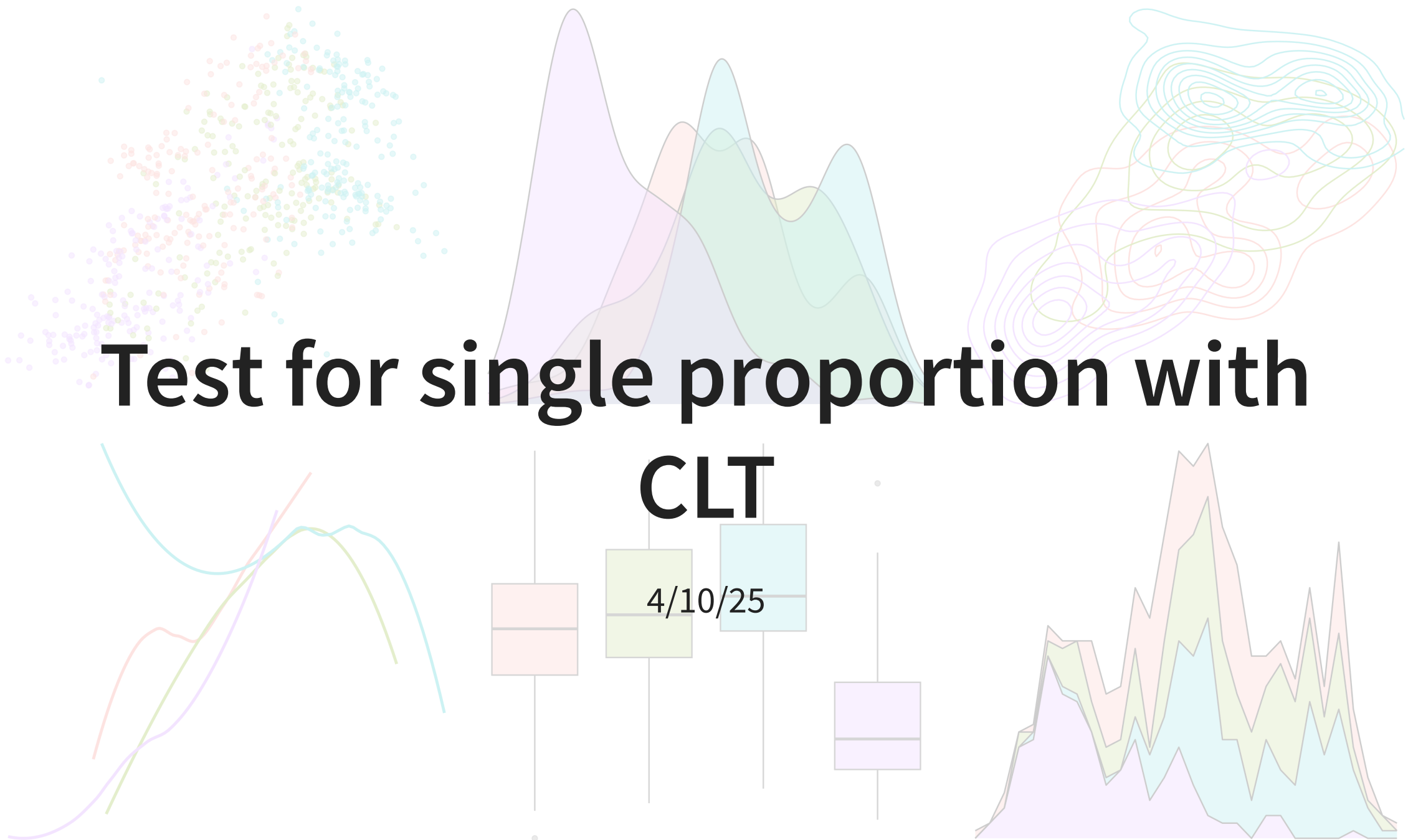# Test for single proportion with CLT

4/10/25

Housekeeping

- Office hours today 2:30-3:30pm!

- Project proposals due Wednesday

CLT recap

Start off with a sample of size n. Assuming **independent** observations,

CLT for sample means

Assume population with mean μ and standard deviation σ. If **normality condition** met:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad \text{or} \quad \bar{X} \mathbin{\dot\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- Normality condition

  1. Exactly normal

  2. Not normal and $n < 30$: no outliers

  3. Not normal and $n \geq 30$: no particularly extreme outliers

CLT for sample proportions

Assume population proportion p. If **success-failure condition** met:

$$\hat{p} \mathbin{\dot\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

- Success-failure condition: $np \geq 10$ and $n(1-p) \geq 10$

Hypothesis testing recap

- Now we're returning to hypothesis testing!

    1. Two sets of hypotheses (competing claims)

    2. Collect data, calculate a statistic from the observed data, set significance level

    3. Obtain p-value from the null distribution

        - p-value: probability of observing data as or more extreme as our own, assuming H_0 true

    4. Make a decision

Hypothesis testing using mathematical model

- We learned how to conduct hypothesis tests (HTs) using simulation to obtain null distribution

- But we can also use CLT to obtain null distribution!

- So the only step that will "look different" is #3: how we obtain our null distribution and p-value

  - Looks different depending on type of data

- Make a conclusion in terms of H_A

Hypothesis test for single proportion

Steps 1 and 2 don't change

Want to conduct a hypothesis test about a population proportion.

1. Define hypotheses

- $H_0 : p = p_0$
- $H_A : p \neq p_0$ (or $H_A : p > p_0$ or $H_A : p < p_0$)
- Remember, p_0 is our "null hypothesized value"

2. Collect data, set significance level

- Obtain observed sample proportion $\hat{p}_{obs}$
- Set α significance level

Null distribution via CLT

Recall CLT for sample proportion: if we have n independent binary observations that satisfy the success-failure condition, then

$$\hat{p} \overset{\cdot}{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

- This is the **sampling distribution** of $\hat{p}$

- But we want the **null distribution** of $\hat{p}$: the sampling distribution under H_0

- We should operate in a world where H_0 is true

- So to use CLT for null distribution, we must satisfy:

  - Independence

  - **Success-failure condition under H_0: $np_0 \geq 10$ and $n(1 - p_0) \geq 10$**

Null distribution via CLT (cont.)

If succes-failure and independence conditions are met and H_0 is true, then CLT tells us our **null distribution** is:

$$\hat{p} \overset{\cdot}{\sim} N \left(p_0, \sqrt{\frac{p_0(1 - p_0)}{n}} \right)$$

- This is how the sample proportion would behave assuming H_0 is true!

- We will use this to see how crazy our $\hat{p}_{obs}$ is

Test statistic

- p-value requires us to compare our observed data to the null distribution

- We calculate a **test statistic**: a quantity that assesses how consistent our sampled data are with H_0

  - Our test statistic is always of the form:

$$\frac{\text{point estimate} - \text{null value}}{\text{SE}_0}$$

where SE_0 represents the standard error under H_0

  - For this specific test, our test statistic is the z-score of $\hat{p}_{obs}$:

$$z = \frac{\hat{p}_{\text{obs}} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

  - Because null distribution is Normal, this z-score follows $N(0, 1)$ distribution!

Obtain p-value

- If $|z|$ large, then that usually means observed value is extremely unusual for H_0, which is convincing evidence against H_0

- p-value is then $\Pr(Z \geq z)$ or $\Pr(Z \leq z)$ (or both), depending on H_A

  - Easily obtained using `pnorm()`

- Step 4: doesn't change!

Example: taste test

Some people claim that they can tell the difference between a diet soda and a regular soda in the first sip. A researcher wanted to test this claim using a hypothesis test at the 0.05 significance level.

- He randomly sampled 80 people.

- He then filled 80 plain white cups with soda, half diet and half regular through random assignment, and asked each person to take one sip from their cup and identify the soda as diet or regular.

- 53 participants correctly identified the soda.

Let p be the proportion of people who correctly identify soda type.

Try defining H_0 and H_A.

Example: taste test (cont.)

1. **Define hypotheses**

   - $H_0 : p = 0.5$ (random guessing)

   - $H_A : p > 0.5$ (better than random guessing)

   - Note: $p_0 = 0.5$ is our null hypothesized value!

2. **Collect data**

   - $\hat{p}_{\text{obs}} = \frac{53}{80} = 0.6625$

- Note: significance level already determined to be 0.05

Example: taste test (cont.)

3. **Obtain null distribution and p-value**

   i. Check conditions for inference satisfied

   - Independence: random sample
   - success-failure: $np_0 = 80(0.5) = 40 \geq 10$ and $n(1 - p_0) = 40 \geq 10$
   - **Because conditions for CLT are met, we can obtain the null distribution using CLT.**

   ii. Null distribution

   $$\hat{p} \overset{.}{\sim} N\left(0.5, \sqrt{\frac{0.5(1 - 0.5)}{80}} = 0.056\right)$$
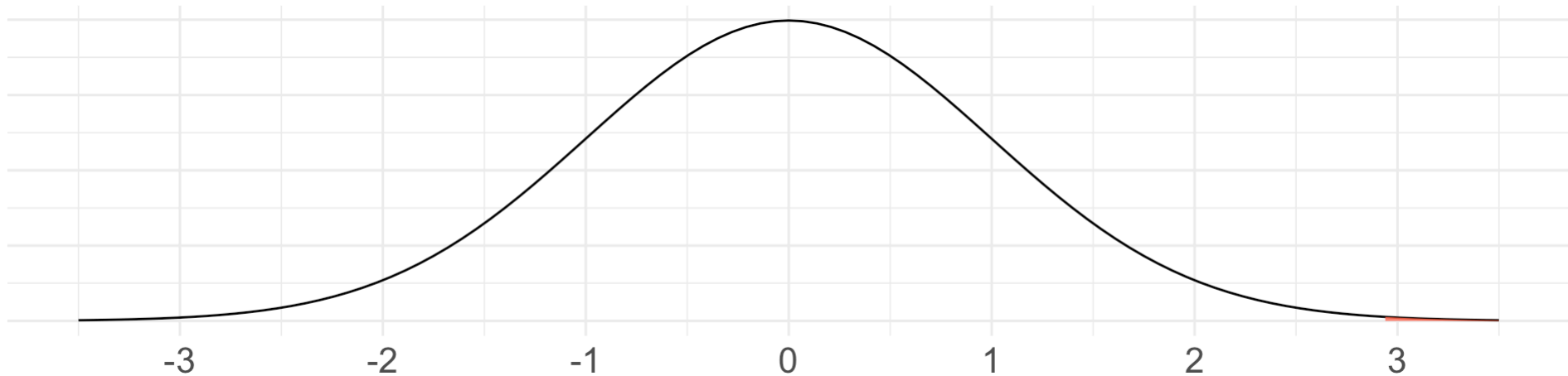
Example: taste test (cont.)

Test statistic:

$$z = \frac{\hat{p}_{obs} - p_0}{\text{SE}_0} = \frac{0.6625 - 0.5}{0.056} = 2.90$$

- This means that if H_0 true, our observed $\hat{p}_{obs}$ is 2.90 SDs above the mean

- Does this seem like "convincing evidence"?

Example: taste test (cont.)

Calculate p-value (i.e. draw picture, set-up the calculation, and/or write code)

$$\text{p-value} = \Pr(Z \geq z) = \Pr(Z \geq 2.90)$$

```
1  1 - pnorm(2.90, 0, 1)
```

```
[1] 0.001865813
```

Example: taste test (cont.)

4. **Decision and conclusion**

- Since our p-value of 0.0019 is less than our significance level of 0.05, we reject H_0.

- The data provide strong evidence that the rate of correctly identifying a soda for these people is better than random guessing.

Example: M&M's

M&M's reported that 14% of its candies are yellow. We are interested in testing this claim. In a random sample of 100 M&M's, 9 were found to be yellow. Conduct a hypothesis test via CLT at the 0.10 level.

1. Write out null and alternative hypotheses in statistical notation (don't forget to define quantities)

2. Summarise data and set α

3. Verify conditions for CLT are met

1. $H_0 : p = 0.14$ vs $H_A : p \neq 0.14$, where $p =$ true proportion of yellow M&M's

2. $\hat{p}_{obs} = \frac{9}{100} = 0.09$

   $\alpha = 0.10$

3. **Independence**? Yes, via random sample

   Success-failure: Yes:
   $np_0 = 100(0.14) = 14 \geq 10$ and
   $n(1 - p_0) = 86 \geq 10$

- **Because conditions for CLT are met, we can obtain the null distribution using CLT.**
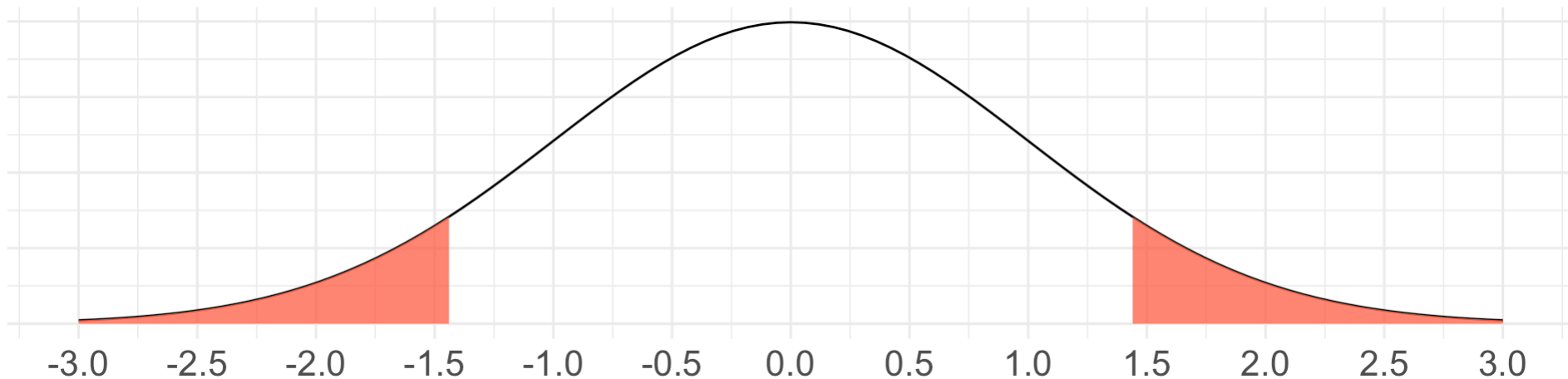
Example: M&M's (cont.)

By CLT, our null distribution is $\hat{p} \stackrel{\cdot}{\sim} N\left(0.14, \sqrt{\frac{0.14(1-0.14)}{100}} = 0.035\right)$

Test statistic:

$$z = \frac{\hat{p}_{obs} - p_0}{\text{SE}_0} = \frac{0.09 - 0.14}{0.035} = -1.43$$

Example: M&M's (cont.)

Since H_A is two-sided, we want

$$\Pr(Z \leq -1.43 \cup Z \geq 1.43)$$
$$= \Pr(Z \leq -1.43) + \Pr(Z \geq 1.43)$$
$$= 2 \times \Pr(Z \leq -1.43)$$

```
1  2 * pnorm(-1.43, 0, 1)
```
```
[1] 0.152717
```

```
1  2 * (1-pnorm(1.43, 0, 1))
```
```
[1] 0.152717
```

Example: M&M's (cont.)

4. <mark>Make a decision and conclusion in context.</mark>

- Since our p-value of 0.153 is greater than our significance level of 0.10, we fail to reject H_0.

- The data do not provide strong enough evidence to suggest that the true proportion of yellow M&Ms is different from 14%.

Comprehension questions

- Why/how do the CLT conditions and/or statements change when obtaining:
    - Sampling distribution of $\hat{p}$
    - Null distribution of $\hat{p}$