Lecture 13

Hypothesis tests for categorical data

Fisher’s exact test

- Ronald Fisher offered lady Muriel Bristol, a cup of tea.
- She declined after watching Fisher prepare it, saying that she preferred the taste when the milk was poured in the cup first.
- Fisher and others scoffed at this and a colleague, William Roach, suggested a test.

- 4 cups with milk poured first, 4 cups with milk poured after.
- Otherwise the cups were the same (temperature, appearance etc).

Fisher's exact test

- The lady is offered the tea, and for every cup she guesses:
  - Milk first (MF) or Tea first (TF)

		Guess		
		MF	TF	Total
Prep	MF	4	0	4
	TF	0	4	4
	Total	4	4	8

Contingency table

Fisher’s exact test

- The lady is offered the tea, and for every cup she guesses:
  - Milk first (MF) or Tea first (TF)

Once you fix one of the values, all the rest are fixed because the marginals are fixed

		Guess		
		MF	TF	Total
Prep	MF	4	0	4
	TF	0	4	4
	Total	4	4	8

Contingency table

Fisher's exact test

- The lady is offered the tea, and for every cup she guesses:
  - Milk first (MF) or Tea first (TF)

- H_0: The lady has no ability of distinguishing the method of preparation (the woman selects randomly).
- x: The number of MF she got right.
- P-value: The probability of observing data at least as extreme (unfavorable to H_0) under the null hypothesis.

		Guess: MF	Guess: TF	Total
Prep	MF	4	0	4
Prep	TF	0	4	4
	Total	4	4	8

Contingency table

Fisher's exact test

- The lady is offered the tea, and for every cup she guesses:
  - Milk first (MF) or Tea first (TF)

- H_0: The lady has no ability of distinguishing the method of preparation (the woman selects randomly).
- x: The number of MF she got right.
- P-value: The probability of observing data at least as extreme (unfavorable to H_0) under the null hypothesis.

- $P\ (X \geq\ x|H_0)$

		Guess: MF	Guess: TF	Total
Prep	MF	4	0	4
	TF	0	4	4
	Total	4	4	8

Contingency table

$P(X = 4|H_0)$

Fisher's exact test

- H_0: The lady has no ability of distinguishing the method of preparation (the woman selects randomly).
- x: The number of MF she got right.
- P-value: The probability of observing data at least as extreme (unfavorable to H_0) under the null hypothesis.

- Under the null hypothesis, the lady picks 4 cups at random, without replacement, from a population of 4 MF and TF cups
- X: number of MF cups
- $X \sim Hypergeometric(N, K, n)$
  - N is the population size
  - K is the number of success states in the population
  - n is the number of draws
- P(X=x) = $\frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$

		Guess		
		MF	TF	Total
Prep	MF	4	0	4
	TF	0	4	4
	Total	4	4	8

Contingency table

$P(X = 4|H_0)$

Fisher's exact test

- Under the null hypothesis, the lady picks 4 cups at random, without replacement, from a population of 4 MF and TF cups
- X: number of MF cups
- $X \sim Hypergeometric(N, K, n)$
  - N is the population size
  - K is the number of success states in the population
  - n is the number of draws
- P(X=x) = $\frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$

For $X \sim \; Hypergeometric(8, 4, 4)$

- P(X=0)=1/70
- P(X=1) = 16/70
- P(X=2) = 36/70
- P(X=3) = 16/70
- P(X=4) =1/70

		Guess: MF	Guess: TF	Total
Prep	MF	4	0	4
	TF	0	4	4
	Total	4	4	8

Contingency table

$$P(X = 4|H_0) = \frac{1}{70} = 0.014$$

Fisher’s exact test

- Under the null hypothesis, the lady picks 4 cups at random, without replacement, from a population of 4 MF and TF cups
- X: number of MF cups
- $X \sim Hypergeometric(N, K, n)$
  - N is the population size
  - K is the number of success states in the population
  - n is the number of draws
- P(X=x) = $\frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$

For $X \sim \ Hypergeometric(8, 4, 4)$

- P(X=0)=1/70
- P(X=1) = 16/70
- P(X=2) = 36/70
- P(X=3) = 16/70
- P(X=4) =1/70

		Guess		
		MF	TF	Total
Prep	MF	3	1	4
	TF	1	3	4
	Total	4	4	8

Contingency table

$$P(X = 3|H_0) + P(X = 4|H_0) = \frac{16}{70} + \frac{1}{70} = 0.242$$

The χ^2 test

- Assume that you have a large population of items of k different types, and let p_i denote the probability of an item selected at random will be of type $i = 1, \ldots, k$

- Let $p_1^o, \ldots, p_k^o$ be numbers such that $p_i^o > 0 \; \sum p_i^o = 1$
- We want to test the hypothesis:
  - H_0: $p_i = p_i^o \; \forall \; i$ vs
  - H_1: $p_i \neq p_i^o$ *for at least one i*
- *Assume we have a data set of n observations, and N_i is the number of observations of type i.*
- *The expected number of observations of type i under the null hypothesis is np_i^0*
- *Define the statistic* $Q = \sum_{i=1}^{k} \frac{(N_i - np_i^o)^2}{np_i^0}$
- *Under the null, when $n \to \infty$ $Q \sim \chi^2$ with k-1 degrees of freedom.*

Example: Independence

- You have a population of 520 people
  - 160/520 smoke.
  - 210/520 have CVD.

		CVD		
		Y	N	Total
Smoking	Y	120	40	160
	N	90	270	360
	Total	210	310	520

Contingency table

Example: Independence

Null Hypothesis ($\mathbf{H}_0$) : Smoking is independent of CVD

Alternative Hypothesis ($\mathbf{H}_1$) : Smoking is dependent of CVD

Mathematically:

$$\mathbf{H}_0 = \forall\, i,j \;\; p_{ij} = p_{i+} \times p_{+j}$$

$$\mathbf{H}_1 = \exists\, i,j\!: \;\; p_{ij} \neq p_{i+} \times p_{+j}$$

	CVD=0	CVD=1	
S=0	p_{00}	p_{01}	p_{0+}
S=1	p_{10}	p_{11}	p_{1+}
	p_{+0}	p_{+1}	1

$$p_{ij} = P(X = i,\; Y = j)$$
$$p_{i+} = P(X = i)$$
$$p_{+j} = P(Y = j)$$

Reminder: Independence:

$$\forall x, y\; P(\mathrm{Y} = \mathrm{y}, \mathrm{X} = \mathrm{x}) = P(Y = \mathrm{y})\mathrm{P}(\mathrm{X} = \mathrm{x})$$

Statistical Dependence

Smoking \ CVD	Y	N	Total
Y	120	40	160
N	90	270	360
Total	210	310	520

Contingency table

Smoking \ CVD	Y	N	Total
Y	.75	.25	1
N	.25	.75	1

Conditional Probability Distribution

$P(CVD|Smoking)$

Smoking \ CVD	Y	N	Total
Y	.2308	.0769	.3077
N	.1731	.5192	.6923
Total	.4038	.5962	1

Joint Probability Distribution

$P(CVD, Smoking)$

Smoking \ CVD	Y	N
Y	.5714	.1290
N	.4286	.8710
Total	1	1

Conditional Probability Distribution

$P(Smoking|CVD)$

Statistical Dependence

Smoking \ CVD	Y	N	Total
Y	120	40	160
N	90	270	360
Total	210	310	520

Contingency table

Smoking \ CVD	Y	N	Total
Y	.75	.25	1
N	.25	.75	1

Conditional Probability Distribution

$P(CVD|Smoking)$

P(Smoking)

Smoking \ CVD	Y	N	Total
Y	.2308	.0769	.3077
N	.1731	.5192	.6923
Total	.4038	.5962	1

Joint Probability Distribution

$P(CVD, Smoking)$

P(Smoking|CVD=yes)

Smoking \ CVD	Y	N
Y	.5714	.1290
N	.4286	.8710
Total	1	1

Conditional Probability Distribution

$P(Smoking|CVD)$

P(Smoking)≠P(Smoking|CVD=yes)

Test statistic: Expected counts

		CVD: Y	CVD: N	Total
Smoking	Y	.2308	.0769	.3077
	N	.1731	.5192	.6923
	Total	.4038	.5962	1

in your data

		CVD: Y	CVD: N	Total
Smoking	Y			.3077
	N			.6923
	Total	.4038	.5962	1

If Smoking and CVD were independent?

Are Smoking and CVD independent?

		CVD		
		Y	N	Total
Smoking	Y	.2308	.0769	.3077
	N	.1731	.5192	.6923
	Total	.4038	.5962	1

in your data

		CVD		
		Y	N	Total
Smoking	Y			.3077
	N			.6923
	Total	.4038	.5962	1

If Smoking and CVD were independent?

$$\boxed{P(Smoking = Yes, CVD = Yes) = P(Smoking = Yes) * P(CVD = Yes)}$$

Are Smoking and CVD independent?

		CVD Y	CVD N	Total
Smoking	Y	.2308	.0769	.3077
	N	.1731	.5192	.6923
	Total	.4038	.5962	1

in your data

		CVD Y	CVD N	Total
Smoking	Y			.3077
	N			.6923
	Total	.4038	.5962	1

If Smoking and CVD were independent?

$$\boxed{P(Smoking = Yes, CVD = Yes) = P(Smoking = Yes) * P(CVD = Yes) = 0.4038 * 0.3077}$$

Are Smoking and CVD independent?

		CVD: Y	CVD: N	Total
Smoking	Y	.2308	.0769	.3077
	N	.1731	.5192	.6923
	Total	.4038	.5962	1

in your sample

		CVD: Y	CVD: N	Total
Smoking	Y	.1242	.1835	.3077
	N	.2796	.4127	.6923
	Total	.4038	.5962	1

If Smoking and CVD were independent?

Are Smoking and CVD independent?

		CVD Y	CVD N
Smoking	Y	120	40
	N	90	270

counts in your data

		CVD Y	CVD N
Smoking	Y	65	95
	N	145	215

Expected counts If Smoking and CVD were independent

$$\boxed{P(Smoking = Yes, CVD = Yes) * \#\ samples = .1242*52}$$

- n_{ij}: Counts in your data (# observations in cell i,j)
- e_{ij}: Expected counts under H_0
- Summarize the difference of n_{ij} from e_{ij} for all i, j.

X^2 statistic:

$$t = \sum_{i,j} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

What is the probability of observing a value t at least as extreme as the one you observed in your data?

p-value: $P(|T| > |t_{obs}| | \mathrm{H}_0)$

Theoretical distribution of t under the null hypothesis

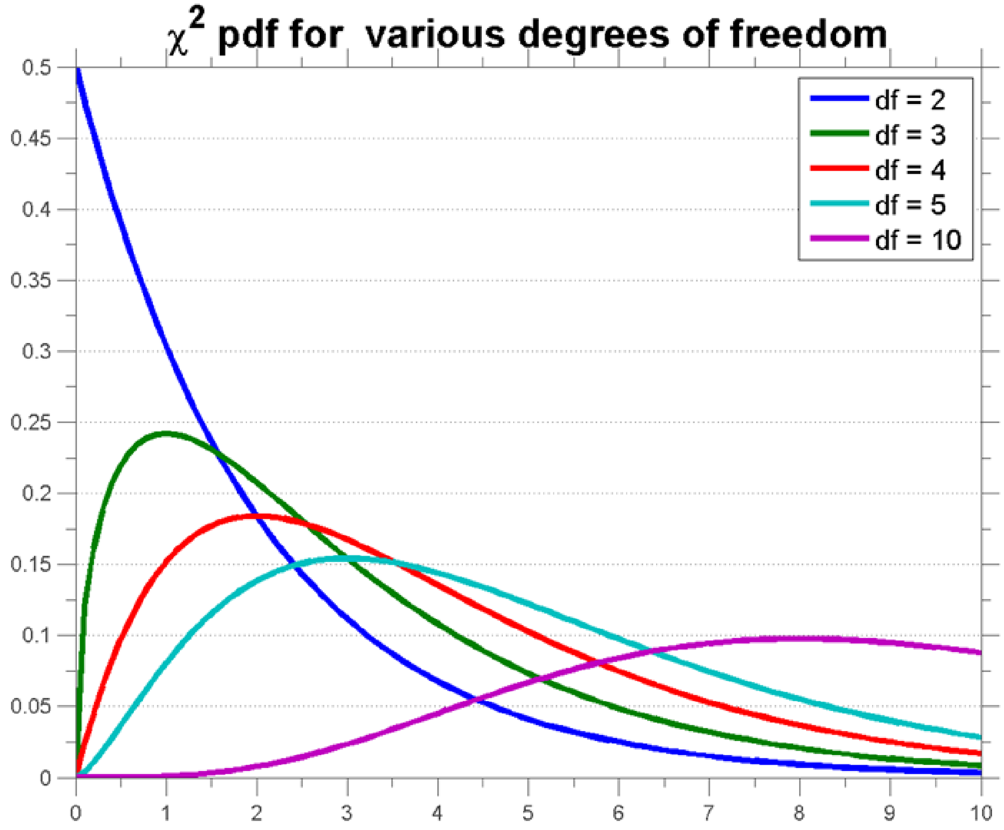
$$P(T = t|H_0) \sim \frac{t^{\frac{df-2}{2}} e^{-\frac{t}{2}}}{2^{\frac{df}{2}} \Gamma\left(\frac{df}{2}\right)},$$

where df are the degrees of freedom, i.e. the number of parameters that are free to vary

For testing X ⫫ Y

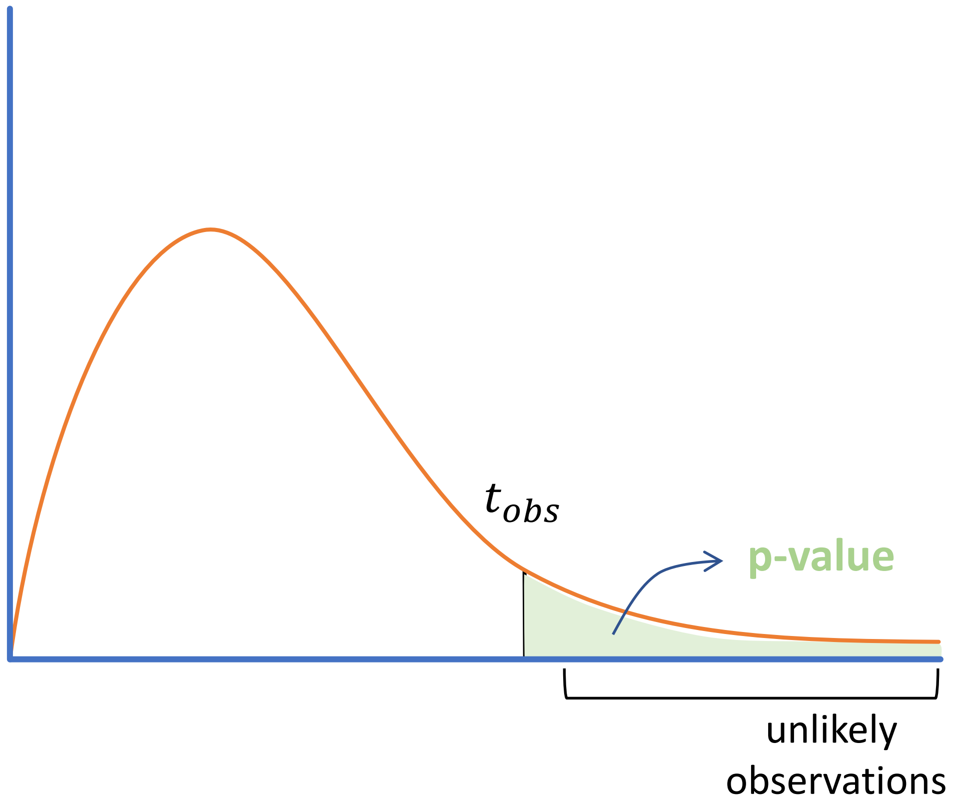
$df = (\#\text{ possible values of } X - 1) \times$
$(\#\text{ of possible values of } Y - 1)$

in our example $df = (2-1) \times (2-1) = 1$



- Check in the pdf
- If the p-value is less than a significance threshold α, reject the null hypothesis.

p-value: $P(|T| > |t_{obs}||\mathrm{H}_0)$



Permutation testing

- What if you do not know the distribution.
- Use permutation to estimate the distribution of t under H_0

Sample (Person)	Smoking	CVD
1	Yes	Yes
2	No	No
3	Yes	Yes
4	No	No
5	Yes	No
6	No	Yes

Sample (Person)	Smoking	CVD
1	Yes	No
2	No	Yes
3	Yes	No
4	No	Yes
5	Yes	No
6	No	Yes

Sample (Person)	Smoking	CVD
1	Yes	No
2	No	No
3	Yes	No
4	No	Yes
5	Yes	Yes
5	No	Yes

Under the null, the columns in your data are independent

Matrices with permuted rows for one of your variables are equally probable (given H_0).

Re-sampling techniques

Sample (Person)	Smoking	CVD
1	Yes	Yes
2	No	No
3	Yes	Yes
4	No	No
5	Yes	No
52	No	Yes

t_1

Sample (Person)	Smoking	CVD
1	Yes	No
2	No	Yes
3	Yes	No
4	No	Yes
5	Yes	No
52	No	Yes

t_2

Sample (Person)	Smoking	CVD
1	Yes	No
2	No	No
3	Yes	No
4	No	Yes
5	Yes	Yes
52	No	Yes

t_3

• • •

Sample (Person)	Smoking	CVD
1	Yes	Yes
2	No	No
3	Yes	No
4	No	No
5	Yes	Yes
52	No	Yes

t_{1000}

Randomly permute the samples for one of your variables and calculate t.
Do that 1000 times.
Estimate the pdf*.
You have an estimate of the distribution of t under H_0.