

PRASC



**Project for the Regional
Advancement of Statistics
in the Caribbean**

**Projet régional pour
l'avancement de la statistique
dans les Caraïbes**

Funded by the
Government
of Canada

Canada



Total Survey Error: Non-Sampling and Sampling Errors

Project for the Regional Advancement of Statistics in the Caribbean (PRASC)

Gavin Thompson
April 27, 2021, Virtual, from Ottawa, Canada



Delivering insight through data for a better Canada



Total Survey Error

- What is total survey error
- Why it is important
- How we report it
- How it is used for analysis

Overarching Processes

Specify needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Reuse or build collection instruments	4.1 Create frame and select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult and confirm needs	2.2 Design variable descriptions	3.2 Reuse or build processing and analysis components	4.2 Set up collection	5.2 Classify and code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collection	3.3 Reuse or build dissemination components	4.3 Run collection	5.3 Review and validate	6.3 Interpret and explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame and sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit and impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing and analysis	3.5 Test production systems		5.5 Derive new variables and units	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare and submit business case	2.6 Design production systems and workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production systems		5.7 Calculate aggregates			
				5.8 Finalise data files			

Measuring Error

- Error has an impact on the quality of the results being produced
- Reporting error allows users to judge the quality of your data
 - Be wary research results that don't share their errors and methods!
- Sampling error helps researchers understand the usability of your data for their research

Total Survey Error

Non-Sampling Errors

- All errors that can occur in all aspects of the survey process other than sampling (e.g. coverage, measurement, non-response, coding, data capture, editing, imputation).

Sampling Errors

- Arise from estimating a population characteristic by measuring only a portion of the population rather than the entire population.
- Measure the extent to which estimates from different possible samples of the same size and design differ from one another.

Total Survey Error

Non-Sampling Errors

- All errors that can occur in all aspects of the survey process other than sampling (e.g. coverage, measurement, non-response, coding, data capture, editing, imputation).

Sampling Errors

- Arise from estimating a population characteristic by measuring only a portion of the population rather than the entire population.
- Measure the extent to which estimates from different possible samples of the same size and design differ from one another.

~~~ Vote Gavin for Chief Statistician of the World! ~~~

# Non-Sampling Errors

**Non-sampling error comes from all sources**

**Planning phase:** coverage errors, slippage rates, screen designs

**Interviewing phase:** interviewer effect, non-response bias

**Processing phase:** edit failure rates, imputation rates, coding errors

## Reporting non-sampling error

- Allows users to understand what may have effected outcomes
- Allows for comparability with other data sources
- Increases trustworthiness of the data

# Non-Sampling Errors – Some Examples

## Type – Category - Example

Coverage errors - Frame defects – units left out, units included that shouldn't be

# Non-Sampling Errors – Some Examples

## Type – Category - Example

Coverage errors - Frame defects – units left out, units included that shouldn't be

Measurement errors – Poor questionnaire design – respondents giving incorrect answers

# Non-Sampling Errors – Some Examples

## Type – Category - Example

Coverage errors - Frame defects – units left out, units included that shouldn't be

Measurement errors – Poor questionnaire design – respondents giving incorrect answers

Interviewing errors – Poor training – Interviewer accidentally biases answers

# Non-Sampling Errors – Some Examples

## Type – Category - Example

Coverage errors - Frame defects – units left out, units included that shouldn't be

Measurement errors – Poor questionnaire design – respondents giving incorrect answers

Interviewing errors – Poor training – Interviewer accidentally biases answers

Non-response errors – Subpopulations with higher non-response – poor follow-up

# Non-Sampling Errors – Some Examples

## Type – Category - Example

Coverage errors - Frame defects – units left out, units included that shouldn't be

Measurement errors – Poor questionnaire design – respondents giving incorrect answers

Interviewing errors – Poor training – Interviewer accidentally biases answers

Non-response errors – Subpopulations with higher non-response – poor follow-up

Processing errors – Code changes answer – incorrect age ranges

# Non-Sampling Errors – Some Examples

## Type – Category - Example

Coverage errors - Frame defects – units left out, units included that shouldn't be

Measurement errors – Poor questionnaire design – respondents giving incorrect answers

Interviewing errors – Poor training – Interviewer accidentally biases answers

Non-response errors – Subpopulations with higher non-response – poor follow-up

Processing errors – Code changes answer – incorrect age ranges

Edit errors – Correct answers marked as incorrect – edit rules don't reflect reality

# Sampling Error

**Sampling error occurs when we sample from a population**

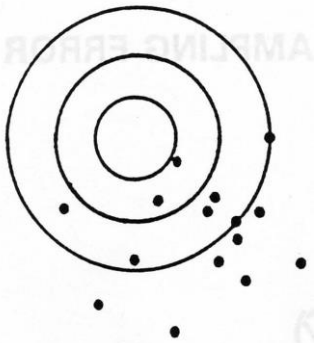
**Common measures of sampling error**

- Sampling variance
- Standard error
- Confidence interval
- Coefficient of variation

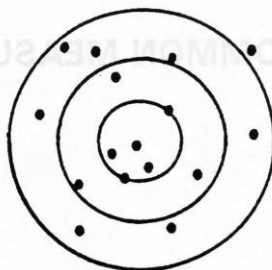
## Reporting Sampling error

- Allows users to understand how accurate the data is
- Allows for comparability with other data sources
- Gives data necessary for statistical testing

# Types of Sampling Results

**A**

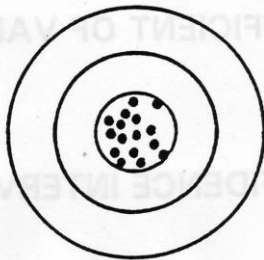
Imprecise and biased

**B**

Imprecise and unbiased

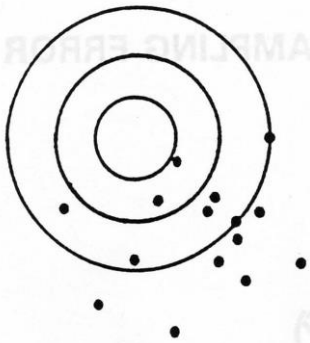
**C**

Precise and biased

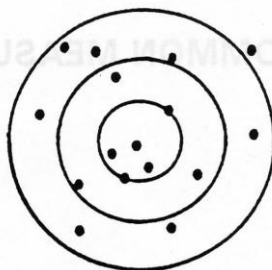
**D**

Precise and unbiased

# Types of Sampling Results

**A**

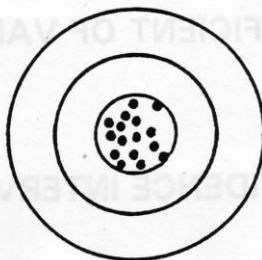
Imprecise and biased

**B**

Imprecise and unbiased

**C**

Precise and biased

**D**

Precise and unbiased

Unbiasedness is a necessary and sufficient condition for Accuracy

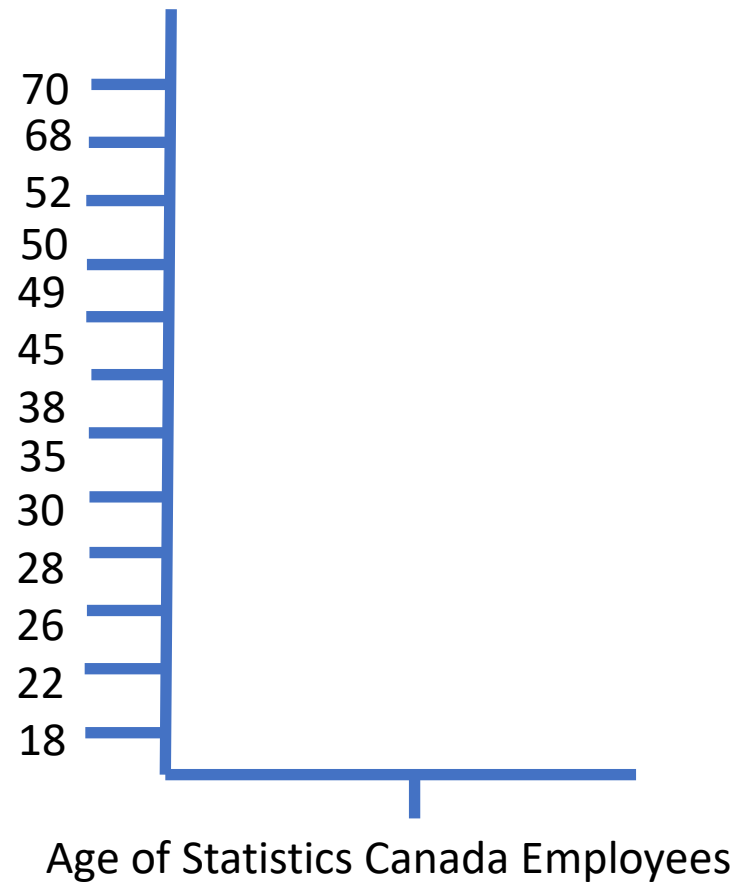
Both B and D are accurate because they give an unbiased estimate of the target value, although D could be described as “very accurate” and B could be described as “not very accurate”

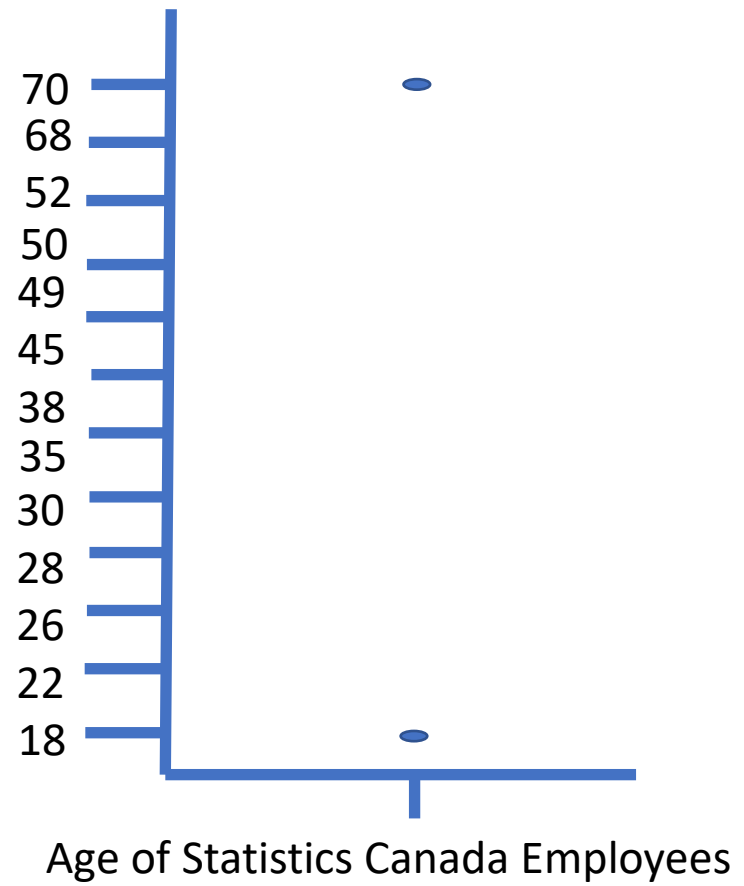
# Variance, Standard Deviation and Standard Error

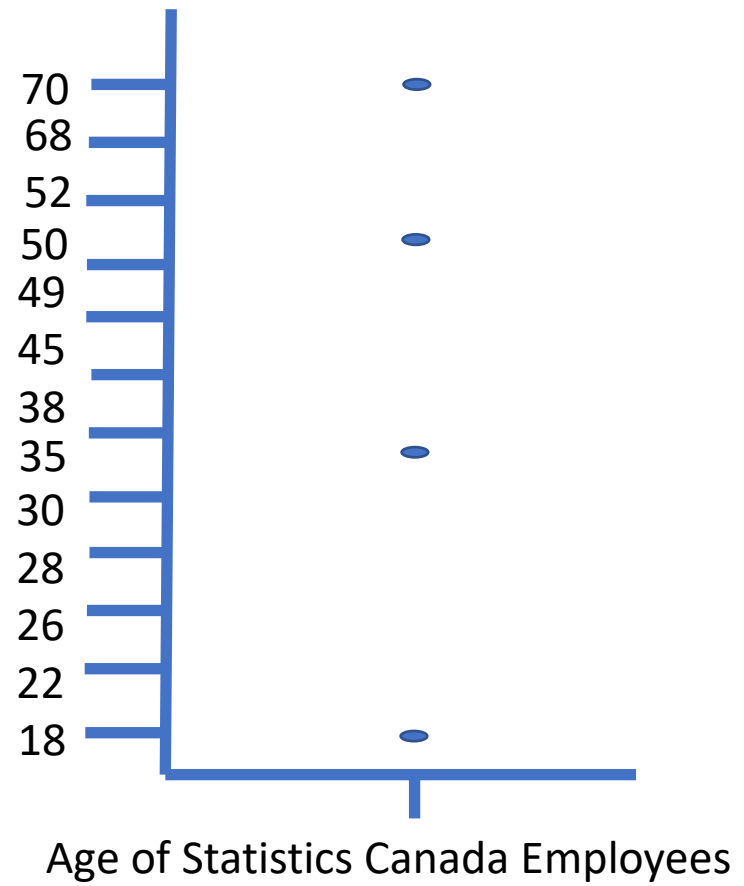
Suppose we are tasked with doing a study to find the average age at Statistics Canada

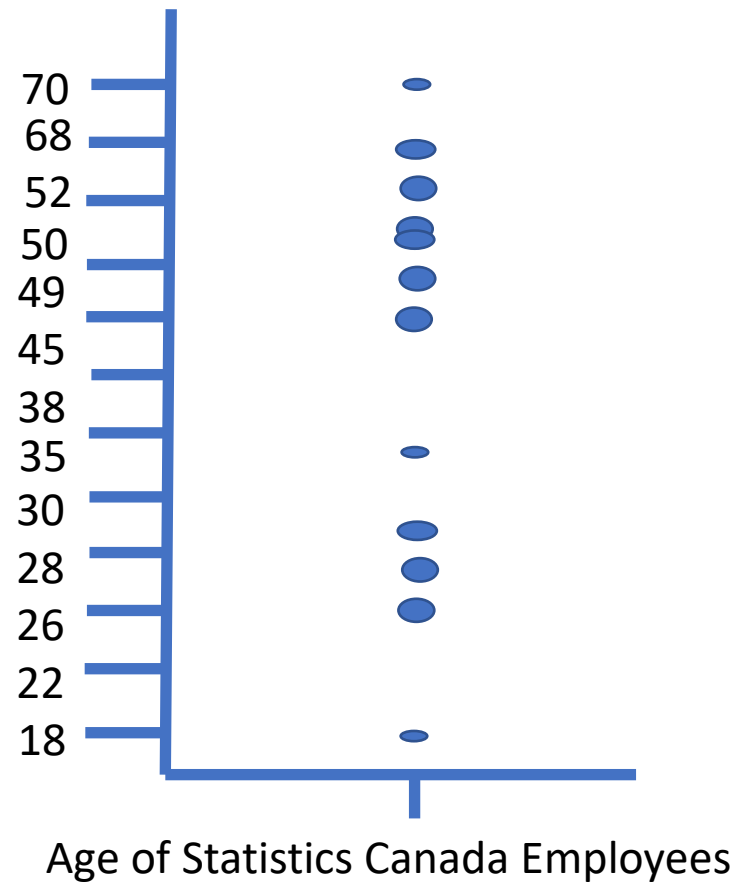
Everyone taking this course is asked to go and sample 10 people...and we repeat this a lot of times, say 500.

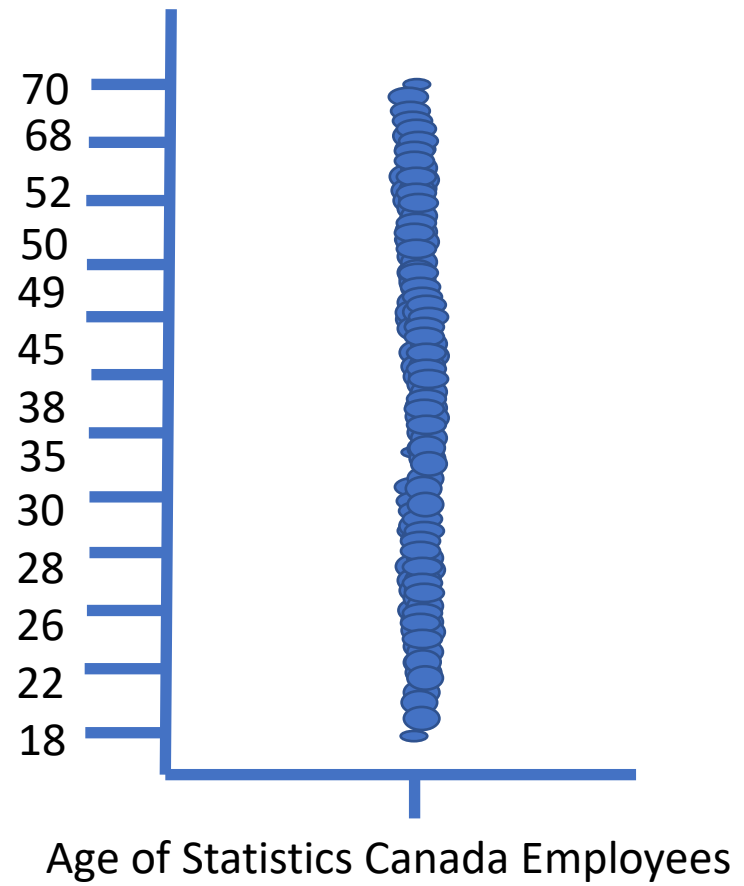
I'll be in charge of collecting the results and analyzing them.

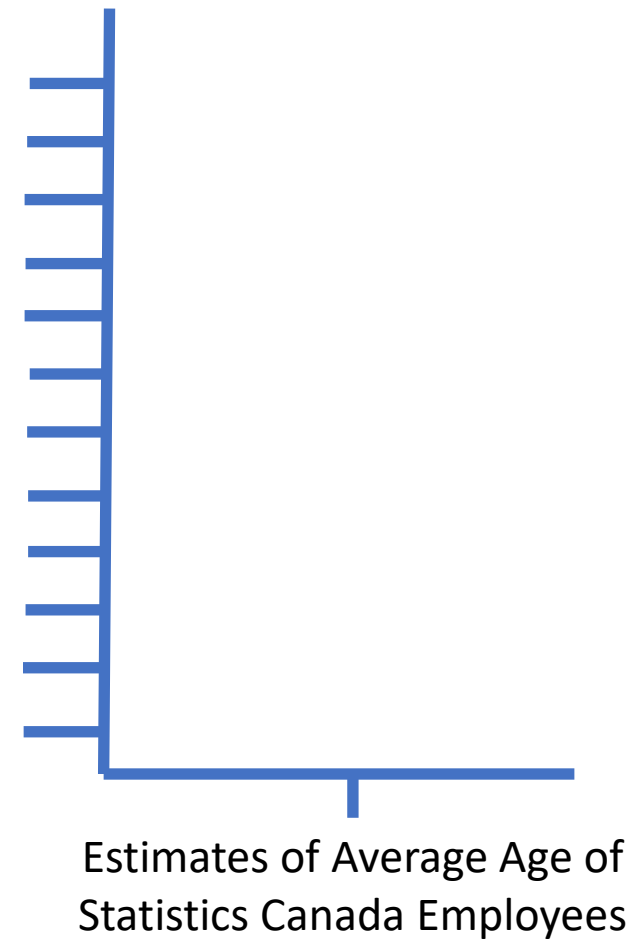
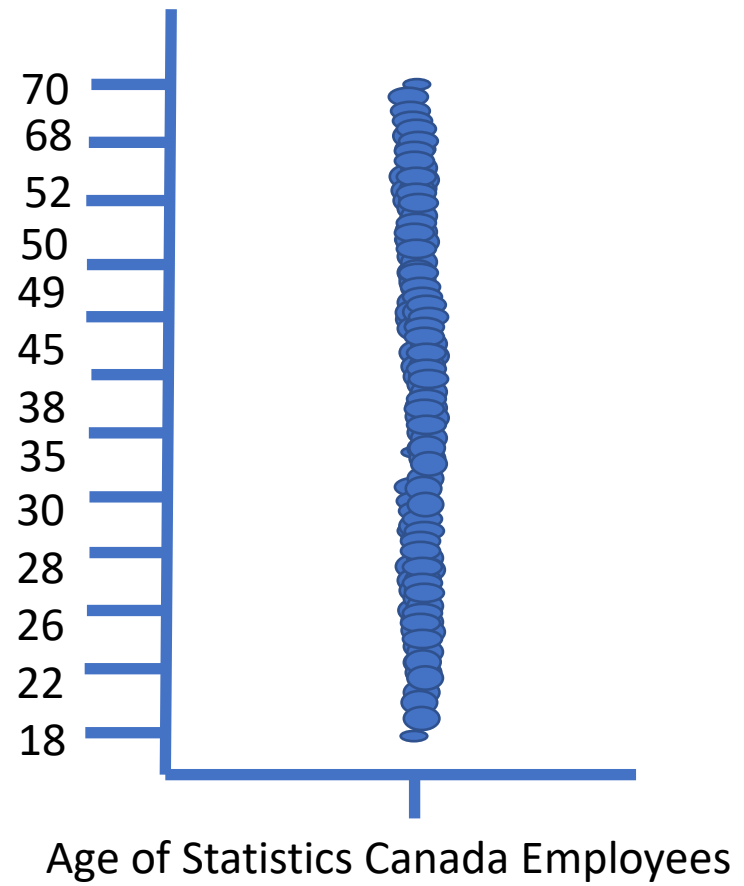


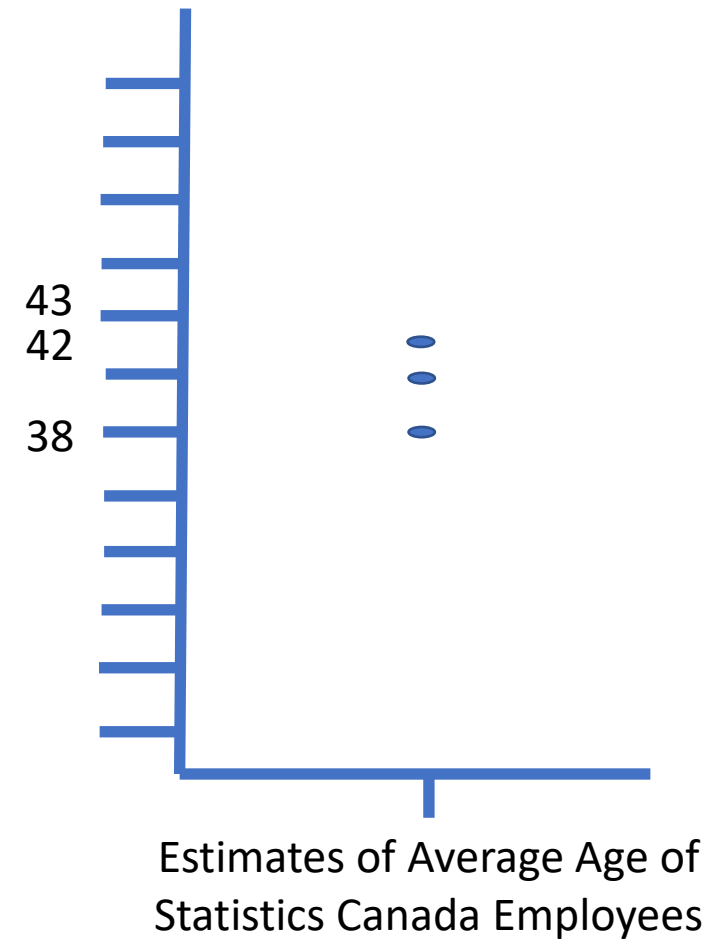
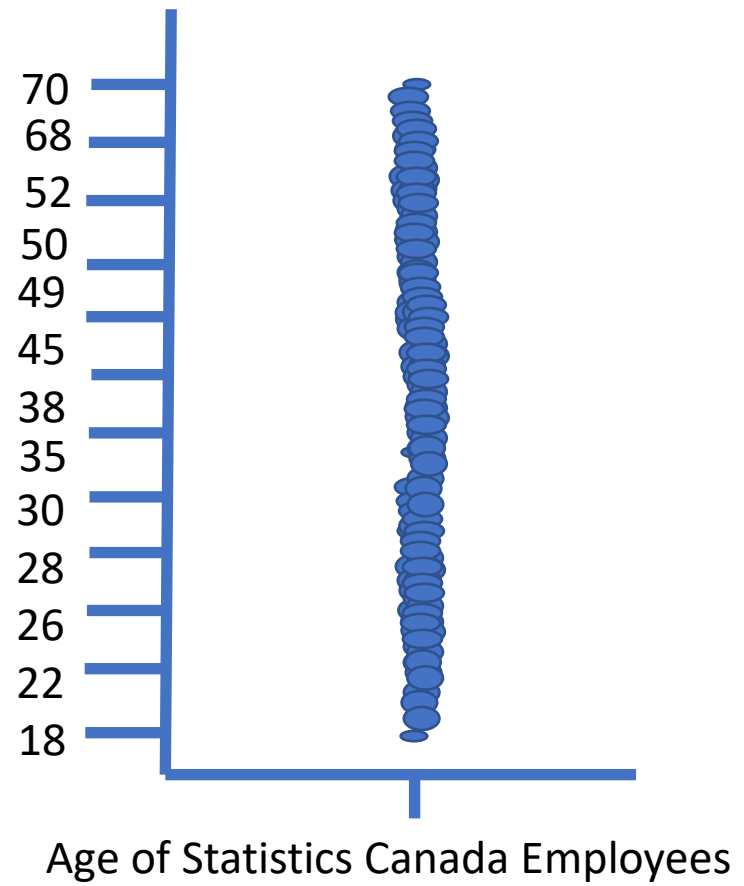


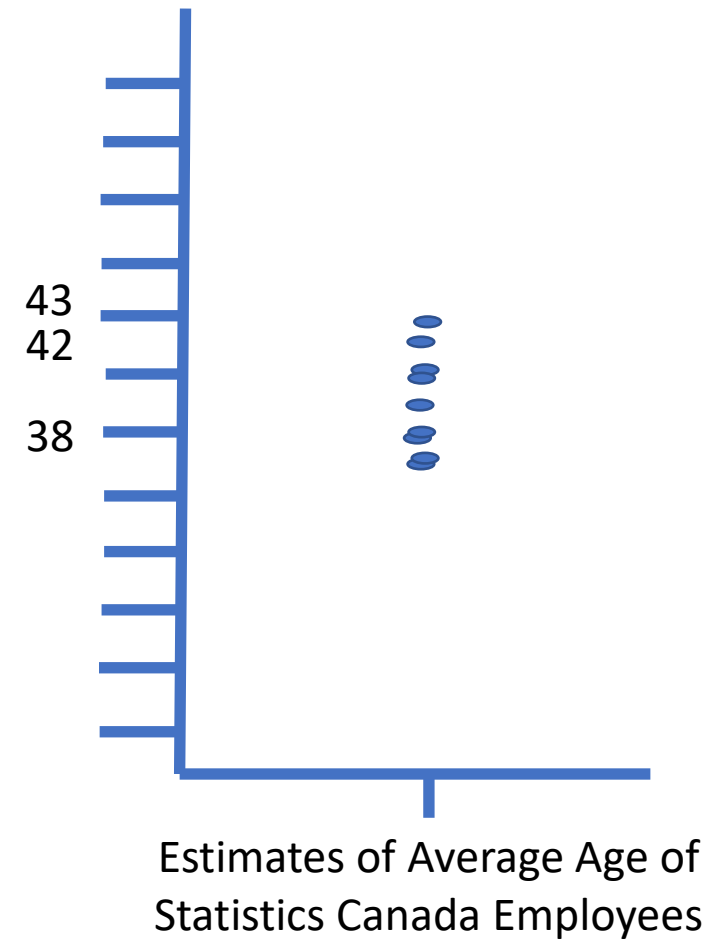
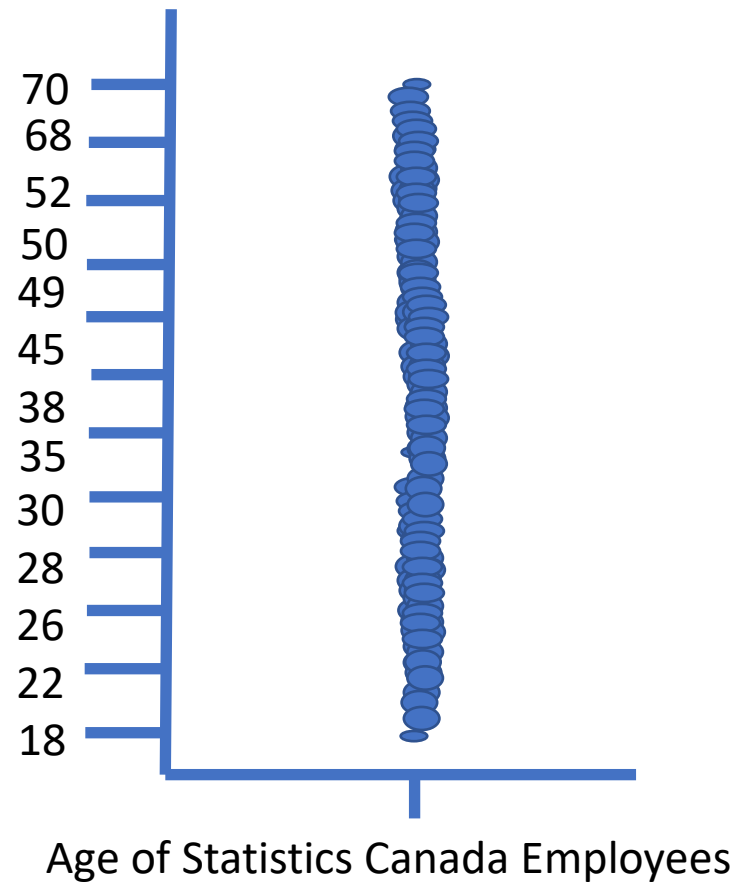


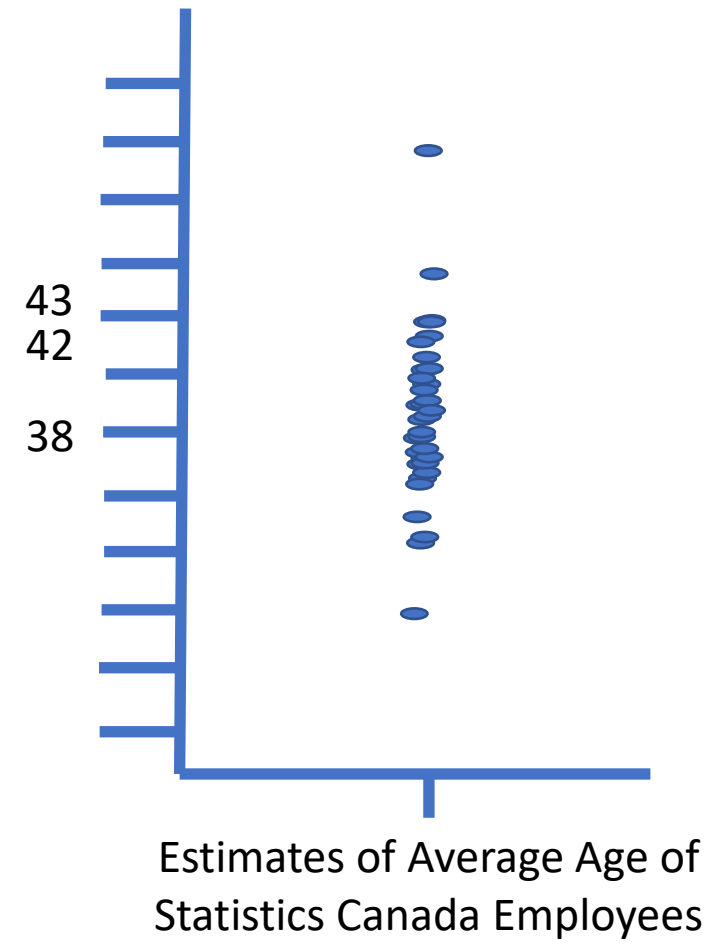
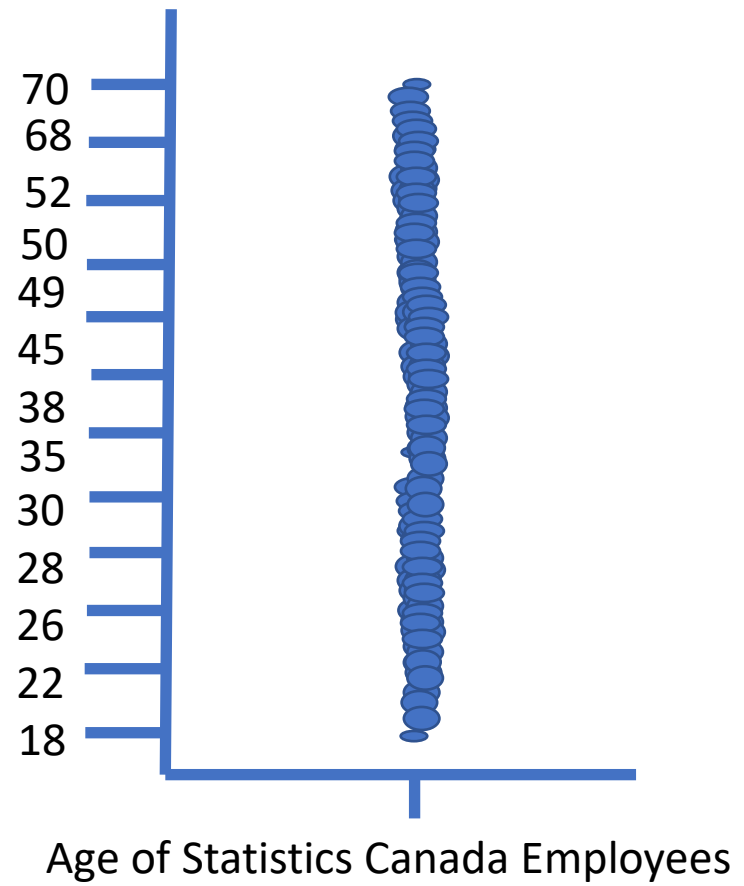


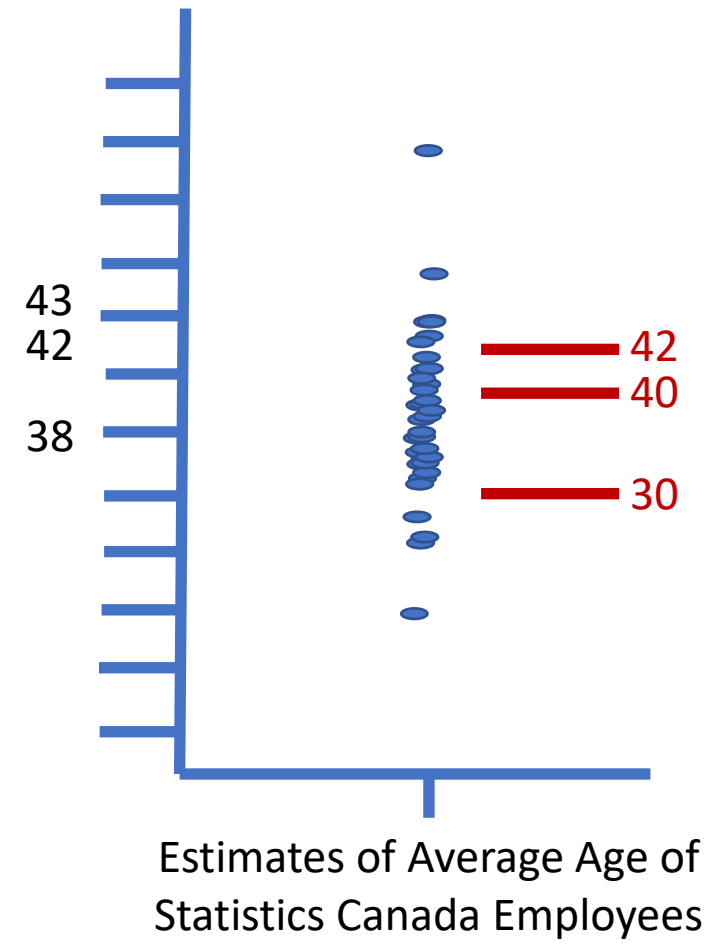
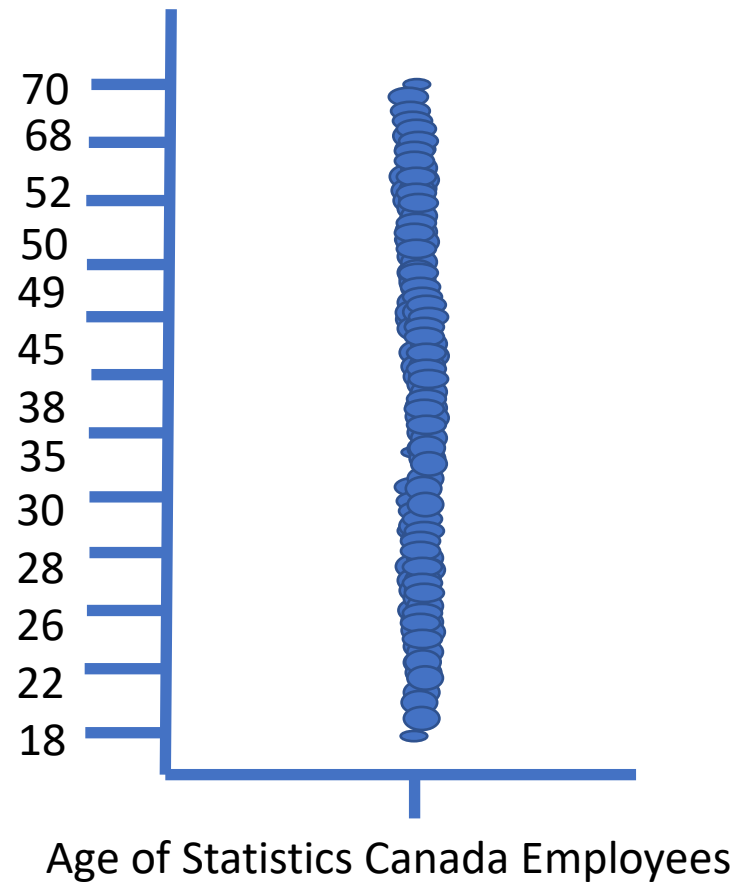




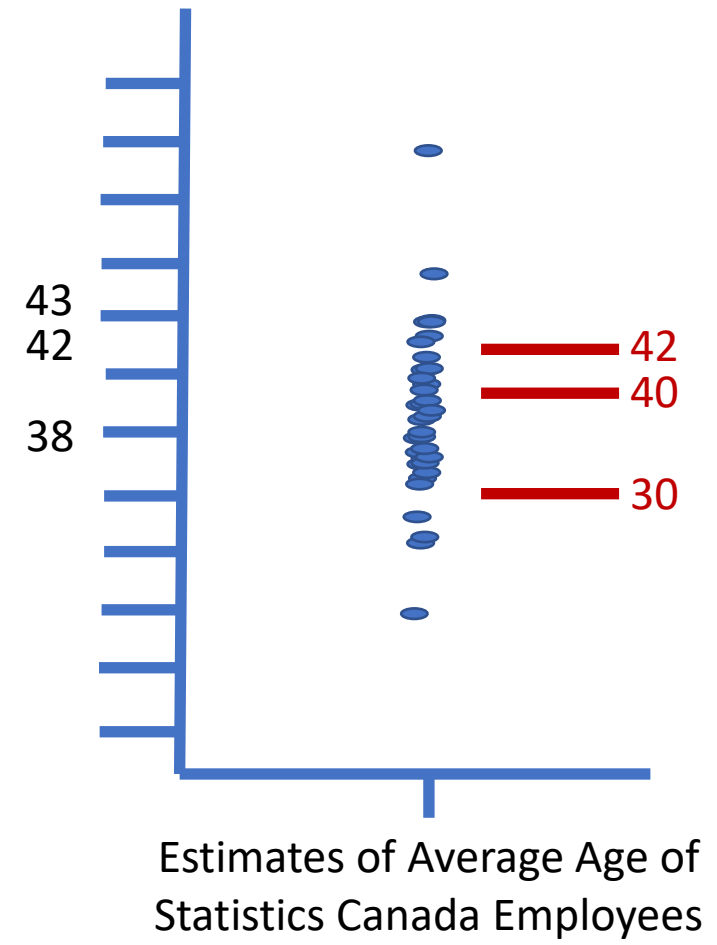
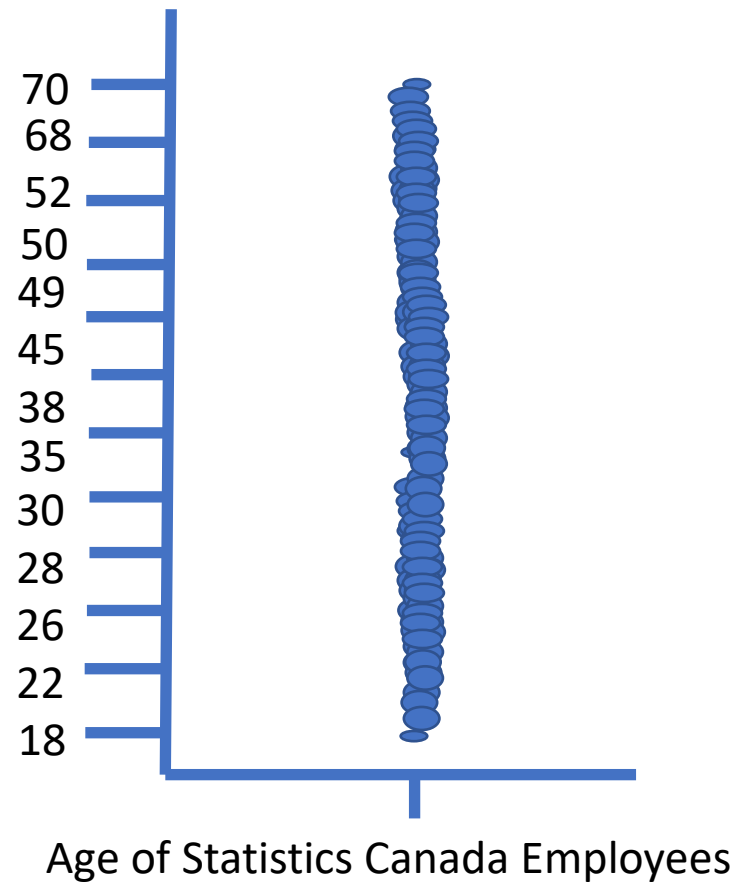








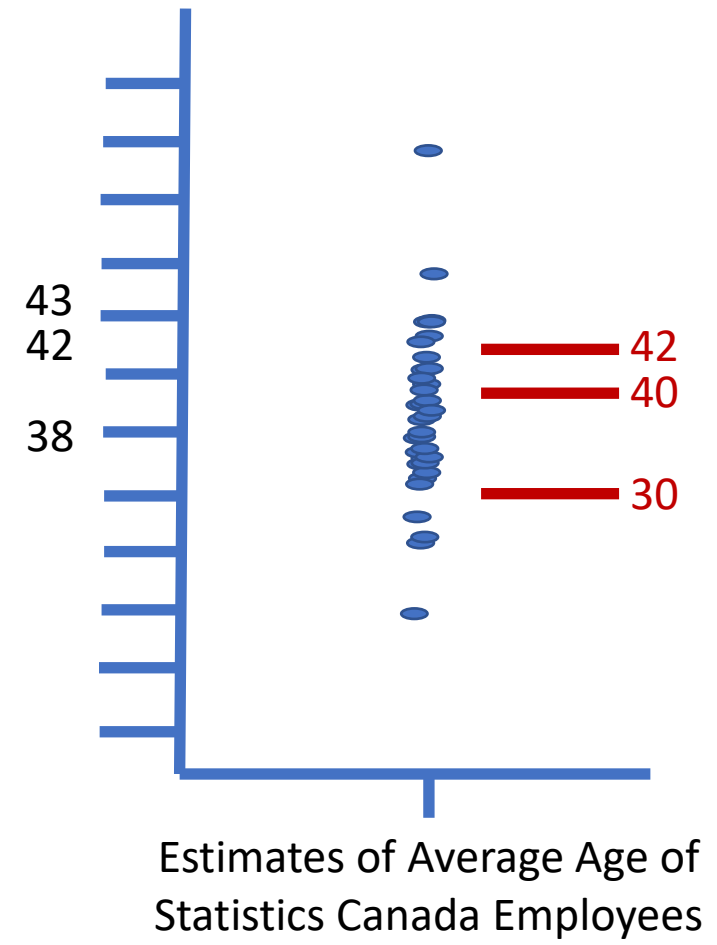
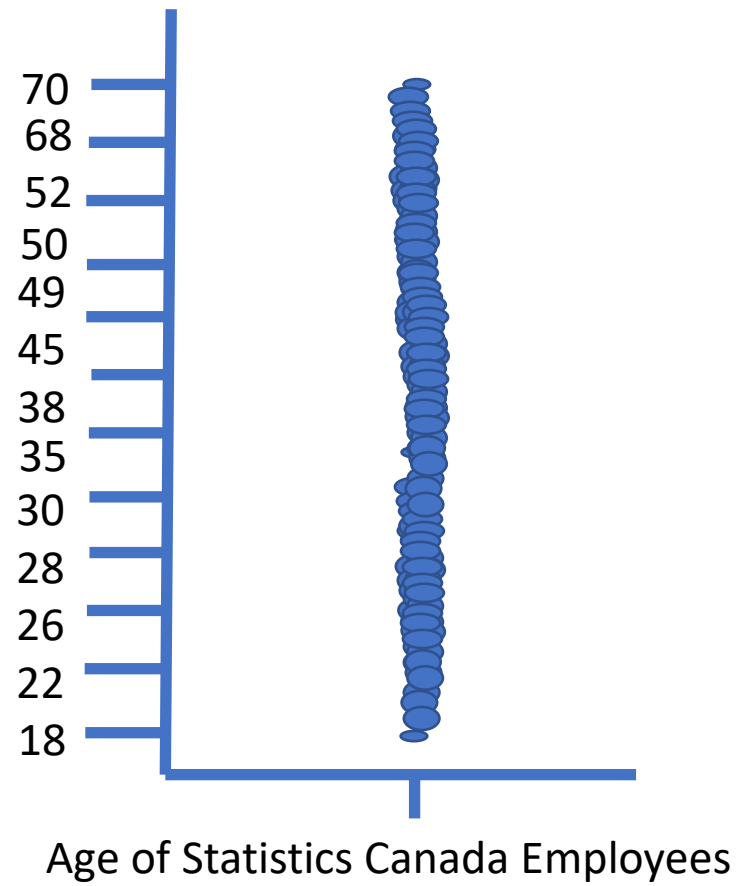
The sampling variance is the average squared deviation about the estimator's average value



The sampling variance is the average squared deviation about the estimator's average value

$(42-40)^2 = 4$

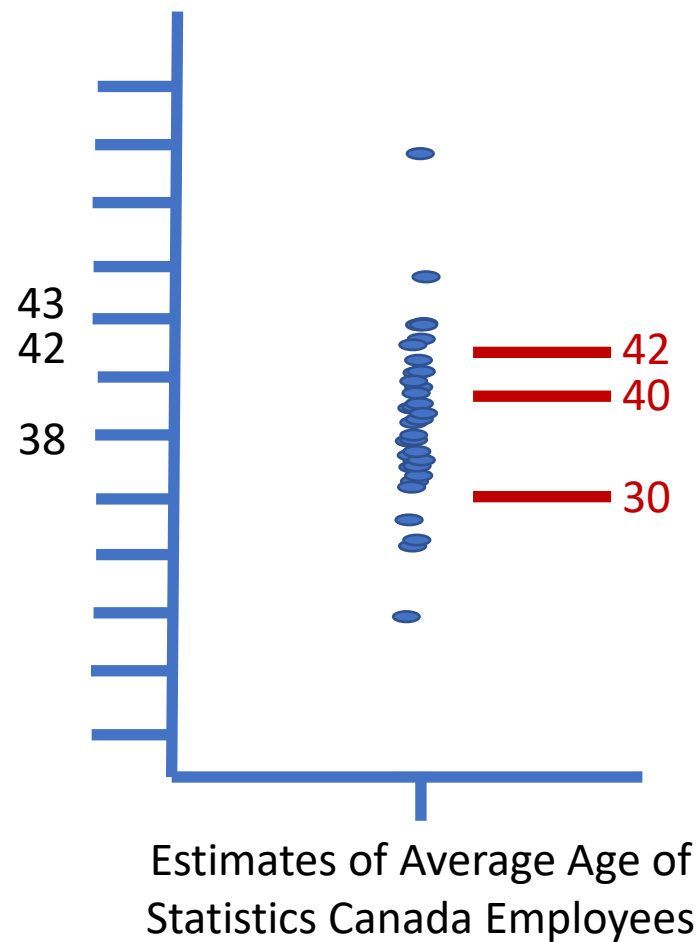
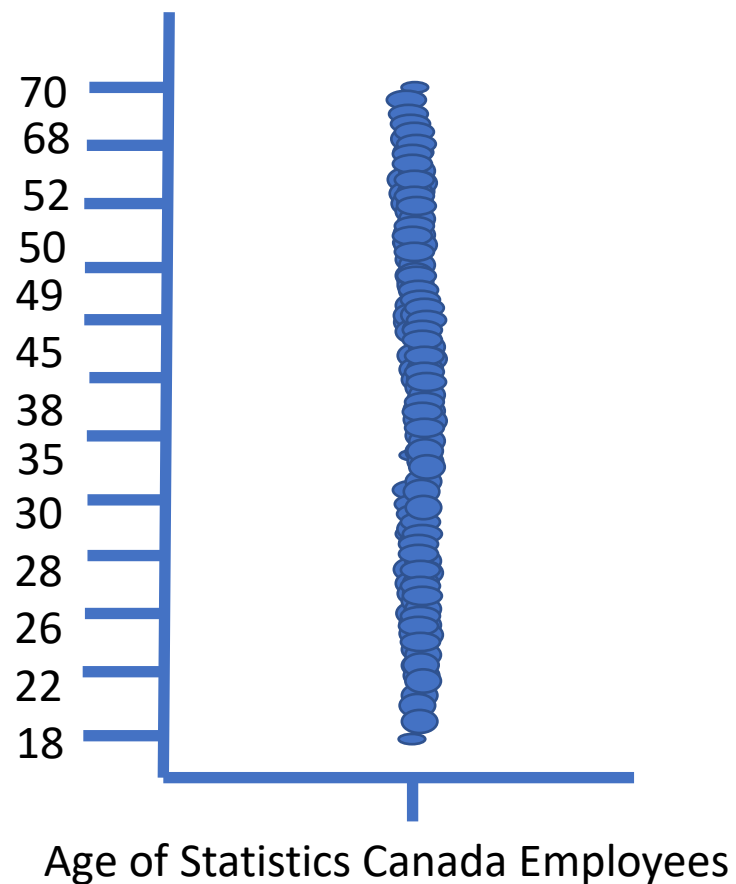
$(30-40)^2 = 100$



The sampling variance is the average squared deviation about the estimator's average value

$(42 \text{ years} - 40 \text{ years})^2 = 4 \text{ years}^2$

$(30 \text{ years} - 40 \text{ years})^2 = 100 \text{ years}^2$

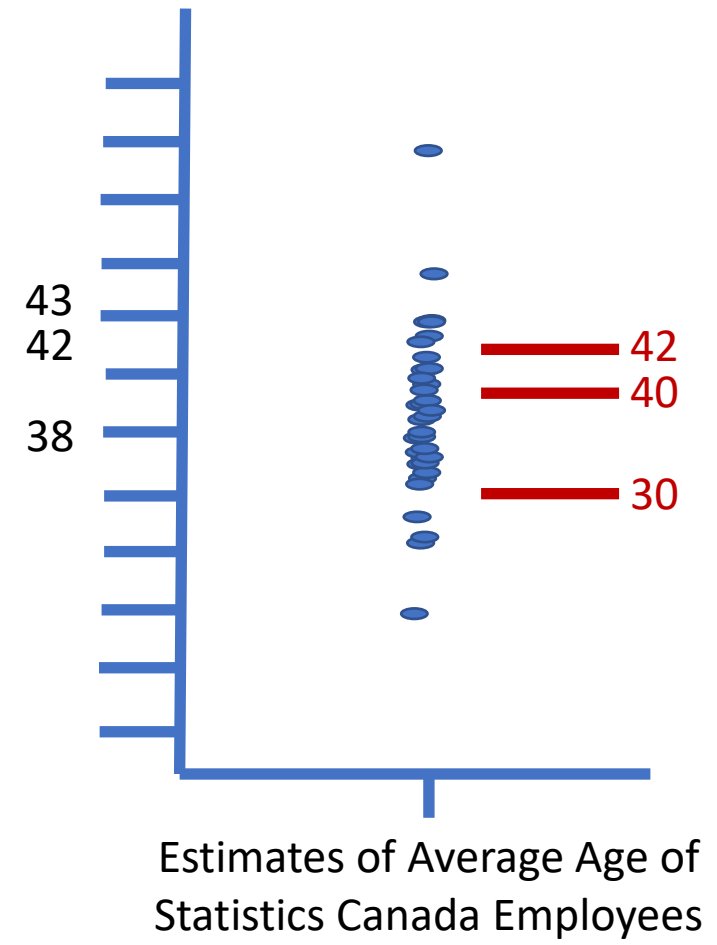
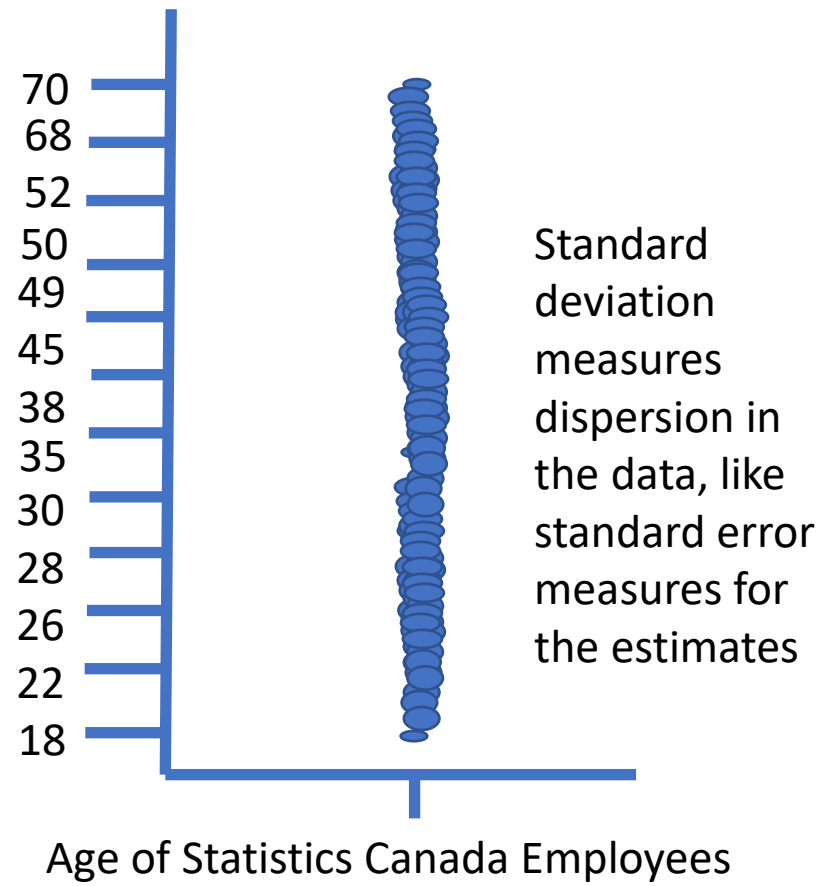


The sampling variance is the average squared deviation about the estimator's average value

$$(42 \text{ years} - 40 \text{ years})^2 = 4 \text{ years}^2$$

$$(30 \text{ years} - 40 \text{ years})^2 = 100 \text{ years}^2$$

The sampling variance is measured in square units...the standard error is the square root



The sampling variance is the average squared deviation about the estimator's average value

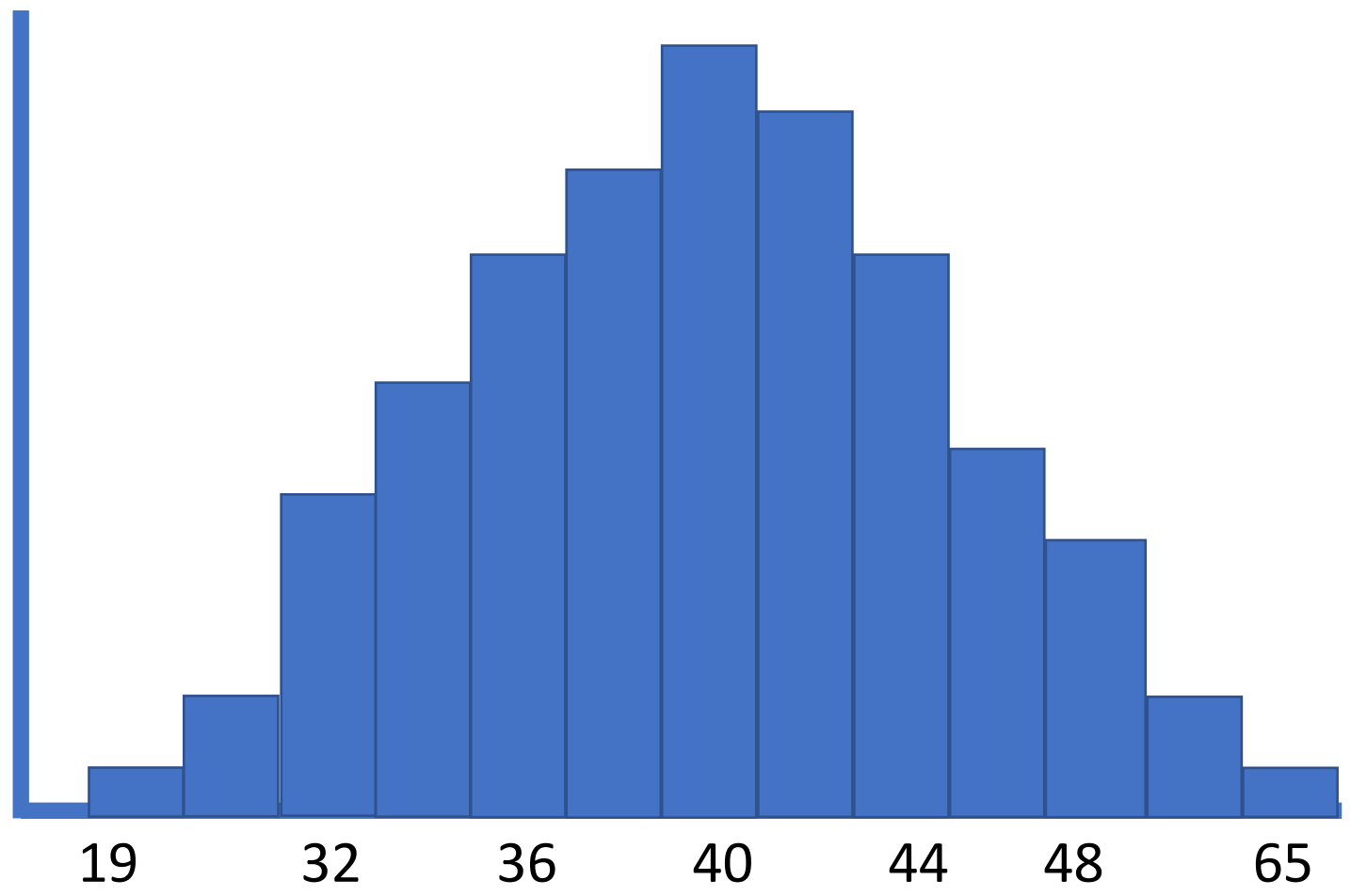
$(42 \text{ years} - 40 \text{ years})^2 = 4 \text{ years}^2$

$(30 \text{ years} - 40 \text{ years})^2 = 100 \text{ years}^2$

The sampling variance is measured in square units...the standard error is the square root



Count

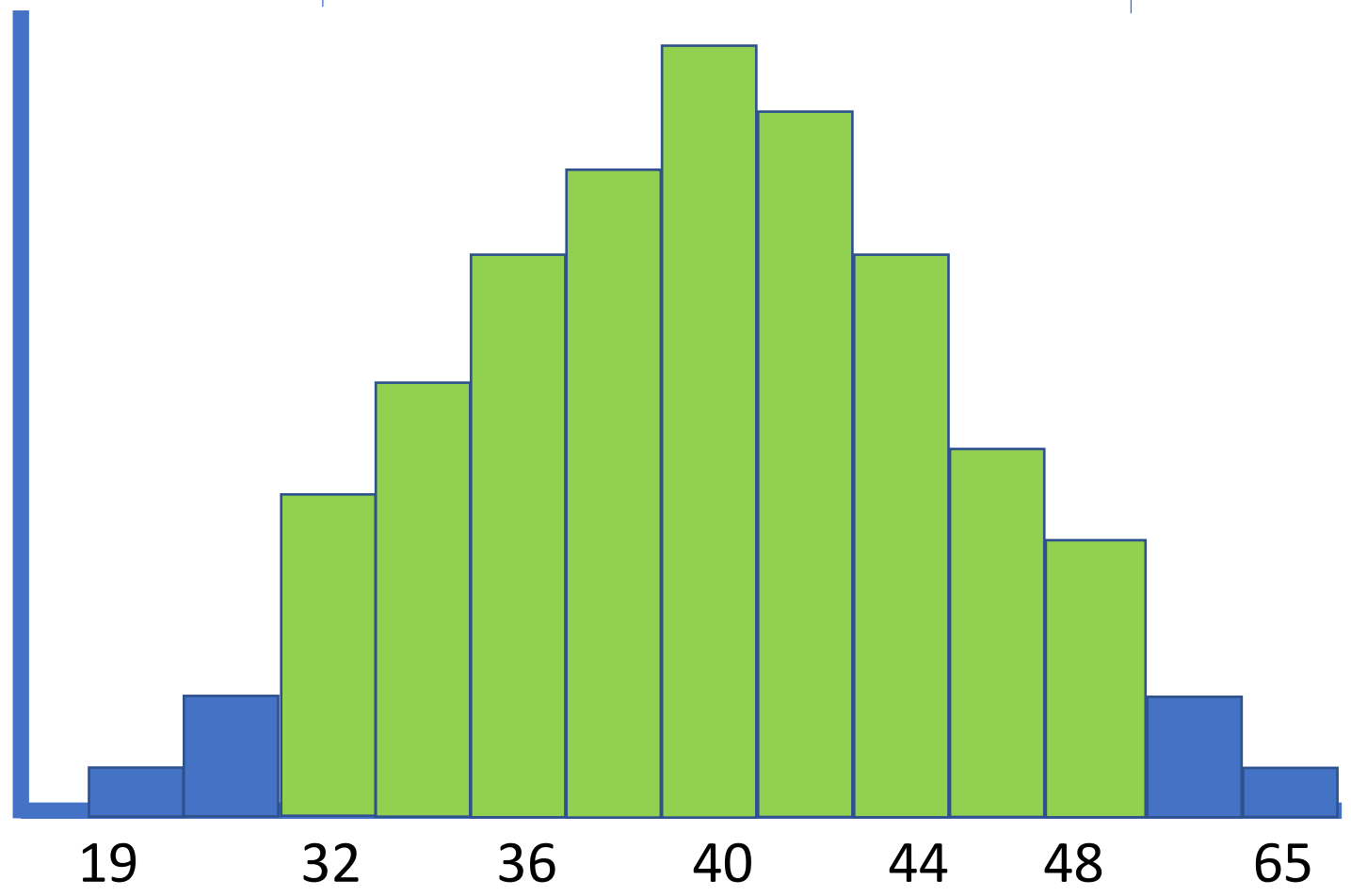


Average 40  
Standard Error 4



95%

Count

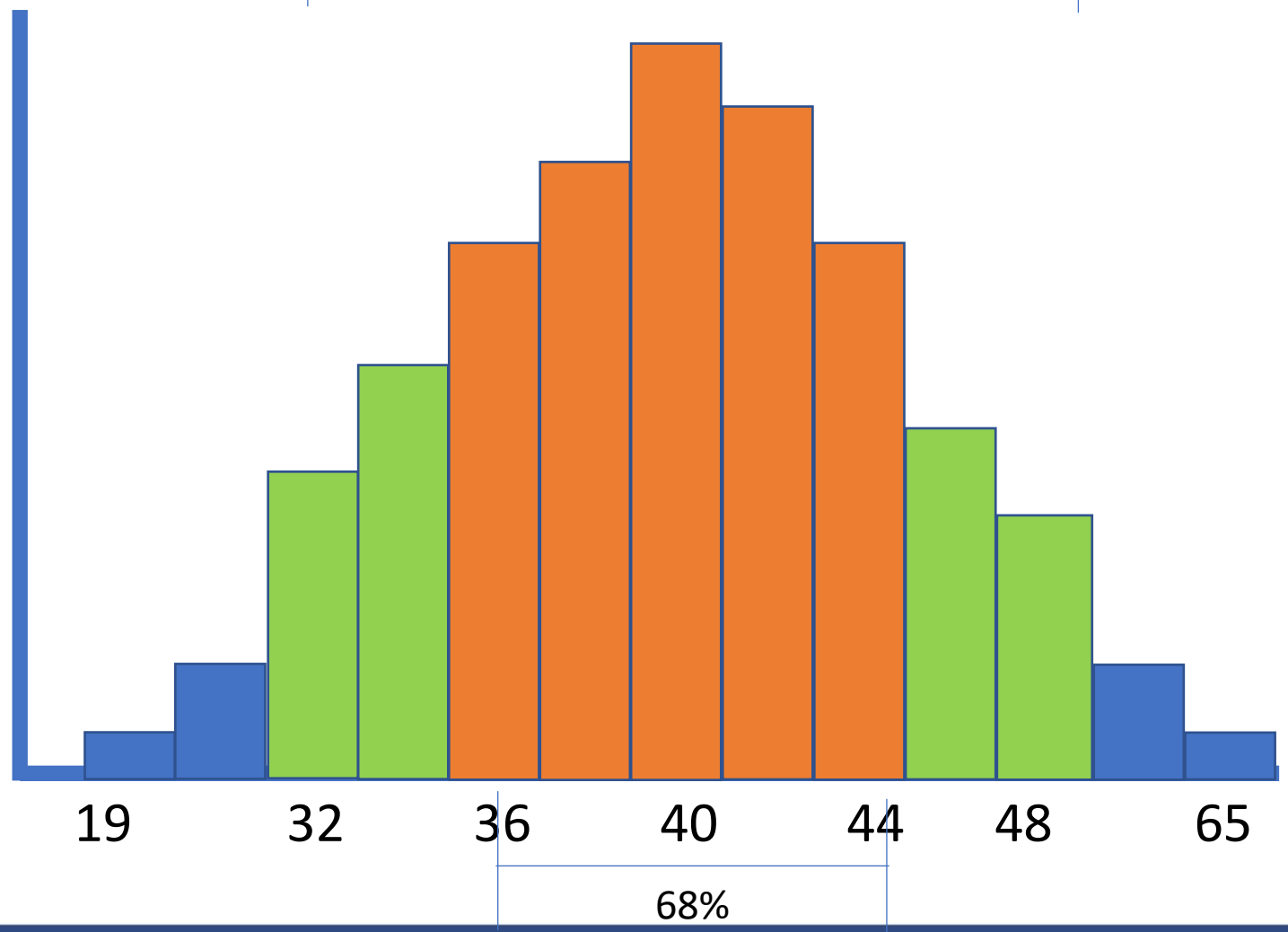


Average 40  
Standard Error 4

About 95% of the  
time the true  
average will be  
between 32 and 48



Count



Average 40  
Standard Error 4

About 95% of the time the true average will be between 32 and 48

About 68% of the time the true average will be between 36 and 44

# Confidence Interval (CI)

Example: Results of a survey in the newspapers

*According to a recent survey, 15% of Ottawa dwellings don't have a landline phone. The results, based on a sample of 1,234 dwellings, are considerate accurate within plus or minus 3 percentage points 19 times out of 20.*

What does this statement mean?

- If the survey was repeated many times, then 19 out of 20 (or 95% of the time), the confidence interval would cover the true population value
- The margin of error is 3 percentage points
- The confidence interval is the range 12% to 18%

# Coefficient of Variation (CV)

“My estimate has a standard error of 4!”

# Coefficient of Variation (CV)

“My estimate has a standard error of 4!”

Statistics Canada average age of 40 example

95% confidence interval -> (32 ; 48)

# Coefficient of Variation (CV)

“My estimate has a standard error of 4!”

Statistics Canada average age of 40 example

95% confidence interval -> (32 ; 48)

Metric tons of wheat produced in Canada, 2019, (32,347,800 metric tons)

95% confidence interval -> (32,347,792 ; 32,347,808)

# Coefficient of Variation (CV)

“My estimate has a standard error of 4!”

Statistics Canada average age of 40 example

95% confidence interval -> (32 ; 48)

Metric tons of wheat produced in Canada, 2019, (32,347,800 metric tons)

95% confidence interval -> (32,347,792 ; 32,347,808)

Having a standard error of 4 makes a big difference, depending on what measurements we are talking about.

## Coefficient of Variation (CV)

- Ratio of the standard error of the survey estimate to the estimate itself.
- Useful to assess the size of the standard error relative to the estimate of the characteristic being measured.
  - “Scale independent”

$$CV = \frac{\textit{standard error}}{\textit{Estimate}} \times 100\%$$

### Example: Guidelines for the Canadian Health Measures Survey (CHMS)

| Category | CV Interval                 | Guidelines   |
|----------|-----------------------------|--------------|
| 1        | $0.0\% \leq CV \leq 16.5\%$ | Publishable  |
| 2        | $16.6\% < CV \leq 33.3\%$   | Restricted   |
| 3        | $CV > 33.3\%$               | Unreleasable |

## Coefficient of Variation (CV)

- Ratio of the standard error of the survey estimate to the estimate itself.
- Useful to assess the size of the standard error relative to the estimate of the characteristic being measured.
  - “Scale independent”

$$CV = \frac{\text{standard error}}{\text{Estimate}} \times 100\%$$

Statistics Canada age example:  
CV = (4/40) x 100% = 10%

### Example: Guidelines for the Canadian Health Measures Survey (CHMS)

| Category | CV Interval                 | Guidelines   |
|----------|-----------------------------|--------------|
| 1        | $0.0\% \leq CV \leq 16.5\%$ | Publishable  |
| 2        | $16.6\% < CV \leq 33.3\%$   | Restricted   |
| 3        | $CV > 33.3\%$               | Unreleasable |

## Coefficient of Variation (CV)

- Ratio of the standard error of the survey estimate to the estimate itself.
- Useful to assess the size of the standard error relative to the estimate of the characteristic being measured.
  - “Scale independent”

$$CV = \frac{\text{standard error}}{\text{Estimate}} \times 100\%$$

Statistics Canada age example:  
 $CV = (4/40) \times 100\% = 10\%$

Wheat example is about 0.00001%

### Example: Guidelines for the Canadian Health Measures Survey (CHMS)

| Category | CV Interval                 | Guidelines   |
|----------|-----------------------------|--------------|
| 1        | $0.0\% \leq CV \leq 16.5\%$ | Publishable  |
| 2        | $16.6\% < CV \leq 33.3\%$   | Restricted   |
| 3        | $CV > 33.3\%$               | Unreleasable |

# Conclusion

- Measuring survey error serves many purposes:
  - Informs users of the validity of the data produced
  - Gives users the necessary information to do statistical testing
  - Allows for review, and improvement, of survey procedures



# Questions?



You can contact the PRASC team at:

[statcan.prasc-prasc.statcan@canada.ca](mailto:statcan.prasc-prasc.statcan@canada.ca)

4  
5