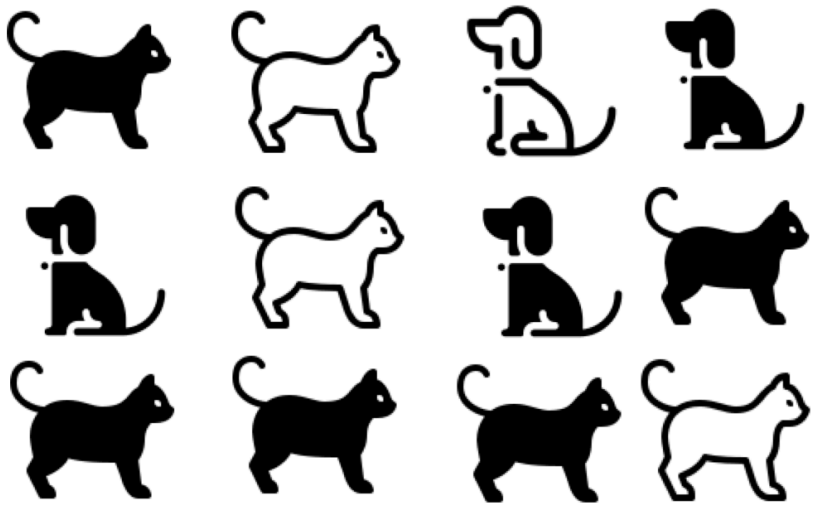
OI 3.4 and 3.5: Random Variables and Sampling Distributions

Random Processes

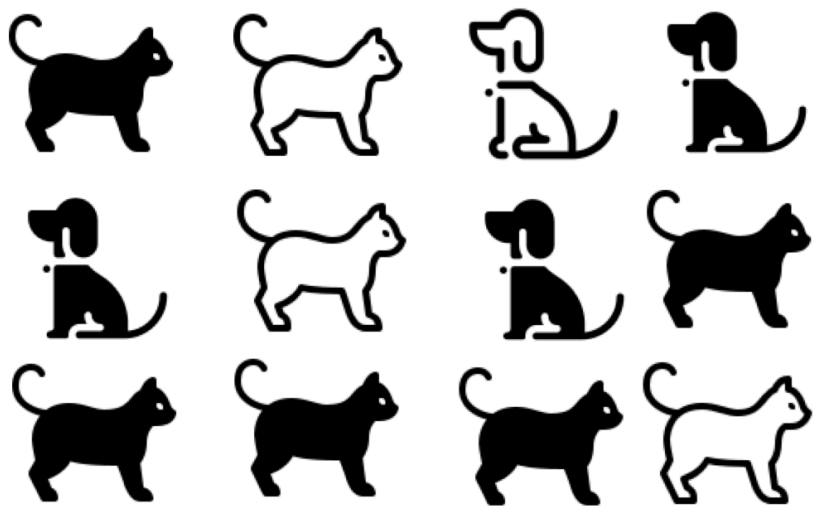
Up to this point we have been discussing probability of a particular outcome(s) of random process(es).



- What is the probability a black cat selected from the animals above?
  - $A =$ pet's coat is black
  - $B =$ pet species is cat
  - $P(A \text{ and } B) = P(\text{black cat})$

Random Variables

- Now we will switch to *random variables*, as opposed to *random processes*.
- A **random variable** is a random process with a *numeric outcome*.
- Some of our previous questions already where random variables. Others need to be adjusted to fit the new definition.



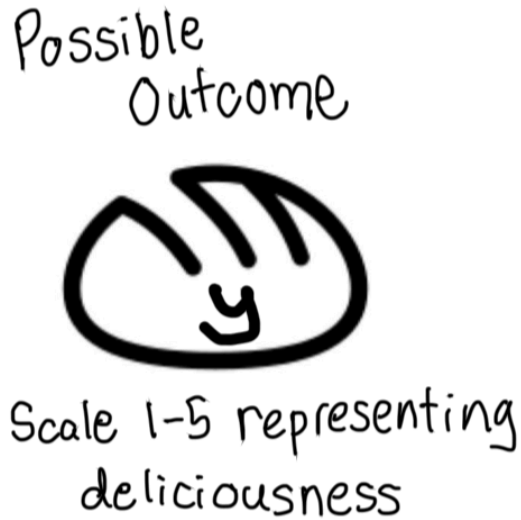
Let $X = 1$ if a cat is selected and $X = 0$ if otherwise. What is $P(X = 1)$?

Taxonomy of a Random Variable

- Recall we have two types of numerical variables.
- The same is true for random variables! We have *discrete random variables* and *continuous random variables*.
- We typically denote a random variable with capital letters: X, Y, Z
- We refer to the possible values of a random variable as the **support**.
- We denote observed (or hypothetical) values for a random variable by using it's respective lower case letter: x, y, z

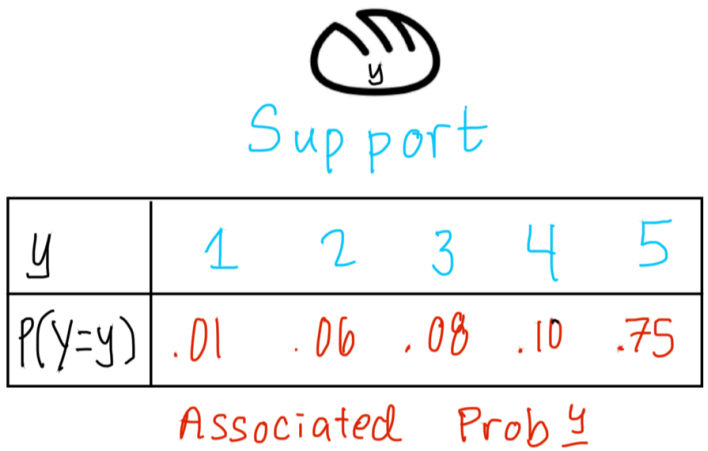
Notation for Random Variables

I like to bake bread. I have a favorite recipe, and each time I make a loaf it comes out a little different. Suppose there is a scale of deliciousness, 1-5: 1 (not delicious), .., 5 (very delicious). Each time I make my loaf of bread I can evaluate the deliciousness for that loaf.



Notation for Random Variables

A random variable Y is like the bread recipe. When I follow the recipe (Y), it results in something with varying levels of deliciousness (y) with different probabilities ($P(Y = y)$).



Notation for Random Variables

Y and y are short hand for a particular random variable and a particular observable value for that variable.

What is the probability that a loaf of bread made using my bread recipe would receive a deliciousness rating of 5?

$$P(Y = 5) = P(\text{Deliciousness rating from my bread recipe} = 5)$$

Random Variable Examples

Random Process	**Random Variable**
Flip 10 coins	$X = \#$ of heads
Roll a d6	$X = \#$ of pips
Body dimensions	$X =$ neck circumfrance $Y =$ waist circumfrance
Medical treatment assignment and status	$X = 0$ if placebo, 1 if treatment $Y = 0$ if no symptoms, 1 if symptoms
Cat body variables	$Y =$ heart wieght in g $X = 0$ if female, 1 if male $Z =$ body wieght in kg

What type of variables are these? Continuous or discrete?

Probability Mass Functions

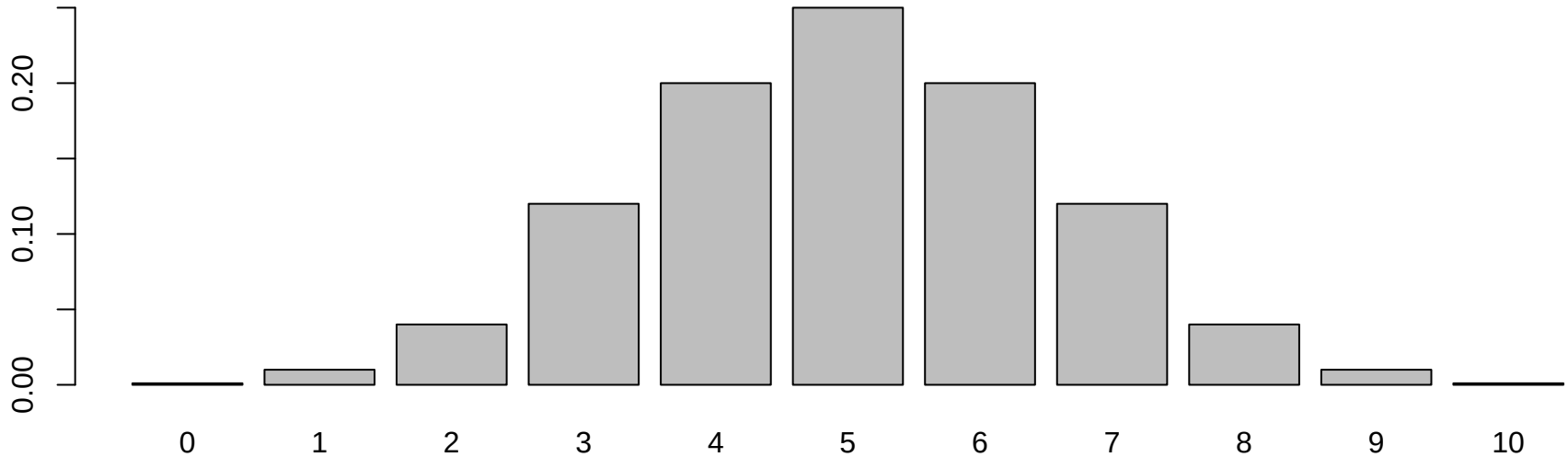
Discrete random variables have **probability mass functions (pmf)**, which for us is the same concept as a distribution:*a map of all possible values and their associated probabilities*

For example, let $X = \#$ of heads when flipping ten coins. We can represent this pmf using a:

- plot
- table
- function

Probability Mass Functions

Example with a plot:



Example with a table:

x	0	1	2	3	4	5	6	7	8	9	10
$P(X = x)$	0.001	0.01	0.04	0.12	0.20	0.25	0.20	0.12	0.04	0.01	0.001

Probability Mass Functions

Example with a function:

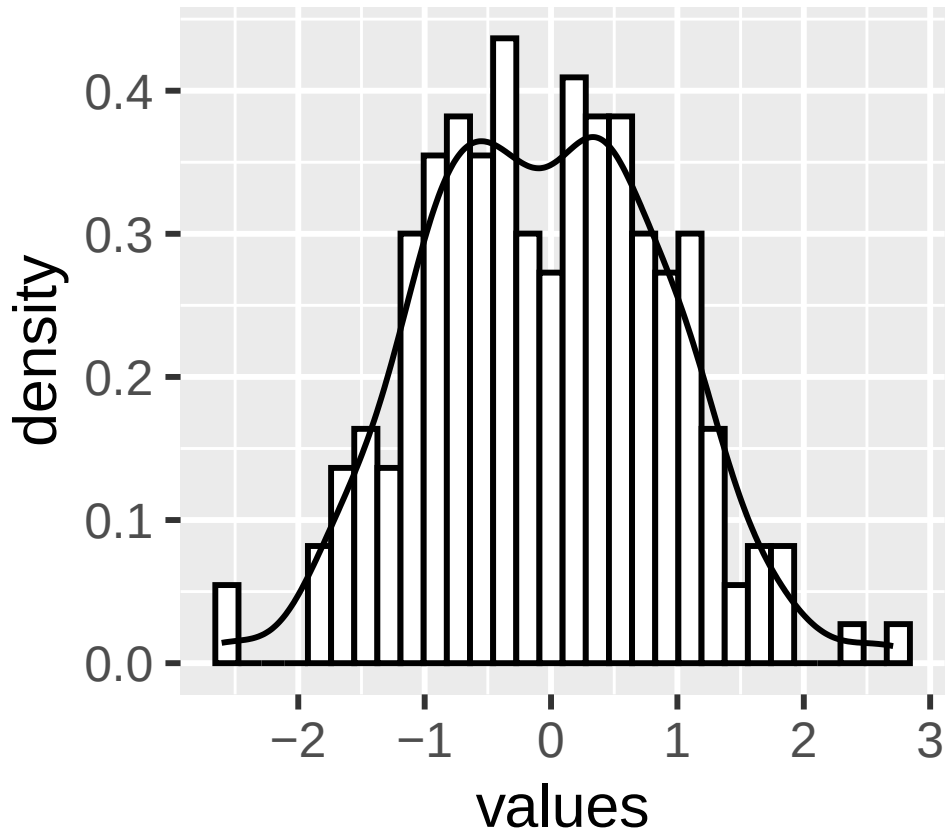
$$P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^x$$

when $x = 0, \ldots, n$, n is the total number of trials (coins flipped) and p is the probability of the event observed (heads)

Probability Density Function

Continuous random variables have **probability density functions (pdf)**, which is: *a map of all possible values and their relative probabilities*

We can estimate pdfs with histograms and the ridge/density plots from before.



Probability Density Function

We can also represent a pdf with functions. For example, here is a very famous pdf (you do not have to know this formula, it is just an example):

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(\frac{(x-\mu)^2}{\sigma^2}\right)$$

If we take the integral over all possible values (support), the area under the curve for a pdf is 1. That is,

$$\int f(x) = 1$$

PDF versus PMF

Why have separate definitions for discrete (pmf) and continuous (pdf) random variables?

PDF versus PMF

Why have separate definitions for discrete (pmf) and continuous (pdf) random variables?

- pdfs tell us *relative probability*
- pmfs have a direct relationship with probability
- pdf often involve calculus, and pmfs do not.

Summarizing Probability Distributions: Center

We can characterize the center and spread of a random variable, just like we did for observed data.

The **expected value** of a discrete random variable X is the weighted average of the possible values that X might take on.

$$E[X] = \sum_x x \times P(X = x)$$

Summarizing Probability Distributions: Center

- We often write $\mu = E(X)$.
- The expected value is the *mean* or *average* based on mathematical properties.
- It is our mean *in theory* based on *mathematical relationships* rather than observations.
- Recall our estimated mean ($\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$) is our *empirical* mean based on the *sample*, and nothing else!
- $\overline{x}$ and $E(X)$ are often similar, but are different concepts.

Example

Suppose that d6 has been loaded so that, rather than landing on each face with equal probability, it lands on a face with x pips with the following probabilities:

$$P(X = x) = \begin{cases} 0.1 & \text{for } x = 1, 3, 5 \\ 0.2 & \text{for } x = 4, 6 \\ 0.3 & \text{for } x = 2 \end{cases} .$$

What is the expected (average) number of pips on the observed face?

Example

What is the expected (average) number of pips on the observed face?

$$
\begin{aligned}
E[X] &= \sum_x x \times P(X = x) \\
&= 1(0.1) + 2(0.3) + 3(0.1) + 4(0.2) + 5(0.1) + 6(0.2) \\
&= 3.5
\end{aligned}
$$

Example

What is the expected value (average) deliciousness rating for breads created from my bread recipe?

y	1	2	3	4	5
$P(Y = y)$	0.01	0.06	0.08	0.10	.75

Properties of Expected Values

- The expected value of a constant c is the constant,

$$E[c] = c$$

- Let X be a random variable and let c be some constant. Then

$$E[cX] = cE[X].$$

- Let X and Y be two random variables. Then the expected value of their sum is

$$E[X + Y] = E[X] + E[Y]$$

Put this together to get the linearity of expectations property:

$$E[aX + bY] = aE[X] + bE[Y]$$

Summarizing Probability Distributions: Spread

We characterizing the center and spread of a random variable, just like we did for observed data.

The **variance** of a discrete random variable X captures its average squared distance from the mean:

$$Var(X) = E[(X - \mu)^2] = \sum_x (x - \mu)^2 \times P(X = x)$$

The **standard deviation** is $SD(X) = \sqrt{Var(X)} = \sqrt{\sigma^2} = \sigma$.

Summarizing Probability Distributions: Spread

- We often write $\sigma^2 = Var(X)$.
- This is the variance *in theory* based on *mathematical relationships* rather than observations.
- Recall our estimated variance ($\hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$) is our *empirical* variance based on the *sample*, and nothing else!
- $\hat{s}^2$ and σ^2 are often similar, but are different concepts.

Example

Suppose that d6 has been loaded so that, rather than landing on each face with equal probability, it lands on a face with x pips with the following probabilities:

$$P(X = x) = \begin{cases} 0.1 & \text{for } x = 1, 3, 5 \\ 0.2 & \text{for } x = 4, 6 \\ 0.3 & \text{for } x = 2 \end{cases} .$$

What is the variance in the number of pips on the observed face?

Example

What is the variance in the number of pips on the observed face?

$$
\begin{aligned}
Var(X) &= \sum_x (x - E[X])^2 \times P(X = x) \\
&= (1 - 3.5)^2(0.1) + (2 - 3.5)^2(0.3) + (3 - 3.5)^2(0.1) \\
&\quad + (4 - 3.5)^2(0.2) + (5 - 3.5)^2(0.1) + (6 - 3.5)^2(0.2) \\
&= 2.85
\end{aligned}
$$

Example

What is the variance of the deliciousness rating for breads created from my bread recipe?

y	1	2	3	4	5
$P(Y = y)$	0.01	0.06	0.08	0.10	.75

Properties of Variance

- The variance of a constant c is

$$Var(c) = 0$$

- Let X be a random variable and let c be some constant. Then

$$Var(X + c) = Var(X)$$

- Let X be a random variable and let c be some constant. Then

$$Var(cX) = c^2 Var(X)$$

- Let X and Y be two random variables. Then

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

Properties of Variance

Putting information together from the previous slide:

- Let X and Y be two random variables, and a and b be two constants. The linear property of variance is

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$$

- Let X and Y be two *independent* random variables, and a and b be two constants. The linear property of variance for independent random variables is

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$$

Additional Practice

1. **[OI 3.30]** Consider the following card game with a well-shuffled deck of cards. If you draw a red card, you win nothing. If you get a spade, you win \$5. For any club, you win \$10 plus an extra \$20 for the ace of clubs.
   - Determine the probabilities for each amount you might win.
   - Find the expected winnings for a single game and the standard deviation of the winnings.
   - What is the maximum amount you would be willing to pay to play this game? Explain.

2. A 2021 survey conducted by the University of New Hampshire found that about 10% of Americans agreed with the statement that "the Earth is flat." Suppose you conduct a second survey of 250 randomly selected Americans.
   - What is the expected number of respondents in your survey who agree with the statement that "the Earth is flat"?
   - What is the standard deviation of this number?