Class 23

DATA1220-55, Fall 2024

Sarah E. Grabinski

2024-10-28

Chi-Square Test for Independence in 2-Way Tables

1. Assume the 2 variables are ***independent***. (H_0: Independence)

Chi-Square Test for Independence in 2-Way Tables

1. Assume the 2 variables are ***independent***. (H_0: Independence)
2. Calculate the ***expected*** counts under the null hypothesis of independence.

Chi-Square Test for Independence in 2-Way Tables

1. Assume the 2 variables are ***independent***. (H_0: Independence)
2. Calculate the ***expected*** counts under the null hypothesis of independence.
3. Find the ***test statistic***.

Chi-Square Test for Independence in 2-Way Tables

1. Assume the 2 variables are ***independent***. (H_0: Independence)
2. Calculate the ***expected*** counts under the null hypothesis of independence.
3. Find the ***test statistic***.
4. Compute the ***degrees of freedom***.

Chi-Square Test for Independence in 2-Way Tables

1. Assume the 2 variables are ***independent***. (H_0: Independence)
2. Calculate the ***expected*** counts under the null hypothesis of independence.
3. Find the ***test statistic***.
4. Compute the ***degrees of freedom***.
5. Determine the probability of the ***observed*** counts under the null hypothesis.

Chi-Square Test for Independence in 2-Way Tables

1. Assume the 2 variables are ***independent***. (H_0: Independence)
2. Calculate the ***expected*** counts under the null hypothesis of independence.
3. Find the ***test statistic***.
4. Compute the ***degrees of freedom***.
5. Determine the probability of the ***observed*** counts under the null hypothesis.
6. If it is sufficiently unlikely to have gotten the ***observed*** data under the null hypothesis of independence, reject H_0 and accept H_A: Dependence.

Example: Foodborne Illness

- There have been 430 cases of E. coli in your region.
- You interviewed these patients and 570 of their close associates about what they ate.
- 235 people ate at McDonald's, 415 people ate at Chipotle, and 350 ate at Arby's.
- 125, 165, and 140 of the people who ate at McDonald's, Chipotle, and Arby's respectively got sick.

Example: Foodborne Illness

- There have been 430 cases of E. coli in your region.
- You interviewed these patients and 570 of their close associates about what they ate.
- 235 people ate at McDonald's, 415 people ate at Chipotle, and 350 ate at Arby's.
- 125, 165, and 140 of the people who ate at McDonald's, Chipotle, and Arby's respectively got sick.

Research question: Is whether or not a person got sick dependent on what restaurant they ate at?

The Data

Table 1: Illness by Restaurant

Restaurant	Sick	Not Sick	Total
McDonald's	125	110	235
Chipotle	165	250	415
Arby's	140	210	350
Total	430	570	1000

Step 1: Assume independence

The Multiplication Rule for Independent Events

The probability of event A ***and*** event B occurring is the product of the probability that A occurs and the probability that B occurs.

Step 1: Assume independence

The Multiplication Rule for Independent Events

The probability of event A ***and*** event B occurring is the product of the probability that A occurs and the probability that B occurs.

$$\begin{aligned} \text{Expected}_{\text{AandB}} &= \frac{\text{count}(A) \times \text{count}(B)}{n} \\ &= P(A) \times P(B) \times n \end{aligned}$$

Step 2: Calculate expected counts

Table 2: Illness by Restaurant

Restaurant	Sick	Not Sick	Total
McDonald's	125	110	235
Chipotle	165	250	415
Arby's	140	210	350
Total	430	570	1000

$$\begin{aligned}\mathrm{Exp}_{\mathrm{McD,S}} &= \frac{n_{\mathrm{McD}} \times n_{\mathrm{S}}}{n} \\ &= \frac{235 \times 430}{1000} \\ &= 101.1\end{aligned}$$

Step 2: Calculate expected counts

Table 3: Illness by Restaurant

Restaurant	Sick	Not Sick	Total
McDonald's	125	110	235
Chipotle	165	250	415
Arby's	140	210	350
Total	430	570	1000

$$\begin{aligned}
\mathrm{Exp}_{\mathrm{McD,NS}} &= \frac{n_{\mathrm{McD}} \times n_{\mathrm{NS}}}{n} \\
&= \frac{235 \times 570}{1000} \\
&= 134.0
\end{aligned}$$

Step 2: Calculate expected counts

Table 4: Illness by Restaurant

Restaurant	Sick	Not Sick	Total
McDonald's	125	110	235
Chipotle	165	250	415
Arby's	140	210	350
Total	430	570	1000

$$\begin{aligned}\mathrm{Exp}_{\mathrm{Chi,S}} &= \frac{n_{\mathrm{Chi}} \times n_{\mathrm{S}}}{n} \\ &= \frac{415 \times 430}{1000} \\ &= 178.5\end{aligned}$$

Step 2: Calculate expected counts

Table 5: Illness by Restaurant

Restaurant	Sick	Not Sick	Total
McDonald's	125	110	235
Chipotle	165	250	415
Arby's	140	210	350
Total	430	570	1000

$$\begin{aligned}\mathrm{Exp}_{\mathrm{Chi,NS}} &= \frac{n_{\mathrm{Chi}} \times n_{\mathrm{NS}}}{n} \\ &= \frac{415 \times 570}{1000} \\ &= 236.6\end{aligned}$$

Step 2: Calculate expected counts

Table 6: Illness by Restaurant

Restaurant	Sick	Not Sick	Total
McDonald's	125	110	235
Chipotle	165	250	415
Arby's	140	210	350
Total	430	570	1000

$$\begin{aligned}\mathrm{Exp}_{\mathrm{Arb,S}} &= \frac{n_{\mathrm{Arb}} \times n_{\mathrm{S}}}{n} \\ &= \frac{350 \times 430}{1000} \\ &= 150.5\end{aligned}$$

Step 2: Calculate expected counts

Table 7: Illness by Restaurant

Restaurant	Sick	Not Sick	Total
McDonald's	125	110	235
Chipotle	165	250	415
Arby's	140	210	350
Total	430	570	1000

$$
\begin{aligned}
\mathrm{Exp}_{\mathrm{Arb,NS}} &= \frac{n_{\mathrm{Arb}} \times n_{\mathrm{NS}}}{n} \\
&= \frac{350 \times 570}{1000} \\
&= 199.5
\end{aligned}
$$

Step 2: Calculate expected counts

Restaurant	Observed	Expected	Difference
Sick			
McDonald's	125	101	24
Chipotle	165	178	-13
Arby's	140	151	-11
Not Sick			
McDonald's	110	134	-24
Chipotle	250	237	13
Arby's	210	199	11

Step 3: Find the test statistic

$$
\begin{aligned}
\chi^2_{\mathrm{df}} &= \sum_{i=1}^{k} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\
&= \frac{(120-101)^2}{101} + \frac{(115-134)^2}{134} + \frac{(165-178)^2}{178} + \\
&\quad \frac{(250-237)^2}{237} + \frac{(140-151)^2}{151} + \frac{(210-199)^2}{199} \\
&= 9.34
\end{aligned}
$$

Step 4: Compute the degrees of freedom

For 2 categorical variables in a 2-way contingency table where R is the number of rows and C is the number of columns, the degrees of freedom for a chi-square test of independence is...

$$\begin{aligned} \mathrm{df} &= (R-1) \times (C-1) \\ &= (3-1) \times (2-1) \\ &= 2 \end{aligned}$$

Step 5: Determine p-value

We always use the upper tail of the probability distribution for a chi-square test, so we use the parameter `lower.tail = F` in the function.

```
pchisq(9.34, df = 2, lower.tail = F)
```

```
[1] 0.00937227
```

Step 6: Decide to reject H_0

If our significance threshold is $\alpha = 0.05$ and the p-value of 0.0094, should we reject H_0?

Step 6: Decide to reject H_0

If our significance threshold is $\alpha = 0.05$ and the p-value of 0.0094, should we reject H_0?

Yes! It is unlikely that we would observe the data that we did if the null hypothesis were true.

Inference & Hypothesis Testing with Means

- The distribution of sample means $\bar{x}$ calculated from samples of size n from the same population approximates a normal distribution (i.e. the *sampling distribution*)

Inference & Hypothesis Testing with Means

- The distribution of sample means $\bar{x}$ calculated from samples of size n from the same population approximates a normal distribution (i.e. the *sampling distribution*)
- Observations in sample assumed to be ***independent and identically distributed*** (***i.i.d.***)

Inference & Hypothesis Testing with Means

- The distribution of sample means $\bar{x}$ calculated from samples of size n from the same population approximates a normal distribution (i.e. the *sampling distribution*)
- Observations in sample assumed to be ***independent and identically distributed*** (***i.i.d.***)
- Need $n \geq 30$ observations in sample

Inference & Hypothesis Testing with Means

- The distribution of sample means $\bar{x}$ calculated from samples of size n from the same population approximates a normal distribution (i.e. the *sampling distribution*)
- Observations in sample assumed to be ***independent and identically distributed*** (***i.i.d.***)
- Need $n \geq 30$ observations in sample
- Underlying population distribution is normal (less strict as sample n increases)

Sample Means & The Standard Normal (z) Distribution

- As n increases, the sampling distribution of $\bar{x}$ approximates the distribution $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right))$

Sample Means & The Standard Normal (z) Distribution

- As n increases, the sampling distribution of $\bar{x}$ approximates the distribution $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right))$
- When assumptions met, $\bar{x} \approx \mu$ and $s \approx \sigma$

Sample Means & The Standard Normal (z) Distribution

- As n increases, the sampling distribution of $\bar{x}$ approximates the distribution $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right))$
- When assumptions met, $\bar{x} \approx \mu$ and $s \approx \sigma$
- $s \approx \sigma$ is a strong assumption!

The t distribution

- Better than z when the population standard deviation σ is unknown (almost always)

The t distribution

- Better than z when the population standard deviation σ is unknown (almost always)
- Appears normal, but is flatter to allow more uncertainty about $SE = \frac{s}{\sqrt{n}}$ of μ

The t distribution

- Better than z when the population standard deviation σ is unknown (almost always)
- Appears normal, but is flatter to allow more uncertainty about $SE = \frac{s}{\sqrt{n}}$ of μ
- Centered at 0 with the single parameter ***degrees of freedom*** $(\mathrm{df} = n - 1)$

The t distribution

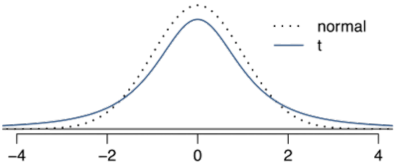


Figure 1: The t distribution versus the standard normal (z) distribution

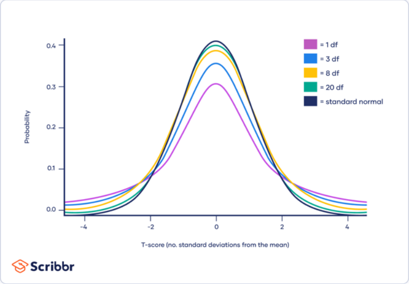
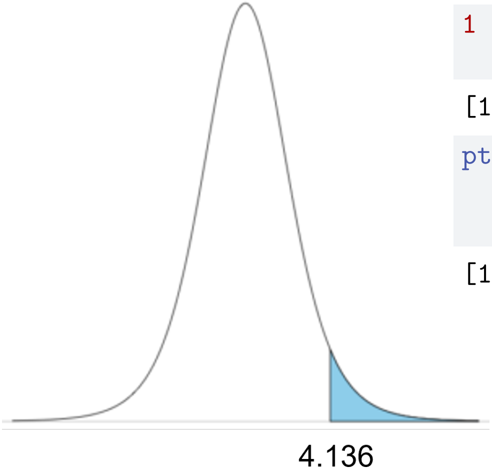


Figure 2: The t distribution is centered at 0 and has the parameter *degrees of freedom* (df)

t Distribution Test Statistic

$$\begin{aligned} T_{\text{df}} &= \frac{\text{point estimate} - \text{null value}}{SE} \\ T_{n-1} &= \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \end{aligned}$$

P-Values in R




```r
1 - pt(4.135,
       df = 205)
```

```
[1] 2.587688e-05
```

```r
pt(4.135,
   df = 205,
   lower.tail = F)
```

```
[1] 2.587688e-05
```

Confidence Intervals

- When $s \approx \sigma$, the confidence interval is
  point estimate $\pm Z^* \times SE$

Confidence Intervals

- When $s \approx \sigma$, the confidence interval is point estimate $\pm Z^* \times SE$
- When σ is unknown, we use point estimate $\pm T^* \times SE$

Confidence Intervals

- When $s \approx \sigma$, the confidence interval is point estimate $\pm Z^* \times SE$
- When σ is unknown, we use point estimate $\pm T^* \times SE$
- $T^* = T_{\alpha/2}$