Day 12

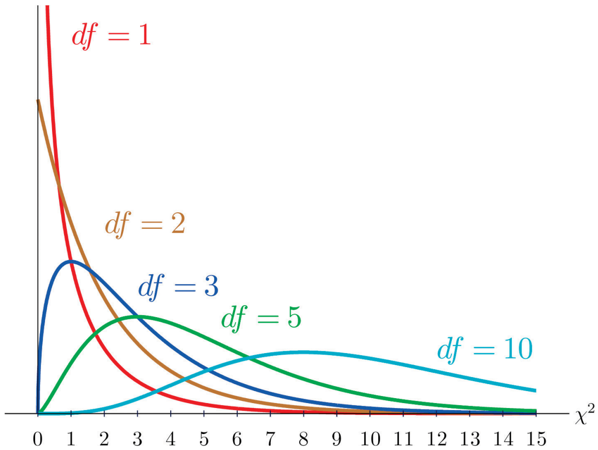
Outline

1. Computing p-values with software using χ^2 distraction
2. Test of independence
3. Measures of association between two categorical variables

Computing P-Value

Given χ^2

P-Value is always above or equal to degrees of freedom.



p ___

- interval values → probability

q ___

- probability → values

P-Values = probability of getting our data or data that "disagree" as much or more with our model, if model is correct.

```
arbitrary_val <- 4.8
pchisq(arbitrary_val, df = 5, lower.tail = FALSE)
```

Test of Independence

What I'm recording: two categorical variables

What I want to know: whether a suspected association between the variables will hold when generalized to the population.

Test of Homogeneity

What I'm recording: 1 categorical variable in samples form multiple populations

What I want to know: Is the variable's distribution the same in all populations?

Both tests use data summarised in two-way tables

We use Fisher's significance testing approach.

Test of independence: we model assuming that the two variables are not actually associated

H_0

- [Variable 1] does not affect [Variable 2]
  - [Variable 1] and [Variable 2] are independent/not associated/ not related

Testing of homogeneity: we model assuming the distribution is the same in every population

- H_0: the distribution of [variable] is the same in [list of population]

More simply put: H_a: not H_0

In test of homogeneity, we consider "population" to be an explanatory variable & run a test of independence

Observed counts = number in sample of each cell of table.

Example (Book Example 9.12)

	Low Salt	High Salt	Total
CVD			200
NO CVD			2215
Total	1169	1246	2415

Figure 1: Base Table

Estimated probability of Cardiovascular Disease(CVD) = $\frac{200}{2415}$

If independent (according to chart):

- $P(\text{CVD} \mid \text{low salt}) = 1169 \times \frac{200}{2415} = 96.81$
- $P(\text{CVD} \mid \text{high salt}) = 1246 \times \frac{200}{2415} = 103.19$
- $P(\text{NO CVD} \mid \text{low salt}) = 1169 \times \frac{2215}{2415} = 1072.19$
- $P(\text{NO CVD} \mid \text{high salt}) = 1246 \times \frac{2215}{2415} = 1142.81$

Pearson Residuals

$\frac{O-E}{\sqrt{E}} \rightarrow$ for each cell

Contribution of a cell to χ^2: residual^2 $= \frac{(O-E)^2}{E}$

$\chi^2 = \Sigma \frac{(O-E)^2}{E}$

- P(CVD | low salt) $= \frac{88-96.81}{\sqrt{96.81}} = -0.895$
- P(CVD | high salt) $= \frac{112-103.19}{\sqrt{103.19}} = 0.867$
- P(NO CVD | low salt) $= \frac{1081-1072.19}{\sqrt{1072.19}} = 0.269$
- P(NO CVD | high salt) $= \frac{1134-1142.81}{\sqrt{1142.81}} = -0.261$

$\chi^2 = (-0.895)^2 + (0.867)^2 + (0.269)^2 + (-0.261)^2 = 1.69$

finish second chart from picture

To get a P-Value

- Option 1: Our $\chi^2{}_{\text{observed}}$ value comes from a χ^2 distribution with degrees of freedom. Find $P(\chi^2 \geq \chi^2{}_{\text{observed}})$
- Option 2: Simulate a bunch of samples assuming independence, then find proportion of simulated χ^2 statistic $\geq \chi^2{}_{\text{observed}}$

Fisher: df (degrees of freedom) = (r - 1)(c -1)

- r: rows
- c: columns

"Sample size assumptions" method (2) always works but different people can get different values.

Method 1 always gives some value, but that value can be inaccurate at small sample sizes.

When all expected counts ≥ 5, use method 1.

When any expected count < 5, use method 2.

Alternate method when `n` is really small: Fisher's exact test

Condition on marginal totals being fixed, get a test statistic with hypergeometric distribution.

P-Value = $P(\chi^2 \geq 1.69)$ from χ^2 distribution with 1 degree of freedom = 0.193

Not on test but may show up in context:

3 "Measures of association" between categorical variables

1. Difference in proportions

- Population: P_1 - P_2
- Samples: $\hat{P}_1 - \hat{P}_2$

2. Relative risk (RR)

- Population : $\frac{P_1}{P_2}$