FROM SAMPLE TO POPULATION: MAKING INFERENCE

Gbenga Ogunfowokan
Lead, Nigerian Regional Faculty
The Global Health Network
19th May 2017

Objectives

- Know what is Inference
- Know what is parameter estimation
- Understand hypothesis testing & the "types of errors" in decision making.
- Know what the α-level means.
- Learn how to use test statistics to examine hypothesis about population mean, proportion

What is inference

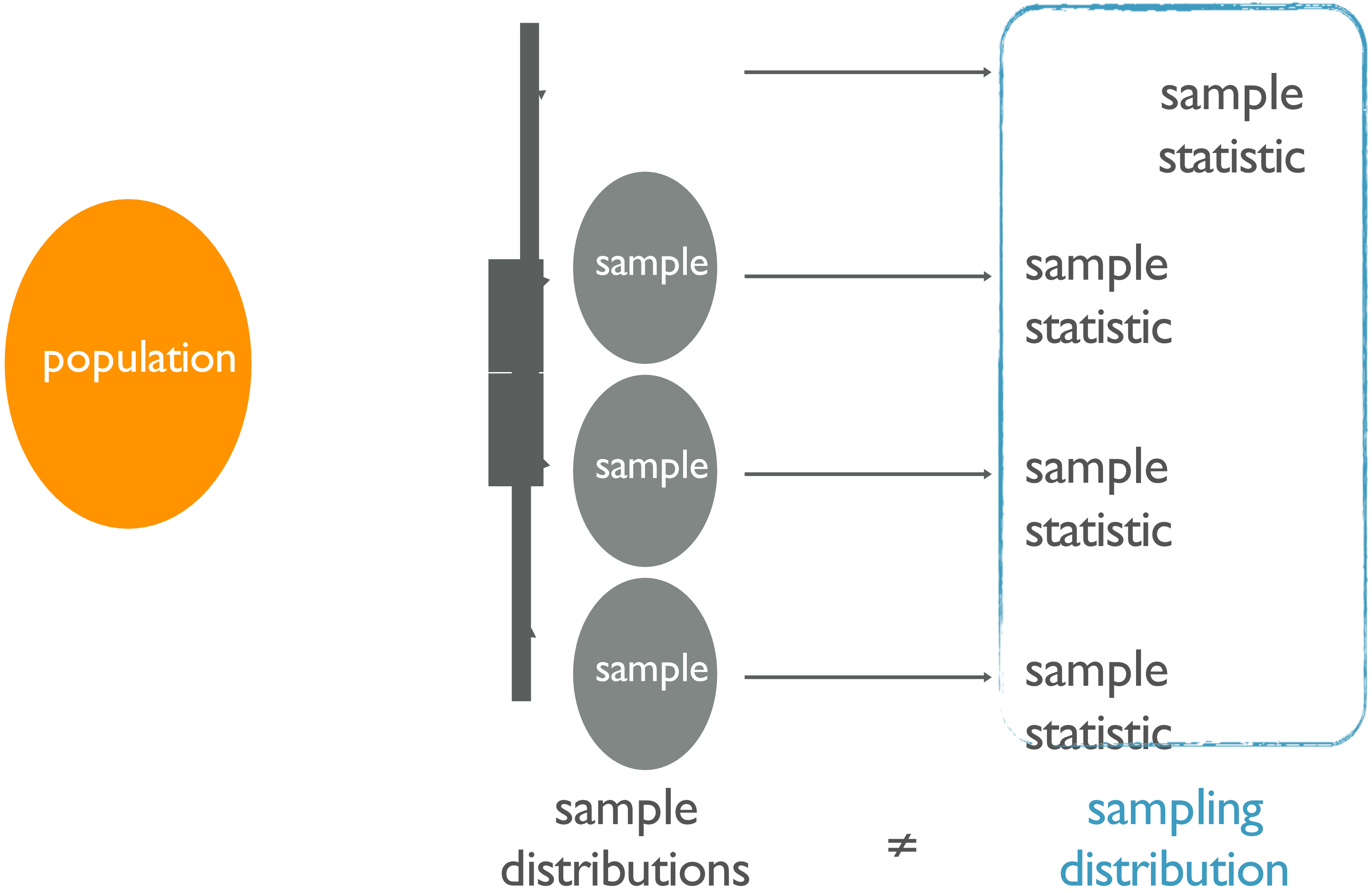
Statistic		Parameter	
Mean:	$\bar{X}$	estimates	_μ___
Standard deviation:	s	estimates	_σ___
Proportion:	p	estimates	_π___

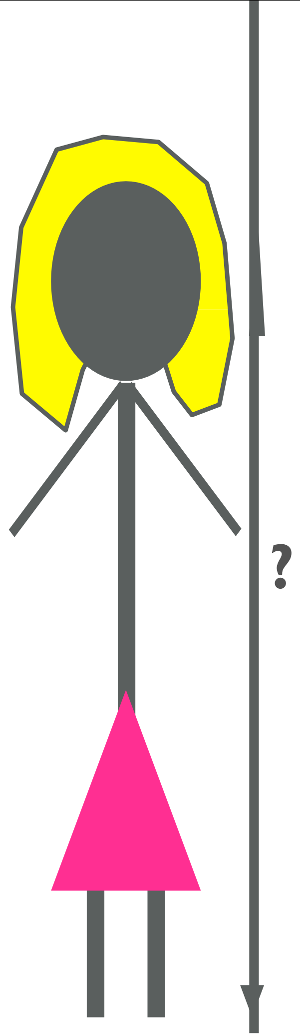
from sample

from entire population

Inference

- **Two ways to make inference**
  - Estimation of parameters
    - * Point Estimation ($\bar{X}$ or p)
    - * Intervals Estimation
  - Hypothesis Testing


population
sample
sample
sample
sample statistic
sample statistic
sample statistic
sample statistic
sample distributions
≠
sampling distribution



Nigerian women
N = pop size
µ

LA $x_{AL,1}, x_{AL,2}, \cdots, x_{AL,1000}$ → $\bar{x}_{AL}$

$\cdots$ $\cdots$

KA: $x_{NC,1}, x_{NC,2}, \cdots, x_{NC,1000}$ → $\bar{x}_{NC}$

$\cdots$ $\cdots$

PLATEAU $x_{WY,1}, x_{WY,2}, \cdots, x_{WY,1000}$ → $\bar{x}_{WY}$

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}$$

sampling distribution

$$mean(\bar{x}) \approx \mu$$

$$SD(\bar{x}) < \sigma$$

n ↑ ↓ standard error

Central Limit Theorem (CLT): The distribution of sample statistics is nearly normal, centered at the population mean, and with a standard deviation equal to the population standard deviation divided by square root of the sample size.

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\cancel{\sigma}\ s}{\sqrt{n}}\right)$$

center

Shape spread

Conditions for the CLT:

1. ***Independence:*** Sampled observations must be independent
   - random sample/assignment
   - if sampling without replacement, n < 10% of population
2. ***Sample size/skew:*** Either the population distribution is normal, or if the population distribution is skewed, the sample size is large (rule of thumb: n > 30).

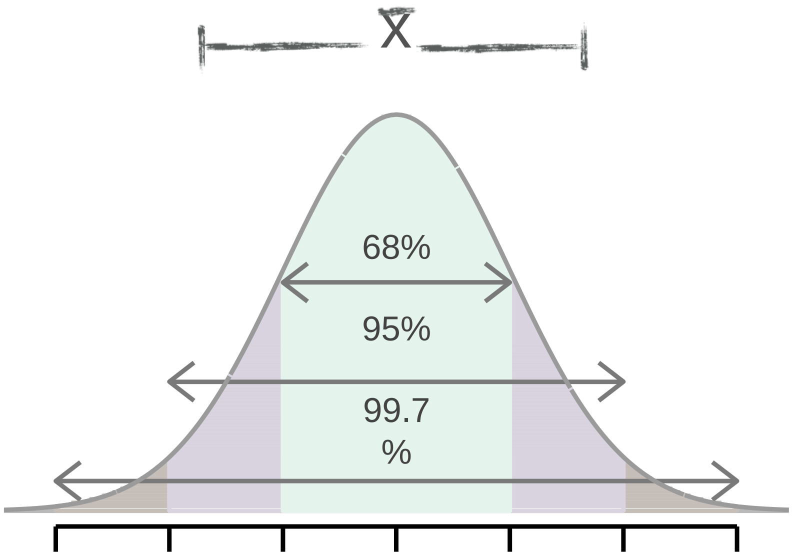
A plausible range of values for the population parameter is called a

confidence interval.

- If we report a point estimate, we probably won't hit the exact population parameter.
- If we report a range of plausible values we have a good shot at capturing the parameter.

Central Limit Theorem (CLT):

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$



approximate 95% CI: $\bar{x} \pm \boxed{2SE}$

margin of error (ME)

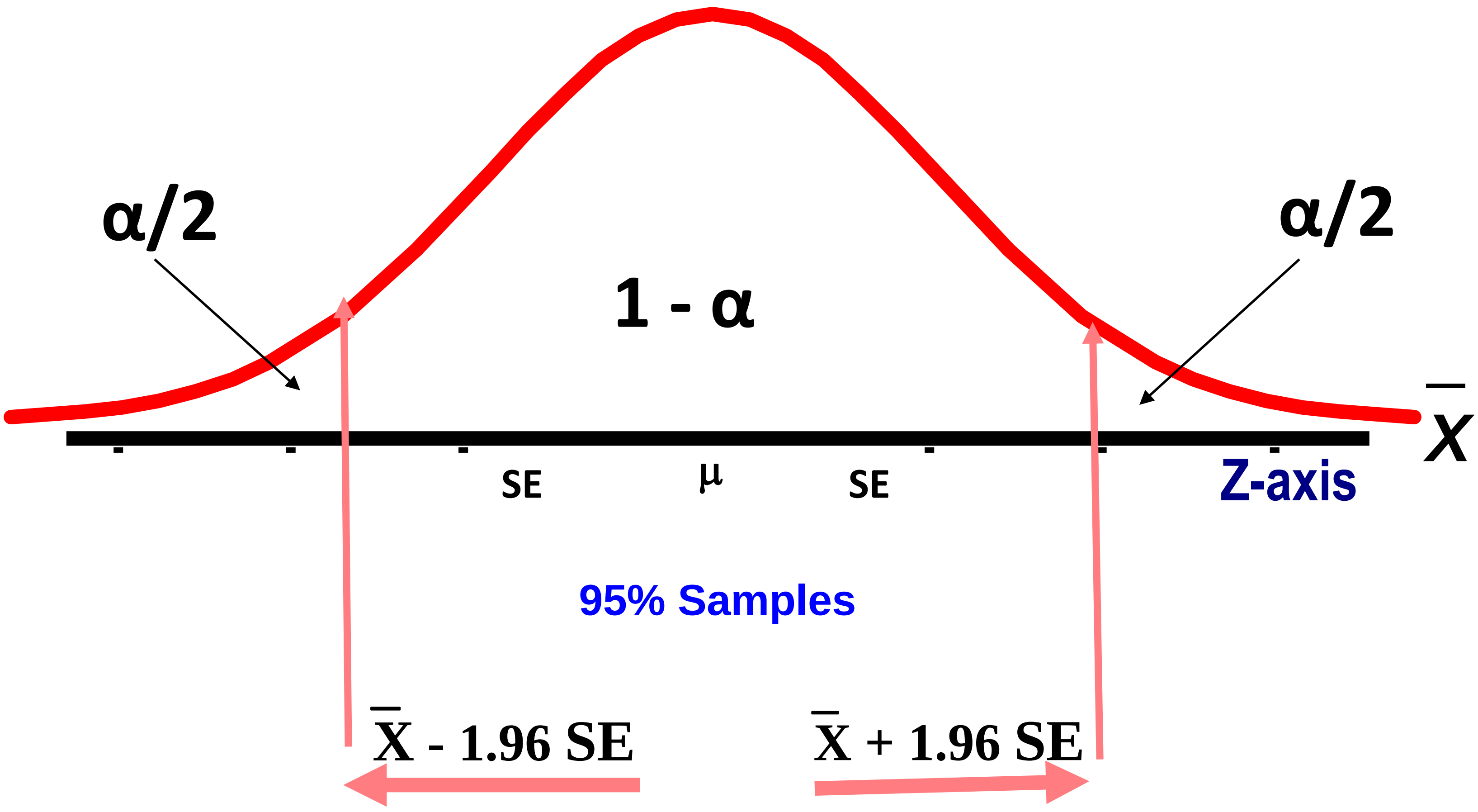
Confidence interval for a population mean: Computed as the sample mean plus/minus a margin of error (critical value corresponding to the middle XX% of the normal distribution times the standard error of the sampling distribution).

$$\bar{x} \pm z^{\star} \frac{s}{\sqrt{n}}$$

Conditions for this confidence interval:

1. ***Independence:*** Sampled observations must be independent.
   - random sample/assignment
   - if sampling without replacement, n < 10% of population
2. ***Sample size/skew:*** $n \geq 30$, larger if the population distribution is very skewed.

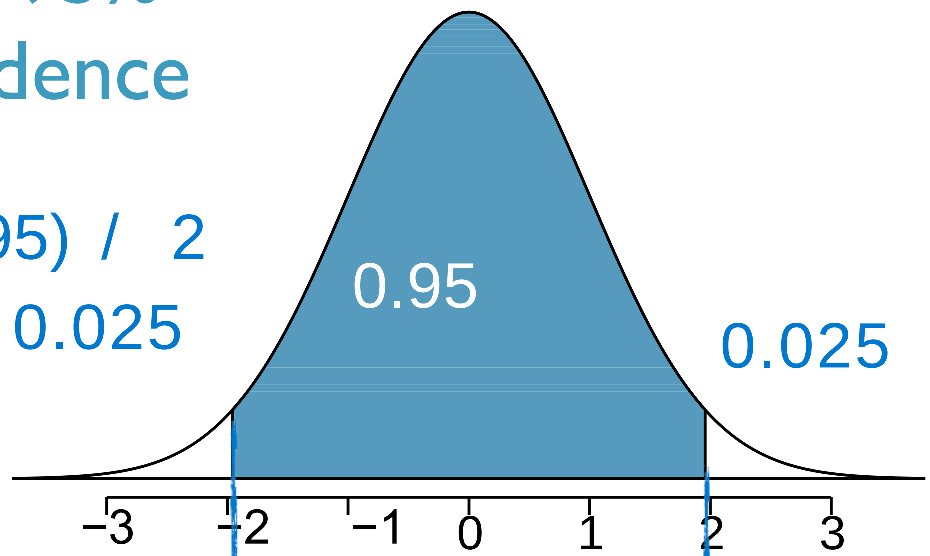
Confidence Interval



finding the critical value 95% confidence

(1 - 0.95) / 2

= 0.025



-1.96

z* = 1.96

R

```
> qnorm(0.025)
[1] -1.96
```

Second decimal place					
0.07	0.06	0.05	0.04	0.00	Z
0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0004	0.0004	0.0004	0.0004	0.0005	−3.3
0.0005	0.0006	0.0006	0.0006	0.0007	−3.2
0.0008	0.0008	0.0008	0.0008	0.0010	−3.1
0.0011	0.0011	0.0011	0.0012	0.0013	−3.0
0.0015	0.0015	0.0016	0.0016	0.0019	−2.9
0.0021	0.0021	0.0022	0.0023	0.0026	−2.8
0.0028	0.0029	0.0030	0.0031	0.0035	−2.7
0.0038	0.0039	0.0040	0.0041	0.0047	−2.6
0.0051	0.0052	0.0054	0.0055	0.0062	−2.5
0.0068	0.0069	0.0071	0.0073	0.0082	−2.4
0.0089	0.0091	0.0094	0.0096	0.0107	−2.3
0.0116	0.0119	0.0122	0.0125	0.0139	−2.2
0.0150	0.0154	0.0158	0.0162	0.0179	−2.1
0.0192	0.0197	0.0202	0.0207	0.0228	−2.0
0.0244	0.0250	0.0256	0.0262	0.0287	−1.9
0.0307	0.0314	0.0322	0.0329	0.0359	−1.8

Confidence Interval

- CI for Mean: given by
  - x ± 1.96*Se(x) for 95% CI where SE(x)=σ/√n
  - X± 1.65*SE(x) for 90% CI
- E.g mean=24.2, Standard deviation 5.9, sample size 100.
  - SE= 5.9/√100 = 0.6
  - 95% CI; 24.2±1.96*0.6; 1.96*0.6=1.2
  - 95% CI; 24.2-1.2 to 24.2+1.2 = 23.0 to 25.4
- Interpretation: we are 95% confident that the mean in the population lies within this interval.

Example (Proportion)

In a survey of 140 asthmatics, 35% had allergy to house dust. Construct the 95% CI for the population proportion.

$$\pi = p \pm Z \sqrt{\frac{P(1-p)}{n}} \qquad SE = \sqrt{\frac{0.35(1-0.35)}{140}} = 0.04$$

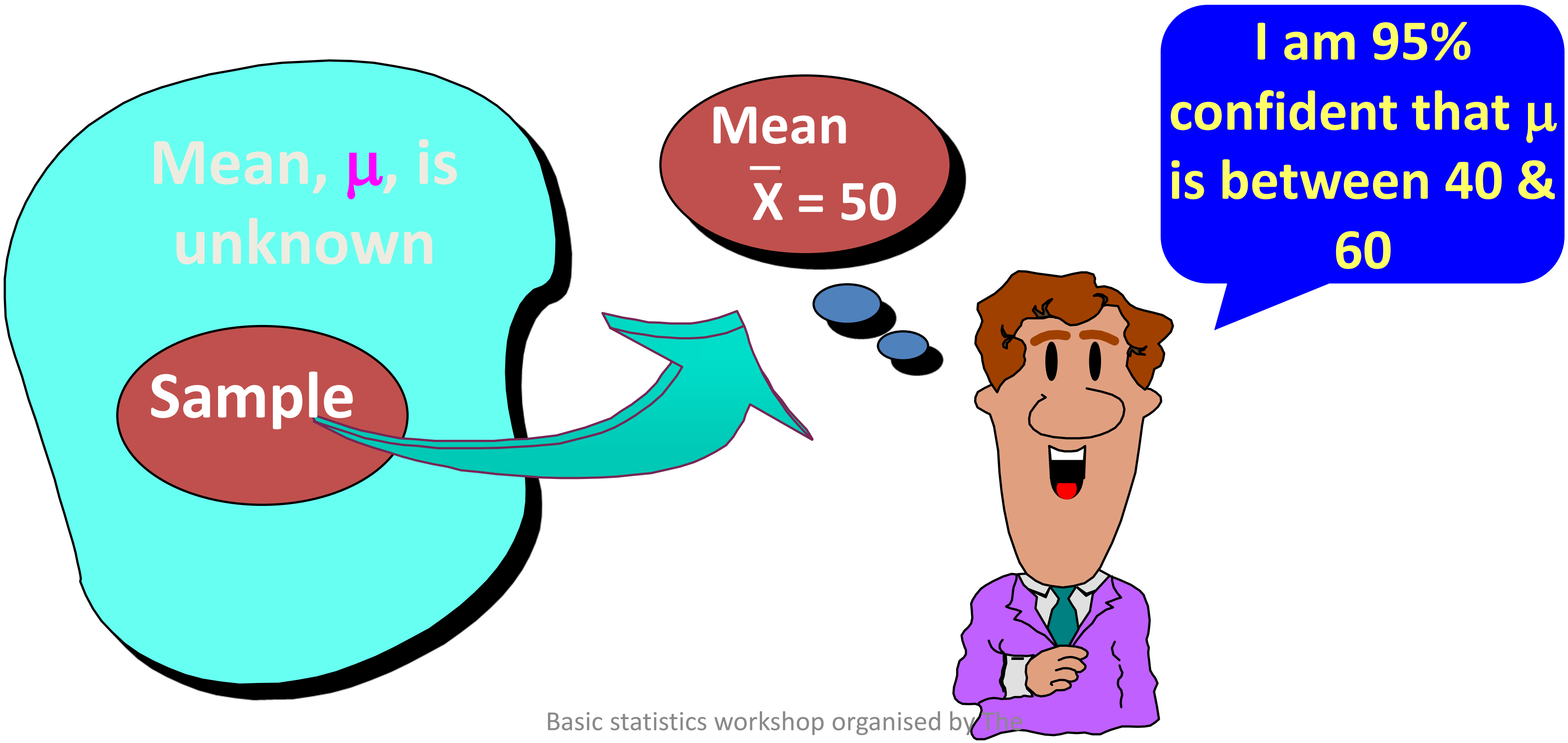
$$0.35 - 1.96 \times 0.04 \leq \pi \geq 0.35 + 1.96 \times 0.04$$

$$0.27 \leq \pi \geq 0.43$$

$$27\% \leq \pi \geq 43\%$$

Estimation of parameters

Population **Point estimate** **Interval estimate**



Interpretation of CI

Probabilistic

In repeated sampling 100(1-α)% of <u>all intervals</u> around sample means will in the long run include μ

Practical

We are 100(1-α)% confident that the <u>single</u> computed CI contains μ

Standard Error

Quantitative Variable

$$SE\ (Mean) = \frac{S}{\sqrt{n}}$$

Qualitative Variable

$$SE\ (p) = \sqrt{\frac{p(1-p)}{n}}$$

Confidence Interval

- CI for proportions
  - P± 1.96*SE(p), for 95 CI where SE(p)= $\sqrt{p(1-p)/n}$
- For small samples, the t distribution is used to estimate CI:
  - Multiplier will be the value of t corresponding to t two-sided p=0.05 with df=n-1
- CI also calculated for RR & OR: estimates that the true association lies within the interval:
  - OR $\exp\{\pm 1.96\sqrt{(1/a + 1/b + 1/c + 1/d)}\}$

Practical

- PRACTICAL DEMONSTRATION WITH EPI INFO SOFTWARE
- 1) Quantitative Variable
- Point Estimate
- Measure of dispersion
- Inference using 95% CI

Practical

- PRACTICAL DEMONSTRATION USING EPI INFO SOFTWARE
- Qualitative variable
- Point estimate using Proportion, Percentage
- Measure of dispersion
- Inference using 95% CI

HYPOTHESIS TESTING

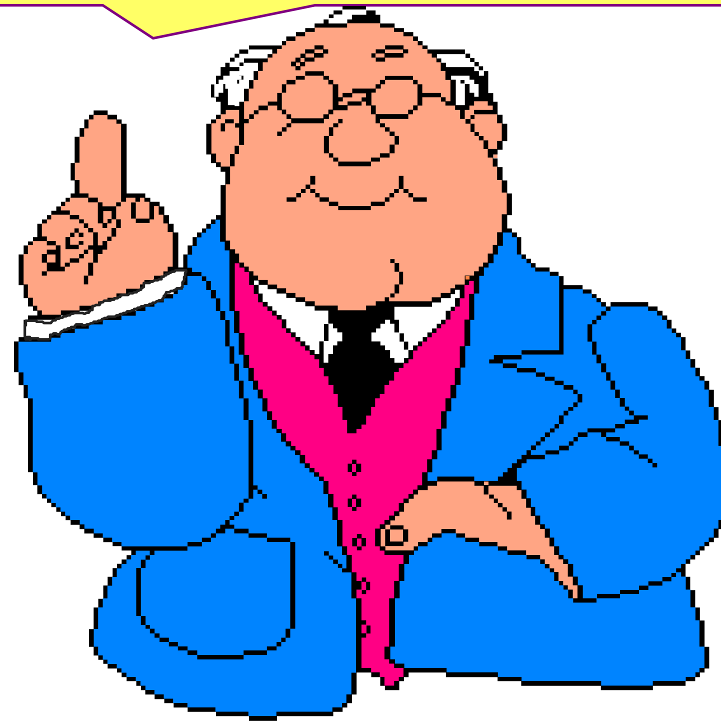
Hypothesis testing

A statistical method that uses sample data to evaluate a hypothesis about a population parameter. It is intended to help researchers differentiate between real and random patterns in the data.

What is a Hypothesis?

I assume the mean SBP of participants is 120 mmHg

An assumption about the population parameter.



agreement of CI and HT

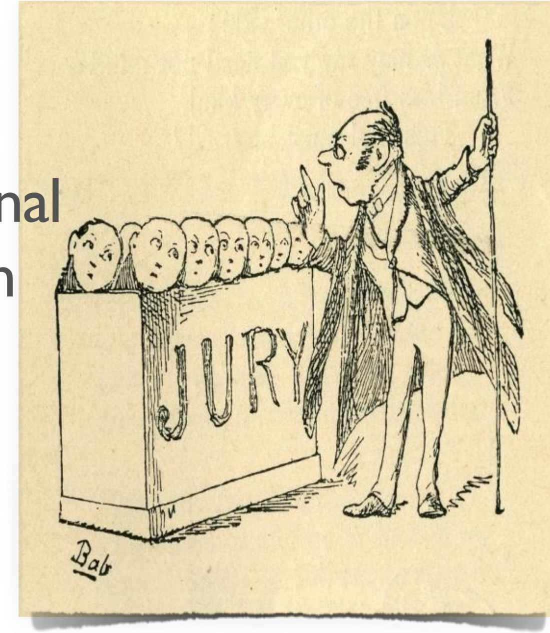
- A two sided hypothesis with threshold of α is equivalent to a confidence interval with CL = 1 − α.
- A one sided hypothesis with threshold of α is equivalent to a confidence interval with CL = 1 − (2 x α).
- If H_0 is rejected, a confidence interval that agrees with the result of the hypothesis test should not include the null value.
- If H_0 is failed to be rejected, a confidence interval that agrees with the result of the hypothesis test should include the null value.

hypothesis test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0 : Defendant is innocent

H_A : Defendant is guilty



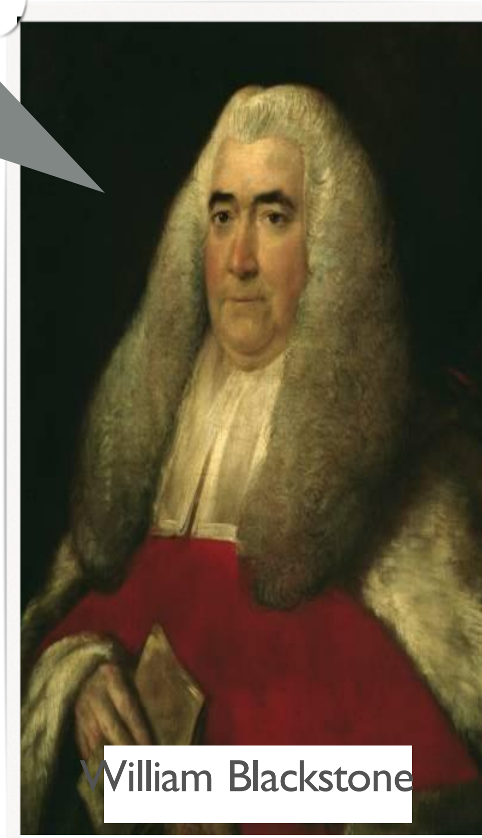
Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty — Type 2 error
- Declaring the defendant guilty when they are actually innocent — Type 1error

"better that ten guilty persons escape than that one innocent suffer"

Which error is the worst error to make?

- Type 2 : Declaring the defendant innocent when they are actually guilty
- Type 1 : Declaring the defendant guilty when they are actually innocent



type I error rate

- We reject H_0 when the p-value is less than 0.05 ($\alpha = 0.05$).
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about 5% chance of making a Type I error if the null hypothesis is true.

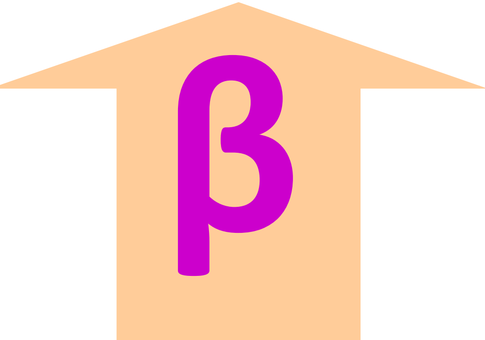
$$P(\text{Type I error} \mid H_0 \text{ true}) = \alpha$$

- This is why we prefer small values of α – increasing α increases the Type I error rate.

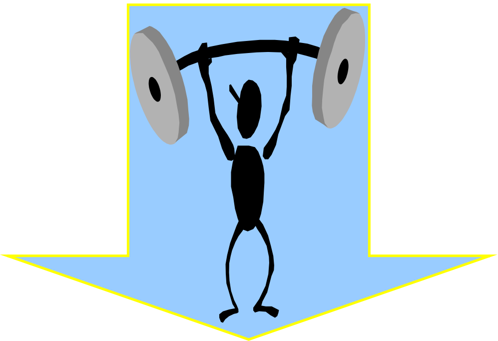
The goal is
o
Keep α
and β low

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1-\alpha$	Type 1 error, α
	H_A true	Type 2 error, β	$1-\beta$

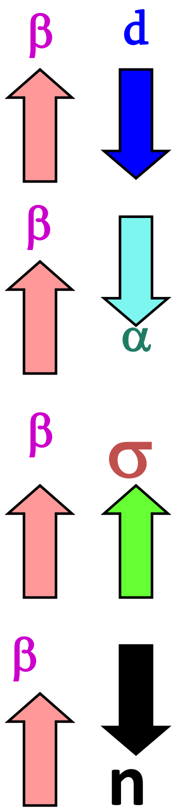
- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level).
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β.
- Power of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1-\beta$



Factors Increasing Type II Error

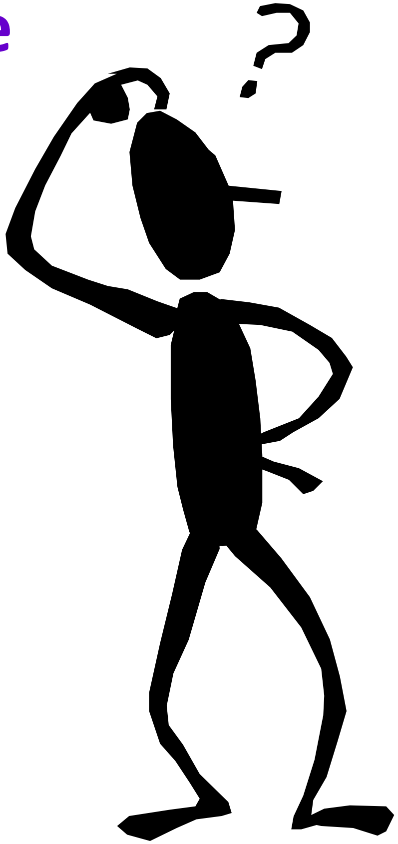


- True Value of Population Parameter
  - Increases When Difference Between Hypothesized Parameter & True Value Decreases
- Significance Level α
  - Increases When α Decreases
- Population Standard Deviation σ
  - Increases When σ Increases
- Sample Size n
  - Increases When n Decreases



p Value Test

- Probability of Obtaining a Test Statistic More Extreme ($\leq$ or $\geq$) than Actual Sample Value Given H_0 Is True
- Called Observed Level of Significance
- Used to Make Rejection Decision
  - If *p* value $\geq \alpha$, Do Not Reject H_0
  - If *p* value $< \alpha$, Reject H_0



Hypothesis Testing: Steps

Test the Assumption that the true mean SBP of participant is 120 mmHg.

State H_0	$H_0: \mu \neq 120$
State H_1	$H_1: \mu = 120$
Choose α	$\alpha = 0.05$
Choose *n*	**n = 100**
Choose Test:	***Z, t, X^2 Test (or p Value)***

Hypothesis Testing: Steps

Compute Test Statistic ***(or compute P value)***

Search for Critical Value

Make Statistical Decision rule

Express Decision

One sample-mean Test

- **Assumptions**
  - **Population is normally distributed**

- **t test statistic**

$$t = \frac{\text{sample mean} - \text{null value}}{\text{standard error}} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

Example Normal Body Temperature

What is **normal body temperature**? Is it actually 37.6^{o}C (on average)?

State the null and alternative hypotheses

$$H_0: \mu = 37.6^{o}C$$

$$H_a: \mu \neq 37.6^{o}C$$

Example Normal Body Temp (cont)

Data: random sample of n = 18 normal body temps

37.2	**36.8**	**38.0**	**37.6**	**37.2**	**36.8**	**37.4**	**38.7**	**37.2**
36.4	**36.6**	**37.4**	**37.0**	**38.2**	**37.6**	**36.1**	**36.2**	**37.5**

Summarize data with a test statistic

Variable	n	Mean	SD	SE	t	P
Temperature	18	37.22	0.68	0.161	2.38	0.029

$$t = \frac{\text{sample mean} - \text{null value}}{\text{standard error}} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

STUDENT'S t DISTRIBUTION TABLE

Degrees of freedom	Probability (p value)		
	0.10	0.05	0.01
1	6.314	12.706	63.657
5	2.015	2.571	4.032
10	1.813	2.228	3.169
17	1.740	2.110	2.898
20	1.725	2.086	2.845
24	1.711	2.064	2.797
25	1.708	2.060	2.787
∞	1.645	1.960	2.576

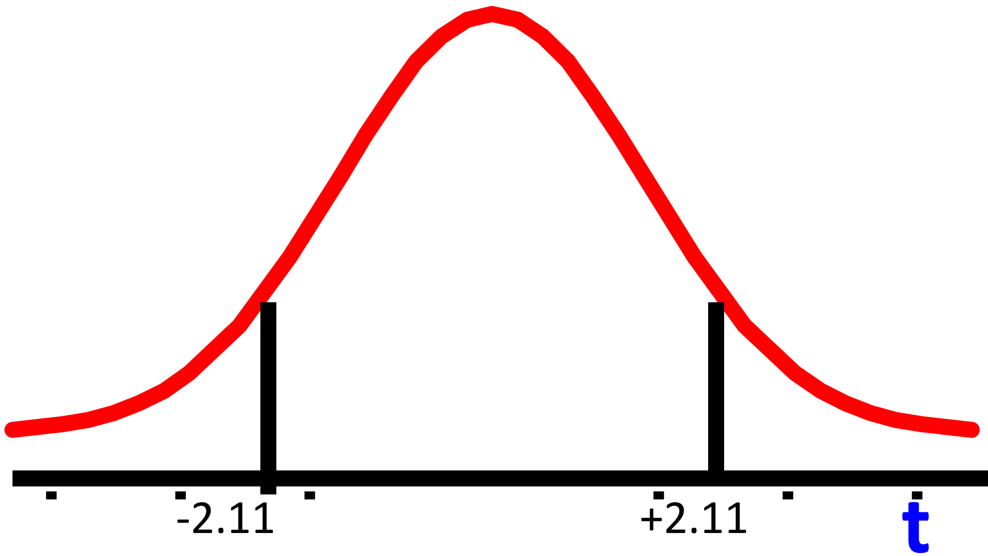
Example Normal Body Temp (cont)

Find the *p*-value

Df = n – 1 = 18 – 1 = 17

From SPSS: *p*-value = 0.029

From t Table: *p*-value is between 0.05 and 0.01.



Area to left of *t* = -2.11 equals area to right of *t* = +2.11.

The value *t* = 2.38 is between column headings 2.110& 2.898 in table, and for df =17, the *p*-values are 0.05 and 0.01.

Example Normal Body Temp (cont)

Decide whether or not the result is statistically significant based on the *p*-value

Using $\alpha = 0.05$ as the level of significance criterion, the results are **statistically significant** because 0.029 is less than 0.05. In other words, we can reject the null hypothesis.

Report the Conclusion

We can conclude, based on these data, that the mean temperature in the human population does not equal 37.6.

One-sample test for proportion

- Involves categorical variables
- Fraction or % of population in a category
- Sample proportion (*p*)

$$p = \frac{X}{n} = \frac{number\ of\ successes}{sample\ size}$$

- Test is called Z test
  where:
- Z is computed value
- π is proportion in population
  (null hypothesis value)

$$Z = \frac{p - \pi}{\sqrt{\dfrac{\pi(1-\pi)}{n}}}$$

Critical Values: 1.96 at α=0.05

2.58 at α=0.01

Example

- In a survey of diabetics in a large city, it was found that 100 out of 400 have diabetic foot. Can we conclude that 20 percent of diabetics in the sampled population have diabetic foot.
- Test at the α =0.05 significance level.

Solution

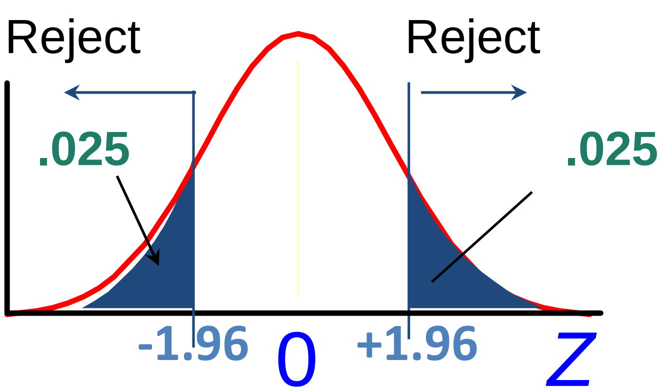
H_o: $\pi = 0.20$

H_1: $\pi \neq 0.20$

$$Z = \frac{0.25 - 0.20}{\sqrt{\dfrac{0.20\,(1-0.20)}{400}}}$$

= 2.50

Critical Value: 1.96

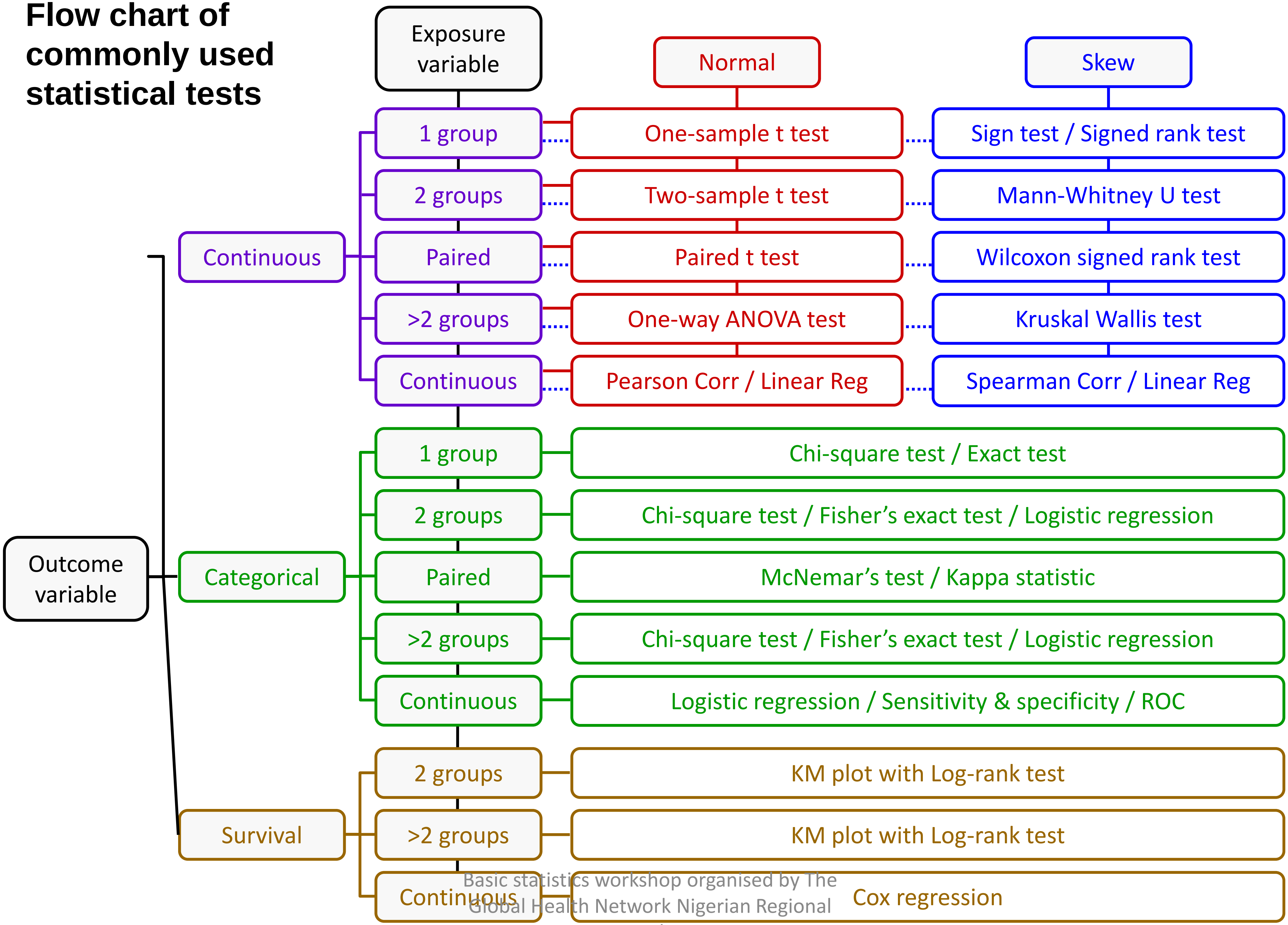


Decision:

We have sufficient evidence to reject the Ho value of 20%

We conclude that in the population of diabetic the proportion who have diabetic foot does not equal 0.20

Flow chart of commonly used statistical tests



Which Statistical Test?

Use the table to obtain informaton on how to carry out the test in SPSS and how to report and present the results.

Move the cursor over the boxes that classify the tests for further details. Click on the statistical tests for more details.

Number of groups / Exposure (Independent) variable	Outcome(Dependent) variable			
	Continuous and Normally distributed (Parametric)	Continuous and skewed / Ordinal (Non-parametric)	Binary (2 categories)	Survival Time to event
1 group	One-sample t test	Sign test / Signed rank test	Chi-square test / Exact test	Life tables analysis
2 independent groups	Two-sample t test Linear regression	Mann-Whitney U test	Chi-square test / Fisher's Exact Logistic regression	Log-rank test Cox regression
Paired (related) sample (2 time points)	Paired t test Bland-Altman method	Wilcoxon signed rank test	McNemar's Test Kappa statistic	*Not covered*
>2 independent groups	One-way ANOVA test Linear regression	Kruskal-Wallis test	Chi-square test / Fisher's Exact Test Logistic regression	Log-rank test Cox regression
>2 related samples (>2 time points)	Repeated measures ANOVA	Friedman's Test	*Not covered*	*Not covered*
Continuous	Pearson's correlation Linear Regression	Spearman's rank correlation Linear regression	Logistic regression	Cox regression
Epidemiological data			Sensitivity & specificity PPV & NPV ROC	

UNIVERSITY OF Southampton

Comparing a binary outcome between two groups – data presented as a 2x2 table

	Unfit	Fit	Total
Standard	80 (a)	140 (b)	220 (a+b)
Enhanced	20 (c)	220 (d)	240 (c+d)

Chi-square test and Fisher's exact test show if there is any **association** between the two independent variables, but it doesn't provide the **effect size** between the groups regarding the outcome of interest, e.g. Fit

- **Table shows results from our trial (number of patients)**
- Difference in proportion of Fit between groups (absolute difference):

$$d/(c+d) - b/(a+b)$$

- An alternative parameter is the relative risk (multiplicative difference):

$$\frac{d/(c+d)}{b/(a+b)}$$

- Another alternative is the odds ratio:

$$\frac{d/c}{b/a} = \frac{a \times d}{b \times c}$$

Percentage of Fit in standard group: 140/220 (63.6%)
Percentage of Fit in enhanced group: 220/240 (91.7%)

	Parameter (95% CI)
Absolute difference in proportions $d/(c+d) - b/(a+b)$	28.1% (21%, 35%)*
Relative risk $\dfrac{d/(c+d)}{c/(a+b)}$	1.44 (1.29, 1.60)†
Odds ratio $\dfrac{a \times d}{b \times c}$	6.29 (3.69, 10.72)†

* Asymptotic 95% confidence intervals (calculated in CIA)
† 95% confidence intervals calculated in SPSS

- **Reminder:** Report confidence intervals for <u>ALL</u> key parameter estimates
  - If 95% confidence interval for a difference excludes 0 ➔ statistically significant e.g. Absolute difference
  - If 95% confidence interval for a ratio excludes 1 ➔ statistically significant e.g. Relative risk and Odds ratio

Advantages and disadvantages of absolute and relative changes, and odds ratios

Absolute difference	• simplest to calculate and to interpret • when applied to number of subjects in a group gives number of subjects expected to benefit • 1/(absolute difference) gives NNT – ‘number needed to treat’ to see one additional positive response
Relative risk	• intuitively appealing • a multiplicative effect – proportion (risk) of failure in the treatment group examined relative to (or compare to) that in the reference group • different result depending on whether risks of ‘Fit’ or ‘Unfit’ are examined and whether ‘Standard exercise’ group is selected as the reference level • natural parameter for cohort studies
Odds ratio	• difficult to understand – unless you’re a betting person! • ratio of ‘number of successes expected per number of failures’ between the treatment group of interest and the reference group • invariant to whether rate of ‘Fit’, ‘Unfit’, or rate of taking ‘Enhanced exercise’ are examined • logistic regression in terms of odds ratios • natural parameter for case-control studies

- PRACTICAL DEMONSTRATION USING EPI INFO AND EXCEL

Practicals

- RESEATCH TITLE
- RESEARCH OBJECTIVES
- DESCRIPTIVE STSTISTICS
- INFERENTIAL STATISTICS

CORBIS/Brian Leng (05065)