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 CIGE Institute – Jacksonville, FL

Course Objectives

- 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

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

Definition

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.

Why Sample Statistically?

- 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?

"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

- 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

- 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

- 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

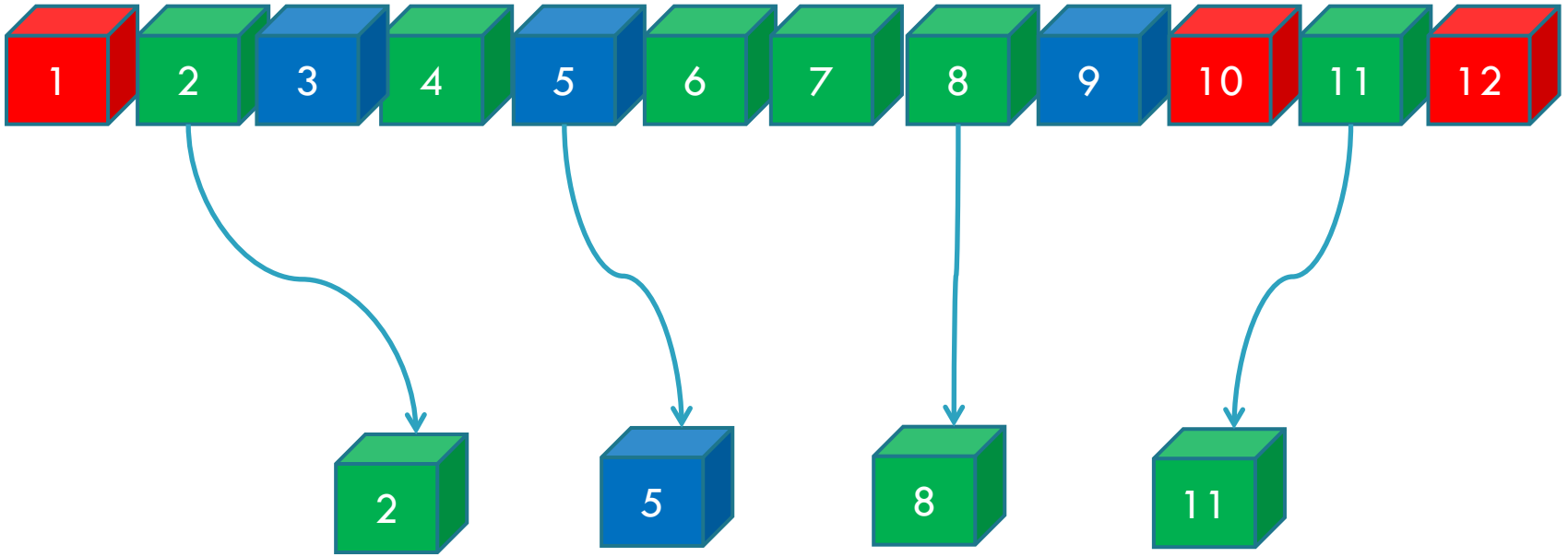
- Select a sampling strategy that minimizes selection bias.

(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

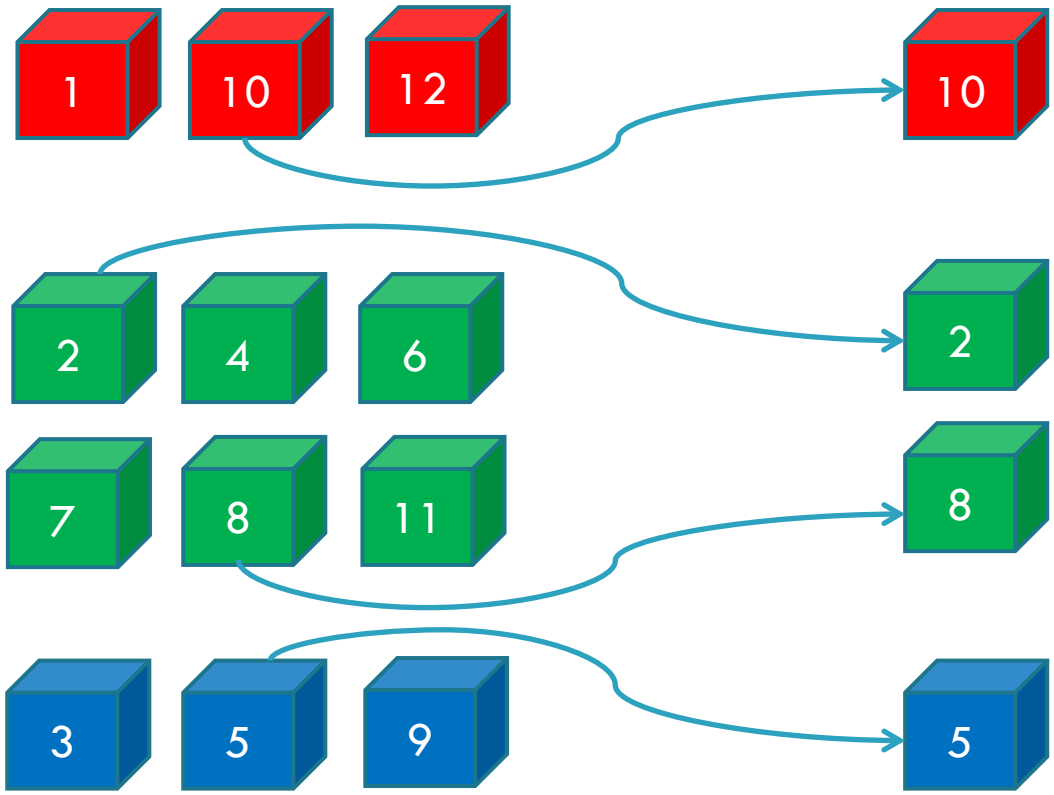
Simple Random Sampling

Systematic Random Sampling



Every nth item is selected after a random start.

Stratified Random Sampling



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

- 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

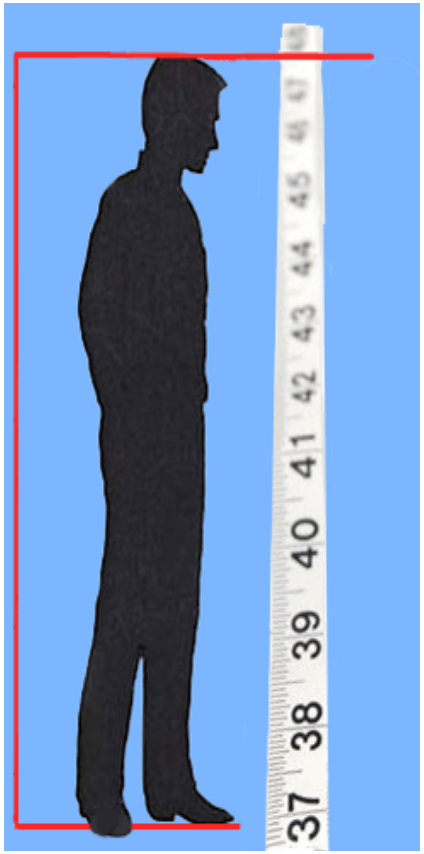
Confidence Level and Margin of Error

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?)

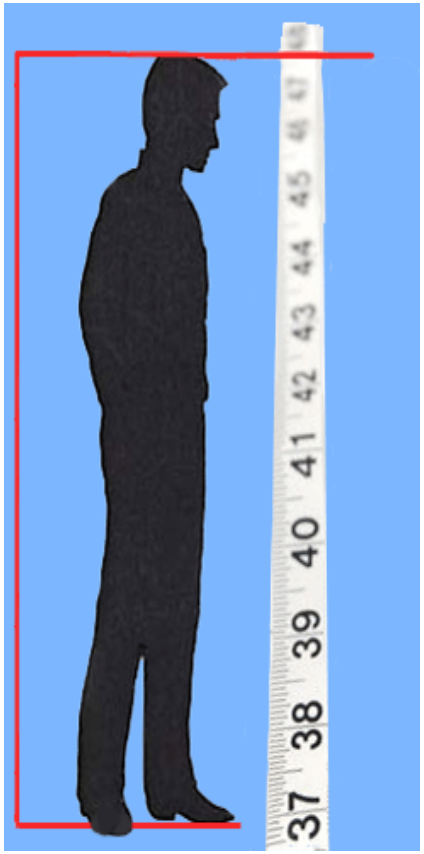
(There is always a balance between how certain one is that the sample is representative of the population and how large the sample should be.)

Example: Height



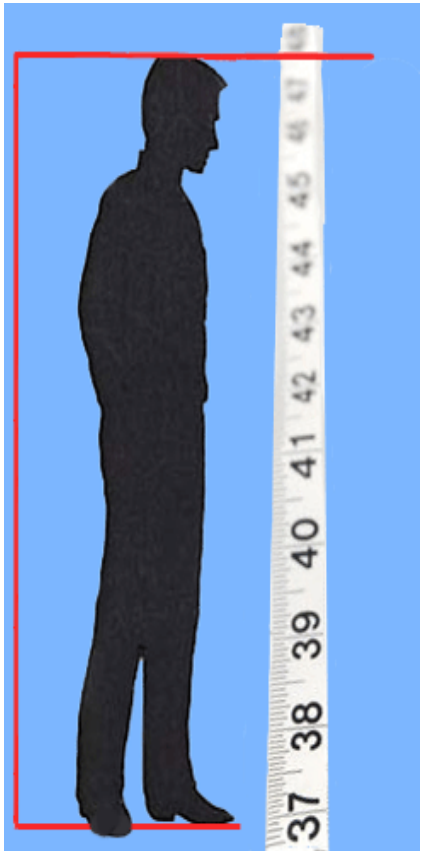
- Office is made up of 55 people. We want to determine the average height.
- If we take a *random sample of 20 people* and take the average height, we can say we are *80% confident* that the average height of people in the office is 5'3" with a *2" margin of error*.
- -OR – We are 80% confident that 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



- 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)



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

Certainty vs. Precision

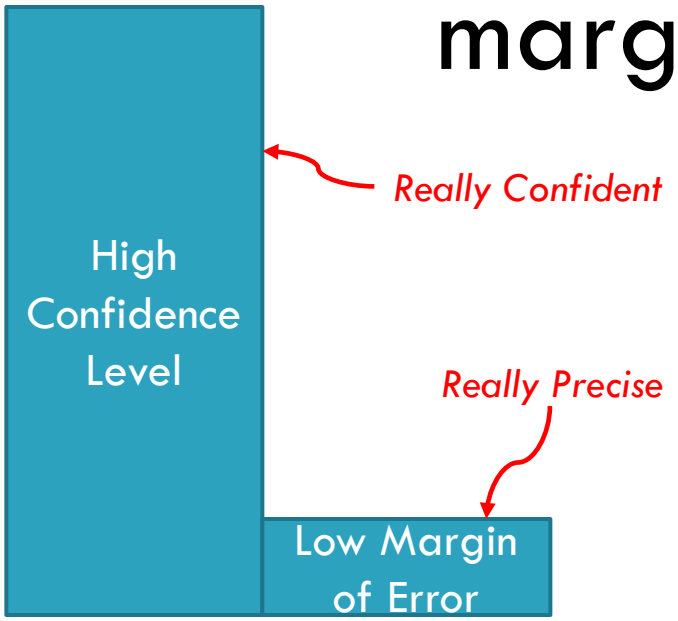
- Confidence Interval (Margin of Error) – A range of values estimated to contain the unknown population parameter.
  - Expresses the precision of an estimate.

- 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

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



Sample Size (Categorical Variables)

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

z-score for 95% Confidence Level

Required sample size

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

5% Margin of Error

Sample Size (Categorical Variables)

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

Increasing your Confidence Level

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

Increases your sample size

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

Decreases your sample size

Increasing your Margin of Error (Wiggle Room)

Types of Variables

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.

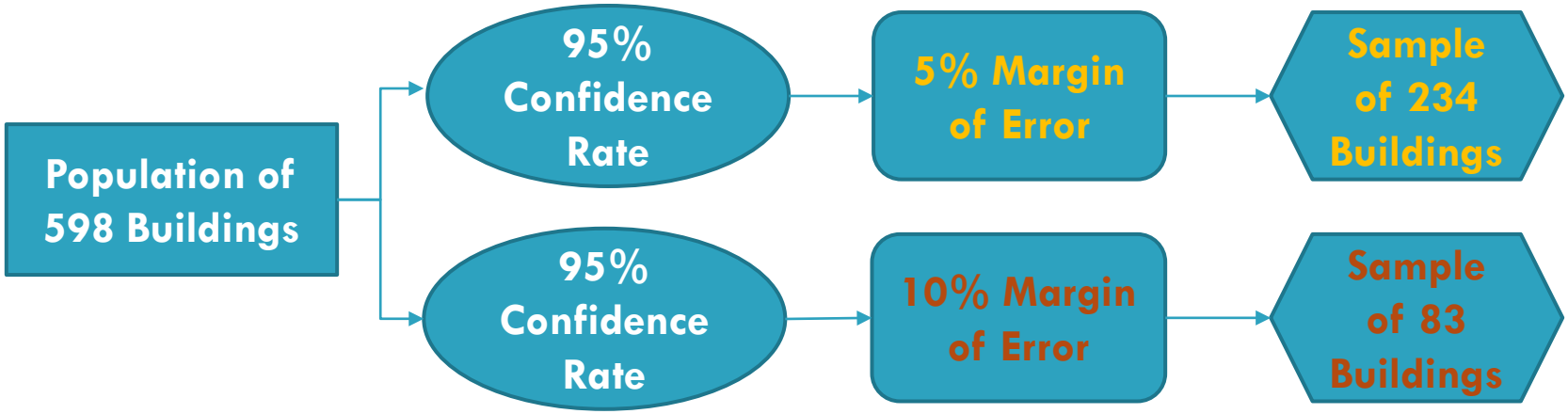
Real World Example

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

- 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

Population = 598
Confidence Level = 95%
Margin of Error = 10%

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

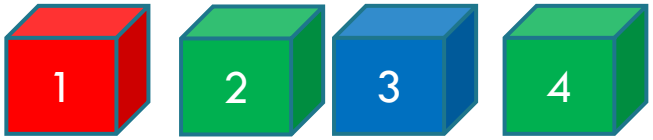
Applying Population Correction Formula:

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

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?

With Replacement:



Without Replacement:

1st Pull = 25%

2nd Pull = 25%

3rd Pull = 25%

4th Pull = 25%

Population Correction Formula

1st Pull = 25%

2nd Pull = 33%

3rd Pull = 50%

4th Pull = 100%

Real World Example - RESULTS

Of the

83

buildings sampled,

Therefore, of the

598

total buildings,

38 did not meet minimum housing quality standards; and

45.8%

274 did not meet minimum housing quality standards; and

51 had unresolved building code violations.

61.4%

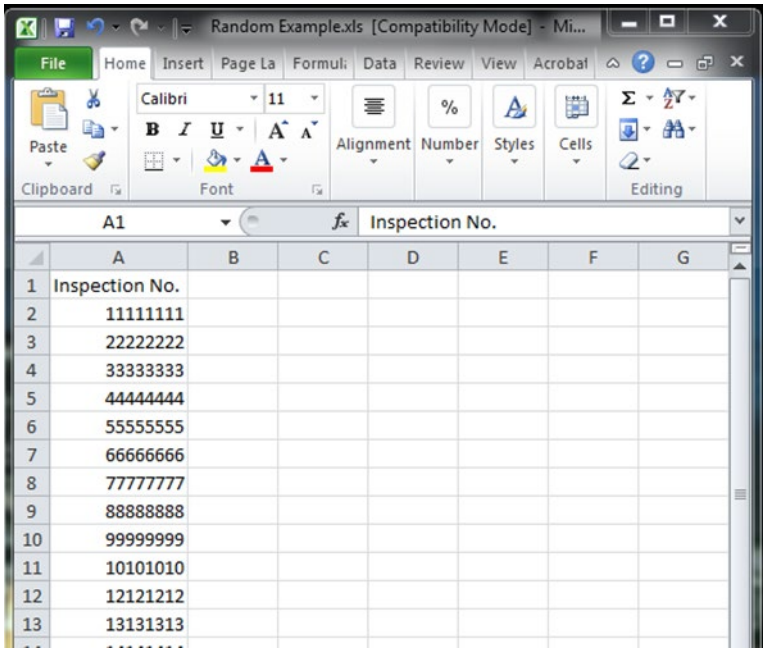
367 had unresolved building code violations.

Taking a Random Sample in Excel

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.

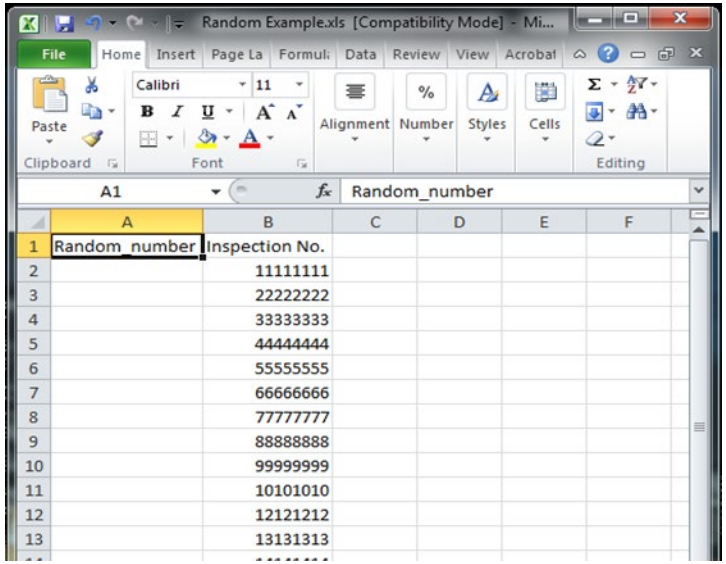
Taking a Random Sample in Excel

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



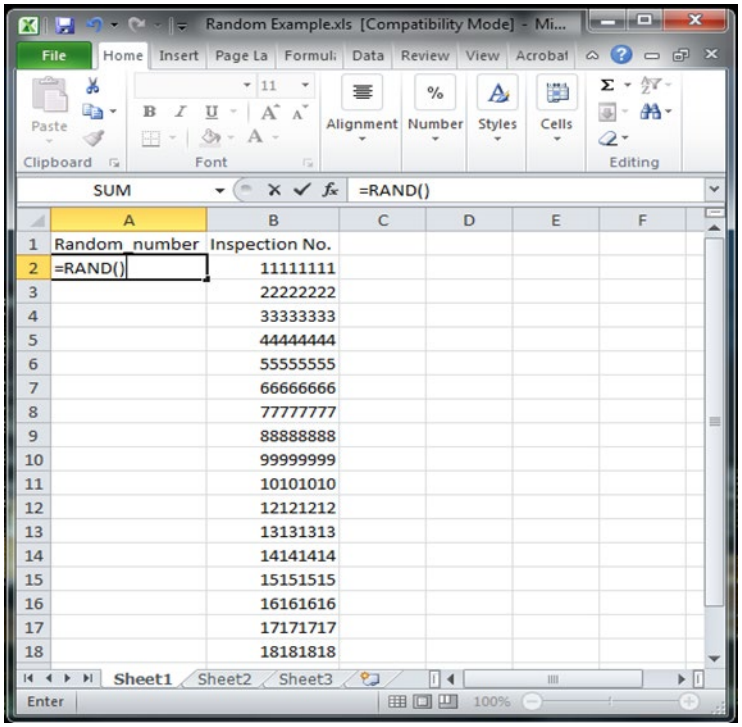
Taking a Random Sample in Excel

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



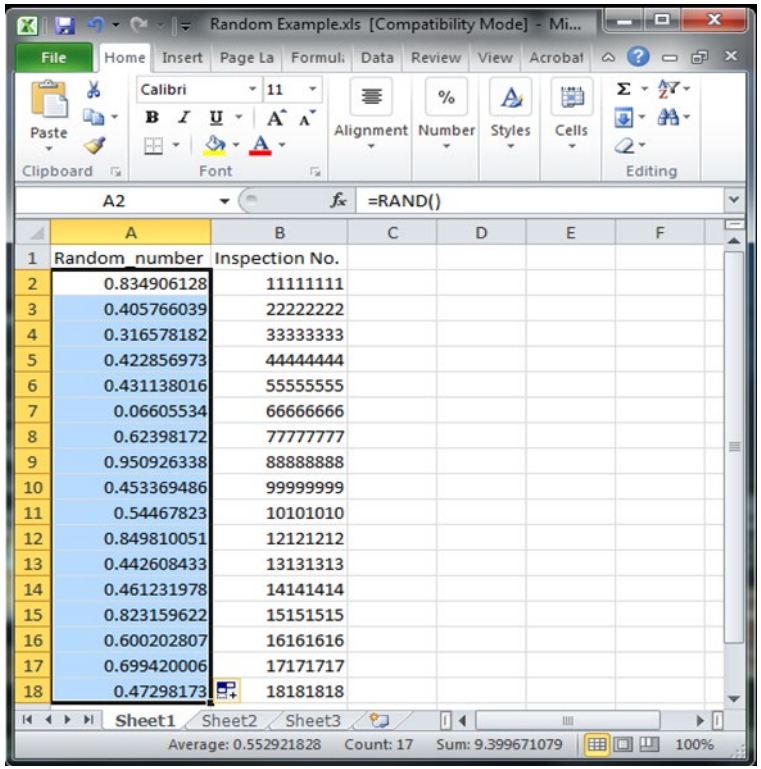
Taking a Random Sample in Excel

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



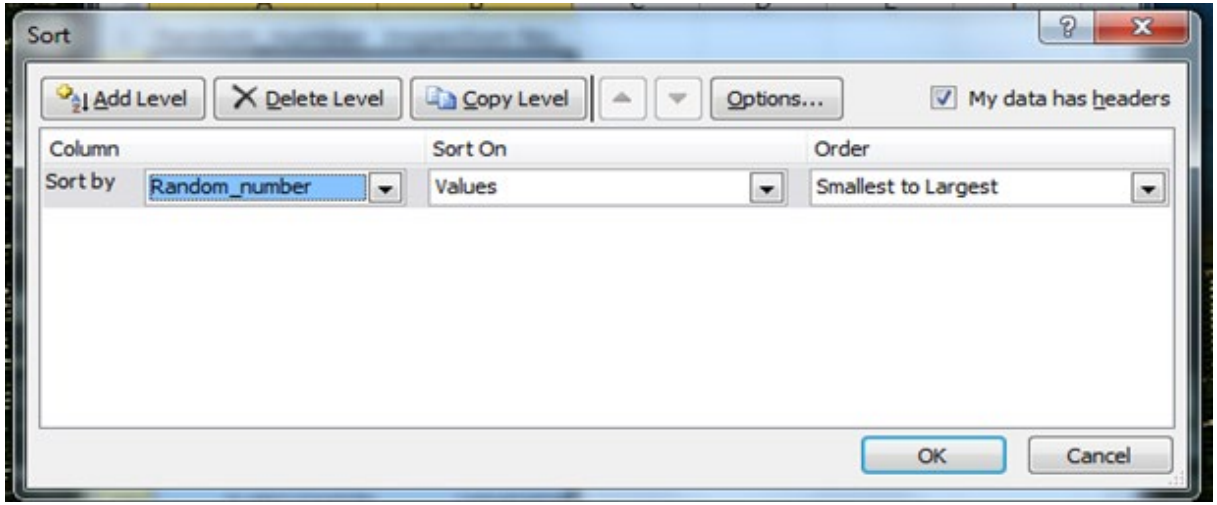
Taking a Random Sample in Excel

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



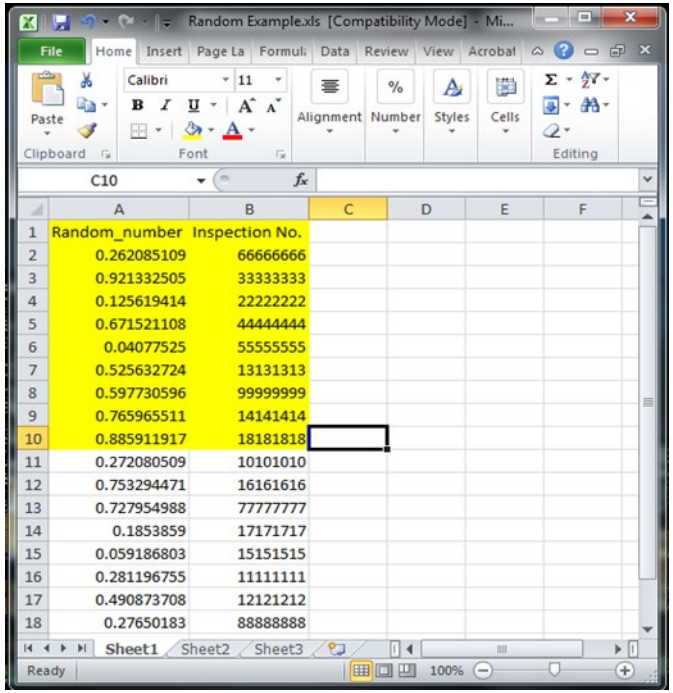
Taking a Random Sample in Excel

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



Taking a Random Sample in Excel

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



Questions?