Descriptive Statistics and Exploratory Data Analysis

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, don't believe it!

Inferential/Confirmatory Statistics

- Generalization of conclusions:

  sample ⟶ population

- Assess strength of evidence
- Make comparisons
- Make predictions

Tools:

- Modeling
- Estimation and Confidence Intervals
- Hypothesis Testing

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)

Types of Data

- Categorical (qualitative)
  - 1) Nominal scale - no natural order
    - yes/no, nationality, gender…
  - 2) Ordinal scale - natural order exists
    - good/better/best, low/medium/high…
- Numerical (quantitative)
  - 1) Discrete - (few) integer values
    - number of children in a family
  - 2) Continuous - measure to arbitrary precision
    - blood pressure, weight

Different types of data demand different analysis and graphics tools

Think: Categorise zip code

QUIZ

(2 mins; complete in pairs)

- Categorise the following variables into nominal, ordinal, discrete, or continuous
    1. Time since you were born
    2. Age measured in years
    3. Price of your lunch
    4. Zipcode of your residence

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.

e.g. $X_1 = 60$, $X_2 = 33$, $X_3 = 41$

THE ABSTRACTION MONSTER

helps us deal with lots

of different settings at once

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.

$X_1= 60, X_2=33, X_3=41$

$X_{(1)}= 33, X_{(2)}=41, X_{(3)}=60$

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

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

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

QUIZ

(2 mins; complete in pairs)

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

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

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

QUIZ: part 2

(2 mins; complete in pairs)

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

2. 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?

3. 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:

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

Mode

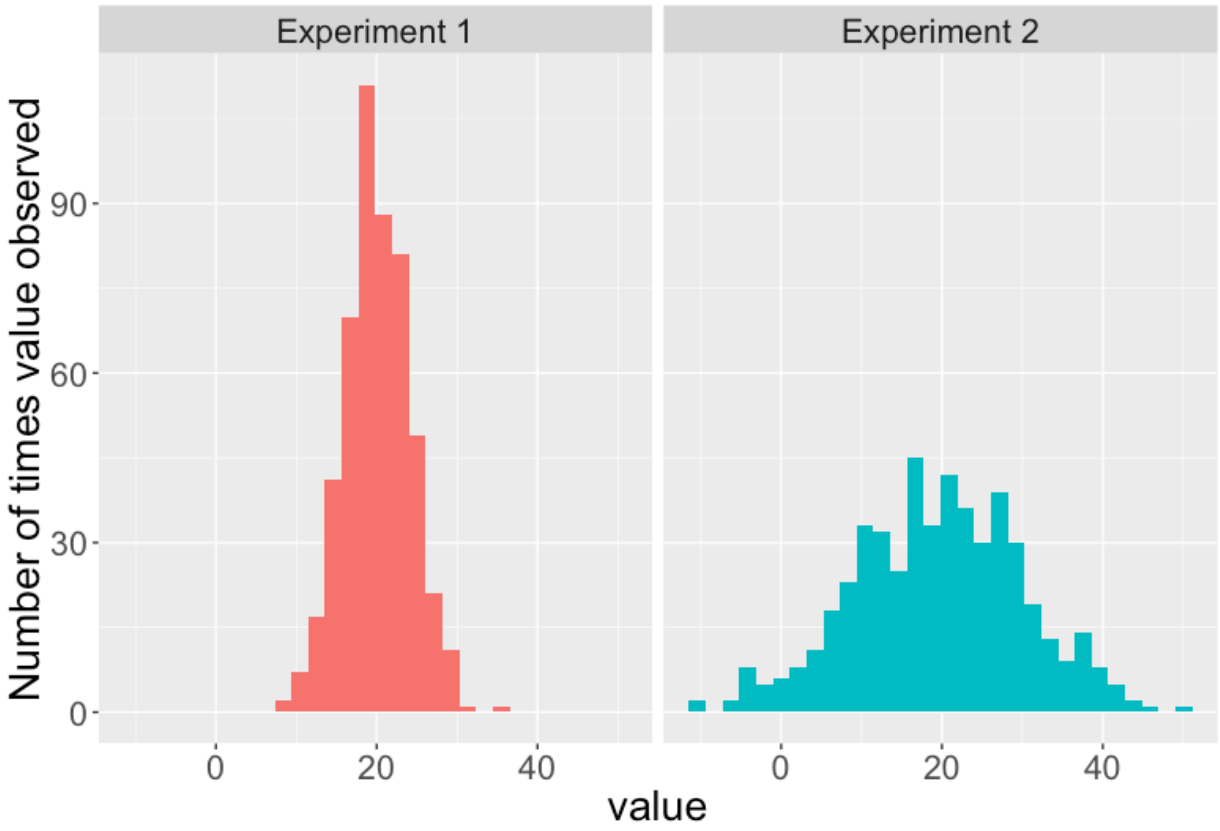
The **mode** is the most frequently occurring value in the sample.

Comparison of Mean and Median

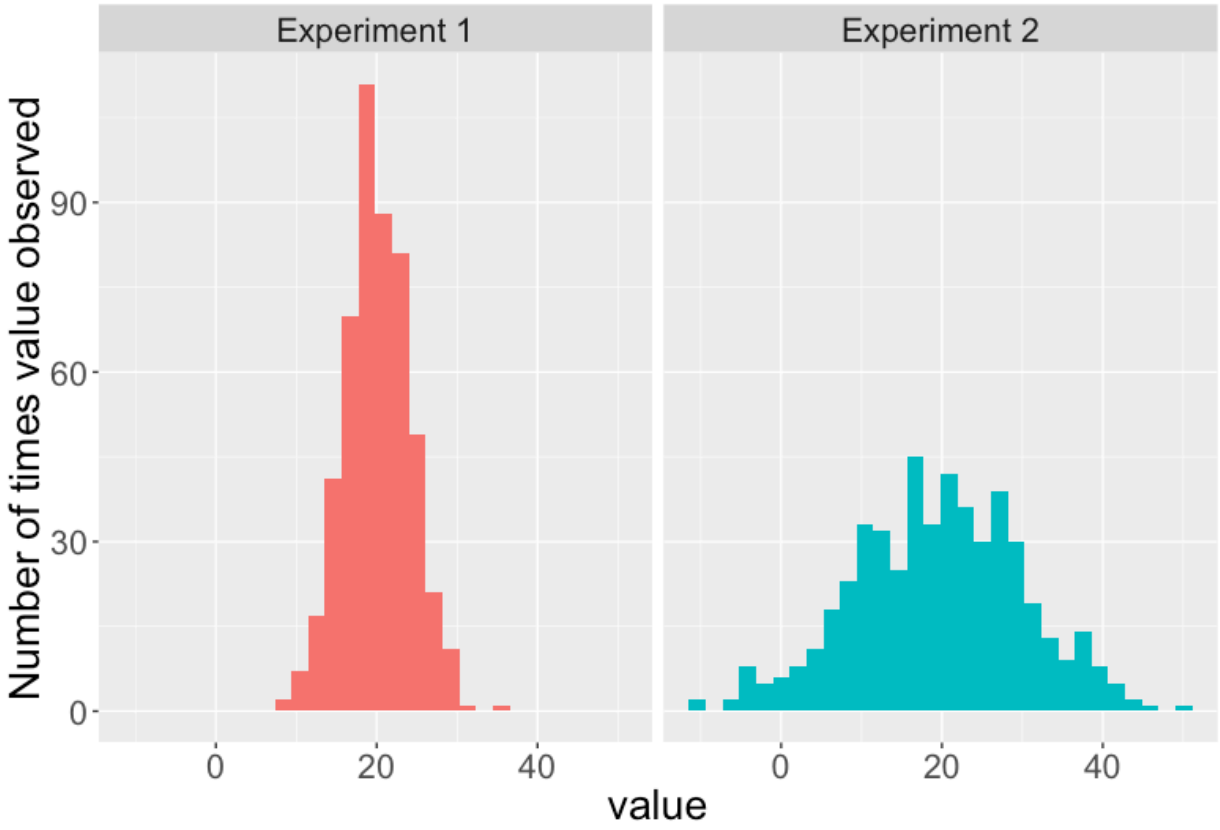
- Mean is sensitive to a few very large (or small) values - “outliers”
- Median is “resistant” to outliers
- Mean is attractive mathematically
- 50% of sample is above the median,

  50% of sample is below the median.

What's the difference?



What's the difference?



- Variance (also called spread) is how we assess relativity in statistics

Measures of Spread: Range

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

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

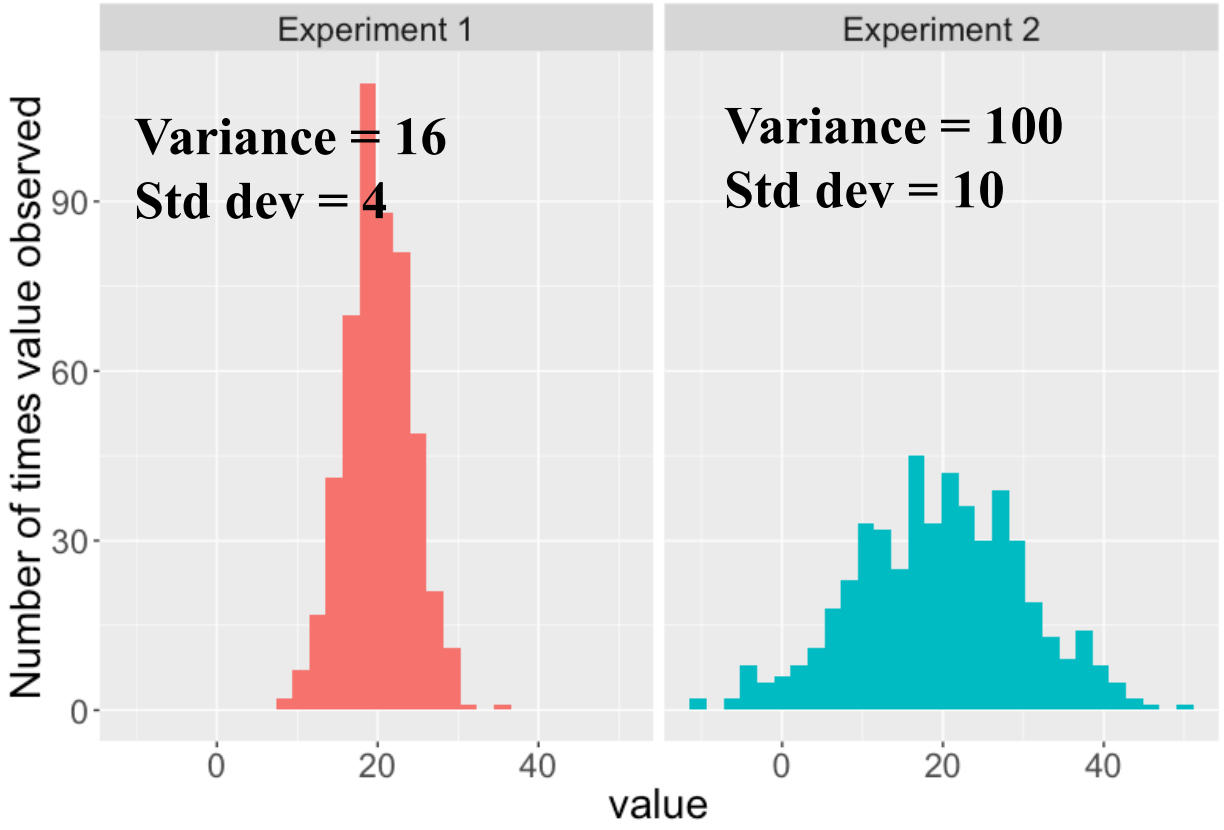
Alternatively, the range may be denoted as the pair of observations:

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

The latter form is useful for data quality control.

Disadvantage: the sample range increases with increasing sample size.

Measures of Spread: Variance



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

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

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

QUIZ

(2 mins; complete in pairs)

1. 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)

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

  Quiz: If the variance in metres was $1m^2$, what's the variance in centimetres?

Measures of Spread: Quantiles and Percentiles

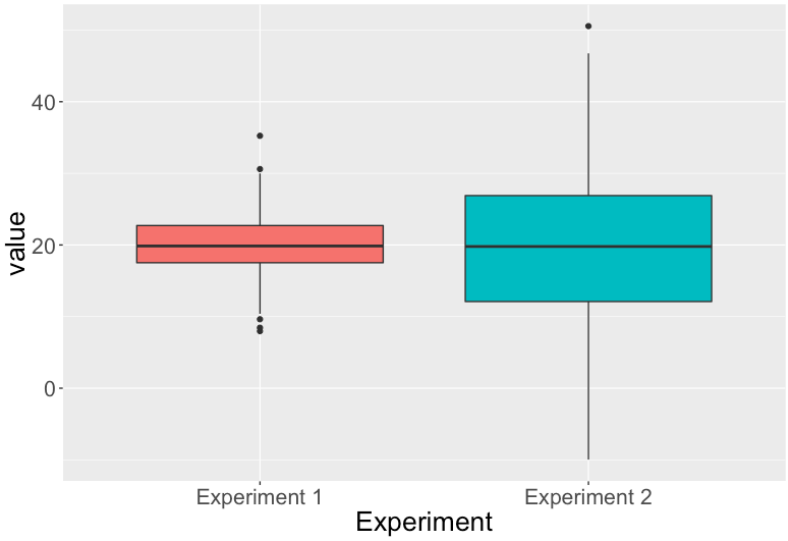
The median was the sample value that had 50% of the data below it.

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

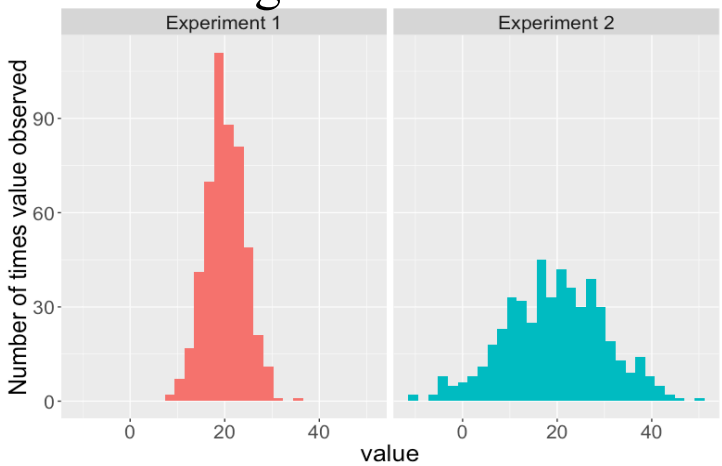
Quartiles are the (25,50,75) percentiles. The **interquartile range** is $Q_{.75}$-$Q_{.25}$ and is another useful measure of spread. The middle 50% of the data is found between $Q_{.25}$ and is $Q_{.75}$.

Boxplot

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



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

Probability Distributions

I

Probability: Why bother?

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 generalise to a broader population, we need to know how reliable/representative/variable our samples were.

Population	⟷	Sample
Probability dist.	⟷	Frequency dist.
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

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

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)?

Binomial Random Variable

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

Experiments succeeding

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 out of 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.

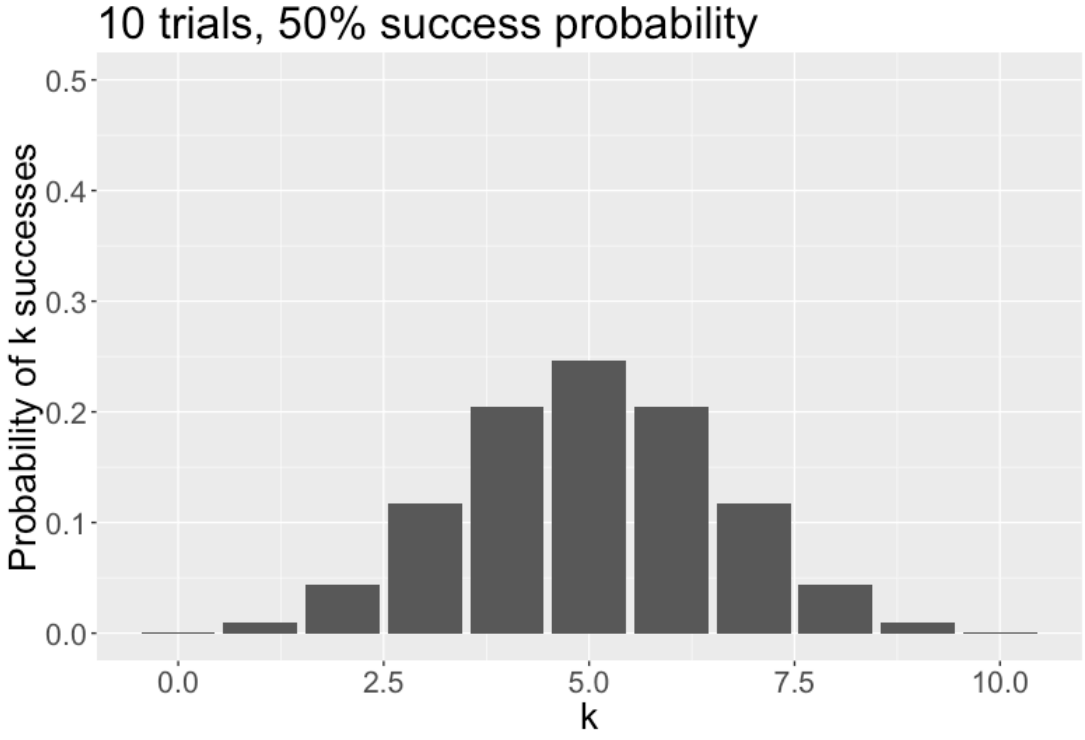
Assumptions:

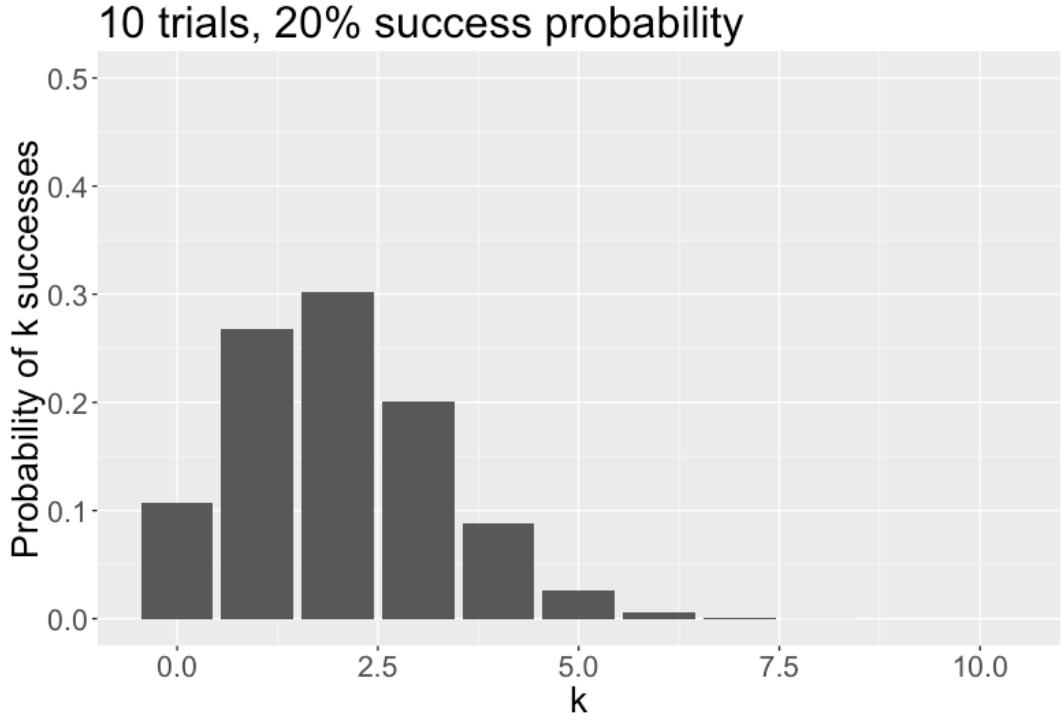
1) Two possible outcomes - success (1) or failure (0) - for each of n trials.

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.


10 trials, 50% success probability
Probability of k successes
0.5
0.4
0.3
0.2
0.1
0.0
0.0
2.5
5.0
7.5
10.0
k


10 trials, 20% success probability
Probability of k successes
0.5
0.4
0.3
0.2
0.1
0.0
0.0
2.5
5.0
7.5
10.0
k

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.

Quiz: 6 mins

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 exactly 1 experiment fails?
- What's the probability that at least 1 experiment fails?
- 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

Quiz Qtn 1: solution

- What's the probability that exactly 1 experiment fails?

- X = number of failures

$$X \sim Bin(n = 3, p = 1/4)$$

$$Pr(X = 1) = \binom{3}{1} \times 0.25^1 \times 0.75^2$$

$$= 0.421875$$

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**.

$$X \sim \text{bin}(n,p)$$

MEAN:

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

VARIANCE:

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

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 (Exercise 3).

Binomial Distribution Summary

Binomial

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

Exercises

1. The current powerball jackpot is $140 million, and your probability of winning it is 1 in 175 million. If it costs $2 to play, what is your expected payoff?
2. A couple intends to have 5 children and both are carriers of myotonic dystrophy, a dominant trait. What is the probability that at least 1 child will have the trait?
3. Calculate the mean and variance of a binomially distributed random variable with n trials and success probability p.

Ex 1. Solution

X = Powerball payoff in dollars

There are 2 possible values for X: X= 140000000-2 (which occurs with probability 1/175000000), and X= -2 (occurs with probability 1-1/175000000).

EX = (140000000-2)*1/175000000 +

-2*(1-1/175000000) = -1.2

So the expected payoff is a loss of $1.20

Ex 2. Solution

The probability of any single child having the trait is 0.75, and the carrier status of each child is independent of every other. The number of children with the trait (X) is therefore a binomially-distributed random variable with n = 5 and p = 0.75.

$$
\begin{aligned}
&Pr(X = 1 \text{ or more}) \\
&= 1\text{–}Pr(X = 0) \\
&= 1 - \binom{5}{0} \times 0.75^0 \times 0.25^5 \\
&= 1 - 1 \times 1 \times 0.0009765625 \\
&= 0.9990
\end{aligned}
$$

Ex 3. Solution

If X ~ Bin(n,p) and Y1, Y2…Yn are independent Bernoulli random variables with success probability p, then X has the same distribution as Y1 + Y2 + … + Yn. So

$$\begin{aligned} EX &= E(Y_1 + Y_2 + \ldots + Y_n) \\ &= EY_1 + EY_2 + \ldots + EY_n \\ &= p + p + \ldots + p \\ &= np \\ VarX &= Var(Y_1 + Y_2 + \ldots + Y_n) \\ &= VarY_1 + VarY_2 + \ldots + VarY_n \\ &= p(1-p) + p(1-p) + \ldots + p(1-p) \\ &= np(1-p) \end{aligned}$$