test of associations btwn categorical variables. "categorical version of corr coeff"

Comparing Multiple Proportions with a χ^2-test

- As is always the case when comparing proportions is of interst, we assume that our response variable is binary:

$$Y_{ij} = \begin{cases} 1 & \text{if unit } i \text{ in condition } j \text{ performs an action of interest} \\ 0 & \text{if unit } i \text{ in condition } j \text{ does not perform an action of interest} \end{cases}$$

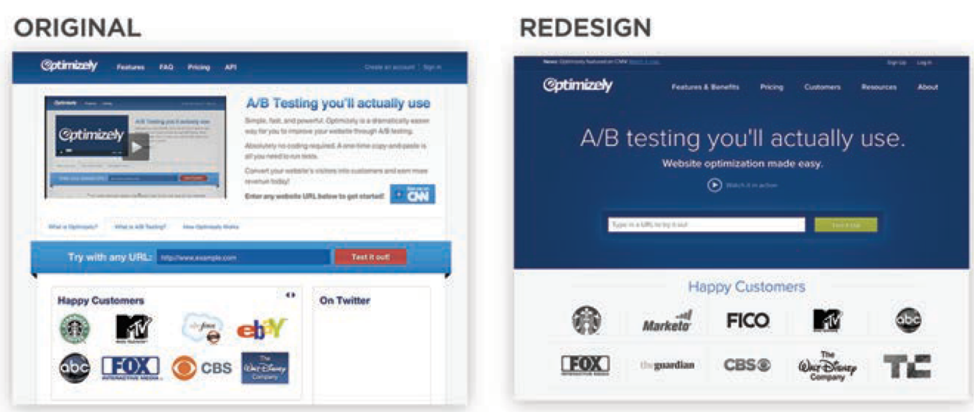
for $i = 1, 2, \ldots, n_j$ and $j = 1, 2, \ldots, m$.

- The "gatekeeper" test for proportions is tested using the **chi-squared test of independence** (also known as Pearson's χ^2-test)

- The chi-squared test of independence is typically used as a test for 'no association' between two categorical variables that are summarized in a *contingency table.*

- We apply this methodology here to test the independence of the binary outcome (whether a unit performs the action of interest) and the particular condition they are in.

- To start, let's assume that $m = 2$, and let's use the Optimizely experiment (previously discussed) as a reference.



		Condition		
		1	2	
Conversion	Yes	280	399	679
	No	8592	8243	16835
		8872	8642	17514

- If $\pi_1 = \pi_2 = \pi$ then we would expect the conversion rate in each condition to be the same.
- An estimate of the pooled conversion rate in this case is $\hat{\pi} = 679/17514 = 0.0388$

- Therefore we would expect $n_1\hat{\pi} = 8872 \times 0.0388 = 343.96$ conversions in condition 1 and $n_2\hat{\pi} = 8642 \times 0.0388 = 335.04$ conversions in condition 2.

- The chi-squared test formally evaluates if the difference between what was observed and what is expected under the null hypothesis is large enough to be considered *significantly* different.

- The *general* 2×2 contingency table for a scenario like this is shown below.

		Condition 1	Condition 2	
Conversion	Yes	$O_{1,1}$	$O_{1,2}$	O_1
	No	$O_{0,1}$	$O_{0,2}$	O_0
		n_1	n_2	$n_1 + n_2$

- So

$$\hat{\pi} = \frac{O_1}{n_1 + n_2} \text{ and } 1 - \hat{\pi} = \frac{O_0}{n_1 + n_2}$$

represent the overall proportions of units that did or did not convert

- Let $E_{1,j}$ and $E_{0,j}$ represent the expected number of conversions and non-conversions in condition $j = 1, 2$

$$E_{1,j} = n_j\hat{\pi} \text{ and } E_{0,j} = n_j(1 - \hat{\pi}).$$

under H_0, $\pi_1 = \pi_2 = \pi$

- The χ^2 test statistic compares the observed count in each cell to the corresponding expected count, and is defined as

$$T = \sum_{l=0}^{1} \sum_{j=1}^{2} \frac{(O_{l,j} - E_{l,j})^2}{E_{l,j}}.$$

- The p-value for this test is calculated by

$$\text{p-value} = P(T \geq t)$$

where $T \sim \chi^2_{(1)}$ $2-1=1$

- Returning to the Optimizely example, the *expected table* is

		Condition 1	Condition 2	
Conversion	Yes	343.96	335.04	679
	No	8528.04	8306.96	16835
		8872	8642	17514

- And the resultant test statistic and p-value are:

- Let's now extend this for $m > 2$
  - We've seen the chi-squared test is a test of 'no association' between the binary outcome (whether a unit performs the action of interest) and the particular condition they are in.
    * But there is no requirement that there be only two conditions.
    * Here we generalize the test to any number of experimental conditions.
  - The information associated with this test can be summarized in a $2 \times m$ contingency table:

		Condition				
		1	2	$\cdots$	m	
Conversion	Yes	$O_{1,1}$	$O_{1,2}$	$\cdots$	$O_{1,m}$	O_1
	No	$O_{0,1}$	$O_{0,2}$	$\cdots$	$O_{0,m}$	O_0
		n_1	n_2	$\cdots$	n_m	$N = \sum_{j=1}^{m} n_j$

  - We compare each of the observed frequencies $O_{l,j}$ with the corresponding expected frequency $E_{l,j}$

$E_{1,j} = n_j \hat{\pi}$, $E_{0,j} = n_j(1-\hat{\pi})$

Z_j is $\frac{O_{l,j} - E_{l,j}}{\sqrt{E_{l,j}}} \xrightarrow{D} N(0, 1-\pi_j)$

  - The χ^2 test statistic compares the observed count in each cell to the corresponding expected count, and is defined as

$$T = \sum_{l=0}^{1} \sum_{j=1}^{m} \frac{(O_{l,j} - E_{l,j})^2}{E_{l,j}}$$

  - The p-value associated with this test is calculated as p-value $= P(T \geq t)$ where $T \sim \chi^2_{(m-1)}$.

But Z_j's are not indep.

$Cov(Z_i Z_j) = -\sqrt{p_i p_j} = E(Z_i Z_j)$

- **Example: Nike SB Ads**
  - Suppose that Nike is running an ad campaign for Nike SB, their skateboarding division, and the campaign involves $m = 5$ different video ads the are being shown in Facebook newsfeeds.
  - A video ad is 'viewed' if it is watched for longer than 3 seconds, and interest lies in determining which ad is most popular and hence most profitable by comparing the viewing rates of the five different videos.
  - Each of the 5 video ads is shown to $n_1 = 5014$, $n_2 = 4971$, $n_3 = 5030$, $n_4 = 5007$, and $n_5 = 4980$ users, and the results are summarized in the table below.

		Condition					
		1	2	3	4	5	
View	Yes	160	95	141	293	197	886
	No	4854	4876	4889	4714	4783	24116
		5014	4971	5030	5007	4980	25002

– The overall watch rate (and its complement) are:

– The expected cell frequencies are found by multiplying n_j by $\hat{\pi}$ and $(1 - \hat{\pi})$, $j = 1, 2, 3, 4, 5$

		Condition					
		1	2	3	4	5	
View	Yes	177.68	176.16	178.25	177.43	176.48	886
	No	4836.32	4794.84	4851.75	4829.57	4803.52	24116
		5014	4971	5030	5007	4980	25002

– The resultant test statistic and p-value are:

The Multiple Comparison Problem

- We have seen that "gatekeeper" tests of overall equality such as

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_m \text{ versus } H_A : \theta_j \neq \theta_k \text{ for some } j \neq k$$

  are often rejected.

- We may follow this up with a series of pairwise comparisons to determine which condition(s) is (are) optimal.
  - We already know how to do this!

- HOWEVER, when doing multiple comparisons like this, we encounter the **multiple comparison** or **multiple testing problem**

Type I errors are more likely to occur in a family of tests than an individual test

- To frame this discussion, let's define some notation:
  - M: the number of hypothesis tests
  - M_0: the number true null hypotheses
  - M_A: the number false null hypotheses
  - R: the number of null hypotheses we reject
  - $M - R$: the number of null hypotheses we do not reject
  - V: the number of true null hypotheses that we incorrectly reject # TI errors
  - S: the number of false null hypotheses that we correctly reject
  - U: the number of true null hypotheses that we correctly do not reject
  - T: the number of false null hypotheses that we incorrectly do not reject # T2 errors

- The outcomes of these M decisions are nicely summarized in the following table:

unobservable randoms

unknown constants

		Decision		
		Reject H_0	Fail to Reject H_0	
Truth	H_0 is True	V	U	M_0
	H_0 is False	S	T	M_A
		R	$M - R$	M

known constants

- Ideally we would like V and T to be small

Optional Exercises:

- Calculations: 8
- Proofs: 3
- R Analysis: 7, 16(not g), 23 (i,j,k)
- Communication: 1(a)