Review: 3-22

- Mediation, path models
- HLM
- Moderation
- Ecological Fallacy

Discrete Variable Comparison Metrics

Examples:

Single class:

- X_1: Smoker or not(0/1)
- X_2: has cancer? (0/1)
- Y: Picture of goat? (0/1)
- $\hat{Y}$: prediction from a logistic model (0/1)
  or any model (e.g. a *gradient boosting deep bayes neural forest)*

Multi-class

- Y: word is subject, direct object, or indirect object (1, 2, or 3 but order means nothing)
  $\hat{Y}$: prediction from a multi-class model
  (a “multinomial” distribution)

Discrete Variable Comparison Metrics

- Chi-Square test for independence
- (true|false) (positive|negative) based metrics:

Discrete Variable Comparison Metrics

Single class:

- X_1: Smoker or not(0/1) X_2: has cancer? (0/1)

N = 100 people sampled from cancer screening center population

	no cancer	cancer	
not smoker	60	10	
smoker	22	8	

Discrete Variable Comparison Metrics

Single class:

- X_1: Smoker or not(0/1) X_2: has cancer? (0/1)

N = 100 people sampled from cancer screening center population

	no cancer	cancer	
not smoker	60	10	
smoker	22	8	

Chi-Squared Test for Independence

H_0: Y and Z are independent H_1: Y and Z are dependent

$$U = \sum_{i=0}^{classes_{x1}} \sum_{j=0}^{classes_{x2}} \frac{(X_{ij} - E_{ij})^2}{E_{ij}}$$

$$\text{where } E_{ij} = \frac{X_{i*} X_{*j}}{n}$$

	no cancer	cancer	
not smoker	60	10	*70*
smoker	22	8	*30*
	82	*18*	*100*

Chi-Squared Test for Independence

H_0: Y and Z are independent H_1: Y and Z are dependent

$$U = \sum_{i=0}^{classes_{x1}} \sum_{j=0}^{classes_{x2}} \frac{(X_{ij} - E_{ij})^2}{E_{ij}}$$

Observed count

Expected count

$$\text{where } E_{ij} = \frac{X_{i*}X_{*j}}{n}$$

	no cancer	cancer	
not smoker	60	10	*70*
smoker	22	8	*30*
	82	*18*	*100*

	Expected distribution	
	70*82/100=**57.4**	12.6
	24.6	5.4

Chi-Squared Test for Independence

k = *df* (degrees of freedom)
= $(classes_{x1} - 1)(classes_{x2} - 1)$

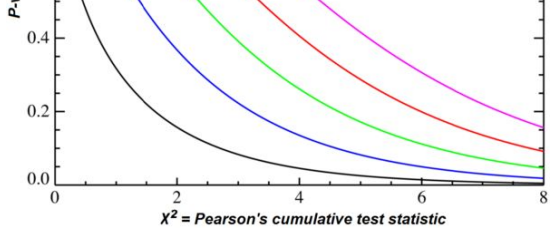
H_0: Y and Z are independent H_1: Y and Z are dependent

$$U = \sum_{i=0}^{classes_{x1}} \sum_{j=0}^{classes_{x2}} \frac{(X_{ij} - E_{ij})^2}{E_{ij}}$$

Observed count

Expected count

$$\text{where } E_{ij} = \frac{X_{i*}X_{*j}}{n}$$



	no cancer	cancer	
not smoker	60	10	*70*
smoker	22	8	*30*
	82	*18*	*100*

	Expected distribution	
	70*82/100=**57.4**	12.6
	24.6	5.4

Discrete Variable Comparison Metrics

- Chi-Square test for independence
- (true|false) (positive|negative) based metrics:

Discrete Variable Comparison Metrics

- Chi-Square test for independence
- **(true|false) (positive|negative) based metrics:**

(Thank you, Wikipedia!)

		True condition			
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive** (Type I error)	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$
	Predicted condition negative	**False negative** (Type II error)	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$
	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	True positive rate (TPR), Sensitivity, Recall $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	

Discrete Variable Comparison Metrics

- Chi-Square test for independence
- **(true|false) (positive|negative) based metrics:**

		True condition		
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**	**False positive** (Type I error)	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$
	Predicted condition negative	**False negative** (Type II error)	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$
	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	True positive rate (TPR), Sensitivity, Recall = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	