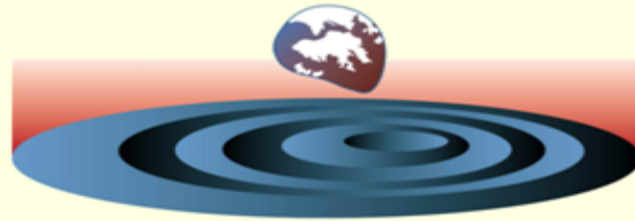


Area of Excellence (AoE) Centre for
Marine Environmental Research & Innovative Technology (MERIT)
and School of Biological Sciences (SBS), The University of Hong Kong (HKU)

Postgraduate Research Training Workshop 2011



Basic Statistics for Research

Kenneth M. Y. Leung

**Why do we
need statistics?**

Statistics

- Derived from the Latin for “state” - governmental data collection and analysis.
- Study of data (branch of mathematics dealing with numerical facts i.e. data).
- The analysis and interpretation of data with a view toward objective evaluation of the reliability of the conclusions based on the data.
- Three major types: **Descriptive, Inferential and Predictive Statistics**

Variation - Why statistical methods are needed

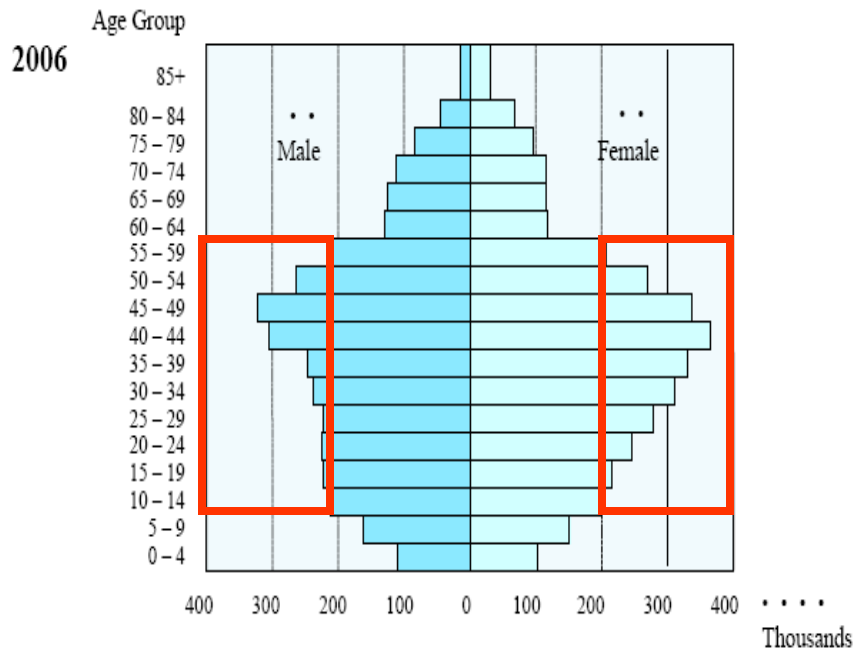
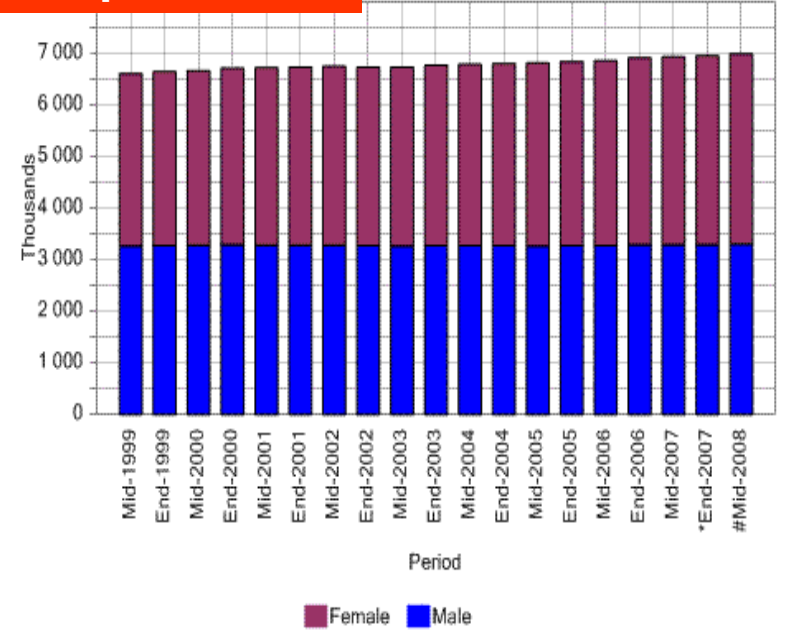
<http://www.youtube.com/watch?v=fsRYkRqQqgg&feature=related>

By UCMSCI

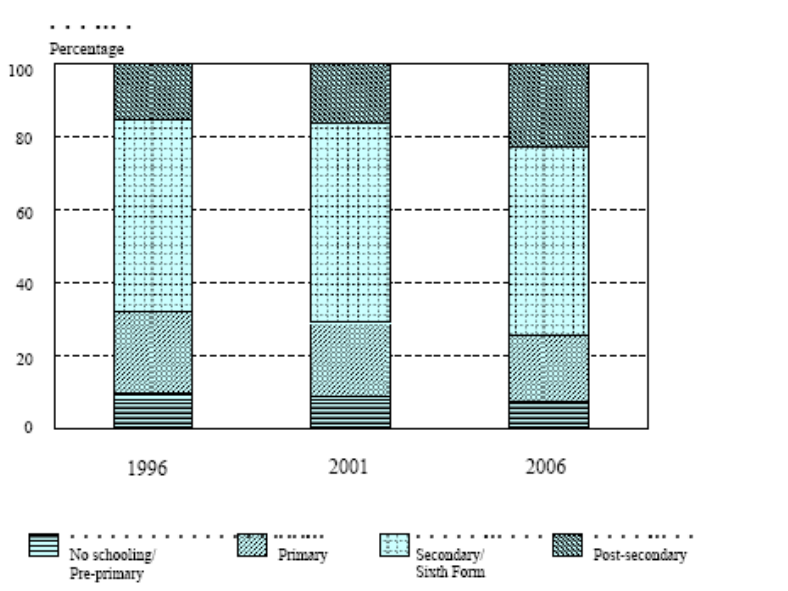
3 Major Types of Stats

- **Descriptive statistics** (i.e., data distribution – central tendency and data dispersion)
- **Inferential statistics** (i.e., hypothesis testing)
- **Predictive statistics** (i.e., modelling)

Descriptive Stats



Distribution of Population Aged 15 and Over by Educational Attainment (Highest Level Attended), 1996, 2001 and 2006



2.3 In terms of the proportion of older persons in the total population, its percentage rose continuously over the past 45 years from 2.8% in 1961 to 12.4% in 2006. The elderly dependency ratio, defined as the number of persons aged 65 and over per 1 000 persons aged between 15 and 64, increased from 50 in 1961 to 168 in 2006.

2.4 The sex ratio (i.e. number of males per 1 000 females) of the older persons in 2006 was 856, as compared with the overall sex ratio of 911. There were more female older persons than male older persons, mainly because of higher life expectancy for female.

From observation to scientific questioning:

Why do females generally live longer than males in human and other mammals?

Setting hypothesis (theory) for testing:

Hypothesis: Metabolic rate of males is faster than that of females, leading to shorter life span in males.

Hypothesis: Males consume more food than females, leading to a higher chance of exposure to toxic substances.

A Hypothesis

- A statement relating to an observation that may be true but for which a proof (or disproof) has not been found.
- The results of a well-designed experiment may lead to the proof or disproof of a hypothesis (i.e. accept or reject of the corresponding null hypothesis).

For example, Heights of male vs. female at age of 30.

Our observations: male $H >$ female H ; it may be linked to genetics, consumption and exercise etc.

Is that true for the hypothesis (H_A): **male $H >$ female H** ?

A corresponding Null hypothesis (H_0): **male $H \leq$ female H**

Scenario I: Randomly select 1 person from each sex.

Male: 170

Female: 175

Then, Female $H >$ Male H . Why?

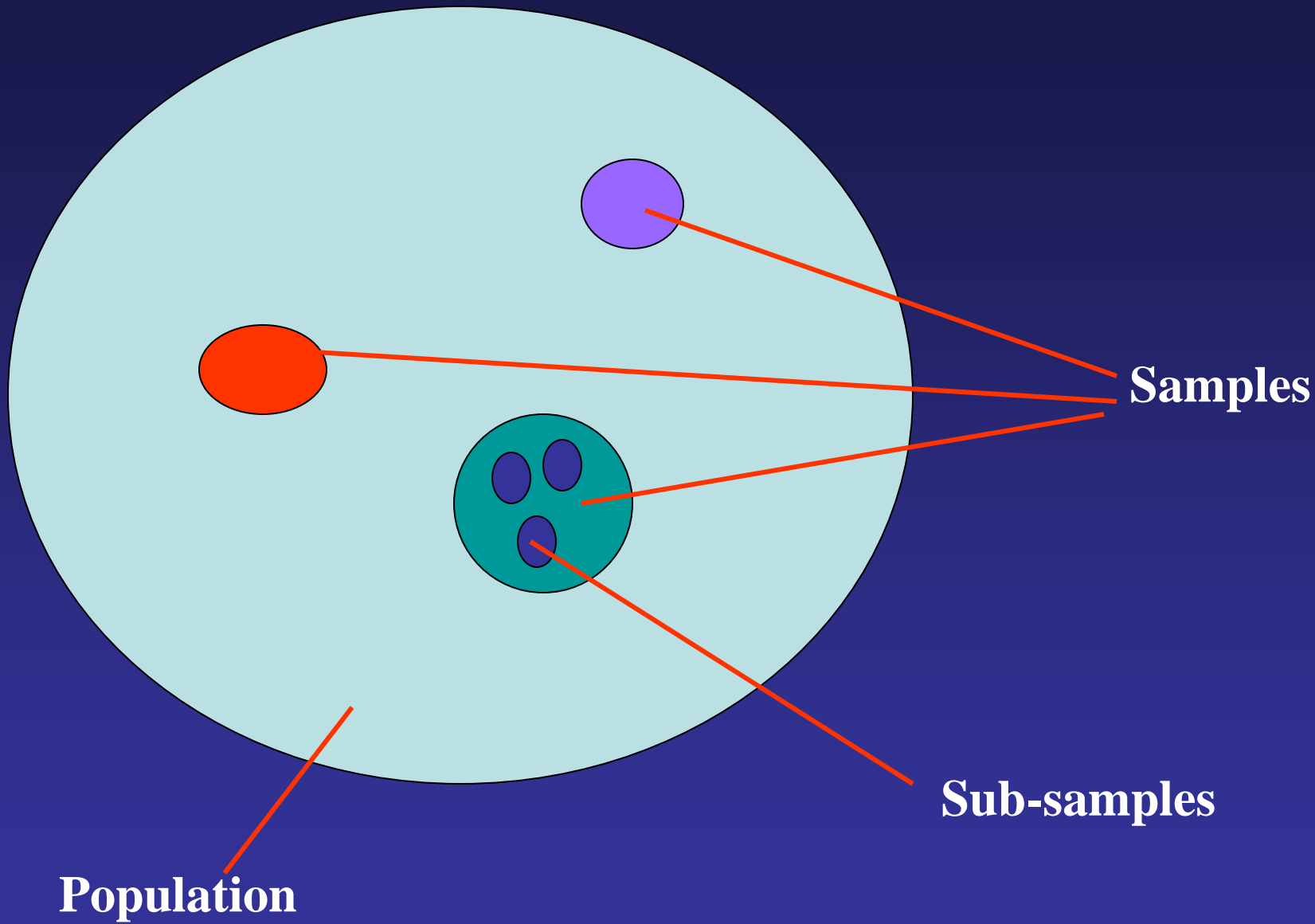
Scenario II: Randomly select 3 persons from each sex.

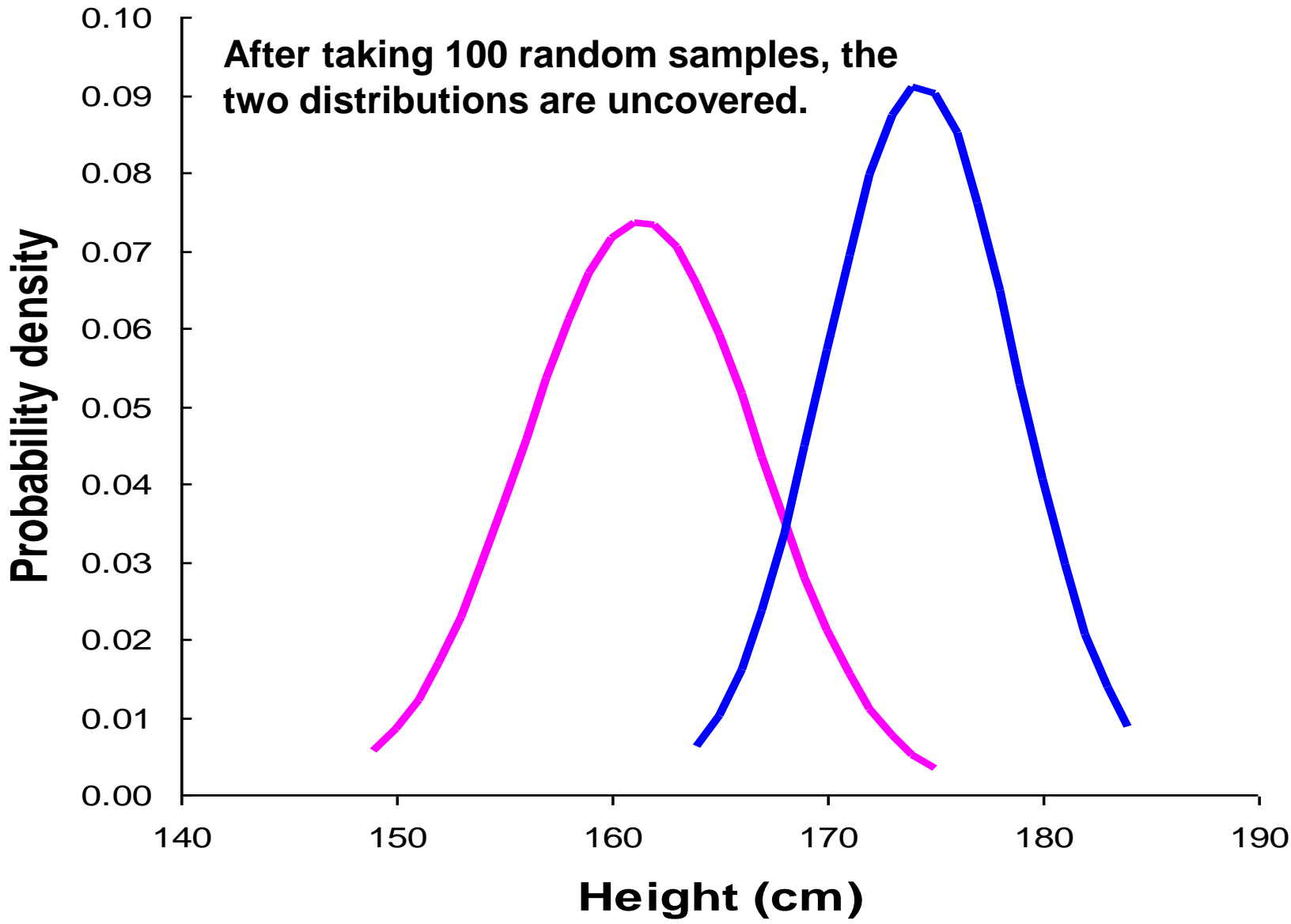
Male: 171, 163, 168

Female: 160, 172, 173

What is your conclusion then?

Inferential Stats



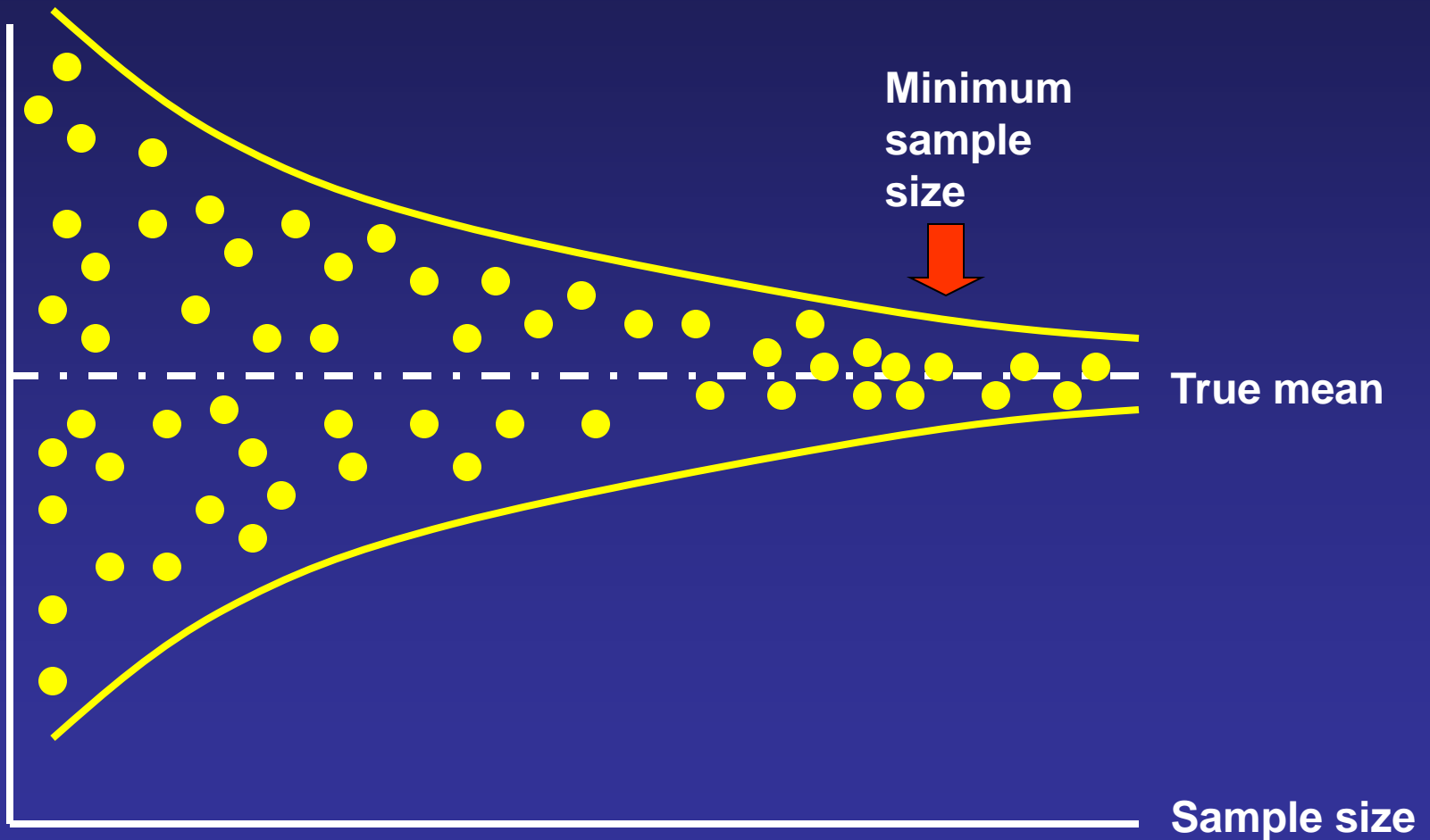


Important Take-Home Messages:

- (1) Sample size is very important and will affect your conclusion.
- (2) Measurement results vary among samples (or subjects) – that is “variation” or “uncertainty”.
- (3) Variation can be due to measurement errors (random or systematic errors) and variation inherent within samples; e.g., at age 30, female height varies between 148 and 189 cm. **Why?**
- (4) Therefore, we always deal with distributions of data rather than a single point of measurement or event.

How many samples are needed?

Mean values



Minimum
sample
size

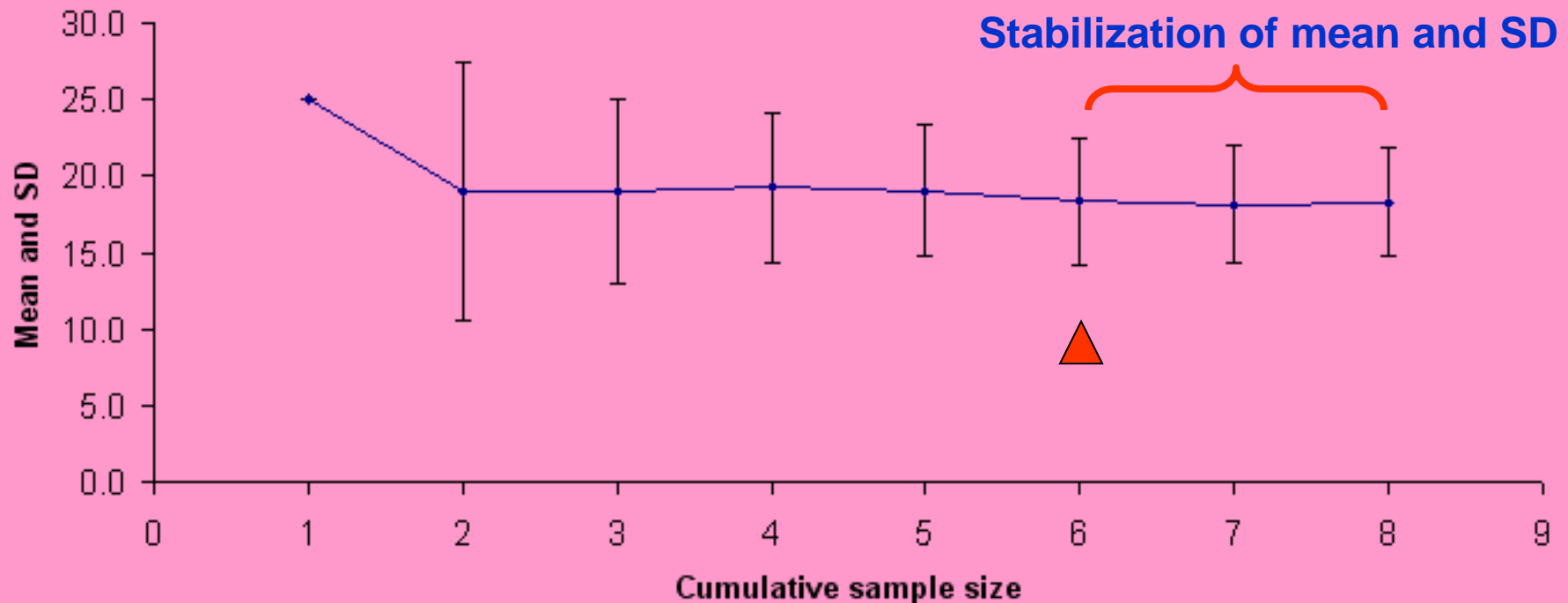
True mean

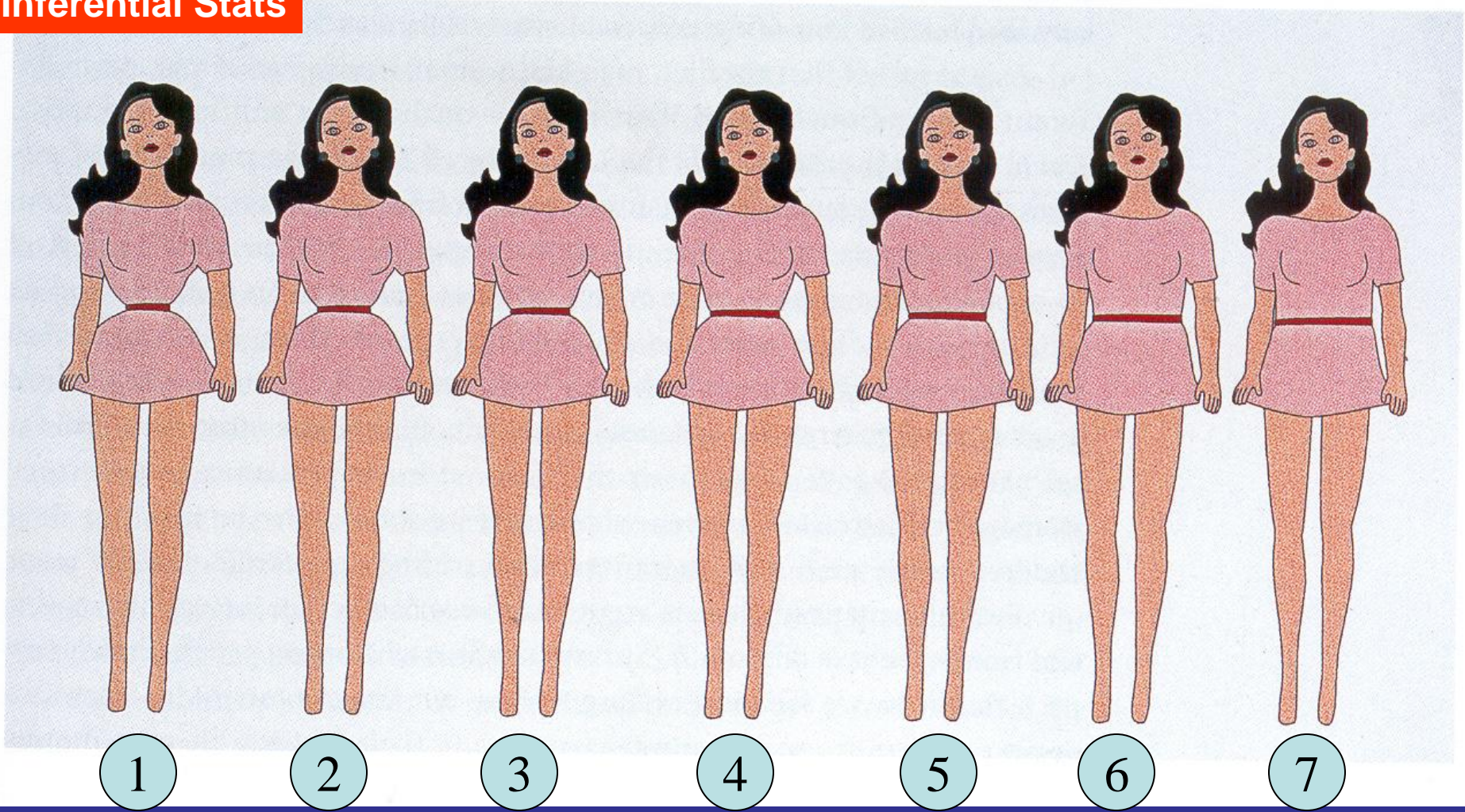
Sample size

*Assuming data follow the normal distribution

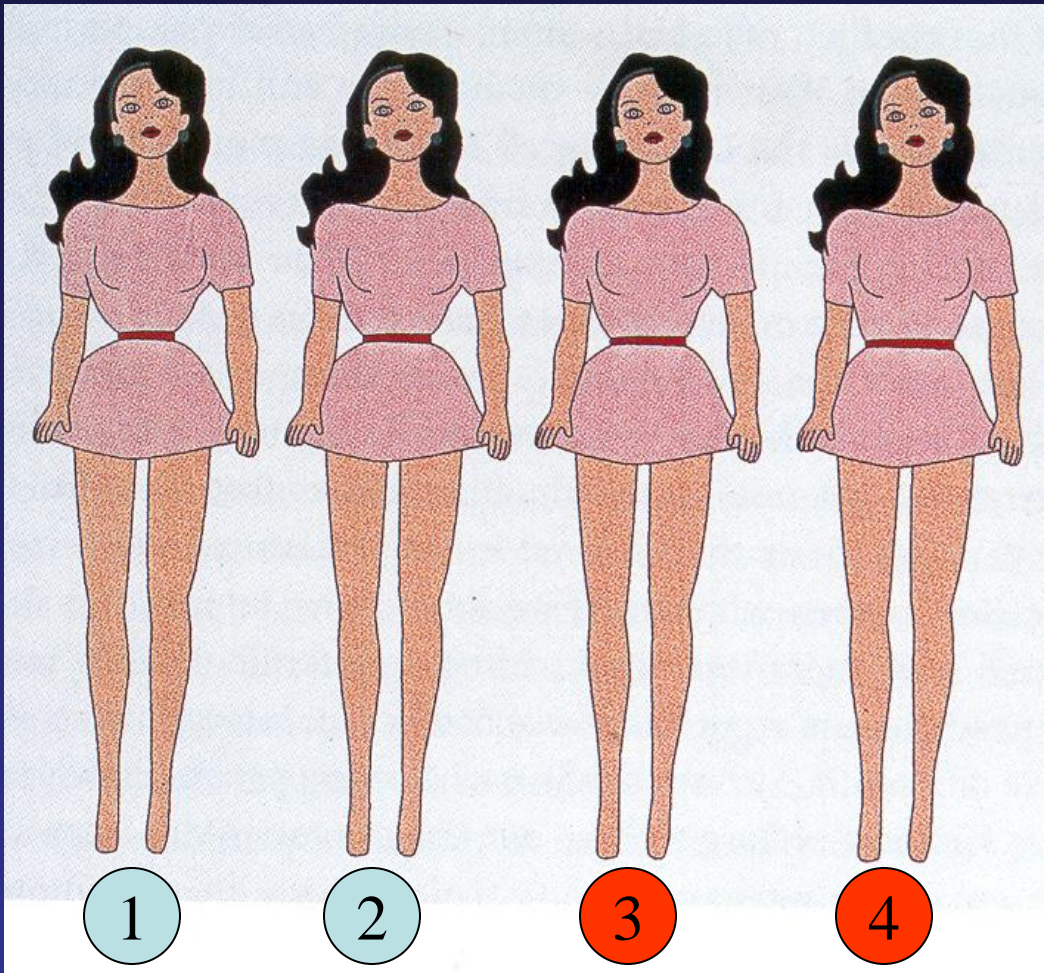
Determine the minimum sample size by plotting the running means

Random sample (<i>E. coli</i> count per 100 ml)										
N	1	2	3	4	5	6	7	8	Mean	SD
1	25.0								25.0	ND
2	25.0	13.0							19.0	8.5
3	25.0	13.0	19.0						19.0	6.0
4	25.0	13.0	19.0	20.0					19.3	4.9
5	25.0	13.0	19.0	20.0	18.0				19.0	4.3
6	25.0	13.0	19.0	20.0	18.0	15.0			18.3	4.2
7	25.0	13.0	19.0	20.0	18.0	15.0	17.0		18.1	3.8
8	25.0	13.0	19.0	20.0	18.0	15.0	17.0	19.5	18.3	3.6





Which one do you prefer?



Men tend to find a hip-to-waist ratio between 60 and 70 percent to be most attractive in women (shown here as the third and fourth from the left). Is this a cultural fluke or an adaptation for finding good mates that evolved in the brains of hominids a million years ago?

We can infer if the observed “preference” frequencies are identical to the hypothetical “preference” frequencies (e.g. 1:2:10:11:3:2:1) using a Chi-square test.

$$Chi-square = \sum (O_i - E_i)^2 / E_i$$

How can we test the following hypotheses?

Ho 1: The water sample A is cleaner than the water sample B in terms of *E. coli* count.

Ho 2: Water quality in Site A is better than Site B in terms of *E. coli* count during the swimming season.

Ho 3: Water quality in Site A is better than Site B in terms of *E. coli* count at all times.

Healthy life expectancy in Hong Kong Special Administrative Region of China

Bulletin of the World Health Organization 2003, 81 (1)

C.K. Law^{1,2} & P.S.F. Yip^{1,2,3}

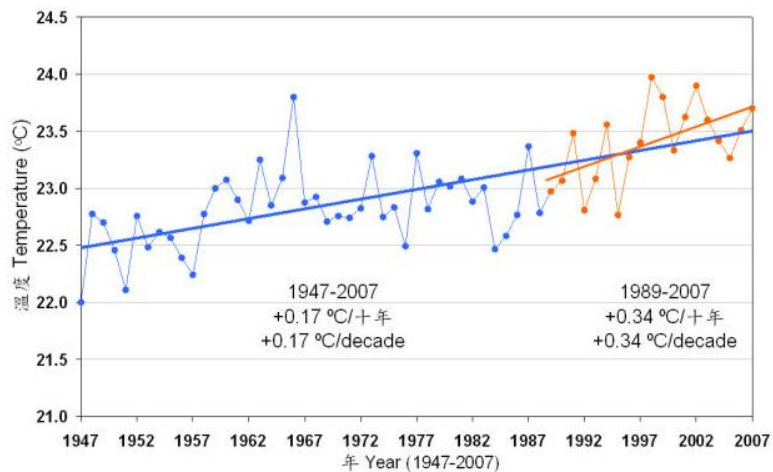
Table 2. Relative ranking of healthy life expectancy (HALE) in Hong Kong Special Administrative Region (SAR) of China at birth, by sex, relative to the top 30 of 191 WHO Member States in 2000^a

Male			Female		
Rank	Country	HALE	Rank	Country	HALE
1	Japan	71.2	1	Japan	76.3
2	Switzerland	70.4		Hong Kong SAR^b	75.7
	Hong Kong SAR^b	70.3	2	San Marino	74.3
3	Sweden	70.1	3	Monaco	73.9
4	Andorra	69.8	4	Andorra	73.7
5	Iceland	69.8	5	Switzerland	73.7
6	San Marino	69.7	6	Australia	73.3
7	Greece	69.7	7	France	72.9
8	Australia	69.6	8	Italy	72.8
9	Italy	69.5	9	Sweden	72.7
10	New Zealand	69.5	10	Iceland	72.6
11	Monaco	69.4	11	Spain	72.5
12	Israel	69.3	12	Austria	72.5
	Hong Kong SAR^c	69.3		Hong Kong SAR^c	72.4
13	Denmark	68.9	13	Norway	72.3
14	Norway	68.8	14	Greece	72.3
15	Malta	68.7	15	New Zealand	72.1
16	Spain	68.7	16	Malta	72.1
17	France	68.5	17	Luxembourg	72
18	Canada	68.3	18	Canada	71.7
19	United Kingdom	68.3	19	Finland	71.5
20	Netherlands	68.2	20	Germany	71.5

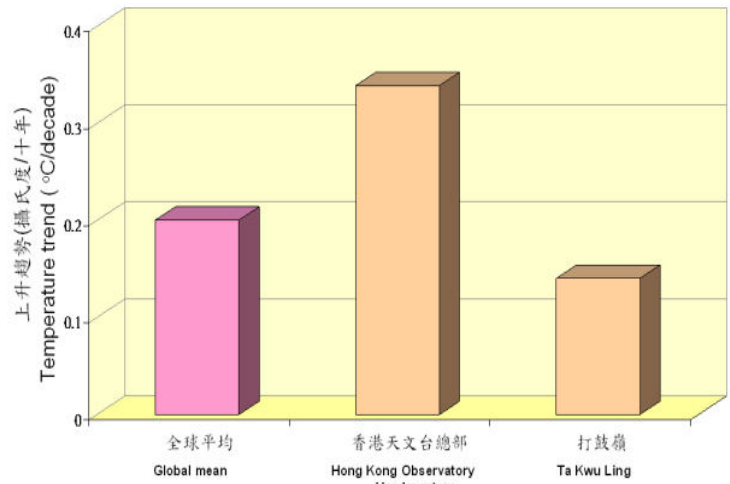
b: Sullivan's method

c: A regression model

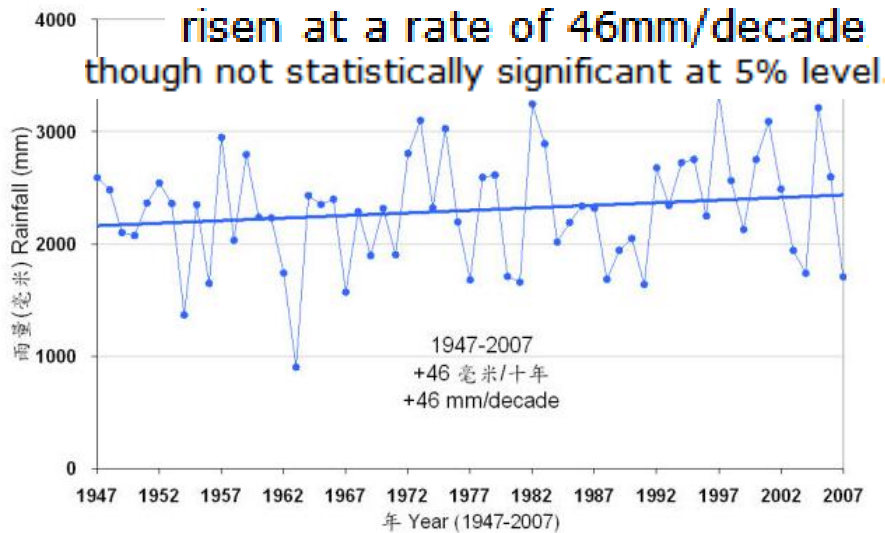
Predictive Stats



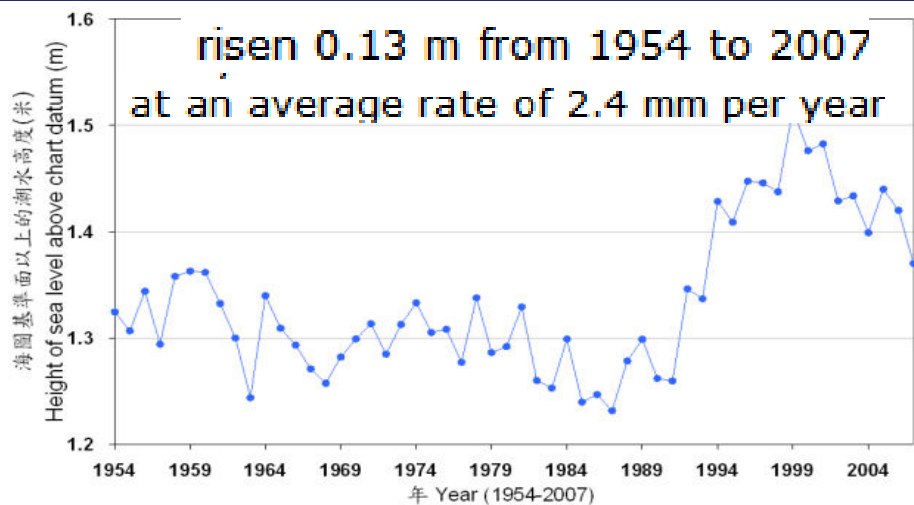
Annual mean temperature recorded at the Hong Kong Observatory Headquarters (1947-2007)



Comparison of recent trends in annual mean temperature in Hong Kong (1989-2007)



Annual rainfall at the Hong Kong Observatory Headquarters (1947-2007)



Annual mean sea level at North Point/Quarry Bay (1954-2007)

Basic Descriptive Statistics

Measurement Theory

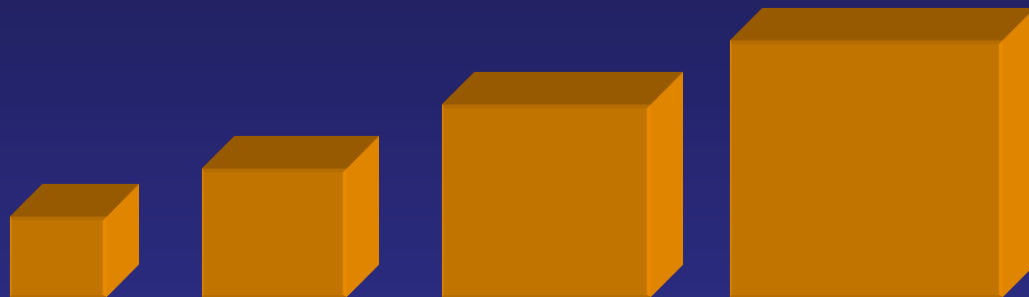
- Environmental scientists use measurements routinely in Lab or field work by assigning numbers or groups (classes).
- Mathematical operations may be applied to the data, e.g. predicting fish mass by their length through an established regression
- **Different levels of measurements:**
 - nominal, ordinal, interval scale, ratio

1



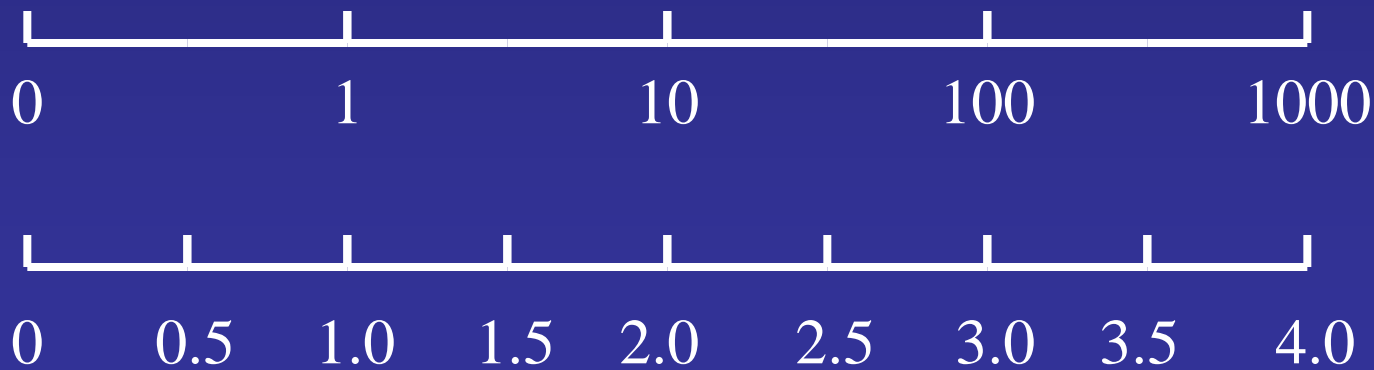
Nominal

2



Ordinal

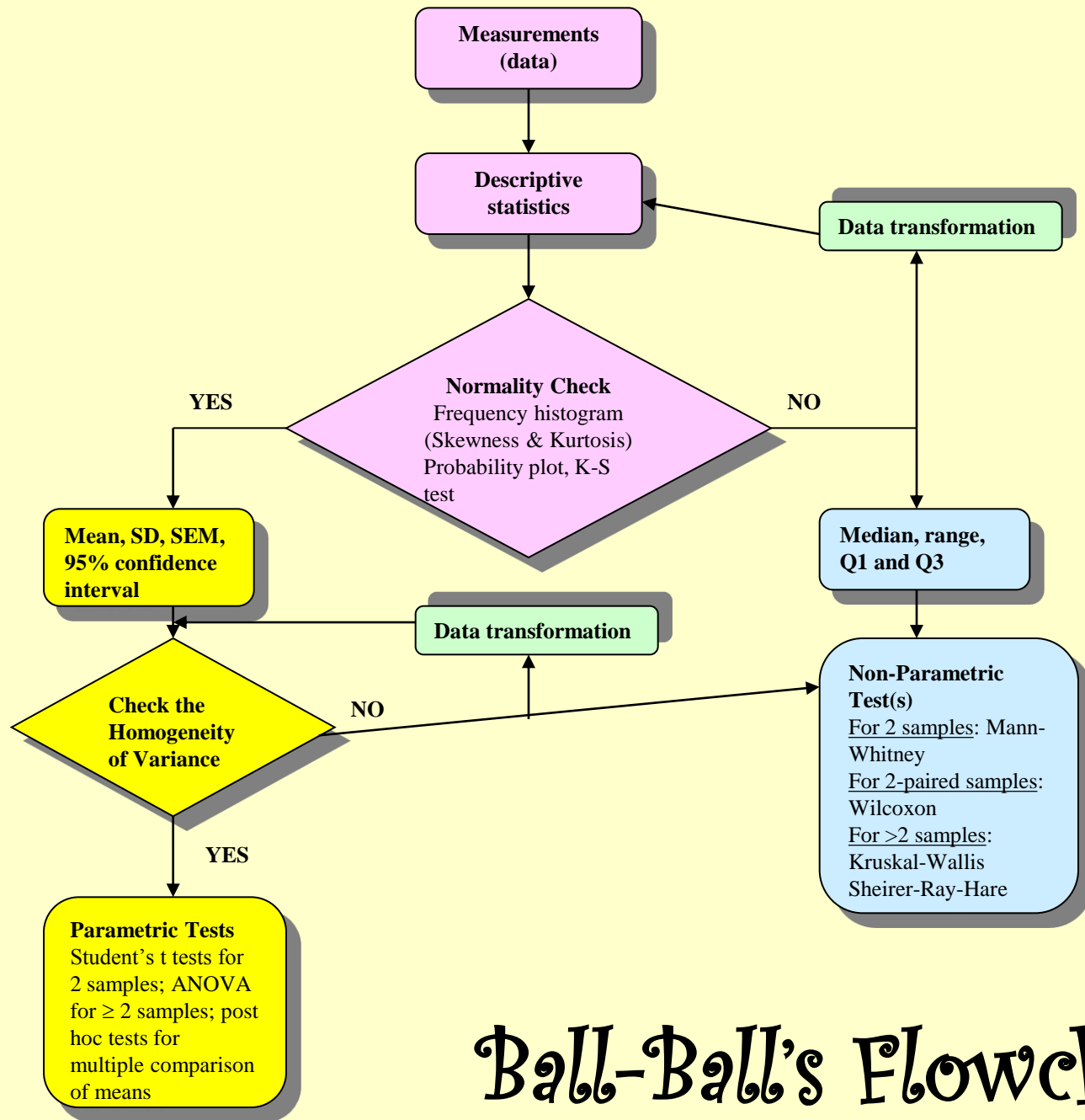
3



Scale

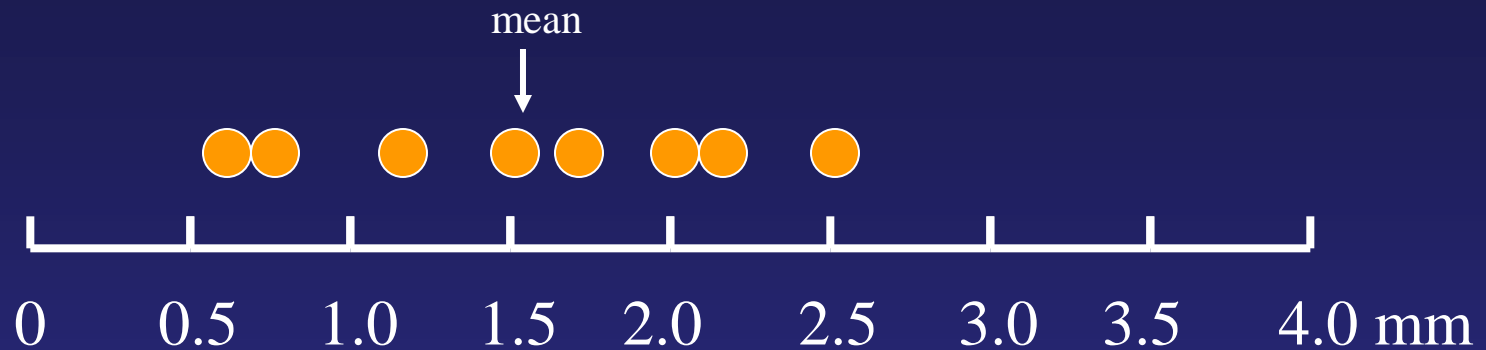
How to Describe the Data Distribution?

- Central tendency
 - Mean for normally distributed data
 - Median for non-normally distributed data
- Dispersal pattern
 - Standard deviation for normally distributed data
 - Range and/or Quartiles for non-normally distributed data



Ball-Ball's Flowchart

Measurements of Central Tendency

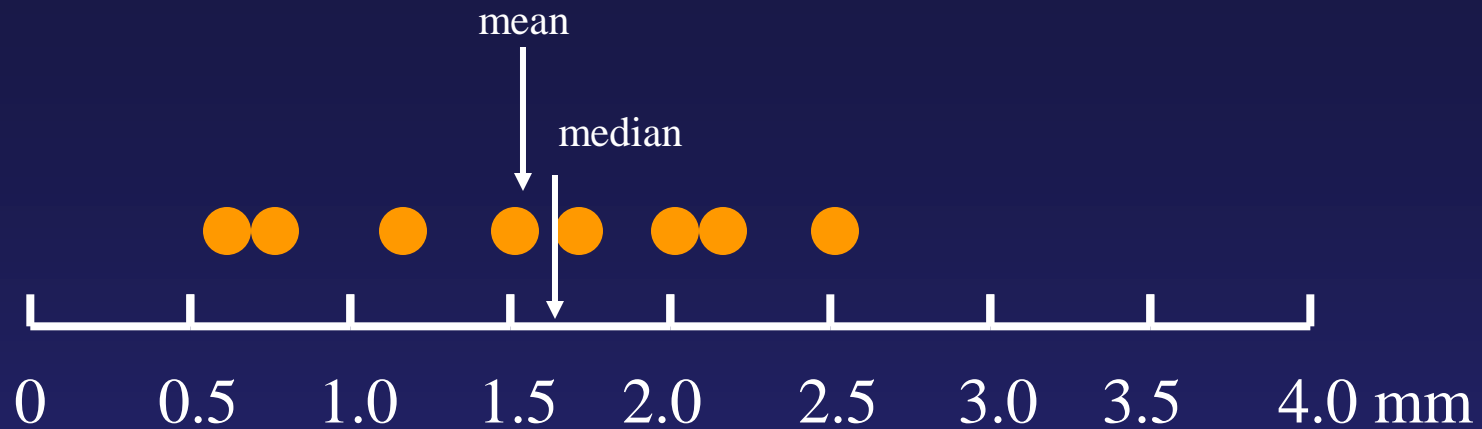


$$\text{Mean} = \text{Sum of values}/n = \sum X_i/n$$

e.g. length of 8 fish larvae at day 3 after hatching:

0.6, 0.7, 1.2, 1.5, 1.7, 2.0, 2.2, 2.5 mm

$$\begin{aligned} \text{mean length} &= (0.6+0.7+1.2+1.5+1.7+2.0+2.2+2.5)/8 \\ &= 1.55 \text{ mm} \end{aligned}$$



Median, Percentiles and Quartiles

- $Order = (n+1)/2$

e.g. 0.6, 0.7, 1.2, 1.5, 1.7, 2.0, 2.2, 2.5 mm

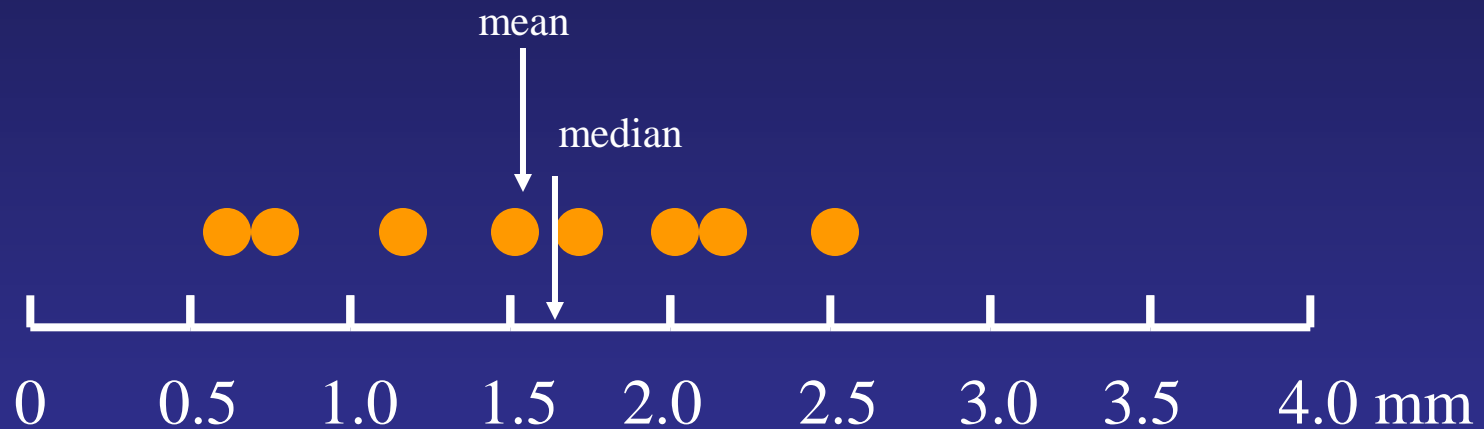
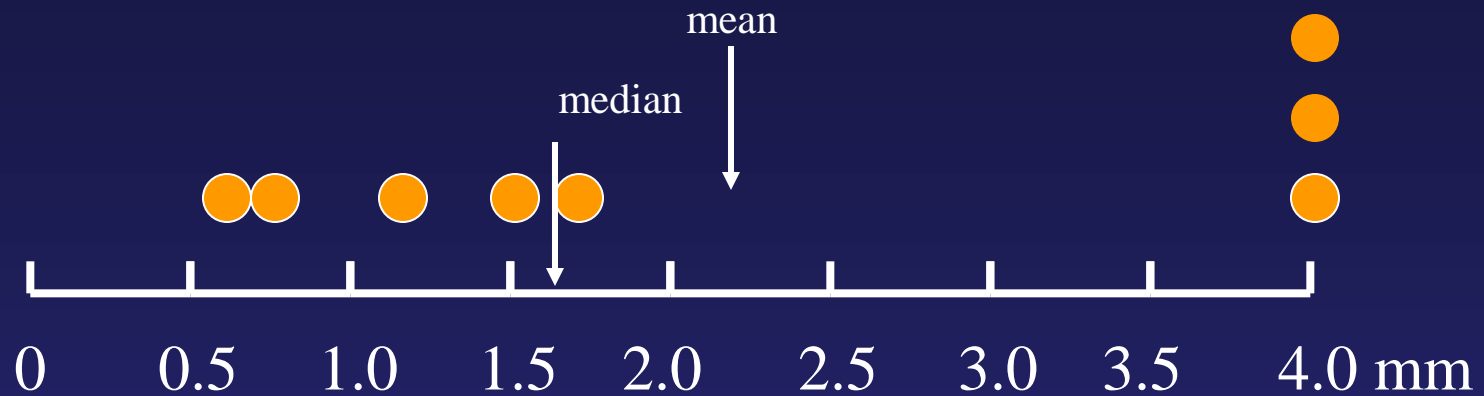
order 1 2 3 4 5 6 7 8

order = $(8+1)/2 = 4.5$

Median = 50th percentile = $(1.5 + 1.7)/2 = 1.6$ mm

order for Q1 = 25th percentile = $(8+1)/4 = 2.25$

then Q1 = $0.7 + (1.2 - 0.7)/4 = 0.825$ mm

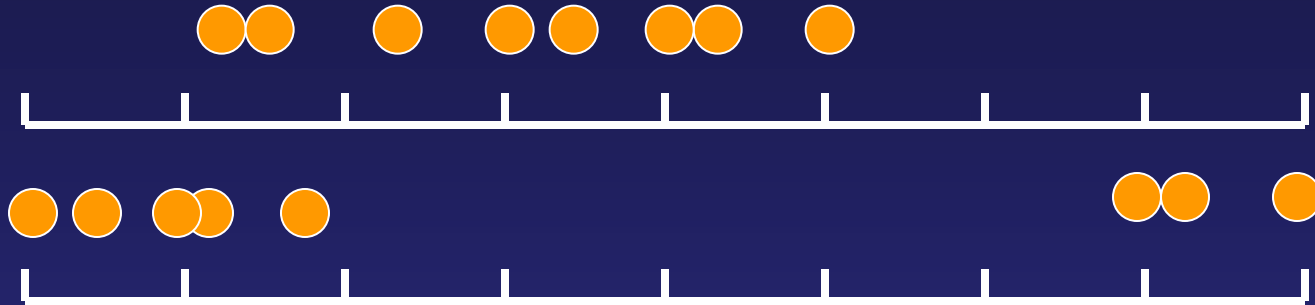


- Median is often used with mean.
- Mean is, however, used much more frequent.
- Median is a better measure of central tendency for data with skewed distribution or outliers.

Other Measures of Central Tendency

- Range midpoint or range = $(\text{Max value} - \text{Min value})/2$
 - not a good estimate of the mean and seldom-used
- Geometric mean = $\sqrt[n]{x_1 x_2 x_3 x_4 \dots x_n}$
= $10^{\text{[mean of } \log_{10}(x_i)]}$
 - Only for positive ratio scale data
 - If data are not all equal, geometric mean < arithmetic mean
 - Use in averaging ratios where it is desired to give each ratio equal weight

Measurements of Dispersion



Range

e.g. length of 8 fish larvae at day 3 after hatching:

0.6, 0.7, 1.2, 1.5, 1.7, 2.0, 2.2, 2.5 mm

Range = $2.5 - 0.6 = 1.9$ mm (or say from 0.6 to 2.5mm)

Percentile and quartiles

Population Standard Deviation (σ)

- Averaged measurement of deviation from mean

$$x_i - \bar{x}$$

- e.g. five rainfall measurements, whose mean is 7

Rainfall (mm)	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
12	$12 - 7 = 5$	25
0	$0 - 7 = -7$	49
2	$2 - 7 = -5$	25
5	$5 - 7 = -2$	4
16	$16 - 7 = 9$	81
	Sum = 184	Sum = 184

- Population variance: $\sigma^2 = \sum (x_i - \bar{x})^2 / n = 184 / 5 = 36.8$
- Population SD: $\sigma = \sqrt{\sum (x_i - \bar{x})^2 / n} = 6.1$

Sample SD (s)

$$s = \sqrt{[\sum(x_i - \bar{x})^2] / (n - 1)}$$

$$s = \sqrt{[\sum x_i^2 - ((\sum x_i)^2 / n)] / (n - 1)}$$

- Two modifications:
 - by dividing $\sqrt{[\sum(x_i - \bar{x})^2]}$ by $\sqrt{(n-1)}$ rather than \sqrt{n} , gives a better unbiased estimate of σ (however, when n increases, difference between s and σ declines rapidly)
 - the sum of squared (SS) deviations can be calculated as $\sum (x_i^2) - (\sum x_i)^2 / n$

Sample SD (s)

- e.g. five rainfall measurements, whose mean is 7.0*

Rainfall (mm)	x_i^2	x_i
12	144	12
0	0	0
2	4	2
5	25	5
16	256	16
	$(\sum x_i^2) = 429$	$\sum x_i = 35$
		$(\sum x_i)^2 = 1225$

- $s^2 = [\sum x_i^2 - (\sum x_i)^2 / n] / (n - 1) = [429 - (1225/5)] / (5 - 1) = 46.0 \text{ mm}$
- $s = \sqrt{46.0} = 6.782 = 6.8 \text{ mm}$

Basic Experimental Design for Environmental Research

1. Setting environmental questions into statistical questions [e.g. spatial and temporal variations]
2. Setting hypotheses and then statistical null hypotheses
3. Statistical consideration (treatment groups, sample size, true replication, confounding factors etc.)
4. Statistical consideration (treatment groups, sample size, true replication, confounding factors etc.)
5. Sampling design (independent, random, samples)
6. Data collection & measurement (Quality Control and Quality Assurance Procedures)
7. Data analysis
 - Too few data: cannot obtain reliable conclusions
 - Too many data: extra effort (time and money) in data collection

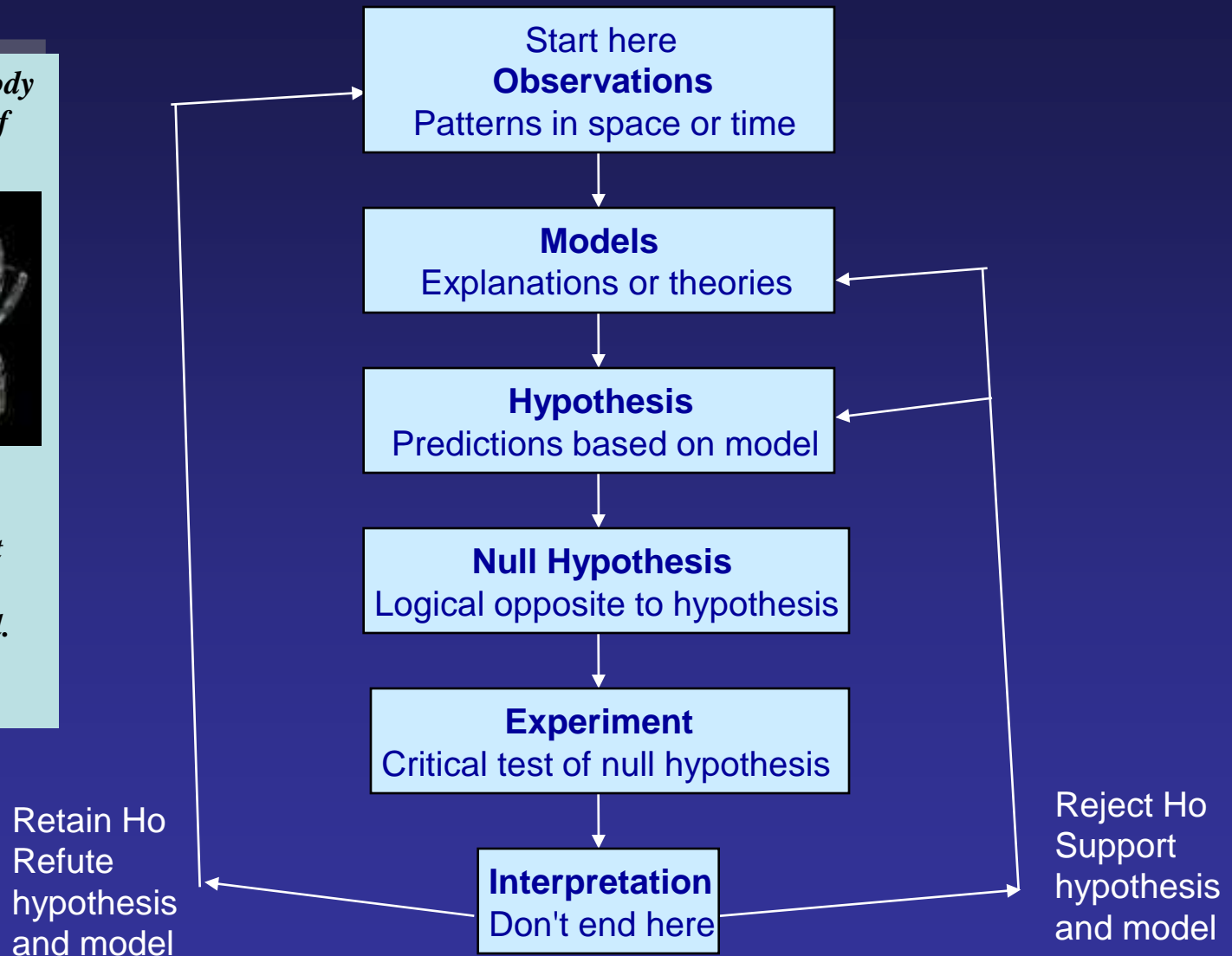
Generalized scheme of logical components of a research programme (Underwood 1997)

Weapon size versus body size as a predictor of winning fights



Carcinus maenas

Reference: Sneddon et al. 1997, in *Behav. Ecol. Sociobiol.* 41: 237 - 242



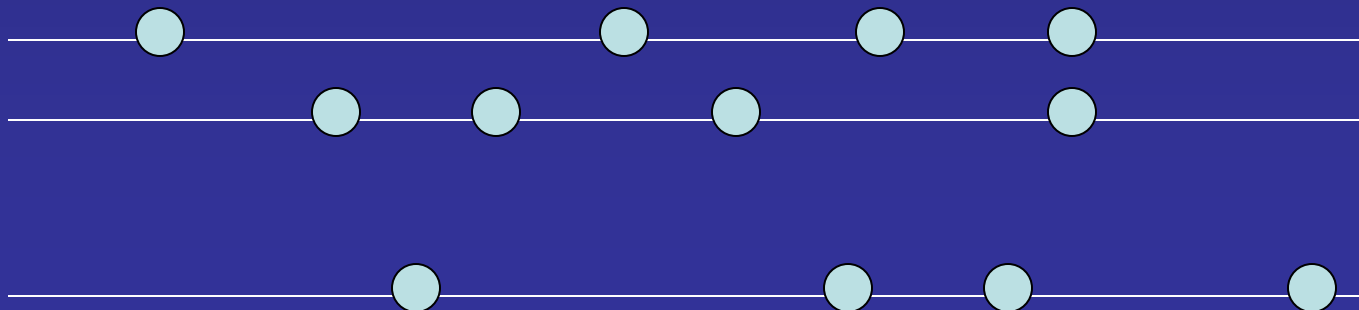
Randomized Sampling

- **Lucky Draw Concept**

- To randomly select 30 out of 200 sampling stations in Hong Kong waters, you may perform a lucky draw.
- So, the chance for selecting each one of them for each time of drawing would be more or less equal (unbiased). It can be done with or without replacement.

- **Sampling with Transects and a Random Number Table**


- Randomly lay down the transects based on random nos.
- Randomly take samples along each transect.



Randomized Sampling

- **Spatial Comparison – Clustered Random Sampling**

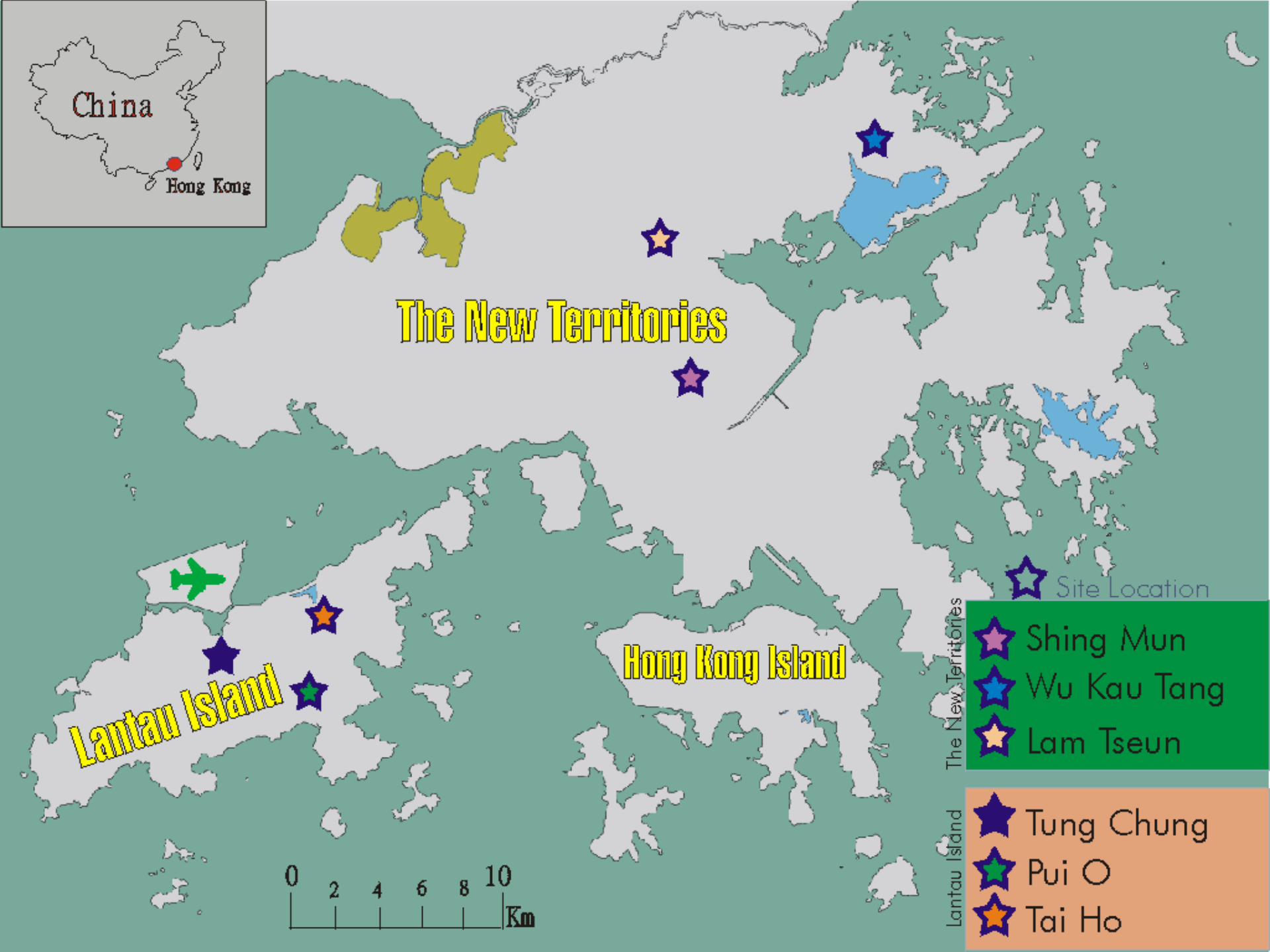
Randomly take
e.g. 10 samples
from each
randomly
selected site



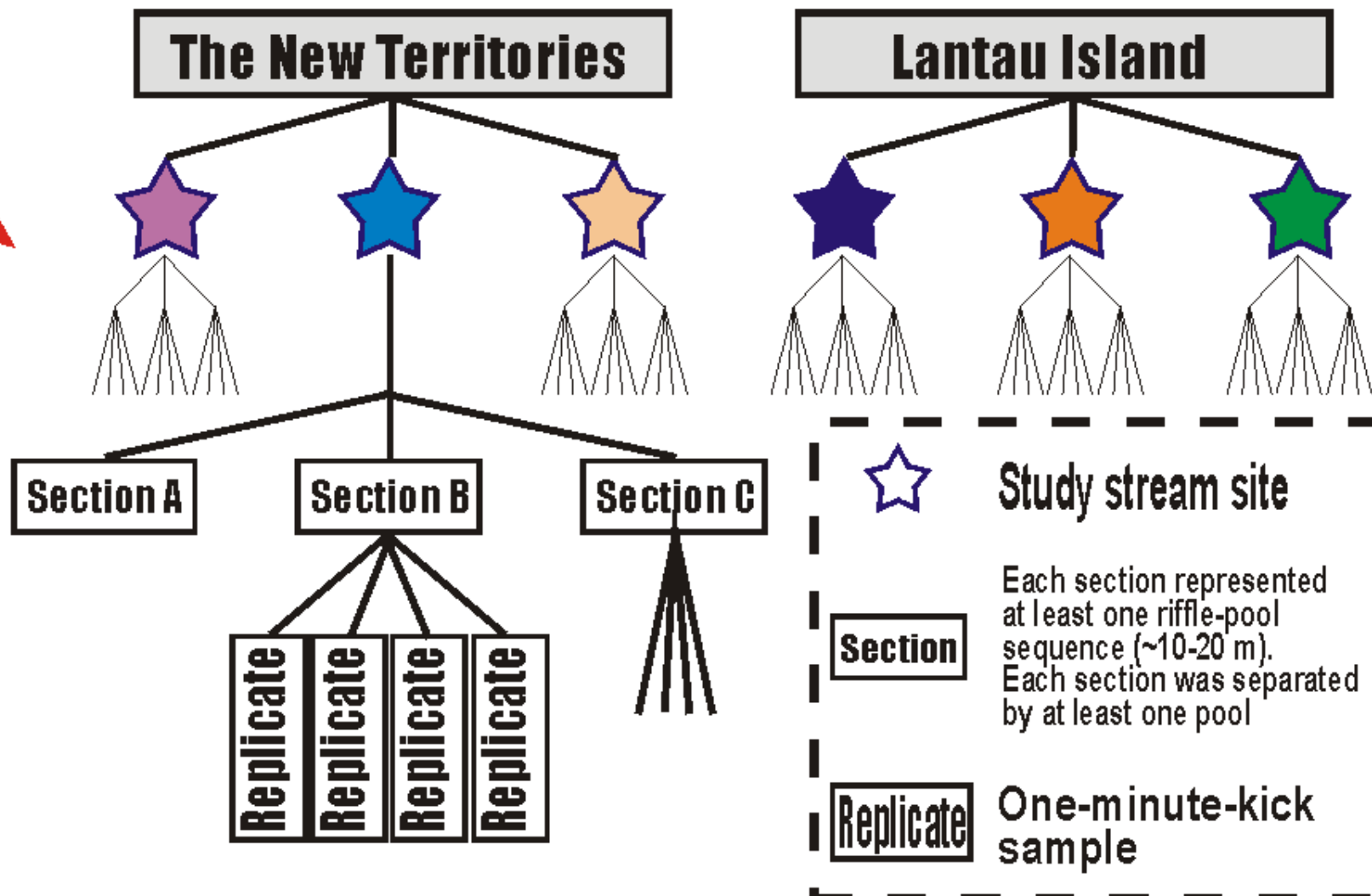
S1	S2	S3	S4	S5	S6	S7
S8	S9	S10	S11	S12	S13	S14
S15	S16	S17	S18	S19	S20	S21

- **Temporal Comparison**

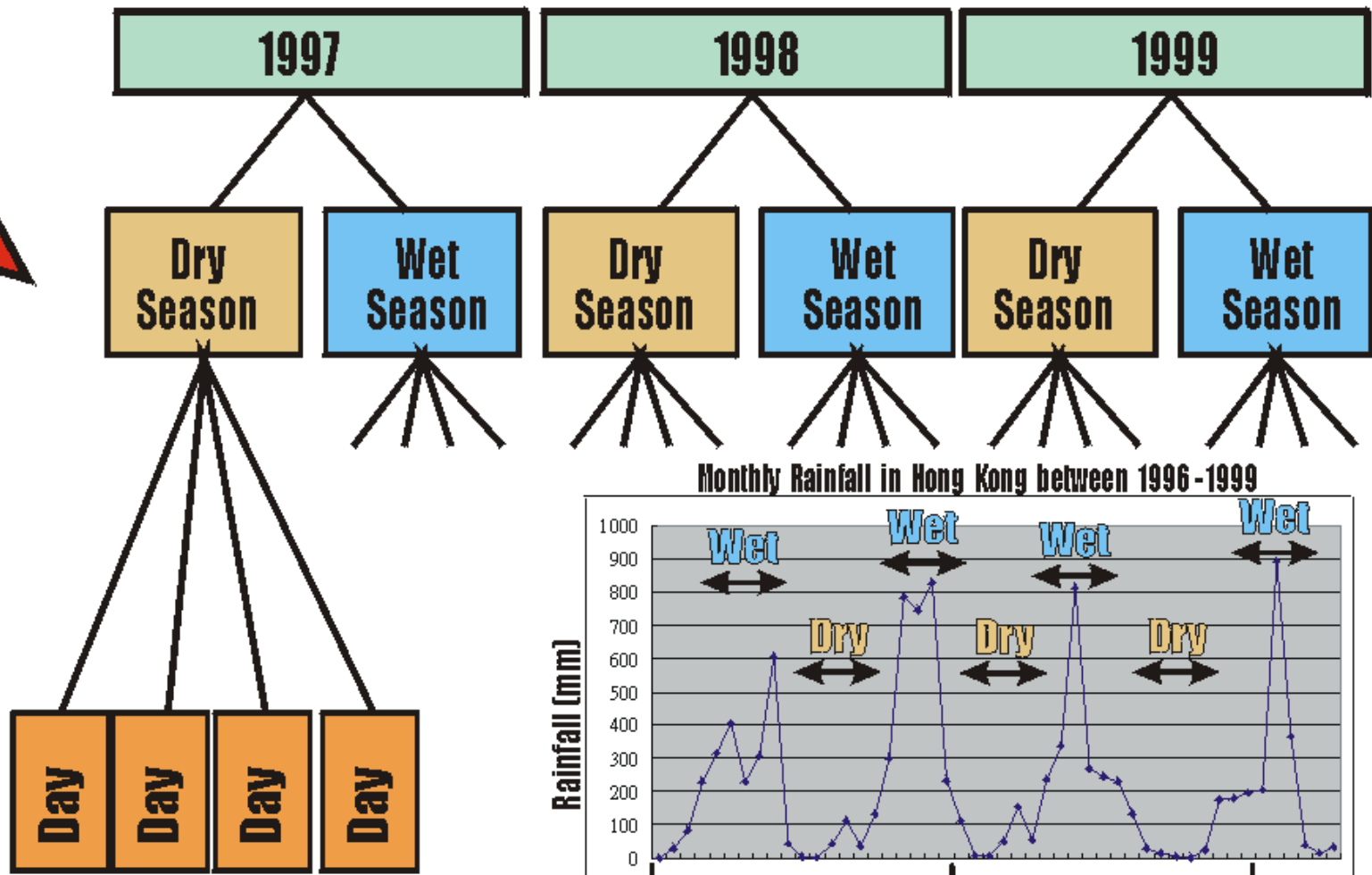
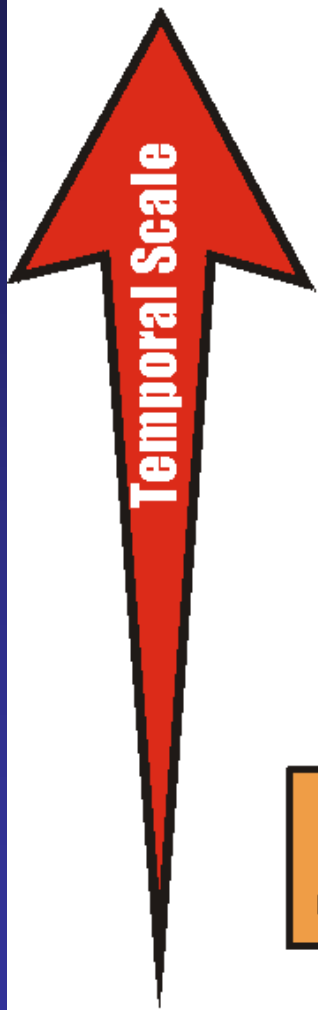
- Wet Season vs. Dry Season
- Randomly select sampling days within each season (assuming each day is independent to other days) covering both neap and spring tides.
- Transitional period should not be selected to ensure independency of the two seasons.



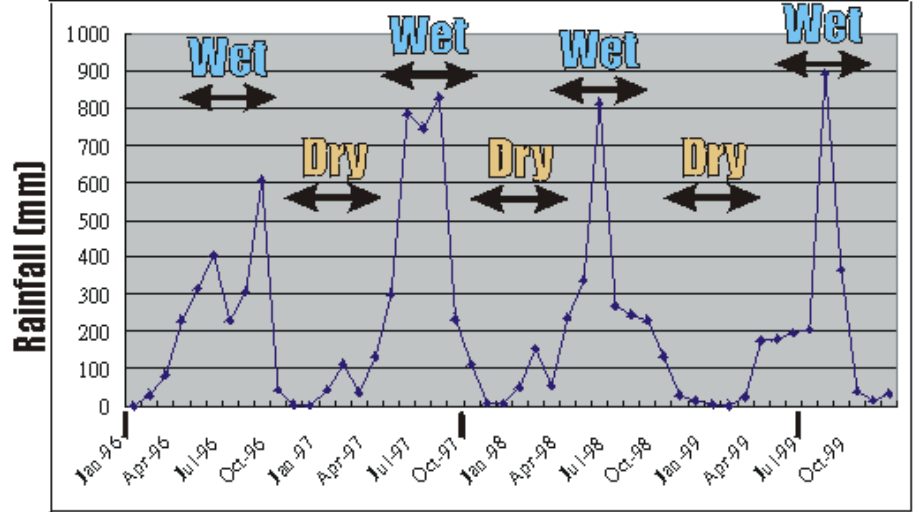
Sampling design for investigating the spatial scale variabilities in and among stream benthic communities in Hong Kong



Sampling Design for Investigating the Temporal Scale Variability In and Among Stream Benthic Communities in Hong Kong



Monthly Rainfall in Hong Kong between 1996-1999

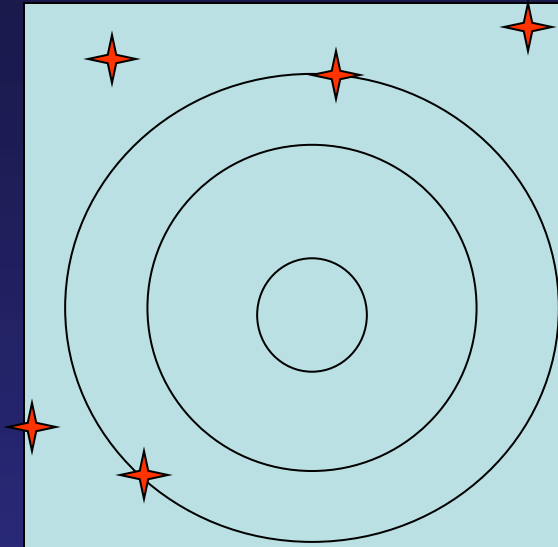


Stratified (Random) Sampling

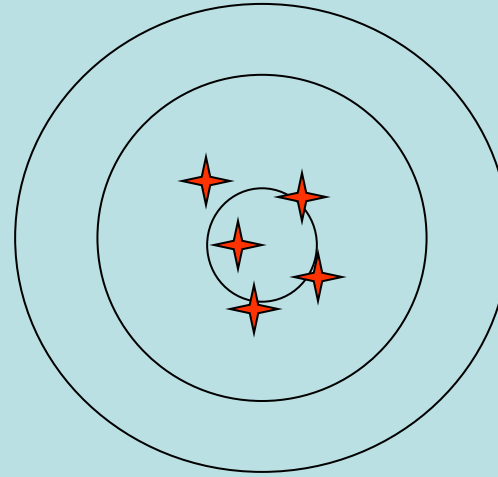
- The population is first divided into a number of parts or 'strata' according to some characteristic, chosen to be related to the major variables being studied.
 - Water samples from three different water depths (1 m from the surface, mid-depth, 1 m above seabed).
 - Water samples from a point source of pollution using a transect (set away from the source to open sea) with fixed sampling intervals (e.g. 1, 5, 10, 20, 50, 100, 500, 1000, 2000 m).
 - Sediment samples from the high (2 m of Chart Datum), mid (1 m CD) and low intertidal zones (0.5 m CD).
 - Sediment and water samples from different beneficial uses in Hong Kong waters.

Precision and Accuracy

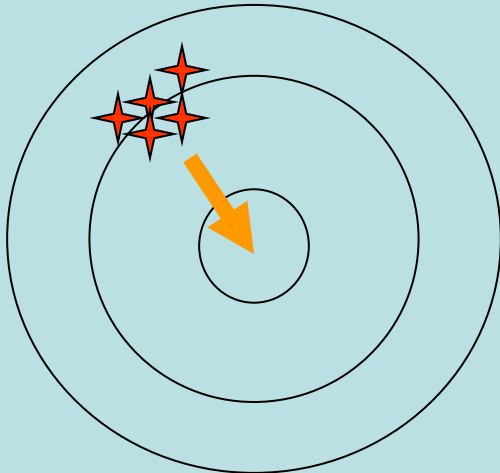
Neither
precise
nor
accurate



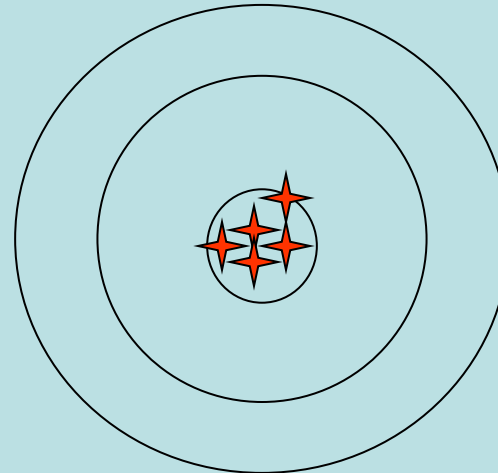
Moderately
precise
and
accurate



Highly
precise
but not
accurate

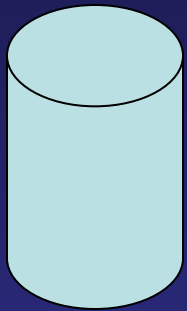


Highly
precise
and
accurate

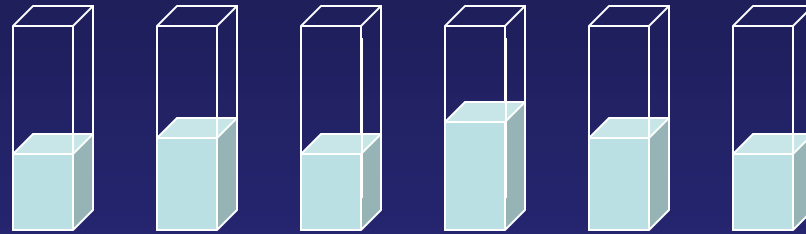


Quality Control & Quality Assurance

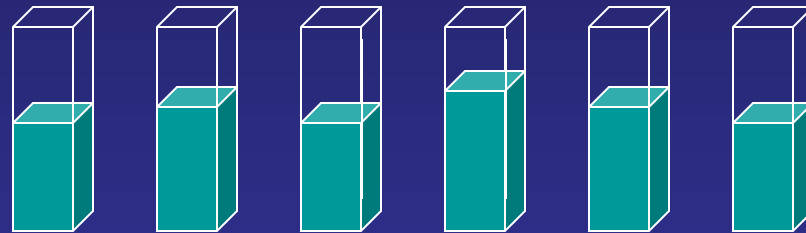
- e.g. Total phosphate measurements for a water sample



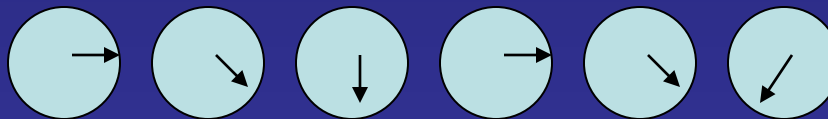
Precision can be estimated using procedure replicates.



Step 1: Pipette 1 ml sample to a cuvette

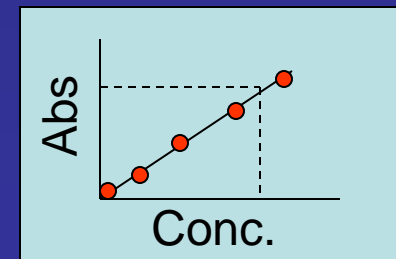


Step 2: Pipette 0.5 ml colour reagent



Step 3: Reaction for 15 minutes

Accuracy can be checked with certified standard reference solutions.





DORM-2

Dogfish Muscle Certified Reference Material for Trace Metals

The following table shows those elements for which certified values have been established for the dogfish (*Squalus acanthias*) reference material. Certified values are based on results of determinations by at least two independent methods of analysis. The uncertainties represent 95 percent tolerance limits for an individual sub-sample of 250 mg or greater.

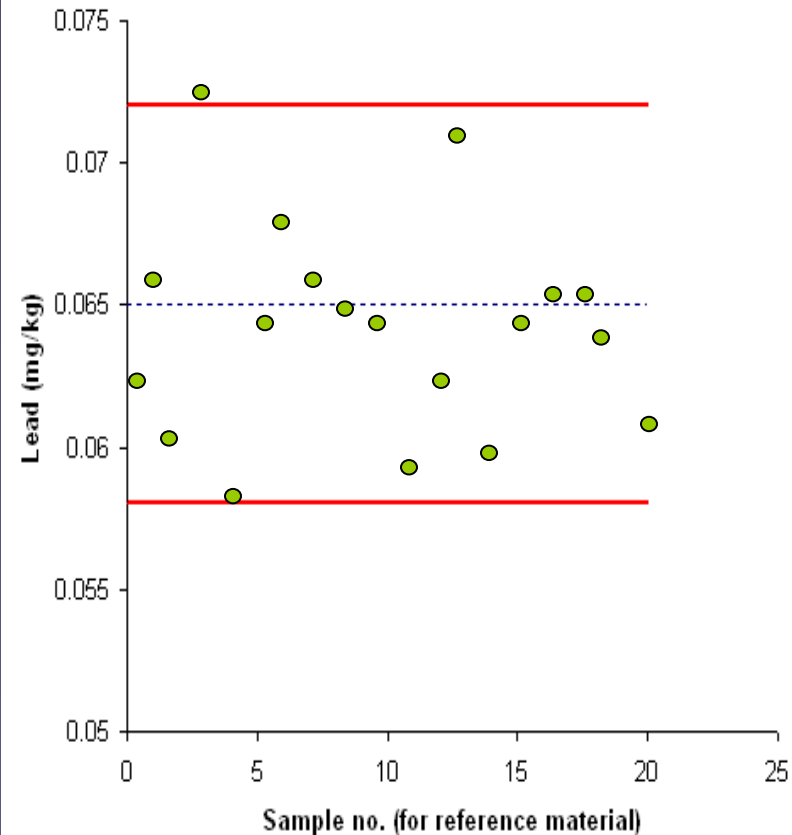
TRACE ELEMENTS (milligrams/kilogram)

Aluminum (d,g,l)	10.9	±	1.7
Arsenic (d,g,h,x)	18.0	±	1.1
Cadmium (g,p)	0.043	±	0.008
Cobalt (d,g)	0.182	±	0.031
Chromium (g,l,p)	34.7	±	5.5
Copper (g,l,p,x)	2.34	±	0.16
Iron (g,l,p,x)	142	±	10
Lead (g,p)	0.065	±	0.007
Manganese (d,g,l)	3.66	±	0.34
Mercury (c,p)	4.64	±	0.26
Nickel (g,l,p)	19.4	±	3.1
Selenium (g,p)	1.40	±	0.09
Silver (g,p)	0.041	±	0.013
Thallium (p)	(0.004)*		
Tin (p)	(0.023)*		
Zinc (f,g,l,p)	25.6	±	2.3
Methylmercury (as Hg) (e,t)	4.47	±	0.32
Arsenobetaine (as As) (l,m)	16.4	±	1.1
Tetramethylarsonium (as As) (l)	0.248	±	0.054

* Information value only

Lead 0.065 ± 0.007

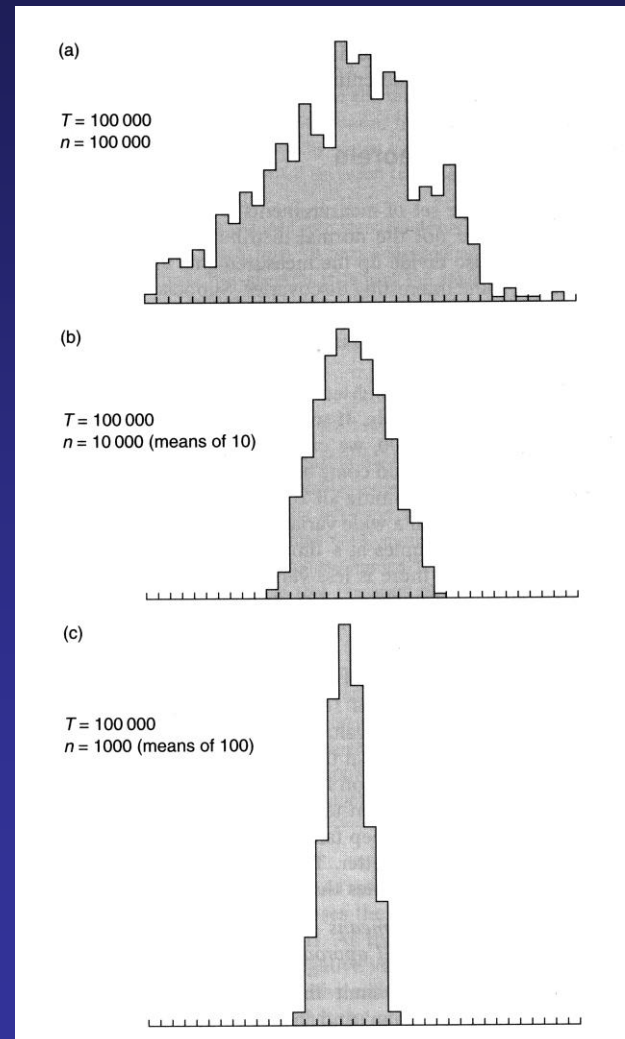
QC & QA: Control Chart



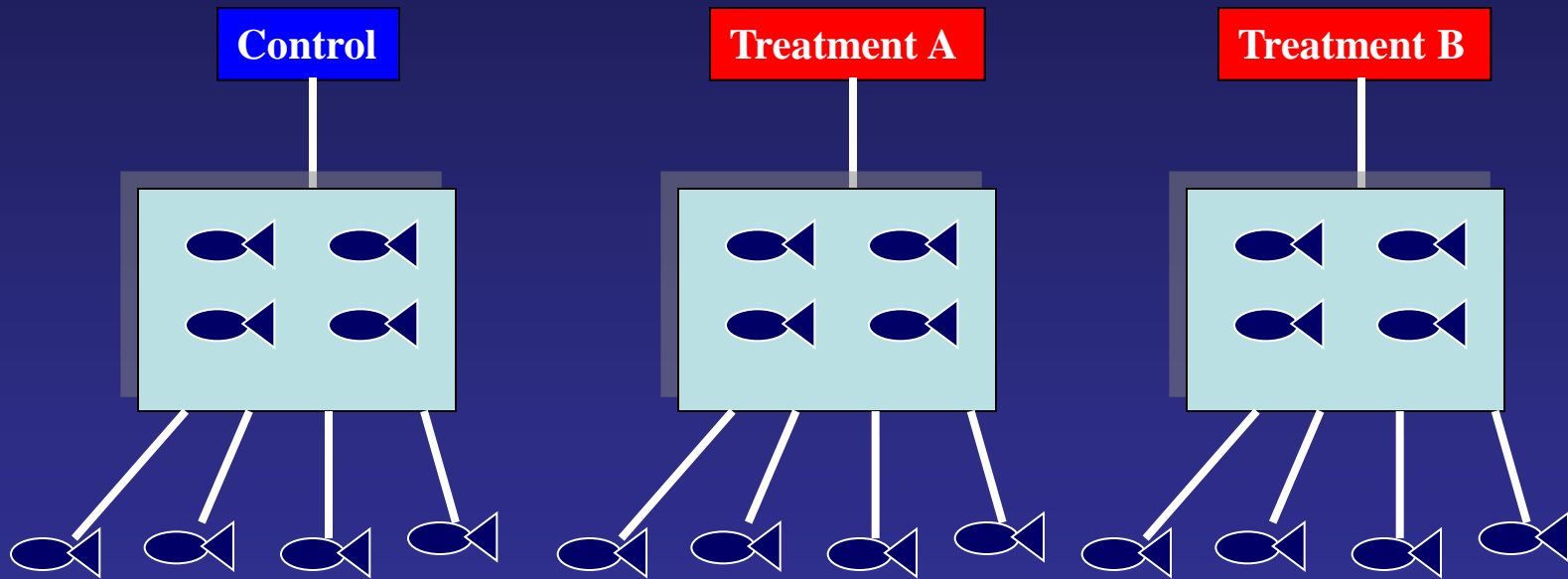
The measured mean value can be compared with the certified mean value using one-sample *t*-test.

Why is it so important to use the “mean of the means” in the experimental design?

- **Central Limit Theorem**
- The mean will remain the same if a mean of the means is used instead of taking a simple mean but the SD of the means will be substantially smaller than the original sample SD.
- For each water body, 50 samples are taken. It is advantageous if they are grouped into 5 groups of 10 samples to compute the mean of the means. This will increase the power for subsequent comparison with other sites.

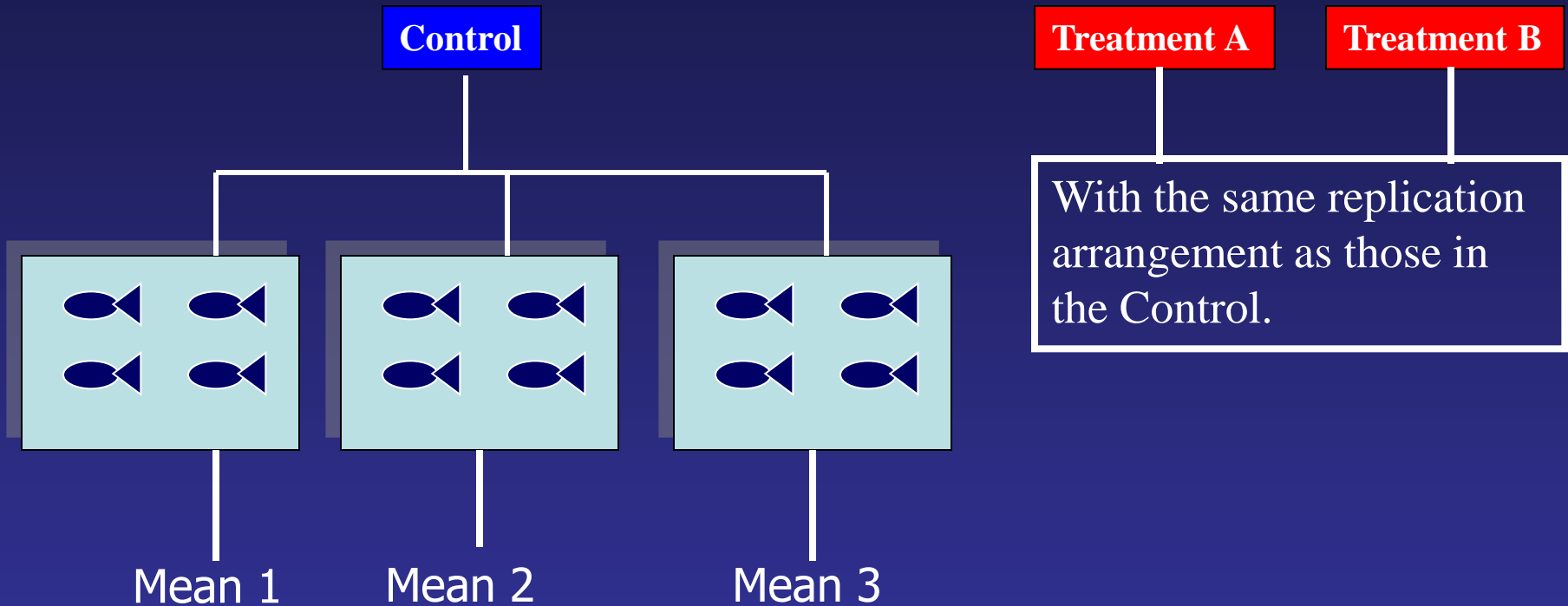


True Replication vs. Pseudo-Replication

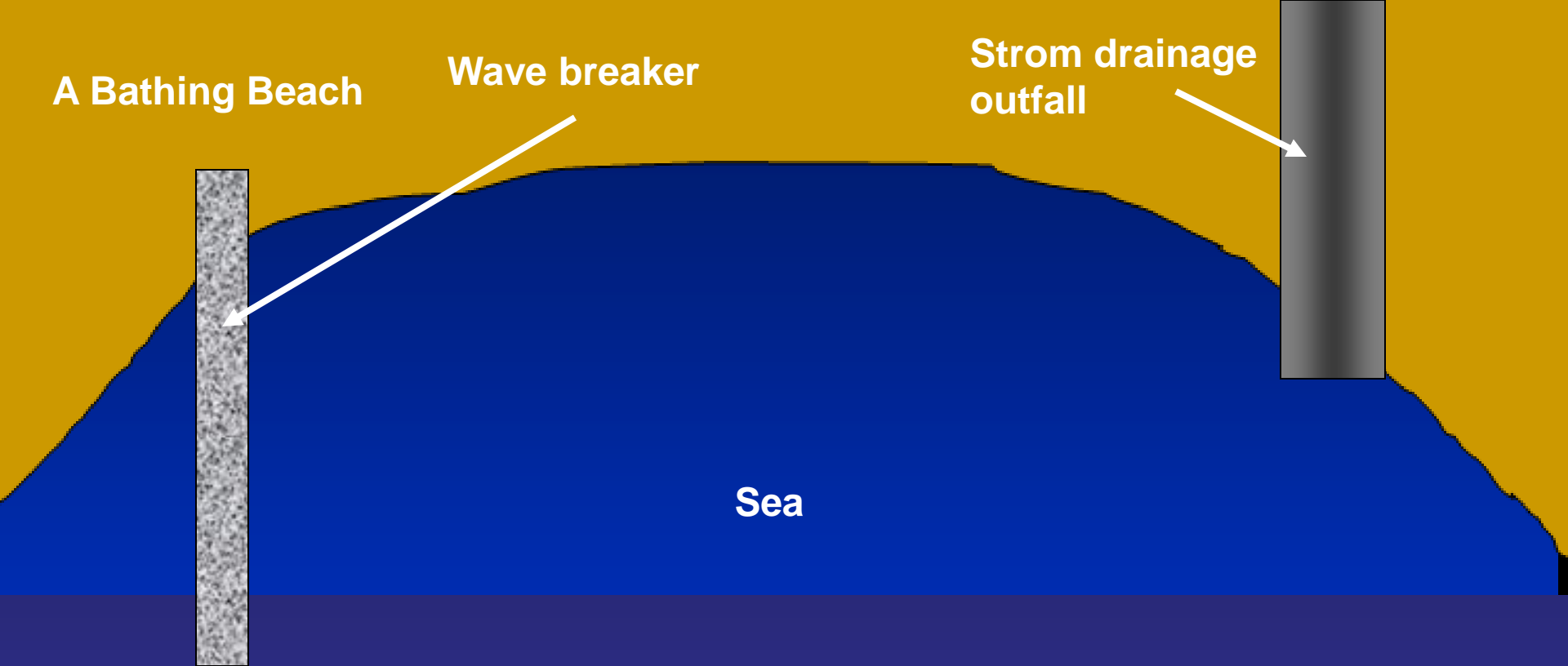


Will it be correct to say that there are four replicates per group? If not, why?

True Replication vs. Pseudo-Replication

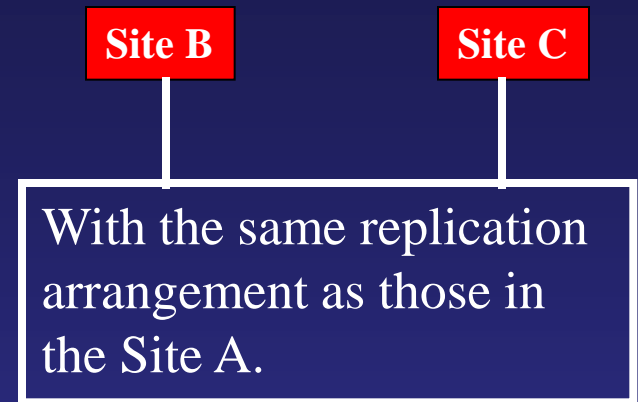
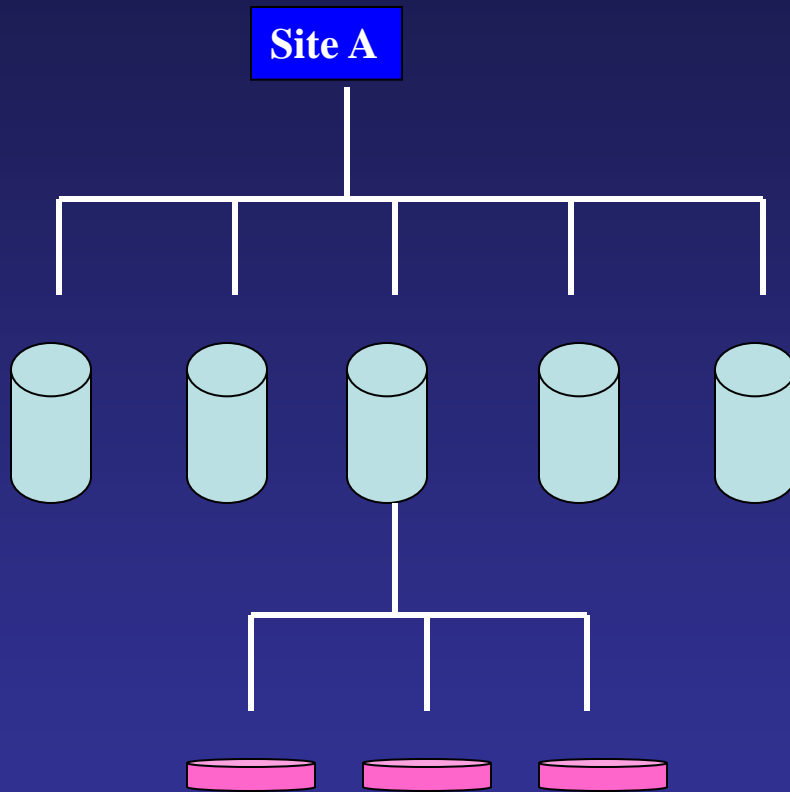


Will it be correct to say that there are three replicates per group? If yes, why?



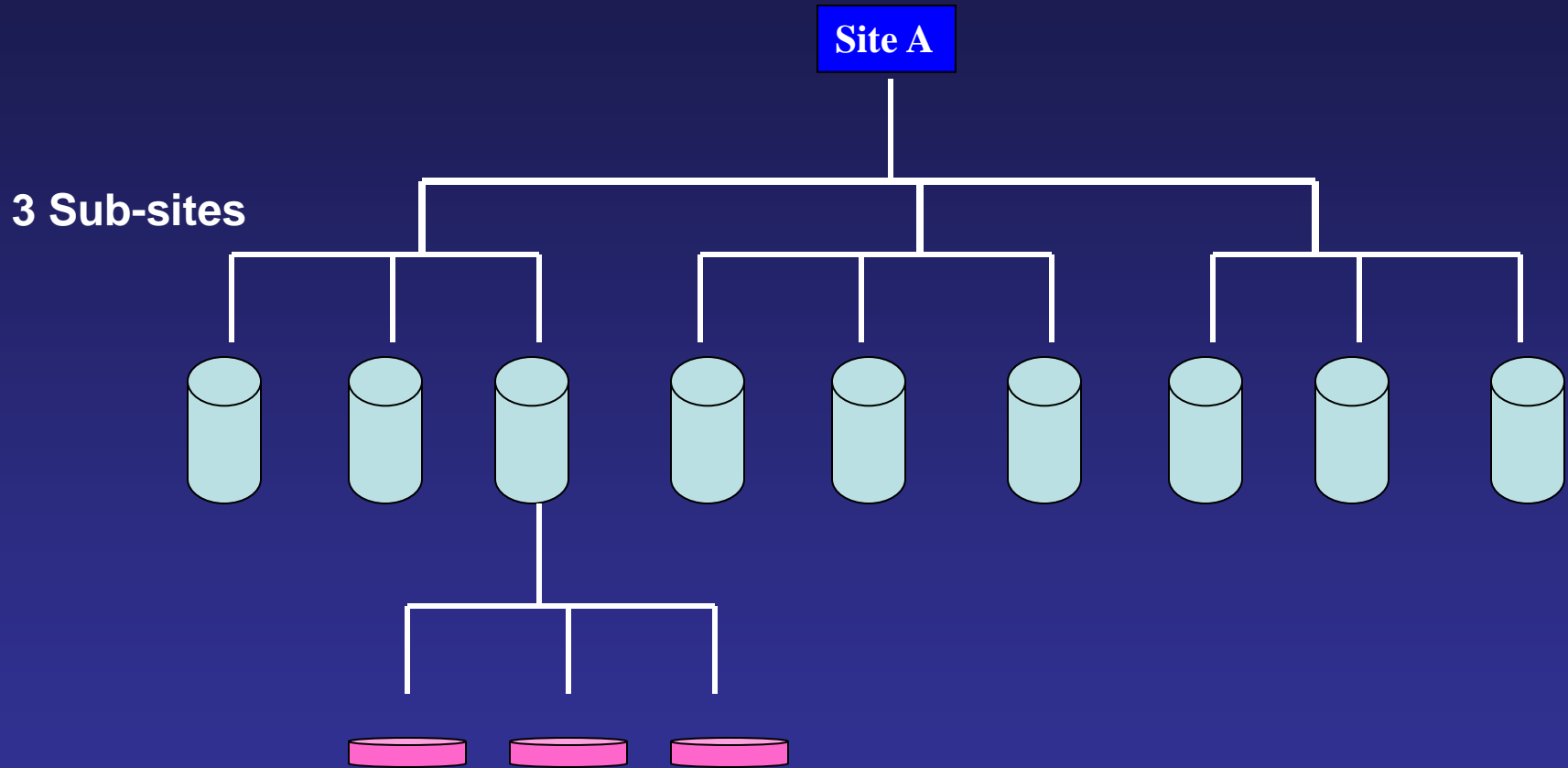
How can we obtain a statistically sound figure of *E. coli* count for this bathing beach?

True Replication vs. Pseudo-Replication



Five replicates per group and each replicate with three 'procedure replicates' to ascertain the measurement precision.

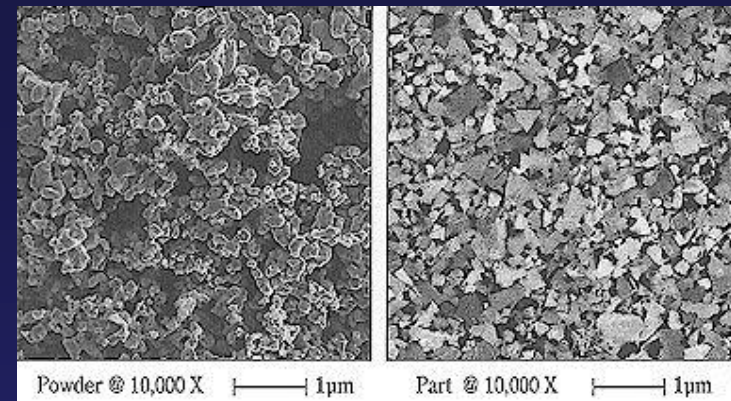
True Replication vs. Pseudo-Replication



Three replicated sites per site, each replicated site with three replicate samples and each sample with three 'procedure replicates' to ascertain the measurement precision.

Inferential Statistics

Sediment grain sizes



e.g. The particle sizes (μm) of 37 grains from a sample of sediment from an estuary

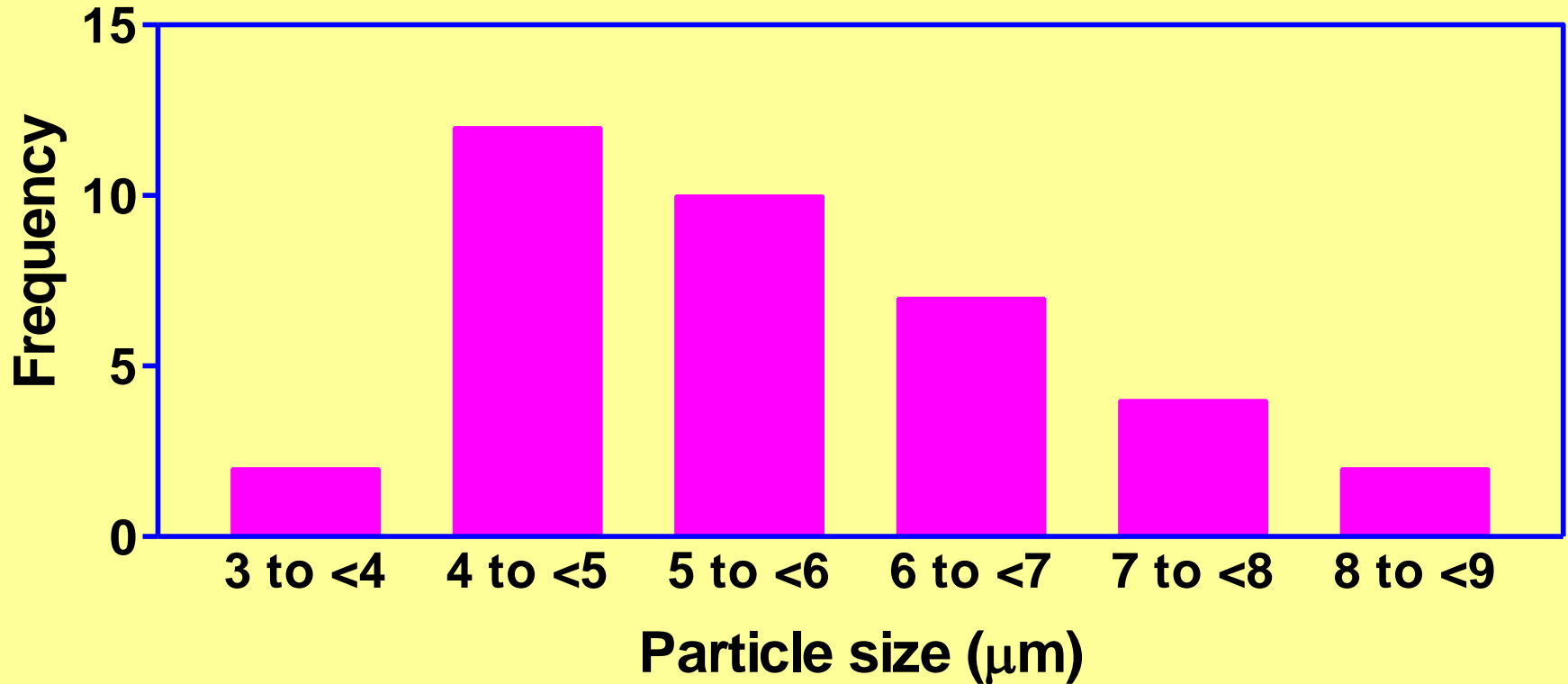
8.2	6.3	6.8	6.4	8.1	6.3	
5.3	7.0	6.8	7.2	7.2	7.1	
5.2	5.3	5.4	6.3	5.5	6.0	
5.5	5.1	4.5	4.2	4.3	5.1	
4.3	5.8	4.3	5.7	4.4	4.1	
4.2	4.8	3.8	3.8	4.1	4.0	4.0

Define convenient classes (equal width) and class intervals e.g. $1 \mu\text{m}$

e.g. A frequency distribution table for the size of particles collected from the estuary

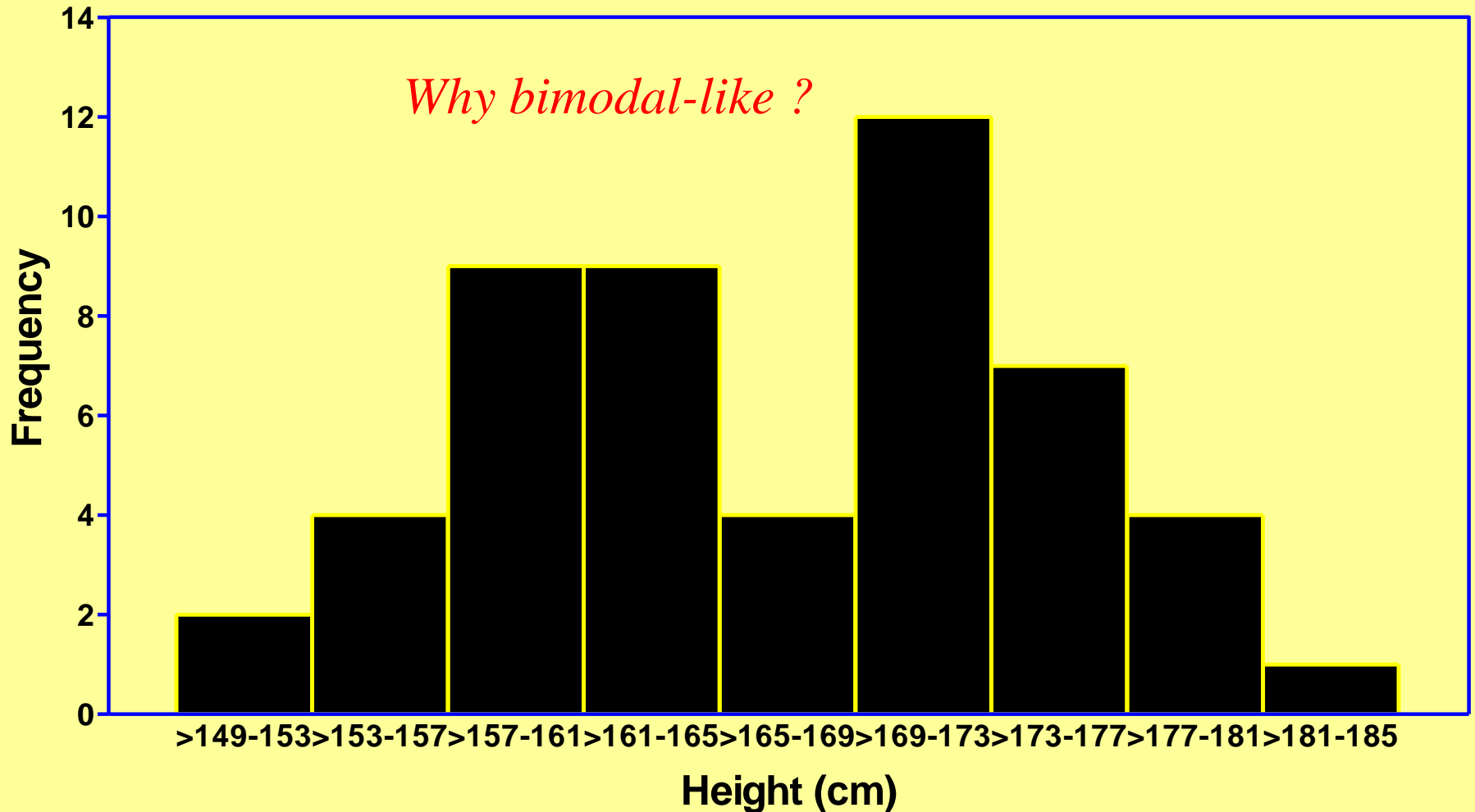
Particle size (μm)	Frequency
3.0 to under 4.0	2
4.0 to under 5.0	12
5.0 to under 6.0	10
6.0 to under 7.0	7
7.0 to under 8.0	4
8.0 to under 9.0	2

Frequency Histogram



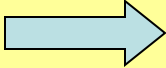
e.g. A frequency distribution for the size of particles collected from the estuary

e.g. A frequency distribution of height of the 30 years old people ($n = 52$: 30 females & 22 males)



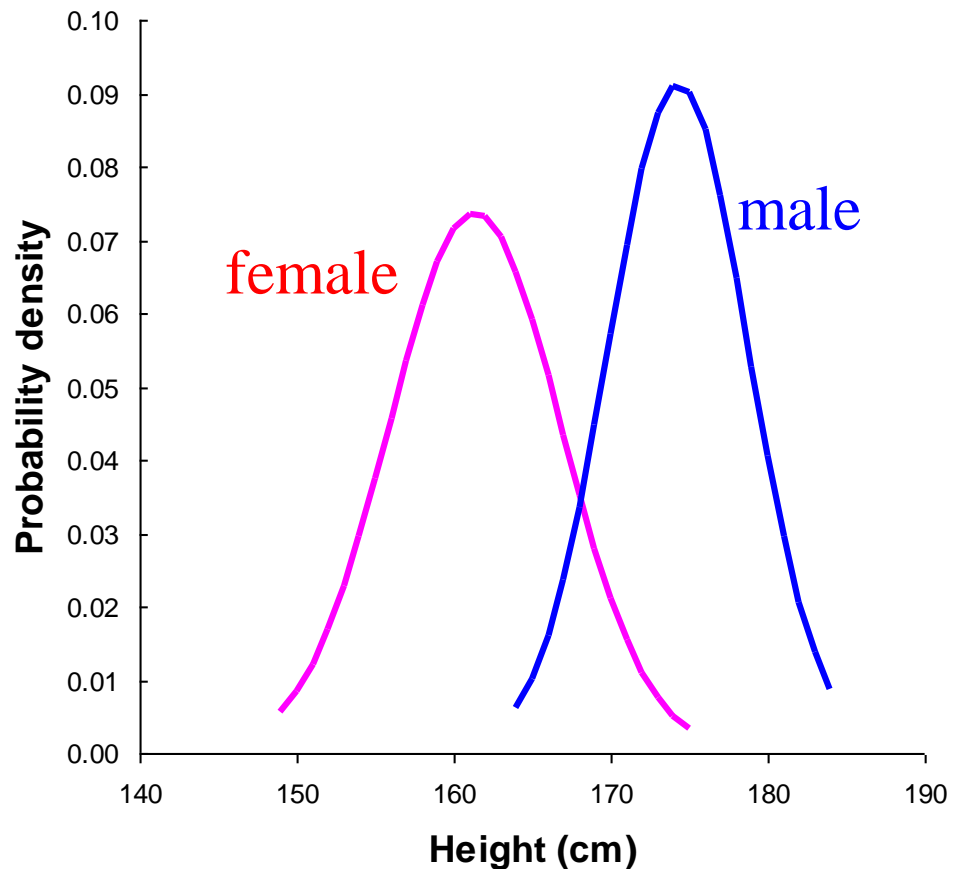
The Normal Curve

- $f(x) = [1/\sigma\sqrt{(2\pi)}]exp[-(x - \mu)^2/(2\sigma^2)]$

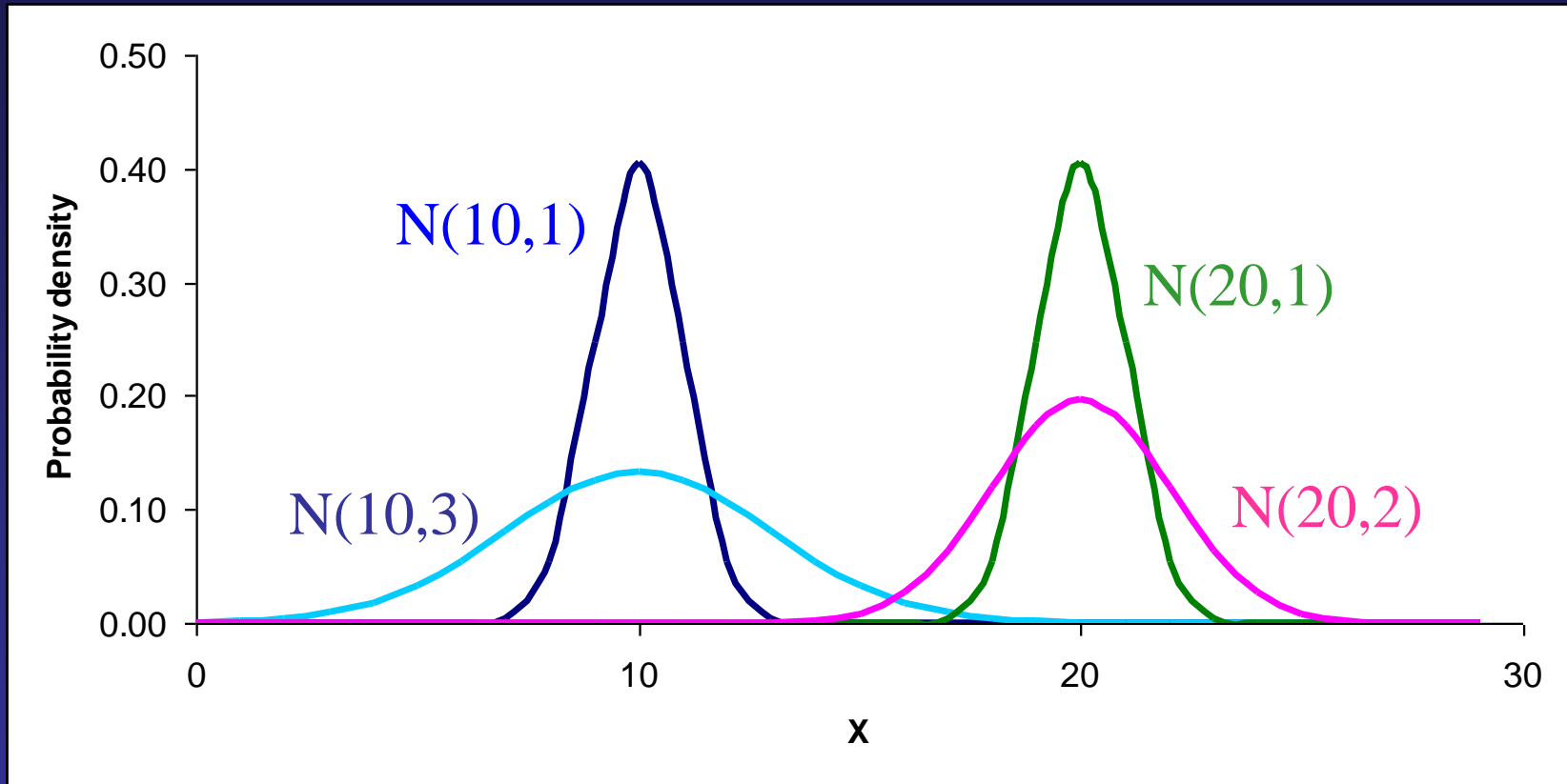
Parameters μ and σ determine the position of the curve on the x-axis and its shape. 

Normal curve was first expressed on paper (for astronomy) by A. de Moivre in 1733.

Until 1950s, it was then applied to environmental problems. (P.S. non-parametric statistics were developed in the 20th century)



$$f(x) = [1/\sigma\sqrt{2\pi}] \exp[-(x - \mu)^2/(2\sigma^2)]$$



- Normal distribution $N(\mu, \sigma)$
- Probability density function: the area under the curve is equal to 1.

The Standard Normal Curve with a Mean = 0

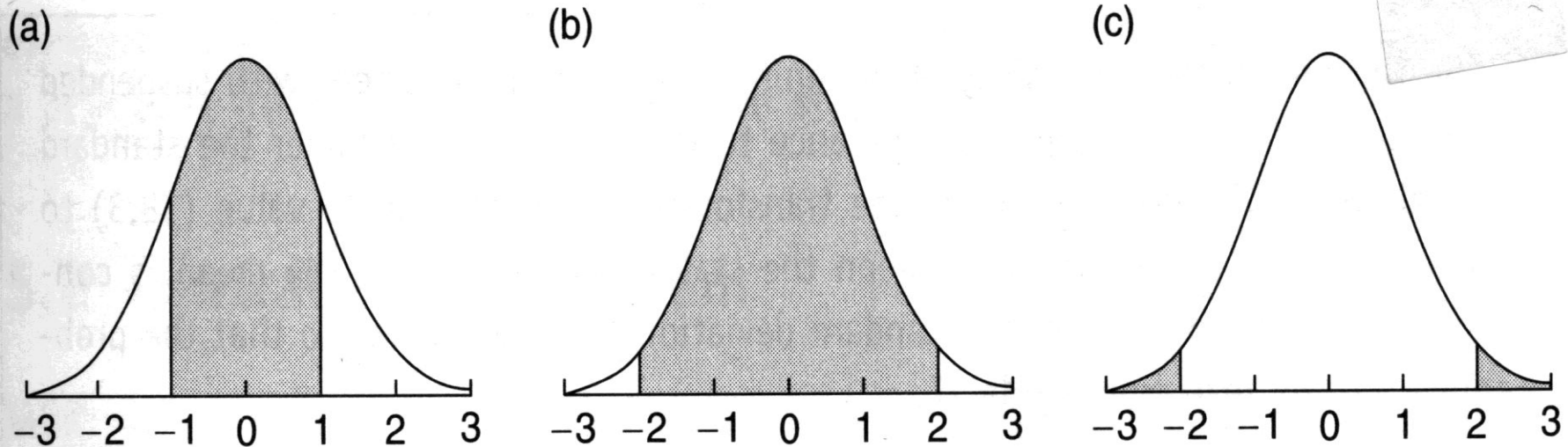


Figure 5.2 A series of standard normal curves: (a) the shaded area lies within one standard deviation of the standardised mean (0); (b) the shaded area is within two standard deviations of the mean; (c) the shaded area lies outside two standard deviations from the mean.

- $\mu = 0$, $\sigma = 1$ and with the total area under the curve = 1
- units along x-axis are measured in σ units
- Figures: (a) for 1 σ , area = 0.6826 (68.26%); (b) for 2 $\sigma \rightarrow$ 95.44%; (c) the shaded area = 100% - 95.44%

Inferential statistics - *testing the null hypothesis*

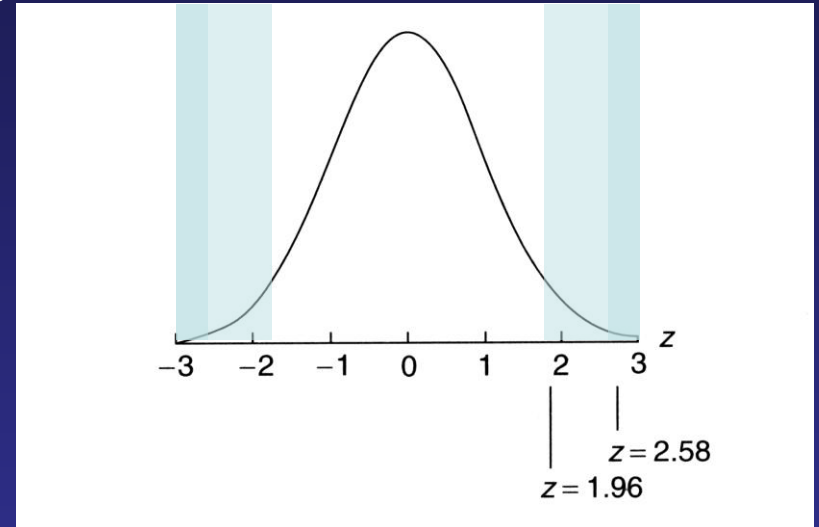
Alternatively, we can state the null hypothesis as that a random observation of Z will lie outside the limit -1.96 or $+1.96$.

There are 2 possibilities:

Either we have chosen an '*unlikely*' value of Z , or our hypothesis is incorrect.

Conventionally, when performing a significant test, we make the rule that if Z values lies outside the range ± 1.96 , then the null hypothesis is rejected and the Z value is termed *significant at the 5% level or $\alpha = 0.05$ (or $p < 0.05$) - critical value of the statistics.*

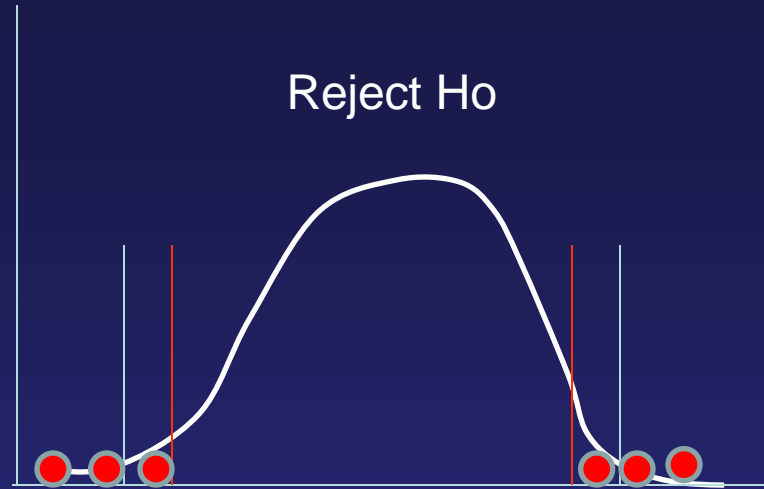
For $Z = \pm 2.58$, the value is termed significant at the 1% level.



Accept H_0



Reject H_0

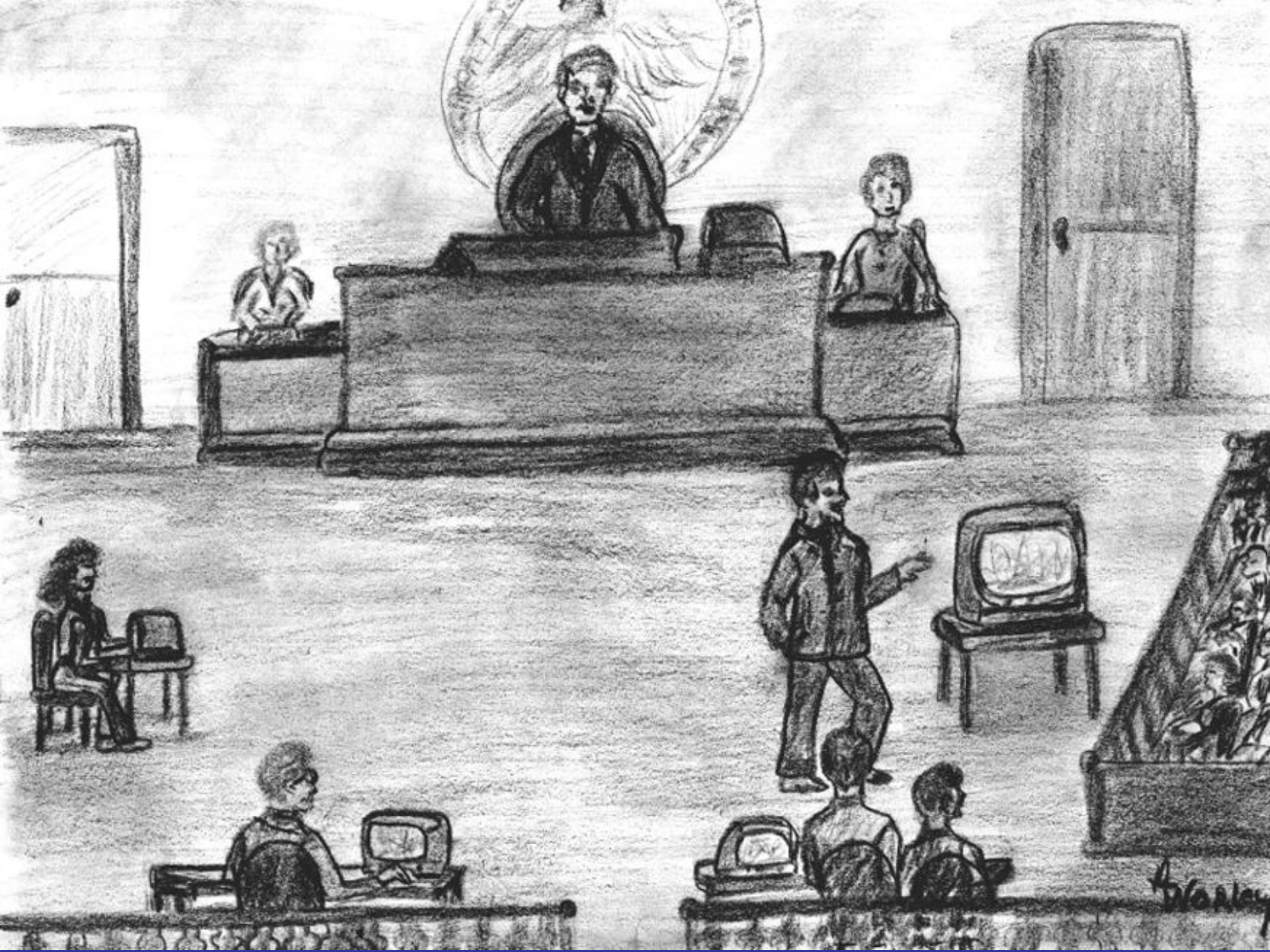


Accept H_0



Reject H_0





Abelley

Statistical Errors in Hypothesis Testing

- Consider court judgements where the accused is presumed innocent until proved guilty beyond reasonable doubt (I.e. $H_0 = \text{innocent}$).

	If the accused is truly innocent (H_0 is true)	If the accused is truly guilty (H_0 is false)
Court's decision: Guilty	Wrong judgement	OK
Court's decision: Innocent	OK	Wrong judgement

Statistical Errors in Hypothesis Testing

- Similar to court judgements, in testing a null hypothesis in statistics, we also suffer from the similar kind of errors:

	If H_0 is true	If H_0 is false
If H_0 is rejected	Type I error	No error
If H_0 is accepted	No error	Type II error

Statistical Errors in Hypothesis Testing

For example, H_0 : The average ammonia concentrations are similar between the suspected polluted Site A and the reference clean Site B, i.e. $A = B$

- **If H_0 is indeed a true statement** about a statistical population, it will be concluded (erroneously) to be false 5% of time (in case $\alpha = 0.05$).
 - Rejection of H_0 when it is in fact true is a **Type I error** (also called an α error).
- **If H_0 is indeed false**, our test may occasionally not detect this fact, and we accept the H_0 .
 - Acceptance of H_0 when it is in fact false is a **Type II error** (also called a β error).

Minimization of Type II error is vitally essential for environmental management.

Power of a Statistical Test

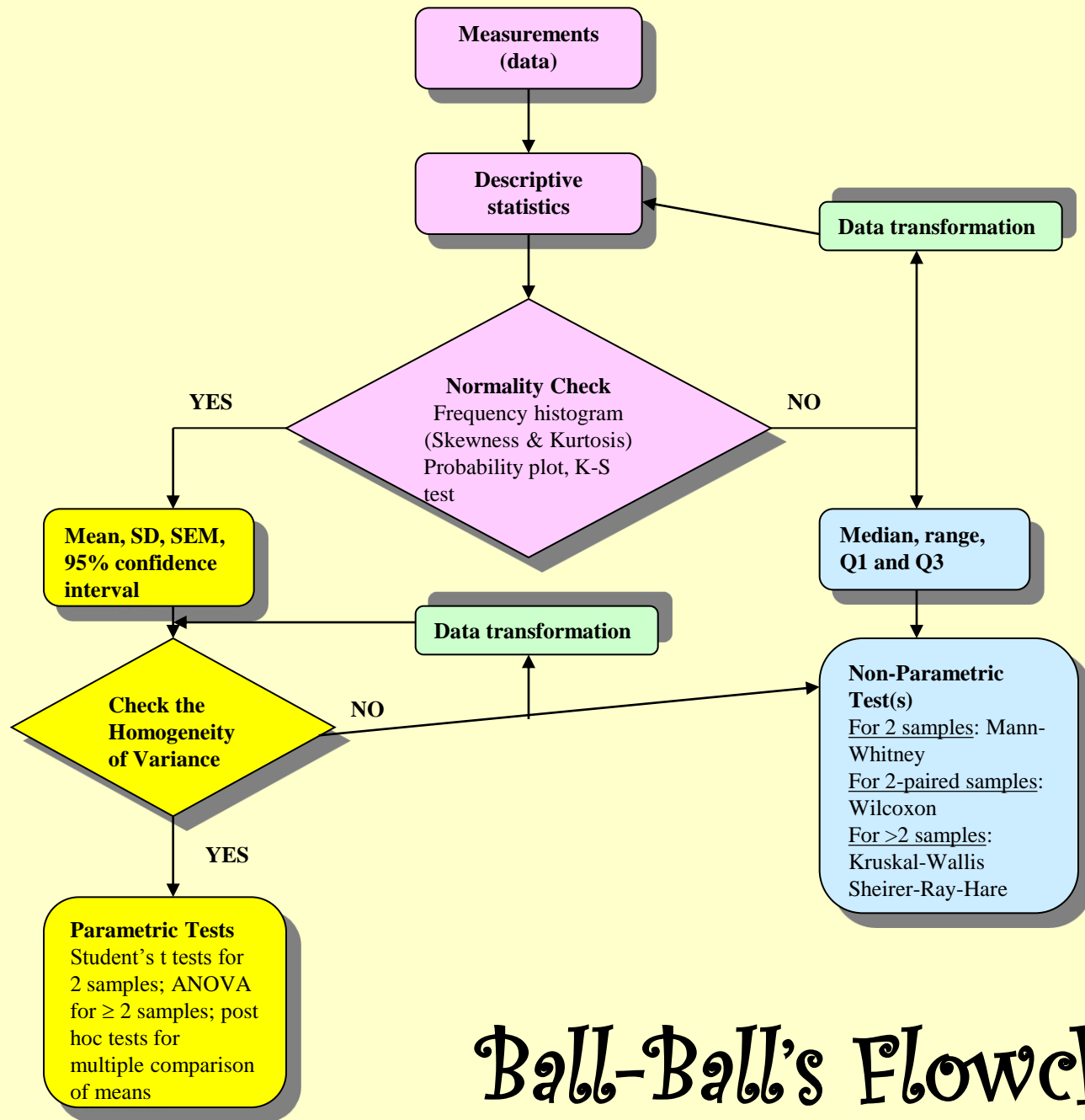
- Power is defined as $1-\beta$.
- β is the probability to have Type II error.
- Power ($1-\beta$) is the probability of rejecting the null hypothesis when it is in fact false and should be rejected.
- Probability of Type I error is specified as α .
- But β is a value that we neither specify nor known.

Power of a Statistical Test

- However, for a given sample size n , α value is related inversely to β value.
- Lower p of committing a Type I error is associated with higher p of committing a Type II error.
- **The only way to reduce both types of error simultaneously is to increase n .**
- For a given α , a large n will result in statistical test with greater power $(1 - \beta)$.

What is next?

- 1. Group Discussion on the Experimental Design for a Case Study**
- 2. Introduction to Two Classes of Basic Statistical Techniques:**
 - (1) correlation based methods and**
 - (2) group comparison methods**
- 3. Power Analysis**



Ball-Ball's Flowchart

Power Analysis with G*Power

A. Comparing Two Samples
Independent Samples t test

B. Comparing More than 2 Samples
Analysis of Variance (ANOVA)

G*Power 3 – Free Software

▶ Download and register

▶ Literature

▶ Program handling

▶ Scientific probability calculator

▶ User guide: Analyses by design

▶ User guide: Analyses by distribution

▶ User guide: Type of Power Analysis

▶ Who we are

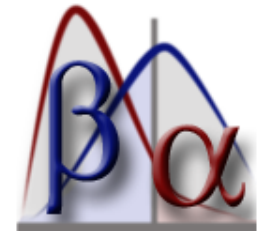
G*Power 3

G*Power 3 is a major extension of, and improvement over, G*Power 2. It covers statistical power analyses for many different statistical tests of the

- F test,
- t test,
- χ^2 -test and
- z test families and some
- exact tests.

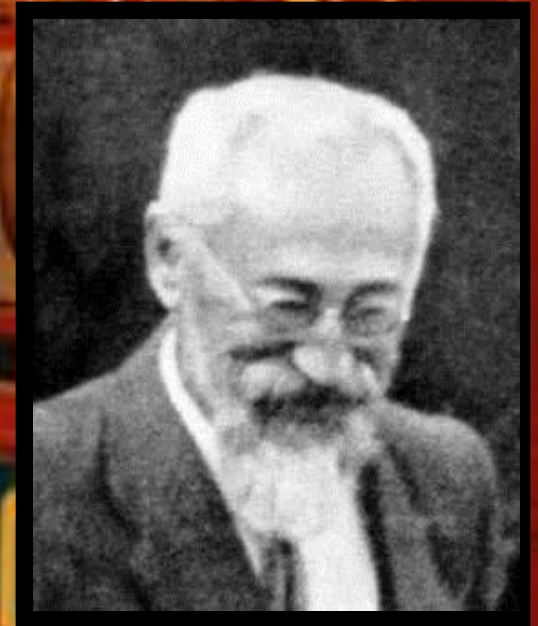
G*Power 3 offers five different types of statistical power analysis:

- A priori (sample size N is computed as a function of power level $1-\beta$, significance level α , and the to-be-detected population effect size)
- Compromise (both α and $1-\beta$ are computed as functions of effect size, N , and an error probability ratio $q = \beta/\alpha$)
- Criterion (α and the associated decision criterion are computed as a function of $1-\beta$, the effect size, and N)
- Post-hoc ($1-\beta$ is computed as a function of α , the population effect size, and N)
- Sensitivity (population effect size is computed as a function of α , $1-\beta$, and N)



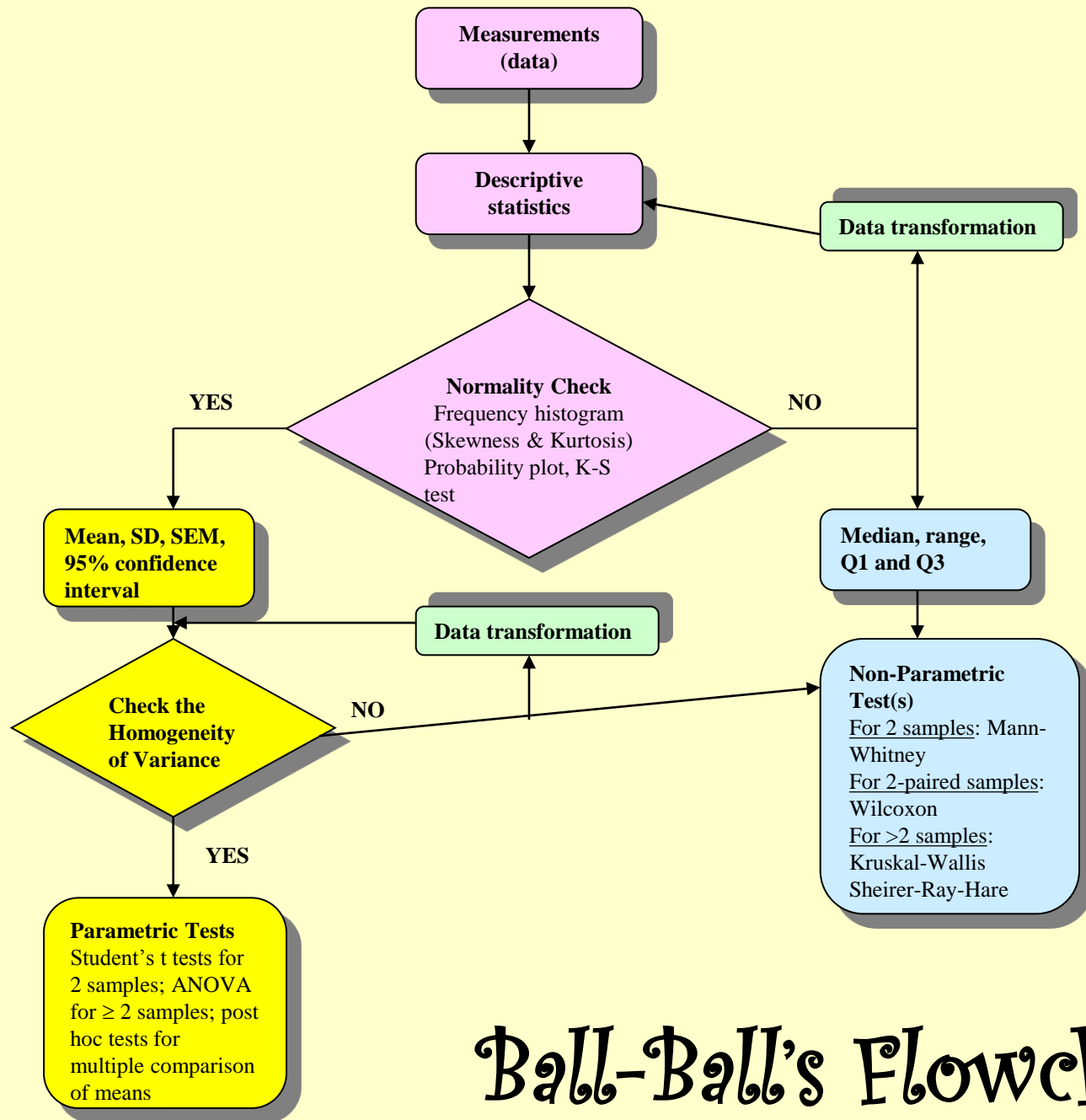
Questions about this website? Contact

Axel Buchner



Mr. Student = Mr. William Sealey Gosset (1876 – 1937)

Photo source: <http://www-groups.dcs.st-and.ac.uk/~history/PictDisplay/Gosset.html>



Ball-Ball's Flowchart

The difference between two sample means with limited data

- If $n < 30$, the above method gives an unreliable estimate of z
- This problem was solved by 'Student' who introduced the t -test early in the 20th century
- Similar to z -test, but instead of referring to z , a value of t is required (Table B3 in Zar)
- $df = 2n - 2$ for $n_1 = n_2$
- For all degrees of freedom below infinity, the curve appears leptokurtic compared with the normal distribution, and this property becomes extreme at small degrees of freedom.

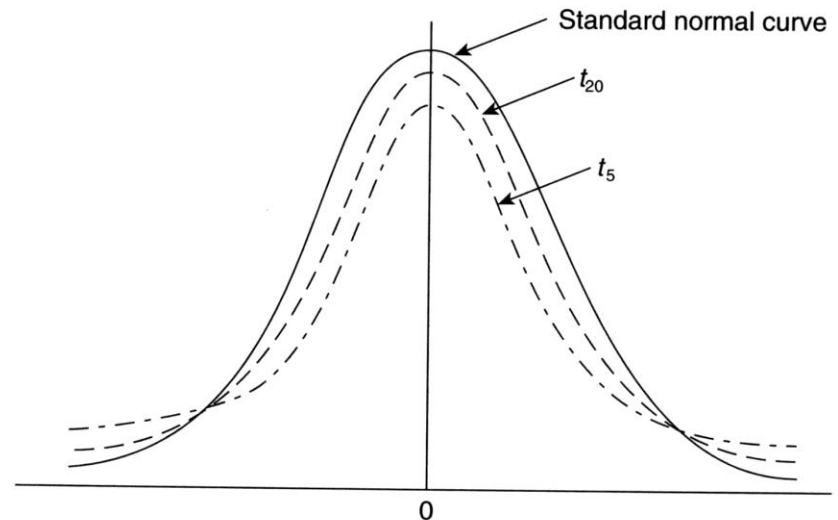
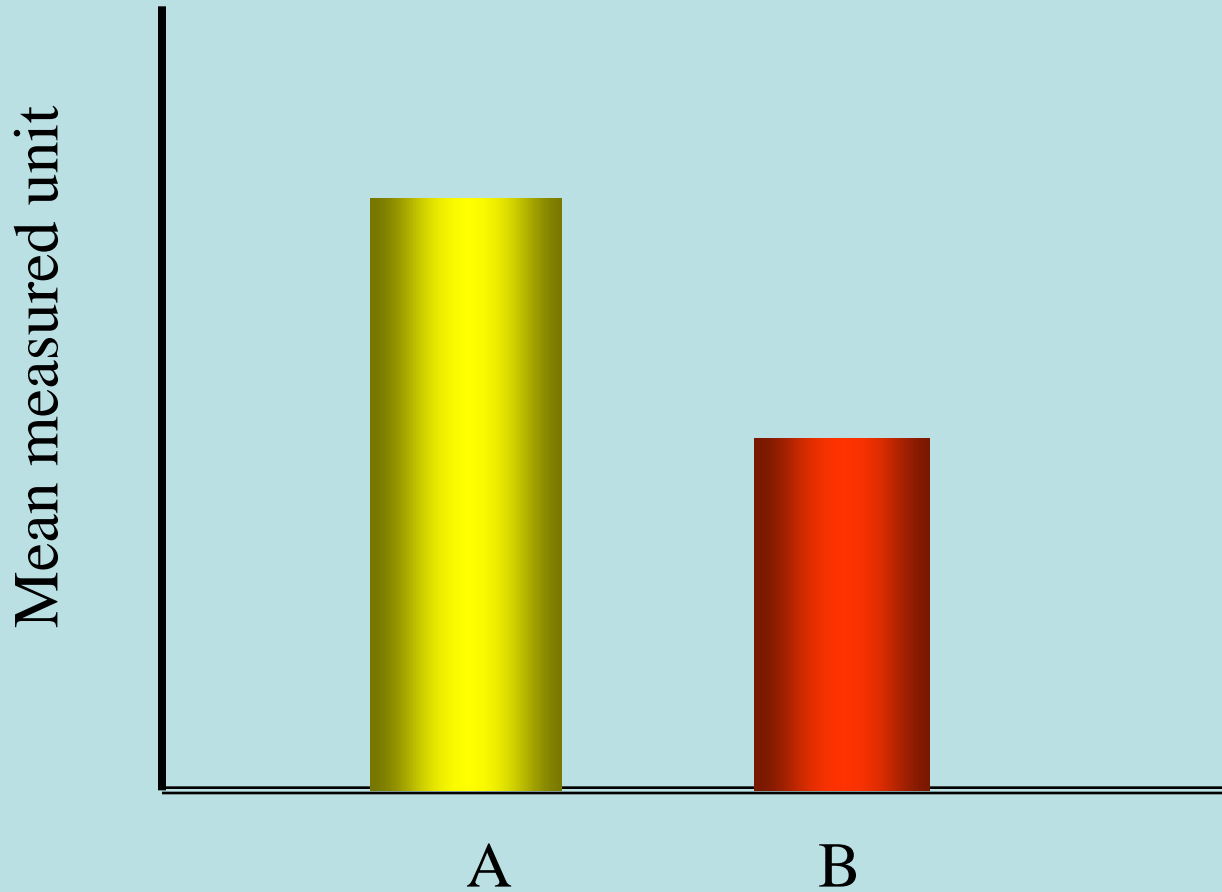
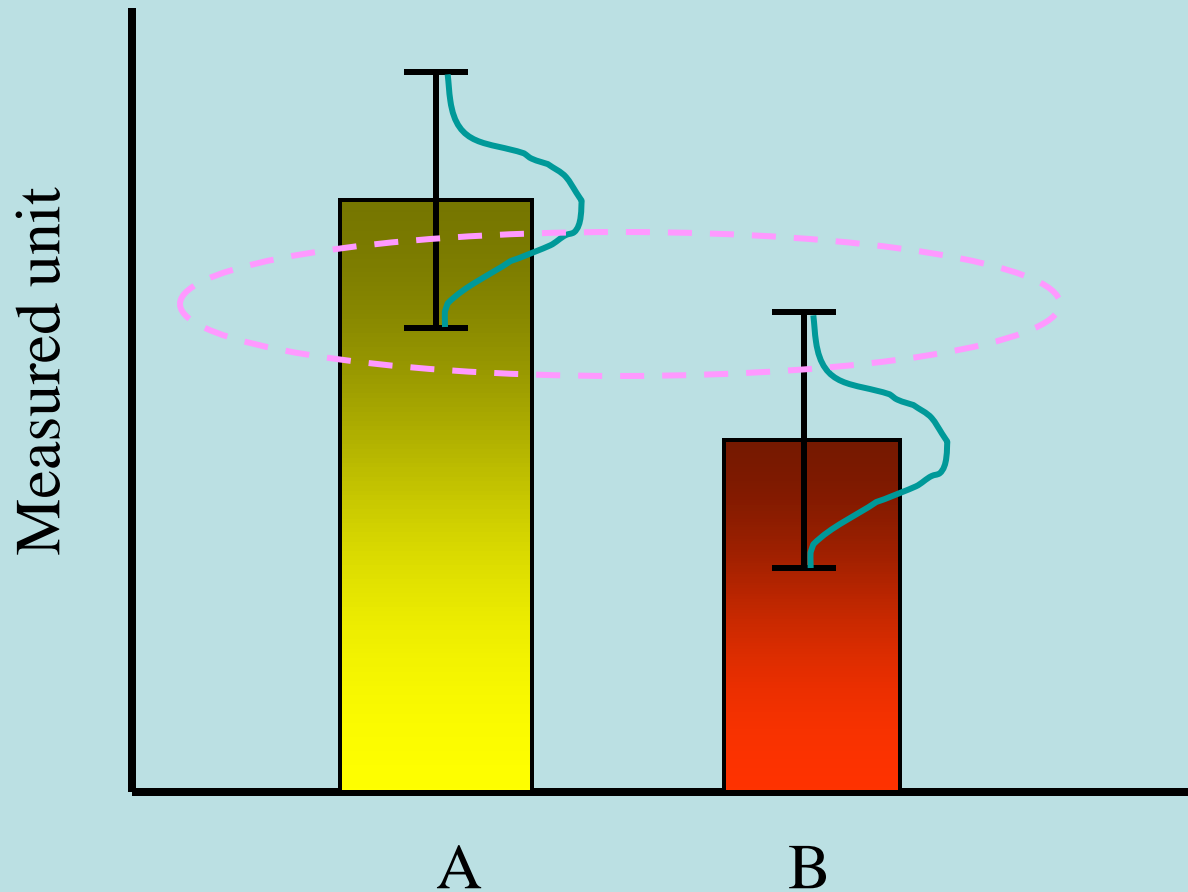


Figure 8.2 Two t curves for 5 and 20 degrees of freedom plotted with the normal curve.

Comparison of 2 Independent Samples

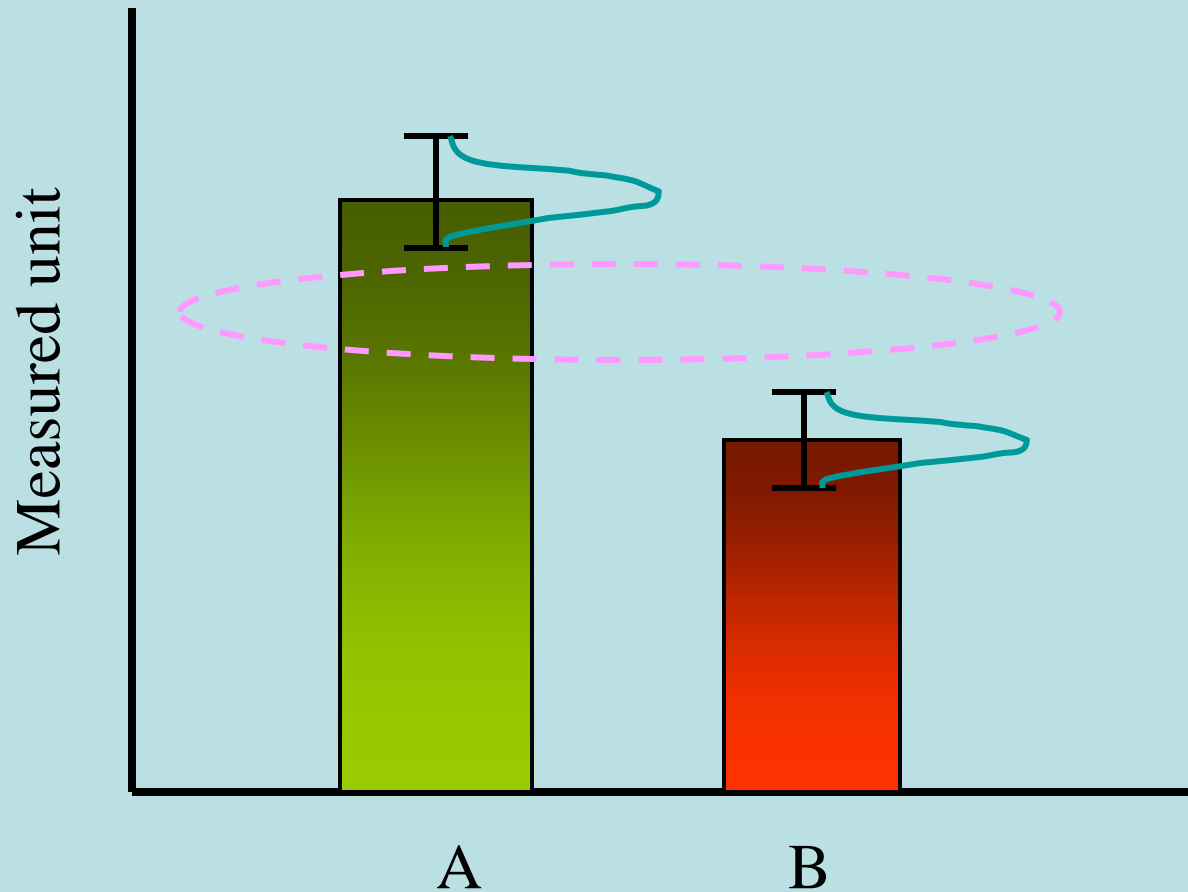


Comparison of 2 Independent Samples



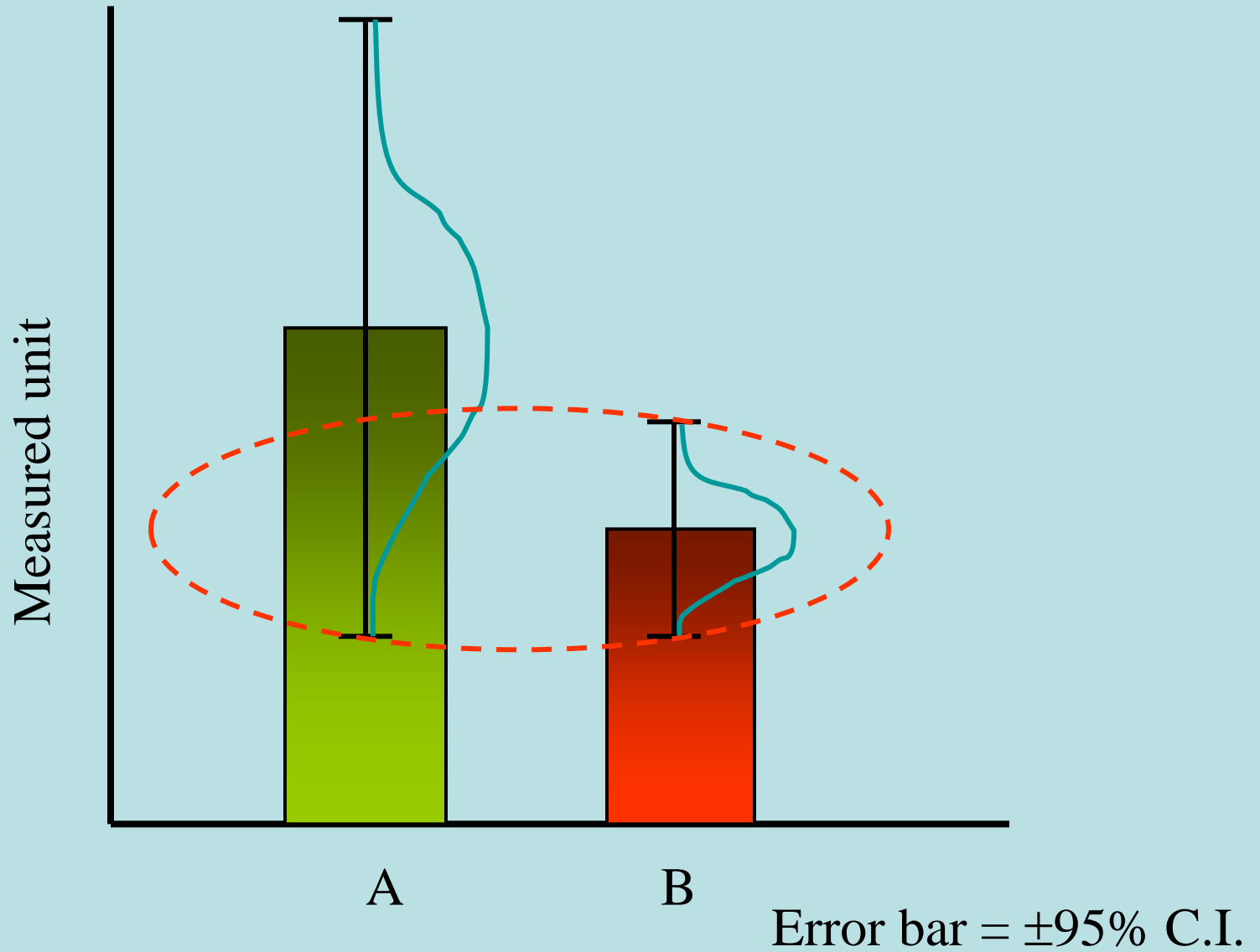
Error bar = $\pm 95\%$ C.I.

Comparison of 2 Independent Samples



Error bar = $\pm 95\%$ C.I.

Comparison of 2 Independent Samples



Power and sample size for Student's t test

- We can estimate the minimum sample size to use to achieve desired test characteristics:
- $n \geq (2S_p^2/\delta^2)(t_{\alpha, v} + t_{\beta(1), v})^2$
- where δ is the smallest population difference we wish to detect: $\delta = \mu_1 - \mu_2$
- Required sample size depends on δ , population variance (σ^2), α , and power ($1-\beta$)
- If we want to detect a very small δ , we need a larger sample.
- If the variability within samples is great, a large n is required. The results of pilot study or pervious study of this type would provide such an information.

Estimation of minimum detectable difference

$$n \geq (2S_P^2/\delta^2)(t_{\alpha, \nu} + t_{\beta(1), \nu})^2$$

- The above equation can be rearranged to ask how small a population difference (δ) is detectable with a given sample size:

$$\delta \geq [\sqrt{(2S_P^2/n)}](t_{\alpha, \nu} + t_{\beta(1), \nu})$$

Some Notes about Effect Size



- If aliens were to land on earth, how long would it take for them to realise that, on average, human males are taller than females?
- The answer relates to the effect size (ES) of the difference in height between men and women.
- The larger the ES, the quicker they would suspect that men are taller.
- Cohen (1992) suggested where 0.2 is indicative of a small ES, 0.5 a medium ES and 0.8 a large ES.

A Student's *t* Test with $n_a = n_b$

- e.g. The chemical oxygen demand (COD) is measured at two industrial effluent outfalls, a and b, as part of consent procedure. Test the null hypothesis: $H_0: \mu_a = \mu_b$ while $H_A: \mu_a \neq \mu_b$

a	b
3.48	3.89
2.99	3.19
3.32	2.80
4.17	4.31
3.78	3.42
4.00	3.41
3.20	3.55
4.40	2.40
3.85	2.99
4.52	3.08
3.09	3.31
3.62	4.52

<i>n</i>	12	12
mean	3.701	3.406
S^2	0.257	0.366

SS = sum of square = $S^2 \times v$

$$s_p^2 = (SS_1 + SS_2) / (v_1 + v_2) = [(0.257 \times 11) + (0.366 \times 11)] / (11 + 11) = 0.312$$

$$s_{X_1 - X_2} = \sqrt{(s_p^2/n_1 + s_p^2/n_2)} = \sqrt{(0.312/12) \times 2} = 0.228$$

$$t = (\bar{X}_1 - \bar{X}_2) / s_{X_1 - X_2} = (3.701 - 3.406) / 0.228 = 1.294$$

$$df = 2n - 2 = 22$$

$$t_{\alpha = 0.05, df = 22, 2-tailed} = 2.074 > |t_{observed}| = 1.294, p > 0.05$$

The calculated *t*-value < the critical *t* value.

Thus, accept H_0 .

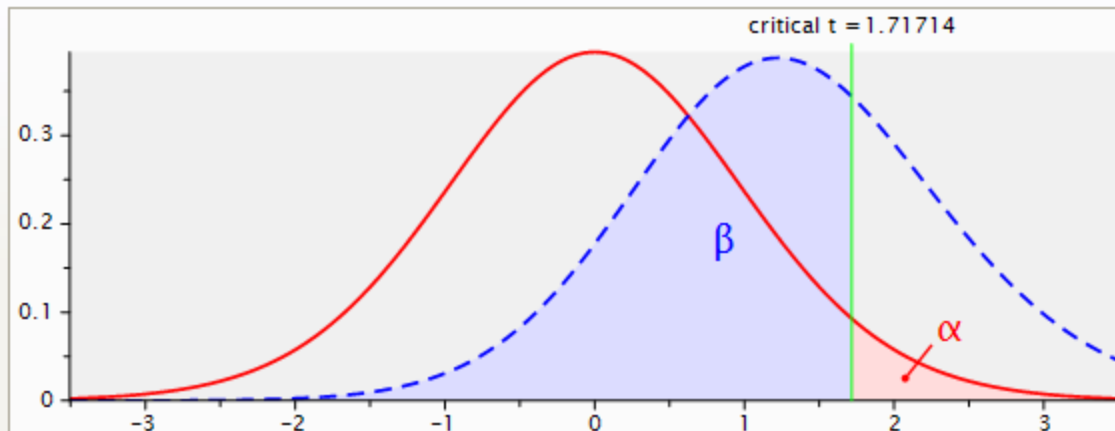
Need to check Power.



Remember to always check the homogeneity of variance before running the *t* test.

File Edit View Tests Calculator Help

Central and noncentral distributions Protocol of power analyses



Test family

t tests

Statistical test

Means: Difference between two independent means (two groups)

Type of power analysis

Post hoc: Compute achieved power - given α , sample size, and effect size

Input Parameters

Tail(s) One

Determine =>

Effect size d 0.5158052

 α err prob 0.05

Sample size group 1 12

Sample size group 2 12

Output Parameters

Noncentrality parameter δ 1.263460

Critical t 1.717144

Df 22

Power ($1 - \beta$ err prob) 0.337150

X-Y plot for a range of values

Calculate

 $n1 \neq n2$

Mean group 1 0

Mean group 2 1

SD σ within each group 0.5 $n1 = n2$

Mean group 1 3.70

Mean group 2 3.41

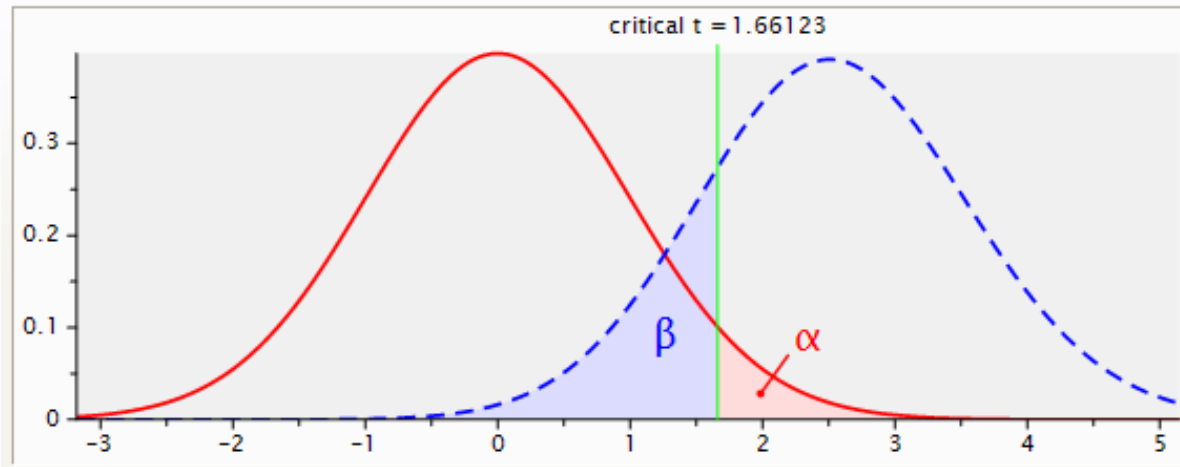
SD σ group 1 0.51SD σ group 2 0.61

Calculate

Effect size d 0.5158052

Calculate and transfer to main window

Close



Test family

t tests

Statistical test

Means: Difference between two independent means (two groups)

Type of power analysis

A priori: Compute required sample size - given α , power, and effect size

Input Parameters

Determine =>

Tail(s)	One
Effect size d	0.5158052
α err prob	0.05
Power (1- β err prob)	0.8
Allocation ratio N2/N1	1

Output Parameters

Noncentrality parameter δ	2.526919
Critical t	1.661226
Df	94
Sample size group 1	48
Sample size group 2	48
Total sample size	96
Actual power	0.806154

X-Y plot for a range of values

Calculate

critical t = 1.71714

Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

Type of power analysis: Post hoc: Compute achieved power - given α , sample size, and effect size

Input Parameters: Tail(s) One, Effect size d 0.5158052, α err prob 0.05, Sample size group 1 12, Sample size group 2 12

Output Parameters: Noncentrality parameter δ 1.263460, Critical t 1.717144, Df 22, Power (1- β err prob) 0.937150

$$N = 2 \times 48 = 96$$

- Growth of 8 months old non-transgenic and transgenic tilapia was determined by measuring the body mass (wet weight). Since transgenic fish cloned with growth hormone (GH) related gene *opAFPcsGH* are known to grow faster in other fish species (Rahman et al. 2001), it is hypothesized that $H_A: \mu_{\text{transgenic}} > \mu_{\text{non-transgenic}}$ while the null hypothesis is given as $H_0: \mu_{\text{transgenic}} \leq \mu_{\text{non-transgenic}}$

Transgenic tilapia carrying *opAFP-csGH* gene

(Rahman et al., 1998)

8 months old



Non-transgenic



Transgenic

$H_0: \mu_{\text{transgenic}} \leq \mu_{\text{non-transgenic}}$

$H_A: \mu_{\text{transgenic}} > \mu_{\text{non-transgenic}}$

Given that mass (g) of tilapia are normally distributed.

transgenic	non-transgenic
700	305
680	280
500	275
510	250
670	490
670	275
620	275
650	300

$$s_p^2 = (SS_1 + SS_2) / (u_1 + u_2) = 5913.4$$

$$s_{X_1 - X_2} = \sqrt{(s_p^2/n_1 + s_p^2/n_2)} = 38.45$$

$$t = (\bar{X}_1 - \bar{X}_2) / s_{X_1 - X_2} = 8.29$$

$$df = 2n - 2 = 14$$

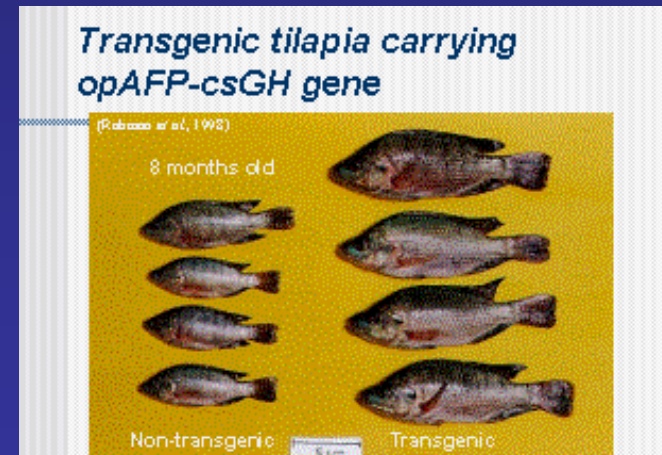
$$t_{\alpha = 0.05, df = 14, 1\text{-tailed}} = 1.761 \ll 8.29 ; p < 0.001$$

The t -value is greater than the critical t value.

Thus, reject H_0 .

n	8	8
mean	625.0	306.25
S^2	6028.6	5798.2

If we are going to repeat this study, can we reduce the sample size? How many?



G*Power 3.0.10

File Edit View Tests Calculator Help

Central and noncentral distributions Protocol of power analyses

critical t = 1.76131

Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

Type of power analysis: Post hoc: Compute achieved power - given α , sample size, and effect size

Input Parameters

Determine =>

Tail(s): One

Effect size d: 4.1468422

α err prob: 0.05

Sample size group 1: 8

Sample size group 2: 8

Output Parameters

Noncentrality parameter δ : 8.293684

Critical t: 1.761310

Df: 14

Power (1- β err prob): 1.000000

X-Y plot for a range of values Calculate

n1 != n2

Mean group 1: 0

Mean group 2: 1

SD σ within each group: 0.5

n1 = n2

Mean group 1: 625

Mean group 2: 306.3

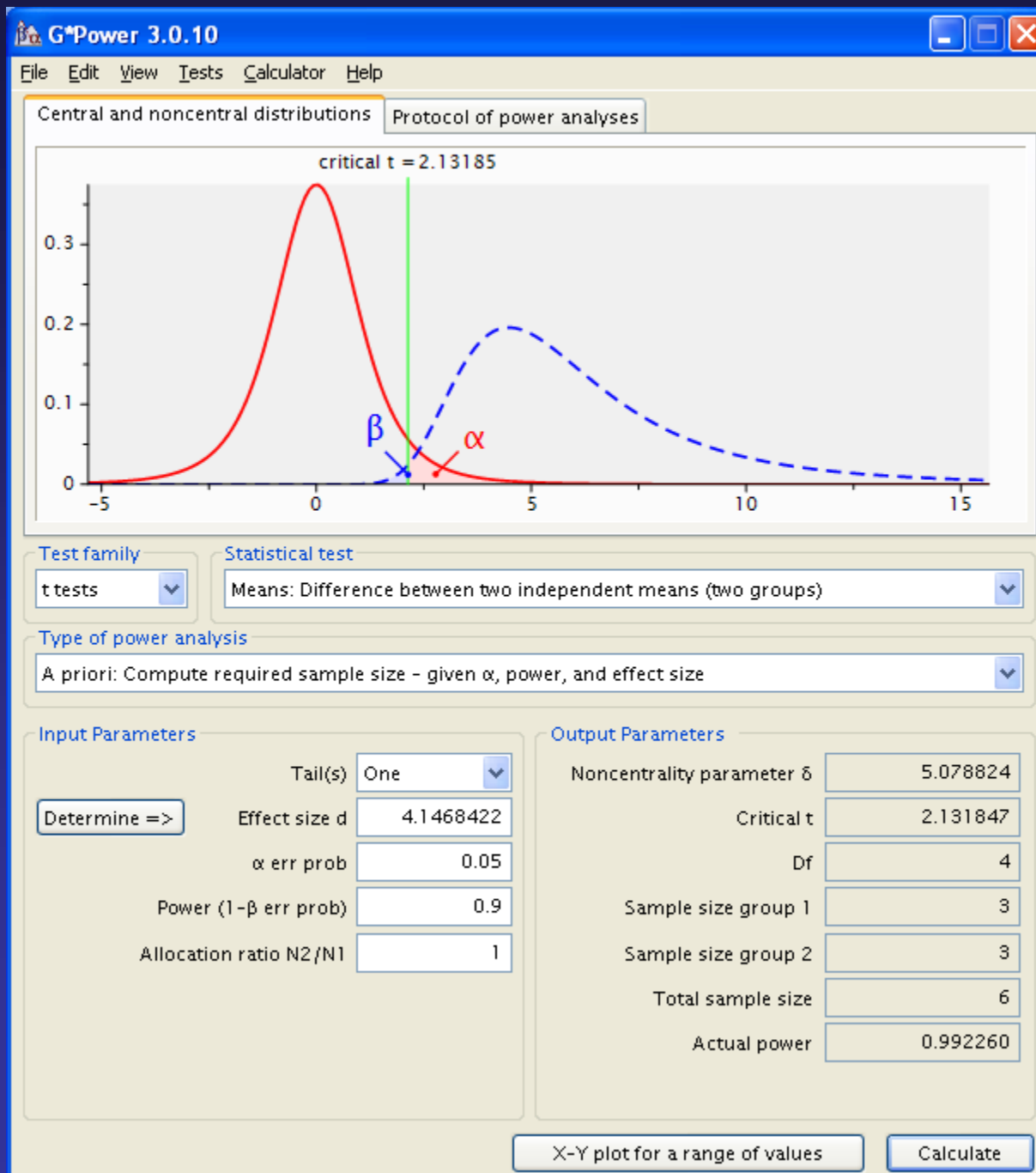
SD σ group 1: 77.6

SD σ group 2: 76.1

Calculate Effect size d: 4.146842

Calculate and transfer to main window

Close



$$N = 2 \times 3 = 6$$

Comparison of [PBDEs] in tissues of transplanted mussels collected from 6 sites along a anticipated pollution gradient

- Expected that high [PBDEs] in samples from polluted sites than clean sites
- H_a : unequal means
- H_o : equal means



[PBDEs] in mussels from various sites (ng/g)

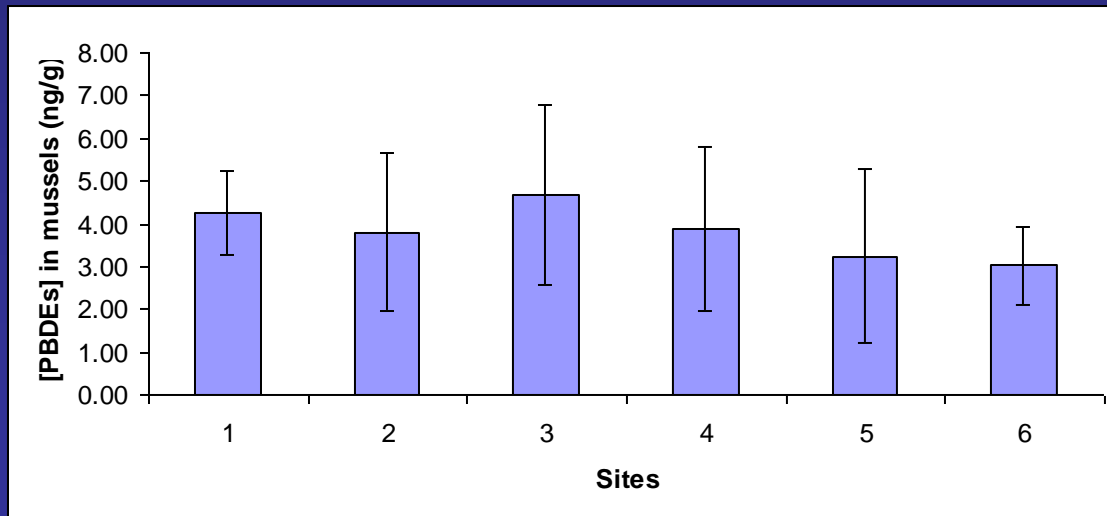
P1	P2	P3	P4	C5	C6
4.25	3.50	7.20	4.00	0.50	2.50
3.45	3.80	6.50	5.50	5.50	2.50
4.75	4.70	4.00	2.20	2.25	2.30
5.60	1.01	2.20	1.70	3.00	3.30
3.20	6.00	3.50	6.00	5.00	4.50

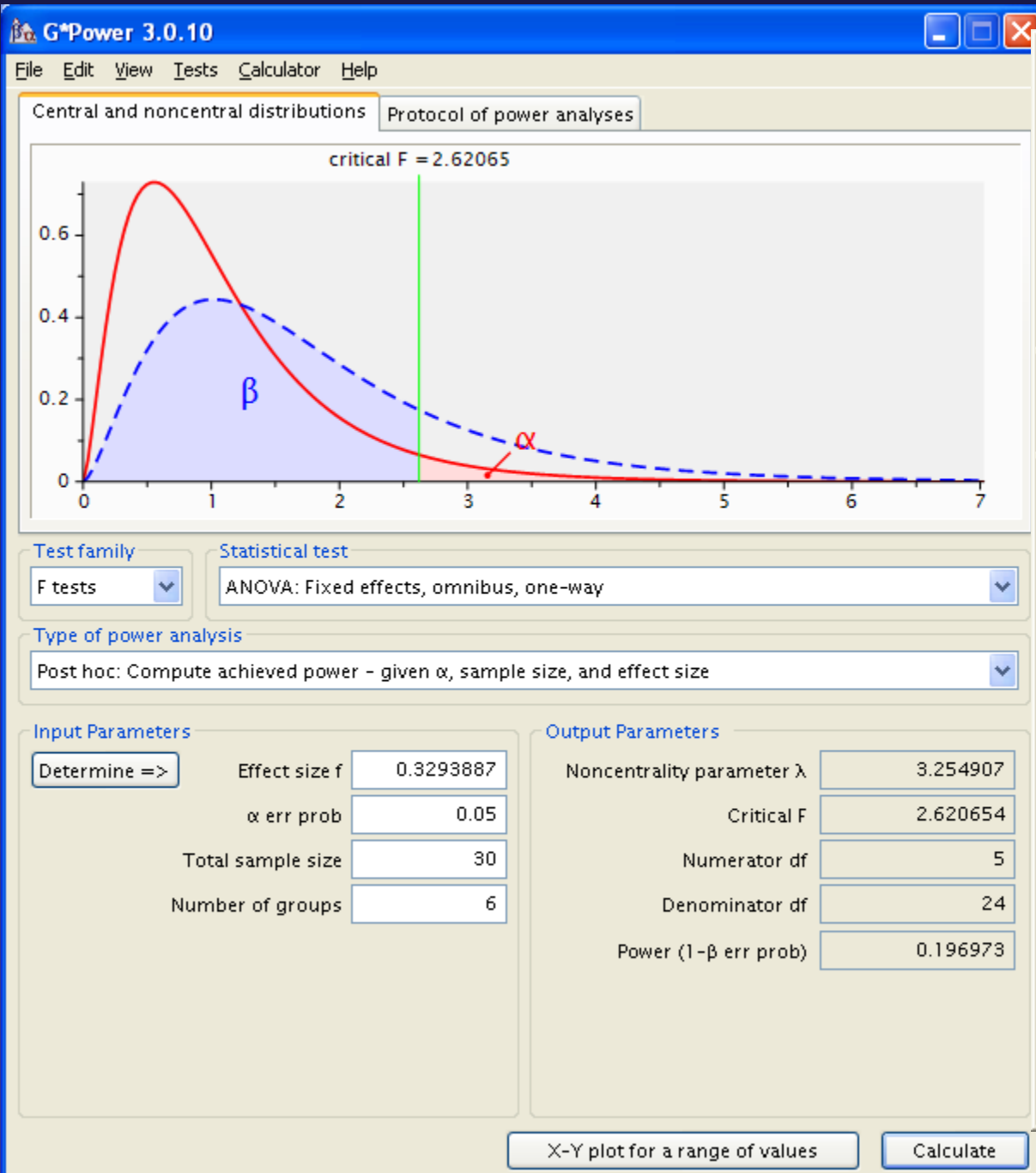
Comparison of [PBDEs] in tissues of transplanted mussels collected from 6 sites along a anticipated pollution gradient

ANOVA

Source of Variation	SS	df	MS	F	P-value
Between Groups	9.465417	5	1.893083	0.650981	0.663547
Within Groups	69.79308	24	2.908045		
Total	79.2585	29			

common SD 1.705299
= $\sqrt{2.908}$





Select procedure

Effect size from means

Number of groups: 6

SD σ within each group: 1.705299

Group	Mean	size
1	4.25	5
2	3.802	5
3	4.68	5
4	3.88	5
5	3.25	5
6	3.02	5

Equal n: 5

Total sample size: 30

Calculate Effect size f: 0.3293887

Calculate and transfer to main window

Close

X-Y plot for a range of values

Calculate

G*Power 3.0.10

File Edit View Tests Calculator Help

Central and noncentral distributions Protocol of power analyses

critical F = 2.28985

Test family: F tests

Statistical test: ANOVA: Fixed effects, omnibus, one-way

Type of power analysis: A priori: Compute required sample size - given α , power, and effect size

Input Parameters

Determine =>

Effect size f	0.3293887
α err prob	0.05
Power (1- β err prob)	0.80
Number of groups	6

Output Parameters

Noncentrality parameter λ	13.670611
Critical F	2.289851
Numerator df	5
Denominator df	120
Total sample size	126
Actual power	0.808094

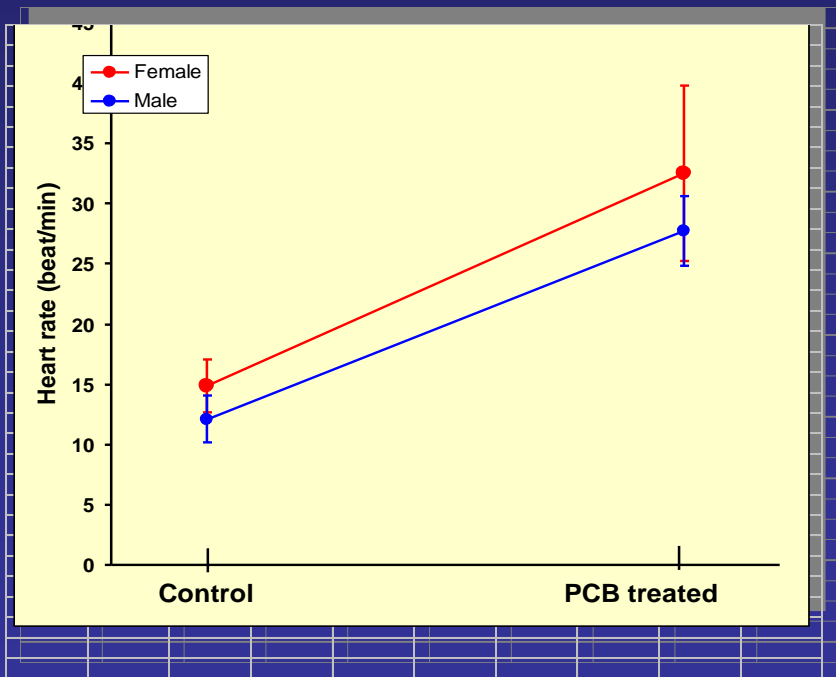
X-Y plot for a range of values Calculate

$$N = 6 \times 21 = 126$$

Example 4

2-Way ANOVA: Effects of dietary PCBs and sex on heart rate in birds

Source of variance	SS	DF	MS = SS/DF	F	F critical, 0.05(1), 1, 16	P
Total	1827.7	19				
Cells	1461.3	3				
PCB	1386.1	1	1386.10	60.53	4.49	< 0.001
Sex	70.31	1	70.31	3.07	4.49	> 0.05
PCB x Sex	4.900	1	4.90	0.21	4.49	> 0.05
Within cells (error)	366.4	16	22.90			



- There was a significant effect of chemical treatment on the heart rate in the birds ($P < 0.001$).
- There was no interaction between sex and hormone treatment while the sex effect was not significant (likely due to inadequate power).
- $N = 2 \times 2 \times 4$

G*Power 3.0.10

File Edit View Tests Calculator Help

Central and noncentral distributions Protocol of power analyses

Source of variance	SS	DF	MS = SS/DF	F	F critical, 0.05(1), 1, 16	P
Total	1827.7	19				
Cells	1461.3	3				
PCB	1386.1	1	1386.10	60.53	4.49	< 0.001
Sex	70.31	1	70.31	3.07	4.49	> 0.05
PCB x Sex	4.900	1	4.90	0.21	4.49	> 0.05
Within cells (error)	366.4	16	22.90			

Test family: F tests

Statistical test: ANOVA: Fixed effects, special, main effects and interactions

Type of power analysis: A priori: Compute required sample size - given α , power, and effect size

Input Parameters

Determine =>

Effect size f: 1.7522288

α err prob: 0.05

Power (1- β err prob): 0.9

Numerator df: 1

Number of groups: 2

Output Parameters

Noncentrality parameter λ : 21.492140

Critical F: 6.607891

Denominator df: 5

Total sample size: 7

Actual power: 0.954634

X-Y plot for a range of values Calculate

For the sex effect

Variance for sex = 70.31

Error variance = 22.90

N should be $2 \times 2 \times 7 = 28$

From Variances

Variance explained by special effect: 70.31

Error variance: 22.9

Direct

Partial η^2 : 0.7543182

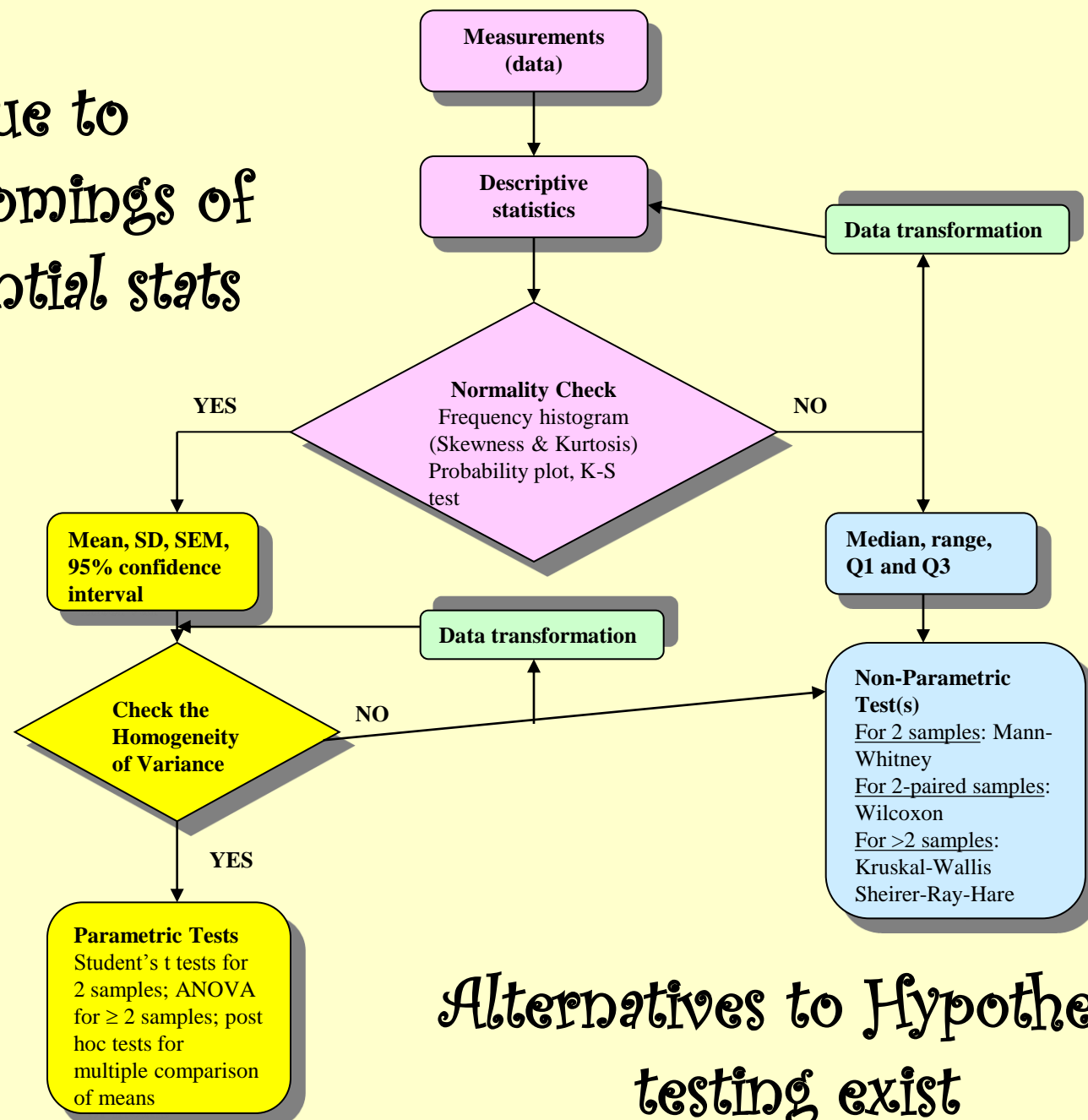
Calculate

Effect size f: 1.752229

Calculate and transfer to main window

Close

Due to shortcomings of inferential stats



Alternatives to Hypothesis testing exist

There are problems in the conventional hypothesis testing:

<http://www.youtube.com/watch?v=ez4DgdurRPg>

YouTube - Bayes' Formula

<http://www.youtube.com/watch?v=pPTLK5hFGnQ&feature=channel>

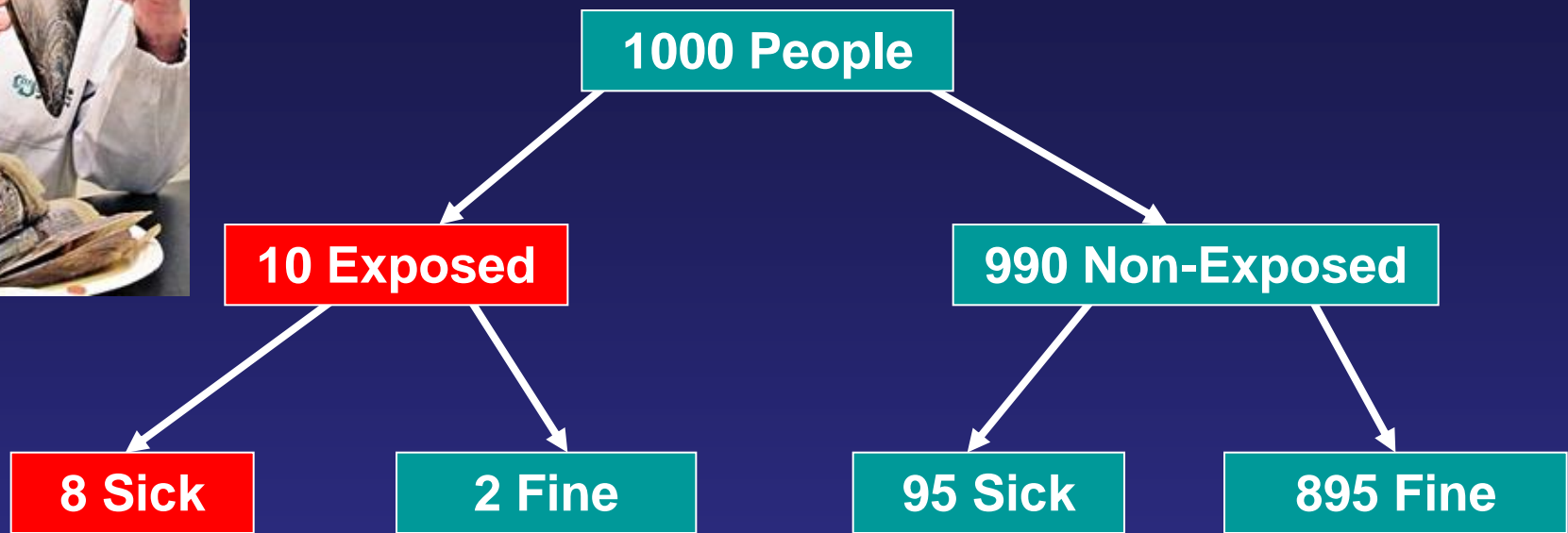
By bionicturtledotcom

YouTube - Bayes' Theorem - Part 2

<http://www.youtube.com/watch?v=bcALcVmLva8&feature=related>

By westofvideo

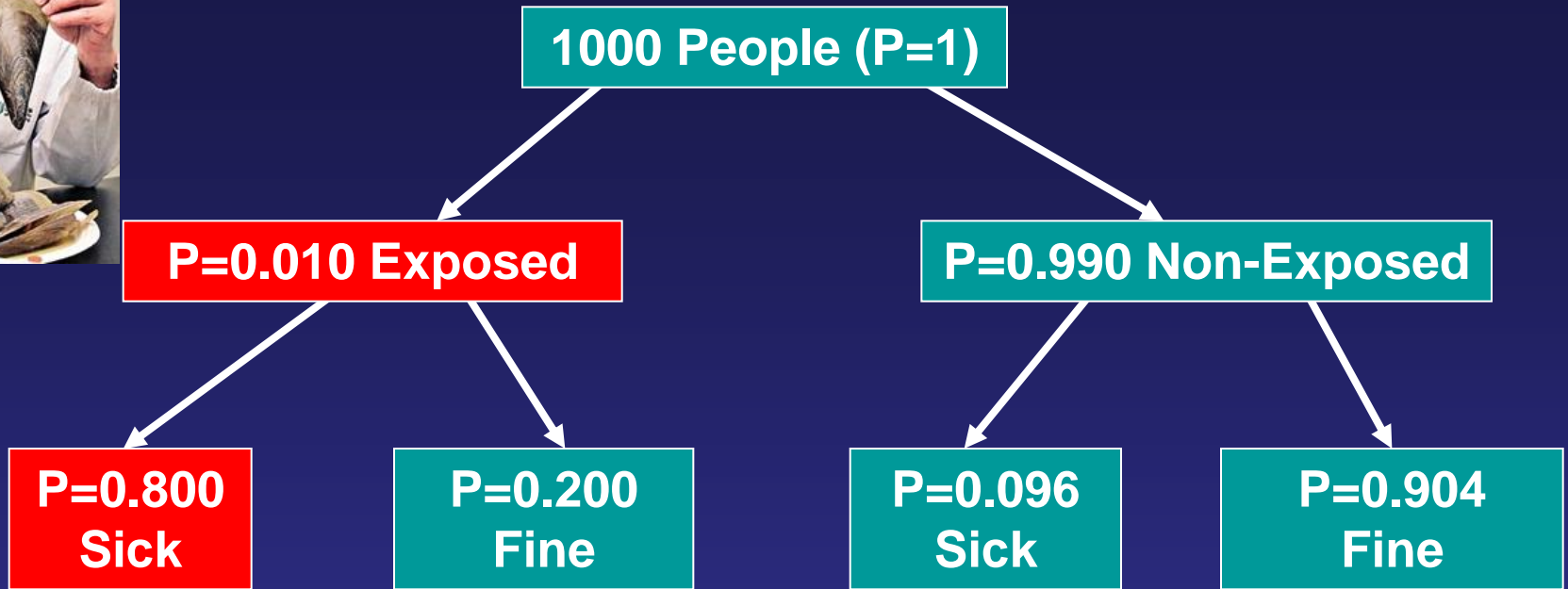
A Simple Example



What is the chance to be sick after eating scallops (i.e. exposed)?

$$\begin{aligned} \text{Probability} &= 8 \text{ exposed with illness} / (\text{total of } 103 \text{ with illness}) \\ &= 0.078 \end{aligned}$$

A Probability Diagram – Bayesian Approach



What is the chance to be sick after eating scallops (i.e. exposed)?

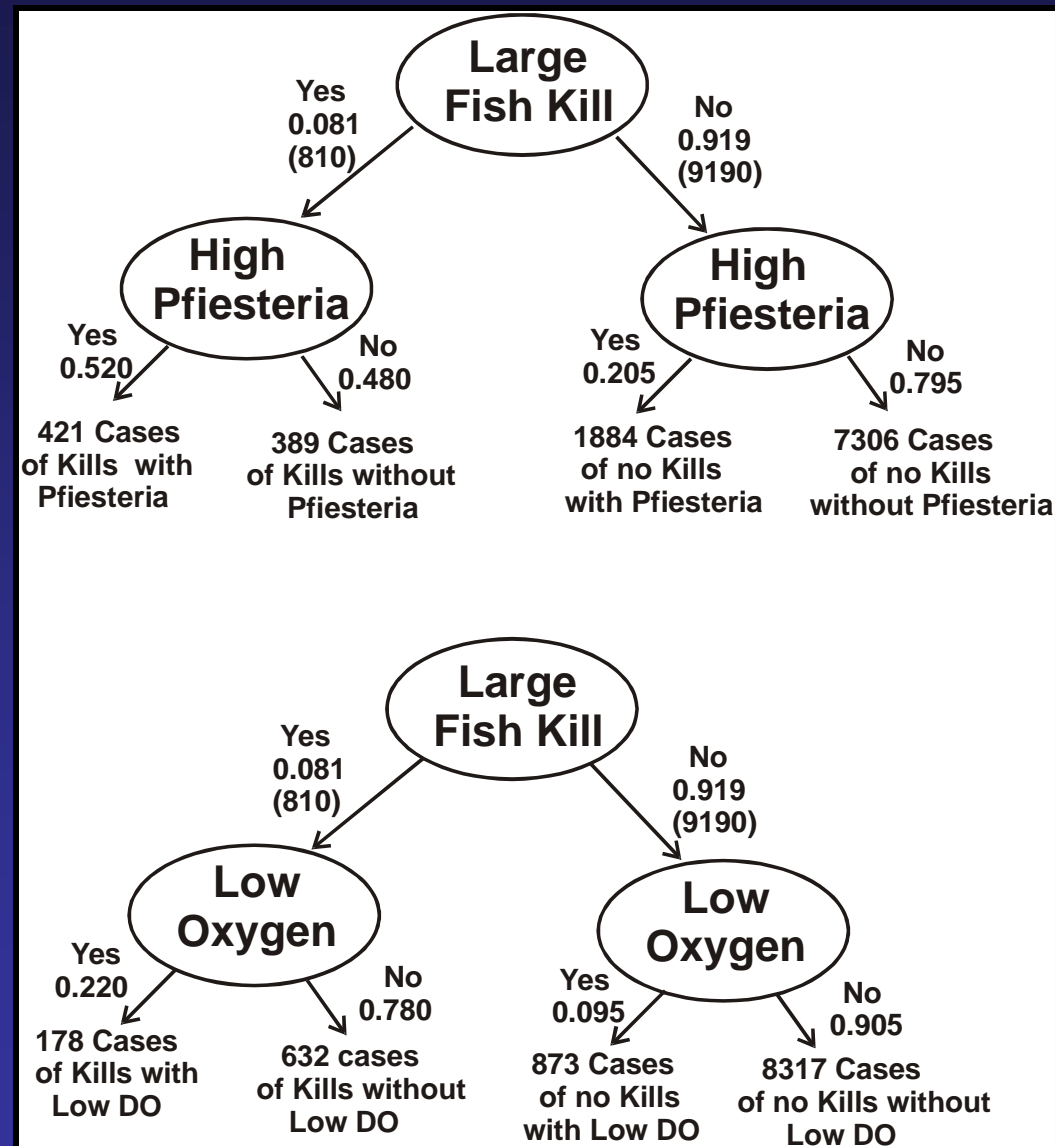
$$\begin{aligned} P(\text{Exposed}|\text{Sick}) &= \frac{P(\text{Exposed}) P(\text{Sick}|\text{Exposed})}{P(\text{Sick})} \\ &= \frac{(0.010)(0.800)}{(0.010*0.800+0.990*0.096)} = 0.078 \end{aligned}$$

Example: Fishkills

This figure illustrates how the natural frequency approach can lead to these same inferences using the $p(\text{Pfiesteria})$ estimate of 0.205. From the figure, the likelihood ratio can be calculated.

Mike Newman, et al. 2007. Coastal and estuarine ecological risk assessment: the need for a more formal approach to stressor identification. Hydrobiologia 577: 31-40.

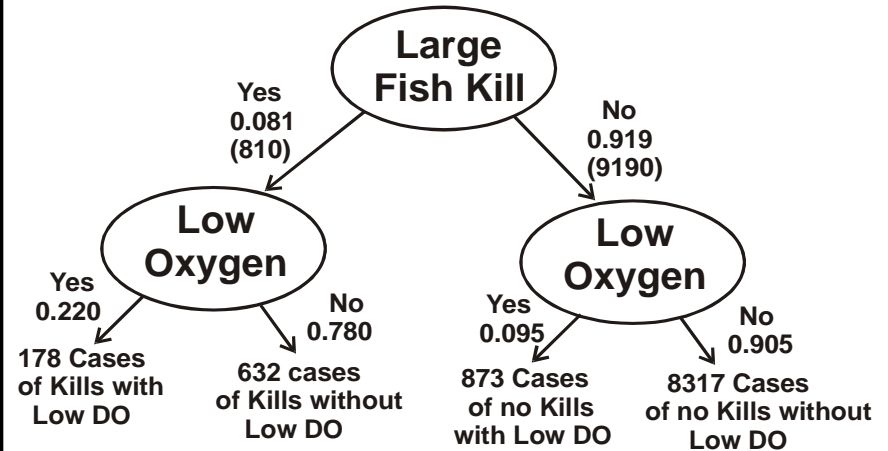
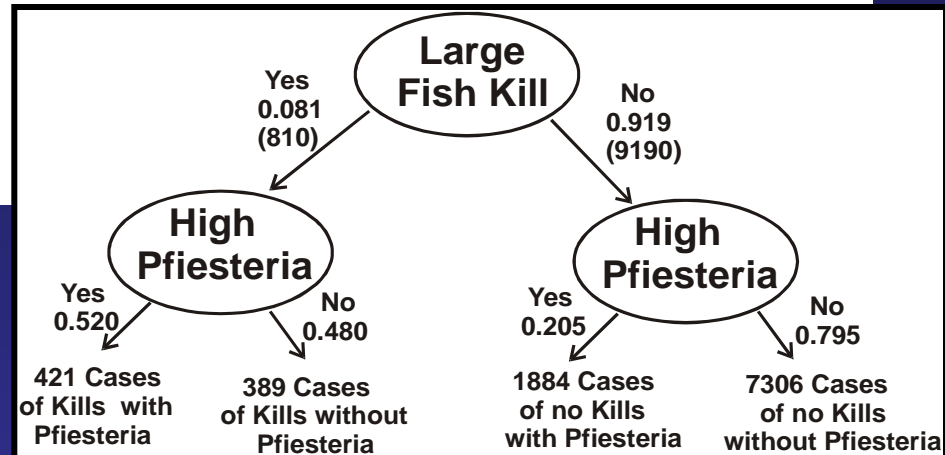
Credit: M.C. Newman



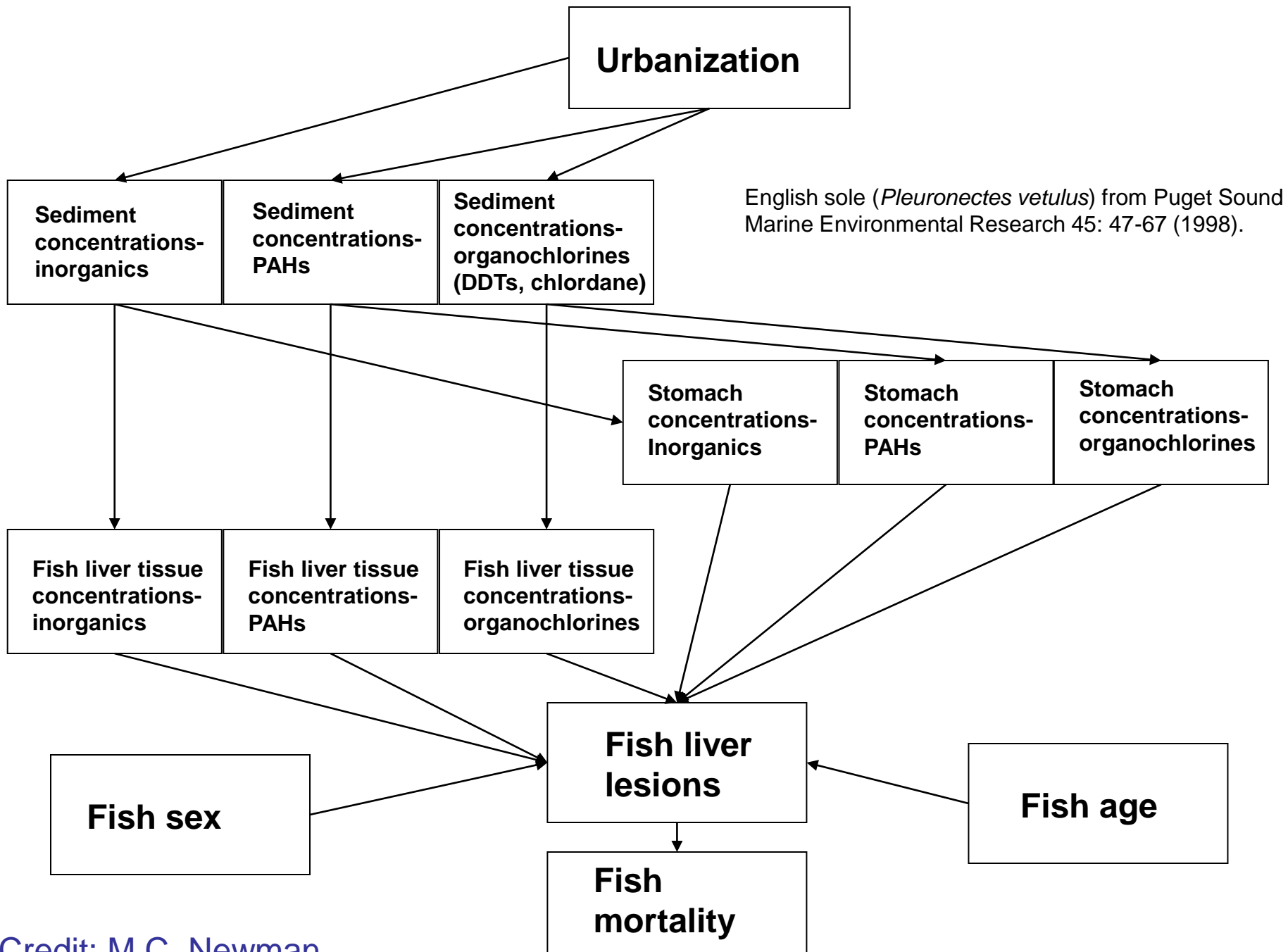
$$\frac{421 \text{ Cases of large fish kills with high Pfiesteria levels}}{1884 \text{ Cases of no large fish kills with high Pfiesteria levels}} = 0.22346$$

$$\frac{178 \text{ Cases of large fish kills with low dissolved oxygen concentrations}}{873 \text{ Cases of no large fish kills with low dissolved oxygen concentrations}} = 0.20389$$

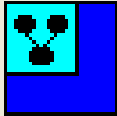
$$\text{Likelihood Ratio} = \frac{0.22346}{0.20389} = 1.096$$



$$\frac{p(\text{Fish Kill} \mid \text{Pfiesteria})}{p(\text{Fish Kill} \mid \text{Low DO})} = 1.095$$



Software Exists for More Complex Situations



Netica 1.12

for Windows 95 and Windows NT 4.0

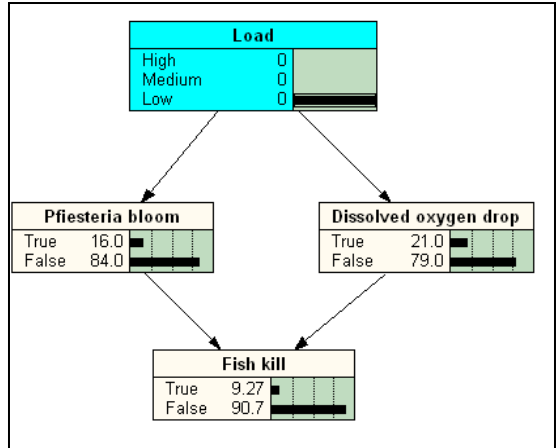
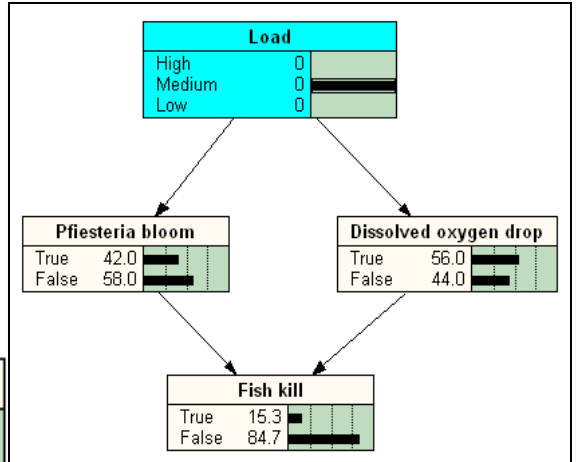
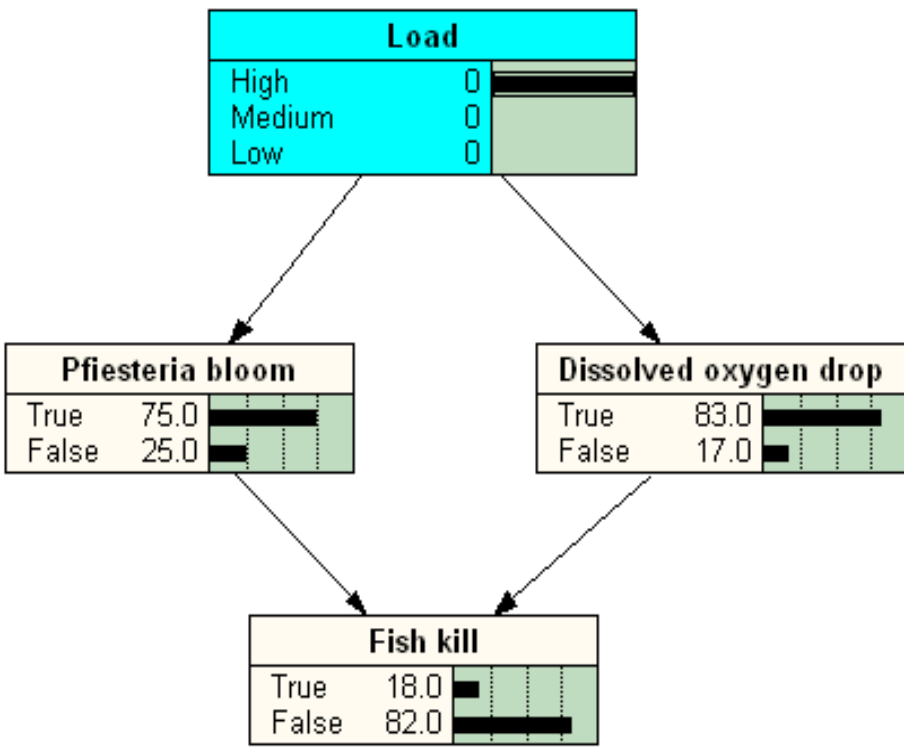
Netica is a complete software package to work with (Bayesian) belief networks, decision networks and influence diagrams.

A linkable software library providing much of the same functionality is also available from Norsys.

For more information on Netica, or to get the manual or latest version, see: www.norsys.com

Norsys and Netica are trademarks of Norsys Software Corp.
Copyright © 1990-98 by Norsys Software Corp.

Okay



Credit: M.C. Newman

Supplemental Readings

- Aven, T. & J.T. Kvaløy, 2002. Implementing the Bayesian paradigm in risk analysis. *Reliability Engineering and System Safety* 78: 195-201.
- Bacon, P.J., J.D. Cain & D.C. Howard, 2002. Belief network models of land manager decisions and land use change. *Journal of Environmental Management* 65: 1-23.
- Belousek, D.W., 2004. Scientific consensus and public policy: the case of *Pfiesteria*. *Journal Philosophy, Science & Law* 4: 1-33.
- Borsuk, M.E., 2004. Predictive assessment of fish health and fish kills in the Neuse River estuary using elicited expert judgment, *Human and Ecological Assessment* 10: 415-434.
- Borsuk, M.E., D. Higdon, C.A. Stow & K.H. Reckhow, 2001. A Bayesian hierarchical model to predict benthic oxygen demand from organic matter loading in estuaries and coastal zones. *Ecological Modelling* 143: 165-181.
- Garbolino, P. and F. Taroni. 2002. Evaluation of scientific evidence using Bayesian networks. *Forensic Sci Intern.* 125:149-155.
- Newman, M.C. and D. Evans. 2002. Causal inference in risk assessments: Cognitive idols or Bayesian theory? In: *Coastal and Estuarine Risk Assessment*. CRC Press LLC, Boca Raton, FL, pp. 73-96.
- Newman, M.C., Zhao, Y., and J.F. Carriger. 2007. Coastal and estuarine ecological risk assessment: the need for a more formal approach to stressor identification. *Hydrobiologia* 577: 31-40.
- Uusitalo, L. 2007. Advantages and challenges of Bayesian networks in environmental modeling. *Ecol. Modelling* 203:312-318.

YouTube - Bayes' Theorem Introduction

http://www.youtube.com/watch?v=0NGmrwu_BkY&feature=related

By westofvideo

Error Type (Type I & II)

<http://www.youtube.com/watch?v=taEmnrTxuzo&feature=related>

By bionicturtledotcom