0.945 + 2.33 \cdot Diameter, \quad r^2 = 0.957 \]
Probability

- Basics
- Laws
- Permutations
- Combinations
- Failure Rate
• Estimated probability of an event is

\[
\text{estimated experimental probability} = \frac{\text{number of successful trials}}{\text{total number of trials}}
\]

For example, if you have 10 successful trials out of 23 trials, then the probability is

\[
\frac{10}{23} = 0.4347826
\]
Probability Laws 1

- **Law 1 (General Character of Probability)**
  - Probability $P(E)$ of an event $(E)$ is a real number ranging from 0 (impossible event) to 1 (certain event).

- **Law 2 (Law of Total Probability)**
  - $P(A + B) = P(A) + P(B) - P(A, B)$
  
  - $P(A + B)$ – Probability that either $A$ or $B$ occur alone or that both occur together
  
  - $P(A)$ – Probability that $A$ occurs
  
  - $P(B)$ – Probability that $B$ occurs
  
  - $P(A,B)$ – Probability that both $A$ and $B$ occur simultaneously
Law 3 (Law of Compound or Joint Probability)

- If neither \( P(A) \) nor \( P(B) = 0 \)

\[
P(A, B) = P(A)P(B \lor A) = P(B)P(A \lor B)
\]

- \( P(B | A) \) = Probability that \( B \) occurs given the fact that \( A \) has occurred

- \( P(A | B) \) = Probability that \( A \) occurs given the fact that \( B \) has occurred.

- If either \( P(A) \) or \( P(B) \) is zero, then \( P(A, B) = 0 \)
Three standard 52-card decks are used in a probability experiment. One card is drawn from each deck. What is the probability that a diamond is drawn from the first deck, an ace from the second, and the ace of hearts from the third?
Solution:

The trials are independent of each other because different decks are used.

Use Probability Law 3 as the probability that all three events will occur is the product of the three individual probabilities.

\[ P(\text{diamond, ace, and ace of hearts}) = P(\text{diamond}) \times P(\text{ace}) \times P(\text{ace of hearts}) \]

\[ P(3) = \left( \frac{13}{52} \right) \left( \frac{4}{52} \right) \left( \frac{1}{52} \right) = 0.00037 \]
Probability Example 1

- Find the probability using R

Solution:

```r
library(prob)
cds <- cards(makespace = TRUE) # include the probability column in the cards function and create a data.frame called cds of the cards function

# What is the probability that a diamond is drawn from the first deck, an ace from the second, and the ace of hearts from the third?

Diamond <- subset(cds, suit == "Diamond") # subset cds with only the Diamond suit
Diamondprob <- Prob(Diamond) # Calculates the probability

Ace <- subset(cds, rank == "A") # subset cds with only Aces
Aceprob <- Prob(Ace) # Calculates the probability

AceHearts <- subset(cds, rank == "A" & suit == "Heart") # subset cds with only the Heart Ace
AceHeartsprob <- Prob(AceHearts) # Calculates the probability

probcards <- Diamondprob * Aceprob * AceHeartsprob
probcards
[1] 0.0003698225
```
Two students are working independently on a problem. Their respective probabilities of solving the problem are 1/3 and ¾. What is the probability that at least one of them will solve the problem?
Probability Example 2
(Lindeburg)

- Solution:

The probability that either or both of the students solve the problem is given by Probability Law 2.

\[ P(A) = \left(\frac{1}{3}\right) \]
\[ P(B) = \left(\frac{3}{4}\right) \]
\[ P(A+B) = P(A) + P(B) - P(A, B) \]
\[ P(A+B) = \frac{1}{3} + \frac{3}{4} - \left(\frac{1}{3}\right)\left(\frac{3}{4}\right) \]
\[ P(A+B) = \frac{10}{12} = \frac{5}{6} \]
Probability Example 2

- Find the probability using R
- Solution:

source("probAB.R")

A <- 1/3
B <- 3/4
probAB(A, B)

The probability is 0.83.

[1] 0.83333333
What is the probability of picking an orange ball and a white ball out of a bag containing seven orange balls, eight green balls, and two white balls?
Solution:
The possible successful outcomes are that either a white ball is picked and then an orange ball, or that an orange ball is picked and then a white ball.

\[ P(WOOW) = \left( \frac{7}{17} \right) \left( \frac{2}{16} \right) + \left( \frac{2}{17} \right) \left( \frac{7}{16} \right) \]

\[ P(WOOW) = 0.05147 + 0.05147 \]

\[ P(WOOW) = 0.1029 \]
Find the probability using R

Solution:

```r
library(prob)

bag <- rep(c("Orange", "Green", "White"), times = c(7, 8, 2)) # Create a bag of 7 orange, 8 green, and 2 white balls

bagsample <- urnsamples(bag, size = 2, replace = FALSE, ordered = FALSE) # Create a sample space of picking 2 balls where there is no replacement and the order does not matter

bagsampleProbspace <- probspace(bagsample) # Probability space of all events occurring

bagsampleProb <- Prob(bagsampleProbspace, isin(bagsampleProbspace, c("White", "Orange"))) # The probability of picking a white or orange ball

bagsampleProb

[1] 0.1029412
```
Permutation (Lindeburg)

- A particular sequence of a given set of objects, i.e. the order does matter.
- Number of different permutations of $n$ distinct objects taken $r$ at a time is

$$P(n,r) = nPr = \frac{n!}{(n-r)!}$$
If we start with 5 people, but we wish to make portraits of only 3 of them at a time. How many distinct portraits are possible?

- \( n = 5 \)
- \( r = 3 \)

\[
P(5,3) = 5 P_3 = \frac{5!}{(5-3)!} = \frac{5!}{2!} = \frac{120}{2} = 60
\]
Permutation Example

- Calculate the permutation using R
- Solution:

```r
library(prob)
ncol(permsn(5, 3))
[1] 60
permsn(5, 3)
```
Number of different combinations of \( n \) distinct objects taken \( r \) at a time is

\[
C(n,r) = \binom{n}{r} = \frac{P(n,r)}{r!}
\]

\[
C(n,r) = \frac{n!}{[r!(n-r)!]}
\]

In a combination the order does not matter.
Suppose that we wish to form committees of 3 people from a group of 5 people. How many such distinct committees are possible?

\[ C(5,3) = \frac{5!}{3!(5-3)!} \]

\[ C(5,3) = \frac{5!}{3!(2)!} = \frac{120}{6 \times 2} = 10 \]
Combination Example

• Compute the combination using R
• Solution:

library(combinat)
ncol(combn(5, 3))

[1] 10
combn(5, 3)
Suppose it is desired to estimate the failure rate of a certain component. A test can be performed to estimate its failure rate. Ten identical components are each tested until they either fail or reach 1000 hours, at which time the test is terminated for that component. (The level of statistical confidence is not considered in this example.)
The results are as follows:

- Estimated failure rate is

\[
\frac{6 \text{ failures}}{7502 \text{ hours}} = 0.0007998 \text{ failures} = 799.8 \times 10^{-6} \frac{\text{failures}}{\text{hour}}
\]

There are 799.8 failures for every million hours of operation.
Failure Rate 3

- Find the failure rate using R
- Solution:

```r
failure <- 6
hours <- 7502
est_failure_rate <- failure / hours
[1] 0.0007997867 = 799.8 * 10^-6
```

There are 799.8 failures for every million hours of operation.
Function files from today’s lecture can be found online:

Lecture Materials 2

- http://www.ecoccs.com/probAB.R
Reviewing the following pages from the *Fundamentals of Engineering Reference Handbook* (posted online at [http://www.ecoccs.com/tsuteach.html#engr1020](http://www.ecoccs.com/tsuteach.html#engr1020)) will be helpful for the remainder of this semester

- iii – 17,
- 19 – 24,
- 28 – 29,
- 40 – 48,
- 109, and
- 114 – 120


