Day 13: Chi-squared tests (Sections 8.3-8.4)

BSTA 511/611

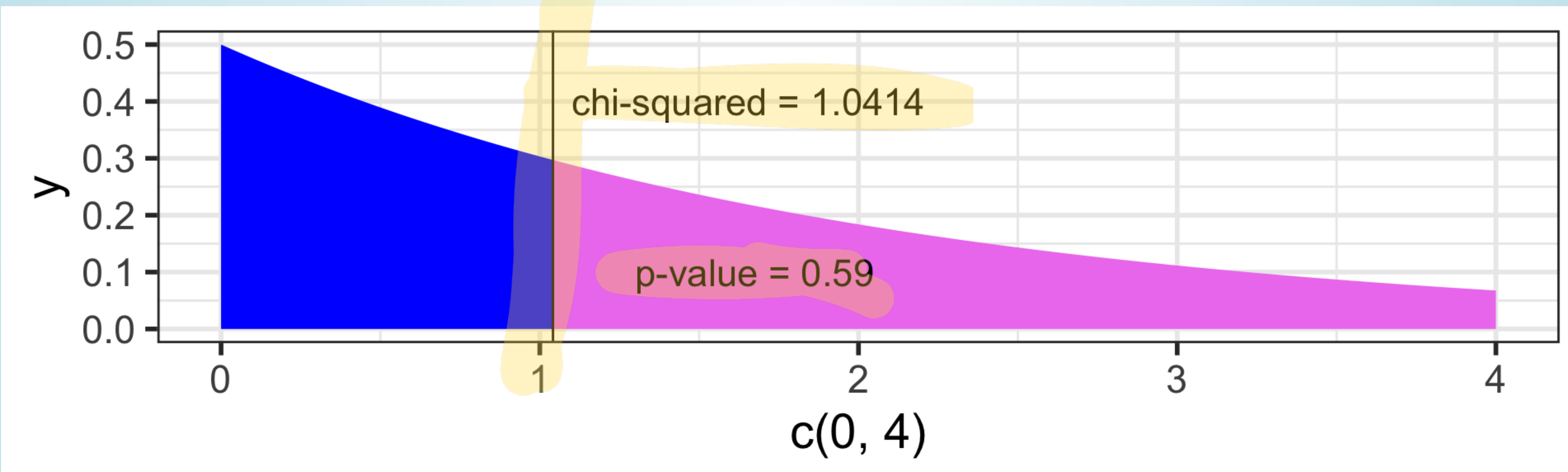
Meike Niederhausen, PhD
OHSU-PSU School of Public Health

2023-11-13

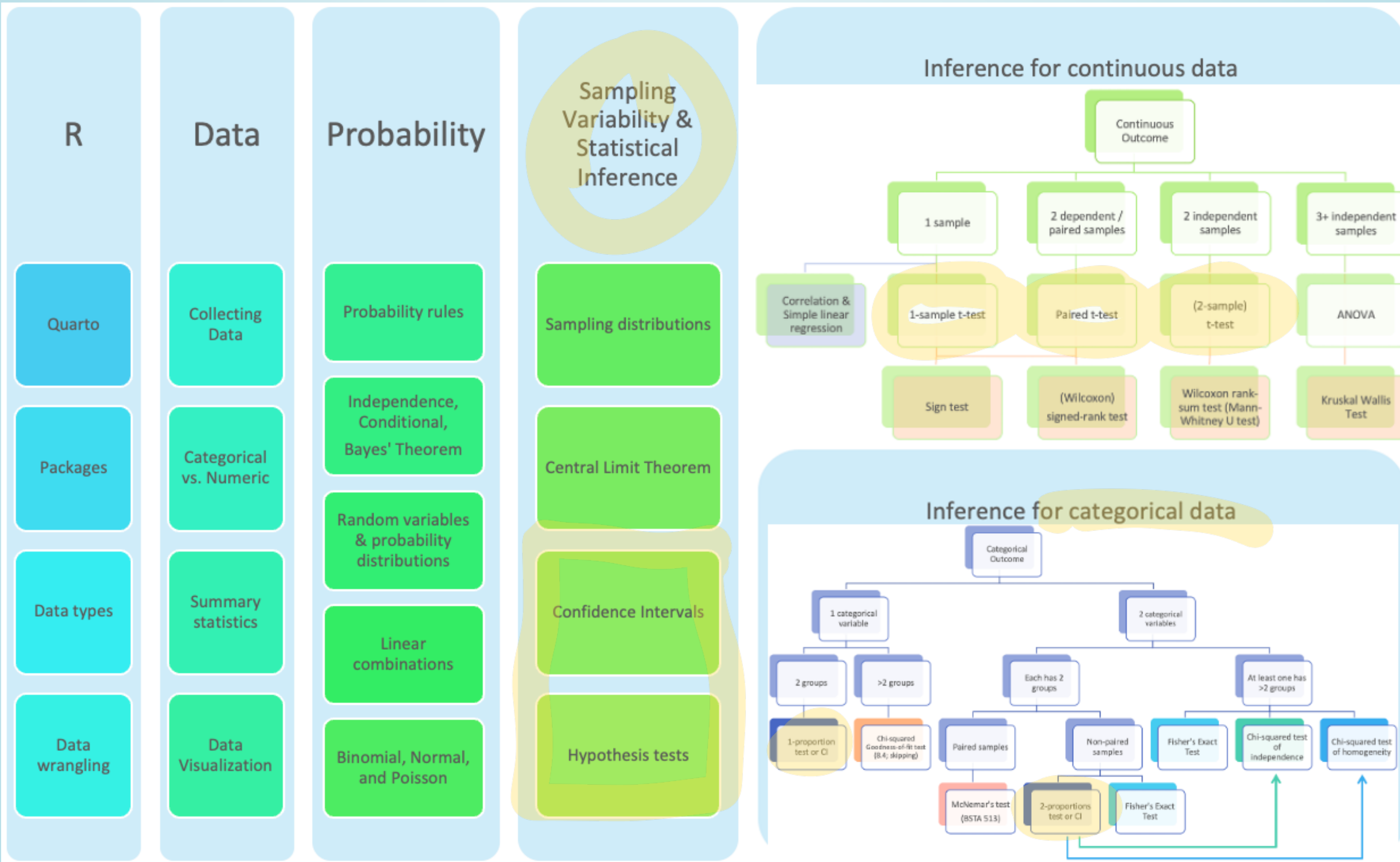
MoRitz's tip of the day

Add text to a plot using `annotate()`:

```r
ggplot(NULL, aes(c(0,4))) +  # no dataset, create axes for x from 0 to 4
  geom_area(stat = "function", fun = dchisq, args = list(df=2),
            fill = "blue", xlim = c(0, 1.0414)) +
  geom_area(stat = "function", fun = dchisq, args = list(df=2),
            fill = "violet", xlim = c(1.0414, 4)) +
  geom_vline(xintercept = 1.0414) +  # vertical line at x = 1.0414
  annotate("text", x = 1.1, y = .4, # add text at specified (x,y) coordinate
           label = "chi-squared = 1.0414", hjust=0, size=6) +
  annotate("text", x = 1.3, y = .1,
           label = "p-value = 0.59", hjust=0, size=6)
```

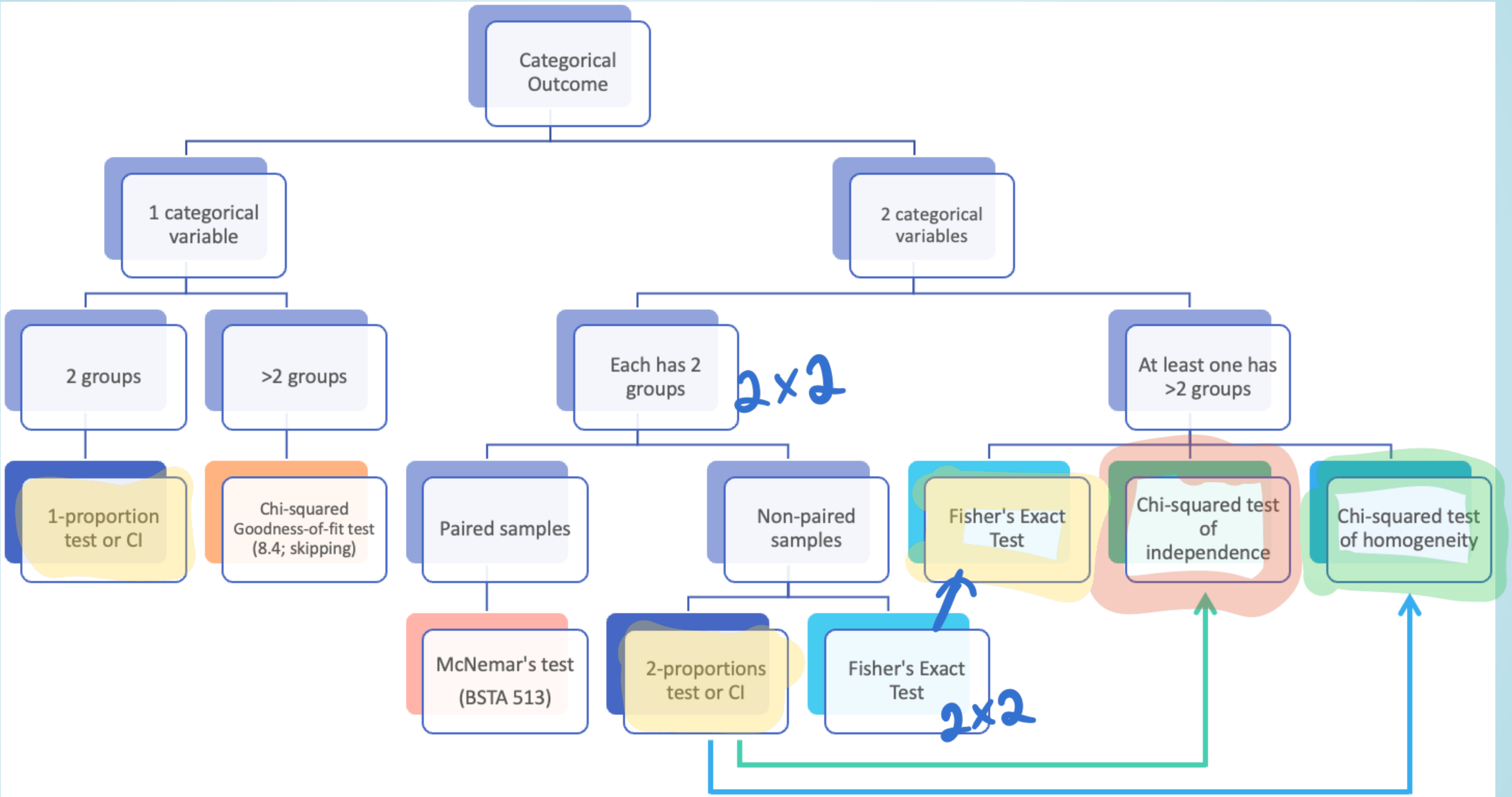


Where are we?



R
Quarto
Packages
Data types
Data wrangling
Data
Collecting Data
Categorical vs. Numeric
Summary statistics
Data Visualization
Probability
Probability rules
Independence, Conditional, Bayes' Theorem
Random variables & probability distributions
Linear combinations
Binomial, Normal, and Poisson
Sampling Variability & Statistical Inference
Sampling distributions
Central Limit Theorem
Confidence Intervals
Hypothesis tests
Inference for continuous data
Continuous Outcome
1 sample
2 dependent / paired samples
2 independent samples
3+ independent samples
Correlation & Simple linear regression
1-sample t-test
Paired t-test
(2-sample) t-test
ANOVA
Sign test
(Wilcoxon) signed-rank test
Wilcoxon rank-sum test (Mann-Whitney U test)
Kruskal Wallis Test
Inference for categorical data
Categorical Outcome
1 categorical variable
2 categorical variables
2 groups
>2 groups
Each has 2 groups
At least one has >2 groups
1-proportion test or CI
Chi-squared Goodness-of-fit test (8.4; skipping)
Paired samples
Non-paired samples
Fisher's Exact Test
Chi-squared test of independence
Chi-squared test of homogeneity
McNemar's test (BSTA 513)
2-proportions test or CI
Fisher's Exact Test

Where are we? Categorical outcome zoomed in



Goals for today (Sections 8.3-8.4)

- Statistical inference for **categorical data** when either are
  - comparing **more than two groups**,
  - or have categorical outcomes that have **more than 2 levels**,
  - or both
- Chi-squared tests of association (independence)
  - Hypotheses
  - test statistic
  - Chi-squared distribution
  - p-value
  - technical conditions (assumptions)
  - conclusion
  - R: `chisq.test()`
- Fisher's Exact Test
- Chi-squared test vs. testing difference in proportions
  - Test of Homogeneity

Chi-squared tests of association (independence)

Testing the association (independence) between two categorical variables

Is there an association between depression and being physically active?

- Data sampled from the NHANES R package:
  - American National Health and Nutrition Examination Surveys
  - Collected 2009-2012 by US National Center for Health Statistics (NCHS)
  - `NHANES` dataset: 10,000 rows, resampled from `NHANESraw` to undo oversampling effects
    - Treat it as a simple random sample from the US population (for pedagogical purposes)
- **`Depressed`**
  - Self-reported *number of days where participant felt down, depressed or hopeless*.
  - One of None, Several, or Most (more than half the days).
  - Reported for participants aged 18 years or older.
- **`PhysActive`**
  - *Participant does moderate or vigorous-intensity sports, fitness or recreational activities* (Yes or No).
  - Reported for participants 12 years or older.

Hypotheses for a Chi-squared test of association (independence)

correlation: between 2 continuous variables

Generic wording:

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 variables are independent
- H_A: The variables are not independent

For our example:

Test of "**association**" wording

- H_0: There is no association between depression and physical activity
- H_A: There is an association between depression and physical activity

Test of "**independence**" wording

- H_0: The variables depression and physical activity are independent
- H_A: The variables depression and physical activity are not independent

No symbols

For chi-squared test hypotheses we do not have versions using "symbols" like we do with tests of means or proportions.

Data from NHANES

- Results below are from
  - a random sample of 400 adults ($\geq$ 18 yrs old)
  - with data for both the depression `Depressed` and physically active (`PhysActive`) variables.

	Days with Depression			
Physical Activity	None	Several	Most	Total
Yes	199	26	1	226
No	115	32	27	174
Total	314	58	28	400

- What does it mean for the variables to be independent?

H_0: Variables are Independent

- Recall from Chapter 2, that events A and B are independent if and only if

$$P(A|B) = P(A)$$

$$P(A \text{ and } B) = P(A)P(B)$$

$$P(B|A) = P(B)$$

- If depression and being physically active are independent variables, then *theoretically* this condition needs to hold for *every combination of levels*, i.e.

$$
\begin{aligned}
P(None \text{ and } Yes) &= P(None)P(Yes) \\
P(None \text{ and } No) &= P(None)P(No) \\
P(Several \text{ and } Yes) &= P(Several)P(Yes) \\
P(Several \text{ and } No) &= P(Several)P(No) \\
P(Most \text{ and } Yes) &= P(Most)P(Yes) \\
P(Most \text{ and } No) &= P(Most)P(No)
\end{aligned}
$$

Days with Depression				
Physical Activity	None	Several	Most	Total
Yes	199	26	1	226
No	115	32	27	174
Total	314	58	28	400

$$P(None \text{ and } Yes) = \frac{314}{400} \cdot \frac{226}{400} \stackrel{?}{=} \frac{199}{400}$$

$$\cdots$$

$$P(Most \text{ and } No) = \frac{28}{400} \cdot \frac{174}{400} \stackrel{?}{=} \frac{27}{400}$$

With these probabilities, for each cell of the table we calculate the **expected** counts for each cell under the H_0 hypothesis that the variables are independent

Expected counts (if variables are independent)

- The expected counts (if H_0 is true & the variables are independent) for each cell are
  - np = total table size · probability of cell

Physical Activity	Days with Depression: None	Several	Most	Total
Yes	199	26	1	226
No	115	32	27	174
Total	314	58	28	400

Expected count of Yes & None:

$$\begin{aligned} 400 \cdot P(None\ and\ Yes) &= 400 \cdot P(None)P(Yes) \\ &= \cancel{400} \cdot \frac{314}{\cancel{400}} \cdot \frac{226}{400} \\ &= \frac{314 \cdot 226}{400} \\ &= 177.41 \\ &= \frac{\text{column total} \cdot \text{row total}}{\text{table total}} \end{aligned}$$

- If depression and being physically active are **independent** variables
  - (as assumed by H_0),
- then the **observed counts should be close to the expected counts** for each cell of the table

Observed vs. Expected counts

- The **observed** counts are the counts in the 2-way table summarizing the data

Physical Activity	Days with Depression			
	None	Several	Most	Total
Yes	199	26	1	226
No	115	32	27	174
Total	314	58	28	400

- The **expected** counts are the counts the we would expect to see in the 2-way table if there was no association between depression and being physically activity

Physical Activity	Days with Depression			
	None	Several	Most	Total
Yes	199 0.565*314 = 226/400*314 **177.41**	26 0.565*58 = 226/400*58 **32.77**	1 0.565*28 = 226/400*28 **15.82**	226 226/400 = 0.565
No	115 0.435*314 = 174/400*314 **136.59**	32 0.435*58 = 174/400*58 **25.23**	27 0.435*28 = 174/400*28 **12.18**	174 174/400 = 0.435
Total	314	58	28	400

Expected count for cell i, j :

$$\text{Expected Count}_{\text{row } i, \text{ col } j} = \frac{(\text{row } i \text{ total}) \cdot (\text{column } j \text{ total})}{\text{table total}}$$

The χ^2 test statistic

Test statistic for a test of association (independence):

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

- When the variables are independent, the observed and expected counts should be close to each other

Observed (Expected)	Days with Depression			
Physical Activity	None	Several	Most	Total
Yes	199 (177.41)	26 (32.77)	1 (15.82)	226
No	115 (136.59)	32 (25.23)	27 (12.18)	174
Total	314	58	28	400

$$\begin{aligned}\chi^2 &= \sum \frac{(O-E)^2}{E} \\ &= \frac{(199 - 177.41)^2}{177.41} + \frac{(26 - 32.77)^2}{32.77} + \ldots + \frac{(27 - 12.18)^2}{12.18} \\ &= 41.2\end{aligned}$$

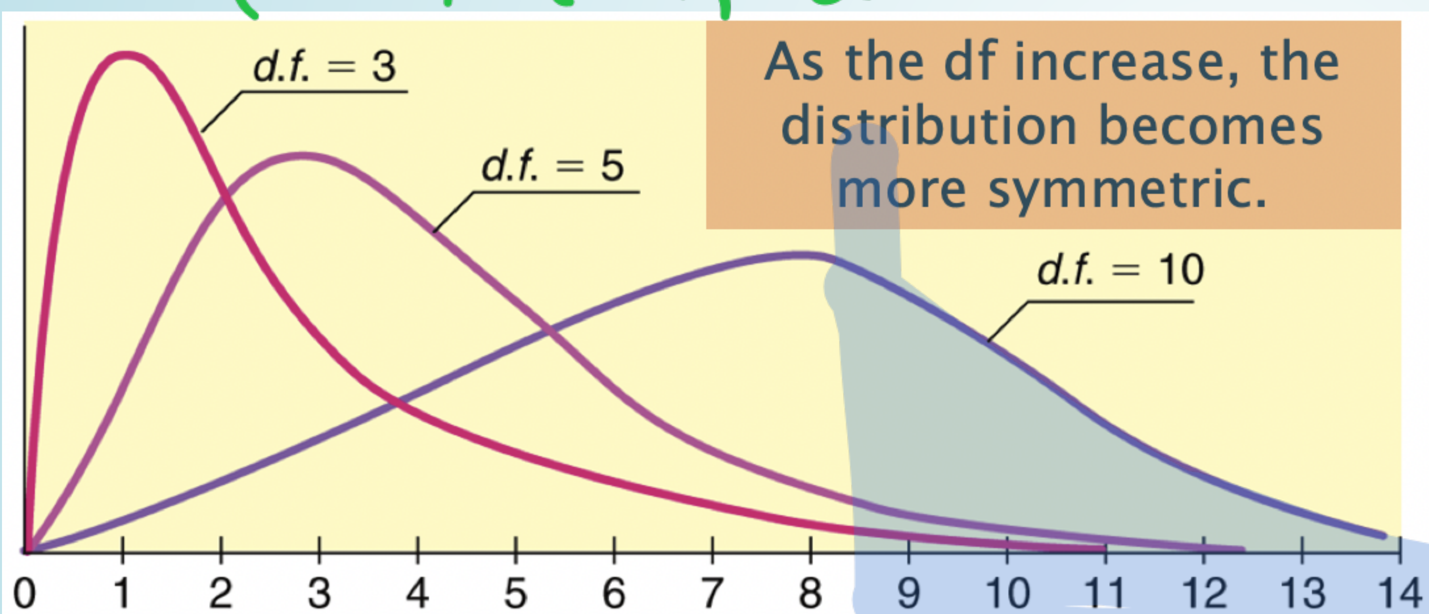
Is this value big? Big enough to reject H_0?

The χ^2 distribution & calculating the *p*-value

The χ^2 distribution shape depends on its degrees of freedom

- It's skewed right for smaller df,
  - gets more symmetric for larger df
- **df = (# rows-1) x (# columns-1)**

$(2-1) \times (3-1) = 2$



- The **p-value** is always the **area to the right** of the test statistic for a χ^2 test.
- We can use the `pchisq` function in R to calculate the probability of being at least as big as the χ^2 test statistic:

```
pv <- pchisq(41.2, df = 2,
             lower.tail = FALSE)
pv
```

```
[1] 1.131185e-09
```

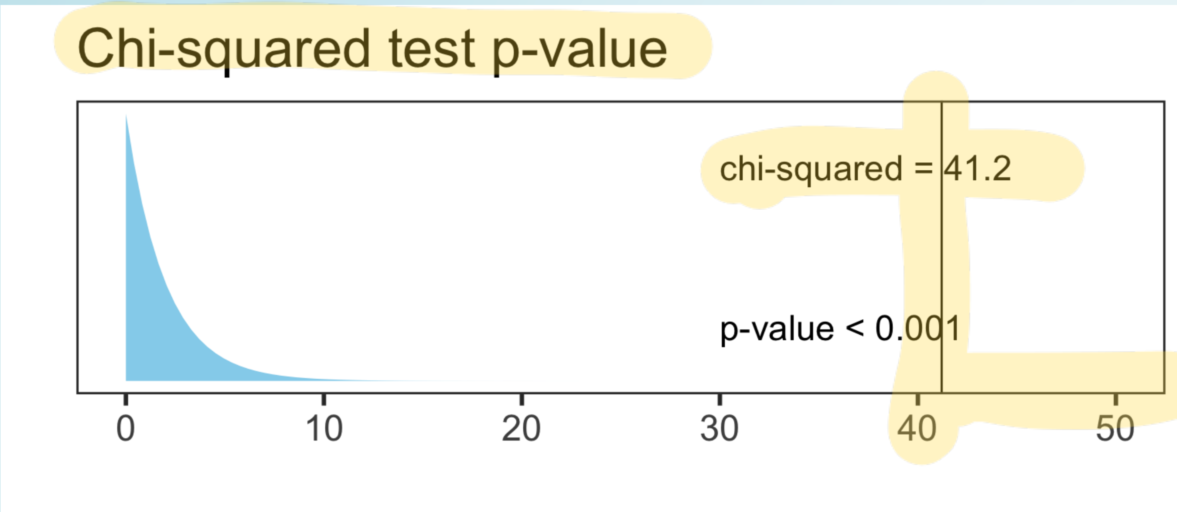
What's the conclusion to the χ^2 test?

OR 1-pchisq(41.2,df=2)

Conclusion

Recall the hypotheses to our χ^2 test:

- H_0: There is **no association** between depression and being physically activity
- H_A: There is **an association** between depression and being physically activity



Conclusion:

Based a random sample of 400 US adults from 2009-2012, there is sufficient evidence that there is an association between depression and being physically activity (p-value < 0.001).

Warning

If we fail to reject, we DO NOT have evidence of no association.

Do not say variables are independent!

Correct: insufficient evidence there is an association.
(not independent)

Technical conditions

- **Independence**
  - Each case (person) that contributes a count to the table must be independent of all the other cases in the table
    - In particular, observational units cannot be represented in more than one cell.
    - For example, someone cannot choose both “Several” and “Most” for depression status. They have to choose exactly one option for each variable.

- **Sample size**
  - In order for the distribution of the test statistic to be appropriately modeled by a chi-squared distribution we need
  - **2×2 table:** → df=1
    - **expected counts are at least 10 for each cell**
  - **larger tables:**
    - **no more than 1/5 of the expected counts are less than 5**, and } ≥5
    - **all expected counts are greater than 1** ≥2

Observed (Expected)	Days with Depression			
Physical Activity	None	Several	Most	Total
Yes	199 (177.41)	26 (32.77)	1 (15.82)	226
No	115 (136.59)	32 (25.23)	27 (12.18)	174
Total	314	58	28	400

Chi-squared tests in R

Depression vs. physical activity dataset

Create dataset based on results table:

```r
DepPA <- tibble(
  Depression = c(rep("None", 314),
          rep("Several", 58),
          rep("Most", 28)),
  PA = c(rep("Yes", 199),  # None
          rep("No", 115),
          rep("Yes", 26), # Several
          rep("No", 32),
          rep("Yes", 1), # Most
          rep("No", 27))
)
```

Observed (Expected)	**Days with Depression**			
Physical Activity	None	Several	Most	Total
Yes	199 (177.41)	26 (32.77)	1 (15.82)	226
No	115 (136.59)	32 (25.23)	27 (12.18)	174
Total	314	58	28	400

Summary table of data:

```r
DepPA %>%
  tabyl(Depression, PA)
```

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

```r
# base R:
table(DepPA)
```

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

χ^2 test in R using dataset

If only have 2 columns in the dataset:

```r
(ChisqTest_DepPA <-
   chisq.test(table(DepPA)))
```

```
	Pearson's Chi-squared test

data:  table(DepPA)
X-squared = 41.171, df = 2, p-value = 1.148e-09
```

If have >2 columns in the dataset, we need to specify which columns to table:

```r
(ChisqTest_DepPA <-
   chisq.test(table(
     DepPA$Depression, DepPA$PA)))
```

```
	Pearson's Chi-squared test

data:  table(DepPA$Depression, DepPA$PA)
X-squared = 41.171, df = 2, p-value = 1.148e-09
```

The tidyverse way (fewer parentheses)

```r
table(DepPA$Depression, DepPA$PA) %>%
  chisq.test()
```

```
	Pearson's Chi-squared test

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

`tidy()` the output (from `broom` package):

```r
table(DepPA$Depression, DepPA$PA) %>%
  chisq.test() %>%
  tidy() %>% gt()
```

statistic	p.value	parameter	method
41.17067	1.147897e-09	2	Pearson's Chi-squared test

Pull *p*-value

```r
table(DepPA$Depression, DepPA$PA) %>%
  chisq.test() %>%
  tidy() %>% pull(p.value)
```

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

Observed & expected counts in R

You can see what the **observed** and **expected** counts are from the saved chi-squared test results:

```
1  ChisqTest_DepPA$observed
```

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

```
1  ChisqTest_DepPA$expected
```

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

Observed (Expected)	Days with Depression			
Physical Activity	None	Several	Most	Total
Yes	199 (177.41)	26 (32.77)	1 (15.82)	226
No	115 (136.59)	32 (25.23)	27 (12.18)	174
Total	314	58	28	400

- Why is it important to look at the expected counts?
- What are we looking for in the expected counts?

χ^2 test in R with 2-way table

Create a base R table of the results:

```r
(DepPA_table <- matrix(c(199, 26, 1, 115, 32, 27), nrow = 2, ncol = 3, byrow = T))
```

```
     [,1] [,2] [,3]
[1,]  199   26    1
[2,]  115   32   27
```

```r
dimnames(DepPA_table) <- list("PA" = c("Yes", "No"),   # row names
                              "Depression" = c("None", "Several", "Most"))  # column names
DepPA_table
```

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

Run χ^2 test with 2-way table:

```r
chisq.test(DepPA_table)
```

```

	Pearson's Chi-squared test

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

```r
chisq.test(DepPA_table)$expected
```

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

(Yates') Continuity correction

- For a **2x2** contingency table,
  - the χ^2 test has the option of including a continuity correction
  - just like with the proportions test
- The **default includes a continuity correction**
- There is no CC for bigger tables

```r
(DepPA_table2x2 <- matrix(c(199, 27, 115, 59), nrow = 2, ncol = 2, byrow = T))
```

```
     [,1] [,2]
[1,]  199   27
[2,]  115   59
```

```r
dimnames(DepPA_table2x2) <- list("PA" = c("Yes", "No"),    # row names
                                 "Depression" = c("None", "Several/Most"))  # column names
DepPA_table2x2
```

```
     Depression
PA    None Several/Most
  Yes  199           27
  No   115           59
```

Output **without a CC**

```r
chisq.test(DepPA_table2x2, correct = FALSE)
```

```

	Pearson's Chi-squared test

data:  DepPA_table2x2
X-squared = 28.093, df = 1, p-value = 1.156e-07
```

Compare to output **with CC**:

```r
chisq.test(DepPA_table2x2)
```

```

	Pearson's Chi-squared test with Yates'
continuity correction

data:  DepPA_table2x2
X-squared = 26.807, df = 1, p-value = 2.248e-07
```

Fischer's Exact Test

Use this if expected cell counts are too small

Example with smaller sample size

- Suppose that instead of taking a random sample of 400 adults (from the NHANES data), a study takes a random sample of 100 such that
  - 50 people that are physically active and
  - 50 people that are not physically active

```r
(DepPA100_table <- matrix(c(43, 5, 2, 40, 4, 6), nrow = 2, ncol = 3, byrow = T))
```

```
     [,1] [,2] [,3]
[1,]   43    5    2
[2,]   40    4    6
```

```r
dimnames(DepPA100_table) <- list("PA" = c("Yes", "No"),    # row names
                                 "Depression" = c("None", "Several", "Most"))  # column names

DepPA100_table
```

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

Chi-squared test warning

```r
chisq.test(DepPA100_table)
```

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


	Pearson's Chi-squared test

data:  DepPA100_table
X-squared = 2.2195, df = 2, p-value = 0.3296
```

```r
chisq.test(DepPA100_table)$expected
```

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

     Depression
PA    None Several Most
  Yes 41.5     4.5    4
  No  41.5     4.5    4
```

- Recall the **sample size** condition
  - In order for the test statistic to be modeled by a chi-squared distribution we need
  - **2 $\times$ 2 table: expected counts are at least 10 for each cell**
  - **larger tables:**
    - **no more than 1/5 of the expected counts are less than 5**, and
    - **all expected counts are greater than 1**

Fisher’s Exact Test

- Called an exact test since it
  - calculates an exact probability for the p-value
    - instead of using an asymptotic approximation, such as the normal, t, or chi-squared distributions
  - For 2x2 tables the p-value is calculated using the **hypergeometric** probability distribution (see book for details)

```r
fisher.test(DepPA100_table)
```

```
	Fisher's Exact Test for Count Data

data:  DepPA100_table
p-value = 0.3844
alternative hypothesis: two.sided
```

Comments

- Note that there is no test statistic
- There is also no CI
- This is always a two-sided test
- There is no continuity correction since the hypergeometric distribution is discrete

Simulate p-values: another option for small expected counts

From the `chisq.test` help file:

- Simulation is done by random sampling from the set of all contingency tables with the same margin totals
    - works only if the margin totals are strictly positive.
- For each simulation, a χ^2 test statistic is calculated
- *P*-value is the proportion of simulations that have a test statistic at least as big as the observed one.
- No continuity correction

```r
set.seed(567)
chisq.test(DepPA100_table, simulate.p.value = TRUE)
```

```
	Pearson's Chi-squared test with simulated p-value (based on 2000
	replicates)

data:  DepPA100_table
X-squared = 2.2195, df = NA, p-value = 0.3893
```

χ^2 test vs. testing proportions

χ^2 test vs. testing differences in proportions

If there are only 2 levels in both of the categorical variables being tested, then the *p*-value from the χ^2 test is equal to the *p*-value from the differences in proportions test.

Example: Previously we tested whether the proportion who had participated in sports betting was the same for college and noncollege young adults:

$$H_0: p_{coll} - p_{noncoll} = 0$$
$$H_A: p_{coll} - p_{noncoll} \neq 0$$

```r
SportsBet_table <- matrix(
  c(175, 94, 137, 77),
  nrow = 2, ncol = 2, byrow = T)

dimnames(SportsBet_table) <- list(
  "Group" = c("College", "NonCollege"), # row r
  "Bet" = c("No", "Yes"))  # column names

SportsBet_table
```

```
            Bet
Group         No Yes
  College    175  94
  NonCollege 137  77
```

```r
chisq.test(SportsBet_table) %>% tidy() %>% gt()
```

statistic	p.value	parameter	method
0.01987511	0.8878864	1	Pearson's Chi-squared test with Yates' continuity correction

```r
prop.test(SportsBet_table) %>% tidy() %>% gt()
```

estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alternative
0.6505576	0.6401869	0.01987511	0.8878864	1	-0.07973918	0.1004806	2-sample test for equality of proportions with continuity correction	two.sided

```r
2*pnorm(sqrt(0.0199), lower.tail=F) # p-value
```

```
[1] 0.8878167
```

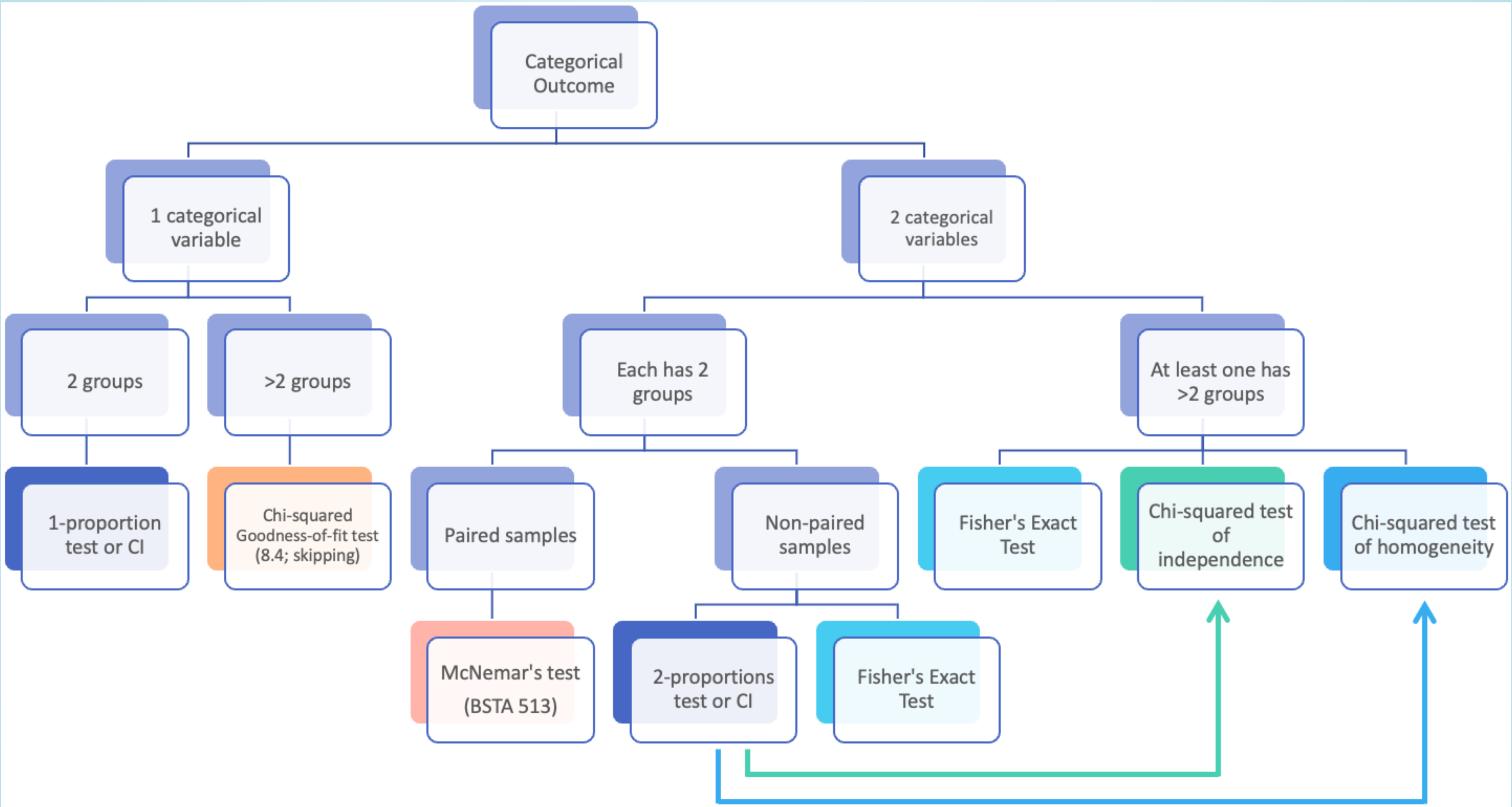
Test of Homogeneity

- Running the sports betting example as a chi-squared test is actually an example of a **test of homogeneity**
- In a test of homogeneity, proportions can be compared between many groups

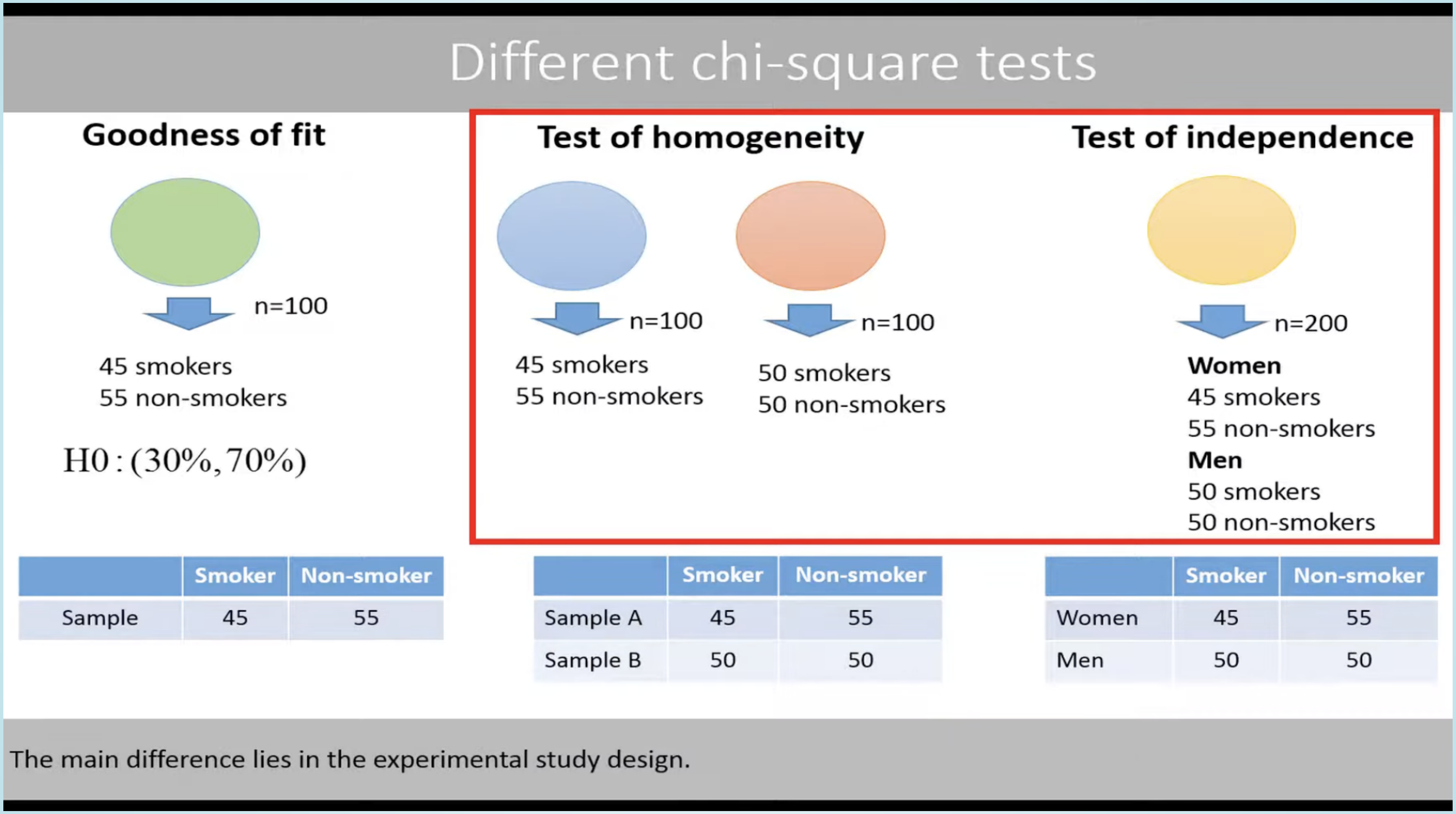
$$H_0 : p_1 = p_2 = p_2 = \ldots = p_n$$
$$H_A : p_i \neq p_j \text{for at least one pair of } i, j$$

- It’s an extension of a two proportions test.
- The test statistic & p-value are calculated the same was as a chi-squared test of association (independence)
- When we fix the margins (whether row or columns) of one of the “variables” (such as in a cohort or case-control study)
    - the chi-squared test is called a **Test of Homogeneity**

Overview of tests with categorical outcome

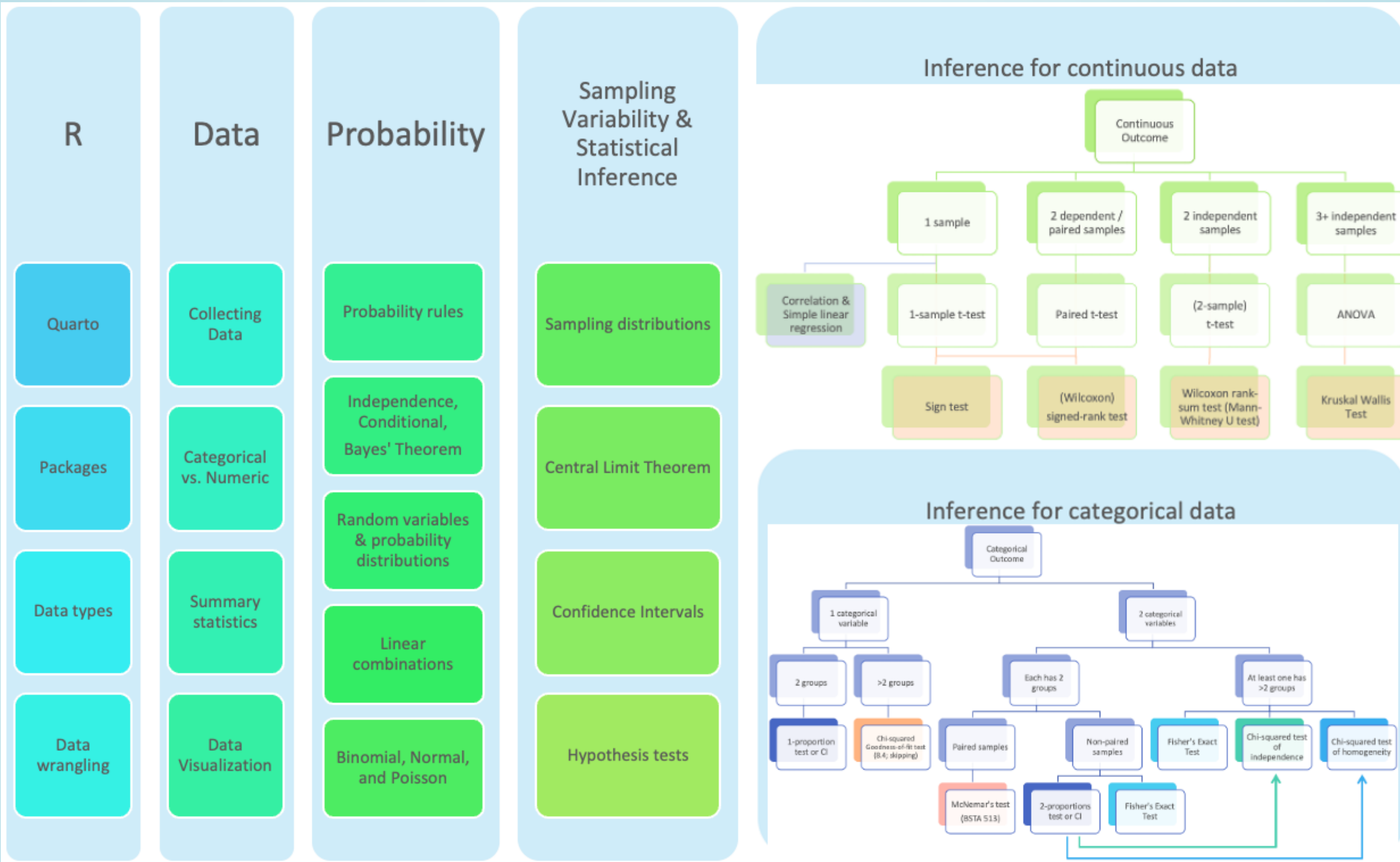


Chi-squared Tests of Independence vs. Homogeneity vs. Goodness-of-fit



- See YouTube video from TileStats for a good explanation of how these three tests are different: https://www.youtube.com/watch?v=TyD-_1JUhxw
- UCLA's INSPIRE website has a good summary too: http://inspire.stat.ucla.edu/unit_13/

What's next?



R
Quarto
Packages
Data types
Data wrangling
Data
Collecting Data
Categorical vs. Numeric
Summary statistics
Data Visualization
Probability
Probability rules
Independence, Conditional, Bayes' Theorem
Random variables & probability distributions
Linear combinations
Binomial, Normal, and Poisson
Sampling Variability & Statistical Inference
Sampling distributions
Central Limit Theorem
Confidence Intervals
Hypothesis tests
Inference for continuous data
Continuous Outcome
1 sample
2 dependent / paired samples
2 independent samples
3+ independent samples
Correlation & Simple linear regression
1-sample t-test
Paired t-test
(2-sample) t-test
ANOVA
Sign test
(Wilcoxon) signed-rank test
Wilcoxon rank-sum test (Mann-Whitney U test)
Kruskal Wallis Test
Inference for categorical data
Categorical Outcome
1 categorical variable
2 categorical variables
2 groups
>2 groups
Each has 2 groups
At least one has >2 groups
1-proportion test or CI
Chi-squared Goodness-of-fit test (8.4; skipping)
Paired samples
Non-paired samples
Fisher's Exact Test
Chi-squared test of independence
Chi-squared test of homogeneity
McNemar's test (BSTA 513)
2-proportions test or CI
Fisher's Exact Test