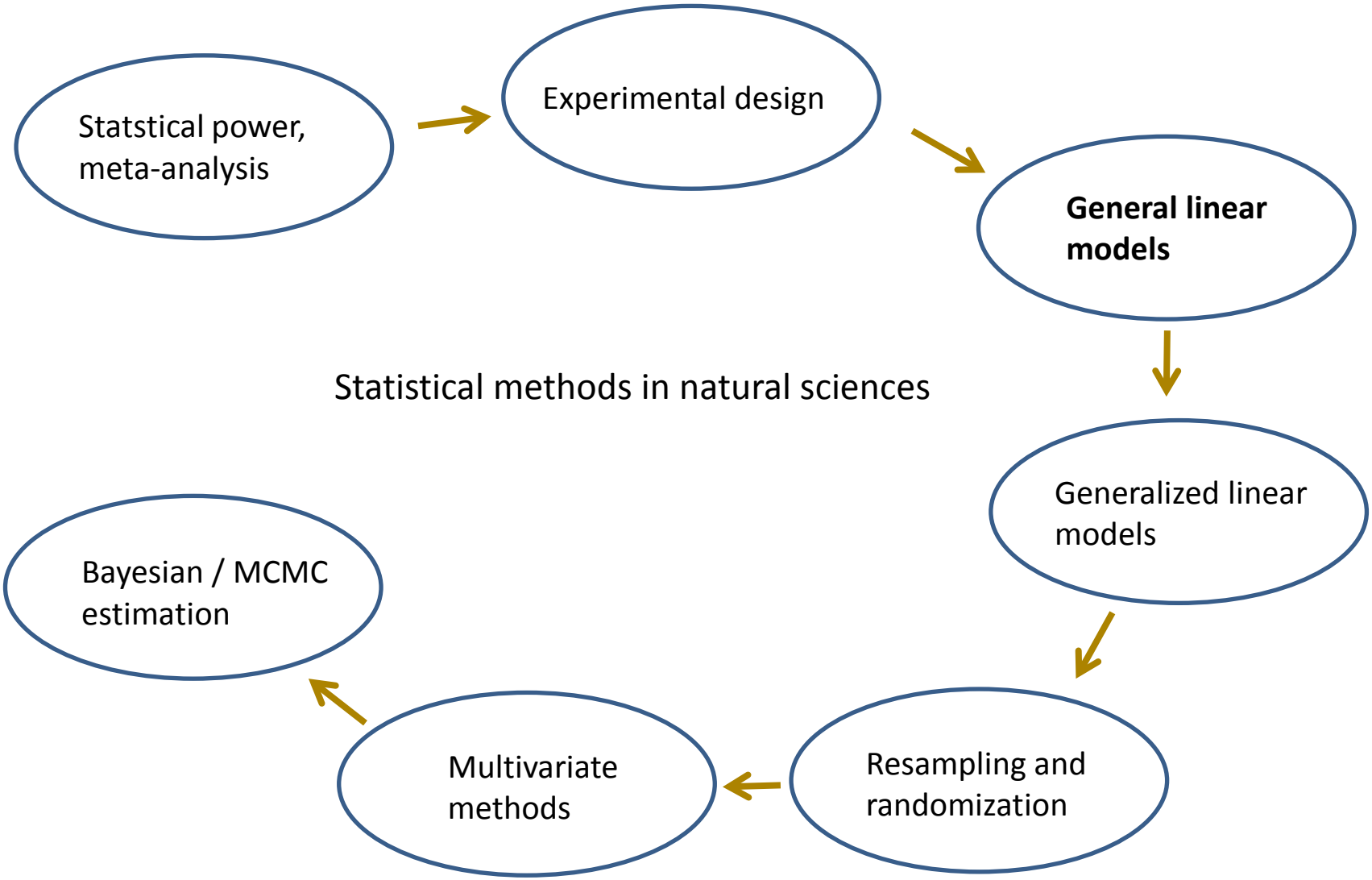
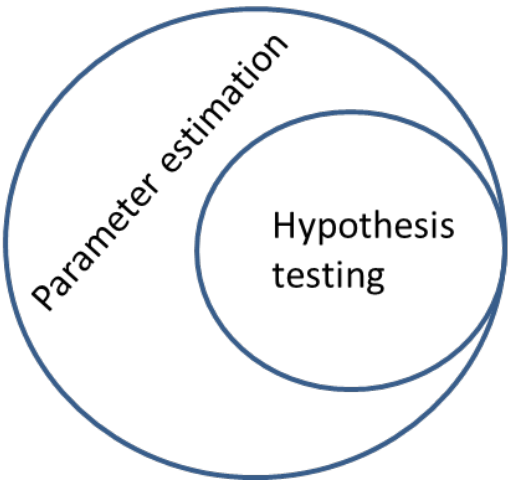


Statistical methods in natural sciences

Statistical efforts can be portrayed as:



For empiricists, statistics suffers from two limitations...

1. Statistical and mechanistic inferences are different "things", rooted in different logical domains…

2. The dualistic nature of hypothesis testing (and our mind) can lead us astray…

E.g…in which study is factor A most important?

"…we saw a very important effect of A on B in our experiments ($P < 0.001$), which shows that…"

"…we saw that A only increased B by 2.1% which shows that…"

- To know it is to beat it (**this lecture**)!

The dualistic nature of our concept of reality

MIND ---------- MATTER

RELIGION ---------- SCIENCE

God ----- Devil

Good ------ Evil

Heaven ------- Hell

Truth -------- Lie

Real ------- Illusory/Unreal

Existing ----- Nonexisting

True ------- False

Accept ----- Reject

Yes ---------- No

Right --------- Wrong

Win --------- Loose

The dualistic scientific method and statistical power

1) Formulate or state a null hypothesis and an alternate hypothesis

H_0: a neutral or "pessimistic" statement (e.g.; no effect; no difference)
H_A: a "positive" statement (e.g.; there is an effect; a difference)

Note: It is *VERY* important to ponder about H_A:
Is $\mu \neq 0$ or is $\mu > 0$ / < 0 under H_A? In the former case, two-tailed tests should be used, in the latter one-tailed tests (is you have a directional H_A and are using two-tailed tests, you are using $\alpha = 0.025$).

2) Given certain assumptions (distributions of variables, types of data etc), H_0 is tested with probabilistic theory. Our P value expresses the probability of our observation, or one more extreme, if H_0 is true. A certain error rate α is accepted (commonly 5%).

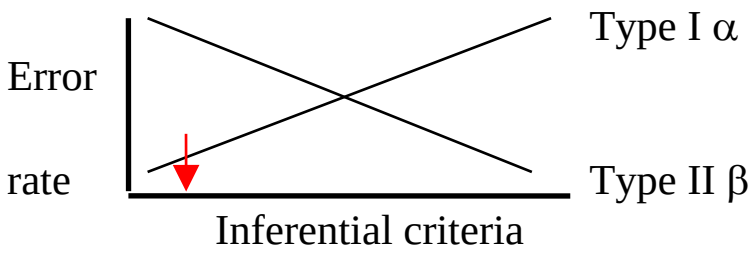
H_0:
- $P < \alpha$: We reject H_0, and hence accept H_A
- $P > \alpha$: We do not reject H_0 (but we do ***NOT*** accept H_0!!)

Hence; simple tests *cannot* lead to the acceptance of H_0! The only hypothesis that can be accepted is H_A

3) In conventional statistical inferences, two types of errors (type I and II) will be made with certain rates/risks (α and β):

	If H_0 is true	***If H_0 is false***
If H_0 is rejected	*Type I error (α)*	*No error*
If H_0 is not rejected	*No error*	*Type II error (β)*

4) The risk of conducting type I and type II errors (the values of α and β) are inversely related to one another. Decreasing α means increasing β, and vice versa. α is always specified in tests, but β is typically unknown/unspecified.



Which error is made most often?

* Type I error is made in about 5% of all tests
* Type II errors is made in about 40 - 80% of all tests!

5) Family-wide type I error rate increases with number of tests performed (e.g., on a given data or in a given study) [equals 1-(1-α)^number of tests]. Called the problem of "**multiple testing**" or "mass inference". The level of α can be adjusted according to the number of tests performed (preferentially by a **sequential bonferroni procedure** or by **false discovery rate** compensation), but this inevitably increases the rate of type II errors.....

STATISTICAL POWER

The statistical power of an inferential test is:
* the probability that (given certain conditions) it will yield statistically significant results ("find an effect") *or in other words:*
* the probability that it will lead to rejection of a false null hypothesis

Power = 1 - β

The power of a test depends on;

1) The chosen α, or the "significance" criterion.

2) The reliability & quality of the test (e.g., N or sample variance).

3) The effect size, or the degree to which a phenomenon is present in a population. The effect size equals 0 under H_0. *So, if H_0 is false, it can be false to different degrees*...!

Power = ***f*** {1,2,3}. Any of these four can be "solved for" (or found) in a power analysis. Pa are available for all kinds of statistical models/test.

Why power analysis?
1) Experimental planning: "tolerable" power = 0.8 (guidline)
2) Can help interpret. inabilitites to reject H_0: a lack of power or is H_0 "true"?
[**Note**: *retrospective power values provided by some programs are bogus – do **not** do ad hoc power analysis!*]
3) Very important indeed where a failure to reject a false null (type II errors) is taken to motivate conclusions or actions (e.g, medical sciences, applied ecology; conservation biology; ecological monitoring programs)!

Most of all: by introducing the concept of effect size, it helps us see beyond our dualistic logic, by acknowledging *the importance of degrees*! H_0 is not true or not, it can be false to different degrees. Helps shift focus from statistical P values to magnitudes of biological effects.

”...We were able to document a very strong correlation between X and Y (P<0.001) in *R. habilis*. This contrasts with the study of Smith (1985) of the congeneric *R. assimilis*, who did not find any such corrlelation (P>0.5). This may be due to...bla...bla...bla”

Error1: A P-value in itself is NOT informative of the strenght of an effect.

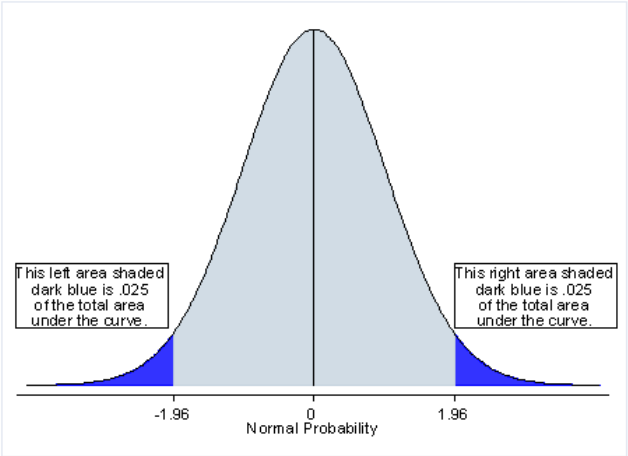
Error2: The fact that one study reject the null hypothesis and another is unable to do so, does NOT imply that the results ”contrast”, are ”conflicting” or even ”different”! In other words, Smith (1985) did not accept the null hypothesis....she/he just failed to reject it.

	R	N	Power (given a true r=0.5)
”We”	0.48	298	0.99
Smith	0.51	13	0.43

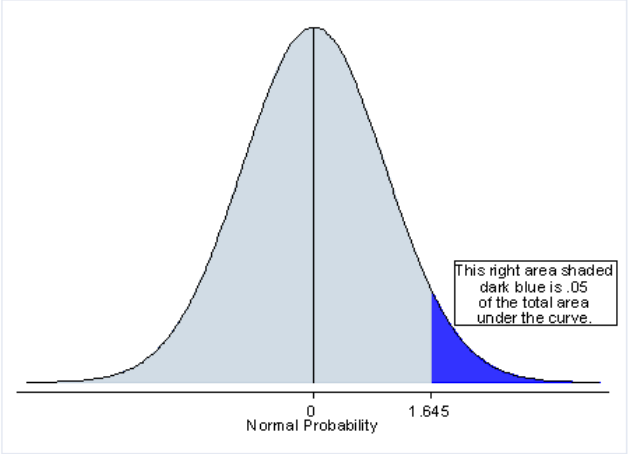
When looked at closer, the studies do not seem different at all....

Important notes on tests and power!

* **Two-tailed tests** (H_A: $\mu \neq 0$) are the norm...but often poorly motivated:



* **One-tailed tests** (e.g. H_A: $\mu < 0$) are rare and often just used the effect can only occur in one direction (one outcome is essentially impossible)...but should be used when relevant: **statistical power is higher**.



* **Directed tests** fall in between these ”extremes”: use when a priori predictions concerning direction of effects are present (see Rice and Gaines 1994). Improved statistical power: one can set, e.g., the lower critical area to 0.01 and the higher to 0.04. Under-utilized: enable detection of patterns that are opposite to predictions while retaining much of the **statistical power** of one-tailed tests!

EXPERIMENTAL DESIGN

Experimental or observational unit: the unit at which observations are made and/or experimental treatments are imposed

Key basic concepts in experimental design:
1) Replication of units
    * Sufficient number of replicates
    * Appropriate scale (cf. pseudoreplication)
2) Controls or control treatments
3) Randomization of units (→ use random numbers!)
    * Random sample from a population (representative)
    * Random allocation of experimental units to treatments (confounding)
5) Independence of units

Treatment factors can have either random or fixed effects:
* **Random effects factors**: levels represent a random sample of an infinite number of possible levels (e.g.; individual, lakes, etc).
* **Fixed effects factors**: levels are fixed and determined by the experimentalist (e.g.; temperature, food availability, concentration of toxins, etc).

One-way designs:

20 replicates	20 replicates	20 replicates	20 replicates	20 replicates

Factor A treatment level: 1 2 3 4 5
(one *can* be control)

* Balanced and unbalanced designs…

Two-way designs:

Factor A treatment level:

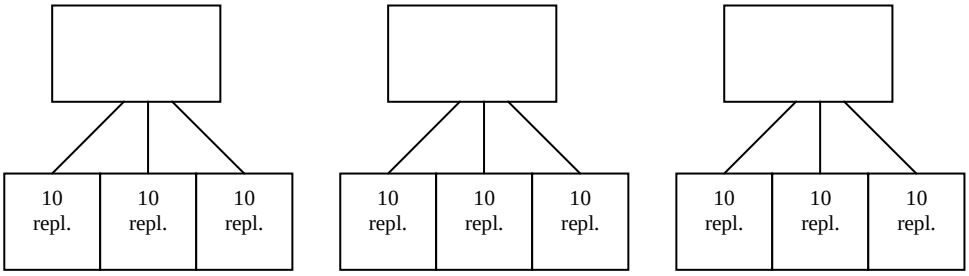
1	20 replicates	20 replicates	20 replicates	20 replicates	20 replicates
2	20 replicates	20 replicates	20 replicates	20 replicates	20 replicates
3	20 replicates	20 replicates	20 replicates	20 replicates	20 replicates
4	20 replicates	20 replicates	20 replicates	20 replicates	20 replicates

Factor B treatment level: 1 2 3 4 5

* Fully crossed, or orthogonal, design strongly preferable!
* Balanced and unbalanced designs…
* Empty cells (missing treatment combinations)…
* Three-way (or multiway) designs have additional factors (i.e., dimension).

Nested, or hierarchical, designs:

Factor A treatment levels (typically fixed): 1 2 3



Factor B levels (typically random): A B C D E F G H I

* The levels of the nested factor (e.g., B above) within each level of the main factor (e.g., A above) are unique to that main factor level.

Instead of randomization of units, “systematic” allocation to “strata” but random within strata is sometimes needed or desirable!

* **Stratified random sampling** from a population (e.g. different areas)
* Systematic allocation of experimental units to **randomized blocks**

Randomized blocked designs:

Factor A treatment levels: 1 2 1 2

Factor B treatment levels:					
	1	1 replicates	1 replicates	1 replicates	1 replicates
	2	1 replicates	1 replicates	1 replicates	1 replicates
	3	1 replicates	1 replicates	1 replicates	1 replicates
	4	1 replicates	1 replicates	1 replicates	1 replicates

Block: A B

* Any design (e.g. one-way or two-way) can be blocked.
* Common when space/time is limited (exp has to be repeated), or when discrete confounding “blocking factors” exists (examples; day, room in greenhouse, enclosure, technician).
* Often **very efficient** and can greatly increase precision of exp and analysis!
* **Randomized complete block design** most common and preferable, but several **incomplete designs** also exists.
* **Latin squares designs** are extensions that can be used when more than one blocking factor exists. Lsd’s are efficient, but analytically restrictive.

Repeated measures designs and split-plot designs:

* Closely related designs in which one (or more) factor is applied ***within*** subjects/units and one (or more) factor is applied ***between*** subjects/units: factors apply to two "levels".
* Observations within subjects **not independent** – special models needed.
* Often **very useful** in biology, but **must be complete (no missing cells)**!
* Examples:

Subject:	***Between*** **subjects/plots factor:**	***Within*** **subjects/plots factor level (amount fertilizer in enclosures):**				
Lake 1	Calcified	1	2	3	4	5
Lake 2	Calcified	1	2	3	4	5
Lake 3	Calcified	1	2	3	4	5
Lake 4	Calcified	1	2	3	4	5
Lake 5	Not calcified	1	2	3	4	5
Lake 6	Not calcified	1	2	3	4	5
Lake 7	Not calcified	1	2	3	4	5
Lake 8	Not calcified	1	2	3	4	5

Subject:	***Between*** **subjects/plots factor:**	***Within*** **subjects/plots factor level (time):**				
Rat/cage 1	Food type A	Jan	Feb	Mar	April	May
Rat/cage 2	Food type A	Jan	Feb	Mar	April	May
Rat/cage 3	Food type A	Jan	Feb	Mar	April	May
Rat/cage 4	Food type A	Jan	Feb	Mar	April	May
Rat/cage 5	Food type B	Jan	Feb	Mar	April	May
Rat/cage 6	Food type B	Jan	Feb	Mar	April	May
Rat/cage 7	Food type B	Jan	Feb	Mar	April	May
Rat/cage 8	Food type B	Jan	Feb	Mar	April	May