Natural Language Processing

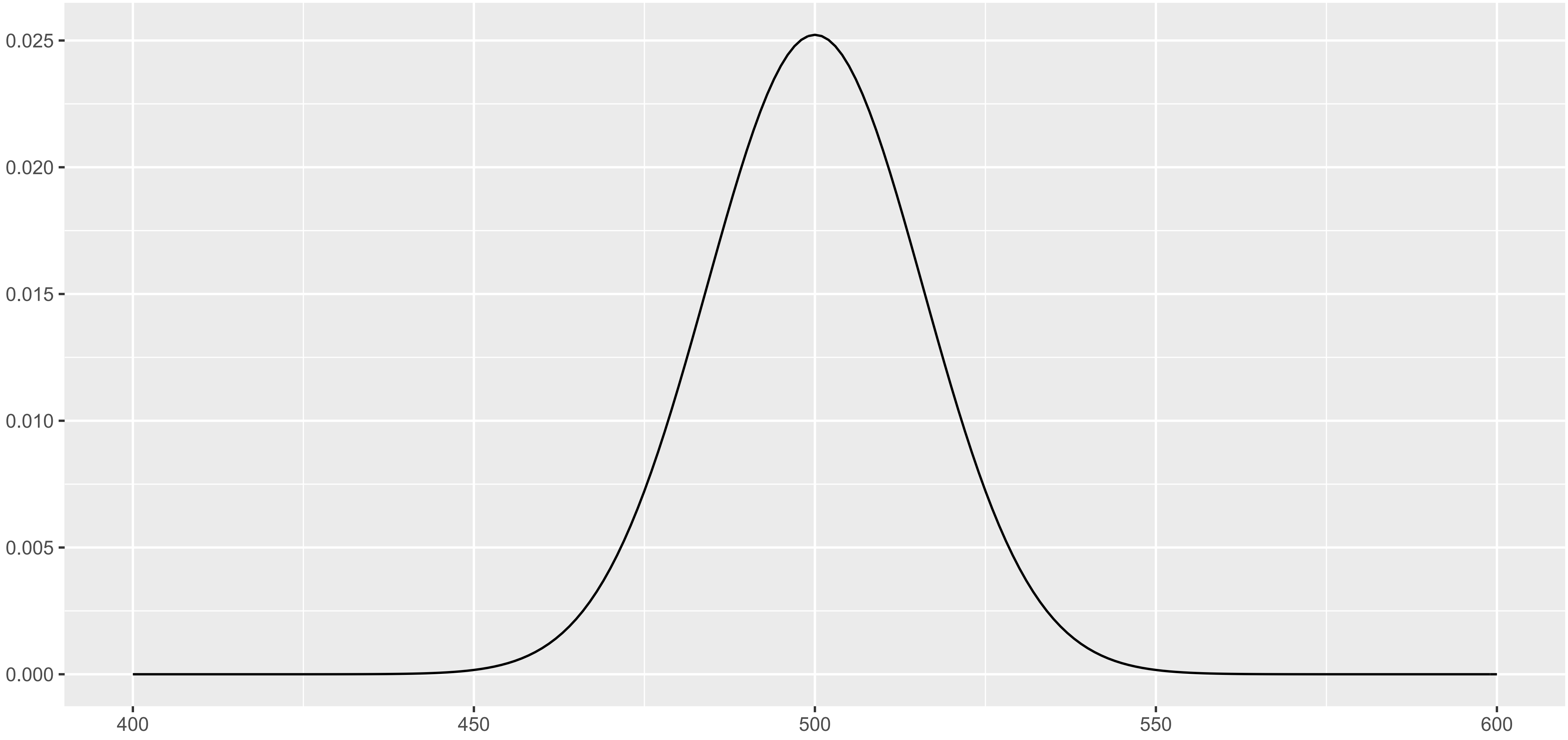
Mehmet Can Yavuz, PhD

Adapted from Info 256 - David Bamman, UC Berkeley

Hypothesis tests

- At what point is the sample statistic so unusual that we can reject the null hypothesis as being too unlikely to have generated it?

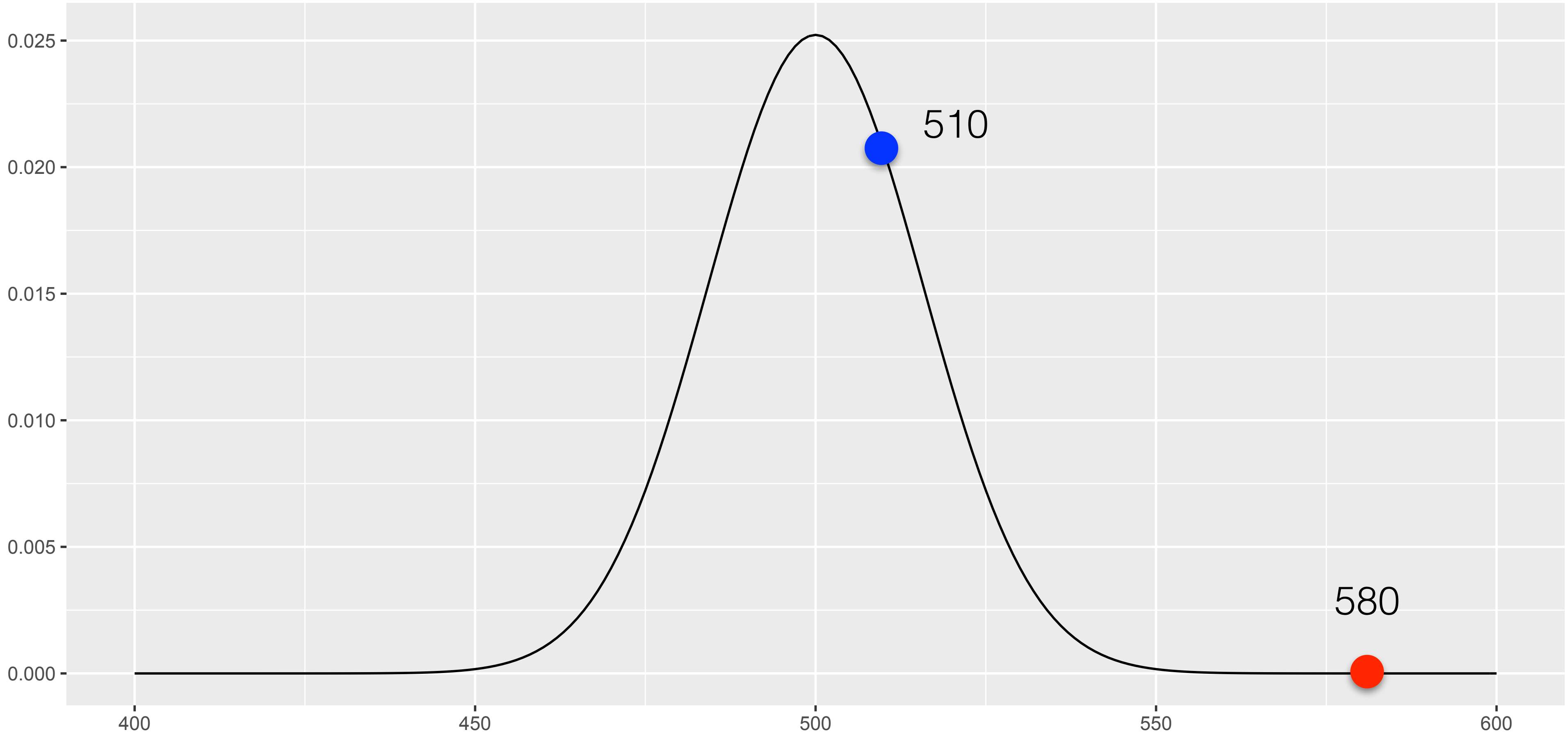
Example



Binomial probability distribution for number of correct predictions in n=1000 with p = 0.5

Example

At what point is a sample statistic unusual enough to reject the null hypothesis?



Hypothesis tests

- How do we define what "too unusual" means?
- Parametric tests state that the null hypothesis follows a probability distribution with a fixed set of parameters:
  - Binomial (parameterized by the success rate p and number of trials n)
  - Normal (parametrized by mean μ and standard deviation σ)

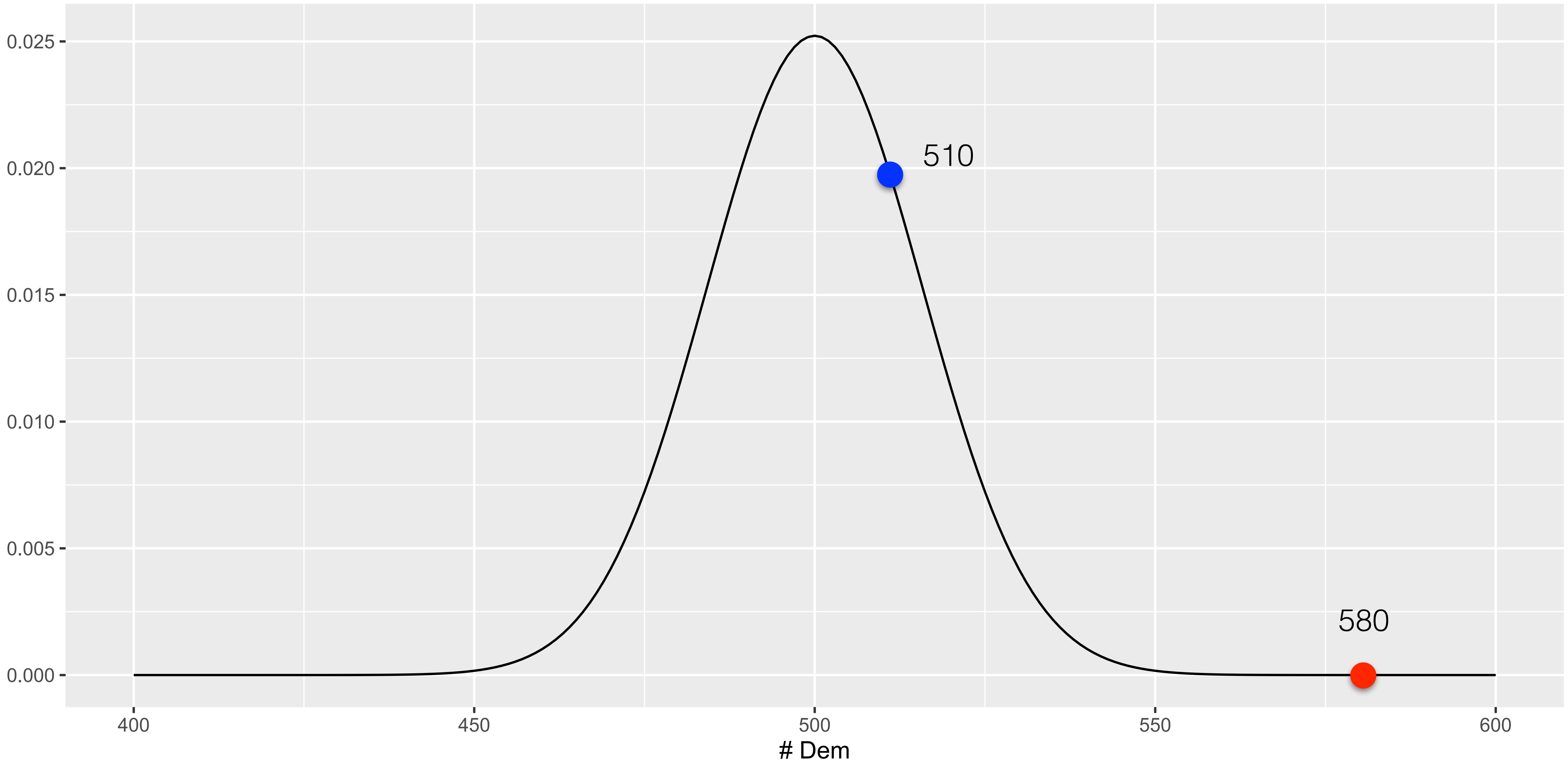
Hypothesis tests

- How do we define what “too unusual” means?

- Parametric tests state that the null hypothesis follows a probability distribution with a fixed set of parameters

- In these tests, we can calculate the probability of the statistic by just looking it up

  - e.g., $P(x=580 \mid p=0.50, n=1000)$ in Binomial distribution.


0.025
0.020
0.015
0.010
0.005
0.000
400
450
500
550
600
# Dem
510
580

Parametric tests

- Parametric tests often rely on a normal approximation for large sample sizes, using the central limit theorem (CLT)

- CLT: the average of independent random variables tends toward a normal distribution, even if the original variables themselves are not normally distributed.

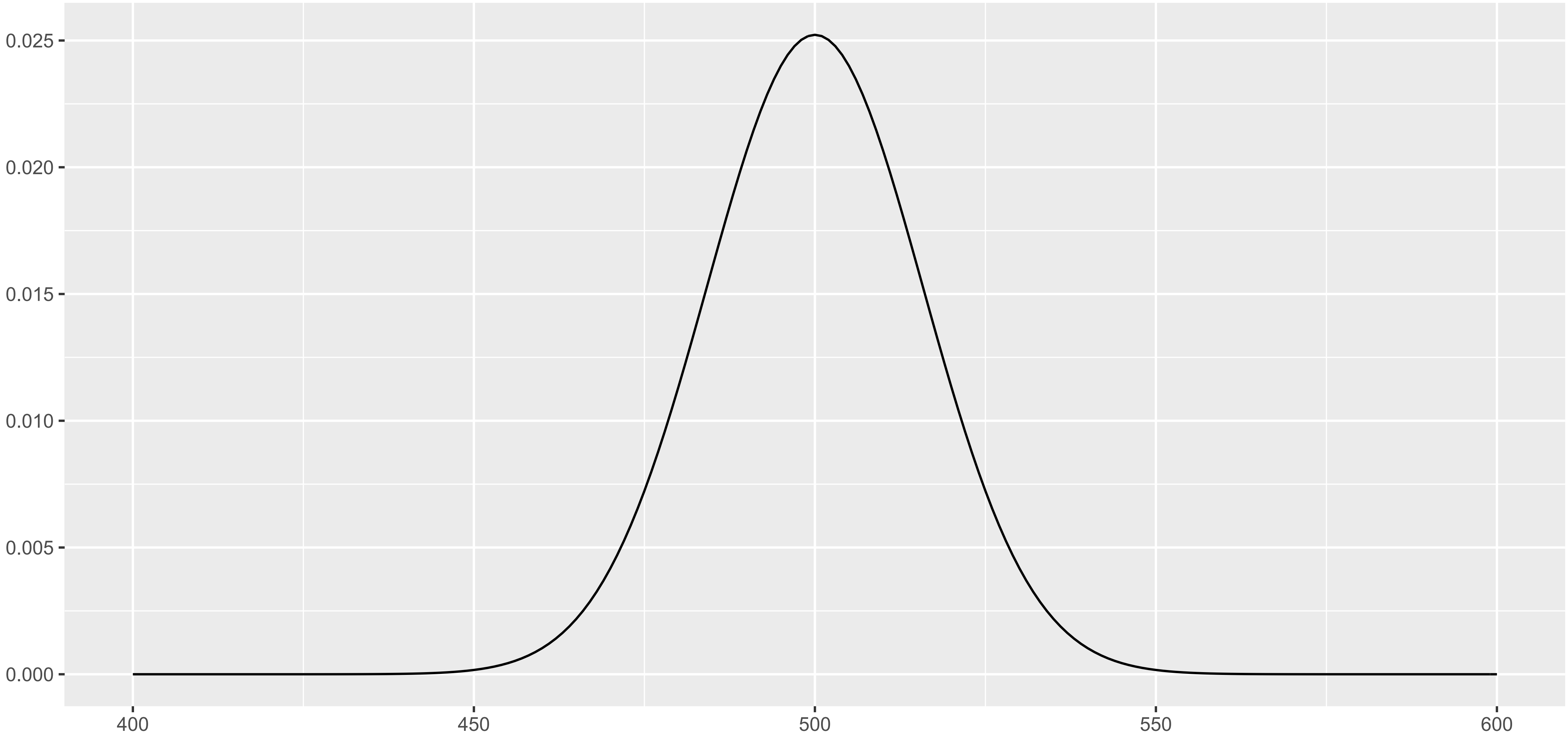
Metrics

Metric	Simple averaging?
Accuracy	✔
Precision	✔
Recall	✔
F1	✖

Nonparametric tests

- The big question: if we can’t make a parametric assumption (e.g., that accuracy follows a normal distribution), how can we say how like a given test statistic is?

- How do we construct a null distribution?


0.025
0.020
0.015
0.010
0.005
0.000
400
450
500
550
600

Nonparametric tests

- Many hypothesis tests rely on parametric assumptions (e.g., normality)
- Alternatives that don’t rely on those assumptions:
    - permutation test
    - the bootstrap

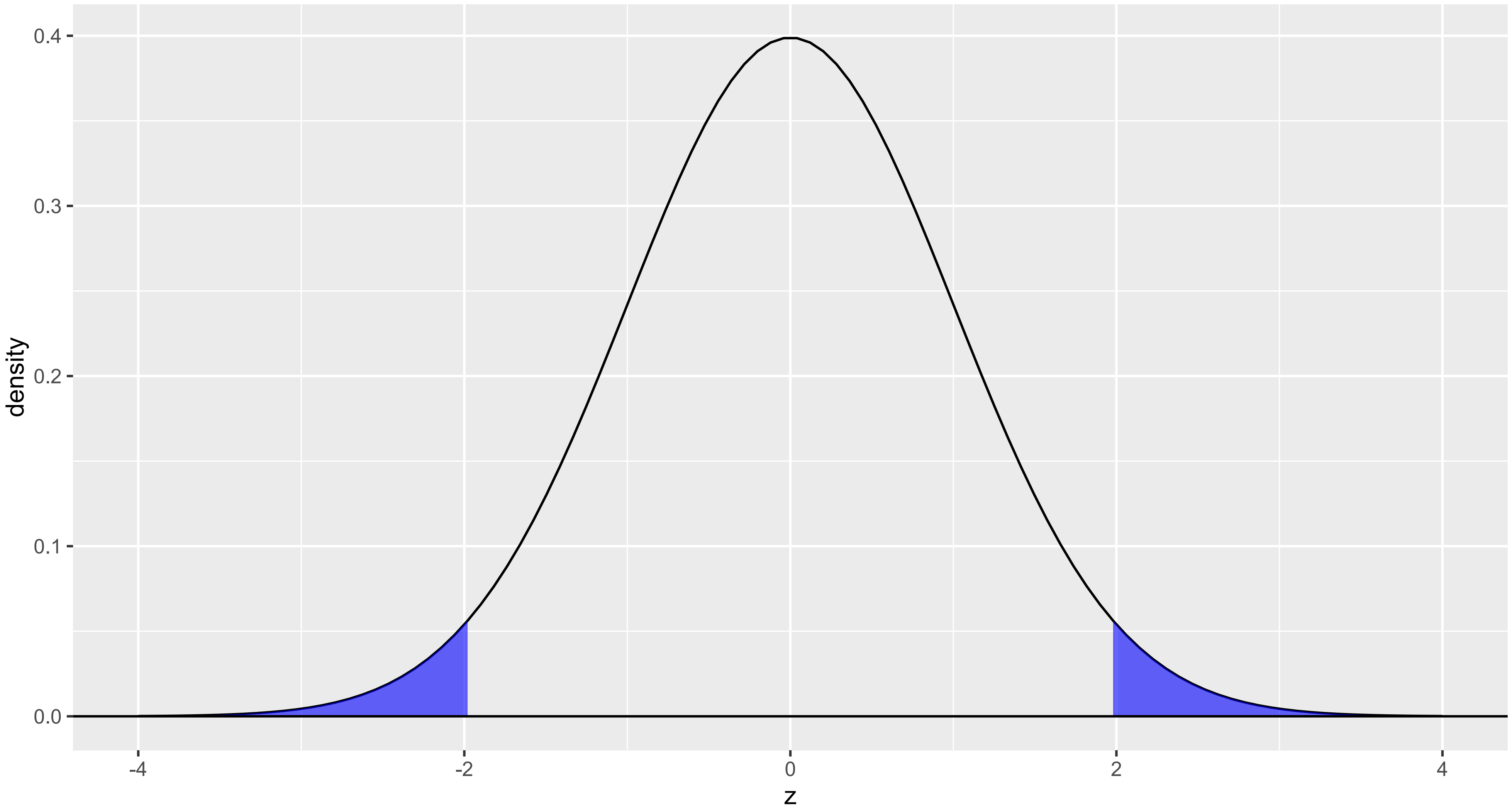
Back to logistic regression

β	change in odds	feature name
2.17	8.76	Eddie Murphy
1.98	7.24	Tom Cruise
1.70	5.47	Tyler Perry
1.70	5.47	Michael Douglas
1.66	5.26	Robert Redford
…	…	…
-0.94	0.39	Kevin Conway
-1.00	0.37	Fisher Stevens
-1.05	0.35	B-movie
-1.14	0.32	Black-and-white
-1.23	0.29	Indie

Significance of coefficients

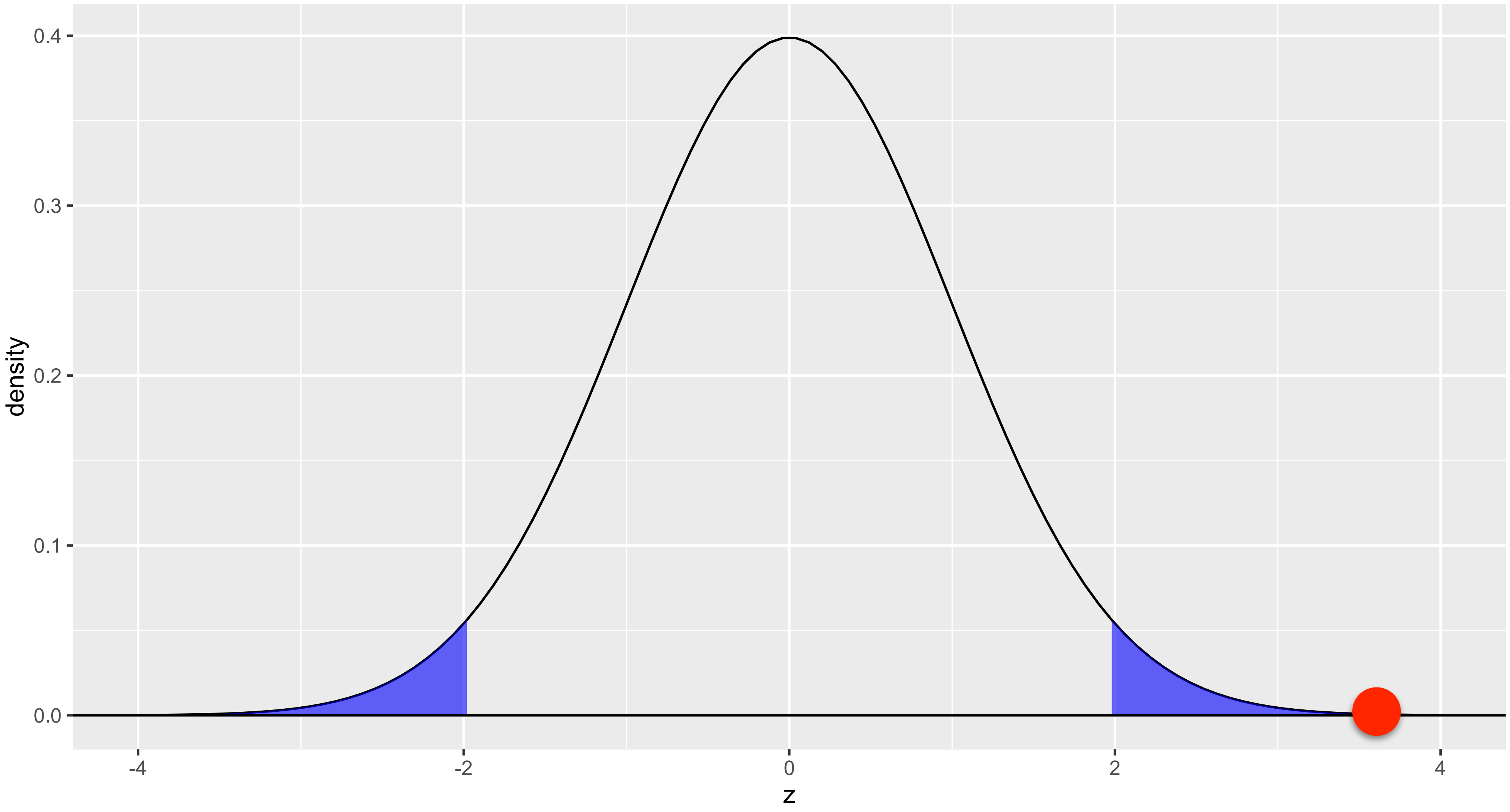
- A β_i value of 0 means that feature x_i has no effect on the prediction of y
- How great does a β_i value have to be for us to say that its effect probably doesn't arise by chance?
- People often use parametric tests (coefficients are drawn from a normal distribution) to assess this for logistic regression, but we can use it to illustrate another more robust test.

Hypothesis tests



Hypothesis tests measure how (un)likely an observed statistic is under the null hypothesis

Hypothesis tests



Permutation test

- Non-parametric way of creating a null distribution (parametric = normal etc.) for testing the difference in two populations A and B

- For example, the median height of men (=A) and women (=B)

- We shuffle the labels of the data under the null assumption that the labels don’t matter (the null is that A = B)

Permutation test

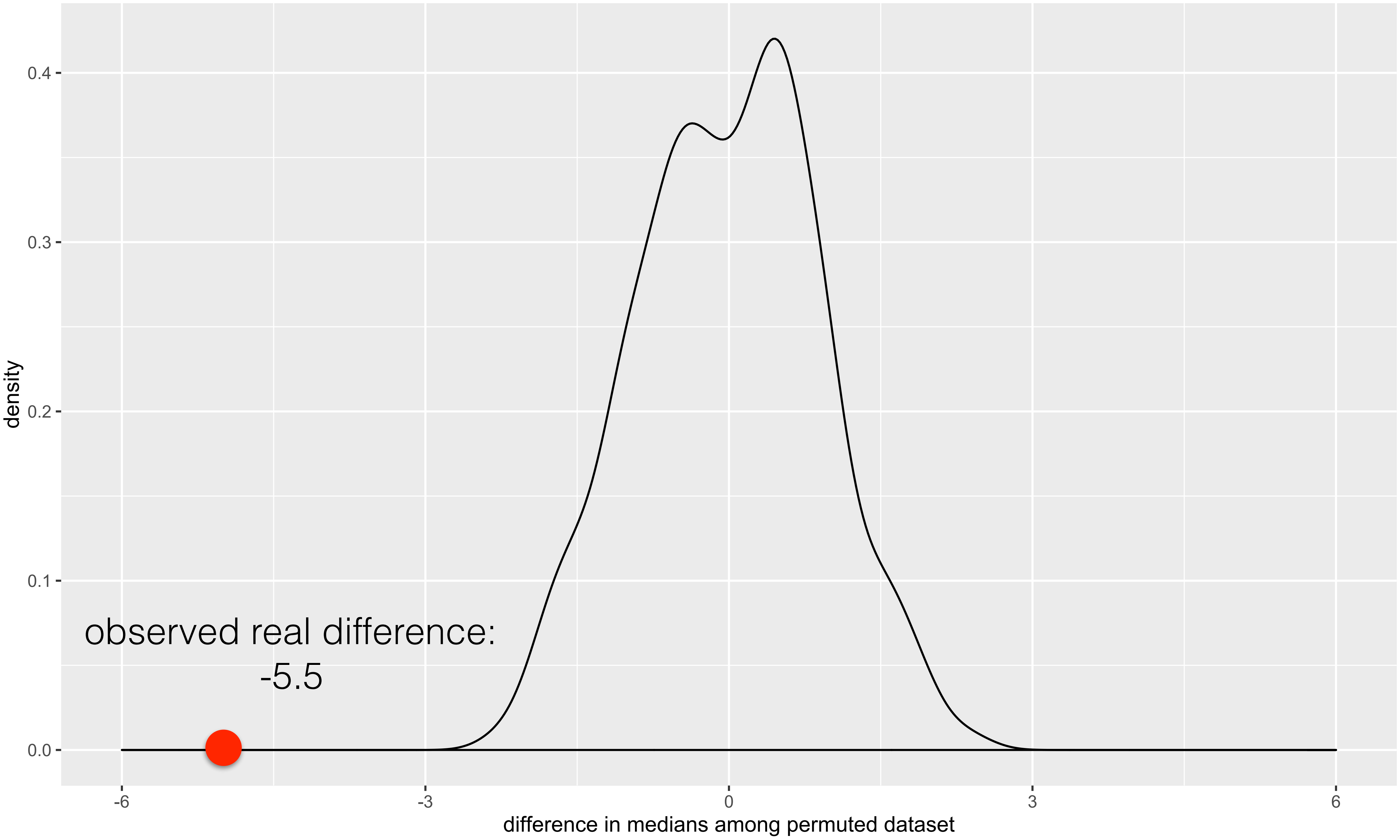
- Core idea: if the null hypothesis were true and there's no difference between groups, then it doesn't matter which label each data point has.

		true labels	perm 1	perm 2	perm 3	perm 4	perm 5
x1	62.8	woman	man	man	woman	man	man
x2	66.2	woman	man	man	man	woman	woman
x3	65.1	woman	man	man	woman	man	man
x4	68.0	woman	man	woman	man	woman	woman
x5	61.0	woman	woman	man	man	man	man
x6	73.1	man	woman	woman	man	woman	woman
x7	67.0	man	man	woman	man	woman	man
x8	71.2	man	woman	woman	woman	man	man
x9	68.4	man	woman	man	woman	man	woman
x10	70.9	man	woman	woman	woman	woman	woman

observed true difference in medians: -5.5

		true	perm 1	perm 2	perm 3	perm 4	perm 5
x1	62.8	woman	man	man	woman	man	man
x2	66.2	woman	man	man	man	woman	woman
…	…	…	…	…	…	…	…
x9	68.4	man	woman	man	woman	man	woman
x10	70.9	man	woman	woman	woman	woman	woman
difference in medians:			-0.8	0.3	1.4	1.2	-2.0

how many times is the difference in medians between the permuted groups greater than the observed difference?



observed real difference:
-5.5

A=100 samples from Norm(70,4)

B=100 samples from Norm(65, 3.5)

Permutation test

The p-value is the number of times the permuted test statistic t_p is more extreme than the observed test statistic t:

$$\hat{p} = \frac{1}{B} \sum_{i=1}^{B} I[abs(t) < abs(t_p)]$$

p values

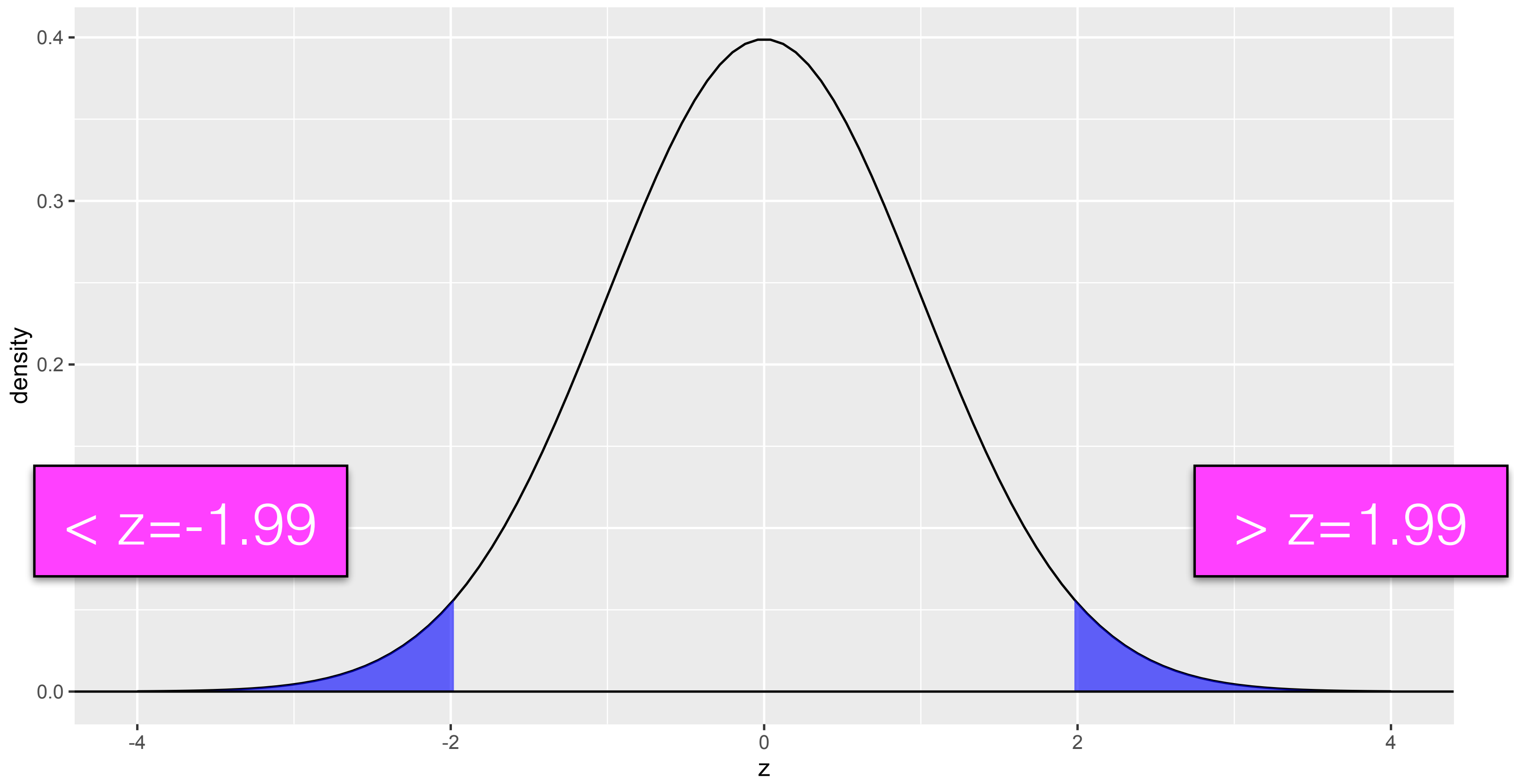
A p value is the probability of observing a statistic at least as extreme as the one we did if the null hypothesis were true.

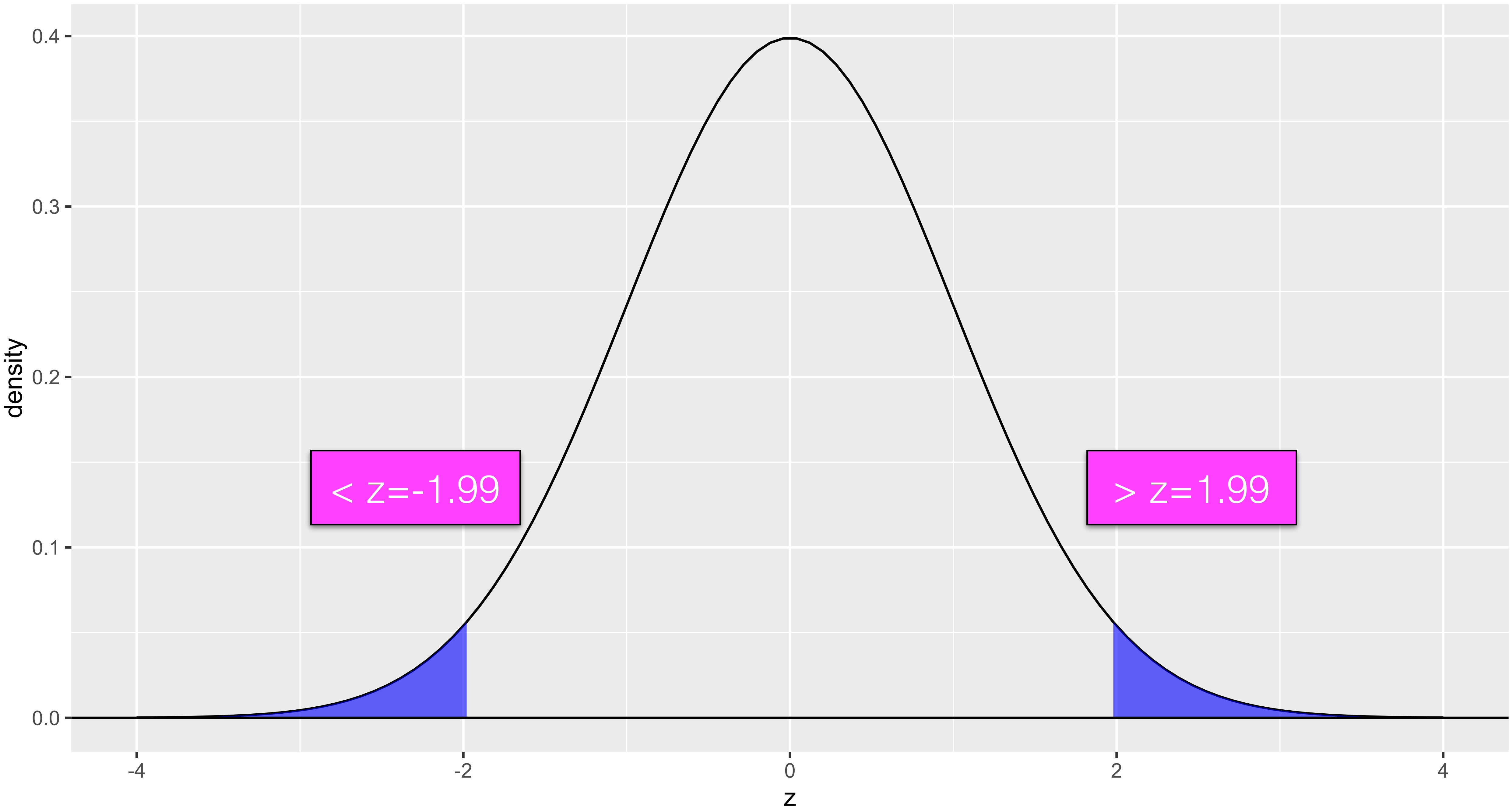
- Two-tailed test $\text{p-value}(z) = 2 \times P(Z \leq -|z|)$
- Lower-tailed test $\text{p-value}(z) = P(Z \leq z)$
- Upper-tailed test $\text{p-value}(z) = 1 - P(Z \leq z)$

P-value

- If our test statistic is 1.99, then the two-tailed p-value is the sum of the shaded probability mass in the extremes.

- In parametric tests, we can calculate this using the CDF P(X < x) of the null distribution.



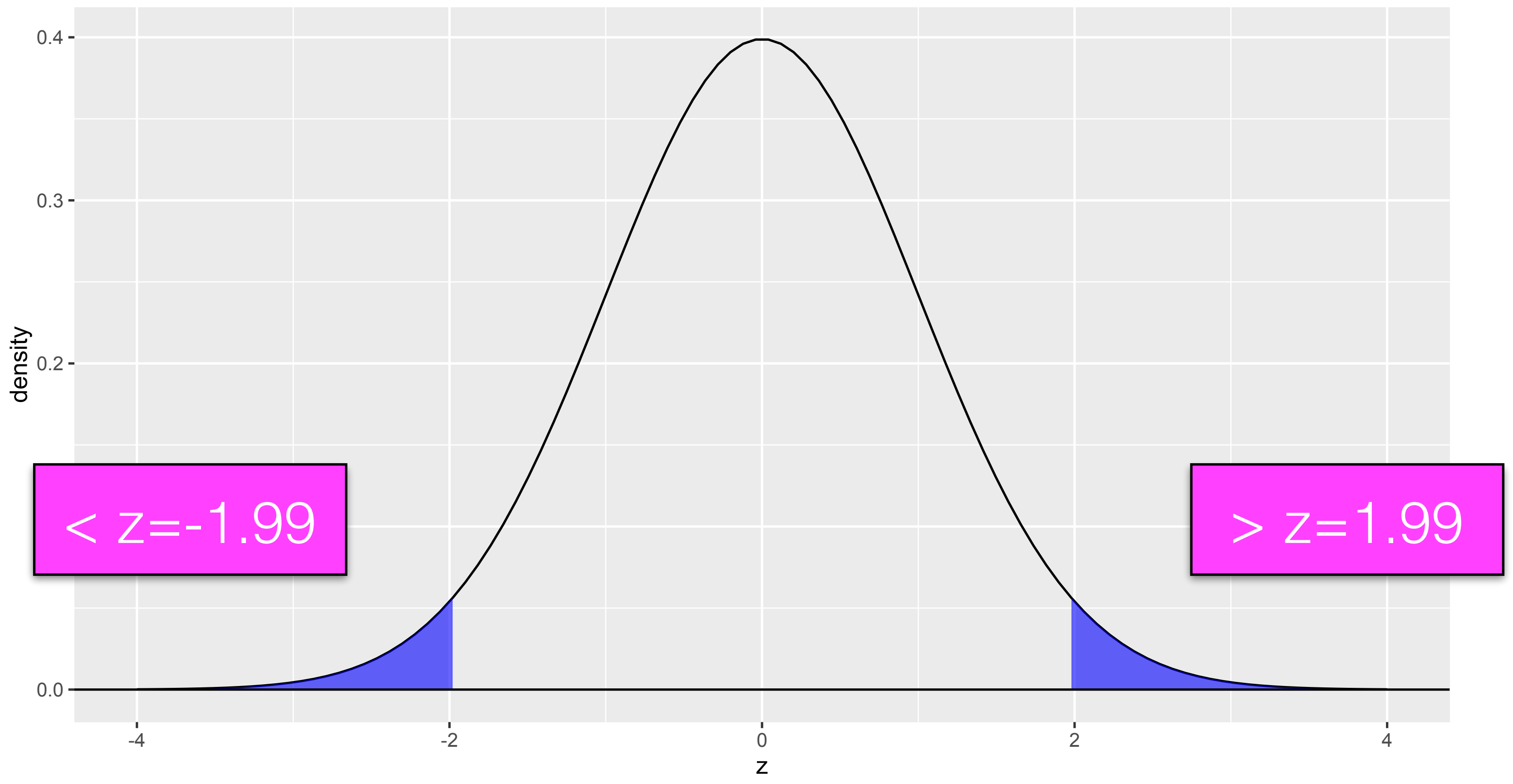


density
0.4
0.3
0.2
0.1
0.0
< z=-1.99
> z=1.99
-4
-2
0
2
4
z

P-value

- If our test statistic is 1.99, then the two-tailed p-value is the sum of the shaded probability mass in the extremes.

- In non-parametric tests, we can calculate this by just counting how many times we observe a permuted statistic fall in these extremes.



Permutation test

- The permutation test is a robust test that can be used for many different kinds of test statistics, including coefficients in logistic regression.

- How?

    - A = members of class 1
    - B = members of class 0
    - β are calculated as the (e.g.) the values that maximize the conditional probability of the class labels we observe; its value is determined by the data points that belong to A or B

Permutation test

- To test whether the coefficients have a statistically significant effect (i.e., they're not 0), we can conduct a permutation test where, for B trials, we:

  1. shuffle the class labels in the training data

  2. train logistic regression on the new permuted dataset

  3. tally whether the absolute value of β learned on permuted data is greater than the absolute value of β learned on the true data

Permutation test

The p-value is the number of times the permuted β_p is more extreme than the observed β_t:

$$\hat{p} = \frac{1}{B}\sum_{i=1}^{B} I[abs(\beta_t) < abs(\beta_p)]$$

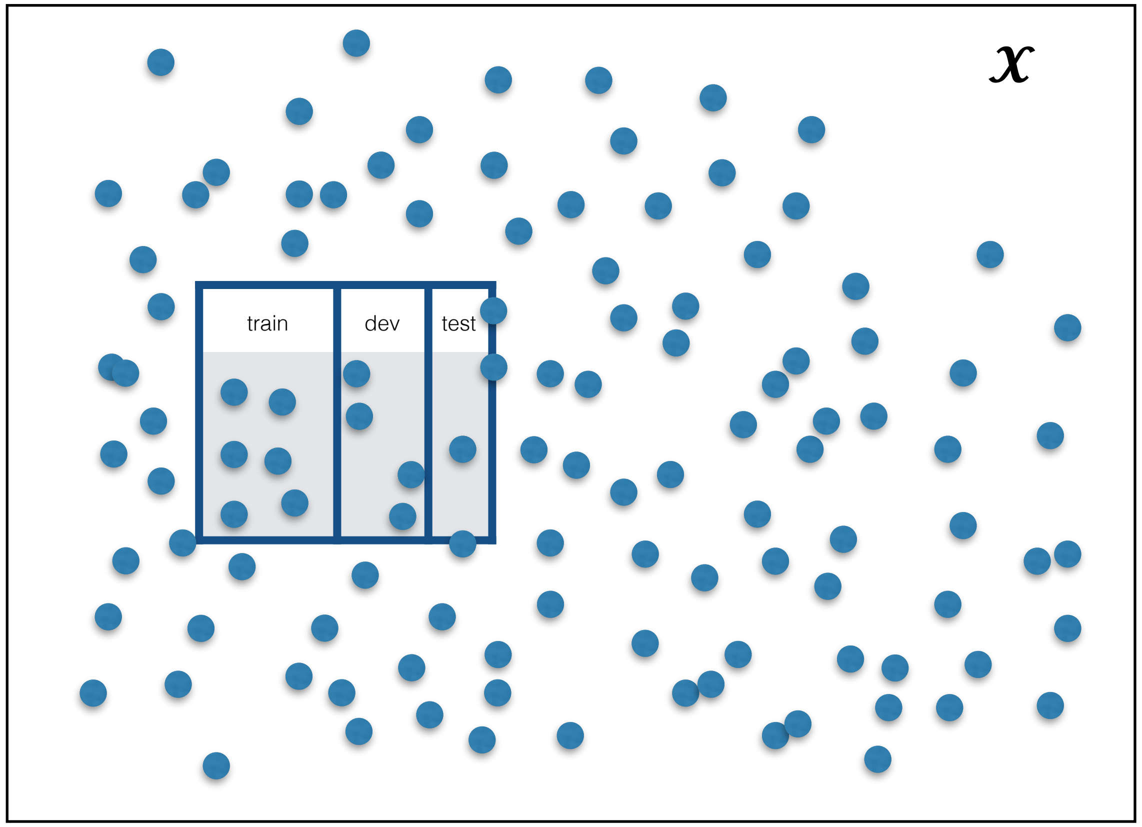
Bootstrap

- The permutation test assesses significance conditioned on the test data you have (we rearrange the labels to form the null distribution, but the data itself doesn't change).

- To also model the variability in the data we have, we can use the statistical bootstrap (Efron 1979).

Bootstrap

- Core idea: the data we happen to have is a sample from all data that could exist; let's sample from our sample to estimate the variability.



Bootstrap

- Core idea: the data we happen to have is a sample from all data that could exist; let’s sample from our sample to estimate the variability.
- Our estimate of the point value of the metric itself won’t change, but we can infer something about the variability of the population from the variable in the resamples.

Bootstrap

- Start with test data x of size n
- Draw b bootstrap samples x(i) of size n by sampling with replacement from x
- For each x(i)
    - Let m(i) = the metric of interest calculated from x(i)

Bootstrap

- At the end of the process, you end up with a vector of values $m = [m(1), \ldots, m(b)]$ (for b bootstrap samples)

m	Utility
Classification accuracy for System A	Estimate confidence intervals
I[System A > System B]	Calculate significance of difference

Bootstrap percentile interval

- m = [m(1), …, m(b)]
- We can define a 95% confidence interval as the middle 95% of m
- e.g., α = 0.05 (95% confidence intervals) = [2.5, 97.5] percentile
- Accurate for larger sample sizes