Biostat 200 – Lab 2
Lecture Review

September 26, 2013

Logistics

- Assignment 1 due next week via email by 10:30am (to gmail account)
- Datasets from textbook posted on Forum (9/24 lecture)
- Post questions to Forum or send emails directly to us (gmail account just for HW)
- Next labs: review previous lecture, review HW

Data cleaning

- Today's lab
- Always necessary because datasets always have errors
- Document your decisions
- Keep original dataset, and save a copy of the cleaned dataset

Brief Stata review

- Recommend always using .do files
- To set your directory
  - `cd "Documents\Biostat 200\Labs"`
  - \ for PC, / for Mac
- To read in a dataset
  - `use "classsurvey_2013_v1.dta", clear`
- Comments with * or /* */

Looking at variables (last week)

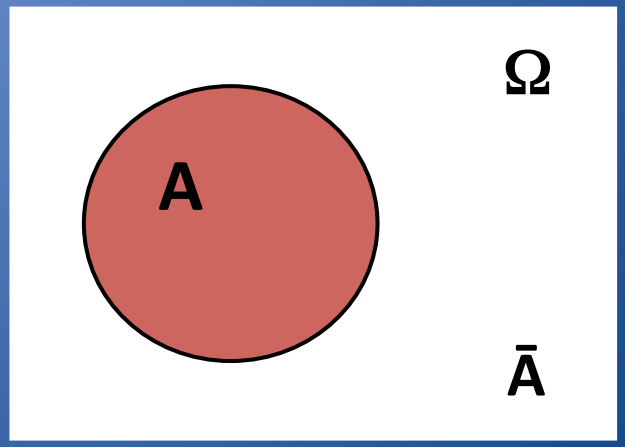
- `summarize [summ] var, detail`
  - percentiles, range, smallest, largest, mean, sd
- `codebook var`
  - same plus type of variable, missing
- `describe var`
  - type of variable, labels
- `tabulate [tab] var`
  - lists all values of the variable

More data cleaning

- To change values (e.g. “92 cents” to “92”)
  - `generate…replace`
- Outliers
  - Consider feasibility
  - Drop values if not possible
  - Document changes in a .do file
  - Keep original copy, save changes to new file

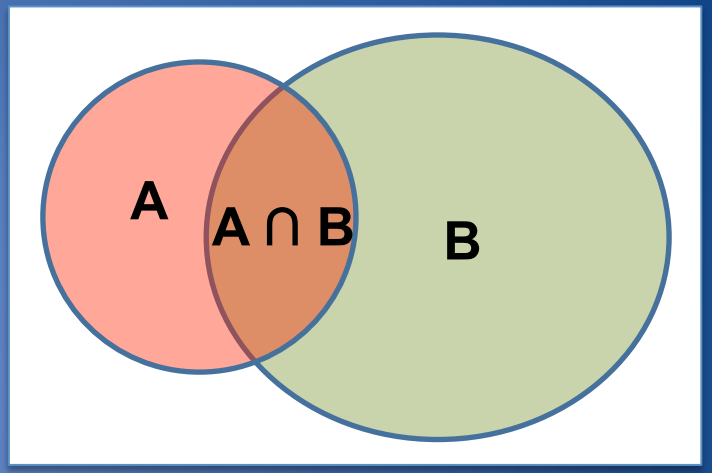
Probability

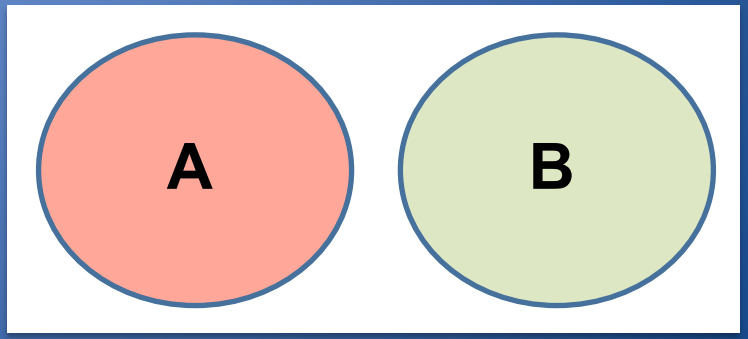
- Event: result of experiment or observation (A)
- Probability of A: P(A)
  - Experiment performed n times, event A occurs m times, then as n increases, m/n approaches P(A)
  - Relative frequency of A in large number of trials
  - Always between 0-1
- Complement: $\bar{A}$ or A^C
  - Probability of not A
  - $P(A) = 1 - P(\bar{A})$
  - $P(\Omega) = P(A) + P(\bar{A}) = 1$



Probability

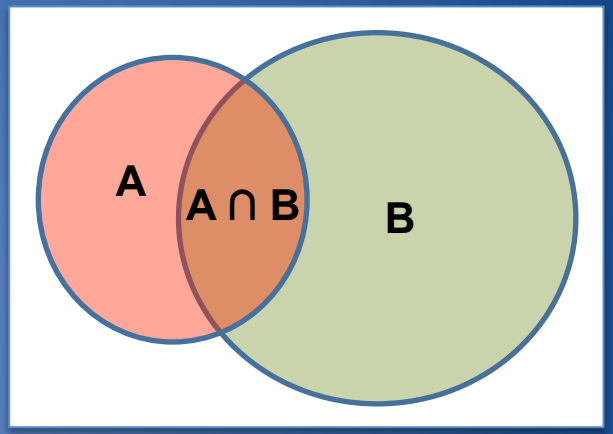
- Intersection: P(A ∩ B)
  - Probability of both A and B
- Union: P(A U B)
  - Probability of A or B or both
  - P(A U B) = P(A) + P(B) – P(A ∩ B)
- Mutual exclusivity
  - Events don't overlap, intersection P(A ∩ B) = 0
  - Additive rule: P(A U B) = P(A) + P(B)





Probability

- Conditional probability: P(B|A)
  - Probability that event B will occur given that event A has occurred
  - P(B|A) = P(A ∩ B) / P(A)
  - Relative size of probability of the intersection (both A and B) compared to probability of A
  - Multiplicative rule: P(A ∩ B) = P(A) P(B|A)
  - Applies to relative risk, odds ratios



Probability

- Independence: $P(B|A) = P(B)$
- If the occurrence of B does not depend on A, then the probabilities of the events are independent
- Independence ≠ mutual exclusivity
- By the multiplicative rule, for independent events, $P(A \cap B) = P(A)\,P(B|A) = P(A)\,P(B)$

Brief word on diagnostic tests

- Sensitivity = $P(T^+ | D^+) = P(T^+ \cap D^+)/P(D^+)$
  - Probability of true positives = TP/(TP+FN)
- Specificity = $P(T^- | D^-) = P(T^- \cap D^-)/P(D^-)$
  - Probability of true negatives = TN/(FP+TN)

		TRUTH	
		Disease D^+	No Disease D^-
Test	T^+	TP	FP
	T^-	FN	TN

Brief word on diagnostic tests

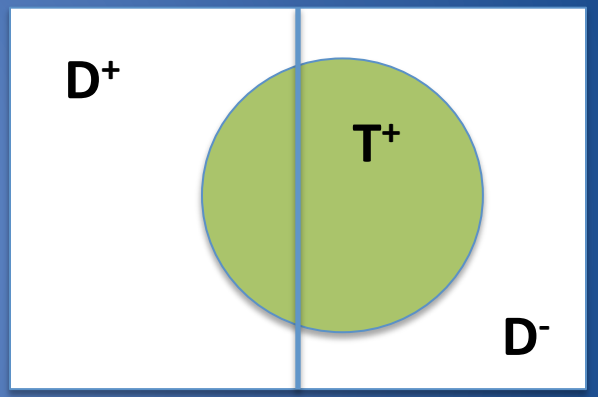
- The probability of results of diagnostic tests (in general) are independent from each other
- If you perform a test twice, the probability of the results is the product of the two independent probabilities (multiplicative rule)
  - $P(A \cap B) = P(A)\, P(B|A) = P(A)\, P(B)$
- All possibilities added together should = 1

Bayes' theorem and diagnostic tests

- You know
  - sensitivity = $P(T^+|D^+)$
  - specificity = $P(T^-|D^-)$
- You want
  - probability that T^+ is truly D^+ = $P(D+|T+)$ = PPV
- Bayes' theorem: $P(A|B) = P(B|A)*P(A) / P(B)$
- Or PPV = $P(D^+|T^+) = P(T^+|D^+)*P(D^+) / P(T^+)$

Bayes' theorem applied

- PPV = $P(D^+|T^+) = P(T^+|D^+)*P(D^+) / P(T^+)$
- $P(T^+|D^+)$ = sensitivity
- $P(D^+)$ = disease prevalence
- $P(T^+) = P(T^+ \cap D^+) + P(T^+ \cap D^-)$ by law of total probability
- Based on the multiplicative rule:
- $P(T^+) = P(T^+|D^+) * P(D^+) + P(T^+|D^-) * P(D^-)$



sensitivity

disease prevalence

1 - specificity

1 - disease prevalence