Chapter 4

Statistics

Statistics is about “educated guessing”. It is about drawing conclusions from incomplete information. It is about collecting, organizing and analyzing data, where data could be thought of as observed values of random variables. One important aspect of statistics is to do with the idea of gathering together a random sample of data, calculating a statistic and using this statistic to infer something about, typically, a parameter of the population, more specifically a parameter from a probability distribution model describing the population, from which this sample is taken. This is generally called an inferential statistical analysis and this is what we will be concentrating on for the remainder of this course.

4.1 What is Statistics?

Types of data, types of studies and sampling techniques are reviewed in this section.

Types of Data

Data are observed from random variables. Consider the following related definitions.

- population, parameter, sample and statistic

  population: set of measurements or observations of a collection of objects.

  Sometimes “population” refers to the objects themselves, but more frequently it refers to the measurements or observations of these objects. These measurements or observations are modelled by approximate distribution functions of random variables.

  parameter: numerical quantity calculated from a population.

  Parameters are constants in approximate mathematical distribution functions of random variables.

  sample selected subset of a population.

  Sometimes “sample” refers to the objects themselves, but more frequently it refers to the measurements
or observations of these objects. This subset of measurements or observations are also modelled by approximate distribution functions of random variables.

**statistic** numerical quantity calculated from a sample.

Statistics are functions of a sample of random variables and so are random variables themselves, with distribution functions. The term “statistic” may refer to the random variable or to the observed value of this random variable.

- nominal, ordinal, interval or ratio (random) variable or data

  **nominal variable (data):** Variable where data cannot be ordered; data consists of names or labels.

  **ordinal (ranked) variable (data):** Variable where data can be ordered but cannot be added or subtracted.

  **interval variable (data):** Variable where data can be both ordered and added or subtracted, but cannot be divided or multiplied.

  When two data points are subtracted from one another, difference between two is an interval, which explains why this is called interval variable (data).

  **ratio variable (data):** Variable where data can be ordered, added or subtracted, and also divided or multiplied.

  When two data points are divided, a ratio is formed, which explains why this is called ratio variable (data).

- qualitative (categorical) versus quantitative

  **qualitative (categorical):** nominal and ordinal (ranked) variables (data)

  **quantitative:** interval and ratio variables (data), measurements or counts

- quantitative: discrete versus continuous

  **discrete variable (data):** Discrete variable where quantitative data is finite or countable infinite.

  **continuous variable (data):** Continuous variable where quantitative data takes any value in some range.

**Types of Studies**

Point of many statistical studies is to determine if and how variables, called explanatory variables (factors), influence an outcome or response variable.

**observational study:** subject decides whether or not to be given treatment; subjects are not changed or modified in any way
**Section 1. What is Statistics? (LECTURE NOTES 7)**

*cross-sectional:* data collected at one time  
*retrospective:* data collected from the past  
*prospective:* data collected over time, into the future

*experiment:* experimenter decides who is to be given treatment and who is to be the control; subjects are changed in this way

*factor:* explanatory variable  
*nuisance factor:* factors of no interest to researcher  
*confounding factor:* when it is not clear how much each a nuisance factor or explanatory factor contributes in the prediction of a response variable  
*lurking factor:* a confounding factor unknown to the researcher  
*block:* subset of population with similar characteristics  
*randomized block design:* population divided into blocks, subjects selected from each block at random to receive treatment

**Sampling techniques**

A *simple random sample (SRS)* involves selecting a group of \( n \) units out of \( N \) population units where every distinct sample has an *equal* chance of being drawn whereas for a *random sample* each group is chosen with some given chance but not necessary *equal* chance. SRSs produce *representative* samples of population. Other (sometimes poor) sampling techniques:

*convenience:* choose easiest-to-get-at sample  
*voluntary response:* subjects in study decide whether or not to participate  
*systematic:* selecting every \( k \)th item from population, where first item is chosen at random  
*cluster:* involves dividing population into *heterogeneous* clusters, choosing a *subset* of these clusters at random, and then using all items in the selected clusters  
*stratified:* involves dividing population into *homogeneous* strata and choosing a simple random sample (SRS) from each/all strata

**Exercise 4.1 (What is Statistics?)**

1. *Types of data: variable and data.*
(a) (i) **True**  (ii) **False** A data point for (random) variable *height of a man* is 5.6 feet tall. Another data point for this variable is 5.8 feet tall.

(b) (i) **True**  (ii) **False** A data point for (random) variable *person’s country of birth* is Sweden. Another data point for this variable is U.S.A.

(c) (i) **True**  (ii) **False** A data point for variable *shoulder height of a cow* is 45 inches. Another data point for this variable is “Argentina”.

(d) A data point for variable *woman’s length of time of exposure in the sun* is (i) $9/24/98$  (ii) 4 hours.

(e) A data point for variable *person’s date of exposure to the sun* is (i) $9/24/98$  (ii) 4 hours.

(f) (i) **True**  (ii) **False** Data point “45” could be a particular instance of the variable *age of a elephant*. Data point “45” could also be a particular instance of the variable *number of marbles in a bag*.

(g) Data point “silver” is a particular instance of variable (circle none, one or more)
   i. length of football field
   ii. medal achieved at a track meet
   iii. color choice of a car
   iv. name of a horse

2. **Types of data: population, sample, statistic and parameter (commute distance)**

   At PNW, 120 students are randomly selected from entire 10,500 and asked their commute distance to campus. Average of 9.8 miles is computed from 120 selected. We infer from data all students have 9.8 average commute.

   (i) **True**  (ii) **False**. An appropriate analogy here would be to think of a box of 10,500 tickets where each ticket has commute distance written on it as *population*; average of all of these tickets would be value of *parameter*. A random sample of 120 tickets taken from this population box would be a *sample*; average of sampled tickets would be value of *statistic*.

   “Sample” may also refer to 10,500 tickets themselves, whatever is written on them.

Match columns.
Section 1. What is Statistics? (LECTURE NOTES 7)

<table>
<thead>
<tr>
<th>terms</th>
<th>travel example</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) data point</td>
<td>(A) average commute distance for 120 students</td>
</tr>
<tr>
<td>(b) variable</td>
<td>(B) all students at PNW</td>
</tr>
<tr>
<td>(c) parameter</td>
<td>(C) commute distances for all students at PNW</td>
</tr>
<tr>
<td>(d) population</td>
<td>(D) commute distance for any PNW student</td>
</tr>
<tr>
<td>(e) sample</td>
<td>(E) average commute distance for all students</td>
</tr>
<tr>
<td>(f) statistic</td>
<td>(F) 120 students</td>
</tr>
<tr>
<td></td>
<td>(G) 120 commute distances</td>
</tr>
<tr>
<td></td>
<td>(H) 8 mile commute distance for a particular student</td>
</tr>
</tbody>
</table>

Some items in first column have more than one match; for example, (d) population matches with both (b) all students at PNW and (c) commute distances for all students at PNW.

3. Types of data: nominal, ordinal (ranked), interval and ratio (milk yield) Measurements are given below on a number of cows taken during a study on effect of a hormone, given in tablet form, on daily milk yield. Eight variables, including “Cow”, “Test Date”, ..., “After Yield”, are listed at top of columns in table. Seven observations (data points) are listed in seven rows below variables. Specify milk yield variables (data) as nominal, ordinal, interval or ratio.

<table>
<thead>
<tr>
<th>Cow</th>
<th>Test Date</th>
<th>Farm</th>
<th>Height</th>
<th>Health</th>
<th>Tablets</th>
<th>Before Yield</th>
<th>After Yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>9/11/98</td>
<td>M</td>
<td>41</td>
<td>poor</td>
<td>2</td>
<td>100.7</td>
<td>100.3</td>
</tr>
<tr>
<td>18</td>
<td>9/11/98</td>
<td>F</td>
<td>40</td>
<td>bad</td>
<td>1</td>
<td>97.8</td>
<td>98.1</td>
</tr>
<tr>
<td>14</td>
<td>9/03/98</td>
<td>F</td>
<td>49</td>
<td>fair</td>
<td>3</td>
<td>98.8</td>
<td>99.6</td>
</tr>
<tr>
<td>15</td>
<td>9/01/98</td>
<td>M</td>
<td>45</td>
<td>good</td>
<td>3</td>
<td>100.9</td>
<td>100.0</td>
</tr>
<tr>
<td>16</td>
<td>9/10/98</td>
<td>F</td>
<td>42</td>
<td>poor</td>
<td>1</td>
<td>101.1</td>
<td>100.1</td>
</tr>
<tr>
<td>19</td>
<td>9/25/98</td>
<td>M</td>
<td>45</td>
<td>good</td>
<td>2</td>
<td>100.0</td>
<td>100.4</td>
</tr>
<tr>
<td>20</td>
<td>9/25/98</td>
<td>M</td>
<td>37</td>
<td>good</td>
<td>3</td>
<td>101.5</td>
<td>100.8</td>
</tr>
</tbody>
</table>

Match columns.

<table>
<thead>
<tr>
<th>milk yield example</th>
<th>level of measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) cow ID number</td>
<td>(A) nominal</td>
</tr>
<tr>
<td>(b) test date of cow</td>
<td>(B) ordinal</td>
</tr>
<tr>
<td>(c) cow’s farm</td>
<td>(C) interval</td>
</tr>
<tr>
<td>(d) shoulder height of cow</td>
<td>(D) ratio</td>
</tr>
<tr>
<td>(e) cow’s health</td>
<td></td>
</tr>
<tr>
<td>(f) number of tablets given to cow</td>
<td></td>
</tr>
<tr>
<td>(g) milk yield before hormone</td>
<td></td>
</tr>
<tr>
<td>(h) milk yield after hormone</td>
<td></td>
</tr>
</tbody>
</table>
4. **Types of data: qualitative (categorical) versus quantitative (milk yield)**

<table>
<thead>
<tr>
<th>Cow</th>
<th>Test Date</th>
<th>Farm</th>
<th>Height</th>
<th>Health</th>
<th>Tablets</th>
<th>Before Yield</th>
<th>After Yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>9/11/98</td>
<td>M</td>
<td>41</td>
<td>poor</td>
<td>2</td>
<td>100.7</td>
<td>100.3</td>
</tr>
<tr>
<td>18</td>
<td>9/11/98</td>
<td>F</td>
<td>40</td>
<td>bad</td>
<td>1</td>
<td>97.8</td>
<td>98.1</td>
</tr>
<tr>
<td>14</td>
<td>9/03/98</td>
<td>F</td>
<td>49</td>
<td>fair</td>
<td>3</td>
<td>98.8</td>
<td>99.6</td>
</tr>
<tr>
<td>15</td>
<td>9/01/98</td>
<td>M</td>
<td>45</td>
<td>good</td>
<td>3</td>
<td>100.9</td>
<td>100.0</td>
</tr>
<tr>
<td>16</td>
<td>9/10/98</td>
<td>F</td>
<td>42</td>
<td>poor</td>
<td>1</td>
<td>101.1</td>
<td>100.1</td>
</tr>
<tr>
<td>19</td>
<td>9/25/98</td>
<td>M</td>
<td>45</td>
<td>good</td>
<td>2</td>
<td>100.0</td>
<td>100.4</td>
</tr>
<tr>
<td>20</td>
<td>9/25/98</td>
<td>M</td>
<td>37</td>
<td>good</td>
<td>3</td>
<td>101.5</td>
<td>100.8</td>
</tr>
</tbody>
</table>

Specify whether the milk yield variables are either quantitative or qualitative.

<table>
<thead>
<tr>
<th>milk yield example</th>
<th>type of variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) cow ID number</td>
<td>(A) qualitative</td>
</tr>
<tr>
<td>(b) test date of cow</td>
<td>(B) quantitative</td>
</tr>
<tr>
<td>(c) cow’s farm</td>
<td></td>
</tr>
<tr>
<td>(d) shoulder height of cow</td>
<td></td>
</tr>
<tr>
<td>(e) cow’s health</td>
<td></td>
</tr>
<tr>
<td>(f) number of tablets given to cow</td>
<td></td>
</tr>
<tr>
<td>(g) milk yield before study</td>
<td></td>
</tr>
<tr>
<td>(h) milk yield after study</td>
<td></td>
</tr>
</tbody>
</table>

5. **Types of data: discrete versus continuous (milk yield)**

<table>
<thead>
<tr>
<th>Cow</th>
<th>Test Date</th>
<th>Farm</th>
<th>Height</th>
<th>Health</th>
<th>Tablets</th>
<th>Before Yield</th>
<th>After Yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>9/11/98</td>
<td>M</td>
<td>41</td>
<td>poor</td>
<td>2</td>
<td>100.7</td>
<td>100.3</td>
</tr>
<tr>
<td>18</td>
<td>9/11/98</td>
<td>F</td>
<td>40</td>
<td>bad</td>
<td>1</td>
<td>97.8</td>
<td>98.1</td>
</tr>
<tr>
<td>14</td>
<td>9/03/98</td>
<td>F</td>
<td>49</td>
<td>fair</td>
<td>3</td>
<td>98.8</td>
<td>99.6</td>
</tr>
<tr>
<td>15</td>
<td>9/01/98</td>
<td>M</td>
<td>45</td>
<td>good</td>
<td>3</td>
<td>100.9</td>
<td>100.0</td>
</tr>
<tr>
<td>16</td>
<td>9/10/98</td>
<td>F</td>
<td>42</td>
<td>poor</td>
<td>1</td>
<td>101.1</td>
<td>100.1</td>
</tr>
<tr>
<td>19</td>
<td>9/25/98</td>
<td>M</td>
<td>45</td>
<td>good</td>
<td>2</td>
<td>100.0</td>
<td>100.4</td>
</tr>
<tr>
<td>20</td>
<td>9/25/98</td>
<td>M</td>
<td>37</td>
<td>good</td>
<td>3</td>
<td>101.5</td>
<td>100.8</td>
</tr>
</tbody>
</table>
Match columns.

<table>
<thead>
<tr>
<th>milk yield example</th>
<th>type of variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) cow ID number</td>
<td>(A) discrete</td>
</tr>
<tr>
<td>(b) test date of cow</td>
<td>(B) continuous</td>
</tr>
<tr>
<td>(c) cow’s farm</td>
<td>(C) qualitative</td>
</tr>
<tr>
<td>(d) shoulder height of cow</td>
<td></td>
</tr>
<tr>
<td>(e) cow’s health</td>
<td></td>
</tr>
<tr>
<td>(f) number of tablets given to cow</td>
<td></td>
</tr>
<tr>
<td>(g) milk yield before hormone</td>
<td></td>
</tr>
<tr>
<td>(h) milk yield after hormone</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>example</th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
<th>(e)</th>
<th>(f)</th>
<th>(g)</th>
<th>(h)</th>
</tr>
</thead>
</table>

6. Types of studies: observational study or experiment.

(a) Effect of air temperature on rate of oxygen consumption (ROC) of four mice is investigated. ROC of one mouse at 0°F is 9.7 mL/sec for example.

<table>
<thead>
<tr>
<th>temperature (F°)</th>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROC (mL/sec)</td>
<td>9.7</td>
<td>10.3</td>
<td>11.2</td>
<td>14.0</td>
</tr>
</tbody>
</table>

Since experimenter (not a mouse!) decides which mice are subjected to which temperature, this is (choose one)

- (i) observational study
- (ii) designed experiment.

(b) Indiana police records from 1999–2001 on six drivers are analyzed to determine if there is an association between drinking and traffic accidents. One heavy drinker had 6 accidents for example.

<table>
<thead>
<tr>
<th>drinking →</th>
<th>heavy</th>
<th>light</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

This is an observational study because

- i. police decided who was going to drink and drive and who was not.
- ii. drivers decided who was going to drink and drive and who was not.

(c) A recent study was conducted to compare academic achievement (measured by final examination scores) of Internet students with classroom students. This is an observational study because

- i. instructor assigned students to classroom or internet.
ii. students decided to attend classroom or Internet class.

(d) *Effect of drug on patient response.* Response from one patient given drug A is 120 units for example.

\[
\text{drug} \rightarrow \begin{array}{ccc}
A & B & C \\
120 & 97 & 134 \\
140 & 112 & 142 \\
125 & 100 & 129 \\
133 & 95 & 137
\end{array}
\]

If this is a designed experiment, then

i. experimenter assigns drugs to patients.

ii. patients assigns drugs to themselves.

7. *Types of studies (observational): cross-sectional, retrospective, prospective.*

(a) *Effect of drinking on traffic accidents.*

Indiana police records from 1999–2001 are analyzed to determine if there is an association between drinking and traffic accidents.

\[
\begin{array}{ccc}
\text{drinking} & \text{heavy drinker} & 3 & 6 & 2 \\
& \text{light drinker} & 1 & 2 & 1
\end{array}
\]

If number of traffic accidents of drunk drivers is compared with sober drivers who both have similar characteristics such as age, gender, health and so on, this is a

i. cross-sectional observational study

ii. retrospective observational study

iii. prospective observational study

iv. experiment

(b) *Explanatory variables influencing traffic accidents.*

If a large group of individuals are observed over an extended period of time to determine explanatory variables contributing to traffic accidents, this is a (choose one)

i. cross-sectional observational study

ii. retrospective (case-control) observational study

iii. prospective (cohort) observational study

iv. designed experiment

(c) *Effect of teaching method on academic achievement.*

A recent study compares academic achievement (measured by final examination scores) of Internet students with classroom students. If data is collected for one set of exams given at one time, this is a
Section 1. What is Statistics? (LECTURE NOTES 7)

i. cross-sectional observational study
ii. retrospective observational study
iii. prospective observational study
iv. experiment

(d) Effect of temperature on mice rate of oxygen consumption.

<table>
<thead>
<tr>
<th>temperature (°F)</th>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROC (mL/sec)</td>
<td>9.7</td>
<td>10.3</td>
<td>11.2</td>
<td>14.0</td>
</tr>
</tbody>
</table>

This is a
i. cross-sectional observational study
ii. retrospective observational study
iii. prospective observational study
iv. experiment

8. Types of studies (experimental): explanatory variables, responses, confounding and lurking variables (mice).

<table>
<thead>
<tr>
<th>temperature (°F)</th>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROC (mL/sec)</td>
<td>9.7</td>
<td>10.3</td>
<td>11.2</td>
<td>14.0</td>
</tr>
</tbody>
</table>

(a) Explanatory variable (factor) considered in study is
i. temperature
ii. rate of oxygen consumption
iii. mice
iv. mouse weight

(b) Response is
i. temperature
ii. rate of oxygen consumption
iii. mice
iv. room temperature

(c) Possible explanatory variable not considered in study (choose two!)
   i. temperature
   ii. rate of oxygen consumption
   iii. noise level
   iv. mouse weight

(d) Mouse weight is lurking variable if confounded with temperature in, for example, following way.
Chapter 4. Statistics (LECTURE NOTES 7)

<table>
<thead>
<tr>
<th>temperature ($^\circ$F)</th>
<th>0$^\circ$</th>
<th>10$^\circ$</th>
<th>20$^\circ$</th>
<th>30$^\circ$</th>
</tr>
</thead>
<tbody>
<tr>
<td>mouse weight (oz)</td>
<td>10</td>
<td>14</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>ROC (mL/sec)</td>
<td>9.7</td>
<td>10.3</td>
<td>11.2</td>
<td>14.0</td>
</tr>
</tbody>
</table>

Hotter temperatures are associated with heavier mice. Hottest temperature, 30$^\circ$ F, is associated with heaviest mouse with weight

(i) 9.7 (ii) 14.0 (iii) 20 ounces

9. Types of studies (experimental): completely randomized, randomized block. Consider experiment to determine effect of temperature on mice ROC.

(a) Completely Randomized Design.

<table>
<thead>
<tr>
<th>temperature $\rightarrow$</th>
<th>70$^\circ$ F</th>
<th>$-10^\circ$ F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10.3</td>
<td>9.7</td>
</tr>
<tr>
<td></td>
<td>14.0</td>
<td>11.2</td>
</tr>
<tr>
<td></td>
<td>15.2</td>
<td>10.3</td>
</tr>
</tbody>
</table>

This is a completely randomized design with (choose one or more!)

i. one factor (temperature) with two treatments,
ii. three pair of mice matched by age,
iii. mice assigned to temperatures at random.

(b) Randomized Block Design.

<table>
<thead>
<tr>
<th>age $\downarrow$ temperature $\rightarrow$</th>
<th>70$^\circ$ F</th>
<th>$-10^\circ$ F</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 days</td>
<td>10.3</td>
<td>9.7</td>
</tr>
<tr>
<td>20 days</td>
<td>14.0</td>
<td>11.2</td>
</tr>
<tr>
<td>30 days</td>
<td>15.2</td>
<td>10.3</td>
</tr>
</tbody>
</table>

This is a randomized block design with (choose one or more!)

i. one factor (temperature) with two treatments,
ii. one block (age) with three levels,
iii. mice assigned to temperatures within each age block at random.

10. Sampling techniques: simple random sample versus random sample (decayed teeth)

(a) Small number of 20 children with decayed teeth is represented by box of tickets below. Child 17 is ticket with 3 decayed teeth for example.
Section 2. Summarizing Data (LECTURE NOTES 7) 

Choosing four children at random from the 20 is an example of a (i) random sample (ii) simple random sample.

(b) Small number of 20 children, broken into 5 batches of 4, with decayed teeth is represented by box of tickets below.

<table>
<thead>
<tr>
<th>Child 1</th>
<th>Child 2</th>
<th>Child 3</th>
<th>Child 4</th>
<th>Child 5</th>
<th>Child 6</th>
<th>Child 7</th>
<th>Child 8</th>
<th>Child 9</th>
<th>Child 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>9</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>10</td>
<td>2</td>
</tr>
</tbody>
</table>

Choosing a batch of children at random from the 5, then using all children in the batch is an example of a (i) random sample (ii) simple random sample.

11. Sampling techniques: simple, stratified, cluster or systematic? Match racing horses examples with sampling technique, “SRS” means “simple random sample”.

(a) (i) Simple (ii) Stratified (iii) Cluster (iv) Systematic

An SRS is taken from all racing horses.

(b) (i) Simple (ii) Stratified (iii) Cluster (iv) Systematic

All racing horses are listed from lightest to heaviest. Sample consists of taking every seventh racing horse from this list.

(c) (i) Simple (ii) Stratified (iii) Cluster (iv) Systematic

All racing horses are listed alphabetically, by name. Sample consists of taking every third racing horse from this list.

(d) (i) Simple (ii) Stratified (iii) Cluster (iv) Systematic

All racing horses are classified as light, middle or heavy weight horses. Sample consists of taking SRSs from the racing horses in each weight class.

(e) (i) Simple (ii) Stratified (iii) Cluster (iv) Systematic

Horses racing occurs in many cities in the U.S.A. Sample consists of taking SRSs from the racing horses in New York, Los Angeles and Chicago.

4.2 Summarizing Data

Bar graphs describe qualitative data, whereas histograms describe quantitative data. Some summary (sample) statistics include:

mean (average)

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$
with corresponding population mean parameter, $\mu$ and where, if $x_i \in \{0, 1\}$, $\bar{x} = \hat{p}$, sample proportion

**mode**: most frequent observation in sample

**variance**: 
\[ s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \]

with corresponding population variance $\sigma^2$; also, **standard deviation** $s = \sqrt{s^2}$ and with population standard deviation $\sigma^2$

**range**: maximum $-$ minimum value in sample

(100$p$)th percentile, $\pi_p$: number which is greater than (100$p$)$\%$ of data values.

**calculation**: arrange data $x_1 \leq x_2 \leq \ldots \leq x_n$;
if $L = np$ is **not** an integer, round up to next-larger integer, $\pi_p = X_L$;
otherwise, if it is, then $\pi_p = \frac{1}{2} (x_L + x_{L+1})$

**quartiles**: 25th, 50th, 75th percentiles are also called first, second and third quartiles, $p_1 = \pi_{0.25}$, $p_2 = \pi_{0.50}$, $p_3 = \pi_{0.75}$

**median**: the 25th percentile or second quartile

**five-number summary**: \[
\{ \text{min, } p_1 = \pi_{0.25}, \text{ median } = p_2 = \pi_{0.50}, p_3 = \pi_{0.75}, \text{ max} \} , \]

graphed using **boxplot**

**percentile rank**: 
\[ \text{percentile rank of } x = \frac{\text{number of data values less than } x}{\text{total number of data values}} \times 100\% \]

**statistic (sample) versus parameter (population)**:

<table>
<thead>
<tr>
<th>statistic (estimator, estimate)</th>
<th>parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean $X$, $\bar{x}$</td>
<td>$\mu = E[X]$</td>
</tr>
<tr>
<td>variance $S^2$, $s^2$</td>
<td>$\sigma^2 = Var[X] = E[X^2] - \mu^2$</td>
</tr>
<tr>
<td>standard deviation $S$, $s$</td>
<td>$\sigma = SD[X]$</td>
</tr>
<tr>
<td>proportion $\hat{P}$, $\hat{p}$</td>
<td>$p$</td>
</tr>
<tr>
<td>histogram/barplot</td>
<td>pdf/pmf</td>
</tr>
</tbody>
</table>

Exercise 4.2 (Summarizing Data)
1. **Bar graphs: patient health.** Health of random sample of twenty patients in a high blood pressure study are:

   good, good, fair, poor, bad, poor, great, fair, good, good, 
good, fair, fair, fair, good, poor, poor, bad, good, good.

Distribution table, bar graph and Pareto chart for this data is given below.

<table>
<thead>
<tr>
<th>health</th>
<th>frequency</th>
<th>relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>bad</td>
<td>2</td>
<td>(\frac{2}{20} = 0.10)</td>
</tr>
<tr>
<td>poor</td>
<td>4</td>
<td>(\frac{4}{20} = 0.20)</td>
</tr>
<tr>
<td>fair</td>
<td>5</td>
<td>(\frac{5}{20} = 0.25)</td>
</tr>
<tr>
<td>good</td>
<td>8</td>
<td>(\frac{8}{20} = 0.40)</td>
</tr>
<tr>
<td>great</td>
<td>1</td>
<td>(\frac{1}{20} = 0.05)</td>
</tr>
<tr>
<td>total</td>
<td>20</td>
<td>1.0</td>
</tr>
</tbody>
</table>

(a) **Review:** This qualitative data is
   (i) **nominal**  (ii) **ordinal**  (iii) **interval**  (iv) **ratio**
   because this data is ordered.
(b) Of 20 patients, (i) 2 (ii) 4 (iii) 5 (iv) 8 are in good health, there is (i) **more** (ii) **less** chance of being in good health than other categories.

(c) *Frequency* means (choose *one or more*)
   (i) **number** (ii) **count** (iii) **proportion** (iv) **percentage** (v) **category**

(d) *Relative frequency* means (choose *two*)
   i. number in a particular category
   ii. count in a particular category
   iii. proportion of observations that fall into each category
   iv. percentage of observations that fall into each category
   v. proportion of observations that fall into at least two categories

(e) Height of each vertical bar in bar graph corresponds to the relative frequency for each category. For example, vertical bar for “good” category has a height (or relative frequency) of (i) **0.30** (ii) **0.35** (iii) **0.40**.

(f) Adding all heights of vertical bars in five categories together, we get (i) **0.40** (ii) **0.75** (iii) **1.00**.

(g) **True** (ii) **False** Although height of each vertical bar gives relative frequency of a particular category of health occurring for twenty patients, *width* of each vertical bar has **no** meaning.

(h) A second random sample would give (i) **the same** (ii) **different** patients with (i) **the same** (ii) **different** patient health ratings, with (i) **the same** (ii) **different** bar plot.

(i) In general, every random sample gives (i) **the same** (ii) **different** patients with (i) **the same** (ii) **different** patient health ratings, with (i) **the same** (ii) **different** bar plots.

(j) Since each sample of patients is chosen at random, every patient has (i) **the same** (ii) **a different** chance of appearing in any sample but since there are only five health different health categories, every health rating has (i) **the same** (ii) **a different** chance of appearing in a sample.


<table>
<thead>
<tr>
<th>age class</th>
<th>frequency</th>
<th>relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>(30,35]</td>
<td>1</td>
<td>$\frac{1}{20} = 0.05$</td>
</tr>
<tr>
<td>(35,40]</td>
<td>3</td>
<td>0.15</td>
</tr>
<tr>
<td>(40,45]</td>
<td>10</td>
<td>0.50</td>
</tr>
<tr>
<td>(45,50]</td>
<td>5</td>
<td>0.25</td>
</tr>
<tr>
<td>(50,55]</td>
<td>1</td>
<td>0.05</td>
</tr>
</tbody>
</table>
Section 2. Summarizing Data (LECTURE NOTES 7)

Figure 4.2: Histogram and boxplot for patient ages

(a) Review: This quantitative data is
   (i) **nominal** (ii) **ordinal** (iii) **interval** (iv) **ratio**
   because this data can be divided in a meaningful way.

(b) Review: This quantitative data is
   (i) **discrete** (ii) **continuous**
   because this data can take any value in a (positive) range.

(c) Number of classes is (i) 3 (ii) 4 (iii) 5 (iv) 6.

(d) First class includes (i) 30 (ii) 35

(e) First class excludes (i) 30 (ii) 35

(f) Number of patients in first class is (i) 1 (ii) 2 (iii) 3 (iv) 4.

(g) Percentage of patients in first class is
   (i) 5% (ii) 10% (iii) 35% (iv) 40%.

(h) Shape of histogram roughly
   (i) **symmetric** (ii) **skewed right** (iii) **skewed left**

(i) Boxplot indicates there (i) is (ii) is no outlier.

(j) A second random sample would give (i) the same (ii) different patients
    with (i) the same (ii) different patient ages,
    with (i) the same (ii) different histogram and boxplot.
(k) If each sample of patients is chosen at random, every patient has
(i) the same (ii) a different chance of appearing in any sample but since
there are finite number of discrete ages, every patient age
has (i) the same (ii) a different chance of appearing in a sample.

(l) Summary statistics measuring the “central” age:

\[
\text{sample mean (average)} \quad \bar{x} = \frac{32 + 37 + \cdots + 51}{20} = 41.0
\]

(i) 41.0 (ii) 43.4 (iii) 43.5 (iv) 45.0

Sample median is middle age or average of middle two ages:

(i) 41.0 (ii) 43.4 (iii) 43.5 (iv) 45.0

Sample mode is most frequent age (choose two):

(i) 41.0 (ii) 43.4 (iii) 43.5 (iv) 45.0

Data are bimodal, although this is hidden in histogram

The sample mean \( \bar{x} = 43.35 \) is an example of the value of a
(i) statistic (ii) parameter based on patient ages,
were every random sample gives
(i) the same (ii) a different statistic value \( \bar{x} \); in general,
random variable \( \bar{X} \) has a number of different values \( \bar{x} \) occurring with
(i) the same (ii) different probability.

The median and mode (i) are (ii) are not (random variables) statistics.

\[
\text{md <- function(x) \{ # function calculating mode}
  ux <- unique(x)
  tab <- tabulate(match(x, ux)); ux[tab == max(tab)]
\}

mean(data); median(data); md(data)

[1] 43.35
[1] 43.5
[1] 41 45

(m) Summary statistics measuring “variability” of ages:

\[
\text{sample variance} \quad s^2 = \frac{1}{20-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \approx
\]

(i) 1.0 (ii) 4.6 (iii) 19.0 (iv) 20.9

Sample standard deviation \( s = \sqrt{s^2} \approx \)

(i) 1.0 (ii) 4.6 (iii) 19.0 (iv) 20.9

Sample range = maximum - minimum 51 - 32 = :

(i) 1.0 (ii) 4.6 (iii) 19.0 (iv) 20.9

The sample variance \( s^2 \approx 20.9 \) is an example of the value of a
(i) statistic (ii) parameter based on patient ages; in general,
random variable \( S^2 \) has a number of different values \( s^2 \) occurring with
(i) the same (ii) different probability

and is used as an estimator of population parameter
(i) \( p \) (ii) \( \mu \) (iii) \( \sigma \) (iv) \( \sigma^2 \).

Sample statistic standard deviation, \( S \), (i) is (ii) is not used as an estimator for population parameter \( \sigma \).
var(data); sd(data); range(data)

[1] 20.87105
[1] 4.568485
[1] 32 51

(n) Five-number summary which gives the boxplot is, since

\[
\text{min} = 32, \quad p_{1} = \pi_{0.25} = 41, \quad p_{2} = \pi_{0.50} = 44, \quad p_{3} = \pi_{0.75} = 47, \quad \text{max} = 51
\]

(i) \{41, 43.5, 46.5\}
(ii) \{32, 41, 43.5, 46.5, 51\}
(iii) \{32, 41, 44, 47, 51\}

For example, for \( p_{3} = \pi_{0.75} \), since \( L = np = 20 \cdot 0.75 = 15 \), average 15th and 16th numbers, \( \pi_{P} = \frac{1}{2}(x_{L} + x_{L+1}) = 0.5(46 + 47) = 46.5 \); if \( L \) had not been an integer like 15, the rule would be to round up to next largest integer, \( \pi_{P} = x_{L} \)

quantile(data,c(0.25,0.50,0.75),type=2)

25% 50% 75%
41.0 43.5 46.5

(o) percentile rank of age 47 is

(i) 70%  (ii) 75%  (iii) 80%  (iv) 85%

perc.rank <- function(x, xo) length(x[x < xo])/length(x)*100
perc.rank(data,47)

[1] 75

(p) The five-number summary (minimum, first quartile, second quartile, third quartile, maximum) and percentile rank calculated for this group of patients are all values of different

(i) statistics  (ii) parameters used to estimate corresponding population

4.3 Maximum Likelihood Estimates

One important aspect of statistics is to do with the idea of gathering together a random sample of data, calculating a statistic and using this statistic to infer something about, typically, a parameter of the population, more specifically a parameter from a probability distribution model describing the population, from which this sample is taken.

So a question is: when does a statistic “fit” the parameter, when is a statistic a “good estimate” of the parameter? One criterion, called unbiasedness, assesses whether or not a statistic is an appropriate fit for any given parameter from a distribution function. A random variable \( \Theta \) whose values estimate parameter \( \theta \) is called an estimator or \( \theta \). A value of \( \Theta \), \( \hat{\theta} \), is called an estimate of \( \theta \) and \( \Theta \) is an unbiased estimator of \( \theta \) if

\[
E(\Theta) = \theta,
\]
and a biased estimator of $\theta$ otherwise.

And another question is: where do statistics come from, how are they created? One method, called the maximum likelihood estimation (MLE) method, is described in this section on how to create possible statistics which estimate parameters from distribution functions. The MLE of $\theta$, denoted $\hat{\theta}$, is the value at which $L(\theta)$ is a maximum,

$$L\left(\hat{\theta}\right) \geq L\left(\theta\right),$$

for all $\theta$. It is often easier to work with the natural log of $L(\theta)$, $\ln L(\theta)$, rather than $L(\theta)$ itself. In general, let $X$ be a random variable with pdf (pmf) $f(x; \theta)$ where $\theta$ is an unknown parameter to be estimated and let $x_1, x_2, \ldots, x_n$ be $n$ random and independent observed values of $X$. The likelihood function is the product of $f(x; \theta)$ evaluated at each observed value,

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

Exercise 4.3 (Maximum Likelihood Estimates)

1. **Density, $f(x)$, versus likelihood, $L(\theta)$: Geometric.**

   The geometric density is
   
   (i) $f(x) = p(1 - p)^{x-1}$, $x = 1, 2, \ldots$
   
   (ii) $L(p) = p(1 - p)^{x-1}$, $0 \leq p \leq 1$

   given in figure (a) where $p = 0.7$ (b) where $x = 3$

   whereas the geometric likelihood is
   
   (i) $f(x) = p(1 - p)^{x-1}$, $x = 1, 2, \ldots$
   
   (ii) $L(p) = p(1 - p)^{x-1}$, $0 \leq p \leq 1$

   given in figure (a) where $p = 0.7$ (b) where $x = 3$

   with maximum value at $\hat{p} \approx$ (i) 0.25 (ii) 0.35 (iii) 0.45

![Figure 4.3: Geometric density and likelihood](image)
2. **Density, \( f(x) \), versus likelihood, \( L(\theta) \): Binomial.**

The binomial density is

(i) \( f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \ x = 0, 1, \ldots, n \)

(ii) \( L(p) = \binom{n}{x} p^x (1-p)^{n-x}, \ 0 \leq p \leq 1 \)

given in figure (a) where \( n = 5, p = 0.7 \) (b) where \( n = 5, x = 3 \)

whereas the binomial likelihood is

(i) \( f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \ x = 0, 1, \ldots, n \)

(ii) \( L(p) = \binom{n}{x} p^x (1-p)^{n-x}, \ 0 \leq p \leq 1 \)

given in figure (a) where \( n = 5, p = 0.7 \) (b) where \( n = 5, x = 3 \)

with maximum value at \( \hat{p} \approx (i) 0.25 \) (ii) 0.35 (iii) 0.45

![Figure 4.4: Binomial density and likelihood](image)

3. **Density, \( f(x) \), versus likelihood, \( L(\theta) \): Normal.**

The normal density is

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

(ii) \( L(\mu) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(1/2)\left[(x-\mu)/\sigma\right]^2}, \ -\infty < \mu < \infty \)

(iii) \( L(\sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(1/2)\left[(x-\mu)/\sigma\right]^2}, \ \mu > 0 \)

given in figure

(a) where \( \mu = 167, \sigma = 20.1 \)
(b) where \( x = 150, \sigma = 20.1 \)
(c) where \( x = 150, \mu = 167 \)

whereas the normal likelihood for \( \mu \) is

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

(ii) \( L(\mu) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(1/2)\left[(x-\mu)/\sigma\right]^2}, \ -\infty < \mu < \infty \)
(iii) \( L(\sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(1/2)((x-\mu)/\sigma)^2}, \mu > 0 \)
given in figure
(a) where \( \mu = 167, \sigma = 20.1 \)
(b) where \( x = 150, \sigma = 20.1 \)
(c) where \( x = 150, \mu = 167 \)
with maximum value at \( \hat{\mu} \approx (i) 100 \) (ii) 150 (iii) 200

and the normal likelihood for \( \sigma \) is
(i) \( f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(1/2)((x-\mu)/\sigma)^2}, -\infty < x < \infty \)
(ii) \( L(\mu) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(1/2)((x-\mu)/\sigma)^2}, -\infty < \mu < \infty \)
(iii) \( L(\sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(1/2)((x-\mu)/\sigma)^2}, \mu > 0 \)
given in figure
(a) where \( \mu = 167, \sigma = 20.1 \)
(b) where \( x = 150, \sigma = 20.1 \)
(c) where \( x = 150, \mu = 167 \)
with maximum value at \( \hat{\sigma} \approx (i) 15 \) (ii) 19 (iii) 21

4. Credit cards: statistic \( \hat{p} \) estimator of parameter \( p \).
Since 54 of 180 randomly selected from all credit card purchase slips are made
with Visa, point estimate \( \hat{p} = \frac{54}{180} \) might be used to estimate true (actual,
population) parameter \( p \). Let random variable \( X \) represent the number of
purchase slips made with Visa and parameter \( p \) be the probability of a purchase
slip made with Visa.

An appropriate probability mass function to model \( X \) is the
(i) binomial pmf:
\[
f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n,
\]
where \( n = 180 \) trials.

(ii) geometric pmf:
\[
f(x) = p(1-p)^{x-1}, \quad x = 1, 2, \ldots
\]

(iii) negative binomial pmf:
\[
f(x) = \binom{x-1}{r-1} p^r q^{x-r}, \quad x = r, r+1, \ldots
\]
where \( r = 180 \) trials.

The point estimate of \( p \), \( \hat{p} = \frac{1}{180} \sum_{i=1}^{180} x_i = (i) 0.3 \) (ii) 54 (iii) 180, where \( x_i = 1, 0 \) if purchase slip \( i \) made with Visa or not.

Statistic \( \hat{p} \) is an (i) estimator (ii) point estimate of parameter \( p \).

Statistic \( \hat{p} \) is a sensible estimator of \( p \), \( \hat{p} \) is (i) unbiased (ii) biased,
\[
E[\hat{p}] = p,
\]
whereas, for example, \( 1 - \hat{p} \) is a poor estimate of \( p \) because \( 1 - \hat{p} \) is biased,
\[
E[1 - \hat{p}] = 1 - p \neq p.
\]

5. Average student weight: statistic \( \bar{x} \) estimator of parameter \( \mu \).

Average weight of simple random sample of 11 PNW students point estimate \( \bar{x} = 167 \) pounds (\( s = 20.1 \) pounds) might be used to estimate true (actual, population) parameter \( \mu \). Let random variable \( X \) represent the weight of a PNW student and parameter \( \mu \) represent the mean weight of PNW students.

It is (i) sensible (ii) unsensible to model \( X \) with the normal pdf:
\[
f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2}[(x-\mu)/\sigma]^2}.
\]

The point estimate of \( \mu \), \( \bar{x} = \frac{1}{11} \sum_{i=1}^{11} x_i = (i) 11 \) (ii) 20.1 (iii) 167 where \( x_i \) is weight of student \( i \).

Statistic \( \bar{x} \) is an (i) estimator (ii) point estimate of parameter \( \mu \).

Statistic \( \bar{x} \) is a sensible estimator of \( \mu \), \( \bar{x} \) is (i) unbiased (ii) biased,
\[
E[\bar{X}] = \mu,
\]
whereas, for example, \( \bar{x} + 1 \) is a poor estimate of \( \mu \) because \( \bar{x} + 1 \) is biased,
\[
E[\bar{X} + 1] = \mu + 1 \neq \mu.
\]
6. Of the two possible estimators (statistics)

\[ S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \quad \text{or} \quad \hat{\Sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2, \]

the statistic \( S^2 \) is generally considered a (i) superior (ii) inferior estimator to the statistic \( \hat{\Sigma}^2 \) for \( \sigma^2 \) from the normal distribution because \( S^2 \) is unbiased whereas \( \hat{\Sigma}^2 \) is biased,

\[ E[S^2] = \sigma^2 \quad \text{but} \quad E[\hat{\Sigma}^2] \neq \sigma^2. \]

7. Geometric distribution, special case. A discrete random variable \( X \) is said to have a geometric distribution if its pmf is

\[ f(x; p) = p(1 - p)^{x-1}, \quad x = 1, 2, \ldots \]

where \( 0 < p < 1 \) is the parameter. If the values \( x_1 = 2, x_2 = 5, \) and \( x_3 = 7 \) are observed, find the (i) likelihood function of \( p, L(p), \) and (ii) MLE of \( p, \hat{p}. \)

![Figure 4.6: Likelihood and ln-Likelihood of Geometric, \( x = (2,5,7) \)](image)

(a) likelihood function, \( L(p). \)

\[ L(p) = f(2) \cdot f(5) \cdot f(7) = p(1 - p)^{2-1} \cdot p(1 - p)^{5-1} \cdot p(1 - p)^{7-1} = \]

(i) \( p^2(1 - p)^{11} \) (ii) \( p^3(1 - p)^{10} \) (iii) \( p^3(1 - p)^{11} \) (iv) \( p^2(1 - p)^{10} \)

(b) maximum likelihood estimate, \( \hat{p}, \) version 1. Since

\[ L'(p) = \frac{dL}{dp} = 3p^2(1-p)^{11} + p^3(11)(1-p)^{10}(-1) = p^2(1-p)^{10}(3-14p) = 0 \]

then \( p = 0, \frac{3}{14} \) or 1, but since \( L(0) = L(1) = 0, \) MLE is \( \hat{p} = \frac{3}{14}. \) Notice \( \frac{3}{14} \approx 0.21, \) which is maximum of plot of likelihood above.
(c) **maximum likelihood estimate, \( \hat{p} \), version 2.** Since
\[
L(p) = p^3(1-p)^{11}
\]
then
\[
\ln L(p) = \ln \left( p^3(1-p)^{11} \right) = 3 \ln p + 11 \ln(1-p)
\]
so
\[
\frac{d \ln L}{dp} = \frac{3}{p} + \frac{11}{1-p} = 0,
\]
then MLE is \( \hat{p} = (i) \frac{2}{14}, \) \( \frac{3}{12}, \) \( \frac{3}{14}\) \( \frac{3}{11} \)
Notice again \( \frac{3}{14} \approx 0.21\), which is maximum of plot of ln-likelihood above.

(d) The MLE is that value of the parameter which maximizes the likelihood (parameter version of distribution) for observed sample of data.

(i) **True**  \( \) (ii) **False**

8. **Geometric distribution, general case.** A discrete random variable \( X \) with geometric distribution has \( f(x; p) = p(1-p)^{x-1}, \) \( x = 1, 2, \ldots \) where \( 0 < p < 1 \) is the parameter. For \( x_1, x_2, \ldots, x_n \), find the (i) \( \ln L(p) \) (ii) MLE \( \hat{p} \) and (iii) whether or not \( \hat{p} \) is unbiased. Recall, mean of geometric is \( E(X) = E(X_i) = \frac{1}{p} \).

(a) **\( L(p) \).**
\[
L(p) = \prod_{i=1}^{n} p(1-p)^{x_i-1} = \]
(i) \( p^n(1-p)^{\sum x_i-n} \) (ii) \( p^{nx}(1-p)^{n-\sum x_i} \) (iii) \( p^{nx}(1-p)^{n-nx} \)

(b) **\( \ln L(p) \).**
\[
\ln L(p) = \ln \left( p^n(1-p)^{\sum x_i-n} \right) = \]
(i) \( p^{\sum x_i}(1-p)^{n-\sum x_i} \) (ii) \( n \ln p + (\sum x_i - n) \ln(1-p) \)

(c) **\( \hat{p} \).** Since
\[
\frac{d \ln L}{dp} = \frac{n}{p} + \frac{1}{1-p} \frac{\sum x_i - n}{1-p} = 0,
\]
\( \hat{p} = (i) \frac{n}{p}, \) \( \frac{n}{\sum x_i}, \) \( \frac{p}{\sum x_i}, \) \( \frac{\sum x_i}{n} \)
Notice \( \frac{\sum x_i}{n} = \frac{3}{3+5+7} = \frac{1}{14}, \) if \( n = 3, x = (2, 5, 7) \), which matches previous special case result.

(d) **unbiased?** First notice
\[
E(X) = E \left( \frac{1}{n} \sum X_i \right) = \frac{1}{n} \cdot n \cdot E(X) = \frac{1}{n} \cdot n \cdot \frac{1}{p} = \frac{1}{p},
\]
so, since \( \hat{p} = \frac{n}{\sum x_i} = \frac{1}{X} \) is convex, Jensen’s inequality tells us
\[
E(\hat{p}) = E \left( \frac{1}{X} \right) > \frac{1}{E(X)} = \frac{1}{1/p} = p
\]
and so \( \hat{p} \) is (i) **unbiased** (ii) **biased**
(e) asymptotically unbiased? In fact it can be shown

$$E(\hat{p}) = p + O\left(\frac{1}{n}\right),$$

where

$$O\left(\frac{1}{n}\right) \to 0, \text{ as } n \to \infty,$$

so \(\hat{p}\) is (i) asymptotically unbiased (ii) asymptotically biased, becomes more unbiased as sample size \(n\) increases.

### 4.4 Sampling Distributions

A sampling distribution is a probability distribution of a statistic. According to the central limit theorem (CLT), if \(X_1, X_2, \ldots, X_n\) are mutually independent random variables where each which common \(\mu\) and \(\sigma^2\), then as \(n \to \infty\),

$$\bar{X}_n \to N\left(\mu, \frac{\sigma^2}{n}\right),$$

no matter what the distribution of the population. Often \(n \geq 30\) is “large enough” for the CLT to apply. So, for large enough \(n\), \(\bar{X}_n\) is said to have a normal sampling distribution. In fact, the CLT applies to statistics other than just \(\bar{X}\). Specifically, let \(X\) be \(b(n, p)\), then as \(n \to \infty\), the sampling distribution of

$$\hat{P} = \frac{X}{n} \to N\left(p, \frac{p(1-p)}{n}\right),$$

where the approximation is reasonable if \(np \geq 5\), \(n(1-p) \geq 5\).

Not all sampling distributions of statistics are normal; for example, the sampling distribution of the sample variance,

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2,$$

has a chi-square, not normal, distribution, but, even in this case, as \(n \to \infty\), this chi-square distribution tends to a normal. We simulate (numerical approximate) the sampling distributions of a couple of statistics to demonstrate an important property; namely,

- sampling (experimental) error: unavoidable differences between distribution parameter and sample statistic due to random fluctuations.
Nonsampling errors, on the other hand, are differences due to improperly recorded collected, recorded or analyzed data.

Exercise 4.4 (Sampling Distributions)

1. CLT and proportion: chance lawyer wins.

Lawyer estimates she wins 40% of her cases \((p = 0.4)\), and currently represents \(n = 50\) defendants. Let \(X\) represent number of wins (of 50 cases) and so \(\hat{p} = \frac{X}{n}\) proportion of wins (of 50 cases). Use CLT to approximate chance she wins at least one–half of her cases,

\[
P \left( \hat{p} > \frac{1}{2} \right) = P (\hat{p} > 0.5).
\]

(a) Check assumptions. Since \(np = 50(0.4) = 20 \geq 5\) and \(np(1-p) = 50(0.4)(1-0.4) = 12 \geq 5\), assumptions necessary for approximation are (i) satisfied (ii) violated.

(b) \(\mu_{\hat{p}} = p = (i)\ 0.3 \ (ii)\ 0.4 \ (iii)\ 0.5\).

(c) \(\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.4(1-0.4)}{50}} \approx (i)\ 0.0012 \ (ii)\ 0.0385 \ (iii)\ 0.06928\).

(d) \(P (\hat{p} > 0.5) \approx 0.07 \ (ii)\ 0.11 \ (ii)\ 0.13\).

(e) Since \(P (\hat{p} > 0.5) \approx 0.07 > 0.05\), \(\hat{p} = 0.5\) is (i) typical (ii) unusual.

Anything greater than 0.05 is usually (arbitrarily) considered “typical”.

2. Simulating sample means: game payoffs. Video game payoff points, \(X\), are uniform between -3 and 3 points, \(U(-3, 3)\), where \(\mu = \frac{3+(-3)}{2} = 0\), \(\sigma^2 = \frac{(3-(-3))^2}{12} = 3\). The average and variance of 5, 10 and 20 plays of this game are simulated (numerically approximated) 1000 times and the results are given in the graphs and table.

<table>
<thead>
<tr>
<th>(n)</th>
<th>(\bar{x}_n)</th>
<th>(s^2_n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.041</td>
<td>3.048</td>
</tr>
<tr>
<td>5</td>
<td>0.038</td>
<td>0.588 (\approx \frac{s^2}{n}) = (\frac{3.048}{5})</td>
</tr>
<tr>
<td>10</td>
<td>0.002</td>
<td>0.297 (\approx \frac{s^2}{n}) = (\frac{3.048}{10})</td>
</tr>
<tr>
<td>20</td>
<td>-0.008</td>
<td>0.146 (\approx \frac{s^2}{n}) = (\frac{3.048}{20})</td>
</tr>
</tbody>
</table>

(a) The \(\bar{x}\), for 1000 simulations, (i) decreases (ii) remains the same (iii) increase as \(n\) increases.
(b) The average of $s^2$, for 1000 simulations,
   (i) decreases  (ii) remains the same  (iii) increase as $n$ increases
   and is approximately equal to $\frac{\sigma^2}{n}$.

(c) Observed values of $\bar{x}_n$ are closer to $\mu = 0$ for (i) smaller $n$  (ii) larger $n$.

(d) Observed values of $\bar{x}_n$ are closer to $\mu = 0$ for
   (i) smaller  (ii) larger $\sigma^2 \approx s^2$.

3. Simulating unbiased $\bar{x}$, $s^2$ and biased $s$. Let $X$ be payoff of game with values
$2, 5$ and $8$ with equal probabilities of $f(x) = \frac{1}{3}$. The mean $\mu$, variance $\sigma^2$
and SD $\sigma$ are
\[
\mu = \sum f(x)x = \frac{1}{3} (2 + 5 + 8) = 5,
\]
\[
\sigma^2 = \sum f(x)(x - \mu)^2 = \frac{1}{3} \left[[2 - 5] + (5 - 5)^2 + (8 - 5)^2]\right] = 6
\]
\[
\sigma = \sqrt{6}.
\]

All possible small samples of size $n = 2$ are chosen (without replacement) from
the distribution of $X$ with results given in the following table.
When comparing simulated parameters with actual parameters,

(a) The sample \( \bar{x} = 5.0 \), for 9 simulations, is
   (i) less than (ii) equal to (iii) greater than
   \( \mu = 5 \), so \( \bar{x} \) is unbiased.

(b) The sample \( s^2 = 6.0 \), for 9 simulations, is
   (i) less than (ii) equal to (iii) greater than
   \( \sigma^2 = 6 \), so \( s^2 \) is unbiased.

(c) The sample \( s \approx 1.89 \), for 9 simulations, is
   (i) less than (ii) equal to (iii) greater than
   \( \sigma = \sqrt{6} \approx 2.45 \), so \( s \) is biased.

4. Simulating sample variances: temperatures. Temperature \( X \) is \( N(0,1) \), where \( \mu = 0 \) and \( \sigma = 1 \), on any given winter day. The variance of 5, 10 and 100 winter days are simulated (numerically approximated) 1000 times and results are given in the red graphs below. The shape of the histogram for the sample variance, at least to begin with, for \( n = 5,10 \) (i) is (ii) is not normal, but when \( n = 100 \) becomes (i) more (ii) less normal-shaped.
Figure 4.8: Sampling distributions of $s^2$ for temperature