Probability Distributions

Special distributions and visualizing probabilities

Dr Andrew Mitchell

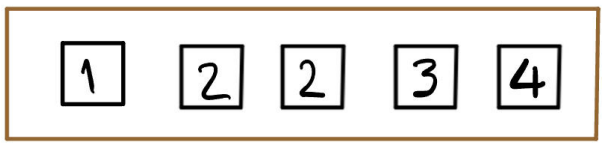
a.j.mitchell@ucl.ac.uk

Lecturer in AI and Machine Learning for Sustainable Construction

2025-01-30

Part 3: Probability Distributions

Special distributions and visualizing probabilities



Box Contents:

- One ticket marked 1
- Two tickets marked 2
- One ticket marked 3
- One ticket marked 4

Probability Distribution:

Outcome	1	2	3	4
Probability	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

- Shows how total probability (100%) is distributed

Special distributions and visualizing probabilities

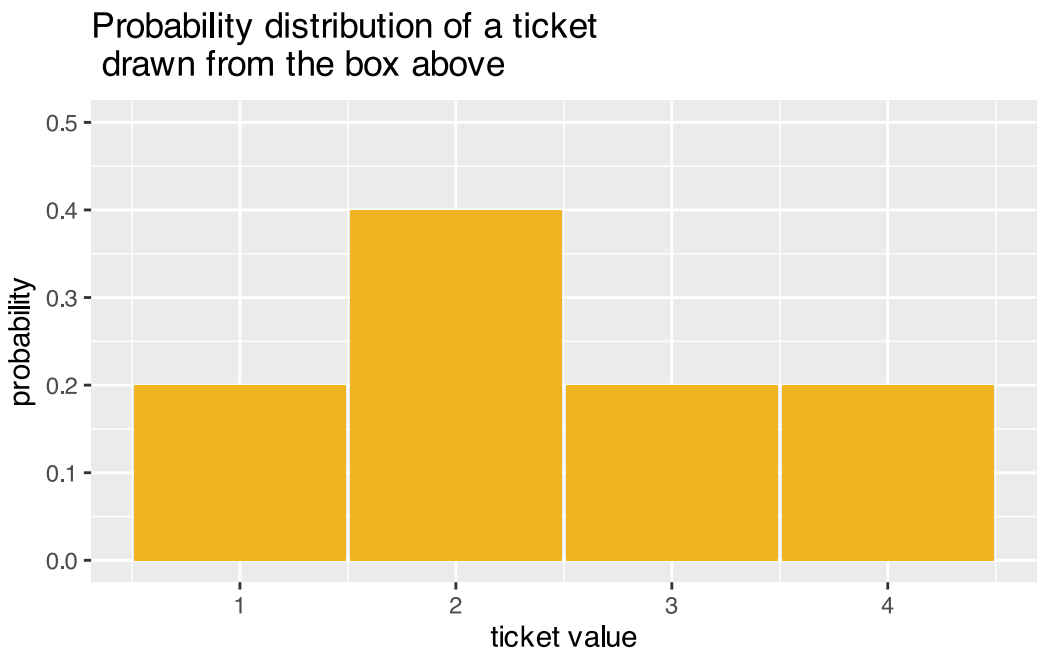
- Each outcome has an associated probability
- Sum of all probabilities equals 1

So far we have seen examples of outcome spaces, and descriptions of how we might compute probabilities, along with tabular representations of the probabilities. In this set of notes, we are going to talk about how to visualize probabilities using tables and histograms, as well as how to visualize simulations of outcomes from actions such as tossing coins or rolling dice.

If we draw one ticket at random from this box, we know that the probabilities of the four distinct outcomes can be listed in a table. What we have described in the table above is a probability distribution. We have shown how the total probability of one or 100% is distributed among all the possible outcomes. Since the ticket marked 2 is twice as likely as any of the other outcomes, it gets twice as much of the probability.

Visualizing Probability Distributions

Probability Histogram



Key Features

- Heights represent probabilities

Visualizing Probability Distributions

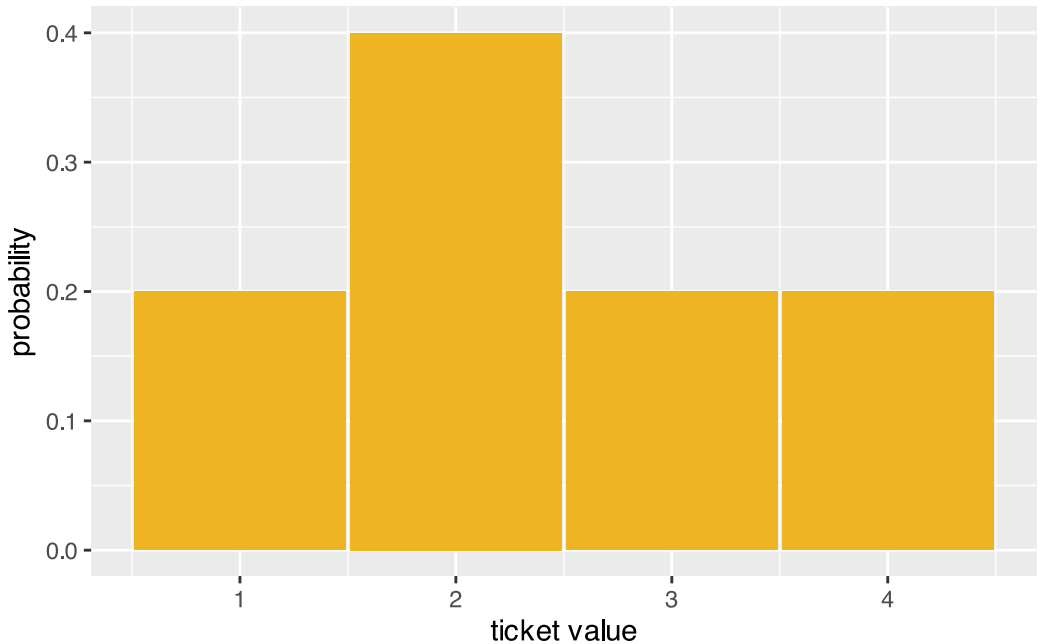
- Area of bars = probability
- Total area = 1
- Theoretical distribution
- No data collection needed

Code Example

```
tkts_box <- c(1, 2, 3, 4)
prob_box <- c(1 / 5, 2 / 5, 1 / 5, 1 / 5)

data.frame(tkts_box) |>
  ggplot(aes(x = tkts_box, y = prob_box)) +
  geom_col(width = 0.98, fill = "goldenrod2") +
  labs(x = "ticket value", y = "probability")
```

Visualizing Probability Distributions



A table is nice, but a visual representation would be even better. We have represented the distribution in the form of a histogram, with the areas of the bars representing probabilities. Notice that this histogram is different from the ones we have seen before, since we didn't

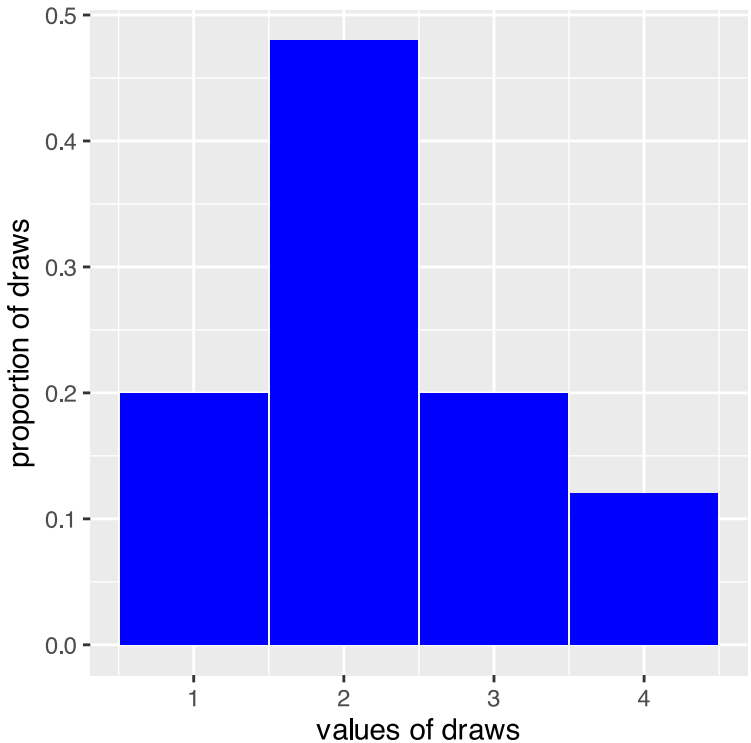
collect any data. We just defined the probabilities based on the outcomes, and then drew bars with the heights being the probabilities. This type of theoretical histogram is called a probability histogram.

Empirical vs Theoretical Distributions

Empirical Distribution (50 draws)

```
Warning: The dot-dot notation (`..prop..`) was deprecated in ggplot2 3.4.0.
ℹ Please use `after_stat(prop)` instead.
```

Empirical vs Theoretical Distributions



Results:

Empirical vs Theoretical Distributions

Ticket	Proportion
1	0.2
2	0.48
3	0.2
4	0.12

Key Differences:

1. Theoretical Distribution:
   - Based on probability model
   - Perfect proportions
   - No randomness
2. Empirical Distribution:
   - Based on actual data
   - Varies with each sample

Empirical vs Theoretical Distributions

- Approaches theoretical as $n \to \infty$

3. Long-run Behavior:
   - Empirical proportions converge
   - Law of Large Numbers
   - Basis for simulation

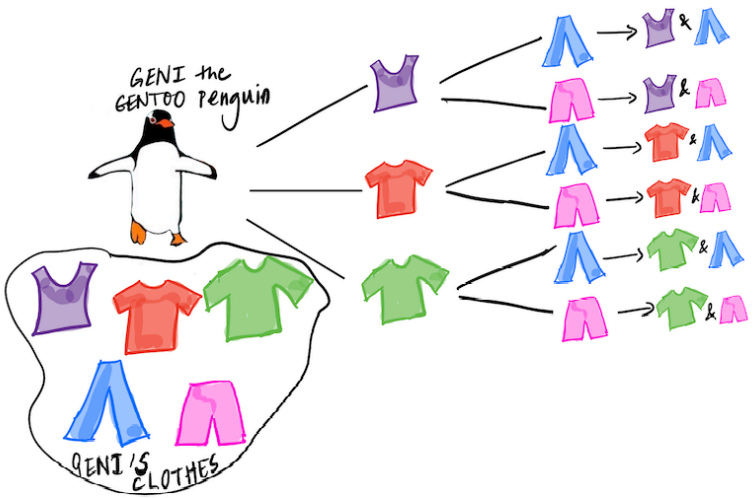
What about if we don't know the probability distribution of the outcomes of an experiment? For example, what if we didn't know how to compute the probability distribution above? What could we do to get an idea of what the probabilities might be? Well, we could keep drawing tickets over and over again from the box, with replacement (that is, we put the selected tickets back before choosing again), keep track of the tickets we draw, and make a histogram of our results. This kind of histogram, which is the kind we have seen before, is a visual representation of data, and is called an empirical histogram.

Empirical vs Theoretical Distributions

On the x-axis of this histogram, we have the ticket values; on the y-axis, we have the proportion of times that this ticket was selected out of the 50 with-replacement draws we took. We can see that the sample proportions looks similar to the values given by the probability distribution, but there are some differences. For example, we appear to have drawn more 3s and less 4s than what was to be expected.

What we have seen here is how when we draw at random, we get a sample that resembles the population, that is, a representative sample, but it isn't exactly the true probabilities. If we increase our sample, however, say to 500, we will get something that more closely aligns with the truth.

Basic Counting Rule



Example:

- 3 t-shirts
- 2 pairs of pants
- Total outfits = $3 \cdot 2 = 6$

General Rule:

Basic Counting Rule

For n steps where: - Step 1 has k_1 outcomes - Step 2 has k_2 outcomes - …and so on

Total outcomes = $k_1 k_2 k_n $

Applications:

- Drawing tickets
- Rolling dice
- Selecting committees

If we have multiple steps (say n) of some action, such that the jth step has k_j outcomes; then the total number of outcomes is $k_1 \times k_2 \times \ldots \times k_n$ and is obtained by multiplying the number of outcomes at each step. This principle is illustrated in the picture: Geni the Gentoo penguin is trying to count how many outfits they have, if each outfit consists of a t-shirt and a pair of pants. The tree diagram shows the number of possible outfits Geni can wear. In this example,

Basic Counting Rule

n=2, $k_1 = 3$, and $k_2 = 2$, since Geni has three t-shirts to choose from, and for each t-shirt, they have two pairs of pants, leading to a total of $3 \times 2 = 6$ outfits.

This example seems trivial, but it illustrates the basic principle of counting: we get the total number of possible outcomes of an action that has multiple steps, by multiplying together the number of outcomes for each step. All the counting that follows in our notes applies this rule.

Permutations and Combinations

Permutations:

Order matters!

$\frac{n!}{(n-k)!}$

Example:

- Drawing letters C,R,A,T,E
- 3 letters, order matters
- $5 \times 4 \times 3 = 60$ outcomes

Combinations:

Order doesn't matter!

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

Permutations and Combinations

Example:

- Selecting committee members
- 5 people, choose 3
- $\binom{5}{3} = 10$ possibilities

Tip

When to use which?

- Permutations: Arrangements, sequences, ordered selections
- Combinations: Groups, teams, unordered selections

When we draw k items from n items without replacement, we have two cases: either we care in what order we draw the k items and the different arrangements of the same set of k items

Permutations and Combinations

have to be counted separately. The number of such arrangements is called the permutations of n things taken k at a time.

For example, let's suppose we are drawing tickets from a box which has tickets marked with the letters C, R, A, T, E, and say we draw three letters with replacement. How many possible sequences of three letters can we get? Using the counting rule, since we have 5 choices for the first letter, 5 for the second, and 5 for the third, we will have a total of $5 \times 5 \times 5 = 125$ possible words, allowing for repeated letters.

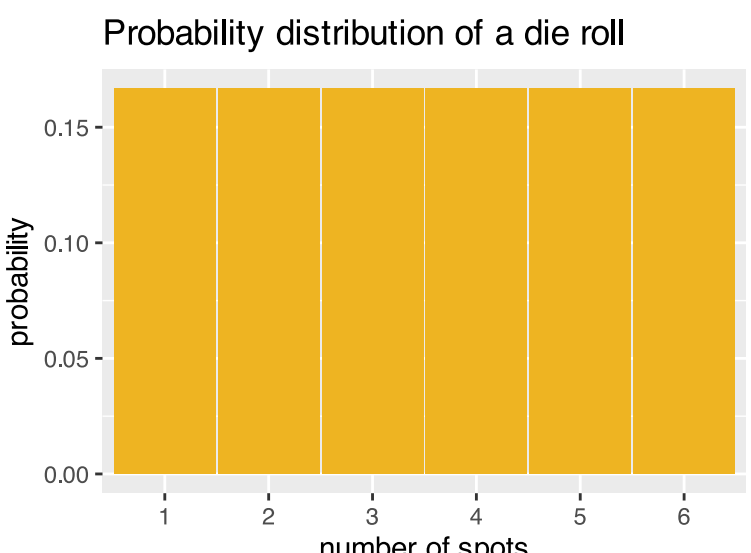
The number of combinations of n things taken k at a time is just the number of distinct arrangements or permutations of n things taken k at a time divided by the number of arrangements of k things. It is denoted by (n k), which is read as "n choose k".

Special Distributions: Discrete Uniform

Definition

- All outcomes equally likely
- Probability = 1/n for each outcome
- Parameter: n (number of outcomes)

Example: Fair Die



Special Distributions: Discrete Uniform

Simulation

```
set.seed(123)
die <- 1:6
rolls <- sample(die, size=1000, replace=TRUE)
table(rolls)/1000
```

```
rolls
    1     2     3     4     5     6
0.170 0.176 0.171 0.157 0.162 0.164
```

This is the probability distribution over the numbers 1, 2, 3, ..., n. We have seen it for dice above. This probability distribution is called the discrete uniform probability distribution, since each possible outcome has the same probability, that is, 1/n. We call n the parameter of the discrete uniform distribution.

Special Distributions: Discrete Uniform

The simulation shows how the empirical distribution approaches the theoretical distribution with a large number of trials. Each outcome occurs approximately 1/6 of the time in our 1000 rolls.

Special Distributions: Bernoulli

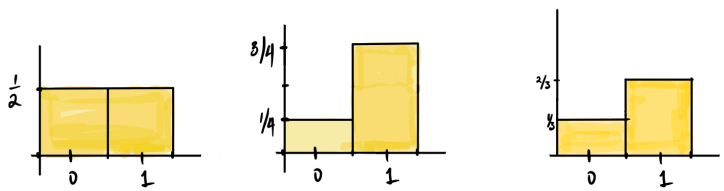
Properties:

- Binary outcomes (Success/Failure)
- One trial only
- Parameter p = P(Success)
- P(Failure) = 1 - p

Examples:

- Single coin flip
- Pass/Fail test
- Win/Lose game

Special Distributions: Bernoulli



Three Bernoulli distributions:

- p = 1/2 (fair coin)
- p = 3/4 (biased)
- p = 2/3 (biased)

This is a probability distribution describing the probabilities associated with binary outcomes that result from one action, such as one coin toss that can either land Heads or Tails. We can represent the action as drawing one ticket from a box with tickets marked 1 or 0, where the probability of 1 is p, and therefore, the probability of 0 is (1-p). We have already seen some examples of probability histograms for this distribution. We usually think of the possible

Special Distributions: Bernoulli

outcomes of a Bernoulli distribution as success and failure, and represent a success by 1 and a failure by 0.

For the Bernoulli distribution, our parameter is p = P(1). If we know p, we also know the probability of drawing a ticket marked 0.

In the figure, the first histogram is for a Bernoulli distribution with parameter p = 1/2, the second p=3/4, and the third has p = 2/3.

Special Distributions: Binomial

Definition

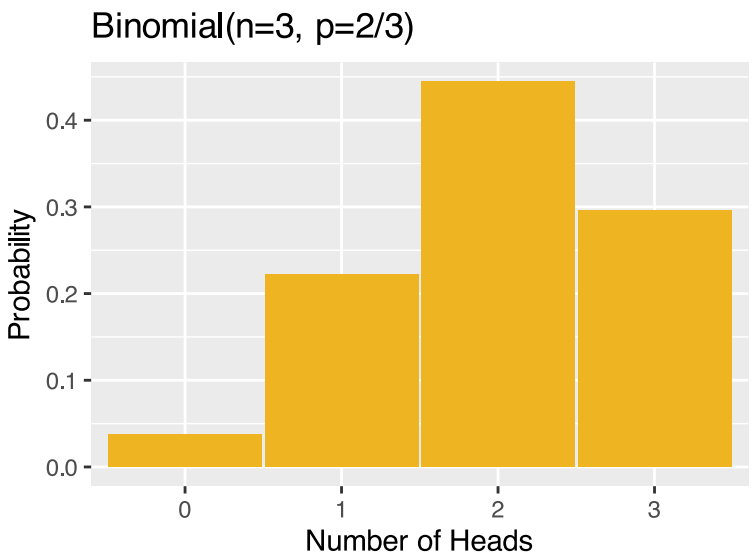
Properties:

- n independent Bernoulli trials
- Each trial has probability p of success
- Count total number of successes
- Parameters: n (trials) and p (probability)

Formula: $P(k \text{ successes}) = \binom{n}{k} p^k (1-p)^{n-k}$

Special Distributions: Binomial

Example: Biased Coin



Simulation

```
set.seed(123)
n <- 3
```

Special Distributions: Binomial

```
p <- 2 / 3
trials <- 1000
results <- replicate(trials, sum(rbinom(n, 1, p)))
table(results) / trials
```

```
results
    0     1     2     3
0.034 0.213 0.465 0.288
```

The binomial distribution, which describes the total number of successes in a sequence of n independent Bernoulli trials, is one of the most important probability distributions. For example, consider the outcomes from tossing a coin n times and counting the total number of heads across all n tosses, where the probability of heads on each toss is p. Each toss is one Bernoulli trial, where a success would be the coin landing heads.

Special Distributions: Binomial

The multiplication rule for independent events tells us how to compute the probability of a sequence that consisted of the first k trials being successes and the rest of the n-k trials being failures. The probability of this particular sequence of k successes followed by n-k failures is given by $p^k \cdot (1-p)^{(n-k)}$.

The probability distribution described by this formula is called the binomial distribution. It is named after the binomial coefficient. The binomial distribution has two parameters: the number of trials n and the probability of success on each trial, p.

Special Distributions: Hypergeometric

Properties:

- Drawing without replacement
- N total items
- G success items
- n draws
- Parameters: N, G, n

Formula:

$$P(k \text{ successes}) = \frac{\binom{G}{k}\binom{N-G}{n-k}}{\binom{N}{n}}$$

Key Differences from Binomial:

1. Probability changes after each draw
2. Draws are dependent

Special Distributions: Hypergeometric

3. No replacement
4. Three parameters instead of two

Example:

- Box with 10 tickets
- 6 marked success
- Draw 3 tickets
- P(2 successes) = $\frac{\binom{6}{2}\binom{4}{1}}{\binom{10}{3}} = \frac{1}{2}$

In the binomial scenario described above, we had n independent trials, where each trial resulted in a success or a failure. This is like sampling with replacement from a box of 0's and 1's. Now consider the situation when we have a box with N tickets marked with either 0 or 1.

Special Distributions: Hypergeometric

As usual, the ticket marked 1 represents a success. Say the box has G tickets marked 1 (and therefore N-G tickets marked 0 representing failures). Suppose we draw a simple random sample of size n from this box. A simple random sample is a sample drawn without replacement, and on each draw, every ticket is equally likely to be selected from among the remaining tickets. Then, the probability of drawing a ticket marked 1 changes from draw to draw.

The probability distribution that gives us this answer is called the hypergeometric distribution. It has three parameters, n, N and G. This is a wacky formula, so let's explain it piece by piece!

Binomial vs Hypergeometric: Key Differences

Binomial Distribution:

- With replacement
- Independent trials
- Constant probability
- Parameters: n, p
- Example: Coin flips

Hypergeometric Distribution:

- Without replacement
- Dependent trials
- Changing probability
- Parameters: N, G, n
- Example: Drawing cards

Binomial vs Hypergeometric: Key Differences

When to Use Which?

- Use Binomial when:
  - → Population is large relative to sample
  - → Replacement is used
  - → Trials are independent
- Use Hypergeometric when:
  - → Sampling significantly affects probabilities
  - → No replacement
  - → Trials are dependent

Both these distributions deal with:

- a fixed number of trials, or instances of the random experiment

Binomial vs Hypergeometric: Key Differences

- outcomes that are deemed either successes or failures

The difference is that for a binomial random variable, the probability of a success stays the same for each trial, and for a hypergeometric random variable, the probability changes with each trial.

If we use a box of tickets to describe these random variables, both distributions can be modeled by sampling from boxes with each ticket marked with 0 or 1, but for the binomial distribution, we sample n times with replacement and count the number of successes; and for the hypergeometric distribution, we sample n times without replacement, and count the number of successes in our sample.

Code Implementation: Key Functions

rep()

```
# Creating vector with repeated values
die <- 1:6
# Repeat each number according to probability
sum_dice <- rep(die, times = c(1,2,3,4,5,6))
head(sum_dice, 10)
```

```
 [1] 1 2 2 3 3 3 4 4 4 4
```

replicate()

```
# Simulate rolling two dice 10 times
set.seed(214)
replicate(10, sum(sample(1:6, 2, replace = TRUE)))
```

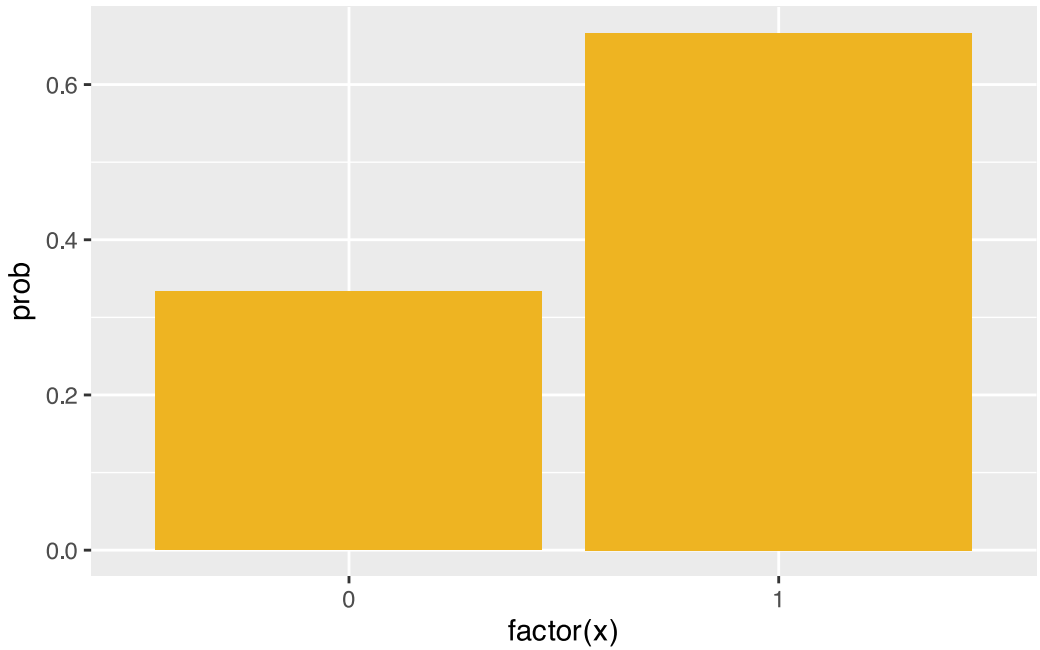
Code Implementation: Key Functions

```
 [1]  7 11  8  8 12  5  7  8  7 10
```

geom_col()

```
# Create probability histogram
x <- c(0,1)
px <- c(1/3, 2/3)
data.frame(x, prob = px) |>
  ggplot(aes(x = factor(x), y = prob)) +
  geom_col(fill = "goldenrod2")
```

Code Implementation: Key Functions



Three useful functions for working with probability distributions:

1. rep(): replicates values in a vector

Code Implementation: Key Functions

   - Arguments: x (vector to repeat), times (number of repetitions), each (repeat each element)
2. replicate(): repeat a specific set of tasks
   - Arguments: n (number of repetitions), expr (task to repeat)
3. geom_col(): plotting with probability
   - Creates bar chart where heights represent specified values
   - Useful for probability histograms

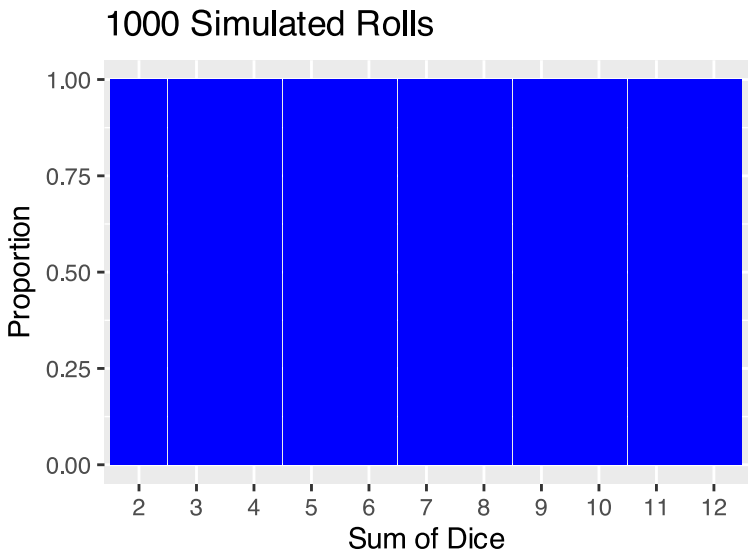
Simulating Probability Distributions

Die Rolling Example:

```
set.seed(123)
# Simulate 1000 rolls
rolls <- replicate(1000,
  sum(sample(1:6, 2, replace = TRUE)))

# Plot empirical distribution
data.frame(rolls) |>
  ggplot(aes(x = factor(rolls))) +
  geom_bar(aes(y = after_stat(prop)),
    fill = "blue", width = 0.98) +
  labs(x = "Sum of Dice",
       y = "Proportion",
       title = "1000 Simulated Rolls")
```

Simulating Probability Distributions



Key Points:

1. Simulation Process:
   - Define possible outcomes
   - Set number of repetitions

Simulating Probability Distributions

- Use appropriate sampling
- Calculate proportions

2. Visualization:
   - Use geom_bar for empirical
   - Use geom_col for theoretical
   - Compare distributions
   - Observe convergence

Let's simulate rolling two die and summing the spots. We can accomplish this task and examine our results using the functions we have just introduced. First, we will make a vector representing a fair, six-sided die.

We can use the sample() function to roll the die twice; this will output a vector with two die numbers. Then, we can take the sum of this vector by nesting the call to sample() inside of sum.

Simulating Probability Distributions

If we would like to repeat this action many times (for instance, in a game of Monopoly, each player has to roll two dice on their turn and sum the spots), the replicate() function will come in handy.

With only 50 experiments run, we see that the empirical histogram doesn't quite match. However, modify the above code by increasing the number of repetitions, and you will see the empirical histogram begin to resemble more closely true probability distribution. This is an example of long-run relative frequency.

Summary

1. Probability Distributions:

Summary

Summary

Summary

1. Probability Distributions:
   - Theoretical vs Empirical
   - Visualization methods
   - Long-run behavior
2. Special Distributions:

Summary

Summary

Summary

1. Probability Distributions:
   - Theoretical vs Empirical
   - Visualization methods
   - Long-run behavior
2. Special Distributions:
   - Discrete Uniform
   - Bernoulli
   - Binomial
   - Hypergeometric
3. Implementation Tools:

Summary

Summary

Summary

1. Probability Distributions:
   - Theoretical vs Empirical
   - Visualization methods
   - Long-run behavior
2. Special Distributions:
   - Discrete Uniform
   - Bernoulli
   - Binomial
   - Hypergeometric
3. Implementation Tools:
   - rep() for repeated values
   - replicate() for simulations
   - geom_col() for visualization
4. Key Concepts:

Summary

- Counting principles
- Permutations vs Combinations
- With vs Without replacement
- Dependent vs Independent trials

In this lecture, we covered:

- Defined probability distributions
- Stated the basic counting principle and introduced permutations and combinations
- Defined some famous named distributions (Bernoulli, discrete uniform, binomial, hypergeometric)
- Visualized probability distributions using probability histograms
- Looked at the relationship between empirical histograms and probability histograms
- Introduced functions rep(), replicate(), geom_col()

Summary

- Simulated random experiments such as die rolls and coin tosses to visualize the distributions