

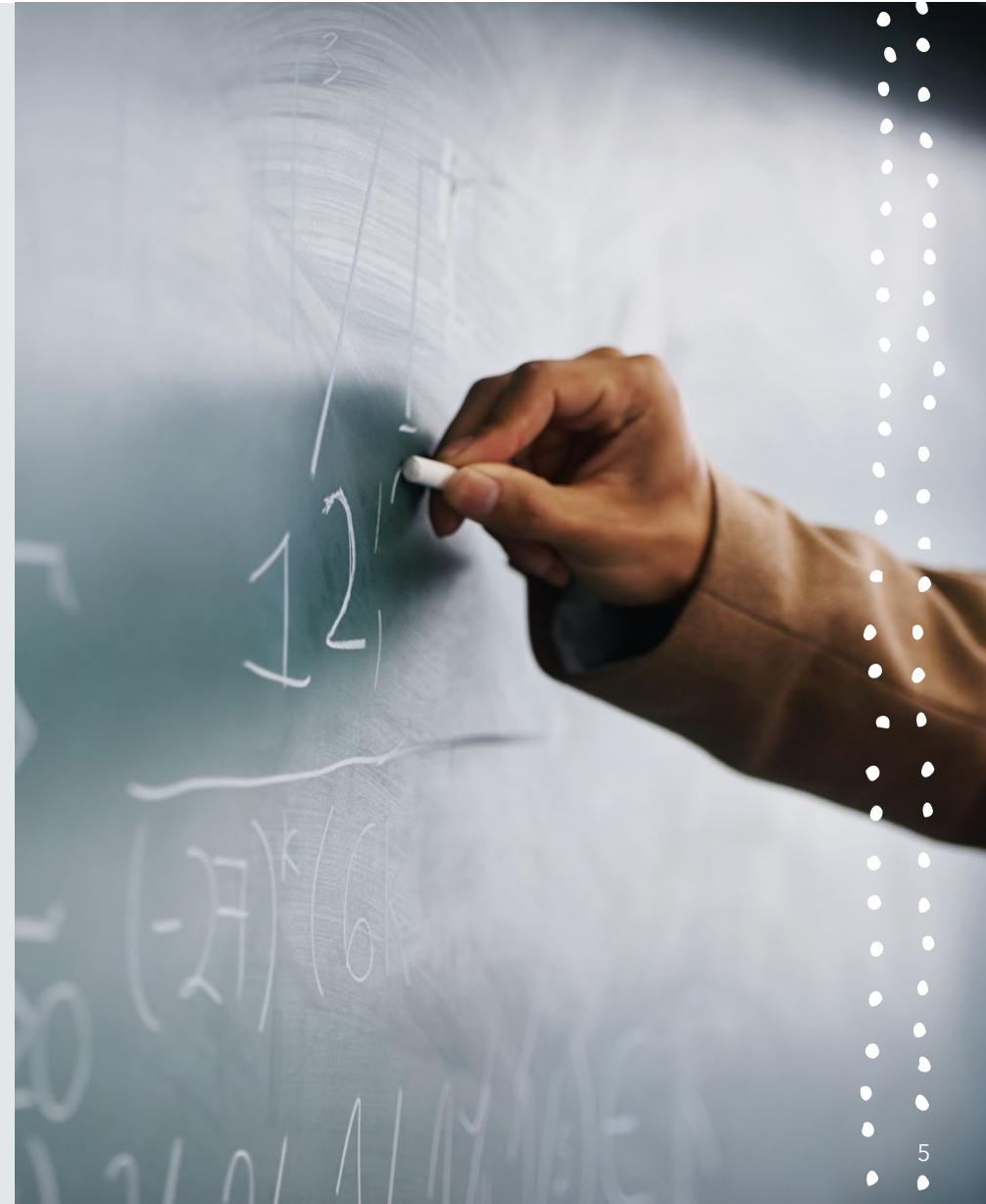
The principles of biostatistics

PROF. DR. JAMALLUDIN AB RAHMAN MD MPH
DEPARTMENT OF COMMUNITY MEDICINE, KULLIYYAH OF
MEDICINE

Learning outcomes

At the end of this workshop, you should be able to

1. Install and use R for,
2. Descriptive, and
3. Analytical statistics.



Truth

We observe.

We believe what we observe.

What we observed might not be the truth.

What we do with statistics

1. Describe data/variable
2. Compare data/variables/estimates
 1. Detect differences
 2. Quantify degree of differences
 3. Identify trend
 4. Make prediction

Is there any difference
between A & B?

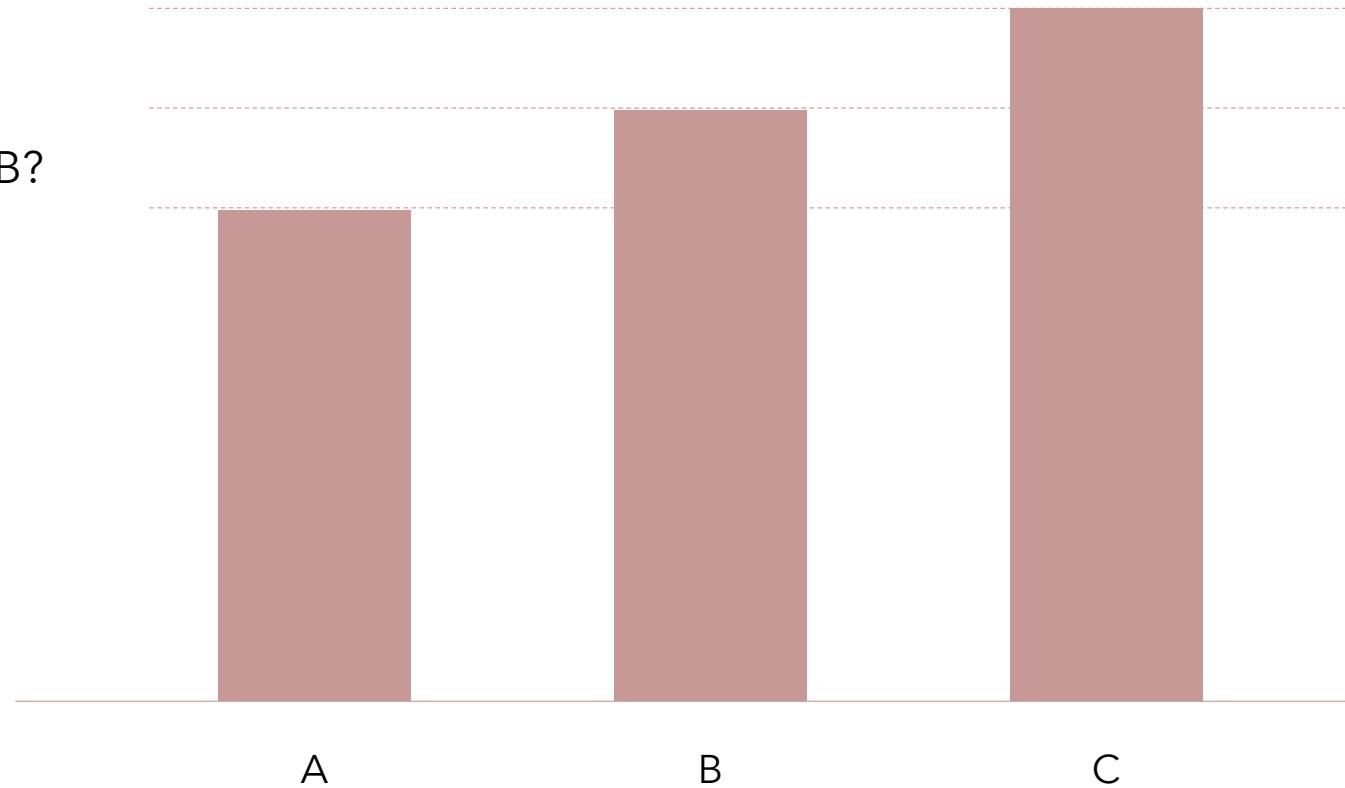
Which one is taller? A or B?

How big is the difference
between A & B?

Is C different from A & B?

Is there any pattern now?

If there will be D, can you
predict how tall is D?



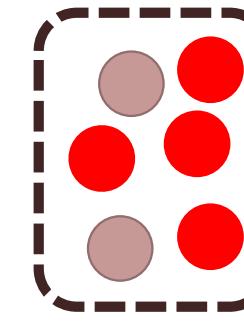
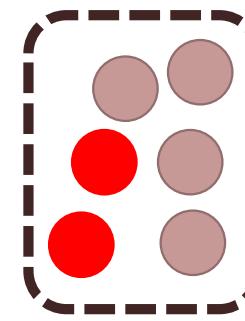
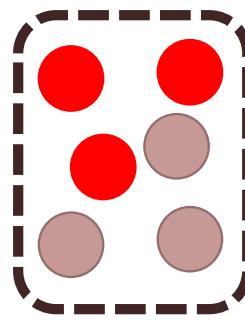
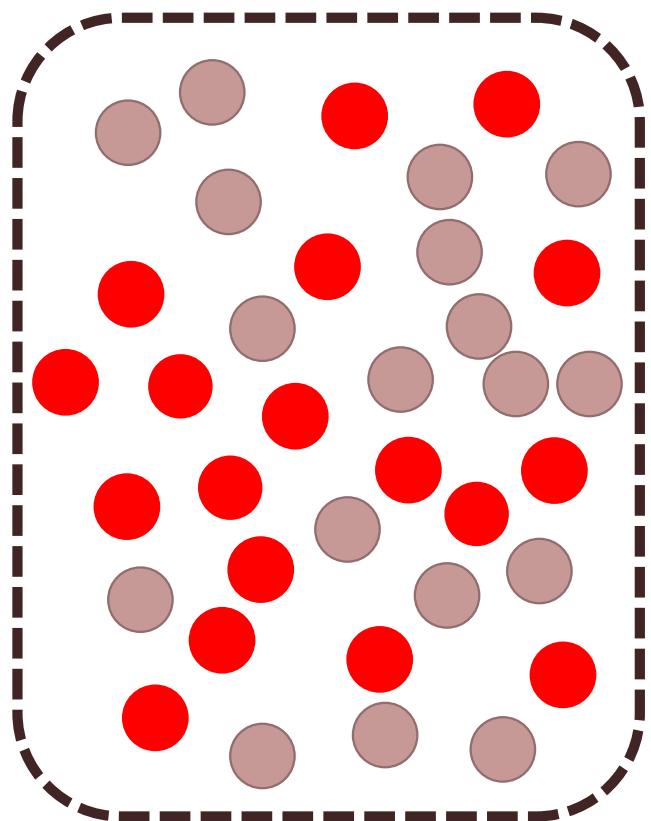
Sample statistics

In research, we usually analyse from samples, not the entire population

Statistics (from samples)
can estimate parameters
(of the population) if

- The samples represent population - low sampling error
- But different samples yield different estimates
- How statistics estimate population then?





Statistics

$$\text{Average \% Red} = \frac{50\% + 33.3\% + 66.7\%}{3} = 50\%$$

Estimated
Parameter

Parameter



Causation

- Relation of two variables
- Correlation does not imply causation
- Rooster's crow does not **cause** the sun to rise
- Switch does not **cause** the bulb to light

Meeting January 14 1965

President's Address

The En
Associa
by Sir A
(Profess
Universi

Among
of Occ
means
physic
of th
and
lems
coll

ings with other Sections, secondly, 'to make available information on the physical, chemical and psychological hazards of occupation, and in particular about those that are rare or not easily recognized'.

have to be
It will depend upon circum-

I have no wish, nor the skill, to embark upon a philosophical discussion of the meaning of 'causation'. The 'cause' of illness may be immediate and direct, it may be remote and indirect underlying the observed association. But with the aims of occupational, and almost synonymously preventive, medicine in mind the decisive question is whether the frequency of the undesirable event B will be influenced by a change in the environmental feature A. How such a

Hill's criteria

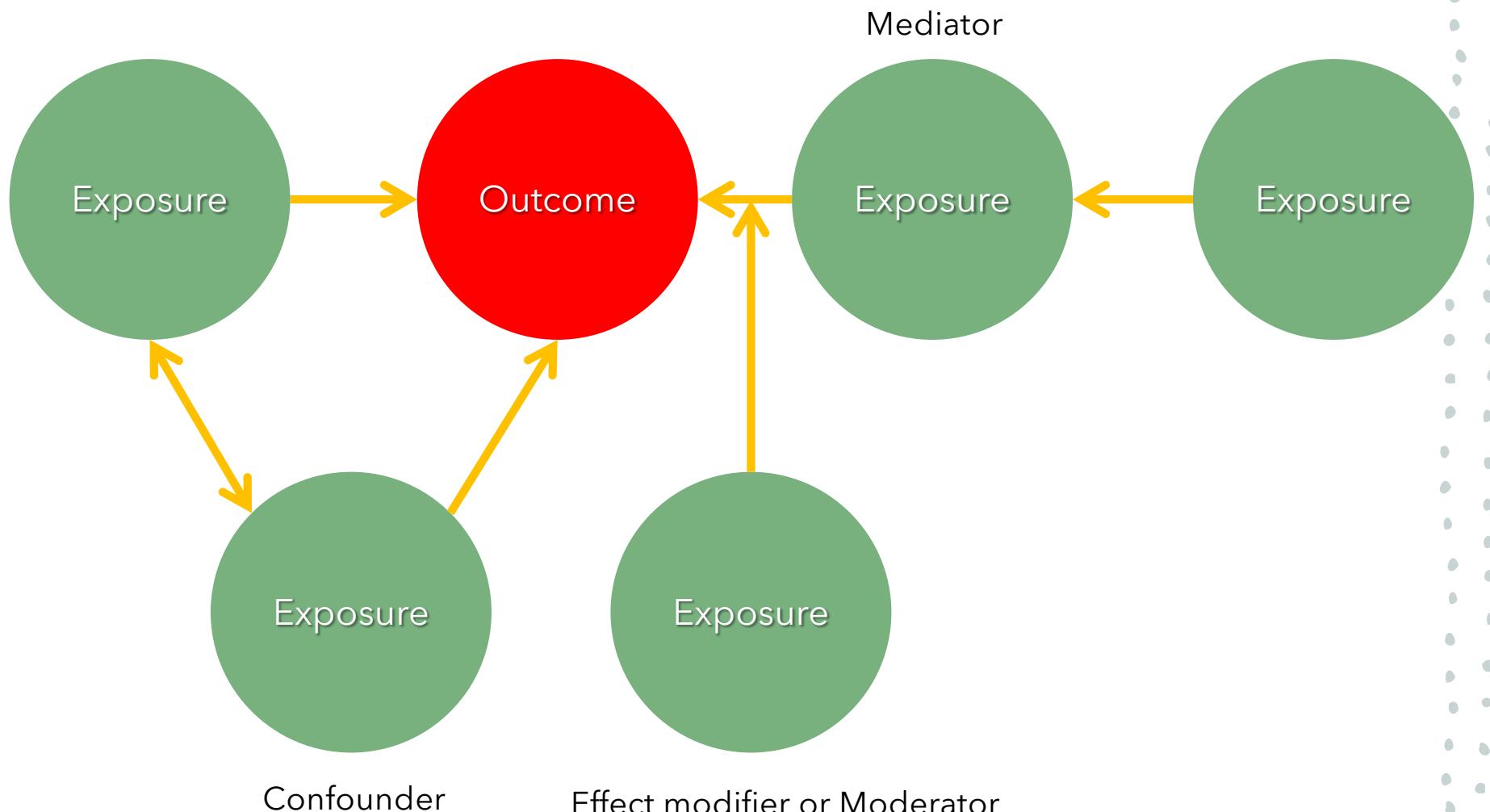
- 1. Strength of association
- 2. Consistency
- 3. Specificity
- 4. Temporality
- 5. Biological gradient
- 6. Plausibility
- 7. Coherence
- 8. Experiment
- 9. Analogy

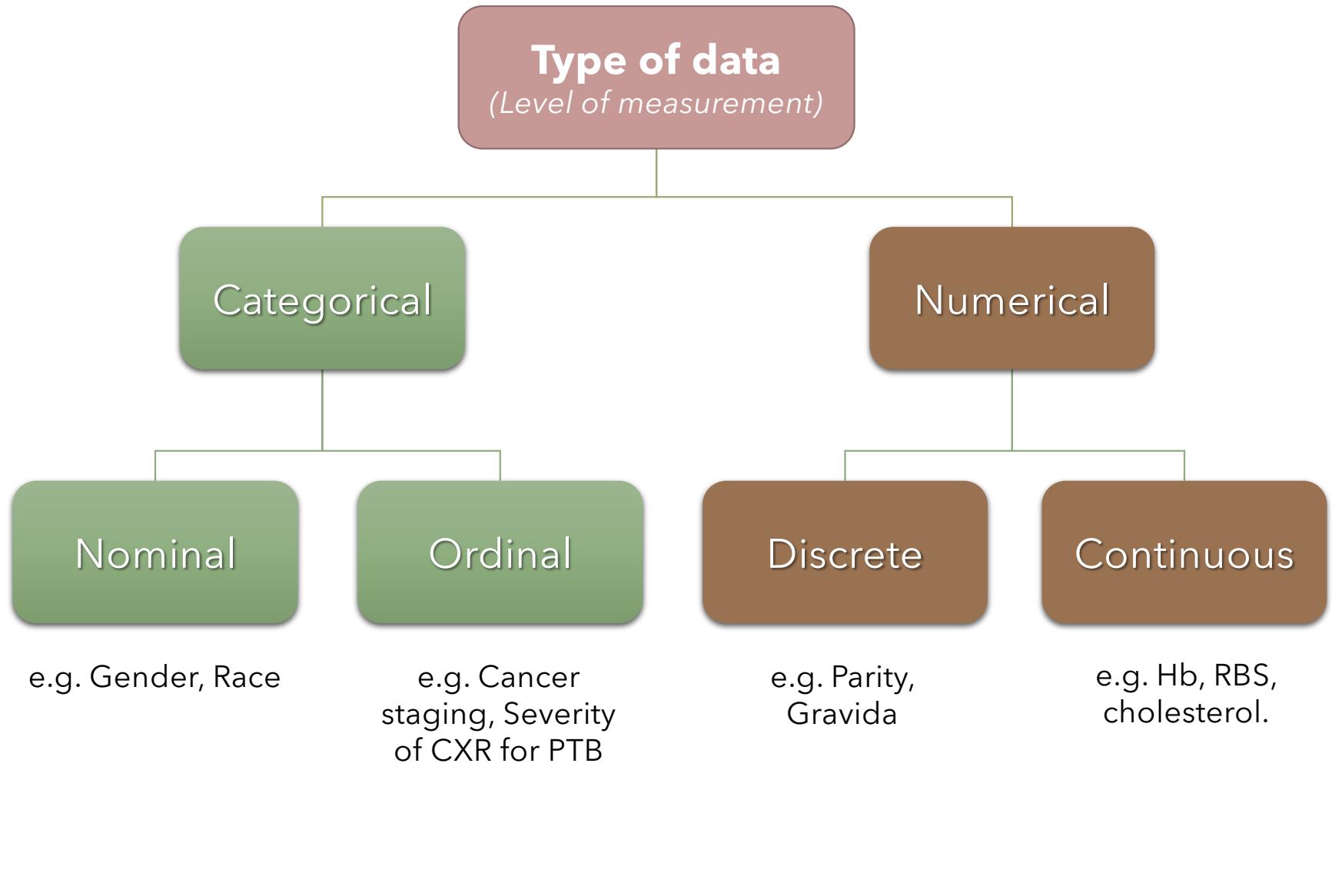
Hill AB. The environment and disease:
Association or causation? Proceed Roy Soc
Medicine – London. 1965;58:295–300.

Causal web

- Web of causation
- Conceptual framework
- Path analysis/web
- Relationship between variables
- Cause and effect







Distribution (shape) of data

- Applicable to **numerical** value
- Discrete or Continuous
- Discrete \sim Binomial, Poisson, Negative Binomial, Hypergeometry, Multinomial etc.
- Continuous \sim Normal, t, chi-square, F etc.

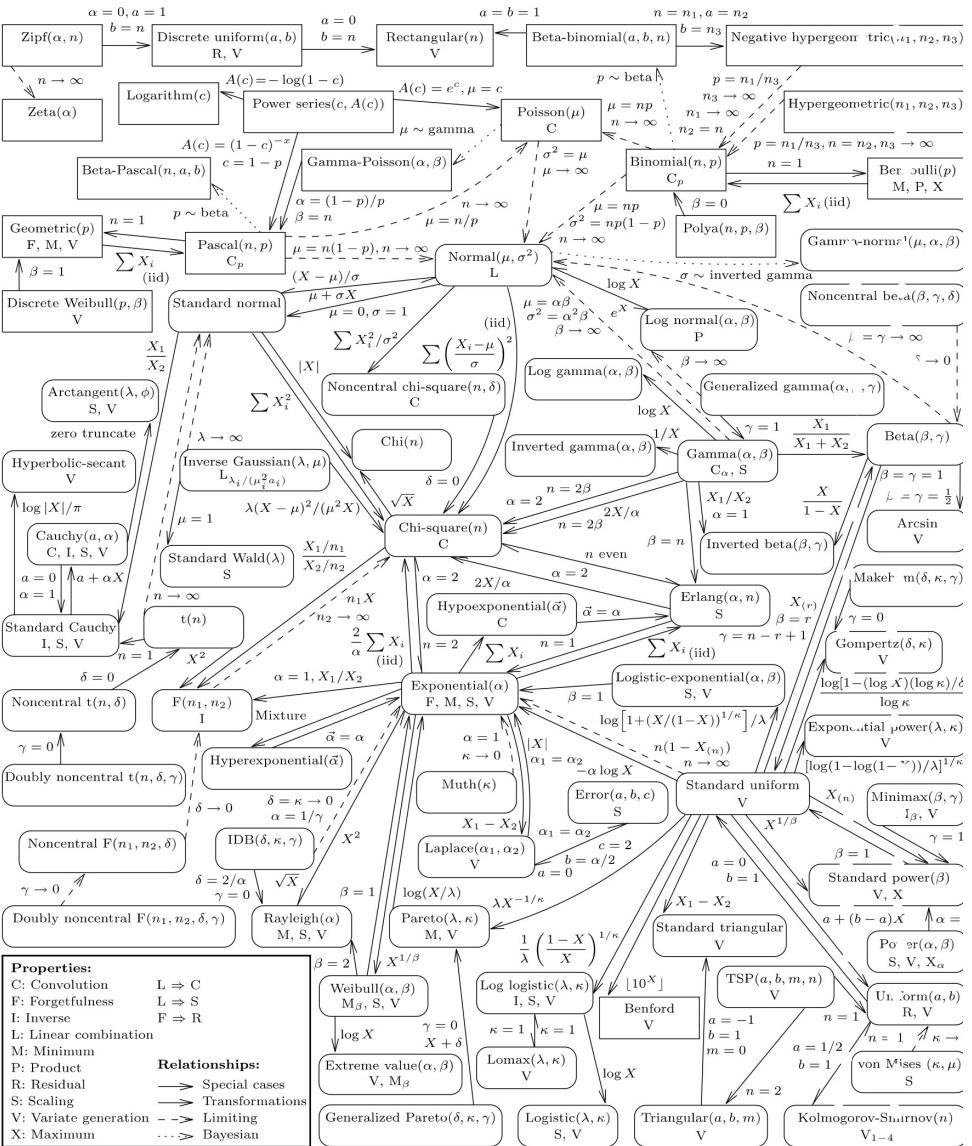
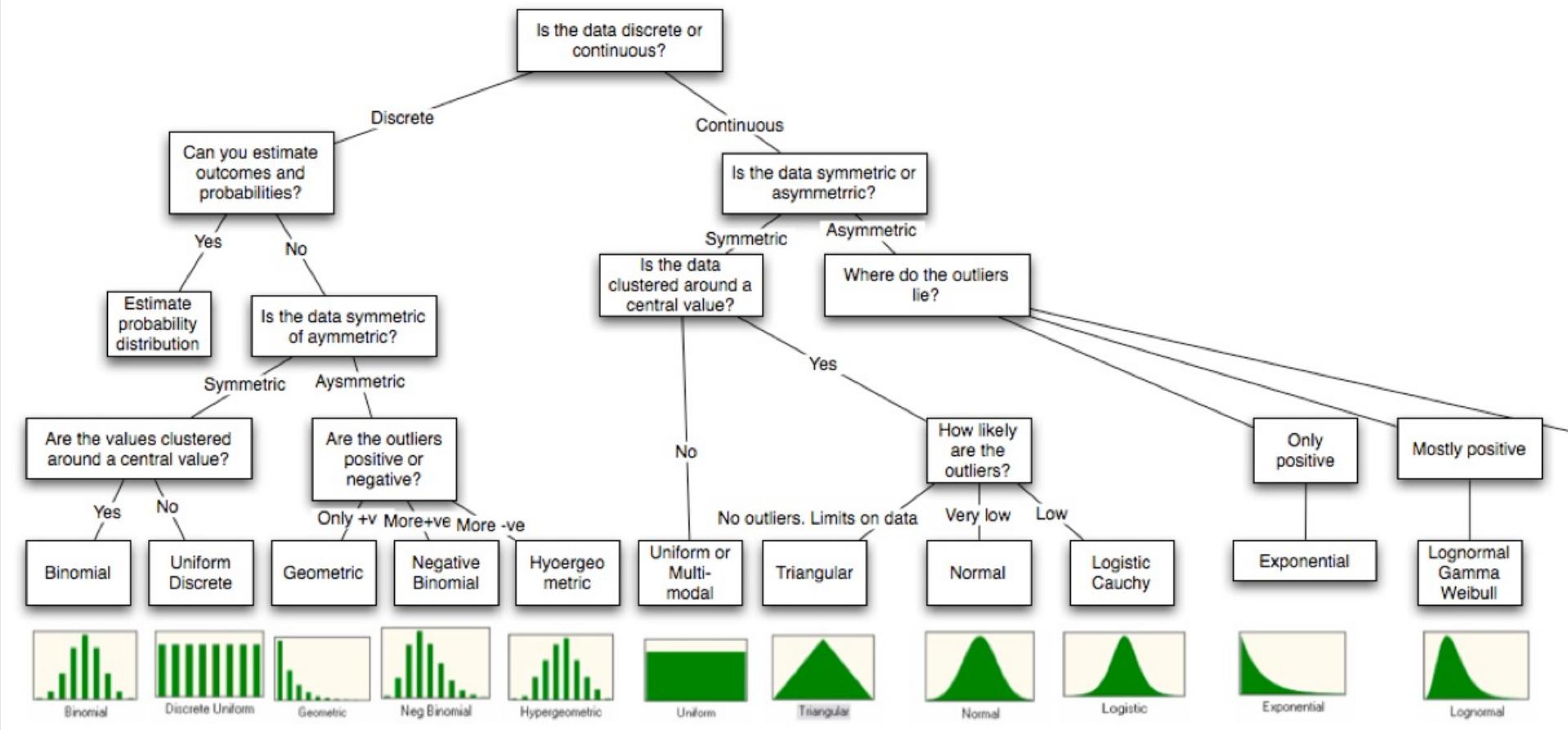
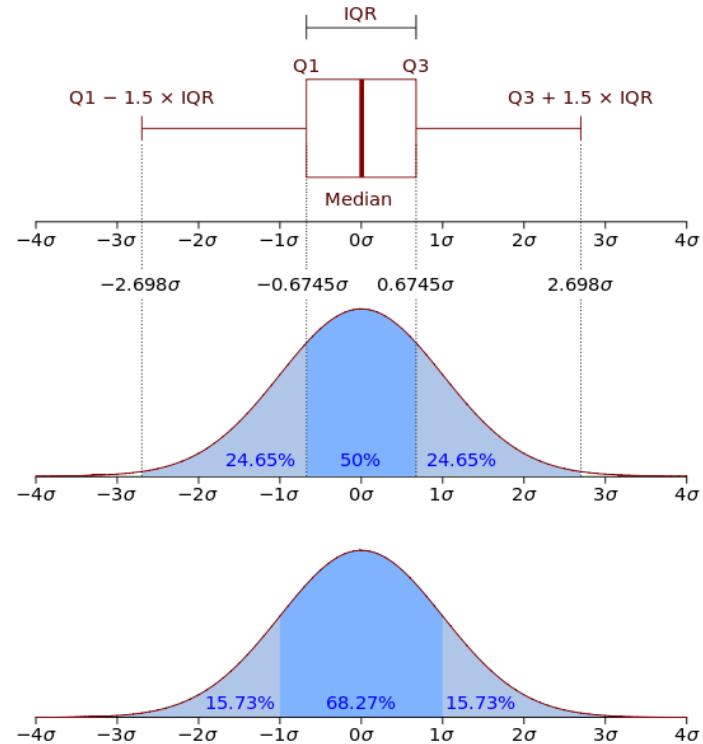


Figure 1. Univariate distribution relationships.

Figure 6A.15: Distributional Choices



Normal Distribution



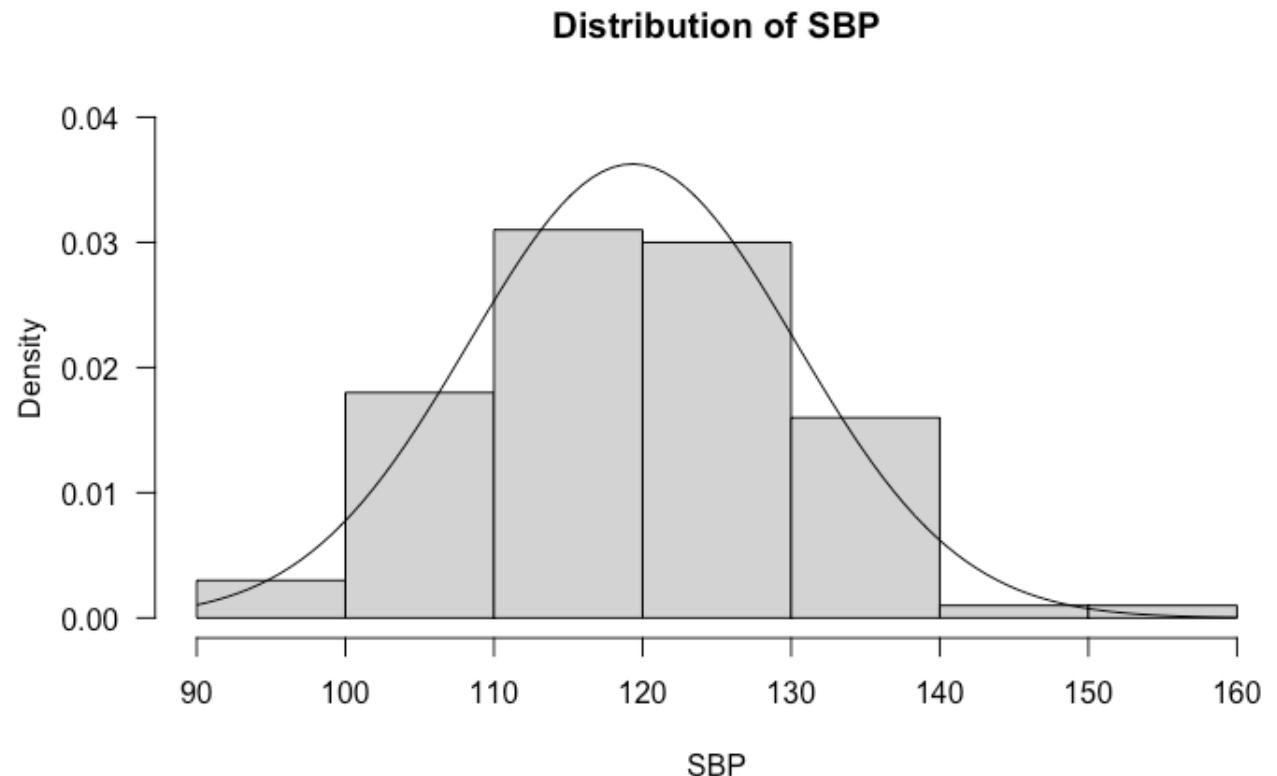
$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

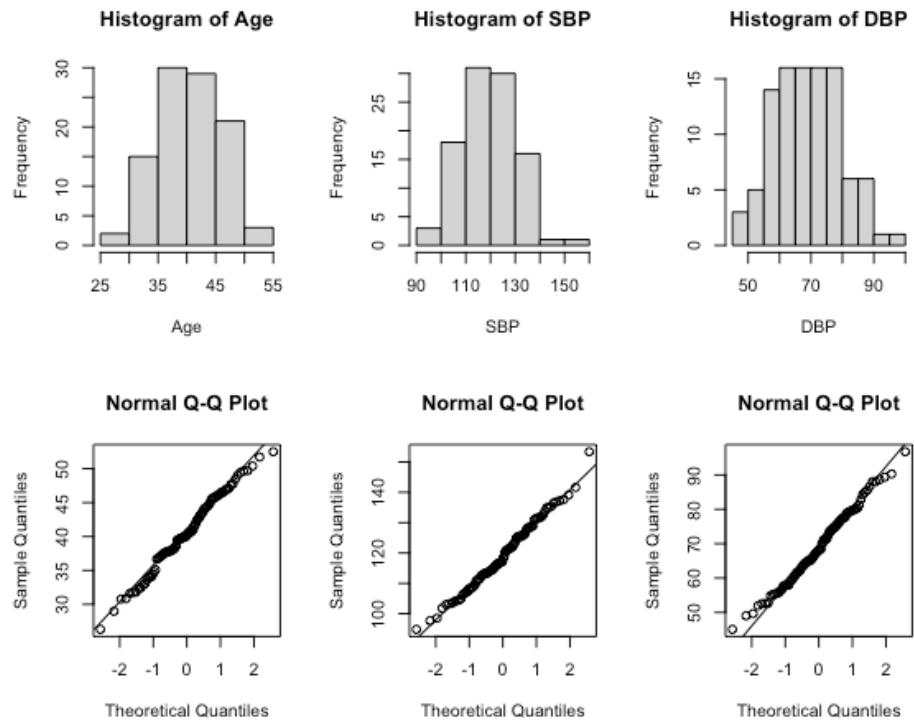
Why Normal?

- Because many biological & psychological variables are distributed normally
- Many predictive models assume Normal distribution

Characteristics

- Bell shaped curve
- Symmetrical
- Unimodal





```
# Checking distribution
```

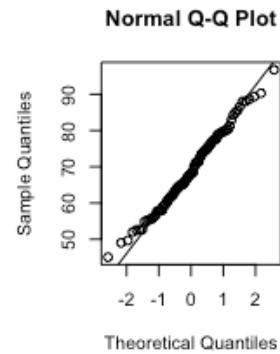
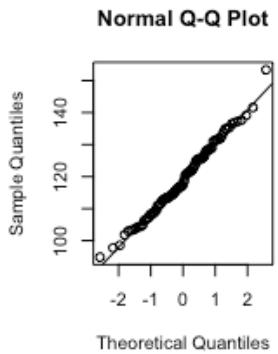
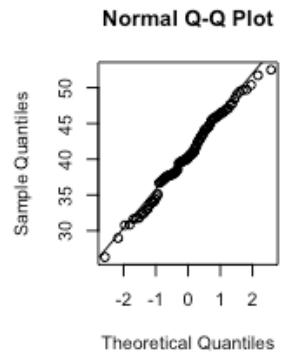
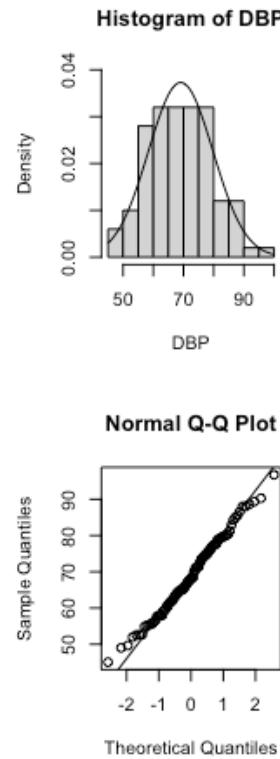
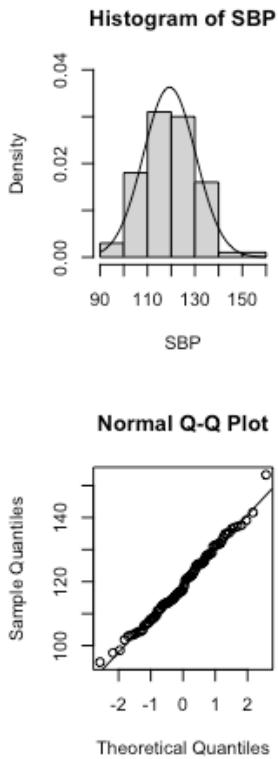
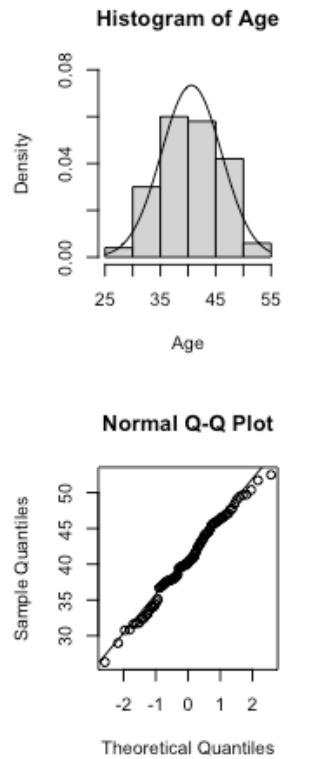
```
## Visually
```

```
par(mfrow= c(2,3))
hist(Age)
hist(SBP)
hist(DBP)
```

```
qqnorm(Age)
qqline(Age)
```

```
qqnorm(SBP)
qqline(SBP)
```

```
qqnorm(DBP)
qqline(DBP)
```



```
# Showing Normal curve
```

```
hist(Age, ylim=c(0,.08), probability = TRUE)
curve(dnorm(x, mean=mean(Age), sd=sd(Age)), add=TRUE)
```

```
hist(SBP, ylim=c(0,0.04), probability = TRUE)
curve(dnorm(x, mean=mean(SBP), sd=sd(SBP)), add=TRUE)
```

```
hist(DBP, ylim=c(0,0.04), probability = TRUE)
curve(dnorm(x, mean=mean(DBP), sd=sd(DBP)), add=TRUE)
```

```
# Q-Q Plot
```

```
qqnorm(Age)
qqline(Age)
```

```
qqnorm(SBP)
qqline(SBP)
```

```
qqnorm(DBP)
qqline(DBP)
```

Characteristics of a normal distribution

1. Shape of the curve - smooth symmetrical bell-shaped
2. Mean=Median
3. Skewness ~ 0 (usually ± 2)
4. Kurtosis ~ 0 (usually ± 2)

```
> library(psych)
> describe(Age)
  vars   n   mean    sd median trimmed   mad   min   max range skew kurtosis    se
X1     1 100 40.62 5.44   40.18    40.7 5.18 26.31 52.5  26.2 -0.12   -0.47 0.54
> describe(SBP)
  vars   n   mean    sd median trimmed   mad   min   max range skew kurtosis    se
X1     1 100 119.32 11 117.81   119.12 11.49 94.85 153.37 58.52  0.24   -0.23 1.1
> describe(DBP)
  vars   n   mean    sd median trimmed   mad   min   max range skew kurtosis    se
X1     1 100 69.09 10.7 68.12    68.84 11.7 45.07 96.75 51.68  0.15   -0.58 1.07
```

Test of Normality

- Anderson-Darling Test
- Corrected Kolmogorov-Smirnov Test (Lilliefors Test)
- Cramér-von-Mises Criterion
- D'agostino's K-squared Test
- Jarque-Bera Test
- Pearson's Chi-square Test
- Shapiro-Francia
- Shapiro-Wilk Test

```
> # Normality test  
>  
> shapiro.test(Age)
```

Shapiro-Wilk normality test

```
data: Age  
W = 0.98956, p-value = 0.6295
```

```
> shapiro.test(SBP)
```

Shapiro-Wilk normality test

```
data: SBP  
W = 0.98957, p-value = 0.6297
```

```
> shapiro.test(DBP)
```

Shapiro-Wilk normality test

```
data: DBP  
W = 0.99029, p-value = 0.688
```

Use Normality test with caution

- Small samples almost always pass a normality test. *Normality tests have little power to tell whether or not a small sample of data comes from a Gaussian distribution.*
- With large samples, minor deviations from normality may be flagged as statistically significant, *even though small deviations from a normal distribution won't affect the results of a t test or ANOVA.*

What would you do if the distribution is not normal?

01

Validate data entry

02

Transform the data

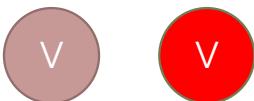
03

Use a non-parametric test

Statistical analysis

Descriptive

Univariable

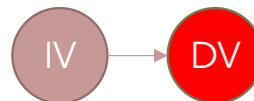


e.g., Describe socio-demographic characteristics - Age, Sex, Race etc.

e.g., Prevalence of hypertension.

Analytical

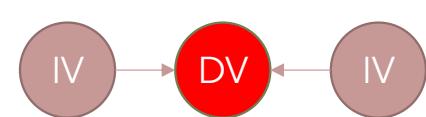
Bivariable



e.g., Compare demographic characteristics between two population - Compare age between male & female

e.g., Distribution of gender by hypertension status

Multivariable



e.g., How demographic characteristics (more than one factors) explain hypertension

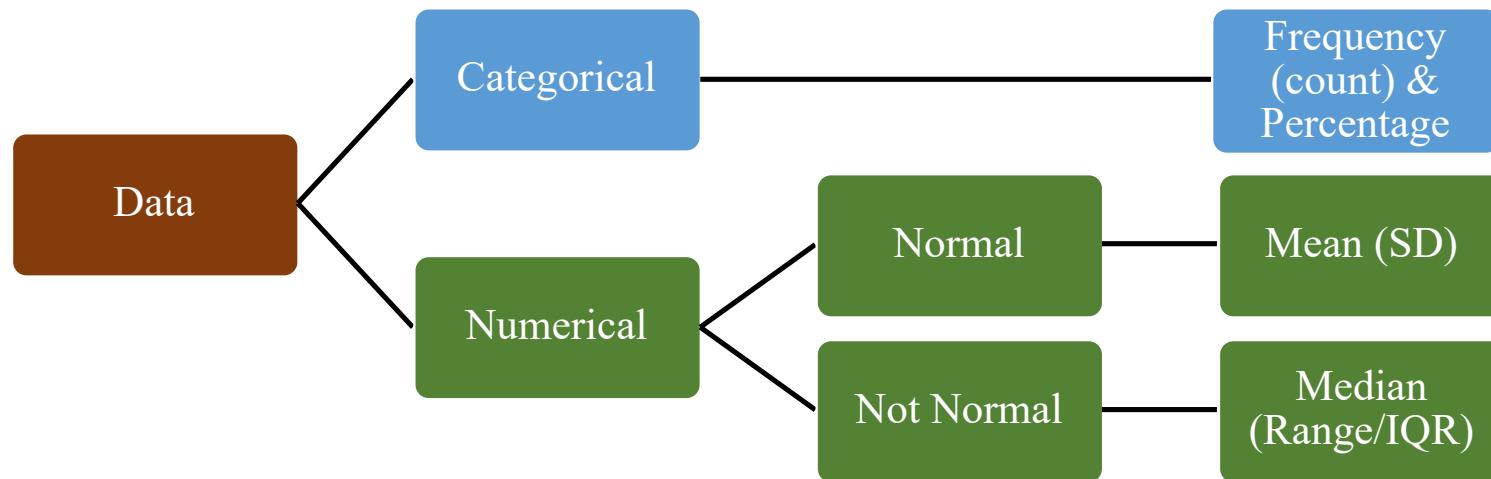
Descriptive statistics

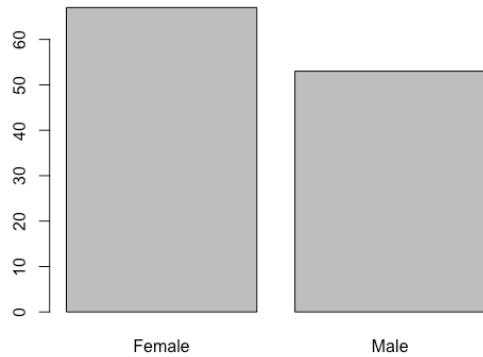
BASIC BIOSTATISTICS
JAMALLUDIN AB RAHMAN

What is descriptive statistics?

- Explain **one variable** at one time
- Method based on level of measurement
 - Categorical
Frequency (Percentage)
 - Numerical
Central measures (e.g. mean, median) & Dispersion (e.g. variance, standard deviation, range, min-max, interquartile range)

How to describe a data





```
> # Describe categorical variable  
> table(Sex)  
Sex  
Female    Male  
67        53  
> round(prop.table(table(Sex))*100,  
digit=1)  
Sex  
Female    Male  
55.8     44.2  
>  
> par(mfrow= c(1,1))  
> plot(Sex)
```

Describe categorical variable

```

> # Describe numerical variables (Age, SBP & DBP)
> library(psych)
> describe(Age)
  vars   n   mean     sd median trimmed   mad   min   max range skew kurtosis    se
X1     1 120 40.01 4.48   40.02   39.94 4.35 29.9 52.43 22.53 0.19      -0.15 0.41
> describe(SBP)
  vars   n   mean     sd median trimmed   mad   min   max range skew kurtosis    se
X1     1 120 119.93 9.21 120.19  119.95 9.34 98.47 142.03 43.56 -0.02      -0.44 0.84
> describe(DBP)
  vars   n   mean     sd median trimmed   mad   min   max range skew kurtosis    se
X1     1 120 70.54 10.72  70.97   70.91 10.55 44.83 100.84 56.01 -0.16      0.05 0.98

```

Describe numerical variable

Characteristic	N = 120 [†]
ID	60.5 (34.8), 60.5 (30.8, 90.2), 1.0 - 120.0
Age	40.0 (4.5), 40.0 (37.1, 43.0), 29.9 - 52.4
Sex	
Female	67 (56%)
Male	53 (44%)
SBP	119.9 (9.2), 120.2 (113.3, 125.7), 98.5 - 142.0
DBP	70.5 (10.7), 71.0 (63.9, 77.9), 44.8 - 100.8

[†]Statistics presented: Mean (SD), Median (IQR), Range; n (%)

```
# Using gtsummary to describe all variables
library(gtsummary)
library(tidyverse)

detach(data2)
data2 %>%
 tbl_summary(
    statistic = all_continuous() ~ c("{mean} ({sd}),",
                                      "{median} ({p25}, {p75}),",
                                      "{min} - {max}"),
    digits = all_continuous() ~ 1
  )
```

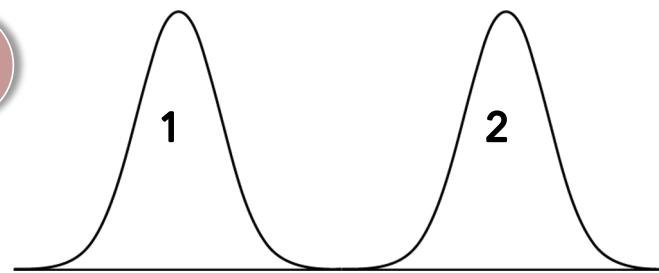
“Table 1”

Analytical statistics

BASIC BIOSTATISTICS
JAMALLUDIN AB RAHMAN

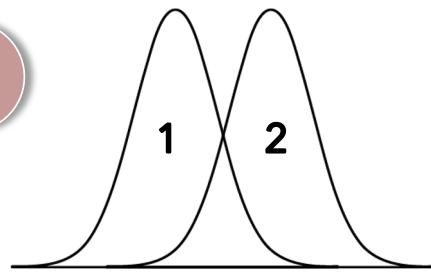
Comparing difference

A

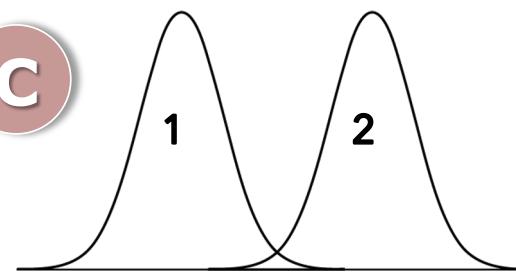


Which of the following shows true difference between the two populations?

B



C



3 statistics to compare values

1. P-value
2. Confidence interval
3. Effect size



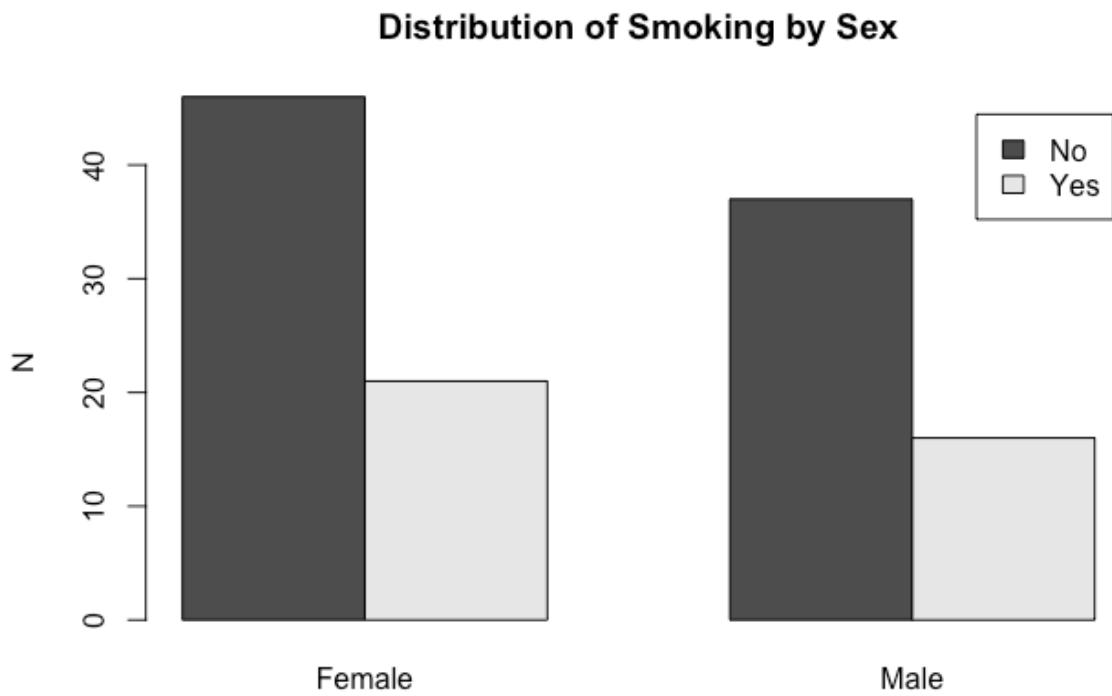
P value

- P-value is 'likely' or 'unlikely' that H_0 is true
- Taking 0.05 as the cut-off point (α), if $P \leq 0.05$, it is then 'unlikely' H_0 is true, therefore reject H_0



Hypothesis Testing using bivariable analysis

- Try to prove that there is relationship between Smoking and Sex
- H_0 : No difference in proportion of smoking between male and female



```
> # Cross-tabulate Sex with Smoking
> x <- table(Smoking, Sex)
> barplot(x, ylab="N",
+           beside = TRUE,
+           legend.text = TRUE,
+           main = "Distribution of Smoking by Sex")
> ftable(x)
   Sex Female Male
Smoking
No          46   37
Yes         21   16
> round(prop.table(x)*100, digit=1)
   Sex
Smoking Female Male
No      38.3 30.8
Yes     17.5 13.3
> chisq.test(x)

Pearson's Chi-squared test with Yates'
continuity correction
```

Confidence Interval

- Range of plausible values
- Narrow interval → high precision
Wide interval → poor precision
- How narrow is narrow? And how wide is wide? Use your clinical judgment, or in some cases can use RSE (Relative Standard Error)

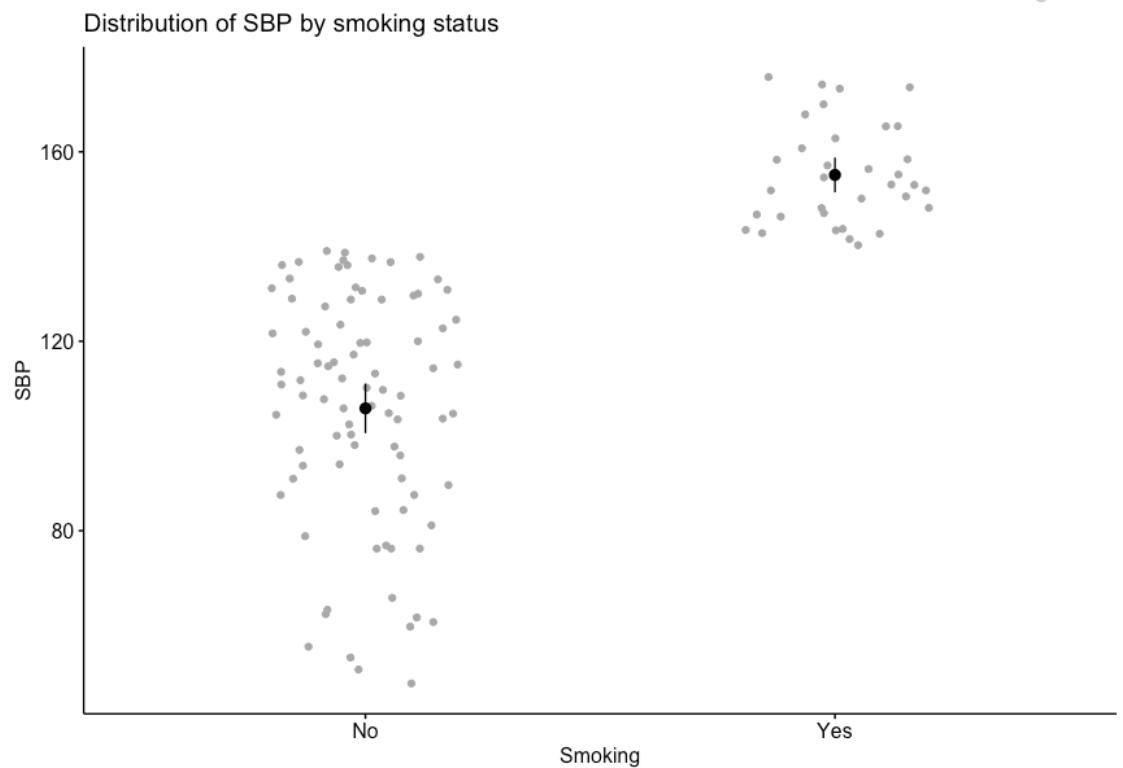
```

> # Error bars
> library(tidyverse)
> library(ggpubr)
>
> ggerrorplot(data=data3, x="Smoking",
y="SBP", main = "Distribution of SBP by
smoking status",
+             desc_stat = "mean_ci",
+             add = "jitter",
+             add.params = list(color =
"darkgray"))
>
> t.test(SBP~Smoking)

Welch Two Sample t-test

data: SBP by Smoking
t = -15.392, df = 117.07, p-value < 2.2e-16
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
-55.64642 -42.95939
sample estimates:
mean in group No mean in group Yes
105.8336      155.1365

```



42

Effect size

- The measure of effect irrespective of sample size
- Cohen (1988) classify effect size into
 - Low (<0.3)
 - Medium (0.3-0.7)
 - Large (> 0.7)
- Manual calculation or web-based calculation

EDITORIAL

Using Effect Size—or Why the P Value Is Not Enough

GAIL M. SULLIVAN, MD, MPH

RICHARD FEINN, PhD

Statistical significance is the least interesting thing about the results. You should describe the results in terms of measures of magnitude—not just, does a treatment affect people, but how much does it affect them.

-Gene V. Glass¹

The primary product of a research inquiry is one or more measures of effect size, not P values.

-Jacob Cohen²

appears clear, the effect size in the second example is less apparent. Is a 0.4 change a lot or trivial? Accounting for variability in the measured improvement may aid in interpreting the magnitude of the change in the second example.

Thus, effect size can refer to the raw difference between group means, or absolute effect size, as well as standardized measures of effect, which are calculated to transform the effect to an easily understood scale. Absolute effect size is useful when the variables under study have intrinsic

```
> # Effect size
> library(effectsize)
>
> model <- aov(SBP ~ Smoking, data=data3)
> omega_squared(model)
For one-way between subjects designs, partial omega squared is
equivilant to omega squared.
Returning omega squared.
Parameter | Omega2 |         90% CI
-----
Smoking   |   0.51 | [0.41, 0.60]
> interpret_omega_squared(0.51, rules = "field2013")
[1] "large"
(Rules: field2013)
```

What test to use?

Variable 1	Variable 2	Test
Categorical	Categorical	Chi-square
Categorical (2 pop)	Numerical (Normal)	Independent sample t-test
Categorical (2 pop)	Numerical (Not Normal)	Mann-Whitney U test
Categorical (> 2 pop)	Numerical (Normal)	One-way ANOVA
Categorical (> 2 pop)	Numerical (Not Normal)	Kruskal-Wallis test
Numerical (Normal)	Numerical (Normal)	Pearson Correlation Coefficient Test
Numerical (Normal/ Not Normal)	Numerical (Not Normal)	Spearman Correlation Coefficient Test
Numerical (Normal)	Numerical (Normal) - Paired	Paired t-test
Numerical (Not Normal)	Numerical (Not Normal) - Paired	Wilcoxon Signed Rank Test

1. Identify & define variables
2. Type - independent vs. dependent
3. Level of measurements - nominal, ordinal or continuous
4. Check distribution - Normal vs. Not Normal
5. Decide what to do - descriptive vs. analytical

Summary

$$= \lim_{h \rightarrow 0} \frac{x^2 + 2xh + h^2}{h}$$

$$= \lim_{h \rightarrow 0} \frac{2xh + h^2}{h}$$

$$\begin{aligned} & g(x+h) - g(x) \\ & = \lim_{h \rightarrow 0} \frac{h(2x+h)}{h} \end{aligned}$$