Introduction to Descriptive Statistics

17.871

Spring 2015
Reasons for paying attention to data description

- Double-check data acquisition
- Data exploration
- Data explanation
# Key measures

Describing data

<table>
<thead>
<tr>
<th>Center</th>
<th>Moment</th>
<th>Non-moment based location parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Mode, median</td>
</tr>
<tr>
<td>Spread</td>
<td>Variance (standard deviation)</td>
<td>Range, Interquartile range</td>
</tr>
<tr>
<td>Skew</td>
<td>Skewness</td>
<td>--</td>
</tr>
<tr>
<td>Peaked</td>
<td>Kurtosis</td>
<td>--</td>
</tr>
</tbody>
</table>
## Key distinction
### Population vs. Sample Notation

<table>
<thead>
<tr>
<th>Population</th>
<th>vs.</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greeks</td>
<td></td>
<td>Romans</td>
</tr>
<tr>
<td>$\mu, \sigma, \beta$</td>
<td></td>
<td>s, b</td>
</tr>
</tbody>
</table>
Mean

$$\sum_{i=1}^{n} x_i \equiv \mu \equiv \bar{X}$$
Guess the Mean

Source: CCES
Guess the Mean

Source: CCES
Guess the Mean

Source: CCES
Guess the Mean
Guess the Mean

3.3
Variance, Standard Deviation of a Population

\[
\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n} \equiv \sigma^2, \\
\sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}} \equiv \sigma
\]
Variance, S.D. of a Sample

\[
\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n-1} \equiv s^2,
\]

Degrees of freedom

\[
\sqrt{\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n-1}} \equiv s
\]
What was the mean and standard deviation of the age of the MIT undergraduate population on Registration Day, Fall 2014?
What was the mean and standard deviation of the MIT undergraduate population on Registration Day, Fall 2014?

My guess:

Mean probably ~ 19.5 (if everyone is 18, 19, 20, or 21, and they are evenly distributed.

s.d. probably ~ 1
Guess the Standard Deviation

Source: CCES
Guess the Standard Deviation

\[ \sigma = 0.89 \]
Guess the Standard Deviation

3.3
Guess the Standard Deviation

3.3  \quad \sigma = 7.2
Binary data

\[ \bar{X} = \text{prob}(X) = 1 = \text{proportion of time } x = 1 \]

\[ s_x^2 = \bar{x}(1 - \bar{x}) \implies s_x = \sqrt{\bar{x}(1 - \bar{x})} \]
Example of this, using the most recent Gallup approval rating of Pres. Obama

- `gen o_approve = 1 if gallup=="Approve"
- `replace o_approve = 0 if gallup=="Disapprove"
- `the command `summ
- `o_approve` produces
- Mean = 0.51
- Var = 0.51(1-0.51)=.2499
- S.d. = .49989999
Therefore, reporting the standard deviation (or variance) of a binary variable is redundant information. **Don’t do it** for papers written for 17.871.
Non-moment base measures of center or spread

• Central tendency
  – Mode
  – Median

• Spread
  – Range
  – Interquartile range
Mode

• The most common value
Guess the Mode

Source: CCES

2.8
Guess the Mode

3.3
Guess the Mode

Number of years the respondent has lived in his/her current home
**Guess the Mode**

<table>
<thead>
<tr>
<th>pew religion</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>protestant</td>
<td>26,241</td>
<td>47.40</td>
<td>47.40</td>
</tr>
<tr>
<td>roman catholic</td>
<td>12,348</td>
<td>22.30</td>
<td>69.70</td>
</tr>
<tr>
<td>mormon</td>
<td>931</td>
<td>1.68</td>
<td>71.38</td>
</tr>
<tr>
<td>eastern or greek orthodox</td>
<td>275</td>
<td>0.50</td>
<td>71.88</td>
</tr>
<tr>
<td>jewish</td>
<td>1,678</td>
<td>3.03</td>
<td>74.91</td>
</tr>
<tr>
<td>muslim</td>
<td>164</td>
<td>0.30</td>
<td>75.21</td>
</tr>
<tr>
<td>buddhist</td>
<td>445</td>
<td>0.80</td>
<td>76.01</td>
</tr>
<tr>
<td>hindu</td>
<td>89</td>
<td>0.16</td>
<td>76.17</td>
</tr>
<tr>
<td>agnostic</td>
<td>2,885</td>
<td>5.21</td>
<td>81.38</td>
</tr>
<tr>
<td>nothing in particular</td>
<td>7,641</td>
<td>13.80</td>
<td>95.18</td>
</tr>
<tr>
<td>something else</td>
<td>2,667</td>
<td>4.82</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Total | 55,364 | 100.00 |
The mode is rarely an informative statistic about the central tendency of the data. It’s most useful in describing the “typical” observation of a categorical variable.
Median

- The numerical value separating the upper half of a distribution from the lower half of the distribution
  - If N is odd, there is a unique median
  - If N is even, there is no unique median --- the convention is to average the two middle values
Guess the Median

Source: CCES
Guess the Median

[Diagram of a bar chart showing distribution with institution approval on the x-axis and density on the y-axis.]

Source: CCES
Guess the Median

3.3
Guess the Median

Number of medals

0 3.3
Guess the Median

Number of years the respondent has lived in his/her current home

Mode = 0
Mean = 11.8

Number of years the respondent has lived in his/her current home
Guess the Median

Number of years the respondent has lived in his/her current home

Mode = 0
Mean = 11.8
Median = 8

Number of years the respondent has lived in his/her current home
Guess the Median

Mode = 0
Mean = 11.8
Median = 8

Note with right-skewed data:
Mode < median < mean

Number of years the respondent has lived in his/her current home
Median frequently preferred for income data

Table 1.1 All Returns: Selected Income and Tax Items, by Size and Accumulated Size of Adjusted Gross Income, Tax Year 2012
(All figures are estimates based on samples—money amounts are in thousands of dollars except as indicated)

<table>
<thead>
<tr>
<th>Size and accumulated size of adjusted gross income</th>
<th>Number of returns</th>
<th>Percent of total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td><strong>Size of adjusted gross income</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All returns</td>
<td>144,928,472</td>
<td>100.0</td>
</tr>
<tr>
<td>No adjusted gross income</td>
<td>2,128,548</td>
<td>1.5</td>
</tr>
<tr>
<td>$1 under $5,000</td>
<td>10,378,183</td>
<td>7.2</td>
</tr>
<tr>
<td>$5,000 under $10,000</td>
<td>11,958,135</td>
<td>8.3</td>
</tr>
<tr>
<td>$10,000 under $15,000</td>
<td>12,032,192</td>
<td>8.7</td>
</tr>
<tr>
<td>$15,000 under $20,000</td>
<td>11,615,575</td>
<td>8.0</td>
</tr>
<tr>
<td>$20,000 under $25,000</td>
<td>10,168,831</td>
<td>7.0</td>
</tr>
<tr>
<td>$25,000 under $30,000</td>
<td>8,734,480</td>
<td>6.0</td>
</tr>
<tr>
<td>$30,000 under $40,000</td>
<td>14,451,152</td>
<td>10.0</td>
</tr>
<tr>
<td>$40,000 under $50,000</td>
<td>10,873,672</td>
<td>7.5</td>
</tr>
<tr>
<td>$50,000 under $75,000</td>
<td>18,865,371</td>
<td>13.1</td>
</tr>
<tr>
<td>$75,000 under $100,000</td>
<td>12,103,591</td>
<td>8.4</td>
</tr>
<tr>
<td>$100,000 under $200,000</td>
<td>15,646,848</td>
<td>10.8</td>
</tr>
<tr>
<td>$200,000 under $500,000</td>
<td>4,154,113</td>
<td>2.9</td>
</tr>
<tr>
<td>$500,000 under $1,000,000</td>
<td>705,029</td>
<td>0.5</td>
</tr>
<tr>
<td>$1,000,000 under $1,500,000</td>
<td>169,413</td>
<td>0.1</td>
</tr>
<tr>
<td>$1,500,000 under $2,000,000</td>
<td>71,874</td>
<td>[2]</td>
</tr>
<tr>
<td>$2,000,000 under $5,000,000</td>
<td>106,711</td>
<td>0.1</td>
</tr>
<tr>
<td>$5,000,000 under $10,000,000</td>
<td>27,167</td>
<td>[2]</td>
</tr>
<tr>
<td>$10,000,000 or more</td>
<td>17,585</td>
<td>[2]</td>
</tr>
</tbody>
</table>
The (uninformative) graph

Mean = 68,735
Median = 35,000
Mode = 0 (probably)
Spread

• Range
  – Max(x) – Min(x)

• Interquartile range (IQR)
  – Q₃(x) – Q₁(x)

\[Q₁ = \text{CDF}^{-1}(0.25)\]
\[Q₃ = \text{CDF}^{-1}(0.75)\]
Guess the IQR

\[ \sigma = 0.89 \]

Source: CCES
Guess the IQR

Source: CCES

$\sigma = 0.89$

$IQR = 2$
Guess the IQR

Mode = 0
Mean = 11.8
Median = 8

σ = 11.7

Number of years the respondent has lived in his/her current home
Number of years the respondent has lived in his/her current home

Mode = 0
Mean = 11.8
Median = 8
σ = 11.7
IQR = 14 (17-3)
Don’t guess the IQR

Mean = 68,735
Median = 35,000
Mode = 0 (probably)

\[ \sigma = 371,799 \]

IQR = 50,000 (62,500-12,500)
Lopsidedness and peakedness
Normal distribution example

- IQ
- SAT
- Height
- Symmetrical
- Mean = median = mode

\[ f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-\mu)/2\sigma^2} \]
Skewness
Asymmetrical distribution

- Income
- Contribution to candidates
- Populations of countries
- Age of MIT undergraduates
- “Positive skew”
- “Right skew”
Distribution of the average $$ of dividends/tax return (in K’s)

Hyde County, SD

Fuel economy of cars for sale in the US

Mitsubishi i-MiEV (which is supposed to be all electric)
Skewness
Asymmetrical distribution

- GPA of MIT students
- Age of MIT faculty
- “Negative skew”
- “Left skew”
Placement of Republican Party on 100-point scale
Skewness

Frequency

Value

Mode
Median
Mean
Guess the sign of the skew

Source: CCES
Guess the sign of the skew

\[ \gamma = 0.89 \]

Source: CCES
Guess the sign of the skew

Number of years the respondent has lived in his/her current home

How long lived in current residence - Years
Guess the sign of the skew

Number of years the respondent has lived in his/her current home

\[ \gamma = 1.5 \]
Note: It is really rare to find a naturally occurring variable with a negative skew

Mean = 68.3
s.d. = 8.7
Skew: -0.80
Kurtosis

Kurtosis measures the "peakedness" or "tailedness" of the distribution. The value of the kurtosis depends on the value of k:

- $k > 3$: leptokurtic (peaked distribution)
- $k = 3$: mesokurtic (normal distribution)
- $k < 3$: platykurtic (flat distribution)
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-placement</td>
<td>4.5</td>
<td>1.9</td>
<td>-0.28</td>
<td>1.9</td>
</tr>
<tr>
<td>Dem. pty</td>
<td>2.2</td>
<td>1.4</td>
<td>1.1</td>
<td>3.9</td>
</tr>
<tr>
<td>Rep. pty</td>
<td>5.6</td>
<td>1.4</td>
<td>-0.98</td>
<td>3.7</td>
</tr>
<tr>
<td>Tea party</td>
<td>6.1</td>
<td>1.3</td>
<td>-2.1</td>
<td>7.5</td>
</tr>
</tbody>
</table>

Source: CCES, 2012
Normal distribution

- Skewness = 0
- Kurtosis = 3

\[ f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]
Commands in STATA for univariate statistics (K&K pp. 176-186)

• `summarize varlist`
• `summarize varlist, detail`
• `histogram varname, bin() start() width() density/fraction/frequency normal discrete`
• `table varname, contents(clist)`
• `tabstat varlist, statistics(statname…)`

• `tabulate`
### summ time_1

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>time_1</td>
<td>10153</td>
<td>11.78371</td>
<td>11.70837</td>
<td>0</td>
<td>89</td>
</tr>
</tbody>
</table>

### summ time_1, det

How long lived in current residence - Years

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>Smallest</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>0</td>
</tr>
<tr>
<td>5%</td>
<td>0</td>
</tr>
<tr>
<td>10%</td>
<td>1</td>
</tr>
<tr>
<td>25%</td>
<td>3</td>
</tr>
<tr>
<td>50%</td>
<td>8</td>
</tr>
<tr>
<td>75%</td>
<td>17</td>
</tr>
<tr>
<td>90%</td>
<td>29</td>
</tr>
<tr>
<td>95%</td>
<td>36</td>
</tr>
<tr>
<td>99%</td>
<td>50</td>
</tr>
</tbody>
</table>

- Obs: 10153
- Sum of Wgt.: 10153
- Mean: 11.78371
- Std. Dev.: 11.70837
- Variance: 137.086
- Skewness: 1.470977
- Kurtosis: 5.21861

Data from 2012 Survey of the Performance of American Elections
<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>time_1</td>
<td>10153</td>
<td>11.7837</td>
<td>11.70837</td>
<td>0</td>
<td>89</td>
</tr>
</tbody>
</table>

Data from 2012 Survey of the Performance of American Elections
. hist time_1, discrete
  (start=0, width=1)
. table pid3

<p>| 3 point     |      |</p>
<table>
<thead>
<tr>
<th>party ID</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrat</td>
<td>3,808</td>
</tr>
<tr>
<td>Republican</td>
<td>3,036</td>
</tr>
<tr>
<td>Independent</td>
<td>2,825</td>
</tr>
<tr>
<td>Other</td>
<td>234</td>
</tr>
<tr>
<td>Not sure</td>
<td>297</td>
</tr>
<tr>
<td>-------------</td>
<td>----------</td>
</tr>
</tbody>
</table>
. tabstat time_1 age

    stats |     time_1       age
----------+-------------------
     mean |  11.78371  49.33363
----------

. tabstat time_1 age,stats(mean sd skew kurt)

    stats |     time_1       age
----------+-------------------
     mean |  11.78371  49.33363
     sd  |  11.70837  15.89716
   skewness |  1.470977  -.0152461
  kurtosis |  5.21861   2.177523
----------
. tabstat time_1 age,by(pid3) s(mean sd)

Summary statistics: mean, sd
by categories of: pid3 (3 point party ID)

<table>
<thead>
<tr>
<th>pid3</th>
<th>time_1</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrat</td>
<td>11.28602</td>
<td>47.72348</td>
</tr>
<tr>
<td></td>
<td>11.7268</td>
<td>15.88458</td>
</tr>
<tr>
<td>Republican</td>
<td>13.1379</td>
<td>52.27569</td>
</tr>
<tr>
<td></td>
<td>12.17941</td>
<td>15.69504</td>
</tr>
<tr>
<td>Independent</td>
<td>11.66335</td>
<td>50.07646</td>
</tr>
<tr>
<td></td>
<td>11.35228</td>
<td>15.51778</td>
</tr>
<tr>
<td>Other</td>
<td>8.457265</td>
<td>42.52991</td>
</tr>
<tr>
<td></td>
<td>9.328546</td>
<td>13.84282</td>
</tr>
<tr>
<td>Not sure</td>
<td>8.084459</td>
<td>38.19865</td>
</tr>
<tr>
<td></td>
<td>9.559606</td>
<td>14.14754</td>
</tr>
<tr>
<td>Total</td>
<td>11.78371</td>
<td>49.33363</td>
</tr>
<tr>
<td></td>
<td>11.70837</td>
<td>15.89716</td>
</tr>
</tbody>
</table>
```
.table pid3,c(mean time_1 sd time_1)

+--------------------------------------------------+
<table>
<thead>
<tr>
<th>3 point</th>
<th>mean(time_1)</th>
<th>sd(time_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrat</td>
<td>11.286</td>
<td>11.7268</td>
</tr>
<tr>
<td>Republican</td>
<td>13.1379</td>
<td>12.17941</td>
</tr>
<tr>
<td>Independent</td>
<td>11.6634</td>
<td>11.35228</td>
</tr>
<tr>
<td>Other</td>
<td>8.45726</td>
<td>9.328546</td>
</tr>
<tr>
<td>Not sure</td>
<td>8.08446</td>
<td>9.559606</td>
</tr>
</tbody>
</table>
+--------------------------------------------------+
```
Univariate graphs
Commands in STATA for univariate statistics

- `histogram varname, bin() start() width() density/fraction/frequency normal`
- `graph box varnames`
- `graph dot varnames`
Example of Florida voters

• Question: does the age of voters vary by race?

• Combine Florida voter extract files, 2010

• \texttt{gen new\_birth\_date= date(birth\_date, "MDY")}
• \texttt{gen birth\_year= year(new\_b)}
• \texttt{gen age= 2010-birth\_year}
Look at distribution of birth year
Explore age by race

```
. table race if birth_year>1900,c(mean age)

<table>
<thead>
<tr>
<th>race</th>
<th>mean(age)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>45.61229</td>
</tr>
<tr>
<td>2</td>
<td>42.89916</td>
</tr>
<tr>
<td>3</td>
<td>42.6952</td>
</tr>
<tr>
<td>4</td>
<td>45.09718</td>
</tr>
<tr>
<td>5</td>
<td>52.08628</td>
</tr>
<tr>
<td>6</td>
<td>44.77392</td>
</tr>
<tr>
<td>9</td>
<td>40.86704</td>
</tr>
</tbody>
</table>

3 = Black
4 = Hispanic
5 = White
```
. hist age if birth_year>1900
(bin=71, start=9, width=1.3802817)
Graph birth year

`. hist age if birth_year>1900
(bin=71, start=9, width=1.3802817)`
Divide into “bins” so that each bar represents 1 year

\[ \text{hist age if birth_year>1900, width(1)} \]
Add ticks at 10-year intervals

. hist age if birth_year>1900, width(1) xlabel(20 (10) 100)
Superimpose the normal curve
(with the same mean and s.d. as the empirical distribution)

```plaintext
hist age if birth_year>1900, wid(1) xlabel(20 (10) 100)
normal
```
. summ age if birth_year>1900, det

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>Smallest</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>18</td>
</tr>
<tr>
<td>5%</td>
<td>21</td>
</tr>
<tr>
<td>10%</td>
<td>24</td>
</tr>
<tr>
<td>25%</td>
<td>34</td>
</tr>
<tr>
<td>50%</td>
<td>48</td>
</tr>
<tr>
<td>75%</td>
<td>63</td>
</tr>
<tr>
<td>90%</td>
<td>77</td>
</tr>
<tr>
<td>95%</td>
<td>83</td>
</tr>
<tr>
<td>99%</td>
<td>91</td>
</tr>
</tbody>
</table>

Obs: 12612114
Sum of Wgt: 12612114

Mean: 49.47549
Std. Dev: 19.01049
Variance: 361.3986
Skewness: 0.2629496
Kurtosis: 2.222442
Histograms by race

hist age if birth_year>1900 & race>=3 & race<=5, wid(1)
xlabel(20 (10) 100) normal by(race)

3 = Black
4 = Hispanic
5 = White
Draw the previous graph with a box plot

graph box age if birth_year>1900

Upper quartile
Median
Lower quartile

1.5 x IQR

Inter-quartile range
Draw the box plots for the different races

graph box age if birth_year>1900 & race>=3 & race<=5, by(race)

3 = Black
4 = Hispanic
5 = White
Draw the box plots for the different races using “over” option

```plaintext
graph box age if birth_year>1900 & race>=3 & race<=5, over(race)
```

3 = Black
4 = Hispanic
5 = White
Main issues with histograms

• Proper level of aggregation
• Non-regular data categories
Stop and think:
What should the distribution of length-of-current residency look like?
(Hint: the median is around 4 years)
A note about histograms with unnatural categories

From the Current Population Survey (2000), Voter and Registration Survey

How long (have you/has name) lived at this address?

-9  No Response  
-3  Refused  
-2  Don't know  
-1  Not in universe  
1  Less than 1 month  
2  1-6 months  
3  7-11 months  
4  1-2 years  
5  3-4 years  
6  5 years or longer
Solution, Step 1
Map artificial category onto “natural” midpoint

recode live_length (min/-1 = .) (1=.042) (2=.29) (3=.75) (4=1.5) ///
(5=3.5) (6=10)

-9 No Response → missing
-3 Refused → missing
-2 Don’t know → missing
-1 Not in universe → missing
1 Less than 1 month → 1/24 = 0.042
2 1-6 months → 3.5/12 = 0.29
3 7-11 months → 9/12 = 0.75
4 1-2 years → 1.5
5 3-4 years → 3.5
6 5 years or longer → 10 (arbitrary)
Graph of recoded data

histogram longevity, fraction
Graph of recoded data

Why doesn’t…

...look like …
Density plot of data

Total area of last bar = .557
Width of bar = 11 (arbitrary)
Solve for: \( a = w \cdot h \) (or)
\[ .557 = 11h \Rightarrow h = .051 \]
## Density plot template

<table>
<thead>
<tr>
<th>Category</th>
<th>Fraction</th>
<th>X-min</th>
<th>X-max</th>
<th>X-length</th>
<th>Height (density)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 1 mo.</td>
<td>.0156</td>
<td>0</td>
<td>1/12</td>
<td>.082</td>
<td>.19*</td>
</tr>
<tr>
<td>1-6 mo.</td>
<td>.0909</td>
<td>1/12</td>
<td>½</td>
<td>.417</td>
<td>.22</td>
</tr>
<tr>
<td>7-11 mo.</td>
<td>.0430</td>
<td>½</td>
<td>1</td>
<td>.500</td>
<td>.09</td>
</tr>
<tr>
<td>1-2 yr.</td>
<td>.1529</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>.15</td>
</tr>
<tr>
<td>3-4 yr.</td>
<td>.1404</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>.07</td>
</tr>
<tr>
<td>5+ yr.</td>
<td>.5571</td>
<td>4</td>
<td>15</td>
<td>11</td>
<td>.05</td>
</tr>
</tbody>
</table>

* = \(\frac{.0156}{.082}\)
Three words about pie charts: don’t use them
So, what’s wrong with them

• For non-time series data, hard to get a comparison among groups; the eye is very bad in judging relative size of circle slices
• For time series, data, hard to grasp cross-time comparisons
Some words about graphical presentation

- Aspects of graphical integrity (following Edward Tufte, *Visual Display of Quantitative Information*)
  - Main point should be readily apparent
  - Show as much data as possible
  - Write clear labels on the graph
  - Show data variation, not design variation
Some bad graphs
Some good graphs
Download and use the “Tufte” scheme

• ssc install scheme_tufte
scatter demrespct2000 demrespct1964 [aw=total], xsize(6.5) ysize(6.5) xscale(range(.2 .8)) yscale(range(.2 .8))
scatter demprespct2000 demprespct1964 [aw=total], xsize(6.5)
ysize(6.5) xscale(range(.2 .8)) yscale(range(.2 .8)) scheme(tufte)
There is a difference between graphs in research and publication.

Graphs by race:
- Density
- Normal age

![Graphs](image-url)