DEEP: Data and Probabilities

Lecturer: Dimitar Kazakov

A Game of Dice

Red:	6	·	6	·	2	·	2	·	2	·	2
Green:	5	·	5	·	5	·	1	·	1	·	1
Purple:	4	·	4	·	4	·	4	·	0	·	0
Yellow:	3	·	3	·	3	·	3	·	3	·	3

Player 1 chooses a die.
Player 2 chooses a die.
Both roll.
The larger number wins.

Is there a winning strategy for Player 1? Is there one for Player 2?

Audience Poll

Which die would you choose?

Go to: **menti.com**

Enter code: **3290 2739**

What's the point of the mean?

		Average
Red:	6 · 6 · 2 · 2 · 2 · 2	. . .
Green:	5 · 5 · 5 · 1 · 1 · 1	. . .
Purple:	4 · 4 · 4 · 4 · 0 · 0	. . .
Yellow:	3 · 3 · 3 · 3 · 3 · 3	. . .

What's the point of the mean?

		Average
Red:	$\boxed{6} \cdot \boxed{6} \cdot \boxed{2} \cdot \boxed{2} \cdot \boxed{2} \cdot \boxed{2}$	$20/6$
Green:	$\boxed{5} \cdot \boxed{5} \cdot \boxed{5} \cdot \boxed{1} \cdot \boxed{1} \cdot \boxed{1}$	$18/6$
Purple:	$\boxed{4} \cdot \boxed{4} \cdot \boxed{4} \cdot \boxed{4} \cdot \boxed{0} \cdot \boxed{0}$	$16/6$
Yellow:	$\boxed{3} \cdot \boxed{3} \cdot \boxed{3} \cdot \boxed{3} \cdot \boxed{3} \cdot \boxed{3}$	$18/6$

What's the point of the mean?

		Average
Red:	6 · 6 · 2 · 2 · 2 · 2	20/6
Green:	5 · 5 · 5 · 1 · 1 · 1	18/6
Purple:	4 · 4 · 4 · 4 · 0 · 0	16/6
Yellow:	3 · 3 · 3 · 3 · 3 · 3	18/6

Let's play this game

Red vs Green

Let's count all possible outcomes and their probability:

6 vs 5: $\frac{2}{6} \times \frac{3}{6} = \frac{6}{36}$

Red vs Green

Let's count all possible outcomes and their probability:

6 vs 5: $\frac{2}{6} \times \frac{3}{6} = \frac{6}{36}$

6 vs 1: $\frac{2}{6} \times \frac{3}{6} = \frac{6}{36}$

Red vs Green

Let's count all possible outcomes and their probability:

6 vs 5: $\frac{2}{6} \times \frac{3}{6} = \frac{6}{36}$

6 vs 1: $\frac{2}{6} \times \frac{3}{6} = \frac{6}{36}$

2 vs 1: $\frac{4}{6} \times \frac{3}{6} = \frac{12}{36}$

Red vs Green

Let's count all possible outcomes and their probability:

- 6 vs 5: $\frac{2}{6} \times \frac{3}{6} = \frac{6}{36}$
- 6 vs 1: $\frac{2}{6} \times \frac{3}{6} = \frac{6}{36}$
- 2 vs 1: $\frac{4}{6} \times \frac{3}{6} = \frac{12}{36}$
- 2 vs 5: $\frac{4}{6} \times \frac{3}{6} = \frac{12}{36}$

Red vs Green

Let's count all possible outcomes and their probability:

6 vs 5: $\frac{2}{6} \times \frac{3}{6} = \frac{6}{36}$

6 vs 1: $\frac{2}{6} \times \frac{3}{6} = \frac{6}{36}$

2 vs 1: $\frac{4}{6} \times \frac{3}{6} = \frac{12}{36}$

2 vs 5: $\frac{4}{6} \times \frac{3}{6} = \frac{12}{36}$

Red wins 24/36=2/3rds of the time. Green wins 1/3rd of the time.

Red beats Green.

Yellow vs Red

Let's count all possible outcomes and their probability:

3 vs 6: $\frac{\cdots}{6} \times \frac{\cdots}{6} = \frac{\cdots}{36}$

3 vs 2: $\frac{\cdots}{6} \times \frac{\cdots}{6} = \frac{\cdots}{36}$

Yellow wins . . . of the time. Red wins . . . of the time.

. . . beats . . .

The meaning of numbers

Would Red beat Yellow if I replaced 6s on the Red with 8s?

		Average
Red:	8 · 8 · 2 · 2 · 2 · 2	24/6
Green:	5 · 5 · 5 · 1 · 1 · 1	18/6
Purple:	4 · 4 · 4 · 4 · 0 · 0	16/6
Yellow:	3 · 3 · 3 · 3 · 3 · 3	18/6

The meaning of numbers

Would Red beat Yellow if I replaced 6s on the Red with 8s?

		Average
Red:	8 · 8 · 2 · 2 · 2 · 2	24/6
Green:	5 · 5 · 5 · 1 · 1 · 1	18/6
Purple:	4 · 4 · 4 · 4 · 0 · 0	16/6
Yellow:	3 · 3 · 3 · 3 · 3 · 3	18/6

Here the numbers are used only to establish **an ordering** between outcomes – on their own, they have no meaning.

Ordinal Data

- When numbers on their own have meaning, so does the mean. E.g. average monthly salary.
- When dealing with ordinal data (e.g. A vs B preferences, Likert scale data), the mean is of no use. Instead, the median is reported as the 'central tendency measure', e.g. *half of the participants rate this ice cream between very good and excellent*.
- What's so special about the median?

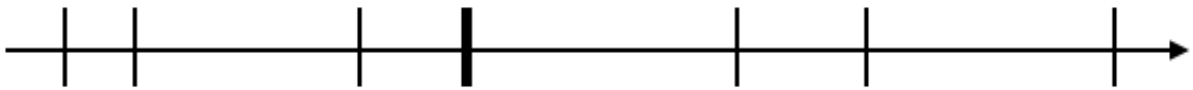
Median minimises sum of absolute differences

Let's consider the sum of absolute differences for a data set of N observations w.r.t. the parameter $\tilde{x}$, $S = \sum_i^N |x_i - \tilde{x}|$. Assume $\tilde{x}$ is set to the median, then consider the effect of changing the value of $\tilde{x}$ by an amount δ in either direction.

Median minimises sum of absolute differences

Let's consider the sum of absolute differences for a data set of N observations w.r.t. the parameter $\tilde{x}$, $S = \sum_i^N |x_i - \tilde{x}|$. Assume $\tilde{x}$ is set to the median, then consider the effect of changing the value of $\tilde{x}$ by an amount δ in either direction.

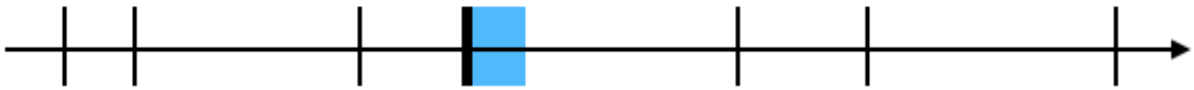
- When $N = 2k + 1$, moving $\tilde{x}$ away from the median by δ would change the sum S by $k.\delta - k.\delta + |\delta| > 0$



Median minimises sum of absolute differences

Let's consider the sum of absolute differences for a data set of N observations w.r.t. the parameter $\tilde{x}$, $S = \sum_i^N |x_i - \tilde{x}|$. Assume $\tilde{x}$ is set to the median, then consider the effect of changing the value of $\tilde{x}$ by an amount δ in either direction.

- When $N = 2k + 1$, moving $\tilde{x}$ away from the median by δ would change the sum S by $k.\delta - k.\delta + |\delta| > 0$



Median minimises sum of absolute differences

Let's consider the sum of absolute differences for a data set of N observations w.r.t. the parameter $\tilde{x}$, $S = \sum_i^N |x_i - \tilde{x}|$. Assume $\tilde{x}$ is set to the median, then consider the effect of changing the value of $\tilde{x}$ by an amount δ in either direction.

- When $N = 2k + 1$, moving $\tilde{x}$ away from the median by δ would change the sum S by $k.\delta - k.\delta + |\delta| > 0$
- When $N = 2k$, the median is a value between the k^{th} and $k + 1^{th}$ data point. S does not change for any $\tilde{x}$ in this interval, but increases if $\tilde{x}$ moves outside it by $(k + 1).|\delta| - (k - 1).|\delta| = 2.|\delta| > 0$

Comparing two sets of ordinal data

- When we compare samples of data, we often use statistical tests to establish if one data set (its mean and/or its variance) can be seen as different from another (with some level of confidence).
- When the two distributions are of known type (e.g. Gaussian), then it's easy to estimate how much they overlap.
- For ordinal data, we can use non-parametric tests for the same purpose. These tests only count the relative ordering of each item in one data set with respect to the other: my best 100m run time is better than all but two of your times, etc.
- To choose the appropriate test (=calculate the appropriate statistic), it is important to know whether the data is paired on unpaired.