PREVIOUSLY

use known probability distributions (binomial, Poisson, normal) and know population parameters (mean, variance) to answer questions such as ...

... given 20 births and a probability of low birth weight of p=0.10, what is the probability that 3 or more low birth weight infants are born ...

... given that birth weight is distributed normally with a mean of 3750 grams and a standard deviation of 500 grams, what is the probability that an infant will have a birth weight of at least 2724 grams (6 pounds) ...

NEW PROBLEMS

- how can a sample be used to estimate the unknown parameters of a population
- how can we estimate central tendency (mean, median, mode)
- how can we estimate variability (variance, quartiles, range)
- how can we set intervals around the estimates (how "sure" are we our estimates)
- of all the measures of central tendency and variability, which are the "best"

EXAMPLES

previously ... hypertension ...

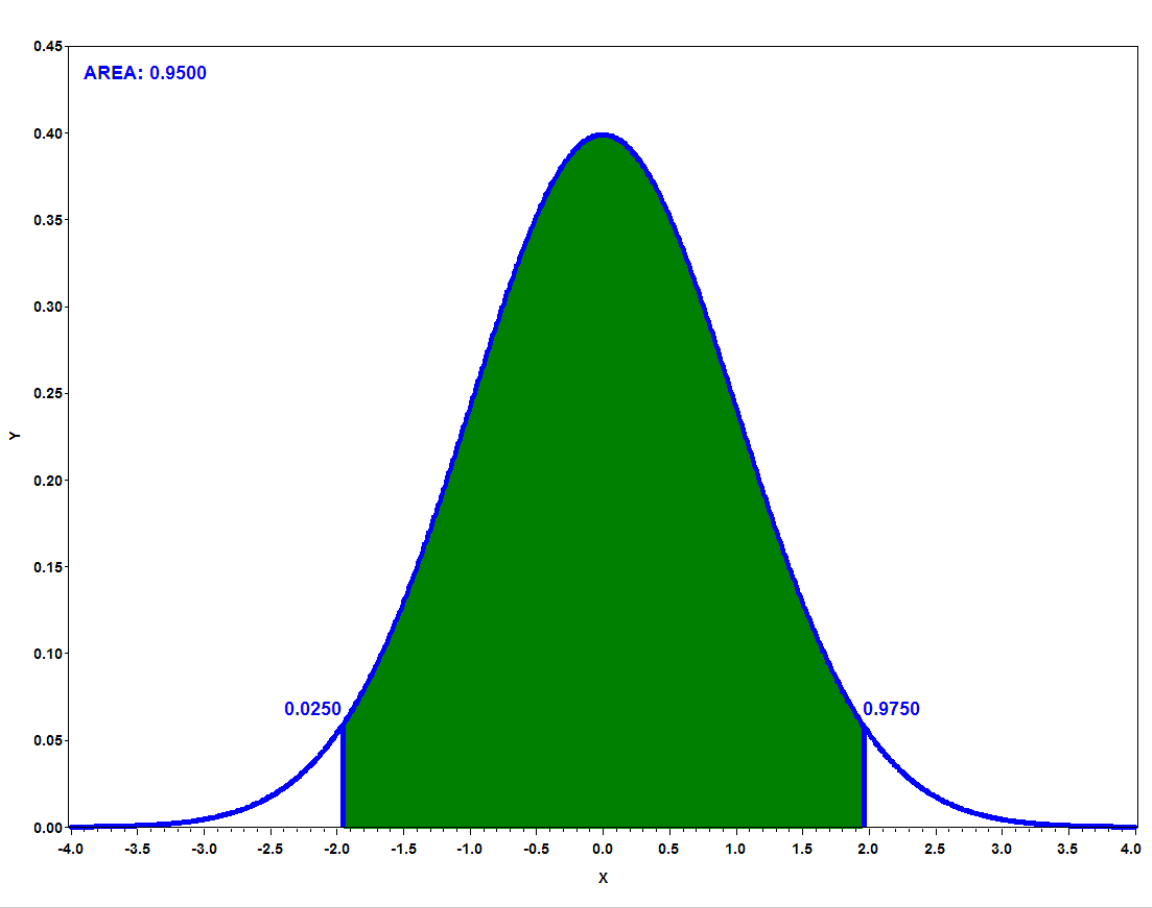
assume that the distribution of diastolic blood pressure (DBP) in 35-44 year old men is NORMAL with mean μ = 80 and standard deviation σ = 12 ...

one can use this information to determine that 95% of all diastolic blood pressures among 35-44 year old men should fall between 56 and 104

$z = \frac{\overline{X} - \mu}{s} = \frac{104 - 80}{12} = 2 \sim 1.96$	$z = \frac{\overline{X} - \mu}{s} = \frac{56 - 80}{12} = -2 \sim -1.96$

95% of area lies between ±2 standard deviations (more accurately, between ±1.96 standard deviations) of the mean ... so ...

P(56<X<140) = 0.95



new task ... hypertension (continuous data) ...

assuming that the distribution of DBP in 34-44 year old men is NORMAL, what is the mean DBP in the population (what is μ and how precise is the estimate of μ)

new task ... hypertension (continuous data) ...

assume that the distribution of DBP in 34-44 year old men is NORMAL with mean $\mu = 80$ and standard deviation $\sigma = 12$ ... you measure DBP on a sample of thirty (not 1) 34-44 year old men ... is there evidence to say that the mean of your sample differs from the population mean

new task ... infectious disease (discrete data) ...

assuming that the number of cases in the population follows a binomial distribution, what is the prevalence of HIV-positive people in a low income census tract in an urban area (what is P, how precise is the estimate of P)

new task ... infectious disease (discrete data)...

assume that in treating gonorrhea, a daily dose of penicillin of 4.8 megaunits has a failure rate of 10% (P = 0.10) ... you treat 46 patients with a daily dose of 4.0 megaunits of penicillin and find that after a week, 6 patients still have gonorrhea ... is there evidence to show a difference in the failure rate of the two different doses

WHEN A POPULATION IS LARGE ... SAMPLING

given that one cannot measure DBP in ALL 35-44 year old men or obtain results of an HIV test for ALL people in the census tract, one must rely on a SAMPLE to estimate POPULATION parameters ... important considerations ...

- reliability

  the sample reflects the population ... sample(s) are collected in a manner that allows you to estimate how much sample results might differ from the results that would be obtained if the entire population was used

- economics

- random

  unbiased ... each unit (person, place, ...) has the same chance of being chosen

  independence ... selection of one unit (person, place, ...) has no influence on selection of other units

types of samples ... from Cartoon Guide, Chapter 6 ...

- **simple random** ... a sample of size N from a population such that each possible sample of size N has an equal probability of being chosen (Rosner ... each group member has an equal probability of being chosen, plus discussion of random number tables --- chance of using them, P <0.05)

- **stratified** ... divide the population into homogeneous groups (strata) based on some attribute of population members (gender, age, ...), then select a simple random sample within each subgroup

- **cluster** ... divide the population in groups (clusters), randomly select a clusters, then select all members (or randomly sample) within the selected clusters

- **systematic** ... select every K^{th} member of the population (every 5^{th}, 10^{th}, 100^{th}, ...)

- **opportunity** (convenience) ... whatever is available

ESTIMATES OF CENTRAL TENDENCY AND VARIATION

central tendency ...	variation ...
mean	variance
median	quartiles
mode	range

what's best ... one quality of "best" is "unbiased" ...

the average value of the estimate over a large number of repeated samples is the population value

example (Triola) ... using a population with only three members ... 1, 2, 5 ... take all possible samples of size 2, compute statistics, find average value of statistics over all samples ... compare the average values to population values ... which average values equal population values ...

SAMPLE	MEAN	MEDIAN	RANGE	VAR	SD
1,1	1.0	1.0	0	0.0	0.000
1,2	1.5	1.5	1	0.5	0.707
1,5	3.0	3.0	4	8.0	2.828
2,1	1.5	1.5	1	0.5	0.707
2,2	2.0	2.0	0	0.0	0.000
2,5	6.5	3.5	3	4.5	2.121
5,1	3.0	3.0	4	8.0	2.828
5,2	6.5	3.5	3	4.5	2.121
5,5	5.0	5.0	0	0.0	0.000
MEAN	2.7	2.7	1.8	2.9	1.3
POPULATION	2.7	2	4	2.9	1.7
UNBIASED	Y	N	N	Y	N

- another quality … estimator of central tendency with the minimum variance … Rosner uses an example with 200 samples of N=200 from a population of 1000 births … the distribution of the sample means, medians, mean of the high/low values are shown (figure 6.2)

- the distribution of the sample means is said to have the minimum spread (conclusion … minimum variance)

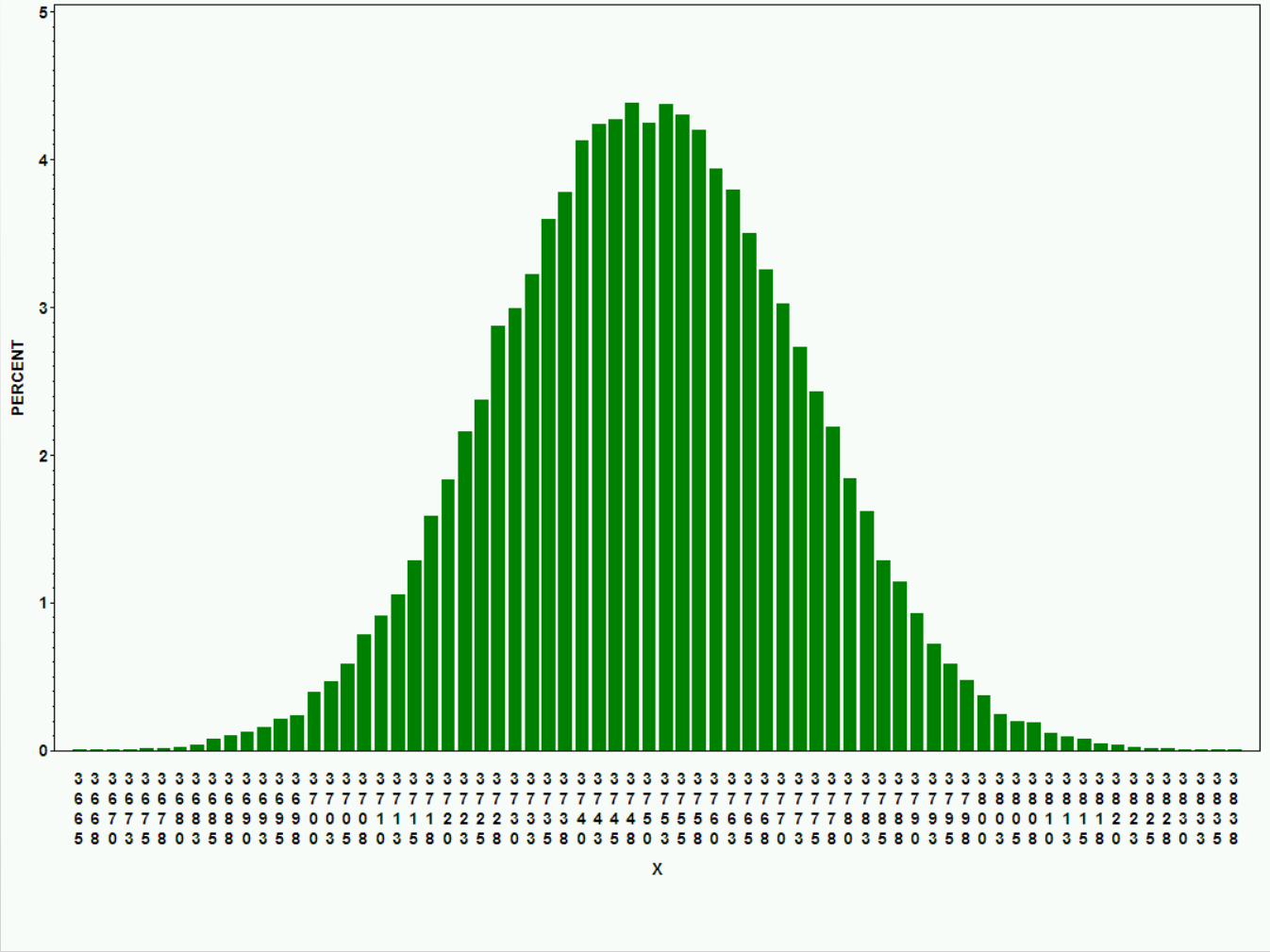
- the same is true for sample estimates of proportions

- another illustration … use SAS to generate a population of 10,000 numbers that are normally distributed with a mean of 3750 and a standard deviation of 500

- in notation terminology … $X \sim N(3750,250000)$

- the mean and standard deviation were chosen to approximate a birth weight distribution

the 10,000 values of x are distributed as shown on the right (looks like a NORMAL distribution) and they have the following characteristics ...



Mean	Std Dev	Variance	Minimum	Maximum
3750	501	251729	1831	5728

- take 100 simple random samples of N=10 from the 10,000 values

- compute the variance of the 100 sample means, medians, and means of the high/low values

- results...

Variable	Mean	Variance
mean	3755	23715
median	3759	33945
mean high/low	3757	38955

- all point estimates means are close to the population mean, but the mean has the smallest variance ... minimum variance, unbiased estimator

STANDARD ERROR

- the variability of a population is measured by the variance (or standard deviation)
- the variability of a set of sample means obtained by repeated random samples of size N from a population is measured by the standard error of the mean, defined as ... $\sigma / \sqrt{N}$
- the standard error IS NOT the standard deviation of individual values ... it IS the standard deviation of sample means ... this is repeated a number of times in Rosner (also discussed in BMJ handout, *Standard deviations and standard errors*)
- standard error a function of population variability (σ) and sample size (N)

Rosner ... review questions 6A,B ...

what is a random (simple random) sample

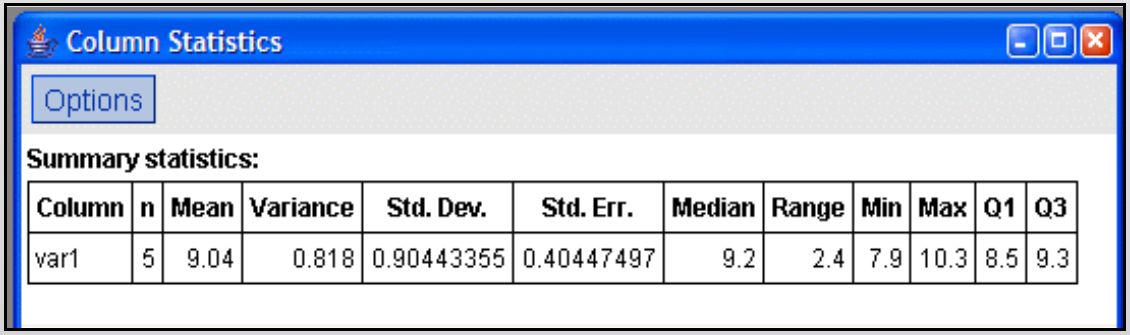
what is a sampling distribution (definition 6.10)

why is the sample mean used to measure the population mean (better question: why use the sample mean as a measure of central tendency of a population)

what is the difference between a standard deviation and a standard error

review continued ...

given 5 sample values obtained from a patient (8.5, 9.3, 7.9, 9.2, 10.3) ... what is the standard deviation? standard error?



Column	n	Mean	Variance	Std. Dev.	Std. Err.	Median	Range	Min	Max	Q1	Q3
var1	5	9.04	0.818	0.90443355	0.40447497	9.2	2.4	7.9	10.3	8.5	9.3

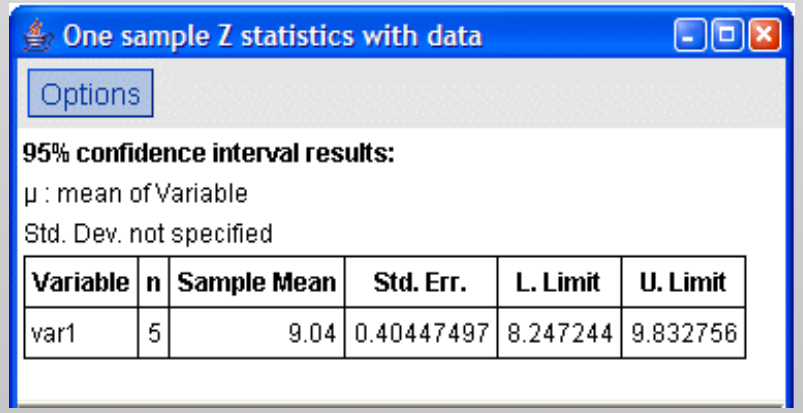
if sample size is increased to 20, would you expect the standard deviation to increase, decrease, stay the same? same questions for the standard error ...

estimation ... you take a sample of size N and use it to estimate the population mean ... you also want to make some statement as to the variability of your estimate

Rosner ... if the variable of interest is distributed normally with mean μ and variance σ^2, then the distribution of sample means is also normal with mean μ and variance σ^2 / N

so ... you can use what you already know about the normal distribution to compute an estimate of the variability of your estimate

using Rosner data from review question 6B4 ... the confidence interval uses z = ±1.96 from Table 3 (2.5% of area in each tail of the normal distribution) ... NOTE: this is NOT CORRECT since you are estimating the standard deviation from your data, it is just for illustration



CENTRAL-LIMIT THEOREM

- if a sample is large enough, the distribution of sample means can be approximated by a normal distribution … allows us to make quantitative statements about the variability of the sample mean

- the above holds regardless of the distribution of the underlying population

  so … what is large enough? … Rosner gives no number … Triola and Daniel suggest $N \geq 30$ … all state "the larger N, the better" and that large enough is a function of the degree of "non-normality" in the underlying population (so … a lot of unknowns !!!)

Triola … given a random variable x that may or may not be distributed normally with mean μ and standard deviation σ … and … simple random samples all of the same size N … then

the distribution of the sample means approaches a normal distribution as the sample size increases … and … the mean of all the samples is the μ, the population mean … and …the standard deviation of all the sample means is $\sigma / \sqrt{N}$ (also referred to as the standard error)

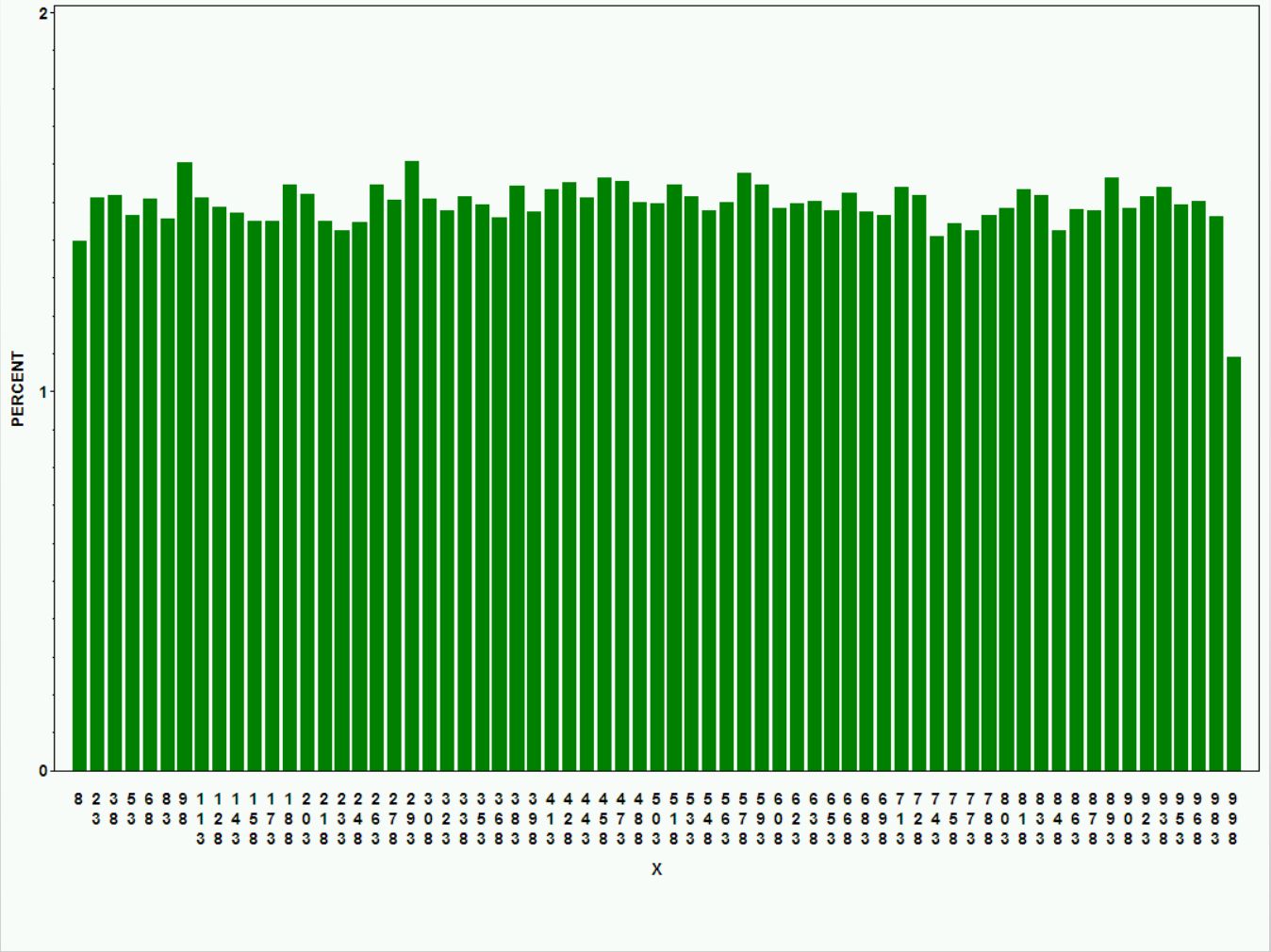
common guideline …

if the original population IS NOT normally distributed, sample sizes greater than 30 result in a distribution of sample means that can be approximated well by a normal distribution

if the original population IS distributed normally, the sample means will be distributed normally for any size sample

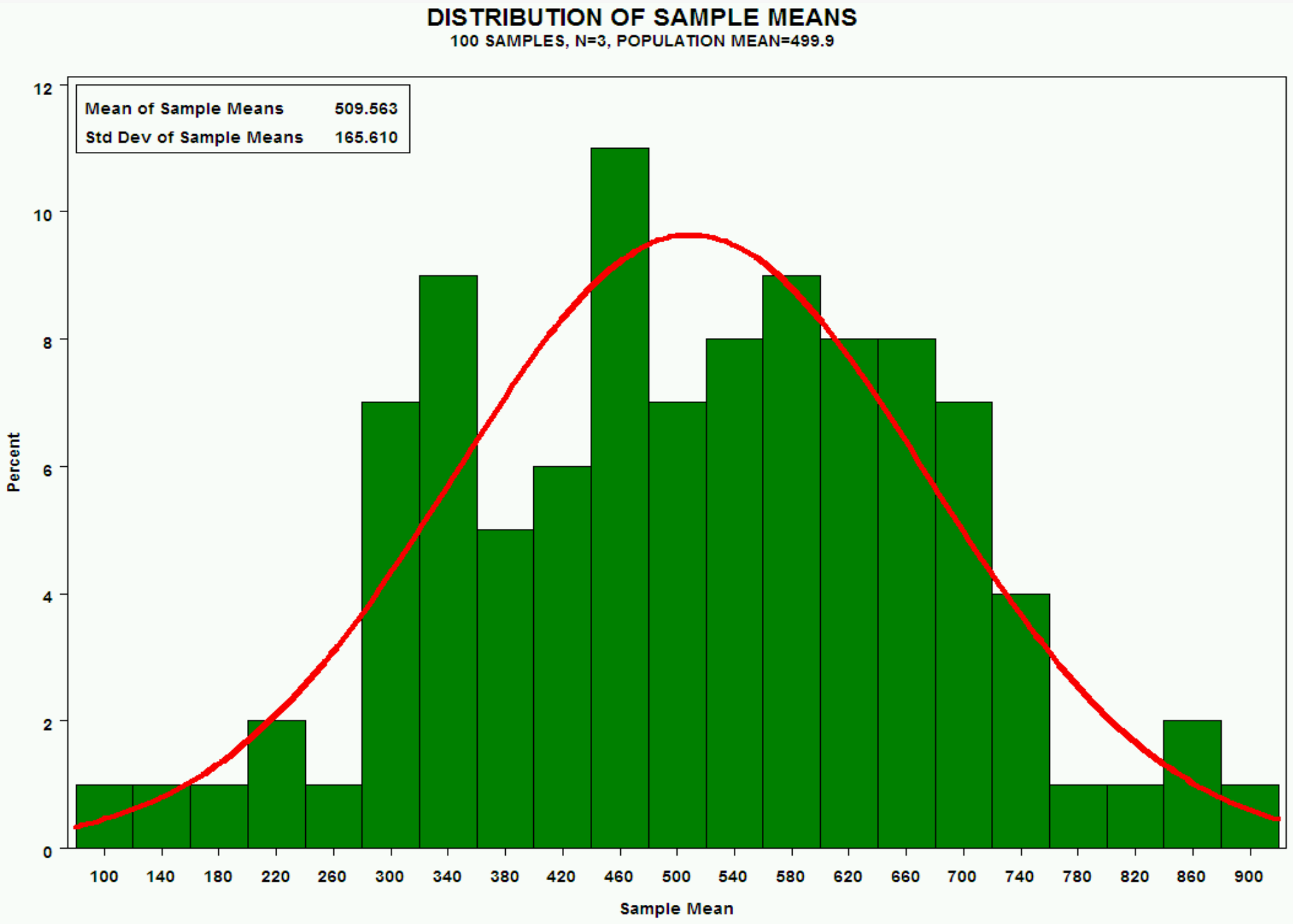
illustration of the central-limit theorem

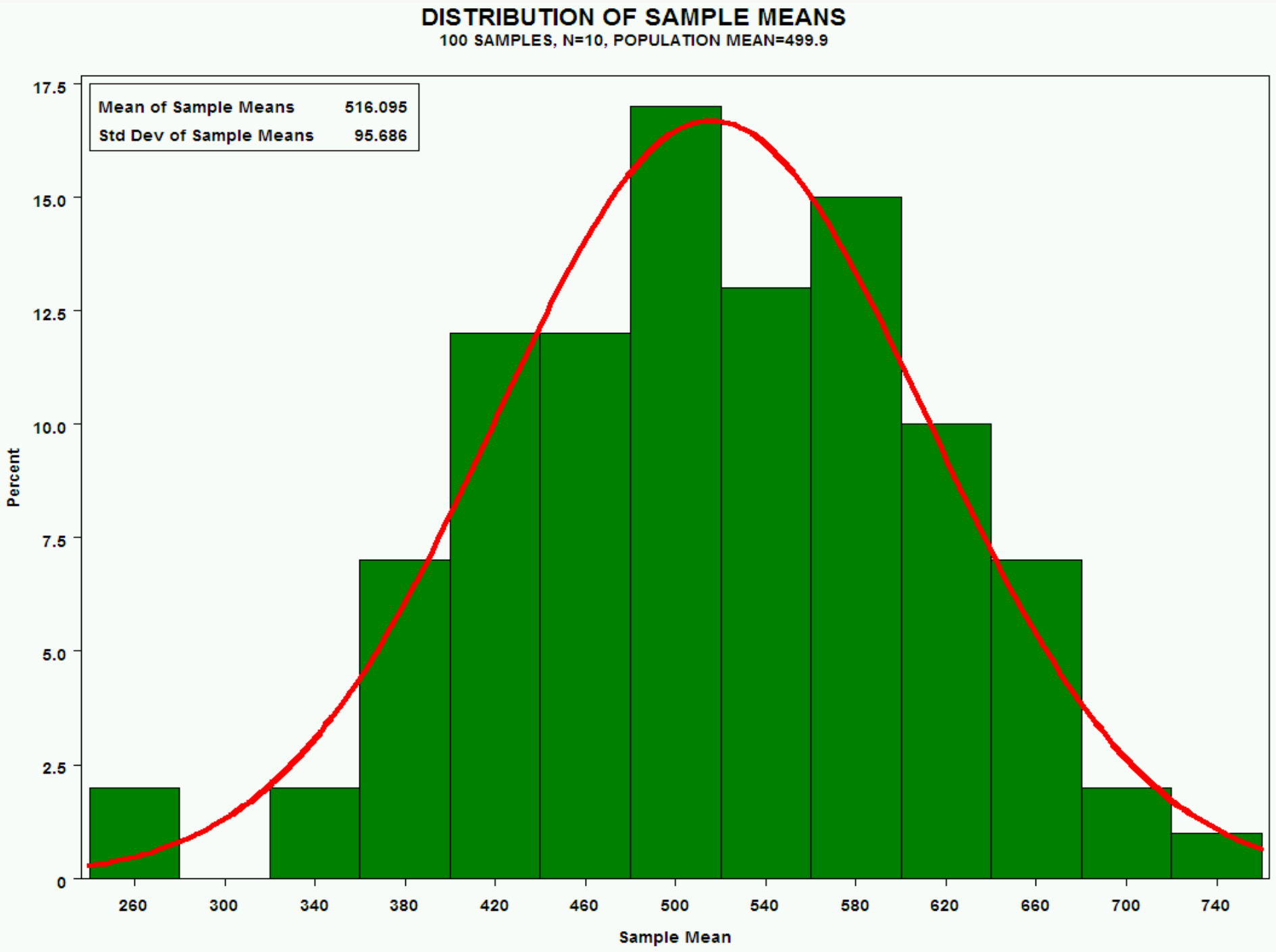
sampling from a population that IS NOT normally distributed (on right) ... 10,000 values of X with a UNIFORM distribution

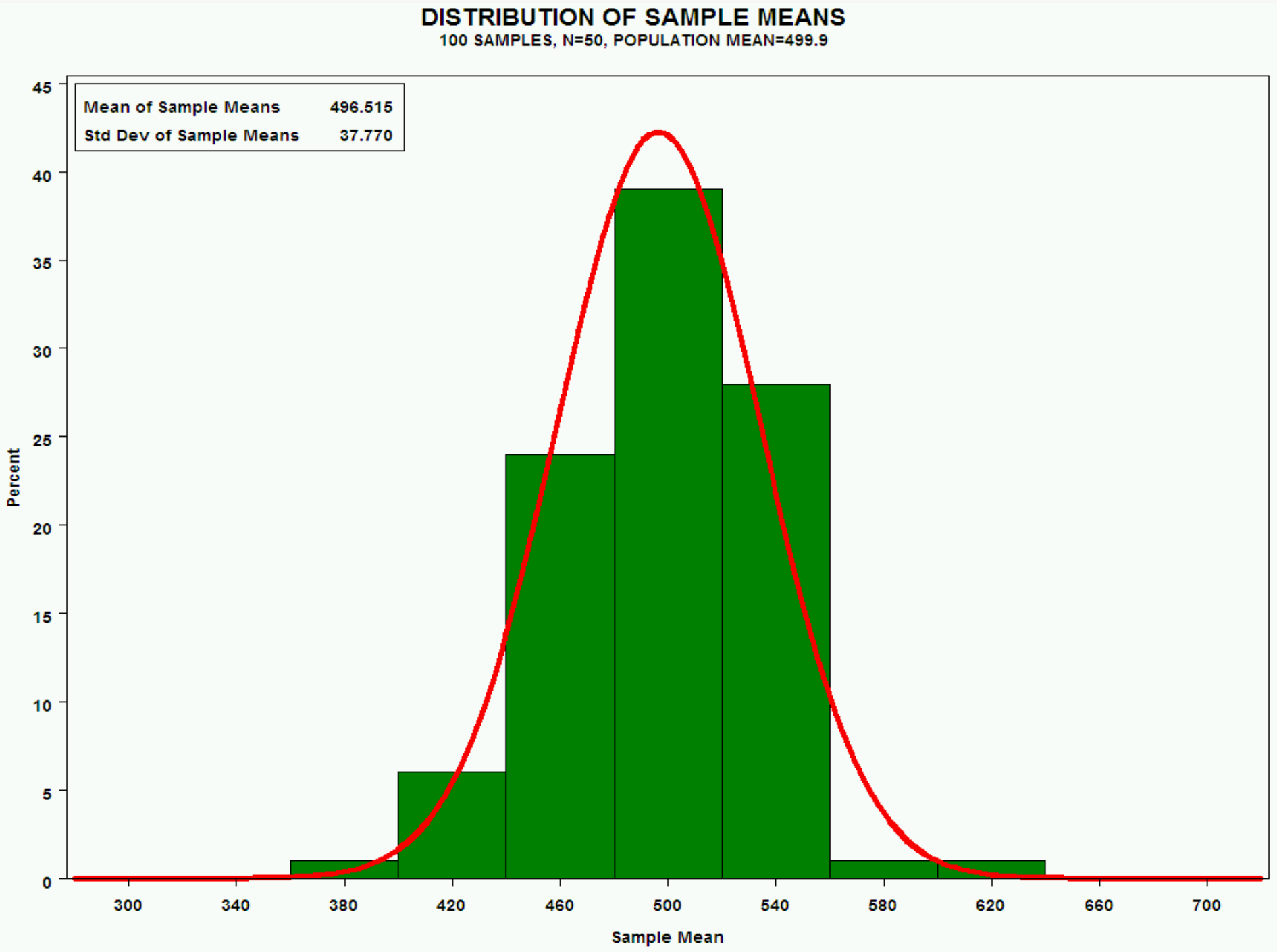


Mean	Std Dev	Variance	Minimum	Maximum
499	288	83071	1	1000

- use SAS to take a series of 100 simple random samples ... three different sizes ... N=3, N=10, N=50
- what is the distribution of sample means


DISTRIBUTION OF SAMPLE MEANS
100 SAMPLES, N=3, POPULATION MEAN=499.9
Mean of Sample Means 509.563
Std Dev of Sample Means 165.610
Percent
0
2
4
6
8
10
12
100 140 180 220 260 300 340 380 420 460 500 540 580 620 660 700 740 780 820 860 900
Sample Mean

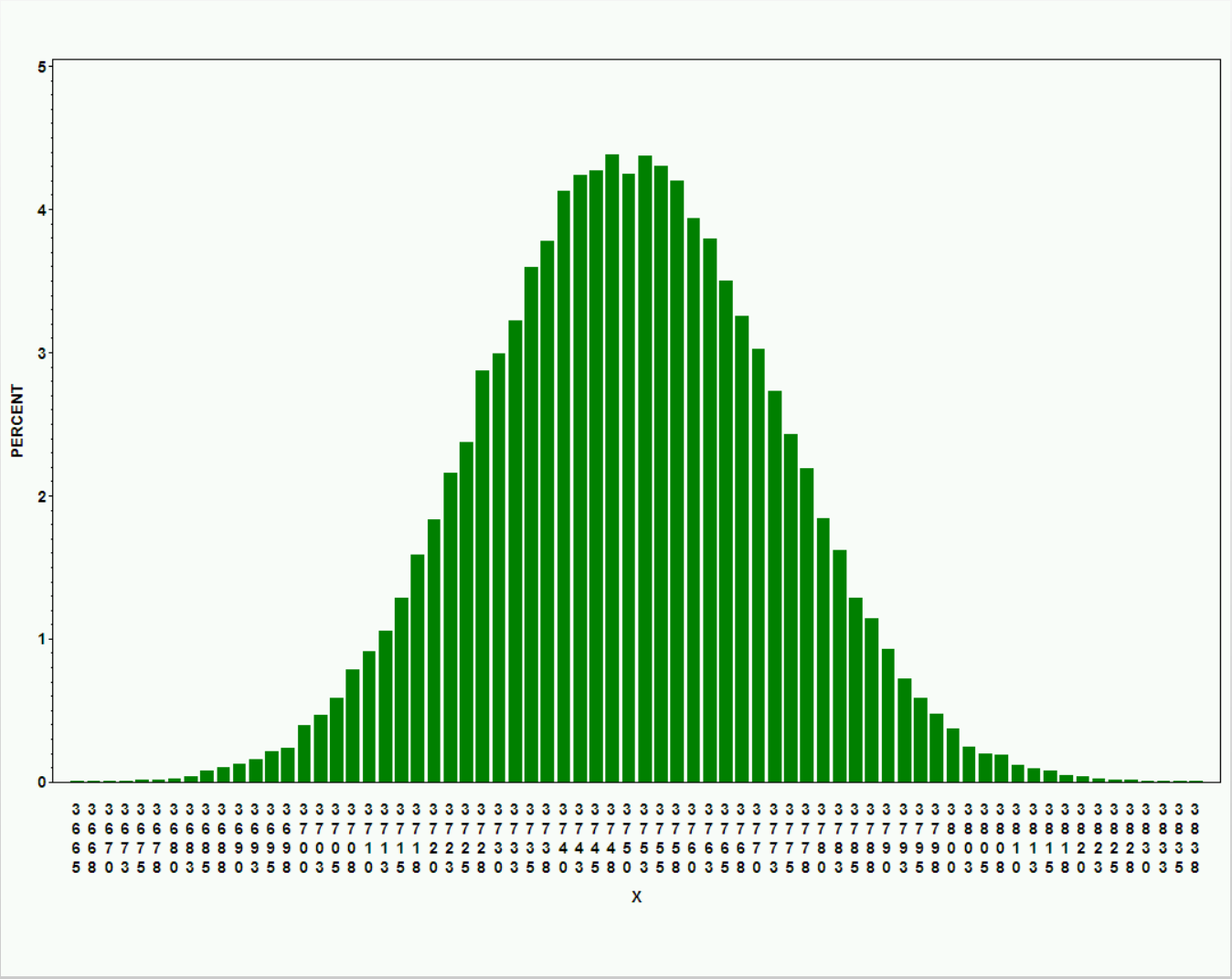

DISTRIBUTION OF SAMPLE MEANS
100 SAMPLES, N=10, POPULATION MEAN=499.9
Mean of Sample Means 516.095
Std Dev of Sample Means 95.686
Percent
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0
260
300
340
380
420
460
500
540
580
620
660
700
740
Sample Mean


DISTRIBUTION OF SAMPLE MEANS
100 SAMPLES, N=50, POPULATION MEAN=499.9
Mean of Sample Means 496.515
Std Dev of Sample Means 37.770
Percent
0
5
10
15
20
25
30
35
40
45
300
340
380
420
460
500
540
580
620
660
700
Sample Mean

repeat the process

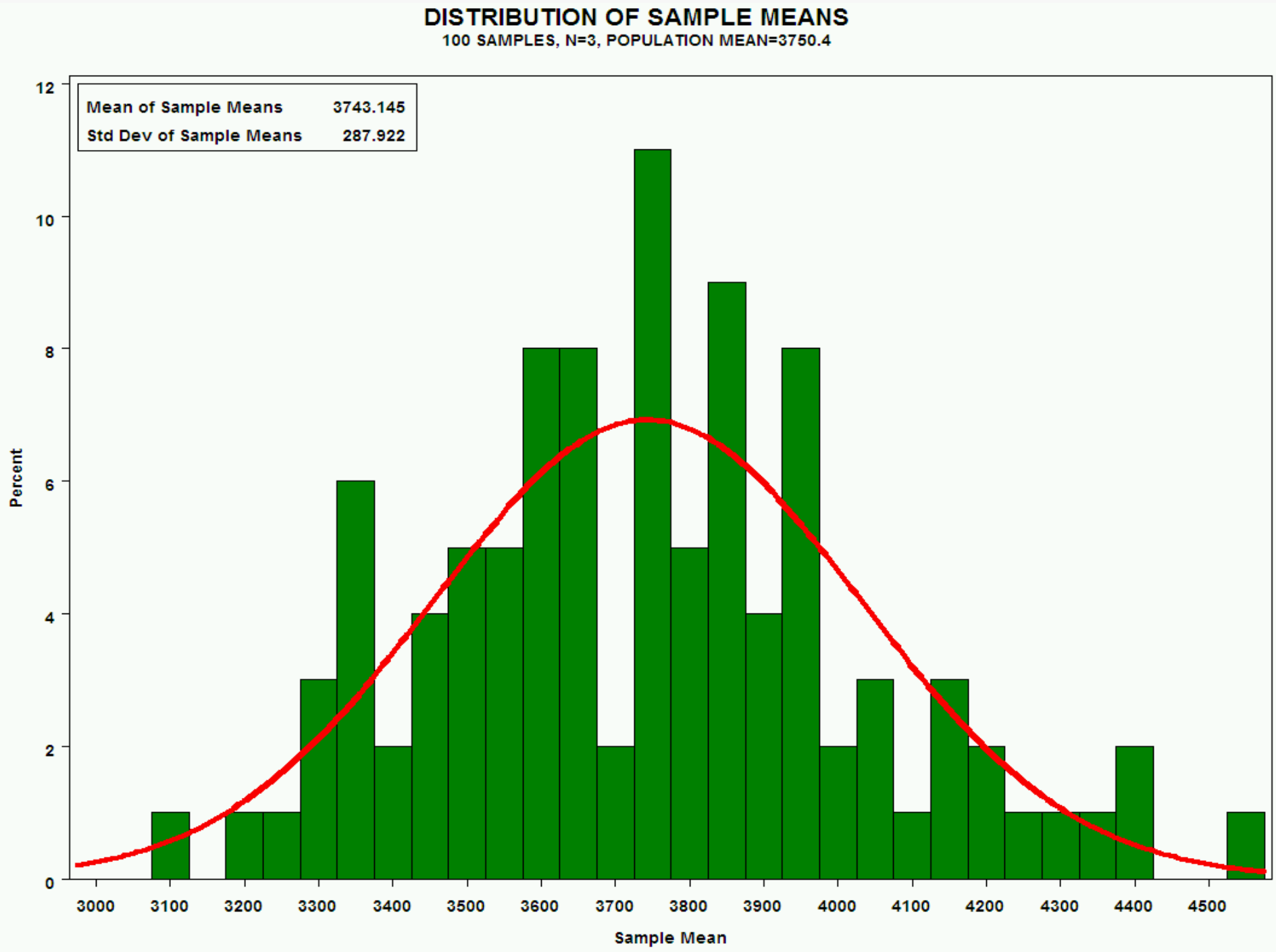
sample from a population of 10,000

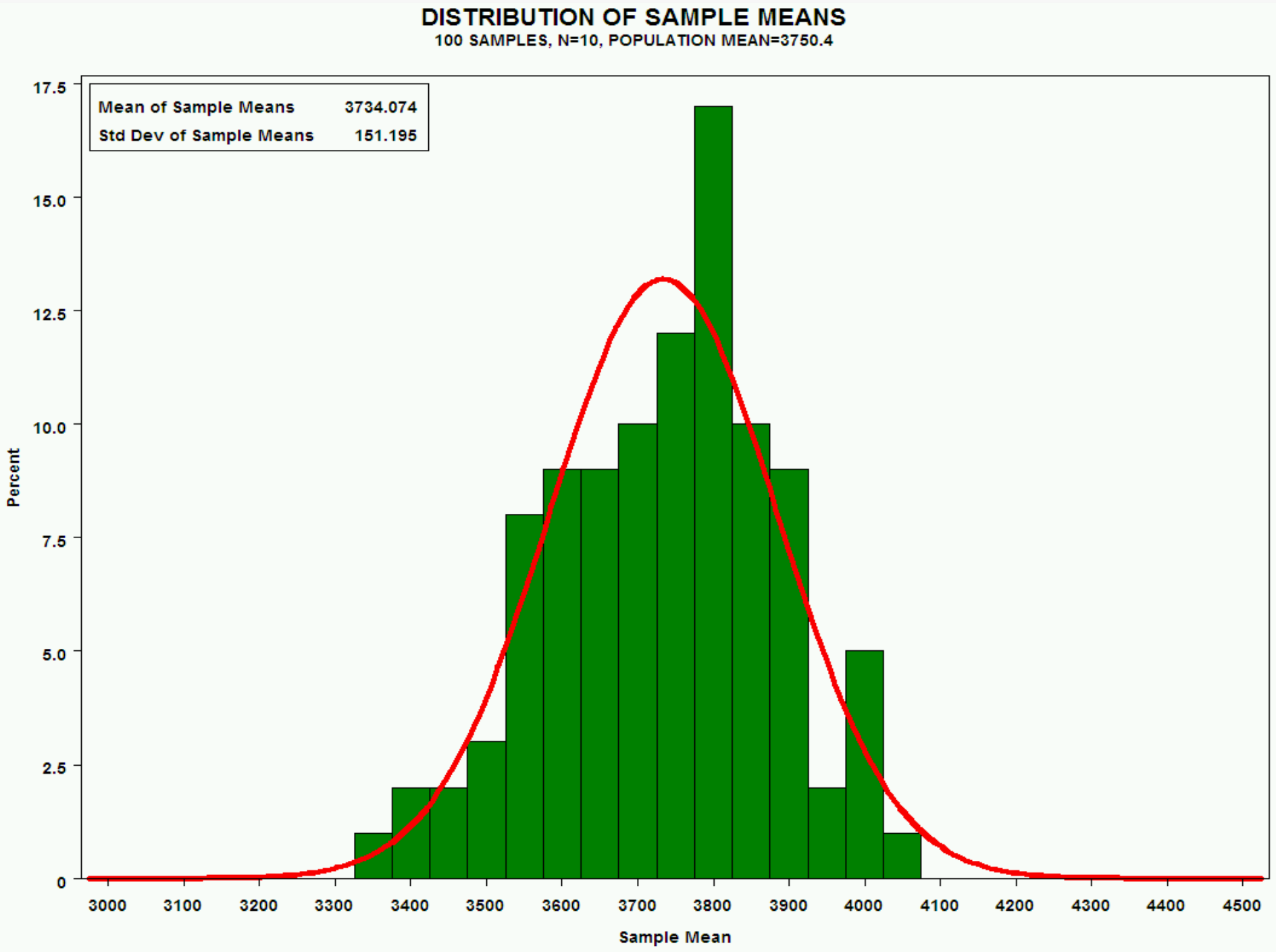
normally distributed with a mean of 3750 and a standard deviation of 500

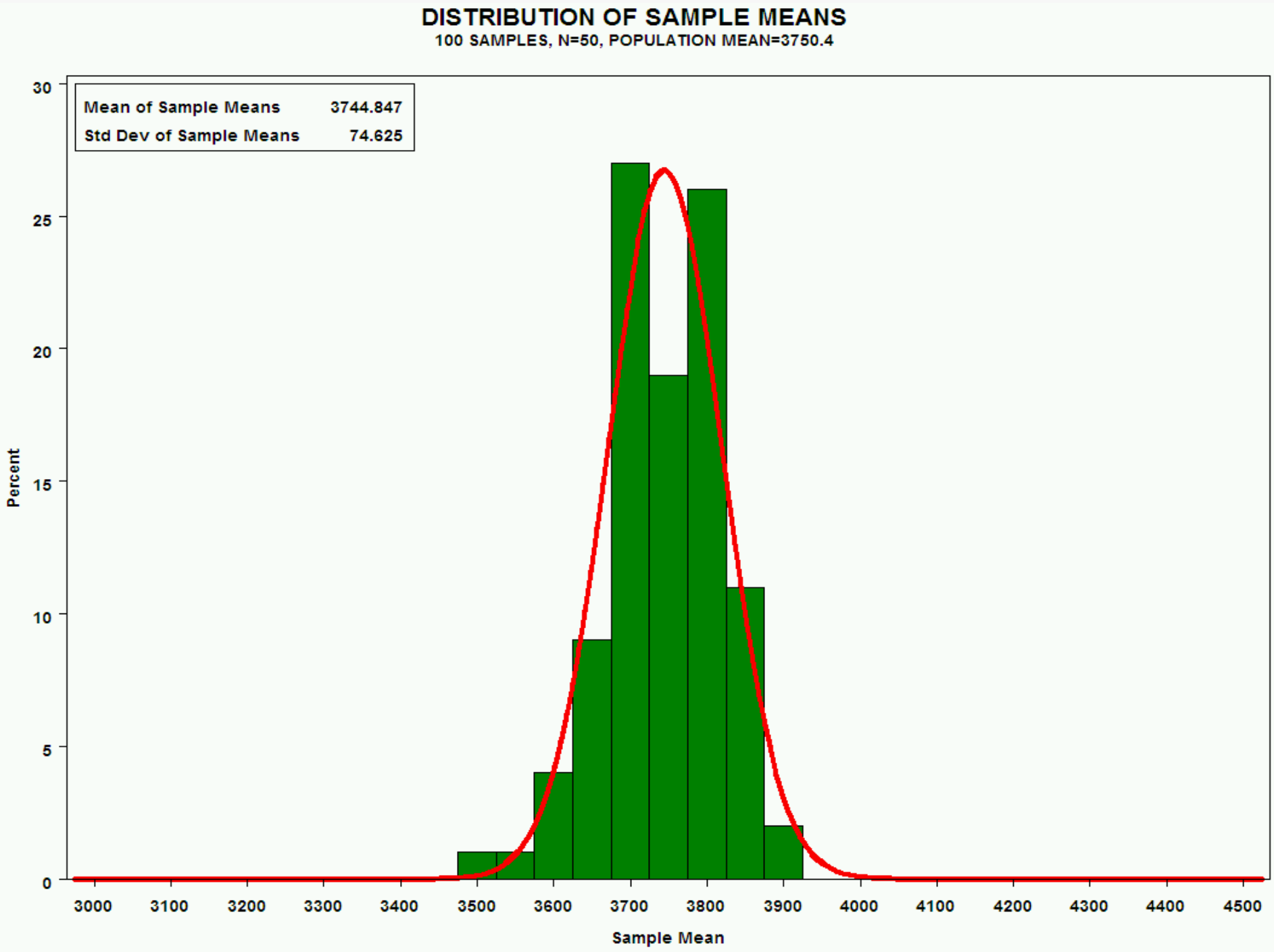


Mean	Std Dev	Variance	Minimum	Maximum
3750	501	251729	1831	5728

- use SAS to take a series of 100 simple random samples ... three different sizes ... N=3, N=10, N=50
- what is the distribution of sample means


DISTRIBUTION OF SAMPLE MEANS
100 SAMPLES, N=3, POPULATION MEAN=3750.4
Mean of Sample Means 3743.145
Std Dev of Sample Means 287.922
Percent
0
2
4
6
8
10
12
3000
3100
3200
3300
3400
3500
3600
3700
3800
3900
4000
4100
4200
4300
4400
4500
Sample Mean


DISTRIBUTION OF SAMPLE MEANS
100 SAMPLES, N=10, POPULATION MEAN=3750.4
Mean of Sample Means 3734.074
Std Dev of Sample Means 151.195
Percent
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0
3000
3100
3200
3300
3400
3500
3600
3700
3800
3900
4000
4100
4200
4300
4400
4500
Sample Mean


DISTRIBUTION OF SAMPLE MEANS
100 SAMPLES, N=50, POPULATION MEAN=3750.4
Mean of Sample Means 3744.847
Std Dev of Sample Means 74.625
Percent
0
5
10
15
20
25
30
3000
3100
3200
3300
3400
3500
3600
3700
3800
3900
4000
4100
4200
4300
4400
4500
Sample Mean

ESTIMATING A POPULATION MEAN

- sample mean best estimate of the population mean (shown earlier as unbiased, minimum variance as compared to other measures of central tendency)

- two scenarios

  population variation (σ) known

    unusual ... hard to think of a situation where you know σ, but do not know μ

  population variation (σ) unknown

    more likely ... no knowledge of σ or μ

POPULATION VARIATION (σ) KNOWN

NORMAL DISTRIBUTION

- all the requirements about sampling are met (reliable, random)
- normality ... or, remember the CENTRAL LIMIT THEOREM

  population is normally distributed
  population is not normally distributed, but sample size is >30

- confidence intervals use the standard error of the mean ($\sigma / \sqrt{N}$)
- Triola term ... margin of error ... E = z * standard error

Triola body temperature example ... given 106 measurements of temperature with a mean of 98.2F and a known population σ of 0.62F, what is a 95% confidence interval for μ

standard error = $\sigma / \sqrt{N} = 0.62 / \sqrt{106} = 0.0602$
for a 95% confidence interval, use ... Z = 1.96
margin of error, E = 1.96 * 0.0602 = 0.1179

mean - E = 98.2 - 0.1179 = 98.08
mean + E = 98.2 + 0.1179 = 98.32
95% confidence interval ... 98.08 < μ < 98.32

what does this mean? what does it say about what is considered a normal temperature, 98.6F?

Rosner example 6.27 … the 1,000 birth weights in table 6.2 are your population … μ=112 oz, σ=20.6 oz (known σ) …

what is the probability that the mean birth weight of a sample of 10 infants will fall between 98 and 126 oz

the range is μ±14 oz, use what you know about z-scores

$$z = \frac{\overline{X} - \mu}{(\sigma / \sqrt{N})} = \frac{112 - 98}{(20.6 / \sqrt{10})} = \frac{14}{6.51} = 2.15$$

from Table 3 in Rosner, z=2.15, column D shows that 0.9684 of the are of the curve is within ± 2.15 z-scores of the mean …

there is a 96.8% chance that the mean birth weight of a sample of 10 infants will fall between 98 and 126 oz … what is the chance the one infant will fall between 98 and 126 (answer is ~50%)

POPULATION VARIATION (σ) UNKNOWN

t DISTRIBUTION

- all the requirements about sampling are met (reliable, random)
- normality ... or, remember the CENTRAL LIMIT THEOREM

  population is normally distributed
  population is not normally distributed, but sample size is >30

- confidence intervals use the standard error of the mean ($\sigma / \sqrt{N}$)
- Triola term ... margin of error ... E = t * standard error

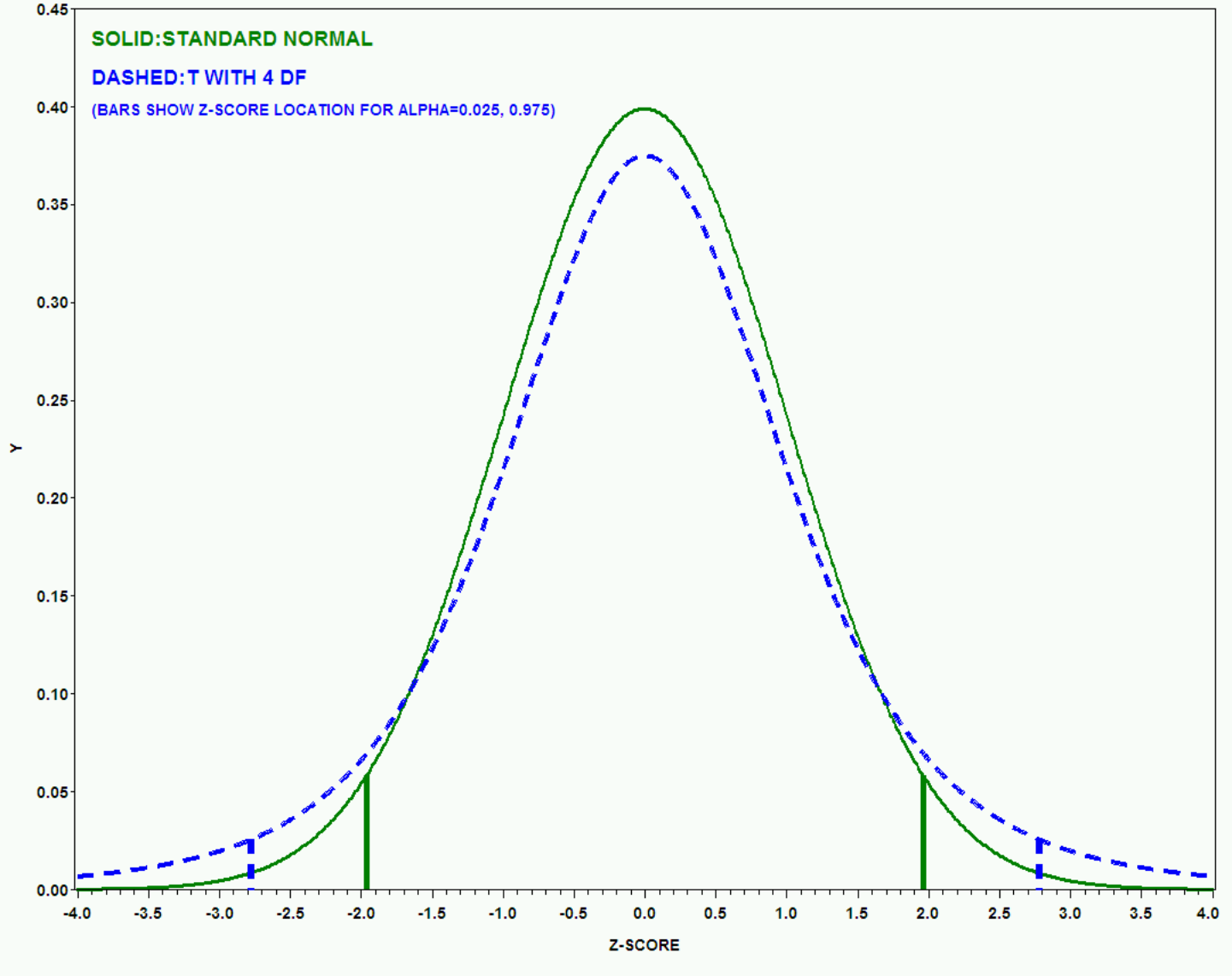
t DISTRIBUTION

- central-limit theorem … sample means are distributed normally with mean μ and standard error, $\sigma / \sqrt{N}$
- 95% of all sample means over a large number of samples of size N will fall between $\mu - 1.96(\sigma / \sqrt{N}), \mu + 1.96(\sigma / \sqrt{N})$
- convert to sample means to z-scores (subtract μ and divide by the standard error)
- assumes that the population standard deviation σ is known

- σ rarely known … use sample data to estimate σ
- z-scores computed using an estimate of the population standard deviation are NOT NORMALLY distributed
- z-scores computed using an estimate of the population standard deviation follow a t-distribution (Student's t) and there are multiple t-distributions that are a function of N, the sample size

the shape of the t distribution is symmetric and similar to that of the standard normal distribution

the shape depends on degrees of freedom (DF)



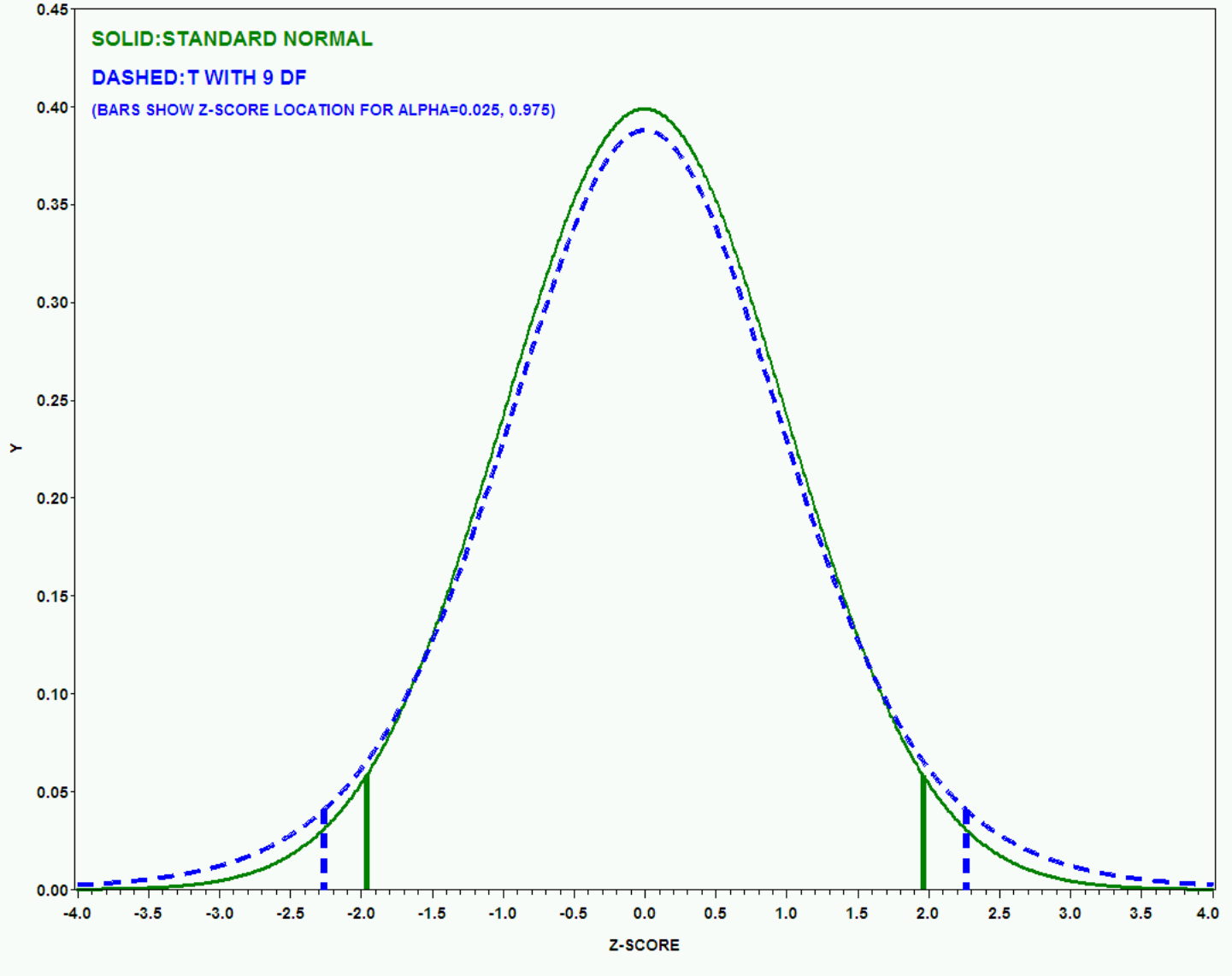
Rosner ... degrees of freedom (DF) ... no definition, just a mention that DF = N -1 when using a t distribution

Triola ... DF for a collection of sample data is the number of sample values that can vary after certain restrictions have been imposed on all data values

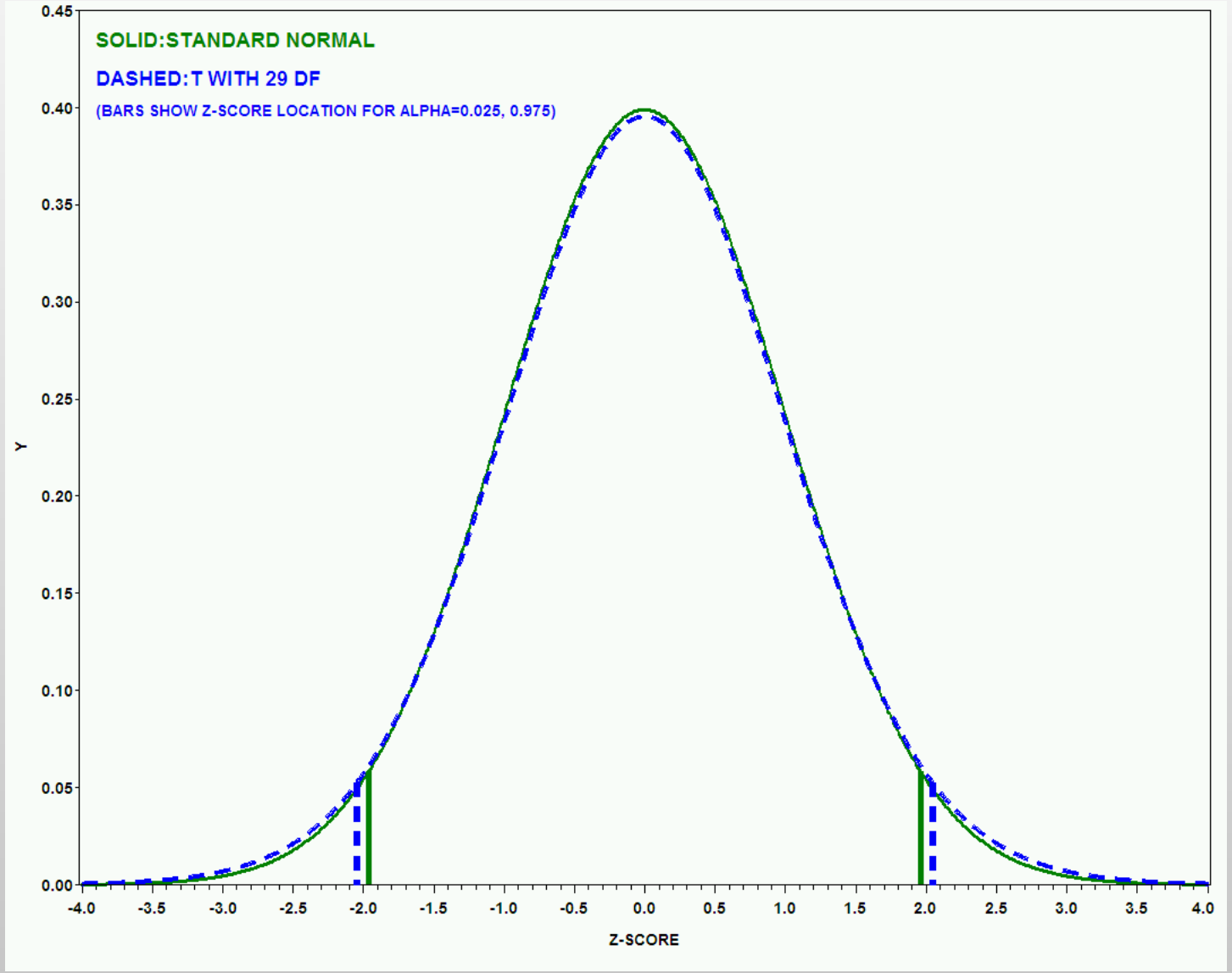
example ... quiz scores of a sample of 10 students have a mean of 80 ... you can assign any scores to 9 of the values, but once 9 are selected, the 10th is fixed since your restriction is that the mean must be 80 (another way to think of this is the sum must be 800 ... if you select 9 numbers at random, the 10th must make the sum 800)

the t distribution is more spread out than the standard normal distribution

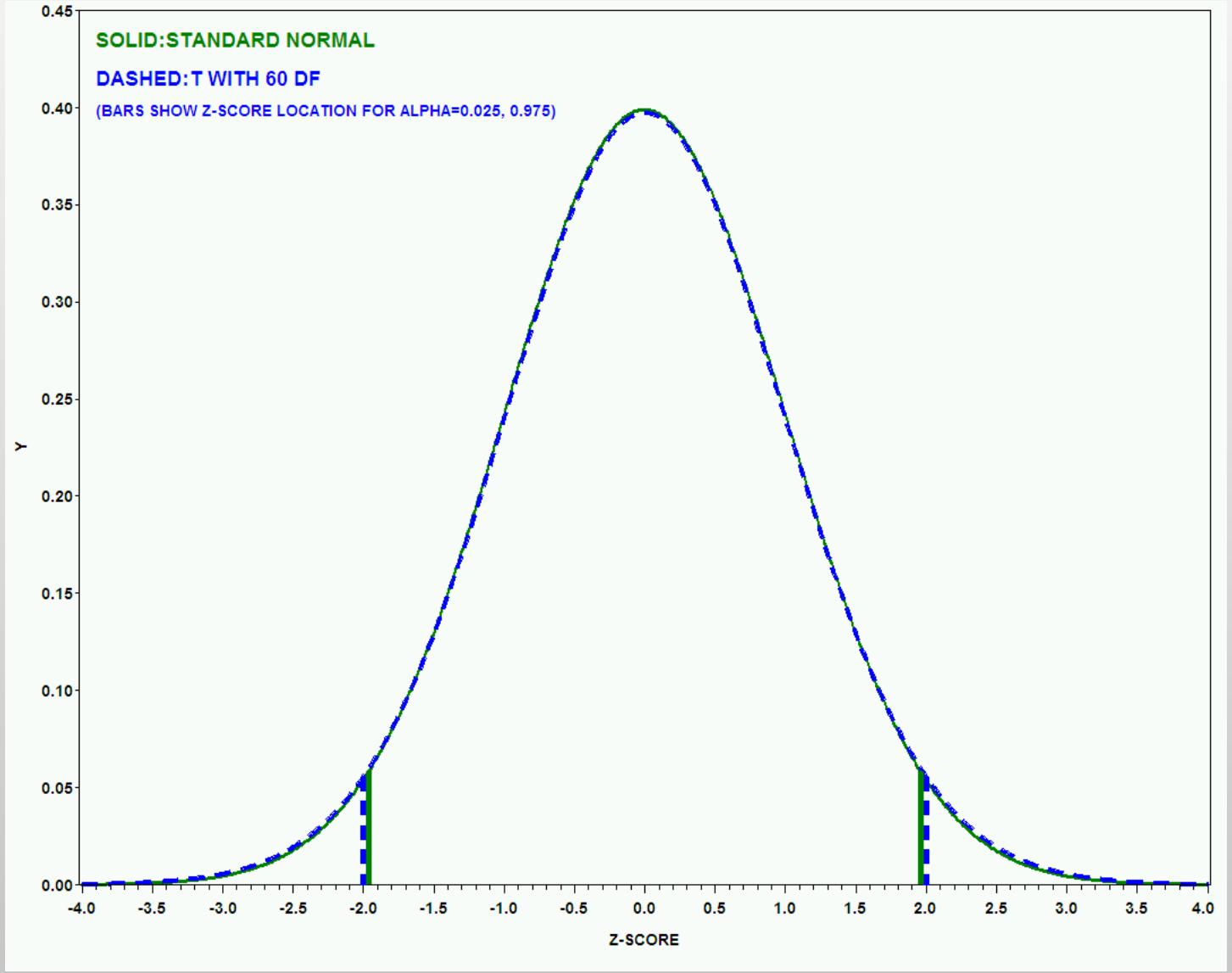
one has to move further to the left (or right) to encompass the same area (probability) as the standard normal curve



as the degrees of freedom increase, the shape of the t distribution more closely resembles that of the standard normal curve



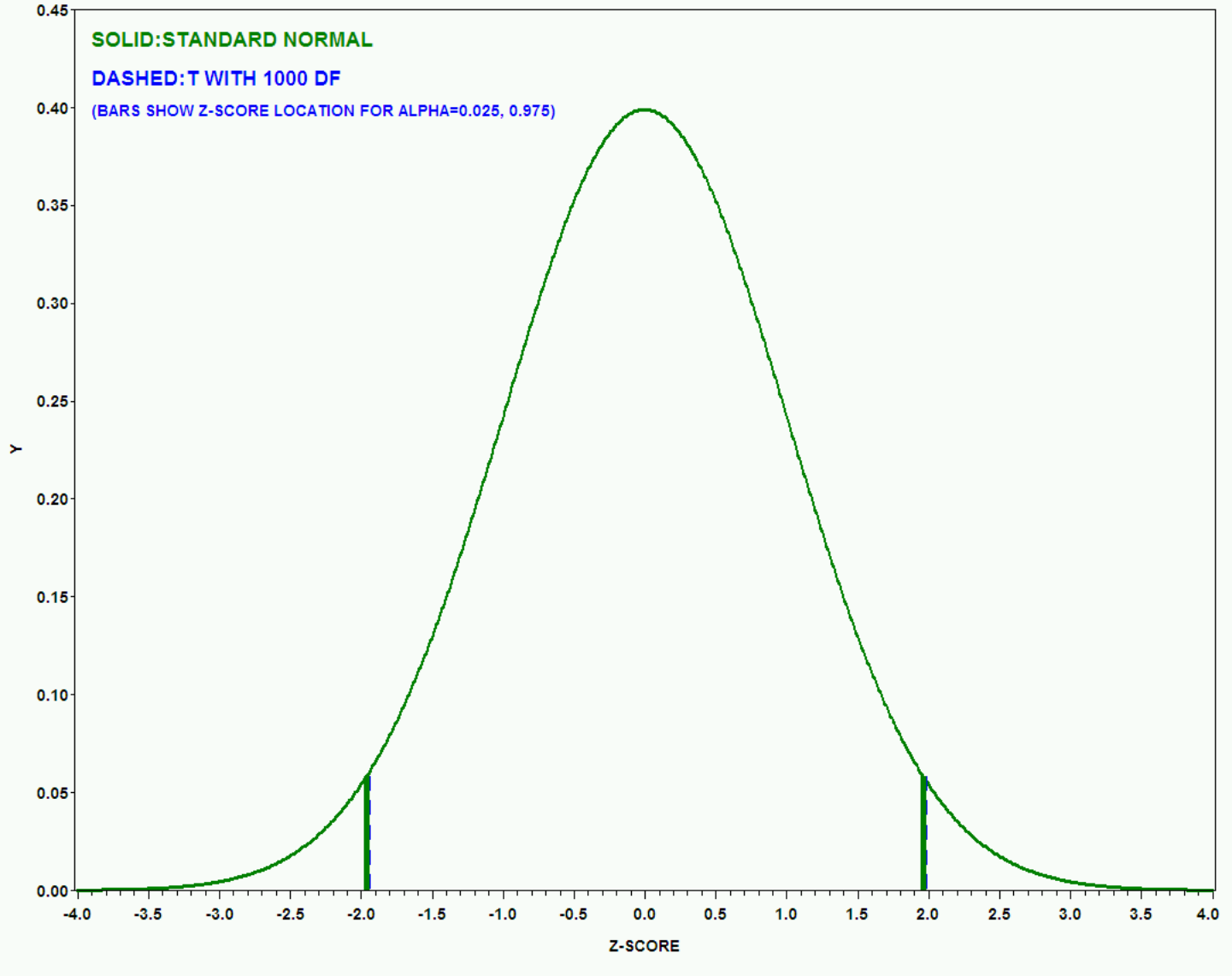
reason ... as DF increase, s (the sample standard deviation) becomes a better estimate of σ (the population standard deviation)



using the standard normal curve, z with p=0.975 is 1.96

using the t distribution, t with p=0.975 varies with DF...

DF	z
4	2.776
9	2.262
29	2.045
60	2.000
∞	1.960



main reason to learn about t distribution … compute a measure of variability for an estimate of the population mean computed from a sample of size N (what type of sample should it be?)

the first measure of variability is the standard error and that value plus a value from a t distribution with N-1 degrees of freedom is used to construct a confidence interval for the mean

Rosner equation 6.6 … confidence interval for the mean of a normal distribution

$$(\bar{x} - t_{n-1,1-\alpha/2} s/\sqrt{n}, \bar{x} + t_{n-1,1-\alpha/2} s/\sqrt{n})$$

or … $$\bar{x} \pm t_{n-1,1-\alpha/2} s/\sqrt{n}$$

earlier … using Rosner data from review question 6B4 … given 5 sample values obtained from a patient (8.5, 9.3, 7.9, 9.2, 10.3)

the confidence interval uses z = ±1.96 from Table 3 (2.5% of area in each tail of the normal distribution) … NOTE: this is NOT CORRECT since you are estimating the standard deviation from your data, it is just for illustration

One sample Z statistics with data

Options

95% confidence interval results:

μ : mean of Variable

Std. Dev. not specified

Variable	n	Sample Mean	Std. Err.	L. Limit	U. Limit
var1	5	9.04	0.40447497	8.247244	9.832756

the confidence interval uses t = ±2.776 from Table 5 (2.5% of area in each tail of the normal distribution) … NOTE: this is CORRECT since you are estimating the standard deviation from your data … it is WIDER

One sample T statistics with data

Options

95% confidence interval results:

μ : mean of Variable

Variable	Sample Mean	Std. Err.	DF	L. Limit	U. Limit
var1	9.04	0.40447497	4	7.9169974	10.163003

95% confidence interval 9.04±2.776(0.4045) (7.917, 10.163)

Rosner equation 6.9 … what factors affect the length (maybe width is a better term) of a confidence interval

- N as N increases, standard error decreases and the value from the t distribution decreases --- width decreases
- s as standard deviation decreases (less variability in the data), standard error decreases --- width decreases
- α as α decreases (desired confidence increases) --- width increases

Triola body temperature example ... given 106 measurements of temperature with a mean of 98.2F and an sample estimate (s) of the population σ, what is a 95% confidence interval for μ

standard error = $s / \sqrt{N} = 0.62 / \sqrt{106} = 0.0602$
for a 95% confidence interval, use ... 105df, t = 1.984
margin of error, E = 1.984 * 0.0602 = 0.1195

mean - E = 98.2 - 0.1195 = 98.08
mean + E = 98.2 + 0.1195 = 98.32
95% confidence interval ... 98.08 < μ < 98.32

yes, it looks the same as the known σ example ... but that is ONLY due to rounding

Triola corn example ... given 11 estimates of corn yield (pounds per acre), construct a 95% confidence interval estimate of the mean yield...

1903 1935 1910 2496 2108 1961 2060 1444 1612 1316 1511

mean = 1841.5, standard deviation = 342.7
standard error = $342.7 / \sqrt{11}$ = 103.328
95% confidence interval with 10 DF, t = 2.228
margin of error, E = 2.228 * 103.328 = 230.215
mean - E = 1841.5 - 230.215 = 1611.3
mean + E = 1841.5 + 230.215 = 2071.7

95% confidence interval ... $1611.3 < \mu < 2071.7$

WHAT DOES THIS MEAN? WHAT IS A CONFIDENCE INTERVAL?

Rosner ... equation 6.8 ... over the collection of all 95% confidence intervals that could be constructed from repeated random samples of size n, 95% will contain the parameter μ

Triola (paraphrasing) ... if we were to conduct many different experiments and construct the 95% confidence interval for the results of each of them, 95% of them would actually contain the population value

Daniel (paraphrasing)... probabilistic ... in repeated sampling, 95% of all 95% confidence intervals will, in the long run, contain the population mean μ

Triola ... it is wrong to say that there is a 95% chance that the true population value falls within the particular 95% confidence limits

Rosner ... we cannot say there is a 95% chance that the parameter μ will fall within a particular 95% confidence interval

Daniel (paraphrasing)... practical probabilistic ... we are 95% confident that the single computed 95% confidence interval contains the population mean μ

following ... 100 samples taken from a normally distributed population with a mean of 3750 ... N=3, N=10, N=50

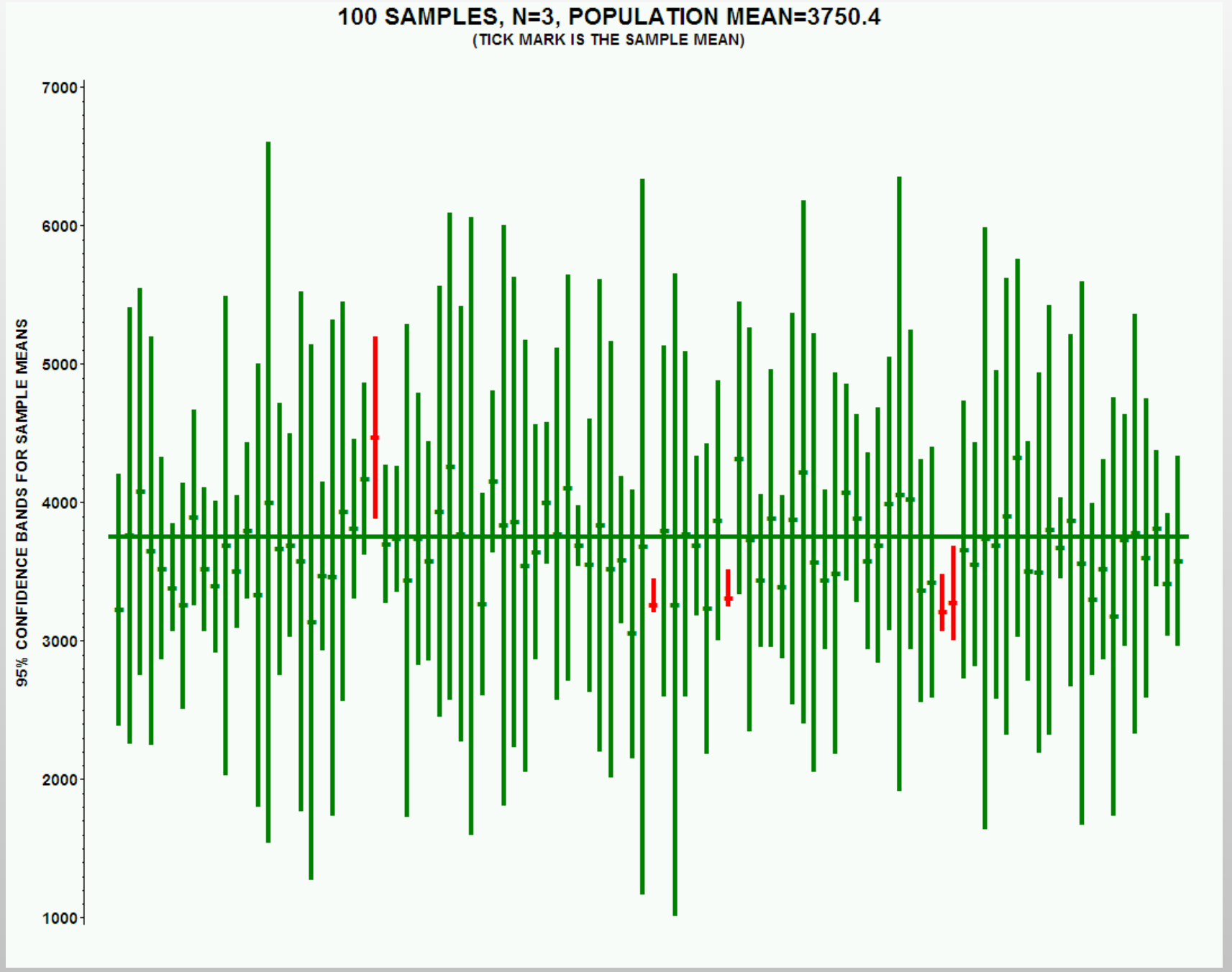
compute the mean of each sample and a 95% confidence band

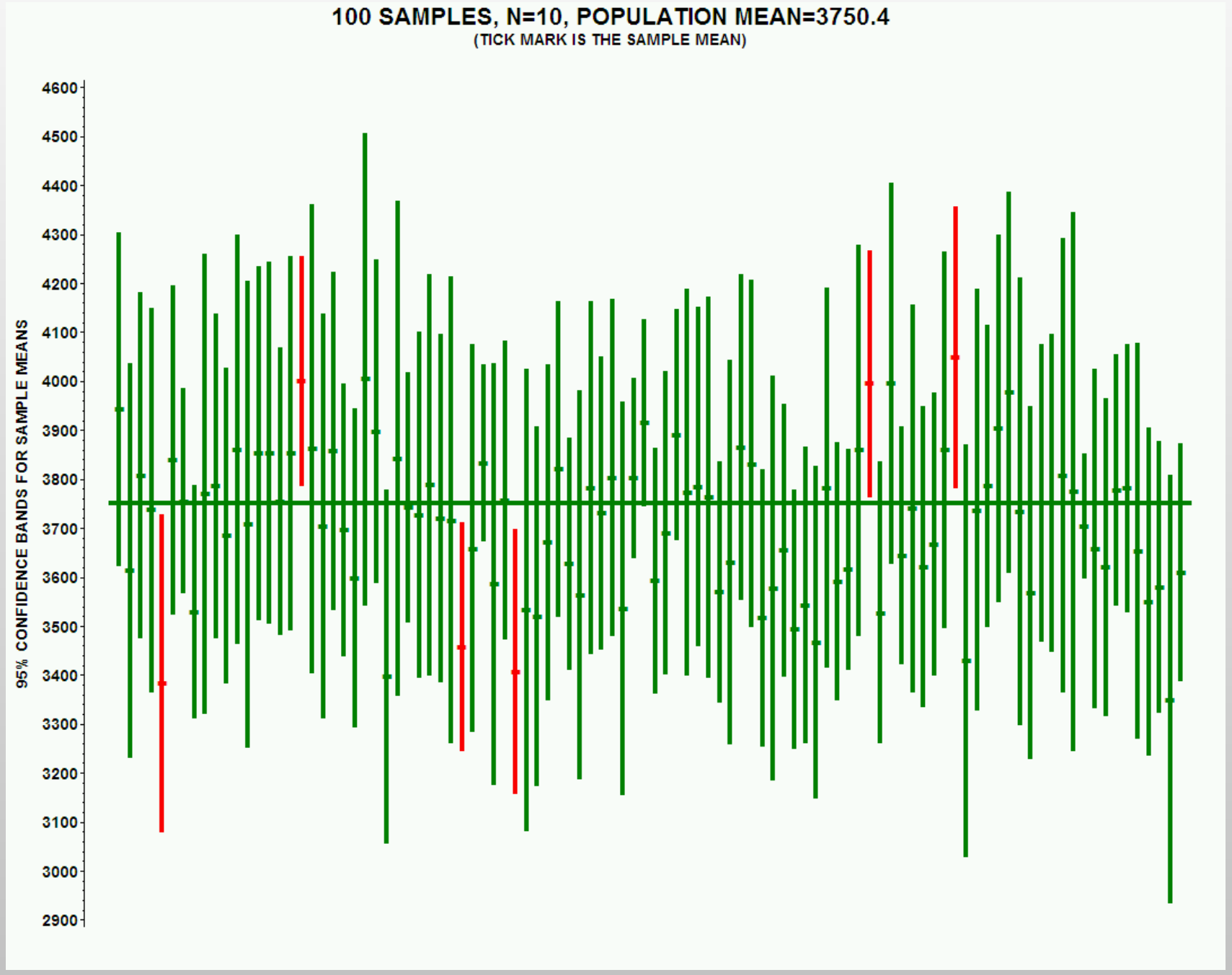
plot the results ... 100 samples in each plot ... out of the 100 samples, how often does the 95% confidence band contain the population mean (look for RED results)?

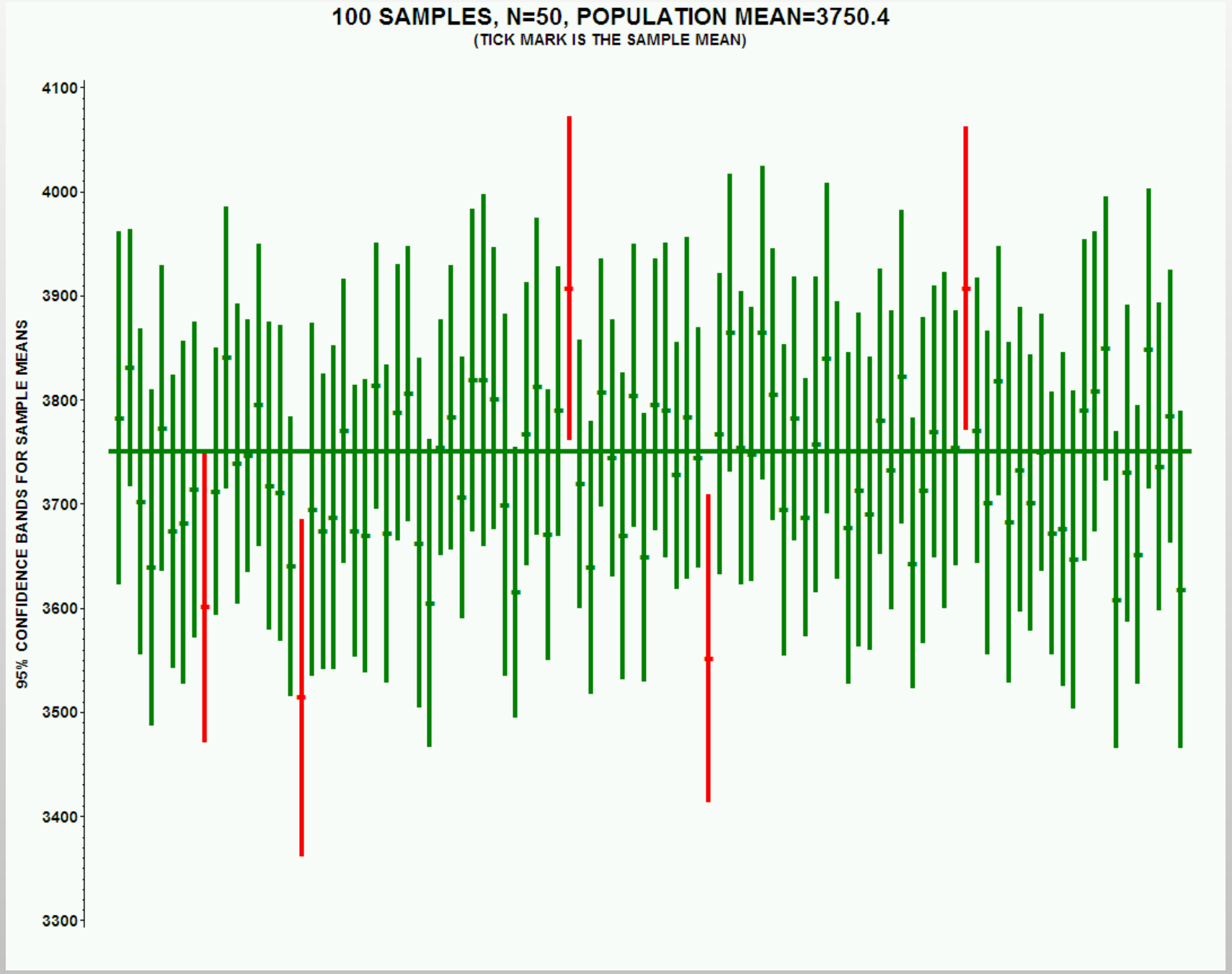
in the plots ... look at the left axis ... as N increase, size of confidence bands decreases ... do you know why?

REMEMBER ... you only take one sample ... notice the variability in the location of the sample means as N increase

(similar to Rosner figure 6.7 based on 5 samples from data in table 6.2, 1000 birth weights with μ=112)


100 SAMPLES, N=3, POPULATION MEAN=3750.4
(TICK MARK IS THE SAMPLE MEAN)
95% CONFIDENCE BANDS FOR SAMPLE MEANS
7000
6000
5000
4000
3000
2000
1000


100 SAMPLES, N=10, POPULATION MEAN=3750.4
(TICK MARK IS THE SAMPLE MEAN)
95% CONFIDENCE BANDS FOR SAMPLE MEANS
4600
4500
4400
4300
4200
4100
4000
3900
3800
3700
3600
3500
3400
3300
3200
3100
3000
2900


100 SAMPLES, N=50, POPULATION MEAN=3750.4
(TICK MARK IS THE SAMPLE MEAN)
95% CONFIDENCE BANDS FOR SAMPLE MEANS
4100
4000
3900
3800
3700
3600
3500
3400
3300

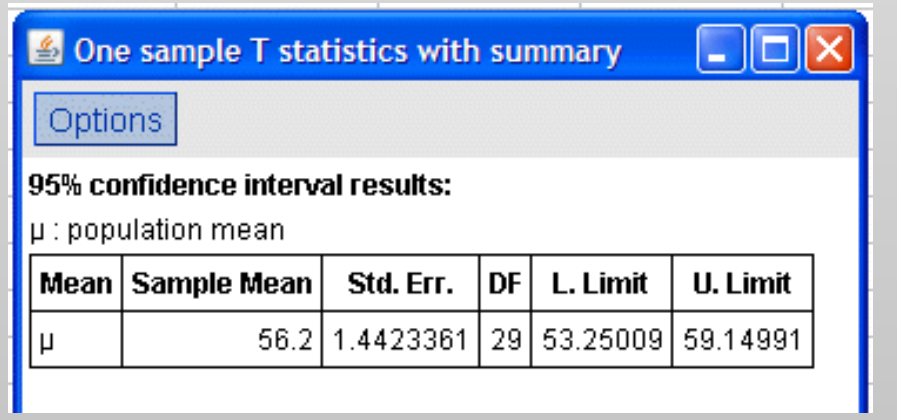
another use for confidence intervals ...

is the mean of a population (estimated from a sample) different from a 'standard' ... assigned problem 6.31 ... sample of 30 children age 5-6 in a specific community

based on the sample, mean SBP=56.2, standard deviation=7.9

is the community different from the nationwide average DBP for 5-6 year old children, 64.2

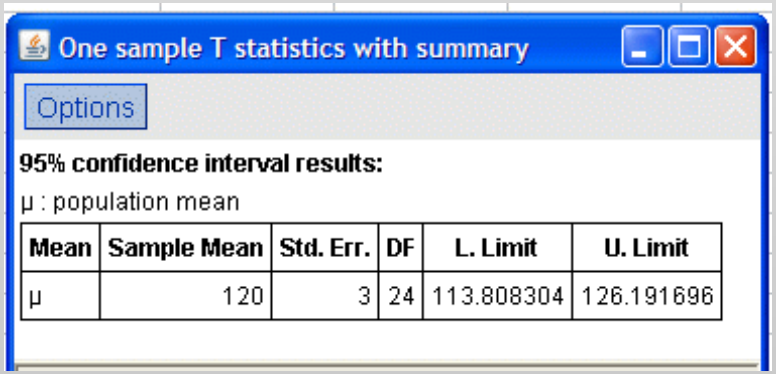
one way to answer ... is 64.2 inside or outside a 95% confidence interval on the sample mean ... it is outside ... conclude that the community is different

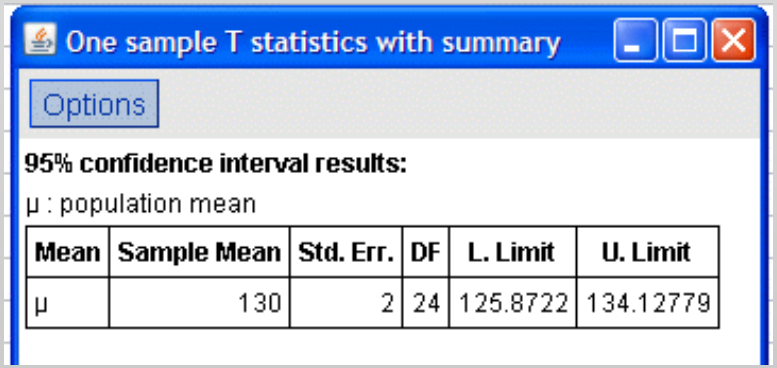


are two means (estimated from samples) different (online course midterm question) … you conduct health exams on samples of 25 men and 25 women … one measurement you make is systolic blood pressure (SBP) and you calculate the following statistics …

gender	mean	standard deviation
women	120	15
men	130	10

do you think that men have a higher mean SBP than women?



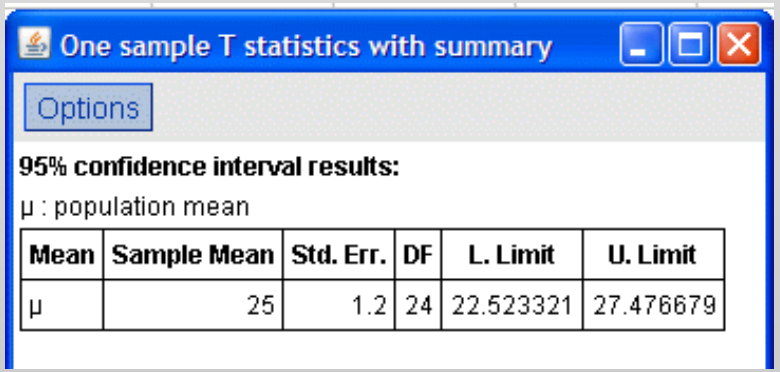


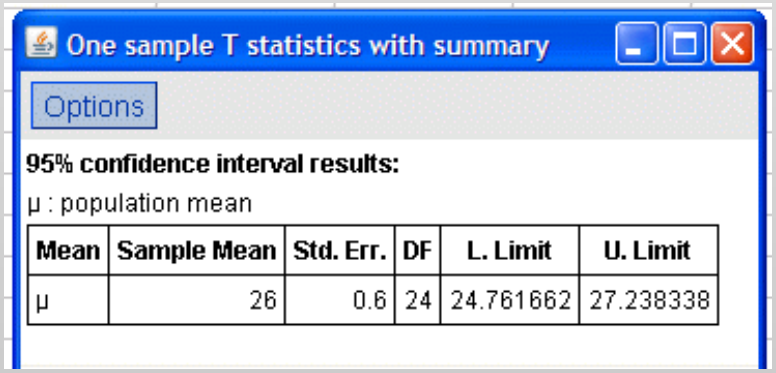
95% confidence intervals overlap … conclude no difference

you also measure body mass index (BMI) for the same 25 men and 25 women and calculate the following statistics...

gender	mean	standard deviation
women	25	6
men	26	3

the upper-limit for a NORMAL BMI for adults is 23 ... based on the above information, can you say that either the group of women or group of men have 'above normal' BMIs?





95% confidence interval for women includes 23 ... no difference
95% confidence interval for men all above 23 ... above normal

ESTIMATING POPULATION VARIABILITY

- point estimate of variance (standard deviation)
- interval estimate of variance (standard deviation) … confidence intervals
- sample size required to estimate variance (standard deviation)

REQUIREMENTS

- same sampling rules apply as when estimating central tendency (mean, proportion) ... random sample, sample represents the population

- normality ... no Central Limit Theorem for estimating the variance ... departures from normality result in "gross" errors

DISTRIBUTIONS

- central tendency ... normal and t-distribution
- variability ... chi-square (χ^2)

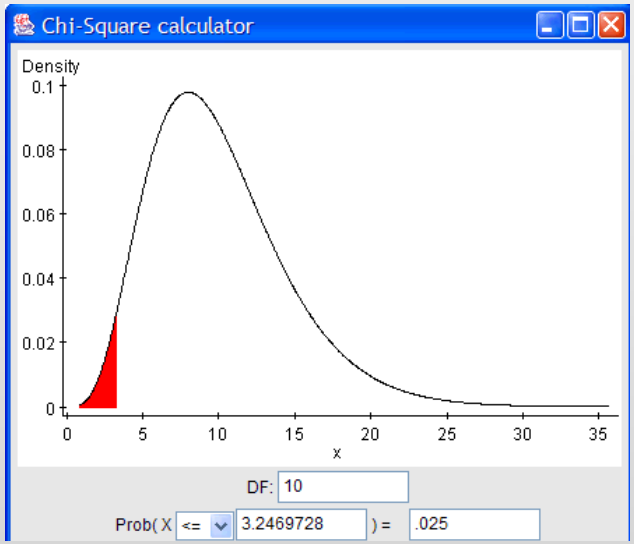
  repeated samples from a normal distribution

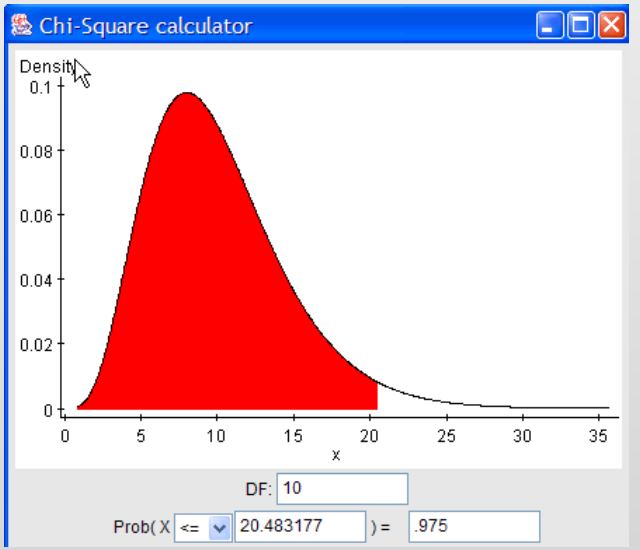
  calculate the variance of each sample

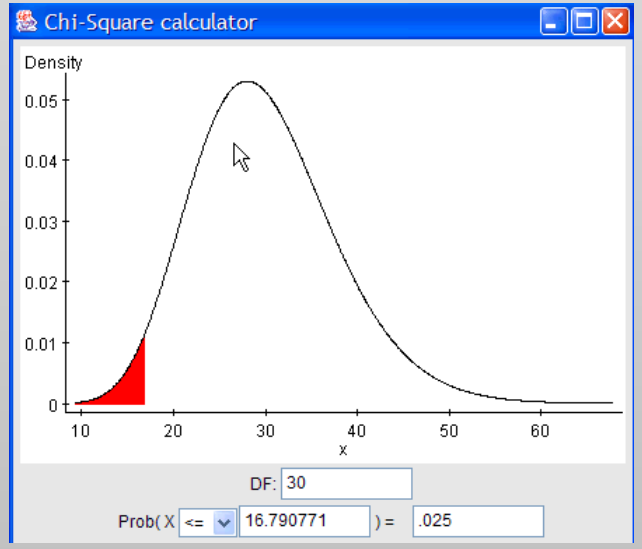
  variance follow a χ^2 distribution

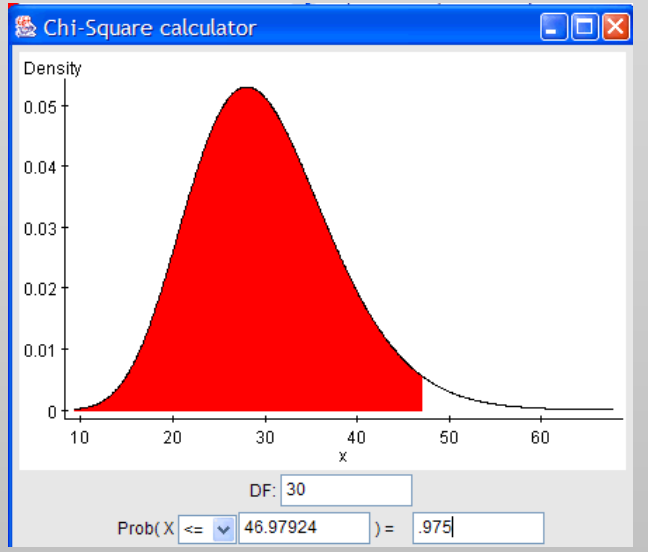
- χ^2 not symmetric, shape varies by degrees of freedom

- values are 0 or positive









- lack of symmetry… cannot use $\sigma^2 \pm$ E (margin of error)
- confidence interval…

$$\left[(n-1)s^2 / \chi^2_{n-1,1-\alpha/2}, (n-1)s^2 / \chi^2_{n-1,\alpha/2}\right]$$

- example in Triola … body temperatures …

n = 106, s = 0.62, s^2 = 0.3844

Table 6 in Rosner, no 105 DF, use 100 DF

(106 - 1) 0.3844 / 129.56 < σ^2 < (106 - 1) 0.3844 / 74.22

0.31 < σ^2 < 0.54 0.56 < σ < 0.74

ESTIMATING A POPULATION PROPORTION

- sample proportion used to estimate the population proportion (a point estimate, central tendency)
- normal approximation to the binomial involved in estimating precision of the point estimate
- construct a confidence interval based (0.90, 0.95, 0.99)
- all requirements for using the normal approximation to the binomial are met (binomial: fixed number of trial, independent trials, two outcomes, P constant over trials, and NPQ $\geq$5)

- confidence intervals use the standard error of a proportion ...

$$\sqrt{\frac{pq}{n}}$$

- Triola margin of error ... E = z * standard error

Triola genetics example…

given 580 offspring peas and 26.2% (N=152) with yellow pods

based on the above, what should one conclude about Mendel's theory that 25% of peas will have yellow pods

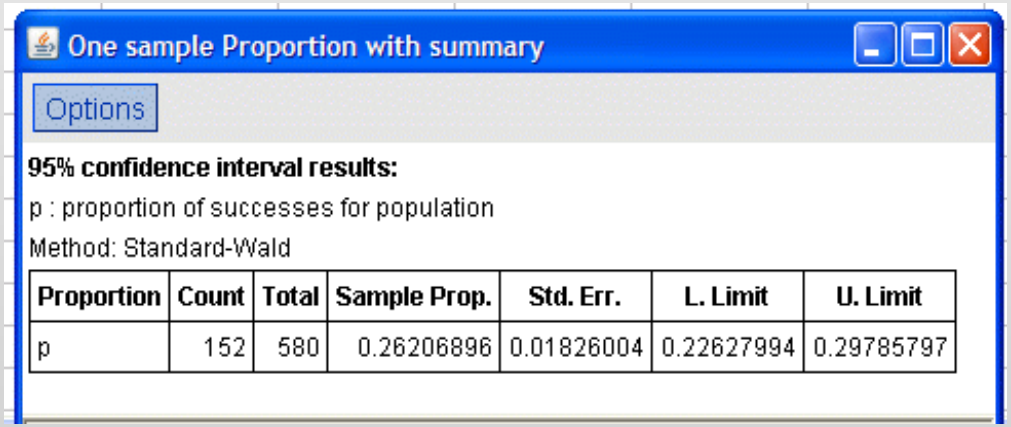
fixed number of trial = 580
independent trials
two outcomes (yellow/not yellow)
P constant over trials
NPQ = 580(0.262)(0.738) = 112

standard error = $\sqrt{\frac{pq}{n}} = \sqrt{\frac{0.262(0.738)}{580}} = 0.01826$

95% confidence interval, Z = 1.96
margin of error, E = 1.96 * 0.01826 = 0.03579
P - E = 0.262 - 0.03579 = 0.226
P+ E = 0.262 + 0.03570 = 0.298

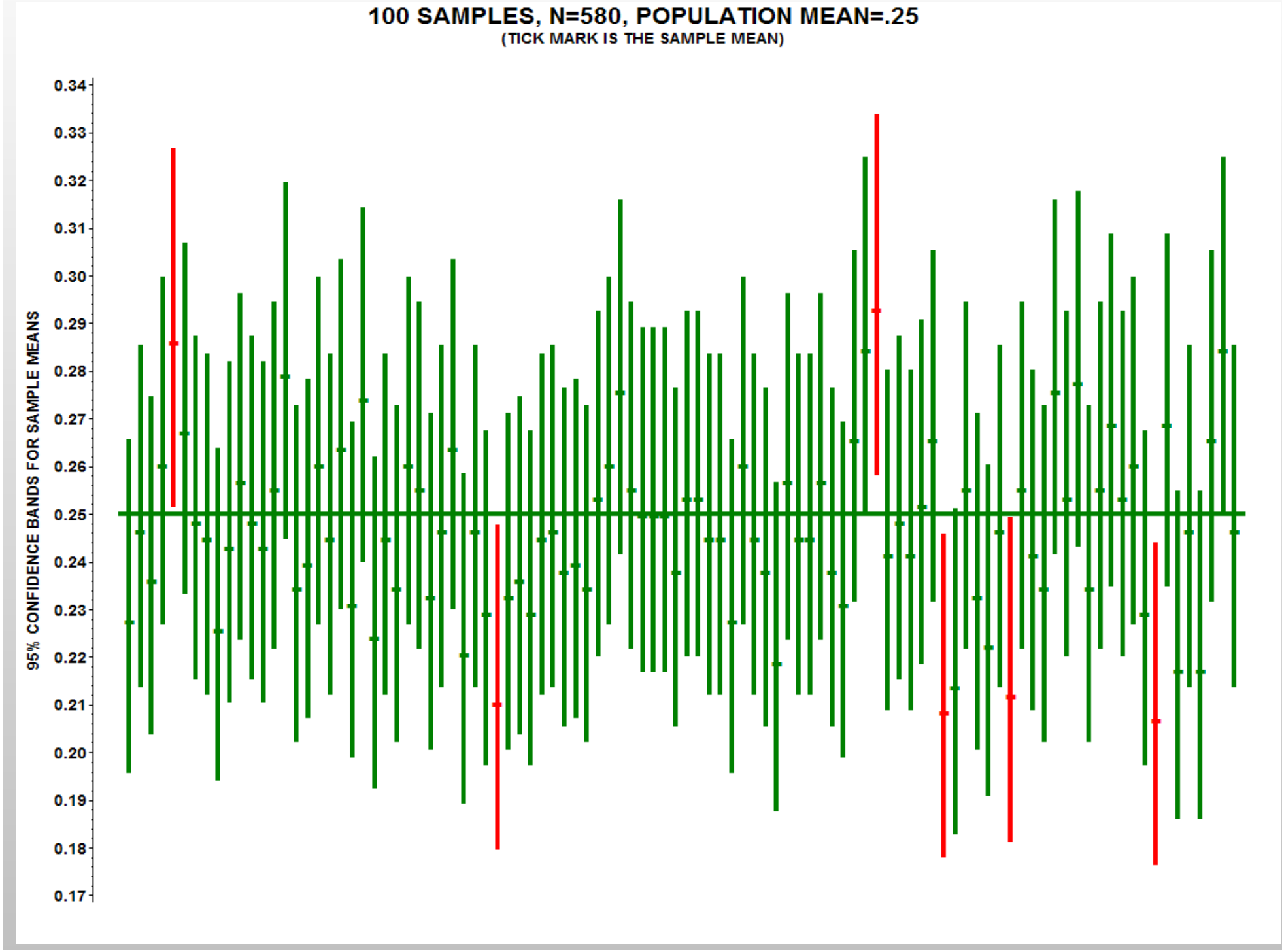
95% confidence interval …
0.226 < P < 0.298

WHAT DOES THIS MEAN?
WHAT ABOUT MENDEL?



Proportion	Count	Total	Sample Prop.	Std. Err.	L. Limit	U. Limit
p	152	580	0.26206896	0.01826004	0.22627994	0.29785797

following … confidence intervals have the same interpretation for proportions as they do for means … 100 samples with N=580, construct 95% confidence intervals


100 SAMPLES, N=580, POPULATION MEAN=.25
(TICK MARK IS THE SAMPLE MEAN)
95% CONFIDENCE BANDS FOR SAMPLE MEANS
0.34
0.33
0.32
0.31
0.30
0.29
0.28
0.27
0.26
0.25
0.24
0.23
0.22
0.21
0.20
0.19
0.18
0.17

- critical values for z in determining confidence intervals ...

  two-sided confidence intervals

  equal area in tails of normal distribution

90%	5% of area in each tail	z =1.645
95%	2.5% of area in each tail	z= 1.96
99%	0.5% of area in each tail	z = 2.575

EXACT BINOMIAL INTERVALS

Rosner … example 6.50 … what is the rate of bladder cancer in mice fed a diet that is high in saccharin … 20 rats … 2 develop bladder cancer … npq = 1.8 < 5 … cannot use the normal approximation method

Rosner suggests using curves in Table 7a or a trial-and-error approach with Excel … use Statcrunch … below/left uses normal approximation … below right are exact (close to Rosner values of 0.01 to 0.32 … not symmetric)

One sample Proportion with summary

Options

95% confidence interval results:

p : proportion of successes for population

Method: Standard-Wald

Proportion	Count	Total	Sample Prop.	Std. Err.	L. Limit	U. Limit
p	2	20	0.1	0.06708204	-0.031478383	0.23147838

One sample Proportion with summary

Options

95% confidence interval results:

p : proportion of successes for population

Method: Agresti-Coull

Proportion	Count	Total	Sample Prop.	Std. Err.	L. Limit	U. Limit
p	2	20	0.1	0.06708204	0.015656238	0.3132439

what happens as you increase sample size … leave estimate of proportion at 0.10 … approximation on left, exact on right with n=200 (npq=18) and n=2000 (npq=180) …

One sample Proportion with summary

Options

95% confidence interval results:

p : proportion of successes for population

Method: Standard-Wald

Proportion	Count	Total	Sample Prop.	Std. Err.	L. Limit	U. Limit
p	20	200	0.1	0.021213204	0.058422886	0.14157711

One sample Proportion with summary

Options

95% confidence interval results:

p : proportion of successes for population

Method: Agresti-Coull

Proportion	Count	Total	Sample Prop.	Std. Err.	L. Limit	U. Limit
p	20	200	0.1	0.021213204	0.06500984	0.15006642

One sample Proportion with summary

Options

95% confidence interval results:

p : proportion of successes for population

Method: Standard-Wald

Proportion	Count	Total	Sample Prop.	Std. Err.	L. Limit	U. Limit
p	200	2000	0.1	0.006708204	0.08685216	0.11314784

One sample Proportion with summary

Options

95% confidence interval results:

p : proportion of successes for population

Method: Agresti-Coull

Proportion	Count	Total	Sample Prop.	Std. Err.	L. Limit	U. Limit
p	200	2000	0.1	0.006708204	0.08758694	0.1139467

as n increases, difference between normal approximation and exact limits decreases

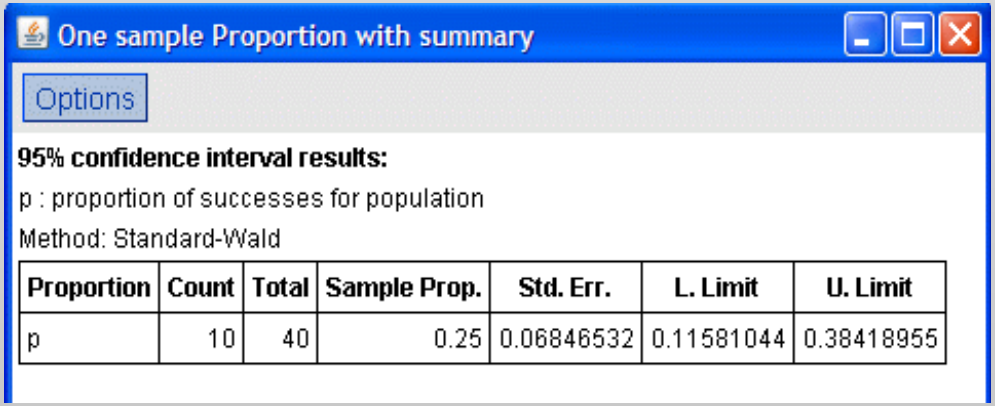
Rosner … review questions 6E … 10 out of 40 participants in a weight loss program are partially successful … what is an estimate of the partial success rate … what is a 95% confidence interval

best estimate of rate = 10/40 = 0.25

npq = 40(0.25) (0.75) = 7.5 > 5, use normal approximation

margin of error … $E = z * \sqrt{\frac{pq}{n}} = 1.96 * \sqrt{\frac{0.25(0.75)}{40}} = 0.134$

95% interval (0.116, 0.384)

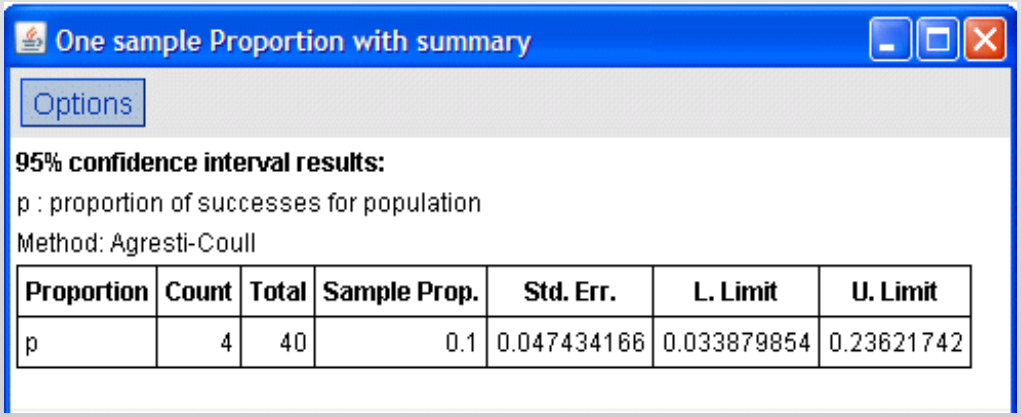


Proportion	Count	Total	Sample Prop.	Std. Err.	L. Limit	U. Limit
p	10	40	0.25	0.06846532	0.11581044	0.38418955

Rosner … review questions 6E … 4 out of 40 participants in a weight loss program are partially successful … what is an estimate of the partial success rate … what is a 95% confidence interval

best estimate of rate = 4/40 = 0.10

npq = 40(0.1) (0.9) = 3.6 < 5, use exact method



Proportion	Count	Total	Sample Prop.	Std. Err.	L. Limit	U. Limit
p	4	40	0.1	0.047434166	0.033879854	0.23621742

note … try curves in Table 7a … find 0.1 on bottom axis, read up to two curves with n=40 … read interval on left axis

SAMPLE SIZE TO ESTIMATE A POPULATION PROPORTION

- how large should a sample be to have a pre-selected margin of error
- margin of error... E = z * standard error
- standard error ... $\sqrt{\frac{pq}{n}}$
- margin of error ... E = z * $\sqrt{\frac{pq}{n}}$
- solve for n ... $n = z^2 * pq / E^2$
- sample size is a function of z (confidence level), p and q (estimate of proportion in population), E (margin of error)

EXAMPLE FROM TRIOLA

- what proportion of US households use e-mail (95% confident that the estimate is in error by no more than 4%)

- $n = z^2 * pq / E^2$

- two scenarios...

  previous study estimated p = 0.169
  $n = 1.96^2 * (0.169 * 0.831) / 0.04^2 = 337.194 = 338$

  no previous estimate of p
  $n = 1.96^2 * (0.5 * 0.5) / 0.04^2 = 600.25 = 601$

- how does sample size vary with margin of error…assuming no prior knowledge of the population proportion...

	SAMPLE SIZE	
E	95%	99%
0.005	38,416	66,307
0.010	9,604	16,577
0.015	4,269	7,368
0.020	2,401	4,145
0.025	1,537	2,653
0.030	1,068	1,842
0.035	784	1,354
0.040	601	1,037
0.045	475	819
0.050	385	664

- how does sample size vary with the assumed population proportion and a given margin of error…assuming margin of error is 3%…

	SAMPLE SIZE	
P	95%	99%
0.1	385	664
0.2	683	1,179
0.3	897	1,548
0.4	1,025	1,769
0.5	1,068	1,842
0.6	1,025	1,769
0.7	897	1,548
0.8	683	1,179
0.9	385	664

BOTH TABLES ASSUME A RANDOM SAMPLE THAT IS REPRESENTATIVE OF THE POPULATION --- A LARGE SAMPLE CANNOT FIX BAD SAMPLING

SAMPLE SIZE TO ESTIMATE A POPULATION MEAN

- how large should a sample be to have a pre-selected margin of error
- margin of error… E = z * standard error
- standard error… $\sqrt{\frac{\sigma}{n}}$
- margin of error… E = z * $\sqrt{\frac{\sigma}{n}}$
- solve for n… $n = (z * \sigma / E)^2$
- sample size is a function of z (confidence level), σ (variability in the population), E (margin of error)

EXAMPLE FROM TRIOLA

- sample size to estimate the mean IQ score of statistics professors with 95% confidence that the sample mean is within 2 IQ points of the population mean

  $z = 1.96$ (95% confidence)

  $\sigma = 15$ (IQ test designed to have: $\mu = 100$, $\sigma = 15$)

  $E = 2$ IQ points

  $n = (1.96 * 15 / 2)^2$

  $n = 216.09 = 217$

  (easy to see the effect of changing E)