Descriptive Statistics and Exploratory Data Analysis

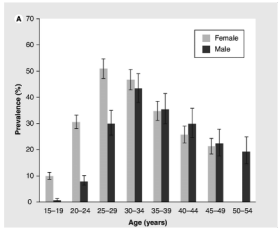
1

Exploratory/Descriptive Statistics

"Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone- the first step"

John Tukey, founder of EDA "school"

- Summarization and presentation of data
- Generally one of first steps to scientific discovery
- Definitely one of first steps to scientific understanding
  - If you can't see it, question it



2

Inferential/Confirmatory Statistics

- Generalization of conclusions:

  sample ⟶ population
- Assess strength of evidence
- Make comparisons
- Make predictions

Tools:

- Modeling
- Estimation and Confidence Intervals
- Hypothesis Testing

3

Exploratory vs Inferential Data Analysis

Exploratory (Descriptive)

- Forming ideas/hypotheses

Inferential (Confirmatory)

- Investigating predefined ideas/hypotheses

Historically these approaches have been studied separately, but there is ongoing modern research into unifying them (2010 – present)

4

Types of Data

- Categorical (qualitative)
  1) Nominal scale - no natural order
     - sex at birth, gender identity, nationality, …
  2) Ordinal scale - natural order exists
     - low/medium/high, BMI categories …
- Numerical (quantitative)
  1) Discrete - (few) integer values
     - number of children in a family
  2) Continuous - measure to a given level of precision
     - blood pressure, weight

Different types of data require different analysis and graphics tools

e.g., categorize zip code

Samples

In statistics, we usually analyze a **sample** of observations or measurements.

We will denote a sample of N numerical values as:

$$X_1, X_2, X_3, \ldots, X_N$$

where X_1 is the first sampled datum, X_2 is the second, etc.

e.g. Ages of 3 people:

$X_1 = 60$

$X_2 = 33$

$X_3 = 41$

Ordered Samples

Sometimes it is useful to order the measurements. We denote the ordered sample as:

$$X_{(1)}, X_{(2)}, X_{(3)}, \ldots, X_{(N)}$$

where $X_{(1)}$ is the smallest value and $X_{(N)}$ is the largest.

Consider age example:

$X_1 = 60$ ⟶ $X_{(1)} = 33$

$X_2 = 33$ ⟶ $X_{(2)} = 41$

$X_3 = 41$ ⟶ $X_{(3)} = 60$

Arithmetic Mean

The **arithmetic mean** is the most common measure of the **central location** of a sample. We use $\overline{X}$ to refer to the mean and define it as:

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

The symbol $\sum$ is shorthand for "*sum*" over a specified range. For example:

$$\sum_{i=1}^{4} X_i = X_1 + X_2 + X_3 + X_4$$

$$\sum_{k=1}^{3} Z_k^2 = Z_1^2 + Z_2^2 + Z_3^3$$

Pause-break time then work on exercises

Question 1

Categorize the following variables into nominal, ordinal, discrete, or continuous

a) Viral load

b) Age measured in years

c) Price of your lunch

d) Zip code of your residence

Questions 2, 3, 4

2. What is $\sum_{j=10}^{12} j$?

3. What is $\sum_{j=1}^{3} j^2$?

4. What is the mean of -5, 10, and 0?

Question 5

a) If I buy a bag of 3 bagels, and they weigh 85g, 95g and 90g, what is the mean weight?

b) If I buy a bag of 3 bagels and they weigh 0.085 kg, 0.095 kg and 0.09 kg, what is the mean weight?

c) If I add 20 grams of cream cheese to each of my bagels, what is the mean (combined) weight of my breakfast?

Some Properties of the Mean

Often we wish to **transform** variables. Linear changes to variables impact the mean in a predictable way:

(1) Adding a constant to all values adds that constant to the mean

(2) Multiplication by constant multiplies the mean by that constant

CAREFUL: This does not happen for all transformations. For example, the logarithm of the mean is not the mean of the logarithms.

Median

Another measure of central tendency is the **median** - the "middle one". Half the values are below the median and half are above. Given the ordered sample, $X_{(i)}$, the median is:

N odd: $\text{Median} = X_{\left(\frac{N+1}{2}\right)}$

N even: $\text{Median} = \frac{1}{2}\left(X_{(N/2)} + X_{(N/2+1)}\right)$

Mode

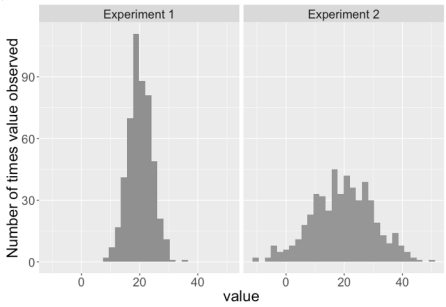
The **mode** is the most frequently occurring value in the sample.

5'5, 5'7, 5'7, 5'7

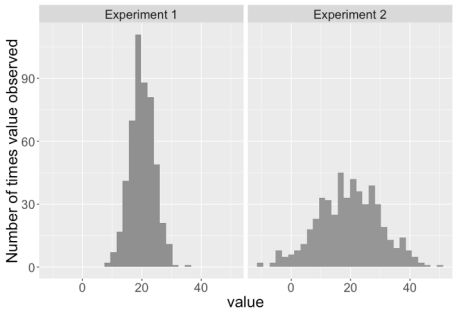
Comparison of Mean and Median

- Mean is sensitive to a few very large (or small) values - "outliers"
- Median is "resistant" to outliers
- Mean has useful mathematical properties

What is different between these 2 experiments and what is the same?



Same mean but the variance differs.



Variance is a way to assess the spread of the data.

17

Measures of Spread: Range

The **range** is the difference between the largest and smallest observations:

$$\begin{aligned} \text{Range} &= \text{Maximum} - \text{Minimum} \\ &= X_{(N)} - X_{(1)} \end{aligned}$$

Alternatively, the range may be denoted as the pair of observations:

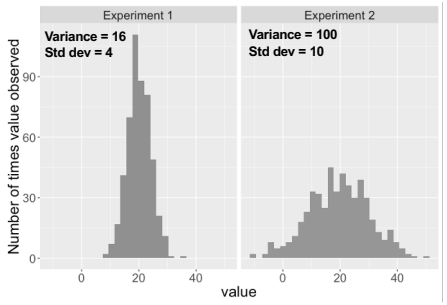
$$\begin{aligned} \text{Range} &= (\text{Minimum}, \text{Maximum}) \\ &= \left(X_{(1)}, X_{(N)}\right) \end{aligned}$$

The latter form is useful for data quality control.

Disadvantage: the sample range increases with increasing sample size.

18

Measures of Spread: Variance



- Most common way to assess spread: variance
- Variance is a measure of the distance from each observation to the center of the observations

$$\text{Variance} = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - X)^2$$

$$\text{Standard deviation} = s = \sqrt{s^2}$$

19

Properties of the variance/standard deviation

- Variance and standard deviation are **ALWAYS** greater than or equal to zero.
- Linear changes are a little trickier than they were for the mean:

(1) Adding a constant to all values does not change the variance or standard deviation

(2) Multiplying by a constant changes the standard deviation by that constant

(3) Multiplying by a constant changes the variance by that constant-squared

20

Measures of Spread: Percentiles and Quartiles

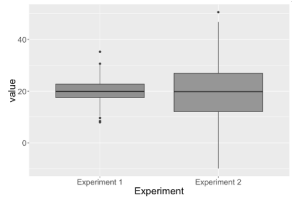
The median is the sample value that has 50% of the data below it.

More generally, we define the **p^{th} percentile** as the value which has p% of the sample values less than or equal to it.

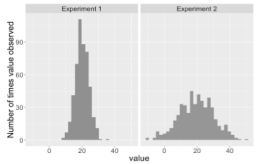
Quartiles are the (25, 50, 75) percentiles. The **interquartile range** is $Q_{.75}$-$Q_{.25}$ and is another useful measure of spread. The middle 50% of the data is found between $Q_{.25}$ and is $Q_{.75}$.

Boxplot

A graphics display of the quartiles of a dataset, as well as the range. Extremely large or small values are also identified.



Note that this is the same data as previously plotted as a histogram:



Summary

- Numerical Summaries
  1. location - mean, median, mode.
  2. spread - range, variance, standard deviation, IQR
- Graphical Summaries
  1. Boxplot
  2. Histograms

Pause-break time then work on exercises

Question 6

a) If I buy a bag of 3 bagels, and they weigh 85g, 95g and 90g, what is the variance and standard deviation of the weight?

(Recall that the mean was 90g)

b) What is the variance and standard deviation of the weight of the 3 bagels if 20g of cream cheese is added to each bagel?

Probability Distributions

I

Probability: Why is it of interest?

Most of the time we are not interested in the samples that we obtained. We are interested in using the samples to inform a more general understanding.

To understand how well our samples generalize to a broader population, we need to know how reliable/representative/variable our samples were.

Population	⟷	Sample
Probability distribution	⟷	Frequency distribution
Parameters	⟷	Estimates

Probability Distribution

Definition: A **random variable** is a characteristic whose obtained values arise as a result of chance factors.

Definition: A **probability distribution** gives the probability of obtaining all possible (sets of) values of a random variable. It gives the probability of the outcomes of an experiment.

Theoretical Distributions

Used to provide a mathematical description of outcomes.
Examples include…

A. Discrete variables

1. Binomial - sums of 0/1 outcomes

   - underlies many epidemiologic applications

   - basic model for logistic regression

2. Multinomial – generalization of binomial

   - a basic model for log-linear analysis

B. Continuous variables

1. Normal - bell-shaped curve; many data summaries are *approximately* normally distributed.

2. t-distribution – similar to the Normal distribution

3. Chi-square distribution (χ^2)

Binomial Distribution - Motivation

Suppose a new student has joined your lab and is learning how to culture cells. Their reference letter says that 25% of the new student's experiments fail. They only have time to create 3 cultures.

- What's the probability that all experiments succeed?
- What's the probability that only 1 experiment fails?
- What's the probability that at least 1 experiment fails?

Bernoulli Trial

A **Bernoulli trial** is an experiment with only 2 possible outcomes, which we denote by 0 or 1 (e.g. coin toss)

Assumptions:

1) Two possible outcomes - success (1) or failure (0).

2) The probability of success, p, is the same for each trial.

3) The outcome of one trial has no influence on later outcomes (independent trials).

Binomial Random Variable

A **binomial random variable** is simply the total number of successes in *n* Bernoulli trials.

Example: number of successful experiments out of 3

To assign probabilities to outcomes of binomial random variables, we first need to know

1. How many ways are there to get k successes (k=0,…3) in n trials?

2. What's the probability of any given outcome with exactly k successes (does order matter)?

Example of a Binomial Random Variable

How many ways are there to get k successes (k=0,…3) in 3 trials?

Experiment succeeds = +
Experiment fails = -

Experiment number 1	2	3	Outcomes
+	+	+	3 successful
+	+	-	2 successful
+	-	+	2 successful
-	+	+	2 successful
+	-	-	1 successful
-	+	-	1 successful
-	-	+	1 successful
-	-	-	0 successful

Combinations

Fortunately there is a general formula:

If C_k^n is the number of ways to get k successes in n attempts, then

$$C_k^n = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

where

"n factorial" = n! = n × (n-1) × … × 1

What are the probabilities of these outcomes?

Experiment number 1	2	3	Outcomes	# ways
p	p	p	3 successful	1
p	p	1-p	2 successful	
p	1-p	p	2 successful	3
1-p	p	p	2 successful	
p	1-p	1-p	1 successful	
1-p	p	1-p	1 successful	3
1-p	1-p	p	1 successful	
1-p	1-p	1-p	0 successful.	1

sequence of k +'s (0, 1, 2, or 3) and (3-k) –'s will have probability

$$p^k(1-p)^{3-k}$$

But there are $\frac{3!}{k!(3-k)!}$ such sequences, so in general…

Binomial Probabilities

What is the probability that a binomial random variable with **n** trials and success probability **p** will yield exactly **k** successes?

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

This formula is called the **probability mass function** for the binomial distribution.

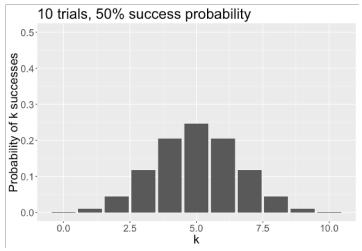
Assumptions:

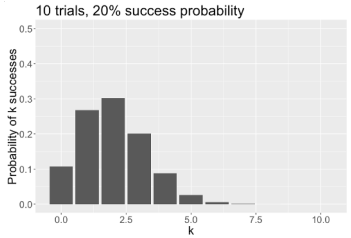
1) Two possible outcomes - success (1) or failure (0) - for each of n trials.

2) The probability of success, p, is the same for each trial.

3) The outcome of one trial has no influence on later outcomes (independent trials).

4) The random variable of interest is the total number of successes.





37

Binomial Models

Important Assumptions:

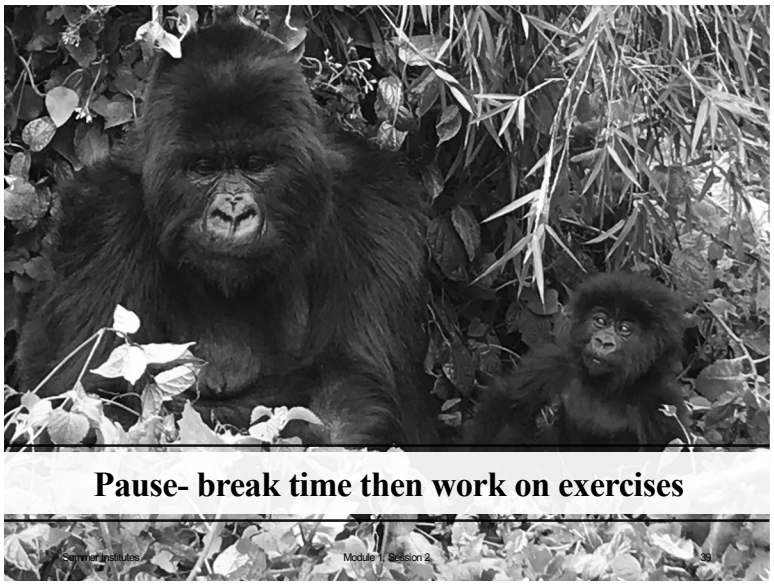
1) Two possible outcomes - success (1) or failure (0) - for each of n trials.

2) The probability of success, p, is the same for each trial.

3) The outcome of one trial has no influence on later outcomes (independent trials).

4) The random variable of interest is the total number of successes.

38



39

Questions 7, 8, 9

Suppose a new student has joined your lab and is learning how to culture cells. Their reference letter says that 25% of the new student's experiments fail. They only have time to create 3 cultures.

7. What's the probability that exactly 1 experiment fails?
8. What's the probability that at least 1 experiment fails?
9. What's the probability that all experiments succeed?

Recall: $P(X = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$

where, e.g., 4! = 4 x 3 x 2 x 1 = 24

40

Mean and Variance of a Discrete Random Variable

Given a **theoretical** probability distribution we can define the **mean and variance of a random variable** which follows that distribution. These concepts are analogous to the summary measures used for samples except that these now describe the value of these summaries in the limit as the sample size goes to infinity (i.e. the **parameters** of the **population**).

Suppose a random variable X can take the values $\{x_1,x_2,\ldots\}$ with probabilities $\{p_1,p_2,\ldots\}$. Then

Mean: $\mu = \mathrm{E}(\mathrm{X}) = \sum_{\mathrm{j}} p_j x_j$

Variance: $\sigma^2 = \mathrm{V}(\mathrm{X}) = \mathrm{E}\left[(\mathrm{X}-\mu)^2\right] = \sum_j p_j (x_j - \mu)^2$

Mean and Variance Example

Consider a Bernoulli random variable with success probability **p**.

P[X=1] = p

P[X=0]=1-p

Mean:
$$\begin{aligned}\mu = E[X] &= \sum_{j=0}^{1} p_j x_j \\ &= (1-p)\times 0 \quad + \quad p\times 1 \\ &= p\end{aligned}$$

Variance:
$$\begin{aligned}\sigma^2 = \mathrm{V}[X] &= \sum_{j=0}^{1} p_j (x_j-\mu)^2 \\ &= (1-p)\times(0-p)^2 \quad + \quad p\times(1-p)^2 \\ &= p(1-p)\end{aligned}$$

Mean and Variance of Binomial Random Variable

Consider a binomial random variable with success probability **p** and sample size **n**.

X ~ Bin(n,p)

Mean:
$$\begin{aligned}\mu = E[X] &= \sum_{j=0}^{n} p_j x_j \\ &= \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} \times j \\ &= \text{???}\end{aligned}$$

Variance:
$$\begin{aligned}\sigma^2 = \mathrm{V}[X] &= \sum_{j=0}^{n} p_j (x_j-\mu)^2 \\ &= \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} \times (j-\mu)^2 \\ &= \text{???}\end{aligned}$$

Means and Variance of the Sum of Independent Random Variables

Recall that a binomial RV is just the sum of **n** independent Bernoulli random variables.

If $X_1, X_2, \ldots, X_n$ are **independent** random variables and if we define $Y= X_1+ X_2+ \ldots+X_n$

1. Means add:

$$E[Y]= E[X_1]+E[X_2]+ \ldots+E[X_n]$$

2. Variances add:

$$V[Y]= V[X_1]+V[X_2]+ \ldots+V[X_n]$$

We can use these results, together with the properties of the mean and variance that we learned earlier, to obtain the mean and variance of a binomial random variable (Question 10 of the exercises).

Binomial Distribution Summary

1. Binomial RVs are discrete.
2. Parameters - **n** (sample size), **p** (probability of success)
3. Sum of n independent 0/1 outcomes
4. Sample proportions, logistic regression

45

46

Questions 10, 11

10. A couple intends to have 5 children and both are carriers of myotonic dystrophy, a dominant trait. Therefore, the probability that a child has the trait is 0.75. What is the probability that at least 1 child will have the trait?
11. Calculate the mean and variance of a binomially distributed random variable with n trials and success probability p.

47