

Applied Epidemiology I: Summary Statistics and Graphs

Enoch Yi-Tung Chen

Department of Medical Epidemiology and Biostatistics, Karolinska Institutet

November 20, 2020

Acknowledgements

This course material in data clearance is based on my learning from Anastasia Lam's teachings in last year's Applied Epidemiology I lab sessions, and readings from *A First Course in Probability and Statistics* by Goldsman and Goldsman [1], *Principles of Biostatistics* by Pagano and Gauvreau [2], and *Biostatistics I* by Gabriel and Frumento [3].

Outline

① Summary statistics

- Bad example

- Measures of Central Tendency: mean, median, mode

- Measures of Dispersion: range, IQR, variance, standard deviation

② Graphs

- Bad examples

- Learning from errors

- Study map

- Histogram

- Bar chart

- Scatter plot

- Box plot

- Line graph

- Stratification

- Putting graphs together

- Export

③ References

Summary statistics: Bad example

What is the problem here?

Table 5
Simulation results for using full data, CRs only, and proposed method under four missing mechanisms

Method	Bias ^a		Variance ^b		95% CI ^c	
	$(\hat{\beta}_W)$	$(\hat{\beta}_X)$	$(\hat{\beta}_W)$	$(\hat{\beta}_X)$	$(\hat{\beta}_W)$	$(\hat{\beta}_X)$
(M.1) $P(R = 1) = 0.66$						
Full	0.01346	0.02229	0.04008	0.03685	0.955	0.950
Comp	0.03062	-0.003561	0.1149	0.06732	0.960	0.955
Impu	0.01431	0.021	0.04088	0.05169	0.980	0.975
(M.2) $\text{logit } P(R = 1) = 2Y$						
Full	0.007908	-0.02116	0.03838	0.03624	0.975	0.925
Comp	0.01945	0.07096	0.107	0.06581	0.960	0.950
Impu	0.006966	0.01597	0.04227	0.05226	0.975	0.985
(M.3) $\text{logit } P(R = 1) = 2X$						
Full	0.007908	-0.02116	0.03838	0.03624	0.975	0.925
Comp	0.01225	0.0589	0.08856	0.06818	0.980	0.975
Impu	0.009563	-0.04699	0.03865	0.04923	0.985	0.970
(M.4) $\text{logit } P(R = 1) = X + Y$						
Full	0.01346	0.02229	0.04008	0.03685	0.955	0.950
Comp	0.02404	1.613	0.1102	0.08202	0.955	0.580
Impu	0.01814	0.08289	0.0578	0.06075	0.955	0.970

^aBias = $(\hat{\beta} - \beta_0)/\beta_0$.

Summary statistics: Measures of Central Tendency: mean, median, mode

- Mean: the sum of the values of a variable and dividing by number of the observations
- Median: the middle (the 50th centile) observation
- Mode: the value that occurs most frequently

```
. tabstat age // will only give you mean
```

variable	mean
age	56.41176

```
. tabstat age, s(count mean median) // s stands for statistics
```

variable	N	mean	p50
age	34	56.41176	56

Summary statistics: Measures of Dispersion: range, IQR, variance, standard deviation

- Range: the difference between the maximum and the minimum
- Interquartile range: the absolute difference between the 25th percentile of the observations and the 75th.
- Variance, standard deviation (sd): a measure of spread of the data

$$s^2 = \widehat{Var}(x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

```
. tabstat age, s(count range min max iqr var sd)
```

variable	N	range	min	max
age	34	20	47	67

variable	iqr	variance	sd
age	10	36.12834	6.010686

Graphs: Bad examples

Graphs can say more than texts! But it depends.....

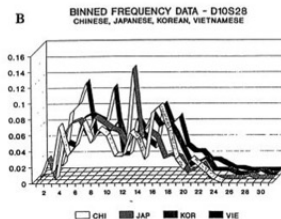
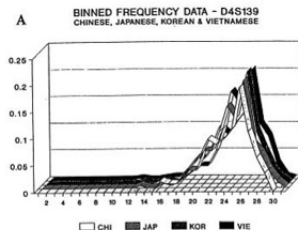
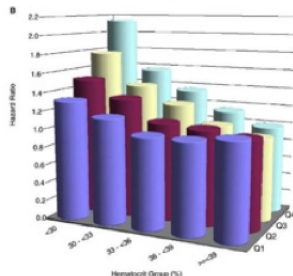
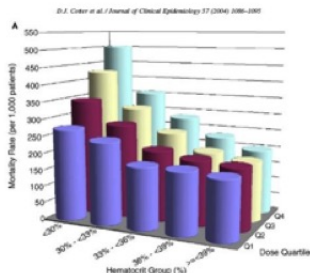


FIG. 4. Fixed bin distribution (histogram) for two loci and four Asian subpopulations (used with permission from John Hartmann): the boundaries of the 30 bins (vertical axis) are determined by the FBI; these bins are not of equal length. Sample sizes (numbers of individuals) for Chinese, Japanese, Korean and Vietnamese are 103, 125, 93 and 215 for D4S139 and 120, 137, 100 and 193 for D10S28. The horizontal axis is the bin number; bins are not of equal length.

Graphs: Bad examples



Graphs: Bad examples

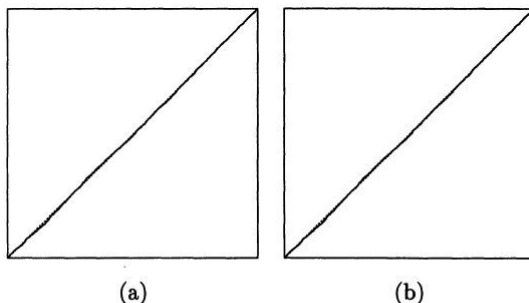
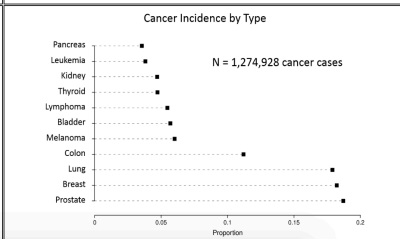
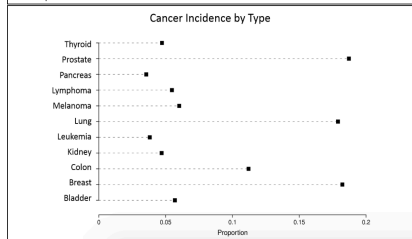
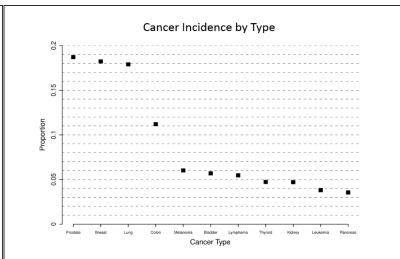
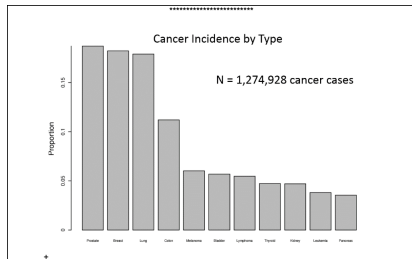


Figure 1. SRQ Plots of T_i/T_n (Vertical Axes) Against i/n (Horizontal Axes) for the Gibbs Sampler (a) and an Alternating Gibbs/Independence Sampler (b) for the Pump Failure Data Based on Runs of Length 5,000. Lines through the origin with unit slope are shown dashed; axis ranges are from 0 to 1 for all axes.

Graphs: Bad examples

Sometimes there is no right nor wrong, it just depends on your interest.

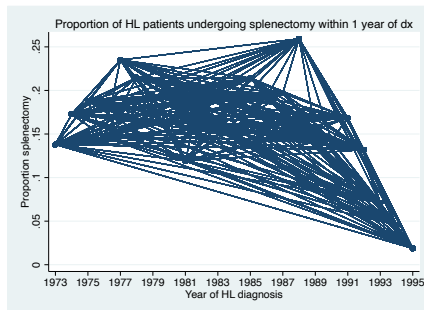


Graphs: Learning from errors

Which part went wrong here?

Hint: something was missed in the code.

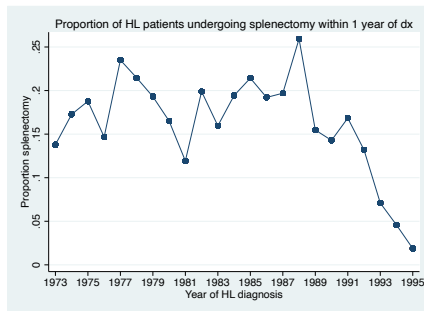
```
twoway connected prop diagyear, ///  
  subtitle("Proportion of HL patients") ///  
  ytitle(Proportion splenectomy) ///  
  xlabel(1973(2)1995)
```



Graphs: Learning from errors

It makes such a big difference if you missed sort!

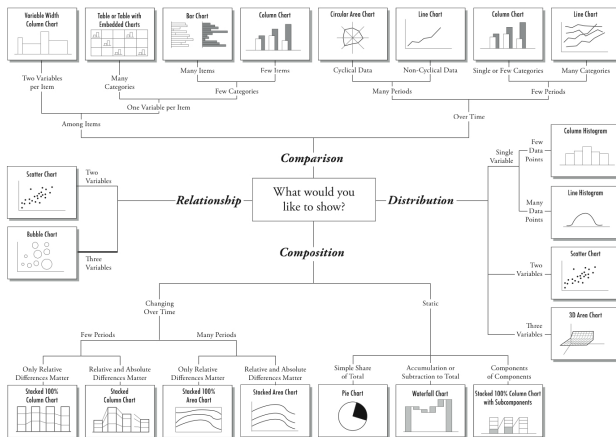
```
twoway connected prop diagyear, ///  
  subtitle("Proportion of HL patients") ///  
  ytitle(Proportion splenectomy) ///  
  xlabel(1973(2)1995) ///  
  sort
```



Graphs: Study map

Check the webpage: <https://extremepresentation.com/tools/>

Chart Suggestions—A Thought-Starter

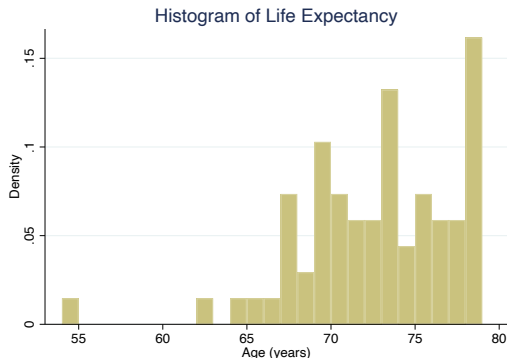


www.ExtremePresentation.com
© 2009 A. Ahela — a.ahela@gmail.com

Graphs: Histogram

Histogram depicts the distribution of data, where x-axis is usually a continuous variable.

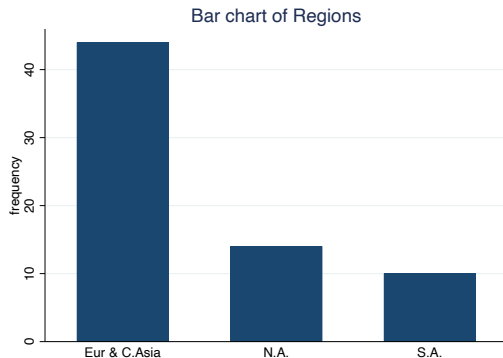
```
hist lexp, title("Histogram of Life Expectancy") ///  
    xtitle(Age (years)) width(1) /// By each age  
    graphregion(color(white)) //
```



Graphs: Bar chart

Bar chart shows the distribution of discrete (categorical) data.

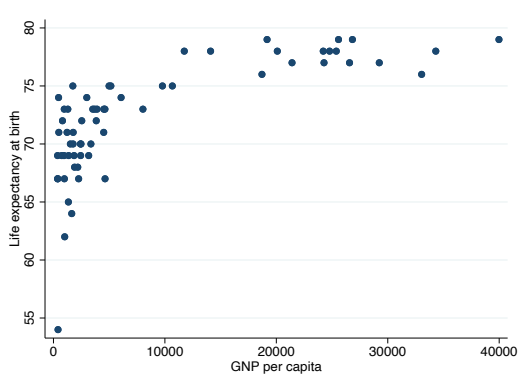
```
graph bar (count), over(region) ///  
    title("Bar chart of Regions") ///  
    graphregion(color(white)) //
```



Graphs: Scatter plot

Scatter plot demonstrates the relationship between two continuous variables.

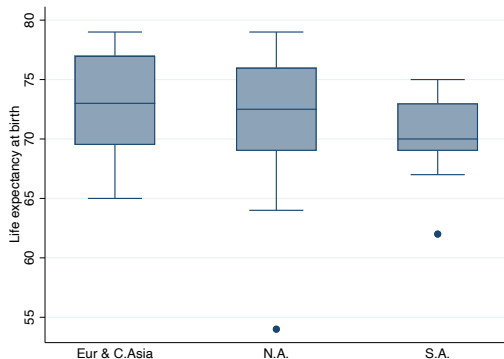
```
twoway scatter lexp gnppc, graphregion(color(white))
```



Graphs: Box plot

Box plot summarises the distribution of the data, with the 25th, 50th, and 75th percentile and 1.5 IQR.

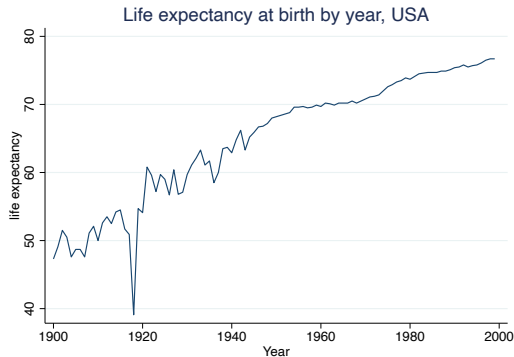
```
graph box lexp, over (region) ///  
graphregion(color(white))
```



Graphs: Line graph

Line graph functions similarly as scatter plots, with time as x-axis usually..

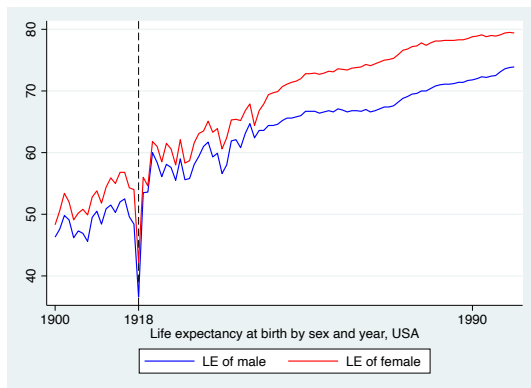
```
sysuse uslifeexp, clear  
twoway line le year, title("Life expectancy at birth by year, USA")  
graphregion(color(white)) }
```



Graphs: Stratification

Data is already in separate columns. Or using `by()`.

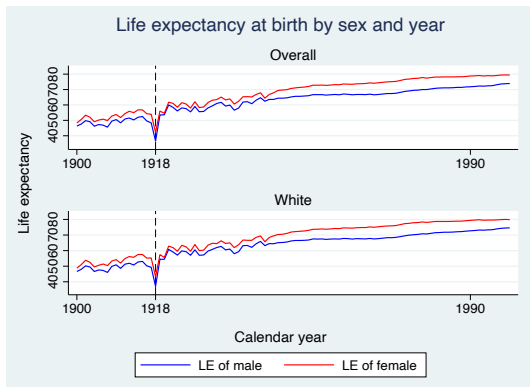
Hint: `by()` is often used in individual-level data.



Graphs: Putting graphs together

`grc1leg2` plays the role in plotting graphs together.

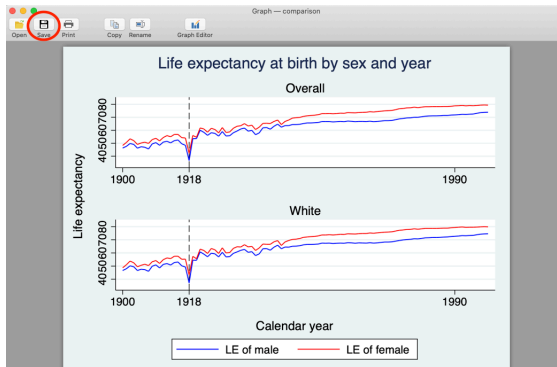
Hint: `grc1leg2` is not a default Stata command. See `help grc1leg2` to install it.



Graphs: Export

- A standard way:

```
graph export "location" /// assign the location  
, as(pdf) name("")
```
- An intuitive way:



References

1. David Goldsman PG. *A First Course in Probability and Statistics*. Georgia Institute of Technology, 2020.
2. Marcello Pagano KG. *Principles of Biostatistics*. Cengage Learning, Inc, 2000. ISBN 0534229026.
3. Erin Gabriel PF. Epidemiology PhD program, Karolinska Institutet, 2020.