Probability II: Distributions and Random Variables

Justin D. Pierce

Department of Political Science
University of Illinois

8/19/2021

Classes of Data

We'll look at two classes of data:

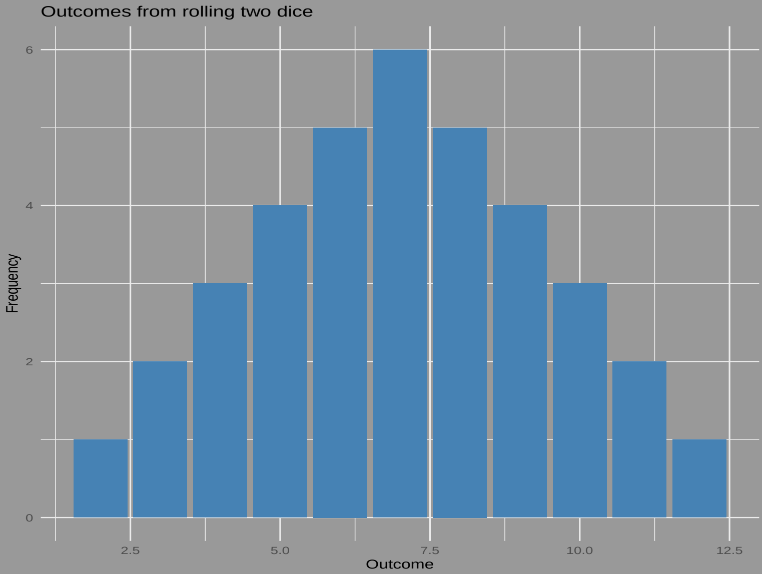
- Discrete
- Continuous

Discrete Data

Example: Rolling two d6's

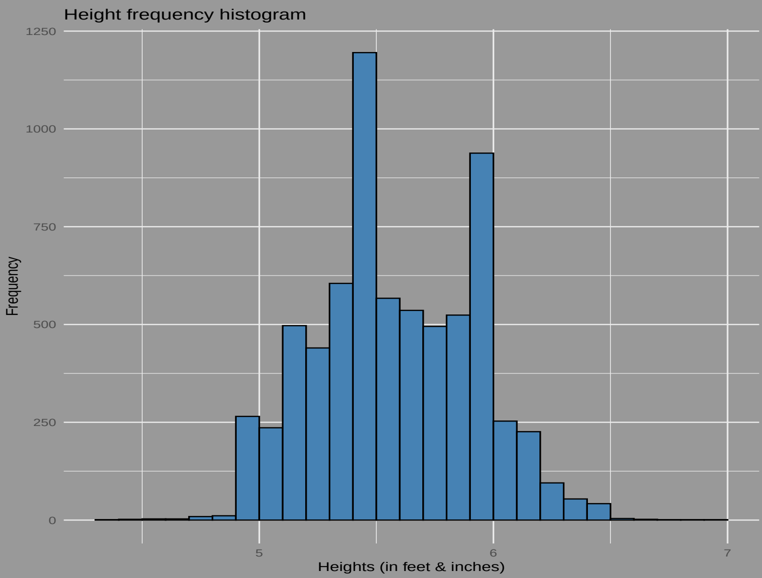
- Roll can be anywhere from 2 through 12
- Are **discrete** outcomes: can't have 4.5
- We can graph these discrete outcomes, and their frequency, in a barplot

Dice Roll outcomes and their frequency



Continuous Data

- Discrete data can only have certain outcomes; Continuous however can take on any value within range.
- Example: height
- Typically measure in feet & inches or centimeters, but can take on any value between full inches or centimeters.
- We graph these outcomes and their frequency with a histogram


Height frequency histogram
1250
1000
750
500
250
0
Frequency
5
6
7
Heights (in feet & inches)

Types of Data

In addition to the classes, there are four types of data

- Categorical(or nominal)
- Ordinal
- Interval
- Ratio

Categorical Data

Aka Nominal Data.
Has 2+ categories which data can fall into.
Examples:

- hair or eye color
- Party identification

NOTE: die rolling not. Instead, is....

Ordinal Data

Similar to categorical, BUT categories have an order/ranking to them.
Example: Presidential job approval:
Do you approve of Joe Biden's job performance?

1. I approve a great deal
2. I approve a moderate amount
3. I neither approve or disapprove
4. I disapprove a moderate amount
5. I disapprove a great deal

Interval Data

Similar to ordinal, but the groups are equally spaced.
Can be either continuous or discrete; different ways of classifying.
Example: What is your income?

- Ranges: \$40-45k, \$46-50K, \$51-55K, etc
- Continuous at \$1 intervals

Ratio Data

Similar to interval, but has a meaningful 0 value.
Likewise, can be continuous or discrete.
Example: How much time do you spend watching TV?

- Ranges: 0-2 hours, 3-5 hours, 6-8 hours, etc
- Or continuous with 1 hour intervals

Thus,

- Categorical and ordinal data is always discrete
- Interval and ratio data can be discrete or continuous

Types	Description	Classes
Categorical	Distinct Categories, w/o order	Dis Only
Ordinal	Distinct Categories w/ order	Dis Only
Interval	Ordered with equally spaced groups	Dis or Cont
Ratio	Ordered with eqaully spaced groups AND a meaningful 0	Dis or Cont

Why do we care??
Changes how we study our data

Random Variables

- Def: a variable whose value is the outcome of a random event.

Random Variables

- Def: a variable whose value is the outcome of a random event.
- **More formally**: a random variable is a function that assigns a number to each point in a sample space.

Random Variables

- Def: a variable whose value is the outcome of a random event.
- **More formally**: a random variable is a function that assigns a number to each point in a sample space.
- For social science purposes, more intuitive definition:

Random Variables

- Def: a variable whose value is the outcome of a random event.
- **More formally**: a random variable is a function that assigns a number to each point in a sample space.
- For social science purposes, more intuitive definition:
- A random variable is a *process* or *mechanism* that assigns the value of a variable in each case.
- We write this as X(x), where X=our function, x=our variable.

Political Science Example[1]

Let's say we have a sample space, S, which is a list of countries, i.e. S=a list of countries, with an individual country $s \in S$.

Political Science Example[1]

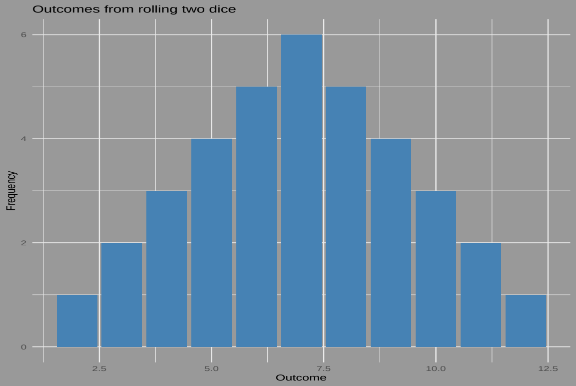
Let's say we have a sample space, S, which is a list of countries, i.e. S=a list of countries, with an individual country $s \in S$.

Country, s, GDP per capita in 1990 can be thought of as a random variable: X(s) = 1990 GPD per capita:

- X(Ghana)=$902
- X(France)=$13,904

Distributions

A list of all the possible outcomes of a random variable along with the corresponding probability with which they occur.



This is a probability distribution for rolling two dice.

Discrete Distribution functions

Probability Mass Function (pmf): For a discrete random variable, the pmf f(x) tells the probability that a given value will occur P(X=x).

Key properties:

- $0 \geq f(x) \leq 1$
- $\sum f(x) = 1$

Discrete Distribution functions II

Cumulative Distribution Function (CDF): For a discrete random variable, the CDF F(x) is the function that tells us the cumulative probability that a given value, x, or any value smaller than x will occur.

$F(X)=P(X \leq x)$

For discrete variables, the equation for a CDF is:

$F(X)=\sum_{i=0}^{x} f(x)$

Continuous Distribution functions

Probability Density Functions (PDF): For a continuous random variable, the PDF f(X) is a function that tells us the probability that a random variable will fall within a particular range of values.
Key properties:

- $P(a \leq X \leq b) = \int_a^b f(x)dx$
- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- $P(X=x) = 0$ (because the integral at a single point is 0)

Cont. Distribution functions II

Cumulative Distribution Function (CDF): For a continuous random variable, the CDF F(x) is the function that tells us the cumulative probability that a given value, x, or any value smaller than x will occur.

Note, same name and concept as the CDF of a discrete variable. The difference is how we calculate it:

F(x)=$\int_{-\infty}^{x} f(\mu) d(\mu)$

Describing Distributions

Describe the shape of distribution through key *parameters*

- Mean (aka expected value)
- Variance
- Standard Deviation

Describing Distributions: the Mean

A description of the central tendency of a distribution or variable.

An expectation: expected value or weighted average that X will take on after many trials.

Represented by E[X] or μ.

Calculated:

- Discrete: $\mu = \sum xf(x)$
- Continuous: $\mu = \int xf(x)dx$

Describing Distributions: the Variance

A measure of the spread of a distribution/variable.

Also an expectation: it is the weighted average of the squares of the distance between X and E[X].

Represented by σ^2 (More on that in a second).

Calculated:

- Discrete: $Var[X] = E[(X - E(X)^2)] = E[X^2] - (E[X])^2$
- Continuous: $Var[X] = E[(X - \mu)^2] = \int (x - \mu)^2 f(x) dx$

Describing Distributions: the Standard Deviation

Another measure of how spread out the numbers are in a distribution or variable.

Represented by σ

Calculated:

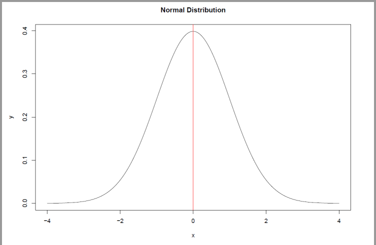
- SD[X]=$\sqrt{Var[X]}$

Distribution Example: The Normal Distribution

Aka "Gausssian" Distribution

Key features:

- "Bell Shaped" and symmetrical
- The mean, median and mode are all equal
- Has mean $= \mu$, standard deviation $= \sigma$ and variance $= \sigma^2$



The Normal Distribution continued

Why so important?

- Differences of means tests, such as t-Tests and ANOVA assume normal distributions.
- Regression assumes residuals normally distributed.
- Need to account if different.
- Often fine assumption because of the **Central Limit Theorem**.

Central Limit Theorem

Def: The sampling distribution of the sample means approaches a normal distribution as the sample size gets larger.

$\overline{X_n} = \frac{x_1+x_2+...+x_n}{n} \to \mathcal{N}(\mu,\ \frac{\sigma^2}{n}),\ n \to \infty$

Typically works in sample sizes larger than 30.
Part of why sample sizes are important.

Law of Large Numbers

As sample size grows, its mean gets closer to the population mean.

$\overline{X_n} = \frac{x_1+x_2+\ldots+x_n}{n} \to \mu, n \to \infty$

Another reason why we like large samples!

Samples vs Populations

- **Population:** the group that we want to draw conclusions about.
- **Sample:** the group of observations we will collect data from.

For example: Might be interested in population of US voters.

Too large a group to distribute survey to.

Instead get a sample which is ideally:

- Randomly drawn and representative.
- Large enough to make inferences from.

Questions?
Justin- jdpierc2@illinois.edu
Also see go.illinois.edu/surveystatsdata for CITL Data Analytics Services.