SNOW
LEOPARD
NETWORK

– Module 12 –

Fundamental Statistical Tools I

03.06.2021

Anne Heloise Theo & Guillaume Demare

WHY STATISTICS

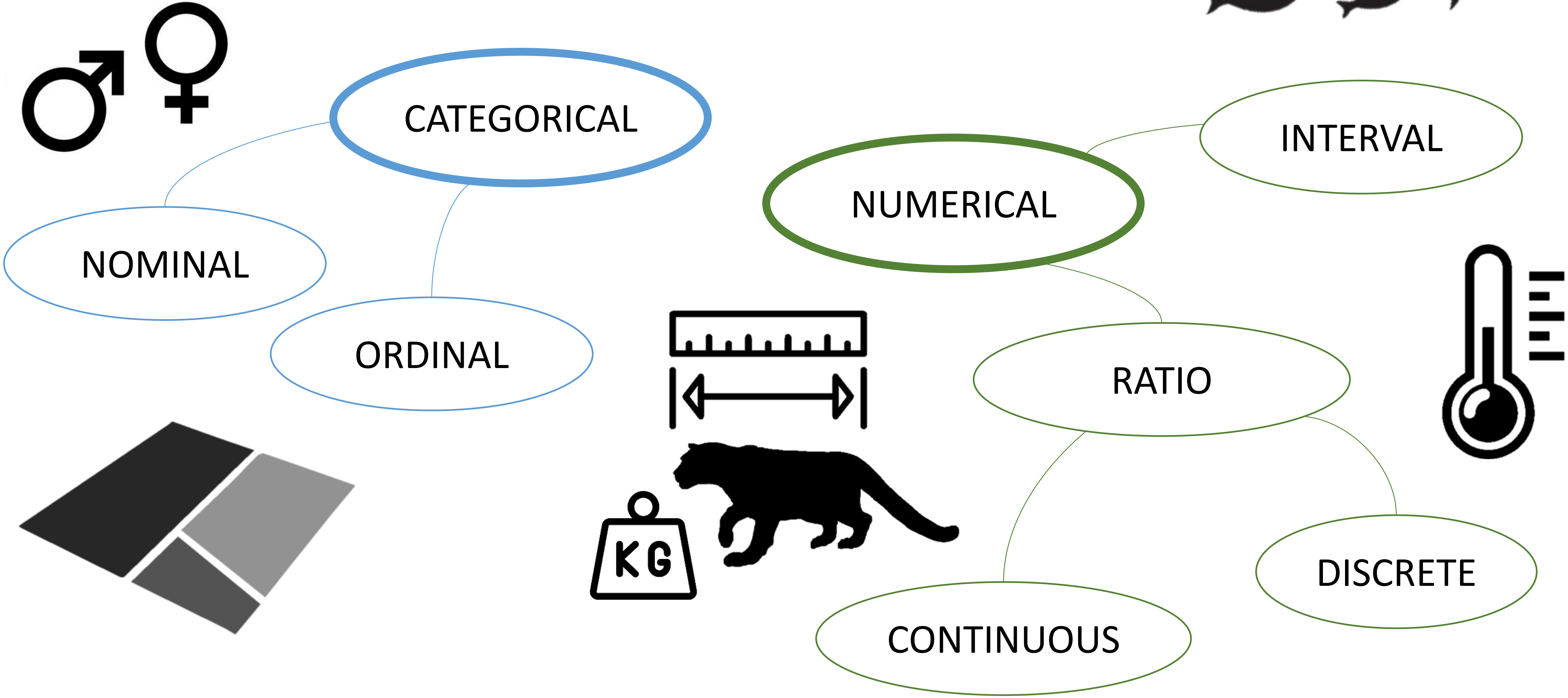
- You collected **data**, NOW WHAT

Statistics is the science
of **learning from data**

- Gain knowledge
- Make inference
- Predict the future

109 106
124 123 109 112 103
105 118 119 116 124 120 128 104
125 130 121 133 130 116 150 119 119 124 119
81 122 121 126 111 94 113 115 110 98 86 101
113 99 83 104 104 108 101 98 99 79 87 88
103 111 105 108 106 118 85 116 90 99 108 96
115 116 92 108 91 101 106 127 100 102 111
123 122 118 107 106 112 113 116 127 109 114
122 102 110 112 106 102 95 97 86 77
81 94 91 78 97 100 94 83 100
90 71 93 100 83
92 88

DATA TYPES



CATEGORICAL

NOMINAL

ORDINAL

NUMERICAL

INTERVAL

RATIO

CONTINUOUS

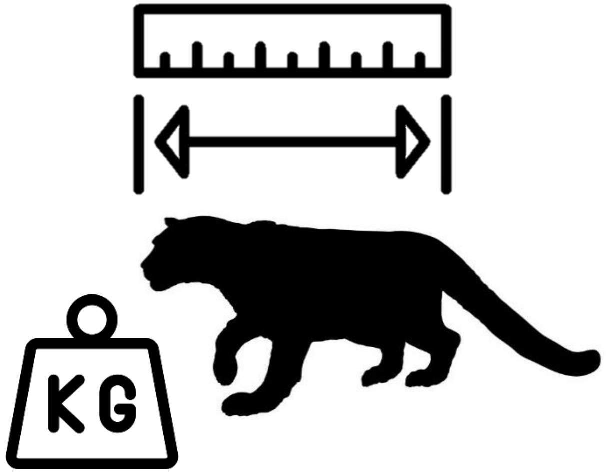
DISCRETE

THE R LANGUAGE

- ✓ Data analysis and statistics
- ✓ Scholars and R&D
- ✓ Difficult at the beginning
- ✓ Beautiful graphs
- ✓ Free open-source
- ✓ Many R packages
  - e.g. Comprehensive R Archive Network (CRAN), R-Forge, GitHub, etc.

- ✓ RStudio IDE (integrated development environment)

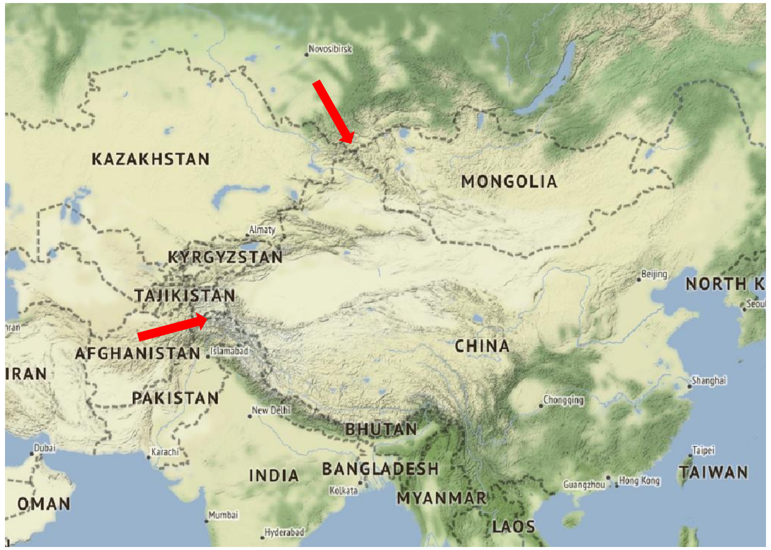
UNCIA DATASET



- 110 snow leopards
- Length (cm) and Weight (kg) of each individual
- Sex (M and F) and Location (Wakhan and Altai)

```
# load data into R from csv file
uncia <- read.csv("uncia.csv", header = TRUE)
str(uncia)

'data.frame':    110 obs. of  5 variables:
 $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Location : chr  "Wakhan" "Wakhan" "Wakhan" "Wakhan" ...
 $ Sex      : chr  "M" "M" "M" "M" ...
 $ Length.cm: int  109 106 124 123 109 112 103 105 118 119 ...
 $ Weight.kg: int  30 29 35 38 32 28 28 32 33 36 ...
```



Map by Stamen Design

DESCRIPTIVE STATISTICS

	Population	Sample
	Parameter	**Statistic**
mean	μ	$\bar{x}$
standard deviation	σ	s
size	N	n

109 106
124 123 109 112 103
105 118 119 116 124 120 128 104
125 130 121 133 130 116 150 119 119 124 119
81 122 121 126 111 94 113 115 110 98 86 101
113 99 83 104 104 108 101 98 99 79 87 88
103 111 105 108 106 118 85 116 90 99 108 96
115 116 92 108 91 101 106 127 100 102 111
123 122 118 107 106 112 113 116 127 109 114
122 102 110 112 106 102 95 97 86 77
81 94 91 78 97 100 94 83 100
90 71 93 100 83
92 88

MEAN

109 106 124 123 109 112 103 105 118 119 116 124 47
51 41 49 51 47 133 130 116 150 119 119 124 119 81
122 121 126 111 94 113 115 110 98 86 101 113 99 83
104 104 108 101 98 99 79 87 88 103 111 105 108 106
118 85 116 90 99 108 96 115 116 92 108 91 101 106
127 100 102 111 123 122 118 107 106 112 113 116
127 109 114 122 102 110 112 106 102 95 97 86 77 81
94 91 78 97 100 94 83 100 90 71 93 100 83 92 88

(sum of all observations)

$$\bar{x} = \frac{\sum x}{n}$$

(number of observations)

102.1

```
# calculate mean length
# note: these all give the same answer
sum(uncia$Length.cm)/nrow(uncia)
sum(uncia$Length.cm)/length(uncia$Length.cm)
mean(uncia$Length.cm)
```

MEDIAN

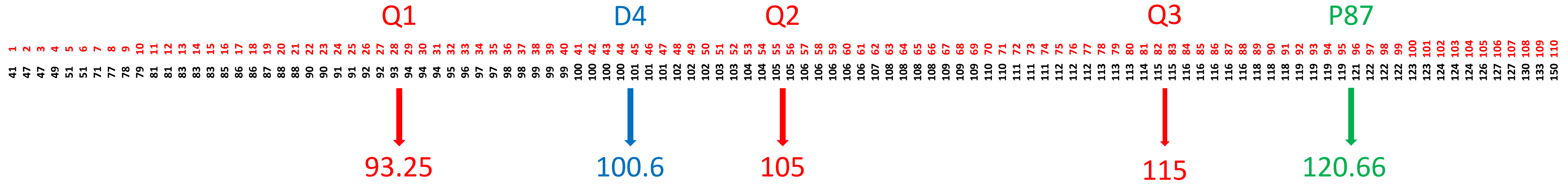
109 106 124 123 109 112 103 105 118 119 116 124 47
51 41 49 51 47 133 130 116 150 119 119 124 119 81
122 121 126 111 94 113 115 110 98 86 101 113 99 83
104 104 108 101 98 99 79 87 88 103 111 105 108 106
118 85 116 90 99 108 96 115 116 92 108 91 101 106
127 100 102 111 123 122 118 107 106 112 113 116
127 109 114 122 102 110 112 106 102 95 97 86 77 81
94 91 78 97 100 94 83 100 90 71 93 100 83 92 88

41 47 47 49 51 51 71 77 78 79 81 81 83 83 83 85 86 86
87 88 88 90 90 91 91 92 92 93 94 94 94 95 96 97 97 98
98 99 99 99 100 100 100 100 101 101 101 102 102 102
103 103 104 104 105 105 106 106 106 106 106 107
108 108 108 108 109 109 109 110 110 111 111 111
112 112 112 113 113 113 114 115 115 116 116 116
116 116 118 118 118 119 119 119 119 121 122 122
122 123 123 124 124 124 126 127 127 130 133 150

```
# calculate median length
length_index <- order(uncia$Length.cm)
length_order <- uncia$Length.cm[length_index]
(length(length_order)+1)/2
median(uncia$Length.cm)
```

QUANTILES

Quartiles | Percentiles | Deciles



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
41	47	47	49	51	51	71	77	78	79	81	81	83	83	83	85	86	86	87	88	88	90

23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44
90	91	91	92	92	93	94	94	94	95	96	97	97	98	98	99	99	99	100	100	100	100

45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66
101	101	101	102	102	102	103	103	104	104	105	105	106	106	106	106	106	107	108	108	108	108

67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88
109	109	109	110	110	111	111	111	112	112	112	113	113	113	114	115	115	116	116	116	116	116

89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110
118	118	118	119	119	119	119	121	122	122	122	123	123	124	124	124	126	127	127	130	133	150

Q1 93.25

D4 100.6

Q2 105

Q3 115

P87 120.66

```
# median (aka Q2)
quantile(uncia$Length.cm, 0.5)
# quartiles Q1 and Q3
c(quantile(uncia$Length.cm, 0.25), quantile(uncia$Length.cm, 0.75))
# fourth decile (i.e. at 40%)
quantile(uncia$Length.cm, 0.4)
# percentile 87 (i.e. at 87%)
quantile(uncia$Length.cm, 0.87)
# quantile summary
quantile(uncia$Length.cm)
```

VARIATION

Range

$$\mathrm{range}(x) = \max(x) - \min(x) = 109$$

Variance

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1} = 364.1$$

Standard Deviation

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} = 19.1$$

- The range is the difference between the highest and smallest value, BUT not always informative

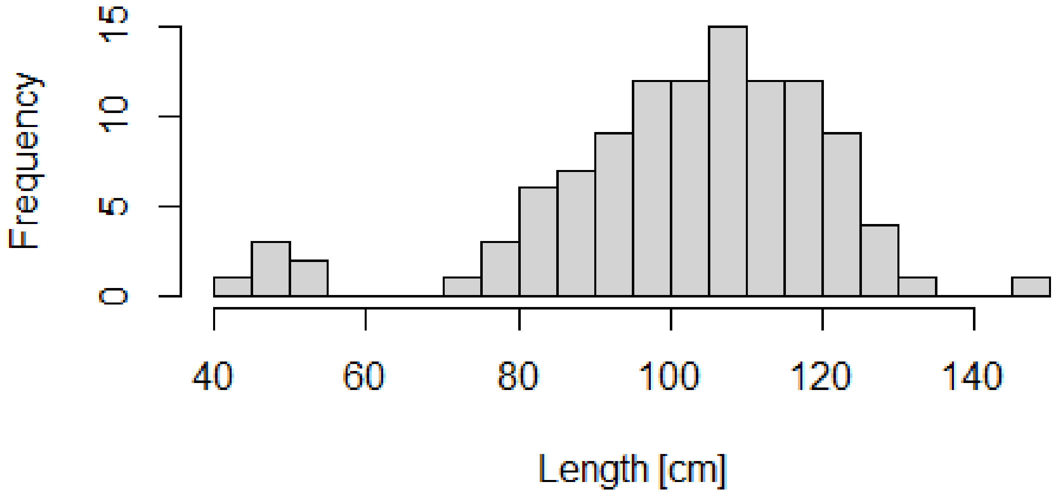
- The variance measures the average of the squared differences from the mean

- The standard deviation is a measure of how widely scattered measurements are from the mean (calculated as the square root of the variance)

```
# range
max(uncia$Length.cm) - min(uncia$Length.cm)
# variance
sum((uncia$Length.cm - mean(uncia$Length.cm)^2) / (nrow(uncia) - 1))
var(uncia$Length.cm)
# standard deviation
sqrt(var(uncia$Length.cm))
sd(uncia$Length.cm)
```

MODE

109 106 124 123 109 112 103 105 118 119 116 124 47
51 41 49 51 47 133 130 116 150 119 119 124 119 81
122 121 126 111 94 113 115 110 98 86 101 113 99 83
104 104 108 101 98 99 79 87 88 103 111 105 108 106
118 85 116 90 99 108 96 115 116 92 108 91 101 106
127 100 102 111 123 122 118 107 106 112 113 116
127 109 114 122 102 110 112 106 102 95 97 86 77 81
94 91 78 97 100 94 83 100 90 71 93 100 83 92 88



```
# frequency table
table(uncia$Length.cm)
table(uncia$Length.cm)[table(uncia$Length.cm) == max(table(uncia$Length.cm))]
# histogram
hist(uncia$Length.cm, breaks = 109, xlab = “Length [cm]", ylab = "Frequency", main = "")
hist(uncia$Length.cm, breaks = 20, xlab = "Length [cm]", ylab = "Frequency", main = "")
```