SOEE1475 Statistics and Data Analysis
Lecture 2: Univariate statistics

Graeme T. Lloyd

Today

- Defining statistical terms
- Introduce ammonite quadrat data set
- First statistical test (the t-test)

Variable variables

Variable (computing)

Data storage; can change over as time as program executes

Variable (statistics)

Any characteristic, number, or quantity that can be measured or counted

Variable types

Univariate

A single type of measurement

Bivariate

Two types of measurement

Multivariate

Three or more types of measurement

Data types

Discrete

Only integer values (e.g., counts)

Continuous

Decimals/fractions possible

Ratio scale

Terminate at zero (e.g., mass, Kelvin)

Interval scale

Do not terminate at zero (e.g., Celsius)

Closed scale

Fixed terminals at both ends (e.g., %ages)

Directional

Modular (e.g., compass bearings)

Ordinal scale

Data only have rank order (e.g., Moh's)

Data types

Discrete

Only integer values (e.g., counts)

Continuous

Decimals/fractions possible

Nominal/Categorical

Names (e.g., Limestone, Basalt)

Example: class heights

Univariate

Continuous

Ratio scale

Populations vs samples

Population

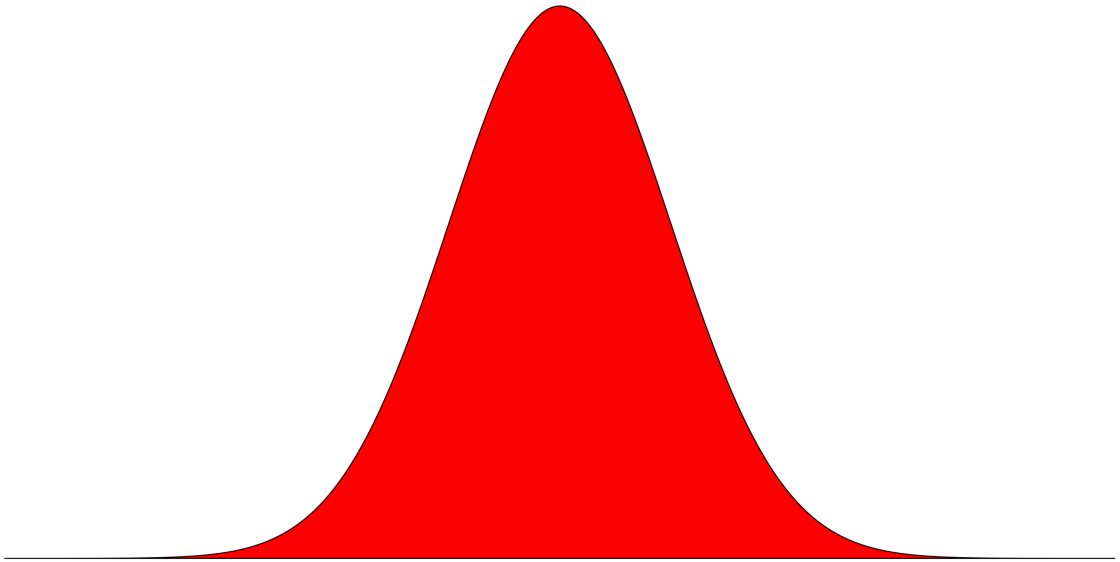
All possible values (may no longer exist, or be accessible)

Sample

A subset of the above and the data we actually have

Populations vs samples

Population
(e.g., all humans ever)

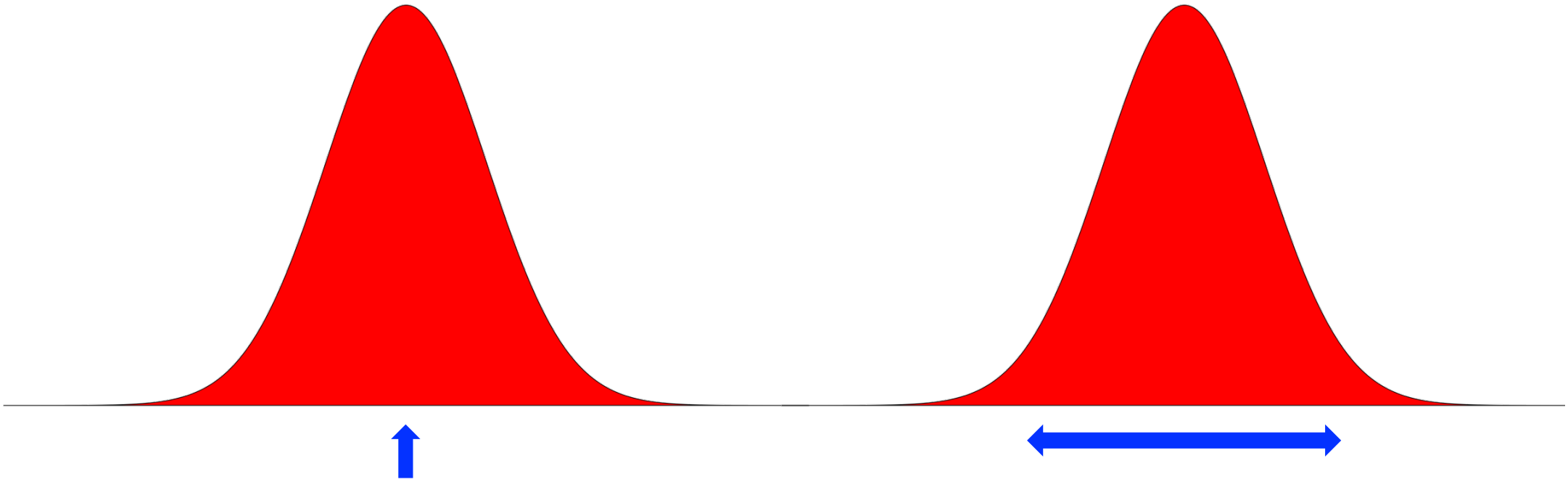


Sample
(e.g., the people in this room)

Summarising univariate data

Location
(captures “centre” of distribution)

Dispersion
(captures “spread” of distribution)



Summarising univariate data

Location

Mean

Sum of values / N values

```
mean(x = ClassHeights)
```

```
68.02273
```

Median

Middle value

```
median(x = ClassHeights)
```

```
68.25
```

Mode

Most frequently occurring value(s)

```
# See below!
```

```
69
```

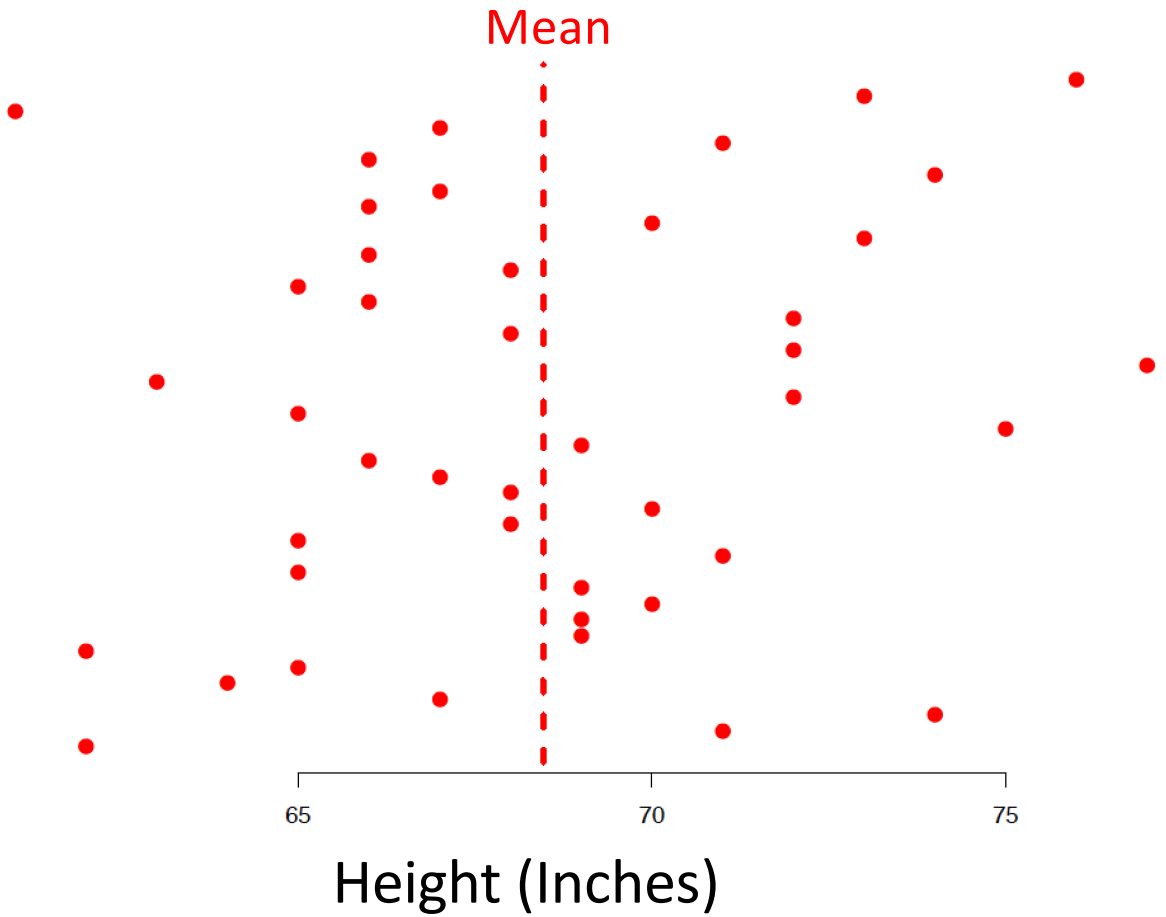
```
rle(sort(ClassHeights))$values[which(rle(sort(ClassHeights))$lengths == max(rle(sort(ClassHeights))$lengths))]
```

Summarising univariate data

Dispersion

Variance

Mean squared distance from mean

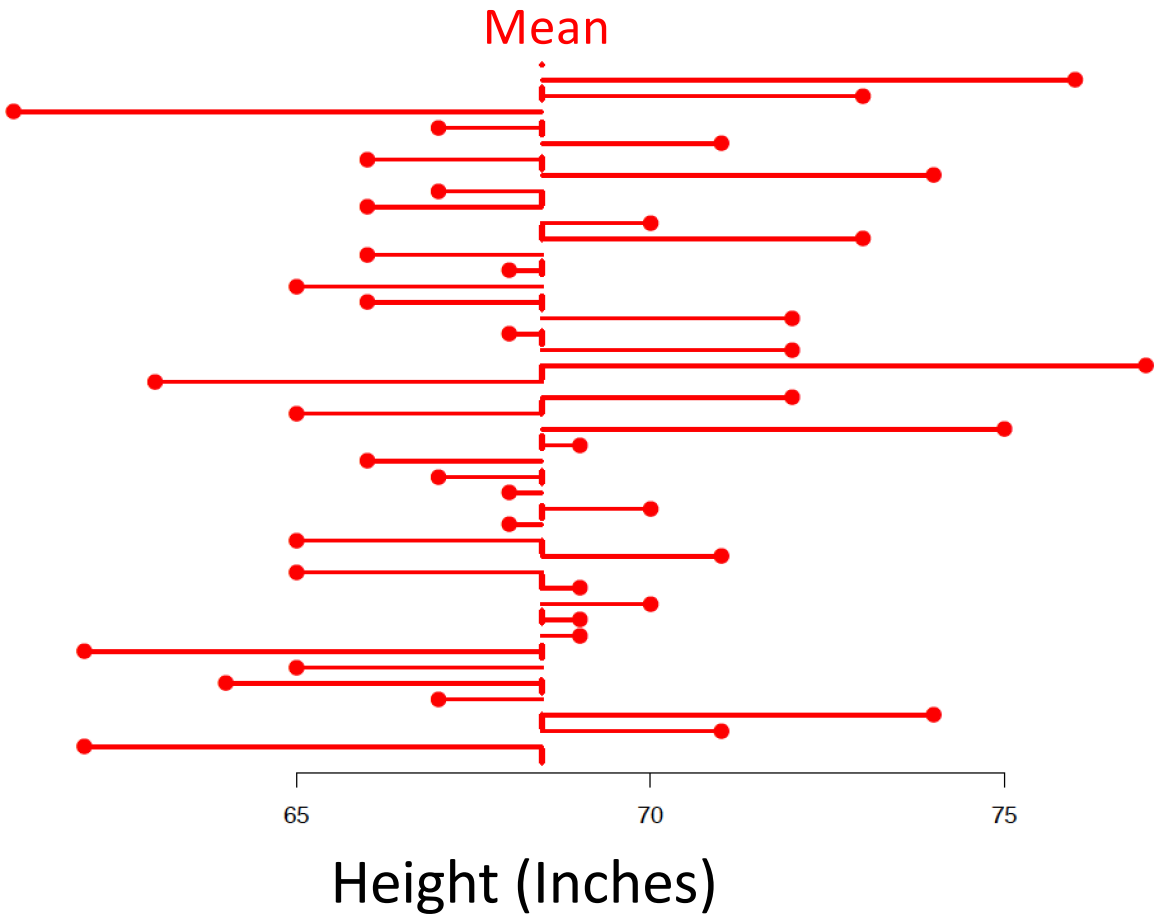


Summarising univariate data

Dispersion

Variance

Mean squared distance from mean

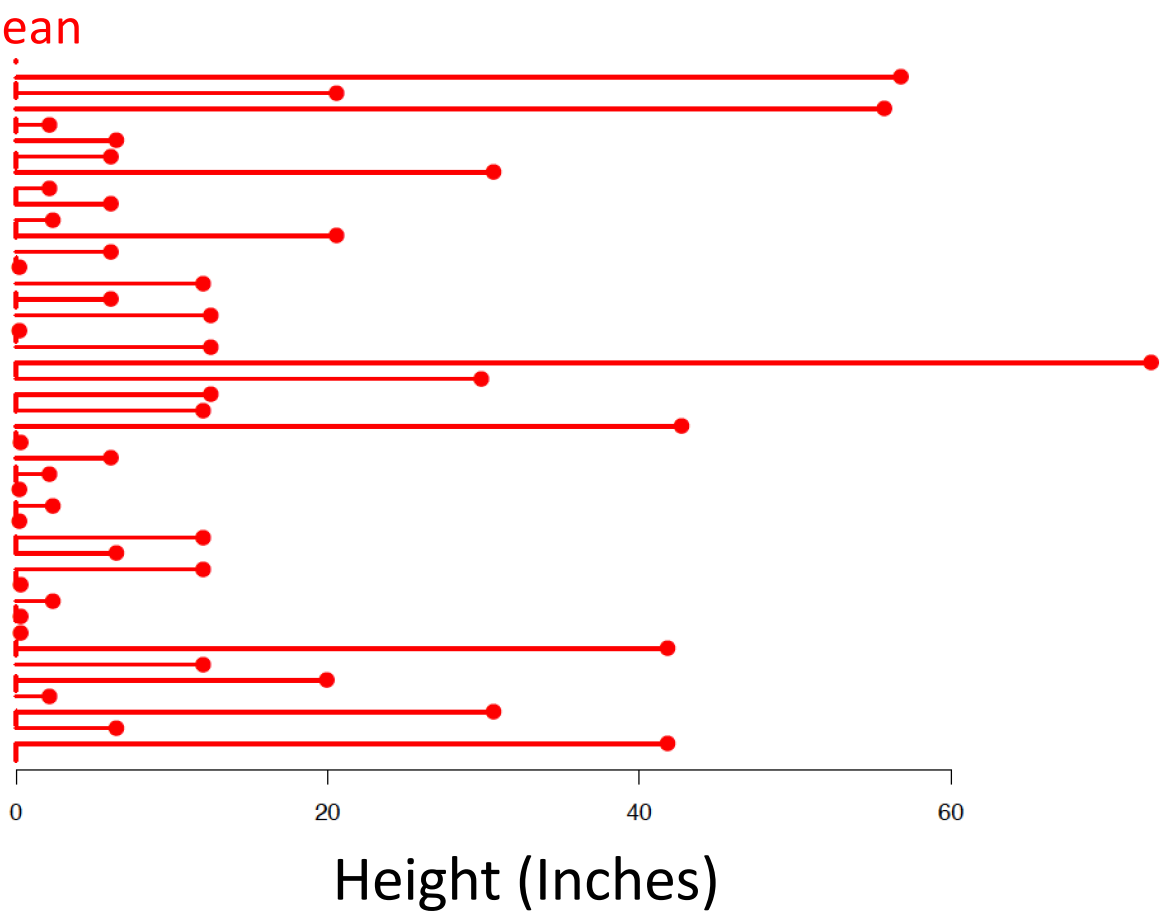


Summarising univariate data

Dispersion

Variance

Mean squared distance from mean



Summarising univariate data

Dispersion

Variance

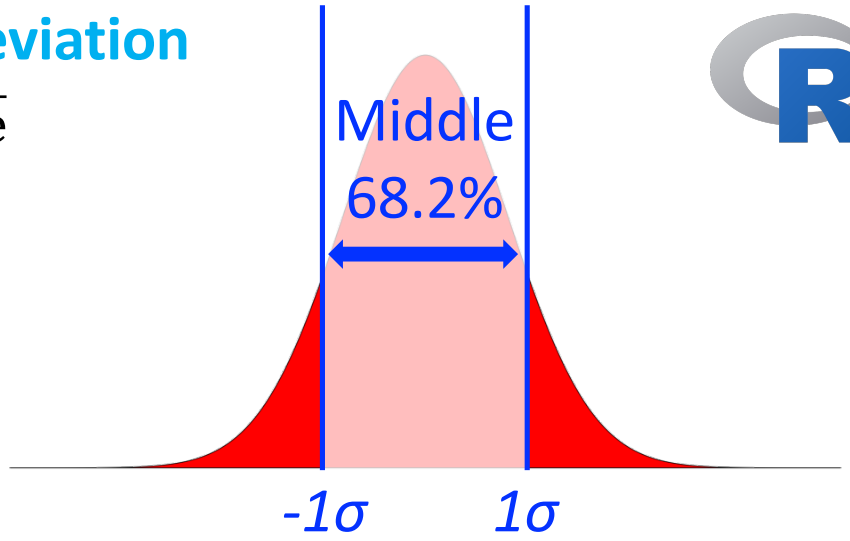
Mean squared distance from mean

```
var(x = ClassHeights)
```

```
16.96374
```

Standard deviation

$\sqrt{\text{Variance}}$

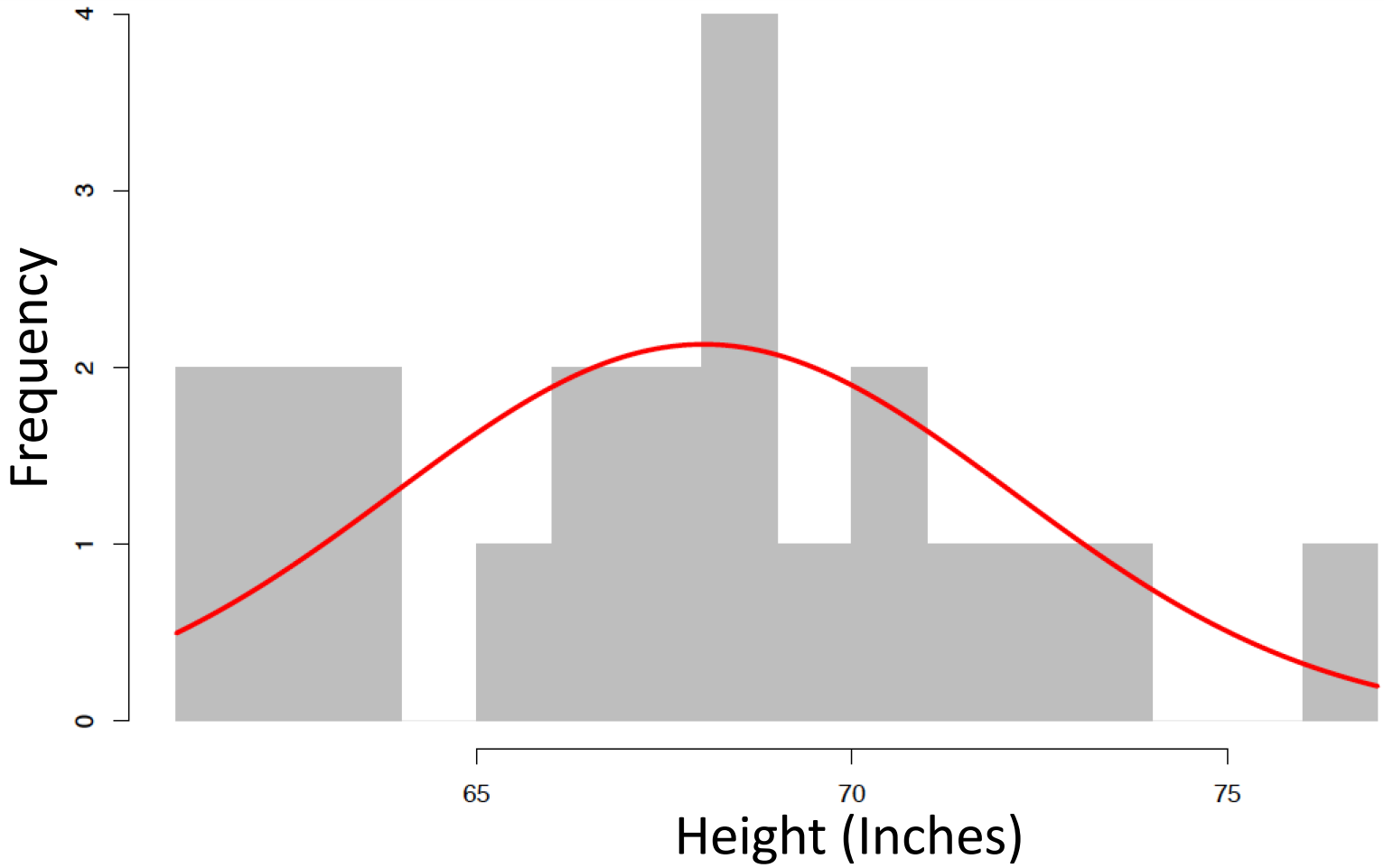



```
sd(x = ClassHeights)
```

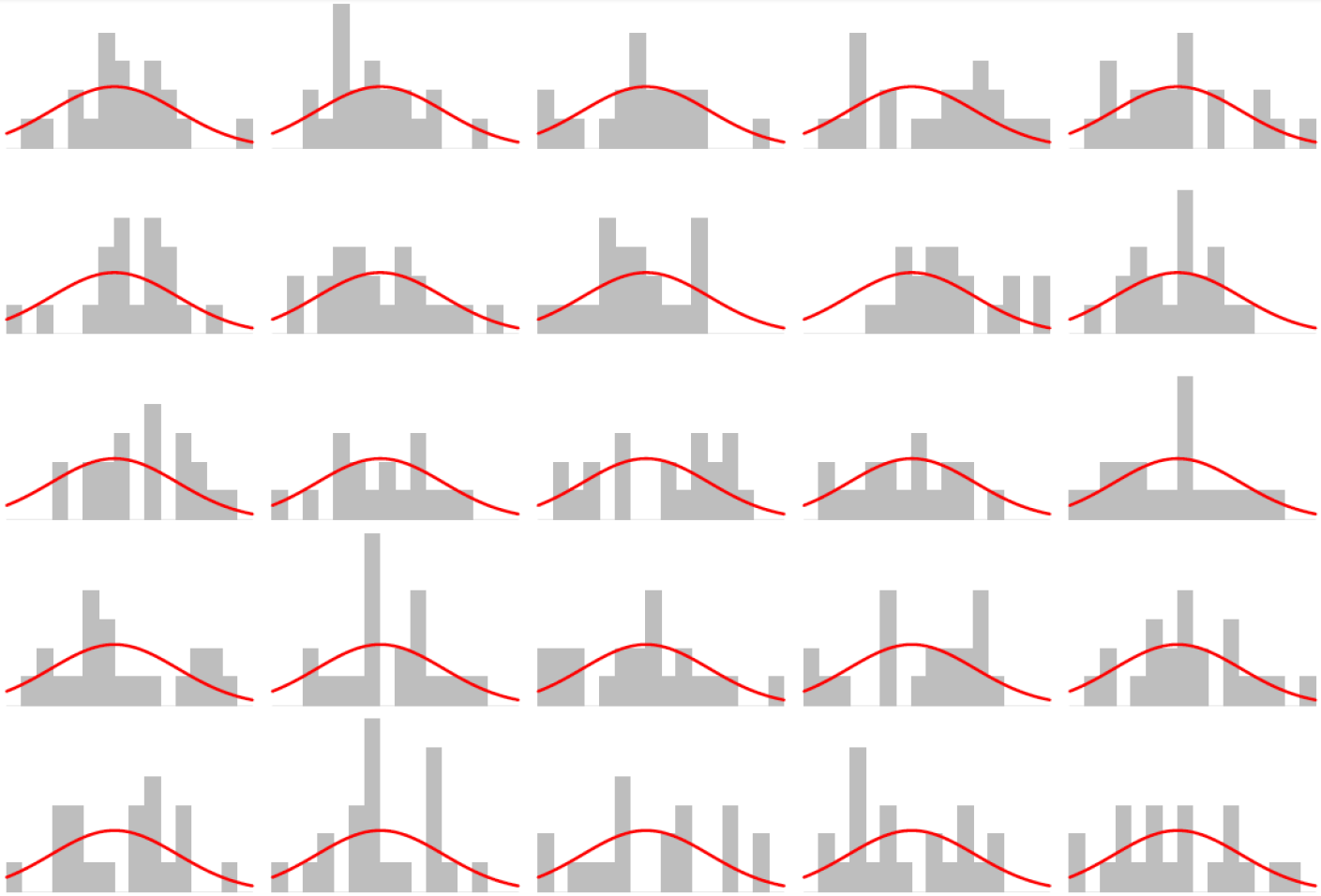
```
4.118707
```

```
68.02273 ± 4.118707
(Actual: 68.2%)
```

Are class heights normally distributed?



Are class heights normally distributed?



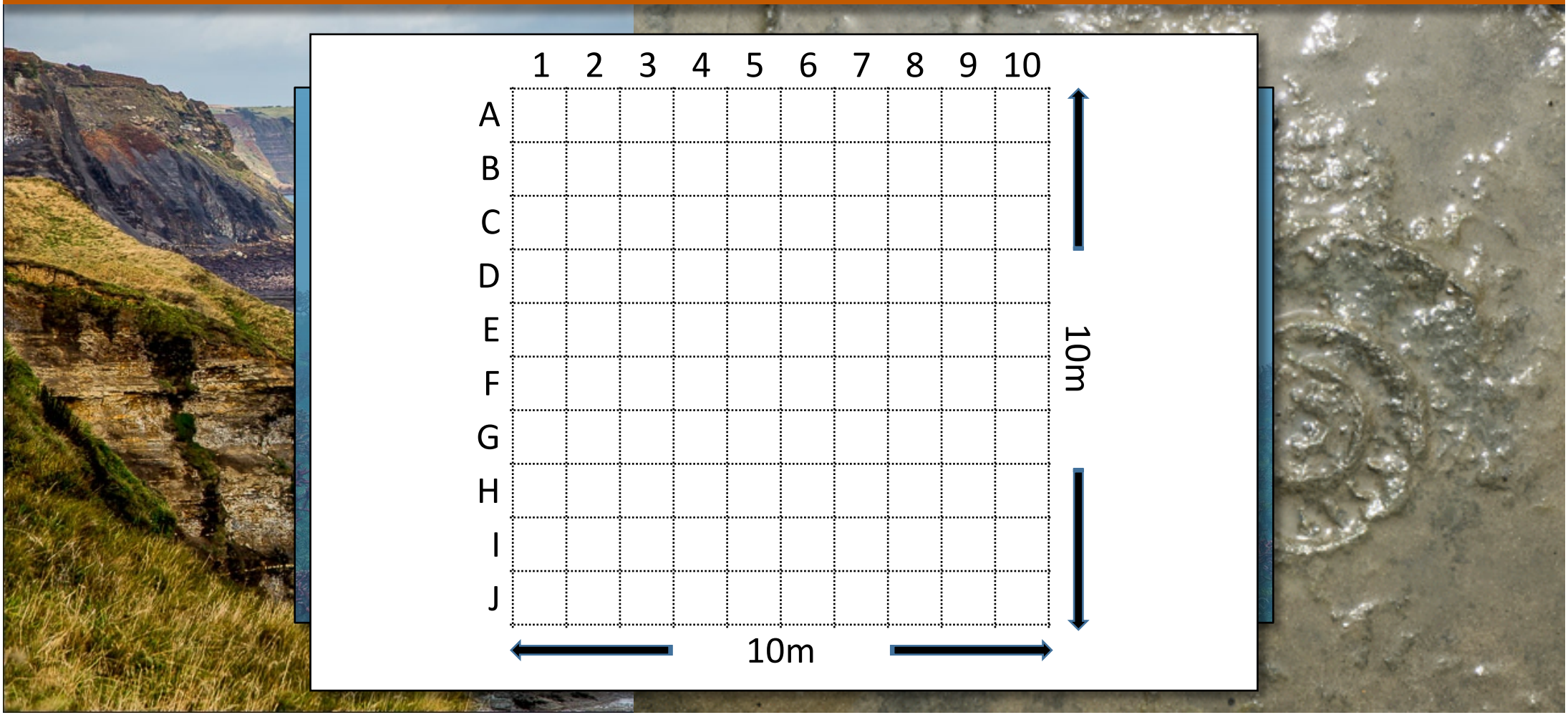
Are class heights normally distributed?

When sample size is small normally distributed data won’t always appear normal

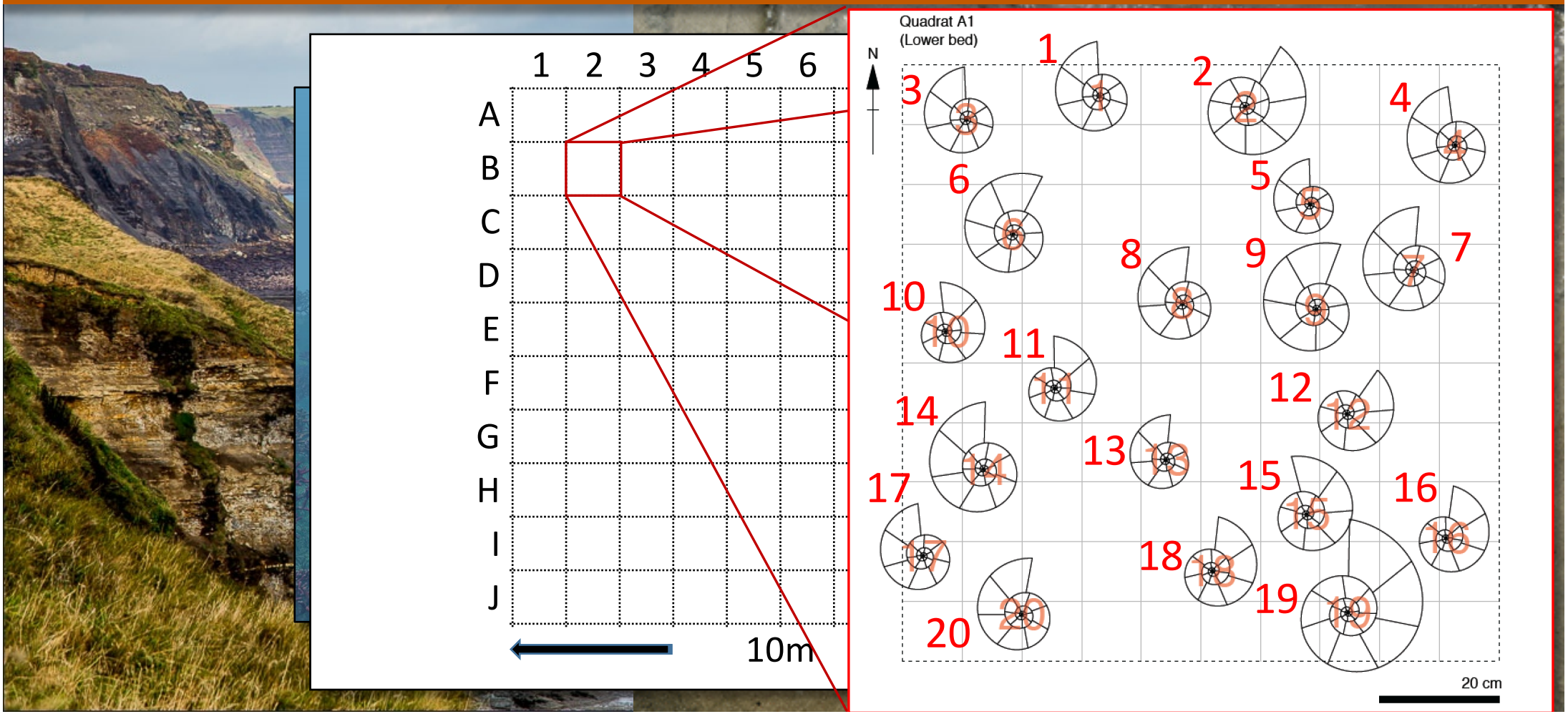
Exercise

Exercise

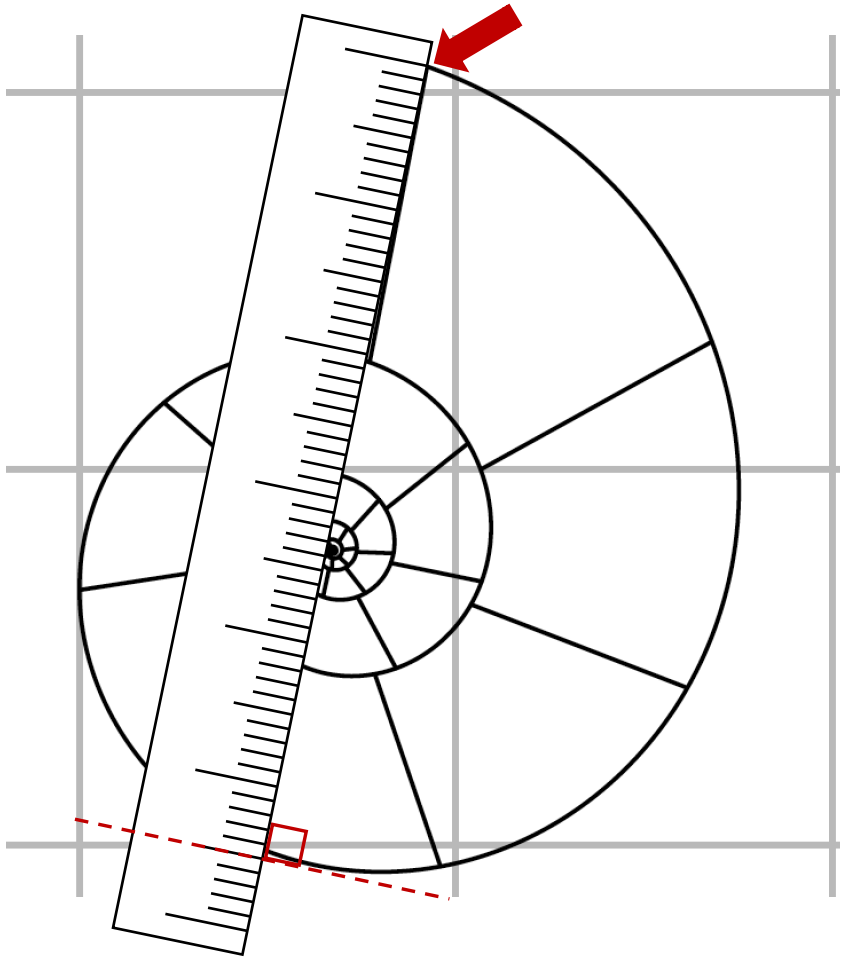
Exercise



Exercise


1 2 3 4 5 6
A
B
C
D
E
F
G
H
I
J
10m
Quadrat A1
(Lower bed)
N
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
20 cm

Exercise

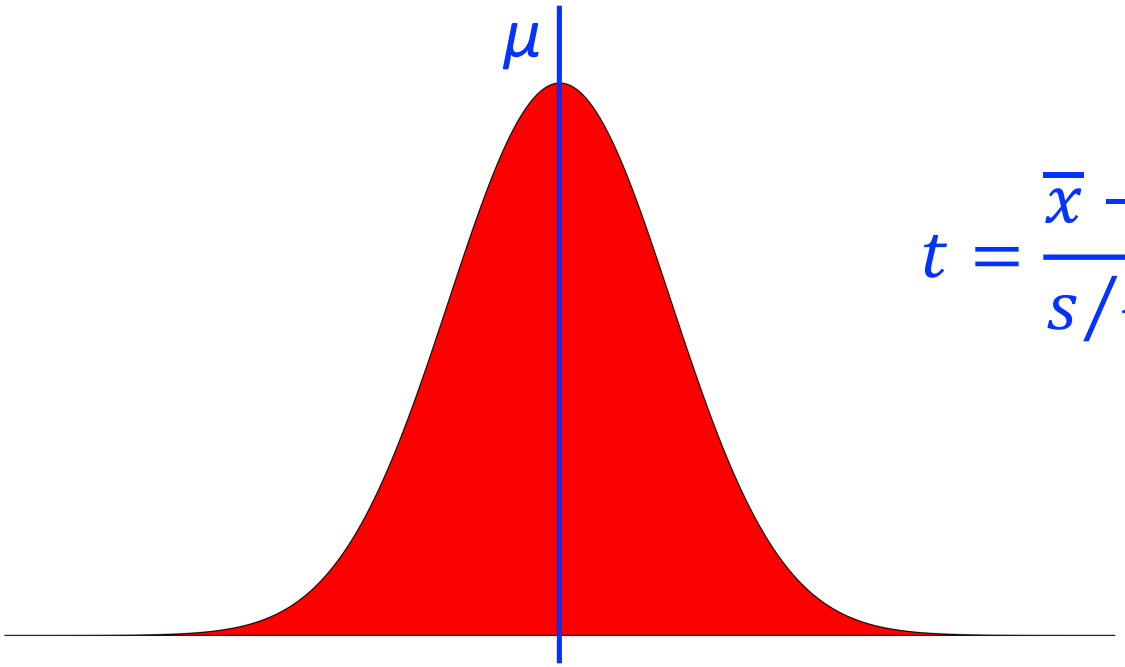


1. Measure ammonites 4 and 17
2. Units of ruler are mm
3. Bin your data (for a histogram):
    1. 51-75 mm
    2. 76-100 mm
    3. 101-125 mm
    4. 126-150 mm
    5. 151-175 mm
    6. 176-200 mm
    7. 201-225 mm
    8. 226-250 mm
    9. 251-275 mm

Estimating the population mean

Population

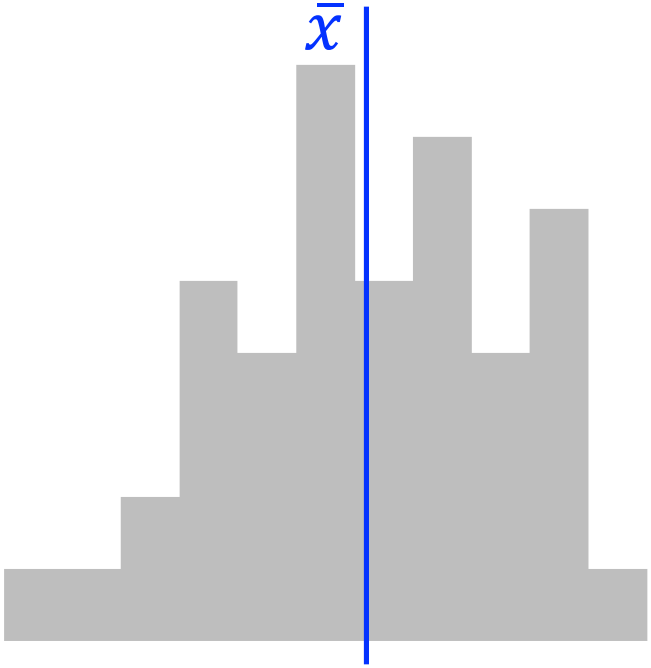
Current Leeds undergraduates



$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

Sample

SOEE1475



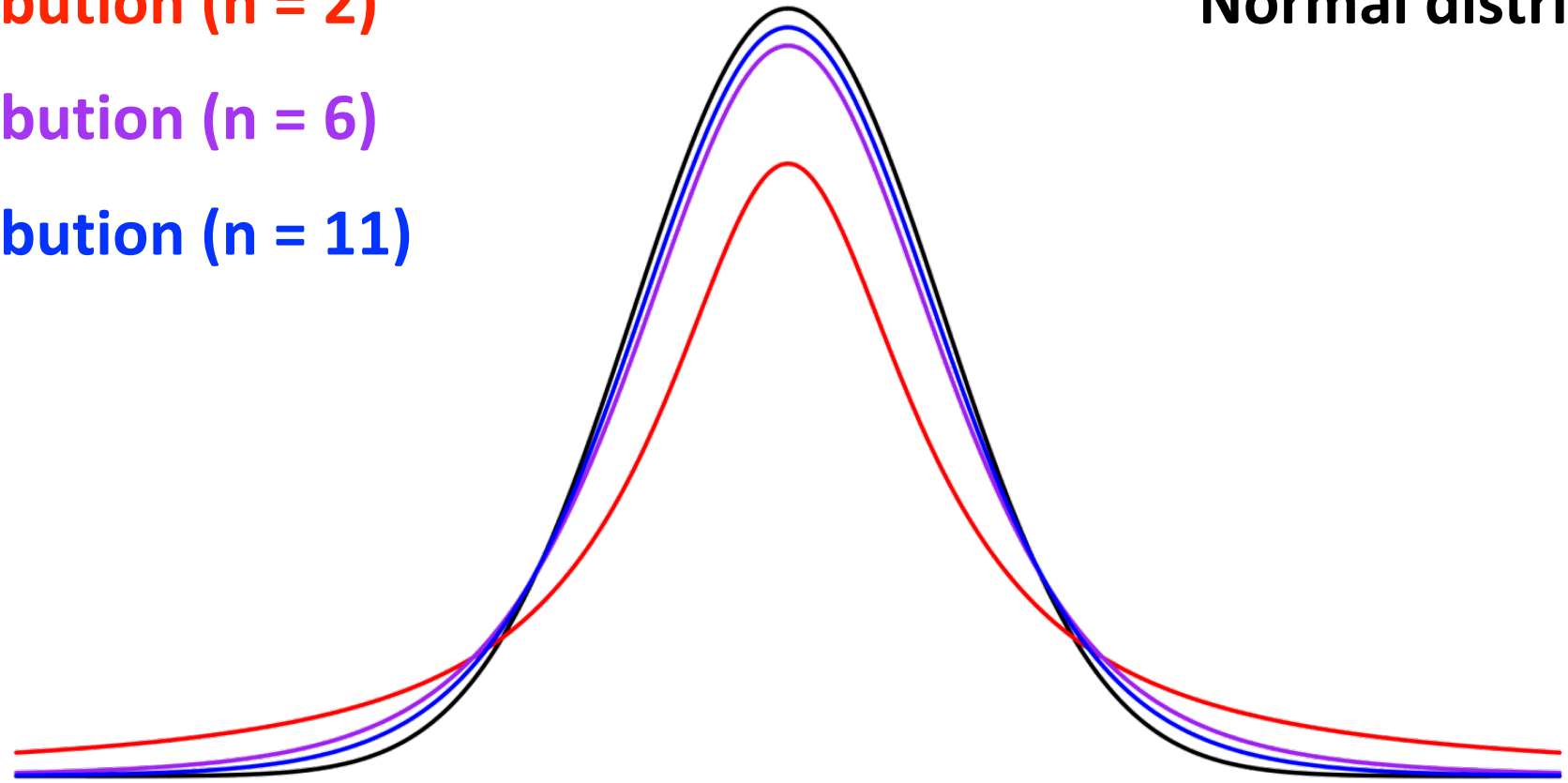
Estimating the population mean

***t*-distribution (n = 2)**

***t*-distribution (n = 6)**

***t*-distribution (n = 11)**

Normal distribution



Estimating the population mean

One sample t-test

Probability a hypothesised population mean (mu) is correct

```
t.test(ClassHeights, mu = 72)$p.value
```

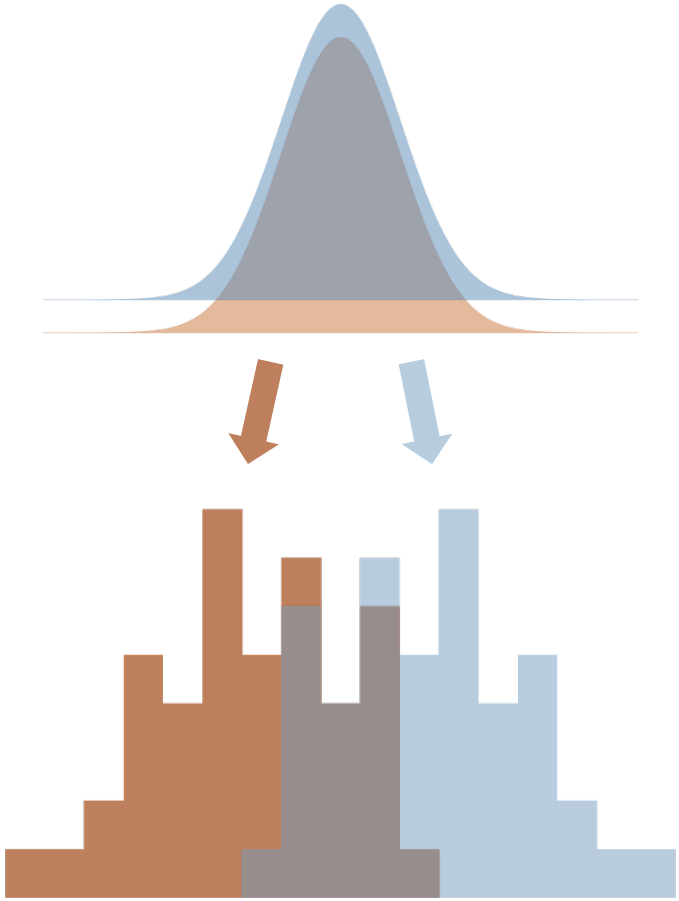
```
0.0001833178
```

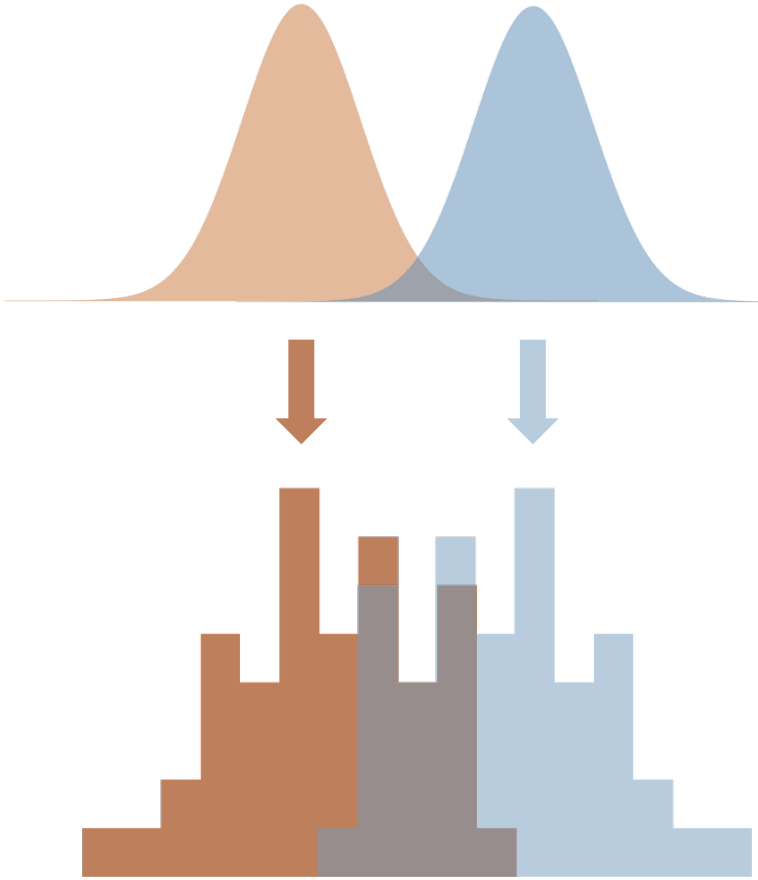
Range of values we can be N% confident population mean falls between

```
t.test(ClassHeights, conf.level = 0.99)$conf.int[1:2]
```

```
65.53648 70.50898
```

Comparing two samples



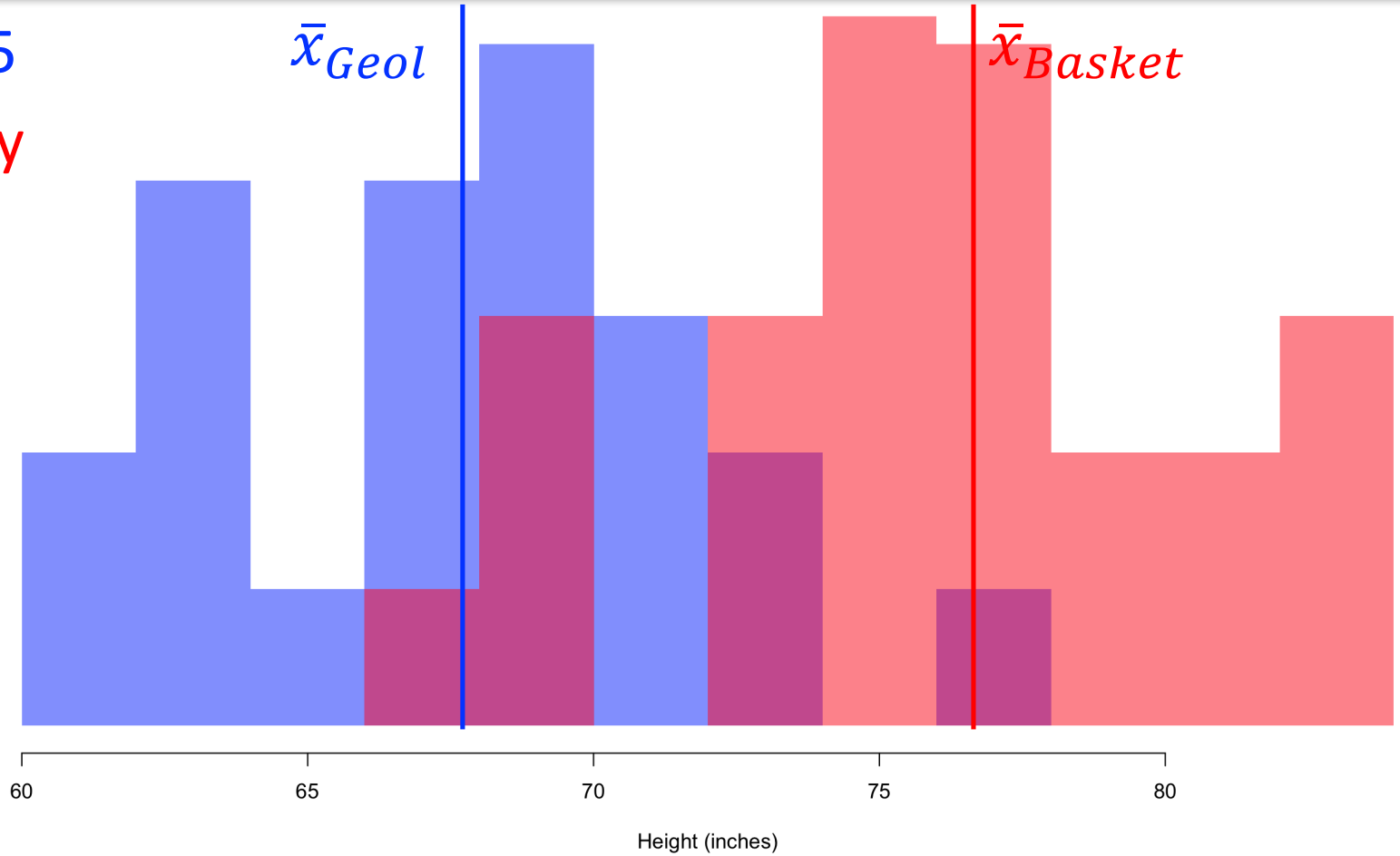


Comparing two samples

Comparing two samples

SOEE1475

Bulls + Sky



Comparing two samples

Two sample t-test

Probability two samples are drawn from populations with same mean (mu)

```
t.test(x = GeolHeights, y = BasketHeights)$p.value
```

```
7.314495e-09
```

Lesson #2: data quality