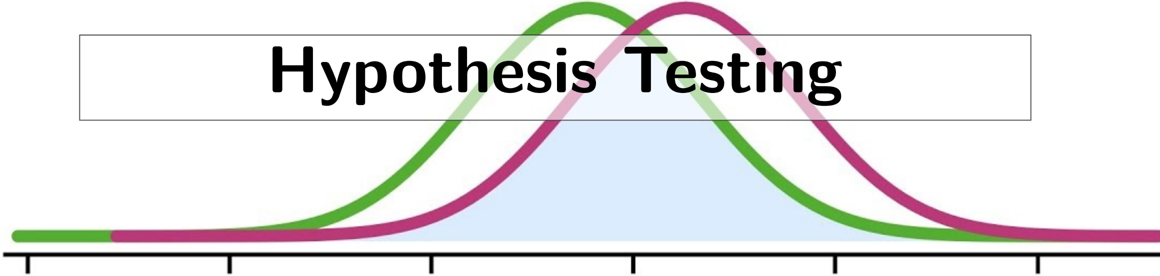
Hypothesis Testing



Learning outcomes:

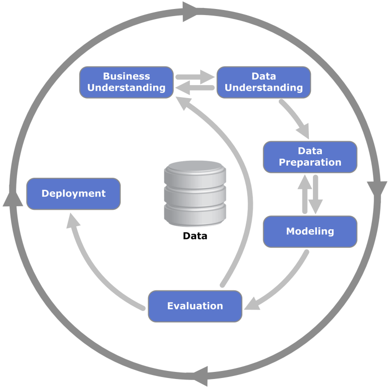
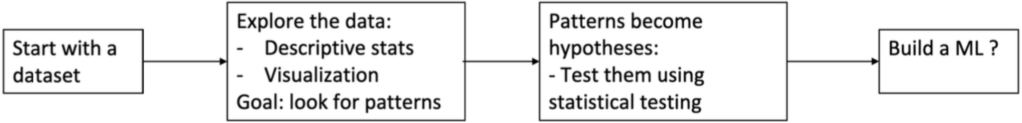


Figure: The CRISP-DM process.

- List good/bad qualities in research questions and hypotheses;
- Define the parts of a hypothesis test;
- Define *p*-values, understand what they mean, and relate them to type I errors;
- Work through an example of a hypothesis test in a gaming context.



Start with a dataset

Explore the data:
- Descriptive stats
- Visualization

Goal: look for patterns

Patterns become hypotheses:
- Test them using statistical testing

Build a ML ?

Research questions

- Before working with data, we should identify research questions and hypotheses;

- A good research question is...
  - ...**Clear** so that it provides enough specifics that one's audience can easily understand its purpose without needing additional explanation;
  - ...**Focused** so that it is narrow enough that it can be answered thoroughly;
  - ...**Specific** so that it explicitly identify variables of interest and their relationship;

Research questions

- Before working with data, we should identify research questions and hypotheses;
- A good research question is...
  - ...**Clear** so that it provides enough specifics that one's audience can easily understand its purpose without needing additional explanation;
  - ...**Focused** so that it is narrow enough that it can be answered thoroughly;
  - ...**Specific** so that it explicitly identify variables of interest and their relationship;

Examples:

How does the education level of the parent impact childhood obesity rates in Phoenix, AZ?

What is the childhood obesity rate in Phoenix, AZ?

What are the effects of childhood obesity in the United States?

How does childhood obesity correlate with academic performance in elementary school children?

What is the relationship between physical activity levels and childhood obesity?

How much time do young children spend doing physical activity per day?

Research questions

- Before working with data, we should identify research questions and hypotheses;
- A good research question is...
  - ...**Clear** so that it provides enough specifics that one's audience can easily understand its purpose without needing additional explanation;
  - ...**Focused** so that it is narrow enough that it can be answered thoroughly;
  - ...**Specific** so that it explicitly identify variables of interest and their relationship;

Examples:

How does the education level of the parent impact childhood obesity rates in Phoenix, AZ?

What is the childhood obesity rate in Phoenix, AZ?

What are the effects of childhood obesity in the United States?

How does childhood obesity correlate with academic performance in elementary school children?

What is the relationship between physical activity levels and childhood obesity?

How much time do young children spend doing physical activity per day?

Hypotheses

- An “educated guess” that answers a research question;

- A well constructed hypothesis is...
  - ...a statement in an “if-then” form;
  - ...specific (mentions variables and the direction of their relationship);
  - ...testable (falsifiable);

- We will always be seeking to DISPROVE a hypothesis!

Hypothesis testing is a model that helps us decide between different hypotheses using **falsification**.

Parts of a hypothesis test

1. A **Null hypothesis $\mathbf{H_0}$**;
    - Usually a claim that there is no effect or nothing of interest;
    - We will be looking to decide if the data falsifies the null hypothesis;

2. An **Alternative hypothesis $\mathbf{H_A}$**;
    - Usually a claim that there IS an effect;
    - We want to create support for this by falsifying its converse;

Parts of a hypothesis test

1. A **Null hypothesis $\mathbf{H_0}$**;
   - Usually a claim that there is no effect or nothing of interest;
   - We will be looking to decide if the data falsifies the null hypothesis;

2. An **Alternative hypothesis $\mathbf{H_A}$**;
   - Usually a claim that there IS an effect;
   - We want to create support for this by falsifying its converse;

3. A **test statistic**;
   - A point estimate summary of the data that we can use to decide between H_0 and H_A;
   - Observed test statistic: the actual value of the test stat we observed IRL;
   - Null distribution: sampling distribution of the test statistic assuming H_0 is true;

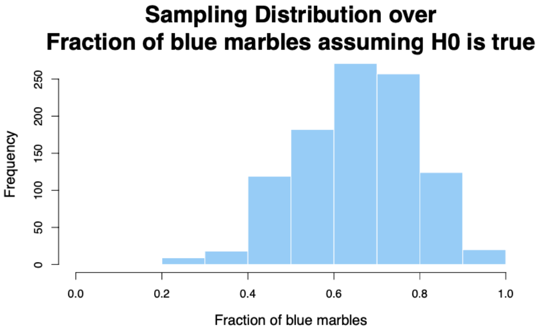
Parts of a hypothesis test

1. A **Null hypothesis $\mathbf{H_0}$**;
   - Usually a claim that there is no effect or nothing of interest;
   - We will be looking to decide if the data falsifies the null hypothesis;
2. An **Alternative hypothesis $\mathbf{H_A}$**;
   - Usually a claim that there IS an effect;
   - We want to create support for this by falsifying its converse;
3. A **test statistic**;
   - A point estimate summary of the data that we can use to decide between H_0 and H_A;
   - Observed test statistic: the actual value of the test stat we observed IRL;
   - Null distribution: sampling distribution of the test statistic assuming H_0 is true;
4. A **rejection criterion** (also called the significance level);
   - If we see this happen to the test statistic then we will decide we have falsified or rejected H_0;
   - Largely an arbitrary choice.

Null distribution

- **Remember the marbles in the bag?** We didn't talk about it but our test statistic was the fraction of blue marbles in the sample;

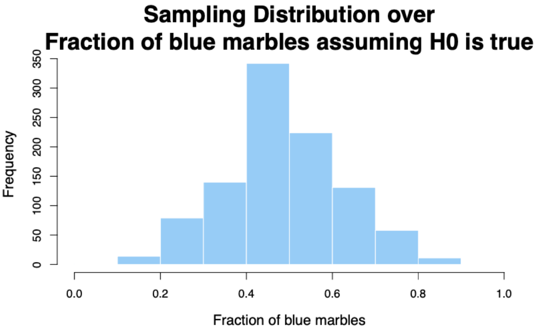
- Imagine when we did that our null hypothesis was that the fraction of blue marbles = 66%. Then the sampling (null) distribution ought to look like:



Null distribution

- **Remember the marbles in the bag?** We didn't talk about it but our test statistic was the fraction of blue marbles in the sample;

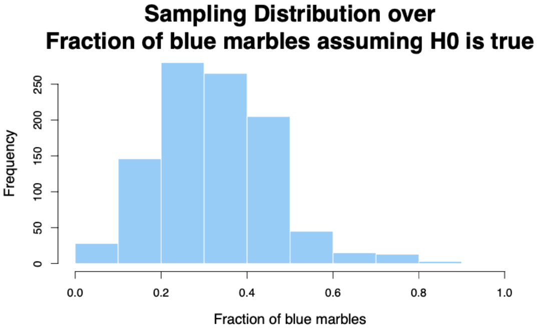
- Imagine when we did that our null hypothesis was that the fraction of blue marbles = 50%. Then the sampling (null) distribution ought to look like:



Null distribution

- **Remember the marbles in the bag?** We didn't talk about it but our test statistic was the fraction of blue marbles in the sample;

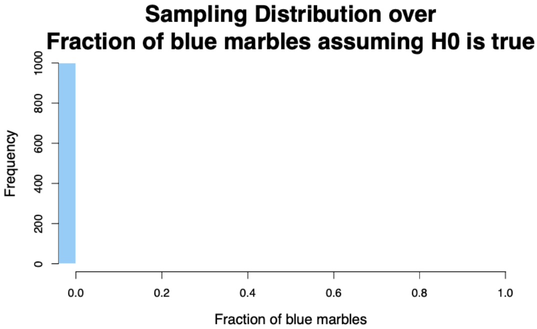
- Imagine when we did that our null hypothesis was that the fraction of blue marbles $= 33\%$. Then the sampling (null) distribution ought to look like:



Null distribution

- **Remember the marbles in the bag?** We didn't talk about it but our test statistic was the fraction of blue marbles in the sample;

- Imagine when we did that our null hypothesis was that the fraction of blue marbles $= 0\%$. Then the sampling (null) distribution ought to look like:



Choosing the rejection criterion: Type I errors and *p*-values

- A **type I error** is when we reject H_0 when it is true;
  - This would mean mistakenly endorsing H_A;
  - A similar idea is that of a false positive;
  - The probability of a type I error is usually called α;

- A **p-value** is defined as the probability of getting a more extreme value than the observed test statistic given that H_0 is true;
  - Measures surprise;
  - Comes from the null distribution;
  - The lower the *p*-value the lower the risk of a type I error;

Choosing the rejection criterion: Type I errors and p-values

- A **type I error** is when we reject H_0 when it is true;
  - This would mean mistakenly endorsing H_A;
  - A similar idea is that of a false positive;
  - The probability of a type I error is usually called α;

- A **p-value** is defined as the probability of getting a more extreme value than the observed test statistic given that H_0 is true;
  - Measures surprise;
  - Comes from the null distribution;
  - The lower the p-value the lower the risk of a type I error;

- Choosing a rejection criterion or significance level is equivalent to choosing a cutoff for α – by convention this is usually 0.05:
  - We reject H_0 when we are sufficiently unlikely to be making a type I error.
  - We reject H_0 when the p-value is less than our α cutoff of 0.05.

Motivation – reasoning about loot boxes...

- Imagine a player in CounterStrike claiming that opening loot boxes gives them a exotic item 50% of the time.
- The player insists this isn't just luck—it's consistent.
- How can we assess this claim? Conduct an experiment:
    - **Sample** (open) n loot boxes in the game;
    - Record the number of exotic items obtained.

Applying this to reason about loot boxes – let's build a hypothesis test

- Suppose we collect a sample of 15 loot boxes and we see that 6 of them are exotics. Do we believe the player's loot box percentage claim?

- H_0:

- H_A:

- Test statistic:

Applying this to reason about loot boxes – let's build a hypothesis test

- Suppose we collect a sample of 15 loot boxes and we see that 6 of them are exotics. Do we believe the player's loot box percentage claim?

- H_0: the player's claim is true, i.e. the fraction of loot boxes that comes out as exotics $\pi = 0.5$;

- H_A: the player's claim is false, i.e. the fraction of loot boxes that comes out as exotics is $\pi < 0.5$;

- Test statistic:

Applying this to reason about loot boxes – let's build a hypothesis test

- Suppose we collect a sample of 15 loot boxes and we see that 6 of them are exotics. Do we believe the player's loot box percentage claim?

- H_0: the player's claim is true, i.e. the fraction of loot boxes that comes out as exotics $\pi = 0.5$;

- H_A: the player's claim is false, i.e. the fraction of loot boxes that comes out as exotics is $\pi < 0.5$;

- Test statistic: the number of exotics in our sample, in this case 6;

Applying this to reason about loot boxes – let's build a hypothesis test

- Suppose we collect a sample of 15 loot boxes and we see that 6 of them are exotics. Do we believe the player's loot box percentage claim?

- Our hypothesis test model says we reject H_0 if:
  - The probability of a more extreme result than our observed test statistic assuming H_0 is true is less than 0.05;

Applying this to reason about loot boxes – let's build a hypothesis test

- Suppose we collect a sample of 15 loot boxes and we see that 6 of them are exotics. Do we believe the player's loot box percentage claim?

- Our hypothesis test model says we reject H_0 if:
    - The probability of a more extreme result than our observed test statistic assuming H_0 is true is less than 0.05;

      The probability of a more extreme result than our observed test statistic according to the null distribution is less than 0.05;

Applying this to reason about loot boxes – let's build a hypothesis test

- Suppose we collect a sample of 15 loot boxes and we see that 6 of them are exotics. Do we believe the player's loot box percentage claim?

- Our hypothesis test model says we reject H_0 if:
  - The probability of a more extreme result than our observed test statistic assuming H_0 is true is less than 0.05;
    The probability of a more extreme result than our observed test statistic according to the null distribution is less than 0.05;
    The probability of 6 or fewer exotics in our sample according to the null distribution is less than 0.05;

Applying this to reason about loot boxes – let's build a hypothesis test

- Suppose we collect a sample of 15 loot boxes and we see that 6 of them are exotics. Do we believe the player's loot box percentage claim?

- Our hypothesis test model says we reject H_0 if:
  - The probability of a more extreme result than our observed test statistic assuming H_0 is true is less than 0.05;
    The probability of a more extreme result than our observed test statistic according to the null distribution is less than 0.05;
    The probability of 6 or fewer exotics in our sample according to the null distribution is less than 0.05;

- So what is the probability of 6 or fewer exotics in our sample? It would be:

$$
\begin{aligned}
& p(0 \text{ exotics}) + p(1 \text{ exotics}) + p(2 \text{ exotics}) \\
& \quad p(3 \text{ exotics}) + p(4 \text{ exotics}) + p(5 \text{ exotics}) \\
& \quad + p(6 \text{ exotics}).
\end{aligned}
$$

Applying this to reason about loot boxes – let's build a hypothesis test

- Suppose we collect a sample of 15 loot boxes and we see that 6 of them are exotics. Do we believe the player's loot box percentage claim?

- Our hypothesis test model says we reject H_0 if:
  - The probability of a more extreme result than our observed test statistic assuming H_0 is true is less than 0.05;
    The probability of a more extreme result than our observed test statistic according to the null distribution is less than 0.05;
    The probability of 6 or fewer exotics in our sample according to the null distribution is less than 0.05;
  - The p-value is less than 0.05;

- So what is the probability of 6 or fewer exotics in our sample? It would be:

$$\begin{aligned} &p(0 \text{ exotics}) + p(1 \text{ exotics}) + p(2 \text{ exotics}) \\ &\quad p(3 \text{ exotics}) + p(4 \text{ exotics}) + p(5 \text{ exotics}) \\ &\quad + p(6 \text{ exotics}). \end{aligned}$$

Applying this to reason about loot boxes – let’s build a hypothesis test

- Where do these probabilities come from?

$$
\begin{aligned}
p(0 \text{ exotics}) &= ???\\
p(1 \text{ exotic}) &= ???\\
p(2 \text{ exotics}) &= ???\\
p(3 \text{ exotics}) &= ???\\
p(4 \text{ exotics}) &= ???\\
p(5 \text{ exotics}) &= ???\\
p(6 \text{ exotics}) &= ???
\end{aligned}
$$

Applying this to reason about loot boxes – let's build a hypothesis test

- Where do these probabilities come from? Well, from the null distribution.

$$
\begin{aligned}
p(0 \text{ exotics}) &= ???\\
p(1 \text{ exotic}) &= ???\\
p(2 \text{ exotics}) &= ???\\
p(3 \text{ exotics}) &= ???\\
p(4 \text{ exotics}) &= ???\\
p(5 \text{ exotics}) &= ???\\
p(6 \text{ exotics}) &= ???
\end{aligned}
$$

Applying this to reason about loot boxes – let's build a hypothesis test

- Where do these probabilities come from? Well, from the null distribution.
- Remember the **binomial distribution** that models the probability of k successes in n trials given that each trial has probability of success π?

$$
\begin{aligned}
p(0\text{ exotics}) &= ???\\
p(1\text{ exotic}) &= ???\\
p(2\text{ exotics}) &= ???\\
p(3\text{ exotics}) &= ???\\
p(4\text{ exotics}) &= ???\\
p(5\text{ exotics}) &= ???\\
p(6\text{ exotics}) &= ???
\end{aligned}
$$

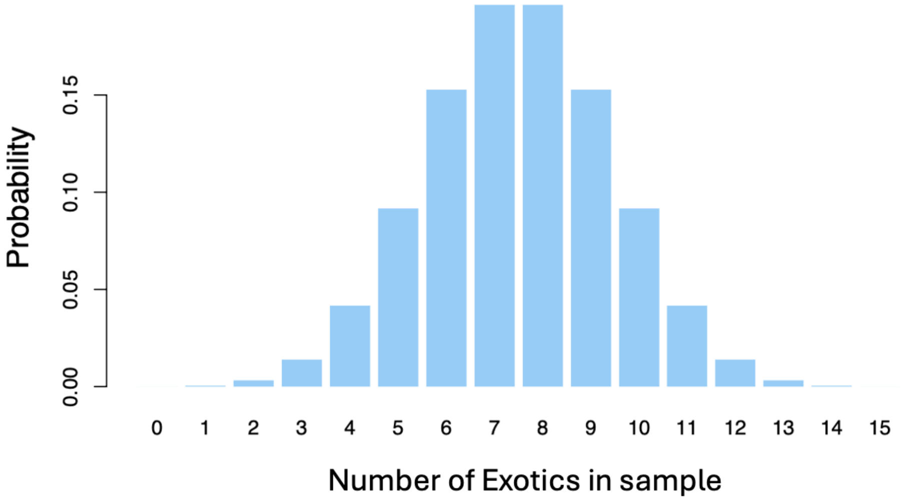
Applying this to reason about loot boxes – let's build a hypothesis test

- Where do these probabilities come from? Well, from the null distribution.

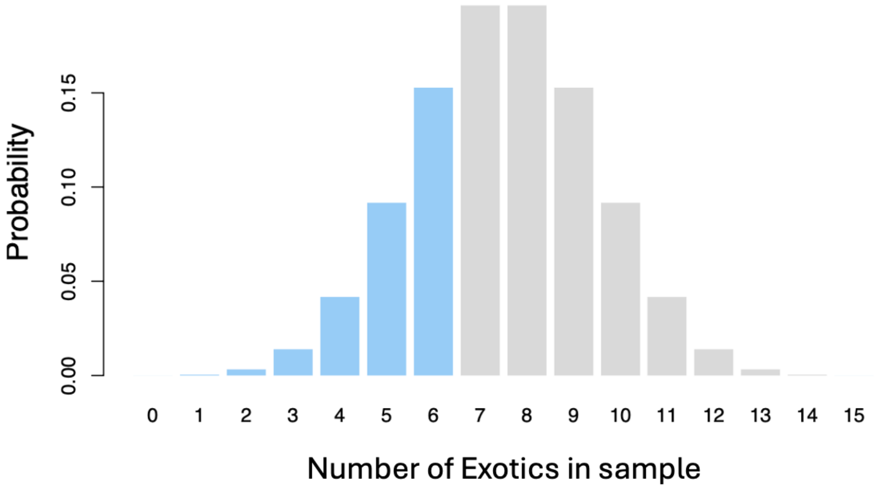
- Remember the **binomial distribution** that models the probability of k successes in 15 trials given that each trial has probability of success $\pi = 0.5$?

$$
\begin{aligned}
p(0 \text{ exotics}) &= p(k = 0, n = 15, \pi = 0.5) \approx 0.00003 \\
p(1 \text{ exotic}) &= p(k = 1, n = 15, \pi = 0.5) \approx 0.00046 \\
p(2 \text{ exotics}) &= p(k = 2, n = 15, \pi = 0.5) \approx 0.00320 \\
p(3 \text{ exotics}) &= p(k = 3, n = 15, \pi = 0.5) \approx 0.01389 \\
p(4 \text{ exotics}) &= p(k = 4, n = 15, \pi = 0.5) \approx 0.04166 \\
p(5 \text{ exotics}) &= p(k = 5, n = 15, \pi = 0.5) \approx 0.09164 \\
p(6 \text{ exotics}) &= p(k = 6, n = 15, \pi = 0.5) \approx 0.15274
\end{aligned}
$$

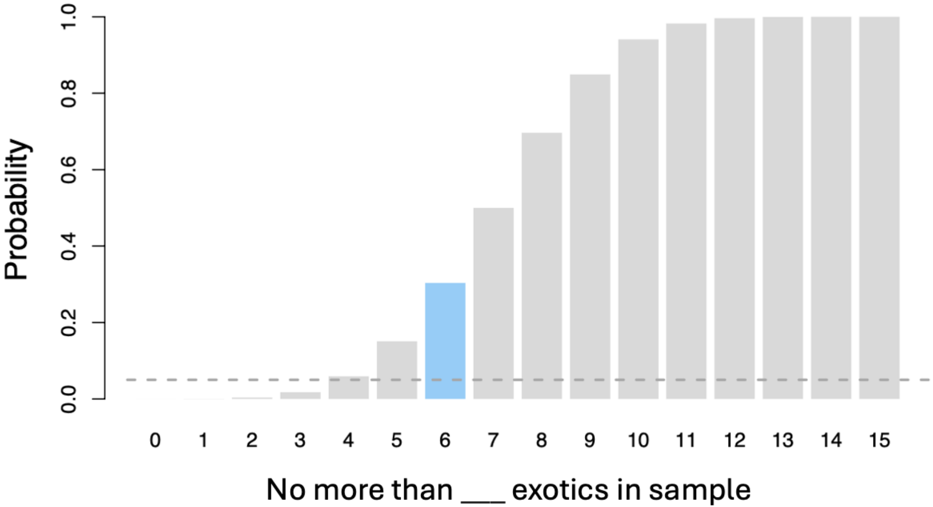
The null distribution in our hypothesis test (comes from the Binomial)



The null distribution in our hypothesis test (comes from the Binomial)



The possible *p*-values in our hypothesis test



So...

- In our Counter Strike player example the p-value is $\approx 0.3 > 0.05$ meaning we fail to reject H_0 that we will pull an exotic;

- If we had seen $4, 5, \ldots, 15$ exotics in our sample of 15 the p-value would have been greater than 0.05 and so we would have failed to reject H_0 the player's statement is true;

- If we had seen 0, 1, 2, or 3 exotics in our sample of 15 the p-value would have been less than 0.05 and so we would have rejected H_0 that the player's statement is true;

- Whenever the p-value < 0.05 we would reject H_0.