Population - entire collection of objects or individuals about which information is desired

Sample - part of the population that is selected for analysis
- Watch out for sample bias

Simple random sampling - every possible sample of a certain size has the same chance of being selected

Mean - the point of balance on a seesaw; the arithmetic average of the data
- susceptible to extreme values (outliers) by moving towards extreme values

Median - in an ordered vector, the median is the middle number
- not affected by extreme values

Quartiles - split the ranked data into 4 equal groups

Variance - the average distance, squared

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

- squaring gets rid of the negative value problem

Standard deviation - shows variation about the mean

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

- highly affected by outliers
- has same units as original data

Observational study - there can always be lurking variables affecting results

Experimental study - lurking variables can be controlled; can give good evidence for causation (if internal validity is high)

Summary measures

Describing data numerically

Center & location	Other Measures of location	Variation
Mean	Quartiles	Range
Median		Interquartile Range
		Variance
		Standard Deviation

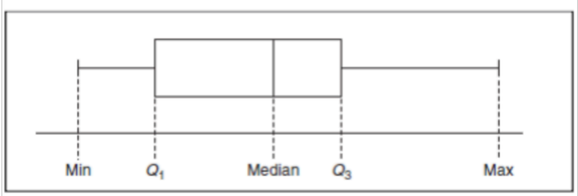
Linear transformations

$$Y = a + bx$$

a: shifts data by a
b: changes scale

Linear transformations change the centre and spread of data.
- *Var(a + bX) = b²Var(X)*
- *Average(a+bX) = a+b[Average(X)]*

Box and Whisker Plot



Range = X_{max} - X_{min}
- disadvantages: ignores the way in which data are distributed; sensitive to outliers

Interquartile range (IQR): 3rd quartile - 1st quartile
- not used often
- not affected by outliers

Effects of Linear Transformations:
- mean $_{new}$: a + b * mean
- median $_{new}$: a + b * median
- stdev $_{new}$ = |b| * stdev
- IQR $_{new}$ = |b| * IQE

z-score: new data set will have mean of 0, and variance of 1

$$z = \frac{x - \bar{x}}{S}$$

Detecting outliers

- Classic outlier detection

$$|z| = |\frac{x - \overline{x}}{S}| \geq 2$$

- The Boxplot Rule

$$X < Q1 - 1.5(Q3 - Q1)$$

or

$$X > Q3 + 1.5(Q3 - Q1)$$

Skewness

- measures the degree of asymmetry exhibited by the data
  - negative values = skewed left
  - positive values = skewed right
  - if |skewness| < 0.8 = no need to transform the data

Measures of association

- Covariance

Covariance > 0 = larger x, larger y
Covariance < 0 = larger x, smaller y

$$cov_{xy} = \frac{1}{n-1} \sum (x - \overline{x})(y - \overline{y})$$

Units = Units of x · Units of y
Covariance is only +, -, or 0
(can be any number)

Central Limit Theorem

- as n increases, $\overline{x}$ (sample mean) should get closer to μ (population mean)

- mean of $\overline{x} = \mu$
- variance of $\overline{x} = \sigma^2/n$

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

if population is normally distributed, *n* can be any value

any population, *n* needs to be ≥ 30

$$Z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

Probability

- measure of uncertainty
- all outcomes have to be exhaustive (all options possible) and mutually exhaustive (no 2 outcome can occur at the same time)

Rules

- Probabilities range from

$$0 \leq Prob(A) \leq 1$$

- The probabilities of all outcomes must add up to 1
- The complement rule = A happens or A doesn't happen

$$P(\overline{A}) = 1 - P(A)$$
$$P(A) + P(\overline{A}) = 1$$

- Addition rule:

$$P(A or B) = P(A) + P(B) - P(A or B)$$

Correlation - measures strength of a ***linear*** relationship between two variables

$$r_{xy} = \frac{cov_{xy}}{(s_x)(x_y)}$$

correlation is between +1 and -1
sign: direction of relationship
absolute value: strength of relationship (-0.6 stronger than +0.4)

0 - 0.2: very weak
0.2 - 0.4: weak to moderate
0.4 - 0.6: medium to substantial
0.6 - 0.8: very strong
0.8 - 1.0: extremely strong

correlation doesn't imply causation
correlation of a variable with itself is 1

Confidence intervals - tells us how good our estimate is
- Want high confidence, narrow interval
- As confidence increases, interval also increases

Margin of Error
- The confidence interval is given by

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\,\hat{q}}{n}}$$

- The standard form of any confidence interval is estimate $\pm$(margin of error)

Effect size - *Cohen's d*

$$d = \frac{\mu_{treatment} - \mu_{control}}{s}$$

where d is Cohen's d, μ is a mean, and s is the standard deviation

Standard Error of the Mean (S.E.M.)

$$\sigma_M = \frac{\sigma}{\sqrt{(n)}}$$

where σ_M is the S.E.M., σ is the standard deviation and n is the sample size.

Hypothesis testing
- Null hypothesis
- H_0, a statement of no change and is assumed true until evidence indicates otherwise
- Alternative hypothesis: H_a, is a statement that we are trying to find evidence to support
- Type I error: reject the null hypothesis when the null hypothesis is true
- Type II error: do not reject the null hypothesis when the alternative hypothesis is true

	Hypothesis testing (p-values)	Effect size	Power
Type of conclusion	Threshold based, yes/no	Gradient, small to large	Gradient, small to large
Sample size	Sensitive	Not sensitive	Sensitive
Alpha level	Sensitive	Not sensitive	Sensitive
# of tails	Sensitive	Not sensitive	Sensitive

Example of Type I and Type II errors
- Null hypothesis
- H_0, a statement of no change and is assumed true until evidence indicates otherwise
- Alternative hypothesis: H_a is a statement that we are trying to find evidence to support
- Type I error: reject the null hypothesis when the null hypothesis is true
- Type II error: do not reject the null hypothesis when the alternative hypothesis is true

Methods of Hypothesis testing
- Confidence intervals
- Test statistics
- p-values
- C.I. and p-values always safe to do because don't need to worry about size of n (can be bigger or smaller than 30)

Cheat Sheet

R Functions (supplement)

For more information on each function, use the question mark and the name of the function in R Studio, e.g.:

?sum()

Check data frame type

is.integer(), is.character(), is.logical()

as.integer(), as.character(), as.logical()

class()

Glimpse into the data

names()

head()

summary()

Formatting the data

sort()

rle()

table()

round()

Basics

length()

sum()

Central Tendency

median()

mean()

Variability

min(), max()

range()

quantile()

IQR()

var()

cor()

sd()

Graphs

plot()

hist()

boxplot()

```
barplot(
  meansVector,
  ylim = c(0,0),
  names.arg = c("list","labels"),
  main = "title"
)
```

Arrows on graphs

```
arrows(
  middle, mean-sem,
  middle, mean+sem,
  length = 0.05,
  angle = 90,
  code=3
)
```

Multiple graph facets

par(mfrow=c(1,2))

Other

unique()

sqrt()

scale()

Subsetting Vectors

```
variable[variable == "value"]
variable[variable > 0]
```

Subsetting Data Frames

```
df[df$col == 0,]
df[df$col > 0,]
```

Probability

pnorm()

Programming Concepts

```
for(i in  1:10) {
 your code goes here
}

myfunc <- function(variable) {
  your code goes here
}

if(condition == TRUE) {
    do something
}
```