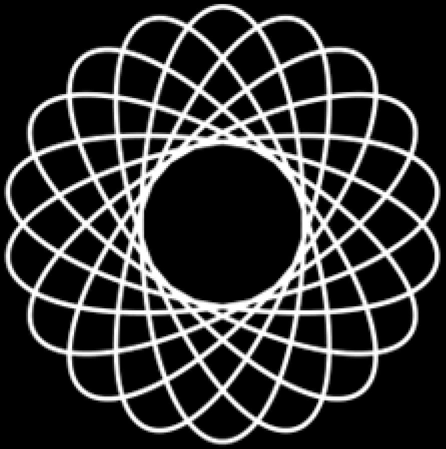
DATA SCIENCE



DATA SCIENCE

★ Continued Lab Sessions ★

Hypothesis Tests

HYPOTHESIS TESTING

Hypothesis testing is a widely used data analytics technique to assess effectiveness and impact of business decisions and to provide confidence for future business decisions

In the bank telemarketing case, using the data that is available, the bank can test multiple hypothesis and report results with confidence that can help it guide strategy

Let’s start with single sample tests:

Is an observed sample mean statistically significantly different from an expected population mean?

HYPOTHESIS TESTING

Supposing we take a random sample of 100 customers from the underlying data because we want to call customers for a short survey, and we don’t want to call every customer

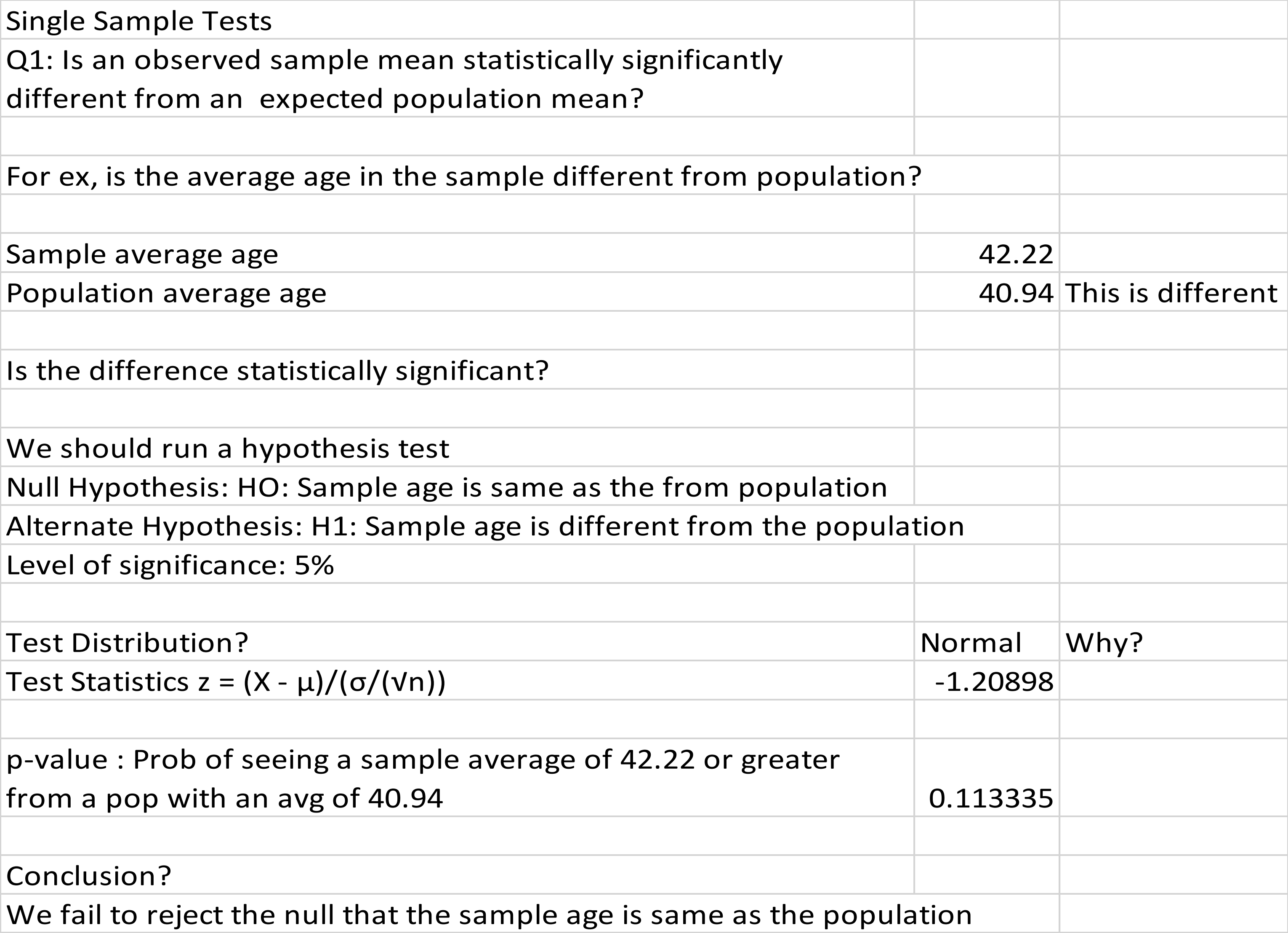
First thing we should establish after taking a random sample is that it is representative

Let’s start with one attribute – age

Once we choose a random sample of 100 customers, compute sample average age

Is the sample average age the same as the population? If not, how much is the difference? And if there is a difference, should you worry? (That is, is the difference statistically significant?)

HYPOTHESIS TESTING



Single Sample Tests		
Q1: Is an observed sample mean statistically significantly different from an expected population mean?		
For ex, is the average age in the sample different from population?		
Sample average age	42.22	
Population average age	40.94	This is different
Is the difference statistically significant?		
We should run a hypothesis test		
Null Hypothesis: HO: Sample age is same as the from population		
Alternate Hypothesis: H1: Sample age is different from the population		
Level of significance: 5%		
Test Distribution?	Normal	Why?
Test Statistics z = (X - μ)/(σ/(√n))	-1.20898	
p-value : Prob of seeing a sample average of 42.22 or greater from a pop with an avg of 40.94	0.113335	
Conclusion?		
We fail to reject the null that the sample age is same as the population		

HYPOTHESIS TESTING

Small sample tests:

Now imagine another scenario where a random sample of 25 customers was chosen for an in depth interview and focus group process

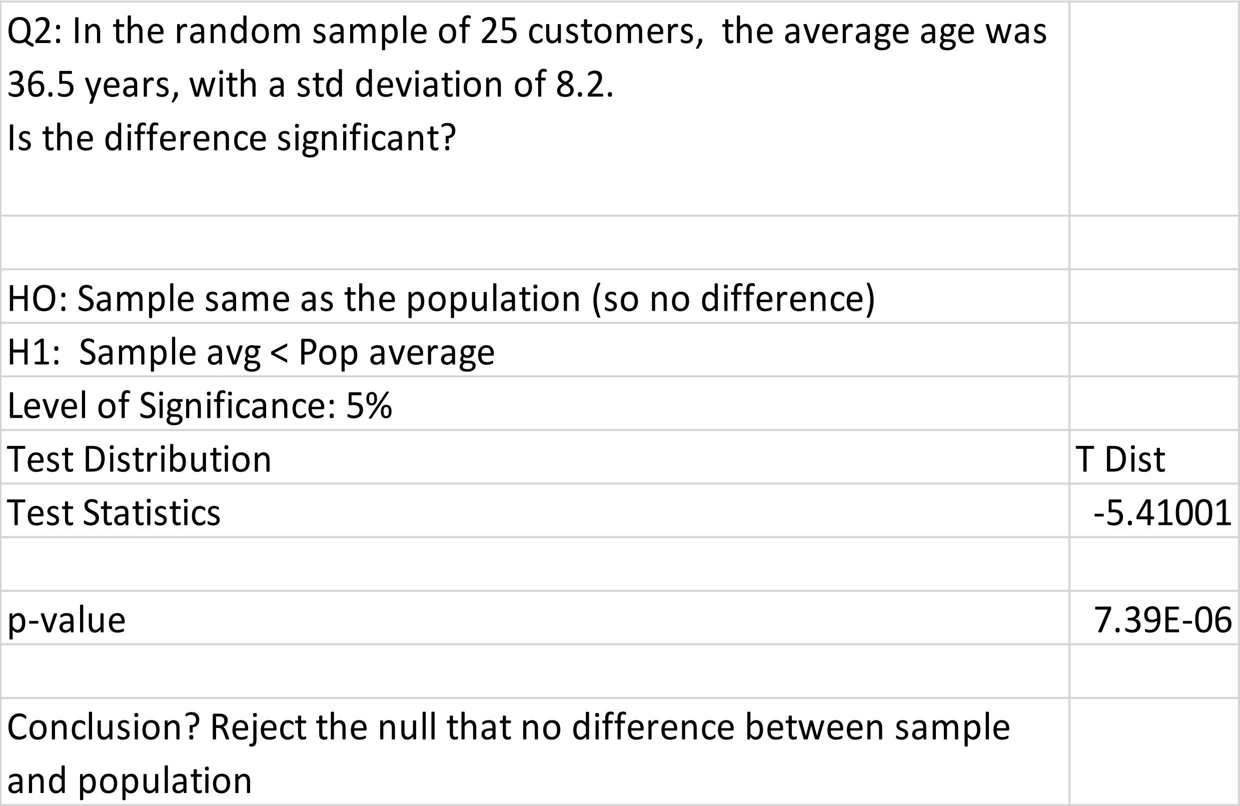
In the random sample of 25 customers, the average age was 39.5 years, with a standard deviation of 8.2.

Is the difference significant?

HYPOTHESIS TESTING

In the random sample of 25 customers, the average age was 39.5 years, with a standard deviation of 8.2.

Is the difference significant?



Q2: In the random sample of 25 customers, the average age was 36.5 years, with a std deviation of 8.2. Is the difference significant?	
HO: Sample same as the population (so no difference)	
H1: Sample avg < Pop average	
Level of Significance: 5%	
Test Distribution	T Dist
Test Statistics	-5.41001
p-value	7.39E-06
Conclusion? Reject the null that no difference between sample and population	

HYPOTHESIS TESTING

Two Sample Tests

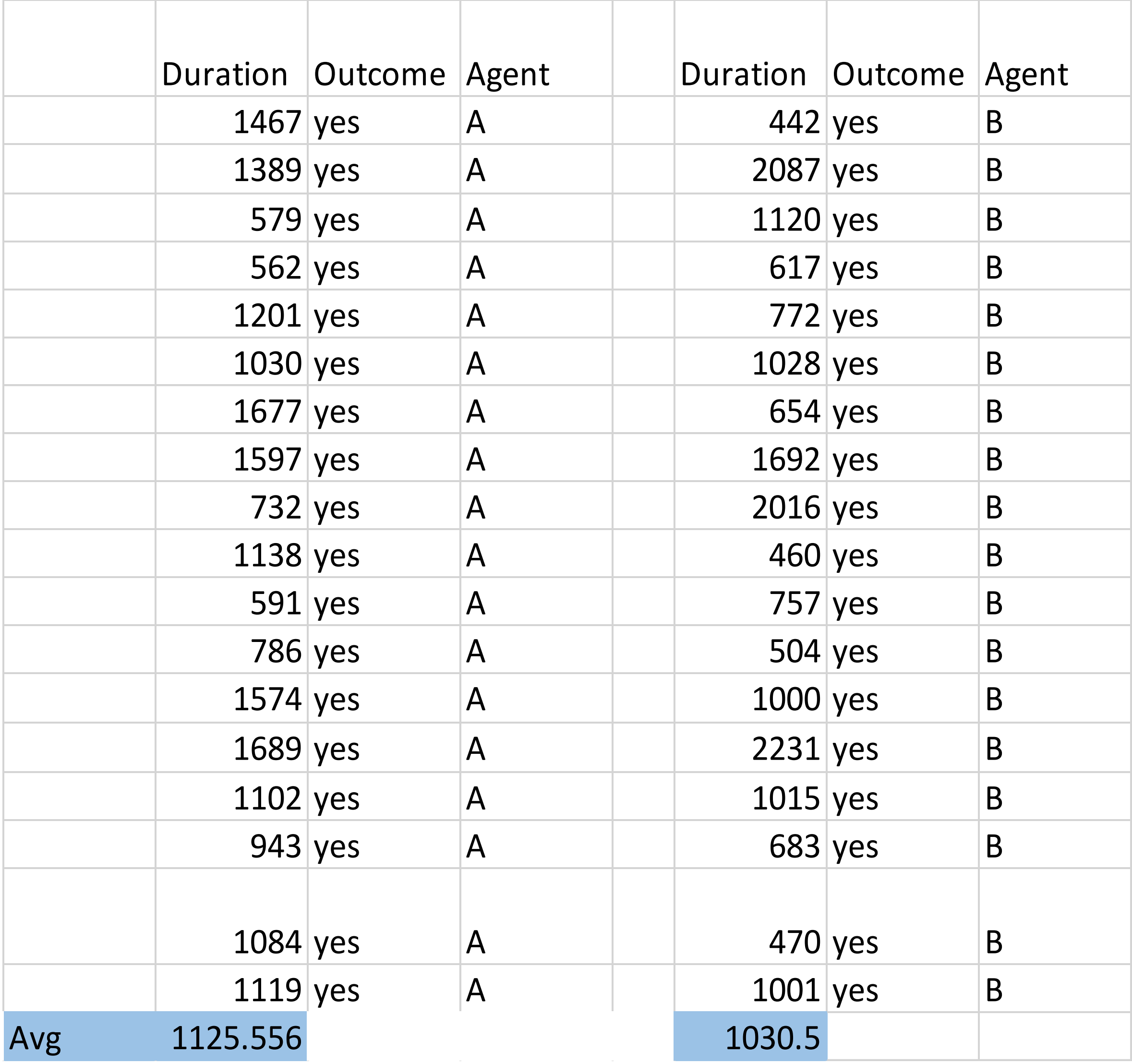
Supposing the company wants to test effectiveness of agents, by checking average time spent on the phone with a customer that results in a yes

They choose two agents, and record their calls, and choose 17 calls each that resulted in a conversion

For Agent A: average call duration is: 1125 secs
For Agent B: average call duration is: 1030 secs

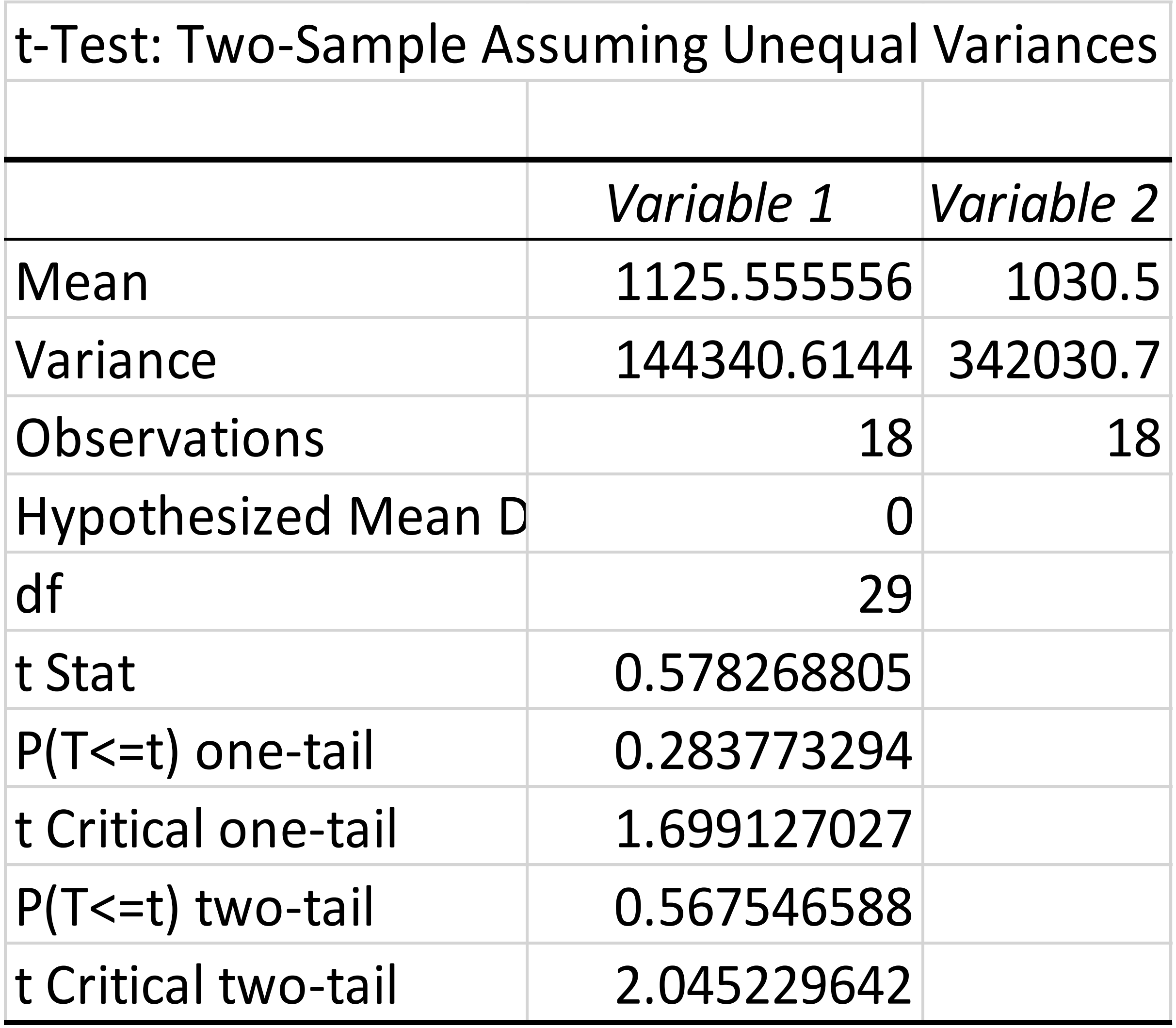
Can it be concluded that Agent B is more efficient? Or is the variation simply random chance variation?

HYPOTHESIS TESTING



	Duration	Outcome	Agent		Duration	Outcome	Agent
	1467	yes	A		442	yes	B
	1389	yes	A		2087	yes	B
	579	yes	A		1120	yes	B
	562	yes	A		617	yes	B
	1201	yes	A		772	yes	B
	1030	yes	A		1028	yes	B
	1677	yes	A		654	yes	B
	1597	yes	A		1692	yes	B
	732	yes	A		2016	yes	B
	1138	yes	A		460	yes	B
	591	yes	A		757	yes	B
	786	yes	A		504	yes	B
	1574	yes	A		1000	yes	B
	1689	yes	A		2231	yes	B
	1102	yes	A		1015	yes	B
	943	yes	A		683	yes	B
	1084	yes	A		470	yes	B
	1119	yes	A		1001	yes	B
Avg	1125.556				1030.5		

HYPOTHESIS TESTING



t-Test: Two-Sample Assuming Unequal Variances

	Variable 1	*Variable 2*
Mean	1125.555556	1030.5
Variance	144340.6144	342030.7
Observations	18	18
Hypothesized Mean D	0	
df	29	
t Stat	0.578268805	
P(T<=t) one-tail	0.283773294	
t Critical one-tail	1.699127027	
P(T<=t) two-tail	0.567546588	
t Critical two-tail	2.045229642	