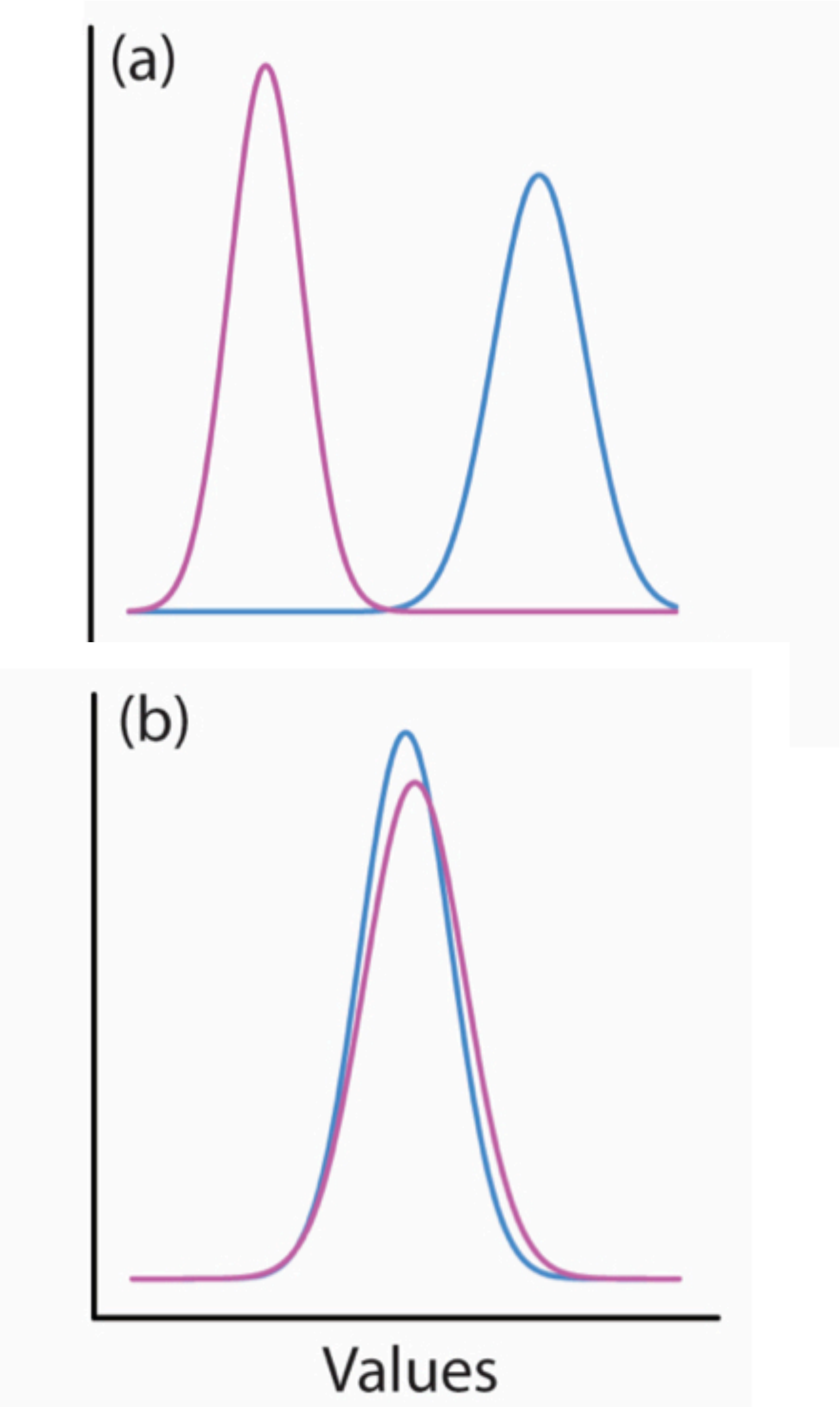
Foundational Statistics

Statistical Hypothesis Tests



STATISTICAL CONCLUSION

	Reject H_0	Retain H_0
Effect	Correct decision Effect detected	Type II error Effect not detected
No effect	Type I error Effect detected; none exists	Correct decision No effect detected; none exists

broad-sense hypotheses
(a.k.a. “scientific hypotheses”)

overarching explanations for specific phenomena

Measurable observed phenomenon

Heritable traits change over time in populations because

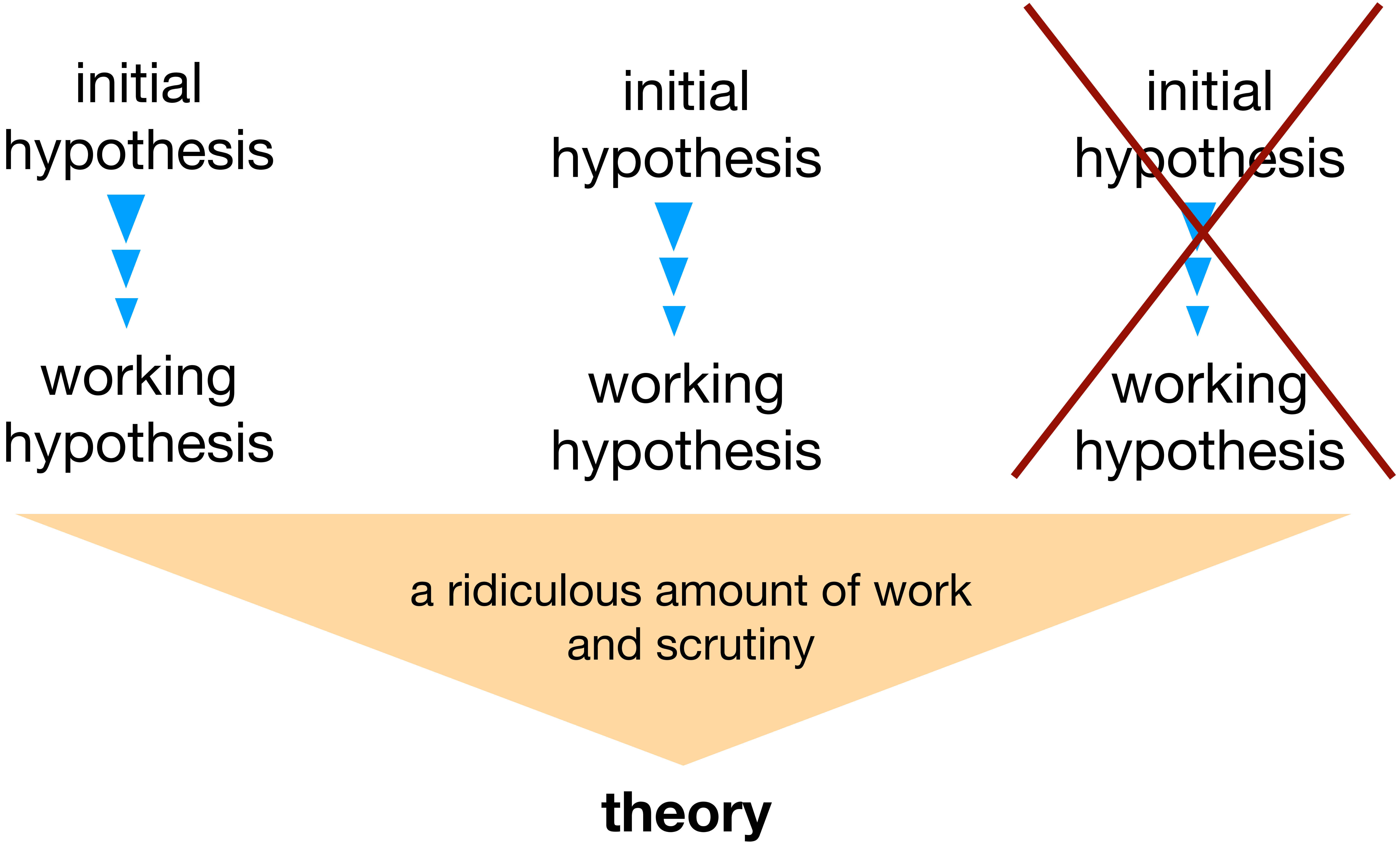
certain variants confer higher survival or reproductive rates to their bearers,

resulting in changes in allele or genotype frequency.

Tentative, mechanistic explanation

broad-sense hypotheses

(a.k.a. “scientific hypotheses”)



broad-sense hypotheses

(a.k.a. “scientific hypotheses”)

initial
hypothesis

working
hypothesis

broad-sense (scientific) hypothesis

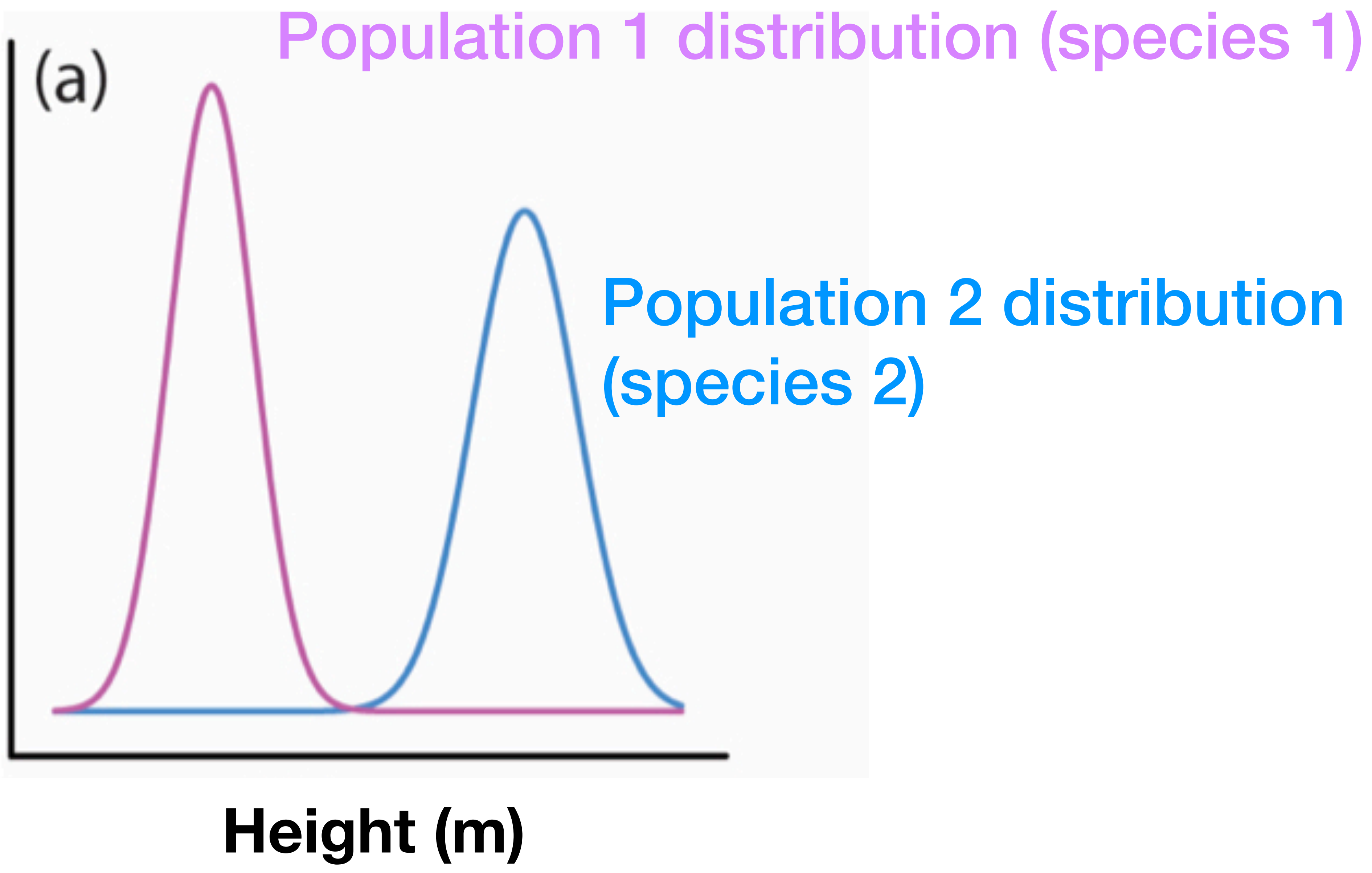
- “prediction” 1 (supported by trend in data?)
    - statistical hypothesis test 1
    - statistical hypothesis test 2
- “prediction” 2 (supported by trend in data?)
    - statistical hypothesis test 1
    - statistical hypothesis test 2
    - statistical hypothesis test 3

⋮

narrow-sense hypotheses
(a.k.a. “statistical hypotheses”)

Formal evaluations of equality or quantity of a parameter, usually based on a sample of measurements from the population of interest and theoretical sampling distributions

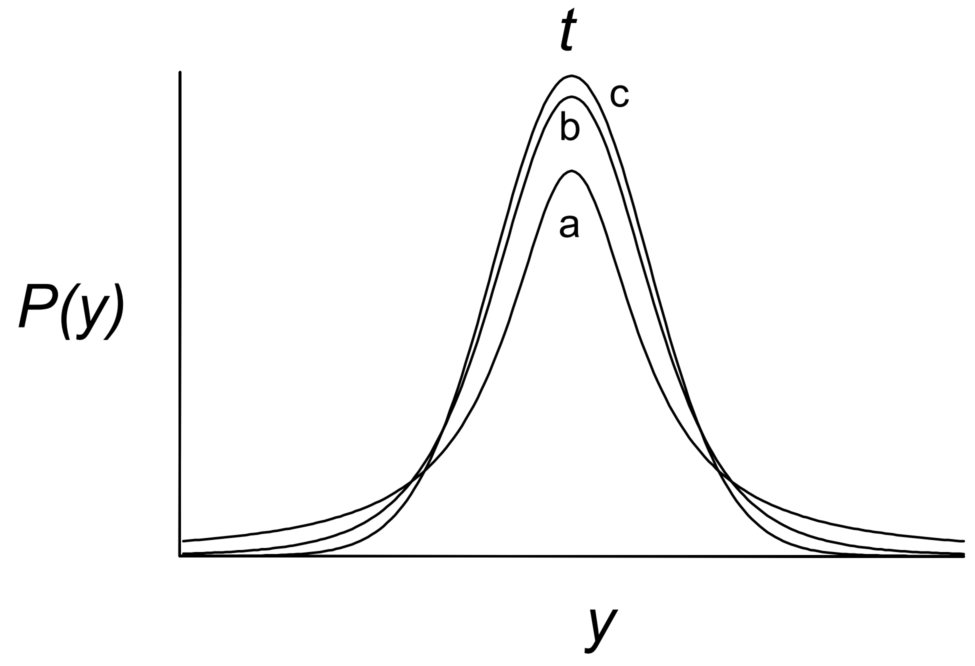
Example: mean height is different between 2 tree species



A statistical hypothesis is a precisely stated belief about the world

- Paired with a “null hypothesis”
- Need a **critical test** to
  - Reject or fail to reject the null hypothesis
  - Compare the likelihood of different models
- **sampling distributions** for “test statistics” form the basis for critical tests

$$t = \frac{(\bar{y}_1 - \bar{y}_2)}{s_{\bar{y}_1 - \bar{y}_2}}$$



***t* statistic and its sampling distribution**

Null and alternative hypotheses

Null Hypothesis (*H_0*): Ponderosa pine trees are the same height, on average, as Douglas fir trees.

Alternative Hypothesis (*H_A*): Ponderosa pine trees are *not* the same height, on average, as Douglas fir trees.

$$H_0 : \mu_1 = \mu_2$$

$$H_A : \mu_1 \neq \mu_2$$

"two-sided" or "two-tailed"

Null and alternative hypotheses

Null Hypothesis (*H_0*): Ponderosa pine trees are **not shorter**, on average, than Douglas fir trees.

Alternative Hypothesis (*H_A*): Ponderosa pine trees are **shorter**, on average, than Douglas fir trees.

$$H_0 : \mu_1 \geq \mu_2$$

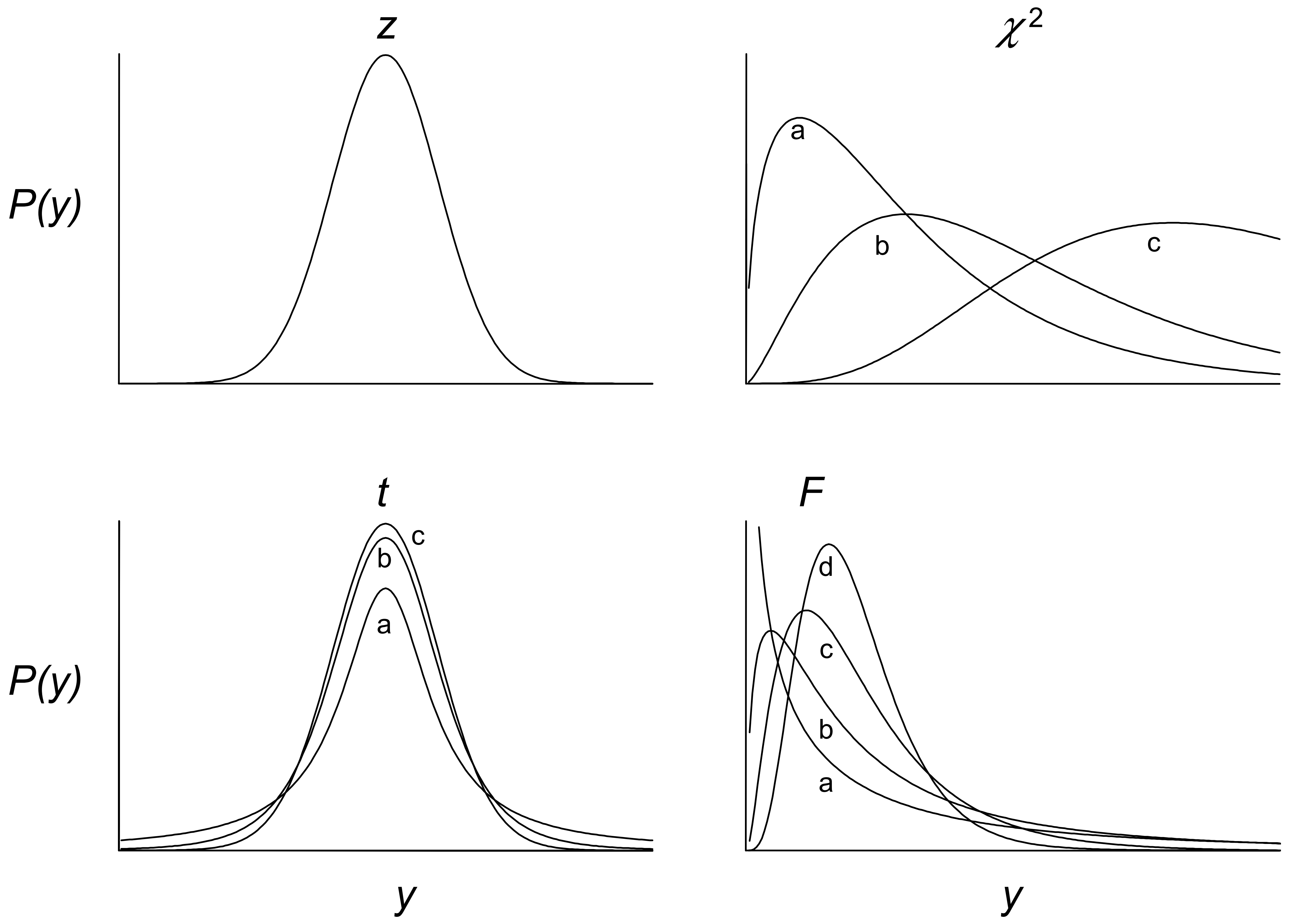
$$H_A : \mu_1 < \mu_2$$

"one-sided" or "one-tailed"

“critical tests” of the null hypothesis

- **We use a “test statistic” calculated from a sample**

- **The test statistic is compared to a probability distribution to assess how extreme it is relative to the null expectation (all outcomes that are not the alternative)**



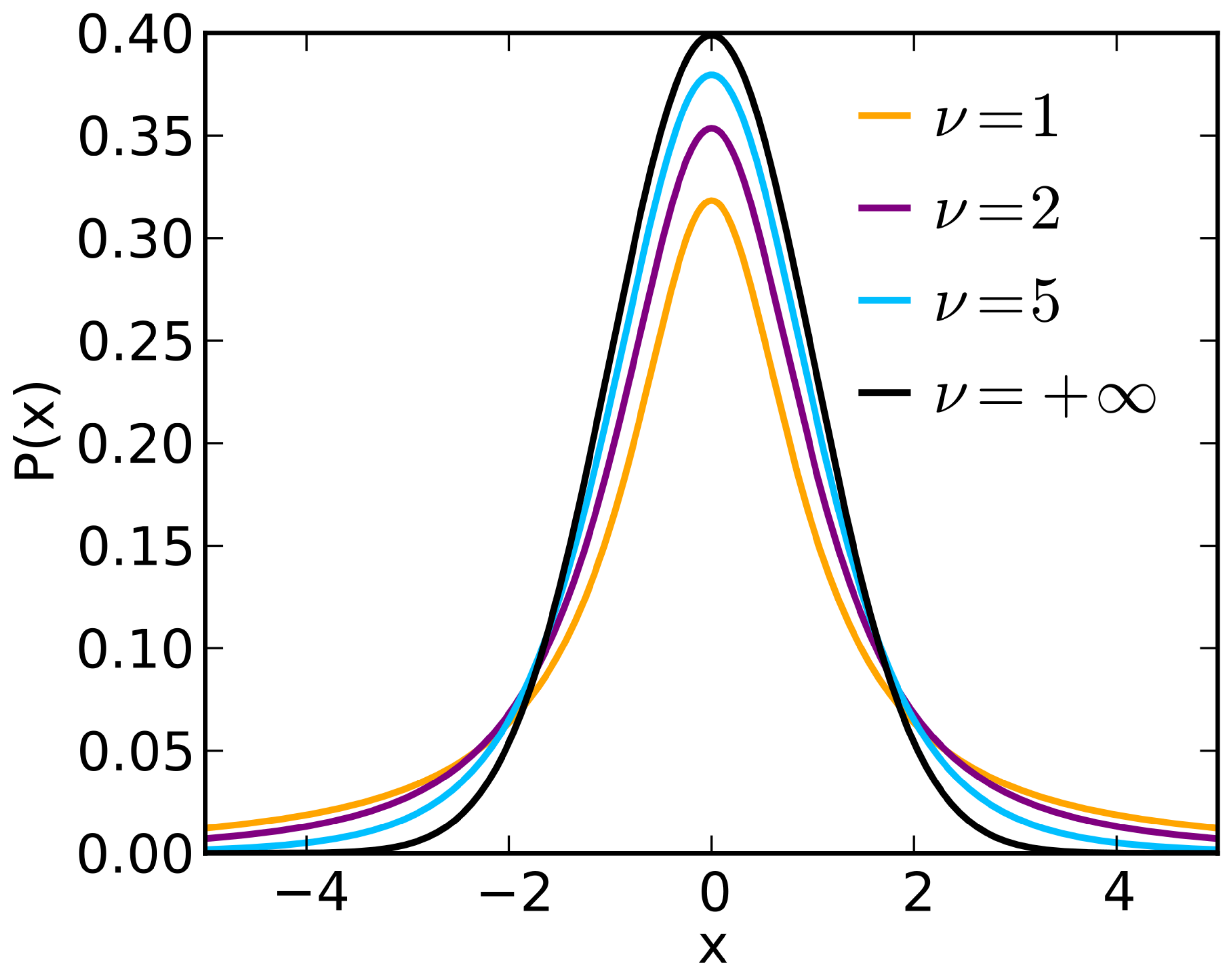
“degrees of freedom” (df) in hypothesis tests

- **Parameters that determine the shape of test statistic prob. distributions**
- **Choosing the correct df for a test depends on sample and group size**
- **Df are determined by the number of independent observations required to estimate a parameter**

$\bar{x}$ **= (15.2 + 13.5 + 20.2) / 3 = 16.3**

For a given estimated mean,
only need 2 of the 3 observations
df = n-1

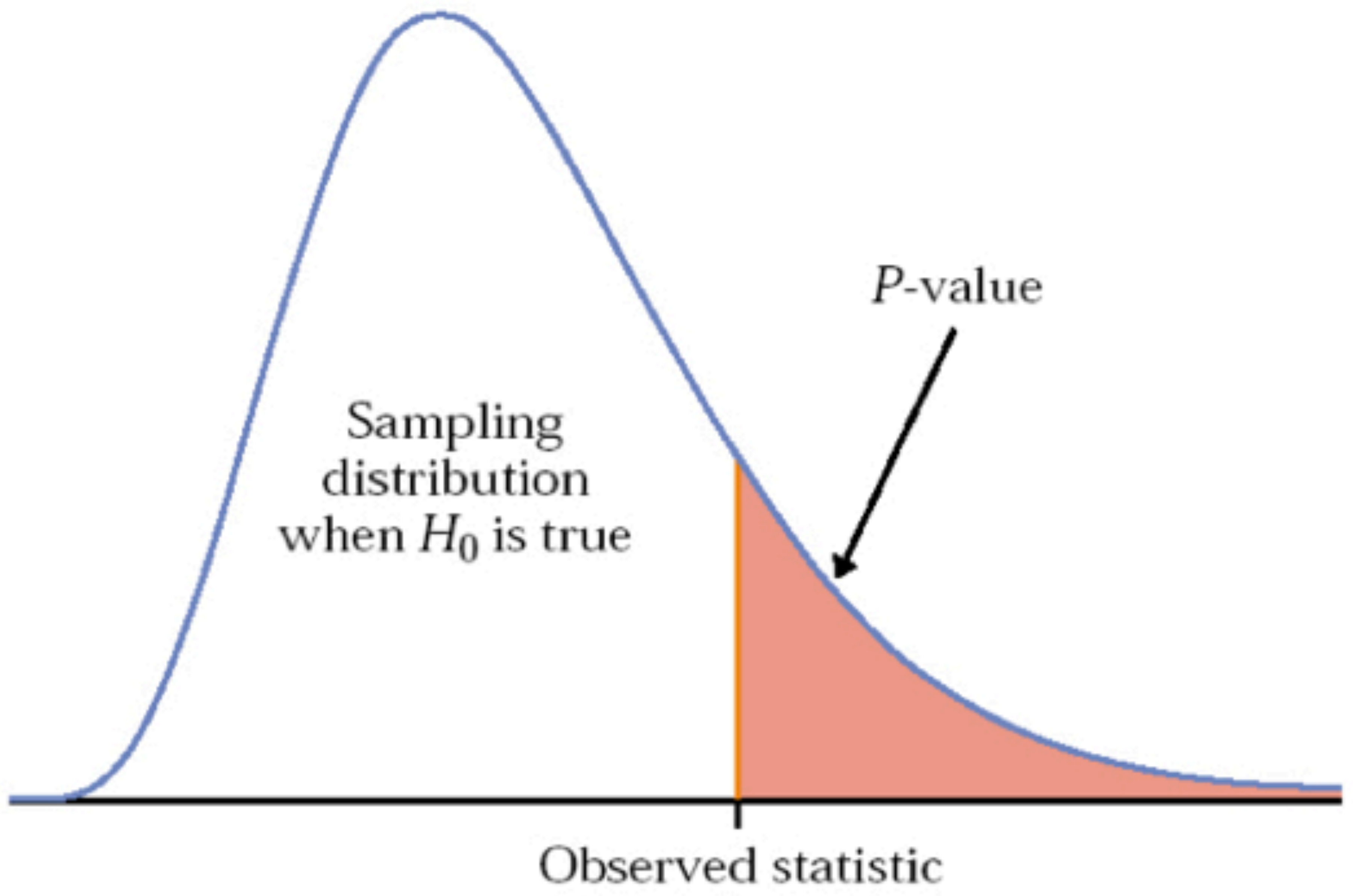
PDF for *t* with various df



The “p-value” in hypothesis tests

***p*-value** = the **long run** probability of observing a result as surprising as the one obtained from our sample, assuming that the null hypothesis is true.

The further our observed test statistic is in the tail, the less likely (lower *p*-value) we observed it by chance under the null.



Hypothesis test errors (type I and II)

STATISTICAL CONCLUSION

Truth at Population Level		Reject H_0	Retain H_0
	Effect	Correct decision Effect detected	Type II error Effect not detected
	No effect	Type I error Effect detected; none exists	Correct decision No effect detected; none exists

"False Negative"

β = accepted prob. of Type II error

"False Positive"

α = accepted prob. of Type I error

Statistical Power

How good is our test at finding a difference when that difference is real (i.e. when H_A is true)?

$$Power \propto \frac{(ES)(\alpha)(\sqrt{n})}{\sigma}$$

(1-β)

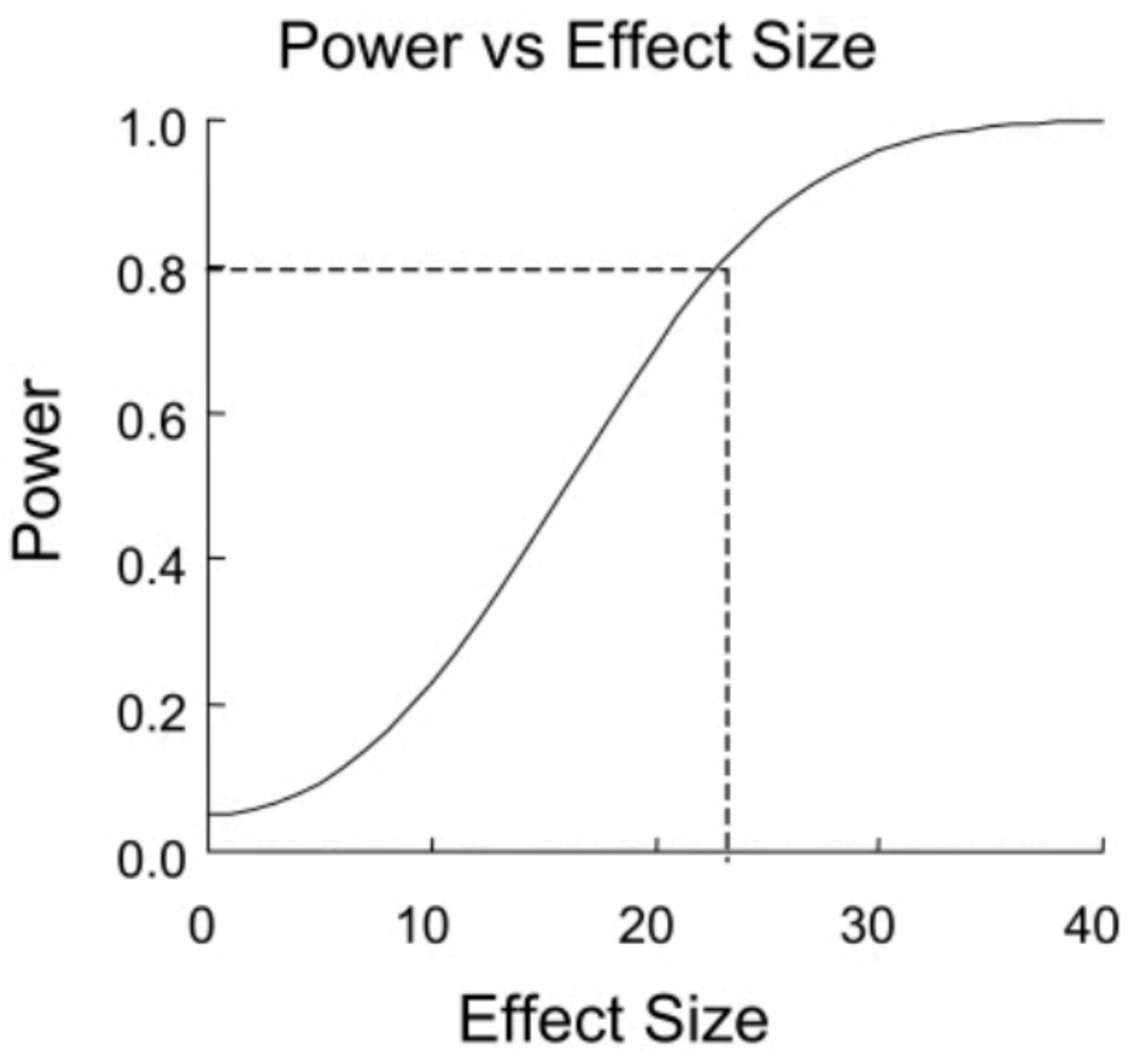
*　*　*

ES = effect size

α = critical p-value

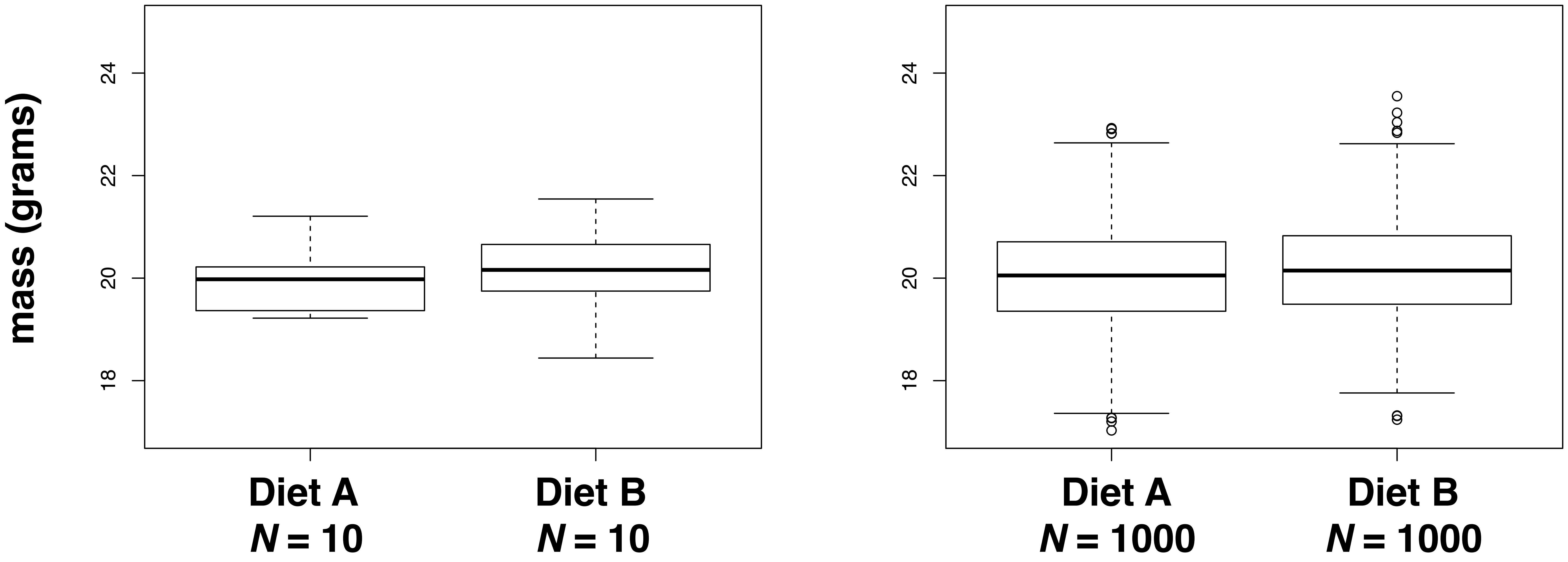
n = sample size

σ = population standard deviation



Interpretation:
Practical vs. Statistical Significance

2 data sets simulated from the same normal distribution.
For both, the mean mass for Diet B is 1% greater than the mean mass for Diet A.



Diet A vs. B
(t-test)

$p = 0.6413$

$p = 0.008922$

Key hypothesis test terms

***p*-value** = the **long run** probability of observing a result as surprising as the one obtained from our sample, given that the null hypothesis is true

alpha (α) = critical value of p-value cutoff for experiments. The Type I error we are willing to tolerate

beta (β) = cutoff for probability of accepting Type II error

Power = the probability that a test will correctly reject the null hypothesis (1 - β). It depends on effect size, sample size, chosen alpha, and population standard deviation

Multiple testing = performing the same or similar tests to address the same fundamental question - need to apply a correction