LIÈGE université

Statistical Power and Related Biases

Statistics Workshop for Zoologists:
Analytical Insights and Practical Applications

Arnout Van Messem

December 12, 2023

Some topics

- What is statistical power and why is it important?
- How to determine the power?
- Some useful software
- Power in practice

Hypothesis testing

Research question

Is the average length of the male resplendent quetzal in San Gerardo de Dota larger than 38 cm?

- X: length (in cm) of the male resplendent quetzal in San Gerardo de Dota
- $X \sim \mathcal{N}(\mu, \sigma)$
- $H_0 : \mu = 38$
  $H_1 : \mu > 38$
- Sample: $X_1, \ldots, X_n$

Hypothesis testing

Research question

Is the average length of the male resplendent quetzal in San Gerardo de Dota larger than 38 cm?

- X: length (in cm) of the male resplendent quetzal in S[illegible]o de Dota
- $X \sim \mathcal{N}(\mu, \sigma)$
- $H_0 : \mu = 38$
  $H_1 : \mu > 38$
- Sample: $X_1, \ldots, X_n$

Hypothesis testing

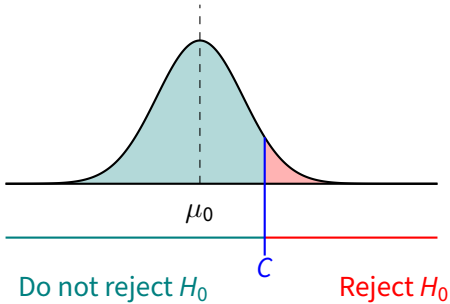
Decide whether or not to reject H_0

- $\bar{X} \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$
- Test statistic **under** H_0: $Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$
- Reject if test statistic becomes too large
- Define significance level α and associated quantile $z_{1-\alpha}$
- Reject if

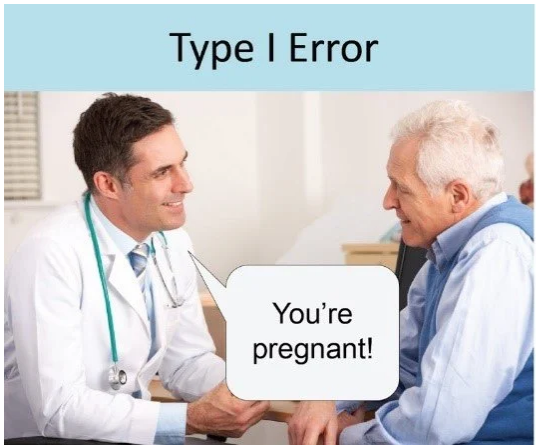
$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha} \Leftrightarrow \bar{X} > \mu_0 + \frac{z_{1-\alpha}\sigma}{\sqrt{n}} := C$$

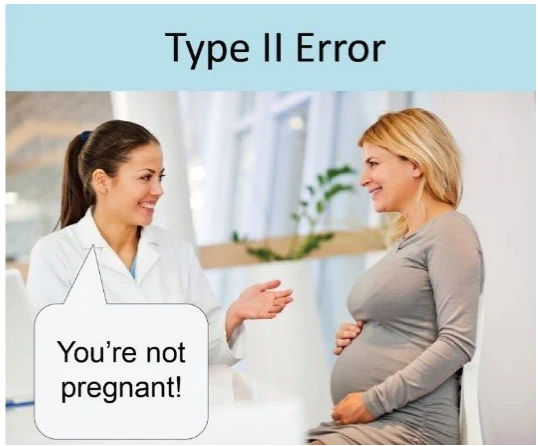
where C is the **critical value**

Hypothesis testing



Errors in hypothesis testing





Errors in hypothesis testing

- **Type I error:** reject H_0 based on the sample while, in reality, H_0 is true in the population

  Probability of type I error limited by **significance level** α

- **Type II error:** Not reject H_0 based on the sample while, in reality, H_0 is not true in the population

  = Not being able to detect a true alternative H_1

  Probability of type II error **for a specific alternative** noted as β. $1 - \beta$ is called the **power** of the test

A well executed statistical test will try to control the probability of both errors!

Errors in hypothesis testing: an example

Research question

Assessing the impact of noise pollution on bird nesting success (f.e. for a rare or endangered species)

- H_0: Noise pollution has no effect on the bird nesting success
  H_1: Noise pollution decreases the bird nesting success
- **Type I error**:
  - Outcome: Researchers incorrectly reject the null hypothesis
  - Result: Conclude that noise pollution significantly reduces nesting success when, in reality, it has no impact
- **Type II error**:
  - Outcome: Researchers incorrectly fail to reject the null hypothesis
  - Result: Conclude that noise pollution has no effect on nesting success when, in reality, it does negatively affect the bird species' nesting success

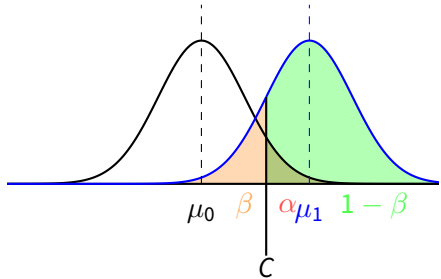
Errors in hypothesis testing

Decision	*Reality*	
	H_0 true	H_1 true
Reject H_0	Type I error probability α	Correct probability $1 - \beta$
Not reject H_0	Correct probability $1 - \alpha$	Type II error probability β

Other terminology:

- False positive (reject H_0 if H_0 is true) – High specificity = low false positives
- False negative (do not reject H_0 if H_1 is true) – High sensitivity = low false negatives

Type I error, type II error and power



It is impossible to minimize the probability of both errors at the same time!

Trade-off between α and β

The power: an overview so far

- Power = the probability of correctly rejecting H_0 a false null hypothesis

  In other words: **the probability of detecting an effect, given that the effect exists**

- More power = better chance of rejecting the null hypothesis and discovering a significant effect

- Thus reducing the likelihood of a type II error

The power: why is it useful

It will help answer questions such as:

- What is the minimal required sample size?
- How to design an experiment?
- Which test to use?
- I can collect a sample of size n, will my test be sufficiently powerful?
- Will I be able to detect an effect that is not only statistically significant, but also scientifically meaningful?

In scientific research it will also

- Give information to grant reviewers to assess the quality of a proposal and its potential for success
- Make results more reliable and reproducible

Significance and power

- Normally, the restriction on the probability of a type I error will be more strict than that on the probability of a type II error
- Common choices: $\alpha \in \{5\%, 2, 5\%, 1\%\}$, $\beta \in \{20\%, 10\%\}$
- If α diminishes, the null hypothesis is less often rejected and thus β automatically becomes larger
- Therefore, first fix α and then look for a test that makes β as small as possible and thus the power as large as possible

Influencing the power

- Find an expression of the power **under the specific alternative** $H_1 : \mu = \mu_1$ or $\bar{X} \sim \mathcal{N}(\mu_1, \sigma/\sqrt{n})$:

$$
\begin{aligned}
1-\beta &= P(\text{reject } H_0 | H_1 \text{ is true}) \\
&= P\left(\bar{X} > \mu_0 + \frac{z_{1-\alpha}\sigma}{\sqrt{n}} | H_1\right) \\
&= P\left(\frac{\bar{X}-\mu_1}{\sigma/\sqrt{n}} > \frac{\mu_0 + z_{1-\alpha}\sigma/\sqrt{n} - \mu_1}{\sigma/\sqrt{n}} | H_1\right) \\
&= 1 - \Phi\left(z_{1-\alpha} + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right) \\
&= \Phi\left(-z_{1-\alpha} + \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right)
\end{aligned}
$$

Influencing the power

- $1 - \beta = \Phi\left(-z_{1-\alpha} + \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right)$
- Power tells us beforehand how (at least) large the probability is to detect a true alternative μ_1 (or a more extreme alternative)
- Power depends on
  - Significance level α: if $\alpha \searrow$, then $1 - \beta \searrow$
  - Effect size $\mu_1 - \mu_0$: if $|\mu_1 - \mu_0| \nearrow$, then $1 - \beta \nearrow$
  - Standard deviation σ: if $\sigma \searrow$, then $1 - \beta \nearrow$
  - Sample size n: if $\mu_1 - \mu_0 \nearrow$, then $1 - \beta \nearrow$

$$\textbf{power} \propto \frac{\textbf{effect size}\ \alpha\sqrt{n}}{\sigma}$$

Scientific versus statistical significance

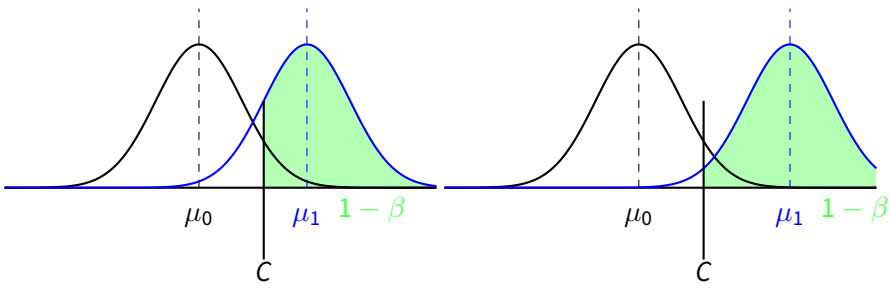
- Statistical significance = rejecting the null hypothesis
- Reject H_0 if

$$\bar{X} > \mu_0 + \frac{z_{1-\alpha}\sigma}{\sqrt{n}}$$

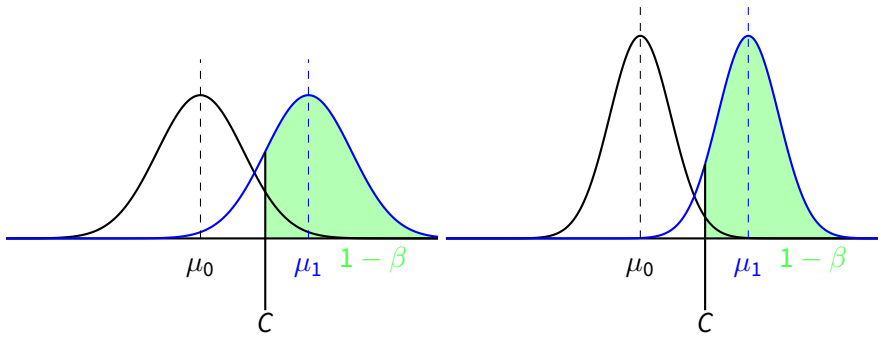
- If n becomes large, easy to pick up on small deviations from μ_0
- Are such deviations scientifically significant?
- For the power, look at the **minimal scientifically significant effect** that you want to detect[1]

[1] Len, T. and Thomas, F.J. (1996). *The importance of statistical power analysis: an example from Animal Behaviour*, Animal Behaviour, **52** (4), 856-859.

Influencing the power: effect size



Influencing the power: standard deviation



Influencing the power: standardized effect size

- Can be difficult to determine a minimal scientifically significant effect
- Jacob Cohen[2] defined ‘small’, ‘medium’ and ‘large’ effects for different tests
- Standardized effect size = $\frac{|\mu_1 - \mu_0|}{\sigma}$

Test	Relevant effect size	Effect Size Threshold		
		Small	Medium	Large
t-test for means	d	0.2	0.5	0.8
F-test for ANOVA	f	0.1	0.25	0.4
t-test for correlation	r	0.1	0.3	0.5
Chi-square	w	0.1	0.3	0.5
2 proportions	h	0.2	0.5	0.8

[2] Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd edn. Hillsdale, New Jersey: Lawrence Erlbaum.

Example power calculation: one-sided test

Research question

Is the average length of the male resplendent quetzal in San Gerardo de Dota larger than 38 cm?

- Determine minimal significant effect: $\mu_1 - \mu_0 = 1$, hence $\mu_1 = 39$
- Assume known standard deviation: $\sigma = 2$
- Sample size: $n = 25$
- Significance level: $\alpha = 5\%$

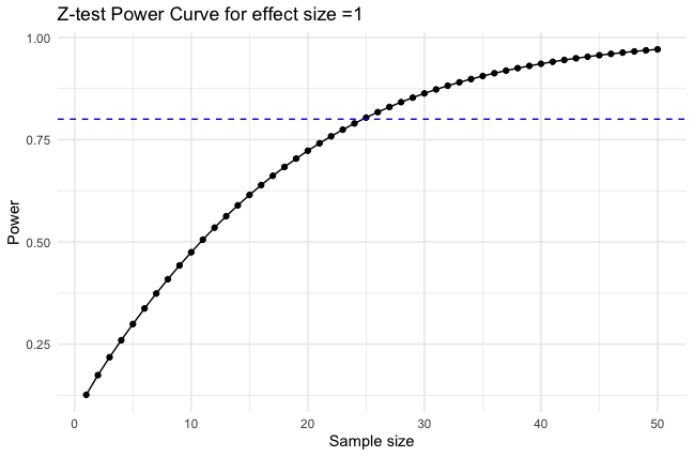
Power : $1 - \beta = \Phi\left(-1.645 + \frac{1}{2/\sqrt{25}}\right) = 80.38\%$

Power curves

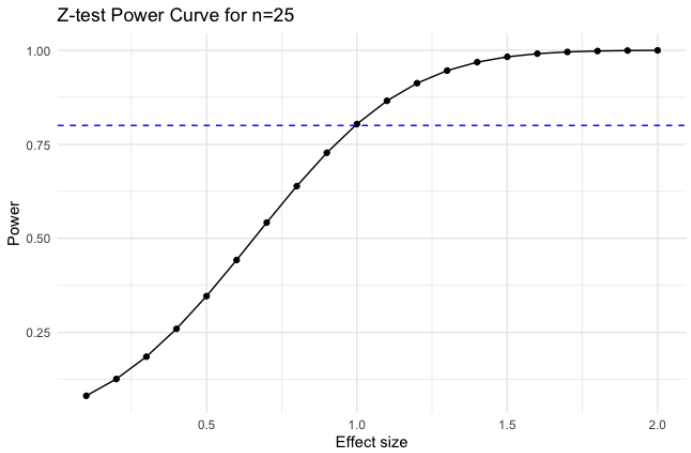
- Plot power in function of
  - $\mu_1 - \mu_0$ (keeping α, σ, and n fixed)
  - n (keeping α, σ, $\mu_1 - \mu_0$ fixed)

- **Examine power in function of sample size for an important alternative μ_1**
- Determine minimal sample size, beforehand, in order to detect an important alternative with sufficient power
  - Exact calculation possible by rewriting formula of the power
- If this sample size is not attainable, review your study!

Power curves



Power curves



R-code

- Package `pwr`, function `pwr.norm.test`
- ! Effect size = $\frac{\mu_1 - \mu_0}{\sigma}$! – “standardized effect size” or “Cohen’s d”

```
library(pwr)
library(ggplot2)

# Calculate the power of the test
pwr.norm.test(d=1/2,n=25,sig.level=0.05,alternative="greater")$power
```

R-code

```r
# Define power curve in function of effect size
power.curve.effect <- function(n, sigma){
  cd <- seq(.1,2,.1)  #Vector of effect size
  samp.out <- NULL
  for(i in 1:length(cd)){
    power <-  pwr.norm.test(d=cd[i]/sigma,n=n,sig.level=.05,alternative="greater")$power
    power <-  data.frame(effect.size=cd[i],power=power)
    samp.out <- rbind(samp.out,power)
  }
  ggplot(samp.out, aes(effect.size,power))+
    geom_line() + geom_point() + theme_minimal() +
    geom_hline(yintercept = .8,lty=2, color='blue') +
    labs(title=paste0("Z-test Power Curve for n=", n),
         x="Effect size",
         y="Power")
}

power.curve.effect(n=25, sigma = 1)
```

R-code

```r
# Define power curve in function of sample size
power.curve.n <- function(d, sigma){
  cn <- seq(1,30,1)  #Vector of sample size
  samp.out <- NULL
  for(i in 1:length(cn)){
    power <-  pwr.norm.test(d=d/sigma, n=cn[i],sig.level=.05,alternative="greater")$power
    power <-  data.frame(n.size=cn[i],power=power)
    samp.out <- rbind(samp.out,power)
  }
  ggplot(samp.out, aes(n.size,power))+
    geom_line() + geom_point() + theme_minimal() +
    geom_hline(yintercept = .8,lty=2, color='blue') +
    labs(title=paste0("Z-test Power Curve for effect size =", d),
         x="Sample size",
         y="Power")
}

power.curve.n(d=2, sigma=1)
```

Example power calculation: two-sided test

Research question

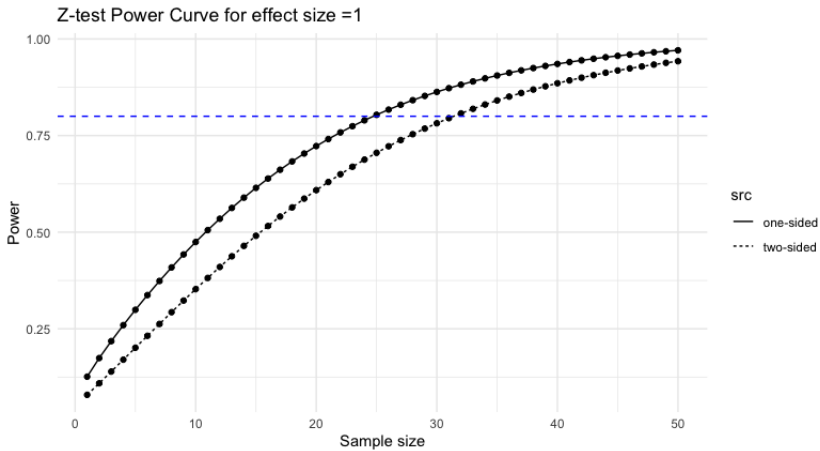
Is the average length of the male resplendent quetzal in San Gerardo de Dota **different** from 38 cm?

- Minimal significant effect: $\mu_1 - \mu_0 = 1$, hence $\mu_1 = 39$
- Known standard deviation: $\sigma = 2$
- Sample size: $n = 25$
- Significance level: $\alpha = 5\%$

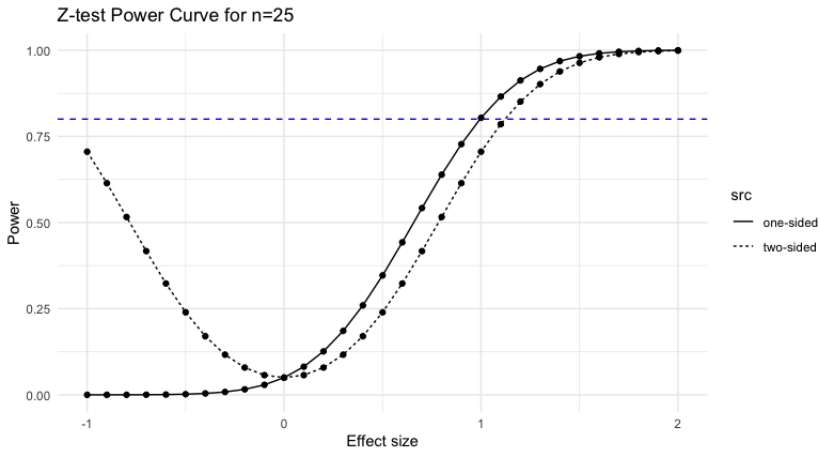
$$\text{Power: } 1 - \beta = \Phi\left(z_{\alpha/2} + \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right) + \Phi\left(z_{\alpha/2} + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right)$$

$$\approx \Phi\left(z_{\alpha/2} + \frac{|\mu_1 - \mu_0|}{\sigma/\sqrt{n}}\right) = \Phi\left(-1.96 + \frac{1}{2/\sqrt{25}}\right) = 70.54\%$$

Power curves



Power curves



Power software

Commercial	SAS	POWER procedure
	STATA	Power, Precision, and Sample-Size
	Power Analysis and Sample Size (PASS)	
	Power and Precision	
	Nquery	
Free	`R`	`pwr`
	G*Power	
	GeoGebra	(selected power curves)
	https://homepage.divms.uiowa.edu/ rlenth/Power/	(Java applets)

And many other website, often for very specific purposes/tests/methods

In practice: flow chart

Hypothesis

↓

Experimental design

Statistical test

↓↑

Power analysis

↓

Sample size

↓

Experiment

↓

Statistical analysis

In practice: power analysis

- Usually, statistical tests fix α to control the probability of a type I error
- It is also important to have an idea about the type II error for a certain, given alternative
  - If the power is not high, the test might be a waste of time, money, resources, ...
- Perform a **power analysis** at the start of the study (during the design phase)
  - Select design of the study and statistical test
  - Select minimal sample size to attain certain power
  - Determine minimal detectable effect in the presence of limitations (financial, logistic, time-wise, …)
  - Post-hoc analysis after non-significant result to evaluate if power was sufficiently high given the sample size and realistic effect sizes

In practice: evaluating the limitations of an experiment

- In practice, power analysis will often lead to (relatively) large n, and it will not always be possible to obtain that sample size
- In that case, **sample size takes priority**: determine smallest effect that could be detected with sufficient probability

Back to the quetzal

Based on 5 observed quetzal in the San Gerardo de Dota valley, the smallest difference in mean length that can be detected with 80% probability at the 5% significance level through a one-sided Z-test, is 2.23 cm (given a standard deviation of 2 cm)

Similar calculation: the test only has 29.91% power to detect a difference of 1 cm if the sample size is 5

In practice: some random remarks

- Perform an ANOVA test instead of multiple pairwise comparisons (with Bonferroni correction) to avoid loss of power
- Non-parametric tests are usually less powerful than their parametric counterpart, but often the difference is negligible
  - However, power calculation for non-parametric tests is not always easy
  - Can be performed through simulations using, f.e., bootstrapping
  - Crude rule of thumb: compute the sample size for the corresponding parametric test and add 15%

In practice: p-value and power

- p-value: probability that the observed test statistic (in the sample) occurred by chance alone
- Decision rule: p-value vs. significance level
- Good practice: never interprete p-value without looking at the power
  - **Significant result ($p < \alpha$)**: Nice! But what is the corresponding difference in means?
    - $\geq$ minimal scientifically significant effect: exciting
    - $<$ minimal scientifically significant effect: not really exciting
      Most likely very big sample and too much power
  - **Not significant result ($p > \alpha$)**: Bummer. But how big was the sample?
    - Sufficiently large = enough power: there really is no effect
    - Too small = not enough power: there might be a meaningful effect that we did not pick up on

In practice: *n* too big

For very large values of *n*, the power will increase to 100% (and p-value to 0). This means that any effect, no matter how small, has a high probability to be discovered.[3]

Strategies to ensure more meaningful and reliable statistical results include:

- Consider and report effect size
- Consider practical significance
- Power analysis
- Significance level
- Use robust statistical methods
- Replication/resampling: for example through bootstrapping or jackknife
- Avoid p-hacking

[3] Lin, M., et al. (2013). *Too Big to Fail: Large Samples and the p-Value Problem*. Information Systems Research, 24(4), 906–917.

In practice: *n* too big

- Bootstrapping
  - Original sample of size n
  - Create a large number (N) of bootstrap samples of size $n_B = n$: random selection, with replacement, from the original dataset
  - For each bootstrap sample, perform hypothesis test
  - Use bootstrap distribution to construct confidence intervals or mean values
- Jackknife
  - Special case of bootstrapping where you leave one observation out
- In this context: create samples of smaller size than original

In practice: n too big

Comparing daytime and night-time[4]

Compare a specific type of sound (referred to as /kwa/) during two different time periods (night-time and peak sound production).

- $21,604$ sound selections for night-time
- $7,286$ sound selections for peak sound production period
- Subsampling: $n_S = 100$ over $N = 1000$ replications
- Mean values are calculated

[4] Di Iorio, L., et al. (2018), *Posidonia meadows calling: a ubiquitous fish sound with monitoring potential*. Remote Sens Ecol Conserv, 4, 248-263.

In practice: *n* too big

Table S2. Comparison of the /*kwa*/ features during the subsampling period (i.e. the two hours of peak production) vs. the rest of the night.

	Mean of 1000 *p* values (paired Student's t-test, $n_{obs} = 100$)	**Median of 1000 *p* values (paired Student's t-test, $n_{obs} = 100$)**
HI	0.46	0.45
ΔFsp	0.50	0.52
ΔFep	0.51	0.53
ΔFcp	0.52	0.53
F_{min}	0.48	0.48
F_{peak}	0.47	0.46
H1	0.50	0.49
H2	0.49	0.48
H3	0.50	0.50
H4	0.51	0.53
T	0.49	0.48

In practice: *n* too big

Results

By subsampling the dataset, the authors have been able to find more realistic p-values and draw biologically relevant conclusions (instead of finding that all tests are significant due to the sample size)

In practice: n too big

Comparing shrimp sounds[5]

Compare a number of features of shrimp sounds between 4 different sites.

- $32,928$ observations for daytime and $37,633$ for night-time
- Subsampling over 999 replications on various subsampled data sizes ($n = 100, 500, 1,000, 5,000$, and $10,000$)

[5] In preparation

In practice: *n* too big

Table 19 Inter-sites comparisons of the BTS features. The P-values were obtained through bootstrap analyses with 999 permutations performed on varying subsampled data sizes (n = 100, 500, 1000, 5,000, and 10,000). For each acoustic feature and for temporal period (daytime and night-time), the P-value of the Kruskal-Wallis (KW) test is displayed, followed by the P-values of the Dunn post-hoc comparisons. Smaller sample sizes revealed pronounced differences that aligned with visual observations, whereas larger sample sizes exhibited increased statistical power but also highlighted minimal differences that may be irrelevant to the study. Cells with non-significant P-values (P-values ≥ 0.05) are highlighted.

night-time BTS ANL					
	N =100	N = 500	N = 1,000	N = 5,000	N = 10,000
KW	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Dunn, S1 vs. S2	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Dunn, S1 vs. S3	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Dunn, S2 vs. S3	0.44	0.15	0.049	< 0.0001	< 0.0001
Dunn, S1 vs. S4	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Dunn, S2 vs. S4	0.46	0.18	0.076	< 0.0001	< 0.0001
Dunn, S3 vs. S4	0.21	0.0017	< 0.0001	< 0.0001	< 0.0001
daytime BTS ANL					
	N =100	N = 500	N = 1000	N = 5000	N = 10000
KW	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001

In practice: *n* too big

Results

The smaller sample sizes revealed noticeable differences that correspond with visual observations, while the larger sample sizes provide increased statistical power but also highlight minimal differences that might be insignificant.

In practice: *n* too small

For small n, the power of the test will be low and thus it becomes very difficult to discriminate between the null and alternative hypothesis (except for large differences).

Strategies to increase power include:

- Increase sample size
- Use a more powerful test
- Optimize experimental design
- Reduce variability
- Focus on Strong effects
- Use a one-sided test
- Increase α (type I error rate)
- Combine similar studies

In practice: *n* too small

Dietary supplement

Study the effect of a new dietary supplement on the average weight gain of a particular species of laboratory mice over a four-week period.
Hypothesis: the new supplement will lead to a significant increase in weight.
Constraint: sample of 10 mice (5 in the test group and 5 in the control group).

- **Effect size estimation:** pilot study / literature review to estimate expected effect size
- **Increase homogeneity:** make mice in both groups as homogeneous as possible to decrease variability
- **Use a one-sided test:** if you have strong reason to expect weight gain and not loss
- **Increase measurement precision:** accurate measurements decrease variability
- **Consider collaboration:** combine data from multiple studies to increase sample size

In practice: *n* too small

Influence of a stressor

Understand how exposure to a specific environmental stressor, such as temperature variation, affects the heart rate of a rare species of turtle.

Constraint: sample size is limited to 10 (5 exposed to stressor, 5 control) due to limited availability of the species and ethical considerations.

- **Repeated measures design:** measure each turtle before and after stressor (each turtle is its own control)
- **Precise measurement tools:** accurate measurements decrease variability
- **Time-of-day controls:** standardize time of data collection to reduce influence of natural fluctuations in heart rate during the day
- **Statistical modeling:** consider statistical modeling techniques instead of testing

In practice: *n* too small

Caution

Remember, while these strategies can be helpful, it is important to transparently report on the limitations of a small sample size, to acknowledge the potential for type II errors, and to interpret the results carefully. Emphasize the biological relevance of your findings (even though the power might be low).