Medical StaTISTICS

Mr. Kongmany Chaleunvong

GFMER - WHO - UNFPA - LAO PDR

Training Course in Reproductive Health Research

Vientiane, 22 October 2009

Why Medical Statistics?

- Medicine is a quantitative science but not exact
  - Not like physics or chemistry
- Statistics is about handling and quantifying variation and uncertainty
- Statistics is the field of science concerned with the collection, classification, summarising & interpretation of numerical data
- Biostatistics: application of statistical methods in biological sciences to deal with living things

Medical Statistics

Application of statistics in medicine

- To compare efficacy of a drug
- The % cured, relieved or died in experiment
- To find an association B/W 2 attributes
- In epidemiological studies

Variation

Variation characterises much of medicine

- Humans differ in response to exposure to adverse effects

  Example: not every smoker dies of lung cancer

  some non-smokers die of lung cancer

- Humans differ in response to treatment

  Example: penicillin does not cure all infections

- Humans differ in disease symptoms

  Example: sometimes cough and sometimes wheeze are presenting features for asthma

Probability

Thus:

Diagnosis and treatment are probabilistically based

We talk of the *risks* due to exposure, the *chances* of cure, the *probability* that a patient has a disease

Data

- **Quantitative :** data with characteristics which have a magnitude or which can be measured.

  E.g.: height, weight, B.P., pulse rate, etc.

- **Qualitative:** data where the characteristics have no magnitude of size & which can't be measured. Do not possess numerical values.

  E.g.: colour of hair, gender, blood group, etc.

Types of Data

- Discrete Data-limited number of choices
  - Binary: two choices (yes/no)
    - Dead or alive
    - Disease-free or not
  - Categorical: more than two choices, not ordered
    - Race
    - Age group
  - Ordinal: more than two choices, ordered
    - Stages of a cancer
    - Likert scale for response
      - E.g., strongly agree, agree, neither agree or disagree, etc.

Types of data

- Continuous data
  - Theoretically infinite possible values (within physiologic limits), including fractional values
    - Height, age, weight
  - Can be interval
    - Interval between measures has meaning
    - Ratio of two interval data points has no meaning
    - Temperature in Celsius, day of the year)
  - Can be ratio
    - Ratio of the measures has meaning
    - Weight, height

Types of Data

- Why important?
- The type of data defines:
  - The summary measures used
    - Mean, Standard deviation for continuous data
    - Proportions for discrete data
  - Statistics used for analysis:
    - Examples:
      - T-test for normally distributed continuous
      - Wilcoxon Rank Sum for non-normally distributed continuous

The outcome determine statistics

Continuous

Mean
Median

Linear Reg.

Categorical

Proportion
(Prevalence
Or
Risk)

Logistic Reg.

Count

Rate per “space”

Poisson Reg.

Survival

Median survival
Risk of events at $T_{(t)}$

Cox Reg.

Dependency of the study outcome required special statistical methods to handle it

Continuous	Categorical	Count	Survival
Mean Median	Proportion (Prevalence Or Risk)	Rate per “space”	Median survival Risk of events at $T_{(t)}$
Linear Reg.	Logistic Reg.	Poisson Reg.	Cox Reg.

Mixed model, multilevel model, GEE

Back to the conclusion

Continuous | Categorical | Count | Survival

Appropriate statistical methods

- Continuous: Mean, Median
- Categorical: Proportion (Prevalence or Risk)
- Count: Rate per "space"
- Survival: Median survival, Risk of events at $T_{(t)}$

Magnitude of effect
95% CI
P-value

Answer the research question based on lower or upper limit of the CI

Descriptive Statistics

- Characterize data set
  - Graphical presentation
    - Histograms
    - Frequency distribution
    - Box and whiskers plot
  - Numeric description
    - Mean, median, SD, interquartile range

Sources of data collection

1. Experiments
2. Surveys
3. Records

Presentation of data

- Tabulation
- Diagrams

SAMPLING

- The process of selecting a representative part from the whole
- Sample: the representative part
- Population: the whole from which sample is drawn

Measures of central Tendency

A measurement that is used to describe the tendency of all the individual observations to be in the centre

- Mean
- Median
- Mode

Mean

Arithmetic average obtained by summing up all the observation & dividing the total by the No: of observations

$$Mean(\overline{X}) = \frac{\sum_{i=1}^{n} x_i}{n}$$

Eg: 7 subjects : 7, 5, 3, 4, 6, 4, 5

$$Mean = \frac{7+5+3+4+6+4+5}{7} = 4.86$$

Median

- The middle observation when all the observation are arranged in ascending or descending order
- E.g.: 7 subjects arranged in ascending order are, 3, 4, 4, 5, 5, 6,7

  The 4th observation i.e. ‘5’ is the median

Mode

Most frequently occurring observation in a series

E.g.: the size of induration in tuberculin test of 10 boys – 3, 5, 7, 7, 8, 8, 8, 10 ,11, 12

The mode here is '8'

Measures of variability

- Variability: biological data collected by measurement or counting are variable
- No 2 measurements in man are absolutely equal
- Measures of variability of observations help to find how individual observations are dispersed around the mean

Measures of variability

- Ra**n**ge
- Mean deviation
- Standard deviation
- Standard error

Range

- The normal limits of a biological characteristic
  - E.g.: systolic B.P. 100-140mm of Hg

Mean deviation

Mean of the absolute deviation from the central tendency

$$MD = \frac{\sum_{i=1}^{n} \left| x_i - \overline{X} \right|}{n}$$

Mean deviation about mean

= sum of the absolute deviation from the mean / No: of observation

Standard deviation

- Square root of the average of the sum of the squares of deviations taken from the mean

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}}$$

- Indicates whether the variation of difference of an individual from the mean is by chance or real

Standard error

- Difference between sample & population values

- Is a measure of chance variation

SE = SD / $\sqrt{n}$

Touch the variability (uncertainty) to understand statistical inference

id	A	$(x-\bar{X})$	$(x-\bar{X})^2$
1	2	-2	4
2	2	-2	4
3	0	-4	16
4	2	-2	4
5	14	10	100
Sum (Σ)	20	0	128
Mean($\bar{X}$)	4	0	
SD			5.66
Median	2		

2+2+0+2+14 = 20

$$\frac{2+2+0+2+14}{5} = \frac{20}{5} = 4$$

0 2 2 2 14

Variance = SD^2

Standard deviation = SD

Touch the variability (uncertainty) to understand statistical inference

id	A	$(x-\bar{X})$	$(x-\bar{X})^2$
1	2	-2	4
2	2	-2	4
3	0	-4	16
4	2	-2	4
5	14	10	100
Sum (Σ)	20	0	128
Mean($\bar{X}$)	4	0	32.0
SD			5.66
Median	2		

Measure of central tendency

Measure of variation

Same mean BUT different variation

id	A
1	2
2	2
3	0
4	2
5	14
Sum (Σ)	20
Mean	4
SD	5.66
Median	2

Heterogeneous data

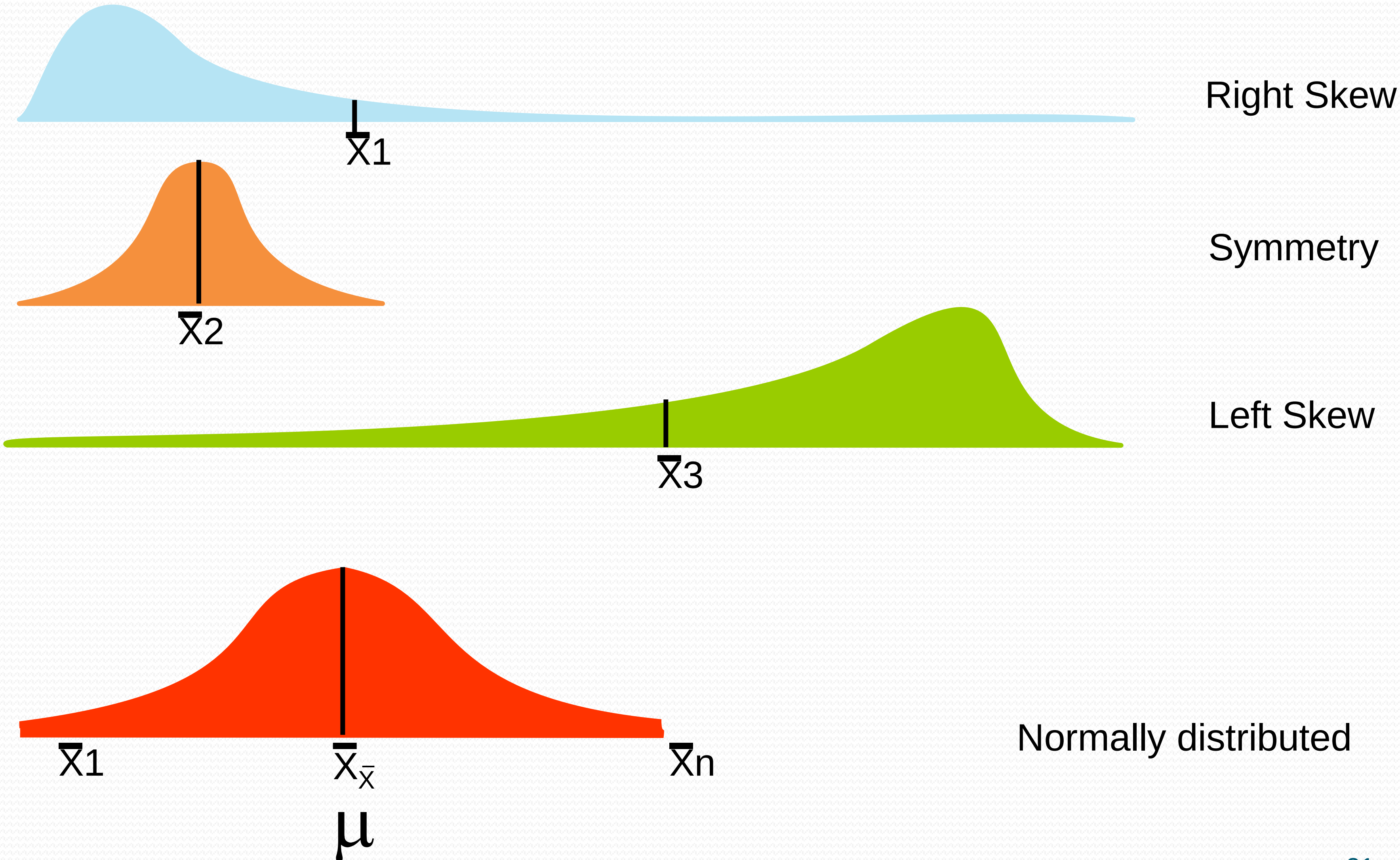
id	B
1	4
2	3
3	5
4	4
5	4
Sum (Σ)	20
Mean	4
SD	0.71
Median	4

Homogeneous data

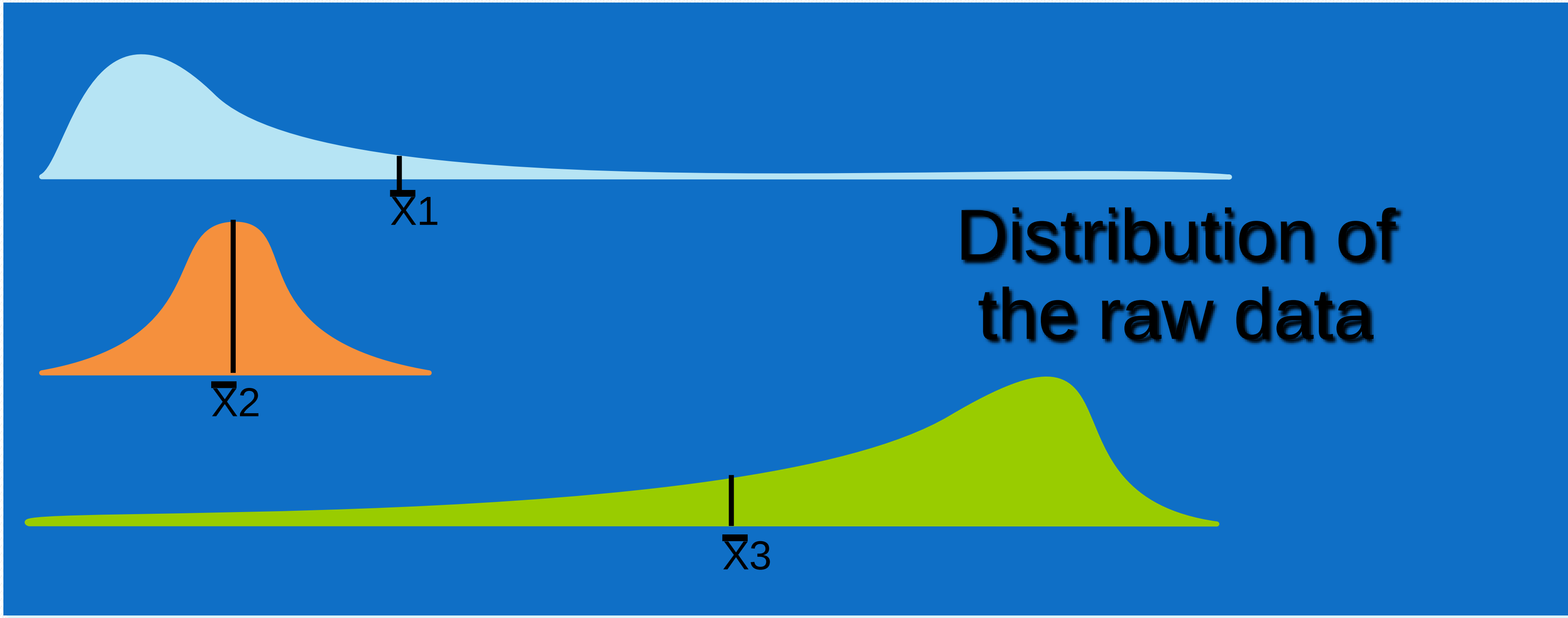
Facts about Variation

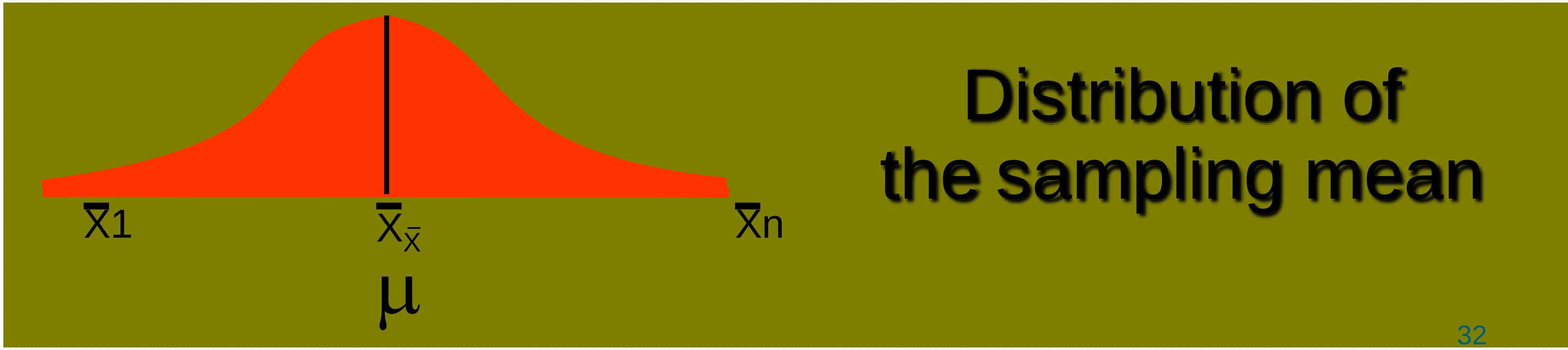
- Because of variability, repeated samples will NOT obtain the same statistic such as mean or proportion:
  - Statistics varies from study to study because of the role of chance
  - Hard to believe that the statistic is the parameter
  - Thus we need statistical inference to estimate the parameter based on the statistics obtained from a study
- Data varied widely = heterogeneous data
- Heterogeneous data requires large sample size to achieve a conclusive finding

Central Limit Theorem

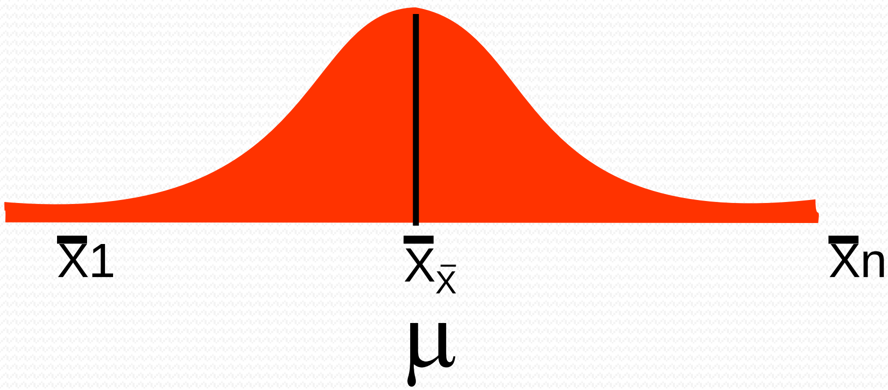


Central Limit Theorem





Central Limit Theorem



Distribution of the raw data

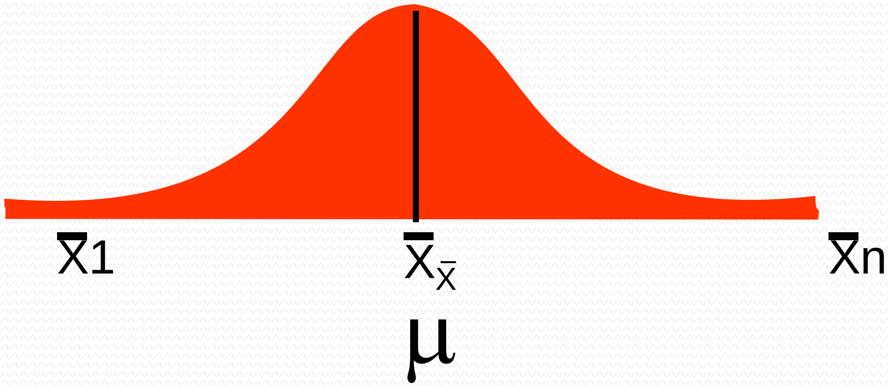
Distribution of the sampling mean

Large sample

(Theoretical) Normal Distribution

Central Limit Theorem

Many X, $\bar{X}$, SD



Many $\bar{X}$, $\bar{X}_{\bar{X}}$, SE

Large sample

Standardized for whatever n,
Mean = 0, Standard deviation = 1

Standard deviation of the sampling mean
Standard error (SE)
Estimated by

$$SE = \frac{SD}{\sqrt{n}}$$

Report and interpret p-value appropriately

- Example of over reliance on p-value:
  - Real results: n=5900; $OR_{\text{Drug A vs Drug B}} = 1.02$ ($P<0.001$)
  - Inappropriate: Quote p-value as < 0.05 or put * or **** (star) to indicate significant results
  - Wrong: Drug A is highly significantly better than Drug B ($P<0.001$)
  - What if 95%CI: 1.001 to 1.300?
  - This is no clinical meaningful at all….!

Report and interpret p-value appropriately

- Example of over reliance on p-value:
  - Real results: n=30; $OR_{\text{Drug A vs Drug B}} = 9.2$ ($P=0.715$)
  - Inappropriate: Quote p-value as > 0.05
  - Wrong: There is no statistical significant difference of the treatment effect ($P<0.05$). Thus Drug A is as effective as Drug B
  - What if 95%CI: 0.99 to 28.97?
  - This is study indicated a low power, NOT suggested an equivalence...!
  - Correct: There was no sufficient information to concluded that . . . => inconclusive findings

P-value is the magnitude of chance NOT magnitude of effect

- P-value < 0.05 = Significant findings
- Small chance of being wrong in rejecting the null hypothesis
- If in fact there is no [effect], it is unlikely to get the [effect] = [magnitude of effect] or more extreme
- Significance DOES NOT MEAN importance
- Any extra-large studies can give a very small P-value even if the [magnitude of effect] is very small

P-value is the magnitude of chance NOT magnitude of effect

- P-value > 0.05 = Non-significant findings
- High chance of being wrong in rejecting the null hypothesis
- If in fact there is no [effect], the [effect] = [magnitude of effect] or more extreme can be occurred chance
- Non-significance DOES NOT MEAN no difference, equal, or no association
- Any small studies can give a very large P-value even if the [magnitude of effect] is very large

P-value vs. 95%CI (1)

An example of a study with dichotomous outcome

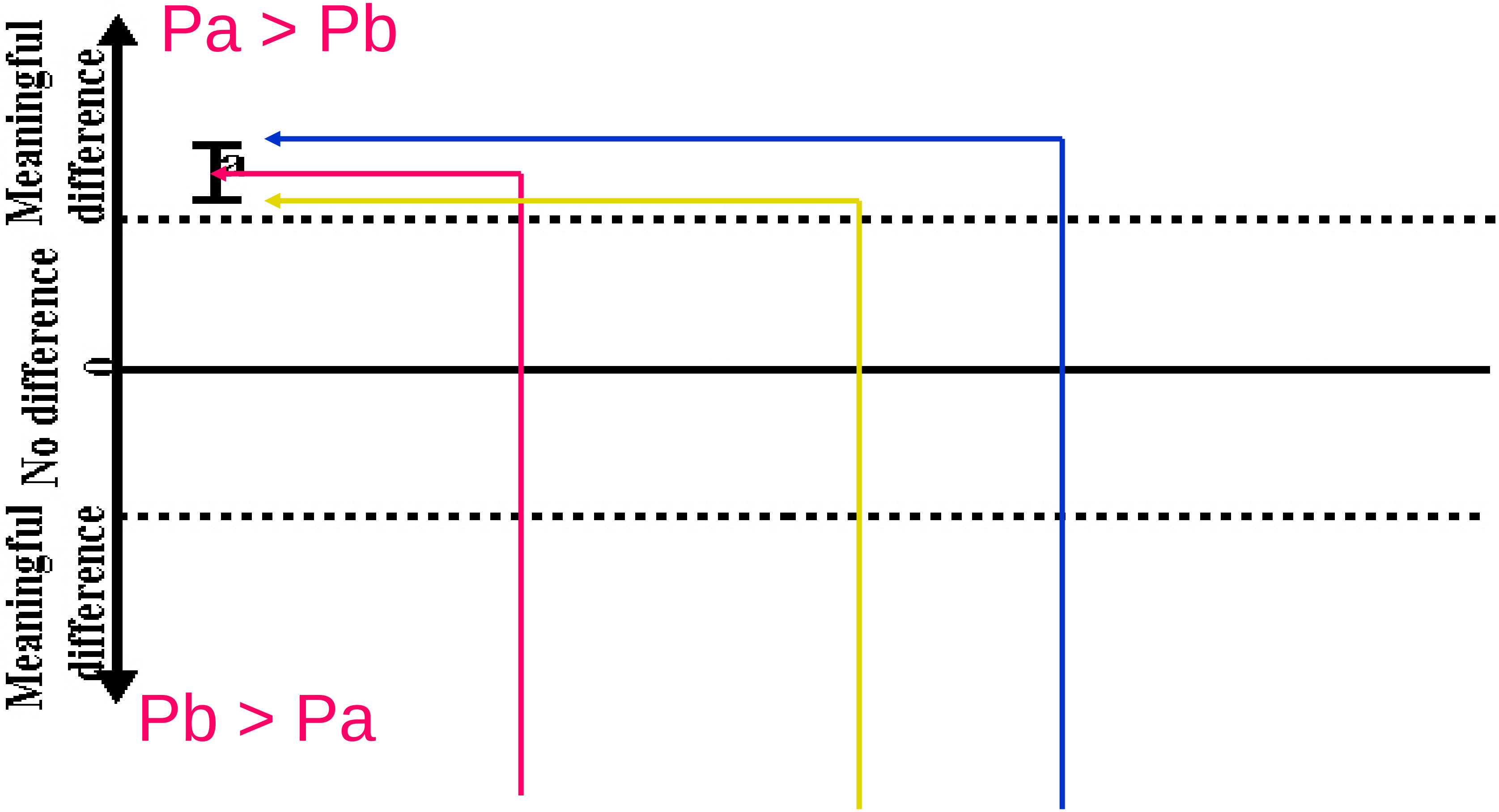
A study compared cure rate between Drug A and Drug B

Setting:
Drug A = Alternative treatment
Drug B = Conventional treatment

Results:
Drug A: n1 = 50, Pa = 80%
Drug B: n2 = 50, Pb = 50%

Pa-Pb = 30% (95%CI: 26% to 34%; P=0.001)

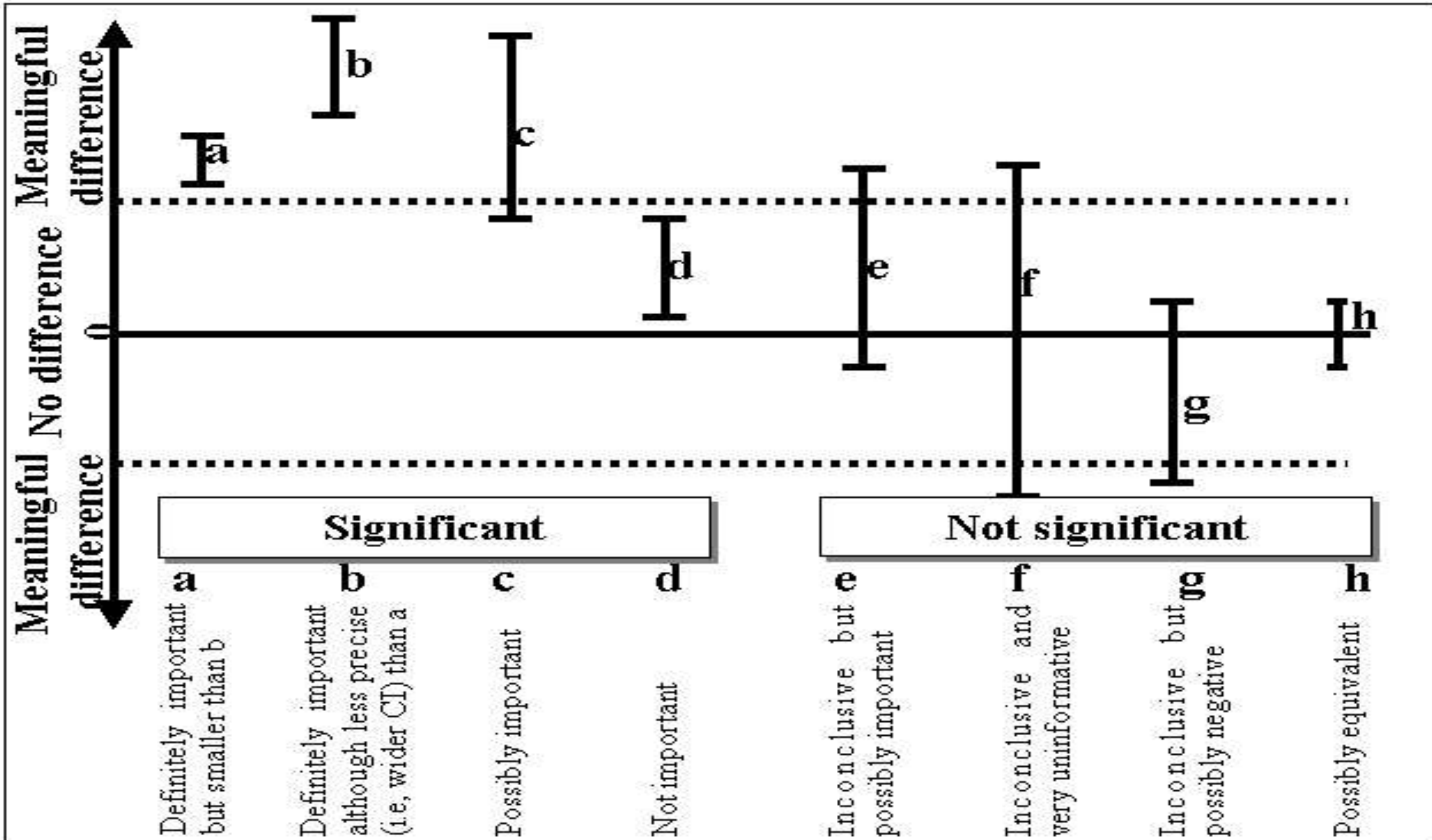
P-value vs. 95%CI (2)



Pa-Pb = 30% (95%CI: 26% to 34%; P< 0.05)

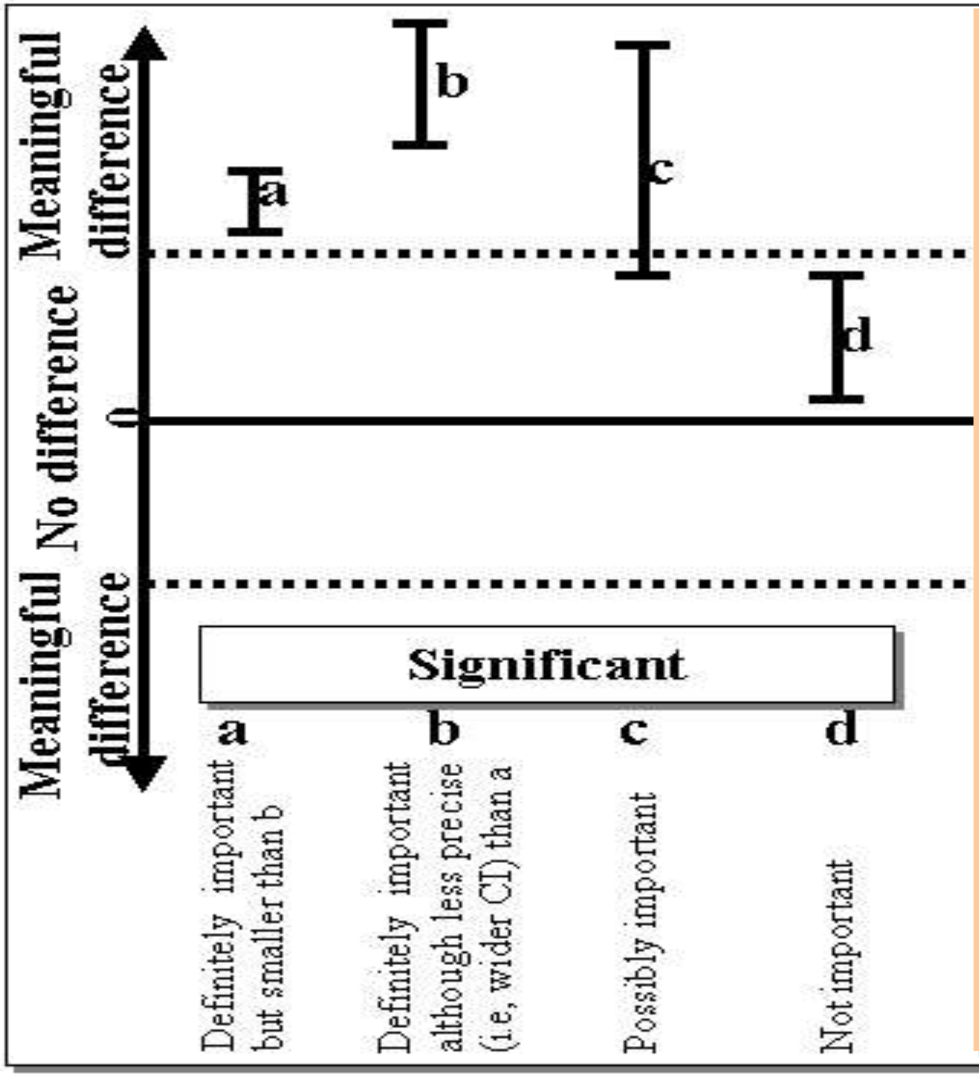
P-value vs. 95%CI (3)

Adapted from: Armitage, P. and Berry, G. Statistical methods in medical research. 3rd edition. Blackwell Scientific Publications, Oxford. 1994. page 99



P-value vs. 95%CI (4)

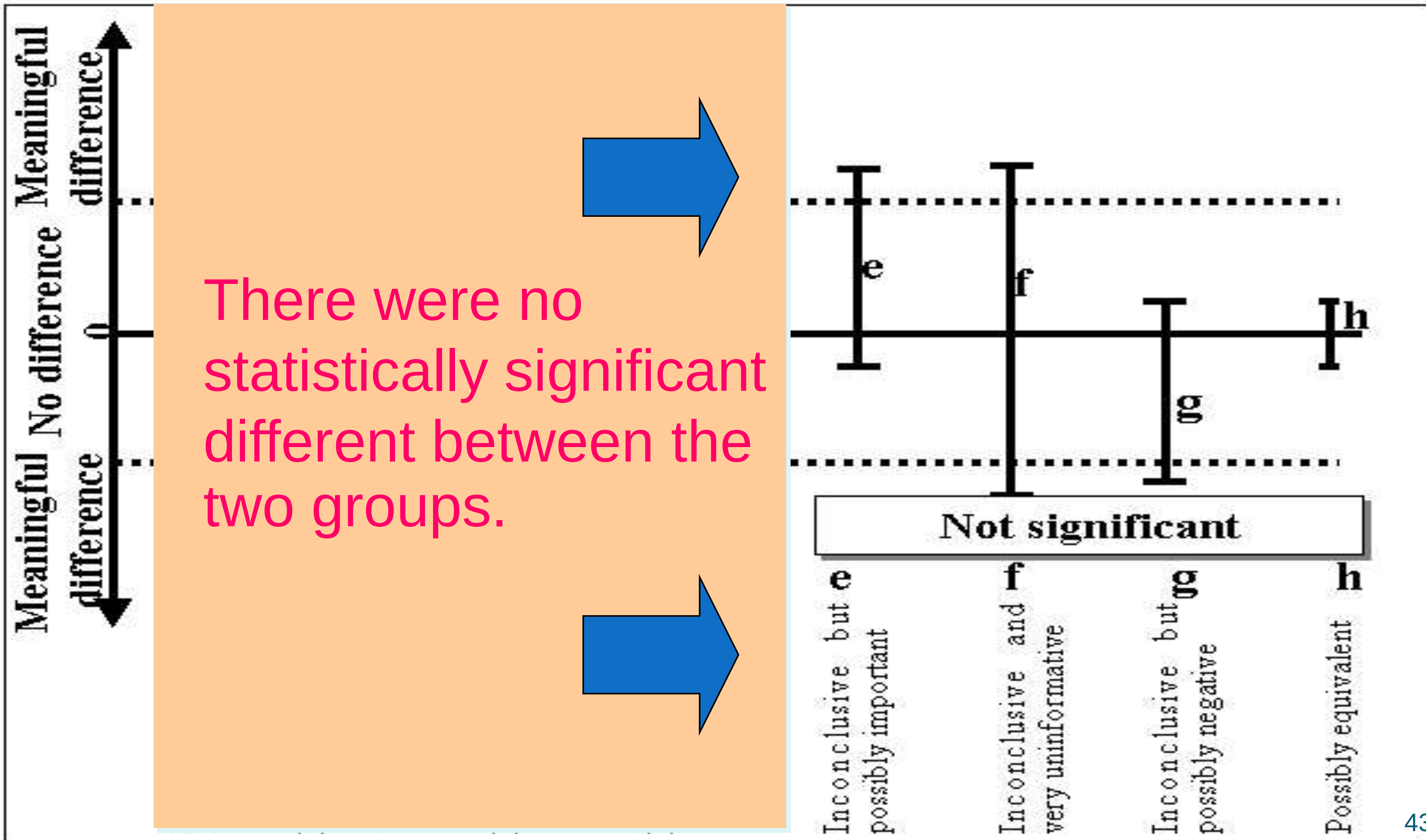
Adapted from: Armitage, P. and Berry, G. Statistical methods in medical research. 3rd edition. Blackwell Scientific Publications, Oxford. 1994. page 99



There were statistically significant different between the two groups.

P-value vs. 95%CI (5)

Adapted from: Armitage, P. and Berry, G. Statistical methods in medical research. 3rd edition. Blackwell Scientific Publications, Oxford. 1994. page 99



There were no statistically significant different between the two groups.

P-value vs. 95%CI (4)

- Save tips:
  - Always report 95%CI with p-value, NOT report solely p-value
  - Always interpret based on the lower or upper limit of the confidence interval, p-value can be an optional
  - Never interpret p-value > 0.05 as an indication of no difference or no association, only the CI can provide this message