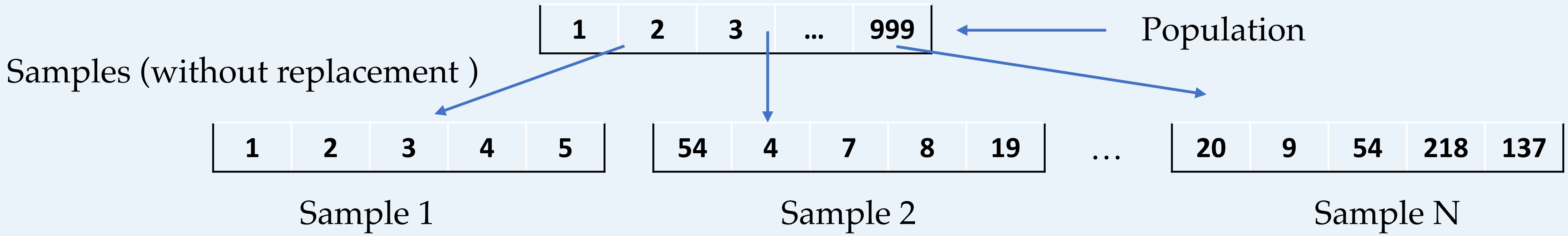
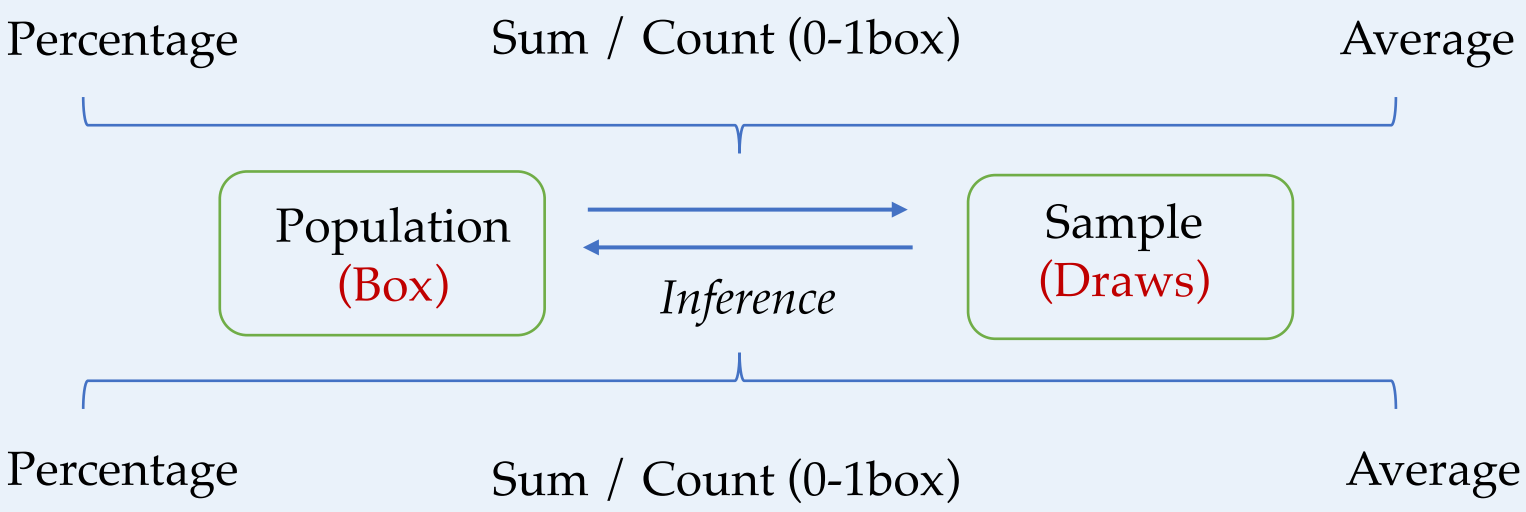
SECTION 6

Chapter 19, 20, 21 and 23

	Sample		Population
Relationship	Sample is the part of the population		
Values	Can be different for each draws		Fixed
2 main ways to sample	With replacement	Without replacement (a simple random sampling)	
Measurement	Statistic		Parameter
Chance Error	Sum in sample = sum in population + chance error Percentage in sample = percentage in population + chance error Average in sample = average in population + chance error Chance error is different from sample to sample SE measures the likely size of the chance error		




Percentage
Sum / Count (0-1box)
Average
Population
(Box)
Sample
(Draws)
Inference
Percentage
Sum / Count (0-1box)
Average

Percentage

Population (Box) → Sample (Draws)

Size of sample = Number of draws

Eg: 0-1 box model:

1. The expected value for percentage of 1's **in the sample** = (is exactly equal to) Percentage of 1's **in the box** $= \dfrac{\text{number of 1's in the population}}{\text{size of population}} \times 100\%$

---

2. The SE for percentage of 1's **in the sample**

(SE is always for sample)

With replacement

$$\frac{SE\ \text{for the number (count)}}{\text{size of sample}} \times 100\% = \frac{SD \times \sqrt{n}}{n} \times 100\% = \frac{SD}{\sqrt{n}} \times 100\% =$$

$$\frac{\sqrt{\text{Percentage of 1's in the box} \times \text{Percentage of 0's in the box}}}{\sqrt{n}} \times 100\%$$

Note:

- When n (size of the draws) ↑, SE for the number ↑, SE for the percentage ↓
- When n_new = a × n

  SE_new for the number = $\sqrt{a}$ × SE for the number

  SE_new for the percentage = SE for the percentage / $\sqrt{a}$
- When estimating percentage, the accuracy is determined by **the absolute value of size of sample**, but not the size relative to the population (when the size of population is large relative to the size of sample, correlation factor is close to 1, which can be ignored)

Without replacement

Correction factor × SE when drawing with replacement

Correction factor $= \sqrt{\dfrac{\text{size of box} - \text{size of sample}}{\text{size of box} - 1}}$

---

3. The percentage of 1's **in the sample** will be around **the expected value**, give or take the **SE** or so.

---

4. The chance (approximately) that the percentage of 1's in the sample will be in some range

Percentage a — Exp — Percentage b $\xrightarrow[\text{Transfer to Unit}]{\text{Normal Approx}}$ Unit a' — 0 — Unit b'

---

5. The probability histogram for the percent of 1's **in the sample** $\xrightarrow{\text{Area}}$ Get the chance of the percentage of 1's in the sample for some ranges

Percentage

Eg: 0-1 box model:

Population (Box) $\xleftarrow{\textit{Inference}}$ Sample (Draws)

1 The estimated SD of the box $\underset{\text{Bootstrap}}{=}$ $\sqrt{\text{Percentage of 1's in the sample} \times \text{Percentage of 0's in the sample}}$

(Bootstrap is a good estimate when sample size is large)

2 The estimated SE for percentage of 1's **in the sample** (See the last slide, substitute the SD to the estimated SD of the box)

3 The estimated percentage of 1's **in the population** is $\boxed{\frac{\textit{number of 1 in the sample}}{\textit{size of sample}} \times 100\%}$, and this estimate is likely to be off by **the estimated SE** or so.

Sample percentage

4 The 95%-confidence interval for the percentage of 1's **in the population** Sample percentage ± 2 SE or [Sample percentage – 2 SE, Sample percentage – 2 SE]

Note:

- The confidence interval is derived from the sample, but aims to estimate the population percentage in a range (interval)
- The 95%-confidence interval has 95% of chance to cover the population percentage (the true value)
- Should only be used with large samples, the confidence level is read off the normal curve

5 The above formulas apply to simple random samples, but may not apply to other kinds of samples

Averages

Population (Box) → Sample (Draws)

Eg: consider a box model:

1. The expected value for the average of **the draws** = (<u>is exactly equal to</u>) Average of **the box**

---

2. The SE for the average of **the draws**

<u>(SE is always for sample)</u>

- With replacement: $\frac{SE\ for\ the\ sum}{number\ of\ draws} = \frac{SD \times \sqrt{n}}{n} = \frac{SD}{\sqrt{n}}$

  Note:
  - When n (size of the draws) ↑, SE for the sum ↑, SE for the average ↓
  - When n_new = a × n

    SE_new for the sum = $\sqrt{a}$ × SE for the sum

    SE_new for the average = SE for the average / $\sqrt{a}$

- Without replacement: Correction factor × SE when drawing with replacement

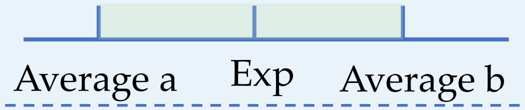
  Correction factor $= \sqrt{\frac{size\ of\ box\ - number\ of\ draws}{size\ of\ box\ - 1}}$

---

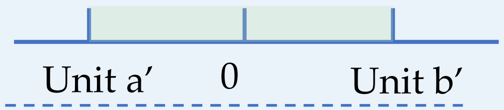
3. The average of **the draws** will be around **<u>the expected value</u>**, give or take the **<u>SE</u>** or so.

---

4. The chance (approximately) that the average of the draws will be in some range



$\xrightarrow[\text{Transfer to Unit}]{\text{Normal Approx}}$



---

5. The probability histogram for the average of **the draws** $\xrightarrow{\text{Area}}$ Get the chance of the average of the draws for some ranges

Averages

Eg: consider a box model:

Population (Box) ← *Inference* — Sample (Draws)

1. The estimated SD of the box = The SD of the sample (The estimate is good when sample size is large)

2. The estimated SE for the average of **the draws** (See the last slide, substitute the SD to the estimated SD of the box)

3. The estimated average of **the population** is Average of sample, and this estimate is likely to be off by **the estimated SE** or so.

4. The 95%-confidence interval for the average **of the population** Sample average $\pm$ 2 SE or [Sample average – 2 SE, Sample average – 2 SE]

Note:
- The confidence interval is derived from the sample, but aims to estimate the population average in a range (interval)
- The 95%-confidence interval has 95% of chance to cover the population average (the true value)
- Should only be used with large samples, the confidence level is read off the normal curve

5. The above formulas apply to simple random samples, but may not apply to other kinds of samples

Exercises:

Chapter 20

2. A university has 25,000 students, of whom 10,000 are older than 25. The registrar draws a simple random sample of 400 students.
   (a) Find the expected value and SE for the number of students in the sample who are older than 25.
   (b) Find the expected value and SE for the percentage of students in the sample who are older than 25.
   (c) The percentage of students in the sample who are older than 25 will be around ______, give or take ______ or so.

2. In a certain town, there are 30,000 registered voters, of whom 12,000 are Democrats. A survey organization is about to take a simple random sample of 1,000 registered voters.
   (a) The expected value for the percentage of Democrats in the sample is ______. The SE for the percentage of Democrats in the sample is ______.
   (b) The percentage of Democrats in the sample is likely to be around ______, give or take ______ or so.
   (c) Find the chance that between 39% and 41% of the registered voters in the sample are Democrats.

Chapter 20

5. A box contains 2 red marbles and 8 blue ones. Four marbles are drawn at random. Find the SE for the percentage of red marbles drawn, when the draws are made

    (a) with replacement. (b) without replacement.

6. The Census Bureau is planning to take a sample amounting to 1/10 of 1% of the population in each state in order to estimate the percentage of the population in that state earning over $100,000 a year. Other things being equal:

    (i) The accuracy to be expected in California (population 35 million) is about the same as the accuracy to be expected in Nevada (population 2 million).
    (ii) The accuracy to be expected in California is quite a bit higher than in Nevada.
    (iii) The accuracy to be expected in California is quite a bit lower than in Nevada.

    Explain.

Chapter 20

7. Five hundred draws are made at random from the box

$$\boxed{60{,}000\ \boxed{0}\text{'s}\quad 20{,}000\ \boxed{1}\text{'s}}$$

True or false, and explain:

(a) The expected value for the percentage of 1's among the draws is exactly 25%.
(b) The expected value for the percentage of 1's among the draws is around 25%, give or take 2% or so.
(c) The percentage of 1's among the draws will be around 25%, give or take 2% or so.
(d) The percentage of 1's among the draws will be exactly 25%.
(e) The percentage of 1's in the box is exactly 25%.
(f) The percentage of 1's in the box is around 25%, give or take 2% or so.

Chapter 21

2. The Residential Energy Consumption Survey found in 2001 that 47% of American households had internet access.[10] A market survey organization repeated this study in a certain town with 25,000 households, using a simple random sample of 500 households: 239 of the sample households had internet access.

   (a) The percentage of households in the town with internet access is estimated as ______; this estimate is likely to be off by ______ or so.
   (b) If possible, find a 95%-confidence interval for the percentage of all 25,000 households with internet access. If this is not possible, explain why not.

Chapter 21

5. A box contains a large number of red and blue marbles, but the proportions are unknown; 100 marbles are drawn at random, and 53 turn out to be red. Say whether each of the following statements is true or false, and explain briefly.
   - (a) The percentage of red marbles in the box can be estimated as 53%; the SE is 5%.
   - (b) The 5% measures the likely size of the chance error in the 53%.
   - (c) The 53% is likely to be off the percentage of red marbles in the box, by 5% or so.
   - (d) A 95%-confidence interval for the percentage of red marbles in the box is 43% to 63%.
   - (e) A 95%-confidence interval for the percentage of red marbles in the sample is 43% to 63%.