Review Session 7

API-201, 11.05.21
Sophie Hill

What are you finding difficult with this week's material?

Agenda

Review:

- PS7 Q3a - do we have one proportion or two?
- Sophie's magic rule for remembering Type I/II errors
- Understanding statistical power
- Hints for PS8 Q1:
  - Converting vote share into **2-party** vote share
  - the "coverage rate"

PS7 Q3a

Suppose that you have a sample from the US population indicating, for each person, any cancer diagnosis. Among patients diagnosed with cancer, you would like to test the hypothesis that the fraction of them over 65 is equal to the fraction of them under 65.

Write the null hypothesis in math.

PS7 Q3a

Suppose that you have a sample from the US population indicating, for each person, any cancer diagnosis. Among patients diagnosed with cancer, you would like to test the hypothesis that the fraction of them over 65 is equal to the fraction of them under 65.

Write the null hypothesis in math.

This looks like a difference-in-proportions...

PS7 Q3a

Suppose that you have a sample from the US population indicating, for each person, any cancer diagnosis. Among patients diagnosed with cancer, you would like to test the hypothesis that the fraction of them over 65 is equal to the fraction of them under 65.

Write the null hypothesis in math.

*… but it's actually a **single** proportion. Why?*

			Sampling Distribution		
Target parameter	**Sample Size(s)**	**Estimate**	**Shape**	**Mean of Estimator**	**Standard Error**
q	n	$\hat{q}$	Approximately normal	q	$\sqrt{\frac{q(1-q)}{n}}$
μ	n	$\bar{x}$	Approximately normal*	μ	$\frac{s}{\sqrt{n}}$
$q_1 - q_2$	$n_1 and\ n_2$	$\hat{q}_1 - \hat{q}_2$	Approximately normal	$q_1 - q_2$	$\sqrt{\frac{q_1(1-q_1)}{n_1} + \frac{q_2(1-q_2)}{n_2}}$
$\mu_1 - \mu_2$	$n_1 and\ n_2$	$\bar{x}_1 - \bar{x}_2$	Approximately normal*	$\mu_1 - \mu_2$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

CLT Conditions: Large enough random samples (all 4 rows); independent samples (last 2 rows)

* Since σ is usually unknown, we use s instead and the sampling distribution is t (instead of Normal). When sample sizes are large enough, the t and Normal distributions are approximately equal.

From Handout 15:

Conditions for Central Limit Theorem (CLT)

- Large enough sample sizes
- Samples are randomly drawn from two groups
- Two samples are independent
    - Sampling procedure for the first sample does not depend on who is in the second sample

PS7 Q3a

Suppose that you have a sample from the US population indicating, for each person, any cancer diagnosis. Among patients diagnosed with cancer, you would like to test the hypothesis that the fraction of them over 65 is equal to the fraction of them under 65.

Write the null hypothesis in math.

PS7 Q3a

Suppose that you have a sample from the US population indicating, for each person, any cancer diagnosis. Among patients diagnosed with cancer, you would like to test the hypothesis that the fraction of them over 65 is equal to the fraction of them under 65.

These two samples are **not independent**. As the proportion of cancer patients who are 65+ goes up, the proportion of cancer patients who are <65 *must go down*.

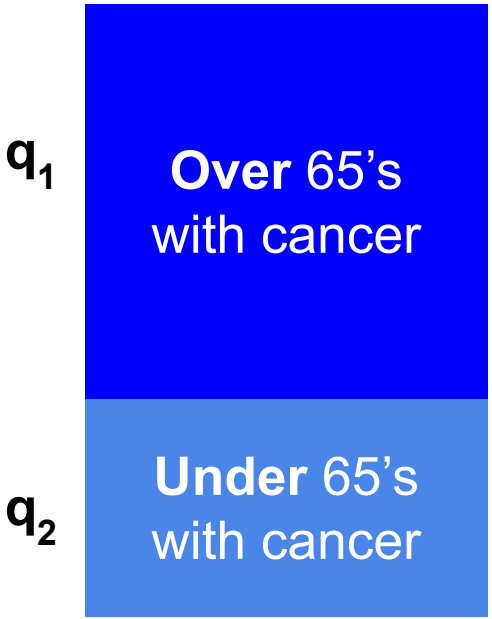
PS7 Q3b

In contrast, in part (b) we *did* have two independent samples:

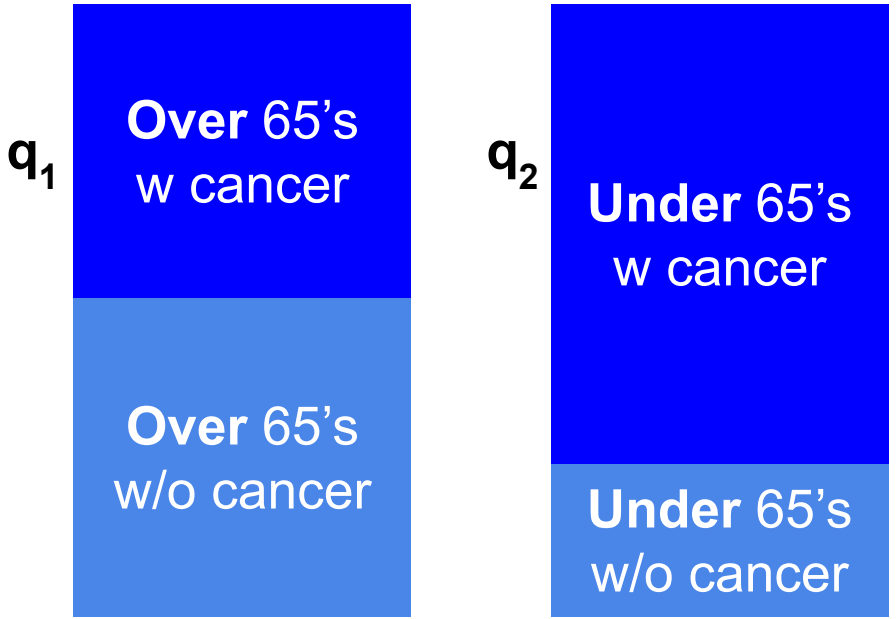
You would now like to test the hypothesis that the fraction of people diagnosed with cancer is equal among those aged over 65 than those under 65.

If the rate of cancer among the 65+'s goes up, there is no necessary mathematical implication for the rate of cancer among the <65's.

Visual interpretation



(a)



(b)

Mathematical interpretation

$$q_1 = \frac{\text{over 65s with cancer}}{\text{everyone with cancer}}$$

$$q_2 = \frac{\text{under 65s with cancer}}{\text{everyone with cancer}}$$

(a)

$$q_1 = \frac{\text{over 65s with cancer}}{\text{everyone who is over 65}}$$

$$q_2 = \frac{\text{under 65s with cancer}}{\text{everyone who is under 65}}$$

(b)

Statistical interpretation

For two random variables *X* and *Y*, the variance of the difference, *X* - *Y*, is given by:

$$\mathbf{Var(X-Y) = Var(X) + Var(Y) - 2\,Cov(X, Y)}$$

If X and Y are **independent**, then the covariance, Cov(X, Y) = 0, so Var(X-Y) = Var(X) + Var(Y).

This is where we get the formula for the SE of a difference in proportions:

$$\sqrt{\frac{q_1(1-q_1)}{n_1} + \frac{q_2(1-q_2)}{n_2}}$$

* We haven't talked about the covariance before, but you can think of it as being an "unstandardized" version of the correlation coefficient (i.e. not always between -1 and +1).

Statistical interpretation

If *X* and *Y* are **not** independent, then we need to take the covariance into account.

For part (a), the two proportions were complements (i.e., q1 = 1 - q2) and so they must be *negatively* correlated.

Since the covariance Cov(q1, q2) is *negative* the variance of the difference q1-q2 is:

$$\mathbf{Var(q1 - q2) = Var(q1) + Var(q2) - 2\,Cov(q1, q2)}$$

...which will be **larger** than **Var(q1) + Var(q2)**

Sophie’s magic rule for remembering Type I/II errors

Type I

Type I I

Sophie's magic rule for remembering Type I/II errors

Type P

Type N

Sophie's magic rule for remembering Type I/II errors

Type P ~ False Positive

Type N ~ False Negative


1
2
Α α
alpha
Β β
beta
Γ γ
gamma
Δ δ
delta
Ε ε
epsilon
Ζ ζ
zeta
Η η
eta
Θ θ
theta
Ι ι
lota
Κ κ
kappa
Λ λ
lambda
Μ μ
mu
Ν ν
nu
Ξ ξ
xl
Ο ο
omikron
Π π
pi
Ρ ρ
rho
Σ σ/ς
sigma
Τ τ
tau
Υ υ
upsilon
Φ φ
phi
Χ χ
hi
Ψ ψ
psi
Ω ω
omega

Sophie's magic rule for remembering Type I/II errors

Type I ~ False Positive

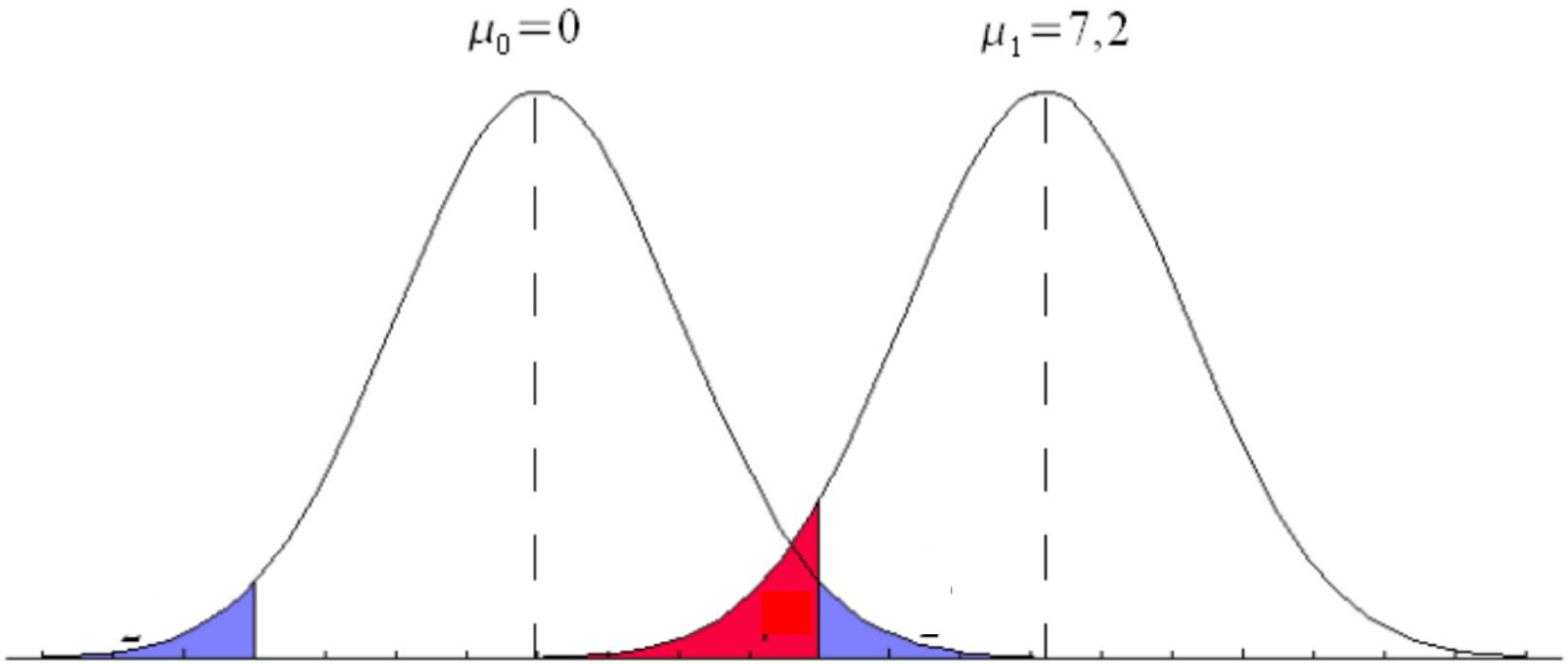
α = P(Type I error)

Type I I ~ False Negative

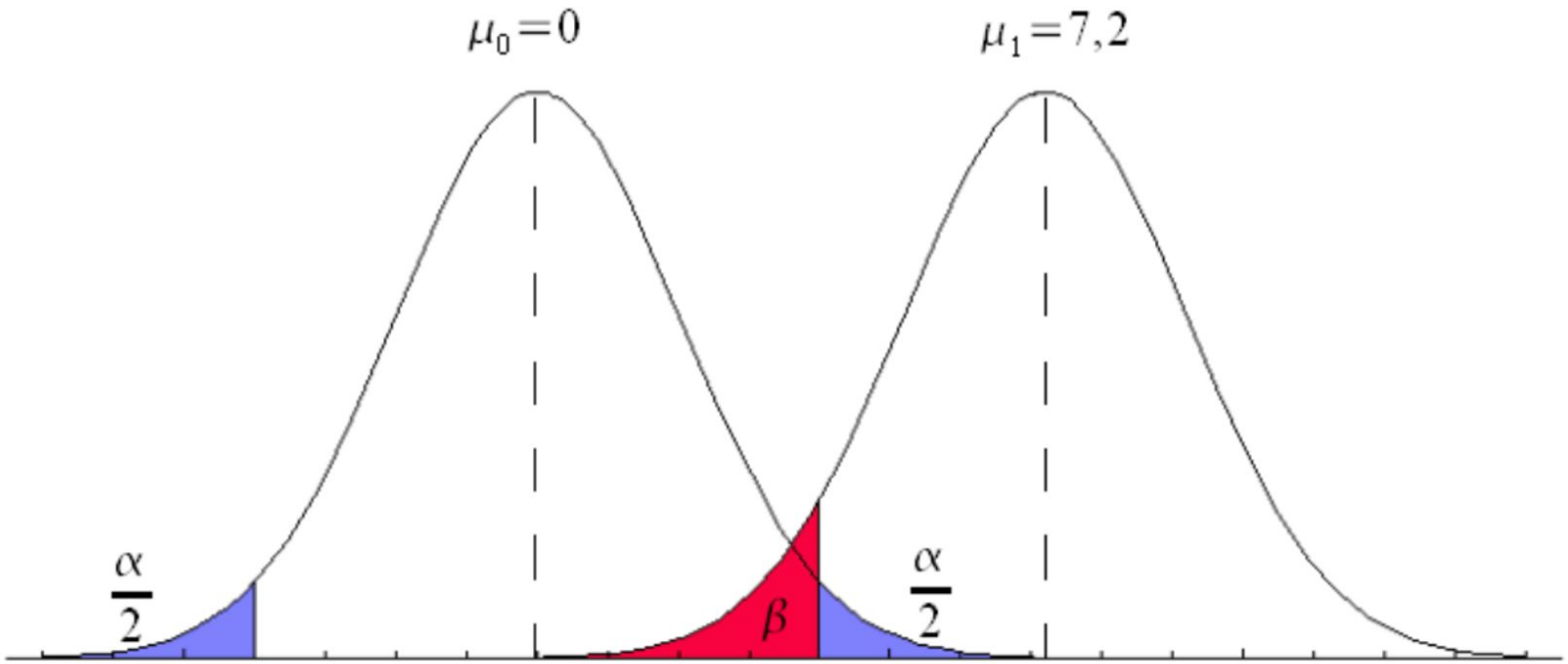
β = P(Type II error)

Statistical power

Which quantity is represented by the red area? And which by the blue area?

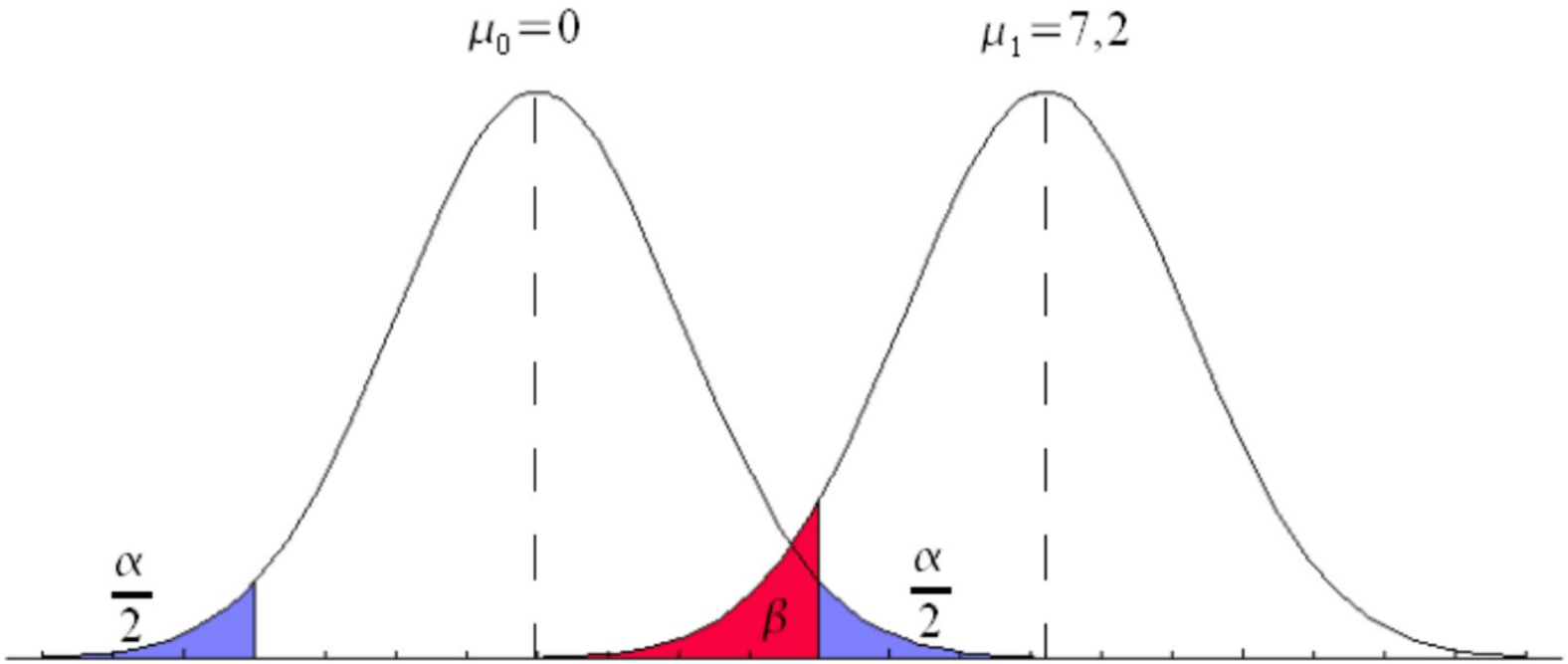


Statistical power



Statistical power

What happens to alpha and beta if the alternative hypothesis moves **closer** to the null?

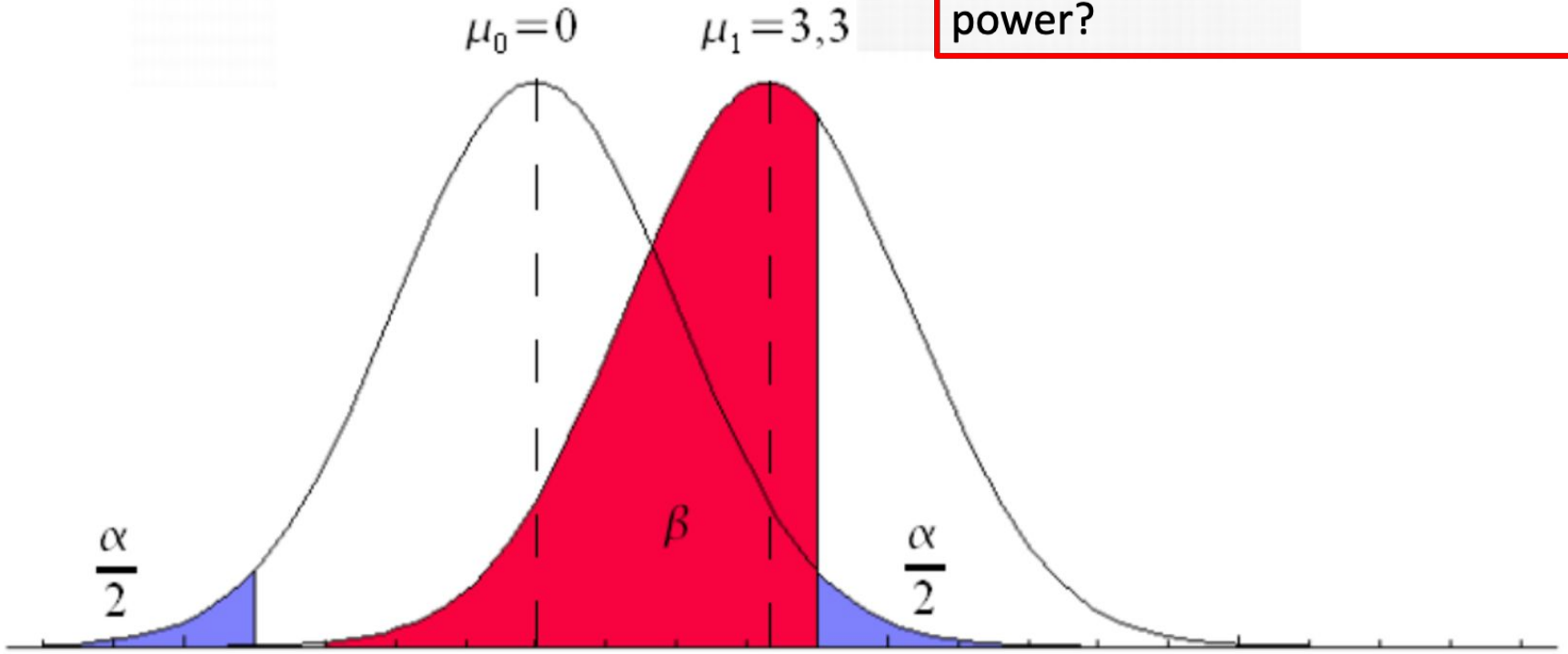


Statistical power

Alpha unchanged

Beta increases

How does this relate to statistical power?



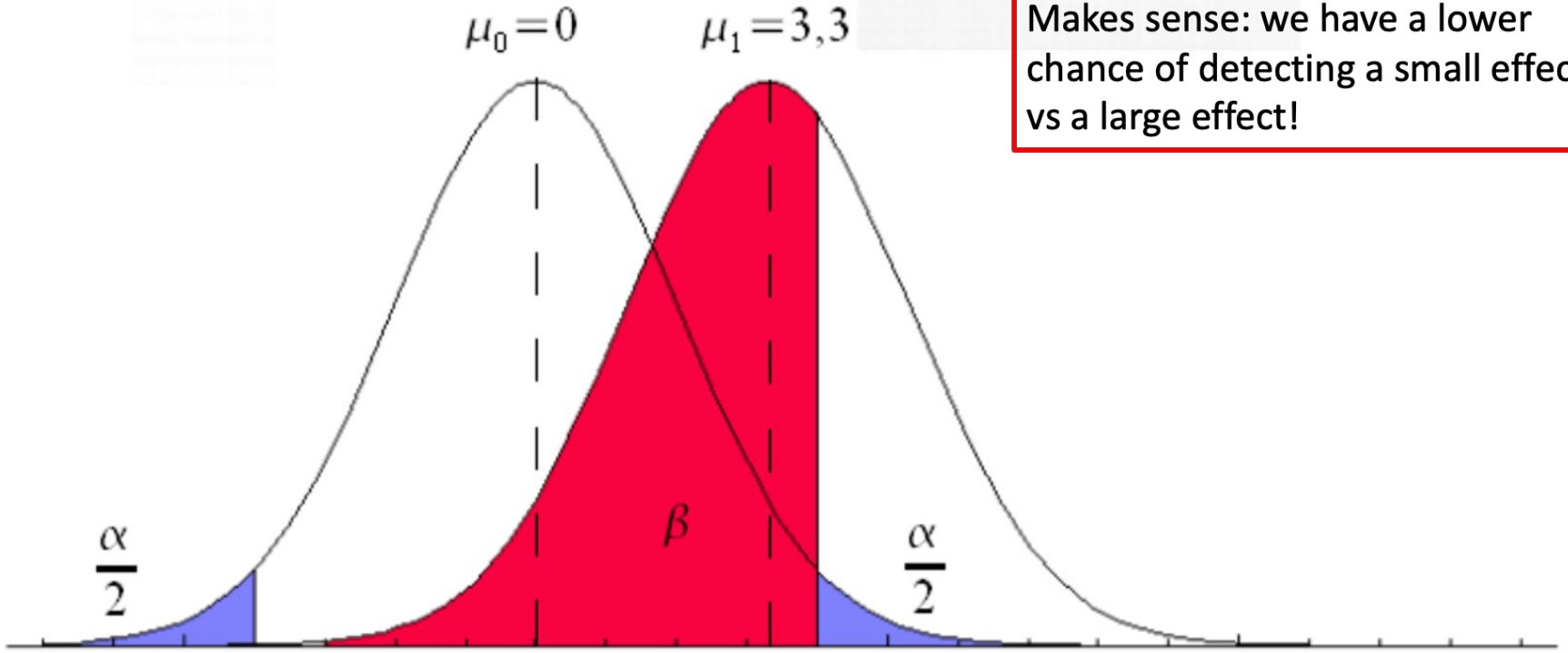
Statistical power

Alpha unchanged

Beta increases

Power = 1 – Beta.

Makes sense: we have a lower chance of detecting a small effect vs a large effect!



Statistical power

Fill in the blanks! [higher / lower]

1. Larger **sample size** means _____ power
2. Larger **effect size** means **higher** power
3. Larger **sample variance** means _____ power
4. Larger **alpha** means _____ power

Statistical power

Fill in the blanks! [higher / lower]

1. Larger **sample size** means **higher** power
2. Larger **effect size** means **higher** power
3. Larger **sample variance** means **lower** power
4. Larger **alpha** means **higher** power

Key takeaways

Alpha

- We can choose whatever alpha we like
- To think about alpha, first **assume the null**, i.e., centre your sampling distribution around the null value

Beta

- Beta is determined by our study design and hypothesized effect size
- To think about beta, first **assume a specific alternative hypothesis**, i.e. centre your sampling distribution around an alternative value

PS8 Q1

3 For example, in 2016, Hillary Clinton was ahead in national polls. Therefore, the polls from 2016 list Clinton's vote share in the poll. We ignore votes for third-party candidates in this exercise and compute the two-party share for each candidate. That means, for example, that if Candidate A has 50% of the vote, Candidate B has 44% of the vote, and other candidates have the remaining 6% of the vote, the two-party share is 94%, Candidate A's share is 50% / 94% = 53.2%, and Candidate B's share is 44% / 94% = 46.8%.

PS8 Q1

3 For example, in 2016, Hillary Clinton was ahead in national polls. Therefore, the polls from 2016 list Clinton's vote share in the poll. We ignore votes for third-party candidates in this exercise and compute the two-party share for each candidate. That means, for example, that if Candidate A has 50% of the vote, Candidate B has 44% of the vote, and other candidates have the remaining 6% of the vote, the two-party share is 94%, Candidate A's share is 50% / 94% = 53.2%, and Candidate B's share is 44% / 94% = 46.8%.

It makes life much easier to think about US political polls in terms of the 2-party vote share... why?

Now we only have a **single proportion** to worry about, since $q_d = 1 - q_r$

PS8 Q1

3 For example, in 2016, Hillary Clinton was ahead in national polls. Therefore, the polls from 2016 list Clinton's vote share in the poll. We ignore votes for third-party candidates in this exercise and compute the two-party share for each candidate. That means, for example, that if Candidate A has 50% of the vote, Candidate B has 44% of the vote, and other candidates have the remaining 6% of the vote, the two-party share is 94%, Candidate A's share is 50% / 94% = 53.2%, and Candidate B's share is 44% / 94% = 46.8%.

It makes life much easier to think about US political polls in terms of the 2-party vote share... why?

Now we only have a **single proportion** to worry about, since $q_d = 1 - q_r$

Life is not so simple in countries with genuine multi-party systems!!

PS8 Q1

Let's repeat this assessment for each of the polls in the database. For each poll, use Excel to calculate the margin of error (at the 95% confidence level) for the leading candidate's vote share using the formula from class (Column J). Then use Excel to indicate whether the result in each election was within the margin of error (Column K).[6]

(c) For what fraction of the polls in the database was the actual result within the calculated margin of error?

PS8 Q1

Let's repeat this assessment for each of the polls in the database. For each poll, use Excel to calculate the margin of error (at the 95% confidence level) for the leading candidate's vote share using the formula from class (Column J). Then use Excel to indicate whether the result in each election was within the margin of error (Column K).[6]

(c) For what fraction of the polls in the database was the actual result within the calculated margin of error?

We call this quantity the **coverage rate**: how often do our confidence intervals contain the true value?

PS8 Q1

We haven't talked about the coverage rate too often in class, but this concept should already be familiar because it is in the definition of a confidence interval!

Confidence Interval: A range around an estimate representing our best guess of the population parameter. For a $(1-\alpha)\%$ confidence interval, $(1-\alpha)\%$ of such intervals contain the true value of the population parameter.

PS8 Q1

Let's build some intuition!! Go to: https://sophieehill.shinyapps.io/ci-sims/