

# Probability Distributions, Permutation Tests, and You

Grinnell College

October 29, 2025

# Motivation

Real fast, everyone write down a random ten digit number (don't use a computer/phone/etc...)

We will double back to this one later one

# Overview

Today we will discuss...

- ▶ Probability Distribution
  - ▶ What are they?
  - ▶ What do we use them for
- ▶ Permutation Tests
  - ▶ First use of our probability distributions
  - ▶ Uses the distribution to discuss the “weirdness” of something happening
  - ▶ Sufficiently unlikely means something is wrong

## Previously....

On Monday we talked about probability and looked at a couple examples...

- ▶ Rolling pairs of dice
- ▶ Writing a random string of digits

Something you may or may not have noticed is that often times probability forms a distribution that we can (coherently) discuss

# Probability Distributions

A probability distribution is a valid assignment of probabilities to all possible, disjoint outcomes. For example the roll of a single die....

|             |               |               |               |               |               |               |
|-------------|---------------|---------------|---------------|---------------|---------------|---------------|
| Roll        | 1             | 2             | 3             | 4             | 5             | 6             |
| Probability | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ |

The above is a uniform distribution

- ▶ each outcome is equally likely
- ▶ talked about on Monday
- ▶ Generally common and intuitive

# Probability Distributions

We can define a distribution for something more complex too. For example, the distribution for the sum of two dice rolls...

| Roll        | 2              | 3              | 4              | 5             | 6              | 7             | 8              | 9             | 10             | 11             | 12             |
|-------------|----------------|----------------|----------------|---------------|----------------|---------------|----------------|---------------|----------------|----------------|----------------|
| Probability | $\frac{1}{36}$ | $\frac{1}{18}$ | $\frac{1}{12}$ | $\frac{1}{9}$ | $\frac{5}{36}$ | $\frac{1}{6}$ | $\frac{5}{36}$ | $\frac{1}{9}$ | $\frac{1}{12}$ | $\frac{1}{18}$ | $\frac{1}{36}$ |

How did I make the above?

# Probability Distributions

I could've counted the number of dice rolls who's sum would be that number

|  |   |   |   |   |   |
|--|---|---|---|---|---|
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

# Probability Distributions: General Info

There are a few rules for a probability distribution to be valid:

- ▶ The outcomes must be disjoint
  - ▶ **Disjoint** means an observation must be in one outcome or another outcome but cannot be considered in both
    - ★ Math-y definition: Two sets which have no common element
    - ★ Eg I roll a die and get a 1 or a 2 but not both
    - ★ Eg I am wearing a belt or I am not wearing a belt.
    - ★ INCORRECT: I roll a six on a dice or I am wearing a belt.
- ▶ Any outcome's probability is between 0 and 1 (inclusive)
- ▶ The probability for all outcomes sum to 1
- ▶ The probability for an impossible event is 0

(Probability distributions are more complicated than this but this is the basic idea and will work for us in this course; look up probability density function and probability mass function for more info or talk to me later)

## Probability Distributions: Practice

In Dungeons and Dragons, dice have a particular abbreviation. To indicate the number of dice rolled (N) and the number of sides on a dice (S), we write NdS.

- ▶ Eg 2d8 means you roll two 8-sided dice
- ▶ Eg 1d6 mean you roll one 6-sided dice (like slide 5)

Write out the probability distribution for the sum of a 3d3 (three 3-sided dice roll). HINT: There are 27 possible combinations

# Probability Distributions: Practice

In Dungeons and Dragons, dice have a particular abbreviation. To indicate the number of dice rolled (N) and the number of sides on a dice (S), we write NdS.

- ▶ Eg 2d8 means you roll two 8-sided dice
- ▶ Eg 1d6 mean you roll one 6-sided dice (like slide 5)

Write out the probability distribution for the sum of a 3d3 (three 3-sided dice roll). HINT: There are 27 possible combinations

|             |                |                |                |                |                |                |                |
|-------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Sum         | 3              | 4              | 5              | 6              | 7              | 8              | 9              |
| Probability | $\frac{1}{27}$ | $\frac{3}{27}$ | $\frac{6}{27}$ | $\frac{7}{27}$ | $\frac{6}{27}$ | $\frac{3}{27}$ | $\frac{1}{27}$ |

# Probability Distributions: Why Care?

We can do two main things with a probability distribution:

1. Give a range of values that are "likely"
  - ▶ Eg giving the distribution of grades on an exam and discussing the most likely score a student would get
  - ▶ Eg reporting the location of  $Q_1$  and  $Q_3$
  - ▶ Confidence and Prediction Intervals (not this slide deck)
2. Discuss the probability of an event happening \*under our assumptions\*
  - ▶ Eg The probability that a random string of numbers would have a given number of odd digits
  - ▶ Eg That two satellites on a given trajectory will collide (astro-statistics)
  - ▶ Eg Call bs on something happening "randomly"
  - ▶ Hypothesis Testing

## Another Example: Set Up

Did you know that most people, when writing random numbers, overestimate how many odd numbers there should be?

Let's find the probability distribution for the total number of "odd" digits in a 10 digit number!

## Another Example: How To

Did you know that most people, when writing random numbers, overestimate how many odd numbers there should be?

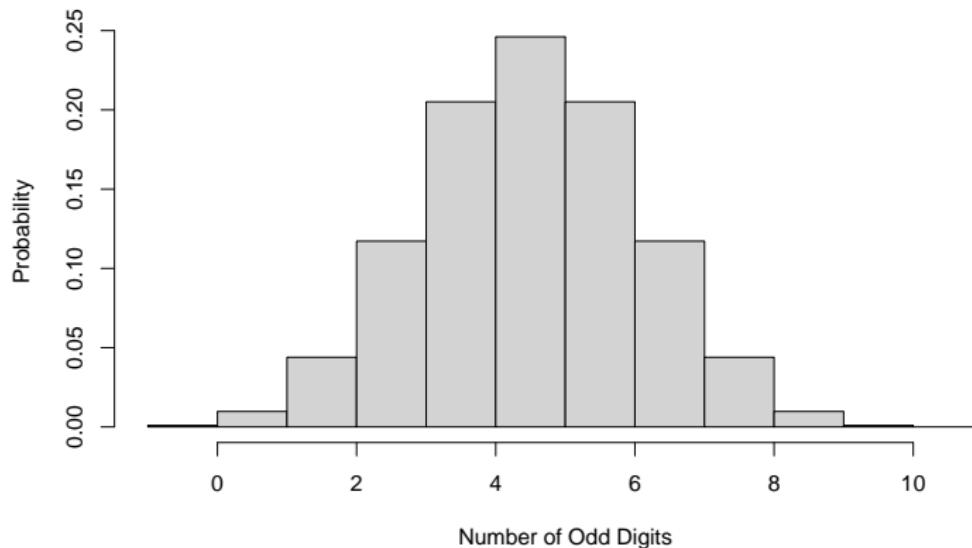
Let's find the probability distribution for the total number of "odd" digits in a 10 digit number!

---

First step is to write all the numbers between 0 and 9999999999 so let's get started

0000000000, 0000000001, 0000000002, and I'm bored already so let's make R do it

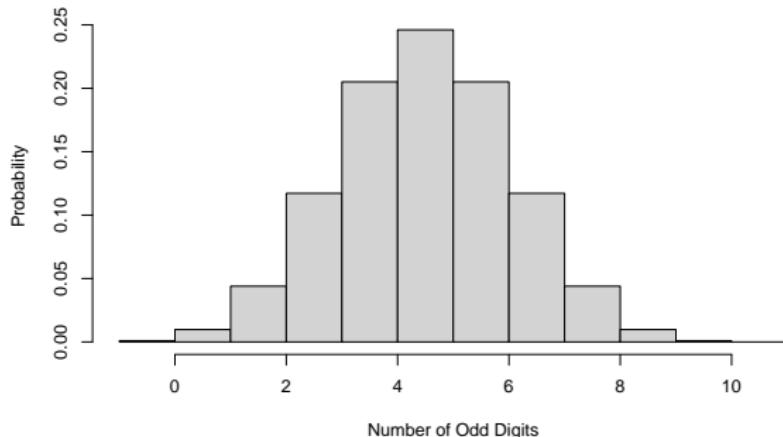
## Another Example: Probability Distribution



This distribution has a mean of 5, is symmetric, unimodal, and bell shaped

Is this theoretical, empirical, or subjective probability knowing that these are the exact proportions?

## Another Example: Probability Distribution



This distribution has a mean of 5, is symmetric, unimodal, and bell shaped

Is this theoretical, empirical, or subjective probability knowing that these are the exact proportions? It's theoretical as it's based on assumptions (equally likely numbers) and all possible outcomes

# Permutation Test

1. We have some starting point
  - ▶ Eg I believe this dice is fair
2. We calculate the probabilities of different outcomes
  - ▶ Eg the probabilities that we'd get one 6 in 10 dice rolls, two 6's in 10 dice rolls, three 6's.....
3. We compare those probabilities against what was actually observed and comment if it seems probable that it could happen
  - ▶ Eg HOW has this guy rolled 6's the last ten rolls straight?
  - ▶  $(1/6)^{10} = 1.65 \times 10^{-8}$

# Permutation Test

The **permutation test** calculates the exact theoretical probability we'd see something as or more weird than we did given our starting belief is true.

- ▶ We have a belief about something
- ▶ All possible outcomes are listed
- ▶ We observe the process in real life
- ▶ The proportion of outcomes weirder than our result/observation is found
- ▶ That proportion is reported
  - ▶ The proportion weirder than our sample is the probability we'd see something weirder
  - ▶ Very low probability means that what we observed was already very weird \*or\* our starting assumptions/hypothesis was wrong

## Permutation Test

(Based on Robert Lee Tarver's "trial") In 1993, Russell County Alabama had a population that was roughly 40% African American. In a racially charged crime, a jury pool of 35 participants were gathered with 14 black and 21 white community members.

He was convicted by a jury of 11 white people and 1 black person.

Based on this information, do we think the jury was a randomly chosen jury of his peers?

(I can't find how many white people were actually in the jury pool but I'm working off the assumption the amount would reflect Russell County's racial demographics)

## Permutation Test

Starting Point: I believe all possible jury combinations were equally likely and not dependent on race

There are many many many combinations of 12 people out of 35 but we can work them out slowly

$$P(0 \text{ black jurors}) = .000352$$

$$P(1 \text{ black juror only}) = .00592$$

$$\text{Alternatively, } P(4 \text{ or } 5 \text{ black jurors}) = .5230$$

Do you believe that this jury happened by chance?

## Probability Distribution: Hypergeometric Distribution

There are many many many combinations of 12 people out of 35....how many?

## Probability Distribution: Hypergeometric Distribution

There are many many many combinations of 12 people out of 35....how many?

- ▶ 834,451,800 combinations exactly of 12 person juries each
- ▶ I can't write that out
- ▶ My laptop doesn't have space for it
- ▶ Solution: Use the hypergeometric distribution

# Probability Distribution: Hypergeometric Distribution

I used a “hypergeometric” distribution which is another probability distribution (already seen uniform and normal)

- ▶ I'm not teaching it as it's complicated and in the weeds
- ▶ Describes the probability of pulling  $N$  observations from Subpopulation 1 and  $M$  observations from Subpopulation 2 for a total observation size of  $N + M$  (out of the size of the population)
- ▶ Playing cards usually are hypergeometric distributions
  - ▶ Eg “I need to draw at least one of the two remaining aces and any two other cards”
  - ▶  $N = 1$  (out of the 2 aces)
  - ▶  $M = 2$  (out of any of the remaining cards in the deck)
  - ▶  $N + M = 3$  (number of cards we need to draw in total)
  - ▶ For more details [Wiki](#) is a good place to start

# Formalizing Testing

Alright, on that happy note let's walk through a generic hypothesis test set-up and then we can redo the Tarver example with proper notation

For those who have seen t-tests, this will all look similar

For those who haven't, is a lot but taken step by step its not bad

# The Fundamentals of Hypothesis Testing

What is the point of hypothesis testing in statistics?

# The Fundamentals of Hypothesis Testing

What is the point of hypothesis testing in statistics?

**Hypothesis tests** allows us to judge how likely seeing our results are if our starting assumption/place (the null hypothesis,  $H_0$ ) is true.

- ▶ If the results are unlikely (small probability) we don't believe our starting point ( $H_0$  we think is wrong)
- ▶ If the results are not unlikely (not small probability) we can't really comment if we think  $H_0$  is correct or not
  - ▶ We just don't have evidence to say it's wrong

# Outline of Hypothesis Tests

All(?) hypothesis tests follow the same general format. Outside of visualizing your data and the predetermined threshold I know of no violations of this

1. Null and Alternative Hypothesis are stated
2. A predetermined threshold (critical level) for action is decided upon (optional, see step 7)
3. Visualize your data
4. Assumptions are checked for whatever test you are using
  - ▶ If failed find a different test/do something else
5. Test Statistic is calculated
6. The probability we see that test statistic if the Null Hypothesis is true is calculated
7. A decision is made

# Hypothesis Statements

We need to list the null and alternative hypothesis so we can all agree what we are actually testing.

- ▶ The **null hypothesis** is the hypothesis we want to disprove
  - ▶ Indicated by  $H_0$ :
  - ▶ Eg Our new fertilizer doesn't do better than the competitor's fertilizer
  - ▶ Eg The probability I get a head is .5
- ▶ The **alternative hypothesis** is what the null hypothesis isn't.
  - ▶ Indicated by  $H_A$ :
  - ▶ Eg Our fertilizer is better than the competitors
  - ▶ Eg The probability I get a head is not .5

All possible situations have to be put into one of the two hypothesis. Ie  $H_0$  and  $H_A$  are disjoint

## Hypothesis Statements: All possible situations?

Yes, ALL possible situations. Our fertilizer can be better, the same \*or\* worse than the competitors.

- ▶ NULL: Our new fertilizer doesn't do better than the competitor's fertilizer
  - ▶  $H_0: \mu_{newfert.} \leq \mu_{oldfert.}$
- ▶ ALTERNATIVE Our fertilizer is better than the competitors
  - ▶  $H_A: \mu_{newfert.} > \mu_{oldfert.}$
- ▶ Note the “less than or equal to” in the null hypothesis.
  - ▶ Without the “less than” part we have no rational way to deal with the situation where our fertilizer is inferior
  - ▶ The idea of the hypothesis is the same as you have seen before

# Threshold

This is contentious so I kicked to later in the slidedeck

# Visualize Your Data

We graph to allow quick and easy interpretation of the data

- ▶ Done this extensively
- ▶ Not traditionally taught as part of the steps in a hypothesis test
- ▶ Can be a critical step in my opinion

# Assumptions

Most tests come with some form of assumptions

- ▶ Independence between samples is super common
- ▶ Having a constant spread/variance is also common
- ▶ Some are more odd
  - ▶ Some tests for proportions strongly suggest at least 10 realizations of each category
- ▶ Permutation Tests have an “exchangability” requirement
  - ▶ I’m not getting into this
  - ▶ Permutation tests were motivation for the next part of the unit
  - ▶ Idea being results we saw could be switched around/taken in different order and the test is still valid

# Test Statistic

A test statistic is a function of the statistic you are interested in that will be used to judge the "weirdness" of your results

- ▶ Some are easy to calculate
  - ▶ Eg the number of black jurors
- ▶ Others take a few steps to get to
  - ▶ Eg A z-score or t-score (for z- and t- tests)
  - ▶ Eg Log odds ratio
  - ▶ Eg Taking the log of the ratio of likelihood functions
  - ▶ Many tests require you to normalize your result

## Choose a threshold (contentious)

Traditionally the conclusions of hypothesis tests either say  $H_0$  is wrong or that we don't know if  $H_0$  is wrong.

- ▶ Based on if the probability we calculated is smaller than our threshold
- ▶ Eg your p-value is below .05 so we “Reject the Null Hypothesis”
- ▶ Statisticians generally have been pulling away from this hard cutoff, why?

## Choose a Threshold: Why Not To

Both statistics as a field and science generally are pulling away from using a cutoff value to decide statistical significance

- ▶ The cutoff is often arbitrary
- ▶ And it doesn't have meaning if there isn't a repeated experiment/study
  - ▶ No repetition - no long term behavior - no probability
- ▶ A hard cutoff for statistical sig. encourages “p-hacking”
  - ▶ Forcing results to look better than they are
  - ▶ “Publish or perish” incentives this
  - ▶ John Oliver has a bit on this
- ▶ American Statistical Association has an entire statement on this.

## Choose a threshold: Counter Argument

In this class we will NOT be choosing a threshold and instead will discuss the strength of evidence (weird, super weird, or kind of expected?)

Why bring it up then?

- ▶ Still common and you'll almost surely see it
- ▶ So I wanted to expose you to it
- ▶ Still has a place
  - ▶ This version of stat testing took off during WW2
  - ▶ Quality control for munitions and war materials
  - ▶ Because they would test similar things regularly that 5% threshold value actually means something probabilistically

Hard Question to Answer for a Zealot: Doesn't the 5% have meaning for a journal as a whole? Many, many tests.....

## Decision is Made

Instead of talking about if our weirdness (probability we'd see this) is less than some arbitrary number we can talk about the strength of evidence we have against  $H_0$

- ▶ We use our strength of evidence to say something coherent
- ▶ Eg We have very strong evidence that the proportion of white jurors in the trial was too high for random chance
- ▶ We no longer say we reject/fail to reject the null hypothesis
  - ▶ INCORRECT: We reject the null hypothesis the the juries were randomly selected regardless of racial make up
- ▶ Personally I feel like this kicks the bucket down the road some but that's just me

## Permutation Test Example Revisited

(Based on Robert Lee Tarver's "trial") In 1993, Russell County Alabama had a population that was roughly 40% African American. In a racially charged crime, a jury pool of 35 participants were gathered with 14 black and 21 white community members.

$H_0$ :

$H_A$ :

## Permutation Test Example Revisited: Hypothesis

(Based on Robert Lee Tarver's "trial") In 1993, Russell County Alabama had a population that was roughly 40% African American. In a racially charged crime, a jury pool of 35 participants were gathered with 14 black and 21 white community members.

$H_0$ : Proportion of African American jurors  $\geq .4$

- ▶ The same as the demographic make up of Russell County AL (or higher)

$H_A$ : Proportion of African American jurors  $< .4$

- ▶ What we are interested in checking

## Visualize and Check Assumptions

Not much to visualize here to be honest. We could plot the probability distribution of getting 0, 1,...,12 black jurors but that might be overkill for this example.

Again, permutation tests have that esoteric assumption that we won't worry about.

# Test Statistic and Probability

Our test statistic is going to be the number of black jurors in the jury

- ▶ 1 person

We want to find the probability we see something as or more extreme than what we saw

- ▶ Need the probability that there is 1 black person on the jury or 0 black people on the jury
- ▶ These two events are disjoint so we can add their probabilities
- ▶ The probability there would only be 1 or fewer black jurors on the jury if jurors were randomly selected is  $.00592 + .000352 = .00627$ ; less than 1%

## Decision

We have strong evidence that the jurors were not randomly selected from Tarver's peers in the community and that African Americans were more likely to be struck from jury duty (not picked to be a juror)

- ▶ .00627 comes off as too small to be random chance if the juries were randomly selected
- ▶ The competing theory would be that jurors were selected based on ethnicity which seems most likely
  - ▶ (My assumed jury pool could be wrong in real life but let's sweep that under the rug for this example)

## Next Time...

We will work on figuring out distributions that are built with help from the data and not purely theoretical

- ▶ Sampling Distribution for a mean
  - ▶ Assumptions
  - ▶ Behavior
  - ▶ Effect of Known vs Unknown Variance