STATS 111/202

Lecture 2: One-way and two-way tables

1/8/2025

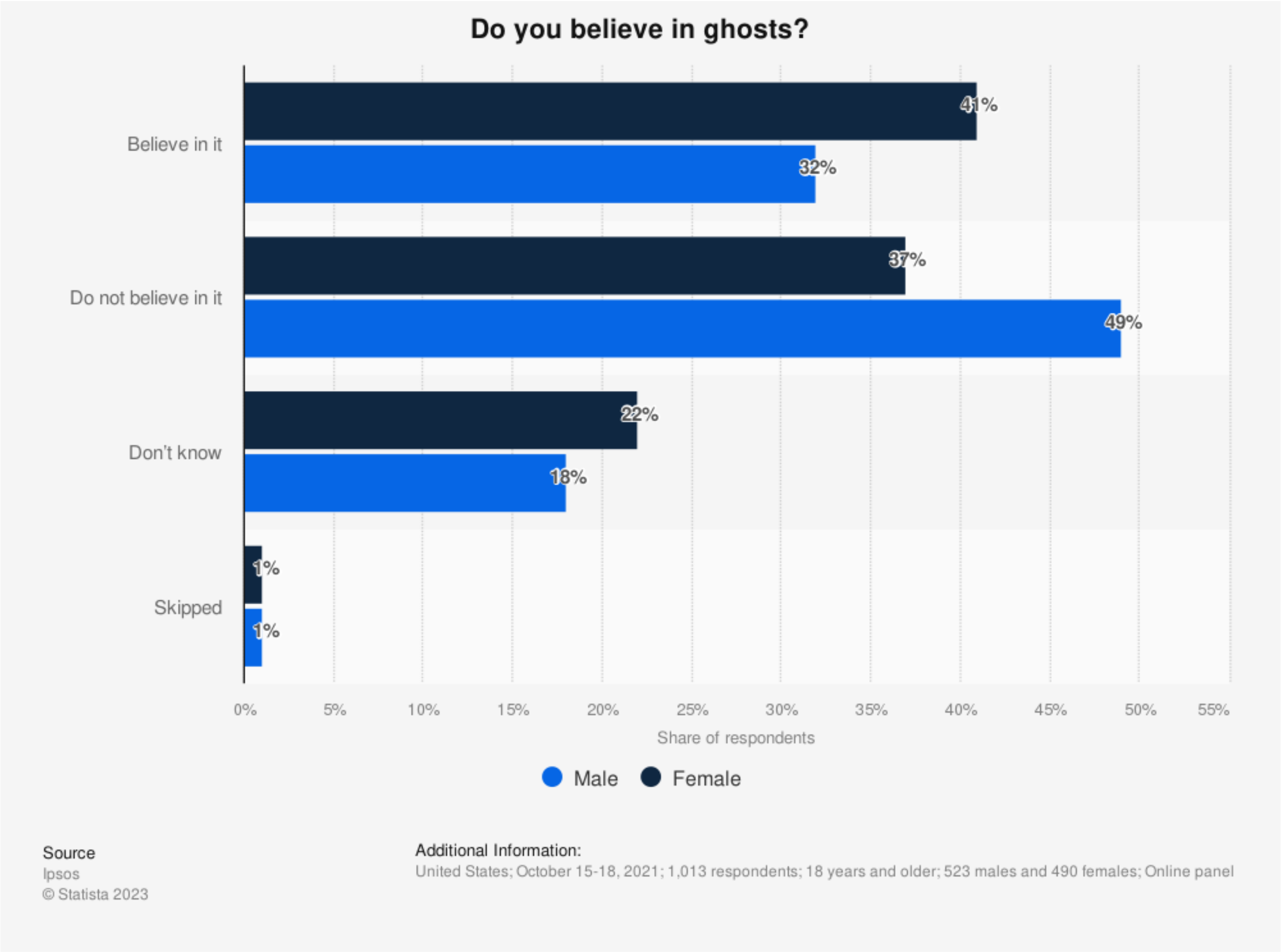
Ana Maria Kenney

UC Irvine Department of Statistics

One-way tables

1.3

Contingency tables



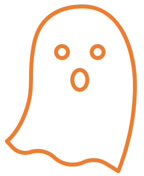
Characteristic	Male	Female
Believe in it	32%	41%
Do not believe in it	49%	37%
Don't know	18%	22%
Skipped	1%	1%

4x2 contingency table

How to we compare across these different groups?

Statista Research Department, published Jan 9th 2023

One-way table



	Believe in it	Do not believe in it
No. of responses	365	436

- Only have 2 outcomes (e.g., believe/don't believe)
- Say we have n observations, call each Y_i , where each one is a yes or no (1/0)
- Take n_0 as the no. of 0's and n_1 as the no. of 1's
  - $n_0 + n_1 = n$
- Then we want to estimate and provide inference on the probability a given trial is a 1 or 0 (e.g., believe in ghosts or not)
  - $p = P(Y_i = 1)$

Statista Research Department, published Jan 9th 2023

One-way table: connection to Bernoulli/Binomial

- Note that if Y_i values are 0 or 1, then $n_1 = \sum_{i=1}^{n} Y_i$
- Recall, we have the sum of n independent Bernoulli random variables, so $n_1 \sim Binomial(n, p)$
- Then to estimate the probability of success p, we take $\hat{p} = \frac{n_1}{n}$

Note: this is the maximum likelihood estimate of p

- Again, from properties of the Binomial, we have
  - $E(\hat{p}) = p$
  - $Var(\hat{p}) = \frac{p(1-p)}{n}$

Note: if we have a constant c and random variable Y then $var(cY) = c^2 var(Y)$

One-way table: confidence intervals

Now that we have a distribution to work off of, we can create **confidence intervals** and conduct **hypothesis tests** around p

- Let SE denote the estimated standard error of p. Then a large-sample $100(1-\alpha)\%$ confidence interval for p follows:

$$\hat{p} \pm z_{\frac{\alpha}{2}}(SE),\ SE = \sqrt{\hat{p}(1-\hat{p})/n}$$

- Where is the z-multiplier (right-tail probability is equal to $\alpha/2$ under a standard normal distribution)
  - E.g., for a 95% confidence interval: $\alpha = 0.05, z_{\alpha/2} = z_{0.025} = 1.96$

One-way table: confidence intervals

Now that we have a distribution to work off of, we can create **confidence intervals** and conduct **hypothesis tests** around p

- Let SE denote the estimated standard error of p. Then a large-sample $100(1-\alpha)\%$ confidence interval for p follows:

$$\hat{p} \pm z_{\frac{\alpha}{2}}(SE),\ SE = \sqrt{\hat{p}(1-\hat{p})/n}$$

- Where is the z-multiplier (right-tail probability is equal to $\alpha/2$ under a standard normal distribution)
  - E.g., for a 95% confidence interval: $\alpha = 0.05, z_{\alpha/2} = z_{0.025} = 1.96$

Q: why are we using a standard normal distribution when considering binomial outcomes?

One-way table: confidence intervals

Now that we have a distribution to work off of, we can create **confidence intervals** and conduct **hypothesis tests** around p

- Let SE denote the estimated standard error of p. Then a large-sample $100(1-\alpha)\%$ confidence interval for p follows:

$$\hat{p} \pm z_{\frac{\alpha}{2}}(SE),\ SE = \sqrt{\hat{p}(1-\hat{p})/n}$$

- Where is the z-multiplier (right-tail probability is equal to $\alpha/2$ under a standard normal distribution)
    - E.g., for a 95% confidence interval: $\alpha = 0.05, z_{\alpha/2} = z_{0.025} = 1.96$

Q: why are we using a standard normal distribution when considering binomial outcomes?

A: As n increases, the distribution of the sample proportion, $\hat{p}$, is approximately normal (CLT!)
The rule of sampling proportions sets thresholds of $np \geq 10$ and $n(1-p) \geq 10$

One-way table: hypothesis tests

- Now suppose we want to test a specific proportion p_0 (e.g, “I want to know whether 60% of people believe in ghosts”).

- Then we take the null hypothesis $H_0: p = p_0$ vs the alternative $H_a: p \neq p_0$ $(or\ p < p_0, p > p_0)$

- We consider the test statistic

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_o(1 - p_0)}{n}}}$$

One-way table: hypothesis tests in R

- To do these computations in R, we can use the function prop.test(n_1, n, p_0,alternative) where
  - n_1: total number of 1's or successes (no. of people who believe in ghosts)
  - n: total number of samples (no. of people in study)
  - π_0: the hypothesized proportion ("I think $p_0 = 0.6$ of the population believes in ghosts")
  - Alternative: type of alternative hypothesis "two.sided" ($\neq$), "less" (<), or "greater" (>)

One-way table: hypothesis tests in R example

	Believe in it	Do not believe in it
No. of responses	365	436

Here we have:

- $n = 365 + 436 = 801$
- $n_1 = 365$
- $p_0 = 0.6$ (whatever is being hypothesized)
- $\hat{p} = \frac{365}{801} = 0.4556804$
- $H_0: p = 0.6 \; vs \; H_a: p \neq 0.6$

```
> prop.test(365,801,0.6,alternative="two.sided")

        1-sample proportions test with continuity correction

data:  365 out of 801, null probability 0.6
X-squared = 68.914, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.6
95 percent confidence interval:
 0.4208666 0.4909249
sample estimates:
        p
0.4556804
```

One-way table: hypothesis tests in R example

	Believe in it	Do not believe in it
No. of responses	365	436

Here we have:

- $n = 365 + 436 = 801$
- $n_1 = 365$
- $p_0 = 0.6$ (whatever is being hypothesized)
- $\hat{p} = \frac{365}{801} = 0.4556804$
- $H_0: p = 0.6 \; vs \; H_a: p \neq 0.6$

```
> prop.test(365,801,0.6,alternative="two.sided")

	1-sample proportions test with continuity correction

data:  365 out of 801, null probability 0.6
X-squared = 68.914, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.6
95 percent confidence interval:
 0.4208666 0.4909249
sample estimates:
        p
0.4556804
```

Conclusion: we reject the null that the proportion is equal to 0.6 and have evidence to support the alternative

One-way table: hypothesis tests in R example

	Believe in it	Do not believe in it
No. of responses	365	436

Here we have:

- $n = 365 + 436 = 801$
- $n_1 = 365$
- $p_0 = 0.6$ (whatever is being hypothesized)
- $\hat{p} = \frac{365}{801} = 0.4556804$
- $H_0 : p = 0.6 \; vs \; H_a : p \neq 0.6$

```
> prop.test(365,801,0.6,alternative="two.sided")

	1-sample proportions test with continuity correction

data:  365 out of 801, null probability 0.6
X-squared = 68.914, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.6
95 percent confidence interval:
 0.4208666 0.4909249
sample estimates:
        p
0.4556804
```

One-way table: hypothesis tests in R example

	Believe in it	Do not believe in it
No. of responses	365	436

Here we have:

- $n = 365 + 436 = 801$
- $n_1 = 365$
- $p_0 = 0.6$ (whatever is being hypothesized)
- $\hat{p} = \frac{365}{801} = 0.4556804$
- $H_0: p = 0.6 \; vs \; H_a: p \neq 0.6$

```
> prop.test(365,801,0.6,alternative="two.sided")

        1-sample proportions test with continuity correction

data:  365 out of 801, null probability 0.6
X-squared = 68.914, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.6
95 percent confidence interval:
 0.4208666 0.4909249
sample estimates:
        p
0.4556804
```

Conclusion: we reject the null and have evidence supporting the alternative that less than 60% of people believe in ghosts

One-way table: hypothesis tests in R

If the p-value < significance level α, then reject the null and conclude evidence for the alternative

If the p-value > significance level α, then fail to reject the null and do not conclude evidence for the alternative

One-way table: more than two categories (k > 2)

	Believe in it	Do not believe in it	Don't know	Skipped
No. of responses	365	436	203	10

- Can extend this approach beyond two categories
- Now each Y_i (e.g., individual surveyed about ghosts) can select one of $k > 2$ choices
- Then instead of just one (true population) proportion p, we have k, $p_1, p_2, \dots, p_k$
- Now we can actually test different proportions for each one!
  - E.g., "I think 40% believe, 20% don't, 10% don't know, and 30% ignored the ghost question"
- Null - $H_0 : p_1 = p_{0_1}, p_2 = p_{0_2}, \dots, p_k = p_{0_k}$
- Alternative is trickier, any option that is no the null. So at least one category where $p_i \neq p_{0_i}$

One-way table: more than two categories (k > 2)

- Now we consider a new test statistic

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}$$

- Where we have
  - O_j : no. of observations/samples in category j
  - E_j: expected no. of observations in category j under the null, more specifically this is computed taking $E_j = np_{0\,j}$ (n is the sample size, $p_{0\,j}$ is the hypothesized proportion for the jth category)

- This follows (approximately) a chi-square distribution with $k - 1$ degrees of freedom (d.f.)

One-way table: more than two categories (k > 2)

- Now we consider a new test statistic

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}$$

- Where we have
  - O_j : no. of observations/samples in category j
  - E_j: expected no. of observations in category j under the null, more specifically this is computed taking $E_j = np_{0\,j}$ (n is the sample size, $p_{0\,j}$ is the hypothesized proportion for the jth category)

- This follows (approximately) a chi-square distribution with $k-1$ degrees of freedom (d.f.)

E.g., $O_{Believe} = 365, E_{Believe} = 1014 \times 0.4$
(if we think 40% believe)

One-way table: more than two categories (k > 2) in R

	Believe in it	Do not believe in it	Don’t know	Skipped
No. of responses	365	436	203	10

- In R, we use chisq.test(x,p) where
  - x is a vector of counts for each category (the above table)
  - p is a vector of the ${p_0}_i$ hypothesized proportions we want to test against
- Let’s say we want to test whether the true proportion of believers is 40%, do not believe 50%, don’t know is 8%, and skipped is 2%
  - $H_0: p_{Believer} = 0.4, p_{don't\ believe} = 0.5, p_{don't\ know} = 0.08, p_{skip} = 0.02$
  - Equivalent to $H_0: p_{Believer} = 0.4, p_{don't\ believe} = 0.5, p_{don't\ know} = 0.08$
  - $H_a: at\ least\ one\ category\ is\ not\ equal\ to\ its\ hypothesized\ proportion$

One-way table: more than two categories (k > 2) in R

	Believe in it	Do not believe in it	Don’t know	Skipped
No. of responses	365	436	203	10

```
> chisq.test(x=c(365,436,203,10),p=c(0.4,0.5,0.08,0.02))

        Chi-squared test for given probabilities

data:  c(365, 436, 203, 10)
X-squared = 202.34, df = 3, p-value < 2.2e-16
```

Conclusion: at a 0.05 significance level, we can reject the null and have evidence to support the alternative that at least one proportion is not equal to the hypothesized value

Two-way tables

2.1.2, 2.3.6

Two-way tables

	Believe in it	Do not believe in it
Male	100	372
Female	138	309

- Now we can add a possible explanatory variable to account for a single categorical variable (with $k \geq 2$ categories).

- The goal is to investigate whether a statistically signification relationship exists between the response and categorical response variable.

Two-way tables

General example

	Y=1	Y=0	Total
X=1	n_{11}	n_{12}	$n_{1+} = n_{11} + n_{12}$
X=0	n_{21}	n_{22}	$n_{2+} = n_{21} + n_{22}$
Total	$n_{+1} = n_{11} + n_{21}$	$n_{+2} = n_{12} + n_{22}$	n

Consider the scenario where a fixed number of subjects with X=0 and X=1 are sampled, and then we observe to see if they result in a Y=0 or Y=1

- Here n_{1+} and n_{2+} are fixed
- Take $p_0 = P(Y = 1|X = 0)$ and $p_1 = P(Y = 1|X = 1)$
- Then $n_{11} \sim Binomial(n_{1+}, p_1)$ and $n_{21} \sim Binomial(n_{2+}, p_0)$
- We compare p_0 to p_1

Two-way tables

Other important quantities:

- Risk difference $RD = p_1 - p_0$
- Relative risk (risk ratio): $RR = \frac{p_1}{p_0}$
- Odds ratio OR$= \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$
- Odds is the probability of an event divided by the probability of no event