Two variable tests I

- Matched pair test. Random sample with two measurements per unit $X_1, Y_1, \ldots, X_N, Y_N$. Pairs of measurements are matched. We are interested in the difference. Compute $d_n = X_n - Y_n$ and perform inference on d_n as a single sample.

Two variable tests II

- Two different populations, two **independent** samples: $X_{1,1}, \ldots, X_{1,N_1}$ and $X_{2,1}, \ldots, X_{2,N_2}$. We are interested in the difference of the means $\mu_1 - \mu_2$.

  The estimate: $\bar{X}_1 - \bar{X}_2$.

  The SE:

  - Different variances $\sigma_1 \neq \sigma_2$
    $SE = \sqrt{S_1^2/N_1 + S_2^2/N_2}, \quad$ DF - $\min(N_1 - 1, N_2 - 1)$.
  - Same variances: $\sigma_1 = \sigma_2$. Pooled estimate:

$$S = \sqrt{\frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2}{N_1 + N_2 - 2}}.$$

$SE = S\sqrt{1/N_1 + 1/N_2}, \quad$ DF - $N_1 + N_2 - 2$.

We use t_{DF} for critical values and p-values.

Special case: Comparing two proportions I

Suppose we have two populations A and B with unknown proportions p_1 and p_2 respectively. A SRS of size N_1 from A yields $\hat{p}_1$, and an independent SRS of size N_2 from B yields $\hat{p}_2$. Then,

$$(\hat{p}_1 - \hat{p}_2) \overset{\cdot}{\sim} N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{N_1} + \frac{p_2(1-p_2)}{N_2}}\right)$$

when N_1 and N_2 are large.

Estimate: $\hat{p}_1 - \hat{p}_2$. **SE** $= \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{N_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{N_2}}$.

An approximate 95% CI for $p_1 - p_2$ is then given by

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96SE.$$

We use the standard normal for critical values and p-values.

Special case: Comparing two proportions II

To test $H_0 : p_1 = p_2$, we compute the test statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

where $\hat{p}$ is the *combined* proportion of successes in both samples

$$\hat{p} = \frac{X_1 + X_2}{N_1 + N_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{N_1 + N_2}$$

with X_1, X_2 denoting the number of successes in each sample.
Under H_0, the Z-statistic has approximately a standard normal distribution (using the normal approximation to the binomial).

Special case: Comparing two proportions III

The ability of question wording to affect the outcome of a survey can be a serious issue. Consider the following two questions:

1. Would you favor or oppose a law that would require a person to obtain a police permit before purchasing a gun?
2. Would you favor or oppose a law that would require a person to obtain a police permit before purchasing a gun, or do you think such a law would interfere too much with the right of citizens to own guns?

Let N_i denote the number of people who were asked question $i = 1, 2$, and $X_i = \sum_{n=1}^{N_i} X_{i,n}$ denote the number of these who favor the permit law.

Ques.	X_i	N_i
1	463	615
2	403	585

Special case: Comparing two proportions IV

Is the true proportion of people favoring the permit law the same in both groups or not?

```
prop.test(c(463,403),c(615,585),correct=FALSE)
```

```
2-sample test for equality of proportions without
continuity correction

data:  c out of c463 out of 615403 out of 585
X-squared = 6.1, df = 1, p-value = 0.01
alternative hypothesis: two.sided
95 percent confidence interval:
 0.01327 0.11465
sample estimates:
prop 1 prop 2
0.7528 0.6889
```

Example: Trumps tweets I

- Tweet from android: “The dishonest media didn’t mention that Bernie Sanders was very angry looking during Crooked’s speech. He wishes he didn’t make that deal!”
- Tweet from iPhone: “Join me in Fayetteville, North Carolina tomorrow evening at 6pm. Tickets now available at:”
- Let’s see if these differences are actually statistically meaningful.

Two-way table I

```
glimpse(trump)
```

```
Rows: 1,208
Columns: 7
$ source   <chr> "Android", "iPhone", "iPhone", "Androi...
$ text     <chr> "My economic policy speech will be car...
$ hour     <int> 10, 8, 19, 18, 16, 8, 21, 21, 20, 15, ...
$ quote    <chr> "no_quote", "no_quote", "no_quote", "n...
$ picture  <chr> "no_picture", "picture", "picture", "n...
$ positive <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, FA...
$ negative <lgl> FALSE, FALSE, FALSE, TRUE, TRUE, TRUE,...
```

```
tabdat <- table(trump$negative, trump$source)
rownames(tabdat) <- c("Non-negative", "Negative")
tabdat
```

```
               Android iPhone
  Non-negative     245    456
  Negative         341    165
```

Two-way table II

Test if proportions of negative in two phones is the same?

```
stats::prop.test(tabdat)
```

```
2-sample test for equality of proportions with
continuity correction

data:  tabdat
X-squared = 120, df = 1, p-value <2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.3801 -0.2687
sample estimates:
prop 1 prop 2
0.3495 0.6739
```

Hypotheses

- We label these hypotheses H_0 and H_A.
- H_0: The proportions are the same $=$ variables `source` and `negative` are independent. They have no relationship, and the observed difference in negative proportions was due to chance.
- H_A: The variables `source` and `negative` are not independent (they are associated). The observed difference in negative proportions is not due to chance.

Different way to test this Hypothesis - without any assumptions on distributions.

How are tweets generated under H_0?

- Under H_0, Trump chooses a tweet, then randomly chooses a phone to send out the tweet, regardless of it being negative or not.
- We can actually perform this randomization!
- I.e., randomly assign 586 of the tweets (whose negativity we know) to be sent from the Android phone and the rest (622) to be sent from the iPhone.
- Why these numbers?

```
table(trump$source)
```

```
Android  iPhone
    586     622
```

One such simulation

```
tabdat <- table(trump$negative, sample(trump$source))
```

Here `sample` is called with the default no-replacement and default sample size same as source. So it a permutation of the `tummp$source` column.

```
propdat <- prop.table(tabdat, margin = 2)
propdat
```

```
        Android iPhone
  FALSE  0.5761 0.5852
  TRUE   0.4239 0.4148
```

So in this case, 0.5761 of the Android tweets are negative and 0.5852 of the iPhone tweets are negative.
This difference -0.0091 is much smaller than in the original dataset.

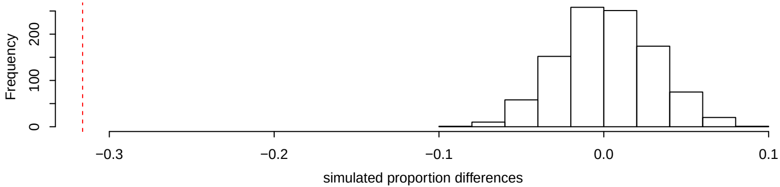
We can repeat this

Repeating this many times will tell us what the “likely” values of the difference are under H_0.

```
simdat <- rep(NA, length = 1000)
for (index in 1:1000) {
  tabdat <- table(trump$negative, sample(trump$source))
  propdat <- prop.table(tabdat, margin = 2)
  simdat[index] <- propdat[1, 1] - propdat[1, 2]
}
realtab <- prop.table(table(trump$negative, trump$source),
                      margin = 2)
realstat <- realtab[1, 1] - realtab[1, 2]
```

Plot the simulations

```
hist(simdat, xlim = c(realstat, max(simdat)),main="",xlab='simulated proportion differences') abline(v
= realstat, col = 2, lty = 2)
```



Since the data we observe is incredibly unlikely under H_0, we reject H_0 and conclude H_A.

Two way tables I

In a study conducted in a Northern Ireland supermarket, researchers counted the number of bottles of French, Italian, and other wine purchased while shoppers were subject to one of three "treatments": no music, French accordion music, and Italian string music.

The following **two-way table** summarizes the data:

	Music			
Wine	None	French	Italian	Total
French	30	39	30	99
Italian	11	1	19	31
Other	43	35	35	113
Total	84	75	84	243

Two way tables II

The table of counts looks like the joint distribution tables we studied earlier. Indeed, from these counts, we can ascertain the (empirical) joint distribution, marginal distributions, and conditional distributions of wine type and music type:

	Music			
Wine	None	French	Italian	Total
French	0.123	0.160	0.123	0.407
Italian	0.045	0.004	0.078	0.128
Other	0.177	0.144	0.144	0.465
Total	0.346	0.309	0.346	1.000

Two way tables III

We are interested in determining whether there is relationship between the row variable (wine type) and the column variable (music type).

If this were the *true distribution*, then the answer would be clear: music and wine are not independent, so there *is* a relationship.

However, this table is *random*, and we want to know whether or not music and wine are independent *under the true distribution*.

This requires a statistical test.

Two way tables IV

Hypotheses

- H_0: the row and column variables are independent (i.e. there is no relationship between the two).
- H_a: the row and column variables are dependent.

Intuition for the Test

Suppose H_0 is true, and the two variables are independent. What counts would we expect to observe?
Recall that under the independence assumption,
$P(x, y) = P(x)P(y)$, for all x, y.
We estimate the marginals from the data.
$\hat{P}(x) = \frac{\text{row total for x}}{\text{total count}}$, $\hat{P}(y) = \frac{\text{col total for y}}{\text{total count}}$

Two way tables V

Thus, for each cell, we have

$$\text{Expected Cell Count} = \text{total count} \cdot \hat{P}(x)\hat{P}(y) = \frac{\text{row total} \ \times \ \text{col total}}{\text{total count}}$$

Our test will be based on a measure of *how far the observed table is from the expected table.*

For the supermarket example, the expected counts are:

	Music			
Wine	None	French	Italian	Total
French	34.22	30.56	34.22	99
Italian	10.72	9.57	10.72	31
Other	39.06	34.88	39.06	113
Total	84	75	84	243

Two way tables VI

The X^2 (Chi-Squared) Statistic

	Observed			Expected			
	Music			Music			
Wine	None	Fr.	It.	None	Fr.	It.	Tot.
French	30	39	30	34.22	30.56	34.22	99
Italian	11	1	19	10.72	9.57	10.72	31
Other	43	35	35	39.06	34.88	39.06	113
Total	84	75	84	84	75	84	243

To measure how far the *expected* table is from the *observed* table, we will use the following test statistic:

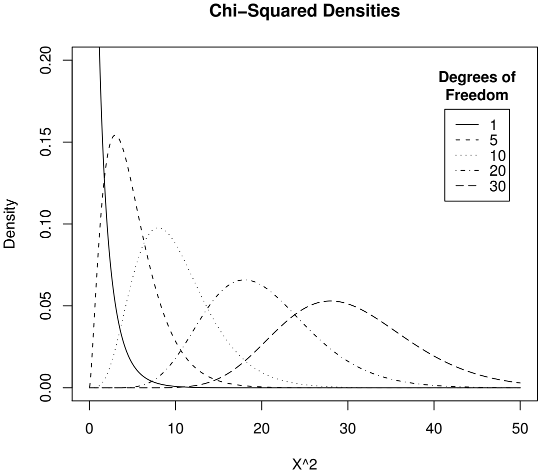
$$X^2 = \sum_{\text{all cells}} \frac{(\text{Observed} \;-\; \text{Expected})^2}{\text{Expected}}$$

Two way tables VII

The χ^2 Distribution

Under H_0, the X^2 test statistic has an approximate χ^2 distribution with $(r-1)(c-1)$ degrees of freedom, denoted $\chi^2_{(r-1)(c-1)}$.
Why $(r-1)(c-1)$?
Recall that our "expected" table is based on some quantities estimated from the data: namely the row and column totals. Once these totals are known, filling in any $(r-1)(c-1)$ undetermined table entries actually gives us the whole table. Thus, there are only $(r-1)(c-1)$ freely varying quantities in the table.
What does the χ^2 distribution look like?

Two way tables VIII



Two way tables IX

- Unlike the Normal or t distributions, the χ^2 distribution takes values in $(0, \infty)$.
- As with the t distribution, the exact shape of the χ^2 distribution depends on its degrees of freedom.

If the observed and expected counts are very different, X^2 will be large, indicating evidence against H_0. Thus, the p-value is always based on the right-hand tail of the distribution.
There is no notion of a two-tailed test in this context.
The p-value is therefore

$$P(\chi^2_{(r-1)(c-1)} \geq X^2)$$

Recall that X^2 has an *approximate* $\chi^2_{(r-1)(c-1)}$ distribution. When is the approximation valid?

Two way tables X

For any two-way table larger than 2 × 2, we require that the average expected cell count is at least 5 and each expected count is at least one.
For 2 × 2 tables, we require that each expected count be at least 5.

Let's get back to our example...
Recall the observed and expected counts:

	Observed			Expected			
	Music			Music			
Wine	None	Fr.	It.	None	Fr.	It.	Tot.
French	30	39	30	34.22	30.56	34.22	99
Italian	11	1	19	10.72	9.57	10.72	31
Other	43	35	35	39.06	34.88	39.06	113
Total	84	75	84	84	75	84	243

Two way tables XI

$$\begin{aligned} X^2 &= \frac{(30-34.22)^2}{34.22} + \frac{(39-30.56)^2}{30.56} + \frac{(30-34.22)^2}{34.22} \\ &\quad + \cdots + \frac{(35-34.88)^2}{34.88} + \frac{(35-39.06)^2}{39.06} \\ &= 18.28 \end{aligned}$$

The table is 3×3, so there are $(r-1)(c-1) = 2 \times 2 = 4$ degrees of freedom.
Finally, the p-value is found from the χ^2_4 table:

$$0.001 \leq P(\chi^2_4 \geq 18.28) \leq 0.0025$$

Two way tables XII

```
# Enter the data
wine=c(30,11,43,39,1,35,30,19,35)
# Reshape it as a table
dim(wine)=c(3,3)
wine
```

```
     [,1] [,2] [,3]
[1,]   30   39   30
[2,]   11    1   19
[3,]   43   35   35
```

```
# Run the test
chisq.test(wine)
```

```
Pearson's Chi-squared test

data:  wine
X-squared = 18, df = 4, p-value = 0.001
```

Two way tables XIII

Models for Two-Way Tables

The χ^2-test for the presence of a relationship between two directions in a two-way table is valid for data produced by several different study designs, although the exact null hypothesis varies.

1. Examining independence between variables

Suppose we select an SRS of size n from a population and classify each individual according to 2 categorical variables. Then, a χ^2-test can be used to test

H_0: The two variables are independent

H_a: Not independent

Example: We collect an SRS of 114 college students, and categorize each by major and GPA (e.g. $(0, 0.5], (0.5, 1], \ldots, (3.5, 4])$. Then, we can use a χ^2 test to ascertain whether grades and major are independent.

Two way tables XIV

2. Comparing several populations

Suppose we select *independent* SRSs from each of c populations, of sizes $n_1, n_2, \ldots, n_c$. We then classify each individual according to a categorical response variable with r possible values (the same across populations). This yields a $r \times c$ table, and a χ^2-test can be used to test

H_0: Distribution of the response variable is the same in all populations.

H_a: Distributions of response variables are not all the same.

Example: we select independent SRSs of Psychology, Biology and Math majors, of sizes 40, 39, 35, and classify each individual by GPA range. Then, we can use a χ^2 test to ascertain whether or not the distribution of grades is the same in all three populations.

2x2 tables and comparing proportions I

Back to the survey example. There are two populations and for each there is a 2 category (bernoulli) variable. So this is a 2x2 table. One variable is the polulation label and one variable is the response. Saying they have the same proportions is the same as saying the two variables are independent. And the chisq-statistic for independence is exactly the **square** of the z-statistic for equality of the proportions. The original data comes as two sample sizes and two counts of yes responses. It can be rewritten

Ques.	X_i	n_i
1	463	615
2	403	585

Ques.	Yes	No
1	463	152
2	403	182

2x2 tables and comparing proportions II

R-code for the chi-squared test for independence

```
# Enter data as two way table.
X=c(463,403,615-463,585-403)
dim(X)=c(2,2)
chisq.test(X,correct=FALSE)
```

```
Pearson's Chi-squared test

data:  X
X-squared = 6.1, df = 1, p-value = 0.01
```

2x2 tables and comparing proportions III

Explicit z-test for equal proportions

```
n1=X[1,1]+X[1,2]
n2=X[2,1]+X[2,2]
p1=X[1,1]/n1
p2=X[2,1]/n2
p=(X[1,1]+X[2,1])/(n1+n2)

z=(p1-p2)/sqrt(p*(1-p)*(1/n1+1/n2))
c(z,2*pnorm(-z))
```

```
[1] 2.47093 0.01348
```

2x2 tables and comparing proportions IV

R proportion test using the X^2 statistic

```
prp=prop.test(X[,1],c(n1,n2),correct=FALSE)
c(prp$statistic,prp$p.value)
```

```
X-squared
  6.10547   0.01348
```

```
z^2
```

```
[1] 6.105
```

Note that X-squared $= z^2$