Chapter 1
Basic Biostatistics

Jamalludin Ab Rahman MD MPH
Department of Community Medicine
Kulliyyah of Medicine

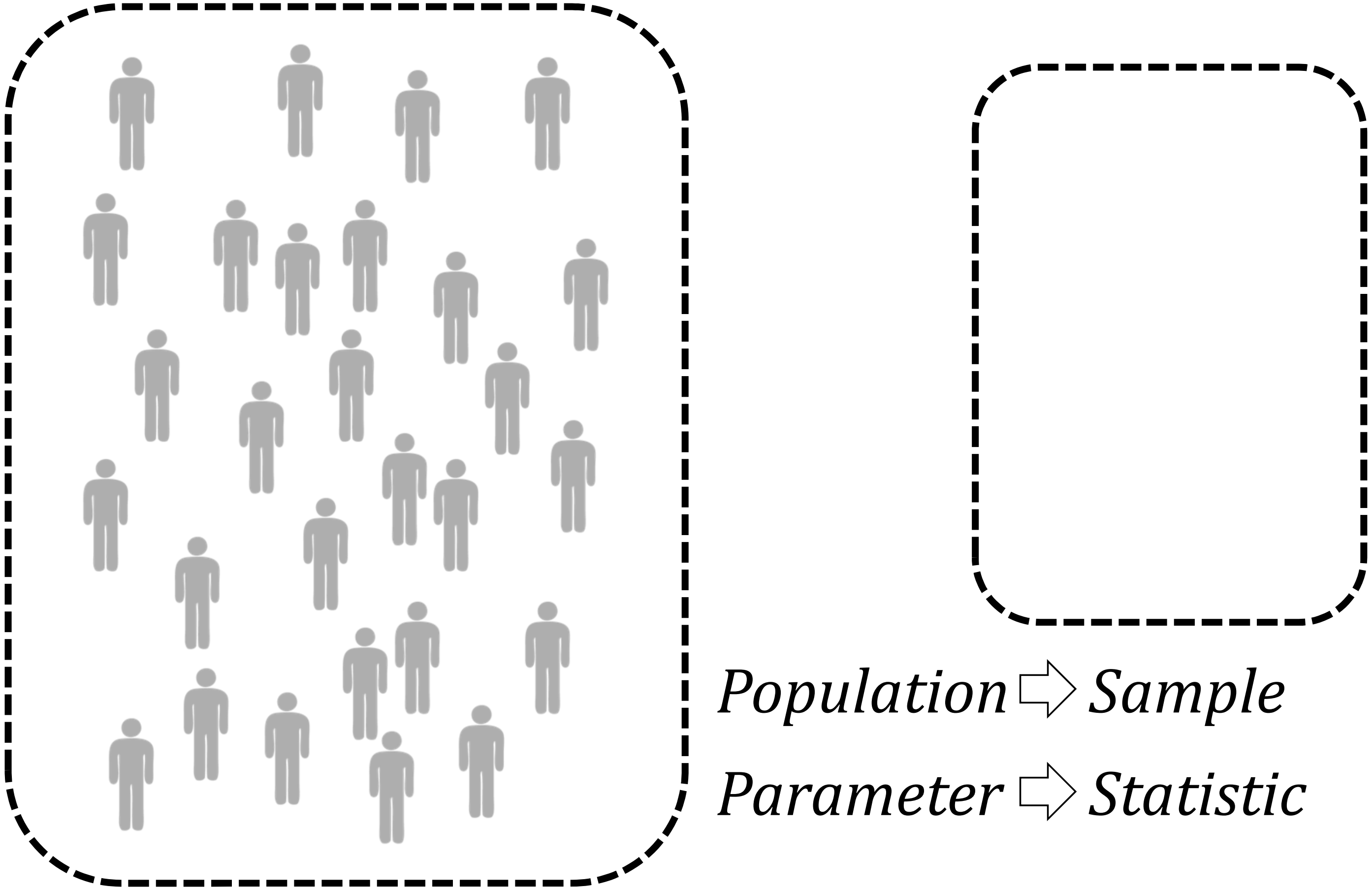
Content

- Basic premises – variables, level of measurements, probability distribution
- Descriptive statistics
- Inferential statistics, hypothesis testing

We observe, we believe.

What we observe might not be the truth

...and we can’t observe all. We sample.

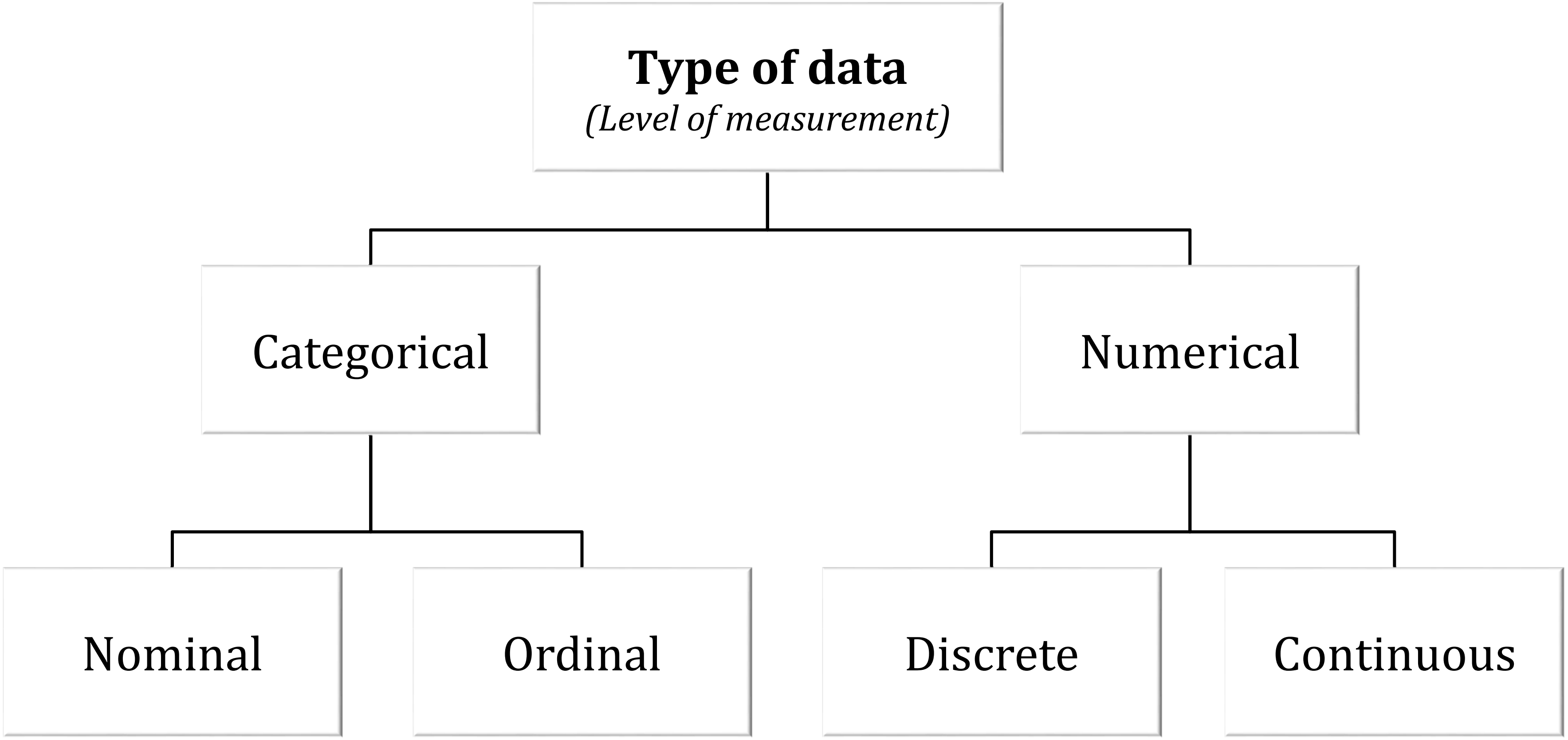


Population ⇨ *Sample*

Parameter ⇨ *Statistic*

Variable

- Characteristic of a population
- Can take different values
- Data = measurements collected/observed



Type of data
(Level of measurement)

Categorical

Numerical

Nominal

Ordinal

Discrete

Continuous

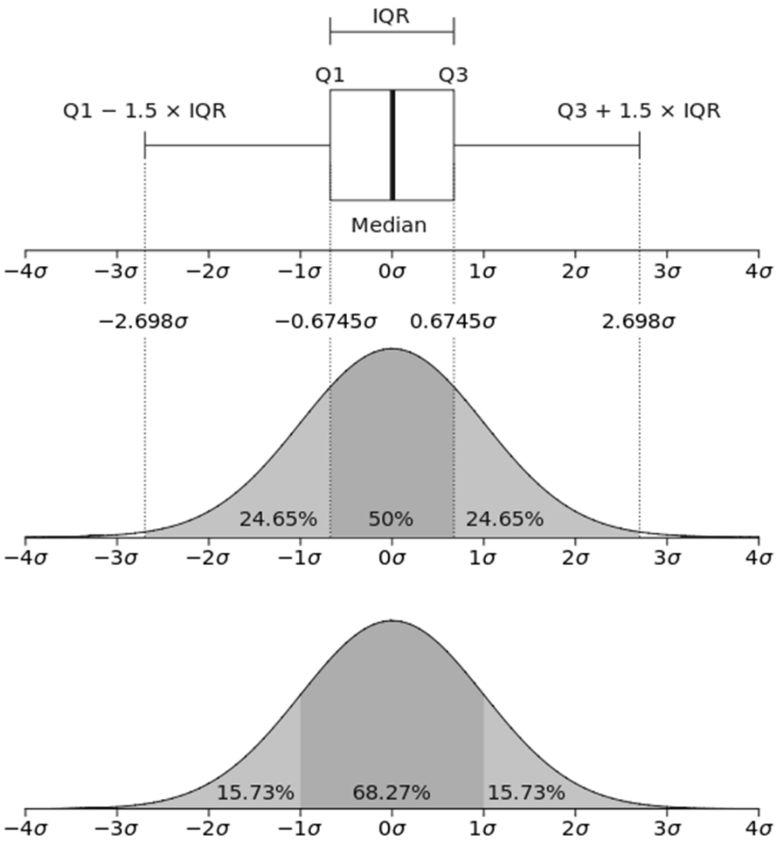
e.g. Gender, Race

e.g. Cancer staging, Severity of CXR for PTB

e.g. Parity, Gravida

e.g. Hb, RBS, cholesterol.

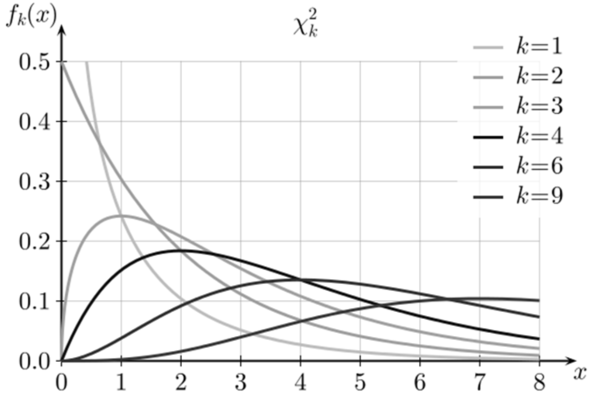
Normal distribution

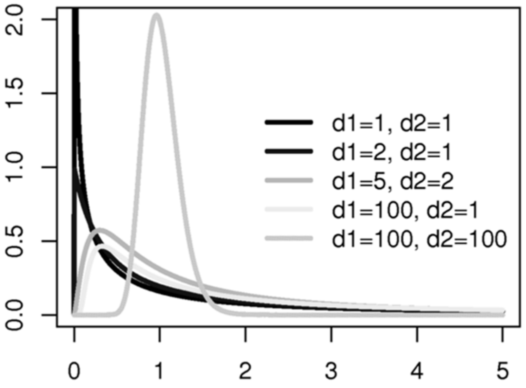


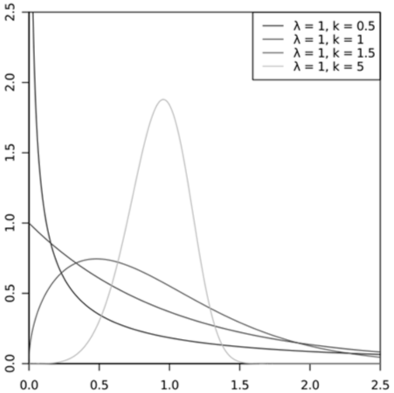
$$f(x;\mu,\sigma^2)=\frac{1}{\sqrt[\sigma]{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2})$$

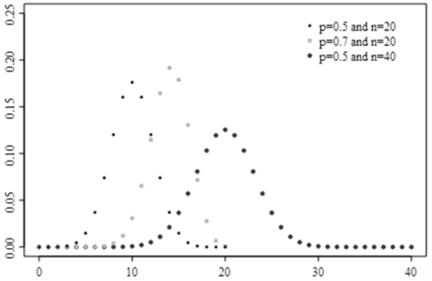
Other distributions

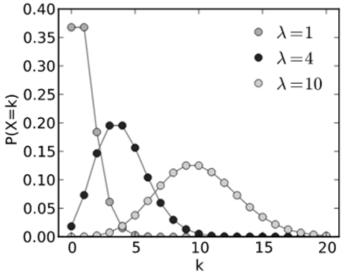
- Discrete vs. continuous probability distribution











χ^2, F, Weibull,
Binomial & Poisson

Characteristics of Normal distribution

- Smooth, symmetrical (around the mean), uni-modal, bell shaped curve
- Mean = Median = Mode
- Skewness = 0
- Kurtosis = 0
- The total area under the curve (AUC) = 1
- Asymptotic to the x-axis – never touch x-axis

Test of Normality

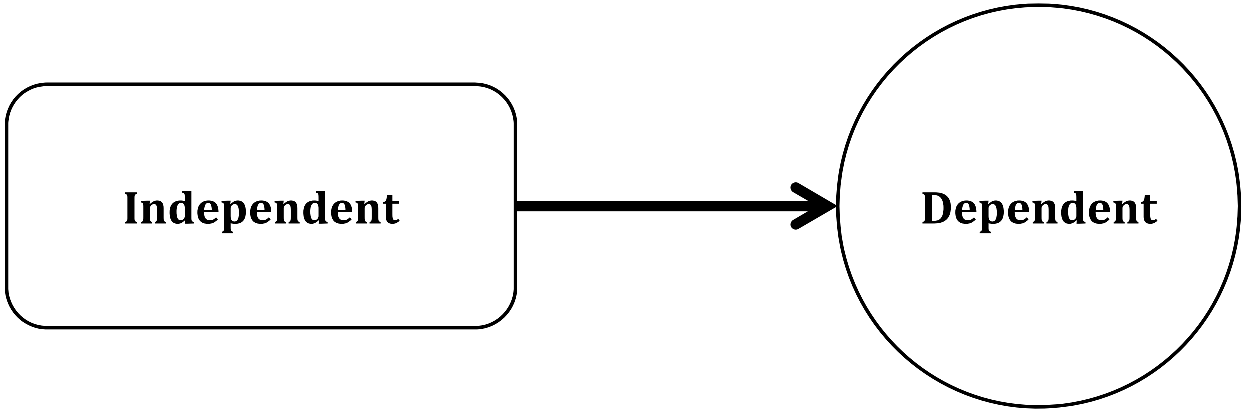
- Anderson–Darling Test
- Corrected Kolmogorov–Smirnov Test (Lilliefors Test)
- Cramér–von-mises Criterion
- D'agostino's K-squared Test
- Jarque–Bera Test
- Pearson's Chi-square Test
- Shapiro–Francia
- Shapiro–Wilk Test

Use Normality test with caution

- ***Small samples almost always pass a normality test***. Normality tests have little power to tell whether or not a small sample of data comes from a Gaussian distribution.
- With ***large samples, minor deviations from normality may be flagged as statistically significant***, even though small deviations from a normal distribution won't affect the results of a t test or ANOVA.

Association

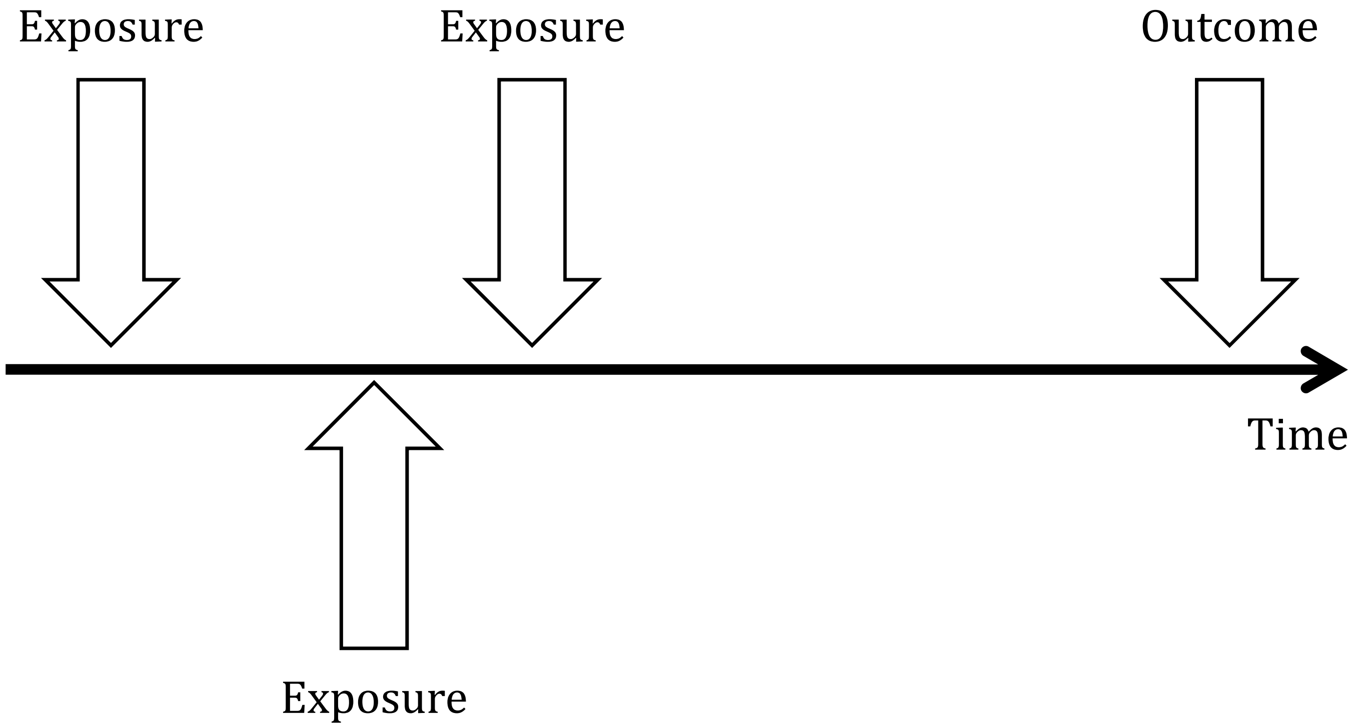
- a value and whose associated value may be changed



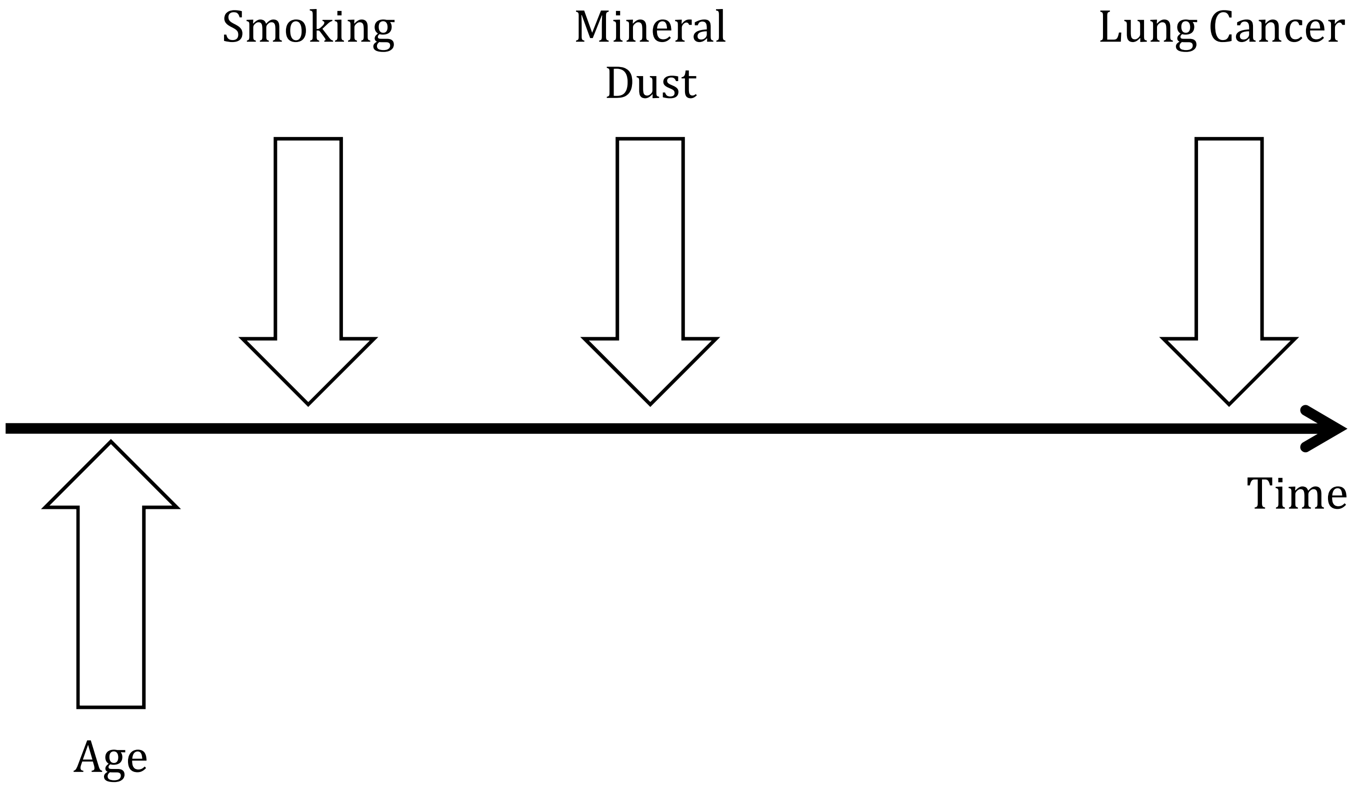
e.g. Smoking

e.g. Lung Cancer

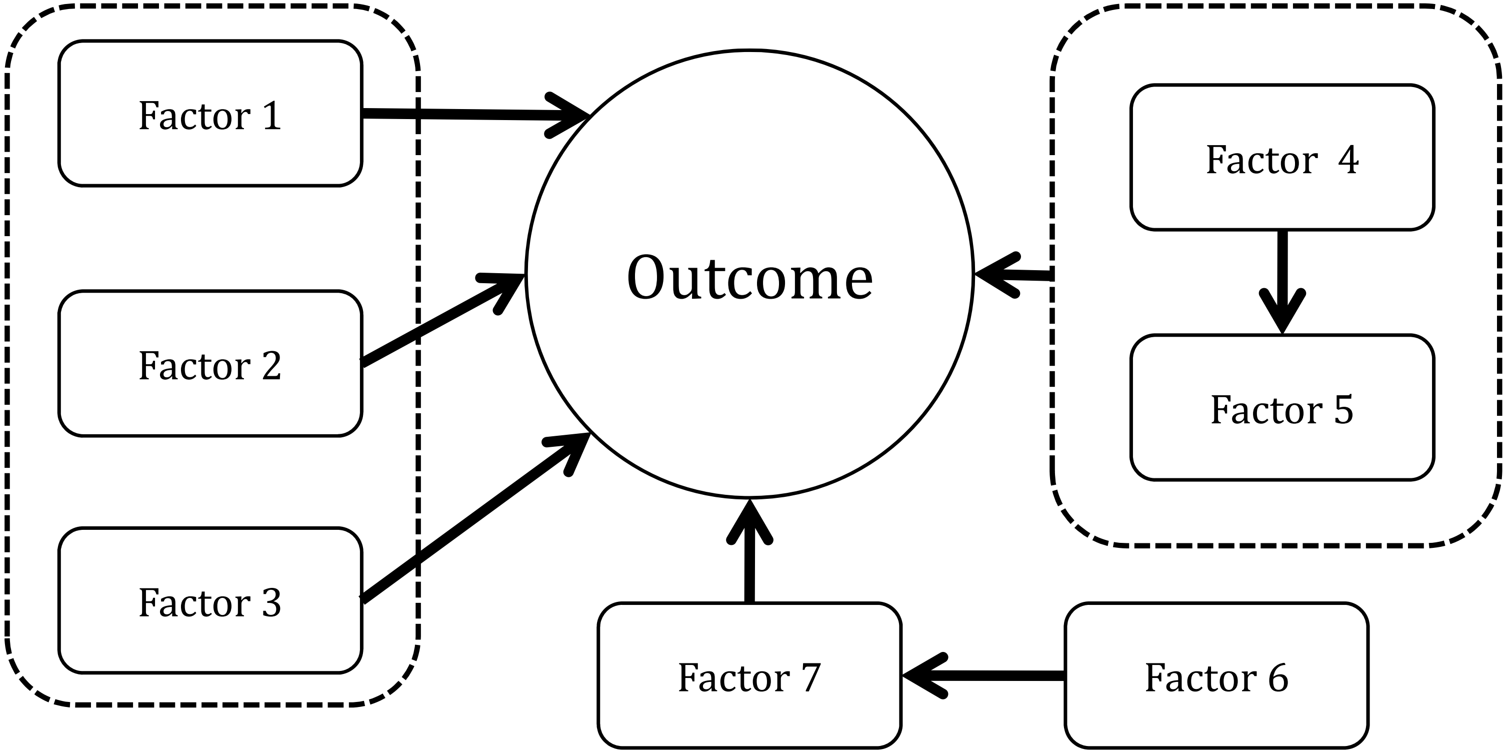
Disease model



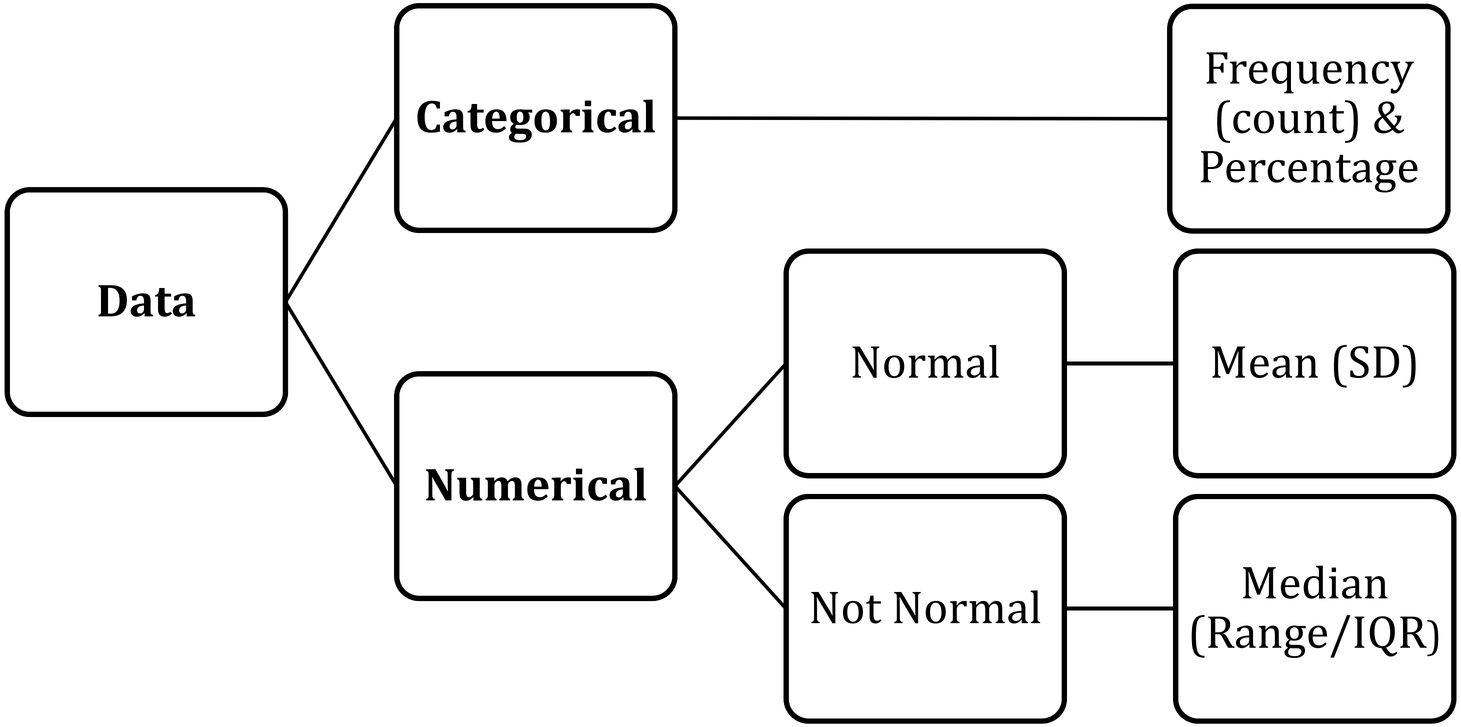
Disease model (example)



Causal relationship



Descriptive statistics


Data

Categorical

Numerical

Normal

Not Normal

Frequency (count) & Percentage

Mean (SD)

Median (Range/IQR)

Hypothesis Testing

- Involve more than one variables
  - exposure & outcome,
  - predictor & criterion,
  - risk & disease
- Try to prove that
  Exposure causes the Disease
  e.g. Smoking causing Lung Cancer
- Example ~ Ho: No difference of risk to get Lung Cancer between smoker and non-smoker

	Lung Cancer	No Lung Cancer
Smoking	20 (18.2%)	90 (81.8%)
Not Smoking	5 (4.5%)	105 (95.5%)

χ^2 *(df=1)= 10.150, p =0.001, OR = 4.7 (CI95% 1.7 – 13.0)*

Because $p < 0.05$, we reject H_0. Therefore there is a different between smoker & non smoker

Statistical Test

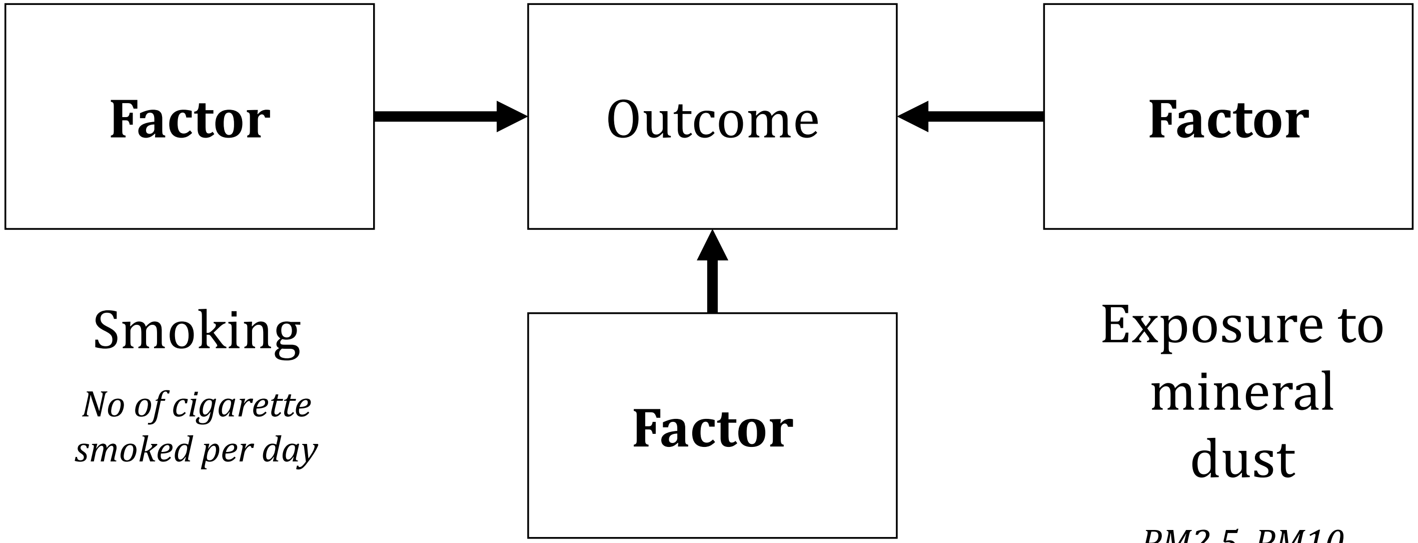
- Univariate ~ One dependent & one independent
- Multivariate ~ Multiple dependent & multiple independent variable

What test to use?

Variable 1	Variable 2	Test
Categorical	Categorical	Chi-square
Categorical (2 pop)	Numerical (Normal)	Independent sample t-test
Categorical (2 pop)	Numerical (Not Normal)	Mann-Whitney U test
Categorical (> 2 pop)	Numerical (Normal)	One-way ANOVA
Categorical (> 2 pop)	Numerical (Not Normal)	Kruskal-Wallis test
Numerical (Normal)	Numerical (Normal)	Pearson Correlation Coefficient Test
Numerical (Normal/ Not Normal)	Numerical (Not Normal)	Spearman Correlation Coefficient Test
Numerical (Normal)	Numerical (Normal) – Paired	Paired t-test
Numerical (Not Normal)	Numerical (Not Normal) – Paired	Friedman test

But life is not simple!

Lung Cancer



Factor

Outcome

Factor

Factor

Smoking

No of cigarette smoked per day

Exposure to mineral dust

PM2.5, PM10 (g/m^3)

Radiation

Radiation Absorbed dose (mGy) per day

The multivariate model

Lung CA = Smoking + Radiation + Mineral dust + Others

$$y = \underbrace{\beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}_{\text{Regression}} + \underbrace{\varepsilon}_{\text{Residual}}$$

A good model is when Regression > Residual

Multivariate Analysis

- Hypothesis testing & control for confounders
  – e.g. General Linear Model, Logistic Regression
- Modeling
  – e.g. Linear Regression
- Data reduction
  – e.g. Factor Analysis, Cluster Analysis

Writing plan for statistical analysis #1

Data were analyzed using the complex sample function of SPSS (version 13.0). Sampling errors were estimated using the primary sampling units and strata provided in the data set. Sampling weights were used to adjust for nonresponse bias and the oversampling of blacks, Mexican Americans, and the elderly in NHANES. The prevalence of hypertension, as well as the awareness, treatment, and control rates, were age adjusted by direct standardization to the US 2000 standard population.10 To analyze differences over time, the 2003–2004 data were compared with the 1999–2000 data. Estimates with a coefficient of variation >0.3 were considered unreliable. A 2-tailed P value <0.05 was considered statistically significant.

(Ong et al. 2009)

Writing plan for statistical analysis #2

To assess the effect of the selection process on the characteristics of the cases, we compared cases included in the final analysis to the rest of the cases. Since controls included in the present analysis were different from the rest of the diabetes free participants by design, no similar comparisons were performed for that group. To compare baseline characteristics of cases and controls appropriate univariate statistics were used. Similar binary logistic and multiple linear regression models were built with incident diabetes or HbA1c as respective outcomes and additive block entry of adiponectin and potential confounders. For linear regression CRP and triglycerides were log transformed. Since HbA1c could be modified by drug treatment, we ran a sensitivity analysis excluding all participants on antidiabetic medication. A p-value of <0.05 was considered significant. Analyses were performed with SPSS 14.0 for Windows.

Reporting analysis (example)

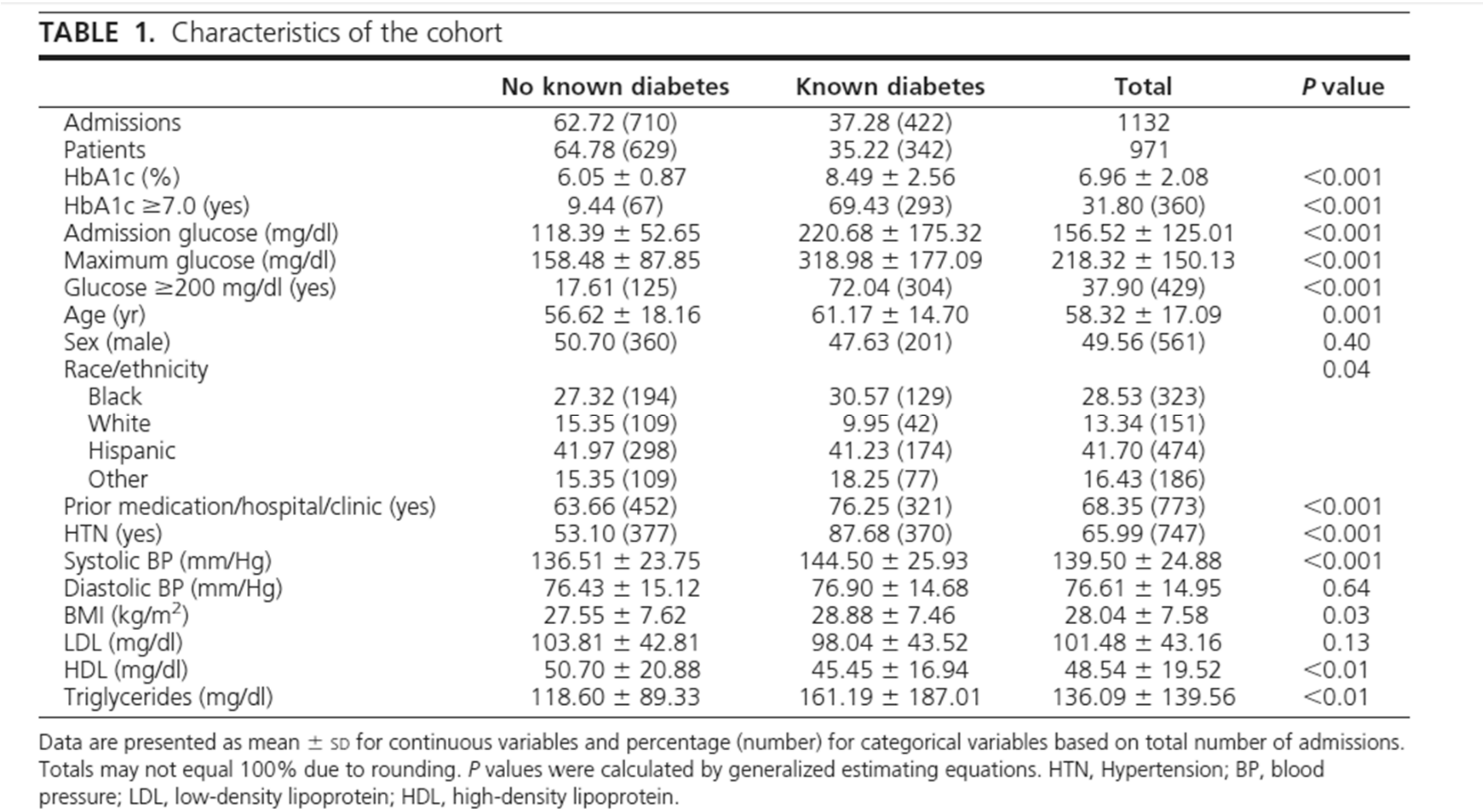


TABLE 1. Characteristics of the cohort

	No known diabetes	Known diabetes	Total	*P* value
Admissions	62.72 (710)	37.28 (422)	1132	
Patients	64.78 (629)	35.22 (342)	971	
HbA1c (%)	6.05 ± 0.87	8.49 ± 2.56	6.96 ± 2.08	<0.001
HbA1c ≥7.0 (yes)	9.44 (67)	69.43 (293)	31.80 (360)	<0.001
Admission glucose (mg/dl)	118.39 ± 52.65	220.68 ± 175.32	156.52 ± 125.01	<0.001
Maximum glucose (mg/dl)	158.48 ± 87.85	318.98 ± 177.09	218.32 ± 150.13	<0.001
Glucose ≥200 mg/dl (yes)	17.61 (125)	72.04 (304)	37.90 (429)	<0.001
Age (yr)	56.62 ± 18.16	61.17 ± 14.70	58.32 ± 17.09	0.001
Sex (male)	50.70 (360)	47.63 (201)	49.56 (561)	0.40
Race/ethnicity				0.04
Black	27.32 (194)	30.57 (129)	28.53 (323)	
White	15.35 (109)	9.95 (42)	13.34 (151)	
Hispanic	41.97 (298)	41.23 (174)	41.70 (474)	
Other	15.35 (109)	18.25 (77)	16.43 (186)	
Prior medication/hospital/clinic (yes)	63.66 (452)	76.25 (321)	68.35 (773)	<0.001
HTN (yes)	53.10 (377)	87.68 (370)	65.99 (747)	<0.001
Systolic BP (mm/Hg)	136.51 ± 23.75	144.50 ± 25.93	139.50 ± 24.88	<0.001
Diastolic BP (mm/Hg)	76.43 ± 15.12	76.90 ± 14.68	76.61 ± 14.95	0.64
BMI (kg/m^2)	27.55 ± 7.62	28.88 ± 7.46	28.04 ± 7.58	0.03
LDL (mg/dl)	103.81 ± 42.81	98.04 ± 43.52	101.48 ± 43.16	0.13
HDL (mg/dl)	50.70 ± 20.88	45.45 ± 16.94	48.54 ± 19.52	<0.01
Triglycerides (mg/dl)	118.60 ± 89.33	161.19 ± 187.01	136.09 ± 139.56	<0.01

Data are presented as mean ± SD for continuous variables and percentage (number) for categorical variables based on total number of admissions. Totals may not equal 100% due to rounding. *P* values were calculated by generalized estimating equations. HTN, Hypertension; BP, blood pressure; LDL, low-density lipoprotein; HDL, high-density lipoprotein.

Reporting analysis (example)

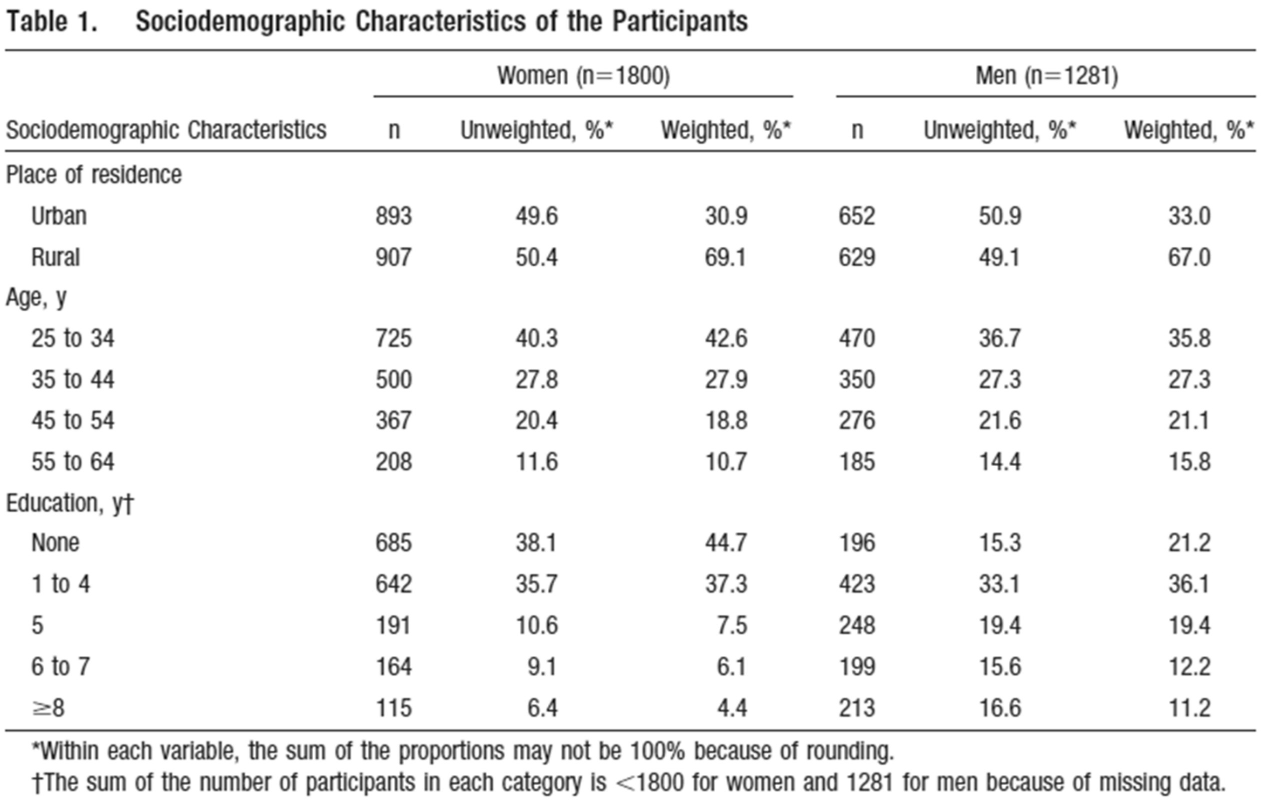


Table 1. Sociodemographic Characteristics of the Participants

	Women (n=1800)			Men (n=1281)		
Sociodemographic Characteristics	n	Unweighted, %*	Weighted, %*	n	Unweighted, %*	Weighted, %*
Place of residence						
Urban	893	49.6	30.9	652	50.9	33.0
Rural	907	50.4	69.1	629	49.1	67.0
Age, y						
25 to 34	725	40.3	42.6	470	36.7	35.8
35 to 44	500	27.8	27.9	350	27.3	27.3
45 to 54	367	20.4	18.8	276	21.6	21.1
55 to 64	208	11.6	10.7	185	14.4	15.8
Education, y†						
None	685	38.1	44.7	196	15.3	21.2
1 to 4	642	35.7	37.3	423	33.1	36.1
5	191	10.6	7.5	248	19.4	19.4
6 to 7	164	9.1	6.1	199	15.6	12.2
≥8	115	6.4	4.4	213	16.6	11.2

*Within each variable, the sum of the proportions may not be 100% because of rounding.

†The sum of the number of participants in each category is <1800 for women and 1281 for men because of missing data.

Reporting analysis (example)

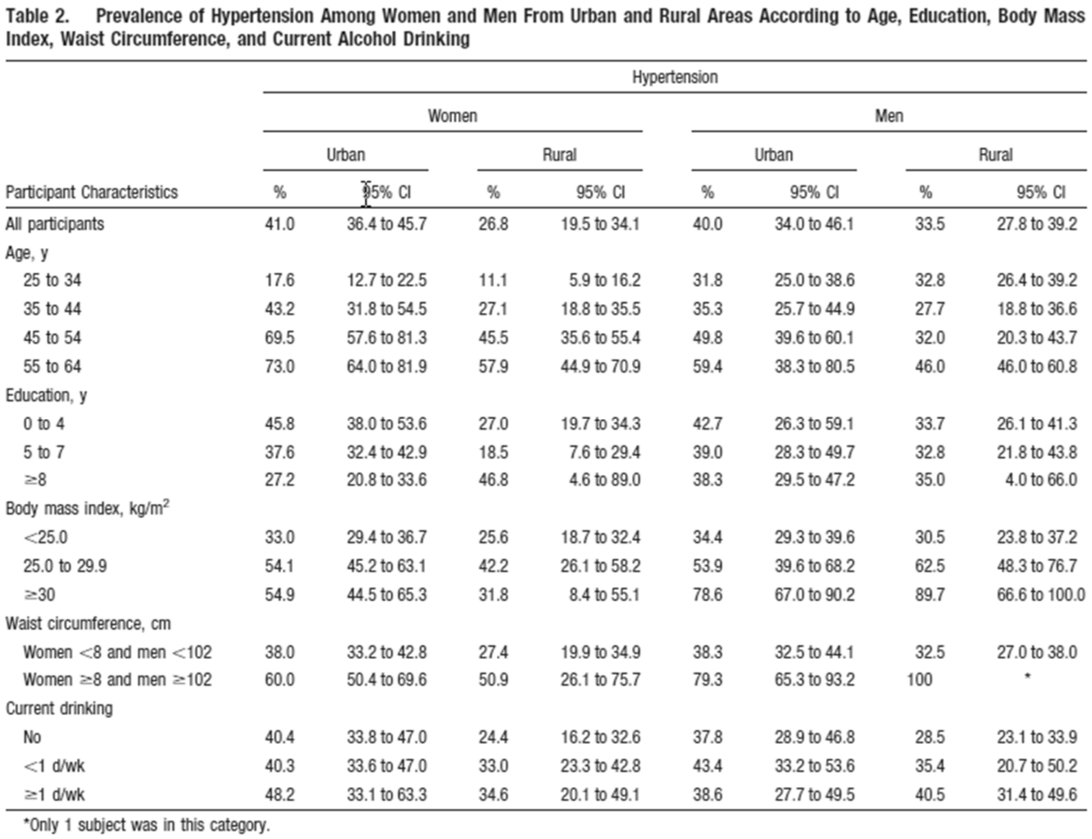


Table 2. Prevalence of Hypertension Among Women and Men From Urban and Rural Areas According to Age, Education, Body Mass Index, Waist Circumference, and Current Alcohol Drinking

	Hypertension							
	Women				Men			
	Urban		Rural		Urban		Rural	
Participant Characteristics	%	95% CI	%	95% CI	%	95% CI	%	95% CI
All participants	41.0	36.4 to 45.7	26.8	19.5 to 34.1	40.0	34.0 to 46.1	33.5	27.8 to 39.2
Age, y								
25 to 34	17.6	12.7 to 22.5	11.1	5.9 to 16.2	31.8	25.0 to 38.6	32.8	26.4 to 39.2
35 to 44	43.2	31.8 to 54.5	27.1	18.8 to 35.5	35.3	25.7 to 44.9	27.7	18.8 to 36.6
45 to 54	69.5	57.6 to 81.3	45.5	35.6 to 55.4	49.8	39.6 to 60.1	32.0	20.3 to 43.7
55 to 64	73.0	64.0 to 81.9	57.9	44.9 to 70.9	59.4	38.3 to 80.5	46.0	46.0 to 60.8
Education, y								
0 to 4	45.8	38.0 to 53.6	27.0	19.7 to 34.3	42.7	26.3 to 59.1	33.7	26.1 to 41.3
5 to 7	37.6	32.4 to 42.9	18.5	7.6 to 29.4	39.0	28.3 to 49.7	32.8	21.8 to 43.8
≥8	27.2	20.8 to 33.6	46.8	4.6 to 89.0	38.3	29.5 to 47.2	35.0	4.0 to 66.0
Body mass index, kg/m^2								
<25.0	33.0	29.4 to 36.7	25.6	18.7 to 32.4	34.4	29.3 to 39.6	30.5	23.8 to 37.2
25.0 to 29.9	54.1	45.2 to 63.1	42.2	26.1 to 58.2	53.9	39.6 to 68.2	62.5	48.3 to 76.7
≥30	54.9	44.5 to 65.3	31.8	8.4 to 55.1	78.6	67.0 to 90.2	89.7	66.6 to 100.0
Waist circumference, cm								
Women <8 and men <102	38.0	33.2 to 42.8	27.4	19.9 to 34.9	38.3	32.5 to 44.1	32.5	27.0 to 38.0
Women ≥8 and men ≥102	60.0	50.4 to 69.6	50.9	26.1 to 75.7	79.3	65.3 to 93.2	100	*
Current drinking								
No	40.4	33.8 to 47.0	24.4	16.2 to 32.6	37.8	28.9 to 46.8	28.5	23.1 to 33.9
<1 d/wk	40.3	33.6 to 47.0	33.0	23.3 to 42.8	43.4	33.2 to 53.6	35.4	20.7 to 50.2
≥1 d/wk	48.2	33.1 to 63.3	34.6	20.1 to 49.1	38.6	27.7 to 49.5	40.5	31.4 to 49.6

*Only 1 subject was in this category.

Summary

1. Identify & define variables
2. Type – independent vs. dependent
3. Level of measurements – nominal, ordinal or continuous
4. Check distribution – Normal vs. Not Normal
5. Decide what to do - descriptive vs. analytical