IEE 380 Midterm 1 Review

Dr. Clough - Chapters 2, 3, 4, 6, 7

UGTAs: Kyle Otstot, Miguel Garcia, Jessica Huber

Chapter 2: Probability

Main Ideas
Chapter 2

- Sample Spaces & Events in a Random Experiment
  - <u>Sample space (S) :</u> the set of all possible outcomes in a random experiment
    - S is **discrete** if it contains a (1) finite; or (2) countably infinite set of outcomes
    - S is **continuous** otherwise (containing an interval of real numbers)
  - <u>Event (E):</u> a subset of the sample space of a random experiment
    - Know the following event operations:
      - **union** ($E_1 \cup E_2$) ; **intersection** ($E_1 \cap E_2$) ; **complement** ($\bar{E}$ or E^c)

Main Ideas
Chapter 2

- Probability of an Event
  - <u>Probability ($[0,1]$):</u> the likelihood that a particular outcome will occur
    - If E is an event, then $P(E) = \sum_{x \in E} P(x)$ *(the sum over all outcomes in E)*
  - <u>Addition rule:</u> $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - Two events A and B are…
    - **Mutually exclusive** if $P(A \cap B) = 0$ *(cannot occur at the same time)*
    - **Independent** if $P(A \cap B) = P(A) \cdot P(B)$ *(don’t influence each other’s likelihood)*

Main Ideas
Chapter 2

- Conditional Probability
  - $P(A\,|\,B)$: The probability that event A occurs, given that event B has already occurred.
    - Know that $P(A \cap B) = P(A)P(B\,|\,A) = P(B)P(A\,|\,B)$
  - Total Probability Rule: S : sample space, $E_1, E_2, \ldots, E_k$: partition of S
    - Then $P(B) = P(B\,|\,E_1)P(E_1) + P(B\,|\,E_2)P(E_2) + \ldots + P(B\,|\,E_k)P(E_k)$
  - Bayes' Theorem: $P(A\,|\,B) = \dfrac{P(B\,|\,A) \times P(A)}{P(B)}$ (*when* $P(B) > 0$)

Main Ideas
Chapter 2

- Random Variables
  - Definition: A function that assigns a real number to each outcome in the sample space of a random experiment.
  - Notation: (1) an uppercase letter (X) for the random variable; (2) a lowercase letter (x) for a measured value of the random variable.
  - Know that a random variable is…
    - **Discrete** when it has a (1) finite; or (2) countably infinite range.
    - **Continuous** when it has an interval of real numbers for its range.

Exercises

Chapter 2

It is estimated that 50% of emails are spam emails. Some software has been applied to filter these spam emails before they reach your inbox. A certain brand of software claims that it can detect 99% of spam emails, and the probability for a false positive (a non-spam email detected as spam) is 5%. Now if an email is detected as spam, then what is the probability that it is in fact a non-spam email?

Exercises

Chapter 2

(1)

It is estimated that 50% of emails are spam emails. Some software has been applied to filter these (2) spam emails before they reach your inbox. A certain brand of software claims that it can detect 99% of spam emails, and the probability for a false positive (a non-spam email detected as spam) is 5%. (3) Now if an email is detected as spam, then what is the probability that it is in fact a non-spam email?

Define the following events: D = email detected as spam, S = email is spam, S' = email is not spam

We also know that (1) $P(S) = P(S') = 0.5$; (2) $P(D \mid S) = 0.99$; and (3) $P(D \mid S') = 0.05$

From TPR we can find $P(D) = P(S)P(D \mid S) + P(S')P(D \mid S') = (0.5)(0.99) + (0.5)(0.05) = 0.52$

And lastly we calculate $P(S' \mid D)$ with Bayes' Theorem

$$P(S' \mid D) = \frac{P(S') \times P(D \mid S')}{P(D)} = \frac{(0.5)(0.05)}{(0.52)} = 0.0481$$

Chapter 3: Discrete Random Variables & Probability Distributions

Main Ideas

Chapter 3

- Probability Mass Function (pmf)
  - $p(x)$: the probability function of a discrete random variable
  - Three properties to know *(if $x_1, \ldots, x_n$ are possible values of X)*

1. $p(x_i) \geq 0$
2. $\displaystyle\sum_{i=1}^{n} p(x_i) = 1$
3. $p(x_i) = P(X = x_i)$ *(the definition of pmf)*

Main Ideas
Chapter 3

- Cumulative Distribution Function (cdf)
    - $F(x)$: probability that a random variable X is *less than or equal to* x.
    - Three properties to know:
        1. $F(x) = P(X \leq x)$ *(definition of cdf)*
        2. $0 \leq F(x) \leq 1$
        3. If $x \leq y$, then $F(x) \leq F(y)$ *(cdf's are monotonically increasing)*

Main Ideas
Chapter 3

- Expected Value & Variance & Standard Deviation of DRVs
  - Expected value (population mean) of X is given by:
    - $\mu = E(x) = \sum_{i=1}^{n} p(x_i) \cdot x_i$
  - Variance of X is given by:
    - $\sigma^2 = V(X) = \sum_{i=1}^{n} (x_i - \mu)^2 p(x_i)$
  - Standard deviation of X is given by:
    - $\sigma = \sqrt{\sigma^2} = \sqrt{V(X)}$

Main Ideas

Chapter 3

- Types of DRV Probability Distributions
  - **Discrete Uniform Distribution**
    - DRV X assumes the values $a, a+1, \ldots, b-1, b$ with equal probabilities.
    - pmf: $p(x) = 1/(b-a+1)$
    - Mean: $\mu = E(X) = (a+b)/2$
    - Variance: $\sigma^2 = V(X) = \dfrac{(b-a+1)^2 - 1}{12}$

Main Ideas
Chapter 3

- Types of DRV Probability Distributions
  - **Bernoulli Trial**
    - DRV X assumes one of two outcomes (*success* = 1, *failure* = 0); *success* has probability p and *failure* has probability $1 - p$
    - <u>pmf:</u> $p(x) = p^x(1 - p)^{1-x}$
    - <u>Mean:</u> $\mu = E(X) = p$
    - <u>Variance:</u> $\sigma^2 = V(X) = p(1 - p)$

Main Ideas

Chapter 3

- Types of DRV Probability Distributions
  - **Binomial Distribution**
    - DRV X is the number of successes in n Bernoulli Trials
    - pmf: $p(x) = \binom{n}{x} p^x (1-p)^{n-x}$ *(for $x = 1,2,\ldots,n$)*
    - Mean: $\mu = E(X) = np$
    - Variance: $\sigma^2 = V(X) = np(1-p)$

Main Ideas

Chapter 3

- Types of DRV Probability Distributions
  - **Poisson Distribution**
    - DRV X is the number of events in a Poisson process *(where events occur randomly but at a constant rate)*
    - x : quantity; T : duration ; λ : rate (quantity / duration)
    - pmf: $p(x) = \dfrac{e^{-\lambda T}(\lambda T)^x}{x!}$ *(for $x = 1,2,3,\ldots$)*
    - Mean: $\mu = E(X) = \lambda T$
    - Variance: $\sigma^2 = V(X) = \lambda T$

Exercises

Chapter 3

The probability that a student is accepted to a prestigious college is 0.3. If ten students from the same school apply, what is the probability that at least four, but less than seven, are accepted?

Exercises

Chapter 3

The probability that a student is accepted to a prestigious college is 0.3. If ten students from the same school apply, what is the probability that at least four, but less than seven, are accepted?

Here we want to calculate cumulative probabilities of a binomial distribution with $p = 0.3$ and $n = 10$. Let X be the number of students accepted.

$$P(4 \leq X < 7) = P(4 \leq X \leq 6) = P(X \leq 6) - P(X \leq 3)$$

$$= \text{binomcdf}(10{,}0.3{,}6) - \text{binomcdf}(10{,}0.3{,}3) = 0.9894 - 0.6494 = 0.3398$$

*use the “binomcdf” function on your calculator for convenience

Exercises

Chapter 3

Suppose it has been observed that, on average, 180 cars per hour pass a specified point on a particular road in the morning rush hour. Due to impending roadworks it is estimated that congestion will occur closer to the city centre if more than 5 cars pass the point in any one minute. What is the probability of congestion occurring?

Exercises

Chapter 3

Suppose it has been observed that, on average, 180 cars per hour pass a specified point on a particular road in the morning rush hour. Due to impending roadworks it is estimated that congestion will occur closer to the city centre if more than 5 cars pass the point in any one minute. What is the probability of congestion occurring?

We will use the Poisson distribution with (1) $\lambda = 180$ cars/hour; and (2) $T = 1$ minute $= 1/60$ hours. Let X be the number of cars passing the point.

$P(X > 5) = 1 - P(X \leq 5) = 1 -$ poissoncdf$((180)(1/60),5) = 1 - 0.9161 = 0.0839$

*use the “poissoncdf” function on your calculator for convenience

Chapter 4: Continuous Random Variables & Probability Distributions

Main Ideas

Chapter 4

- Probability Density Function (pdf)
  - $f(x)$: the probability function of a continuous random variable
  - Three properties to know:

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$

Main Ideas
Chapter 4

- Cumulative Distribution Function (cdf)
    - $F(x)$: probability that a random variable X is *less than or equal to* x.
    - Two properties to know:
        1. $F(x) = P(X \leq x) = P(X < x)$ *(definition of cdf)*
        2. $F(x) = \int_{-\infty}^{x} f(x)dx$
    - Notice that $P(x_1 \leq X \leq x_2) = P(X \leq x_2) - P(X \leq x_1) = F(x_2) - F(x_1)$

Main Ideas
Chapter 4

- Expected Value & Variance & Standard Deviation of CRVs
  - Expected value (population mean) of X is given by:
    - $\mu = E(x) = \int_{-\infty}^{\infty} xf(x)dx$
  - Variance of X is given by:
    - $\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx$
  - Standard deviation of X is given by:
    - $\sigma = \sqrt{\sigma^2} = \sqrt{V(X)}$

Main Ideas

Chapter 4

- Types of CRV Probability Distributions
  - **Continuous Uniform Distribution**
    - CRV X assumes values in (a, b) with equal probabilities.
    - pdf: $f(x) = 1/(b-a)$
    - Mean: $\mu = E(X) = (a+b)/2$
    - Variance: $\sigma^2 = V(X) = \dfrac{(b-a)^2}{12}$

Main Ideas

Chapter 4

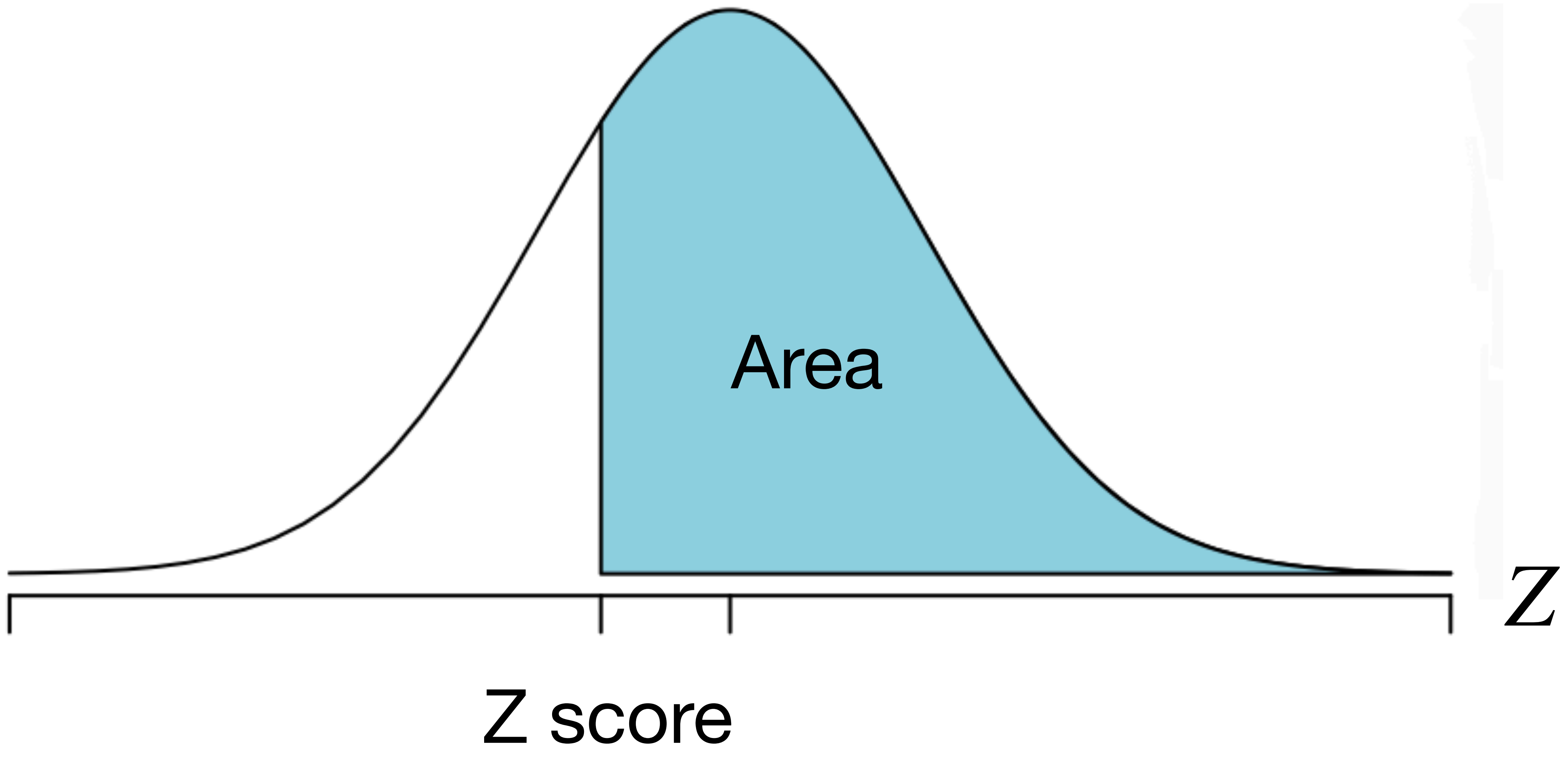
- Types of CRV Probability Distributions
  - **Normal Distribution**
    - $X \sim N(\mu, \sigma^2)$ is used to denote that CRV X is drawn from the normal distribution. μ is the mean and σ^2 is the variance.
    - **normalcdf(** L, U, μ, σ **) :** finds the probability between two values L and U on a calculator. No need to integrate the pdf.

Main Ideas
Chapter 4

- Types of CRV Probability Distributions
  - **Standard Normal Distribution**
    - Normal distribution with $\mu = 0$ and $\sigma^2 = 1$
    - The standard normal random variable is denoted as Z , and for any $X \sim N(\mu, \sigma^2)$, we can find Z with the following equation:
    - $Z = \dfrac{X - \mu}{\sigma}$ *(and vice versa:* $X = \mu + \sigma Z$ *)*

Main Ideas

Chapter 4



- Types of CRV Probability Distributions
  - **Standard Normal Distribution**
    - Z-score: value on the Z-axis where the area to the right of the value is equal to the given probability.
    - Find *area* given *Z-score*: normalcdf$(zscore, \infty, \mu = 0, \sigma = 1)$
    - Find *Z-score* given *area*: invNorm$(area, \mu = 0, \sigma = 1, Tail = Right)$

Main Ideas

Chapter 4

- Types of CRV Probability Distributions
  - **Exponential Distribution**
    - CRV X assumes a distance between successive events in a *Poisson process* with mean number of events $\lambda > 0$ per unit interval.
    - pdf: $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$
    - Mean: $\mu = E(X) = 1/\lambda$
    - Variance: $\sigma^2 = V(X) = 1/\lambda^2$
    - cdf: $F(x) = 1 - e^{-\lambda x}$ for $x \geq 0$

Exercises

Chapter 4

The Edwards's Theater chain has studied its movie customers to determine how much money they spend on concessions. The study revealed that the spending distribution is approximately normally distributed with a mean of \$4.11 and a standard deviation of \$1.37. What percentage of customers will spend less than \$3.00 on concessions? What spending amount corresponds to the top 87th percentile?

Exercises

Chapter 4

The Edwards's Theater chain has studied its movie customers to determine how much money they spend on concessions. The study revealed that the spending distribution is approximately normally distributed with a mean of $4.11 and a standard deviation of $1.37. What percentage of customers will spend less than $3.00 on concessions? What spending amount corresponds to the top 87th percentile?

We will take the cumulative probability of a normal distribution with $\mu = 4.11$ and $\sigma = 1.37$. Let X be the dollars spend on concessions.

$$P(X < 3) = P(X \leq 3) = \text{normalcdf}(-\infty, 3, \mu = 4.11, \sigma = 1.37) = 0.2089$$

Therefore 20.89% of customers will spend less than $3 on concessions.

* With a calculator use "1E-10" in place of $-\infty$

Exercises

Chapter 4

The Edwards's Theater chain has studied its movie customers to determine how much money they spend on concessions. The study revealed that the spending distribution is approximately normally distributed with a mean of \$4.11 and a standard deviation of \$1.37. What percentage of customers will spend less than \$3.00 on concessions? What spending amount corresponds to the top 87th percentile?

"Top 87th percentile" means that the area to the right of the value we want to find is 0.87.

Therefore we can calculate $x = \text{invNorm}(0.87, \mu = 4.11, \sigma = 1.37, Tail = Right) = 2.567$

The spending amount of \$2.56 corresponds to the top 87th percentile.

Chapter 6: Descriptive Statistics

Main Ideas

Chapter 6

- Terminology
  - <u>Population:</u> The totality of *all* observations
  - <u>Sample:</u> A *portion* of the population used for analysis
  - <u>Mean:</u> A measure representing the “center” of the data
  - <u>Standard deviation:</u> A measure representing the “spread” of the data

Main Ideas
Chapter 6

- Population Measures
  - If a **population** consists of N observations $x_1, x_2, \ldots, x_N$, then we can calculate the following:

Population mean	Population variance	Population standard deviation
$\mu = \sum_{i=1}^{N} \frac{x_i}{N}$	$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$	$\sigma = \sqrt{\sigma^2}$

Main Ideas
Chapter 6

- Sample Measures
  - If a **sample** consists of n observations $x_1, x_2, \ldots, x_n$, then we can calculate the following:

Sample mean	Sample variance	Sample standard deviation
$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$	$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$	$s = \sqrt{s^2}$

  - Know that the **degrees of freedom** in a sample is $n - 1$

Main Ideas

Chapter 6

- Frequency Distributions
  - Compact summaries of our data
  - Data is grouped into **bins**
  - If n is the sample size, then…
    - # of bins $\approx \sqrt{n}$
    - range = Max. data − Min. data
- bin width = range / # of bins

Main Ideas
Chapter 6

- Histograms
    - Used to visualize frequency distributions
    - <u>Categorical data:</u> bins are grouped by categories, not numbers.
        - Includes two types:
            - <u>Ordinal:</u> the categories have a natural order *(e.g. Mon, Tues, …)*
                - Example: a **Pareto chart** places categories in decreasing order
            - <u>Nominal:</u> the categories don't have a natural order *(e.g. Red, Blue, …)*

Exercises

Chapter 6

Sam decides to sample 5 rose bushes from his backyard, and the number of flowers on each bush are 9, 2, 5, 4, and 12. Work out the standard deviation.

Exercises

Chapter 6

Sam decides to sample 5 rose bushes from his backyard, and the number of flowers on each bush are 9, 2, 5, 4, and 12. Work out the standard deviation.

First we can find the sample mean: $\bar{x} = \frac{1}{5}(9 + 2 + 5 + 4 + 12) = 6.4$

Then we can find the sample variance:

$$s^2 = \frac{1}{5-1}\left((9-6.4)^2 + (2-6.4)^2 + (5-6.4)^2 + (4-6.4)^2 + (12-6.4)^2\right) = 16.3$$

Lastly we find the sample standard deviation: $s = \sqrt{16.3} = 4.037$

Chapter 7: Methods of Point Estimation of Parameters and Sampling Distributions

Main Ideas

Chapter 7

- Statistic vs. Point Estimate
  - A **statistic** is any function of the observations in a random sample.
    - e.g. $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ is the statistic corresponding to population parameter μ.
    - We use random variables X_i instead of observations x_i because statistics are *probability distributions*, not single values.

Main Ideas

Chapter 7

- Statistic vs. Point Estimate
  - A **point estimate** is a single reasonable value / representation of a population parameter.
    - e.g. any sample mean $\bar{x}$ we calculate is a point estimate of μ
  - Key takeaway: The **statistic** is a *function / probability distribution*, while the **point estimate** is a *specific observed instance* of the statistic (e.g. output of a function when we have a specific input).

Main Ideas
Chapter 7

- Examples of statistics for different population parameters

Parameter	Measure	Statistic
μ	Mean of single population	$\bar{X}$
σ^2	Variance of single population	S^2
σ	Standard deviation of single population	S
p	Proportion of single population	$\hat{P}$
$\mu_1 - \mu_2$	Difference in means of two populations	$\bar{X}_1 - \bar{X}_2$
$p_1 - p_2$	Difference in proportions of two populations	$\hat{P}_1 - \hat{P}_2$

Main Ideas
Chapter 7

- Central Limit Theorem
  - If $X_1, X_2, \ldots, X_n$ is a random sample of size n taken from a population with mean μ and variance σ^2, and if $\bar{X}$ is the sample mean, then

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

becomes a normal random variable as n increases.

Exercises

Chapter 7

The length of time, in hours, it takes a group of people to play one soccer match is normally distributed with a mean of 2 hours and a standard deviation of 0.5 hours. A sample of size $n = 50$ is drawn randomly from the population. Find the probability that the sample mean is between 1.8 hours and 2.3 hours.

Exercises

Chapter 7

The length of time, in hours, it takes a group of people to play one soccer match is normally distributed with a mean of 2 hours and a standard deviation of 0.5 hours. A sample of size $n = 50$ is drawn randomly from the population. Find the probability that the sample mean is between 1.8 hours and 2.3 hours.

We will apply the CLT to find the sample mean distribution and take a cumulative probability.

Let $\bar{X}$ denote the average number of hours it takes to play a soccer match. Then from CLT we know that $\bar{X} \sim N(\mu, \sigma^2/n) = N(2, 0.5^2/50)$.

So $P(1.8 \leq \bar{X} \leq 2.3) =$ normalcdf$(1.8, 2.3, \mu = 2, \sigma = 0.5/\sqrt{50}) = 0.9977$