# Analyzing Literacy Data

Griffith Feeney

# Literacy as an Example of Census Data

- Binary (simplest possible) response, literate or not, yet poses significant difficulties of *definition* and *interpretation*
- Importance of standard tabulation not always recognized
- Value comes only after computing *proportions literate* from tabulations

# Social and Economic Significance of Literacy

- Knowledge has become the key resource for economic development; literacy is the first step in gaining knowledge
- Literacy of *women* in particular is hypothesized to have important effects
- Literacy *stimulates imagination*
- The population census can provide literacy information for *every* population subgroup

# Minimal Tabulation of Literacy Data

- Literacy data should ***always*** be tabulated by **age** and **sex**; five year age groups will do, but the open-ended age interval should be *high*, typically at least 75+
- The sex dimension is important for gender differences, the age dimension for the study of historical change

# Numbers and Proportions

- Census reports provide *counts* of persons in various categories, and this is good census practice (*why?*)
- Counts numbers are only a *first step*, however, toward useful information
- In the case of literacy data, we want to know *proportions (or percentages) literate*

# The Importance of Age in Connection with Literacy

- Tremendous increases in literacy in developing countries in the past century
- Most people become literate at around age 10 or remain illiterate
- Very large differences in literacy by age group, perhaps as much as 10 *vs* 90 percent
- The *overall* level of literacy and its rate of change are therefore *very incomplete*

# Time-Plotting Literacy Data

- Identify a typical age ***l*** at which literacy is attained; exactitude not required
- Identify the proportion literate among persons aged ***x*** with time ***t - (x-l)***, where ***t*** is the reference time of the census
- Plot proportions literate against time

# Example: Vietnam 1989

- *Detailed Analysis of Sample Results* report shows percentages literate on page 51
- Age groups are 10-14 to 55-59, 60+ (***too low!***), census was October 1
- Time points are 1989.25 - (12.5 - 10) = 1986.75, 1989.25 - (17.5 - 10) = 1981.75, and so on; for *analysis plot* plot 60+ as though it were 60-64

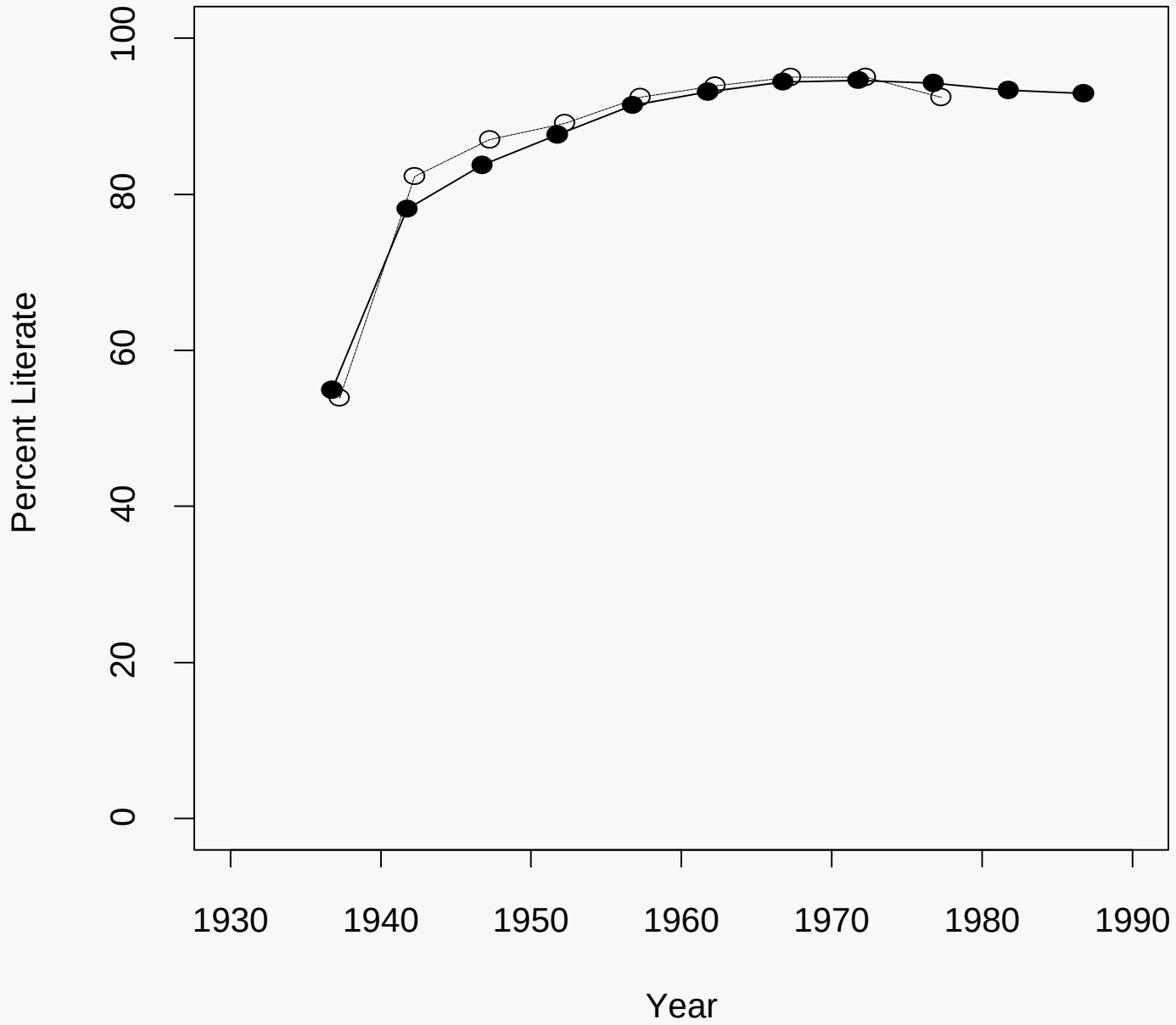
# ***Never fail to plot and look!***

(Statistician John W. Tukey)

- Plots that tell us what we already know ("security blankets") ***vs*** plots that teach us something we didn't know
- Plots we make to learn about the data ***vs*** plots for publication
- Spreadsheet programs remove any last excuse for not plotting; ***never fail to plot and look***

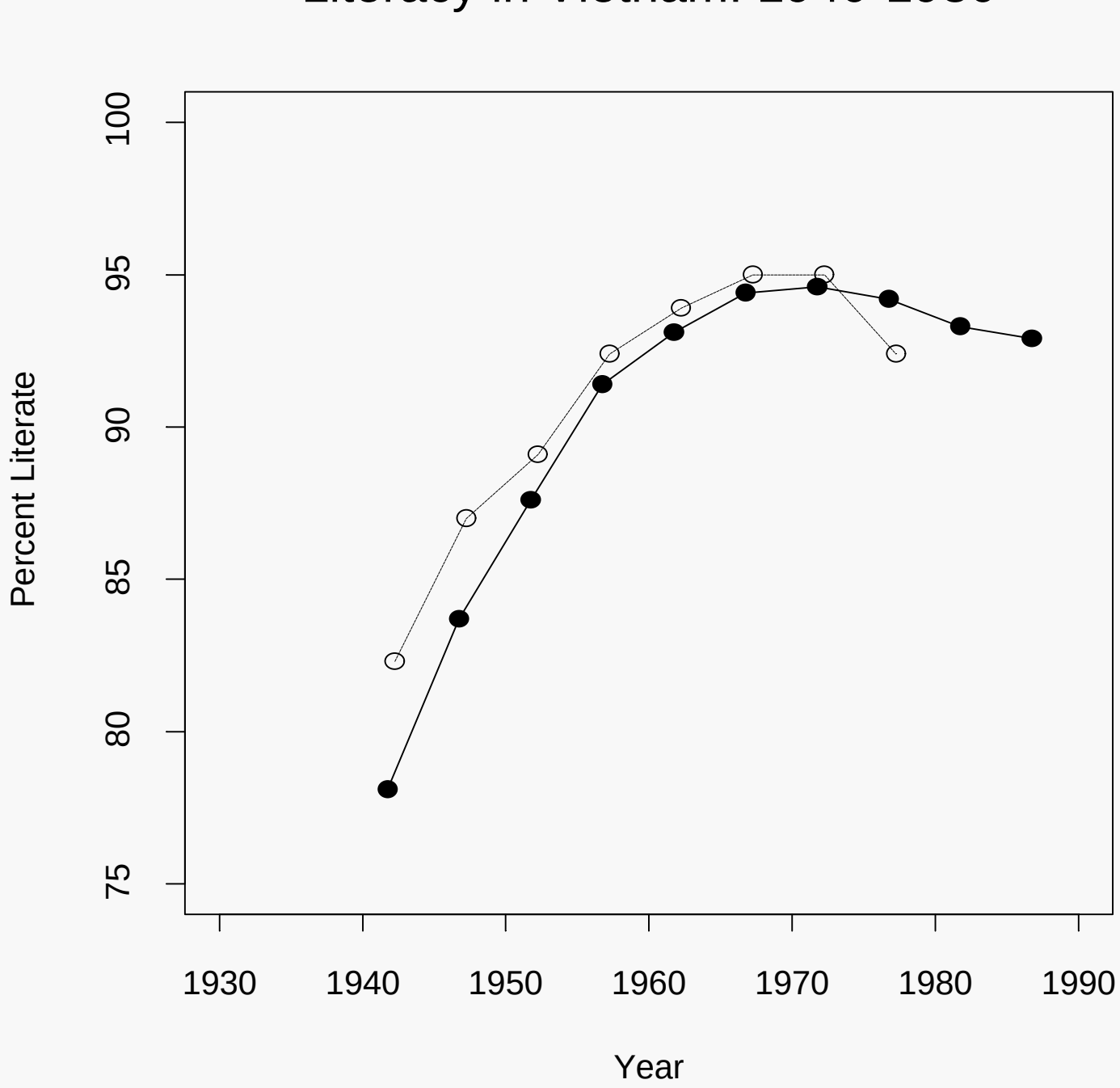
# Elementary Plotting Technique

- Simple but critical points of technique make the difference between good and bad plots
- Select a good **aspect ratio**; ‘banking’
- Select good **scales**; try several choices
- Select good **plotting symbols**; help the viewer see
- **Focus** on good plotting technique; don’t be satisfied with bad plots


Literacy in Vietnam: 1940-1980
Percent Literate
0
20
40
60
80
100
1930
1940
1950
1960
1970
1980
1990
Year

# What's Wrong with this Plot?

- Almost all the data is plotted in a small fraction of the plotting area, *i.e.*, most of the plotted area is *wasted*
- As a result, we can't see very well the differences we want to see, between the 1979 and 1989 census results
- Solution: ***change the vertical scale***; how?


Literacy in Vietnam: 1940-1980
Percent Literate
100
95
90
85
80
75
1930
1940
1950
1960
1970
1980
1990
Year

# Much Better, but ...

- How could the picture be improved?
- By larger and clearer plotting points
- Software limitations may keep us from getting what we want with reasonable effort

# What Have We Learned about *Literacy in Vietnam*?

- Overall level is very high, about 95 percent (if we believe the data)
- Has been high for many decades; already 80 percent when our series begins *circa* 1940
- We haven't learned anything about gender differences in literacy, but we can; all we have to do is look

# What Have We Learned about the Vietnamese Literacy *Data*?

- What's the difference? **REALITY = DATA + ERROR**
- The consistency between the 1979 and 1989 data suggests reasonably consistent interpretation of question and reasonably reliable reporting
- Which census showed higher literacy in birth cohorts? How to interpret this observation?

# What Have We Learned About *Census Tabulation Technique*?

- We often need to process census tabulations in some way to get useful information
- Setting the open-ended age group too low needlessly discards potentially useful information
- For both Vietnam censuses, an open-ended group of 85+ rather than 51+ or 65+ would have given valuable information on the early development of literacy

# What Have We Learned about *Plotting*?

- Never fail to plot and look!
- Use plots for consistency checks
- Pay attention to **aspect ratio**
- Pay attention to **scales**
- Pay attention to **plotting symbols**

# Future Changes in Literacy

- How will overall literacy change in future decades?
- What does this data tell us about *future* literacy?
- What do we *know* will happen in the future that will tend to increase overall literacy?
- What must be happen for this increase in literacy to occur; how far will it go?

# Review of Key Points

- The **importance of literacy data**; social, demographic, political
- **Tabulation of literacy data**; importance of open-ended age group
- **Analysis of literacy data**; time-trend analysis, consistency analysis
- The importance of **plotting** data
- ‘The future that has already happened’

Questions?
Comments?
Discussion?