Example 1

The 1991 General Social Survey asked respondents: Do you have in your home any guns or revolvers?" 393 individuals replied "Yes" and 583 individuals replied "No". Estimate the population proportion p of U.S. citizens who keep firearms at home.

The estimate for the proportion of U.S. citizens who keep firearms would be

$$\hat{p} = \frac{393}{393 + 583} = .403$$

A 95% CI for the population proportion p is given by

$$0.403 \pm 1.96\sqrt{\frac{0.403(1-0.403)}{976}} = (0.372,\ 0.434)$$

Example 2

Physicians' Health Study Research Group at Harvard Medical School investigated the relationship between aspirin use and Myocardial Infarction (heart attacks). A Five-year double-blind randomized study was conducted. Every other day physicians took either one aspirin or one placebo.

	MI		
Group	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

$$\hat{p_1} = \frac{189}{11,034} = 0.017,\ \hat{p_2} = \frac{104}{11,037} = 0.009,\ \hat{p_1} - \hat{p_2} = 0.008$$

A 95% CI for $p_1 - p_2$ is given by

$$0.008 \pm 1.96\sqrt{\frac{(0.017)(0.983)}{11,034} + \frac{(0.009)(0.991)}{11,037}}$$

$$= (0.008 \pm 0.0033)$$

Example 3

The following table from the 1991 General Social Survey cross-classifies respondents by their gender and their belief in an afterlife

	Belief		
Gender	Yes	No or Undecided	Total
Female	435	147	582
Male	375	134	509

$$\hat{p_1} = \frac{435}{582} = 0.747,\ \hat{p_2} = \frac{375}{509} = 0.737,\ \hat{p_1} - \hat{p_2} = 0.01$$

$$SE[\hat{p_1} - \hat{p_2}] = \sqrt{\frac{(0.747)(1-0.747)}{582} + \frac{(0.737)(1-0.737)}{509}}$$

$$= 0.027$$

The 95% CI margin of error is 0.052. Conclusion?

Brief Review of Significance Tests

Proof by Contradiction in Math and Logic

1. Do some reasoning
2. Always assume the opposite of what we are trying to prove
3. If we reason to a contradiction, we conclude that our assumption could not possibly be true.

How about Statistics?

The same strategy is used. We want to use data to provide evidence for a hypothesis, i.e., the **alternative hypothesis** H_a.

1. Assume the hypothesis is not true, i.e., assume the **null hypothesis** H_0 is true
2. Show that the null hypothesis H_0 is very **unlikely** to be true based on the observed data (by using a test statistic).

Rejection and Acceptance

1. If we do not "Reject H_0" should we say that we "Accept H_0"? It is better to use terminology like "Fail to Reject H_0 since H_0 may not be true but there isn't enough evidence in the data for such a conclusion.
2. A common convention is to reject H_0 if the P-value $< .05$
3. What if we get a P-value slightly larger than our threshold, e.g. a P-value $= .06$? This is why we use the above terminology.

Review of some basic inference ideas

1. For both hypothesis testing and confidence intervals, the key is to think about the distribution of the random process whose outcome we have just observed.
2. Need to use information about variability, i.e., the SD of the random variable whose values we observe.
3. In hypothesis testing, we see how far away our observation is from what we would expect under the null hypothesis H_0, with "how far" being expressed as a **multiple of the SD (or the SE)**.
4. For Confidence Intervals, we add or subtract some multiple of the **SD (or SE)**
5. Conversion of these **multiples** into probabilistic measures of strength of evidence (P-values, confidence intervals) is done with the appropriate distributions: Normal Distribution, t distribution, etc...

Analysis of Two-Way Tables

1. We continue our study of methods for analyzing categorical data.
2. We learned methods for statistical inference about proportions in one-sample and two-sample settings.
3. We now study how to compare two or more populations when the response variable has two or more possible values.

Two-way Tables

E.g. Cross-classification of a sample of 980 Americans by gender and party affiliation (data from the 1991 General Social Survey)

rows: Party (D,I,R) *columns: Gender (M,F)*

	F	M	Total
D	279	165	444
I	73	47	120
R	225	191	416
Total	577	403	980

What is the total number of female Democrats in the sample?

What is the total number of males in the sample?

What is the total sample size?

Notation

	F	M	Total
D	279	165	444
I	73	47	120
R	225	191	416
Total	577	403	980

Party affiliation = **Row Variable**

Gender = **Column Variable**

Each combination of values of two variables = **cell**

What is the total number of cells in the above table?

Joint Distribution

	F	M	Total
D	279	165	444
I	73	47	120
R	225	191	416
Total	577	403	980

A two-way table with proportions (or percentages) describes the **joint distribution** of the two variables. Each cell gives the proportion of the total sample size.

	F	M	Total
D	28.5%	16.8%	45.3%
I	7.4%	4.8%	12.2%
R	23.0%	19.5%	42.5%
Total	58.9%	41.1%	100.0%

Marginal Distribution

The distribution of a single variable in a two-way table is called the **marginal distribution.**

	F	M	Total
D	28.5%	16.8%	45.3%
I	7.4%	4.8%	12.2%
R	23.0%	19.5%	42.5%
Total	58.9%	41.1%	100.0%

What is the marginal distribution of Party affiliation?

D= 45.3%

I=12.2%

R=42.5%

What is the marginal distribution of Gender?

F= 58.9%

M=41.1%

Conditional Distribution

The distribution of one variable after we condition on the value of the other variable in a two way table is called the **conditional distribution.**

	F	M	Total
D	279	165	444
I	73	47	120
R	225	191	416
Total	577	403	980

What is the distribution of party affiliation for females in our sample of 980? (condition on Gender=F)

D=$\frac{279}{577} = 48.64\%$

I=$\frac{73}{577} = 12.6\%$

R=$\frac{225}{577} = 39.0\%$