

Data Analytics in Clinical Data Management using Stata

Jaya Kumawat

Head Biometrics

PPCE Pvt Ltd, Pune



STATA Conference

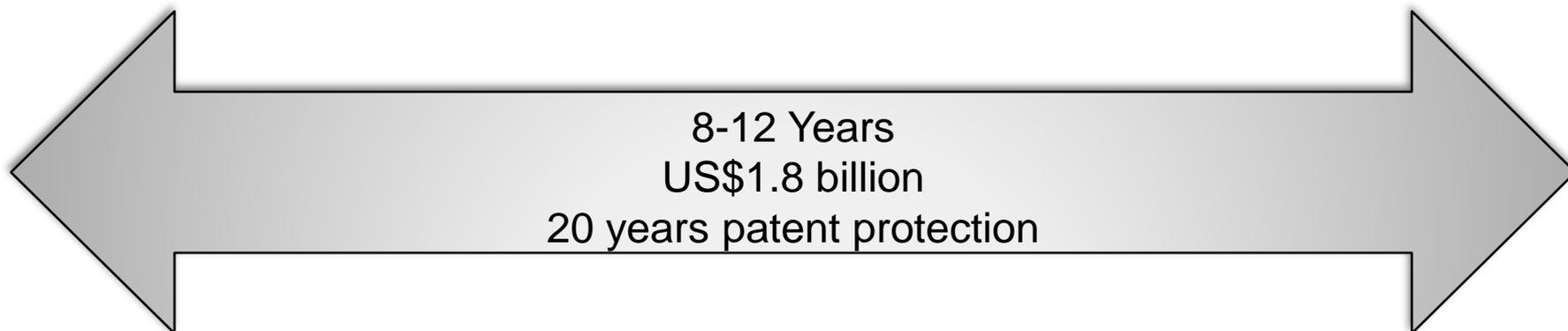
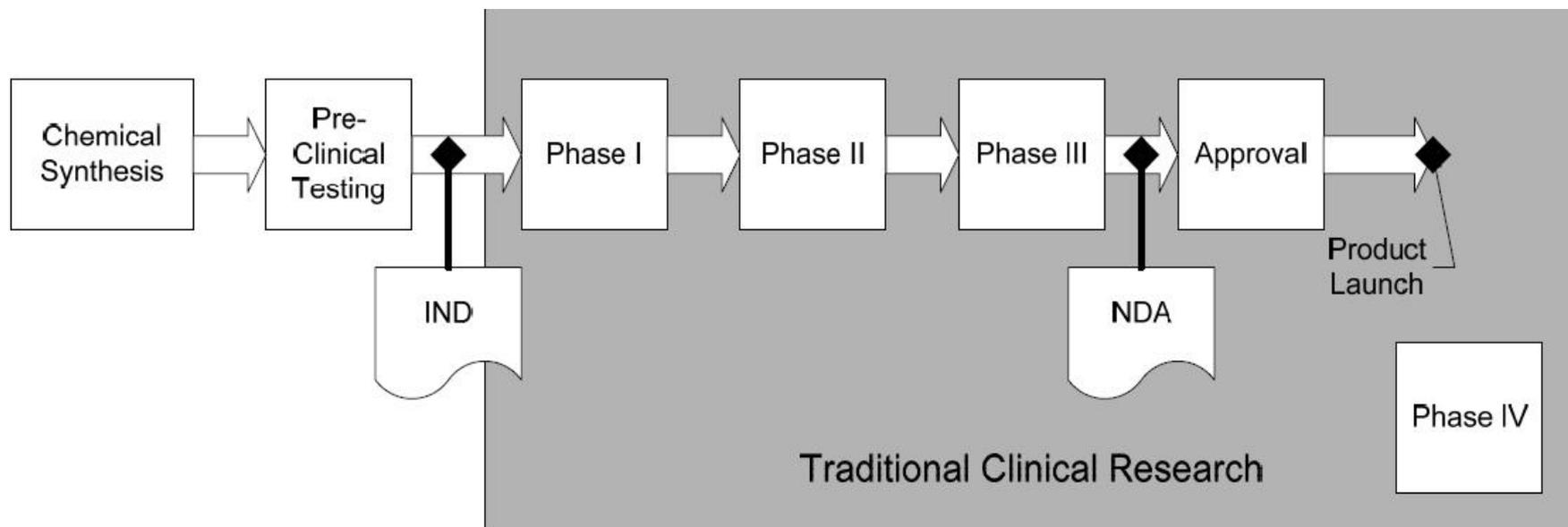
1st -3rd August, 2013
Mumbai India

AGENDA

- 1 Understanding Clinical Trials
- 2 About Clinical Data Management
- 3 Statistical Analysis in Clinical Research
- 4 Stats Software – Evaluation Criteria
- 5 STATA – Features and Benefits
- 6 Conclusion

Understanding Clinical Trials

Drug development life cycle



Drug discovery and preclinical development

Pre-discovery

Goal: Understand the disease and choose a target molecule.

How: Scientists in pharmaceutical research companies, government, academic and for-profit research institutions contribute to basic research.

Discovery

Goal: Find a drug candidate.

How: Create a new molecule or select an existing molecule as the starting point. Perform tests on that molecule and then optimize (change its structure) it to make it work better.

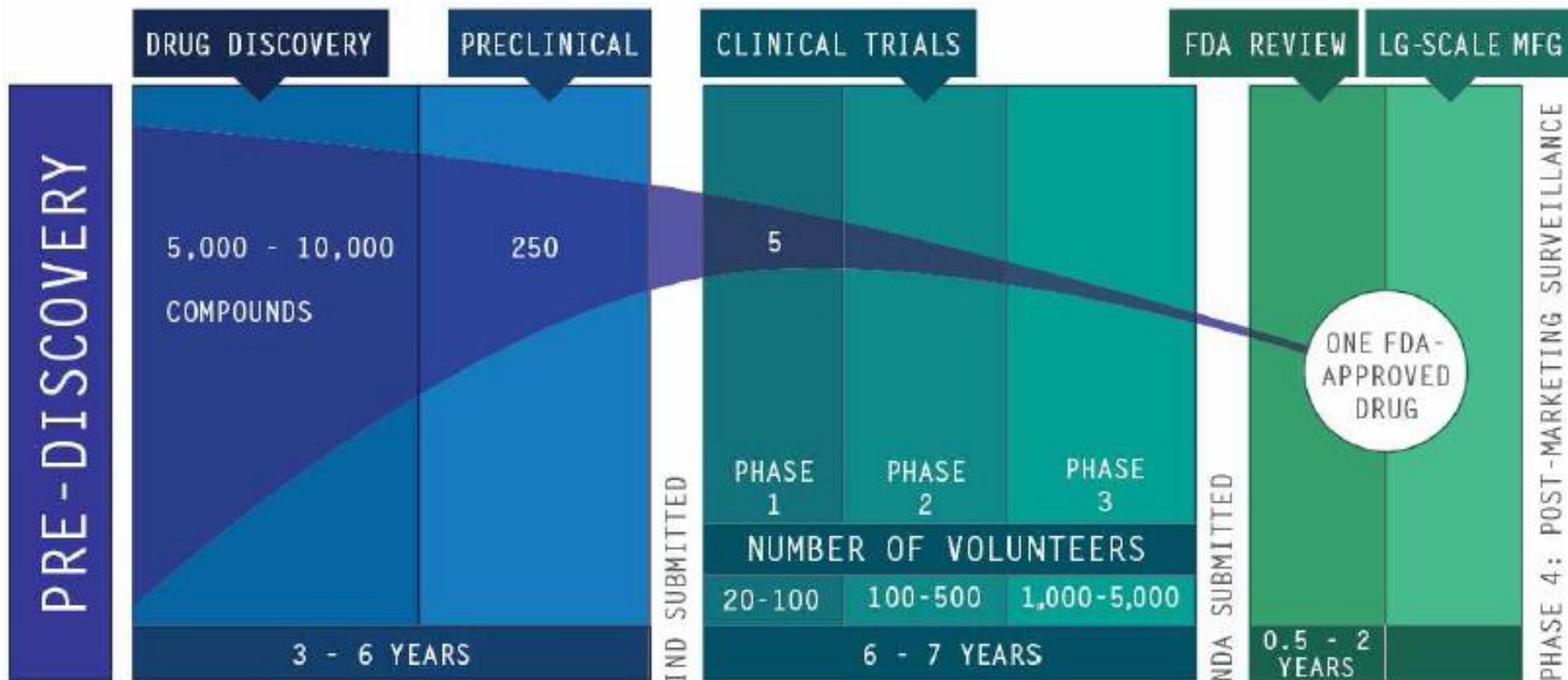
Preclinical

Goal: Test extensively to determine if the drug is safe enough for human testing.

How: Researchers test the safety and effectiveness in the lab and in animal models.

3 - 6 YEARS

Drug Development Process



Basel Institute for Clinical Epidemiology and Biostatistics, University of Zürich

Clinical Research

- Clinical research involves working with human subjects to answer questions relevant to their well-being.
- Patient oriented research is where the ‘rubber meets the road’!

Hypothesis

- Hypothesis (proposed explanation for a phenomenon) is a tentative construct to be proved or disproved according to the evidence
- The hypothesis is sometimes expressed as a null hypothesis

Study Types

- **Observational** : *the investigators observe the subjects and measure their outcomes. The researchers do not actively manage the study.*
 - Longitudinal(Cohort)
 - Cross-sectional

- **RCT** : Randomized, double-blind, parallel group, placebo controlled trial

Study Design

■ Study Population

- Age
- Gender
- Ethnicity/Race
- Disease characteristics
- Exclusions
- Number
- Stratification
- Randomization



Human Subjects

- The safety and rights of human subjects must be protected
 - Study Design
 - Institutional Review Board
 - Informed consent
 - Data Safety Monitoring/Medical Monitors

Phase I Studies

- Assess drug safety and tolerability
- Healthy volunteers, then those with target disease
- Pharmacokinetics
 - Absorption
 - Metabolism
 - Excretion
- Dose escalation
- 70% of new drugs pass this phase



Phase II Studies

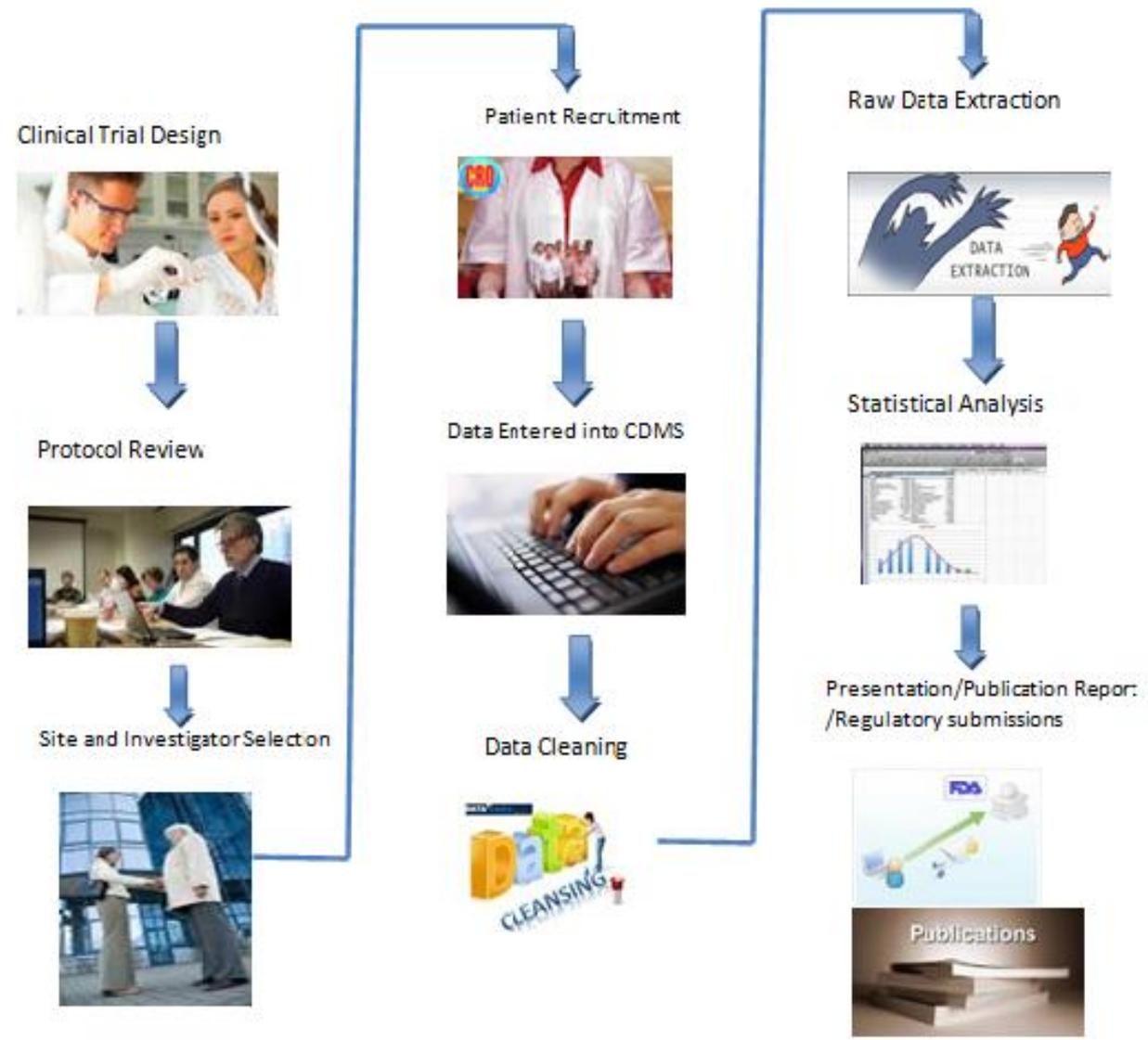
- Assess drug efficacy
- Usually randomized, controlled trials with smaller numbers up to several hundred subjects
- Test different therapeutic strategies
- Use surrogate variables and are usually short term
- Only 1/3 get past phase II

Phase III Trial

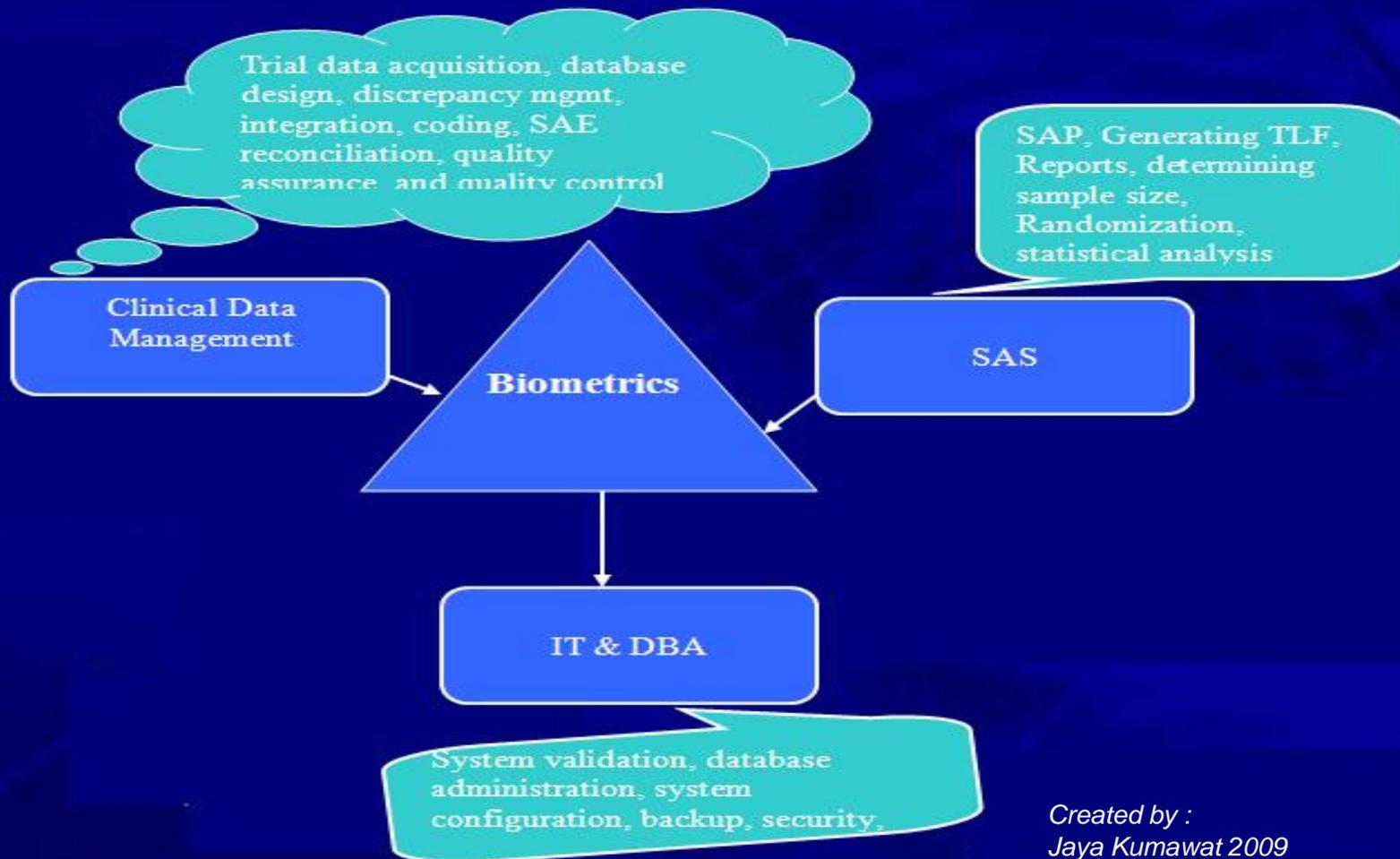
- **Objective** : To compare experimental or new therapies with standard therapy or competitive therapies.
- Very large, expensive studies
- Required by FDA for drug approval
- If drug approved, usually followed by Phase IV trials to follow-up on long-range adverse events – concern is safety

About Clinical Data Management

CONCEPTUAL MODEL OF CLINICAL DATA MANAGEMENT



Biometrics – A Broad Perspective



Technology

The CDM Industry is guided by the philosophy that technology is not an add-on to clinical trial processes; it is the

backbone.

Technology - CDMS

What is a CDMS?

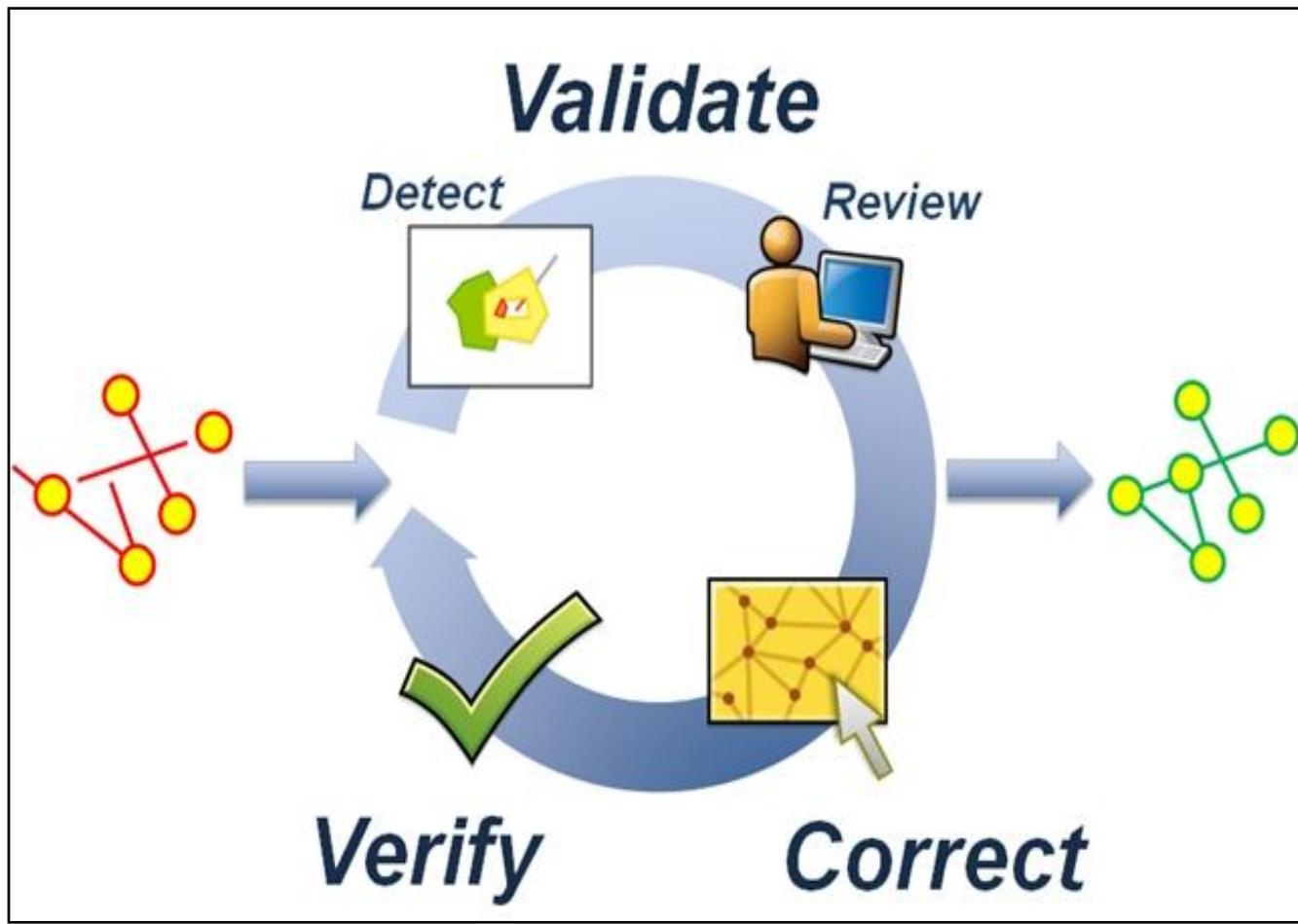
- A flexible relational database system for
 - Capturing,
 - Storing, and
 - Processing clinical trial data

HE WILL BE SHOT DEAD/KILLED!!

Data Entry



Data Validation



Data

- Data are the facts you measure
- They should be carefully recorded in an unbiased manner
- They should be measured in a manner that minimizes random variation
- They should be derived from the operational definitions you have developed

Data Validation

- Do the data make sense?
- Look critically at the data
 - Highest and lowest values
 - Data entry errors
 - Distribution: Normal or skewed
- Check selected data entries with original data forms



Data Interpretation

- Do not interpret/analyze data until after study is completed
- Do not 'unblind' subjects until the study is completed other than for safety reasons
- Do not interpret/analyze data until after data has been validated and the data set closed



CDM – Compliance, Standards & Regulatory

Understanding the Environment

International Regulations/ Guidelines

- **Code of Federal Regulations**

- HIPAA

- FDA

- 21 CFR Part 11

- IRB

-

- **ICH**

GCP

Standards for
data



- **GCP/GCDMP**

- **Sponsor specific**

- **Local regulations**

- **Etc, etc, etc**

GCDMP

Quality Assurance
ICH E6

Compliance

- **compliance** with [21 CFR part 11](#): security, audit trail, version control
- **Validation** - accuracy, reliability, consistent intended performance and the means to discern invalid or altered records
- **Restriction of system access** to only authorized individuals
- **Secure, computer-generated, time-stamped audit trails** to record operator entries and action to create, modify and delete electronic records – retained for required period and available for agency review and copying
- **Operational system checks** to enforce permitted sequencing of steps and events as appropriate
- **Authority checks** for use, e-signature, access of input and output device, altering a record and performing operation at hand
- **PDUFA III** - The '**Prescription Drug User Fee Act**' (PDUFA) was a [law](#) passed by the [United States Congress](#) in 1992 which allowed the [Food and Drug Administration](#) (FDA) to collect fees from [drug manufacturers](#) to fund the new [drug](#) approval process.

Related Standards

SDTM

Define.xml

ADaM

But why standards??????

- Reviewer's efficiency.
- FDA Repository
- Standard software tools for viewing and analyzing.
- Streamline the flow of data from collection through submission, facilitating data interchange b/w partners and providers.

Statistical Analysis in Clinical Research

Statistical Considerations in Clinical Research

- Sample Size determination.
- Randomization
- SAP(Statistical Analysis Plan)
- Manual Discrepancy Management
- Generating Table, Listings, Figures.
- Data Conversions into CDISC (SDTM & ADaM) Models.
- Analysis and Reporting (including XMLs for FDA submissions).

Stats Software – Evaluation Criteria

Criteria 1 : 21CFR Part 11

- **Stata is verifiably accurate**
- When you submit new drug applications (NDAs), the U.S. Food and Drug Administration (FDA) requires you to verify the validity of your data and analyses.
- According to this document, to demonstrate the validity of the software used for an NDA, you must have :
 - a written design specification
 - a test plan
 - test results.

Stata provides all the three

Criteria 2: Used for Clinical Research ??

Stata is widely used in the medical community, including many hospitals and medical schools. In addition, the Centers for Disease Control uses Stata frequently.

Online resource to STATA users :

- ❖ *Dalla Lana School of Public Health, University of Toronto*
- ❖ *Department of Epidemiology, University of Albany*
- ❖ *Department of Psychology, Stockholm University & National Institute for Psychosocial Medicine*
- ❖ *Department Community Health & General Practice, Trinity College, Dublin, Ireland*
- ❖ *Ontario Cancer Institute, Princess Margaret Hospital*
- ❖ *University of North Carolina Chapel Hill*
- ❖ *Harvard School of Public Health*

Criteria 3: Programming Required?

- No in-depth programming skills are necessary for the primary usage of Stata. If a user wishes to write their own command, they will need some experience.
- There are many useful topics here and they are presented in an intuitive way. The manuals provided with your license (in PDF format) are also very valuable.

Criteria 4: Support for Regulatory/FDA Submission

Can STATA generate SAS transport files(.XPT) required for FDA submissions?

Answer:

- Stata can read and write SAS XPORT format datasets natively, using the `fdause` and `fdasave` commands.
- STATA can export a dataset in FDA (SAS XPORT) format using the `-fdasave-` command.
- You can also read in a SAS XPORT file into Stata with the `-fdause-` command.

Criteria 5: CDISC Compliance

- How will the software support ADaM. Does it have built in CDISC SDTM and ADaM libraries?
- Stata has no direct link to ADaM, and it hasn't been tested in this environment though no problems anticipated.

Criteria 6: Generation of XML files

- How will the software support generation of Data Definition Files like the `define.xml`.
- Stata can write datasets to `.xml` format using `-xmlsave-`, or you can write to text format for spreadsheets using `-out sheet-`.

Sample xml file generated from STATA

```
<?xml version="1.0" encoding="US-ASCII" standalone="yes"?>
<dta>
<header>
<ds_format>113</ds_format>
<byteorder>LOHI</byteorder>
<filetype>1</filetype>
<nvar>12</nvar>
<nobs>74</nobs>
<data_label>1978 Automobile Data</data_label>
<time_stamp>26 Jul 2013 12:55</time_stamp>
</header>
<descriptors>
<typelist>
<type varname='make'>str18</type>
<type varname='price'>int</type>
<type varname='mpg'>int</type>
<type varname='rep78'>int</type>
<type varname='headroom'>float</type>
<type varname='trunk'>int</type>
<type varname='weight'>int</type>
<type varname='length'>int</type>
<type varname='turn'>int</type>
<type varname='displacement'>int</type>
<type varname='gear_ratio'>float</type>
<type varname='foreign'>byte</type>
</typelist>
<varlist>
```

Criteria 7: Cost Assessment

- Breaks the market monopoly of other statistical tools.
- Recurring License Fee.
- Reasonably priced.

Criteria 8: Training/learning resources.

- STATA can arrange for an online training at corporate level on a need basis.
- Huge amounts of online training documentations available.
- a pdf help file has been installed within the software.

Criteria 9: Support for biostatistics/epidemiology features

- Comprehensive coverage of biostatistics/epidemiology .
- The capabilities include all the standard functionalities required viz. dataset filtering, transformation, merging, statistical models, great graphics and other biostatistical tests.

Criteria 10: Miscellaneous Factors

- Maintenance & support.
- Ethics/ principles/ honesty.
- GUI , ease of use, intuitive.
- Updates/patches.
- Multi platform support i.e. Linux, Macintosh, windows.

STATA – Features and Benefits as explored and supported for Clinical Research

1. Data imports

- Imports data in all possible formats:
 - a) MsExcel.
 - b) CSV.
 - c) ASCII
 - d) .xpt
 - e) .xml

 - f) *(perhaps some others.....)*

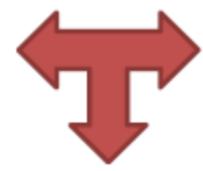
2. Data Transformations (E.g.: merge , append)

mydata1

| | country | year | y | y_bin | x1 | x2 | x3 |
|----|---------|------|-------|-------|------|-------|------|
| 1 | A | 2000 | 1343 | 1 | .28 | -1.11 | .28 |
| 2 | A | 2001 | -1900 | 0 | .32 | -.95 | .49 |
| 3 | A | 2002 | -11 | 0 | .36 | -.79 | .7 |
| 4 | A | 2003 | 2646 | 1 | .25 | -.89 | -.09 |
| 5 | B | 2000 | -5935 | 0 | -.08 | 1.43 | .02 |
| 6 | B | 2001 | -712 | 0 | .11 | 1.65 | .26 |
| 7 | B | 2002 | -1933 | 0 | .35 | 1.59 | -.23 |
| 8 | B | 2003 | 3073 | 1 | .73 | 1.69 | .26 |
| 9 | C | 2000 | -1292 | 0 | 1.31 | -1.29 | .2 |
| 10 | C | 2001 | -3416 | 0 | 1.18 | -1.34 | .28 |
| 11 | C | 2002 | -358 | 0 | 1.26 | -1.26 | .37 |
| 12 | C | 2003 | 1225 | 1 | 1.42 | -1.21 | -.28 |

mydata2

| | country | year | x4 | x5 | x6 | order |
|----|---------|------|----|----|----|-------|
| 1 | A | 2000 | 10 | 1 | 9 | 1 |
| 2 | A | 2001 | 7 | 1 | 9 | 2 |
| 3 | A | 2002 | 7 | 9 | 4 | 3 |
| 4 | A | 2003 | 1 | 2 | 3 | 4 |
| 5 | B | 2000 | 0 | 5 | 6 | 5 |
| 6 | B | 2001 | 5 | 8 | 5 | 6 |
| 7 | B | 2002 | 9 | 4 | 5 | 7 |
| 8 | B | 2003 | 1 | 5 | 1 | 8 |
| 9 | C | 2000 | 4 | 5 | 4 | 9 |
| 10 | C | 2001 | 6 | 9 | 6 | 10 |
| 11 | C | 2002 | 6 | 5 | 3 | 11 |
| 12 | C | 2003 | 7 | 3 | 3 | 12 |



```
merge 1:1 country year using mydata2
```

```
Result                                     # of obs.
-----
not matched                                0
matched                                    12  (_merge==3)
```

- Make sure one dataset is loaded into Stata (in this case mydata1), then use merge.
 - Make sure to map where the using data is located (in this case mydata2, for example "c:\folders\data\mydata2.dta").
- NOTE: For Stata 10 or older:
- 1) Remove the 1:1
 - 2) Sort both datasets by all the ids and save before merging

| | country | year | y | y_bin | x1 | x2 | x3 | x4 | x5 | x6 | order | _merge |
|---|---------|------|-------|-------|------|-------|------|----|----|----|-------|-------------|
| 1 | A | 2000 | 1343 | 1 | .28 | -1.11 | .28 | 10 | 1 | 9 | 1 | matched (3) |
| 2 | A | 2001 | -1900 | 0 | .32 | -.95 | .49 | 7 | 1 | 9 | 2 | matched (3) |
| 3 | A | 2002 | -11 | 0 | .36 | -.79 | .7 | 7 | 9 | 4 | 3 | matched (3) |
| 4 | A | 2003 | 2646 | 1 | .25 | -.89 | -.09 | 1 | 2 | 3 | 4 | matched (3) |
| 5 | B | 2000 | -5935 | 0 | -.08 | 1.43 | .02 | 0 | 5 | 6 | 5 | matched (3) |
| 6 | B | 2001 | -712 | 0 | .11 | 1.65 | .26 | 5 | 8 | 5 | 6 | matched (3) |
| 7 | B | 2002 | -1933 | 0 | .35 | 1.59 | -.23 | 9 | 4 | 5 | 7 | matched (3) |

3. Ability to handle Diff -> Types of Data – (1)

- **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 – (2)

■ *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

Support for 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

Descriptive Statistics

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

Numeric Descriptive Statistics

- Measures of central tendency of data
 - Mean
 - Median
 - Mode
- Measures of variability of data
 - Standard Deviation
 - Interquartile range

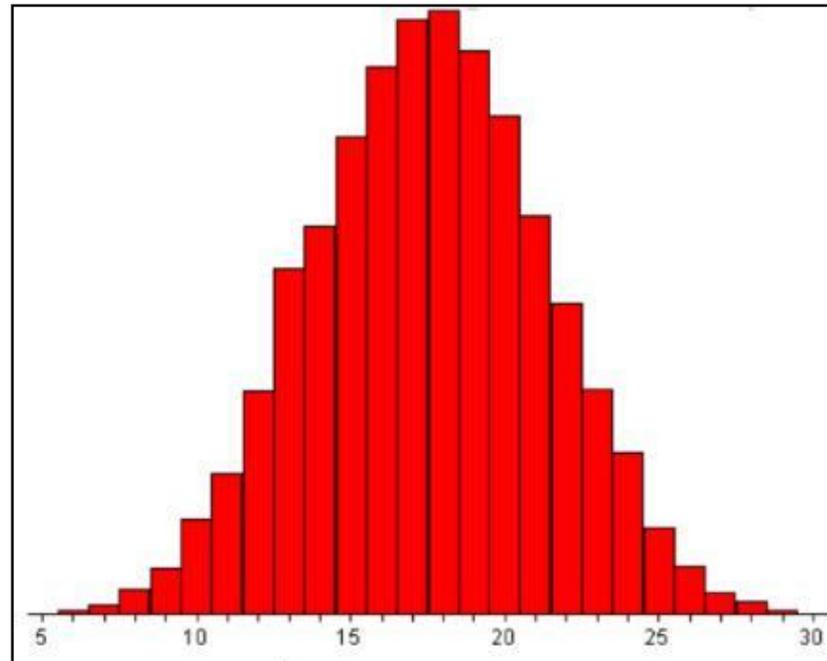
Descriptive Statistics Example

```
. summarize
```

| variable | obs | Mean | Std. Dev. | Min | Max |
|---------------------------------|-----|----------|-----------|------|------|
| id | 30 | 15.5 | 8.803408 | 1 | 30 |
| lastname | 0 | | | | |
| firstname | 0 | | | | |
| city | 0 | | | | |
| state | 0 | | | | |
| Zeros indicate string variables | | | | | |
| gender | 0 | | | | |
| studentstatus | 0 | | | | |
| major | 0 | | | | |
| country | 0 | | | | |
| age | 30 | 25.2 | 6.870226 | 18 | 39 |
| sat | 30 | 1848.9 | 275.1122 | 1338 | 2309 |
| averagescore | 30 | 80.36667 | 10.11139 | 63 | 96 |
| heightin | 30 | 66.43333 | 4.658573 | 59 | 75 |
| newspaperwk | 30 | 4.866667 | 1.279368 | 3 | 7 |

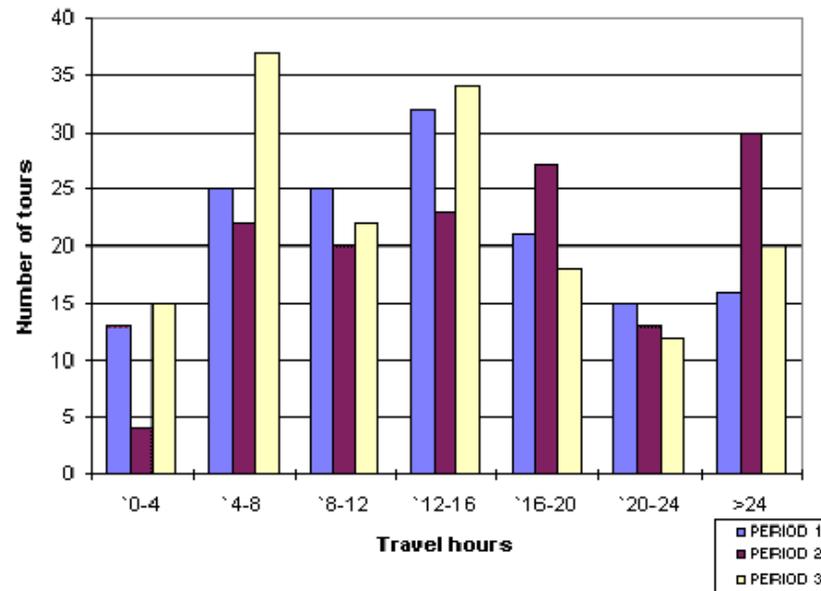
Use 'min' and 'max' values to check for a valid range in each variable. For example, 'age' should have the expected values ('don't know' or 'no answer' are usually coded as 99 or 999).

Histogram - Continuous Data



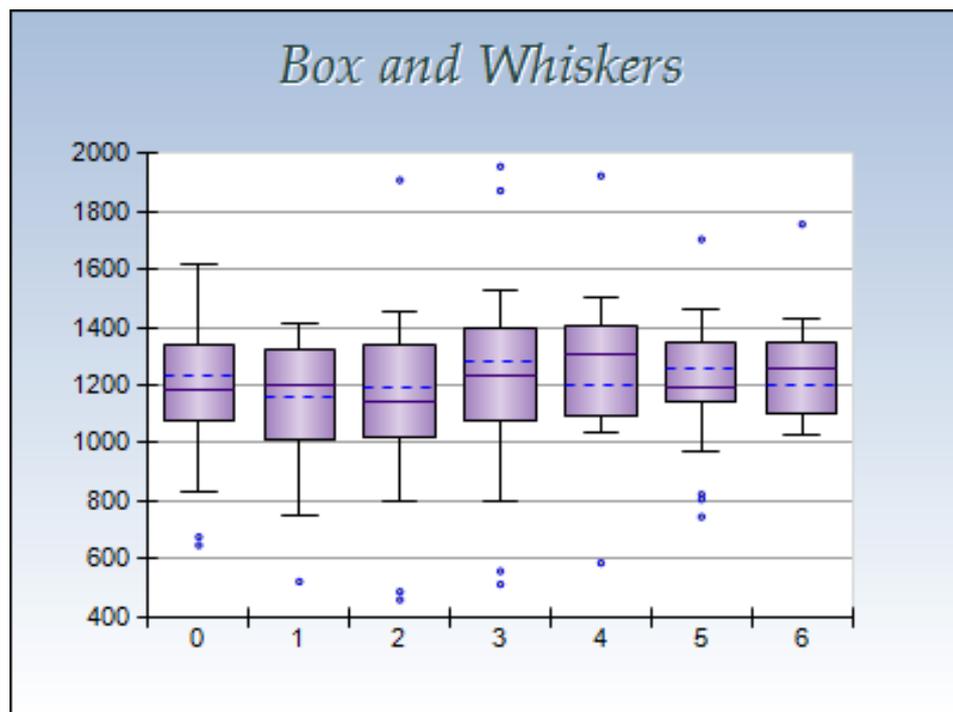
No segmentation of data into groups

Frequency Distribution



Segmentation of data into groups
Discrete or continuous data

Box and Whiskers Plots



Useful for presenting comparative data graphically

Standard Error

- Standard error measures the precision of the estimate of a population parameter provided by the sample mean or proportion.

Standard Error Significance

- Significance:
 - Is the basis of confidence intervals
 - A 95% confidence interval is defined by
 - Sample mean (or proportion) $\pm 1.96 \times$ standard error
 - Since standard error is inversely related to the sample size:
 - The larger the study (sample size), the smaller the confidence intervals and the greater the precision of the estimate.

Confidence Intervals

- May be used to assess a single point estimate such as mean or proportion.
- Most commonly used in assessing the estimate of the difference between two groups.

Confidence Interval

Table 1
Mean Values of Baseline Characteristics and Short-term Outcomes

| | No. of Patients | Mean ± SE | 95% CI |
|---|-----------------|-------------|-----------|
| Baseline characteristics | | | |
| Age (y) | 100 | 43.0 ± 0.52 | 41.9–44.0 |
| Body mass index (kg/m ²) | 83 | 26.9 ± 0.44 | 26.0–27.7 |
| Baseline uterine volume (cm ³) | 96 | 628 ± 34.1 | 560–695 |
| Baseline fibroid volume (cm ³) | 96 | 150 ± 15.7 | 118–181 |
| Baseline fibroid-specific QOL symptom score | 98 | 54.1 ± 2.19 | 49.8–58.5 |
| Baseline fibroid-specific QOL total score | 98 | 52.3 ± 2.24 | 47.9–56.8 |
| Ethnic background (%) | | | |
| African-American | 61 | | |
| Caucasian | 34 | | |
| Other | 5 | | |
| Short-term outcome measures | | | |
| Maximum VAS score in hospital | 99 | 3.03 ± 0.26 | 2.50–3.55 |
| Maximum VAS score first week | 92 | 4.89 ± 0.26 | 4.38–5.4 |
| Maximum temperature in hospital (°C) | 91 | 37.1 ± 0.05 | 37.1–37.2 |
| Maximum temperature first week (°C) | 93 | 37.4 ± 0.05 | 37.4–37.5 |
| Symptom summary score | | | |
| First week | 90 | 26.6 ± 1.73 | 23.2–30.1 |
| Week 2 | 96 | 5.93 ± 0.34 | 5.25–6.60 |
| Week 3 | 87 | 4.68 ± 0.38 | 3.92–5.44 |
| Week 4 | 90 | 4.86 ± 0.41 | 4.04–5.68 |
| Weeks 2–4 | 83 | 15.3 ± 0.85 | 13.6–17.0 |
| Number of PCA doses attempted | 96 | 70.6 ± 6.72 | 57.2–83.9 |
| Number of PCA doses given | 97 | 28.1 ± 1.62 | 25.6–32.0 |
| Total PCA dose (normalized to morphine mg) | 98 | 46.7 ± 3.48 | 39.8–53.6 |
| Total ondansetron dose (mg) | 98 | 3.43 ± 0.36 | 2.71–4.15 |
| Total promethazine dose (mg) | 98 | 12.3 ± 1.41 | 9.53–15.1 |
| Total number of oxycodone/acetaminophen tablets | 92 | 10.7 ± 1.19 | 8.32–13.0 |
| Total number of ibuprofen tablets | 91 | 17.9 ± 0.58 | 16.8–19.0 |

P values

- The probability that any observation is due to chance alone assuming that the null hypothesis is true
 - Typically, an estimate that has a p value of 0.05 or less is considered to be “statistically significant” or unlikely to occur due to chance alone.

ERRORS

- Type I error
 - Claiming a difference between two samples when in fact there is none. (Rejecting the null hypothesis, when it is not true : **concern of Regulator**)
 - Remember there is variability among samples- they might seem to come from different populations but they may not.
 - Also called the α error.
 - Typically 0.05 is used

ERRORS

■ Type II error

- Results show there is no difference between two samples when in fact there is. (**concern of Sponsor**)
- Also called a β error.
- The probability of not making a Type II error is $1 - \beta$, which is called the power of the test.
- Hidden error because can't be detected without a proper power analysis

(Pearson's) Chi-Squared (χ^2) Test

- Used to compare observed proportions of an event compared to expected.
- Used with nominal data (better/ worse; dead/alive)

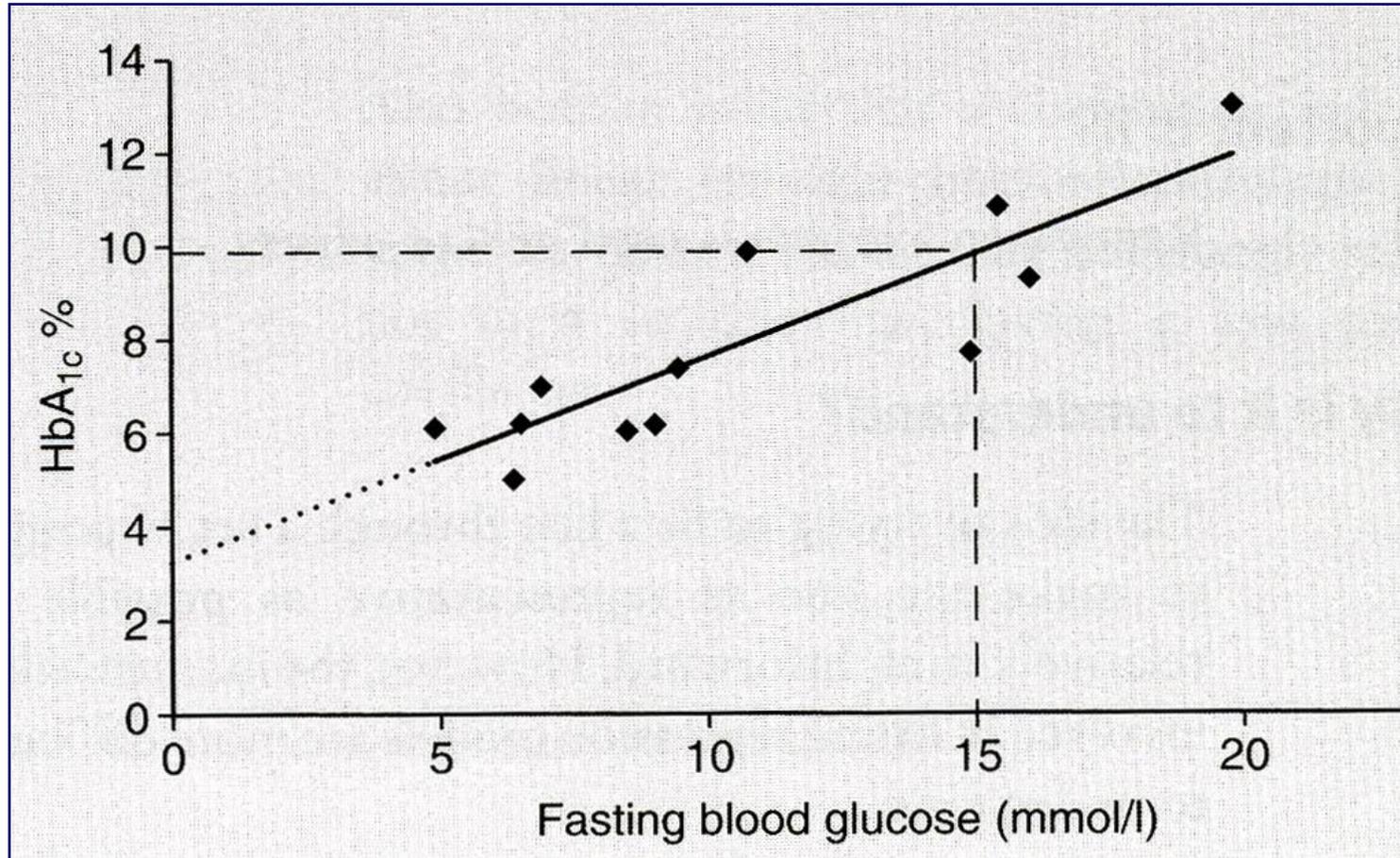
Example that uses 4 test in a single command:

“ . tab ecostatu1 gender, column row nokey
chi2 lrchi2 V exact gamma taub ”

- Chi-Squared (χ^2) Test
- Fisher's exact
- Gamma and taub (*Kruskal's γ (gamma)*)
- Cramer's V

| RECODE of ecostatu (Status of Nat'l Eco) | Gender of Respondent | | Total |
|--|----------------------|--------|--------|
| | Male | Female | |
| Well | 427 | 392 | 819 |
| | 52.14 | 47.86 | 100.00 |
| | 67.99 | 52.62 | 59.65 |
| Bad | 196 | 343 | 539 |
| | 36.36 | 63.64 | 100.00 |
| | 31.21 | 46.04 | 39.26 |
| Not sure/ref | 5 | 10 | 15 |
| | 33.33 | 66.67 | 100.00 |
| | 0.80 | 1.34 | 1.09 |
| Total | 628 | 745 | 1,373 |
| | 45.74 | 54.26 | 100.00 |
| | 100.00 | 100.00 | 100.00 |
| Pearson chi2(2) = 33.5266 Pr = 0.000 Likelihood-ratio chi2(2) = 33.8162 Pr = 0.000 Cramér's V = 0.1563 gamma = 0.3095 ASE = 0.050 Kendall's tau-b = 0.1553 ASE = 0.026 Fisher's exact = 0.000 | | | |

Correlation



Conclusions

Conclusion

- Do **NOT** follow the rest of the world because your requirements and budgets may be altogether different.
- Understand your statistical requirements first and then make the choice for an appropriate and best fit stats software.

The Roads not Taken

***Two roads diverged in
a yellow wood,
And sorry I could not
travel both***

***I took the one less
traveled by,
And that has made all
the difference.”***

- Robert Frost

