Outline

1. Analysis of categorical data
   - Goodness or fit: fit of data to a claim
   - Test of independence
   - Assumptions (and applicability) of the method - Merging cells

Goodness or fit test

Example:

- A form of contamination in ground water is fecal coliforms. Officials claim that 20% wells are contaminated. Evaluate the claim.
- You randomly select 100 wells. 27 are contaminated.

Question: How do these data fit with the claim?

Hypotheses:

- H_0: claim is true, proportion $p = .20$ of contaminated wells.
- H_A: $p \neq .20$.

Test: evaluate the fit between claim and data.

Goodness or fit test

Observed counts:

contam.	clear	total
27	73	100

Expected counts under claim H_0:

contam.	clear	total
20	80	100

Build the table with expected counts: keep same total, use proportion(s) from the claim H_0.

Expected counts: $E_i = \text{Row total} * p_i = np_i$

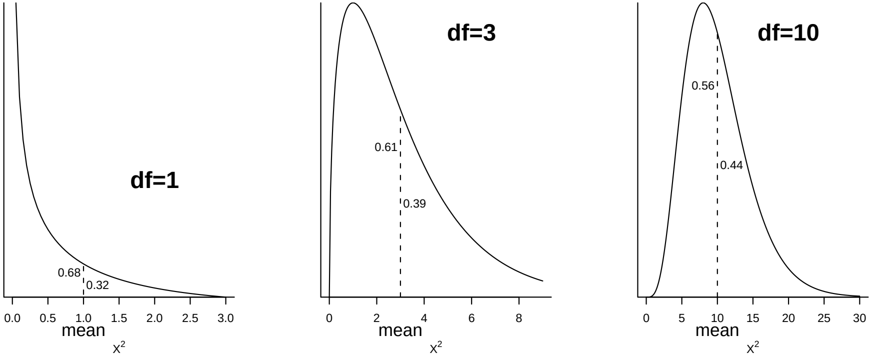
Test statistic:

$$X^2 = \sum_{\text{all cells}} \frac{(\text{obs} - \text{exp})^2}{\text{exp}}$$ (use counts, not proportions)

- If H_0 is true, X^2 tends to be small: it has a χ^2 distribution on 1 degree of freedom. df = # cells -1 in general.
- If H_A is true, X^2 will be bigger. More extreme = larger

The χ^2 distribution

- $X^2 \geq 0$ always
- $X^2 = 0$ when observed = expected counts: data in perfect agreement with the claim. X^2 close to 0: supports H_0.
- X^2 large: supports H_A.



benchmark: $X^2 \leq$ df supports H_0.

Goodness or fit test

- Here $X^2 = \dfrac{(27-20)^2}{20} + \dfrac{(73-80)^2}{80} = 3.0625$
- Use Table B to bracket the p-value: $.05 < p < .10$ at df=1. Or use R:

```
> 1 - pchisq(3.0625, df=1)
[1] 0.08061297
```

- There is only weak evidence that the claim $p = .20$ is false.

Same conclusion as with a z test. (remember?) We get $X^2 = z^2$ and exact same p-value.

Goodness or fit test: Assumptions

- Random sampling!
- The χ^2 distribution comes from the Normal approximation to the binomial. A large sample size is needed for this approximation to be good enough. What is large enough? Recall it meant $np \geq 5$ and $n(1-p) \geq 5$ before.

We need

expected counts ≥ 5 in all cells

for the chi-square distribution to be a good approximation to the exact distribution of X^2, and for the p-value to be correct. Or:

Expected counts ≥ 1 in all cells and ≥ 5 in at least 80% of the cells (less conservative).

A genetic example

Under a genetic model, a cross of white and yellow summer squash will yield a progeny with colors white, yellow and green with probabilities 12/16, 3/16 and 1/16.
Expected ratios are 12:3:1.

We have a total of 200 plants. Observations:

white	yelow	green	total
153	39	8	200

H_0: genetic model is true, i.e. $p_{\text{white}} = 12/16$ and $p_{\text{yellow}} = 3/16$.
H_A: the genetic model is not true (many different possibilities!)

A genetic example

Observed:

white	yelow	green	total
153	39	8	200

Expected under genetic model:

white	yelow	green	total
150	37.5	12.5	200

Expected counts: np_i in cell i. Here $n =$ total # plants.
Degree of freedom: # pieces of information needed to fill in the table (total is known from the design of the experiment)
Here df$=$ 2.
Test statistic and p-value:

$$X^2 = \frac{(153-150)^2}{150} + \frac{(39-37.5)^2}{37.5} + \frac{(8-12.5)^2}{12.5} = 1.74$$

With table: $p > .20$. With R:

```
> mycounts = c(153, 39, 8)
> chisq.test(mycounts, p=c(12/16, 3/16, 1/16))
      Chi-squared test for given probabilities
data:  mycounts
X-squared = 1.74, df = 2, p-value = 0.4190
```

A genetic example

Validity: expected counts (150, 37.5 and 12.5) are all ≥ 5. Good!

Conclusion: There is no evidence that the genetic model is false. The data are consistent with the genetic model (p=0.4). The difference between the data and the model can easily be due to sampling error.

Test of independence

We want to compare the performance of 2 drugs on rats.

	drug 1	drug 2	total
success	71	45	116
failure	34	42	76
total	105	87	192

$p_1 = \mathbb{P}\{\text{success} \,|\, \text{drug 1}\}$, probability of success with drug 1
$p_2 = \mathbb{P}\{\text{success} \,|\, \text{drug 2}\}$

He want to test H_0: **drugs perform equally**, i.e $p_1 = p_2$
against H_A: one drug is better than the other, i.e $p_1 \neq p_2$.

Equivalently, we have
H_0: drug and success are **independent**.
H_A: drug and success are not independent.

Test of independence

1. Build table of expected counts under H_0.

If H_0 is true, $p_1 = p_2$, but we don't know this value. So we estimate it. Best guess is

$$\hat{p} = \frac{\text{total \# successes}}{\text{total \# rats}} = \frac{116}{192} = .60$$

somewhat in between $\hat{p}_1 = \frac{71}{105} = .68$ and $\hat{p}_2 = \frac{45}{87} = .52$.

Expected # successes with drug 1: $105 * \hat{p} = 105 * \frac{116}{192}$.
In general,

$$E = \frac{\text{Row total * Column total}}{\text{Grand total}}$$

Test of independence

Observed counts:

	drug 1	drug 2	total
☺	71	45	116
☹	34	42	76
total	105	87	192

Expected counts when drug and success are independent:

	drug 1	drug 2	total
☺	63.44	52.56	116
☹	41.56	34.44	76
total	105	87	192

2. Calculate the test statistic X^2

$$
\begin{aligned}
X^2 &= \sum_{\text{all cells}} \frac{(\text{obs} - \text{exp})^2}{\text{exp}} \\
&= \frac{(71 - 63.44)^2}{63.44} + \cdots + \frac{(42 - 34.44)^2}{34.44} \\
&= 5.026
\end{aligned}
$$

Test of independence

3. **Calculate the p-value.** If there is independence (success does not depend on drug) then X^2 has a χ^2 distribution with df$= 1$ here.
   Using Table B, we get $.02 < p < .05$.
4. **Conclusion:** There is moderate evidence that the drugs have different success rates ($p = 0.025$, chi-square test of independence).
   Furthermore, in the data we have $\hat{p}_1 = .68 > \hat{p}_2 = .52$.
   There is evidence that drug 1 has a higher success rate.

Same conclusion as a z test (remember?) for testing $p_1 = p_2$.
We have $X^2 = z^2$ and exact same p-value.

Using R: chisq.test()

```
> rats = matrix( c(71,34,45,42),  2,2)
> rats
     [,1] [,2]
[1,]   71   45
[2,]   34   42
> chisq.test(rats)
       Pearson’s Chi-squared test with
       Yates’ continuity correction
data:  rats
X-squared = 4.3837, df = 1, p-value = 0.03628

> chisq.test(rats, correct=FALSE)
        Pearson’s Chi-squared test
data:  rats
X-squared = 5.0264, df = 1, p-value = 0.02496
```

χ^2 test with a bigger table

Wisconsin corn farmers: does use of pesticides depend on age?

	Observed counts		
Response	never use	use sparingly	use as needed
25 and under	15	15	6
26-40	39	73	28
41-55	46	90	41
56 and over	25	59	29

H_0: Response to pestidice use and age are independent
H_A: They are not!

Expected values

	never use	use sparingly	use as needed	total
≤ 25	15	15	6	36
26-40	39	73	28	140
41-55	46	90	41	177
≥ 56	25	59	29	113
total	125	237	104	466

X^2, degree of freedom and p-value

$$X^2 = \sum_{\text{all cells}} \frac{(\text{obs} - \text{exp})^2}{\text{exp}} = \underbrace{\frac{(15 - 9.7)^2}{9.7} + \cdots + \frac{(29 - 25.2)^2}{25.2}}_{\text{12 cells, so 12 terms}}$$

$$= \mathbf{6.149}$$

Degree of freedom: # pieces of information (cells) needed to fill in entire table. Marginals (totals in the margins) are known.

$$\text{df} =$$

Here df$=$ 6.

p-value: From Table B, we get $.25 < p < .50$. There is no evidence that pesticide use is dependent upon farmer's age. No association.

Applicability of the method

- Random samples
- Expected counts of cells ≥ 1 in all cells, and ≥ 5 in at least 80% cells for the χ^2 distribution to be a good approximation. If all expected counts ≥ 5, it's even better.

If some cells have small counts, what can be done?

- Fisher's exact test: but we won't cover this method.
- Group cells together. Pesticide use example with a 10-fold decrease in sample size.

Grouping cells

	never use	use sparingly	use as needed	total
≤ 25	2 (1.10)	1 (1.96)	1 (.94)	4
26-40	4 (3.87)	7 (6.85)	3 (3.28)	14
41-55	5 (4.98)	9 (8.81)	4 (4.21)	18
≥ 56	2 (3.05)	6 (5.38)	3 (2.57)	11
total	13	23	11	47

1 cell with E$<$ 1 and 9 cells (out of 12) have E$<$ 5. Chi^2 test not to be trusted.

Can we merge cells to increase observed counts and expected counts in turn?

Grouping cells

	never use	use sparingly	use as needed	total
≤ 40	6 ( )	8 ( )	4 ( )	18
≥ 41	7 ( )	15 ( )	7 ( )	29
total	13	23	11	47

Now all expected counts are > 1 and only 2 are < 5, although close to 5. We get df$=$ 2, $X^2 = 0.477$ and $.75 < p < .90$.

All ages seem to make the same use of pesticides: no evidence of change before/after age of 40.

Smoking cessation example

	no contact	group counseling	total
☺ quit smoking	1	26	27
☹ resumed within a year	30	69	99
Total	31	95	126

$p_1 = \mathbb{P}\{☺ |$ no contact$\}$ and $p_2 = \mathbb{P}\{☺ |$ group counseling$\}$

$\hat{p}_1 = \frac{1}{31} = .032 \quad < \quad \hat{p}_2 = \frac{26}{95} = .274.$

Question: Does group counseling increase the chances of quitting for longer than a year? Does success rate depend on counseling strategy?

Smoking cessation example

Observed and expected counts:

	no contact	group counseling	total
☺ quit smoking	1 (6.64)	26 (20.36)	27
☹ resumed within a year	30 (24.36)	69 (74.64)	99
Total	31	95	126

Is a chi-square test of independence okay to use here? Is it valid/applicable?

Smoking cessation example

We get $X^2 = 8.09$, with df $= 1$ so $.001 < p < .01$ and we conclude that there is strong evidence that success is dependent on counseling strategy ($p_1 \neq p_2$). Since the data show $\hat{p}_1 < \hat{p}_2$, there is strong evidence that $p_1 < p_2$, i.e. better success rate with group counseling.

Exact p-value in R:

```
> 1-pchisq(8.09, df=1)
[1] 0.004451016
```