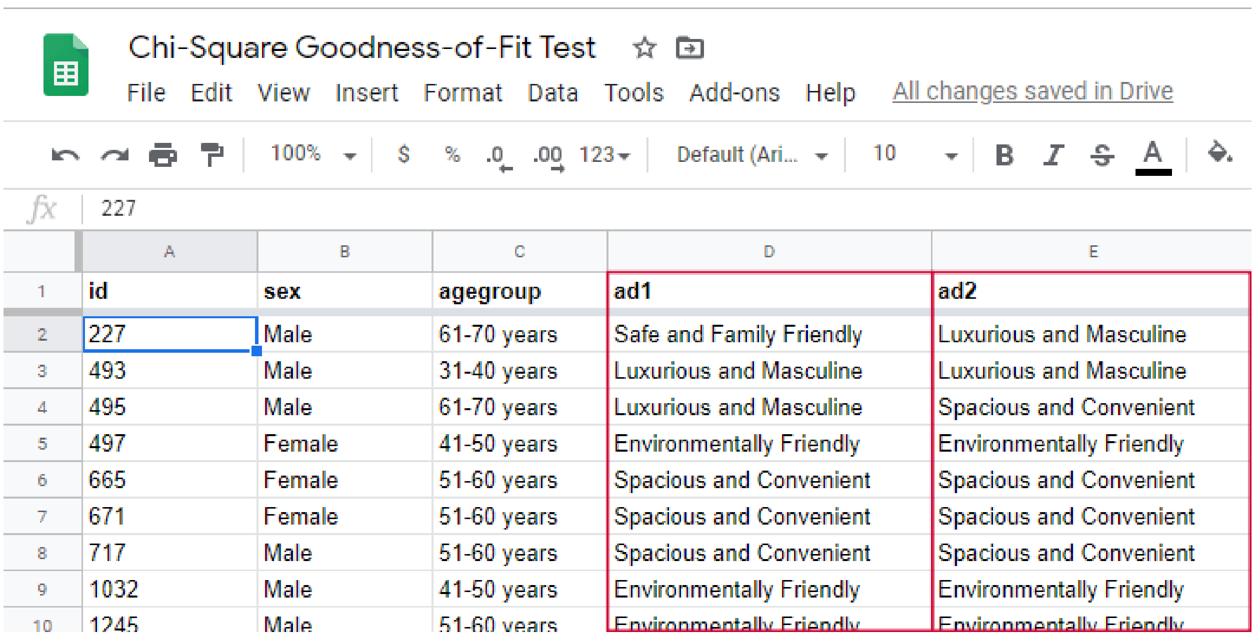
A chi-square goodness-of-fit test examines if a categorical variable has some hypothesized frequency distribution in some population.

Example - Testing Car Advertisements

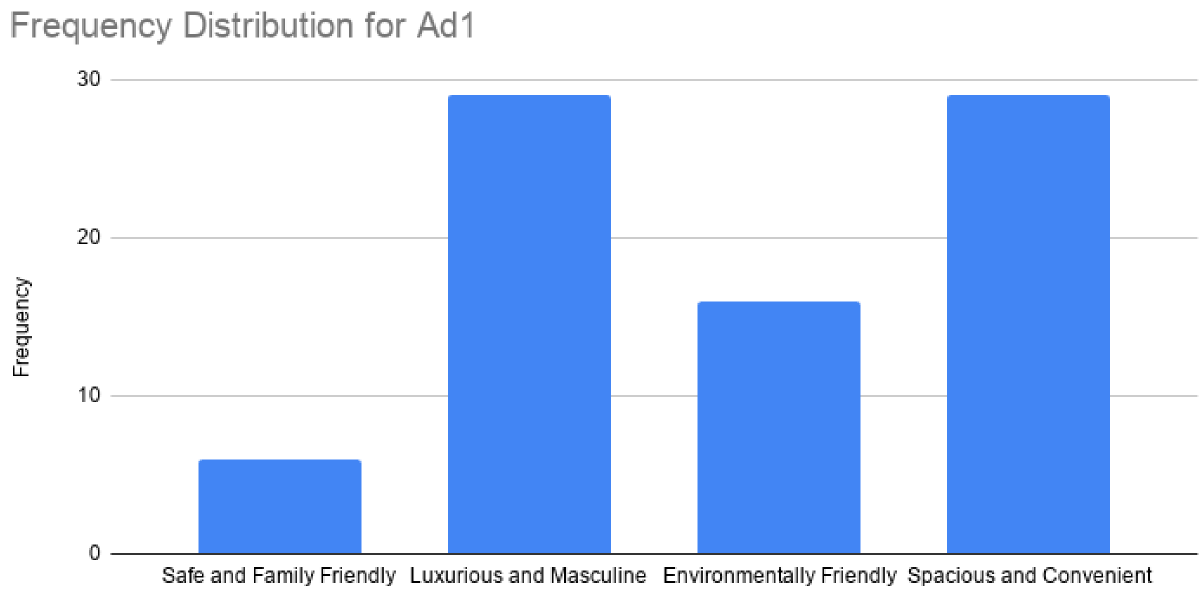
A car manufacturer wants to launch a campaign for a new car. They'll show advertisements -or “ads”- in 4 different sizes. For ad each size, they have 4 ads that try to convey some message such as “this car is environmentally friendly”. They then asked N = 80 people which ad they liked most. The data thus obtained are in this Googlesheet, partly shown below.


Chi-Square Goodness-of-Fit Test

	A	B	C	D	E
1	id	sex	agegroup	ad1	ad2
2	227	Male	61-70 years	Safe and Family Friendly	Luxurious and Masculine
3	493	Male	31-40 years	Luxurious and Masculine	Luxurious and Masculine
4	495	Male	61-70 years	Luxurious and Masculine	Spacious and Convenient
5	497	Female	41-50 years	Environmentally Friendly	Environmentally Friendly
6	665	Female	51-60 years	Spacious and Convenient	Spacious and Convenient
7	671	Female	51-60 years	Spacious and Convenient	Spacious and Convenient
8	717	Male	51-60 years	Spacious and Convenient	Spacious and Convenient
9	1032	Male	41-50 years	Environmentally Friendly	Environmentally Friendly
10	1245	Male	51-60 years	Environmentally Friendly	Environmentally Friendly

So which ads performed best in our sample? Well, we can simply look up which ad was preferred by most respondents: the ad having the highest frequency is the mode for each ad size. So let's have a look at the frequency distribution for the first ad size -ad1- as visualized in the bar chart shown below.

Observed Frequencies and Bar Chart



The **observed frequencies** shown in this chart are

1. Safe and Family Friendly: 6
2. Luxurious and Masculine: 29
3. Environmentally Friendly: 16
4. Spacious and Convenient: 29

Note that ad1 has a bimodal distribution: ads 2 and 4 are both winners with 29 votes. However, our data only hold a sample of N = 80. So**can we conclude that ads 2 and 4 also perform best in the entire population?**The chi-square goodness-of-fit answers just that. And for this example, it does so by trying to reject the null hypothesis that all ads perform equally well in the population.

Null Hypothesis

Generaly, the null hypothesis for a chi-square goodness-of-fit test is simply

$$H_0: P_{01}, P_{02}, \ldots, P_{0m}, \sum_{i=0}^{m} (P_{0i}) = 1$$

H0:P01,P02,...,P0m,∑i=0m(P0i)=1

where P_{0i} denote population proportions for m categories in some categorical variable. You can choose any set of proportions as long as they add up to one. In many cases, all proportions being equal is the most likely null hypothesis.
For a dichotomous variable having only 2 categories, you're better off using

- a binomial test because it gives the exact instead of the approximate significance level or
- a z-test for 1 proportion because it gives a confidence interval for the population proportion.

Anyway, for our example, we'd like to show that some ads perform better than others. So we'll try to refute that our 4 population proportions are all equal and -hence- 0.25.

Expected Frequencies

Now, if the 4 population proportions really are 0.25 and we sample N = 80 respondents, then we expect each ad to be preferred by 0.25 · 80 = 20 respondents. That is,**all 4 expected frequencies are 20.**We need to know these expected frequencies for 2 reasons:

- computing our test statistic requires expected frequencies and
- the assumptions for the chi-square goodness-of-fit test involve expected frequencies as well.

Assumptions

The chi-square goodness-of-fit test requires 2 assumptions[2,3]:

1. independent observations;

2. for 2 categories, each expected frequency E_i must be at least 5.
   For 3+ categories, each E_i must be at least 1 and no more than 20% of all E_i may be smaller than 5.

The observations in our data are independent because they are distinct persons who didn't interact while completing our survey. We also saw that all E_i are (0.25 · 80 =) 20 for our example. So this second assumption is met as well.

Formulas

We'll first compute the χ^2 test statistic as

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where

- O_i denotes the observed frequencies and
- E_i denotes the expected frequencies -usually all equal.

For ad1, this results in

$$\chi^2 = \frac{(16-20)^2}{20} + \frac{(29-20)^2}{20} + \frac{(9-20)^2}{20} + \frac{(29-20)^2}{20} = 18.7$$

If all assumptions have been met, χ^2 approximately follows a chi-square distribution with df degrees of freedom where

$$df = m - 1$$

for m frequencies. Since we have 4 frequencies for 4 different ads,

$$df = 4 - 1 = 3$$

for our example data. Finally, we can simply look up the significance level as

$$P(\chi^2(3) > 18.7) \approx 0.00032$$

We ran these calculations in this Googlesheet shown below.

	A	B	C	D	E
1	*ad1*	Observed Frequency	Expected Frequency	Residual	Chi-Square Points
2	Environmentally Friendly	16	20	-4	0.8
3	Luxurious and Masculine	29	20	9	4.05
4	Safe and Family Friendly	6	20	-14	9.8
5	Spacious and Convenient	29	20	9	4.05
6	**Grand Total**	**80**	80	0	18.7
7					
8	Chi-Square Value	18.7			
9	DF	3			
10	P	0.00032			
11	Conclusion	Since p < 0.05, we **reject the null hypothesis** that all 4 ads have equal population pro			
12					

"Ad preferences were distributed *unequally*, $\chi^2(3) = 18.7$, p = 0.000."

So what does this mean? Well, if all 4 ads are equally preferred in the population, there's a 0.00032 chance of finding our observed frequencies. Since $p < 0.05$, we reject the null hypothesis. **Conclusion**: some ads are preferred by more people than others in the entire population of readers.

Right, so it's safe to assume that the population proportions are not all equal. But precisely how different are they? We can express this in a single number: the effect size.

Effect Size - Cohen's W

The effect size for a chi-square goodness-of-fit test -as well as the chi-square independence test- is Cohen's W. Some rules of thumb[1] are that

- Cohen's W = **0.10** indicates a **small** effect size;
- Cohen's W = **0.30** indicates a **medium** effect size;
- Cohen's W = **0.50** indicates a **large** effect size.

Cohen's W is computed as

$$W=\sqrt{\sum_{i=1}^{m}\frac{(P_{oi}-P_{ei})^2}{P_{ei}}}$$

where

- P_{oi}Poi denote observed proportions and
- P_{ei}Pei denote expected proportions under the null hypothesis for
- mm cells.

For ad1, the null hypothesis states that all expected proportions are 0.25. The observed proportions are computed from the observed frequencies (see screenshot below) and result in

$$W=\sqrt{\frac{(0.2-0.25)^2}{0.25}+\frac{(0.3625-0.25)^2}{0.25}+\frac{(0.075-0.25)^2}{0.25}+\frac{(0.3625-0.25)^2}{0.25}}=$$

W=(0.2−0.25)20.25+(0.3625−0.25)20.25+(0.075−0.25)20.25+(0.3625−0.25)20.25=

$$W=\sqrt{0.234}=0.483$$

W=0.234=0.483

We ran these computations in this Googlesheet shown below.

	A	B	C	D	E	F
1	*ad1*	Observed Frequency	Observed Proportion	Expected Proportion	Residual Proportion	W Points
2	Environmentally Friendly	16	0.2	0.25	-0.050	0.010
3	Luxurious and Masculine	29	0.3625	0.25	0.113	0.051
4	Safe and Family Friendly	6	0.075	0.25	-0.175	0.123
5	Spacious and Convenient	29	0.3625	0.25	0.113	0.051
6	**Grand Total**	**80**	1	1	0.000	0.234
7						
8	Sum W points	0.234				
9	Cohen's W	0.483	EFFECT SIZE			
10	Conclusion	This is roughly a **large effect size**: overall, the observed proportions differ quite a lot from the presumed eq				
11						

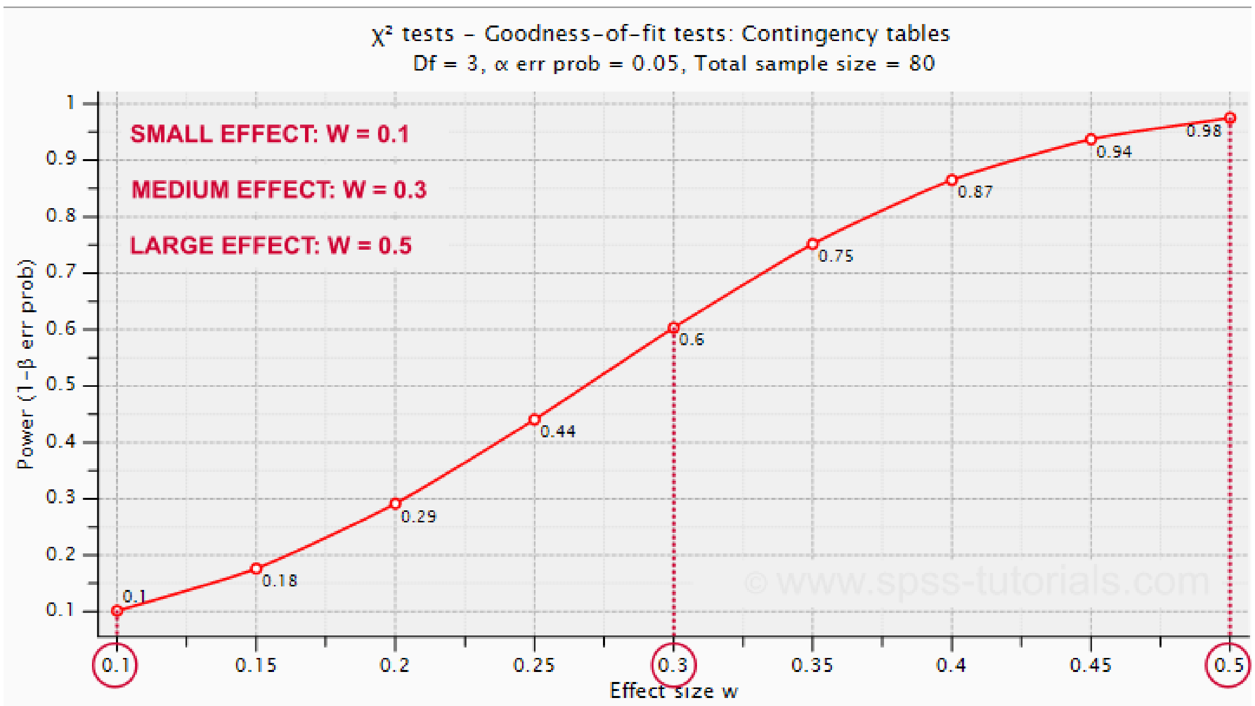
For ad1, the effect size WW = 0.483. This indicates a large overall difference between the observed and expected frequencies.

Power and Sample Size Calculation

Now that we computed our effect size, we're ready for our last 2 steps. First off, what about power? What's the probability demonstrating an effect if

- we test at $\alpha = 0.05$;
- we have a sample of N = 80;
- df = 3 (our outcome variable has 4 categories);
- we don't know the population effect size WW?

The chart below -created in G*Power- answers just that.



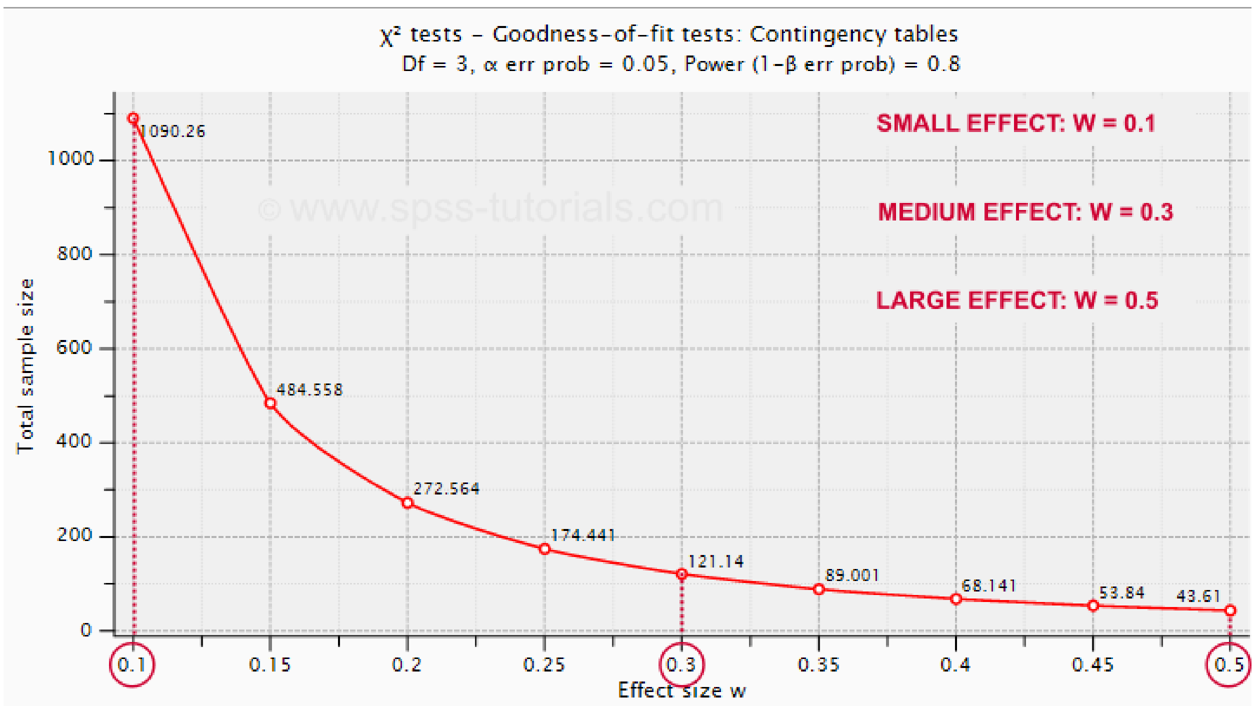
Some basic conclusions are that

- power = **0.98** for a **large** effect size;
- power = **0.60** for a **medium** effect size;
- power = **0.10** for a **small** effect size.

These outcomes are not too great: we only have a 0.60 probability of rejecting the null hypothesis if the population effect size is medium and N = 80. However, we can increase power by increasing the sample size. So **which sample sizes do we need** if

- we test at α = 0.05;
- we want to have power = 0.80;
- df = 3 (our outcome variable has 4 categories);
- we don't know the population effect size WW?

The chart below shows how required sample sizes decrease with increasing effect sizes.



Under the aforementioned conditions, we have power ≥ 0.80

- for a **large** effect size if **N = 44**;
- for a **medium** effect size if **N = 122**;
- for a **small** effect size if **N = 1091**.