SNOW LEOPARD NETWORK

– Module 12 –

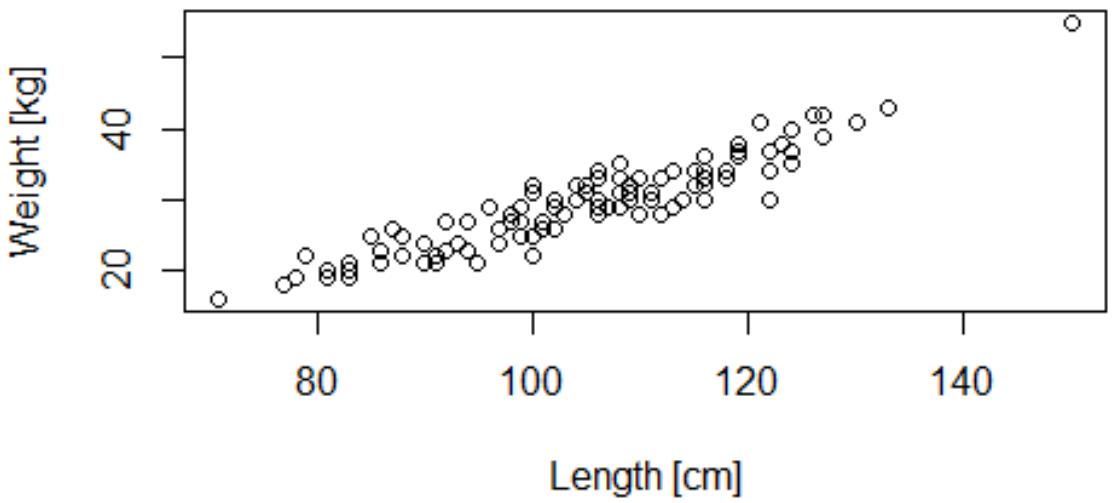
Fundamental Statistical Tools II

10.06.2021

Anne Heloise Theo & Guillaume Demare

DESCRIPTIVE STATISTICS (RECAP)

- **Central tendency** measures
  - Mean
  - Median
  - Mode
- **Dispersion** measures
  - Range and quantiles
  - Variance
  - Standard deviation




```
# mean length
mean(uncia$Length.cm)
[1] 105.2308
# median length
median(uncia$Length.cm)
[1] 106
# mode of length
table(uncia$Length.cm)
hist(uncia$Length.cm, breaks = 109)
hist(uncia$Length.cm, breaks = 20)
# range of length
max(uncia$Length.cm) - min(uncia$Length.cm)
[1] 79
# quantiles of length
quantile(uncia$Length.cm)
    0%    25%    50%    75%   100%
 71.00  95.75 106.00 116.00 150.00
# variance of length
var(uncia$Length.cm)
[1] 202.121
# standard deviation of length
sd(uncia$Length.cm)
[1] 14.21693
# plot length against weight
plot(x = uncia$Length.cm, y = uncia$Weight.kg)
# remove the "fat dwarves" outliers
uncia <- subset(uncia, uncia$Length.cm > 60)
```

DEGREES OF FREEDOM

Standard deviation

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

??

- Standard deviation of sample underestimates population $\boldsymbol{\sigma}$
  → solution: Bessel's correction
- Related to how **degrees of freedom (DF)** are calculated

- DF are the number of independent data points in the calculation of the statistic
- EXAMPLE: If sample mean $\bar{x}$ = 45, only 2 observations (i.e. n – 1) are *free* to take any value, while the 3rd observation is *fixed*

	Observation x	Deviations from mean $x - \bar{x}$
1	47	+2
2	35	-10
3	53	+8
SUMS	135	0

STANDARD ERROR

Standard Deviation

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

Standard Error

$$\mathrm{SE}(\bar{x}) = \frac{s}{\sqrt{n}}$$

- The standard deviation is a measure of how widely scattered measurements are from the mean
  BUT this does not tell us how confident we are with our estimate of the population mean
- The standard error (of the mean) is a measure of how confident we are with our population mean estimate
- The higher the sample size, the lower the standard error

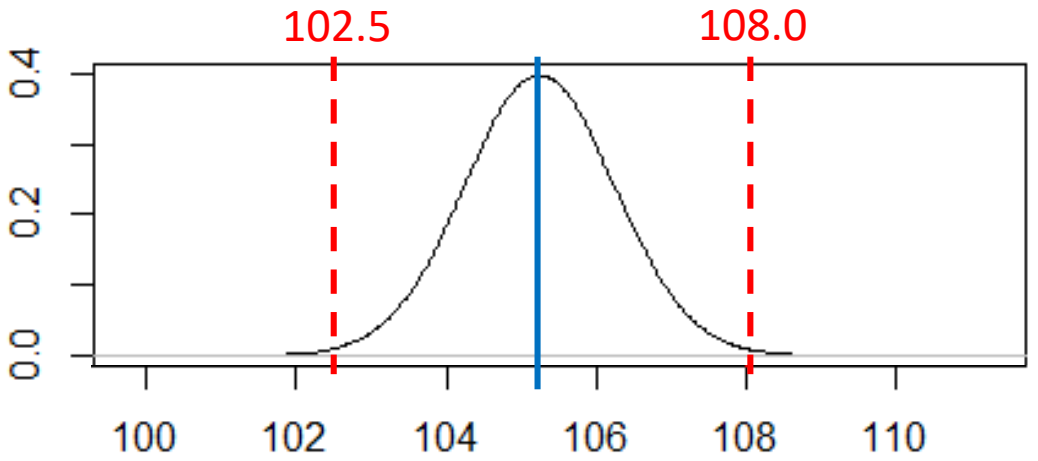
```
# create function to calculate standard error
SE <- function(x) { sd(x)/sqrt(length(x)) }
# take 10 random observations of leopard length and calculate SE
x <- sample(uncia$Length.cm, 10)
SE(x)
# repeat on all observations available (n = 104)
SE(uncia$Length.cm)
```

CONFIDENCE INTERVALS

Mean $\bar{x} = \frac{\sum x}{n} = 105.2$

- We can use the **standard error** to generate confidence intervals

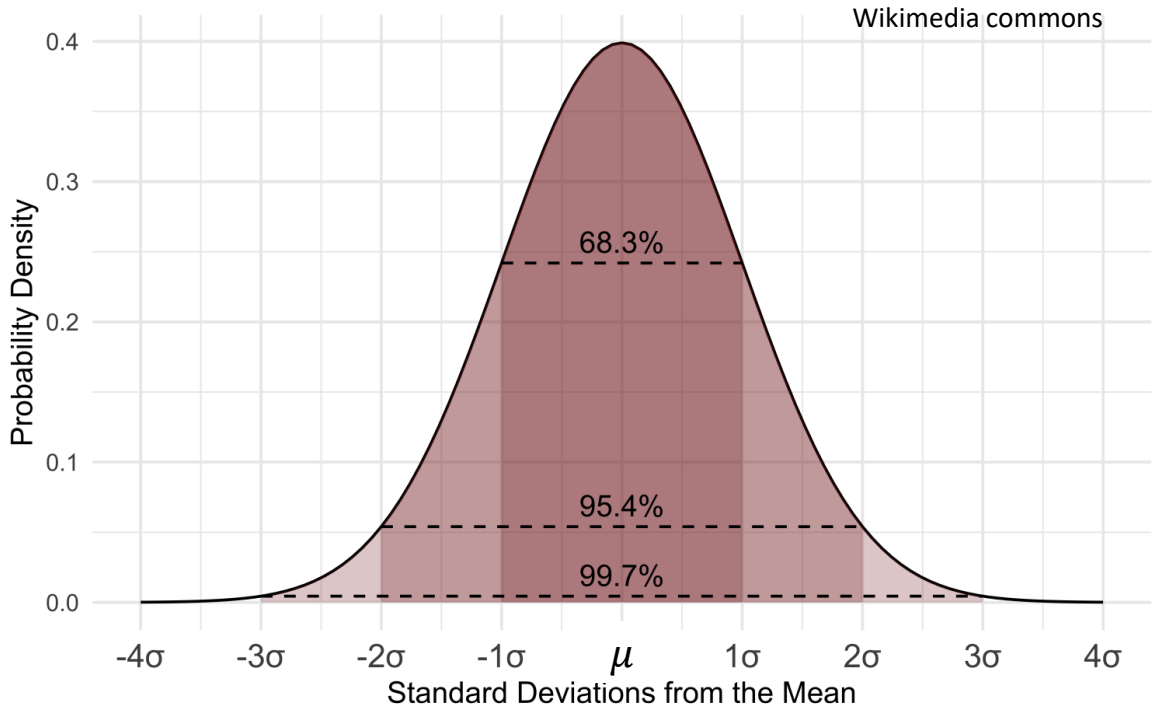
  E.g. 95% CI




```
# calculate 95% confidence intervals around sample mean
xbar <- mean(uncia$Length.cm)
se <- SE(uncia$Length.cm)
upper = xbar + se * qnorm(0.975)
lower = xbar - se * qnorm(0.975)
```

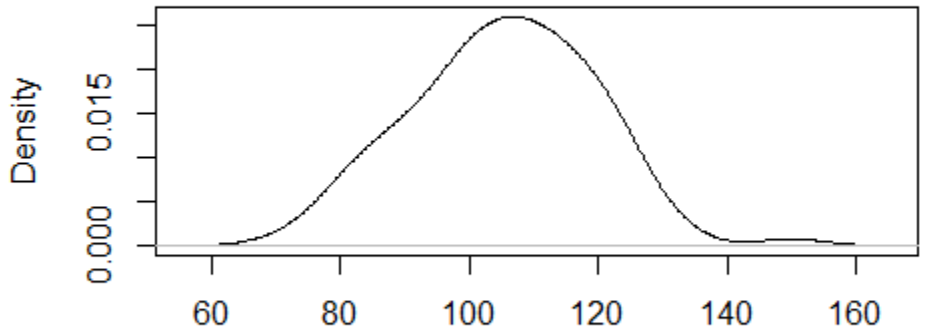
NORMAL DISTRIBUTION

aka Gaussian distribution



- Hypothetical distribution
- Symmetric
- Mean = Median = Mode
- Area under curve (of pdf) = 1

```
# histogram for length (uncia dataset)
hist(uncia$Length.cm)
# corresponding density plot
plot(density(uncia$Length.cm))
```



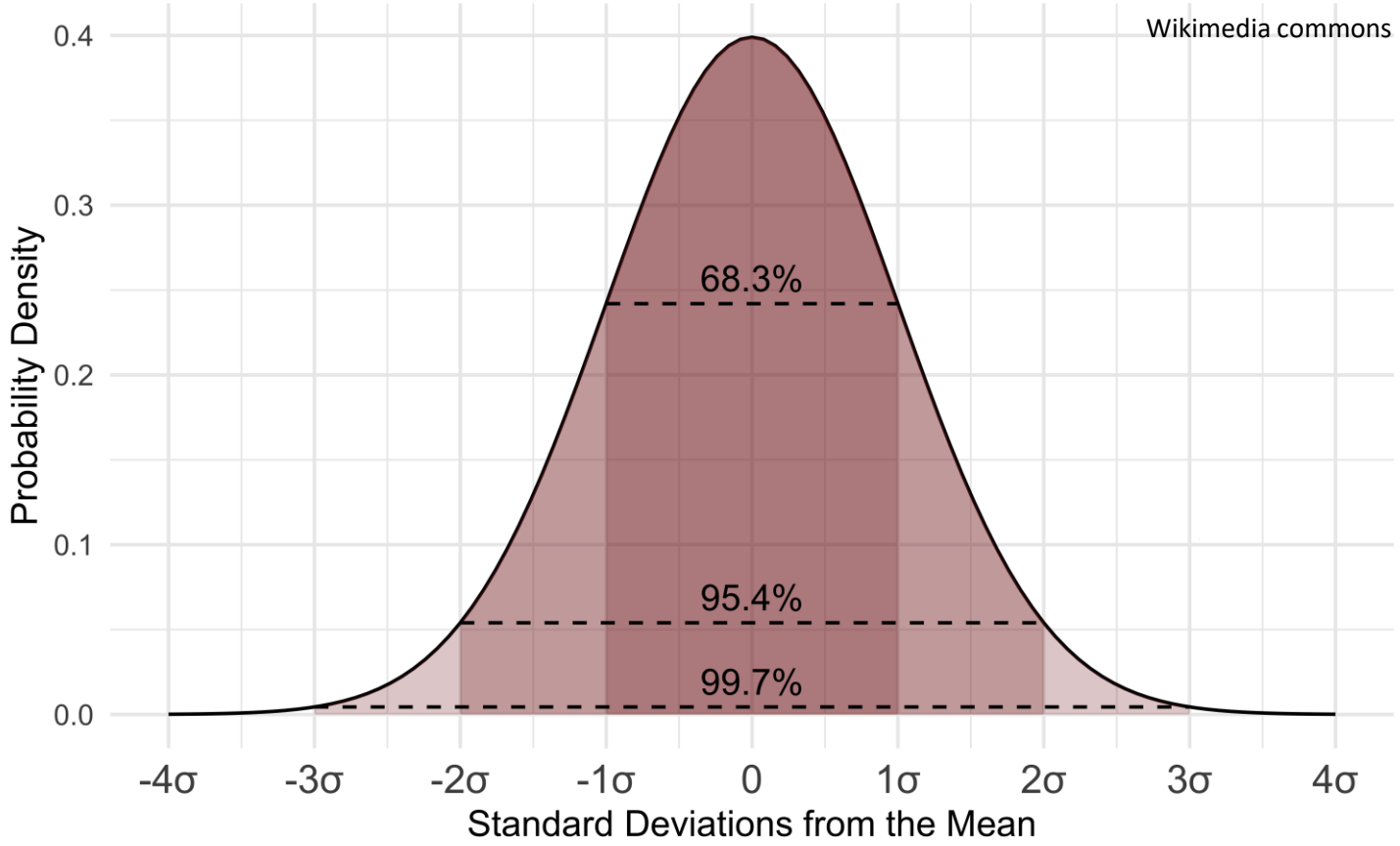
UNIT NORMAL DISTRIBUTION

aka Z distribution

- The **Z-distribution** is a standardized normal distribution

Population Mean $\mu = 0$

Population SD $\sigma = 1$

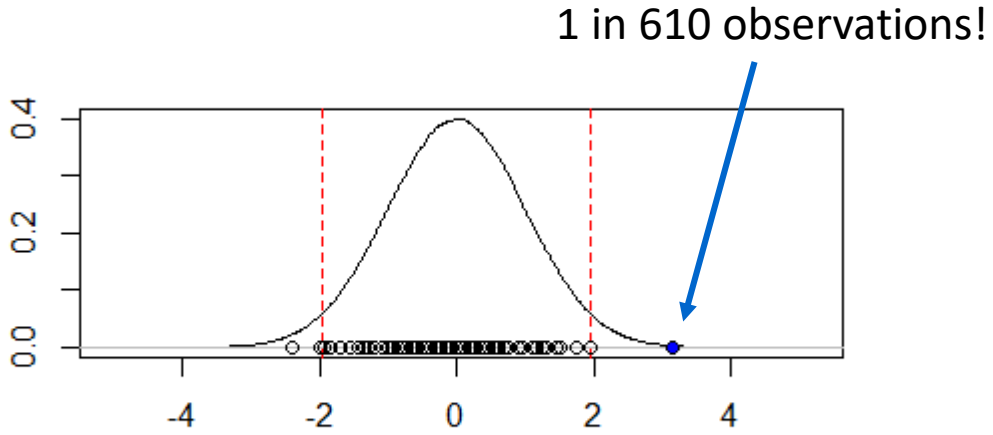


Z-SCORE

- **Z-score** is also known as the standard score:

$$\textbf{Z-score} = \frac{x-\mu}{\sigma}$$

- It measures the distance of individual observations from the mean, in terms of number of standard deviations




```
# calculate z-scores for variable "Length.cm"
x <- uncia$Length.cm
x <- x - mean(x)
x <- x / sd(x)
uncia$Length.Z <- x

# inspect outlier (length = 150 cm)
uncia[uncia$Length.cm==150,]
    X Location Sex Length.cm Weight.kg Length.scaled
22 22   Wakhan   M       150        55      3.149009
```

CENTRAL LIMIT THEOREM

- **Sampling distribution of means**

- For any population, regardless of the underlying distribution, it *approximates* normal distribution with increasing sample size

  Mean = μ (Population mean)

  SD = $\frac{\sigma}{\sqrt{n}}$ (Standard error)

- **Implications**:

  - Applicable even for variables that are originally not normally distributed

  - Methods that work for normal distributions can apply to many problems involving other types of distributions

HYPOTHESIS TESTING

STEP 1: State the null hypothesis:

$\mathbf{H_0}$**:** $\mu_1 = \mu_2$ (the two population means are equal)

STEP 2: State the alternative hypothesis:

$\mathbf{H_1}$ **(two-tailed):** $\mu_1 \neq \mu_2$

$\mathbf{H_1}$ **(left-tailed):** $\mu_1 < \mu_2$

$\mathbf{H_1}$ **(right-tailed):** $\mu_1 > \mu_2$

STEP 3: Set alpha (level of significance)

$\alpha = 0.05$

STEP 4: Compare value of test statistic to critical value

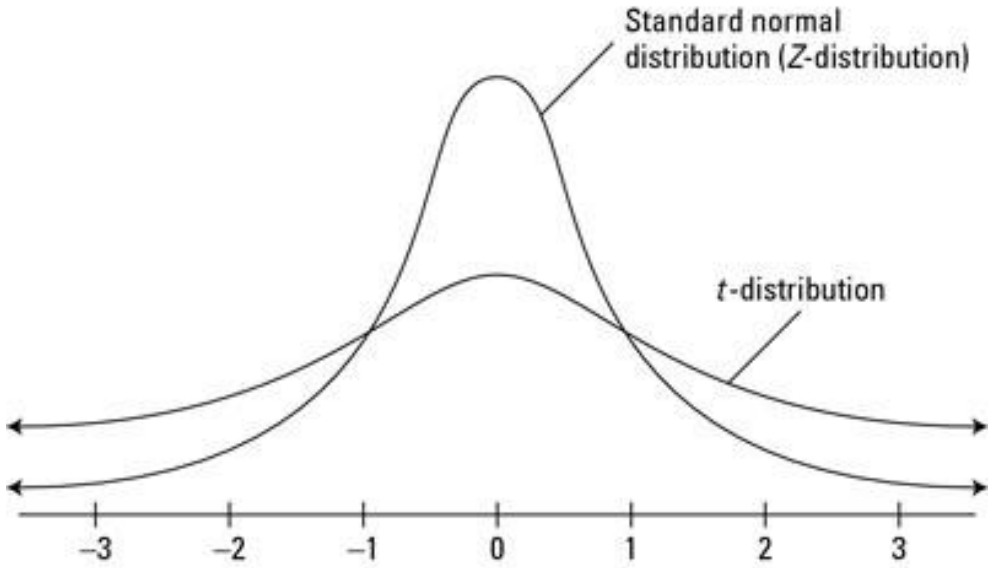
Example of critical value table

Degrees of freedom	Significance level					
	20% (0.20)	10% (0.10)	5% (0.05)	2% (0.02)	1% (0.01)	0.1% (0.001)
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850

Jimgrange/wordpress

T-TEST

- When σ is not known
- For small sample sizes (< 30)
- For larger sample size: t distribution approximates the z distribution



ASSUMPTIONS:

- Data points are independent
- Data is (approximately) normally distributed.
- Samples being compared have a similar amount of variance within each group being compared (a.k.a. homogeneity of variance)

Different types:

- One sample t-test $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$
- Two sample t-test $t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{(s^2(\frac{1}{n_1} + \frac{1}{n_2}))}}$
- Paired t-test $t = \frac{x_{diff}}{s_{diff}/\sqrt{n}}$

UNCIA BCI

Condition index	Abbreviation	References
Body mass divided by body length	BM/BL	Lunn and Boyd 1993, Huot et al. 1995
Log-transformed body mass divided by log-transformed body length	logBM/logBL	Hayes and Shonkwiler 2001

Labocha , Schutz and Hayes (Oikos 123: 111–119, 2014)

```
# first method : body mass divided by body length
uncia$BCI_1 <- uncia$Weight.kg / uncia$Length.cm
# second method : log (body mass) divided by log (body length)
uncia$Weight.log <- log(uncia$Weight.kg)
uncia$Length.log <- log(uncia$Length.cm)
uncia$BCI_2 <- uncia$Weight.log / uncia$Length.log

# boxplots for BCI_2 for each location
boxplot(subset(uncia$BCI_2, uncia$Location == "Wakhan")) # boxplot for Wakhan
boxplot(subset(uncia$BCI_2, uncia$Location == "Altai")) # boxplot for Altai
boxplot(BCI_2 ~ Location, data = uncia) # both combined

# t-test to compare mean of BCI_2 between the two locations
bci.wakhan <- subset(uncia$BCI_2, uncia$Location == "Wakhan")
bci.altai <- subset(uncia$BCI_2, uncia$Location == "Altai")
t.test(x = bci.wakhan, y = bci.altai, alternative = "two.sided", var.equal = TRUE)
```

NON-PARAMETRIC TESTS

Parametric tests

- Data is normally distributed or can be approximated using normal distribution
- Statistics used: Mean, standard deviation
- Numeric variables
- Influenced by outliers
- Higher statistical power

Non-Parametric tests

- No assumptions about distribution
- a.k.a distribution free methods
- Statistic used: Median
- Both numeric and categorical
- Not influenced by outliers
- Lower statistical power