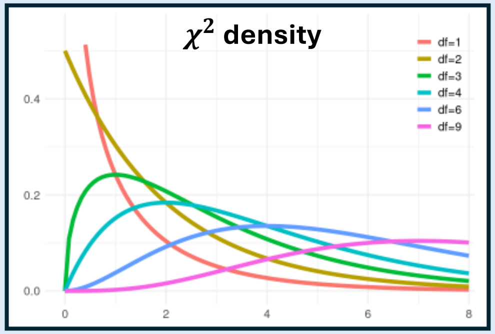
Chapter 11. Chi-Squared Tests (χ^2 Tests)

χ^2 distribution

- indexed by *"degrees of freedom"* written *"df" (like the t distribution)*
- result of summing ***squares*** of normal random variables
- **main example:** variance is χ^2
- **big idea:** if two things are equal, then difference has variance 0!



$\mu = df \qquad \sigma = \sqrt{2(df)}$

Uses:

Main use!

- Hypothesis tests on **variance** (**one-sample only**)
- *[Pearson's]* tests on **tables of counts** ("**contingency**" or "**frequency**" tables)

Requirement.
All expected counts ≥ 5

Usually **better** than z proportion test

Pearson tests on tables of counts *(Categorical variables)*

Goodness-of-fit

Compare **observed** frequency counts with **expected** distribution

Observed counts:

group	A	B	C	$\cdots$
#	x_A	x_B	x_C	$\cdots$

Expected counts:

group	A	B	C	$\cdots$
#	e_A	e_B	e_C	$\cdots$

Statistic: $\sum \frac{(x-e)^2}{e}$ is χ^2

degrees of freedom, df $=$ *# groups* -1

***p*-value:** right-tail probability

$H_A : \chi^2 > 0$ *(cannot be negative)*

Common application.
compare to **uniform** distribution
$e = \frac{n}{k}$, where $n =$ # observations
$k =$ # groups

Test for Independence

Check if **rows** and **columns** are **independent** in 2D table of overlapping counts

group	A	B	C	$\cdots$	Total
1	x_{A1}	x_{B1}	x_{C1}	$\cdots$	n_1
2	x_{A2}	x_{B2}	x_{C2}	$\cdots$	n_2
3	x_{A3}	x_{B3}	x_{C3}	$\cdots$	n_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$		
Total	n_A	n_B	n_C		N

Expected counts: $e = \frac{(n_{row})(n_{col})}{N}$

Statistic: $\sum \frac{(x-e)^2}{e}$ is χ^2

with df $= (\#\text{rows} - 1)(\#\text{cols} - 1)$

***p*-value:** right-tail probability

Test for Homogeneity

Check if **two frequency counts** came from **same** distribution

- "nonparametric" – do not need to know underlying distribution!
- equivalent to independence test with only two rows!

group	A	B	C	$\cdots$	Total
X	x_A	x_B	x_C	$\cdots$	n_X
Y	y_A	y_B	y_C	$\cdots$	n_Y
Total	n_A	n_B	n_C		N

Expected counts: $e = \frac{(n_{row})(n_{col})}{N}$

degrees of freedom, df $=$ # groups -1

***p*-value:** right-tail probability

Chapter 11. More Detail and Examples

Q: Why do **goodness-of-fit** test?
Why not just a bunch of t **proportion** tests?

A: Performing multiple tests is **BAD** due to **"α-inflation"**!!!

If you do **six** tests, each with **5%** probability of *Type I Error*, then the total probability of at least one *Type I Error* is $1-(.95)^6 \approx \mathbf{26\%}$!!!

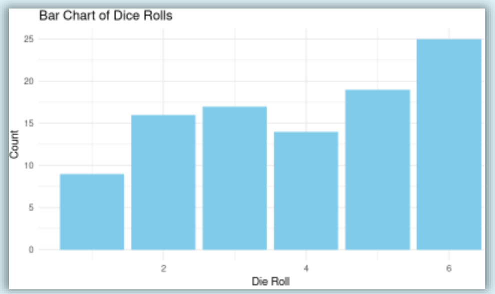
This is a type of **"*p-hacking*"**! *(...don't do it!!!)*

Example: Fair dice?

Experiment: Roll die 100 times and record results.

Data is table with 100 values (numbers 1,2,3,4,5,6)

	A	B
1	die roll	
2	6	
3	5	
4	2	
5	2	
6	6	
7	6	
8	3	



Expected: $\frac{100}{6} \approx 16$ rolls of each value

Proportion of 1's and 6's looks suspicious... but t-test on only those would have α-inflation!

Goodness of fit test against $p = \frac{1}{6}$ uniform distribution!

```
# Create a contingency table for the dice data
table(dice$`die roll`)

 1  2  3  4  5  6
 9 16 17 14 19 25
```

Roll	⚀	⚁	⚂	⚃	⚄	⚅
Count	9	16	17	14	19	25

```
# Perform chi-squared goodness-of-fit test
chisq.test(dice_counts$count, p = rep(1/6, 6))

   Chi-squared test for given probabilities

data:  dice_counts$count
X-squared = 8.48, df = 5, p-value = 0.1317
```

Fail to Reject Null Hypothesis

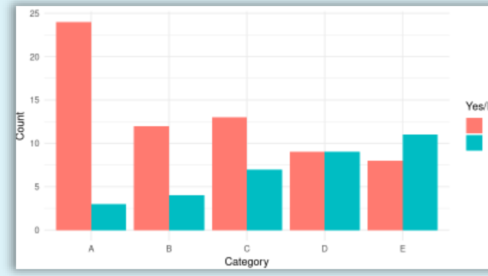
*Note: This die **was** actually fair.*

Example: Correlation in Survey Responses?

Experiment: Collect survey data yes/no value and category A-E.

Data is table with 100 rows containing survey responses.

	A	B
1	yes/no	category
2	no	C
3	yes	E
4	yes	E
5	no	B
6	no	D
7	no	C
8	no	A



It looks like people choosing ***no*** were more likely to pick ***A***; and people choosing ***yes*** were more likely to pick ***E***.

Homogeneity test to verify if yes/no influences category choice!

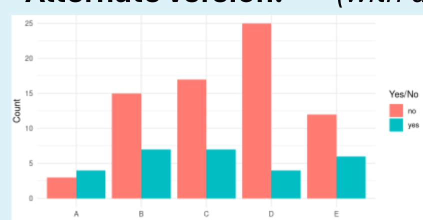
```
# Create a contingency table for the dep_cat data
table(dep_cat$`yes/no`, dep_cat$category)

      A  B  C  D  E
 no  24 12 13  9  8
 yes  3  4  7  9 11
```

```
# Perform a Chi-squared test for homogeneity
chisq.test(contingency_table_dep_cat)

   Pearson's Chi-squared test

data:  contingency_table_dep_cat
X-squared = 13.778, df = 4, p-value = 0.008039
```

Reject Null Hypothesis

Note: They were actually different!

Alternate version: *(with data from identical distributions....)*




```
Warning message in chisq.test(contingency_table):
"Chi-squared approximation may be incorrect"

   Pearson's Chi-squared test

data:  contingency_table
X-squared = 6.2816, df = 4, p-value = 0.1791
```

Fail to Reject Null Hypothesis

*Note: The warning is because there were $< \mathbf{5}$ values for both **no** and **yes** choosing **A***