Inference about a population proportion

BPS chapter 20

Objectives (BPS chapter 20)

Inference for a population proportion

- The sample proportion $\hat{p}$
- The sampling distribution of $\hat{p}$
- Large sample confidence interval for *p*
- Accurate confidence intervals for *p*
- Choosing the sample size
- Significance tests for a proportion

The two types of data — *reminder*

- ***Quantitative***
  - Something that can be counted or measured and then added, subtracted, averaged, etc., across individuals in the population.
  - Example: How tall you are, your age, your blood cholesterol level
- ***Categorical***
  - Something that falls into one of several categories. What can be counted is the proportion of individuals in each category.
  - Example: Your blood type (A, B, AB, O), your hair color, your family health history for genetic diseases, whether you will develop lung cancer

How do you figure it out? Ask:

- What are the *n* individuals/units in the sample (of size "*n*")?
- What's being recorded about those *n* individuals/units?
- Is that a number (→ quantitative) or a statement (→ categorical)?

The sample proportion $\hat{p}$

We now study categorical data and draw inference on the proportion, or percentage, of the population with a specific characteristic.

If we call a given categorical characteristic in the population "success," then the sample proportion of successes, $\hat{p}$,is:

$$\hat{p} = \frac{\text{count of successes in the sample}}{\text{count of observations in the sample}}$$

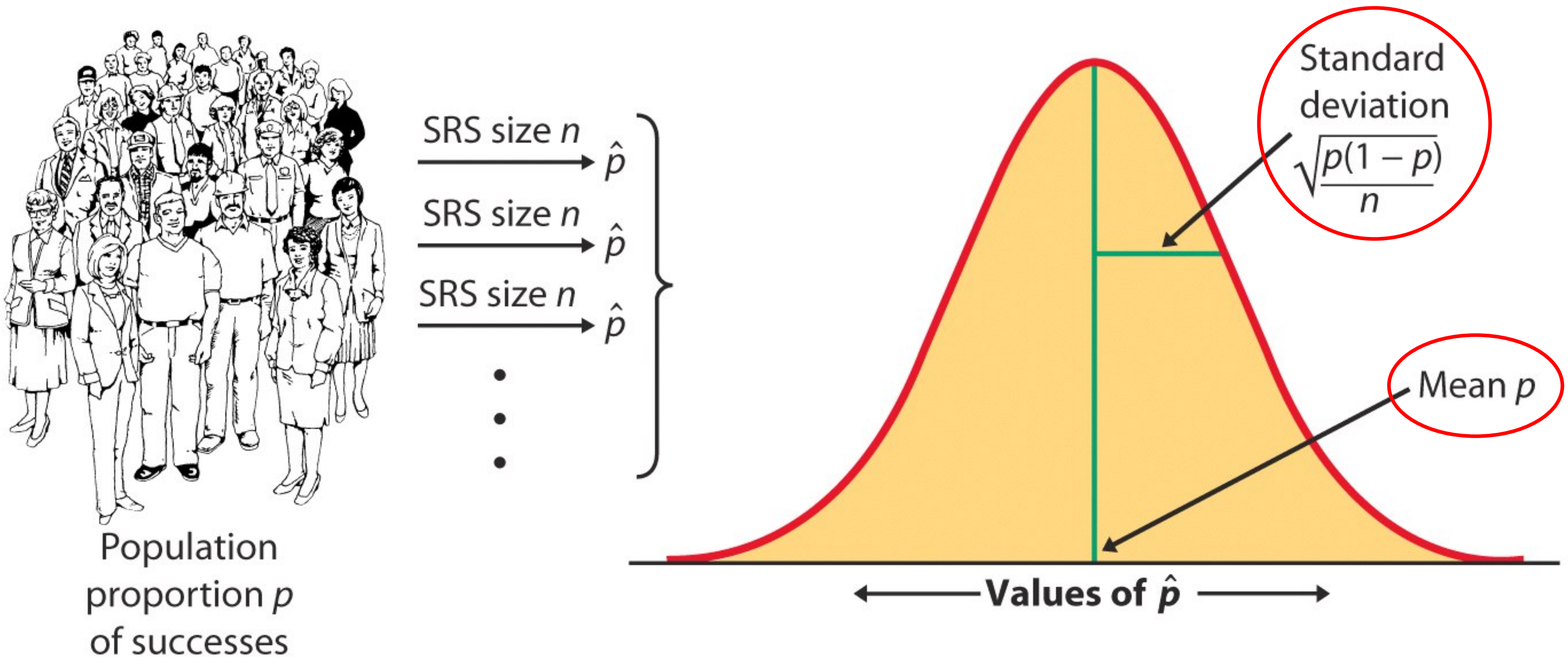
- We choose 50 people in an undergrad class, and find that 10 of them are Hispanic:

  $\hat{p}$ = (10)/(50) = 0.2 *(proportion of Hispanics in sample)*
- You treat a group of 120 Herpes patients given a new drug; 30 get better:

  $\hat{p}$ = (30)/(120) = 0.25 *(proportion of patients improving in sample)*

Sampling distribution of $\hat{p}$

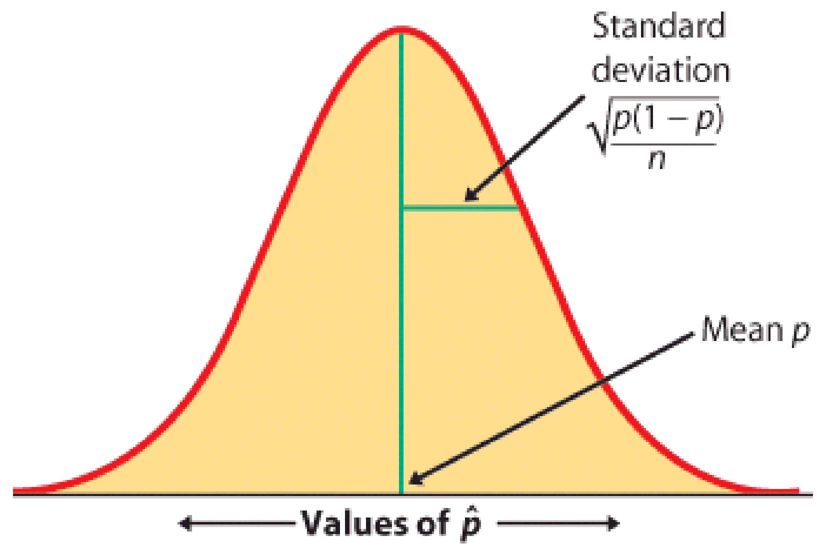
The sampling distribution of $\hat{p}$ is never exactly normal. But as the sample size increases, the sampling distribution of $\hat{p}$ becomes approximately normal.



Implication for estimating proportions

The mean and standard deviation (width) of the sampling distribution are both completely determined by *p* and *n*.

$$N\left(p, \sqrt{p(1-p)/n}\right)$$



Thus, we have only one population parameter to estimate, *p*.

Therefore, inference for proportions can rely directly on the normal distribution *(unlike inference for means, which requires the use of a t distribution with a specific degree of freedom)*.

Conditions for inference on *p*

Assumptions:

1. We regard our data as a **simple random sample** (SRS) from the population. That is, as usual, the most important condition.

2. The **sample size *n* is large enough** that the sampling distribution is indeed normal.

   How large a sample size is enough? Different inference procedures require different answers (*we'll see what to do practically*).

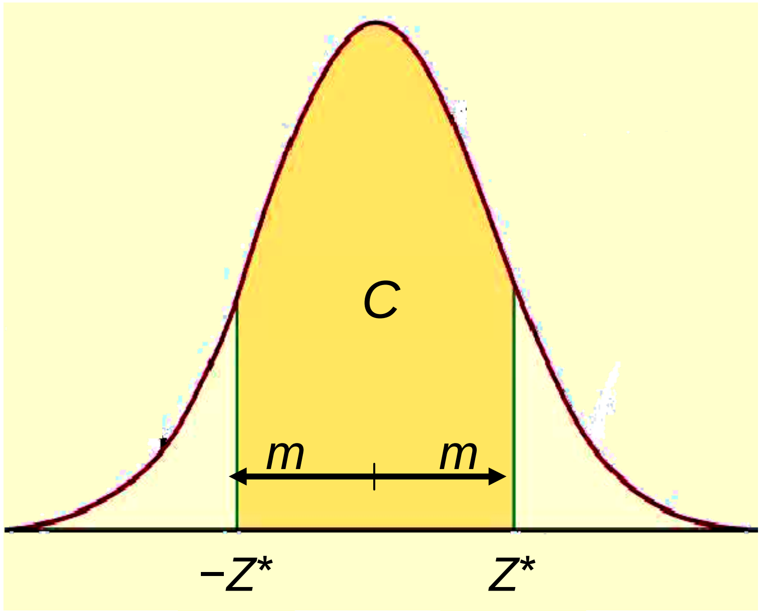
Large-sample confidence interval for *p*

Confidence intervals contain the population proportion *p* in C% of samples. For an SRS of size *n* drawn from a large population and with sample proportion $\hat{p}$ calculated from the data, an **approximate level *C* confidence interval** for *p* is:

$$\hat{p} \pm m, \;\; m \text{ is the margin of error}$$

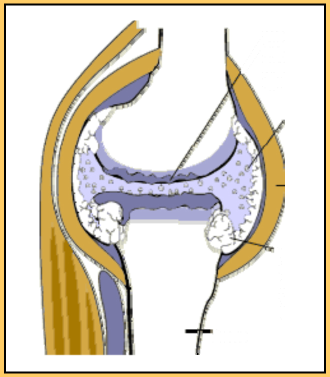
$$m = z^* SE = z^* \sqrt{\hat{p}(1-\hat{p})/n}$$

Use this method when the number of successes and the number of failures are both at least 15.

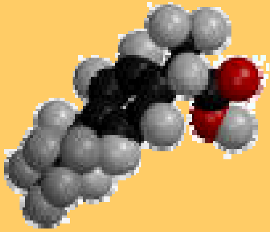


C is the area under the standard normal curve between −*z** and *z**.

Medication side effects



Arthritis is a painful, chronic inflammation of the joints. An experiment on the side effects of pain relievers examined arthritis patients to find the proportion of patients who suffer side effects.



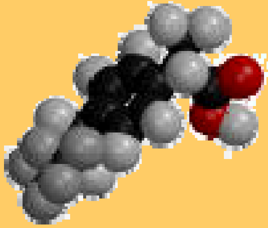
What are some side effects of ibuprofen?

Serious side effects (seek medical attention immediately):

- Allergic reactions (difficulty breathing, swelling, or hives)
- Muscle cramps, numbness, or tingling
- Ulcers (open sores) in the mouth
- Rapid weight gain (fluid retention)
- Seizures
- Black, bloody, or tarry stools
- Blood in your urine or vomit
- Decreased hearing or ringing in the ears
- Jaundice (yellowing of the skin or eyes)
- Abdominal cramping, indigestion, or heartburn

Less serious side effects (discuss with your doctor):

- Dizziness or headache
- Nausea, gaseousness, diarrhea, or constipation
- Depression
- Fatigue or weakness
- Dry mouth
- Irregular menstrual periods



Let's calculate a 90% confidence interval for the population proportion of arthritis patients who suffer some "adverse symptoms."

What is the sample proportion $\hat{p}$?

$$\hat{p} = \frac{23}{440} \approx 0.052$$

What is the sampling distribution for the proportion of arthritis patients with adverse symptoms for samples of 440?

$$\hat{p} \approx N(p, \sqrt{p(1-p)/n})$$

For a 90% confidence level, z^* = 1.645.

z*	0.67	0.841	1.036	1.282	1.645	1.960	2.054	2.326
	50%	60%	70%	80%	90%	95%	96%	98%
					Confidence level C			

Using the large sample method, we calculate a margin of error *m*:

$$m = z^* \sqrt{\hat{p}(1-\hat{p})/n}$$

$$m = 1.645 * \sqrt{0.052(1-0.052)/440}$$

$$m = 1.645 * 0.014 \approx 0.023$$

$$90\% \text{ CI for } p: \hat{p} \pm m$$

$$\text{or } 0.052 \pm 0.023$$

➔ With 90% confidence level, between 2.9% and 7.5% of arthritis patients taking this pain medication experience some adverse symptoms.

Because we have to use an estimate of *p* to compute the margin of error, confidence intervals for a population proportion are not very accurate.

$$m = z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Specifically, the actual confidence interval is usually less than the confidence level you asked for in choosing z. But there is no systematic amount (because it depends on p).*

Use with caution!

“Plus four” confidence interval for *p*

A simple adjustment produces more accurate confidence intervals. We act as if we had four additional observations, two being successes and two being failures. Thus, the new sample size is $n + 4$ and the count of successes is $X + 2$.

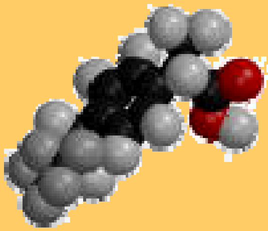
The “plus four” estimate of p is: $$\tilde{p} = \frac{\text{counts of successes} + 2}{\text{count of all observations} + 4}$$

And an approximate level C confidence interval is:

$$CI: \quad \tilde{p} \pm m \text{ , with}$$

$$m = z^* SE = z^* \sqrt{\tilde{p}(1 - \tilde{p})/(n + 4)}$$

Use this method when C is at least 90% and sample size is at least 10.



We now use the "plus four" method to calculate the 90% confidence interval for the population proportion of arthritis patients who suffer some "adverse symptoms."

What is the value of the "plus four" estimate of *p*? $\tilde{p} = \frac{23+2}{440+4} = \frac{25}{444} \approx 0.056$

An approximate 90% confidence interval for *p* using the "plus four" method is:

$$m = z^* \sqrt{\tilde{p}(1-\tilde{p})/(n+4)}$$
$$m = 1.645 * \sqrt{0.056(1-0.056)/444}$$
$$m = 1.645 * 0.011 \approx 0.018$$

$$90\%\ \text{CI for}\ p: \quad \tilde{p} \pm m$$
$$\text{or} \quad 0.056 \pm 0.018$$

➔ With 90% confidence level, between 3.8% and 7.4% of arthritis patients taking this pain medication experience some adverse symptoms.

	Upper tail probability P											
	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.003	0.001	0.0005
z*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

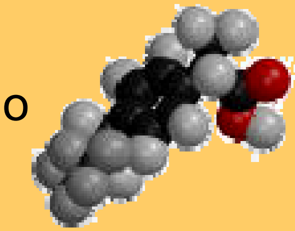
Choosing the sample size

You may need to choose a sample size large enough to achieve a specified margin of error. However, because the sampling distribution of $\hat{p}$ is a function of the population proportion *p* this process requires that you guess a likely value for *p*: *p**.

$$p \sim N\left(p, \sqrt{p(1-p)/n}\right) \;\Rightarrow\; \boxed{m = \left(\frac{z^*}{m}\right)^2 p^*(1-p^*)}$$

The margin of error will be less than or equal to *m* if *p** is chosen to be 0.5.

Remember, though, that sample size is not always stretchable at will. There are typically costs and constraints associated with large samples.



What sample size would we need in order to achieve a margin of error no more than 0.01 (1%) for a 90% confidence interval for the population proportion of arthritis patients who suffer some "adverse symptoms?"

We could use 0.5 for our guessed *p**. However, since the drug has been approved for sale over the counter, we can safely assume that no more than 10% of patients should suffer "adverse symptoms" (a better guess than 50%).

For a 90% confidence level, *z** = 1.645.

z*	0.67	0.841	1.036	1.282	1.645	1.960	2.054	2.326
	50%	60%	70%	80%	90%	95%	96%	98%
					Confidence level C			

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1-p^*) = \left(\frac{1.645}{0.01}\right)^2 (0.1)(0.9) \approx 2434.4$$

➔ To obtain a margin of error of no more than 1% we would need a sample size *n* of at least 2435 arthritis patients.

Significance test for p

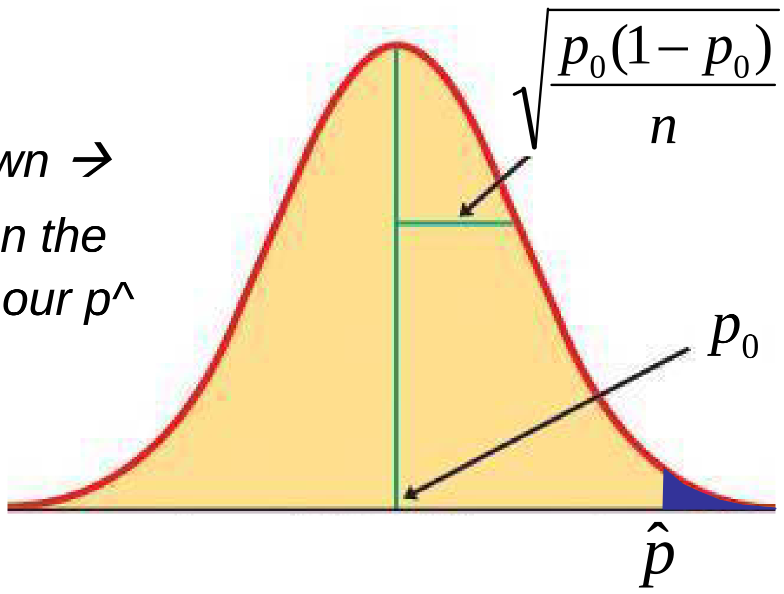
The sampling distribution for $\hat{p}$ is approximately normal for large sample sizes, and its shape depends solely on p and n.

Thus, we can easily test the null hypothesis:

H_0: $p = p_0$ (a given value we are testing)

If H_0 is true, the sampling distribution is known →
The likelihood of our sample proportion given the null hypothesis depends on how far from p_0 our p^ is in units of standard deviation.

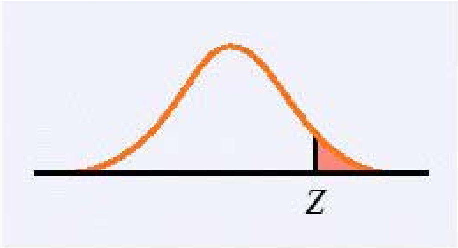
$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$



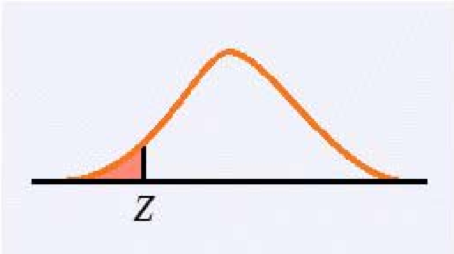
This is valid when both expected counts — expected successes np_0 and expected failures $n(1 - p_0)$ — are each 10 or larger.

P-values and one- or two-sided hypotheses — *reminder*

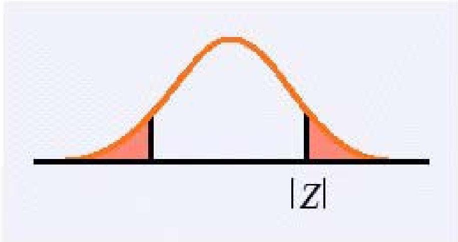
$H_a: p > p_0$ is $P(Z \geq z)$



$H_a: p < p_0$ is $P(Z \leq z)$



$H_a: p \neq p_0$ is $2P(Z \geq |z|)$



And as always, if the *P*-value is smaller than the chosen significance level α, then the difference is statistically significant and we reject H_0.

A national survey by the National Institute for Occupational Safety and Health on restaurant employees found that 75% said that work stress had a negative impact on their personal lives.

You investigate a restaurant chain to see if the proportion of all their employees negatively affected by work stress differs from the national proportion $p_0 = 0.75$.

$$H_0: p = p_0 = 0.75 \text{ vs. } H_a: p \neq 0.75 \text{ (two-sided alternative)}$$

In your SRS of 100 employees, you find that 68 answered “Yes” when asked, “Does work stress have a negative impact on your personal life?”

The expected counts are 100 × 0.75 = 75 and 25.
Both are greater than 10, so we can use the z-test.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

The test statistic is:

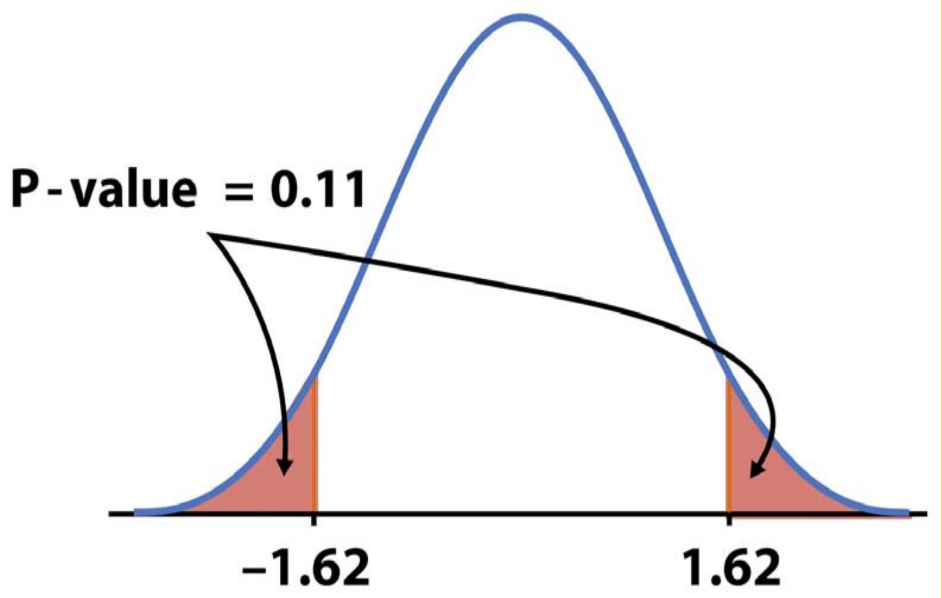
$$= \frac{0.68 - 0.75}{\sqrt{\frac{(0.75)(0.25)}{100}}} = 1.62$$

From Table A we find the area to the left of z 1.62 is 0.9474.

Thus $P(Z \geq 1.62)$ = 1 − 0.9474, or 0.0526. Since the alternative hypothesis is two-sided, the *P*-value is the area in both tails, and *P* = 2 × 0.0526 = 0.1052.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

➔ The chain restaurant data are compatible with the national survey results ($\hat{p}$ = 0.68, *z* = 1.62, *P* = 0.11).



Interpretation: magnitude versus reliability of effects

The **reliability** of an interpretation is related to the strength of the evidence. The smaller the ***P*****-value,** the stronger the evidence against the null hypothesis and the more confident you can be about your interpretation.

The **magnitude** or **size** of an effect relates to the real-life relevance of the phenomenon uncovered. The *P*-value does NOT assess the relevance of the effect, nor its magnitude.

A **confidence interval** will assess the magnitude of the effect. However, magnitude is not necessarily equivalent to how theoretically or practically relevant an effect is.