Statistics: Changes since I was an undergrad

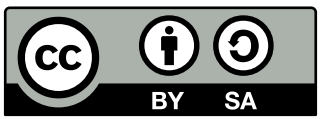
Part 2: Advances in theory and methods

Pedro J. Aphalo

Department of Biosciences, University of Helsinki

18 November 2014

Department of Biosciences, University of Helsinki, Finland.
http://blogs.helsinki.fi/aphalo/



Lectures in this series

Changes in statistics: 1975–2014

This lecture is the second in a series of four related lectures.

1. Computing capacity has dramatically increased and programming languages have evolved.
2. New statistical methods have been developed.
3. Statistical software is dramatically more powerful and varied.
4. Accountability has become a requirement in scientific research.

Outline

Statistics

- Statistics is **not** an exact science.
- Different methods make different assumptions about the data.
- Different methods give different results ⇒ conclusions may depend on method.
- Selecting the method used is a key step in any analysis and assumptions play an important role when we choose a method.
- The number of methods to choose from has increased enormously in the last 35 years. (I started my university studies in 1975).
- This talk has two objectives: to familiarize you with general families of methods that have become usable on personal computers during this time.
- To highlight that some things that we frequently do, are just workarounds or kludges that are no longer needed.

Statistics

- Statistics is **not** an exact science.
- Different methods make different assumptions about the data.
- Different methods give different results → conclusions may depend on method.
- Selecting the method used is a key step in any analysis and assumptions play an important role when we choose a method.
- The number of methods to choose from has increased enormously in the last 35 years. (I started my university studies in 1975).
- This talk has two objectives: to familiarize you with general families of methods that have become usable on personal computers during this time.
- To highlight that some things that we frequently do, are just workarounds or kludges that are no longer needed.

Statistics

- Statistics is **not** an exact science.
- Different methods make different assumptions about the data.
- Different methods give different results → conclusions may depend on method.
- Selecting the method used is a key step in any analysis and assumptions play an important role when we choose a method.
- The number of methods to choose from has increased enormously in the last 35 years. (I started my university studies in 1975).
- This talk has two objectives: to familiarize you with general families of methods that have become usable on personal computers during this time.
- To highlight that some things that we frequently do, are just workarounds or kludges that are no longer needed.

Statistics

- Statistics is **not** an exact science.
- Different methods make different assumptions about the data.
- Different methods give different results → conclusions may depend on method.
- Selecting the method used is a key step in any analysis and assumptions play an important role when we choose a method.
- The number of methods to choose from has increased enormously in the last 35 years. (I started my university studies in 1975).
- This talk has two objectives: to familiarize you with general families of methods that have become usable on personal computers during this time.
- To highlight that some things that we frequently do, are just workarounds or kludges that are no longer needed.

Statistics

- Statistics is **not** an exact science.
- Different methods make different assumptions about the data.
- Different methods give different results → conclusions may depend on method.
- Selecting the method used is a key step in any analysis, and assumptions play an important role when we choose a method.
- The number of methods to choose from has increased enormously in the last 35 years. (I started my university studies in 1975).
- This talk has two objectives: to familiarize you with general families of methods that have become usable on personal computers during this time.
- To highlight that some things that we frequently do, are just workarounds or kludges that are no longer needed.

Statistics

- Statistics is **not** an exact science.
- Different methods make different assumptions about the data.
- Different methods give different results → conclusions may depend on method.
- Selecting the method used is a key step in any analysis, and assumptions play an important role when we choose a method.
- The number of methods to choose from has increased enormously in the last 35 years. (I started my university studies in 1975).
- This talk has two objectives: to familiarize you with general families of methods that have become usable on personal computers during this time.
- To highlight that some things that we frequently do, are just workarounds or kludges that are no longer needed.

Statistics

- Statistics is **not** an exact science.
- Different methods make different assumptions about the data.
- Different methods give different results → conclusions may depend on method.
- Selecting the method used is a key step in any analysis, and assumptions play an important role when we choose a method.
- The number of methods to choose from has increased enormously in the last 35 years. (I started my university studies in 1975).
- This talk has two objectives: to familiarize you with general families of methods that have become usable on personal computers during this time.
- To highlight that some things that we frequently do, are just workarounds or kludges that are no longer needed.

Statistics

- Statistics is **not** an exact science.
- Different methods make different assumptions about the data.
- Different methods give different results → conclusions may depend on method.
- Selecting the method used is a key step in any analysis, and assumptions play an important role when we choose a method.
- The number of methods to choose from has increased enormously in the last 35 years. (I started my university studies in 1975).
- This talk has two objectives: to familiarize you with general families of methods that have become usable on personal computers during this time.
- To highlight that some things that we frequently do, are just workarounds or kludges that are no longer needed.

Statistics

- Statistics is **not** an exact science.
- Different methods make different assumptions about the data.
- Different methods give different results → conclusions may depend on method.
- Selecting the method used is a key step in any analysis, and assumptions play an important role when we choose a method.
- The number of methods to choose from has increased enormously in the last 35 years. (I started my university studies in 1975).
- This talk has two objectives: to familiarize you with general families of methods that have become usable on personal computers during this time.
- To highlight that some things that we frequently do, are just workarounds or kludges that are no longer needed.

P-values

- Where did the risk levels (α) 10%, 5%, 1%, and 0.1% originate?
- Probabilities from the Normal and other distributions are not easy to calculate by hand. It became easy and fast to do so on the desktop around 1980s.
- Before 1990's we almost always reported P-values based on where the calculated t-value or F-value fell on a printed table. Interpolating to approximate the actual P-value was time consuming, and most researchers did not do it.

P-values

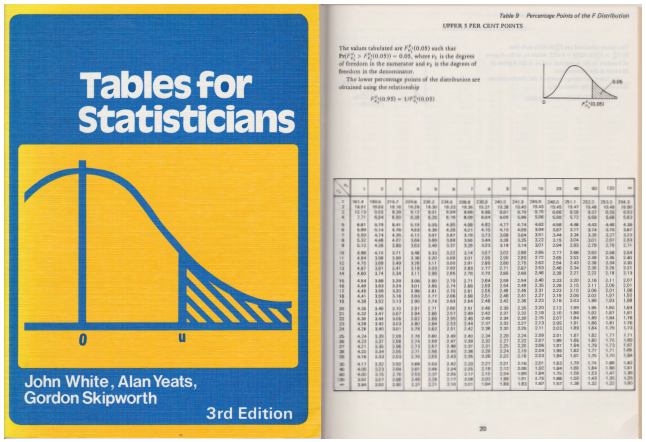
- Where did the risk levels (α) 10%, 5%, 1%, and 0.1% originate?
- Probabilities from the Normal and other distributions are not easy to calculate by hand. It became easy and fast to do so on the desktop around 1980s.
- Before 1990's we almost always reported P-values based on where the calculated t-value or F-value fell on a printed table. Interpolating to approximate the actual P-value was time consuming, and most researchers did not do it.

P-values

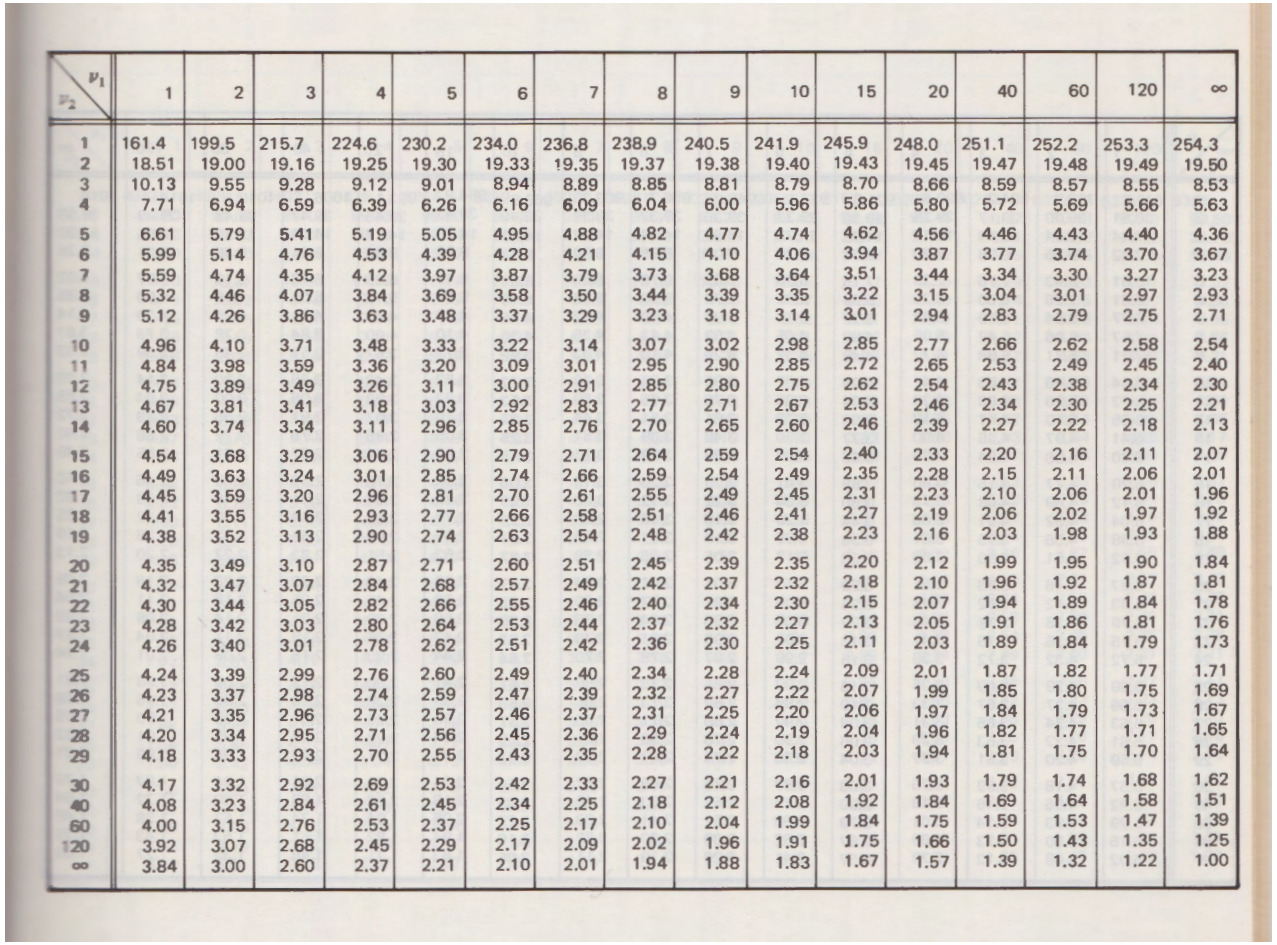
- Where did the risk levels (α) 10%, 5%, 1%, and 0.1% originate?
- Probabilities from the Normal and other distributions are not easy to calculate by hand. It became easy and fast to do so on the desktop around 1980s.
- Before 1990's we almost always reported P-values based on where the calculated *t*-value or *F*-value fell on a printed table. Interpolating to approximate the actual P-value was time consuming, and most researchers did not do it.

P-values

- Where did the risk levels (α) 10%, 5%, 1%, and 0.1% originate?
- Probabilities from the Normal and other distributions are not easy to calculate by hand. It became easy and fast to do so on the desktop around 1980s.
- Before 1990's we almost always reported P-values based on where the calculated *t*-value or *F*-value fell on a printed table. Interpolating to approximate the actual P-value was time consuming, and most researchers did not do it.



Tabulated values for $F_{\nu_2}^{\nu_1}(0.05)$



ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	15	20	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	245.9	248.0	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.45	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.28	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31	2.23	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.19	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.16	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.18	2.10	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15	2.07	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.13	2.05	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.11	2.03	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.09	2.01	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.07	1.99	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.06	1.97	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.04	1.96	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.03	1.94	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.93	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92	1.84	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.84	1.75	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.75	1.66	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.67	1.57	1.39	1.32	1.22	1.00

Risk levels

Problem

Reporting $P < 0.05$ for both $P = 0.049$ and $P = 0.011$ discards important information.

Problem

Reporting an effect as significant when $P = 0.048$ and not significant when $P = 0.051$ without further details is misleading.

Solution

Always report actual P-values.

Risk levels

Problem

Reporting $P < 0.05$ for both $P = 0.049$ and $P = 0.011$ discards important information.

Problem

Reporting an effect as significant when $P = 0.048$ and not significant when $P = 0.051$ without further details is misleading.

Solution

Always report actual P-values.

Risk levels

Problem

Reporting $P < 0.05$ for both $P = 0.049$ and $P = 0.011$ discards important information.

Problem

Reporting an effect as significant when $P = 0.048$ and not significant when $P = 0.051$ without further details is misleading.

Solution

Always report actual P-values.

Adjusting P-values

Bonferroni's method

Bonferroni's correction is very simple, and easy to calculate

$$\alpha_{\text{adj}} = \frac{\alpha}{n}$$

where α is the 'risk level' and n the number of contrasts.

Problem

But is extremely conservative! $\Rightarrow$ tests have low power (too low sensitivity to real treatment effects).

Solution

Do not use Bonferroni's method.

Adjusting P-values

Bonferroni's method

Bonferroni's correction is very simple, and easy to calculate

$$\alpha_{\text{adj}} = \frac{\alpha}{n}$$

where α is the 'risk level' and n the number of contrasts.

Problem

But is extremely conservative! $\Rightarrow$ tests have low power (too low sensitivity to real treatment effects).

Solution

Do not use Bonferroni's method.

Adjusting P-values

Bonferroni's method

Bonferroni's correction is very simple, and easy to calculate

$$\alpha_{\text{adj}} = \frac{\alpha}{n}$$

where α is the 'risk level' and n the number of contrasts.

Problem

But is extremely conservative! $\Rightarrow$ tests have low power (too low sensitivity to real treatment effects).

Solution

Do not use Bonferroni's method.

Not adjusting P-values

Problem

LSD is very easy to calculate but ignores simultaneous testing (like trying to get a given value by rolling a die several times).

Recommendation

Never use LSD (least significant difference) for multiple comparisons.

Solution

Keep the number of tests to the minimum (decide *a priori* which ones are most interesting) and use a modern and well-behaved for adjusting P-values.

Not adjusting P-values

Problem

LSD is very easy to calculate but ignores simultaneous testing (like trying to get a given value by rolling a die several times).

Recommendation

Never use LSD (least significant difference) for multiple comparisons.

Solution

Keep the number of tests to the minimum (decide *a priori* which ones are most interesting) and use a modern and well-behaved for adjusting P-values.

Not adjusting P-values

Problem

LSD is very easy to calculate but ignores simultaneous testing (like trying to get a given value by rolling a die several times).

Recommendation

Never use LSD (least significant difference) for multiple comparisons.

Solution

Keep the number of tests to the minimum (decide *a priori* which ones are most interesting) and use a modern and well-behaved for adjusting P-values.

Types of "fitted" functions

- Functions linear in the parameters: e.g. polynomials.

$$y = a_0 + a_1 \cdot x + a_2 \cdot x^2 + a_3 \cdot x^3$$

  Exact analytical methods, but with strict assumptions.
- Functions not linear in the parameters: e.g. exponential.

$$y = a_0 + \mathrm{e}^{a_1 \cdot x}$$

  Approximate iterative numerical methods.
- Splines (piece-wise functions): e.g. cubic splines.
  Approximate iterative numerical methods.

Models linear in the parameters

- Using iterative methods was almost impossible early on,
  1) because of the time needed for repeated calculation, and
  2) when using slide rules and the like, accumulation or errors.
- Using analytical solutions to least squares fits of linear models could be calculated already more than 200 years ago, if strict assumptions on data properties where adopted.
- When the assumptions did not hold, we had to modify the data using **transformations**.
- Nowadays we can easily fit non-linear and linear mixed models, and also additive models, assuming different distributions, or even not assuming any specific shape for the distribution. Many other assumptions can also be relaxed, even if we do calculations on a desktop or laptop computer.

Models linear in the parameters

- Using iterative methods was almost impossible early on, 1) because of the time needed for repeated calculation, and 2) when using slide rules and the like, accumulation or errors.
- Using analytical solutions to least squares fits of linear models could be calculated already more than 200 years ago, if strict assumptions on data properties where adopted.
- When the assumptions did not hold, we had to modify the data using **transformations**.
- Nowadays we can easily fit non-linear and linear mixed models, and also additive models, assuming different distributions, or even not assuming any specific shape for the distribution. Many other assumptions can also be relaxed, even if we do calculations on a desktop or laptop computer.

Models linear in the parameters

- Using iterative methods was almost impossible early on, 1) because of the time needed for repeated calculation, and 2) when using slide rules and the like, accumulation or errors.
- Using analytical solutions to least squares fits of linear models could be calculated already more than 200 years ago, if strict assumptions on data properties where adopted.
- When the assumptions did not hold, we had to modify the data using **transformations**.
- Nowadays we can easily fit non-linear and linear mixed models, and also additive models, assuming different distributions, or even not assuming any specific shape for the distribution. Many other assumptions can also be relaxed, even if we do calculations on a desktop or laptop computer.

Models linear in the parameters

- Using iterative methods was almost impossible early on, 1) because of the time needed for repeated calculation, and 2) when using slide rules and the like, accumulation or errors.
- Using analytical solutions to least squares fits of linear models could be calculated already more than 200 years ago, if strict assumptions on data properties where adopted.
- When the assumptions did not hold, we had to modify the data using **transformations**.
- Nowadays we can easily fit non-linear and linear mixed models, and also additive models, assuming different distributions, or even not assuming any specific shape for the distribution. Many other assumptions can also be relaxed, even if we do calculations on a desktop or laptop computer.

When to use data transformations

Problem

In a factorial ANOVA the use of transformations changes the interpretation of interactions and additivity of effects. H_0 for untransformed data: additive effects, but for log transformed data: multiplicative effects on original scale.

Solution

Decide use of transformations based on what is best for interpreting the data. Avoid use of transformations for other reasons, use instead a statistical model whose assumptions match the properties of the data.

When to use data transformations

Problem

In a factorial ANOVA the use of transformations changes the interpretation of interactions and additivity of effects. H_0 for untransformed data: additive effects, but for log transformed data: multiplicative effects on original scale.

Solution

Decide use of transformations based on what is best for interpreting the data. Avoid use of transformations for other reasons, use instead a statistical model whose assumptions match the properties of the data.

Choosing models: LM

- Linear models should be familiar to all of you.
- They include linear and polynomial regression, ANOVA and ANCOVA.
- They can be fitted using least squares as the optimization criterium using an exact analytical method.
- Usual assumptions are: Normal distribution of residuals, and homogeneity of variance among treatments.
- Calculations are easy to do and only in rather extreme cases one needs to worry about loss of numerical precision. Estimation of degrees of freedom is easy and exact if assumptions hold.

Choosing models: LM

- Linear models should be familiar to all of you.
- They include linear and polynomial regression, ANOVA and ANCOVA.
- They can be fitted using least squares as the optimization criterium using an exact analytical method.
- Usual assumptions are: Normal distribution of residuals, and homogeneity of variance among treatments.
- Calculations are easy to do and only in rather extreme cases one needs to worry about loss of numerical precision. Estimation of degrees of freedom is easy and exact if assumptions hold.

Choosing models: LM

- Linear models should be familiar to all of you.
- They include linear and polynomial regression, ANOVA and ANCOVA.
- They can be fitted using least squares as the optimization criterium using an exact analytical method.
- Usual assumptions are: Normal distribution of residuals, and homogeneity of variance among treatments.
- Calculations are easy to do and only in rather extreme cases one needs to worry about loss of numerical precision. Estimation of degrees of freedom is easy and exact if assumptions hold.

Choosing models: LM

- Linear models should be familiar to all of you.
- They include linear and polynomial regression, ANOVA and ANCOVA.
- They can be fitted using least squares as the optimization criterium using an exact analytical method.
- Usual assumptions are: Normal distribution of residuals, and homogeneity of variance among treatments.
- Calculations are easy to do and only in rather extreme cases one needs to worry about loss of numerical precision. Estimation of degrees of freedom is easy and exact if assumptions hold.

Choosing models: LM

- Linear models should be familiar to all of you.
- They include linear and polynomial regression, ANOVA and ANCOVA.
- They can be fitted using least squares as the optimization criterium using an exact analytical method.
- Usual assumptions are: Normal distribution of residuals, and homogeneity of variance among treatments.
- Calculations are easy to do and only in rather extreme cases one needs to worry about loss of numerical precision. Estimation of degrees of freedom is easy and exact if assumptions hold.

Choosing models: GLM

- Generalized linear models lift the assumption of Normal distribution of residuals.
- We can specify the distribution that is assumed (e.g. Poisson, Binomial, etc.) and the link function (e.g. log or linear). This makes it possible to analyse count and frequency data directly, without need for transformations or special handling of zeros.
- Some assumptions remain: e.g. homogeneity of variance.
- Fitting is by approximate iterative methods, but there are rather fast and reliable methods available. We can still easily get good estimates of degrees of freedom.

Choosing models: GLM

- Generalized linear models lift the assumption of Normal distribution of residuals.
- We can specify the distribution that is assumed (e.g. Poisson, Binomial, etc.) and the link function (e.g. log or linear). This makes it possible to analyse count and frequency data directly, without need for transformations or special handling of zeros.
- Some assumptions remain: e.g. homogeneity of variance.
- Fitting is by approximate iterative methods, but there are rather fast and reliable methods available. We can still easily get good estimates of degrees of freedom.

Choosing models: GLM

- Generalized linear models lift the assumption of Normal distribution of residuals.
- We can specify the distribution that is assumed (e.g. Poisson, Binomial, etc.) and the link function (e.g. log or linear). This makes it possible to analyse count and frequency data directly, without need for transformations or special handling of zeros.
- Some assumptions remain: e.g. homogeneity of variance.
- Fitting is by approximate iterative methods, but there are rather fast and reliable methods available. We can still easily get good estimates of degrees of freedom.

Choosing models: GLM

- Generalized linear models lift the assumption of Normal distribution of residuals.
- We can specify the distribution that is assumed (e.g. Poisson, Binomial, etc.) and the link function (e.g. log or linear). This makes it possible to analyse count and frequency data directly, without need for transformations or special handling of zeros.
- Some assumptions remain: e.g. homogeneity of variance.
- Fitting is by approximate iterative methods, but there are rather fast and reliable methods available. We can still easily get good estimates of degrees of freedom.

Choosing models: GLS, LME and GLMM

- Nowadays it is relatively easy to fit LME (linear mixed effects) models, GLS (general linear) models, and somehow more difficult to fit GLMM (generalized linear mixed) models.
- In all these models the error variance can be estimated based on variance covariates, variance structures, and correlations (including auto-correlations).
- Mixed effects models allow for partitioning of the random variation into (nested) components (e.g. those corresponding to blocks and plants).
- Fitting is by approximate iterative methods, but although methods frequently converge to a good answer, one should check for non-convergence. Estimates of degrees of freedom are approximations, consequently so are P-values.
- It is usual to select the best model with criteria like AIC or BIC that penalize the goodness of fit based on the number of parameters fitted.

Choosing models: GLS, LME and GLMM

- Nowadays it is relatively easy to fit LME (linear mixed effects) models, GLS (general linear) models, and somehow more difficult to fit GLMM (generalized linear mixed) models.
- In all these models the error variance can be estimated based on variance covariates, variance structures, and correlations (including auto-correlations).
- Mixed effects models allow for partitioning of the random variation into (nested) components (e.g. those corresponding to blocks and plants).
- Fitting is by approximate iterative methods, but although methods frequently converge to a good answer, one should check for non-convergence. Estimates of degrees of freedom are approximations, consequently so are P-values.
- It is usual to select the best model with criteria like AIC or BIC that penalize the goodness of fit based on the number of parameters fitted.

Choosing models: GLS, LME and GLMM

- Nowadays it is relatively easy to fit LME (linear mixed effects) models, GLS (general linear) models, and somehow more difficult to fit GLMM (generalized linear mixed) models.
- In all these models the error variance can be estimated based on variance covariates, variance structures, and correlations (including auto-correlations).
- Mixed effects models allow for partitioning of the random variation into (nested) components (e.g. those corresponding to blocks and plants).
- Fitting is by approximate iterative methods, but although methods frequently converge to a good answer, one should check for non-convergence. Estimates of degrees of freedom are approximations, consequently so are P-values.
- It is usual to select the best model with criteria like AIC or BIC that penalize the goodness of fit based on the number of parameters fitted.

Choosing models: GLS, LME and GLMM

- Nowadays it is relatively easy to fit LME (linear mixed effects) models, GLS (general linear) models, and somehow more difficult to fit GLMM (generalized linear mixed) models.
- In all these models the error variance can be estimated based on variance covariates, variance structures, and correlations (including auto-correlations).
- Mixed effects models allow for partitioning of the random variation into (nested) components (e.g. those corresponding to blocks and plants).
- Fitting is by approximate iterative methods, but although methods frequently converge to a good answer, one should check for non-convergence. Estimates of degrees of freedom are approximations, consequently so are P-values.
- It is usual to select the best model with criteria like AIC or BIC that penalize the goodness of fit based on the number of parameters fitted.

Choosing models: GLS, LME and GLMM

- Nowadays it is relatively easy to fit LME (linear mixed effects) models, GLS (general linear) models, and somehow more difficult to fit GLMM (generalized linear mixed) models.
- In all these models the error variance can be estimated based on variance covariates, variance structures, and correlations (including auto-correlations).
- Mixed effects models allow for partitioning of the random variation into (nested) components (e.g. those corresponding to blocks and plants).
- Fitting is by approximate iterative methods, but although methods frequently converge to a good answer, one should check for non-convergence. Estimates of degrees of freedom are approximations, consequently so are P-values.
- It is usual to select the best model with criteria like AIC or BIC that penalize the goodness of fit based on the number of parameters fitted.

Choosing models: GAM and GAMLSS

- GAM can be used when the functional relationship between response and explanatory variable is unknown, and when we do not want to assume a functional form.
- In GAM the response is fitted to a spline, and also interaction between the spline and factors (e.g. treatments) can be included in the model.
- GAMLSS are extremely flexible as in addition to the mean (μ) the parameters of error distribution functions can be flexibly fit as a function of other variables (or splines based on these variables).
- Fitting is by iterative methods. When using these models one needs to fit several parameters, so they are best suited to relatively large data sets, with simple treatment/groups structure and at least one continuous covariate.

Choosing models: GAM and GAMLSS

- GAM can be used when the functional relationship between response and explanatory variable is unknown, and when we do not want to assume a functional form.
- In GAM the response is fitted to a spline, and also interaction between the spline and factors (e.g. treatments) can be included in the model.
- GAMLSS are extremely flexible as in addition to the mean (μ) the parameters of error distribution functions can be flexibly fit as a function of other variables (or splines based on these variables).
- Fitting is by iterative methods. When using these models one needs to fit several parameters, so they are best suited to relatively large data sets, with simple treatment/groups structure and at least one continuous covariate.

Choosing models: GAM and GAMLSS

- GAM can be used when the functional relationship between response and explanatory variable is unknown, and when we do not want to assume a functional form.
- In GAM the response is fitted to a spline, and also interaction between the spline and factors (e.g. treatments) can be included in the model.
- GAMLSS are extremely flexible as in addition to the mean (μ) the parameters of error distribution functions can be flexibly fit as a function of other variables (or splines based on these variables).
- Fitting is by iterative methods. When using these models one needs to fit several parameters, so they are best suited to relatively large data sets, with simple treatment/groups structure and at least one continuous covariate.

Choosing models: GAM and GAMLSS

- GAM can be used when the functional relationship between response and explanatory variable is unknown, and when we do not want to assume a functional form.
- In GAM the response is fitted to a spline, and also interaction between the spline and factors (e.g. treatments) can be included in the model.
- GAMLSS are extremely flexible as in addition to the mean (μ) the parameters of error distribution functions can be flexibly fit as a function of other variables (or splines based on these variables).
- Fitting is by iterative methods. When using these models one needs to fit several parameters, so they are best suited to relatively large data sets, with simple treatment/groups structure and at least one continuous covariate.

Choosing models: NLM and NLME

- Non-linear mixed models can be useful and can be thought as being in-between empirical and mechanistic models.
- They have the same kind of assumptions than LM and LME models with respect to the error variance, but allow the fitting of a non linear function.
- Instead of say fitting an exponential growth response curve to data for individual plants, we do a joint fit for all replicates and all treatments, in a model where the values of the fitted parameters are a function of the treatments.
- Fitting is by approximate iterative methods and usually one has to tweak starting values to achieve convergence. They have the advantage that as the whole data set is used, the correlations among the parameters are taken into account in the calculation of P-values.

Choosing models: NLM and NLME

- Non-linear mixed models can be useful and can be thought as being in-between empirical and mechanistic models.
- They have the same kind of assumptions than LM and LME models with respect to the error variance, but allow the fitting of a non linear function.
- Instead of say fitting an exponential growth response curve to data for individual plants, we do a joint fit for all replicates and all treatments, in a model where the values of the fitted parameters are a function of the treatments.
- Fitting is by approximate iterative methods and usually one has to tweak starting values to achieve convergence. They have the advantage that as the whole data set is used, the correlations among the parameters are taken into account in the calculation of P-values.

Choosing models: NLM and NLME

- Non-linear mixed models can be useful and can be thought as being in-between empirical and mechanistic models.
- They have the same kind of assumptions than LM and LME models with respect to the error variance, but allow the fitting of a non linear function.
- Instead of say fitting an exponential growth response curve to data for individual plants, we do a joint fit for all replicates and all treatments, in a model where the values of the fitted parameters are a function of the treatments.
- Fitting is by approximate iterative methods and usually one has to tweak starting values to achieve convergence. They have the advantage that as the whole data set is used, the correlations among the parameters are taken into account in the calculation of P-values.

Choosing models: NLM and NLME

- Non-linear mixed models can be useful and can be thought as being in-between empirical and mechanistic models.
- They have the same kind of assumptions than LM and LME models with respect to the error variance, but allow the fitting of a non linear function.
- Instead of say fitting an exponential growth response curve to data for individual plants, we do a joint fit for all replicates and all treatments, in a model where the values of the fitted parameters are a function of the treatments.
- Fitting is by approximate iterative methods and usually one has to tweak starting values to achieve convergence. They have the advantage that as the whole data set is used, the correlations among the parameters are taken into account in the calculation of P-values.

Choosing models: Mixture Models

- Mixture models are related to classification approaches.
- They are used when we have reasons to expect that the observations originate from two or more statistical populations.
- As with other models there are different degrees of flexibility: some approaches assume a mix of Normally distributed populations, and the means and variances are fitted from the data (although they can be constrained).
- Fitting is by iterative approximate methods, and confidence of the fitted values is estimated by re-sampling (bootstrapping).
- Confidence intervals calculated by re-sampling are reliable.

Choosing models: Mixture Models

- Mixture models are related to classification approaches.
- They are used when we have reasons to expect that the observations originate from two or more statistical populations.
- As with other models there are different degrees of flexibility: some approaches assume a mix of Normally distributed populations, and the means and variances are fitted from the data (although they can be constrained).
- Fitting is by iterative approximate methods, and confidence of the fitted values is estimated by re-sampling (bootstrapping).
- Confidence intervals calculated by re-sampling are reliable.

Choosing models: Mixture Models

- Mixture models are related to classification approaches.
- They are used when we have reasons to expect that the observations originate from two or more statistical populations.
- As with other models there are different degrees of flexibility: some approaches assume a mix of Normally distributed populations, and the means and variances are fitted from the data (although they can be constrained).
- Fitting is by iterative approximate methods, and confidence of the fitted values is estimated by re-sampling (bootstrapping).
- Confidence intervals calculated by re-sampling are reliable.

Choosing models: Mixture Models

- Mixture models are related to classification approaches.
- They are used when we have reasons to expect that the observations originate from two or more statistical populations.
- As with other models there are different degrees of flexibility: some approaches assume a mix of Normally distributed populations, and the means and variances are fitted from the data (although they can be constrained).
- Fitting is by iterative approximate methods, and confidence of the fitted values is estimated by re-sampling (bootstrapping).
- Confidence intervals calculated by re-sampling are reliable.

Choosing models: Mixture Models

- Mixture models are related to classification approaches.
- They are used when we have reasons to expect that the observations originate from two or more statistical populations.
- As with other models there are different degrees of flexibility: some approaches assume a mix of Normally distributed populations, and the means and variances are fitted from the data (although they can be constrained).
- Fitting is by iterative approximate methods, and confidence of the fitted values is estimated by re-sampling (bootstrapping).
- Confidence intervals calculated by re-sampling are reliable.

Robust methods

- Nowadays there are alternatives to traditional non-parametric methods.
- Bootstrapping. Re-sampling methods estimate P-values and distributions based on the data. They are computationally intensive.
- WLE (Weighted Likelihood Estimation). A distribution is assumed, and 'outliers' auto-magically ignored, totally or partially, during least squares calculations. There are implementations of LM (linear models), GLM (generalized linear models), t-test, mixture model, etc.

Robust methods

- Nowadays there are alternatives to traditional non-parametric methods.
- Bootstrapping. Re-sampling methods estimate P-values and distributions based on the data. They are computationally intensive.
- WLE (Weighted Likelihood Estimation). A distribution is assumed, and 'outliers' auto-magically ignored, totally or partially, during least squares calculations. There are implementations of LM (linear models), GLM (generalized linear models), t-test, mixture model, etc.

Robust methods

- Nowadays there are alternatives to traditional non-parametric methods.
- Bootstrapping. Re-sampling methods estimate P-values and distributions based on the data. They are computationally intensive.
- WLE (Weighted Likelihood Estimation). A distribution is assumed, and 'outliers' auto-magically ignored, totally or partially, during least squares calculations. There are implementations of LM (linear models), GLM (generalized linear models), t-test, mixture model, etc.

Choosing models

The good and the bad

The models to choose from are many. This is good, but it makes model selection a key and rather time-consuming step in the analysis.

Example 1

We are interested in RGR (relative growth rate), then fitting a NLME to log-transformed dry-weight data may be best.

Example 2

We are interested in germination percent, although it has been customary to use an arcsine data transformation an ANOVA for this type of data, fitting a GLM assuming a binomial distribution is a better approach.

Choosing models

The good and the bad

The models to choose from are many. This is good, but it makes model selection a key and rather time-consuming step in the analysis.

Example 1

We are interested in RGR (relative growth rate), then fitting a NLME to log-transformed dry-weight data may be best.

Example 2

We are interested in germination percent, although it has been customary to use an arcsine data transformation an ANOVA for this type of data, fitting a GLM assuming a binomial distribution is a better approach.

Choosing models

The good and the bad

The models to choose from are many. This is good, but it makes model selection a key and rather time-consuming step in the analysis.

Example 1

We are interested in RGR (relative growth rate), then fitting a NLME to log-transformed dry-weight data may be best.

Example 2

We are interested in germination percent, although it has been customary to use an arcsine data transformation an ANOVA for this type of data, fitting a GLM assuming a binomial distribution is a better approach.

Statistics and the research process

- Statistics is not just about tests of significance and fitting models.
- The design of an experiment must be matched to the analysis method.
- How an experiment is designed determines which methods (if any) can be used to assess the strength of the evidence.
- How an experiment is carried out, and how measurements are taken, may cause bias, or make impossible a good estimate of random variation.
- How well and in how much detail we describe the data and its statistical analysis affects the credibility of our conclusions and the long-term usefulness and impact of our work.

Statistics and the research process

- Statistics is not just about tests of significance and fitting models.
- The design of an experiment must be matched to the analysis method.
- How an experiment is designed determines which methods (if any) can be used to assess the strength of the evidence.
- How an experiment is carried out, and how measurements are taken, may cause bias, or make impossible a good estimate of random variation.
- How well and in how much detail we describe the data and its statistical analysis affects the credibility of our conclusions and the long-term usefulness and impact of our work.

Statistics and the research process

- Statistics is not just about tests of significance and fitting models.
- The design of an experiment must be matched to the analysis method.
- How an experiment is designed determines which methods (if any) can be used to assess the strength of the evidence.
- How an experiment is carried out, and how measurements are taken, may cause bias, or make impossible a good estimate of random variation.
- How well and in how much detail we describe the data and its statistical analysis affects the credibility of our conclusions and the long-term usefulness and impact of our work.

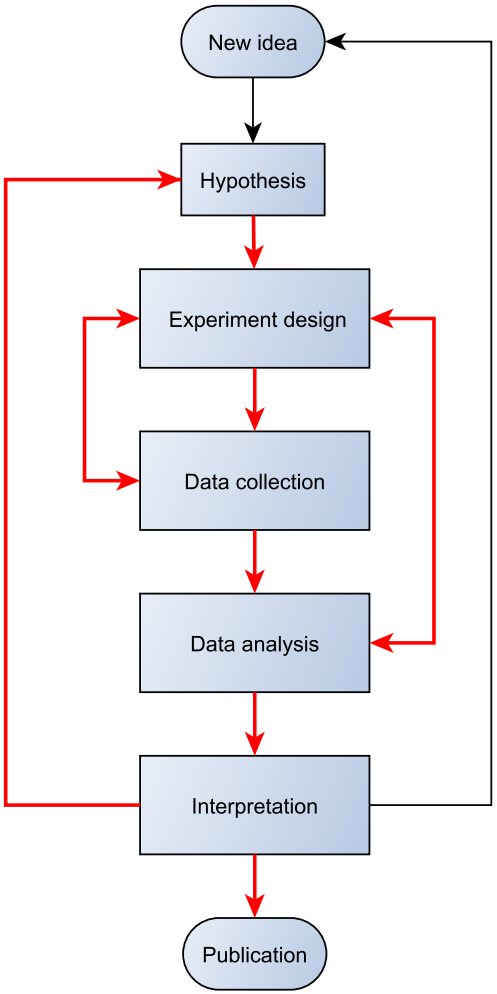
Statistics and the research process

- Statistics is not just about tests of significance and fitting models.
- The design of an experiment must be matched to the analysis method.
- How an experiment is designed determines which methods (if any) can be used to assess the strength of the evidence.
- How an experiment is carried out, and how measurements are taken, may cause bias, or make impossible a good estimate of random variation.
- How well and in how much detail we describe the data and its statistical analysis affects the credibility of our conclusions and the long-term usefulness and impact of our work.

Statistics and the research process

- Statistics is not just about tests of significance and fitting models.
- The design of an experiment must be matched to the analysis method.
- How an experiment is designed determines which methods (if any) can be used to assess the strength of the evidence.
- How an experiment is carried out, and how measurements are taken, may cause bias, or make impossible a good estimate of random variation.
- How well and in how much detail we describe the data and its statistical analysis affects the credibility of our conclusions and the long-term usefulness and impact of our work.

Flowchart: statistics and research process



Recommendations on design

- Design your experiment taking into account how the data could be analysed, and considering the different alternatives available.
- Design your experiment so that you maximize the chances of reliably testing the hypothesis of interest.
- Things to consider during the design phase include size of response that should be detectable, expected amount and type of random variation, possible confounding sources of variation, ways of minimizing the unaccounted variation, and of improving the estimates of the parameters that are most interesting.
- Also take into consideration efficiency in use of resources.

Recommendations on design

- Design your experiment taking into account how the data could be analysed, and considering the different alternatives available.
- Design your experiment so that you maximize the chances of reliably testing the hypothesis of interest.
- Things to consider during the design phase include size of response that should be detectable, expected amount and type of random variation, possible confounding sources of variation, ways of minimizing the unaccounted variation, and of improving the estimates of the parameters that are most interesting.
- Also take into consideration efficiency in use of resources.

Recommendations on design

- Design your experiment taking into account how the data could be analysed, and considering the different alternatives available.
- Design your experiment so that you maximize the chances of reliably testing the hypothesis of interest.
- Things to consider during the design phase include size of response that should be detectable, expected amount and type of random variation, possible confounding sources of variation, ways of minimizing the unaccounted variation, and of improving the estimates of the parameters that are most interesting.
- Also take into consideration efficiency in use of resources.

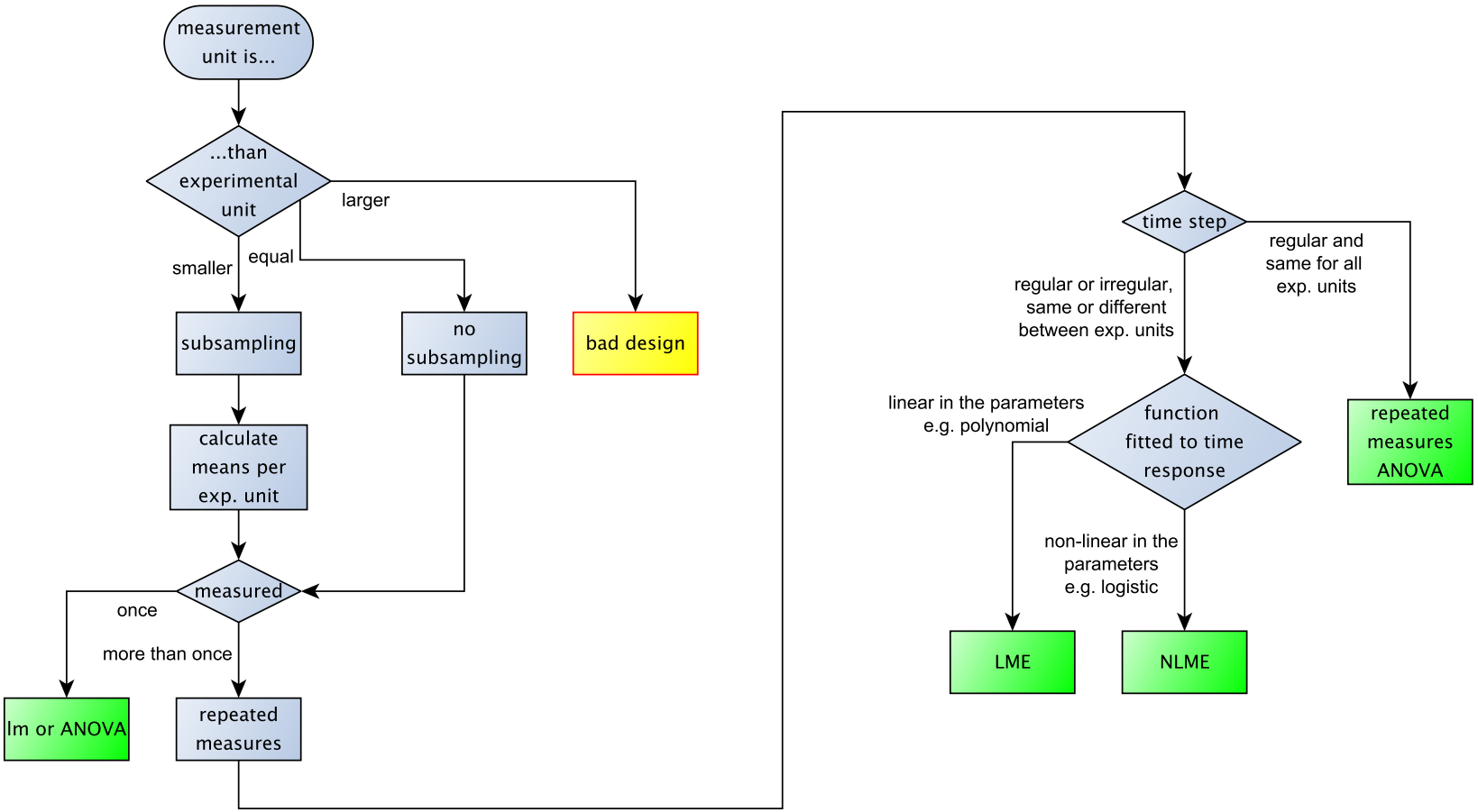
Recommendations on design

- Design your experiment taking into account how the data could be analysed, and considering the different alternatives available.
- Design your experiment so that you maximize the chances of reliably testing the hypothesis of interest.
- Things to consider during the design phase include size of response that should be detectable, expected amount and type of random variation, possible confounding sources of variation, ways of minimizing the unaccounted variation, and of improving the estimates of the parameters that are most interesting.
- Also take into consideration efficiency in use of resources.

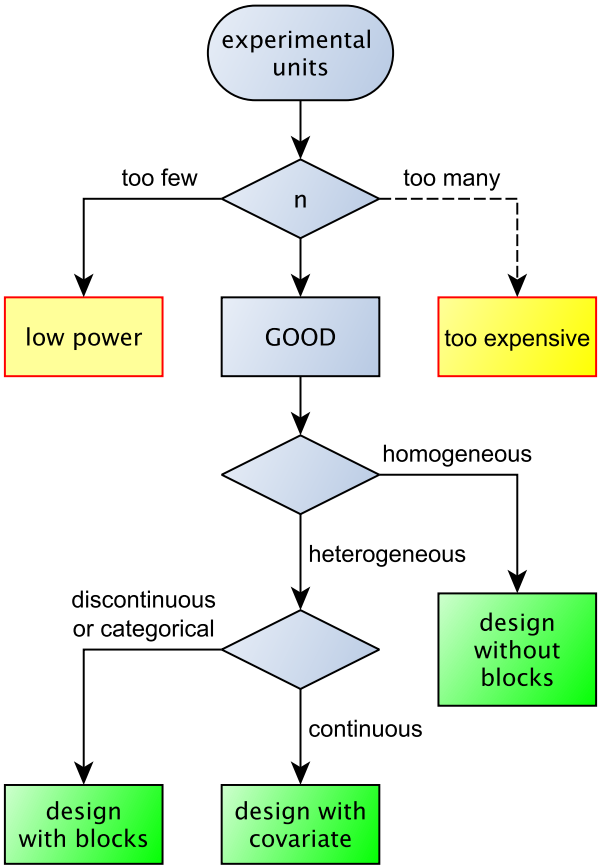
Concept: experimental vs. measurement units

- A correct statistical analysis is only possible if we correctly identify the true experimental units in our experiments.
- First step when analysing data is to make sure that the model used matches the replication used.
- An *experimental unit* is the unit or 'thing' to which the treatment is assigned (at random) (e.g. a plant in a pot).
- An experimental unit is not necessarily the unit that is measured, which can be smaller (e.g. a leaf from the treated plant ⇒ measuring several leaves separately does not increase the number of replicates).
- A measured object (= measurement unit) which is smaller than an experimental unit is called a sub-sample.

Flowchart: experimental vs. measurement units



Flowchart: replication



Recommendations on choosing models

- Think twice before using transformations of your data.
- Choose carefully the model fitted, and if you have an equation describing the hypothesis under test, then design an experiment where you can test the hypothesis by actually fitting this and competing equations.
- When designing experiments and analysing data, remind yourself of the often quoted phrase by John Tukey: 'An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question'.

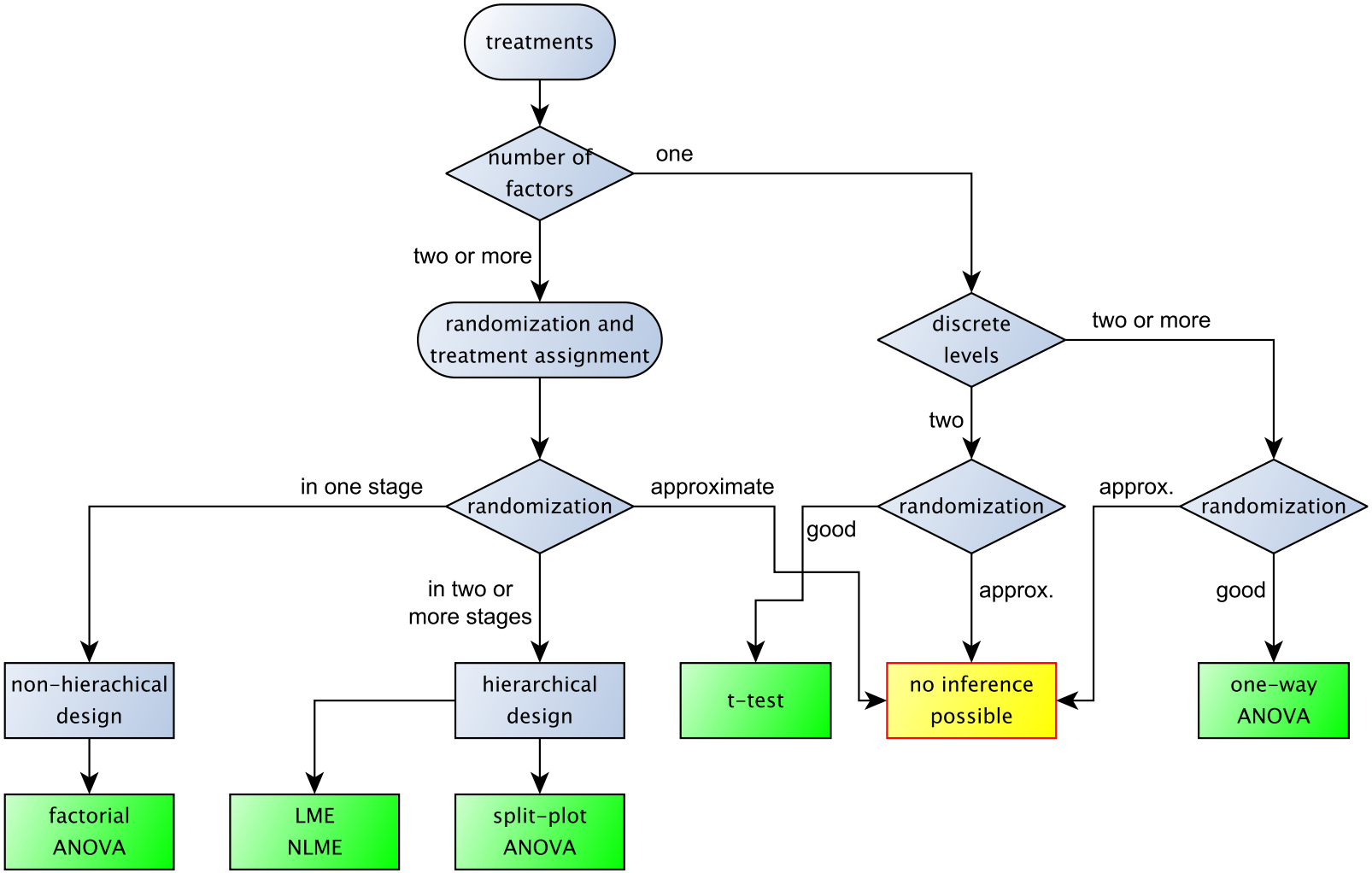
Recommendations on choosing models

- Think twice before using transformations of your data.
- Choose carefully the model fitted, and if you have an equation describing the hypothesis under test, then design an experiment where you can test the hypothesis by actually fitting this and competing equations.
- When designing experiments and analysing data, remind yourself of the often quoted phrase by John Tukey: 'An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question'.

Recommendations on choosing models

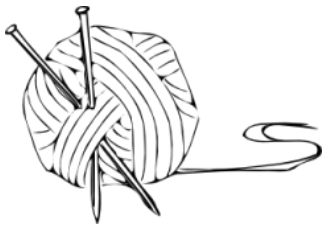
- Think twice before using transformations of your data.
- Choose carefully the model fitted, and if you have an equation describing the hypothesis under test, then design an experiment where you can test the hypothesis by actually fitting this and competing equations.
- When designing experiments and analysing data, remind yourself of the often quoted phrase by John Tukey: 'An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question'.

Flowchart: choosing the statistical model



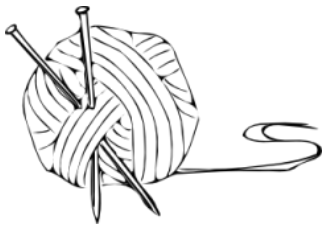
The End

- I prepared this slide presentation with LaTeX and R, on the RStudio IDE. I used Beamer and knitr, ggplot2 and other packages. This is all free open-source software, available for MS-Windows, OS-X, Linux, and Unix.
- This whole presentation (including example) is coded in a single text file (except for flowcharts, 2 figures and logos).
- To be continued. . . Part 3: Examples in R.
- Thanks for listening!



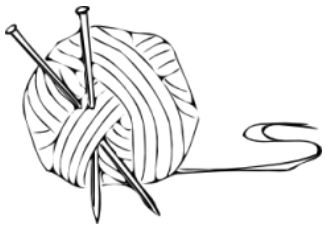
The End

- I prepared this slide presentation with LaTeX and R, on the RStudio IDE. I used Beamer and knitr, ggplot2 and other packages. This is all free open-source software, available for MS-Windows, OS-X, Linux, and Unix.
- This whole presentation (including example) is coded in a single text file (except for flowcharts, 2 figures and logos).
- To be continued. . . Part 3: Examples in R.
- Thanks for listening!



The End

- I prepared this slide presentation with LaTeX and R, on the RStudio IDE. I used Beamer and knitr, ggplot2 and other packages. This is all free open-source software, available for MS-Windows, OS-X, Linux, and Unix.
- This whole presentation (including example) is coded in a single text file (except for flowcharts, 2 figures and logos).
- To be continued. . . Part 3: Examples in R.
- Thanks for listening!



The End

- I prepared this slide presentation with LaTeX and R, on the RStudio IDE. I used Beamer and knitr, ggplot2 and other packages. This is all free open-source software, available for MS-Windows, OS-X, Linux, and Unix.
- This whole presentation (including example) is coded in a single text file (except for flowcharts, 2 figures and logos).
- To be continued. . . Part 3: Examples in R.
- Thanks for listening!