The opinions expressed in this presentation are the author's own and do not reflect the view of the Department of Health and Human Services, Office of Inspector General or the United States government.

Application of Statistics to Auditing

Jared Smith, PhD

Three Sections

- Part 0: Steps to Execute a Statistical Sample
- Part 1: Key Concepts in Statistics
- Part 2: Deciding on a Sample Size

Steps to Execute a Statistical Sample

The nuts and bolts of things.

Sampling Steps

- Step 0: Define Your Problem
- Step 1: Select a Sampling Frame
- Step 2: Select a Sample Design
- Step 3: Select a Sample Size
- Step 4: Generate Your Random Numbers
- Step 5: Select Your Sample

Step 1: Select a Sampling Frame

- Your statistical estimates only apply to your frame!
- Quality of frame directly impacts your estimate.
  - E.g., duplicates, irrelevant items, items accidently left out of frame, etc.
- As a best practice, the data file used for sampling should be saved in a read only format so the sample can easily be recovered.
- The frame should be numbered given the ordering used to pull the actual sample.

Specific Audit Considerations

- Retain your frame and the files used to create it.
- Review any frame related work papers prior to pulling your sample.
- Limit filtering of the frame if your goal is to provide a representative error rate (Purchase Card Example).
- Take steps to ensure the reliability of the frame (Inventory Example).

Step 2: Select a Sample Design

- Splitting your population into pieces (strata) can increase the accuracy of your estimates.
- Clustering can decrease the time required to complete your sample (Property Example).
- Most mistakes for new users come from using stratification and/or clustering!
- If you do not have a technical expert on hand, it is best to start with a simple random sample.

Audit Considerations

- There is a significant risk of mistakes with more complex designs (two examples).
- Clustered designs can produce extremely poor results (County example).
- Stratification can produce poor results if the samples are not split across the strata appropriately (Purchase Card example).
- You should decide on the analyses that you plan to run prior to pulling your sample.

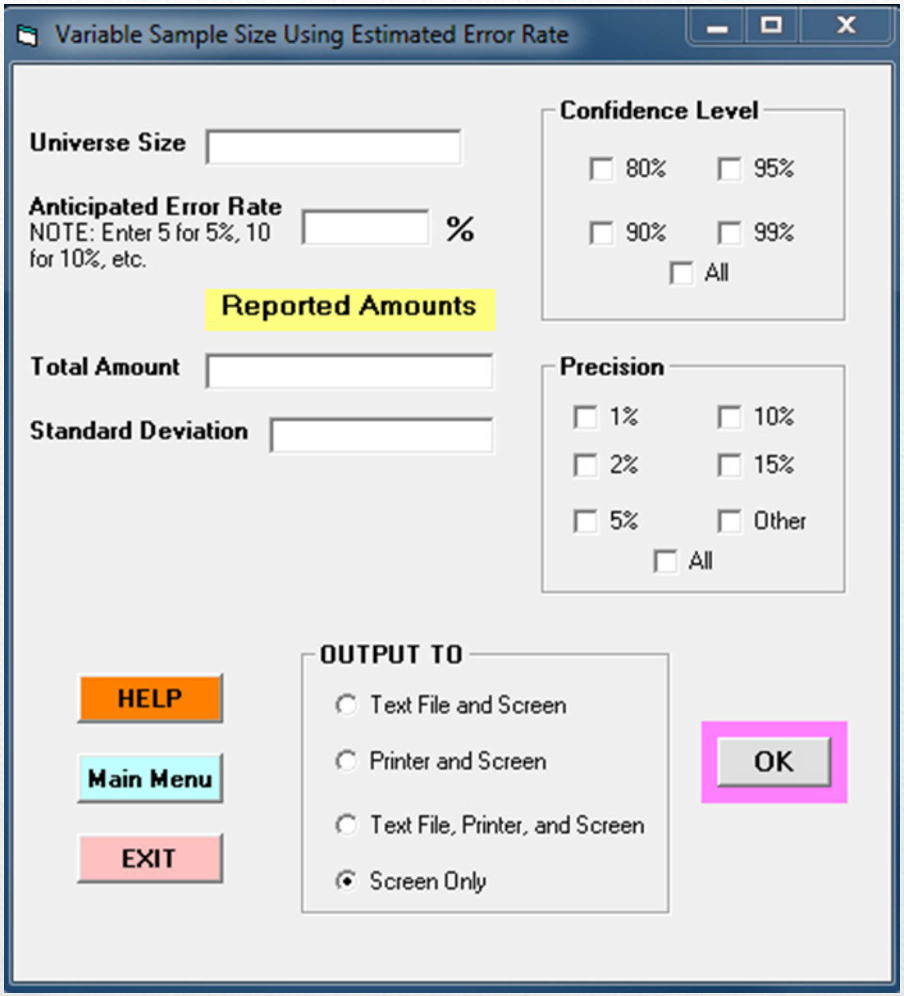
Step 3: Select a Sample Size

- We will discuss this topic at length later today.
- **Key point**: Deciding on a sample size involves judgment and is not simply driven by equations (Contractor story).
- When agencies set a minimum sample size that often becomes the default sample size.
- You must decide on your confidence level before pulling your sample.

Audit Considerations

- Resource constraints often put strict limits on the sample size.
- Larger sample sizes help protect against unforeseen events (Personnel Record Example).
- While it may seem efficient, choosing your sample size as you go is against the rules of statistics!

Variable Sample Size Determination



Calculated:

Universe Size

Total Amount

Standard Deviation of Amount

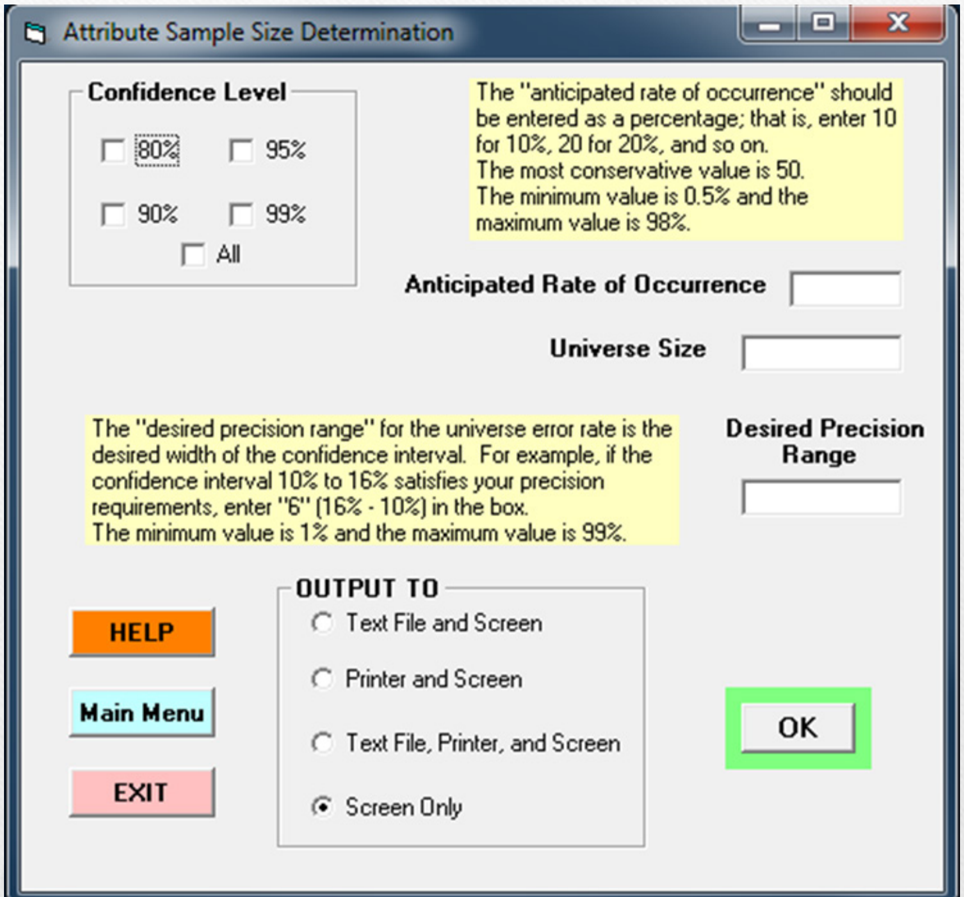
Decided On:

Confidence Level

Precision (Half witdth)

Anticipated Error Rate

Attribute Sample Size Determination



Calculated:
Universe Size

Decided On:
Confidence Level
<u>Precision (Full Width)</u>
Anticipated Error Rate

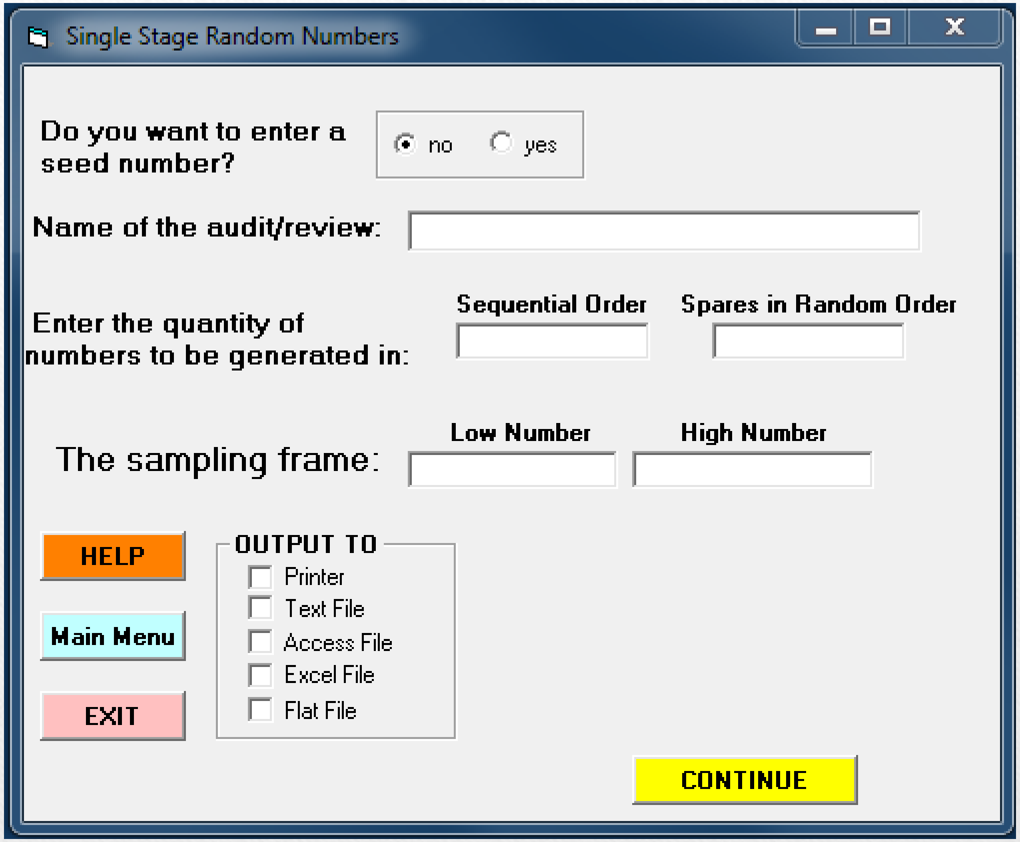
Step 4: Generate Random Numbers

- Some packages like SAS will select the sample directly. Other tools such as RAT-STATS generate random numbers that can then be matched against a numbered sampling frame.

Audit Considerations

- You should use a software package that allows you to replicate your sample through the use of a random seed (e.g., SAS, RAT-STATS, R).
- In addition, be aware that some packages have flawed random number generators (e.g. Rnd in Excel).
- I recommend not using spares without guidance from a statistician.

Random Number Generator



Calculated:
Low Number
High Number

Decided On:
Sequential Order (Sample Size)
Spares in Random Order

Step 5: Select Your Sample

- If you use a package that does not select your sample directly, then you will have to match up the random numbers generated by your software against your frame.

Audit Considerations

- Save the numbered version of the frame used to select your sample.

Analysis Steps

- Step 6: Review Your Sample
- Step 7: Analyze Your Results
- Step 8: Interpret Your Results
- Step 9: Reporting

Step 7: Review Your Sample

- You must review all items selected. Your estimates will be invalid if your stop your review early.
- For dollar estimates, code irrelevant transactions as non-errors (IT Contract Example).
- For percent estimates, you need to distinguish between irrelevant and inaccessible items.
  - Irrelevant – Outside your scope
  - Inaccessible – Inside your scope but cannot be tested (e.g., Fuel Tank Example).

Audit Considerations

- The validity of a statistical estimate depends on the soundness of the underlying audit work (Training Example).
- While performing your audit tests, it can be useful to also check the reliability of your data.

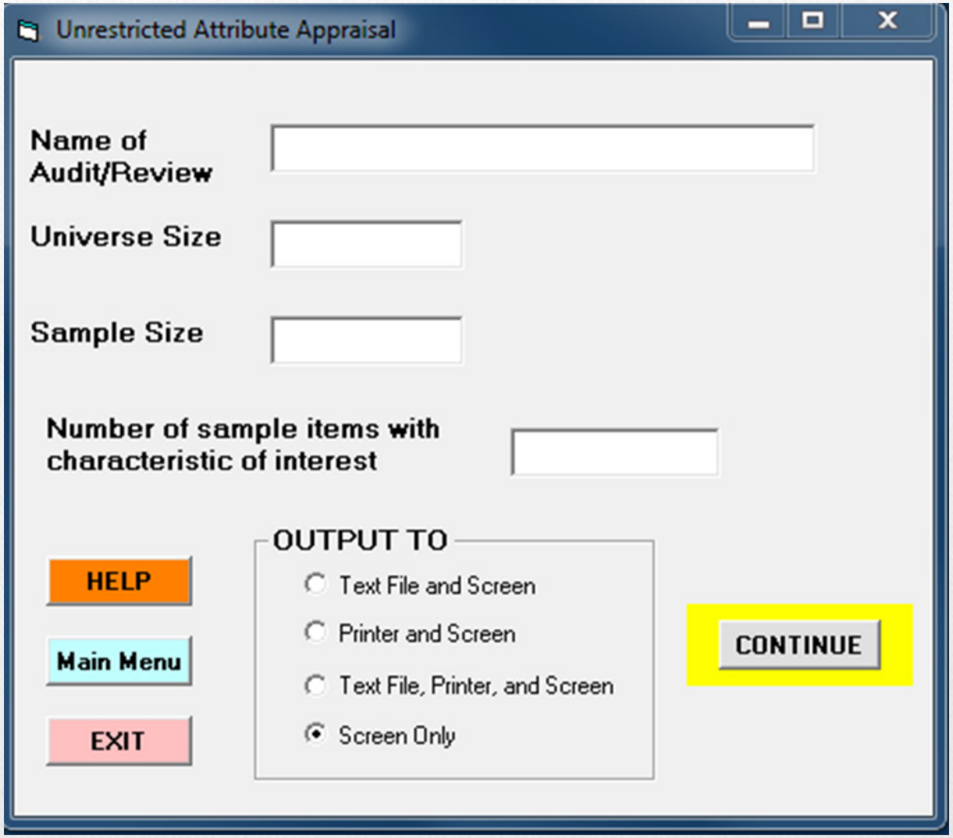
Step 7: Analyze Your Results

- Do no try calculating results by hand or informally in Excel; instead, use a formal software package.
- Ensure your analysis accounts for all aspects of your sample design (e.g. stratification, clustering, spares, etc.).
- For dollar and count estimates do not exclude any sampled items.
- If you did not use a simple random sample, be very careful when referencing sample percentages.

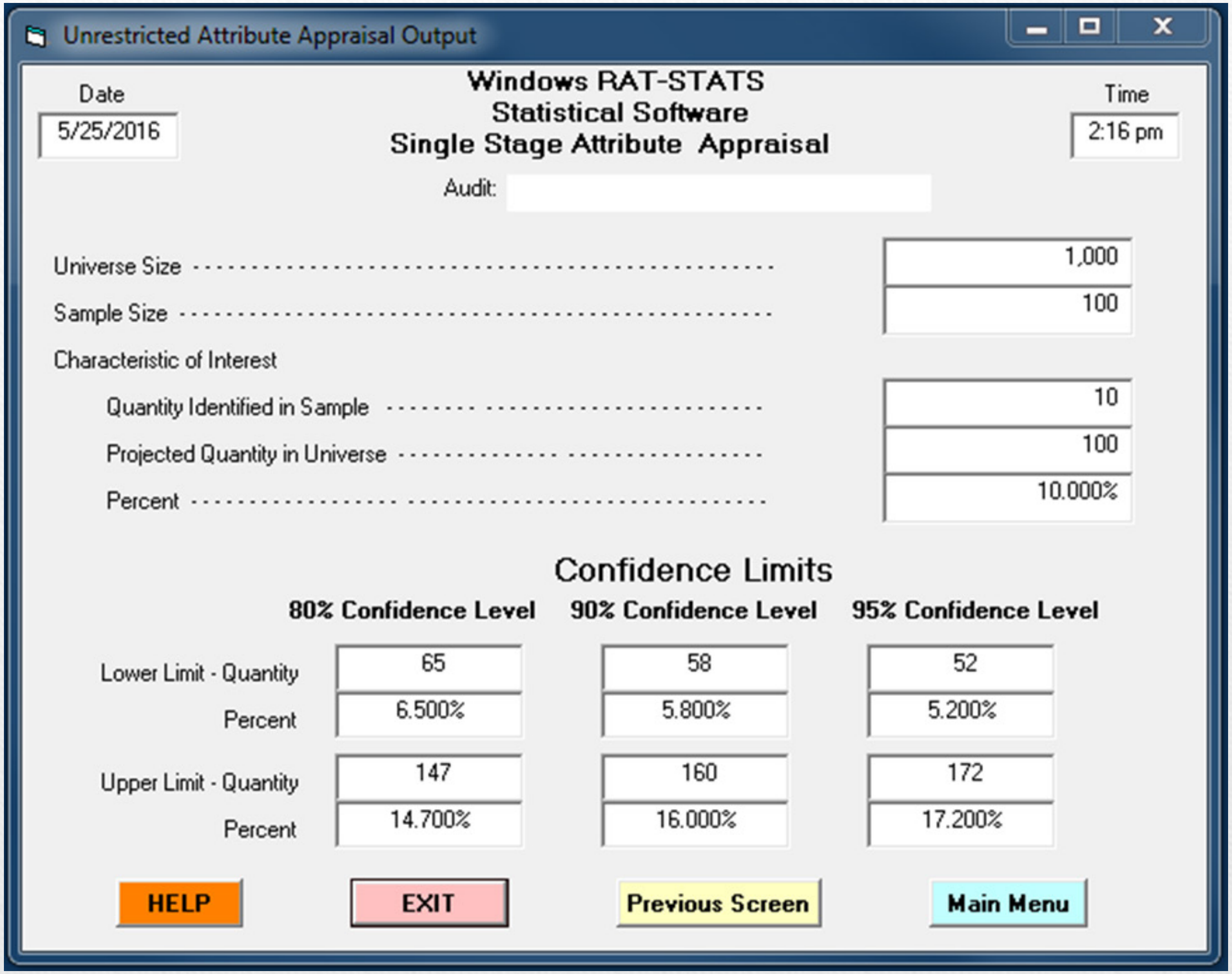
Audit Considerations

- When calculating error percentages, think very carefully about what they mean in reference to your frame (Safety Check Example).
- Use the confidence level you decided on prior to pulling the sample.
- Remember to update your statistical estimates whenever your sample results change!

Attribute Testing Example



Attribute Testing Results



Unrestricted Attribute Appraisal Output

Windows RAT-STATS
Statistical Software
Single Stage Attribute Appraisal

Date: 5/25/2016

Time: 2:16 pm

Audit:

Universe Size	1,000
Sample Size	100
Characteristic of Interest	
Quantity Identified in Sample	10
Projected Quantity in Universe	100
Percent	10.000%

Confidence Limits

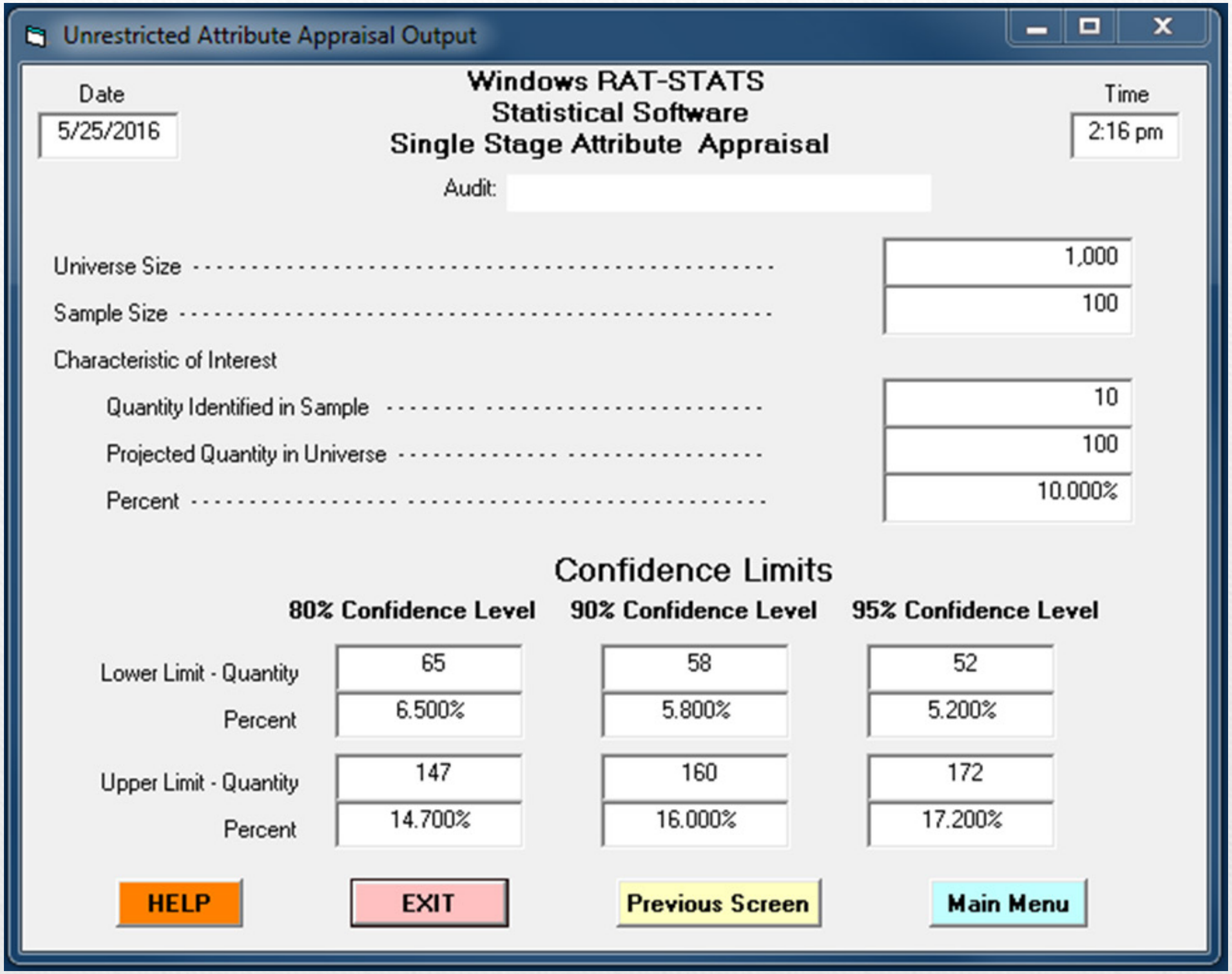
	80% Confidence Level	90% Confidence Level	95% Confidence Level
Lower Limit - Quantity	65	58	52
Percent	6.500%	5.800%	5.200%
Upper Limit - Quantity	147	160	172
Percent	14.700%	16.000%	17.200%

HELP | EXIT | Previous Screen | Main Menu

Step 8: Interpreting Your Results

- **Point Estimate** - Your best guess (might not be very good so how do we get assurance?)
- **Confidence Interval** – A range of values that gives you the assurance you are looking for.
- **Confidence Level** – The amount of assurance you are looking for.
- *****Precision** – The width of the confidence interval.

Attribute Testing Results



Audit Considerations

- Is the lower limit materially different from the upper limit? If there is materially difference, it is a problem since your sample results cannot rule out two materially different conclusions (Wait Times example).
- A limit is usually used when an audit agency wants to use a conservative estimate (Financial example).
- The point estimate is preferred when a conservative estimate is not necessary and precision is reasonable, which is a judgment call.

Step 9: Reporting

- **Point Estimate** – We estimate that…
- **Lower Limit** – We estimate that at least…
- **Upper Limit** – We estimate that at most
- Regardless of what is presented in text, it is standard practice to include the point estimate and confidence interval somewhere in the report (e.g. an appendix, footnote, etc.).
- Agencies differ on amount of detail they include in their reports.

Step 9: Reporting

- Commonly presented information (not definitive):
  - Description of sample frame (e.g. record count, dollar value)
  - Manner which sampling frame differs from target group of interest
  - Confidence level
  - Sample design (with strata breakdown when relevant)
  - Sample unit
  - Estimation method

Concept Overview

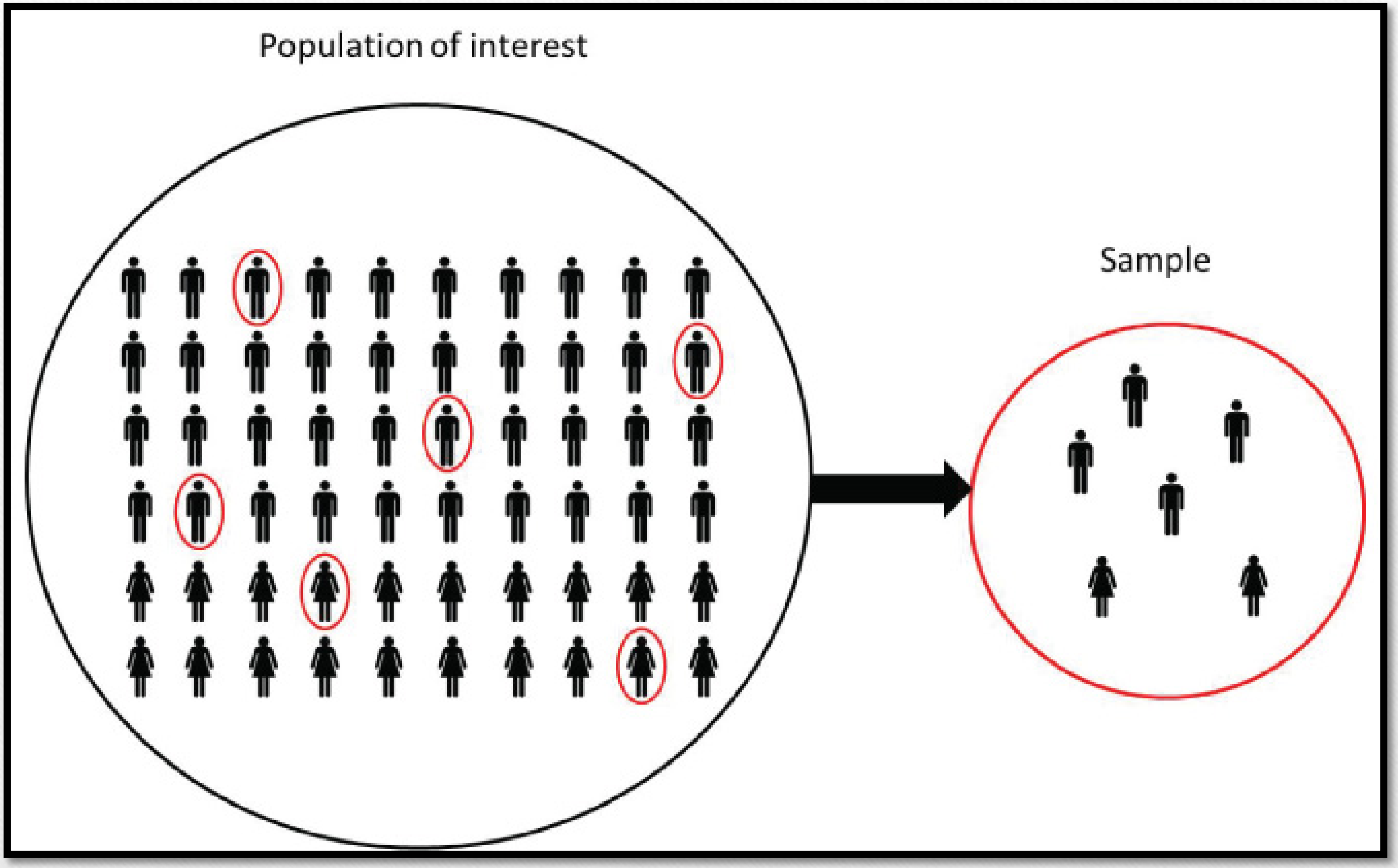
Taking it Higher

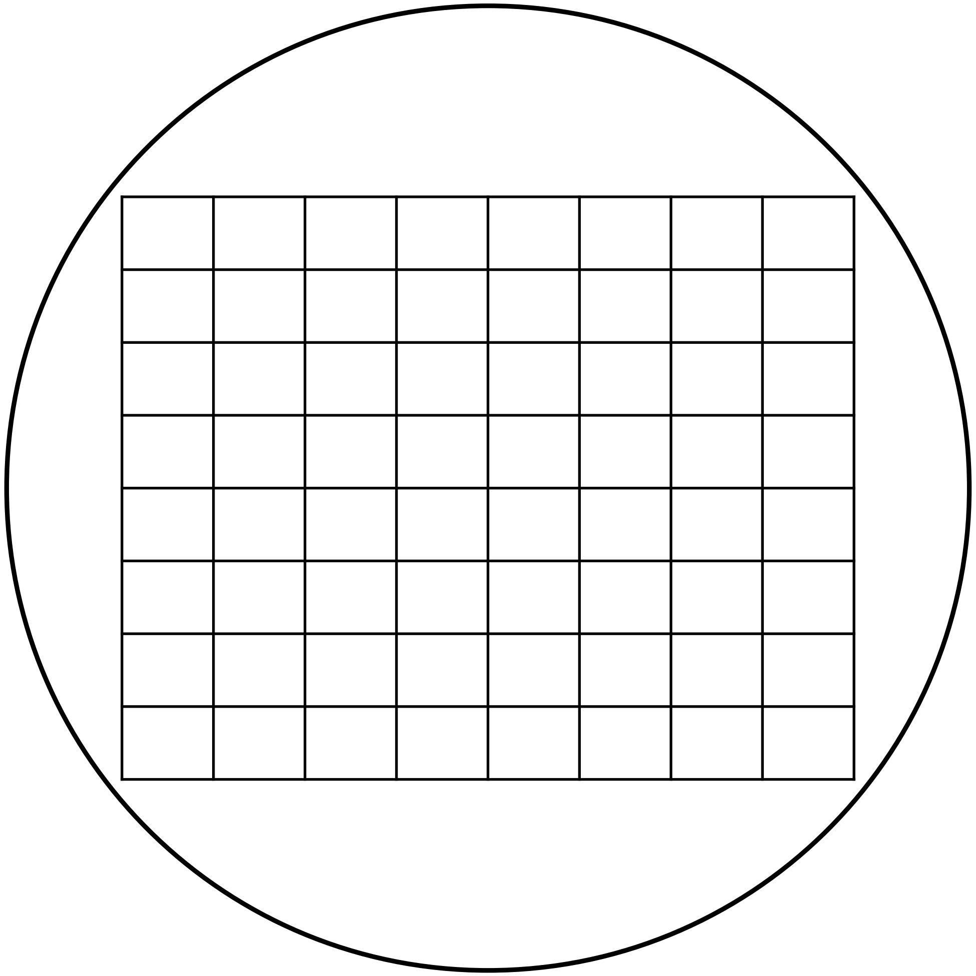
Concepts Roadmap

- General Concepts
  - Overview of Sampling
  - Why we need statistics
- Common Misconceptions
  - Sample Coverage (A Drop in the Bucket)
  - Snowflakes (Bundles of Uniqueness)
  - Frame Distribution (Central Limit Magic)
  - Randomness (How Random is Random)

General Concepts

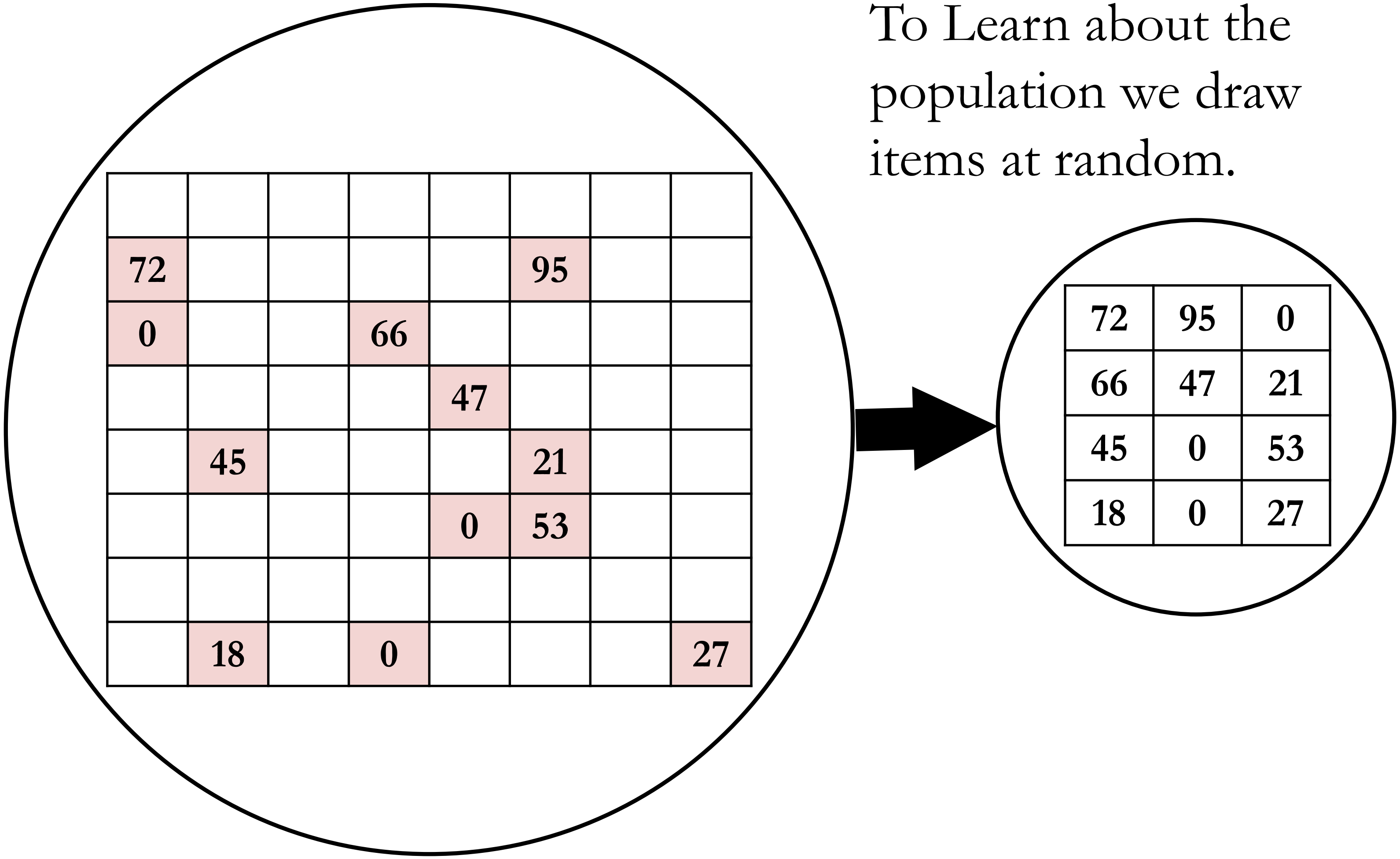
A Little Theory


Population of interest
Sample



We start with a group of items that we want to learn about.

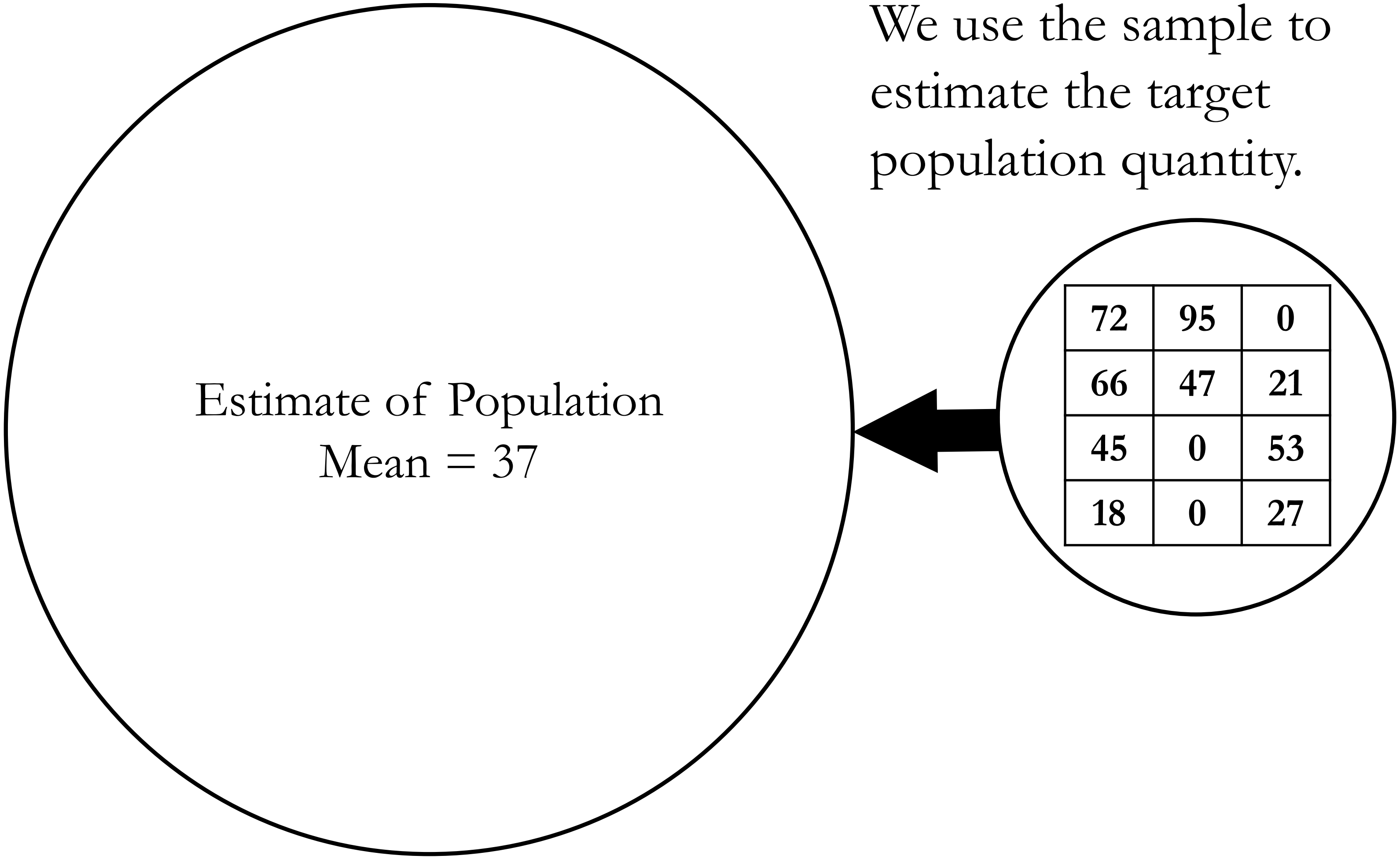
To Learn about the population we draw items at random.



Critical Moment

- The critical moment is **when the sample is selected**.
- At that moment we could have any random sample of items.
- We do not want our results to depend on which random sample we might happen to pull.

We use the sample to estimate the target population quantity.



72	**95**	**0**
66	**47**	**21**
45	**0**	**53**
18	**0**	**27**

The real magic is the ability to assess the quality of our estimate.

Estimate of Population
Mean = 37

Estimated Precision = ± 14

72	95	0
66	47	21
45	0	53
18	0	27

Red shading indicates sampled items.

90	0	16	64	0	32	44	31
72	99	22	96	71	95	51	0
0	62	62	66	57	73	59	46
64	0	89	36	47	20	0	10
2	45	94	11	0	21	11	35
28	43	67	66	0	53	44	0
0	2	66	98	64	88	0	88
50	18	88	0	87	93	95	27

72	95	0
66	47	21
45	0	53
18	0	27

Red shading indicates sampled items.

90	0	16	64	0	32	44	31
72	99	22	96	71	95	51	0
0	62	62	66	57	73	59	46
64	0	89	36	47	20	0	10
2	45	94	11	0	21	11	35
28	43	67	66	0	53	44	0
0	2	66	98	64	88	0	88
50	18	88	0	87	93	95	27

99	0	10
2	94	21
0	66	64
87	95	27

Red shading indicates sampled items.

90	0	16	64	0	32	44	31
72	99	22	96	71	95	51	0
0	62	62	66	57	73	59	46
64	0	89	36	47	20	0	10
2	45	94	11	0	21	11	35
28	43	67	66	0	53	44	0
0	2	66	98	64	88	0	88
50	18	88	0	87	93	95	27

16	64	72
95	0	0
47	20	53
98	88	50

Red shading indicates sampled items.

90	0	16	64	0	32	44	31
72	99	22	96	71	95	51	0
0	62	62	66	57	73	59	46
64	0	89	36	47	20	0	10
2	45	94	11	0	21	11	35
28	43	67	66	0	53	44	0
0	2	66	98	64	88	0	88
50	18	88	0	87	93	95	27

90	0	16
72	99	22
0	62	62
64	0	89

Why we need statistics

What ill is statistics trying to protect us from?

First Hand Randomness Example

- We will go over a brief demo.
- Central Limit Magic

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	**Day**	**Date**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
2	Wednesday	1/1/2014																						
3	Thursday	1/2/2014						1																
4	Friday	1/3/2014																						
5	Saturday	1/4/2014																						2
6	Sunday	1/5/2014																						
7	Monday	1/6/2014																						
8	Tuesday	1/7/2014														5								
9	Wednesday	1/8/2014																						
10	Thursday	1/9/2014																						
11	Friday	1/10/2014																						
12	Saturday	1/11/2014																						
13	Sunday	1/12/2014																						
14	Monday	1/13/2014																						
15	Tuesday	1/14/2014																						
16	Wednesday	1/15/2014																						
17	Thursday	1/16/2014																						

Common Points of Confusion

Confusion Roadmap

- Sample Coverage (A Drop in the Bucket)
- Snowflakes (Bundles of Uniqueness)
- Frame Distribution (Central Limit Magic)
- Randomness (How Random is Random)

Sample Coverage

How can a small sample

- capture the complexity in a frame,
- account for the potential subjectivity of an audit test,
- result in a fair estimate of overpayment?

Sample Coverage: A Minor Example

The red dot makes up roughly 1/30,000th of the of the above image. Not much. Samples this small are frequently used in auditing and in society as a whole. For example, election polls often sample 3,000 individuals from frames as big as 200 million.

Another Way to Frame The Issue

If you need to pull 100 transactions from a frame of 1,000 records, how could it possibly make sense to pull just 100 transactions from a frame of 1,000,000 records?

Sample Coverage : The Reality

- Coverage of the frame has virtually no impact on the quality of the sample. In fact, statistics works even if a frame is infinitely big!
- First an example and then the why.

Sample Coverage: What Matters?

- Frame A: 1,000,000 transactions that are 1 dollar each.
- Frame B: 10,000 transactions that are 100 dollars each.
- Given a 50 percent error rate, both frames would have total error amounts of 500,000 dollars.
- How do you think a sample of 100 will compare across the two frames?

Sample Coverage: What Matters?

- Frame A: Precision = +/- $83,434
- Frame B: Precision = +/- $83,020

- Frame A has 100 times more items then Frame B but the precision differs by only 0.6 percent!

- This example shows that the frame size doesn't matter, but why doesn't it matter?

Sample Coverage: Explanation

- Critically, we aren't trying to learn about all million items; we are just trying to learn about the average and the variance.

- Each sample item provides a similar amount of information about these two numbers regardless of the size of the frame.

Sample Coverage: Relative Precision

- The second key point is that we judge our success in estimating a number against the total error amount.
- Plus or minus 10 percent means something very different for a frame of 100,000,000 and a frame of 10,000.

Sample Coverage: Relative Example

- Frame A: 1,000,000 transactions that are 100 dollars each.
- Frame B: 10,000 transactions that are 100 dollars each.
- Given a 50 percent error rate, the larger frame would be 100 times larger than the smaller one.
- How do you think a sample of 100 will compare across the two frames?

Sample Coverage: Relative Example

- Frame A: Precision = +/- $8,343,362 = +/- 8.3%
- Frame B: Precision = +/- $83,020 = +/- 8.3%

- The precision for the larger frame is 100 times worse than the smaller frame, but as a fraction of the frame total it is virtually identical.

Snowflakes

Unique Little Snowflakes

Concern: What do we do when each item in the frame is a unique little snowflake that has nothing to do with the other frame items.

Uniqueness and Sampling

- While the uniqueness of the sample items may cause audit or legal concerns, it does not impact the validity of the statistical projection.

- The next example is crucial to understanding why statistical methods work in these cases.

Snowflakes: Unique Example A

Overpayment	Description
5	Claim 1
8	Claim 2
1	Claim 3
12	Claim 4
50	Claim 5
4	Claim 6
12	Claim 7
15	Claim 8
10	Claim 9

Snowflakes: Unique Example B

Value	Description
5	Number of fingers on your right hand
8	Number of Supreme Court Justices
1	The number of moons around our planet
12	Maximum rating on the Beaufort wind force scale
50	Fifth magic number in nuclear physics
4	Current House of Cards Season
79	Year Mt. Vesuvius Erupts and buries Pompeii
15	Common tax rate for long term capital gains
10	Number of teams in the Big 12

Snowflakes: Unique Example C

Value	Description
5	Overpayment associated in incorrectly billed dental claim
8	Capitation payment for in eligible beneficiary
1	Duplicate venipuncture
12	Supply code for an incorrectly billed immunosuppressive drug claim
50	Payment after death of beneficiary
4	Unsupported NEMT payment
79	Hospital inpatient claim should have been billed outpatient
15	Claim coded with an incorrect DRG
10	Insulin sent by supplier goes beyond beneficiaries prescribed need

Sample of the 1st, 4th, 8th claim

- Values are: 5, 12, 18.
- Mathematics of the statistical calculation are the same regardless of whether these items represent:
  - Claim 1, Claim 3, Claim 8;
  - Number of fingers on your right hand, Maximum rating on the Beaufort wind force scale, Common tax rate for long term capital gains; or
  - Overpayment associated in incorrectly billed dental claim, Supply code for an incorrectly billed immunosuppressive drug claim, Claim coded with an incorrect DRG.

Snowflakes: Uniqueness Doesn't Matter

- When calculating overpayments the reasons for the errors play absolutely no role in the validity or interpretation of the lower limit. All that matters is the mean and variance of the target amount.

- **Key Point**: The diversity of sample items only impacts the validity of the statistics in so far as it impacts the audit team's ability to reliably code the sample. If the samples can be coded accurately, then the extrapolation will be statistically valid.

Distribution of the Frame (aka, Central Limit Magic)

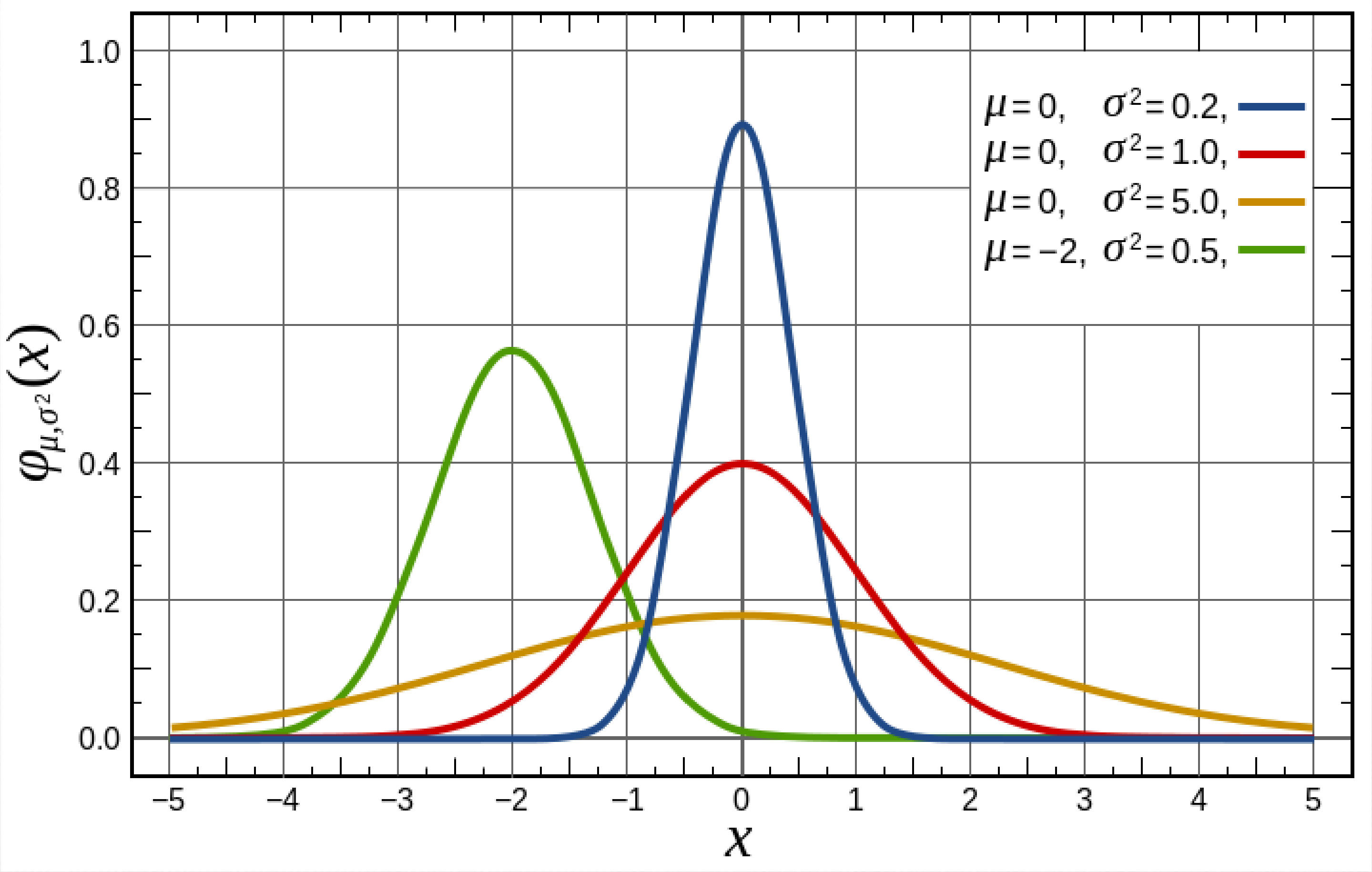
What happens when your data just isn't normal?

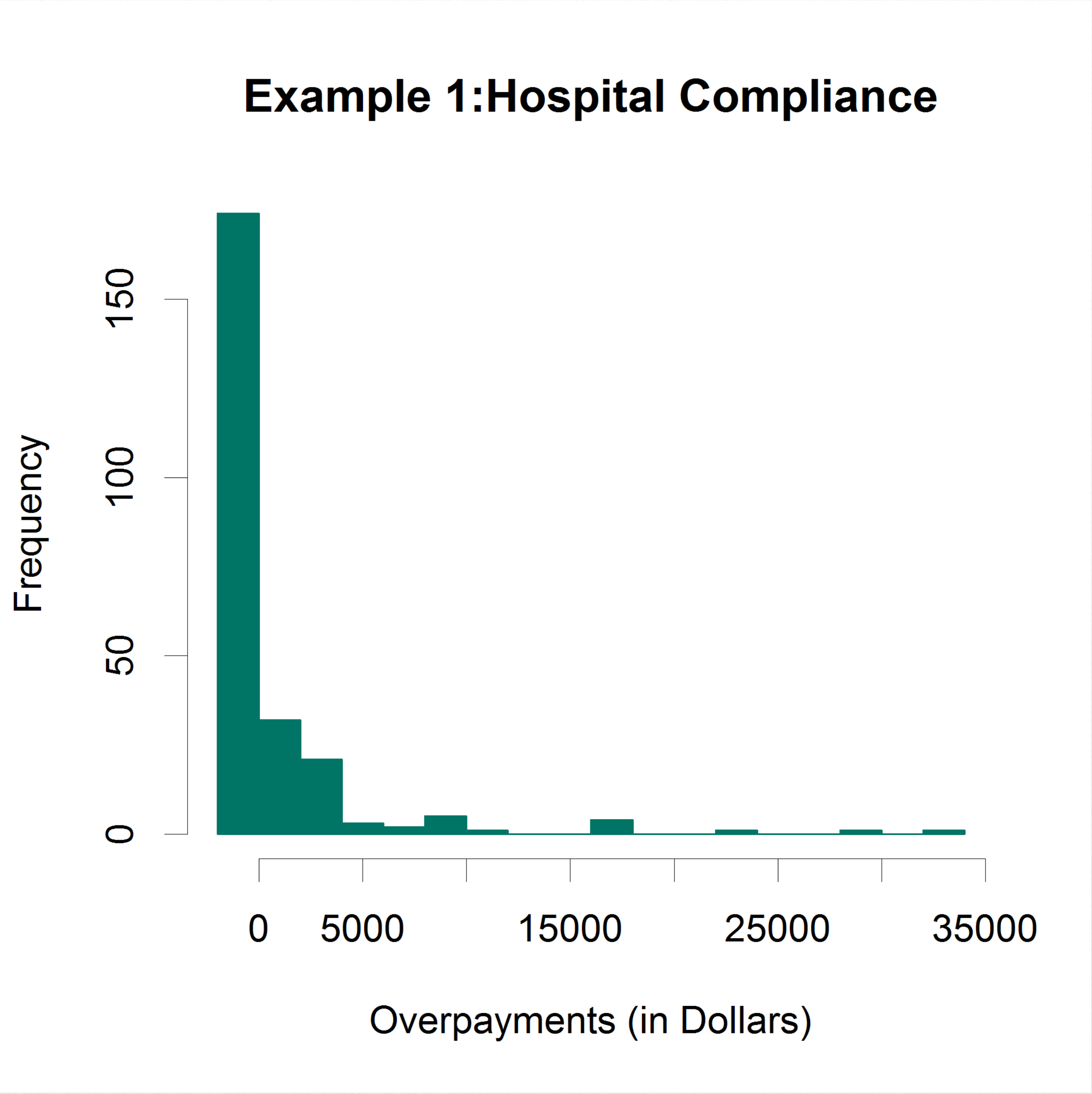
General Considerations

Question: Does the validity of our results depend on the amount of the various error within our sampling frame.

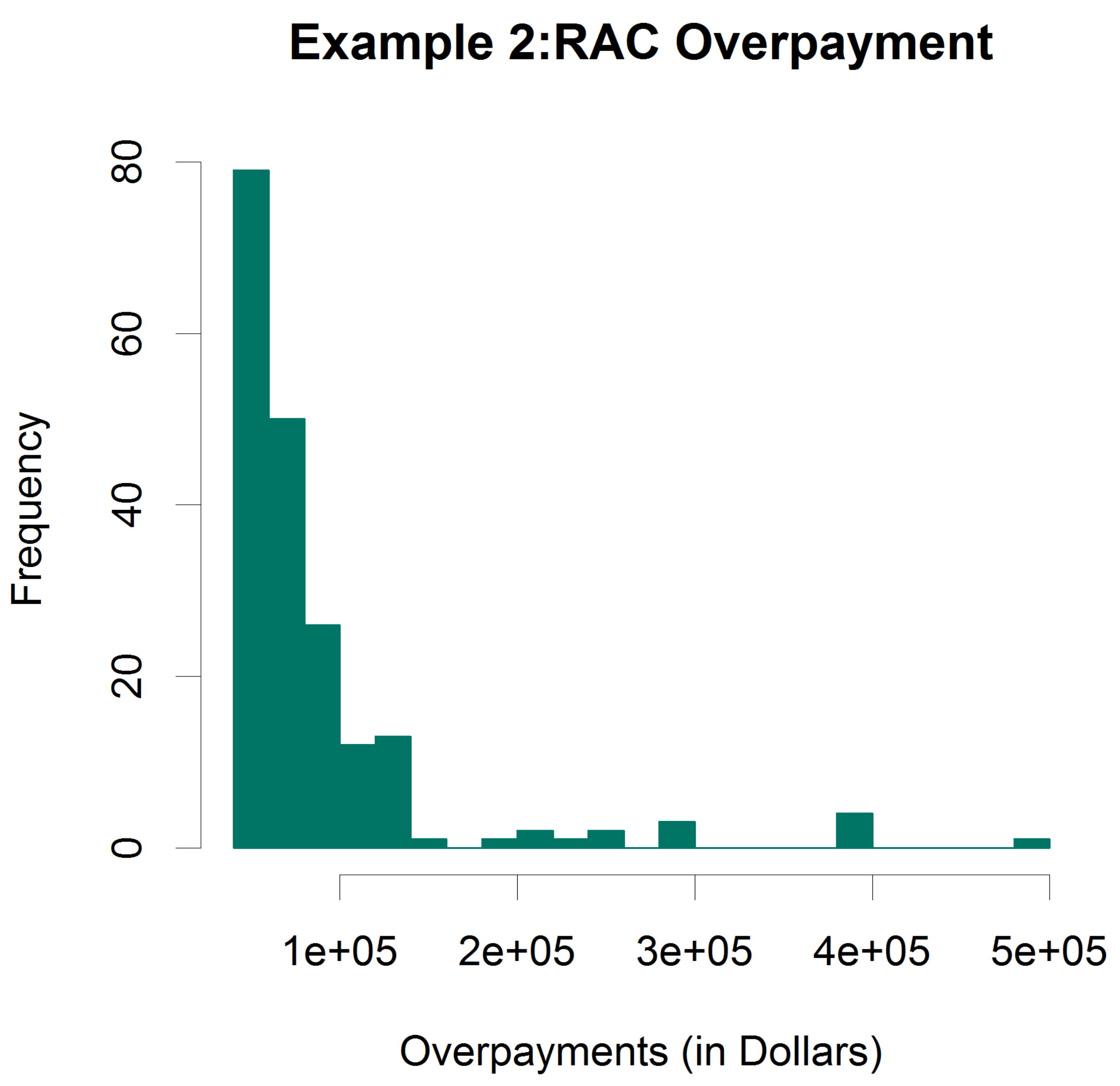
- E.g., what if the data is skewed or includes outliers.

Normal Distribution

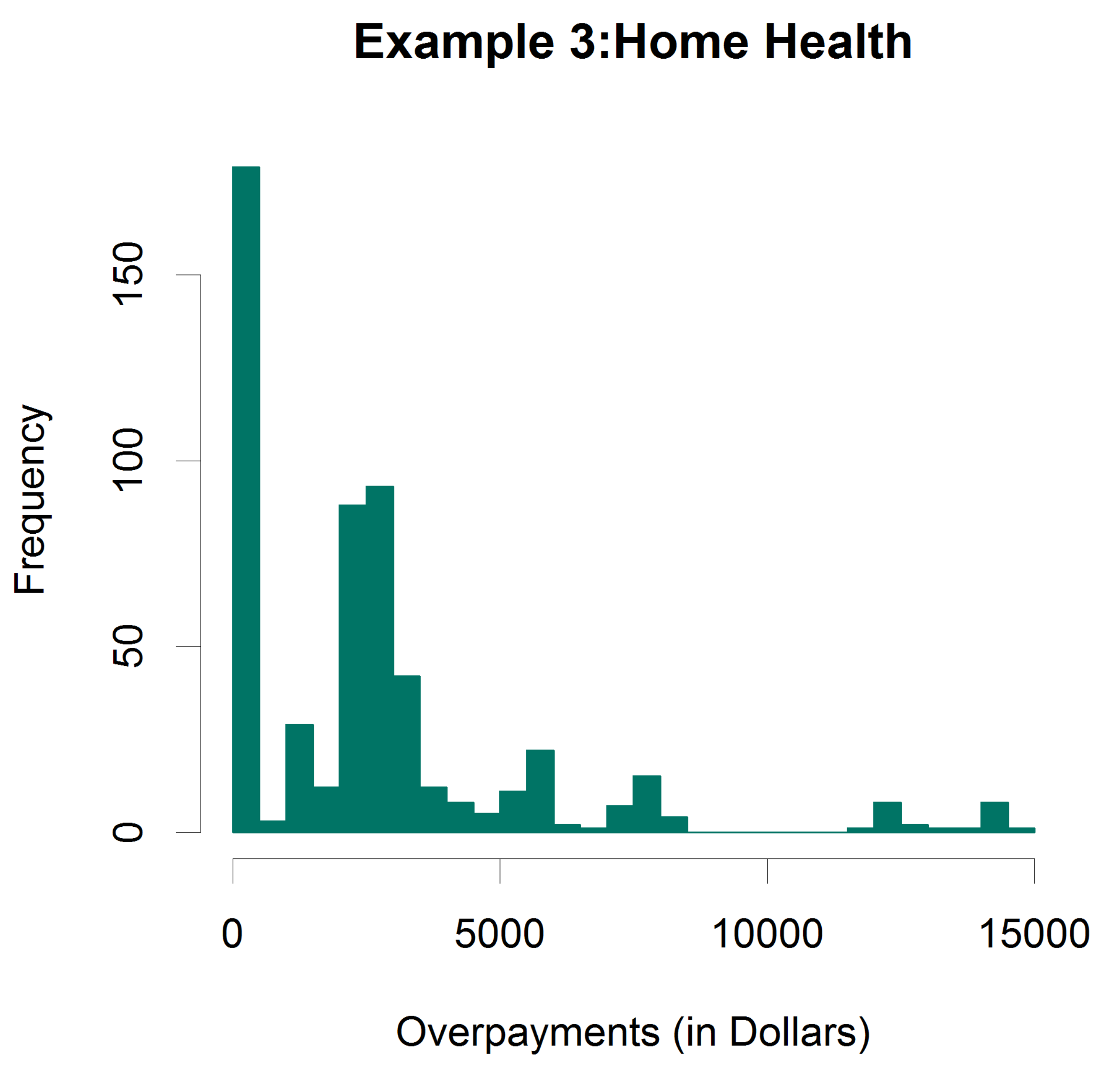




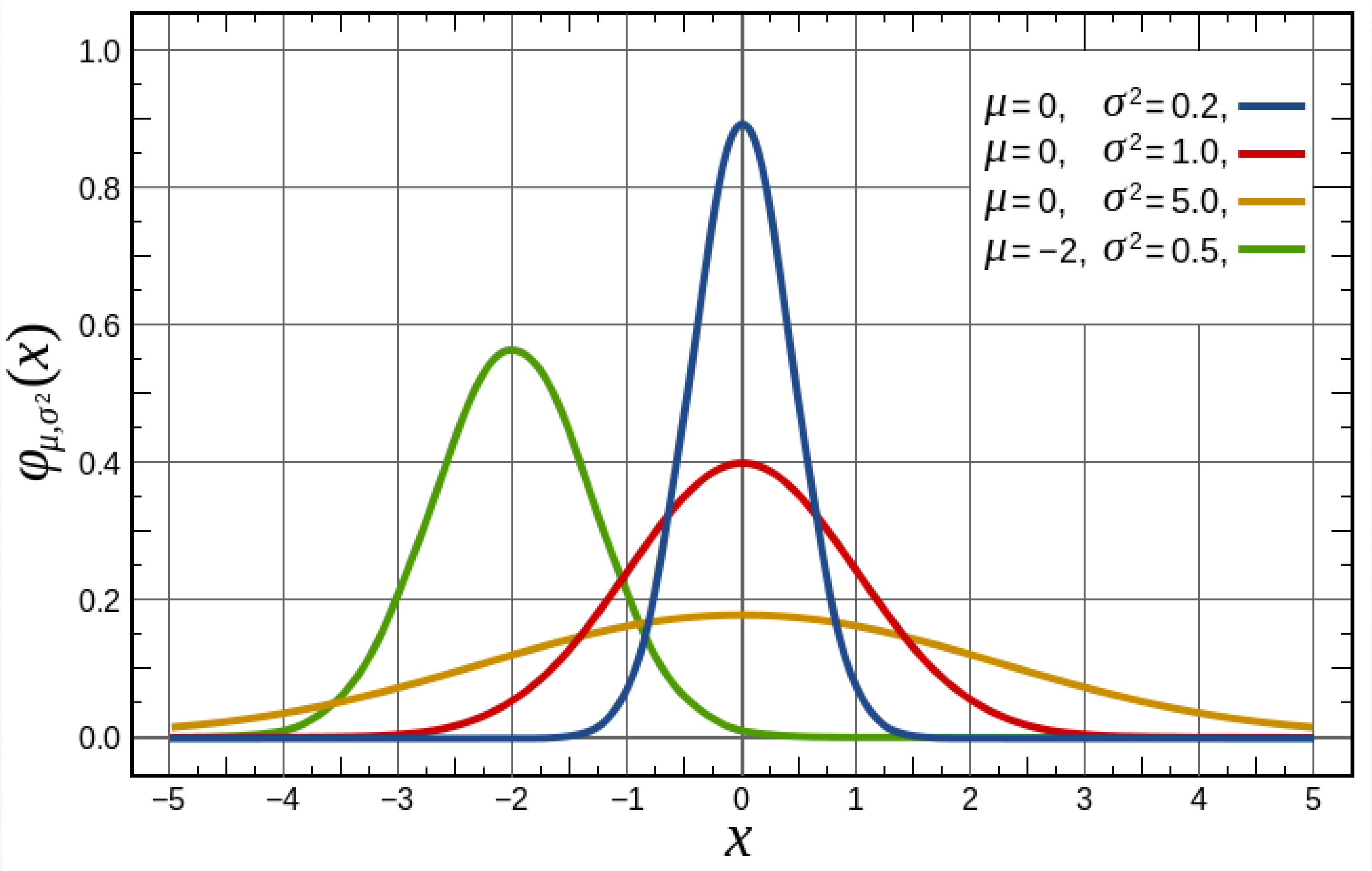
Example 1:Hospital Compliance
Frequency
0
50
100
150
0
5000
15000
25000
35000
Overpayments (in Dollars)



Example 2:RAC Overpayment
Frequency
0
20
40
60
80
1e+05
2e+05
3e+05
4e+05
5e+05
Overpayments (in Dollars)


Example 3:Home Health
Frequency
0
50
100
150
0
5000
10000
15000
Overpayments (in Dollars)

Recall We Want Things That Look Like This…

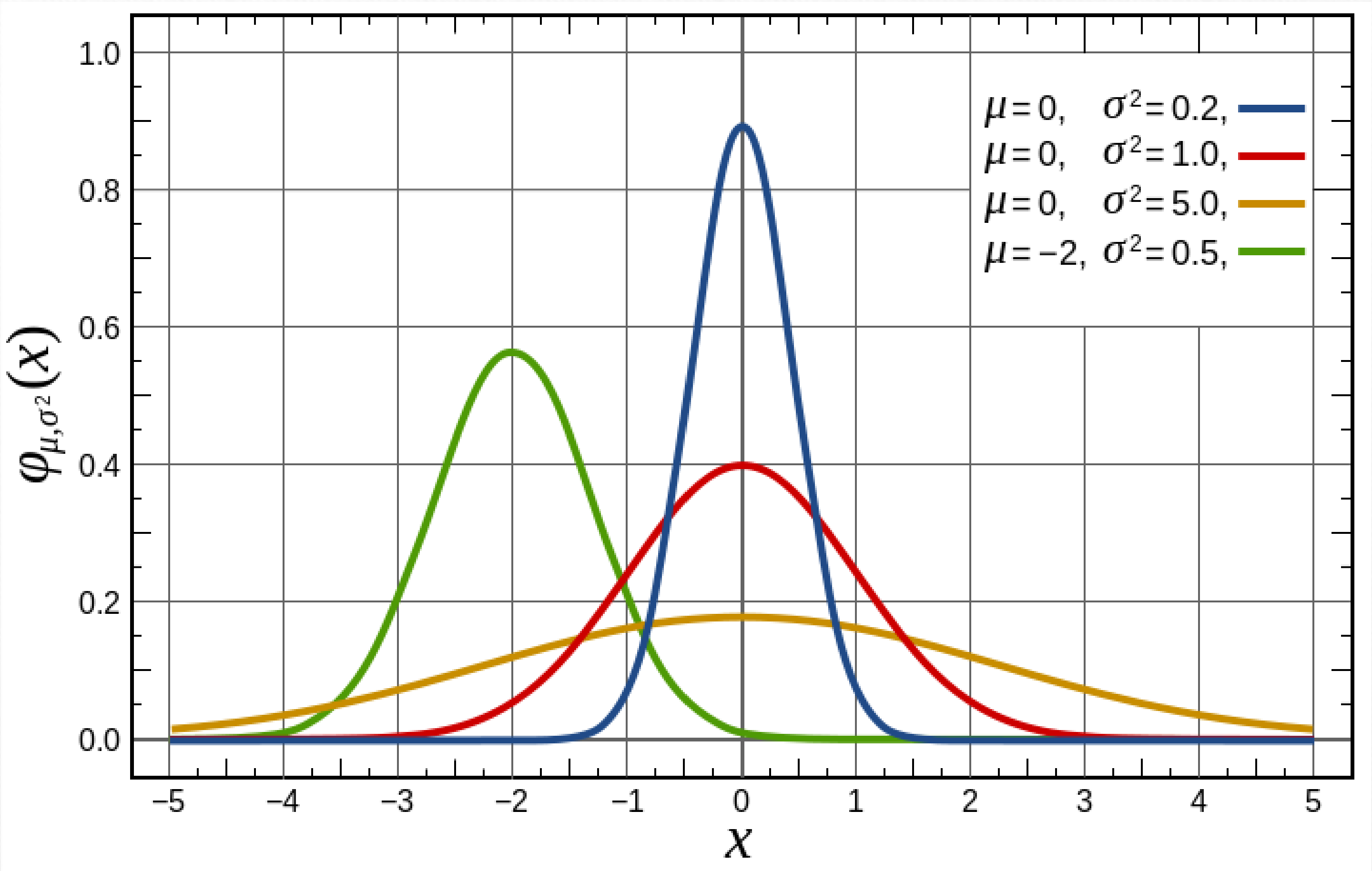


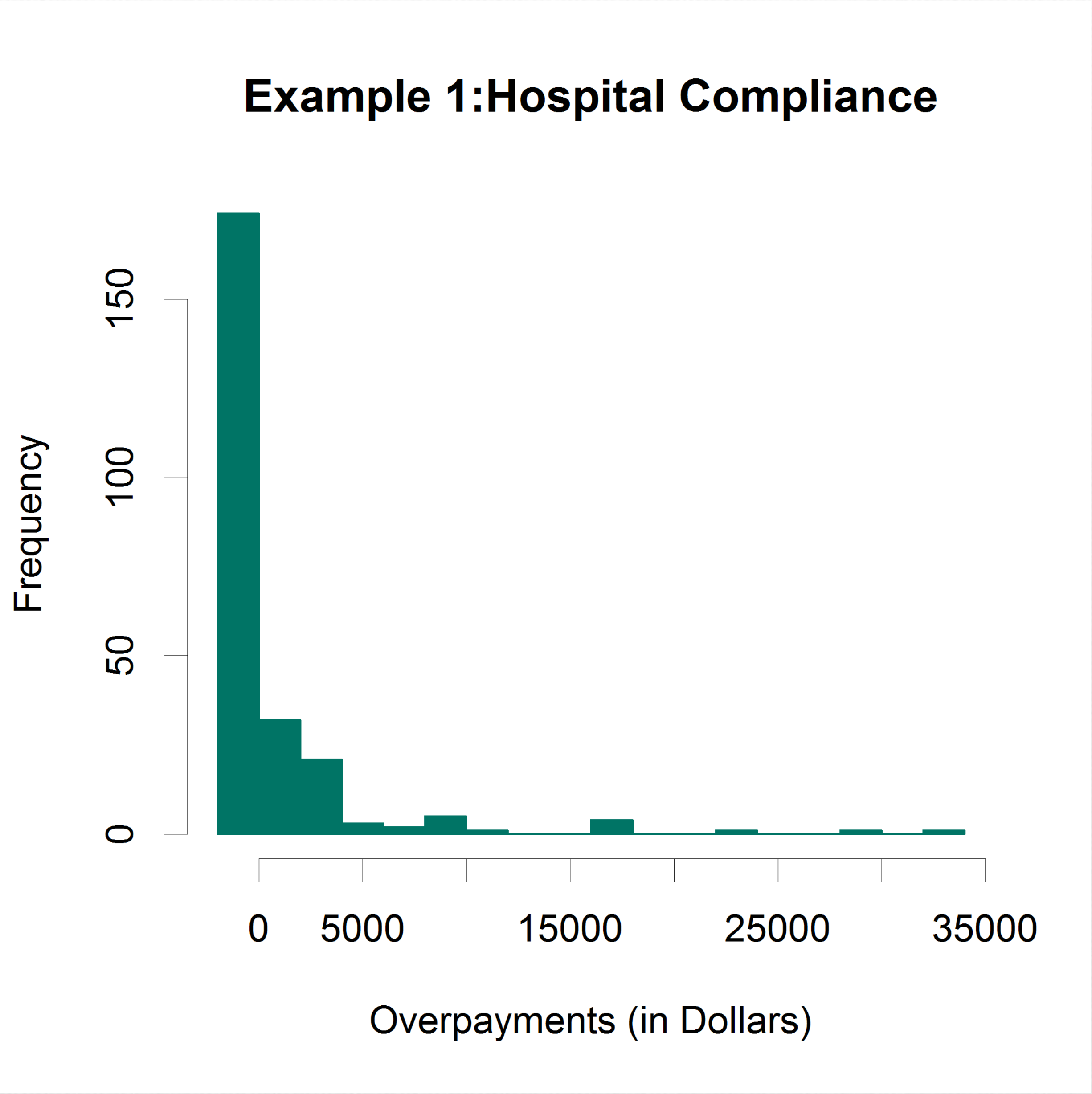
Central Limit Magic

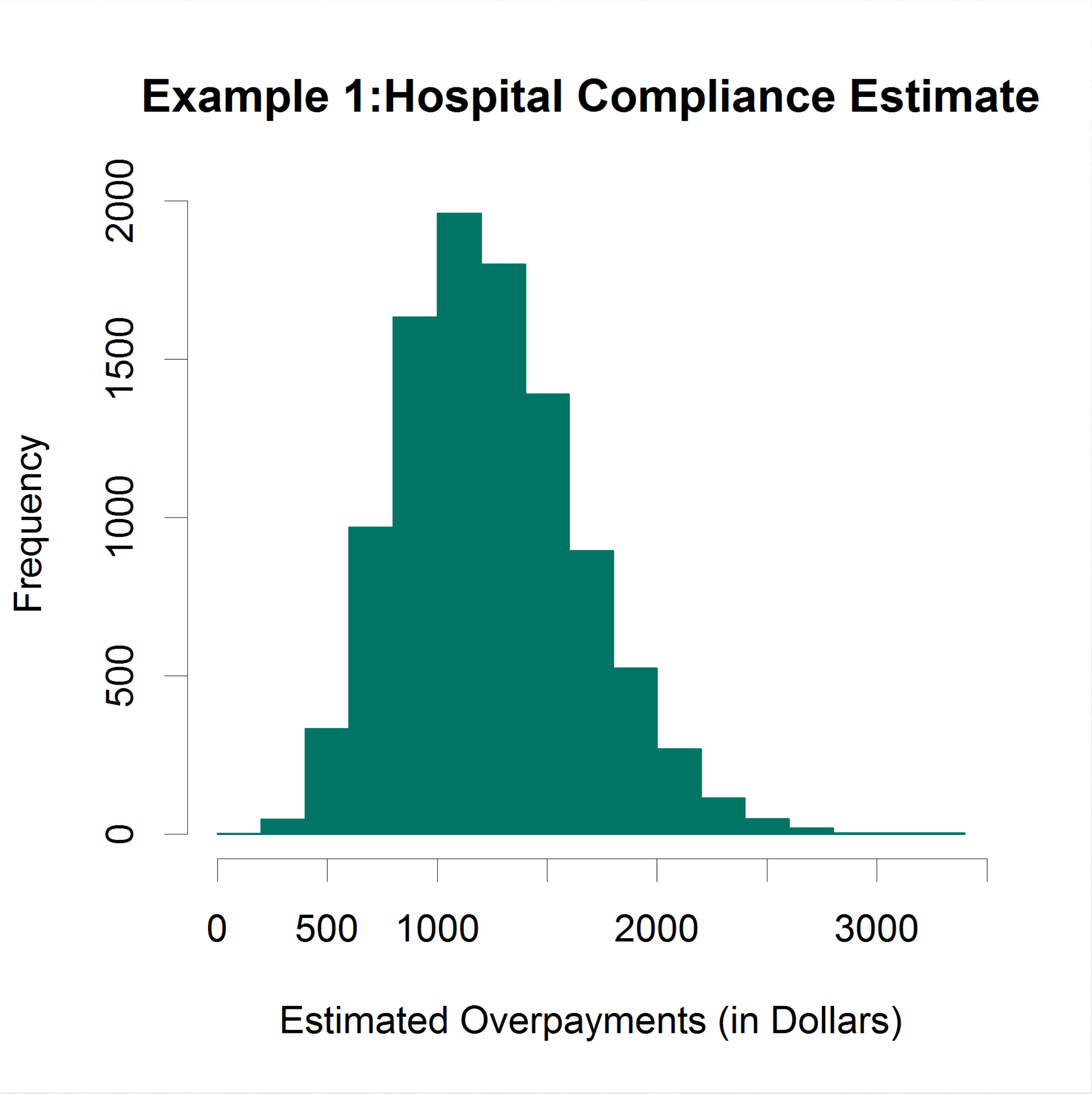
- Even though averages may not perfectly follow a normal distribution they are far closer than the original data.

- Lets return to the previous examples and see how the distribution of their estimates look across repeated audits (i.e., suppose we were to repeat the audit with different sets of random sample and then plot how much the results differ across audits).

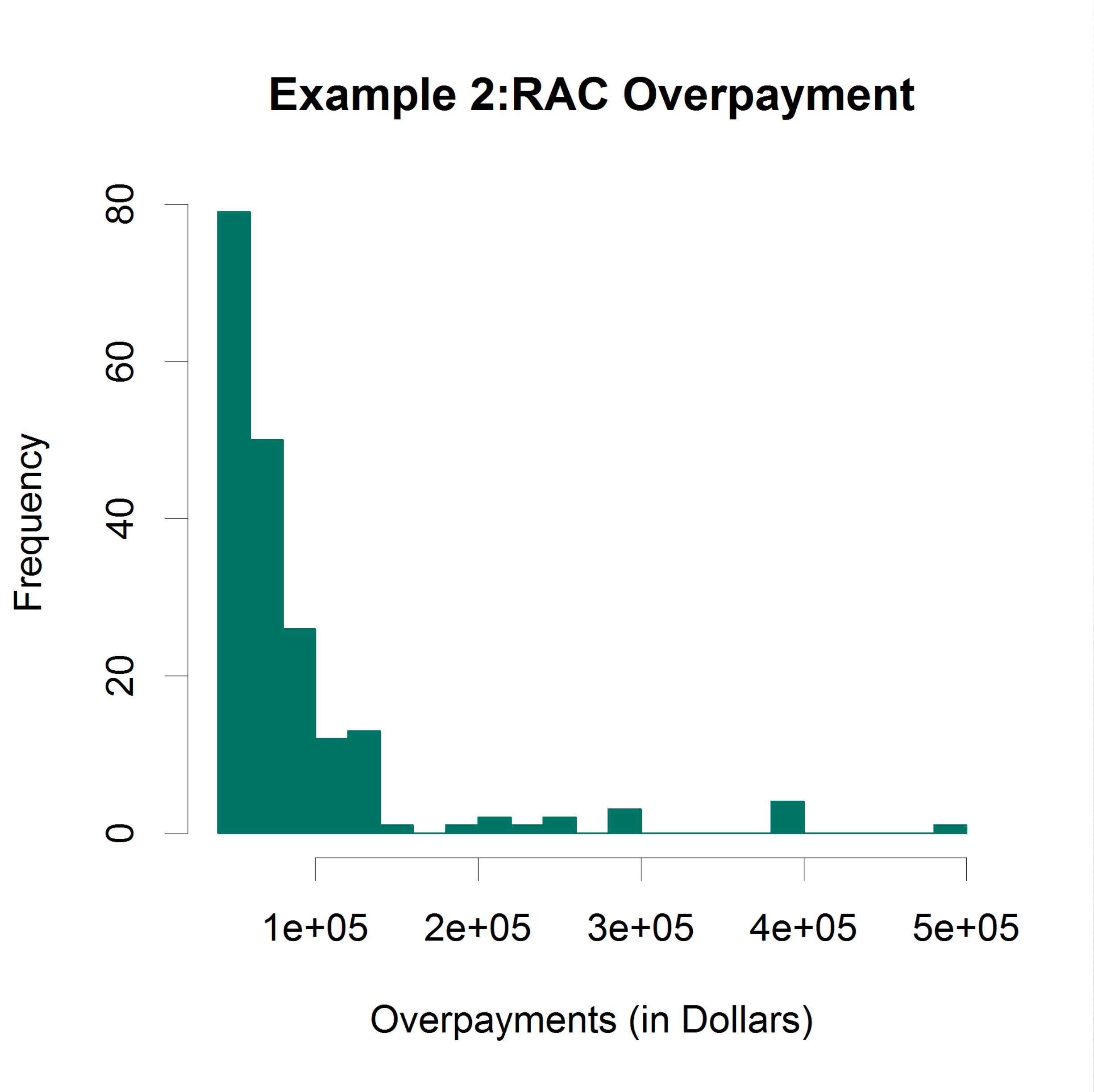
Recall We Want Things That Look Like This…

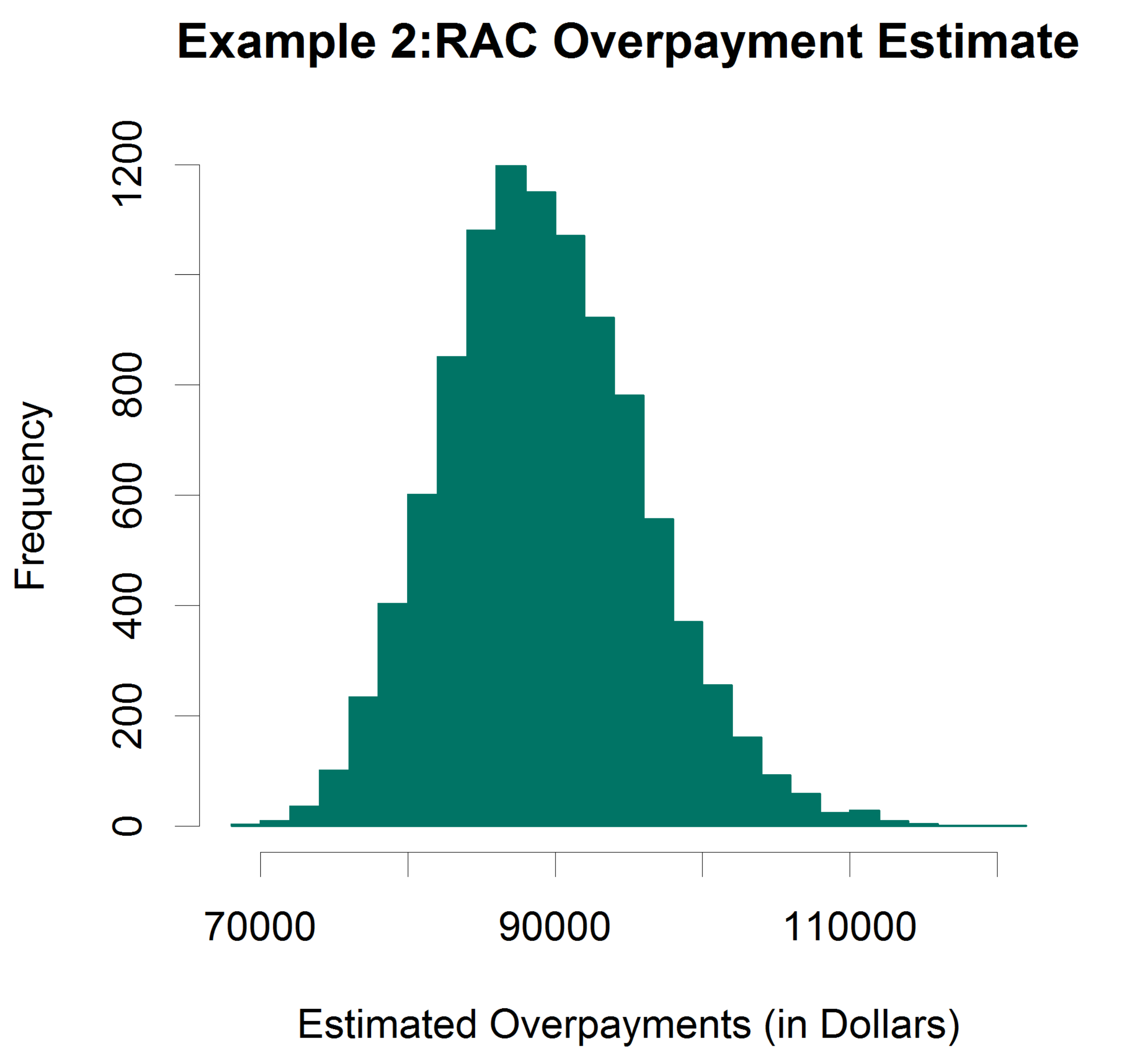


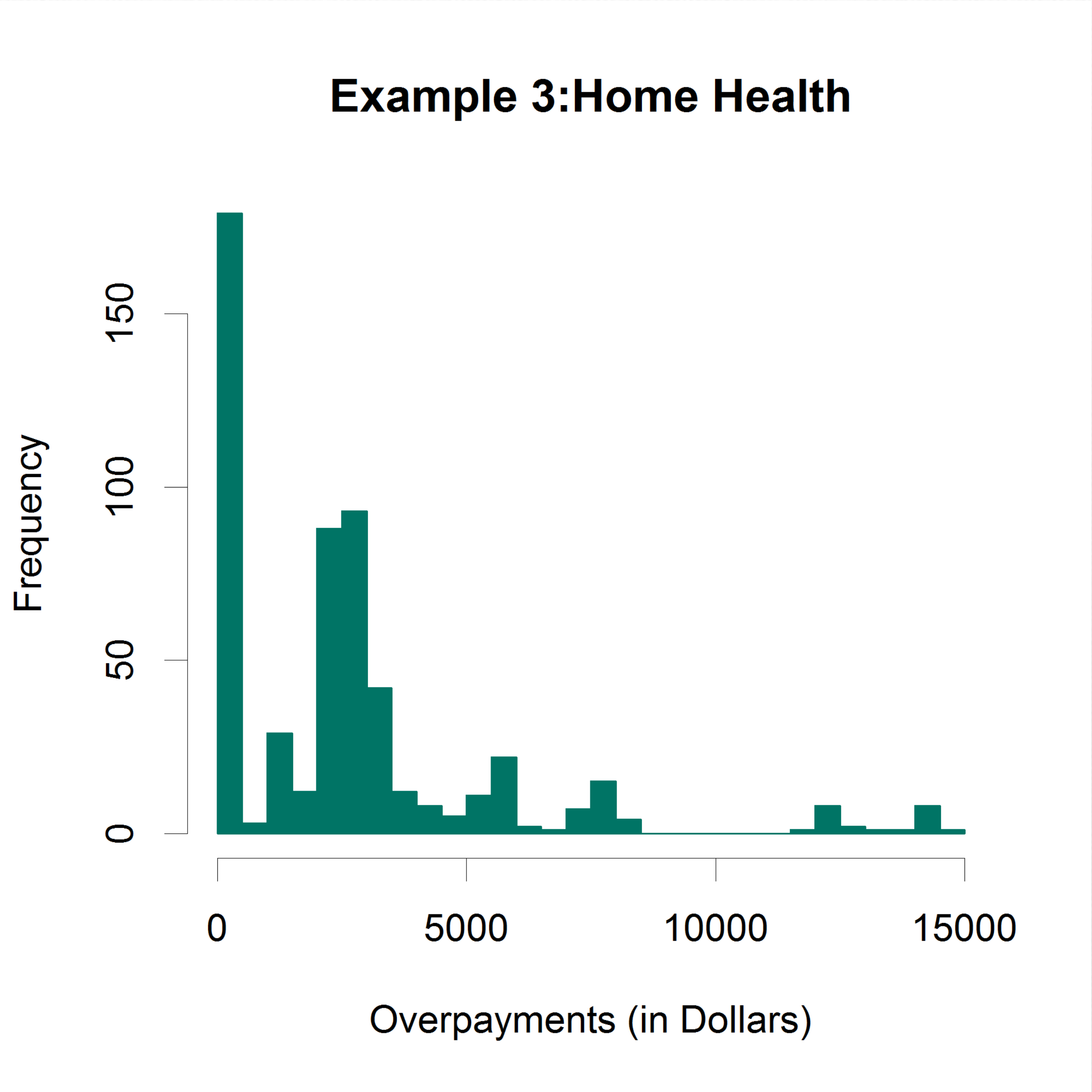

Example 1:Hospital Compliance
Frequency
0
50
100
150
0
5000
15000
25000
35000
Overpayments (in Dollars)

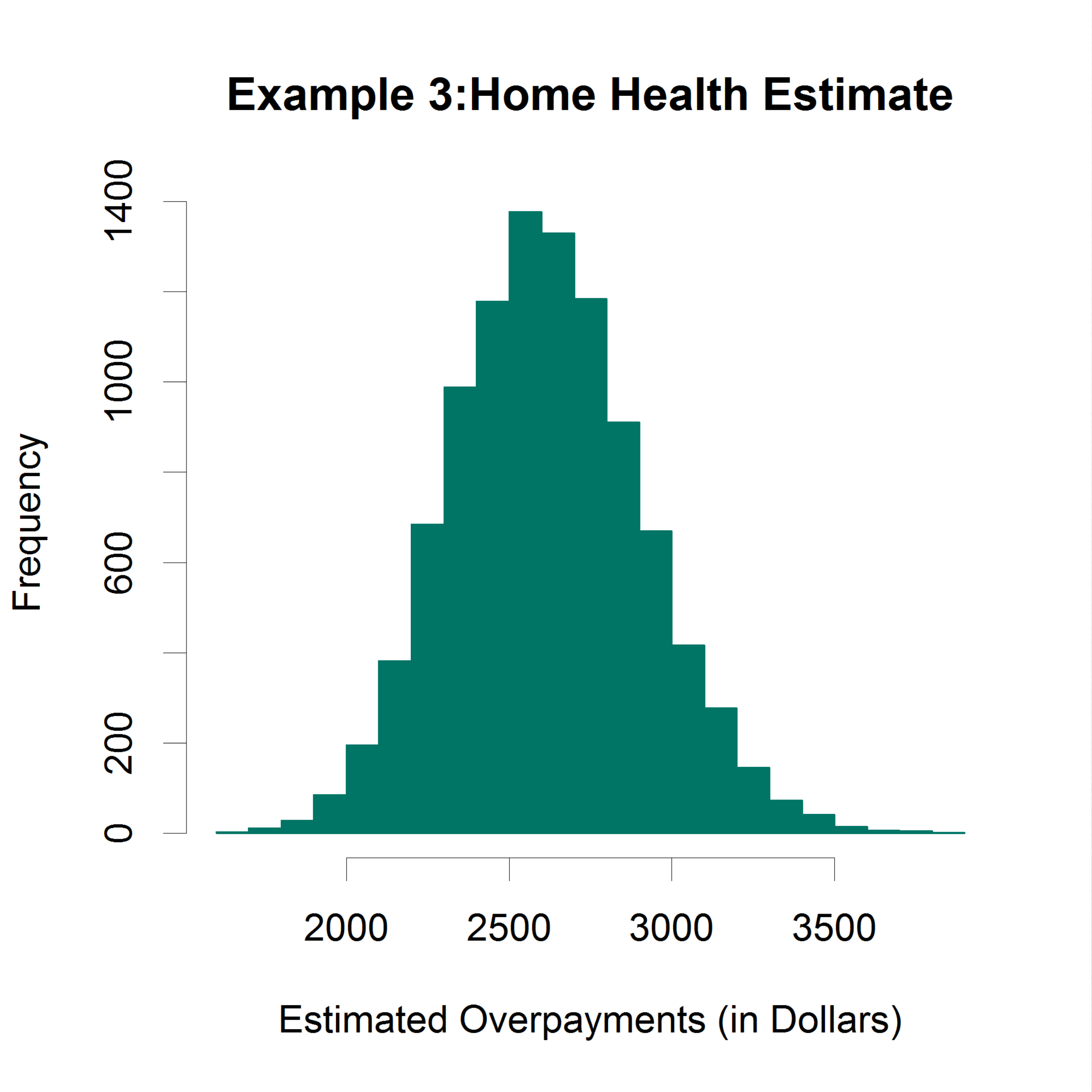


Example 1:Hospital Compliance Estimate
Frequency
0
500
1000
1500
2000
0
500
1000
2000
3000
Estimated Overpayments (in Dollars)


Example 2:RAC Overpayment
Frequency
0
20
40
60
80
1e+05
2e+05
3e+05
4e+05
5e+05
Overpayments (in Dollars)


Example 2:RAC Overpayment Estimate
Frequency
0
200
400
600
800
1200
70000
90000
110000
Estimated Overpayments (in Dollars)


Example 3:Home Health
Frequency
0
50
100
150
0
5000
10000
15000
Overpayments (in Dollars)

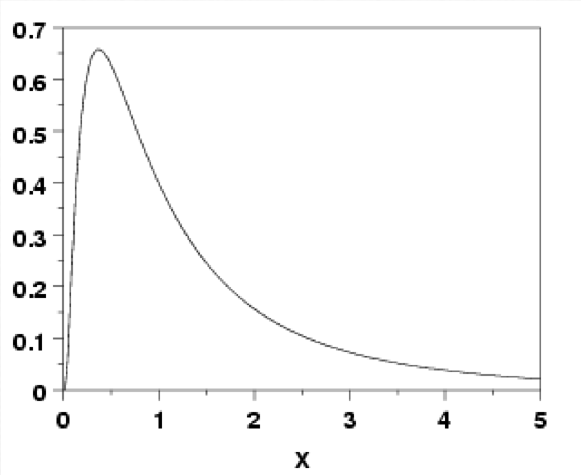

Example 3:Home Health Estimate
Frequency
0
200
600
1000
1400
2000
2500
3000
3500
Estimated Overpayments (in Dollars)

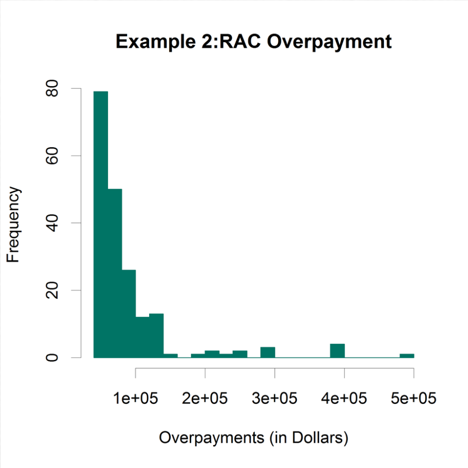
Magic Part 2

- The distribution of the estimates across repeated audits is much closer to normally distributed in all cases. The match, however, isn't perfect.
- What are the implications that the Central Limit Theorem doesn't hold exactly. In other words, what does it mean that one of the assumptions behind our analysis is, in some cases, violated?

Impact of CLT Limitations

- **The Kicker**: If data is positively skewed then the impact is that the lower limit is more conservative than intended. In other words the error works in favor of the auditee. No impact on point estimate





Impact of CLT Limitations

- There are more advanced methods that do not use the central limit theorem. We can see impact on the violation of the central limit theory by comparing estimates that require it with those that do not. report.

Method	90 Percent Lower Limit
Empirical Likelihood	$1,142,090
Standard Estimate (CLT)	**$820,096**
Bootstrap BCA	$1,185,852

Magic: Conclusion

- The fact that your data does not follow a nice bell curve is in itself not relevant to the validity of your estimate.
- Issues with the distribution of the data are often favorable to the auditee.
- You could have an issue if…
  - Your Data is **positively skewed** and you are relying on the **upper limit** or
  - your data is **negatively skewed** and you are relying on the **lower limit**.

Impact of the randomness of randomness

How do We Ensure Randomness?

Key Question: Even though the sample is pulled at random it may not necessarily appear so. For example, we may pull more high dollar transactions than expected.

Response

- A sample is random as long as it was drawn from a well validated pseudorandom number generator.
- For example, the random number generator in RAT-STATS passed a battery of tests showing is validity.
- Therefore, samples drawn from RAT-STATS are random.

An Extreme Situation

What if you pulled 10 sample numbers from a frame of a million and they were 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. How would that make you feel? Wouldn't such an anomaly imply that our sample results may not actually match up with the target population total. Could it be that sampling works regardless of which items are pulled?

Any ideas what statistical idea matches with the above intuition?

Limits of Our Knowledge

We expect lower limits calculated at the 90 percent confidence level to be less than the actual population amount 95 percent of the time. What about the other 5 percent? What if the provider has additional information about the sample indicating it might very well fall into that 5 percent category?

When the Unexpected Unexpectedly Occurs

Suppose a weather man has an extremely well calibrated model that says there is a 95 percent chance it will not rain today. The model is so good that on 95 percent of the days where it says won't rain, it doesn't rain. Now suppose you wake up, go outside (without your umbrella), and look up to see a large storm cloud approaching. What is the chance of rain? Is it still 5 percent?

Handling Randomness

- There are literally thousands of ways one could look for "anomalies" in the data. Given that the data is random, some anomalies will likely exist. The presence of such anomalies does not undermine the validity of the sample.
- The mathematics behind the calculation of the confidence interval assumes that all samples have the same chance of being selected. If we throw a sample out, post hoc, we will bias our results.

The Caveat

- There may be rare occasions where a given anomaly is so significant and so closely related to the target error rate that it may make sense to adjust the sample results. One sanity check is to estimate the transaction (not the error) amount in the frame.
- If your confidence interval fails seriously for the transaction amounts (which are known) then it may imply that an alternate analysis method may be more reasonable for the error totals.

Cost and Benefit Decisions in Sampling

Sample Size Determination – The Long Version

Overview

- **Defining Costs**
- **Introduction to Audit Types for Estimates**
- **Cost/Benefit: Rate Identification**
- **Cost/Benefit: Impact Identification**
- **Cost/Benefit: Dollar Recovery (Cost Savings)**

Introduction

Costs Eternal

Overview

Overall Goal: Discuss cost benefit considerations across a range of audit topics:

Section Organization: The course is organized based on audit type. For each audit type we will discuss the cost benefit considerations involved in making the final decision. First though we will talk about costs and identify the key audit types.

Costs

Costs versus Benefits: In many ways the cost side of the equation is more straightforward.

Key Question: How much time and money does it require to look at each additional sample.

Special Case: There is a special case when multiple locations are involved but we will address that only in passing here.

Opportunity Costs

Measuring time and money: Costs calculations must include adjustments for what else could have been done with the resources.

Key Question: What is the value to society of one hour of audit work?

- Could use Agency ROI as a multiplier
- Could attempt to quantify value of audits left on the table

Dodging the Problem

A Simpler Approach: Sometimes sufficient information is not available to perform a true cost-benefit analysis. In these cases sample sizes can sometimes be set through a minimum acceptable rather than an optimization approach.

Minimum Acceptable: Select the cheapest possible design that has a high chance of meeting the audit objective.

Overview

First Step in Cost Benefit Analysis: Identify exactly what you are trying to achieve with your sample!

Audit Types

Rate Identification: What is the error rate across the entire population of interest?

Impact Identification: Is there an issue that is sufficiently large to motivate a change?

Dollar recovery: Are there improper payments to be recovered and if so how much?

Keep in Mind…

These audit types are informal so you will not necessarily find them if you go look in a book.

Rate Identification

- Applies when primary purpose is to identify the actual magnitude of a phenomenon within an frame.
- Applies to any case when making comparisons across entities.
- Applies when follow up work will be performed in order to identify whether there has been an improvement in a system.

Rate Identification Examples

CERT error rate testing: Goal is to identify the actual improper error rate amount for all Part A and Part B claims.

Some Program Eligibility Audits: If part of the goal is to show differences between states or to inform Congress and other stakeholders about the magnitude of the comparative underlying issue.

Some Inventory Audits

Rate Identification Note

For rate identification your frame must be meaningful. Description of filters becomes absolutely critical!

Impact Identification

- Applies when error rate or amount is used only to show impact.
- Applies when the recommendations and findings are not sensitive to the error rate as long as the error rate is above a given level.
- **Note**: Does not apply to dollar recovery audits since the recommended recovery depends directly on the estimated error amount.

Impact Identification Examples

Foster care audits: If the goal of the audit is not to provide an overall error rate for the foster care system but rather identify weaknesses that need to be fixed.

Purchase card audits when there is no monetary recoveries

Dollar Recovery (or Cost Savings)

Applies when report is recommending a dollar recovery or cost savings amount.

Dollar Recovery Examples

Medicare and Medicaid overpayment audits.

Contract requirement audits.

Are Things Really Cut and Dry

A single audit may include multiple estimate types. However it is rare to have both Rate identification and Dollar recovery.

When audit includes Impact identification and another type, usually impact identification will be superseded by concerns associated with rate identification or dollar recovery.

Where are We?

We will now address the specific cost benefit considerations associated with each of the three audit types in the following order:

Rate Identification:

Impact Identification:

Dollar recovery:

Rate Identification

What you get for your sample: Increased accuracy in your point estimate. Example is for attribute estimate.

Sample Size	Plus or Minus Precision
10	28%
30	16%
100	9%
300	5%
1000	3%
7000	1%

Rate Identification

Key question 1: What represents an inherently material difference between error rates. A plus or minus 5 percent difference? A plus or minus 10 percent difference?

Key question 2: Who is the consumer of this data and what do they plan to use it for?

Key question 3: Are there any statutory requirements?

Key question 4: How different are areas within or outside the frame being compared?

Rate Identification

Steps:

1. For key questions, write down what precision you think would be necessary.
2. Calculate the required sample size for various precision levels.
3. Compare cost of sample size to nature of questions that can be answered.

Example 1: CERT

CERT Dollar Error Rate Testing – Sampling to estimate the overall error rate across all Medicare FFS providers

Example 1: CERT

What represents an inherently material difference between error rates? **No clear answer**

Who is the consumer of this data and what do they plan to use it for? **Agencies use data to identify sub areas to fix. Benchmarks involve changes of 0.5 to 1.5 percent**

Are there any statutory requirements? **Yes, 2.5 percent**

How different are areas that are being compared? **Key comparison is over time. In the past changes have ranged from .5 to 3 percent.**

Example 1: CERT

Benefit A: Agencies use data to identify sub areas to fix. Benchmarks involve changes of 0.5 to 1.5 percent

Cost A: Between 10k and 92k

Benefit B: Meet Statutory Requirements of 2.5 percent

Cost B: 3.7k

Benefit C: Accurately track changes over time

Cost C: 258k

Example 2: Polling

Election Polling

It is election season so as you know there are many polls reported every day (e.g. entail)

Example 2: Polling

What represents an inherently material difference between error rates? **No clear answer**

Who is the consumer of this data and what do they plan to use it for? **Public wants to identify the next president. Biggest concern is how everyone compares to the top individual.**

Are there any statutory requirements? **No, but industry standard is 3 percent for presidential election.**

How different are areas that are being compared? **Most candidates more than 5 percent from the top candidate. However dynamic process with candidates at times being in a dead heat.**

Example 2: Polling

Benefit A: See what individuals are generally in the lead (plus or minus 5 to 15 percent precision)

Cost A: Between 40 and 360 samples

Benefit B: Meet industry standard plus or minus 3 percent

Cost B: 1,000

Benefit C: Identify the individual leading in all but the closest races (plus or minus 0.5 percent)

Cost C: 36k

Example 3: Medicaid Eligibility

Audit to determine the percent of ineligible Medicaid beneficiaries along with the amount improperly paid to them. Audits performed across multiple states without any dollar recovery requested.

Example 3: Medicaid Eligibility

What represents an inherently material difference between error rates? **Grey area, but a precision 20 percent of the frame (2 billion) would be unacceptable. Even a precision of 10 percent (1 billion) could be problematic.**

Who is the consumer of this data and what do they plan to use it for? **Primarily the state agency; however, Congress and CMS could in theory be interested as well.**

Are there any statutory requirements? **No.**

How different are areas that are being compared? **Perhaps compare across states. States range from almost perfect to having dollar error rates of 50 percent.**

Example 3: Medicaid Eligibility

Benefit A: Avoid possible issues with the face validity of the confidence interval (plus or minus 10 to 20 percent).

Cost A: Between 288 and 1,152 sample items

Benefit B: Separate the best and worst states (plus or minus 30 percent).

Cost B: 128 sample items

Benefit C: Provide detailed resolution on differences across states (plus or minus 5 percent)

Cost C: 4.6k sample items

Rate Identification: Conclusion

Conclusions:

1. Rate identifications are some of the most expensive audits in regards to required sample size.
2. A sample of 100 will often only answer the most basic questions. Samples in the thousands may be necessary.
3. More of an art than a science. Need to take some leaps of faith about what precision will be necessary for what questions.

Where are We?

We just covered Rate Identification, now we are moving on to Impact Identification:

Rate Identification:

Impact Identification:

Dollar recovery:

Impact Identification

What do your sample units buy you?: Increased chance of identifying an impact of an issue if an issue exists.

Relative to Rate Identification: Impact Identification can be thought of as a simple case of Rate identification where the only question is whether your sample size is large enough to identify a problem.

Impact Identification: Analysis A

Steps:

1. Identify the tolerable error rate/amount.
2. Identify the expected error rate/amount.
3. Identify a sample size that can reliably detect the difference between the tolerable and expected error rate (roughly the precision must be less than the difference than the tolerable and expected error rate).

Analysis A Limitation…

Limitation: If I knew what the error rate was beforehand I wouldn't have to do the audit! How do I pick an expected error rate.

Sources of Information: Probe samples, previous audits, initial analysis of the control environment, findings in similar areas.

Insurance: Under this formulation, additional sample units buy you insurance in case your expectation is incorrect.

Impact Identification: Analysis B

Steps:

1. Identify the tolerable error rate/amount.
2. Identify a range of plausible sample sizes
3. Identify the minimum error rate that can be reliably detected given the plausible range of sample sizes.

Analysis B Limitation…

Limitation: What if the largest sample that you can reasonably complete cannot detect error rates/amounts that appear on the face material?

The Reality of Reality: The cost benefit decision in impact identification primarily focuses on reducing the risk of not detecting material errors.

Practice Note

The Real Reality of Reality: Auditors often end up applying an Analysis B approach due the difficulty in identifying an expected error rate.

What is Tolerable: For many audits the tolerable error rate is effectively zero, since the report will cover whatever issues are identified in the sample.

Example 1: VA Enrollment

Background: VA OIG receives multiple allegations that there is a huge backlog of unprocessed enrollments for veteran benefits. VA OIG is tasked with identify whether this is true.

Audit involves combination of analytic and statistical testing.

Target Estimate: Number of Veterans with locked records who have applied or received medical care.

Example 1: VA Enrollment

Tolerable Error Rate: The best context number to gauge the materiality of an error is likely the total number of individuals leaving the military each year (roughly 300k). Any substantial fraction of that amount would likely be considered intolerable (e.g. 150k). 150k would represent 3 percent of the frame.

Expected Error Rate: Likely to be very high given the number of whistleblowers. For the purpose of this example, we will consider expected error rates of 10, 20, and 50 percent. For context the frame contains 6 million records.

Example 1: VA Enrollment

Benefit A: Very good chance of identifying at least 150k errors if error rate is 20 percent or above

Cost A: 30 samples

Benefit B: Very good chance of identifying at least 150k errors if the error rate is 50 percent or above

Cost B: 15 samples

Benefit C: Very good chance of identifying at least 150k errors if the error rate is 10 percent or above

Cost C: 125 samples

Example 1: VA Enrollment

Actual Choice: 24 samples

Observed error rate was 34 percent. Lower limit was 774k which was clearly material.

Example 2: Medicaid Waivers

Background: Audit concerning whether state is properly granting waivers that allow disabled individuals to receive Medicaid benefits.

Target Estimate: Total dollar value of benefits paid to individuals who were inappropriately granted a waiver.

Example 2: Medicaid Waivers

Tolerable Error Rate: Unclear. Can set at zero if this is a process that should be error free or can attempt set higher if audit team believe there is some non-zero error that is acceptable.

Expected Error Rate: No idea!

Analysis B: If we have no idea of the error rate we can start at a plausible range of sample sizes and see what type of results they might give. Lets try 100, 150, and 300.

Example 2: Medicaid Waivers

Benefit A: Very likely detect error rates of 17 percent. Moderate chance of detecting rates of 5 percent.

Cost A: 100 samples

Benefit B: Very likely detect error rates of 11 percent. Moderate chance of detecting rates of 4 percent.

Cost B: 150 samples

Benefit C: Very likely detect error rates of 7 percent. Moderate chance of detecting rates of 2 percent.

Cost C: 300 samples

Impact Identification

Conclusions:

1. Sample size depends heavily on difference between Expected and Tolerable Error rate/amount
2. You can reduce risk by learning more about your frame and/or by increasing your sample size.
3. No sample size is perfect (not even a 100 percent review).

Compare Impact and Rate

Impact Identification: Increasing the sample size reduces the chance that you do not detect the impact of an actual error.

Rate Identification: Increasing the sample size increases your ability to answer key audit questions, increases the face validity in your design, and may help meet statutory requirements.

Dollar Recovery (Cost Savings)

What do your sample units buy you?: Money when using the lower limit or credibility if using the point estimate (big difference between the two!).

Relative to Impact Identification: When using the point estimate the issues is more all or nothing. Either you have the credibility to support your findings or you don't. When using the lower limit every sample will, on average, increase your recovery.

Average recovery is the same when using the point estimate, but increases when using the lower limit.

Dollar Recovery

Key Point: In health care auditors often recommend recovery based on the statistical lower limit. In these cases, smaller sample sizes mean leaving more money on the table.

Key Point: If your agency uses the lower limit, poor precision can undermine the credibility of your results. Effect of sample size is more subjective in this situation.

Dollar Recovery: Analysis A

Steps for a Value of One More Sample Analysis:

1. Design your sample to the best of your ability.
2. Review the value of adding an additional sample.
3. Check whether the value of adding an addition sample is worth the cost to select that sample.

Dollar Recovery: Example 2B

Background: Suppose we are going to perform a series of DME audits and we are trying to figure out what sample size to use. We want to keep the audits relatively consistent and we do not expect providers to be relatively similar in size.

See Example file.

Dollar Recovery: Comparisons

Versus Impact Identification: Notice that underlying any Dollar Recovery is an Impact Identification audit. Namely, if the lower limit is negative it may be difficult to support findings (regardless of the recovery).

Versus Rate Identification: Rate Identification audits focus on the point estimate. Dollar Recovery on the lower limit. Rate Identification audits tend to require much higher sample sizes. Can you figure out why?

Conclusions

The end is near!

What we covered

- The steps needed to pull a sample and analyze your results.
- Some common misconceptions about statistical sampling.
- A conceptual framework for thinking about sample size decisions.

Sampling Steps

- Step 0: Define Your Problem
- Step 1: Select a Sampling Frame
- Step 2: Select a Sample Design
- Step 3: Select a Sample Size
- Step 4: Generate Your Random Numbers
- Step 5: Select Your Sample

Analysis Steps

- Step 6: Review Your Sample
- Step 7: Analyze Your Results
- Step 8: Interpret Your Results
- Step 9: Reporting

Concepts Roadmap

- General Concepts
  - Overview of Sampling
  - Why we need statistics
- Common Misconceptions
  - Sample Coverage (A Drop in the Bucket)
  - Snowflakes (Bundles of Uniqueness)
  - Frame Distribution (Central Limit Magic)
  - Randomness (How Random is Random)

Sample Size Considerations

- Defining Costs
- Introduction to Audit Types for Estimates
- Cost/Benefit: Rate Identification
- Cost/Benefit: Impact Identification
- Cost/Benefit: Dollar Recovery (Cost Savings)

Questions?

Jared Smith, PhD

Jared.Smith@oig.hhs.gov