Insightful Descriptives

Ian Watson

SPRC Training
UNSW
14 June 2011

Overview

- Insightful descriptives tell a story. Narrative structured around:
  - comparison sub-groups;
  - drilling down to reveal complexity;
  - dispelling first impressions.
- Take the reader on a journey—hence the importance of presentation.
- Some contrasts where boundaries blur:
  - Descriptive and inferential statistics
  - Analysis and presentation

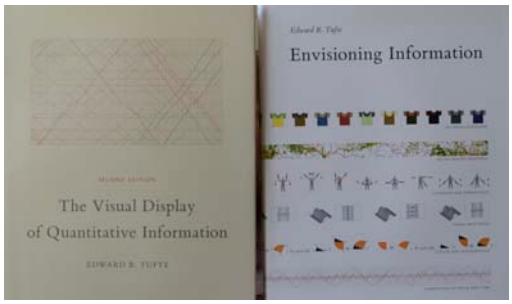
Descriptive and inferential statistics

- Less of a distinction than first appears
- Status issue around regression and assumptions of 'causality'
- Much modelling is descriptive—net 'effects' are still associations
- Modelling for confounding—issue of sample size
- Presenting results (such as predicted probabilities) requires same presentation strategies

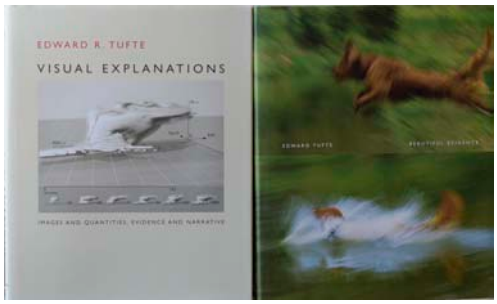
Presentation: exemplar of Edward Tufte

- Telling a quantitative story well
- Website: http://www.edwardtufte.com/

Edward Tufte's books



Edward Tufte's books



Tufte's Principles of Graphical Excellence

- Defined as: the well-designed presentation of interesting data—a matter of *substance*, of *statistics*, and of *design*.
- Consists of: complex ideas communicated with clarity, precision and efficiency.
- Gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.
- Nearly always multivariate.

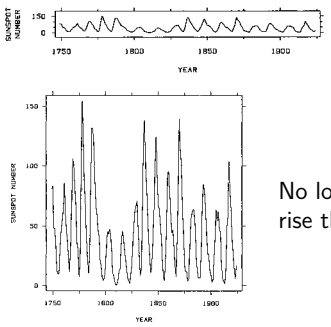
Analysis: exemplar of William Cleveland

- William Cleveland, **Visualizing Data**, Hobart Press, 1993
- Website: http://www.stat.purdue.edu/ wsc/

Cleveland's encoding and decoding

- Encoding information when making graph: position, length, slope, area, texture, colour
- Visual decoding of information when studying graph: graphical perception
- Knowledge from graphical methods combined with knowledge of visual perception
- Example from studies of vision: Weber's Law - visual detection of differences in line length depends on ratio, not overall size
- William Cleveland and Robert McGill, 'The Visual Decoding of Quantitative Information on Graphical Displays of Data', Journal of the Royal Statistical Society, Series A, Vol. 150, No.3, 1987, pp.192–229.

Cleveland's sunspot cycles



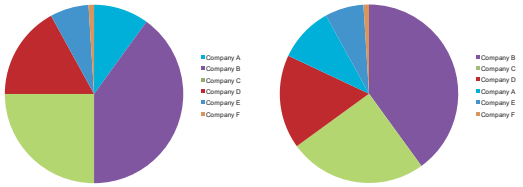
Can see that rises more rapid than falls

No longer see more rapid rise than fall

Ranking of 'elementary codes'

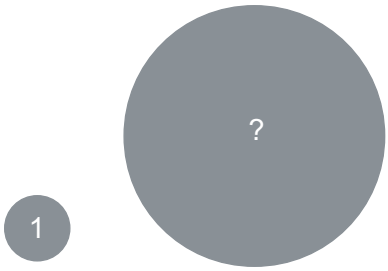
Rank	Code
1	Positions along a common scale
2	Positions along identical, nonaligned scales
3	Lengths
4	Angles
4–10	Slopes
6	Areas
7	Volumes

Stephen Few's pie chart examples

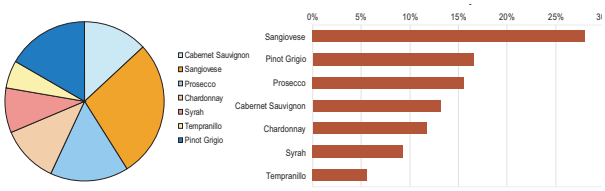


From Stephen Few, "Save the Pies for Dessert", **Perceptual Edge**, August 2007.

Stephen Few's pie chart examples



Stephen Few's pie chart examples



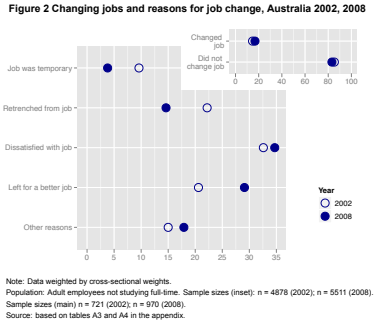
Graphs for analysis

- Graphs as a method for exploratory data analysis – dynamic and static
- "the direct goal is to see **patterns in the data** and understand the **overall behaviour**, but the more accurately the data are visually decoded, the better our chance to detect and properly understand the patterns and behaviour of the data" (Cleveland & McGill, 1987, p.198.)
- Value of density estimation and multivariate dot plots

Graphs for presentation

- Tufte scorns 'chart junk', common in business graphics
- Reasons for bad graphics:
  - Designers not researchers
  - Assumption statistics boring
  - Assumption audience ignorant
- ABS standards: dot plots, line charts and bar charts.
- Readership:
  - *Generalist*: graphs in chapters, tables in appendix
  - *Specialist*: graphs and key tables in chapter, detailed tables in appendix

Dot plots



Implications for tables

- Three key principles from Tufte relevant to tables:
  - present many numbers in a small space;
  - encourage the eye to compare different pieces of data;
  - make the process of decoding efficient for the reader.
- Contrast with the usual stats package output:
  - separate individual tables;
  - unnecessary additional information (DKs or the NO when only YES really relevant)
- Contrast with 'lazy tables':
  - missing bits of information which make the reader undertake tedious mental calculations (eg. no 100%)

Example table

- Shows population estimates and percentages
- Population estimates give readers a feel for the numbers involved

Example table

- Always show 100s, so instant awareness that dealing with column percentages

Example table

- Show sample sizes, so that cell counts can be calculated and reader can sense the precision of the estimates

Example table

- Show **notes**, **population** and **source** (unless obvious)
- Notes should explain decision rules, definitions and weighting
- Source should explain where data items came from (unless obvious)

Analysis: continuous data

Inspection of the variable

Annual household gross income (thousands)

	Percentiles	Smallest		
1%	1.4	0		
5%	16.402	0		
10%	26.016	0	Obs	17630
25%	49.55	0	Sum of Wgt.	17630
50%	88.289		Mean	100.8699
		Largest	Std. Dev.	84.96582
75%	130.264	994.416		
90%	180.559	994.416	Variance	7219.19
95%	223.838	994.416	Skewness	4.55088
99%	381.455	994.416	Kurtosis	42.73259

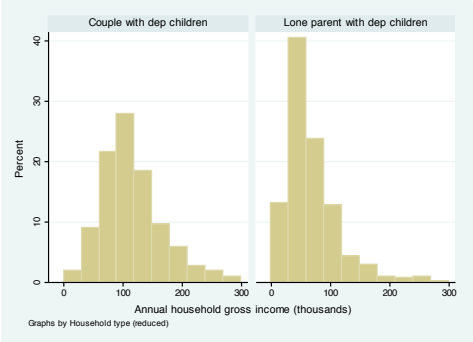
Household income: all households

Summary statistics
Comparing two household types:

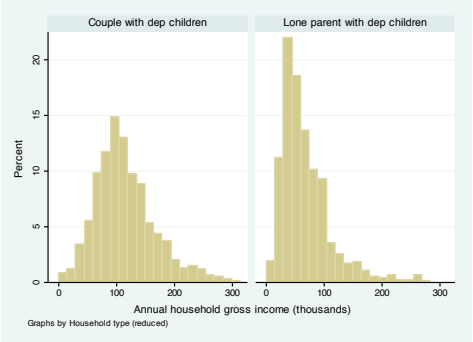
hhtype	mean	p50	iqr	s	k	N
Couple with dep children	127	110	65	5	45	7,748
Lone parent with dep children	71	56	47	7	100	1,643
Other	82	65	79	4	37	8,239
Total	101	88	81	5	43	17,630

p50 = median
iqr = inter-quartile range (diff 25th & 75th pciles)
sk = skewness (tail)
k = kurtosis (peakiness)
N = sample size

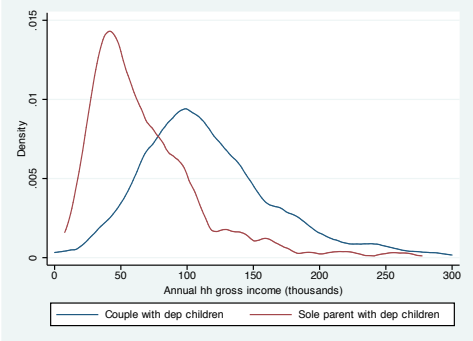
Histograms: hh income (all hhs)



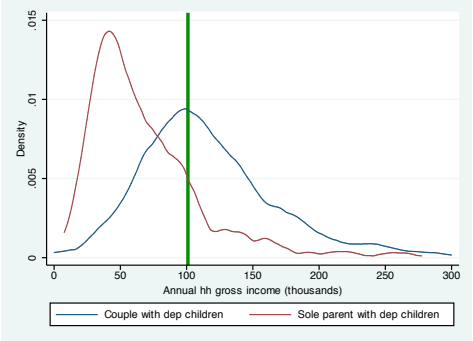
Histograms: hh income (all hhs)



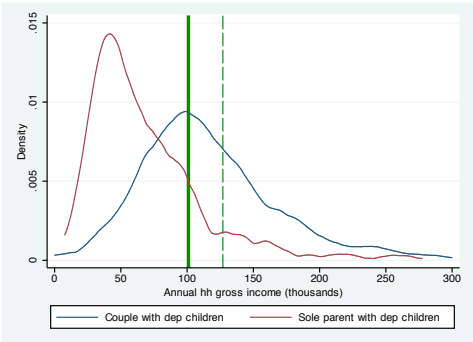
Density estimation: hh income (all hhs)



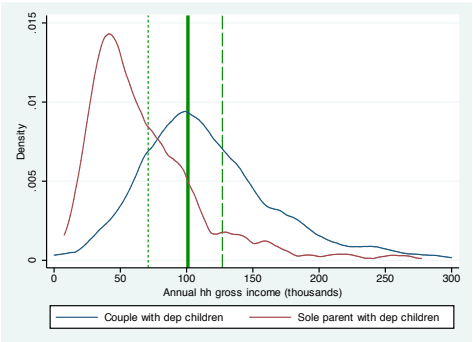
Density estimation: hh income (all hhs)



Density estimation: hh income (all hhs)



Density estimation: hh income (all hhs)



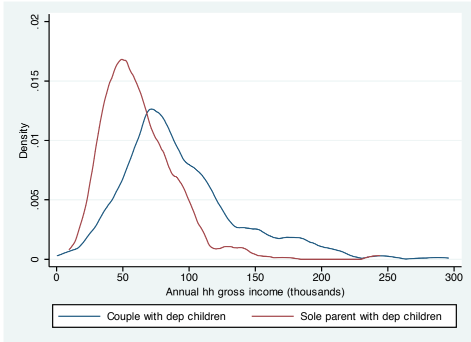
Household income: comparable households

Summary statistics
Comparing two household types:

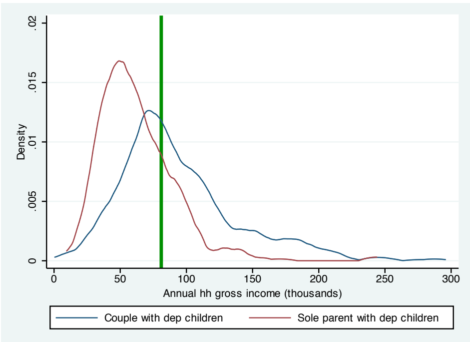
hhtype	mean	p50	iqr	sk	k	N
Couple dep ch	101	84	52	6	64	1,834
Lone parent dep ch	68	57	35	10	127	708
Other	69	55	47	7	85	2,472
Total	81	67	51	7	75	5,014

p50 = median
iqr = inter-quartile range (diff 25th & 75th pciles)
sk = skewness (tail)
k = kurtosis (peakiness)
N = sample size

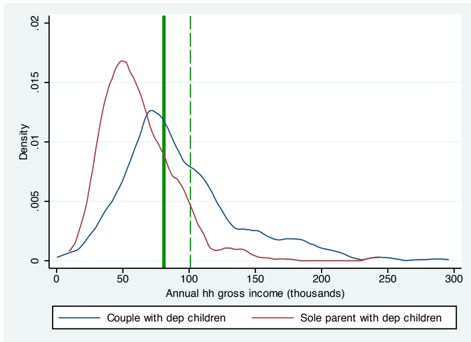
Density: hh income (comparable hhs)



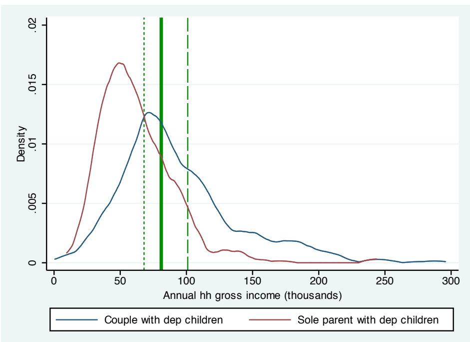
Density: hh income (comparable hhs)



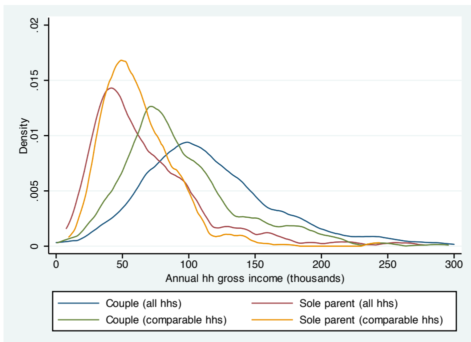
Density: hh income (comparable hhs)



Density: hh income (comparable hhs)



Density: final comparison



Summing up: continuous data

- Density estimation tells you:
  - where the differences occur
  - the magnitude of the differences
  - facilitates direct comparisons of sub-groups (bulge)
- Move sub-groups towards comparability but:
  - don't jump there immediately
  - take reader with you, as a narrative device
  - helps explain the *nature* of the differences
- Limitations of this approach:
  - sample size – hence regression techniques
  - compositional differences – apples and oranges problem – again regression
- Moral of the story: never be content with a single number

Acknowledging uncertainty

Unweighted data:

Household type (reduced)	Mean	CI
Couple with dep children	101	[98-105]
Lone parent with dep children	68	[62-73]
Other	69	[67-72]
Total	81	[79-83]

Weighted data

Household type (reduced)	Mean	CI
Couple with dep children	96	[91-101]
Lone parent with dep children	73	[61-86]
Other	73	[70-76]
Total	79	[77-82]

Survey data and weights

- Importance of weighting data for descriptives (debate when it comes to regressions)
- Yet, results should not differ that much with good survey data
- Where weights come from:
  - sampling design;
  - non-response.
- Examine the range of weights—should not be too large

Categorical data: row or column % ?

- Comparing sub-groups side-by-side: column percentages allow immediate comparison

	Couple	**Sole**	**Other**	**Total**
Age group	%	%	%	%
Under 25	52.7	66.5	15.0	36.4
25 to 34	10.8	8.6	15.0	12.5
35 to 44	19.1	11.4	9.0	13.7
45 to 54	14.4	10.3	15.0	14.3
55 to 64	2.4	2.4	20.1	10.7
65 and older	0.5	0.8	26.0	12.4
Total	100.0	100.0	100.0	100.0
N	7,748	1,643	8,239	17,630

Categorical data: row or column % ?

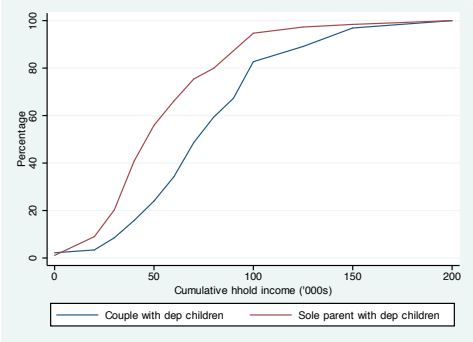
- Row percentages: require comparison by reference to the group average - more tedious to express

	Couple	**Sole**	**Other**	**Total**	**N**
Age group	%	%	%	%	
Under 25	63.7	17.0	19.2	100.0	6,414
25 to 34	37.8	6.4	55.7	100.0	2,210
35 to 44	61.6	7.8	30.7	100.0	2,407
45 to 54	44.4	6.7	48.9	100.0	2,523
55 to 64	10.0	2.1	87.9	100.0	1,888
65 and older	1.6	0.6	97.8	100.0	2,188
Total	43.9	9.3	46.7	100.0	17,630

Turning continuous into categorical data

Income grps	**Couple**		**Sole**		**Other**		**Total**	
('000s)	*%*	*Cum*	*%*	*Cum*	*%*	*Cum*	*%*	*Cum*
Under 20,000	2.2	2.2	1.1	1.1	6.0	6.0	4.6	4.6
20,000–29,999	1.2	3.4	7.8	9.0	7.0	13.0	5.5	10.0
30,000–39,999	5.1	8.5	11.2	20.2	12.4	25.4	10.3	20.3
40,000–49,999	7.3	15.8	20.6	40.8	13.5	38.9	12.4	32.8
50,000–59,999	8.2	24.0	15.1	55.9	12.5	51.4	11.6	44.3
60,000–69,999	10.2	34.2	10.2	66.1	11.5	62.9	11.0	55.4
70,000–79,999	14.4	48.6	9.3	75.4	7.3	70.3	9.4	64.8
80,000–89,999	10.7	59.3	4.5	79.9	7.2	77.5	7.9	72.7
90,000–99,999	7.9	67.3	7.4	87.3	5.7	83.2	6.5	79.2
100,000–124,999	15.5	82.7	7.4	94.7	6.5	89.8	9.0	88.3
125,000–149,999	6.4	89.1	2.6	97.3	4.5	94.2	4.8	93.1
150,000–199,999	7.8	96.9	1.2	98.4	2.2	96.5	3.7	96.7
200,000 and over	3.1	100.0	1.6	100.0	3.5	100.0	3.3	100.0
Total	100.0		100.0		100.0		100.0	

Cumulative data mirroring continous



Acknowledging uncertainty: full sample

	Couple	Sole	Other	Total
Emp full-time	47	26	43	44
	[46–49]	[23–30]	[42–45]	[42–45]
Emp part-time	27	28	14	20
	[26–29]	[25–32]	[13–15]	[19–21]
Unemp, look FT	2	4	3	3
	[2–3]	[2–6]	[2–3]	[2–3]
Unemp, look PT	1	3	1	1
	[1–2]	[2–5]	[0–1]	[1–1]
NILF, marg attach	7	15	4	6
	[6–8]	[12–18]	[4–5]	[5–6]
NILF, not marg	15	24	35	27
	[13–16]	[20–29]	[33–36]	[26–28]
Other	0	0	0	0
	[0–1]		[0–0]	[0–0]
Total	100	100	100	100
N	4,621	862	7,744	13,227

Acknowledging uncertainty: subset

	Couple	Sole	Other	Total
Emp full-time	34	15	67	47
	[31–37]	[11–21]	[65–70]	[45–49]
Emp part-time	32	31	16	24
	[29–34]	[26–37]	[14–18]	[23–26]
Unemp, look FT	3	5	5	4
	[2–5]	[3–8]	[4–6]	[4–5]
Unemp, look PT	3	5	1	2
	[2–4]	[3–8]	[1–2]	[2–3]
NILF, marg attach	11	20	5	9
	[9–13]	[15–25]	[4–6]	[8–10]
NILF, not marg	18	24	5	13
	[16–20]	[19–30]	[4–7]	[11–14]
Other	0	0	0	0
	[0–1]		[0–1]	[0–0]
Total	100	100	100	100
N	1,947	469	2,285	4,701

Overlapping CIs or SE difference?

- Rory Wolfe and James Hanley (2002), "If we're so different, why do we keep overlapping? When 1 plus 1 doesn't make 2 ", *Canadian Medical Association Journal*, Vol. 161, No.1, pp.65–66.
- "A frequently encountered misconception is that if 2 independent 95% CIs overlap each other ...then a statistical test of the difference will not be statistically significant at the 5% level.".
- In practice, overlapping CIs can be too conservative. Hence preference for the standard error of the difference: $1.96 * sqrt((SE_A{}^2 + SE_B{}^2))$

Cross-tabulations: CIs vs SE difference

	Couple	SE	Sole	SE	Other	SE	Total	SE
Emp full-time	33.8	(1.5)	15.4	(2.4)	67.5	(1.3)	47.2	(1.0)
Emp part-time	31.7	(1.3)	31.3	(2.7)	15.9	(0.9)	24.4	(0.8)
Unemp, look FT	3.3	(0.6)	4.8	(1.3)	5.1	(0.6)	4.3	(0.4)
Unemp, look PT	2.7	(0.5)	4.9	(1.2)	1.2	(0.3)	2.2	(0.3)
NILF, marg attach	10.7	(0.9)	19.7	(2.5)	4.8	(0.7)	9.0	(0.6)
NILF, not marg	17.6	(1.1)	23.9	(2.9)	5.4	(0.8)	12.7	(0.7)
Other	0.3	(0.2)	0.0	(0.0)	0.1	(0.1)	0.2	(0.1)
Total	100.0		100.0		100.0		100.0	

Overlapping CIs for NILF, not marginally attached:
15.5–19.9 (couple) and 18.7–29.9 (sole).
Conclusion: two groups are not statistically significantly different.

SE of the difference between 23.9 and 17.6 (diff of 6.3) is:
$1.96 * (sqrt(1.1^2 + 2.9^2)) = 6.1$
Diff of 6.3 is greater than SE of diff of 6.1.
Conclusion: two groups are statistically significantly different.

Statistical significance: a minefield

- Need to be wary about unduly emphasising statistical significance.
- Caveats commonly made:
  - substantive significance more important than statistical significance;
  - arbitrariness of 5% figure; Adrian Raftery (1995), 'Bayesian model selection in social research', *Sociological Methodology 1995*, pp.111–163, Blackwell Publishers;
  - Andrew Gelman and Hal Stern (2006) 'The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant', *The American Statistician*, Vol.60, No.4, pp.328–331.
- If your argument about sub-groups hinges on small differences, think again about your argument.
- BUT worth emphasising uncertainty, hence value of thinking in CIs.