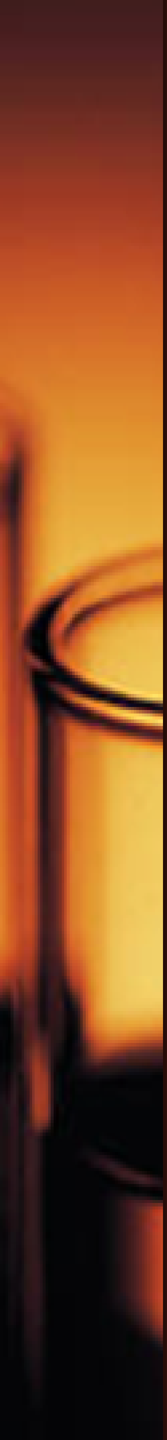
Biostatistics

Week #9 4/19/2022

Ch 8 - Sampling Distribution of the Mean



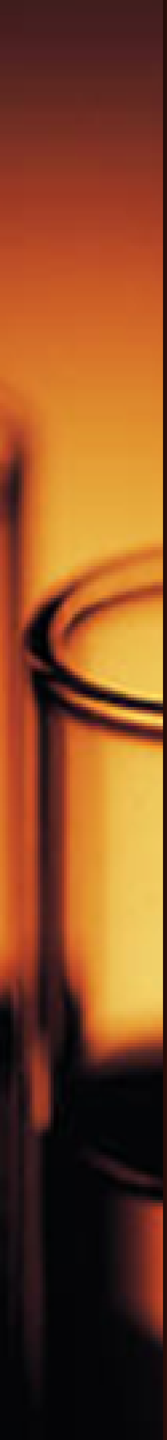
Chapter 8

- 8.1 Sampling Distribution
- 8.2 The Central Limit Theorem
- 8.3 More Examples

8.1 Sampling Distributions

Introduction

- In previous chapter we have seen a number of ***theoretical*** distributions, in which we supply (via a large-size observations) certain parameters (such as n and p in binomial, λ=np in Poisson, or μ and σ in normal distribution) to fully describe such distribution.
- ***In reality***, when large-scale observation is not possible, one must resort to the process called ***sampling* (取樣)** to get these needed parameters.



Sampling (wiki)

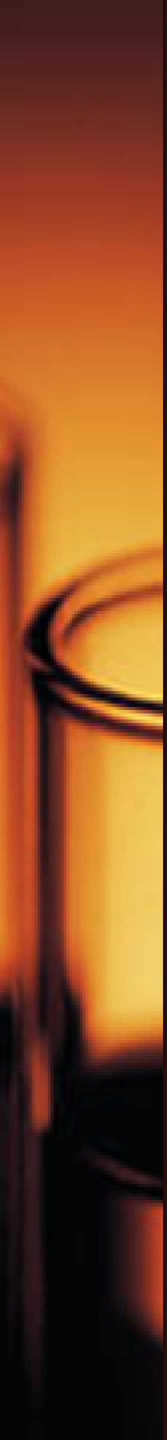
- **Sampling** is the part of statistical practice concerned with **the selection of individual observations** intended to yield some knowledge about a population of concern, especially for the purposes of statistical inference.

Sample vs ***Population***

Example 1

- To describe the cholesterol level in US male from age 20- to 74-yr-old, *we need to supply μ and σ to characterize the normal distribution*.
- The sample used to yield these quantities must provide an accurate representation of the population. (For example, if we sampled only 60-yr-old men, μ would be too large.)

Why?



Important issues - 1

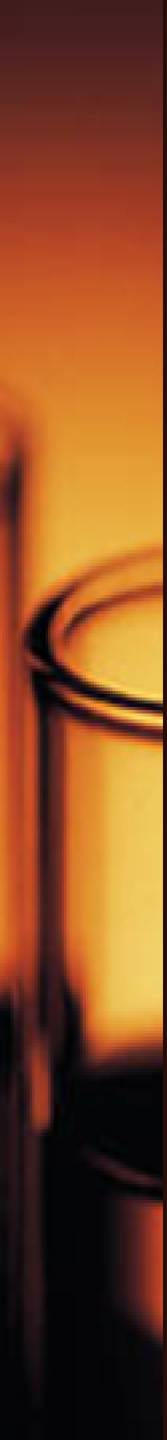
- The sample selection must be **random**; each individual in the population should have an **equal chance** of being selected.
  - How would we know if one particular batch of selections is **not** good?
  - That is, statistical characteristics based on this particular sample ***would be much different*** than some other samples we may possibly get from this population.

Important issues - 2

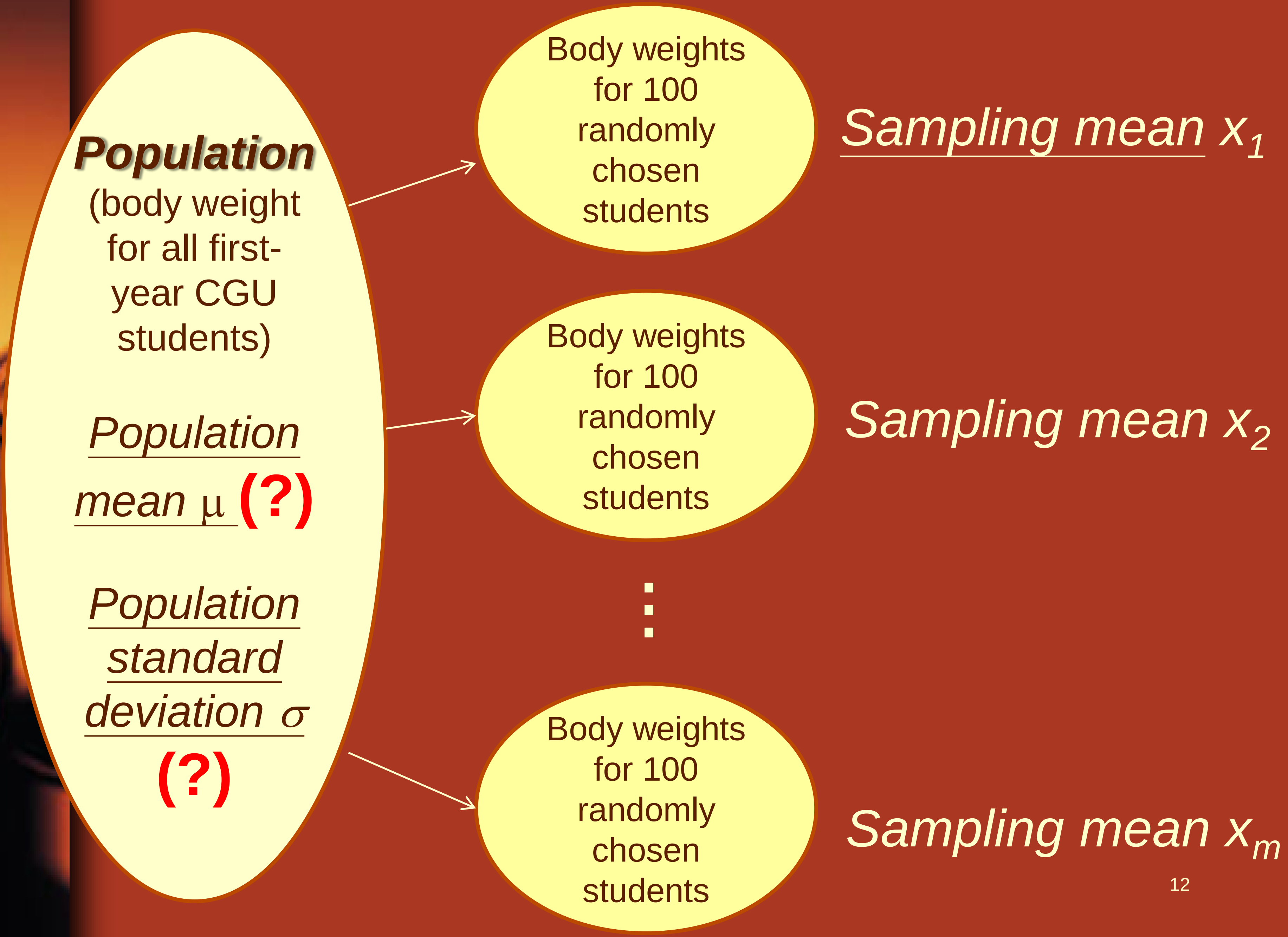
- The **sample size** should also be **large enough** to yield a more reliable estimation of these parameters (e.g., μ and σ if the underlying population is assumed to have a normal distribution).
- What sample size is considered large enough?

Sampling Distributions

- Following previous example (cholesterol level in US male from age 20- to 74-yr-old), where μ is the mean and σ the standard deviation.
- Let's randomly select **a sample of *n* observations** and compute the mean out of this sample, calling it x_1.
- Let's randomly select **another sample of *n* observations** and compute the mean out of this sample, calling it x_2.



- If the process continues, we’d have **a collection of x_i, each of them is a sample mean for n randomly selected observations**.
- **Consider this collection of x_i the outcomes of a random variable X** (representing the *mean cholesterol levels* of the *entire population*), the probability distribution of X itself is called ***a sampling distribution of means of samples of size n***.



Population

(body weight for all first-year CGU students)

Population mean μ **(?)**

Population standard deviation σ **(?)**

Body weights for 100 randomly chosen students

Sampling mean x_1

Body weights for 100 randomly chosen students

Sampling mean x_2

⋮

Body weights for 100 randomly chosen students

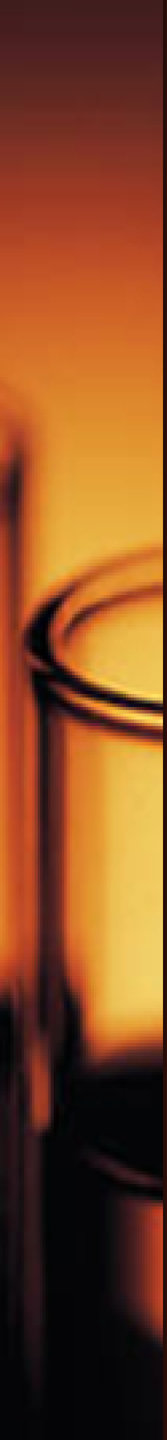
Sampling mean x_m

Cont’d

- In practice, of course, we don’t do that many samplings and choose one best sample out of them.
- Understanding some important properties of **the theoretical distribution of their mean values (those various sampling means)**, however, allows us to make inference based on a **single** sample of size *n*.

Summary

- We need to know, for example, two parameters μ and σ in order to make use of the theoretical probability density distribution in describing a general population (a **complete** set of observations) following a normal distribution.
- **These parameters (μ and σ), however, are generally unknown.**



Cont'd

- Although a sample of population can be used to compute for these parameters, we did not know whether the sampling is a good one or not.
- A **sampling distribution** can help in choosing an "ideal" parameter to use.
- For example, getting an "ideal" population mean value μ to use, chosen from a distribution of **many mean values** from many samples. **(Sample means now become the new random variable.)**

Population vs Sample

- Population statistics
  - **Population mean** and **population STD**
  - For example, the mean cholesterol level for all 20 to 65-year male in US is 145.5.
- Sample statistics
  - **Sample mean** and **sample STD**
  - For example, the mean cholesterol level for a group of 100 persons chosen from all 20 to 65-year male in US is 123.5.

Are they comparable?

- The sampling is a good one if these two statistics are **comparable**.

Example 2

- An extreme case with only 3 observed values of {1, 2, 5}. This is the *entire population*, with μ=2.6667 and σ=1.7*

Observation X	X-mu	(X-mu)^2
1	-1.6667	2.778
2	-0.6667	0.444
5	2.3333	5.444
		sum=8.667
μ=**2.6667**	**average**	
	stdev	2.0817
	stdevp*	σ=**1.7***

$$\sum_{i=1}^{n}(x_i-\overline{x})^2$$

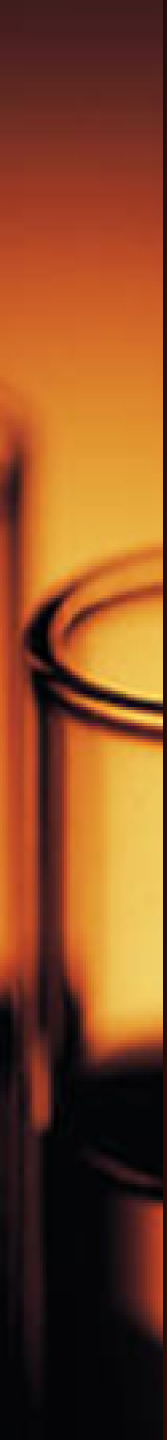
Variance of the "sample"

$$s^2=\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

Variance of the "population"

$$\sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

Observing that the variance based on "sample" and "population" are different!!



- Consider samples of n = 2.
- There are 9 such samples (see next slide).
- Observing various statistics from these samples, including the sample means.
- Observing how these ***sample-based parameters*** may ‘target’ the ***population parameters***.

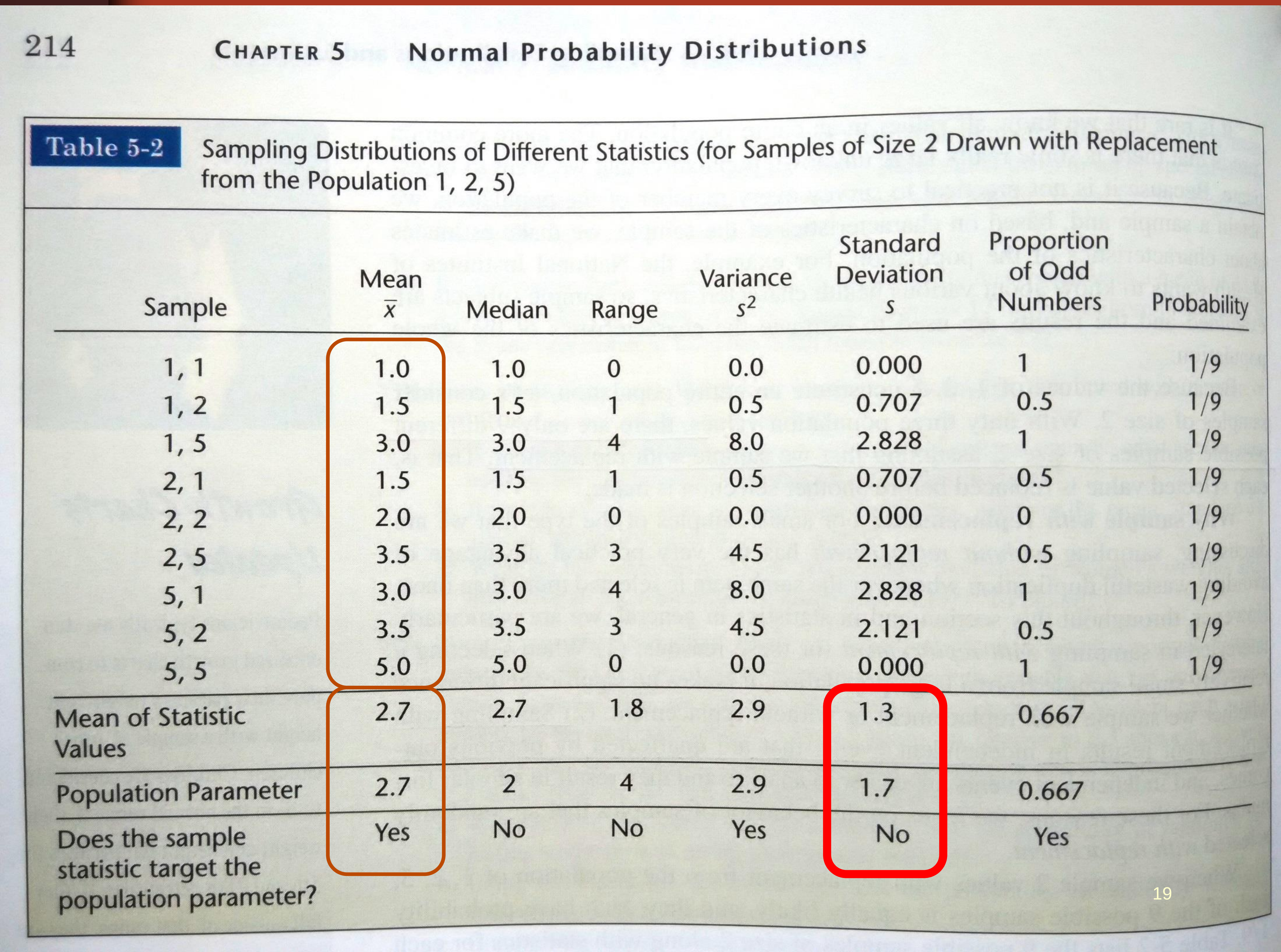


Table 5-2 Sampling Distributions of Different Statistics (for Samples of Size 2 Drawn with Replacement from the Population 1, 2, 5)

Sample	Mean $\bar{x}$	Median	Range	Variance s^2	Standard Deviation s	Proportion of Odd Numbers	Probability
1, 1	1.0	1.0	0	0.0	0.000	1	1/9
1, 2	1.5	1.5	1	0.5	0.707	0.5	1/9
1, 5	3.0	3.0	4	8.0	2.828	1	1/9
2, 1	1.5	1.5	1	0.5	0.707	0.5	1/9
2, 2	2.0	2.0	0	0.0	0.000	0	1/9
2, 5	3.5	3.5	3	4.5	2.121	0.5	1/9
5, 1	3.0	3.0	4	8.0	2.828	1	1/9
5, 2	3.5	3.5	3	4.5	2.121	0.5	1/9
5, 5	5.0	5.0	0	0.0	0.000	1	1/9
Mean of Statistic Values	2.7	2.7	1.8	2.9	1.3	0.667	
Population Parameter	2.7	2	4	2.9	1.7	0.667	
Does the sample statistic target the population parameter?	Yes	No	No	Yes	No	Yes	

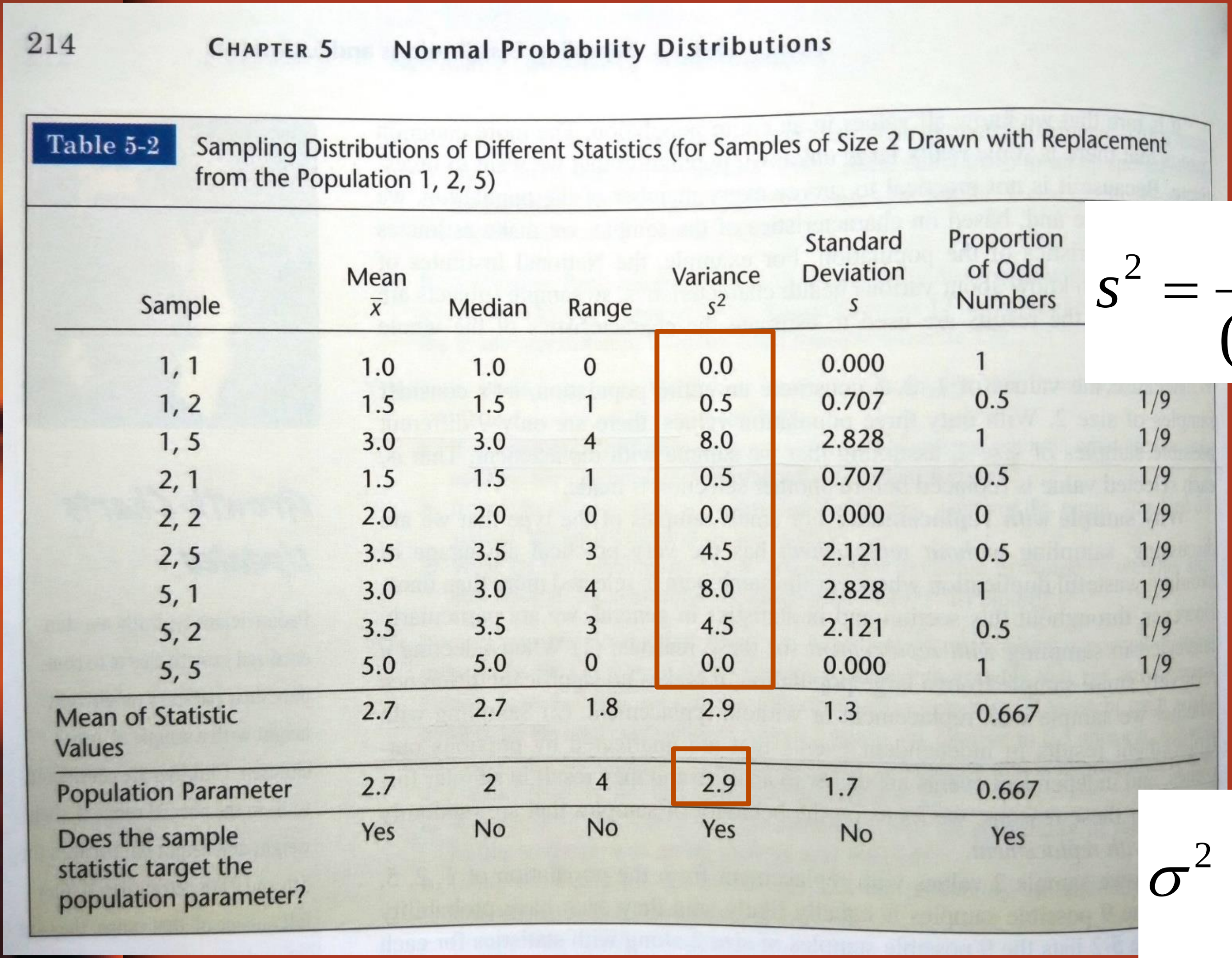


Table 5-2 Sampling Distributions of Different Statistics (for Samples of Size 2 Drawn with Replacement from the Population 1, 2, 5)

Sample	Mean $\bar{x}$	Median	Range	Variance s^2	Standard Deviation s	Proportion of Odd Numbers	
1, 1	1.0	1.0	0	0.0	0.000	1	
1, 2	1.5	1.5	1	0.5	0.707	0.5	1/9
1, 5	3.0	3.0	4	8.0	2.828	1	1/9
2, 1	1.5	1.5	1	0.5	0.707	0.5	1/9
2, 2	2.0	2.0	0	0.0	0.000	0	1/9
2, 5	3.5	3.5	3	4.5	2.121	0.5	1/9
5, 1	3.0	3.0	4	8.0	2.828	1	1/9
5, 2	3.5	3.5	3	4.5	2.121	0.5	1/9
5, 5	5.0	5.0	0	0.0	0.000	1	1/9
Mean of Statistic Values	2.7	2.7	1.8	2.9	1.3	0.667	
Population Parameter	2.7	2	4	2.9	1.7	0.667	
Does the sample statistic target the population parameter?	Yes	No	No	Yes	No	Yes	

$$s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Variance for each sample (n=2) was computed by this formula.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Variance for the population (n=3) was computed by this formula.

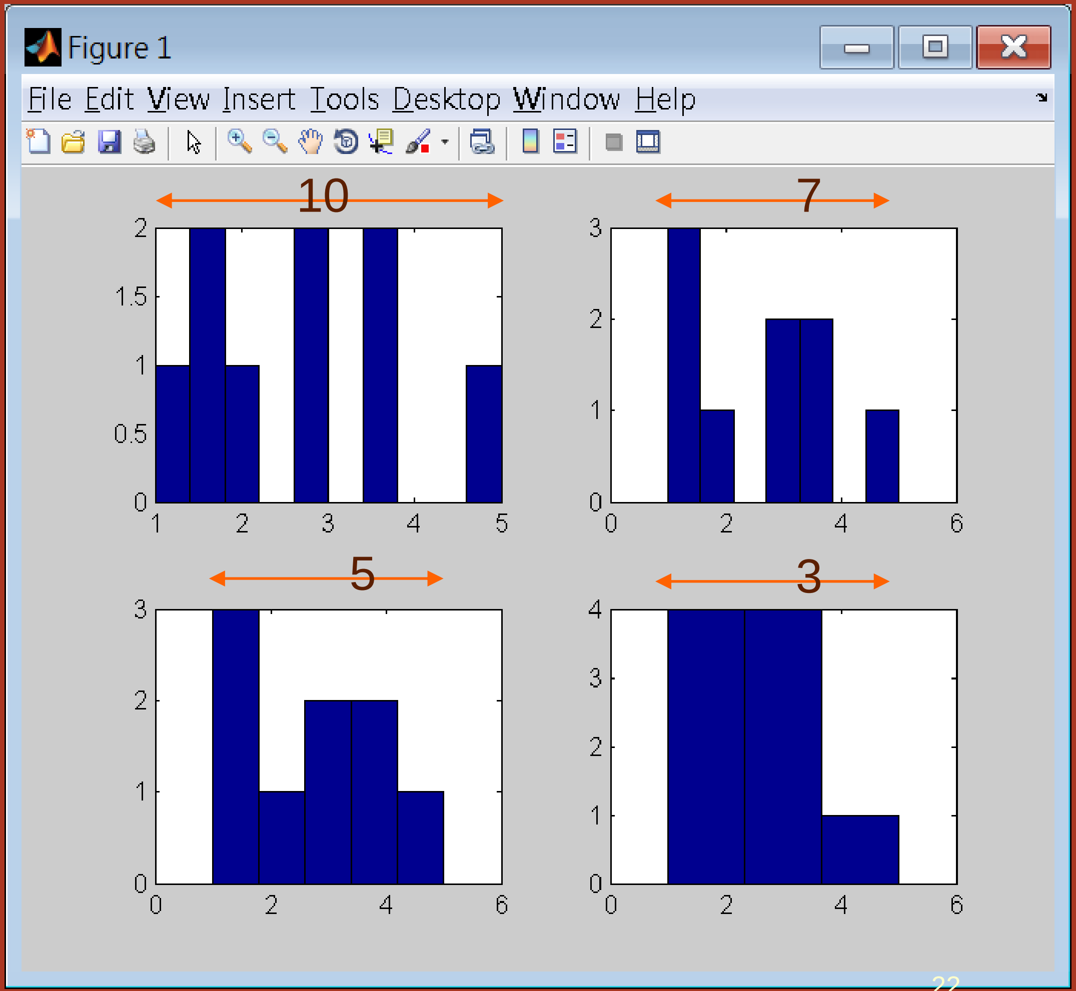
Conclusion

- ***Sampling variability***: A sample mean depends on the particular values included in the sample, generally varies from sample to sample.
- **The mean of all possible sample means *is equal to* the mean of the original population**.
- The means for all possible sample standard variations (*s*=1.3), however, is ***different and smaller*** than the population standard deviation (σ=1.7)

Distribution for the sample means

Mean $\bar{x}$
1.0
1.5
3.0
1.5
2.0
3.5
3.0
3.5
5.0

```
>> X=[1.0  1.5  3.0  1.5  2.0  3.5  3.0  3.5  5.0]
>> subplot(2,2,1)
>> hist(X)
>> subplot(2,2,2)
>> hist(X,7)
>> subplot(2,2,3)
>> hist(X,5)
>> subplot(2,2,4)
>> hist(X,3)
```



Although it is not obvious here, we expect to see a normal distribution for these sample means.

8.2 Central Limit Theorem

Central Limit Theorem

- For a certain population, where μ is the mean and σ the standard deviation, this theorem states the followings:
  - (1) The **mean of sampling distribution of sample means** is the same as the population mean μ.

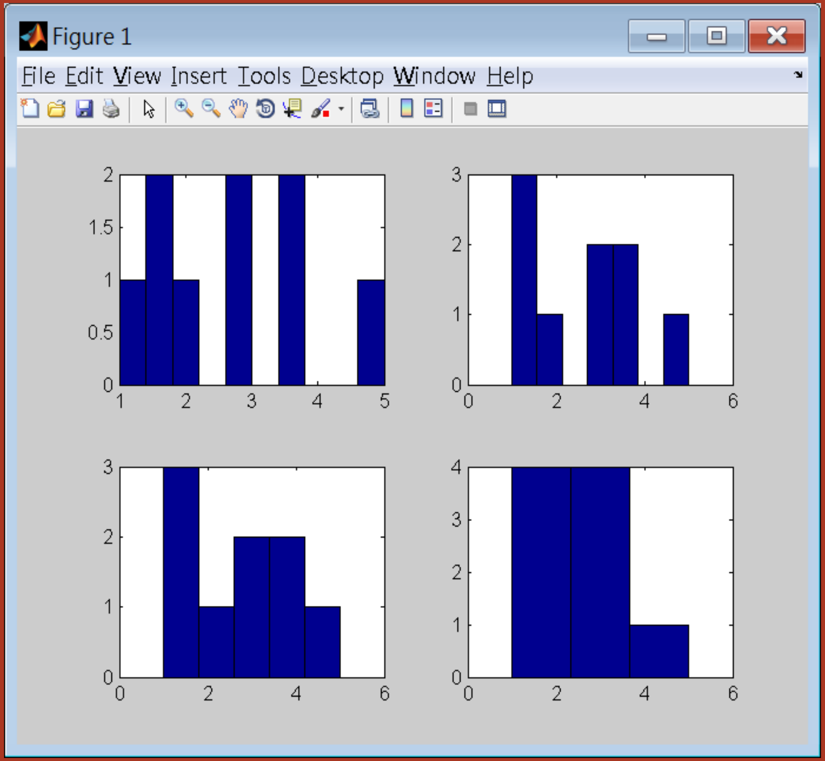
Recall that in example 2, we have the mean of sampling means as 2.7, which is equal to the population mean 2.7.

- (cont'd):
  - (2) Standard deviation of the sampling distribution of sample means is **σ/sqrt(n)**. (This is called the **standard error of the mean (SEM)** upon sampling)

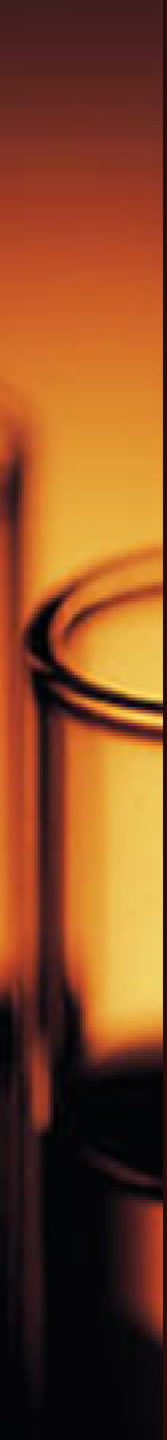
Recall that in example 2, we have s = 1.3 for the mean of sampling means, and σ = 1.7 for the population. Indeed, 1.3 is roughly the same as 1.7 divided by the square root of 2.

This tells us why we'd prefer larger sample size (to make SEM small), to better make sure the mean value we chose from a sample is good enough to represent the general population.

- (cont’d):
  - (3) When n is large, the sampling distribution of the sample means is approximately **normal**.



Note that these are the sample means with n=2.



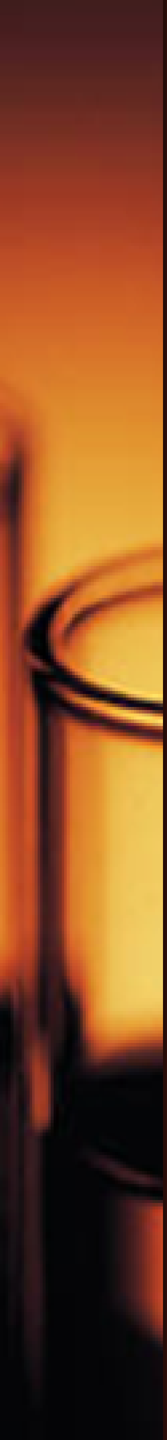
Comment

- The aforementioned central theorem applies to ***any*** **population with a** ***finite*** **standard deviation (σ)**, regardless of the shape of the underlying distribution.
- The farther departure from being normally distributed, however, ***would require larger n*** to ensure the normality of the sampling distribution. (Usually **a sample size of 30** would be large enough.)

8.3 More Examples

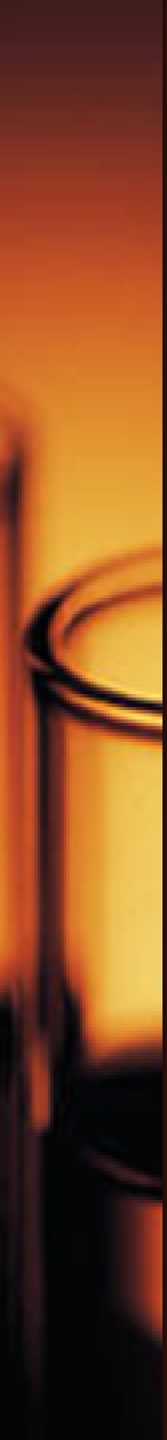
Example 3

- Consider the distribution of serum cholesterol levels for **all** 20- to 74-yr-old males in US, with mean **μ=211** and σ=46.
- Remember these are called **population mean** and **population standard deviation**, in contrast to the mean and standard deviation derived from a ***particular*** sample of ***limited*** size.



- Consider a sampling size of n=25. (We may have **many** such samples, each contains 25 observations.)
- ***Question - What proportion of these many samples (each of n=25) will have a mean value of 230 or higher?*** (Recall that the population mean is 211.)

Note that we are not asking “what proportion of the population will have cholesterol level of 230 or higher”.

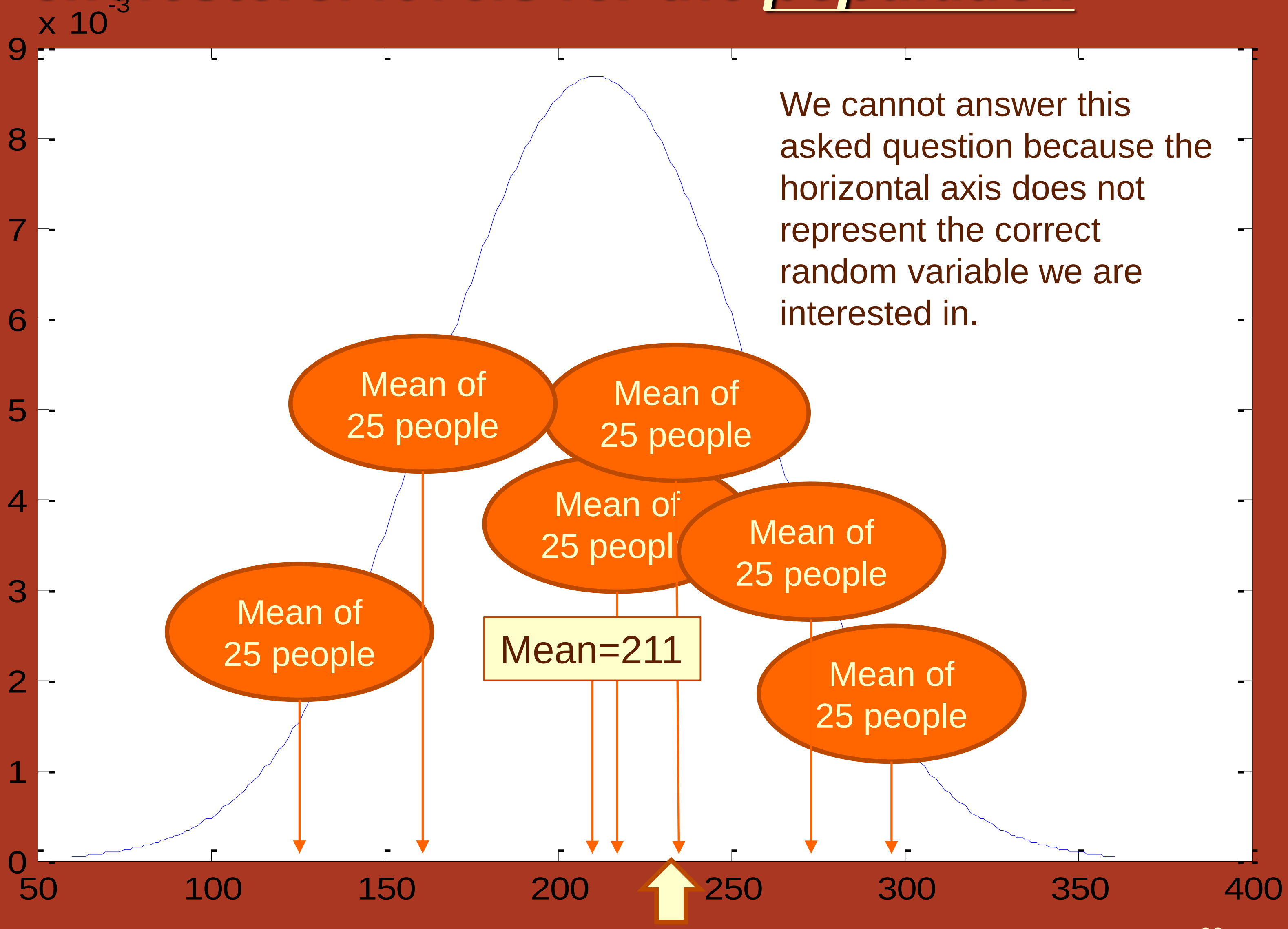


Analysis before solving

- Recall that mean **μ=211** and σ=46 for the entire population.
- If our answer is small (e.g., only 10 % of such samples of n=25 may have mean values of 230 or higher), then **many*** of these samples (each of n=25) will have a mean value of **230 or smaller that are closer to the actual mean 211.**.

* According the central limiting theorem, ***the distribution of these sample means*** is approximately normal. So there would be 40% of the sample means fall between 211 and 230.

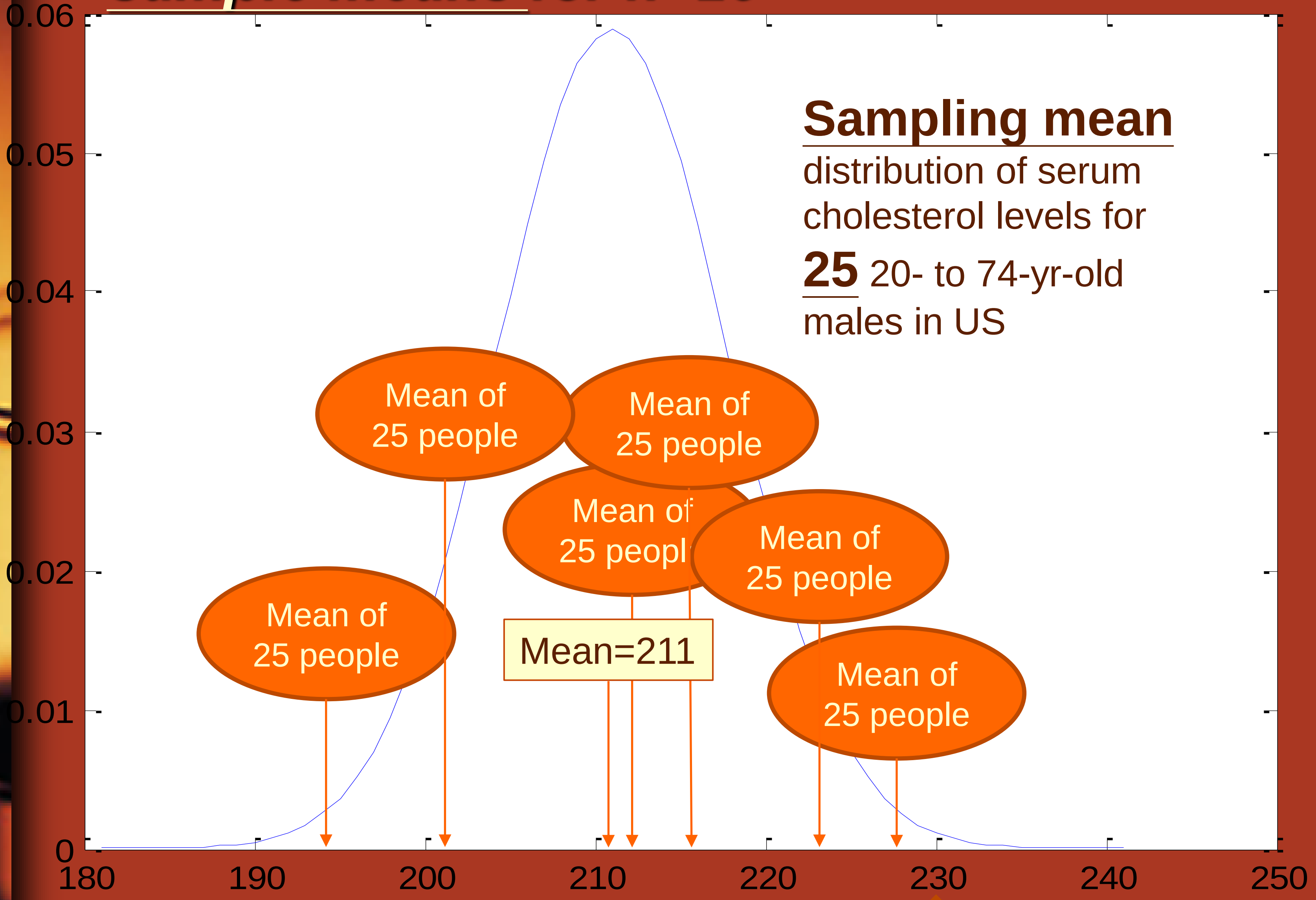
Given the distribution of serum cholesterol levels for the population~~



We cannot answer this asked question because the horizontal axis does not represent the correct random variable we are interested in.

230 or higher

Given the sampling distribution of the sample means for n=25~~



Sampling mean distribution of serum cholesterol levels for **25** 20- to 74-yr-old males in US

Mean of 25 people

Mean of 25 people

Mean of 25 peopl

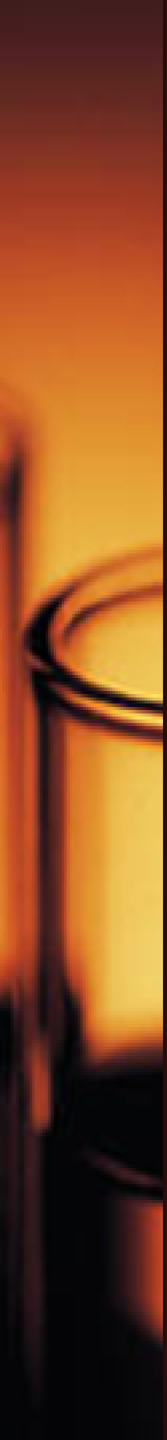
Mean of 25 people

Mean of 25 people

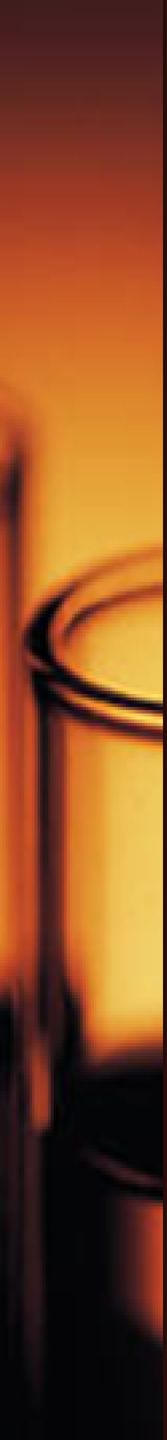
Mean of 25 people

Mean=211

230 or higher



- So it will be **doubtful** for this particular sample (of n=25) with mean value=230 to represent the entire population. [Because there are many with smaller means that are more close to the actual mean of 211, which are apparently better samples.]



- Imagine our 25 observations are mostly drawn from **older people**, the mean value would likely to be higher than the population mean 211, and potentially higher than the stated 230 threshold.
- A sampling size of 25 in this particular case **will not be too reliable** to give us accurate estimated μ to use in representing the ***whole*** population (20 to 74-yr old).

- If the particular sample mean is **too far off (already far enough)** the actual mean 211 in the sampling mean distribution, it is not easy to find other sampling means which are further far off.
- ***This is why we asked the question for “the probability that our sampling mean will be 230 or greater”.*** [Recall that sampling mean is also a random variable, and it follows a normal distribution.]

Solution

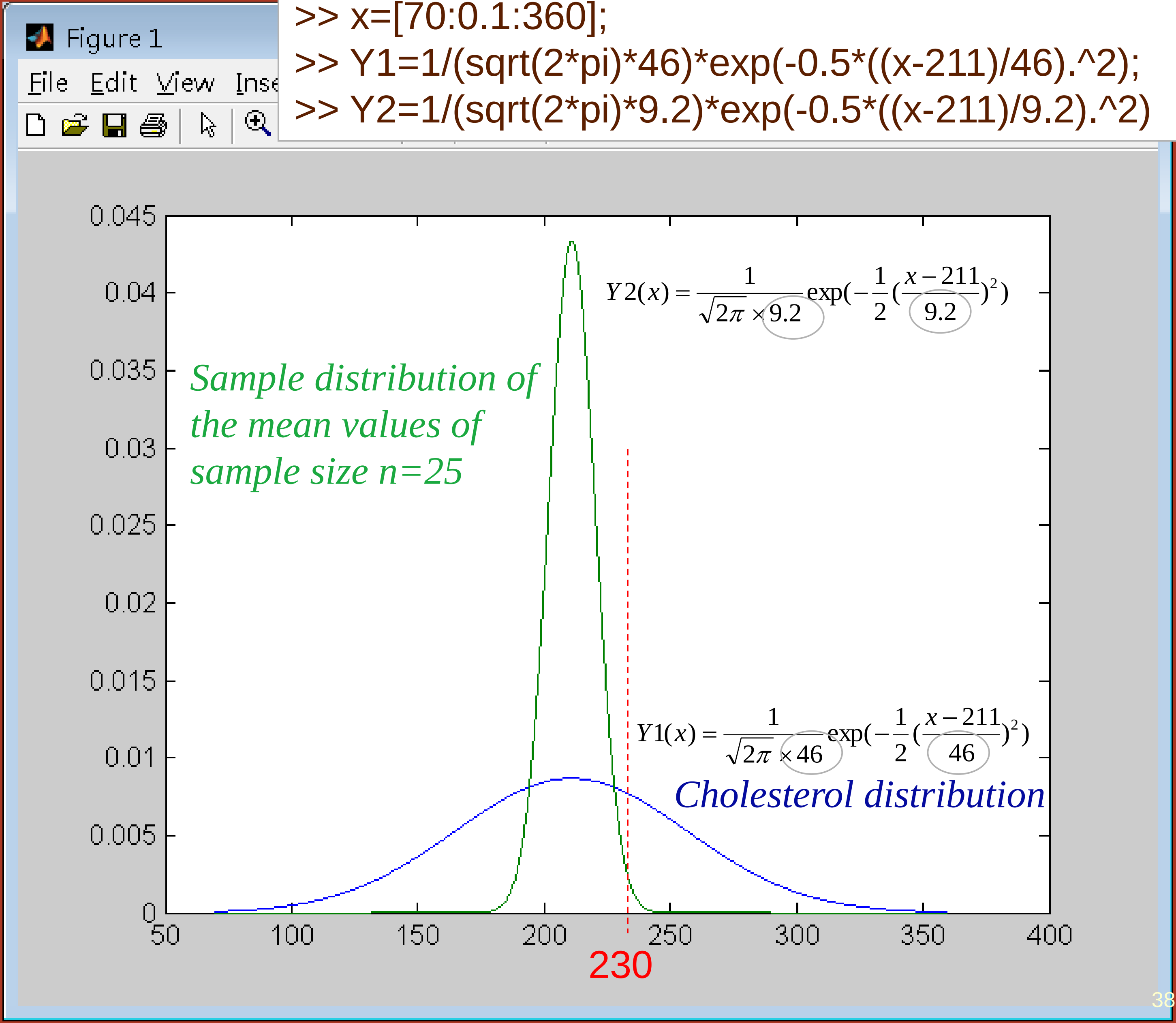
- According to the central limit theorem, we will have **a normal sampling distribution of the means, with the mean μ = 211 and standard deviation σ = 46/25$^{1/2}$ = 46/5 = $\underline{9.2}$.**
- As a result, the following transformation will bring this sampling distribution to a standard normal distribution with μ=0 and σ=1.

Note this is the sampling distribution of the sample means, not the cholesterol level distribution.

$$Z=\frac{X-211}{9.2}$$

Use 9.2, not 46!!!

```
>> x=[70:0.1:360];
>> Y1=1/(sqrt(2*pi)*46)*exp(-0.5*((x-211)/46).^2);
>> Y2=1/(sqrt(2*pi)*9.2)*exp(-0.5*((x-211)/9.2).^2)
```



$$Y2(x)=\frac{1}{\sqrt{2\pi}\times 9.2}\exp(-\frac{1}{2}(\frac{x-211}{9.2})^2)$$

Sample distribution of the mean values of sample size n=25

$$Y1(x)=\frac{1}{\sqrt{2\pi}\times 46}\exp(-\frac{1}{2}(\frac{x-211}{46})^2)$$

Cholesterol distribution

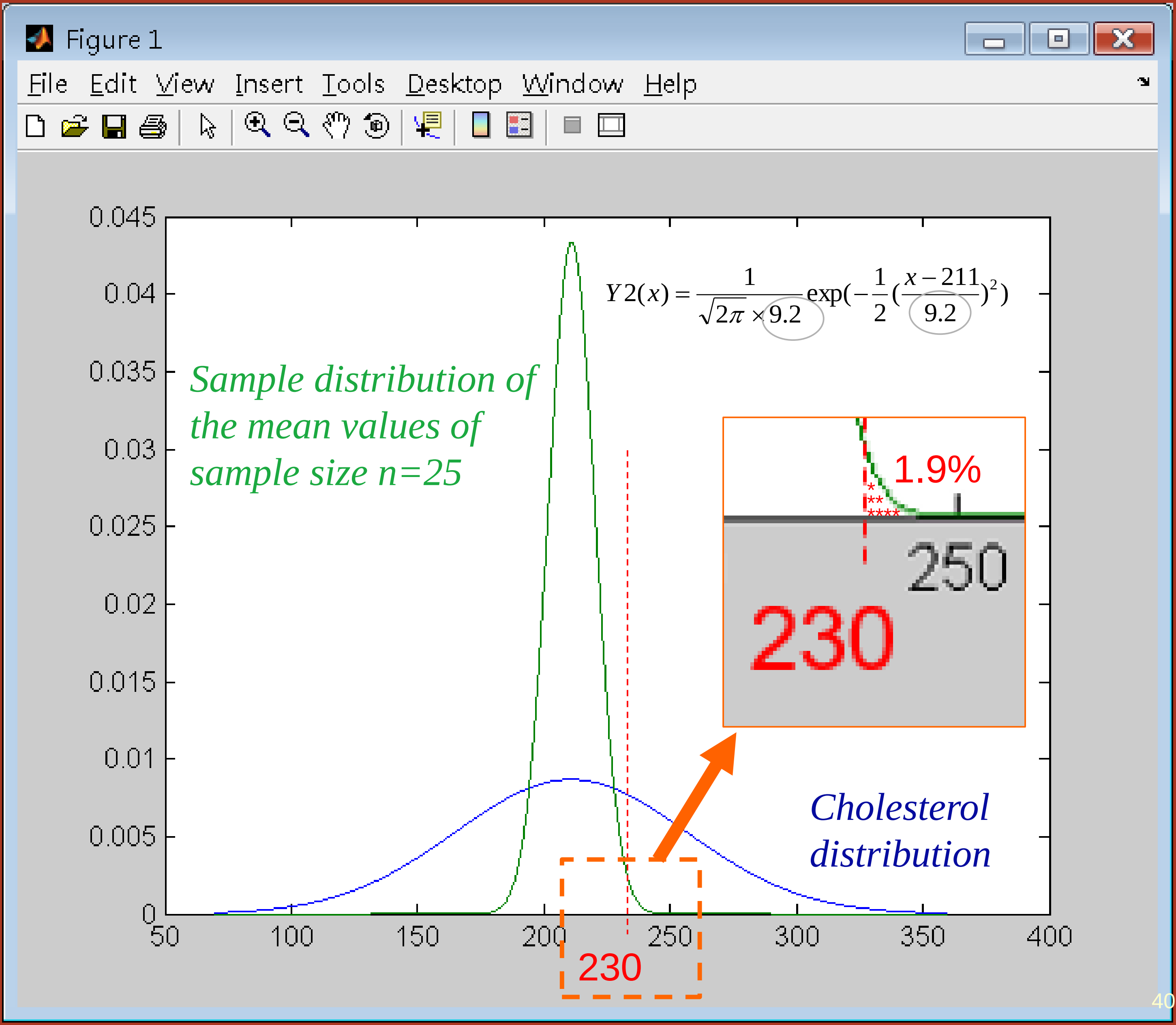
230

- Now, to get the solution, we may convert X=230 to a Z-score, and find the area under the right-tail of the standard normal distribution, as we did in the last lecture.
- As a result, approximately only **1.9%** chance the mean value would go beyond 230 for this sampling process. (This is expected by the **narrow** bell shape seen from previous slide.)

```
>> F='1/(sqrt(2*pi))*exp(-0.5*x^2)‘;
>> z=(230-211)/9.2
z =    2.0652
>> int(F,z,inf)
ans =.19451217852135832736501724576047e-1
>>
```

A standard normal distribution

Integration from 2.0652 to ∞ to get the area under the right-tail.



$$Y2(x)=\frac{1}{\sqrt{2\pi}\times 9.2}\exp(-\frac{1}{2}(\frac{x-211}{9.2})^2)$$

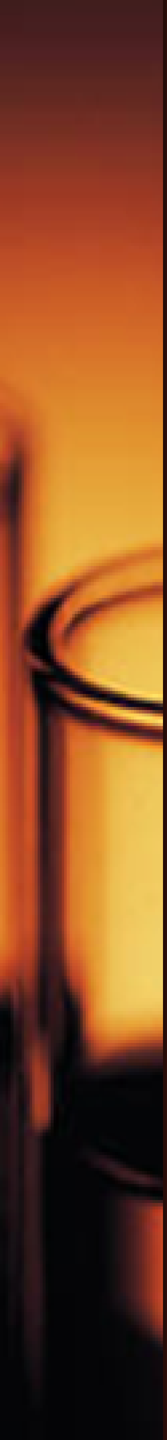
Sample distribution of the mean values of sample size n=25

Cholesterol distribution

- Alternatively, when conveniently using MATLAB shown below, we don’t even have to convert X into a Z-score. We can directly perform the integration over the original sampling distribution, from **230 to inf**. The answer is exactly the same.

```
>> F2 =‘1/(sqrt(2*pi)*9.2)*exp(-0.5*((x-211)/9.2)^2)’;
>> int(F2,230,inf)
ans =.19451217852135832736501724576025e-1
>>
```

Note to use 9.2 (subject to sample size of 25), not the population standard deviation of 46.

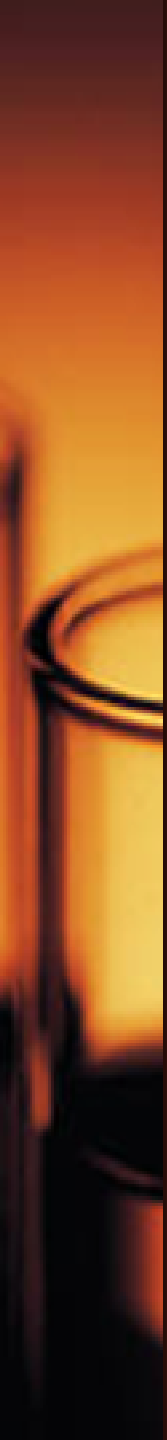


Solution summary

- There is only a chance of 1.9% that the other samples would get a mean value of 230 or bigger.
- Our sample with mean=230 is far off the actual population mean.
- It’s hard to find other samples having a more inaccurate mean cholesterol value than this one.
- ***Conclusion : This is NOT a good sampling.***

Applications

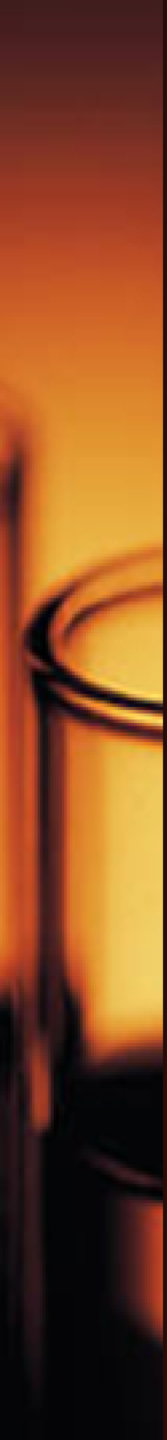
- Similarly, we may compute, for example, the **range of mean value** to give 2.5% under both-end tails.
- This covers **95%** of the mean values, all can be considered ***"good enough"*** (for these samples) to represent this entire population.



- It can be shown that Z-value between **-1.96** and **+1.96**, or mean cholesterol value between **193.0** and **229.0** would render this 95% proportion, under the sampling size of n=25.
- ***Try obtaining these intervals using MATLAB!***

Discussion - 1

- Why would we say "covering **95%** of the mean values is considered ***"good enough"*** (for these samples) to represent this entire population?
- Wouldn't 90%, even 50% much better?

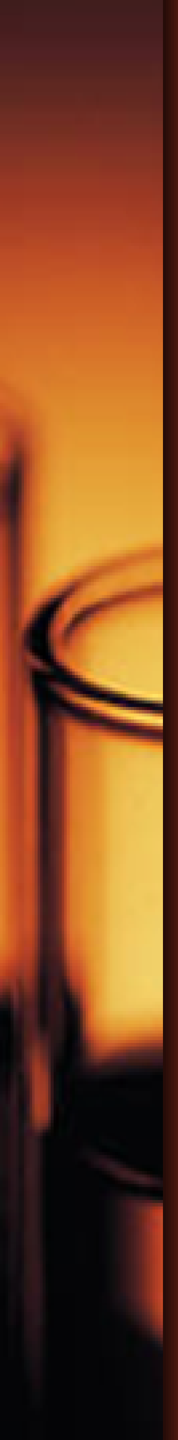


Discussion - 2

- In fact, here we are seeing this “good enough” from thinking the other way around.
- Anything that are ***“not extreme”*** are considered “good”.
- In the world of statistics, 5% is in general considered an “extreme”.

Discussion - 3

- Two tails vs one tail?
- Are very low and extremely high values are both considered extremes? (e.g., blood pressure too low and too high are both bad)
- Or only very low (e.g., dying young) or only very high (e.g., air pollution) is considered extreme?



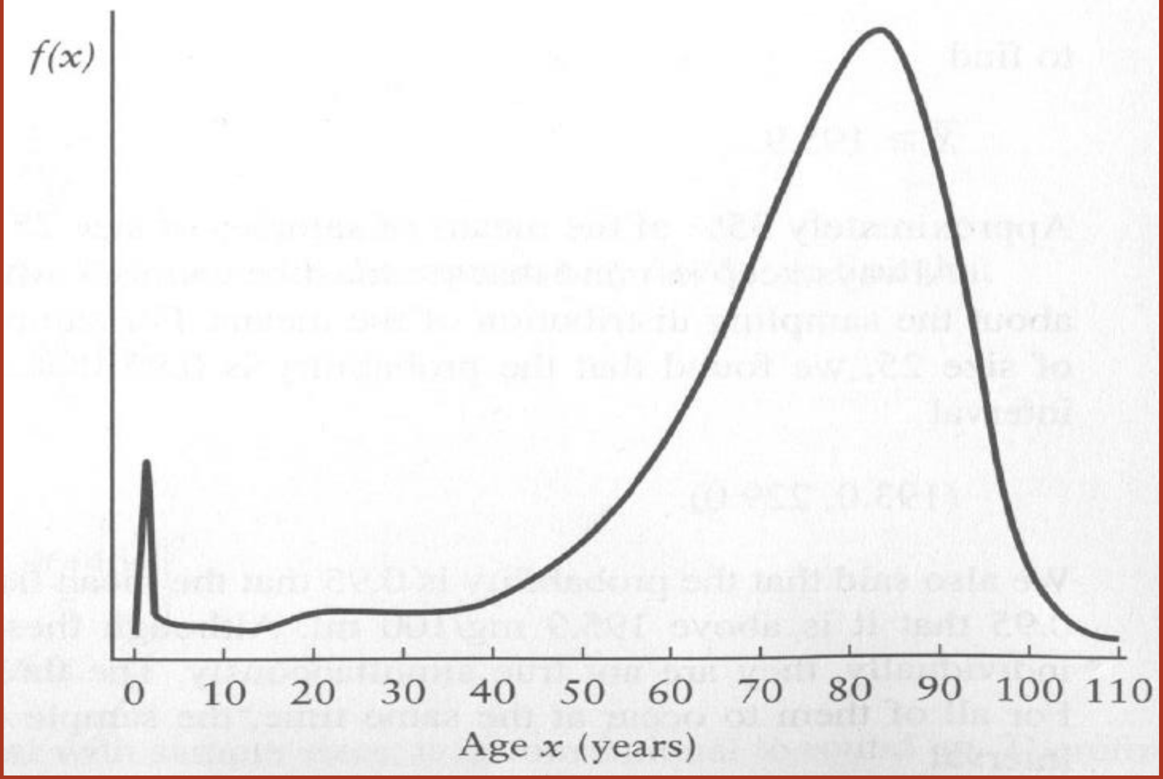
How large is good enough for the sampling size n?

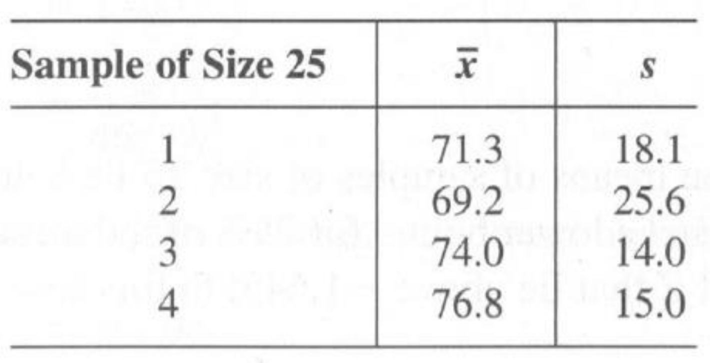
n	σ/sqrt(n)	Interval covering 95% of the means	Length of interval
1	46.0	120.8 – 301.2	180.4
10	14.5	182.5 – 239.5	57.0
25	9.2	193.0 – 229.0	36.0
50	6.5	198.2 – 223.8	25.6
100	4.6	202.0 – 220.0	18.0

Preferred sample sizes to get satisfactory sample mean to represent the actual population mean. The narrower the interval is, the better the sampling is.

Example 4

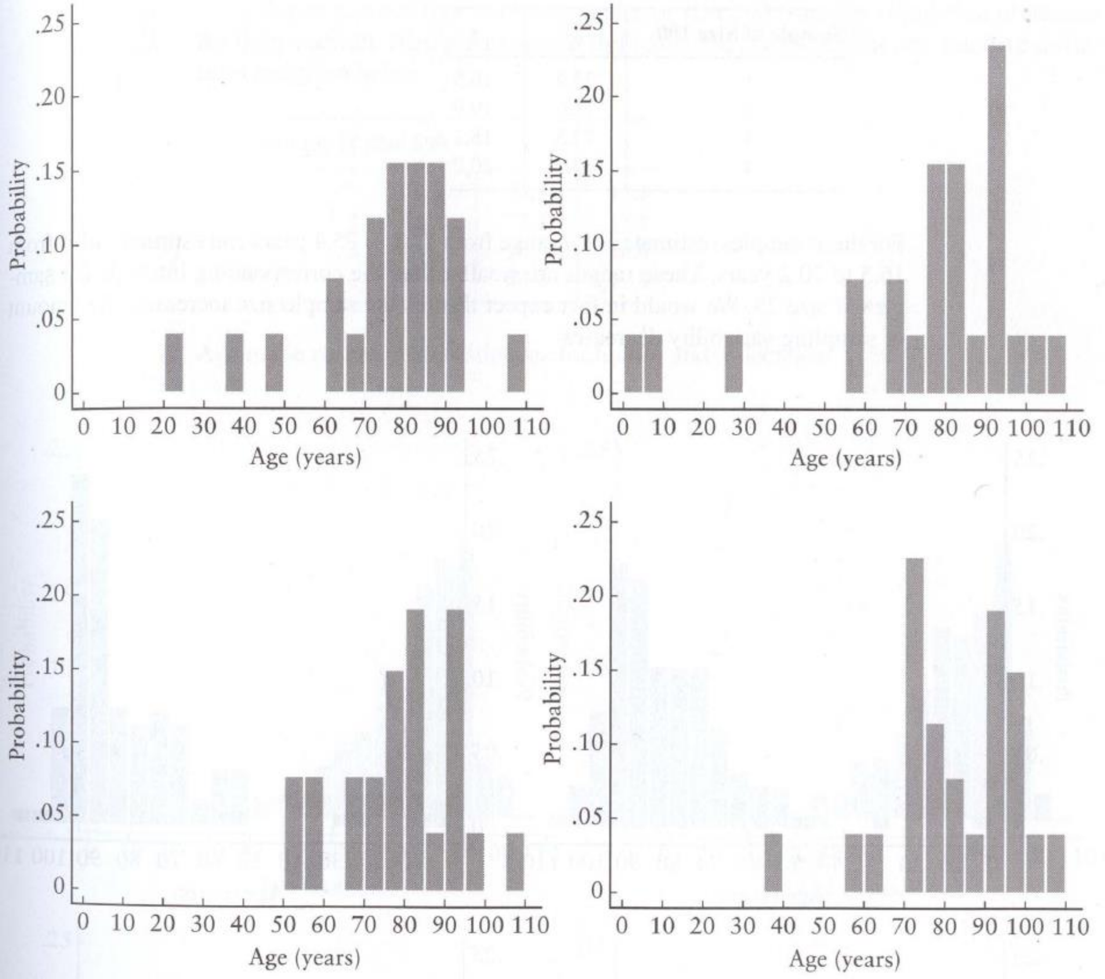
This graph represents the distribution of age at the time of death in US in 1979-1981. It has **μ=73.9** and **σ=18.1**. Note that **we do not have a normal distribution here**. What do we expect to happen when we sample from this population of ages?



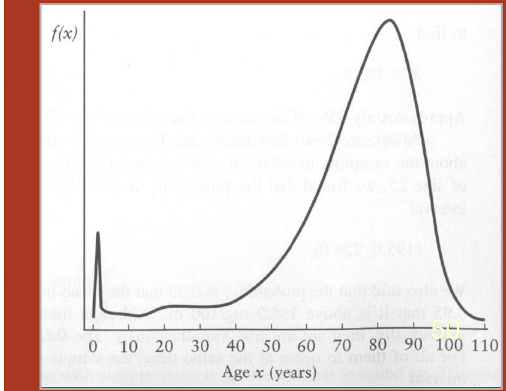


Sample of Size 25	$\bar{x}$	s
1	71.3	18.1
2	69.2	25.6
3	74.0	14.0
4	76.8	15.0

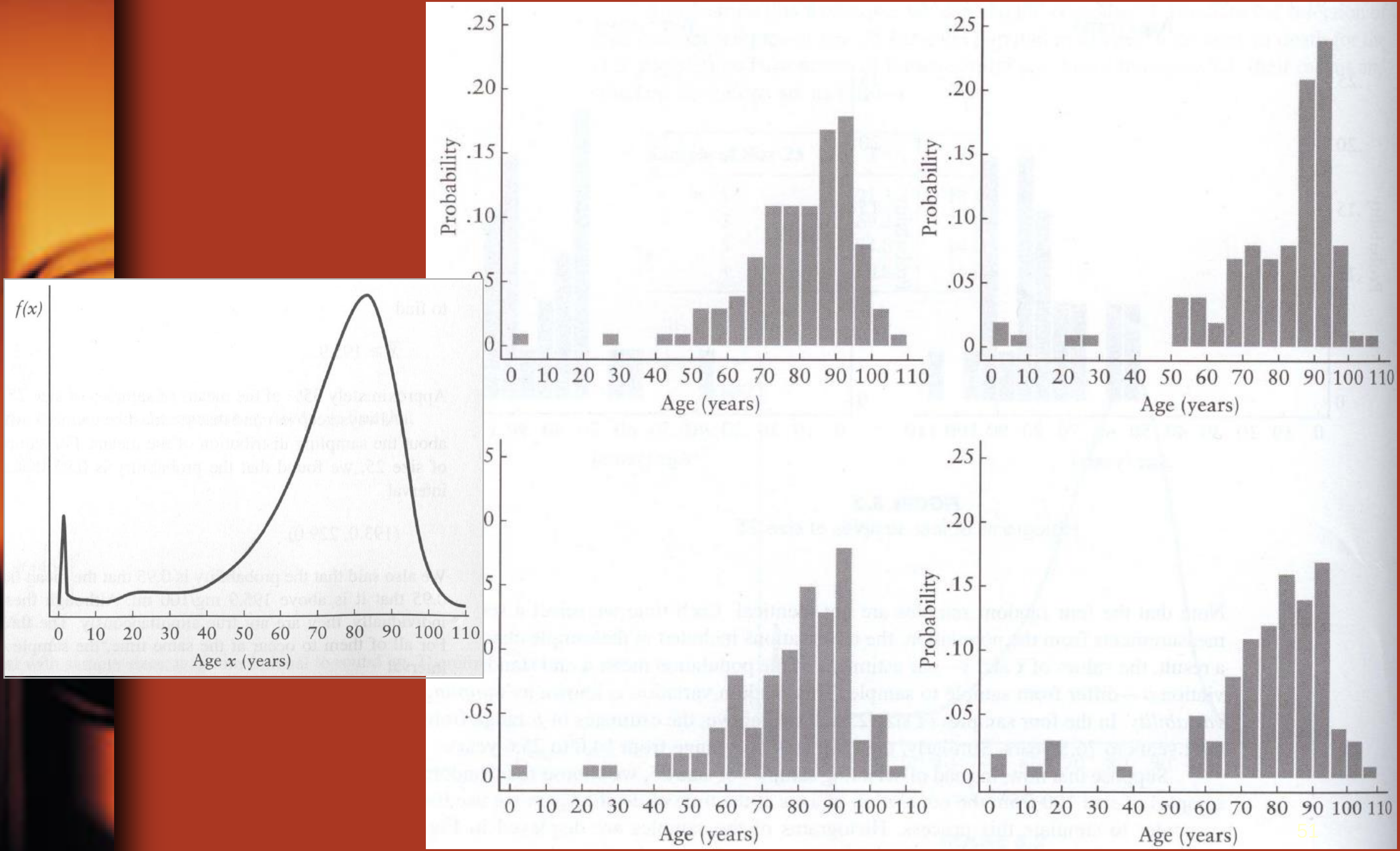
4 randomly selected samples were formed, with their mean and STD indicated on the left. Corresponding probability histograms are shown below. You can see the **variability** from such sampling.



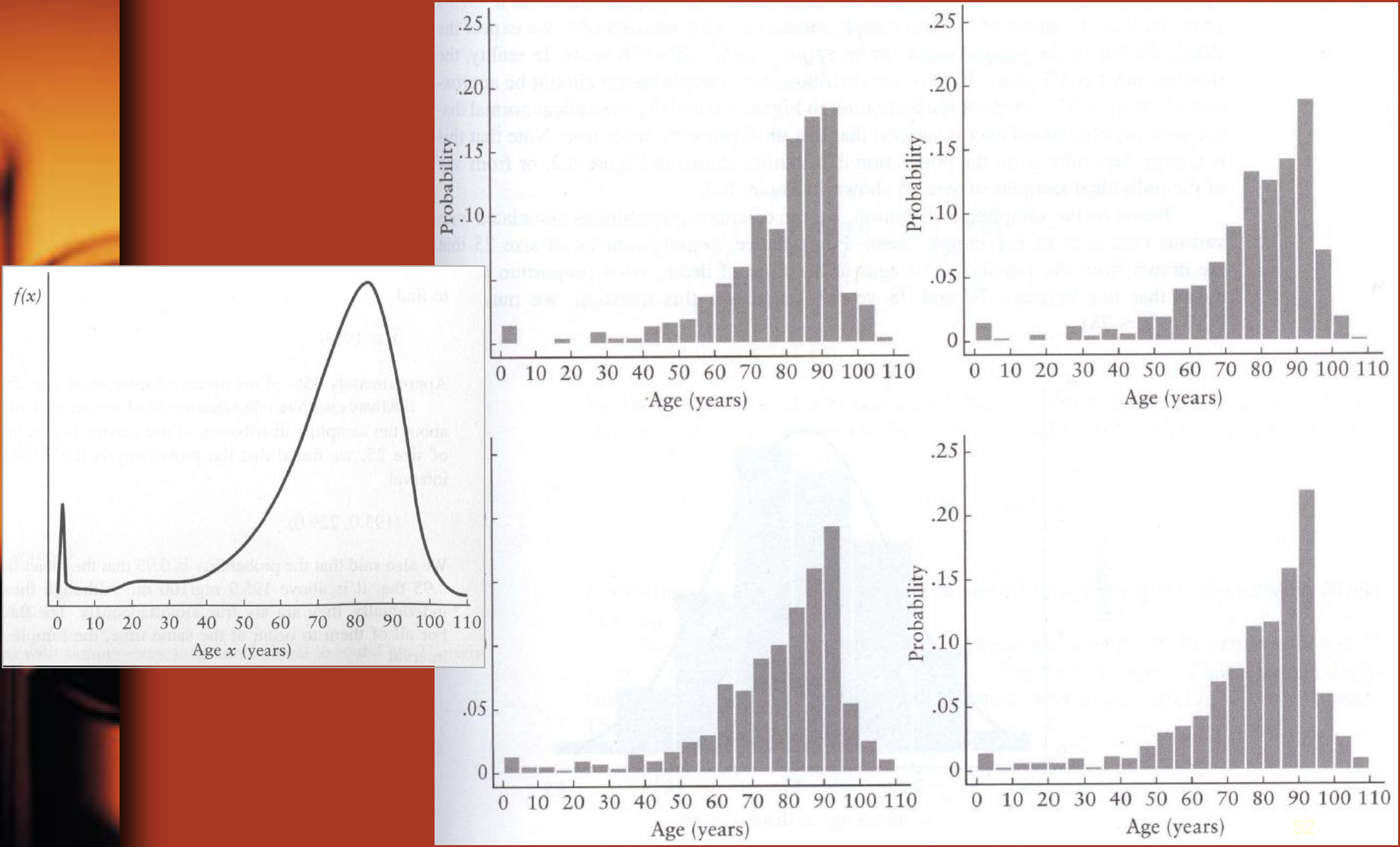
μ=73.9 and σ=18.1 for the population.

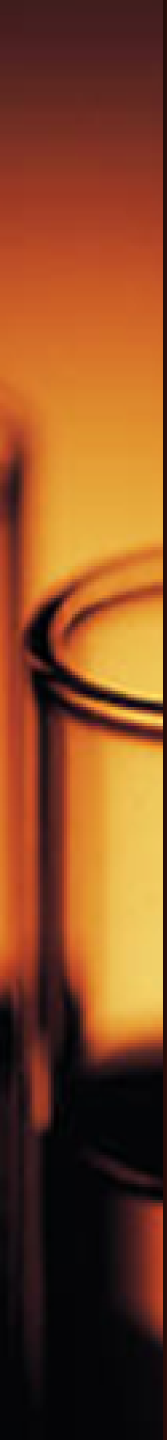


Histogram of 4 samples of **size 100**



Histogram of 4 samples of **size 500**





Conclusion

- This example shows us that, even if the underlying distribution is not normal, the samplings with bigger *n* would still have very good chance to get a sample statistics which is similar to the real population statistics.
- As stated earlier, *n* can be smaller if the underlying distribution is already normal, and must be bigger if the underlying distribution is not normal.

Example 5

- Assume that the population of ages at the time of death. The mean value μ=73.9 years and σ=18.1 years.
- Q1: Among samples of size 25 that are drawn from this population, what proportion have a ***mean age*** of death that lies between 70 and 78 years?
- Q2: What will be the proportion in Q1 if the sample size is 100?

- According to central limit theorem, the mean values X of these samples will have a normal distribution ***centered at*** the population mean 73.9, with a standard deviation 18.1 divided by the square root of the sample size 25.
- This normal distribution (mean values from many n=25 samples) can be converted to a standard normal distribution as:

$$Z=\frac{X-73.9}{18.1/\sqrt{25}}=\frac{X-73.9}{3.62}$$

$$P(70 \le X \le 78)$$

$$P\left(\frac{70-73.9}{3.62} \le \frac{X-73.9}{3.62} \le \frac{78-73.9}{3.62}\right)$$

$$P(-1.08 \le Z \le 1.13)$$

```
>> F='1/(sqrt(2*pi))*exp(-0.5*x^2)‘;
>> int(F,-1.08,1.13)
ans = 0.73069079767121312631047998477208
```

For n = 100:

$$Z = \frac{X - 73.9}{18.1/\sqrt{100}} = \frac{X - 73.9}{1.81}$$

$$P(70 \le X \le 78)$$

$$P(\frac{70 - 73.9}{1.81} \le \frac{X - 73.9}{1.81} \le \frac{78 - 73.9}{1.81})$$

$$P(-2.15 \le Z \le 2.27)$$

```
>> F='1/(sqrt(2*pi))*exp(-0.5*x^2)‘;
>> int(F,-2.15, 2.27)
ans = 0. 97261860108700595588800257362372
```

The percentage gets much higher that n=25 because the bell-shape becomes narrower.