(Null) Hypothesis Testing

Ling250/450: Data Science for Linguistics

C.M. Downey

Spring 2025

Overall Idea

Overall Idea

- Hypothesis testing is the backbone of the **Experimental Scientific Method**

Overall Idea

- Hypothesis testing is the backbone of the **Experimental Scientific Method**
- A hypothesis is a **claim** that can be **supported or refuted by evidence**, usually from a controlled experiment

Overall Idea

- Hypothesis testing is the backbone of the **Experimental Scientific Method**
- A hypothesis is a **claim** that can be **supported or refuted by evidence**, usually from a controlled experiment
  - Experimentalism deals with **evidence** for a claim, rather than proof

Overall Idea

- Hypothesis testing is the backbone of the **Experimental Scientific Method**
- A hypothesis is a **claim** that can be **supported or refuted by evidence**, usually from a controlled experiment
  - Experimentalism deals with **evidence** for a claim, rather than proof
  - Acknowledges that there's **always uncertainty** in the process

Overall Idea

- Hypothesis testing is the backbone of the **Experimental Scientific Method**
- A hypothesis is a **claim** that can be **supported or refuted by evidence**, usually from a controlled experiment
  - Experimentalism deals with **evidence** for a claim, rather than proof
  - Acknowledges that there's **always uncertainty** in the process
  - A hypothesis has to be **falsifiable**: if it can't be shown to be wrong, it's **experimentally useless**

Overall Idea

- Hypothesis testing is the backbone of the **Experimental Scientific Method**
- A hypothesis is a **claim** that can be **supported or refuted by evidence**, usually from a controlled experiment
    - Experimentalism deals with **evidence** for a claim, rather than proof
    - Acknowledges that there's **always uncertainty** in the process
    - A hypothesis has to be **falsifiable**: if it can't be shown to be wrong, it's **experimentally useless**
- In practice, we test new claims against a **null hypothesis** (the hypothesis that our new claim **isn't true**)

Statistical Hypotheses

Statistical Hypotheses

- To test claims with **data**, we have to translate research hypotheses into **statistical hypotheses**

Statistical Hypotheses

- To test claims with **data**, we have to translate research hypotheses into **statistical hypotheses**
  - **Research hypothesis**: clairvoyance/ESP is real, allowing people to "see" things that aren't physically present

Statistical Hypotheses

- To test claims with **data**, we have to translate research hypotheses into **statistical hypotheses**
  - **Research hypothesis**: clairvoyance/ESP is real, allowing people to "see" things that aren't physically present
  - **Experiment**: 100 people guess the result of a coin flip from a separate room

Statistical Hypotheses

- To test claims with **data**, we have to translate research hypotheses into **statistical hypotheses**
  - **Research hypothesis**: clairvoyance/ESP is real, allowing people to "see" things that aren't physically present
  - **Experiment**: 100 people guess the result of a coin flip from a separate room
  - **Statistical hypothesis**: the proportion of subjects that guess the flip correctly will be different from what is expected **by chance** ($\theta = 0.5$)

Statistical Hypotheses

- To test claims with **data**, we have to translate research hypotheses into **statistical hypotheses**
  - **Research hypothesis**: clairvoyance/ESP is real, allowing people to "see" things that aren't physically present
  - **Experiment**: 100 people guess the result of a coin flip from a separate room
  - **Statistical hypothesis**: the proportion of subjects that guess the flip correctly will be different from what is expected **by chance** ($\theta = 0.5$)
- The statistical hypothesis **only** supports the research hypothesis **if the experiment is well-designed** (set up to support or refute ESP)

The Null Hypothesis

The Null Hypothesis

- Counterintuitively, we focus on the **negation** of the hypothesis, called the **Null Hypothesis** or H_0

The Null Hypothesis

- Counterintuitively, we focus on the **negation** of the hypothesis, called the **Null Hypothesis** or H_0
- The Null Hypothesis often represents a **simpler** or **less-informative** state of affairs than the actual hypothesis
  - E.g. H_0 for the ESP experiment is "subjects will guess the coin flip according to **random chance**"

The Null Hypothesis

- Counterintuitively, we focus on the **negation** of the hypothesis, called the **Null Hypothesis** or H_0
- The Null Hypothesis often represents a **simpler** or **less-informative** state of affairs than the actual hypothesis
  - E.g. H_0 for the ESP experiment is "subjects will guess the coin flip according to **random chance**"
- The goal of statistical hypothesis testing is to **refute the Null Hypothesis**, which is otherwise **assumed to be true**

"Null until proven otherwise"

"Null until proven otherwise"

- The book uses the metaphor of a **legal trial**, where a defendant is presumed to be **innocent until proven guilty**

"Null until proven otherwise"

- The book uses the metaphor of a **legal trial**, where a defendant is presumed to be **innocent until proven guilty**
  - The idea: it's **worse to convict an innocent person** than it is to let a guilty person go free

"Null until proven otherwise"

- The book uses the metaphor of a **legal trial**, where a defendant is presumed to be **innocent until proven guilty**
  - The idea: it's **worse to convict an innocent person** than it is to let a guilty person go free
  - (It's worse to **incorrectly refute the Null Hypothesis**)

"Null until proven otherwise"

- The book uses the metaphor of a **legal trial**, where a defendant is presumed to be **innocent until proven guilty**
  - The idea: it's **worse to convict an innocent person** than it is to let a guilty person go free
  - (It's worse to **incorrectly refute the Null Hypothesis**)
- Why? This gives us a **rigorous standard of evidence** for new claims

"Null until proven otherwise"

- The book uses the metaphor of a **legal trial**, where a defendant is presumed to be **innocent until proven guilty**
  - The idea: it's **worse to convict an innocent person** than it is to let a guilty person go free
  - (It's worse to **incorrectly refute the Null Hypothesis**)
- Why? This gives us a **rigorous standard of evidence** for new claims
  - Imagine believing everything you heard was true until proven otherwise

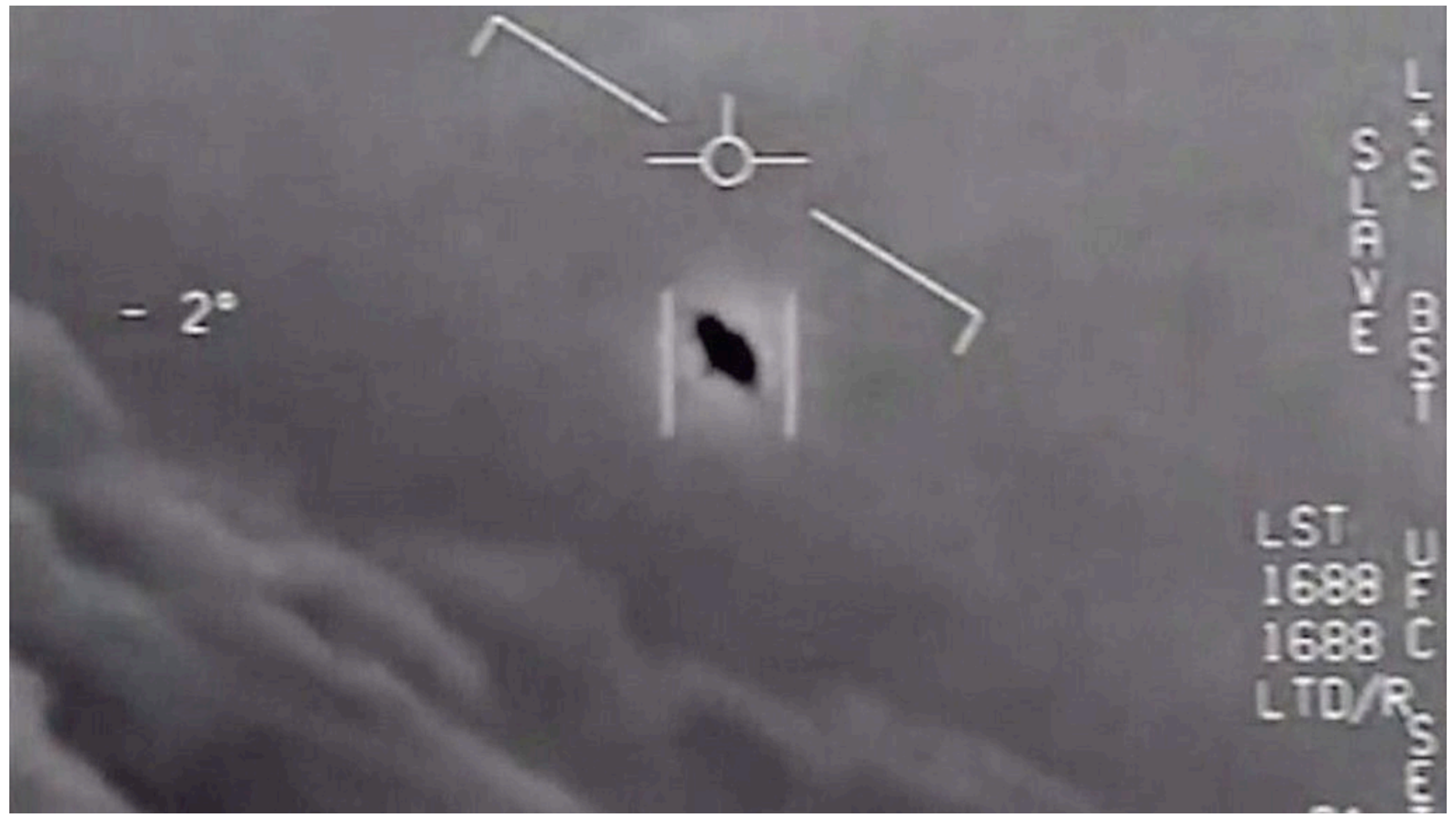
"Null until proven otherwise"

- The book uses the metaphor of a **legal trial**, where a defendant is presumed to be **innocent until proven guilty**
  - The idea: it's **worse to convict an innocent person** than it is to let a guilty person go free
  - (It's worse to **incorrectly refute the Null Hypothesis**)
- Why? This gives us a **rigorous standard of evidence** for new claims
  - Imagine believing everything you heard was true until proven otherwise
  - The **onus is on the scientist** to support their claim

Aliens and the Null Hypothesis

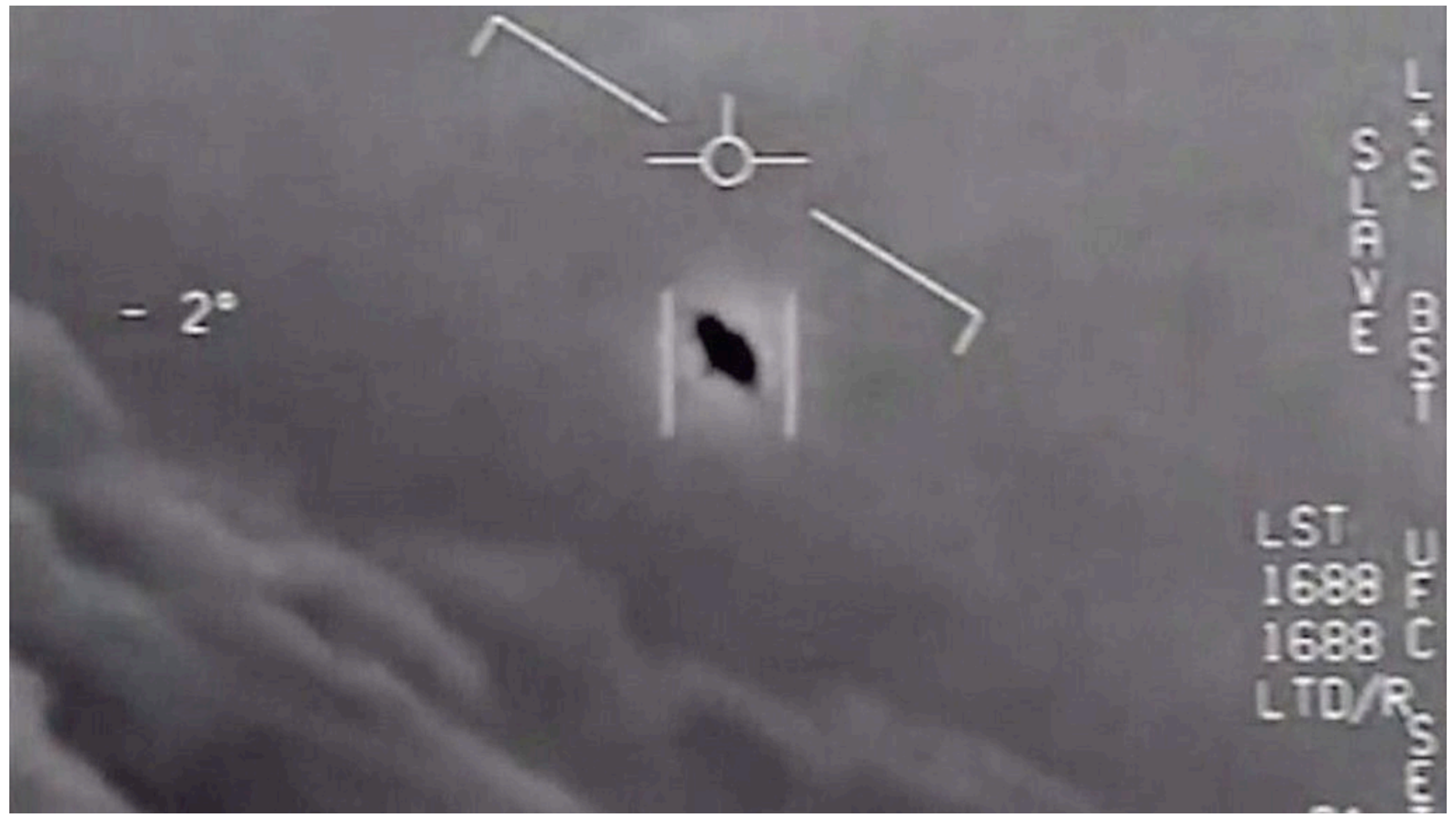
Aliens and the Null Hypothesis

- What's the **Null Hypothesis** when someone says they "saw an alien spacecraft" or that "aliens built the pyramids"?



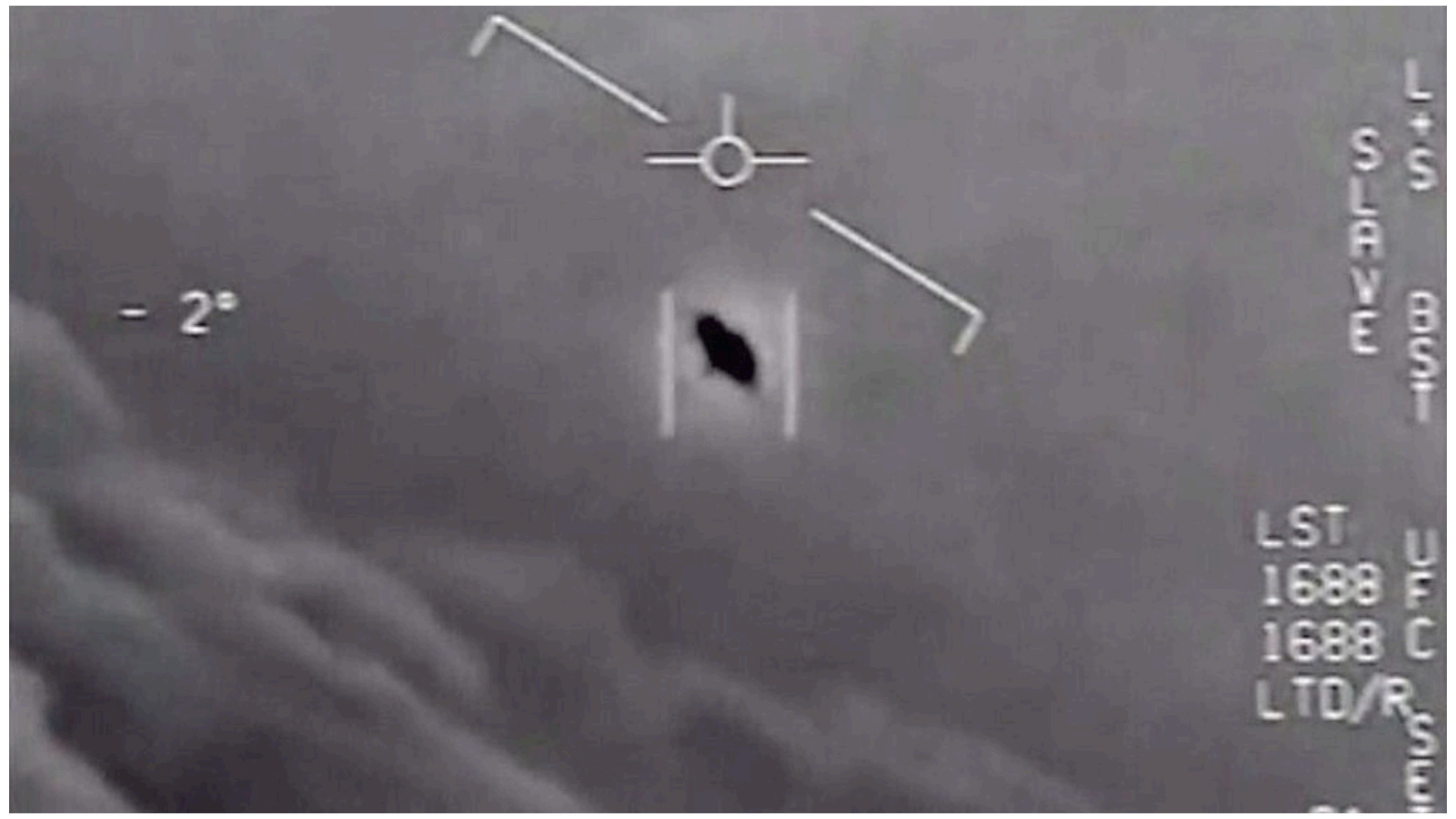
Aliens and the Null Hypothesis

- What's the **Null Hypothesis** when someone says they "saw an alien spacecraft" or that "aliens built the pyramids"?
  - H_0: **not aliens**



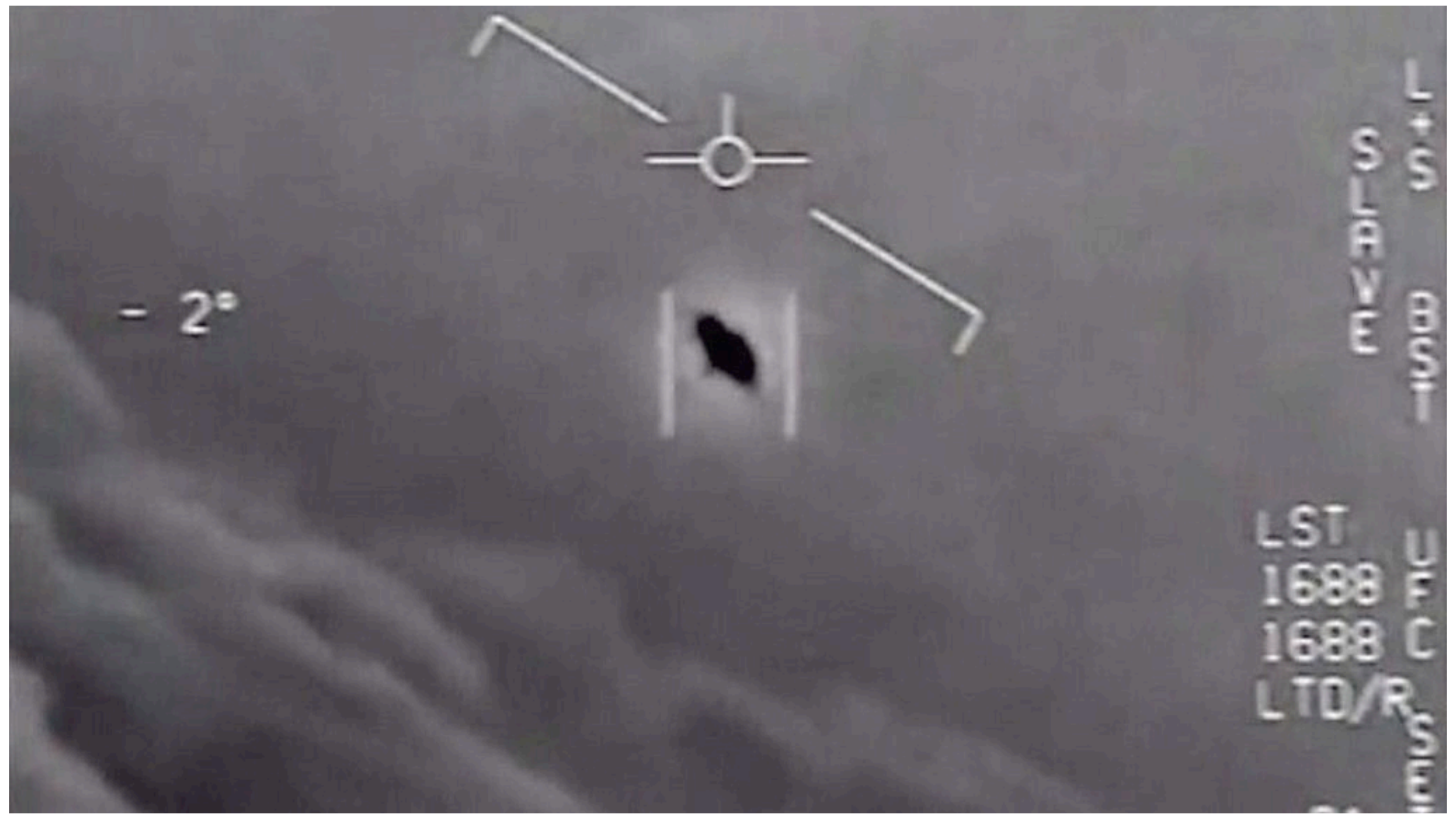
Aliens and the Null Hypothesis

- What's the **Null Hypothesis** when someone says they "saw an alien spacecraft" or that "aliens built the pyramids"?
  - H_0: **not aliens**
  - This is the fallacy behind shows like *Ancient Aliens* or similar "documentaries"



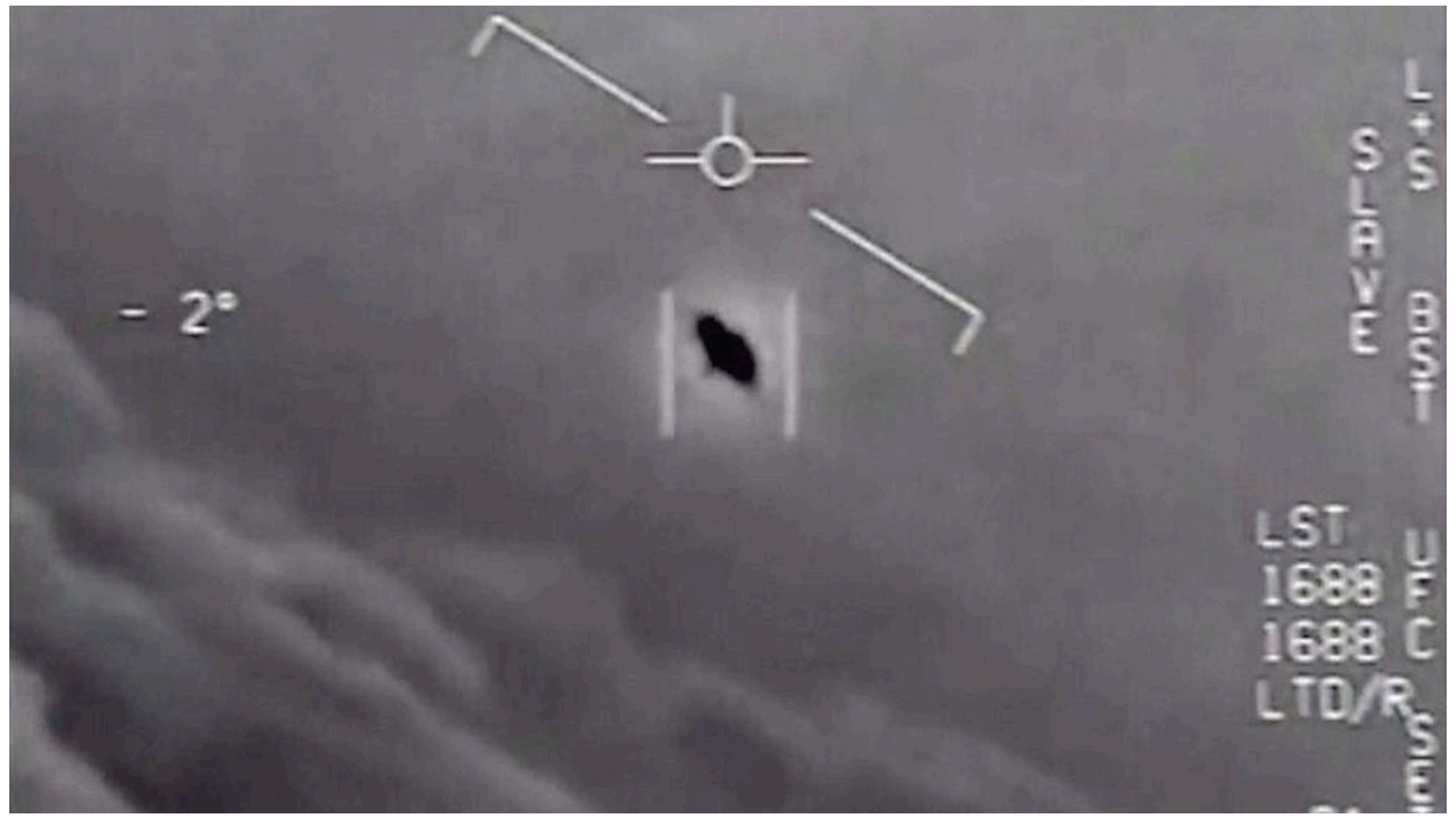
Aliens and the Null Hypothesis

- What's the **Null Hypothesis** when someone says they "saw an alien spacecraft" or that "aliens built the pyramids"?
  - H_0: **not aliens**
  - This is the fallacy behind shows like *Ancient Aliens* or similar "documentaries"
- Important: H_0 is **NOT** an alternative hypothesis like "you saw an airplane"



Aliens and the Null Hypothesis

- What's the **Null Hypothesis** when someone says they "saw an alien spacecraft" or that "aliens built the pyramids"?
    - H_0: **not aliens**
    - This is the fallacy behind shows like *Ancient Aliens* or similar "documentaries"
- Important: H_0 is **NOT** an alternative hypothesis like "you saw an airplane"
- Carl Sagan: **"Extraordinary claims require extraordinary evidence"**



Null Hypothesis for ESP study

Null Hypothesis for ESP study

- Null Hypothesis
  - Non-statistically: "ESP will not factor into subjects guessing the coin"
  - Statistically: "The subjects will only be able to guess the outcome of the coin flip by chance, with a probability of $\theta = 0.5$"

Null Hypothesis for ESP study

- Null Hypothesis
  - Non-statistically: "ESP will not factor into subjects guessing the coin"
  - Statistically: "The subjects will only be able to guess the outcome of the coin flip by chance, with a probability of $\theta = 0.5$"
- The goal of the test is to **retain** (keep) or **reject** the Null Hypothesis

Null Hypothesis for ESP study

- Null Hypothesis
  - Non-statistically: "ESP will not factor into subjects guessing the coin"
  - Statistically: "The subjects will only be able to guess the outcome of the coin flip by chance, with a probability of $\theta = 0.5$"
- The goal of the test is to **retain** (keep) or **reject** the Null Hypothesis
- "Alternative" hypothesis (what we call the **non-null** hypothesis)
  - "Subjects will be able to guess the outcome of the coin flip with a probability different from chance $\theta \neq 0.5$"
  - The **experimental design** suggests that this means ESP

Types of Error

Types of Error

- Hypothesis testing can result in **two types of error**
  - **Type I:** the null is **incorrectly rejected**
  - **Type II:** the null is **incorrectly retained**

	retain H_0	reject H_0
H_0 is true	correct decision	error (type I)
H_0 is false	error (type II)	correct decision

Types of Error

- Hypothesis testing can result in **two types of error**
  - **Type I:** the null is **incorrectly rejected**
  - **Type II:** the null is **incorrectly retained**
- Goal is to **minimize Type I errors**
  - "Null until proven otherwise"
  - Type 1 error rate of a test is its **significance level** (α)

	retain H_0	reject H_0
H_0 is true	correct decision	error (type I)
H_0 is false	error (type II)	correct decision

Significance and Power

	retain H_0	reject H_0
H_0 is true	$1 - \alpha$ (probability of correct retention)	α (type I error rate)
H_0 is false	β (type II error rate)	$1 - \beta$ (power of the test)

Significance and Power

- α: probability the test gives a **Type I error** (incorrectly reject the null)
  - AKA: significance

	retain H_0	reject H_0
H_0 is true	$1-\alpha$ (probability of correct retention)	α (type I error rate)
H_0 is false	β (type II error rate)	$1-\beta$ (power of the test)

Significance and Power

- α: probability the test gives a **Type I error** (incorrectly reject the null)
  - AKA: significance
- β: probability the test gives a **Type II error** (incorrectly retain the null)
  - We don't talk about β so much as $1-\beta$, which is called the **power** of the test
  - The power is the probability that we **correctly reject the null**
  - Power is closely tied to the **sample size** of the test

	retain H_0	reject H_0
H_0 is true	$1-\alpha$ (probability of correct retention)	α (type I error rate)
H_0 is false	β (type II error rate)	$1-\beta$ (power of the test)

Significance and Power

- α: probability the test gives a **Type I error** (incorrectly reject the null)
  - AKA: significance
- β: probability the test gives a **Type II error** (incorrectly retain the null)
  - We don't talk about β so much as $1-\beta$, which is called the **power** of the test
  - The power is the probability that we **correctly reject the null**
  - Power is closely tied to the **sample size** of the test
- Significance is **prioritized** over power!

	retain H_0	reject H_0
H_0 is true	$1-\alpha$ (probability of correct retention)	α (type I error rate)
H_0 is false	β (type II error rate)	$1-\beta$ (power of the test)

Elements of a test

Elements of a test

- **Test statistic**: the data variable used to support or refute the hypothesis
  - I.e. what are we measuring?

Elements of a test

- **Test statistic**: the data variable used to support or refute the hypothesis
  - I.e. what are we measuring?
- **Significance level**: what is the **highest Type I error rate** that we're willing to accept?
  - What chance of **incorrectly rejecting** H_0 are we okay with?

Elements of a test

- **Test statistic**: the data variable used to support or refute the hypothesis
    - I.e. what are we measuring?
- **Significance level**: what is the **highest Type I error rate** that we're willing to accept?
    - What chance of **incorrectly rejecting** H_0 are we okay with?
- **Sampling distribution** of the test statistic:
    - How do we **expect** the data to be distributed **assuming the null is true?**

ESP test

ESP test

- **Test statistic**: the number of **correct guesses** out of the total trials (say 100 total subjects, 1 test each)

ESP test

- **Test statistic**: the number of **correct guesses** out of the total trials (say 100 total subjects, 1 test each)
- **Significance level**: let's say **0.05 (5%)** to start wtih
  - This is a little **arbitrary**, but it's a common threshold in science

ESP test

- **Test statistic**: the number of **correct guesses** out of the total trials (say 100 total subjects, 1 test each)
- **Significance level**: let's say **0.05 (5%)** to start wtih
  - This is a little **arbitrary**, but it's a common threshold in science
- **Sampling distribution**: the number of correct guesses will follow the **Binomial Distribution**
  - $\theta = 0.5$ if subjects are guessing randomly (Null Hypothesis)

ESP test

- **Test statistic**: the number of **correct guesses** out of the total trials (say 100 total subjects, 1 test each)
- **Significance level**: let's say **0.05 (5%)** to start wtih
  - This is a little **arbitrary**, but it's a common threshold in science
- **Sampling distribution**: the number of correct guesses will follow the **Binomial Distribution**
  - $\theta = 0.5$ if subjects are guessing randomly (Null Hypothesis)

$$X \sim \text{Binomial}(\theta, N)$$

Sampling Distribution

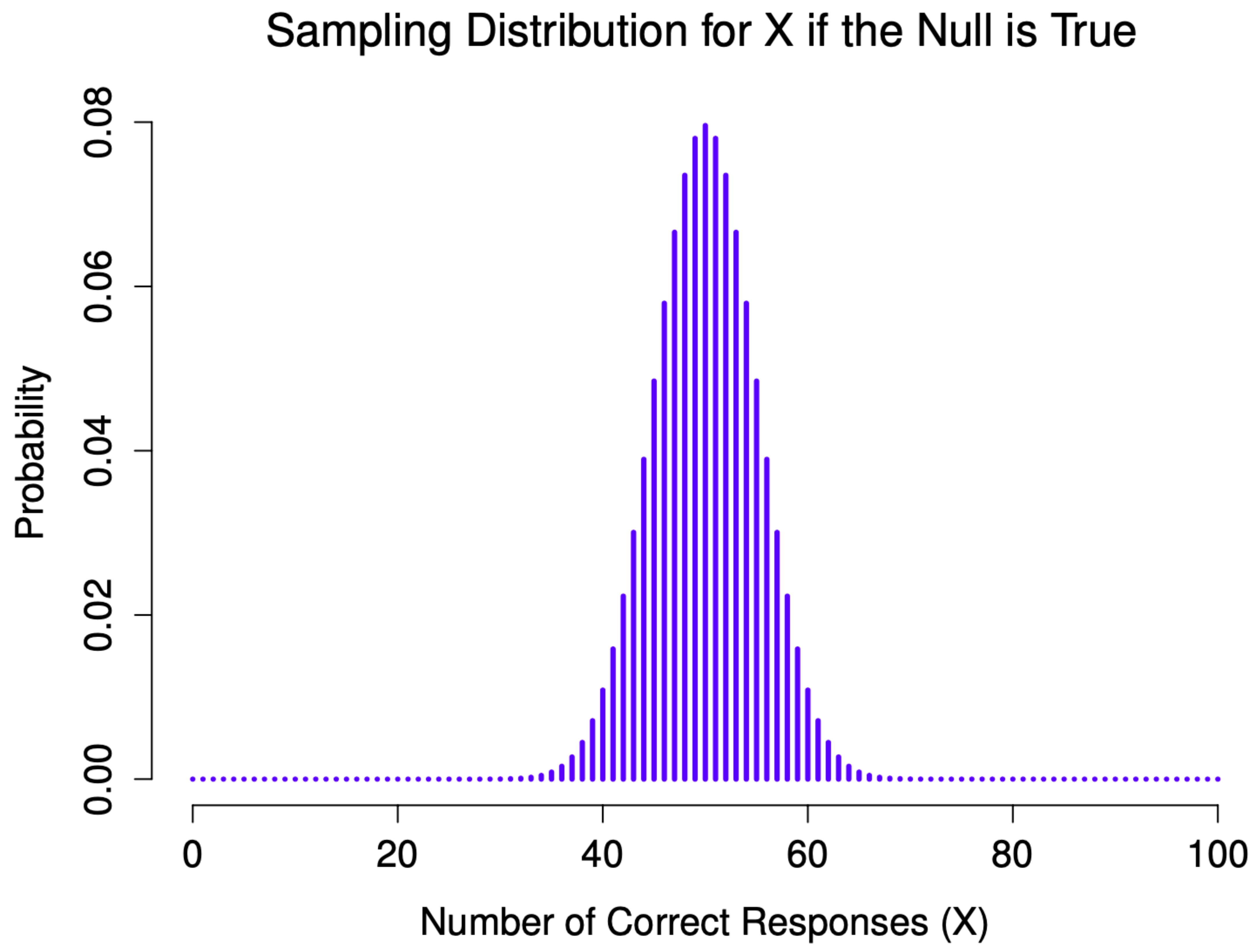


Figure 11.1: The sampling distribution for our test statistic X when the null hypothesis is true. For our ESP scenario, this is a binomial distribution. Not surprisingly, since the null hypothesis says that the probability of a correct response is $\theta = .5$, the sampling distribution says that the most likely value is 50 (our of 100) correct responses. Most of the probability mass lies between 40 and 60.

Expected results

Expected results

- About how many correct guesses do we expect **assuming the Null is true?**

Expected results

- About how many correct guesses do we expect **assuming the Null is true?**
  - **About 50** (the binomial distribution gives our expected results)

Expected results

- About how many correct guesses do we expect **assuming the Null is true?**
  - **About 50** (the binomial distribution gives our expected results)
- How many correct guesses do we expect **if the Null is false?**

Expected results

- About how many correct guesses do we expect **assuming the Null is true?**
  - **About 50** (the binomial distribution gives our expected results)
- How many correct guesses do we expect **if the Null is false?**
  - Somewhere **not close to 50!**

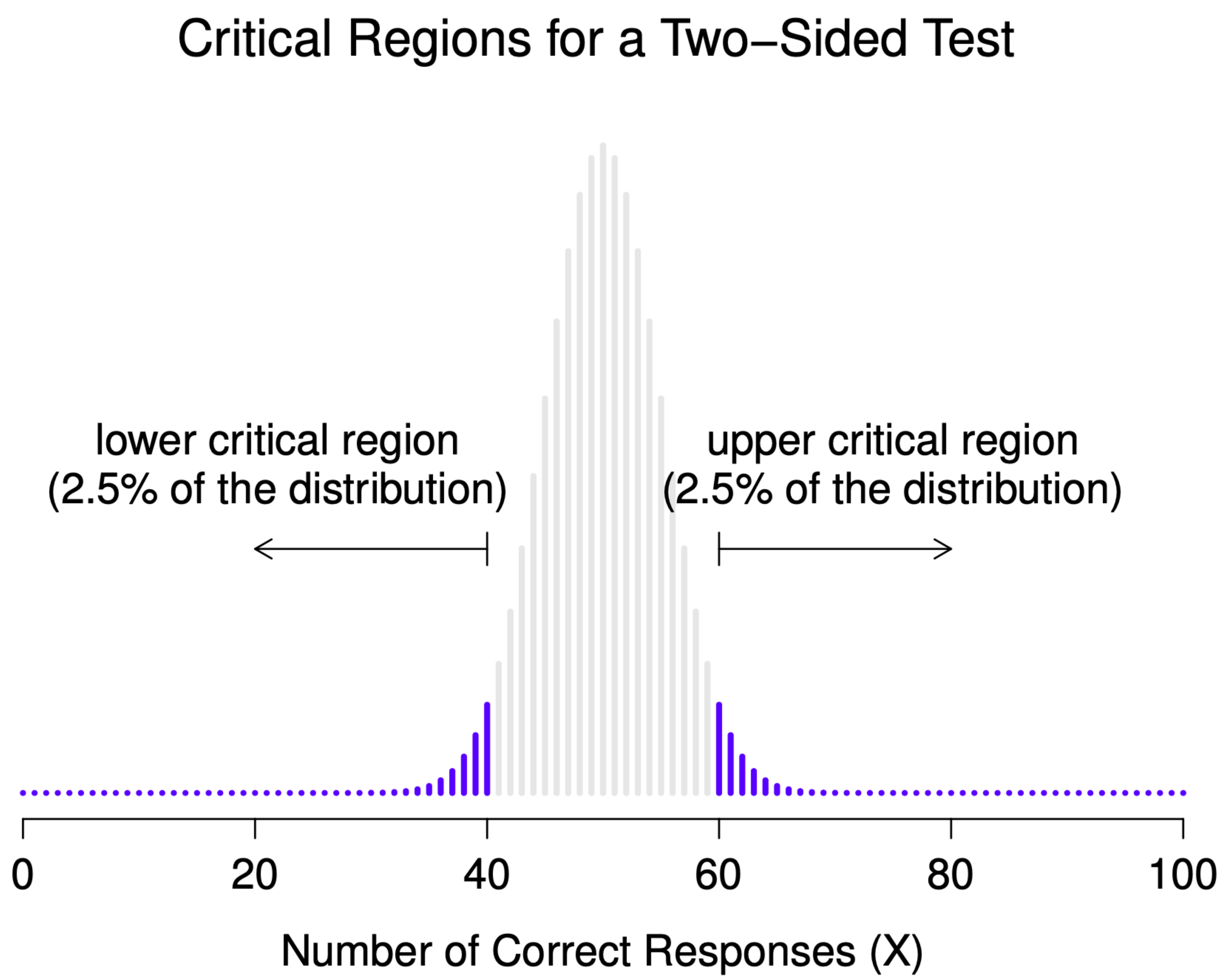
Expected results

- About how many correct guesses do we expect **assuming the Null is true?**
  - **About 50** (the binomial distribution gives our expected results)
- How many correct guesses do we expect **if the Null is false?**
  - Somewhere **not close to 50!**
  - Our alternative hypothesis is **not specific** about whether to expect less or more

Expected results

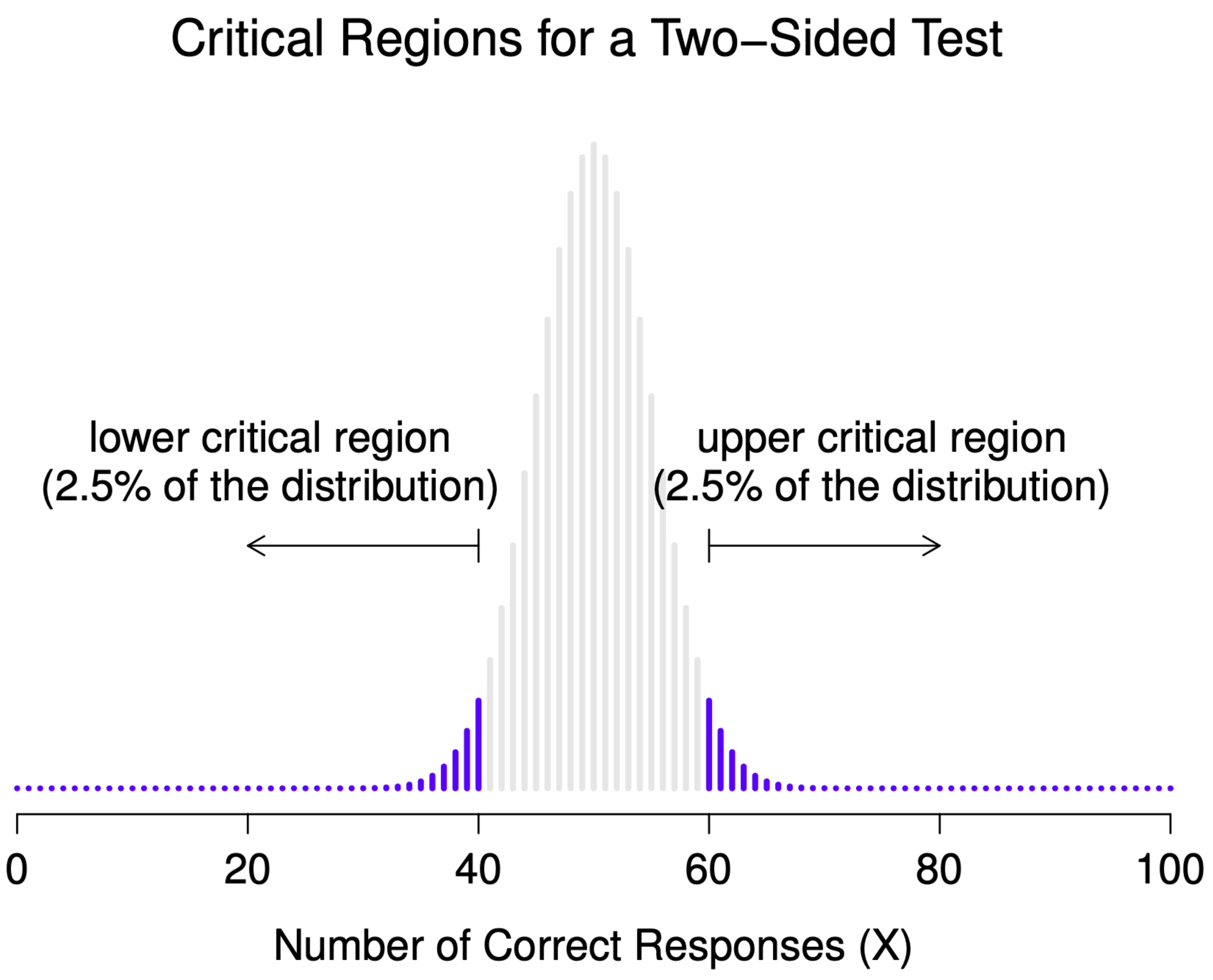
- About how many correct guesses do we expect **assuming the Null is true?**
  - **About 50** (the binomial distribution gives our expected results)
- How many correct guesses do we expect **if the Null is false?**
  - Somewhere **not close to 50!**
  - Our alternative hypothesis is **not specific** about whether to expect less or more
- How much do our results need to **diverge from expectation** in order to declare them **significant**?

Critical regions



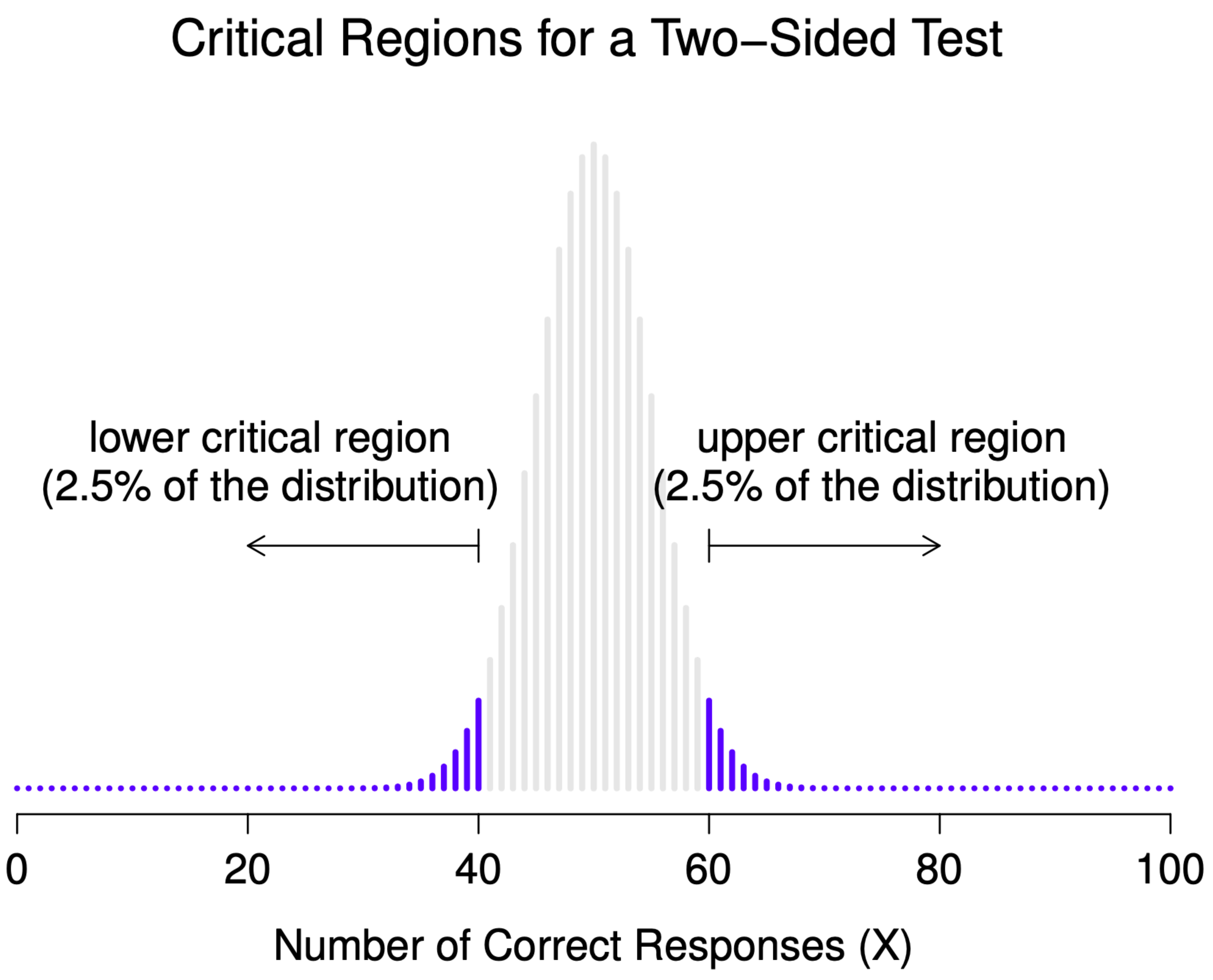
Critical regions

- **Critical regions** are the values of the test statistic (X) which **lead us to reject the Null**



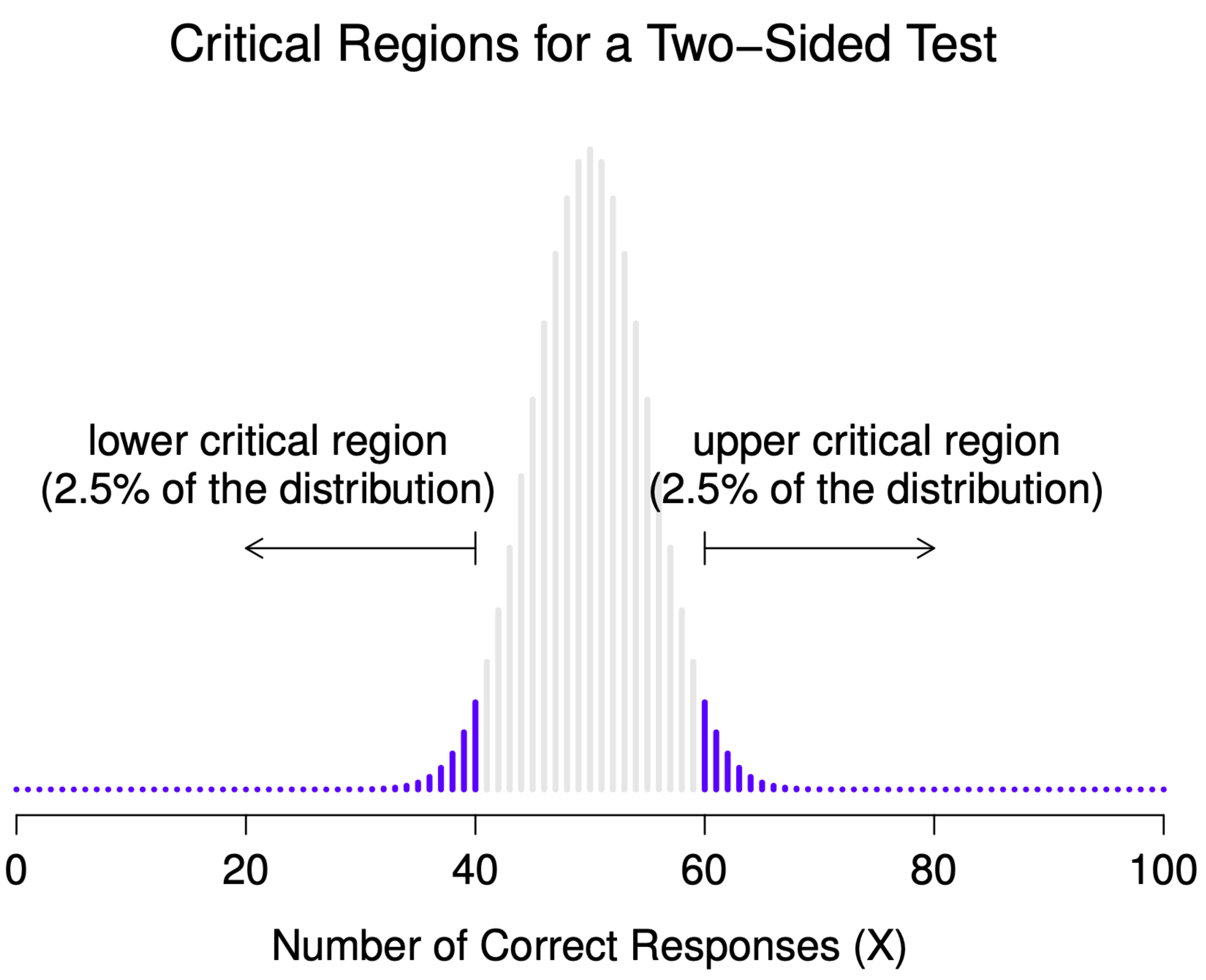
Critical regions

- **Critical regions** are the values of the test statistic (X) which **lead us to reject the Null**
- For significance $\alpha = 0.05$, this is the **outer 5%** of the sampling distribution



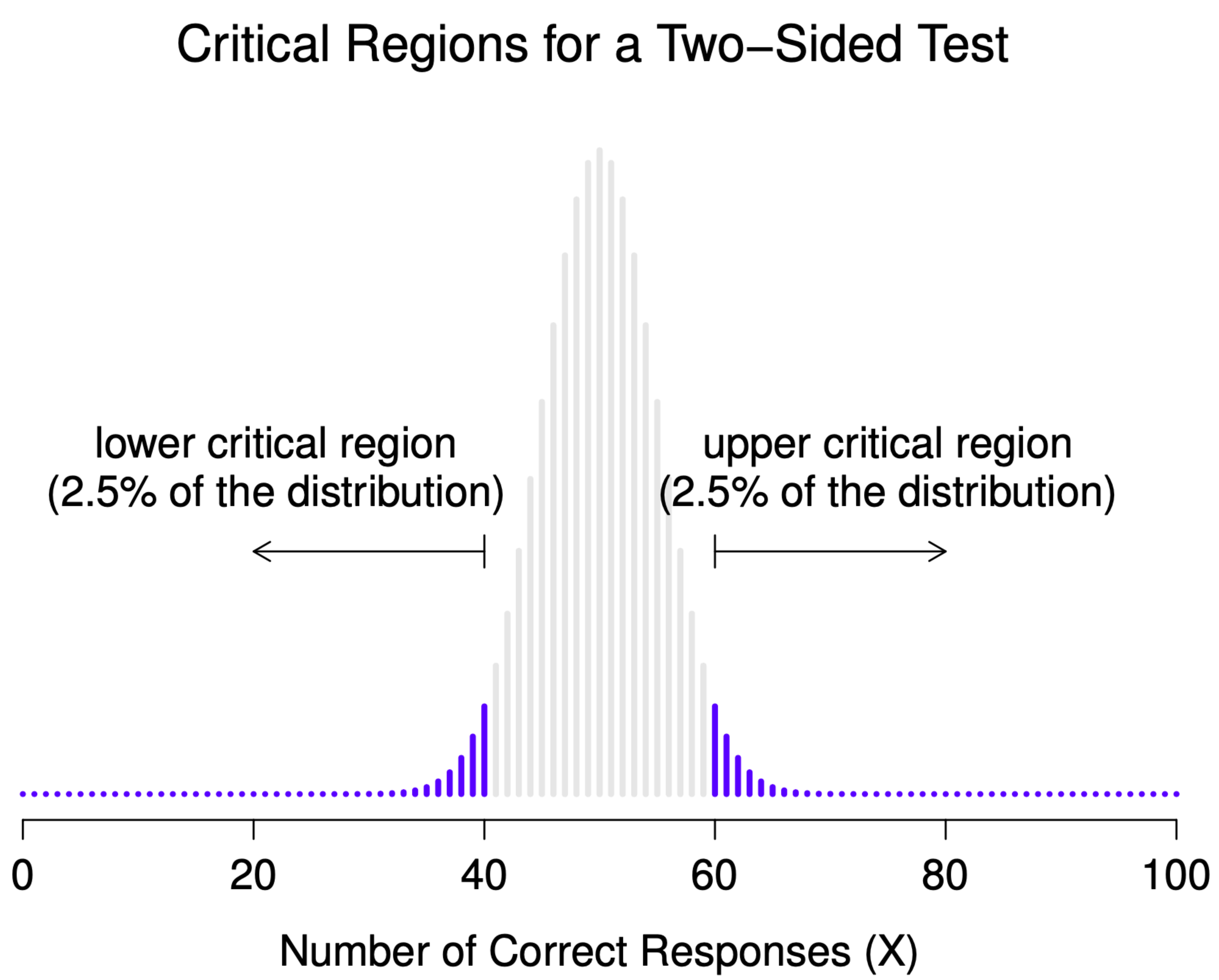
Critical regions

- **Critical regions** are the values of the test statistic (X) which **lead us to reject the Null**
- For significance $\alpha = 0.05$, this is the **outer 5%** of the sampling distribution
    - These values are **possible but unlikely** under the Null

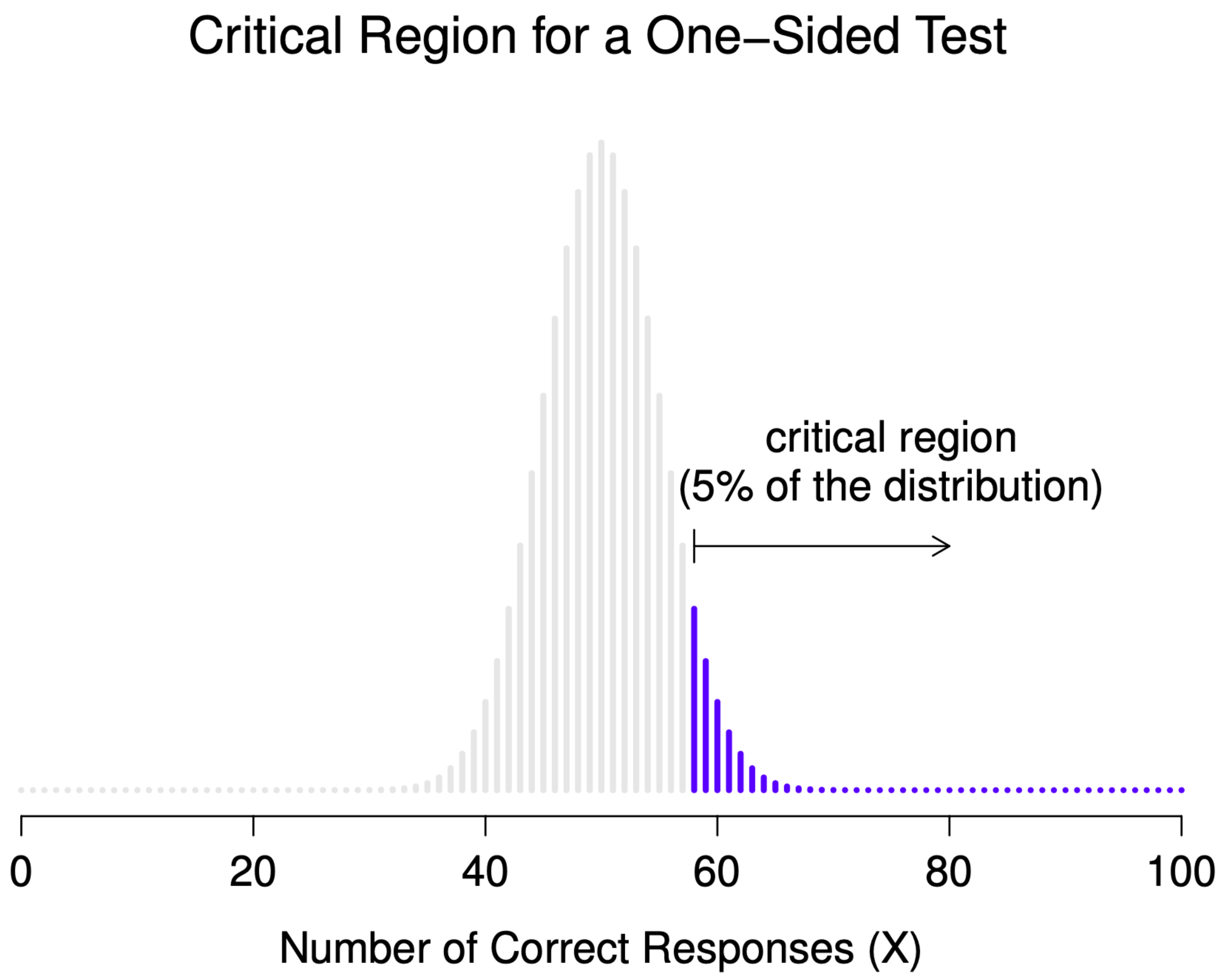


Critical regions

- **Critical regions** are the values of the test statistic (X) which **lead us to reject the Null**
- For significance $\alpha = 0.05$, this is the **outer 5%** of the sampling distribution
  - These values are **possible but unlikely** under the Null
  - Thus: α is the chance of **incorrectly rejecting** the Null

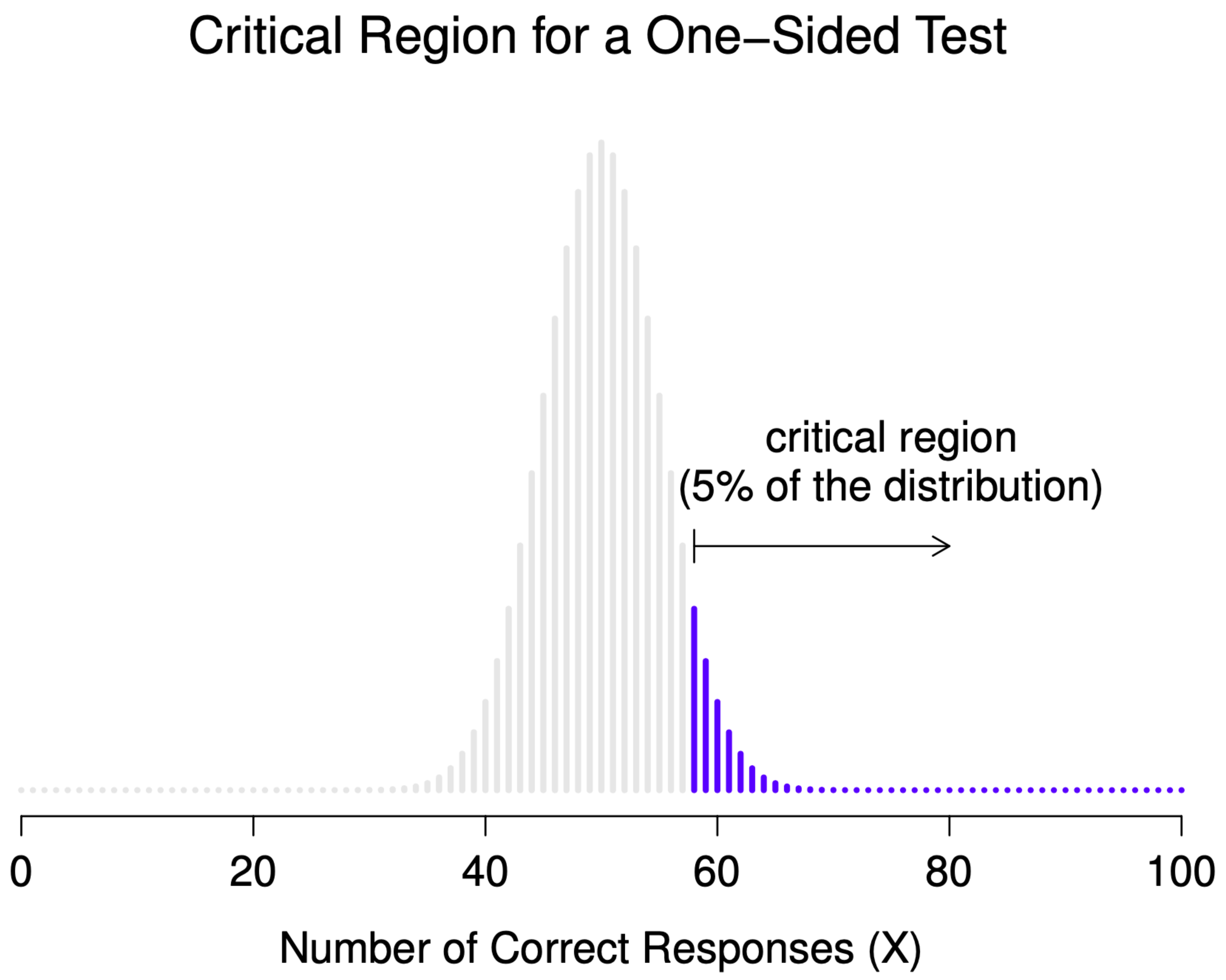


One-sided test



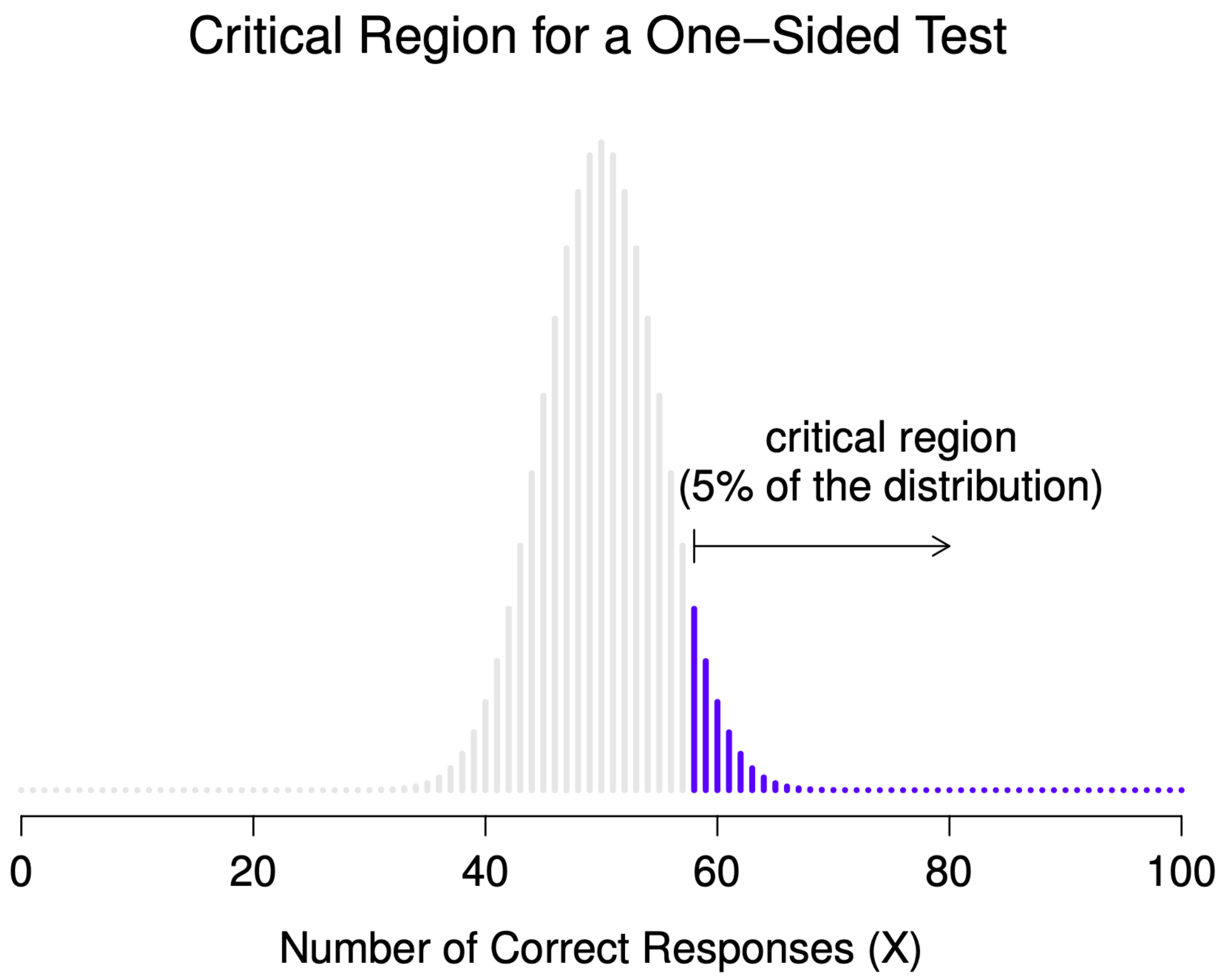
One-sided test

- A **one-sided** test has a critical region only to one side of the mean



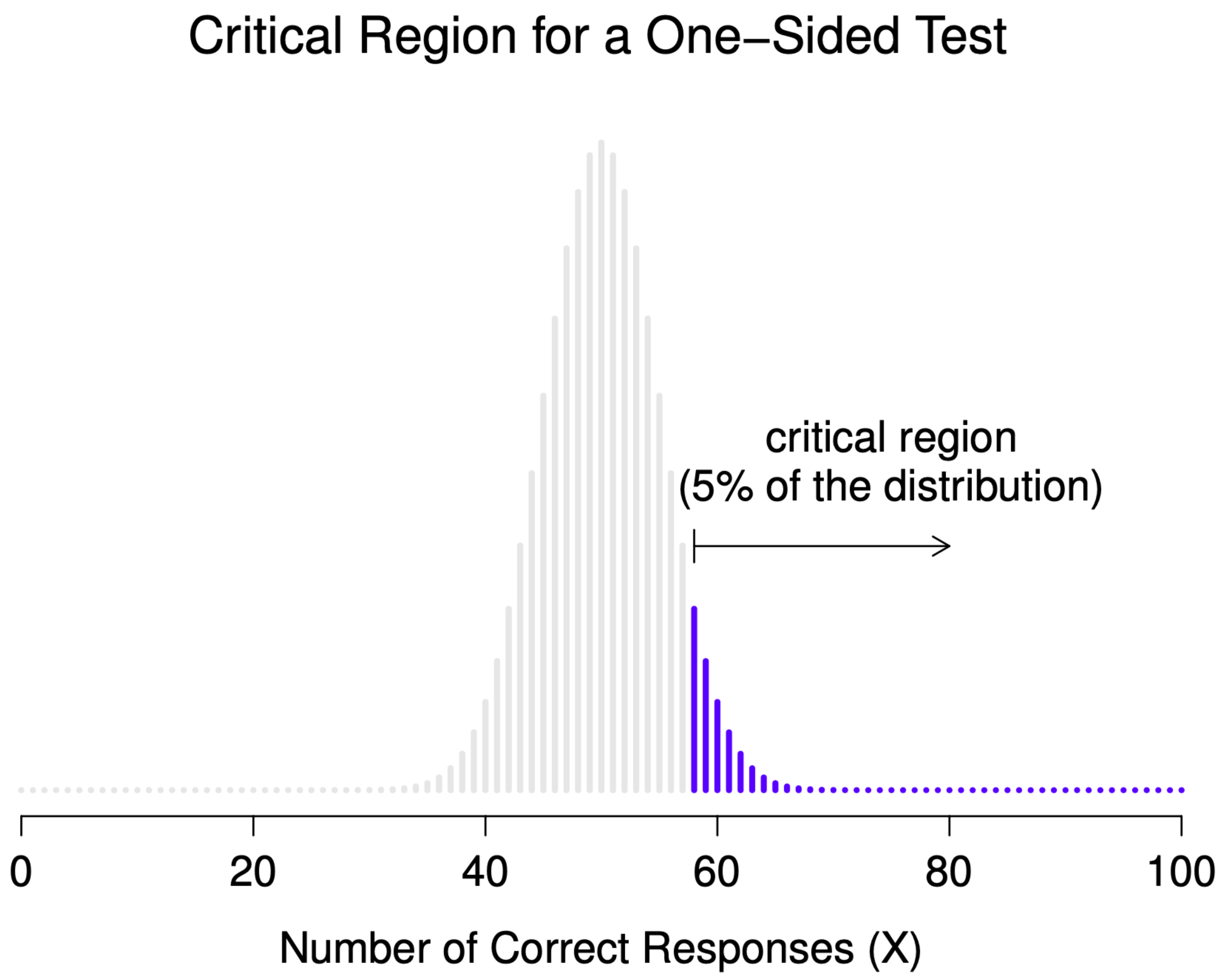
One-sided test

- A **one-sided** test has a critical region only to one side of the mean
- This represents the hypothesis that ESP **increases** accuracy (rather than just having an effect)
  - $H_0 : X \leq 0.5$
  - $H_1 : X > 0.5$

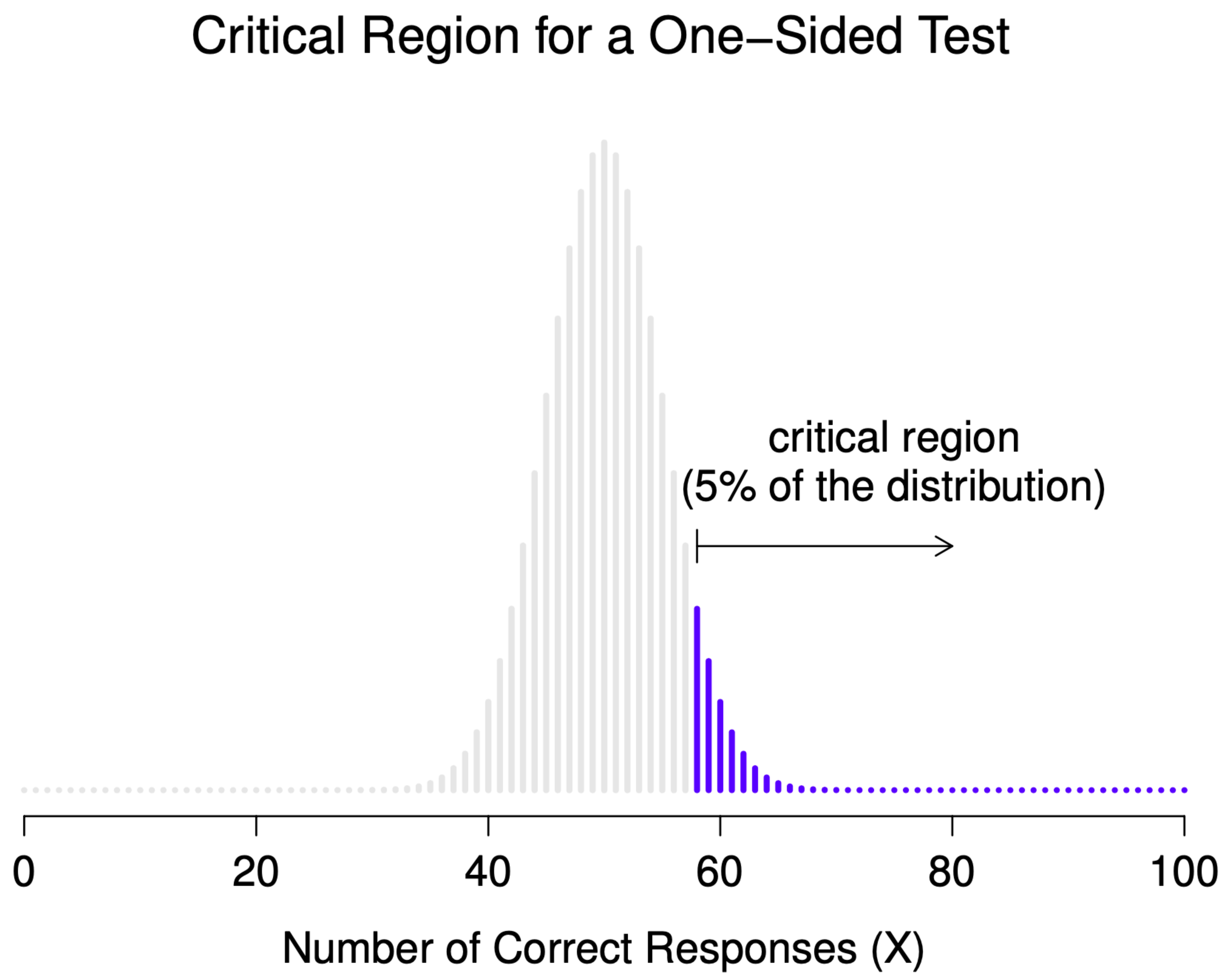


One-sided test

- A **one-sided** test has a critical region only to one side of the mean
- This represents the hypothesis that ESP **increases** accuracy (rather than just having an effect)
  - $H_0 : X \leq 0.5$
  - $H_1 : X > 0.5$
- The alternative is **two-sided**

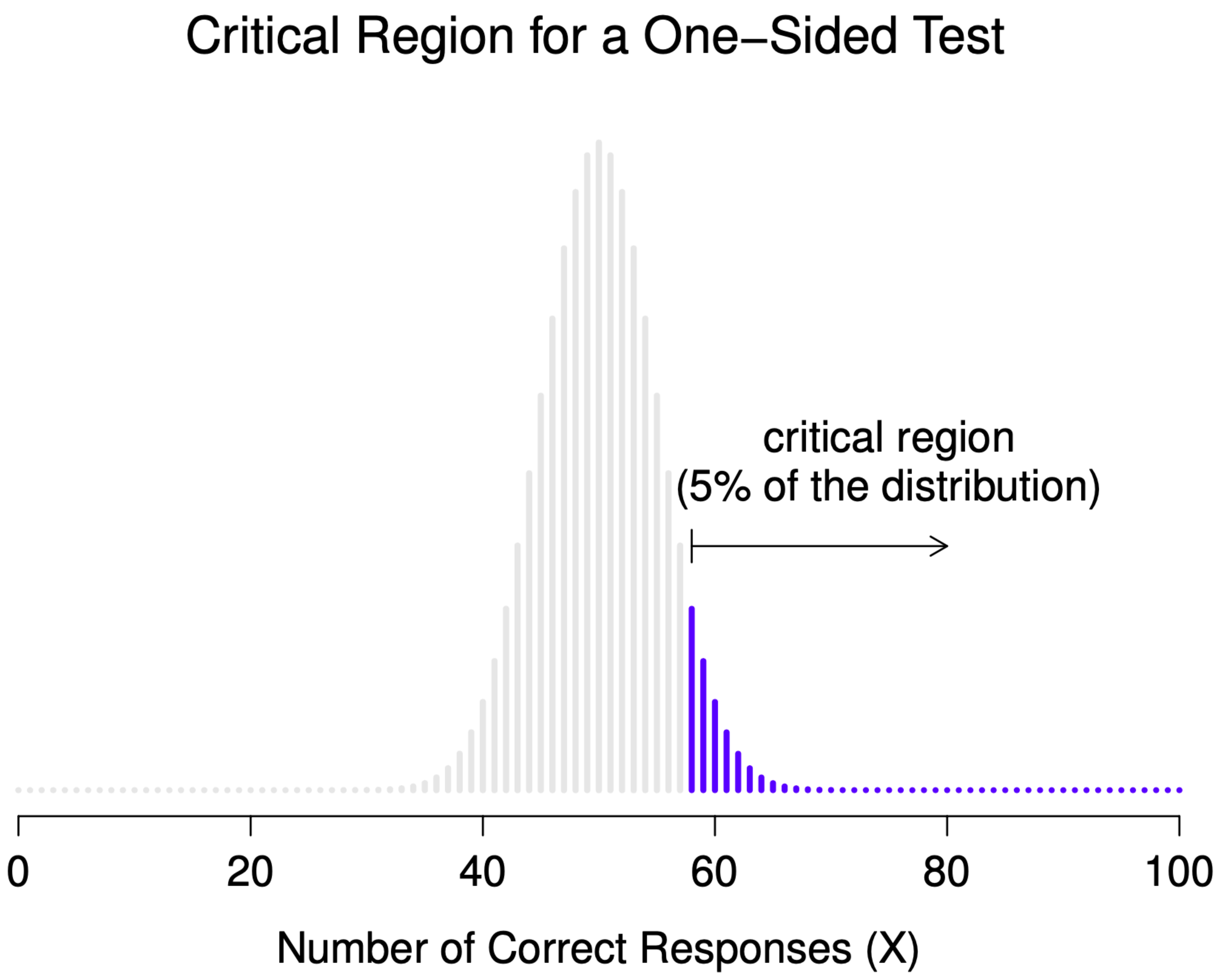


Are you convinced?



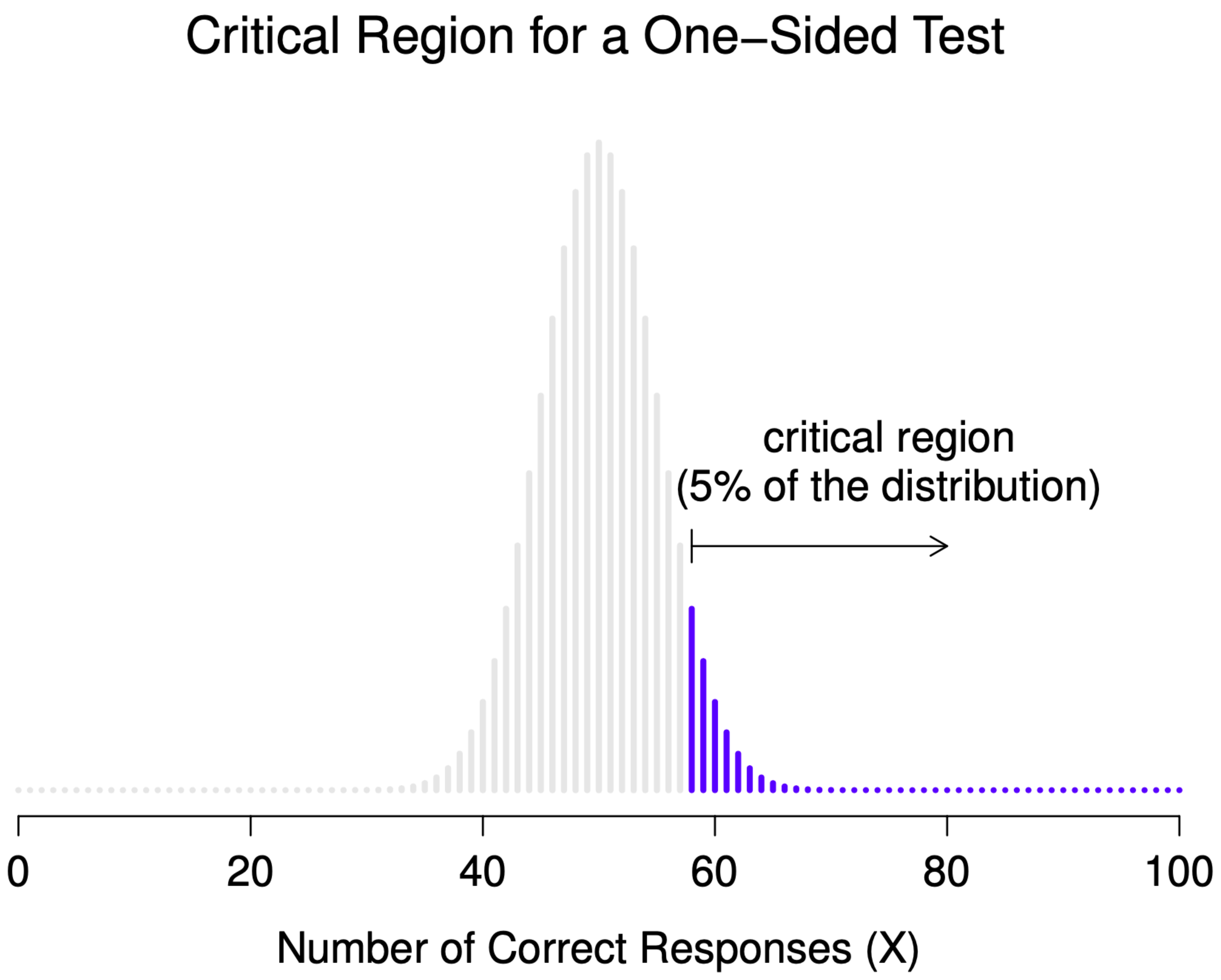
Are you convinced?

- Let's say we're doing a **one-sided** test, with a **significance threshold** of $\alpha = 0.05$



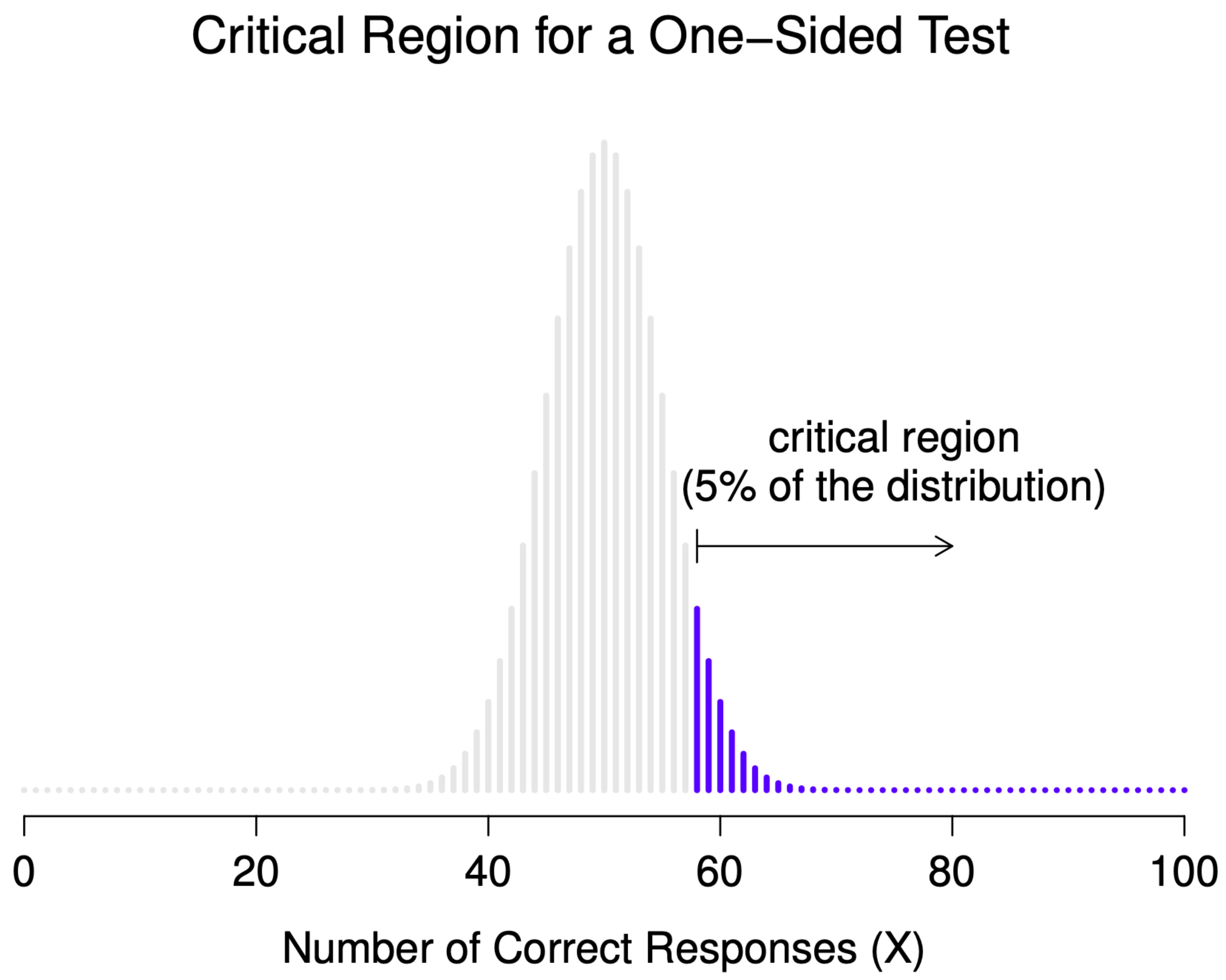
Are you convinced?

- Let's say we're doing a **one-sided test**, with a **significance threshold** of $\alpha = 0.05$
- Let's say that **60 subjects guess correctly** in the experiment



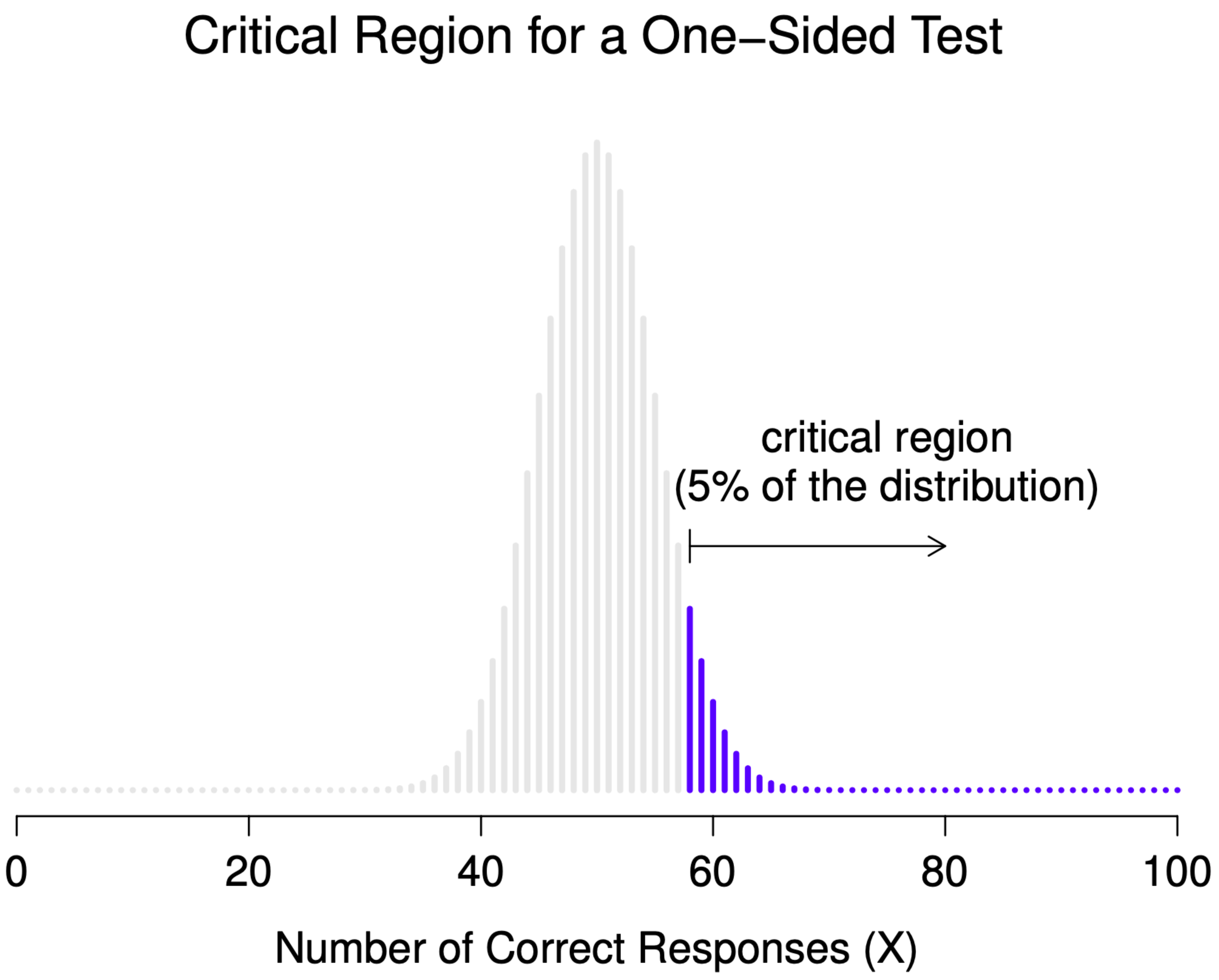
Are you convinced?

- Let's say we're doing a **one-sided** test, with a **significance threshold** of $\alpha = 0.05$
- Let's say that **60 subjects guess correctly** in the experiment
- This is **"statistically significant"** given our definitions



Are you convinced?

- Let's say we're doing a **one-sided** test, with a **significance threshold** of $\alpha = 0.05$
- Let's say that **60 subjects guess correctly** in the experiment
- This is **"statistically significant"** given our definitions
- Do you believe in ESP?



p-values

Usual notation	Signif. stars	English translation
$p > .05$		The test wasn't significant
$p < .05$	*	The test was significant at $\alpha = .05$ but not at $\alpha = .01$ or $\alpha = .001$.
$p < .01$	**	The test was significant at $\alpha = .05$ and $\alpha = .01$ but not at $\alpha = .001$.
$p < .001$	***	The test was significant at all levels

p-values

- The p-value is the **probability** assigned to results that are **at least as extreme** as the one we have, assuming the Null

Usual notation	Signif. stars	English translation
$p > .05$		The test wasn't significant
$p < .05$	*	The test was significant at $\alpha = .05$ but not at $\alpha = .01$ or $\alpha = .001$.
$p < .01$	**	The test was significant at $\alpha = .05$ and $\alpha = .01$ but not at $\alpha = .001$.
$p < .001$	***	The test was significant at all levels

p-values

- The p-value is the **probability** assigned to results that are **at least as extreme** as the one we have, assuming the Null
  - Essentially: "How probable would this data (or more extreme data be assuming the Null is true?"

Usual notation	Signif. stars	English translation
$p > .05$		The test wasn't significant
$p < .05$	*	The test was significant at $\alpha = .05$ but not at $\alpha = .01$ or $\alpha = .001$.
$p < .01$	**	The test was significant at $\alpha = .05$ and $\alpha = .01$ but not at $\alpha = .001$.
$p < .001$	***	The test was significant at all levels

p-values

- The p-value is the **probability** assigned to results that are **at least as extreme** as the one we have, assuming the Null
  - Essentially: "How probable would this data (or more extreme data be assuming the Null is true?"
  - The **more unlikely** a result is assuming the Null, the **more significant**

Usual notation	Signif. stars	English translation
$p > .05$		The test wasn't significant
$p < .05$	*	The test was significant at $\alpha = .05$ but not at $\alpha = .01$ or $\alpha = .001$.
$p < .01$	**	The test was significant at $\alpha = .05$ and $\alpha = .01$ but not at $\alpha = .001$.
$p < .001$	***	The test was significant at all levels

p-values

- The p-value is the **probability** assigned to results that are **at least as extreme** as the one we have, assuming the Null
  - Essentially: "How probable would this data (or more extreme data be assuming the Null is true?"
  - The **more unlikely** a result is assuming the Null, the **more significant**
- Many/most **scientific papers** report the p-value of their result

Usual notation	Signif. stars	English translation
$p > .05$		The test wasn't significant
$p < .05$	*	The test was significant at $\alpha = .05$ but not at $\alpha = .01$ or $\alpha = .001$.
$p < .01$	**	The test was significant at $\alpha = .05$ and $\alpha = .01$ but not at $\alpha = .001$.
$p < .001$	***	The test was significant at all levels

p-values

- The p-value is the **probability** assigned to results that are **at least as extreme** as the one we have, assuming the Null
    - Essentially: "How probable would this data (or more extreme data be assuming the Null is true?"
    - The **more unlikely** a result is assuming the Null, the **more significant**
- Many/most **scientific papers** report the p-value of their result
- The p-value is **NOT** the "probability that the Null is true"

Usual notation	Signif. stars	English translation
$p > .05$		The test wasn't significant
$p < .05$	*	The test was significant at $\alpha = .05$ but not at $\alpha = .01$ or $\alpha = .001$.
$p < .01$	**	The test was significant at $\alpha = .05$ and $\alpha = .01$ but not at $\alpha = .001$.
$p < .001$	***	The test was significant at all levels

ESP p-value (ESP-value?)

```
> binom.test(x=60, n=100, p=0.5)

        Exact binomial test

data:  60 and 100
number of successes = 60, number of
trials = 100, p-value = 0.05689
alternative hypothesis: true probability of succe
ss is not equal to 0.5
95 percent confidence interval:
 0.4972092 0.6967052
sample estimates:
probability of success
                   0.6
```

ESP p-value (ESP-value?)

- We can get the p-value of our ESP experiment by checking the **probability that the Null assigns** to our result

```
> binom.test(x=60, n=100, p=0.5)

        Exact binomial test

data:  60 and 100
number of successes = 60, number of
trials = 100, p-value = 0.05689
alternative hypothesis: true probability of succe
ss is not equal to 0.5
95 percent confidence interval:
 0.4972092 0.6967052
sample estimates:
probability of success
                   0.6
```

ESP p-value (ESP-value?)

- We can get the p-value of our ESP experiment by checking the **probability that the Null assigns** to our result
  - The p-value says there is a 5.7% chance of getting **60 or more OR 40 or fewer** successes (given the Null)

```
> binom.test(x=60, n=100, p=0.5)

        Exact binomial test

data:  60 and 100
number of successes = 60, number of
trials = 100, p-value = 0.05689
alternative hypothesis: true probability of succe
ss is not equal to 0.5
95 percent confidence interval:
 0.4972092 0.6967052
sample estimates:
probability of success
                   0.6
```

ESP p-value (ESP-value?)

- We can get the p-value of our ESP experiment by checking the **probability that the Null assigns** to our result
  - The p-value says there is a 5.7% chance of getting **60 or more OR 40 or fewer** successes (given the Null)
- Can get a **one-sided** test with `binom.test(x, n, p, alternative="greater")`

```
> binom.test(x=60, n=100, p=0.5)

        Exact binomial test

data:  60 and 100
number of successes = 60, number of
trials = 100, p-value = 0.05689
alternative hypothesis: true probability of succe
ss is not equal to 0.5
95 percent confidence interval:
 0.4972092 0.6967052
sample estimates:
probability of success
                   0.6
```

Power

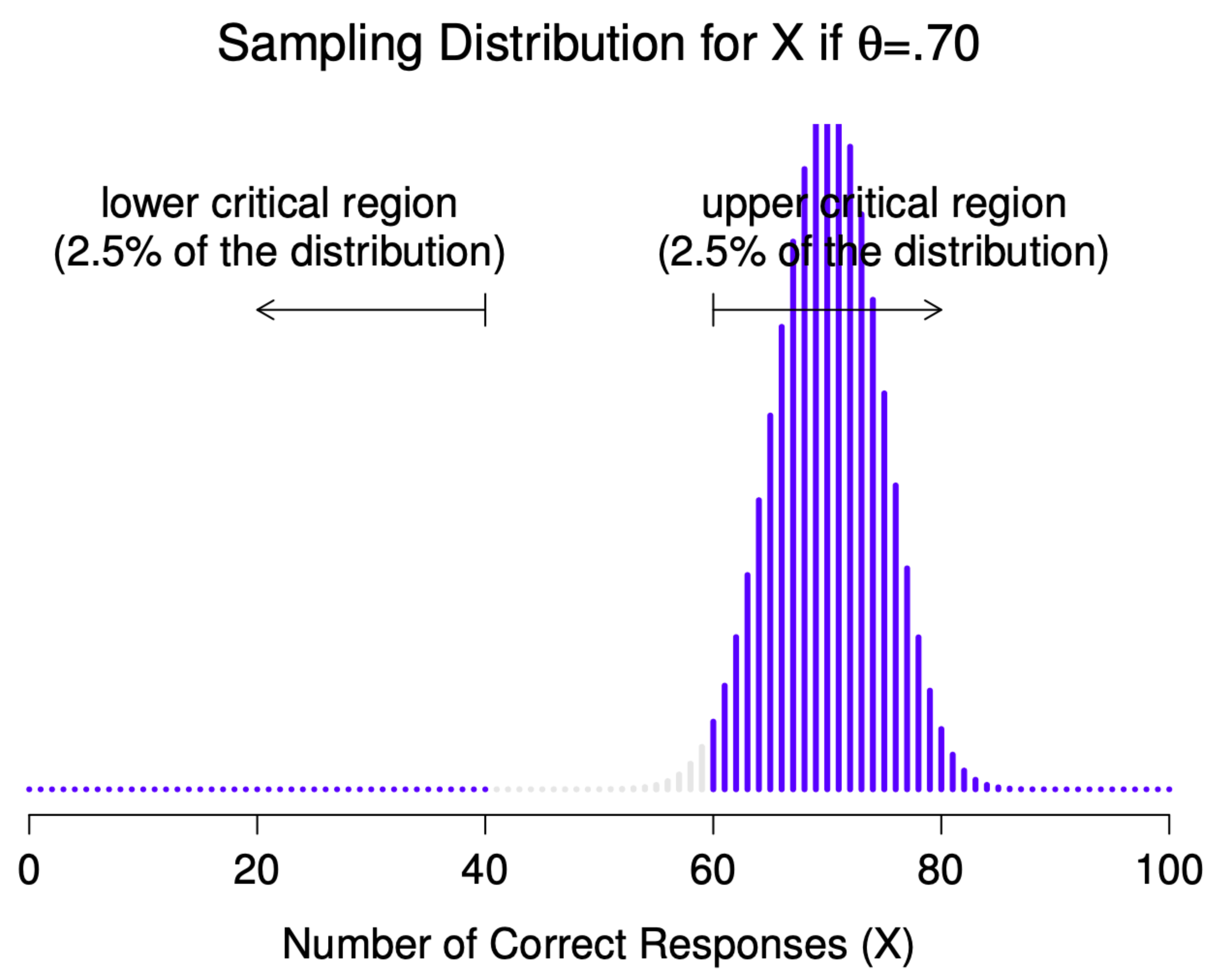


Figure 11.5: Sampling distribution under the *alternative* hypothesis, for a population parameter value of $\theta = 0.70$. Almost all of the distribution lies in the rejection region.