

Chi-Squared Tests and Fisher's Exact Test

Textbook Sections 8.3–8.4

Emile Latour, Nicky Wakim, Meike Niederhausen

February 25, 2026

Learning Objectives

By the end of today's lecture, you will be able to:

1. Explain what a contingency table is and why we need the chi-squared test
2. Calculate expected cell counts under the null hypothesis of independence
3. Conduct a chi-squared test of association/independence using R
4. Check the technical conditions for the chi-squared test
5. Recognize when to use Fisher's exact test instead of the chi-squared test
6. Conduct Fisher's exact test in R and interpret results

Roadmap for Today

Part 1: From 2×2 Tables to Bigger Tables

- Extending two-proportion tests
- What is a contingency table?
- The motivating example: depression and physical activity
- Setting up the hypotheses

Part 2: The Chi-Squared Test

- Concept: observed vs. expected counts
- The chi-squared statistic
- The chi-squared distribution and degrees of freedom
- Assumptions / technical conditions

Part 3: Chi-Squared Test in Practice

- Walking through all 6 steps
- Using `chisq.test()` in R
- Extracting observed and expected counts
- Writing a conclusion

Part 4: Fisher's Exact Test

- When chi-squared breaks down
- The idea behind the exact test
- Using `fisher.test()` in R
- Chi-squared vs. Fisher's: how to decide

Where we are in the course

Last time we covered inference for **proportions**:

Day	Method	Question
13	One proportion test	Is the response rate different from 30%?
13	Two proportion test	Do response rates differ between groups?
Today	Chi-squared test	Is there an association between two categorical variables?
Today	Fisher's exact test	Same question, but with small samples

The key extension: What if one or both variables has **more than 2 levels**? We can no longer rely on a simple difference in proportions — we need a new method.

Part 1: From 2×2 Tables to Bigger Tables

Why do we need something new?

Last class, we tested whether two proportions are equal. For example:

Is the proportion of melanoma patients who respond to immunotherapy the same in treatment A vs. treatment B?

This works great when both variables are **binary** (two levels each). But what about:

Is there an association between depression level (None / Several days / Most days) and physical activity (Yes / No)?

Now the depression variable has **3 levels** — we can't just take a difference in two proportions! We need a way to compare proportions across multiple groups simultaneously.

The motivating example: Depression and Physical Activity

NHANES Data

Data from the National Health and Nutrition Examination Survey (NHANES), 2009–2012. A random sample of **400 US adults** (aged ≥ 18) with data on:

- **Depressed**: Self-reported number of days feeling down, depressed, or hopeless — **None**, **Several** (days), or **Most** (days)
- **PhysActive**: Whether the participant does moderate or vigorous physical activity — **Yes** or **No**

Research Question: Is there an association between depression status and physical activity?

This is a **2-row \times 3-column contingency table** (or 3×2 , depending on how you arrange it).

The contingency table

A **contingency table** shows the frequency distribution of two categorical variables simultaneously.

Physical Activity by Depression Status (n = 400)

PA/Depression	None	Several	Most	Total
Yes	199	26	1	226
No	115	32	27	174
Total	314	58	28	400

Each cell tells us: **how many people had this combination of depression status and physical activity level?**

Creating the contingency table in R

If you have individual-level (tidy) data:

```
1 # Using janitor::tabyl()
2 DepPA %>%
3   tabyl(PA, Depression)
4
5 # Using base R table()
6 table(DepPA$PA, DepPA$Depression)
```

	PA	None	Several	Most
Yes	199	26	1	
No	115	32	27	

If you only have the summary counts:

```
1 DepPA_mat <- matrix(
2   c(199, 26, 1, # PhysActive = Yes
3     115, 32, 27), # PhysActive = No
4   nrow = 2,
5   byrow = TRUE
6 )
7
8 dimnames(DepPA_mat) <- list(
9   PA = c("Yes", "No"),
10  Depression = c("None", "Several", "Most")
11 )
12
13 DepPA_mat
```

	Depression			
PA	None	Several	Most	
Yes	199	26	1	
No	115	32	27	

Setting up the hypotheses

Hypotheses for a Chi-Squared Test

For chi-squared tests, we **do not use symbols** (no $p_1 - p_2 = 0$). We always state hypotheses in words.

Two equivalent ways to write the hypotheses:

Test of “association” wording:

- H_0 : There is **no association** between the two variables
- H_A : There is **an association** between the two variables

Test of “independence” wording:

- H_0 : The two variables are **independent**
- H_A : The two variables are **not independent**

Both wordings are correct — pick one and be consistent. **NOTE** The test is always two-sided.

For our example: H_0 : There is no association between depression status and physical activity level

Part 2: The Chi-Squared Test

The core idea: observed vs. expected counts

The Key Insight

If depression and physical activity are truly **independent** (i.e., H_0 is true), then knowing someone's depression status tells us **nothing** about whether they are physically active.

In that case, the proportion who are physically active should be the **same across all depression groups**.

We test this by asking: **how far do the observed counts deviate from what we'd expect if the variables were independent?**

- **Observed counts** = what we actually see in the data
- **Expected counts** = what we'd expect to see if H_0 is true

Calculating expected counts

Under independence, the probability of being in a given cell is just the product of the row and column probabilities. This leads to the expected count formula:

$$E = \frac{R \times C}{N}$$

where

- R is the marginal total for a given row
- C is the marginal total for a given column
- N is the grand total for the table

Expected counts will sum to the same row and column (marginal) totals.

Observed vs. expected: side by side

Observed counts:

PA/Depression	Observed			
	None	Several	Most	Total
Yes	199	26	1	226
No	115	32	27	174
Total	314	58	28	400

Expected counts (if independent):

PA/Depression	Expected			
	None	Several	Most	Total
Yes	177.41	32.77	15.82	226.00
No	136.59	25.23	12.18	174.00
Total	314.00	58.00	28.00	400.00

Example: Expected count for PhysActive = Yes, Depression = Several:

$$E = \frac{(\text{row total for Yes}) \times (\text{col total for Several})}{\text{table total}} = \frac{226 \times 58}{400} = 32.77$$

Notice

Look at **Most** depression, PhysActive = **Yes**: observed = 1, expected = 15.82. That's a huge discrepancy! People with the most depression days are far less physically active than independence would predict.

Calculating expected counts in R

```
1 observed_table <- DepPA_mat
2
3 row_totals <- rowSums(observed_table)
4 col_totals <- colSums(observed_table)
5 grand_total <- sum(observed_table)
6
7 # Calculate expected counts using the manual formula
8 manual_expected <- outer(row_totals, col_totals) / grand_total
9
10 # View the manually calculated expected counts
11 manual_expected
```

	None	Several	Most
Yes	177.41	32.77	15.82
No	136.59	25.23	12.18

The `outer()` function in R computes the outer product of two vectors, which efficiently creates the matrix of `row_totals * col_totals` for all cells.

The chi-squared test statistic

Once we have observed and expected counts, the **test statistic** quantifies the total discrepancy:

$$\chi^2 = \sum_{\text{all cells}} \frac{(O - E)^2}{E}$$

- O = observed count in a cell
- E = expected count in that cell
- We sum this ratio over **all cells** in the table

Intuition:

- If H_0 is true $\rightarrow O \approx E \rightarrow$ small χ^2
- If H_A is true \rightarrow some O are very different from $E \rightarrow$ large χ^2
- **Large χ^2 = evidence against H_0**

Why $(O - E)^2 / E$?

$$\chi^2 = \sum_{\text{all cells}} \frac{(O - E)^2}{E}$$

- Each term measures how far a cell is from what we'd expect under independence
- We square the difference so that **positive and negative deviations don't cancel out**.
- We divide by E to **standardize** — a deviation of 10 matters more when the expected count is 12 than when it's 300.

χ^2 with depression/activity data

For our depression/physical activity data:

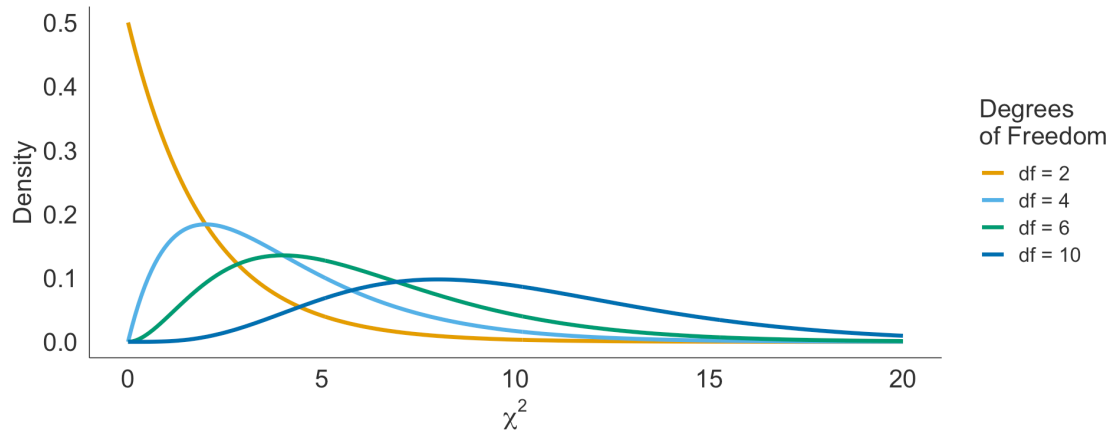
$$\chi^2 = \frac{(199 - 178.13)^2}{178.13} + \frac{(26 - 32.77)^2}{32.77} + \frac{(1 - 15.82)^2}{15.82} + \dots = 41.17$$

Is 41.17 large enough to reject H_0 ? We need to know what distribution this statistic follows!

The chi-squared distribution

Under H_0 , the test statistic follows a **chi-squared distribution** with degrees of freedom:

$$df = (\# \text{ rows} - 1) \times (\# \text{ columns} - 1)$$



Key properties:

- Always **right-skewed**
- Only takes **non-negative** values ($\chi^2 \geq 0$)
- Shape depends entirely on **df**
- As df increases, distribution becomes more symmetric

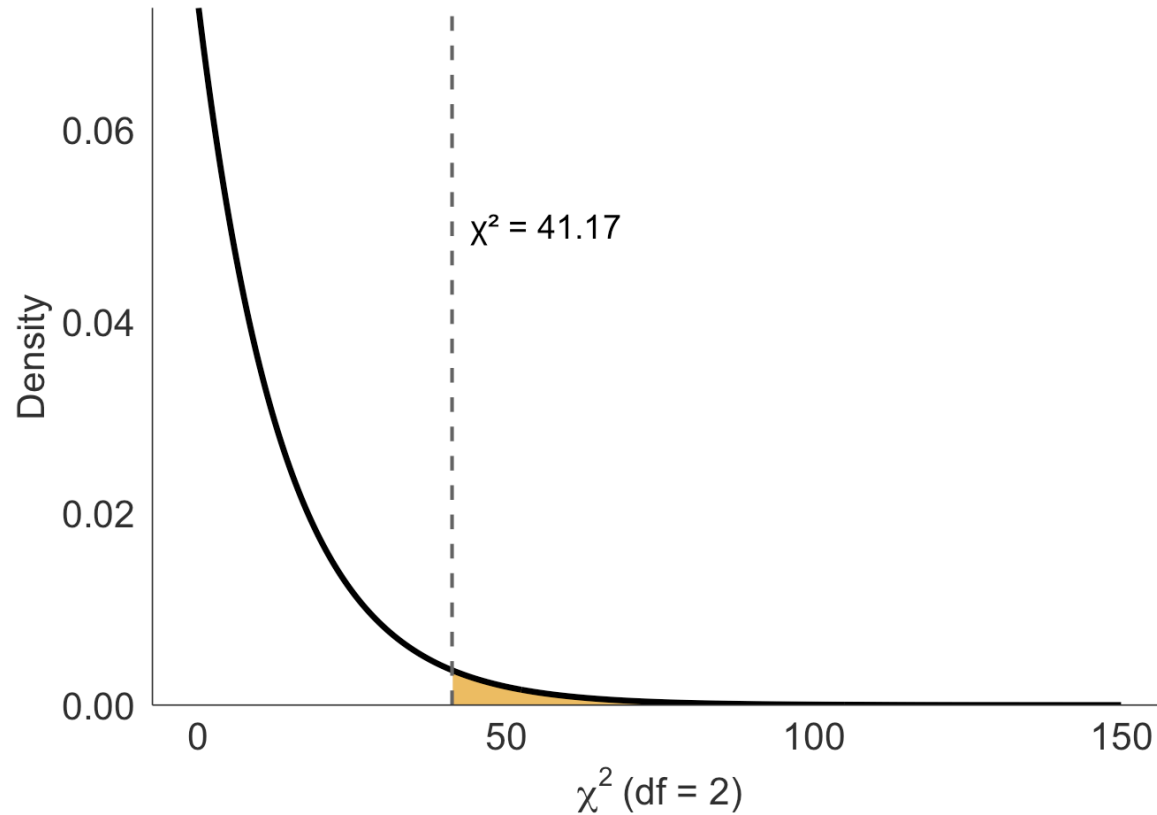
For our 2×3 table:

$$df = (2 - 1)(3 - 1) = 2$$

The **p-value** is always the area to the right of the **observed** χ^2 — it is always a one-sided, upper-tail test.

P-value for the chi-squared test

The p-value is the area to the **right** of our test statistic:



```
1 # Calculate p-value manually
2 pchisq(abs(41.17), df = 2, lower.tail = FALSE)
```

```
[1] 1.148281e-09
```



Technical conditions (assumptions)

Before running the chi-squared test, we need to check two conditions:

Conditions for the Chi-Squared Test

1. Independence

Each observation must be independent. In practice:

- Each person contributes to **exactly one cell** of the table
- Observations are not paired or clustered (no repeated measures)

2. Sample size (expected counts)

The chi-squared approximation is valid when:

- **2×2 tables:** All expected counts ≥ 10
- **Larger tables:** No more than 20% of expected counts < 5 , **and** all expected counts ≥ 1

If these fail → use Fisher's exact test (Part 4)

Part 3: Chi-Squared Test in Practice

The 6-step framework

We'll use our standard 6-step approach for hypothesis testing.

Our question: Is there an association between depression status and physical activity among US adults?

Data: Random sample of 400 US adults from NHANES 2009–2012

Let's go through all 6 steps together.

From matrix to real data

In Part 2 we built `DepPA_mat` by hand to illustrate the concepts. In practice, you always start from your actual dataset.

Base R's `table()` produces the same result:

```
1 # These are equivalent
2 DepPA_mat
```

	Depression		
PA	None	Several	Most
Yes	199	26	1
No	115	32	27

```
1 table(DepPA$PA, DepPA$Depression)
```

	None	Several	Most
Yes	199	26	1
No	115	32	27

For the 6 steps, we'll use `table()` on the real data — which is how you'll do it in your own analyses.

Step 1: State the hypotheses

H_0 : There is no association between depression status and physical activity

H_A : There is an association between depression status and physical activity

Note: No symbols needed

Unlike tests for means or proportions, chi-squared hypotheses are **always stated in words**. There's no simple symbolic representation like $\mu_1 = \mu_2$ because we're simultaneously comparing proportions across multiple groups and levels.

Step 2: Set significance level

$$\alpha = 0.05$$

Step 3: Check assumptions

Independence: Each person contributes to exactly one cell. A person can't be both physically active and not physically active, and can't have both "None" and "Several" depression days. ✓

Expected counts:

In practice, I check expected counts *after* running the test — `chisq.test()` will warn you if expected counts are too small, and you can pull them directly from the result object:

```
1 dep_pa_tab <- table(DepPA$PA, DepPA$Depression)
2
3 chisq_result <- chisq.test(dep_pa_tab)
4
5 # check what is stored in the result
6 names(chisq_result)
```

```
[1] "statistic" "parameter" "p.value"    "method"    "data.name" "observed"
[7] "expected"  "residuals" "stdres"
```

```
1 # Look at the expected counts
2 chisq_result$expected
```

	None	Several	Most
Yes	177.41	32.77	15.82
No	136.59	25.23	12.18

All expected counts > 5. The only borderline cell is PhysActive = Yes, Depression = Most (expected = 15.82), which easily passes. ✓

Step 4: Calculate the test statistic

$$\chi^2 = \sum_{\text{all cells}} \frac{(O - E)^2}{E}$$

```
1 chisq_result <- chisq.test(dep_pa_tab)
2 chisq_result
```

Pearson's Chi-squared test

```
data: dep_pa_tab
X-squared = 41.171, df = 2, p-value = 1.148e-09
```

The test statistic is $\chi^2 = 41.17$ with $df = (2 - 1)(3 - 1) = 2$.

Step 5: Calculate the p-value

```
1 chisq_result$statistic    # test statistic
```

```
X-squared  
41.17067
```

```
1 chisq_result$parameter    # degrees of freedom
```

```
df  
2
```

```
1 chisq_result$p.value      # p-value
```

```
[1] 1.147897e-09
```

- The p-value is essentially 0 (< 0.001).
- This means: **if there were truly no association between depression and physical activity (the null hypothesis is true), the probability of observing data as extreme as ours (or more extreme) is less than 0.001.**

Step 6: Conclusion

Since $p < 0.001 < \alpha = 0.05$, we **reject** H_0 .

Conclusion Statement

Based on a random sample of 400 US adults from NHANES 2009–2012, there is very strong evidence of an association between depression status and physical activity level ($\chi^2 = 41.17, df = 2, p < 0.001$).

Those with more days of depression appear to be substantially less physically active than would be expected if the variables were independent.

Common mistake

If we fail to reject H_0 , we do **NOT** say “the variables are independent.” We say: “We have insufficient evidence of an association.”
Absence of evidence \neq evidence of absence!

Base R: contingency table and expected counts

```
1 # Make the table
2 dep_pa_tab <- table(DepPA$PA, DepPA$Depression)
3
4 dep_pa_tab
```

	None	Several	Most
Yes	199	26	1
No	115	32	27

```
1 # Run the test
2 chisq_result <- chisq.test(dep_pa_tab)
3
4 # Check expected counts
5 chisq_result$expected
```

	None	Several	Most
Yes	177.41	32.77	15.82
No	136.59	25.23	12.18

Base R: test result

```
1 chisq_result
```

Pearson's Chi-squared test

data: dep_pa_tab
X-squared = 41.171, df = 2, p-value = 1.148e-09

```
1 chisq_result %>%  
2 broom::tidy()
```

```
# A tibble: 1 × 4  
  statistic      p.value parameter method  
    <dbl>      <dbl>    <int>   <chr>  
1    41.2 0.00000000115         2 Pearson's Chi-squared test
```


janitor: contingency table and expected counts

```
1 # Make the table
2 dep_pa_tab_janitor <- DepPA %>%
3   tabyl(PA, Depression)
4
5 dep_pa_tab_janitor
```

PA	None	Several	Most
Yes	199	26	1
No	115	32	27

```
1 # Run the test
2 chisq_result_janitor <- dep_pa_tab_janitor %>%
3   chisq.test()
4
5 # Check expected counts
6 chisq_result_janitor$expected
```

PA	None	Several	Most
Yes	177.41	32.77	15.82
No	136.59	25.23	12.18

janitor: test result

```
1 chisq_result_janitor
```

Pearson's Chi-squared test

```
data: .  
X-squared = 41.171, df = 2, p-value = 1.148e-09
```

```
1 chisq_result_janitor %>%  
2   broom::tidy()
```

```
# A tibble: 1 × 4  
  statistic      p.value parameter method  
    <dbl>        <dbl>    <int> <chr>  
1     41.2 0.00000000115         2 Pearson's Chi-squared test
```

rstatix: contingency table and expected counts

```
1 # Make the table
2 dep_pa_tab <- table(DepPA$PA, DepPA$Depression)
3
4 dep_pa_tab
```

	None	Several	Most
Yes	199	26	1
No	115	32	27

```
1 # Run the test
2 chisq_result_rstatix <- dep_pa_tab %>%
3   rstatix::chisq_test()
4
5 # Check expected counts
6 chisq_result_janitor$expected
```

PA	None	Several	Most
Yes	177.41	32.77	15.82
No	136.59	25.23	12.18

rstatix: test result

```
1 chisq_result_rstatix
```

```
# A tibble: 1 × 6
  n statistic      p    df method      p.signif
* <int>     <dbl>   <dbl> <int> <chr>      <chr>
1   400      41.2 0.00000000115     2 Chi-square test ****
```

Note

`rstatix` returns a tidy data frame directly — no need for `broom::tidy()`.

Part 3: Summary

- Chi-squared test asks: **do the observed counts match what we'd expect under independence?**
- Use `table()` or `janitor::tabyl()` to build the contingency table
- Use `chisq.test()`, `janitor::chisq.test()`, or `rstatix::chisq_test()` to run it
- Always check `$expected` — if any expected counts are too small, use Fisher's exact test (Part 4)

Part 4: Fisher's Exact Test

When chi-squared breaks down

The chi-squared test relies on an **approximation** — the test statistic only follows the chi-squared distribution when the **expected counts are large enough**.

What happens when we have small samples? Suppose instead of 400 adults, we had only **100** adults:

Physical Activity by Depression Status (n = 100)

PA/Depression	None	Several	Most	Total
Yes	43	5	2	50
No	40	4	6	50
Total	83	9	8	100

The chi-squared warning

```
1 # Make the table
2 dep_pa_tab <- table(DepPA100$PA, DepPA100$Depression)
3
4 # Run the test
5 chisq_res <- chisq.test(dep_pa_tab)
```

Warning in stats::chisq.test(x, y, ...): Chi-squared approximation may be incorrect

What does this warning mean?

R tells us: *"Chi-squared approximation may be incorrect."* The expected counts for the "Most" depression column are too small (< 5), violating our sample size condition. The chi-squared approximation can't be trusted here. **We need a different approach.**

```
1 chisq_res$expected
```

	None	Several	Most
Yes	41.5	4.5	4
No	41.5	4.5	4

Fisher's Exact Test: the idea

The Core Concept

Fisher's exact test computes the **exact probability** of observing our data (or data more extreme) given the **fixed row and column totals**.

It's called "exact" because instead of using an asymptotic approximation (like the chi-squared distribution), it computes the probability **exactly** using the **hypergeometric distribution** for 2×2 case; multivariate hypergeometric distribution in $r \times c$ case.

Think of it this way: if the row and column totals are fixed, there's only a limited number of ways to fill in the cells. Fisher's test calculates how likely our observed arrangement is under H_0 .

Key features of Fisher's exact test

What's different about Fisher's exact test?

1. **No test statistic** — the p-value is calculated directly from the data
2. **No confidence interval** from the standard output (for tables larger than 2×2)
3. **Always two-sided** — no option to specify direction of the alternative
4. **No continuity correction** needed — the hypergeometric distribution is already discrete
5. **Works for any sample size** — exact regardless of whether expected counts are small or large

When to use it:

- **2×2 tables:** Any expected count < 10
- **Larger tables:** More than 20% of expected counts < 5 , or any expected count < 1

Fisher's exact test in R

```
1 fisher.test(dep_pa_tab)
```

Fisher's Exact Test for Count Data

```
data: dep_pa_tab  
p-value = 0.3844  
alternative hypothesis: two.sided
```

The p-value of 0.384 provides no evidence of an association between depression and physical activity in this smaller sample (Fisher's exact test). We fail to reject H_0 at $\alpha = 0.05$.

Note: the larger sample ($n = 400$) gave very strong evidence of an association. With only $n = 100$, we don't have enough data to detect the association reliably — this is a **power** issue!

Base R: Fisher's exact test

```
1 # Make the table
2 dep_pa_tab100 <- table(DepPA100$PA, DepPA100$Depression)
3 dep_pa_tab100
```

	None	Several	Most
Yes	43	5	2
No	40	4	6

```
1 # Run the test
2 fisher_result <- fisher.test(dep_pa_tab100)
3 fisher_result
```

Fisher's Exact Test for Count Data

```
data: dep_pa_tab100
p-value = 0.3844
alternative hypothesis: two.sided
```

```
1 fisher_result %>%
2   broom::tidy()
```

```
# A tibble: 1 × 3
  p.value method alternative
  <dbl> <chr>      <chr>
1 0.3844 Fisher's Exact Test for Count Data two.sided
```

BMSC 920: Chi-Squared Tests and Fisher's Exact Test

janitor: Fisher's exact test

```
1 # Make the table
2 dep_pa_tab100_janitor <- DepPA100 %>%
3   tabyl(PA, Depression)
4 dep_pa_tab100_janitor
```

	PA None	Several	Most
Yes	43	5	2
No	40	4	6

```
1 # Run the test
2 fisher_result_janitor <- dep_pa_tab100_janitor %>%
3   fisher.test()
4
5 fisher_result_janitor
```

Fisher's Exact Test for Count Data

```
data: .
p-value = 0.3844
alternative hypothesis: two.sided
```

```
1 fisher_result_janitor %>%
2   broom::tidy()
```

```
# A tibble: 1 × 3
#   p.value method alternative
```

rstatix: Fisher's exact test

```
1 # Make the table
2 dep_pa_tab100 <- table(DepPA100$PA, DepPA100$Depression)
3 dep_pa_tab100
```

	None	Several	Most
Yes	43	5	2
No	40	4	6

```
1 dep_pa_tab100 %>%
2   rstatix::fisher_test()
```

```
# A tibble: 1 × 3
      n      p p.signif
* <int> <dbl> <chr>
1   100 0.384 ns
```

Note

`rstatix` returns a tidy data frame directly — no need for `broom::tidy()`.

Another small-sample example

Clinical Example: Adverse Events in a Cancer Drug Trial

A pilot trial of 25 patients tests a new targeted therapy. Investigators are concerned about liver toxicity:

	Elevated enzymes	Normal	Total
Treated	5	8	13
Control	1	11	12
Total	6	19	25

With only 25 patients, expected counts in the “Elevated enzymes” column will be very small. Chi-squared is not appropriate here.

```
1 liver_mat <- matrix(  
2   c(5, 8,  
3     1, 11),  
4   nrow = 2, byrow = TRUE,  
5   dimnames = list(Group = c("Treated", "Control"),  
6                     Enzymes = c("Elevated", "Normal"))  
7 )
```

Interpreting the Fisher's exact test output

```
1 fisher.test(liver_mat)
```

Fisher's Exact Test for Count Data

```
data: liver_mat
p-value = 0.1602
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.5555821 353.2536123
sample estimates:
odds ratio
 6.382758
```

- **p-value = 0.160** → No significant evidence of an association at $\alpha = 0.05$
- **Odds ratio = 6.4** → Treated patients have about 7 times the odds of elevated liver enzymes compared to control
- **95% CI: (0.6 — 353.3)** → Extremely wide — the study is **underpowered** for detecting a difference

Takeaway

With only 25 patients, we lack the statistical power to detect even a potentially large effect. This study should be viewed as exploratory — not definitive. Larger trials are needed.

Chi-Squared vs. Fisher's Exact: how to decide

	Chi-Squared Test	Fisher's Exact Test
When to use	Expected counts large enough	Expected counts too small
2×2 table	All $E \geq 10$	Any $E < 10$
Larger tables	$\leq 20\%$ of $E < 5$, all $E \geq 1$	$> 20\%$ of $E < 5$, or any $E < 1$
Test statistic	χ^2	None
Distribution	Chi-squared (approx.)	Hypergeometric (exact)
Confidence interval	No (from test)	For 2×2 only
One-sided available?	No	No
R function	<code>chisq.test()</code>	<code>fisher.test()</code>

Rule of thumb: Always check expected counts first. If in doubt, Fisher's exact test is conservative and always valid — it just requires more computing time for large tables.

Bonus: simulated p-values for larger tables

For **large contingency tables** with small expected counts, `fisher.test()` can be slow. An alternative is to simulate the p-value using `chisq.test()`:

- This randomly generates many tables with the same row/column totals and calculates what proportion have a χ^2 statistic as large as ours.
- **B = 10000** means we simulate 10,000 random tables
- Results will be slightly different each run (set a seed for reproducibility!)

```
1 set.seed(456)
2 chisq.test(dep_pa_tab, simulate.p.value = TRUE, B = 10000)
```

Pearson's Chi-squared test with simulated p-value (based on 10000 replicates)

```
data:  dep_pa_tab
X-squared = 2.2195, df = NA, p-value = 0.3817
```

Summary and Key Takeaways

What we learned today

1. Chi-squared test of association

- Tests whether two categorical variables are associated (independent)
- Compares **observed** vs. **expected** counts under independence
- Expected count = $\frac{\text{row total} \times \text{column total}}{\text{table total}}$
- Test statistic: $\chi^2 = \sum \frac{(O-E)^2}{E}$, with $df = (r - 1)(c - 1)$
- Use `chisq.test()` in R
- **Always check expected counts first!**

2. Fisher's exact test

- Use when expected counts are too small for chi-squared
- Computes exact p-value using the hypergeometric distribution
- No test statistic, no CI (for tables $> 2 \times 2$), always two-sided
- Use `fisher.test()` in R

Key decision: which test to use?

Decision workflow for categorical association tests

Step	Action
1. Create the contingency table	<code>tabyl()</code> or <code>table()</code> or <code>matrix()</code>
2. Calculate expected counts	<code>chisq.test()\$expected</code>
3a. Expected counts are large enough	→ Use <code>chisq.test()</code>
3b. Expected counts are too small	→ Use <code>fisher.test()</code>