Lecture 3

Comparing Loss Functions

DSC 40A, Spring 2024

Announcements

- Homework 1 is due on **Thursday, April 11th**.
    - Before working on it, watch the Walkthrough Videos on problem solving and using Overleaf.
    - Using the Overleaf template is required for Homework 2 (and only Homework 2).
- Remember that in, general, groupwork worksheets are released on Sunday and due Monday. → solutions will be posted!
- Look at the office hours schedule here and plan to start regularly attending!
- Remember to take a look at the supplementary readings linked on the course website.

Agenda

- Recap: Empirical risk minimization.
- Choosing a loss function.
    - The role of outliers.
- Center and spread.
- Towards linear regression.

Question 🤔

Answer at q.dsc40a.com

Remember, you can always ask questions at q.dsc40a.com!

Recap: Empirical risk minimization

Goal

We had one goal in Lecture 2: given a dataset of values from the past, **find the best constant prediction** to make.

$$y_1 = 72 \qquad y_2 = 90 \qquad y_3 = 61 \qquad y_4 = 85 \qquad y_5 = 92$$

Key idea: Different definitions of "best" give us different "best predictions."

mean

median

both the best, under different conditions!

The modeling recipe

In Lecture 2, we made two full passes through our "modeling recipe."

1. Choose a model.

$$H(x) = h$$

2. Choose a loss function.

$$L_{sq}(y_i, h) = (\overset{\text{actual}}{y_i} - \overset{\text{predicted}}{h})^2 \qquad L_{abs}(y_i, h) = |y_i - h|$$

3. Minimize average loss to find optimal model parameters.

$$\Downarrow \qquad\qquad\qquad\qquad \Downarrow$$

"mean squared error"

$$R_{sq}(h) = \frac{1}{n} \sum_{i=1}^{n} (y_i - h)^2$$

"mean absolute error"

$$R_{abs}(h) = \frac{1}{n} \sum_{i=1}^{n} |y_i - h|$$

$$\Rightarrow h^* = \text{Mean}(y_1, y_2, ..., y_n)$$

$$\Rightarrow h^* = \text{Median}(y_1, y_2, ..., y_n)$$

Empirical risk minimization

- The formal name for the process of minimizing average loss is **empirical risk minimization**.
- Another name for "average loss" is **empirical risk**.
- When we use the squared loss function, $L_{\text{sq}}(y_i, h) = (y_i - h)^2$, the corresponding empirical risk is mean squared error:

$$R_{\text{sq}}(h) = \frac{1}{n} \sum_{i=1}^{n} (y_i - h)^2$$

- When we use the absolute loss function, $L_{\text{abs}}(y_i, h) = |y_i - h|$, the corresponding empirical risk is mean absolute error:

$$R_{\text{abs}}(h) = \frac{1}{n} \sum_{i=1}^{n} |y_i - h|$$

Empirical risk minimization, in general

Key idea: If $L(y_i, h)$ is **any** loss function, the corresponding empirical risk is:

$$R(h) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, h)$$

Question 🤔

Answer at q.dsc40a.com

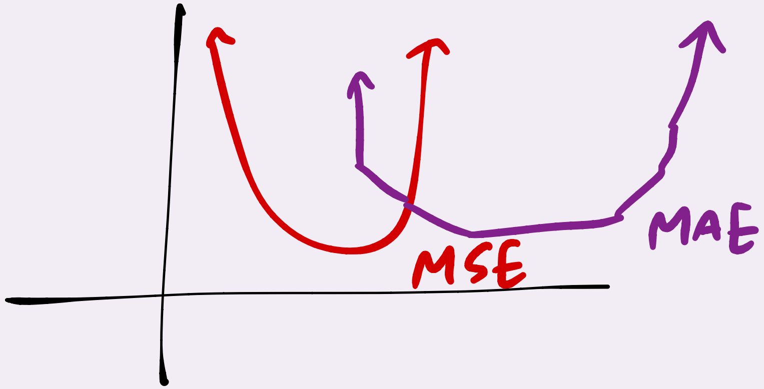
What questions do you have?

$$\left[R_{abs}(h)\right]^2 = \left(\frac{1}{n}\left(|y_1-h| + |y_2-h| + \dots + |y_n-h|\right)\right)^2$$

$$= \frac{1}{n^2}\left((y_1-h)^2 + |y_1-h| \cdot |y_2-h| + \dots\right) \neq R_{sq}(h)$$

Question 🤔

Answer at q.dsc40a.com



$$R_{\text{sq}}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h)^2$$ mean squared error

$$R_{\text{abs}}(h) = \frac{1}{n}\sum_{i=1}^{n}|y_i - h|$$ mean absolute error

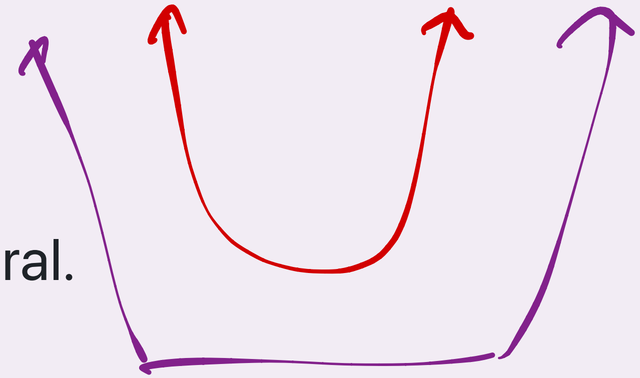
Is the following statement true, for any dataset $y_1, y_2, \dots, y_n$ and prediction h?

NOT TRUE IN GENERAL!

$$(R_{\text{abs}}(h))^2 = R_{\text{sq}}(h)$$

- ~~A. It's true for any h and any dataset.~~
- B. It's true for at least one h for any dataset, but not in general.
- ~~C. It's never true.~~

hard to say



Choosing a loss function

Now what?

- We know that, for the constant model $H(x) = h$, the **mean** minimizes mean **squared** error.
- We also know that, for the constant model $H(x) = h$, the **median** minimizes mean **absolute** error.
- **How does our choice of loss function impact the resulting optimal prediction?**

61, 72, 85, 90, 92 = 400

61, 72, 85, 90, 292

Comparing the mean and median

- Consider our example dataset of 5 commute times.

$$y_1 = 72 \qquad y_2 = 90 \qquad y_3 = 61 \qquad y_4 = 85 \qquad y_5 = 92$$

- As of now, the **median is 85** and the **mean is 80**.
- What if we add 200 to the largest commute time, 92?

$$y_1 = 72 \qquad y_2 = 90 \qquad y_3 = 61 \qquad y_4 = 85 \qquad y_5 = 292$$

- Now, the median is still 85 but the mean is 120 !
- **Key idea**: The mean is quite **sensitive** to outliers.

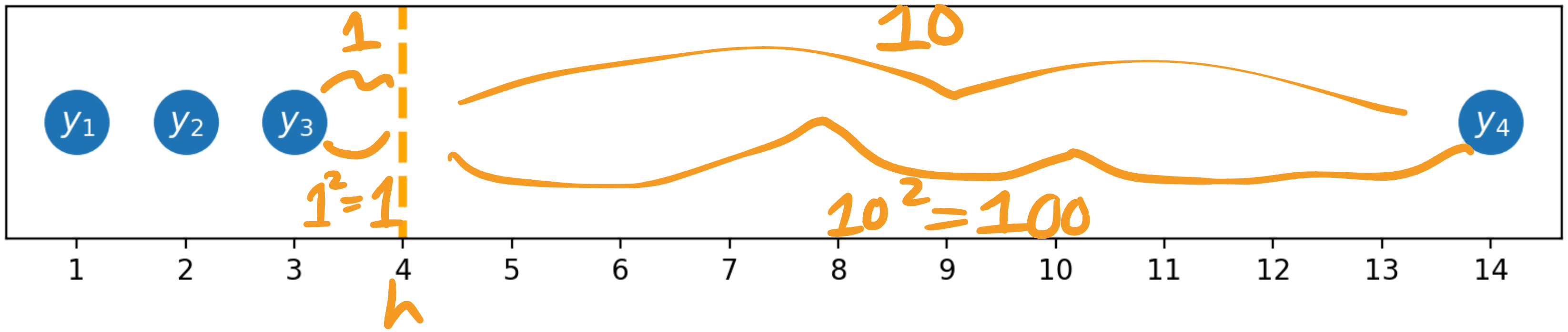
$80 + 40 = 120$

old mean

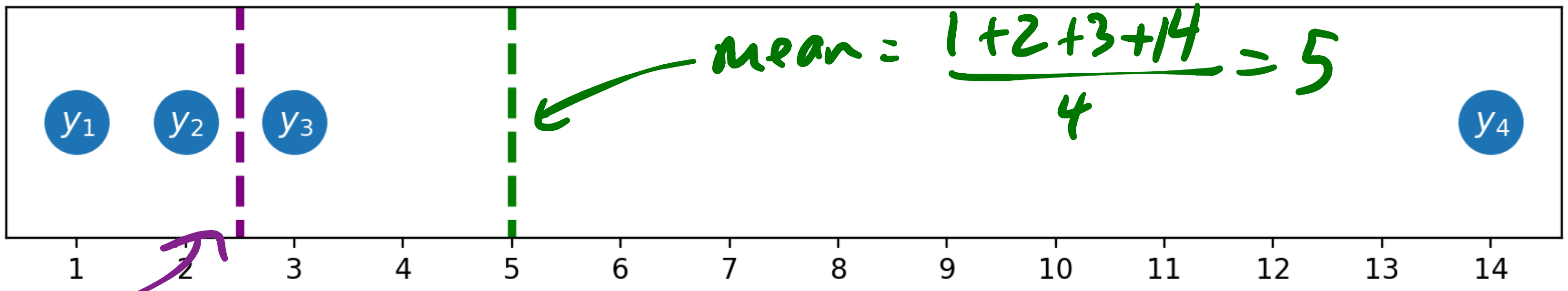
$\frac{200}{5} = 40$

Outliers

Below, $|y_4 - h|$ is 10 times as big as $|y_3 - h|$, but $(y_4 - h)^2$ is 100 times $(y_3 - h)^2$.



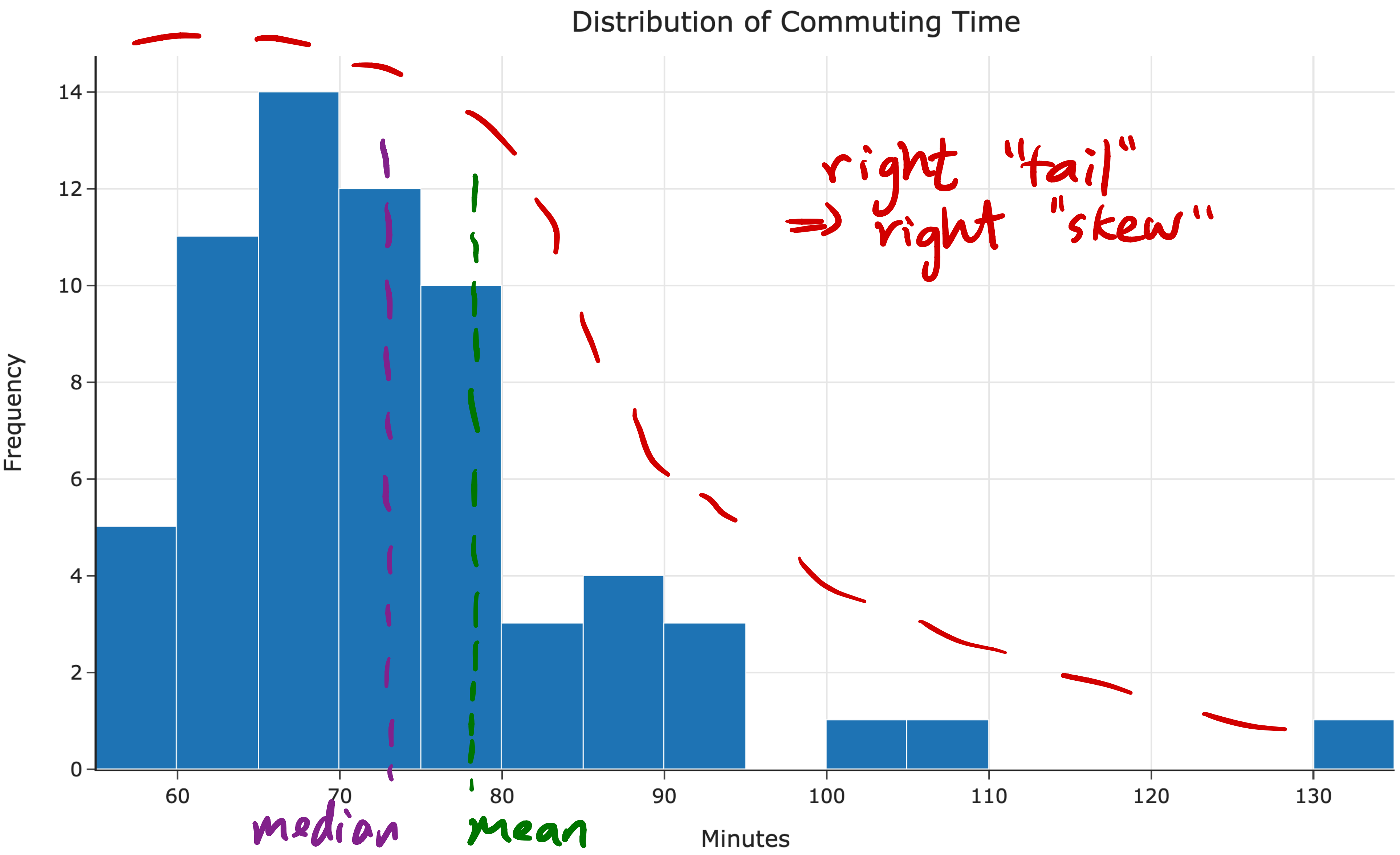
The result is that the **mean** is "pulled" in the direction of outliers, relative to the **median**.



mean: $\frac{1+2+3+14}{4} = 5$

median: 2.5

As a result, we say the **median** is **robust** to outliers. But the **mean** was easier to solve for.

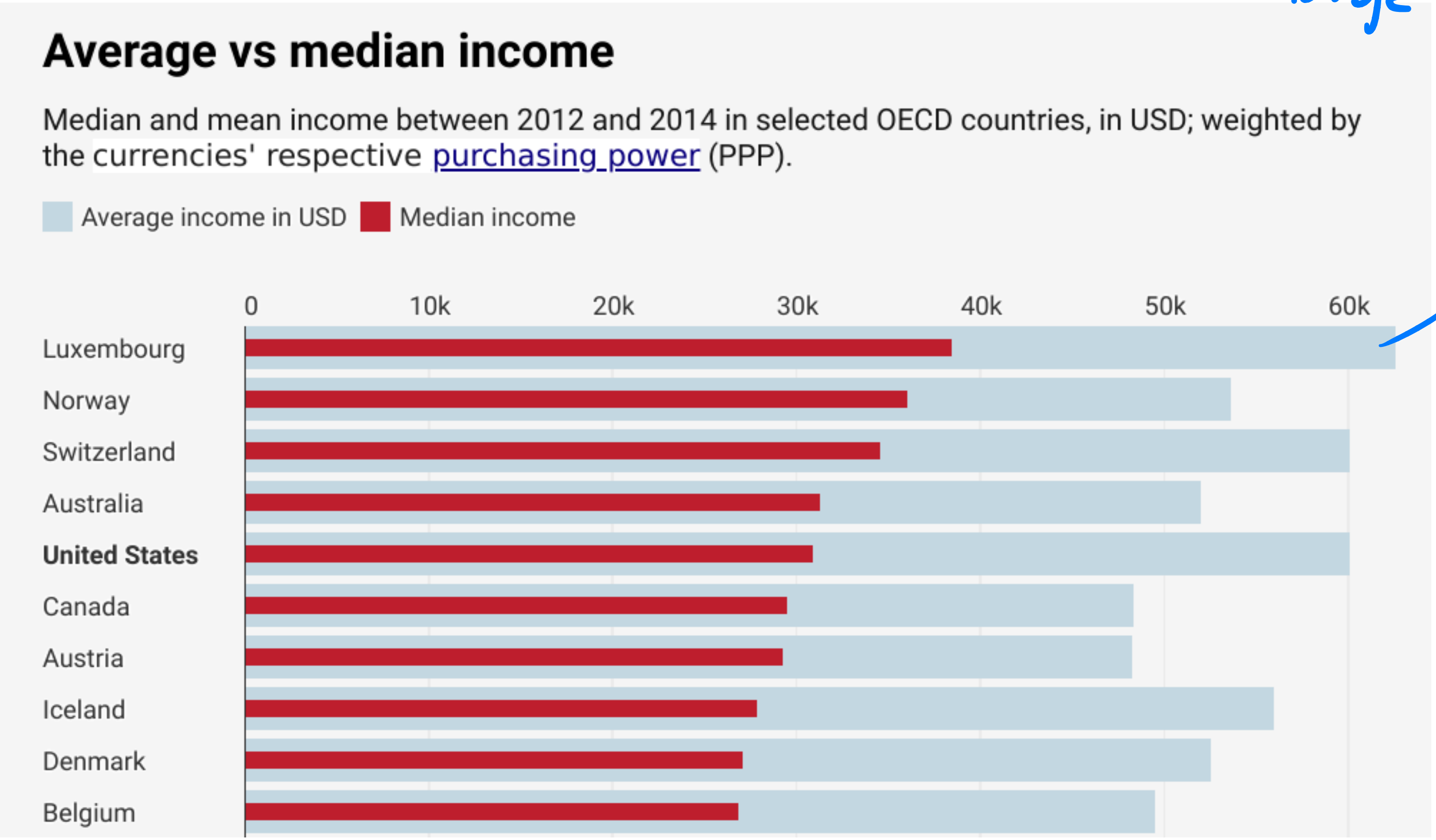


right "tail"

→ right "skew"

Example: Income inequality

mean is influenced by large outliers



Average vs median income

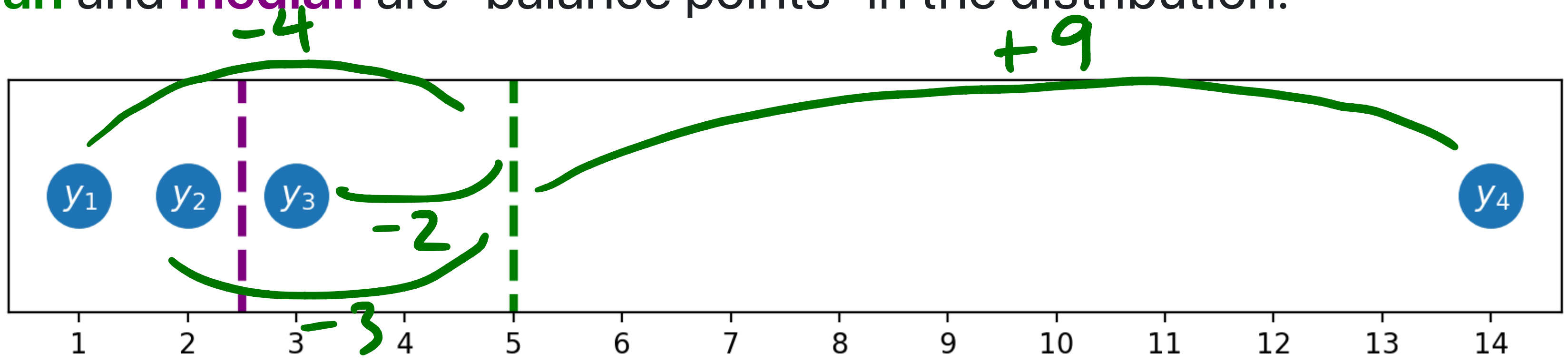
Median and mean income between 2012 and 2014 in selected OECD countries, in USD; weighted by the currencies' respective purchasing power (PPP).

For the mean: sum of distances below = sum of distances above

$$\sum_{y_i<h} |y_i - h| = \sum_{y_i>h} |y_i - h|$$

Balance points

Both the **mean** and **median** are "balance points" in the distribution.



- The **mean** is the point where $\sum_{i=1}^{n} (y_i - h) = 0$.
  - This appears in Homework 1!

- The **median** is the point where $\# (y_i < h) = \# (y_i > h)$.

Here: 2 points to the left of median (y_1, y_2),
2 points to the right of median (y_3, y_4)

Why stop at squared loss?

$\frac{1}{n}\left[\#(y_i < h) - \#(y_i > h)\right]$

mean squared error

Empirical Risk, $R(h)$	Derivative of Empirical Risk, $\frac{d}{dh}R(h)$	Minimizer
$\frac{1}{n}\sum_{i=1}^{n}\|y_i - h\|$	$\frac{1}{n}\left(\sum_{y_i<h} 1 - \sum_{y_i>h} 1\right)$	median
$\frac{1}{n}\sum_{i=1}^{n}(y_i - h)^2$	$\frac{-2}{n}\sum_{i=1}^{n}(y_i - h)$ set this to 0 to find the mean!	mean
$\frac{1}{n}\sum_{i=1}^{n}\|y_i - h\|^3$		???
$\frac{1}{n}\sum_{i=1}^{n}(y_i - h)^4$	$-\frac{4}{n}\sum_{i=1}^{n}(y_i-h)^3 = 0$	???
$\frac{1}{n}\sum_{i=1}^{n}(y_i - h)^{100}$		???
...	...	...

if using an odd exponent, need absolute value because $(-x)^3 \Rightarrow$ negative!

Generalized L_p loss

For any $p \geq 1$, define the L_p loss as follows:

$$L_p(y_i, h) = |y_i - h|^p$$

The corresponding empirical risk is:

$$R_p(h) = \frac{1}{n} \sum_{i=1}^{n} |y_i - h|^p$$

- When $p = 1$, $h^* = \text{Median}(y_1, y_2, \dots, y_n)$.
- When $p = 2$, $h^* = \text{Mean}(y_1, y_2, \dots, y_n)$.
- What about when $p = 3$?
- What about when $p \to \infty$?

"p-norm of a vector"

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}$$

generalized Pythagorean theorem

$$\|x\|_3 = \sqrt[3]{x_1^3 + x_2^3 + \dots + x_n^3}$$

$$\|x\|_{100} = \sqrt[100]{x_1^{100} + x_2^{100} + \dots + x_n^{100}}$$

$$\vdots$$

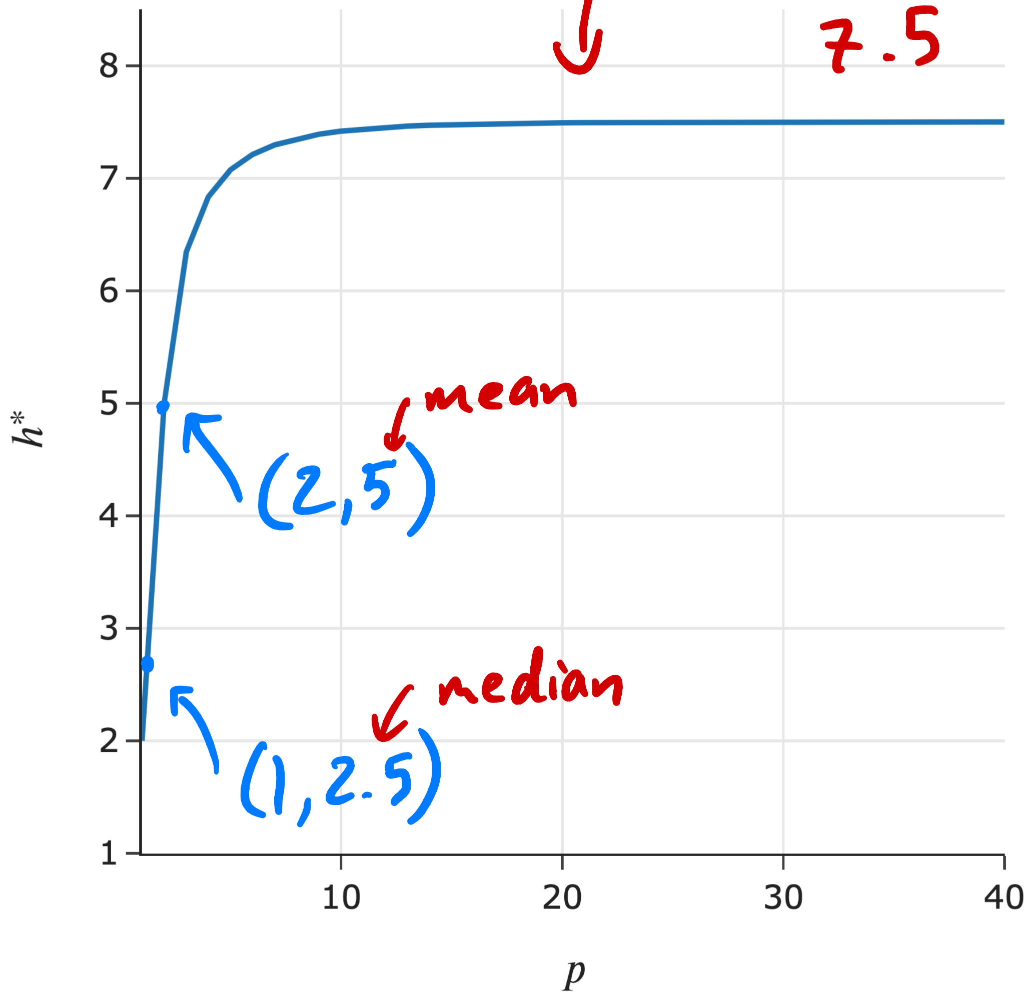
$$\|x\|_\infty = \max(x_1, x_2, \dots, x_n)$$

related idea

What value does h^* approach, as $p \to \infty$?

halfway between 1 and 14!

7.5



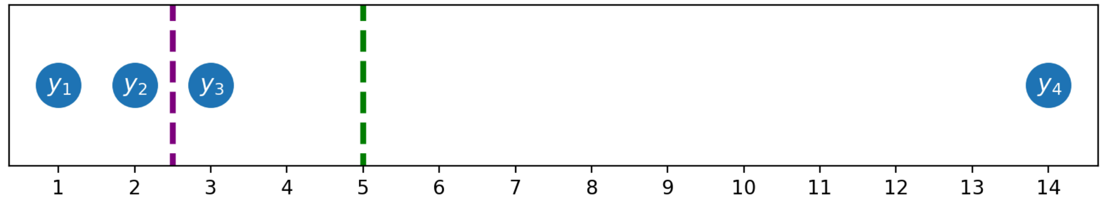
mean

(2,5)

median

(1,2.5)

Consider the dataset $1, 2, 3, 14$:



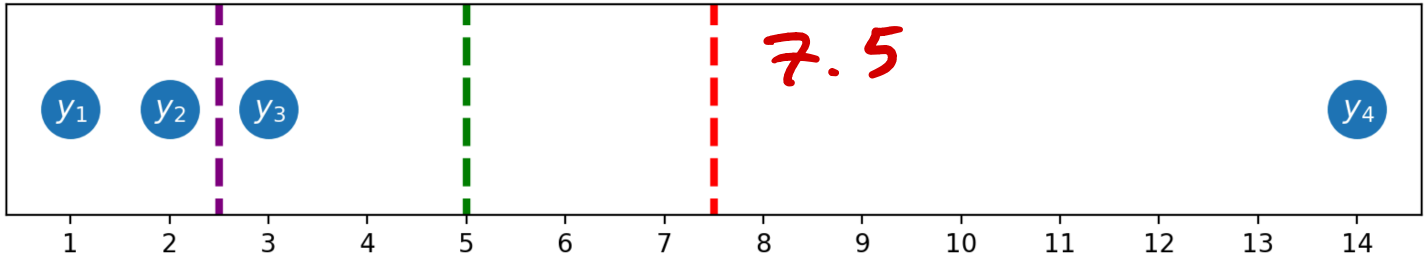
On the left:

- The x-axis is p.
- The y-axis is h^*, the optimal constant prediction for L_p loss:

$$h^* = \underset{h}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} |y_i - h|^p$$

The *midrange* minimizes average $\overbrace{L_\infty}^{\text{"infinity loss"}}$ loss!

On the previous slide, we saw that as $p \to \infty$, the minimizer of mean L_p loss approached **the midpoint of the minimum and maximum values in the dataset**, or the midrange.



- As $p \to \infty$, $R_p(h) = \frac{1}{n}\sum_{i=1}^{n}|y_i - h|^p$ minimizes **the "worst case" distance from any data point"**. (Read more here).
- If your measure of "good" is "not far from any one data point", then the midrange is the best prediction.

mean = 5, worst case distance = $|14-5| = 9$

median = 2.5, worst case distance = $|14-2.5| = 11.5$

midrange = 7.5, worst case distance = $|14-7.5| = 6.5$

Another example: 0-1 loss

Consider, for example, the **0-1 loss**:

$$L_{0,1}(y_i, h) = \begin{cases} 0 & y_i = h \\ 1 & y_i \neq h \end{cases}$$

The corresponding empirical risk is:

$$R_{0,1}(h) = \frac{1}{n} \sum_{i=1}^{n} L_{0,1}(y_i, h)$$

Question 🤔

Answer at q.dsc40a.com

$$R_{0,1}(h) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}0 & y_i = h\\ 1 & y_i \neq h\end{cases}$$ = proportion of points NOT equal to h

Suppose $y_1, y_2, \ldots, y_n$ are all unique. What is $R_{0,1}(y_1)$?

- A. 0.
- B. $\frac{1}{n}$.
- [✓ C. $\frac{n-1}{n}$.]
- D. 1.

$R_{0,1}(y_1)$ = proportion of points $\neq y_1$

$$= \boxed{\frac{n-1}{n}}$$

Minimizing empirical risk for 0-1 loss

$$R_{0,1}(h) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}0 & y_i = h\\1 & y_i \neq h\end{cases}$$

= proportion of points NOT equal to h

1, 2, 3, 14

mode is not usually unique!

Minimized when $y_i = h$ as often as possible

$\Rightarrow$ set $h^* = \text{Mode}(y_1, y_2, \ldots, y_n)$

most common value!

Summary: Choosing a loss function

Key idea: Different loss functions lead to different best predictions, h^*!

Loss	Minimizer	Always Unique?	Robust to Outliers?	Differentiable?
L_{sq}	mean	yes ✅	no ❌	yes ✅
L_{abs}	median	no ❌	yes ✅	no ❌
L_∞	midrange	yes ✅	no ❌	no ❌
$L_{0,1}$	mode	no ❌	yes ✅	no ❌

The optimal predictions, h^*, are all **summary statistics** that measure the **center** of the dataset in different ways.

e.g. 1, 2, 3, 5

mean $= \frac{1+2+3+5}{4}$

$= \frac{11}{4} = 2.75$

median $= 2.5$

2.4

2.7

2.0001

all minimize mean absolute error!

NOT UNIQUE

Center and spread

What does it mean?

- The general form of empirical risk, for any loss function $L(y_i, h)$, is:

$$R(h) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, h)$$

- As we just saw, the input h^* that minimizes $R(h)$ is some measure of the **center** of the dataset.
    - Examples include the mean (L_{sq}), median (L_{abs}), and mode ($L_{0,1}$).
- The minimum output, $R(h^*)$, represents some measure of the **spread**, or variation, in the dataset.

Squared loss

- The empirical risk for squared loss, i.e. mean squared error, is:

$$R_{\text{sq}}(h) = \frac{1}{n} \sum_{i=1}^{n} (y_i - h)^2$$

- $R_{\text{sq}}(h)$ is minimized when $h^* = \text{Mean}(y_1, y_2, \ldots, y_n)$.
- Therefore, the minimum value of $R_{\text{sq}}(h)$ is:

$$\begin{aligned} R_{\text{sq}}(h^*) &= R_{\text{sq}}\left(\text{Mean}(y_1, y_2, \ldots, y_n)\right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \text{Mean}(y_1, y_2, \ldots, y_n)\right)^2 \end{aligned}$$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Variance

- The minimum value of $R_{\text{sq}}(h)$ is the mean squared deviation from the mean, more commonly known as the **variance**.

$$\text{Variance}(y_1, y_2, \ldots, y_n) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}}_{\text{average}} (y_i - \overbrace{\text{Mean}(y_1, y_2, \ldots, y_n)}^{\bar{y}})^2$$

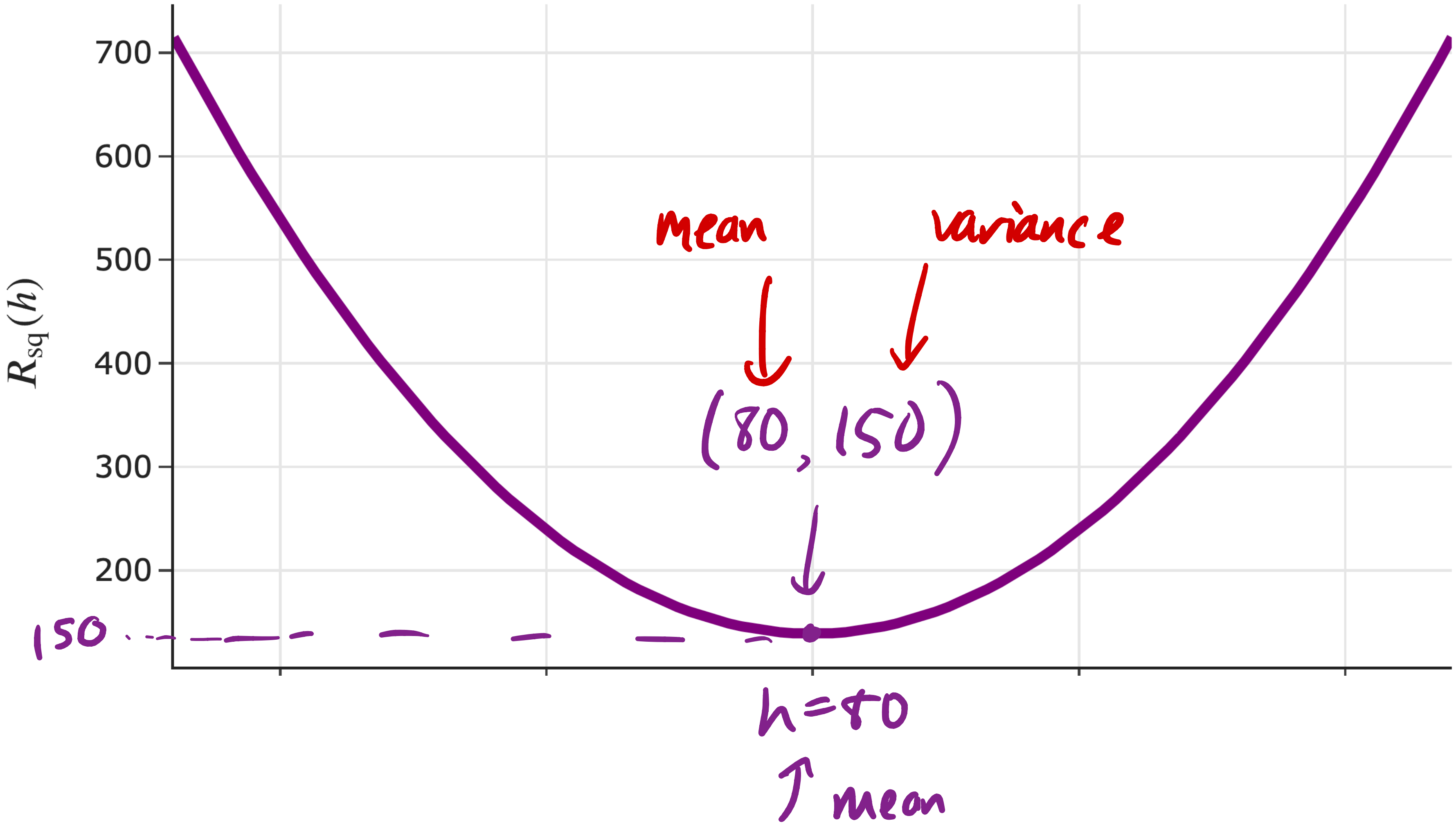
squared distance from mean

- It measures the squared distance of each data point from the mean, on average.
- Its square root is called the **standard deviation**.

$$R_{sq}(h^*) = \text{Variance},$$

$$h^* = \text{Mean}$$

$$R_{\text{sq}}(h) = \frac{1}{5}\left((72-h)^2 + (90-h)^2 + (61-h)^2 + (85-h)^2 + (92-h)^2\right)$$



Absolute loss

- The empirical risk for absolute loss, i.e. mean absolute error, is:

$$R_{\text{abs}}(h) = \frac{1}{n} \sum_{i=1}^{n} |y_i - h|$$

- $R_{\text{abs}}(h)$ is minimized when $h^* = \text{Median}(y_1, y_2, \dots, y_n)$.
- Therefore, the minimum value of $R_{\text{abs}}(h)$ is:

$$\begin{aligned} R_{\text{abs}}(h^*) &= \frac{1}{n} \sum_{i=1}^{n} |y_i - h| \\ &= R_{\text{abs}}(h) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \text{Median}(y_1, y_2, \dots, y_n)| \end{aligned}$$

Mean absolute deviation from the median

- The minimum value of $R_{\text{abs}}(h)$ is the **mean absolute deviation from the median**.

$$\text{MAD from the median}(y_1, y_2, \dots, y_n) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \text{Median}(y_1, y_2, \dots, y_n)|$$

- It measures how far each data point is from the median, on average.
- **Example**: What's the MAD from the median in the dataset $2, 3, 3, 4, 5$?

median: 3

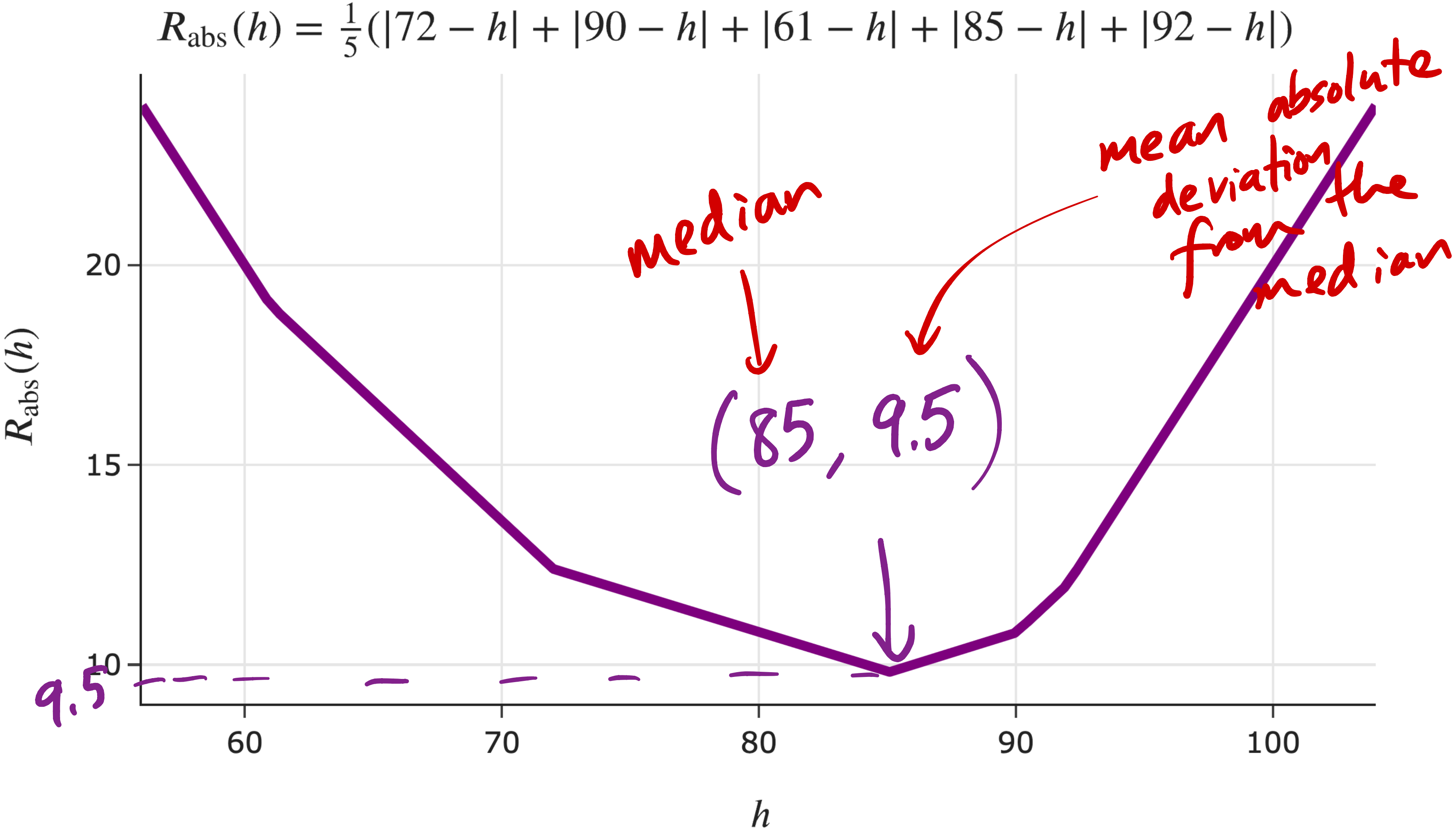
$$|2-3| = 1$$
$$|3-3| = 0$$
$$|3-3| = 0$$
$$|4-3| = 1$$
$$|5-3| = 2$$

$$\rightarrow \frac{1+0+0+1+2}{5} = \boxed{\frac{4}{5}}$$

mean abs dev. from median

$R_{abs}(h^*) =$ mean absolute deviation from median

Mean absolute deviation from the median



0-1 loss

- The empirical risk for the 0-1 loss is:

$$R_{0,1}(h) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}0 & y_i = h\\ 1 & y_i \neq h\end{cases}$$

- This is the proportion (between 0 and 1) of data points not equal to h.
- $R_{0,1}(h)$ is minimized when $h^* = \text{Mode}(y_1, y_2, \dots, y_n)$.
- Therefore, $R_{0,1}(h^*)$ is the proportion of data points not equal to the mode.
- **Example**: What's the proportion of values not equal to the mode in the dataset $2, 3, 3, 4, 5$?

A poor way to measure spread

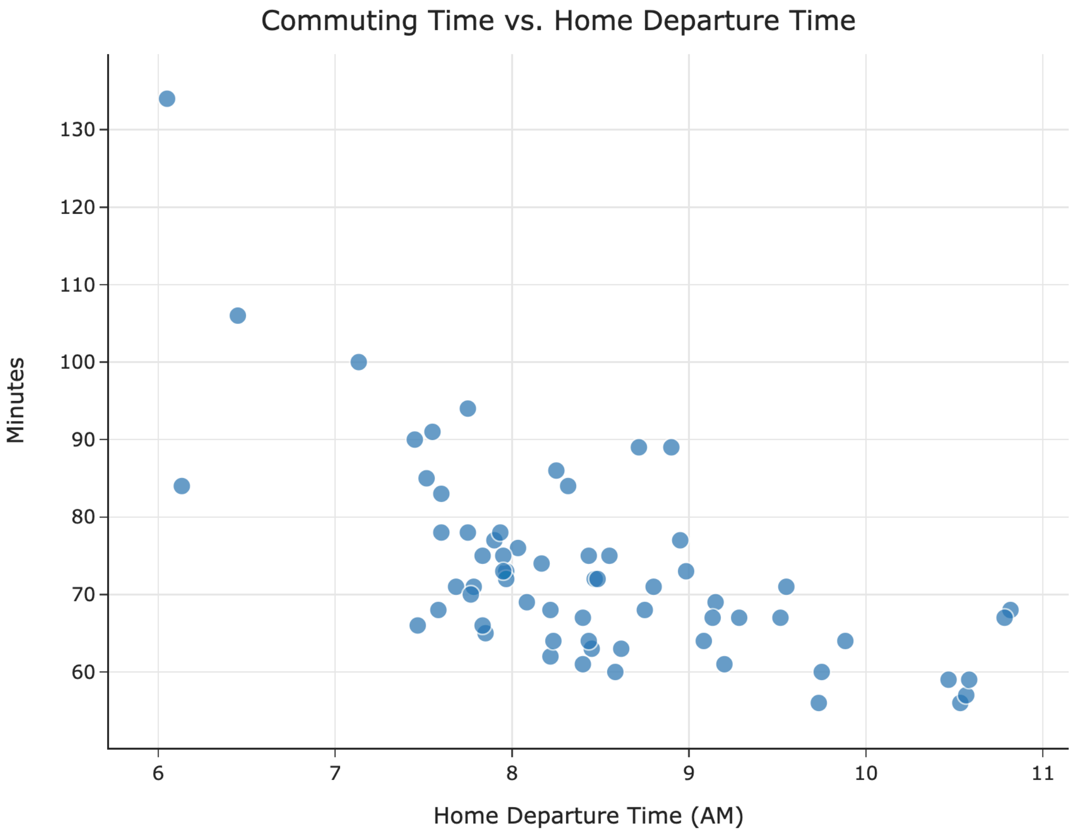
- The minimum value of $R_{0,1}(h)$ is the proportion of data points not equal to the mode.
- A higher value means less of the data is clustered at the mode.
- Just as the mode is a very basic way of measuring the center of the data, $R_{0,1}(h^*)$ is a very basic and uninformative way of measuring spread.

Summary of center and spread

- Different loss functions $L(y_i, h)$ lead to different empirical risk functions $R(h)$, which are minimized at various measures of **center**.
- The minimum values of empirical risk, $R(h^*)$, are various measures of **spread**.
- There are many different ways to measure both center and spread; these are sometimes called **descriptive statistics**.

What's next?

Towards simple linear regression



- In Lecture 1, we introduced the idea of a hypothesis function, $H(x)$.
- We've focused on finding the best **constant model**, $H(x) = h$.
- Now that we understand the modeling recipe, we can apply it to find the best **simple linear regression model**, $H(x) = w_0 + w_1 x$.
- This will allow us to make predictions that aren't all the same for every data point.

The modeling recipe

1. Choose a model.

2. Choose a loss function.

3. Minimize average loss to find optimal model parameters.