INFO 6205
Program Structure and Algorithms

Nik Bear Brown

Sampling Theory

Topics

- Types of samples
- Sampling frames
- Why samples work

Census or sample?

Edinburgh Council wants to know peoples thoughts on its leisure facilities.

A decision is made to send a questionnaire to every household in Edinburgh. A 5% response rate is expected.

This is a census of all households in Edinburgh

Census or sample?

Boxes of questionnaires are placed at the door to every council leisure facility in Edinburgh. There is a large sign next to the door advertising the questionnaire.

This is a census of everyone that uses the leisure facilities

Sampling theory does not apply to census results

What is Sampling?

- Identify population
- Select members of population to sample
- Study selected members (the sample)
- Draw inferences about population from sample

Types of samples

- Non-Probability Samples

  Not all members of the population have a chance to be included in the sample. Selection method not random.

- Probability Samples

  Every member of the population has a known, nonzero chance of being included in the sample. The selection method is random.

Non-Probability Samples

Convenience

- Sample selected simply for ease

Example – Edinburgh Council go to Ocean Terminal and hand questionnaires to everyone they encounter until they have 1000 complete.

- Quick and cheap

Non-Probability Samples

Convenience

Example – Want to know views on Trams project.

- Go to George St at 16:30 on Friday afternoon and ask 1000 people for their views
- Go to airport at 16:30 on Friday afternoon and ask 1000 people for their views

- Bias is a major problem – sample unlikely to be representative

Non-Probability Samples

Quota Sampling

- Subjective choice of sample based on what researcher thinks is representative.
- Example – Edinburgh council send 3 researchers knocking on doors in Granton, Broughton and Gilmerton until they have responses broken down as:

Age	Men	Women
10 to 20	120	100
20 to 30	100	100
30 to 40	55	50
40 to 50	70	65
50 to 60	80	90
60+	60	110

Non-Probability Samples

Quota Sampling

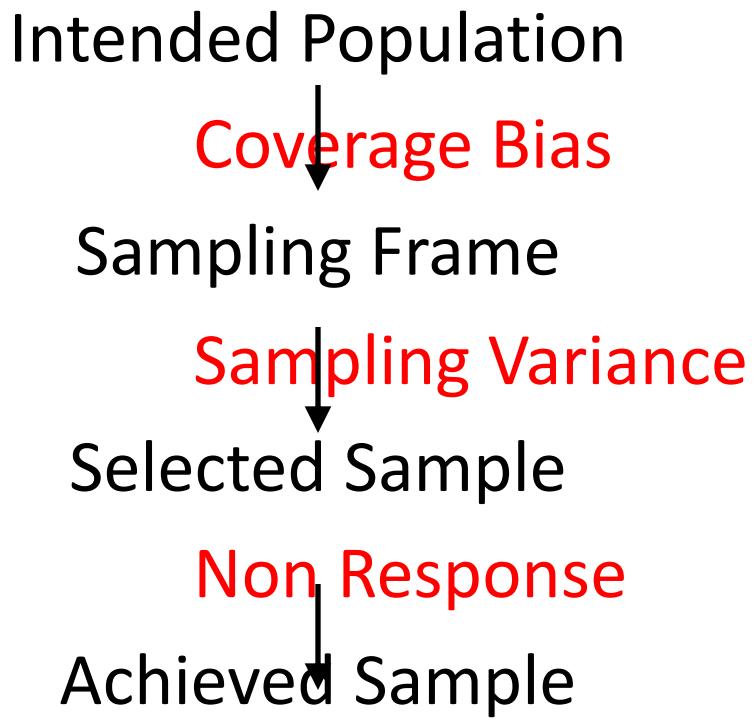
- Quicker and cheaper than probability sampling
- Large non response
- Still may not be representative
- What do base your quota on – age, gender, ethnicity, education, tenure, religion?

Probability Sampling

Example – 1000 households randomly selected across Edinburgh.

- More expensive and slower
- Non response a problem – but resources can be targeted and extent of non response bias can be estimated.
- Enables precision of final statistics to be assessed.
- Sample selection method is objective, specified and replicable.

Intended and achieved populations



Sampling Frames

- List of all units/people that could be included in sample

- Sample is only as good as the sampling frame
  - o Eligible units/people not on frame cannot be selected – leads to coverage error
  - o Units/people on frame more than once changes probability of being selected
  - o Ineligible units/people on frame can lead to final sample being smaller than intended.

Sampling Frames

Frames can be created:

- From pre-existing lists (UK postcode address file)
- Geographically with multi-stages
- Through time

Simple random sampling

- Sampling method completely random based on random numbers.
- Is easy to understand, can be expensive

Example – every household in Edinburgh assigned a number. 1000 random numbers chosen between 1 and 204,683. These numbers identify which households are in sample.

- Tables of random numbers
- www.random.org
- Excel function '=rand()'

Systematic Sampling

- Uses a ‘random’ start on the sampling frame and then selects every i’th unit/person.

- Easy to understand, quick and easy to implement.
- Can lead to some stratification depending on how the list is ordered.
- Can be expensive
- Need to be careful on how list is ordered to avoid bias.

Systematic Sampling

Sampling Frame

1	16	31	46
2	17	32	47
3	18	33	48
4	19	34	49
5	20	35	50
6	21	36	51
7	22	37	52
8	23	38	53
9	24	39	54
10	25	40	55
11	26	41	56
12	27	42	57
13	28	43	58
14	29	44	59
15	30	45	60

<u>Example</u>

Total units on sampling frame = 60

Want sample of 10

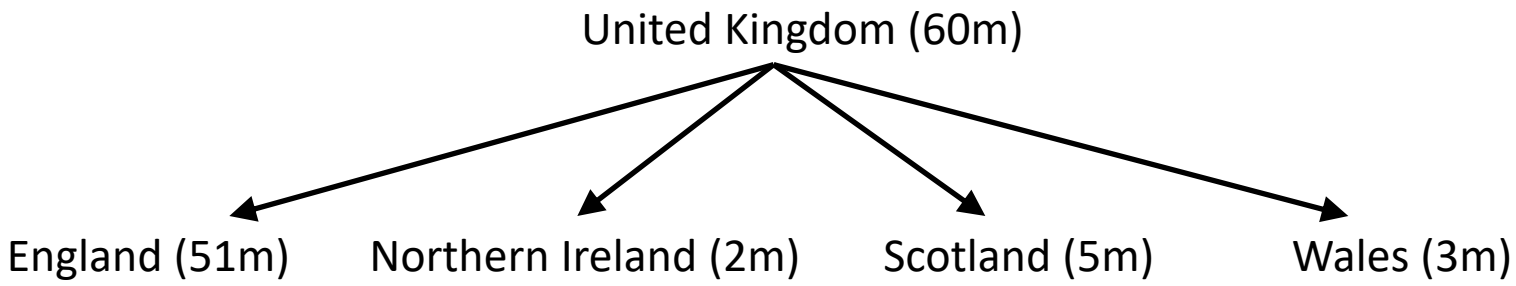
Interval size is 60/10 = 6

Select random start between 1 and 6

Select every sixth unit

Stratified Sampling

- Units/people are aggregated into subgroups called strata. A certain number of units are sampled from each stratum.



- Guards against unusual samples
- Stratification information has to be available

Stratified Sampling

Proportionate Stratification

- Chance of inclusion in sample is same for all units/people regardless of strata

	Population	Sample Size	Sampling Fraction
United Kingdom	61 000 000	6 100	0.01%
England	51 000 000	5 100	0.01%
Northern Ireland	2 000 000	200	0.01%
Scotland	5 000 000	500	0.01%
Wales	3 000 000	300	0.01%

Stratified Sampling

Disproportionate Stratification

- Chance of a unit/person being included in the sample depends on the strata they are in.
- Often used to target small sub groups to help analysis

	Population	Sample Size	Sampling Fraction
United Kingdom	61 000 000	6 100	0.01%
England	51 000 000	3 100	0.01%
Northern Ireland	2 000 000	1 000	0.05%
Scotland	5 000 000	1 000	0.02%
Wales	3 000 000	1 000	0.03%

Cluster Sampling

What if there are a small number of units across a large area?

- Divide population into clusters along geographic boundaries (e.g. wards)
- Randomly sample clusters
- Measure all units within selected clusters
- Reduces costs for face to face interviews
- But bias can be a problem if what you want to measure depends of geographic location.

Cluster Sampling

Example – Select 100 random postcodes within Edinburgh and interview all households within the 100 postcodes

- Reduces costs for face to face interviews
- But bias can be a problem if what you want to measure depends of geographic location.

Multistage Sampling

- Larger units are randomly selected
- Smaller units within the larger unit are then randomly selected

- Examples – Edinburgh randomly selects 100 postcodes and then selects 10 households within each postcode

Multistage Sampling

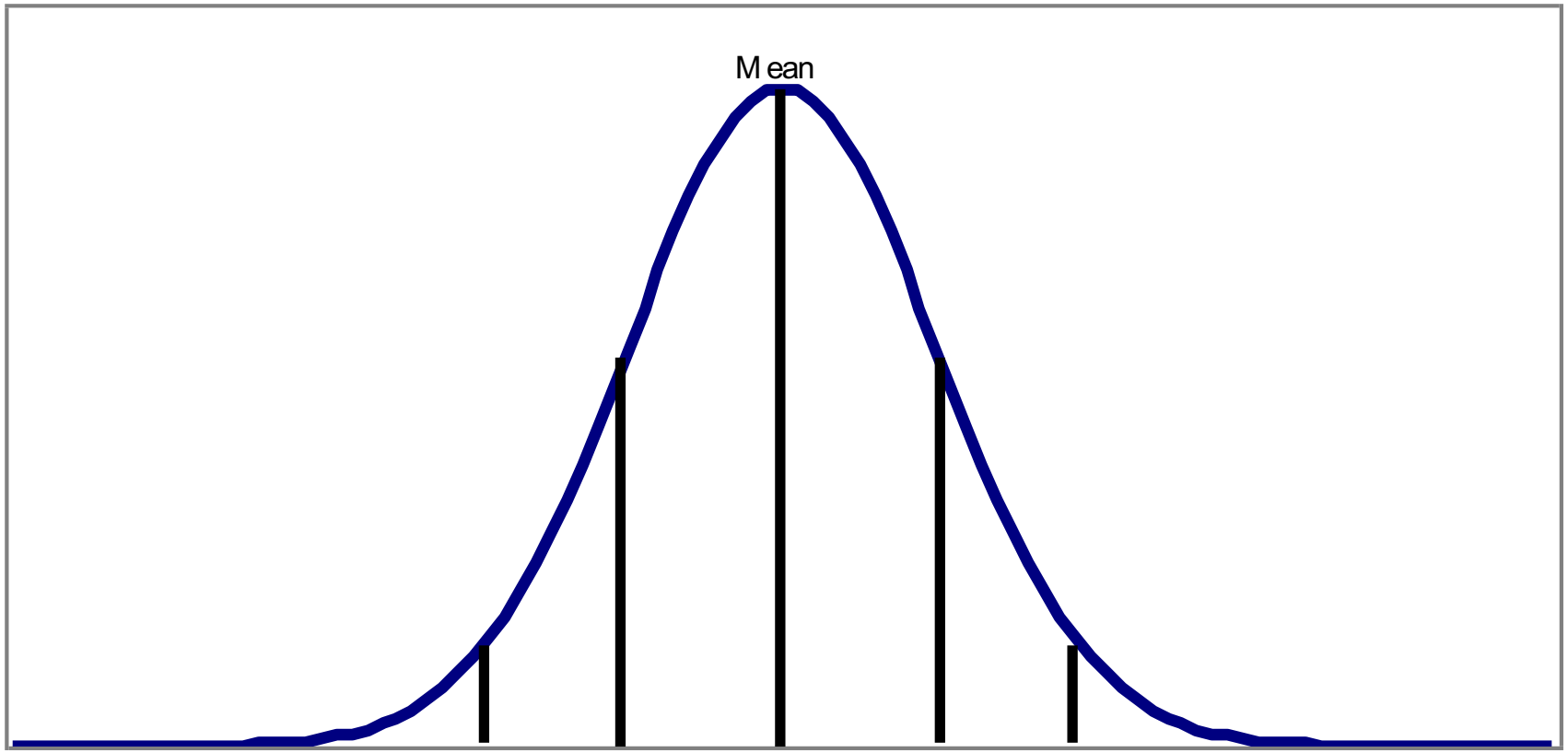
- Can be useful when no sampling frame is available
- Reduces costs for face to face interviews
- Similar bias problems to clustering

Why samples work

Central limit Theorem

The distribution of a sample mean will tend to the normal distribution as sample size increases, regardless of the population distribution

Normal Distribution



Basics of Sampling Theory

$P = \{ x_1, x_2, \ldots\ldots, x_N \}$

where P = population

$x_1, x_2, \ldots\ldots, x_N$ are real numbers

Assuming x is a random variable;

<u>Mean/Average of x</u>,

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Basics of Sampling Theory

Standard Deviation,

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{N} (x_i - x)^{2-}}{N}}$$

Variance,

$$\sigma_x^2 = \frac{\sum_{i=1}^{N} (x_i - x)^{2-}}{N}$$

Basics of Sampling Theory

Theorem About Mean

picking random numbers x, mean = $\bar{x}$
picking random numbers y, mean = $\bar{y}$

$\bar{x} = \bar{y}$

Picking another number z,

mean z = $\bar{x} = \bar{y}$

$z = c_1x + c_2y$; c_1, c_2 are constants

$z = x + y$

$\bar{z}$ $\bar{x}$ $\bar{y}$

Basics of Sampling Theory

Independence

two events are independent if the occurrence of one of the events gives no information about whether or not the other event will occur; that is, the events have no influence on each other

for example a, b and c are independent if:

- a and b are independent; a and c are independent; and b and c are independent

Basics of Sampling Theory

Theorem About Variances/Sampling Theorem

$z = (x + y)/2$; $\sigma_z^2 = ?$ $\sigma_z^2 < \sigma_x^2$

Taking, $z = (x + y)/2$

$\sigma_z^2 = (\sigma_x^{2+} \sigma_y^2)/4$

Taking k sample, $z = (x + x' + x'' + ... + x''^{\ldots k})/k$

$\sigma_z^2 = (k\sigma_x^2)/k^2$

$\sigma_z^2 = \sigma_x^2)/k$

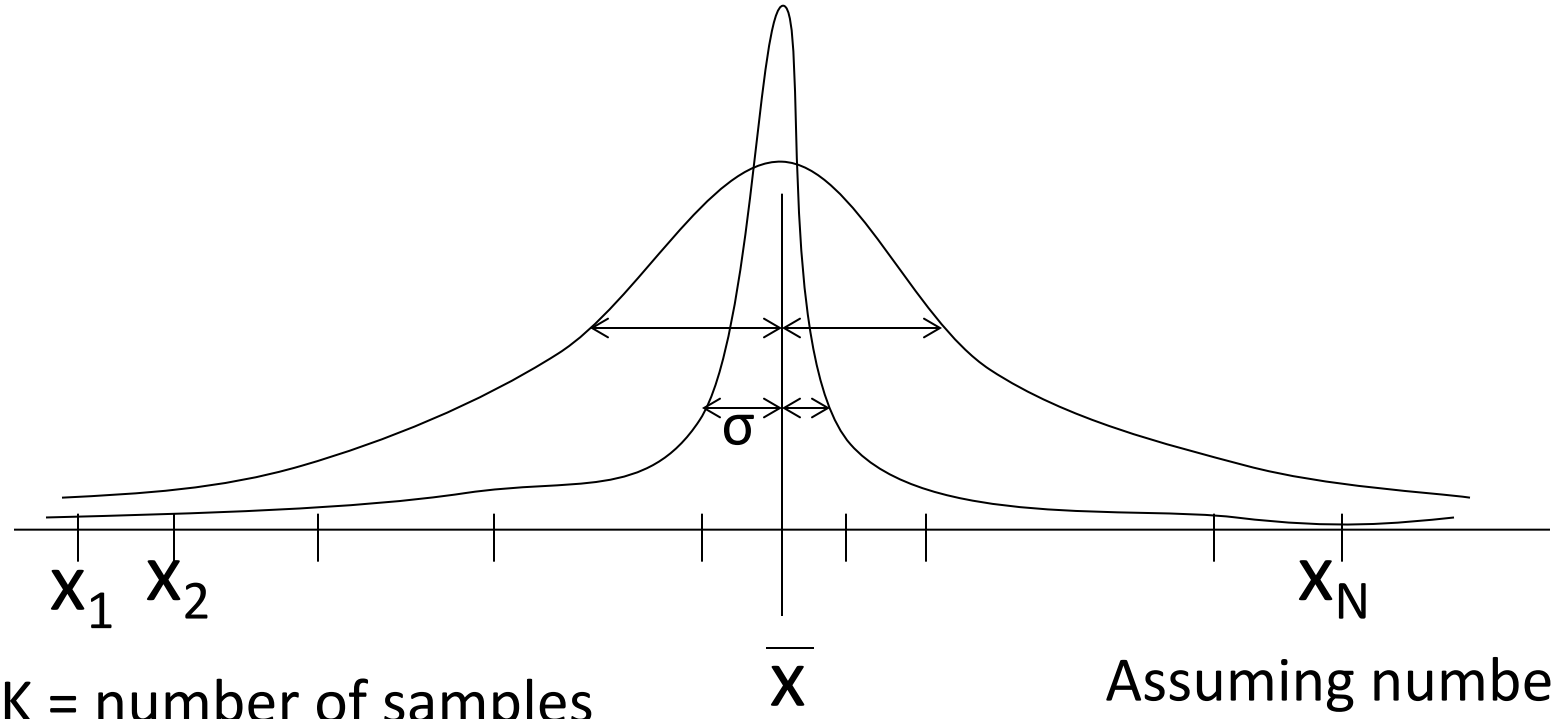
* This theory works only on independent variables; while mean theorem works on dependent variable

* Error depends on number of samples; bigger sample – less error; smaller sample – more error

* This formula is true for sampling with replacement

Basics of Sampling Theory

Normal Distribution curve



- K = number of samples
- $\bar{z}$ = sample mean
- as k increases, z $\bar{c}$omes closer to $\bar{x}$

Assuming numbers are sorted

Sample and population (ASW, 15)

- A **population** is the collection of all the elements of interest.
- A **sample** is a subset of the population.
  - Good or bad samples.
  - Representative or non-representative samples. A researcher hopes to obtain a sample that represents the population, at least in the variables of interest for the issue being examined.
  - Probabilistic samples are samples selected using the principles of probability. This may allow a researcher to determine the sampling distribution of a sample statistic. If so, the researcher can determine the probability of any given sampling error and make statistical inferences about population characteristics.

Why sample?

- Time of researcher and those being surveyed.
- Cost to group or agency commissioning the survey.
- Confidentiality, anonymity, and other ethical issues.
- Non-interference with population. Large sample could alter the nature of population, eg. opinion surveys.
- Do not destroy population, eg. crash test only a small sample of automobiles.
- Cooperation of respondents – individuals, firms, administrative agencies.
- Partial data is all that is available, eg. fossils and historical records, climate change.

Methods of sampling – nonprobabilistic

- Friends, family, neighbours, acquaintances.
- Students in a class or co-workers in a workplace.
- Convenience (ASW, 286).
- Volunteers.
- Snowball sample.
- Judgment sample (ASW, 286).
- Quota sample – obtain a cross-section of a population, eg. by age and sex for individuals or by region, firm size, and industry for businesses. This may be reasonably representative.
- Sampling distribution of statistics cannot be obtained using any of the above methods, so statistical inference is not possible.

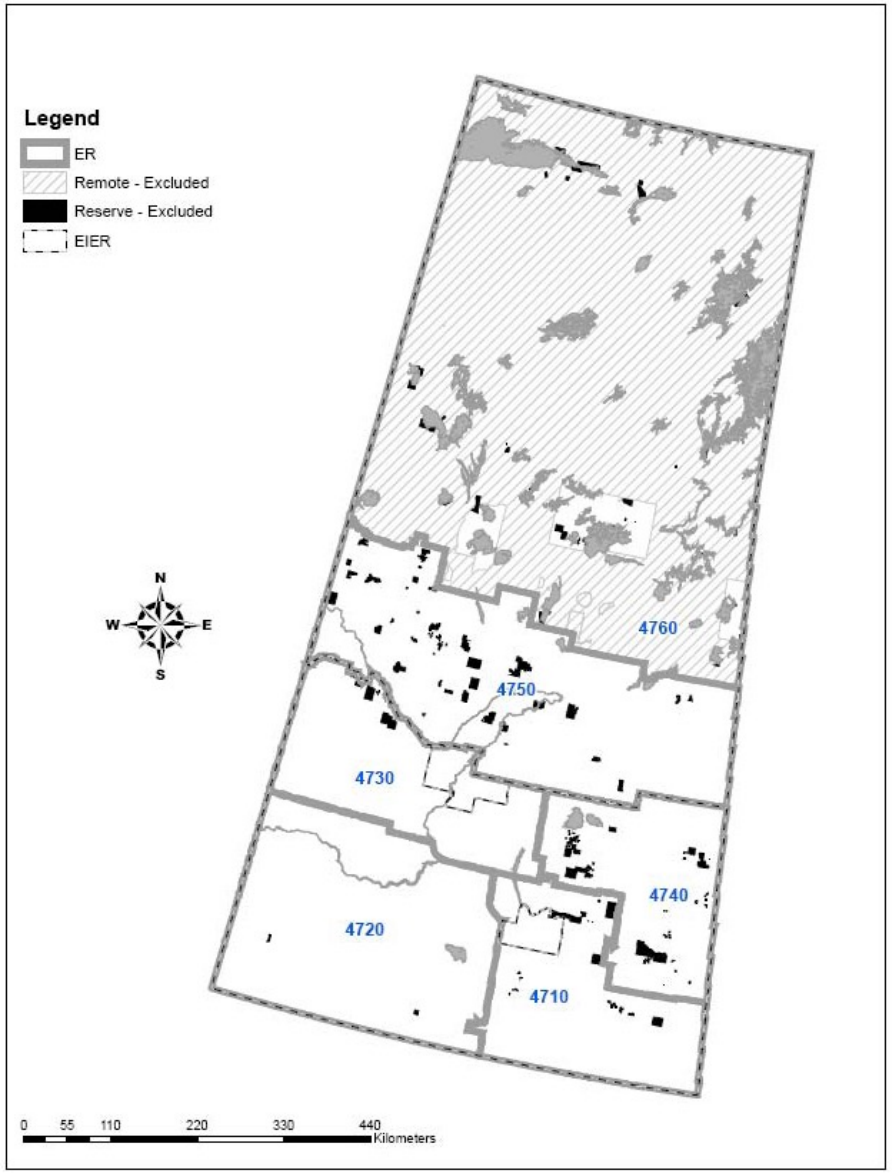
Methods of sampling – probabilistic

- Random sampling methods – each member has an equal probability of being selected.
- Systematic – every k^{th} case. Equivalent to random if patterns in list are unrelated to issues of interest. Eg. telephone book.
- Stratified samples – sample from each stratum or subgroup of a population. Eg. region, size of firm.
- Cluster samples – sample only certain clusters of members of a population. Eg. city blocks, firms.
- Multistage samples – combinations of random, systematic, stratified, and cluster sampling.
- If probability involved at each stage, then distribution of sample statistics can be obtained.

Map of Economic Regions in Saskatchewan for strata used in the monthly Labour Force Survey.

Source: Statistics Canada, catalogue number 71-526-X.

Clusters and individuals are selected from each of the 5 southern economic regions. In addition, the two CMAs of Regina and Saskatoon are strata. Note that the north of the province is treated as a remote region. Remote regions and Indian Reserves are not sampled in the Survey.



Some terms used in sampling

- **Sampled population** – population from which sample drawn (ASW, 258). Researcher should clearly define.
- **Frame** – list of elements that sample selected from (ASW, 258). Eg. telephone book, city business directory. May be able to construct a frame.
- **Parameter** – characteristics of a population (ASW, 259). Eg. total (annual GDP or exports), proportion *p* of population that votes Liberal in federal election. Also, μ or σ of a probability distribution are termed parameters.
- **Statistic** – numerical characteristics of a sample. Eg. monthly unemployment rate, pre-election polls.
- **Sampling distribution** of a statistic is the probability distribution of the statistic.

Selecting a sample (ASW, 259-261)

- *N* is the symbol given for the size of the population or the number of elements in the population.
- *n* is the symbol given for the size of the sample or the number of elements in the sample.
- **Simple random sample** is a sample of size *n* selected in a manner that each possible sample of size *n* has the same probability of being selected.
- In the case of a random sample of size *n* = 1, each element has the same chance of being selected.

Selecting a simple random sample

- **Sample with replacement** – after any element randomly selected, replace it and randomly select another element. But this could lead to the same element being selected more than once.
- More common to **sample without replacement**. Make sure that on each stage, each element remaining in the population has the same probability of being selected.
- Use a random number table or a computer generated random selection process. Or use a coin, die, or bingo ball popper, etc.

Simple random sample of size 2 from a population of 4 elements – without replacement

Population elements are A, B, C, D. N=4, n=2.

1st element selected could be any one of the 4 elements and this leaves 3, so there are 4 x 3 = 12 possible samples, each equally likely: AB, AC, AD, BA, BC, BD, CA, CB, CD, DA, DB, DC.

$$P_n^N = \frac{N!}{(N-n)!} = \frac{4!}{(4-2)!} = 12$$

If the order of selection does not matter (ie. we are interested only in what elements are selected), then this reduces to 6 combination. If {AB} is AB or BA, etc., then the equally likely random samples are {AB}, {AC}, {AD}, {BC}, {BD}, {CD}. This is the number of combinations (ASW, 261, note 1).

$$C_n^N = \frac{N!}{n!(N-n)!} = \frac{4!}{2!(4-2)!} = 6$$

First N = 18 companies

on US 200 list

1. 3M
2. Abbott
3. Adobe
4. Aetna
5. Aflac
6. Air products
7. Alcoa
8. Allergan
9. Allstate
10. Alfria
11. Amazon
12. American Electric
13. American Express
14. American Tower
15. Amgen
16. Andarko
17. Anheuser Busch
18. Apache

Using random number table

Part of Table 7.1:

71744 51102 15141

95436 79115 08303

Suppose you were asked to select a simple random sample of size n =5.

Since 18 cases, two digits required and, in order, these are: 71 74 45 11 02 15 14 19 54 36 79 11 50 83 03.

Select cases 11, 2, 15, 14, and 3.

Keep track of where you last used the table and begin the next selection at that point.

Using Excel(ASW, 292)

- Suppose the data are in rows 2 through 46 in columns A through H.
- To arrange the rows in random order
  - Enter =RAND() in H2
  - Copy cell H2 to cells H3:H46 and each cell has a random number assigned – these later change
  - Select any cell in H
  - For Excel 2003, click **Data**, then **Sort**, and **Sort by Ascending**.
  - For Excel 2007, on the **Home** tab, in the **Editing** group, click **Sort and Filter** and **Sort Smallest to Largest.**
- The rows are now in random order. For a random sample of size *n*, select the data in the first *n* rows.

Sampling from a process (ASW, 261)

- It my be difficult or impossible or to obtain or construct a frame.
  - Larger or potentially infinite population – fish, trees, manufacturing processes.
  - Continuous processes – production of milk or other liquids, transporting commodities to a warehouse.
- Random sample is one where any element selected in the sample:
  - Is selected independently of any other element.
  - Follows the same probability distribution as the elements in the population.
- Careful design for sample is especially important.
  - Sample production of milk at random times.
  - Forest products – randomly select clusters from maps or previous surveys of tree types, size, etc.

Point Estimation (ASW, 263)

- gg

Measure	Parameter	Statistic or point estimator	Sampling error
Mean	μ	$\bar{x}$	$\lvert\bar{x}-\mu\rvert$
Standard deviation	σ	s	$\lvert s-\sigma\rvert$
Proportion	*p*	$\bar{p}$	$\lvert\bar{p}-p\rvert$
No. of elements	*N*	*n*	

The proportion is the frequency of occurrence of a characteristic divided by the total number of elements. The proportion of elements of a population that take on the characteristic is *p* and the proportion of the elements in the sample selected with this same characteristic is $\bar{p}$.

Terms for estimation

- **Parameters** are characteristics of a **population** or, more specifically, a **target population** (ASW, 265). Parameters may also be termed **population values**.
- A **statistic** is also referred to as a **sample statistic** or, when estimating a parameter, a **point estimator** of a parameter. A specific value of a point estimator is referred to as a **point estimate** of a parameter.
- The **sampling error** is the difference between the point estimate (value of the estimator) and the value of the parameter. This is the error caused by sampling only a subset of elements of a population, rather than all elements in a population. A researcher hopes to minimize the sampling error, but all samples have some such error associated with them.

Percentage of respondents, votes, and number of seats by party, November 5, 2003 Saskatchewan provincial election

Political Party	CBC Poll, Oct. 20-26 $\overline{P}$	Cutler Poll, Oct. 29 – Nov. 5 $\overline{P}$	Election Result P	Number of Seats
NDP	42%	47%	44.5%	30
Saskatchewan Party	39%	37%	39.4%	28
Liberal	18%	14%	14.2%	0
Other	1%	2%	1.9%	0
Total	100%	100%	100.0%	58
Undecided	15%	16%		
Sample size *(n)*	800	773		

Sources: CBC Poll results from Western Opinion Research, "Saskatchewan Election Survey for The Canadian Broadcasting Corporation," October 27, 2003. Obtained from web site. http://sask.cbc.ca/regional/servlet/View?filename=poll_one031028, November 7, 2003. Cutler poll results provided by Fred Cutler and from the *Leader-Post*, November 7, 2003, p. A5.

Sampling error in Saskatchewan polls

The actual results from the election are provided in the last two columns, with the second last column giving the parameters for the population. These are percentages, rather than proportions, so I have labelled them as upper case P. The second and third columns provide statistics on point estimators $\overline{P}$ of P from two different polls. For any party, the difference between these two provides a measure of the sampling error.

For example, the Cutler Poll has a sampling error of only 0.2 percentage points for the Liberals, but a sampling error of 2.4 percentage points for the Saskatchewan Party.

Sampling distributions

- A **sampling distribution** is the probability distribution for all possible values of the sample statistic.
- Each sample contains different elements so the value of the sample statistic differs for each sample selected. These statistics provide different estimates of the parameter. The sampling distribution describes how these different values are distributed.
- For the most part, we will work with the sampling distribution of the sample mean. With the sampling distribution of $\bar{x}$, we can "make probability statements about how close the sample mean is to the population mean μ" (ASW, 267). Alternatively, it provides a way of determining the probability of various levels of sampling error.

Sampling distribution of the sample mean

- When a sample is selected, the sampling method may allow the researcher to determine the sampling distribution of the sample mean $\bar{x}$. The researcher hopes that the mean of the sampling distribution will be μ, the mean of the population. If this occurs, then the expected value of the statistic $\bar{x}$ is μ. This characteristic of the sample mean is that of being an unbiased estimator of μ. In this case,

$$E(\bar{x}) = \mu$$

- If the variance of the sampling distribution can be determined, then the researcher is able to determine how variable $\bar{x}$ is when there are repeated samples. The researcher hopes to have a small variability for the sample means, so most estimates of μ are close to μ.

Sampling distribution of the sample mean when random sampling

- If a simple random sample is drawn from a normally distributed population, the sampling distribution of $\bar{x}$ is normally distributed (ASW, 269).
- The mean of the distribution of $\bar{x}$ is μ, the population mean.
- If the sample size *n* is a reasonably small proportion of the population size, then the standard deviation of $\bar{x}$ is the population standard deviation σ divided by the square root of the sample size. That is, samples that contain, say, less than 5% of the population elements, the **finite population correction factor** is not required since it does not alter results much (ASW, 270).

Random sample from a normally distributed population

	Normally distributed population	Sampling distribution of $\bar{x}$ when sample is random
No. of elements	*N*	*n*
Mean	μ	μ
Standard deviation	σ	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Note: If $n/N > 0.05$, it may be best to use the finite population correction factor (ASW, 270).

Central limit theorem – CLT (ASW, 271)

The sampling distribution of the sample mean, $\bar{x}$, is approximated by a normal distribution when the sample is a simple random sample and the sample size, *n*, is large.

In this case, the mean of the sampling distribution is the population mean, μ, and the standard deviation of the sampling distribution is the population standard deviation, σ, divided by the square root of the sample size. The latter is referred to as the **standard error** of the mean.

A sample size of 100 or more elements is generally considered sufficient to permit using the CLT. If the population from which the sample is drawn is symmetrically distributed, $n > 30$ may be sufficient to use the CLT.

Large random sample from any population

	Any population	Sampling distribution of $\bar{x}$ when sample is random
No. of elements	*N*	*n*
Mean	μ	μ
Standard deviation	σ	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

A sample size n of greater than 100 is generally considered sufficiently large to use.

Simulation example

- 192 random samples from population that is not normally distributed.
- Sample size of $n = 50$ for each of the random samples.

- Handouts in Monday's class provide these results.

Sampling distribution in theory and practice

- Population mean μ = 2352 and standard deviation σ = 1485.
- Random sample of size *n = 50.*
- Sample mean $\bar{x}$ is normally distributed with a mean of μ = 2352 and a standard deviation, or standard error, of

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1485}{\sqrt{50}} = \frac{1485}{7.071} = 210$$

In the simulation, the mean of the 192 random samples is 2337 and the standard deviation is 206.

Sampling Theory

Determining the distribution of Sample statistics

Sampling Theory
sampling distributions

Note:It is important to recognize the dissimilarity (variability) we should expect to see in various samples from the same population.

- It is important that we model this and use it to assess accuracy of decisions made from samples.
- A sample is a subset of the population.
- In many instances it is too costly to collect data from the entire population.

Statistics and Parameters

A **statistic** is a numerical value computed from a sample. Its value may differ for different samples. *e.g. sample mean* $\bar{x}$*, sample standard deviation s, and sample proportion* $\hat{p}$*.*

A **parameter** is a numerical value associated with a population. Considered fixed and unchanging. *e.g. population mean* μ*, population standard deviation* σ*, and population proportion p.*

Observations on a measurement X

$$x_1, x_2, x_3, \ldots, x_n$$

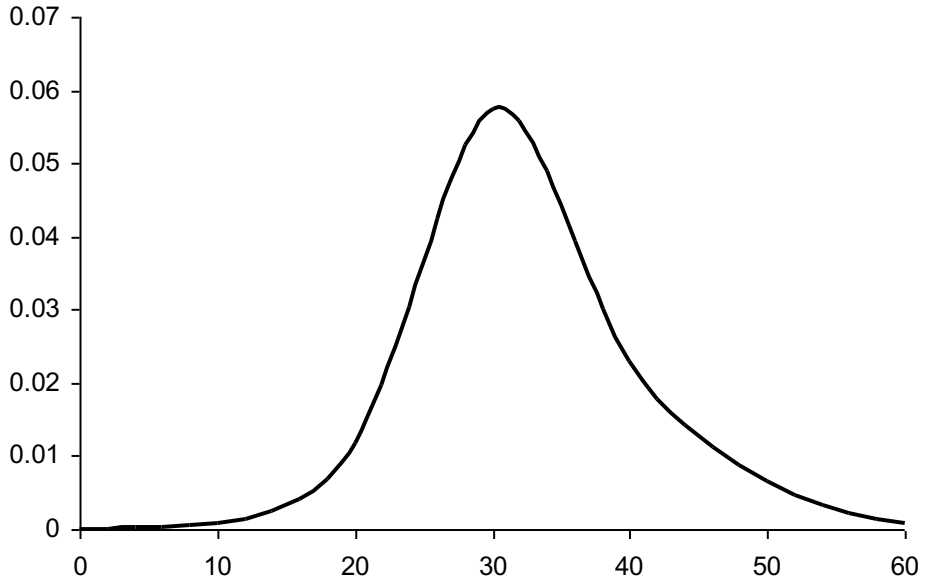
taken on individuals (cases) selected at random from a population are **random variables** prior to their observation.

The observations are numerical quantities whose values are determined by the outcome of a random experiment (the choosing of a random sample from the population).

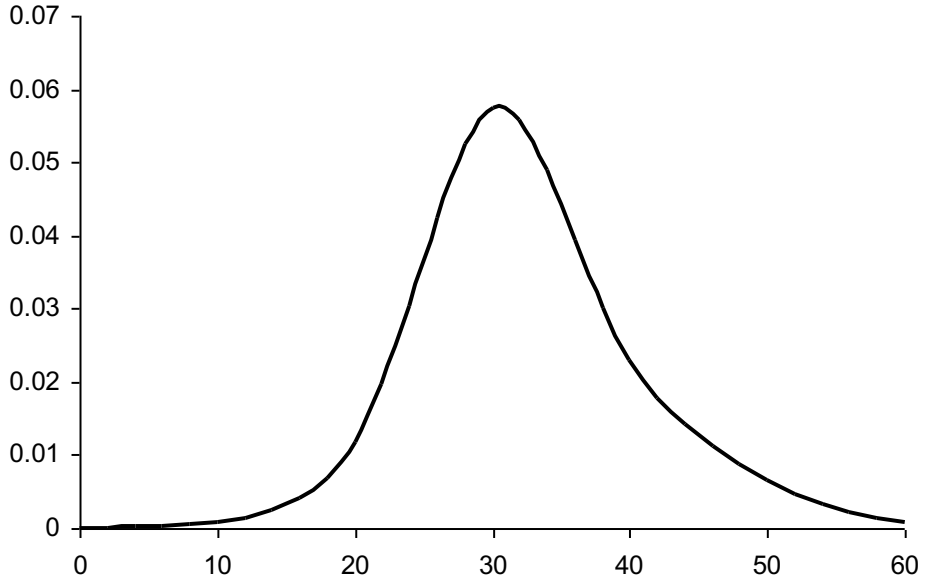
The probability distribution of the observations x_1, x_2, x_3, … , x_n

is sometimes called the **population**.

This distribution is the *smooth* histogram of the the variable *X* for the entire population



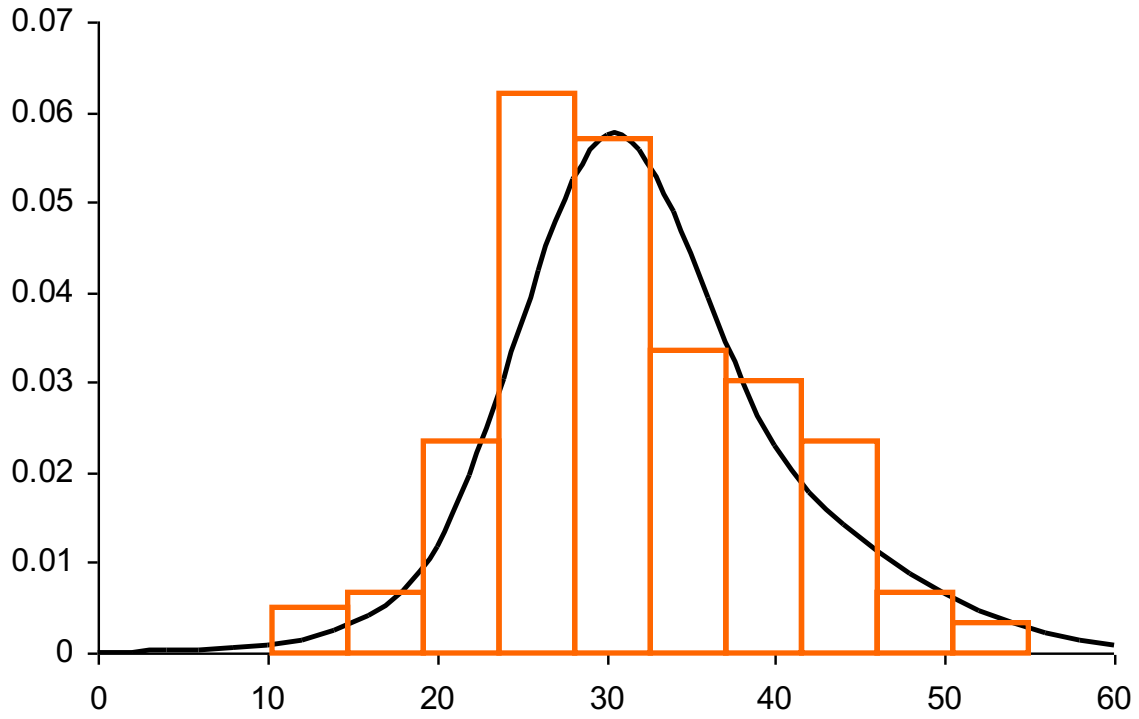
the **population** is unobserved (unless all observations in the population have been observed)



A histogram computed from the observations

$x_1, x_2, x_3, \ldots, x_n$

Gives an *estimate* of the **population**.



A **statistic** computed from the observations

$$x_1, x_2, x_3, \ldots, x_n$$

is also a **random variable** prior to observation of the sample.

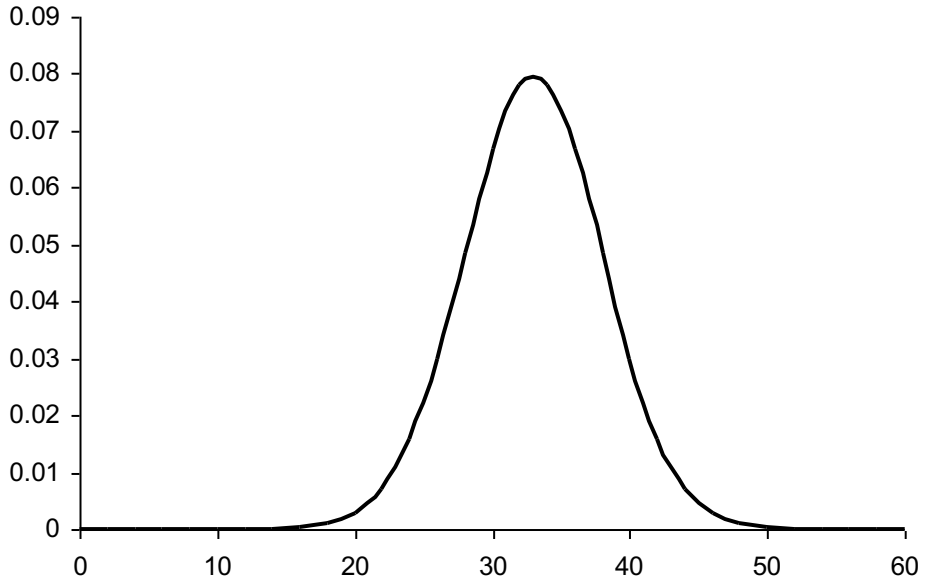
A **statistic** is also a numerical quantity whose value is determined by the outcome of a random experiment (the choosing of a random sample from the population).

The probability distribution of **statistic computed from** the observations

$$x_1, x_2, x_3, \ldots, x_n$$

is sometimes called its **sampling distribution**.

This distribution describes the random behaviour of the statistic



It is important to determine the **sampling distribution** of a statistic.

It will describe its *sampling behaviour*.

The sampling distribution will be used the assess the *accuracy* of the statistic when used for the purpose of estimation.

Sampling theory is the area of Mathematical Statistics that is interested in determining the sampling distribution of various statistics

Many statistics have a **normal** distribution.

This quite often is true if the population is Normal

It is also sometimes true if the sample size is reasonably large. (reason – **the Central limit theorem**, to be mentioned later)

Combining Random Variables

Combining Random Variables

Quite often we have two or more random variables

X, Y, Z etc

We combine these random variables using a mathematical expression.

Important question

What is the distribution of the new random variable?

Example 1: Suppose that one performs two independent tasks (A and B):

$X =$ time to perform task A (normal with mean 25 minutes and standard deviation of 3 minutes.)

$Y =$ time to perform task B (normal with mean 15 minutes and std dev 2 minutes.)

Let $T = X + Y =$ total time to perform the two tasks

What is the distribution of T?

What is the probability that the two tasks take more than 45 minutes to perform?

Example 2:

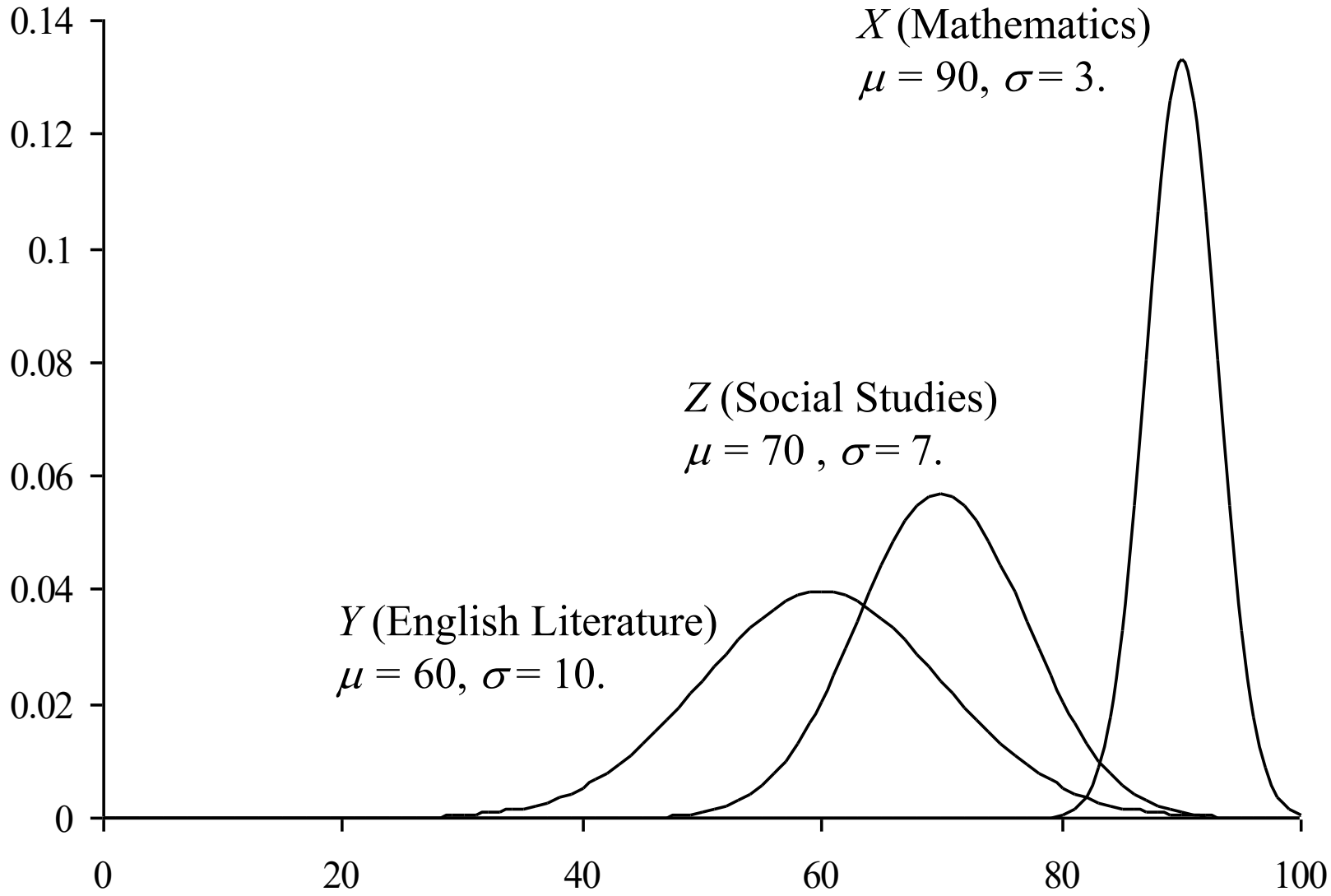
Suppose that a student will take three tests in the next three days

1. Mathematics (X is the score he will receive on this test.)
2. English Literature (Y is the score he will receive on this test.)
3. Social Studies (Z is the score he will receive on this test.)

Assume that

1. X (Mathematics) has a Normal distribution with mean $\mu = 90$ and standard deviation $\sigma = 3$.
2. Y (English Literature) has a Normal distribution with mean $\mu = 60$ and standard deviation $\sigma = 10$.
3. Z (Social Studies) has a Normal distribution with mean $\mu = 70$ and standard deviation $\sigma = 7$.

Graphs


0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100
X (Mathematics)
μ = 90, σ = 3.
Z (Social Studies)
μ = 70 , σ = 7.
Y (English Literature)
μ = 60, σ = 10.

Suppose that after the tests have been written an overall score, *S*, will be computed as follows:

$$S\,(\text{Overall score}) = 0.50\,X\,(\text{Mathematics}) + 0.30\,Y\,(\text{English Literature}) + 0.20\,Z\,(\text{Social Studies}) + 10\,(\text{Bonus marks})$$

What is the distribution of the overall score, *S*?

Sums, Differences, Linear Combinations of R.V.'s

A **linear combination** of random variables, $X, Y, \ldots$ is a combination of the form:

$$L = aX + bY + \ldots + c \text{ (a constant)}$$

where $a, b,$ etc. are numbers – positive or negative.

Most common:

Sum $= X + Y$ **Difference** $= X - Y$

Others

Averages $= {}^1/_3\, X + {}^1/_3\, Y + {}^1/_3\, Z$

Weighted averages $= 0.40\, X + 0.25\, Y + 0.35\, Z$

Sums, Differences, Linear Combinations of R.V.'s

A **linear combination** of random variables, $X, Y, \ldots$ is a combination of the form:

$$L = aX + bY + \ldots + c \text{ (a constant)}$$

where $a, b,$ etc. are numbers – positive or negative.

Most common:

Sum $= X + Y$ **Difference** $= X - Y$

Others

Averages $= {}^1/_3\, X + {}^1/_3\, Y + {}^1/_3\, Z$

Weighted averages $= 0.40\, X + 0.25\, Y + 0.35\, Z$

Means of Linear Combinations

If $L = aX + bY + \ldots + c$

The **mean of *L*** is:

$$\text{Mean}(L) = a\,\text{Mean}(X) + b\,\text{Mean}(Y) + \ldots + c$$

$$\mu_L = a\,\mu_X + b\,\mu_Y + \ldots + c$$

Most common:

$$\text{Mean}(X + Y) = \text{Mean}(X) + \text{Mean}(Y)$$

$$\text{Mean}(X - Y) = \text{Mean}(X) - \text{Mean}(Y)$$

Variances of Linear Combinations

If $X, Y, \ldots$ are ***independent*** random variables and

$$L = aX + bY + \ldots + c \text{ then}$$

$$\text{Variance}(L) = a^2 \text{ Variance}(X) + b^2 \text{ Variance}(Y) + \ldots$$

$$\sigma_L^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + \cdots$$

Most common:

$$\text{Variance}(X + Y) = \text{Variance}(X) + \text{Variance}(Y)$$

$$\text{Variance}(X - Y) = \text{Variance}(X) + \text{Variance}(Y)$$

The constant c has no effect on the variance

Example: Suppose that one performs two independent tasks (A and B):

$X =$ time to perform task A (normal with mean 25 minutes and standard deviation of 3 minutes.)

$Y =$ time to perform task B (normal with mean 15 minutes and std dev 2 minutes.)

X and Y independent so $T = X + Y =$ total time is normal with

mean $\mu = 25 + 15 = 40$

standard deviation $\sigma = \sqrt{3^2 + 2^2} = 3.6$

What is the probability that the two tasks take more than 45 minutes to perform?

$$P(T > 45) = P\left(Z > \frac{45 - 40}{3.6}\right) = P(Z > 1.39) = .0823$$

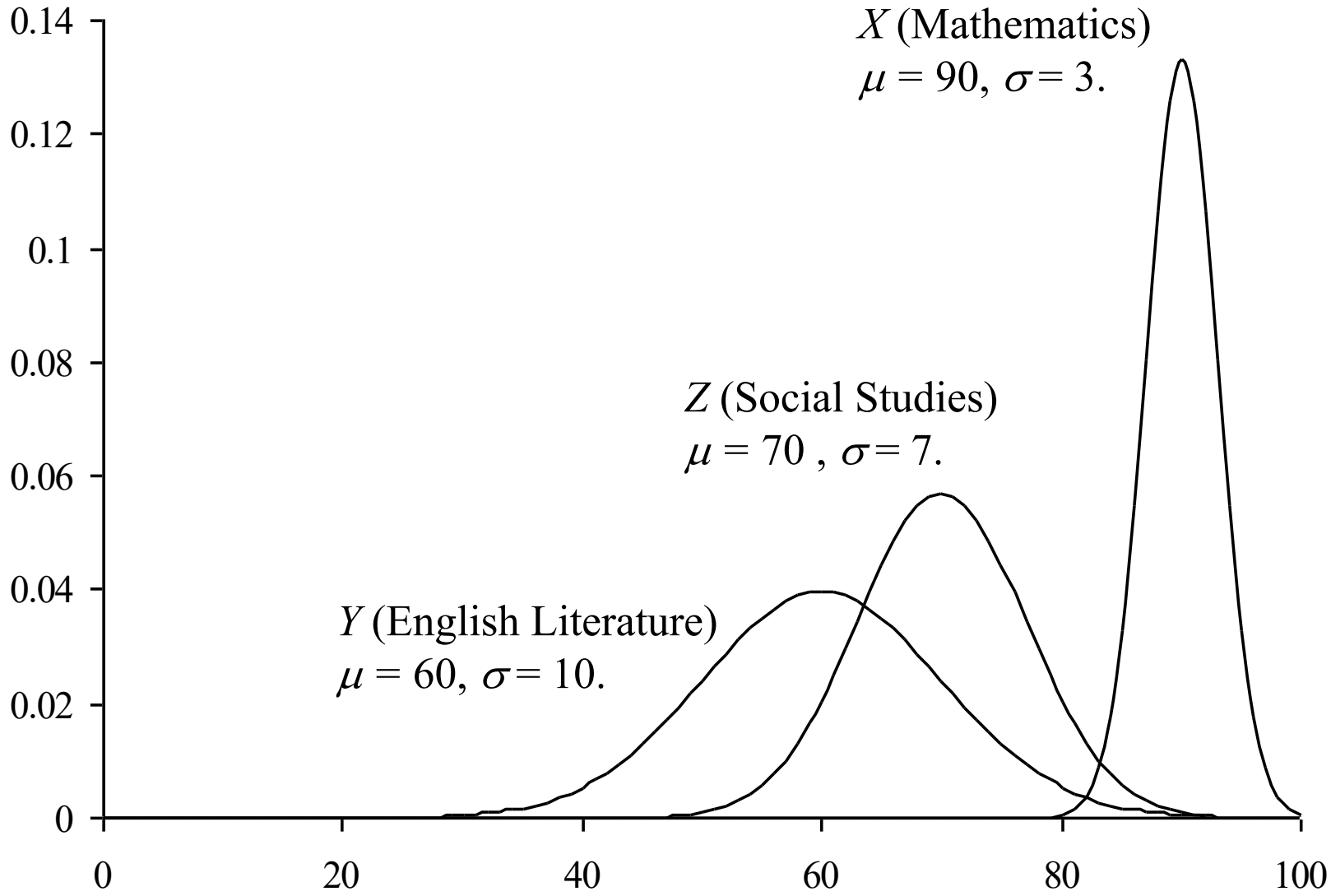
Example 2:

A student will take three tests in the next three days

1. X (Mathematics) has a Normal distribution with mean $\mu = 90$ and standard deviation $\sigma = 3$.
2. Y (English Literature) has a Normal distribution with mean $\mu = 60$ and standard deviation $\sigma = 10$.
3. Z (Social Studies) has a Normal distribution with mean $\mu = 70$ and standard deviation $\sigma = 7$.

Overall score, $S = 0.50\ X$ (Mathematics) + $0.30\ Y$ (English Literature) + $0.20\ Z$ (Social Studies) + 10 (Bonus marks)

Graphs


0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100
X (Mathematics)
μ = 90, σ = 3.
Z (Social Studies)
μ = 70 , σ = 7.
Y (English Literature)
μ = 60, σ = 10.

Determine the distribution of

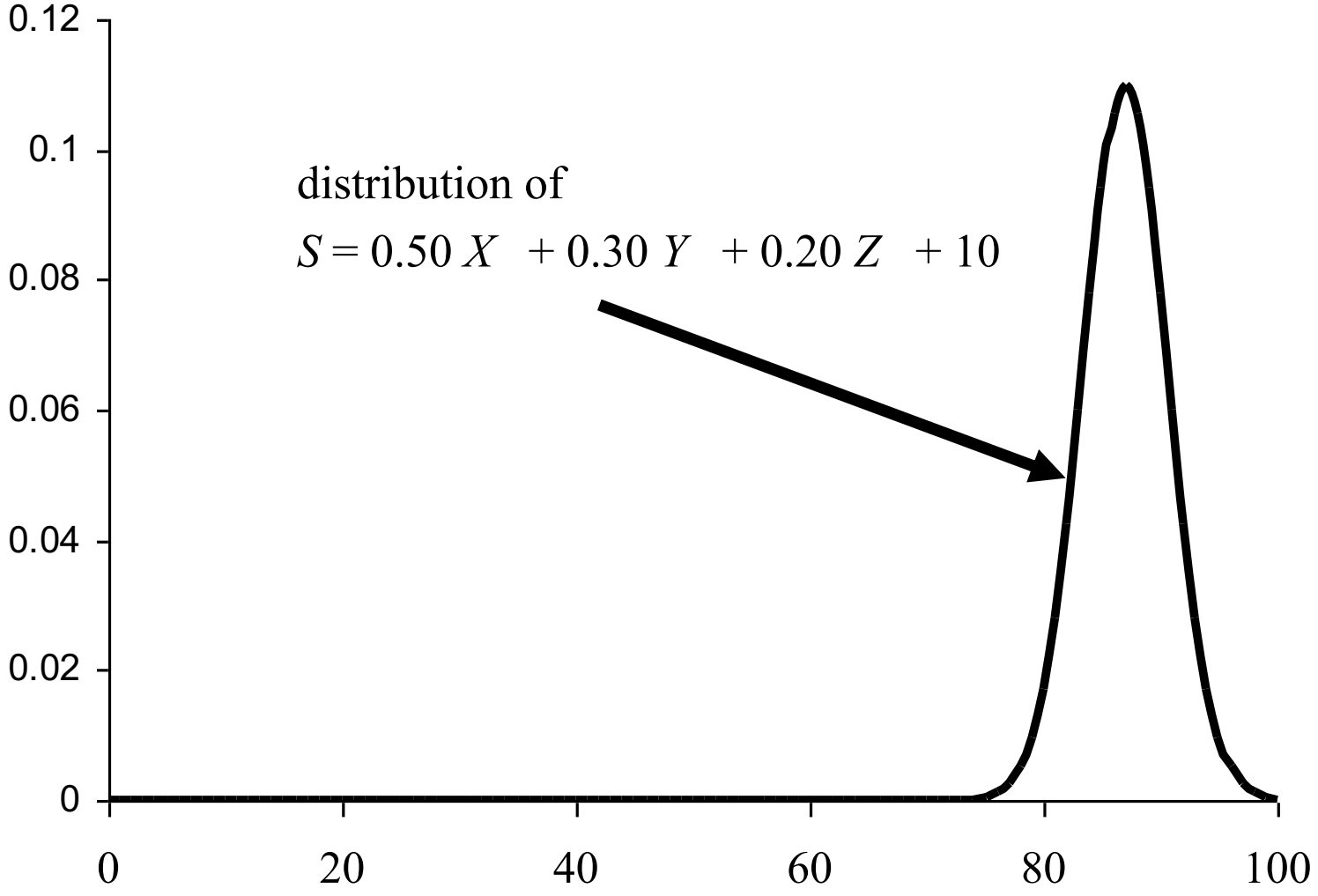
$S = 0.50\,X + 0.30\,Y + 0.20\,Z + 10$

S has a normal distribution with

$$\text{Mean } \mu_S = 0.50\,\mu_X + 0.30\,\mu_Y + 0.20\,\mu_Z + 10$$
$$= 0.50(90) + 0.30(60) + 0.20(70) + 10$$
$$= 45 + 18 + 14 + 10 = 87$$

$$\sigma_s = \sqrt{(0.5)^2\,\sigma_X^2 + (0.3)^2\,\sigma_Y^2 + (0.2)^2\,\sigma_Z^2}$$
$$= \sqrt{(0.5)^2\,3^2 + (0.3)^2\,10^2 + (0.2)^2\,7^2}$$
$$= \sqrt{2.25 + 9 + 1.96} = \sqrt{13.21} = 3.635$$

Graph


0.12
0.1
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100
distribution of
$S = 0.50\,X + 0.30\,Y + 0.20\,Z + 10$

Sampling Theory

Determining the distribution of Sample statistics

Combining Random Variables

Sums, Differences, Linear Combinations of R.V.'s

A **linear combination** of random variables, $X, Y, \ldots$ is a combination of the form:

$$L = aX + bY + \ldots + c \text{ (a constant)}$$

where $a, b,$ etc. are numbers – positive or negative.

Most common:

Sum $= X + Y$ **Difference** $= X - Y$

Others

Averages $= {}^1/_3\, X + {}^1/_3\, Y + {}^1/_3\, Z$

Weighted averages $= 0.40\, X + 0.25\, Y + 0.35\, Z$

Means of Linear Combinations

If $L = aX + bY + \ldots + c$

The **mean of *L*** is:

$$\text{Mean}(L) = a\,\text{Mean}(X) + b\,\text{Mean}(Y) + \ldots + c$$

$$\mu_L = a\,\mu_X + b\,\mu_Y + \ldots + c$$

Most common:

$$\text{Mean}(X + Y) = \text{Mean}(X) + \text{Mean}(Y)$$

$$\text{Mean}(X - Y) = \text{Mean}(X) - \text{Mean}(Y)$$

Variances of Linear Combinations

If $X, Y, \ldots$ are ***independent*** random variables and

$$L = aX + bY + \ldots + c \text{ then}$$

$$\text{Variance}(L) = a^2 \text{ Variance}(X) + b^2 \text{ Variance}(Y) + \ldots$$

$$\sigma_L^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + \cdots$$

Most common:

$$\text{Variance}(X + Y) = \text{Variance}(X) + \text{Variance}(Y)$$

$$\text{Variance}(X - Y) = \text{Variance}(X) + \text{Variance}(Y)$$

The constant c has no effect on the variance

Normality of Linear Combinations

If $X, Y, \ldots$ are ***independent*** **Normal** random variables and

$$L = aX + bY + \ldots + c$$

then L is **Normal** with

mean

$$\mu_L = a\mu_X + b\mu_Y + \ldots + c$$

and standard deviation

$$\sigma_L = \sqrt{a^2\sigma_X^2 + b^2\sigma_X^2 + \ldots}$$

In particular:

$X + Y$ is normal with mean $\mu_X + \mu_Y$

standard deviation $\sqrt{\sigma_X^2 + \sigma_Y^2}$

$X - Y$ is normal with mean $\mu_X - \mu_Y$

standard deviation $\sqrt{\sigma_X^2 + \sigma_Y^2}$

The distribution of the sample mean

The distribution of averages (the mean)

- Let $x_1, x_2, \ldots, x_n$ denote n independent random variables each coming from the same Normal distribution with mean μ and standard deviation σ.
- Let

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \left(\frac{1}{n}\right)x_1 + \left(\frac{1}{n}\right)x_2 + \cdots + \left(\frac{1}{n}\right)x_n$$

What is the distribution of $\bar{x}$?

The distribution of averages (the mean)

Because the mean is a "linear combination"

$$\mu_{\bar{x}} = \left(\frac{1}{n}\right)\mu_{x_1} + \left(\frac{1}{n}\right)\mu_{x_2} + \cdots + \left(\frac{1}{n}\right)\mu_{x_n}$$

$$= \left(\frac{1}{n}\right)\mu + \left(\frac{1}{n}\right)\mu + \cdots + \left(\frac{1}{n}\right)\mu = n\left(\frac{1}{n}\right)\mu = \mu$$

and

$$\sigma^2_{\bar{x}} = \left(\frac{1}{n}\right)^2 \sigma^2_{x_1} + \left(\frac{1}{n}\right)^2 \sigma^2_{x_2} + \cdots + \left(\frac{1}{n}\right)^2 \sigma^2_{x_n}$$

$$= \left(\frac{1}{n}\right)^2 \sigma^2 + \left(\frac{1}{n}\right)^2 \sigma^2 + \cdots + \left(\frac{1}{n}\right)^2 \sigma^2 = n\frac{\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Thus if $x_1, x_2, \ldots, x_n$ denote n independent random variables each coming from the same Normal distribution with mean μ and standard deviation σ.

Then

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \left(\frac{1}{n}\right)x_1 + \left(\frac{1}{n}\right)x_2 + \cdots + \left(\frac{1}{n}\right)x_n$$

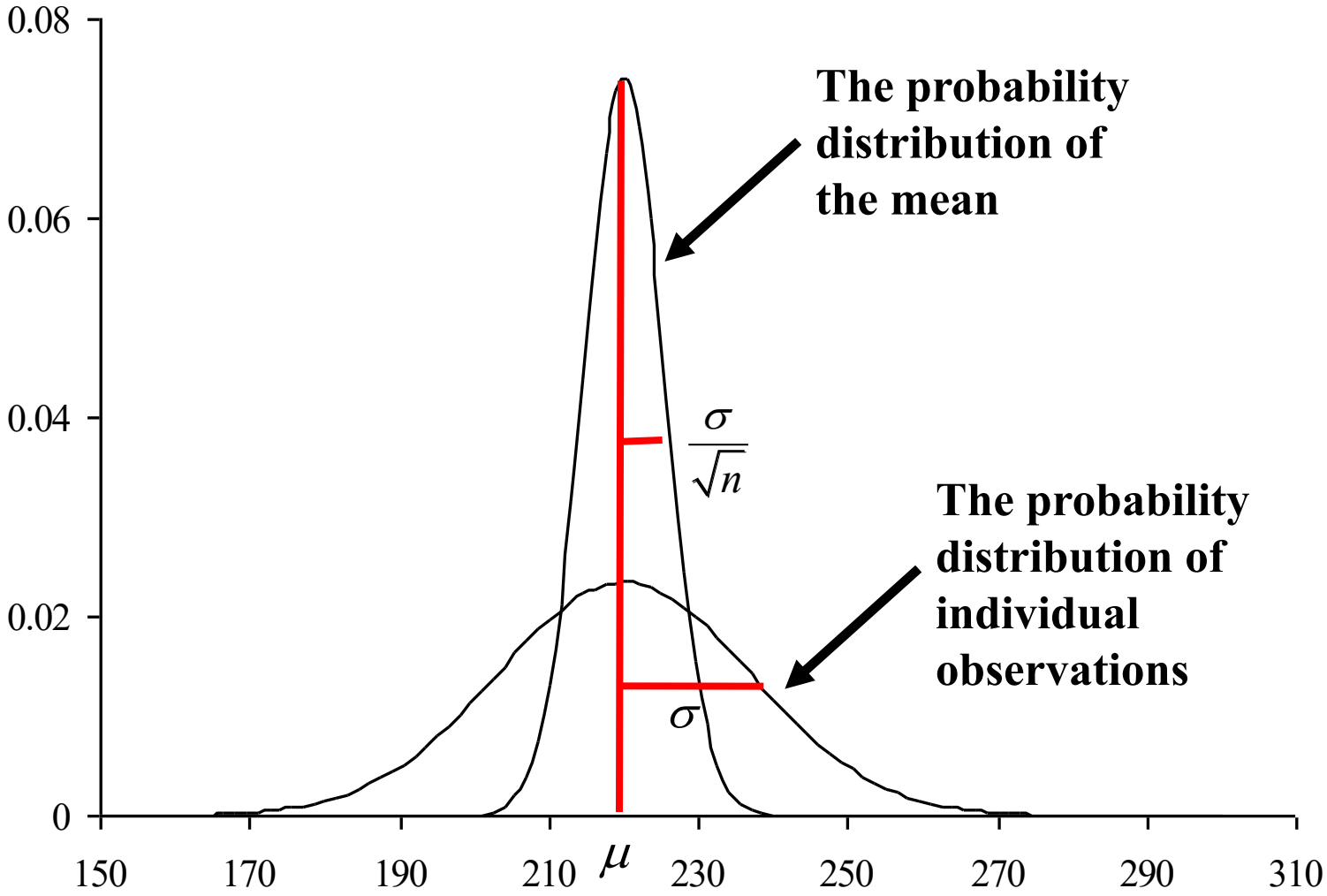
has Normal distribution with

$$\text{mean} \quad \mu_{\bar{x}} = \mu \quad \text{and}$$

$$\text{variance} \quad \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

$$\text{standard deviation} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Graphs


0.08
0.06
0.04
0.02
0
150
170
190
210
230
250
270
290
310
μ
The probability distribution of the mean
σ/√n
The probability distribution of individual observations
σ

Summary

- The distribution of the sample mean $\bar{x}$ is **Normal.**
- The distribution of the sample mean has exactly the same mean as the **population** (μ).
- The distribution of the sample mean has a **smaller standard deviation** than the **population.** $\frac{\sigma}{\sqrt{n}}$ compared to σ.

- Averaging tends to **decrease variability** $\bar{x}$
- An Excel file illustrating the distribution of the sample mean

Example

- Suppose we are measuring the cholesterol level of men age 60-65
- This measurement has a Normal distribution with mean μ = 220 and standard deviation σ = 17.
- A sample of n = 10 males age 60-65 are selected and the cholesterol level is measured for those 10 males.
- $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$, are those 10 measurements

Find the probability distribution of $\bar{x}$?

Compute the probability that $\bar{x}$ is between 215 and 225

Solution

Find the probability distribution of $\overline{x}$

$$\text{Normal with } \mu_{\overline{x}} = \mu = 220$$

$$\text{and } \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{17}{\sqrt{10}} = 5.376$$

$$P\left[215 \le \overline{x} \le 225\right]$$

$$= P\left[\frac{215-220}{5.376} \le \frac{\overline{x}-220}{5.376} \le \frac{225-220}{5.376}\right]$$

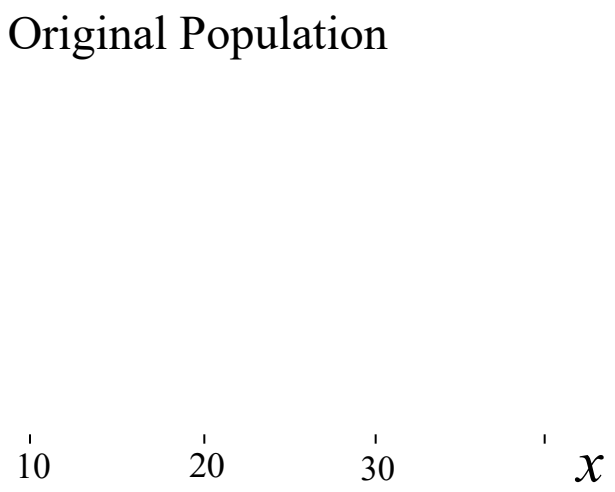
$$= P\left[-0.930 \le z \le 0.930\right] = 0.648$$

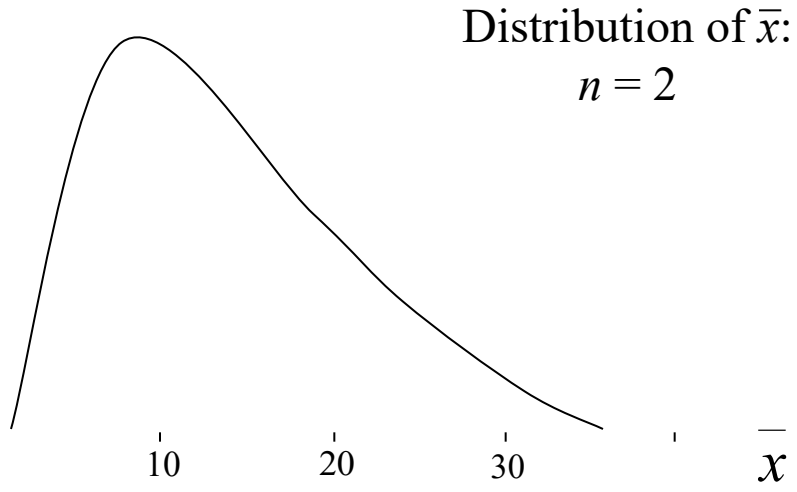
The Central Limit Theorem

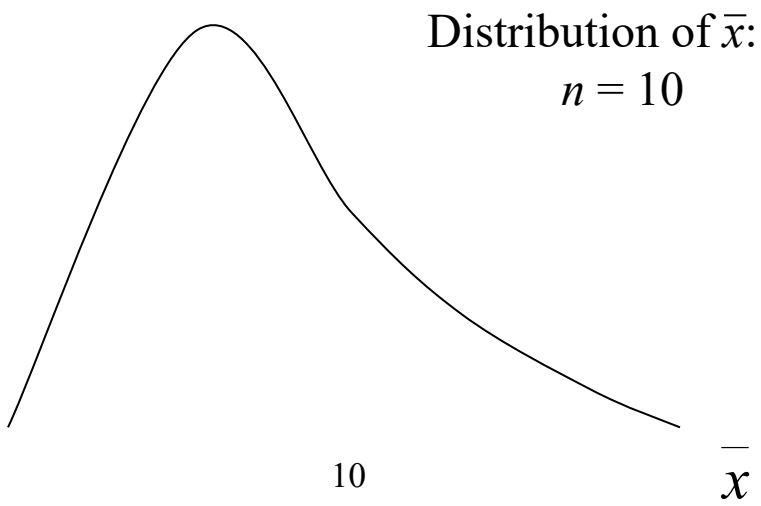
The **Central Limit Theorem (C.L.T.)** states that if n is *sufficiently large*, the **sample means** of random samples from **any** population with mean μ and finite standard deviation σ are **approximately normally distributed** with mean μ and standard deviation $\sigma / \sqrt{n}$

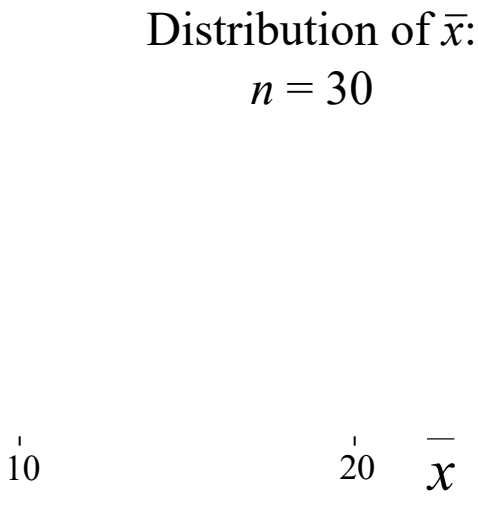
Technical Note:
The mean and standard deviation given in the CLT hold for any sample size; it is only the "approximately normal" shape that requires n to be sufficiently large.

Graphical Illustration of the Central Limit Theorem









Implications of the Central Limit Theorem

- The Conclusion that the sampling distribution of the sample mean is **Normal**, will to **true** if the sample size is large (>30). (even though the population may be non-normal).
- When the population can be assumed to be normal, the sampling distribution of the sample mean is **Normal**, will to **true** for any sample size.
- Knowing the sampling distribution of the sample mean allows to answer probability questions related to the sample mean.

Example

Example: Consider a normal population with $\mu = 50$ and $\sigma = 15$.

Suppose a sample of size 9 is selected at random. Find:

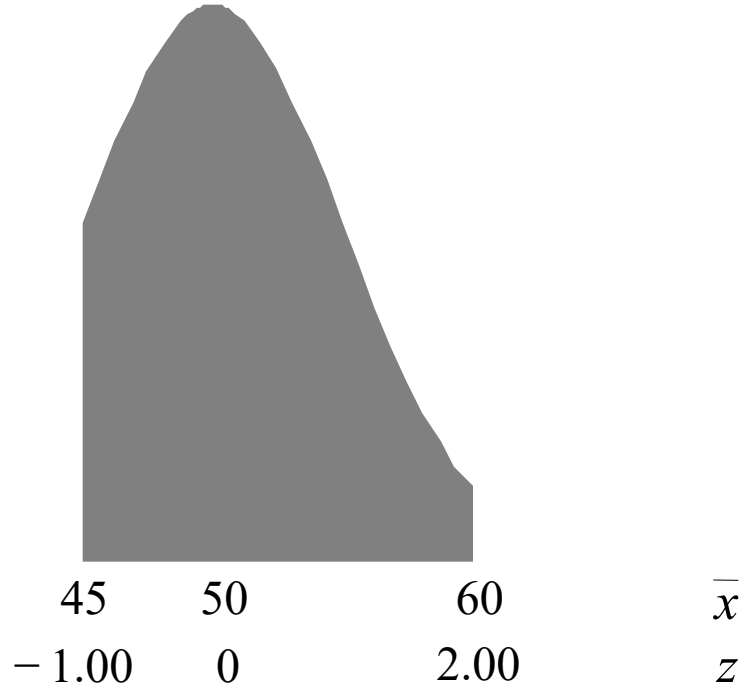
1) $P(45 \le \bar{x} \le 60)$
2) $P(\bar{x} \le 47.5)$

Solutions: Since the original population is normal, the distribution of the sample mean is also (exactly) normal

1) $\mu_{\bar{x}} = \mu = 50$

2) $\sigma_{\bar{x}} = \sigma / \sqrt{n} = 15/\sqrt{9} = 15/3 = 5$

Example

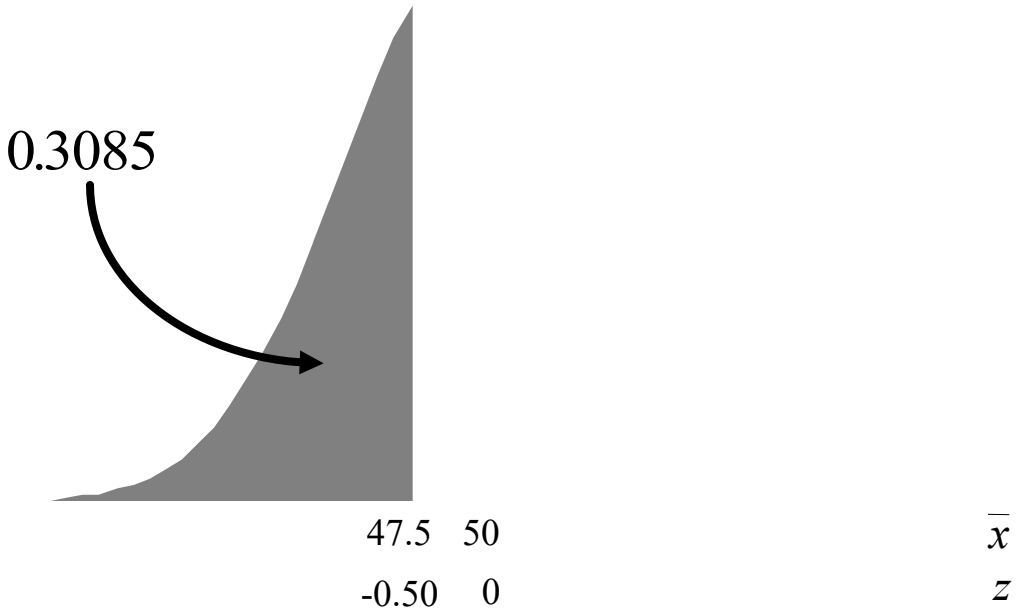


$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}; \qquad P(45 \le \bar{x} \le 60) = P\left(\frac{45-50}{5} \le z \le \frac{60-50}{5}\right)$$

$$= P(-1.00 \le z \le 2.00)$$

$$= 0.8413 - 0.0228 = 0.8185$$

Example



$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}; \qquad P(\bar{x} \le 47.5) = P\left(\frac{\bar{x} - 50}{5} \le \frac{47.5 - 50}{5}\right)$$

$$= P(z \le -.5)$$

$$= 0.5000 - 0.1915 = 0.3085$$

Example

Example: A recent report stated that the average cost of a hotel room in **Toronto** is $109/day. Suppose this figure is correct and that the standard deviation is known to be $20.

1) Find the probability that a sample of 50 hotel rooms selected at random would show a mean cost of $105 or less per day.
2) Suppose the actual sample mean cost for the sample of 50 hotel rooms is $120/day. Is there any evidence to refute the claim of $109 presented in the report?

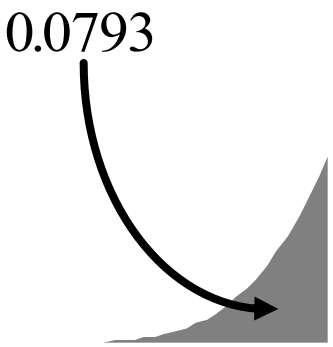
Solution:

- The shape of the original distribution is unknown, but the sample size, *n*, is large. The CLT applies.
- The distribution of $\bar{x}$ is approximately normal

$$\mu_{\bar{x}} = \mu = 109 \qquad \sigma_{\bar{x}} = \sigma / \sqrt{n} = 20 / \sqrt{50} \approx 2.83$$

Example

1)



105	109	$\bar{x}$
-1.41	0	z

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}; \qquad P(\bar{x} \le 105) = P\left(z \le \frac{105-109}{2.83}\right)$$

$$= P(z \le -1.41)$$

$$= 0.0793$$

2)

- To investigate the claim, we need to examine how *likely* an observation is the sample mean of $120

- Consider how far out in the tail of the distribution of the sample mean is $120

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}; \quad P(\bar{x} \geq 120) = P\left(z \geq \frac{120 - 109}{2.83}\right)$$

$$= P(z \geq 3.89)$$

$$= 1.0000 - 0.9999 = 0.0001$$

- Since the probability is so small, this suggests the observation of $120 is very rare (if the mean cost is really $109)

- There is evidence (the sample) to suggest the claim of $\mu = \$109$ is likely wrong

Summary

- The mean of the sampling distribution of $\bar{x}$ is equal to the mean of the original population: $\mu_{\bar{x}} = \mu$
- The standard deviation of the sampling distribution of $\bar{x}$ (also called the standard error of the mean) is equal to the standard deviation of the original population divided by the square root of the sample size:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- The distribution of $\bar{x}$ is (exactly) normal when the original population is normal
- The CLT says: the distribution of $\bar{x}$ is approximately normal regardless of the shape of the original distribution, when the sample size is large enough!

Sampling Distribution of a Sample Proportion

Sampling Distribution for Sample Proportions

Let p = population proportion of interest
or binomial probability of success.

Let

$$\hat{p} = \frac{X}{n} = \frac{\text{no. of succeses}}{\text{no. of bimomial trials}}$$

= sample proportion or proportion of successes.

Then the sampling distribution of $\hat{p}$ is approximately a normal distribution with

$$\text{mean } \mu_{\hat{p}} = p \qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Logic

Recall X = the number of successes in n trials has a **Binomial distribution** with parameters n and p (the probability of success).
Also X has approximately a **Normal distribution** with

mean $\mu = np$ and

standard deviation

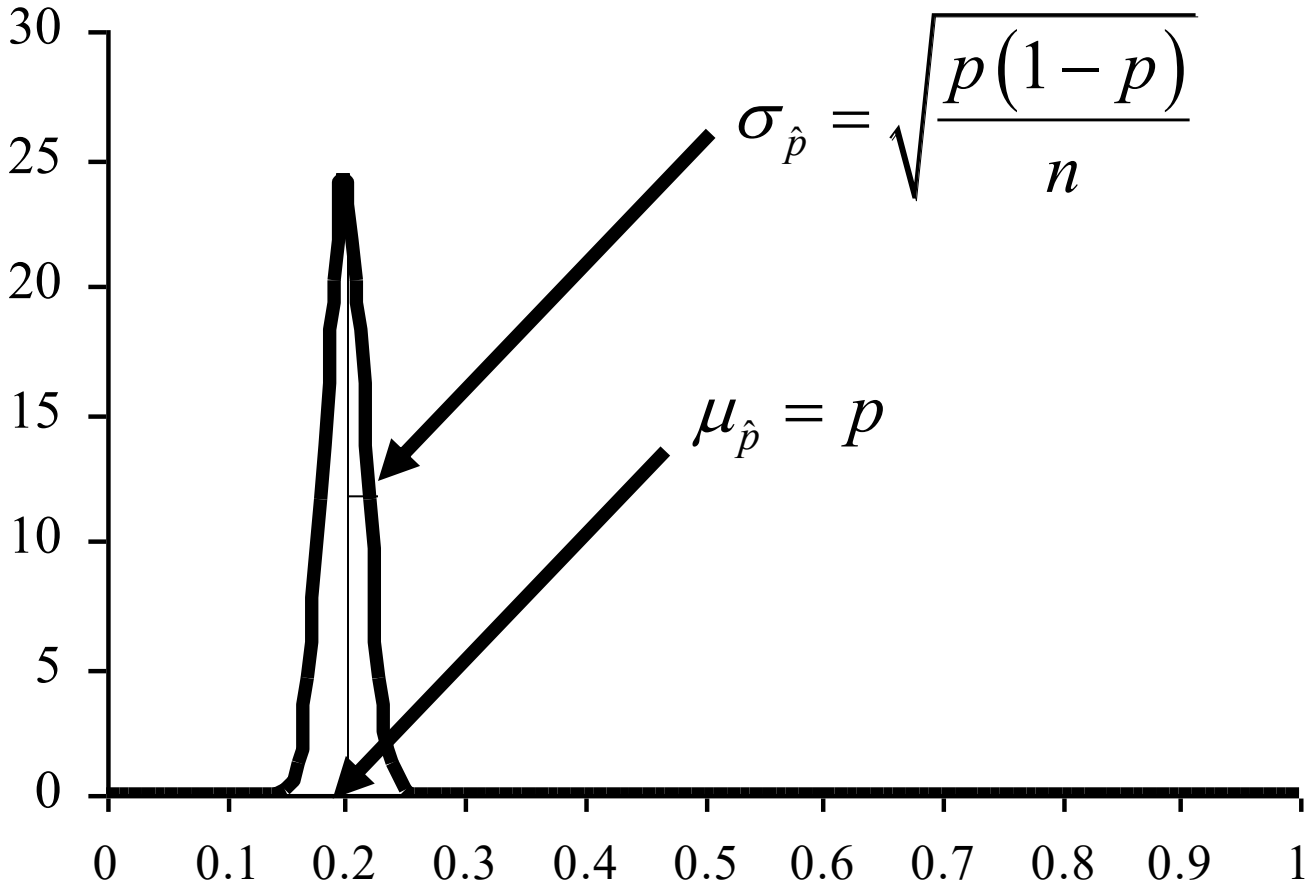
$$\sigma = \sqrt{npq} = \sqrt{np(1-p)}$$

Then the sampling distribution of $\hat{p} = \frac{X}{n} = \left(\frac{1}{n}\right)X$ is a normal distribution with

$$\text{mean } \mu_{\hat{p}} = \left(\frac{1}{n}\right)\mu = \left(\frac{1}{n}\right)np = p$$

$$\text{and } \sigma_{\hat{p}} = \left(\frac{1}{n}\right)\sigma = \left(\frac{1}{n}\right)\sqrt{np(1-p)} = \sqrt{\frac{p(1-p)}{n}}$$

Sampling distribution of $\hat{p}$



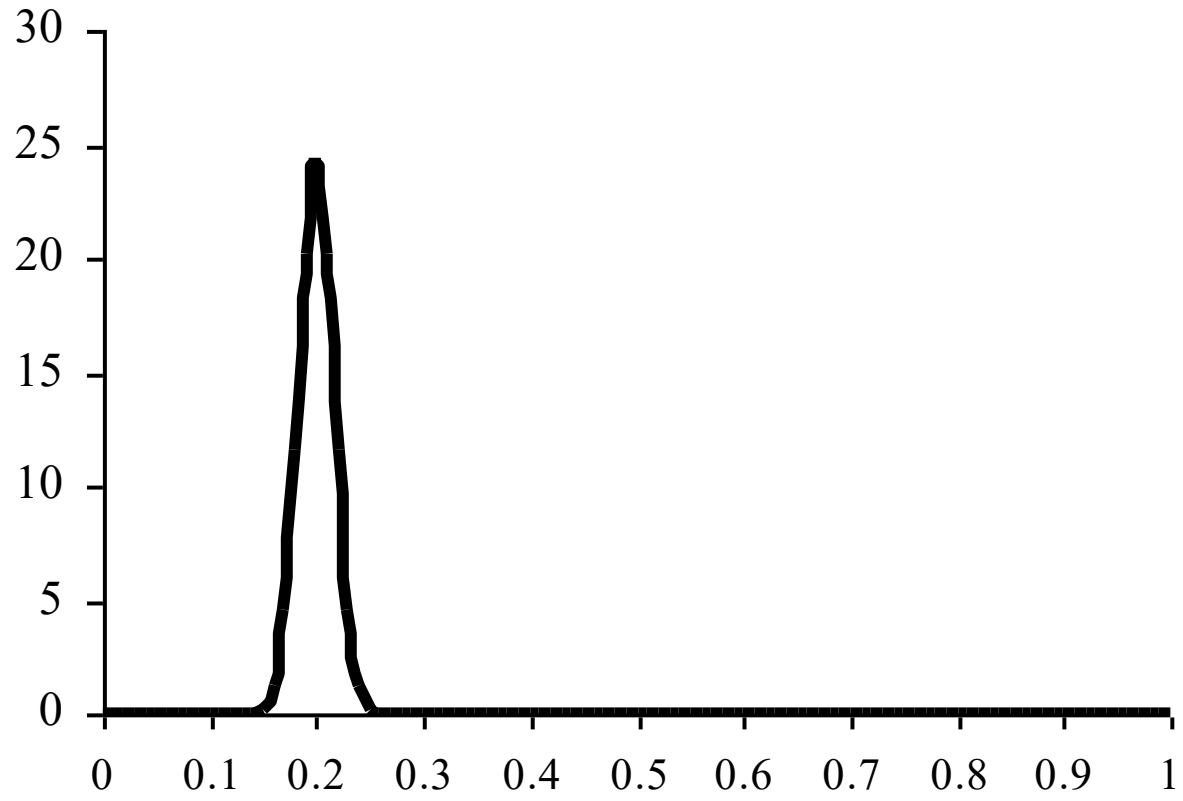
Example *Sample Proportion Favoring a Candidate*

Suppose 20% all voters favor Candidate A. Pollsters take a sample of $n = 600$ voters. Then the sample proportion who favor A will have approximately a normal distribution with

$$\text{mean } \mu_{\hat{p}} = p = 0.20$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.20(0.80)}{600}} = 0.01633$$

Sampling distribution of $\hat{p}$



Using the Sampling distribution:

Suppose 20% all voters favor Candidate A. Pollsters take a sample of $n = 600$ voters.

Determine the probability that the sample proportion will be between 0.18 and 0.22

i.e. the probability, $P\left[0.18 \le \hat{p} \le 0.22\right]$

Solution:

Recall $\mu_{\hat{p}} = p = 0.20$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.20(0.80)}{600}} = 0.01633$$

$$P\left[0.18 \le \hat{p} \le 0.22\right] = P\left[\frac{0.18-0.20}{0.01633} \le \frac{\hat{p}-0.20}{0.01633} \le \frac{0.22-0.20}{0.01633}\right]$$

$$= P\left[-1.225 \le z \le 1.225\right] = 0.8897 - 0.1103 = 0.7794$$

Sampling Theory - Summary

The distribution of sample statistics

Distribution for Sample Mean

If data is collected from a **Normal** distribution with mean μ and standard deviation σ then:

the sampling distribution of $\bar{x}$

is a normal distribution with

mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

The Central Limit Thereom

If data is collected from a distribution (possibly non **Normal**) with mean μ and standard deviation σ then:

the sampling distribution of $\bar{x}$

is a **approximately** normal (for $n > 30$) with

mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

Distribution for Sample Proportions

Let p = population proportion of interest
or binomial probability of success.

Let

$$\hat{p} = \frac{X}{n} = \frac{\text{no. of succeses}}{\text{no. of bimomial trials}}$$

= sample proportion or proportion of successes.

Then the sampling distribution of $\hat{p}$ is approximately a normal distribution with

$$\text{mean } \mu_{\hat{p}} = p \qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Sampling distribution of a **differences**

Sampling distribution of a **difference** in two Sample means

Recall

If X, Y are ***independent*** normal random variables, then :

$X - Y$ is normal with

$$\text{mean } \mu_X - \mu_Y$$

$$\text{standard deviation } \sqrt{\sigma_X^2 + \sigma_Y^2}$$

Comparing Means

Situation

- We have two normal populations (1 and 2)
- Let μ_1 and σ_1 denote the mean and standard deviation of population 1.
- Let μ_2 and σ_2 denote the mean and standard deviation of population 2.
- Let $x_1, x_2, x_3, \ldots, x_n$ denote a sample from a normal population 1.
- Let $y_1, y_2, y_3, \ldots, y_m$ denote a sample from a normal population 2.
- Objective is to compare the two population means

We know that:

$\bar{x}$ is Normal with mean $\mu_{\bar{x}} = \mu_1$ and $\sigma_{\bar{x}} = \dfrac{\sigma_1}{\sqrt{n}}$

and

$\bar{y}$ is Normal with mean $\mu_{\bar{y}} = \mu_2$ and $\sigma_{\bar{y}} = \dfrac{\sigma_2}{\sqrt{m}}$

Thus $D = \bar{x} - \bar{y}$ is Normal with mean

$$\mu_{\bar{x}-\bar{y}} = \mu_{\bar{x}} - \mu_{\bar{y}} = \mu_1 - \mu_2$$

$$\sigma_{\bar{x}-\bar{y}} = \sqrt{\sigma_{\bar{x}}^2 + \sigma_{\bar{y}}^2} = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$$

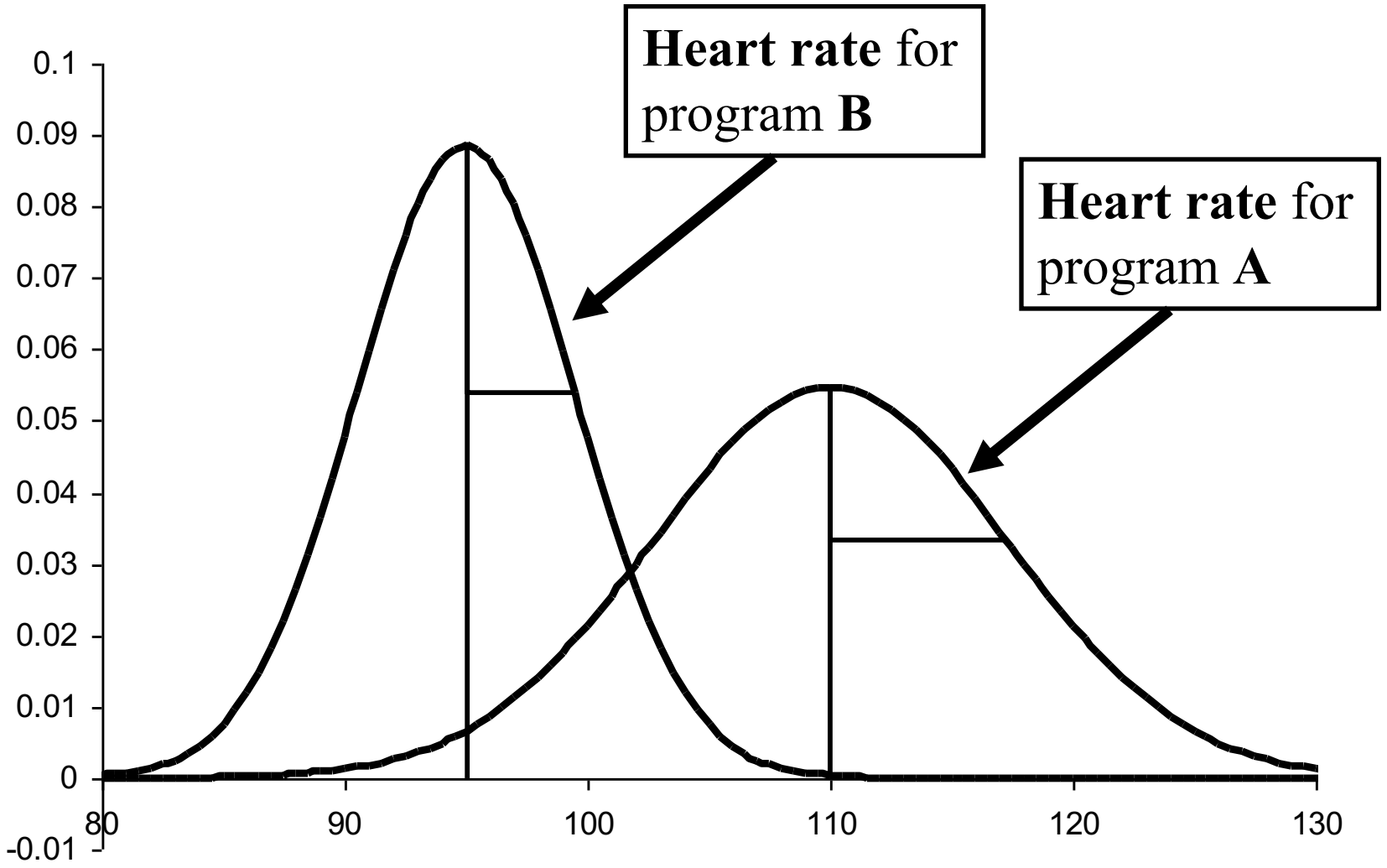
Example

Consider measuring **Heart rate** two minutes after a twenty minute exercise program.

There are two groups of individuals

1. Those who performed exercise program A (considered to be **heavy**).
2. Those who performed exercise program B (considered to be **light**).

The average **Heart rate** for those who performed exercise **program A** was $\mu_1 = 110$ with standard deviation, $\sigma_1 = 7.3$, while the average **Heart rate** for those who performed exercise **program B** was $\mu_2 = 95$ with standard deviation, $\sigma_2 = 4.5$.


Heart rate for program **B**
Heart rate for program **A**
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
-0.01
80
90
100
110
120
130

Situation

- Suppose we observe the **heart rate** of $n = 15$ subjects on **program A**.
- Let $x_1, x_2, x_3, \ldots, x_{15}$ denote these observations.
- We also observe the **heart rate** of $m = 20$ subjects on **program B**.
- Let $y_1, y_2, y_3, \ldots, y_{20}$ denote these observations.
- What is the probability that the sample mean heart rate for Program A is at least 8 units higher than the sample mean heart rate for Program B?

We know that:

$\bar{x}$ is Normal with mean $\mu_{\bar{x}} = 110$ and $\sigma_{\bar{x}} = \frac{7.3}{\sqrt{15}}$

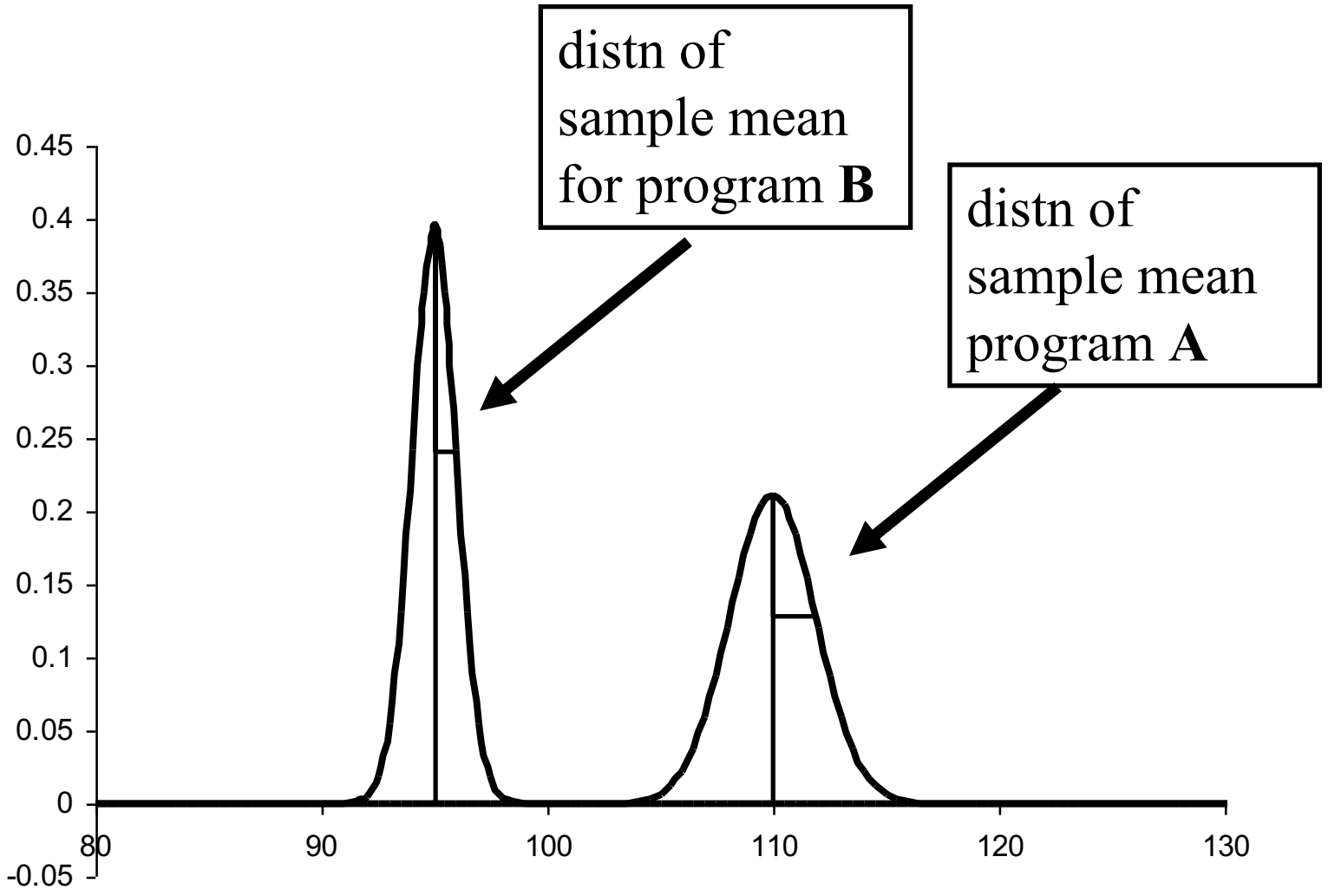
and

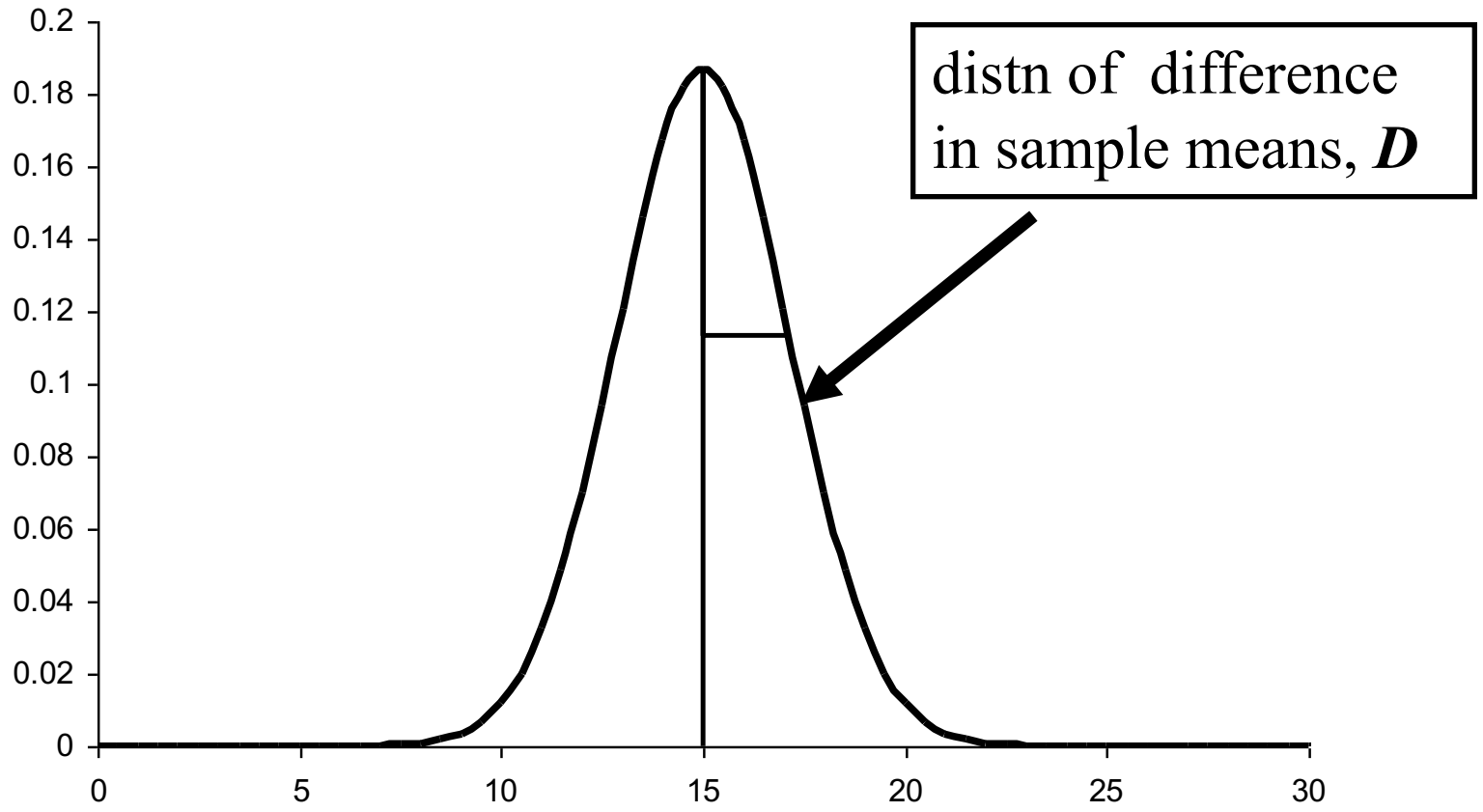
$\bar{y}$ is Normal with mean $\mu_{\bar{y}} = 95$ and $\sigma_{\bar{y}} = \frac{4.5}{\sqrt{20}}$

and $D = \bar{x} - \bar{y}$ is Normal with mean

$$\mu_{\bar{x}-\bar{y}} = \mu_{\bar{x}} - \mu_{\bar{y}} = 110 - 95 = 15$$

$$\sigma_{\bar{x}-\bar{y}} = \sqrt{\sigma_{\bar{x}}^2 + \sigma_{\bar{y}}^2} = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} == \sqrt{\frac{7.3^2}{15} + \frac{4.5^2}{20}} = 2.1366$$


distn of
sample mean
for program **B**
distn of
sample mean
program **A**
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
80
90
100
110
120
130


0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
5
10
15
20
25
30
distn of difference
in sample means, D

- What is the probability that the sample mean heart rate for Program A is at least 8 units higher than the sample mean heart rate for Program B?

Solution

$$\text{want } P\left[\bar{x} \geq \bar{y} + 8\right] = P\left[\bar{x} - \bar{y} \geq 8\right] = P\left[D \geq 8\right]$$

$$= P\left[\frac{D - 15}{2.1366} \geq \frac{8 - 15}{2.1366}\right] = P\left[z \geq -3.28\right]$$

$$= 1 - 0.0005 = 0.9995$$

Sampling distribution of a **difference** in two Sample proportions

Comparing Proportions

Situation

- Suppose we have **two *Success-Failure*** experiments
- Let p_1 = the **probability of success** for experiment 1.
- Let p_2 = the **probability of success** for experiment 2.
- Suppose that experiment 1 is repeated n_1 times and experiment 2 is repeated n_2
- Let x_1 = the no. of **successes** in the n_1 repititions of experiment 1, x_2 = the no. of **successes** in the n_2 repititions of experiment 2.

$$\hat{p}_1 = \frac{x_1}{n_1} \quad \text{and} \quad \hat{p}_2 = \frac{x_2}{n_2}$$

What is the distribution of $D = \hat{p}_1 - \hat{p}_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2}$?

We know that:

$$\hat{p}_1 = \frac{x_1}{n_1} \quad \text{is Normal with mean} \quad \mu_{\hat{p}_1} = p_1$$

$$\text{and} \quad \sigma_{\hat{p}_1} = \sqrt{\frac{p_1(1-p_1)}{n_1}}$$

$$\text{Also} \quad \hat{p}_2 = \frac{x_2}{n_2} \quad \text{is Normal with mean} \quad \mu_{\hat{p}_2} = p_2$$

$$\text{and} \quad \sigma_{\hat{p}_2} = \sqrt{\frac{p_2(1-p_2)}{n_2}}$$

Thus $D = \hat{p}_1 - \hat{p}_2$ is Normal with mean

$$\mu_{\hat{p}_1 - \hat{p}_2} = \mu_{\hat{p}_1} - \mu_{\hat{p}_2} = p_1 - p_2$$

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\sigma^2_{\hat{p}_1} + \sigma^2_{\hat{p}_2}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Example

The **Globe and Mail** carried out a survey to investigate the ***"State of the Baby Boomers".*** **(June 2006)**

Two populations in the study

1. Baby Boomers (age 40 – 59) ($n_1 = 664$)
2. Generation *X* (age 30 – 39) ($n_2 = 342$)

One of questions

"Are you close to your parents? – Yes or No"

Suppose that the proportions in the two populations were:

- Baby Boomers – 40% **yes** ($p_1 = 0.40$)
- Generation *X* – 20% **yes** ($p_2 = 0.20$)

What is the probability that this would be observed in the samples to a certain degree?

What is $P[\hat{p}_1 - \hat{p}_2 \geq 0.15]$?

Solution:

$$\hat{p}_1 = \frac{x_1}{n_1} \text{ is Normal with mean } \mu_{\hat{p}_1} = p_1 = 0.40$$

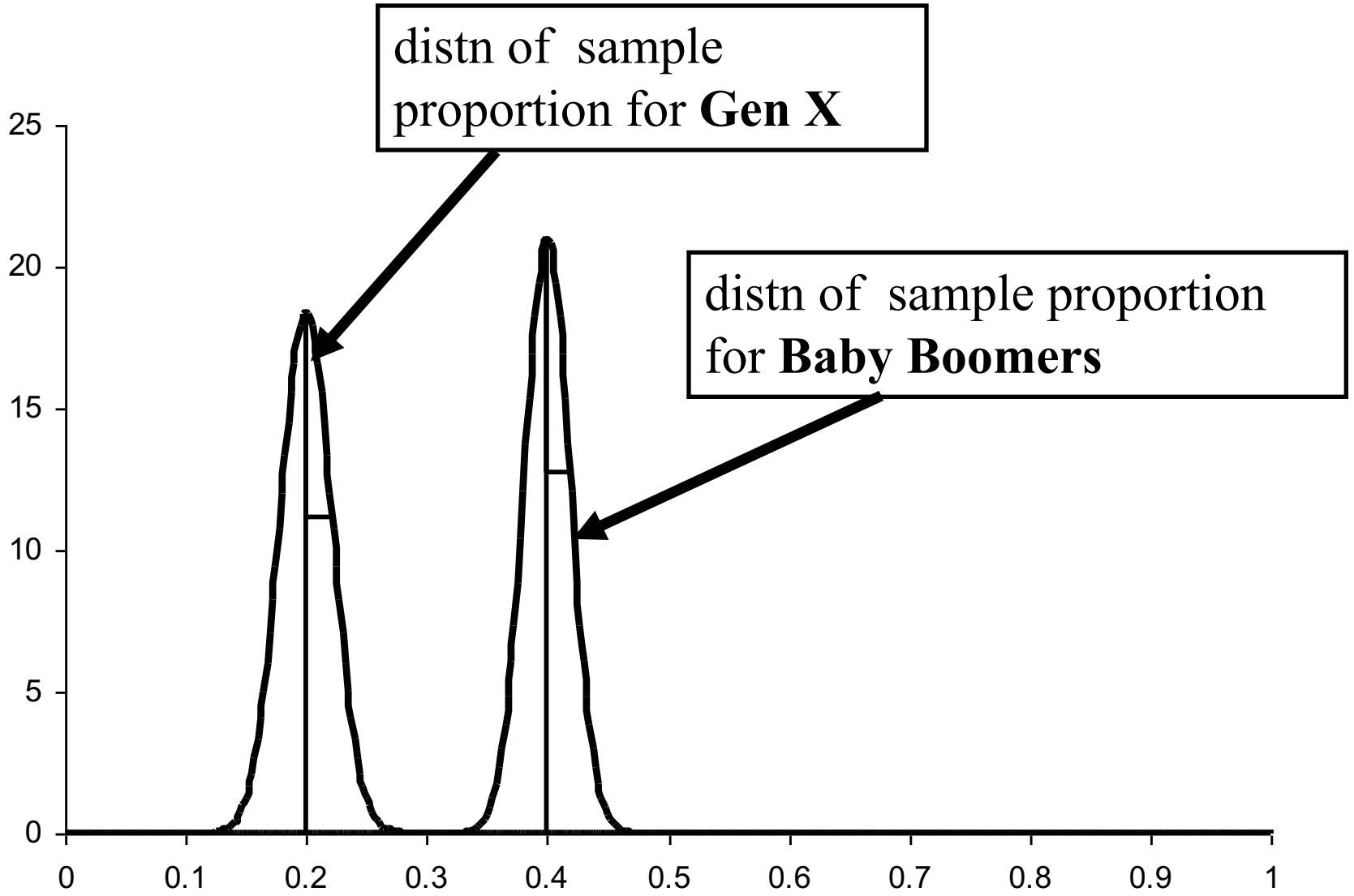
$$\text{and } \sigma_{\hat{p}_1} = \sqrt{\frac{p_1(1-p_1)}{n_1}}$$

$$= \sqrt{\frac{0.40(1-0.40)}{664}} = 0.019012$$

$$\text{Also } \hat{p}_2 = \frac{x_2}{n_2} \text{ is Normal with mean } \mu_{\hat{p}_2} = p_2 = 0.20$$

$$\text{and } \sigma_{\hat{p}_2} = \sqrt{\frac{p_2(1-p_2)}{n_2}}$$

$$= \sqrt{\frac{0.20(1-0.20)}{342}} = 0.02163$$


distn of sample proportion for **Gen X**
distn of sample proportion for **Baby Boomers**
25
20
15
10
5
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Now $D = \hat{p}_1 - \hat{p}_2$ is Normal with mean

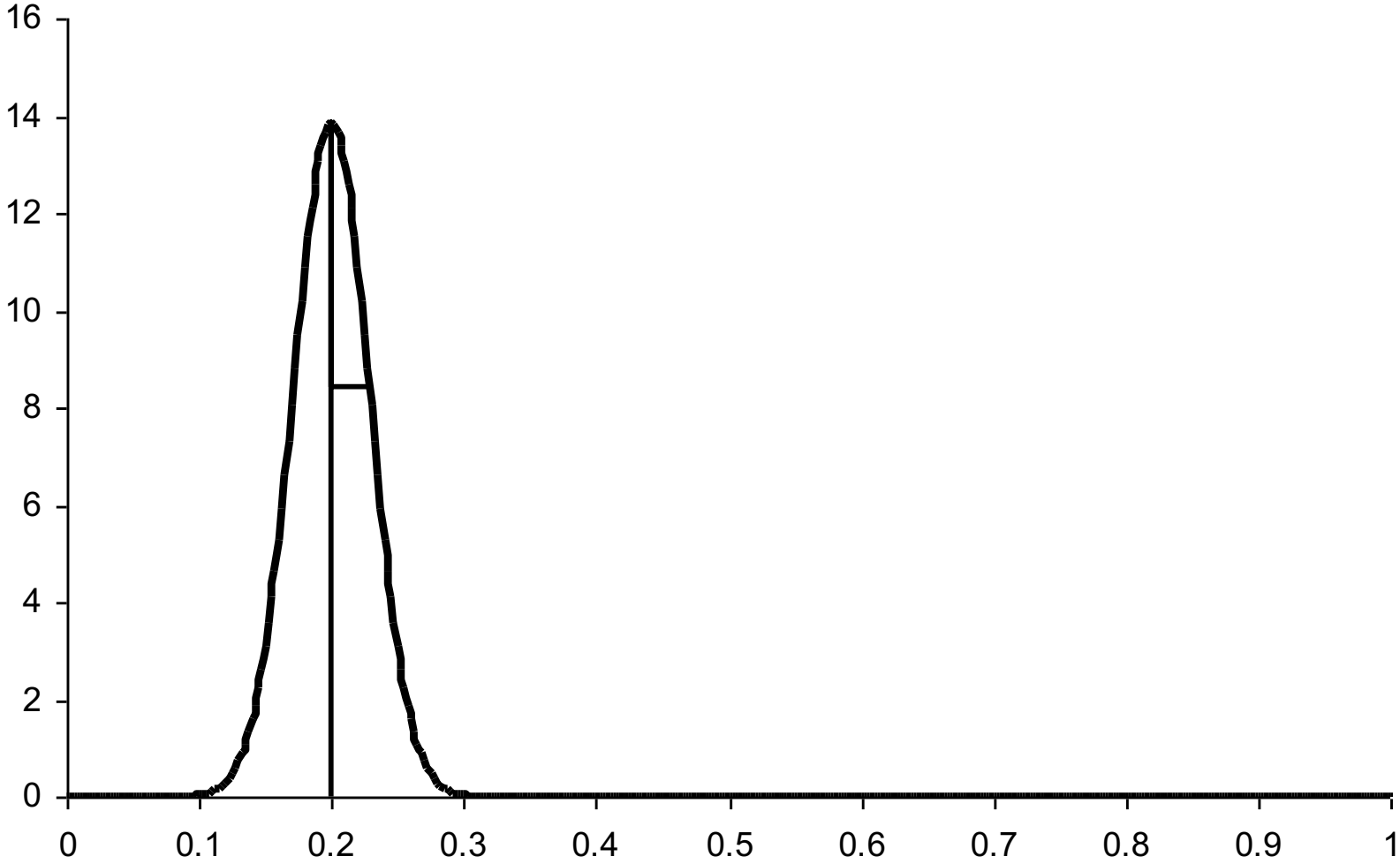
$$\mu_D = \mu_{\hat{p}_1 - \hat{p}_2} = \mu_{\hat{p}_1} - \mu_{\hat{p}_2} = p_1 - p_2 = 0.4 - 0.2 = 0.2$$

$$\sigma_D = \sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\sigma^2_{\hat{p}_1} + \sigma^2_{\hat{p}_2}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$= \sqrt{\frac{0.4(1-0.4)}{664} + \frac{0.2(1-0.2)}{342}}$$

$$= 0.028797$$

Distribution of $D = \hat{p}_1 - \hat{p}_2$



Now

$$P\left[\hat{p}_1 - \hat{p}_2 \geq 0.15\right] = P\left[D \geq 0.15\right]$$

$$= P\left[\frac{D - \mu_D}{\sigma_D} \geq \frac{0.15 - 0.2}{0.028797}\right]$$

$$= P\left[z \geq -1.74\right] = 1 - 0.0409 = .9591$$

Sampling distributions

Summary

Distribution for Sample Mean

If data is collected from a **Normal** distribution with mean μ and standard deviation σ then:

the sampling distribution of $\bar{x}$

is a normal distribution with

mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

The Central Limit Thereom

If data is collected from a distribution (possibly non **Normal**) with mean μ and standard deviation σ then:

the sampling distribution of $\bar{x}$

is a **approximately** normal (for $n > 30$) with

mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

Distribution for Sample Proportions

Let p = population proportion of interest
or binomial probability of success.

Let

$$\hat{p} = \frac{X}{n} = \frac{\text{no. of succeses}}{\text{no. of bimomial trials}}$$

= sample proportion or proportion of successes.

Then the sampling distribution of $\hat{p}$ is approximately a normal distribution with

$$\text{mean } \mu_{\hat{p}} = p \qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Distribution of a difference in two sample Means

$$D = \bar{x} - \bar{y} \quad \text{is Normal with mean}$$

$$\mu_{\bar{x}-\bar{y}} = \mu_{\bar{x}} - \mu_{\bar{y}} = \mu_1 - \mu_2$$

$$\sigma_{\bar{x}-\bar{y}} = \sqrt{\sigma_{\bar{x}}^2 + \sigma_{\bar{y}}^2} = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$$

Distribution of a difference in two sample proportions

$$D = \hat{p}_1 - \hat{p}_2 \quad \text{is Normal with mean}$$

$$\mu_{\hat{p}_1 - \hat{p}_2} = \mu_{\hat{p}_1} - \mu_{\hat{p}_2} = p_1 - p_2$$

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\sigma^2_{\hat{p}_1} + \sigma^2_{\hat{p}_2}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

The Chi-square (χ^2) distribution

The Chi-squared distribution
with
ν degrees of freedom

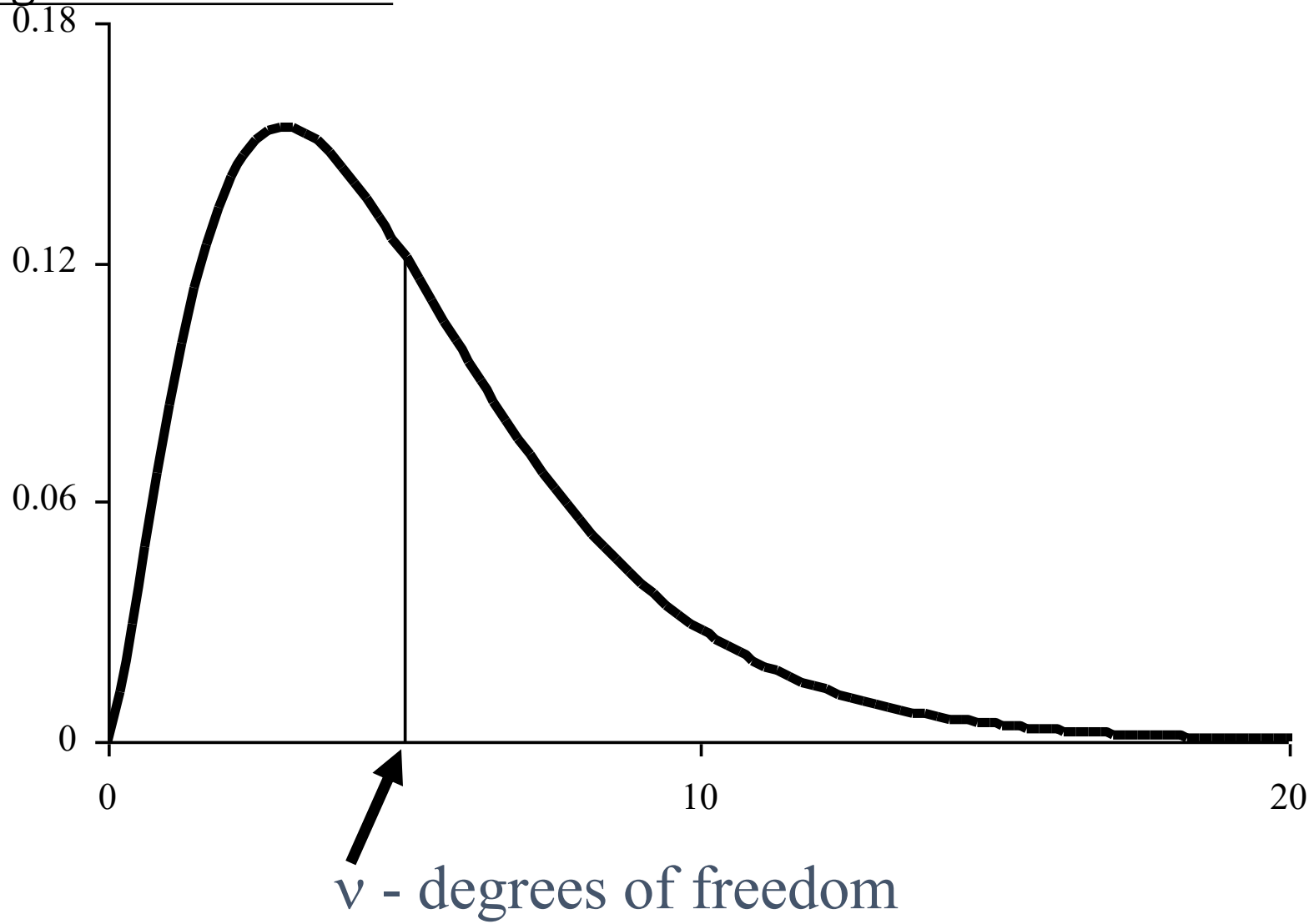
Comment: If $z_1, z_2, \ldots, z_\nu$ are independent random variables each having a standard normal distribution then

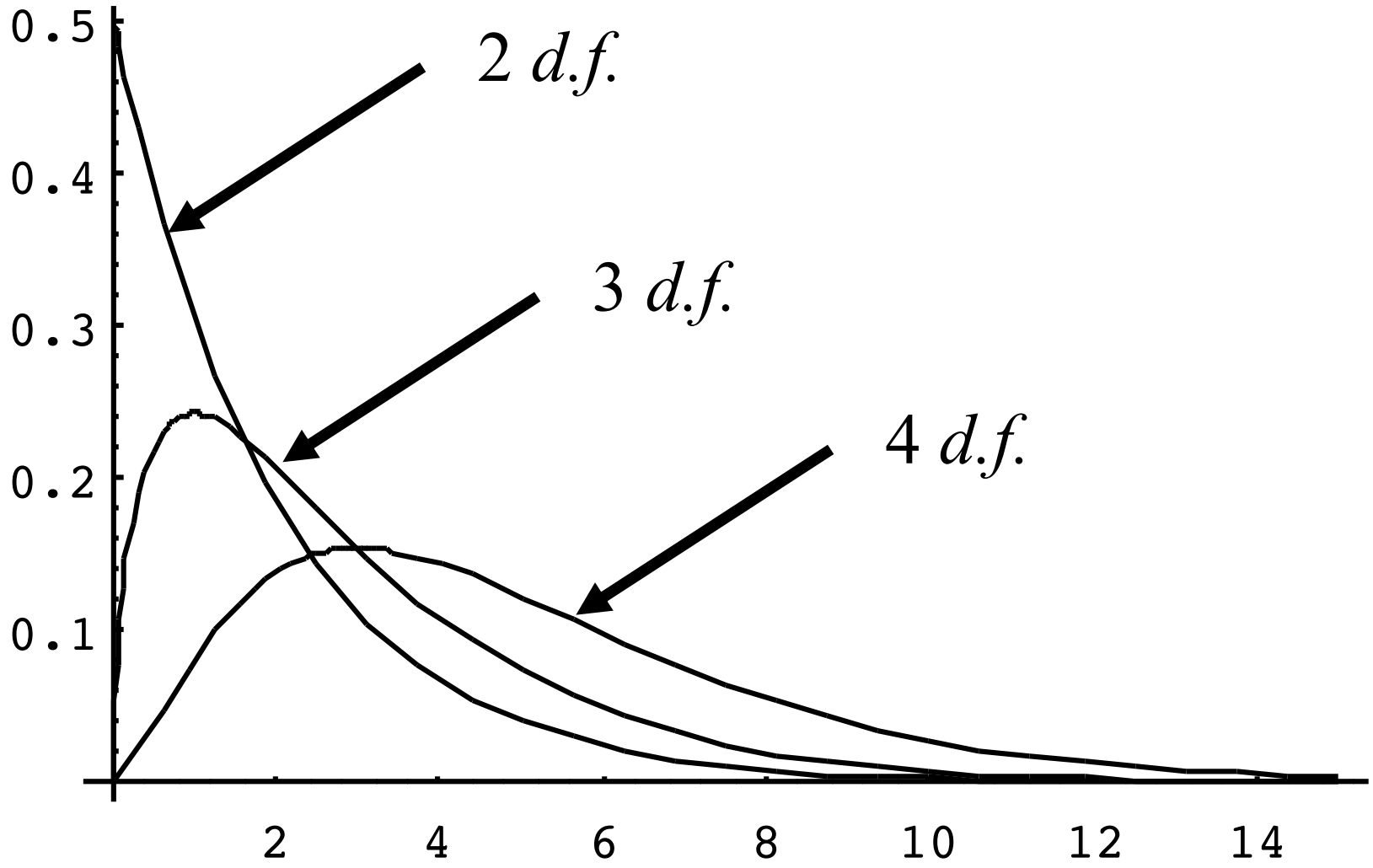
$$U = z_1^2 + z_2^2 + \ldots + z_\nu^2$$

has a chi-squared distribution with ν degrees of freedom.

The Chi-squared distribution with

ν degrees of freedom




0.5
0.4
0.3
0.2
0.1
2 d.f.
3 d.f.
4 d.f.
2
4
6
8
10
12
14

Statistics that have the Chi-squared distribution:

1. $$\chi^2 = \sum_{j=1}^{c}\sum_{i=1}^{r}\frac{\left(x_{ij} - E_{ij}\right)^2}{E_{ij}} = \sum_{j=1}^{c}\sum_{i=1}^{r} r_{ij}^2$$

The statistic used to detect independence between two categorical variables

$$d.f. = (r-1)(c-1)$$

Let $x_1, x_2, \ldots, x_n$ denote a sample from the normal distribution with mean μ and standard deviation σ, then

2. $$U = \frac{\sum_{i=1}^{r} \left(x_i - \bar{x} \right)^2}{\sigma^2}$$

$$= \frac{(n-1)s^2}{\sigma^2}$$

has a chi-square distribution with $d.f. = n - 1$.

Example

Suppose that $x_1, x_2, \ldots, x_{10}$ is a sample of size $n = 10$ from the normal distribution with mean $\mu = 100$ and standard deviation $\sigma = 15$.

Suppose that

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

is the sample standard deviation.

Find $P[10 < s < 20]$.

Note

$$U = \frac{\sum_{i=1}^{r}\left(x_i - \bar{x}\right)^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} = \frac{(9)s^2}{(15)^2}$$

has a chi-square distribution with

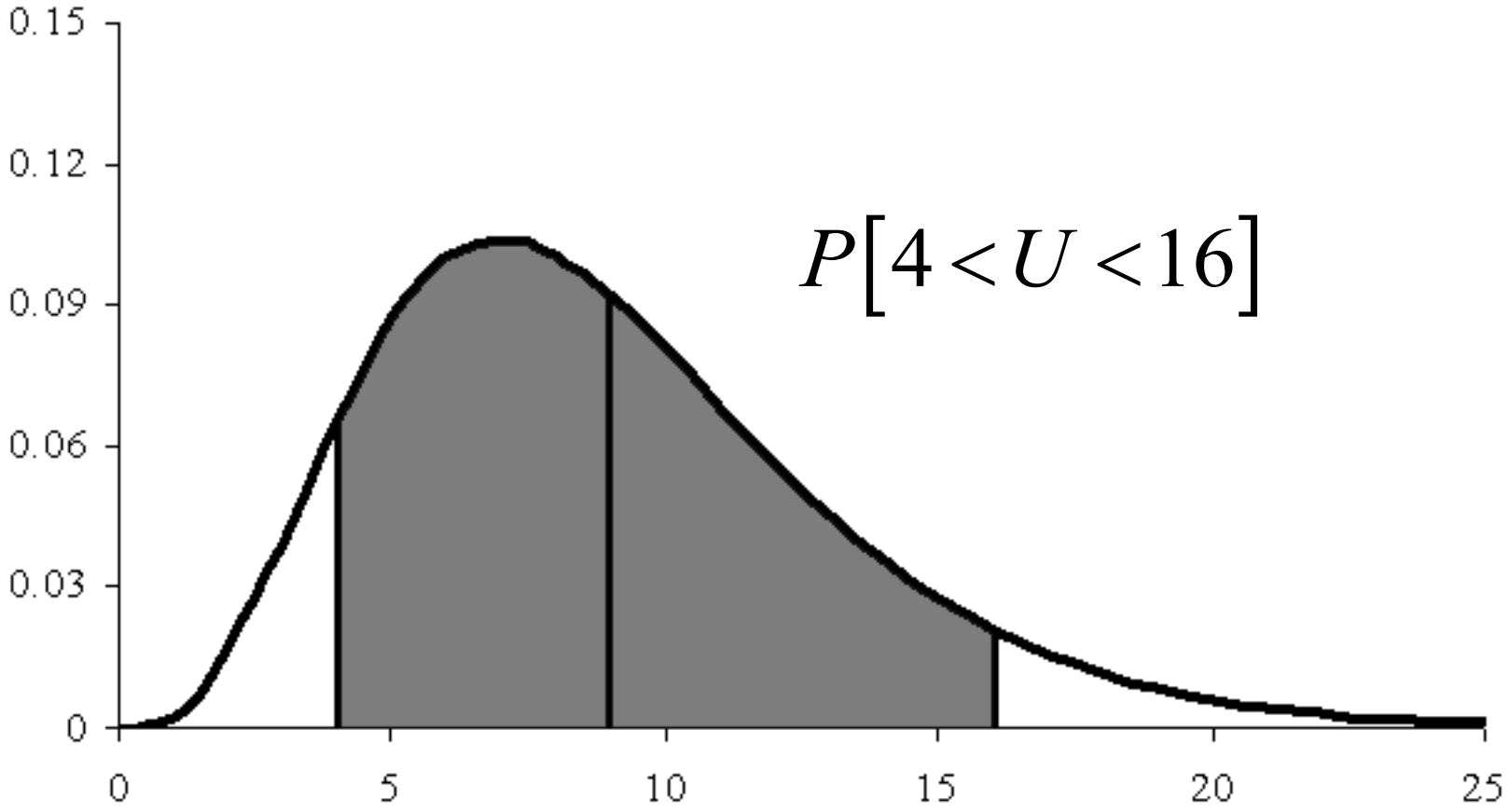
$$d.f. = n - 1 = 9$$

$$P\left[10 < s < 20\right] = P\left[100 < s^2 < 400\right]$$

$$= P\left[\frac{9(100)}{15^2} < \frac{9s^2}{15^2} < \frac{9(400)}{15^2}\right]$$

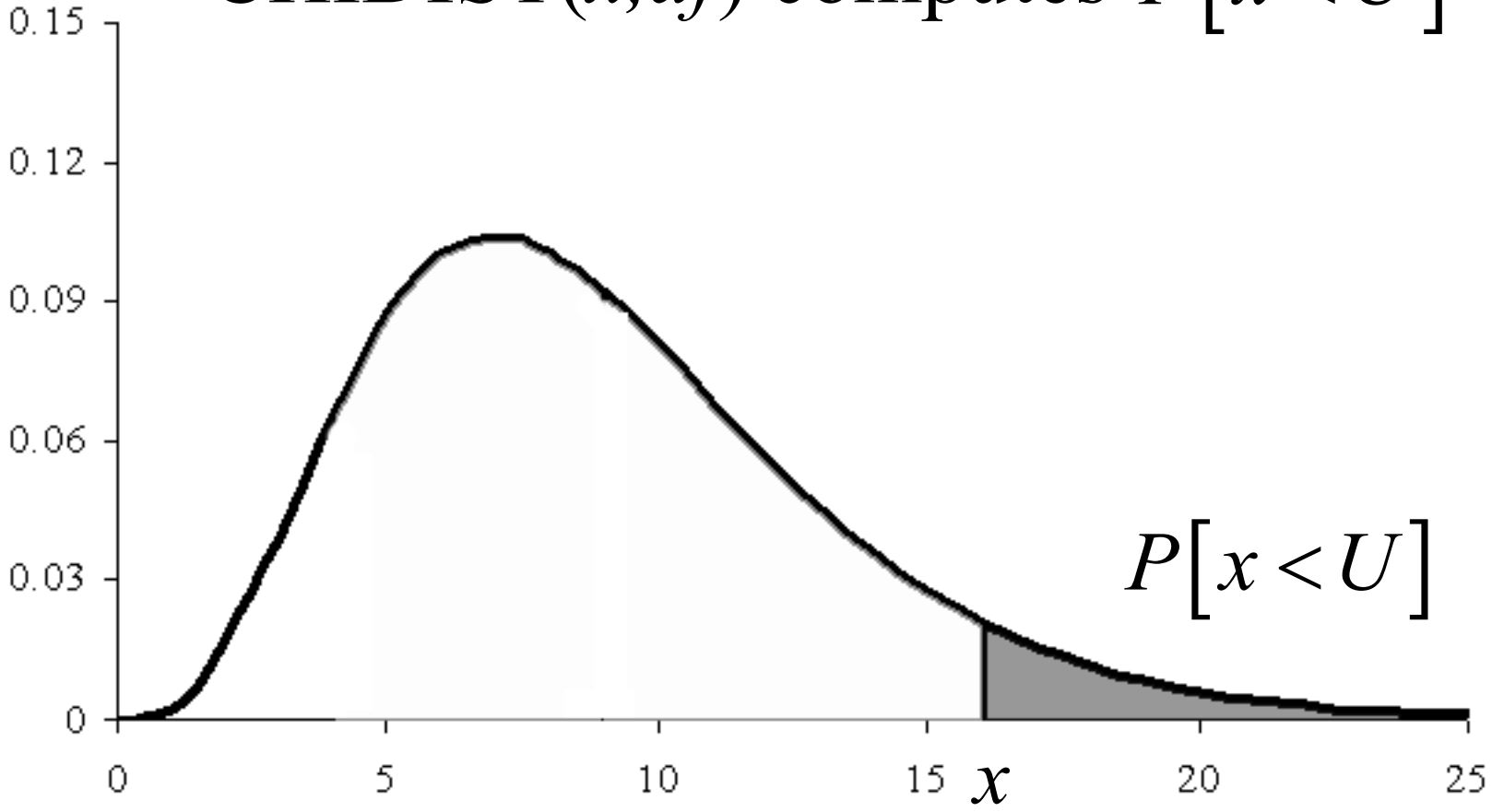
$$= P\left[4 < U < 16\right]$$

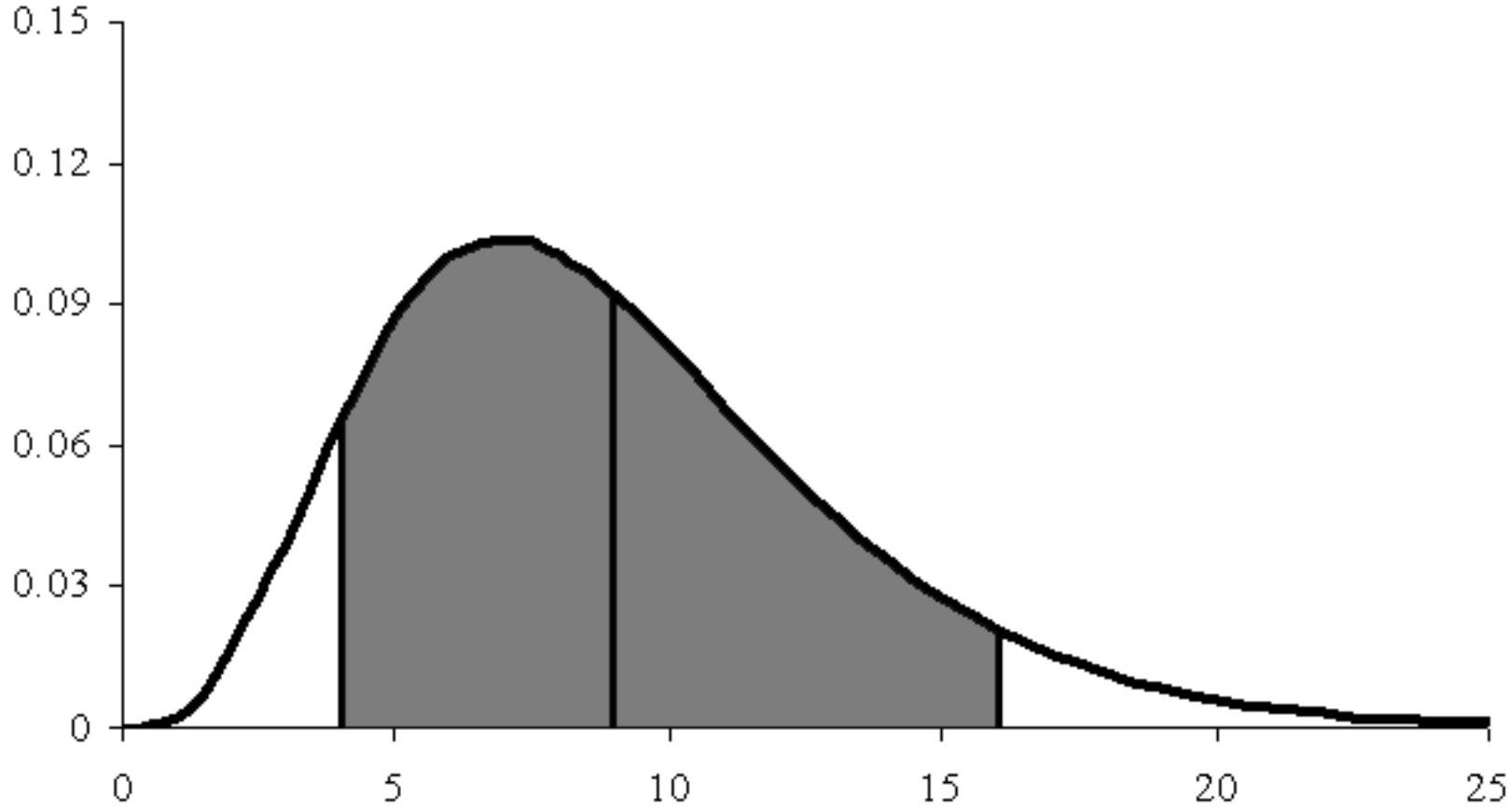
chi-square distribution with $d.f. = n - 1 = 9$



We do not have tables to compute this area

The excel function CHIDIST(x,df) computes $P[x < U]$





$$P\left[4 < U < 16\right] = \text{CHIDIST(4,9)-CHIDIST(16,9)}$$

$$= 0.91141 - 0.06688 = 0.84453$$

Statistical Inference