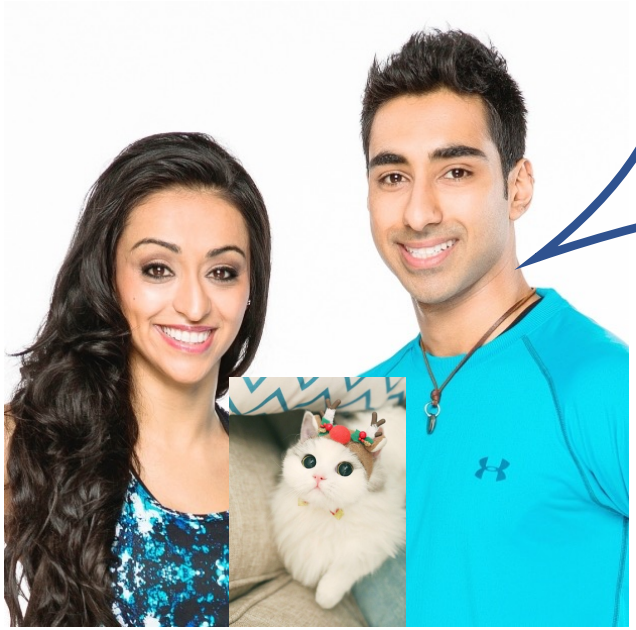
Knowledge advances not because we know the truth, but because we learn how to reason when the truth is unknown.

Who takes Mimy, the cat?
The decisional hardships of a brother and a sister!

I want to take Mimy!



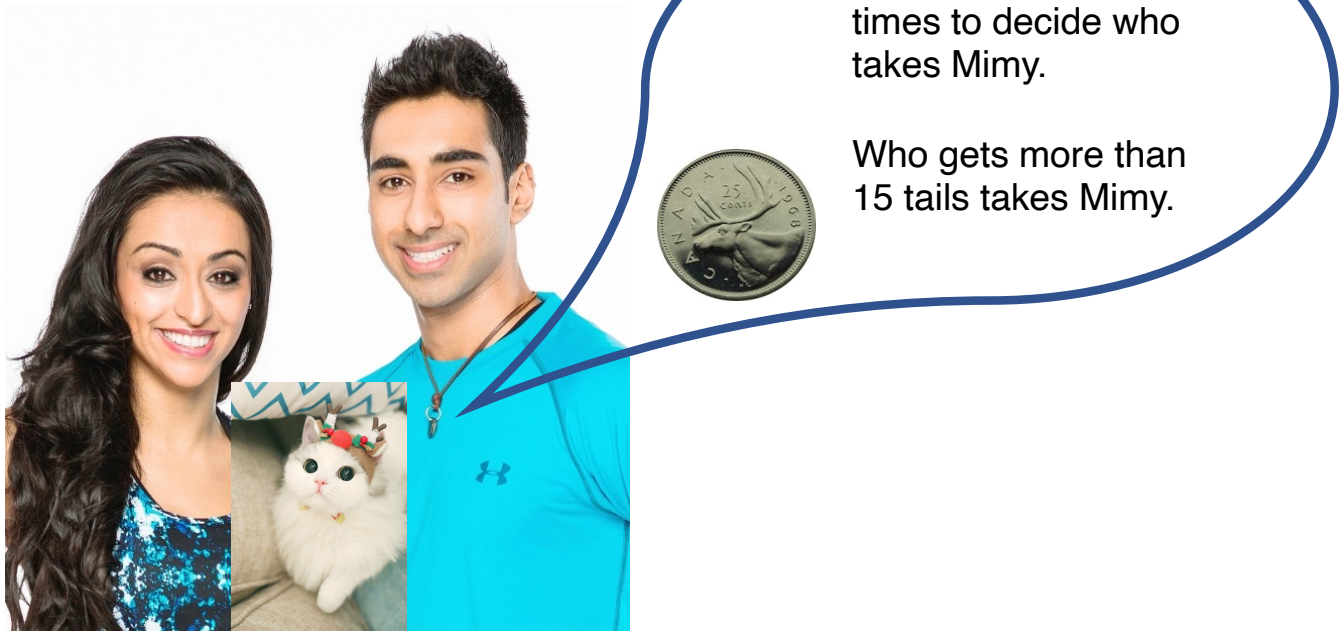
I want to keep Mimy!

Sukhi & Jinder Atwal B.C. brother-sister team off to The Amazing Race. And they are competitive!

They share Mimy, the family cat!

But Jinder is now leaving their hometown and wants to take Mimy with him.

Disclaimer: the story and characters here are made up for pedagogical entertainment ☺



Disclaimer: the story and characters here are made up for pedagogical entertainment ☺

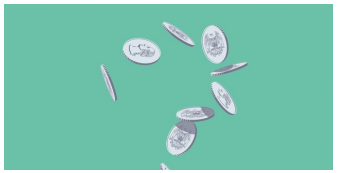
Your coin? You always cheated in games when we were young! Let's use my coin

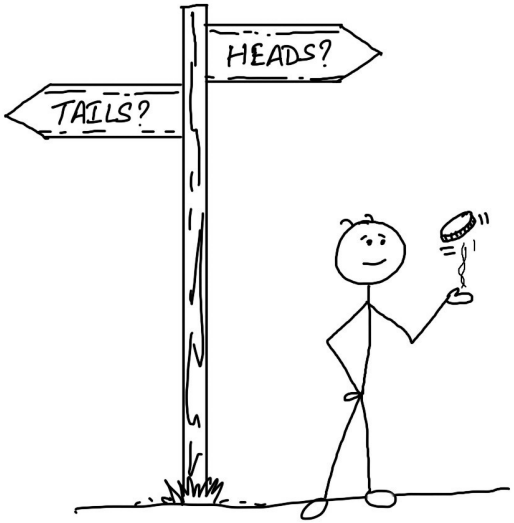
But should I trust you? You also love Mimy too much

Disclaimer: the story and characters here are made up for pedagogical entertainment ☺

Sukhi, who took BIOL322, proposes using statistics to assess how fair each other's coins are!

All right, I propose a statistical experiment to test our coins.





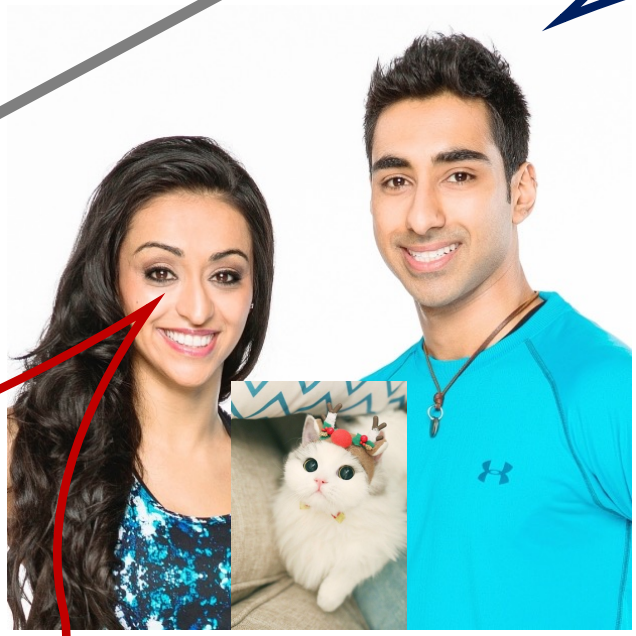
Disclaimer: the story and characters here are made up for pedagogical entertainment ☺

Sukhi proposes that Jinder takes her coin & she takes his.

Each one will flip their own coin in several sets of 30 flips, and then we will graph the results

Sukhi proposes that Jinder takes her coin & she takes his.

You really don't trust me!

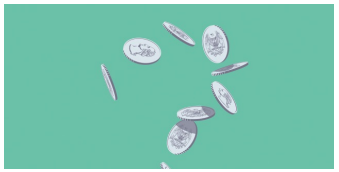


Each one will flip their own coin in several sets of 30 flips, and then we will graph the results

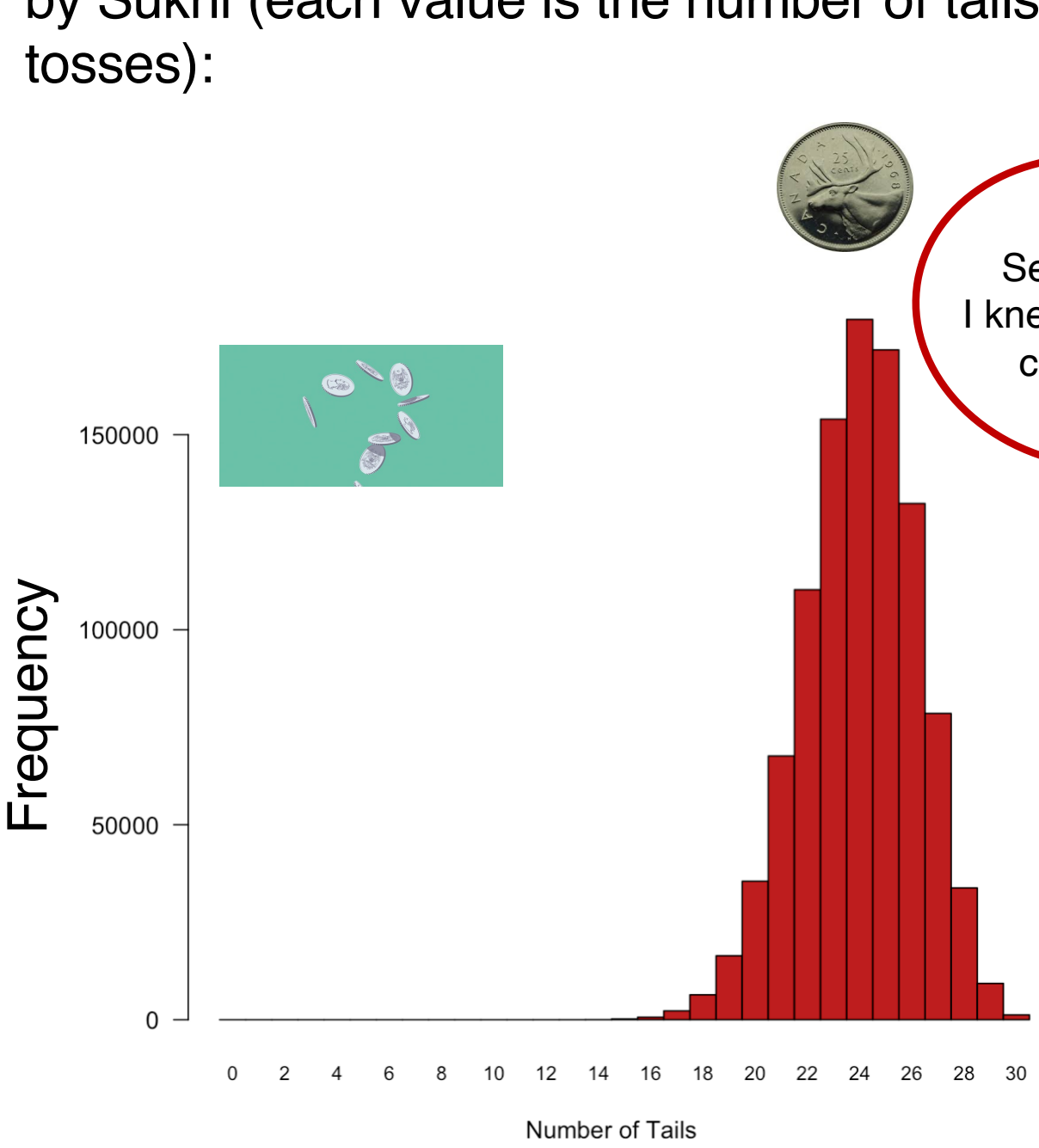
Sukhi proposes that Jinder takes her coin & she takes his.

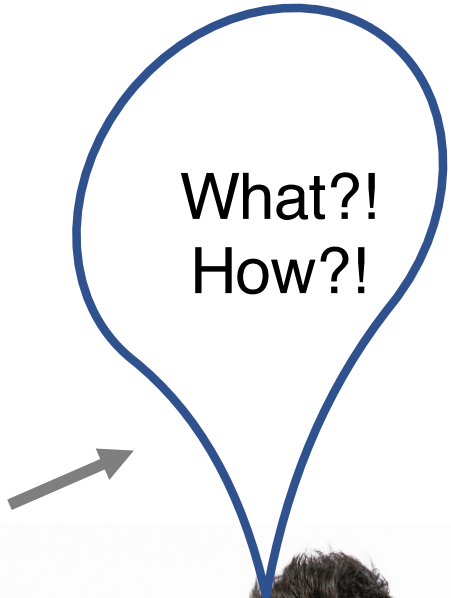
Nope,
I don’t!!!

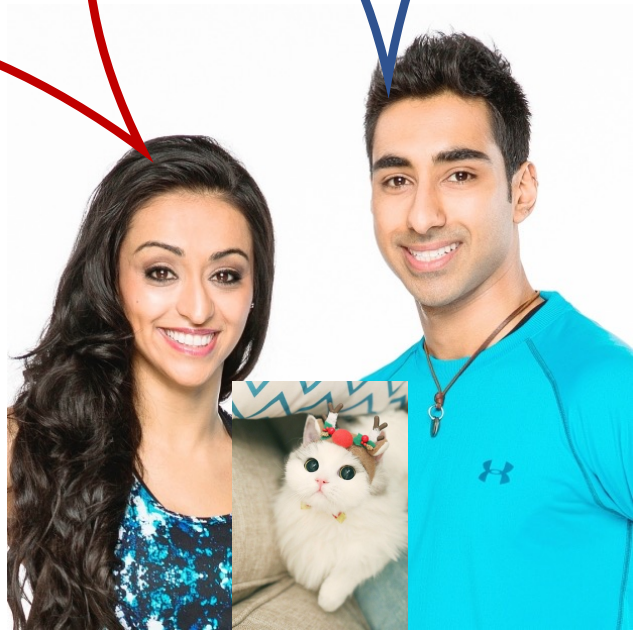
This is the sampling distribution of Sukhi's coin generated by Jinder (each value is the number of tails out of 30 tosses):

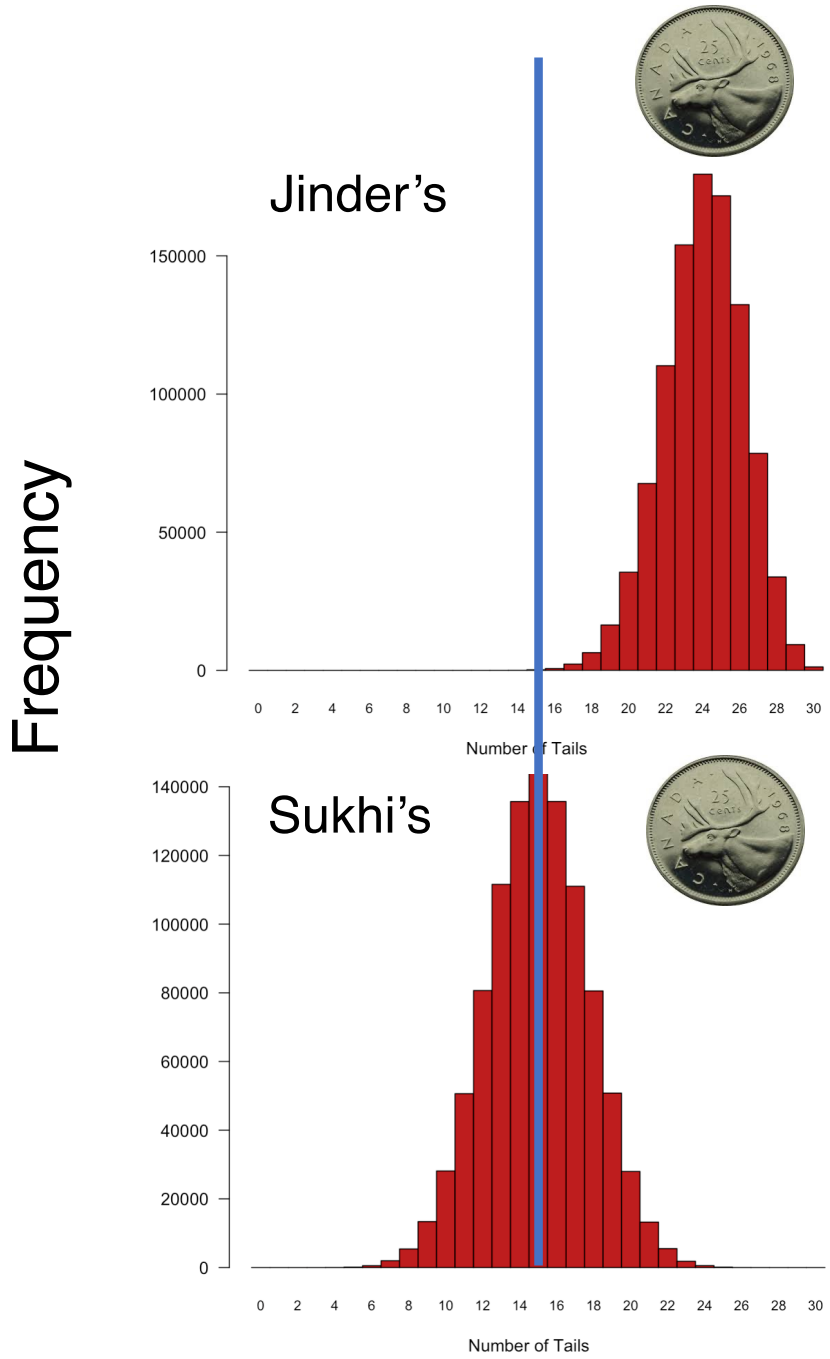


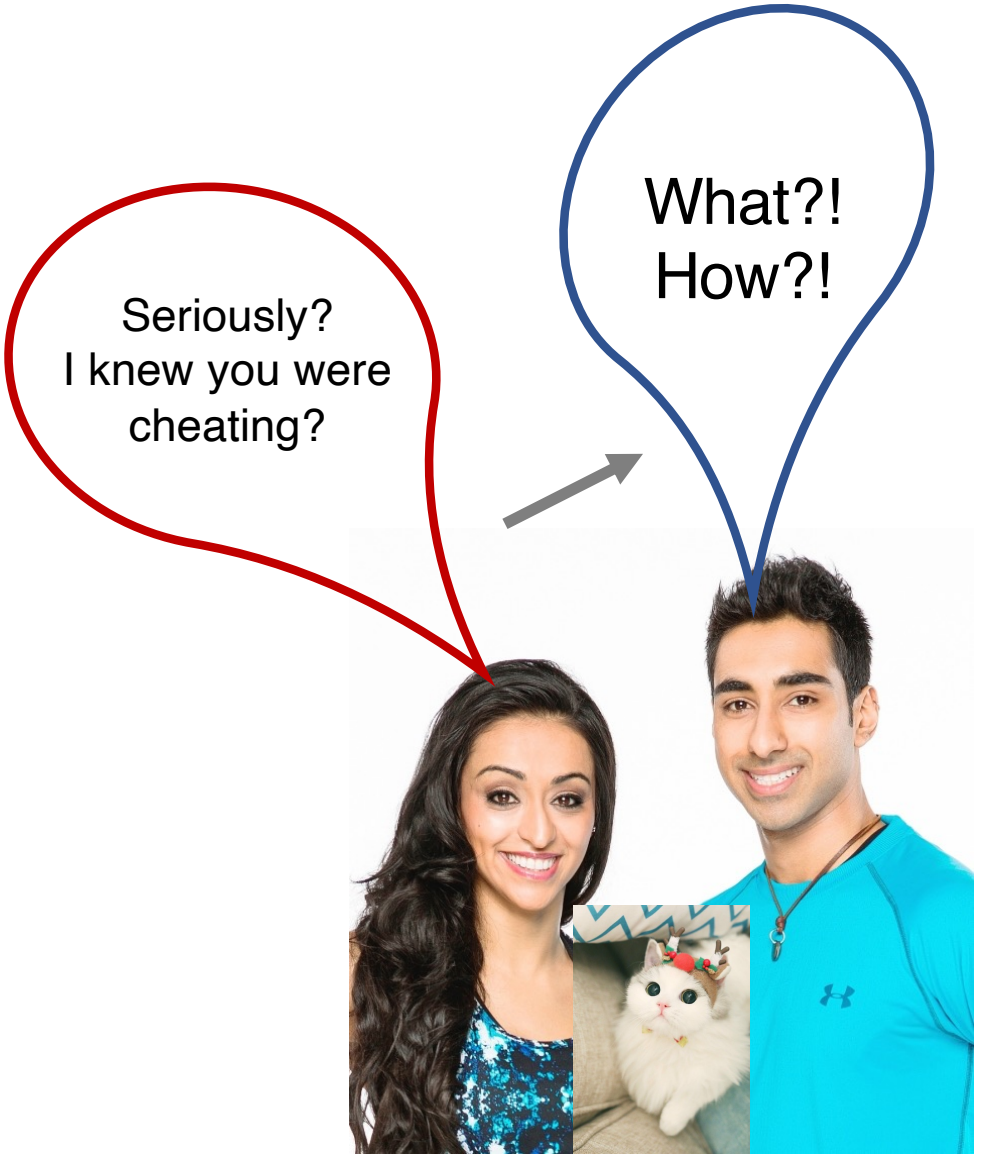
This is the sampling distribution of Jinder's coin generated by Sukhi (each value is the number of tails out of 30 tosses):







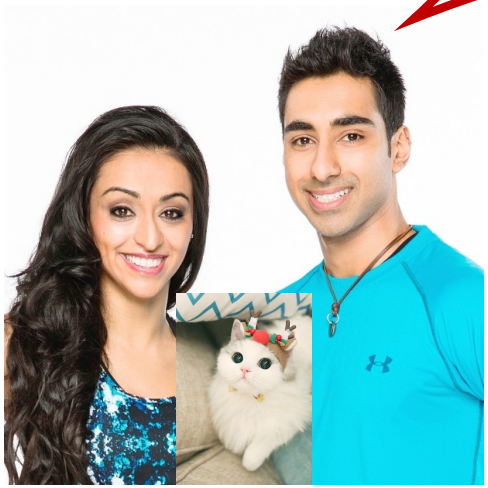

Jinder's
Frequency
150000
100000
50000
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30
Number of Tails
Sukhi's
140000
120000
100000
80000
60000
40000
20000
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30
Number of Tails


Seriously?
I knew you were
cheating?
What?!
How?!

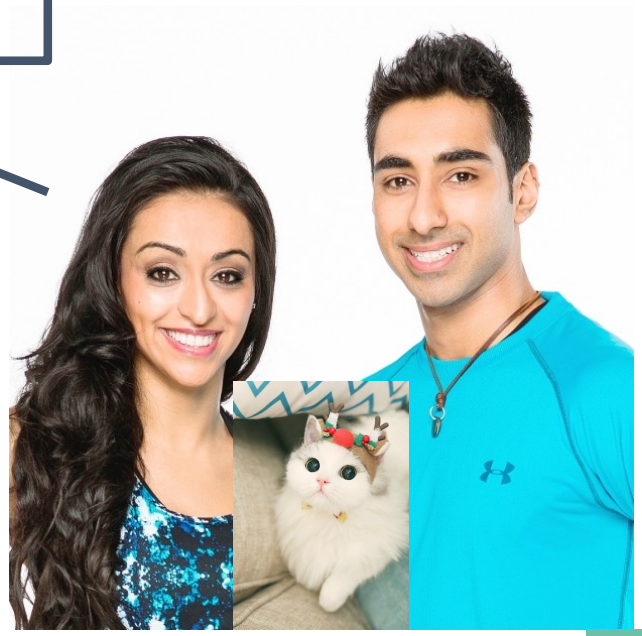
Ok, I'm sorry…I really want to take Mimy

We will use my coin then!!

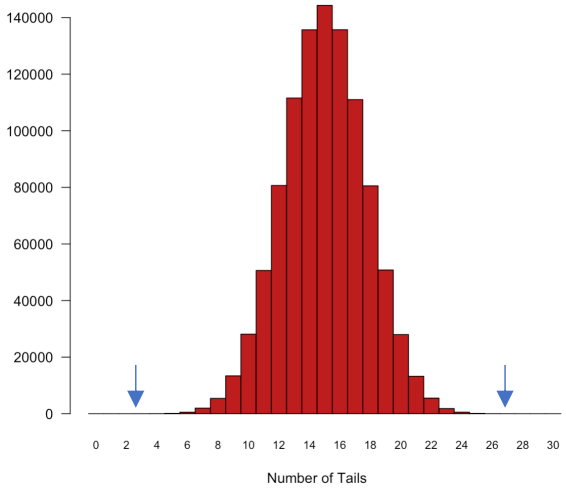
Nice, I got 27 tails!!! Coin was yours…and Mimy is mine!



Hummm……
The probability of obtaining 27 or more heads (or 27 or more tails) in 30 flips with a fair coin is approximately P=0.00000003 according to my fair coin!

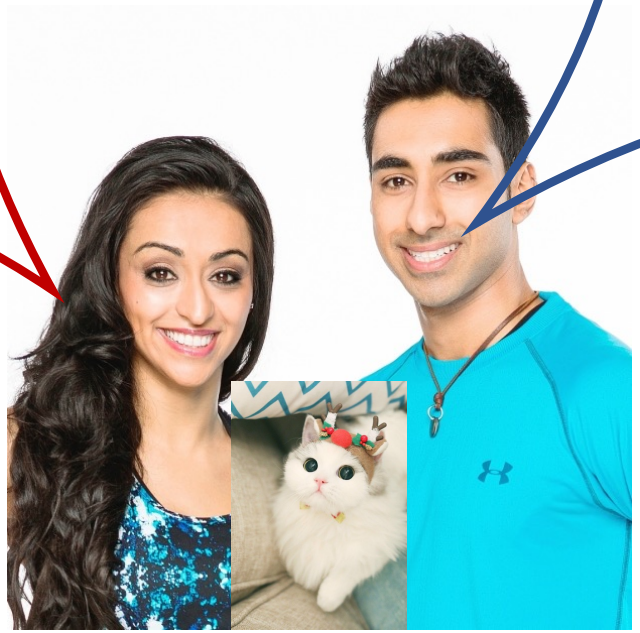


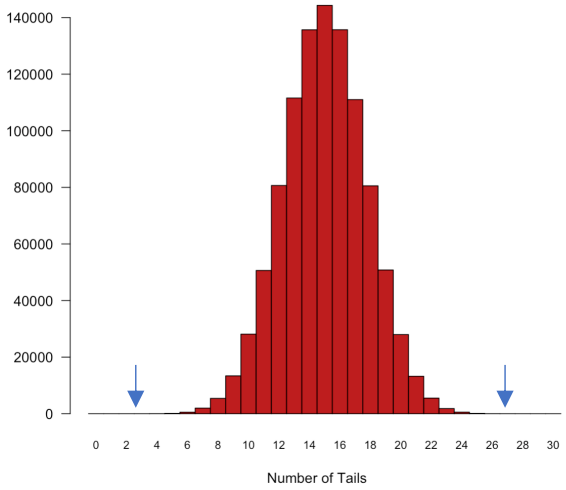
I'm just that lucky, I guess!



No luck, you must have cheated again!!!

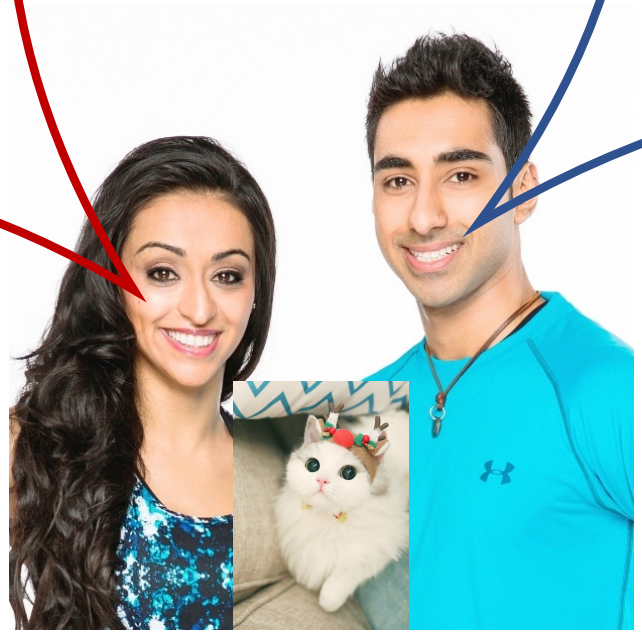
But how? Even in your 1,000,000 tosses there were 8 cases with 27 or more tails.

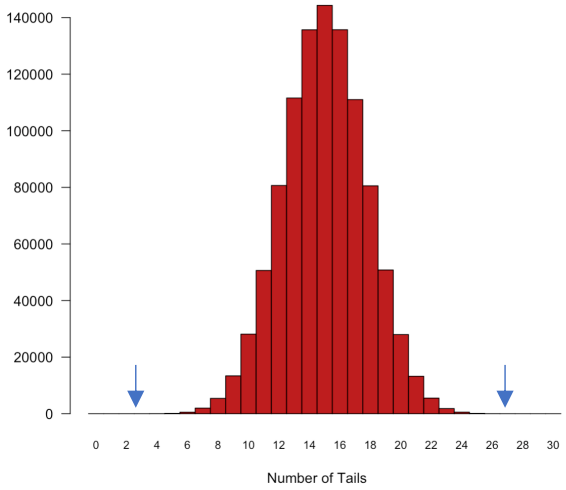




Yes, but it's almost impossible to get the result you got if you didn't cheat! So, I'm reasoning you did cheat.

Sorry… you're right… I did it again. You didn't see it, but I switched our coins without you noticing!





Take Mimy…if you cheated like this, I take that you really love and want Mimy!

Thanks sister!!!!!

These guys are nuts! I'm not going with either of them

The Core Problem: We Never Know the Truth

In real scientific problems, the **true state of nature is unknown**.

We do not know whether a hypothesis is actually true or false (e.g., Jinder's coin is or not); we only observe **sample data** and make a decision based on those data.

Statistics therefore provides a framework for **reasoning and making decisions under uncertainty**, not a way of directly discovering the truth.

The story of **Sukhi and Jinder deciding who gets the cat Mimy** illustrates this idea.

The Statistical Setup in the Story

Sukhi suspects that Jinder may be cheating with his coin. She therefore proposes a statistical experiment:

Null hypothesis (H_0): Jinder's coin is fair.
Alternative hypothesis (H_A): Jinder's coin is not fair (he cheated).

They flip the coin many times and observe the results. If the results are very unlikely under the assumption of a fair coin, Sukhi rejects the null hypothesis and concludes that Jinder probably cheated.

This mirrors the logic of hypothesis testing in science.

Decisions vs. Truth

In statistics, two things exist simultaneously:

[1] The real truth about the hypothesis (which we cannot observe).

[2] Our decision based on the data (reject or not reject H_o).

Because we do not know the truth, our decision can be correct or incorrect.

Reality (truth)	**Decision**	**Result**
H_o true	Do not reject H_o	Correct decision
H_o true	Reject H_o	**Type I error (false positive)**
H_o false	Reject H_o	Correct decision
H_o false	Do not reject H_o	**Type II error (false negative)**

Interpreting the Story with Error Types

The coin story helps visualize these possibilities.

Type I error (false positive):

This would occur if Jinder's coin were actually fair, but Sukhi rejects H_0 and accuses him of cheating.

Type II error (false negative):

This would occur if Jinder actually cheated, but the results of the experiment look normal enough that Sukhi does not reject H_0 and believes the coin is fair.

Thus, statistical testing cannot guarantee correct conclusions - it only controls the probability of these two possible mistakes (errors).

The story highlights a central idea of statistical inference:

Statistics does not reveal the truth directly. It provides a framework for making decisions about hypotheses while controlling the probability of making errors.

Knowledge advances not because we know the truth, but because we learn how to reason when the truth is unknown.

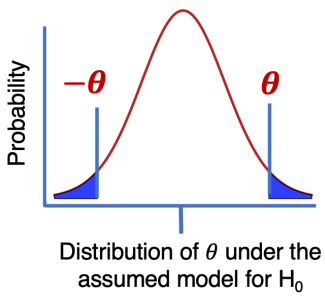
The Frequentist Hypothesis-Testing Framework

Statistical hypothesis testing is a **quantitative inference framework.**

It evaluates how **compatible the data are with an assumed model.**

That model is the **null hypothesis (H_0).**

core idea: we evaluate how surprising the observed data would be if the null hypothesis (H_0) were true.



Evidence for handedness in other animals

Humans are predominantly right-handed. *Do other animals exhibit handedness as well?* Bisazza et al. (1996) tested this possibility on the common toad.

They randomly sampled 18 wild toads, placed a balloon over each one's head, and recorded which forelimb the toads used to remove it to determine their preferred limb.

Translating the research question into a statistical question:

Do right-handed and left-handed toads occur with equal frequency in the (population, or is one more common than the other?

RESULTS: 14 toads were right-handed and four were left-handed. Do these results provide sufficient evidence to demonstrate handedness in toads?

Quantifying Statistical Evidence from Data (i.e., from samples)

If 14 right-handed is already strong evidence that the handedness (i.e., 50%/50% - akin to the fair coin) assumption may not hold, then 15, 16, 17 or 18 right-handed would represent even stronger evidence against handedness, and 4, 3, 2, 1 or 0 right-handed do the same in the opposite direction.

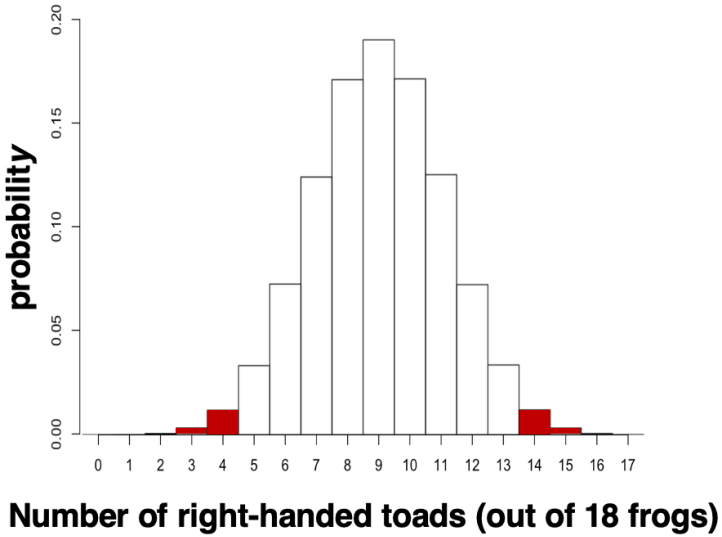
When assessing how surprising a result is, we therefore include all outcomes that contradict the model at least as much as the one we observed.

P-value =

P[0] + P[1] + P[2] + P[3] + P[4] +

P[14] + P[15] + P[16] + P[17] + P[18] =

0.0155 + 0.0155 = **0.031.**



The alpha level is the pre-established risk of a Type I error decided by the researcher (in Biology usually 0.05 or 0.01).

Assuming here an alpha = 0.05, we would reject the null hypothesis that frogs are equally handed.

Deciding on when to reject the null hypothesis

The significance level, denoted by α (alpha), is the threshold we set before analyzing the data to decide how much incompatibility with the null model we are willing to tolerate before rejecting it.

It represents the probability of rejecting the null hypothesis when it is actually true (more on this later).

α is not a law of nature or probabilities. It is a chosen threshold that reflects how cautious we want to be about claiming evidence against the null model.

In Biology, we usually apply 0.05 or 0.01 – tradition and consistency.

Statistical hypothesis testing *versus* estimation

Estimation asks - How large is the effect?

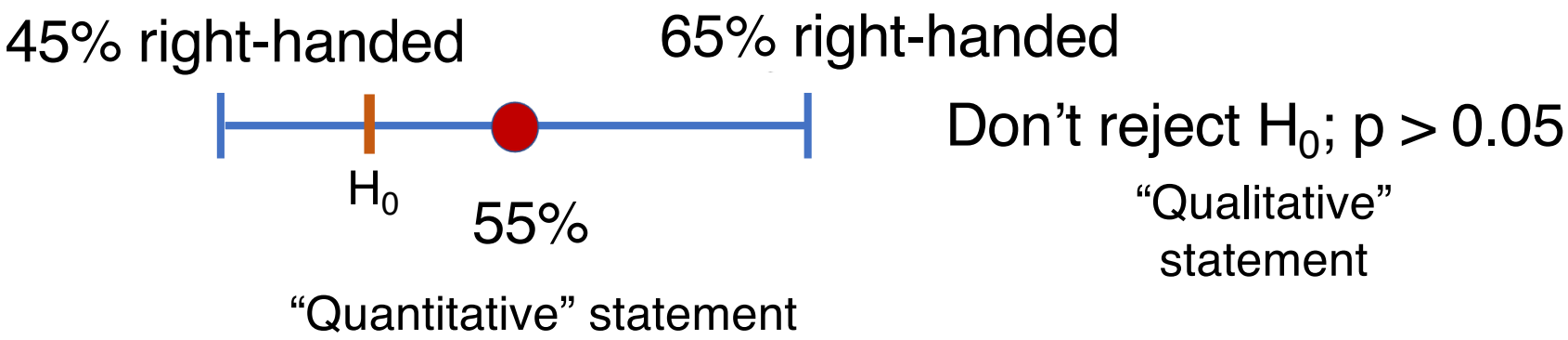
Hypothesis testing asks - Is there any effect at all?

Estimation would ask: What is the proportion of right- and left-handed toads in the population?

Statistical hypothesis testing would ask: Is there a statistically significant difference in the number of toads using their left or right limb to remove the balloon?

Statistical hypothesis testing does not focus on the exact proportion value but on whether there is evidence that the proportion differs from a specified value (commonly 50%/50%, though other values can be tested depending on the hypothesis).

Estimation & associated confidence intervals and statistical hypothesis testing always agree but have different interpretations



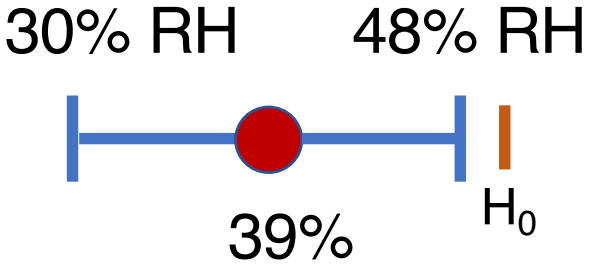
"*The purpose of statistical inference is to develop theory and methods to make inference on the unknown parameters based on observed data*" (Hong, 2017)

Estimation & associated confidence intervals and statistical hypothesis testing always agree but have different interpretations

"Quantitative" statement

Don't reject H_0; $p > 0.05$

"Qualitative" statement



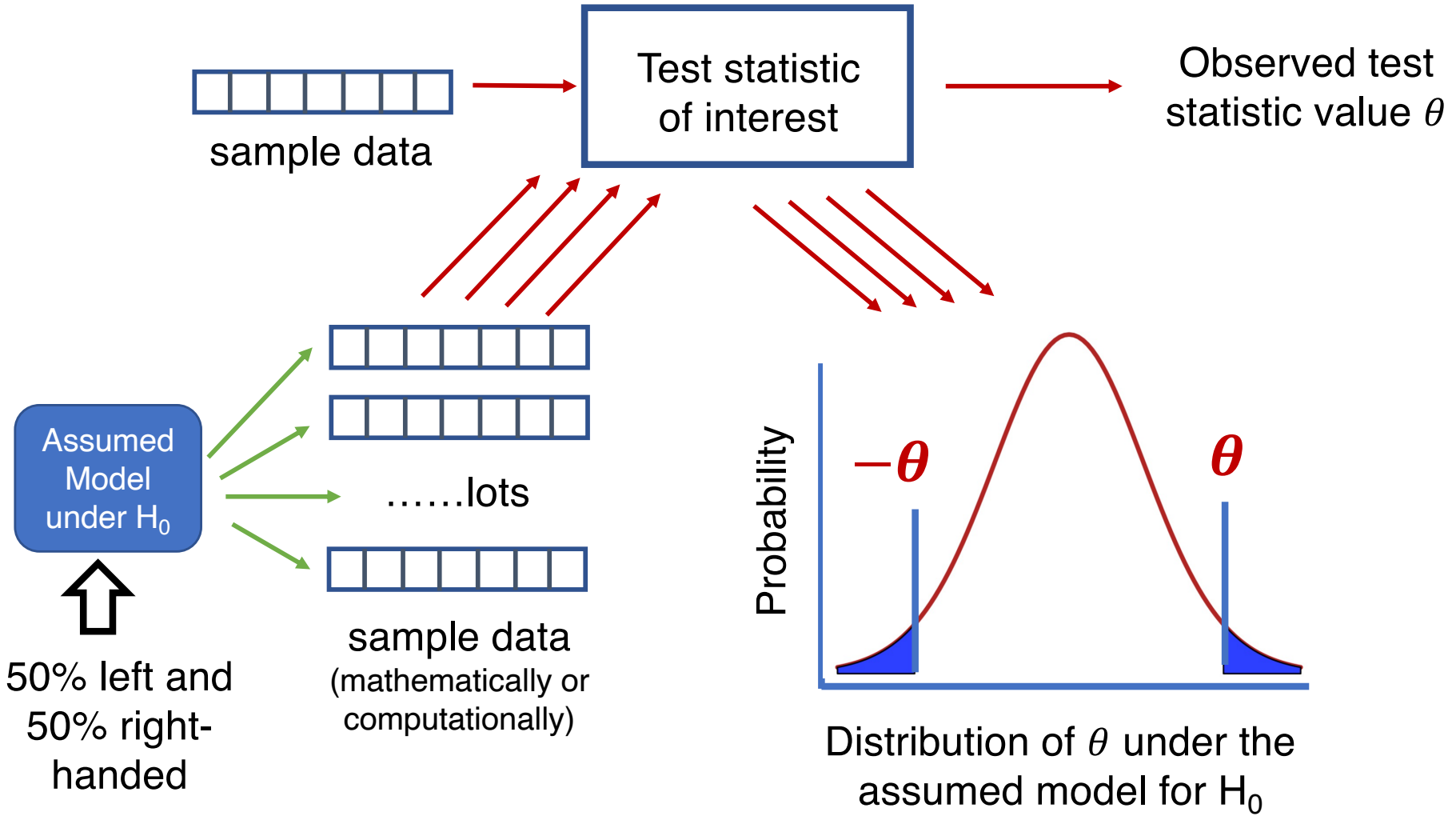
"Quantitative" statement

Reject H_0; $p < 0.05$

"Qualitative" statement

"The purpose of statistical inference is to develop theory and methods to make inference on the unknown parameters based on observed data" (Hong, 2017)

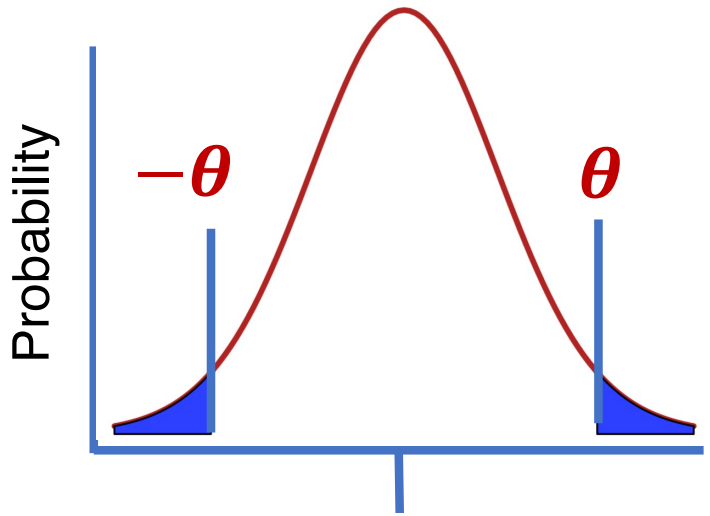
The “machinery” behind the framework of the frequentist statistical hypothesis testing



θ = observed number of right-handed toads in the sample

$-\theta$ = observed number of left-handed toads in the sample

Why we are the ones that set type I error (alpha) but not type II error (beta)



Distribution of θ under the assumed model for H_0

50% left and 50% right-handed

The **p-value is the total probability of those shaded tail regions calculated by assuming** H_0 as true.

SO: The p-value is the probability, calculated under the assumed null hypothesis (H_0), of observing a value of the test statistic (θ) as extreme as, or more extreme than, the one actually observed.

$A = \pi r^2$

θ = number of right-handed toads equal or larger than the observed

$-\theta$ = number of left-handed toads smaller or larger than the observed

Quantifying Statistical Evidence from Data (i.e., from samples)

If 14 right-handed is already strong evidence that the handedness (i.e., 50%/50% - akin to the fair coin) assumption may not hold, then 15, 16, 17 or 18 right-handed would represent even stronger evidence against handedness, and 4, 3, 2, 1 or 0 right-handed do the same in the opposite direction.

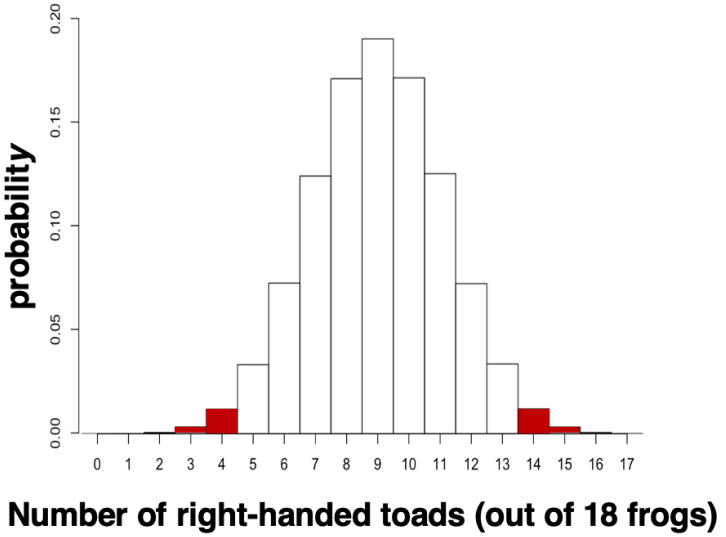
When assessing how surprising a result is, we therefore include all outcomes that contradict the model at least as much as the one we observed.

P-value =

P[0] + P[1] + P[2] + P[3] + P[4] +

P[14] + P[15] + P[16] + P[17] + P[18] =

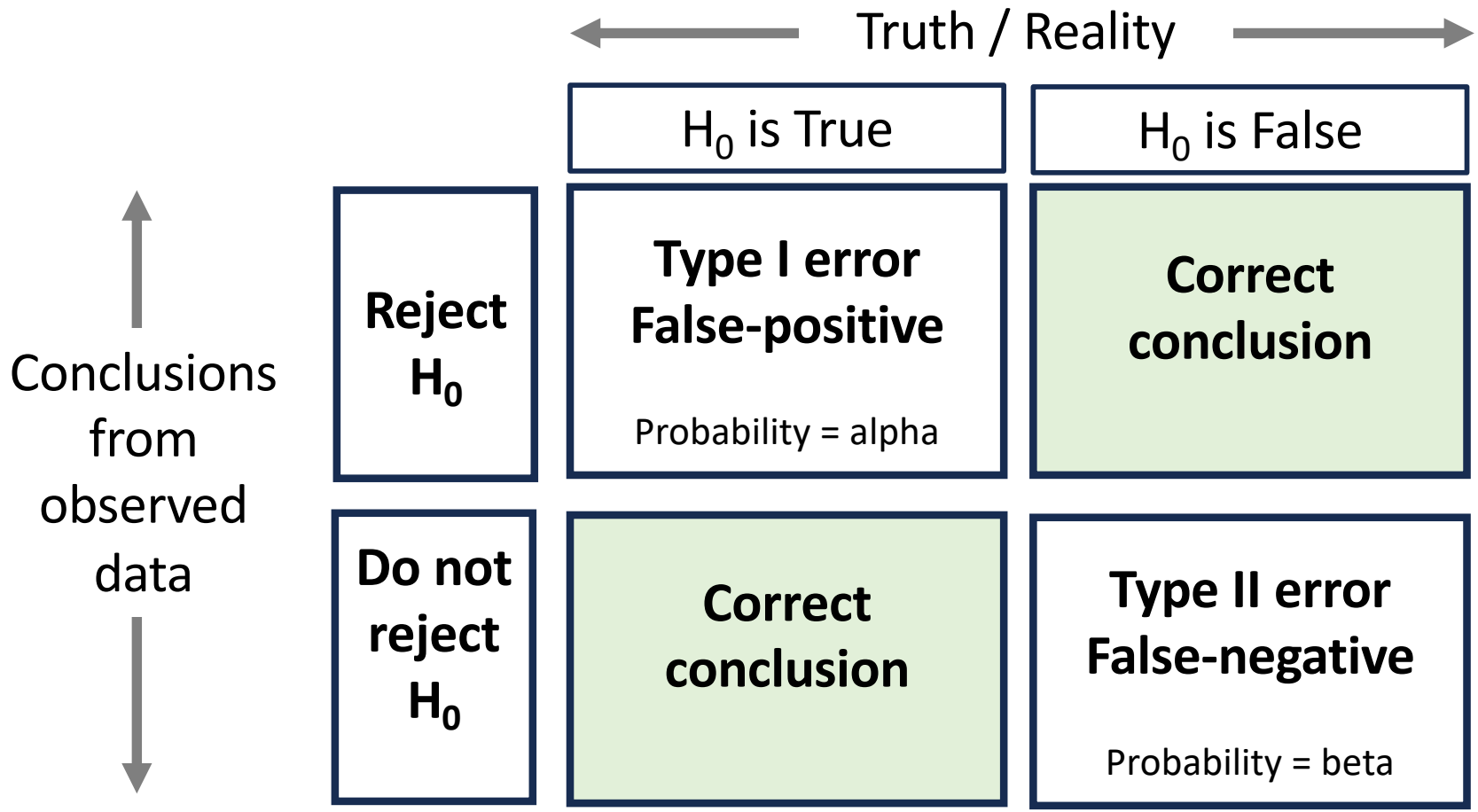
0.0155 + 0.0155 = **0.031.**



The alpha level is the pre-established risk of a Type I error decided by the researcher (in Biology usually 0.05 or 0.01).

Assuming here an alpha = 0.05, we reject the null hypothesis that frogs are equally handed.

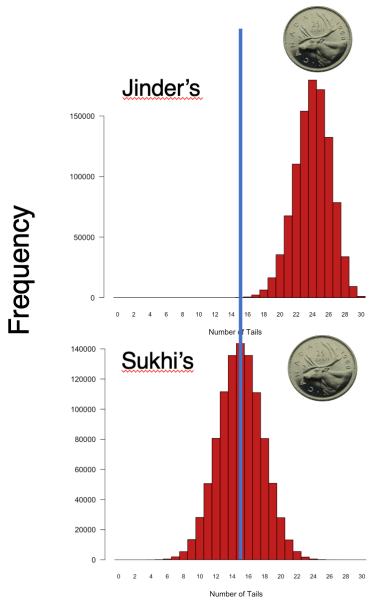
The key idea is that **statistics does not reveal the truth directly**. Instead, it provides a framework for **making decisions under uncertainty**, while controlling the probability of two errors.

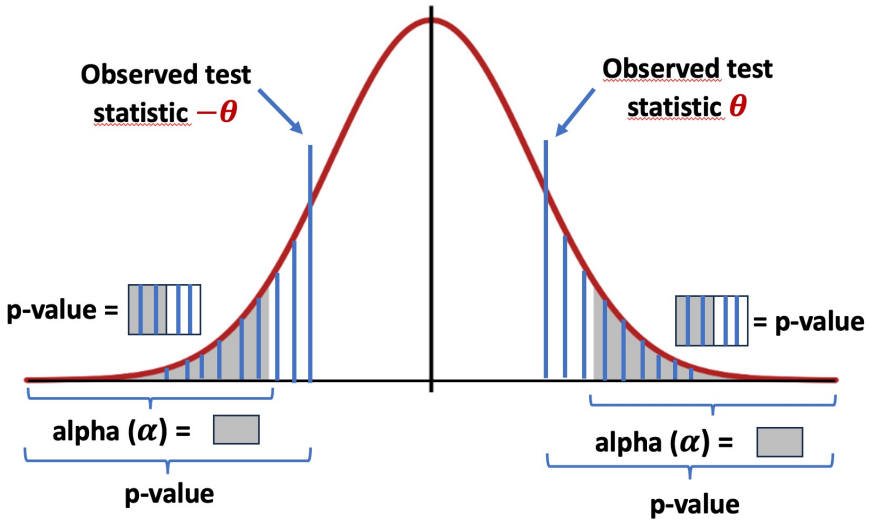


Knowledge advances not because we know the truth (unobservable), but because we learn how to reason when the truth is unknown.

Why do we set the Type I error rate (α) but not the Type II error rate (β)?

Because α is defined from the probability distribution under the assumption that H_0 is true, which we can calculate, i.e., there is only one value under the H_0 and many possible (infinite) under the H_A.

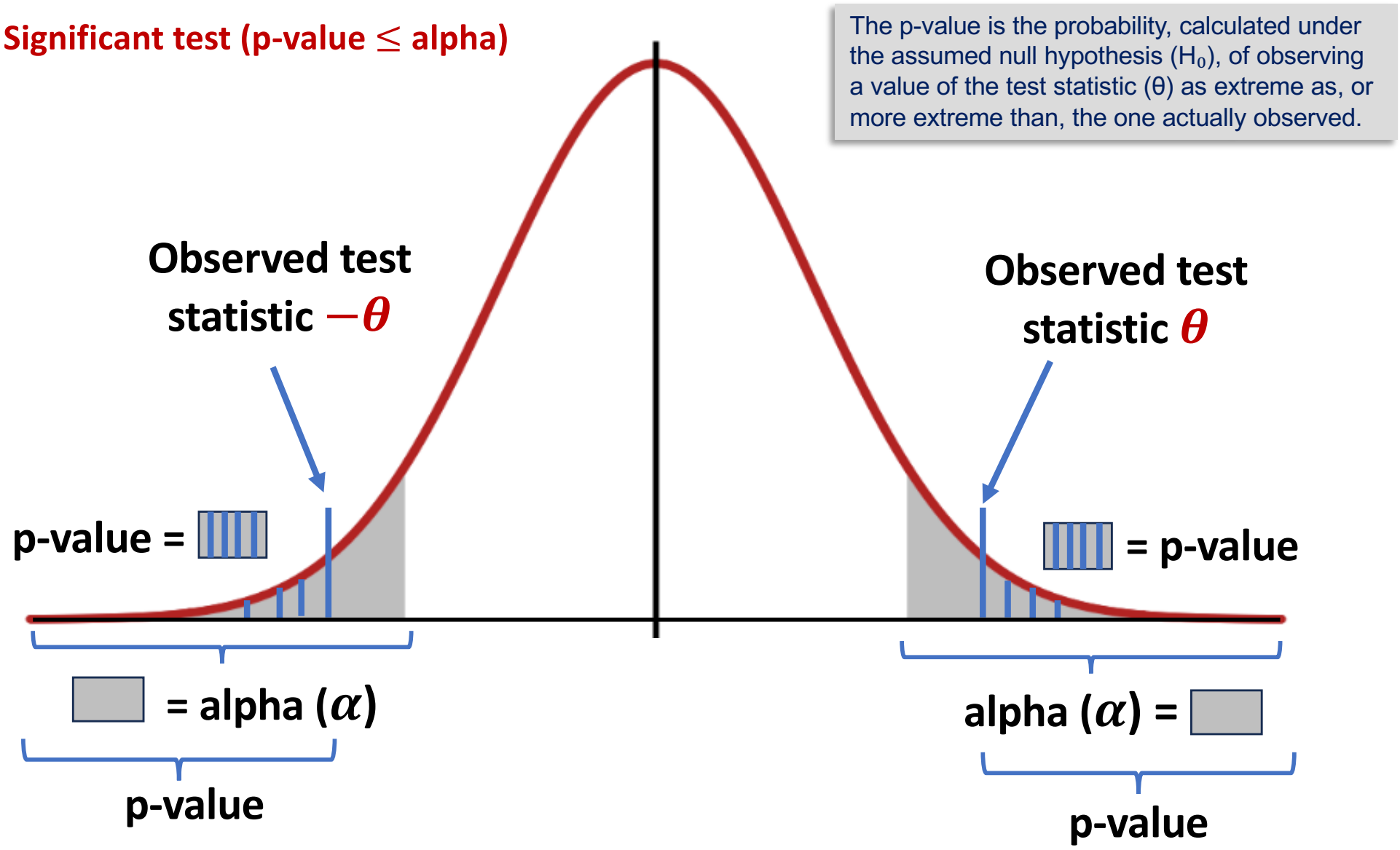




α defines the proportion of outcomes in the probability distribution (assuming H_0 is true) that we decide to treat as "too unlikely by chance."

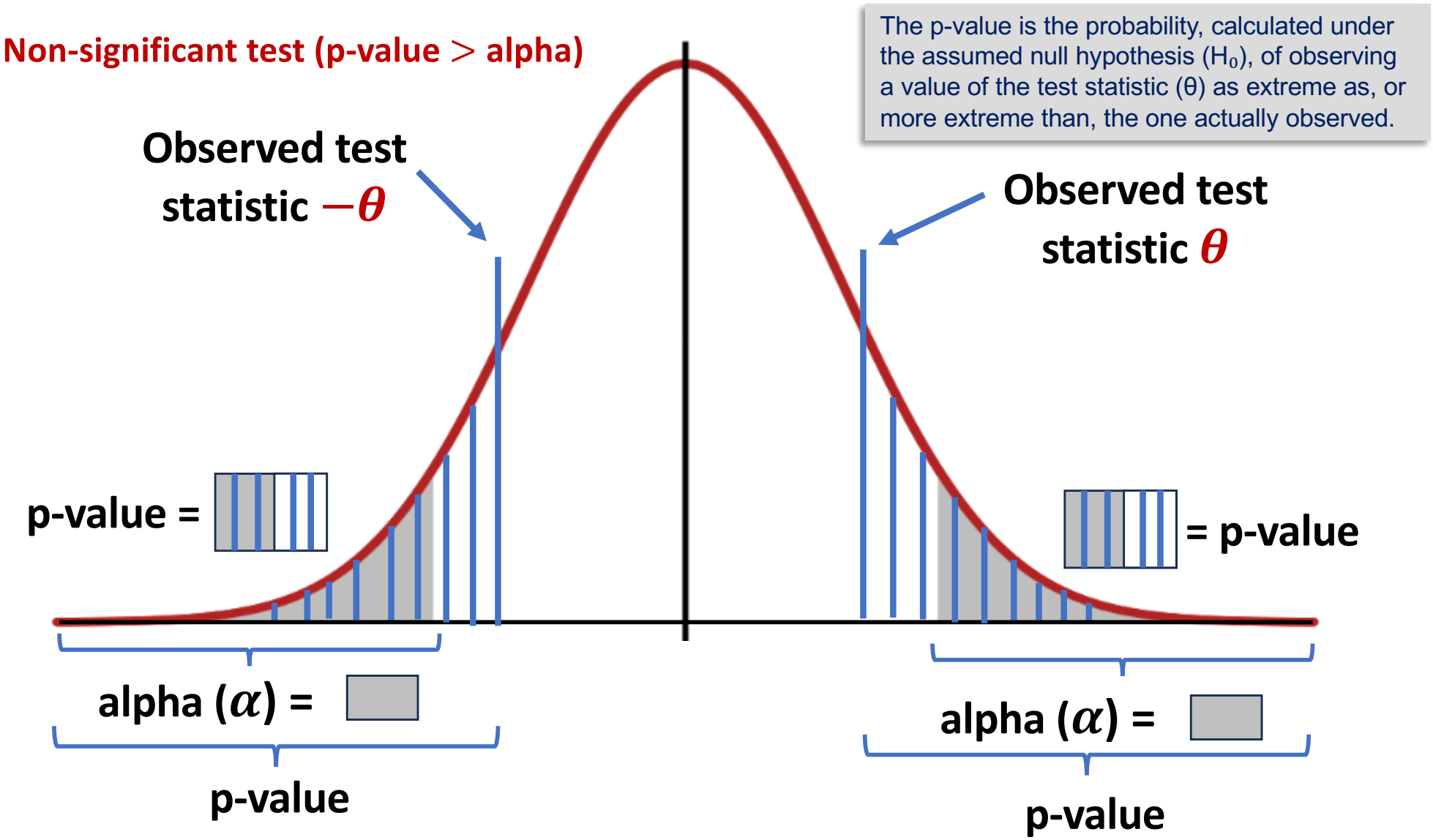
Because α is defined from the probability (sampling) distribution assuming H_0 is true, it represents the probability of committing a Type I error (a false positive); that is, rejecting H_0 when H_0 is actually true.

Alpha, p-values, and statistical decisions



α and p-values are probabilities represented by areas under the probability distribution curve of the test statistic.

Alpha, p-values, and statistical decisions



α and p-values are probabilities represented by areas under the probability distribution curve of the test statistic.

Summary

In statistical hypothesis testing, we construct the sampling distribution under the assumption that the null hypothesis (H_0) is true.

This means that all values in that distribution, including the observed sample value, are outcomes that could occur if H_0 were true.

If the observed sample value lies in a region of the distribution that is sufficiently unlikely under H_0, we conclude that the result is improbable under the null model and reject H_0.

However, because our decision is based on sample data rather than the true state of nature, there is always a possibility of making a mistake, whether we reject or fail to reject H_0.

Summary

In statistical hypothesis testing, we construct the sampling distribution under the assumption that the null hypothesis (H_0) is true.

This means that all values in that distribution, including the observed sample value, are outcomes that could occur if H_0 were true.

If the observed sample value lies in a region of the distribution that is sufficiently unlikely under H_0, we conclude that the result is improbable under the null model and reject H_0.

However, because our decision is based on sample data rather than the true state of nature, there is always a possibility of making a mistake, whether we reject or fail to reject H_0.

The protection against incorrectly rejecting a true null hypothesis (a Type I error) is determined by the chosen alpha level (the significance level).

Reducing the probability of failing to reject a false null hypothesis (Type II error) typically requires increasing statistical power, often by increasing the sample size.

Critical definitions

A Type I error occurs when a true null hypothesis is incorrectly rejected (i.e., rejecting the null hypothesis when it should not be rejected). Its probability is the significance level (α), which is determined by us and remains unaffected by the sample size (n).

Type II error is failing to reject a false null hypothesis (i.e., do not reject the null hypothesis when you should not have). Its probability is β and is more complex to estimate (advanced stats). This probability decreases as sample size increases.

The power of a test (1 - β) is the probability of correctly rejecting the null hypothesis when it is truly false. This probability increases as the sample size grows.

Hypothesis testing involving a continuous variable; the toad problem involved a categorial variable

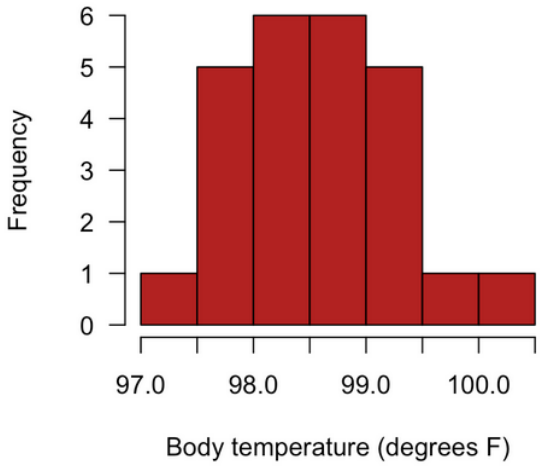
Normal human body temperature, as kids are taught in North America, is 98.6^{o}F (37^{o}C). But how well is this supported by data?

Let's understand this problem under a statistical hypothesis testing framework

Normal human body temperature, as kids are taught in North America, is 98.6°F. But how well is this supported by data? Researchers obtained body-temperature measurements on randomly chosen healthy people (Schoemaker 1996). The data for the 25 people are as follows:

98.4	98.6	97.8	98.8	97.9
99.0	98.2	98.8	98.8	99.0
98.0	99.2	99.5	99.4	98.4
99.1	98.4	97.6	97.4	97.5
97.5	98.8	98.6	100.0	98.4

The data looks relatively symmetric so for now we have a good indication that these data are "normally" distributed. We'll see later in the course how to test this assumption in a more rigorous way.



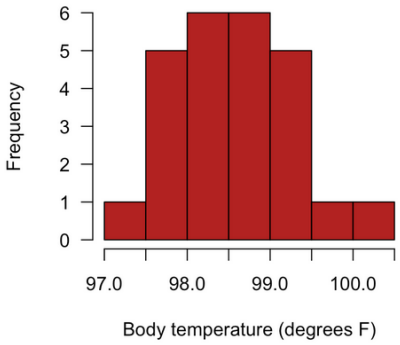
Normal human body temperature, as kids are taught in North America, is 98.6°F. But how well is this supported by data?

Let's understand this problem under a statistical hypothesis testing framework

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 Probability (or P-value), or estimated probability, is the probability of finding the observed, or more *extreme*, assuming that the null hypothesis (H_0) related to a study question (μ = 98.6°F) is true.

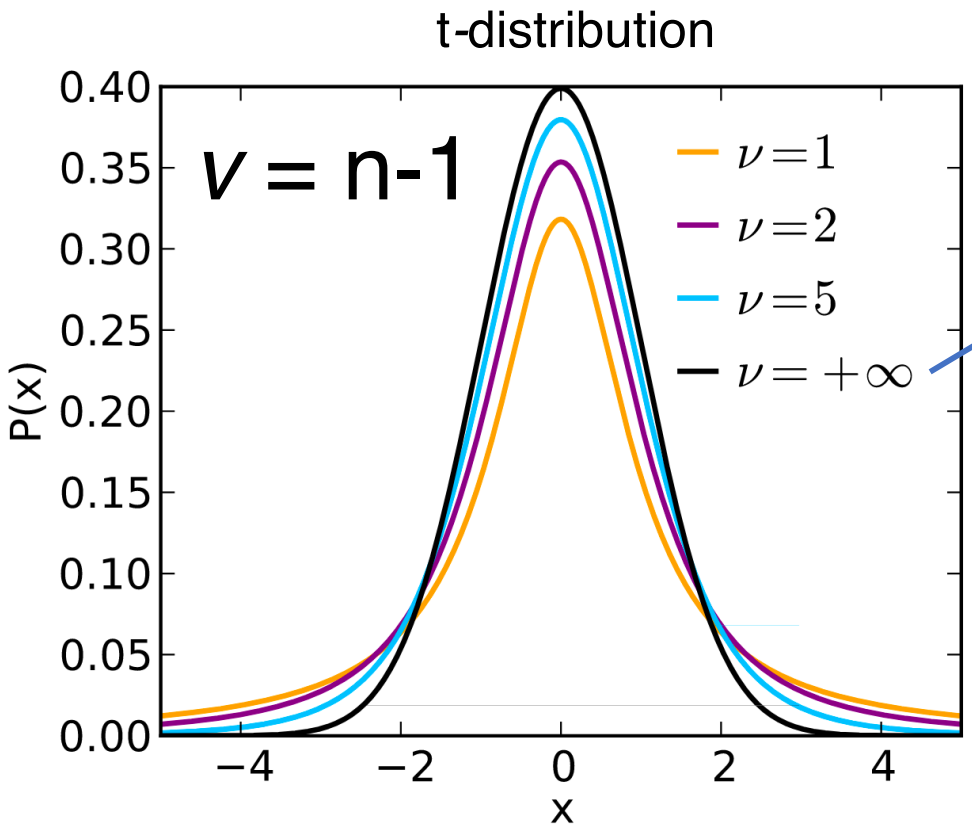


$$\bar{Y} = 98.524$$

H_0 can be stated as "any observed difference between the sample mean (98.524°F) and the theoretical population value (98.6°F) is due to chance alone.

To address the human body temperature example, we will use the t distribution (quick review below).

For small sample sizes (n < 60), the sampling distribution of sample means from a normally distributed population are a bit far from normally distributed – the distribution is wider and depends on the sample size (degrees of freedom) – it's called the t-distribution.



Normally distributed

Read:

BIOL322 - Building the t-Distribution from First Principles

Pedro Peres-Neto, Biology, Concordia University

observed data

test statistic of interest (here t statistic)

± observed t-value

$$t_i = \frac{\bar{Y}_i - \mu}{\frac{s_i}{\sqrt{n}}} = \frac{\bar{Y}_i - 98.6}{\frac{s_i}{\sqrt{n}}}$$

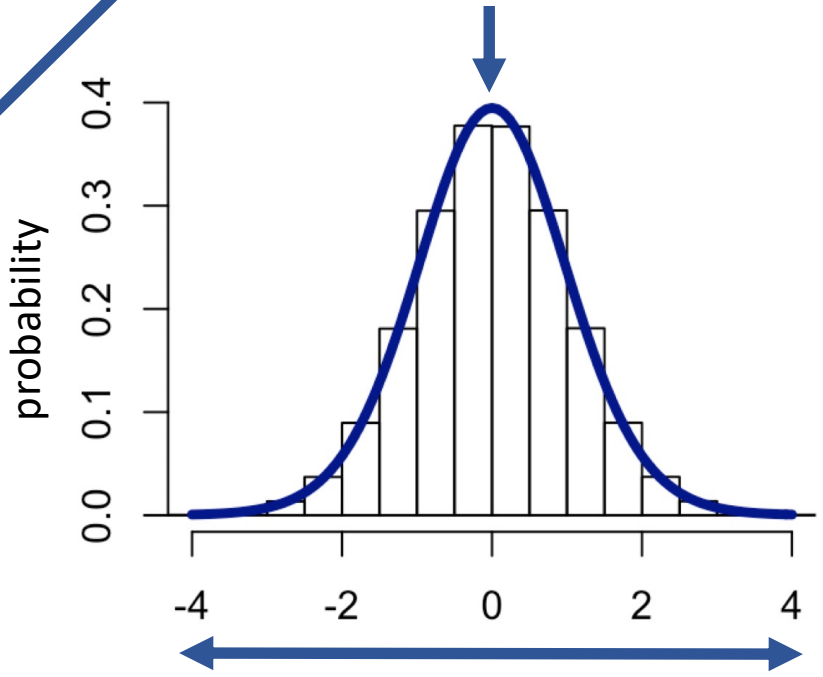
model assuming H_0 as true (mu=98.6F)

...

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=25). Let's assume here μ = 98.6 and σ = 5.

sampling distribution of t values under H_0

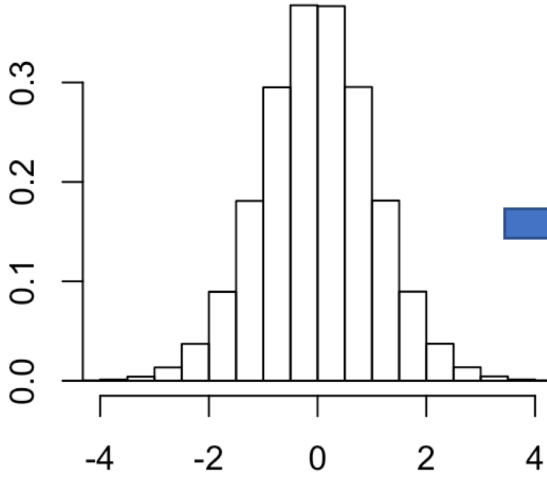


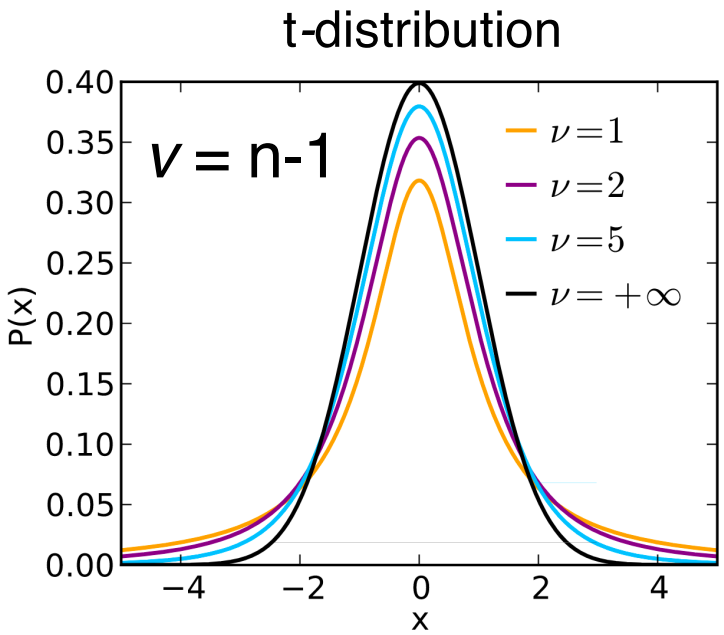
Number of standard deviations away from the theoretical parameter assumed under H_0

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

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=25).

```r
samples.pop.1 <- replicate(1000000,rnorm(n=25,mean=98.6,sd=5))
sampleMeans.Pop1 <- apply(samples.pop.1,MARGIN=2,FUN=mean)
sampleSDs.Pop1 <- apply(samples.pop.1,MARGIN=2,FUN=sd)
standardized.SampDist.Pop1 <- (sampleMeans.Pop1-98.6)/(sampleSDs.Pop1/sqrt(25))
hist(standardized.SampDist.Pop1)
```

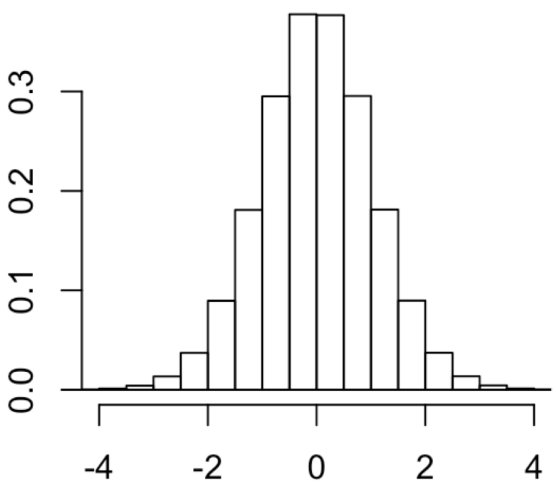




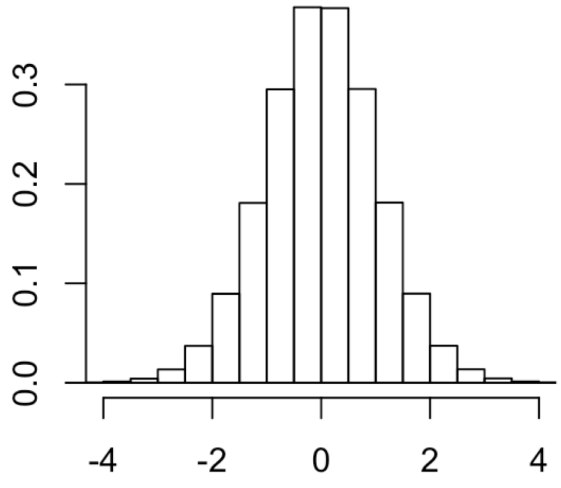
Two populations with different means and standard deviations, but samples drawn from each population have the same size n.

```r
samples.pop.1 <- replicate(1000000,rnorm(n=25,mean=98.6,sd=5))
sampleMeans.Pop1 <- apply(samples.pop.1,MARGIN=2,FUN=mean)
sampleSDs.Pop1 <- apply(samples.pop.1,MARGIN=2,FUN=sd)
standardized.SampDist.Pop1 <- (sampleMeans.Pop1-98.6)/(sampleSDs.Pop1/sqrt(25))
hist(standardized.SampDist.Pop1)
```

```r
samples.pop.2 <- replicate(1000000,rnorm(n=25,mean=13,sd=15))
sampleMeans.Pop2 <- apply(samples.pop.2,MARGIN=2,FUN=mean)
sampleSDs.Pop2 <- apply(samples.pop.2,MARGIN=2,FUN=sd)
standardized.SampDist.Pop2 <- (sampleMeans.Pop2-13)/(sampleSDs.Pop2/sqrt(25))
hist(standardized.SampDist.Pop2)
```



=



Because of standardization, the t-distribution has a mean of 0, and its spread depends only on the sample size (via the degrees of freedom).

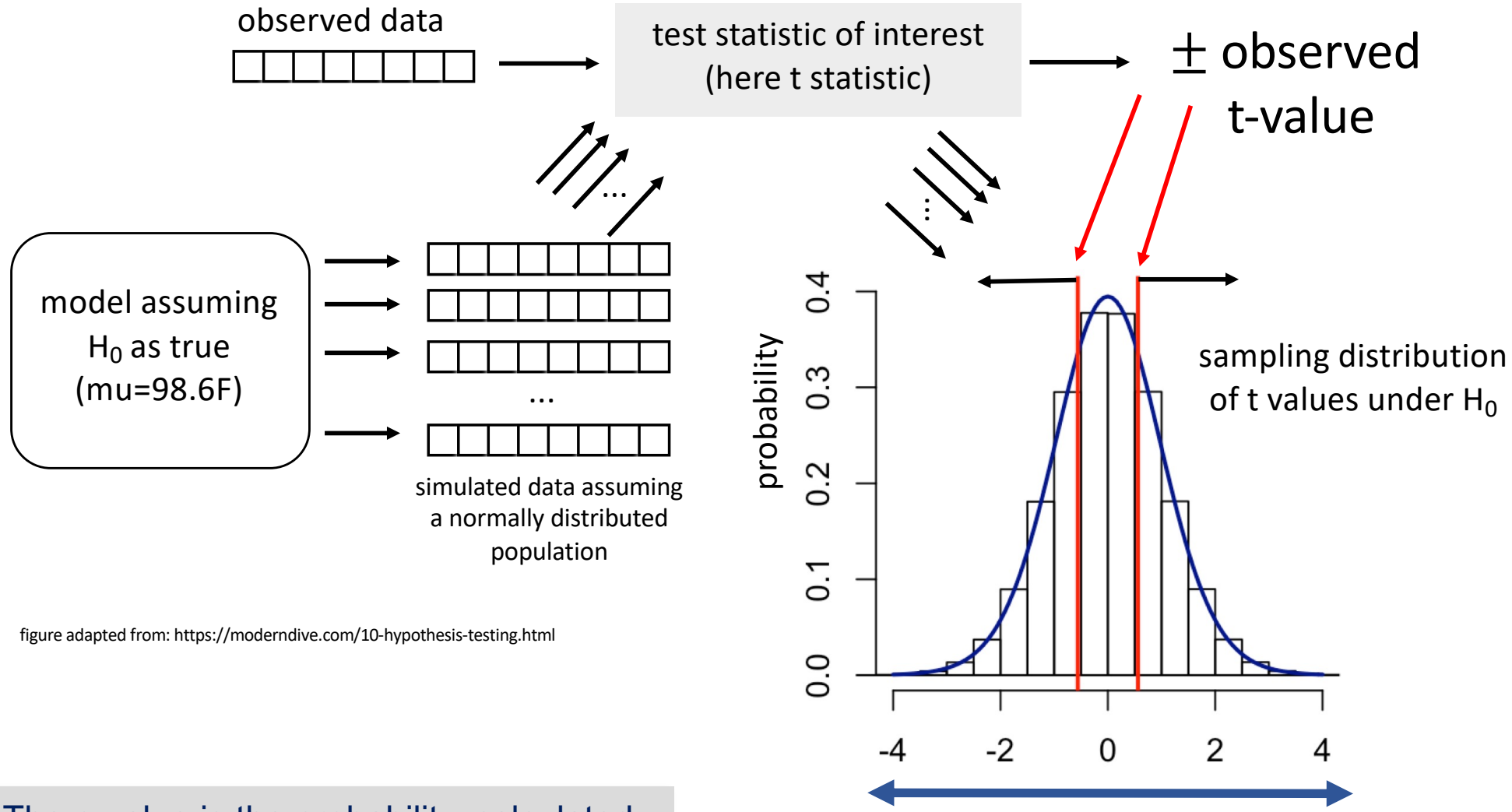


figure adapted from: https://moderndive.com/10-hypothesis-testing.html

The p-value is the probability, calculated under the assumed null hypothesis (H_0), of observing a value of the test statistic (θ) as extreme as, or more extreme than, the one actually observed.

t-values = the number of standard errors that the sample estimate is away from the theoretical parameter assumed under H_0

$$t_i = \frac{\bar{Y}_i - \mu}{\frac{s_i}{\sqrt{n}}}$$

The data appear relatively symmetric, suggesting that the assumption of normality may be reasonable. Later in the course, we will learn more rigorous methods for formally testing this assumption.

We assume a normally distributed population when deriving the t-distribution, either through mathematical theory (infinite sampling) or simulation. If the original population is not normally distributed, the standardized sampling distribution of the sample mean may also deviate from the expected t-distribution, and the test may not perform as assumed (as discussed earlier for confidence intervals).

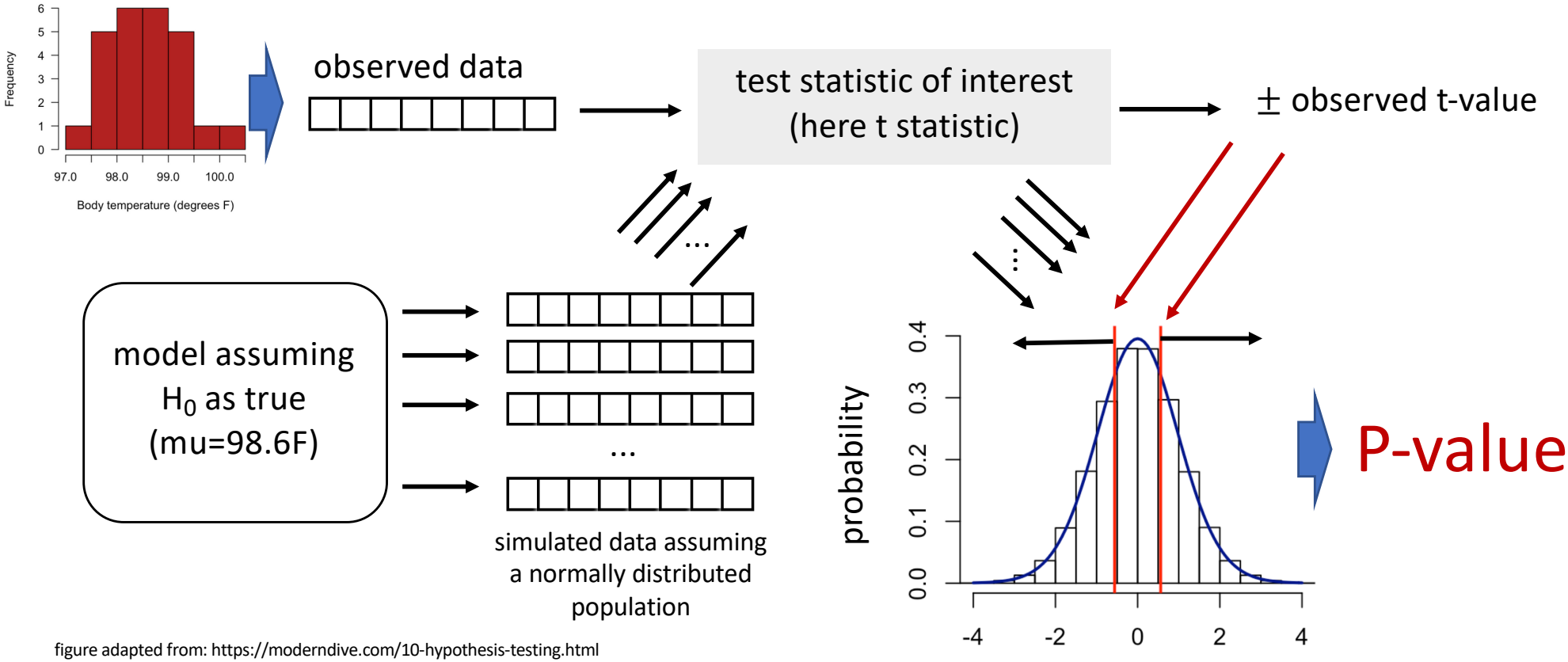


figure adapted from: https://moderndive.com/10-hypothesis-testing.html

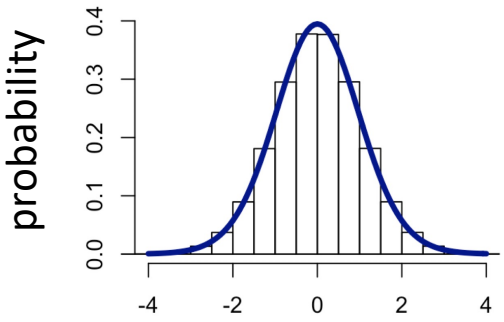
The use of probability distribution functions for continuous variables

Unlike discrete variables (e.g., handedness in toads), the t-distribution is used for continuous variables (e.g., temperature) and is described by a probability density function (pdf). Because the probability of any exact value is zero, probabilities are calculated as the area under the curve between two values.

Suppose a species of bacteria typically lives between 4 and 6 hours. What is the probability that a bacterium lives exactly 5 hours? For a continuous variable, the probability of any exact value is 0. Many bacteria may live approximately 5 hours, but none die at exactly 5.000000… hours.

Instead, probabilities are defined over ranges of values, such as the probability that a bacterium dies between 5.00 and 5.01 hours. The same applies to human body temperature and any other continuous variable.

The bacteria example is adapted from Wikepedia



One sample t-test

Normal human body temperature, as kids are taught in North America, is 98.6ºF. But how well is this supported by data? Researchers obtained body-temperature measurements on randomly chosen healthy people (Schoemaker 1996).
The data for the 25 people are as follows:

98.4	98.6	97.8	98.8	97.9
99.0	98.2	98.8	98.8	99.0
98.0	99.2	99.5	99.4	98.4
99.1	98.4	97.6	97.4	97.5
97.5	98.8	98.6	100.0	98.4

$\bar{Y} = 98.524$ $s = 0.678$ $\mathrm{SE}_{\bar{Y}} \frac{0.678}{\sqrt{25}} = 0.136$

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

Let's "transform" this question into a probabilistic statement:

$$t_i = \frac{\bar{Y}_i - 98.6}{\frac{s_i}{\sqrt{n}}}$$

What is the probability of obtaining a sampling *mean as extreme or more extreme* than 98.524°F given that the theoretical population mean (assumed under H_0) is 98.6°F?

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

The sample mean is -0.56 standard deviations away from the mean of the theoretical population (assumed under H_0)!

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

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

Should we reject or not reject the H_0?

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

The sample mean is -0.56 standard deviations away from the mean of the theoretical population (assumed under H_0)!

In probabilistic terms, the question becomes: What is the probability of observing a sample t-value equal to or more extreme than ±0.56 in the sampling distribution of the theoretical population (i.e., the t-distribution, where $\mu = 0$)?

$$P(t \leq -0.56 \text{ or } t \geq 0.56) = 0.58$$

$$\text{or } P(|t| \geq 0.56 = 0.58$$

The probability of observing a t-value at least as extreme as ±0.56 is 0.58.

The Procedure for a One-Sample Mean Test

1. Specify the mean under H_0.

2. Take a sample from the population of interest.

3. Standardize the sample mean using the t-statistic.

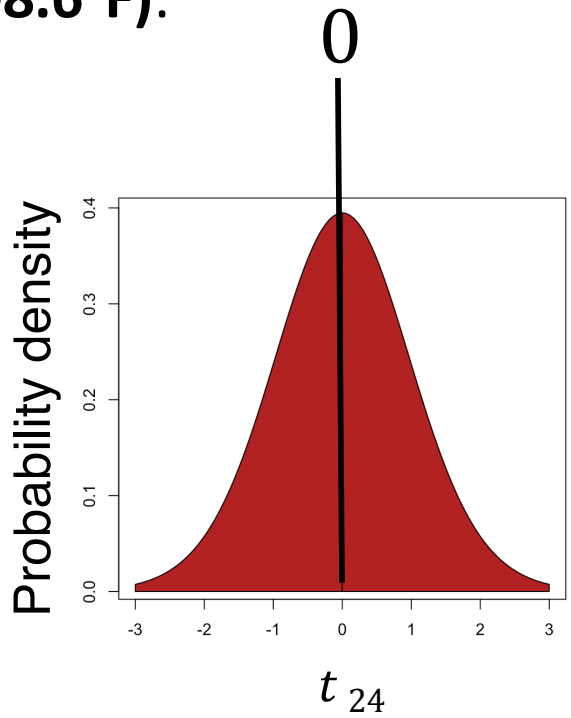
4. Calculate the probability of observing the t-value or a more extreme value in the t-distribution.

5. Compare the probability with alpha and make a decision about H_0.

Imagine the most compatible result possible: If the sample mean were **exactly equal to the null hypothesis (98.6°F)**:

$$t = \frac{98.6 - 98.6}{0.136} = 0$$

This is the **least surprising outcome** under H_0.
Therefore: **p-value = 1.00**



What does that mean? The data are **completely compatible with** H_0.
But even in this case: **We still cannot conclude that** H_0 **is true.**
We can only say: **The data provide no evidence against it.**

Statistical tests do not prove hypotheses, they only evaluate whether the data contradict them.

Summary

We started with: Is the normal human body temperature of 98.6^{o}F, as kids are taught in North America, supported by data?

Then we "translated" the above question into: What is the probability of obtaining a sampling mean as extreme or more than a sample mean of 98.524^{o}F given that the population mean is 98.6^{o}F?

1) In principle, we were not interested in knowing if the sample mean we obtained would be smaller or greater than the true population mean.

2) As such, all we are interested is to state whether we have evidence to say that the sample mean we obtained is ***consistent*** with H_0 or ***inconsistent*** with H_0.

3) If **consistent** (large P-value), then we can state that we have no evidence to state that the human temperature is different from 98.6^{o}F.

4) If **inconsistent** (small P-value), then we would have stated that we have evidence that the *Normal human body temperature is not 98.6^{o}F.*