Sampling Exercise

Extended presentation with notes and results

Alan Brown

Eurosite 5th Natura 2000 Monitoring Workshop, Part 2
Remote Sensing Support Group, Zagreb 23-25 May 2022

Notes

The exercise used a population of smooth, washed stones collected so that they have a smooth, continuous weight distribution from large to small.

Each stone was weighed (primary variable) and measured along the longest axis (secondary, auxiliary variable).

The stones were laid out on a small table, too small to completely spread out every stone, and delegates were given a kitchen scale suitable to accurately weigh single stones – but not big groups of stones at once.

The instructions were given to form groups and weigh 6 stones in a competition for a prize. The actual weight of the stones was also 'accidentally' shown very briefly, to test for confirmation bias.

All simulated draws by AB using macros in Excel

Many thanks to **Marina, Mirjana, Matej** and **Lea** for collecting the stones and making the measurements!

Notes

The workshop focused on observer variation in ground ('in situ') observations and new technologies, notably remote sensing using drones. Here we have a contrast between a small sample of – possibly uncertain – but very targeted and specific measurements versus complete coverage of less specific wall-to-wall observations.

This sampling exercise was meant to look at bias, precision and uncertainty in estimates using partial data, working with a very small sample selected using our judgement from a population which can be see and, to some extent, evaluated qualitatively during the selection process. This is designed to be similar to the common situation in vegetation recording where we have a more or less homogenous area of habitat and try to set out a 'representative' set of plots or quadrats. Notice that the equivalent of 'in situ' measurements, the weighing process, is accurate and repeatable so we are only looking at sampling issues.

The design of the exercise is based on similar demonstrations done in the USA in the 1920s and 1930s when the use of partial data, sampling and what is meant by a 'representative' sample was controversial. There is a good discussion of the history of sampling methods in Tillé (2020) – see further reading.

Please feel free to contact me on alb115@aber.ac.uk if you have any questions about the exercises or simulations, or want a copy of the spreadsheet data.

Notes

Representative plots are often chosen and recorded one at a time, that is, where the observations from earlier plots can influence the choice of later plots. In this case, if we already have in mind a particular result – perhaps we consider the stand of vegetation to meet a certain standard – it is tempting to choose later sample points which compensate for what we see in the earlier plots to try and get the result we expect, sometimes called **confirmation bias**. Of course this might also influence the measurements themselves, notably cover estimates, but here we are only considering the selection of sample units.

Confirmation bias is possible where, for example, we are re-visiting plots to look at changes, and happen to have both the earlier results and a preconception of whether or not we think changes have taken place, tempting us to minimise or discount any chance differences when we make observations. In this exercise, we were interested in whether knowing the result (the population mean weight) influenced the groups selection of stones in a situation where the groups were not certain whether this information was part of the exercise, noting that a second-order effect of confirmation bias might tempt us to add back in some variability to make the results seem more plausible...

Notes

As well as the group competition, we carried out a set of simulated random draws from the population weights and measurements in order to show a) the statistical properties of small samples, and b) how an auxiliary variable (length) can be used to increase the efficiency of sampling and precision of estimation for our variable of interest (weight), when it at least some predictive power.

In the first demonstration, sample sets of stones were drawn at random with **equal probability**, so each stone had the same chance of being included in the sample. This was done 'with replacement' – so each stone is available to be included more than once – but with such a small sample from a large population the results are similar without replacement, as in the group exercises.

A simulated draw of 1000 samples, gradually increasing sample size, shows how these **probability samples** always give (on average) an unbiassed estimate of the mean. However, the frequency histograms showing the 'sampling distribution of sample means' for smaller sample sizes is rather skewed and strongly influenced by the frequency distribution of few large stones and many small stones in the population. As sample size increases, the distribution of sample means becomes more symmetrical, approaching a normal (Gaussian) distribution, and as sample size increases further this becomes narrower as the precision of the estimate improves. This shows how, regardless of the population distribution of the variable (weight), the sampling distribution can be modelled by a normal distribution if the sample size is large enough: in the literature, a figure of >30 is often quoted. This is the **central limit theorem**, and is what allows us to calculate confidence limits, carry out statistical tests and statistical power analyses using the properties of the normal probability density function.

Notes

However, a sample size of at least 30 plots is rather large, especially if these have to include rare species, and many monitoring projects will not have enough. We also have to consider how damaging the actual recording might be in sensitive habitats and for population which are easily disturbed.

In the second part of the demonstration, we showed how an auxiliary variable can be used to enable **unequal probability sampling**. In this approach, some correlated variable which is easy to measure can be used to improve the choice of a small sample for some variable that is more difficult or more time-consuming to measure. All sample units (eg stones) are available to be selected, but are given **inclusion probabilities** proportional to an appropriate transformation of the second, auxiliary variable (here, length^3 , since weight it proportional to the volume). The stones were drawn at random, but those with higher inclusion probabilities are more likely to be selected. The **estimator** for the mean weight is then slightly different, as the sample weights have to be divided by the inclusion probability before calculating the mean.

The frequency histogram of 'sampling distribution of sample means' of 1000 draws shows that this estimator is both unbiassed and far more efficient, and even with as few as 5 stones in the sample the curve is symmetrical and narrow – in other words the estimate of the population is more precise, and even with just a small sample size already has the properties of the normal distribution so we can be confident in carrying out statistical tests and power analysis. The formulas used to calculate mean and variance are called **Horvitz-Thompson** estimators and are extremely important in practical survey-sampling designs (see notes on further reading).

Sampling Exercise

In your Group:

1. Decide how you want to select stones while you wait for your turn

2. Select 6 (six) stones and weight them one at at time to get the sample mean – then if you can, calculate variance, standard deviation, standard error

3. Return the stones to the box / bag for the next group, let them know

4. Estimate 95% confidence intervals for the true population mean weight

5. Don't share your results or weigh more than one stone at once

6. Decide whether some or all of this process is legitimate and is the following rule for awarding a prize fair....?

The group with a **sample mean** closest to the true population mean wins a prize!

RESULTS

Stone-Agers	117.17
Petra	113.50
The Stone Roses	82.5
The Flintstones	53.83
Rolling stones	44.56
8 Stones	39.5 - ‘winner’
The Stone Roses 2	25

Actual mean weight = **35.85** grams

1	**Rolling stones**	Stone											95% Limits ?		
	Used 5 strata (classes)	**1**	**2**	**3**	**4**	**5**	**6**	Total	Mean	Variance	SD	SE	Lower	Upper	Calculation
		230	120	32	30	12	4								
		Large					Small								
	Estimated class weight	230	120	31		12	4								
	Estimated class size	13	100	60		125	125	423							
	Class size x class weight	2990	12000	1860		1500	500	18850	**44.56**						group
2	**Stone-agers**	Stone													
	Chose 6 stones	**1**	**2**	**3**	**4**	**5**	**6**								
		14	21	298	100	178	92	703	**117.17**	11449	107	43.6	**31.71**	**202.62**	group
		10643.36	9248.03	32700.69	294.69	3700.69	633.36		**117.17**	11444.1667	106.977412	43.6733455	31.57	**202.77**	AB
3	**Flintstones**	Stone													
		1	**2**	**3**	**4**	**5**	**6**								
		83	69	17	11	5	138	323	**53.83**						
		850.69	230.03	1356.69	1834.69	2384.69	7084.03		**53.83**	2748.16667	52.4229593	21.4015835	**11.89**	**95.78**	AB
4	**Petra**	Stone													
		1	**2**	**3**	**4**	**5**	**6**								
		283	143	139	96	9	11	681	**113.50**						
		28730.25	870.25	650.25	306.25	10920.25	10506.25		**113.50**	10396.7	101.964209	41.6267142	**31.91**	**195.09**	AB
5	**The Stone Roses**	Stone													
		1	**2**	**3**	**4**	**5**	**6**								
		5	11	21	57	119	282	495	**82.50**						
		6006.25	5112.25	3782.25	650.25	1332.25	39800.25		**82.50**	11336.7	106.47394	43.4678042	**-2.70**	**167.70**	AB
6	**8 Stones**	Stone													
		1	**2**	**3**	**4**	**5**	**6**								
		4	8	20	26	135	44	237	**39.50**						
		1260.25	992.25	380.25	182.25	9120.25	20.25		**39.50**	2391.1	48.8988752	19.9628822	**0.37**	**78.63**	AB

Notes

While the results of the competition shows the 8 stones group to be the winners, were they influenced by confirmation bias, picking a final stone to get closer to the expected result?

Of course, assuming we have an unbiassed sampling design, exactly how close to the true mean an individual sample happens to be is influenced by chance. So the rules of the competition are not exactly fair: if all the groups used exactly the same selection method it would be a lottery.

In fact, in this exercise as in all the previous exercises we have done with similar groups of stones, the groups **always overestimate the population mean weight**, often by a considerable amount. When we try to choose a representative sample in the field using only our judgement, we tend to be strongly biased towards choosing plots which over-represent whatever we are interested in, such as the cover of a protected plant species. We can unconsciously reject 'poorer' plots and locations, perhaps feeling we are wasting time recording them.

The Rolling stones divided the stones into 5 strata and selected representative stones from each (two from one), weighting them by the number in each stratum. Stratification is a good approach, though even here they tended to pick slightly heavier representative stones, overestimating the the population weight by 25%. Because we have only one (or two) stones in each stratum, we cannot so easily calculate confidence limits – but in fact, the simulations show that these calculations aren't really legitimate for *any* of the groups, something hinted at by the lower limit of -2.70 for the Stone Roses.

The **sample variance** formula looks like this:*

Formula	Explanation
$s^2 = \frac{\Sigma (X - \bar{x})^2}{n - 1}$	• s^2 = sample variance • Σ = sum of... • X = each value • $\bar{x}$ = sample mean • n = number of values in the sample

*Nb we use slightly different formulas for variance and standard deviation when we have a lot of values, and to minimise rounding errors

$$s^2 = \frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N - 1}$$

The **sample standard deviation** formula looks like this:

Formula	Explanation
$s = \sqrt{\frac{\Sigma (X - \bar{x})^2}{n - 1}}$	• s = sample standard deviation • Σ = sum of... • X = each value • $\bar{x}$ = sample mean • n = number of values in the sample

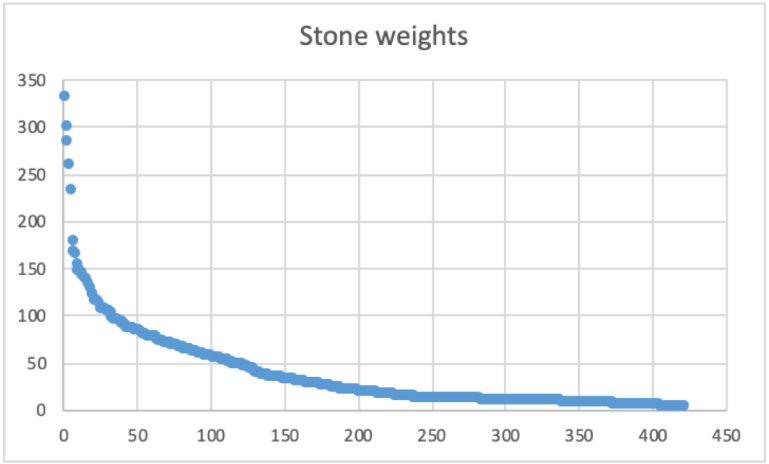
Formula	Explanation
$SE = \frac{s}{\sqrt{n}}$	• SE is standard error • s is sample standard deviation • n is the number of elements in the sample

95% confidence intervals (for **normal distribution**)

Lower limit	Upper limit
$\bar{x} - (1.96 \times SE)$	$\bar{x} + (1.96 \times SE)$

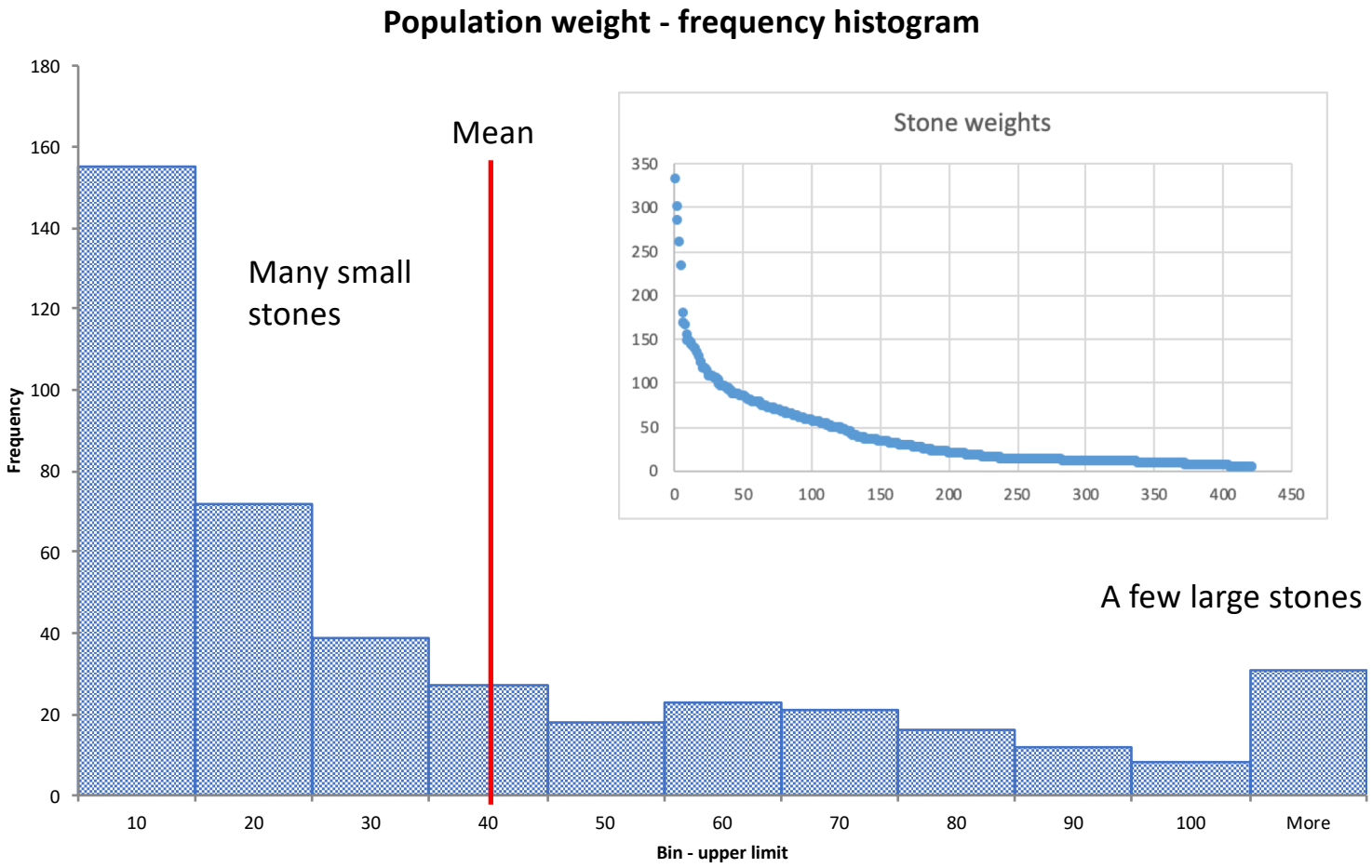
Simple explanations at:
https://www.scribbr.com/statistics

422 Stones



Smallest 1 gram → Largest 330 grams

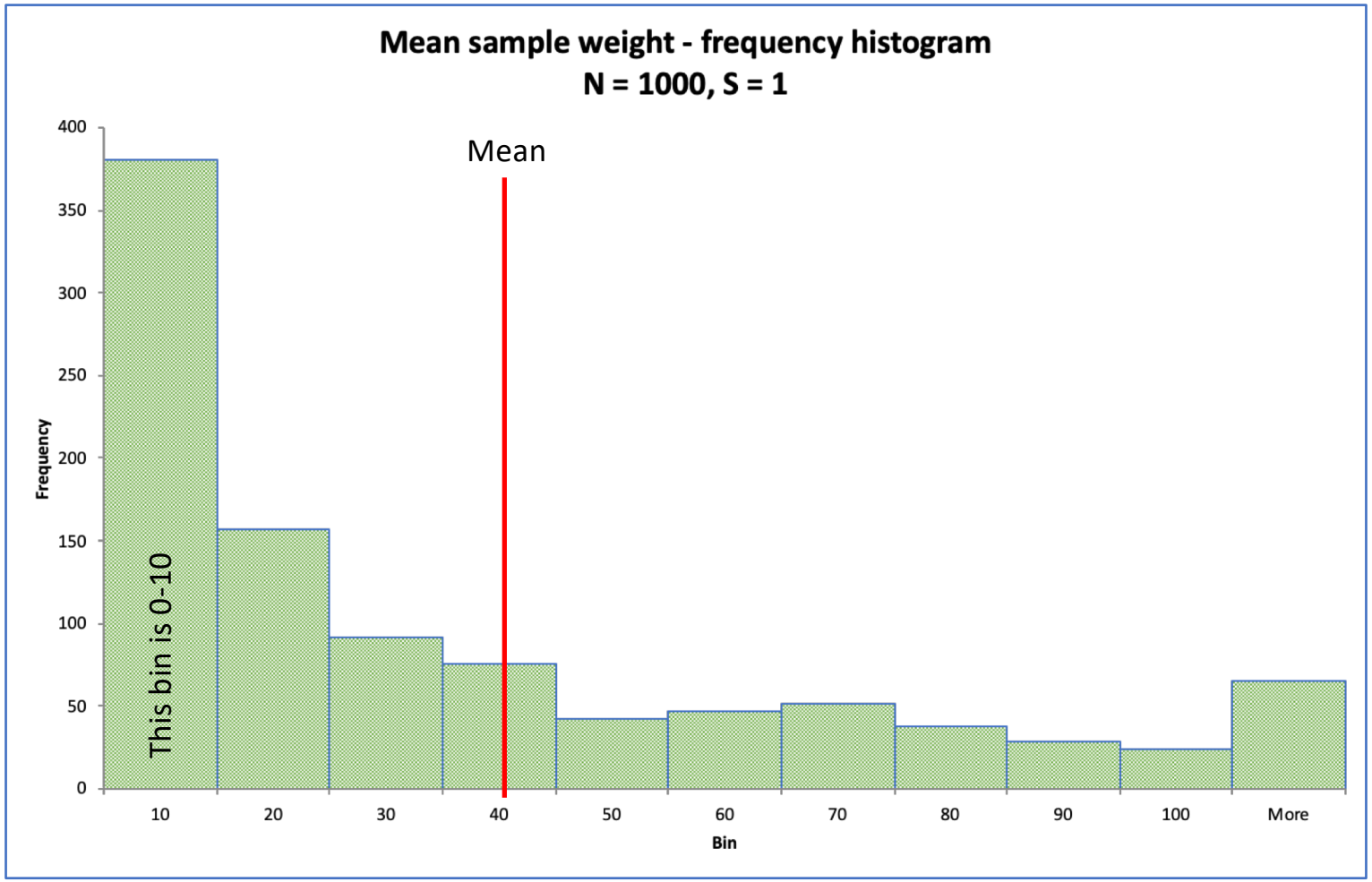
Mean weight = 35.85 grams

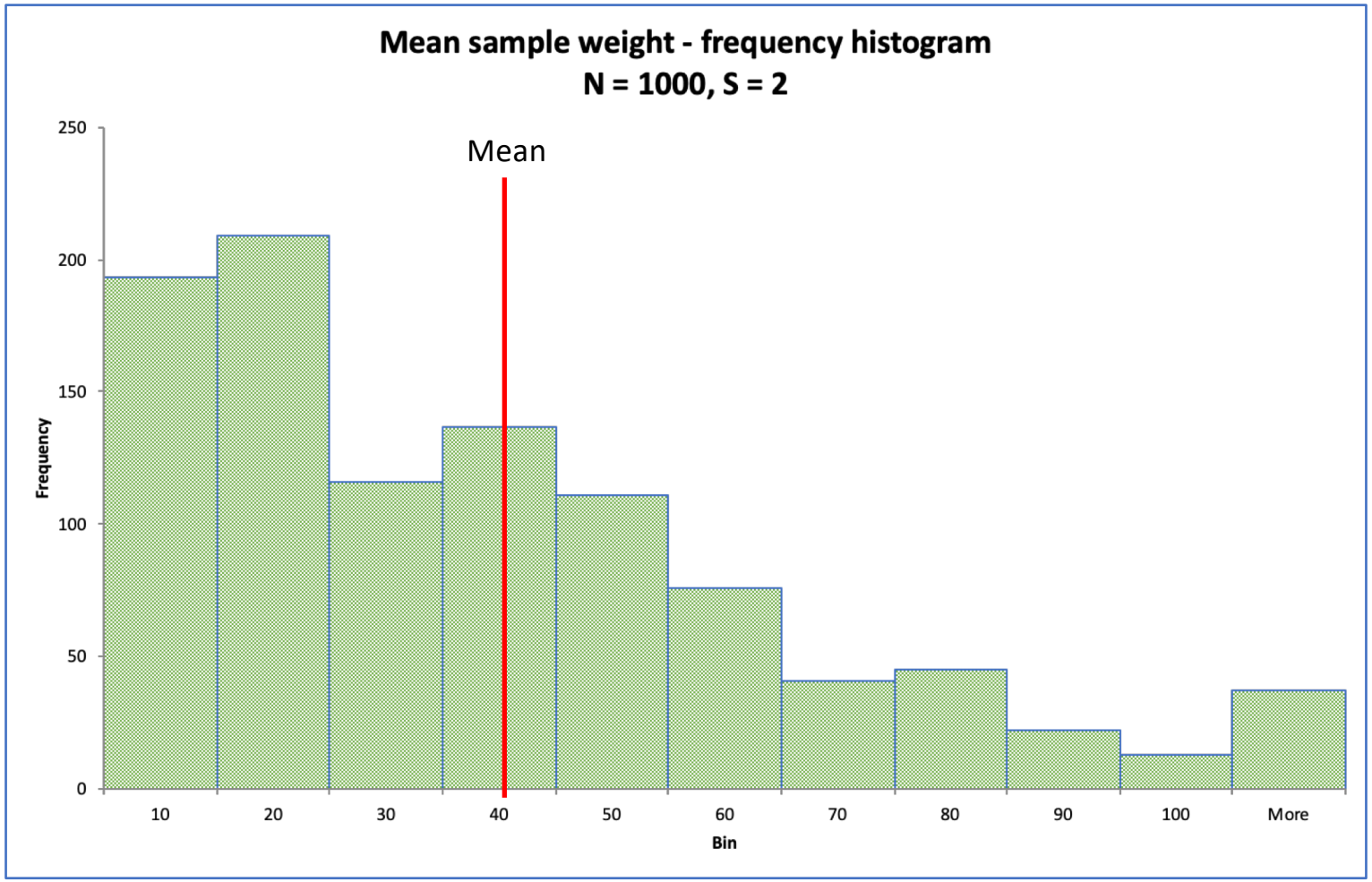

Population weight - frequency histogram
Mean
Many small stones
A few large stones
Stone weights
Frequency
Bin - upper limit
0
20
40
60
80
100
120
140
160
180
10
20
30
40
50
60
70
80
90
100
More
0
50
100
150
200
250
300
350
400
450

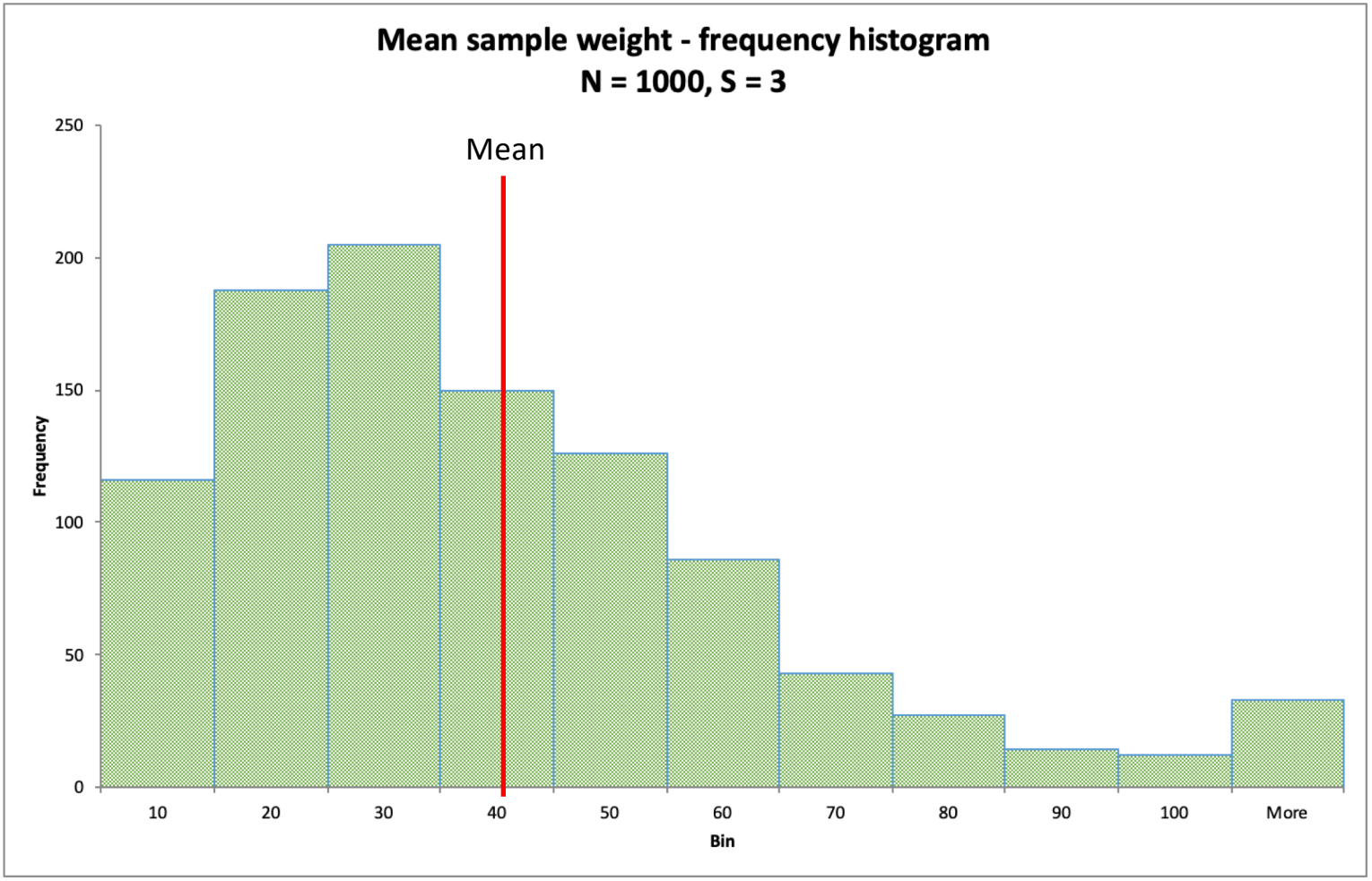
1. Equal Probability sampling

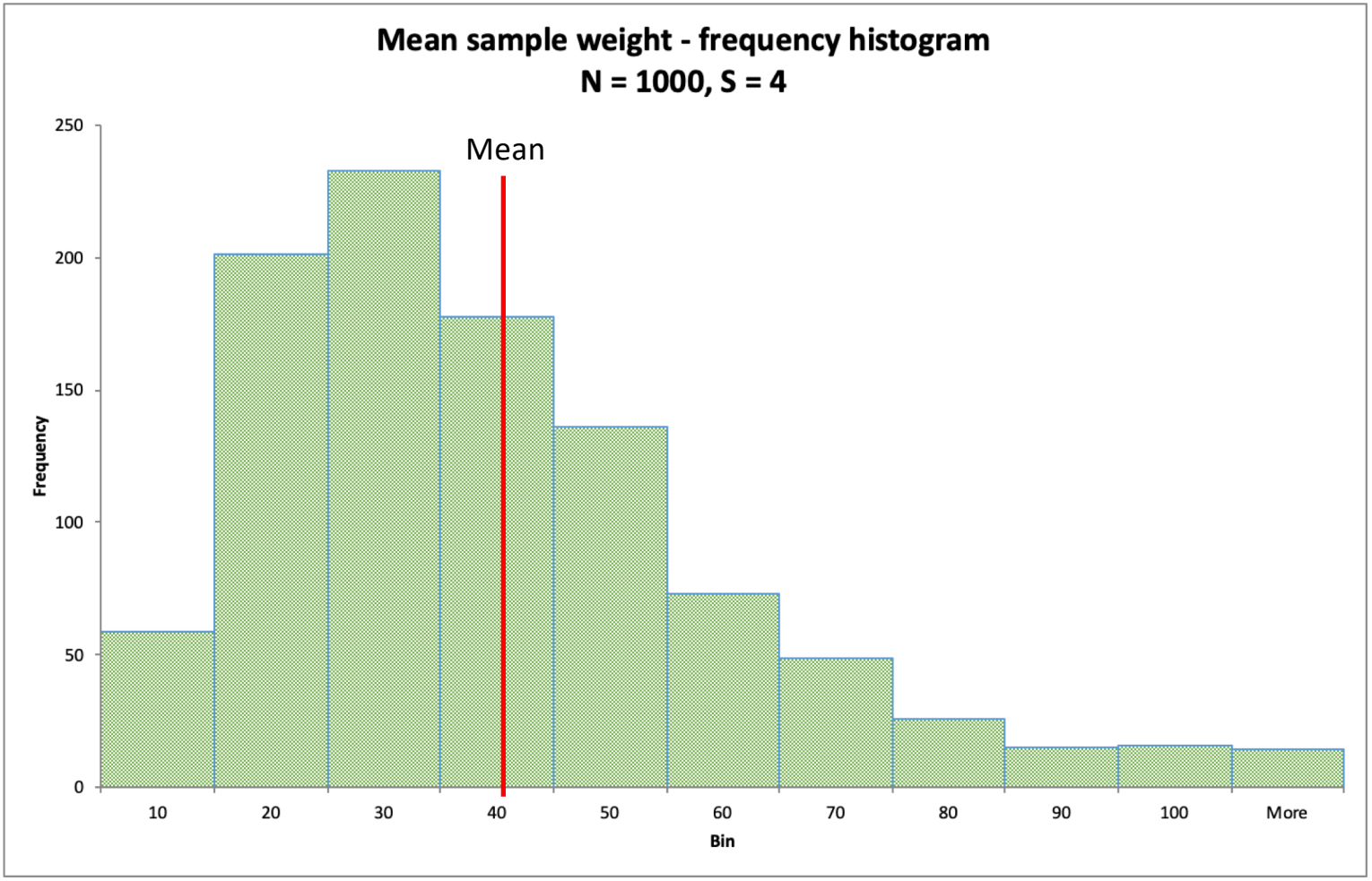
Simple Random Sampling

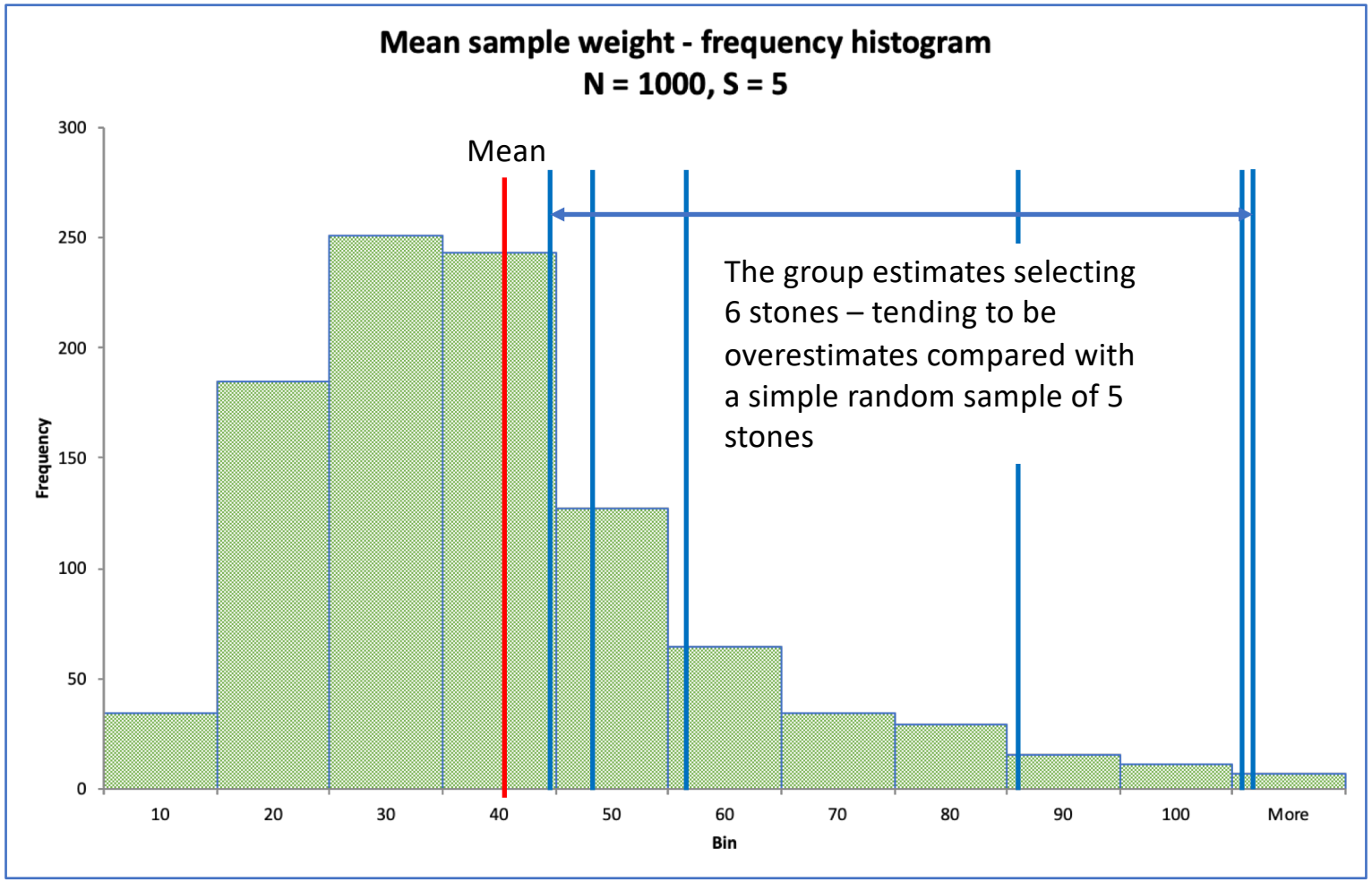
Each population unit (= stone) has an equal chance of being selected for the sample


Mean sample weight - frequency histogram
N = 1000, S = 1
Mean
Frequency
400
350
300
250
200
150
100
50
0
This bin is 0-10
10
20
30
40
50
60
70
80
90
100
More
Bin


Mean sample weight - frequency histogram
N = 1000, S = 2
Mean
250
200
150
100
50
0
Frequency
10
20
30
40
50
60
70
80
90
100
More
Bin


Mean sample weight - frequency histogram
N = 1000, S = 3
Mean
250
200
150
100
50
0
Frequency
10
20
30
40
50
60
70
80
90
100
More
Bin


Mean sample weight - frequency histogram
N = 1000, S = 4
Mean
250
200
150
100
50
0
Frequency
10
20
30
40
50
60
70
80
90
100
More
Bin


Mean sample weight - frequency histogram
N = 1000, S = 5
Mean
The group estimates selecting 6 stones – tending to be overestimates compared with a simple random sample of 5 stones
Frequency
300
250
200
150
100
50
0
10
20
30
40
50
60
70
80
90
100
More
Bin

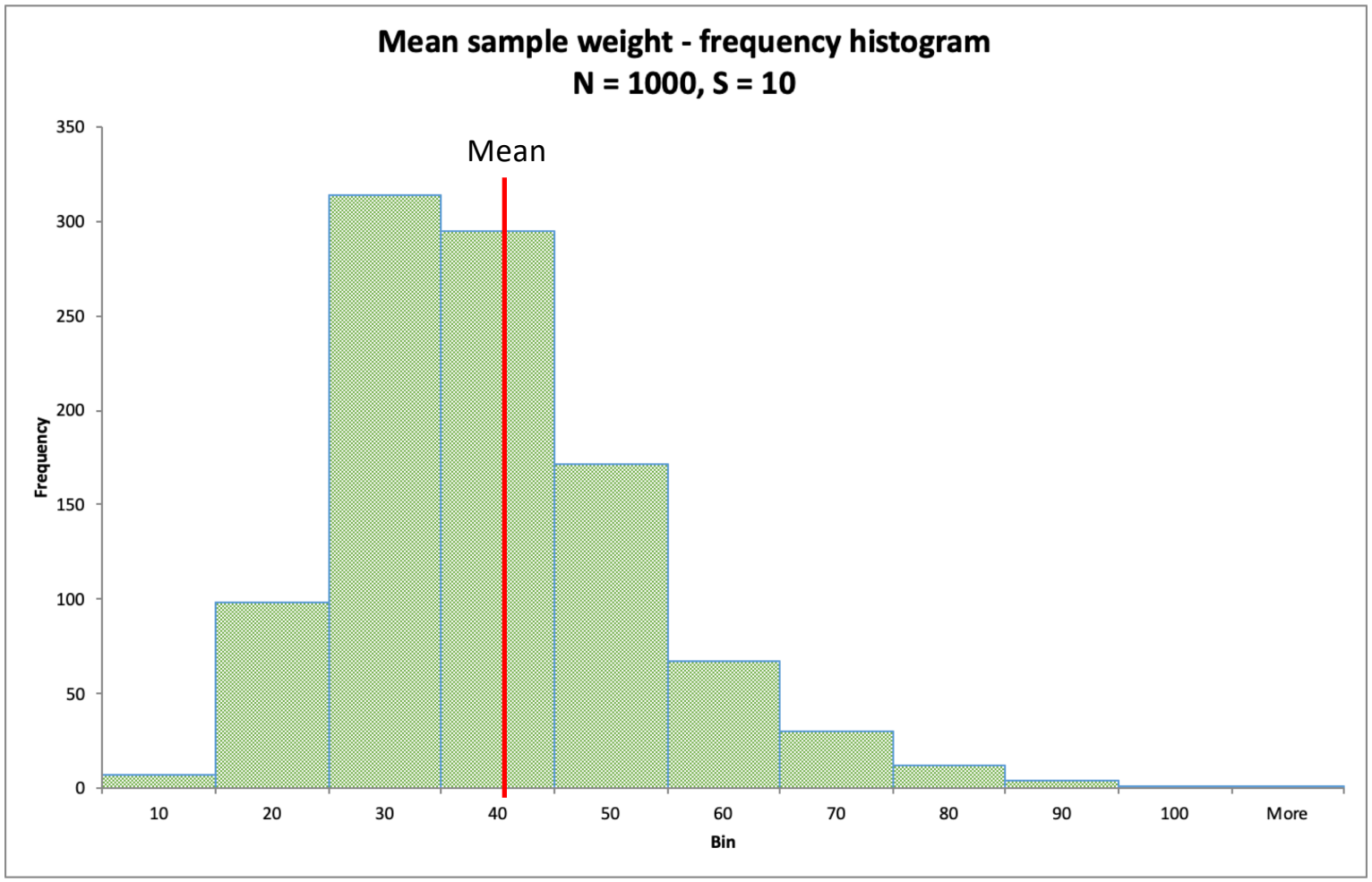
The central limit theorem

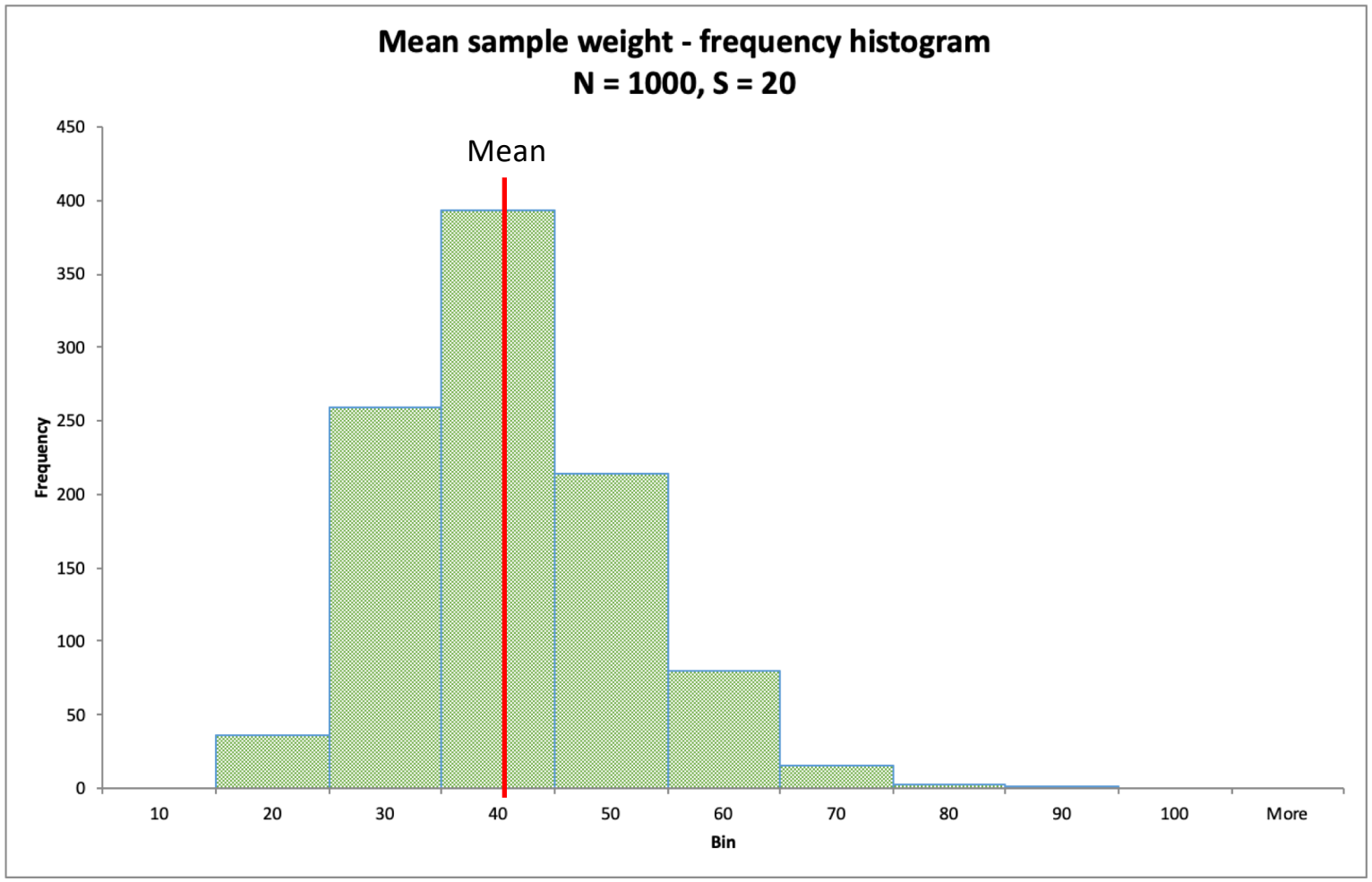
With Probability Sampling, as the sample size increases, the

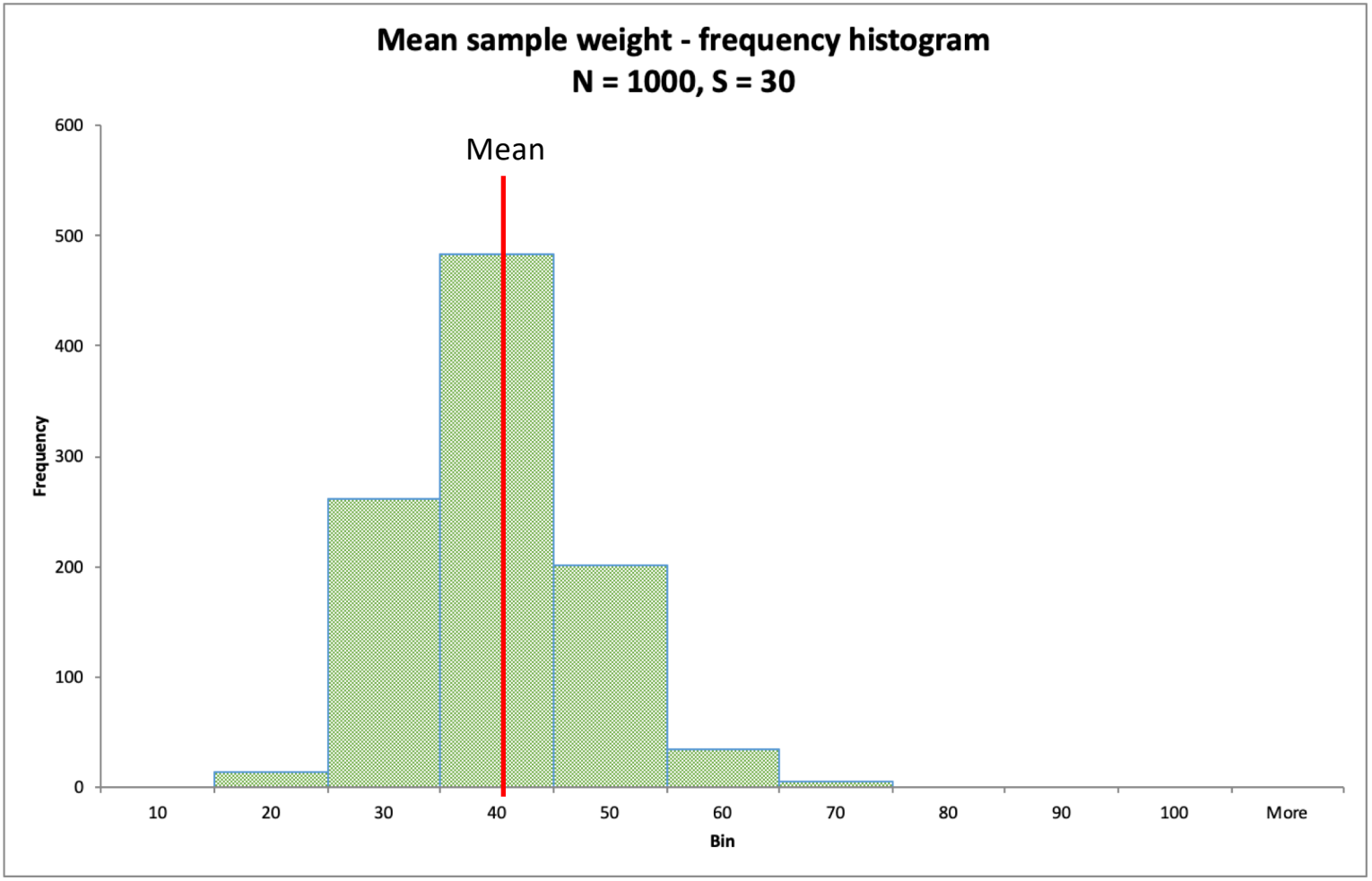
sampling distribution of sample means

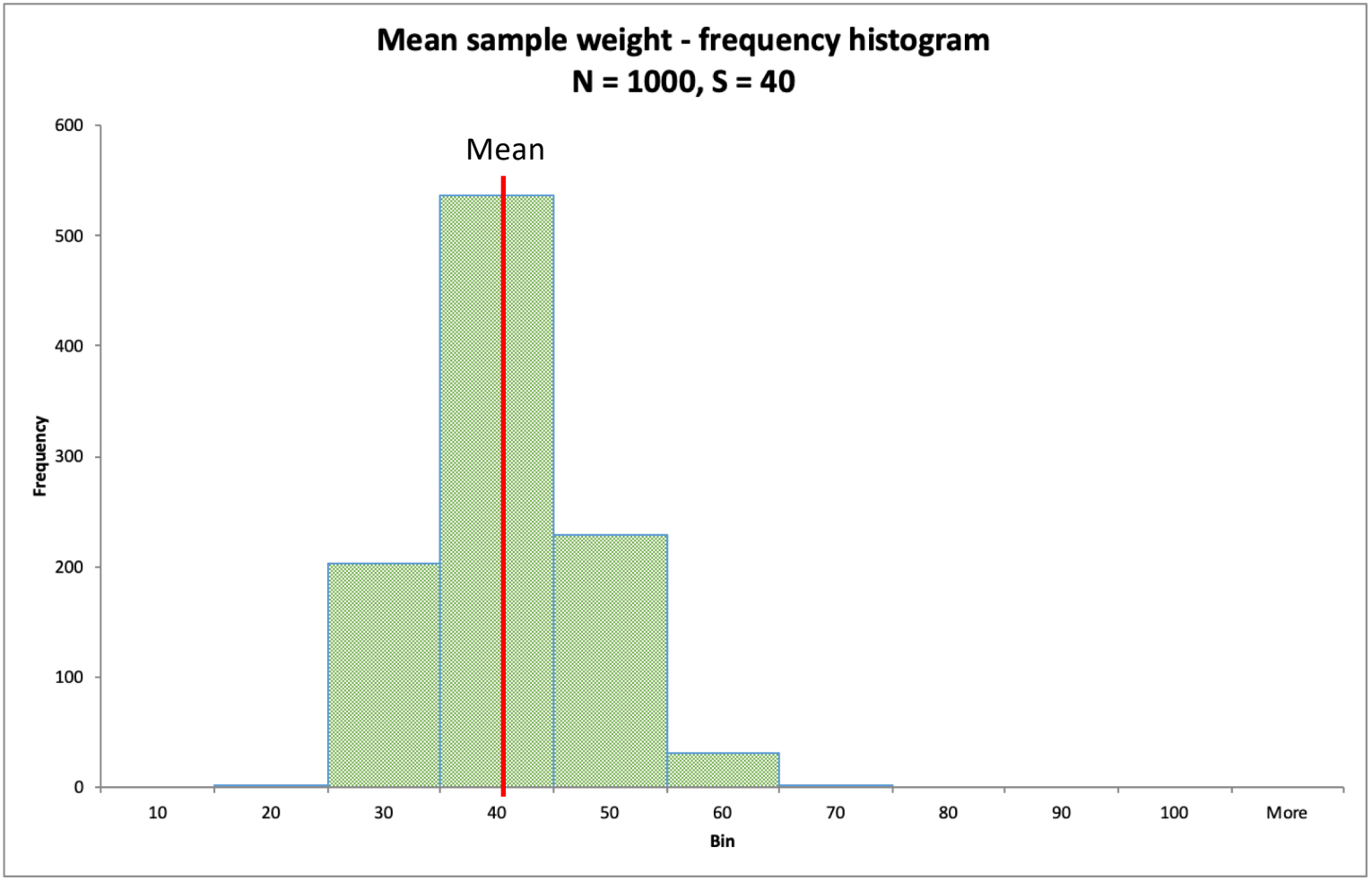
- becomes more symmetrical
- approximates a normal distribution,
- while the variance (spread) decreases

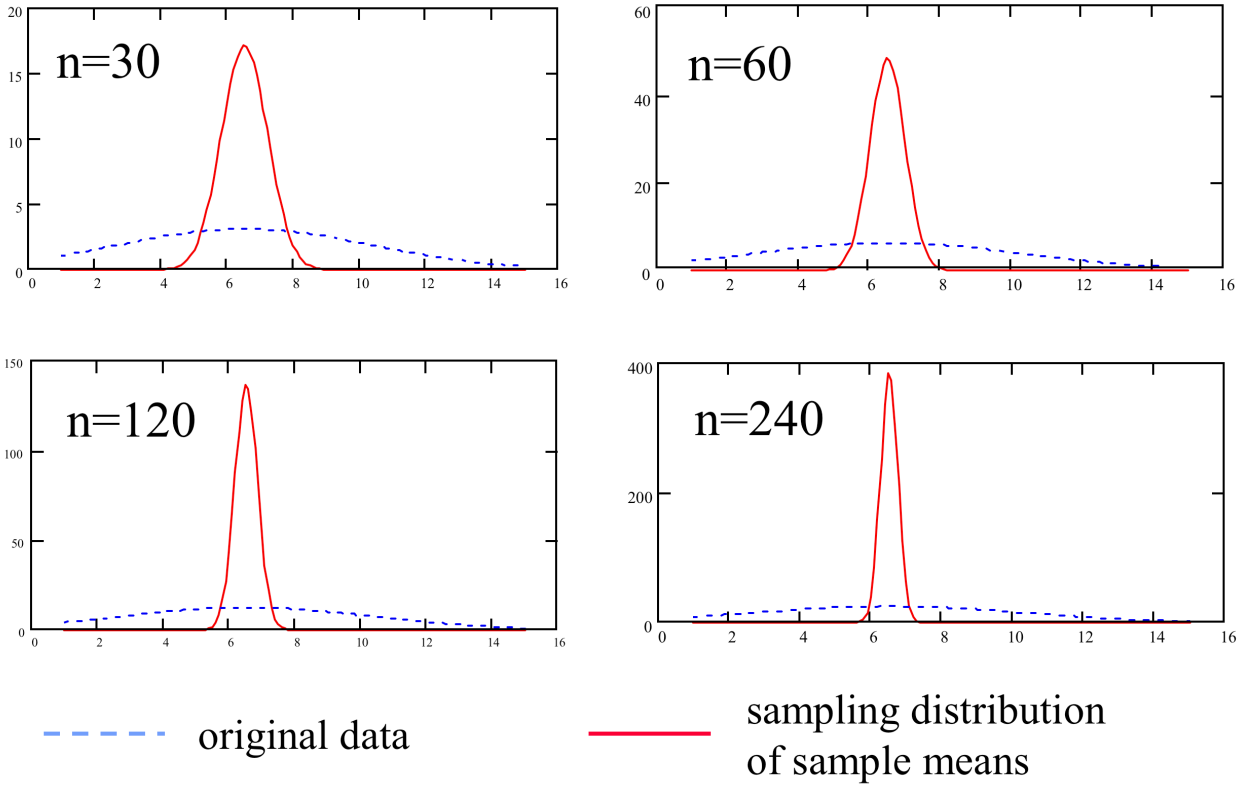
We can use this result to calculate confidence limits, carry out parametric tests (eg T tests) and use statistical power analysis to determine sample size


Mean sample weight - frequency histogram
N = 1000, S = 10
Mean
350
300
250
200
150
100
50
0
Frequency
10
20
30
40
50
60
70
80
90
100
More
Bin

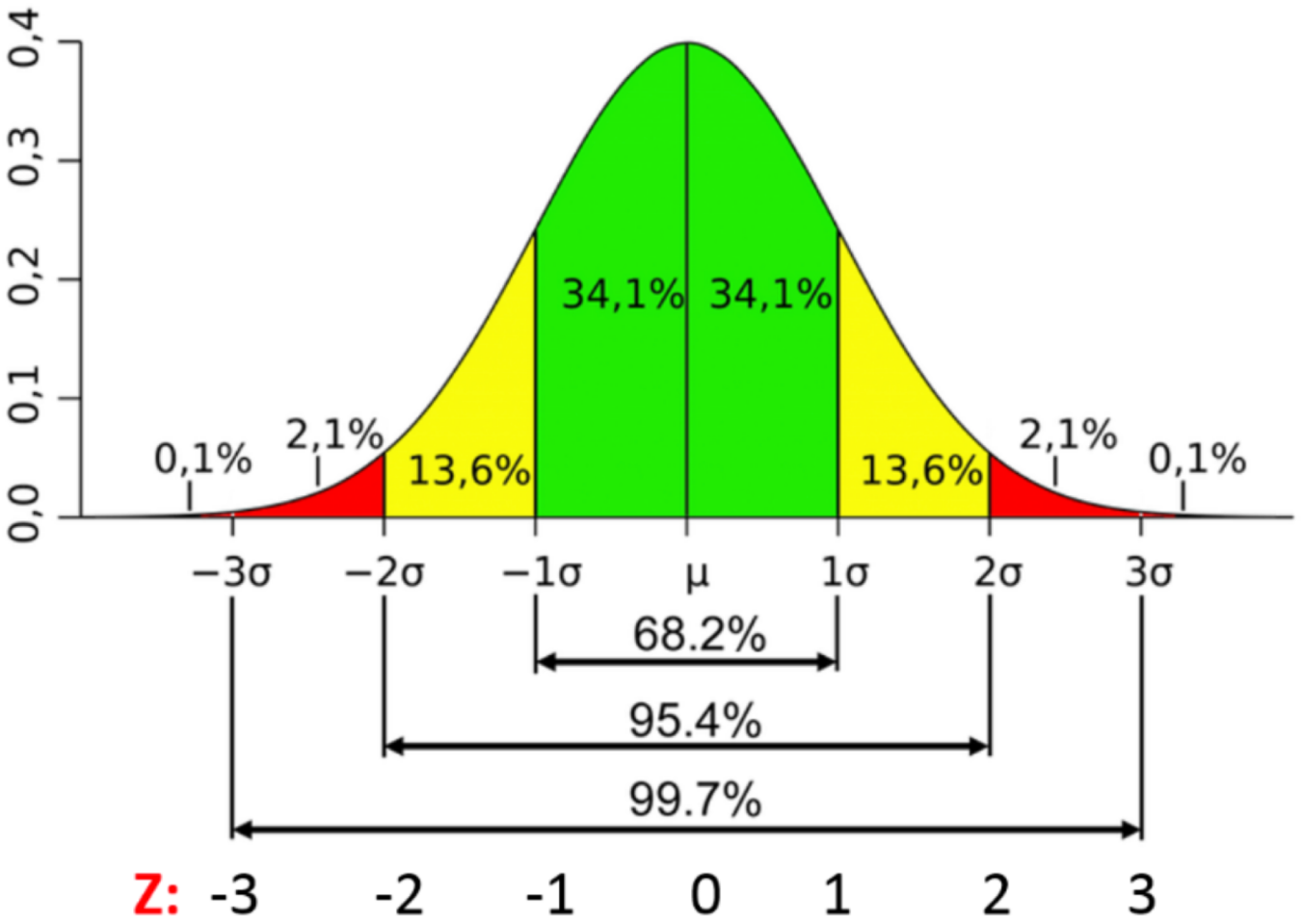

Mean sample weight - frequency histogram
N = 1000, S = 20
Mean
Frequency
450
400
350
300
250
200
150
100
50
0
10
20
30
40
50
60
70
80
90
100
More
Bin


Mean sample weight - frequency histogram
N = 1000, S = 30
Mean
600
500
400
300
200
100
0
Frequency
10
20
30
40
50
60
70
80
90
100
More
Bin


Mean sample weight - frequency histogram
N = 1000, S = 40
Mean
600
500
400
300
200
100
0
Frequency
10
20
30
40
50
60
70
80
90
100
More
Bin



As sample size increases, the standard error (variance of the sampling distribution of sampling means) decreases

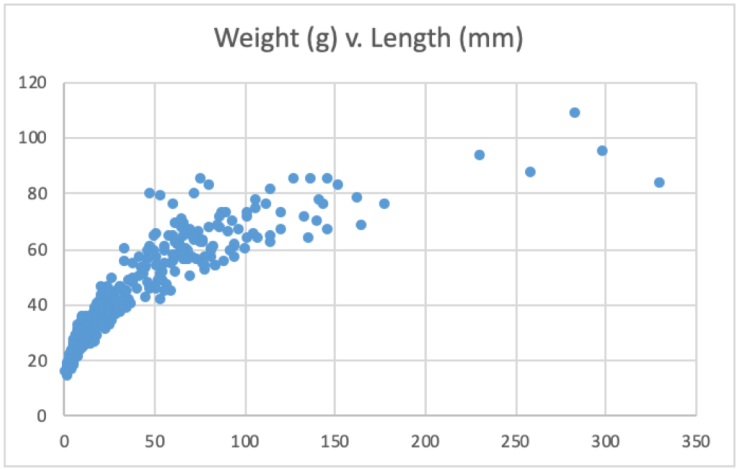

0,4
0,3
0,2
0,1
0,0
34,1%
34,1%
0,1%
2,1%
13,6%
13,6%
2,1%
0,1%
−3σ
−2σ
−1σ
μ
1σ
2σ
3σ
68.2%
95.4%
99.7%
Z: -3 -2 -1 0 1 2 3

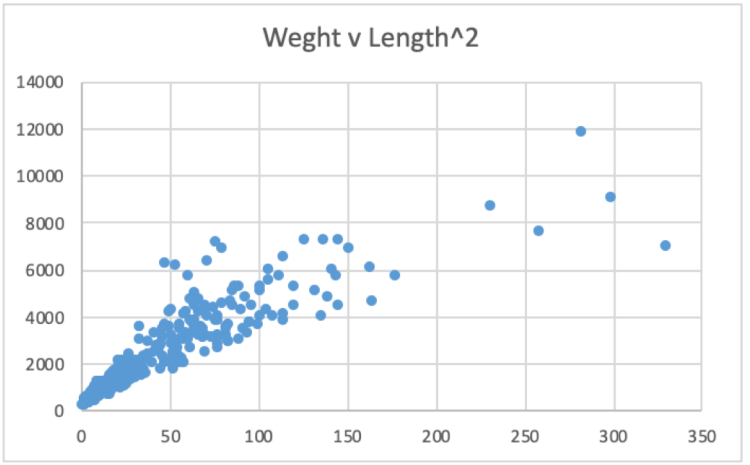
2. Unequal Probability sampling

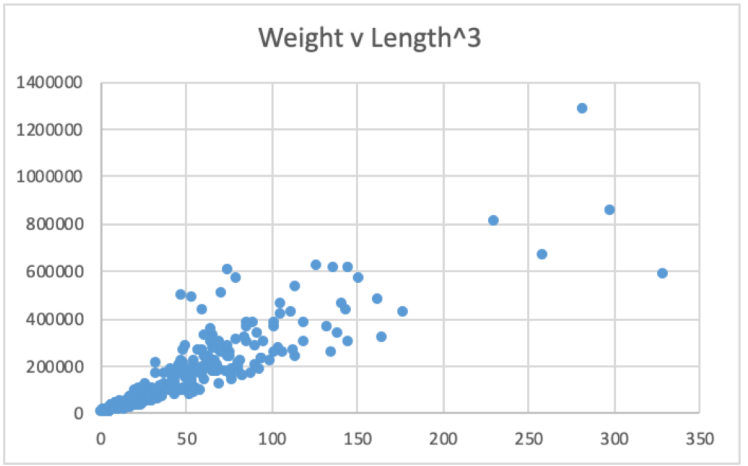
Each population unit (= stone) has a chance of being selected, but not an equal chance

These **selection probabilities** are proportional to some auxiliary variable, one which predicts some of the variance

Length as an auxiliary variable which predicts stone weight







Weight ~ volume ~ length x width x height

Weight ~ length^3

Auxiliary variable

Inclusion probability π_i

Cumulative π_i

Stone	Weight_g	Length_mm	Length^3	IP	Cum_L^3
1	330	83.6	584277.1	0.0129722	0.012972203
2	299	94.9	854670.3	0.01897551	0.031947716
3	283	108.7	1284366	0.02851567	0.060463383
4	259	87.2	663054.8	0.01472124	0.075184621
5	231	93	804357	0.01785845	0.09304307
6	178	75.3	426957.8	0.00947938	0.102522448
7	165	68	314432	0.00698106	0.109503512
.					
.					
.					
417	2	16.8	4741.632	0.00010527	0.999531405
418	2	14.1	2803.221	6.2238E-05	0.999593643
419	2	15.1	3442.951	7.6441E-05	0.999670084
420	2	18.6	6434.856	0.00014287	0.999812951
421	2	17	4913	0.00010908	0.99992203
422	1	15.2	3511.808	7.797E-05	1
SUM:	35.8507109	39.48317536	45040697	1	

Sample size

weight

$$\hat{\tau}_\pi = \sum_{i=1}^{\nu} \frac{y_i}{\pi_i}$$

Inclusion probability

3. Horvitz-Thompson Estimators

Most modern survey-sampling depends on unequal probability sampling and stratification, for example using auxiliary variables from remote sensing.

A good way of understanding unequal probabilities is looking at **Horvitz-Thompson Estimators**.

If anyone wants to know about these in more detail, we can run a smaller informal session on statistical methods

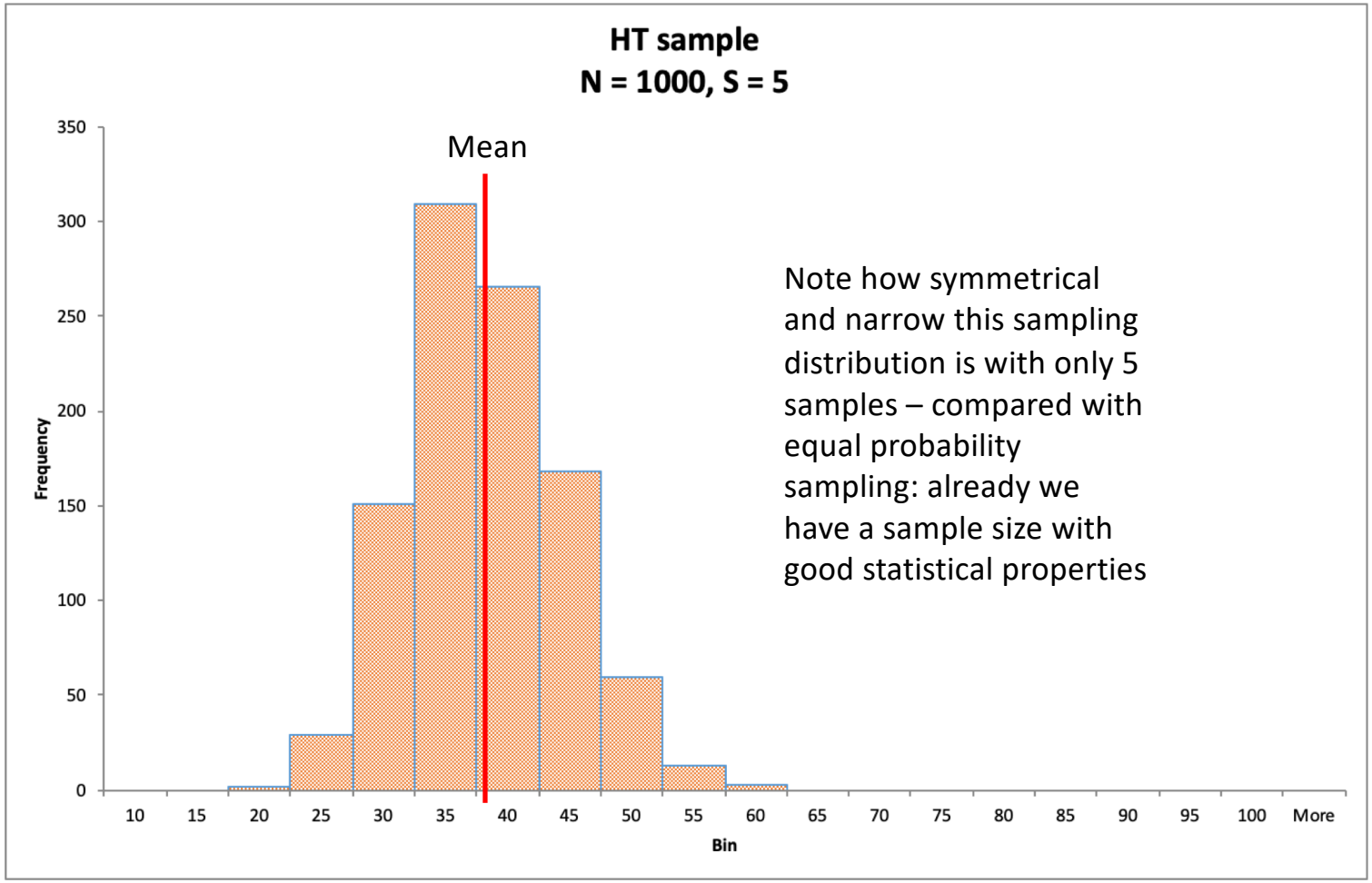
Horvitz-Thompson estimators

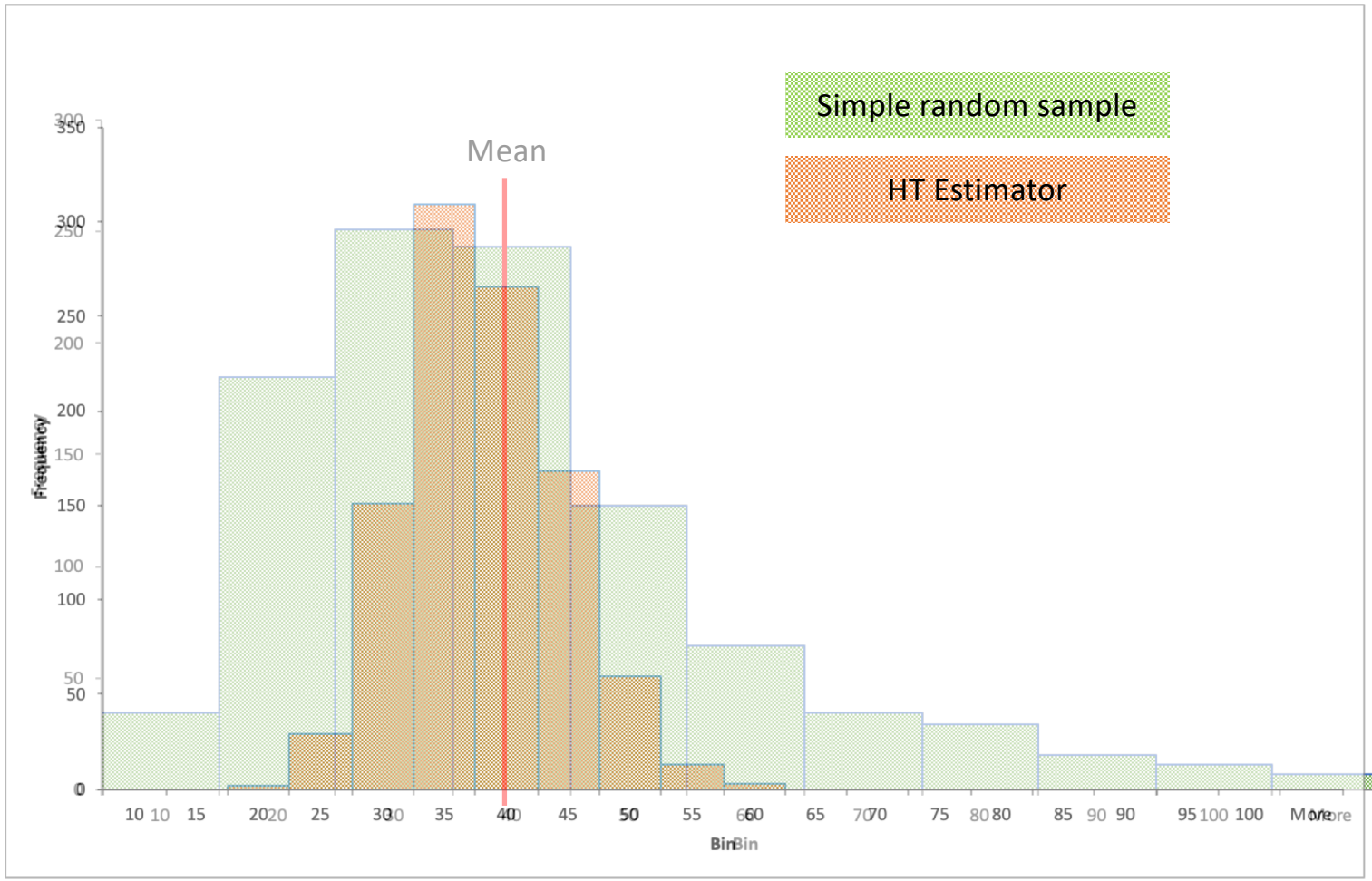
$$\hat{\tau}_\pi = \sum_{i=1}^{\nu} \frac{y_i}{\pi_i}$$

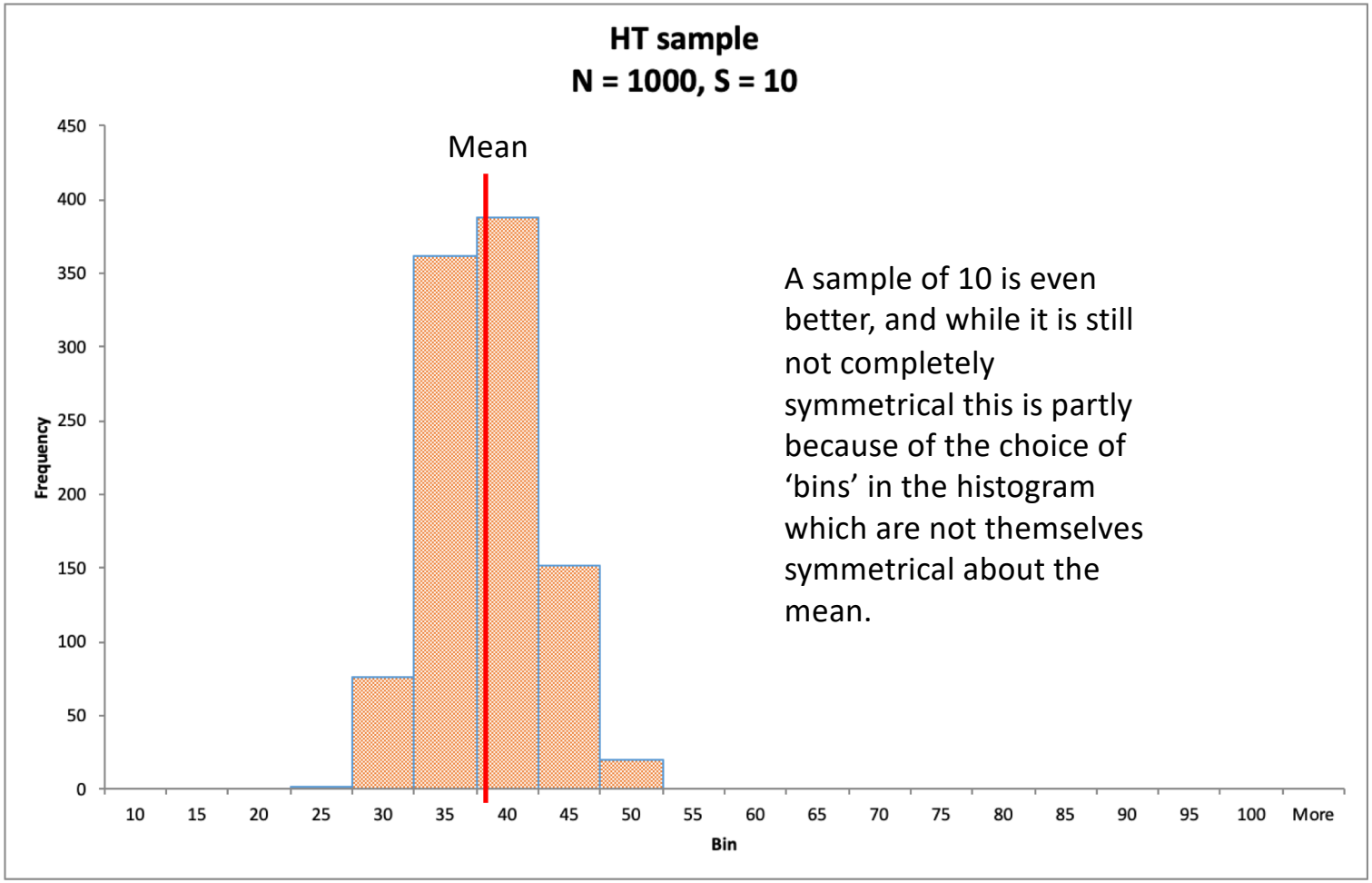
This is the HT estimator for the total, which we can easily convert to the mean by dividing by the sample / population number.

This is the rather more scary equation for the estimated variance (of the estimated total) – but of course all of this can be calculated using statistical software such as R. The $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_{ij}$ elements are inclusion probabilities for single units and pairs of units, the y_i and y_{ij} are the observed variables, in our case the stone weights.

$$\hat{V}ar(\hat{\tau}_\pi) = \sum_{i=1}^{v} \left(\frac{1-\pi_i}{\pi_i^2} \right) y_i^2 + \sum_{i=1}^{v} \sum_{j \neq i} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) \frac{1}{\pi_{ij}} y_i y_j$$

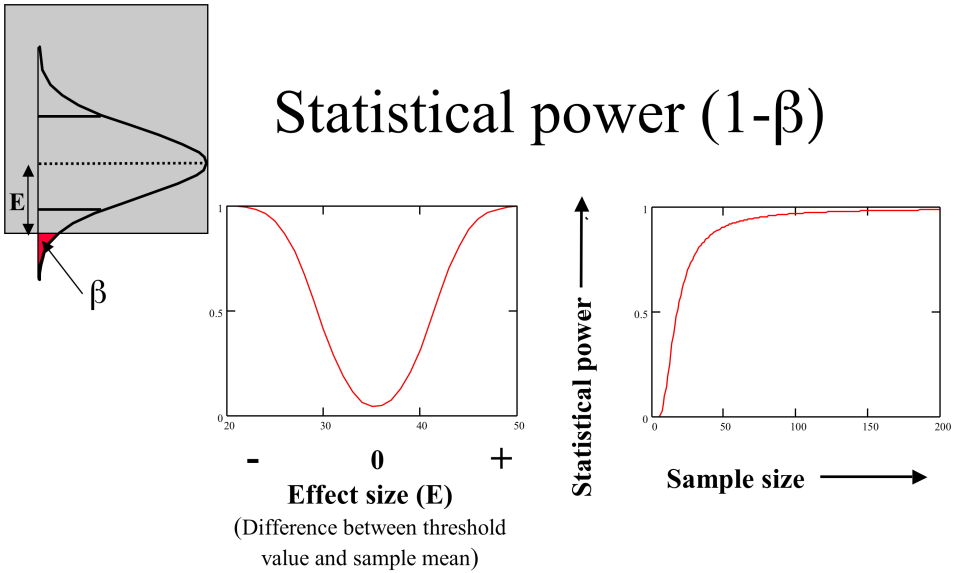

HT sample
N = 1000, S = 5
Mean
Frequency
350
300
250
200
150
100
50
0
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
More
Bin
Note how symmetrical and narrow this sampling distribution is with only 5 samples – compared with equal probability sampling: already we have a sample size with good statistical properties


Simple random sample
HT Estimator
Mean
Frequency
Bin
More



A sample of 10 is even better, and while it is still not completely symmetrical this is partly because of the choice of 'bins' in the histogram which are not themselves symmetrical about the mean.

Statistical power analysis depends on having a sample size big enough to be normally distributed...



- Statistical power increases with effect size (constant sample size)
- Statistical power increases with sample size (constant effect size)

... and is essential for hypothesis testing in order to estimate the type II error rate....

You conclude ↓	Reality: H_0 *true*	Reality: H_0 *false*
Accept H_0	*correct* (1-α)	**type II** error (β)
Reject H_0	**type I** error (α)	*correct* (1-β)

- A **type I** (false change) error means rejecting a true null hypothesis H_0
- A **type II** (missed change) error means accepting a false null hypothesis H_0

Notes and Further Reading

It is really important to understand unequal probability sampling in sample-survey design, especially now we often have auxiliary variables from remote sensing which can be used as inclusion probabilities and for stratification or post-stratification. A good place to start is:

Overton, W.S. & Stehman, S.V. (1995). The Horvitz-Thompson theorem as a unifying perspective for probability sampling: with examples from natural resource sampling. *The American Statistician*, 49(3), pp. 261-268.

And real, practical designs are discussed in considerable detail in the classic text:

Särndal, C-E., Swensson, B. & Wretman, J. (2003) *Model-Assisted Survey Sampling*. Springer. (there are on-line scans of this book if you look for them...)

An excellent recent text is:

Tillé, Y. (2020). *Sampling and Estimation from Finite Populations.* John Wiley & Sons.
see also: https://cran.r-project.org/web/packages/sampling/index.html

Note that there are other approaches to using very small samples, for example using data transformations and Bayesian statistics – but these are for another workshop!