Data Science 1

STAT/CS 287

Jim Bagrow, UVM Dept of Math and Statistics

LECTURE 10

Here's an incredibly useful and powerful idea!

Suppose we have a sample of numeric data (a list of numbers)

$$x_1, x_2, \ldots, x_N$$

and we want to know something about its probability distribution $P(x)$

We could compute the histogram

Here's another idea→

Question

Imagine we take our data and **sort** (or rank) the numbers from smallest to largest.

Meaning we now know that

$$x_1 \leq x_2 \leq \cdots \leq x_N$$

is true

You can think of this sorting as *computing* a new ordering or indexing of the points (replacing the subscripts)

What *else* have we computed?

OK, let's build up a table

i	x_i
1	x_1
2	x_2
3	x_3
$\vdots$	$\vdots$
N	x_N

Let me swap these columns

OK, let's build up a table

x_i	i
x_1	1
x_2	2
x_3	3
$\vdots$	$\vdots$
x_N	N

How does i relate to x_i?

OK, let's build up a table

x_i	i	i-1
x_1	1	0
x_2	2	1
x_3	3	2
$\vdots$	$\vdots$	$\vdots$
x_N	N	N-1

What about $i - 1$?

It's the number of points $< x_i$

OK, let's build up a table

x_i	i	i-1	(i-1)/N
x_1	1	0	0
x_2	2	1	$^1/_N$
x_3	3	2	$^2/_N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_N	N	N-1	1-$^1/_N$

What about $(i-1)/N$?

It's the *fraction* of points $< x_i$

OK, let's build up a table

x_i	i	i-1	(i-1)/N
x_1	1	0	0
x_2	2	1	$^1/_N$
x_3	3	2	$^2/_N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_N	N	N-1	1-$^1/_N$

It's the *fraction* of points < x_i

=

Probability that a *randomly chosen element* of the sample is less than x_i

OK, let's build up a table

x_i	i	i-1	$(i\text{-}1)/N$	$\approx P(X < x_i)$
x_1	1	0	0	0
x_2	2	1	$^1/_N$	$^1/_N$
x_3	3	2	$^2/_N$	$^2/_N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_N	N	N-1	1-$^1/_N$	1-$^1/_N$

It's the *fraction* of points $< x_i$

=

Probability that a *randomly chosen element* of the sample is less than x_i

$P(X < x)$ is the cumulative distribution function (CDF)

For a random variable X with associated probability distribution $P(x)$

If X is continuous

$$P(X < x) = \int_{-\infty}^{x} P(x)\, dx$$

If X is discrete

$$P(X \leq x) = \sum_{x_i \leq x} P(X = x_i) = \sum_{x_i \leq x} P(x_i)$$

Complementary cumulative distribution function (CCDF): $P(X \geq x) = 1 - P(X < x)$

Back to our table

x_i	$(i\text{-}1)/N$	$\approx P(X < x_i)$
x_1	0	0
x_2	$^1/_N$	$^1/_N$
x_3	$^2/_N$	$^2/_N$
$\vdots$	$\vdots$	$\vdots$
x_N	$1\text{-}^1/_N$	$1\text{-}^1/_N$

← What we have is an empirical estimate of the CDF (Empirical CDF = ECDF)

$$P(X < x) \approx \frac{(\text{number of datapoints} < x)}{(\text{number of datapoints})}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{x_i < x}$$

The *fraction* of points $< x_i$

Recall our question

Imagine we take our data and **sort** (or rank) the numbers from smallest to largest.

Meaning we now know that

$$x_1 \leq x_2 \leq \cdots \leq x_N$$

is true

You can think of this sorting as *computing* a new ordering or indexing of the points (replacing the subscripts)

What *else* have we computed?

Recall our question

Imagine we take our data and **sort** (or rank) the numbers from smallest to largest.

You can think of this sorting as *computing* a new ordering or indexing of the points (replacing the subscripts)

Meaning we now know that

$$x_1 \leq x_2 \leq \cdots \leq x_N$$

is true

What *else* have we computed?

We have computed the **cumulative distribution**

sorting = integrating

🤯

Another question

x_i	$(i\text{-}1)/N$	$\approx P(X < x_i)$
x_1	0	0
x_2	${}^{1}/_{N}$	${}^{1}/_{N}$
x_3	${}^{2}/_{N}$	${}^{2}/_{N}$
$\vdots$	$\vdots$	$\vdots$
x_N	$1\text{-}{}^{1}/_{N}$	$1\text{-}{}^{1}/_{N}$

(x-axis) (y-axis)

What happens if we **plot** **these columns**?

$P(X < x)$ as a function of x…

Back to the notebook