Probability and Statistics

Kristel Van Steen, PhD[2]

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

kristel.vansteen@ulg.ac.be

CHAPTER 5: SAMPLING TO APPROXIMATE THE TRUE WORLD

1 Introduction

2 Generating data

2.1 Design of experiments

2.2 Sampling designs and towards inference

2.3 Ethics

3 Sample versus population

3.1 Introduction

3.2 Distribution of a sample

3.3 Statistics and sample moments

4 Sample mean

4.1 Mean and variance

4.2 Law of large numbers revisited

4.3 Central-limit theory revisited

4.4 Bernoulli and Poisson distribution

4.5 Exponential distribution

4.6 Uniform distribution

5 Sampling from the normal distribution

5.1 The role of normal distributions in statistics

5.2 Sample mean

5.3 The chi-square distribution

5.4 The F distribution

5.5 The Student’s t distribution

6 Future highlights

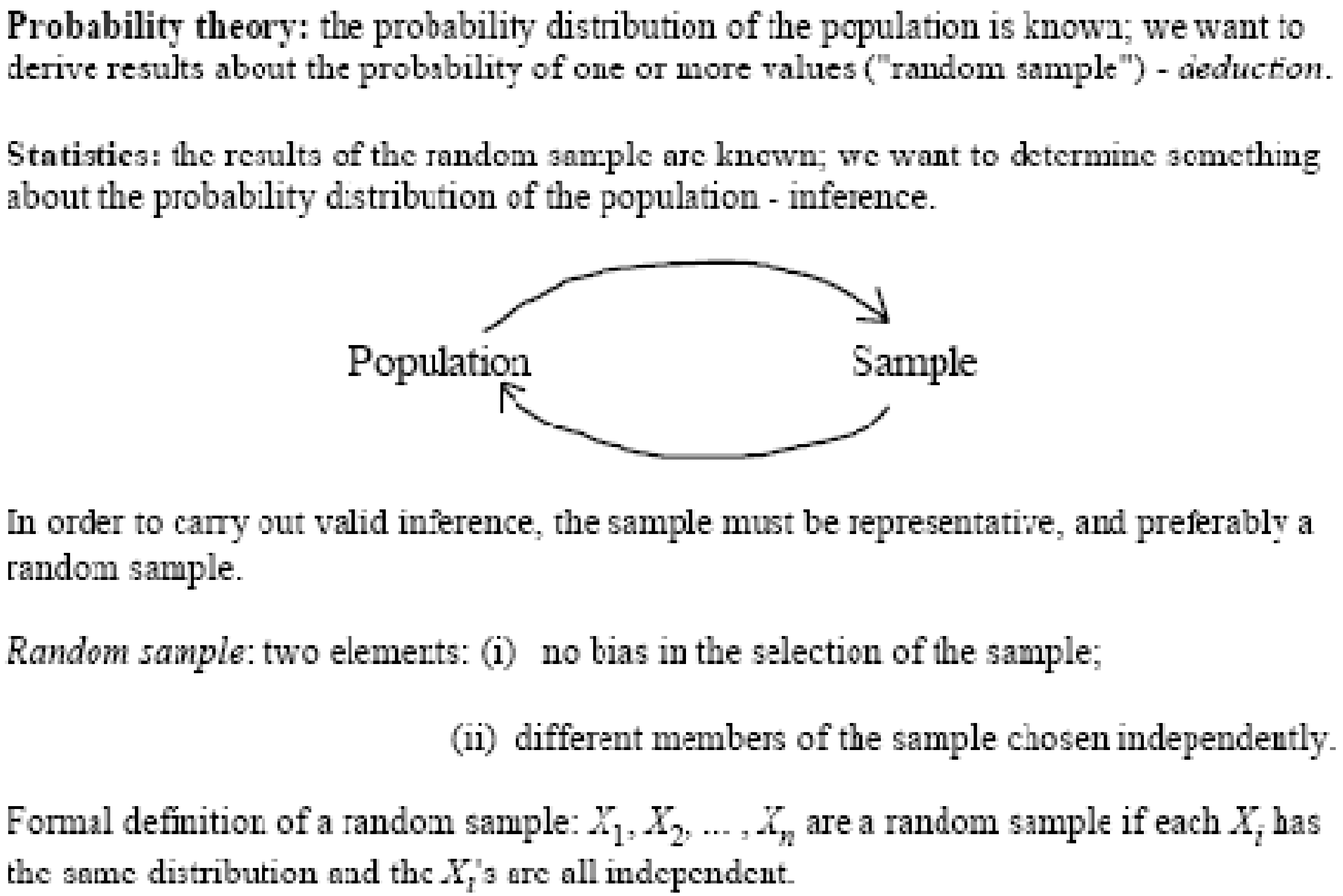
6.1 Estimating parameters

6.2 Order statistics

6.3 Sample size calculations

1 Introduction

Probability theory versus statistics



Probability theory: the probability distribution of the population is known; we want to derive results about the probability of one or more values ("random sample") - *deduction*.

Statistics: the results of the random sample are known; we want to determine something about the probability distribution of the population - inference.

Population
Sample

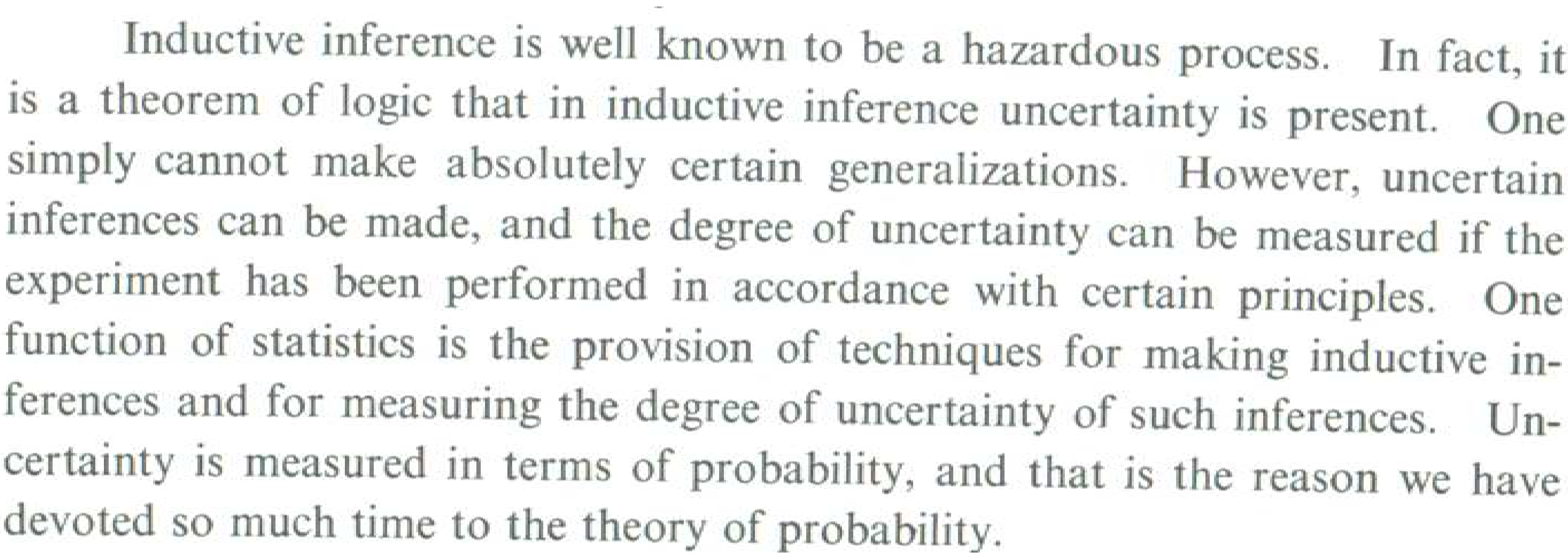
In order to carry out valid inference, the sample must be representative, and preferably a random sample.

Random sample: two elements: (i) no bias in the selection of the sample;

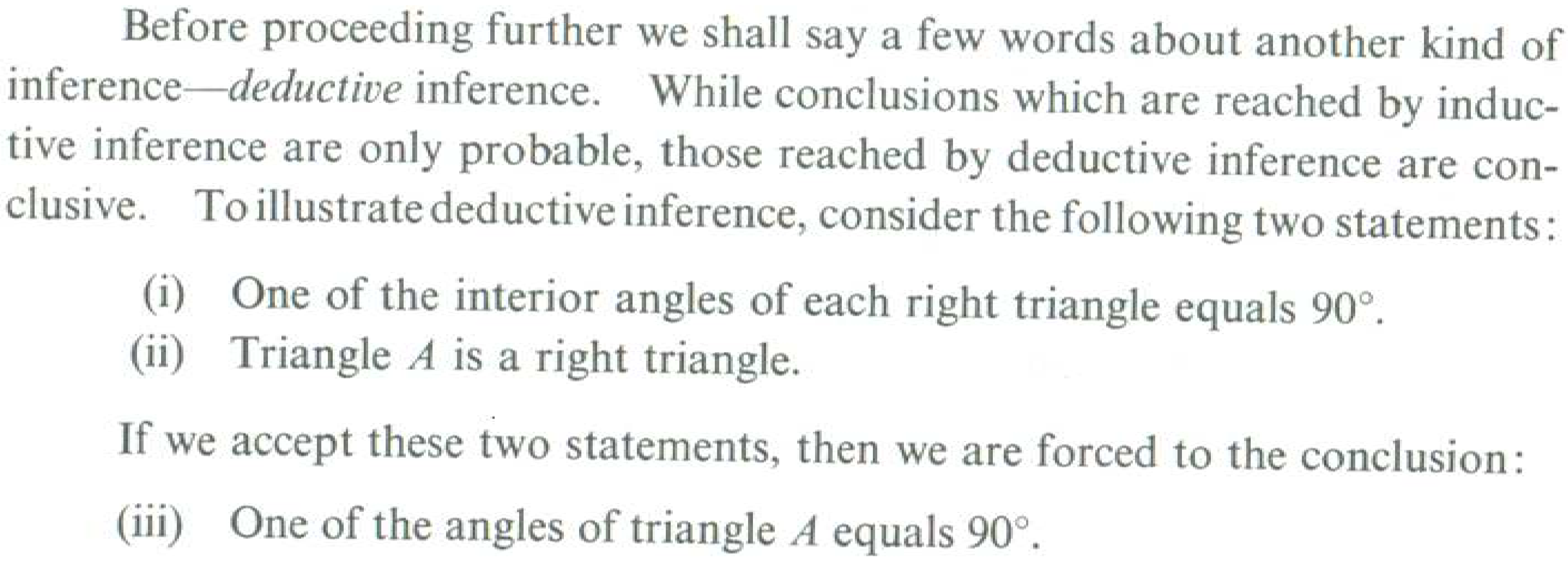
(ii) different members of the sample chosen independently.

Formal definition of a random sample: $X_1, X_2, \ldots, X_n$ are a random sample if each X_i has the same distribution and the X_i's are all independent.

Inductive versus deductive inference



Inductive inference is well known to be a hazardous process. In fact, it is a theorem of logic that in inductive inference uncertainty is present. One simply cannot make absolutely certain generalizations. However, uncertain inferences can be made, and the degree of uncertainty can be measured if the experiment has been performed in accordance with certain principles. One function of statistics is the provision of techniques for making inductive inferences and for measuring the degree of uncertainty of such inferences. Uncertainty is measured in terms of probability, and that is the reason we have devoted so much time to the theory of probability.



Before proceeding further we shall say a few words about another kind of inference—*deductive* inference. While conclusions which are reached by inductive inference are only probable, those reached by deductive inference are conclusive. To illustrate deductive inference, consider the following two statements:

(i) One of the interior angles of each right triangle equals 90°.
(ii) Triangle A is a right triangle.

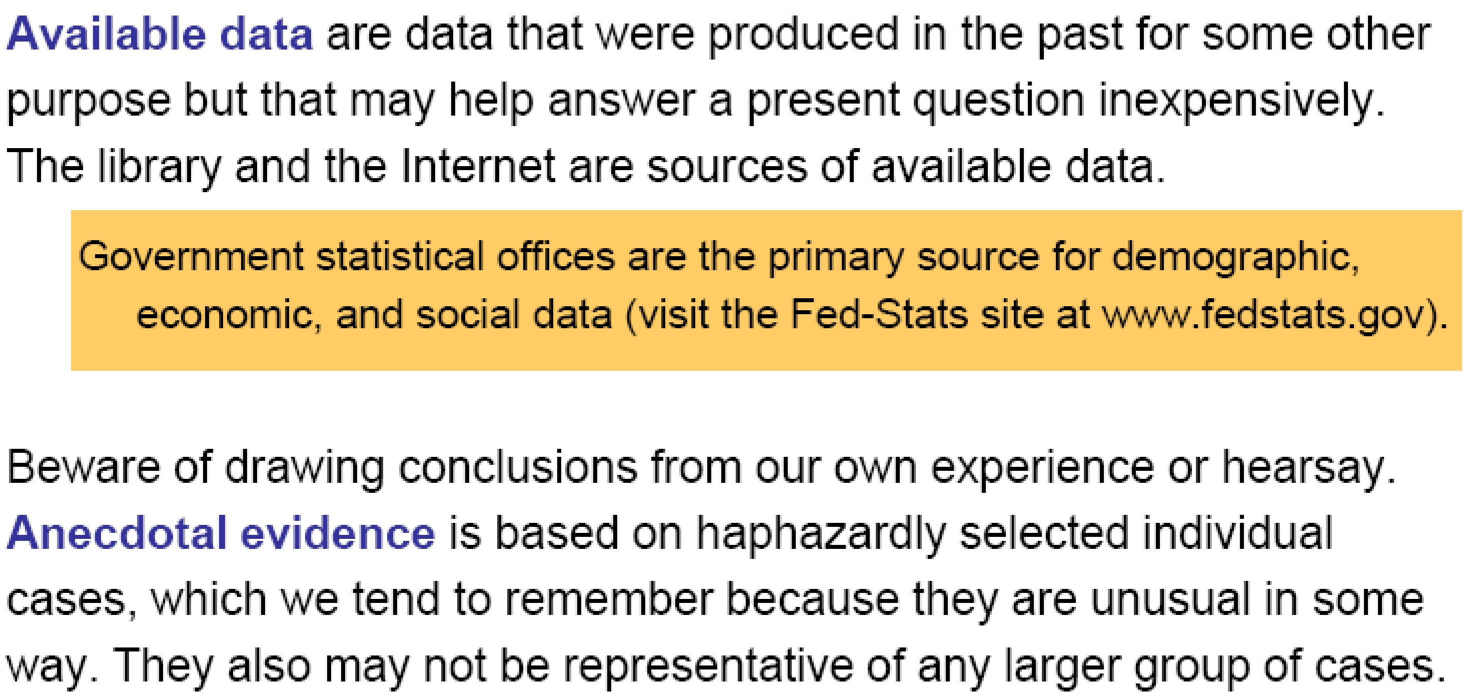
If we accept these two statements, then we are forced to the conclusion:

(iii) One of the angles of triangle A equals 90°.

2 Generating data

2.1 Design of experiments

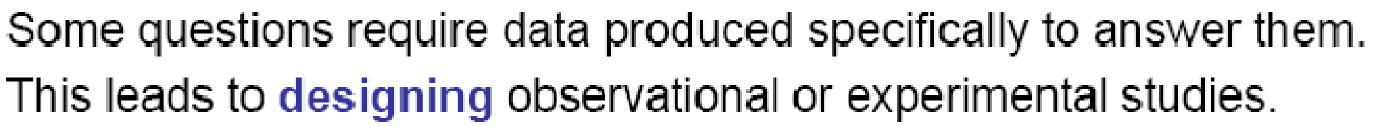
Obtaining data



Available data are data that were produced in the past for some other purpose but that may help answer a present question inexpensively. The library and the Internet are sources of available data.

> Government statistical offices are the primary source for demographic, economic, and social data (visit the Fed-Stats site at www.fedstats.gov).

Beware of drawing conclusions from our own experience or hearsay. **Anecdotal evidence** is based on haphazardly selected individual cases, which we tend to remember because they are unusual in some way. They also may not be representative of any larger group of cases.



Some questions require data produced specifically to answer them.
This leads to designing observational or experimental studies.

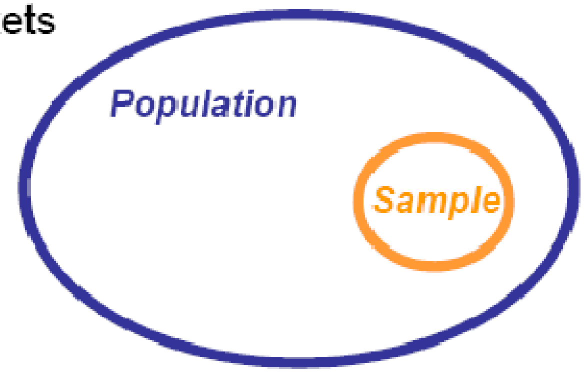
Population versus sample

- **Population**: The entire group of individuals in which we are interested but can't usually assess directly.

  Example: All humans, all working-age people in California, all crickets

- **Sample**: The part of the population we actually examine and for which we do have data.

  How well the sample represents the population depends on the sample design.



- A parameter is a number describing a characteristic of the population.

- A statistic is a number describing a characteristic of a sample.

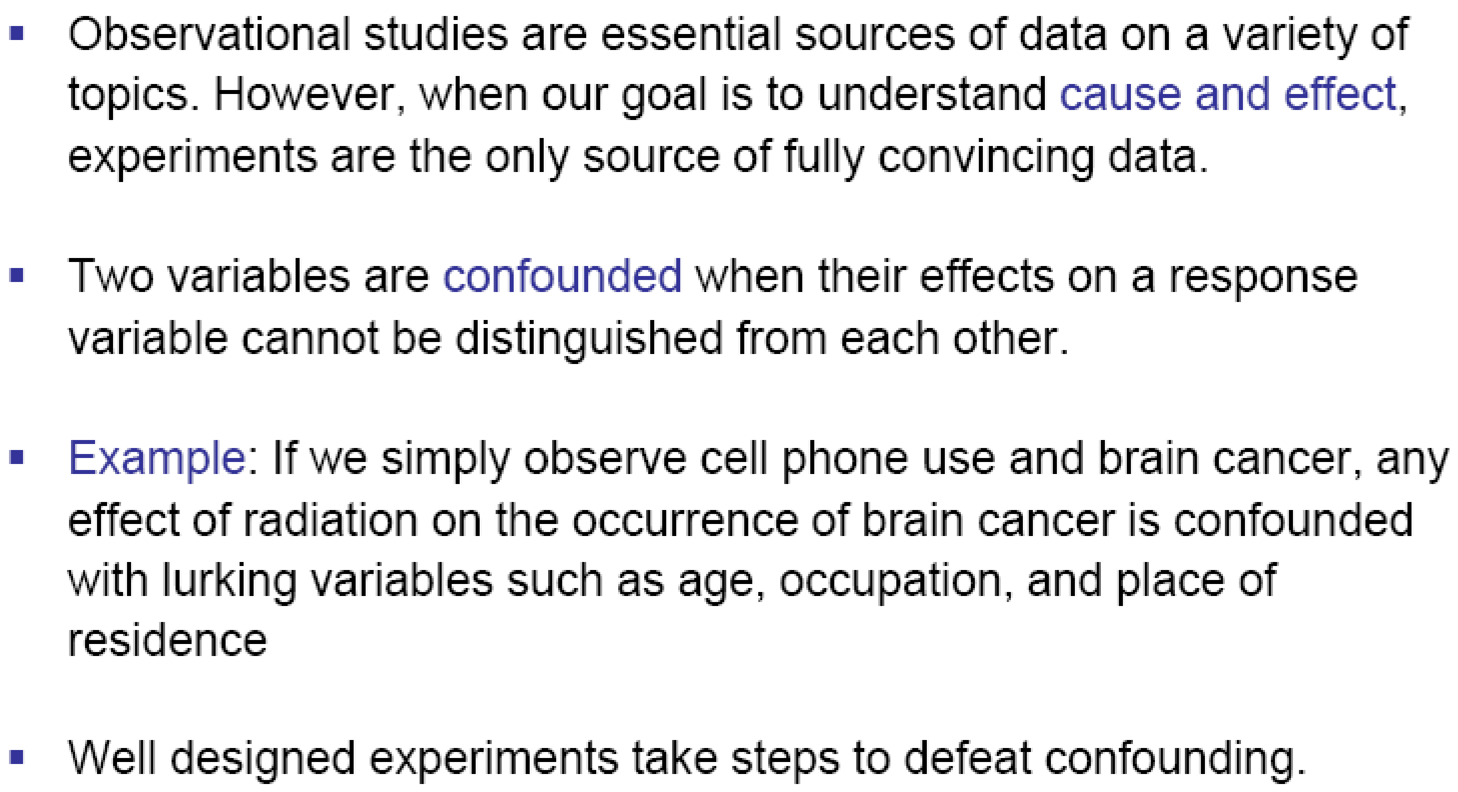
Observational studies versus experiments

Observational study: Record data on individuals without attempting to influence the responses.

Example: Based on observations you make in nature, you suspect that female crickets choose their mates on the basis of their health. → Observe health of male crickets that mated.

Experimental study: Deliberately impose a treatment on individuals and record their responses. Influential factors can be controlled.

Example: Deliberately infect some males with intestinal parasites and see whether females tend to choose healthy rather than ill males.



- Observational studies are essential sources of data on a variety of topics. However, when our goal is to understand cause and effect, experiments are the only source of fully convincing data.

- Two variables are confounded when their effects on a response variable cannot be distinguished from each other.

- Example: If we simply observe cell phone use and brain cancer, any effect of radiation on the occurrence of brain cancer is confounded with lurking variables such as age, occupation, and place of residence

- Well designed experiments take steps to defeat confounding.

Some terminology

- The individuals in an experiment are the **experimental units.** If they are human, we call them **subjects.**
- In an experiment, we do something to the subject and measure the response. The "something" we do is a called a **treatment**, or **factor.**

- The factor may be the administration of a drug.
- One group of people may be placed on a diet/exercise program for six months (treatment), and their blood pressure (response variable) would be compared with that of people who did not diet or exercise.

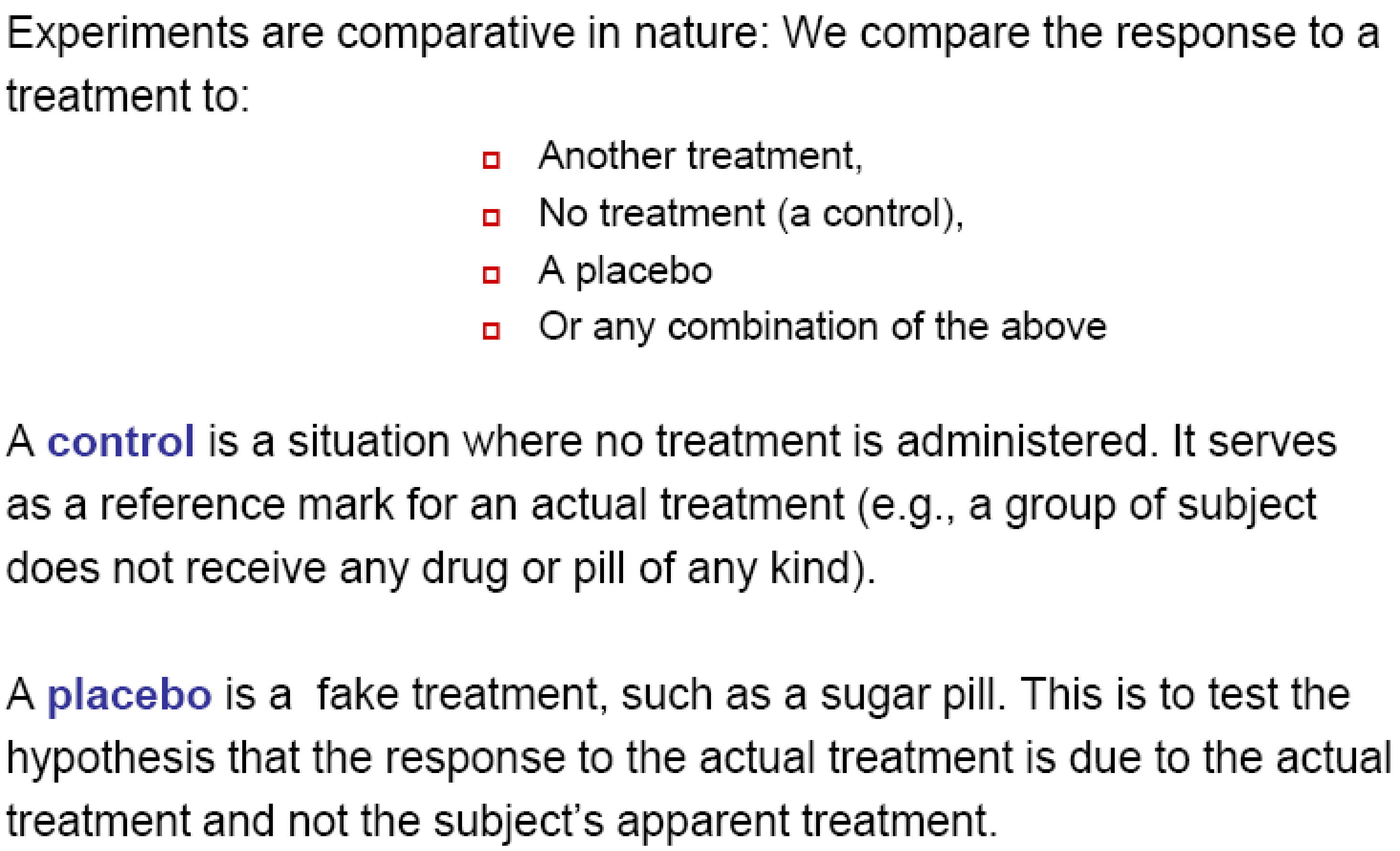
- If the experiment involves giving two different doses of a drug, we say that we are testing two **levels** of the factor.

- A response to a treatment is **statistically significant** if it is larger than you would expect by chance (due to random variation among the subjects). We will learn how to determine this later.

In a study of sickle cell anemia, 150 patients were given the drug hydroxyurea, and 150 were given a placebo (dummy pill). The researchers counted the episodes of pain in each subject. Identify:

• The subjects	• (patients, all 300)
• The factors / treatments	• (hydroxyurea and placebo)
• And the response variable	• (episodes of pain)

Comparative experiments



Experiments are comparative in nature: We compare the response to a treatment to:

- Another treatment,
- No treatment (a control),
- A placebo
- Or any combination of the above

A **control** is a situation where no treatment is administered. It serves as a reference mark for an actual treatment (e.g., a group of subject does not receive any drug or pill of any kind).

A **placebo** is a fake treatment, such as a sugar pill. This is to test the hypothesis that the response to the actual treatment is due to the actual treatment and not the subject's apparent treatment.

Placebo effects

The "placebo effect" is an improvement in health not due to any treatment, but only to the patient's belief that he or she will improve.

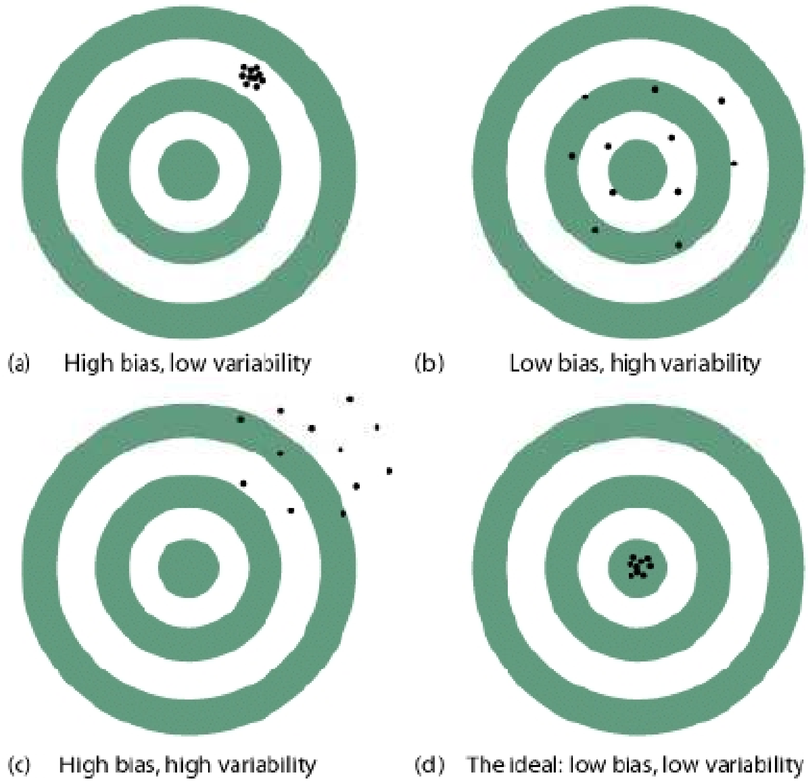
- The "placebo effect" is not understood, but it is believed to have therapeutic results on up to a whopping 35% of patients.
- It can sometimes ease the symptoms of a variety of ills, from asthma to pain to high blood pressure, and even to heart attacks.
- An opposite, or "negative placebo effect," has been observed when patients believe their health will get worse.

The most famous, and maybe most powerful, placebo is the "kiss," blow, or hug—whatever your technique.

Unfortunately, the effect gradually disappears once children figure out that they sometimes get better without help and vice versa.

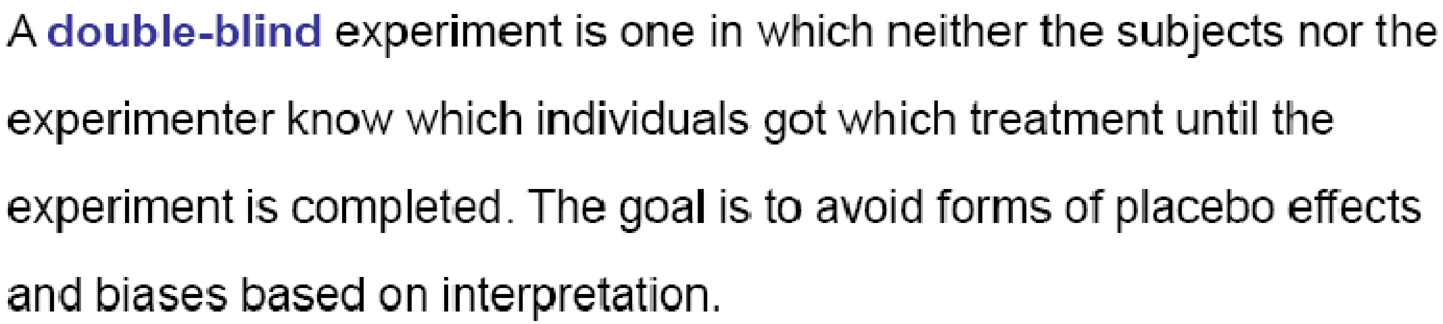
Caution about experimentation

The design of a study is **biased** if it systematically favors certain outcomes.

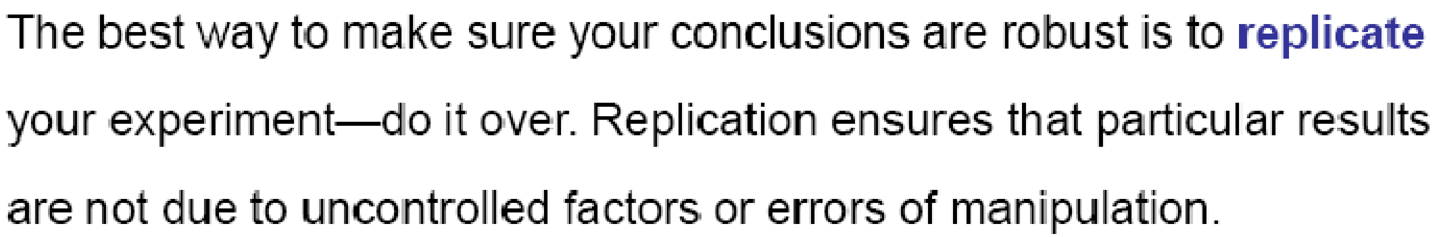

(a) High bias, low variability
(b) Low bias, high variability
(c) High bias, high variability
(d) The ideal: low bias, low variability

The best way to exclude biases from an experiment is to **randomize** the design. Both the individuals and treatments are assigned randomly.

Other ways to remove bias

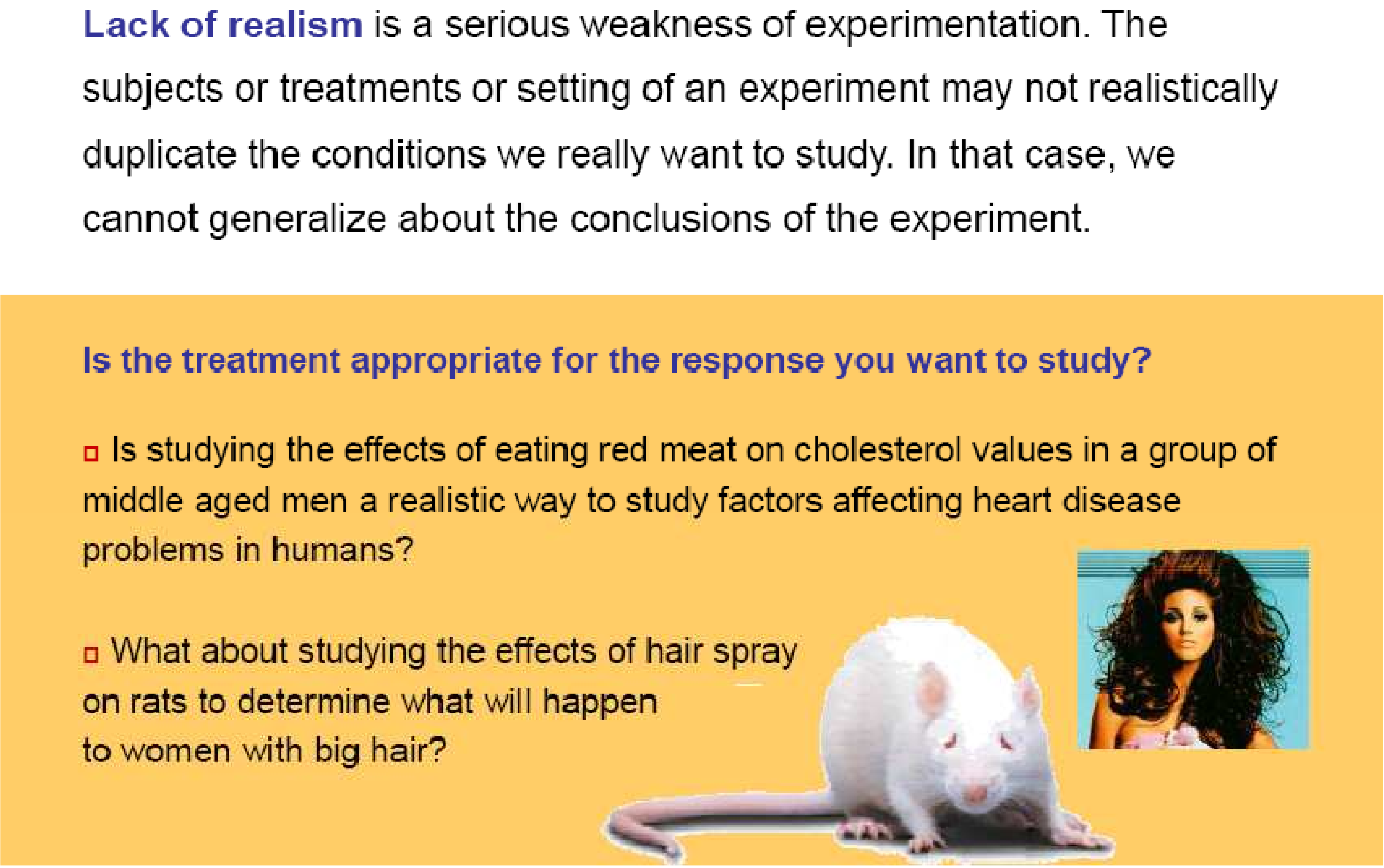


A **double-blind** experiment is one in which neither the subjects nor the experimenter know which individuals got which treatment until the experiment is completed. The goal is to avoid forms of placebo effects and biases based on interpretation.



The best way to make sure your conclusions are robust is to **replicate** your experiment—do it over. Replication ensures that particular results are not due to uncontrolled factors or errors of manipulation.

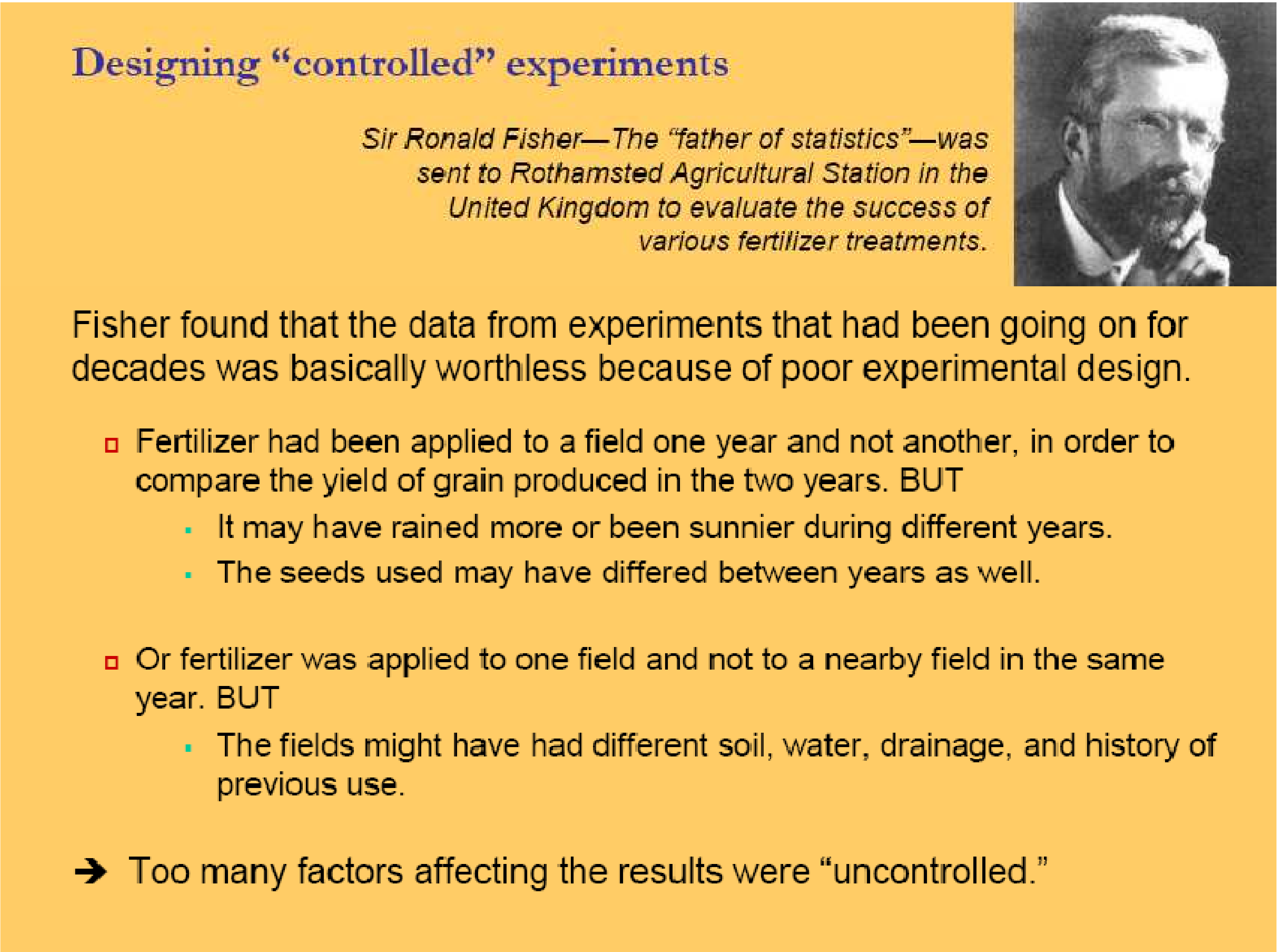
Lack of realism



Lack of realism is a serious weakness of experimentation. The subjects or treatments or setting of an experiment may not realistically duplicate the conditions we really want to study. In that case, we cannot generalize about the conclusions of the experiment.

Is the treatment appropriate for the response you want to study?

- Is studying the effects of eating red meat on cholesterol values in a group of middle aged men a realistic way to study factors affecting heart disease problems in humans?
- What about studying the effects of hair spray on rats to determine what will happen to women with big hair?



Designing "controlled" experiments

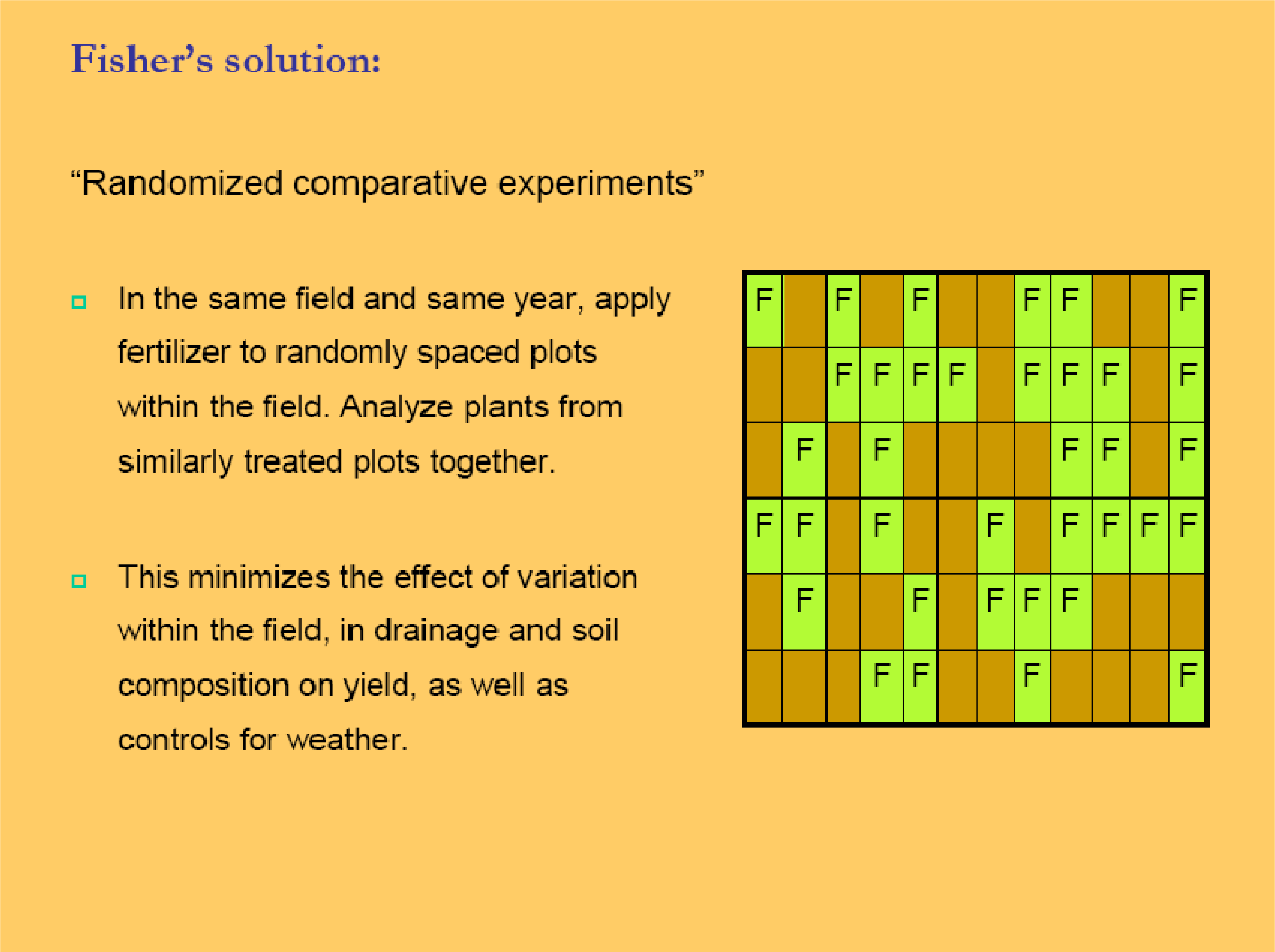
Sir Ronald Fisher—The "father of statistics"—was sent to Rothamsted Agricultural Station in the United Kingdom to evaluate the success of various fertilizer treatments.

Fisher found that the data from experiments that had been going on for decades was basically worthless because of poor experimental design.

- Fertilizer had been applied to a field one year and not another, in order to compare the yield of grain produced in the two years. BUT
  - It may have rained more or been sunnier during different years.
  - The seeds used may have differed between years as well.

- Or fertilizer was applied to one field and not to a nearby field in the same year. BUT
  - The fields might have had different soil, water, drainage, and history of previous use.

➔ Too many factors affecting the results were "uncontrolled."



Fisher's solution:

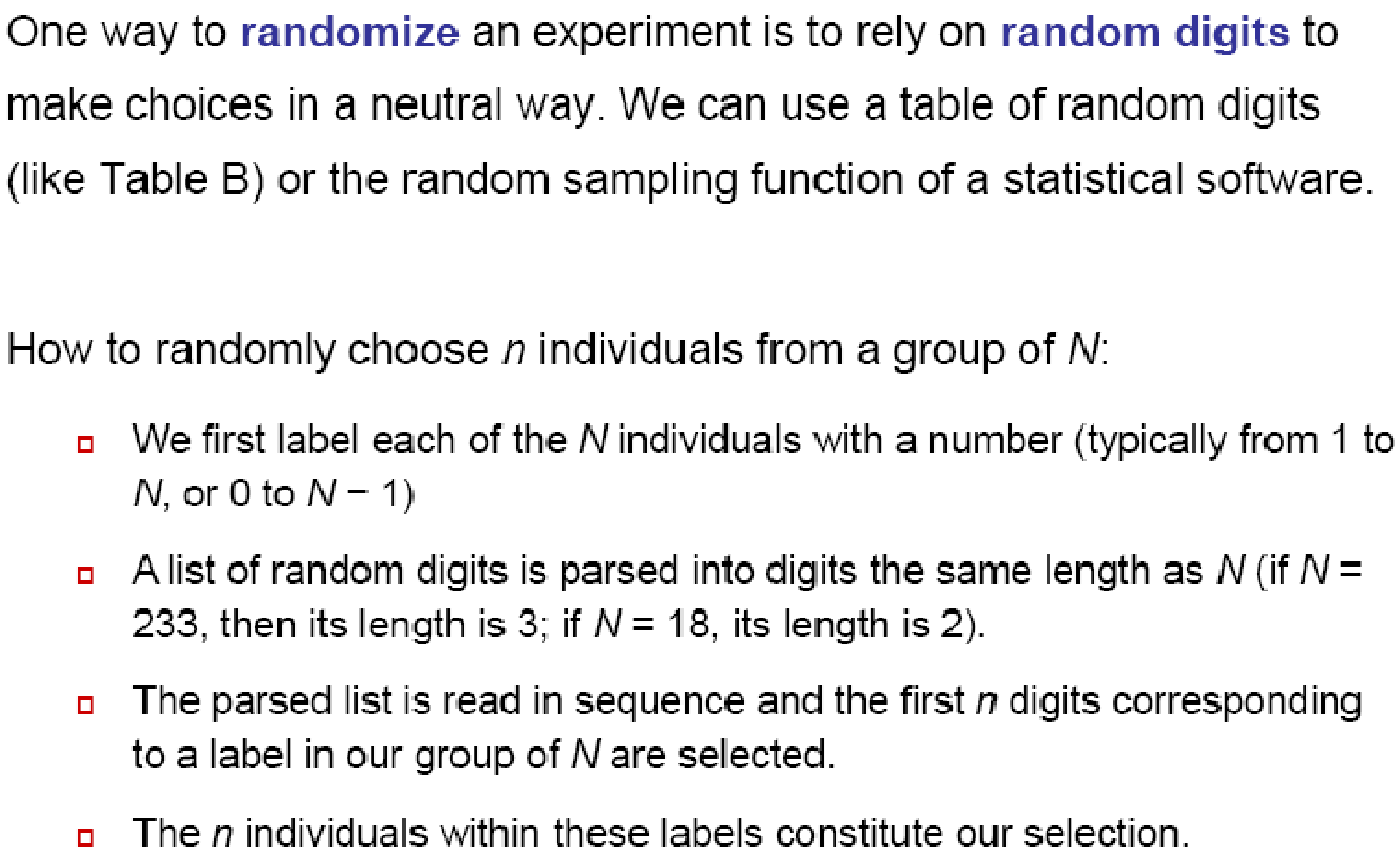
"Randomized comparative experiments"

- In the same field and same year, apply fertilizer to randomly spaced plots within the field. Analyze plants from similarly treated plots together.

- This minimizes the effect of variation within the field, in drainage and soil composition on yield, as well as controls for weather.

F		F		F			F	F			F
		F	F	F	F		F	F	F		F
	F		F					F	F		F
F	F		F			F		F	F	F	F
	F			F		F	F	F			
			F	F			F				F

Randomization



One way to **randomize** an experiment is to rely on **random digits** to make choices in a neutral way. We can use a table of random digits (like Table B) or the random sampling function of a statistical software.

How to randomly choose *n* individuals from a group of *N*:

- We first label each of the *N* individuals with a number (typically from 1 to *N*, or 0 to *N* − 1)
- A list of random digits is parsed into digits the same length as *N* (if *N* = 233, then its length is 3; if *N* = 18, its length is 2).
- The parsed list is read in sequence and the first *n* digits corresponding to a label in our group of *N* are selected.
- The *n* individuals within these labels constitute our selection.

Using Table B

We need to randomly select five students from a class of 20.

1. Since the class is of 20 people, list and number all members as 01,02,…20.

2. The number 20 is two digits long, so parse the list of random digits into numbers that are two digits long. Here we chose to start with line 103 for no particular reason.

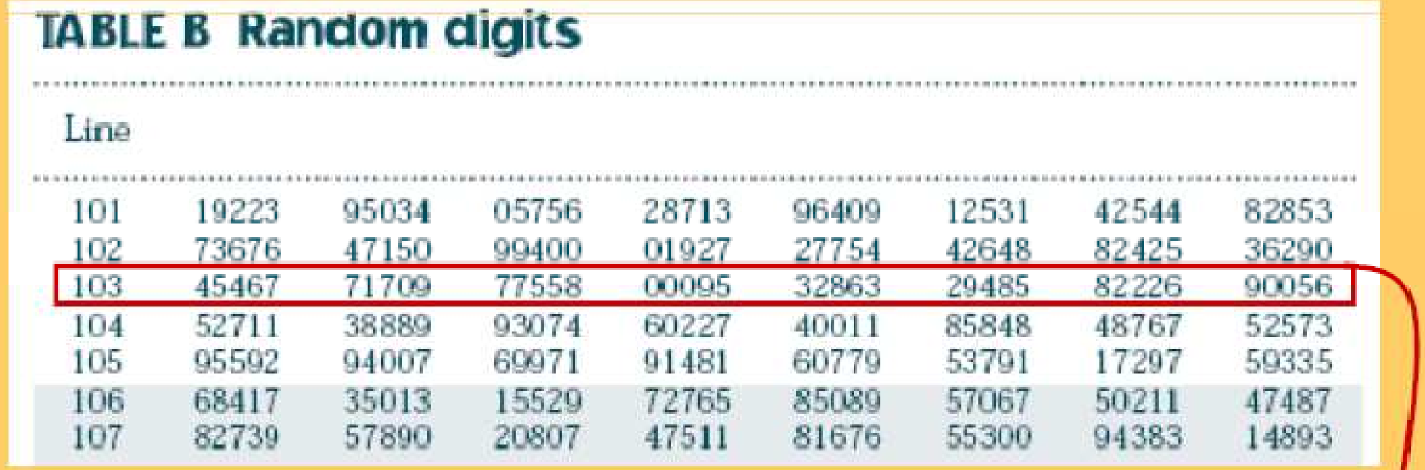
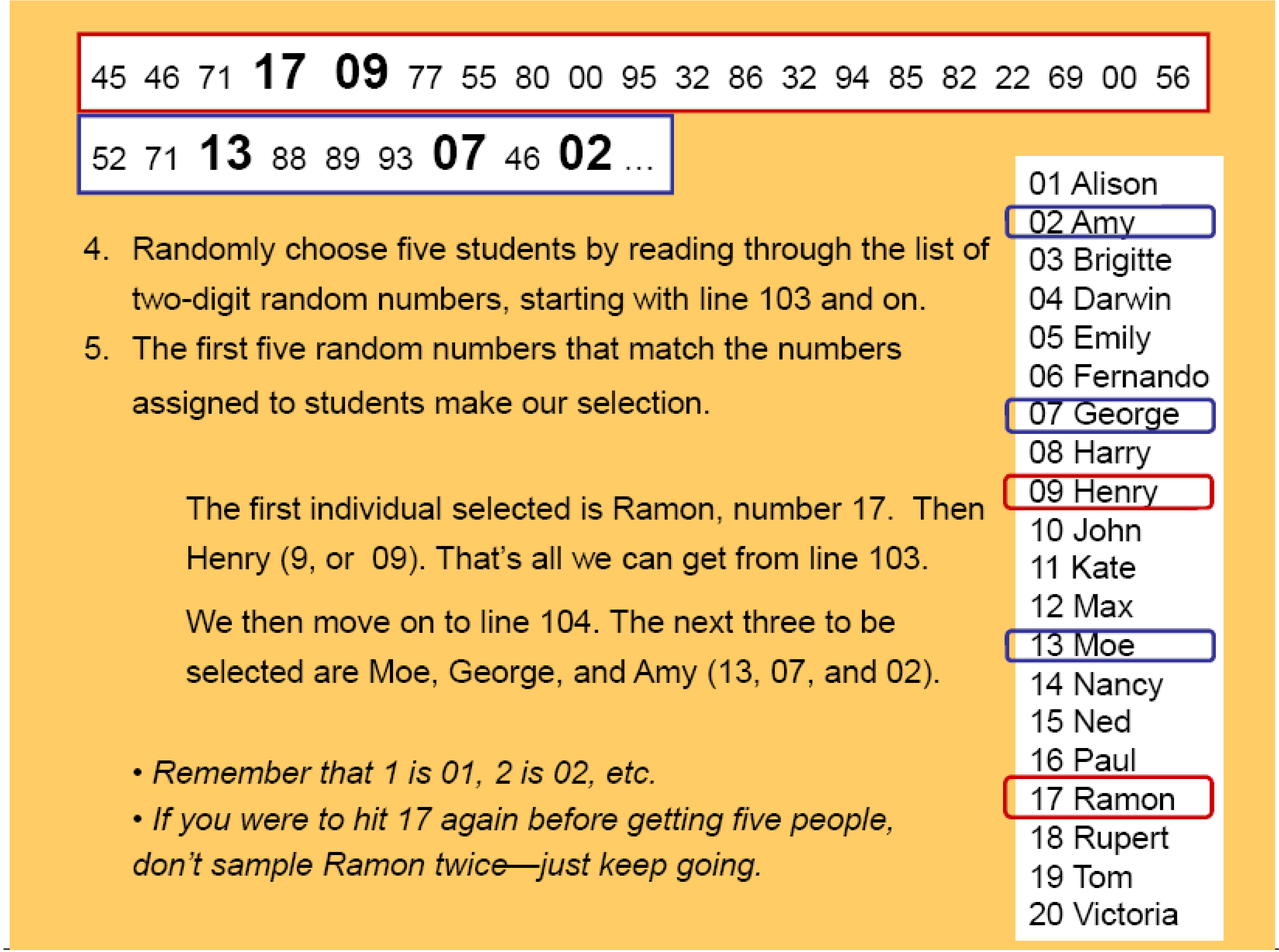


TABLE B Random digits

Line								
101	19223	95034	05756	28713	96409	12531	42544	82853
102	73676	47150	99400	01927	27754	42648	82425	36290
103	45467	71709	77558	00095	32863	29485	82226	90056
104	52711	38889	93074	60227	40011	85848	48767	52573
105	95592	94007	69971	91481	60779	53791	17297	59335
106	68417	35013	15529	72765	85089	57067	50211	47487
107	82739	57890	20807	47511	81676	55300	94383	14893

45 46 71 17 09 77 55 80 00 95 32 86 32 94 85 82 22 69 00 56



45 46 71 **17** **09** 77 55 80 00 95 32 86 32 94 85 82 22 69 00 56

52 71 **13** 88 89 93 **07** 46 **02** …

4. Randomly choose five students by reading through the list of two-digit random numbers, starting with line 103 and on.
5. The first five random numbers that match the numbers assigned to students make our selection.

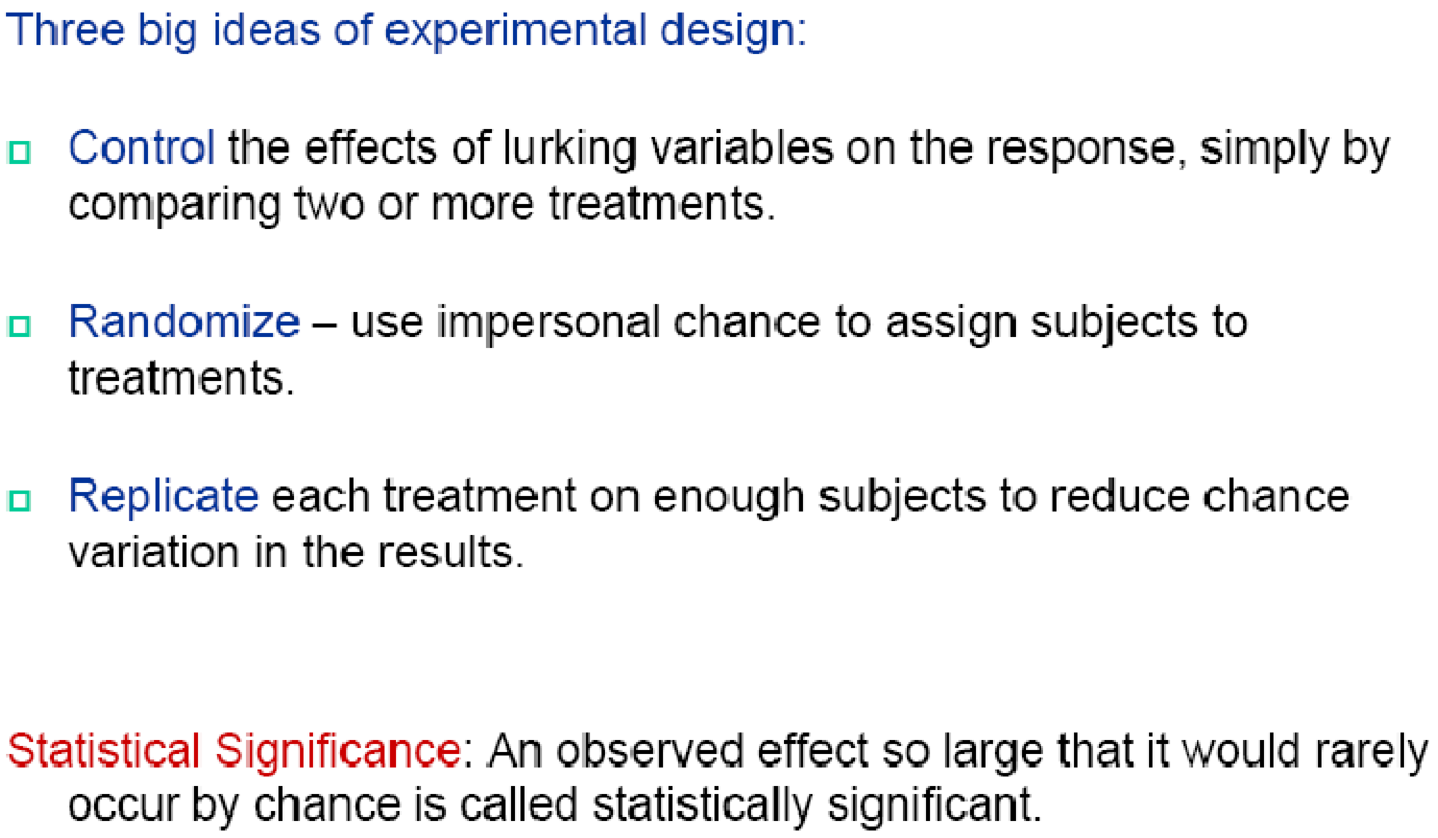
The first individual selected is Ramon, number 17. Then Henry (9, or 09). That's all we can get from line 103.

We then move on to line 104. The next three to be selected are Moe, George, and Amy (13, 07, and 02).

- *Remember that 1 is 01, 2 is 02, etc.*
- *If you were to hit 17 again before getting five people, don't sample Ramon twice—just keep going.*

01	Alison
02	Amy
03	Brigitte
04	Darwin
05	Emily
06	Fernando
07	George
08	Harry
09	Henry
10	John
11	Kate
12	Max
13	Moe
14	Nancy
15	Ned
16	Paul
17	Ramon
18	Rupert
19	Tom
20	Victoria

Principles of experimental design



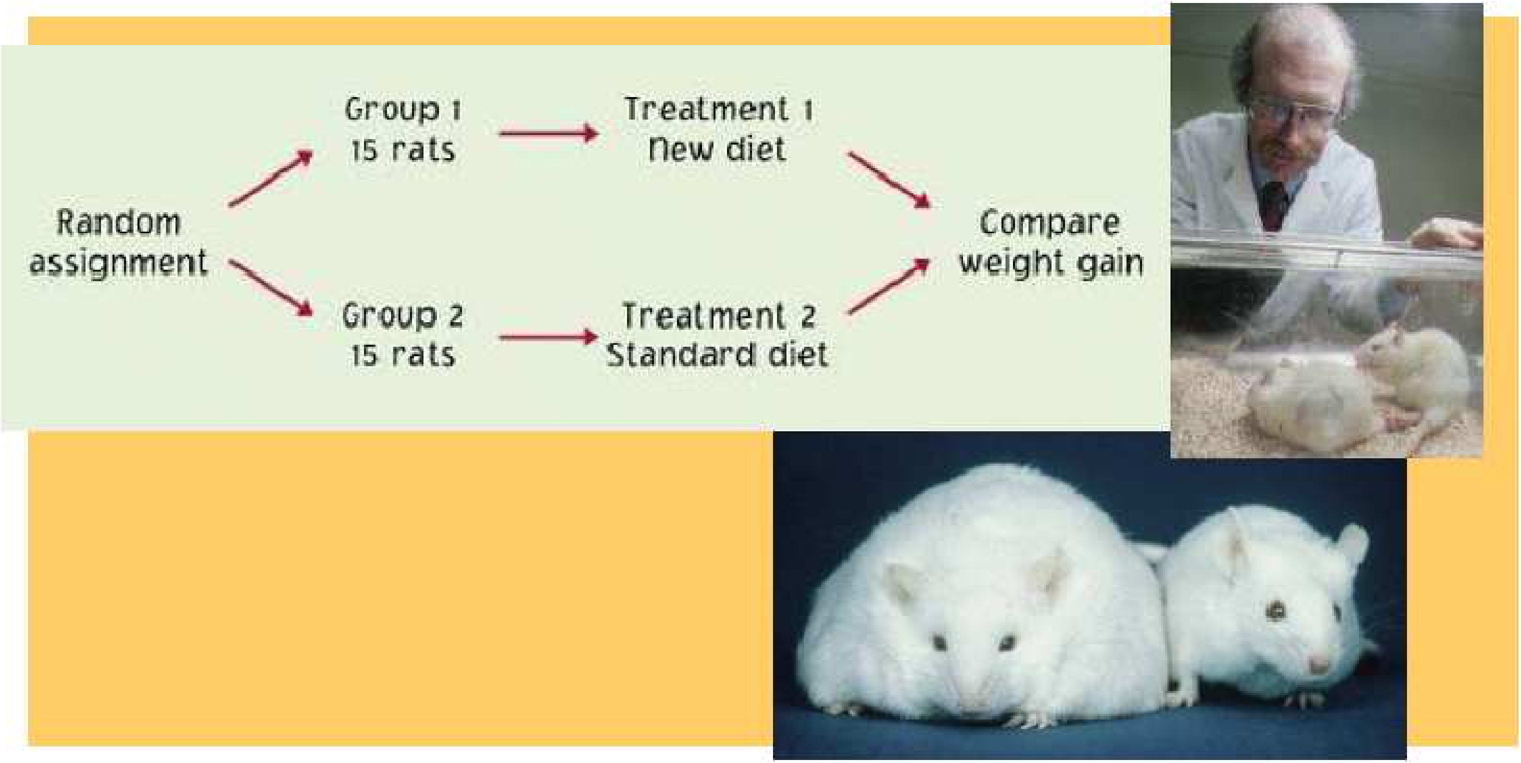
Three big ideas of experimental design:

- Control the effects of lurking variables on the response, simply by comparing two or more treatments.
- Randomize – use impersonal chance to assign subjects to treatments.
- Replicate each treatment on enough subjects to reduce chance variation in the results.

Statistical Significance: An observed effect so large that it would rarely occur by chance is called statistically significant.

Completely randomized designs

Completely randomized experimental designs: Individuals are randomly assigned to groups, then the groups are randomly assigned to treatments.



Random assignment

Group 1
15 rats

Group 2
15 rats

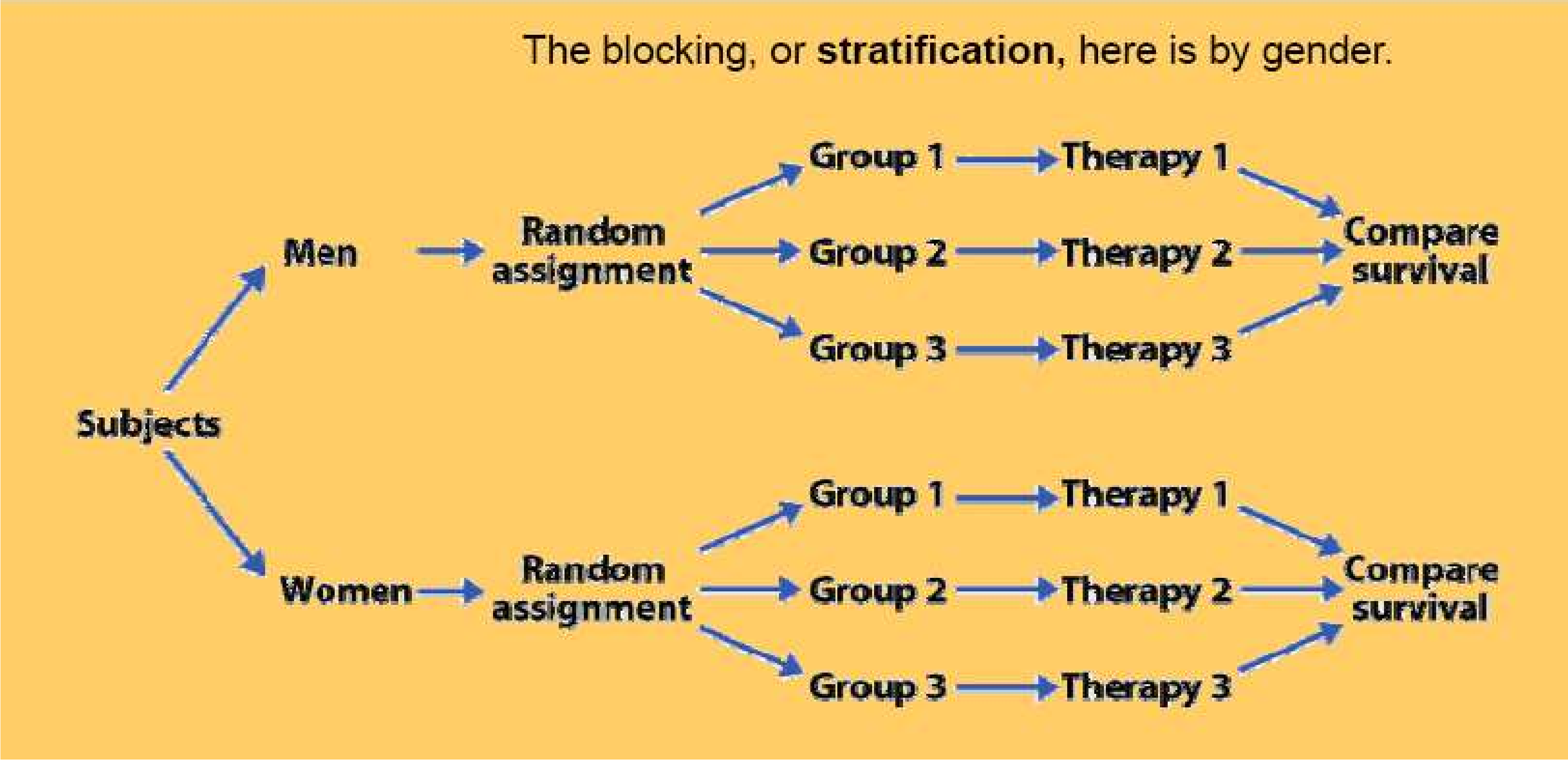
Treatment 1
New diet

Treatment 2
Standard diet

Compare weight gain

Block designs

In a **block,** or **stratified,** design, subjects are divided into groups, or blocks, prior to experiments, to test hypotheses about differences between the groups.



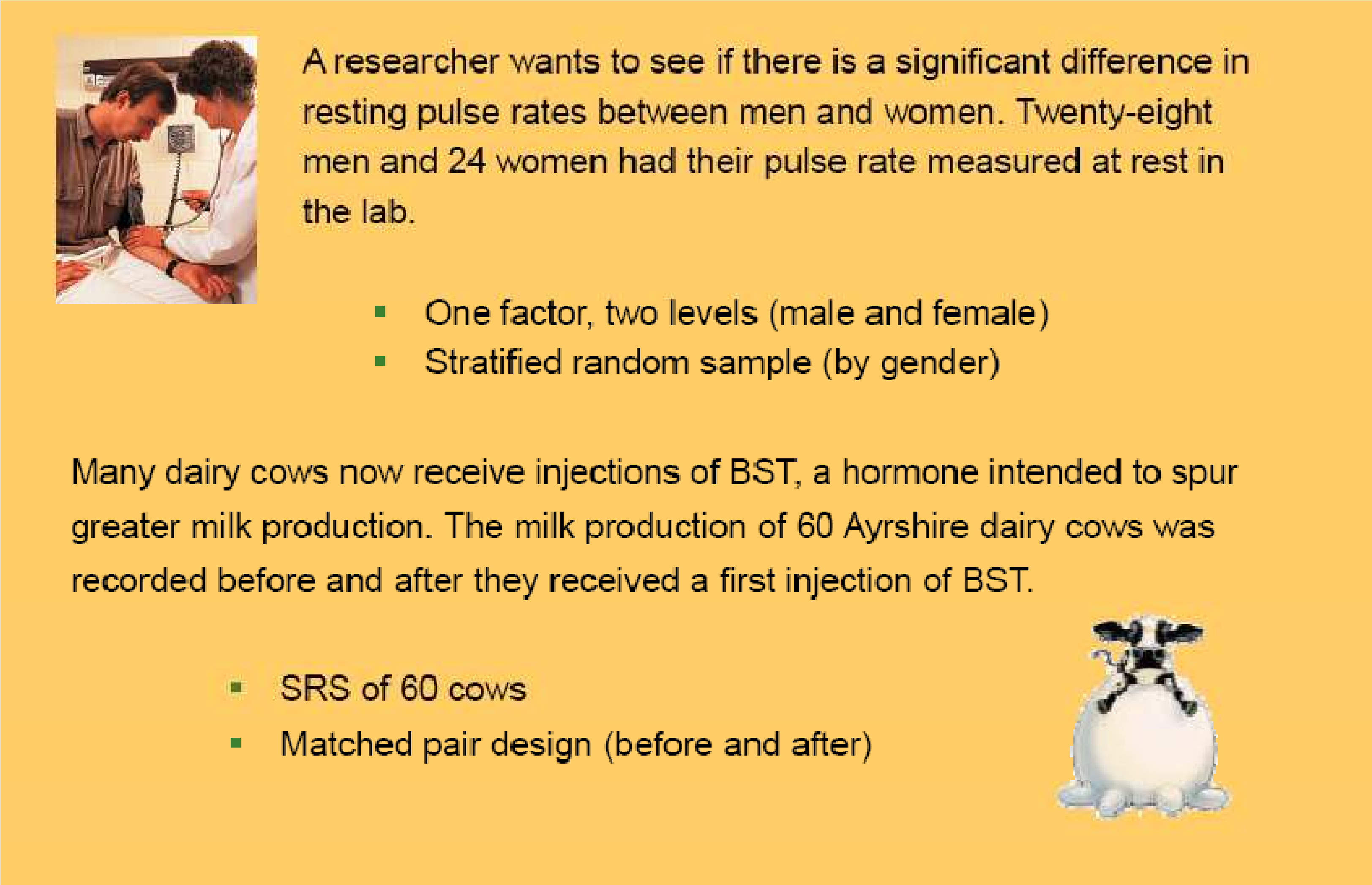
Matched pairs designs

Matched pairs: Choose pairs of subjects that are closely matched—e.g., same sex, height, weight, age, and race. Within each pair, randomly assign who will receive which treatment.

It is also possible to just use a single person, and give the two treatments to this person over time in random order. In this case, the "matched pair" is just the same person at different points in time.

The most closely matched pair studies use identical twins.

Why experimental designs ?



A researcher wants to see if there is a significant difference in resting pulse rates between men and women. Twenty-eight men and 24 women had their pulse rate measured at rest in the lab.

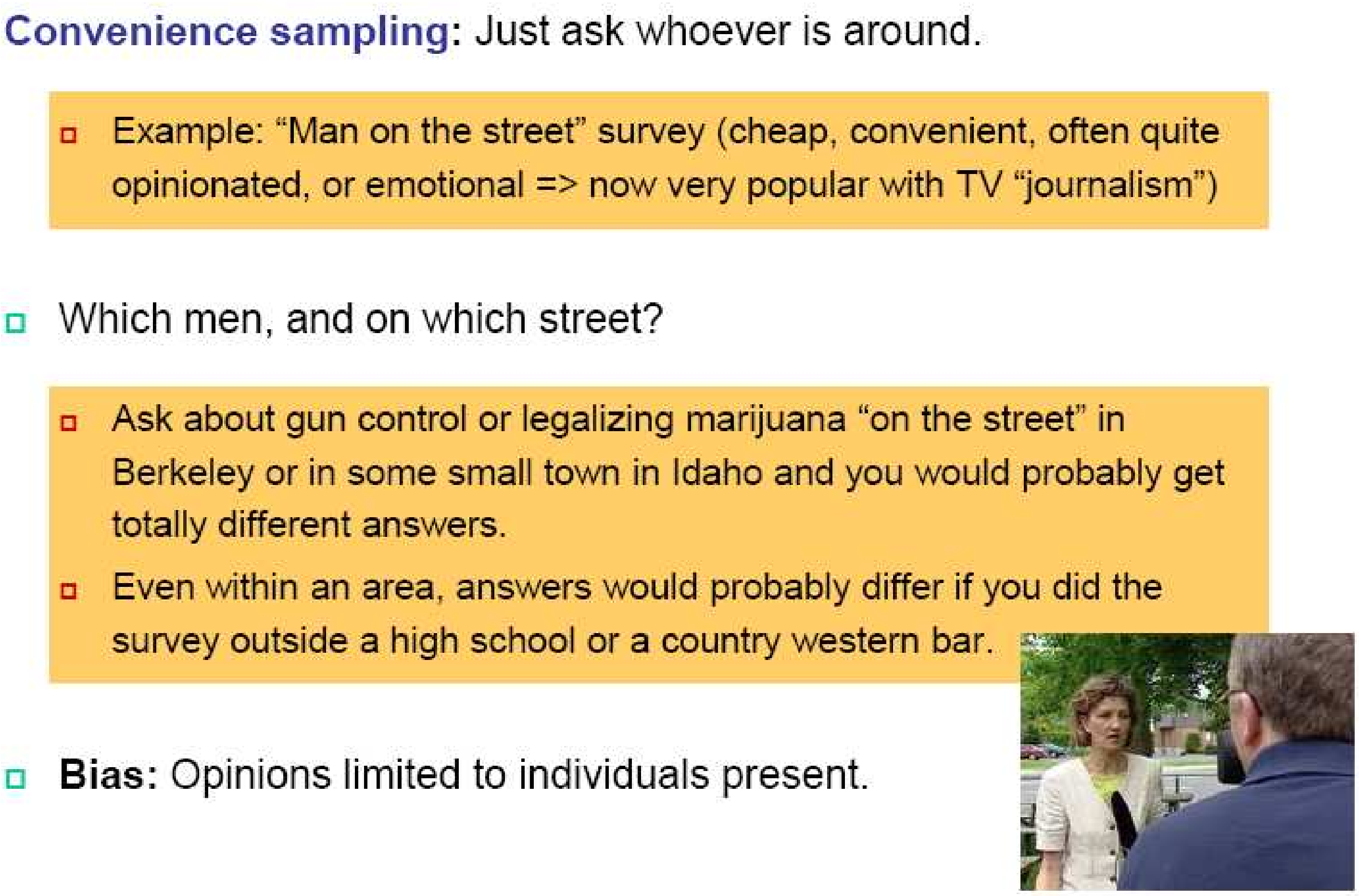
- One factor, two levels (male and female)
- Stratified random sample (by gender)

Many dairy cows now receive injections of BST, a hormone intended to spur greater milk production. The milk production of 60 Ayrshire dairy cows was recorded before and after they received a first injection of BST.

- SRS of 60 cows
- Matched pair design (before and after)

2.2 Sampling designs and towards inference

Sampling methods



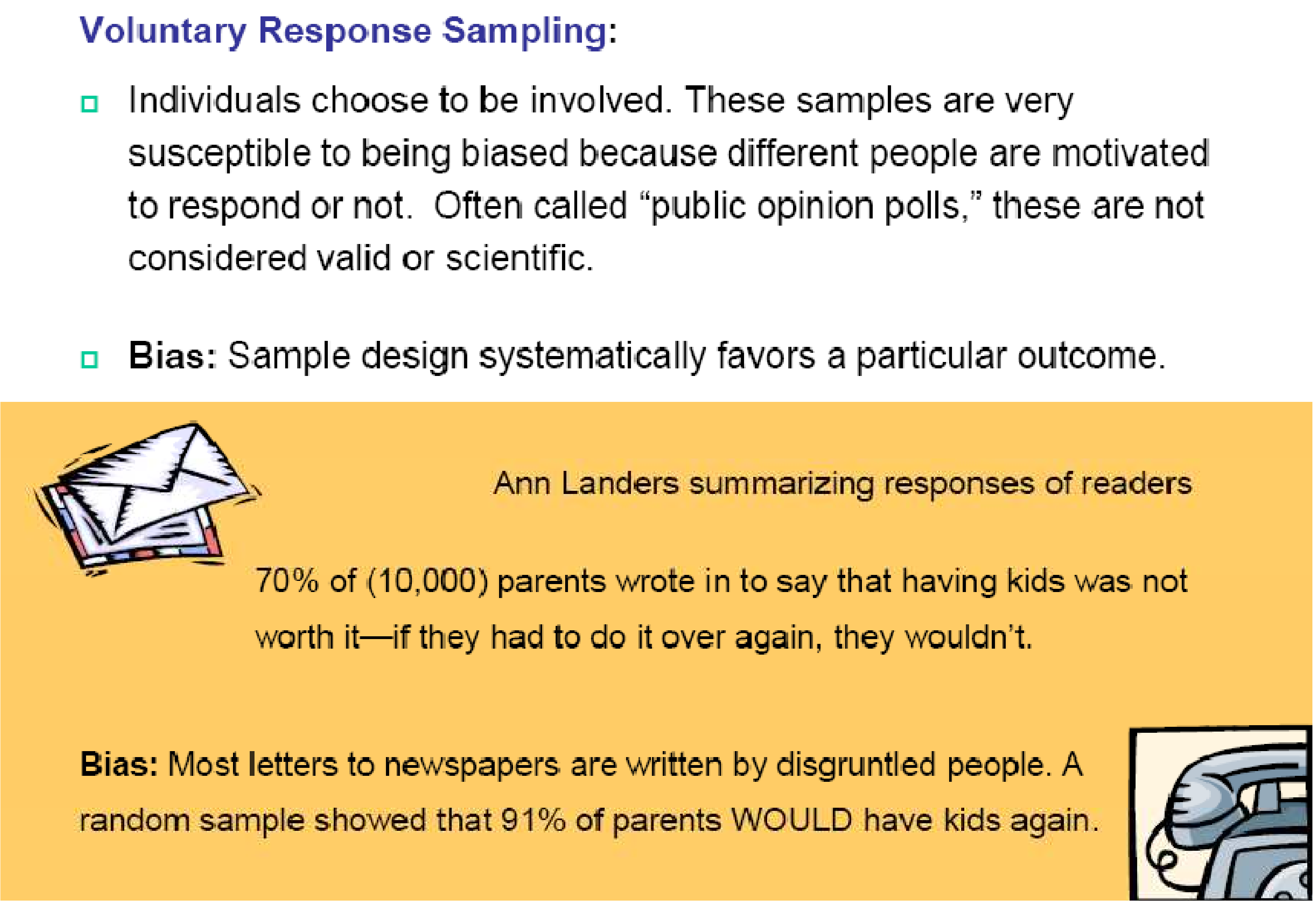
Convenience sampling: Just ask whoever is around.

- Example: "Man on the street" survey (cheap, convenient, often quite opinionated, or emotional => now very popular with TV "journalism")

- Which men, and on which street?

  - Ask about gun control or legalizing marijuana "on the street" in Berkeley or in some small town in Idaho and you would probably get totally different answers.
  - Even within an area, answers would probably differ if you did the survey outside a high school or a country western bar.

- **Bias:** Opinions limited to individuals present.



Voluntary Response Sampling:

- Individuals choose to be involved. These samples are very susceptible to being biased because different people are motivated to respond or not. Often called "public opinion polls," these are not considered valid or scientific.

- **Bias:** Sample design systematically favors a particular outcome.

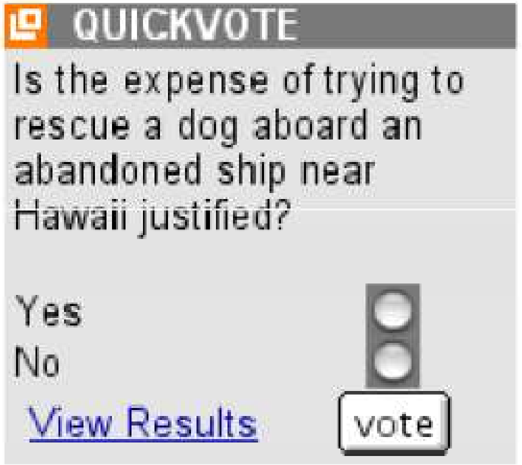
Ann Landers summarizing responses of readers

70% of (10,000) parents wrote in to say that having kids was not worth it—if they had to do it over again, they wouldn't.

Bias: Most letters to newspapers are written by disgruntled people. A random sample showed that 91% of parents WOULD have kids again.

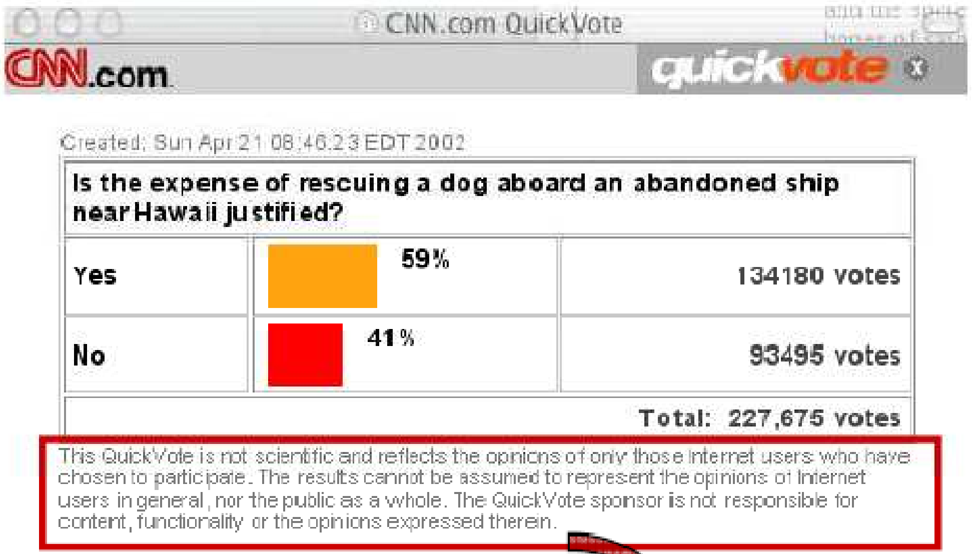
CNN on-line surveys:

Bias: People have to care enough about an issue to bother replying. This sample is probably a combination of people who hate "wasting the taxpayers money" and "animal lovers."



Is the expense of trying to rescue a dog aboard an abandoned ship near Hawaii justified?

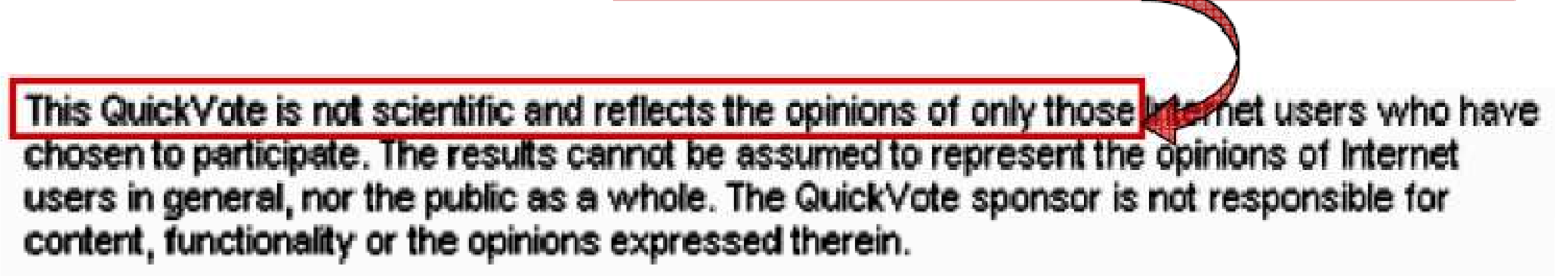
○ Yes
○ No



Created: Sun Apr 21 08:46:23 EDT 2002

Is the expense of rescuing a dog aboard an abandoned ship near Hawaii justified?		
Yes	59%	134180 votes
No	41%	93495 votes
		Total: 227,675 votes

This QuickVote is not scientific and reflects the opinions of only those Internet users who have chosen to participate. The results cannot be assumed to represent the opinions of Internet users in general, nor the public as a whole. The QuickVote sponsor is not responsible for content, functionality or the opinions expressed therein.



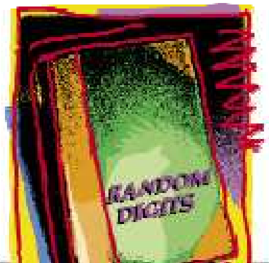
This QuickVote is not scientific and reflects the opinions of only those Internet users who have chosen to participate. The results cannot be assumed to represent the opinions of Internet users in general, nor the public as a whole. The QuickVote sponsor is not responsible for content, functionality or the opinions expressed therein.

In contrast :

Probability or **random sampling**:

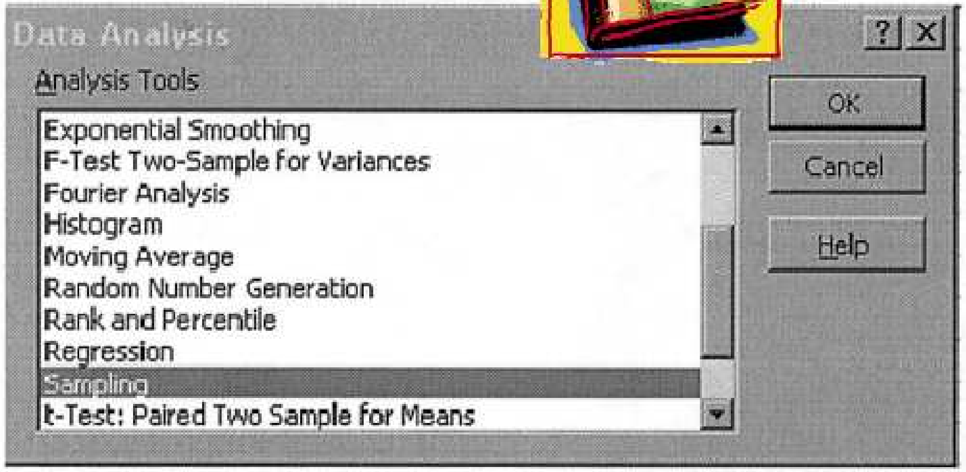
- Individuals are randomly selected. No one group should be over-represented.

Sampling randomly gets rid of bias.

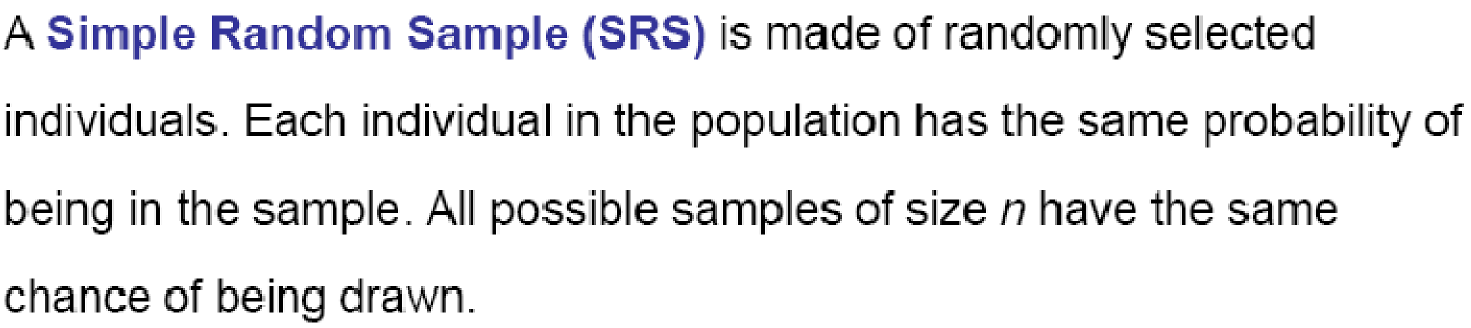


Random samples rely on the absolute objectivity of random numbers. There are tables and books of random digits available for random sampling.

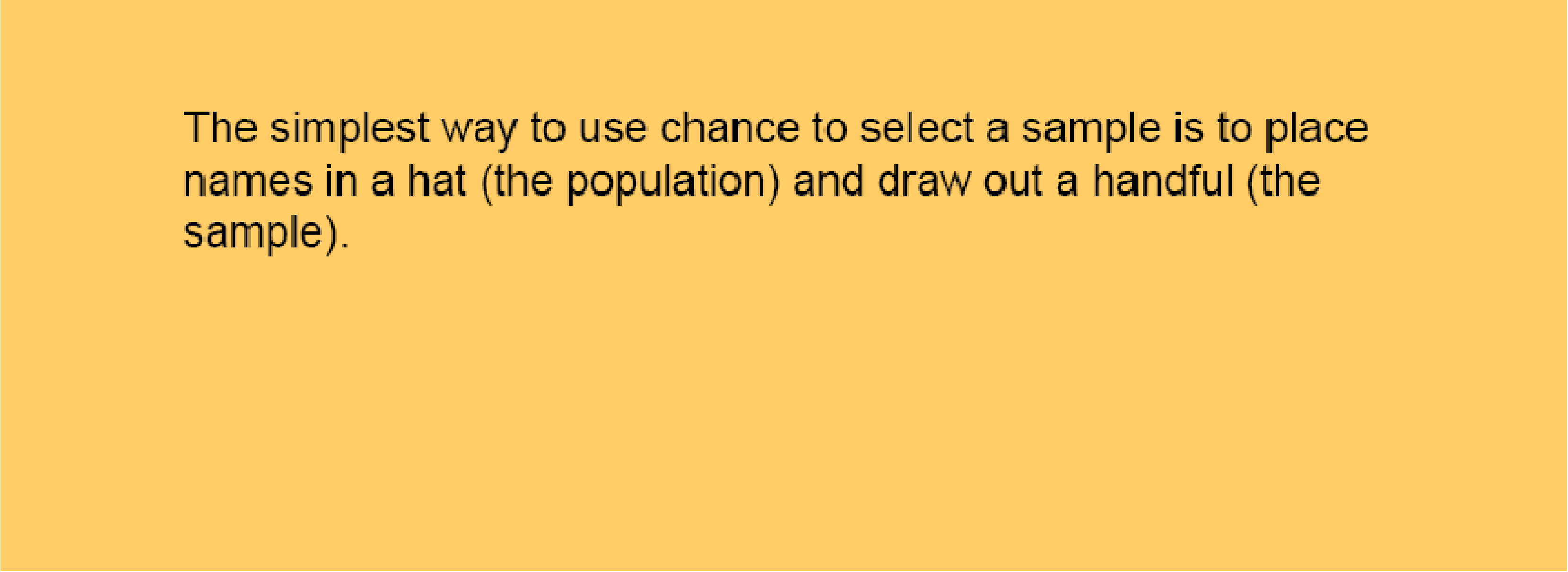
Statistical software can generate random digits (e.g., Excel “=random()”).



Simple random samples

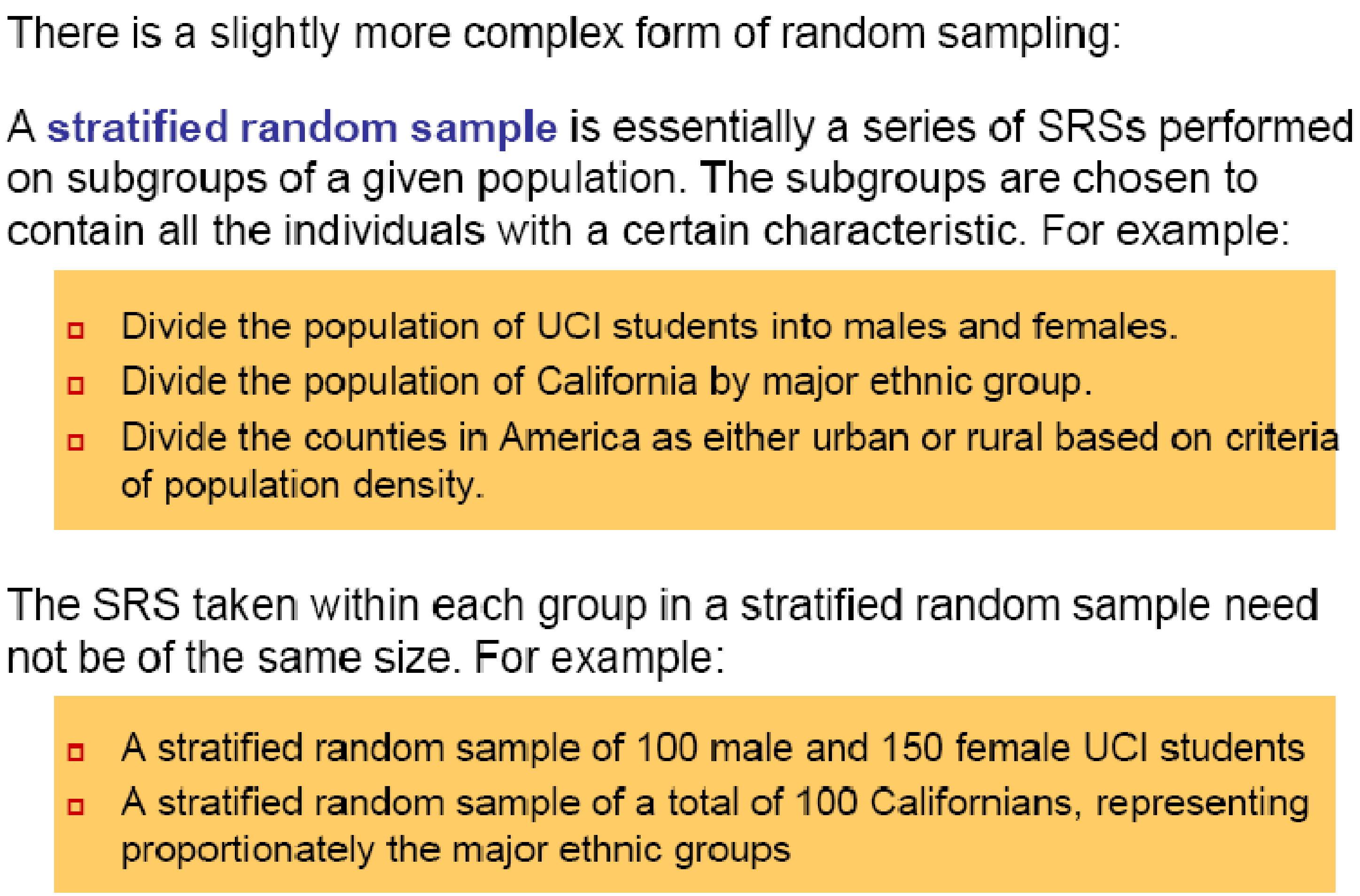


A **Simple Random Sample (SRS)** is made of randomly selected individuals. Each individual in the population has the same probability of being in the sample. All possible samples of size *n* have the same chance of being drawn.



The simplest way to use chance to select a sample is to place names in a hat (the population) and draw out a handful (the sample).

Stratified samples



There is a slightly more complex form of random sampling:

A **stratified random sample** is essentially a series of SRSs performed on subgroups of a given population. The subgroups are chosen to contain all the individuals with a certain characteristic. For example:

- Divide the population of UCI students into males and females.
- Divide the population of California by major ethnic group.
- Divide the counties in America as either urban or rural based on criteria of population density.

The SRS taken within each group in a stratified random sample need not be of the same size. For example:

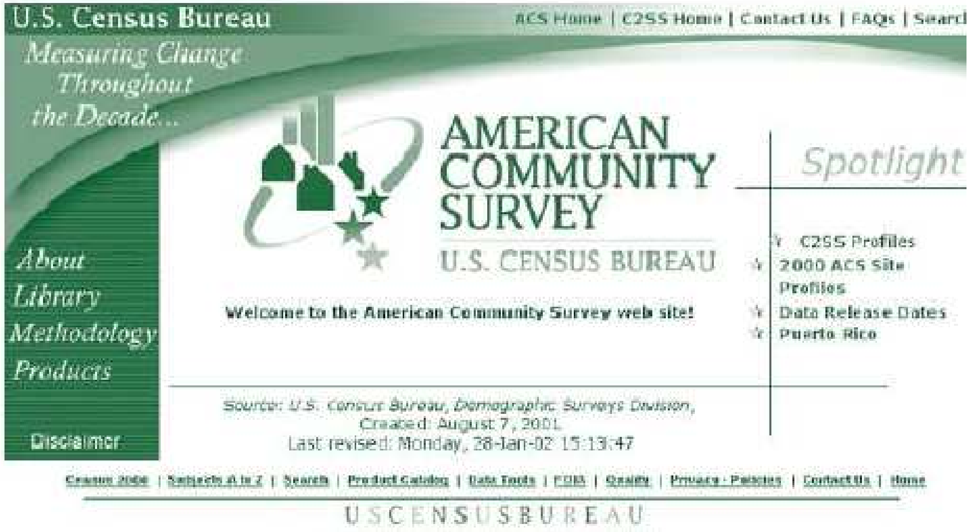
- A stratified random sample of 100 male and 150 female UCI students
- A stratified random sample of a total of 100 Californians, representing proportionately the major ethnic groups

Multistage samples use multiple stages of stratification. They are often used by the government to obtain information about the U.S. population.

Example: Sampling both urban and rural areas, people in different ethnic and income groups within the urban and rural areas, and then within those strata individuals of different ethnicities

Data are obtained by taking an SRS for each substrata.

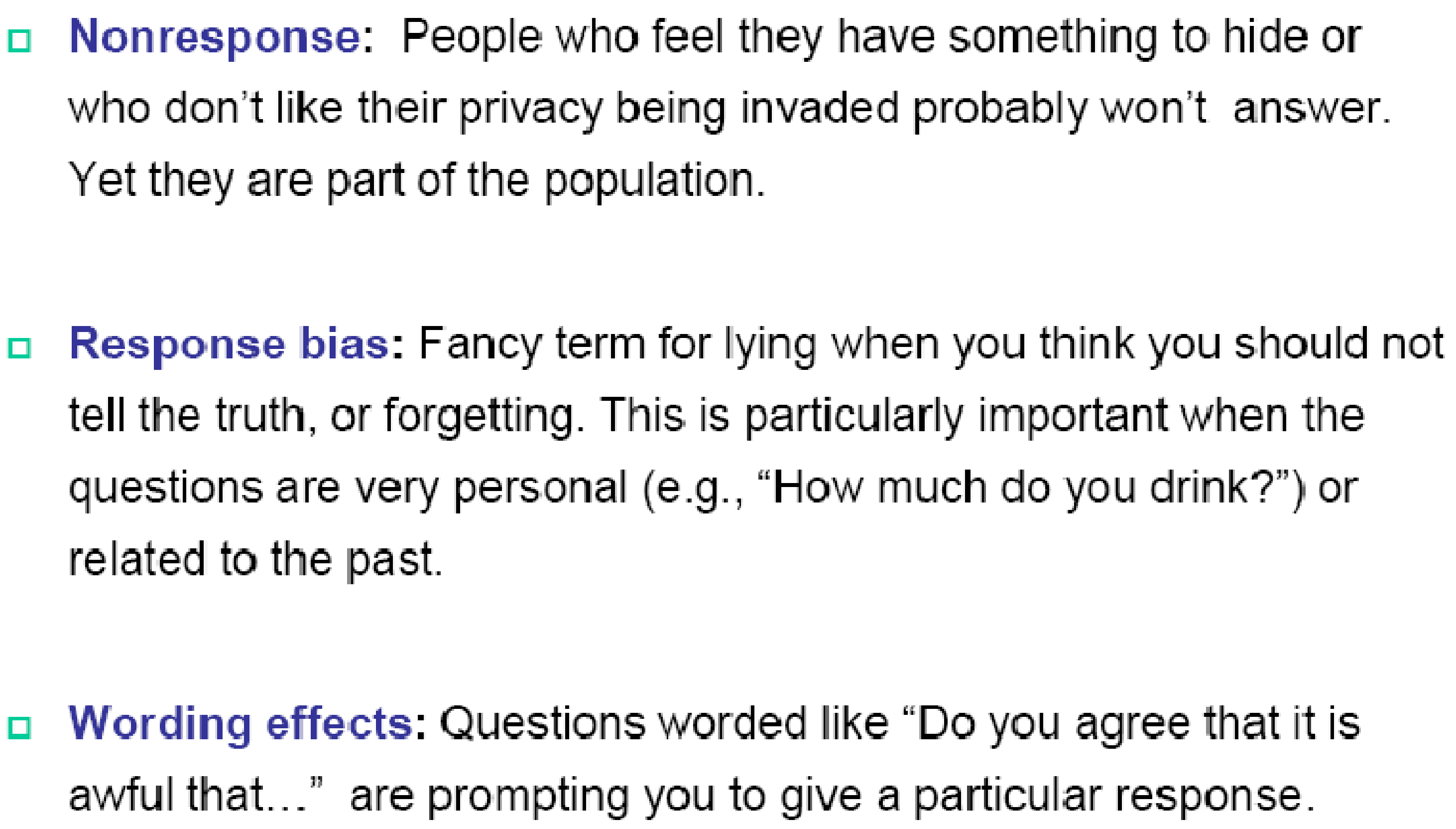
Statistical analysis for multistage samples is more complex than for an SRS.


U.S. Census Bureau
ACS Home | C2SS Home | Contact Us | FAQs | Search
Measuring Change Throughout the Decade...
AMERICAN COMMUNITY SURVEY
U.S. CENSUS BUREAU
Spotlight
C2SS Profiles
2000 ACS Site Profiles
Data Release Dates
Puerto Rico
About
Library
Methodology
Products
Disclaimer
Census 2000 | Subjects A to Z | Search | Product Catalog | Data Tools | FOIA | Quality | Privacy · Policies | Contact Us | Home
USCENSUSBUREAU

Welcome to the American Community Survey web site!

Source: U.S. Census Bureau, Demographic Surveys Division,
Created: August 7, 2001
Last revised: Monday, 28-Jan-02 15:13:47

Caution about sampling surveys



- **Nonresponse**: People who feel they have something to hide or who don't like their privacy being invaded probably won't answer. Yet they are part of the population.

- **Response bias**: Fancy term for lying when you think you should not tell the truth, or forgetting. This is particularly important when the questions are very personal (e.g., "How much do you drink?") or related to the past.

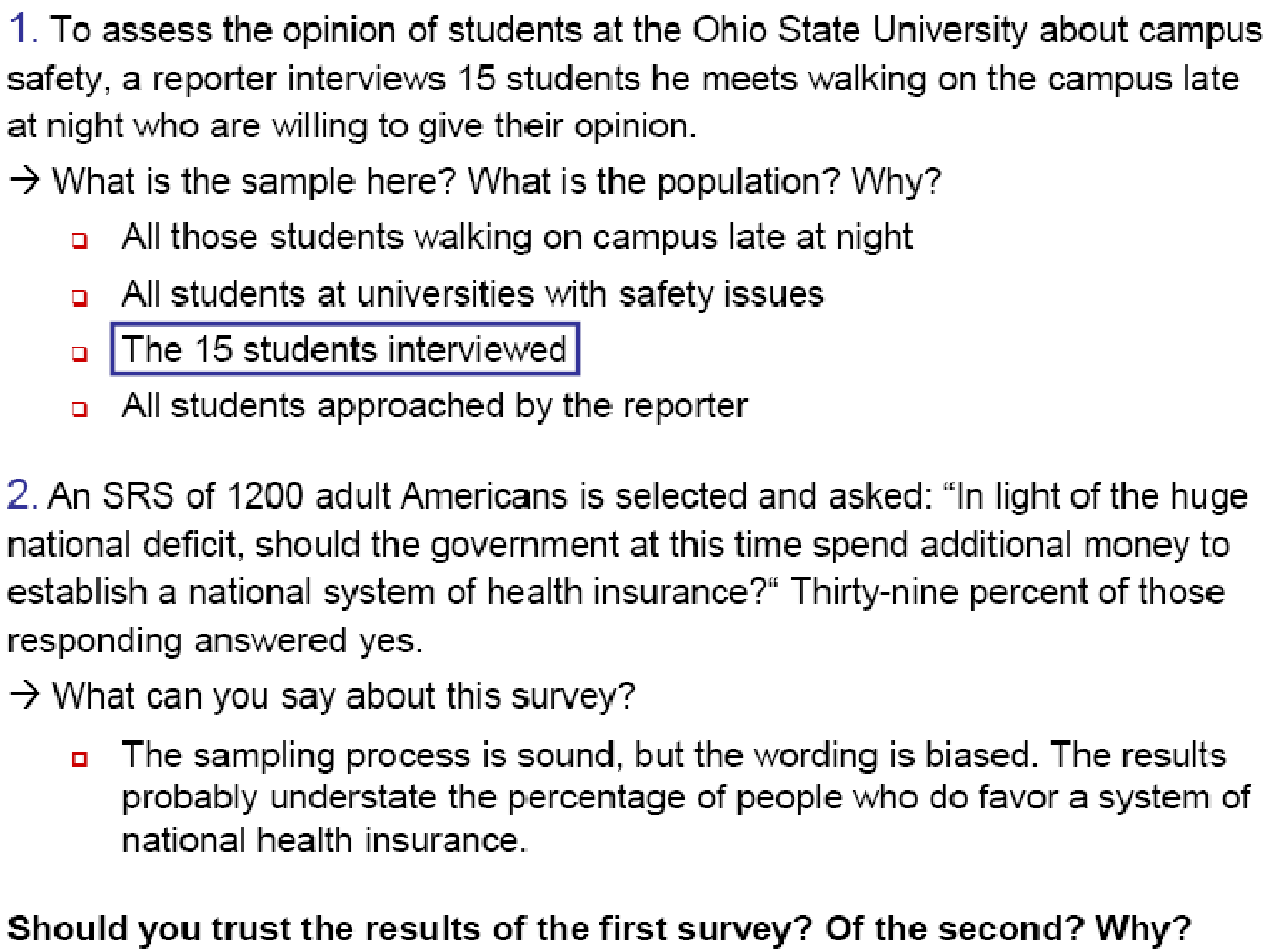
- **Wording effects**: Questions worded like "Do you agree that it is awful that…" are prompting you to give a particular response.

□ **Undercoverage:**

Occurs when parts of the population are left out in the process of choosing the sample.

Because the U.S. Census goes "house to house," homeless people are not represented. Illegal immigrants also avoid being counted. Geographical districts with a lack of coverage tend to be poor. Representatives from wealthy areas typically oppose statistical adjustment of the census.

Historically, clinical trials have avoided including women in their studies because of their periods and the chance of pregnancy. This means that medical treatments were not appropriately tested for women. This problem is slowly being recognized and addressed.



1. To assess the opinion of students at the Ohio State University about campus safety, a reporter interviews 15 students he meets walking on the campus late at night who are willing to give their opinion.

→ What is the sample here? What is the population? Why?

- □ All those students walking on campus late at night
- □ All students at universities with safety issues
- □ [✓ The 15 students interviewed]
- □ All students approached by the reporter

2. An SRS of 1200 adult Americans is selected and asked: “In light of the huge national deficit, should the government at this time spend additional money to establish a national system of health insurance?“ Thirty-nine percent of those responding answered yes.

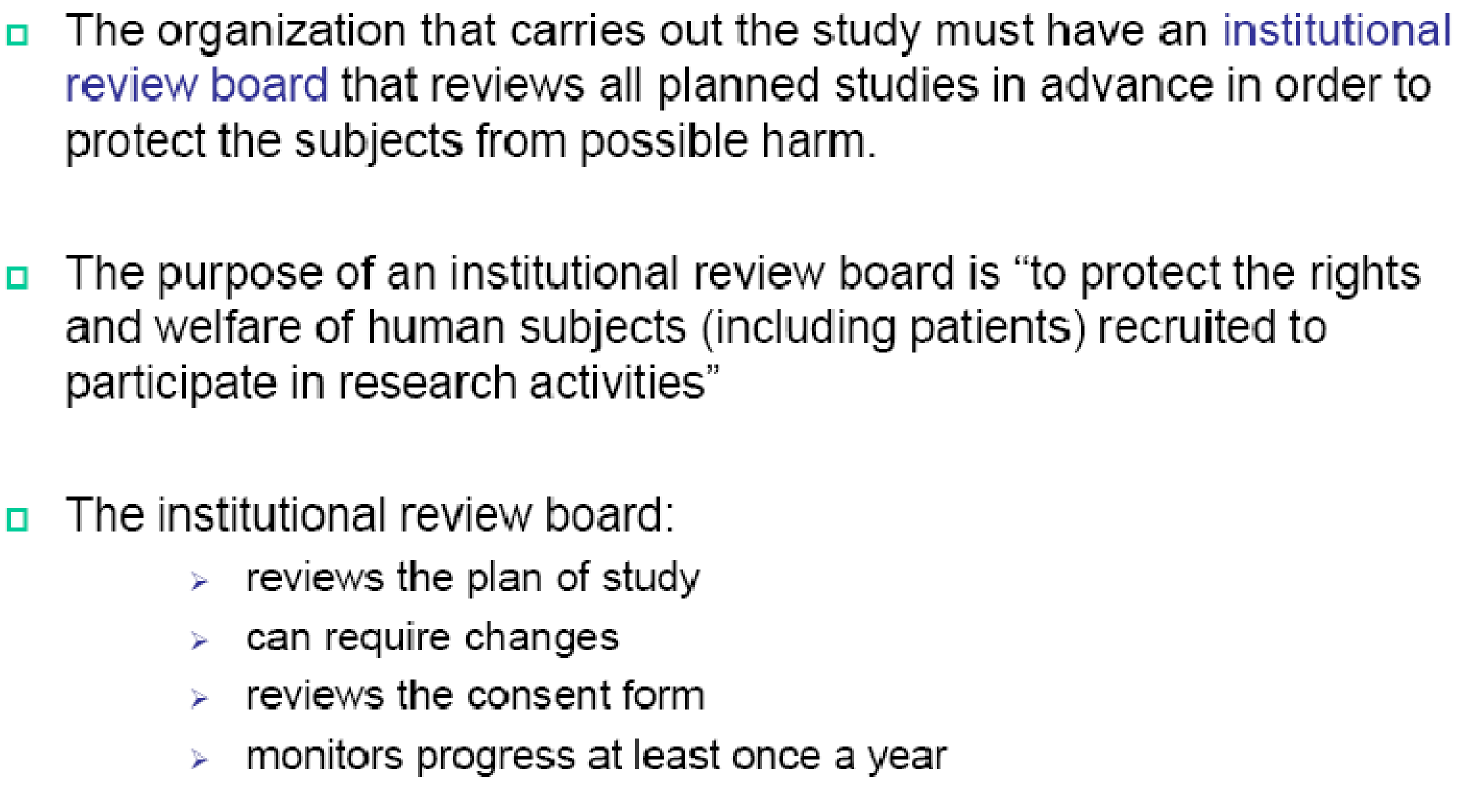
→ What can you say about this survey?

- □ The sampling process is sound, but the wording is biased. The results probably understate the percentage of people who do favor a system of national health insurance.

Should you trust the results of the first survey? Of the second? Why?

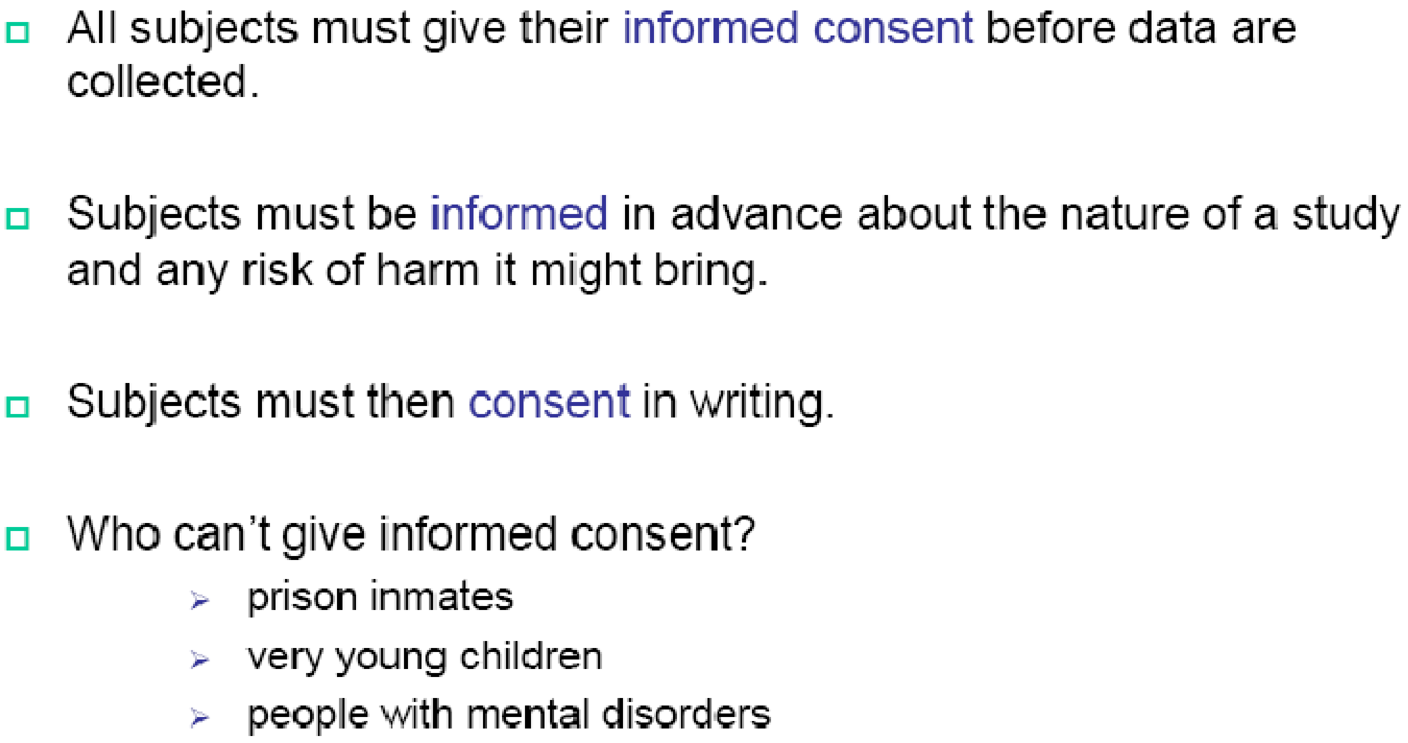
2.3 Ethics

Institutional Review Boards



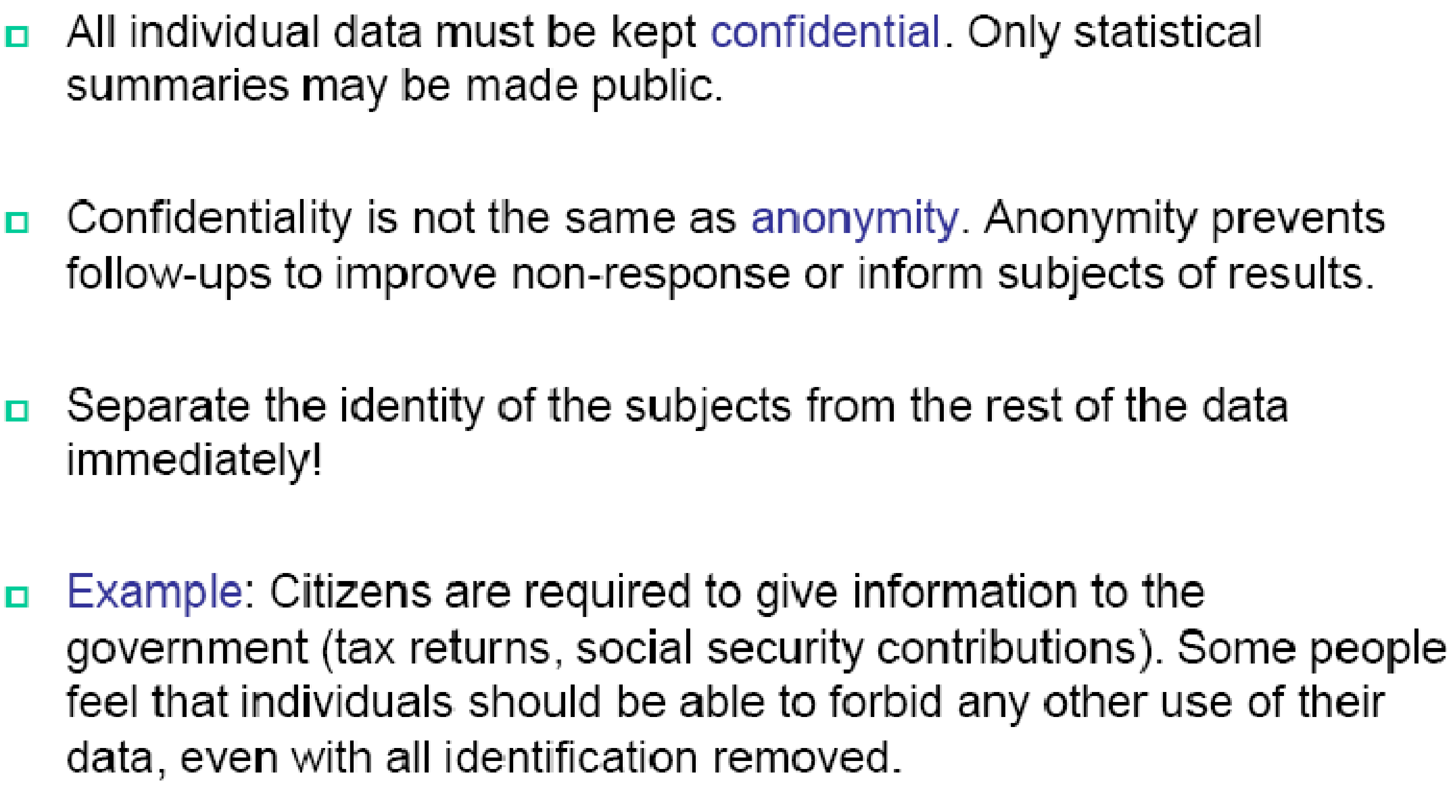
- The organization that carries out the study must have an institutional review board that reviews all planned studies in advance in order to protect the subjects from possible harm.
- The purpose of an institutional review board is “to protect the rights and welfare of human subjects (including patients) recruited to participate in research activities”
- The institutional review board:
    - reviews the plan of study
    - can require changes
    - reviews the consent form
    - monitors progress at least once a year

Informed Consent



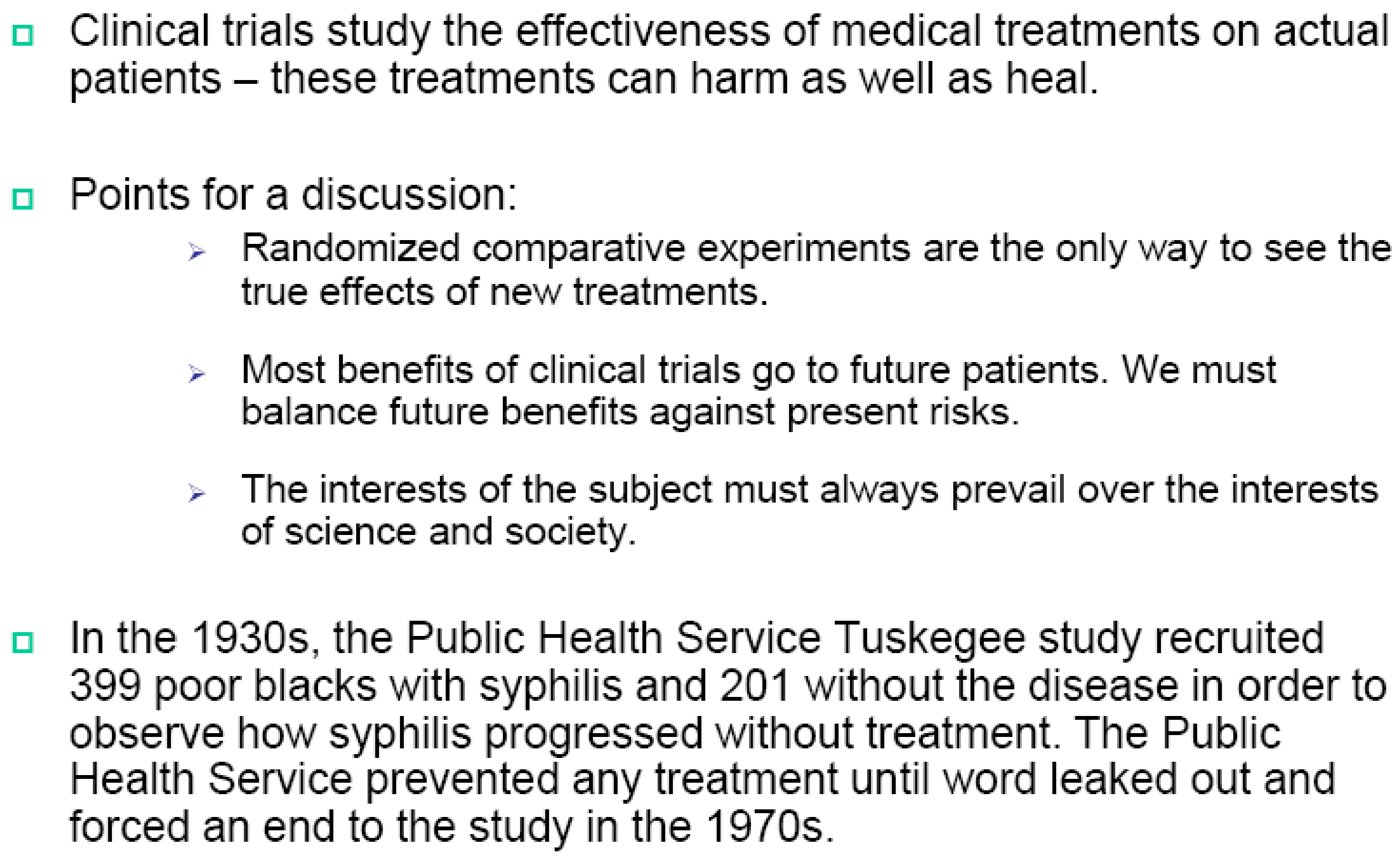
- All subjects must give their informed consent before data are collected.
- Subjects must be informed in advance about the nature of a study and any risk of harm it might bring.
- Subjects must then consent in writing.
- Who can't give informed consent?
  - prison inmates
  - very young children
  - people with mental disorders

Confidentiality



- All individual data must be kept confidential. Only statistical summaries may be made public.
- Confidentiality is not the same as anonymity. Anonymity prevents follow-ups to improve non-response or inform subjects of results.
- Separate the identity of the subjects from the rest of the data immediately!
- Example: Citizens are required to give information to the government (tax returns, social security contributions). Some people feel that individuals should be able to forbid any other use of their data, even with all identification removed.

Clinical trials

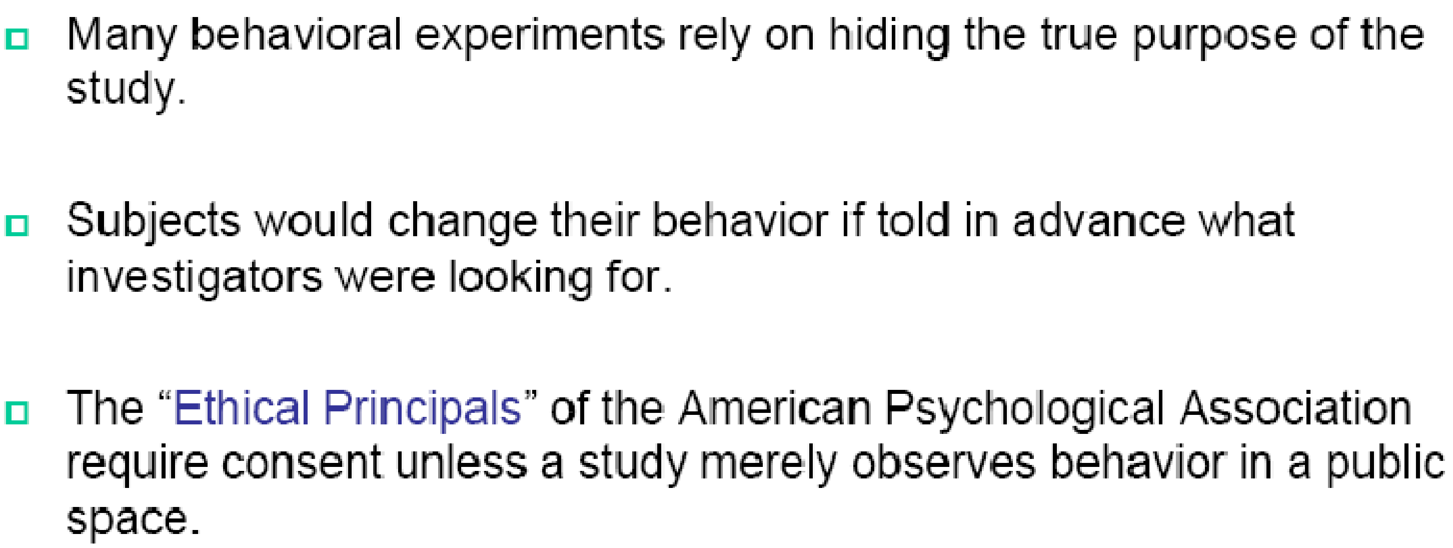


- Clinical trials study the effectiveness of medical treatments on actual patients – these treatments can harm as well as heal.

- Points for a discussion:
  - Randomized comparative experiments are the only way to see the true effects of new treatments.
  - Most benefits of clinical trials go to future patients. We must balance future benefits against present risks.
  - The interests of the subject must always prevail over the interests of science and society.

- In the 1930s, the Public Health Service Tuskegee study recruited 399 poor blacks with syphilis and 201 without the disease in order to observe how syphilis progressed without treatment. The Public Health Service prevented any treatment until word leaked out and forced an end to the study in the 1970s.

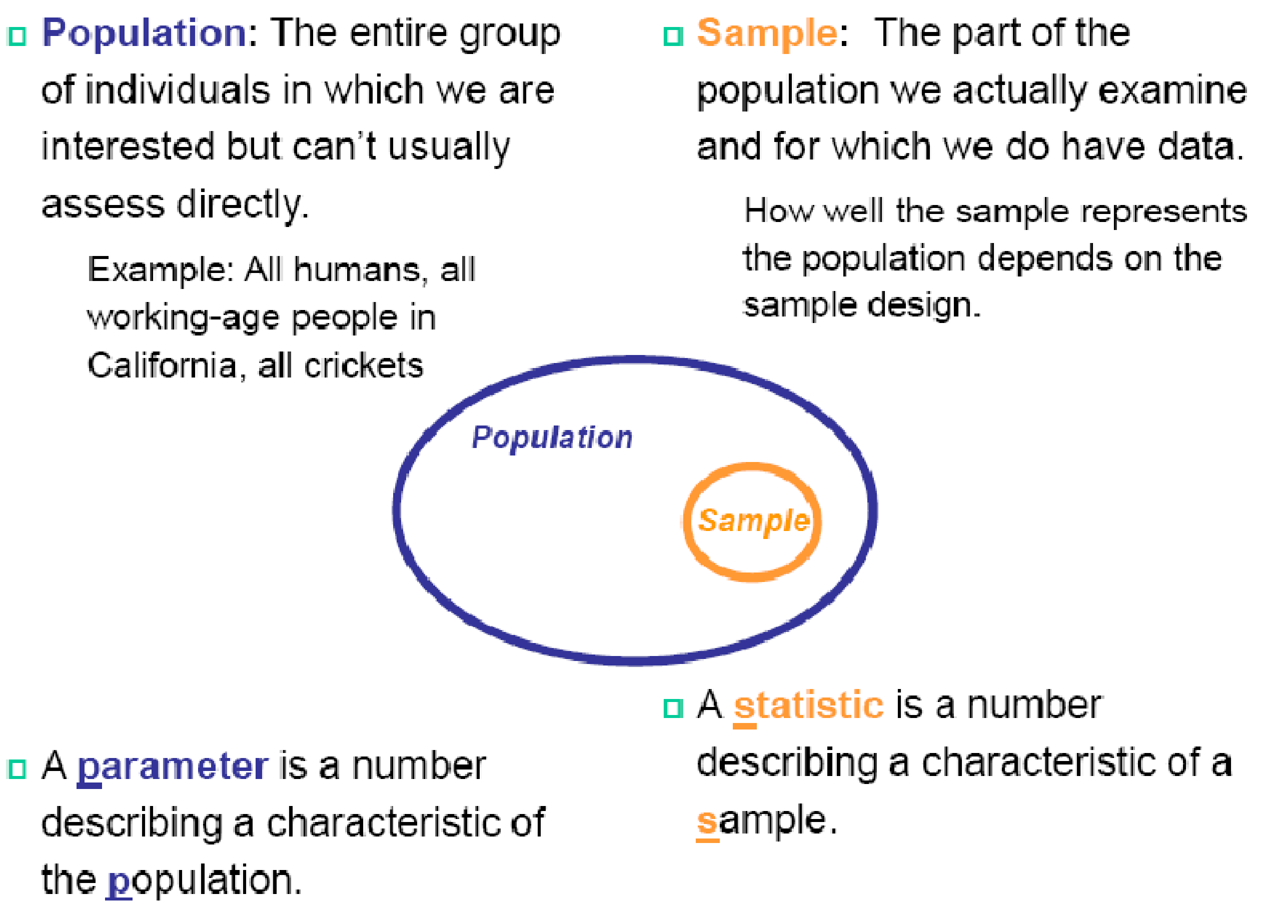
Behavioral and social science experiments



- Many behavioral experiments rely on hiding the true purpose of the study.
- Subjects would change their behavior if told in advance what investigators were looking for.
- The “Ethical Principals” of the American Psychological Association require consent unless a study merely observes behavior in a public space.

3 Sample versus population

3.1 Introduction



- **Population**: The entire group of individuals in which we are interested but can't usually assess directly.

  Example: All humans, all working-age people in California, all crickets

- **Sample**: The part of the population we actually examine and for which we do have data.

  How well the sample represents the population depends on the sample design.

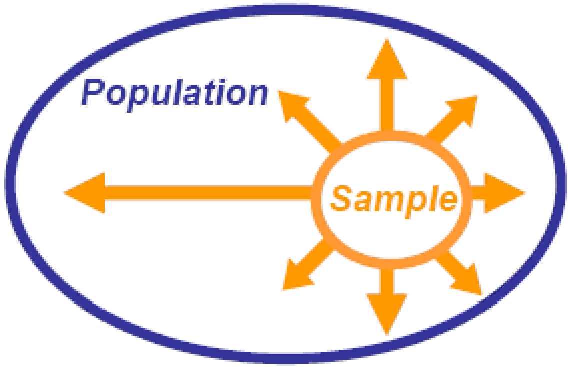
- A **parameter** is a number describing a characteristic of the population.

- A **statistic** is a number describing a characteristic of a sample.

Towards statistical inference

The techniques of inferential statistics allow us to draw inferences or conclusions about a population in a sample.

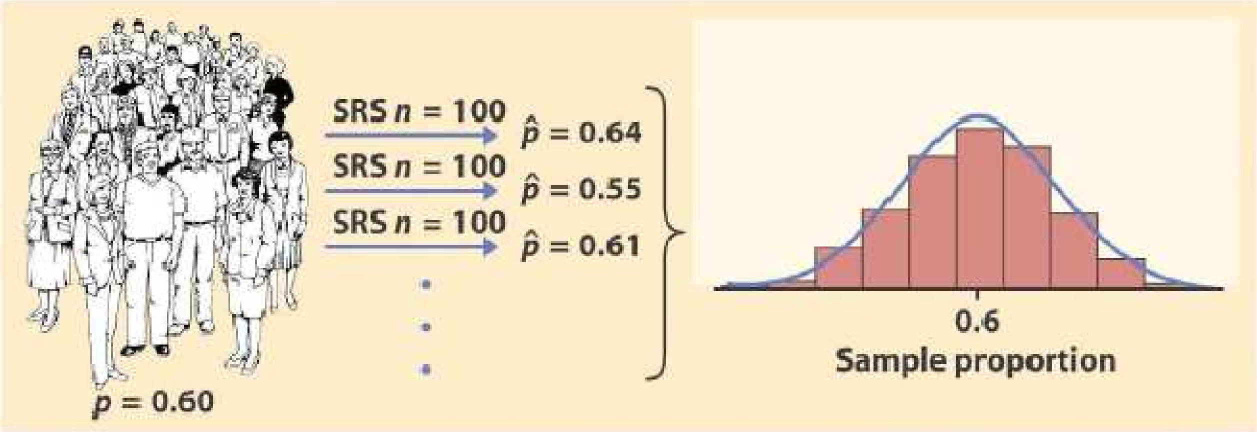
- Your estimate of the population is only as good as your sampling design.
  → Work hard to eliminate biases.
- Your sample is only an estimate—and if you randomly sampled again you would probably get a somewhat different result.
- The bigger the sample the better.



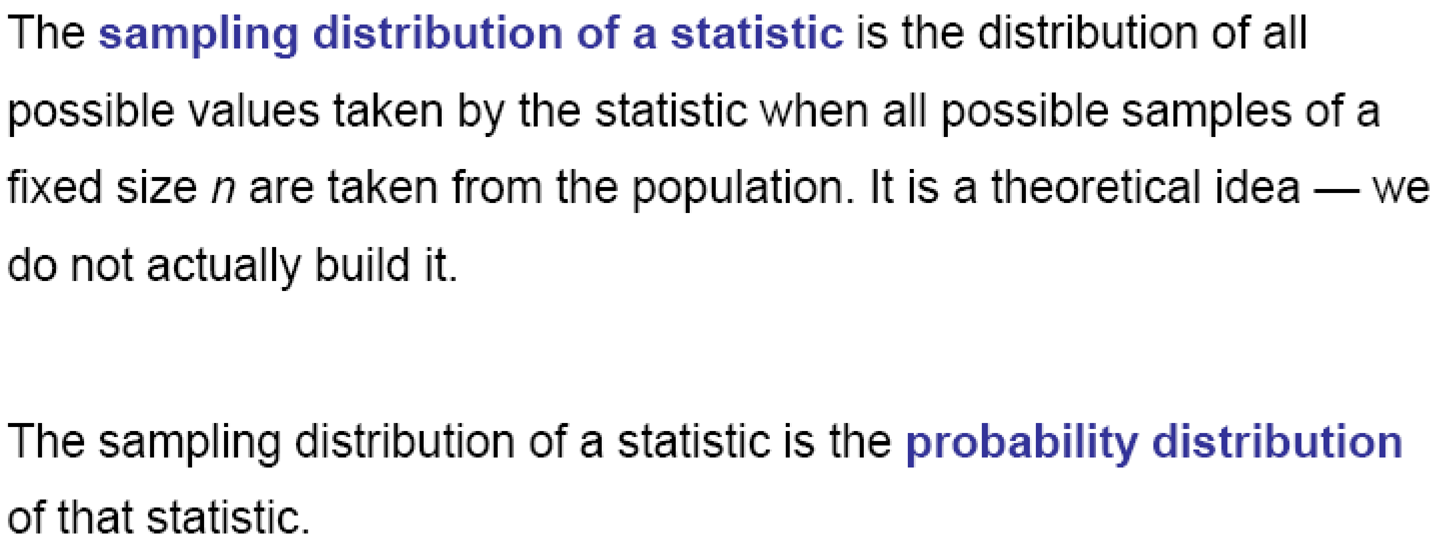
Sampling variability

Each time we take a random sample from a population, we are likely to get a different set of individuals and a calculate a different statistic. This is called **sampling variability.**

The good news is that, if we take lots of random samples of the same size from a given population, the variation from sample to sample—the **sampling distribution**—will follow a predictable pattern. All of statistical inference is based on this knowledge.



3.2 Distribution of a sample



The **sampling distribution of a statistic** is the distribution of all possible values taken by the statistic when all possible samples of a fixed size *n* are taken from the population. It is a theoretical idea — we do not actually build it.

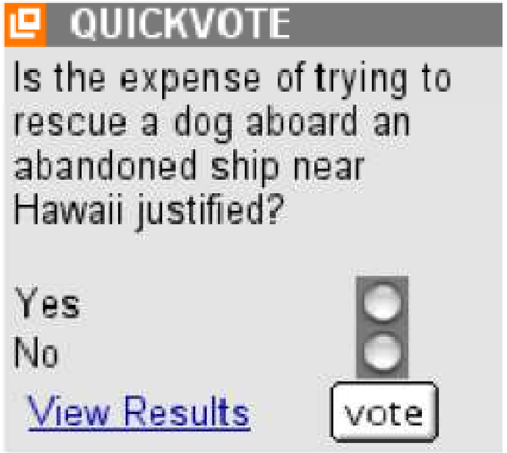
The sampling distribution of a statistic is the **probability distribution** of that statistic.

3.3 Statistics and sample moments

The **variability of a statistic** is described by the spread of its sampling distribution. This spread depends on the sampling design and the sample size *n*, with larger sample sizes leading to lower variability.

➔ Statistics from large samples are almost always close estimates of the true population parameter. However, this only applies to random samples.

Remember the “QuickVote” online surveys. They are worthless no matter how many people participate because they use a voluntary sampling design and not random sampling.



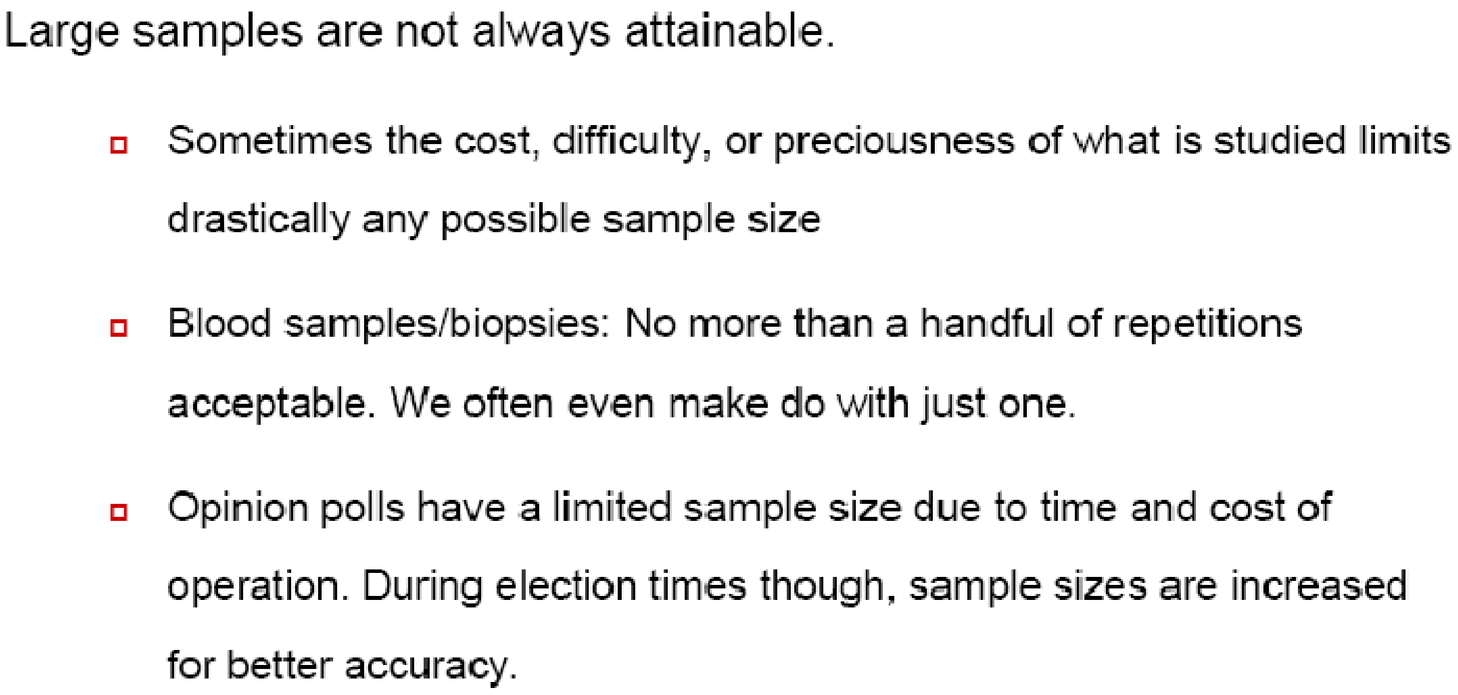
QUICKVOTE

Is the expense of trying to rescue a dog aboard an abandoned ship near Hawaii justified?

Yes ○

No ○

Practical note

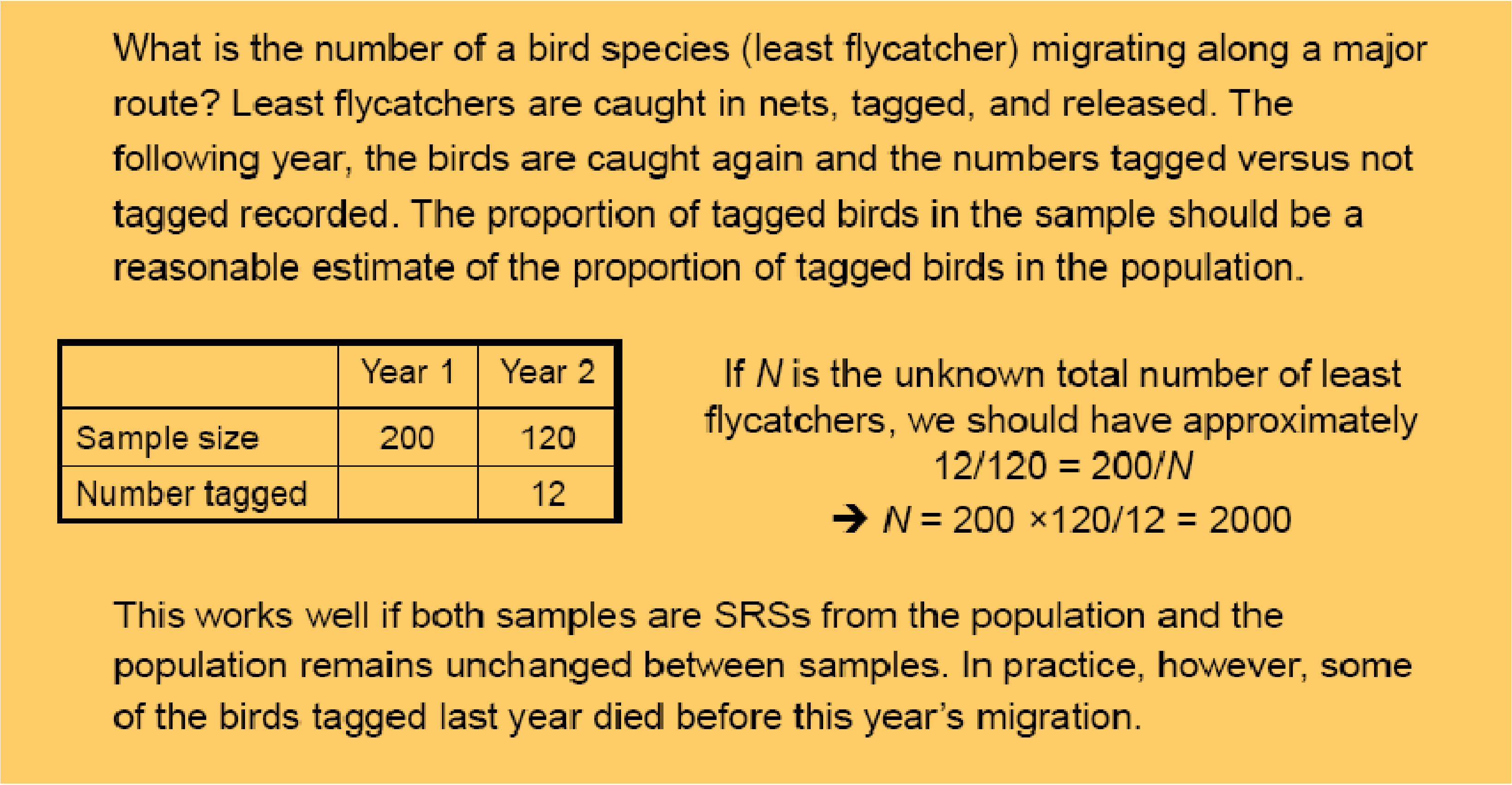


Large samples are not always attainable.

- Sometimes the cost, difficulty, or preciousness of what is studied limits drastically any possible sample size
- Blood samples/biopsies: No more than a handful of repetitions acceptable. We often even make do with just one.
- Opinion polls have a limited sample size due to time and cost of operation. During election times though, sample sizes are increased for better accuracy.

Capture-recapture sampling

Repeated sampling can be used to estimate the size N of a population (e.g., animals). Here is an example of **capture-recapture sampling**:



What is the number of a bird species (least flycatcher) migrating along a major route? Least flycatchers are caught in nets, tagged, and released. The following year, the birds are caught again and the numbers tagged versus not tagged recorded. The proportion of tagged birds in the sample should be a reasonable estimate of the proportion of tagged birds in the population.

	Year 1	Year 2
Sample size	200	120
Number tagged		12

If N is the unknown total number of least flycatchers, we should have approximately

$$12/120 = 200/N$$

➔ $N = 200 \times 120/12 = 2000$

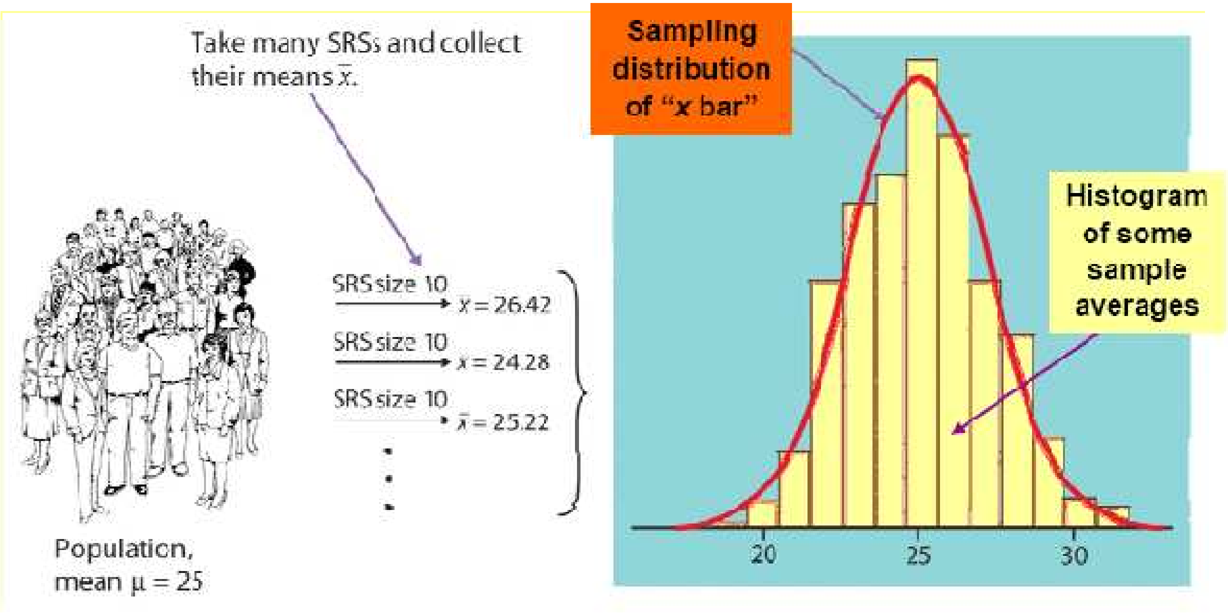
This works well if both samples are SRSs from the population and the population remains unchanged between samples. In practice, however, some of the birds tagged last year died before this year's migration.

4 Sample mean

4.1 Mean and variance

We take many random samples of a given size n from a population with mean μ and standard deviation σ.

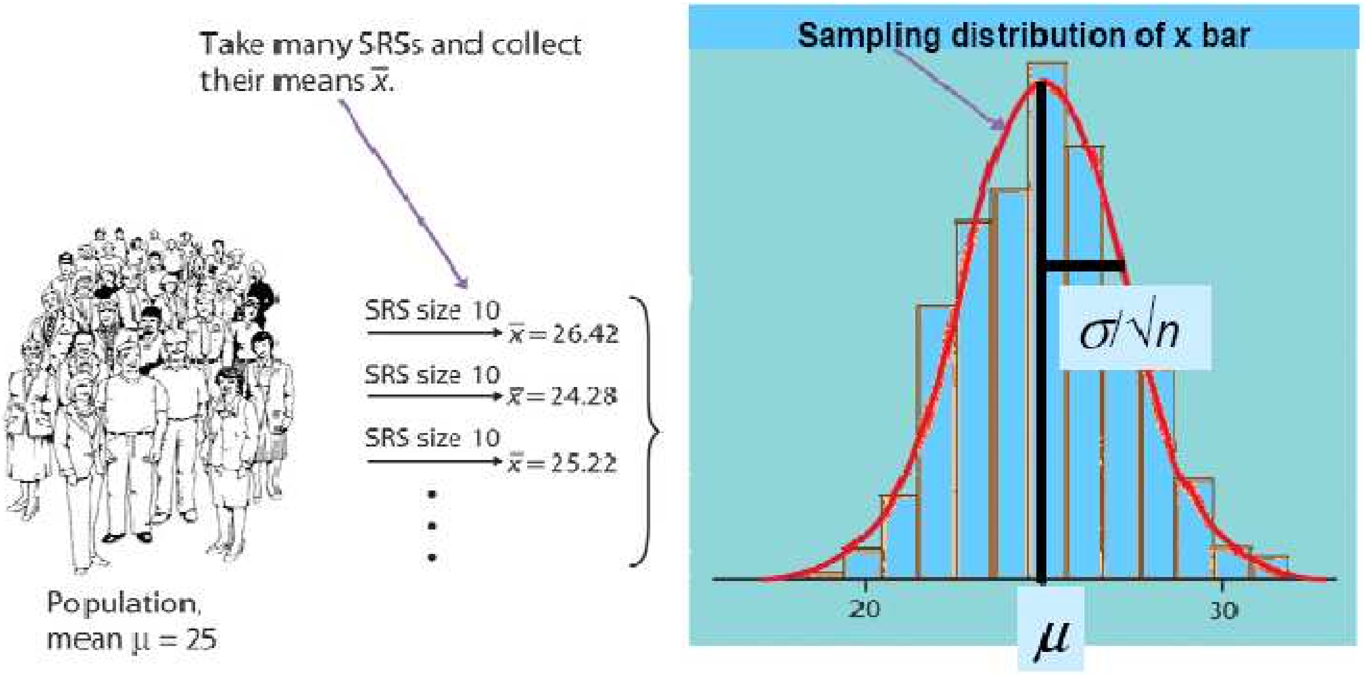
Some sample means will be above the population mean μ and some will be below, making up the sampling distribution.


Take many SRSs and collect their means $\bar{x}$.
SRS size 10
$\bar{x} = 26.42$
SRS size 10
$\bar{x} = 24.28$
SRS size 10
$\bar{x} = 25.22$
Population, mean $\mu = 25$
Sampling distribution of "x bar"
Histogram of some sample averages
20
25
30

For any population with mean μ and standard deviation σ:

□ The **mean**, or center of the sampling distribution of $\bar{x}$, is equal to the population mean μ : $\boldsymbol{\mu_x = \mu}$.

□ The **standard deviation** of the sampling distribution is $\sigma/\sqrt{n}$, where n is the sample size : $\boldsymbol{\sigma_x = \sigma/\sqrt{n}}$.



- Mean of a sampling distribution of $\overline{x}$

  There is no tendency for a sample mean to fall systematically above or below μ, even if the distribution of the raw data is skewed. Thus, the mean of the sampling distribution is an **unbiased estimate** of the population mean μ — it will be "correct on average" in many samples.

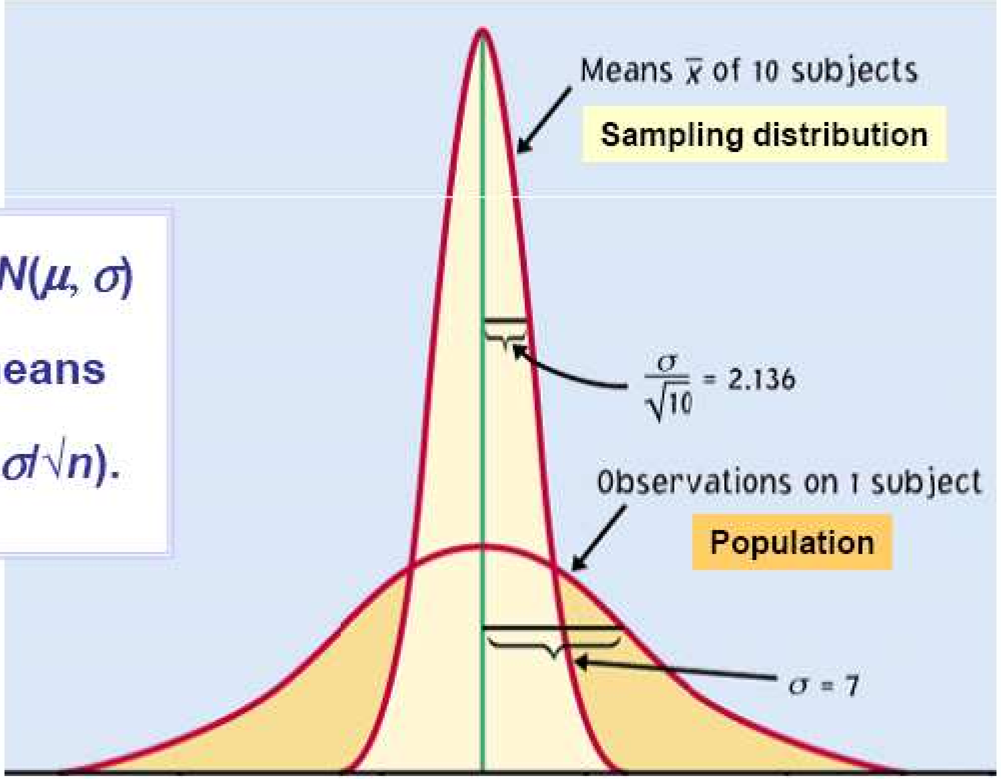
- Standard deviation of a sampling distribution of $\overline{x}$

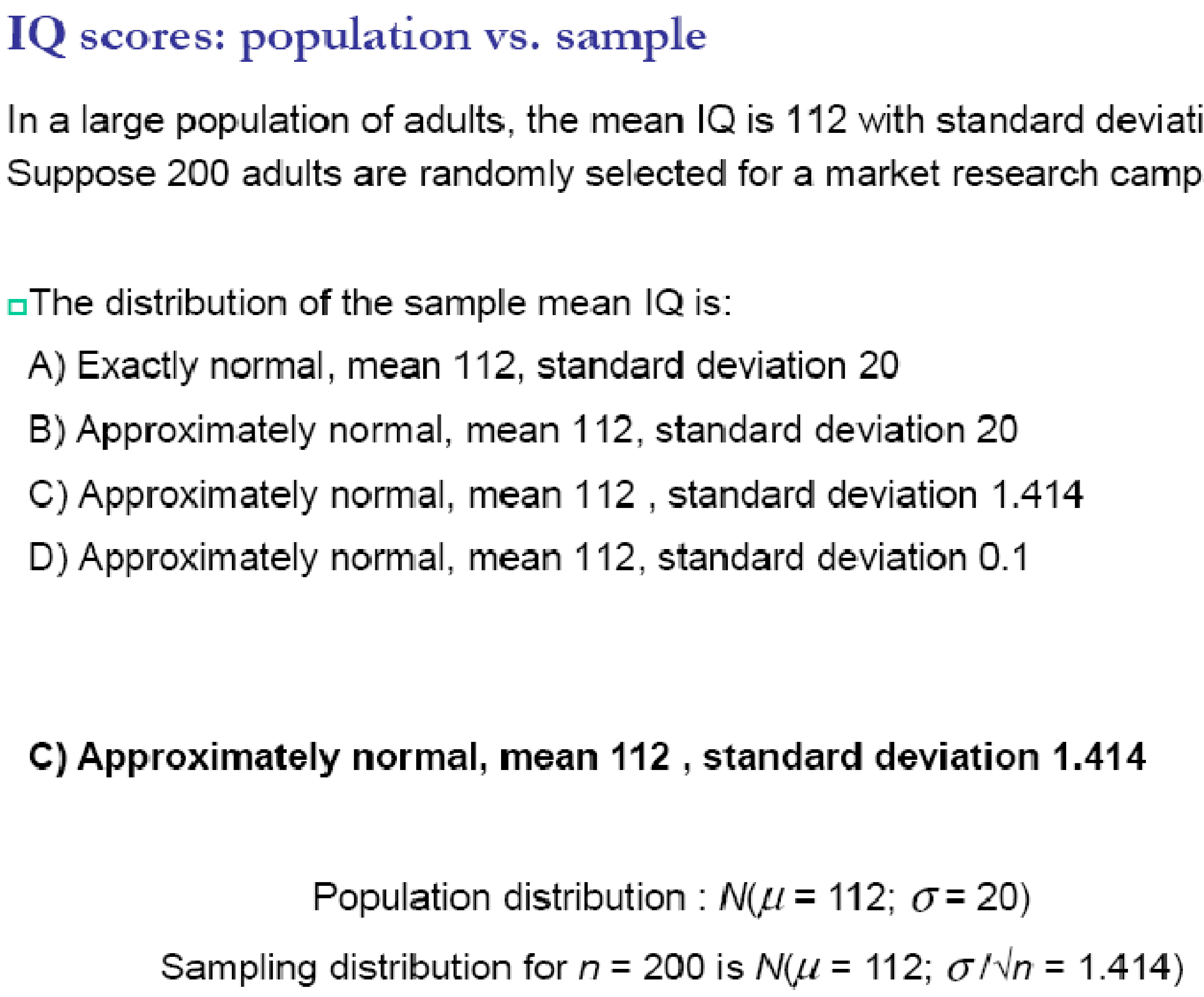
  The standard deviation of the sampling distribution measures how much the sample statistic varies from sample to sample. It is smaller than the standard deviation of the population by a factor of $\sqrt{n}$. ➔ **Averages are less variable than individual observations**.

For normally distributed populations

When a variable in a population is normally distributed, the sampling distribution of $\overline{x}$ for all possible samples of size *n* is also normally distributed.

If the population is *N*(μ, σ) then the sample means distribution is *N*(μ, σ/√*n*).





IQ scores: population vs. sample

In a large population of adults, the mean IQ is 112 with standard deviation 20. Suppose 200 adults are randomly selected for a market research campaign.

□The distribution of the sample mean IQ is:

A) Exactly normal, mean 112, standard deviation 20

B) Approximately normal, mean 112, standard deviation 20

C) Approximately normal, mean 112 , standard deviation 1.414

D) Approximately normal, mean 112, standard deviation 0.1

C) Approximately normal, mean 112 , standard deviation 1.414

Population distribution : $N(\mu = 112;\ \sigma = 20)$

Sampling distribution for n = 200 is $N(\mu = 112;\ \sigma/\sqrt{n} = 1.414)$

Application

Hypokalemia is diagnosed when blood potassium levels are below 3.5mEq/dl. Let's assume that we know a patient whose measured potassium levels vary daily according to a normal distribution $N(\mu = 3.8, \sigma = 0.2)$.

If only one measurement is made, what is the probability that this patient will be misdiagnosed with Hypokalemia?

$$z = \frac{(x-\mu)}{\sigma} = \frac{3.5-3.8}{0.2} \qquad z = -1.5,\ P(z < -1.5) = 0.0668 \approx 7\%$$

Instead, if measurements are taken on 4 separate days, what is the probability of a misdiagnosis?

$$z = \frac{(\bar{x}-\mu)}{\sigma/\sqrt{n}} = \frac{3.5-3.8}{0.2/\sqrt{4}} \qquad z = -3,\ P(z < -1.5) = 0.0013 \approx 0.1\%$$

Note: Make sure to standardize (z) using the standard deviation for the sampling distribution.

On a practical note

- Large samples are not always attainable.
  - Sometimes the cost, difficulty, or preciousness of what is studied drastically limits any possible sample size.
  - Blood samples/biopsies: No more than a handful of repetitions are acceptable. Oftentimes, we even make do with just one.
  - Opinion polls have a limited sample size due to time and cost of operation. During election times, though, sample sizes are increased for better accuracy.

- Not all variables are normally distributed.
  - Income, for example, is typically strongly skewed.
  - Is $\overline{x}$ still a good estimator of μ then?

4.2 Law of large numbers revisited

- Question: using only a finite number of values of X (a random sample of size n, say), can any reliable inference be made about E[X], “the average of an infinite number of values of X”?
- Answer: YES!

- A positive integer n can be determined such that if a random sample of size n or larger is taken from a population with density f(.) (with E[X])=μ), the probability can be made as close to 1 as desired that the sample mean X bar will deviate from μ by less than an arbitrarily specified small quantity:

$$\forall \epsilon > 0, 0 < \delta < 1, \text{ there exists a value } n \text{ such that } \forall m \leq n$$

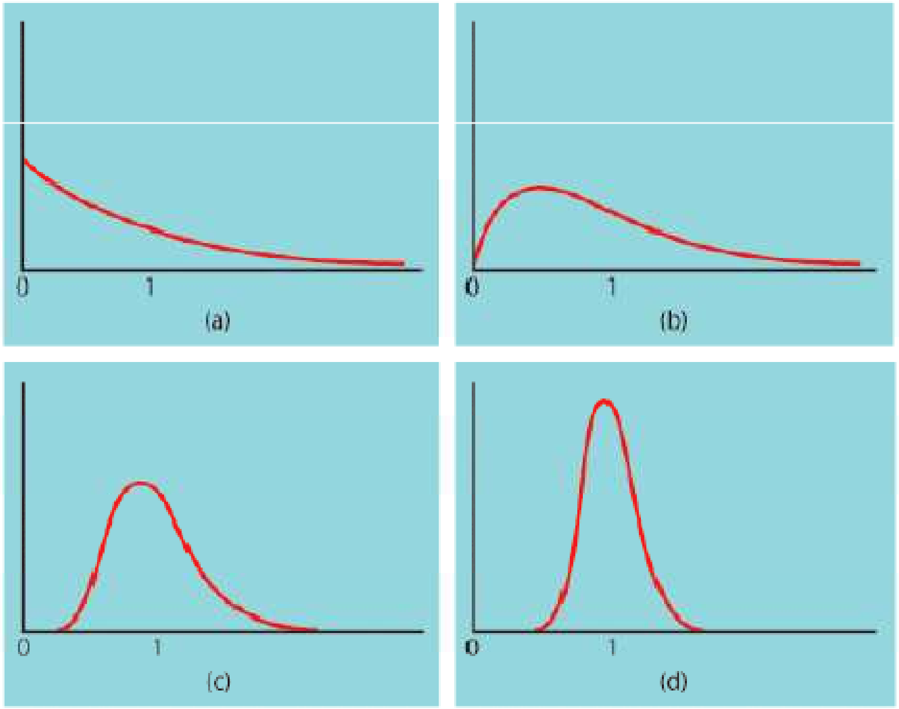
$$P[-\epsilon < \overline{X}_m < \epsilon] \leq 1 - \delta$$

(cfr. weak law of large numbers)

- Proof: Use Chebyshev inequality

4.3 Central-limit theory revisited

Central Limit Theorem: When randomly sampling from **any** population with mean μ and standard deviation σ, **when n is large enough**, the sampling distribution of $\overline{x}$ is approximately normal: ~ $N(\mu, \sigma/\sqrt{n})$.



Population with strongly skewed distribution

Sampling distribution of $\overline{x}$ for n = 2 observations

Sampling distribution of $\overline{x}$ for n = 10 observations

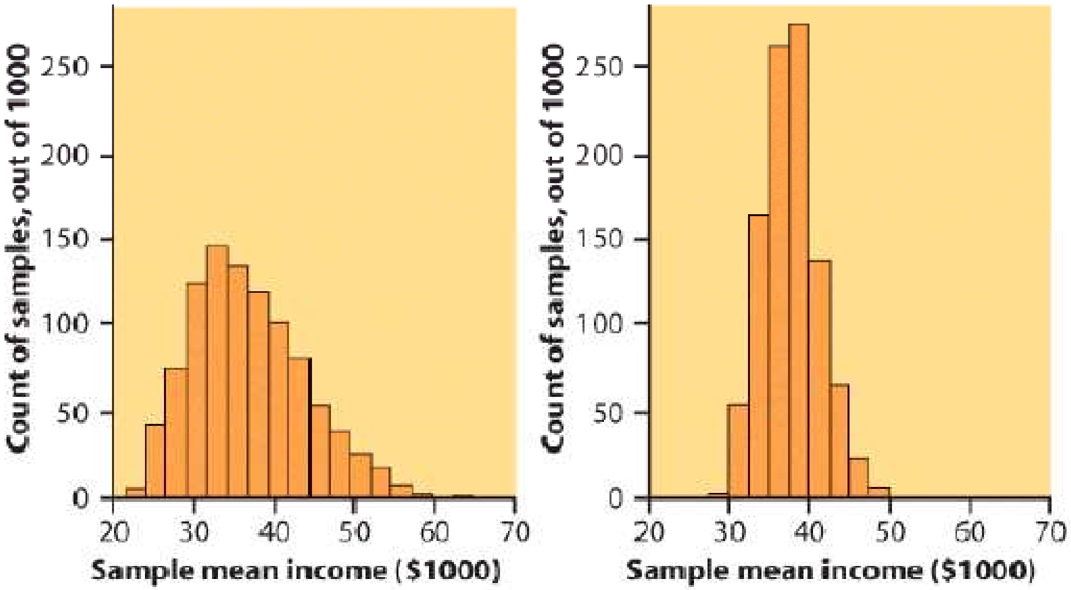
Sampling distribution of $\overline{x}$ for n = 25 observations

Income distribution

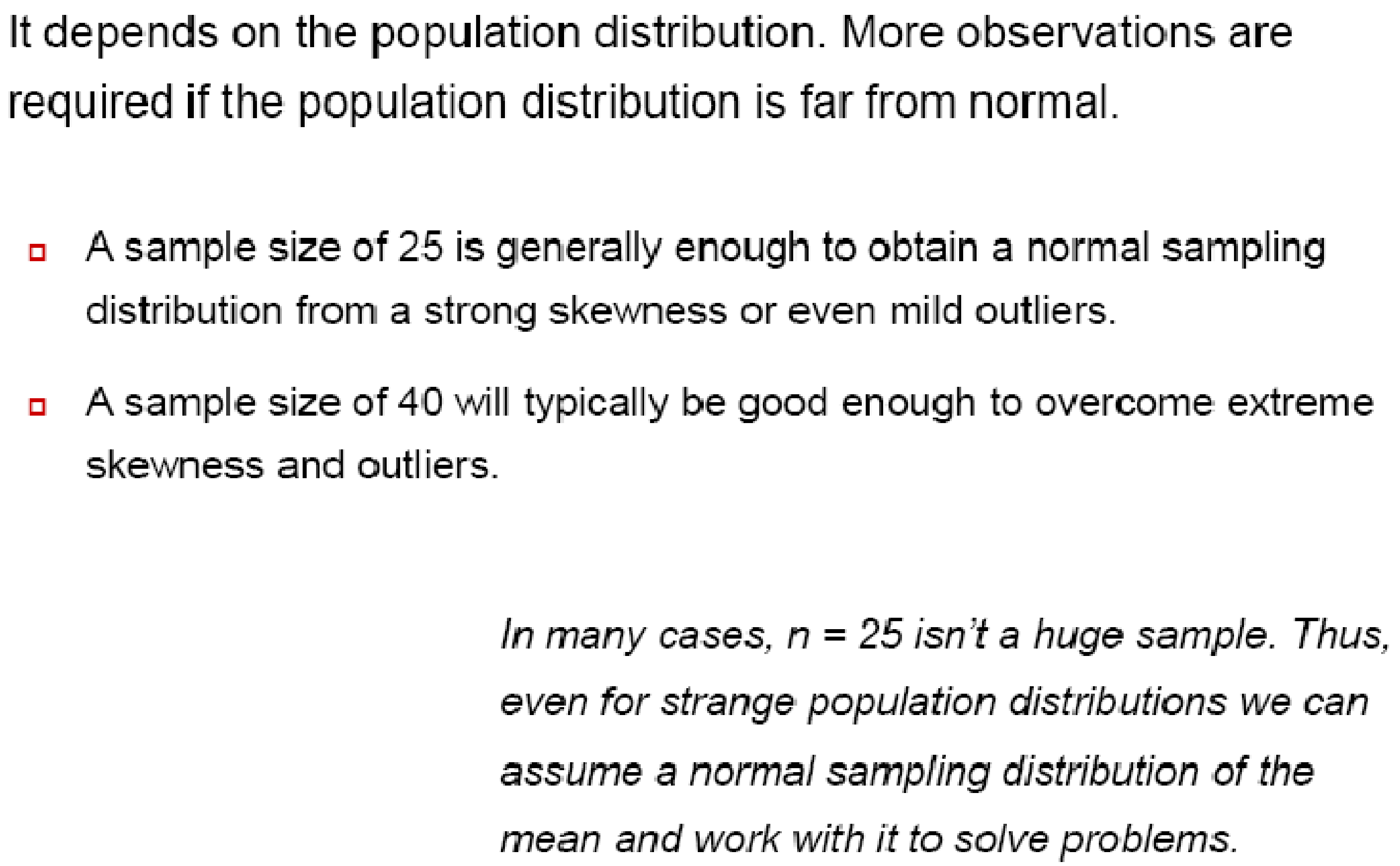
Let's consider the very large database of individual incomes from the Bureau of Labor Statistics as our population. It is strongly right skewed.

- We take 1000 SRSs of 100 incomes, calculate the sample mean for each, and make a histogram of these 1000 means.
- We also take 1000 SRSs of 25 incomes, calculate the sample mean for each, and make a histogram of these 1000 means.

Which histogram corresponds to samples of size 100? 25?



How large a sample size ?



It depends on the population distribution. More observations are required if the population distribution is far from normal.

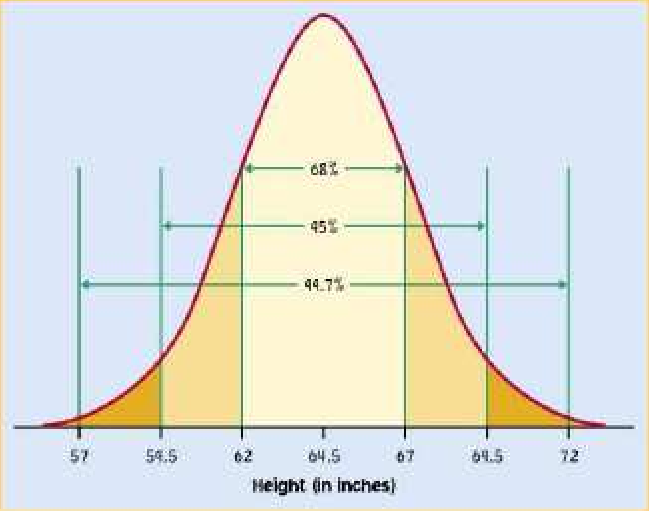
- A sample size of 25 is generally enough to obtain a normal sampling distribution from a strong skewness or even mild outliers.
- A sample size of 40 will typically be good enough to overcome extreme skewness and outliers.

In many cases, n = 25 isn't a huge sample. Thus, even for strange population distributions we can assume a normal sampling distribution of the mean and work with it to solve problems.

Any linear combination of independent random variables is also normally distributed.

More generally, the central limit theorem is valid as long as we are sampling many small random events, even if the events have different distributions (as long as no one random event dominates the others).

Why is this cool? It explains why the normal distribution is so common.

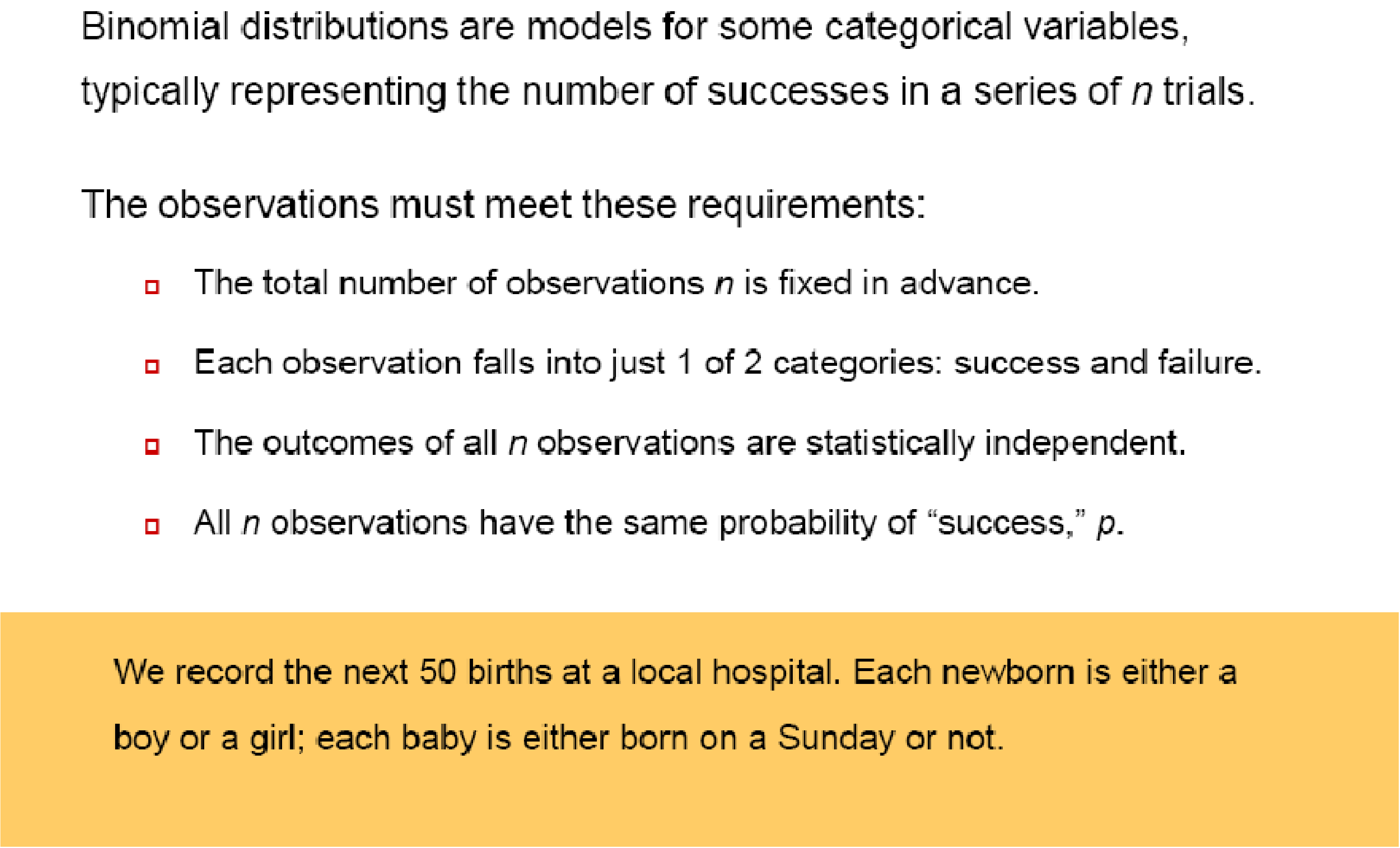


Example: Height seems to be determined by a large number of genetic and environmental factors, like nutrition. The “individuals” are genes and environmental factors. Your height is a mean.

4.4 Bernoulli and Poisson distribution

See before

Binomial distributions for sample counts



Binomial distributions are models for some categorical variables, typically representing the number of successes in a series of *n* trials.

The observations must meet these requirements:

- The total number of observations *n* is fixed in advance.
- Each observation falls into just 1 of 2 categories: success and failure.
- The outcomes of all *n* observations are statistically independent.
- All *n* observations have the same probability of "success," *p*.

We record the next 50 births at a local hospital. Each newborn is either a boy or a girl; each baby is either born on a Sunday or not.

We express a binomial distribution for the count X of successes among n observations as a function of the parameters n and p: $B(n,p)$.

- The parameter n is the total number of observations.
- The parameter p is the probability of success on each observation.
- The count of successes X can be any whole number between 0 and n.

A coin is flipped 10 times. Each outcome is either a head or a tail. The variable X is the number of heads among those 10 flips, our count of "successes."

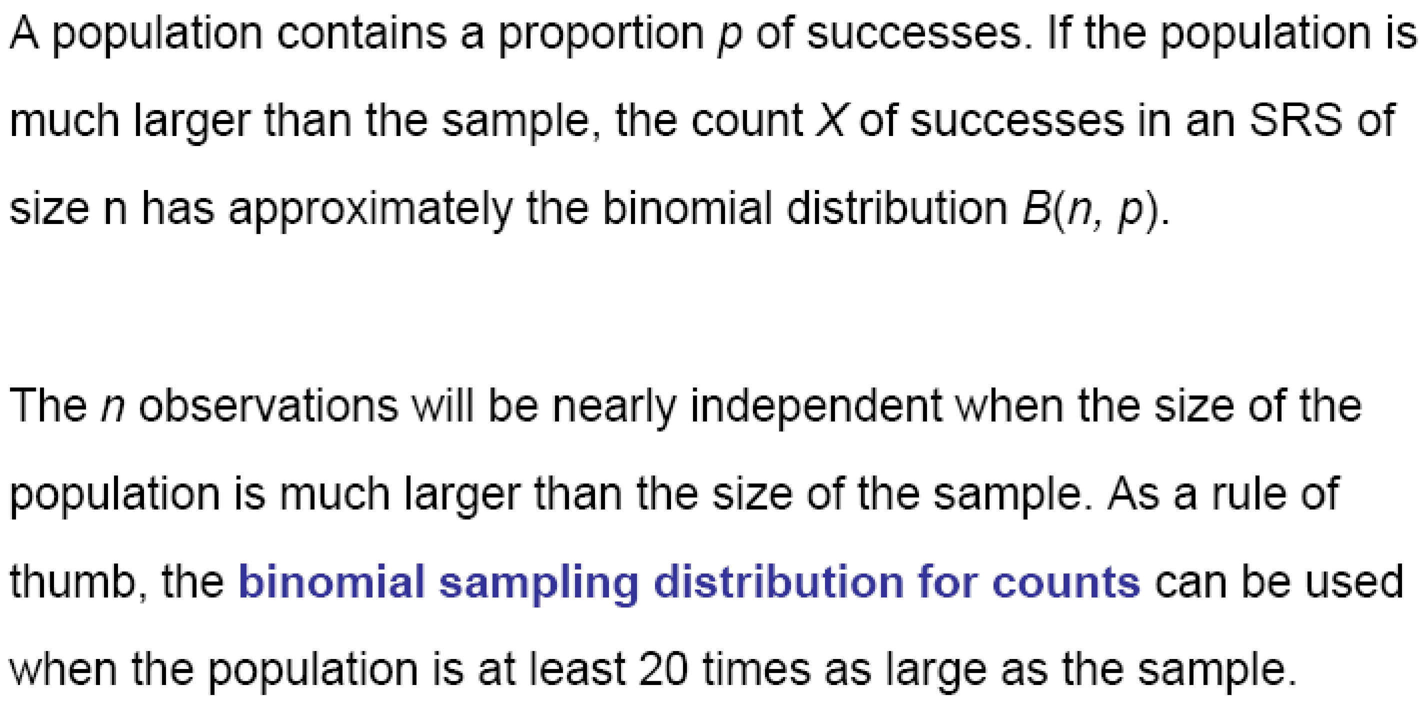
On each flip, the probability of success, "head," is 0.5. The number X of heads among 10 flips has the binomial distribution $B(n = 10, p = 0.5)$.

Imagine that coins are spread out so that half of them are heads up, and half tails up. Close your eyes and pick one. The probability that this coin is heads up is 0.5.

However, if you don't put the coin back in the pile, the probability of picking up another coin and having it be heads up is now less than 0.5. The successive observations are not independent.

Likewise, choosing a simple random sample (SRS) from any population is not quite a binomial setting. However, when the population is large, removing a few items has a very small effect on the composition of the remaining population: successive observations are very nearly independent.

Binomial distribution in statistical sampling



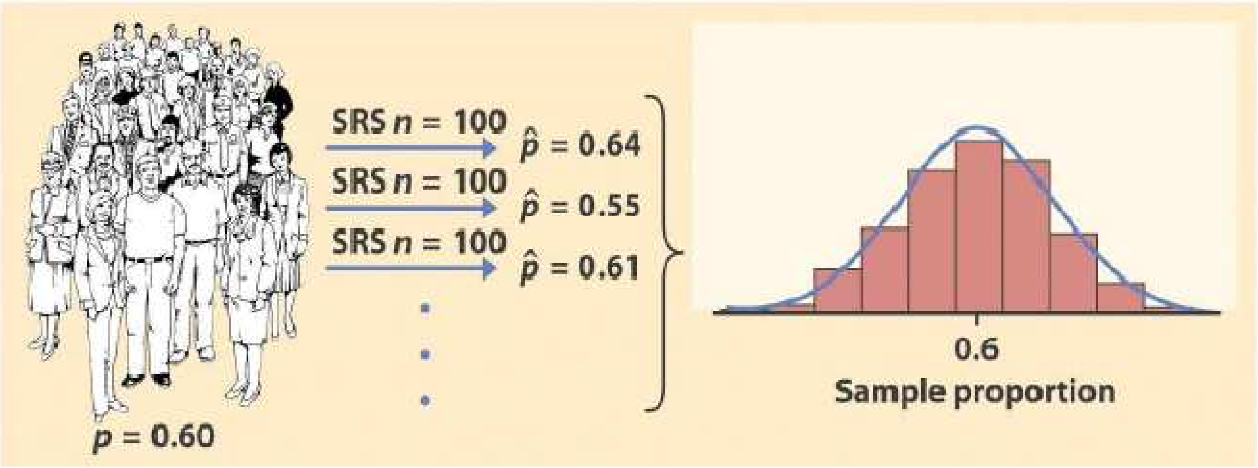
A population contains a proportion *p* of successes. If the population is much larger than the sample, the count *X* of successes in an SRS of size n has approximately the binomial distribution *B*(*n*, *p*).

The *n* observations will be nearly independent when the size of the population is much larger than the size of the sample. As a rule of thumb, the **binomial sampling distribution for counts** can be used when the population is at least 20 times as large as the sample.

Reminder: sampling variability

Each time we take a random sample from a population, we are likely to get a different set of individuals and calculate a different statistic. This is called sampling variability.

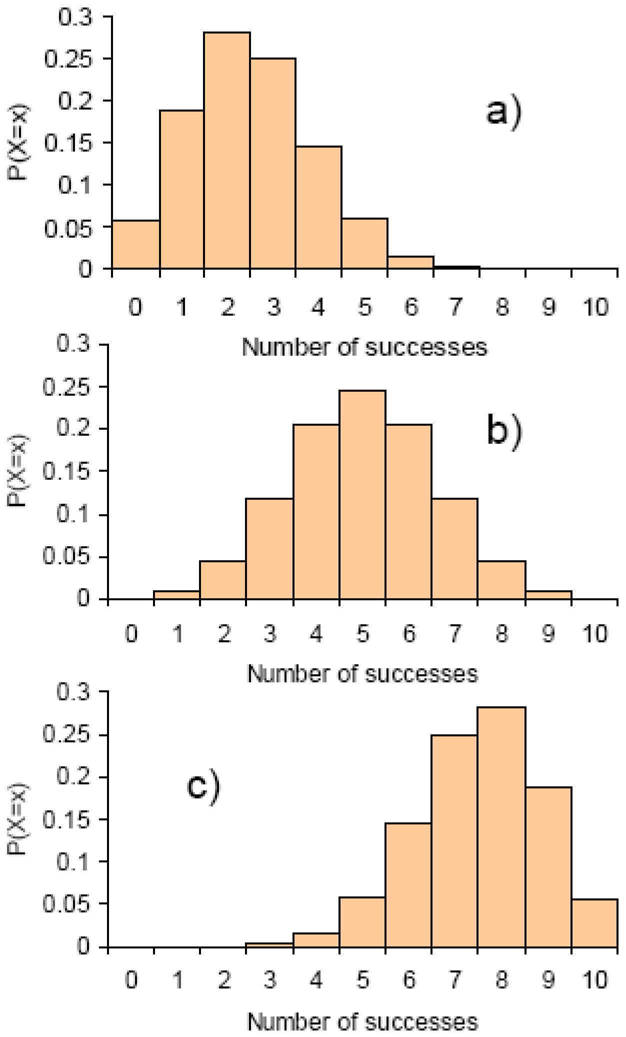
If we take a lot of random samples of the same size from a given population, the variation from sample to sample—the **sampling distribution**—will follow a predictable pattern.



Binomial mean and standard deviation

The center and spread of the binomial distribution for a count *X* are defined by the mean μ and standard deviation σ:

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



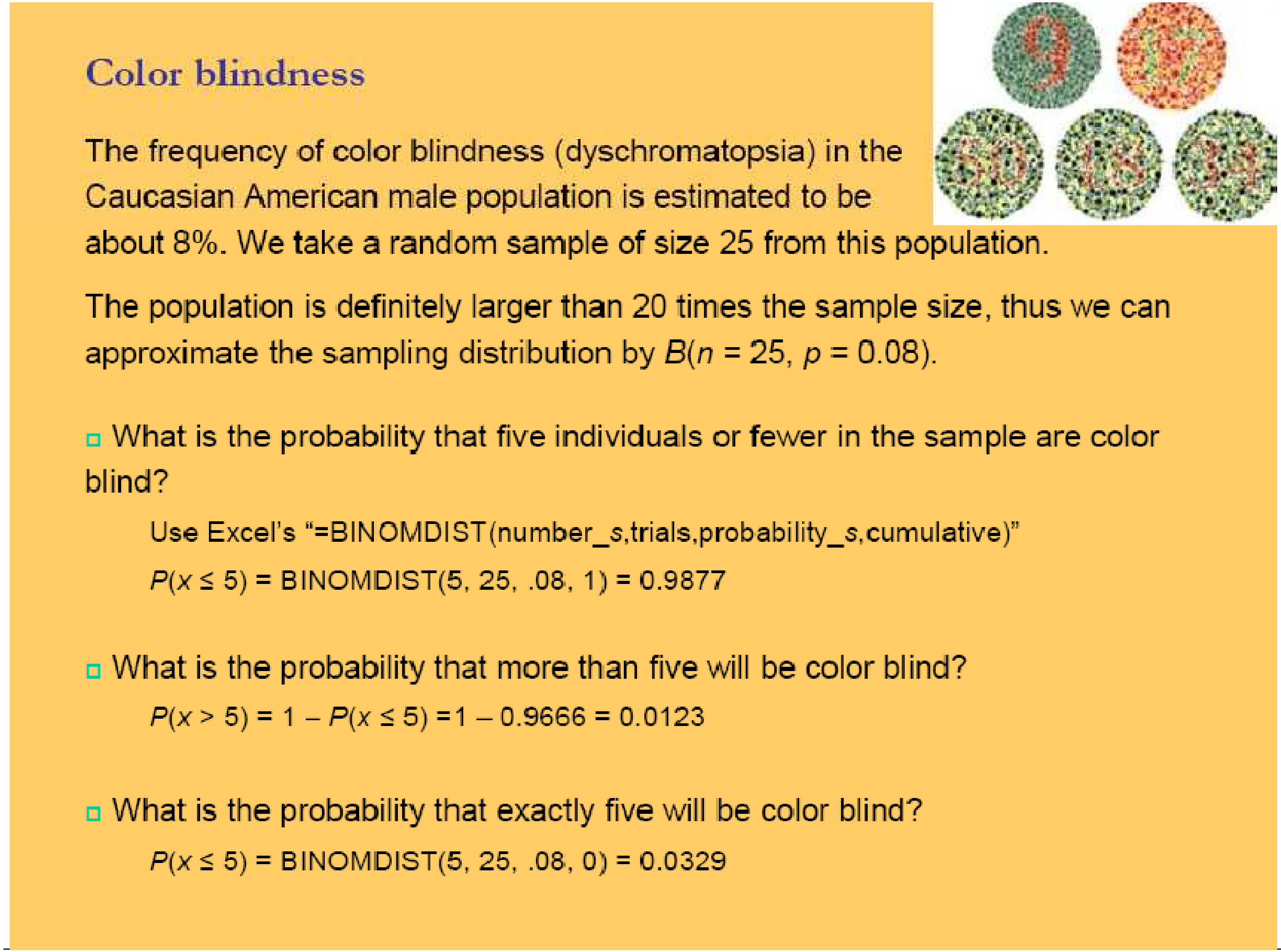
Effect of changing *p* when *n* is fixed.

a) *n* = 10, *p* = 0.25

b) *n* = 10, *p* = 0.5

c) *n* = 10, *p* = 0.75

For small samples, binomial distributions are skewed when *p* is different from 0.5.



Color blindness

The frequency of color blindness (dyschromatopsia) in the Caucasian American male population is estimated to be about 8%. We take a random sample of size 25 from this population.

The population is definitely larger than 20 times the sample size, thus we can approximate the sampling distribution by $B(n = 25, p = 0.08)$.

- What is the probability that five individuals or fewer in the sample are color blind?

  Use Excel's "=BINOMDIST(number_*s*,trials,probability_*s*,cumulative)"

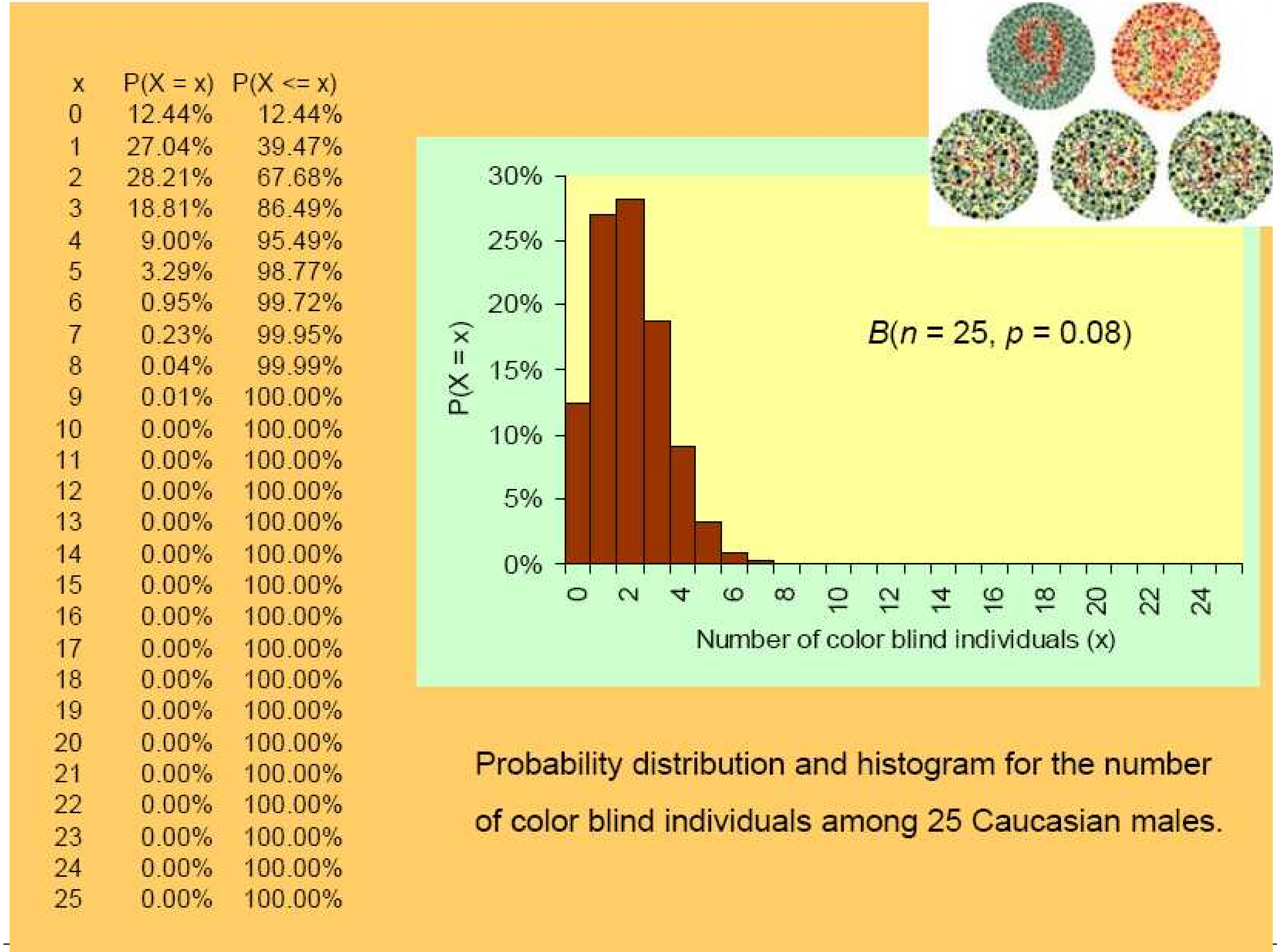
  $P(x \leq 5)$ = BINOMDIST(5, 25, .08, 1) = 0.9877

- What is the probability that more than five will be color blind?

  $P(x > 5) = 1 - P(x \leq 5) = 1 - 0.9666 = 0.0123$

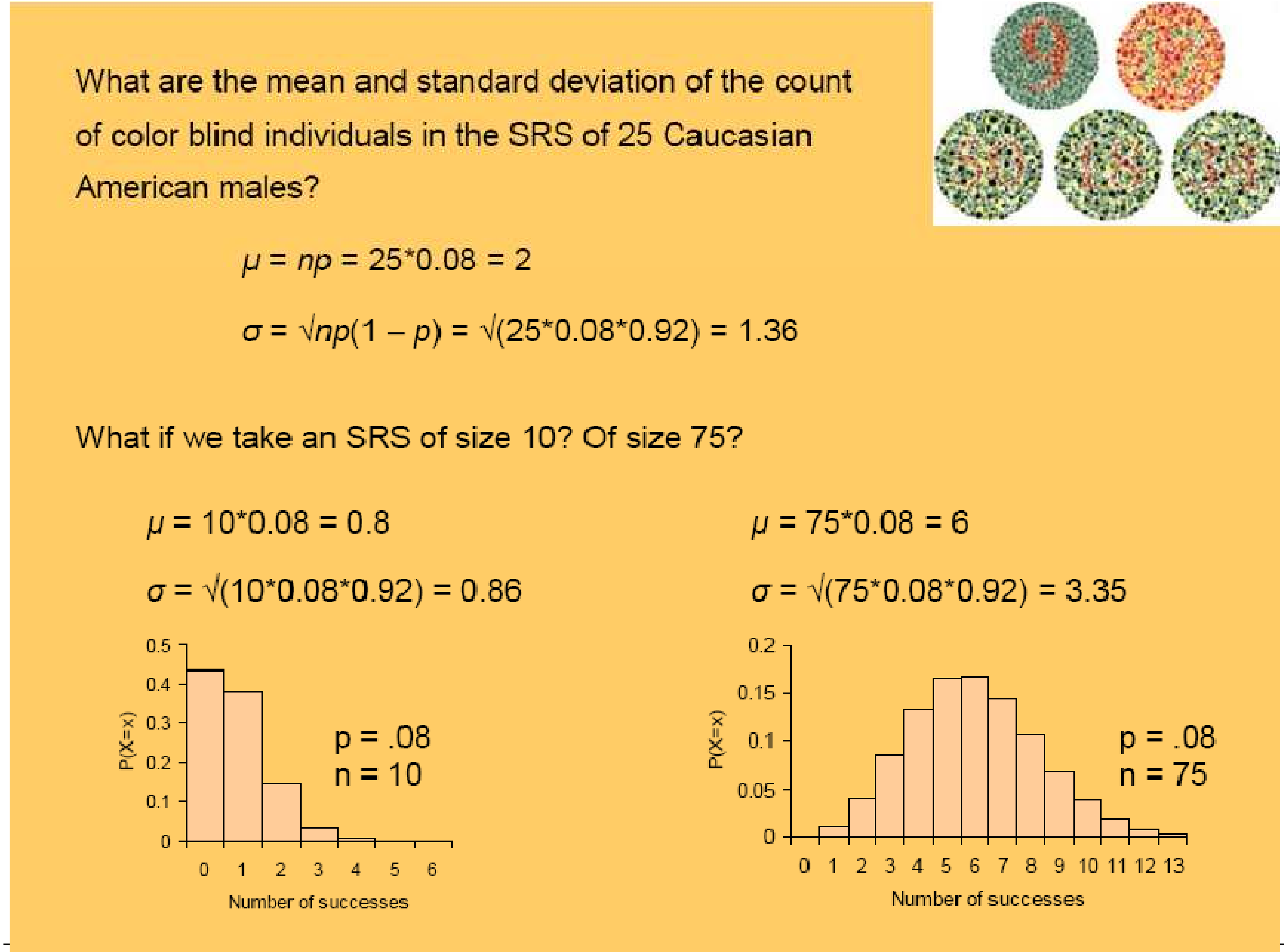
- What is the probability that exactly five will be color blind?

  $P(x \leq 5)$ = BINOMDIST(5, 25, .08, 0) = 0.0329



x	P(X = x)	P(X <= x)
0	12.44%	12.44%
1	27.04%	39.47%
2	28.21%	67.68%
3	18.81%	86.49%
4	9.00%	95.49%
5	3.29%	98.77%
6	0.95%	99.72%
7	0.23%	99.95%
8	0.04%	99.99%
9	0.01%	100.00%
10	0.00%	100.00%
11	0.00%	100.00%
12	0.00%	100.00%
13	0.00%	100.00%
14	0.00%	100.00%
15	0.00%	100.00%
16	0.00%	100.00%
17	0.00%	100.00%
18	0.00%	100.00%
19	0.00%	100.00%
20	0.00%	100.00%
21	0.00%	100.00%
22	0.00%	100.00%
23	0.00%	100.00%
24	0.00%	100.00%
25	0.00%	100.00%

Probability distribution and histogram for the number of color blind individuals among 25 Caucasian males.



What are the mean and standard deviation of the count of color blind individuals in the SRS of 25 Caucasian American males?

$\mu = np = 25*0.08 = 2$

$\sigma = \sqrt{np(1-p)} = \sqrt{(25*0.08*0.92)} = 1.36$

What if we take an SRS of size 10? Of size 75?

$\mu = 10*0.08 = 0.8$

$\sigma = \sqrt{(10*0.08*0.92)} = 0.86$

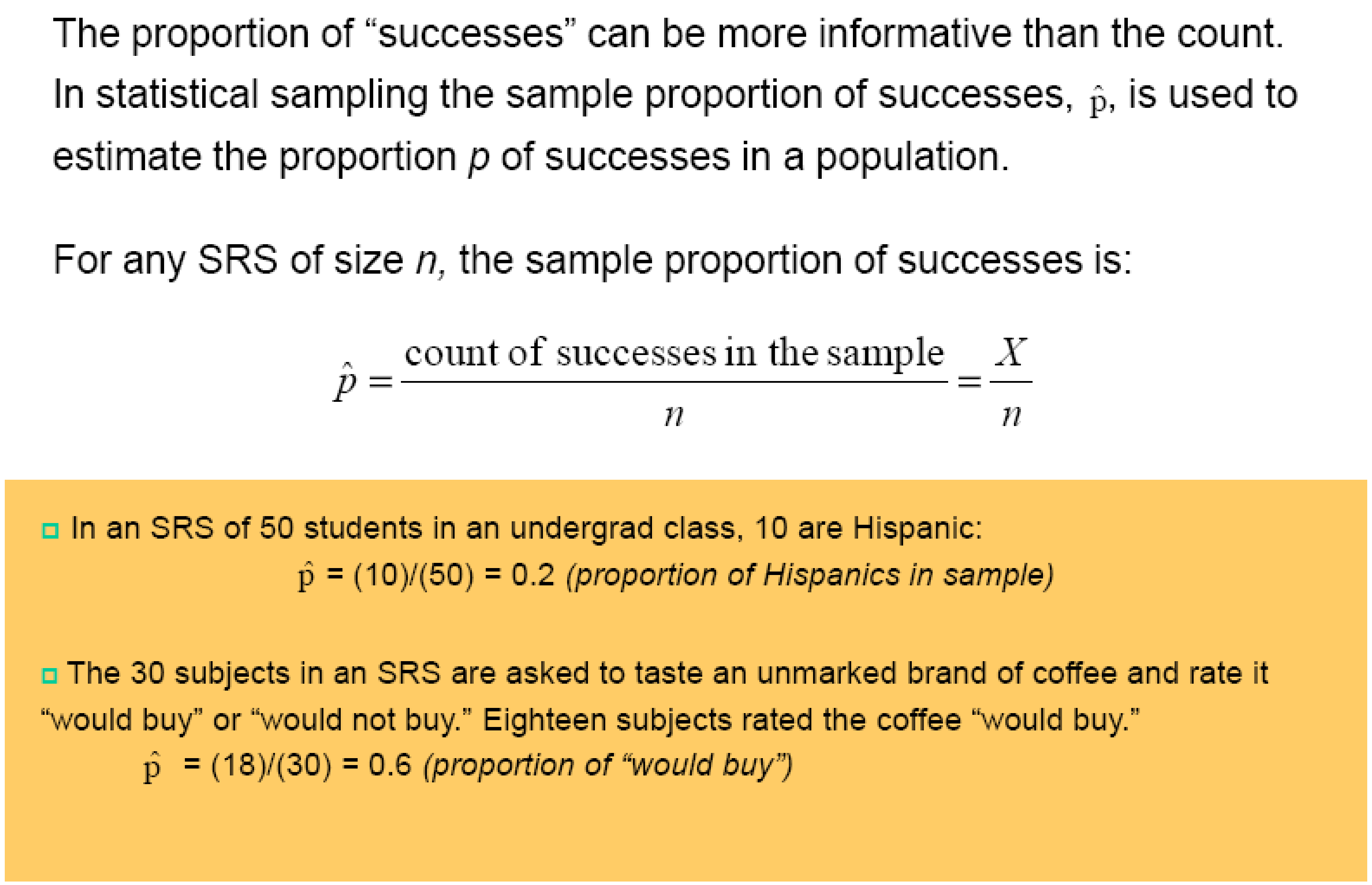
$\mu = 75*0.08 = 6$

$\sigma = \sqrt{(75*0.08*0.92)} = 3.35$

p = .08
n = 10

p = .08
n = 75

Sample proportions



The proportion of "successes" can be more informative than the count. In statistical sampling the sample proportion of successes, $\hat{p}$, is used to estimate the proportion *p* of successes in a population.

For any SRS of size *n*, the sample proportion of successes is:

$$\hat{p} = \frac{\text{count of successes in the sample}}{n} = \frac{X}{n}$$

□ In an SRS of 50 students in an undergrad class, 10 are Hispanic:

$\hat{p}$ = (10)/(50) = 0.2 *(proportion of Hispanics in sample)*

□ The 30 subjects in an SRS are asked to taste an unmarked brand of coffee and rate it "would buy" or "would not buy." Eighteen subjects rated the coffee "would buy."

$\hat{p}$ = (18)/(30) = 0.6 *(proportion of "would buy")*

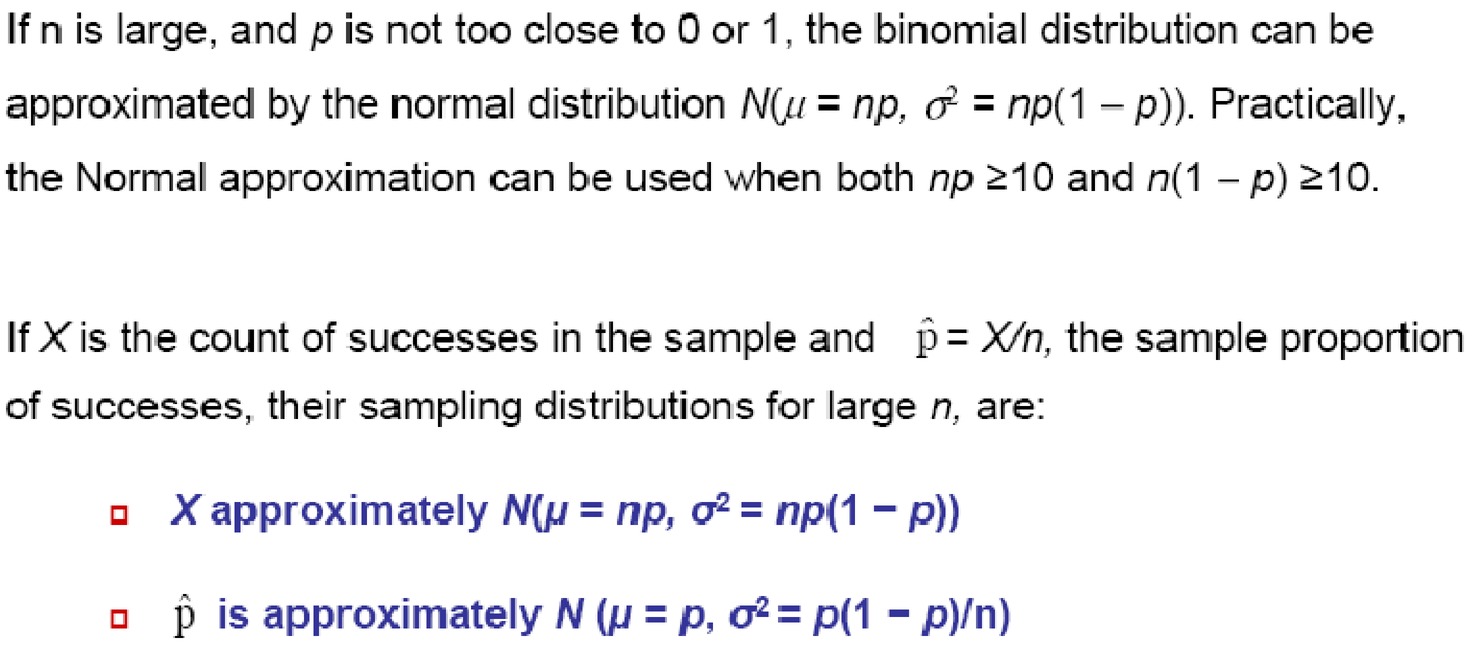
If the sample size is much smaller than the size of a population with proportion *p* of successes, then the mean and standard deviation of $\hat{p}$ are:

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

□ Because the mean is *p,* we say that the sample proportion in an SRS is an **unbiased estimator** of the population proportion *p*.

□ The variability decreases as the sample size increases. So larger samples usually give closer estimates of the population proportion *p*.

Normal approximation



If n is large, and p is not too close to 0 or 1, the binomial distribution can be approximated by the normal distribution $N(\mu = np,\ \sigma^2 = np(1 - p))$. Practically, the Normal approximation can be used when both $np \geq 10$ and $n(1 - p) \geq 10$.

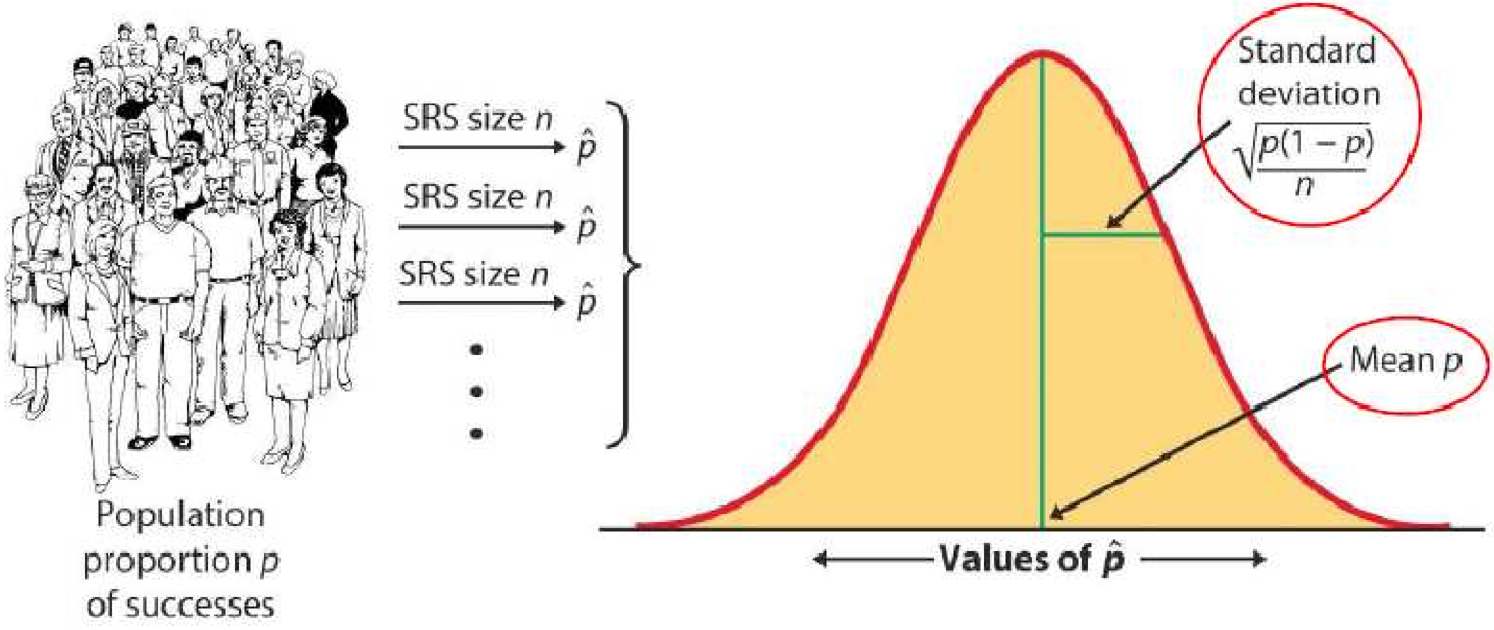
If X is the count of successes in the sample and $\hat{p} = X/n$, the sample proportion of successes, their sampling distributions for large n, are:

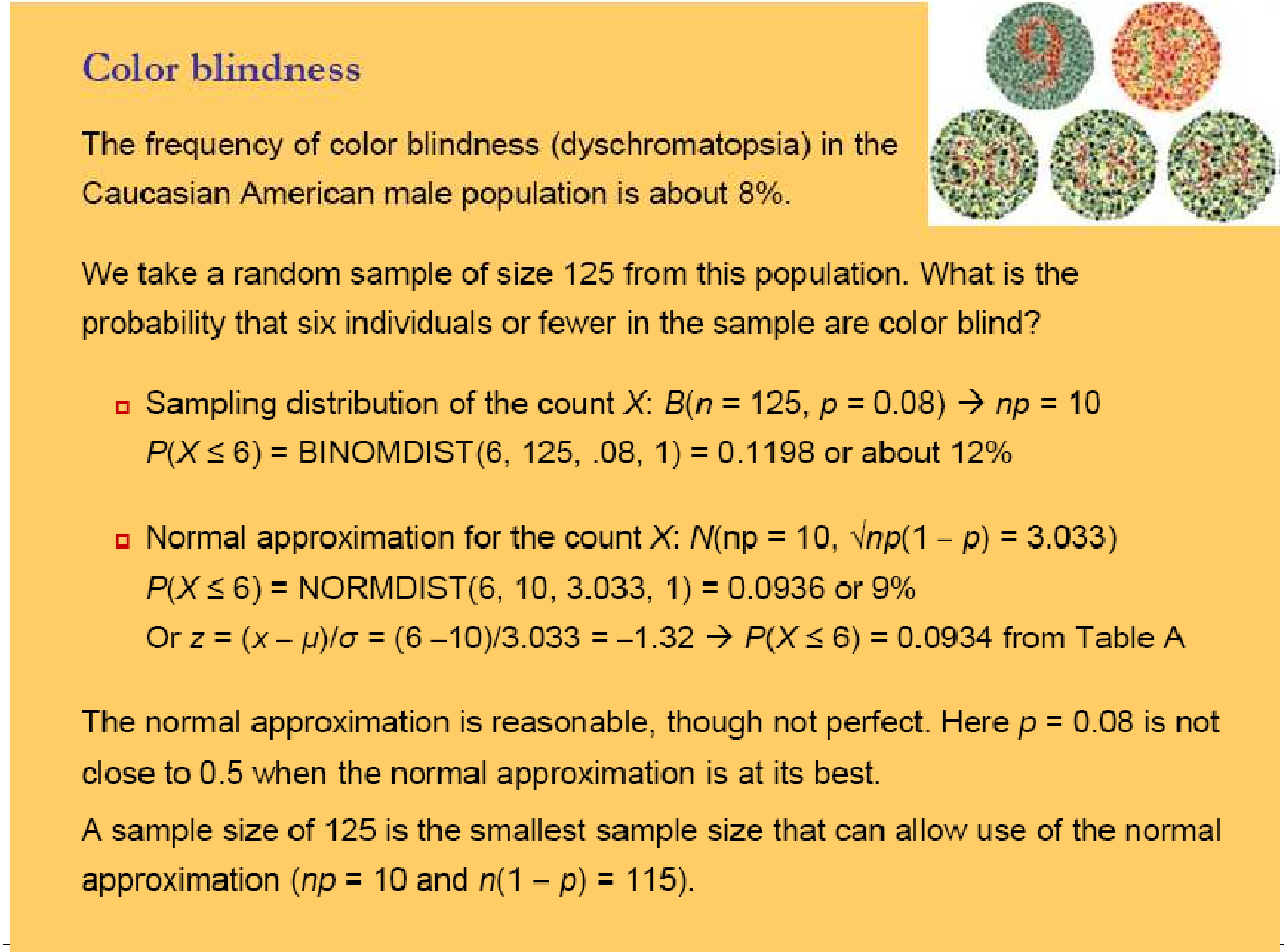
- X approximately $N(\mu = np,\ \sigma^2 = np(1 - p))$
- $\hat{p}$ is approximately $N\ (\mu = p,\ \sigma^2 = p(1 - p)/\text{n})$

Sampling distribution of the sampling proportion

The sampling distribution of $\hat{p}$ is never exactly normal. But as the sample size increases, the sampling distribution of $\hat{p}$ becomes approximately normal.

The normal approximation is most accurate for any fixed *n* when *p* is close to 0.5, and least accurate when *p* is near 0 or near 1.





Color blindness

The frequency of color blindness (dyschromatopsia) in the Caucasian American male population is about 8%.

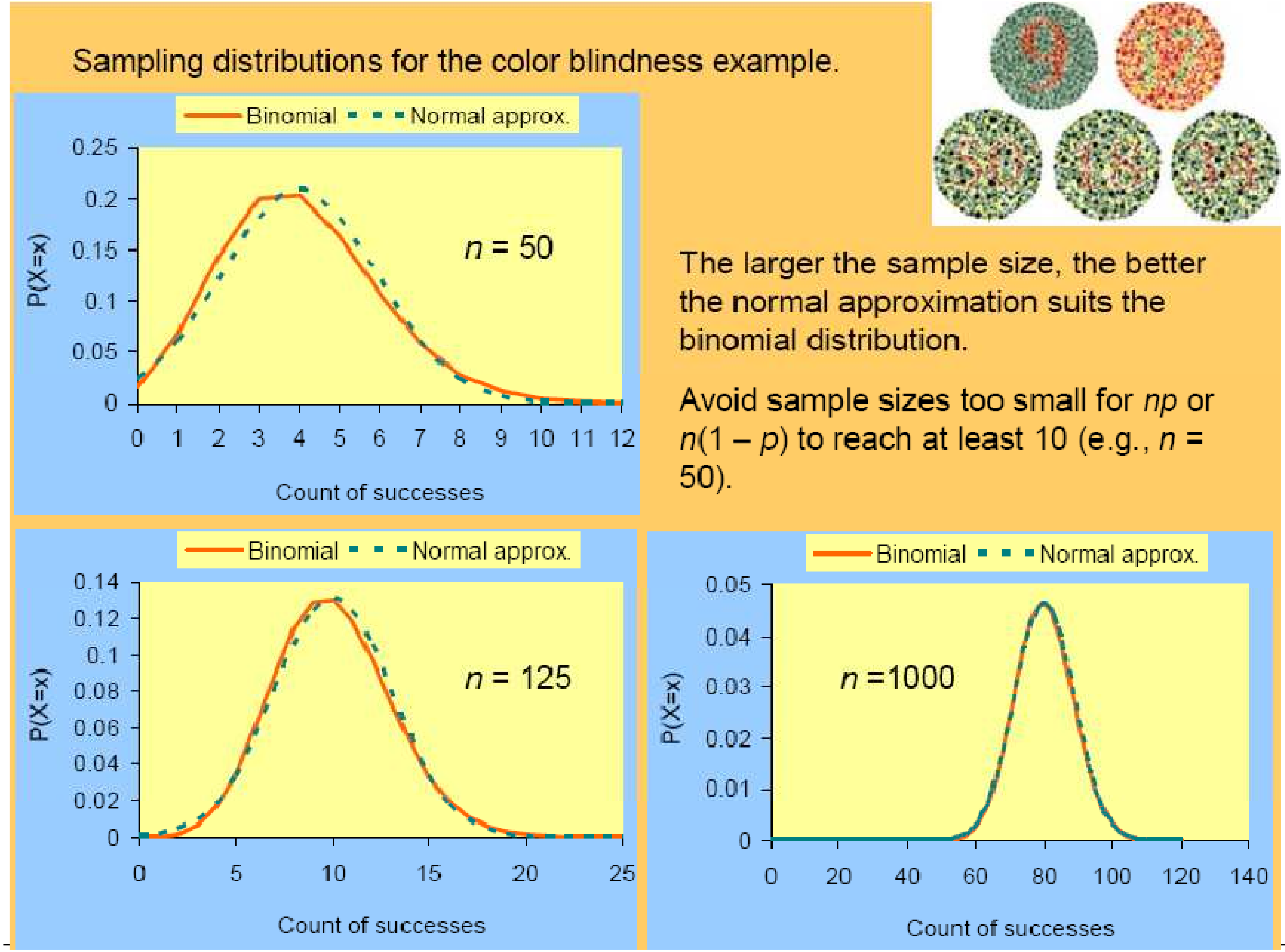
We take a random sample of size 125 from this population. What is the probability that six individuals or fewer in the sample are color blind?

- Sampling distribution of the count X: $B(n = 125, p = 0.08)$ → $np = 10$
  $P(X \leq 6)$ = BINOMDIST(6, 125, .08, 1) = 0.1198 or about 12%

- Normal approximation for the count X: $N(\text{np} = 10, \sqrt{np(1-p)} = 3.033)$
  $P(X \leq 6)$ = NORMDIST(6, 10, 3.033, 1) = 0.0936 or 9%
  Or $z = (x - \mu)/\sigma = (6 - 10)/3.033 = -1.32$ → $P(X \leq 6) = 0.0934$ from Table A

The normal approximation is reasonable, though not perfect. Here $p = 0.08$ is not close to 0.5 when the normal approximation is at its best.

A sample size of 125 is the smallest sample size that can allow use of the normal approximation ($np = 10$ and $n(1 - p) = 115$).



Sampling distributions for the color blindness example.

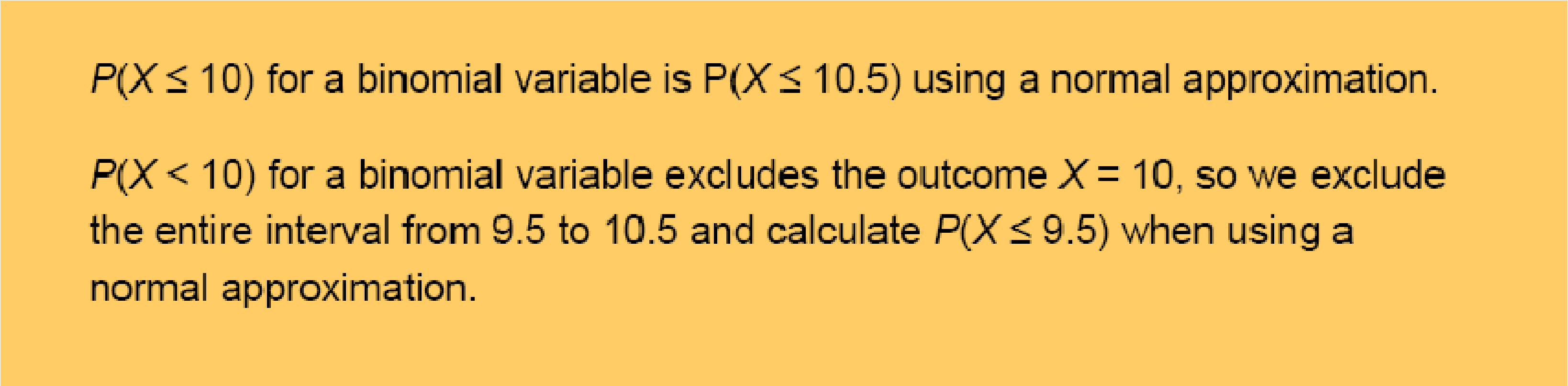
The larger the sample size, the better the normal approximation suits the binomial distribution.

Avoid sample sizes too small for *np* or *n*(1 – *p*) to reach at least 10 (e.g., *n* = 50).

Normal approximation: continuity correction

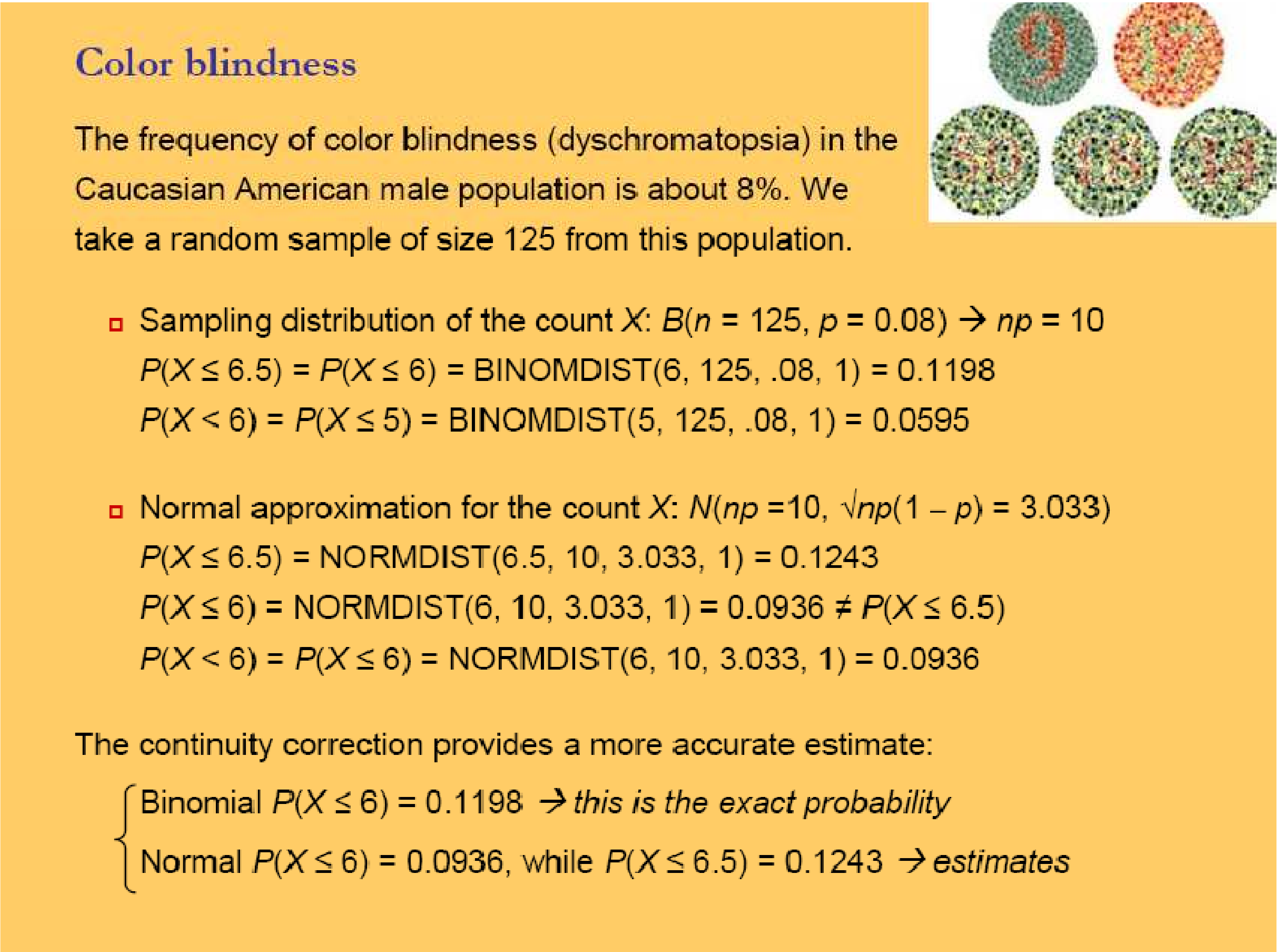
The normal distribution is a better approximation of the binomial distribution, if we perform a continuity correction where $x' = x + 0.5$ is substituted for x, and $P(X \leq x)$ is replaced by $P(X \leq x + 0.5)$.

Why? A binomial random variable is a discrete variable that can only take whole numerical values. In contrast, a normal random variable is a continuous variable that can take any numerical value.



$P(X \leq 10)$ for a binomial variable is $P(X \leq 10.5)$ using a normal approximation.

$P(X < 10)$ for a binomial variable excludes the outcome $X = 10$, so we exclude the entire interval from 9.5 to 10.5 and calculate $P(X \leq 9.5)$ when using a normal approximation.



Color blindness

The frequency of color blindness (dyschromatopsia) in the Caucasian American male population is about 8%. We take a random sample of size 125 from this population.

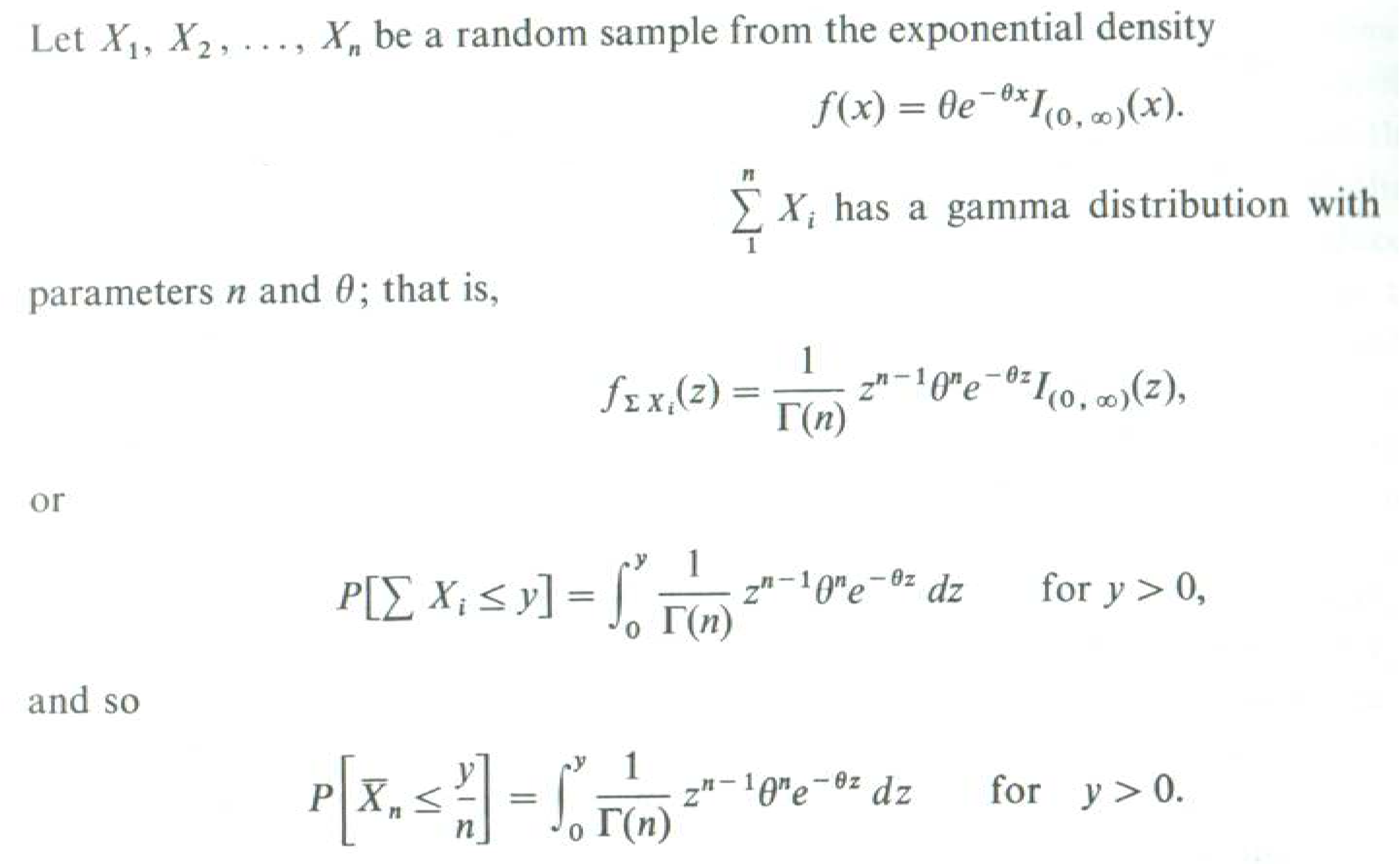
- Sampling distribution of the count X: $B(n = 125, p = 0.08) \rightarrow np = 10$
  $P(X \leq 6.5) = P(X \leq 6) =$ BINOMDIST(6, 125, .08, 1) = 0.1198
  $P(X < 6) = P(X \leq 5) =$ BINOMDIST(5, 125, .08, 1) = 0.0595

- Normal approximation for the count X: $N(np = 10, \sqrt{np(1-p)} = 3.033)$
  $P(X \leq 6.5) =$ NORMDIST(6.5, 10, 3.033, 1) = 0.1243
  $P(X \leq 6) =$ NORMDIST(6, 10, 3.033, 1) = 0.0936 $\neq P(X \leq 6.5)$
  $P(X < 6) = P(X \leq 6) =$ NORMDIST(6, 10, 3.033, 1) = 0.0936

The continuity correction provides a more accurate estimate:

$$\begin{cases} \text{Binomial } P(X \leq 6) = 0.1198 \rightarrow \textit{this is the exact probability} \\ \text{Normal } P(X \leq 6) = 0.0936, \text{ while } P(X \leq 6.5) = 0.1243 \rightarrow \textit{estimates} \end{cases}$$

4.5 Exponential distribution



Let $X_1, X_2, \ldots, X_n$ be a random sample from the exponential density

$$f(x) = \theta e^{-\theta x} I_{(0,\infty)}(x).$$

$\sum_1^n X_i$ has a gamma distribution with parameters n and θ; that is,

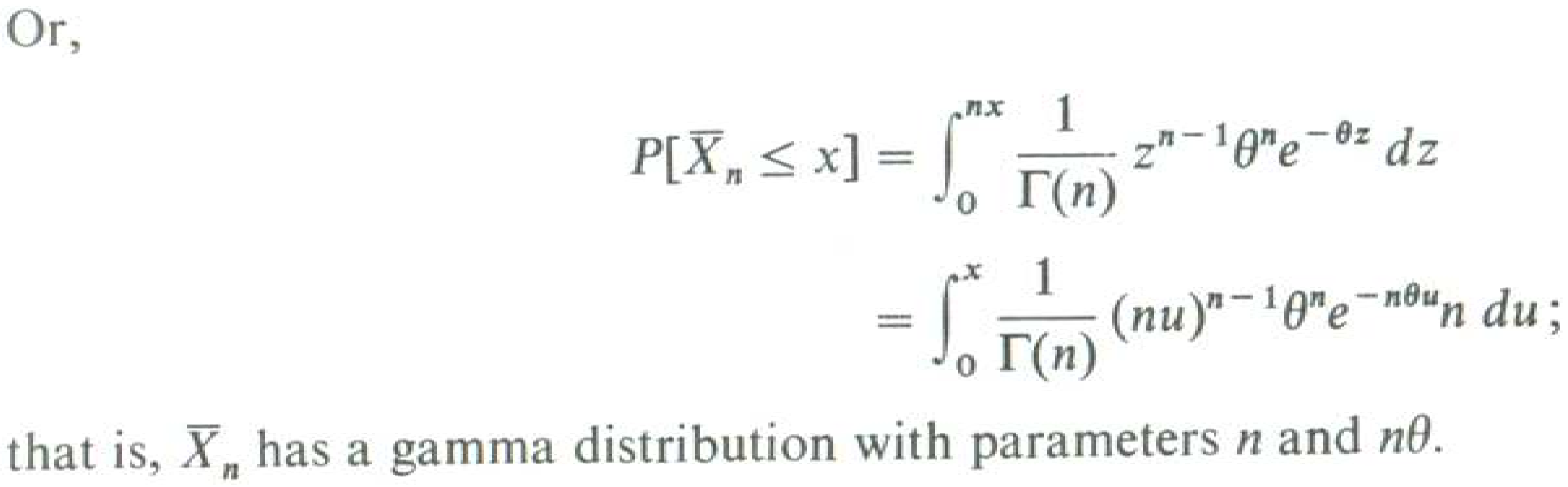
$$f_{\Sigma X_i}(z) = \frac{1}{\Gamma(n)} z^{n-1} \theta^n e^{-\theta z} I_{(0,\infty)}(z),$$

or

$$P[\sum X_i \le y] = \int_0^y \frac{1}{\Gamma(n)} z^{n-1} \theta^n e^{-\theta z}\, dz \qquad \text{for } y > 0,$$

and so

$$P\left[\bar{X}_n \le \frac{y}{n}\right] = \int_0^y \frac{1}{\Gamma(n)} z^{n-1} \theta^n e^{-\theta z}\, dz \qquad \text{for} \quad y > 0.$$



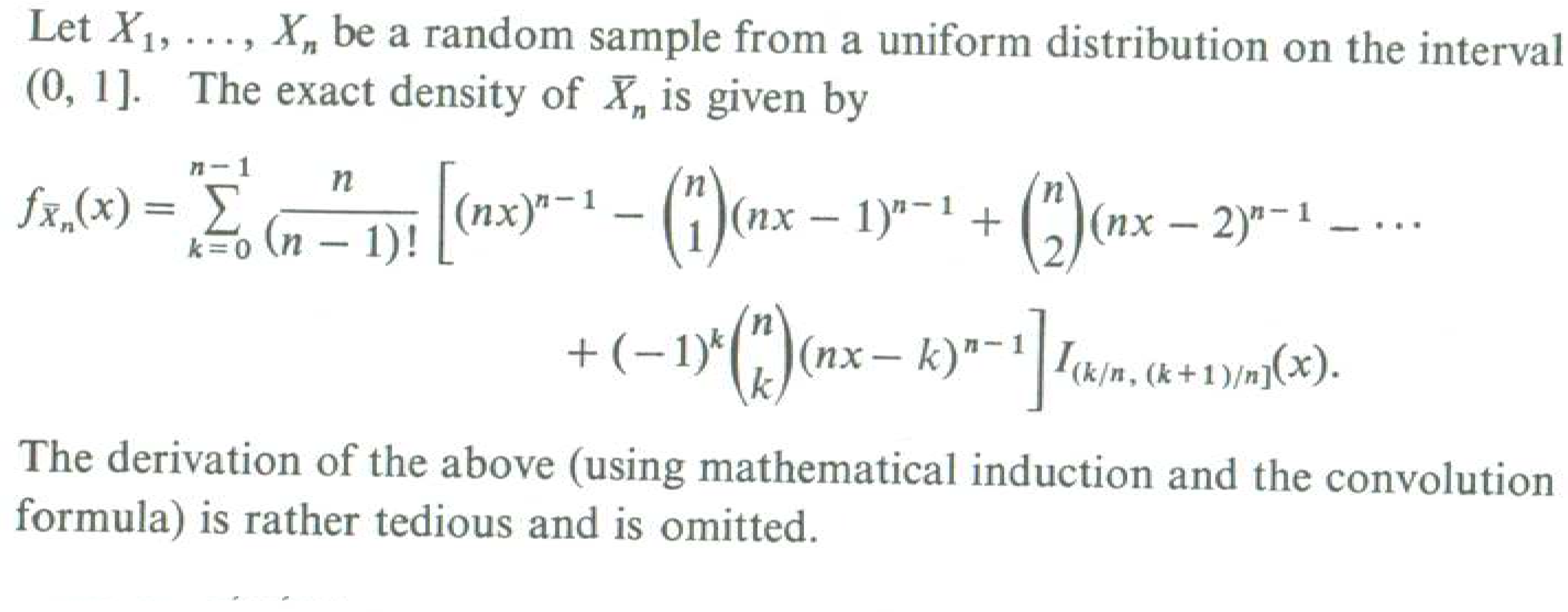
Or,

$$P[\overline{X}_n \leq x] = \int_0^{nx} \frac{1}{\Gamma(n)} z^{n-1}\theta^n e^{-\theta z}\, dz$$

$$= \int_0^{x} \frac{1}{\Gamma(n)} (nu)^{n-1}\theta^n e^{-n\theta u} n\, du\,;$$

that is, $\overline{X}_n$ has a gamma distribution with parameters n and $n\theta$.

4.6 Uniform distribution



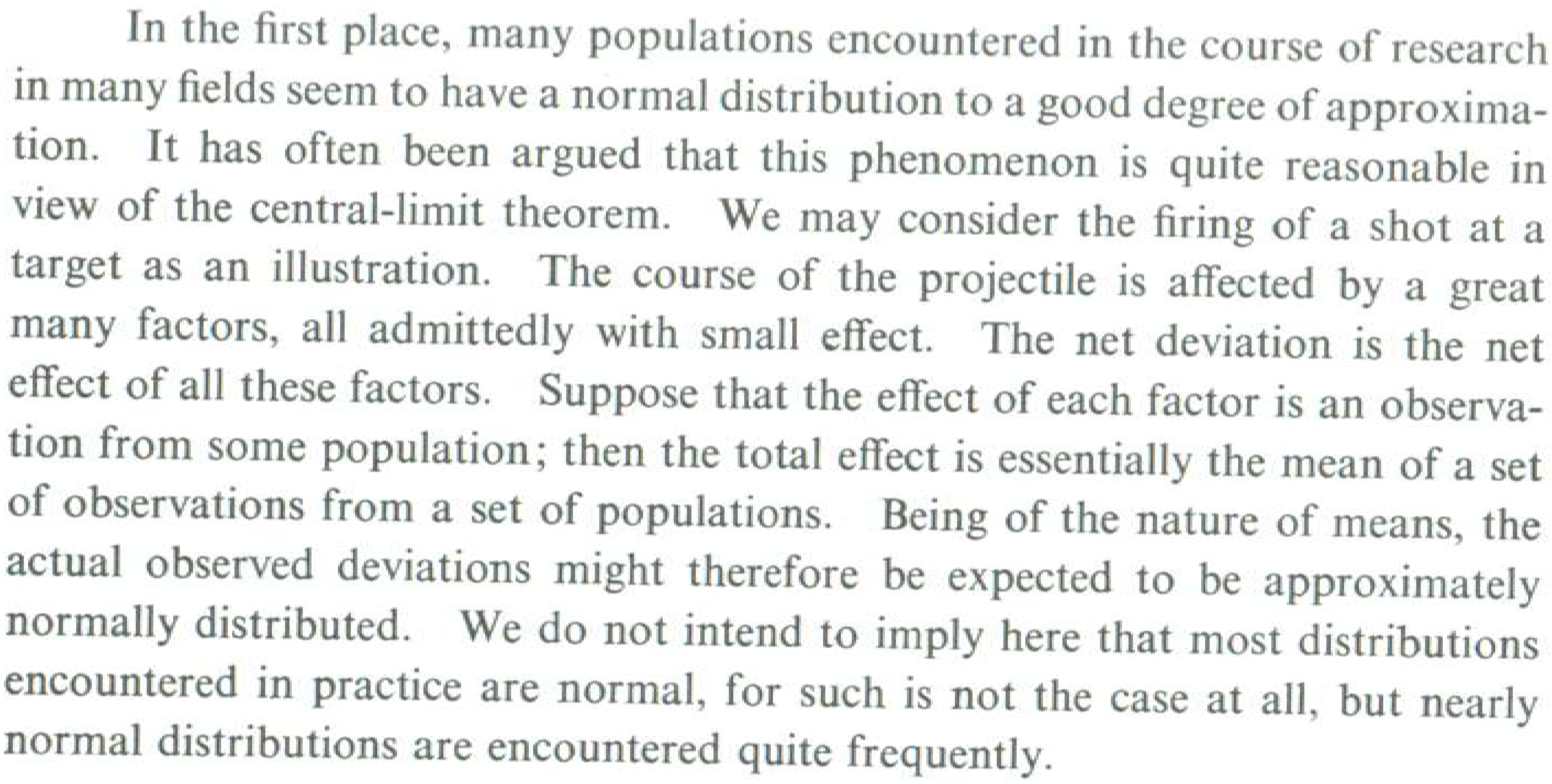
Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution on the interval $(0, 1]$. The exact density of $\bar{X}_n$ is given by

$$f_{\bar{X}_n}(x) = \sum_{k=0}^{n-1} \frac{n}{(n-1)!} \left[(nx)^{n-1} - \binom{n}{1}(nx-1)^{n-1} + \binom{n}{2}(nx-2)^{n-1} - \cdots + (-1)^k \binom{n}{k}(nx-k)^{n-1} \right] I_{(k/n,\,(k+1)/n]}(x).$$

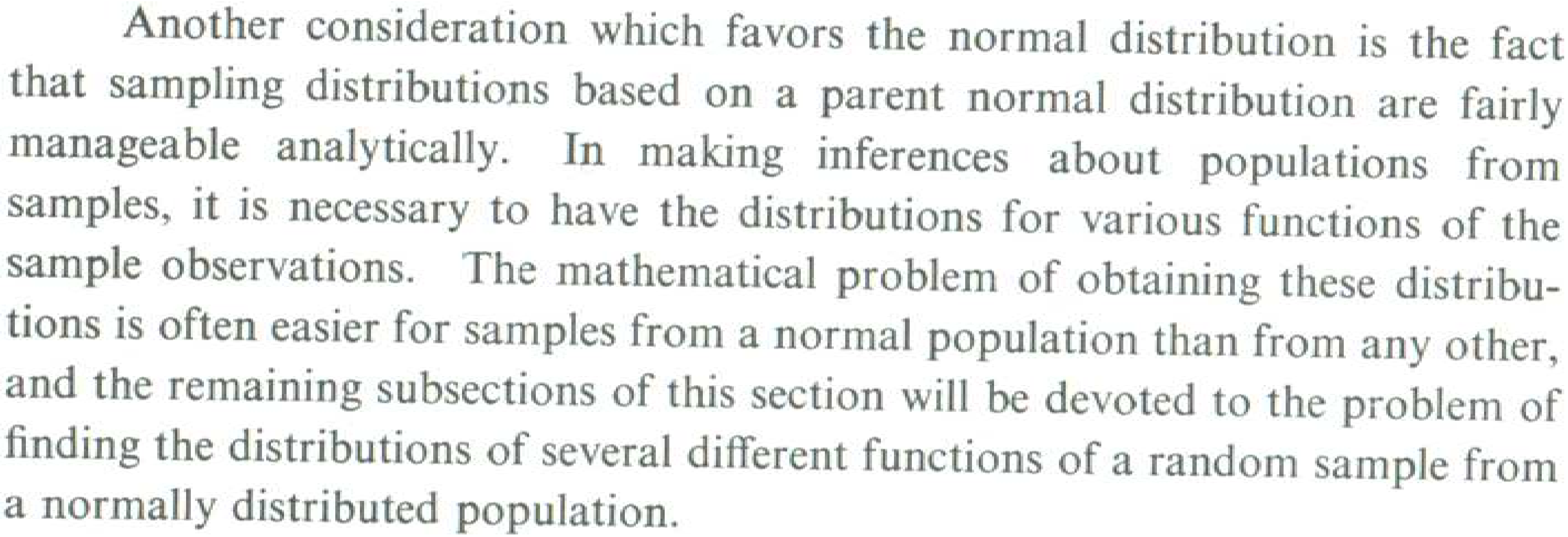
The derivation of the above (using mathematical induction and the convolution formula) is rather tedious and is omitted.

5 Sampling from the normal distribution

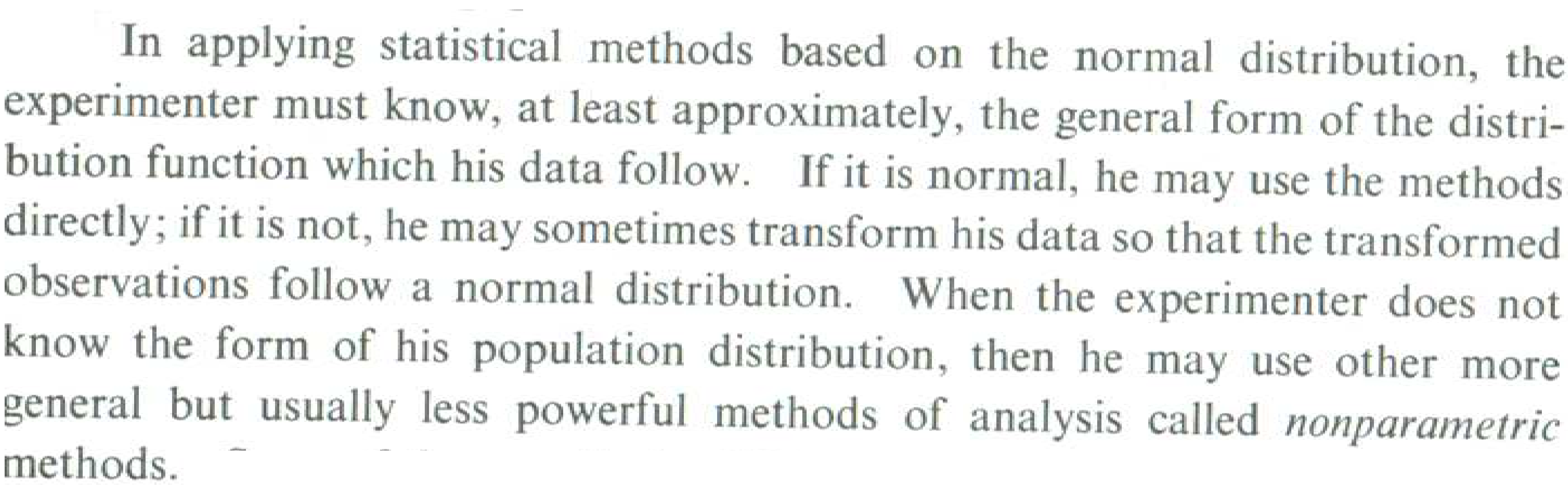
5.1 The role of normal distributions in statistics



In the first place, many populations encountered in the course of research in many fields seem to have a normal distribution to a good degree of approximation. It has often been argued that this phenomenon is quite reasonable in view of the central-limit theorem. We may consider the firing of a shot at a target as an illustration. The course of the projectile is affected by a great many factors, all admittedly with small effect. The net deviation is the net effect of all these factors. Suppose that the effect of each factor is an observation from some population; then the total effect is essentially the mean of a set of observations from a set of populations. Being of the nature of means, the actual observed deviations might therefore be expected to be approximately normally distributed. We do not intend to imply here that most distributions encountered in practice are normal, for such is not the case at all, but nearly normal distributions are encountered quite frequently.

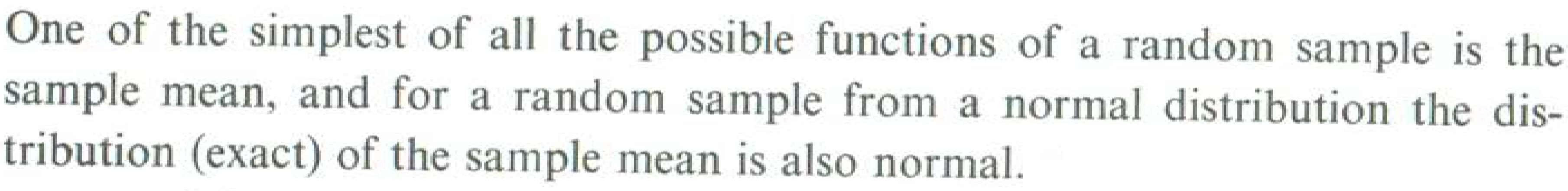


Another consideration which favors the normal distribution is the fact that sampling distributions based on a parent normal distribution are fairly manageable analytically. In making inferences about populations from samples, it is necessary to have the distributions for various functions of the sample observations. The mathematical problem of obtaining these distributions is often easier for samples from a normal population than from any other, and the remaining subsections of this section will be devoted to the problem of finding the distributions of several different functions of a random sample from a normally distributed population.

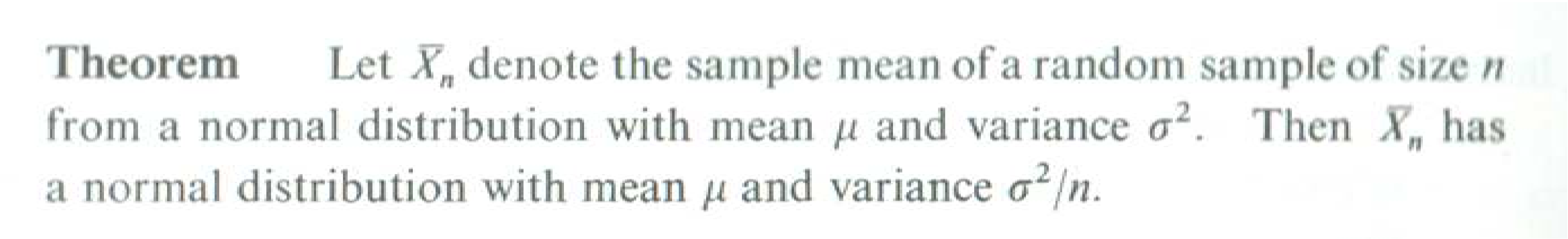


In applying statistical methods based on the normal distribution, the experimenter must know, at least approximately, the general form of the distribution function which his data follow. If it is normal, he may use the methods directly; if it is not, he may sometimes transform his data so that the transformed observations follow a normal distribution. When the experimenter does not know the form of his population distribution, then he may use other more general but usually less powerful methods of analysis called *nonparametric* methods.

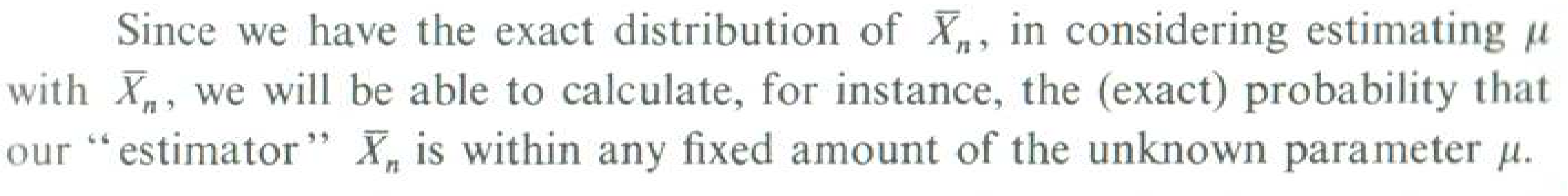
5.2 Sample mean



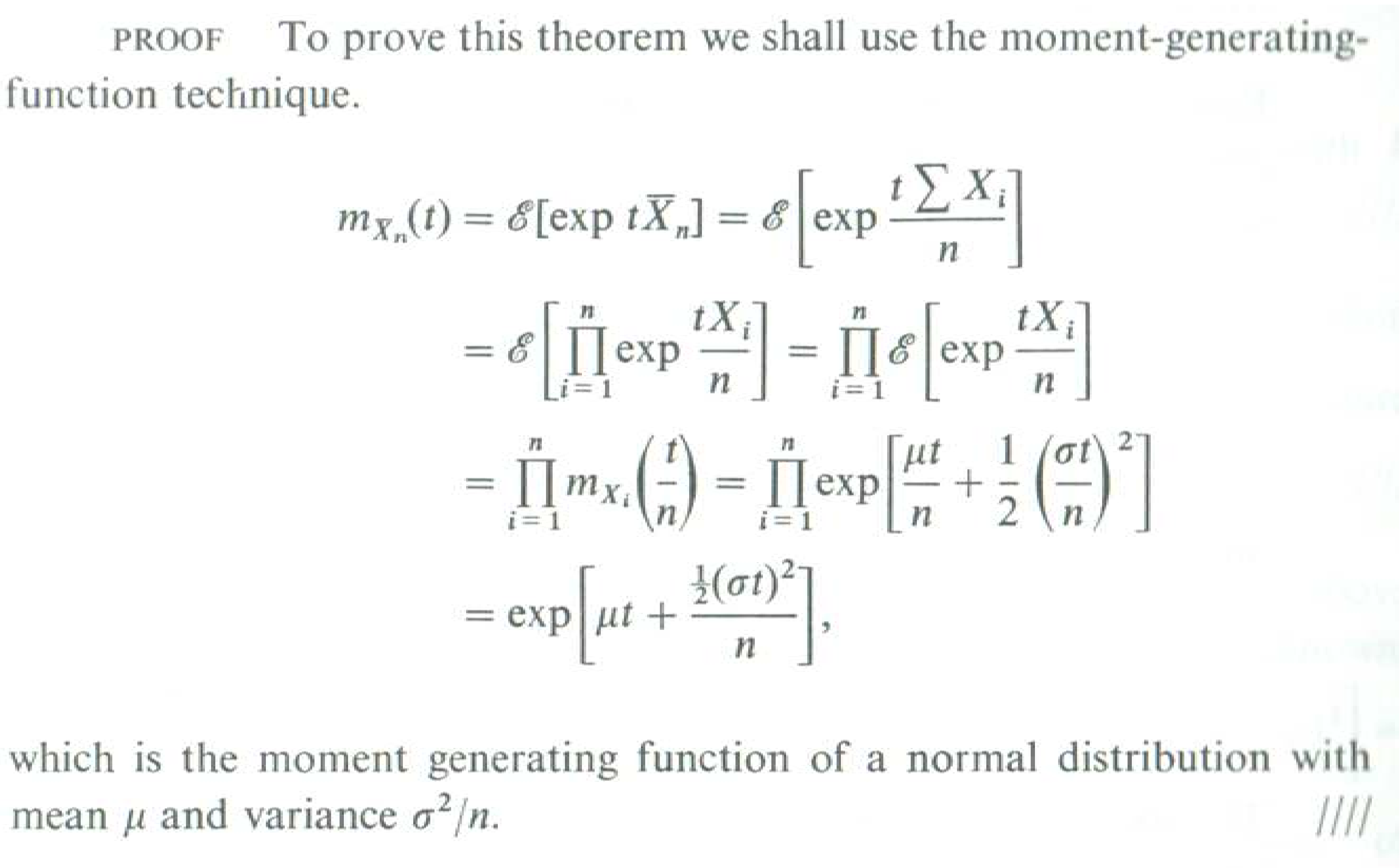
One of the simplest of all the possible functions of a random sample is the sample mean, and for a random sample from a normal distribution the distribution (exact) of the sample mean is also normal.



Theorem Let $\overline{X}_n$ denote the sample mean of a random sample of size n from a normal distribution with mean μ and variance σ^2. Then $\overline{X}_n$ has a normal distribution with mean μ and variance σ^2/n.



Since we have the exact distribution of $\overline{X}_n$, in considering estimating μ with $\overline{X}_n$, we will be able to calculate, for instance, the (exact) probability that our "estimator" $\overline{X}_n$ is within any fixed amount of the unknown parameter μ.

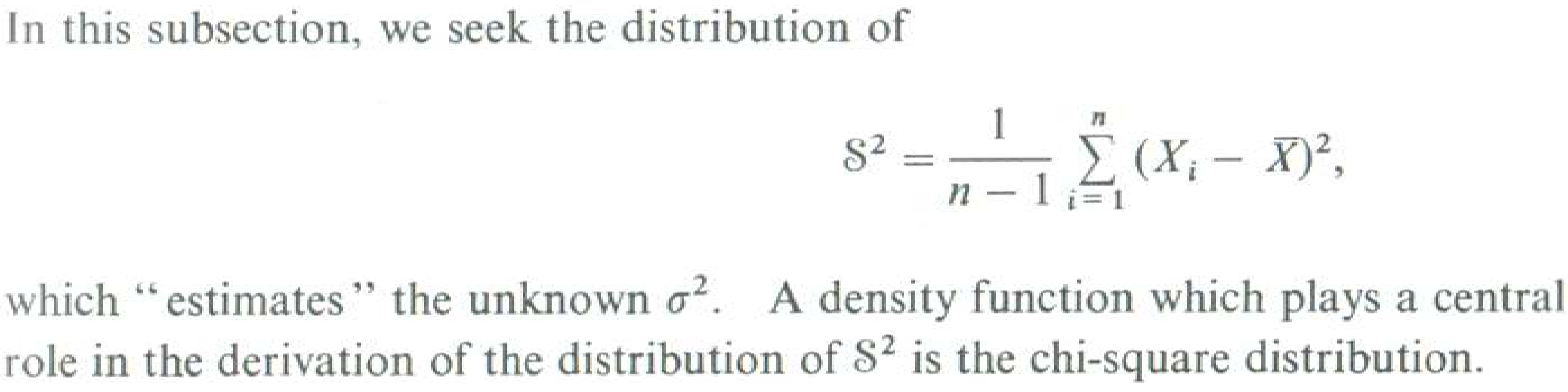


PROOF To prove this theorem we shall use the moment-generating-function technique.

$$\begin{aligned} m_{\bar{X}_n}(t) &= \mathscr{E}[\exp t\bar{X}_n] = \mathscr{E}\left[\exp \frac{t \sum X_i}{n}\right] \\ &= \mathscr{E}\left[\prod_{i=1}^{n} \exp \frac{tX_i}{n}\right] = \prod_{i=1}^{n} \mathscr{E}\left[\exp \frac{tX_i}{n}\right] \\ &= \prod_{i=1}^{n} m_{X_i}\left(\frac{t}{n}\right) = \prod_{i=1}^{n} \exp\left[\frac{\mu t}{n} + \frac{1}{2}\left(\frac{\sigma t}{n}\right)^2\right] \\ &= \exp\left[\mu t + \frac{\frac{1}{2}(\sigma t)^2}{n}\right], \end{aligned}$$

which is the moment generating function of a normal distribution with mean μ and variance σ^2/n. ////

5.3 The chi-square distribution



In this subsection, we seek the distribution of

$$\mathcal{S}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

which "estimates" the unknown σ^2. A density function which plays a central role in the derivation of the distribution of $\mathcal{S}^2$ is the chi-square distribution.

Definition **Chi-square distribution** If X is a random variable with density

$$f_X(x) = \frac{1}{\Gamma(k/2)} \left(\frac{1}{2}\right)^{k/2} x^{k/2-1} e^{-\frac{1}{2}x} I_{(0,\infty)}(x),$$

then X is defined to have a *chi-square distribution with k degrees of freedom*; or the density given in Eq.) is called a *chi-square density with k degrees of freedom*, where the parameter k, called the *degrees of freedom*, is a positive integer. ////

Remark We note that a chi-square density is a particular case of a gamma density with gamma parameters r and λ equal, respectively, to $k/2$ and $\frac{1}{2}$. Hence, if a random variable X has a chi-square distribution,

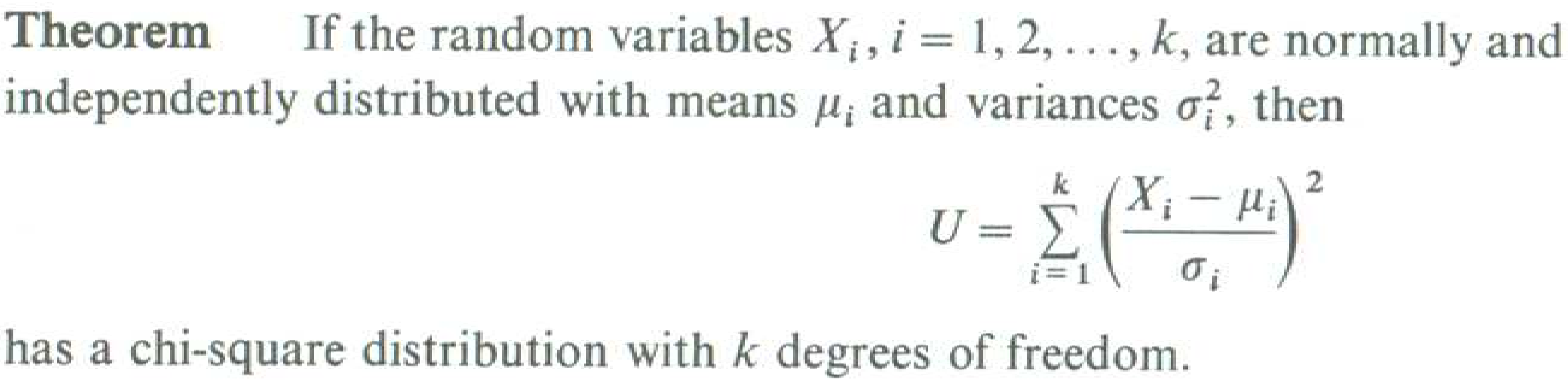
$$\mathscr{E}[X] = \frac{k/2}{\frac{1}{2}} = k,$$

$$\text{var}\,[X] = \frac{k/2}{(1/2)^2} = 2k,$$

and

$$m_X(t) = \left[\frac{\frac{1}{2}}{\frac{1}{2} - t}\right]^{k/2} = \left[\frac{1}{1 - 2t}\right]^{k/2}, \qquad t < 1/2.$$

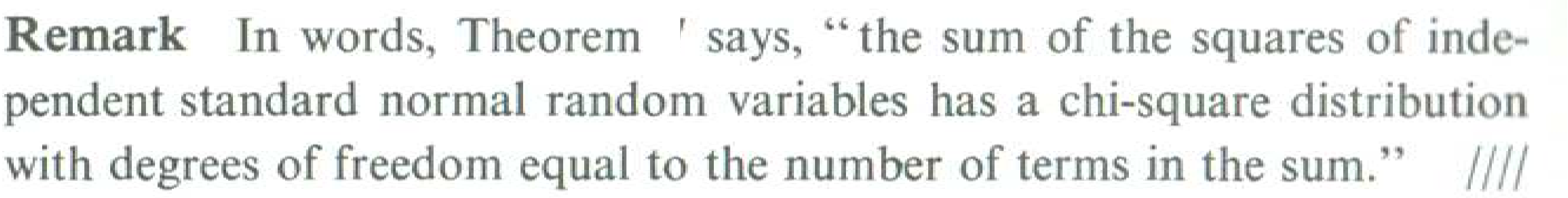
////



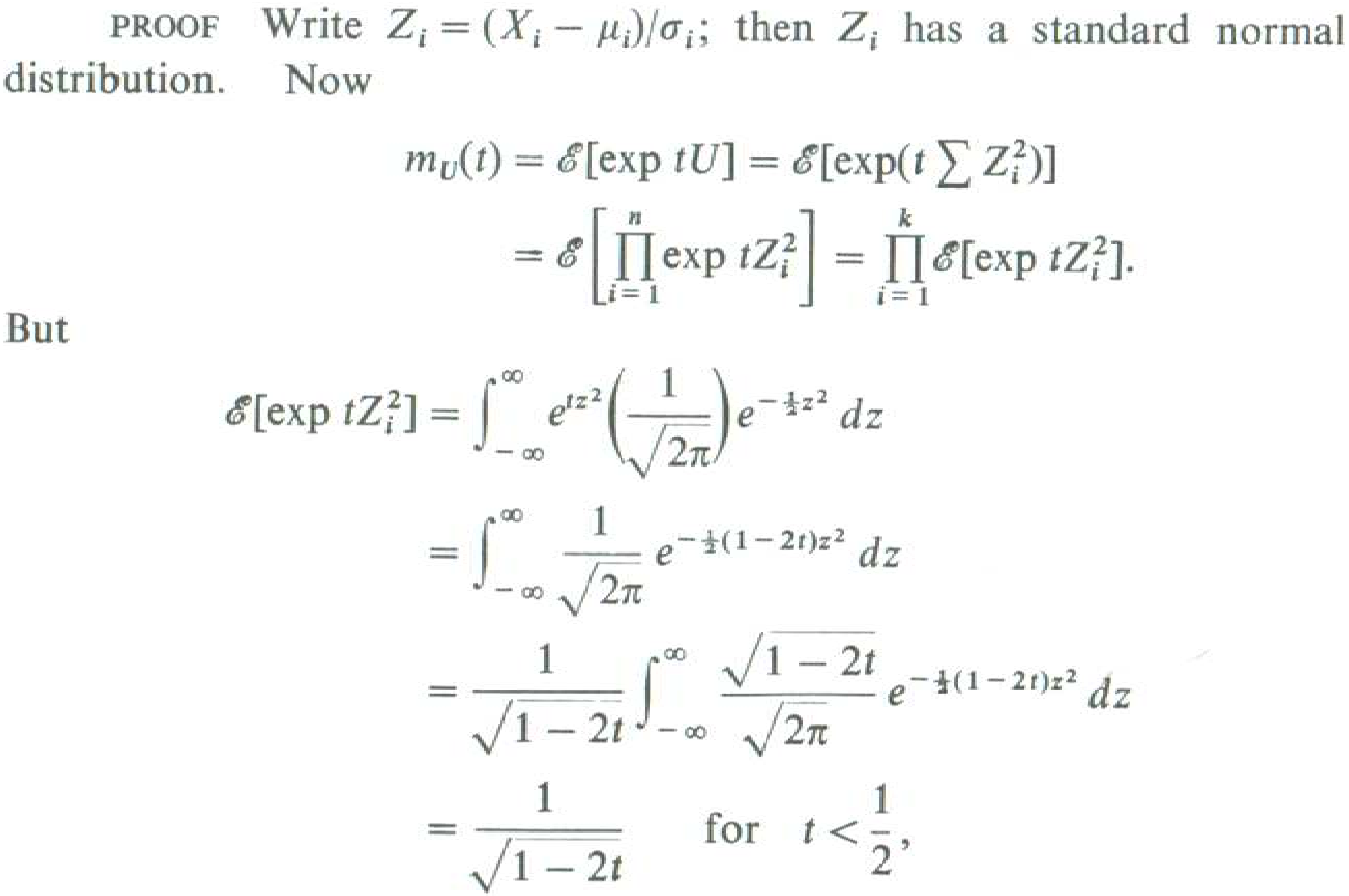
Theorem If the random variables $X_i, i = 1, 2, \ldots, k$, are normally and independently distributed with means μ_i and variances σ_i^2, then

$$U = \sum_{i=1}^{k} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2$$

has a chi-square distribution with k degrees of freedom.



Remark In words, Theorem ' says, "the sum of the squares of independent standard normal random variables has a chi-square distribution with degrees of freedom equal to the number of terms in the sum." ////

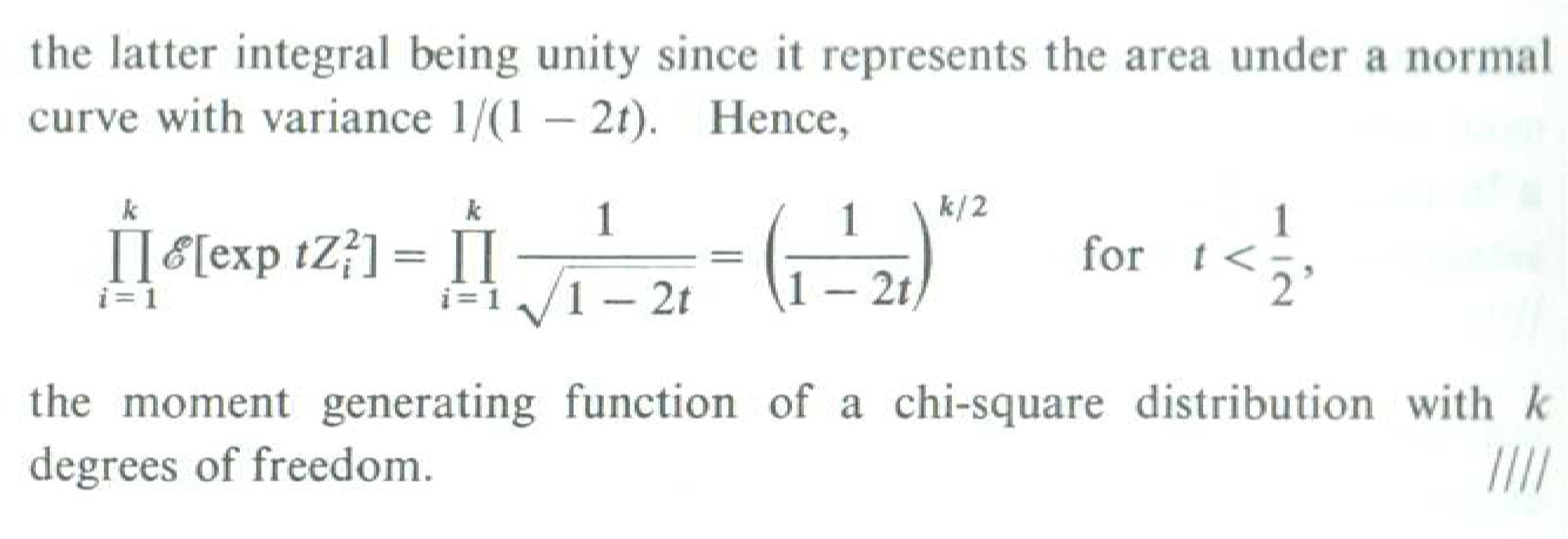


PROOF Write $Z_i = (X_i - \mu_i)/\sigma_i$; then Z_i has a standard normal distribution. Now

$$m_U(t) = \mathscr{E}[\exp tU] = \mathscr{E}[\exp(t \sum Z_i^2)]$$
$$= \mathscr{E}\left[\prod_{i=1}^{n} \exp tZ_i^2\right] = \prod_{i=1}^{k} \mathscr{E}[\exp tZ_i^2].$$

But

$$\mathscr{E}[\exp tZ_i^2] = \int_{-\infty}^{\infty} e^{tz^2}\left(\frac{1}{\sqrt{2\pi}}\right)e^{-\frac{1}{2}z^2}\,dz$$
$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}(1-2t)z^2}\,dz$$
$$= \frac{1}{\sqrt{1-2t}}\int_{-\infty}^{\infty} \frac{\sqrt{1-2t}}{\sqrt{2\pi}}\,e^{-\frac{1}{2}(1-2t)z^2}\,dz$$
$$= \frac{1}{\sqrt{1-2t}} \qquad \text{for} \quad t < \frac{1}{2},$$

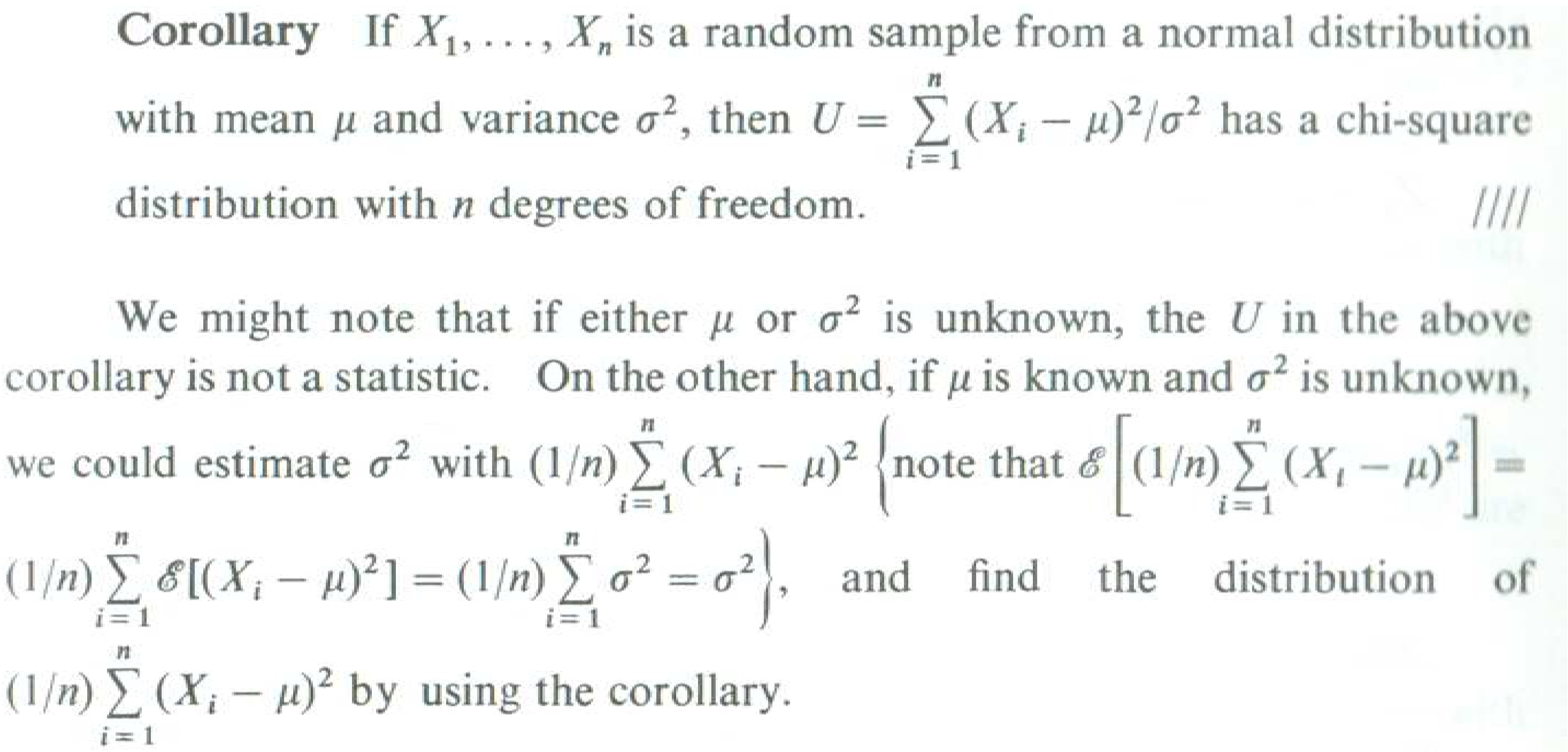


the latter integral being unity since it represents the area under a normal curve with variance $1/(1-2t)$. Hence,

$$\prod_{i=1}^{k} \mathscr{E}[\exp tZ_i^2] = \prod_{i=1}^{k} \frac{1}{\sqrt{1-2t}} = \left(\frac{1}{1-2t}\right)^{k/2} \quad \text{for} \quad t < \frac{1}{2},$$

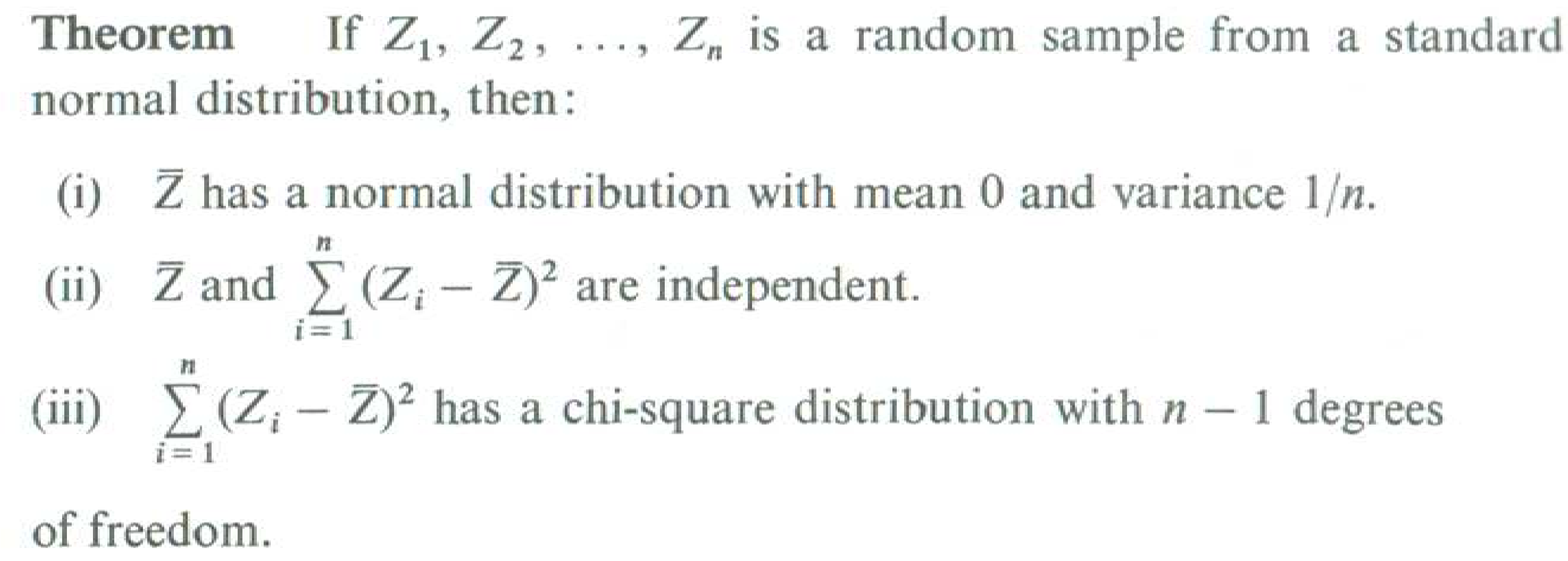
the moment generating function of a chi-square distribution with k degrees of freedom. ////

- Recall: that if two moment generating functions both exist and are equal (they agree), then the corresponding cumulative distribution functions are the same (agree)



Corollary If $X_1, \ldots, X_n$ is a random sample from a normal distribution with mean μ and variance σ^2, then $U = \sum_{i=1}^{n} (X_i - \mu)^2/\sigma^2$ has a chi-square distribution with n degrees of freedom. ////

We might note that if either μ or σ^2 is unknown, the U in the above corollary is not a statistic. On the other hand, if μ is known and σ^2 is unknown, we could estimate σ^2 with $(1/n) \sum_{i=1}^{n} (X_i - \mu)^2$ $\left\{\text{note that } \mathscr{E}\left[(1/n) \sum_{i=1}^{n} (X_i - \mu)^2\right] = (1/n) \sum_{i=1}^{n} \mathscr{E}[(X_i - \mu)^2] = (1/n) \sum_{i=1}^{n} \sigma^2 = \sigma^2\right\}$, and find the distribution of $(1/n) \sum_{i=1}^{n} (X_i - \mu)^2$ by using the corollary.

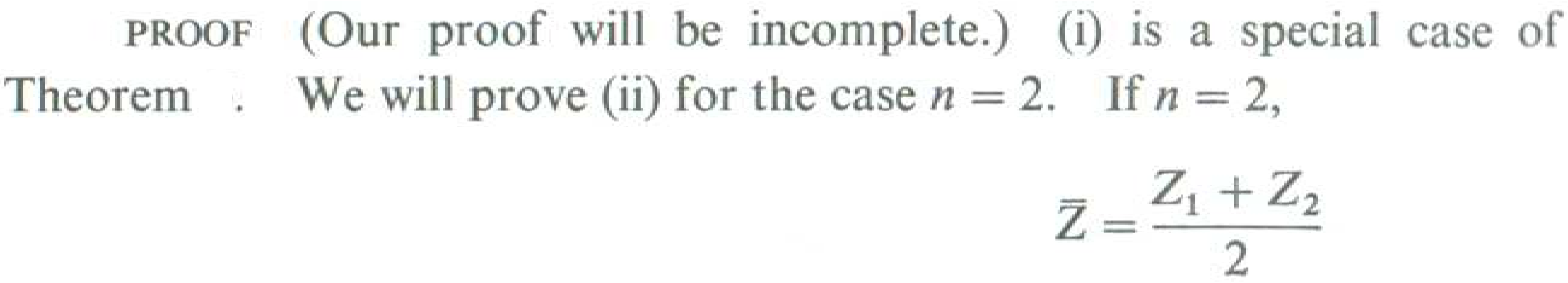


Theorem If $Z_1, Z_2, \ldots, Z_n$ is a random sample from a standard normal distribution, then:

(i) $\bar{Z}$ has a normal distribution with mean 0 and variance $1/n$.

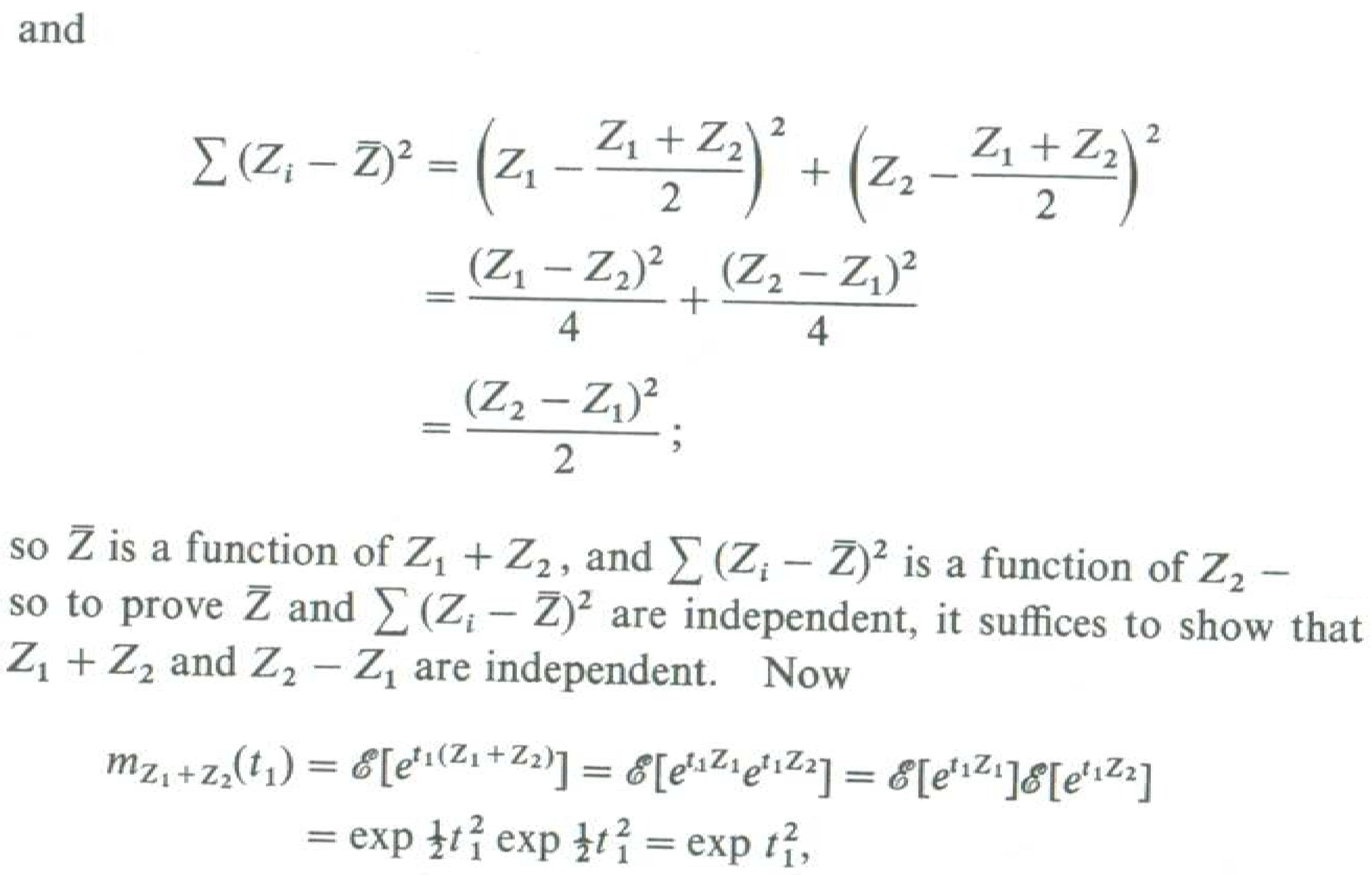
(ii) $\bar{Z}$ and $\sum_{i=1}^{n}(Z_i - \bar{Z})^2$ are independent.

(iii) $\sum_{i=1}^{n}(Z_i - \bar{Z})^2$ has a chi-square distribution with $n - 1$ degrees of freedom.



PROOF (Our proof will be incomplete.) (i) is a special case of Theorem . We will prove (ii) for the case $n = 2$. If $n = 2$,

$$\bar{Z} = \frac{Z_1 + Z_2}{2}$$



and

$$\sum (Z_i - \bar{Z})^2 = \left(Z_1 - \frac{Z_1 + Z_2}{2}\right)^2 + \left(Z_2 - \frac{Z_1 + Z_2}{2}\right)^2$$
$$= \frac{(Z_1 - Z_2)^2}{4} + \frac{(Z_2 - Z_1)^2}{4}$$
$$= \frac{(Z_2 - Z_1)^2}{2};$$

so $\bar{Z}$ is a function of $Z_1 + Z_2$, and $\sum (Z_i - \bar{Z})^2$ is a function of $Z_2 -$ so to prove $\bar{Z}$ and $\sum (Z_i - \bar{Z})^2$ are independent, it suffices to show that $Z_1 + Z_2$ and $Z_2 - Z_1$ are independent. Now

$$m_{Z_1+Z_2}(t_1) = \mathscr{E}[e^{t_1(Z_1+Z_2)}] = \mathscr{E}[e^{t_1 Z_1} e^{t_1 Z_2}] = \mathscr{E}[e^{t_1 Z_1}]\mathscr{E}[e^{t_1 Z_2}]$$
$$= \exp \tfrac{1}{2} t_1^2 \exp \tfrac{1}{2} t_1^2 = \exp t_1^2,$$

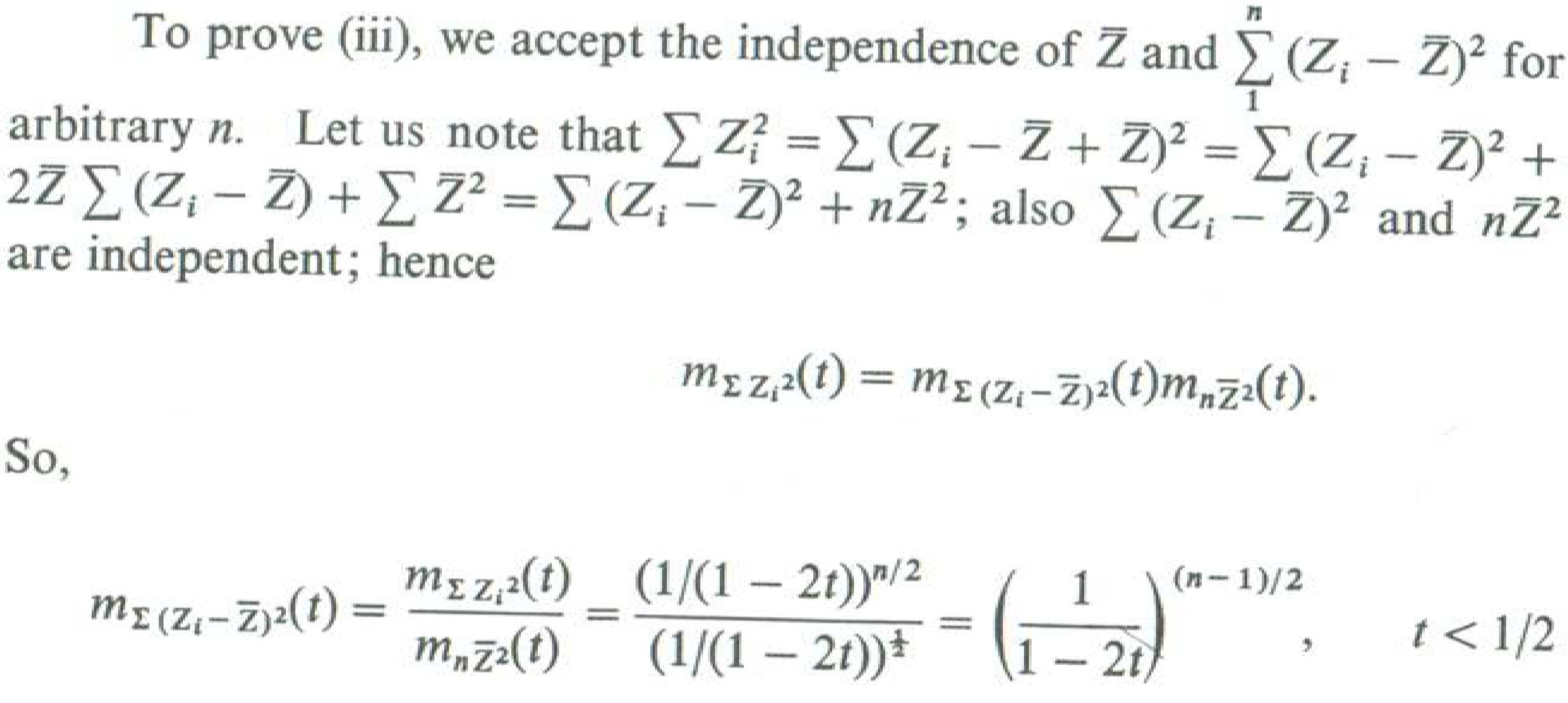
and, similarly,

$$m_{Z_2-Z_1}(t_2) = \exp t_2^2 .$$

Also,

$$\begin{aligned} m_{Z_1+Z_2,\,Z_2-Z_1}(t_1, t_2) &= \mathscr{E}[e^{t_1(Z_1+Z_2)+t_2(Z_2-Z_1)}] \\ &= \mathscr{E}[e^{(t_1-t_2)Z_1}e^{(t_1+t_2)Z_2}] = \mathscr{E}[e^{(t_1-t_2)Z_1}]\mathscr{E}[e^{(t_1+t_2)Z_2}] \\ &= e^{\frac{1}{2}(t_1-t_2)^2}e^{\frac{1}{2}(t_1+t_2)^2} = \exp t_1^2 \exp t_2^2 \\ &= m_{Z_1+Z_2}(t_1)m_{Z_2-Z_1}(t_2); \end{aligned}$$

and since the joint moment generating function factors into the product of the marginal moment generating functions, $Z_1 + Z_2$ and $Z_2 - Z_1$ are independent.

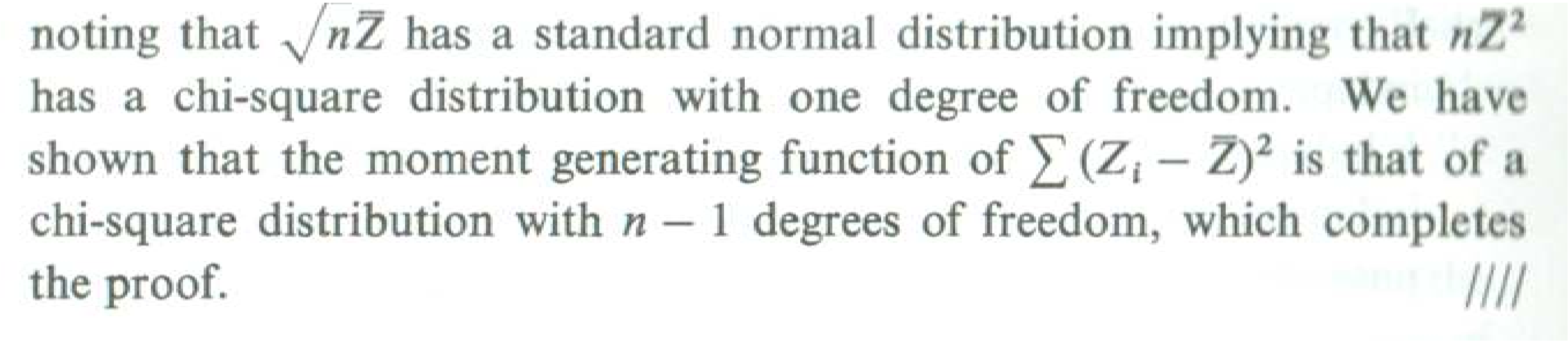


To prove (iii), we accept the independence of $\bar{Z}$ and $\sum_1^n (Z_i - \bar{Z})^2$ for arbitrary n. Let us note that $\sum Z_i^2 = \sum (Z_i - \bar{Z} + \bar{Z})^2 = \sum (Z_i - \bar{Z})^2 + 2\bar{Z} \sum (Z_i - \bar{Z}) + \sum \bar{Z}^2 = \sum (Z_i - \bar{Z})^2 + n\bar{Z}^2$; also $\sum (Z_i - \bar{Z})^2$ and $n\bar{Z}^2$ are independent; hence

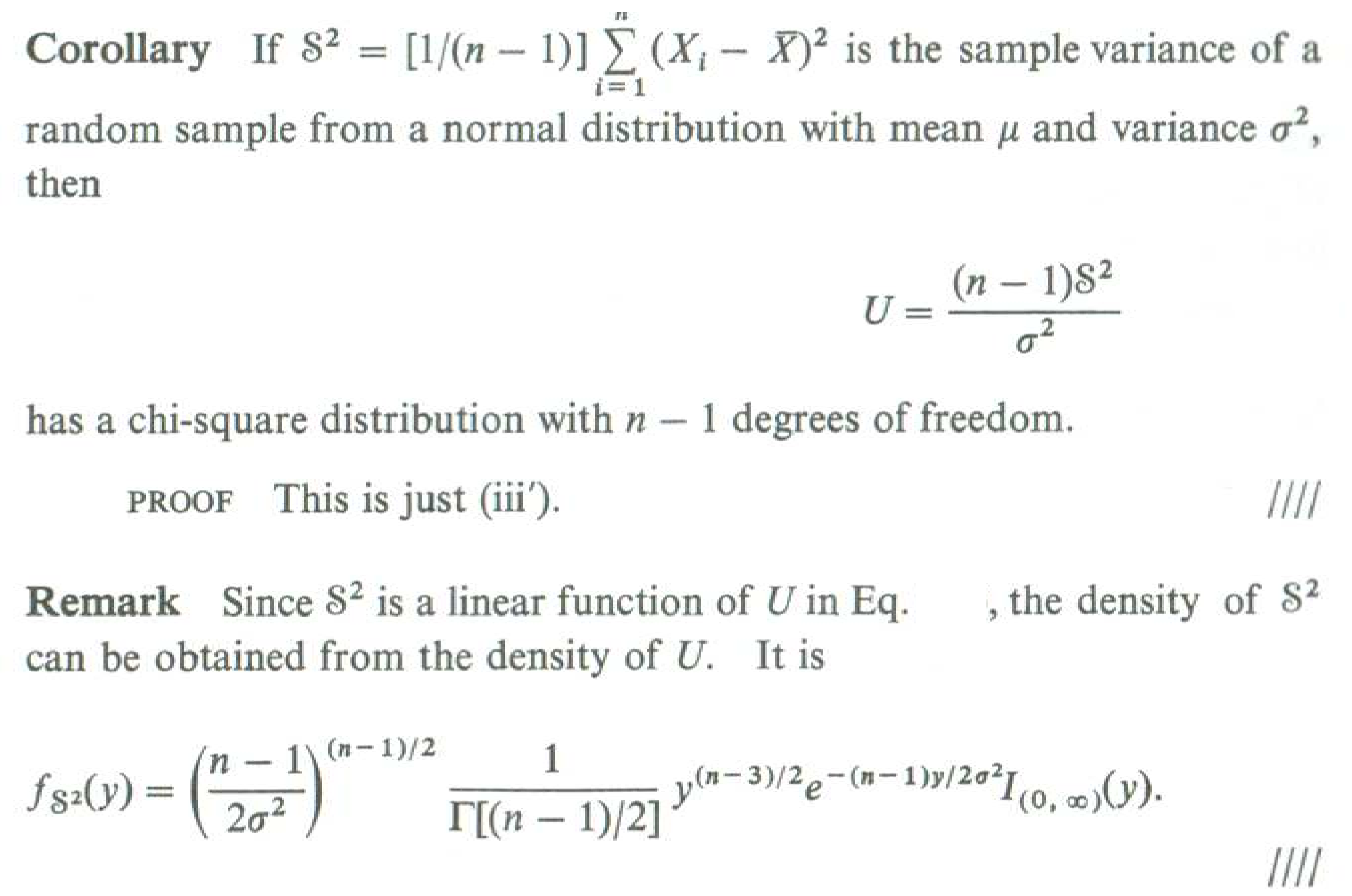
$$m_{\Sigma Z_i^2}(t) = m_{\Sigma (Z_i - \bar{Z})^2}(t) m_{n\bar{Z}^2}(t).$$

So,

$$m_{\Sigma (Z_i - \bar{Z})^2}(t) = \frac{m_{\Sigma Z_i^2}(t)}{m_{n\bar{Z}^2}(t)} = \frac{(1/(1-2t))^{n/2}}{(1/(1-2t))^{\frac{1}{2}}} = \left(\frac{1}{1-2t}\right)^{(n-1)/2}, \qquad t < 1/2$$



noting that $\sqrt{n}\bar{Z}$ has a standard normal distribution implying that $n\bar{Z}^2$ has a chi-square distribution with one degree of freedom. We have shown that the moment generating function of $\sum (Z_i - \bar{Z})^2$ is that of a chi-square distribution with $n - 1$ degrees of freedom, which completes the proof. ////



Corollary If $\mathcal{S}^2 = [1/(n-1)] \sum_{i=1}^{n} (X_i - \bar{X})^2$ is the sample variance of a random sample from a normal distribution with mean μ and variance σ^2, then

$$U = \frac{(n-1)\mathcal{S}^2}{\sigma^2}$$

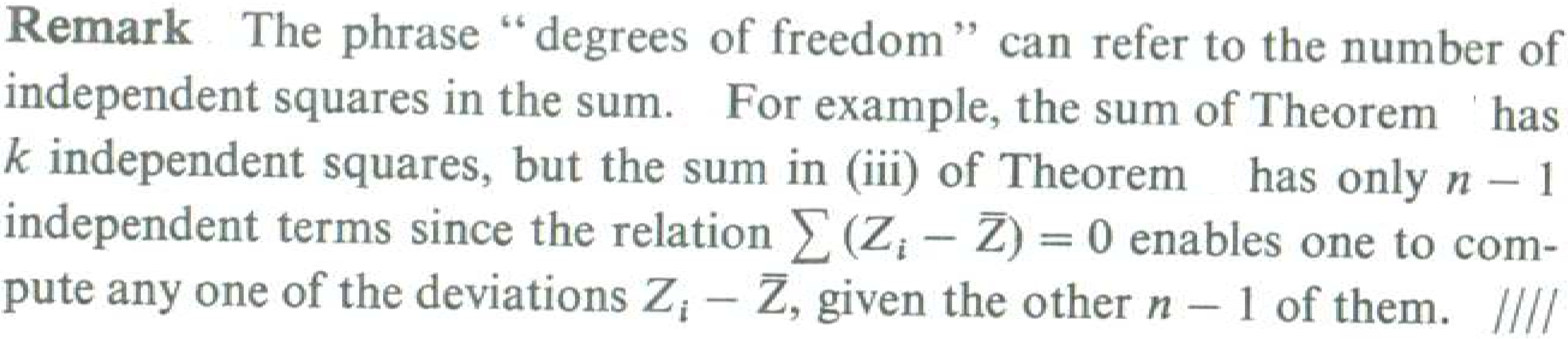
has a chi-square distribution with $n - 1$ degrees of freedom.

PROOF This is just (iii′). ////

Remark Since $\mathcal{S}^2$ is a linear function of U in Eq. , the density of $\mathcal{S}^2$ can be obtained from the density of U. It is

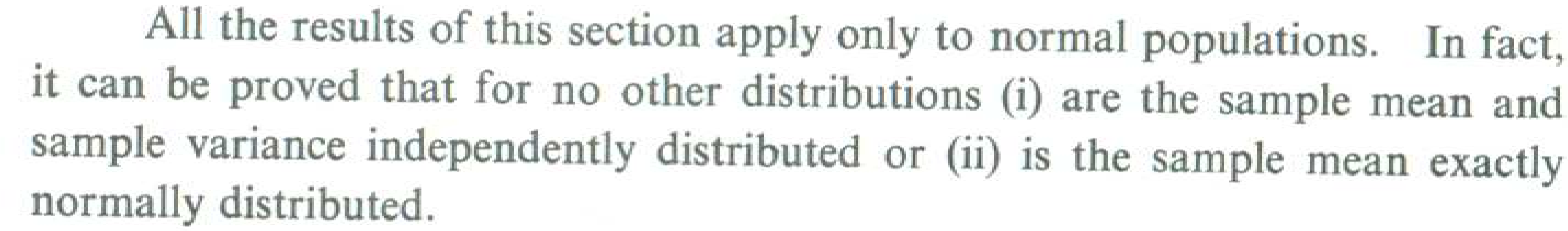
$$f_{\mathcal{S}^2}(y) = \left(\frac{n-1}{2\sigma^2}\right)^{(n-1)/2} \frac{1}{\Gamma[(n-1)/2]} y^{(n-3)/2} e^{-(n-1)y/2\sigma^2} I_{(0,\infty)}(y).$$

////

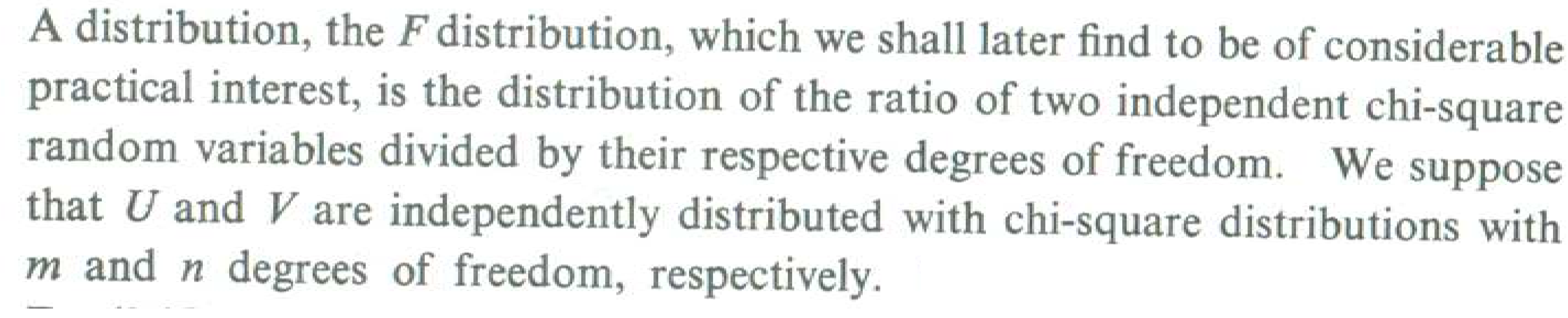


Remark The phrase "degrees of freedom" can refer to the number of independent squares in the sum. For example, the sum of Theorem has k independent squares, but the sum in (iii) of Theorem has only $n - 1$ independent terms since the relation $\sum (Z_i - \bar{Z}) = 0$ enables one to compute any one of the deviations $Z_i - \bar{Z}$, given the other $n - 1$ of them. ////

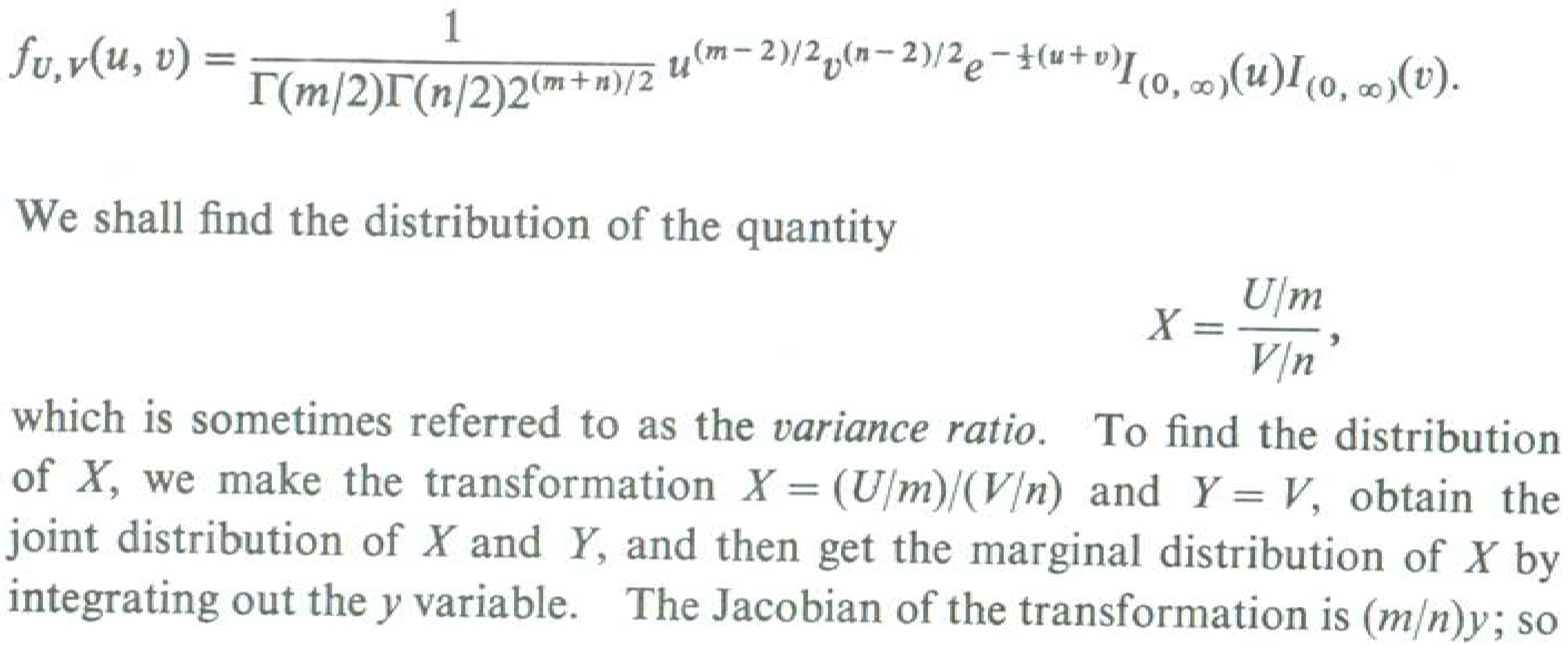
5.4 The F distribution



All the results of this section apply only to normal populations. In fact, it can be proved that for no other distributions (i) are the sample mean and sample variance independently distributed or (ii) is the sample mean exactly normally distributed.



A distribution, the *F* distribution, which we shall later find to be of considerable practical interest, is the distribution of the ratio of two independent chi-square random variables divided by their respective degrees of freedom. We suppose that U and V are independently distributed with chi-square distributions with m and n degrees of freedom, respectively.



$$f_{U,V}(u, v) = \frac{1}{\Gamma(m/2)\Gamma(n/2)2^{(m+n)/2}} u^{(m-2)/2} v^{(n-2)/2} e^{-\frac{1}{2}(u+v)} I_{(0,\infty)}(u) I_{(0,\infty)}(v).$$

We shall find the distribution of the quantity

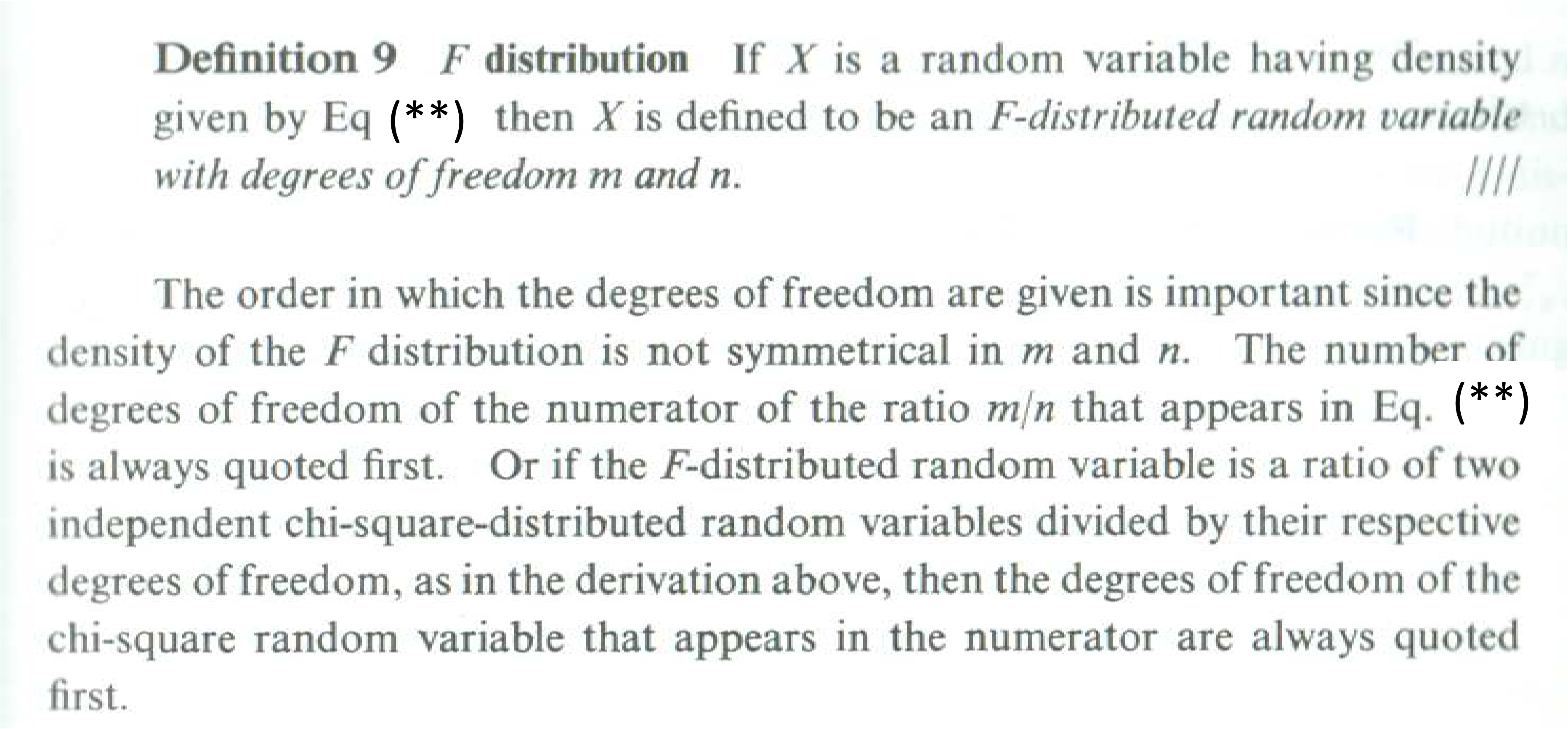
$$X = \frac{U/m}{V/n},$$

which is sometimes referred to as the *variance ratio*. To find the distribution of X, we make the transformation $X = (U/m)/(V/n)$ and $Y = V$, obtain the joint distribution of X and Y, and then get the marginal distribution of X by integrating out the y variable. The Jacobian of the transformation is $(m/n)y$; so

$$f_{X,Y}(x,y) = \frac{m}{n}\, y \,\frac{1}{\Gamma(m/2)\Gamma(n/2)2^{(m+n)/2}} \left(\frac{m}{n}\,xy\right)^{(m-2)/2} y^{(n-2)/2} e^{-\frac{1}{2}[(m/n)xy+y]},$$

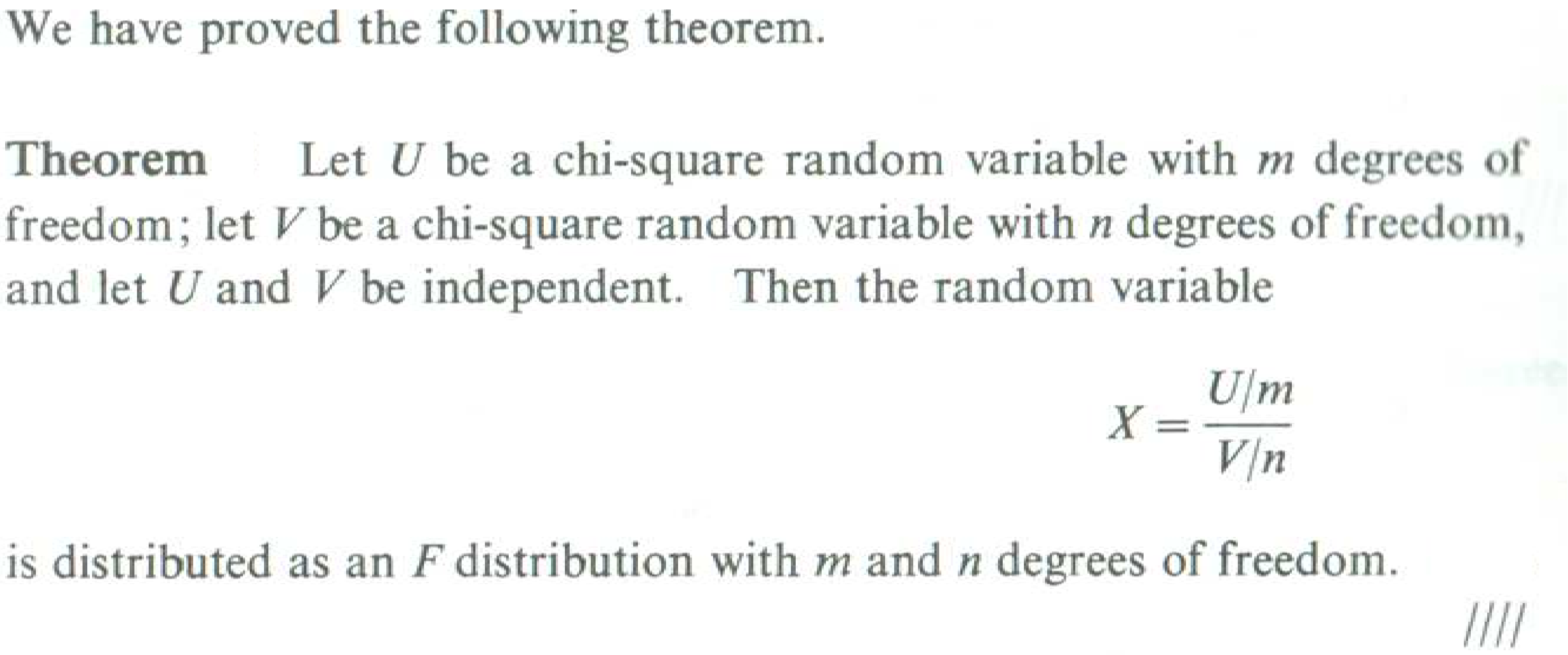
and

$$\begin{aligned} f_X(x) &= \int_0^\infty f_{X,Y}(x,y)\,dy \\ &= \frac{1}{\Gamma(m/2)\Gamma(n/2)2^{(m+n)/2}} \left(\frac{m}{n}\right)^{m/2} x^{(m-2)/2} \int_0^\infty y^{(m+n-2)/2} e^{-\frac{1}{2}[(m/n)x+1]y}\,dy \\ &= \frac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)} \left(\frac{m}{n}\right)^{m/2} \frac{x^{(m-2)/2}}{[1+(m/n)x]^{(m+n)/2}}\, I_{(0,\infty)}(x). \qquad (**) \end{aligned}$$



Definition 9 ***F* distribution** If X is a random variable having density given by Eq (**) then X is defined to be an *F-distributed random variable with degrees of freedom m and n.* ////

The order in which the degrees of freedom are given is important since the density of the F distribution is not symmetrical in m and n. The number of degrees of freedom of the numerator of the ratio m/n that appears in Eq. (**) is always quoted first. Or if the F-distributed random variable is a ratio of two independent chi-square-distributed random variables divided by their respective degrees of freedom, as in the derivation above, then the degrees of freedom of the chi-square random variable that appears in the numerator are always quoted first.



We have proved the following theorem.

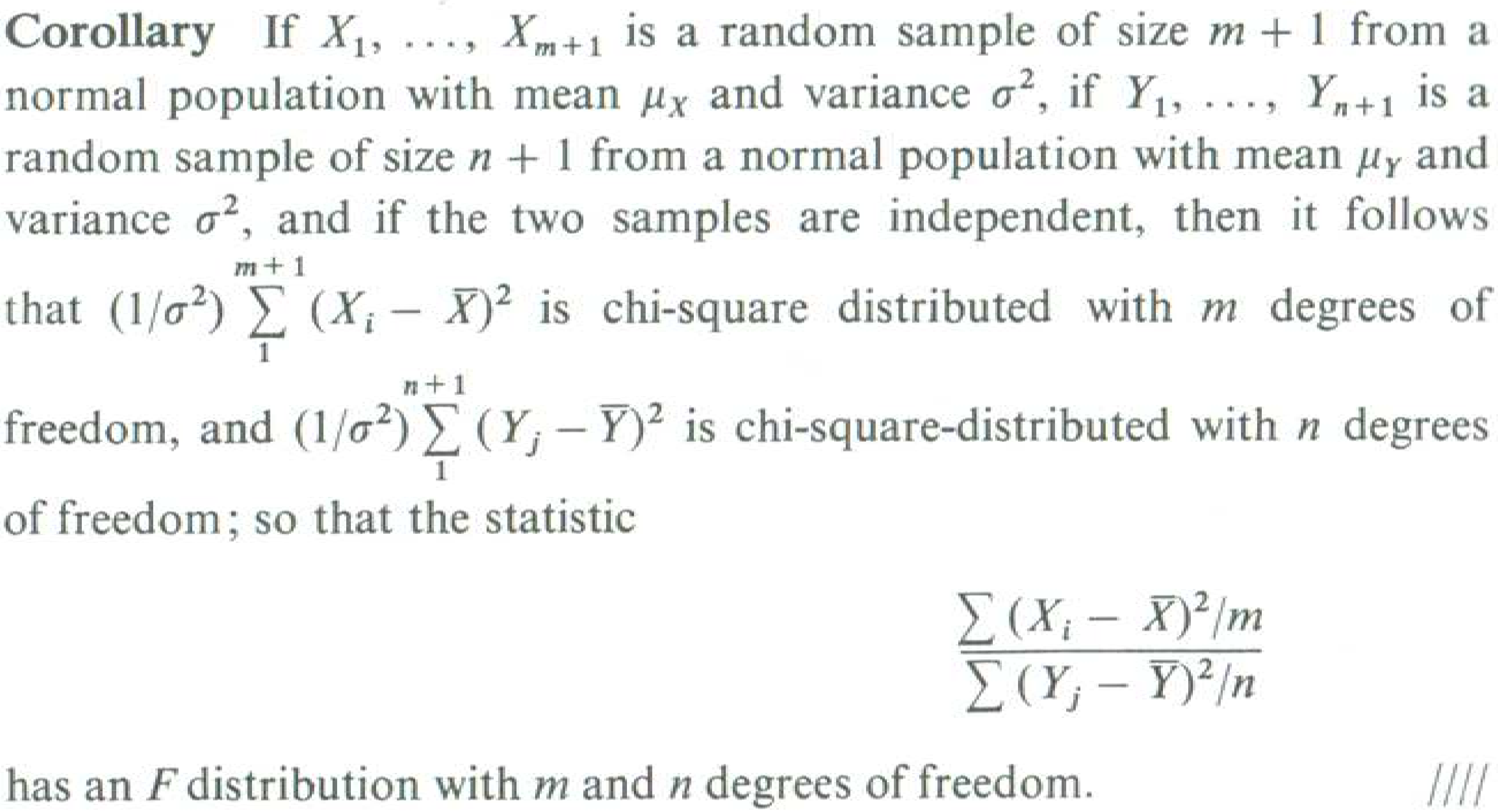
Theorem Let U be a chi-square random variable with m degrees of freedom; let V be a chi-square random variable with n degrees of freedom, and let U and V be independent. Then the random variable

$$X = \frac{U/m}{V/n}$$

is distributed as an F distribution with m and n degrees of freedom.

////

This theorem can be very useful in sampling



Corollary If $X_1, \ldots, X_{m+1}$ is a random sample of size $m+1$ from a normal population with mean μ_X and variance σ^2, if $Y_1, \ldots, Y_{n+1}$ is a random sample of size $n+1$ from a normal population with mean μ_Y and variance σ^2, and if the two samples are independent, then it follows that $(1/\sigma^2) \sum_1^{m+1} (X_i - \bar{X})^2$ is chi-square distributed with m degrees of freedom, and $(1/\sigma^2) \sum_1^{n+1} (Y_j - \bar{Y})^2$ is chi-square-distributed with n degrees of freedom; so that the statistic

$$\frac{\sum (X_i - \bar{X})^2/m}{\sum (Y_j - \bar{Y})^2/n}$$

has an F distribution with m and n degrees of freedom. ////

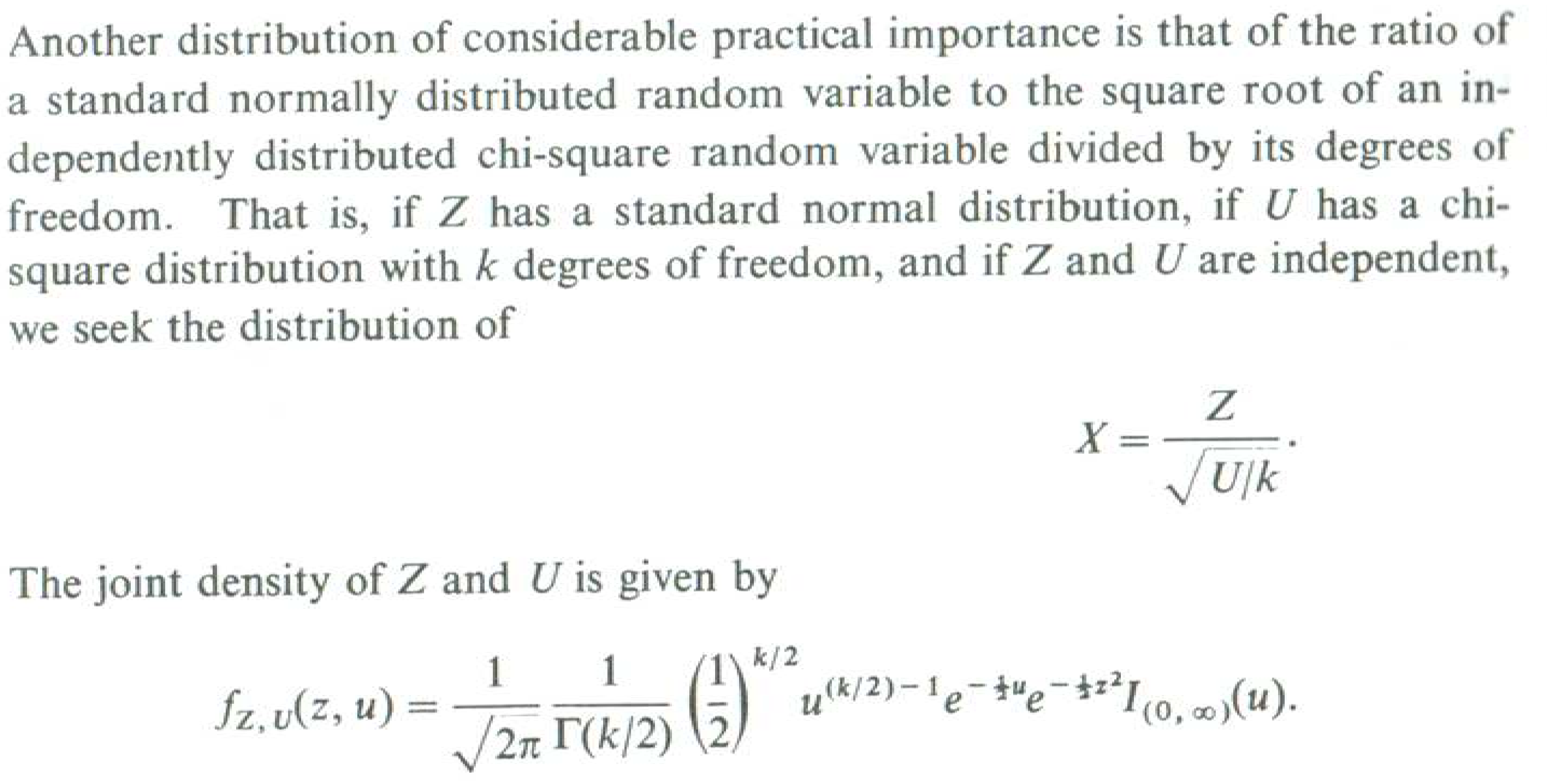
Remark If X is an F-distributed random variable with m and n degrees of freedom, then

$$\mathscr{E}[X] = \frac{n}{n-2} \qquad \text{for } n > 2$$

and

$$\text{var}\,[X] = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \qquad \text{for } n > 4.$$

5.5 The Student’s t distribution

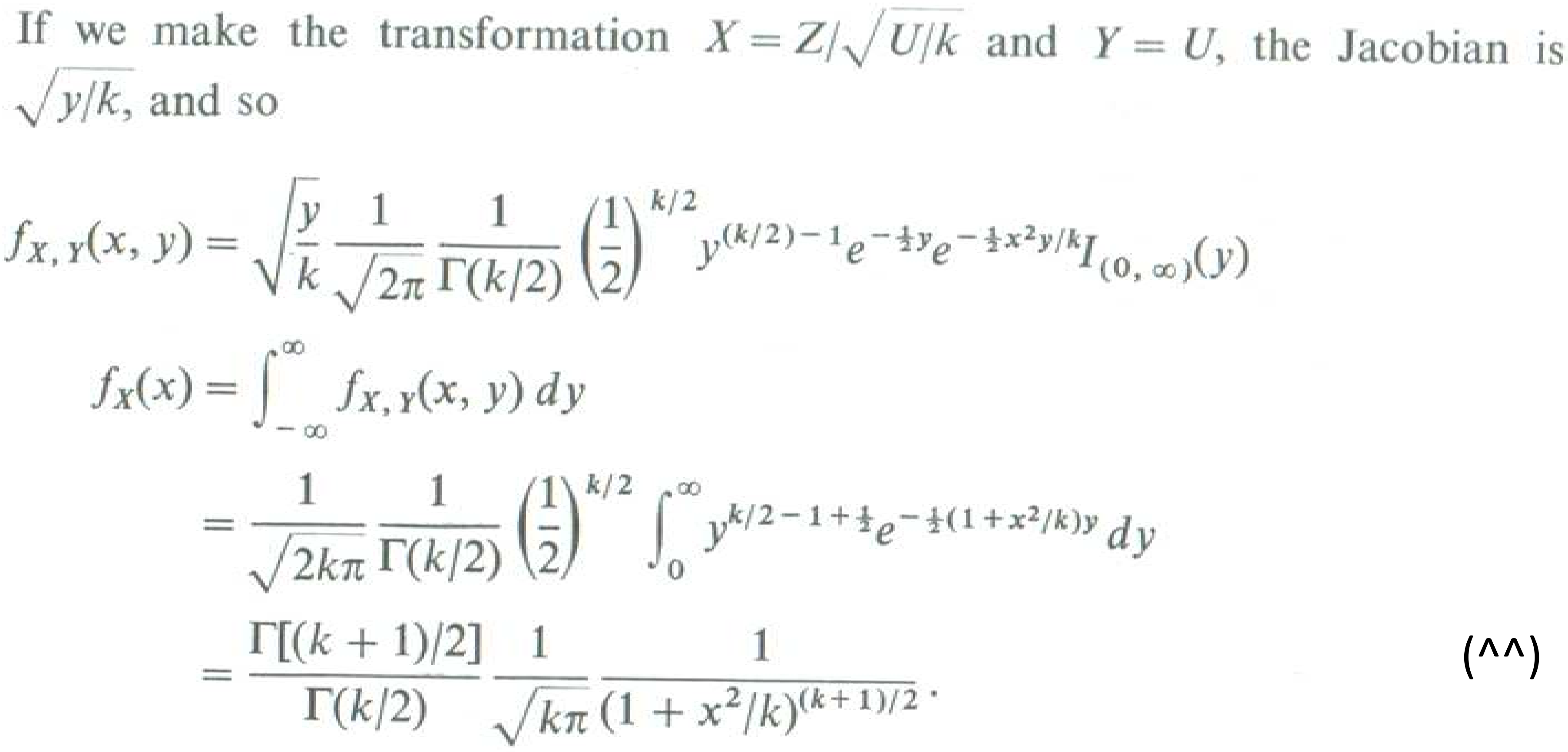


Another distribution of considerable practical importance is that of the ratio of a standard normally distributed random variable to the square root of an independently distributed chi-square random variable divided by its degrees of freedom. That is, if Z has a standard normal distribution, if U has a chi-square distribution with k degrees of freedom, and if Z and U are independent, we seek the distribution of

$$X = \frac{Z}{\sqrt{U/k}}.$$

The joint density of Z and U is given by

$$f_{Z,U}(z, u) = \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(k/2)} \left(\frac{1}{2}\right)^{k/2} u^{(k/2)-1} e^{-\frac{1}{2}u} e^{-\frac{1}{2}z^2} I_{(0,\infty)}(u).$$

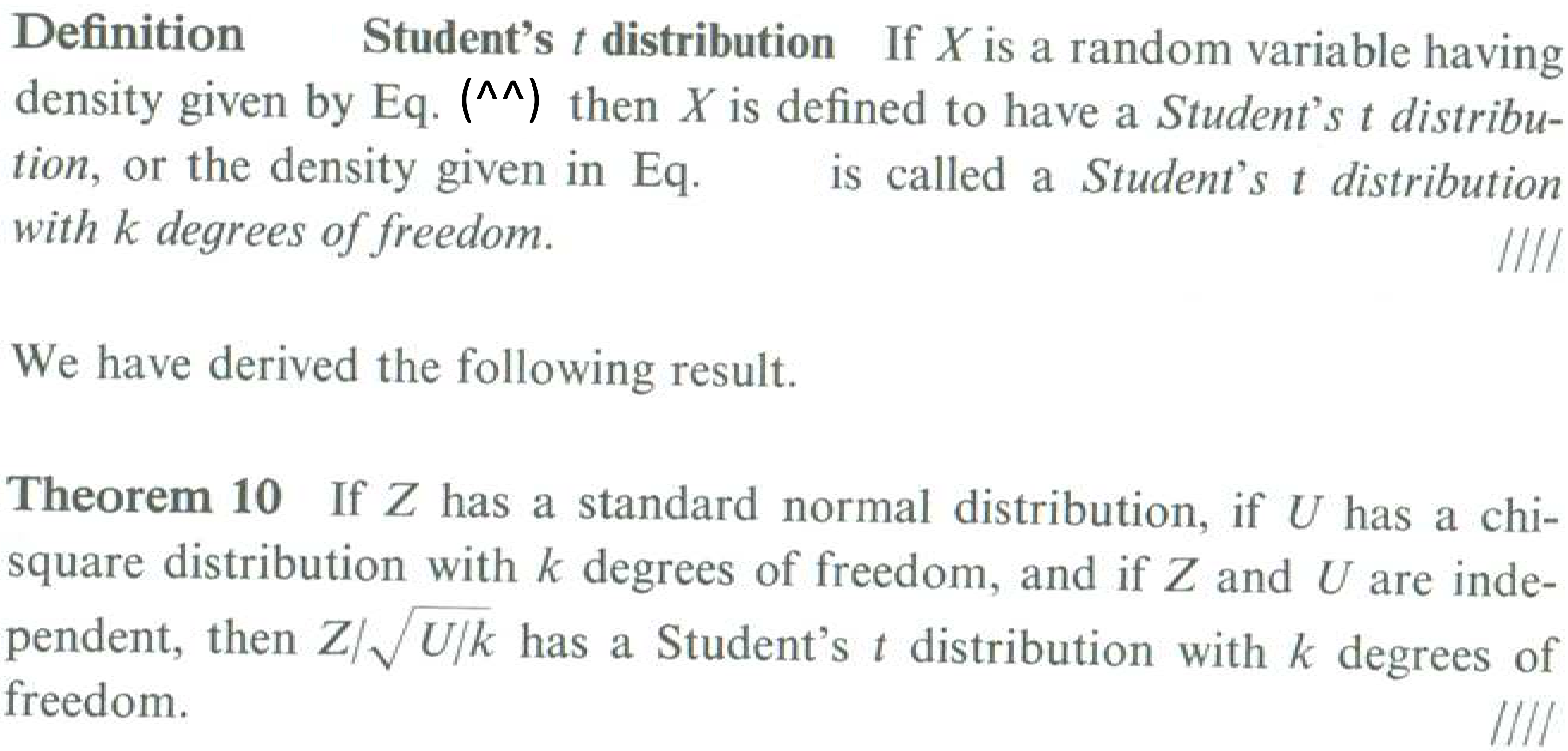


If we make the transformation $X = Z/\sqrt{U/k}$ and $Y = U$, the Jacobian is $\sqrt{y/k}$, and so

$$f_{X,Y}(x, y) = \sqrt{\frac{y}{k}}\,\frac{1}{\sqrt{2\pi}}\,\frac{1}{\Gamma(k/2)}\left(\frac{1}{2}\right)^{k/2} y^{(k/2)-1} e^{-\frac{1}{2}y} e^{-\frac{1}{2}x^2 y/k} I_{(0,\infty)}(y)$$

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y)\,dy \\ &= \frac{1}{\sqrt{2k\pi}}\,\frac{1}{\Gamma(k/2)}\left(\frac{1}{2}\right)^{k/2} \int_0^{\infty} y^{k/2-1+\frac{1}{2}} e^{-\frac{1}{2}(1+x^2/k)y}\,dy \\ &= \frac{\Gamma[(k+1)/2]}{\Gamma(k/2)}\,\frac{1}{\sqrt{k\pi}}\,\frac{1}{(1+x^2/k)^{(k+1)/2}}. \end{aligned}$$

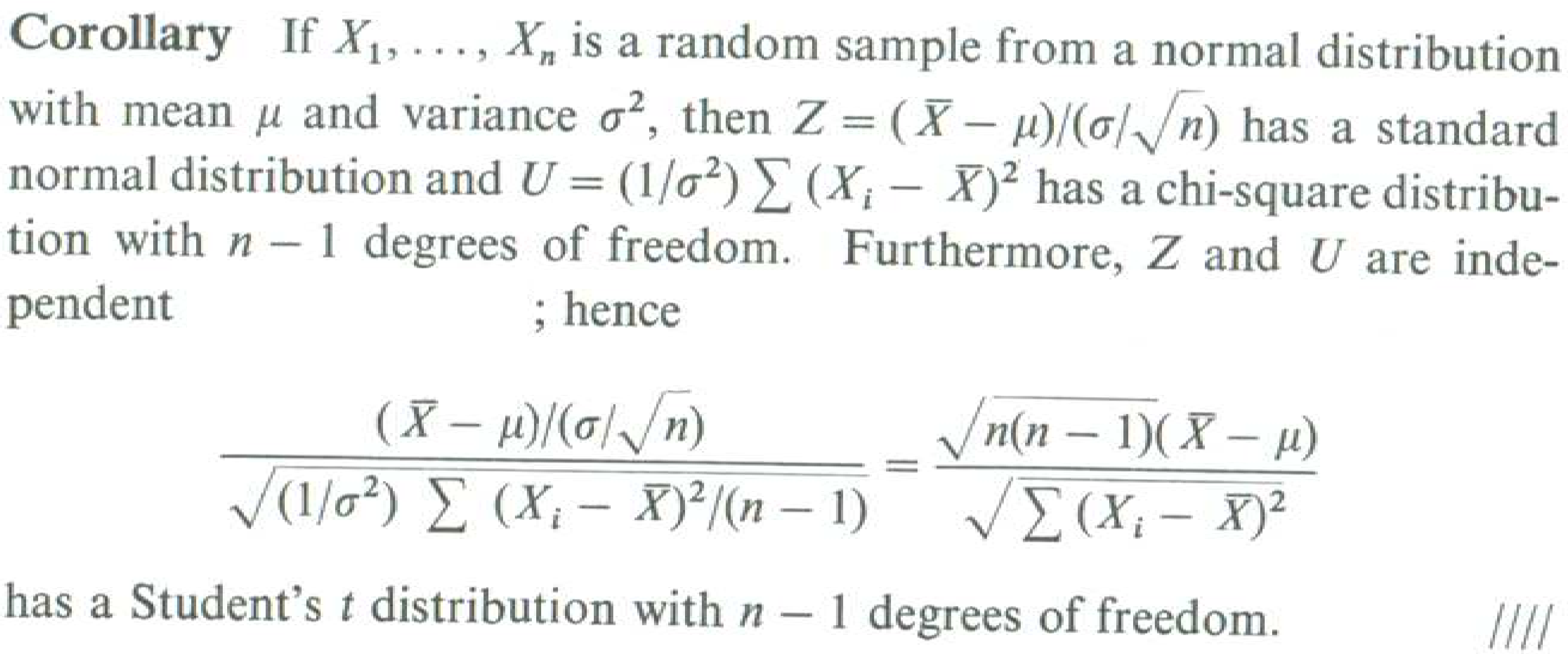
(^^)



Definition **Student's *t* distribution** If X is a random variable having density given by Eq. (^^) then X is defined to have a *Student's t distribution*, or the density given in Eq. is called a *Student's t distribution with k degrees of freedom.* ////

We have derived the following result.

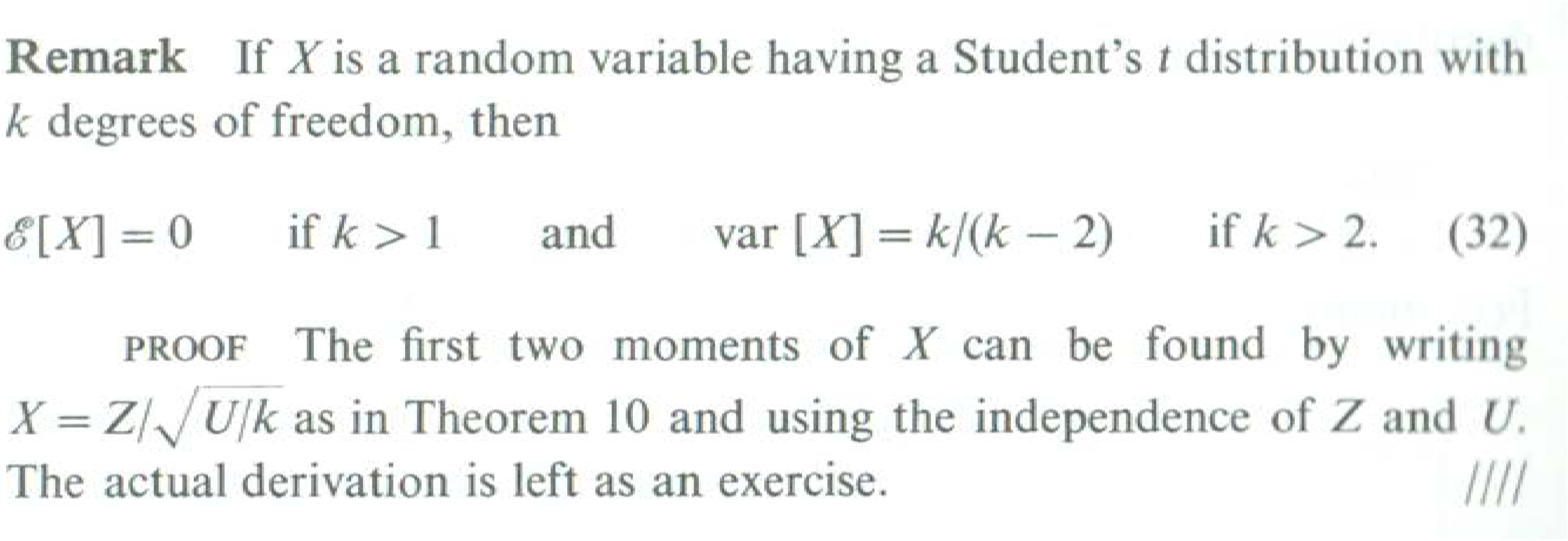
Theorem 10 If Z has a standard normal distribution, if U has a chi-square distribution with k degrees of freedom, and if Z and U are independent, then $Z/\sqrt{U/k}$ has a Student's t distribution with k degrees of freedom. ////



Corollary If $X_1, \ldots, X_n$ is a random sample from a normal distribution with mean μ and variance σ^2, then $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ has a standard normal distribution and $U = (1/\sigma^2) \sum (X_i - \bar{X})^2$ has a chi-square distribution with $n-1$ degrees of freedom. Furthermore, Z and U are independent ; hence

$$\frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{(1/\sigma^2) \sum (X_i - \bar{X})^2/(n-1)}} = \frac{\sqrt{n(n-1)}(\bar{X} - \mu)}{\sqrt{\sum (X_i - \bar{X})^2}}$$

has a Student's t distribution with $n-1$ degrees of freedom. ////



Remark If X is a random variable having a Student's t distribution with k degrees of freedom, then

$$\mathscr{E}[X] = 0 \qquad \text{if } k > 1 \qquad \text{and} \qquad \operatorname{var}[X] = k/(k-2) \qquad \text{if } k > 2. \tag{32}$$

PROOF The first two moments of X can be found by writing $X = Z/\sqrt{U/k}$ as in Theorem 10 and using the independence of Z and U. The actual derivation is left as an exercise. ////