Descriptive Statistics and Exploratory Data Analysis

Descriptive Statistics (Exploratory)

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

  John Tukey
  Exploratory Data Analysis
  Addison-Wesley, 1977

- organization, summarization, and presentation of data
- If you can't see it, don't believe it!

Inferential Statistics (Confirmatory)

- Generalization of conclusions:

  sample ⟶ population

- Assess strength of evidence
- Make comparisons
- Make predictions

Tools:

- Modeling
- Estimation and Confidence Intervals
- Hypothesis Testing

Exploratory vs Confirmatory Data Analysis

Exploratory (Descriptive)

- Detective work
- Open (but directed) mind
- Creative

Confirmatory (Inferential)

- Acting as judge and jury (or at least lawyer)
- Focused on one or a few ideas
- Following principles of inference

Types of Data

- Categorical (qualitative)
  - 1) Nominal scale - no natural order
    - gender, marital status, race
  - 2) Ordinal scale
    - severity scale, good/better/best
- Numerical (quantitative)
  - 1) Discrete - (few) integer values
    - number of children in a family
  - 2) Continuous - measure to arbitrary precision
    - blood pressure, weight

Why bother?

$\Rightarrow$ PROPER DISPLAYS

$\Rightarrow$ PROPER ANALYSIS

Samples

In statistics we usually deal with 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.

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.

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 Σ is shorthand for "*sum*" over a specified range. For example:

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

Some Properties of the Arithmetic Mean

Often we wish to **transform** variables. Linear changes to variables (i.e. Y = a*X+b) impact the mean in a predictable way:

(1) Adding (or subtracting) a constant to all values:

$$Y_i = X_i + c$$
$$\overline{Y} =$$

(2) Multiplication (or division) by a constant:

$$Y_i = cX_i$$
$$\overline{Y} =$$

Does this nice behavior happen for any change? NO! (show that $\log \overline{X} \neq \overline{\log X}$)

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:

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

Mode

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

Comparison of Mean and Median

- Mean is sensitive to a few very large (or small) values - “outliers”
- Median is “resistant” to outliers
- Mean is attractive mathematically
- 50% of sample is above the median,

  50% of sample is below the median.

Variation is important!

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.

Measures of Spread: Variance

Consider the following two samples:

20,23,34,26,30,22,40,38,37

30,29,30,31,32,30,28,30,30

These samples have the same mean and median, but the second is much less variable. The average "distance" from the center is quite small in the second. We use the **variance** to describe this feature:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2$$

$$s^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} X_i^2 - N\overline{X}^2\right)$$

$$s^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} X_i^2 - \left(\sum_{i=1}^{N} X_i\right)^2 / N\right)$$

The standard deviation is simply the square root of the variance:

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

For the first sample, we obtain:

$$\begin{aligned} \overline{X} &= 30 \\ \sum_{i=1}^{9} X_i^2 &= 8574 \end{aligned}$$

$$\begin{aligned} s^2 &= \frac{1}{9-1}\left(8574 - 9 \times 30^2\right) \\ &= \left(8574 - 8100\right)/8 \\ &= 59.25\text{yr}^2 \end{aligned}$$

For the second sample, we obtain:

$$\begin{aligned} \overline{X} &= 30 \\ \sum_{i=1}^{9} X_i^2 &= 8110 \end{aligned}$$

$$\begin{aligned} s^2 &= \frac{1}{9-1}\left(8110 - 9 \times 30^2\right) \\ &= \left(8110 - 8100\right)/8 \\ &= 1.25\text{yr}^2 \end{aligned}$$

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) Add/substract a constant: $Y_i=X_i+c$

$$\begin{aligned} s_Y^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i-\overline{Y}\right)^2 \\ &= \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i+c-(\overline{X}+c)\right)^2 \\ &= s_X^2 \end{aligned}$$

(2) Multiply/divide by a constant: $Y_i=c \times X_i$

$$\begin{aligned} s_Y^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i-\overline{Y}\right)^2 \\ &= \frac{1}{N-1}\sum_{i=1}^{N}\left(cX_i-c\overline{X}\right)^2 \\ &= \frac{1}{N-1}\sum_{i=1}^{N}c^2\left(X_i-\overline{X}\right)^2 \\ &= c^2 \times s_X^2 \end{aligned}$$

So what happens to the standard deviation?

Measures of Spread: Quantiles and Percentiles

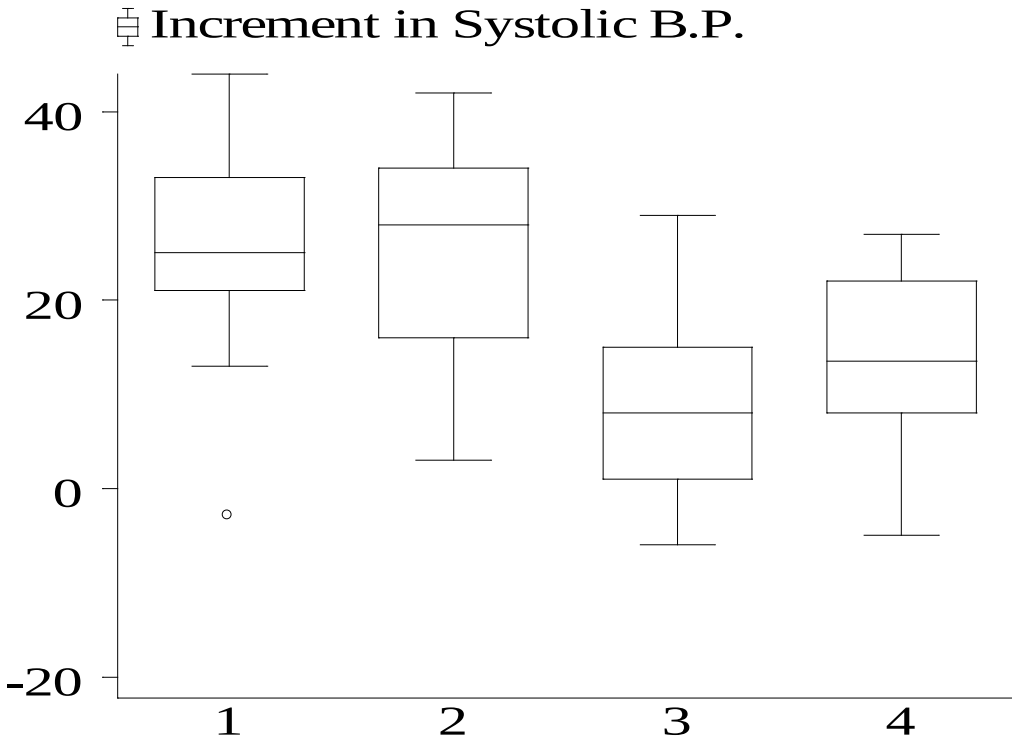
The median was the sample value that had 50% of the data below it.

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.



Summary

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

Probability Distributions

I

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. Note that a probability distribution is an example of the "classical" definition of probability.

Population	⟷	Sample
Random variable	⟷	Measurement
Probability dist.	⟷	Frequency dist.
Parameters	⟷	Statistics (Estimates)

Theoretical Distributions

Used to provide a mathematical description of outcomes of an experiment.

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 measurements are approximately normally distributed.

2. t- distribution

3. Chi-square distribution (χ^2)

Binomial Distribution - Motivation

Question: In a family where both parents are carriers for a recessive trait, what is the probability that in a family of 3 children exactly 1 child would be affected?

What is the probability that at least 1 would be affected?

In a family of 6 children, what is the probability that exactly 1 child is affected?

What if the trait is dominant?

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 affected children in a family of 3.

What we need to know is:

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)?

Combinations

Combinations: number of different arrangements of k objects (successes) taken from a total of n objects (trials) if order doesn't matter.

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

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

E.g.

Child number 1	2	3	Outcomes
+	+	+	3 affected
+	+	-	2 affected
+	-	+	2 affected
-	+	+	2 affected
+	-	-	1 affected
-	+	-	1 affected
-	-	+	1 affected
-	-	-	0 affected

What are the probabilities of these outcomes?

Child number 1	2	3	Outcomes	# ways
p	p	p	3 affected	1
p	p	1-p	2 affected	
p	1-p	p	2 affected	3
1-p	p	p	2 affected	
p	1-p	1-p	1 affected	
1-p	p	1-p	1 affected	3
1-p	1-p	p	1 affected	
1-p	1-p	1-p	0 affected	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:

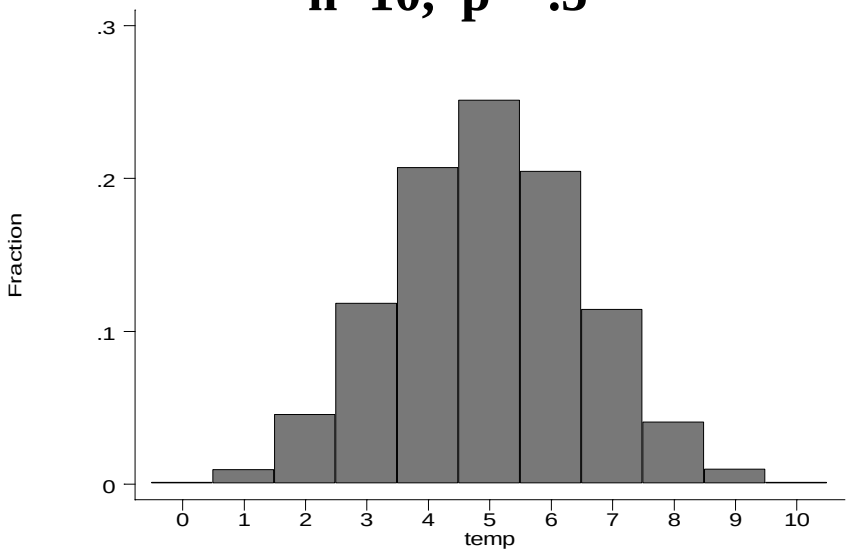
1) Two possible outcomes - success (1) or failure (0) - for each of n trials.

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

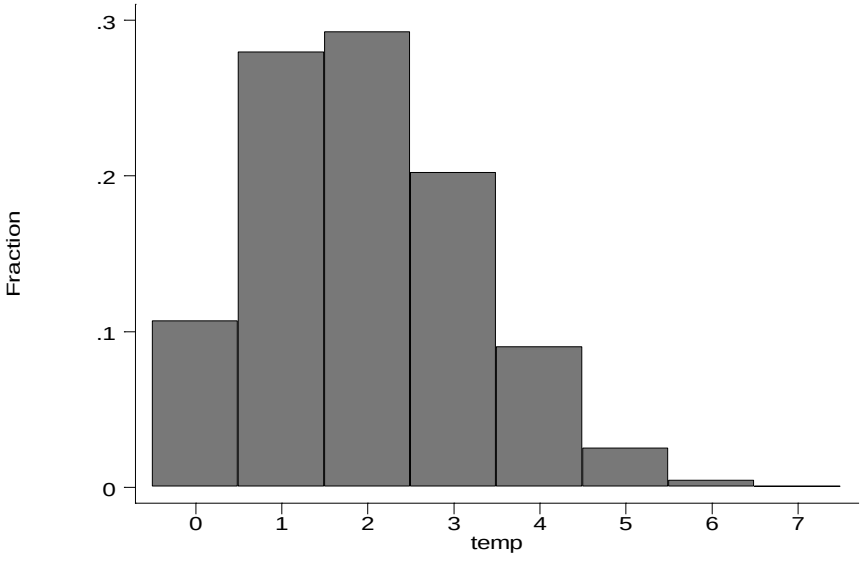
3) The outcome of one trial has no influence on later outcomes (independent trials).

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

n=10, p = .5


Fraction
.3
.2
.1
0
0 1 2 3 4 5 6 7 8 9 10
temp

n=10, p = .2


Fraction
.3
.2
.1
0
0 1 2 3 4 5 6 7
temp

Binomial Probabilities - Example

Returning to the original question: What is the probability of exactly 1 affected child in a family of 3? (recessive trait, carrier parents)

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:

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

Example - Mean and Variance

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 - Binomial

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

$$X \sim bin(n,p)$$

MEAN:

$$\mu = E[X] = \sum_{j=0}^{n} p_j x_j$$

$$= \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} \times j$$

$$= ???$$

VARIANCE:

$$\sigma^2 = V[X] = \sum_{j=0}^{n} p_j (x_j - \mu)^2$$

$$= \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} \times (j-\mu)^2$$

$$= ???$$

Help!

Means and Variance of the Sum of independent RV's

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 (hmwk).

Binomial Distribution Summary

Binomial

1. Discrete, bounded
2. Parameters - **n,p**
3. Sum of n independent 0/1 outcomes
4. Sample proportions, logistic regression