Food and Agriculture Organization of the United Nations

Training course to enhance collection of fisheries and aquaculture statistics

Module 2 – Refresher on biostatistics

2018

Module 2 – Refresher on biostatistics

Outline

1. Why conduct a refresher on biostatistics

2. Basic statistical terms (*population versus sample, variables versus indicator)*

3. Statistics – estimates *(mean, variance, standard deviation)*

4. Reliability, precision and accuracy of estimates *(distributions, confidence limits, relative error, bias)*

5. Basic concepts of data analysis in the context of SSF and aquaculture statistics

6. Statistical software for computing statistics

1

Why a refresher on biostatistics?

1.1 Why a refresher on biostatistics?

Most notions of general statistics (mean, standard deviation, variance, etc.) are probably well known to the participants in this training.

This module refreshes knowledge of the mathematical basics of the design and collection of statistical data (key statistical terminology, sampling theory, stratification, etc.)

The refresher aids understanding of the modules on

- Sampling and survey design for SSF and aquaculture statistics
- Calculation of indicators from collected data

2

Statistical terms

2.1 Statistical terms: population versus sample (1/5)

Fundamental terms in statistics

Target population	**Sample**
• Set of all items to study	• Subset or portion of the target population
• Information on the target population is sought if a complete enumeration or census is practically and financially possible	• Information is taken from the sample to make inferences on the population
Example: all artisanal canoes OR all canoes operating gillnets OR all households practicing SSF/aquaculture	Example: sample of daily catch, income, fishing effort, species compositions (characteristic of target population and/or estimated population average)
• Information about a population of size *N* the "population parameter"	• Information on a sample of size *n* is a "statistic"
Example: population mean $\mu = \frac{\sum_{i=1}^{N} Y_i}{N}$	Example: sample mean $\overline{y} = \frac{\sum_{i=1}^{N} y_i}{n}$

2.1 Statistical terms: population versus sample (2/5)

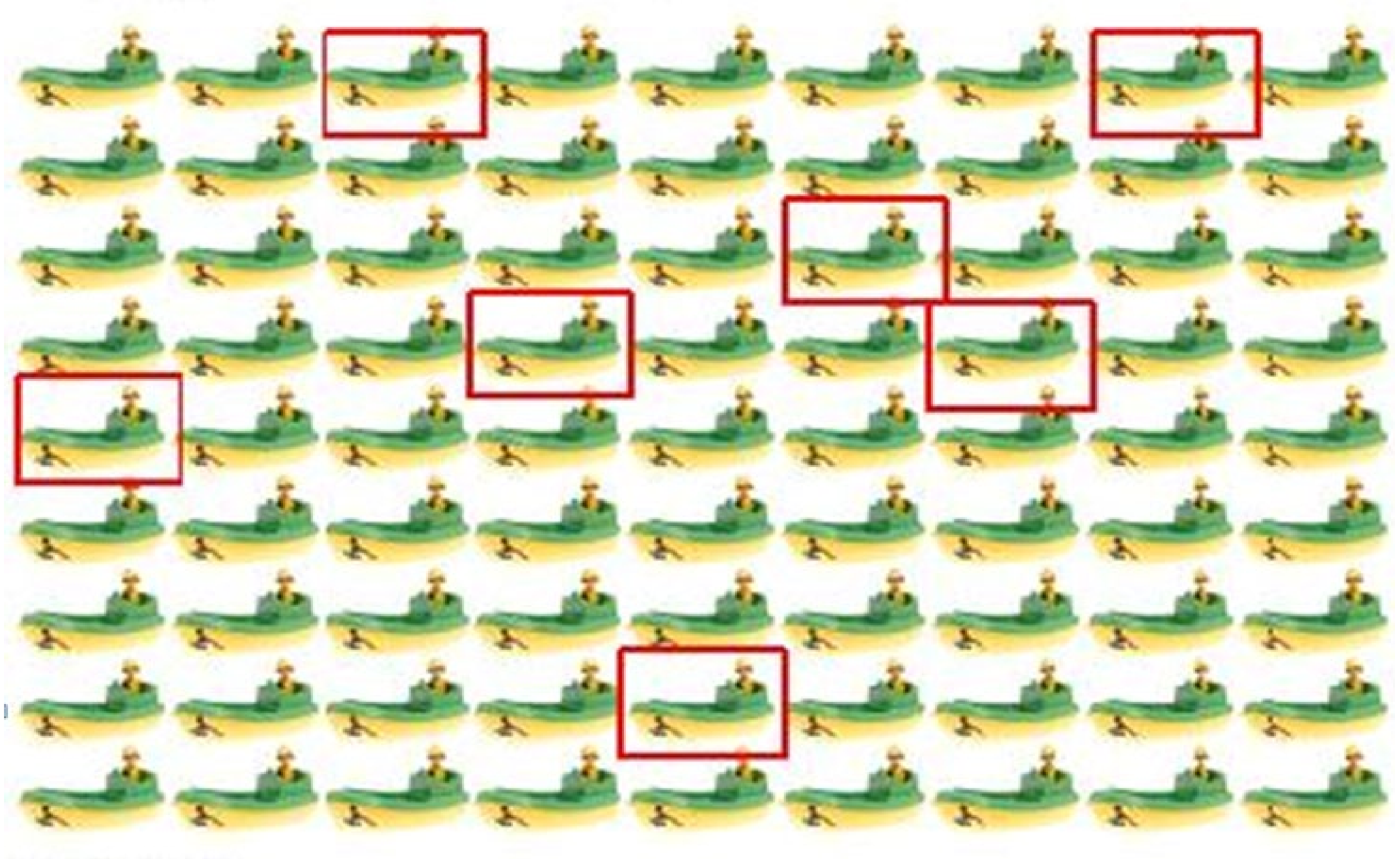
- Generally, NOT possible to "see" entire population
  - must sample to learn

- Values of estimates differ from sample to sample
  - This leads to variability in the sampling
  - Variability is inherent and conceals the "truth"

- Statistical conclusions (inferences) are always uncertain; however, the <u>uncertainty can be quantified</u>

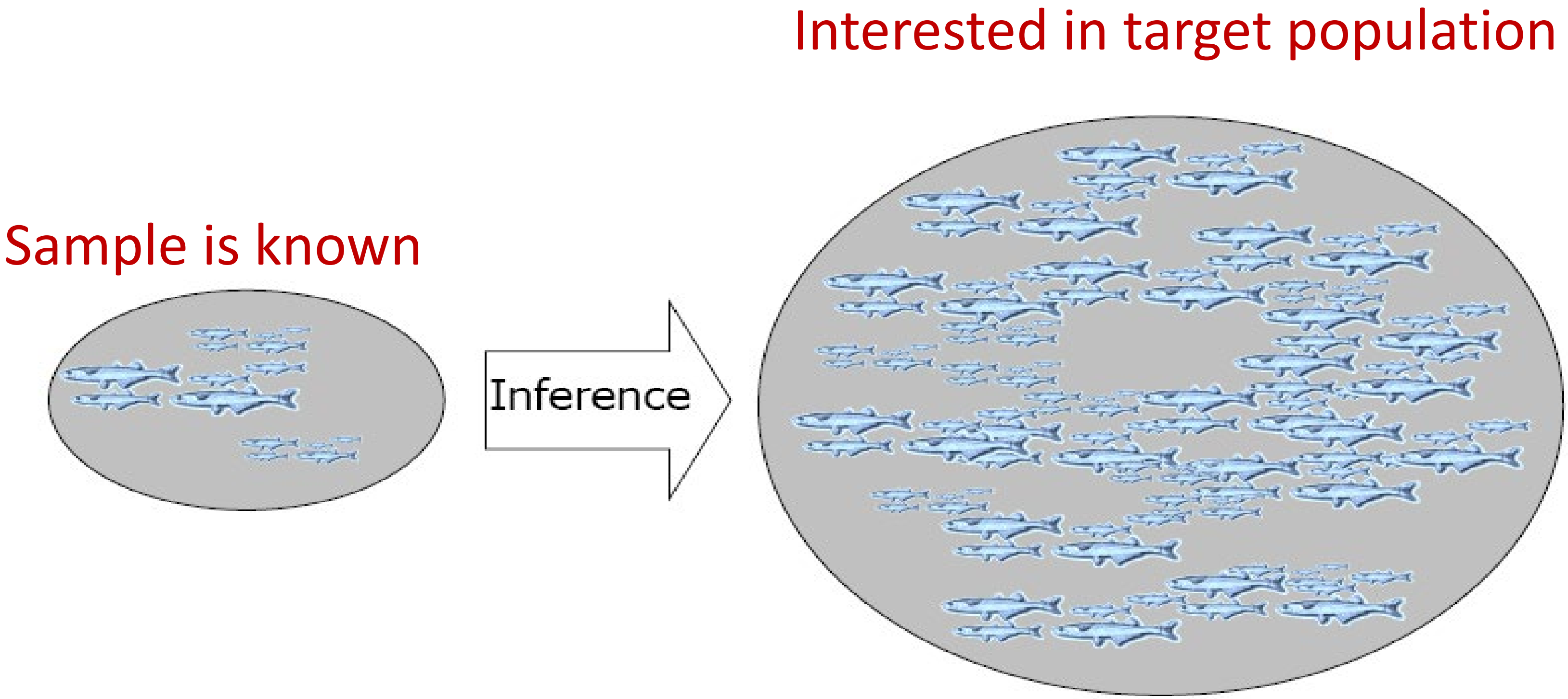
2.1 Statistical terms: population versus sample (3/5)

Example: population (81 vessels); sample (7 vessels)



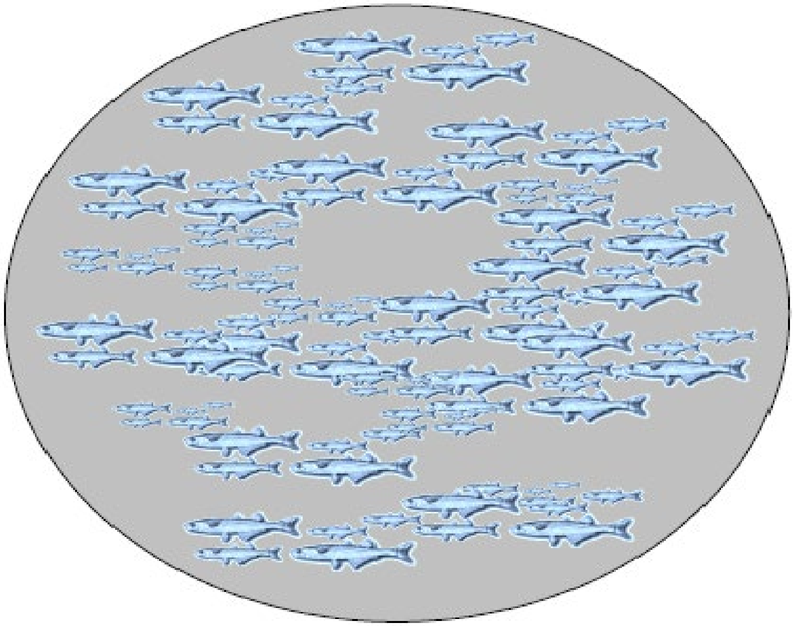
2.1 Statistical terms: population versus sample (4/5)

The sample is drawn from the target population. Statistics are calculated on the sample size and extrapolated to the target population.



Statistical inference is based on a theory and a methodology for generalizing from a sample to a population

2.1 Statistical terms: population versus sample (5/5)



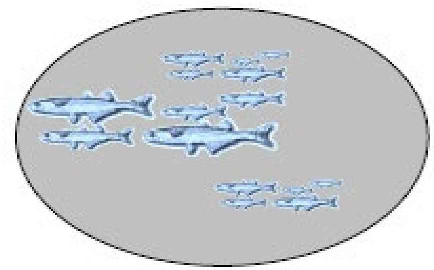
Population

quantity (count) = N

mean = μ

variance = σ^2

standard deviation = σ



Sample

quantity (count) = n

mean = $\bar{x}$

variance = s^2

standard deviation = s

2.2 Statistical terms: variable versus indicator

Fundamental terms in statistics

Variable	**Indicator**
• Information collected from the field	• Measures/follows performance of element under observation
Example: daily catch per fishing vessel, monthly income of fisherman, price of fish caught	Example: Catch Per Unit of Effort (CPUE), fishery contribution to GDP
• A variable takes usually nonconstant values	• An indicator is constructed from single or multiple variables
Example: daily catch is different for different fishermen, even when fishermen use the same type of vessel	Example: the CPUE is calculated from two variables – *daily catch* and *effort*
• Information about a population of size *N* is called the population parameter	• Indicators take nonconstant values

3

Statistics – estimates

3.1 Arithmetic, geometric and harmonic means (1/4)

Also known as the *average*. It is used to show the *central tendency* of a set of '*n*' numbers denoted by a variable y_i

Types	Note
• Arithmetic mean	The harmonic mean is suitable for calculating the average of rates. The geometric mean is suited to calculating averages of ratios or percentage changes.
$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$	
• Geometric mean	
$Geometric\_mean = \sqrt[n]{y_1 y_2 y_3 \ldots y_n}$	
• Harmonic mean	
$Harmonic\_mean = \frac{n}{\sum_{i=1}^{n} \frac{1}{y_i}}$	

3.1 Arithmetic, geometric and harmonic means (2/4)

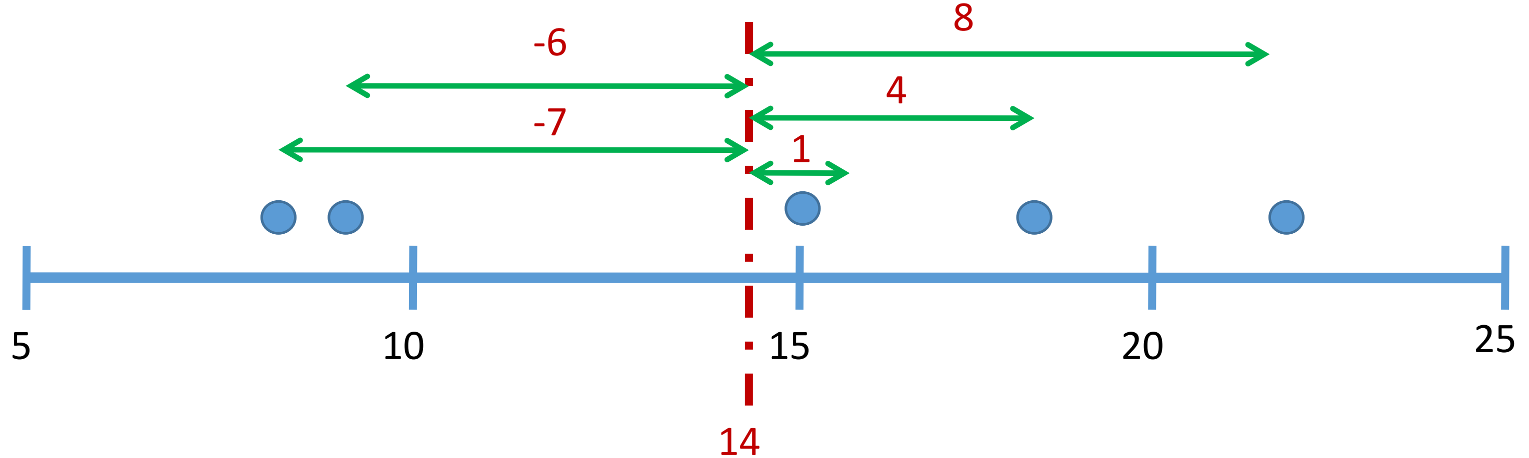
- **Arithmetic mean** - the sum of all measurements divided by the number of observations in the data set
- Weighted mean - an arithmetic mean that incorporates weighting into certain data elements
- Geometric mean - the *n*th root of the product of the data values
- Harmonic mean - the reciprocal of the arithmetic mean of the reciprocals of the data values

3.1 Arithmetic, geometric and harmonic means (3/4)

Properties of arithmetic mean

If numbers have a mean of $\bar{x}$, then, as $x_i - \bar{x}$ is the distance from a given number to the mean, then the numbers (distances) to the left of the mean are balanced by the numbers to the right of the mean.

Sample	y_1	y_2	y_3	y_4	y_5	Mean
Traps	8	15	22	7	18	14



3.1 Arithmetic, geometric and harmonic means (4/4)

Example

Score	Frequency	(Score) * (Frequency)
17	1	17
16	0	0
15	4	60
14	6	84
13	4	52
12	1	12
11	1	11
10	1	10
108	**18**	**246**

$$A := \frac{1}{n}\sum_{i=1}^{n} x_i$$

Arithmetic mean

= 108/8

=13.5

$$A := \frac{1}{n}\sum_{i=1}^{n} x_i\, f_i$$

Weighted mean

= 246/18

=13.67

17	**15**	**15**	**15**	**15**	**14**	**14**	**14**	**14**	**14**	**14**	**13**	**13**	**13**	**13**	**12**	**11**	**10**

3.2 Median and quartiles (1/2)

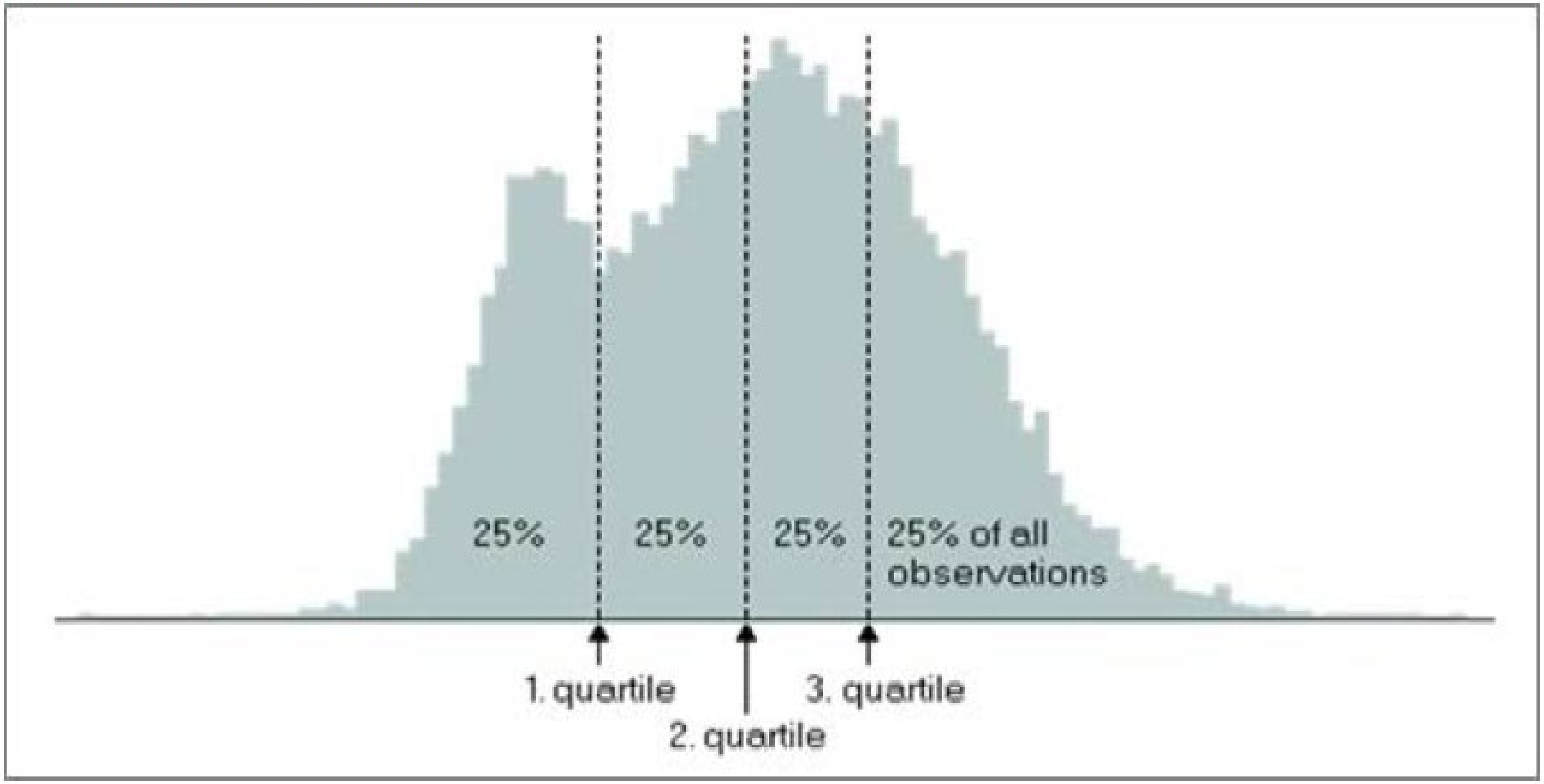
Also used to show the *central tendency* of a set of '*n*' numbers when arranged from smallest to biggest

Types	Issues
• First quartile (Q1)	Apart from the mean, it is also advisable to describe the catch, income, etc. of the "average fisher" by the middle value, the lower 25^{th} percentile value and the upper 75^{th} percentile value
Cuts lowest 25 percent of observations (25^{th} percentile)	
• Second quartile (Q2) = Median	
Cuts lowest 50 percent of observations (50^{th} percentile)	
• Third quartile (Q3)	
• Cuts lowest 75 percent of observations (75^{th} percentile)	

3.2 Median and quartiles (2/2)

Observations could be the catch of canoes of same size, the incomes of fishermen, etc.

Quartiles of a distribution



3.3 Dispersion and variability (1/)

These include the variance, standard deviation, coefficient of variation and standard error of the mean.

Questions	note
1. Is the mean representative?	Close clustering around the mean indicates that the mean represents the observations well. The more spread out the observations are, the higher the variability.
2. Are observations well clustered around the mean?	
3. How spread out is the data?	

3.3 Dispersion and variability (2/3)

Dispersion measures around a mean of y_i calculated from n units

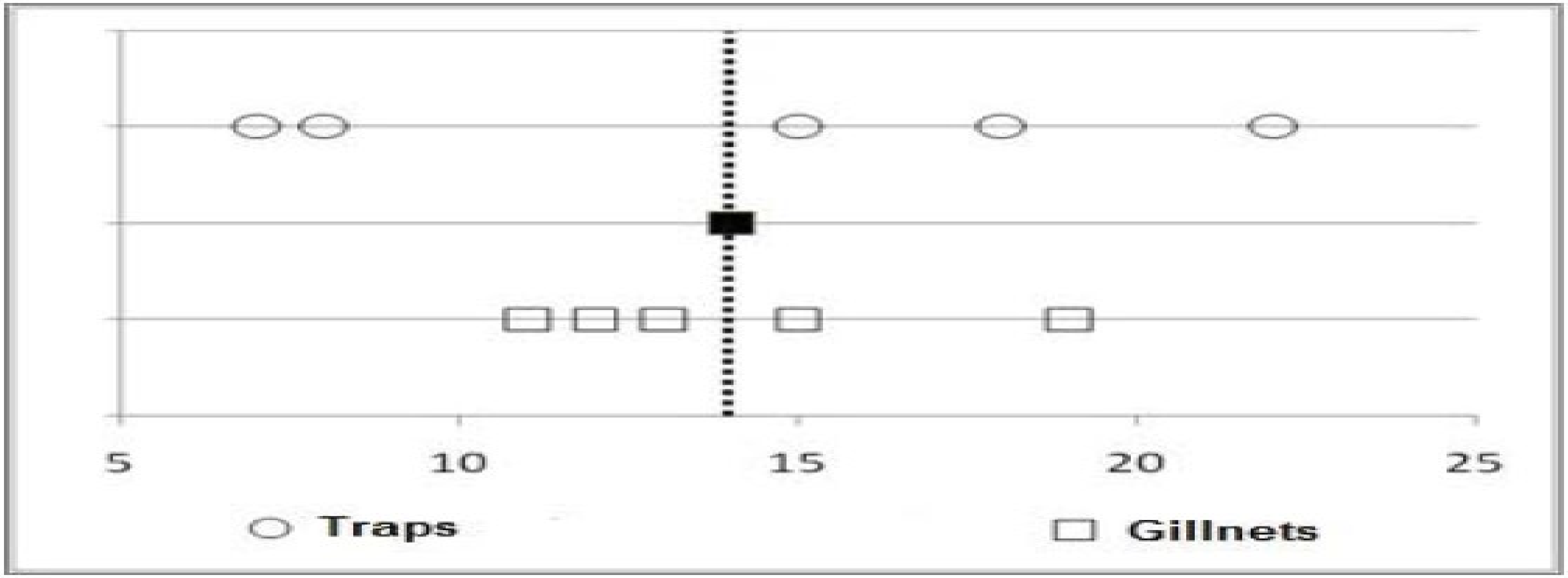
Measure	Formula
1. Variance *NOTE: a high dispersion from the mean leads to a high variance*	$Variance = s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$
2. Standard deviation *NOTE: a high variance leads to a high standard deviation*	$StdDev = s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$
3. Coefficient of Variation (CV) *Example: CV of 0.1 means that observations deviate by 10 percent around the mean*	$CV = \frac{s}{\bar{y}} = \frac{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}}{\bar{y}}$
4. Standard error of the mean *NOTE: a high standard deviation gives a high standard error of the mean*	$SEM = \frac{s}{\sqrt{n}} = \frac{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}}{\sqrt{n}}$

3.3 Dispersion and variability (2/3)
Example (same mean, different dispersion)

Sample observations of daily catches of gillnets and traps

Sample	**y_1**	**y_2**	**y_3**	**y_4**	**y_5**	**Mean**
Gillnets	12	15	19	13	11	**14 kg**
Traps	8	15	22	7	18	**14 kg**

Daily catch of gillnets and traps



4

Reliability, precision and accuracy of estimates

4.1 Reliability of estimates through confidence limits

- After calculating the mean of a set of observations, it is necessary to show how close the estimate probably is to the population ("true") mean
- Confidence limits indicate how accurate the estimate of the mean is likely to be
- For example, setting 95 percent confidence limits means that if repeated random samples were taken from a population and the mean and **confidence limits** calculated for each sample, the **confidence interval** for 95 percent of the samples would include the population ("true") mean
- **Confidence limits** are the numbers at the upper and lower ends of a **confidence interval**; for example, if the mean is 7.4 kg with **confidence limits of 5.4 and 9.4**, the **confidence interval is 5.4kg to 9.4kg**
- Confidence limits are obtained using the **mean, distributions (t-distribution for SSF) and standard error of mean** information

4.2 Distributions (1/2)

These equations model observations using information such as means and variance.

Distribution	Note
Normal distribution	Models continuous variables using population mean and variance. Used for samples sizes above 100
$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	The normal distribution helps to calculate confidence limits.
To obtain 90 percent confidence limit, take	Mean ± 1.64 x Standard Error of Mean $\bar{y}\pm 1.64\frac{s}{\sqrt{n}}$
To obtain 95 percent confidence limit, take	Mean ± 1.96 x Standard Error of Mean $\bar{y}\pm 1.96\frac{s}{\sqrt{n}}$
To obtain 99 percent confidence limit, take	Mean ± 2.58 x Standard Error of Mean $\bar{y}\pm 2.58\frac{s}{\sqrt{n}}$

4.2 Distributions (2/2)

However, samples from SSF operations are small and not normally distributed. Therefore, another distribution is used to obtain confidence limits.

Distribution	Note
t-distribution	Models continuous variables using *degrees of freedom* and the gamma function
$f(t) = \dfrac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi}\,\Gamma(\frac{v}{2})}\left(1+\frac{t^2}{v}\right)^{-\frac{v+1}{2}}$	Preferable for small sample sizes (<100) – Preferred for SSF statistics – Provides means to calculate confidence limits
To obtain confidence limit, take	Mean ± tn-1 x Standard Error of Mean $ConfLimit = \bar{y} \pm t_{n-1}\dfrac{s}{\sqrt{n}}$
tn-1	Is the upper critical value of the t-distribution or the t fractiles from t-distribution table with n-1 degrees of freedom, for sample of size *‘n’* at confidence levels of 90 percent, 95 percent and 99 percent

4.3 Distributions and confidence limits
Example

- Sample: 31 vessels with **mean** daily catch of **87 kg**
- **Standard error of mean: 10.3 kg**
- To obtain confidence limits, use **t-distribution**
  - the t-fractile $\mathbf{t_{n-1}}$ value for sample of 31 at 95 percent confidence level is 2.04
  - the confidence limit is obtained by first multiplying the t-fractile $\mathbf{t_{n-1}}$ value by the standard error mean, thus 2.04 x 10.3 kg = 21.2kg
  - **confidence limits: 87 kg ± 21.2 kg**
  - reported as 87 ±21.2kg (95% confidence limits).
  - Thus, the **confidence interval** is **65.8 kg** (87 - 21.2) **to 108.2 kg** (87 + 21.2)

4.4 The t-distribution (1/3)

Select the t fractiles at 90 percent, 95 percent or 99 percent confidence levels. If fractiles are selected for a sample of size *n* at a 95 percent confidence level, the calculated confidence limits will be called 95 percent confidence limits

No. of samples	Degrees of freedom	90% tf	95% tf	99% tf
2	1	6.31	12.71	63.66
3	2	2.92	4.30	9.93
4	3	2.35	3.18	5.84
5	4	2.13	2.78	4.60
6	5	2.02	2.57	4.03
7	6	1.94	2.45	3.71
8	7	1.90	2.37	3.50
9	8	1.86	2.31	3.36
10	9	1.83	2.26	3.25
11	10	1.81	2.23	3.17
12	11	1.80	2.20	3.11
13	12	1.78	2.18	3.06
14	13	1.77	2.16	3.01

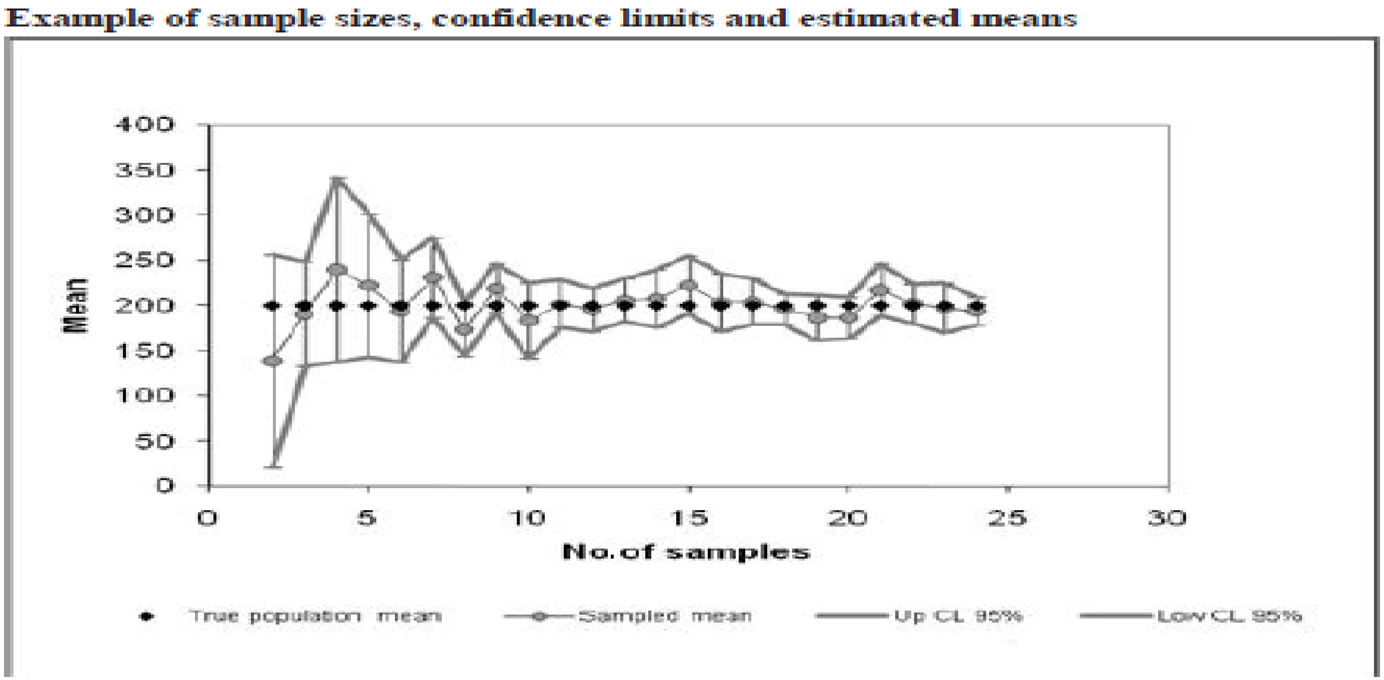
4.4 The t-distribution (2/3)

DO YOU SEE THE 95 PERCENT FRACTILE VALUE FOR A SAMPLE SIZE OF 31?

No. of samples	Degrees of freedom	90% tf	95% tf	99% tf
16	15	1.75	2.13	2.95
17	16	1.75	2.12	2.92
18	17	1.74	2.11	2.90
19	18	1.73	2.10	2.88
20	19	1.73	2.09	2.86
26	25	1.71	2.06	2.79
31	30	1.70	2.04	2.75
41	40	1.68	2.02	2.70
51	50	1.67	2.01	2.68
61	60	1.67	2.00	2.66
81	80	1.67	1.99	2.64
101	100	1.66	1.98	2.63
oo	oo	1.65	1.96	2.58

4.4 The t-distribution (3/3)

An increasing sample size picked from the population, estimating the mean and the standard error of mean, and calculating the confidence limits from fractile values of the t-distribution for the increasing sample sizes, leads to approaching the true population mean.



4.5 Data precision and accuracy (1/2)

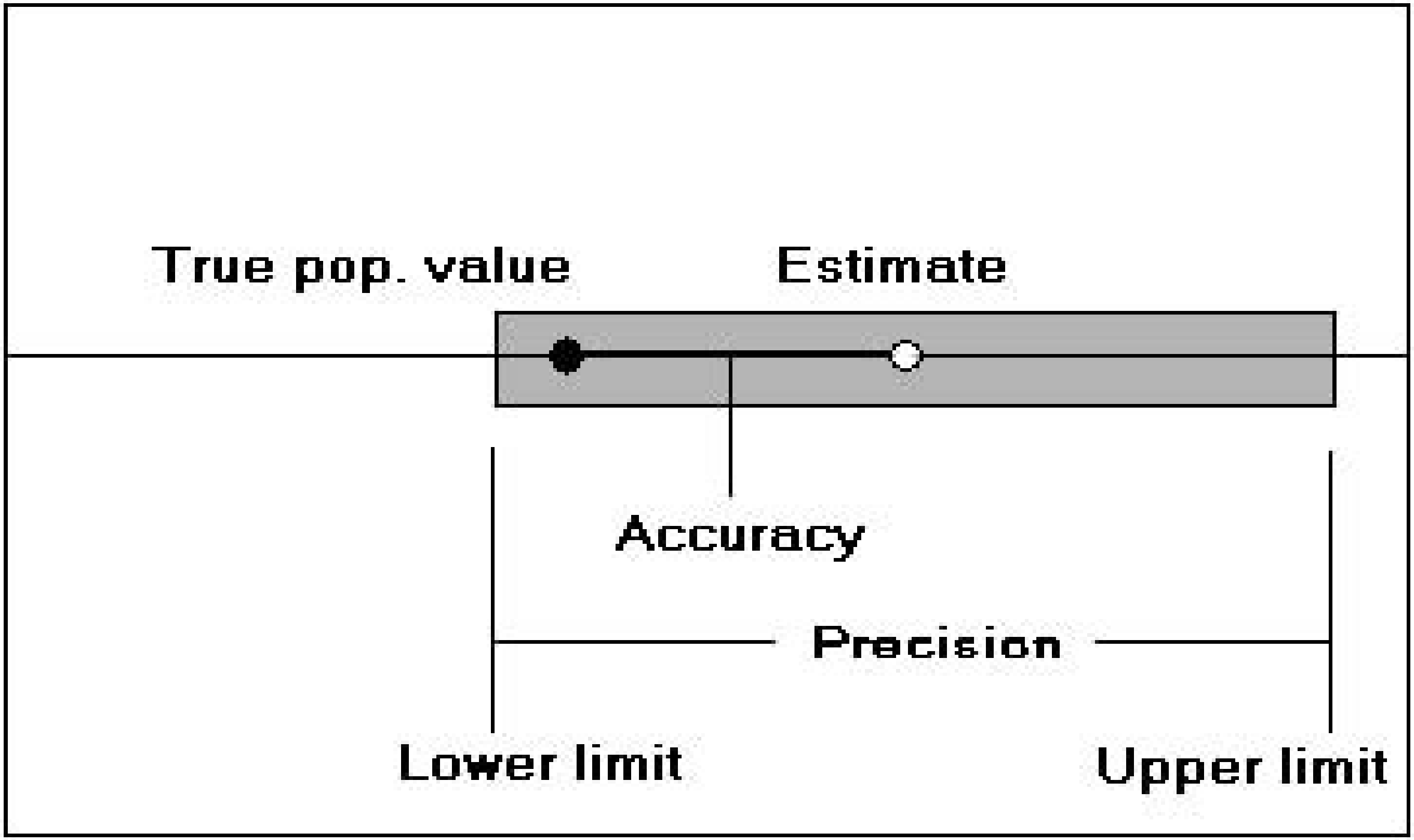
- Precision

  Closeness of repeated measurements of the same quantity. **Precision examines how well-clustered around the mean the sample observations are. Note its relation to the confidence interval (upper and lower limits).**

- Accuracy

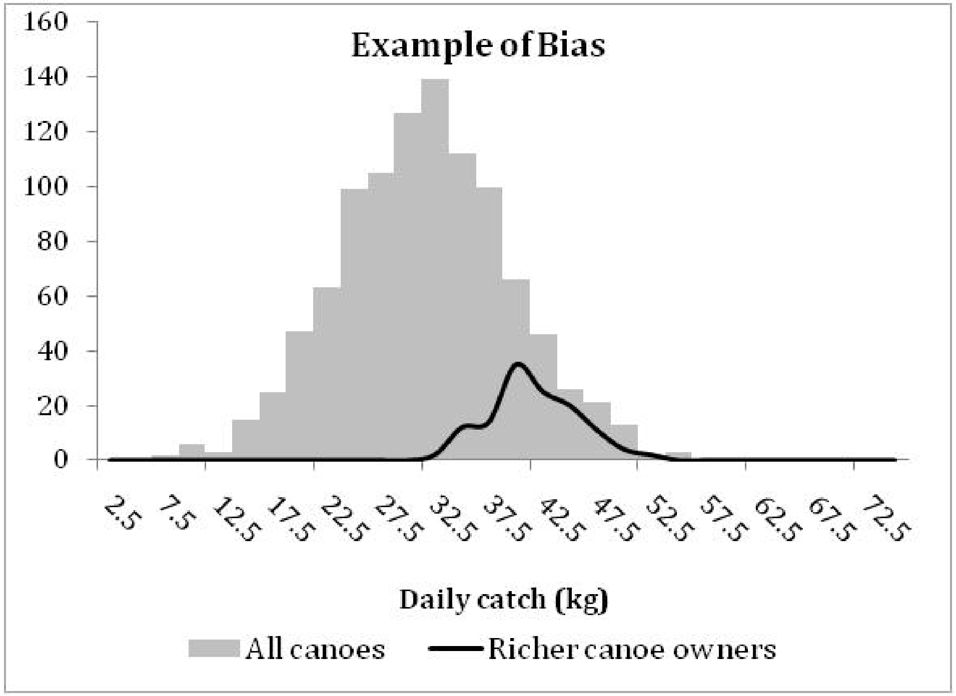
  Closeness of measured or computed value to its **true target population** value. **Accuracy examines whether the estimate mean is close to the true population parameter. A sample that gives an inaccurate estimate is called a BIASED SAMPLE.**

4.5 Data precision and accuracy (2/2)



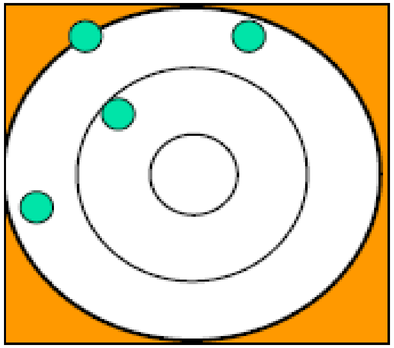
4.6 Bias of estimates

- **Bias** is the tendency for sample estimates to centre upon a value that is different from the true value, as data accumulate
- Systematic errors are made
- Biased data may produce <u>inaccurate</u> but <u>precise</u> estimates

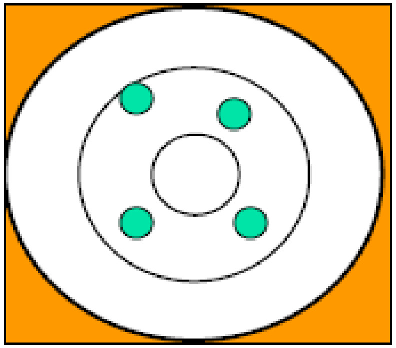


4.7 Accuracy of estimates

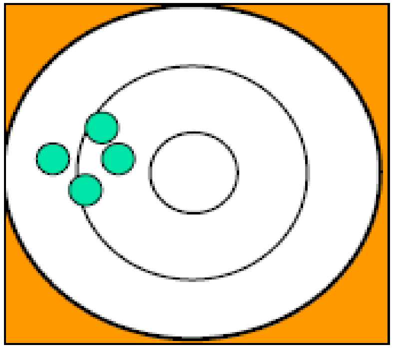
Accuracy = Precision + Bias



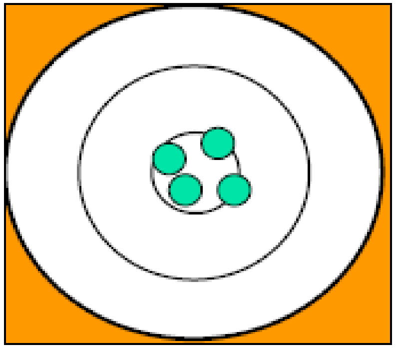
Not accurate and not precise



Accurate but not precise
(Vaguely right)



Precise but not accurate
(Precisely wrong)



Accurate and precise

4.8 How to quantify the accuracy of estimates

- As seen, the standard error of the mean decreases when sample size increases

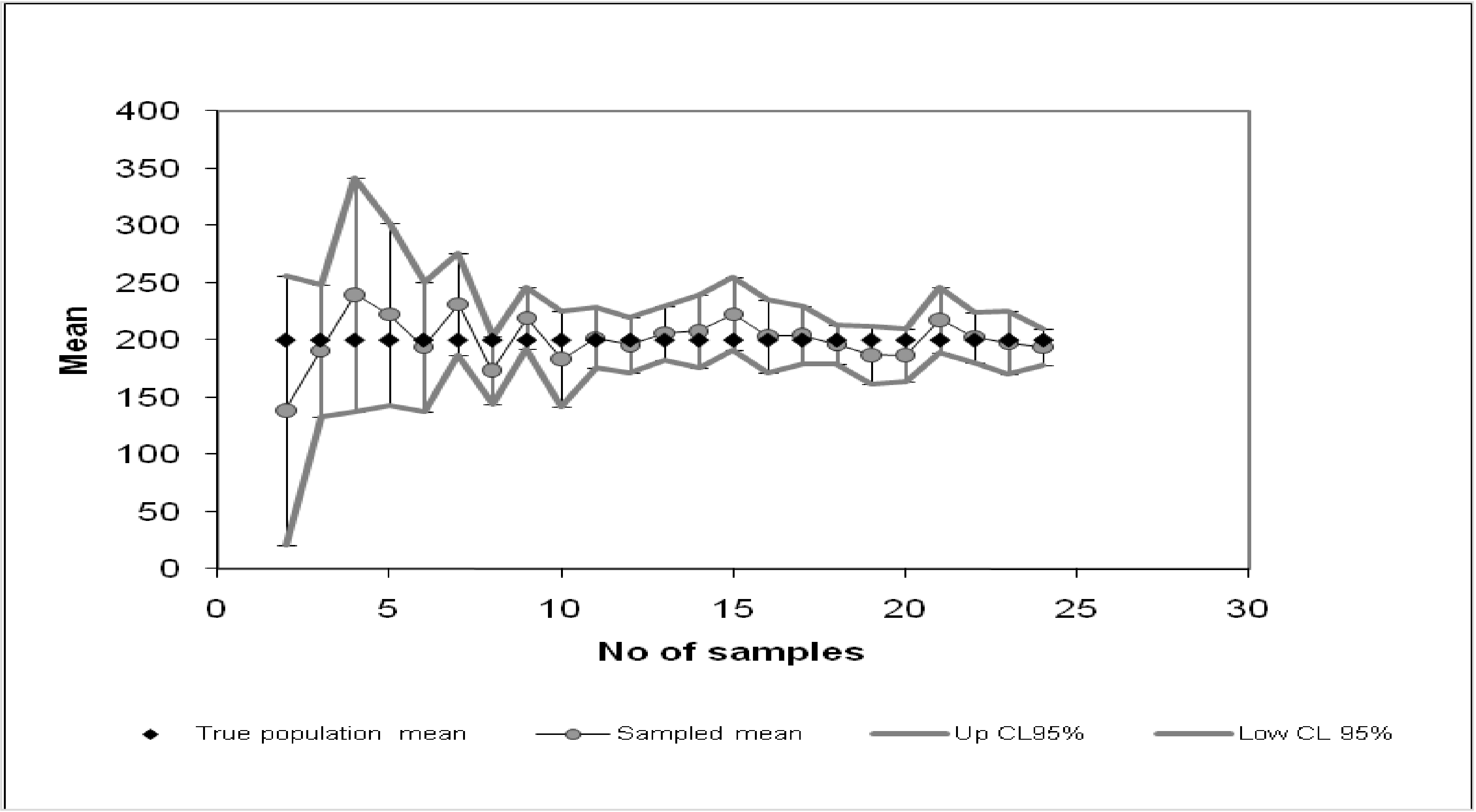
$$se.m = {}^{s}/_{\sqrt{n}}$$

- The **CL decreases** as **sample size increases**; also, both tn-1 and SEM decrease with increasing sample sizes:

CL = $\pm t_{n-1} * \text{sem}$

4.9 Sample sizes, confidence limits and estimated means

Increased sample size -> decreased CL



4.10 Relative error (1/2)

- The Confidence Limit (CL) decreases if sample size increases
- CLs are expressed in <u>absolute</u> values
- Dividing the CL by the sample mean gives the relative error (ε) of the estimate:

$$\varepsilon = \frac{CL}{\bar{x}}$$

A **relative error** of 10% with a 90% confidence limit means that there is a 90% probability that the value of the estimated sample mean deviates from the true target population mean by no more than 10%.

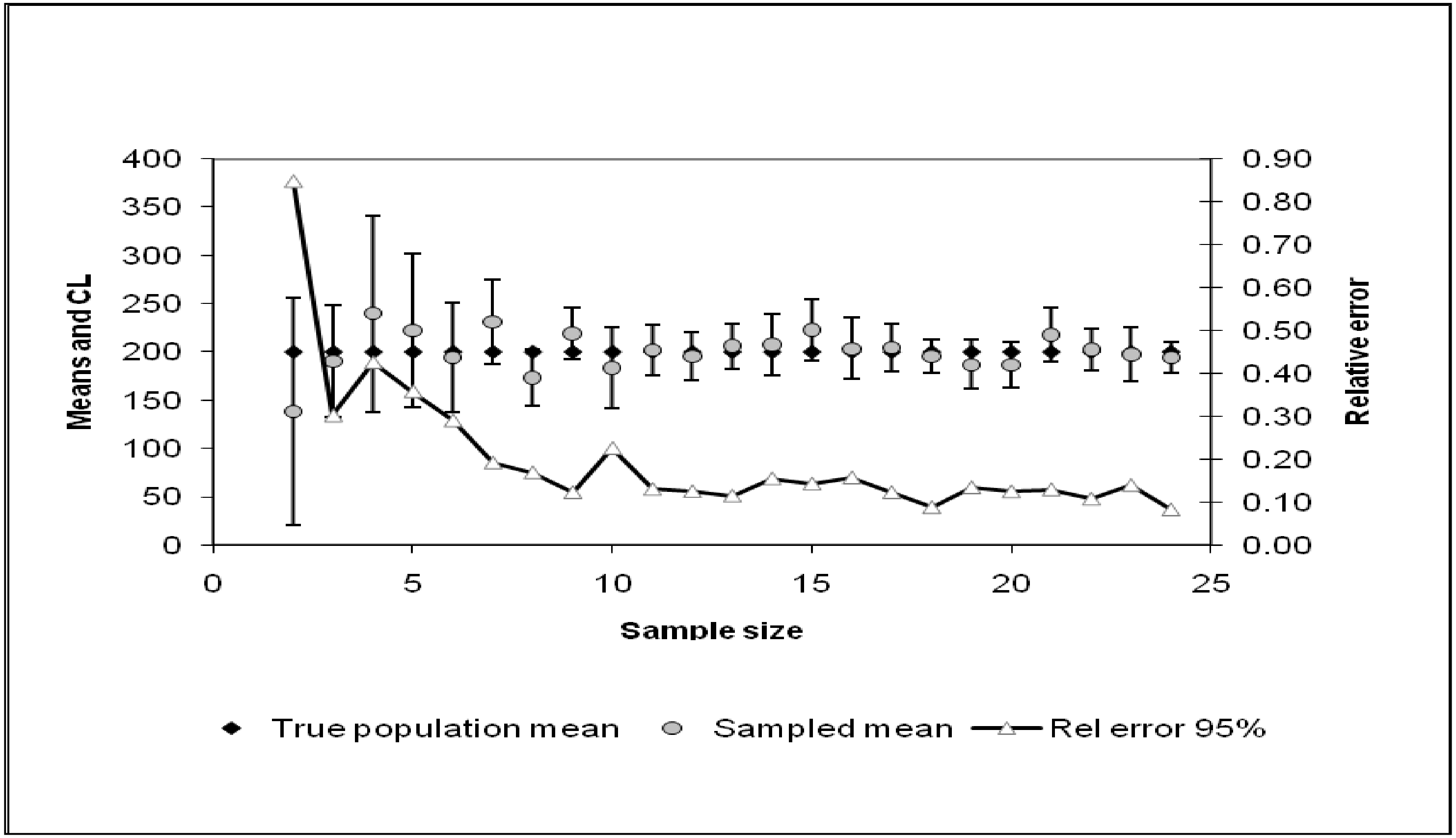
4.10 Relative error (2/2)

The relative error is determined by the characteristics of the variance of the target population.

A target population with a high variance will need more samples to obtain a good estimate.

At a certain point, increasing sample size has little impact on the relative error, and increasing the sample size further is a **waste of staff time and funds (human and financial resources)!**

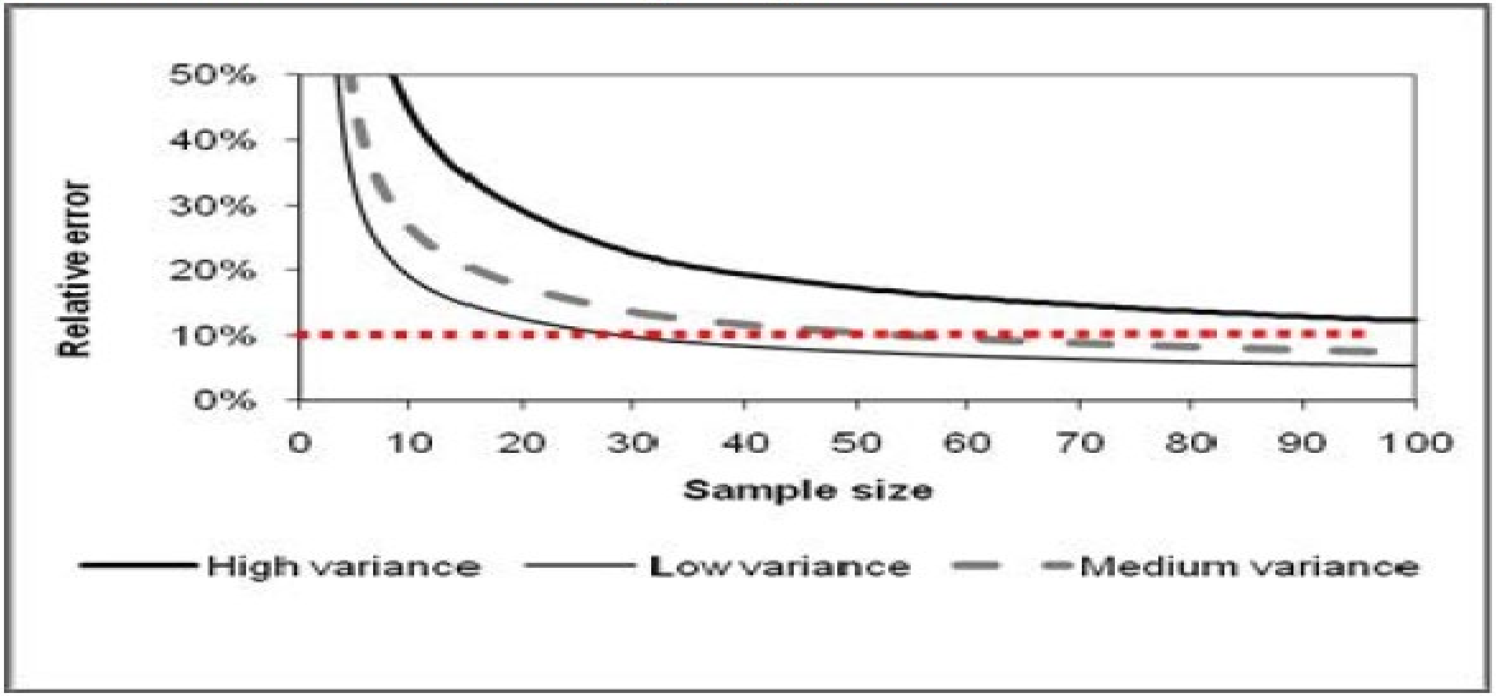
4.11 CL, relative error and sample size



4.12 Precision of estimates (1/7)

The size of the **relative error** is determined by the variance of target population and the sample size. A higher variance produces a small relative error (such as 10%) at higher sample sizes. A low variance achieves a small relative error (such as 10%) at small sample sizes.

Relative error and increasing sample sizes



4.12 Precision of estimates (2/7)

The **relative error is calculated** using the confidence limits taken as absolute values, that is, without the plus and minus signs and dividing by the mean.

Item	Note
Relative error	$MaxRelativ\ eError\ \%, \varepsilon = \frac{ConfidenceLimit}{\bar{y}} = \frac{t_{n-1}s}{\bar{y}\sqrt{n}}$
Uses	– The relative error formula is used to determine the sample size required from the population that meets a desired relative error, that is, the relative error that can be tolerated
	– determining sample size using the relative error formula requires some prior information on the population for which statistics are to be calculated: the variance/standard deviation and the mean
	– in the next module, we shall see how to determine sample size using the relative error formula – first we will examine, on the basis of this formula, the relationship between relative error and variance, involved in determining sample size

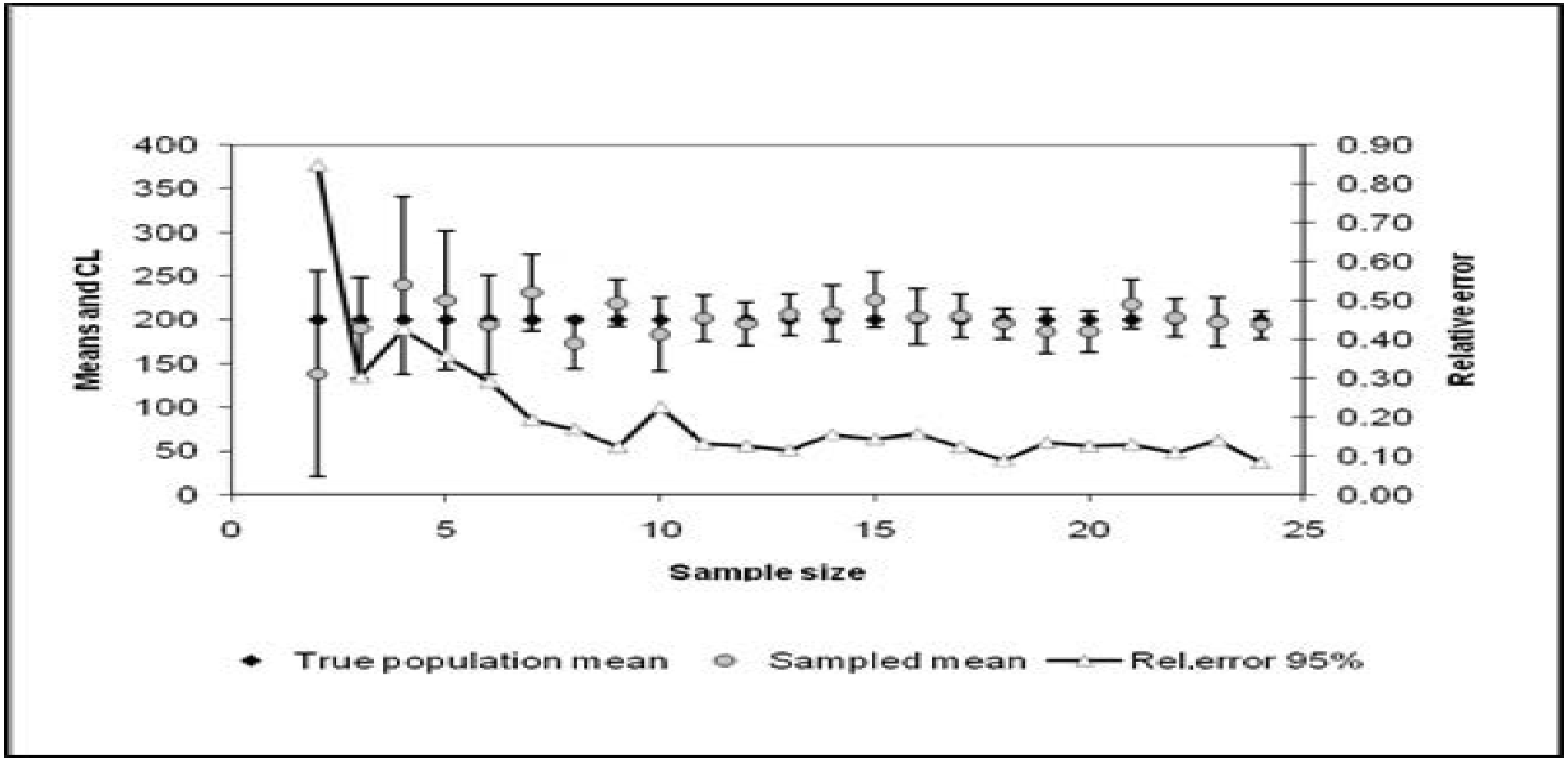
4.12 Precision of estimates (3/7)

- See, in the following figure, how relative error decreases as sample size increases
- See also how the estimated sample mean deviates around the true population mean. The true population mean lies within the confidence intervals of a 95 percent confidence limit.
- A maximum relative error of 10 percent, calculated with a 95 percent confidence limit, indicates a probability of 95 percent that the true mean deviates by no more than 10 percent from the value of estimated mean
- Usually, a low relative error size (precision) is desired (10 percent and below).

4.12 Precision of estimates (4/7)

Note that the relative error decreases from approximately 80%, at low sample sizes, to 10 to 15% with a sample size greater than 10

Example of sample size and relative error



4.12 Precision of estimates (5/7)
Lessons to remember

- From the relationship between relative error, sample size and variance just examined, we should draw the following lessons:
  - **Lesson 1:** increasing sample size reduces the relative error; that is, it increases the precision of estimates. However, the gains occur at low sample sizes. Above a certain sample size, the reduction in relative error size is not great. Thus, **taking more samples will increase operating costs** but will not improve the precision of your estimates to a significant degree.
  - **Lesson 2: Low relative error**, that is, a high precision of estimates is gained quickly at lower sample sizes when the variance or standard deviation of the target population is low. Thus, when measures are taken to reduce variation, such as **stratification** before sampling, there is a gain precision at lower sample sizes, which reduces the costs of conducting the survey.

4.12 Precision of estimates (6/7)
Lessons to remember (example)

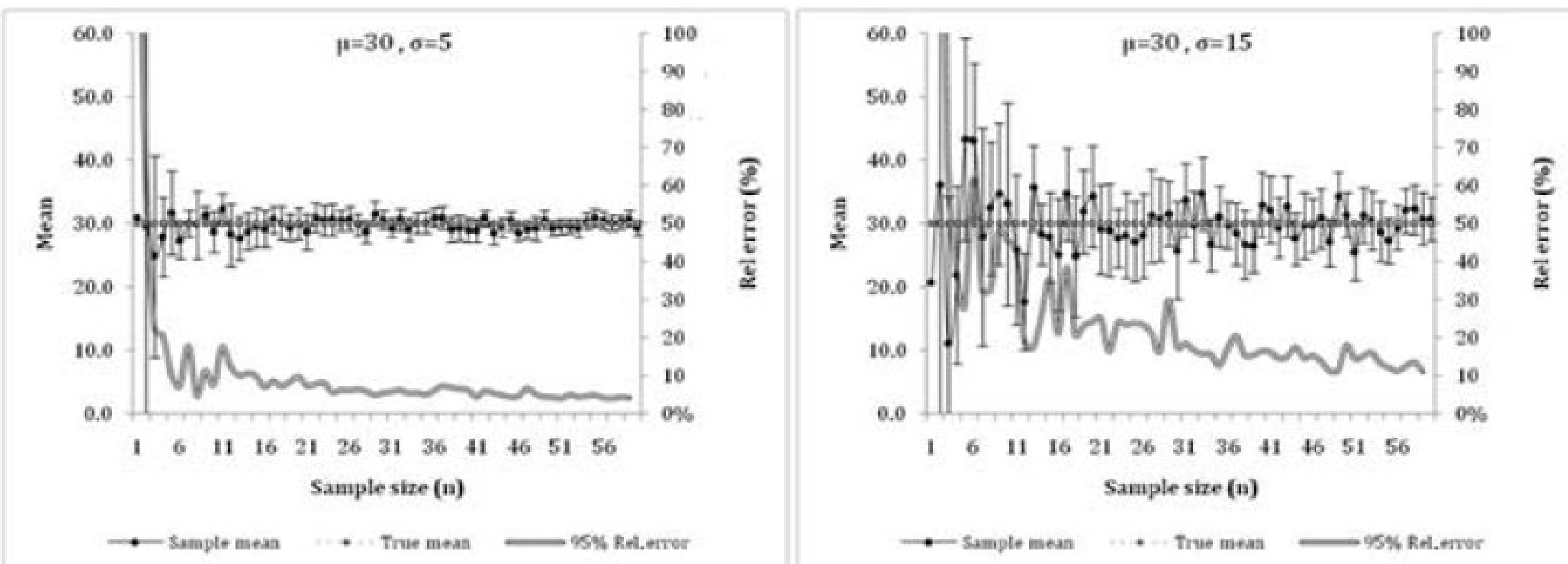
Below are two real-life cases of SSF the data of which was plotted on a graph to show the relationship between variance, sample size and relative error:

- **Case 1:** 1 000 canoes with very similar gears and number of gears, which results in similar daily catches or a low variance of catches in the target population. The mean daily catch of the target is 30 kg.
- **Case 2:** 800 canoes operating different gears, which results in very different daily catches or a high variance of catches in the target population. The mean daily catch of the target population is 30 kg.

4.12 Precision of estimates (7/7)
Lessons to remember (example)

The first sample, because of its low variance, can achieve a relative error lower than 10 % at a very early stage, at sample sizes below 10. In the second case, relative error below 10 % is achieved at sample sizes greater than 30 because of the high variance.

Relative error for sample obtained from two different target populations



Notes: 1) population mean =30, population variance = 25; 2) population mean =30, population variance =225.

5

Basics of SSF and aquaculture statistics data analysis

5.1 Comparing means of groups and strata (1/2)

Key issues

In SSF statistics, stratification is important when setting up a sampling scheme. Strata are used to define different groups with similar characteristics, such as canoe owners who use gillnets versus canoe owners who use traps. How can we know whether daily catch using gillnets differs significantly from daily catch using traps?

1. t-test	– used to compare means of two groups – each group consists of subjects that are normally distributed – the two-sample t-test procedure can be used to test whether the means are equal
2. Analysis of Variance (ANOVA)	– tests whether the means of several groups are all equal – generalizes student's two-sample t-test to more than two groups – possible to test whether three or more different types of fishing gear have equal daily catches

5.1 Comparing means of groups and strata (1/2)
Example

Is the CPUE of gillnets in the north different from gillnets in the south?

CPUE Gillnet north differs from CPUE Gillnet south

Two gear groups → so t-test

Step 1: Test if variance is equal

H_0: Variance CPUE north = Variance CPUE south

Do F test for variance.

$P \leq 0.05$ Then H_0 not correct → Variance not equal

$P > 0.05$ Then H_0 is correct → Variance equal

Step 2: Do t-test if average CPUE north = average CPUE south

Variance is equal → Do t-test assuming equal variance

Variance is not equal → Do t-test unequal variance

T-TEST (with equal or non-equal variance)

H_0: Averages are the same:

H_0: CPUE south = CPUE north

$P \leq 0.05$ H_0 is not correct, CPUE not the same

$P > 0.05$ H_0 is correct, CPUE the same.

ANOVA

H_0: CPUE of all gear groups are the same

$P \leq 0.05$, CPUE gear groups are not the same

$P > 0.05$, CPUE gear groups are the same

5.2 Survey weights and estimation

- **General steps**
  - First, determine the statistical unit's **base weight:** the inverse of the statistical unit's selection probability
  - **(if needed)** adjust base weight to consider possible nonresponse or make the estimates conform to a known population total
  - compute final survey weights for each responding unit as a product of the base weight, the nonresponse adjustment, and the population weighting adjustment
  - use final survey weights in estimation of formulas to produce valid estimates of population parameters

5.3 Adjusting survey weights (1/2)
Example

- For small agricultural holdings and households, the most common form of nonresponse weighting adjustment is a weighting class adjustment.

- The full sample of respondent households and nonrespondent households is divided into a number of weighting classes or cells; nonresponse adjustment factors are computed for each *cell* denoted by the letter *'c'*, as

$$w'_c = \frac{\sum_{i \in rc} w_{di} + \sum_{j \in mc} w_{dj}}{\sum_{i \in rc} w_{di}} = \frac{\sum_{i \in sc} w_{di}}{\sum_{i \in rc} w_{di}}$$

- The denominator of w'_c is the sum of the weights of respondent households (indicated by letter 'r') in cell *c.* The numerator adds the sum of the weights for respondent households and the sum of the weights for eligible non-respondent households (indicated by letter 'm' for missing) in *cell c.*

5.3 Adjusting survey weights (2/2) *Example*

- The two sums in the numerator yield the sum of the weights for the total eligible sample (indicated by letter ‘s’) in cell *c*
- Thus, the nonresponse weight adjustment **w’c** is the inverse of the weighted response rate in cell *c*
- A low response rate results in having to make a large nonresponse adjustment for the cell
- This significantly lowers the precision of survey estimates
- **Therefore, attempts must be made to avoid high nonresponse rates**

5.4 Estimation of population total

- The total is estimated by

$$\hat{Y} = \sum w_i y_i$$

$\hat{Y}$ = estimate of population total

y_i = value of variable *y* for respondent *i*

$w_{i\,=}$ the final weight for respondent *i*

5.5 Estimation of population mean

- The population mean is estimated by

$$\bar{y} = \sum w_i y_i / \sum w_i$$

$\bar{y}$ = estimate of population mean

y_i = value of variable *y* for respondent *i*

w_i = the final weight for respondent *i*

5.6 Estimation of population ratio

- The population ratio $R = Y/X$ is estimated by

$$r = \sum w_i y_i / \sum w_i x_i$$

where
x_i = value of variable x for respondent i
y_i = value of variable y for respondent i
$w_{i\,=}$ the final weight for respondent i

6

Statistical software for computing statistics

6.1 Statistical software for computing statistics

- Statistics include means, standard deviations, coefficients of variation, confidence limits, means comparison using t-tests and ANOVA
- These parameters, when computed, support sampling design, sample size determination and sample allocation
- They also support the analysis of data, such as estimates of population totals, means and ratios, as well as the computation of indicators such as CPUE
- These calculations can be done using any standard statistical software which ensure calculation of the above statistics (**Microsoft Excel, Stata, SPSS, R, E-views, SAS etc.)**
- The choice of software depends on the skills of the technical officer and level of familiarity with the software

References

- **de Graaf, G.J., Nunoo, F., Ofori Danson, P., Wiafe, G., Lamptey, E. & Bannerman, P.** 2015. *International training course in fisheries statistics and data collection.* FAO Fisheries and Aquaculture Circular No. 1091. FAO Publication: Rome.

Thank you