Let’s summarize the steps involved in statistical hypothesis testing

1. Think about how to transform the scientific question into a statistical question.

2. State the null (parameter for a theoretical population of no interest) and alternative hypotheses based on population values.

3. Compute the appropriate test statistic based on the sample (usually involving a combination of mean and standard error - so far).

4. Determine the p-value by contrasting the sample value with a sampling distribution that assumes the null hypothesis to be true (theoretical population), i.e., probability of finding the observed, or a more extreme value in the sampling distribution of the theoretical population.

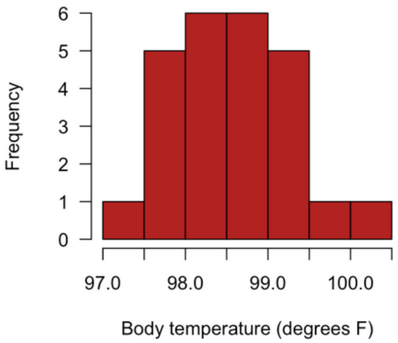
5. Draw a conclusion by comparing the observed P-value against the significance level (α). If P-value greater than α, then do not reject H_0; if P-value smaller than or equal to α, then reject H_0.

Normal human body temperature, as kids are taught in North America, is 98.6°F. But how well is this supported by data?

Because we testing these hypotheses based on a single sample of 25 individuals using the t-test, we refer to this as a **one-sample t test**

H_0 (null hypothesis): the mean human body temperature is 98.6°F.

H_A (alternative hypothesis): the true population is different from 98.6°F.



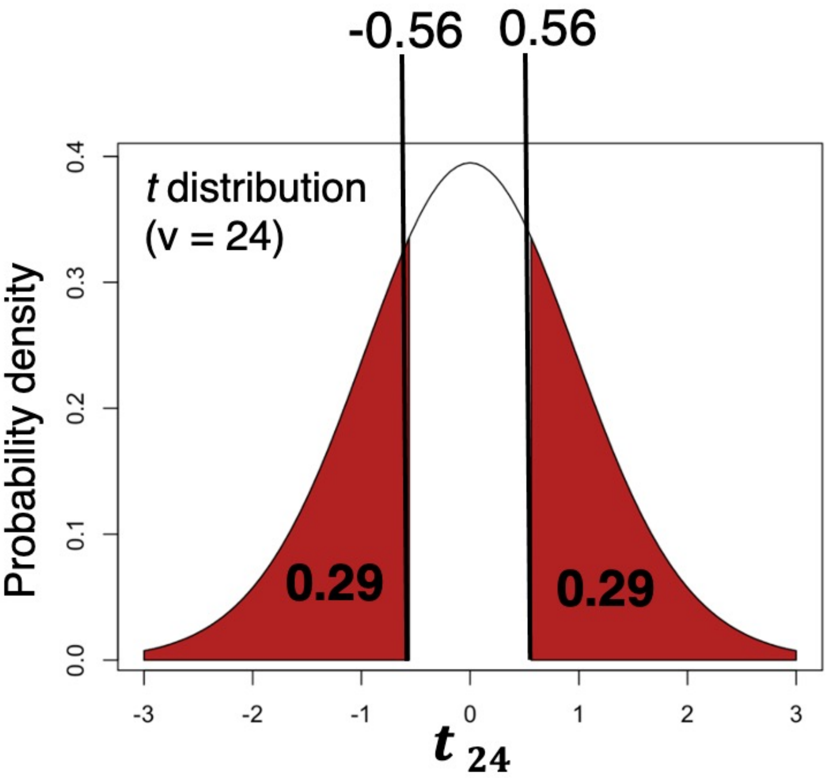
$$\bar{Y} = 98.524$$

We started with: *Normal human body temperature, as kids are taught in North America, is 98.6^{o}F. But how well is this supported by data?*

Then "translated" the above question into: What is the probability of obtaining a sampling *mean as extreme or more extreme* (i.e., *smaller*) than 98.524^{o}F given that the population mean is 98.6^{o}F?

$$t = \frac{98.524 - 98.6}{0.136} = -0.56$$

Pr[t < -0.56] + Pr[t > 0.56] =
2 Pr[t > abs(0.56)] = **0.58**
(*t* is symmetric around μ)



By not rejecting H_0, we cannot state that the true population value is 98.6^{o}F; all we can say is that we have no evidence to state that it does not!

BUT there is always possibility for new evidence to be put together in the future to reject the original conclusion. **HOW?**

The effects of increasing sample size on hypothesis testing: body temperature revisited

The effects of larger sample size on hypothesis testing: body temperature revisited

Let's say that we took a new sample of 130 individuals (instead of 25 as in our previous sample). The values for the new sample are:

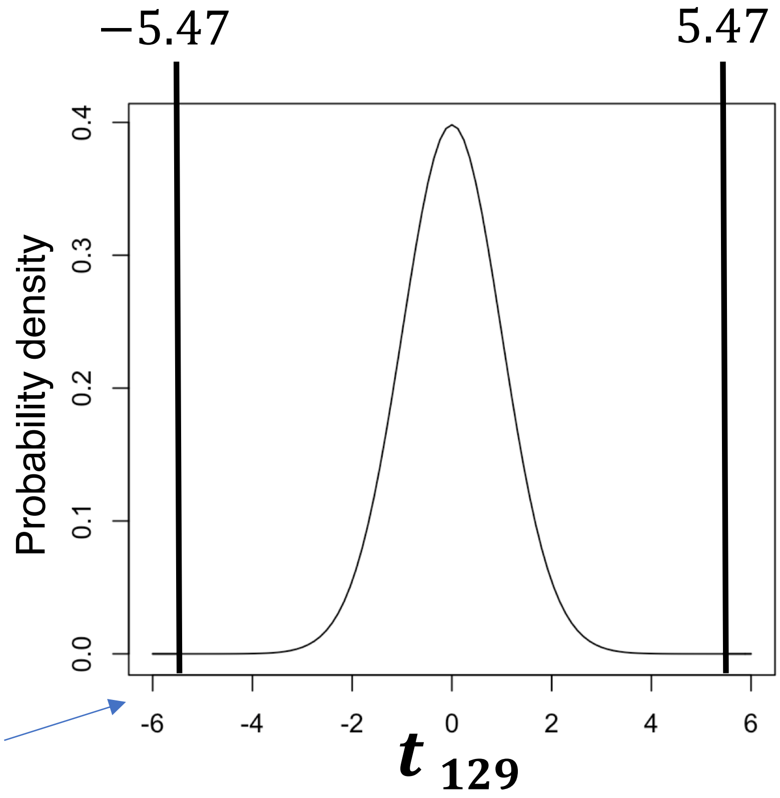
$$\bar{Y} = 98.25^{o}\text{F}$$

$$s = 0.733^{o}\text{F}$$

$$t = \frac{98.25 - 98.6}{0.064} = -5.47$$

$$\text{SE}_{\bar{Y}} \frac{0.733}{\sqrt{130}} = 0.064$$

Pr[t < -5.47] + Pr[t > 5.47] = 2 Pr[t > abs(5.55)] = 0.000002



Can't see the area in red as it is too small

The effects of larger sample size on hypothesis testing: body temperature revisited

$$n = 25$$

$$\bar{Y} = 98.524$$

$$s = 0.678$$

$$\mathrm{SE}_{\bar{Y}} \frac{0.678}{\sqrt{25}} = 0.136$$

$$t = \frac{98.524 - 98.6}{0.136} = -0.56$$

Pr[t < -0.56] + Pr[t > 0.56] =
2 Pr[t > abs(0.56)] =
0.58

$$n = 130$$

$$\bar{Y} = 98.25^{o}\mathrm{F}$$

$$s = 0.733^{o}\mathrm{F}$$

$$\mathrm{SE}_{\bar{Y}} \frac{0.733}{\sqrt{130}} = 0.064$$

$$t = \frac{98.25 - 98.6}{0.064} = -5.47$$

Pr[t < -5.47] + Pr[t > 5.47] =
2 Pr[t > abs(5.55)] =
0.000002

The effects of larger sample size on hypothesis testing: body temperature revisited – *in line of new and stronger evidence*:

H_0 (null hypothesis): the mean human body temperature is 98.6°F.

H_A (alternative hypothesis): the true population is different from 98.6°F.

THE NEW SAMPLE LEADS TO A P-VALUE = 0.0000002 ($P < \alpha = 0.05$), SO WE REJECT THE NULL HYPOTHESIS IN LINE OF THIS NEW EVIDENCE.

Therefore, we have NEW AND STRONGER (larger sample size) evidence to **state that it is very likely that the true value of human body temperature is different from 98.6°F (note that we are not saying that it is not 98.6F)**

The effects of larger sample size on hypothesis testing: body temperature revisited – *in line of new and stronger evidence*:

As we saw in previous lectures, sample size ***decreases*** the standard error, which makes the t value (test statistic) increase, which in turn leads to smaller p-values.

Smaller P values allows rejecting the null hypothesis. As such, increased sample values lead to greater ***statistical power*** (smaller Type II errors) to reject the null hypothesis when it is not true!

$$t_i = \frac{\bar{Y} - \mu}{\frac{s}{\sqrt{n}}} \longleftarrow$$

Remember: The power of a test (1-β) is the probability of rejecting the null hypothesis when is truly false; it is difficult to estimate (advanced stats). This probability increases as sample size increases.

Again, because we only have one sample, we call this a one-sample t test

H_0 (null hypothesis): the mean human body temperature is 98.6^{o}F.

H_A (alternative hypothesis): the true population is different from 98.6^{o}F.

Assumptions of the one-sample t test (very important):

1) The data are a random sample from the population (either from the theoretical) or any of the other possible populations from which the sample may have been sampled from. This assumption is shared by all tests covered in this course and used to test biostatistical hypothesis.

2) The variable (e.g., human temperature) is normally distributed in the population.

Statistical hypothesis testing for comparing two samples described by a quantitative variable

One- and two-sample hypothesis testing

One sample

One sample (frogs) according to a single categorical variable (Left/Right)
Binomial test

One sample (humans) according to a single quantitative variable (temperature)
One-sample t-test

Two samples

Paired-design
Paired t-test

independent-design
Two-sample t-test (equal variance)

independent-design
Two-sample t-test (unequal variance) [coming soon]

Multiple samples

independent-design
Analysis of Variance (equal variance) [coming soon]

Examples of statistical hypothesis testing for comparing two sample means:

Do female hyenas differ from male hyenas in body size?

Do patients treated with a new drug live longer than those treated with an old drug?

Do students perform better on tests if they stay up late studying or get a good night's rest?

Statistical hypothesis testing for comparing two sample means

Scientific question: Does clear-cutting a forest affect the number of salamanders present?

- There are two treatments: clear cutting / no clear-cutting (control).

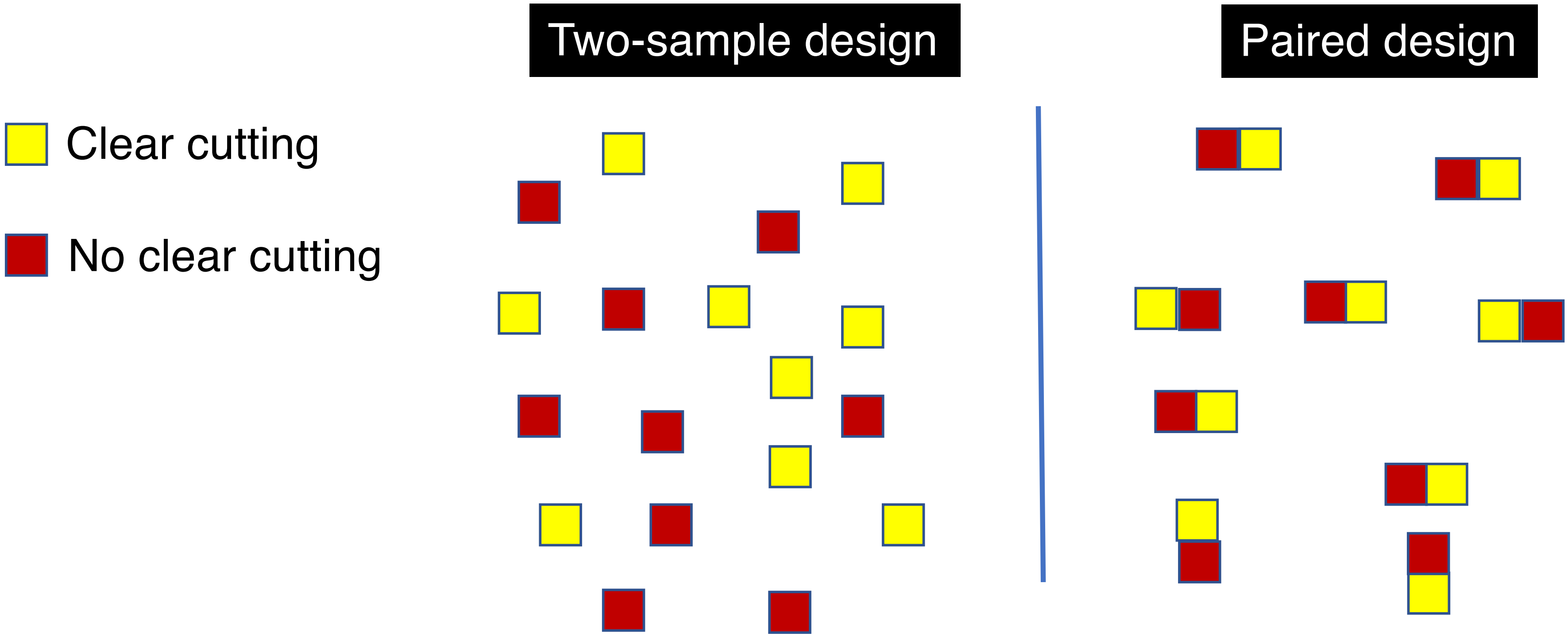
- Statistical question: Does the mean number of salamanders differ between the two treatments?

- Treatment is a *categorical variable* and number of salamanders is a *numerical variable*.

Paired sample *versus* two independent samples

Scientific question: ***Does clear-cutting a forest affect the number of salamanders present?*** There are two main alternative study designs that affect the choice of statistical test:

In the ***two-sample* design**, each treatment group is composed of an independent, random sample unit.
In the ***paired design***, both treatments are applied to every sampled unit (here - forest plots).

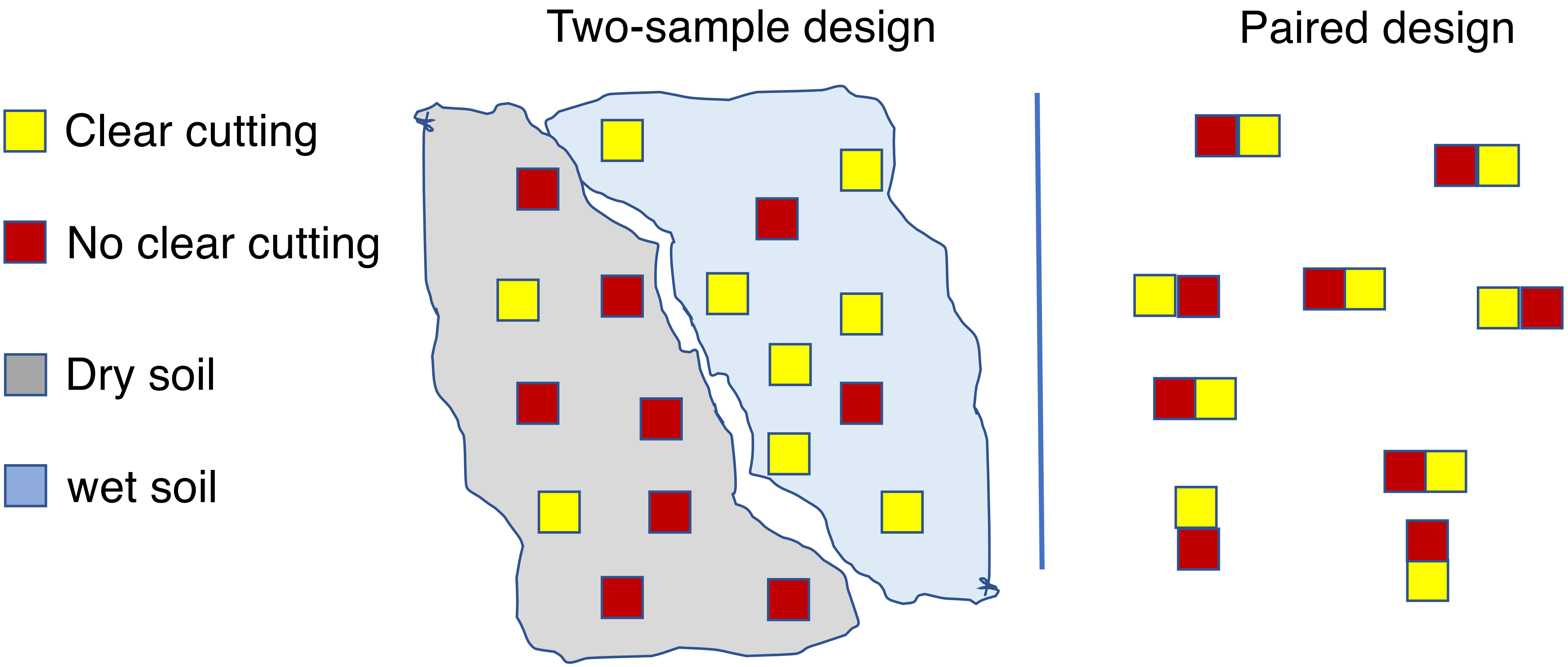


Paired design for comparing two sample means

Paired comparison of two means

Scientific question: ***Does clear-cutting a forest affect the number of salamanders present?***

The advantage of the paired design is that it reduces the effects of variation among **sampling units** that has nothing to do with the treatment itself (e.g., local environmental differences among units). It reduces confounder variables.

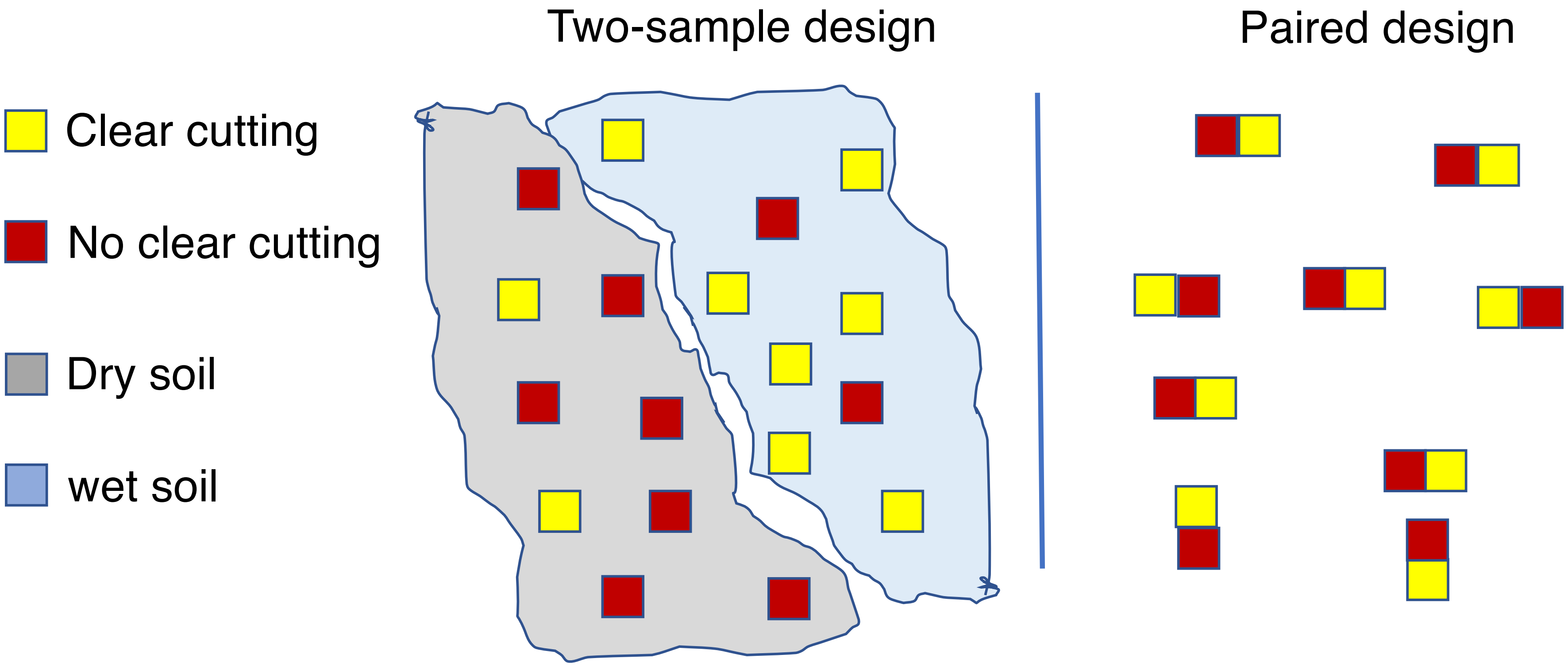


Paired comparison of two means

Scientific question: ***Does clear-cutting a forest affect the number of salamanders present?***

The advantage of the paired design is that it reduces the effects of variation among **sampling units** that has nothing to do with the treatment itself
(e.g., local environmental differences among units). It reduces confounder variables.

Note how clear cutting happened more in wet soil than no clear cutting which predominated dry soils. If soil moisture is important to salamanders, then this non-random distribution of observation units could affect the conclusion.



Paired comparison of two means

The advantage of the paired design is that it reduces the effects of variation among sampling units that has nothing to do with the treatment itself (e.g., local environmental features).

Other examples of paired study designs:

- Comparing patient weight before and after hospitalization.
- Comparing fish species diversity in lakes before and after heavy metal contamination.
- Testing effects of sunscreen applied to one arm of each subject compared with a placebo applied to the other arm.
- Testing effects of smoking in a sample of smokers, each of which is compared with an non-smoker closely matched by age, weight, and ethnic background.
- Testing effects of socioeconomic condition on dietary preferences by comparing identical twins raised in separate adoptive families that differ in their socioeconomic conditions.

A previously seen example of paired design:

Tidarren (spider)

It gives an “arm” (or a pedipalp) for a female spider.

Running speed (cm/s) of male *Tidarren* spiders before and after voluntary amputation of one pedipalp.

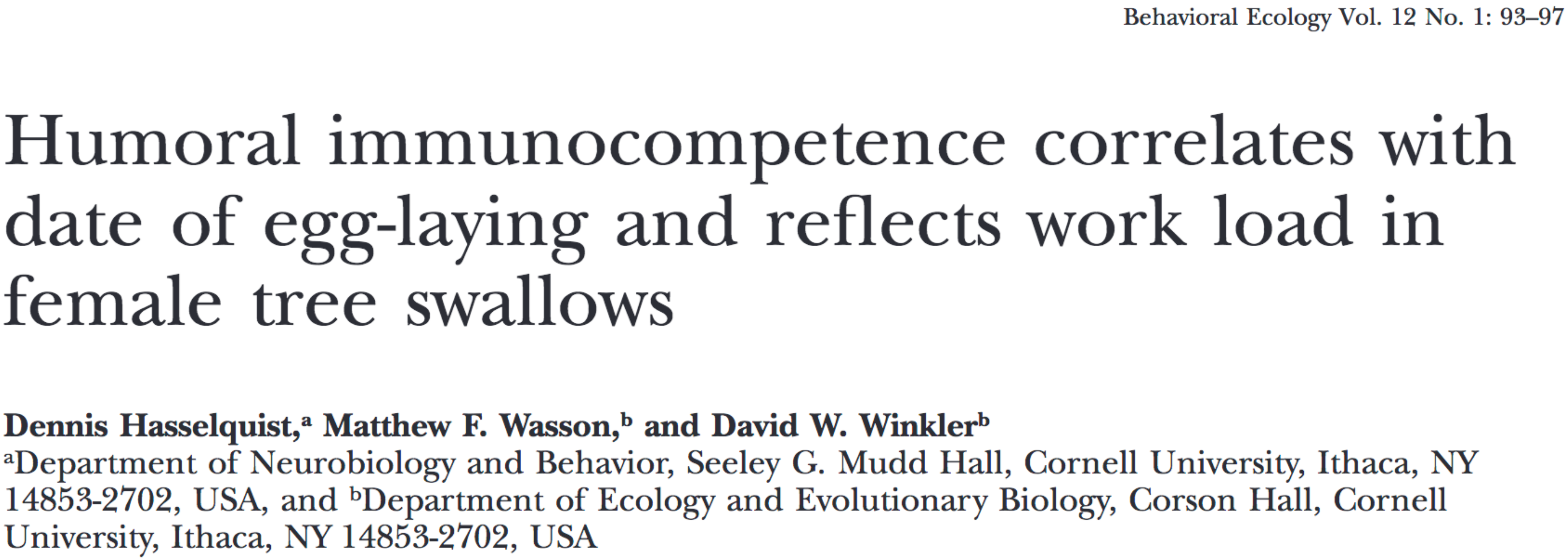
Oxyopes salticus

Spider	Speed before	Speed after
1	1.25	2.40
2	2.94	3.50
3	2.38	4.49
4	3.09	3.17
5	3.41	5.26
6	3.00	3.22
7	2.31	2.32
8	2.93	3.31
9	2.98	3.70
10	3.55	4.70
11	2.84	4.94
12	1.64	5.06
13	3.22	3.22
14	2.87	3.52
15	2.37	5.45
16	1.91	3.40

Paired comparison of two means – an empirical example

- In many species, males are more likely to attract females if males have high testosterone levels.

- **Research question:** Are males with high testosterone paying a cost for this extra mating success in other ways (trade-offs)?



Behavioral Ecology Vol. 12 No. 1: 93–97

Humoral immunocompetence correlates with date of egg-laying and reflects work load in female tree swallows

Dennis Hasselquist,[a] Matthew F. Wasson,[b] and David W. Winkler[b]
[a]Department of Neurobiology and Behavior, Seeley G. Mudd Hall, Cornell University, Ithaca, NY 14853-2702, USA, and [b]Department of Ecology and Evolutionary Biology, Corson Hall, Cornell University, Ithaca, NY 14853-2702, USA

Paired comparison of two means – an empirical example

- In many species, males are more likely to attract females if males have high testosterone levels.

- **Research question:** Are males with high testosterone paying a cost for this extra mating success in other ways (trade-offs)?

- Males with high testosterone might be less able to fight off disease (levels of testosterone reduce their immunocompetence).

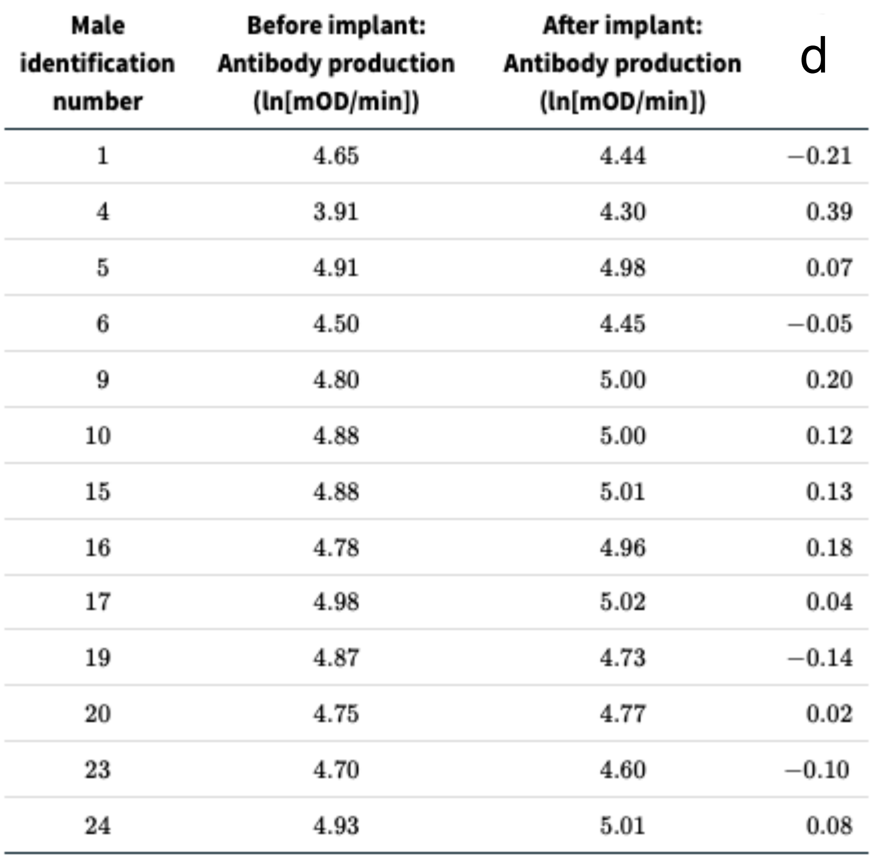
- Hasselquist et al. (1999) experimentally increased the testosterone levels of 13 male red-winged blackbirds (implant of a small tube that releases testosterone).

- Immunocompetence was measured (rate of antibody production in response to a non-pathogenic antigen in each bird’s blood serum both before and after the testosterone implant).

et al. = abbreviation of latin “et alia” = “and others”

Are males with high testosterone paying a cost for this extra mating success in other ways (trade-offs)?

Antibody production rates measure optically
ln[mOD/min] = log optical density per minute



Male identification number	Before implant: Antibody production (ln[mOD/min])	After implant: Antibody production (ln[mOD/min])	d
1	4.65	4.44	−0.21
4	3.91	4.30	0.39
5	4.91	4.98	0.07
6	4.50	4.45	−0.05
9	4.80	5.00	0.20
10	4.88	5.00	0.12
15	4.88	5.01	0.13
16	4.78	4.96	0.18
17	4.98	5.02	0.04
19	4.87	4.73	−0.14
20	4.75	4.77	0.02
23	4.70	4.60	−0.10
24	4.93	5.01	0.08

After – Before difference between treatments (positive difference more antibody production after testosterone implant).

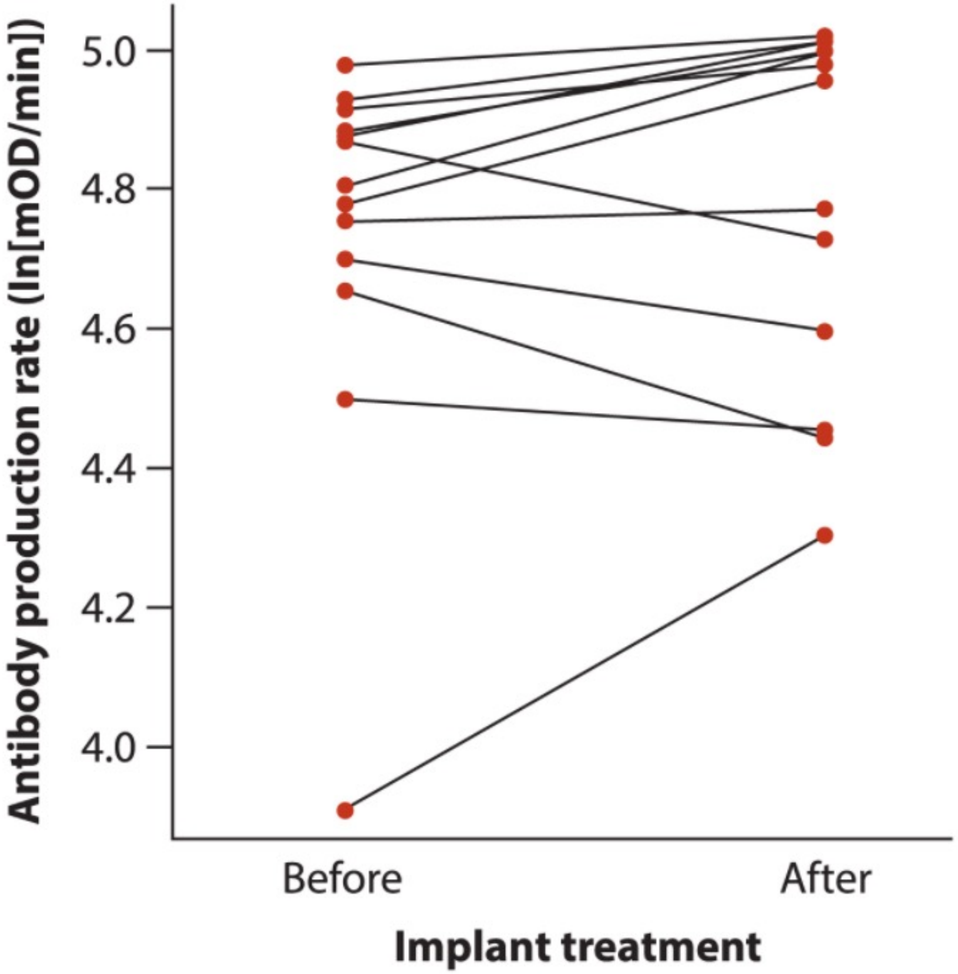
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

Are males with high testosterone paying a cost for this extra mating success in other ways (trade-offs)?

Antibody production rates measure optically
ln[mOD/min] = log optical density per minute

Male identification number	Before implant: Antibody production (ln[mOD/min])	After implant: Antibody production (ln[mOD/min])	d
1	4.65	4.44	−0.21
4	3.91	4.30	0.39
5	4.91	4.98	0.07
6	4.50	4.45	−0.05
9	4.80	5.00	0.20
10	4.88	5.00	0.12
15	4.88	5.01	0.13
16	4.78	4.96	0.18
17	4.98	5.02	0.04
19	4.87	4.73	−0.14
20	4.75	4.77	0.02
23	4.70	4.60	−0.10
24	4.93	5.01	0.08

d is the difference between treatments (positive difference more production after)



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

Are males with high testosterone paying a cost for this extra mating success in other ways (trade-offs)?

H_0: The mean change in antibody production in the population after testosterone implants is zero.

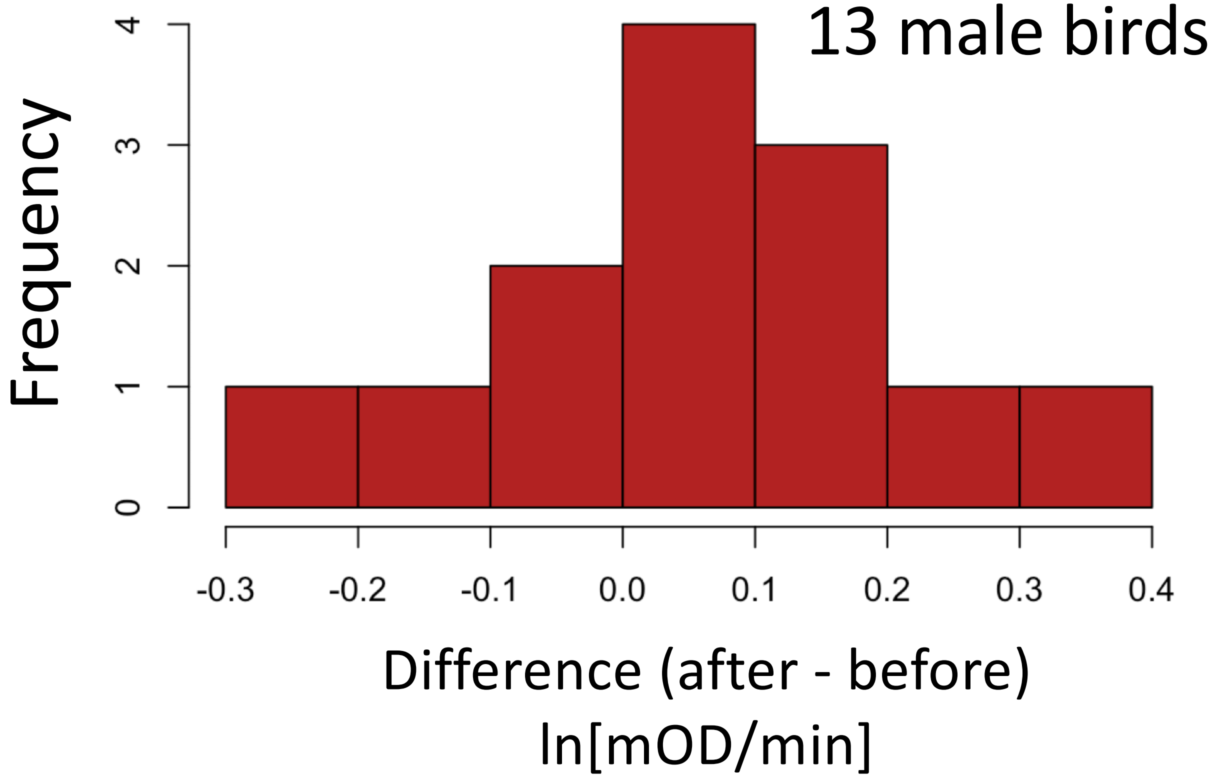
H_A: The mean change in antibody production in the population after testosterone implants is different from zero.

H_0: $\mu_d = 0$

H_A: $\mu_d \neq 0$

μ_d is the population mean difference between treatments

Are males with high testosterone paying a cost for this extra mating success in other ways (trade-offs)?



H$_0$: $\mu_d = 0$

H$_A$: $\mu_d \neq 0$

$\bar{d} = 0.056$

$s_d = 0.159$

$n = 13$

$$\mathrm{SE}_{\bar{d}} = \frac{0.159}{\sqrt{13}} = 0.044$$

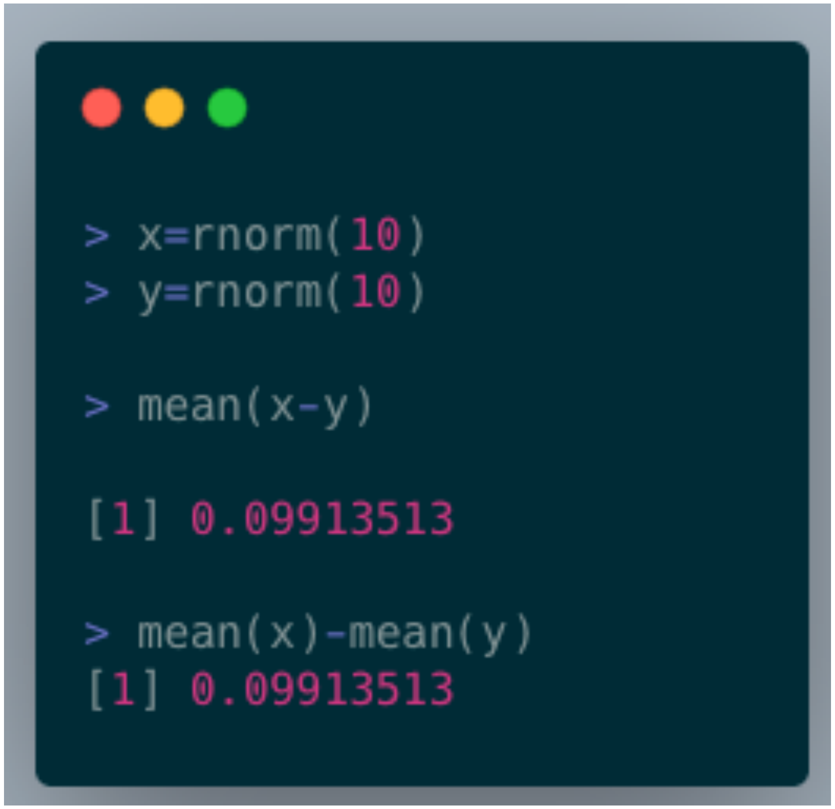
$\bar{\mathrm{d}}$ = mean difference

$\mathrm{s_d}$ = standard deviation of the difference

$\mathrm{SE}_{\bar{d}}$ = standard error of the mean difference

One important thing to note:

Differences between paired observations between two samples is equal to the differences between means (this is a property of means):

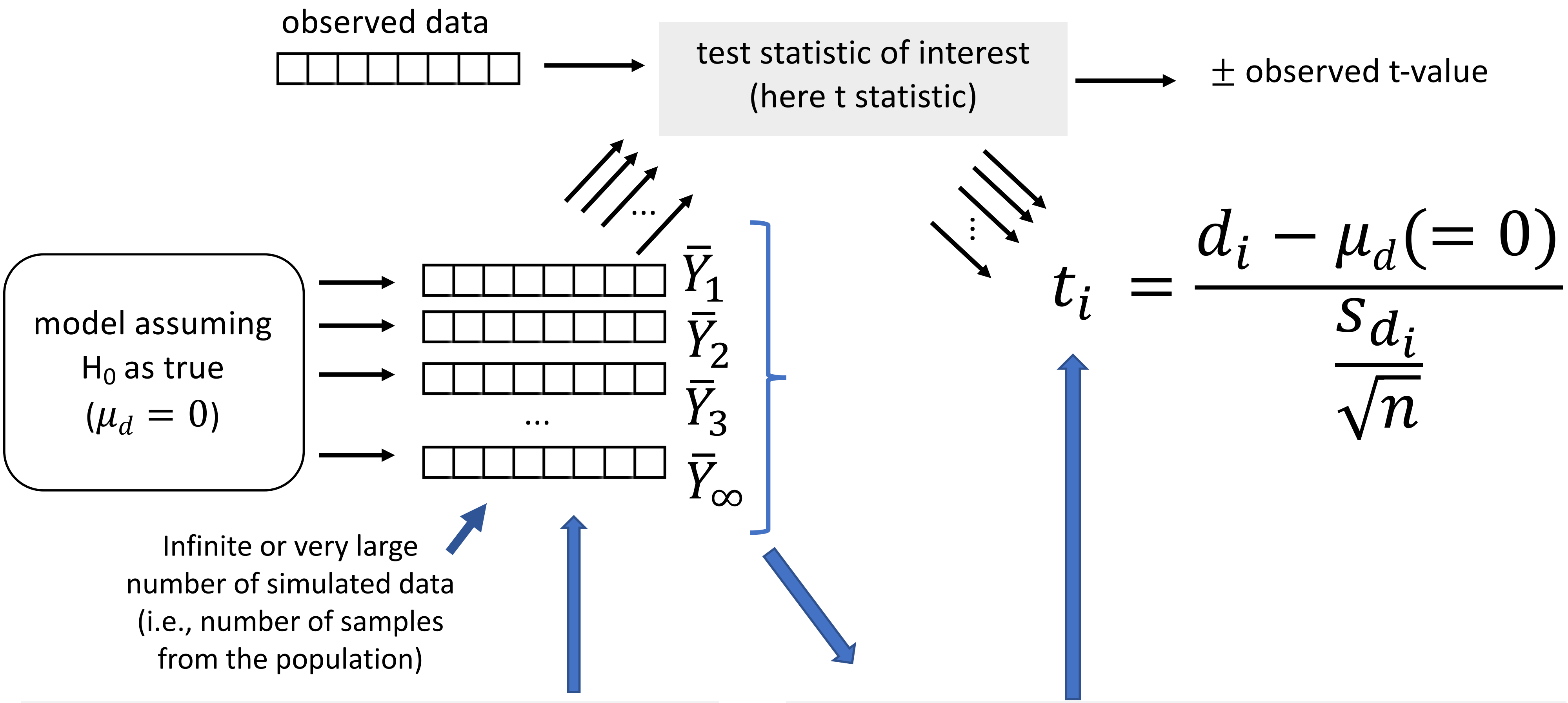


```
> x=rnorm(10)
> y=rnorm(10)

> mean(x-y)

[1] 0.09913513

> mean(x)-mean(y)
[1] 0.09913513
```



observed data

test statistic of interest
(here t statistic)

± observed t-value

model assuming
H_0 as true
($\mu_d = 0$)

$\bar{Y}_1$

$\bar{Y}_2$

$\bar{Y}_3$

…

$\bar{Y}_\infty$

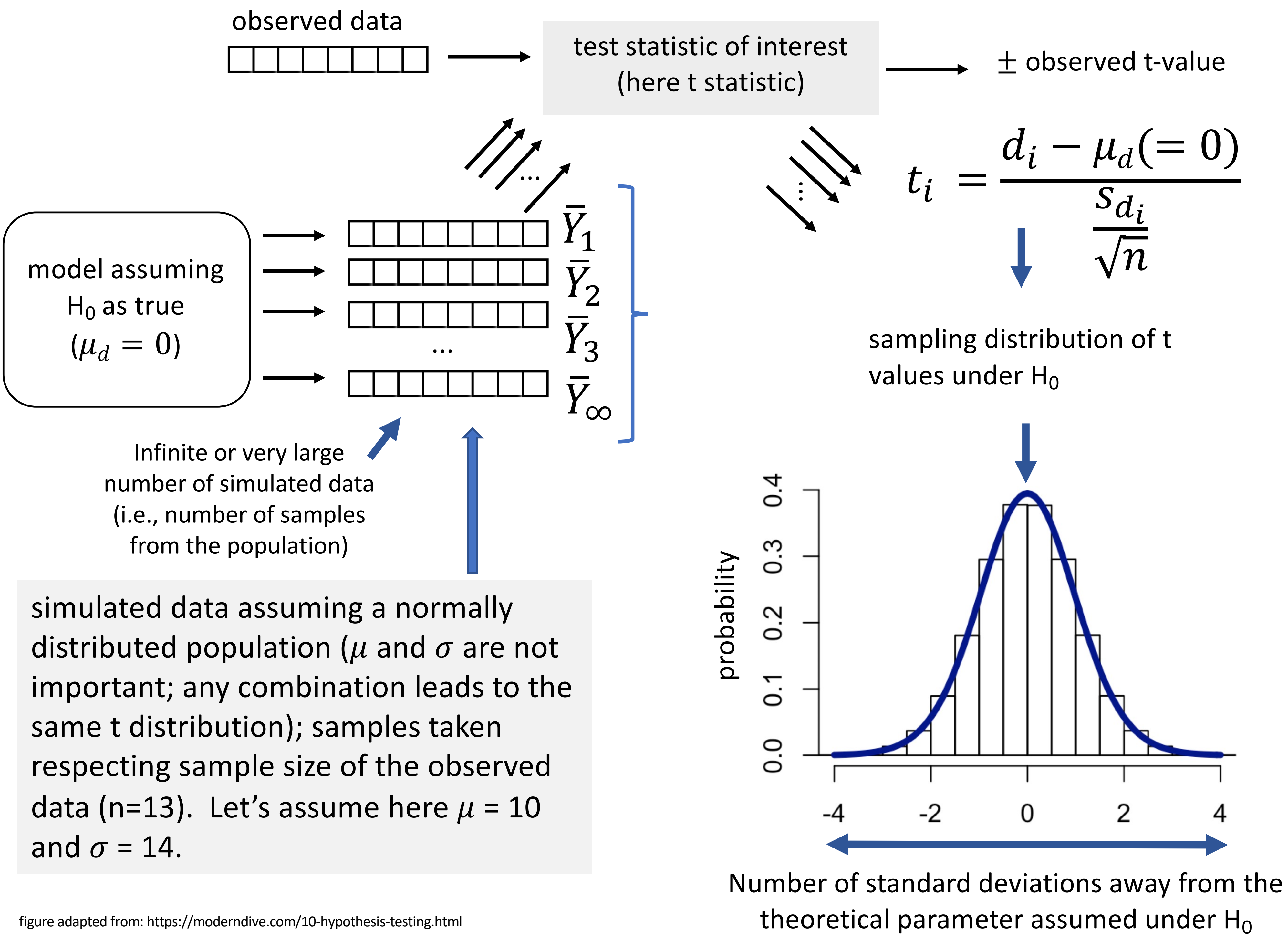
$$t_i = \frac{d_i - \mu_d(= 0)}{\frac{s_{d_i}}{\sqrt{n}}}$$

Infinite or very large
number of simulated data
(i.e., number of samples
from the population)

simulated data assuming a normally distributed population (μ and σ are not important; any combination leads to the same t distribution); samples taken respecting sample size of the observed data (n=13). Let's assume here μ = 10 and σ = 14.

Every single pair of sample means are subtracted from one another to create infinite or very large numbers of differences d_i(e.g.,
$d_1 = \bar{Y}_1$- $\bar{Y}_2$; $d_2 = \bar{Y}_1$- $\bar{Y}_3$;
$d_\infty = \bar{Y}_\infty$- $\bar{Y}_{\infty-1}$

figure adapted from: https://moderndive.com/10-hypothesis-testing.html



observed data

test statistic of interest (here t statistic)

± observed t-value

model assuming H_0 as true ($\mu_d = 0$)

$\bar{Y}_1$

$\bar{Y}_2$

$\bar{Y}_3$

...

$\bar{Y}_\infty$

$$t_i = \frac{d_i - \mu_d(= 0)}{\frac{s_{d_i}}{\sqrt{n}}}$$

Infinite or very large number of simulated data (i.e., number of samples from the population)

simulated data assuming a normally distributed population (μ and σ are not important; any combination leads to the same t distribution); samples taken respecting sample size of the observed data (n=13). Let's assume here μ = 10 and σ = 14.

sampling distribution of t values under H_0

Number of standard deviations away from the theoretical parameter assumed under H_0

figure adapted from: https://moderndive.com/10-hypothesis-testing.html

Paired comparison of two means (**paired t-test**) – an empirical example

H₀: $\mu_d = 0$

H$_A$: $\mu_d \neq 0$

$\bar{d} = 0.056$

$s_d = 0.159$

$n = 13$

$$\mathrm{SE}_{\bar{d}} = \frac{0.159}{\sqrt{13}} = 0.044$$

Degrees of freedom = 13-1=12

$$t = \frac{\bar{d} - 0\ (Ho\colon \mu_d)}{\mathrm{SE}_{\bar{d}}} = \frac{0.056 - 0}{0.044} = 1.27$$

$P = 0.23$

Decision based on alpha = 0.05:
do not reject H₀

Paired comparison of two means (**paired t-test**) – an empirical example

H$_0$: $\mu_d = 0$

H$_A$: $\mu_d \neq 0$

The standardization process in relation to the parameter assumed under H$_0$. The value for the mean of the population is 0 in this case.

For the standardized t-distribution, the parameter value under the H$_0$ is zero.

$$t = \frac{\bar{d} - 0\ (Ho\!: \mu_d)}{\mathrm{SE}_{\bar{d}}} = \frac{0.056 - 0}{0.044} = 1.27$$

To make our sample compatible with the standardized t-distribution, we subtract our value under the H$_0$ which here is the 98.6^{o}C.

Contrast with our one sample test for human body temperature

$$t = \frac{98.524 - 98.6(Ho\!: \mu_d)}{0.136} = -0.56$$

Paired comparison of two means (**paired t-test**) – an empirical example

$$P = 0.23$$

Decision based on alpha = 0.05:
do not reject H_0

H_0: The mean change in antibody production in the population after testosterone implants is zero.

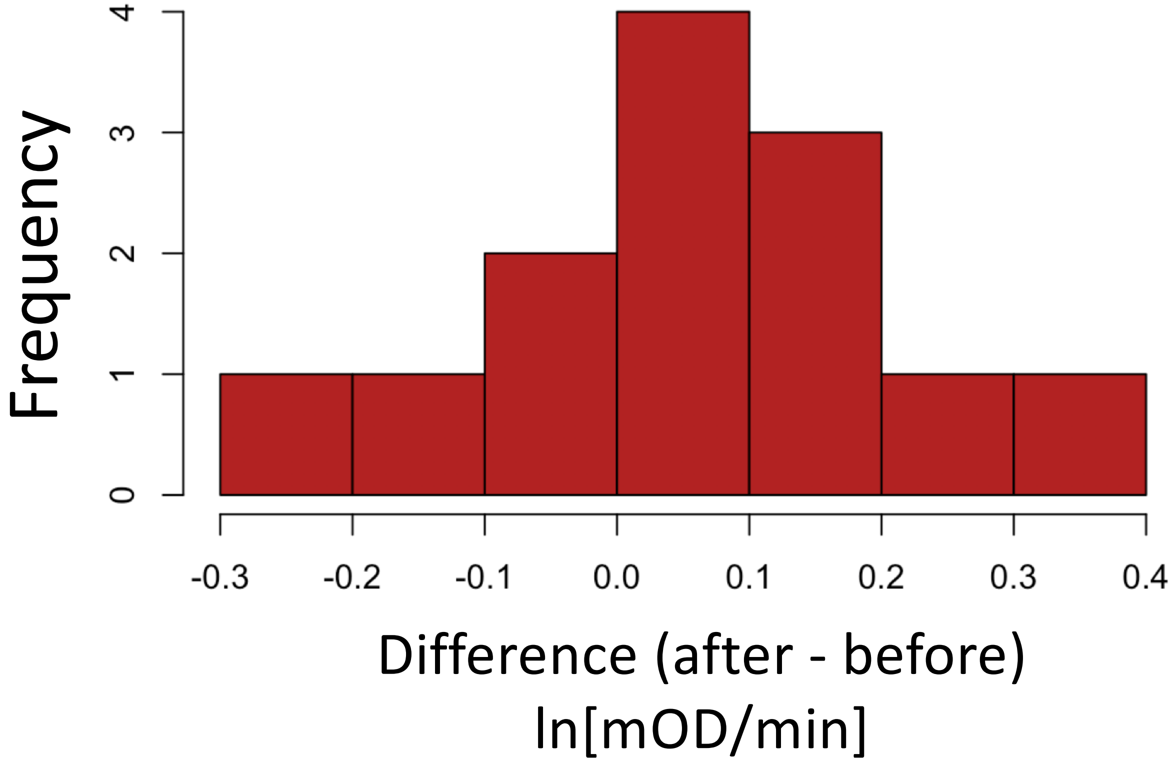
SCIENTIFIC CONCLUSION: We lack evidence that testosterone affects immunocompetence in red-winged blackbirds.

Paired comparison of two means (**paired t-test**)

Assumptions:

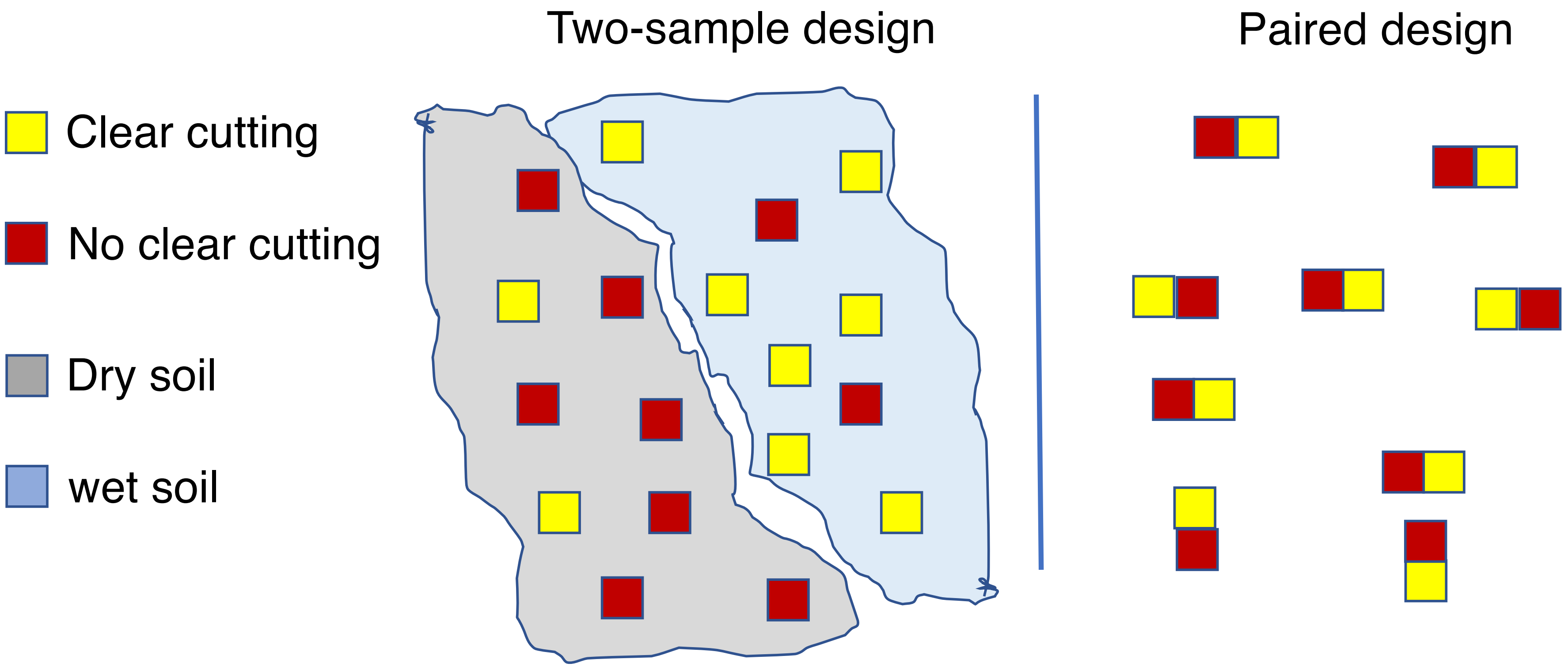
- The observational units are randomly sampled from the population.

- The paired differences have a normal distribution in the population.

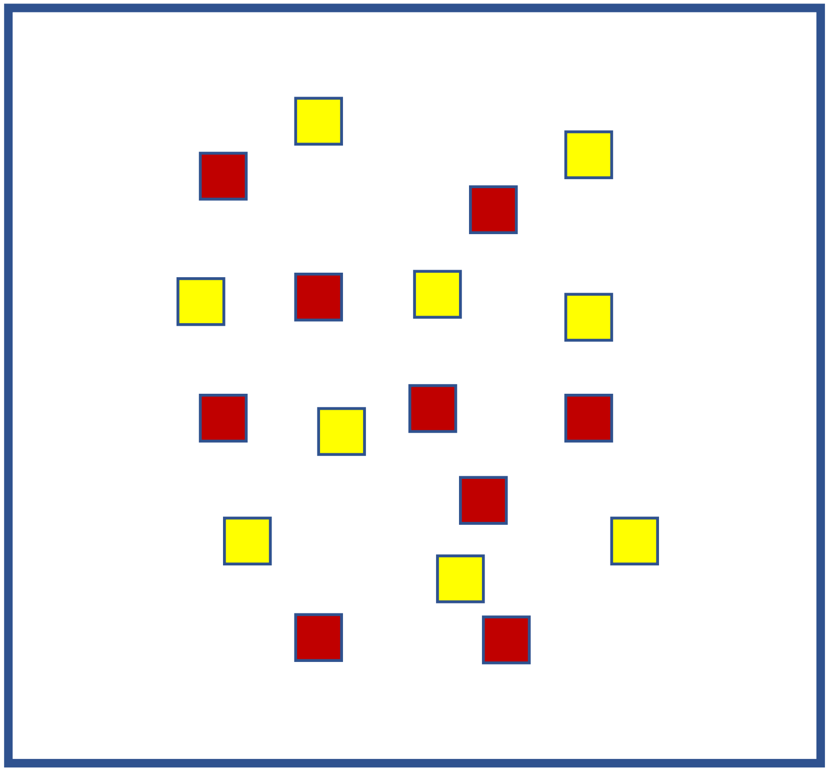


Let’s take a break - 2 minutes

Paired comparison of two means
versus
Two-sample design



Two-sample comparison of means (independent sampling)



Comparison of two independent sample means

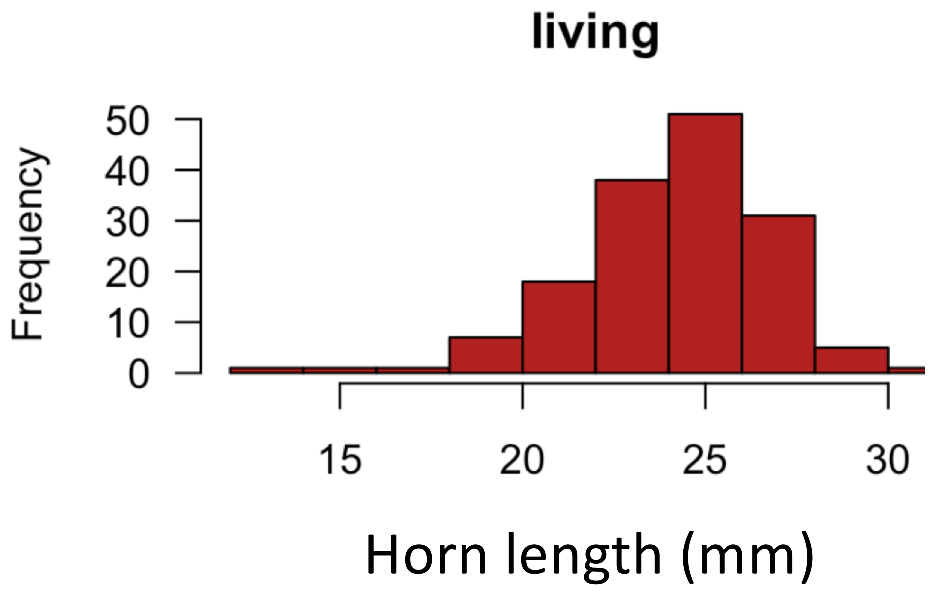
Do spikes help protect horned lizards from predation (being eaten by the loggerhead shrike)?

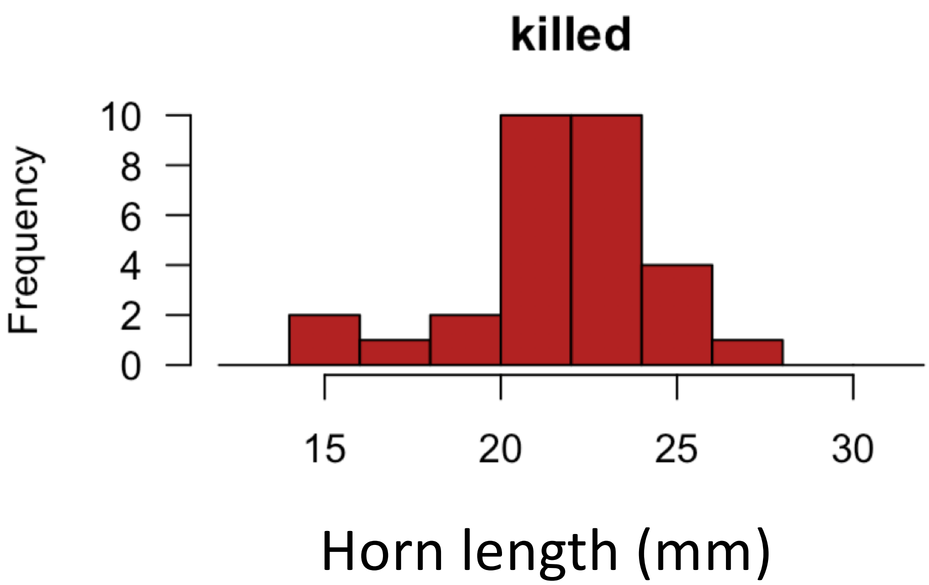
Horned lizard

Loggerhead shrike

Two-sample comparison of means
an empirical example

Lizard group	Sample mean (mm)	Sample standard deviation (mm)	Sample size *n*
Living	24.28	2.63	154
Killed	21.99	2.71	30





Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

Two-sample (means) t-test

Lizard group	**Sample mean (mm)**	**Sample standard deviation (mm)**	**Sample size *n***
Living	24.28	2.63	154
Killed	21.99	2.71	30

H_0: Lizards killed by shrikes and living lizard *do not differ* in mean horn length (i.e., $\mu_1 = \mu_2$).

H_A: Lizards killed by shrikes and living lizard *differ* in mean horn length (i.e., $\mu_1 \neq \mu_2$).

Two sample (means) t-test

Lizard group	Sample mean (mm)	Sample standard deviation (mm)	Sample size *n*
Living	24.28	2.63	154
Killed	21.99	2.71	30

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\mathrm{SE}_{\bar{Y}_1 - \bar{Y}_2}}$$

The sampling distribution of the difference between two sample means is t distributed! Aren’t we lucky?!!

Two sample (means) t-test

Lizard group	Sample mean (mm)	Sample standard deviation (mm)	Sample size *n*
Living	24.28	2.63	154
Killed	21.99	2.71	30

$$t = \frac{(24.28 - 21.99) - 0}{0.527} = \frac{2.29}{0.527} = 4.35$$

$$\mathrm{SE}_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$s_p^2 = \frac{df_1 s_1^2 + df_2 s_2^2}{df_1 + df_2}$$

$$df_1 = n_1 - 1 = 153$$

$$df_2 = n_2 - 1 = 29$$

The quantity s_p^2 is called the pooled sample variance and is the average of the sample variances weighted by their degrees of freedom.

Two sample (means) t-test

Lizard group	Sample mean (mm)	Sample standard deviation (mm)	Sample size *n*
Living	24.28	2.63	154
Killed	21.99	2.71	30

$$t = \frac{(24.28 - 21.99) - 0}{0.527} = \frac{2.29}{0.527} = 4.35$$

$$\mathrm{SE}_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{s_p^2 (\frac{1}{n_1} + \frac{1}{n_2})} = \sqrt{6.98\left(\frac{1}{154} + \frac{1}{30}\right)} = 0.527$$

$$s_p^2 = \frac{df_1 s_1^2 + df_2 s_2^2}{df_1 + df_2} = \frac{153(2.63^2) + 29(2.71^2)}{153 + 29} = 6.98$$

Two sample (means) t-test

Lizard group	Sample mean (mm)	Sample standard deviation (mm)	Sample size *n*
Living	24.28	2.63	154
Killed	21.99	2.71	30

$$t = \frac{(24.28 - 21.99) - 0}{0.527} = \frac{2.29}{0.527} = 4.35$$

$$P = 0.000023$$

Decision based on alpha = 0.05:

reject H_0

Two sample (means) t-test

$$P = 0.000023$$

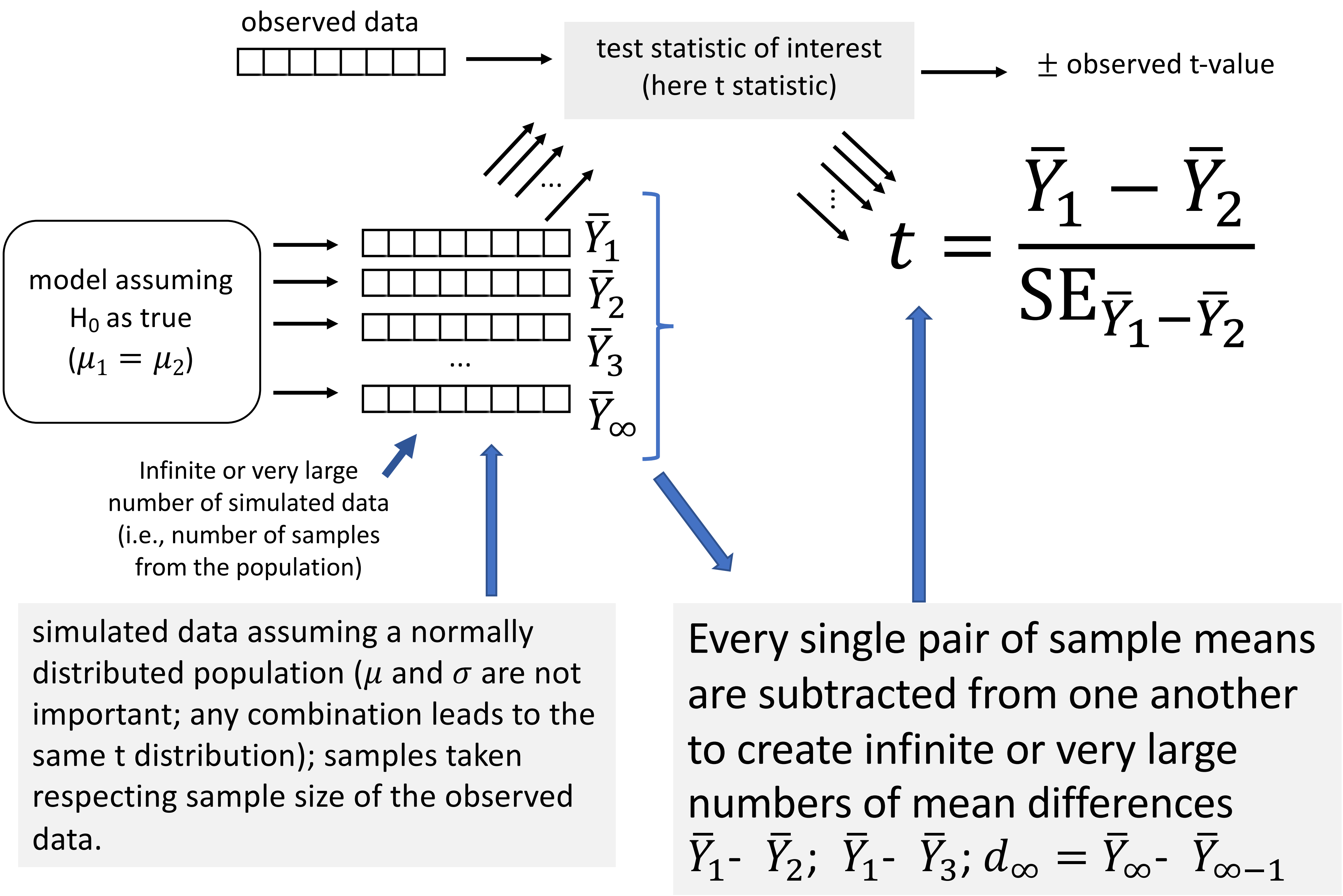
Decision based on alpha = 0.05: ***reject H_0***

H_A: Lizards killed by shrikes and living lizard *differ* in mean horn length (i.e., $\mu_1 \neq \mu_2$).

STASTISTICAL CONCLUSION: we have evidence that lizards killed by shrikes and living lizard *differ* in mean horn length.

↓

SCIENTIFIC CONCLUSION: we have evidence that horn size is a protection against predation.



observed data

test statistic of interest (here t statistic)

± observed t-value

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\mathrm{SE}_{\bar{Y}_1 - \bar{Y}_2}}$$

model assuming H_0 as true ($\mu_1 = \mu_2$)

$\bar{Y}_1$

$\bar{Y}_2$

$\bar{Y}_3$

…

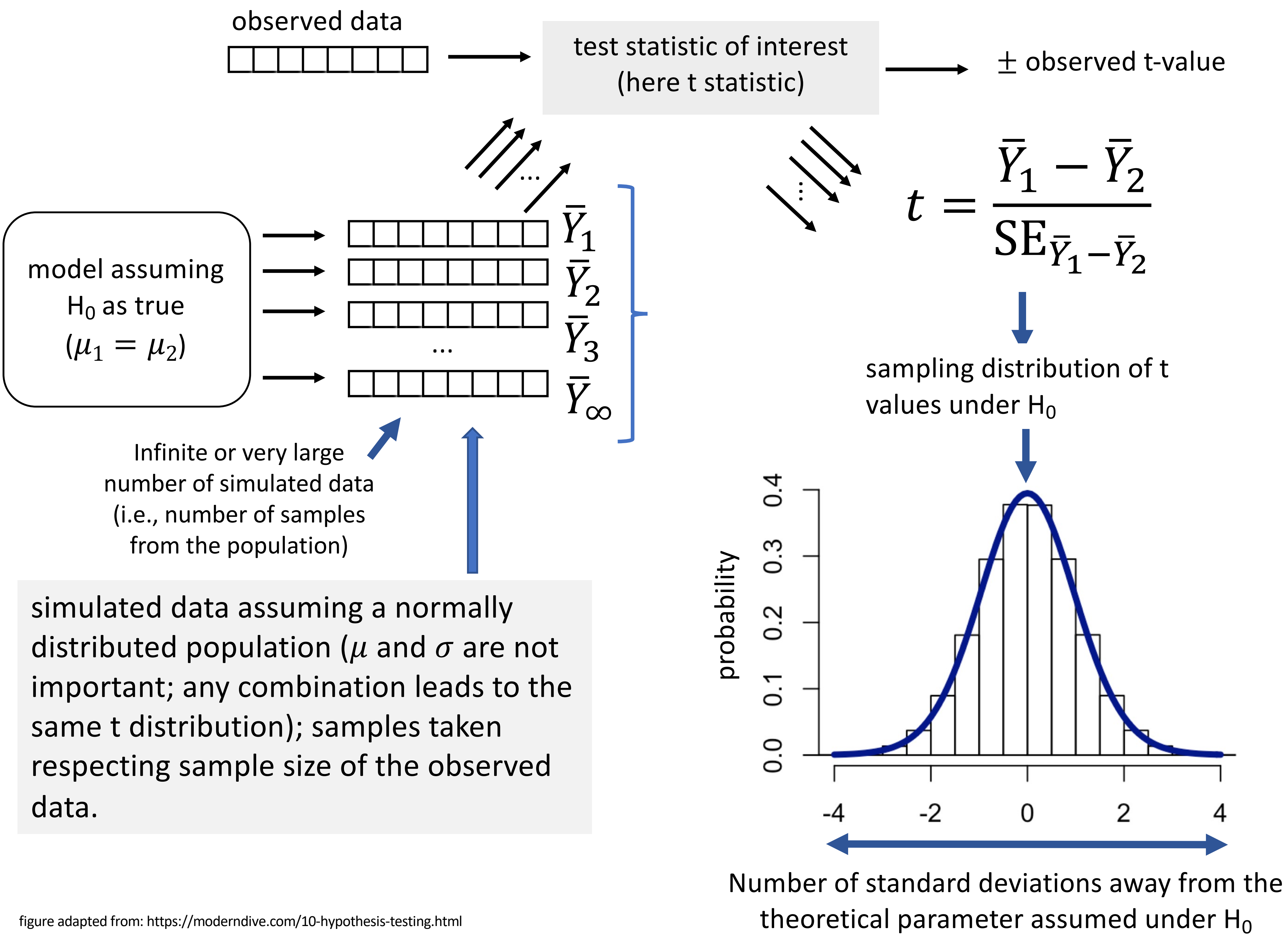
$\bar{Y}_\infty$

Infinite or very large number of simulated data (i.e., number of samples from the population)

simulated data assuming a normally distributed population (μ and σ are not important; any combination leads to the same t distribution); samples taken respecting sample size of the observed data.

Every single pair of sample means are subtracted from one another to create infinite or very large numbers of mean differences $\bar{Y}_1$- $\bar{Y}_2$; $\bar{Y}_1$- $\bar{Y}_3$; $d_\infty = \bar{Y}_\infty$- $\bar{Y}_{\infty-1}$

figure adapted from: https://moderndive.com/10-hypothesis-testing.html



observed data

test statistic of interest (here t statistic)

± observed t-value

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\mathrm{SE}_{\bar{Y}_1 - \bar{Y}_2}}$$

model assuming H_0 as true ($\mu_1 = \mu_2$)

$\bar{Y}_1$

$\bar{Y}_2$

$\bar{Y}_3$

…

$\bar{Y}_\infty$

Infinite or very large number of simulated data (i.e., number of samples from the population)

simulated data assuming a normally distributed population (μ and σ are not important; any combination leads to the same t distribution); samples taken respecting sample size of the observed data.

sampling distribution of t values under H_0

Number of standard deviations away from the theoretical parameter assumed under H_0

figure adapted from: https://moderndive.com/10-hypothesis-testing.html

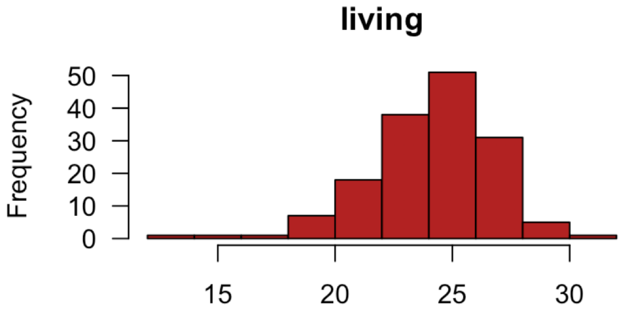
Two sample (means) t-test

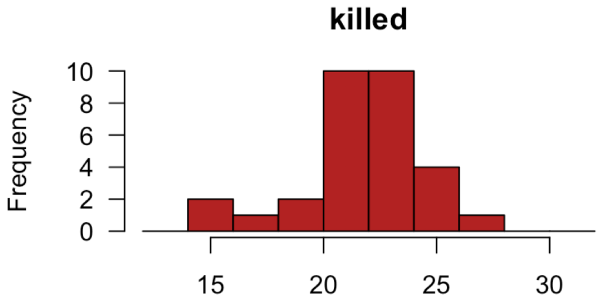
Assumptions (very important)**:**

- Each of the two samples is a random sample from its population.

- The variable (e.g., horn length) is normally distributed in each population.

- The standard deviation (and variance) of the variable is the same in both populations (we will assume this for now but see later on how to test it).


living
Frequency


killed
Frequency