Introduction to Biostatistics (1/3: Basic Concepts)

Introduction to Biostatistics
Part I: Basic Concepts

Helmut Schütz
BEBAC

Biostatistics: Basic concepts & applicable principles for various designs in bioequivalence studies and data analysis | Mumbai, 29 – 30 January 2011
Biometry, Biometrics, and Biostatistics

- Introduced in 1947 by R.A Fisher as ‘Biometry’ and later ‘Biometrics’
  ‘Biometry, the active pursuit of biological knowledge by quantitative methods.’

- The International Biometric Society
  ‘The terms “Biometrics” and “Biometry” have been used since early in the 20th century to refer to the field of development of statistical and mathematical methods applicable to data analysis problems in the biological sciences. Recently, the term “Biometrics” has also been used to refer to the emerging field of technology devoted to identification of individuals […]’

- ‘Biostatistics’ was introduced as a new term…
Introduction to Biostatistics (1/3: Basic Concepts)

Biometry, Biometrics, and Biostatistics

**Statistics.** A subject which most statisticians find difficult but in which nearly all physicians are expert.

**Biostatistician.** One who has neither the intellect for mathematics nor the commitment for medicine but likes to dabble in both.

**Medical statistician.** One who will not accept that Columbus discovered America... because he said he was looking for India in the trial plan.

*Stephen Senn*
Introduction to Biostatistics (1/3: Basic Concepts)

Terminology I

- **high bias**
  - Low variance

- **low bias**
  - Low variance

Bias
Terminology II

- **discrete**
  - nominal scale
  - ordinal scale
  - distinctness
  - rank order

- **continuous**
  - interval scale
  - ratio scale
  - distinctness +
  - rank order +
  - interval +
  - ratio

Increasing information
Data I

- Nominal scale (aka categorial)
  - Sex, ethnicity,…
    - Statistics: mode, $\chi^2$ test
    - Transformations: equality

- Ordinal scale
  - School grades, disease states,…
    - Statistics: median, percentile, sign test, Wilcoxon test
    - Transformations: monotonic increasing order
Data II

● Interval scale
  ■ Calendar dates, temperature in °C, IQ,…
    ■ Statistics: mean, variance (standard deviation), correlation, regression, ANOVA
    ■ Transformations: linear

● Ratio scale
  ■ Measures with true zero point, temperature in K,…
    ■ Statistics: all of the above, geometric and harmonic mean, coefficient of variation
    ■ Transformations: multiplicative, logarithm
Examples from PK

- **Ordinal scale**
  - $t_{\text{max}}, t_{\text{lag}}$
  - Statistics: median, percentile, sign test, Wilcoxon test
  - Transformations: monotonic increasing order

- **Ratio scale**
  - $\text{AUC}, C_{\text{max}}, \lambda_z, \ldots$
  - Statistics: mean, variance (standard deviation), correlation, regression, ANOVA, geometric and harmonic mean, coefficient of variation
  - Transformations: multiplicativive, logarithm
Bell curve – and beyond

- Abraham de Moivre (1667–1754), Pierre-Simon Laplace (1749–1827)
  Central limit theorem 1733, 1812
- Carl F. Gauß (1777–1855)
  Normal distribution 1795
- William S. Gosset, aka Student (1876–1937)
  t-distribution 1908
- Ronald A. Fisher (1890–1962)
  Analysis of variance 1918
- Frank Wilcoxon (1892–1965)
  Nonparametric tests 1945
Statistical Distributions

- Normal distribution: mean = 100, CV = 30%

- Density plots for different sample sizes (n = 12, 48, 128, 1024):
  - n = 12: 0.000 0.005 0.010 0.015 0.020
  - n = 48: 0.000 0.005 0.010 0.015 0.020
  - n = 128: 0.000 0.005 0.010 0.015 0.020
  - n = 1024: 0.000 0.005 0.010 0.015 0.020

- Additional comments or analysis related to statistical distributions.
Introduction to Biostatistics (1/3: Basic Concepts)

Statistical Distributions

lognormal distr.: mean = 100 CV = 30 %

Density

0 50 100 150 200 250

0.000 0.005 0.010 0.015 0.020

n = 12

lognormal distr.: mean = 100 CV = 30 %

Density

0 50 100 150 200 250

0.000 0.005 0.010 0.015 0.020

n = 48

lognormal distr.: mean = 100 CV = 30 %

Density

0 50 100 150 200 250

0.000 0.005 0.010 0.015 0.020

n = 128

lognormal distr.: mean = 100 CV = 30 %

Density

0 50 100 150 200 250

0.000 0.005 0.010 0.015 0.020

n = 1024
**Statistical Distributions**

- **Normal Distribution**
  - Defined by location (aka central tendency) and dispersion
  - Population
    - Location: population mean $\mu$
    - Dispersion: population variance $\sigma^2$
  - Sample
    - Location: sample mean $\bar{x}$
    - Dispersion: sample variance $s^2$
  - Probability $= 1$ within $-\infty$ and $+\infty$

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

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$$
Statistical Distributions

- Lognormal Distribution
  - Defined by location and dispersion
  - Population
    - Location: population mean $\mu$
    - Dispersion: population variance $\sigma^2$
  - Sample
    - Location: sample mean $\bar{x}$
    - Dispersion: sample variance $s^2$
  - Probability $= 1$ within $0$ and $+\infty$

$$f(x) = \begin{cases} \frac{1}{\sigma x \sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

$$F(x) = \frac{1}{\sigma \sqrt{2\pi}} \int_{0}^{x} \frac{1}{t} e^{-\frac{(\ln t - \mu)^2}{2\sigma^2}} dt$$
Introduction to Biostatistics (1/3: Basic Concepts)

Statistical Distributions

normal distr.: mean = 100
sd = 20 CV = 20 %

normal distr.: mean = 100
sd = 30 CV = 30 %

normal distr.: mean = 110
sd = 22.09 CV = 20.08 %

normal distr.: mean = 100
sd = 25.5 CV = 25.5 %

normal distr.: mean = 120
sd = 24 CV = 20 %

n = 1000

n = 2000

n = 1000

n = 2000

Biostatistics: Basic concepts & applicable principles for various designs in bioequivalence studies and data analysis | Mumbai, 29 – 30 January 2011
**Statistical Distributions**

**Population:** mean = 100  
sd = 20.08 CV = 20.08%

**Sample 1:** mean = 101  
sd = 15.94 CV = 15.78%

**Sample 2:** mean = 100.1  
sd = 20.41 CV = 20.38%

**Sample 3:** mean = 100.2  
sd = 19.61 CV = 19.57%

**Sample 4:** mean = 96.69  
sd = 19.31 CV = 19.97%

**Sample 5:** mean = 102.5  
sd = 19.31 CV = 18.85%
Statistical Distributions

population: mean = 100
sd = 20.03 CV = 20.03 %

20 samples drawn from population
Central Limit Theorem

- If samples are drawn by a random process from a population with a normal distribution, distribution of sample means is also normal.
- The mean of the distribution of sample means is identical to the mean of the ‘parent population’ – the population from which the samples are drawn.
- The higher the sample size that is drawn, the ‘narrower’ will be the dispersion of the distribution of sample means.
Normal Distribution I

Standard normal distribution: $\mu = 0$, $\sigma = 1$

$\pm 1\sigma$ p 68.27%

$\pm 2\sigma$ p 95.45%

$\pm 3\sigma$ p 99.73%

$\pm 4\sigma$ p 99.99%
Normal Distribution II

Standard normal distribution: $\mu = 0, \sigma = 1$

- 90% $\pm 1.645\sigma$
- 95% $\pm 1.960\sigma$
- 99% $\pm 2.576\sigma$
- 99.9% $\pm 3.291\sigma$
Normal Distribution III

Standard normal distribution: $\mu = 0, \sigma = 1$

- $[\pm 1.645]$ 90%
- $(-\infty, +1.645]$ 95%
- $[-1.645, +\infty]$ 95%
Confidence Interval I

- If we have drawn a sample from a population, we get the sample mean $\bar{x}$ and the sample standard deviation $s$.

- Can we make a prediction about the population mean?

- Yes. That’s called a Confidence Interval (CI).
  - If $\sigma$ is known:
    $$[\mu] = \bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$
Confidence Interval II

- Example from previous slides: \( \mu = 100, \sigma = 20 \)
- Sample sizes 36, \( z_{0.05} = 1.960 \)

<table>
<thead>
<tr>
<th>Samples’ means</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>101.0</td>
<td>94.47 107.5</td>
</tr>
<tr>
<td>100.1</td>
<td>93.57 106.6</td>
</tr>
<tr>
<td>100.2</td>
<td>93.67 106.7</td>
</tr>
<tr>
<td>96.69</td>
<td>90.16 103.2</td>
</tr>
<tr>
<td>102.5</td>
<td>95.97 109.0</td>
</tr>
</tbody>
</table>

- But generally we don’t know \( \sigma \)!
- Help is on the way…
Student’s $t$ Distribution

- Depends on one parameter, the ‘degrees of freedom $\nu$’. In the most simple case $\text{df} = n - 1$.
- The $t$ Distribution is ‘heavy tailed’ compared to the normal distribution. Small sample sizes are penalized.
- Approaches quickly the normal distribution for $\text{df} \geq 30$.
- Allows calculation of a CI of the sample mean based on the sample standard deviation $s$. 

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

$$\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}dt$$
Student’s \( t \) Distribution

- **Student’s \( t \) distribution: \( v = 1 \)**
  - \( n = 2 \)

- **Student’s \( t \) distribution: \( v = 5 \)**
  - \( n = 6 \)

- **Student’s \( t \) distribution: \( v = 11 \)**
  - \( n = 12 \)

- **Student’s \( t \) distribution: \( v = 35 \)**
  - \( n = 36 \)
**Confidence Interval III**

- Example from previous slides: \( \mu = 100, \sigma = 20 \)
- Sample sizes 36, \( z_{0.05} = 1.960, t_{36-1,0.05} = 2.030 \)

<table>
<thead>
<tr>
<th>Samples’ mean</th>
<th>Samples’ stand. dev.</th>
<th>Confidence Intervals based on ( z )</th>
<th>Confidence Intervals based on ( t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>101.0</td>
<td>15.94</td>
<td>94.47</td>
<td>107.5</td>
</tr>
<tr>
<td>100.1</td>
<td>20.41</td>
<td>93.57</td>
<td>106.6</td>
</tr>
<tr>
<td>100.2</td>
<td>19.61</td>
<td>93.67</td>
<td>106.7</td>
</tr>
<tr>
<td>96.69</td>
<td>19.31</td>
<td>90.16</td>
<td>103.2</td>
</tr>
<tr>
<td>102.5</td>
<td>19.31</td>
<td>95.97</td>
<td>109.0</td>
</tr>
</tbody>
</table>

\[
[\mu] = \bar{x} \pm z \frac{\sigma}{\sqrt{n}} \quad [\mu] = \bar{x} \pm t \frac{s}{\sqrt{n}}
\]
Location parameters

• $x = [91, 72, 141, 119, 92, 124, 92, 101, 90, 145]$ ranks $[3, 1, 9, 7, 4.5, 8, 4.5, 6, 2, 10]$ ordered $[72, 90, 91, 92, 92, 101, 119, 124, 141, 145]$

• **Mode**: 92 (most frequent number)

• **Median**: 96.5 (middle value)
  - If $n$ = odd: value at $x_{n/2}$
  - If $n$ = even: value $(x_{n/2} + x_{n/2+1})/2 = (92+101)/2$
Introduction to Biostatistics (1/3: Basic Concepts)

Location parameters

- **Harmonic mean**: 101.9516
  
  \[ \bar{x}_{\text{harm}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_i}} \]

- **Geometric mean**: 104.2814
  
  \[ \bar{x}_{\text{geom}} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n} = e^{\frac{1}{n} \sum_{i=1}^{n} \ln x_i} \]

- **Arithmetic mean**: 106.7
  
  \[ \bar{x}_{\text{arithm}} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_i}{n} \]
A note on the harmonic mean

- Driving at 100 km/h from A to B; distance is 100 km.
- Driving back at 50 km/h.
- What is the average speed for the round-trip?
  - 75 km/h
  - 70.71 km/h
  - **66.67 km/h**
- 1 h for 100 km (A→B) and 2 h for 100 km (A←B); 200 km/3 h = 66.67 km/h.
- Harmonic mean for rates!
  \[
  \bar{x}_{\text{harm}} = \frac{2}{\frac{1}{100} + \frac{1}{50}} = \frac{2}{0.01 + 0.02} = 66.6
  \]
Location parameters

- Application of *any* location parameter *always* (!) implies an underlying distributional assumption.
  - Median: discrete (or unknown)
  - Arithmetic mean: normal distribution
  - Geometric mean: lognormal distribution
  - Harmonic mean: rates

- Example from above sampled from a lognormal distribution
  - Arithmetic mean: 106.7 (too high!)
  - Geometric mean: 104.3 (correct)
**Location parameters**

![Boxplot](image)

**Boxplot**

- **Linear scale**
- **Logarithmic scale**

---

**Boxplot**

- **Location parameters**
Nitpicking terminology

• If we are estimating parameters of a distribution, we are using **Estimators**; *e.g.*, the arithmetic mean is the unbiased estimator of the central tendency of the normal distribution.

• The numerical outcomes (*i.e.*, values one give in the report) are **Estimates**.

• Don’t write something like

  ‘*The point estimator was 95.34 %.*’

  … when it was actually a maximum likelihood estimator based on least squares means in log-scale 😊
Dispersion parameters

- ordered [72,90,91,92,92,101,119,124,141,145]
- **Quartiles** (25%, 75%): Be cautious! Different methods implemented in software...
  - 90.00, 124.00: SAS
  - 91.25, 122.75: S, R, M$-Excel$
  - 90.75, 128.25: Minitab, SPSS, Phoenix/WinNonlin
  - 90.92, 125.42: Hyndman & Fan (1996)
  - … and many others!
Dispersion parameters

- **Standard deviation (SD) of arithmetic mean:**
  24.2400

\[
SD_{\text{arithm}} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}
\]

- **SD of geometric mean:**
  23.8000

\[
SD_{\text{geom}} = e^{\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (\ln x_i - \ln \bar{x}_{\text{geom}})^2}}
\]
Dispersion parameters

- SD of harmonic mean: 22.8519

\[
SD_{harm} = \sqrt{(n - 1) \sum_{i=1}^{i=n} (\bar{H}_i - \bar{H})^2}
\]

\[
\bar{H} = \frac{1}{n} \sum_{i=1}^{i=n} \bar{H}_i
\]

\[
\bar{H}_i = \frac{n - 1}{\left( \sum_{i=n}^{i=n} \frac{1}{x_j} \right) - \frac{1}{x_i}}
\]
# Dispersion parameters

- **Coefficient of Variation**
  (sometimes given in percent of mean):

\[
CV\% = 100 \cdot \frac{SD}{\bar{x}}
\]

<table>
<thead>
<tr>
<th>Population (N=10^6), parameters</th>
<th>Sample (n=36), parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu )</td>
<td>100.00</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>20.00</td>
</tr>
<tr>
<td>( CV% )</td>
<td>20.00</td>
</tr>
</tbody>
</table>
A remark on Variances

- Whilst means and variances are additive, standard deviations (and CVs as well) are not!

<table>
<thead>
<tr>
<th>sample</th>
<th>mean</th>
<th>s</th>
<th>s²</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>20</td>
<td>400</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>30</td>
<td>900</td>
</tr>
<tr>
<td>1+2</td>
<td>(100+100)/2</td>
<td>25??</td>
<td>[\sum s^2/2]</td>
</tr>
</tbody>
</table>

mean = 100 sd = 20 CV = 20 %
mean = 100 sd = 30 CV = 30 %
mean = 100 sd = 25.5 CV = 25.5 %
Arithm. vs. geom. means

Arithmetic mean (95% CI)

Geometric mean (95% CI)
Median and quantiles

Median (5%, 25%, 75%, 95% quantile)
Degrees of freedom...

For every estimated parameter in a statistical model one degree of freedom is ‘lost’ from the number of samples (df = n – p).

- Any model becomes useless if df=0, and impossible to fit if df<0 (p>n). Example:
  - Linear regression: Two parameters are fit (slope, intercept; since any line is defined by two points \((x_1/y_1|x_2,y_2)\) at least three data points are needed (df=1).
Degrees of freedom...

mean = 100, sd = 20

sample # 1

sample # 2

sample # 3

sample # 4

sample # 5

sample # 6

sample # 7

sample # 8

sample # 9

sample # 10

sample # 11

sample # 12

sample # 13

sample # 14

sample # 15
Degrees of freedom...
Visualize your data!

Anscombe’s Quartet (1973)

All datasets:
- mean $x = 9.0, s^2_x = 10$
- mean $y = 7.5, s^2_y = 3.75$
- Corr$_{yx} = 0.898$
- Regr$_{yx} y = 3 + 0.5x$

Don’t rely *solely* on numerical results.

---

**Anscombe’s Quartet (1973)**

**Visualize your data!**

- **Correct**
  - Correct model:
  - $y = 3 + 0.5x$

- **Wrong model:**
  - Quadratic!

- **Correct model, but biased by outlier**
  - Nonsense

Don’t rely *solely* on numerical results.
Data Transformation?

Clearly in favor of a lognormal distribution. Shapiro-Wilk test highly significant for normal distribution (rejected).
Data set from a real study. Both tests *not* significant (assumed distributions not rejected).

Tests not acceptable according to GLs; log-transformation based on prior knowledge (PK)!
Data Transformation

- BE testing started in the early 1980s with an acceptance range of 80% – 120% of the reference based on the normal distribution.
- Was questioned in the mid 1980s
  - Like many biological variables AUC and $C_{\text{max}}$ do not follow a normal distribution
    - Negative values are impossible
    - The distribution is skewed to the right
    - Might follow a lognormal distribution
  - Serial dilutions in bioanalytics lead to multiplicative errors
Data Transformation: PK

\[ F_T = \frac{AUC_T \cdot CL_T}{D_T}, \quad F_R = \frac{AUC_R \cdot CL_R}{D_R} \]

\[ F_{rel}(BA) = \frac{AUC_T}{AUC_R} \]

Assumption 1: \( D_1 = D_2 \) \((D_1/D_2 = 1^*)\)

Assumption 2: \( CL_1 = CL_2 \)
Data Transformation

‘Problems’ with logtransformation

- If we transform the ‘old’ acceptance limits of 80% – 120%, we get –0.2231, +0.1823.
- These limits are *not symmetrical* around 100% any more, the maximum power is obtained at $e^{0.1823-0.2231} = 96\%$...

- Solution:
  lower limit = 1 – 0.20, upper limit = 1/lower limit
  $ln(0.80) = -0.2231$ and $ln(1.25) = +0.2231$. Symmetrical around 0 in the log-domain and around 100% in the backtransformed domain ($e^0=1$).
Data Transformation

- ‘Problems’ with log-transformation
  - Discussion, whether more bioinequivalent formulations will pass due to ‘5% wider’ limits
    - lower limit = 1 – 0.20, upper limit = 1/lower
      - 80.00% – 125.00% (width 45.00%)
    - instead of keeping the ‘old’ width
      - lower limit = 1 – 0.1802, upper limit = 1/lower
      - 81.98% – 121.98% (width 40.00%)
    - or even become more strict by setting
      - upper limit = 1 + 0.20, lower limit = 1/upper
      - 83.33% – 120.00% (width 36.67%)
  - 80% – 125% was chosen for convenience (!)
F Distribution

- Allows comparison of variances (depending on $\nu$) of two distributions. We will need that in ANOVA.

\[
F(x | \nu_1, \nu_2) = \begin{cases} 
\frac{\frac{\nu_1}{2} x}{x^2} \cdot \frac{\Gamma\left(\frac{\nu_1}{2} + \frac{\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right) \Gamma\left(\frac{\nu_2}{2}\right)} \cdot \frac{\nu_1 x - 1}{\nu_2 x} & x \geq 0 \\
0 & x < 0
\end{cases}
\]

\[
\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt
\]

- Note that if one of the degrees of freedom $= 1$, there is a relationship to the $t$ distribution:

\[
F(\nu_1 = 1, \nu_2 = \nu) = (t(\nu))^2
\]
Significance tests

- In statistics (as well as in science in general) it is not possible to **prove** something.
- We can only state a hypothesis and try to **reject** this so called **null hypothesis** by evaluating data from an experiment.
- Example:
  - $H_0: \mu_1 = \mu_2$ (no difference in means, null hypothesis)
  - $H_a: \mu_1 \neq \mu_2$ (different means; alternative hypothesis)
**α- vs. β-Error**

- All formal decisions are subjected to two types of error:
  - Error Type I (α-Error, Risk Type I)
  - Error Type II (β-Error, Risk Type II)

Example from the justice system:

<table>
<thead>
<tr>
<th>Verdict</th>
<th>Defendant innocent</th>
<th>Defendant guilty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presumption of innocence not accepted</td>
<td>Error type I</td>
<td>Correct</td>
</tr>
<tr>
<td>(guilty)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Presumption of innocence accepted</td>
<td>Correct</td>
<td>Error type II</td>
</tr>
<tr>
<td>(not guilty)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

α-Error: False Positive (Type I Error)

β-Error: False Negative (Type II Error)
---

### Biostatistics: Basic Concepts & Applicable Principles for Various Designs

**Introduction**

Biostatistics is the application of statistical methods to biological data. It involves the design, collection, analysis, and interpretation of biological data.

**Bioequivalence Studies**

Bioequivalence studies are designed to compare the systemic availability of two formulations of a drug.

**Data Analysis**

Analysis of biological data involves the use of statistical methods to draw meaningful conclusions from the data.

---

**α- vs. β-Error**

- **α-Error (Type I Error)**: Rejecting the null hypothesis when it is true.
- **β-Error (Type II Error)**: Failing to reject the null hypothesis when it is false.

**In BE-testing the null hypothesis is bioequivalence ($\mu_1 \neq \mu_2$)!

<table>
<thead>
<tr>
<th>Decision</th>
<th>Null hypothesis true</th>
<th>Null hypothesis false</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null hypothesis rejected</td>
<td></td>
<td>Correct ($H_a$)</td>
</tr>
<tr>
<td>Failed to reject null hypothesis</td>
<td>Correct ($H_0$)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Error type II</td>
<td></td>
</tr>
</tbody>
</table>

---

**Decision**

<table>
<thead>
<tr>
<th>Decision</th>
<th>Null hypothesis true</th>
<th>Null hypothesis false</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null hypothesis rejected</td>
<td>Patients’ risk</td>
<td>Correct (BE)</td>
</tr>
<tr>
<td>Failed to reject null hypothesis</td>
<td>Correct (not BE)</td>
<td>Producer’s risk</td>
</tr>
</tbody>
</table>

---

Biostatistics: Basic concepts & applicable principles for various designs in bioequivalence studies and data analysis | Mumbai, 29 – 30 January 2011

---
**α- vs. β-Error**

- **α-Error**: Patients’ Risk to be treated with a bioinequivalent formulation (H₀ falsely rejected)
  - BA of the test compared to reference in a particular patient is risky *either* below 80% *or* above 125%.
  - If we keep the risk of particular patients at 0.05 (5%), the risk of the entire population of patients (<80% *and* >125%) is $2\times\alpha$ (10%) is: 90% CI = $1 - 2\times\alpha = 0.90$
**α- vs. β-Error**

- **β-Error**: Producer’s Risk to get no approval for a bioequivalent formulation (H₀ falsely not rejected)
  - Set in study planning to ≤0.2, where power = 1 − β = ≥80%
  - If power is set to 80%
    - One out of five studies will fail just by chance!

<table>
<thead>
<tr>
<th>α 0.05</th>
<th>BE</th>
</tr>
</thead>
<tbody>
<tr>
<td>not BE</td>
<td>β 0.20</td>
</tr>
</tbody>
</table>
Significance test \((\alpha - \text{ vs. } \beta)\)
Part I: Basic Concepts

Helmut Schütz

BEBAC

Consultancy Services for Bioequivalence and Bioavailability Studies
1070 Vienna, Austria

helmut.schuetz@bebac.at
To bear in Remembrance...

In these matters the only certainty is that nothing is certain.  

*Gaius Plinius Secundus (Pliny the Elder)*

The theory of probabilities is at bottom nothing but common sense reduced to calculus.  

*Pierre-Simon Laplace*

It is a good morning exercise for a research scientist to discard a pet hypothesis every day before breakfast.  

It keeps him young.  

*Konrad Lorenz*