

STATISTICAL SAMPLING

PRESENTED BY:

DARWYN JONES

CHIEF PERFORMANCE ANALYST - AUDIT AND PROGRAM REVIEW

CITY OF CHICAGO OFFICE OF INSPECTOR GENERAL

(773) 478-4680

DJONES@IGCHICAGO.ORG

August 2021 CIGA Institute – Jacksonville, FL

Course Objectives

2

- ❑ Recognize when to use a statistical sample vs non-statistical sample in the context of an OIG performance audit.
- ❑ Know how to calculate the most appropriate statistical sample size.
- ❑ Know how to extrapolate sample results to the population.

Definition

3

SAMPLE: A subset of the population that the auditor examines.

Definition

4

Audit sampling is the application of audit procedures to less than 100% of the items within the population for the purpose of evaluating some characteristic of that population.

Standards

General Standards: “The audit organization’s management must assign auditors to conduct the engagement who before beginning work on the engagement collectively possess the competence needed to address the engagement objectives and perform their work in accordance with GAGAS.” (4.02)

Field Work Standards: “When sampling is used, the appropriate selection method will depend on the audit objectives. When a representative sample is needed, the use of statistical sampling approaches generally results in stronger evidence than that obtained from nonstatistical techniques. When a representative sample is not needed, a targeted selection may be effective if the auditors have isolated risk factors or other criteria to target the selection.” (8.107)

Standards

6

Reporting Standards: “Auditors should identify significant assumptions made in conducting the audit; describe comparative techniques applied; describe the criteria used; and, when the results of sample testing significantly support the auditors’ findings, conclusions, or recommendations, describe the sample design and state why the design was chosen, including whether the results can be projected to the intended population.” (9.14)

Why Sample Statistically?

7

- If you want to make inferences about a population with a sample, you must sample statistically (randomly).
- When you sample non-statistically, such as a convenience or judgmental sample, you can only speak about the units you observed. You cannot reasonably extrapolate to the whole population.

Audit Risk – What if I’m Wrong?

8

“Audit risk is the possibility that the auditors’ findings, conclusions, recommendations, or assurance may be improper or incomplete as a result of factors such as evidence that is not sufficient or appropriate, an inadequate audit process, or intentional omissions or misleading information because of misrepresentation or fraud.” (8.16)

Relying on statistical sampling can help minimize audit risk and allow others to assess the sufficiency and appropriateness of your evidence.

Statistical vs. Non-Statistical

9

- Statistical sampling allows you to make inferences about a population with a quantifiable level of certainty and precision.
- ★
- Non-statistical may require fewer resources.
 - ▣ If error is rare but would have large impact, a risk-based judgmental sample may be more likely to demonstrate existence of the problem.
 - ▣ When data is incomplete or unreliable, you may not be able to create a statistical sample.

Choosing an Approach

10

- Different methodologies work better depending on the situation. Some things to consider include:
 - ▣ What is the question I want to answer?
 - ▣ Do I have the resources to test the whole population or do limitations mean that's not feasible?
 - ▣ How reliable is the data?
 - ▣ Do I need to be able to extrapolate to the population?
 - ▣ Does the data contain the variables that I need to test?
 - ▣ What type of variable will I be testing?

You should consider these and other factors when developing your methodology.

Obtaining the Best Statistical Sample

11

- Know your population of interest and obtain a sampling frame.

(A sampling frame is a comprehensive list of units that could potentially be selected for your sample.)

- ▣ Define the period covered by the test
- ▣ Define the sample unit
- ▣ Consider the completeness of the population

Obtaining the Best Statistical Sample

12

- Select a sampling strategy that minimizes selection bias.
(Selection bias is a common form of bias where certain data points with common characteristics have a higher probability of being included in a sample. This can lead to an overestimation or underestimation of the true value. Random sampling eliminates selection bias. The sample should be representative of the population.)
- ▣ Simple Random Sampling
- ▣ Systematic Random Sampling
- ▣ Stratified Random Sampling

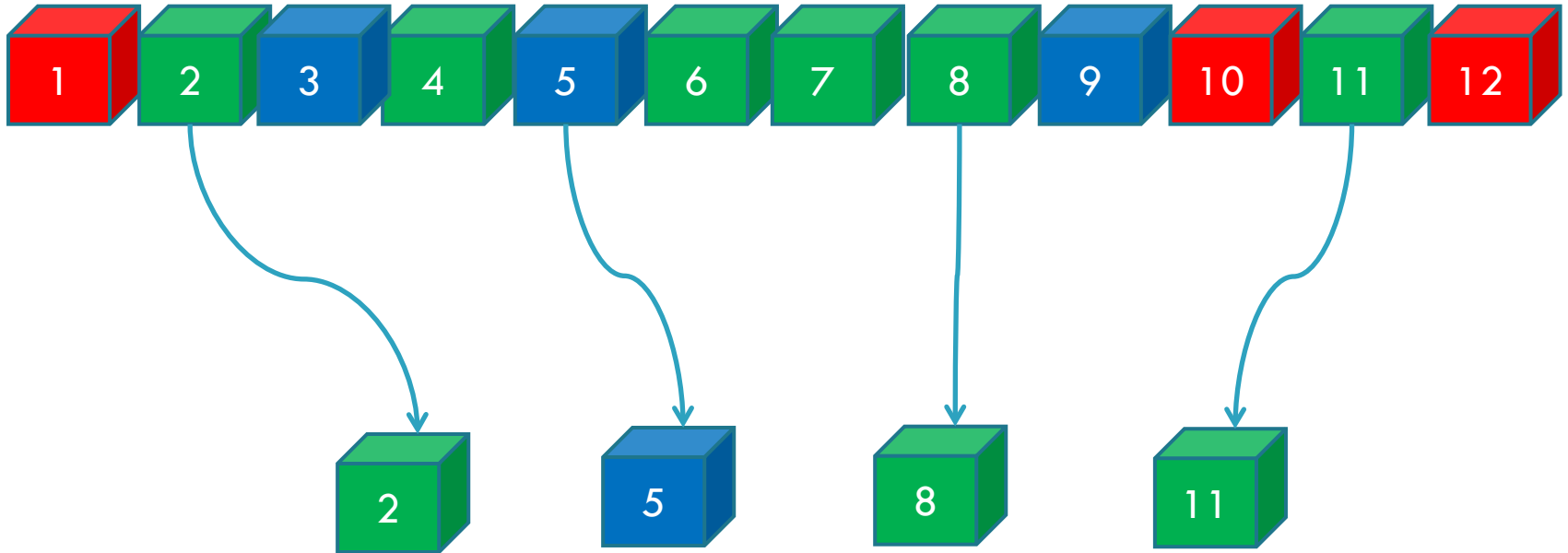
Simple Random Sampling

13



Systematic Random Sampling

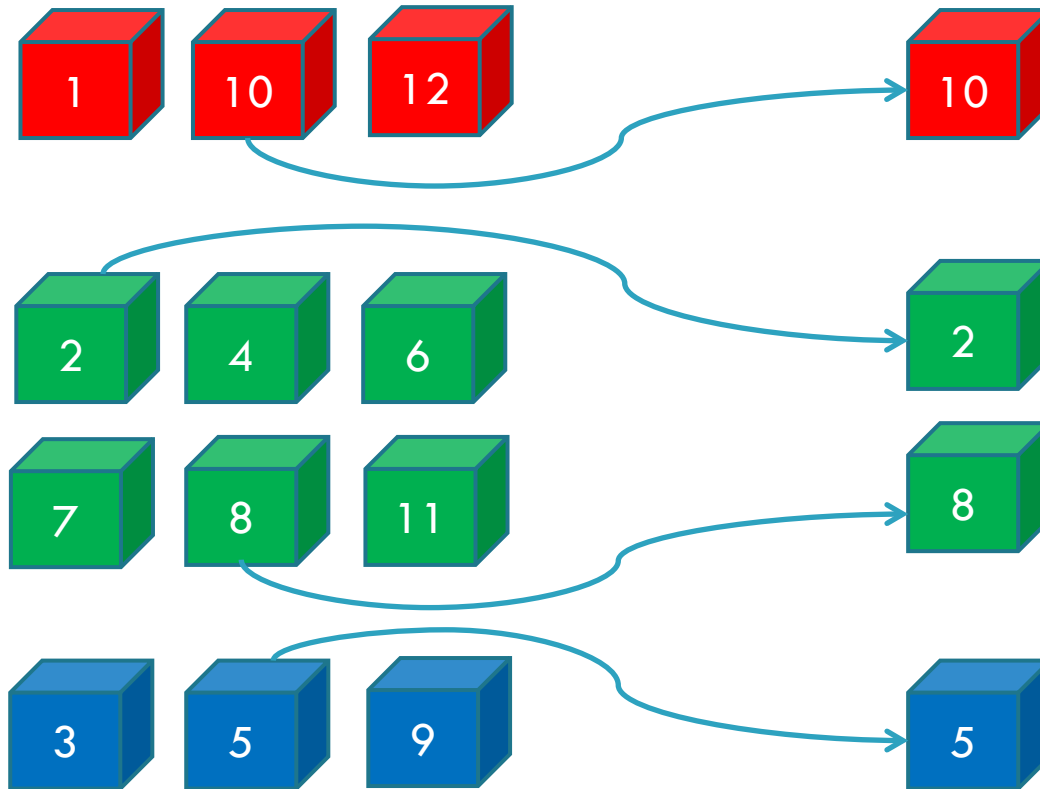
14



Every n th item is selected after a random start.

Stratified Random Sampling

15



Population is divided into subgroups.

A random sample is taken from each subgroup.

The resulting sample should be proportional to population.

Obtaining the Best Statistical Sample

16

- Make sure sample size is large enough to be representative of the population.

(There is always a balance between how certain one is that the sample is representative of the population and how large the sample should be.)

- Determine Acceptable
 - ▣ Confidence Level
 - ▣ Margin of Error

Confidence Level and Margin of Error

17

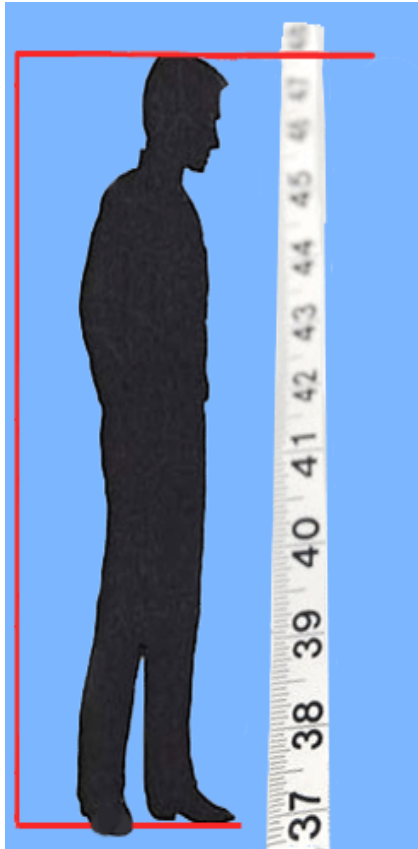
Confidence Level: How confident do you want to be that the sample results are reflective of the population?

Margin of Error: How precise do you want your conclusion to be? (How much wiggle room will you allow?)

(There is always a balance between how certain one is that the sample is representative of the population and how large the sample should be.)

Example: Height

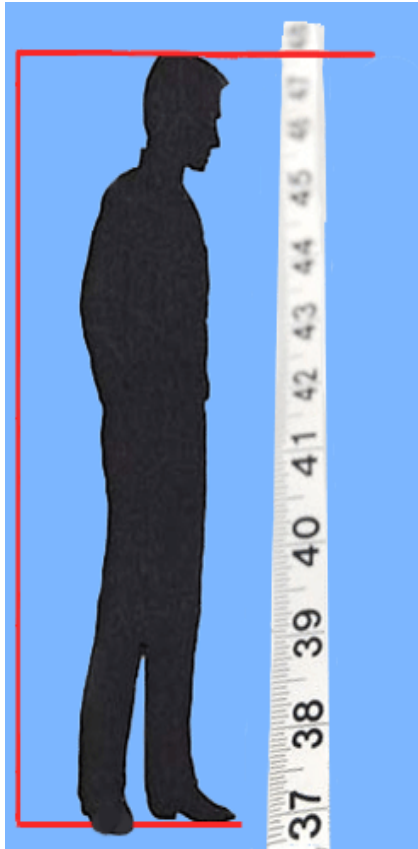
18



- Office is made up of 55 people. We want to determine the average height.
- If we take a random sample of 20 people and take the average height, we can say we are 80% confident that the average height of people in the office is 5'3" with a 2" margin of error.
- -OR – We are 80% confident that the interval 5'1" to 5'5" contains the true average height of all staff
- Conversely, there's a 20% chance that the interval 5'1" to 5'5" does not contain the true average height of all staff.

Example: Height

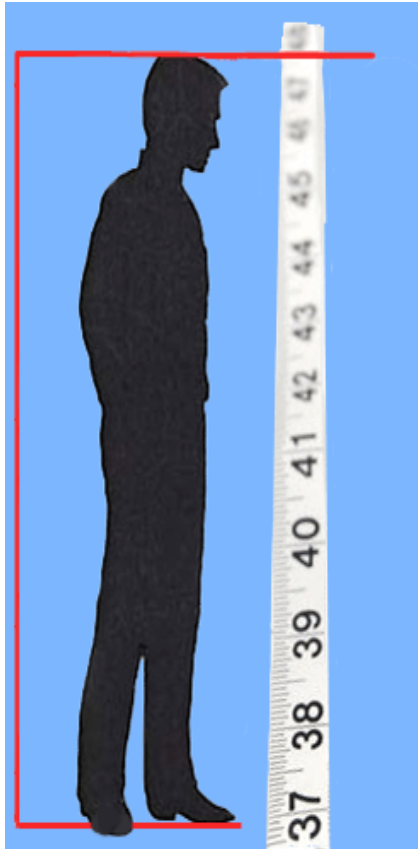
19



- Office is made up of 55 people. We want to determine the average height.
- If we increase the random sample to 30 people and take the average height, we can say we are 93% confident that the average height of people in the office is 5'3" with a 1" margin of error .
- -OR – We are 93% confident that the interval 5'2" to 5'4" contains the true average height of all staff
- Conversely, there's a 7% chance that the interval 5'2" to 5'4" does not contain the true average height of all staff.

Example: Height (Summary)

20



Sample Size:	20 people	30 people	↑
Average Height:	5' 3"	5' 3"	
Confidence Level:	80%	93%	↑
Margin of Error:	2"	1"	↓



Certainty vs. Precision

21

- Confidence Interval (Margin of Error) – A range of values estimated to contain the unknown population parameter.
 - ▣ Expresses the precision of an estimate.

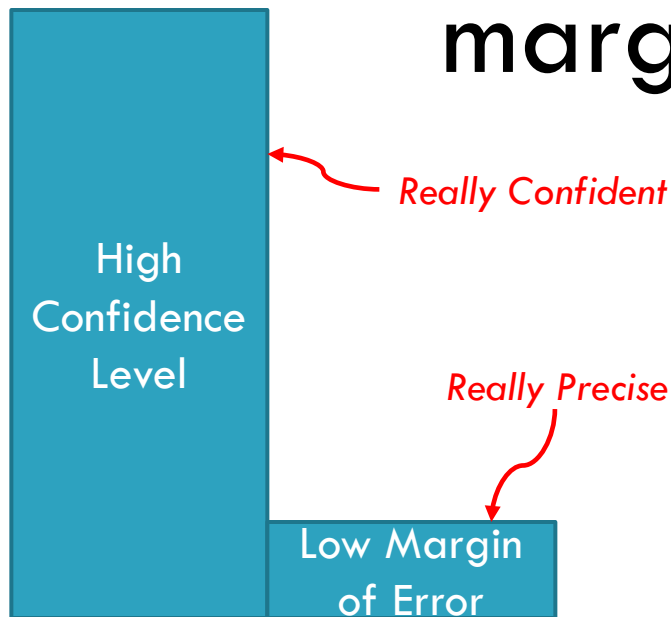
- Confidence Level – The probability that the confidence interval contains the true value of a parameter given many repeated samples.
 - ▣ Expresses certainty of an estimate.

You can CHOOSE how certain and how precise you want to be when creating your sample, but you usually will sacrifice one for the other.

Confidence Level and Margin of Error

22

What is the trade off of having a high confidence level and a small margin of error?



Sample Size (Categorical Variables)

23

$$n_o = \frac{(\text{Confidence Level z-score})^2 (.5)(.5)}{(\text{Margin of Error})^2}$$

z-score for 95%
Confidence Level

$$n_o = \frac{(1.96)^2 (.5)(.5)}{(.05)^2} = 384$$

Required sample
size

5% Margin of
Error

Sample Size (Categorical Variables)

24



$$n_o = \frac{(1.96)^2 (.5)(.5)}{(.05)^2} = 384$$

Increasing your
Confidence Level

$$n_o = \frac{(2.56)^2 (.5)(.5)}{(.05)^2} = 655$$

Increases your
sample size

$$n_o = \frac{(1.96)^2 (.5)(.5)}{(.10)^2} = 96$$

Increasing your
Margin of Error
(Wiggle Room)

Decreases your
sample size

Types of Variables

25

Variable – Any characteristic of an individual or record. Gender, income, inspection status, and age are all variables.

- Categorical Variable – A characteristic of an individual or record that falls into a category, e.g. gender, income range, inspection status.
- Continuous Variable – A characteristic of an individual or record that can be quantified in continuous terms, e.g. days to complete a process, amount of fees paid to a department.

Sample Size (Continuous Variables)

26

- Need standard deviation (σ)
 - ▣ It is unlikely you will know σ unless you conducted a pilot study or have historical data.
 - ▣ σ and margin of error need to be expressed in variable's units (e.g. feet, minutes, etc.)
 - ▣ See online calculator
<http://homepage.stat.uiowa.edu/~rlenth/Power>

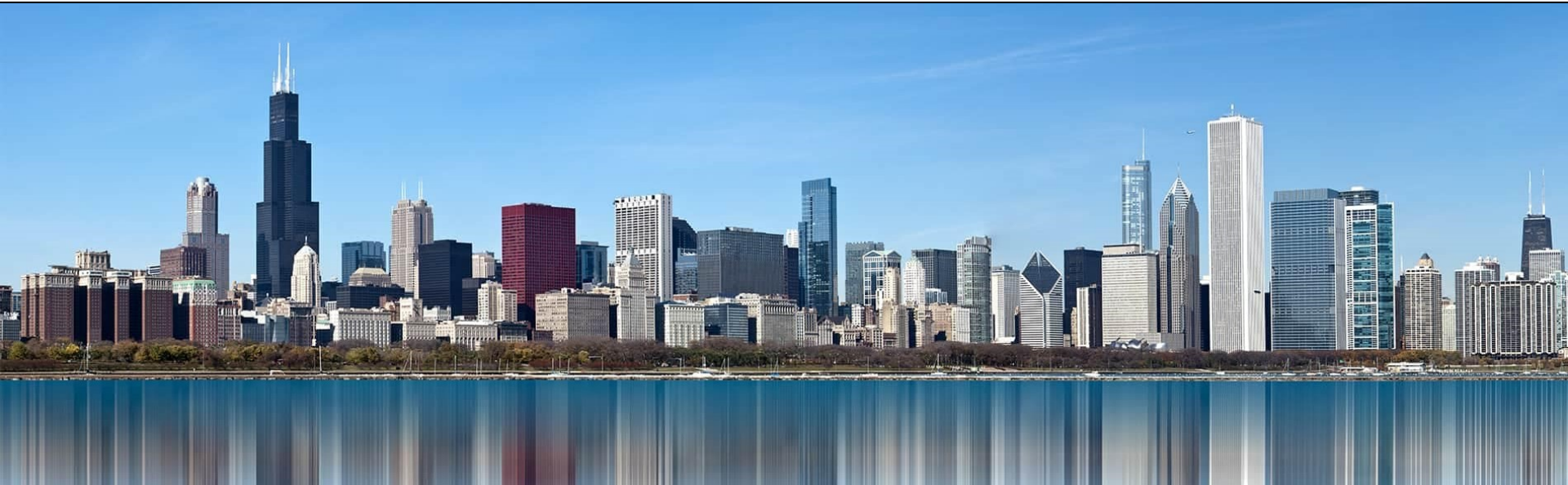
- ▣ Math:
$$n_0 = \frac{1.96 \times \sigma}{\text{Margin of Error}}$$

Real World Example

27

The City of Chicago has a Rental Subsidy Program to help low-income residents meet their housing needs. The City provides rent subsidies, via a Trust Fund, to landlords that provide affordable housing.

Our goal was to determine whether the buildings were inspected for minimum housing quality standards and met the City's Building Code.



Real World Example

28

- Identified the population of interest (598 participating buildings)
- If data was electronically captured, we would have tested all. However, documentation was largely in hard copy form.
- Chose a simple random sampling strategy to avoid bias.
- Calculated the appropriate sample size (including the Population Correction Formula).



Real World Example

29

$$\left. \begin{array}{l} \text{Population} = 598 \\ \text{Confidence Level} = 95\% \\ \text{Margin of Error} = 10\% \end{array} \right\} n_o = \frac{(1.96)^2 (.5)(.5)}{(.10)^2} = 96$$

Applying Population Correction Formula:

$$n = \frac{96}{1 + \left(\frac{(96 - 1)}{598} \right)} = 83$$

Therefore, a sample size of 83 is needed to be 95% confident that the results fall within + or - 10% of the true value in the population.

What are the Chances?

30

With
Replacement:



1st Pull =
25%



2nd Pull =
25%



3rd Pull =
25%



4th Pull =
25%



Without
Replacement:

1st Pull =
25%



2nd Pull =
33%



3rd Pull =
50%



4th Pull =
100%

Population Correction
Formula

Real World Example - RESULTS

31

Of the
83

buildings sampled,

38 did not meet
minimum housing quality
standards; and

51 had unresolved
building code violations.



45.8%

61.4%

Therefore, of the

598

total buildings,

274 did not meet
minimum housing quality
standards; and

367 had unresolved
building code violations.

Taking a Random Sample in Excel

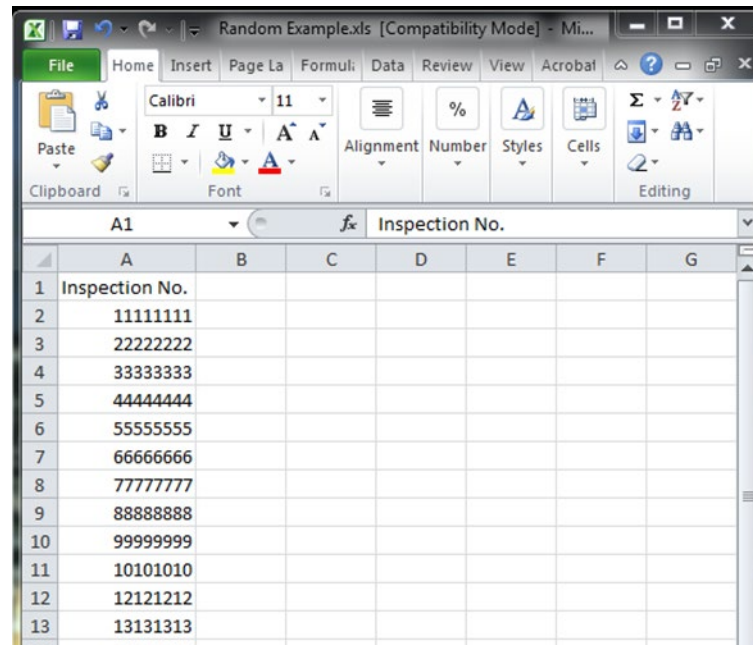
32

1. Open the worksheet containing the whole population that you wish to sample.
2. Add a column in the worksheet. Name it *Random_number*.
3. In the first cell of *Random_number*, enter the formula **=RAND()**. This generates a random number between 0 and 1.
4. Copy the formula to all cells in the column.
5. Sort the entire worksheet by the values in *Random_number*.
6. The first X lines (where X = desired sample size) is your random sample.

Taking a Random Sample in Excel

33

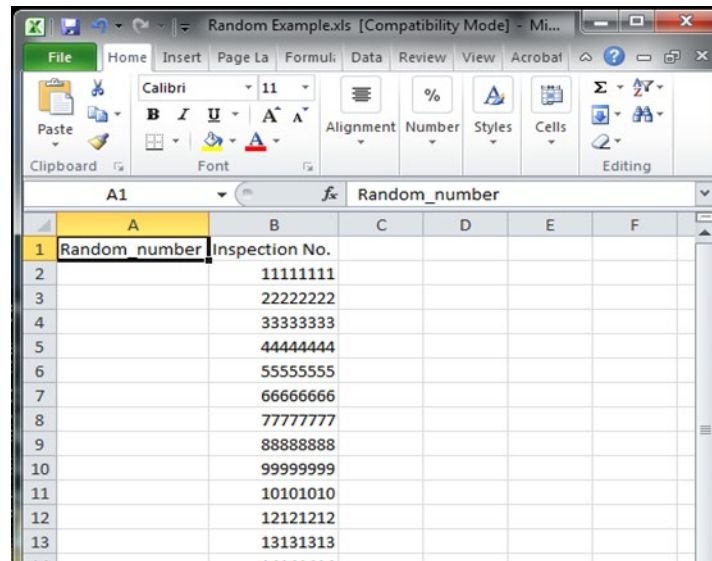
1. Open the worksheet containing the whole population that you wish to sample.



Taking a Random Sample in Excel

34

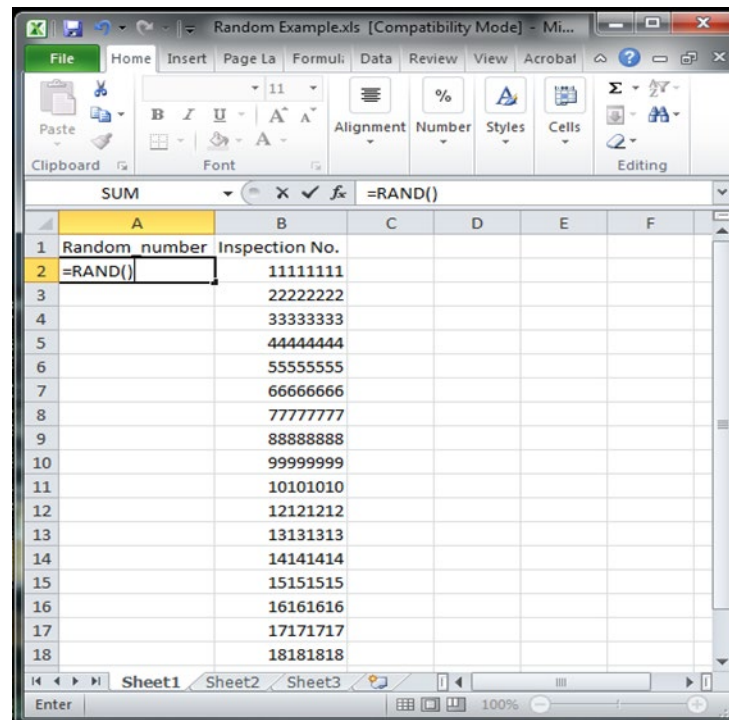
2. Add a column in the worksheet. Name it Random_number.



Taking a Random Sample in Excel

35

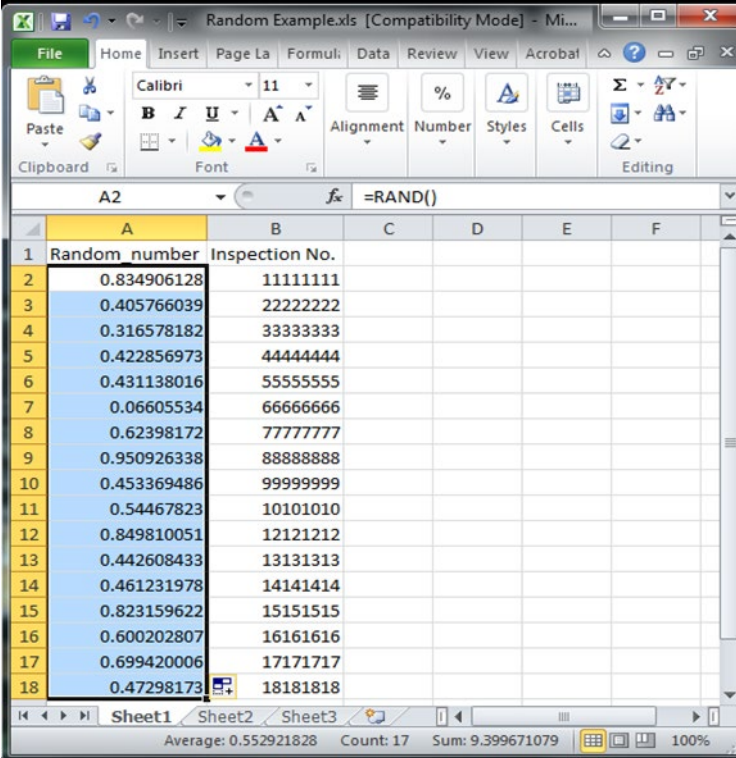
3. In the first cell of Random_number, enter the formula `=RAND()`. This generates a random number between 0 and 1.



Taking a Random Sample in Excel

36

4. Copy the formula to all cells in the column.



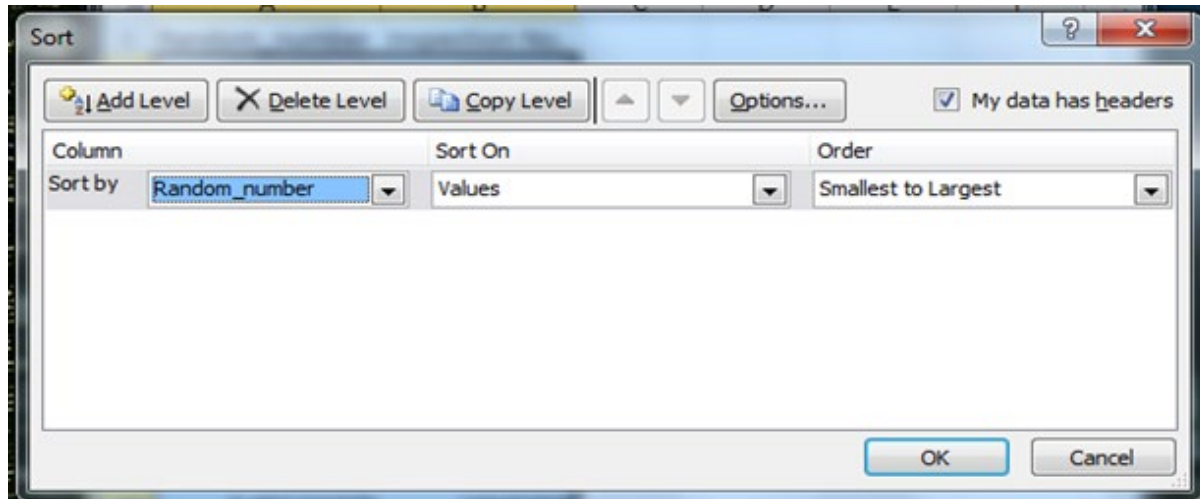
The screenshot shows an Excel spreadsheet titled "Random Example.xls [Compatibility Mode] - Mi...". The ribbon includes File, Home, Insert, Page Layout, Formulas, Data, Review, View, and Acrobat. The Home ribbon is active, showing Font, Paragraph, Styles, and Editing groups. The formula bar shows the formula `=RAND()` in cell A2. The spreadsheet has columns A through F and rows 1 through 18. Column A is labeled "Random number" and column B is labeled "Inspection No.". The data in column A consists of 17 random decimal values, and the data in column B consists of 17 corresponding 7-digit integers. The status bar at the bottom shows "Average: 0.552921828", "Count: 17", "Sum: 9.399671079", and "100%".

	A	B	C	D	E	F
1	Random number	Inspection No.				
2	0.834906128	11111111				
3	0.405766039	22222222				
4	0.316578182	33333333				
5	0.422856973	44444444				
6	0.431138016	55555555				
7	0.06605534	66666666				
8	0.62398172	77777777				
9	0.950926338	88888888				
10	0.453369486	99999999				
11	0.54467823	10101010				
12	0.849810051	12121212				
13	0.442608433	13131313				
14	0.461231978	14141414				
15	0.823159622	15151515				
16	0.600202807	16161616				
17	0.699420006	17171717				
18	0.47298173	18181818				

Taking a Random Sample in Excel

37

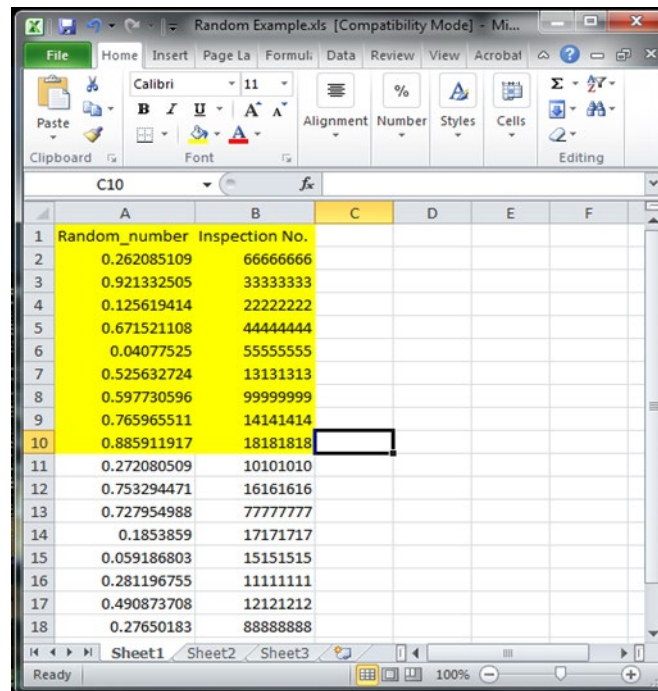
5. Sort the entire worksheet by the values in Random_number.



Taking a Random Sample in Excel

38

6. The first X lines (where X = desired sample size) is your random sample.



	A	B	C	D	E	F
1	Random_number	Inspection No.				
2	0.262085109	66666666				
3	0.921332505	33333333				
4	0.125619414	22222222				
5	0.671521108	44444444				
6	0.04077525	55555555				
7	0.525632724	13131313				
8	0.597730596	99999999				
9	0.765965511	14141414				
10	0.885911917	18181818				
11	0.272080509	10101010				
12	0.753294471	16161616				
13	0.727954988	77777777				
14	0.1853859	17171717				
15	0.059186803	15151515				
16	0.281196755	11111111				
17	0.490873708	12121212				
18	0.27650183	88888888				

Questions?

39

