

Statistical hypothesis testing involve:

How the research hypothesis should be transformed into a statistical question.

State the null (parameter for the theoretical population) and alternative hypotheses.

Compute the observed value for a particular metric of interest (i.e., based on the sample data, i.e., observed summary statistic). This is called *test statistic*. In our toad example it was simply the number of right-handed individuals.

Computer the P-value by contrasting the sample (observed) value against a sampling distribution that assumes the null hypothesis to be true (around the parameter of interest for a theoretical population).

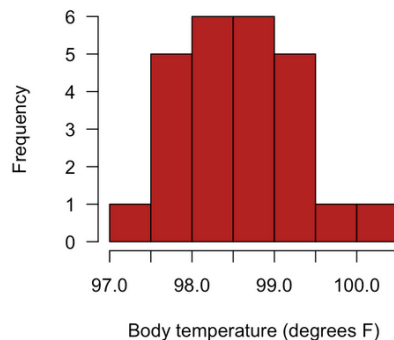
Draw a conclusion by contrasting the p-value against the significance level (α). If the p-value is greater than α , then do not reject H_0 ; if P-value is smaller or equal than α , then reject H_0 .

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.



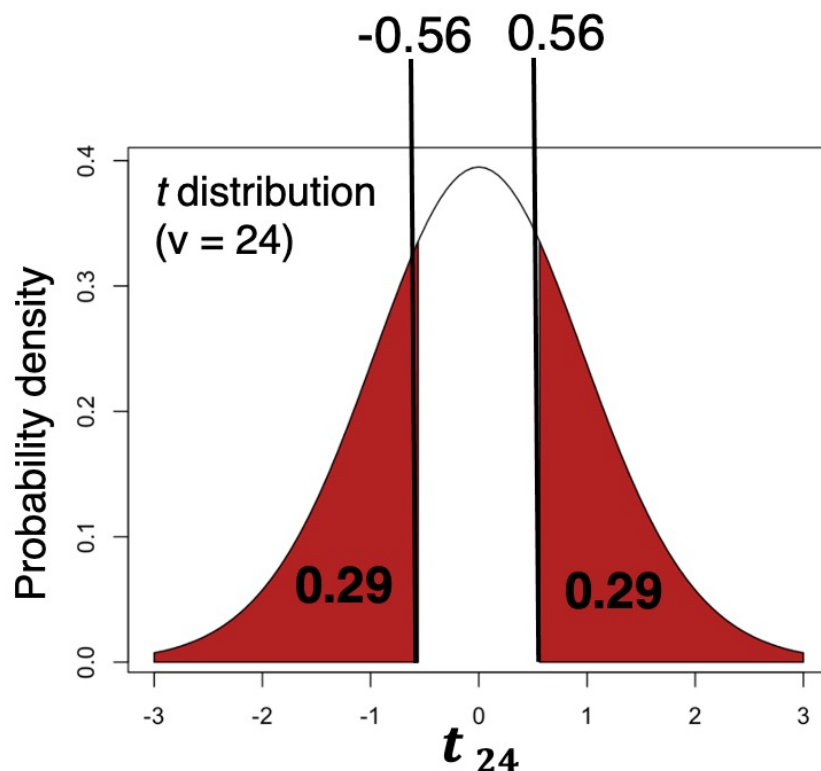
$$\bar{Y} = 98.524$$

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

Then “translated” the above question into: What is the probability of obtaining a *sample mean* as extreme or more extreme (i.e., smaller) than 98.524°F given that the true *population mean* is 98.6°F?

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

$$\begin{aligned} \Pr[t < -0.56] + \Pr[t > 0.56] &= \\ 2 \Pr[t > \text{abs}(0.56)] &= \mathbf{0.58} \\ (t \text{ is symmetric around } \mu) \end{aligned}$$



Failing to reject H_0 does not confirm that the true population mean is 98.6°F; it simply indicates that we lack sufficient evidence to conclude otherwise.

However, new evidence could emerge in the future that challenges and overturns the original conclusion. **How might this happen?**

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



The effects of larger sample sizes 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:

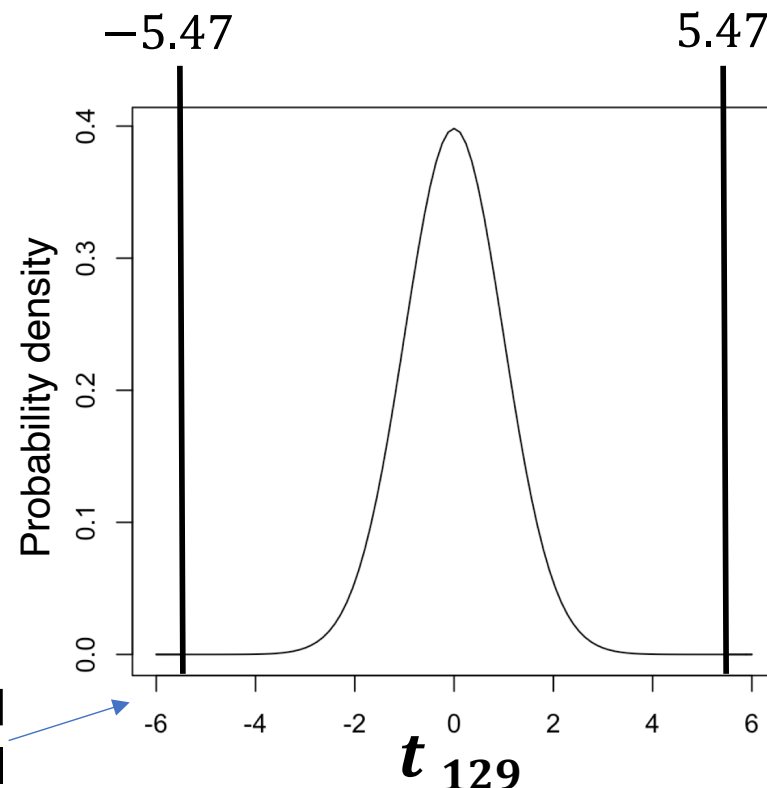
$$\bar{Y} = 98.25^{\circ}\text{F}$$

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

$$SE_{\bar{Y}} = \frac{0.733}{\sqrt{130}} = 0.064$$

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

$$\begin{aligned} \Pr[t < -5.47] + \Pr[t > 5.47] &= \\ 2 \Pr[t > \text{abs}(5.47)] &= \\ 0.000002 \end{aligned}$$



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

$$n = 25$$

$$\bar{Y} = 98.524^{\circ}\text{F}$$

$$s = 0.678^{\circ}\text{F}$$

$$SE_{\bar{Y}} \frac{0.678}{\sqrt{25}} = 0.136$$

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

$$\begin{aligned} \Pr[t < -0.56] + \Pr[t > 0.56] &= \\ 2 \Pr[t > \text{abs}(0.56)] &= \\ 0.58 \end{aligned}$$

$$n = 130$$

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

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

$$SE_{\bar{Y}} \frac{0.733}{\sqrt{130}} = 0.064$$

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

$$\begin{aligned} \Pr[t < -5.47] + \Pr[t > 5.47] &= \\ 2 \Pr[t > \text{abs}(5.55)] &= \\ 0.000002 \end{aligned}$$

The impact of larger sample sizes on hypothesis testing: Revisiting body temperature in light 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 LED TO A P-VALUE = 0.0000002 ($P < \alpha = 0.05$), SO WE REJECT THE NULL HYPOTHESIS IN LINE OF THIS NEW EVIDENCE.

Therefore, with new and stronger evidence from a larger sample size, we can confidently suggest that the true average human body temperature is likely different from 98.6°F - though this does not conclusively rule out the possibility that it could still be 98.6°F.

In this case, we can state that the increased number of observations likely increased the statistical power.

We can't say for sure because we do not know the true state of nature; that is, whether the null hypothesis is actually false.

Statistical power is defined relative to a true effect, and since the truth is unknown in real data, we cannot directly observe or confirm power; we can only infer that, if a real effect exists, a larger sample size would make it more likely to detect it.

The impact of larger sample sizes on hypothesis testing: Revisiting body temperature in light 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 (n) can lead to greater **statistical power** (smaller Type II errors) to reject the null hypothesis when it is not true!

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

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°F.

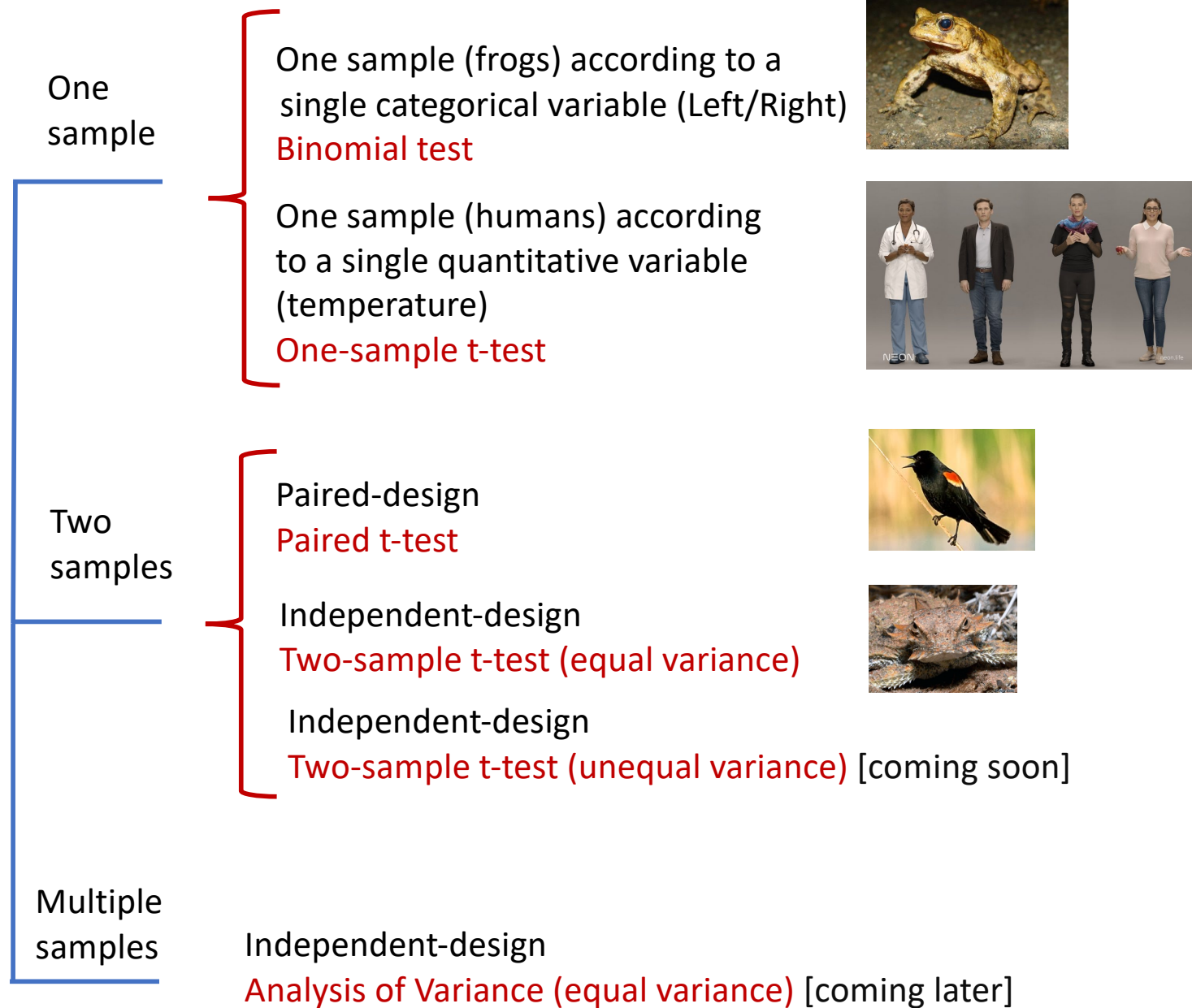
H_A (alternative hypothesis): the true population is different from 98.6°F.

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

- 1) The data are assumed to represent a random and independent sample from the population. This assumption underpins all statistical tests covered in this course and ensures that the sample provides an unbiased and representative view of the population of interest.
- 2) In addition, the variable of interest (e.g., human body temperature) is assumed to follow an approximately normal distribution in the population. This assumption is particularly important for small sample sizes, as it justifies the use of the t-distribution. For larger samples, the test is more robust to deviations from normality

Statistical hypothesis testing
for comparing two samples
based on a quantitative
variable.

One- and two-sample hypothesis testing



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?

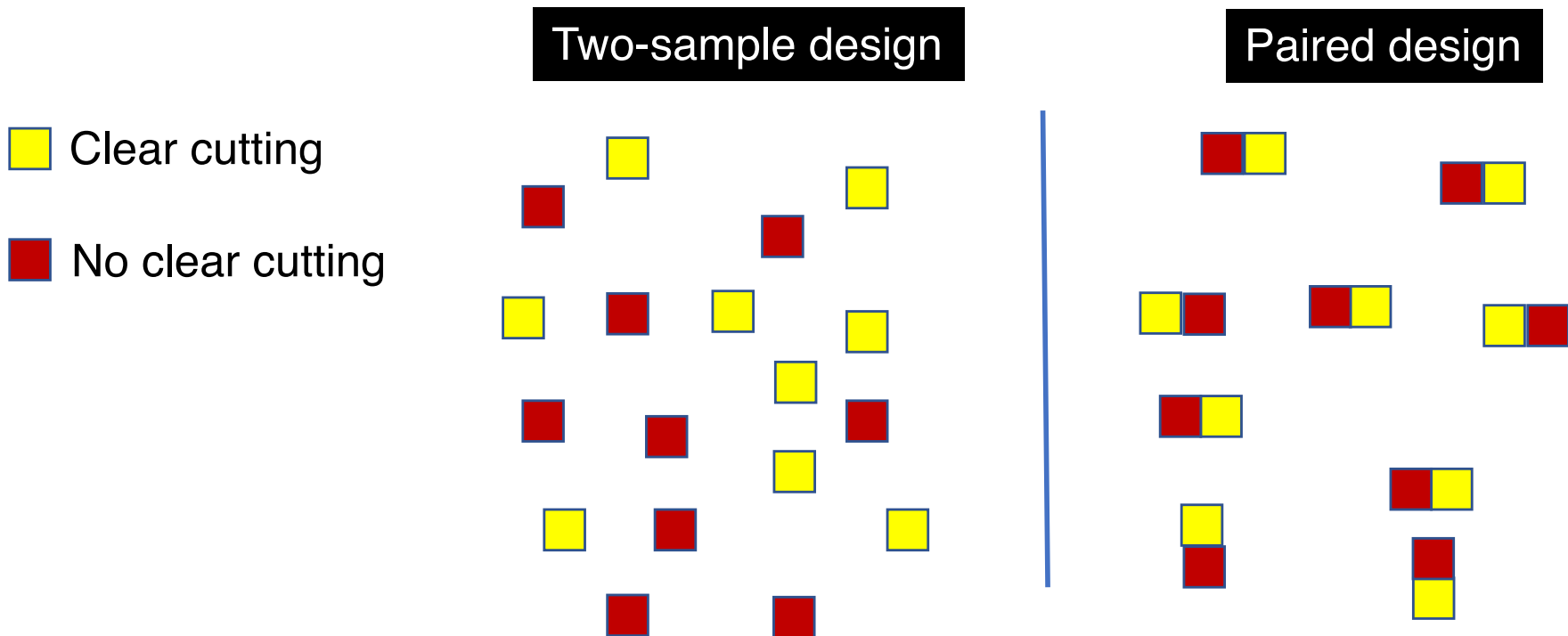
- There are two treatments: **clear cutting** / **no clear-cutting** (control).
- Statistical question: Does the mean number of salamanders differ between the two treatments?
- Treatment is a *categorical variable* and number of salamanders is a *numerical variable*.

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).

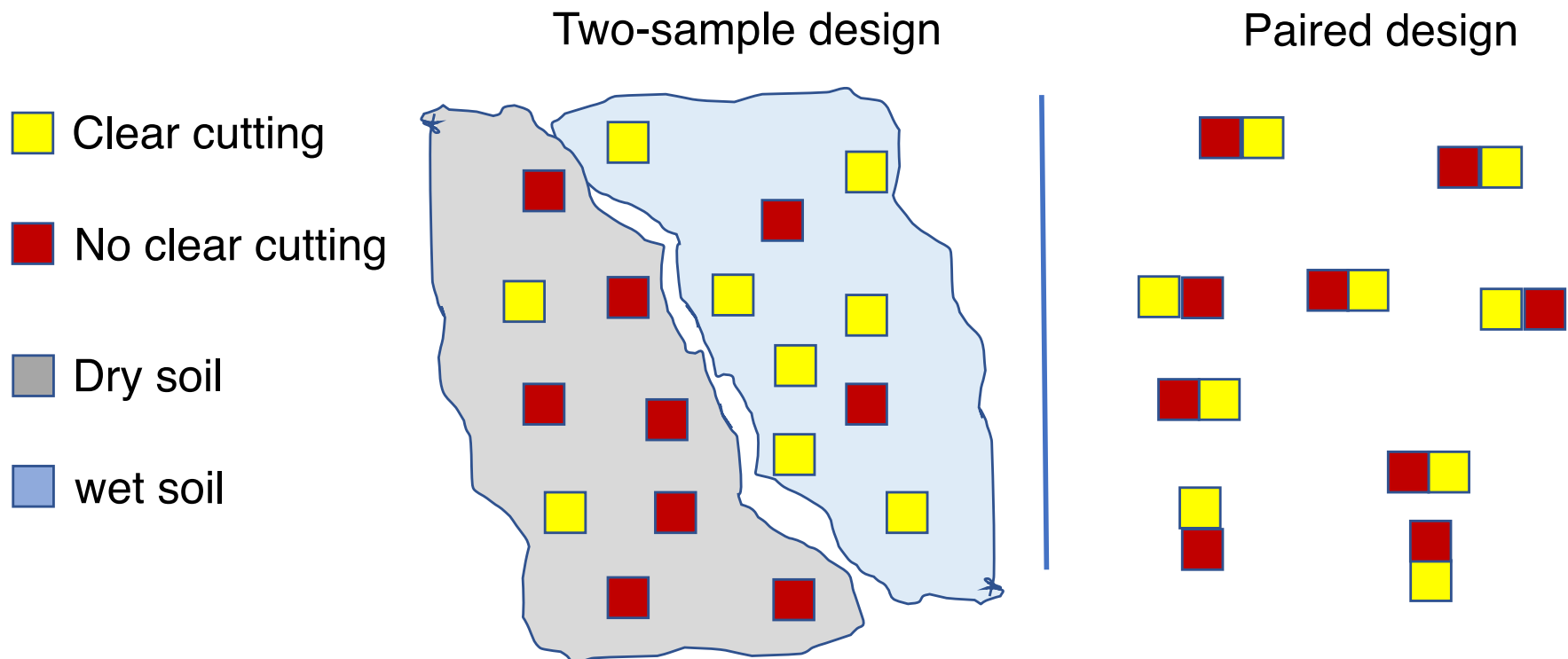


Paired design for
comparing two sample
means

Paired comparison of two means

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

The advantage of a paired design is that it minimizes the impact of variability among sampling units that is unrelated to the treatment, thereby increasing the precision of the results (e.g., local environmental differences among observational units). It reduces confounder variables.

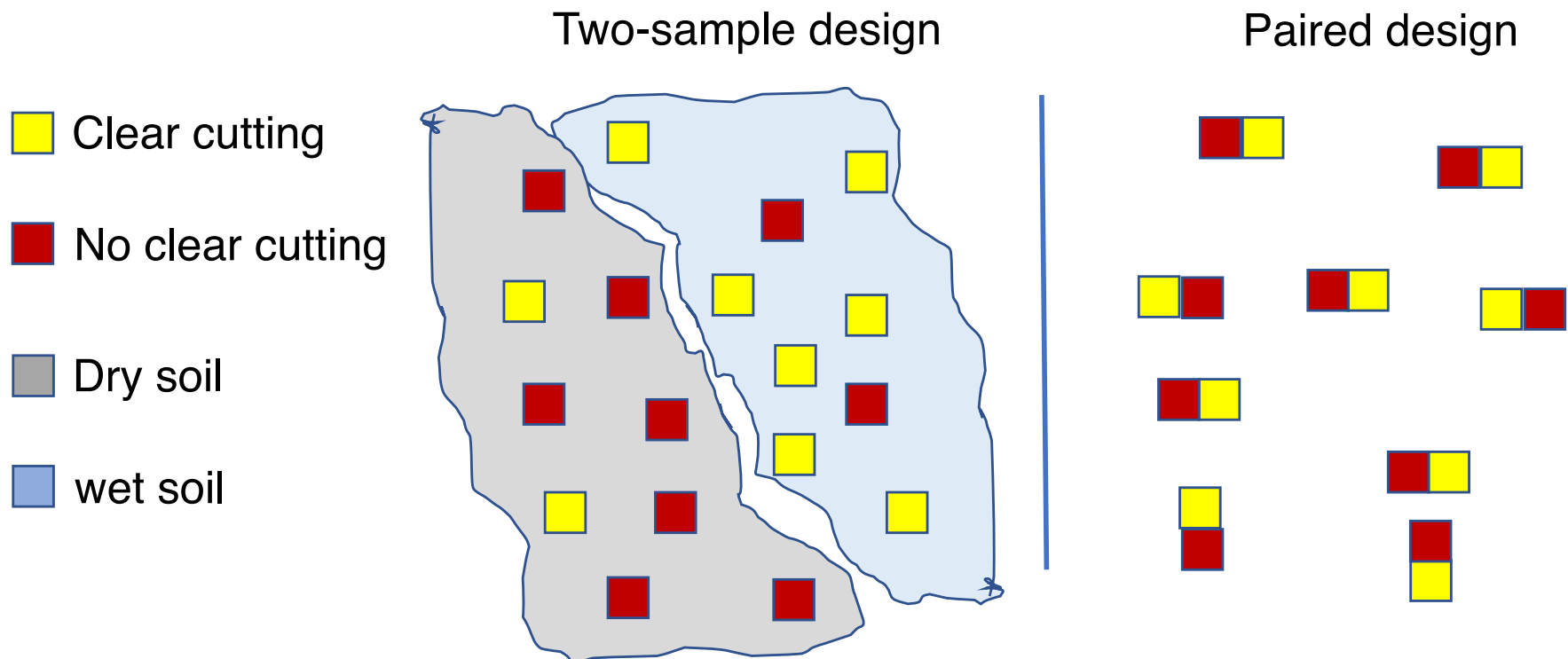


Paired comparison of two means

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

The advantage of a paired design is that it minimizes the impact of variability among sampling units that is unrelated to the treatment, thereby increasing the precision of the results (e.g., local environmental differences among observational units).

Notice that clear-cutting occurred more frequently in wet soils, whereas areas without clear-cutting were predominantly dry. If soil moisture plays a critical role for salamanders, this non-random distribution of sampling units could bias the results and affect the conclusions.



Paired comparison of two means

The advantage of a paired design is that it minimizes the impact of variability among sampling units that is unrelated to the treatment, thereby increasing the precision of the results (e.g., local environmental differences among observational units). It reduces confounder variables.

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 a 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

Spider	Speed before	Speed after
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)? OR

Is avian humoral immunocompetence (i.e., ability of the immune system to produce antibodies to defend against pathogens) suppressed by testosterone?

Behav Ecol Sociobiol (1999) 45: 167–175

© Springer-Verlag 1999

ORIGINAL ARTICLE

Dennis Hasselquist · James A. Marsh
Paul W. Sherman · John C. Wingfield

Is avian humoral immunocompetence suppressed by testosterone?

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 extra mating success in other ways (trade-offs)?

Antibody production rates measure optically
 $\ln[\text{mOD}/\text{min}] = \log \text{ optical density per minute}$

Male identification number	Before implant: Antibody production ($\ln[\text{mOD}/\text{min}]$)	After implant: Antibody production ($\ln[\text{mOD}/\text{min}]$)	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).

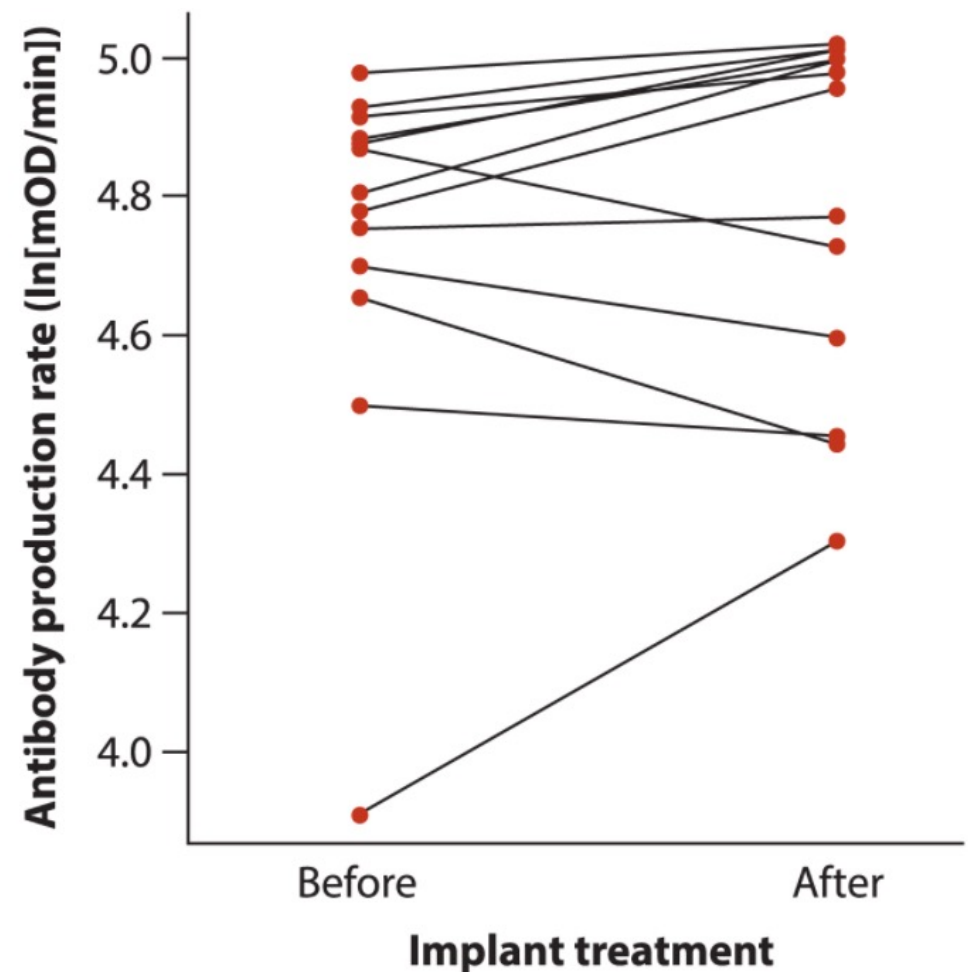


Are males with high testosterone paying a cost for extra mating success in other ways (trade-offs)?

Antibody production rates measure optically
 $\ln[\text{mOD}/\text{min}] = \log \text{ optical density per minute}$

Male identification number	Before implant: Antibody production ($\ln[\text{mOD}/\text{min}]$)	After implant: Antibody production ($\ln[\text{mOD}/\text{min}]$)	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 implant)



Are males with high testosterone paying a cost for 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.

$$H_0: \mu_d = 0$$

μ_d is the population mean difference between treatments

$$H_A: \mu_d \neq 0$$



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

$$H_0: \mu_d = 0$$

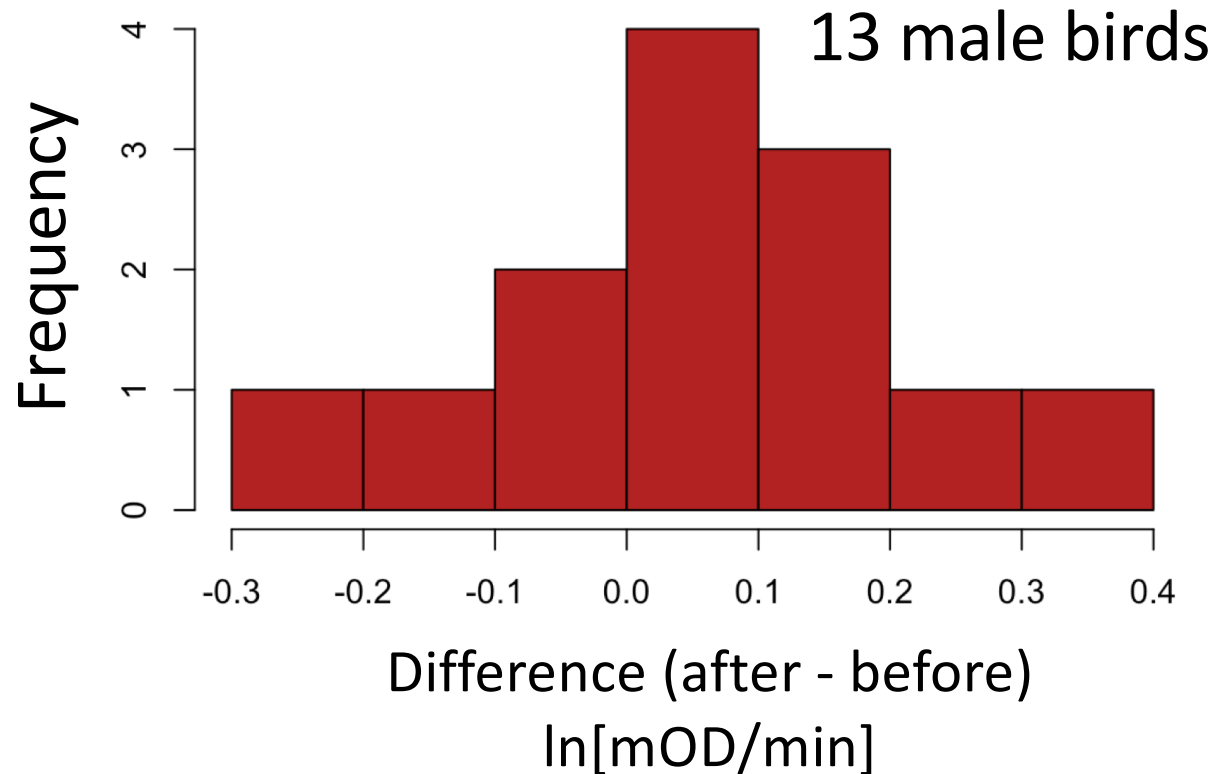
$$H_A: \mu_d \neq 0$$

$$\bar{d} = 0.056$$

$$s_d = 0.159$$

$$n = 13$$

$$SE_{\bar{d}} = \frac{0.159}{\sqrt{13}} = 0.044$$



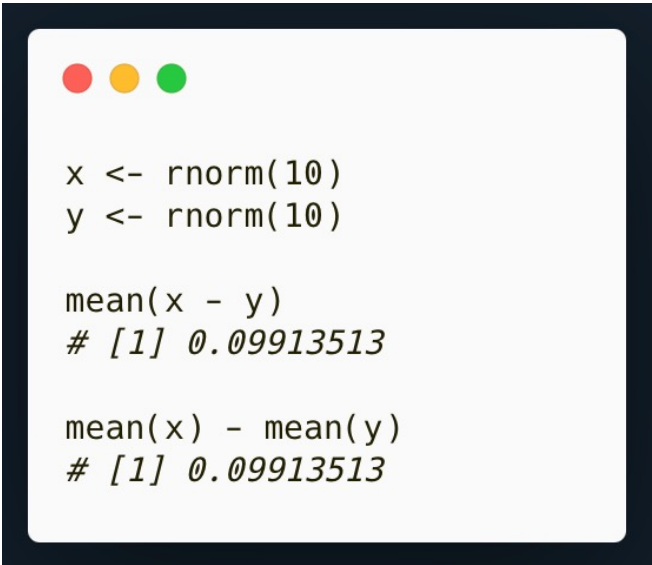
$\bar{d}$ = mean difference

s_d = standard deviation
of the difference

$SE_{\bar{d}}$ = 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)

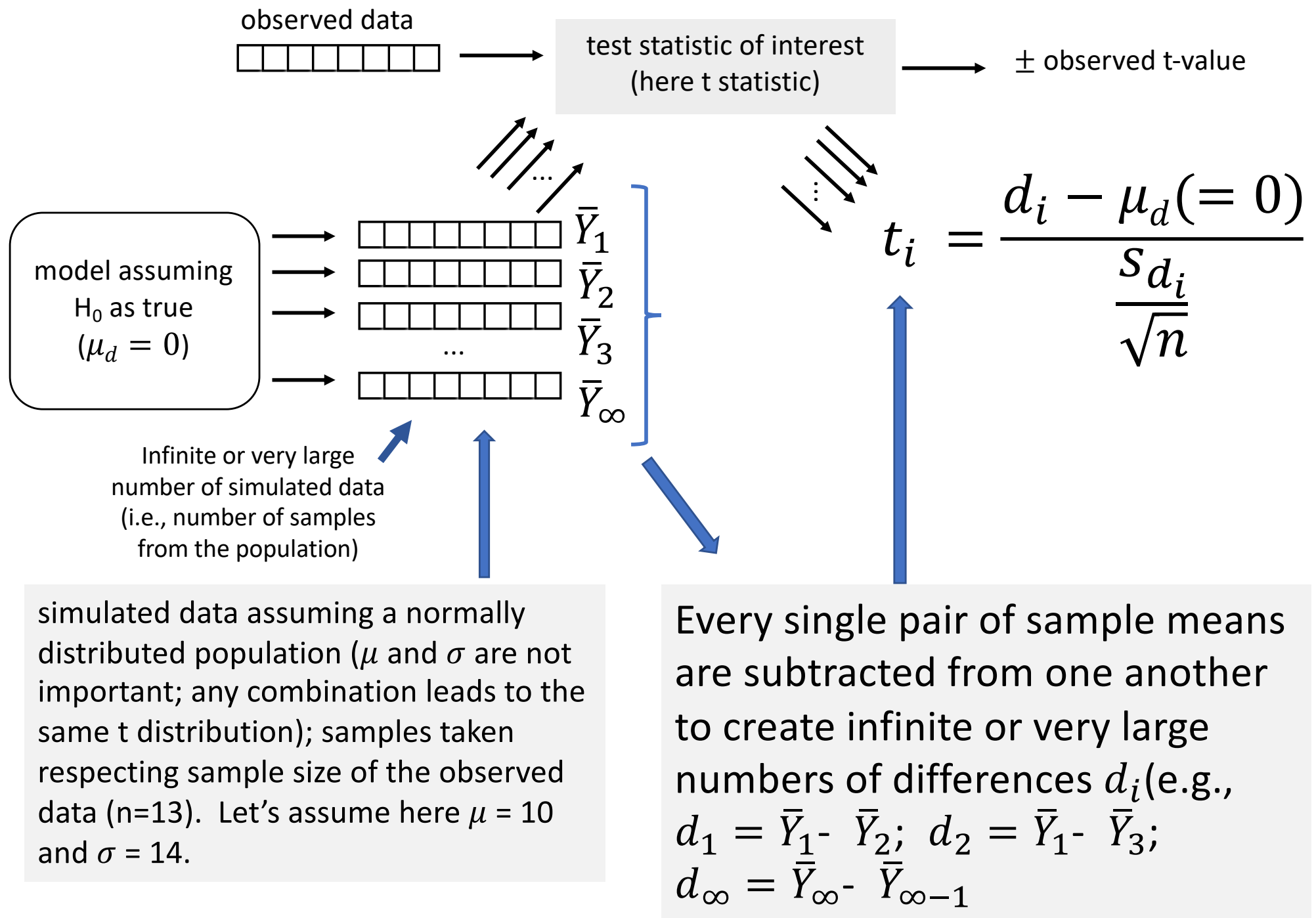
mean(x - y)
# [1] 0.09913513

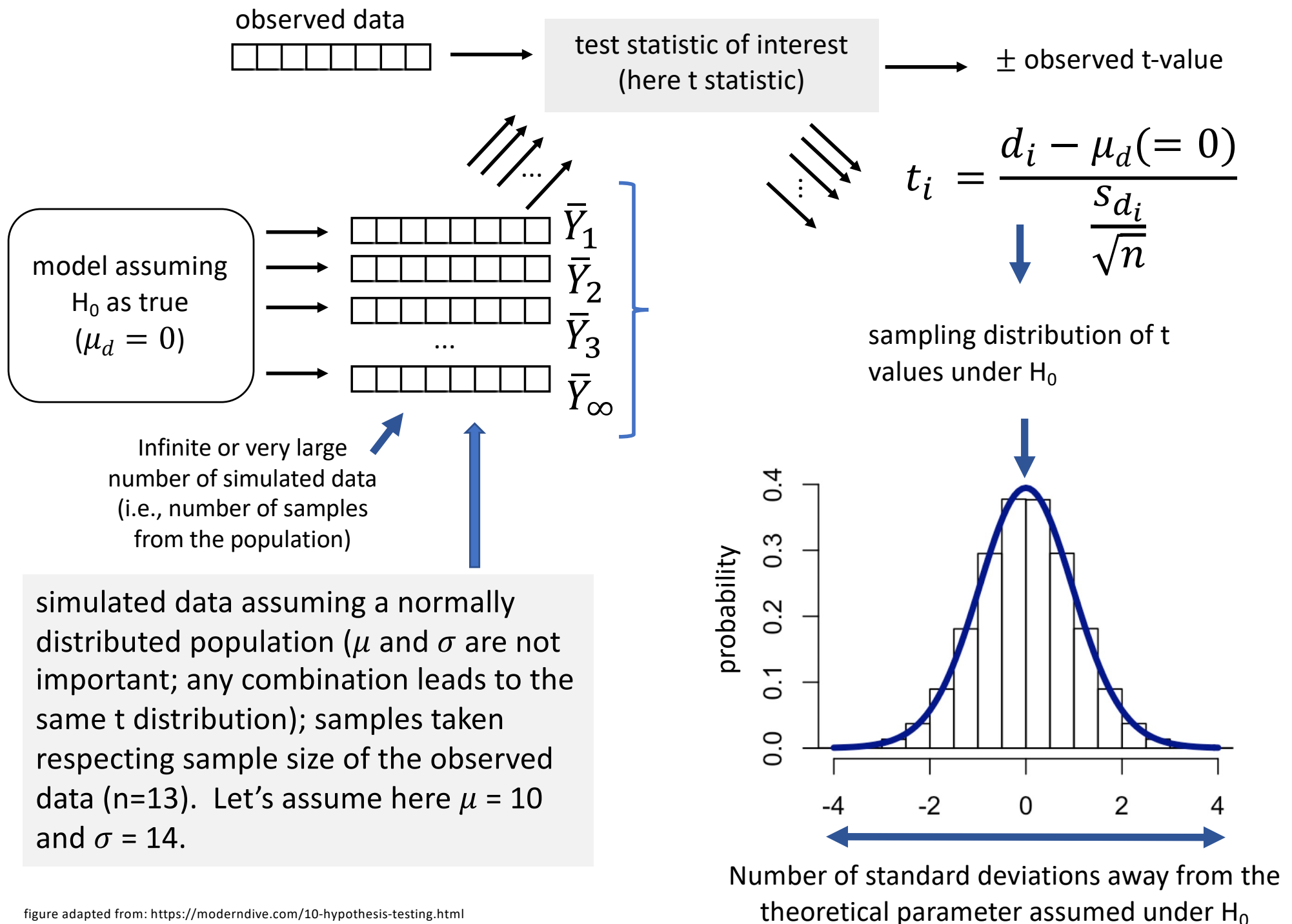
mean(x) - mean(y)
# [1] 0.09913513
```

$$(\overline{X - Y}) = \bar{X} - \bar{Y}$$

This property allows us to analyze within-pair differences directly using a one-sample t-test, which properly accounts for the pairing and reduces variability.

Importantly, if we instead treated the data as two independent samples, we would lose this pairing information and use fewer degrees of freedom (df), resulting in a less powerful test.





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

$$\begin{array}{ll} H_0: \mu_d = 0 & \bar{d} = 0.056 \\ & s_d = 0.159 \\ H_A: \mu_d \neq 0 & n = 13 \\ & SE_{\bar{d}} = \frac{0.159}{\sqrt{13}} = 0.044 \end{array}$$

**Degrees of
freedom = 13-1=12**

$$t = \frac{\bar{d} - 0 \text{ (} H_0: \mu_d \text{)}}{SE_{\bar{d}}} = \frac{0.056 - 0}{0.044} = 1.27$$

$$P = 0.23$$

Decision based on alpha = 0.05:
do not reject H_0

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.


$$t = \frac{\bar{d} - 0 (H_0: \mu_d)}{SE_{\bar{d}}} = \frac{0.056 - 0}{0.044} = 1.27$$

To make our sample compatible with the standardized t-distribution, we subtract our value under the H_0 which here is the 98.6°C.

Contrast with the one sample test for human body temperature


$$t = \frac{98.524 - 98.6(H_0: \mu_d)}{0.136} = -0.56$$

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.

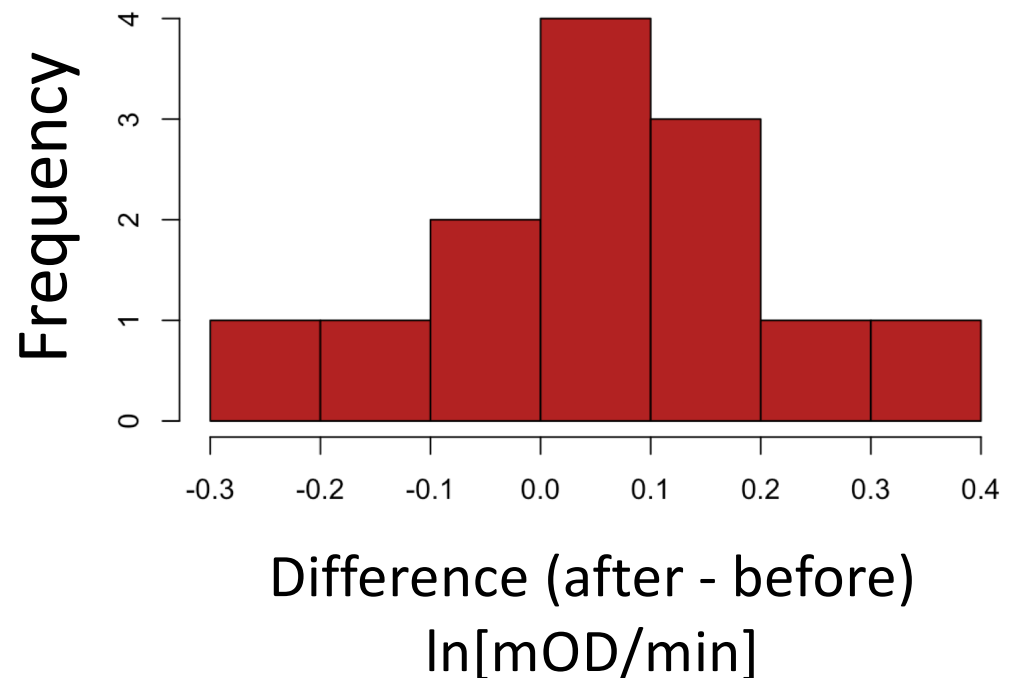
Paired comparison of two means (**paired t-test**)

Assumptions:

The observational units are randomly and independently sampled from the population.

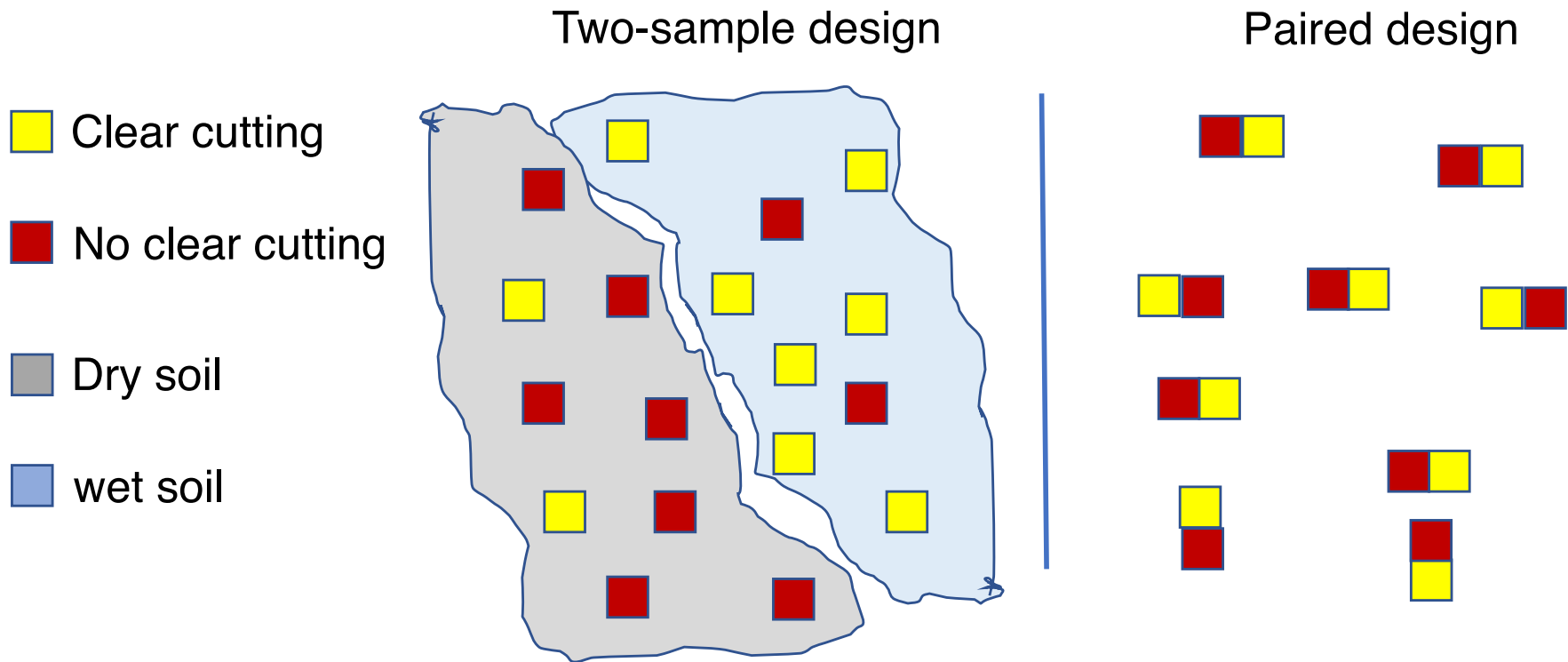
The observations are paired (i.e., measurements are meaningfully linked within units).

The differences within pairs are approximately normally distributed in the population (especially important for small sample sizes).





Paired comparison of two means versus Two-sample design



Two-sample comparison of means (independent sampling)

Comparison of two independent sample means

Do the spikes of horned lizards provide protection against predation from loggerhead shrikes?



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



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

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

→ We include $(\mu_1 - \mu_2)$ because the t-statistic measures how far the observed difference is from what is expected under the null hypothesis.



Most of the time, we test $H_0: \mu_1 - \mu_2 = 0$, so the formula simplifies to:

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - 0}{SE_{\bar{Y}_1 - \bar{Y}_2}}$$

The sampling distribution of the difference between two sample means is also 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

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

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

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

$$df_1 = n_1 - 1 = 153$$

$$df_2 = n_2 - 1 = 29$$

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 (related to sample sizes).

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

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

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

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

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

$$P = 0.000023$$

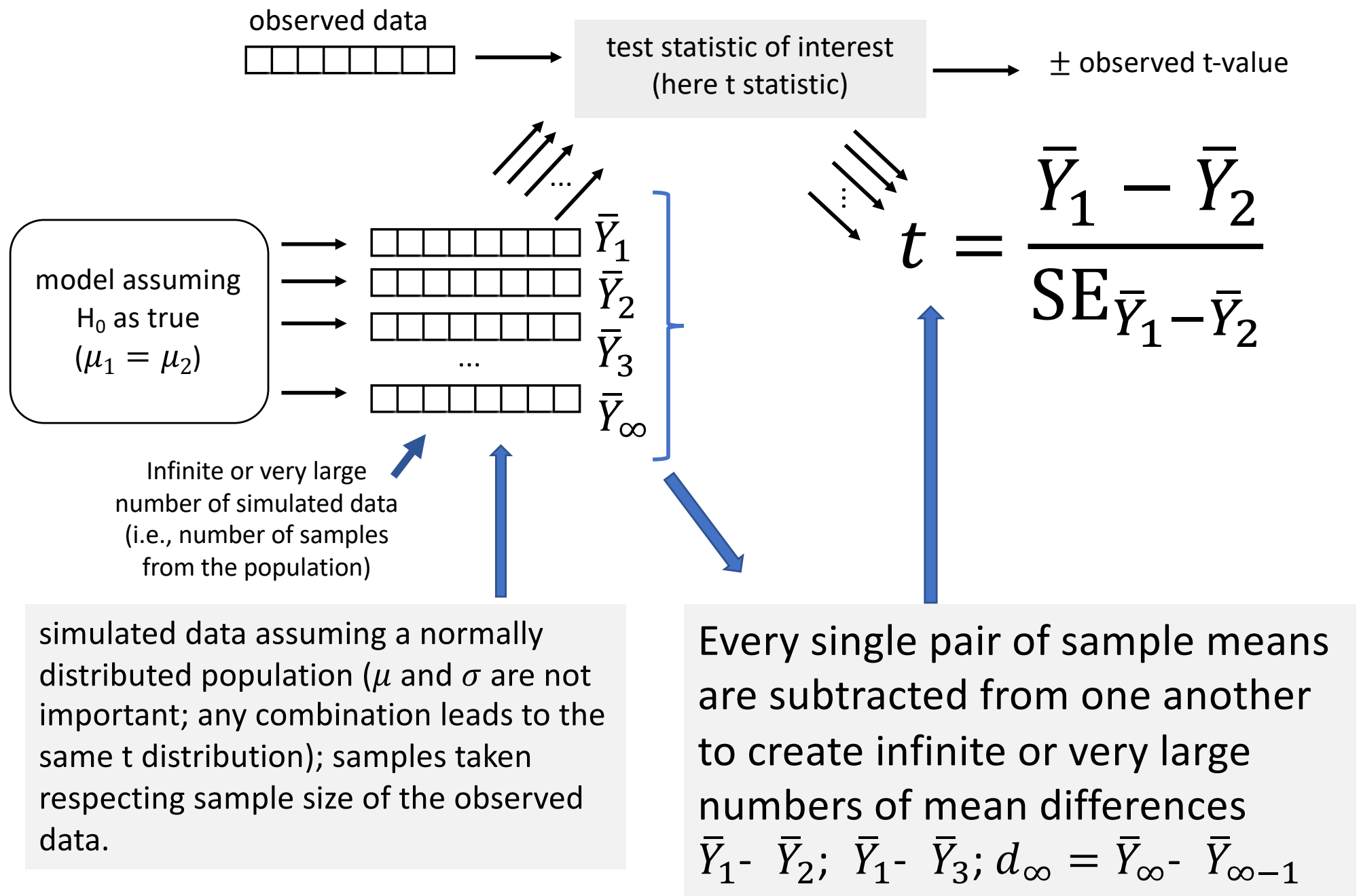
Decision based on alpha = 0.05: **reject H_0**

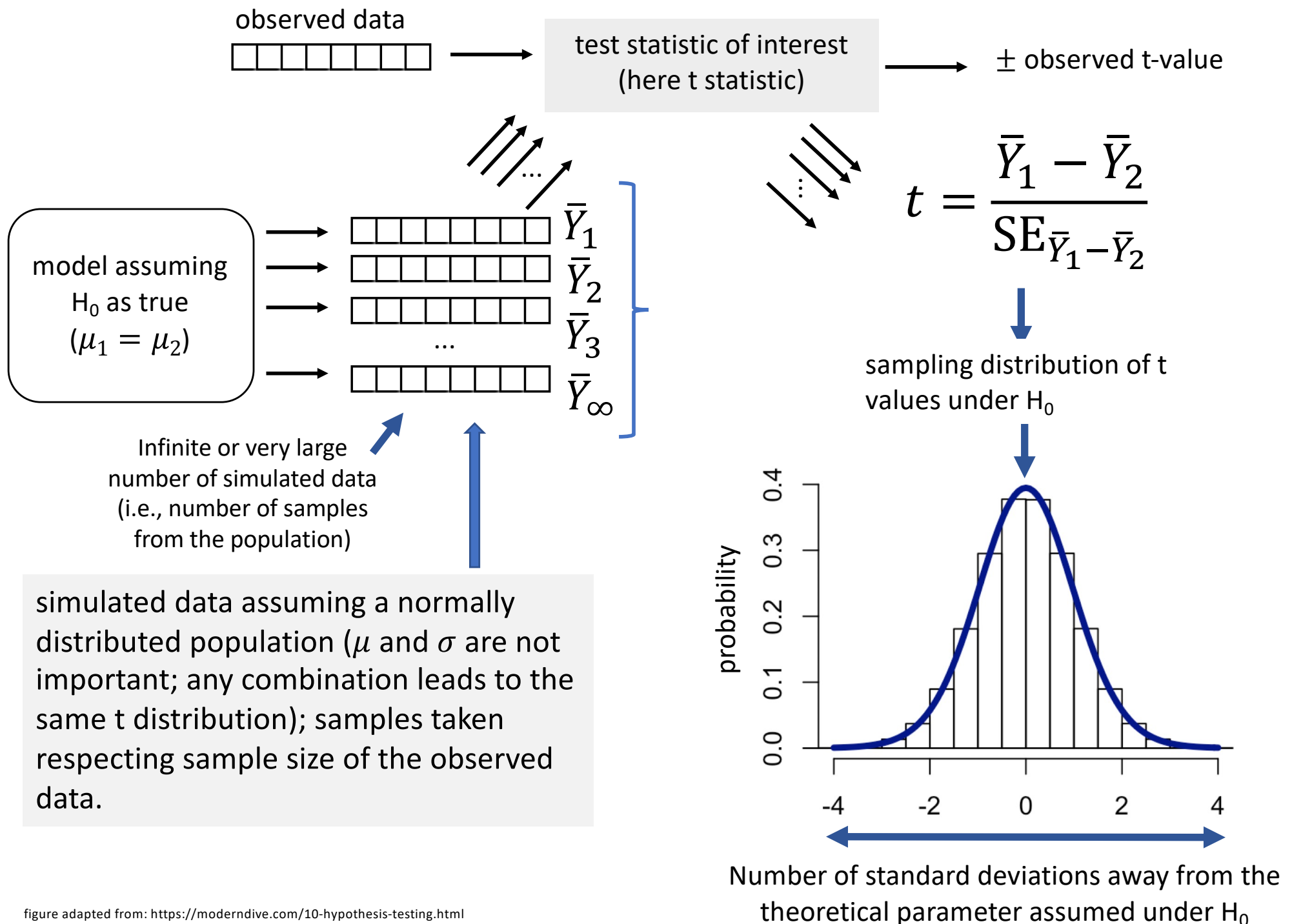
H_A : Lizards killed by shrikes and living lizard *differ* in mean horn length (i.e., $\mu_1 \neq \mu_2$).

STATISTICAL CONCLUSION: Evidence shows that the mean horn length statistically differs between lizards killed by shrikes and those that survive.

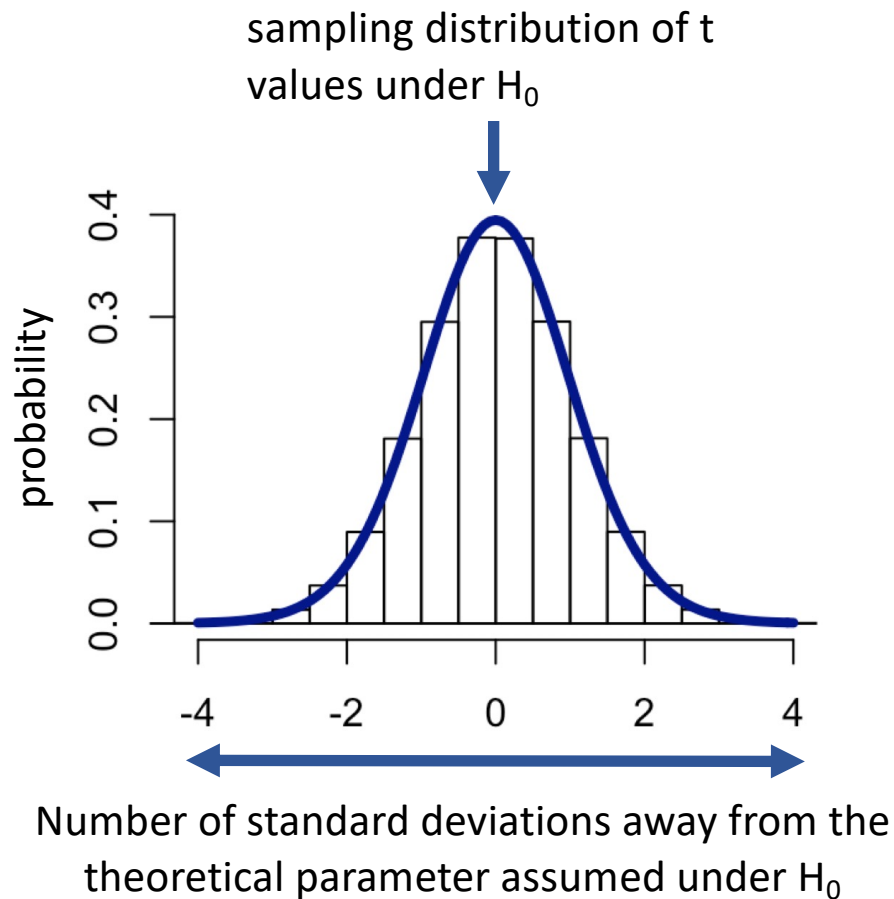


SCIENTIFIC CONCLUSION: We generated evidence suggesting that larger horn size may provide protection against predation.





Two-sample (means) t-test



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

We include $(\mu_1 - \mu_2)$ because the t-statistic measures how far the observed difference is from what is expected under the null hypothesis.

Most of the time, we test $H_0: \mu_1 - \mu_2 = 0$, so the formula simplifies to:

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - 0}{SE_{\bar{Y}_1 - \bar{Y}_2}}$$

Two sample (means) t-test

Assumptions (very important):

- Both samples are independent random samples drawn from their respective populations (e.g., surviving vs. predated individuals).
- The observations are independent within and between samples.
- The variable of interest (e.g., horn length) is approximately normally distributed in each population (especially important for small sample sizes).
- The population variances are equal across groups (homoscedasticity). We assume this here but will later learn how to assess and relax this assumption.

