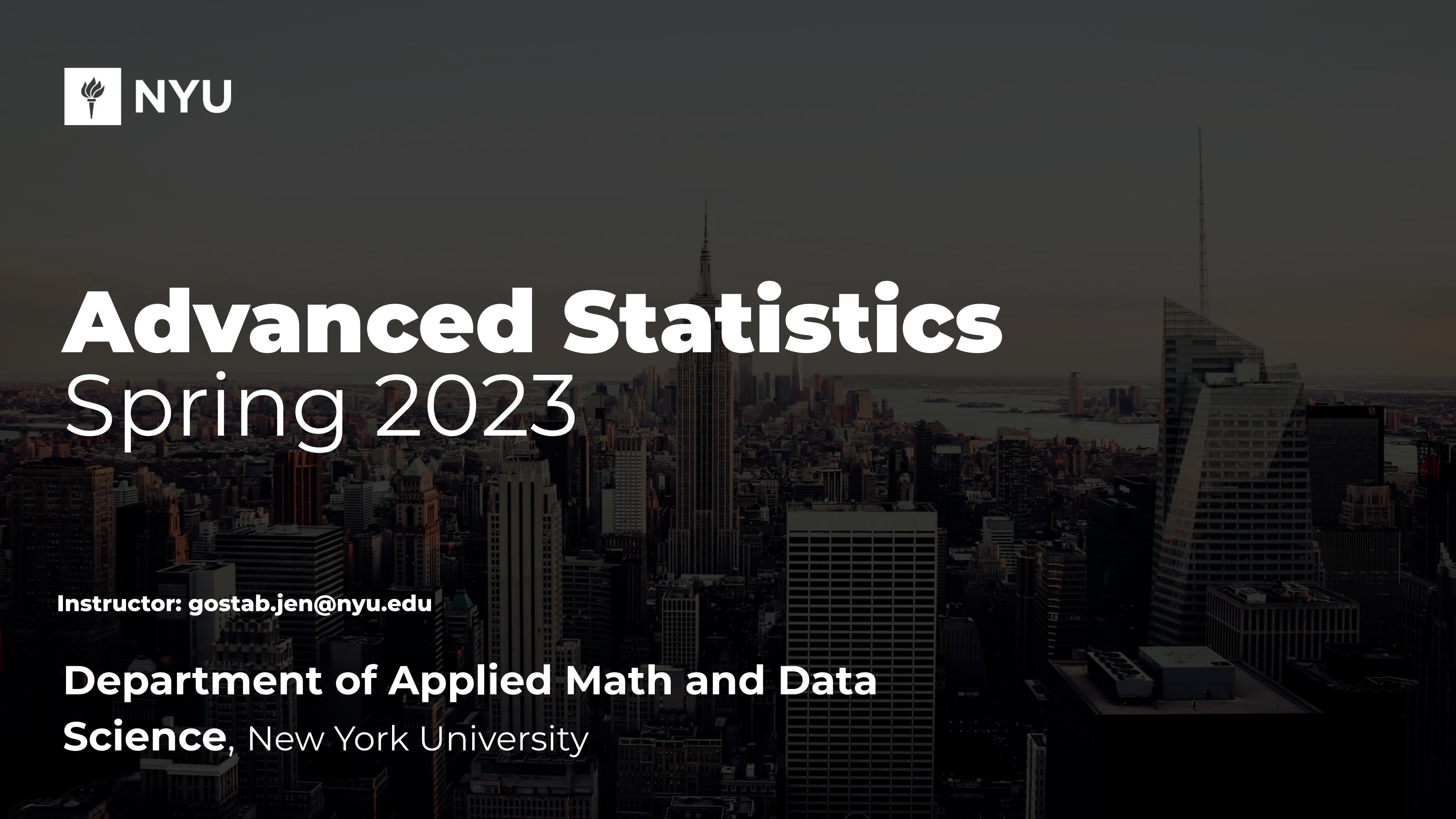


Advanced Statistics

Spring 2023

Instructor: gostab.jen@nyu.edu

Department of Applied Math and Data Science, New York University

Data Assignment 2

- Grade released(with feedback)
- Section Average: 87%, B+.
- Points off: 1) interpretation of your analysis
- 2) Flip the variables

Lab Assignment

Don’t forget to submit your lab activity to brightspace and ensure your full credit

Week 9-10 Plans

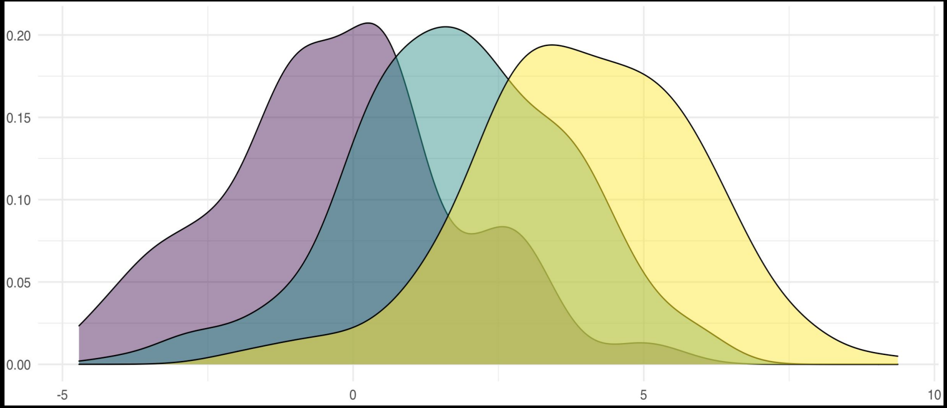
Review:

Recap Logistic Regression with Exam 2

Hypothesis Testing

Intro to Statistical Inference

(No time for practice!)



Week 9-10 Plans

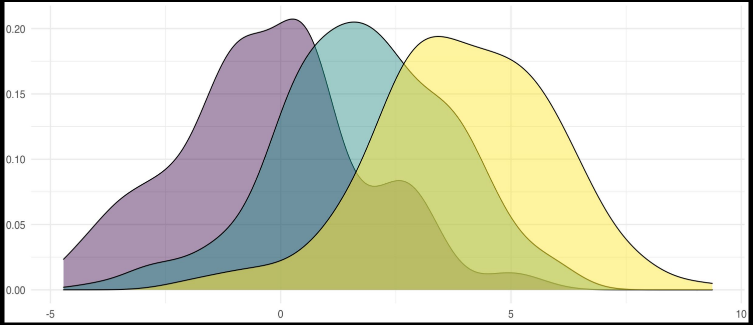
Review:

→Recap Logistic Regression with Exam 2

Hypothesis Testing

Intro to Statistical Inference

(No time for practice!)



NYU

Logistic Regression

Recall some basic concepts of Logistic Regression...
We want to estimate the probability of getting the treatment(or response to the survey) **but we got two problems:**

Probability is not always linear so linear regression will be invalid.

Model a quantity so that if solving the linear regression we can get the probability

➔Survey response indicator variable **(1 for respondent; 0 for non-rep)**

Imagine selected students are required to complete a survey:

If given the data for which student responded to the survey

Think about what estimator we have for this variable *R*?

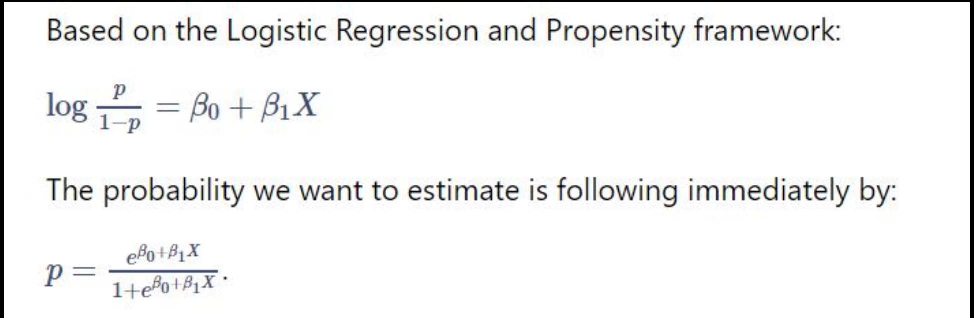
Can we get E(R)? Think about it. -Yes =0.6

Can we get P(R=1) or P(R=1 | X) This is a harder question.

Logistic Regression

If we have a dataset as follows:

	R	P(R=1|X)	*x*
1	1	?	12
2	0	?	17
3	0	?	40
4	1	?	25
5	1	?	29



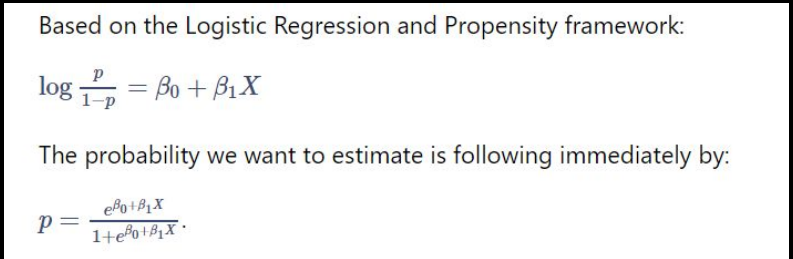
Based on the Logistic Regression and Propensity framework:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 X$$

The probability we want to estimate is following immediately by:

$$p = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$

Logistic Regression



Based on the Logistic Regression and Propensity framework:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 X$$

The probability we want to estimate is following immediately by:

$$p = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$

Now for this question:

In logistic regression, probabilities are converted to log odds

A. So that the Y axis is limited to a range from 0 to 1

B. To allow for residuals to be computed

C. To set the Y intercept to zero (0)

D. So that the Y axis can range from positive to negative infinity

From Logit to Logistic

The question we discussed in the last page indicates:

The linear form serves as a intermediary between the hypothesis space of functions and the probability we would like to estimate.(but we just choose logit function for convenience)

Since the RH formula follows the linear model, it must be visiting any possible values in R.

Week 9-10 Plans

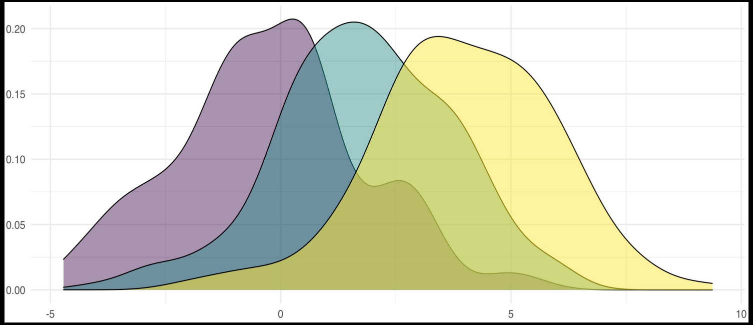
Review:

Recap Logistic Regression with sample questions

→Hypothesis Testing

Intro to Statistical Inference

(No time for practice!)



NYU

Hypothesis Testing

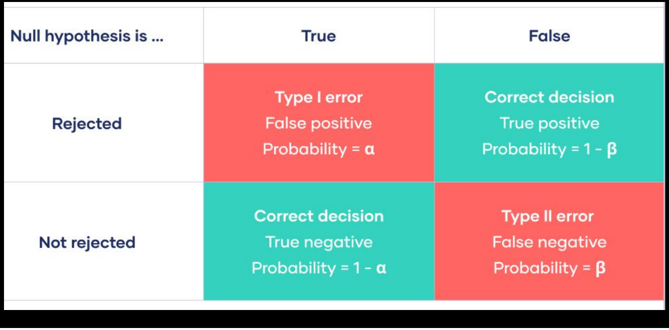
In this section I will walk you through all concepts of hypothesis testing and test of significance. This section is not designated for just doing manual calculations like statistics or critical value comparison **but for being a “deep thinker” of hypothesis testing:**

1) Understanding **Test Statistics and its null/alternative distributions**

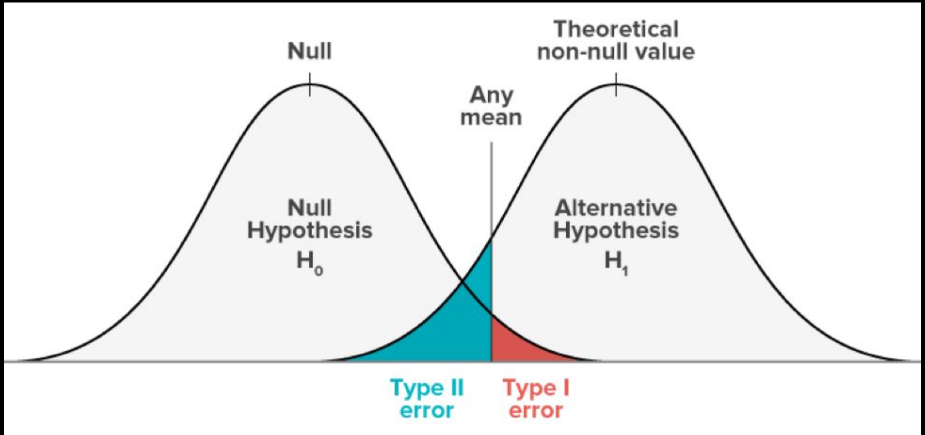
2) Build up a **strong link** to **conditional probability**

3) Have a **thorough examination of** **p value**.

Null hypothesis testing: Type I and II errors



Null hypothesis is ...	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = 1 - β
Not rejected	Correct decision True negative Probability = 1 - α	Type II error False negative Probability = β



Type I error: false positive

- You do find a significant result, but you shouldn't have (H_0 is true)

Type II error: miss

- You don't find a significant result, but you should have (H_0 is not true)

Which do you think is worse?

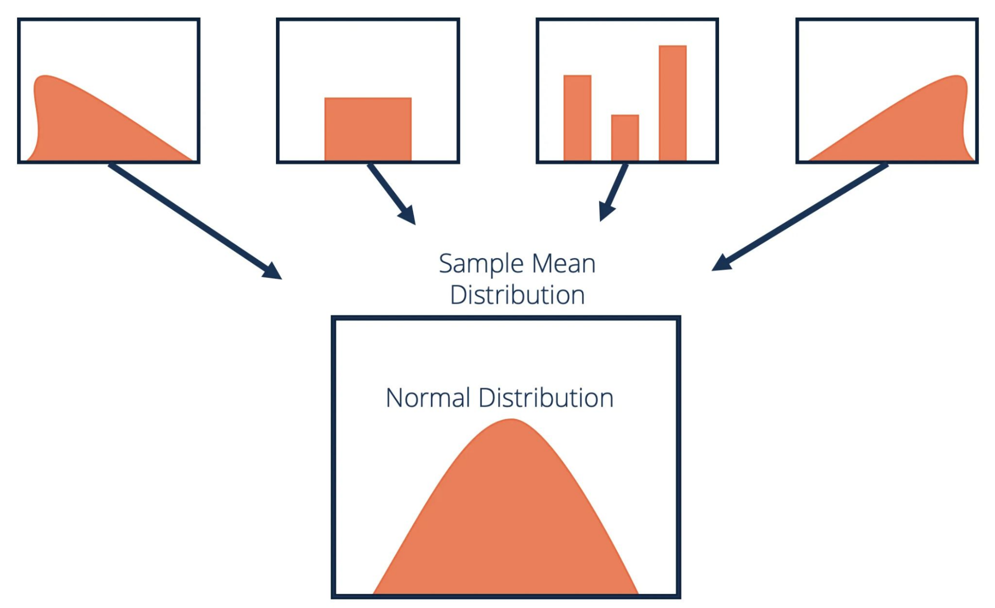
- Typically we think of Type I errors as worse (so we want to minimize Type I errors)

Central limit theorem

If you keep taking a bunch of samples (and take the mean of each sample), the **sample means will form a normal distribution, no matter what the original population distribution looks like**

No matter what the original distribution is, the sample means will form a normal distribution

Bigger sample size → less error, closer to true population mean



NYU