Summarizing Data

Research Methods for Political Science

Thomas Chadefaux

Trinity College Dublin

Measures of central tendency

The median

```
##   [1] 51 51 53 53 54 54 54 54 54 54 55 55 55 55 55 56 56 56 5
##  [26] 57 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 59 59 5
##  [51] 59 59 60 60 60 60 60 60 60 61 61 61 61 61 61 61 61 61 6
##  [76] 62 63 63 63 63 63 63 63 63 64 64 64 64 64 64 64 65 65 6
## [101] 69
```

Median: Value which splits a distribution in half.

```
median(grades)
```

```
## [1] 59
```

What if there is an even number of observations?

```
median(c(1,2,3,4))
```

```
## [1] 2.5
```

The mean

The mean is often written as $\bar{x}$ and is calculated as:

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

In R, you calculate it very simply as:

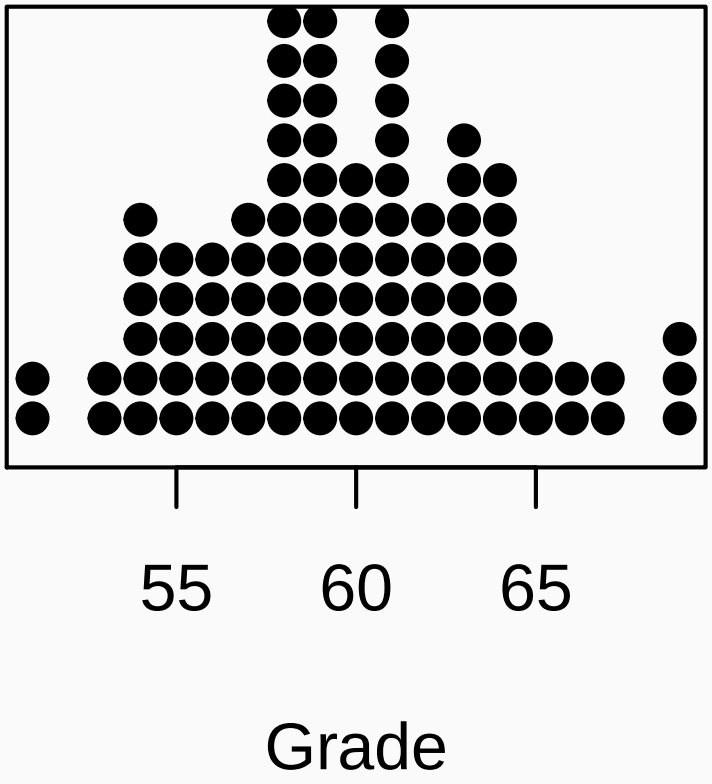
```
mean(grades)
```

```
## [1] 59.71287
```

The Mode

Value that is observed with the greatest frequency

Each dot represents one stude



Choosing measures of central tendency

The choice often depends on the type of variable

- Categorical data: Mode
- Quantity data: mean or median

Mean or median?

The mean is stable: varies little. E.g:

- sample1: $1, 2, \ , 9, 10$
- sample2: $1, 2, 3, 9, 10$

```
# The mean remains stable...
mean(sample1); mean(sample2)
```

```
## [1] 5.5
```

```
## [1] 5
```

```
# ...but the median changes dramatically
median(sample1); median(sample2)
```

```
## [1] 5.5
```

```
## [1] 3
```

Mean vs median

But sometimes the median is more representative. E.g. sample incomes:

- sample1: 30, 50, 70, 75, 80
- sample1: 30, 50, 70, 75, 290

```
mean(sample1); mean(sample2)
```

```
## [1] 61
```

```
## [1] 103
```

```
median(sample1); median(sample2)
```

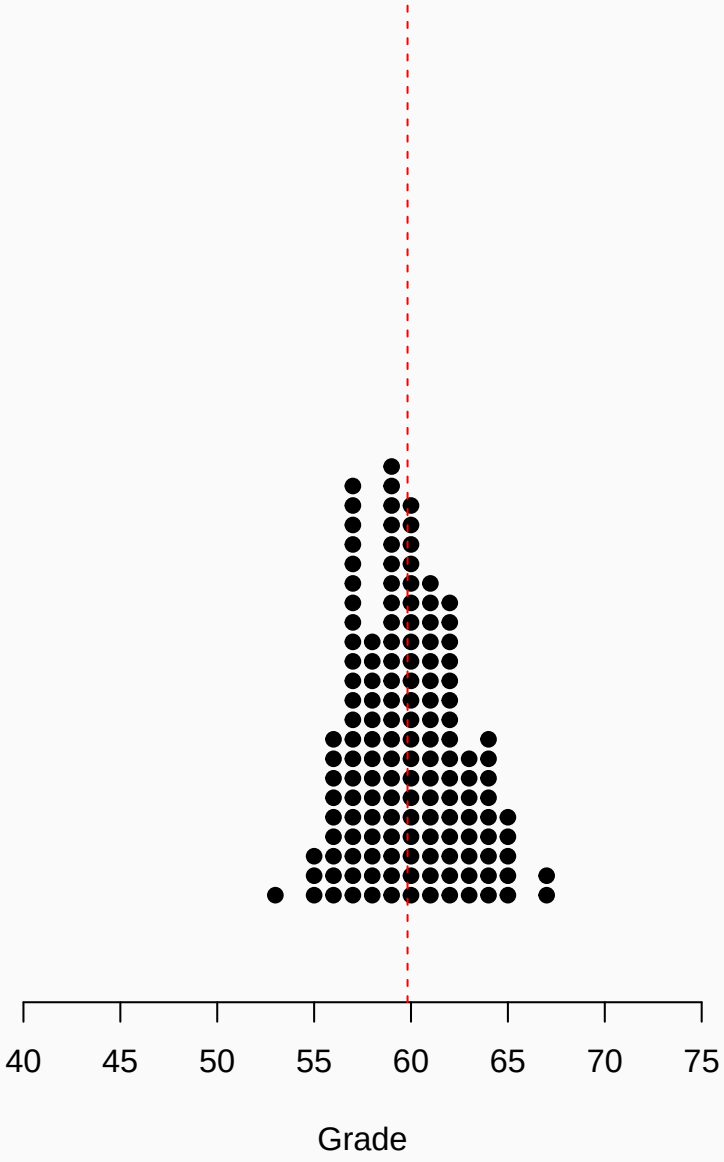
```
## [1] 70
```

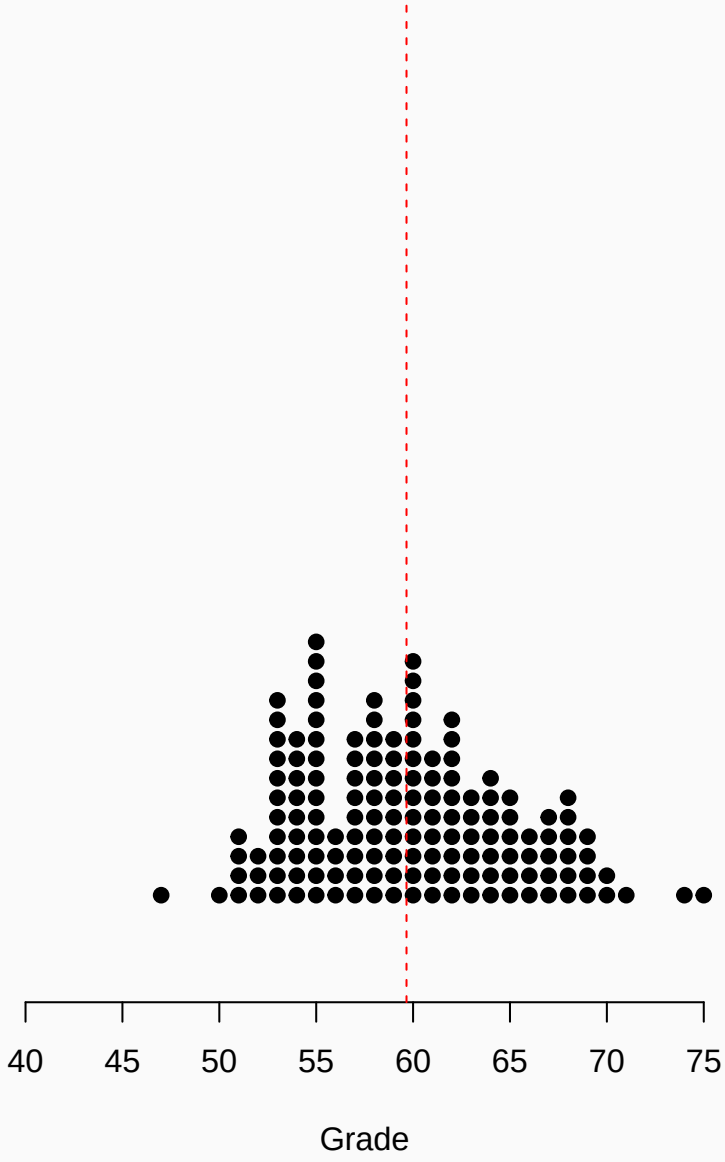
```
## [1] 70
```

Measures of Dispersion

What do we mean by “Dispersion”

Each dot represents one student





Quantifying spread: the range

Difference between largest and smallest value

```
max(sample1) - min(sample1)
```

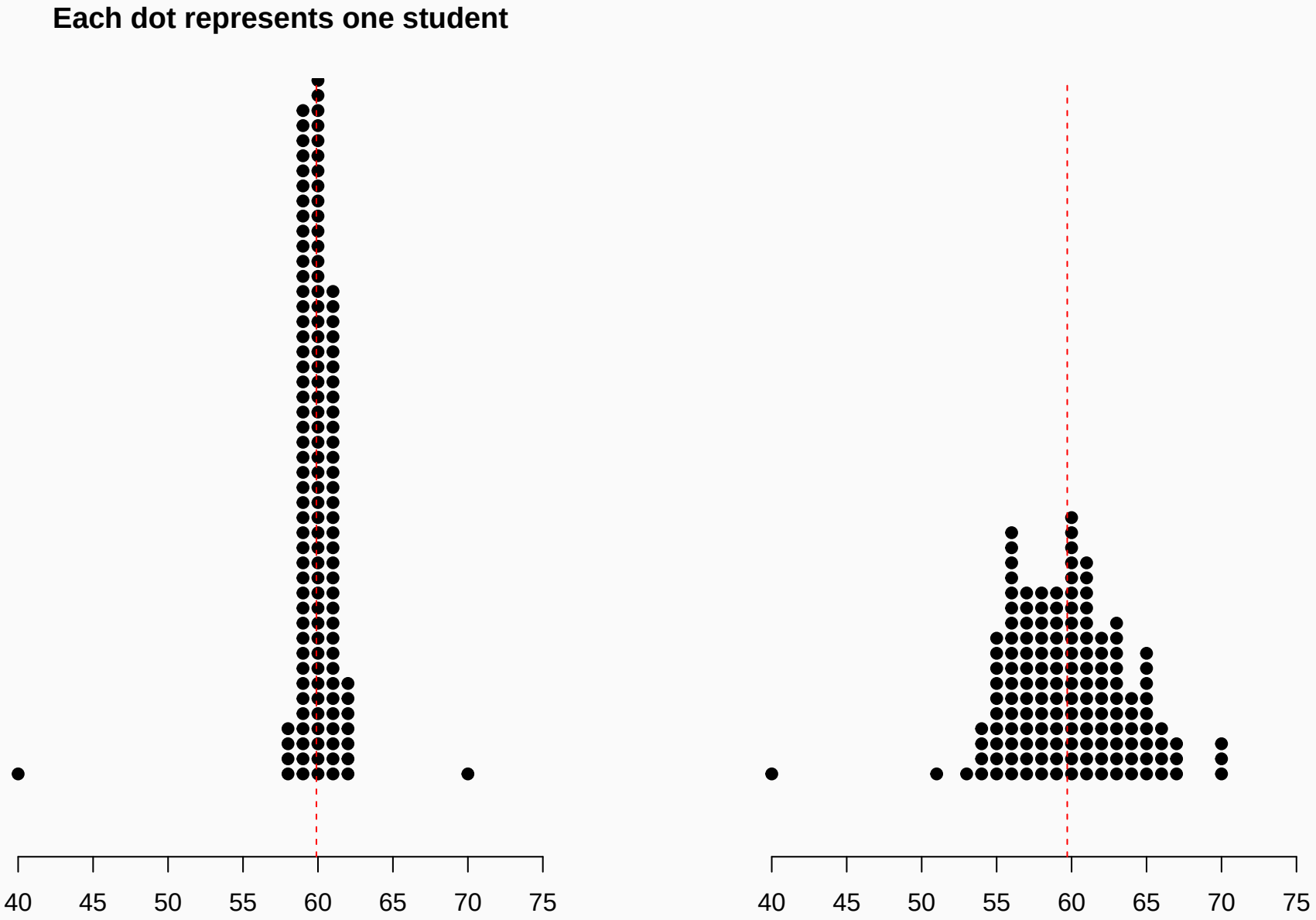
```
## [1] 14
```

```
max(sample2) - min(sample2)
```

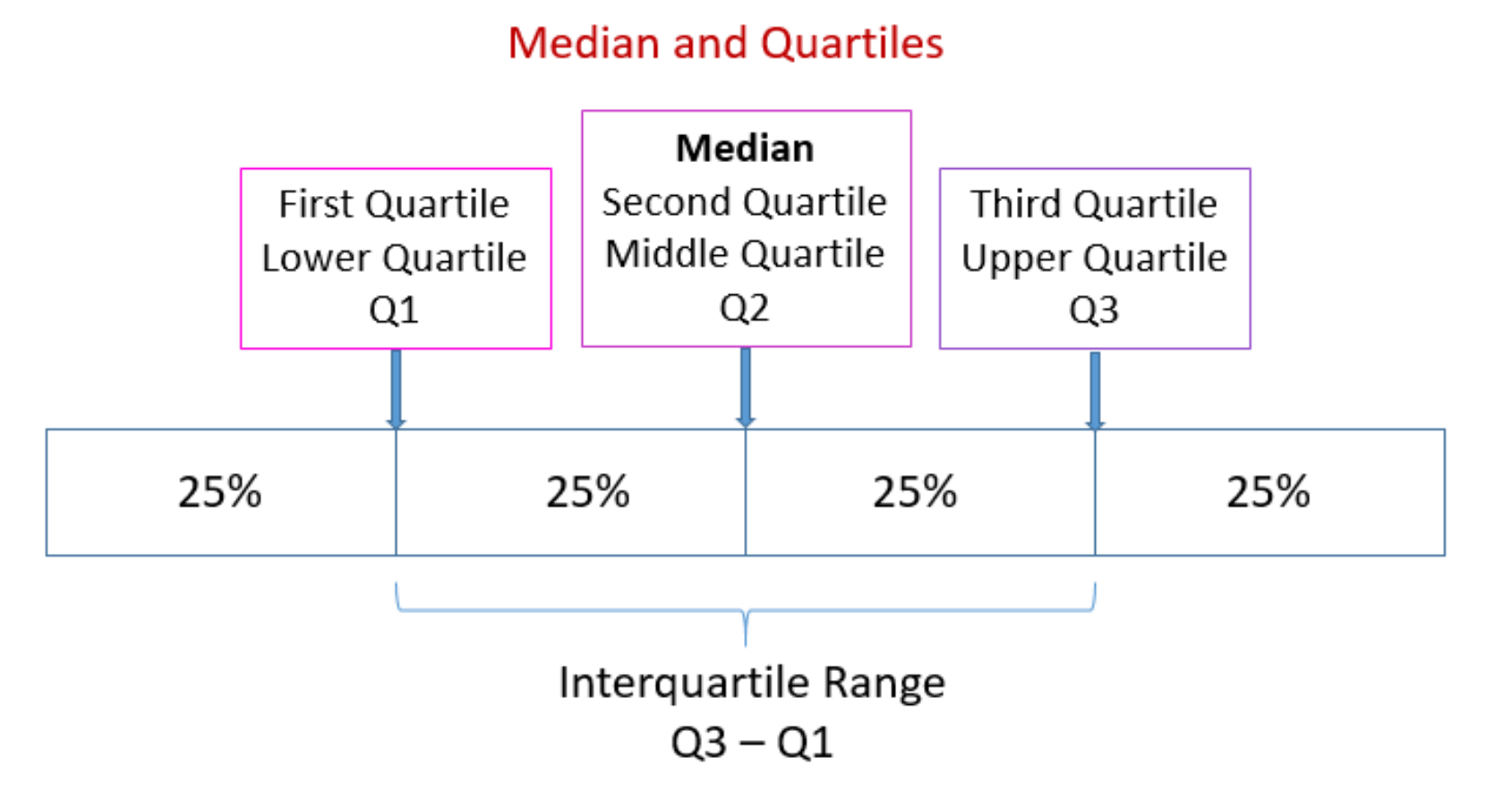
```
## [1] 28
```

Quantifying spread: the range

But the range has limits. E.g., these have the same range:



Quantifying spread: Quartiles



Quantifying spread: Quartiles

There are three quartiles. The second quartile is the median. The range between Q1 and Q3 is called the inter-quartile range

```
quantile(sample1, probs = c(0.25, 0.5, 0.75))
```

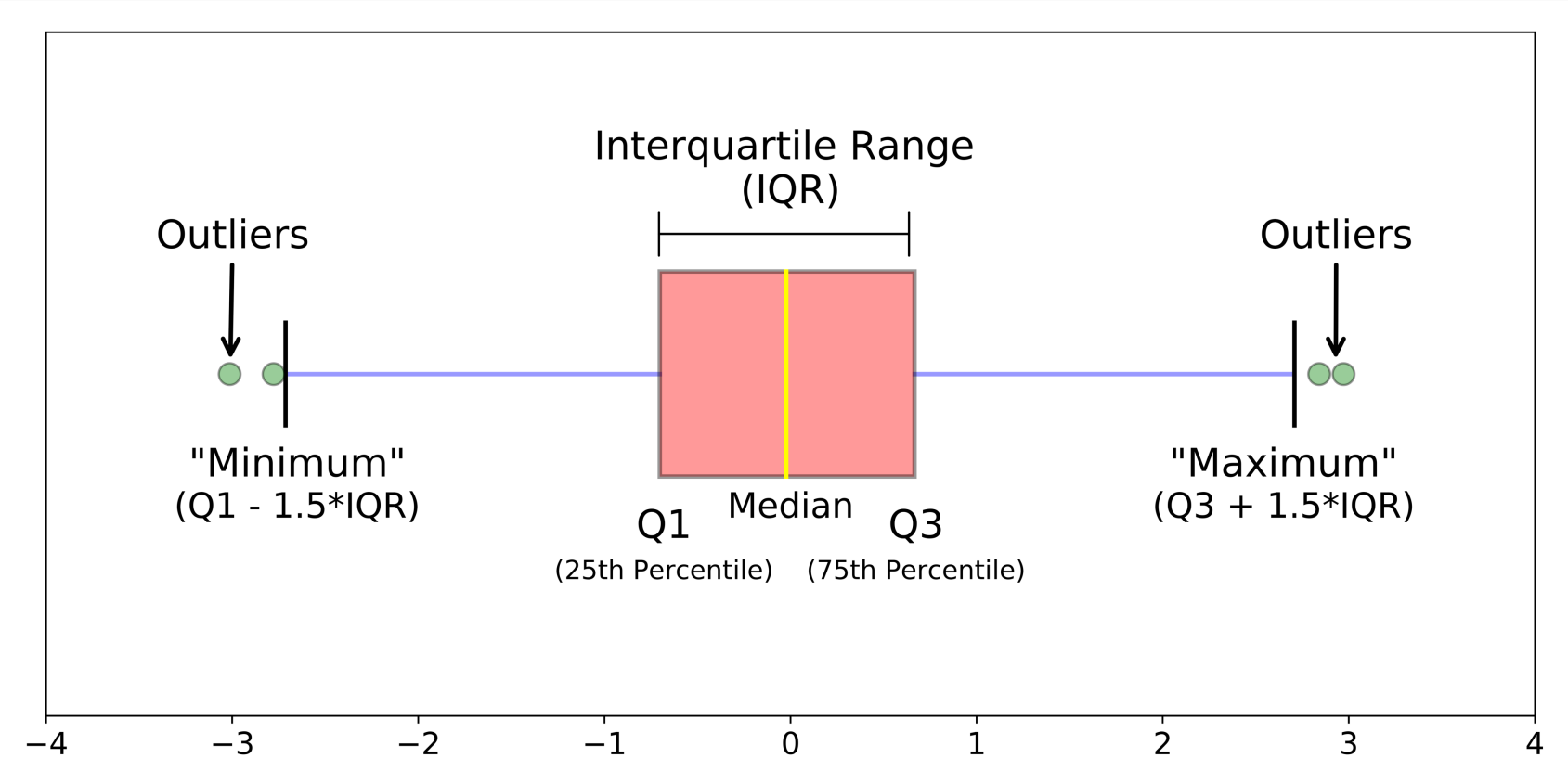
```
## 25% 50% 75%
##  59  60  61
```

```
quantile(sample2, probs = c(0.25, 0.5, 0.75))
```

```
## 25% 50% 75%
##  57  60  62
```

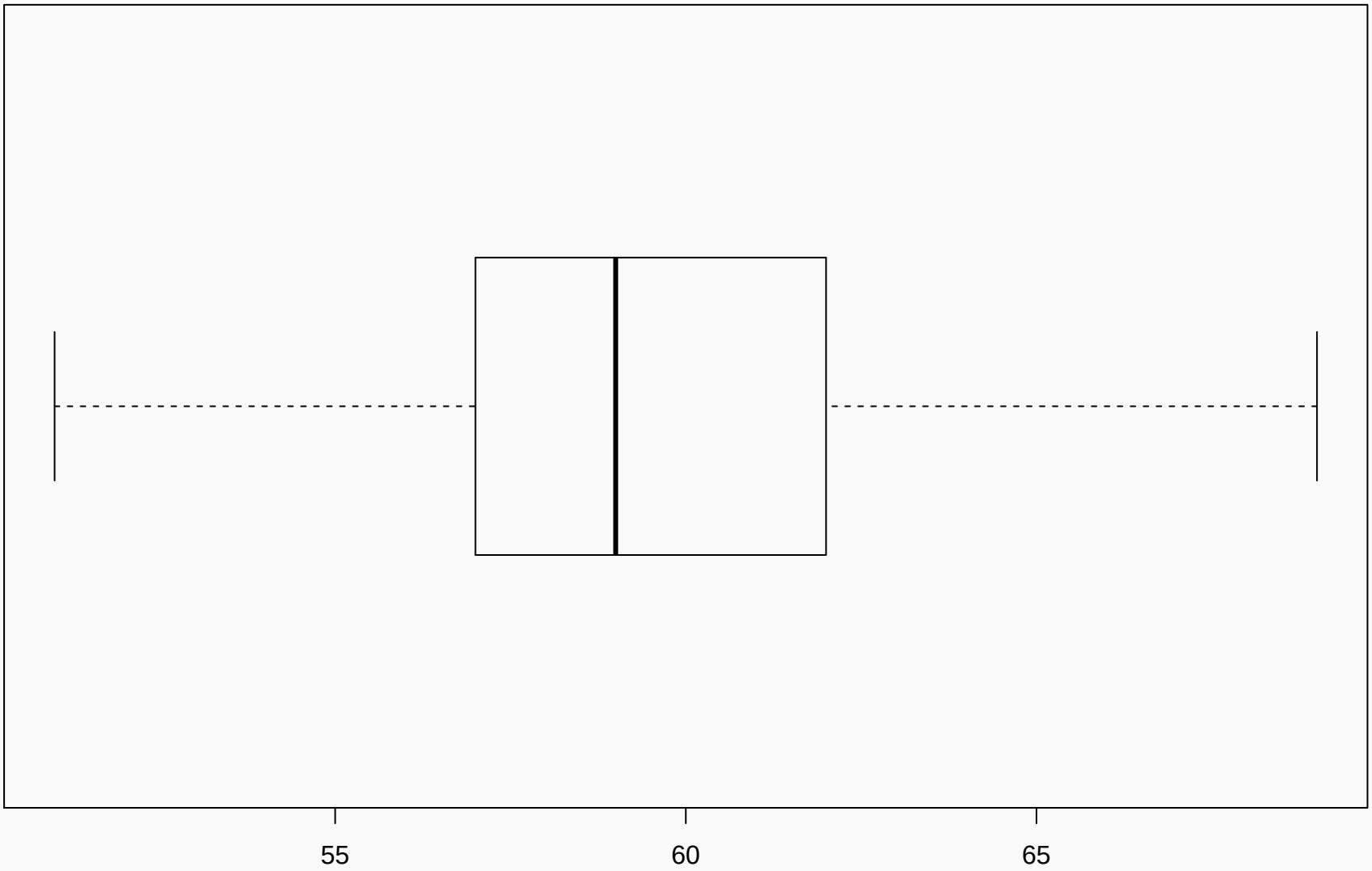
Interquartile range: 4 for sample 1 vs. about 8 for sample 2

Quantifying spread: Boxplots



Quantifying spread: Boxplots

```
boxplot(grades, horizontal = TRUE)
```



Quantifying spread: Boxplots

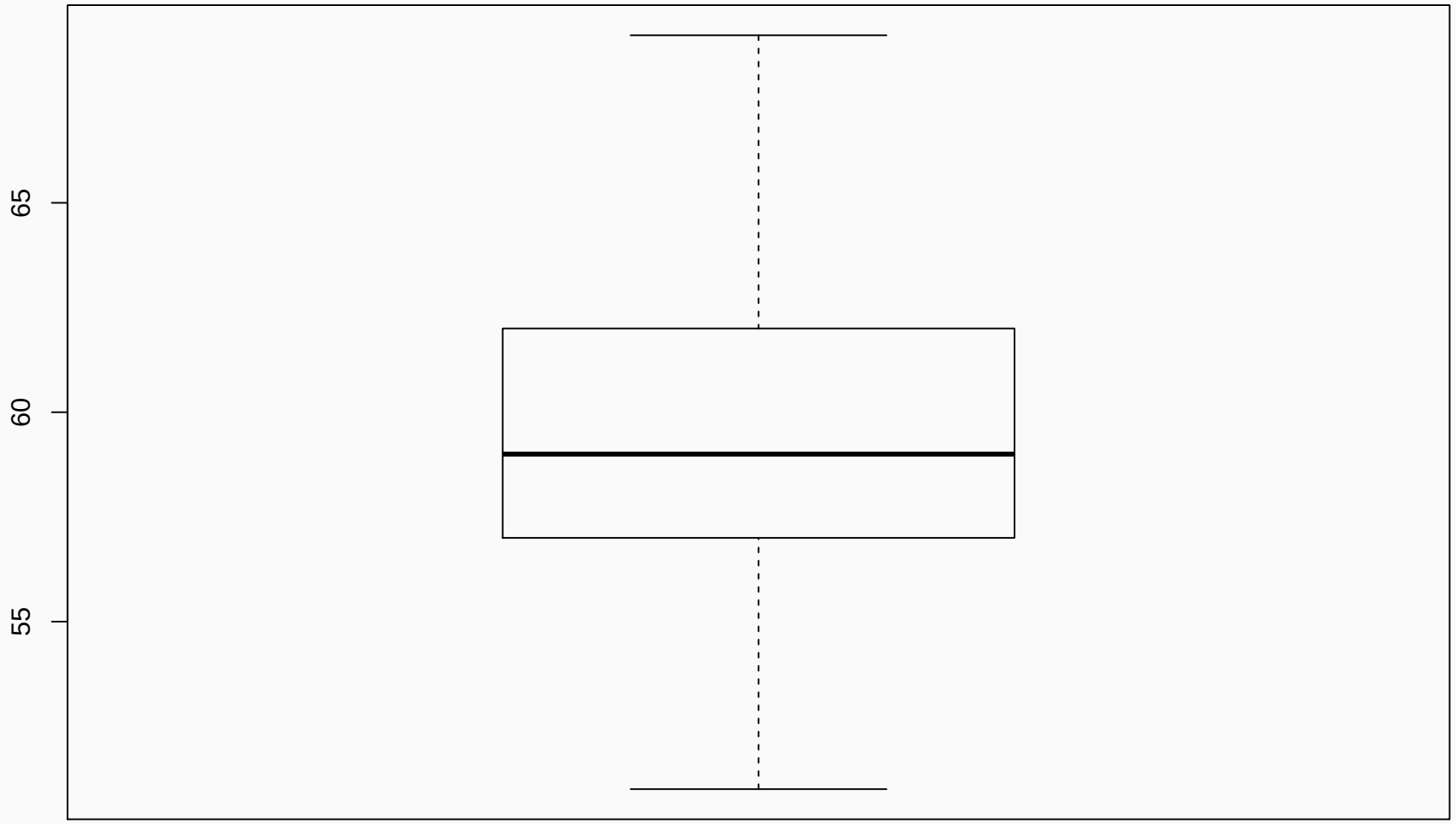
- Line in the middle of the box: Median
- Left and right of the box: Quartiles
- Whiskers:
  - left: lowest case within Q1 - 1.5$\times$ Inter-quartile range
  - right: highest case within Q3 + 1.5$\times$ Inter-quartile range

Quantifying spread: Boxplots

NB: boxplots can be horizontal or vertical, it's a matter of taste!

Same plot:

```
boxplot(grades)
```



Quantifying spread: The standard deviation (IMPORTANT!)

Consider two samples, each with mean 20:

- sample 1: $17, 20, 20, 21, 22$
- sample 2: $7, 14, 18, 26, 35$

Quantifying spread: The standard deviation

Let's look at each sample's deviation from the mean:

- sample 1's deviations from 20: $-3, 0, 0, 1, 2$
- sample 2's deviations from 20: $-13, -6, -2, 6, 15$

So sample 1 and 2's average deviation from the mean is . . . 0. T

What we'd like is an average standard deviation, but the negative deviations will cancel the positive ones so we get an average deviation from the mean of 0. Not useful.

Quantifying spread: The standard deviation

To avoid the positive and negative deviations cancelling out, we will take the **squared** deviations:

- sample 1's squared deviations from 20: $9, 0, 0, 1, 4$
- sample 2's squared deviations from 20: $169, 36, 4, 36, 225$

Quantifying spread: The standard deviation

The mean of these squared deviations is called the *variance* and is written as σ^2. Formally, it is calculated as:

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

A small rabbit hole (well, two). Not that important but if you're curious. . .

Enter the rabbit hole

When we calculate the **sample** variance, as opposed to the **population** variance, we have to use a slightly different formula, namely:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}$$

Rabbit hole 1 (easy)

Why did I write s^2 and not σ^2? Because greek letter (e.g. σ) refer to **population$$ parameters, whereas roman letters (s^2) refer to** sample** parameters.

Rabbit hole 2 (less easy)

Rabbit hole 2 (less easy): Why do we divide by $n - 1$ and not by n, as you'd expect for a mean?

When we calculated the mean, we divided by the number of observations n. That's because we have n different observations, each of which can take on any value.

Rabbit hole 2 (less easy)

But now look at the formula for the sample variance:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

and notice that there is $\bar{x}$ in that formula. But $\bar{x}$ is calculated from the sample. So once we've established $\bar{x}$, we no longer have n observations that can take on any value, but $n-1$ such observations.

Why? Imagine I tell you that I have three observations, and that their mean is 0.

$$1, 2, x$$

Can you guess what x is? Well yes, x MUST be -3, otherwise the mean would not be 0. In other words, x is not "free to vary". In statistical jargon, we say that we have only 2 *degrees of freedom*, not 3.

Rabbit hole 2 (less easy)

So since the mean has been calculated and is given, only $n - 1$ observations are free to move, and that's why the formula divides by $n - 1$ and not n.

Out of the rabbit hole!

Quantifying spread: The standard deviation

Reminder: - sample 1: $17, 20, 20, 21, 22$

Let's now calculate the variance of sample 1 manually:

$$\frac{(17-20)^2+(20-20)^2+(20-20)^2+(21-20)^2+(22-20)^2}{n-1}$$

$$=\frac{9+0+0+1+4}{4}=3.5 \quad (1)$$

In R, you obtain it very simply as:

```
var(sample1)
```

```
## [1] 3.5
```

```
var(sample2)
```

```
## [1] 16.10108
```

Quantifying spread: The standard deviation

The variance has units that are not particularly intuitive, so we often take the square root of the variance and call it the *standard deviation*:

$$\text{Standard deviation} = \sqrt{\text{variance}}$$

We denote the standard deviation by σ, so we can also write:

$$\sigma = \sqrt{\sigma^2}$$

(and $s = \sqrt{s^2}$ when we talk about the sample standard deviation).

Summary

Summary

- Measures of central tendency: a "typical" number to summarize a sample or population
  - Mean (μ or $\bar{x}$)
  - Median
  - Mode
- Measures of spread: how different are the observations from one another
  - Range
  - Quartiles
  - Variance (σ^2) and standard deviation (σ)