Université
de Technologie
Tarbes
Occitanie Pyrénées

THE **COMPLETE GUIDE** TO

INFERENTIAL STATISTICS

< By Raymond HOUE NGOUNA />

Université de Technologie Tarbes Occitanie Pyrénées

</SECTION 2>

CHI-SQUARE TEST

Université de Technologie Tarbes Occitanie Pyrénées

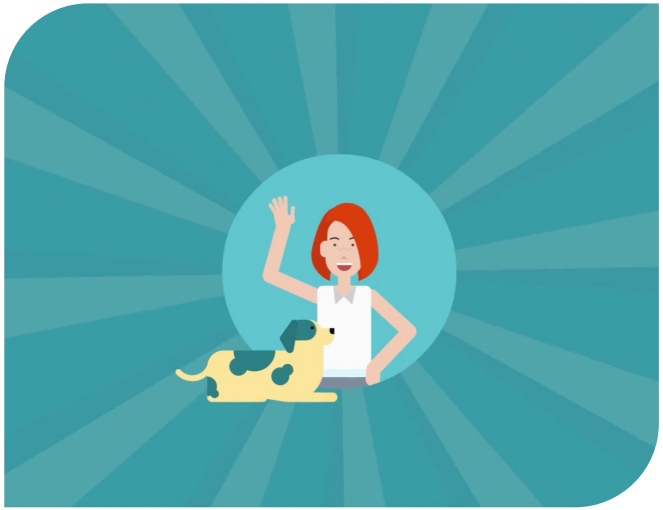
SECTION OUTLINE

- ❑ Case study
- ❑ Key concepts
  - ✓ Sample
  - ✓ Statistic & Parameter
  - ✓ Levels of measurement
  - ✓ Contingency table
  - ✓ Hypothesis testing
  - ✓ Test statistic
  - ✓ p-value
  - ✓ Statistical significance
- ❑ Summary
  - ✓ Chi-square test implementation

CASE STUDY

THE CASE STUDY

- ✓ Kelly was recently is responsible for a leading dog food brand.
- ✓ Brian, her colleague in the packaging department, introduces Kelly to their latest innovation
  - a new type of packaging that he believes will revolutionize the market.



➔ But when Kelly sees this new design for the first time, her reaction is mixed: the packaging, while innovative, appears bulky.

CONTROVERSIAL ASSUMPTION

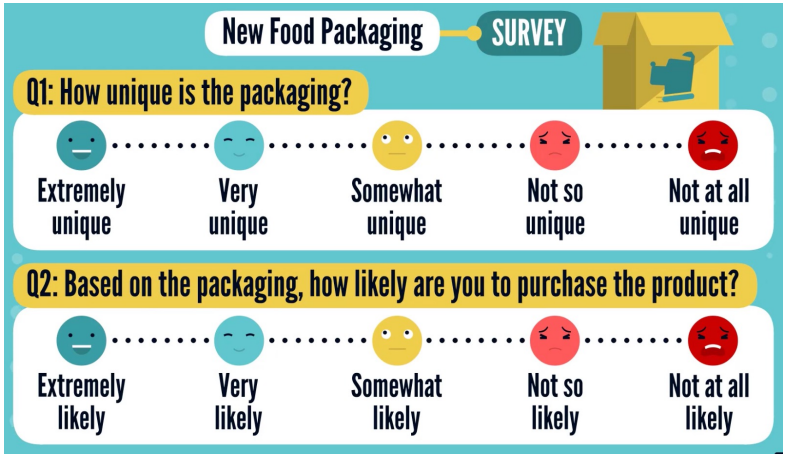
- ✓ Everyone in Kelly's team finds the new packaging unique.
- ✓ Yet Kelly's views differ on whether this uniqueness can boost sales.

➔ Being relatively new to this product category, Kelly chooses not to rely solely on her limited experience and personal feelings:

- she opts for a more objective approach by deciding to conduct a survey,
- this market research should enable her to test the various hypotheses that emerge throughout her exchange with Brian and the packaging team.

KELLY'S APPROACH

- ✓ Kelly's team then randomly selected 523 respondents and conducted a survey.
- ✓ The survey focuses on two pivotal questions:

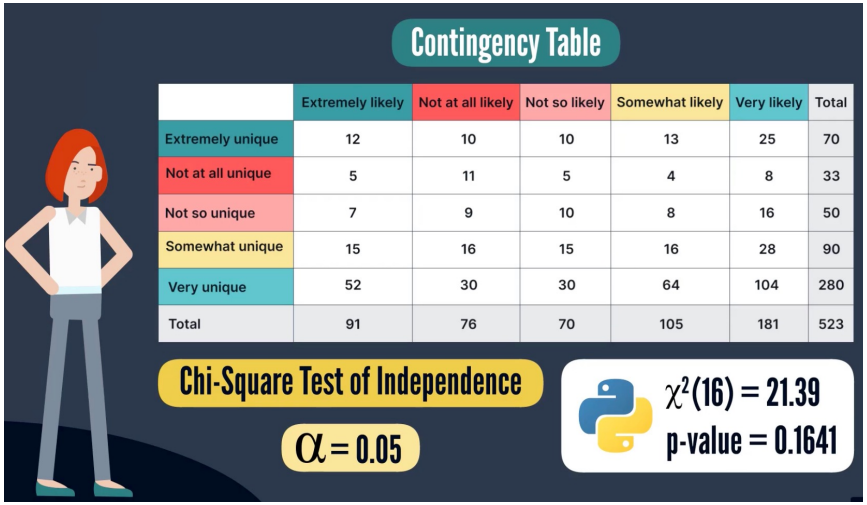


(1) The first examines the distinctiveness of the packaging, requesting participants to assess it on a scale from extremely unique to not at all unique.

(2) The second inquires into the probability of purchasing the product solely on its packaging, offering choices from extremely likely to not at all likely.

THE STUDY OUTCOMES

- ✓ Kelly compiled the data into a contingency table
  - showing how packaging uniqueness is related to purchasing tendencies.
- ✓ She then ran a chi-square test
  - setting the significance level at 0.05
  - exhibiting a chi-square value of 21.39 and a p-value of 0.1641.



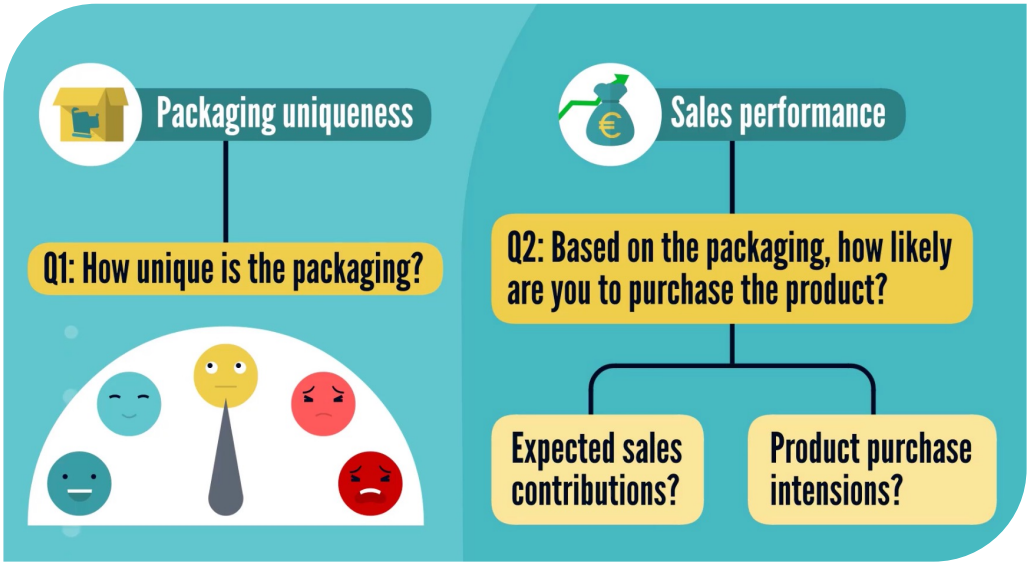
	Extremely likely	Not at all likely	Not so likely	Somewhat likely	Very likely	Total
Extremely unique	12	10	10	13	25	70
Not at all unique	5	11	5	4	8	33
Not so unique	7	9	10	8	16	50
Somewhat unique	15	16	15	16	28	90
Very unique	52	30	30	64	104	280
Total	91	76	70	105	181	523

THE STUDY OUTCOMES

- ✓ Despite the packaging's distinctiveness, Kelley concludes that its uniqueness doesn't necessarily guarantee increased sales.
- ✓ Therefore, she decides not to proceed with the new packaging.

2 CONCEPTS NEEDED TO BE OPERATIONALIZED

(1) packaging uniqueness and (2) sales performance

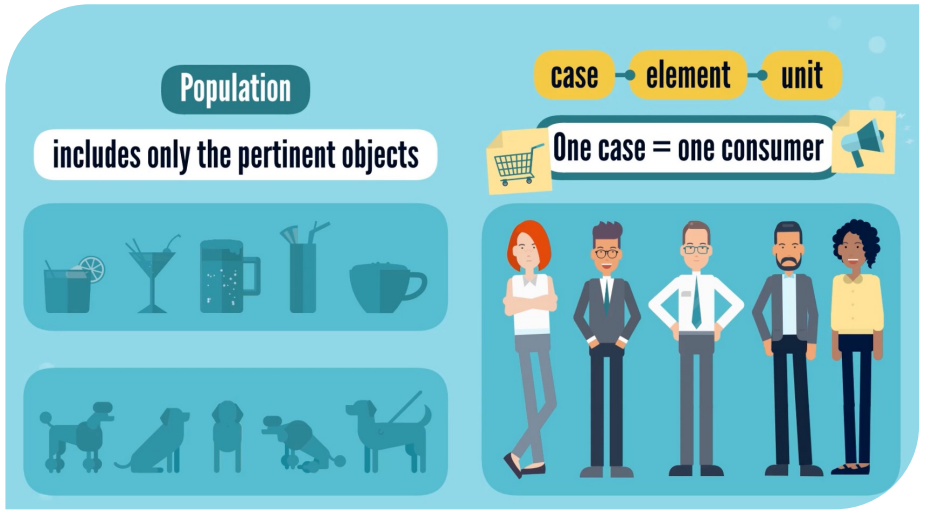


➔ *need to measure the sales performance differently*
➔ *the second question in the survey asks respondents to rate the likelihood of purchasing the product;*
➔ *good indicator of sales performance.*

SAMPLE

POPULATION

- ✓ A population is the collection of all cases the analyst wants to consider.
  - A case or element or unit represents the smallest object of study.
  - The population includes only objects that are pertinent to the examinations.



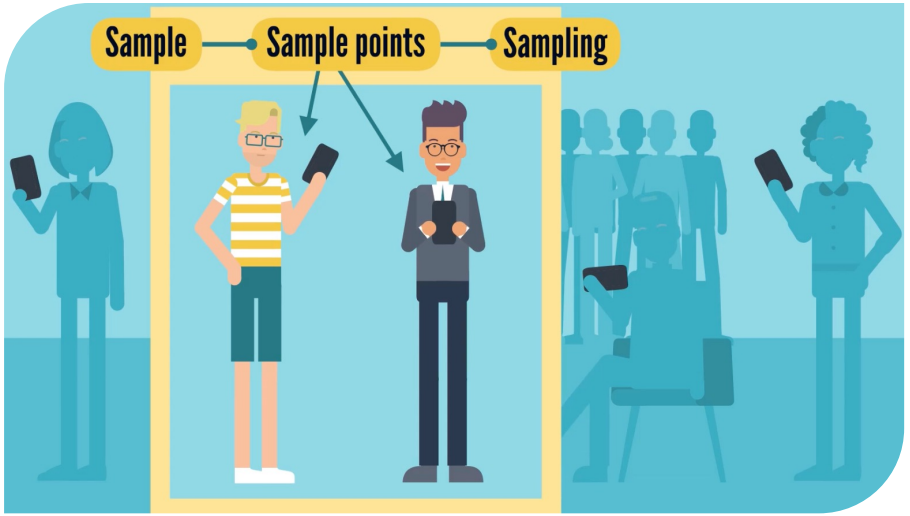
DEFINING A POPULATION

✓ Defining a population for research often presents challenges:

- Surveys typically cannot cover the entire population because not everyone is accessible or willing to engage.
- Furthermore, the exact size of most populations has access nearly impossible.
- Time and cost constraints also make examining every individual in the population unrealistic.

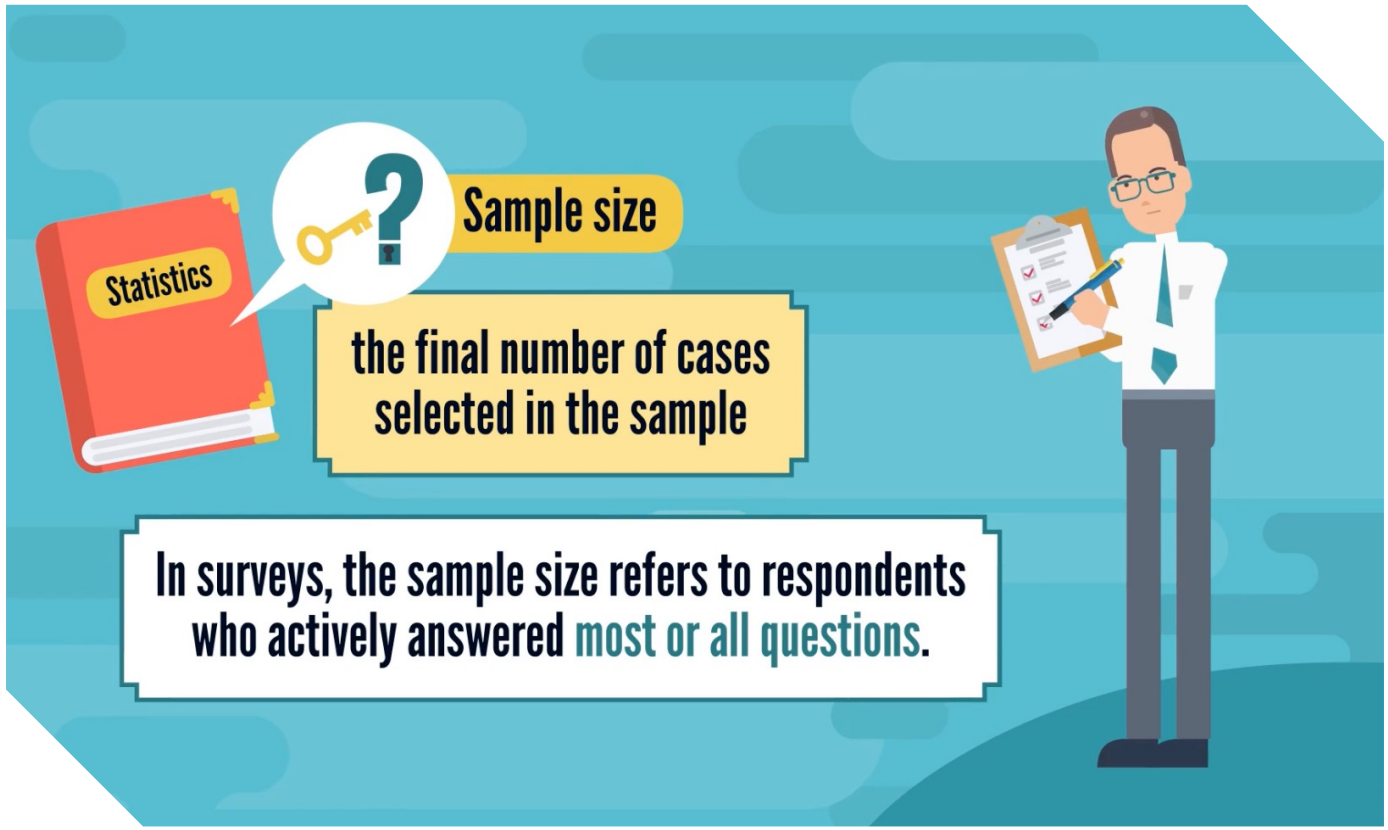
SAMPLE & SAMPLING

- ✓ Since defining a population is challenging, the optimal approach involves collecting data from a population subset:
  - using this smaller group observations to infer about the entire population,
  - this subset is termed a sample.

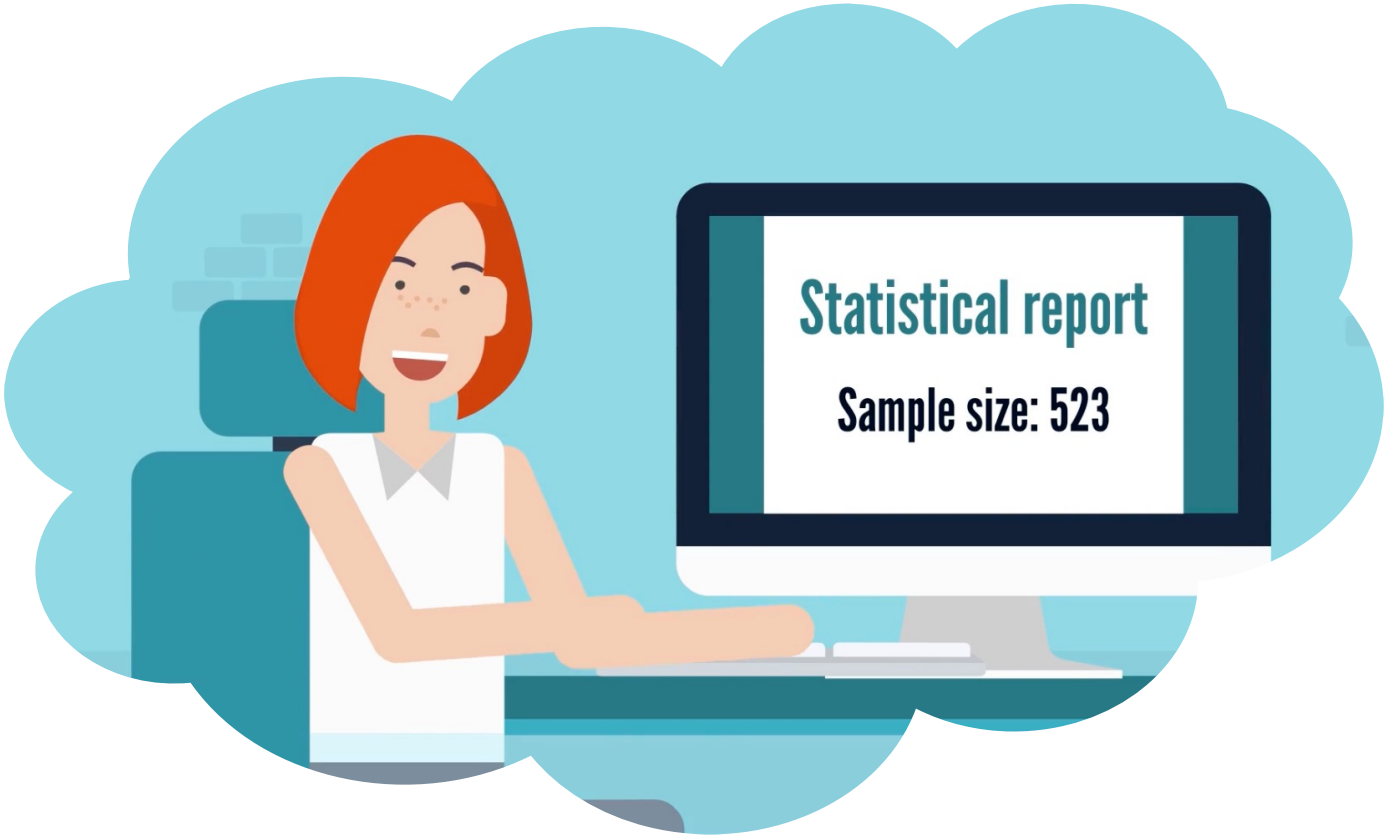


➔The process of choosing this subset from the population is called sampling.

KEY QUESTION IN DATA SAMPLING



SAMPLE SIZE OF THE CASE STUDY



STATISTIC & PARAMETER

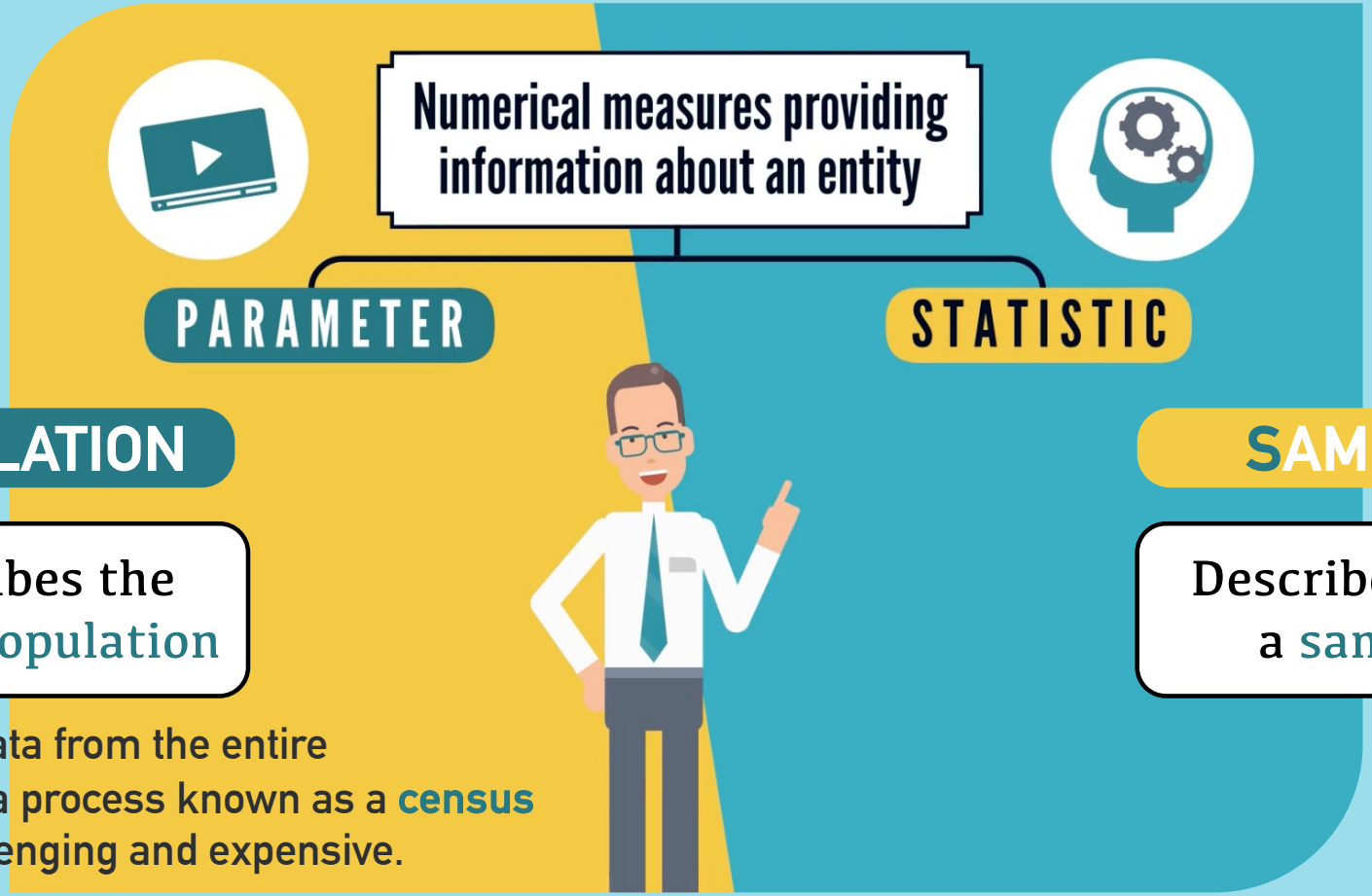
Université de Technologie Tarbes Occitanie Pyrénées

STATISTIC

A single numerical value

- ✓ offering insights into a specific aspect of data distribution,
- ✓ or summarizing a distinct sample trait (characteristic).

STATISTIC VS PARAMETER



LEVELS OF MEASUREMENT

Université de Technologie Tarbes Occitanie Pyrénées

LEVELS OF MEASUREMENT

- ✓ A vital aspect of quantitative research is the distinction between data types.
- ✓ This requires understanding levels of measurement or scales of measure.
  - The importance of these levels lies in their direct influence on choosing the appropriate statistical test for a specific problem.
- ✓ Statistics categorize measurements into four levels: nominal, ordinal, interval, and ratio scales.

NOMINAL DATA

Nominal data divides variables into mutually exclusive, labeled categories.

Examples

Eye color

Blue Brown Green

Smartphone

iPhone Samsung Moto

Transport

Bus Train Car

How is nominal data analyzed?

Descriptive statistics: Frequency distribution and mode

Non-parametric statistical tests

Université de Technologie Tarbes Occitanie Pyrénées

ORDINAL DATA

Ordinal data classifies variables into categories which have a natural order or rank.

Examples

School grades

Education level

Seniority level

How is ordinal data analyzed?

Descriptive statistics: Frequency distribution, mode, median, and range

Non-parametric statistical tests

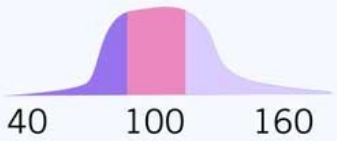
INTERVAL DATA

Interval data is measured along a numerical scale that has equal intervals between adjacent values.

Examples

Temperature

IQ score



Income ranges

How is interval data analyzed?

Descriptive statistics: Frequency distribution; mode, median, and mean; range, standard deviation, and variance

Parametric statistical tests (e.g. t-test, linear regression)

Université de Technologie Tarbes Occitanie Pyrénées

RATIO DATA

Ratio data is measured along a numerical scale that has equal distances between adjacent values, and a true zero.

Examples

Weight in KG

...50 70 90...

Number of staff

...10 30 50...

Income in USD

...20k 40k 60k...

How is ratio data analyzed?

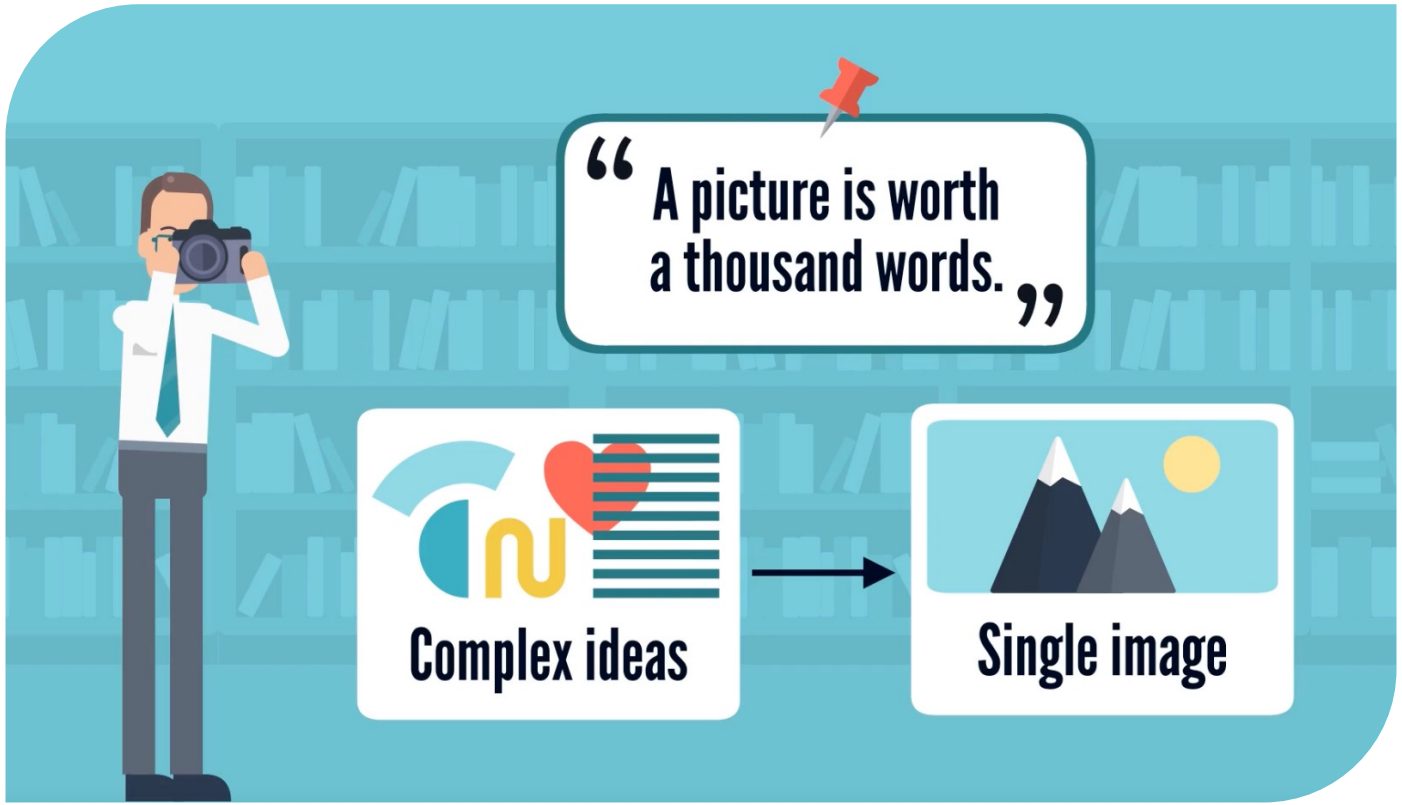
Descriptive statistics: Frequency distribution; mode, median, and mean; range, standard deviation, variance, and coefficient of variation

Parametric statistical tests (e.g. ANOVA, linear regression)

CONTINGENCY TABLE

Université de Technologie Tarbes Occitanie Pyrénées

VISUALIZATION IN THE ANALYTICAL PROCESS



A PICTURE IS WORTH A THOUSAND WORDS

- ✓ A contingency table, also called *frequency table*, *two way or cross-tabulation table*, effectively displays discrete or categorical data.
  - It represents each category alongside the count of data points in that category, each table entry indicating the frequency of occurrences within specific categories.

	Extremely likely	Not at all likely	Not so likely	Somewhat likely	Very likely	Total
Extremely unique	12	10	10	13	25	70
Not at all unique	5	11	5	4	8	33
Not so unique	7	9	10	8	16	50
Somewhat unique	15	16	15	16	28	90
Very unique	52	30	30	64	104	280
Total	91	76	70	105	181	523

A PICTURE IS WORTH A THOUSAND WORDS

- ✓ A relative frequency table shows the popularity or mode of a specific data type based on the population sampled.

	Extremely likely	Not at all likely	Not so likely	Somewhat likely	Very likely	Total
Extremely unique	0.023	0.019	0.019	0.025	0.048	0.134
Not at all unique	0.010	0.021	0.010	0.008	0.015	0.063
Not so unique	0.013	0.017	0.019	0.015	0.031	0.096
Somewhat unique	0.029	0.031	0.029	0.031	0.054	0.172
Very unique	0.099	0.057	0.057	0.122	0.199	0.535
Total	0.174	0.145	0.134	0.201	0.346	1.000

- To convert a standard frequency table showing counts into a relative frequency table displaying proportions, divide each absolute frequency by the total number of data points.
- Multiplying it by 100 gives you the percentage for each category.
- *Examples: 70/523 = 0.134, 12/523 = 0.023*

HYPOTHESIS TESTING

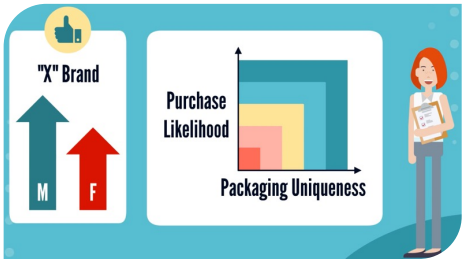
Université de Technologie Tarbes Occitanie Pyrénées

HYPOTHESIS TESTING

- ✓ Hypothesis testing is a fundamental statistical technique used to make inferences about a population based on a sample of data.
  - It's a powerful tool, allowing us to (1) make informed decisions based on empirical evidence, (2) and draw conclusions about population parameters.
  - It is also a tool for researchers and scientists to establish new theories.

EXPRESSING A HYPOTHESIS

- ✓ A hypothesis is a statement interpreting the question or problem at hand, likely a theory in science.
- ✓ Examples
  - *Men find brand X more likable than women.*
  - *The product's packaging uniqueness impacts the likelihood of customer purchase (case study).*

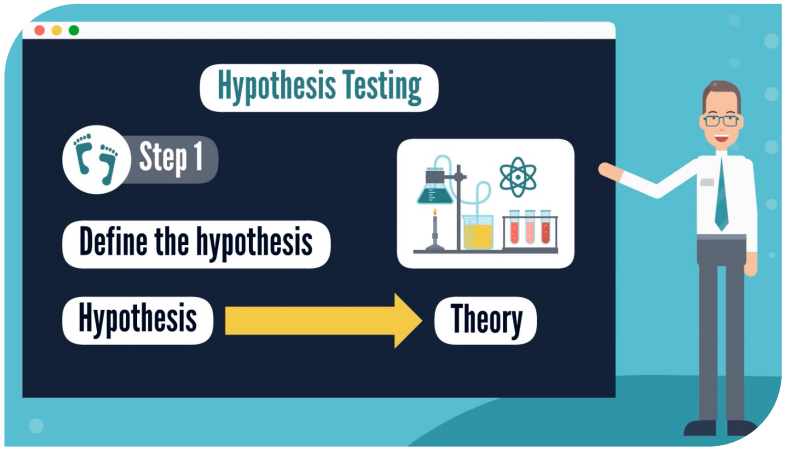


NULL & ALTERNATIVE HYPOTHESES

- ✓ It is essential to propose two contrasting and mutually exclusive statements: the null and alternative hypotheses, of which only one can be correct.
- ✓ The null hypothesis represents the commonly accepted fact or assumption that there is no significant effect or change in the population.

- Null Hypothesis, HO is expressed as a hypothesis of no difference or status quo.
- It signifies the default position and serves as a baseline assertion that the variables or groups being studied do not show any statistically significant variation or impact.

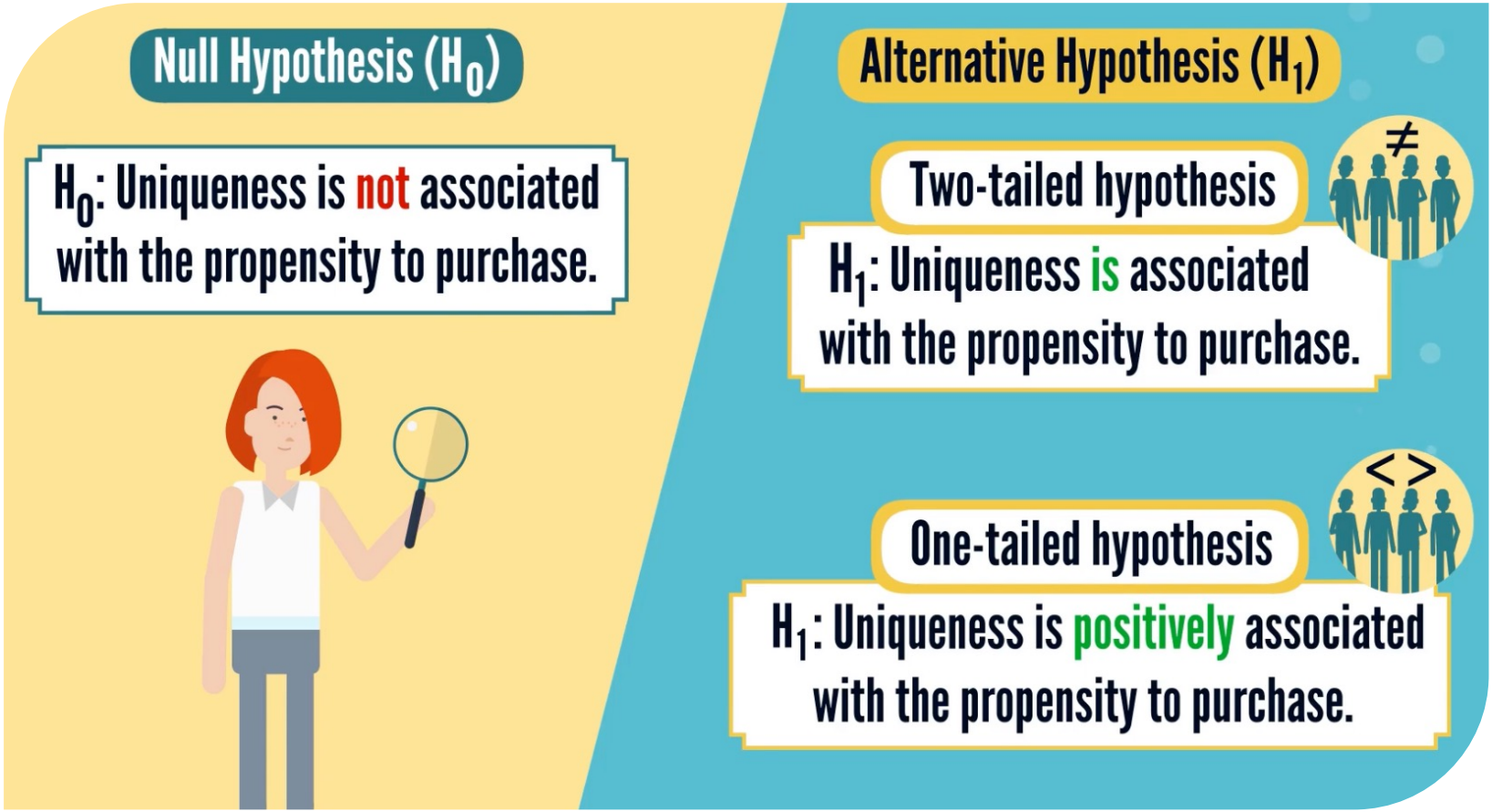
NULL & ALTERNATIVE HYPOTHESES



- ✓ In many scenarios, the aim is to explore beyond the status quo.
- ✓ This involves testing whether a particular action or variable has a significant effect, represented by the alternative hypothesis (H1).

- ○ H0 often serves as a starting point against which H1 is tested to gain new insights or verify the impact of specific interventions.
- ○ H0 can be thought of as the hypothesis that you'll probably want to reject.
- ○ H1 is the focus of your investigation, representing the phenomenon you aim to demonstrate or gather evidence for.

EXAMPLES


Null Hypothesis (H_0)
H_0: Uniqueness is not associated with the propensity to purchase.
Alternative Hypothesis (H_1)
Two-tailed hypothesis
H_1: Uniqueness is associated with the propensity to purchase.
One-tailed hypothesis
H_1: Uniqueness is positively associated with the propensity to purchase.

TEST STATISTIC

TEST STATISTIC

- ✓ A test statistic is the critical value used for deciding a hypothesis,
- ✓ *Quantifying the difference between the sample data and the value predicted by the null hypothesis.*
  - It assesses whether the observed effect is significant enough to support the alternative hypothesis,
  - Providing a basis for potentially rejecting the null hypothesis.
- ✓ The choice of a test statistic depends on the data's nature and the specific research question.

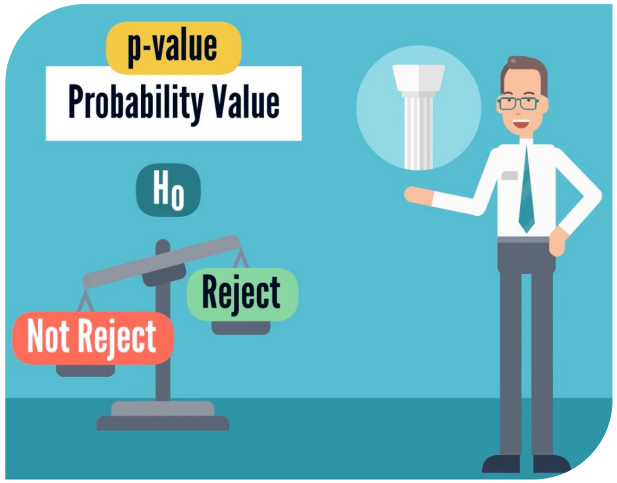
EXAMPLES OF TEST STATISTIC

- ✓ The Chi-Square test:
  - Dependent variable ➔ categorical
  - Independent variables➔ categorical
- ✓ Used for comparing the distribution of categorical variables

  - For instance, compare the distributions of “satisfied” and “dissatisfied” people (uniqueness packaging case study)
- ✓ More examples (upcoming)
  - T-test, Mann Whitney, Wilcoxon, etc.

P-VALUE

P-VALUE

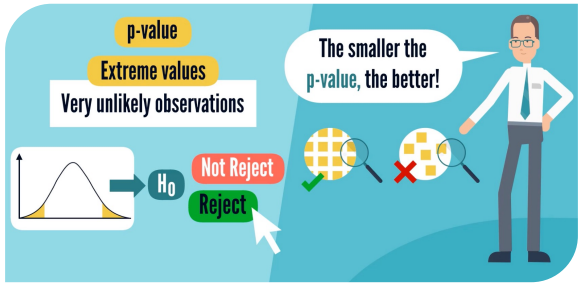


- ✓ The p-value, or probability value is a fundamental concept in statistical hypothesis testing.
  - It quantifies the strength of evidence against the null hypothesis using observed data,
  - Offering analysts the crucial information needed to support or reject it.

➔ It's called the probability value because it measures the probability that an observed result could have occurred just by chance, instead of a specific pattern.

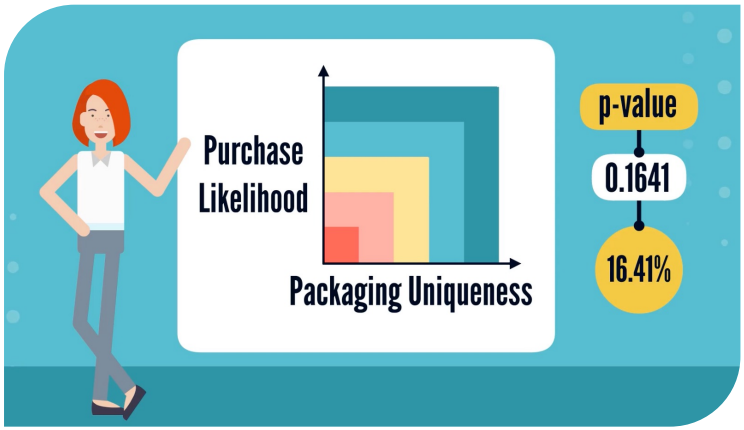
INTERPRETING THE P-VALUE

- ✓ The p-value indicates the rare outcomes
  - *"probability of obtaining a test statistic as extreme as the observed value, assuming the null hypothesis is true".*



- ✓ A small p-value indicates that the observed result is significantly different from what the null hypothesis predicts,
- ✓ Suggesting that the null hypothesis is less likely to be true.

THE P-VALUE OF THE CASE STUDY



- ✓ Assuming that the null hypothesis is true *("no relationship between uniqueness and sales")*
- ✓ There is a 16.41% probability that the observed sample data results are rare, occurred by chance.

- ✓ One might consider 16.41% a relatively low likelihood:
  - Suggesting the null hypothesis be rejected (implying a non-random relationship between the two variables).
  - Yet Kelly did the opposite and accepted the null hypothesis that the two variables are unrelated.

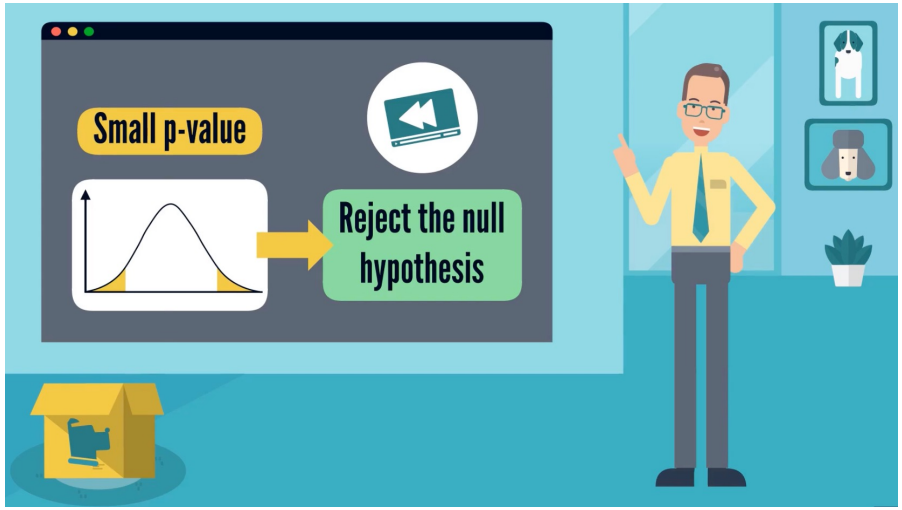
What motivated this decision?

STATISTICAL SIGNIFICANCE

Université de Technologie Tarbes Occitanie Pyrénées

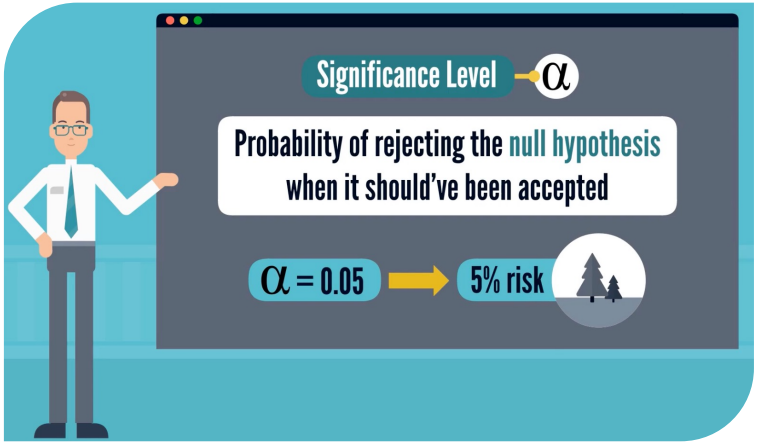
WHAT MOTIVATES THE FINAL DECISION

- ✓ A probability of 16.41% was not small enough for Kelly.
  - This raises the question of *"how small is small enough"* to establish the permissible threshold.
  - Analysts initially define a cutoff value, known as the significance level.



➔ It's called the probability value because it measures the probability that an observed result could have occurred just by chance, instead of a specific pattern.

SIGNIFICANCE LEVEL ALPHA

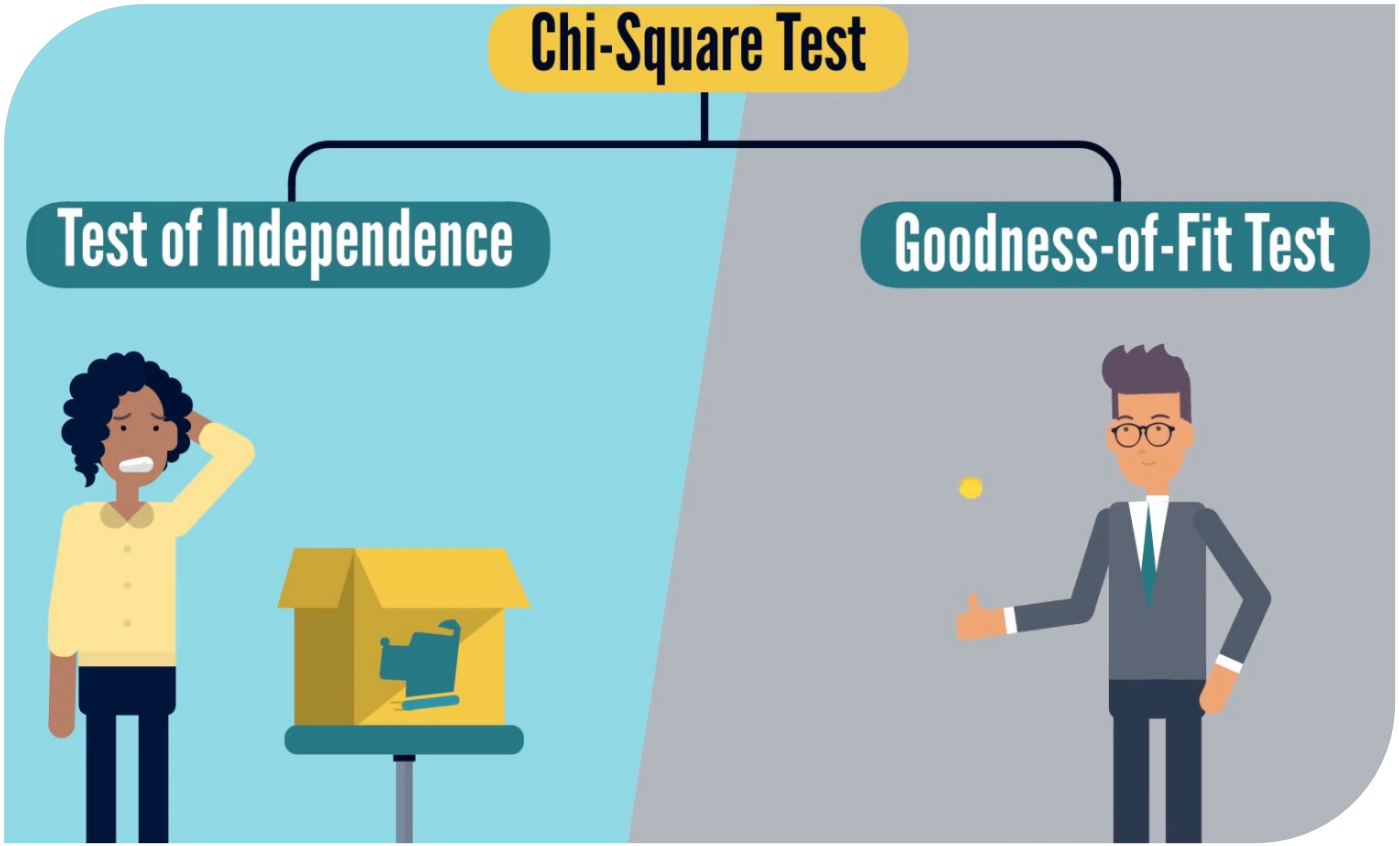


✓ The significance level, generally denoted α, is a number stated in advance (arbitrarily) to determine how small the p-value must be to reject the null hypothesis.

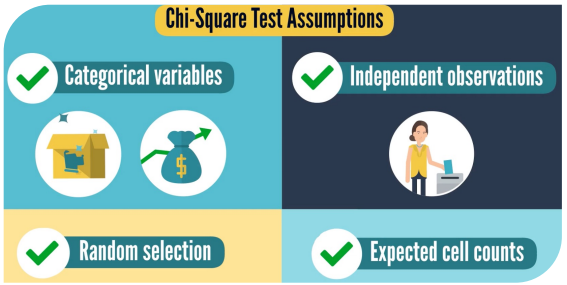
- It corresponds with the probability of rejecting the null hypothesis when it should have been accepted.
  - For example, $\alpha = 0.05$ indicates a 5% risk of concluding that a difference exists when there is no actual difference.
- Unlike the p-value, α is independent of the underlying hypotheses and is not obtained from observational data.

CHI-SQUARE TEST

PURPOSE OF CHI-SQUARE TEST



ASSUMPTIONS OF CHI-SQUARE TEST

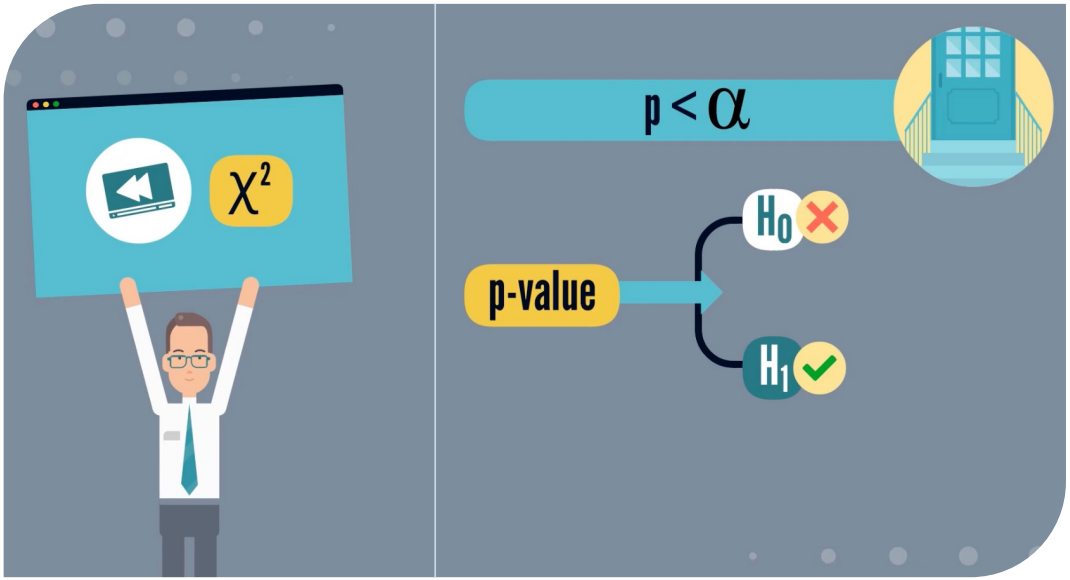


- ✓ **Categorical variables**:
  - Uniqueness and propensity to purchase Kelly's variables are ordinal ➔categorical.
- ✓ **Independent observations**:
  - Respondents completed the same survey only once, with multiple choice questions, allowing a single response, ensuring data independence.
- ✓ **Random selection**:
  - The respondents were randomly selected fulfilling this criterion.
- ✓ **Expected cell counts**:
  - All expected count cells are **5** or more meeting this condition.

DECISION RULE

Use either:

- Chi-square statistics ➔ degree,
- Or p-value ➔ significance level



SUMMARY

KELLY'S RESEARCH

OBSERVED	Extremely likely	Not at all likely	Not so likely	Somewhat lekely	Very likely	ROW TOTALS
Extremely likely	12,00	10,00	10,00	13,00	25,00	70,00
Not at all likely	5,00	11,00	5,00	4,00	8,00	33,00
Not so likely	7,00	9,00	10,00	8,00	16,00	50,00
Somewhat lekely	15,00	16,00	15,00	16,00	28,00	90,00
Very likely	52,00	30,00	30,00	64,00	104,00	280,00
COLUMN TOTALS	91,00	76,00	70,00	105,00	181,00	**523,00**

✓ Observed values
(actual outcomes)

EXPECTED	Extremely likely	Not at all likely	Not so likely	Somewhat lekely	Very likely	ROW TOTALS
Extremely likely	12,18	10,17	9,37	14,05	24,23	70,00
Not at all likely	5,74	4,80	4,42	6,63	11,42	33,00
Not so likely	8,70	7,27	6,69	10,04	17,30	50,00
Somewhat lekely	15,66	13,08	12,05	18,07	31,15	90,00
Very likely	48,72	40,69	37,48	56,21	96,90	280,00
COLUMN TOTALS	91,00	76,00	70,00	105,00	181,00	**523,00**

✓ Expected values
(assuming the null hypothesis)

Each cell =
(rows total * column totals)
/ overall total

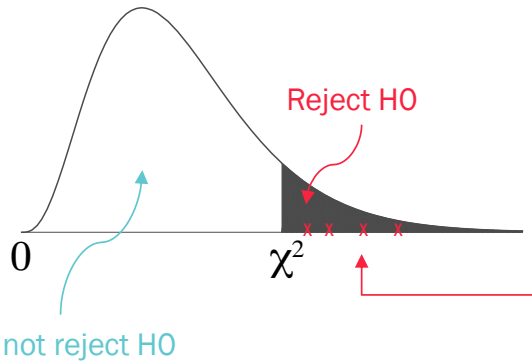
KEELY'S STUDY CHI-SQUARE TEST

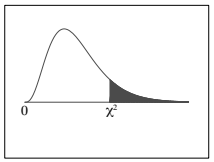
	Extremely likely	Not at all likely	Not so likely	Somewhat lekely	Very likely	ROW TOTALS
Extremely likely	0,00	0,00	0,04	0,08	0,02	0,15
Not at all likely	0,10	8,03	0,08	1,04	1,02	10,27
Not so likely	0,33	0,41	1,64	0,41	0,10	2,89
Somewhat lekely	0,03	0,65	0,72	0,24	0,32	1,96
Very likely	0,22	2,81	1,49	1,08	0,52	6,12
COLUMN TOTALS	0,68	11,91	3,97	2,85	1,99	**21,39**

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where

$$\begin{cases} O \text{ observed value} \\ E \text{ expected value} \end{cases}$$

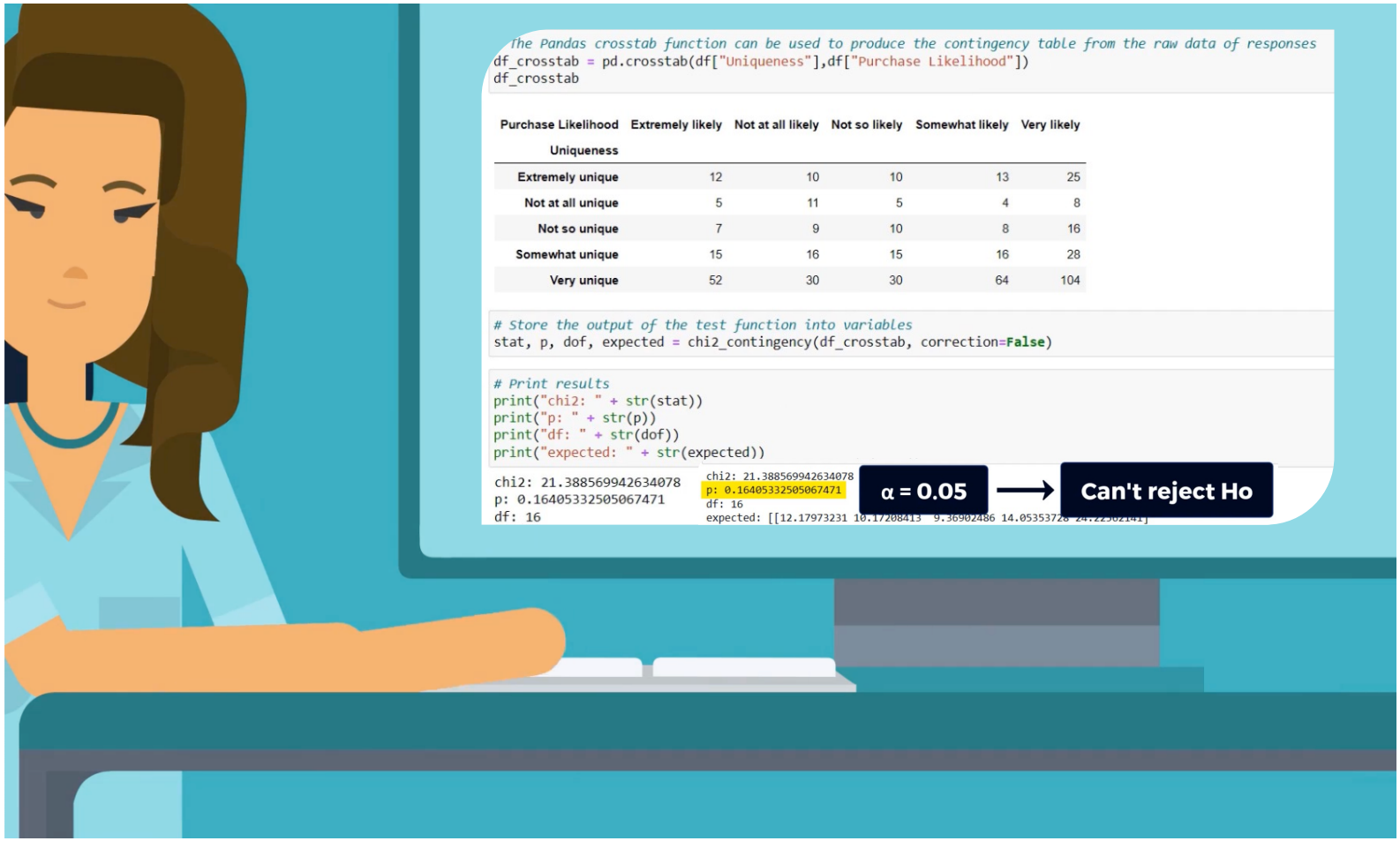




The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$

10% 5% 2.5% 1%

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169


The Pandas crosstab function can be used to produce the contingency table from the raw data of responses
df_crosstab = pd.crosstab(df["Uniqueness"],df["Purchase Likelihood"])
df_crosstab
Purchase Likelihood | Extremely likely | Not at all likely | Not so likely | Somewhat likely | Very likely
Uniqueness
Extremely unique | 12 | 10 | 10 | 13 | 25
Not at all unique | 5 | 11 | 5 | 4 | 8
Not so unique | 7 | 9 | 10 | 8 | 16
Somewhat unique | 15 | 16 | 15 | 16 | 28
Very unique | 52 | 30 | 30 | 64 | 104
Store the output of the test function into variables
stat, p, dof, expected = chi2_contingency(df_crosstab, correction=False)
# Print results
print("chi2: " + str(stat))
print("p: " + str(p))
print("df: " + str(dof))
print("expected: " + str(expected))
chi2: 21.388569942634078
p: 0.16405332505067471
df: 16
chi2: 21.388569942634078
p: 0.16405332505067471
df: 16
α = 0.05
Can't reject Ho