Part I - Basic Statistics

Course Director: Edward Winslow

Systems Analyst

Auditmetrics Inc.

Third Party Administrator/Actuary –Ambulatory Health Services, Financial Auditor

info@auditmetrics.com

Course Director

- Disciplines: statistics, epidemiology, health care finance and economics

- Epidemiologist: “HealthLink Wellness” program

  A CDC funded healthy aging program

  Web Site: www.NewEnglandSenior.com

- Statistical and analytical advisor

  Web Site: www.auditmetrics.com

Text and Materials

- Audimetrics - AI v6.3 Software Recommended.
- Textbook: Statistical Audit - AI

  Applying Artificial Intelligence techniques

- Slide Presentation
  - –Part I Basic Principles
  - –Part II the Statistical Audit

Course Disciplines

- Statistics – Mathematical tools to aid auditors to make decisions under conditions of uncertainty
- Random sampling procedures such as stratification and outlier assessment

Statistics as an Audit Tool

What - use inferential statistics to draw conclusions about populations based on samples of data.

In this course we will discuss:

- The Why – The fundamentals of statistics
- The How - to implement statistical tools
- Some additional more advanced topics

Audit Sampling

American Institute of Certified Public Accountants (AICPA).

- Statistical Auditing Standards (SAS) No. 39, the essential features of statistical sampling are:
  - The sample items should have a known probability of selection, for example random selection
  - The results should be evaluated mathematically - that is in accordance with probability theory

Both features must be present to be considered a statistical audit

Uncertainty??

- We want to determine the amount of error in estimating a book of accounts but cannot examine all transactions.
- Auditor must determine the degree of reliability (assurance) required about the recorded financials.
- Must be able to mathematically determine the extent of testing (sample size) that is necessary to achieve desired reliability.

Uncertainty can lead to errors?

- Sampling Error- the difference between the sample and population values is considered a sampling error, sampling error can be estimated by probabilistic modeling of the sample.

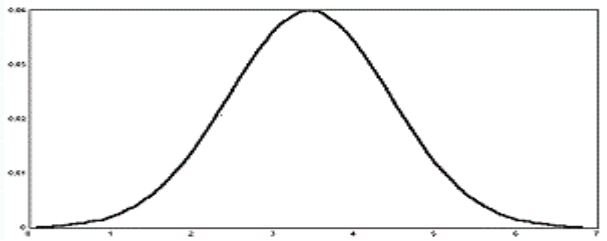
- Non-sampling Error- due to failure in human judgment:
  - Failure to properly define an audit population
  - Not properly defining evaluation criteria

Parameter, Statistic and Random Samples

- A **parameter** is a number that describes the population. It is a fixed number, but in practice we do not know its value. In a book of dollar transactions from a computer file the error rate is an unknown parameter.
- A **statistic** is a function of the sample data, i.e., it is a quantity whose value can be calculated from the sample data. It is a random variable used to make inference about unknown population parameters.
- The random variables $X_1, X_2, \ldots, X_n$ are said to form a (simple) **random sample** if each Xi has the same probability of being selected.

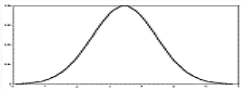
Transaction($): **X**

Population:



→

Sample:



Parameters:

Statistics:

Average		μ	Average		$\overline{x}$
Dispersion		σ	Dispersion		s

The Steps in the Statistical Audit Process

- Define population in terms relevant $x.
- Properly sample the population of $x.
- From that sample determine those values of x where an error was made or not properly handled.
- From the sample draw conclusions about the population

Proceeding in an environment where information is incomplete, can we take into account uncertainty?

- We must have quantifiable information
- Probability of outcomes

  --Likelihood of an event

- Decision Making

  --Choices that affect possible outcomes

  --What are probabilistic impacts of choices

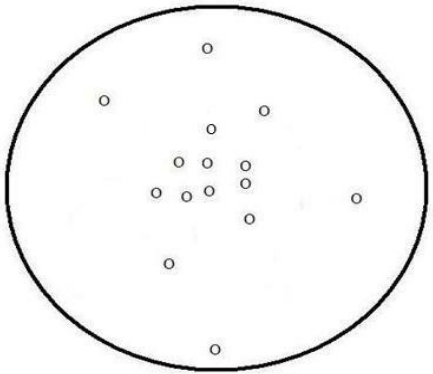
Discovery of the Random Process

- Sampling Errors are due to random chance.
- We can express that chance as a mathematical probability.
- The only way to quantify that probability is to have a mathematical definition of the random process.

Star Position
Multiple Observations by
Carl Friedrich Gauss*

***1777-1855: Number theory, Statistics, Physics, Differential Geometry, Astronomy, Optics.**

"In trying to plot the position of a star he found *random scatter due to varying atmospheric conditions affected his results"*

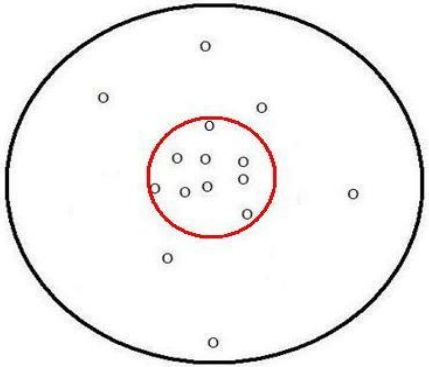


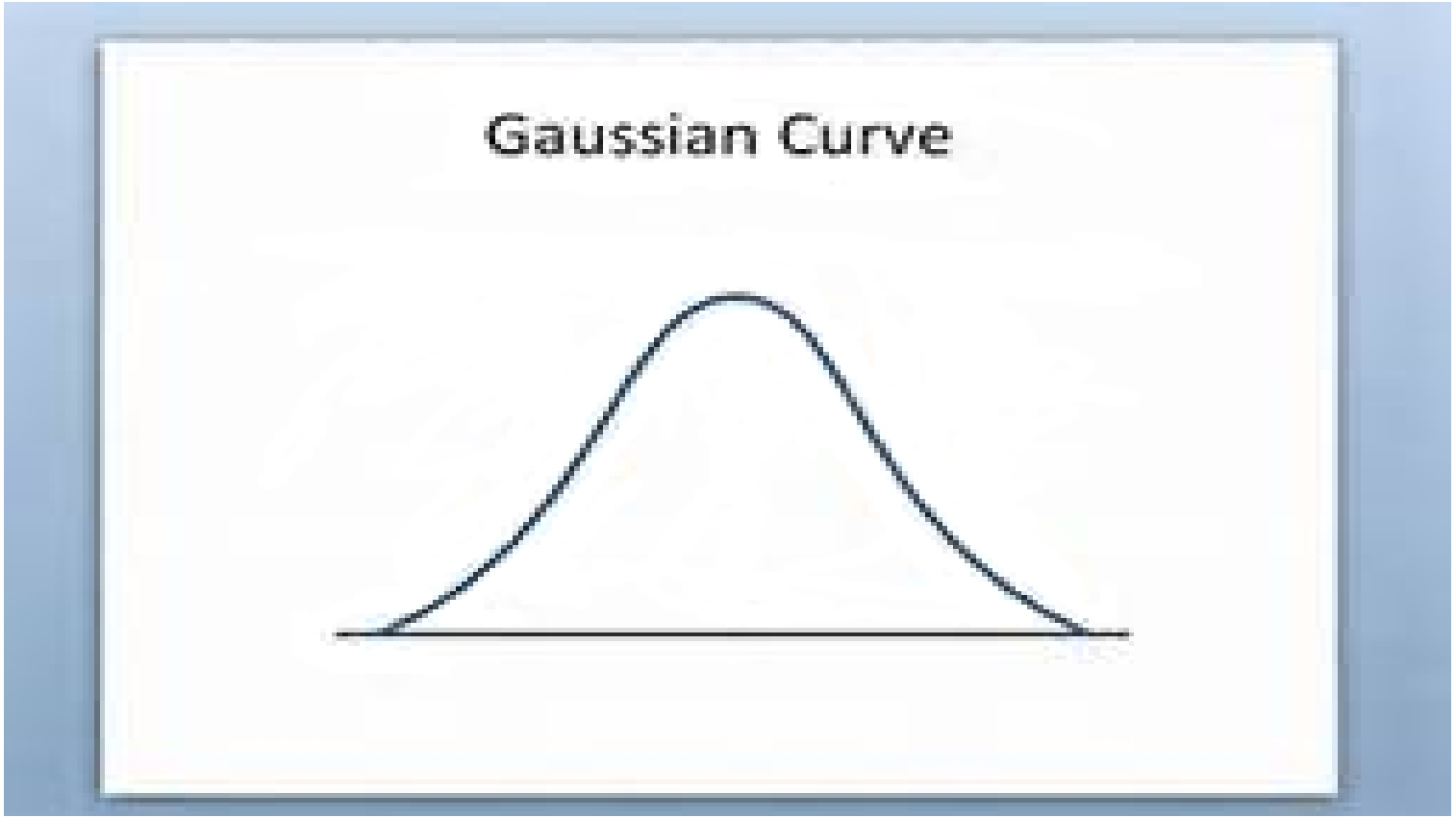
95% Probability of True Star Position

Gauss developed the mathematical formulation to set range where one could be 95% confident of the

True Star Position

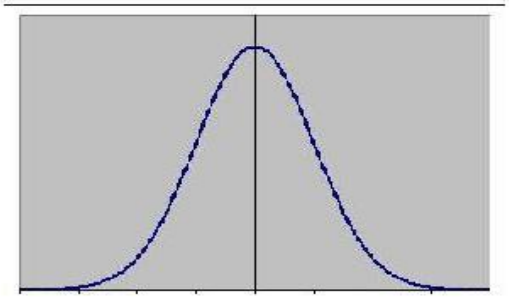
With enough observations there seem to be convergence on the true position




Gaussian Curve

Mathematical Properties of Randomness

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



$$e = \frac{1}{1!} + \frac{2}{2!} + \frac{3}{3!} + \cdots\cdots \infty = 2.71828\ldots\ldots\ldots\ldots$$

π Circumference/Diameter = 3.14159.....

The Benefit of Quantifying Randomness

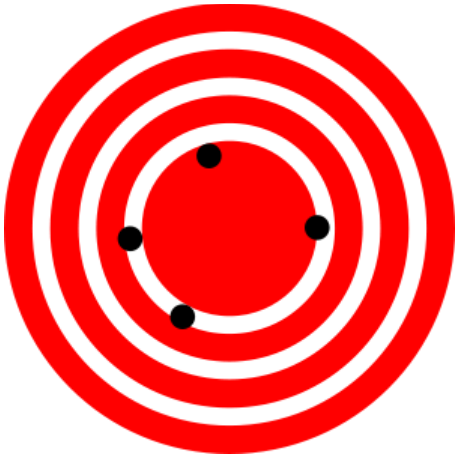
- Gauss had the random process imposed by nature on his observations.

- By mathematically describing that process he was able to correctly project the orbit of dwarf planet Ceres when many others failed.

- The auditor using statistical audit techniques is imposing a random process with the same goal in mind: “to make valid & precise projections”

High Validity – Low Precision

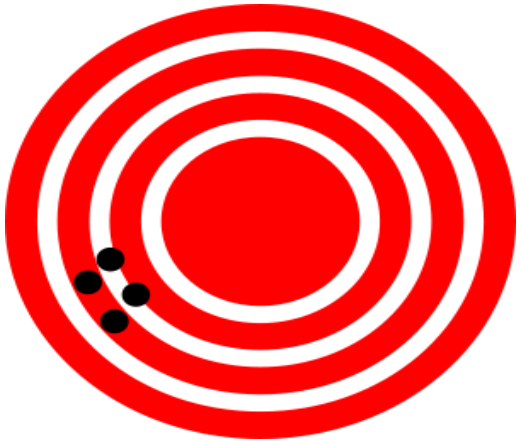
- Below is high sampling error (lack of precision) but has good validity.
- Can be solved with strategies to improve (reduce) sampling error such as increasing sample size.



High Precision – Low Validity

- Below is high precision but has poor validity due to improper sampling or non-sampling errors.

- Can be solved by proper random sampling & accurately defining audit population and clear evaluation criteria.



Imposition of Random Process by Sampling

1. Simple to use

A main random sampling advantage is that it is very easy to assemble the sample. Many expert analysts also consider random sampling like a fair method of taking samples from a certain population because each member is provided equal chances of being chosen. Furthermore, it is cheap to manage, especially in the era of electronic data.

Imposition of Random Process by Sampling (cont.)

2. Represents the population

The other reason why most analysts prefer random sampling over other analytical methods is the fact it represents the whole population. The only factor that may compromise the representativeness of random sampling is luck. Such a compromise usually results in a sampling error and a good method of preventing this error is to make certain that random sampling is properly done.

Imposition of Random Process by Sampling (cont.)

3. Clear conclusions

Because random sampling offers an unbiased selection and it is highly representative, it enables auditors to draw clear conclusions from results derived from the audit. Keep in mind that the main objective of doing an audit is the capability of making conclusions concerning the whole population from results obtained from sampling. This means that the representativeness of samples obtained through using random sampling can be used for making generalizations concerning the specific population.

The building blocks - Data

- Generally the numerical facts that helps in drawing conclusions about the population of interest.
- Numerical facts that can be mathematically analyzed, mathematics being the language of science.
- From our analysis we can generate information as to account accuracy.

Scales of Measurement

- Qualitative
  - Nominal Scale
  - Ordinal Scale
- Quantitative Scale
  - Interval Scale
  - Ratio Scale

Data Types in the Statistical Audit

- Ratio Scale – dollar value of the book of accounts.
- Nominal* – During the audit the auditor is making a yes/no or often referred to as a dichotomous decision.
  - Is a transaction in error -> yes/no.
  - Can estimate a probability (p) of that error.
  - Can apply that probability to a dollar volume.

** Nominal Categorical designations are also attributes*

Attribute Vs. Variable Sampling

Attribute - a characteristic that can be evaluated with a discrete response:

good – bad; yes – no; in error – not in error

Best when projecting frequency counts

Variable - a characteristic that is continuous and a quantity that can be measured:

Weight – length – dollars

Best when projecting a quantity such as dollar volume

Sampling Frame

Stratifying the Universe

- Acct. Type (attribute sampling):
  - Acct. Rec.
  - Acct. Pay.
  - Cash etc.
- Dollar Volume (variable sampling*):
  - 0-$1,000
  - $1,001 to $10,000
  - > $10,000

***These categories are ranges in the sampling of a continuous quantity**

Describing Uncertainty

Statistics is the data we collect to help us understand the world around us, in this course the characteristics of a set of accounts.

Can we have perfect knowledge? – very rarely

Sometimes we want to use statistics from samples to describe the population e.g. expense and income statistics

Sometimes we want to infer a characteristic or value e.g. project the amount of error in a set of accounts.

Estimating the Population

- What is the amount of error in estimating the population?
- If we know the population characteristics called parameters then we can properly judge effectiveness.
- However, due to limited money, access etc. we must use a sample of the population as an estimate of the parameters of our population of interest.

Statistic vs. Parameter

- We use a sample or subset of the population to determine the characteristic of interest
- We collect statistics from the sample to estimate the parameters of the population.
- Statistic → Characteristic of sample
- Parameter → Characteristic of Population

Amount of error is 10% in sample (statistic) so we estimate same effect in whole population (parameter). However, we know there is going to be a certain amount of sampling error. ***What is needed a way to quantify those errors.***

Sample Characteristics

We must have a good "representative sample" of the population or what is called an unbiased sample.

The best method to assure an unbiased sample is to draw random samples.

Random sample- each individual in the population has an equal chance of being selected. In other words each selection is independent of any other selection.

Block Sample

1. Difficult to control sample size because blocks can vary considerably,

2. By not taking sample transactions over the entire audit period, block samples increase sampling risk. If the tax error ratio in the sample time period differs significantly from the time periods not sampled, the block sample will produce results that are not valid. Subject to systematic bias. Another potential source of uncertainty.

Estimation

We are using sample statistics (characteristics/measurements) to estimate population parameters

Uncertainty comes in when we are dealing with one random sample and use that to determine population characteristics.

We cannot say a specific random sample statistic is an exact replica of the population but is one observation that exhibits random scatter around the true population parameter.

Estimation (cont.)

- Just as Gauss's observations were subject to random scatter, auditors who rely on random samples are also observing a world where their estimations are subject to random scatter
- Given the mathematics of randomness (the normal curve) the auditor can provide a probability range for the true population value.

The Mathematics of Randomness

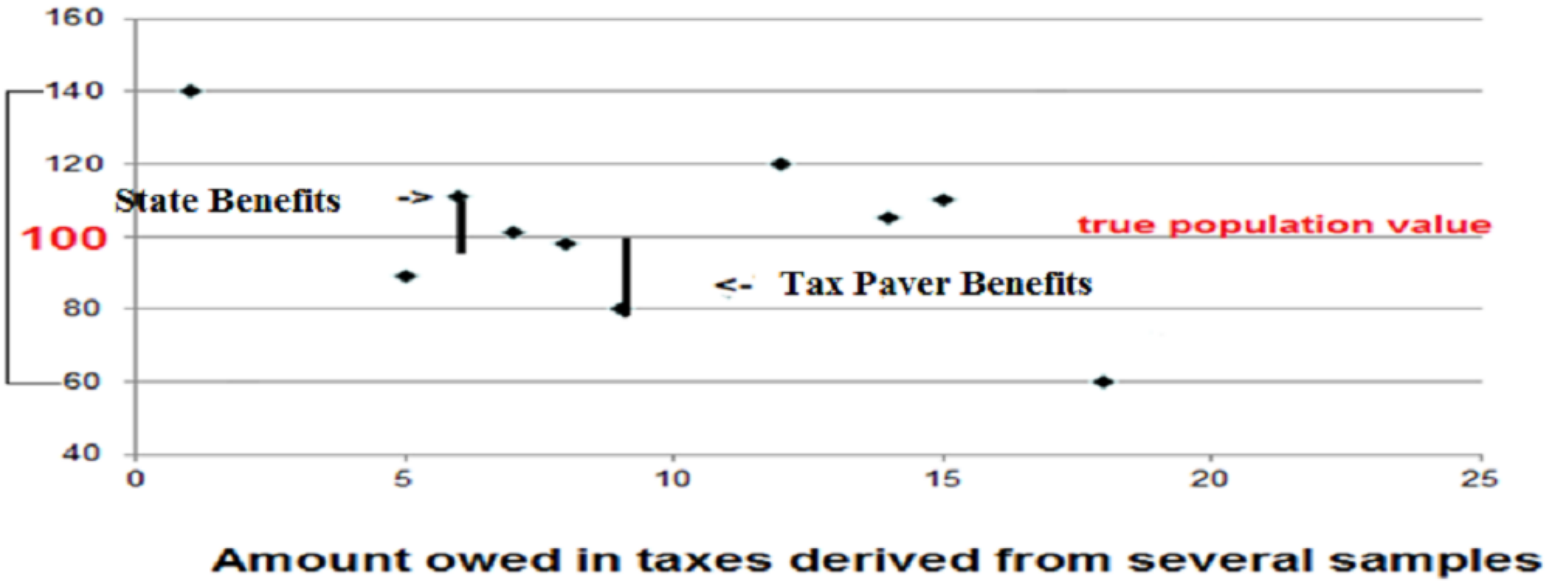
The mathematician/scientist who first described random behavior was Gauss who derived a mathematical formulation called the normal curve or bell shaped curve.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

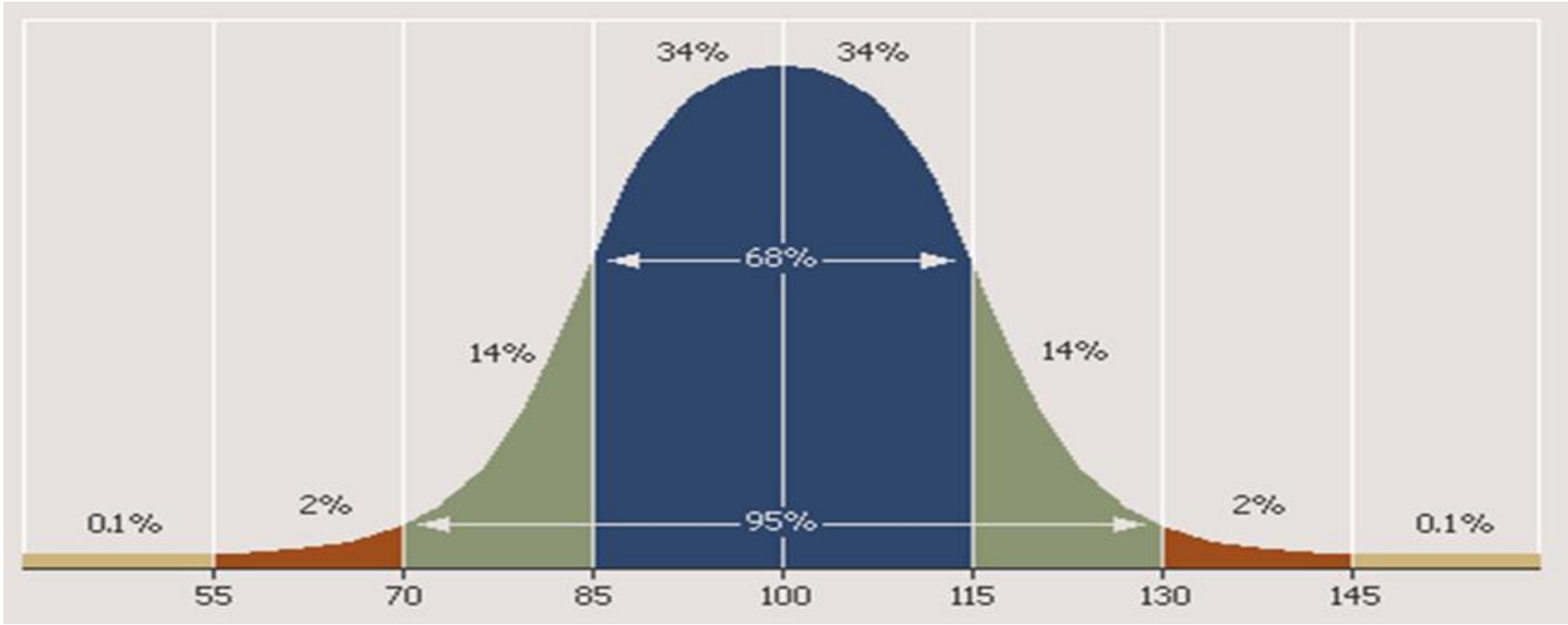
$$-\infty < x < \infty$$

The Normal Curve and Outcome Probability

- Based on the normal curve multiple audit random samples will tend to converge on the true population values.
- In this population 68% samples will be between 85 & 115
- The average of all of possible audit samples will equal true parameter of 100



Probability Ranges Around the True Value



What are those Key Parameters?

- In an efficient manner we need to summarize the characteristics of the population that imparts maximum information. For that we need:
  - a measure of the center of the distribution of interest. A specific single point of interest that describes the dollar magnitude of the typical transaction or a point estimate.
  - a measure of the spread of the distribution of interest. Spread is the uncertainty around that center point estimate.

Measures of Central Tendency

- Mean
  - The ‘average’ score—total score divided by the number of scores
  - has a number of useful statistical properties
  - many statistical inferences based on mean
  - Sensitive to ‘outliers’

    Extreme cases that just happened to end up in your sample by chance

Measures of Central Tendency (Cont.)

These measures also give us an idea what the 'typical' case in a distribution is like:

- Mode: the most frequent score in a distribution

  good for nominal data
- Median: the midpoint or mid-score in a distribution.

  (50% cases above/50% cases below)

  – insensitive to extreme cases

  --Interval or ratio

Average Salary

Annual Salary (x1000)	# Employees	Central Tendency Measures
450	**1**	
150	**1**	
102	**2**	
64	**1**	← Mean 64 is weighted average
52	**3**	
45	**4**	
40	**1**	← Median 12 above 12 below
30	**12**	← Mode most frequent
Total	**25**	

Measures of Dispersion

- Look at how widely scattered over the scale the values are.
- Groups with identical means can be more or less diverse.
- To find out how the group is distributed, we need to know how far or close individual members are to the mean.
- Like mean, only meaningful for interval or ratio-level measures

Measures of Dispersion

- Range

- Distance between the highest and lowest scores in a distribution;

    - sensitive to extreme scores;
    - compensate by calculating interquartile range (distance between the 25th and 75th percentile points) which represents the range of scores for the middle half of a distribution

- Usually used in combination with other measures of dispersion.

	Score on a test	Deviations
Percentiles	X	(X-mean)
P20	0	-50
P40	25	-25
P50	50	0
P80	75	25
P100	100	50
	======	======
Sum	250	0
mean	50	

Score on a test	|Deviations|	
X	|X-mean|	
----------	----------	
0	50	
25	25	
50	0	
75	25	Average Deviation
100	50	=150/5 =30
======	======	
250	150	
Mean=50		

Score on a test	Deviations	Deviations Squared			
X	(X-M)	$(X-M)^2$			
----------	----------	----------			
0	-50	2500	Population:		
25	-25	625	**Variance σ^2 =6250/5 = 1250**		
50	0	0	**Std. Dev. σ= 35.4**		
75	25	625			
100	50	2500	Sample:		
======	======	======	**Variance S^2 =6250/4 = 1562.5**		
250	0	6250	**Std. Dev. S = 39.5**		
50					

Measures of dispersion

Variance (S^2)

- Average of squared distances of individual points from the mean
- sample variance
- High variance means that most scores are far away from the mean. Low variance indicates that most scores cluster tightly about the mean.
- The amount that one score differs from the mean is called its deviation score (deviate)
- The sum of all deviation scores in a sample is called the SUM OF SQUARES

Standard Deviation (s , σ)

A summary statistic of how much scores vary from the mean

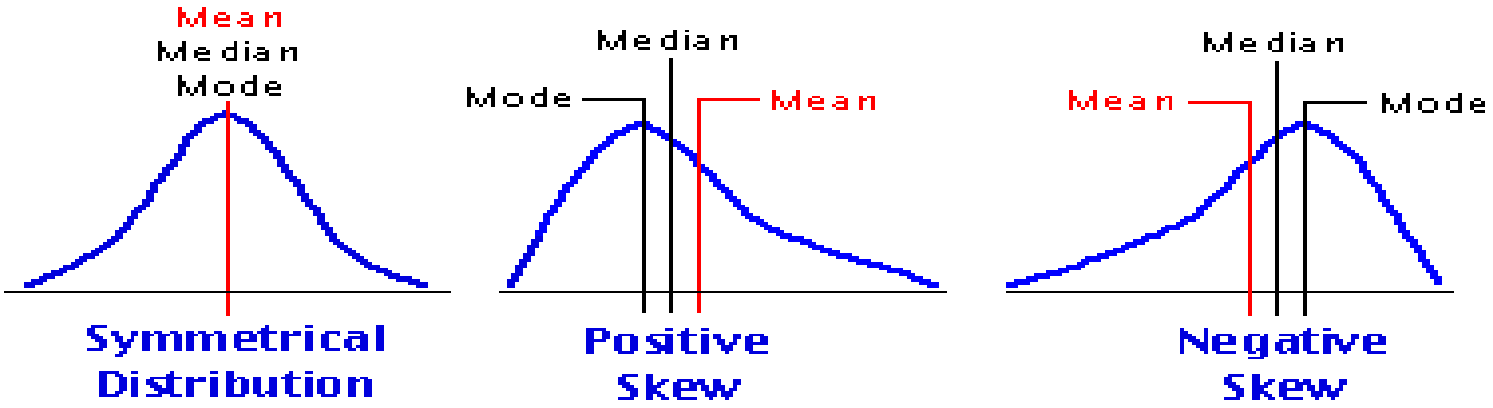
Square root of the Variance

- expressed in the original units of measurement.
- Average amount of dispersion around the mean or indication of uncertainty
- Used in a number of inferential statistics

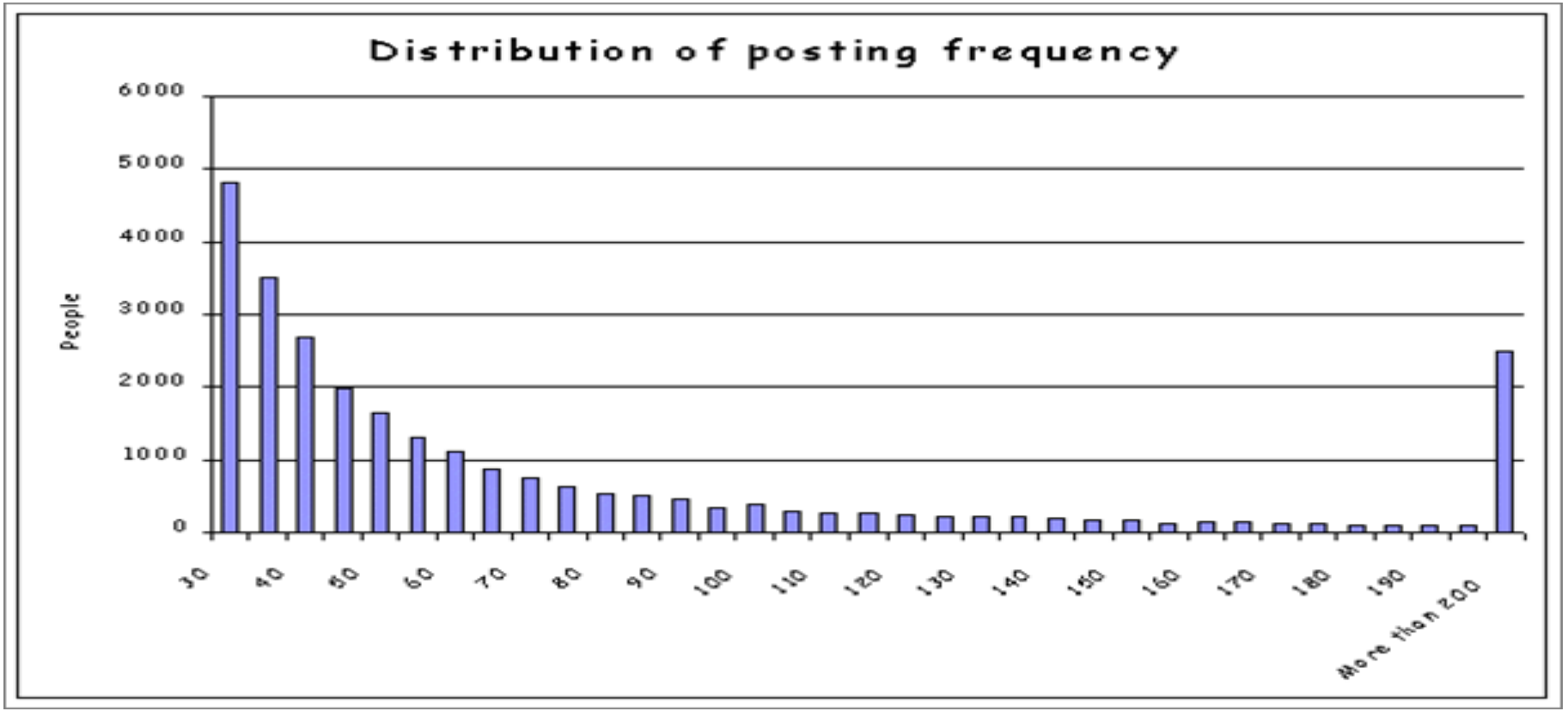
Skewness of Distributions

- Measures look at how lopsided distributions are—how far from the ideal of the normal curve they are
- When the median and the mean are different, the distribution is skewed. The greater the difference, the greater the skew.
- Distributions that trail away to the left are negatively skewed and those that trail away to the right are positively skewed
- If the skewness is extreme, develop strategies to minimize the impact of skewness.

Different Shapes of Distributions



Distribution of posting per month frequency on Facebook



Two Issues that Hinder Estimation

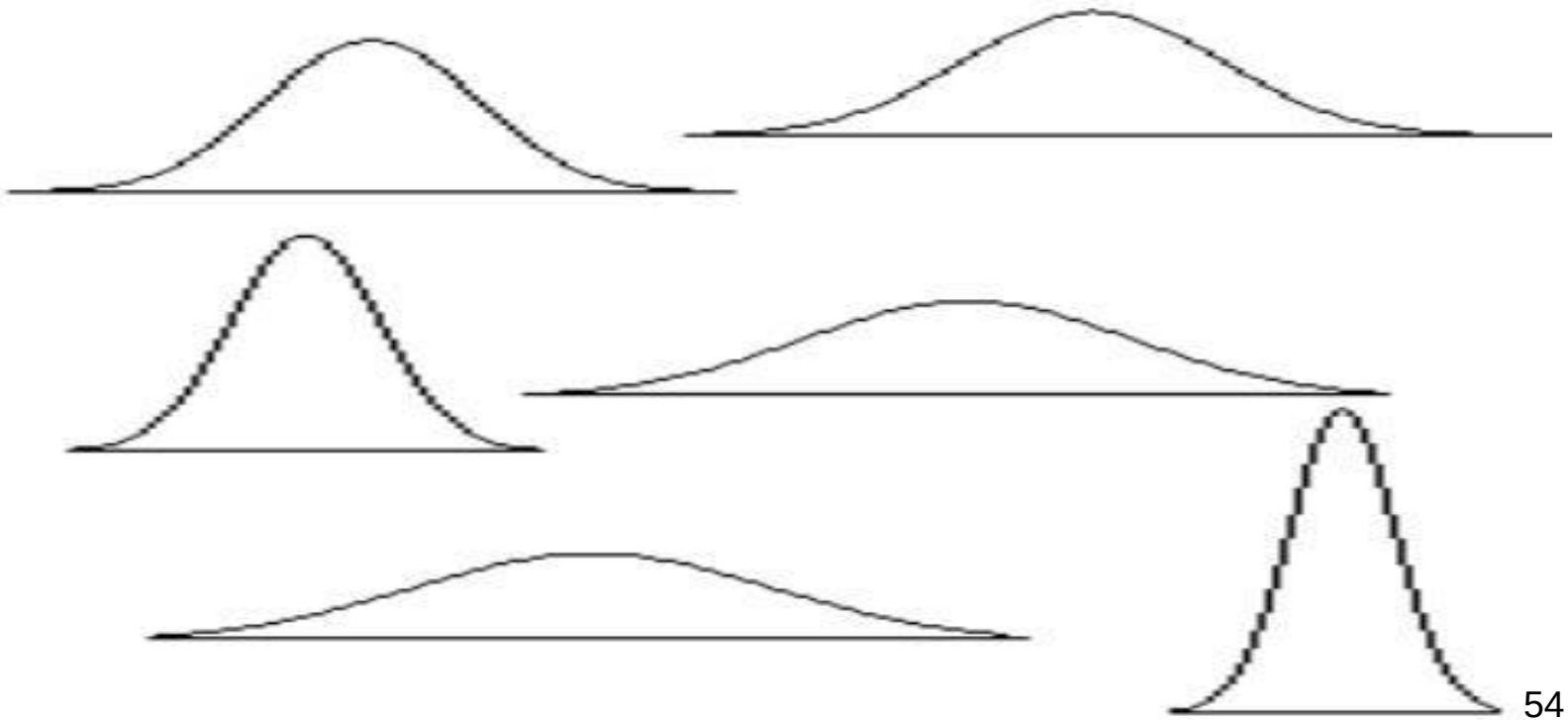
We have made the case that the normal curve is essential in calculating the amount of sampling error around the true parameter value but two outstanding issues must be solved before we can proceed, they are:

1. There are an infinite number of normal curves

1. Financial data tends to be very skewed or very far from normal.

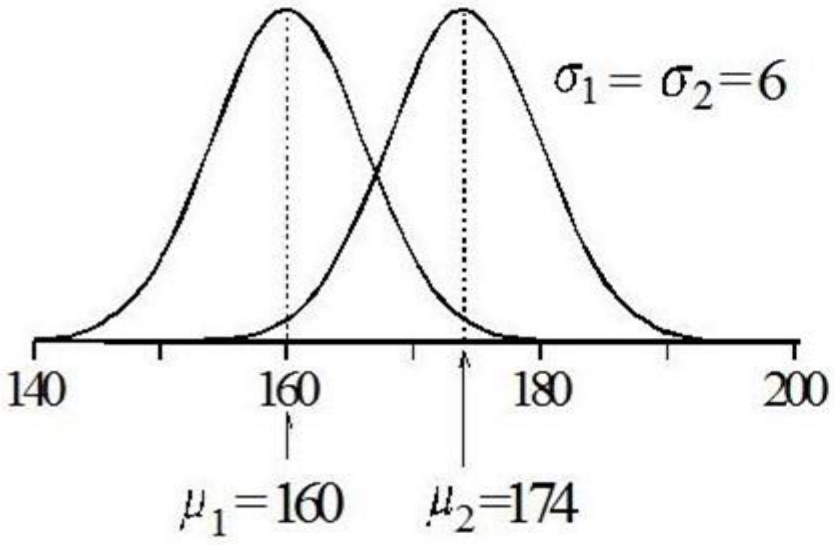
Let’s solve the first problem

Infinite Number of Normal Curves

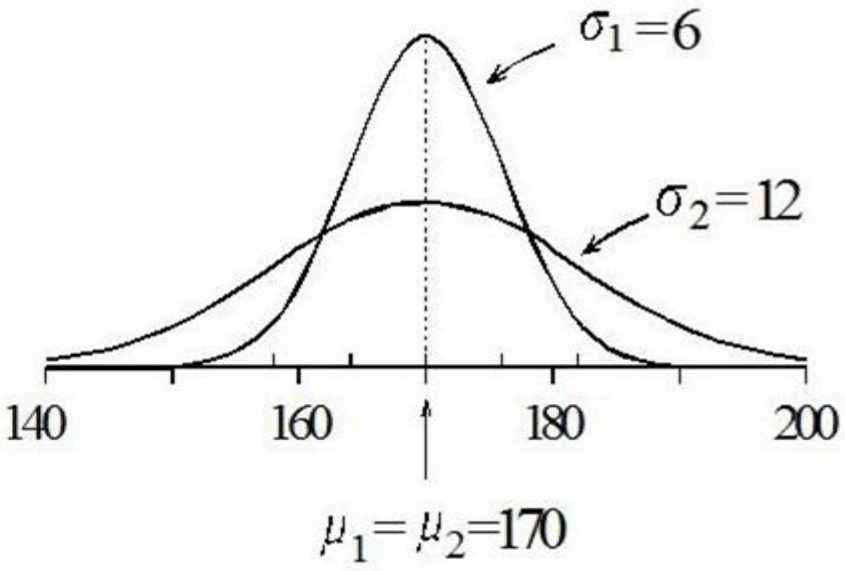


Infinite Number of Normal Distributions (cont.)

Shifts curve along axis



Increases the spread and flattens the curve



Abundance of Normal Curves Down to One

- The universe of normal curves is infinite
- We need to drill down to one normal curve to calculate our probabilities.
- We need a translation process to convert any normal curve into one uniform standard normal curve.
- The translation process involves any normal curve converted to one that has a constant mean and standard deviation.
- For the auditor this leaves only one variable, the dollar value ($X) of the transaction

IQ		μ = 100	σ = 15
X_1		μ = 10	σ = 2
X_2		μ = 1000	σ = 250

IQ	70	85	100	115	130
X_1	6	8	10	12	14
X_2	500	750	1000	1250	1500
Z	-2	-1	0	+1	+2

$$Z = \frac{X_i - \mu}{\sigma}$$

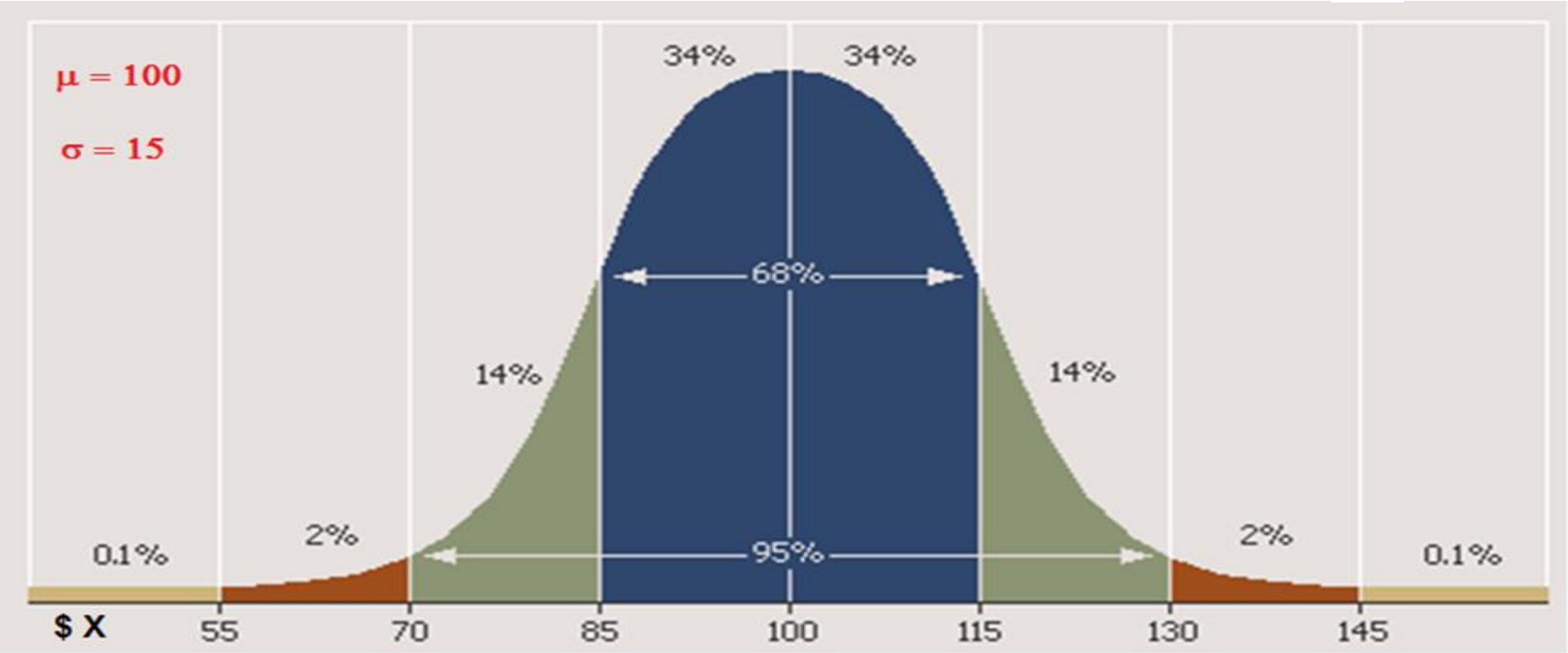
Placing the variable of interest (e.g. $X), mean and standard deviation in the above formula results into a single standard Z distribution

The Math for a Standard Curve is Now in Place

- We can now start to make statistical decisions about our observed results.
- I observe a certain outcome but what is the probability of that outcome.
- With this knowledge is the outcome to be expected or unusual?
- Lets look at the normal curve example again.

Normal Curve Revisited

$$z = \frac{X-\mu}{\sigma}$$



We have solved the issue of infinite normal curves

- Now we must solve the issue of sampling from a skewed distribution.
- There is a very powerful mathematical tool called the "Central Limit Theorem" that solves this second issue.
- The random sample process when properly applied can lead to normal curve based estimates.
- The key is understanding the behavior of random samples.

Sampling Distribution

- Because any statistic varies from sample to sample, it is a random variable and has its own probability distribution.

- The probability distribution of a statistic is called its sampling distribution.

- Often we simply say the sampling distribution of a statistic.

Sampling Distribution of the Mean

- A sampling distribution is the distribution of a sample statistic over all possible samples- in our audits we are interested in the sample mean.

- Definition:
  - Taking several samples from a population
  - Computing the mean of each sample
  - Forming a distribution of those sample means

The Central Limit Theorem

Gives us the extraordinary power to generalize from a well chosen subset of the population (random sample)

1. the sampling distribution of sample means is a normal distribution *if sample size is large enough*.
1. The mean of the sampling distribution of sample means equals the mean of the population,
1. The variance of the sampling distribution of sample means equals the variance of the population distribution divided by n.

Basic Idea of the Central Limit Theorem

- The core principle of the Central Limit Theorem is that a large, properly drawn sample will resemble the population from which it is drawn.
- Obviously there will be variation from sample to sample but the probability that any sample will deviate massively from the population is very low.
- There are formulas that give us a quantitative way of estimating with a certain level of confidence how well our sample is doing.

For example a well chosen sample of 1,200 Americans can tell us a great deal about how the entire country is thinking

Central Limit Theorem:

Sampling Distribution of Mean (SDM):

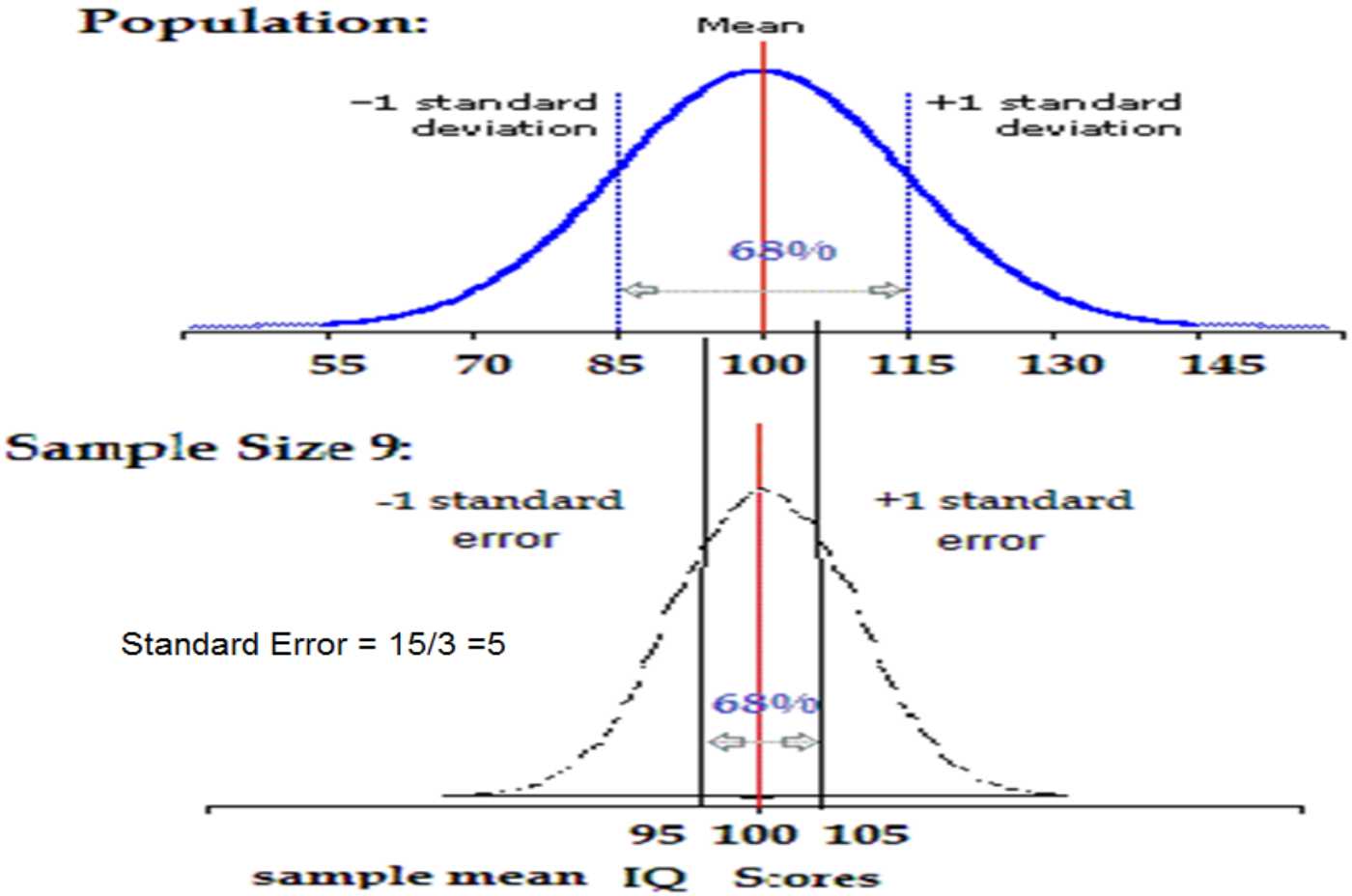
1. $\mu_{\text{sample means}} = \mu_{\text{population}}$
2. $\sigma_{\text{sample means}} = \sigma_{\text{population}} / \sqrt{n}$

   (standard error)

3. When sampling from almost any distribution, SDM is normally distributed:

if sample size is large enough!

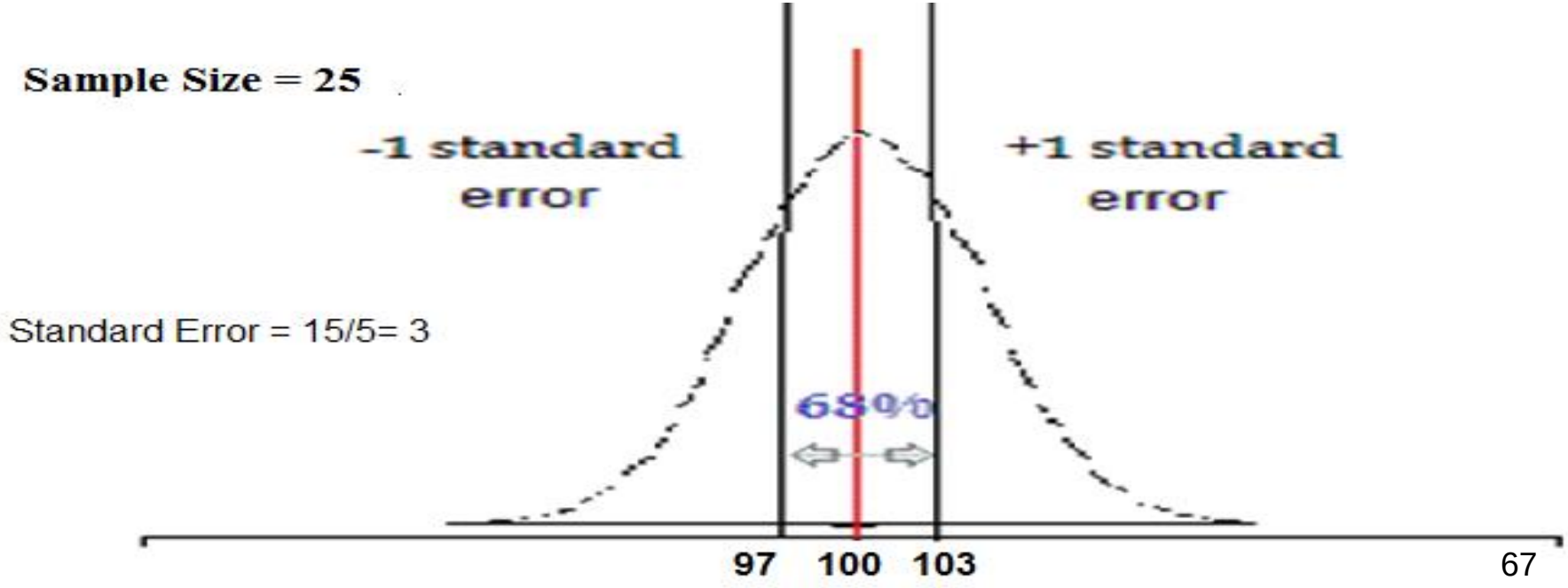
Sample Population with n = 9



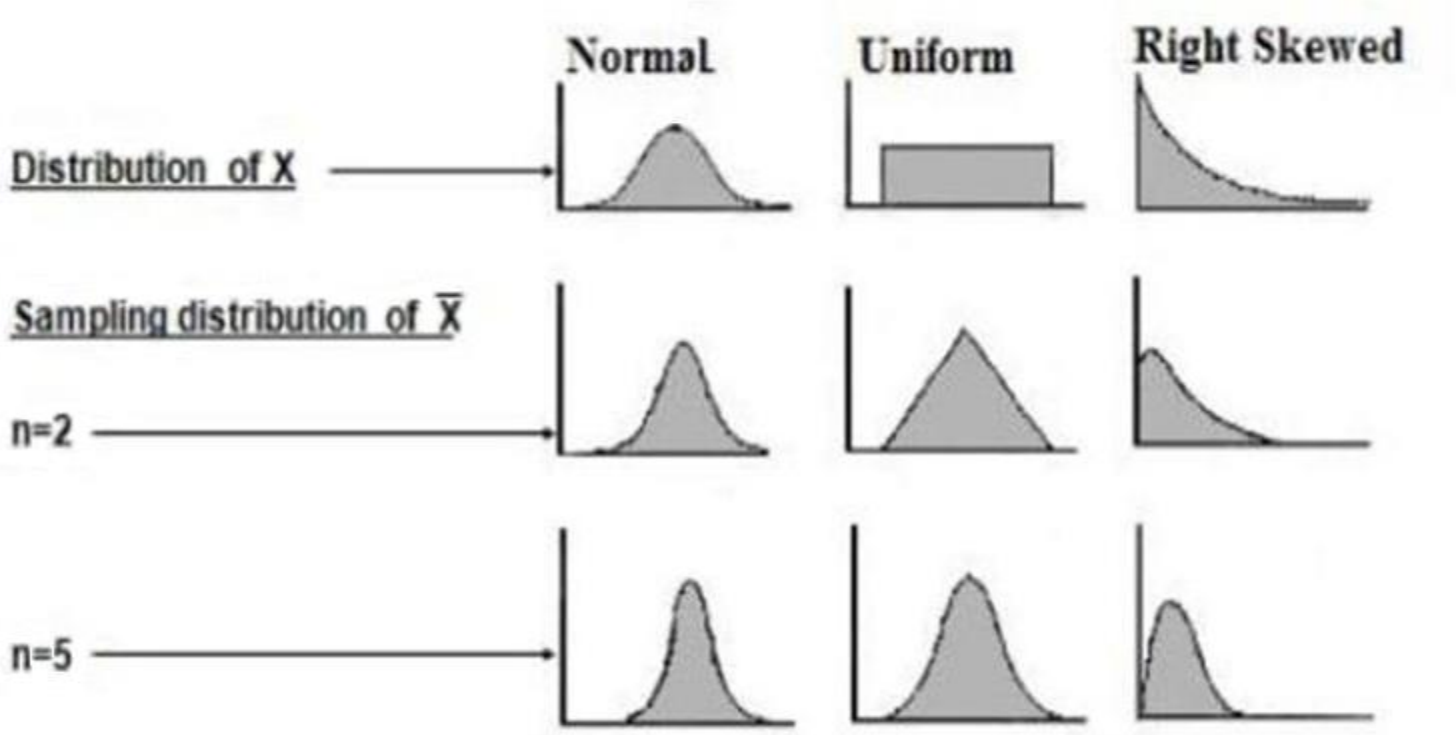
Increase Sample size to 25

Standard error goes from 5 to 3

68% observations from 95/105 range to 97/103



As Sample Size Increases Distribution of Sample Means Approach a Normal Distribution



Standard Deviation vs. Standard Error

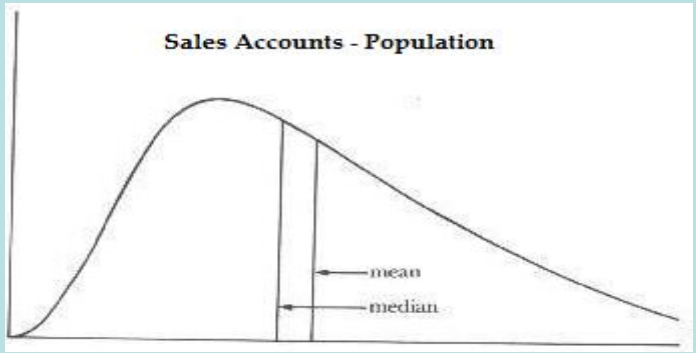
- Standard Deviation - measures dispersion in underlying population e.g. the dispersion of all sales in a chart of accounts.

- Standard Error – measures the dispersion of sample means e.g. if we draw repeated sample of n=100 what will be the dispersion of those sample means.

- Where the two concepts tie together is standard error is standard deviation of sample means.

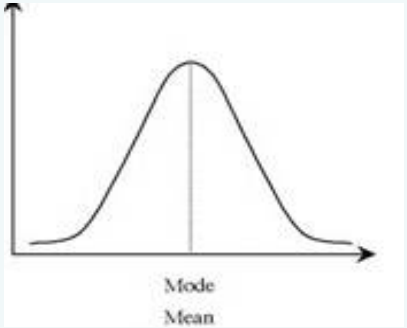
Account vs. Sample Means

Distribution Of Sales Accounts



Distribution of Sample Means

Dependent on Sample Size
n=100



Harnessing the power of the central Limit Theorem

Finally we arrive at the payoff!

Because sample means are distributed normally:

- We expect roughly 68% of all sample means lie within one standard error
- Roughly 95% of all sample means to lie within two standard errors
- Roughly 99.7% of all sample means to lie within three standard errors.

With Properly Drawn Samples:

- We now have a process where the data speaks to us.
- The point is not to do mathematical calculations but for us to gain insight into a population too large to be measured directly.
- It is the merging of data and probability with the help of the Central Limit Theorem.

Defining the Boundaries Around Uncertainty

- I have talked about uncertainty in broad terms but we need to be more precise.
- Let's start with a simple example that involves percentages that is like the error rate you calculate for the audit.
- But let's start with a simple somewhat trivial example in which we know a-priori (beforehand) all the probabilities involved.
- Any error rate is based on what is called a dichotomous decision: success/failure, yes/no, 0/1, taxes due/taxes not due etc.
- If such percentages are derived from a random process, then the observations would follow a distribution called the binomial.

Defining the Boundaries Around Uncertainty (cont.)

- This is a distribution that has similar properties to the normal curve. When you derive a percentage from a fairly large sample, as you do in an audit, then for all practical purposes the binomial merges into a normal curve.

- That is convenient because we can use the properties of the normal curve in making decisions.

- We will cover this in a little more detail later in the course but for now let's start with a binomial example with a small set of observations:

  - We will start with very simple binomial outcomes, the flip of a coin.
  - With its known probability of .5 heads and .5 tails let's set up a test of how to establish precise boundaries around our decision criteria.

Data Testing:
for example is a coin fair or biased

- Suppose we have a game of chance where you win $10 if the coin is tails and pay $10 if it is heads.
- You may be concerned that I doctored the coin to favor me.
- I will not let you see the coin but will allow you to do a statistical experiment- toss the coin 5 times.
- From that data set a decision rule that the probability of the 5 tosses in my favor is less than 5%, you will declare the coin “biased” and reject that heads will at least occur 50% but decidedly in my favor i.e. heads >50%.

Level of risk- Need to make a decision
Result of sample (trials) n = 5

- Define success- heads/ failure- tails (remember tails you lose)
- Decision – is the coin I am given true i.e. pr(heads)=.5?

Set Decision Criteria if probability of event of interest is p<5%
Then I will say that the coin is not fair or true

# Heads	Probability (based on binomial)
0	3.10%
1	15.60%
2	31.30%
3	31.30%
4	15.60%
5	3.10%

- The gray area is the critical decision zone where we say we reject it is a true coin.

Decision Based On Our Results.

- Results 5 heads in a row - probability is 3.1% was in our critical decision zone of <5%
- Therefore we conclude that result is rare and we conclude the coin is not true.
- Can we be mistaken in that decision!
  - Yes, a true coin can come up heads 5 times in a row
  - Its probability would be 3.1%
  - the probability of an error in incorrectly rejecting the coin as true is 3.1%
  - Statisticians call this alpha or type I error, AICPA calls this “error of incorrect rejection”.
  - Our decision has a 3.1% alpha error.

Let's Make this Example Analogous to the Audit Process

- In the audit process we are more interested in how reliable is our sample, what are its reasonable outcomes.
- In this simple case our sample is 5 observations or tosses of the coin, n=5.
- Can we come up with probabilities of many possible samples around a true value?
- We can now look at the previous table and also make a decision about the spread of our possible sample outcomes.
- In the next example we will do what is called a two tailed test.

Possible Sample Outcomes

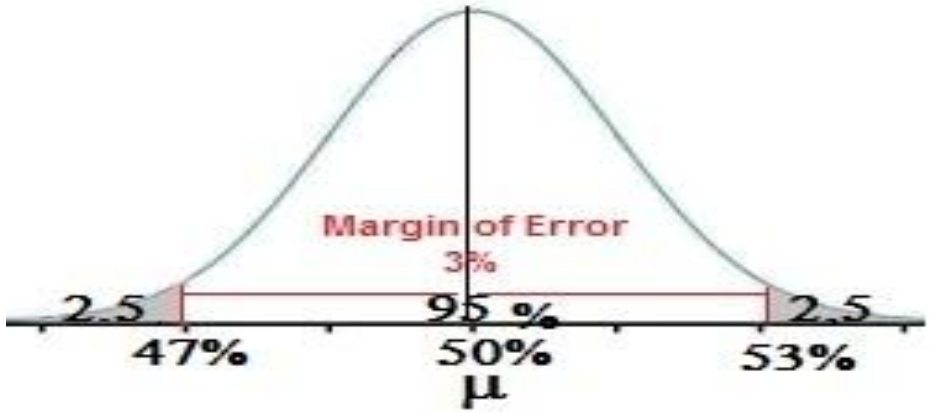
- Our boundaries now tells us in the long run how samples of size 5 behave with probabilities based on a yes/no random event.

	Possible sample Outcomes
# Heads	Probability (based on binomial)
0	3.10%
1	15.60%
2	31.30%
3	31.30%
4	15.60%
5	3.10%

- This is analogous in stating that extreme outcomes (gray alpha error) occurs only 6.2% of the time with a 93.8% confident interval within that alpha error zone.

Audit Sample Specifications

- **Audit sample criteria *(suppose we are talking about account errors*):**
  - 95% confident sample mean is within 3% of true mean **or**
  - 95 samples out of 100 would be within 3% of true mean



- For our μ = 50% then 95% probability that sample means are between 47% and 53%, for the binomial in this example the main driver is sample size.

Making Decisions Under Conditions of Uncertainty

- Allows us to use sample data to determine the characteristics of a population, such as testing whether a population proportion, mean or total value equals some number.
- However we know it is unlikely that a sample would be an exact replica of the population.
- Can we establish a range with probability being outside that range, alpha error, would be <= 5%, within that range 95%?
- Can we build a 3% margin of error around our sample projection e.g. 95% confident sample mean is within 3% of the true population mean?

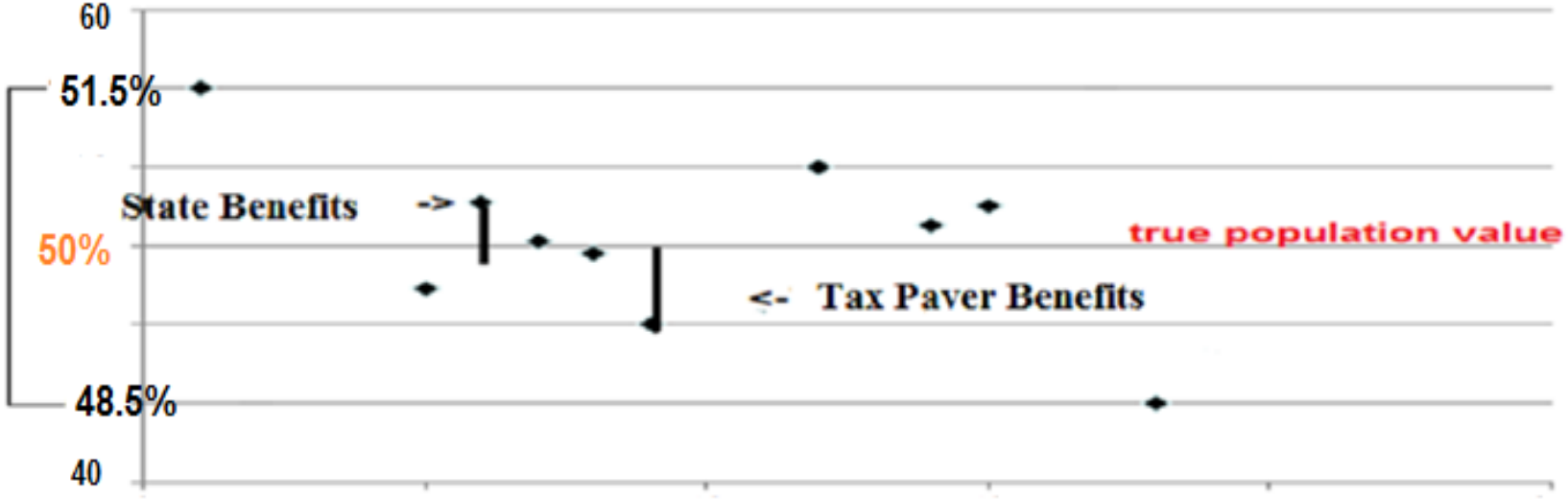
Error Rate for Transactions

- An auditor is making a dichotomous (yes/no) decision whether a transaction is in error or not.
- We can now use the same binomial distribution as in the coin toss to help make a decision.
- Now our example is not so trivial.
- Like in the coin toss lets assume the account error is 50%, a bit extreme but let’s look at the decision consequences.

Error Rate for Transactions (cont.)

- With the help of the Central Limit Theorem we can establish a sample size giving us 95% probability our sample error rate will be within 3% of the true population value.
- If we take multiple samples over time the sample error rates will follow the normal distribution, some greater others less than the true value.
- Over time the collective sample error rates will converge onto the true population value.

Audit Error Rate (multiple sample means)



Each point is a separate sample, 68% of the error rates will be between 48.5% and 51.5%. Collectively they will converge on the true population value of 50%

The Statistical Audit & Bias

- A biased selection for the sample would be some data points are over selected and others under selected e.g. a block sample that may over represent some transaction over another.
- When selecting a sample for the statistical audit, random selection software is used to assure independent selection once the audit population is properly defined.
- The most important first step is to accurately define the audit population to be assured of no inadvertent omission of relevant transactions or the inclusion of unnecessary transactions (dead wood).

The Statistical Audit & Risk (cont.)

- Once the audit population is properly defined and independent selection guaranteed by random selection software then the power of the Central Limit Theorem can be harnessed.
- We can now proceed to calculate a sample size that encompasses the margin of error or precision and level of risk we are willing to accept. That is:
  - 3% Precision also called Margin of Error.
  - 5% Alpha Error or Error of Incorrect Rejection.