Probability & Statistics

Dongjiao Ge

djge@cityu.edu.mo

Session 08

Chapter 5
Inequalities & Sequences of random variables

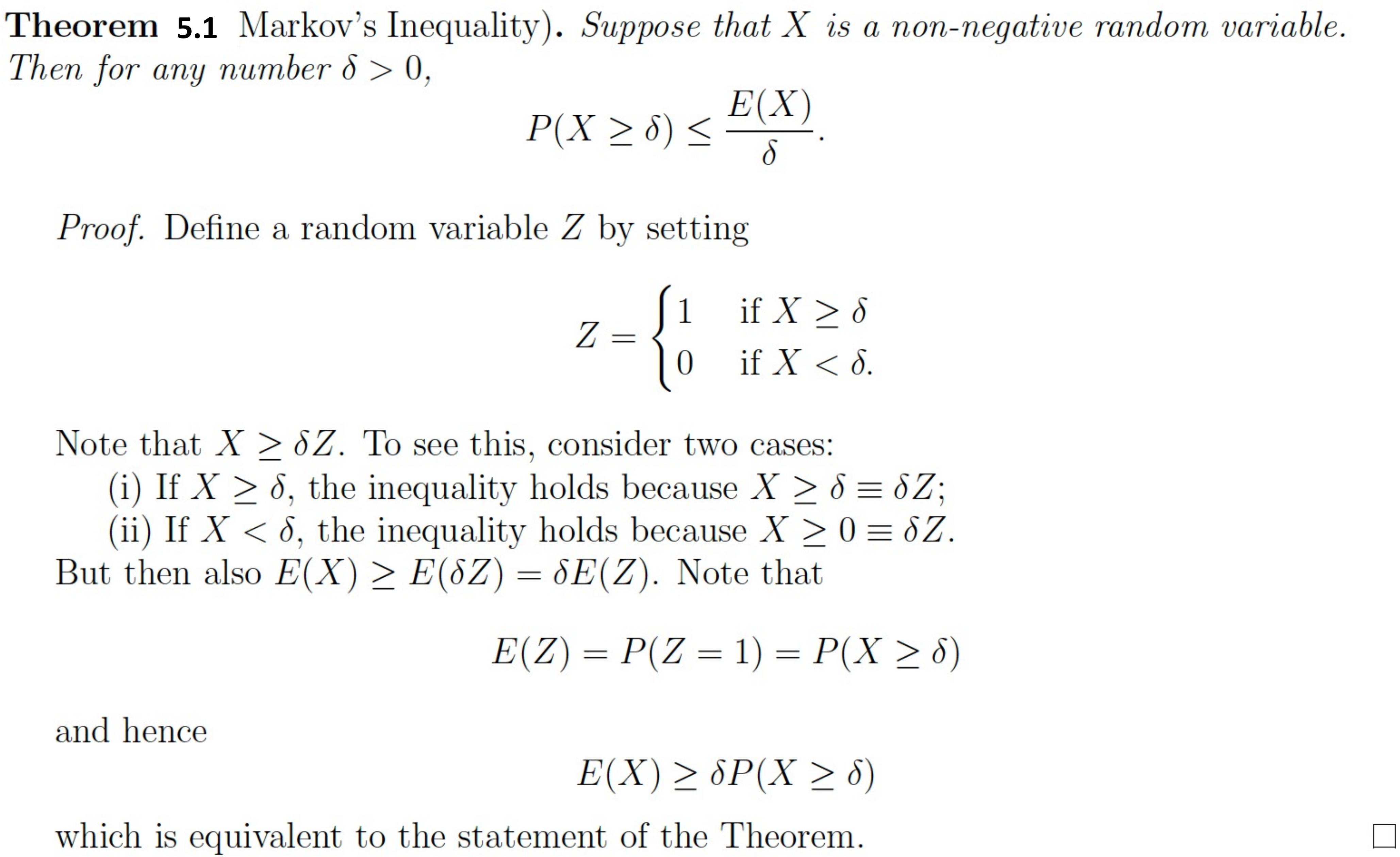
In many cases we cannot do the computations necessary to compute the probabilities, expectations, etc, that we are interested in.

In these cases, it can be very useful to have a rough idea of the size of these things, even if we cannot get a precise answer.

5.1 Inequalities

Example

Suppose that we know the expectation of some random variable is small. We cannot in general say that it is unlikely to be large. E.g., it could be a billion with probability 1/2 and minus a billion with probability 1/2. Then its expectation is zero, but half of the times, it is extremely large. However, we can say something if the random variable is non-negative.



Theorem 5.1 Markov's Inequality). *Suppose that X is a non-negative random variable. Then for any number $\delta > 0$,*

$$P(X \geq \delta) \leq \frac{E(X)}{\delta}.$$

Proof. Define a random variable Z by setting

$$Z = \begin{cases} 1 & \text{if } X \geq \delta \\ 0 & \text{if } X < \delta. \end{cases}$$

Note that $X \geq \delta Z$. To see this, consider two cases:
(i) If $X \geq \delta$, the inequality holds because $X \geq \delta \equiv \delta Z$;
(ii) If $X < \delta$, the inequality holds because $X \geq 0 \equiv \delta Z$.
But then also $E(X) \geq E(\delta Z) = \delta E(Z)$. Note that

$$E(Z) = P(Z = 1) = P(X \geq \delta)$$

and hence

$$E(X) \geq \delta P(X \geq \delta)$$

which is equivalent to the statement of the Theorem. □

Example. *When a person wants to pass a driving test, they need on average 2.5 attempts. Prove that the chance they need 10 or more attempts is at most* $1/4$*.*

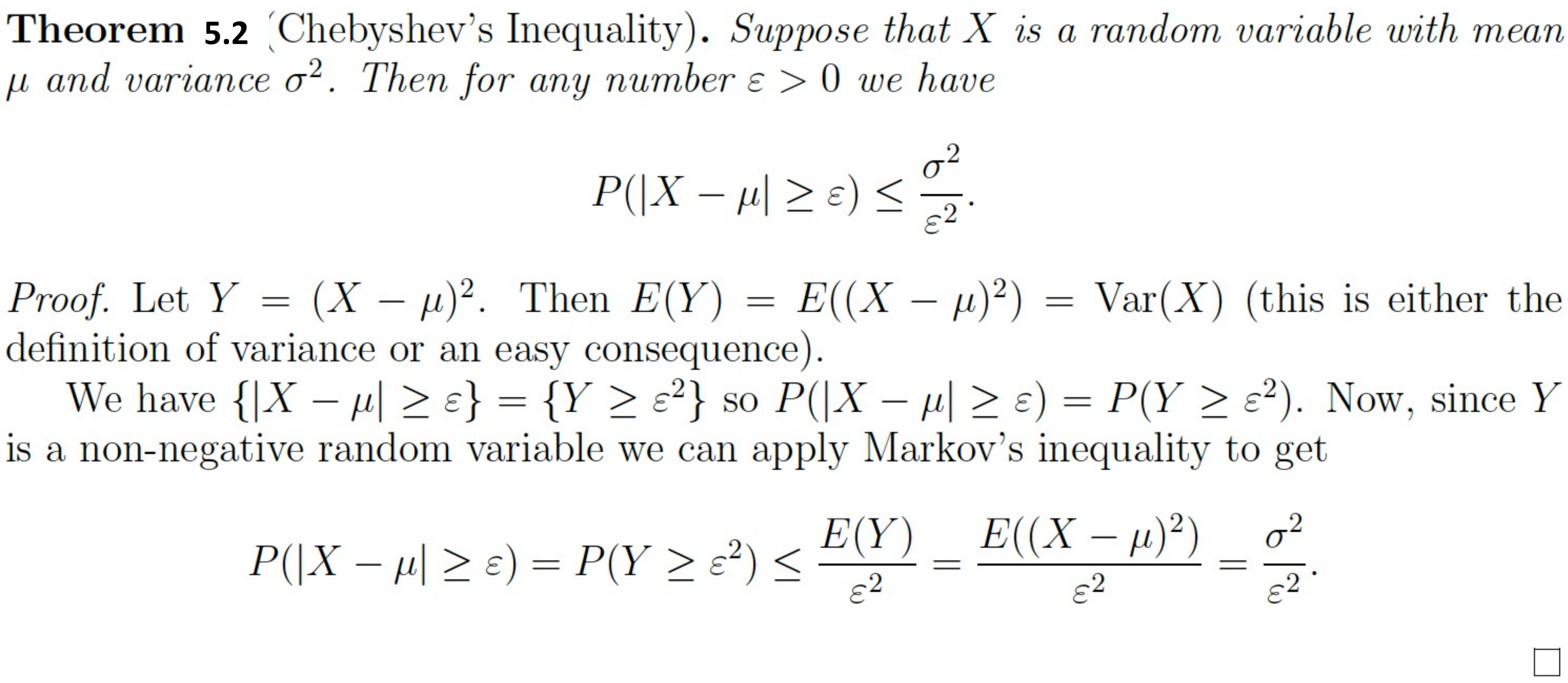
Answer. Let X be the number of attempts it takes to pass. We are told that $E(X) = 2.5$. Hence by Markov's Inequality

$$P(X \geq 10) \leq \frac{E(X)}{10} = \frac{2.5}{10} = \frac{1}{4}$$

□

Q1. In the same situation as above example, what does Markov's Inequality tell you about the probability it takes 2 or more attempts? **(10 minutes)**

Note the Markov's inequality is **only a bound** - it says that the probability is not more than something. It might actually be much smaller than this bound.



Theorem 5.2 (Chebyshev's Inequality). *Suppose that X is a random variable with mean μ and variance σ^2. Then for any number $\varepsilon > 0$ we have*

$$P(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

Proof. Let $Y = (X - \mu)^2$. Then $E(Y) = E((X - \mu)^2) = \mathrm{Var}(X)$ (this is either the definition of variance or an easy consequence).

We have $\{|X - \mu| \geq \varepsilon\} = \{Y \geq \varepsilon^2\}$ so $P(|X - \mu| \geq \varepsilon) = P(Y \geq \varepsilon^2)$. Now, since Y is a non-negative random variable we can apply Markov's inequality to get

$$P(|X - \mu| \geq \varepsilon) = P(Y \geq \varepsilon^2) \leq \frac{E(Y)}{\varepsilon^2} = \frac{E((X - \mu)^2)}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}.$$

□

Example. *Suppose that when you sit an exam, your expected mark is 50 and standard deviation is 10. Show that the probability that you get a first class mark (70 or more) is at most* 1/4.

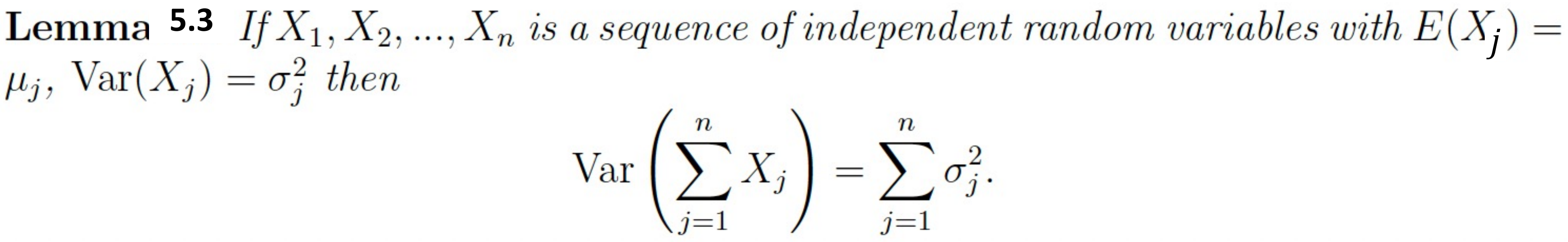
Answer. Let X be your mark. We know that X has mean 50 and variance 100. Hence by Chebychev's Inequality

$$P(X \geq 70) \leq P(|X - 50| \geq 20) = P(|X - E(X)| \geq 20) \leq \frac{100}{20^2} = \frac{1}{4}.$$

不等号两边同时减去均值

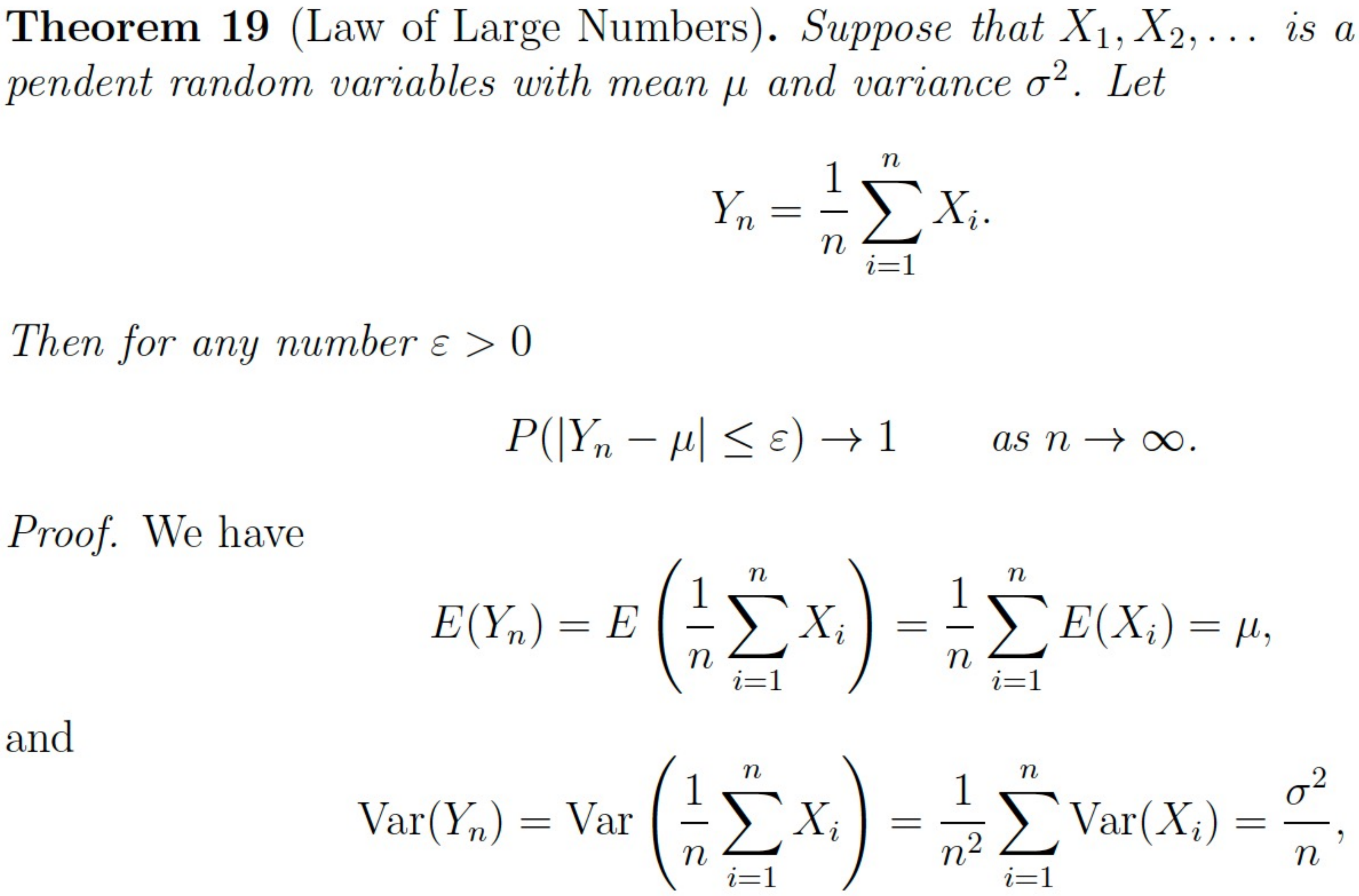
□

5.2 The Law of Large Numbers (LLN)



Lemma 5.3 *If* $X_1, X_2, \ldots, X_n$ *is a sequence of independent random variables with* $E(X_j) = \mu_j$, $\mathrm{Var}(X_j) = \sigma_j^2$ *then*

$$\mathrm{Var}\left(\sum_{j=1}^{n} X_j\right) = \sum_{j=1}^{n} \sigma_j^2 .$$



Theorem 19 (Law of Large Numbers). *Suppose that $X_1, X_2, \ldots$ is a sequence of independent random variables with mean μ and variance σ^2. Let*

$$Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then for any number $\varepsilon > 0$

$$P(|Y_n - \mu| \le \varepsilon) \to 1 \qquad \text{as } n \to \infty.$$

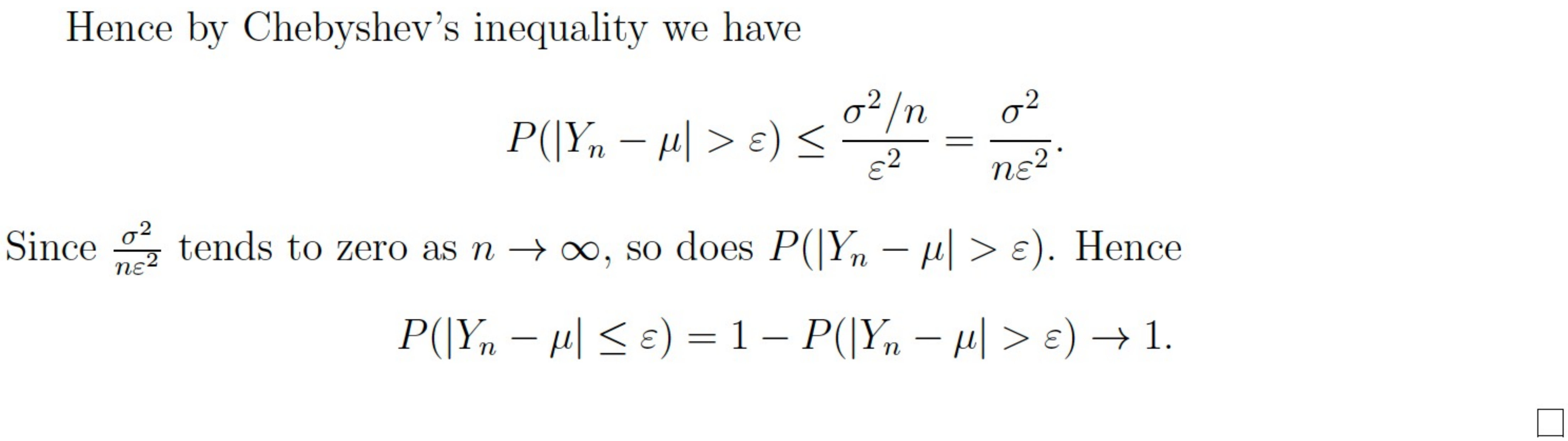
Proof. We have

$$E(Y_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu,$$

and

$$\mathrm{Var}(Y_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{\sigma^2}{n},$$

where we use two properties of variance: $Var(cZ) = c^2 Var(Z)$ and Lemma 5.3.



Hence by Chebyshev's inequality we have

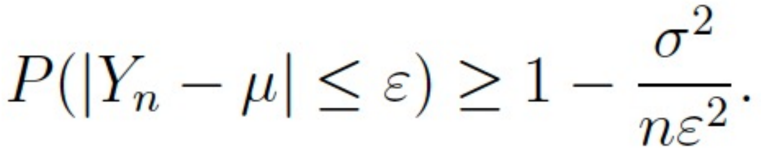
$$P(|Y_n - \mu| > \varepsilon) \leq \frac{\sigma^2/n}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

Since $\frac{\sigma^2}{n\varepsilon^2}$ tends to zero as $n \to \infty$, so does $P(|Y_n - \mu| > \varepsilon)$. Hence

$$P(|Y_n - \mu| \leq \varepsilon) = 1 - P(|Y_n - \mu| > \varepsilon) \to 1.$$

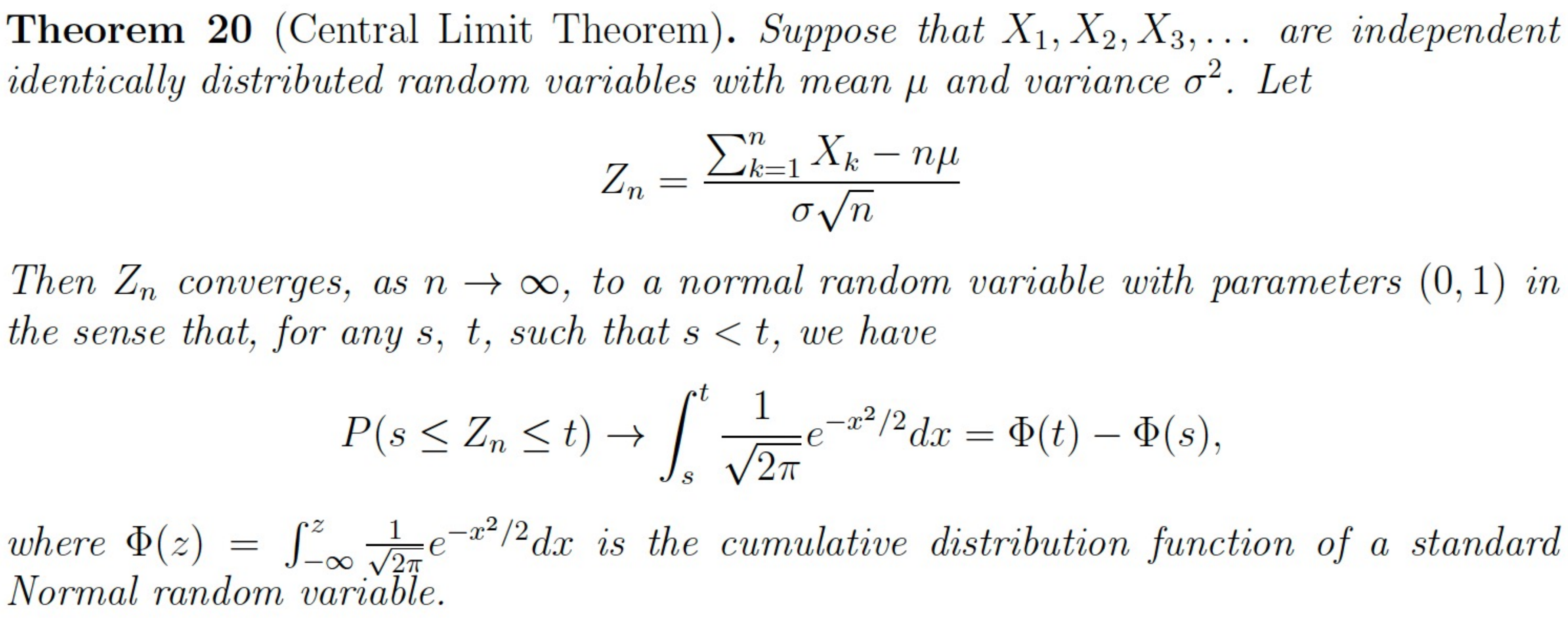
□

Remark. Theorem 19 is also called the weak LLN. It basically says that for some specified "large" n, the average Y_n of the $(X_1, \ldots, X_n)$ is likely to be close to the mean μ. In fact, we can repeat the arguments of the above proof, to also proved a useful estimate:



$$P(|Y_n - \mu| \leq \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}.$$

5.3 Central Limit Theorem



Theorem 20 (Central Limit Theorem). *Suppose that* $X_1, X_2, X_3, \ldots$ *are independent identically distributed random variables with mean* μ *and variance* σ^2. *Let*

$$Z_n = \frac{\sum_{k=1}^{n} X_k - n\mu}{\sigma\sqrt{n}}$$

Then Z_n *converges, as* $n \to \infty$, *to a normal random variable with parameters* $(0,1)$ *in the sense that, for any* s, t, *such that* $s < t$, *we have*

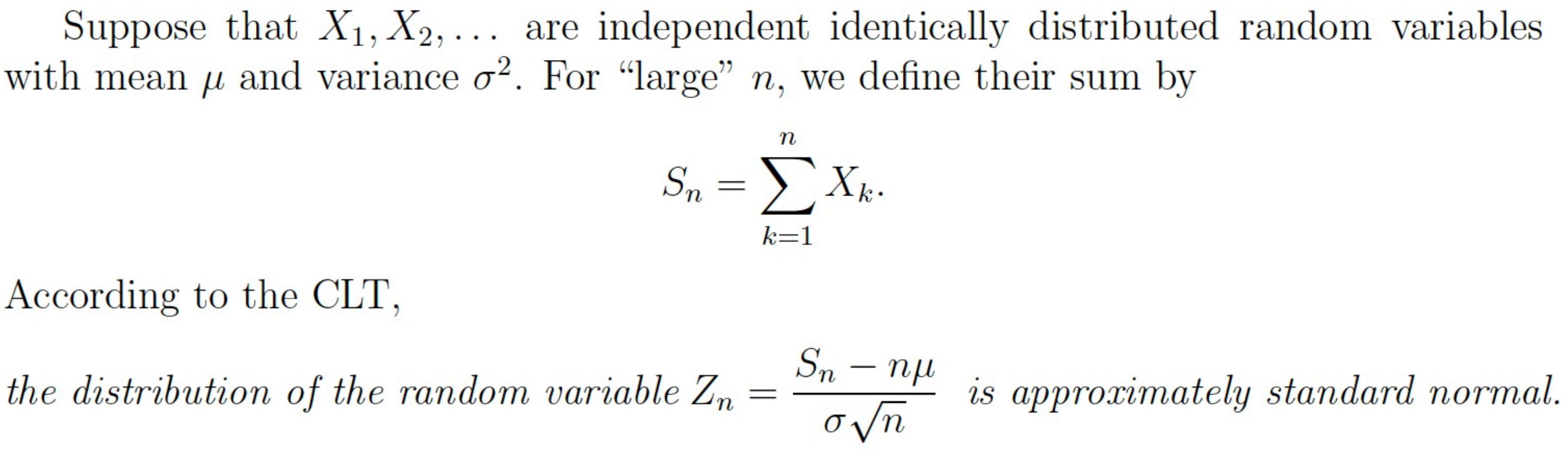
$$P(s \le Z_n \le t) \to \int_s^t \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \Phi(t) - \Phi(s),$$

where $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$ *is the cumulative distribution function of a standard Normal random variable.*

The proof is not required.

Remarks

- The Central Limit Theorem (CLT) only tells you about what happens as $n \to \infty$.
- In Statistics, CLT is commonly (and very conveniently) used for finite but large values of n.



Suppose that $X_1, X_2, \ldots$ are independent identically distributed random variables with mean μ and variance σ^2. For "large" n, we define their sum by

$$S_n = \sum_{k=1}^{n} X_k.$$

According to the CLT,

the distribution of the random variable $Z_n = \dfrac{S_n - n\mu}{\sigma\sqrt{n}}$ *is approximately standard normal.*

Using this we can also characterise the distribution of the average Y_n as n increases, which is a very useful result in statistical applications. Namely, we get that

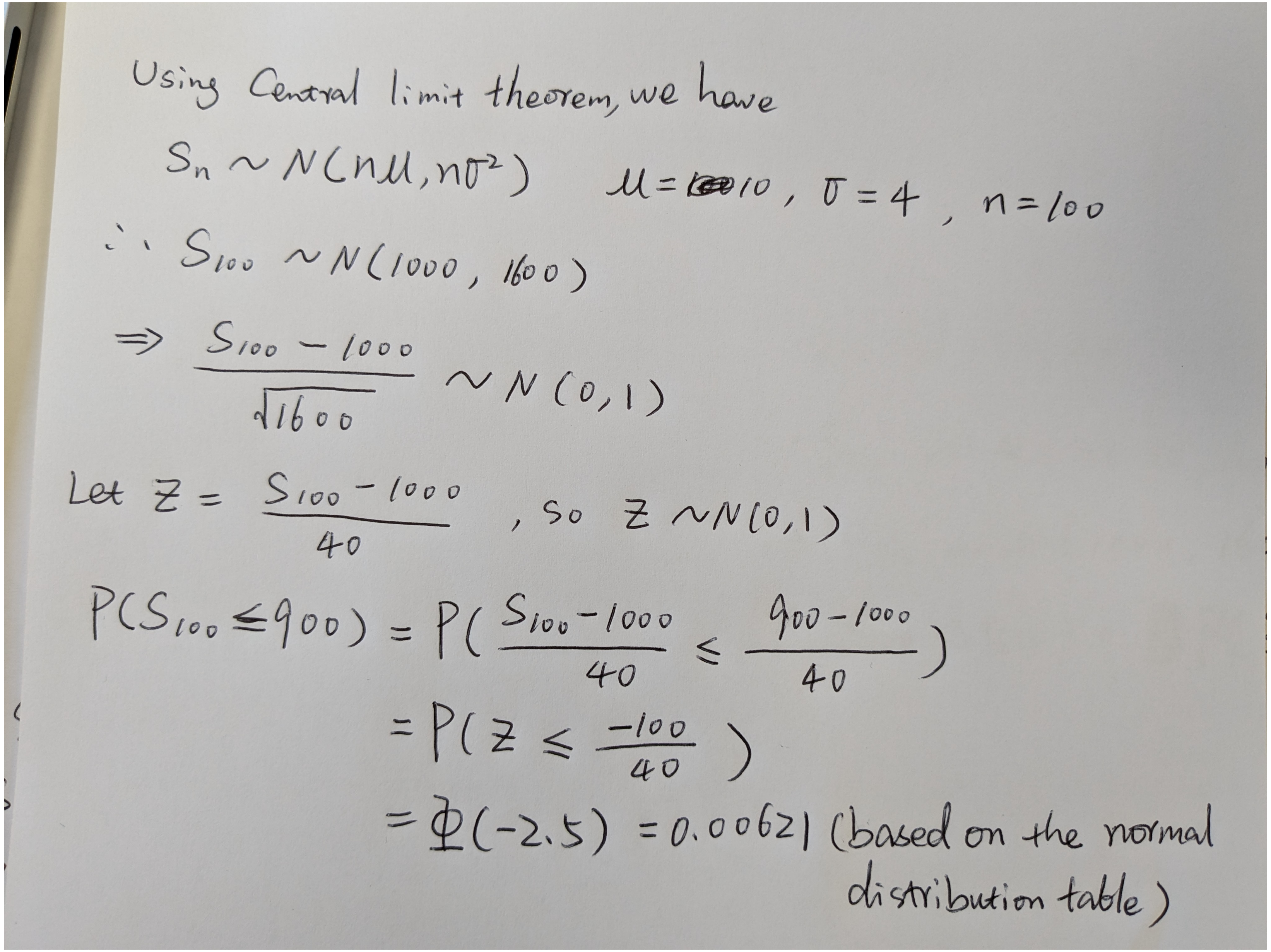
$$\textit{the distribution of the average } Y_n := \frac{S_n}{n} \textit{ is approximately } N(\mu, \frac{\sigma^2}{n}).$$

This result justifies also the extensive use of Normal distributions in real-life applications to model data resulting from many different independent factors (roughly independent) or when the distribution of data is unknown.

Finally, we have

$$\textit{the distribution of the random variable } S_n \ \textit{ is approximately } N(n\mu, n\sigma^2).$$

Example Let X_i be i.i.d random variables with $\mu = 10$ and $\sigma = 4$. Let $S_n = X_1 + \cdots + X_n$, Please compute the probability $P(S_{100} \leq 900)$.



Using Central limit theorem, we have

$S_n \sim N(n\mu, n\sigma^2)$ $\quad \mu = 10, \ \sigma = 4, \ n = 100$

$\therefore S_{100} \sim N(1000, 1600)$

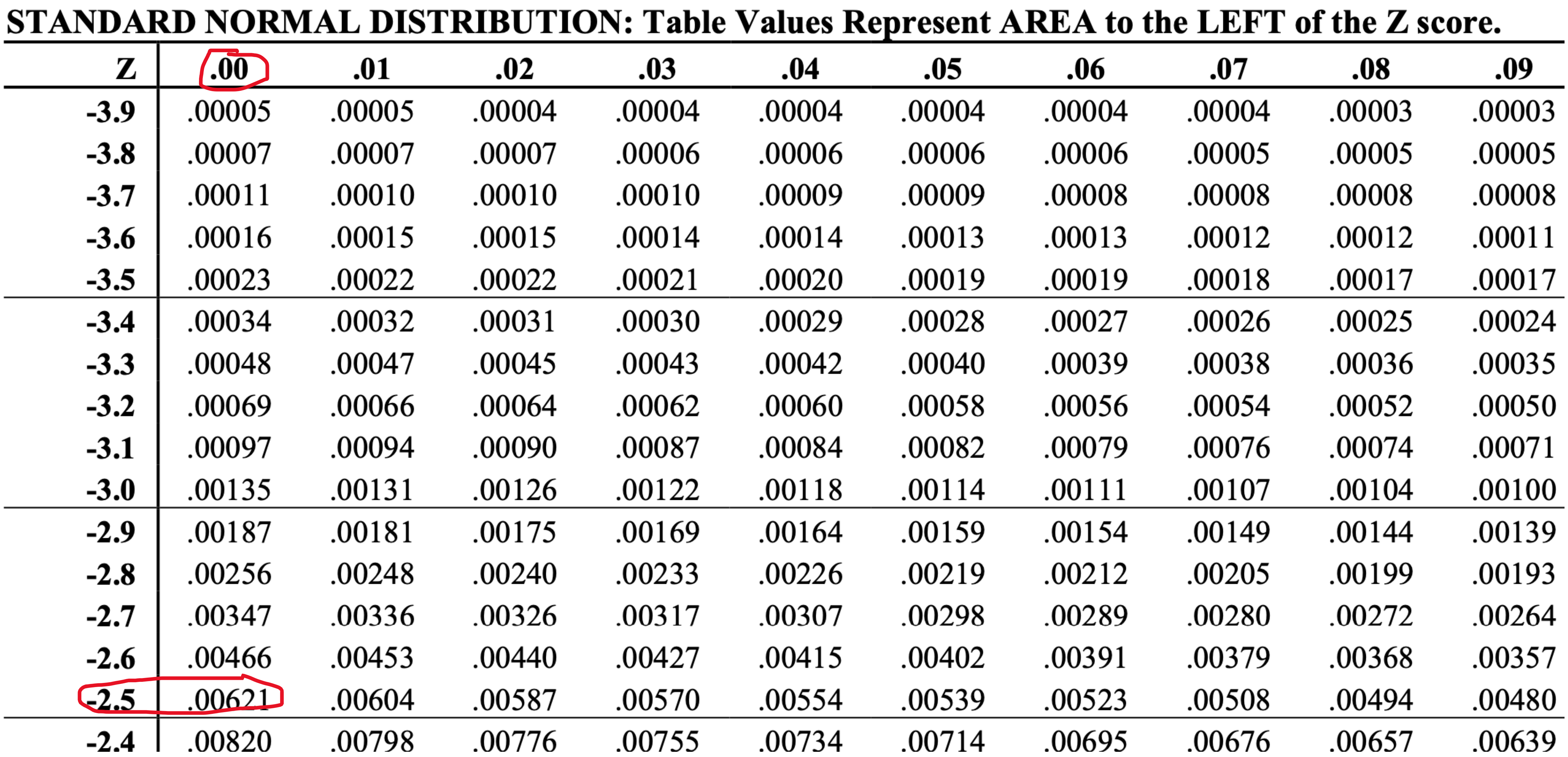
$$\Rightarrow \frac{S_{100} - 1000}{\sqrt{1600}} \sim N(0,1)$$

Let $Z = \frac{S_{100} - 1000}{40}$, so $Z \sim N(0,1)$

$$P(S_{100} \leq 900) = P\left(\frac{S_{100} - 1000}{40} \leq \frac{900 - 1000}{40}\right)$$

$$= P\left(Z \leq \frac{-100}{40}\right)$$

$= \Phi(-2.5) = 0.00621$ (based on the normal distribution table)



STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639

Part 2
Statistics

Reading list:
https://minerva.it.manchester.ac.uk/~saralees/lecturenotes.pdf
Our slides for this part are made mainly based on the above lecture notes.

Chapter 0
Introduction to statistics

1 Introduction: What is Statistics?

Statistics is:
‘the science of learning from data, and of measuring, controlling, and communicating uncertainty; and it thereby provides the navigation essential for controlling the course of scientific and societal advances.’

---------- Davidian, M. and Louis, T.A. (2012), Science.

1 Introduction: What is Statistics?

Statistics

The art and science of answering questions and exploring ideas through the processes of gathering data, describing data, and making generalizations about a population on the basis of a smaller sample.

1 Introduction: What is Statistics?

There are two basic forms: descriptive statistics and inferential statistics.

- **Descriptive Statistics** is primarily about summarizing a given data set through numerical summaries and graphs, and can be used for exploratory analysis to visualize the information contained in the data and suggest hypotheses etc.

- **Inferential Statistics** is concerned with methods for making conclusions about a population using information from a sample, and assessing the reliability of, and uncertainty in, these conclusions.

2 Populations and samples

A **population** is the collection of all individuals or items under consideration in the study.

For a given population there will typically be one or more variables in which we are interested. For example, consider the following populations together with corresponding variables of interest:

1. Car batteries of a particular type manufactured by a particular company; the variable of interest is the lifetime of the battery before failure.
2. All potential possible outcomes of a planned laboratory experiment; the variable of interest is the value of a particular measurement.

2 Populations and samples

In general, the variables of interest may be either **qualitative or quantitative**.

- **Qualitative variables** are either nominal, e.g. gender or political party supported, or ordinal, e.g. a measurement of size grouped into three categories: small, medium or large.
- **Quantitative variables** are either discrete, for example, a count, or continuous, such as the variables income and lifetime above.

We wish to make conclusions, or inferences, about the population characteristics of variables of interest.

2 Populations and samples

One way to do so is to **conduct a census**, i.e. to collect data for **each individual** in the population.

However often this is **not feasible**, due to one or more of the following:

- It may be too expensive or time-consuming to do so
- Testing may be destructive
- The population may be purely conceptual

Instead, we collect data only for a sample, i.e. a subset of the population.

We then use the characteristics of the sample to estimate the characteristics of the population. In order for this procedure to give a good estimate, **the sample must be representative of the population**. Otherwise, if an unrepresentative or ‘biased’ sample is used the conclusions will be systematically incorrect.

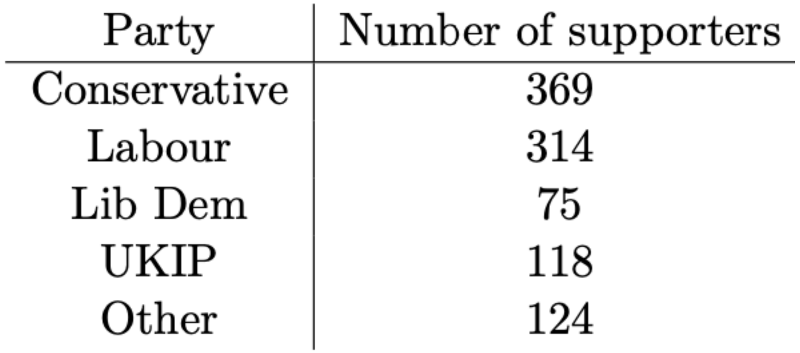
2 Populations and samples

Some examples of samples from populations:

1. A random sample of 40 manufactured car batteries was taken from the production line, and their lifetimes (in years) were determined. The data are as follows, arranged in ascending order for convenience:

   1.6, 1.9, 2.2, 2.5, 2.6, 2.6, 2.9, 3.0, 3.0, 3.1, 3.1, 3.1, 3.1, 3.2, 3.2, 3.2, 3.3, 3.3, 3.3, 3.4, 3.4, 3.4, 3.5, 3.5, 3.6, 3.7, 3.7, 3.7, 3.8, 3.8, 3.9, 3.9, 4.1, 4.1, 4.2, 4.3, 4.4, 4.5, 4.7, 4.7

2. In an opinion poll in May 2015, a sample of 1000 adults was obtained and asked which political party they intended to vote for in the upcoming UK General Election on 7 May 2015. A summary of these responses is:



Party	Number of supporters
Conservative	369
Labour	314
Lib Dem	75
UKIP	118
Other	124

2.1 Finite population sampling

In modern Statistics, the most common way of guaranteeing representativeness is to use a random sample of size n chosen according to a probabilistic sampling rule. This probabilistic sampling is objective and eliminates investigator bias. For a population of finite size N, the most common method is to use simple random sampling.

This takes two main forms: **sampling without replacement** and **sampling with replacement**.

2.1 Finite population sampling

- **Sampling without replacement**: each of the $\binom{N}{n}$ possible samples of n distinct individuals from the population has equal probability of selection, $\binom{N}{n}^{-1}$. No individual appears more than once in the sample.

  This can be implemented by choosing individuals sequentially, one at a time, as follows. For $i = 1, \ldots, n$:

  Step 1. Select an individual at random with equal probability from the remaining population of size $N - i + 1$
  Step 2. Include the selected individual as the i-th member of the sample, and remove the selected individual from the population, leaving $N - i$ individuals remaining.

  The above steps are repeated until a sample of size n is obtained.

- **Sampling with replacement:** each individual may appear any number of times in the sample, leading to N^n possible samples. The probability of selecting any particular sample is N^{-n}. This can be implemented using a similar sequential algorithm to before, where instead in Step 2 the selected individual is not removed from the population.

2.2 Sampling from a general population

The idea of independence is used to define sampling from a general population.

- **Sampling without replacement,** the X_i are **not independent**.
- However, the X_i can be considered to be ***approximately independent***

In the remaining sessions of this module, *we will always assume that* $X_1, \ldots, X_n$ *are* ***sampled independently*** *from a c.d.f.* $F_X\ (x)$*.*

2.2 Sampling from a general population

Random Samples

- We say that $X_1, \ldots, X_n$ are a random sample from X if $X_1, \ldots, X_n \sim F_X(x)$ independently.
- We may also say that $X_1, \ldots, X_n$ is a random sample from $F_X(x), f_X(x)$ or $p_X(x)$.

($F_X(x)$ --- cdf, $f_X(x)$ --- pdf, $p_X(x)$ --- pmf)

3 Probability models for data

Let $x_1, \ldots, x_n$ be the observed values in a particular random sample of the random variable X, whose distribution is unknown.

We want to use these data to estimate the probability of an event $\{X \ \in \ A\}$

- One way is to use the **empirical probability** of the event, in other words the proportion of the sample values that lie in A,

$$\widehat{\mathrm{P}}(X \in A) = \frac{\#\{i : x_i \in A\}}{n} \, .$$

3 Probability models for data

- An alternative approach is to assume that the data were generated as a random sample from a particular parametric probability model, e.g. $N(\mu, \sigma^2)$.
    - Such models usually contain *unknown parameters*, e.g. in the previous example the parameters μ and σ^2 are unknown.
    - We can *use the sample to estimate the parameters of the distribution*, thereby fitting the model to the data.
    - A fitted model can be used to calculate probabilities of events of interest.

Remark. *In practice, you never know the real distribution of the data. No "right" or "wrong" for the fit, there is only "good"(close to real) or "bad" (huge difference from the real).*

4 Sampling distributions of sample statistics

Let $X_1, \ldots, X_n$ be a random sample from a distribution $F_X(x)$. A **statistic** is a function of the data, $h(X_1, \ldots, X_n)$.

- The value of this statistic will usually be different for different samples.
- As the sample data is random, the statistic is also a random variable.
- If we repeatedly drew samples of size n, calculating and recording the value of the sample statistic each time, then we would build up its probability distribution.

The probability distribution of a sample statistic is referred to as its sampling distribution.

4.1 Sample mean

The random variables $X_1, \dots, X_n$ are assumed to be independent and identically distributed (often abbreviated to i.i.d.) random variables, each being distributed as $F_X(x)$.

This means that $E(X_i) = \mu$ for $i = 1, \dots, n$ and $Var(X_i) = \sigma^2$ for $i = 1, \ldots, n$.

The sample mean of the n sample variables is:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

4.1 Sample mean

The mean and the variance of the sampling (probability) distribution of $\bar{X}$ are computed as follows:

$$
\begin{aligned}
\mathrm{E}(\overline{X}) &= \mathrm{E}\left[\frac{1}{n}\left(X_1 + \ldots + X_n\right)\right] \\
&= \frac{1}{n}\left[\mathrm{E}(X_1) + \ldots + \mathrm{E}(X_n)\right] \\
&= \frac{n\mu}{n} = \mu\,,
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(\overline{X}) &= \mathrm{Var}\left[\frac{1}{n}\left(X_1 + \ldots + X_n\right)\right] \\
&= \frac{1}{n^2}\left[\mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n)\right] \\
&= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}\,.
\end{aligned}
$$

These results tell us that the sampling distribution of the sample mean $\bar{X}$ is centered on the common mean μ of each of the sample variables $X_1, \ldots, X_n$ (i.e. the mean of the distribution from which the sample is obtained) and has variance equal to the common variance of the X_i divided by n. Thus, as the sample size n increases, the sampling distribution of $\bar{X}$ becomes more concentrated around the true mean μ.

4.1 Sample mean

In the above:

- No actual distribution from which the X_i have been sampled is given.
- All we are assuming is that the mean and variance of the underlying distribution are both finite.

4.1.1 Normally distributed data

Let the random variable $X \sim N(\mu_X, \sigma_X^2)$ and let the random variable $Y \sim N(\mu_Y, \sigma_Y^2)$, **independently** of X. Then we have the following results:

(i) $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$

(ii) $X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$

(iii) In general, $c_1 X + c_2 Y \sim N(c_1\mu_X + c_2\mu_Y, c_1^2\sigma_X^2 + c_2^2\sigma_Y^2)$; $c_1 \neq 0$, $c_2 \neq 0$.

4.1 Sample mean

4.1.1 Normally distributed data

These results can extend to the linear combination of n independent normal random variables.

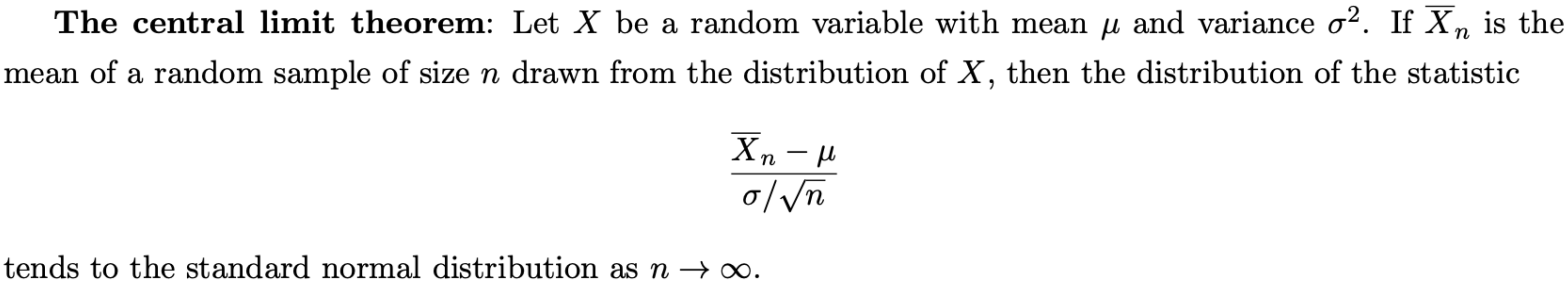
Let $X_1, ... X_n$ be n ***independent normally distributed random variables*** with $E(X_i) = \mu_i$ and $Var(X_i) = \sigma_i^2$ for $i = 1, ..., n$. Thus, here the normal distributions for different X_i may have different means and variances. We then have that

$$\sum_{i=1}^{n} c_i X_i \sim N\left(\sum_{i=1}^{n} c_i \mu_i, \sum_{i=1}^{n} c_i^2 \sigma_i^2\right)$$

where the $c_i \in \boldsymbol{R}$.

4.1 Sample mean

4.1.2 Using the central limit theorem



The central limit theorem: Let X be a random variable with mean μ and variance σ^2. If $\overline{X}_n$ is the mean of a random sample of size n drawn from the distribution of X, then the distribution of the statistic

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

tends to the standard normal distribution as $n \to \infty$.

This means that for **a large random sample** from a population with mean μ and variance σ^2:

- The sample mean $\bar{X}_n$ is **approximately normally distributed** with mean μ and variance $\frac{\sigma^2}{n}$.
- Since, for large n, $\bar{X}_n \sim N(\mu, \frac{\sigma^n}{n})$ approximately we have that $\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$ approximately.

Exercises:

Read the pages 1 to 12 of the following notes
https://minerva.it.manchester.ac.uk/~saralees/lecturenotes.pdf