Chap.1 Introduction

Types of Variables

1. Response vs. Explanatory Variables

   Response (Dependent) variable – the variables we are attempting to predict or explain.

   Explanatory (Independent) variable – the variables which may be used to help predict or explain the response.

2. Continuous vs. Discrete variables

   Continuous variables – a numeric variable which, in principle, may assume any value over same interval collection of intervals

   Discrete variables – a nuueric or non-numeric variable for which the set of possible values is either finite or countably infinite.

3. Measurement scales for variables

   Nominal scale – the levels of the scale have no natural ordering.
   Ex) choice of transport (walk, car, bike, bus), Religion (Christian, jew, muslim, other)

   Ordinal scale – the level of the scale have a natural ordering but there are no numeric distances between the various levels of the scale
   Ex) disease severity (mild, moderate, severe),
   political beliefs (liberal, moderate, conservative)

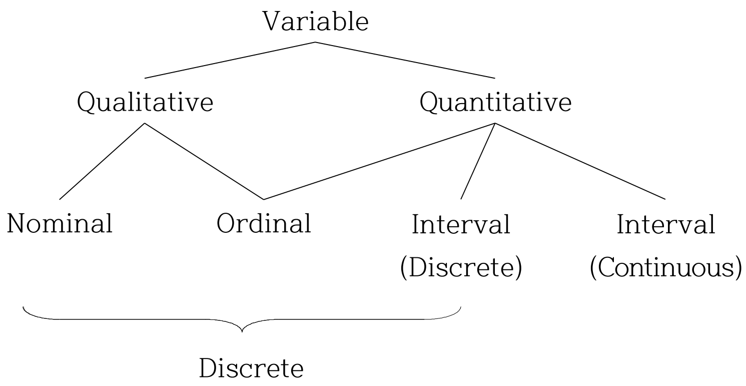
   Interval scale – there are numeric distances between the various levels of the scale; Any measurement of interval scale can be ranked, counted, subtracted, or added, and equal intervals separate each number on the scale.
   Ex) Interval Discrete (count) - Number of cigarettes, teeth, visits etc.
   Interval Continuous - temperature (^oC, F)

   Ratio scale – never fall below zero.
   Ex) Blood pressure, weight, distance

4. Quantitative vs Qualitative variables

Quantitative variable – an interval variable or an ordinal variable where the levels of the scale can be assigned meaningful numeric orders.

Qualitative variable – a nominal variable or an ordinal variable where the levels of the scale cannot be assigned meaningful numeric order.



Types of studies

1. Experimental vs. Observational Study

Experimental Study – the investigator has control over which subjects receive the treatments.

Observational Study – the investigator has no control over the treatment or the control group

2. General Study Designs

Retrospective Designs – choose subjects and look into their past and collect data

Cross-Sectional Designs – choose subjects and observe their present status and collect data

Prospective Designs – choose subjects, and monitor their status into the future and collect data

3. Study Designs in Epidemiology(유행병학)

Notation:

D: has disease (has condition)
$\overline{D}$: does not have disease (Does not have condition) — Response Variable

E: has been exposed (has received treatment)
$\overline{E}$: has not been exposed (has not received treatment) — Explanatory Variable

P: Population
RS: Random Sample
RA : Random Assignment

1) Cross-Sectional Study (Survey) - Classify according to present status

$$P \rightarrow RS \begin{cases} D, \ E \\ D, \ \overline{E} \\ \overline{D}, \ E \\ \overline{D}, \ \overline{E} \end{cases}$$

Results:

	D	$\overline{D}$
E	a	b
$\overline{E}$	c	d

$$P(D|E) \approx \frac{a}{a+b}$$

2) Case-Control Study (Retrospective) - Classify into exposure groups

		classify into exposure groups
Cases:	D	E $\overline{E}$
Controls:	$\overline{D}$	E $\overline{E}$
		look into past

Subjects are matched so that each case has a counterpart control(or controls)

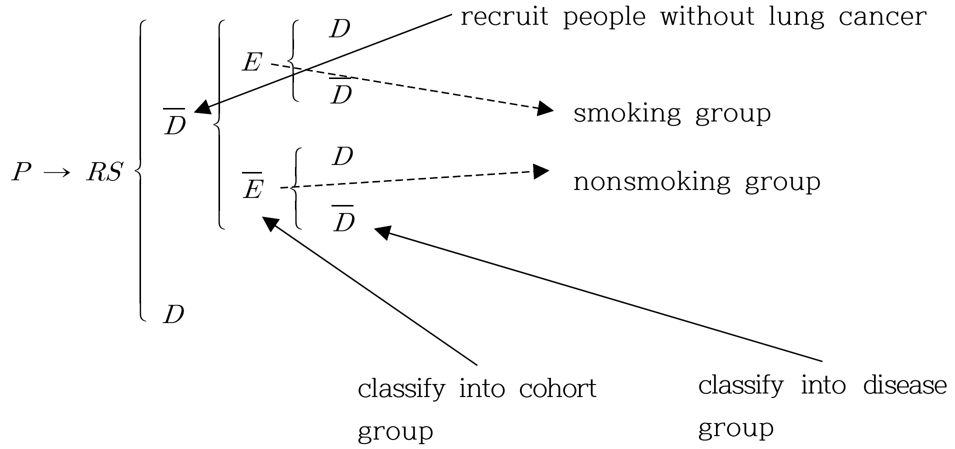
Results:

	D	$\overline{D}$
E	a	b
$\overline{E}$	c	d
	$a+c$	$b+d$
	fixed	fixed

Estimation of $P(D|E)$ requires $P(D)$ and Bayes Rule

$$P(D|E) = \frac{P(D \cap E)}{P(E)} = \frac{P(E|D)P(D)}{P(E \cap D) + P(E \cap \overline{D})}$$
$$= \frac{P(E|D)P(D)}{P(E|D)P(D) + P(E|\overline{D})P(\overline{D})}$$

3) Cohort Study (Prospective)



Results:

	D	$\overline{D}$	
E	a	b	$a+b$ fixed
$\overline{E}$	c	d	$c+d$ fixed

$$P(D|E) \approx \frac{a}{a+b}$$

4) Clinical Trial(Prospective)

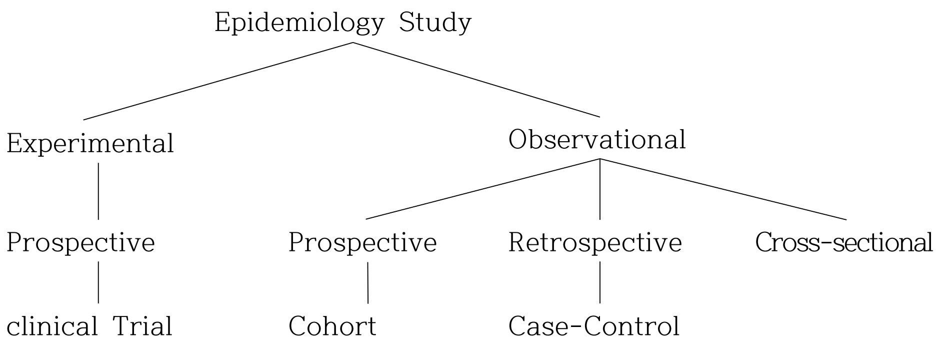
$$P \rightarrow RS \rightarrow RA \begin{cases} E \begin{cases} D \\ \overline{D} \end{cases} \\ \overline{E} \begin{cases} D \\ \overline{D} \end{cases} \end{cases}$$

$$\dashrightarrow$$
$$time$$

Results:

	D	$\overline{D}$	
E	a	b	$a+b$ fixed
$\overline{E}$	c	d	$c+d$ fixed

$$P(D|E) \approx \frac{a}{a+b}$$



Probability Distributions for Categorical Data

The binomial distribution (and its multinomial distribution generalization) plays the role that the normal distribution does for continuous response.

Binomial Distribution

- n Bernoulli trials - two possible outcomes for each(success, failure)
  $\pi = P(sucess)$, $1-\pi = P(failure)$ for each time
  $Y=$ number of successes out of n trials
  Trials are independent
  Y has binomial distribution

$$P(y) = \frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y},\ y=0,1,\cdots,n$$

  where $y \neq y(y-1)(y-2)\cdots(1)$ with $0!=1$ (factorial)

- Example Vote (Democrat, Republican)
  Suppose $\pi = P(Democrat) = 0.50$
  For random sample size $n=3$, let $y=$ number of Democratic votes

$$p(y) = \binom{3}{y}(0.5)^y(0.5)^{3-y}$$

$$\Rightarrow p(0) = \frac{3!}{0!3!}(0.5)^0(0.5)^3 = 0.125$$

$$p(1) = \frac{3!}{1!2!}(0.5)^1(0.5)^2 = 0.375$$

y	$p(y)$
0	0.125
1	0.375
2	0.375
3	0.125
sum	1.0

Note

$$E(Y) = n\pi$$

$$Var(Y) = n\pi(1-\pi),\ \sigma = \sqrt{n\pi(1-\pi)}$$

Let $p = \frac{Y}{n} =$ proportion of success (also denoted $\hat{\pi}$)

$$E(p) = E\left(\frac{Y}{n}\right) = \pi$$

$$\sigma\left(\frac{Y}{n}\right) = \sqrt{\frac{\pi(1-\pi)}{n}}$$

When each trial has >2 possible outcomes, numbers of outcomes in various categories have multinomial distribution.

Inference for a proportion

We conduct inferences about parameters using maximum likelihood

Def. The likelihood function is the probability of the observed data, expressed as a function of the parameter value.

Example : Binomial, $n = 2$, observe $y = 1$

$$p(1) = \frac{2!}{1!1!}\pi^1(1-\pi)^1 = 2\pi(1-\pi) \overset{let}{=} l(\pi)$$

the likelihood function defined for π between 0 and 1.
If $\pi = 0$, probability is $l(0) = 0$ of getting $y = 1$
If $\pi = 0.5$, probability of $l(0.5) = 0.5$ of getting $y = 1$

Def. The Maximum Likelihood Estimate(MLE) is the parameter value at which the likelihood function takes its maximum.

Example : $l(\pi) = 2\pi(1-\pi)$ maximized at $\hat{\pi} = 0.5$
i.e. $y = 1$ in $n = 2$ trials is most likely if $\pi = 0.5$ ML estimate of π is $\hat{\pi} = 0.5$

Note.

- For binomial, $\hat{\pi} = \frac{y}{n}$ = proportion of successes
- *If* $y_1, y_2, \cdots, y_n$ are independent from normal(or many other distribution, such as Poisson), ML estimate $\hat{\mu} = \overline{y}$ (sample mean)
- In ordinary regression ($Y \sim normal$) "least squares" estimates are ML.
- For large n for any distribution, ML estimates are optimal(no other estimator has smaller standard error).
- For large n, ML estimators have approximate normal sampling distribution (under weak conditions)

ML Inference about Binomial Parameter

$$\hat{\pi} = p = \frac{y}{n}$$

Recall $E(p) = \pi$, $\sigma(p) = \sqrt{\frac{\pi(1-\pi)}{n}}$

- Note $\sigma(p) \downarrow$ as $n \uparrow$
  $p \rightarrow \pi$ (law of large numbers, true in general for ML)
- p is a sample mean for (0,1) data, so by central limit Theorem, sampling distribution of p is approximately normal for large n (True in general for ML)

Significance Test for binomial parameter

$$H_0:\ \pi=\pi_0 \qquad H_a:\ \pi\neq\pi_0\ \ (\mathrm{or}\ 1-sided)$$

Test statistic

$$Z=\frac{p-\pi_0}{\sigma(p)}=\frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

has large-sample standard normal (denoted by $N(0,1)$) null distribution. (Note use null SE for test).

$p-$ value= two-tail probability of results at least as extrem as observed (if null were true)

Confidence interval(C.I.) for binomial parameter

Def. Wald C.I for a parameter θ is

$$\hat{\theta}\pm Z_{\frac{\alpha}{2}}(SE).$$

(eg, for 95% confidence level, estimate plus and minus 1.96×estmiated standard errors, where $Z_{0.025}=1.96$)

Example $\theta=\pi,\ \hat{\theta}=\hat{\pi}=p$

$$\sigma(p)=\sqrt{\frac{\pi(1-\pi)}{n}}\ \text{estimated by}\ SE=\sqrt{\frac{p(1-p)}{n}}$$

95% C.I. for π is

$$p\pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

Note. Wald C.I. often has poor performance in categorical data analysis unless n quite large.

Example Estimate π =population proportion of vegetarians

For $n=20$, whe get $y=0$

$$p=\frac{0}{20}=0.0$$

95% C.I. : $0 \pm 1.96\sqrt{\frac{0 \times 1}{20}} = 0 \pm 0 = (0.0)$

- Note what happens with Wald C.I. for π if $p=0$ or 1
- Actual coverage probability is much less than 0.95 if π is near 0 or 1
- Wald 95% C.I. = set of π_0 values for which $p-$ value >0.05 in testing

$$H_0 : \pi = \pi_0 \quad vs. \quad H_a : \pi \neq \pi_0$$

using

$$Z = \frac{p - \pi_0}{\sqrt{\frac{p(1-p)}{n}}}$$ (denominator uses estimated SE)

<u>Def</u> Score test, Score C.I. - use null SE

eg) Score 95% C.I. = set of π_0 values for which $p-$ value >0.05 in testing

$$H_0 : \pi = \pi_0 \quad vs. \quad H_a : \pi \neq \pi_0$$

using

$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

[note null SE in denominator, (known, not estimated)]

Example π = probability of being vegetarian, $y=0$, $n=20$, $p=0$

what π_0 satisfies

$$\pm 1.96 = \frac{0 - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{20}}} \quad ?$$

$$\Leftrightarrow \ 1.96\sqrt{\frac{\pi_0(1-\pi_0)}{20}} = |0 - \pi_0|$$

1. $\pi_0 = 0$ is one solution
2. $\pi_0 = 0.16$ is other solution(solving quadratic equation)

95% score C.I. is (0, 0.16), more sensible than Wald C.I. of (0,0)

- Wald C.I.

$$p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

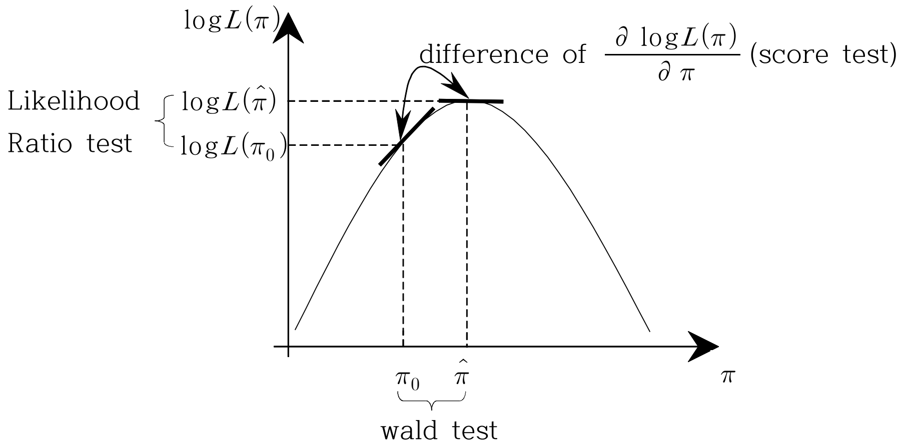
also works well even for small samples if add 2 successes, add 2 failures before appling(this is the “Agresti-coull method”)

- For inference about proportions, score method tends to perform better than Wald method, in terms of having actual error rates closer to the advertised levels
- Another good test, C.I. uses the <u>likelihood function.</u>

(eg. C.I.= values of π for which $l(\pi)$ close to $l(\hat{\pi})$

= values of π_0 not rejected in "likelihood-ratio test")

● For small n, inference uses actual binomial sampling dist. of data instead of nomal approximation for that dist.



2. Multinomial

We have c categories

$$\underline{y_i} = (y_{i1},\ y_{i2},\ \cdots,\ y_{ic})\ \ with\ \ p(y_{ij}=1)=\pi_j,\ \ \sum_{j=1}^{c} y_{ij} = 1$$

Let $n_j = \sum_{i=1}^{n} y_{ij}$ and $n = \sum_j n_j$

$$p(n_1, n_2, \cdots, n_{c-1}) = \left(\frac{n!}{n_1! \cdots n_c!}\right) \pi_1^{n_1} \cdots \pi_c^{n_c}$$

$$E(n_j) = n\pi_j,\quad Var(n_j) = n\pi_j(1-\pi_j)$$

3. Poisson

$$p(y) = \frac{e^{-\mu}\mu^y}{y!},\ \ y = 0, 1, \cdots$$

$$E(Y) = \mu = Var(Y)$$

Note

- Poisson derived from Binomial as $n \to \infty$, $\pi \to 0$, $\mu = n\pi$
- When variation exceeds that predicted by standard dist. there is <u>overdispersion</u>
- For c categories, it assumes counts $(Y_1, Y_2, \cdots, Y_c)$ are indep. Poisson(μ_i), then given $\sum_{j=1}^{c} Y_j = n$, conditional dist. is multinomial with $\pi_j = \dfrac{\mu_j}{\sum_k \mu_k}$

Statistical Inference for Categorical Data

We will use Maximum Likelihood (ML) to illustrate for multinomial. Multinomial log-likelihood is

$$L(\underline{\pi}) = \sum_{j=1}^{c} n_j \log \pi_j$$

MLE of π_j : $\hat{\pi}_j = \frac{n_j}{n}$, $j = 1, 2, \cdots, c$.

		Categories				
		1	2	$\cdots$	c	
Subject	1	y_{11}	y_{12}	$\cdots$	y_{1c}	$\sum_{j=1}^{c} y_{ij} = 1$
	2	y_{21}	y_{22}	$\cdots$	y_{2c}	
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	
	n	y_{n1}	y_{n2}	$\cdots$	y_{nc}	
		n_1	n_2	$\cdots$	n_c	$\sum_{j=1}^{c} n_j = n$
		π_1	π_2	$\cdots$	π_c	

Ex) How can you test $H_0 : \pi_j = \pi_{j0}$, $j = 1, \cdots, c$? (Karl Pearson, 1900)

$$X^2 = \sum_{j=1}^{c} \frac{(n_j - \mu_j)^2}{\mu_j} \xrightarrow[H_0]{d} \chi^2_{c-1}$$

(Pearson Chi-squared statistic)

Where $\mu_j = n\pi_{j0}$ = expected frequency

- For $c = 2$ categories X^2 has $df = 1$, then $X^2 = Z^2$ where

$$Z = \frac{\hat{\pi}_1 - \pi_{10}}{\sqrt{\frac{\pi_{10}(1 - \pi_{10})}{n}}} \xrightarrow[H_0]{d} N(0,1)$$

for testing a binomial proportion