

Center for Statistics and the Social Sciences

Math Camp 2021

Lecture 6: Probability Density Functions

Peter Gao & Jessica Kunke

Department of Statistics
University of Washington

September 16, 2021

Motivation

- Statistics enables us to draw conclusions about a **population** from a **sample**.

For example we use the sample mean to estimate the population mean.

- Our estimator will be **random**: each time we perform the experiment, we will get a different sample and a different value.
- By understanding the **probability distribution** of our estimator, we can express its uncertainty (i.e. confidence intervals).

Motivation

Distributions vs. Data

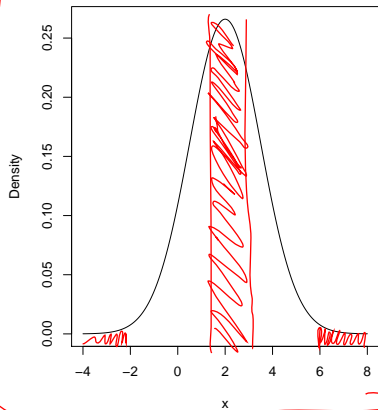
Probability
distribution

unknown
number x

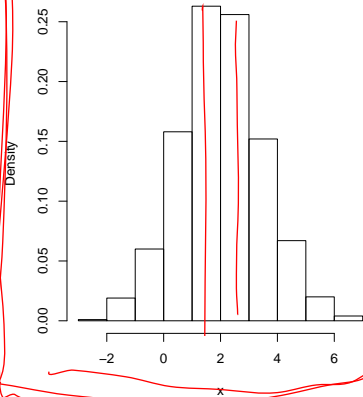
theoretical
PD

empirical
PD

$x \sim N(2, 1.5)$



1,000 Samples from $N(2, 1.5)$



Motivation

theoretical

In general the true probability distribution is unknown.

- We assume a model for the distribution of our data. The estimator we choose and its distribution will depend on that model.
- In lectures 7 and 8, we will learn some examples of different models (discrete and continuous).
- In lecture 9 we will see an introduction to maximum likelihood estimation.

Random Variables

A **random variable** is a function which assigns a number to each element in the sample space. (Think of it as the answer to a question you are asking about each element in the space).

- **Variable**, because the answer will be different for each element.
- **Random** because we can't predict the answer with any great certainty.

Random variables are usually denoted with capital letters, X , Y , Z .

Examples:

- Ask each person in the room their favorite number (integer).
Sample Space: $S = \{\dots, -2, -1, 0, 1, 2, \dots\}$ (all integers).
 - X = The favorite number
Possible Values: All
 - Y = the favorite number squared
Possible Values: $0, 1, 2, 4, 9, \dots$ (square numbers)

Random Variables

Examples

- Experiment: Flip 4 coins:

Sample Space:

$S = \{HHHH, HHHT, HHTH, HTHH, THHH, HHTT, HTHT, HTTH, THTH, TTHH, THHT, TTTH, TTHT, THTT, HTTT, TTTT\}$

- X = The number of Heads
Possible Values: 0, 1, 2, 3, 4
- Y = The number of Tails
Possible Values: 0, 1, 2, 3, 4
- Z = number of Heads - number of Tails
Possible Values: -4, -2, 0, 2, 4

Random Variables

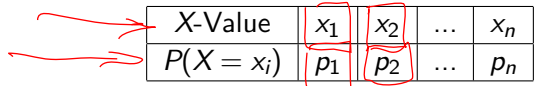
- Experiment: Measure a person's height:
Sample Space: All positive values
 - $X = \text{the height}$
Possible Values: All positive values
- Experiment: Ask how many children a person has:
Sample Space: 0,1,2,3....
 - $X = \text{The number of children}$
Possible Values: 0,1,2,3,4....
- Experiment: Ask the average amount of time a student studies
 - $X = \text{The average amount of time}$
Possible Values: All positive values

Probability Distribution

Now that we know there are all these possible values for the random variable X we want to think about their probability.

The **probability distribution** of a random variable is a function that assigns a probability to each possible value of X .

The probability distribution can be written as:



X-Value	x_1	x_2	...	x_n
$P(X = x_i)$	p_1	p_2	...	p_n

where each possible value x_i for X is listed with its probability p_i . Find each p_i by summing the probabilities of the elements such that $X = x_i$.

Probability Distribution

Example

Rolling a die (6 choices for 1 roll - 6^1).

Possible Values $S = \{1, 2, 3, 4, 5, 6\}$.

X = the values of the roll:

X -Value	1	2	3	4	5	6
$P(X = x_i)$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

Probability Distribution

Example

Flipping 3 Coins (2 choices for each coin - $2^3 = 8$).

Possible Values

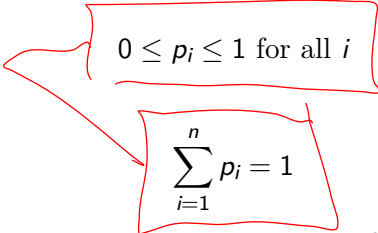
$S = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$.

X = The number of Heads

X-Value	0	1	2	3
$P(X = x)$	$1/8$	$3/8$	$3/8$	$1/8$

Probability Distribution

Two necessary conditions for probability distributions:


$$0 \leq p_i \leq 1 \text{ for all } i$$

$$\sum_{i=1}^n p_i = 1$$

If these conditions are not met, it is not a valid probability distribution.

Probability Distribution

Are these probability distributions?

X-Value	0	1	2	3
$P(X = x)$	0.1	0.25	0.3	0.3

NO

X-Value	0	1	2
$P(X = x)$	$1/3$	$1/3$	$1/3$

YES

X-Value	0	1	2
$P(X = x)$	-0.05	2	0

NO

Means & Expectations

Once we have the distribution of a random variable, we can figure out what value of X we would expect to see.

If all of the values are equally likely, we could just take the average.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

X-Value	0	1	2	3
$P(X = x)$	1/4	1/4	1/4	1/4

$$\bar{X} = \frac{0 + 1 + 2 + 3}{4} = 6/4 = 1.5$$

Note: the value will often not be one of the possible x -values.

Means & Expectations

What if all of the values were not equally likely? The average of the numbers should be closer to the values with the highest probability.

We find the **expected value** or **expectation** or **mean** of x , $E[X]$, using a weighted mean:

$$E[X] = \sum_{i=1}^n x_i \cdot p_i$$

X-Value	0	1	2	3
$P(X = x)$	0.3	0.45	0.2	0.05

$$E[X] = \sum_{i=1}^n x_i \cdot p_i = 0 \cdot 0.30 + 1 \cdot 0.45 + 2 \cdot 0.20 + 3 \cdot 0.05 = 1$$

Variance

The expected value is often the summary we will use of the center of a distribution. The **variance** is the measure we will use of the spread of the distribution.

Like the expectation, the variance can be thought of as a weighted mean.

$$\text{Var}[X] = \sum_{i=1}^n (x_i - E[X])^2 \cdot p_i$$

- a weighted average of the squared-difference (distance) between the x -values and the mean
- values with higher probability have more weight in the average

Variance

Example

X-Value	0	1	2	3
$P(X = x)$	0.3	0.45	0.2	0.05

$$E[X] = \sum_{i=1}^n x_i \cdot p_i = 0 \cdot 0.30 + 1 \cdot 0.45 + 2 \cdot 0.20 + 3 \cdot 0.05 = 1$$

$$\begin{aligned} \text{Var}[X] &= (0 - 1)^2 \cdot 0.30 + (1 - 1)^2 \cdot 0.45 + (2 - 1)^2 \cdot 0.20 + (3 - 1)^2 \cdot 0.05 \\ &= 1^2 \cdot 0.30 + 0 \cdot 0.45 + 1^2 \cdot 0.20 + 2^2 \cdot 0.05 \\ &= 0.30 + 0.20 + 0.20 = 0.70 \end{aligned}$$

Low variance means a tight/narrow distribution. A higher variance means a wider, flatter distribution.

Linear Functions of Random Variables

Sometimes we are interested in functions of random variables.

For example, if we know the mean and variance of X , do we know anything about $Y = 2X$ or $Y = X + 4$?

Or if we have a temperature information in Fahrenheit, can we find the mean and variance in Celsius?

As it turns out, expectations (or averages) are linear. If every number undergoes the same transformation, so does the mean.

Linear Functions of Random Variables

Mean Example

If $E[X] = 1$, then the average of the random variable X is 1. Now let's look at the random variable $Y = X + 4$.

What is $E[Y]$?

$X + 4$ add the constant 4 to every value of X

X-Value	0	1	2	3
$P(X = x)$	0.3	0.45	0.2	0.05

Y-Value	4	5	6	7
$P(Y = y)$	0.3	0.45	0.2	0.05

$$E[Y] = 4 \cdot 0.30 + 5 \cdot 0.45 + 6 \cdot 0.20 + 7 \cdot 0.05 = 1.20 + 2.25 + 1.20 + 0.35 = 5$$

$$E[Y] = E[X + 4] = E[X] + 4 = 1 + 4 = 5$$

Linear Functions of Random Variables

Mean Example

If $E[X] = 1$, then the average of the random variable X is 1. Now let's look at the random variable $Y = 2X$.

What is $E[Y]$?

$2X$ multiplies every X by 2:

X-Value	0	1	2	3
$P(X = x)$	0.3	0.45	0.2	0.05

Y-Value	0	2	4	6
$P(Y = y)$	0.3	0.45	0.2	0.05

$$E[Y] = 0 \cdot 0.30 + 2 \cdot 0.45 + 4 \cdot 0.20 + 6 \cdot 0.05 = 0.90 + 0.80 + 0.30 = 2$$

$$E[Y] = E[2X] = 2E[X] = 2 \cdot 1 = 2$$

Linear Functions of Random Variables

Variance Example, $Y=X+4$

X-Value	0	1	2	3
$P(X = x)$	0.3	0.45	0.2	0.05

Y-Value	4	5	6	7
$P(Y = y)$	0.3	0.45	0.2	0.05

$$\begin{aligned} \text{Var}[X] &= (0 - 1)^2 \cdot 0.30 + (1 - 1)^2 \cdot 0.45 + (2 - 1)^2 \cdot 0.20 + (3 - 1)^2 \cdot 0.05 \\ &= 1^2 \cdot 0.30 + 0 \cdot 0.45 + 1^2 \cdot 0.20 + 2^2 \cdot 0.05 \\ &= 0.30 + 0.20 + 0.20 = 0.70 \end{aligned}$$

$$\begin{aligned} \text{Var}[Y] &= (4 - 5)^2 \cdot 0.30 + (5 - 5)^2 \cdot 0.45 + (6 - 5)^2 \cdot 0.20 + (7 - 5)^2 \cdot 0.05 \\ &= 1^2 \cdot 0.30 + 0 \cdot 0.45 + 1^2 \cdot 0.20 + 2^2 \cdot 0.05 \\ &= 0.30 + 0.20 + 0.20 = 0.70 \end{aligned}$$

If we add 4 to every number, the spread is completely unaffected.
 $\text{Var}[Y] = \text{Var}[X + 4] = \text{Var}[X]$.

Linear Functions of Random Variables

Variance Example, $Y=2X$

X-Value	0	1	2	3
$P(X = x)$	0.3	0.45	0.2	0.05

Y-Value	0	2	4	6
$P(Y = y)$	0.3	0.45	0.2	0.05

$$\begin{aligned} \text{Var}[X] &= (0 - 1)^2 \cdot 0.30 + (1 - 1)^2 \cdot 0.45 + (2 - 1)^2 \cdot 0.20 + (3 - 1)^2 \cdot 0.05 \\ &= 1^2 \cdot 0.30 + 0 \cdot 0.45 + 1^2 \cdot 0.20 + 2^2 \cdot 0.05 \\ &= 0.30 + 0.20 + 0.20 = 0.70 \end{aligned}$$

$$\begin{aligned} \text{Var}[Y] &= (0 - 2)^2 \cdot 0.30 + (2 - 2)^2 \cdot 0.45 + (4 - 2)^2 \cdot 0.20 + (6 - 2)^2 \cdot 0.05 \\ &= 2^2 \cdot 0.30 + 0 \cdot 0.45 + 2^2 \cdot 0.20 + 4^2 \cdot 0.05 \\ &= 1.2 + 0.80 + 0.80 = 2.80 \end{aligned}$$

If we multiply every value by 2 we quadruple the spread.

$$\text{Var}[Y] = \text{Var}[2X] = 2^2 \text{Var}[X].$$

Summary of Expectations & Variances

$$E[aX + b] = aE[X] + b$$

$$\text{Var}[aX + b] = a^2 \text{Var}[X]$$

- $E[X] = 3, \text{Var}[X] = 1, Y = X + 1$
 $E[Y] = E[X] + 1 = 4, \text{Var}[Y] = \text{Var}[X] = 1$
- $E[X] = 4, \text{Var}[X] = 1/2, Y = 3X$
 $E[Y] = 3E[X] = 12, \text{Var}[Y] = 3^2 \text{Var}[X] = 9/2$
- $E[X] = 4, \text{Var}[X] = 1/2, Y = 3X + 1$
 $E[Y] = 3E[X] + 1 = 12 + 1 = 13, \text{Var}[Y] = 3^2 \text{Var}[X] = 9/2$

Probability Distribution Functions

Random Variables can be discrete or continuous.

A **discrete random variable** only gives values that you can list or count.

- Previous examples were all discrete distributions.

A **continuous random variable** gives an infinite number of values in a range.

- We can't list all of the possible values for a continuous random variable.

Examples

- X - number of books in a grad student's office.

DISCRETE $X = 0, 1, 2, 3, 4, \dots$

- Y - the amount of water it takes to fill a pool.

CONTINUOUS $Y \in (0, \infty)$

- Z - the time it takes to finish a task.

CONTINUOUS $Z \in (0, \infty)$

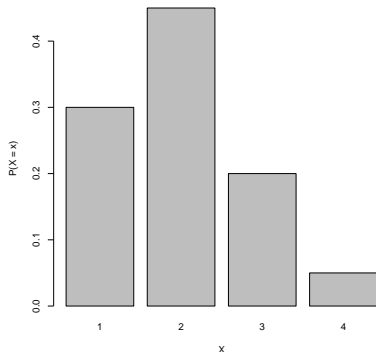
Probability Density Functions & Probability Mass Functions

The **probability distribution function** is defined differently for discrete and continuous random variables.

Discrete Random Variables

- **probability mass function** (pmf)
- Countable number of outcomes n , can write down

$$P(X = x_i) \quad \forall x_i, \quad i = 1, \dots, n$$

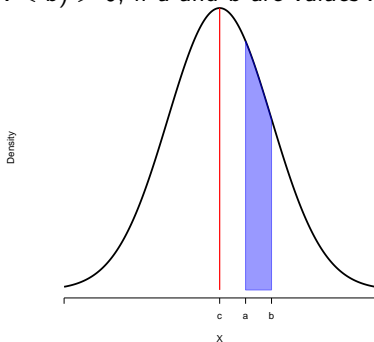


Probability Density Functions & Probability Mass Functions

The **probability distribution function** is defined differently for discrete and continuous random variables.

Continuous Random Variables

- **probability density function (pdf)**
- Too many (∞) possible values to write down.
- $P(X = c) = 0$ for any value of x .
- $P(a < X < b) > 0$, if a and b are values X can take on.



Expectations

Continuous Random Variables

Recall our definition of the expectation:

$$E[X] = \sum_{i=1}^n P(X = x_i) \cdot x_i = \sum_{i=1}^n p_i \cdot x_i$$

We can extend this definition to continuous distributions. For a continuous distribution, $P(X = x_i)$ is always zero, but we can compute the probability that X falls within a certain interval by integrating the pdf over that interval. If we divide up the real line into very small intervals, we can estimate $E[X]$ with

$$E[X] \cong \sum_{i=1}^n P(x_i < X < x_{i+1}) \cdot x_i$$

Expectations

Continuous Random Variables

By letting the number of rectangles approach infinity while their width approaches zero (and taking the limit), we obtain:

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

Example, continuous uniform distribution on $[0,5]$. $f(x) = 1/5$ for $x \in [0, 5]$.

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} xf(x)dx = \int_0^5 x \cdot 1/5 dx = 1/5 \int_0^5 x dx \\ &= 1/5 [x^2/2]_0^5 = 1/10 [5^2 - 0^2] = 25/10 = 2.5 \end{aligned}$$

Variance

Continuous Random Variables

We can extend the formula for the variance to continuous random variables in the same way. Recall the formula for the variance for a discrete distribution:

$$\text{Var}[X] = \sum_{i=1}^n (x_i - E[X])^2 \cdot P(X = x_i) = \sum_{i=1}^n (x_i - E[X])^2 \cdot p_i$$

Following the arguments we used for the expectation, we obtain the following formula for the variance:

$$\text{Var}[X] = \int_{-\infty}^{\infty} (x - E[X])^2 f(x) dx$$

Variance

Continuous Random Variables

Example, continuous uniform distribution on $[0, 5]$. $f(x) = 1/5$ for $x \in [0, 5]$. Previously we found that $E[X] = 2.5$.

$$\begin{aligned}\text{Var}[X] &= \int_{-\infty}^{\infty} (x - E[X])^2 f(x) dx = \int_0^5 (x - 2.5)^2 \cdot 1/5 dx \\&= 1/5 \int_0^5 (x - 2.5)^2 dx = 1/5 \left[\frac{1}{3} (x - 2.5)^3 \right]_0^5 \\&= 1/15 [(5 - 2.5)^3 - (0 - 2.5)^3] = 31.25/15 = 2.0833\end{aligned}$$