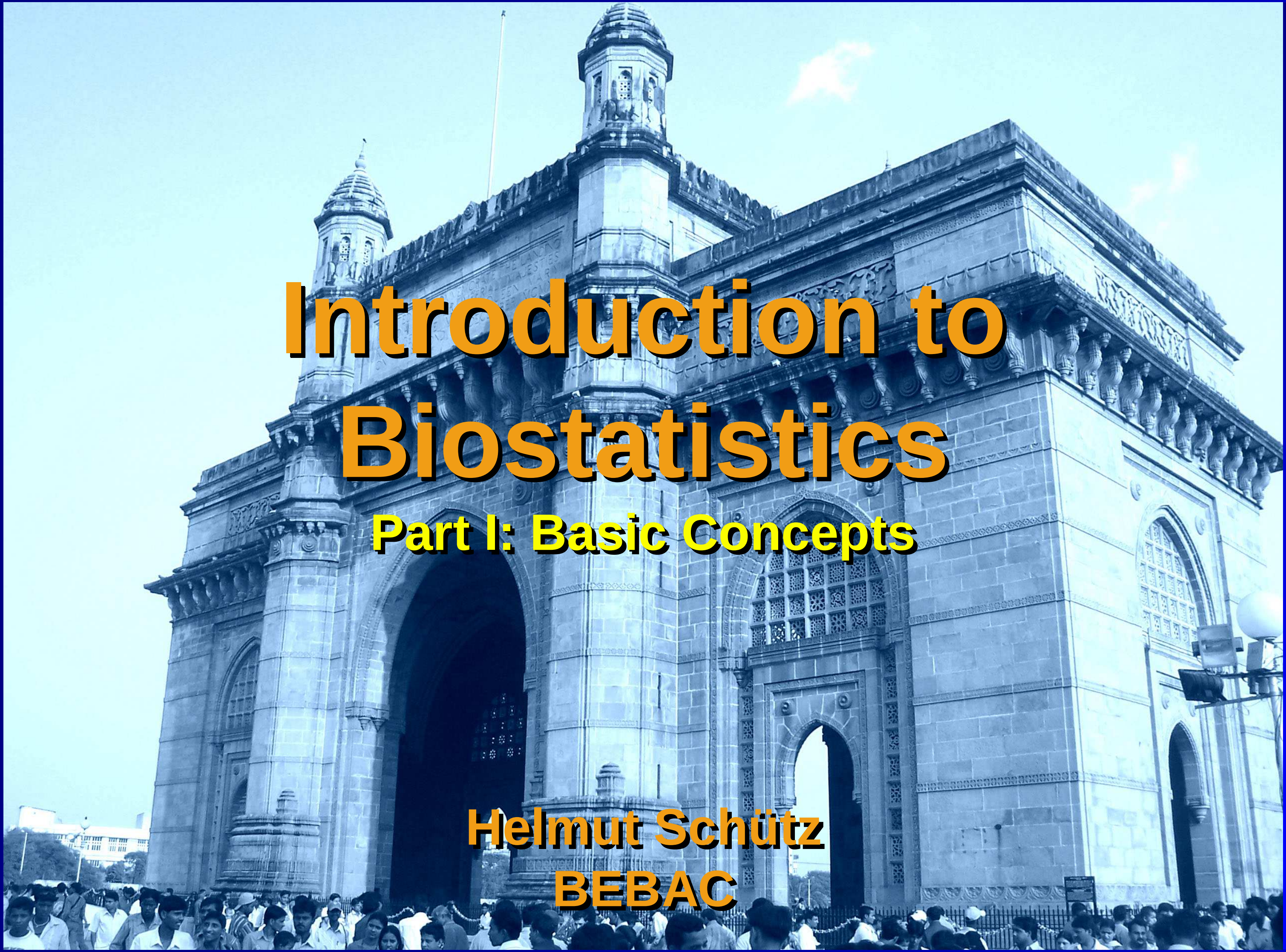


Introduction to Biostatistics

Part I: Basic Concepts

Helmut Schütz
BEBAC

BE
BAC

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…

Biometry, Biometrics, and Biostatistics

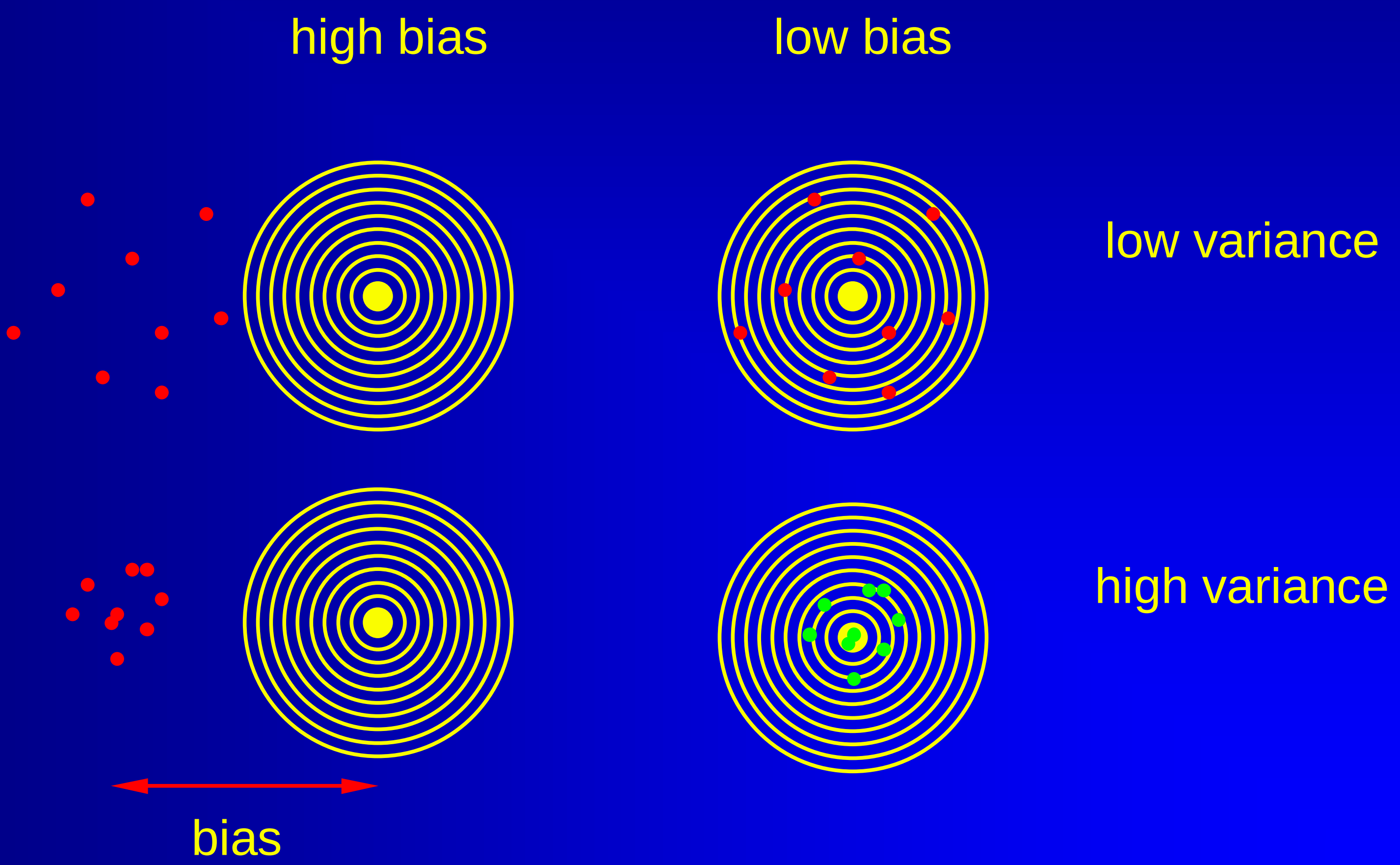
Statistics. A subject which most statisticians find difficult but in which nearly all physicians are expert.

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

Terminology I



Terminology II

- data
  - discrete
    - nominal scale
    - ordinal scale
  - continuous
    - interval scale
    - ratio scale

nominal scale	ordinal scale	interval scale	ratio scale
distictness	distictness +	distictness +	distictness +
	rank order	rank order +	rank order +
		interval	interval +
			ratio

increasing information →

Data I

- Nominal scale (aka categorial)
  - Sex, ethnicity,…
    - Statistics: mode, χ^2 test
    - Transformations: equality
- Ordinal scale
  - School grades, disease states,…
    - Statistics: median, percentile, sign test, Wilcoxon test
    - Transformations: monotonic increasing order

Data II

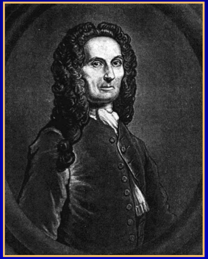
- Interval scale
  - Calendar dates, temperature in ℃, IQ,…
    - Statistics: mean, variance (standard deviation), correlation, regression, ANOVA
    - Transformations: linear
- Ratio scale
  - Measures with true zero point, temperature in K,…
    - Statistics: all of the above, geometric and harmonic mean, coefficient of variation
    - Transformations: multiplicative, logarithm

Examples from PK

- Ordinal scale
  - t_{max}, t_{lag}
    - Statistics: median, percentile, sign test, Wilcoxon test
    - Transformations: monotonic increasing order
- Ratio scale
  - AUC, C_{max}, λ_z,…
    - Statistics: mean, variance (standard deviation), correlation, regression, ANOVA, geometric and harmonic mean, coefficient of variation
    - Transformations: multiplicative, logarithm

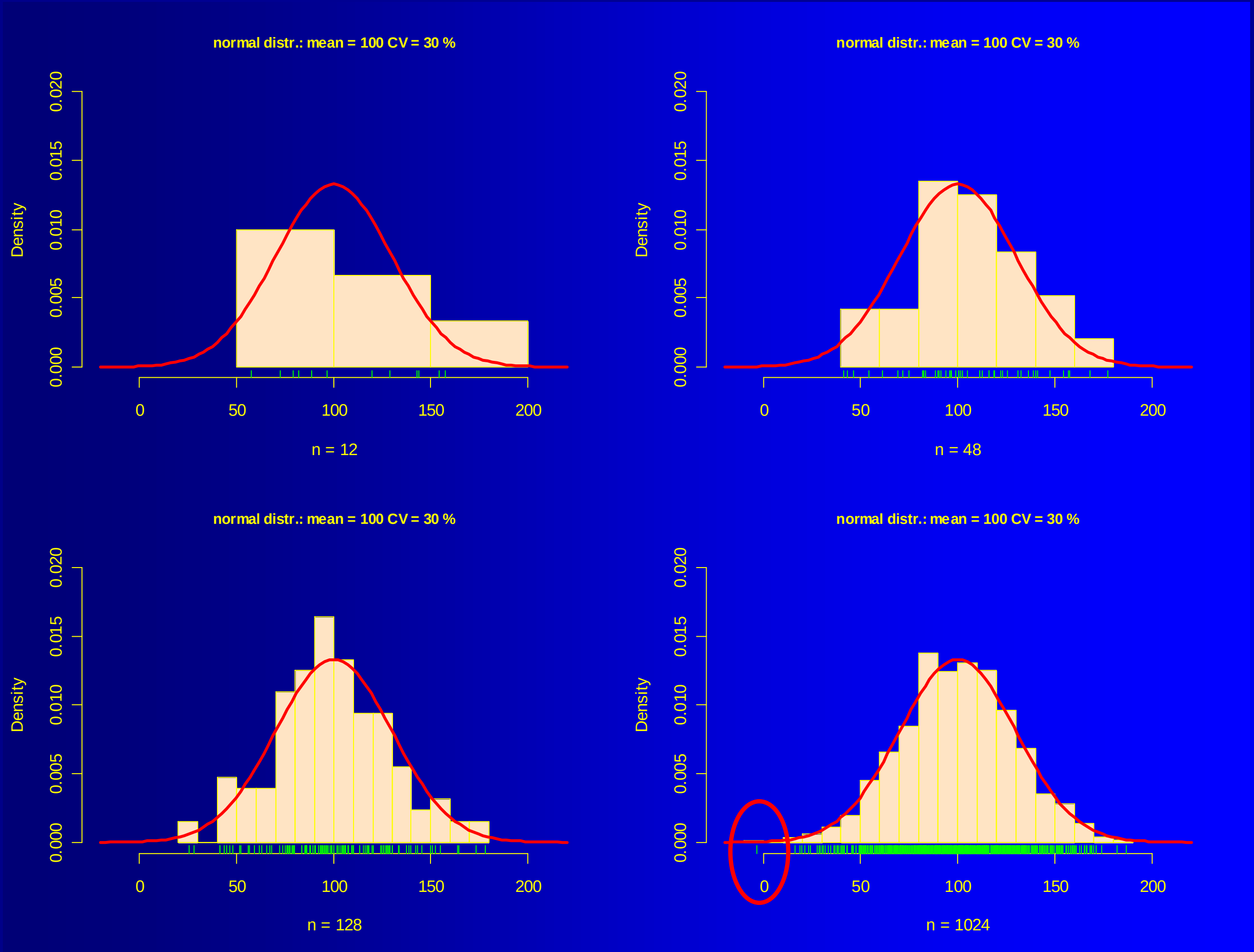
Bell curve – and beyond

- 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

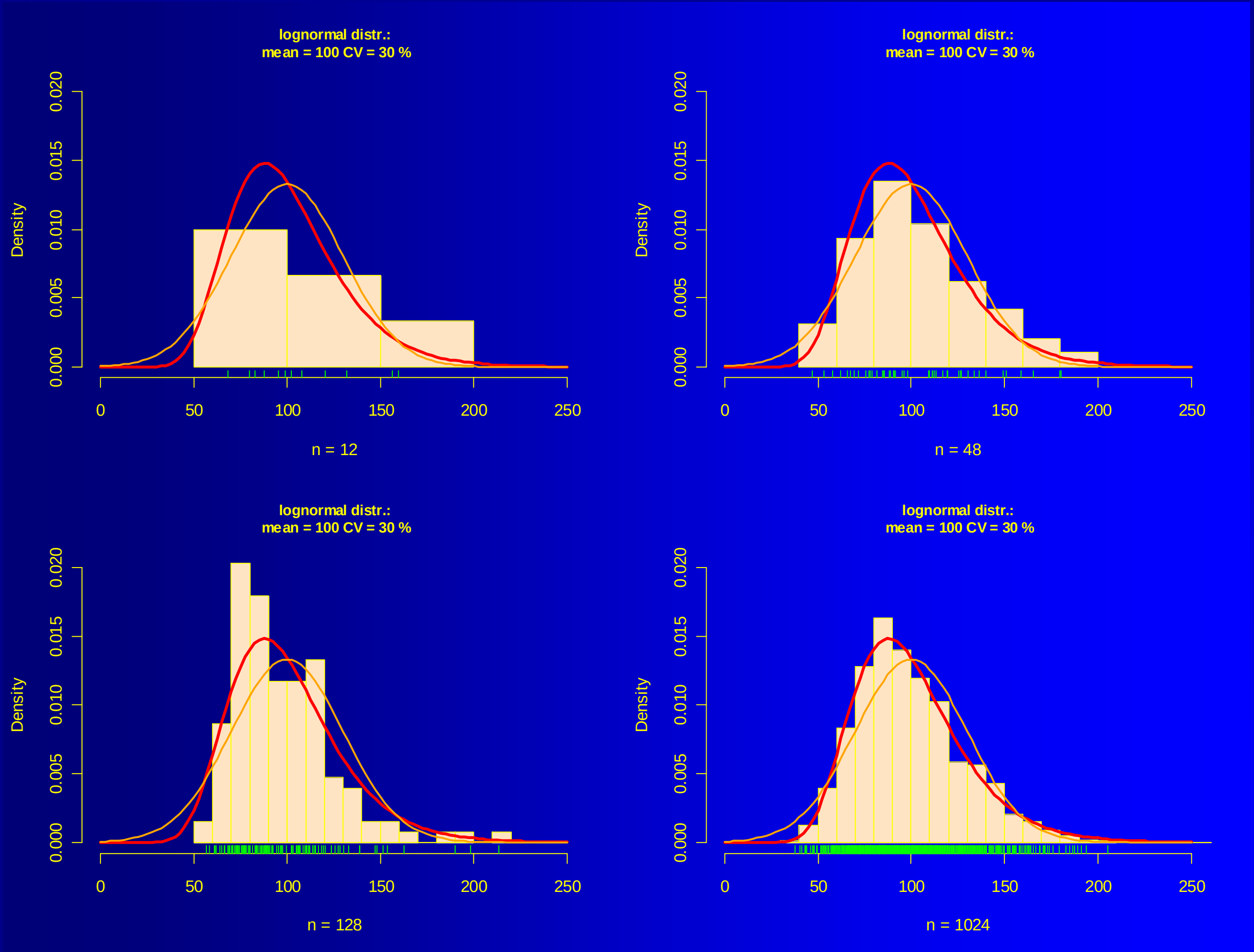





Statistical Distributions



Statistical Distributions



Statistical Distributions

- Normal Distribution

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

  - Defined by location (aka central tendency) and dispersion
  - Population
    - Location: population mean μ
    - Dispersion: population variance σ^2
  - Sample
    - Location: sample mean $\bar{x}$
    - Dispersion: sample variance s^2
  - Probability = 1 within -∞ and +∞

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

BE BAC

Statistical Distributions

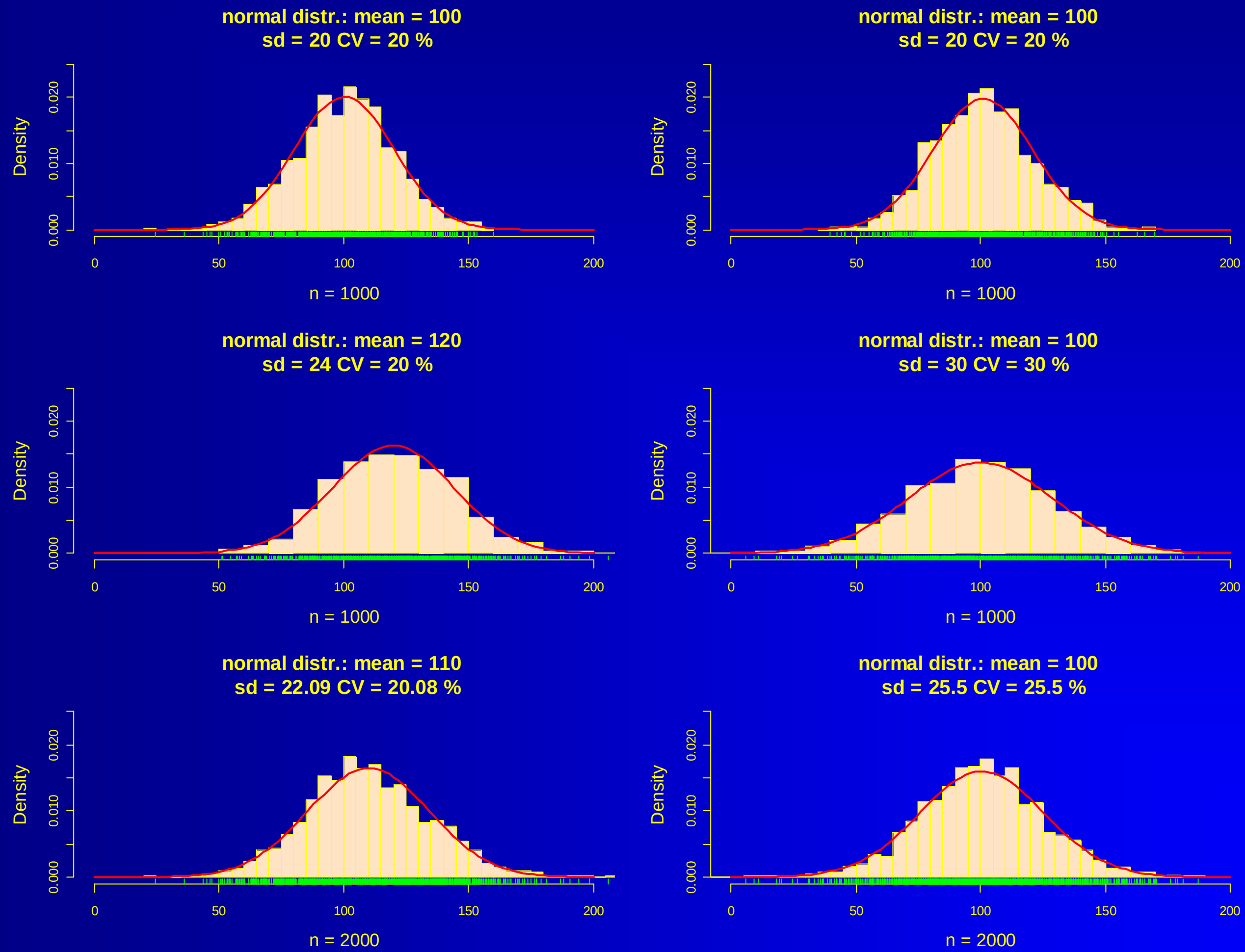
- Lognormal Distribution

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

  - Defined by location and dispersion
  - Population
    - Location: population mean μ
    - Dispersion: population variance σ^2
  - Sample
    - Location: sample mean $\bar{x}$
    - Dispersion: sample variance s^2
  - Probability = 1 within 0 and +∞

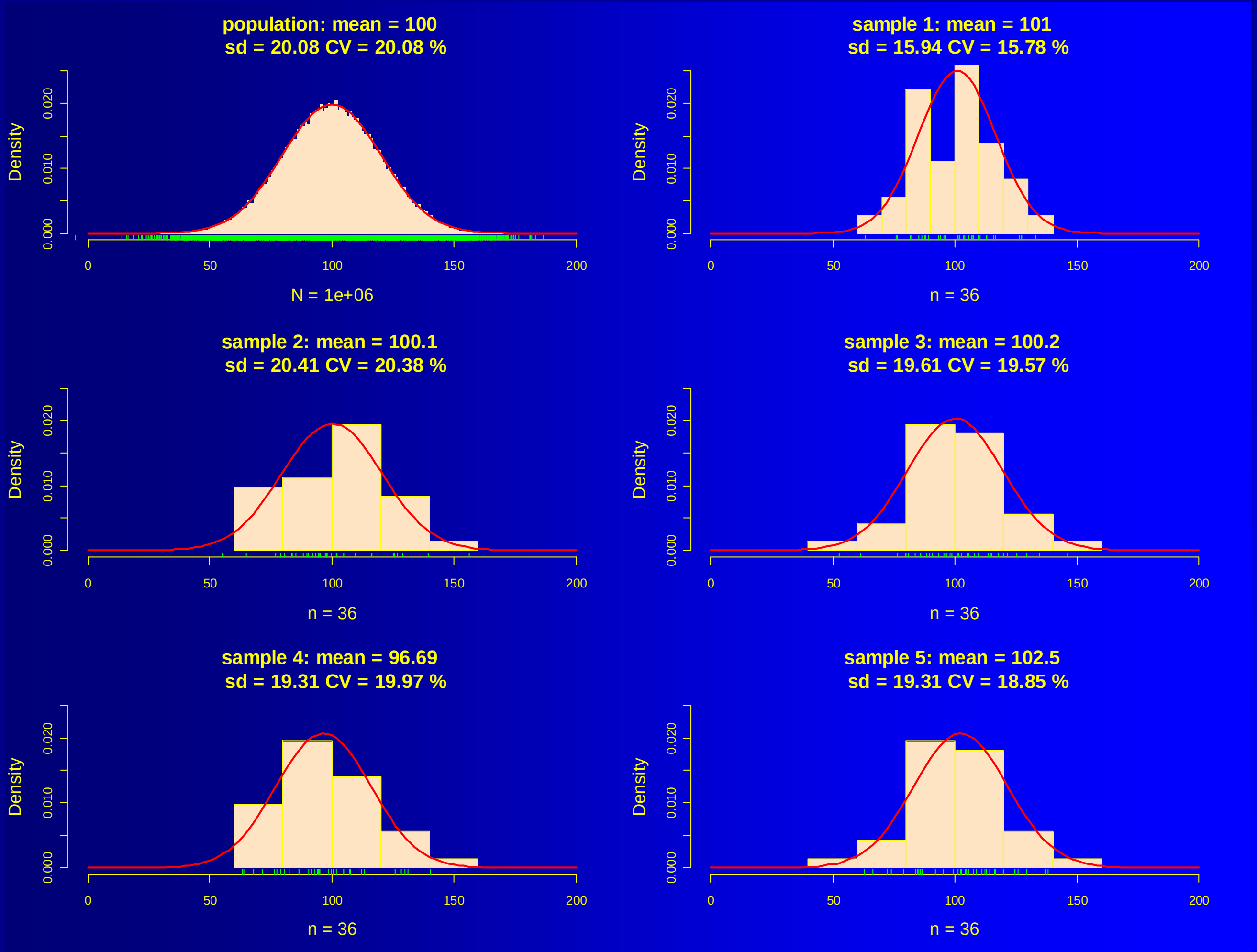
$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_0^x \frac{1}{t}e^{-\frac{(\ln t-\mu)^2}{2\sigma^2}}dt$$

Statistical Distributions

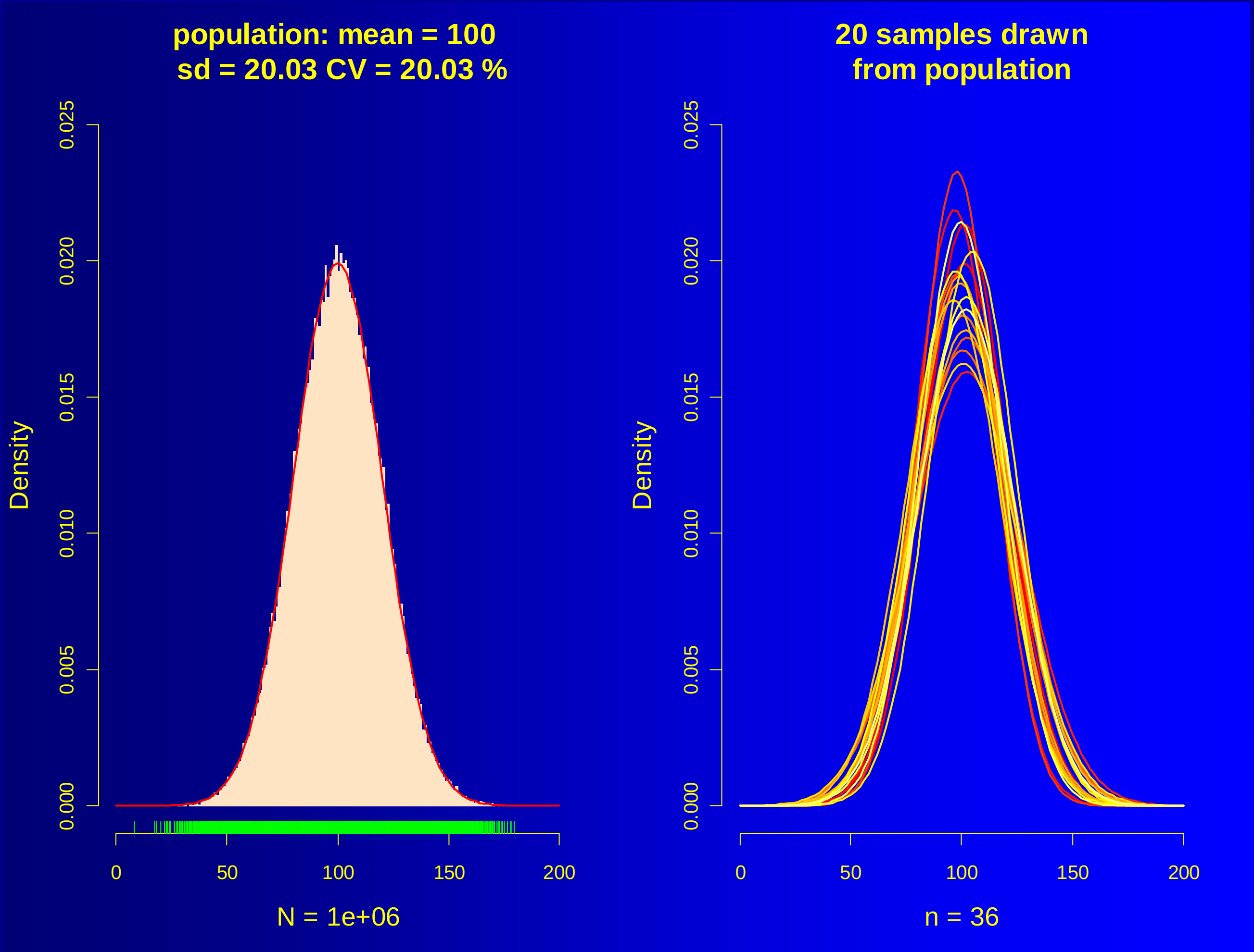


BE
BAC

Statistical Distributions



Statistical Distributions

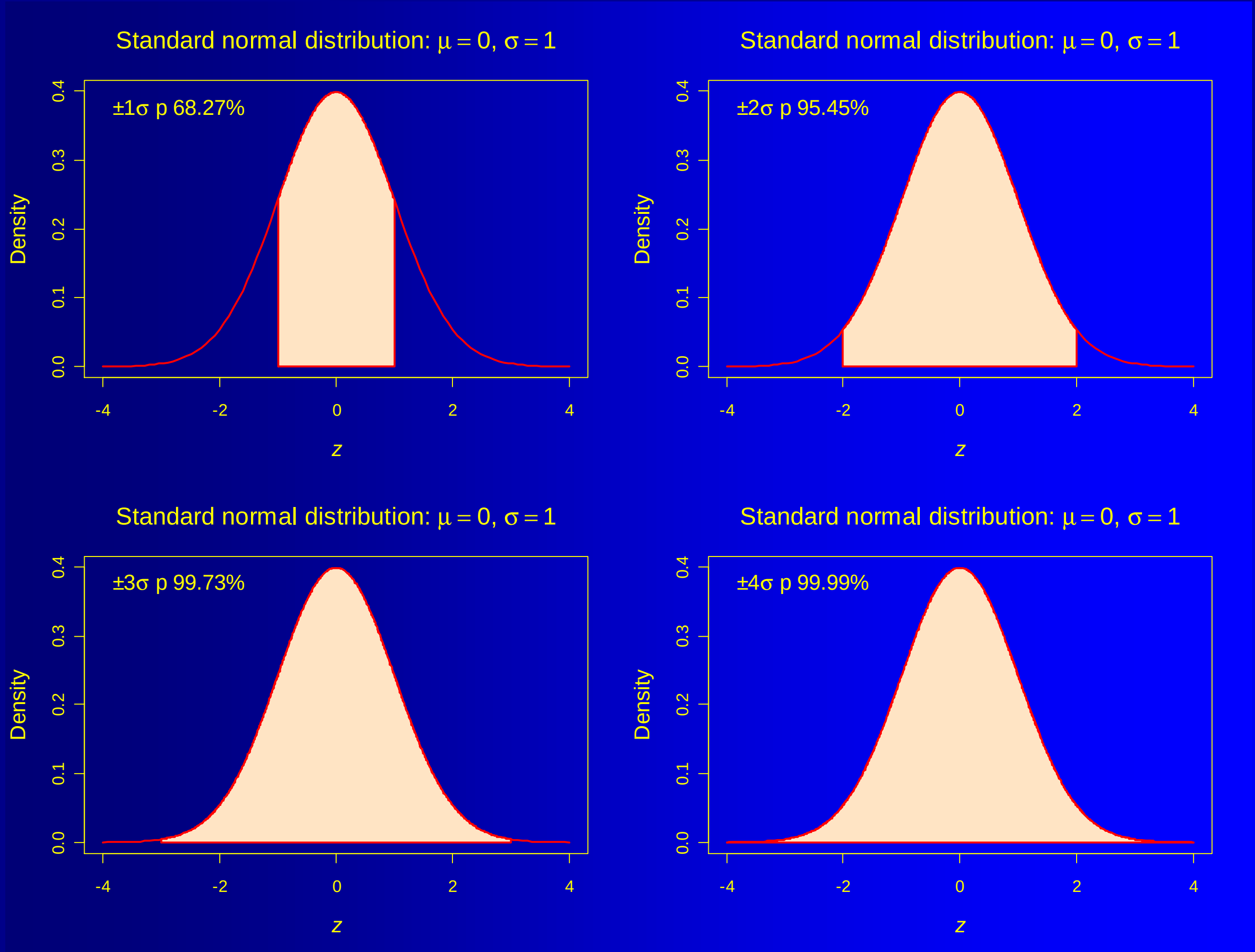


BE BAC

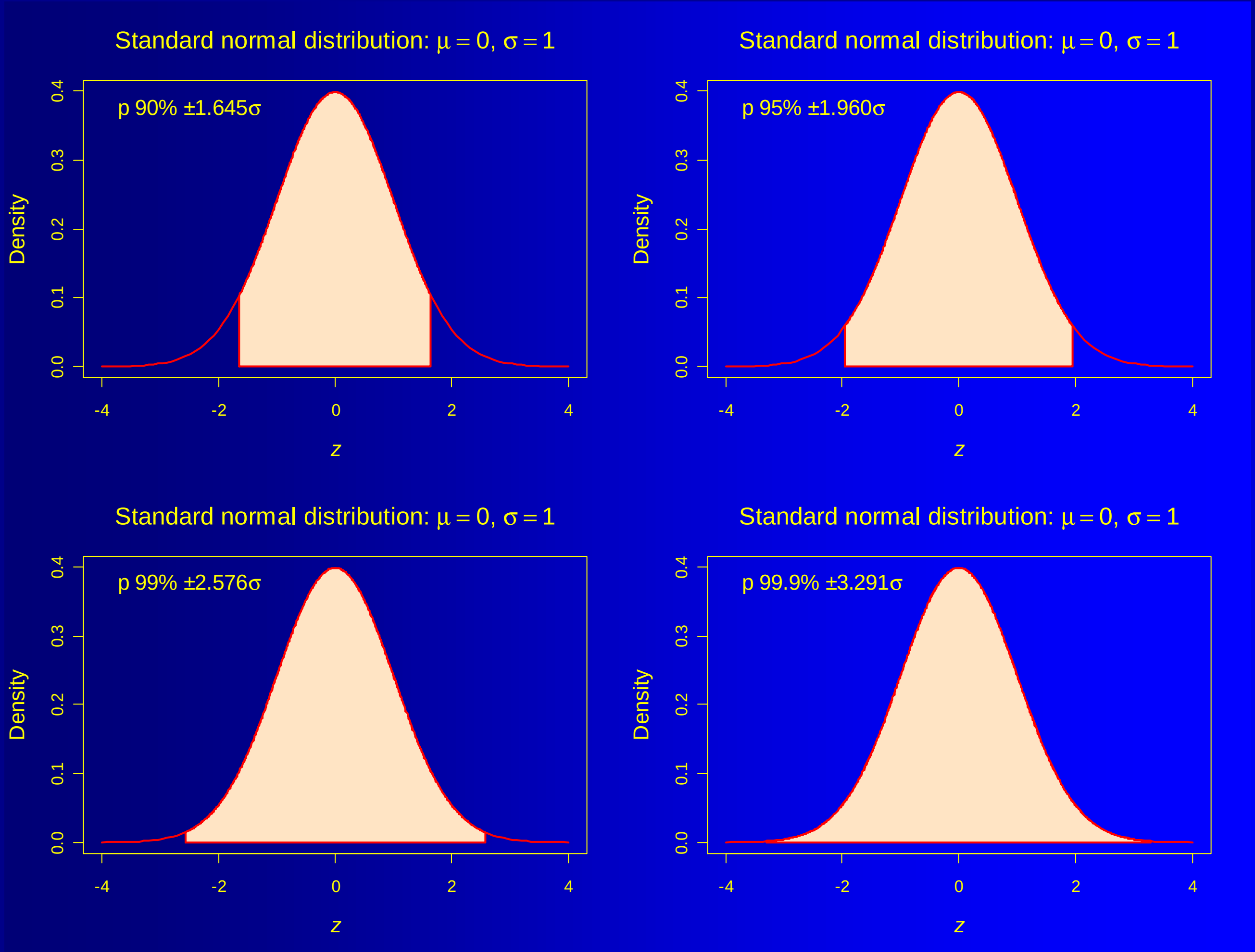
Central Limit Theorem

- 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.

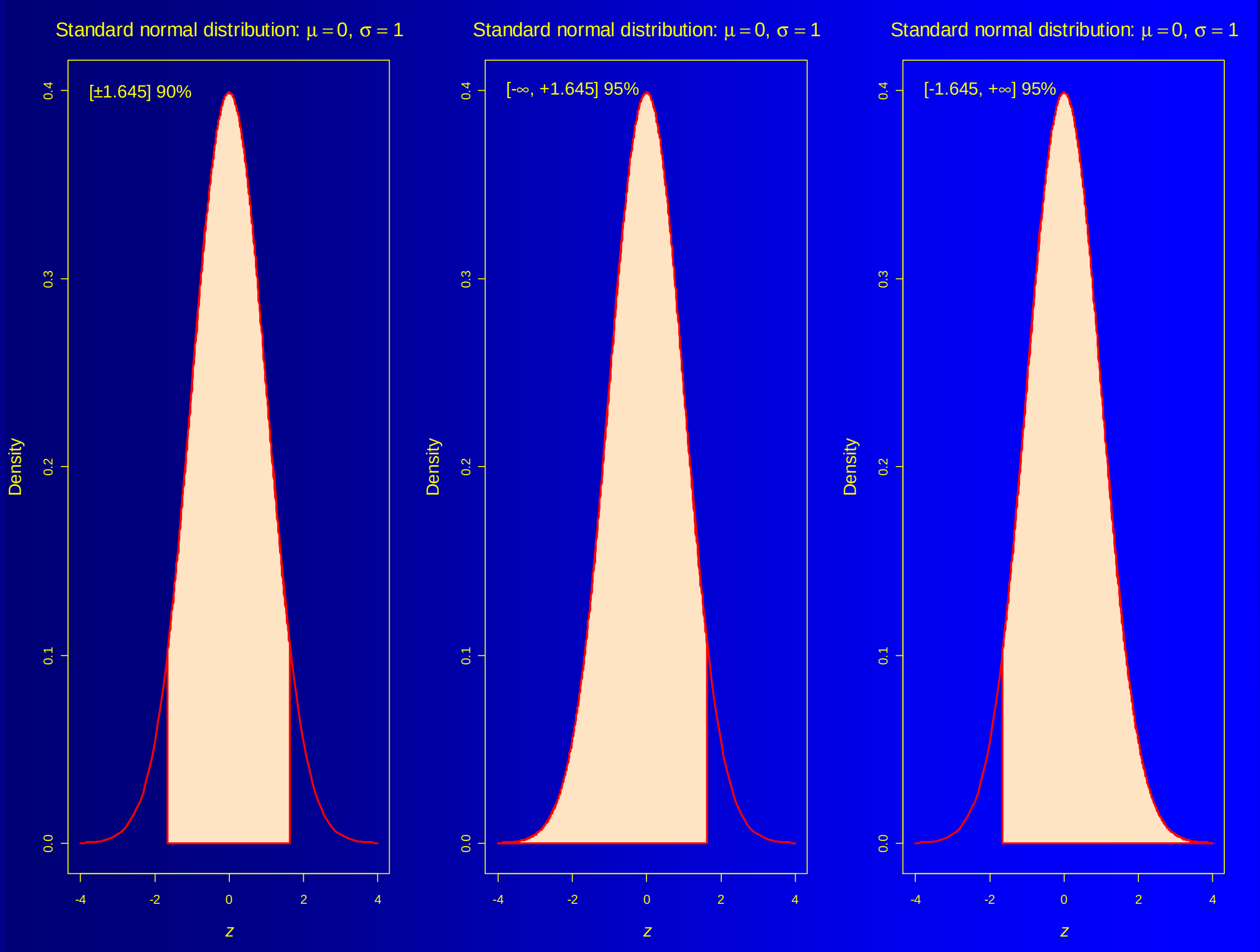
Normal Distribution I



Normal Distribution II



Normal Distribution III



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 σ is known:

$$[\mu] = \bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

Confidence Interval II

- Example from previous slides: μ 100, σ 20
  Sample sizes 36, $z_{0.05}$ 1.960

Samples' means	Confidence Interval	
101.0	94.47	107.5
100.1	93.57	106.6
100.2	93.67	106.7
96.69	90.16	103.2
102.5	95.97	109.0

- But generally we don't know σ!
- Help is on the way…

Student's *t* Distribution

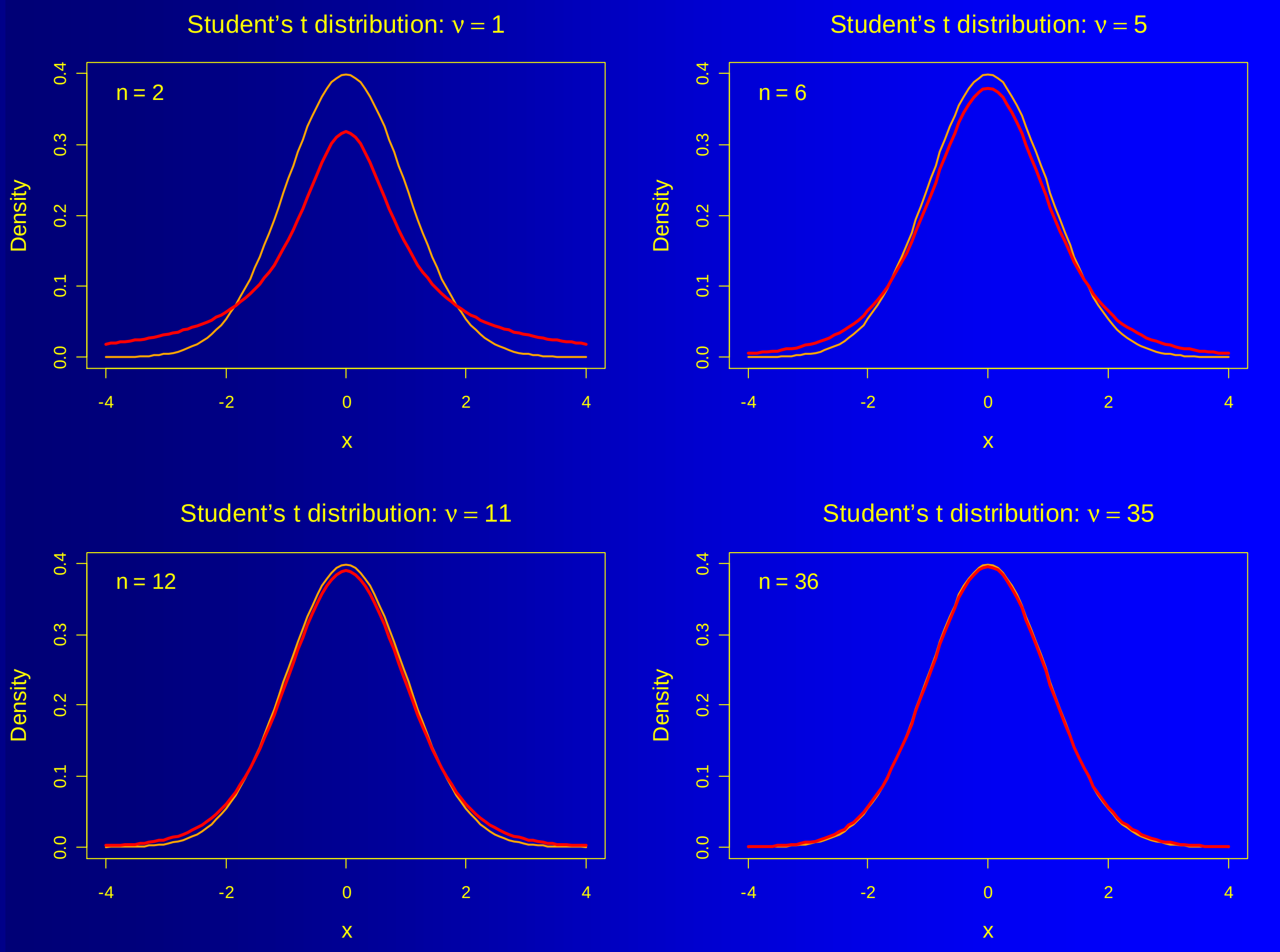
- Depends on one parameter, the 'degrees of freedom ν'. In the most simple case df = n – 1.

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

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

- The *t* Distribution is 'heavy tailed' compared to the normal distribution. Small sample sizes are penalized.
- Approaches quickly the normal distribution for df >≈30.
- Allows calculation of a CI of the sample mean based on the sample standard deviation s.

Student's *t* Distribution



Confidence Interval III

- Example from previous slides: μ 100, σ 20
  Sample sizes 36, $z_{0.05}$ 1.960, $t_{36\text{-}1,0.05}$ 2.030

Samples' mean	Samples' stand. dev.	Confidence Intervals based on z		Confidence Intervals based on t	
101.0	15.94	94.47	107.5	95.61	106.4
100.1	20.41	93.57	106.6	93.19	107.0
100.2	19.61	93.67	106.7	93.57	106.8
96.69	19.31	90.16	103.2	90.16	103.2
102.5	19.31	95.97	109.0	95.97	109.0

$$[\mu] = \bar{x} \pm z\frac{\sigma}{\sqrt{n}} \qquad [\mu] = \bar{x} \pm t\frac{s}{\sqrt{n}}$$

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

Location parameters

- Harmonic mean: 101.9516

$$\bar{x}_{harm} = \frac{n}{\sum_{i=1}^{i=n} \frac{1}{x_i}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} \cdots \frac{1}{x_i}}$$

- Geometric mean: 104.2814

$$\bar{x}_{geom} = \sqrt[n]{\prod_{i=1}^{i=n} x_i} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n} = e^{\frac{1}{n}\sum_{i=1}^{i=n} \ln x_i}$$

- Arithmetic mean: 106.7

$$\bar{x}_{arithm} = \frac{1}{n}\sum_{i=1}^{i=n} x_i = \frac{x_1 + x_2 \cdots x_i}{n}$$

A note on the harmonic mean

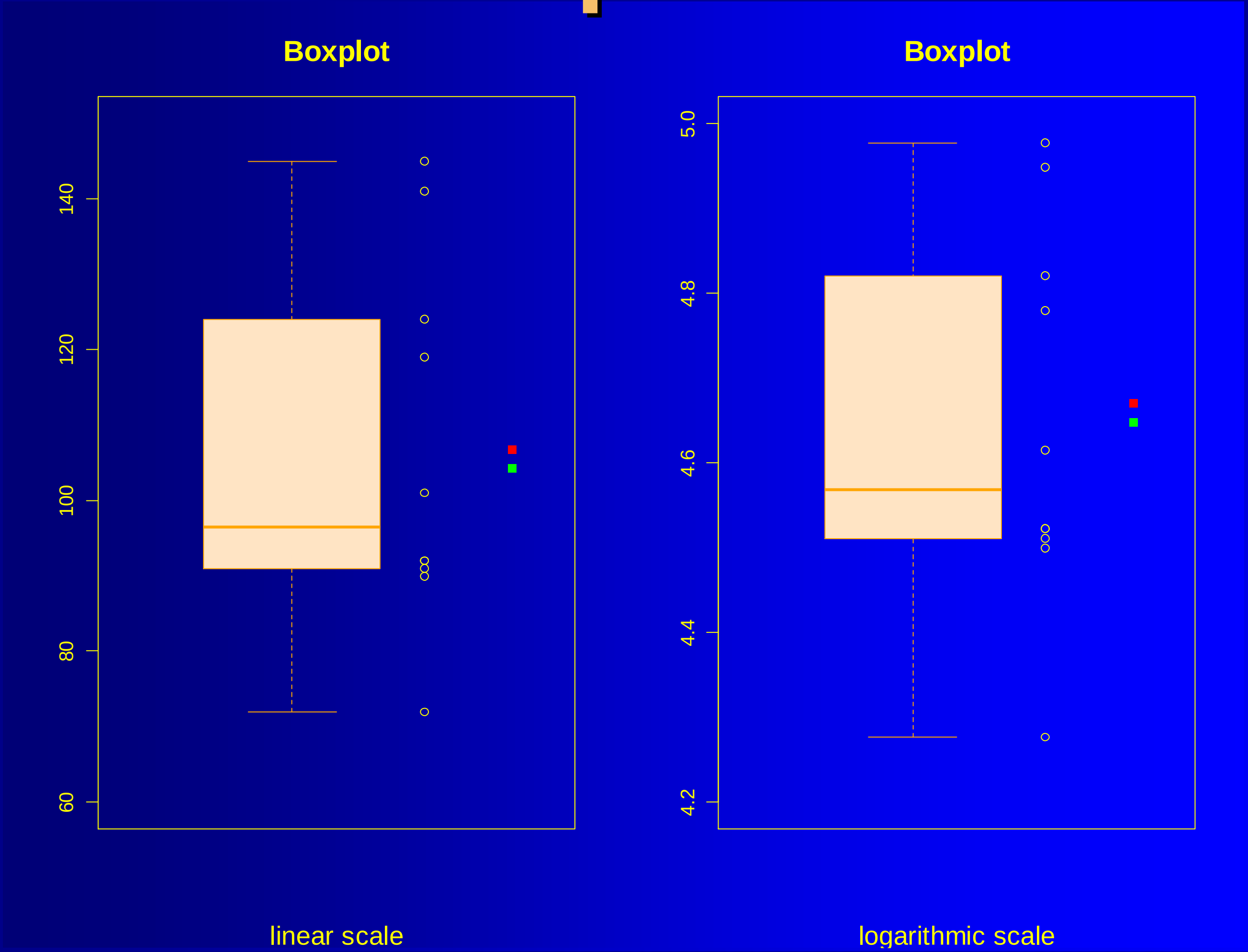
- Driving at 100 km/h from A to B; distance is 100 km.
- Driving back at 50 km/h.
- What is the average speed for the round-trip?
  - ❑ 75 km/h ❑ 70.71 km/h ☑ 66.67 km/h
- 1 h for 100 km (A→B) and 2 h for 100 km (A←B); 200 km/3 h = 66.67 km/h.
- Harmonic mean for rates!

$$\bar{x}_{harm} = \frac{2}{\frac{1}{100}+\frac{1}{50}} = \frac{2}{0.01+0.02} = 66.\dot{6}$$

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



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

$$SD_{arithm} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{i=n}\left(x_i - \overline{x}\right)^2}$$

- SD of geometric mean: 23.8000

$$SD_{geom} = e^{\sqrt{\frac{1}{n-1}\sum_{i=1}^{i=n}\left(\ln x_i - \ln \overline{x}_{geom}\right)^2}}$$

Dispersion parameters

- SD of harmonic mean: 22.8519

$$SD_{harm} = \sqrt{(n-1)\sum_{i=1}^{i=n}\left(\bar{H}_i - \bar{H}\right)^2}$$

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

$$\bar{H}_i = \frac{n-1}{\left(\sum_{j=1}^{i=n}\frac{1}{x_j}\right) - \frac{1}{x_i}}$$

Dispersion parameters

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

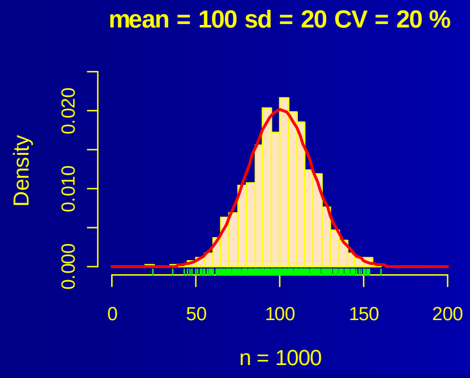
$$CV\% = 100 \cdot SD/\bar{x}$$

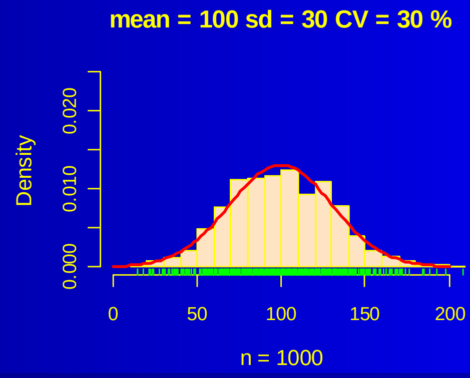
Population (N=10^6), parameters		Sample (n=36), parameters			
μ	100.00	$\bar{x}_{arithm}$	106.70	$\bar{x}_{geom}$	104.28
σ	20.00	SD_{arithm}	24.24	SD_{geom}	23.80
$CV\%$	20.00	$CV\%$	22.72	$CV\%$	22.83

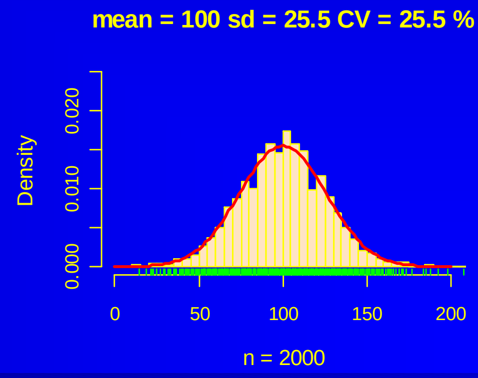
A remark on Variances

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

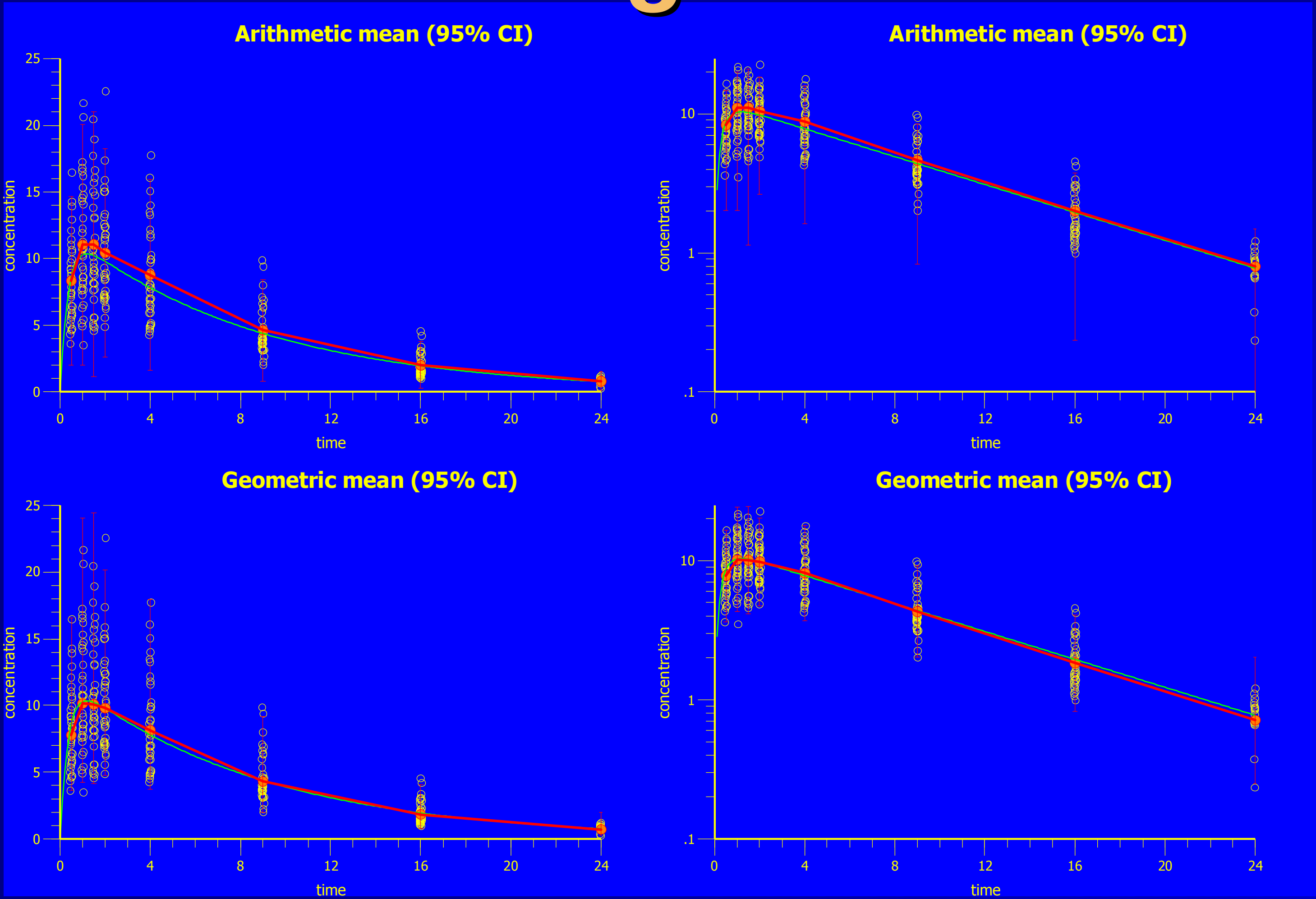
sample	mean	s	s^2	
1	100	20	400	
2	100	30	900	
1+2	(100+100)/2	25??	$\sum s^2/2$	$\sqrt{650}$=25.5!!



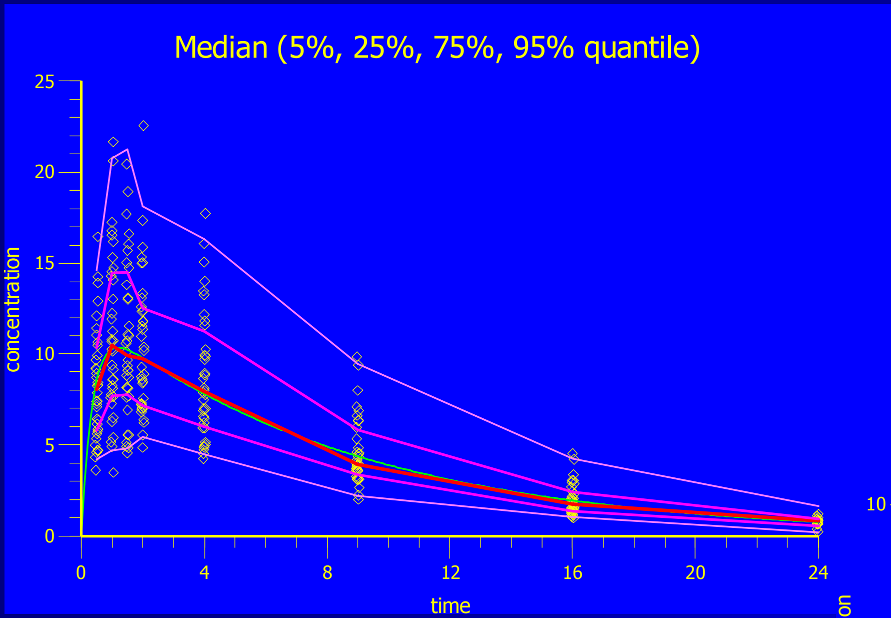


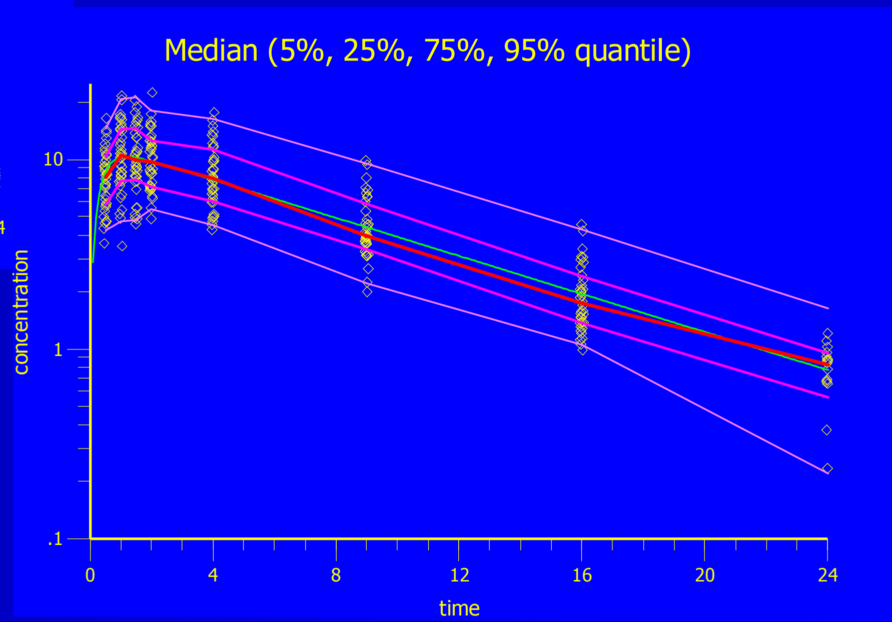


Arithm. *vs.* geom. means



Median and quantiles

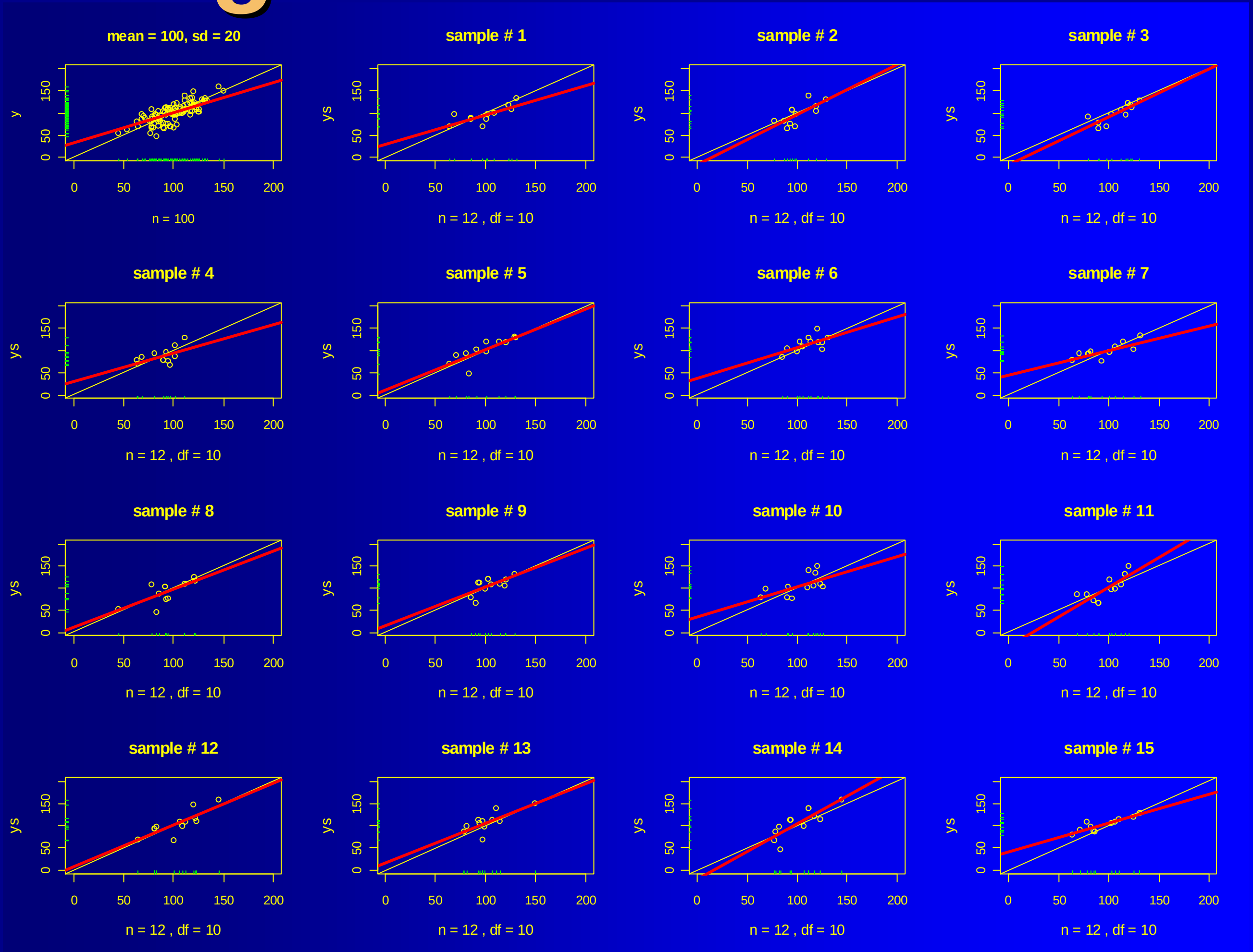




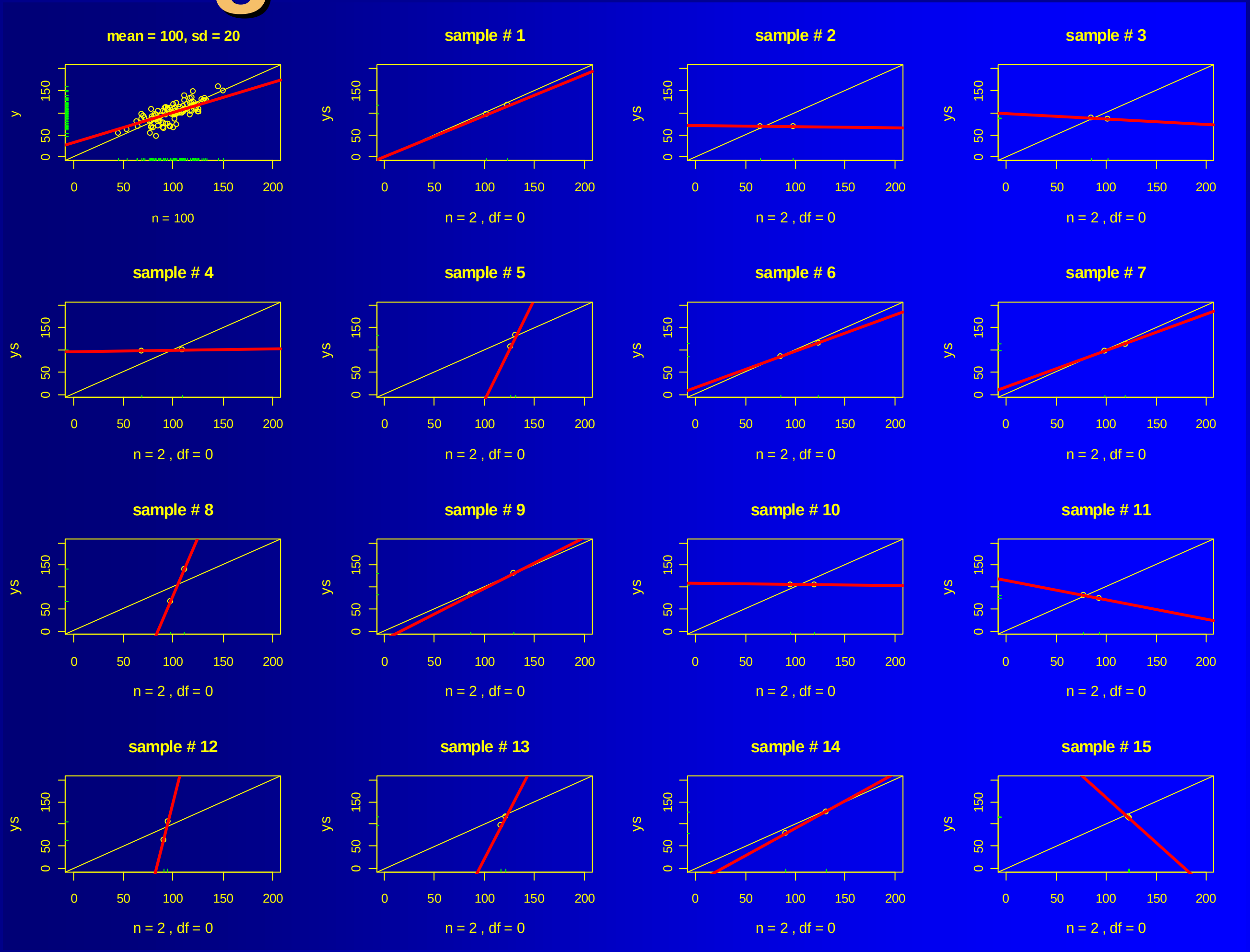
Degrees of freedom…

- 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…



Degrees of freedom…

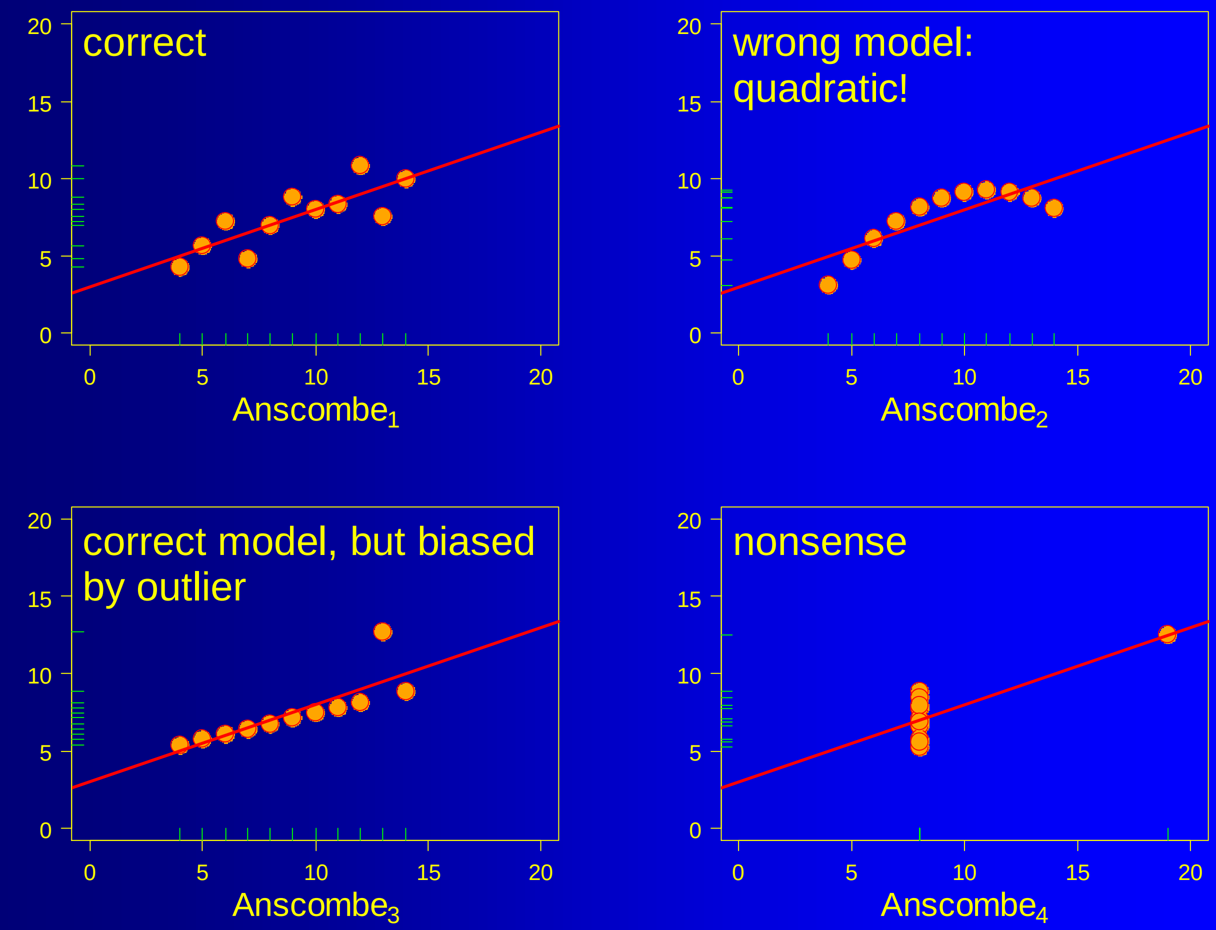


Visualize your data!

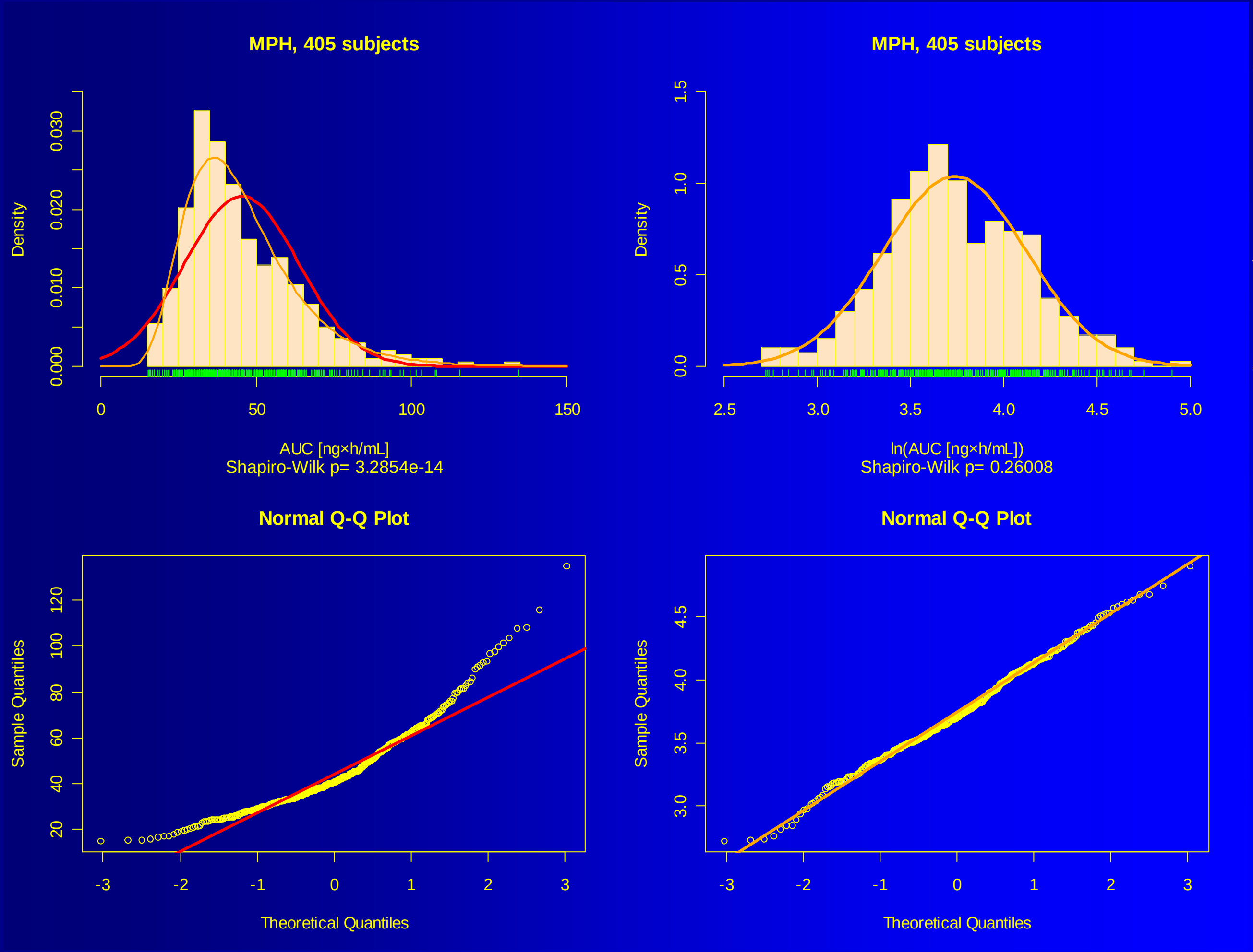
Anscombe's Quartet (1973)

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.

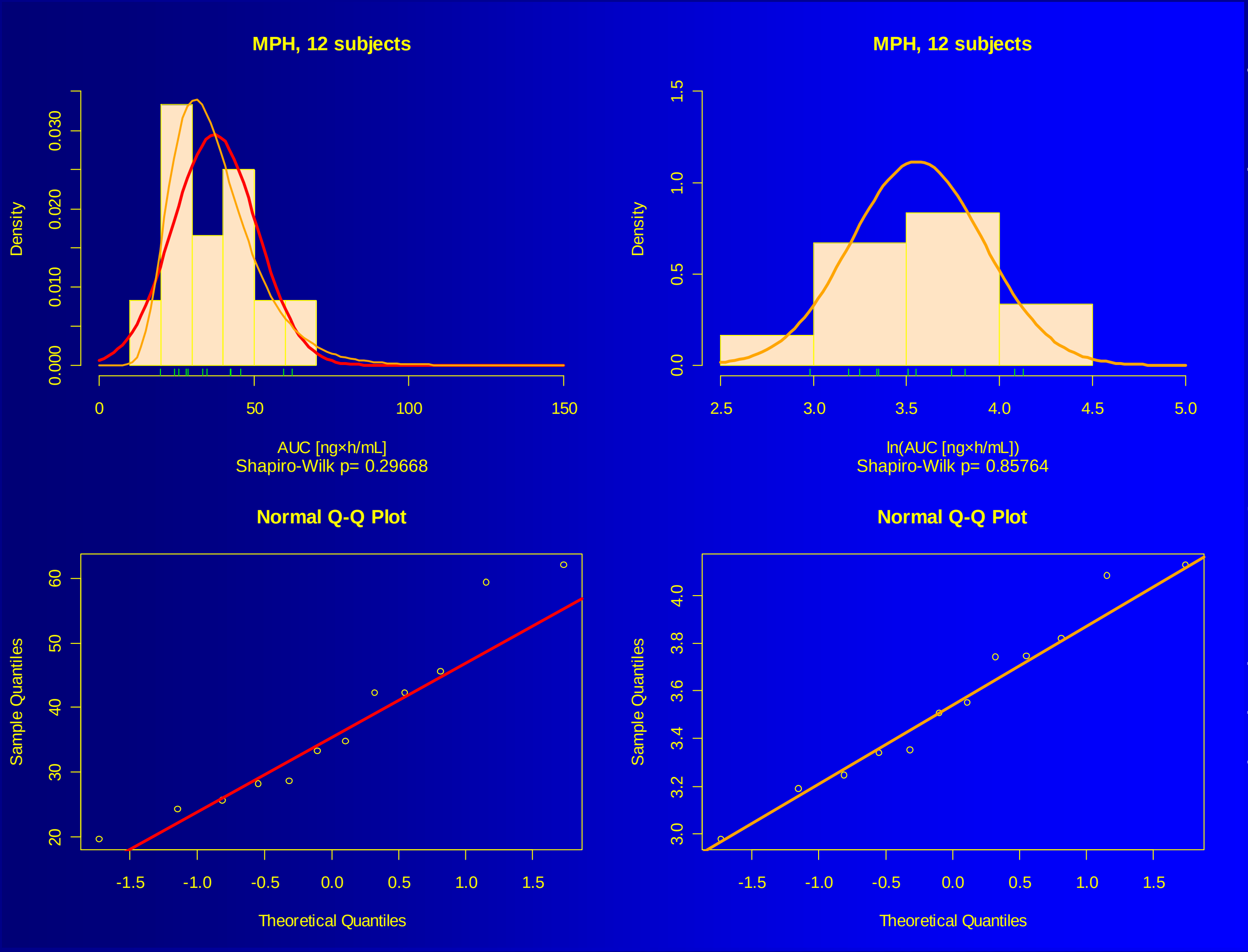


Data Transformation?



Clearly in favor of a lognormal distribution. Shapiro-Wilk test highly significant for normal distribution (rejected).

Data Transformation!



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_{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

$$F_T = \frac{AUC_T \cdot \cancel{CL_T}}{\cancel{D_T}}, F_R = \frac{AUC_R \cdot \cancel{CL_R}}{\cancel{D_R}}$$

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

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 symetrical* 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.

    Symetrical around 0 in the log-domain and around 100% in the backtransformed domain (e^0=1).

Data Transformation

- 'Problems' with logtransformation
  - Discussion, whether more bioinequivalent formula-tions 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 ν) of two distributions. We will need that in ANOVA.

$$F(x\,|\,\nu_1,\nu_2)=\begin{cases}\nu_1^{\frac{\nu_1}{2}}\nu_2^{\frac{\nu_2}{2}}\cdot\dfrac{\Gamma\left(\frac{\nu_1}{2}+\frac{\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_1}{2}\right)}\cdot\dfrac{x^{\frac{\nu_1}{2}-1}}{\left(\nu_1x+\nu_2\right)^{\frac{\nu_1+\nu_2}{2}}} & x\geq 0\\ 0 & x<0\end{cases}$$

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

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

$$F(\nu_1=1,\nu_2=\nu)=\left(t(\nu)\right)^2$$

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) *vs.*
  - 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:

Verdict	Defendant innocent	Defendant guilty
Presumption of innocence not accepted (guilty)	**Error type I**	**Correct**
Presumption of innocence accepted (not guilty)	**Correct**	**Error type II**

α- vs. β-Error

- … in more statistical terms:

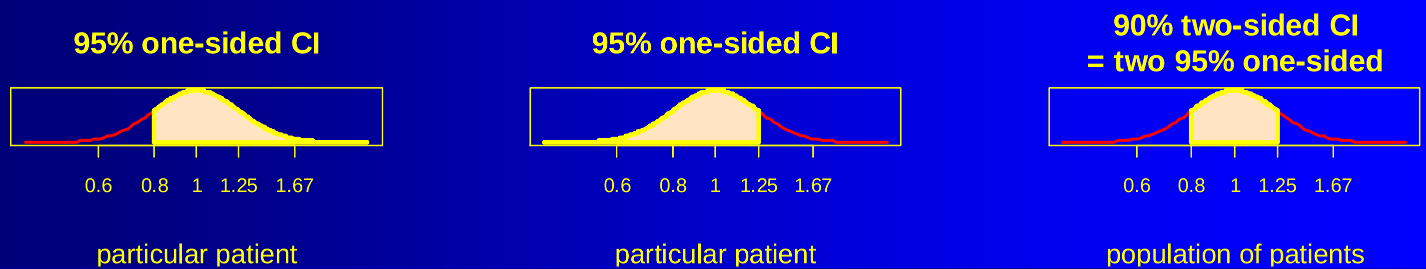
Decision	Null hypothesis true	Null hypothesis false
Null hypothesis rejected	**Error type I**	**Correct (H_a)**
Failed to reject null hypothesis	**Correct (H_0)**	**Error type II**

- In BE-testing the null hypothesis is bio<u>in</u>equivalence ($\mu_1 \neq \mu_2$)!

Decision	Null hypothesis true	Null hypothesis false
Null hypothesis rejected	**Patients' risk**	**Correct (BE)**
Failed to reject null hypothesis	**Correct (not BE)**	**Producer's risk**

α- vs. β-Error

- α-Error: Patients' Risk to be treated with a bioinequivalent formulation (H_0 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

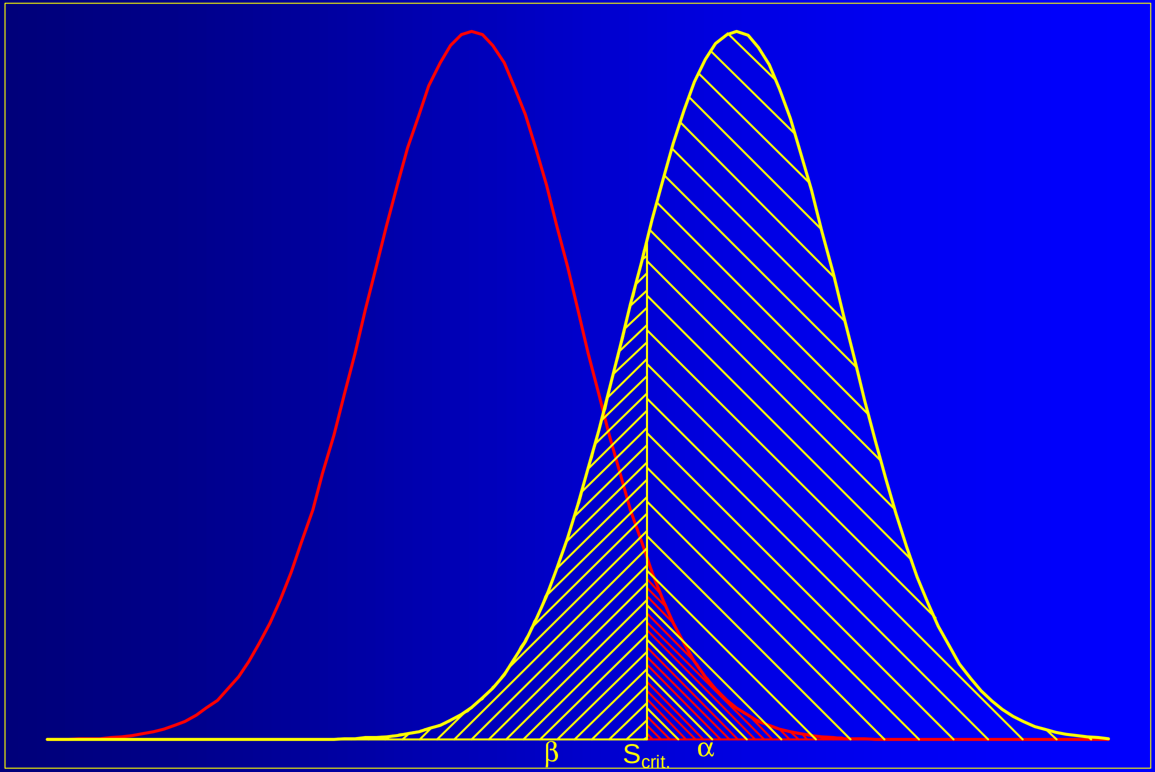
- β-Error: Producer's Risk to get no approval for a bioequivalent formulation (H_0 falsely not rejected)
  - *Set* in study planning to ≤0.2, where power = 1 – β = ≥80%
  - If power is set to 80 %

One out of five studies will fail just by chance!

α 0.05	BE
not BE	β 0.20

Significance test (*α*- vs. *β*)

Significance test: α, β



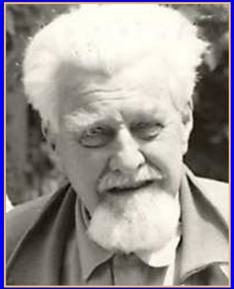
Part I: Basic Concepts

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