## Multiple testing survey (BIOL322); anonymous survey - it will close on Monday Nov. 11 (5pm)

Results will be used to demonstrate the statistical principles of multiple testing

last number of your street address *

- ◯ Odd number
- ◯ Even number

Your birthday is an odd or even number (the actual day; not month or year) *

- ◯ Odd number
- ◯ Even number

Do you like soccer? *

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Deslike | ◯ | ◯ | ◯ | ◯ | ◯ | Love it |

Do you like video games? *

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Deslike | ◯ | ◯ | ◯ | ◯ | ◯ | Love it |

Don’t forget our survey for a in-class statistical demonstration!!

https://docs.google.com/forms/d/e/1FAIpQLSfOB7-yGs2jpcHO4IloiZVRLyTVgwoMVHR9a-0i8LoGh4_Hag/viewform?usp=sf_link

# One-sided *versus* two-sided tests

also known as

# One-tailed *versus* two-tailed tests

# The statistical hypothesis testing framework is an intimate stranger

Most researchers know how to operate it!
But few know how it really works!

***Research question -*** *Do other animals exhibit handedness as well?* (Frog example, 18 individuals)

$H_0$: Right-handed and left-handed toads are equally frequent in the population.

$H_A$: Right-handed and left-handed toads are NOT equally frequent in the population.

The alternative hypothesis $H_A$ is two-sided (or two-tailed). This just means that the alternative hypothesis allows for two possibilities:

**[1]** that the proportion is greater than 0.5, in which case right-handed toads outnumber left-handed toads in the population; OR

**[2]** that the proportion is less than 0.5 (i.e., left-handed toads predominate).

Neither possibility [1 or 2] can be ruled out before gathering the data, so both should be included in the alternative hypothesis.

$H_0$: Right-handed and left-handed toads are equally frequent in the population.

$H_A$: Right-handed and left-handed toads are NOT equally frequent in the population.

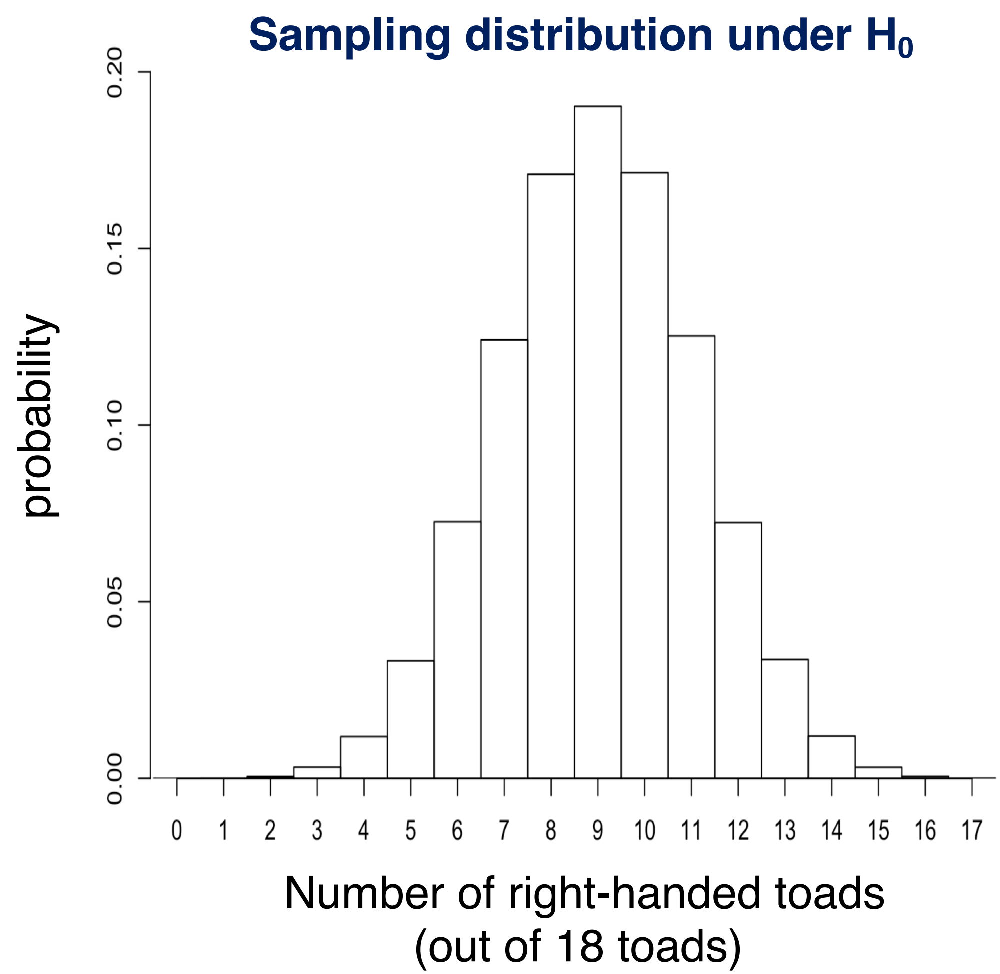
The test statistic that we will use here is the number of right-handed frogs.

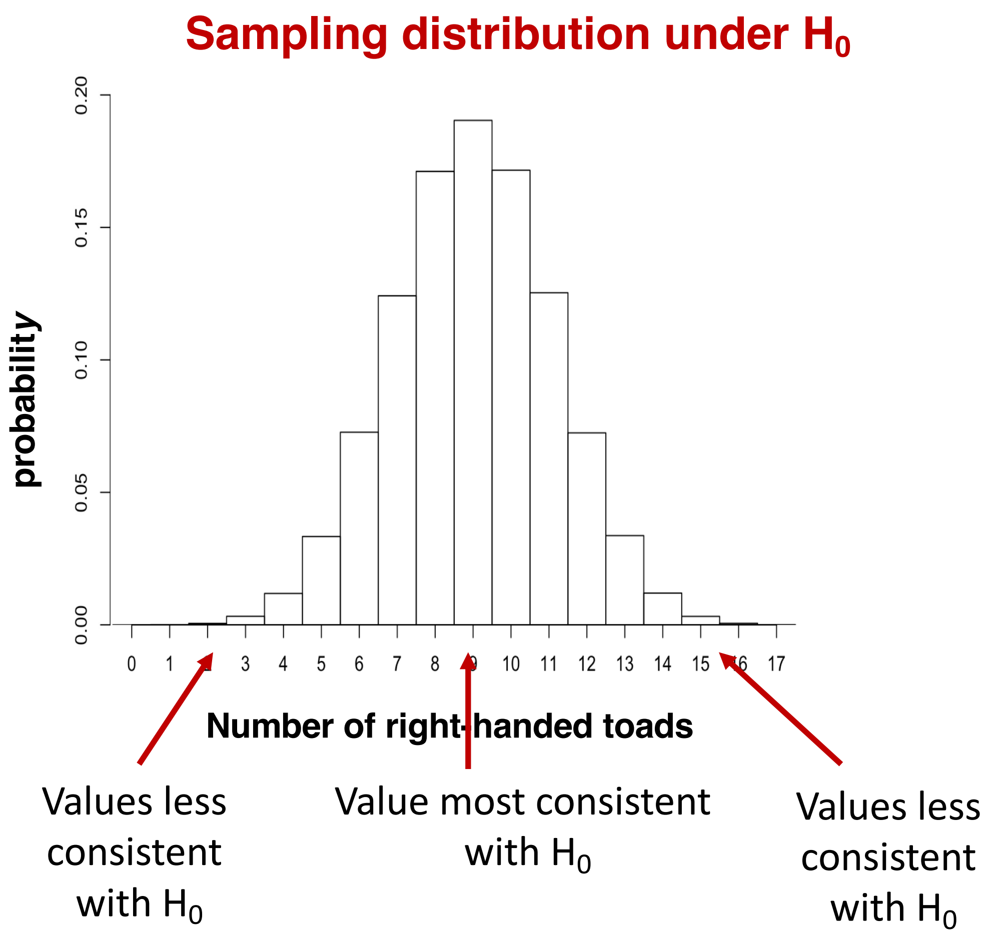
Remember, the test statistic is a calculated value based on sample data, used to assess how well the observed data aligns with expectations under the assumption that the null hypothesis ($H_0$) is true. In other words, it helps determine the compatibility of the observed results with what would be expected if random sampling occurred from a population where the null hypothesis holds.

$H_0$: Right-handed and left-handed toads are equally frequent in the population.

$H_A$: Right-handed and left-handed toads are NOT equally frequent in the population.

A perfect alignment with the null hypothesis would result in 9 right-handed and 9 left-handed frogs. However, even under the assumption that the null hypothesis is true, the majority of observed values (over 82%) deviate from this expectation.






The sampling distribution under the null hypothesis represents the range of test statistic values that are compatible with the null hypothesis.

Even if the null hypothesis ($H_0$) is true, some test statistic values align more closely with $H_0$ (**more consistent**) while others are **less consistent** with it.

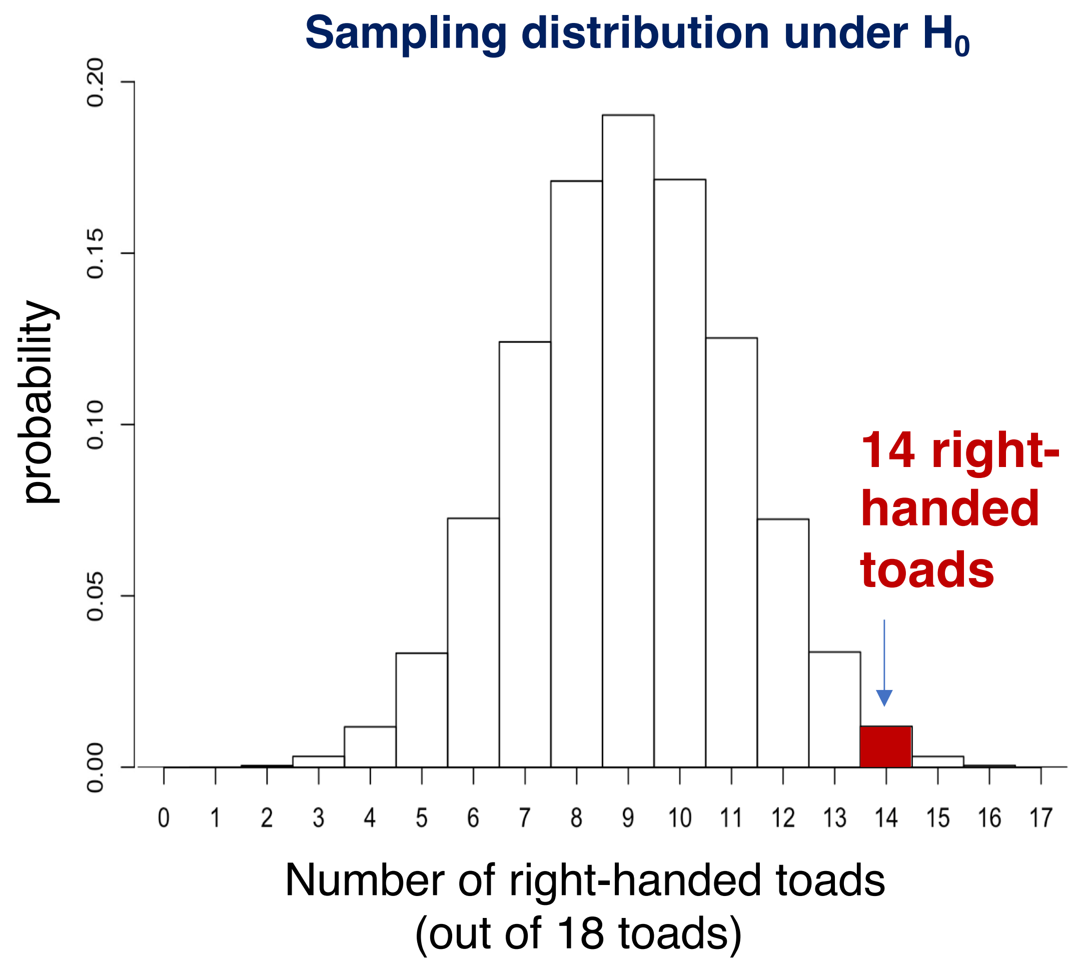
What is **consistency**? It refers to being compatible or in agreement with something (here $H_0$)

**$H_0$:** Right-handed and left-handed toads are equally frequent in the population.

**$H_A$:** Right-handed and left-handed toads are NOT equally frequent in the population.

**RESULTS:** 14 toads were found to be right-handed

Under the sampling distribution that assumes the null hypothesis ($H_0$) is true, observing 14 right-handed toads out of 18 is quite unusual if the null hypothesis were correct.



**$H_0$:** Right-handed and left-handed toads are equally frequent in the population.

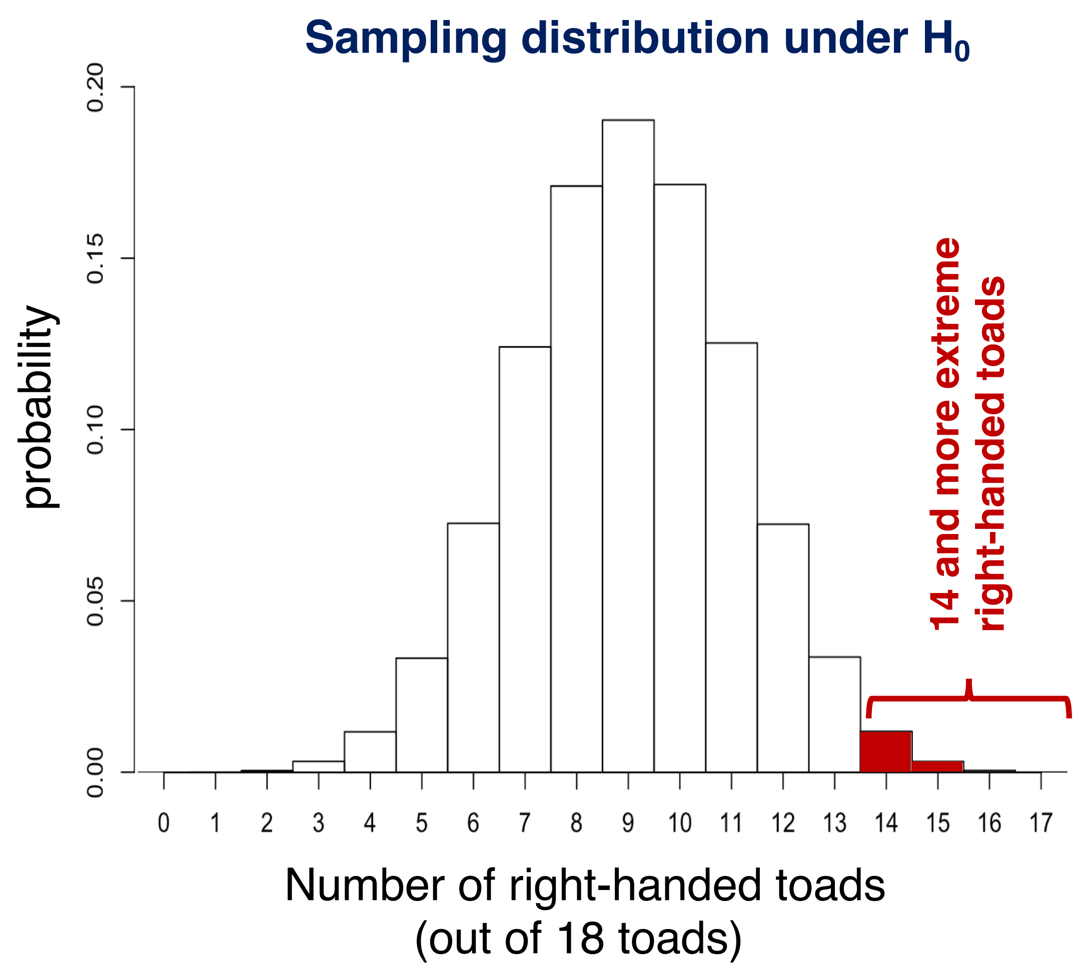
**$H_A$:** Right-handed and left-handed toads are NOT equally frequent in the population.

**RESULTS:** 14 toads were found to be right-handed

Why should we also consider values more extreme than the observed ones, such as 15, 16, 17, and 18?

These values are even rarer in the theoretical sampling distribution under the assumption that the null hypothesis ($H_0$) is true, making them even less consistent with $H_0$.

Thus, values more extreme than the observed outcome provide additional evidence against $H_0$.



**RESULTS: 14 toads were found to be right-handed**

**Why do we also consider the frequency of right-handed toads on the left side of the distribution with values of 4 or more extreme?**

Pr[14 or more right-handed toads] = Pr[14] + P[15] + P[16] + P[17] + P[18] = **0.0155**

**+**

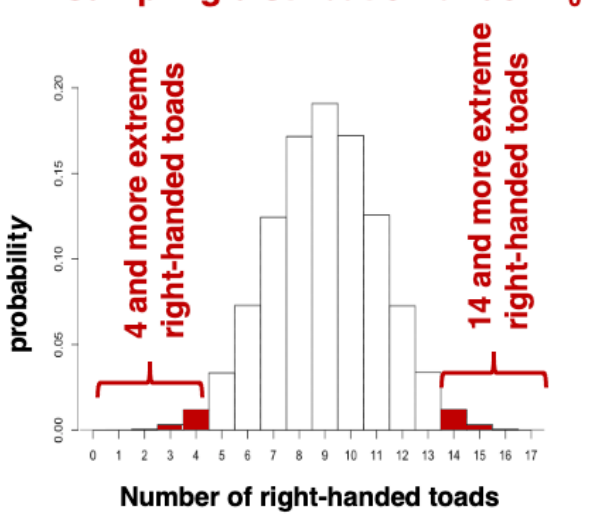
Pr[4 or less right-handed toads] = Pr[4] + P[3] + P[2] + P[1] + P[0] = **0.0155**

**= 0.031**

$H_0$: Right-handed and left-handed toads are equally frequent in the population.

$H_A$: Right-handed and left-handed toads are NOT equally frequent in the population.

**Sampling distribution under $H_0$**




The alternative hypothesis $H_A$ is two-sided (or two-tailed). This just means that the alternative hypothesis allows for two possibilities:

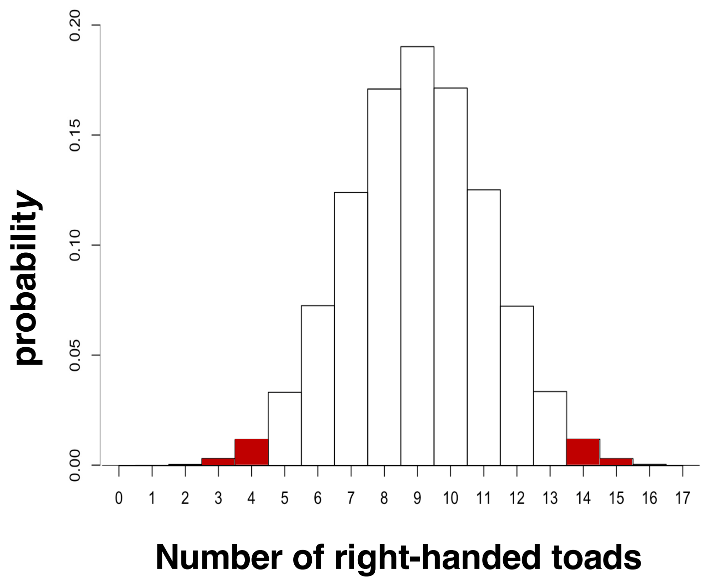
[1] that the proportion is greater than 0.5, in which case right-handed toads outnumber left-handed toads in the population; OR

[2] that the proportion is less than 0.5 (i.e., left-handed toads predominate).

Neither possibility [1 or 2] can be ruled out before gathering the data, so both should be included in the alternative hypothesis.

Let's contrast the observed test statistic with the sampling distribution underlying $H_0$.

**Sampling distribution under $H_0$**




P=0.031

P-value is a **measure of consistency** of the observed test statistic and more extreme values with the sampling distribution underlying $H_0$.

**Why should we also count the more extreme values than the observed?** Because they are even rarer to observe in the sampling distribution assuming $H_0$ as true.

Therefore, values more extreme than the observed count as evidence against $H_0$ as well, thus assisting in measure whether the observed test statistic is consistent or not with $H_0$.

**If the p-value is high**, then the observed sample is consistent with the general proposition of $H_0$ (i.e., number of right- and left-handed toads are the same).

**If the p-value is low**, then the observed sample is inconsistent with the general proposition of $H_0$. And is more consistent with the proposition of $H_A$ (i.e., number of right- and left-handed toads are NOT the same).

As we saw, high and low p-values are decided according to the significance value, alpha.

$H_0$: Right-handed and left-handed toads are equally frequent in the population.

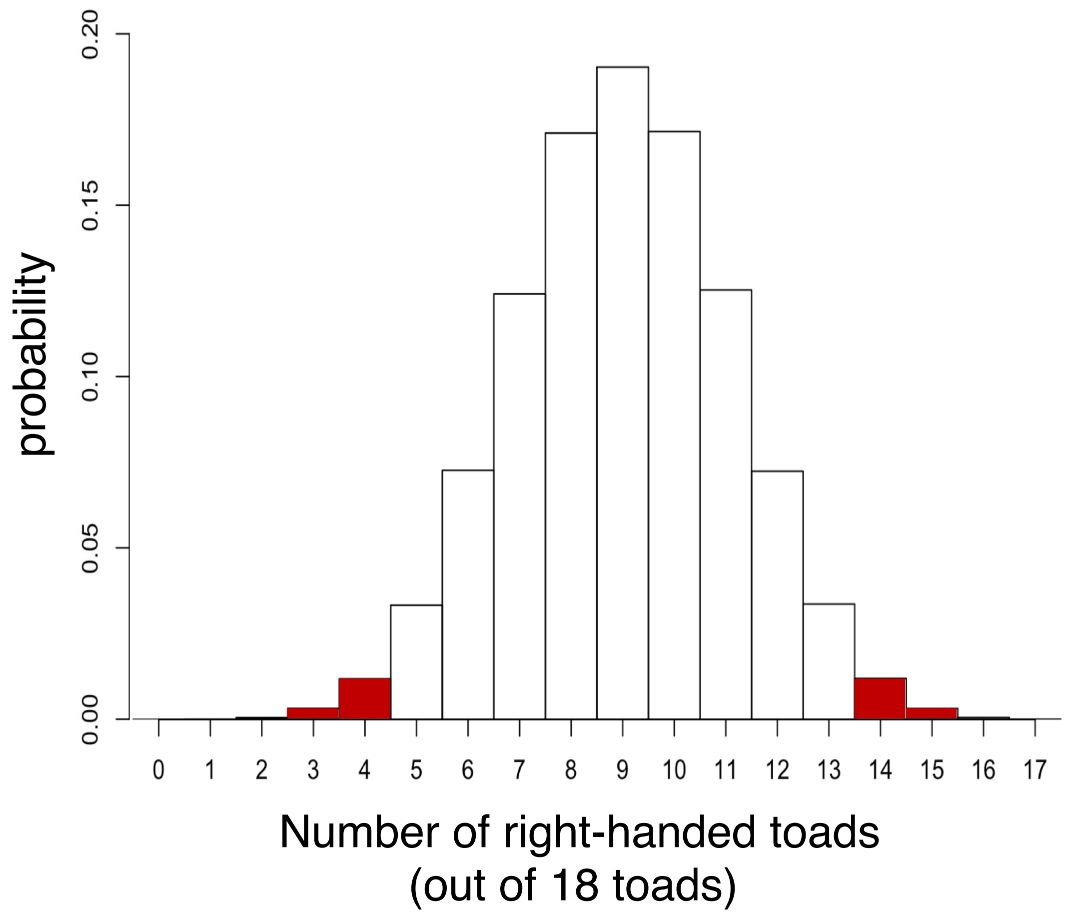
$H_A$: Right-handed and left-handed toads are NOT equally frequent in the population.

**RESULTS:** 14 toads were found to be right-handed

Pr[14 or more right-handed toads] =
Pr[14]+P[15]+P[16]+P[17]+P[18] =
0.0155 x 2 = **0.031**

In summary: this is clearly a two-tailed test:

We have no clear theoretical basis for predicting a deviation from the $H_0$ in one direction over the other direction.



Rule: if you don’t have a clear theoretical basis, always choose a two-tailed test

# Let’s take a break – 1 minute

# One-sided *versus* two-sided tests (toad example)

- In a one-sided (or one-tailed) test, the alternative hypothesis considers values for the test statistic under the null hypothesis on only one side of the value specified by the null hypothesis.

- $H_0$ is rejected only if data depart from it in the direction stated by $H_A$.

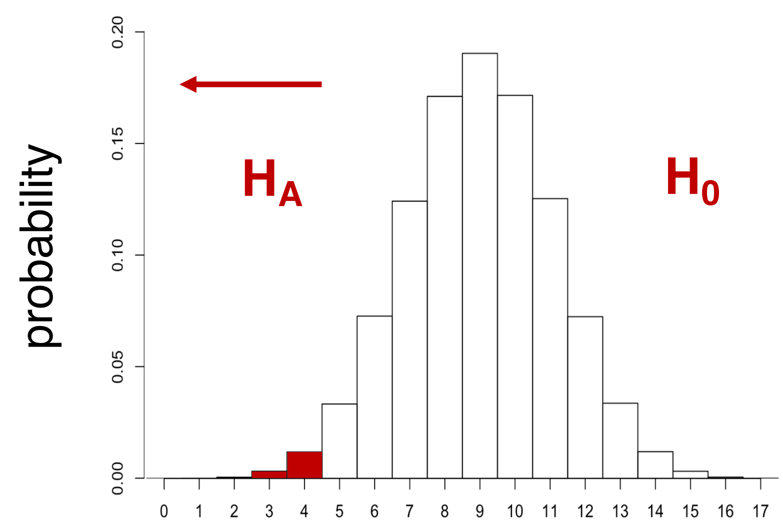
P-value is a **measure of consistency** of the observed test statistic and more extreme values with the sampling distribution underlying $H_0$.

One-sided instead - so that it becomes easier to understand; though there is no clear theoretical basis for $H_0$ & $H_A$ (left side):

**$H_0$:** The number of right-handed *is equal or greater* than left-handed toads in the population.

**$H_A$:** The number of right-handed is smaller than left-handed toads in the population.

Number of right-handed frogs is smaller than expected by chance from a population where toads are 50%/50%




Number of right-handed toads
(out of 18 toads)

# One-sided *versus* two-sided tests (toad example)

- In a one-sided (or one-tailed) test, the alternative hypothesis considers values for the test statistic under the null hypothesis on only one side of the value specified by the null hypothesis.

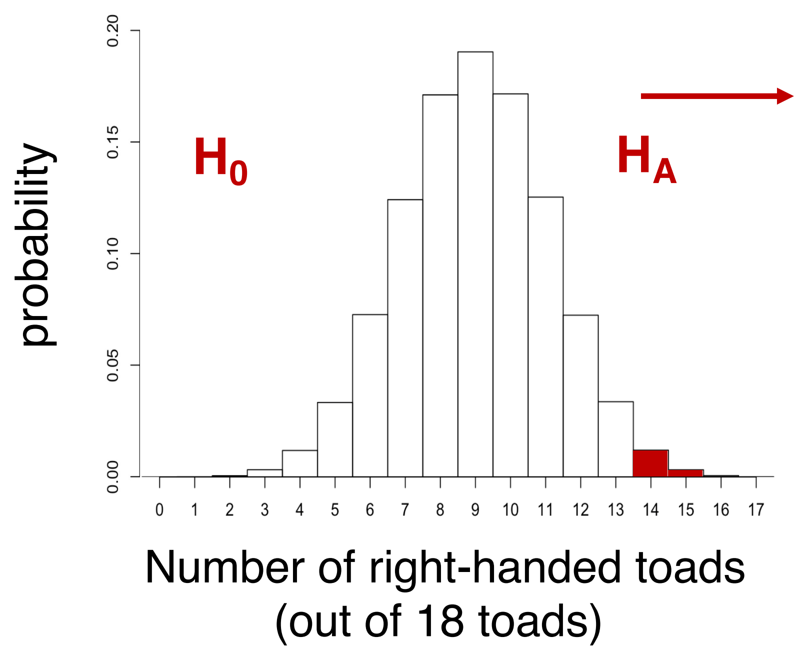
- $H_0$ is rejected only if data depart from it in the direction stated by $H_A$.

One-sided instead - so that it becomes easier to understand; though there is no clear theoretical basis for **$H_0$ & $H_A$** (right side):

**$H_0$:** The number of right-handed *is equal or smaller* than left-handed toads in the population.

**$H_A$:** The number of right-handed is greater than left-handed toads in the population.

Number of right-handed frogs is greater than expected by chance from a population where toads are 50%/50%



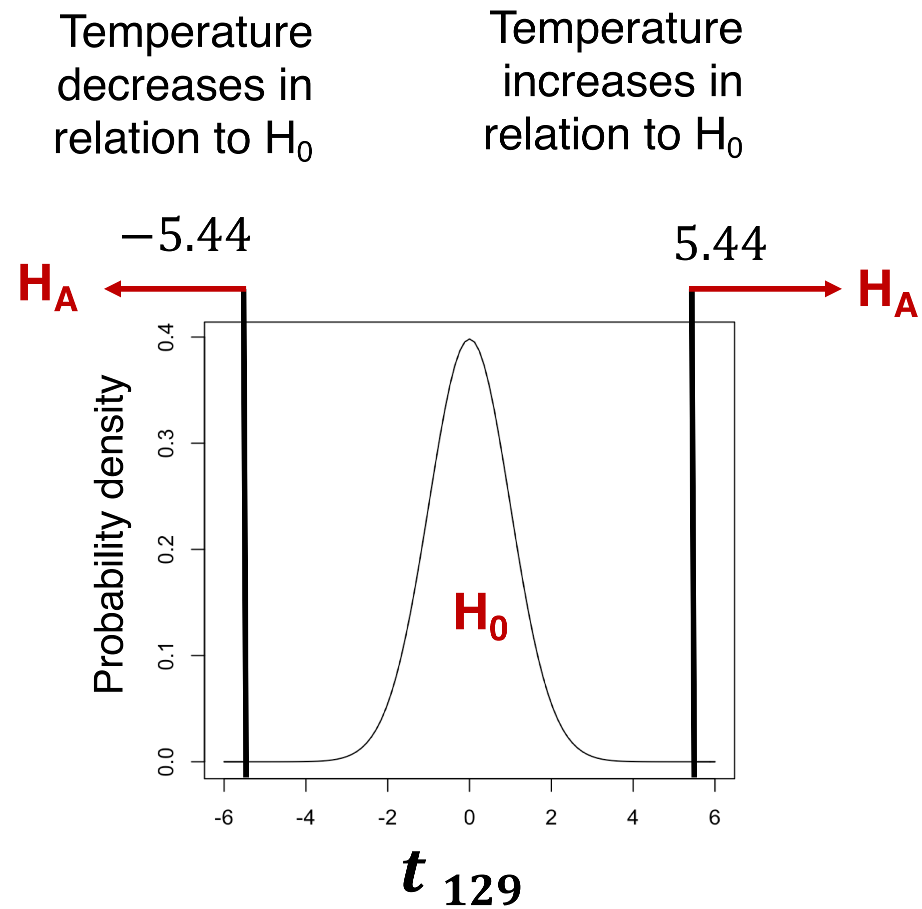
# One-sided *versus* two-sided tests (human body temperature) (based on the increased sample size)

**Two-sided:**

$H_0$: the mean human body temperature is 98.6°F.

$H_A$: the mean human body temperature is different from 98.6°F.

Pr[t < -5.44] + Pr[t > 5.44] =
2 Pr[t > abs(5.44)] = **0.000016**

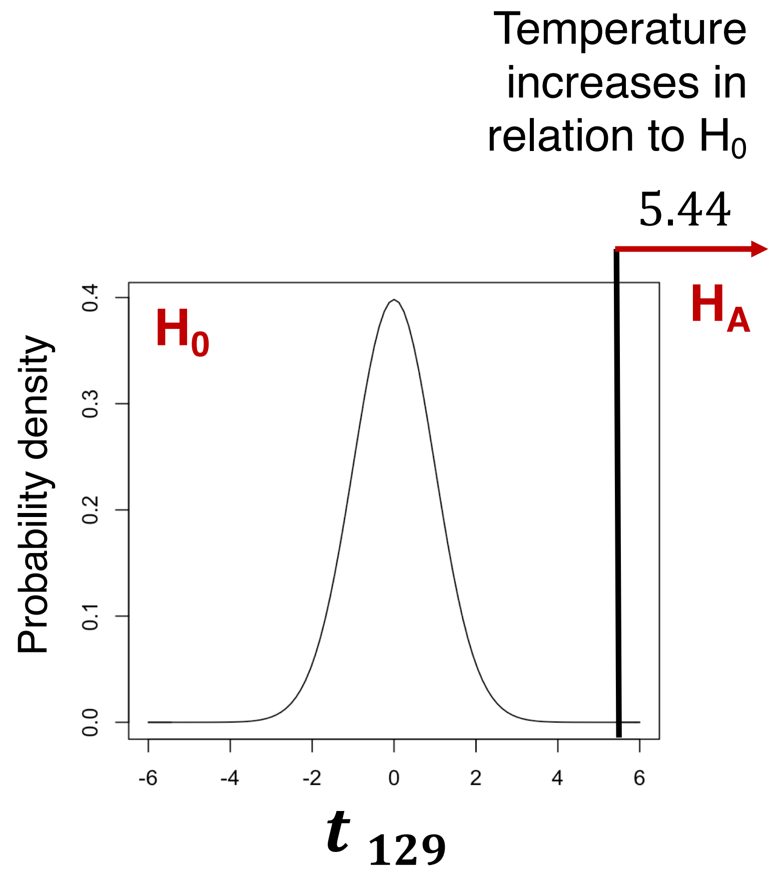


# One-sided *versus* two-sided tests (human body temperature) (based on the increased sample size)

One-sided instead - so that it becomes easier to understand; though there is no clear theoretical basis for **$H_0$ & $H_A$** (right side):

**$H_0$:** the mean human body temperature is smaller or equal to 98.6°F.

**$H_A$:** mean human body temperature is greater than 98.6°F.

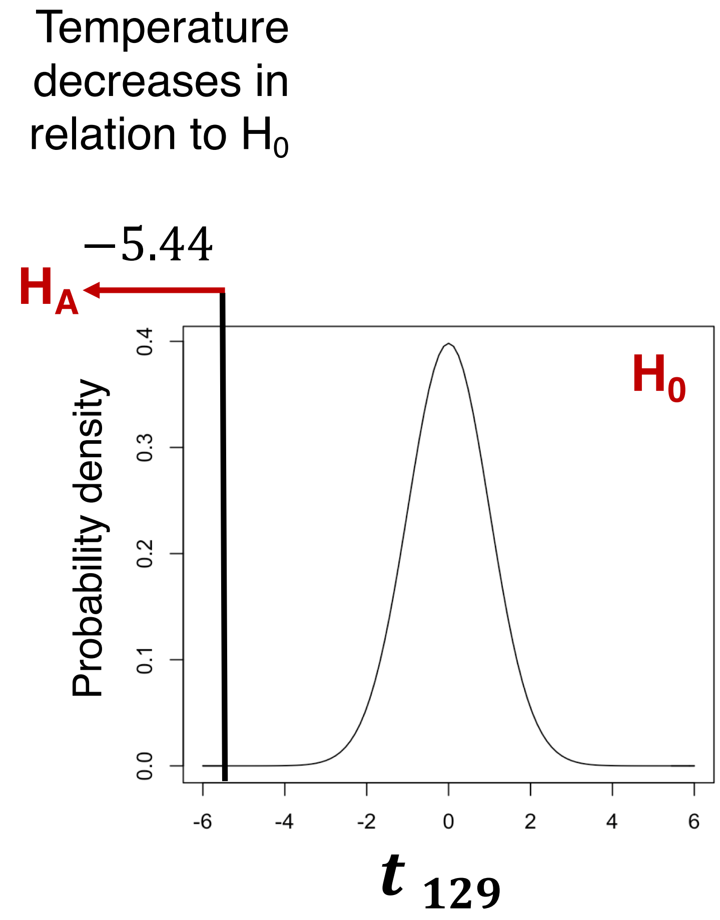


# One-sided *versus* two-sided tests (human body temperature) (based on the increased sample size)

One-sided instead - so that it becomes easier to understand; though there is no clear theoretical basis for $\mathbf{H_0}$ **&** $\mathbf{H_A}$ (left side):

$H_0$: the mean human body temperature is equal or greater than 98.6$^o$F.

$H_A$: mean human body temperature is smaller than 98.6$^o$F.



# One-sided *versus* two-sided tests

## The two-sample test: two- *versus* one-sided tests

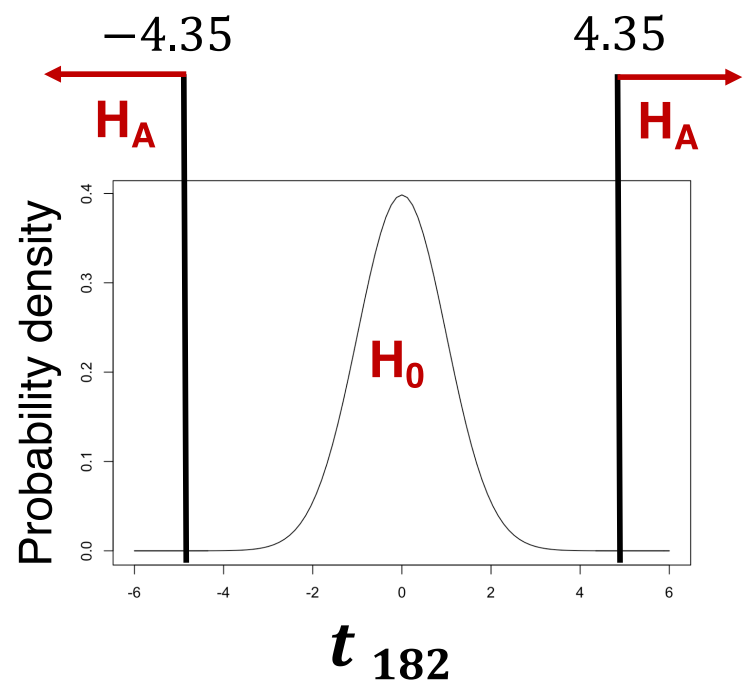
**Research question -** Do spikes help protect horned lizards from being eaten by predators? It's possible that individuals with larger spikes may carry extra weight, which could hinder their ability to escape from predators compared to those with smaller spikes.

$H_0$: 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., $\mu_1 \neq \mu_2$).

Pr[t < -4.35] + Pr[t > 4.35] =
2 Pr[t > abs(4.35)] = **0.000023**

This should be a two-tailed test – we have no clear theoretical basis for predicting a deviation from the $H_0$ in one direction over the other direction.



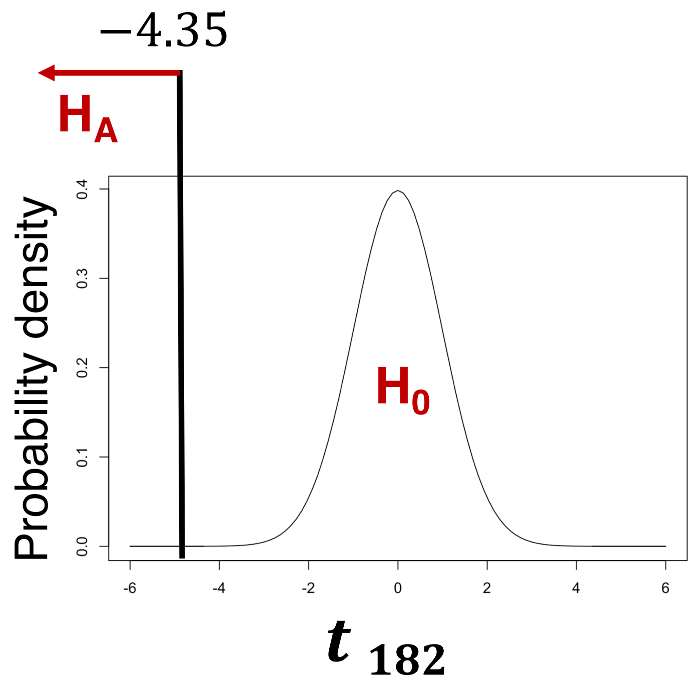
# One-sided *versus* two-sided tests (lizard)

Although there is no theoretical basis to choose a two-sided test in this case, here are the one-tailed possible hypotheses:

One-sided instead – so that you understand though no clear theoretical basis for these (left side): t based on $(\bar{X}_{killed}$- $\bar{X}_{live})$/SE

**$H_0$:** Lizards killed by shrikes have larger or equal mean horn length than living lizard (i.e., $\mu_{killed} \geq \mu_{living}$). t value is positive

**$H_A$:** Lizards killed by shrikes have smaller mean horn length than living lizard (i.e., $\mu_{killed} < \mu_{living}$). t value is negative



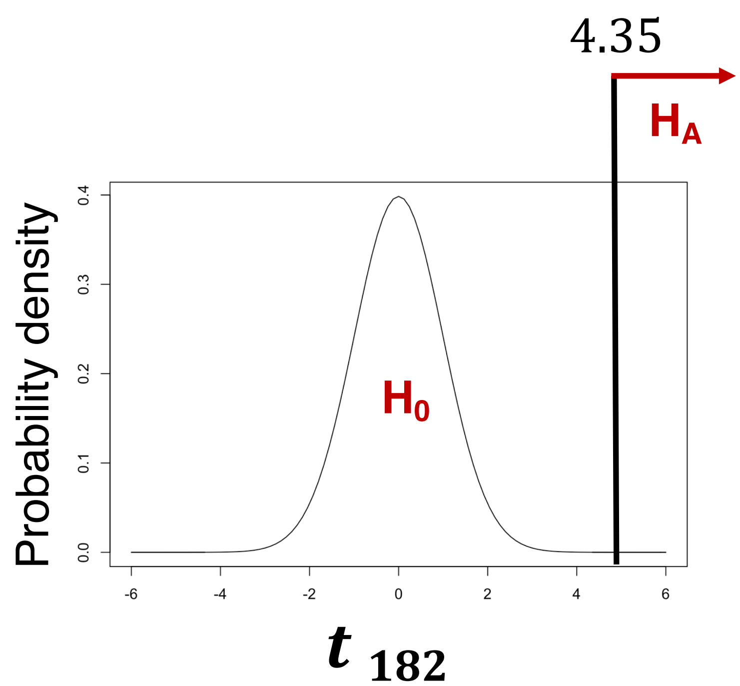
# One-sided *versus* two-sided tests (lizard)

Although there is no theoretical basis to choose a two-sided test in this case, here are the one-tailed possible hypotheses:

One-sided instead – so that you understand though no clear theoretical basis for these (left side): t based on $(\bar{X}_{killed}$- $\bar{X}_{live})$/SE

**$H_0$:** Lizards killed by shrikes have smaller or equal mean horn length than living lizard (i.e., $\mu_{killed} \leq \mu_{living}$). t value is negative

**$H_A$:** Lizards killed by shrikes have greater mean horn length than living lizard (i.e., $\mu_{killed} > \mu_{living}$). t value is positive



# Let’s take a break – 1 minute

Rule: if you don’t have a clear theoretical basis, always choose a two-tailed test

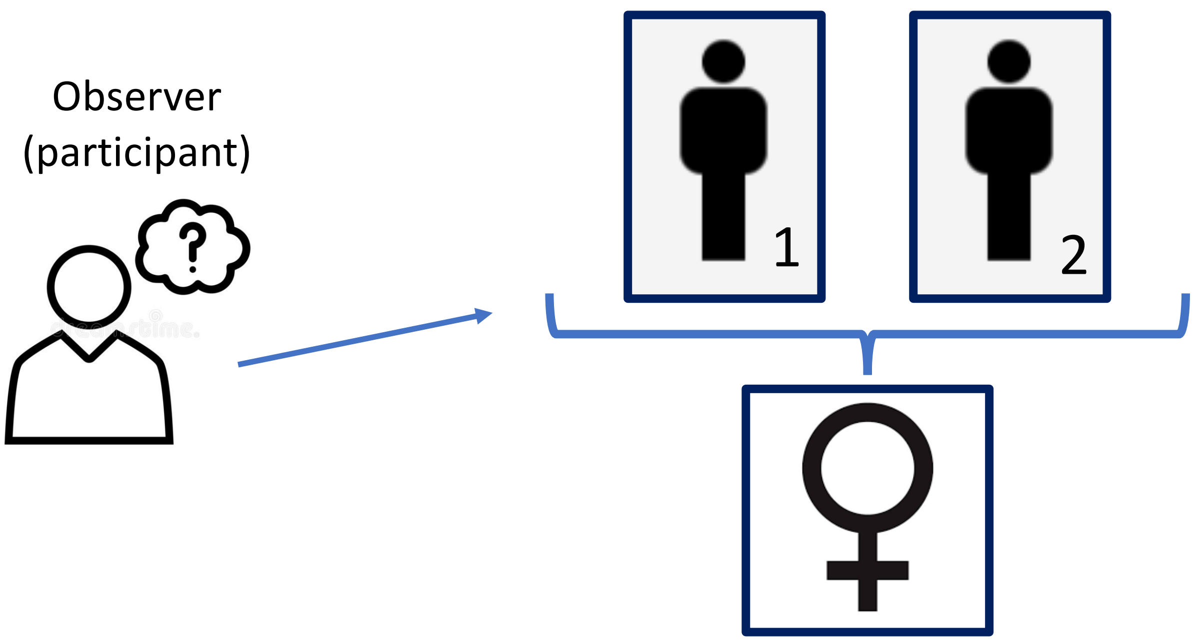
A fictional example where a one-sided test is preferable

# One-sided *versus* two-sided tests

For the three previous examples discussed, there was no clear theoretical basis for predicting a deviation from the null hypothesis ($H_0$) in either direction. Therefore, a two-sided test should be applied.

**Let's describe a fictional study where such theoretical basis exists:**

Imagine a study designed to test whether daughters resemble their fathers. Each out of 18 participants examines a photo of one girl and photos of two adult men (one of whom is the girl's father).



# One-sided *versus* two-sided tests

**Let's describe a fictional study where such theoretical basis exists:**

Imagine a study designed to test whether daughters resemble their fathers. Each out of 18 participants examines a photo of one girl and photos of two adult men (one of whom is the girl's father).

The only reasonable alternative hypothesis is that daughters indeed resemble their fathers more than expected by chance, i.e., why would we expect that daughters resemble their fathers less than other men?

**$H_0$:** Participants pick the father correctly half of the time ($p = 1/2$).
**$H_A$:** Participants pick the father more frequently than half of the time ($p > 1/2$).

$H_0$: expected under pure guess (chance) alone

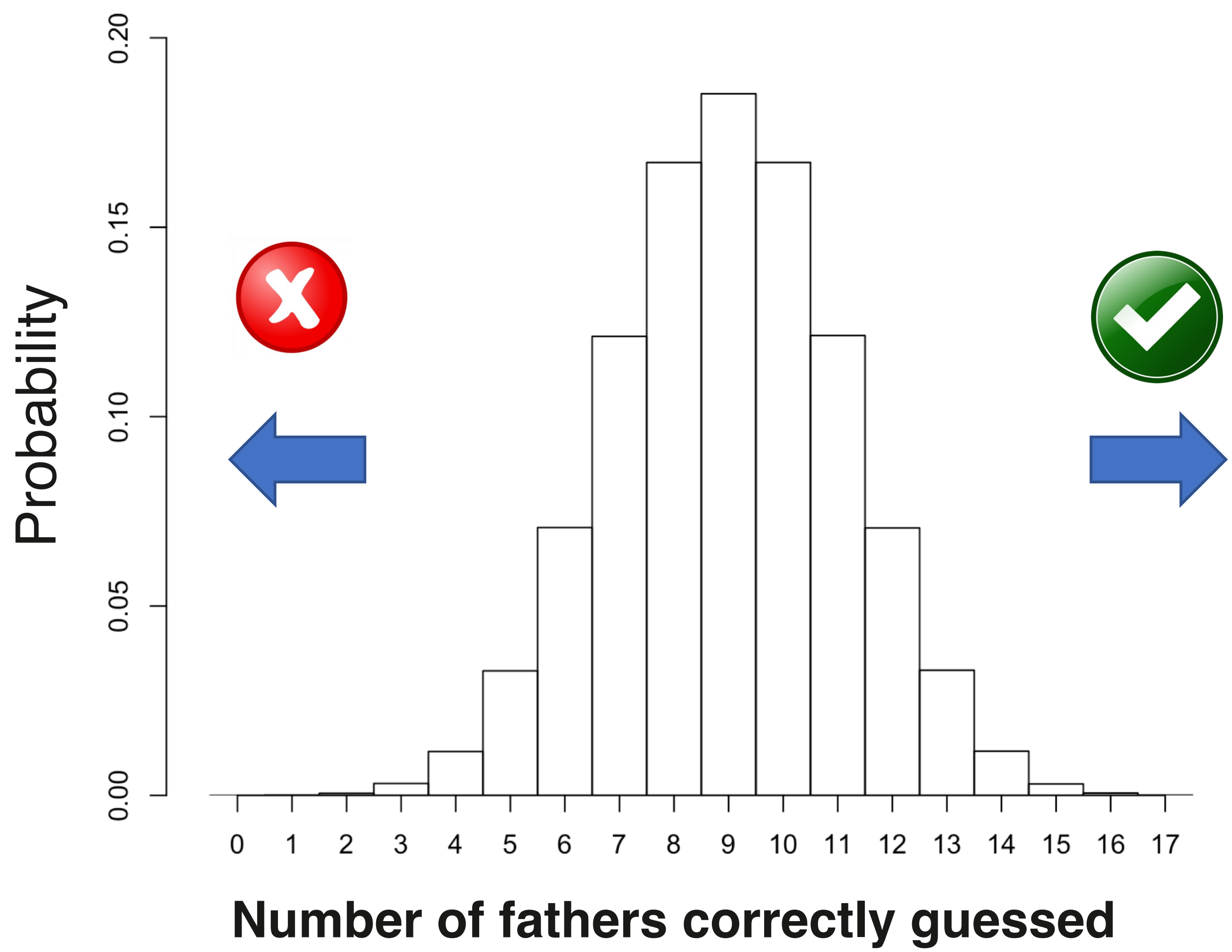
# One-sided *versus* two-sided tests

**$H_0$:** Participants pick the father correctly half of the time ($p = 1/2$).
**$H_A$:** Participants pick the father more frequently than half of the time ($p > 1/2$).

The only reasonable alternative hypothesis is that daughters indeed resemble their fathers more than expected by chance, i.e., why would they resemble their fathers less than other men?

A one-sided test is justifiable in this context because any values on the opposite side of the null hypothesis ($H_0$) value would be inconceivable except by random chance.

Specifically, it is implausible to imagine daughters resembling their fathers less than they would resemble randomly chosen men.



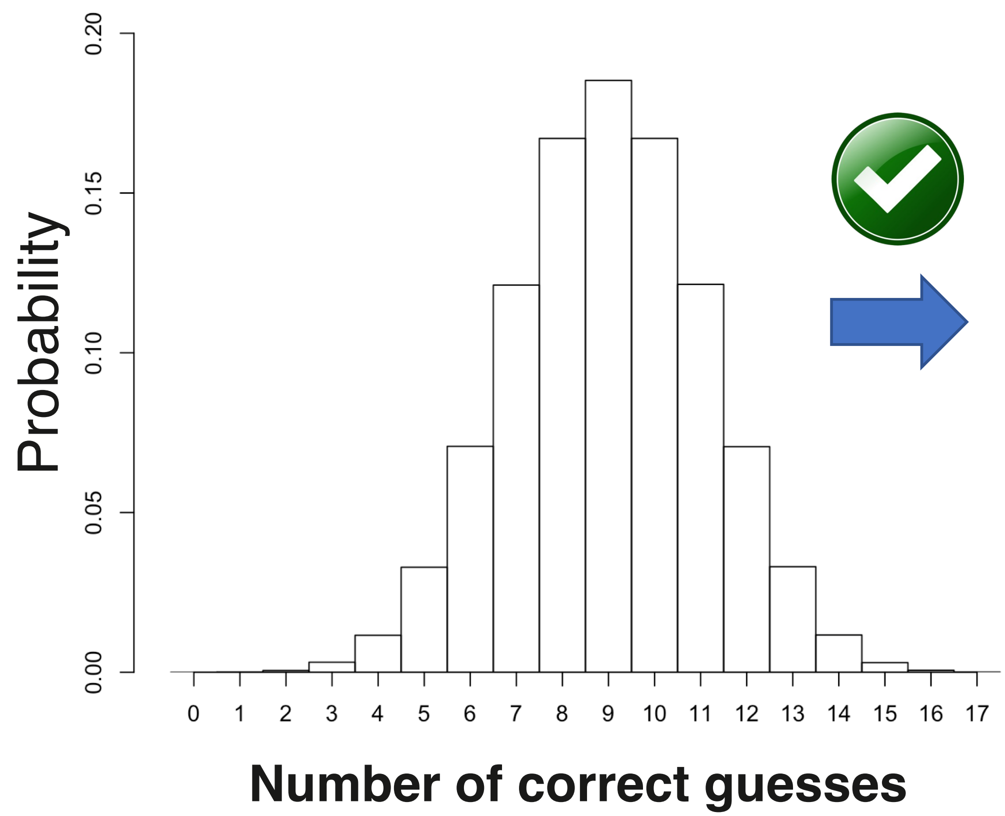
# One-sided *versus* two-sided tests

$H_0$: Participants pick the father correctly half of the time (p = 1/2).
$H_A$: Participants pick the father more frequently than half of the time (p > 1/2).

Let's say that 14 daughters out of 18 were paired correctly with their fathers.

P = Pr[number of correct guesses $\geq$ **14**]
= Pr[14] + … + Pr[18]
= **0.0155** (i.e, assuming that $H_0$ is correct).

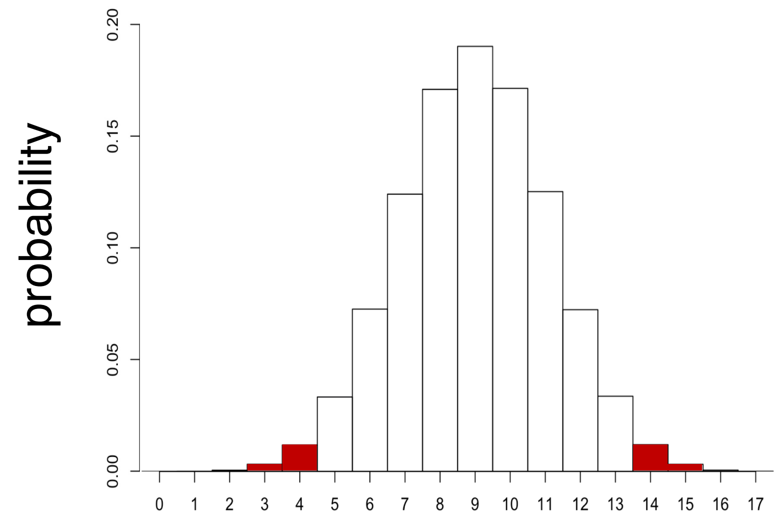
There is no need to multiply this probability by two as is done in two-sided tests, since it only accounts for values in one tail of the distribution under the assumption that the null hypothesis ($H_0$) is true.



# One-sided *versus* two-sided tests – the differences in P-values

Pr[14 or more right-handed toads] =
Pr[14] + P[15] + P[16] + P[17] + P[18] =
0.0155 x 2 = **0.031**

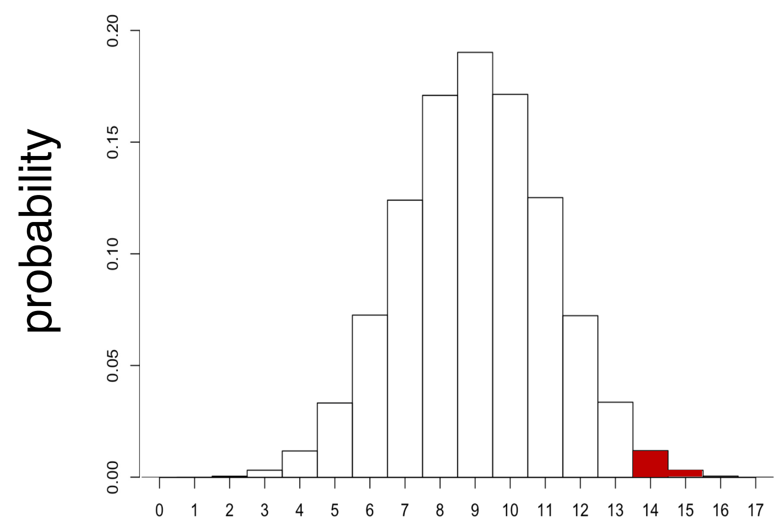
This is a two-tailed test – we have no clear theoretical basis for predicting a deviation from the $H_0$ in one direction over the other direction.




Number of right-handed frogs (out of 18 frogs)

Pr[14 or more right-handed toads] =
Pr[14] + P[15] + P[16] + P[17] + P[18] =
**0.0155**

This is a one-tailed test – we have clear theoretical basis for predicting a deviation from the $H_0$ in one direction over the other direction.




Number of fathers correctly guessed

One-sided tests lead to smaller p-values, which increases statistical power.

# One-sided *versus* two-sided tests

**Two-sided tests keep us honest!**

What if we carried out a subsequent study to test whether daughters, when they marry, choose husbands who resemble their fathers?

The null hypothesis is that there is no resemblance, but what is the alternative hypothesis here then?

Should it be one-sided (husbands resemble fathers) or two-sided (husbands may resemble fathers OR husbands may not resemble fathers in contrast to chance alone)?

We should opt for a two-sided test here because there is no theoretical basis to establish one side over the other.

# One-sided *versus* two-sided tests

**Two-sided tests keep us honest!**

One researcher may have a clear theoretical basis for a particular one-sided hypothesis, but another researcher may not.

We may be tempted to choose the side that provided us with greater probability of significant results (i.e., greater statistical power) - Two-sided tests keep us honest!

CONCLUSION: unless one has a clear theoretical basis to support a one-sided test, use a two-sided test.

**Rule:** if you don't have a clear theoretical basis, always choose a two-tailed test