

STAT 350 Help Session for Midterm 1

Chapter 1: Introduction to Statistics

- Three branches of Statistics
 - Data Collection: the process of gathering information
 - Descriptive Statistics: summarizing and organizing data
 - Drawing conclusions from data
- Population (with parameters): the entire collection of individuals or objects to be considered or studied (e.g. μ)
- Sample (with statistics): a subset of the entire population (e.g. \bar{x} and s^2)
- Probability: assume the population model is known, and answer questions concerned with sampling
- Inferential Statistics: use the information from the sample to answer questions concerning population

Practice Exam

2.6 Which of the following should be used if we know the form and all parameters of the population level model and we want to gain insight into the possible samples?

- Ⓐ Descriptive statistics
- Ⓑ Data collection
- Ⓒ Probability
- Ⓓ Inferential statistics
- Ⓔ Code

Chapter 2: Data Types and Distribution Shapes (Sample level)

- Quantitative (numerical) vs Qualitative (categorical)
- Distribution shapes based on histogram
 - Peaks: unimodal, bimodal, multimodal
 - Symmetry: symmetric, right-skewed (positive skew), left-skewed (negative skew)
- Location of mode, median, mean by distribution shapes

FALL 2023

1.5. An application in senior technology aims to help improve the overall health, balance, and flexibility of the elderly by tracking several variables. One of these variables is the zip code of the participant.

Ⓙ or Ⓢ This variable is a discrete numerical variable.

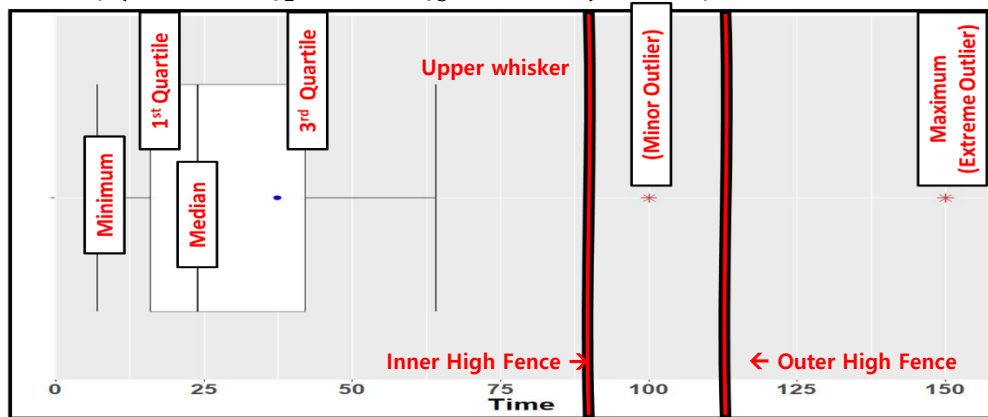
1.6. For a random variable X that follows a normal distribution,

Ⓙ or Ⓢ the mode of X is always greater than its mean.

(Bonus) What if X follows a Poisson/exponential distribution?

Chapter 3: Descriptive Statistic (Sample level)

- Five-number summary (*minimum, Q_1 , Median, Q_3 , maximum*) and boxplot



- Measure of central tendency: mean, median, mode

$$\text{Sample Mean: } \bar{x} = \sum x_i / n \quad \text{Sample Median: } \tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} & \text{if } n \text{ is even} \end{cases}$$

- Measure of spread: variance (standard deviation), Interquartile Range IQR, range

$$\text{Sample Variance: } s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad \text{Sample Standard Deviation: } s = \sqrt{s^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \quad IQR = Q_3 - Q_1$$

- Appropriate measures of location and spread for given data (symmetric vs skewed)
Use median and IQR for skewed distributions or with strong outliers (i.e., median and IQR are resistant/robust)

SPRING 2024

- 1.5. Given a five number summary for a dataset we could compute the interquartile range, identify the fences, and draw a modified box plot to visualize properties of the data.

Ⓓ or Ⓕ The upper whisker of the modified boxplot would be drawn to terminate at the point $Q_3 + 1.5 \times IQR$.

FALL 2023

- 1.6. A manufacturing company, aiming to meet quality standards, assesses the lifespan of its light bulbs. The company claims that the mean lifespan of the bulbs is 1200 hours with a standard deviation of 150 hours. To verify, this a large retailer tests a sample of 215 bulbs, finding a sample mean of 1180 hours with a sample standard deviation of 157 hours.

Ⓓ or Ⓕ The correct symbol to represent the 1200 hours is \bar{x} .

- 2.3. The measure of spread which is the most likely to be influenced by outliers in the dataset is

- Ⓐ sample mean
- Ⓑ sample median
- Ⓒ sample standard deviation
- Ⓓ Inner quartile range (IQR)
- Ⓔ more than one of the above.

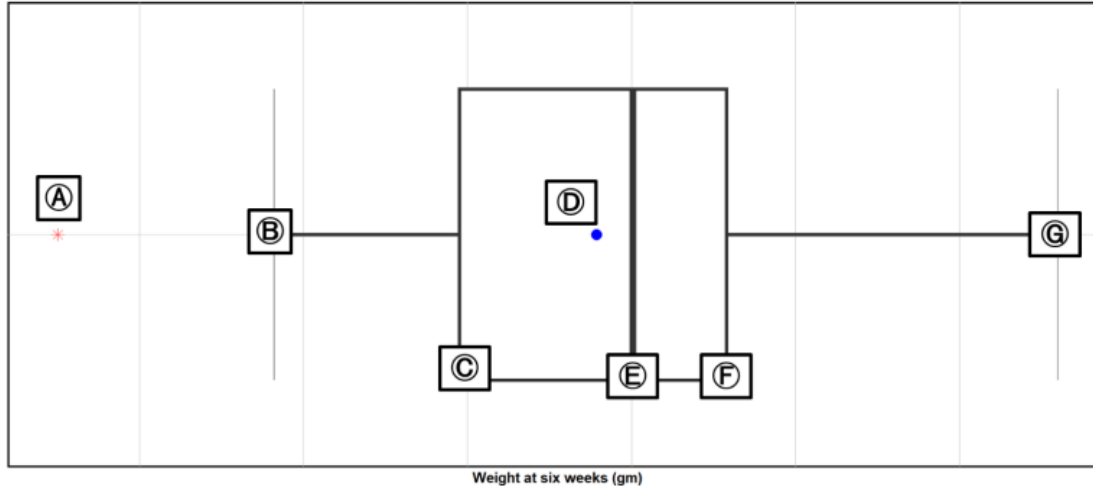
SUMMER 2024

The following dataset records the weights of 24 chickens at six weeks of age, recorded to study the effects of various feeding regimes. Each chicken was fed one of two diets, Meatmeal, or Linseed. Below are the recorded weights (in grams) of the chickens:

75, 141, 148, 153, 169, 181, 203, 206, 213, 229, 242, 244,
257, 257, 258, 260, 263, 271, 303, 309, 315, 325, 344, 380

(Note: the data is stored in a dataframe `chick_weights` with variable `weight`)

Modified Boxplot of Chicken Weight at six weeks (gm)



The R code has provided the following statistical measures:

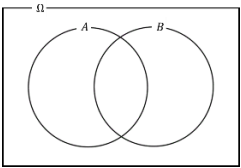
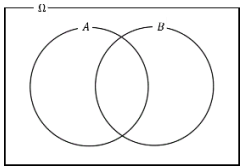
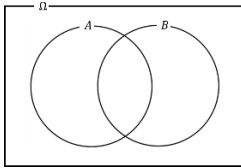
```
> quantile(chick_weights$weight)
 0%   25%   50%   75%  100%
75.0 197.5 250.5 279.0 380.0
> mean(chick_weights$weight)
[1] 239.4167
```

Refer to the modified boxplot provided above.

Based on the output from the R code below and the provided data, determine the numerical values indicated by the labels **A** through **G**.

Chapter 4: Probability (Population level)

- Set theory & Venn Diagram
 - Sample Space Ω : a set of all possible outcomes
 - Event: any set of outcomes from an experiment
 - Empty event/set \emptyset or $\{\}$

Definition	Union	Intersection	Complement	Subset	Disjoint (mutually exclusive)
Notation	$A \cup B$	$A \cap B$	A' or A^c	$A \subset B$	$A \cap B = \emptyset$
Venn Diagram					

- Axioms of probability
 - For any event $E \subseteq \Omega$, $0 \leq P(E) \leq 1$: probability must be between 0 and 1.
 - $P(\Omega) = 1$: the probability of the sample space is always 1.
 - For any event $E \subseteq \Omega$, $P(E) = \sum_{\omega \in E} P(\omega)$
- General Probability Rules
 - General Addition Rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
 - Complement Rule: $P(A') \text{ or } P(A^c) = 1 - P(A)$
 - Law of Partitions (simple case): $P(A) = P(A \cap B) + P(A \cap B^c)$
 - Law of Total Probability (simple case): $P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$
 - Conditional Probability: $P(B|A) = \frac{P(A \cap B)}{P(A)}$
 - General Multiplication Rule (simple case): $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$
- Independence: let two events A and B are independent, then
 - $P(A \cap B) = P(A) \times P(B)$
 - $P(A|B) = P(A)$
 - $P(B|A) = P(B)$
- Dependent Events: Bayes' Rule – Find $P(A|B)$ when $P(B|A)$ is available.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \quad (\text{simple case})$$

SPRING 2024

1.2. Suppose two events A and B are in the sample space Ω with all outcomes of A contained within the event B .

T or **F** In this scenario it must follow that $P(A \cap B) = P(B)$.

1.3. Given two non-empty events A and B of a sample space Ω ,

T or **F** if $P(A|B) = P(A)$ then we are certain that $A \cap B \neq \emptyset$.

Practice Exam

1.1 Let A and B be events in a same sample space Ω , with $P(A) > 0$ and $P(B) > 0$.

T or **F** If $P(A \cap B) = P(A)P(B)$, then the events A and B must be independent.

1.2 Let **A** and **B** be events in a same sample space Ω . It is known that $P(A) > 0$ and $P(B) > 0$.

☐ **T** or ☐ **F** If **A** and **B** are mutually exclusive, then it must be true that $P(A) = 1 - P(B)$.

FALL 2023

1.1. Suppose that events **A** and **B** belong to the same sample space Ω , and that the values of $P(A)$ and $P(B)$ are **non-zero** and known.

☐ **T** or ☐ **F** If $P(A \cap B) = P(A)$, then $P(A' \cap B') = P(A')$.

5. (16 points) Suppose a college student uses the bus 70% of the time to go to his first class of the week on the college campus. Further suppose the student chooses to walk whenever he chooses not to take the bus. Assume that the student can choose only one form of transportation at a time. If the student chooses to walk, he is late 30% of the time. Further, if he chooses to use the bus, he is late 20% of the time. Using this information, answer the following questions.

Summarize the information given in the problem

a) **(6 pts)** What is the probability that the student will be late when he travels to his first class of the week on the college campus?

2. (15 points, 3 points each) **Multiple Choice Questions.** Indicate the correct answer by completely filling in the appropriate circle. If you indicate your answer by any other way, you may be marked incorrect. **For each question, there is only one correct option letter choice.**

2.1. The number of customers arriving at a UPS branch during working hours follows a **Poisson distribution** with an **average rate of 4 customers per hour**. Let X denote the number of customers arriving between **9:00 AM** and **10:00 AM** and let Y denote the number of customers arriving between **10:30 AM** and **12:00 PM**.

What is the **conditional probability** that exactly **3 customers** arrive between **10:30 AM** and **12:00 PM**, given that **6 customers** arrived between **9:00 AM** and **10:00 AM**?

- (A) $P(Y = 3|X = 6) = 0$
- (B) $P(Y = 3|X = 6) = 0.0093$
- (C) $P(Y = 3|X = 6) = 0.0892$
- (D) $P(Y = 3|X = 6) = 0.1954$
- (E) $P(Y = 3|X = 6) = 0.8564$

2.3. Suppose $X \sim \text{Binomial}(n = 10, p = 0.1)$ and $Y \sim \text{Binomial}(n = 10, p = 0.9)$.

Which statement is **not always true** about X and Y ?

- (A) The **mode** of X is less than the **mode** of Y .
- (B) $SD(X) - \sqrt{\text{Var}(Y)} = 0$
- (C) $P(X = 1 \cap Y = 8) = 0.1943$
- (D) $E[X^2] = (10)(0.1)(0.9) + [(10)(0.1)]^2$
- (E) $P(X = 1) = P(Y = 9)$

Chapter 5: Discrete Random Variables (Population level)

- Valid probability distribution (discrete)
 - Each probability $p_X(x)$ satisfies $0 \leq p_X(x) \leq 1$.
 - The sum of probabilities is one: $\sum p_X(x) = 1$.
- Expectation and Variance rules
 - $E[X] = \sum x \times p_X(x)$ (discrete)
 - $Var(X) = \sum x^2 \times p_X(x) = E[X^2] - E[X]^2$
 - $sd(X) = \sqrt{Var(X)}$
 - $E[aX \pm bY] = aE[X] \pm bE[Y]$
 - $Var(a \pm bX) = b^2 Var(X)$
 - When X and Y are independent: $Var(X + Y) = Var(X) + Var(Y)$
- Binomial Distribution $X \sim Bin(n, p)$
 - X : number of successes, n : total number of trials, p : probability of success
 - Pmf: $p_X(x) = P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} = \frac{n!}{(n-x)!x!} p^x (1 - p)^{n-x}$ for $x \in \{0, 1, \dots, n\}$
 - $E[X] = np$, $Var(X) = np(1 - p)$
- Poisson Distribution $X \sim Poisson(\lambda)$
 - X : number of events occurred in a given interval, λ : average number of events in a given interval
 - Pmf: $p_X(x) = P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}$ for $x \in \{0, 1, 2, \dots\}$
 - $E[X] = Var(X) = \lambda$
 - Properties:
 - The rate λ is proportional to an interval of time/length/volume.
 - Two Poisson random variables with non-overlapping intervals are independent.

SPRING 2024

1.1. Let $X \sim \text{Binomial}(n, p = 0.5)$, where n is any positive integer.

Ⓓ or Ⓕ For any value of x in the support of X , $P(X = x) = P(X = n - x)$.

2.1. Let X be a random variable with mean $\mu_X = 7$ and standard deviation $\sigma_X = 9$. Define another random variable $Y = 2X^2 + 5X + 3$. Determine the value of $E[Y]$.

(Bonus) Let $E[\log(Z)] = 0.4$ and $Var[\log(Z)] = 0.1$. Find:

- $E[3 \log(Z) + 5] =$
- $E[\{\log(Z)\}^2] =$
- $E[\log(Z^2)] =$

2.3. In Cerulean city, 3.2 car accidents are reported on average per day. The number of car accidents is known to follow a Poisson distribution. What is the probability that at least one accident occurs per day? Let X denote the Poisson random variable for this situation.

$$P(X \geq 1) =$$

Chapter 6: Continuous Random Variables and Probability Distributions (Population level)

- Calculus Prerequisites

$$\int_a^b x^n dx = \left[\frac{1}{n+1} x^{n+1} \right]_a^b = \frac{1}{n+1} (b^{n+1} - a^{n+1})$$

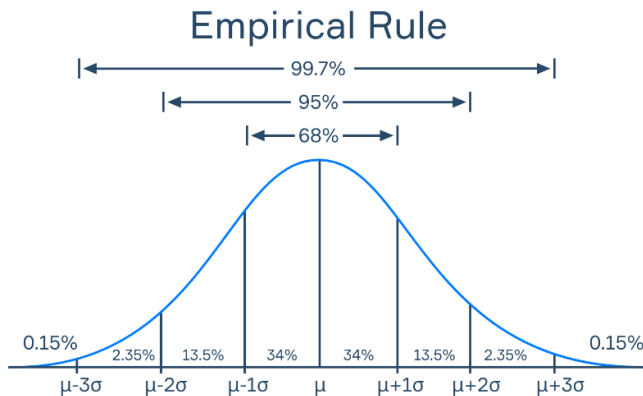
$$\int_a^b e^{nx} dx = \left[\frac{1}{n} e^{nx} \right]_a^b = \frac{1}{n} (e^{nb} - e^{na})$$

- General continuous random variables & Properties

- PDF: $f_X(x) \Rightarrow$ this does NOT mean $P(X = x)$.
 - $f_X(x) \geq 0$: the pdf is nonnegative.
 - $\int_{-\infty}^{\infty} f_X(x) dx = 1$: the total area under the curve is one.
 - $P(X = x) = 0$ for any x when X is a continuous random variable.
- CDF: $F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(z) dz$
 - Non-decreasing Function: $F_X(a) \leq F_X(b)$ when $a < b$
 - Limiting Behavior: $F_X(\infty) = 1$ and $F_X(-\infty) = 0$
 - Probabilities: $P(a < X < b) = F_X(b) - F_X(a)$ and $P(X > a) = 1 - F_X(a)$
 - Percentiles: $(100 * p)$ -th percentile of $X \Leftrightarrow \int_{-\infty}^{x_p} f_X(z) dz = p$ & solve for x_p
- Expectation and Variance
 - $E[X] = \int x * f_X(x) dx$
 - $Var[X] = E[X^2] - E[X]^2$
 - $E[g(X)] = \int g(x) f_X(x) dx$ e.g., $E[X^2] = \int x^2 * f_X(x) dx$

- Normal Distribution $X \sim N(\mu, \sigma)$ or $N(\mu, \sigma^2)$

- Empirical Rule



- CDF of Standard Normal: $\Phi(z) = P(Z \leq z)$ where $Z \sim N(0, 1)$
 - Probabilities: standardize $X \rightarrow$ represent them in $\Phi(\cdot) \rightarrow$ look at the table
 - Percentiles: find percentile in $Z \rightarrow$ convert it to the scale of X
- Check Normality using Histogram or Normal Probability Plots (see Ch.6 slides 64-69)
- Uniform Distribution $X \sim Uniform(a, b)$ where a : lower limit, b : upper limit
- $f_X(x) = \frac{1}{b-a}$, $a \leq x < b$ and $F_X(x) = \frac{x-a}{b-a}$, $a \leq x < b$
- $E[X] = \frac{a+b}{2}$, $Var(X) = \frac{(b-a)^2}{12}$
- Exponential Distribution $X \sim Exp(\lambda)$ with the average of $1/\lambda$
 - $f_X(x) = \lambda e^{-\lambda x}$, $x \geq 0$ and $F_X(x) = 1 - e^{-\lambda x}$, $x \geq 0$
 - $E[X] = \frac{1}{\lambda}$, $Var(X) = \frac{1}{\lambda^2}$

Practice Exam

1.4 Suppose X follows a normal distribution with mean μ and standard deviation σ and $Z \sim N(0, 1)$.

Ⓓ or Ⓕ It follows that $P(\mu - 2\sigma < X < \mu + 2\sigma) = 2P(Z < 2) - 1$

2.1. Suppose X is a continuous random variable with mean $\mu = 5$ and standard deviation $\sigma = 5$. What is the value of $P(X = 10)$?

2.4. Let X and Y be two normal random variables with means μ_X and μ_Y , and standard deviations σ_X and σ_Y , respectively. Select the **correct statement** regarding the relationship between the parameters $(\mu_X, \mu_Y, \sigma_X, \sigma_Y)$ when the **normal curve associated with X is more peaked** than the **normal curve associated with Y** .

- Ⓐ $\sigma_X < \sigma_Y$
- Ⓑ $\sigma_X > \sigma_Y$
- Ⓒ $(\mu_X / \sigma_X) > (\mu_Y / \sigma_Y)$
- Ⓓ $\mu_X < \mu_Y$
- Ⓔ $\mu_X > \mu_Y$

6. Use $k = 1/32$.

$$f_X(x) = \begin{cases} k & -3 \leq x \leq -1 \\ k(-15x + 45) & 1 \leq x \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

$$F_X(x) = \begin{cases} \text{[A]} & x \leq -3 \\ \text{[B]} & -3 \leq x < -1 \\ \text{[C]} & -1 \leq x < 1 \\ -\frac{15}{64}x^2 + \frac{45}{32}x - \frac{71}{64} & 1 \leq x < 3 \\ \text{[E]} & x \geq \text{[D]} \end{cases}$$

b) (6 pts) Determine the missing parts **[A, B, C, D, E]** of the cumulative distribution function.

d) (6 pts) What value represents the 5th percentile of this distribution?

2.4. Suppose a random variable X follows a normal distribution with an unknown mean, μ , and unknown variance, σ^2 . Then, $P(X > \mu + \sigma)$ is approximately equal to?

3. (24 points) Marina orders her dinner from Doordash. When she places an order, it is known to take 35 minutes on average for delivery. Assume each order's delivery time is independent of others.

a) (4 points) Define the continuous random variable X which represents the amount of time (in minutes) Marina waits for her delivery. Write the name of its distribution and provide the value of the parameter, λ or μ .

b) (4 points) What is the probability that Marina will wait exactly 38 minutes for her delivery?

c) (6 points) What is the probability that Marina will wait more than 25 minutes for her delivery?

d) (10 points) If the delivery takes less than 25 minutes, Marina will add an additional tip to the deliverer. Assume she placed 10 orders in January. What is the probability that Marina adds additional tip for 2 orders out of 10 orders?

4. (24 points) Health authorities at Lumina University observed that the weights of their students follow a normal distribution. Furthermore, their assessment revealed that the mean weight of their students is 160 lbs, with a variance of 25 lbs². Using this information, answer the following questions:

b) (8 points) What is the probability that a student at Lumina University weighs between 160 lbs and 168 lbs?

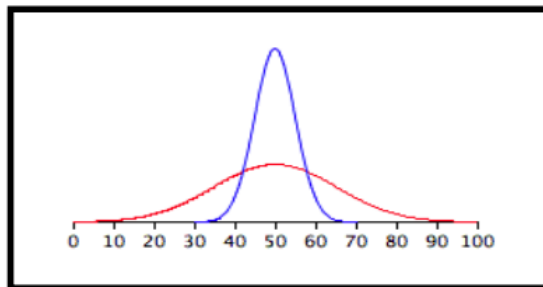
c) (10 points) Lumina University's health officials declared 0.25% of students overweight. What cutoff value was used by them to determine whether a student is overweight or not?

Fall 2024

- 1.4. Let V be a random variable with a **probability density function** $f_V(v)$ that is **non-zero only** on the **interval** $[-5, -2)$. Let $F_V(\cdot)$ denote the **cumulative distribution function (CDF)** of V .

Ⓙ or Ⓕ Then, $F_V(c) = 1$ holds for any $c > 0$.

- 1.6. For the figure below,



Ⓙ or Ⓕ the blue normal distribution has more area underneath its curve than the red normal distribution does.