

Lecture 18: Decisions and Uncertainty

Incomplete Information

An important aspect of data science is to assist the decision-making process when we have incomplete information.

Alabama vs. Swain, we had to choose between:

- The view that the panel was chosen at random (that was the model),
- And the view that it was not (this was the *alternative view*).

In the case of Mendel's pea plants, we had to choose between:

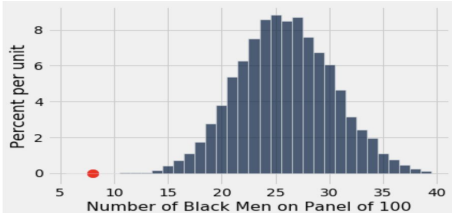
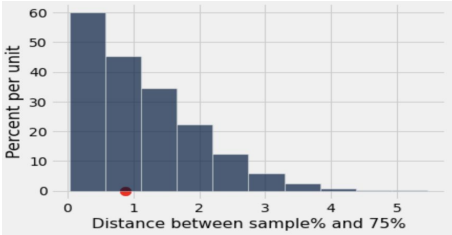
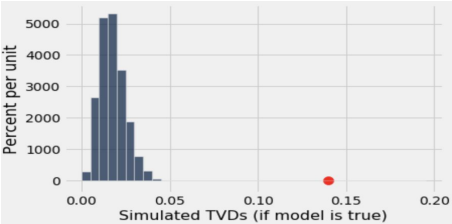
- The view that each plant has a 75% chance of having purple flowers (that was the model), and
- The view that it does not.

In the case of the Alameda jury panels, we had to choose between:

- The view that the jury panels were chosen at random (that was the model), and
- the view that they were not.

In each case, how did we choose between views?

- We chose a statistic to measure possible discrepancy between the model and the observed data
- We simulated this statistic under the assumptions of the model
- We compared the statistic of the observed data to the model's prediction
- We made our decision based upon whether or not the observed data was consistent with the model's prediction.

Example	Statistic	Comparison of observed data with model's prediction
Swain v. Alabama	Number of black men in a jury panel of 100 people	 <p>A histogram showing the distribution of the number of black men on a panel of 100. The x-axis is labeled 'Number of Black Men on Panel of 100' and ranges from 5 to 40. The y-axis is labeled 'Percent per unit' and ranges from 0 to 8. The distribution is bell-shaped, centered around 25. A red dot is placed at 8 on the x-axis.</p>
Mendel's pea plants	<p>Absolute value of the difference in percentage of purple flowers in sample and 75%, or...</p> $ \% \text{ purple in sample} - 75\% $	 <p>A histogram showing the distribution of the absolute difference between the percentage of purple flowers in a sample and 75%. The x-axis is labeled 'Distance between sample% and 75%' and ranges from 0 to 5. The y-axis is labeled 'Percent per unit' and ranges from 0 to 60. The distribution is right-skewed, with a peak at 0. A red dot is placed at 0.5 on the x-axis.</p>
Alameda Jury panel	Total Variation Distance of sample from Eligible population	 <p>A histogram showing the distribution of simulated Total Variation Distances (TVDs) if the model is true. The x-axis is labeled 'Simulated TVDs (if model is true)' and ranges from 0.00 to 0.20. The y-axis is labeled 'Percent per unit' and ranges from 0 to 5000. The distribution is right-skewed, with a peak around 0.01. A red dot is placed at 0.14 on the x-axis.</p>

Testing Hypotheses

hypothesis testing: Choosing between a model and its alternative

- Model = Null Hypothesis
- Alternative = Alternative Hypothesis

A hypothesis test picks the hypothesis that is better supported by, or is consistent with, the observed data.

Example	Null Hypothesis	Alternative Hypothesis	Conclusion
Swain v. Alabama	Jury panel selection was done with a random process	Selection process was <i>not</i> random	Observed data <i>did not</i> support null hypothesis
Mendel's pea plants	A pea plant has a probability of 75% of having purple flowers	The probability that a pea plant's flowers are purple is <i>not</i> 75%	Observed data was consistent with null hypothesis
Alameda Jury panel	Jury panel selection was done with a random process	Selection process was <i>not</i> random	Observed data <i>did not</i> support null hypothesis

Null and Alternative

Hypothesis testing only works if we can simulate data under the null hypothesis. The two hypotheses are formally defined as follows:

- **Null Hypothesis**

- A well-defined chance model about how data are generated.
- We can simulate data under the assumptions of this model - we say that data is simulated **'under the null hypothesis'**.

- **Alternative Hypothesis**

- Is a different view about the origin of the data.

The Test Statistic

In order to test the null hypothesis we must choose a test statistic that will allow us to decide between the two hypotheses, as we saw in our previous cases.

Example	Statistic
Swain v. Alabama	Number of black men in a jury panel of 100 people
Mendel's pea plants	Absolute value of the difference in percentage of purple flowers in sample and 75%, or... $ \% \text{ purple in sample} - 75\% $
Alameda Jury panel	Total Variation Distance of sample from Eligible population

Guidelines to consider before choosing a statistic.

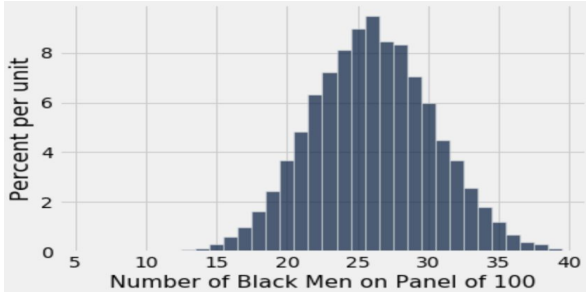
- What values of the statistic will make us lean towards the null hypothesis?
- What values will make us lean towards the alternative?
 - Preferably, the answer should be just “high” or just “low”. Try to avoid “both high and low”.

Prediction 'Under the Null Hypothesis'

After test statistic...

- Simulate the test statistic under the null hypothesis
- Make histogram: **empirical distribution of the statistic under the null hypothesis.**
- This empirical distribution is a prediction about the statistic
 - It shows all the values of the statistic, and their corresponding probabilities
- Probabilities are approximate

Prediction 'Under the Null Hypothesis'

Example	Statistic	Empirical prediction of the statistic under the Null hypothesis																																														
Swain v. Alabama	Number of black men in a jury panel of 100 people	 <p>The histogram displays a normal distribution of the number of black men on a jury panel of 100 people. The x-axis, labeled 'Number of Black Men on Panel of 100', ranges from 5 to 40 with major ticks every 5 units. The y-axis, labeled 'Percent per unit', ranges from 0 to 8 with major ticks every 2 units. The distribution is centered at 25, with the highest frequency (approximately 8.5%) occurring at 25. The data is spread between approximately 15 and 35.</p> <table><caption>Estimated data for the histogram</caption><tr><th>Number of Black Men</th><th>Percent per unit</th></tr><tr><td>15</td><td>0.2</td></tr><tr><td>16</td><td>0.5</td></tr><tr><td>17</td><td>1.0</td></tr><tr><td>18</td><td>1.5</td></tr><tr><td>19</td><td>2.5</td></tr><tr><td>20</td><td>4.0</td></tr><tr><td>21</td><td>5.0</td></tr><tr><td>22</td><td>6.5</td></tr><tr><td>23</td><td>7.5</td></tr><tr><td>24</td><td>8.0</td></tr><tr><td>25</td><td>8.5</td></tr><tr><td>26</td><td>8.0</td></tr><tr><td>27</td><td>7.5</td></tr><tr><td>28</td><td>6.5</td></tr><tr><td>29</td><td>5.0</td></tr><tr><td>30</td><td>4.0</td></tr><tr><td>31</td><td>3.0</td></tr><tr><td>32</td><td>2.0</td></tr><tr><td>33</td><td>1.5</td></tr><tr><td>34</td><td>1.0</td></tr><tr><td>35</td><td>0.5</td></tr><tr><td>36</td><td>0.2</td></tr></table>	Number of Black Men	Percent per unit	15	0.2	16	0.5	17	1.0	18	1.5	19	2.5	20	4.0	21	5.0	22	6.5	23	7.5	24	8.0	25	8.5	26	8.0	27	7.5	28	6.5	29	5.0	30	4.0	31	3.0	32	2.0	33	1.5	34	1.0	35	0.5	36	0.2
Number of Black Men	Percent per unit																																															
15	0.2																																															
16	0.5																																															
17	1.0																																															
18	1.5																																															
19	2.5																																															
20	4.0																																															
21	5.0																																															
22	6.5																																															
23	7.5																																															
24	8.0																																															
25	8.5																																															
26	8.0																																															
27	7.5																																															
28	6.5																																															
29	5.0																																															
30	4.0																																															
31	3.0																																															
32	2.0																																															
33	1.5																																															
34	1.0																																															
35	0.5																																															
36	0.2																																															

Conclusion of the Test

How do we decide?

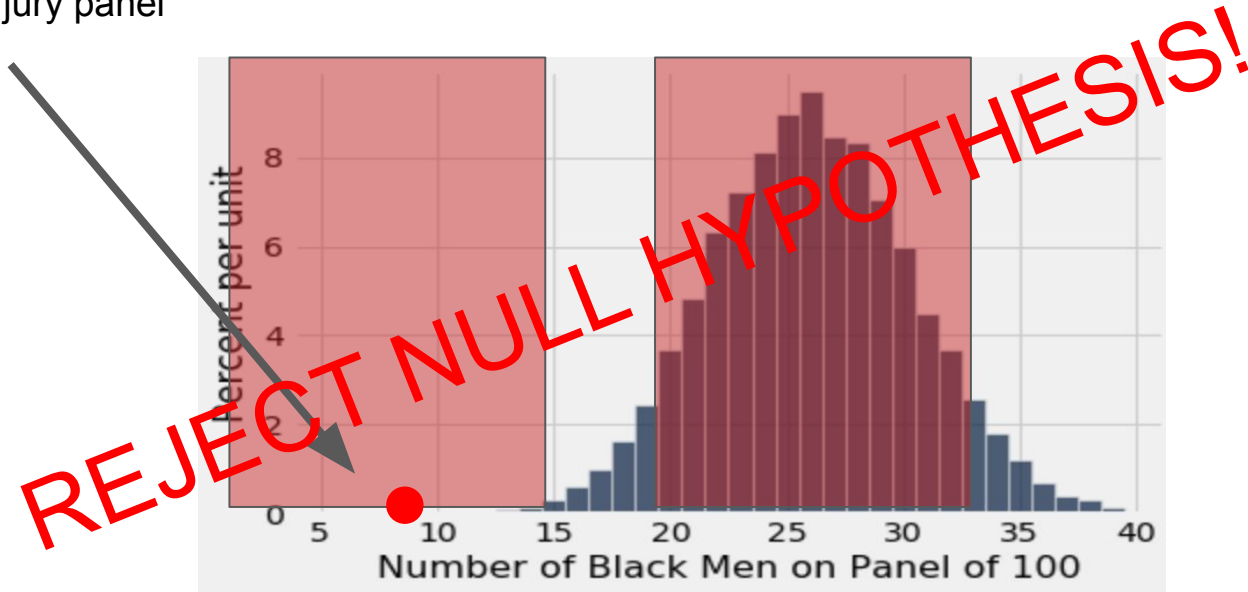
- We compare the **observed test statistic** and the empirical distribution of the statistic under the null hypothesis
- If the observed value is **not consistent** with the distribution, then the test favors the alternative

To determine whether a value is consistent with a distribution:

- A visualization may be sufficient.
- If not, there are conventions about “consistency”.

Swain vs. Alabama

Observed test statistic:
8 black men on jury panel

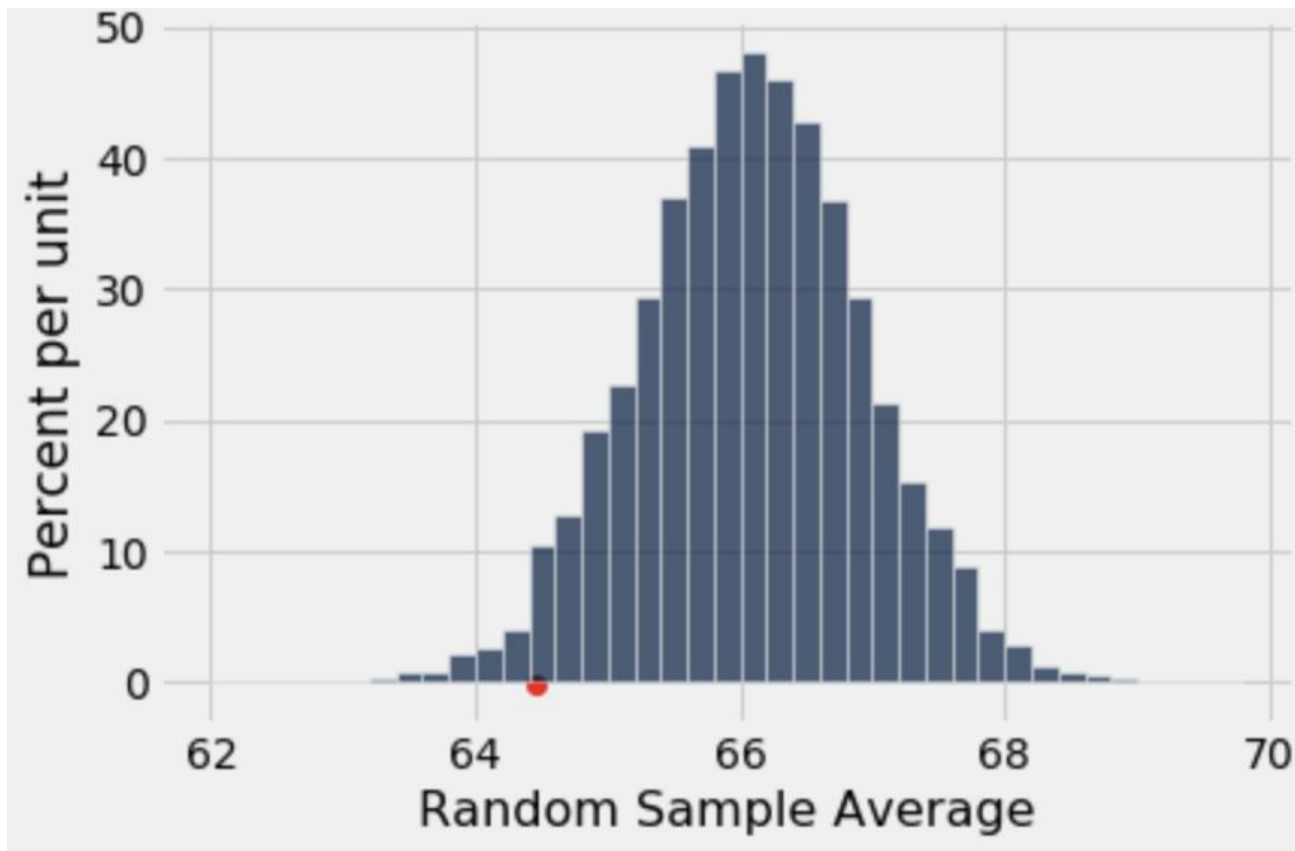


Observed test statistic is
NOT consistent with prediction

Observed test statistic is
consistent with prediction

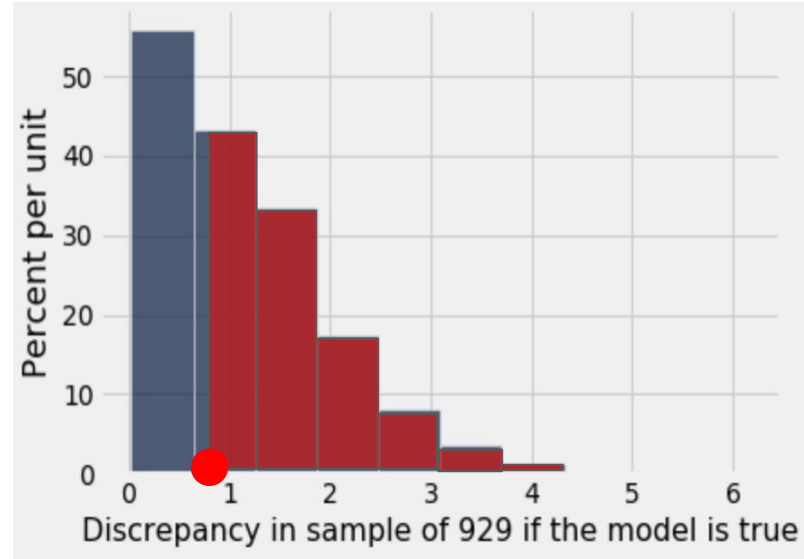
Go to Lecture 18 Jupyter notebook [here](#); pick up [here](#) at the end of the notebook.

Student Scores



Statistical Significance

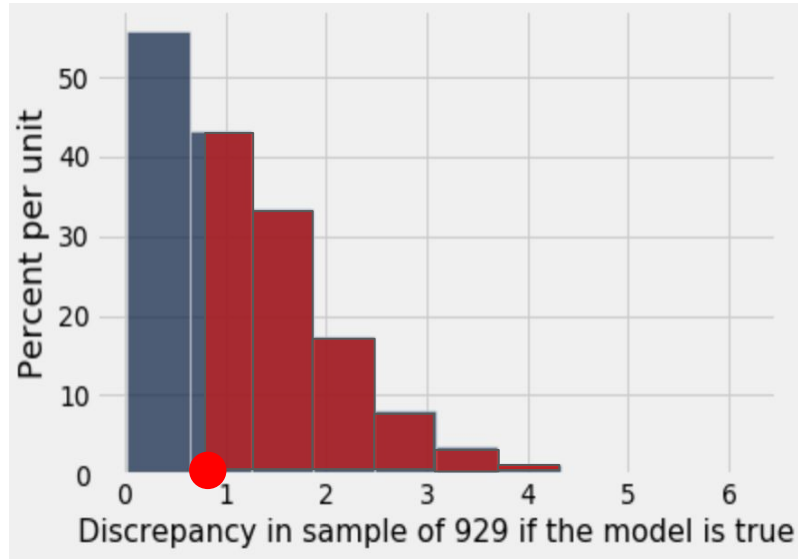
Mendel's pea plants: the *tail* of the observed test statistic of 0.88% is the portion of the histogram *to the right* of 0.88%.



The *tail* is the portion of the histogram starting at the observed statistic and looking in the direction that makes us lean toward the alternative.

Statistical Significance

In Mendel's case, the area in the tail is the percentage of sample averages that are *greater than or equal to* the value of the statistic.



The *smaller* this area, *the more statistically significant* the observed statistic.

Conventions about inconsistency

- If the area in the tail is less than 5%, we say that the result is **statistically significant**.
- If the area in the tail is less than 1%, we say that the result is **highly statistically significant**.

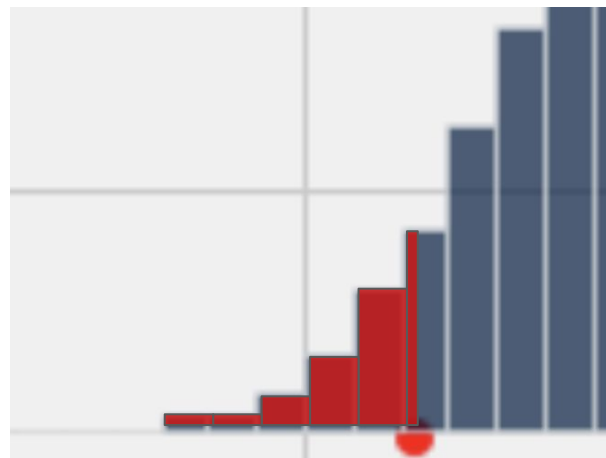
The test statistic is “Inconsistent with the null” if it is in the tail of the empirical distribution under the null hypothesis.

The P-value of a test

Definition: The P-value of a test is the chance, based on the model in the null hypothesis, that the test statistic will be equal to the observed value in the sample or even further in the direction that supports the alternative.

In other words, the P-value is the area of the tail beyond the observed value.

Small P-values: observed test statistic far from prediction; data supports alternative



The P-value of a test

"If one in twenty does not seem high enough odds, we may, if we prefer it draw the line at one in fifty (the 2 percent point), or one in a hundred (the 1 percent point). Personally, the author prefers to set a low standard of significance at the 5 percent point ..."

-Sir Ronald Fisher, 1925