Week 10
Quantitative Data Analysis

PowerPoint presentation developed by:
Allen Rubin, Jennifer Manuel & Jennifer L. Bellamy

Overview

- Coding
- Descriptive Univariate Analysis
- Inferential Analysis- Statistical Significance
  - Relationships Among Variables
    - Bivariate Analysis (Chi-square, t-test, ANOVA)
    - Multivariate Analysis (linear regression and logistic regression)
    - Non-parametric tests

Coding

- Refer to the assignment of a number of numeral to the attributes of a variable
- The goal is the conversion of data items into numerical codes, necessary for statistical analyses
- After entering the data, the next step is to eliminate error – that is, “clean” the data

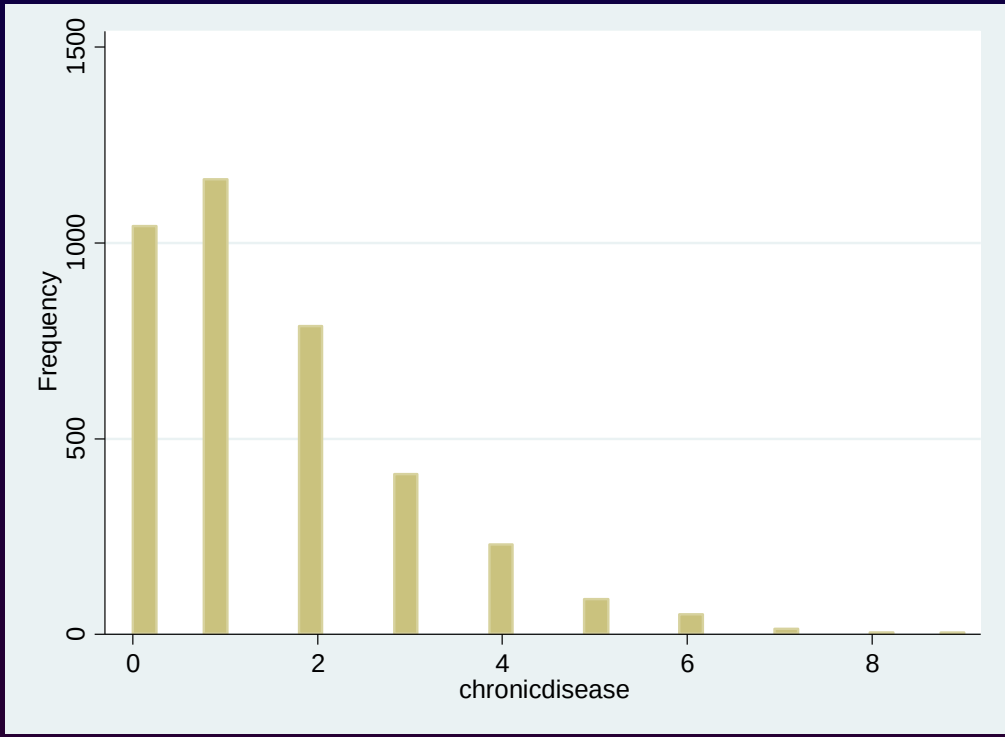
Descriptive Univariate Analysis

- Analysis of a single variable
- Frequency distributions
- Central tendency
- Variability (=Dispersion)
- Report Result Based on Levels of measurement

Descriptive Univariate Analysis

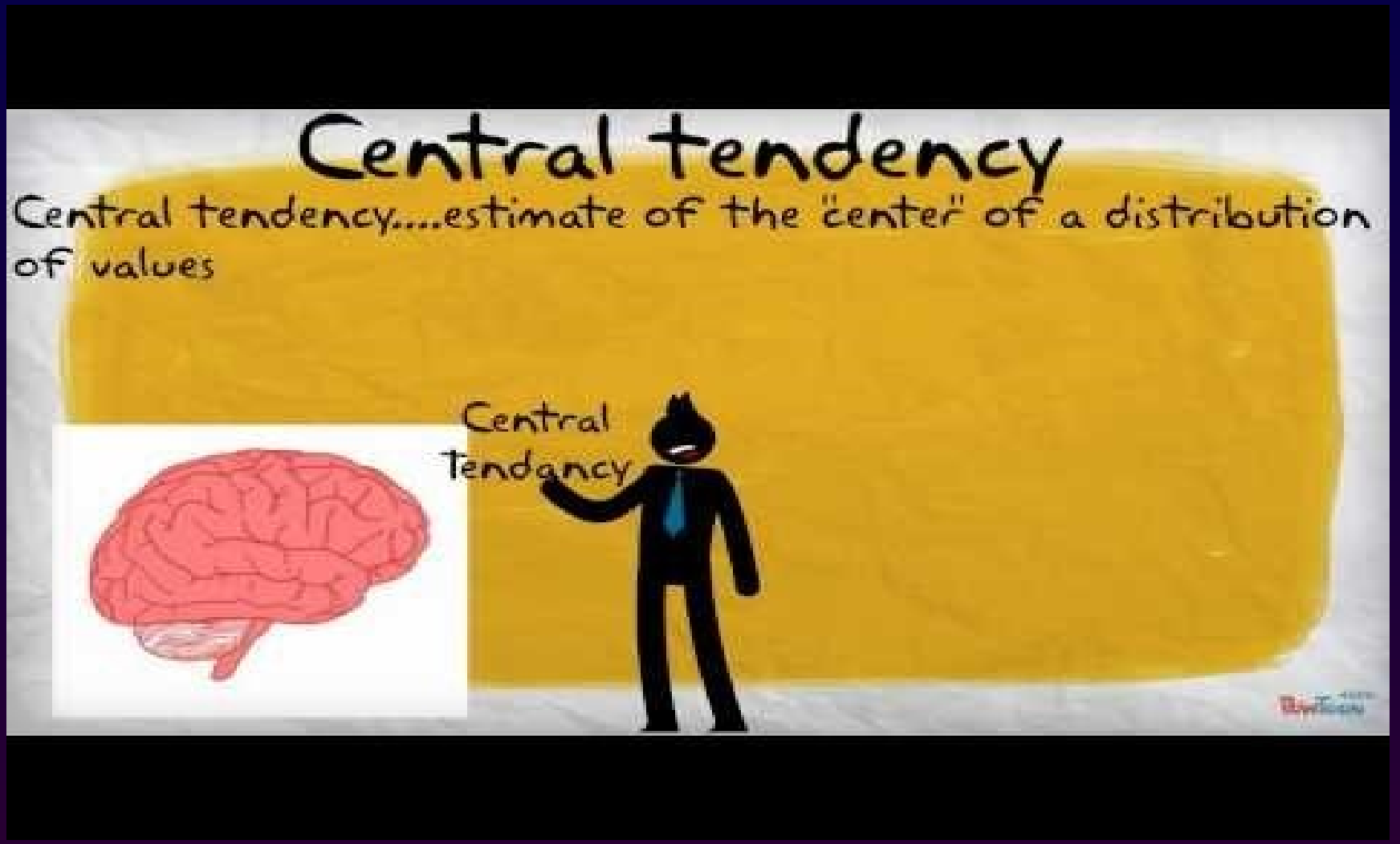
- Frequency distribution

chronicdisease	Freq.	Percent	Cum.
0	1,044	27.50	27.50
1	1,162	30.60	58.10
2	788	20.75	78.85
3	409	10.77	89.62
4	230	6.06	95.68
5	91	2.40	98.08
6	52	1.37	99.45
7	15	0.40	99.84
8	5	0.13	99.97
9	1	0.03	100.00
Total	3,797	100.00	



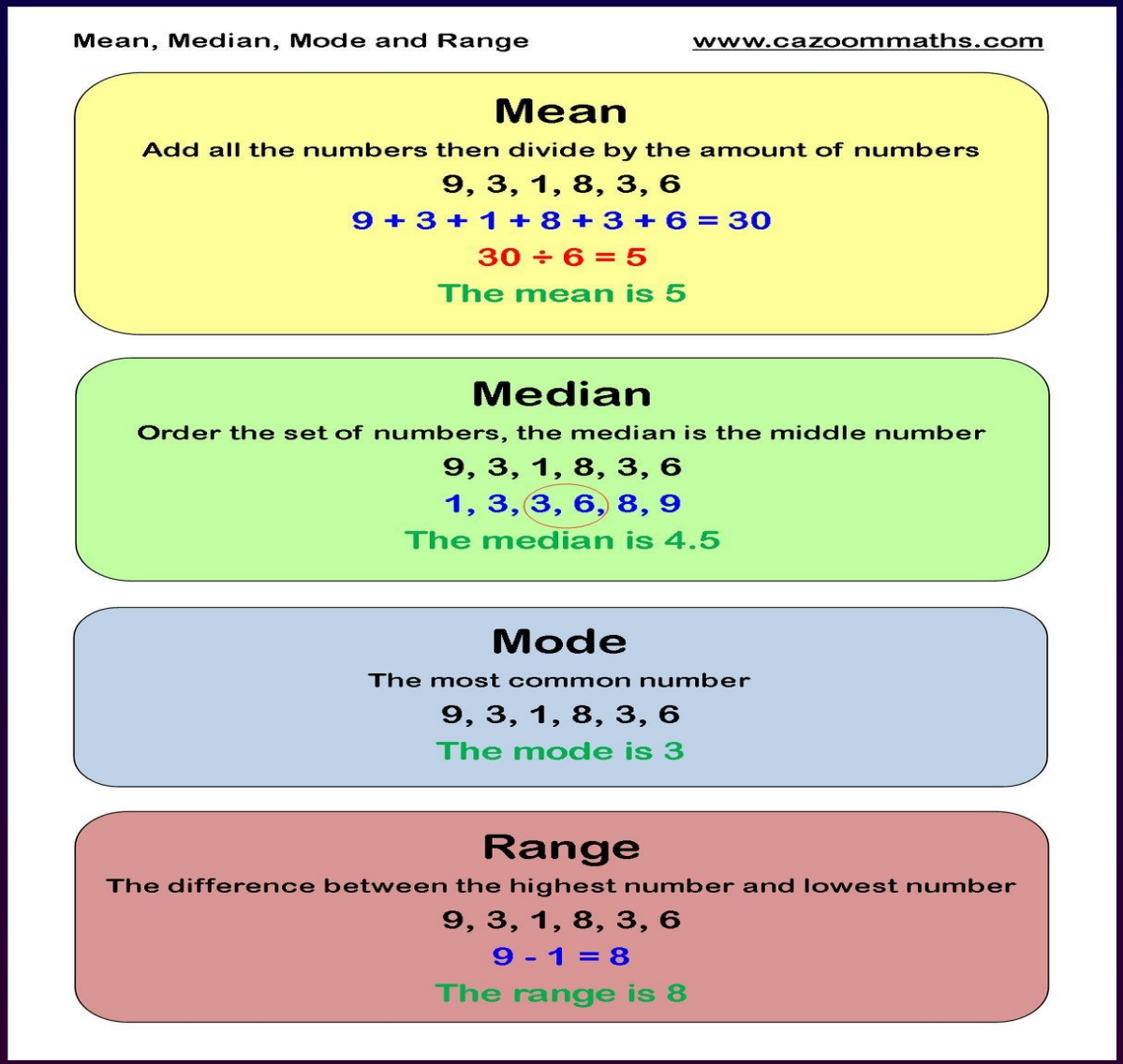
Central Tendency

- a **central tendency** (or measure of **central tendency**) is a **central** or typical value for a probability distribution. It may also be called a center or location of the distribution. Colloquially, measures of **central tendency** are often called averages.



Descriptive Univariate Analysis

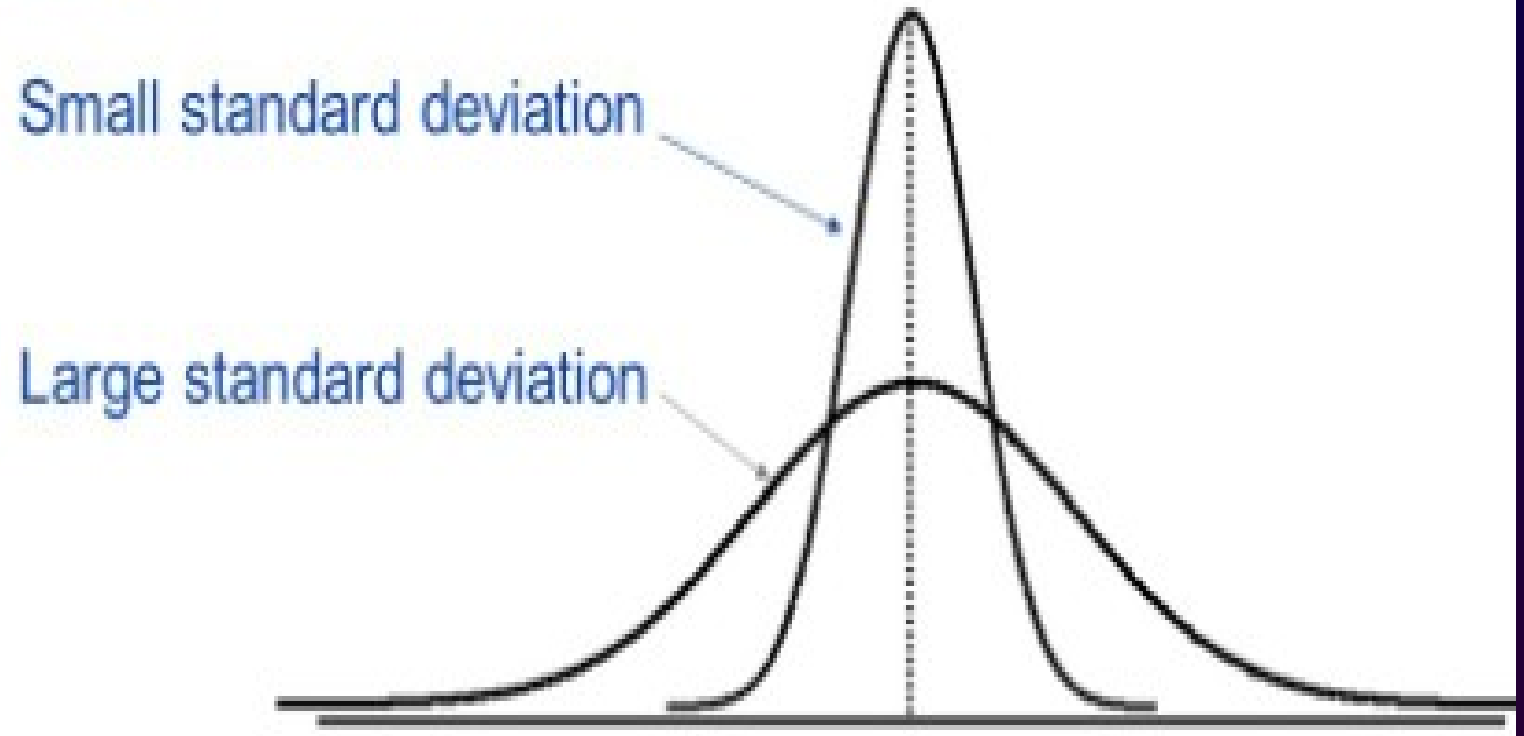
- Measures of central tendency:

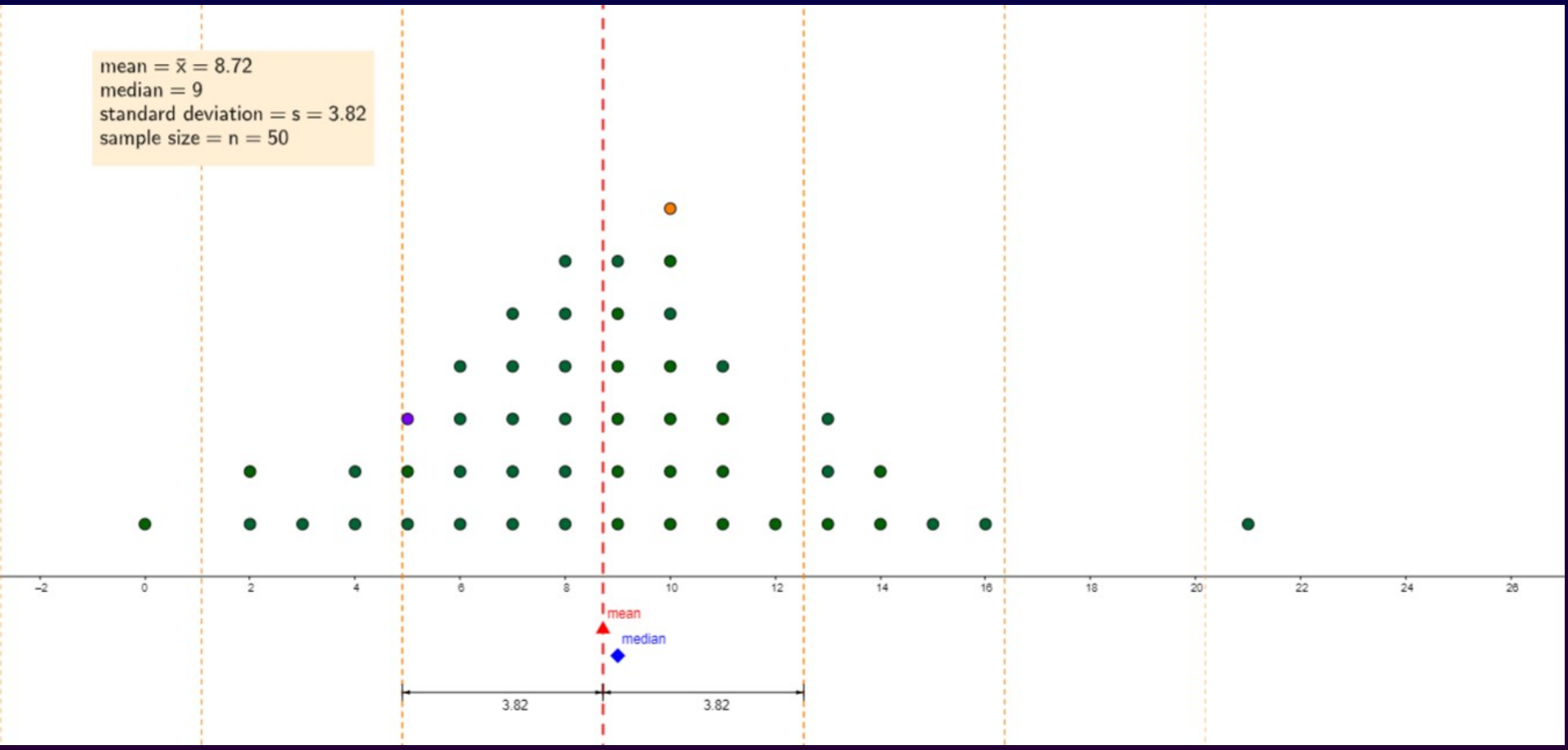


Descriptive Univariate Analysis

- Variability (=Dispersion)
  - Measures of dispersion provide a summary of the distribution of cases around some central value
    - Range, the distance between the highest and lowest value, is the simplest dispersion measure
    - Standard deviation is the most common and is used to get an idea of far away from the mean the values in our data are falling

Larger values of standard deviation indicate greater amounts of variation.




mean = x̄ = 8.72
median = 9
standard deviation = s = 3.82
sample size = n = 50
−2
0
2
4
6
8
10
12
14
16
18
20
22
24
26
mean
median
3.82
3.82

Example of Central Tendency in Report

Descriptive Statistics – Central Tendency – Standard Deviation:

	N	Minimum	Maximum	Mean	Std. Deviation
Child's age in months	4010	.00	215.00	88.1387	58.30834
Valid N	4010				

Report these

	Level of Measurement			
	Categorical	Continuous		
Type of descriptive statistics	**Nominal**	**Ordinal**	**Interval**	**Ratio**
Mode, Count, Frequency	Yes	Yes	Yes	Yes
Median, Minimum, Maximum, range	No	Yes	Yes	Yes
Mean, Variance, Standard Deviation	No	**No**/Yes	Yes	Yes

1. Descriptive Univariate Analysis

- Nominal level of Measurement
  - **D**escribes a variable in terms of the number of cases in each category of that variable.
  - Examples

    gender

    ethnicity

    religious affiliation
- Measures of central tendency and dispersion cannot be used with nominal level variables

- Report Nominal level of Measurement

	People with SMI		No SMI		
	N= 3,797	%	N= 32,621	%	χ2
Gender					496.2***
Male	990	26.1	14,241	43.7	
Female	2,807	73.9	18,380	56.3	
Race					179.0***
White	3,384	89.1	27,555	84.5	
Non-White	413	10.9	5,066	15.5	
Employment					1879.8***
Employed for salary	1,233	32.5	14,612	45.1	
Self-employed	219	5.8	2,545	7.9	
Unemployed 1 year=<	178	4.7	841	2.6	
Unemployed 1 year >	104	2.7	816	2.5	
A homemaker	208	5.5	1,833	5.7	
A student	88	2.3	792	2.5	
Retired	892	23.5	9,374	28.9	
Unable to work	871	23.0	1,574	4.9	

1. Descriptive Univariate Analysis

- Ordinal level of measurement
  - **D**escribes a variable whose categories can be rank-ordered according to how much of that variable they are.
  - Code numbers are assigned to the categories, but the precise differences or distance between the categories is unknown - we only know the *order* of the categories (e.g., high to low, more to less)
  - We know only whether one case has more or less of something than another case, but we don't know precisely *how much* more.
  - Example

  level of physical activity

1. Descriptive Univariate Analysis

Physical Activity (Moderator variable)	**A calculated variable for levels of physical activity from multiple physical activity related questions, ranging from 1 to 4, with a higher number indicating higher level of physical activity**

Physical Activity level of people with SMI	Frequency	Percent
Inactive (= 1)	1,272	34.48 %
Insufficiently inactive (=2)	795	21.55 %
Active (=3)	659	17.86 %
Highly Active (=4)	963	26.10 %
Total	3,689	100 %

However, most researchers treat an ordinal level measurement as interval

	With SMI (n=3,652)	Without SMI (n= 36,230)	
Item	Mean(SD)	Mean(SD)	t-value
Physical Activity (Rage 1-4)	2.25 (1.18)	2.52 (1.18)	-8.00***

1. Descriptive Univariate Analysis

- Interval level of measurement
  - differences between different levels have the same meanings.
  - Example: IQ. The difference between an IQ score of 95 and 100 is the same in magnitude as the difference between 100 and 105.
- Ratio level of measurement
  - Has the same attribute as interval measures, but in addition has a true zero point.
  - Example

    Number of arrests. It's possible to have no arrests, one arrest, and so on. Because there is a true zero point, we know that the person with 4 arrests has been arrested exactly twice as many times as the person with 2 arrests.

1. Descriptive Univariate Analysis

Healthy days (Dependent variable)	**An index of individual's perceived physical and mental health over time, ranging from 0 to 30 unhealthy days, with a higher number indicating higher number of healthy days**

	With SMI (n=3,652)	Without SMI (n= 36,230)
Item	Mean(SD)	Mean(SD)
Healthy days	14.27 (12.49)	24.58 (9.42)

We do not usually provide frequency table for interval or ratio level measurement

However, some

healthydays	Freq.	Percent	Cum.
0	1,209	33.12	33.12
1	3	0.08	33.19
2	22	0.60	33.79
3	15	0.41	34.20
4	4	0.11	34.31
5	51	1.40	35.71
6	16	0.44	36.14
7	16	0.44	36.58
8	31	0.85	37.43
9	21	0.58	38.01
10	89	2.44	40.44
11	14	0.38	40.83
12	26	0.71	41.54
13	33	0.90	42.44
14	18	0.49	42.94
15	142	3.89	46.82
16	45	1.23	48.06
17	19	0.52	48.58
18	37	1.01	49.59
19	17	0.47	50.05
20	143	3.92	53.97
21	31	0.85	54.82
22	49	1.34	56.16
23	113	3.09	59.26
24	79	2.16	61.42
25	186	5.09	66.51
26	102	2.79	69.30
27	145	3.97	73.27
28	201	5.50	78.78
29	99	2.71	81.49
30	676	18.51	100.00

2. Inferential Analysis: Statistical significance

- Purpose: Use inferential statistics to rule out the plausibility of chance as the explanation for the relationships observed in a study's findings

- Predicting/estimating a relationship between some key variables and the expected output variable under consideration.

- If the probability of chance is very low (usually <.05), the results are called statistically significant

Relationships Among Variables

- Bivariate analysis

  The analysis of two variables (ordinal, interval, ratio) finding association or difference

1. Pearson's Correlation analysis

Correlations

		Height	Weight
Height	Pearson Correlation	1	.513**
	Sig. (2-tailed)		.000
	N	A 408	B 354
Weight	Pearson Correlation	.513**	1
	Sig. (2-tailed)	.000	
	N	C 354	D 376

**. Correlation is significant at the 0.01 level (2-tailed).

A Co
observations for height (n=408)

B Correlation of height and weight (r=0.513), based on n=354

C Correlation of height and weight (r=0.513), based on n=354

D Correlation of weight with itself (r=1), and the number of nonmissing observations for weight (n=376).

Relationships Among Variables

Reporting result of correlation analysis- Height and weight are statistically correlated ($r=0.513$, $p<.001$)

- Interpreting measures of association (relationship strength)
  - Measures of association can be in the form of:
    - Correlations (ranging between 0 and +1 or 0 and -1)

- Strong, medium, and weak effect sizes (approximate general guidelines)

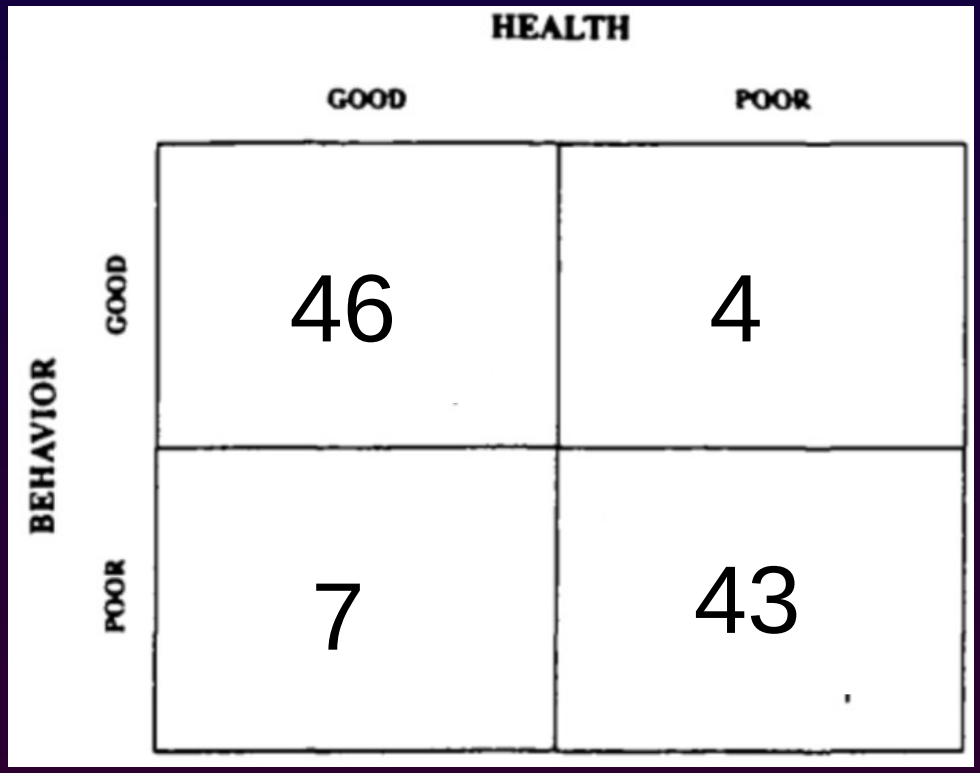
	strong	medium	weak
Correlations:	$\geq .50$	.30	$\leq .10$

Relationships Among Variables

2. Chi-square (χ^2) test - finding association between two variables (nominal) by comparing differences between groups.

N (sample size)= 100

Hypothesis- people with good health behavior rate their health as good.



Relationships Among Variables

Reporting result of Chi-square test

A chi-square test of independence was performed to examine the relationship between good health behaviors and perceived health status. The relation between these variables was significant, χ^2 (2, N = 100) = 14.14, p <.01

Relationships Among Variables

- Bivariate analysis

3. t-test: compares the difference in the means of the dependent variable (ordinal, interval, and ratio) of two groups
   - Paired samples t-test: pre and post test
   - Independent t-test: two different groups

	With SMI (n=3,652)	Without SMI (n= 36,230)	
Item	Mean(SD)	Mean(SD)	t-value
Healthy days	14.27 (12.49)	24.58 (9.42)	-8.00***

the dependent variable (ordinal, interval, ratio) of > 2 groups

Relationships Among Variables

- Multivariate Analysis
  - A more complex method that involves analyzing the relationships among several variables
  - E.g., examining the relationship between an independent and dependent variable while controlling for extraneous variables
  - E.g., Finding contributing factors (IVs) affecting DV while controlling for extraneous variables
  - Statistical methods: Linear regression and logistic regression

Relationships Among Variables

- Multivariate analysis

1. Multiple Regression- examine the associations between multiple IVs and DV; Use this when DV is ordinal, interval, ratio

Source	SS	df	MS
Model	222487.577	6	37081.2629
Residual	318408.938	3479	91.523121
Total	540896.515	3485	155.207035

Number of obs	= 3486
F(6, 3479)	= 405.16
Prob > F	= 0.0000
R-squared	= 0.4113
Adj R-squared	= 0.4103
Root MSE	= 9.5668

Check these

healthydays	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
chronicdisease	-1.359157	.1199171	-11.33	0.000	-1.594272	-1.124042
physicalactivity	.9167306	.139135	6.59	0.000	.6439362	1.189525
psychologicaldistress	-1.230764	.0298529	-41.23	0.000	-1.289295	-1.172233
race	.4688283	.5285922	0.89	0.375	-.5675539	1.50521
gender	.7205735	.371252	1.94	0.052	-.0073202	1.448467
employment	-.12751	.1437638	-0.89	0.375	-.40938	.15436
_cons	23.49185	.7038088	33.38	0.000	22.11193	24.87177

Relationships Among Variables

Reporting results of linear regression

This study found that chronic diseases significantly predicted Healthy Days (B= 1.36, *p* <.05), even after controlling for psychological distress, physical activity levels, race, gender and employment. This result indicates that there was a decrease of 1.36 Healthy Days for each additional chronic disease amongst people with mental illness, after controlling for covariates and other key factors of interest.

Relationships Among Variables

2. Multivariate Analysis

- Logistic Regression- Use this when DV is categorical
  - Odds Ratios (OR): How much more or less likely a certain dependent variable (outcome) is for the categories of the independent variable.

Table 3. Hierarchical Logistic Regression Analysis Results for Association of Unhealthy Lifestyle Behaviors and Perceived Health Status (Incorporating weights) (N=1,277)

Variables	Model 1	Model 2
Socio-demographic Factors	OR [CI]	OR [CI]
Age		
18-34		
35-49	66[0.26-1.65]	0.62 [0.23-1.66]
50-64	66[0.29-1.51]	0.80 [0.31-2.08]
65+	56[0.23-1.41]	0.83 [0.28-2.44]
Gender		
Male		
Female	0.85[0.46-1.58]	0.82[0.46-1.48]
Unhealthy lifestyle behaviors		
Physical inactivity		0.31[0.10-0.96]*
Unhealthy diet/vegetable consumption		0.23[0.10-0.56]***
Binge drinking		2.56 [1.21-5.42]*
Current smoking		0.81[0.43-1.53]

Note. Values are expressed as OR (odds ratios) and CI (95% confidence intervals). * $P \leq 0.05$, ** $P \leq 0.01$, and *** $P \leq 0.001$. All asterisks are compared to the first group.

Check this

Relationships Among Variables

Reporting results of logistic regression

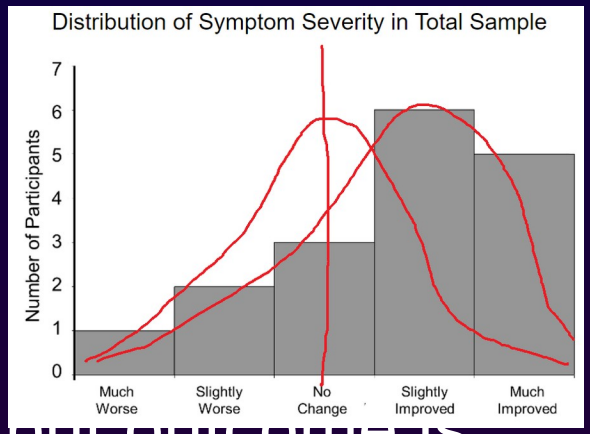
- Physical inactivity was significant predictors of general health condition, indicating individuals who did not walk at least 150 minutes per week for leisure and exercise in past week were less likely to rate their perceived health status as higher than poor, as compared to their counterparts (OR=0.31, 95% CI [0.10, 0.96]).

- The odds of rating their health as poor for people with SMI who does not walk at least 150 minutes per week is 0.31 times that of individuals with SMI who walk at least 150 minutes per week

Non-Parametric Tests

- Called distribution-free tests- Assumptions of parametric test are violated
  - when not normal distribution
  - When outcome is an ordinal

  Variable (relative size)
  - When many outliers exist
  - When sample size less than 30, but outcome is normally distributed



- Advantage- above the reasons
- Disadvantage- Less efficient; Results may not provide an accurate answer

Parametric and Nonparametric Tests

BASIS FOR COMPARISON	PARAMETRIC TEST	NONPARAMETRIC TEST
Meaning	A statistical test, in which specific assumptions are made about the population parameter is known as parametric test.	A statistical test used in the case of non-metric independent variables, is called non-parametric test.
Measurement level	Interval or ratio	Nominal or ordinal
Measure of central tendency	Mean	Median
Information about population	Completely known	Unavailable

Parametric and Nonparametric Tests

Parametric Tests	Nonparametric Tests
Pearson Correlation	Spearman Correlation
Independent Samples T-Test	Mann-Whitney Tests
Paired Samples T-Test	Wilcoxon Signed-Rank Test
One-way ANOVA	Kruskal-Wallis Test
	Chi-square Test