Biostatistics 612
(Statistical Analysis II)

Outline

⇒ Types of Random Variables

⇒ Descriptive Statistics:

- ⇒ Univariate
- ⇒ Bivariate

⇒ Example in JMP.

Types of Random Variables

$\Rightarrow$ Data sets typically include different types of random variables (RVs).

$\Rightarrow$ The type of RV will determine what descriptive statistic/ graph we use to describe it.

Types of Random Variables

⇒ **Quantitative (usually continuous).** Can take on an infinite number of values, and there is a notion of order and distance between values (e.g., height)

⇒ **Categorical**; take on a finite number of values (e.g., presence/absence of disease).

⇒ Categorical RVs can be divided into various subtypes.

Types of Categorical Random Variables

Binary (presence/absence)

Polychotomous/Multinomial (Green, Red, Blue...); there is no natural order between the levels.

Ordered Multinomial (e.g., small/medium/large); there is a natural order between the levels.

Example

CPS5 data set (Goldberger 1998, adapted from Berndt 1991)
http://www.hup.harvard.edu/features/golint/CPS5.txt.

⇒ This data comprise information from 528 people surveyed in 1985.

⇒ The variables included in the data set are:

- years of education (integer, but can be analyzed as continuous)
- years of experience in the labor market (integer, but can be analyzed as continuous)
- wage USD/hour (continuous)
- Sex (Male/Female, binary)
- Region (South/non-South, binary)
- Marital status (married/ not married)

⇒ We illustrate regression analysis to quantify effects of education on wages, after accounting for differences due to sex, and region.

The first rows of the CPS5 data set…

education	south	ehtnicGroup	female	married	experience	unionized	hourlyWage
10	0	White	0	1	27	0	9
12	0	White	0	1	20	0	5.5
12	0	White	1	0	4	0	3.8
12	0	White	1	1	29	0	10.5
12	0	White	0	1	40	1	15
16	0	White	1	1	27	0	9
12	0	White	1	1	5	1	9.57
14	0	White	0	0	22	0	15
8	0	White	0	1	42	0	11

Quantitative

- Education
- Experience and
- Hourly-wage

Multinomial:

- Ethnicity

Binary:

- Region (south=1)
- Sex (female=1)
- Married (yes=1)
- Unionizes (yes=1)

Descriptive Analysis

⇒ The 1st step in any statistical analysis consist on performing descriptive statistics/graphs.

⇒ The objectives of such an analysis is to:

(a) detect potential problems (e.g., coding errors) and

(b) get insights into the associations between variables.

⇒ We will use:

- Statistics (functions of the data, e.g. minimum, maximum, mean..).
- Graphs.

⇒ There are two basic types of descriptive statistics

- Univariate; these are used to describe RVs, one at a time.
- Bivariate; thee are used to describe the association between two RVs.

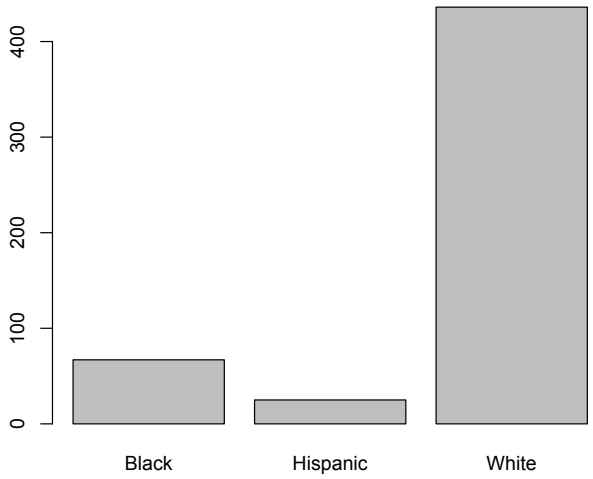
Univariate Analysis for Discrete RVs

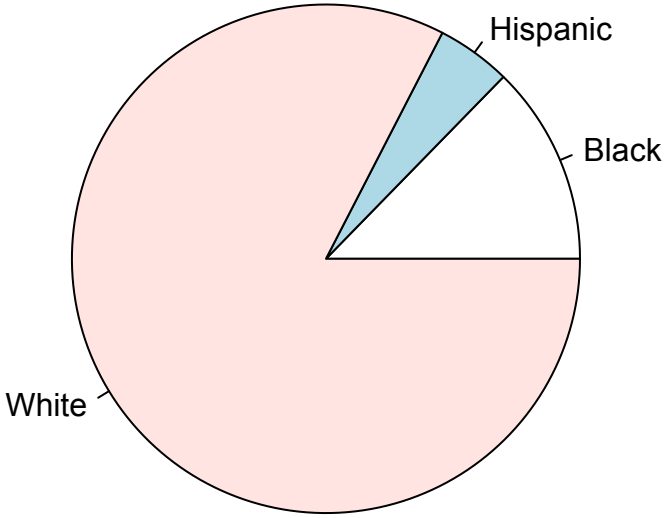
Commonly used statistic: frequency tables

Black	Hispanic	White
67	25	436

Black	Hispanic	White
12.7%	4.7%	82.6%

Commonly used graphs: bar and pie charts.





Univariate Analysis for Continuous RVs

Commonly used Statistics:

- Central measures (e.g., mean, median or other percentiles)

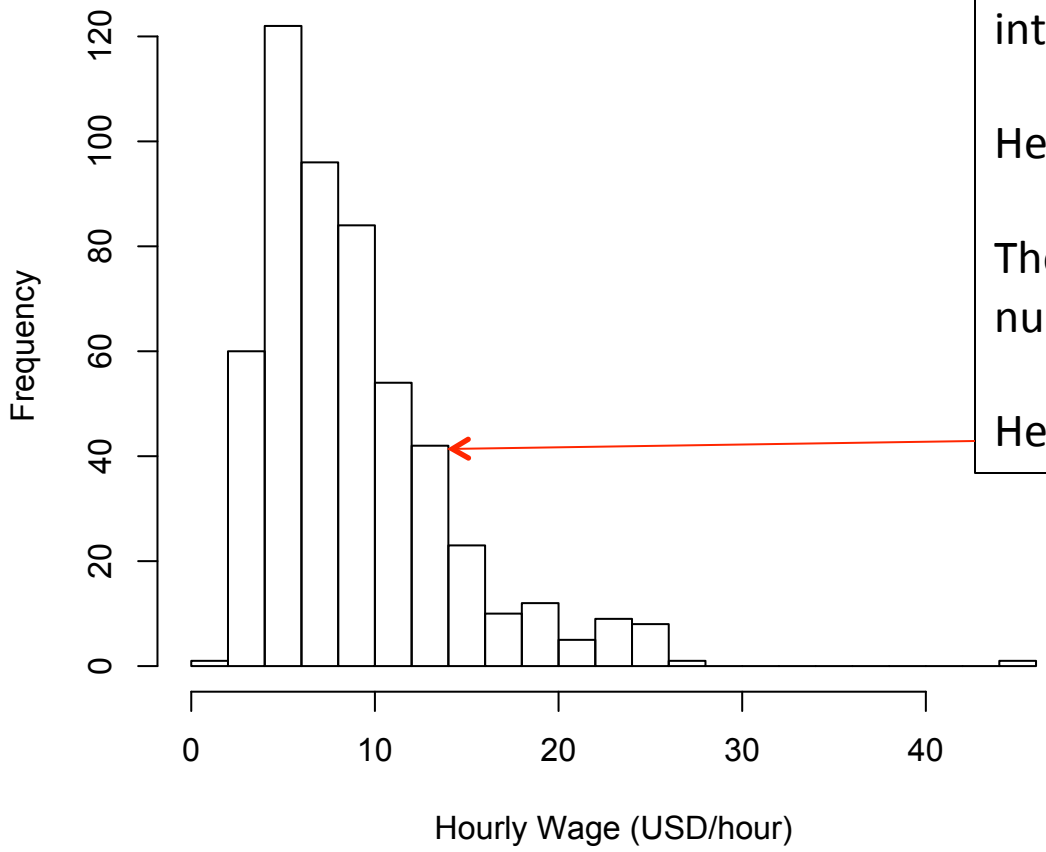
- Measures of dispersion (e.g., range, variance, standard deviation…)

Hourly Wage:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.75	5.25	7.79	9.05	11.25	44.50

Univariate Descriptive Statistics & Graphs: Continuous RVs

Commonly used graph: Histogram



The variable on the x-axis is divided into bins of equal width.

Here, each bin has $2 increments.

The height of each bar reflects the number of subjects falling in each bin.

Here, 40 people make $12-14 USD/hr

Bivariate Descriptive Analysis

⇒ The objective is to describe patterns of associations between two RVs.

⇒ We will use both statistics (e.g., correlation) and graphs.

⇒ In univariate descriptive analysis the focus is on the marginal distribution of a RV.

⇒ In bivariate analysis we focus on the joint distribution of two RV, and on the conditional distribution of one RV given the other RV.

⇒ The type of statistic/graph we use depends on the type of RV.

Bivariate Descriptive Analysis By Type of RV

	Discrete	Continuous
Discrete	Contingency Tables	Box-Plots; Conditional Means
Continuous		Scatter-Plots Co-variance & Correlation

Contingency Tables
(Discrete By Discrete RV)

Contingency Tables
(Discrete By Discrete RV)

Counts

	Black	Hispanic	White	Total
South	27	11	116	154
Other	40	14	320	374
Total	67	25	336	528

Frequencies

	Black	Hispanic	White	Total
South	5.2%	2.1%	22.0%	29.2%
Other	7.6%	2.7%	60.6%	70.8%
Total	12.7 %	4.7%	82.6%	100.0%

Joint Distribution

Marginal Distributions

Conditional Distributions

Ethnic Group Given Region

	Black	Hispanic	White	Total
South	10.7%	3.7%	85.6%	100%
Other	17.5%	7.1%	75.3%	100%

Region Given Ethnic Group

	South	Other
Black	40.3%	59.7%
Hispanic	44.0%	56.0%
White	26.6%	73.4%

Do these descriptive statistics show association between these two RVs?

Conditional Distributions With Barplots

Continuous and Discrete

Conditional Statistics

⇒ **Conditional Mean:** the average value of one RV for a given value of the other RV.

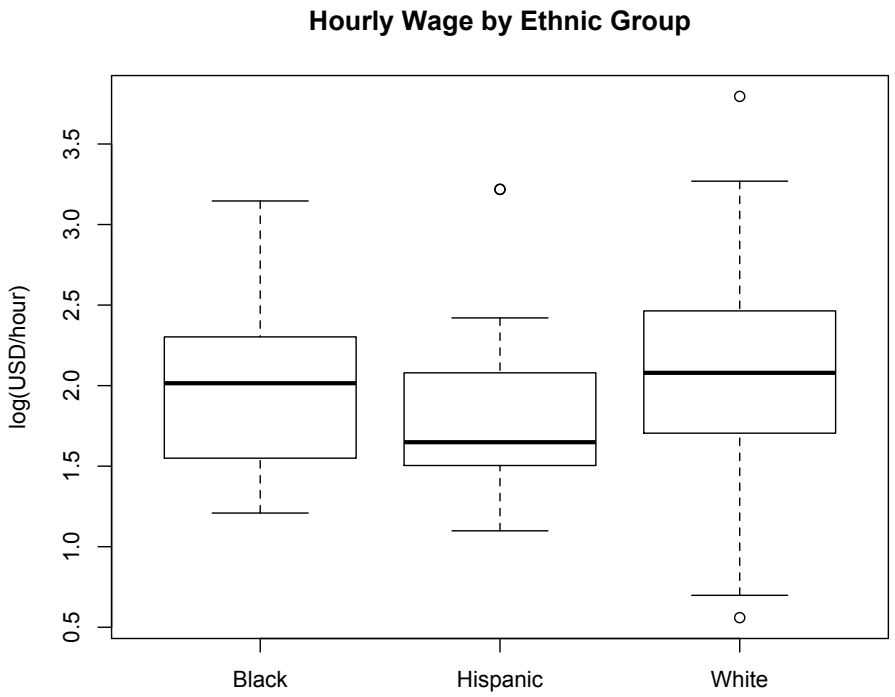
⇒ The same idea can be applied to other statistics (min, max, sd)

⇒ Example: Conditional Min, Mean, Median, Max and SD of Hourly Wage Given Sex.

	Min	Mean	Median	Max	SD
Male	2.01	10.1	9	26.3	5.3
Female	1.75	7.9	6.73	44.5	4.7

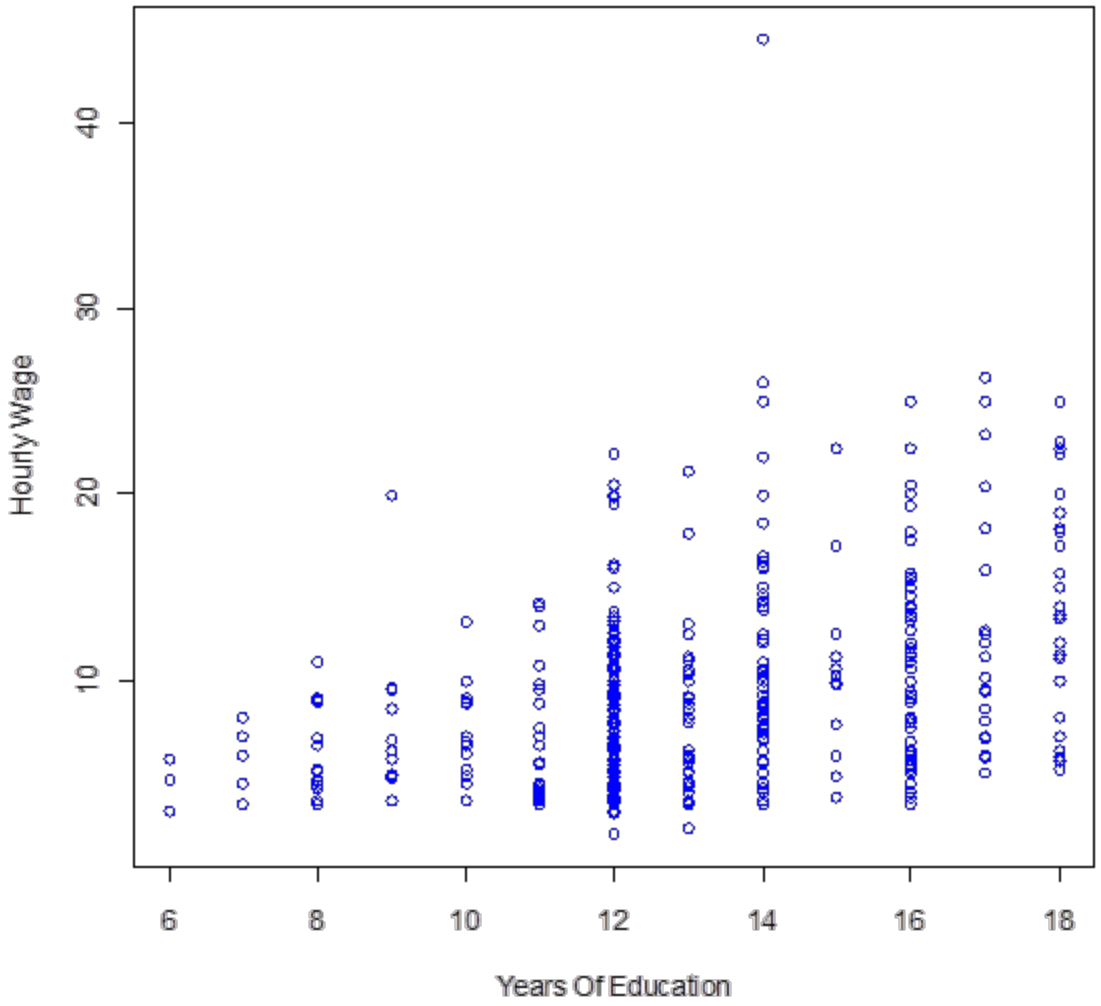
Boxplot

⇒ Displays the percentiles of the conditional distribution of a continuous RV by level of a discrete RV.



Two Continuous RVs

Scatter-plot



What do we see?

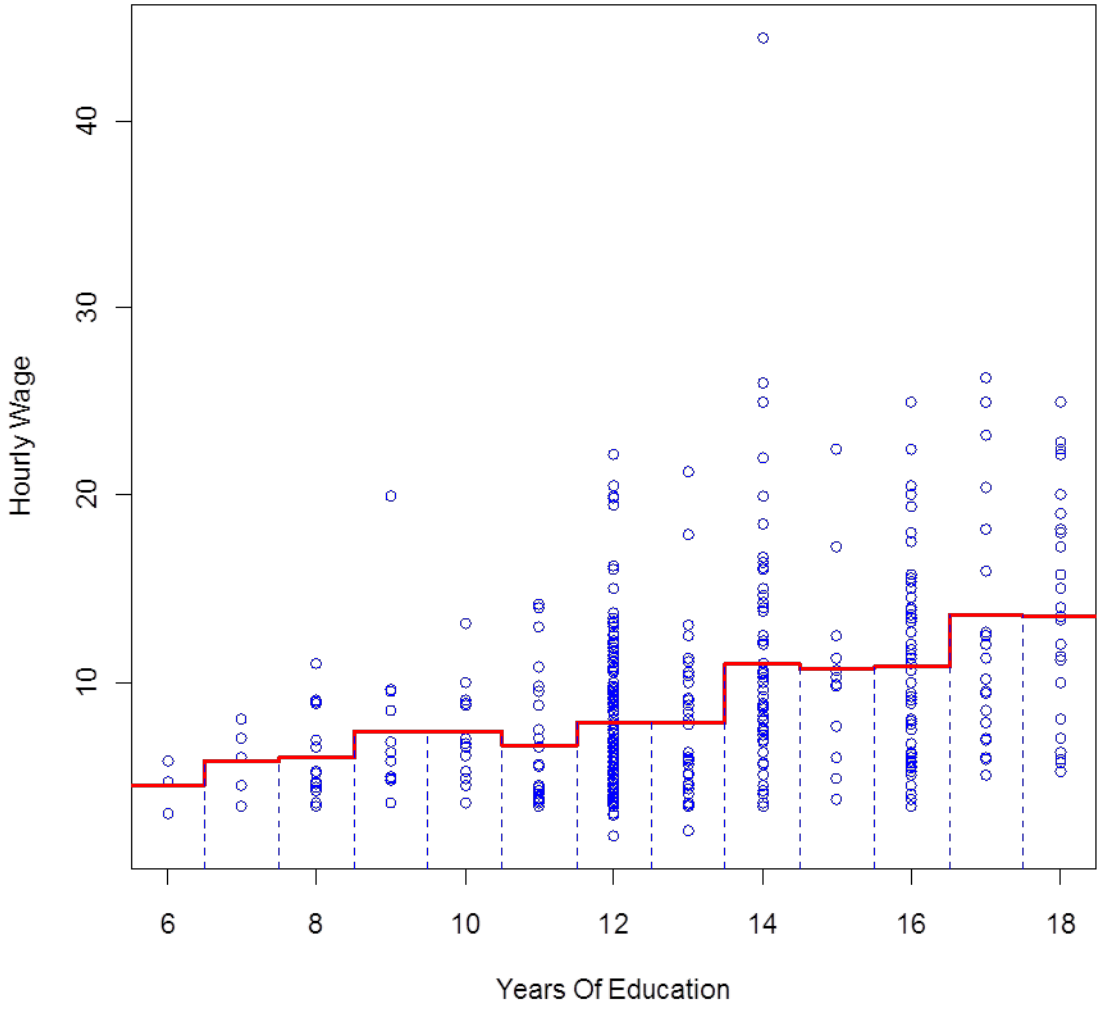
⇒ Variability of wages increases with years of education.

⇒ The average wage also seems to increase with years of education.

⇒ How do we quantify this association?

⇒ How do we estimate conditional means in this case?

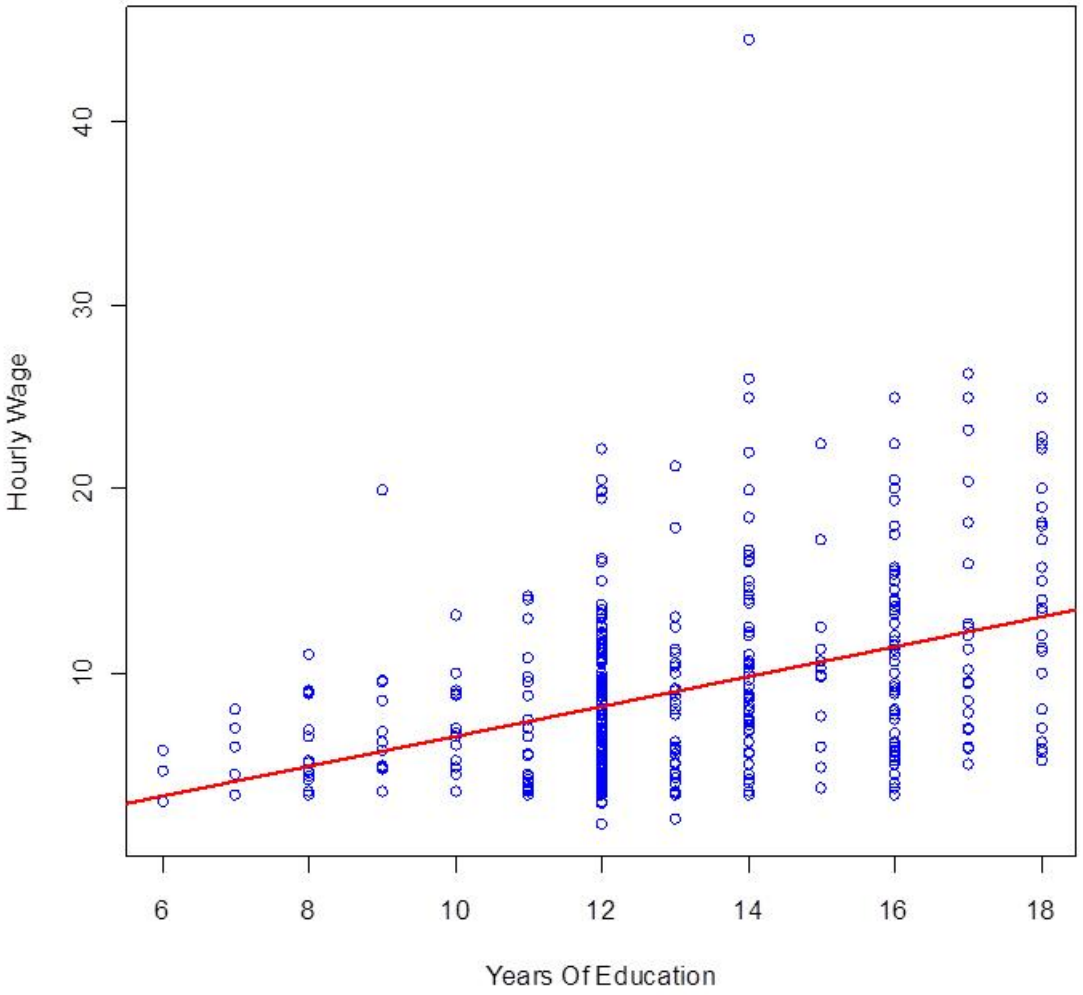
Conditional Mean



⇒ The estimated conditional mean suggest an increase in wage associated to increased years of education.

⇒ One of the central topics of this class will be how to approximate this conditional mean using linear methods (see next)

Linear Approximation to the conditional mean



⇒ Next class we will discuss how to quantify association between two quantitative variables using linear methods (co-variance, correlation, regression)

Summary

⇒ The 1st step of any statistical data analysis is to perform a descriptive statistics analysis.

⇒ To this end we use statistics (means, frequencies, etc.) and graphs.

⇒ The type of statistic/graph we use depends on the type of RV.

⇒ RVs can be classified in quantitative and discrete. Within discrete there are various subtypes.

⇒ We perform univariate and bivariate descriptive analysis.

⇒ There are also multivariate methods, but we will not use them much in this course.

⇒ Univariate descriptive analysis focuses on the marginal distribution of a RV and functions of it (e.g., mean, variance, frequencies, percentiles).

⇒ Bivariate analysis focuses on the association patterns between two RVs.

⇒ Here we focus on the joint, and mainly on the conditional distribution of Y given X.