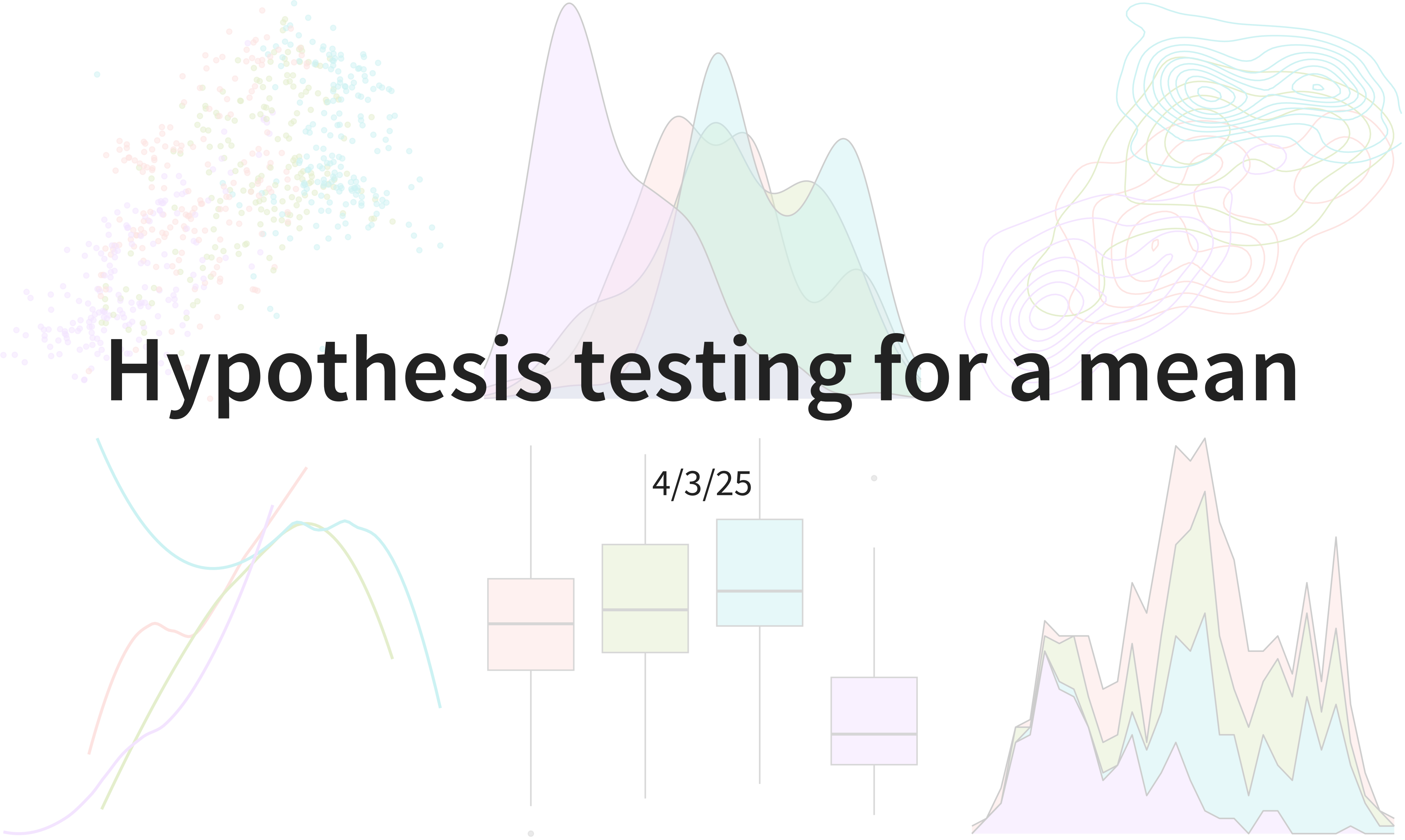


Hypothesis testing for a mean

4/3/25

Housekeeping

- DataFest tomorrow! See you there!

Recap

- We have seen how to perform hypothesis tests for questions involving the following:
    - A single proportion (coffee consumption)
    - Independence of two categorical variables (banker sex discrimination)
        - Think of as one population
    - Difference in two proportions (blood thinner)
        - Think of as two populations
- We are now going to see another hypothesis test, this time for *numerical* data

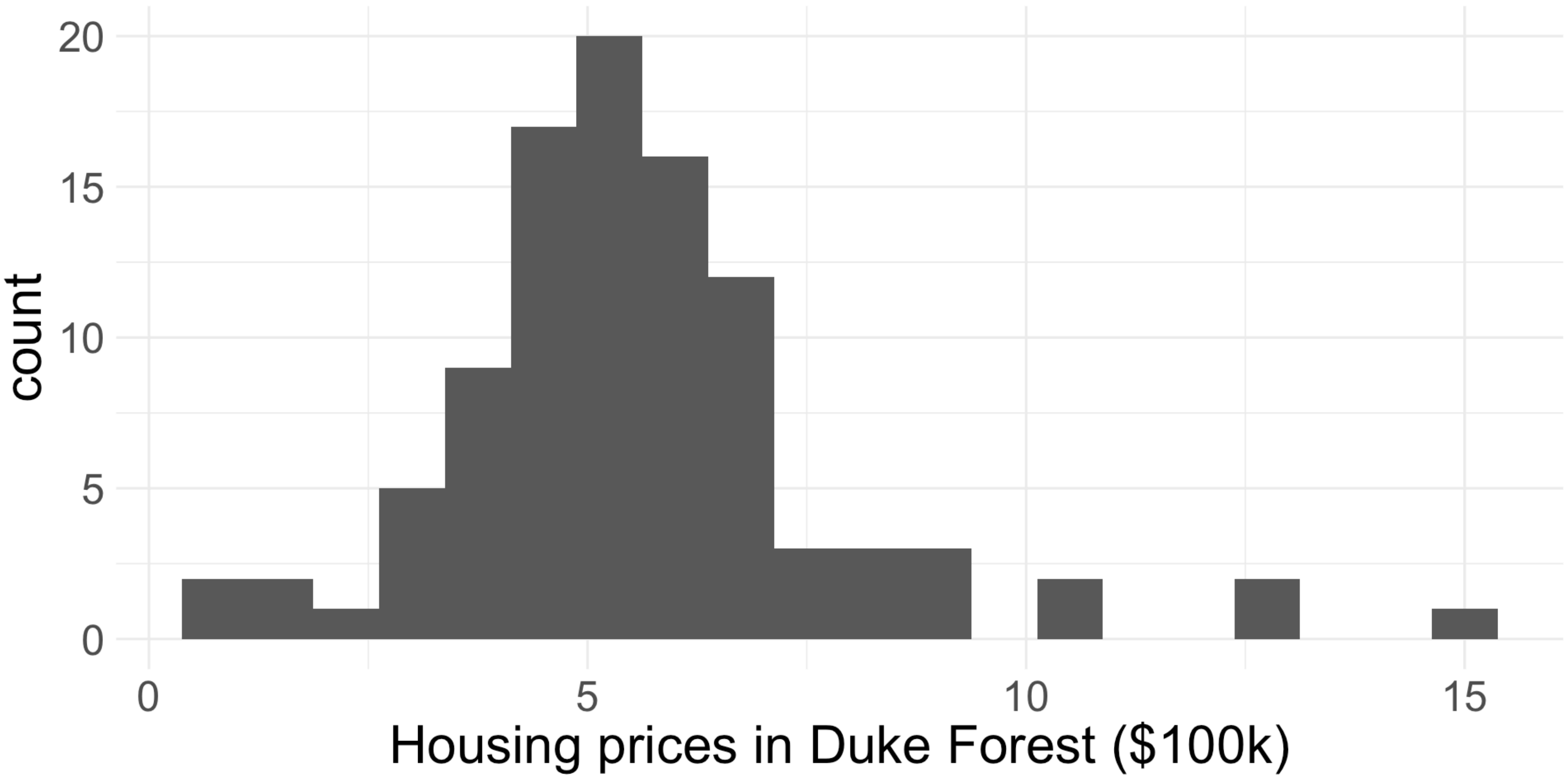
Test for a single mean

Running example + form hypotheses

We will use the `duke_forest` dataset provided in `openintro`. It provides data on some houses that were sold in the Duke Forest neighborhood of Durham, NC in November 2020.

- Before we look at the data, we should form our hypotheses. Suppose I am interested in learning if the average price of houses in Duke Forest is $500,000 or not.
- What might our hypotheses be?
  - $H_0 : \mu = 5$ versus $H_A : \mu \neq 5$, where μ is the true average house of prices in Duke Forest in $100,000
  - Terminology: I will refer to $\mu_0 = 5$ as my "null hypothesized value". (i.e. the specific value of μ in H_0)

Collect and summarise data



The observed/sample mean housing price is $\bar{x} = 5.599$ from a sample of 98 houses.

- Now we must determine if we have “convincing evidence”! Choose $\alpha = 0.05$

Simulating null distribution

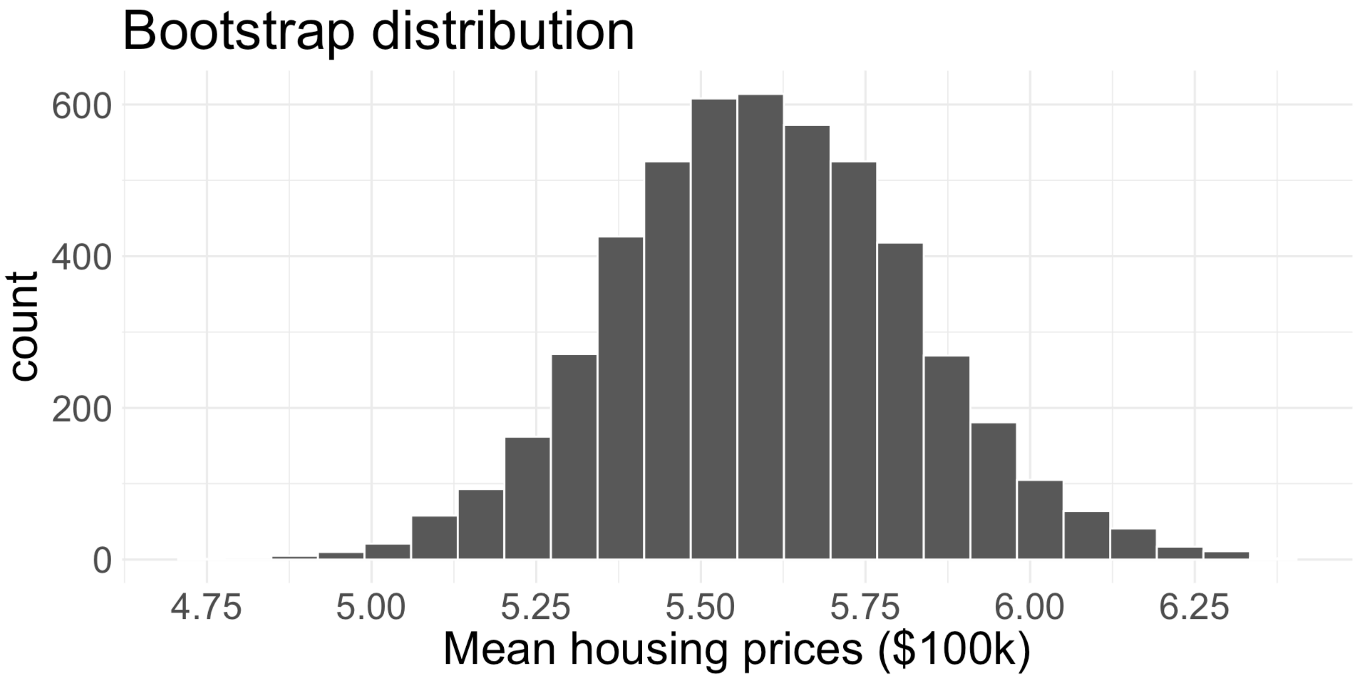
To simulate from the null distribution, we need to operate in a world where H_0 is true

- So, I need to repeatedly simulate data sets of size 98 where the true mean is 5, without changing anything else
- If I don’t want to make any assumptions about how the data behave, how might I do that?

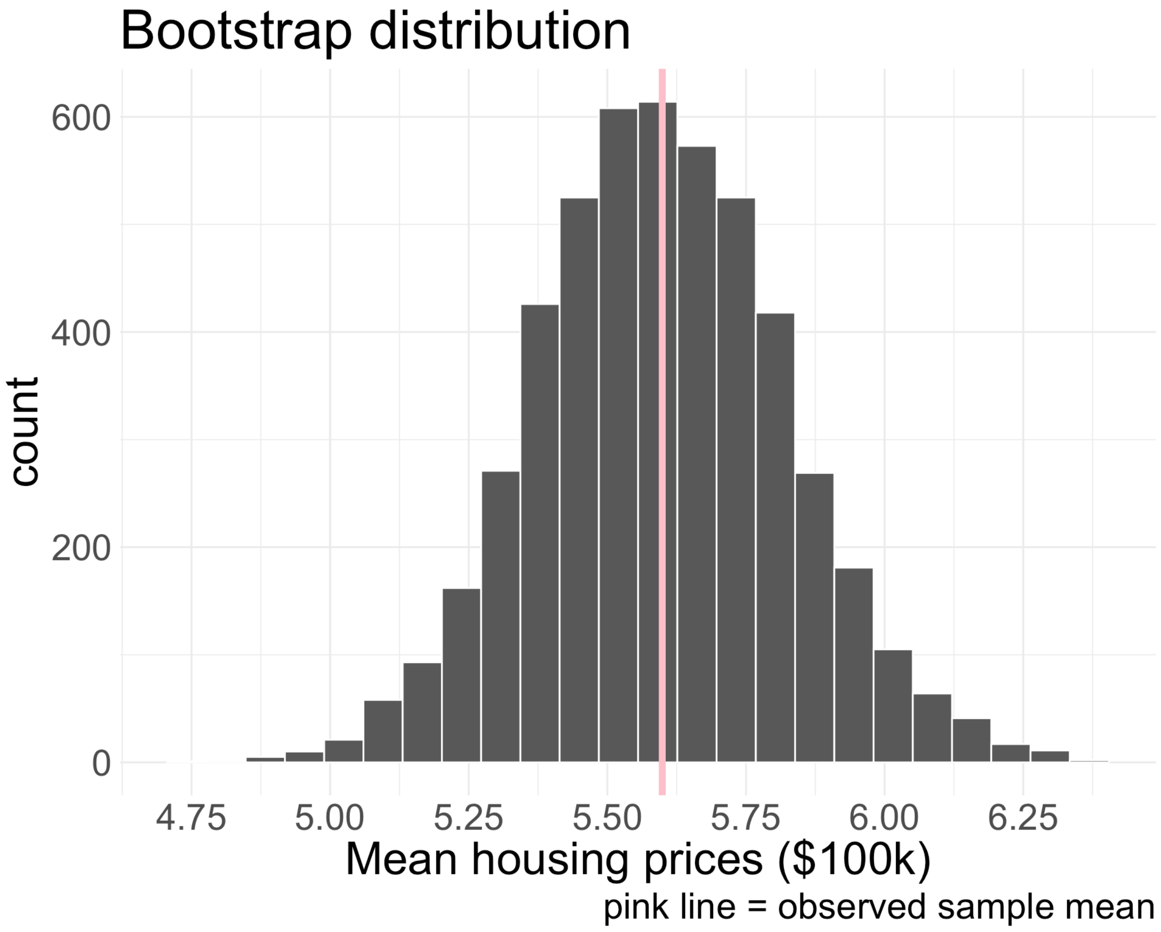
Bootstrap to the rescue

How would I obtain a bootstrap distribution of the sample mean price of houses?

Remind ourselves: Where should the bootstrap distribution be centered?



Bootstrap to null distribution



- This is **not** the null distribution! The null distribution should be centered at $\mu_0 = 5$
- However, the null distribution should have the same variability in $\bar{x}$ as the bootstrap distribution.

- So to get the null distribution, why not just **shift** the bootstrap distribution to be centered where we want it to be?

Shifting the bootstrap distribution

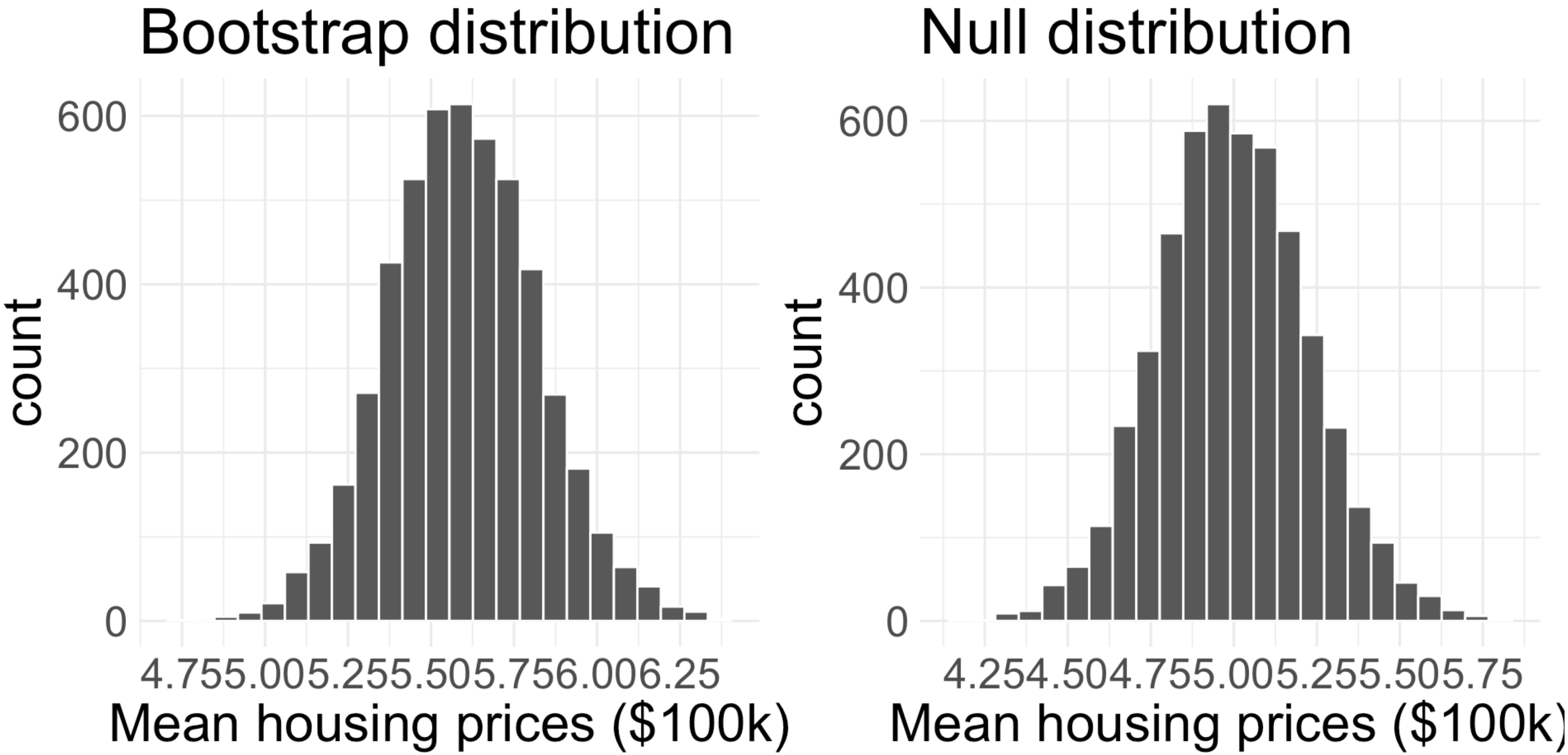
- In this example, bootstrap distribution is centered at $\bar{x} = 5.599$
- In order to center this distribution at $\mu_0 = 5$, just subtract $5.599 - 5 = 0.599$ from every single bootstrapped mean
  - This will give us a simulated distribution for $\bar{x}$ centered at $\mu_0 = 5$, which is exactly the null distribution!
  - We call this “shifting the bootstrap distribution”, because we simply shift where the bootstrap distribution is centered

```r
mu0 <- 5

# xbar holds observed sample mean
shift <- xbar - mu0

# boot_means is a vector holding B bootstrapped sample means
null_dist <- boot_means - shift
```

Null distribution



Bootstrap distribution

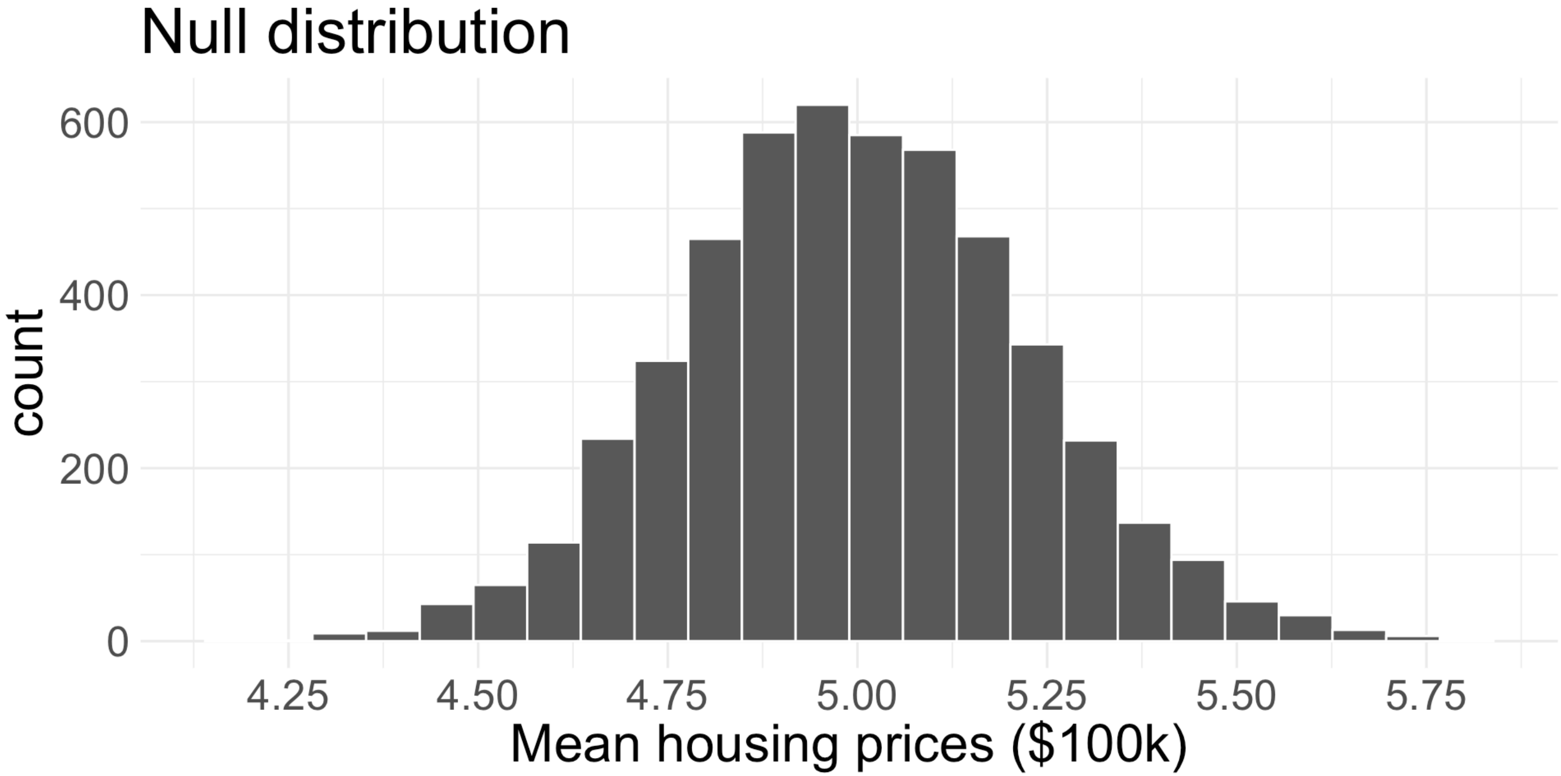
Null distribution

- Notice where the distributions are centered

Obtain the p-value

$H_0: \mu = 5$ versus $H_A: \mu \neq 5$

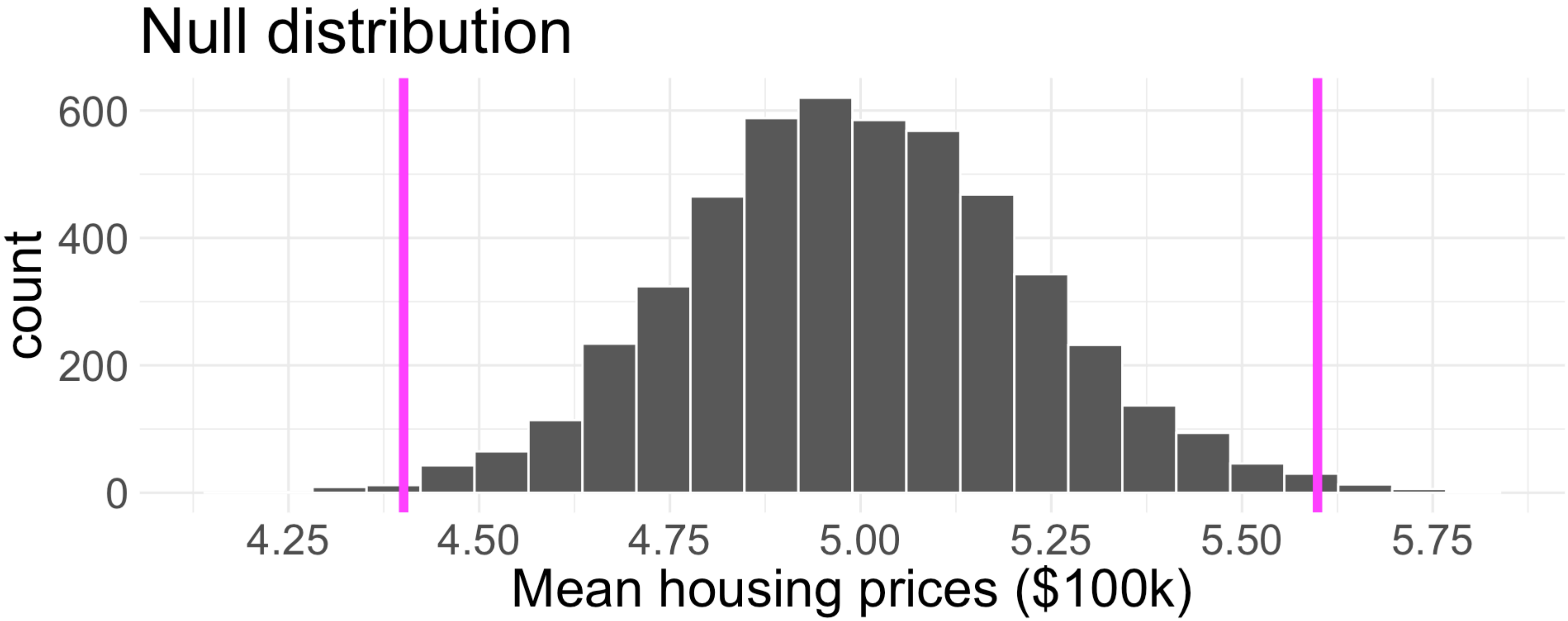
Our observed sample mean housing price is 5.599.



What does it mean to be “as or more extreme” now?

Two-sided alternative hypothesis

- This is the first time we've seen a *two-sided* hypothesis as a class
- Since the alternative is "double sided", we can be extreme in **both** the positive and negative direction!

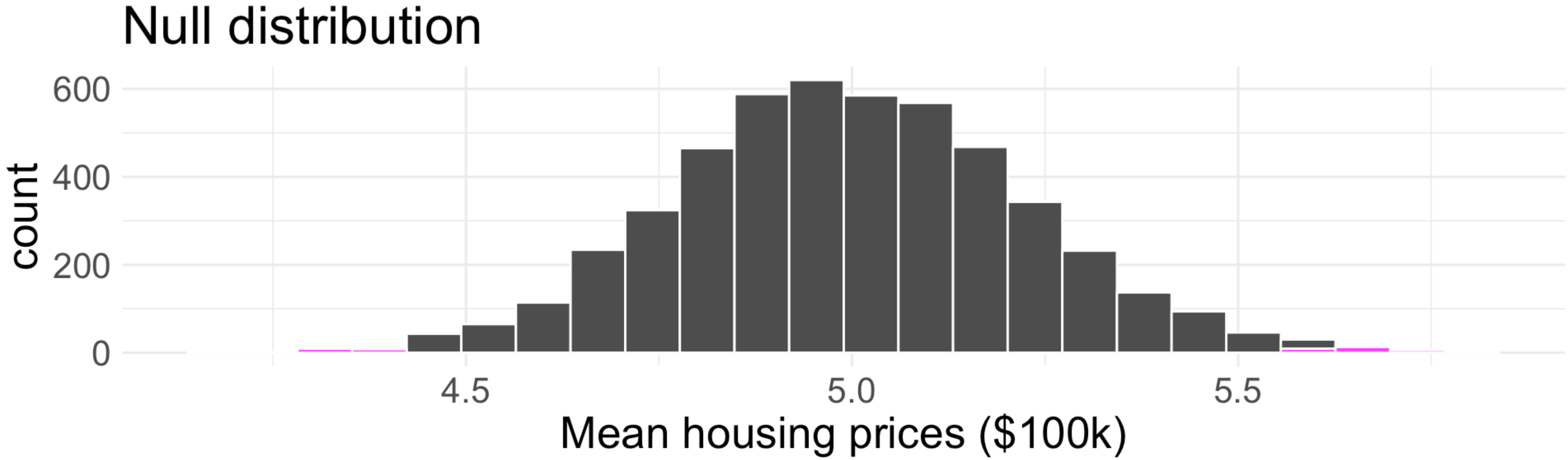


Obtain the p-value (cont.)

Let $shift$ represent the amount we shifted the distribution by:

$$shift = 5.599 - 5 = 0.599$$

Simulated means as or more extreme than $\mu_0 + shift$ or $\mu_0 - shift$ will contribute:



```r
sum( (null_dist >= mu0 + shift) | (null_dist <= mu0 - shift))/B
```

```
[1] 0.0098
```

Make decision and conclusion

Make a decision and conclusion in the context of the research question.

- Since our p-value of 0.0098 is less than the significance level of 0.05, we reject H_0. We have convincing evidence to suggest that the true average housing price of homes in Duke Forest in 2020 was not $500k.

Comprehension questions

- Why did we shift the bootstrap distribution?
- How do we estimate the p-value in the case of a two-sided alternative hypothesis?