Descriptive Statistics and Exploratory Data Analysis

Session 2

Module 1 Probability & Statistical Inference

The Summer Institutes
DEPARTMENT OF BIOSTATISTICS
SCHOOL OF PUBLIC HEALTH
UNIVERSITY of WASHINGTON

Why do we need statistics?

To:

- answer scientific questions
- analyze data from a sample to make generalizations back to the population of interest, to gain insight, reach the truth

Not only to:

- state statistical significance
- quantify findings but to interpret these and make the right conclusions.

Science is about uncovering the truth - statistics provides the framework.

Exploratory vs. Inferential Data Analysis

Form idea/hypothesis

Investigate predefined idea/hypothesis

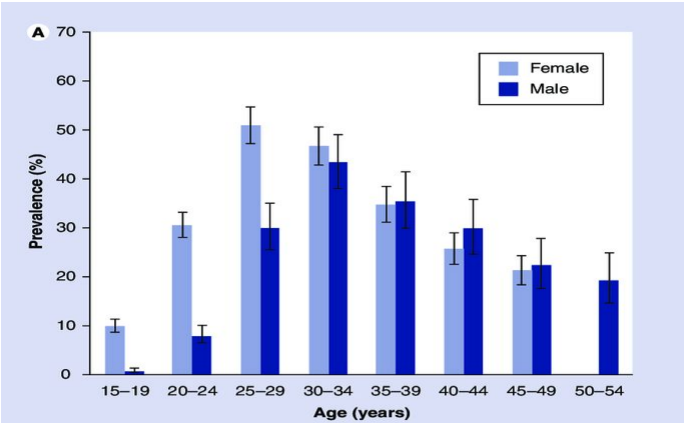
Why Descriptive Statistics First?

All statistical analyses should start with a basic summary and presentation of the data.

Generally a first step towards scientific discovery

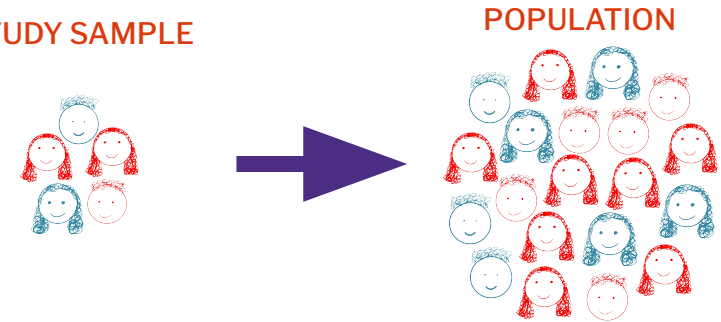
Definitely a first steps towards scientific understanding

"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"



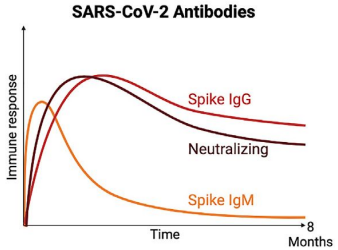
Inferential Statistics

Provides a framework for moving from statements about the sample to the population:



- Assess strength of evidence
- Make comparisons
- Make predictions

Tools:



Cohen, Lindermann, Moodie, et al. Cell Reports Medicine 2021

- Modeling
- Estimation and Confidence Intervals: mean = 25 with 95% CI = (20.5, 24.5)
- Hypothesis Testing: p-value < 0.001

Descriptive and Inferential Statistics Depends on Data Type

1. **Categorical/qualitative data**
   - Nominal scale - no natural order: e.g., sex at birth, gender identity, …
   - Ordinal scale - natural order exists: e.g., low/medium/high, BMI categories …

2. **Numerical/quantitative data**
   - Discrete - (few) integer values: e.g., number of children in a family
   - Continuous - measured to a given level of precision: e.g., blood pressure, weight

Different types of data require different analysis and graphics tools

e.g., categorize zip code

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 data point, X_2 is the second, etc.

e.g. Ages of 3 people:

$X_1 = 60$ $X_2 = 33$ $X_3 = 15$

Order Statistics

Sometimes it is useful to order the measurements:

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

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

Calculate order statistics of age example:

Data		Order Statistics
$X_1 = 60$	→	$X_{(1)} = 15$
$X_2 = 33$		$X_{(2)} = 33$
$X_3 = 15$		$X_{(3)} = 60$

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:

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

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

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

2, 3, 4, 8, 10

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

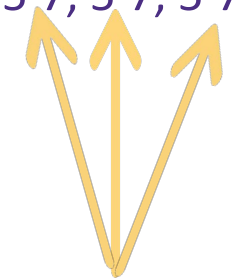
2, 3, 4, 6, 8, 10

(4+6)/2=5

Mode

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

5’5, 5’7, 5’7, 5’7



Mean vs. Median

Mean is sensitive to a few very large (or small) values - “outliers”:

mean of (3, 3, 4, 5, 100) = 23

Median is not sensitive to outliers:

median of (3, 3, 4, 5, 100) = 4

Mean has useful mathematical properties

Pause- break time then work on exercises 1-5

Two Properties of the Mean

Often we wish to transform a variable X (more on this later).

Linear changes to the variable X impact the mean $\overline{X}$ in a predictable way:

1. Adding a constant to all values adds the same constant to the mean.

   ((2+10) + (4+10) + (6+10))/3 = 14

   (2 + 4 + 6)/3 + 10 = 14

2. Multiplication by constant multiplies the mean by the same constant.

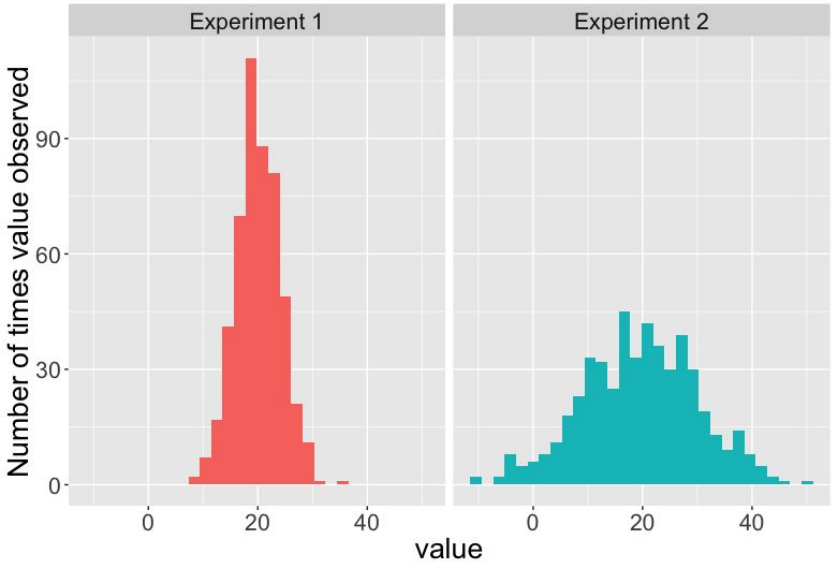
   ((2*10) + (4*10) + (6*10))/3 = 40

   (2 + 4 + 6)/3 * 10 = 40

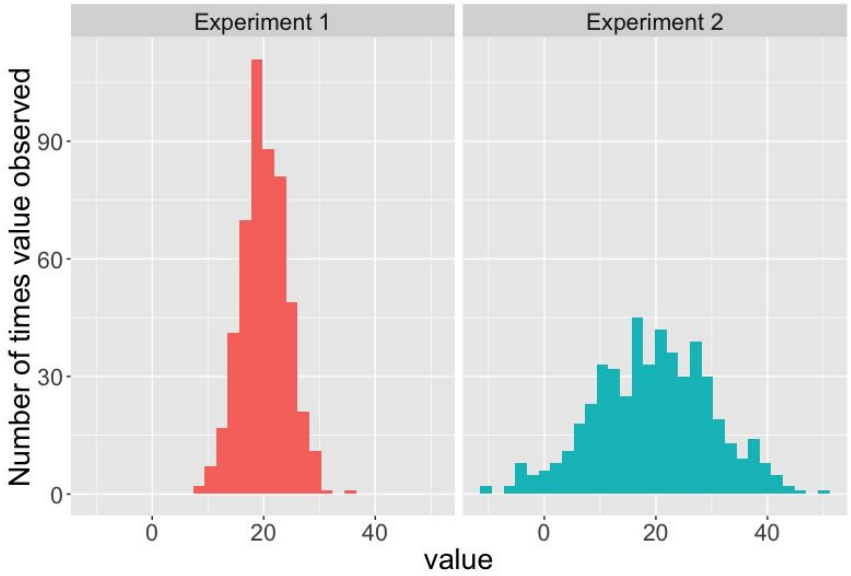
CAUTION: This does not happen for *all* transformations (only linear ones).

e.g., log($\overline{X}$) ≠ mean(log(X))

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



Same mean but different variance



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

Measures of Spread: Range

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

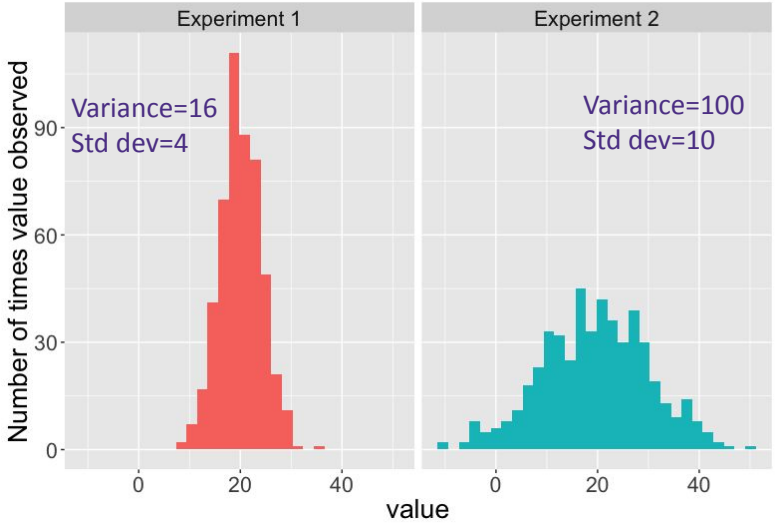
$$\begin{aligned} \text{Range} &= \text{Maximum - Minimum} \\ &= X_{(n)} - X_{(1)} \end{aligned}$$

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

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

Note: the sample range increases with increasing sample size.

Measures of Spread: Variance



Variance is most common summary of spread.

Measure of the distance from each observation to the center of the observations:

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

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

Properties of the variance/standard deviation

Variance and standard deviation always >= 0.

Linear changes are different than changes to the mean:

1. Adding a constant to all values **does not change** the variance or standard deviation.

2. Multiplying by a constant multiplies the standard deviation by that constant.

3. Multiplying by a constant multiplies the variance by that constant-squared.

Measures of Spread: Percentiles and Quartiles

The median is the 50^{th} percentile of the sample data. That is, 50% of the sample data fall below the median.

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

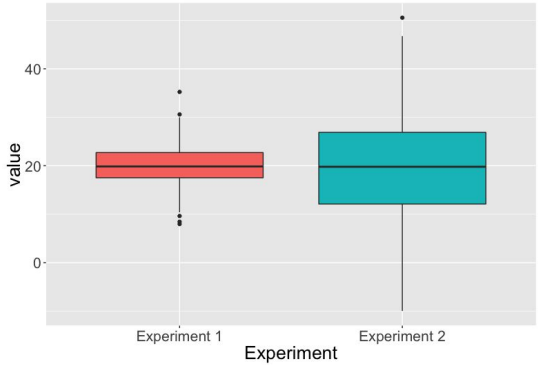
Quartiles are the (25^{th}, 50^{th}, 75^{th}) percentiles.

The **interquartile range** is $Q_{.75}$-$Q_{.25}$ and is another useful measure of spread.

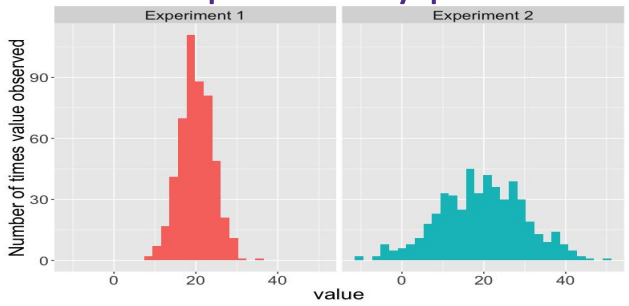
The middle 50% of the data is found between $Q_{.25}$ and is $Q_{.75}$.

Boxplots

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:

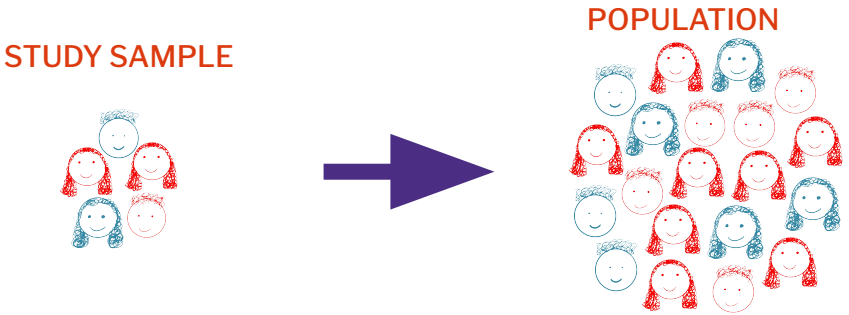


Pause- break time then work on exercise 6

Probability Distributions Part I

Probability: How does it help answer questions?

Most of the time we are not interested in the sample 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.

Sample	⟷	Population
Frequency distribution	⟷	Probability distribution
Estimates	⟷	Parameters

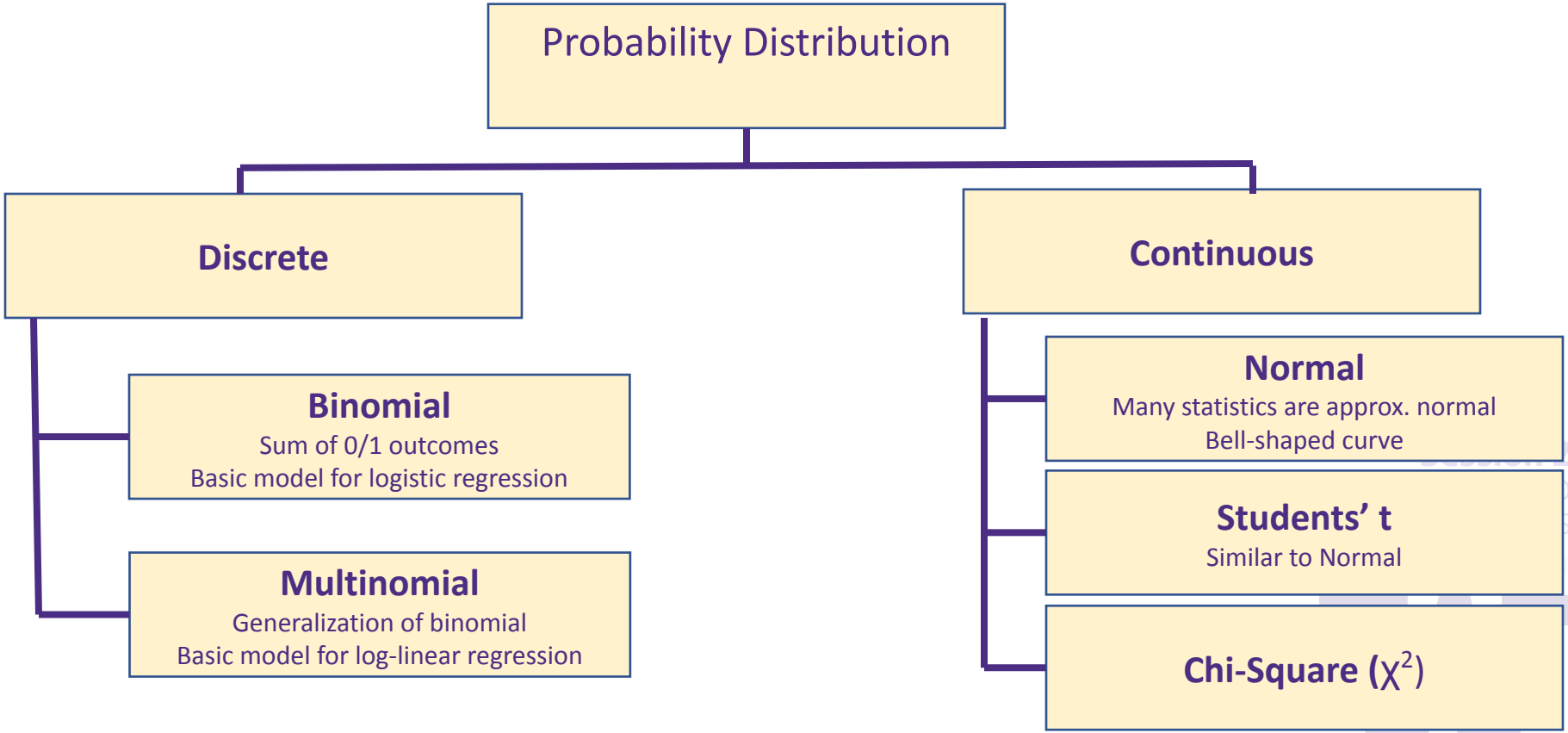
Probability Distribution Definitions

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

A **probability distribution P(X)** gives the probability of obtaining all possible (sets of) values of a random variable X.

It gives the probability of the outcomes of an experiment.

Theoretical Probability Distributions



Binomial Distribution – Example

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.

The binomial distribution helps calculate the following probabilities:

- all experiments succeed
- only 1 experiment fails
- at least 1 experiment fails

First, define concept of a 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 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 = 1
Experiment fails = 0

Experiment number			
1	2	3	Outcomes
1	1	1	3 successes
1	1	0	2 successes
1	0	1	2 successes
0	1	1	2 successes
1	0	0	1 success
0	1	0	1 success
0	0	1	1 success
0	0	0	0 success

Calculating Combinations

The general formula:

$\mathbf{C_k^n}$ is the number of ways to get k successes in n attempts:

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

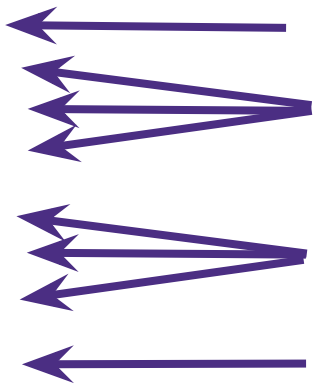
where

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

What are the probabilities of 0, 1, 2, 3 successes?

Experiment number

1	2	3	Outcome	No. of ways
p	p	p	3 successes	1
p	p	1-p	2 successes	
p	1-p	p	2 successes	3
1-p	p	p	2 successes	
p	1-p	1-p	1 successes	
1-p	p	1-p	1 successes	3
1-p	1-p	p	1 successes	
1-p	1-p	1-p	0 successes	1



Sequence of k successes (k=0, 1, 2, or 3) and (3-k) failures will have probability:

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

and there are $\frac{3!}{k!(3-k)!}$ such sequences.

Binomial Probabilities

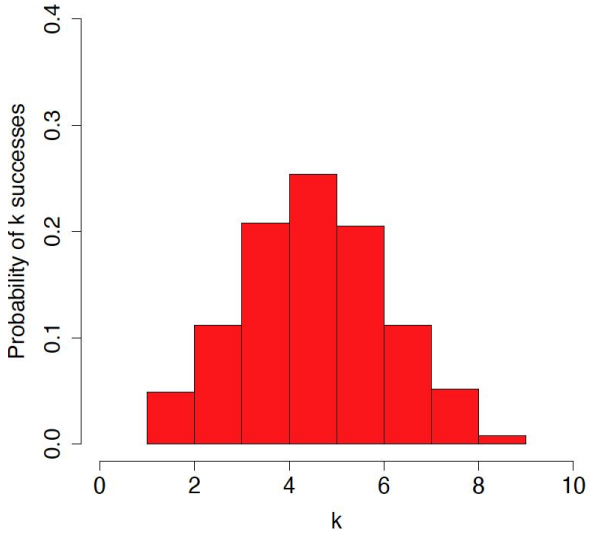
The probability that a binomial random variable with n trials and success probability p will yield exactly k successes is:

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

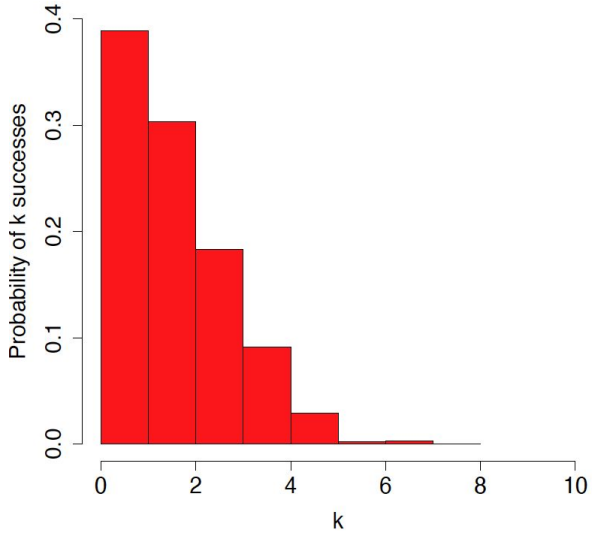
This formula is called the **probability mass function** for the binomial distribution.

Shape of distribution depends on success probability p and number of trials, n

10 trials, 50% success probability



10 trials, 20% success probability



Binomial Model

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 other trials (independent trials).

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

Pause- break time then work on exercises 7-9

Mean and Variance of a Discrete Random Variable

If a random variable is known to follow a theoretical probability distribution, its mean and variance is defined by that distribution.

These concepts are analogous to the sample mean and variance except that these now describe their value 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 a Binomial Random Variable

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

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

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:

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

Properties of Independent Random Variables Simplify the Calculations

Recall that a binomial random variable 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 can be summed:

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

2. Variances also:

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

Binomial Distribution Summary

1. Binomial random variables are discrete.
2. Parameters: n (sample size), p (probability of success)
3. Sum of n independent 0/1 outcomes
4. Seen in sample proportions, logistic regression

Now work on exercises 10-11 (end of Session 2)