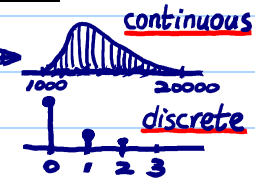
Types of variables

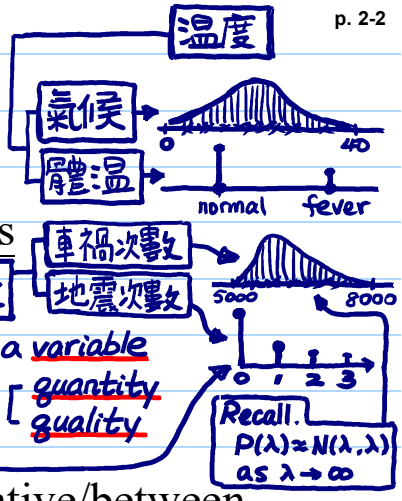
- response-explanatory distinction (dependent-independent, response-predictor; covariate)
  - response variables Y: regarded as *random*.
  - explanatory variables $X_1, \ldots, X_m$: regarded as *deterministic*.
  - causal relationship? not necessary.

① values determined before collecting Y data, e.g., DOE
② if random, analysis conditional on the observed values of $X_1, \ldots, X_m$.

- continuous-discrete distinction
  - whether a variable can take any values within an interval (if yes ⇒ continuous; if no, only countable values ⇒ discrete)
  - Q: does continuous *data* really exist in the real world?
  - No, because of the precision limitation of measuring instrument.
  - a better distinction approach from the viewpoint of data analysis: according the *number* of values a variable can take within the range in which most data fall



- variable taking lots of values (dense) ⇒ continuous; few values ⇒ discrete
  - Q: Should Poisson data (infinite possible values) be treated as discrete or continuous from this data analysis viewpoint?



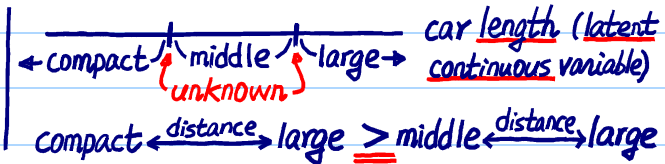
- quantitative-qualitative distinction
  - continuous variable must be quantitative
  - discrete variable could be quantitative/qualitative/between

- categorical (qualitative or between) variables can be further classified into:
  - nominal variable: no natural ordering between categories (e.g., religious affiliation, mode of transportation, favorite type of music, …)
    - values that represent categories have no numeric meaning
    - no value exist between categories
    - analysis irrelevant to the order of listing the categories

➢ ordinal variable: there exist some ordering between categories (e.g., size of automobile, social class, political philosophy, patient condition, …)

In analysis, cannot arbitrarily change the order of categories as in nomial case

e.g., size of car – a (latent) continuous variable hidden behind



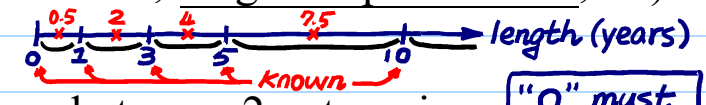
discrete (relatively few intervals)

▪ exact distance between ordered categories are unknown

→ discrete interval or continuous interval

➢ interval variable: categories are non-overlapping intervals (e.g., functional life length of television set, length of prison term, …)

cf. continuous variable (an interval variable with many and very dense intervals)



For variable with 2 categories, does not matter they have order or not, same treatment.

▪ can have numerical distances between 2 categories might be inaccurate

"0" must be well-defined

▪ sometimes, possible to compare the ratio of 2 categories

➢ Sometimes, it is the way that a variable is measured determined its classification, e.g., education:

X_2 has higher "resolution" than X_1. Q: Which is better in analysis? It depends.

▪ nominal when measured as public/private school

$X_1 \equiv$ ▪ ordinal when measured as none/high school/bachelor/…

$X_2 \equiv$ ▪ interval variable when measured by # of year intervals

$X_1 = f(X_2)$, f: non-invertable ⇒ information lost in X_1

more information

➢ hierarchy of measurement scale: interval variable (highest) > ordinal > nominal (lowest)

∵ ignore or disregard some information in the higher level

$X_{ord} = f(X_{int})$, $X_{nom} = g(X_{ord})$

★ ▪ statistical methods for variables of one type can be used for variables at higher level, but not at lower levels

➢ nominal variable ⇒ qualitative; interval variable ⇒ (close to) quantitative; ordinal variable ⇒ between (fuzzy)

- choice of statistical method/model for different types of variables – a rough classification:

	Resp.	Expl.
Cont.		
Disc. Quan.	counts (LNp 2-8)	
Disc. Inte.		
Ordi.		
Nomi.		

digitization

check ★ in LNp 2-6

LNp 2-9

➢ only response variables, no explanatory variable

▪ 1 response variable ⇒ uni-variate analysis

e.g., $Y_1, \cdots, Y_k \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ or $B(n, p)$ or $Exp(\lambda)$…

estimate or test the parameters.

categorical data

assign distribution

use base functions

- principal component
- factor analysis
- canonical correlation

▪ more than 1 response ⇒ multi-variate analysis

e.g., $\underline{Y}_1, \cdots, \underline{Y}_k \overset{i.i.d.}{\sim}$ multivariate normal $(\underline{\mu}, \Sigma)$

information contained in the var-cov matrix

➢ both response and explanatory variables → regression-type

- response: (1) regarded as random variable; (2) modeling depends on the types of variable ← assign distribution
  - continuous & normal ⇒ linear model (LM)
  - continuous but not normal (including exponential family, such as Weibull, gamma, …) ⇒ generalized linear model
  - discrete ⇒ generalized linear model (GLM) ← main focus of this course

assigning probabilities on counts:
- binomial
- multinomial
- Poisson
- negative binomial
- hypergeometric

- explanatory: (1) regarded as deterministic; (2) same treatment (sum of base functions multiplied by their coefficients) for any types of explanatory variables in modeling

Recall. $X\beta$ in LM

  - quantitative: base functions like polynomial or other continuous transformations (e.g., log, exp, sin, cos, …) ← LNp.1-23,24
  - qualitative: base function like dummy variables ← LNp.1-25~32; e.g.,
    - treatment or sum coding for nominal
    - Helmert coding for ordinal

Sufficient statistics of categorical responses

- Q: how to convert categorical observations (symbols or notations) into numerical and computable data without losing any important information? Q: what important information? meaningless if taking difference or ratio or numeric calculation on

- Example 1.
  - scenario
    - observe $Y_{\mathbf{x},i} \in r$ categories $\{1, 2, \ldots, r\}$, $i=1, \ldots, k_{\mathbf{x}}$. ($Y$: response; $\mathbf{x}$: predictors; r: could be nominal, ordinal, or discrete interval; $k_{\mathbf{x}}$: a fixed known value)
    - category j observed with probability $p_{\mathbf{x},j}$, $j=1, \ldots, r$, and $p_{\mathbf{x},1} + p_{\mathbf{x},2} + \cdots + p_{\mathbf{x},r} = 1$. ← Note. $Y_{\mathbf{x}}$ is random
    - statistical modeling: $Y_{\mathbf{x},1}, Y_{\mathbf{x},2}, \ldots, Y_{\mathbf{x},k_{\mathbf{x}}}$ are independent and identically distributed from $multinomial(1; p_{\mathbf{x},1}, \ldots, p_{\mathbf{x},r})$.
  - sufficient statistics
    - the joint pmf of $Y_{\mathbf{x},1}, Y_{\mathbf{x},2}, \ldots, Y_{\mathbf{x},k_{\mathbf{x}}}$ (categorical responses, might not be meaningful if do calculation like $\sum_i Y_{\mathbf{x},i}/k_{\mathbf{x}}$ cf. Linear model) is
$$p_{\boldsymbol{x},1}^{N_{\boldsymbol{x},1}} \cdot p_{\boldsymbol{x},2}^{N_{\boldsymbol{x},2}} \cdots p_{\boldsymbol{x},r}^{N_{\boldsymbol{x},r}}$$
where $N_{\mathbf{x},j}$ = number of category j in the $k_{\mathbf{x}}$ trials, $j=1, \ldots, r$, and $k_{\mathbf{x}}=N_{\mathbf{x},1}+\cdots+N_{\mathbf{x},r}$. ← count
    - by factorization thm, $(N_{\mathbf{x},1}, \ldots, N_{\mathbf{x},r})$ are the sufficient statistics for the parameters $(p_{\mathbf{x},1}, \ldots, p_{\mathbf{x},r})$ and $(N_{\mathbf{x},1}, \ldots, N_{\mathbf{x},r}) \sim multinomial(k_{\mathbf{x}}; p_{\mathbf{x},1}, \ldots, p_{\mathbf{x},r})$. (important information (dim=r-1); numerical data, difference, sum, ratio, … are meaningful)

* can be treated as a predictor with r levels
* can understand how $\mathbf{x}$ influence $p_{\mathbf{x},i}$'s in analysis

  - note: two categories and binomial is a special case of $r=2$.