



USP

---

## **Biostatistics III: Sampling strategies**

Data collection for fisheries assessment:  
Monitoring and sampling strategies

- “Optimistic” approach → Precautionary approach
  - *“States should apply the precautionary approach widely to conservation, management and exploitation of living aquatic resources in order to protect them and preserve the aquatic environment. The **absence of adequate scientific information** should not be used as a reason for postponing or failing to take conservation and management measures”*
  - *“States **should promote the use of research results** as a basis for the setting of management objectives, reference points and performance criteria, as well as for ensuring adequate linkage between applied research and fisheries management”*

- In implementing the precautionary approach, states shall:
  - *"a) Improve decision-making for fishery resource conservation and management by **obtaining and sharing the best scientific information available** and implementing improved techniques for dealing with risk and uncertainty;"*
  - *"d) **Develop data collection programmes** to assess the impact of fishing on non-target and associated or dependent species and their environment, and adopt plans, which are necessary to ensure the conservation of such species and to protect habitats of special concern."*

UN Conservation of straddling & highly migratory fish stocks

**Nation specific**

- Analysis of fisheries, resource status and trends:
  - Catch, landings and discards
    - A fundamental variable of interest
  - Fishing activity (effort)
    - biological assessment:
      - Fishing mortality =  $f(\text{Effort})$ , Fishing mortality = constant \* Effort?
      - Yield =  $f(\text{Effort})$
    - economic assessment:
      - cost of fishing =  $f(\text{effort})$ , cost of fishing = constant \* Effort?
    - socio-cultural assessment
      - number of jobs =  $f(\text{effort})$ , number of jobs = constant \* Effort?
  - Species, catch and stock composition
    - Biological state of the resources:  $C = FB$ 
      - Stock size, production, mortality, growth, ...

- The sampling from a population should be made at the highest resolution possible
  - Effort:
    - Number of hours fished, number of sets, ...
    - Days fishing
    - Days absent from port / days at sea
    - Number of registered boats
    - ...
  - Catch
    - Preferable each setting, where we also get catch information
    - Total catch by a certain strata (Area, fleet, province, national, ..)
    - ...
  - Location
    - Latitude, longitude
    - Statistical rectangle
    - Fishing area
    - Political boundaries (Provinces)
    - ...

# Can it be collected?

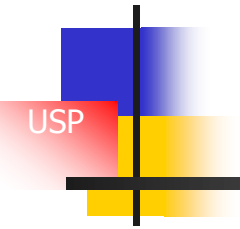
- Collection of data cost money
  - Money is out of “our” control
    - → need to compromise
    - → need to prioritize
    - end result is optimization
  - Optimization
    - Is somebody else out there collecting or in need for the same data as I am?
      - Governmental agencies
        - Total landings, vessel registration, ...
      - Private companies
        - Total yield going through a market, ...
      - NGO's
        - Biological, economical and socioeconomic data
      - Fishermen
        - They THE data collectors and are sometimes the best biologist
    - If so, team up!
      - Within different governmental agency: Not a question
      - Fishermen: Create a cooperation
      - Between agencies: Create a liaison & trust through dialog

# Data collection strategy

- Given the current logistic and monetary resources “we” must optimize.
  - Theoretical scientific approach can help:
    - Optimization theory = stratification
    - How many samples are needed?
  - Part of this will always be pragmatic:
    - Compromise with other objectives already within the system
    - Theoretical approach is, well just theoretical approach!
  - Part of this we can never do!
    - Monitoring cost  $\ll$  Monetary value from the fisheries ?

Next few slides, the science part →

- Fish distribution and fishing activity is not a random process:
  - Spatial
    - Species specific habitat, life cycle
    - Fishing grounds
    - Landings sites
  - Temporal
    - Season, month, week, days, night and day
  - Vessel and gear
    - Fleets
    - Gear
  - ...
- Need to take this into account when thinking about sampling design

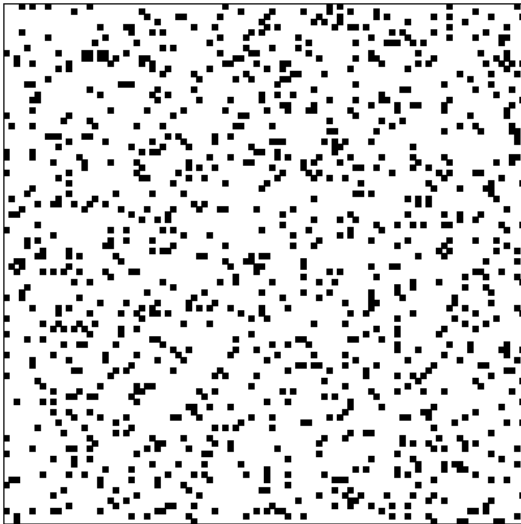


USP

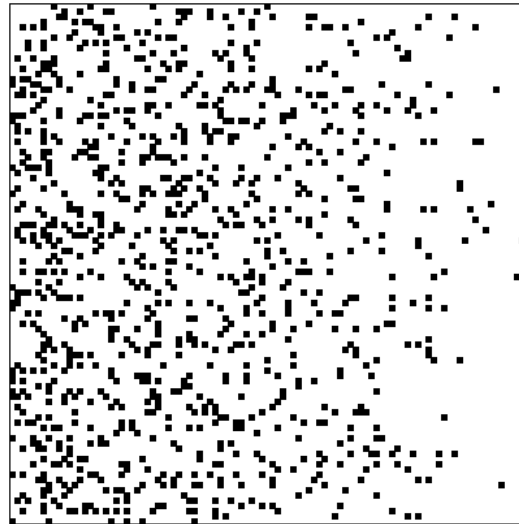
# Sampling design

# Population distributions

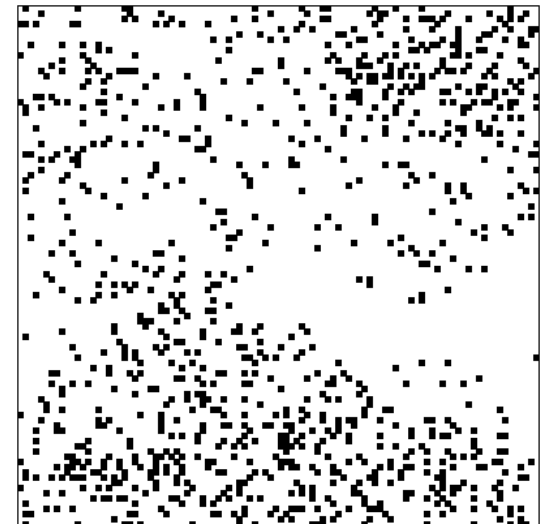
**Random  
(homogenous)**



**Gradient**

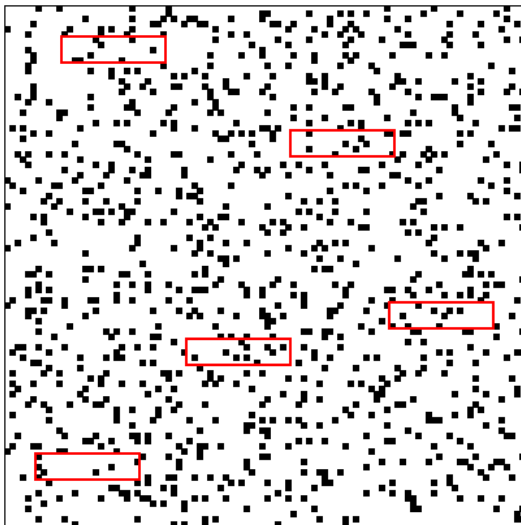


**Patchy  
(heterogeneous)**

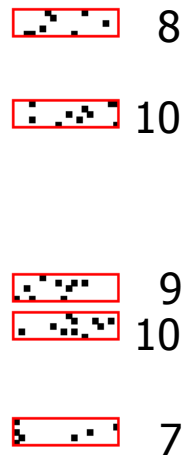


- What type of sampling design gives the most reliable estimation of the population density and population size?

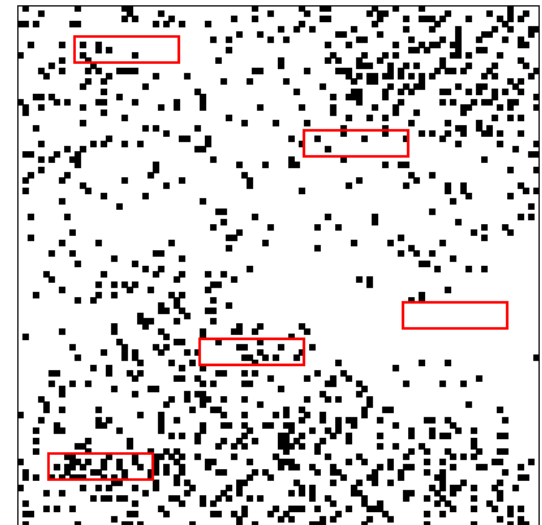
## Homogenous population



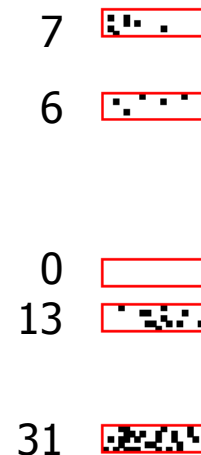
### Sample



## Heterogenous population



### Sample



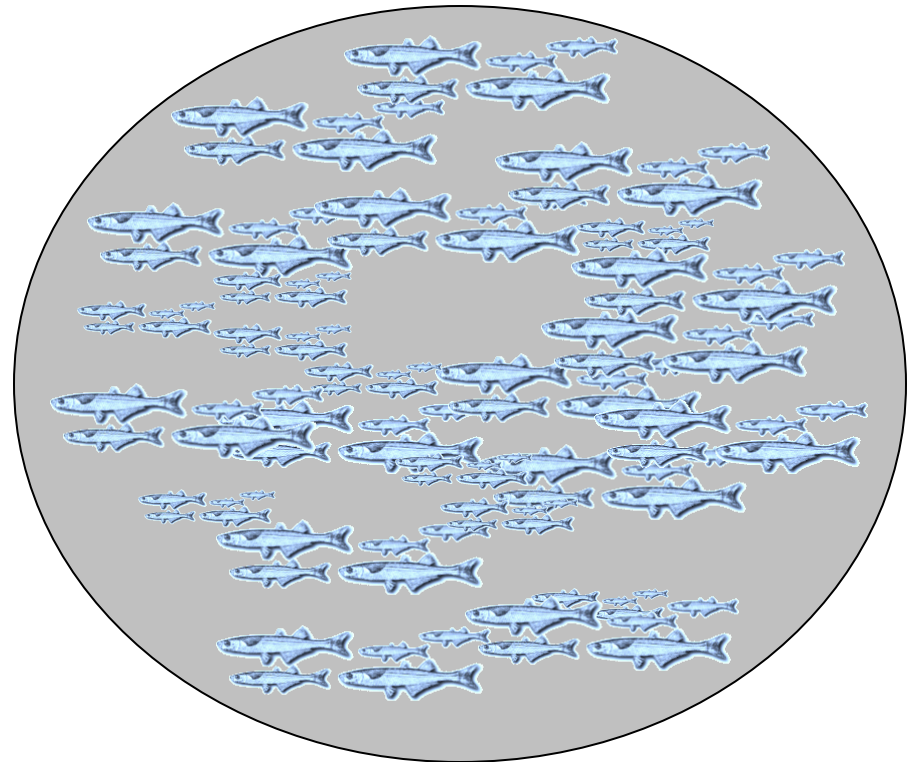
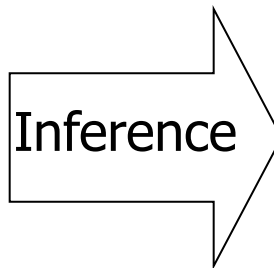
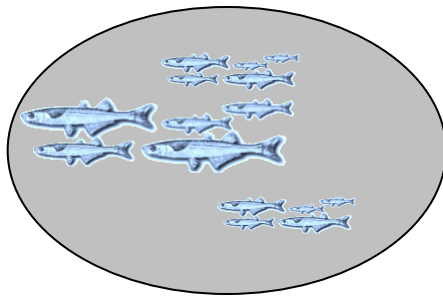
- The number of individuals in a sample are highly variable in random samples from a population that is heterogeneously distributed.
- Such survey sampling design will lead to high variance estimates, meaning that the precision of the mean population estimates is very low.

Each dot represents an animal

- Use information from a sample to make inference about the population

**I'm interested what this is**

**Sample is known**



Can only make inference about the population from the sample if the sample is representative of the population

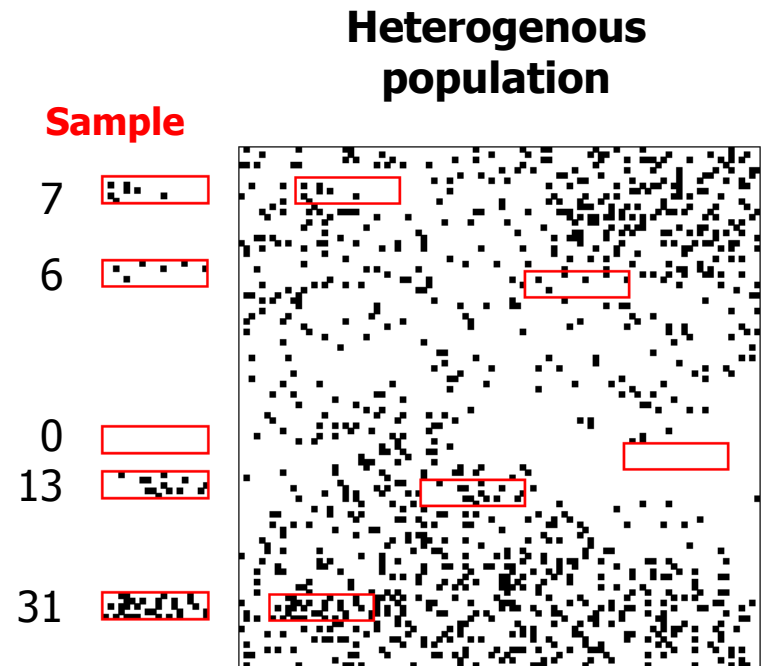
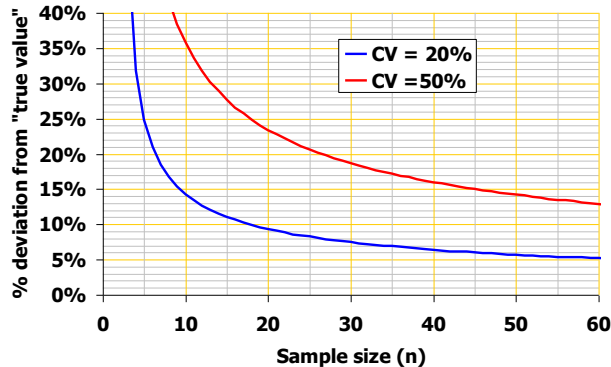
# So what confidence do we have in our data?

**There 95% probability that the interval contains the true population mean value**

	i	Random	Hereogeneous
	1	8	7
	2	10	6
	3	9	0
	4	10	13
	5	7	31
	n	5.00	5.00
	Mean	8.80	11.40
	Standard deviation	1.30	11.89
	CV	0.15	1.04
	CV%	15%	104%
	Standard error	0.58	5.32
	$t_{n-1,0.05}$	2.78	2.78
	Lower 95% Conf.interval	7.18	-3.36
	Upper 95% Conf.interval	10.42	26.16

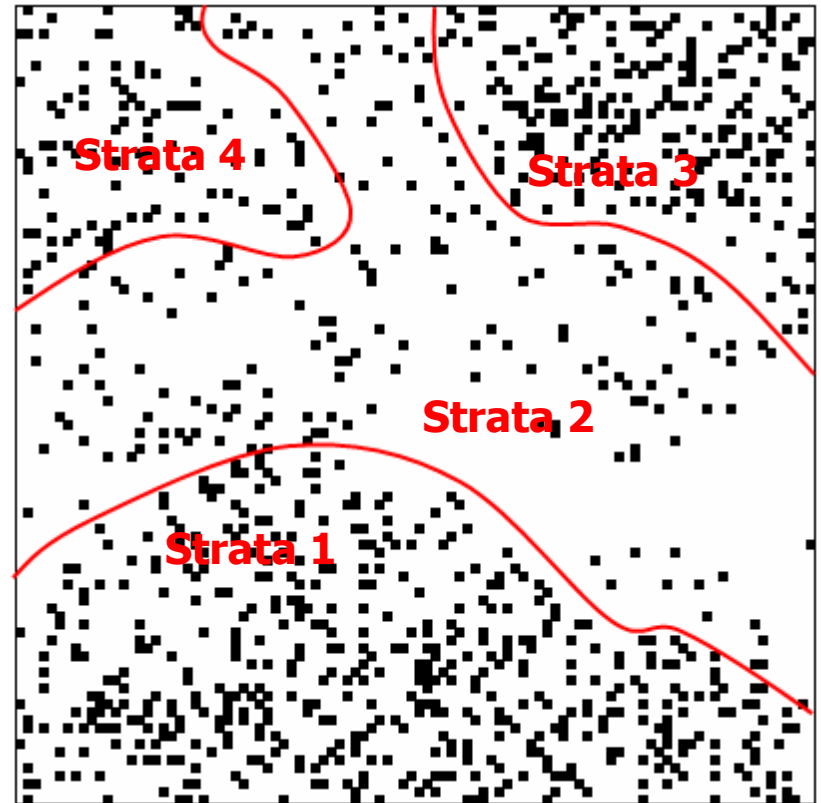
# What are the remedies?

- We could of course increase the sample size.
  - Given the population is distributed like on the left we probably should anyhow.
  - Increasing sample size has to be weighted against increasing cost
  - Theory can **help** in determining sample size:



- But given the heterogenous nature of the data we may need to think about an additional element.

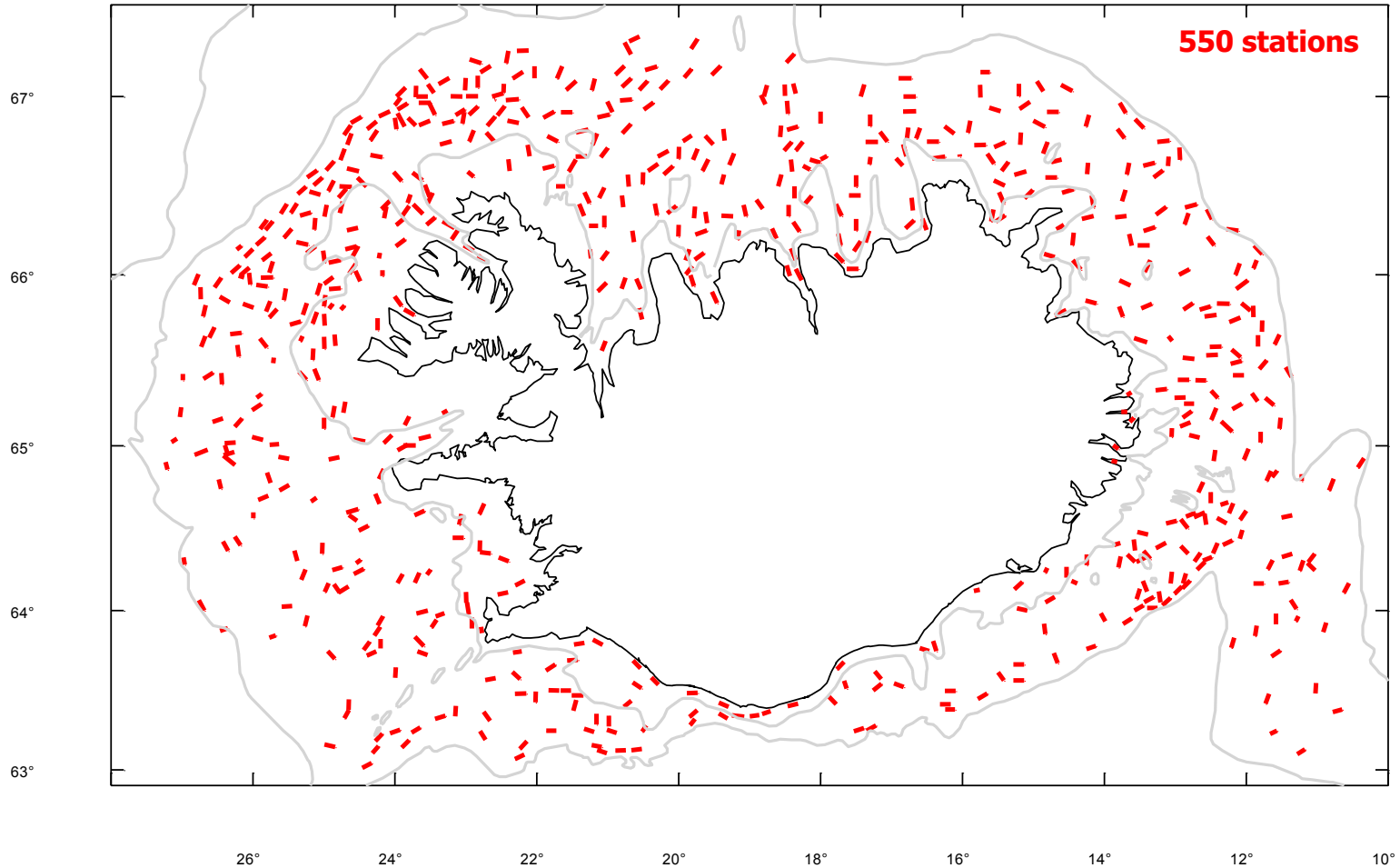
- If population is heterogeneously distributed, then one should divide the sampling area into areas (strata) which are more homogenous.
  - Sampling within each strata should be random
  - By estimating the mean and the variance within each strata and then combining strata values to obtain overall values normally results in reduced variance estimate.
  - Note one strata does not to be geographically continuous
    - Should we make strata 1 & 3 a single strata.



Example of 4 strata allocation based on density distribution

# Station location of Icelandic spring survey

The distribution of sampling stations is a result of a stratified random analysis.



# Allocation and number of strata (Surveys)

- Allocation of strata can be based on:
  - Spatial information of catch rates / density
  - Depth, topography
  - Substrate type
  - Information from other studies
  
- The number of strata depend on:
  - How heterogeneous the population is
  - The total number of stations feasible
  - Sample size within stratum needed to obtain the desired precision
  
  - Generally very case specific

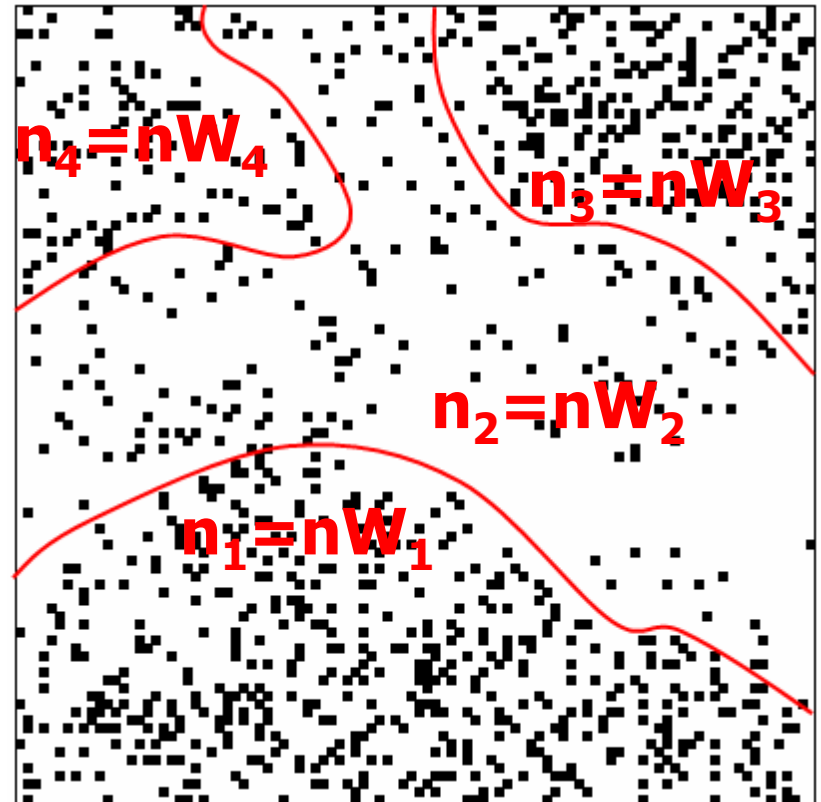
# Sample allocation to strata 1: Area

- Let  $A_j$  stand for area size of stratum  $j$  and  $A_T$  the total survey area. Then the relative weight of strata  $j$  is:

$$W_j = \frac{A_j}{A_T}$$

- Distributing the total sample size ( $n$ ) according to strata weight ( $W_j$ ) gives the number of samples in strata  $j$  by:

$$n_j = nW_j$$



$$n = n_1 + n_2 + n_3 + n_4$$

## Example 1: Area allocation of samples

- Lets imagine that we have 3 strata of known size and that we can only take 20 samples.
- So how do we allocate my 20 samples to the 3 strata?

	Strata 1	Strata 2	Strata 3	Total
Area j	40	60	100	200
Weight j	0.2	0.3	0.5	1
Sample size (nj)	4	6	10	20

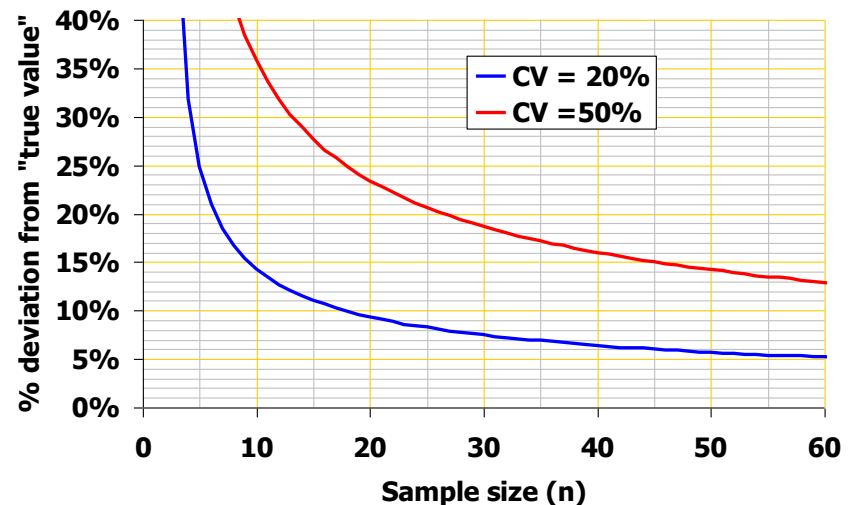
$$W_j = \frac{A_j}{A_T}$$

$$n_j = nW_j$$

## Sample allocation to strata 2: Variance

- If prior information other than strata size are available these can be included in the weighing factor. These normally include:
  - Prior variance estimates
    - As a general rule the sampling intensity should thus be **higher in areas where the variance is greater**
    - Often found that greater variance is observed where density is higher. Thus sampling intensity is often proportional to density

Why? Where the variance is higher we need larger sample size to get a lower confidence interval →



## Example 2: Take variance into account

- Larger variance = larger sample size

$$A_{ej} = A_j \sigma_j$$

$$W_j = \frac{A_{ej}}{\sum_i A_{ej}}$$

$$n_j = nW_j$$

	Strata 1	Strata 2	Strata 3	Total
Area j	40	60	100	200
Std. Deviation j	30	40	5	
Aej	1200	2400	500	4100
Weight j	0.29	0.59	0.12	1
Sample size (nj)	6	12	2	20

# Sample allocation to strata 3: Including cost

- Higher cost of sampling in a strata, lower samples size
  - Distance to location, depth, ..
- Sample number in stratum as a function of stratum size, sample variance and sample cost:

$$A_{ej} = \frac{A_j \sigma_j}{\sqrt{C_j}}$$

$$W_j = \frac{A_{ej}}{\sum_j A_{ej}}$$

$$n_j = n W_j$$

Symbol legends:

$n_j$  Number of samples in stratum  $j$

$n$  Total number of samples in survey

$A_j$  Area of stratum  $j$

$\sigma_j$  Standard deviation of stratum  $j$

$C_j$  Cost per sample in stratum  $j$

## Example 3: Take cost into account

- Higher cost of sampling = smaller sample size

$$A_{ej} = \frac{A_j \sigma_j}{\sqrt{C_j}}$$

$$W_j = \frac{A_{ej}}{\sum_j A_{ej}}$$

$$n_j = nW_j$$

	Strata 1	Strata 2	Strata 3	Total
Area j	40	60	100	200
Std. Deviation j	20	30	5	
Cost	5	15	25	
$A_{ej}$	358	465	100	923
Weight j	0.39	0.50	0.11	1
Sample size (nj)	8	10	2	20

# Example summary

## Area size only

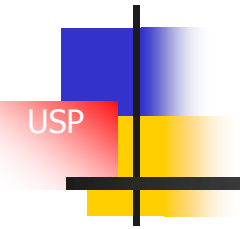
	Strata 1	Strata 2	Strata 3	Total
Area j	40	60	100	200
Weight j	0.2	0.3	0.5	1
<b>Sample size (n<sub>j</sub>)</b>	4	6	10	20

## Add variance

	Strata 1	Strata 2	Strata 3	Total
Area j	40	60	100	200
Std. Deviation j	30	40	5	
A <sub>ej</sub>	1200	2400	500	4100
Weight j	0.29	0.59	0.12	1
<b>Sample size (n<sub>j</sub>)</b>	6	12	2	20

## Add cost

	Strata 1	Strata 2	Strata 3	Total
Area j	40	60	100	200
Std. Deviation j	20	30	5	
Cost	5	15	25	
A <sub>ej</sub>	358	465	100	923
Weight j	0.39	0.50	0.11	1
<b>Sample size (n<sub>j</sub>)</b>	8	10	2	20



# Calculations of stratified mean and standard error

- The formulas

$$\bar{y}_i = \frac{\sum_{i=1}^{n_i} y_{i,j}}{n_i}$$

$$\bar{y}_{st} = \sum_{j=1}^L W_j \bar{y}_j$$

- The legends:

$y_{i,j}$  : catch of the  $i^{\text{th}}$  tow in stratum  $j$

$n_j$  : number of tows in the  $i^{\text{th}}$  stratum

$y_j$  : mean catch rate in the  $j^{\text{th}}$  stratum

$y_{st}$  : estimated stratified mean density in the entire area

$W_j$  : weight of stratum  $j$

# The variance of the stratified mean

- The formulas

$$s_j^2 = \frac{\sum_{i=1}^{n_j} y_{i,j} - \bar{y}_j}{n_j - 1}$$

$$V \bar{y}_{st} = \sum_{j=1}^L W_j^2 \frac{s_j^2}{n_j}$$

$$SE_{\bar{y}_{st}} = \sqrt{V \bar{y}_{st}}$$

- The legends:

$s_j^2$  : variance in stratum j

$V(y_{st})$  : Estimated variance of the stratified mean

$SE_{y_{st}}$  : Standard error estimates of the stratified mean

# The confidence interval

- If the sample is large enough, then there is a 95% chance that the true mean lies in the interval:

$$\bar{y}_{st} \pm t_{n-1} SE_{\bar{y}_{st}}$$

- The t is from the Student's distribution with n-1 degrees of freedom and  $\alpha=0.025$  (this alpha level corresponds to getting 95% confidence interval)

# Example: Annual landings by harbors

100 landings sites in total (the total population)

Stratum 1 (large landings)											
Y1,i	45	59	87	41	71	25	9	69	10	7	
Stratum 2 (medium landings)											
Y2,i	17	13	19	26	1	8	27	11	12	26	
	5	8	10	16	16	4	16	16	13	29	
	14	25	29	27	20	25	2	7	3	12	
Stratum 3 (small landings places)											
	2	6	7	0	1	2	1	5	4	7	
	8	9	3	2	5	4	2	0	2	8	
	5	3	8	9	8	9	1	6	5	3	
	3	4	7	5	5	3	2	4	6	1	
	6	2	5	1	0	3	8	0	4	3	
	3	5	5	0	7	0	9	7	9	0	

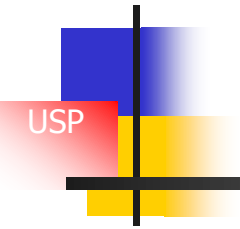
Can only afford to take 20 samples annually: Could take randomly, relative to the number of harbours (n) or optimize by also into account the variance ( $s^2$ )

Stratum #	n	Variance	Sampling strategy				
			Stratum #	Random	Proportional	Optimum	
1	10	835	1	??	2	8	
2	30	73	2	??	6	7	
3	60	8	3	??	12	5	
Sum	100						
Stratified standard error					3.06	2.10	1.20

# Advantages of stratification

- Increase in precision in estimates given a certain sampling intensity
- By stratification we reduce the variance of the samples within each strata relative to the total variance of all the samples
- Usually provides more detailed information on distribution and abundance than simple random sampling because stratification ensures that samples are dispersed over the whole survey area
  - Makes stratification useful even when dealing with basically homogenous distributions (can still be patchy)

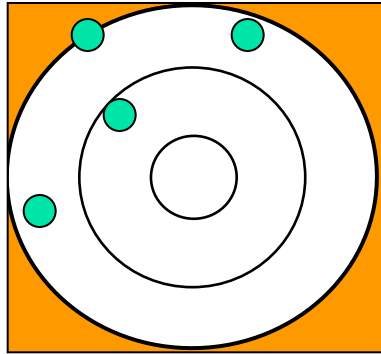
- General rule
  - Should be large if the stratum (area, numbers) is large
  - Should be large if the variance is large
  - Should be large if the sampling is inexpensive



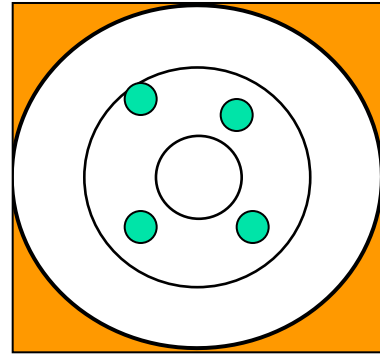
USP

## Bias vs. variance estimate

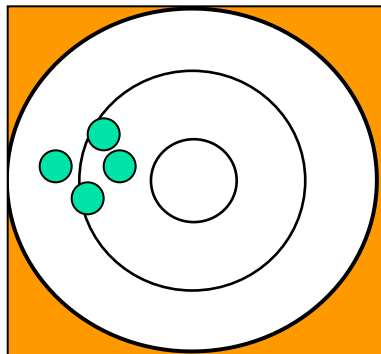
# Accuracy = Precision + Bias



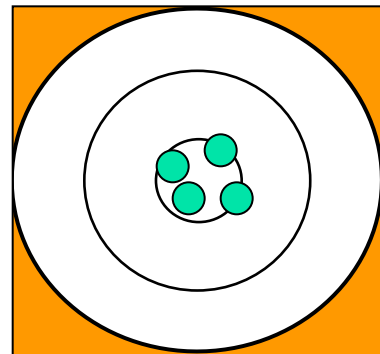
Not accurate and not precise



Accurate but not precise  
(Vaguely right)

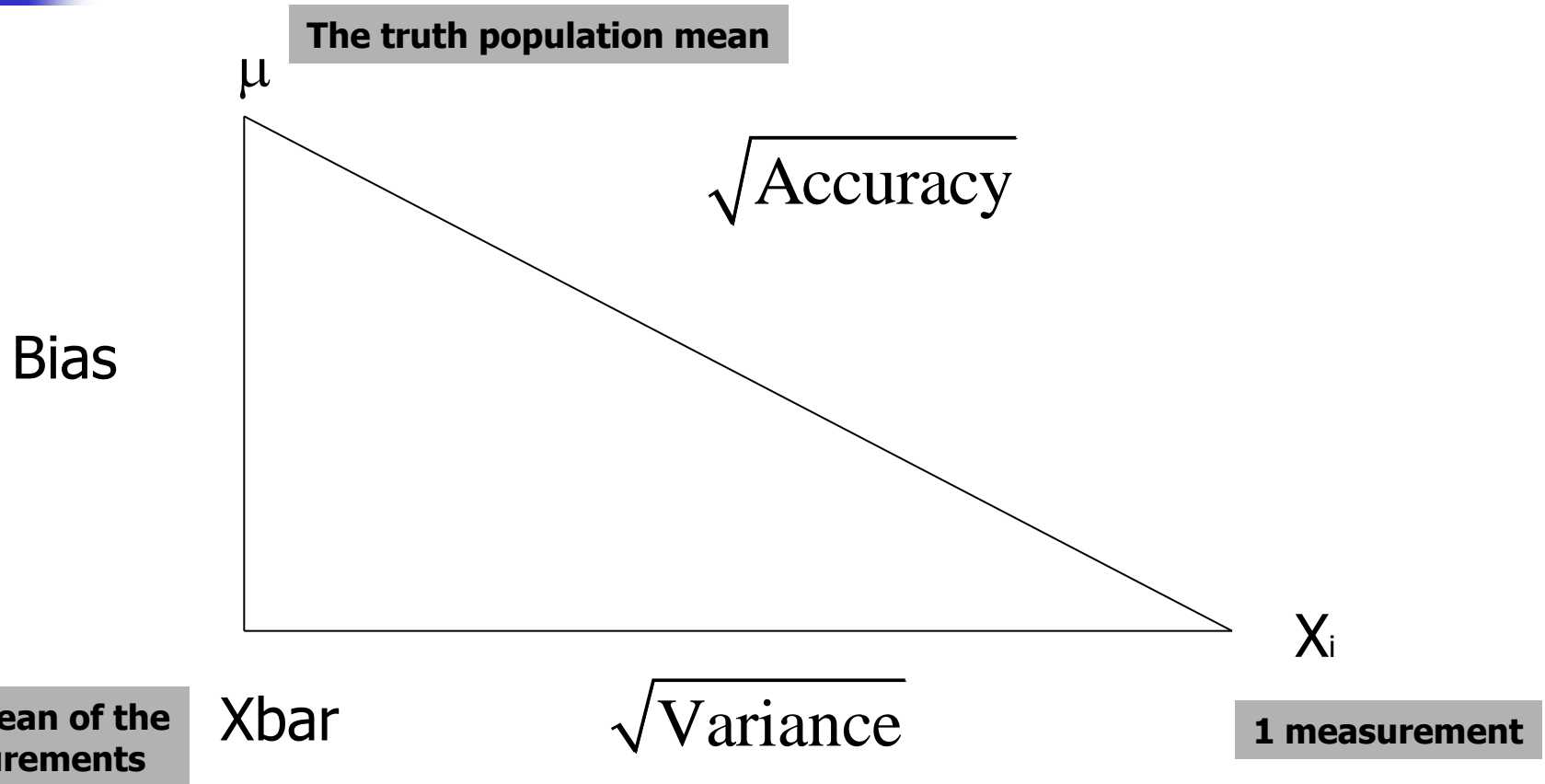


Precise but not accurate  
(Precisely wrong)



Accurate and precise

# Relation between bias, accuracy and variance



$$\text{Accuracy (MSE)} = E (x_i - \mu)^2 = \text{Variance} + \text{Bias}^2$$

$$x_i - \mu^2 = x_i - \bar{x}^2 + \bar{x} - \mu$$

- Variance:
  - Can be measured directly from the observations
  - Is a measure of the distribution of the data, given the data!
  - Can be reduced by stratification and by increasing sample size
- Bias:
  - Can not be estimated from the observations, since we need to know the "truth",  $s^{\text{th}}$  which in most cases is impossible
  - Could get an idea by comparing estimates from independent measurements
    - Seldom done
  - Can potentially be reduced by stratification
  - Can **not** be reduced by increasing sample size

- The sampling design is always a compromise between different objectives
  - Sampling theory can however always help
- Take it for granted that your samples are always biased
  - Meaning: Try to get a independent estimate of your statistics once in a while