Visualizing and summarizing data

Ken Rice, Dept of Biostatistics

HUBIO 530

January 2015

Q. What's your talk about?

Today I will describe:

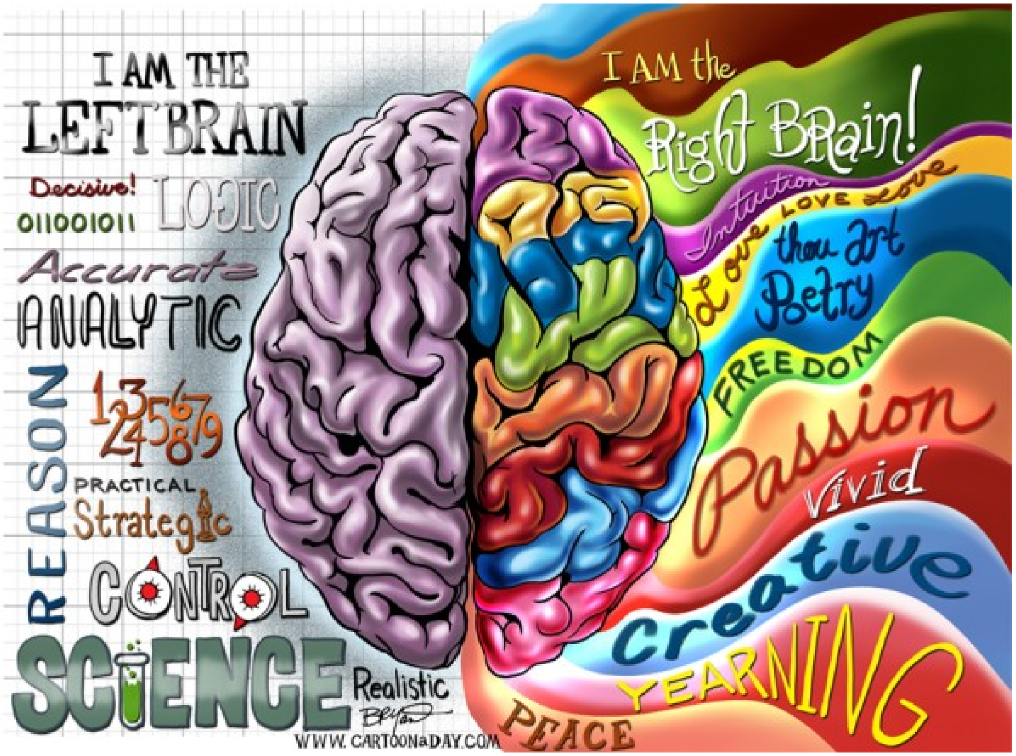
- How to visualize small datasets
- How to summarize small datasets
- Some methods for larger datasets

The 'summary' ideas are *usually* introduced through formulae alone – i.e. mathematical equations. Instead, I will use *only* pictures to describe/explain what's going on.

For those who want/need it ('keen people') the math is given in supplementary slides.

Q. Aren't you meant to just do math?

Very roughly, today's approach is 'right brain':



Cognition research suggests that in humans, thinking is;

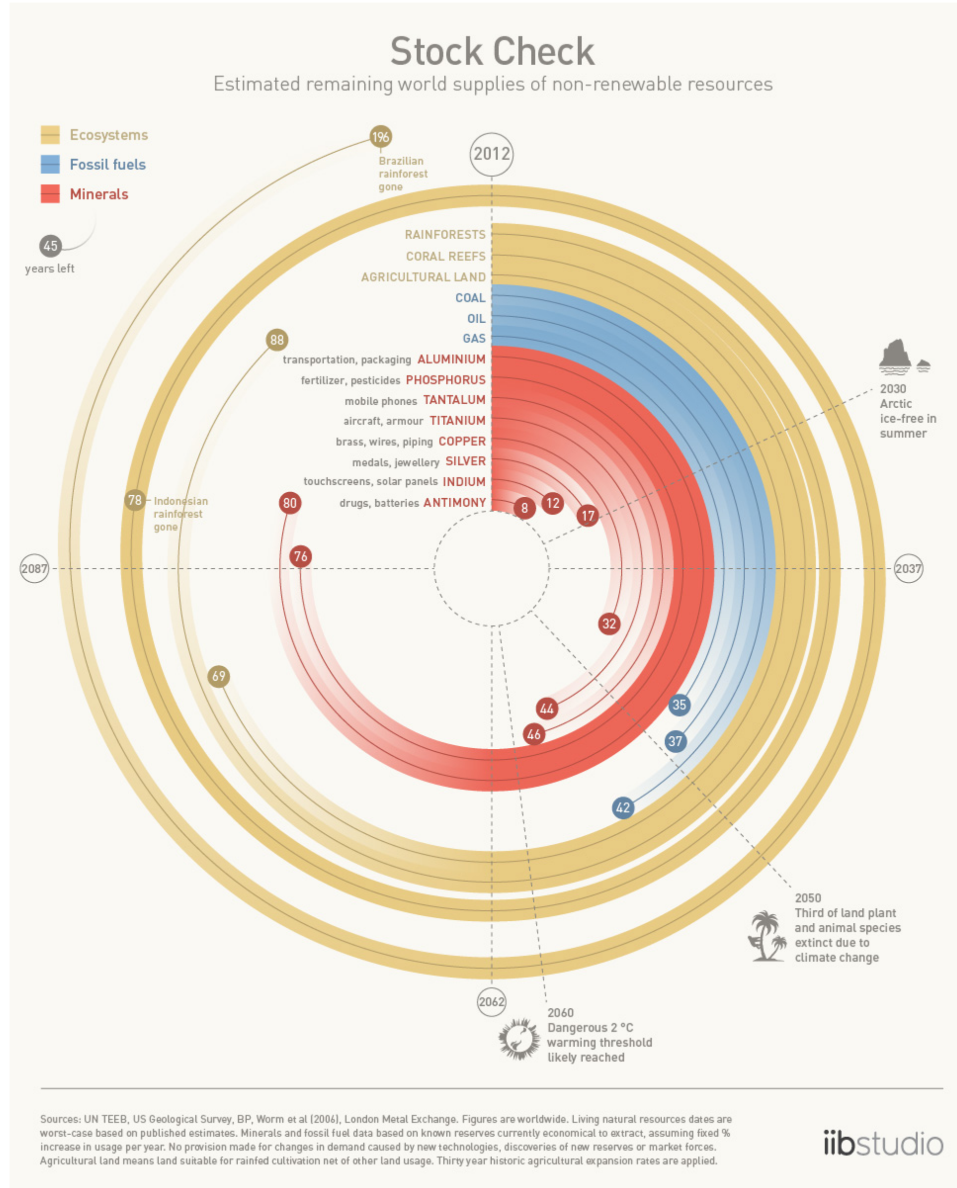
Exclusively in words (25%)	Both visual/spatial & in words (45%)	Strongly visual/ spatial (30%)

Even if you're in the 25% – or already know the ideas – today's explanations may help you communicate with the other 75%.

Visualization: some data

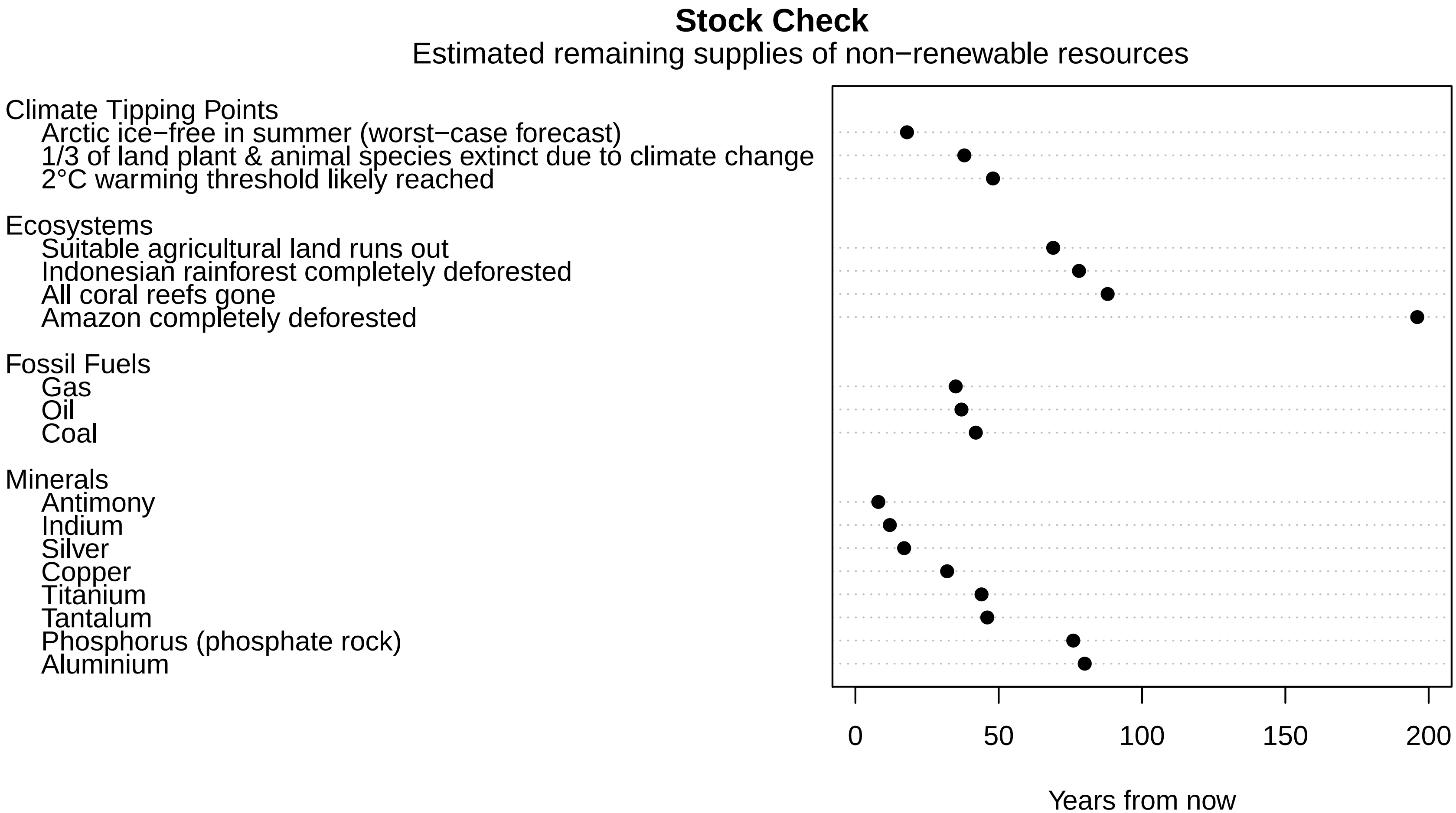
'BBC Future', trying to impress/amaze you with 18 numbers;

Q. What's the message?



Visualization: some data

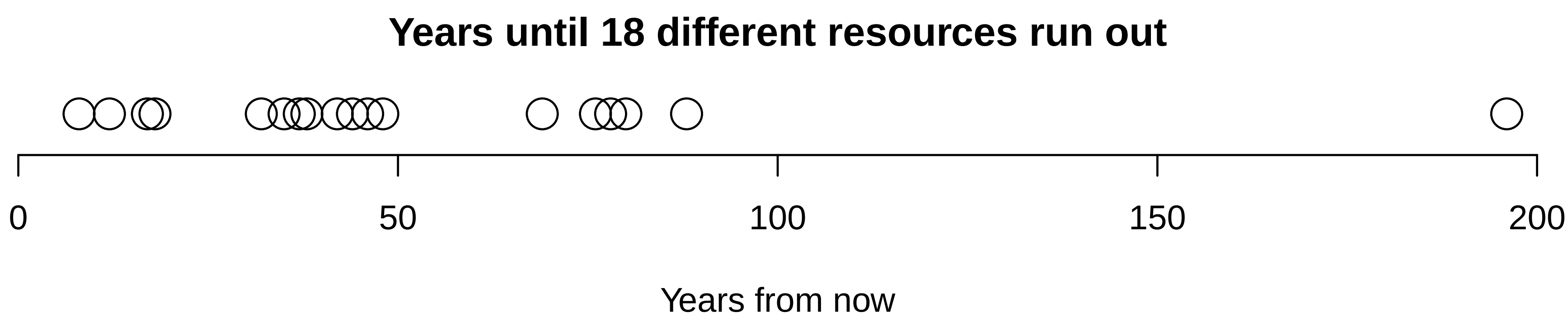
The statistician-approved version – does it impress? amaze?



'Position on a common scale' is known to be the best mode of presentation, for making visual comparisons.

Visualization: some data

Typically, one ‘common scale’ will do;

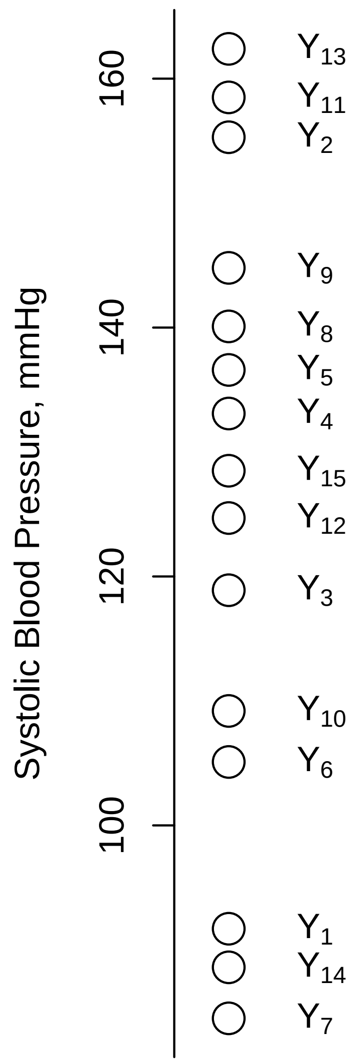


- Known as a *dotplot*, *dotchart* or *stripchart*
- It is a natural choice for displaying multiple blood pressures, or GFRs, BMIs, TLAs, eTLAs, times-to-event, etc etc
- Open circles work better than closed* – consider overlap
- Can be impractical with 100s of data points – we’ll see alternatives later.

* *This has been known since research in Bell Labs in the 70s... though only recently among Microsoft’s Excel team*

Summarizing data: how?

Some blood pressure data, from $n = 15$ unordered subjects...



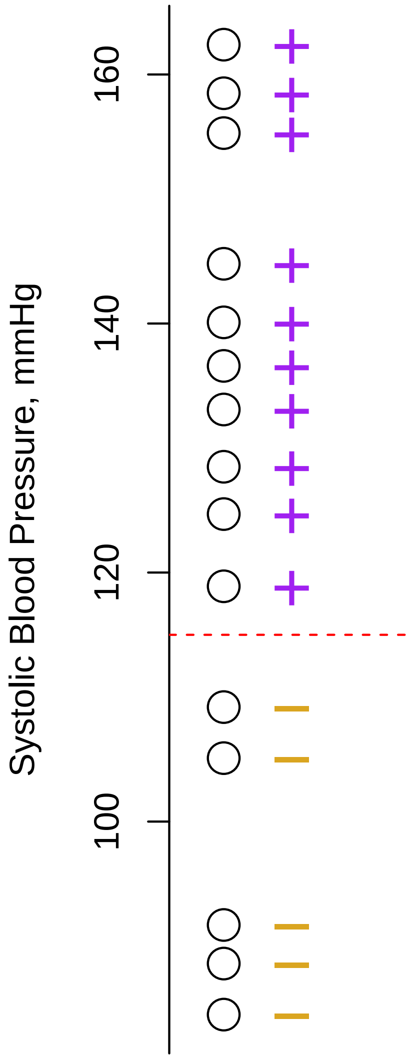
... on a vertical chart – so higher values are higher, lower values are lower

- There are *many* different ways to summarize this data – *none* of them are ‘right’ or ‘wrong’, or ‘require’ that data follow any particular pattern*
- I will motivate different choices as answers to different questions

* *Beware! This is a simple but often-misunderstood topic; so e.g. wikipedia entries – and some textbooks – can be unhelpful*

Summarizing data: find a balance

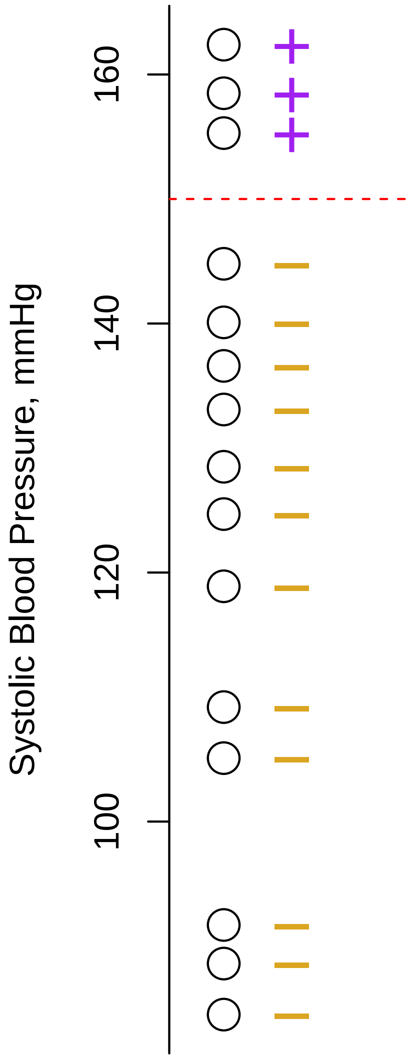
Q. ‘Where’s the balance point in the data’?



- For any value, mark points above ‘+’ and points below ’-’.
- What value balances these?
- Not this one (110 mmHg) ...too low

Summarizing data: find a balance

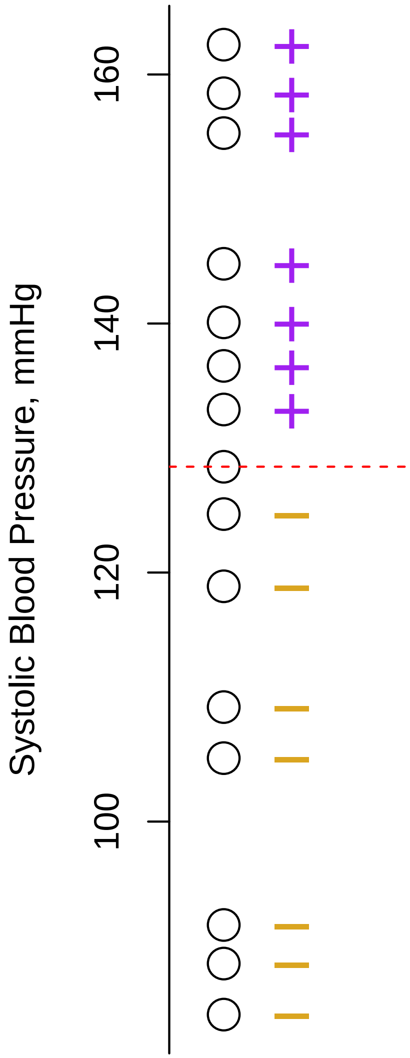
Q. ‘Where’s the balance point in the data’?



- For any value, mark points above ‘+’ and points below ’–’
- What value balances these?
- Not this one (150 mmHg) ...too high

Summarizing data: find a balance

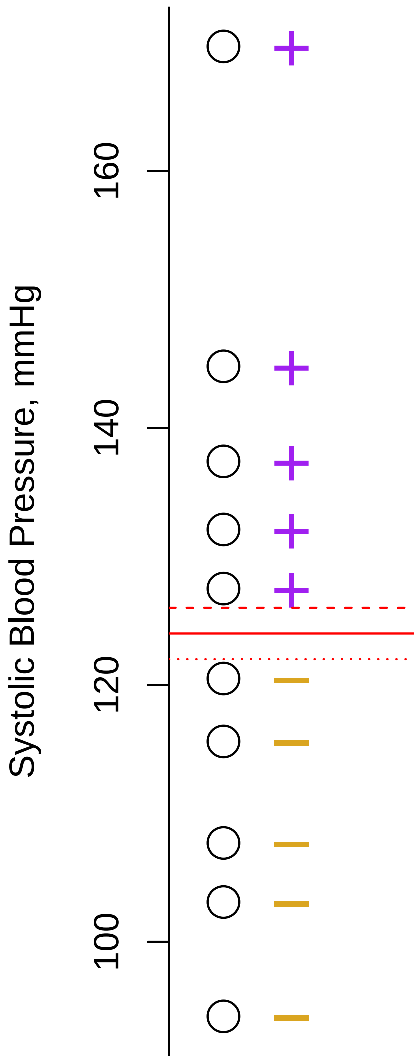
Q. 'Where's the balance point in the data'?



- For any value, mark points above '+' and points below '−'
- What value balances these?
- This one! (128.5 mmHg) − known as the *median* value
- Note that median point is neither '+' nor '-', so giving 7 data on each side

Details: the median for even n

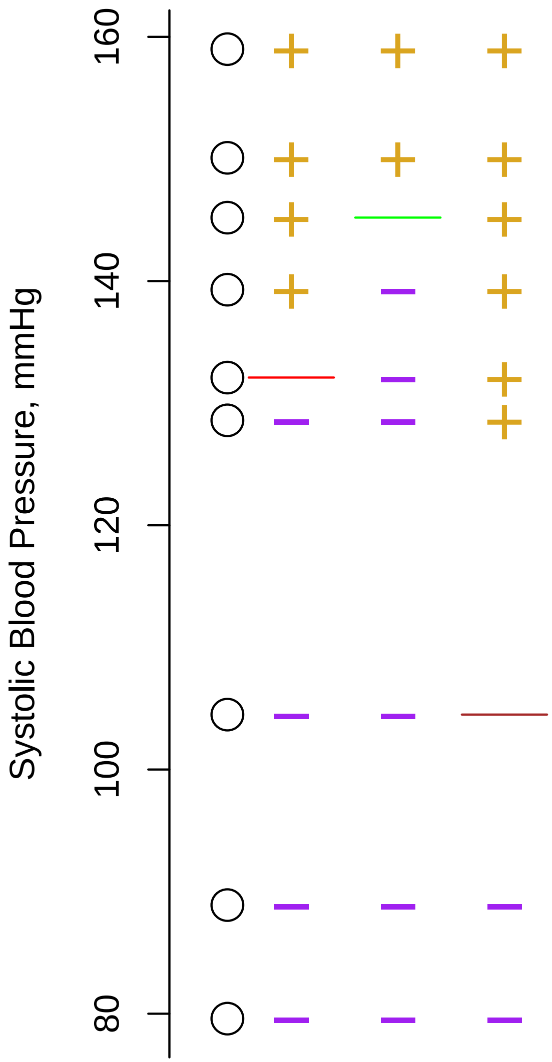
What to do when there is no 'middle'?



- Here n=10 – but would see same issue for any even n
- Here, any value between 5th, 6th points gives the same 'balance' – 5 on each side
- Default solution uses average (halfway point) between two middle data points – here 124 mmHg, the solid red line

Details: other quantiles

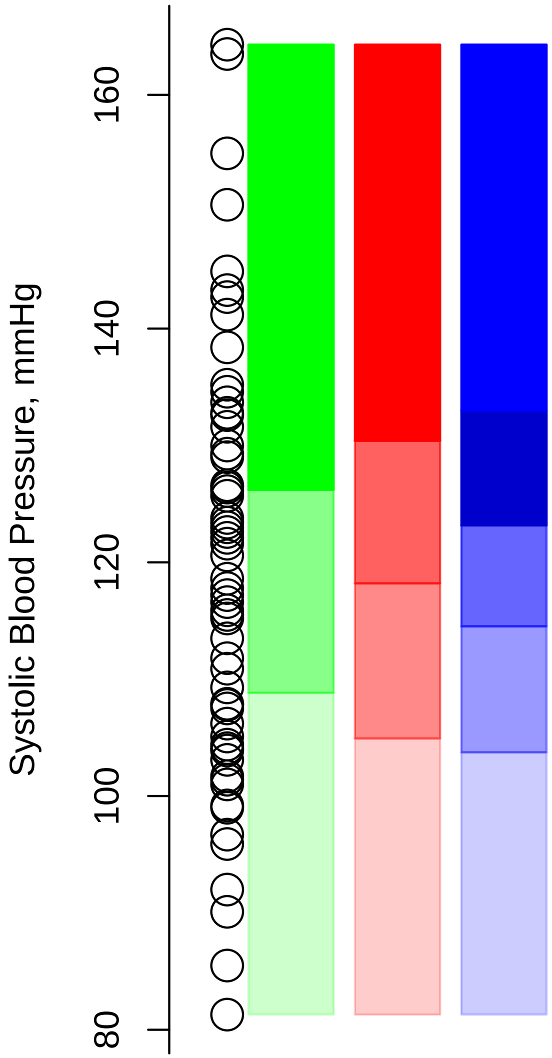
The median is a.k.a. the 50% *quantile*, or 50$^{\text{th}}$ *percentile*



- Shown here for $n = 9$; 50% above/below
- 75% quantile has 75% below, 25% above
- 25% quantile has 25% below, 75% above
- Could use any percentage between 0, 100%

Details: other quantiles

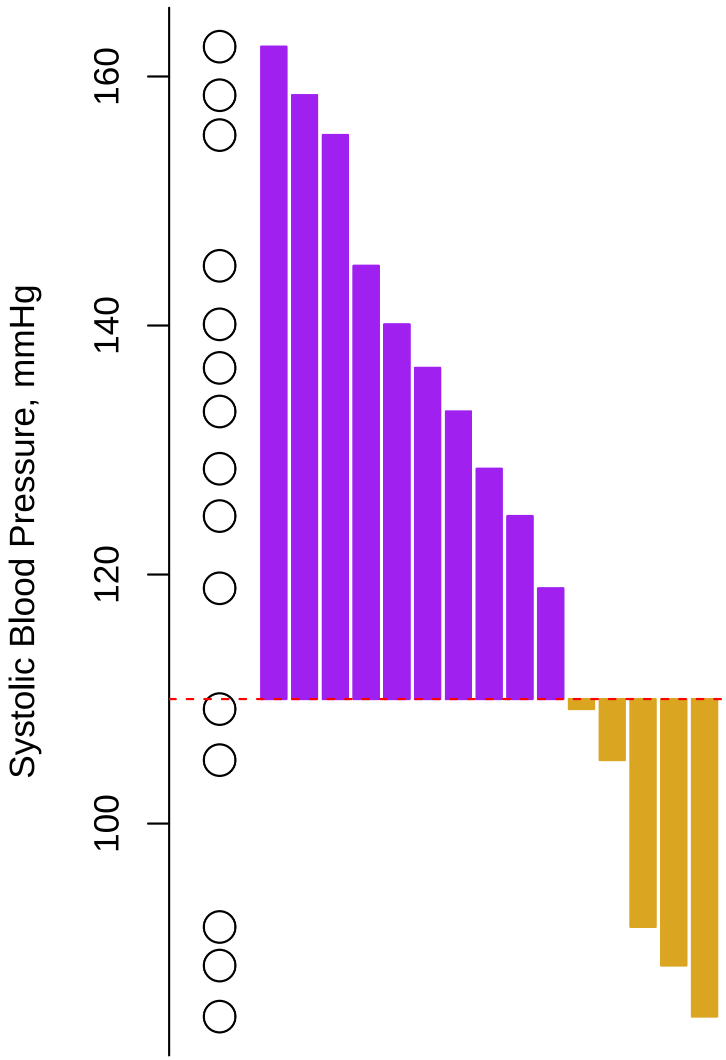
Special names for splitting data at evenly-spaced quantiles:



- Split at 33%, 66%: *tertiles*
- Split at 25%, 50%, 75%: *quartiles*
- Split at 20%, 40%, 60%, 80%: *quintiles*
- Same number of data in each 'bin' – this is NOT equal width bins
- When no exact quantile available, use special methods – not covered here

Summarizing data: find another balance

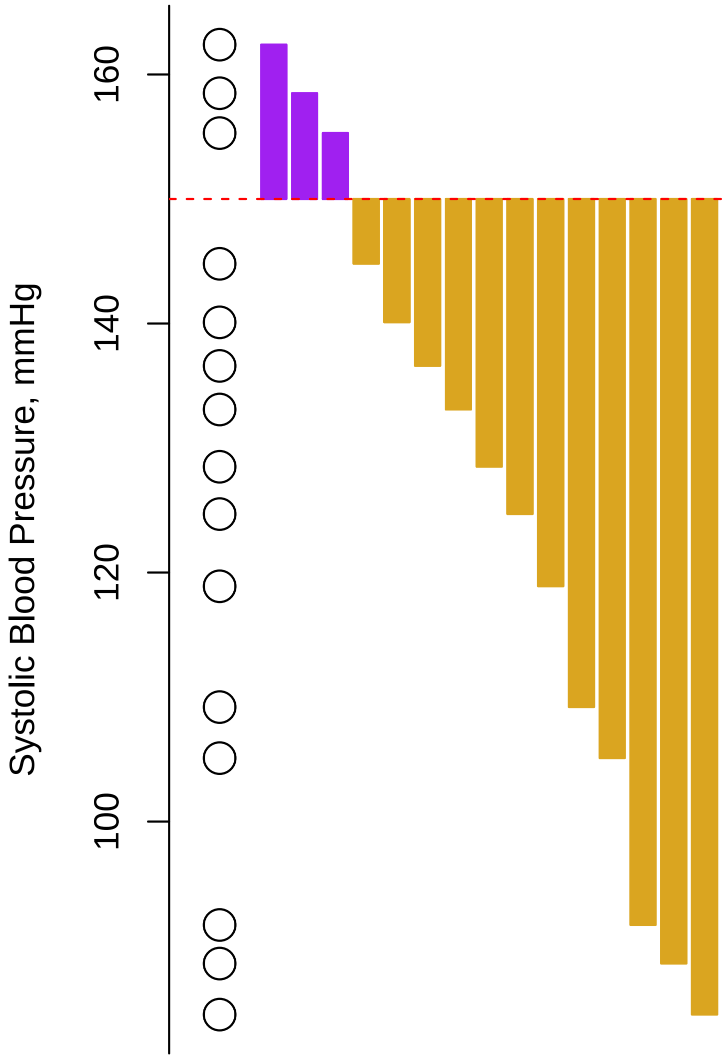
Q. 'Where's the balance point in the data'?



- Back to n=15, data as before
- For any value (red line) bars above are purple, below are gold
- What red line value balances *total* purple vs *total* gold?
- Not this one! (110 mmHg) – too low

Summarizing data: find another balance

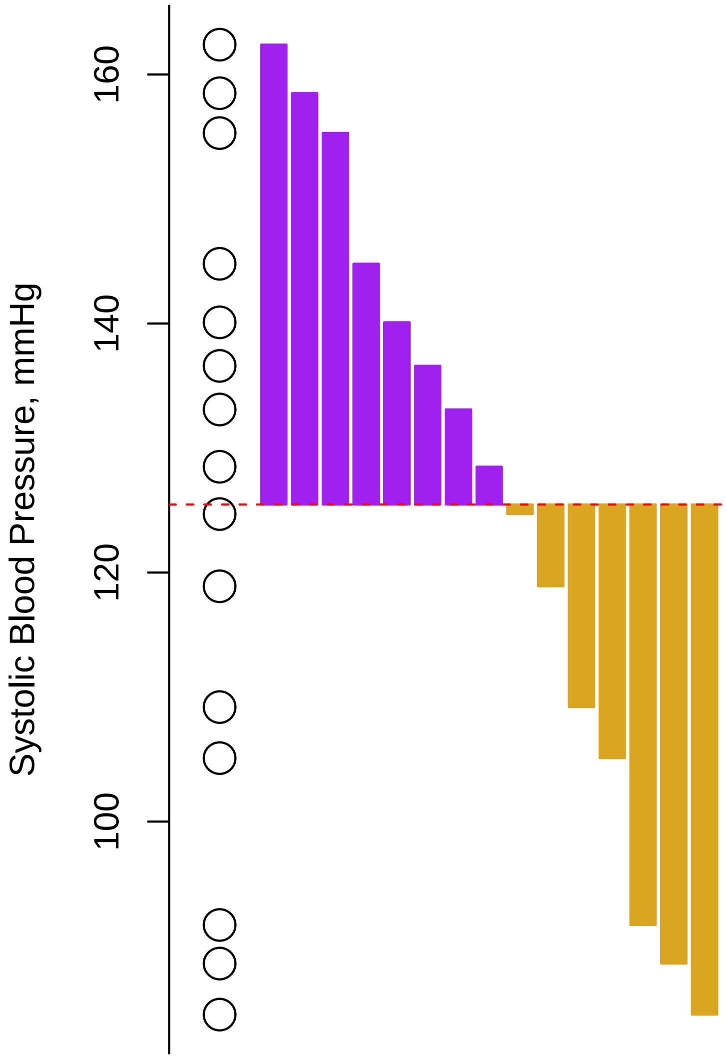
Q. 'Where's the balance point in the data'?



- Back to n=15, data as before
- For any value, bars above are purple, below are gold
- What red line value balances *total* purple vs *total* gold?
- Not this one! (150 mmHg) – too high

Summarizing data: find another balance

Q. 'Where's the balance point in the data'?



- Back to n=15, data as before
- For any value, bars above are purple, below are gold
- What red line value balances *total* purple vs *total* gold?
- This one! (125.5 mmHg) – known as the *mean*

Summary so far

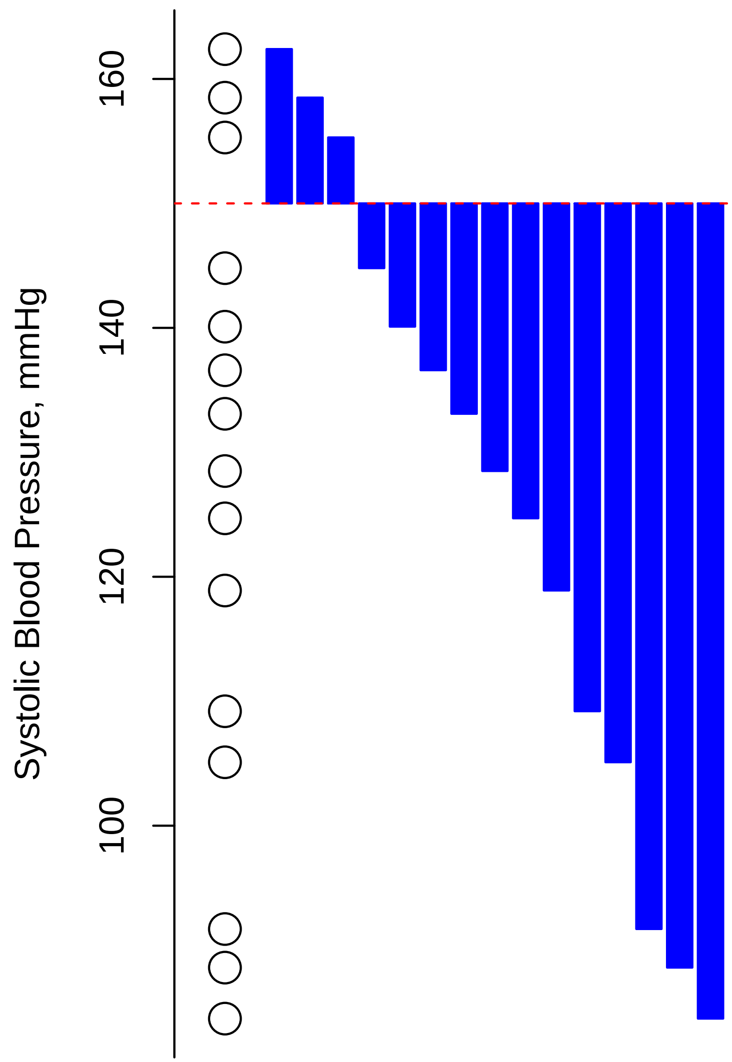
- The median value balances *number* of values above/below
- The mean value balances *deviations* of values above/below
- These are not the same criteria, hence don't give same answers (128 mmHg vs 125.5 mmHg)

Which to give? It's often fine to give both, but if you *must* pick:

- The mean is sensitive to extremes, while the median – depending only on the middle values – is not. Consider e.g. mean/median wealth & "the 1%"
- Means relate directly to totals – e.g. if I drove 10 miles in 30 mins, what was my mean speed? median?
- Means are often used in prediction – e.g. suppose in 1000 gambles each with \$0, \$1 for loss & win, that I win 600. What are my mean winnings per new gamble? Median?
- Pragmatism can be okay: if mean and median are close and you *must* give only one, your choice is unlikely to matter

Summarizing data: ‘best’ summaries

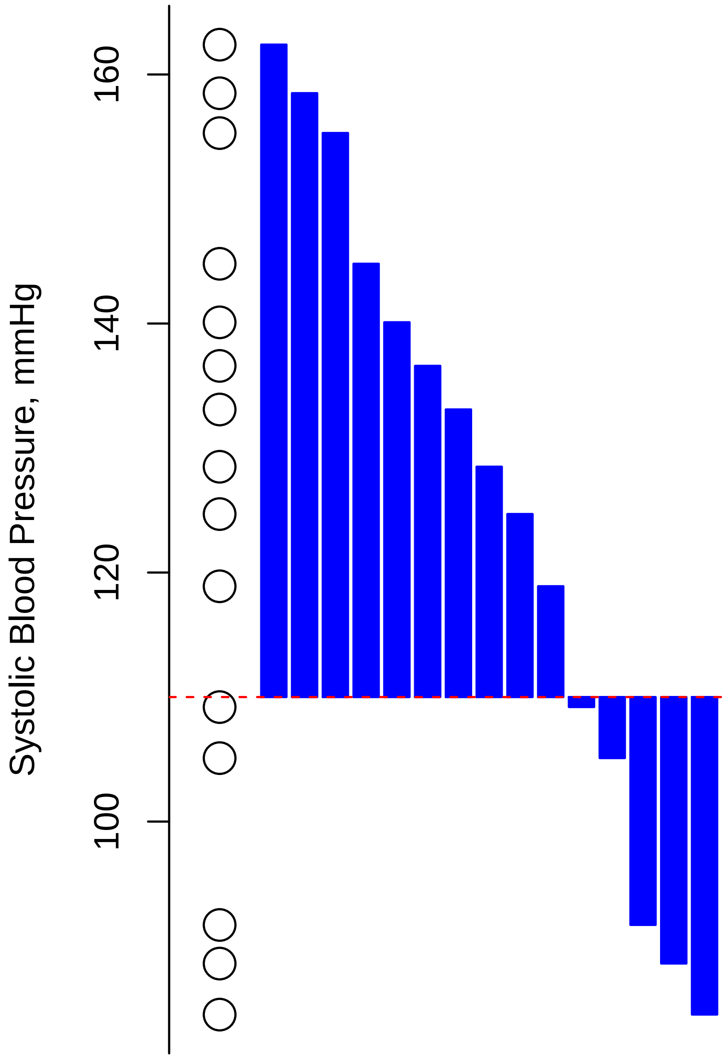
Q. ‘What value is *most* central in the data’?



- To measure ‘centrality’, for a given value (the red line) add up all the *deviations* (blue bars) from there to the data
- Q. What choice of red line *minimizes* the total amount of blue ink?
  (Not this one! – at 150 mmHg)

Summarizing data: ‘best’ summaries

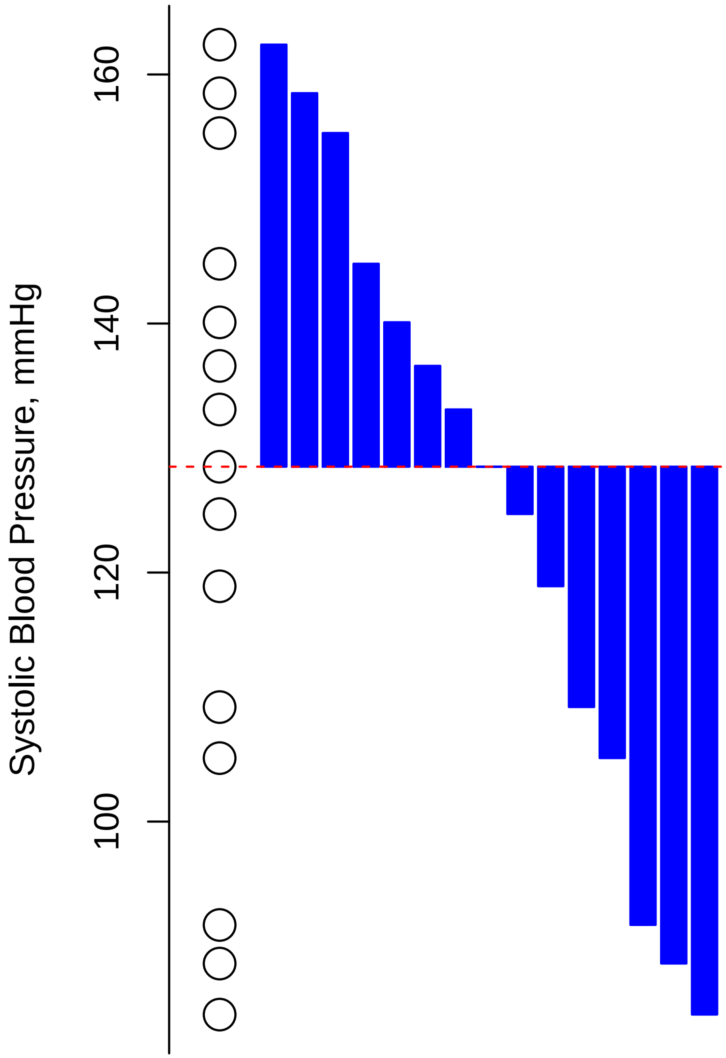
Q. ‘What value is *most* central in the data’?



- To measure ‘centrality’, for a given value (the red line) add up all the *deviations* (blue bars) from there to the data
- Another attempt... 110 mmHg
  Still not optimal!

Summarizing data: ‘best’ summaries

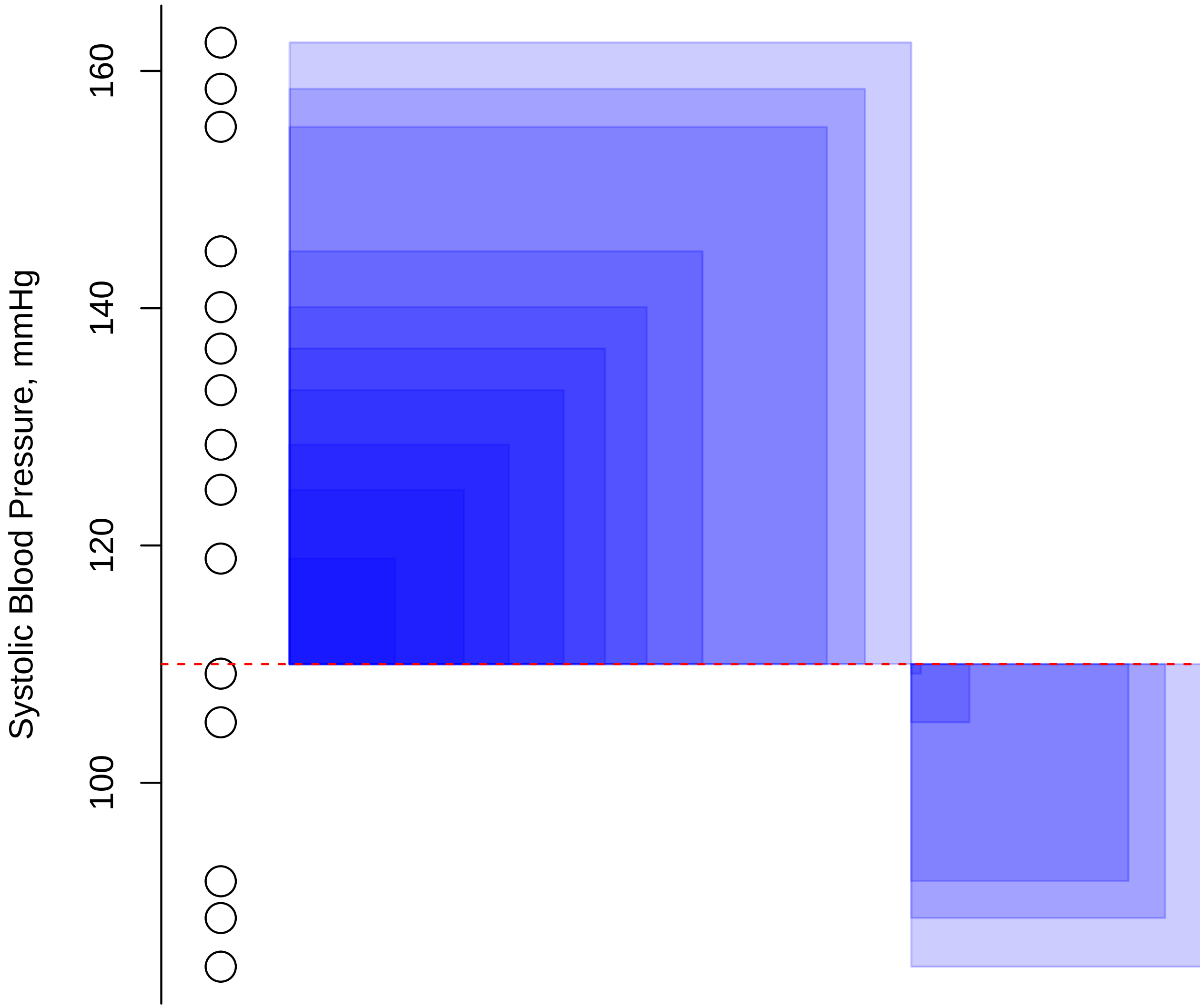
Q. ‘What value is *most* central in the data’?



- Putting the line at the ‘middle’ observation get closest – i.e. at 128 mmHg, for these data
- This is the *median* (again)
- For n even, all points between middle two values are equally most central

Summarizing data: 'best' summaries

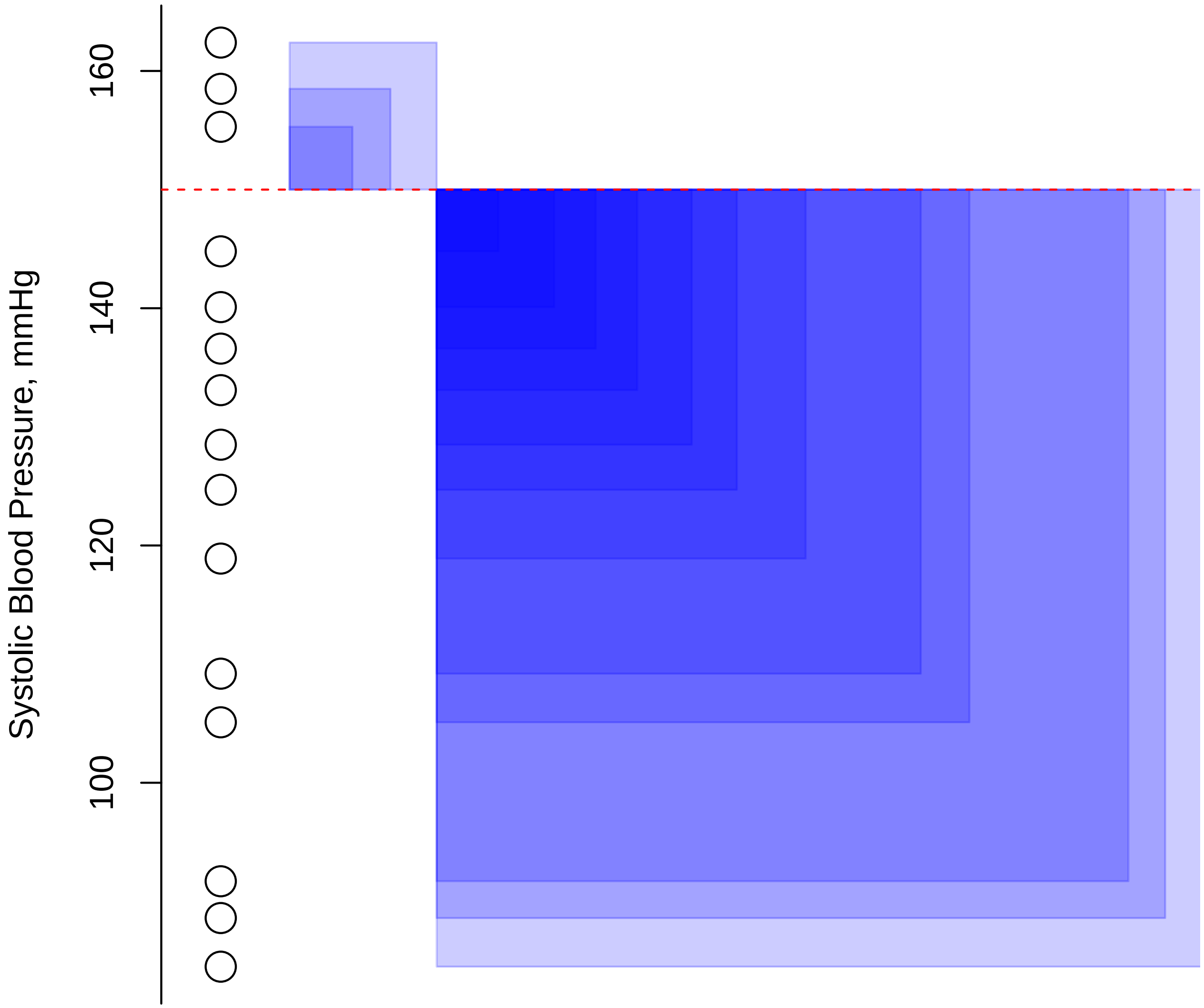
Another measure of 'centrality' uses area – *squared deviations*;



- What red line value minimizes total blue ink?

Summarizing data: 'best' summaries

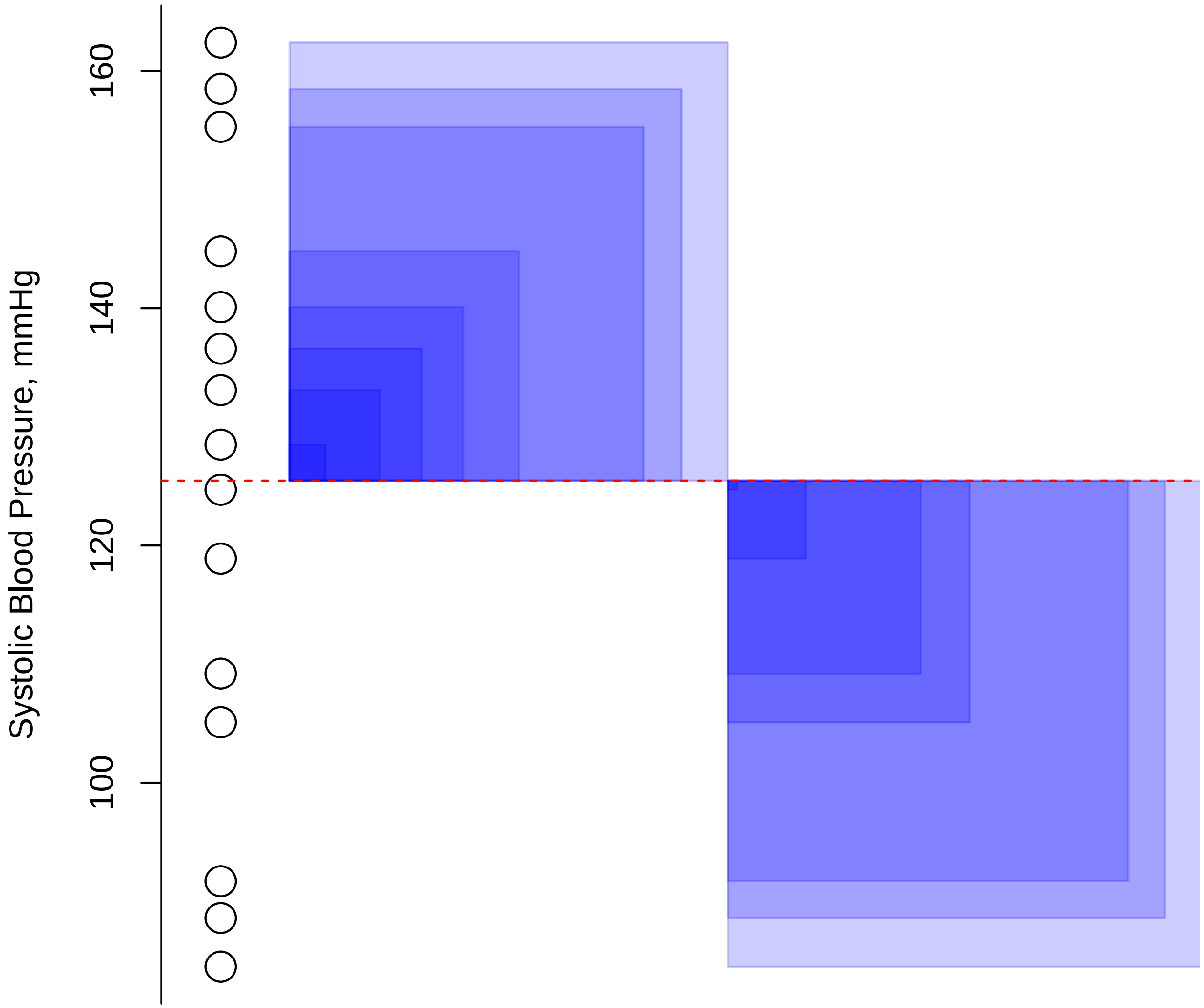
Another measure of 'centrality' uses area – *squared deviations*;



- What red line value minimizes total blue ink?

Summarizing data: ‘best’ summaries

Another measure of ‘centrality’ uses area – *squared deviations*;



- Best choice here is 125.5mmHg
- This is the *mean* (again)

Summarizing data: 'best' summaries

We saw before that median and mean reflect different types of balance. We can *also* interpret...

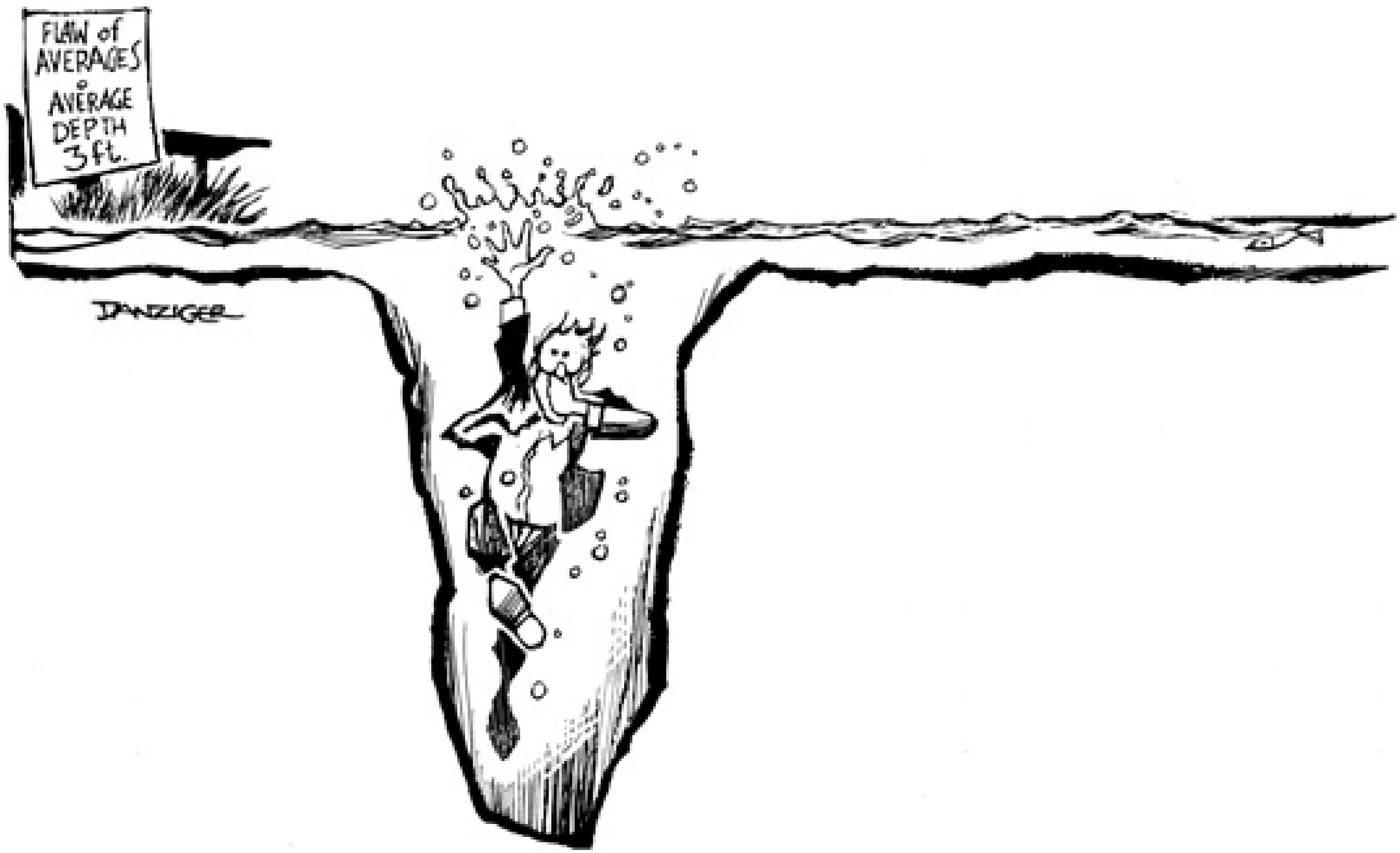
- ...the median as being most central, measured by *absolute deviation* – a measure of length
- ...the mean as being most central, measured by *squared deviation* – a measure of area

As before, these are different criteria – i.e. asking the data different questions – so they provide different answers.

Thinking about these deviations leads to measures of *dispersion* – how spread out is the data?

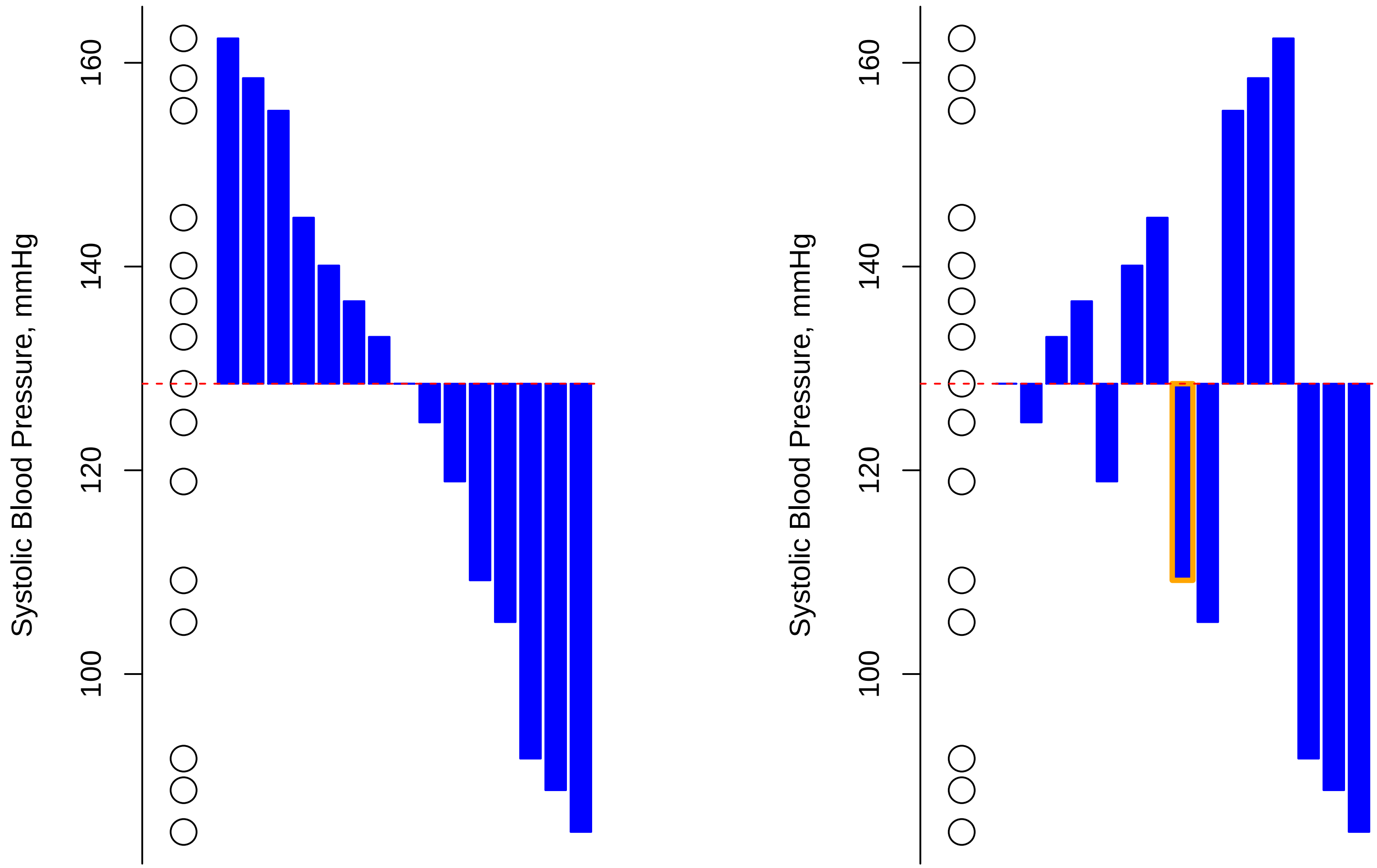
Summarizing data: 'best' summaries

Ignoring spread in the data is the *'flaw of averages'*...



Summarizing data: dispersion

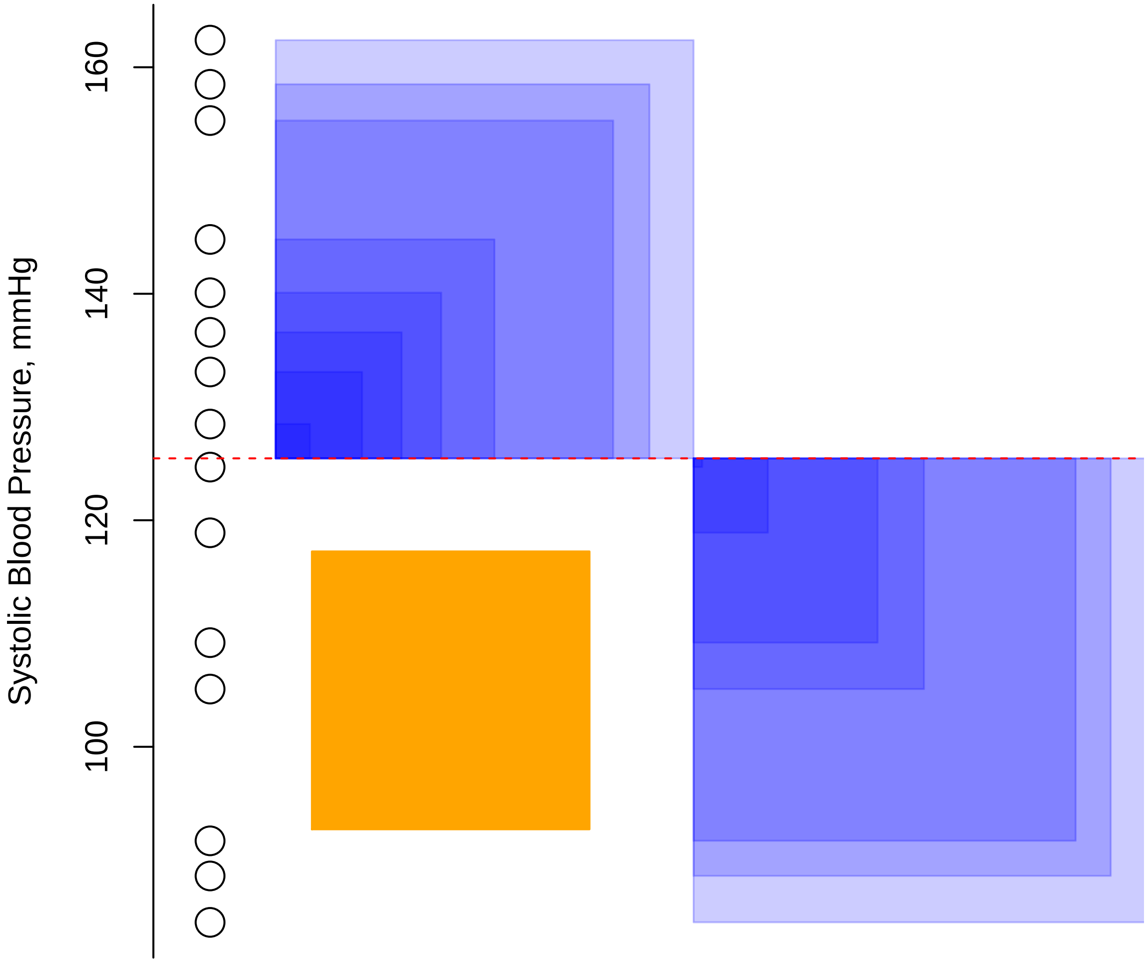
Q. Median length of blue bars around median? (Ordered, in RH)



The orange length (19.3 mmHg) is the *median absolute deviation* about the median – known as the MAD.

Summarizing data: dispersion

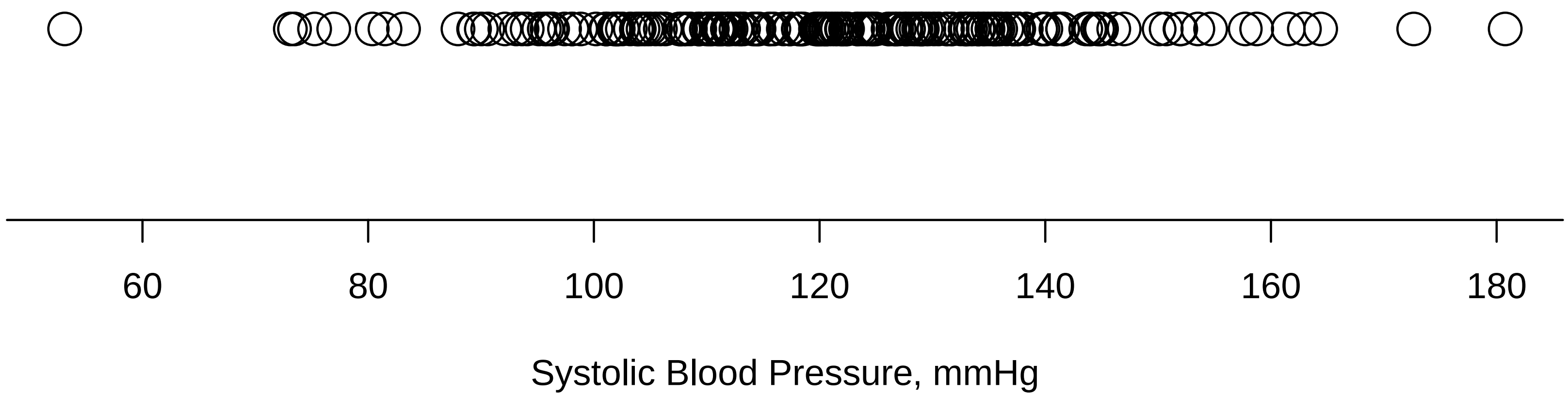
Q. Average area of blue box? (This is harder to ‘eyeball’)



Area of this ‘average box’ (602 mmHg^2, in orange) is the *variance* – its edge length (24.5 mmHg) is the *standard deviation*.

What to do with more data?

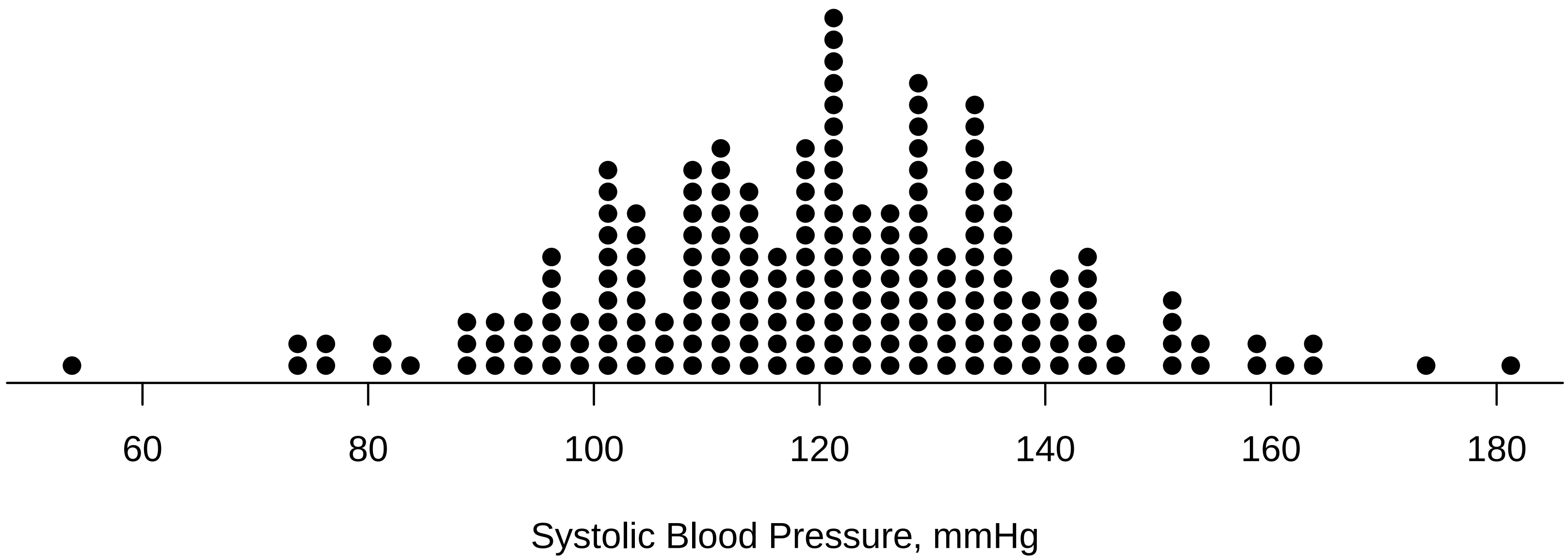
Dotcharts get a bit clumsy beyond $n = 30$ − here is $n = 200$;



- Exact SBP for any individual not important
- Want to get an idea of the location (center) and dispersion (spread) of the data
- *Coarsened* data will do, for a summary

What to do with more data?

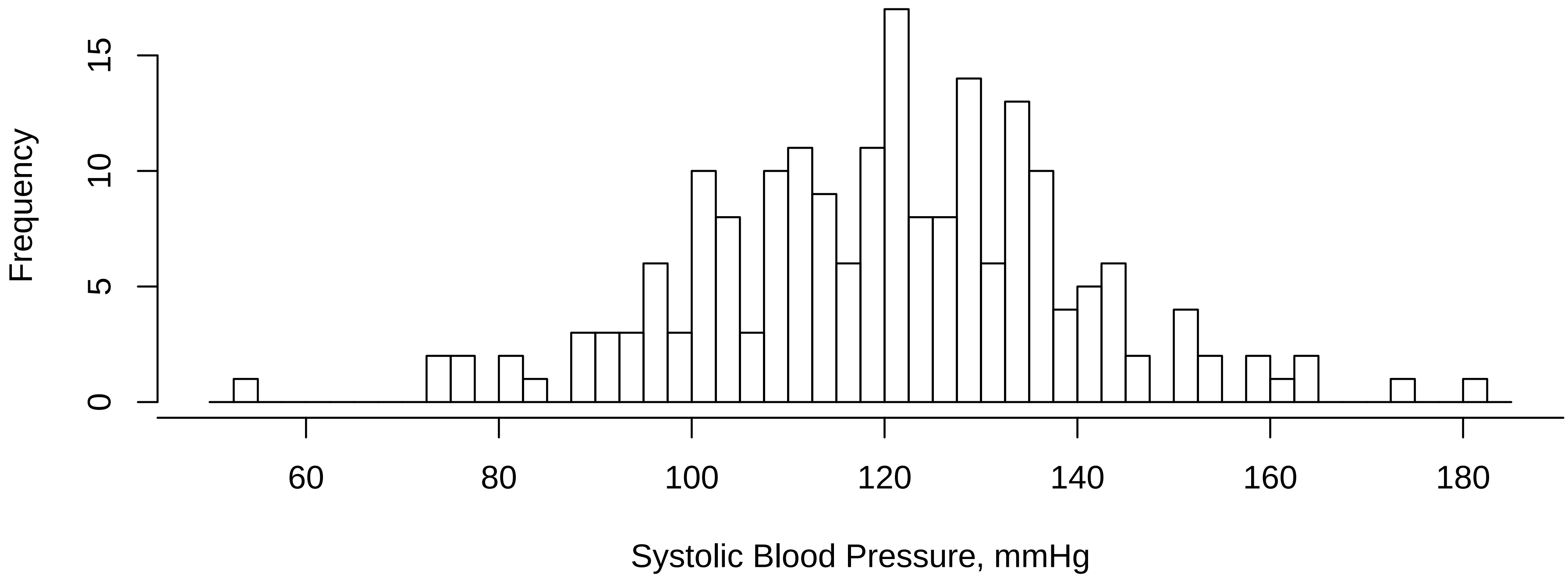
A *stacked dotchart* for the same data;



- 'Bins' every 2.5 mmHg (120, 122.5, 125 etc)
- Count the data points in each bin
- Plot one point per observation, in each bin
- How to read off median? 75% quantile?

What to do with more data?

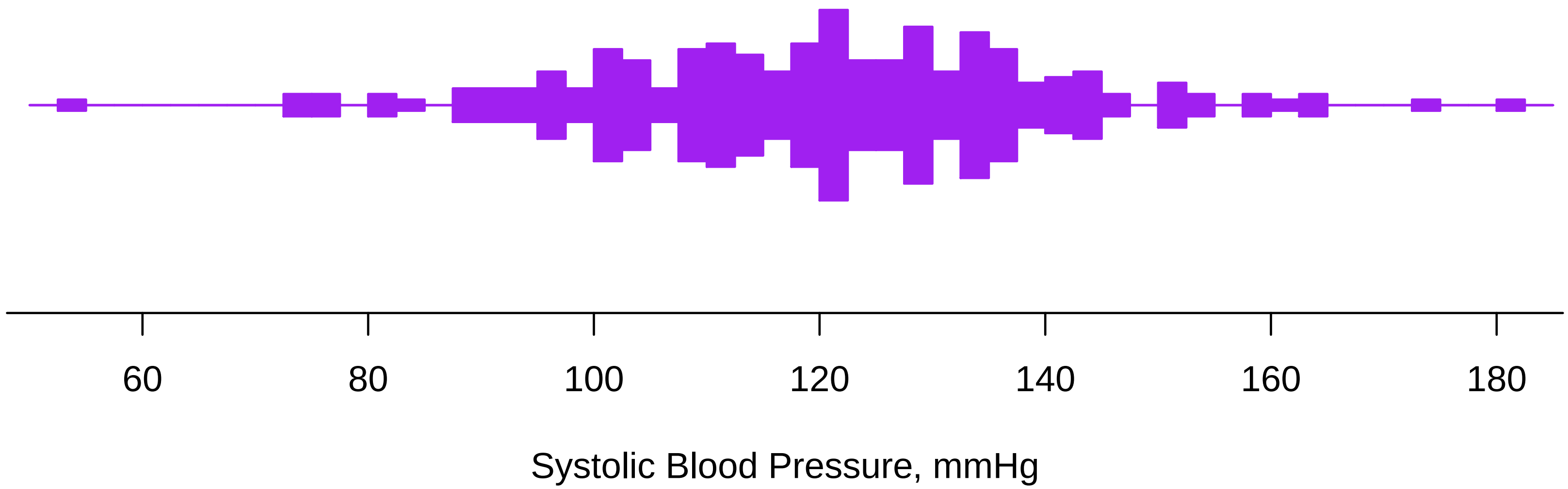
A *histogram* for the same data;



- ‘Bins’ every 2.5 mmHg (120, 122.5, 125 etc)
- Count the data points in each bin
- Bin height proportional to this count, a.k.a. frequency
- Better than stacking, for large n

What to do with more data?

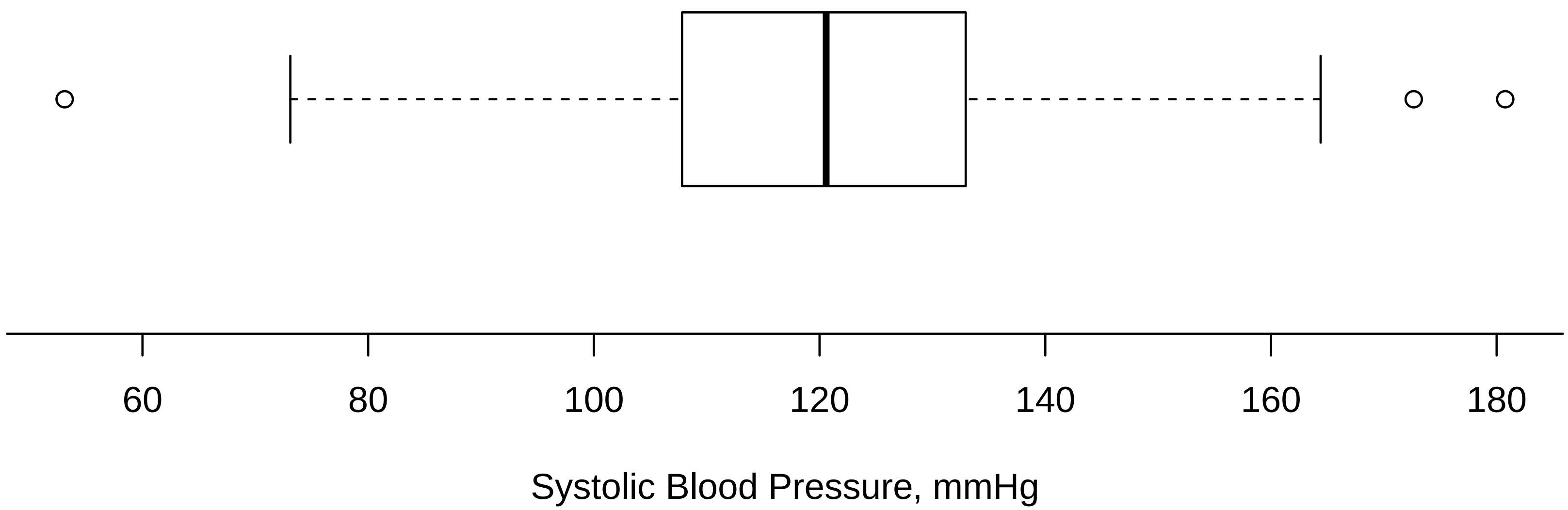
A *violinplot* for the same data;



- 'Bins' every 2.5 mmHg (120, 122.5, 125 etc)
- Count the data points in each bin
- Bin height proportional to this count, a.k.a. frequency
- Better than stacking, for large n

What to do with more data?

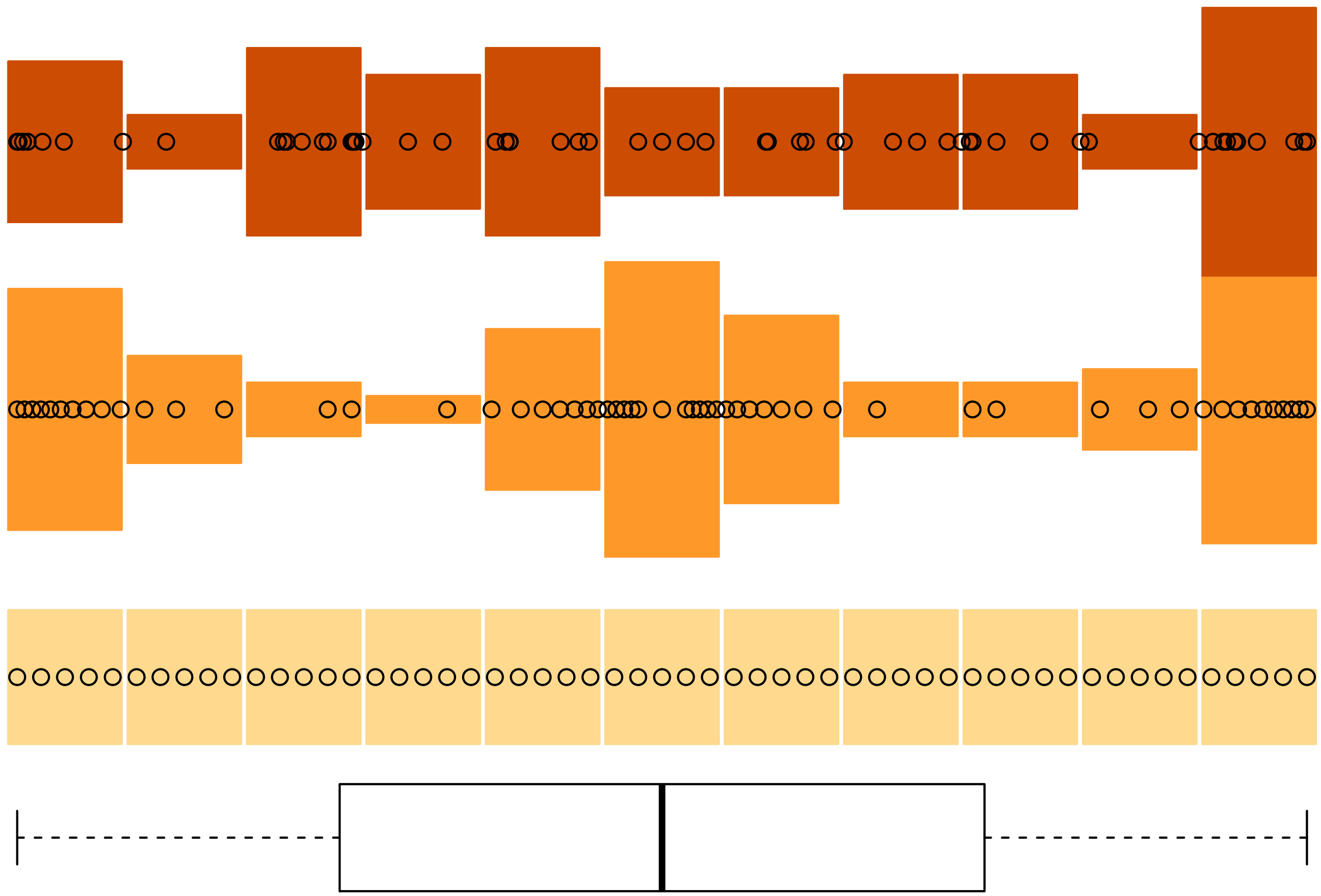
Finally, a *boxplot*; (short for *box-whisker plot*)



- Solid bold line is the median, box edges are 25% and 75% quantiles, box width is the interquartile range (IQR)
- 'Whiskers' go to last point up to 1.5×box width *beyond* box
- Points beyond this plotted individually
- Fancier versions exist, but this is the default

What to do with more data?

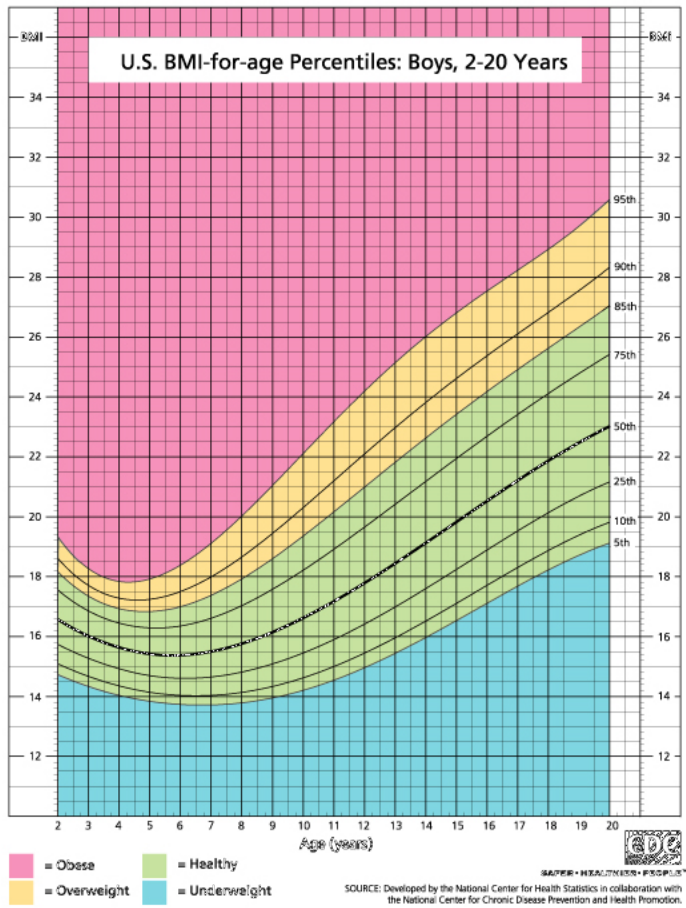
Boxplots are crude – cruder than dotcharts, and violinplots;



The plot shows 3 different datasets: all give the same boxplot.

What to do with more data?

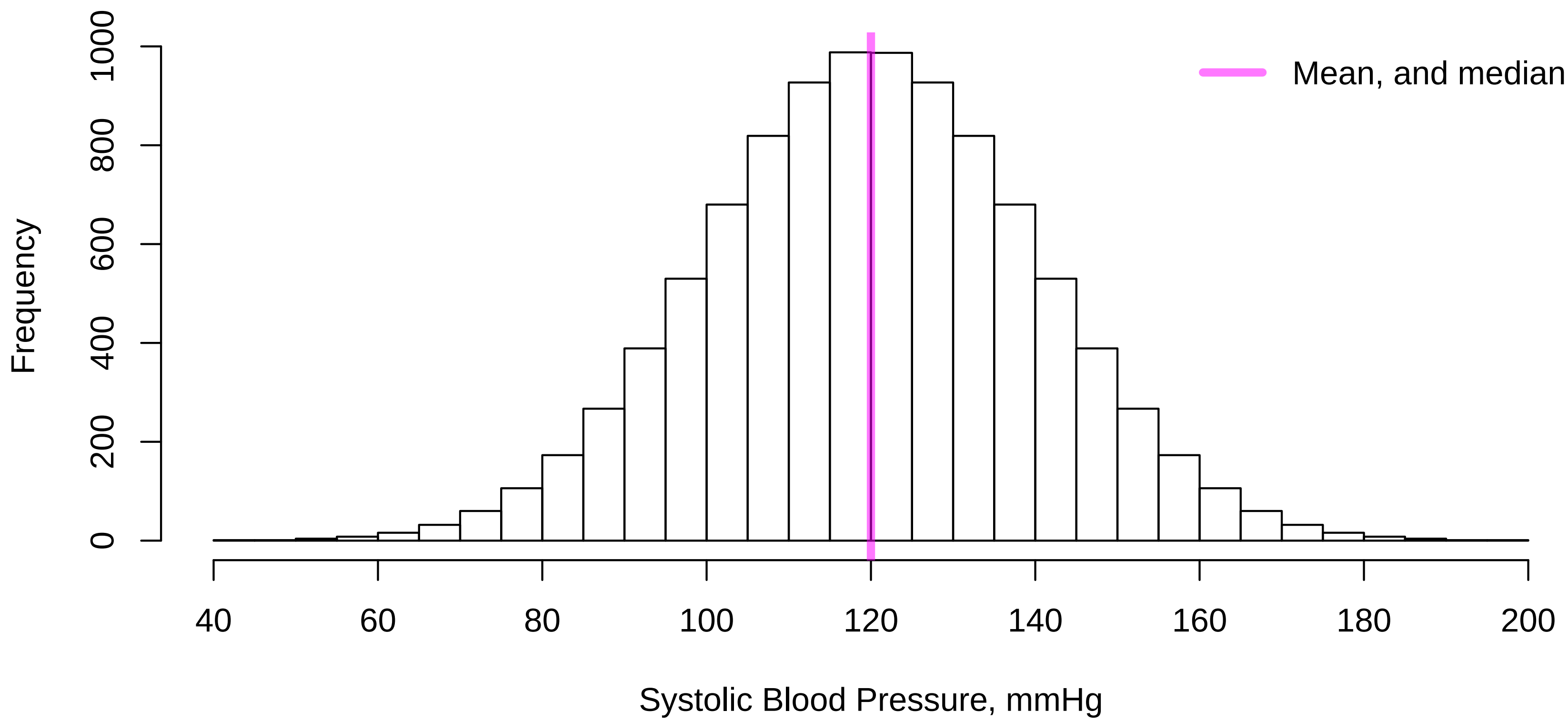
But plotting just quantiles aids comparison of many groups;



A *quantile plot*, showing various percentiles BMI by (many) ages

Details: symmetry

When the histogram (or violinplot) is symmetric, the mean and median *must be* equal*;

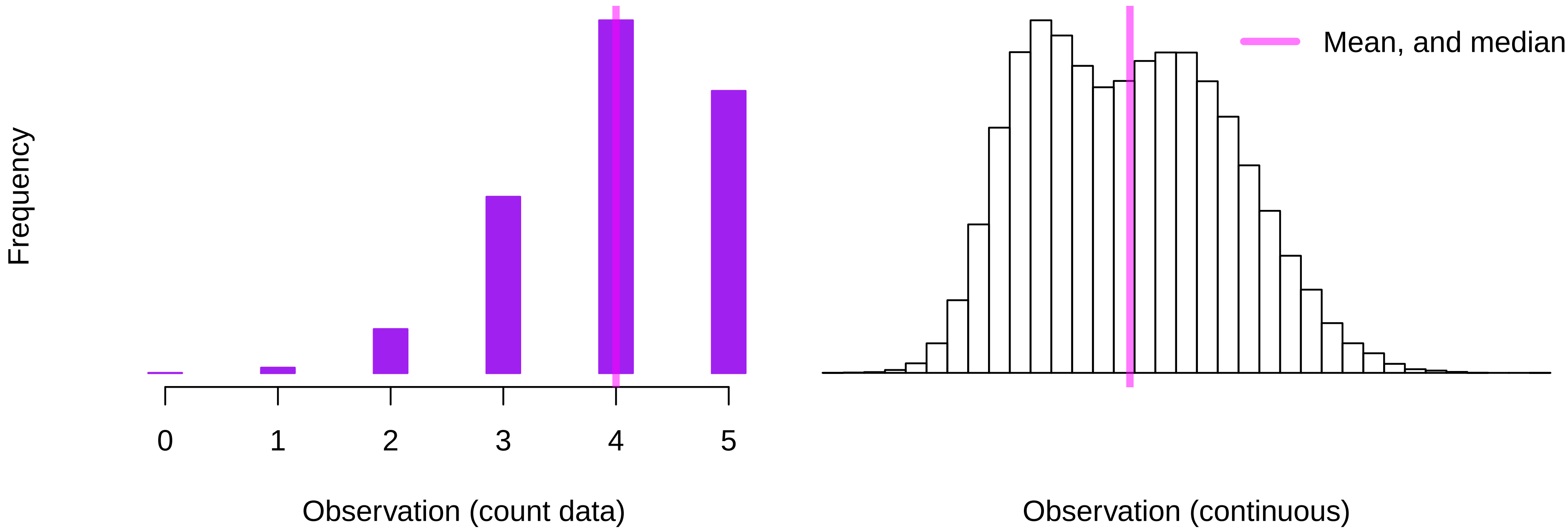


- 50% of values above & below axis of symmetry
- Equal deviations above & below axis of symmetry

* *And if the histogram is approximately symmetric, mean and median must be approximately equal*

Details: symmetry

But mean and median being the same does NOT imply the distribution is symmetric – even approximately*;

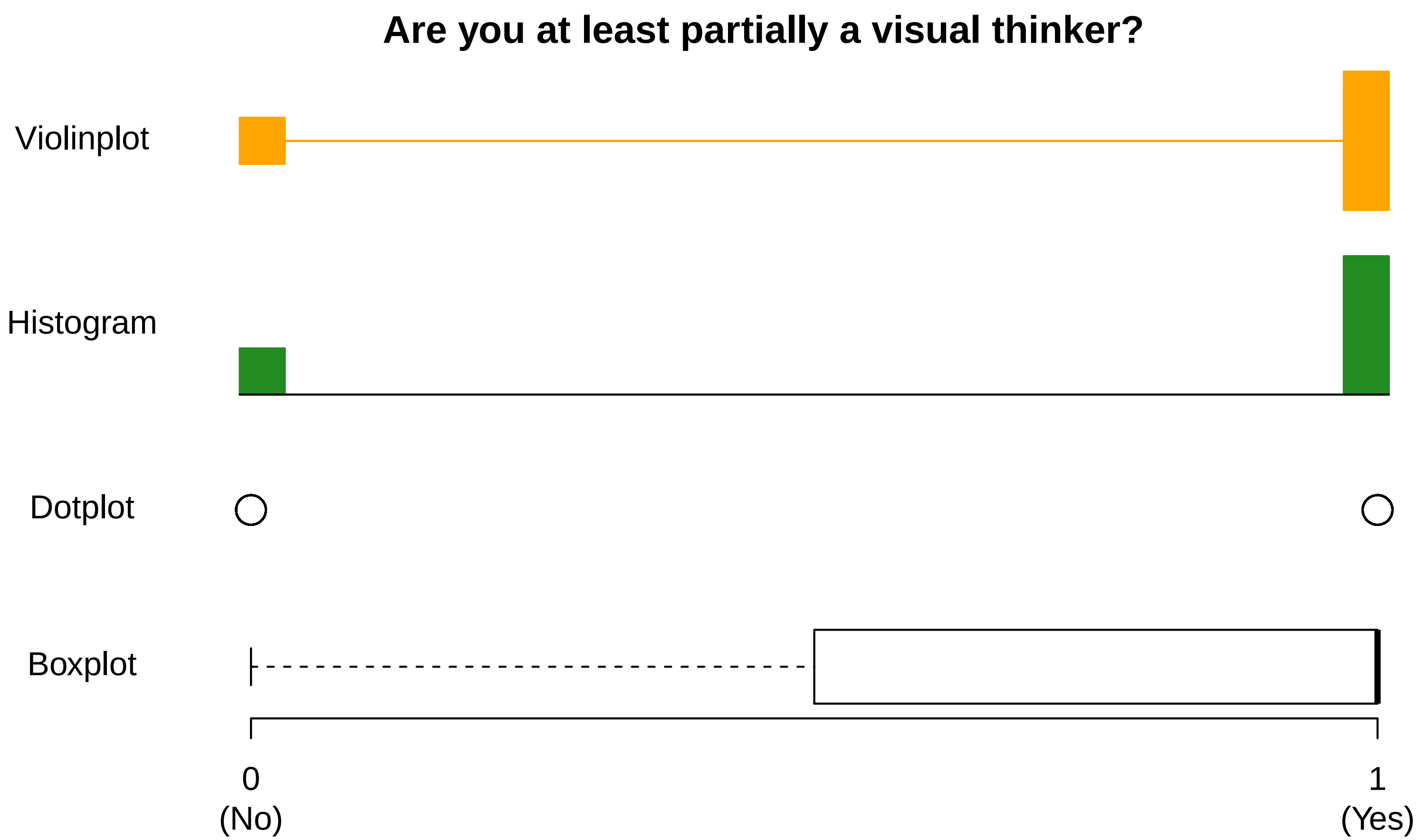


Keen people: many texts claim seeing mean<median or mean>median implies data is skewed to the left/right, respectively. But this is *not true* for standard measures of skewness.

* *LH example is e.g. number of 'successes' from 5 trials, each with 80% chance of success. RH could be e.g. height*

Binary & categorical data

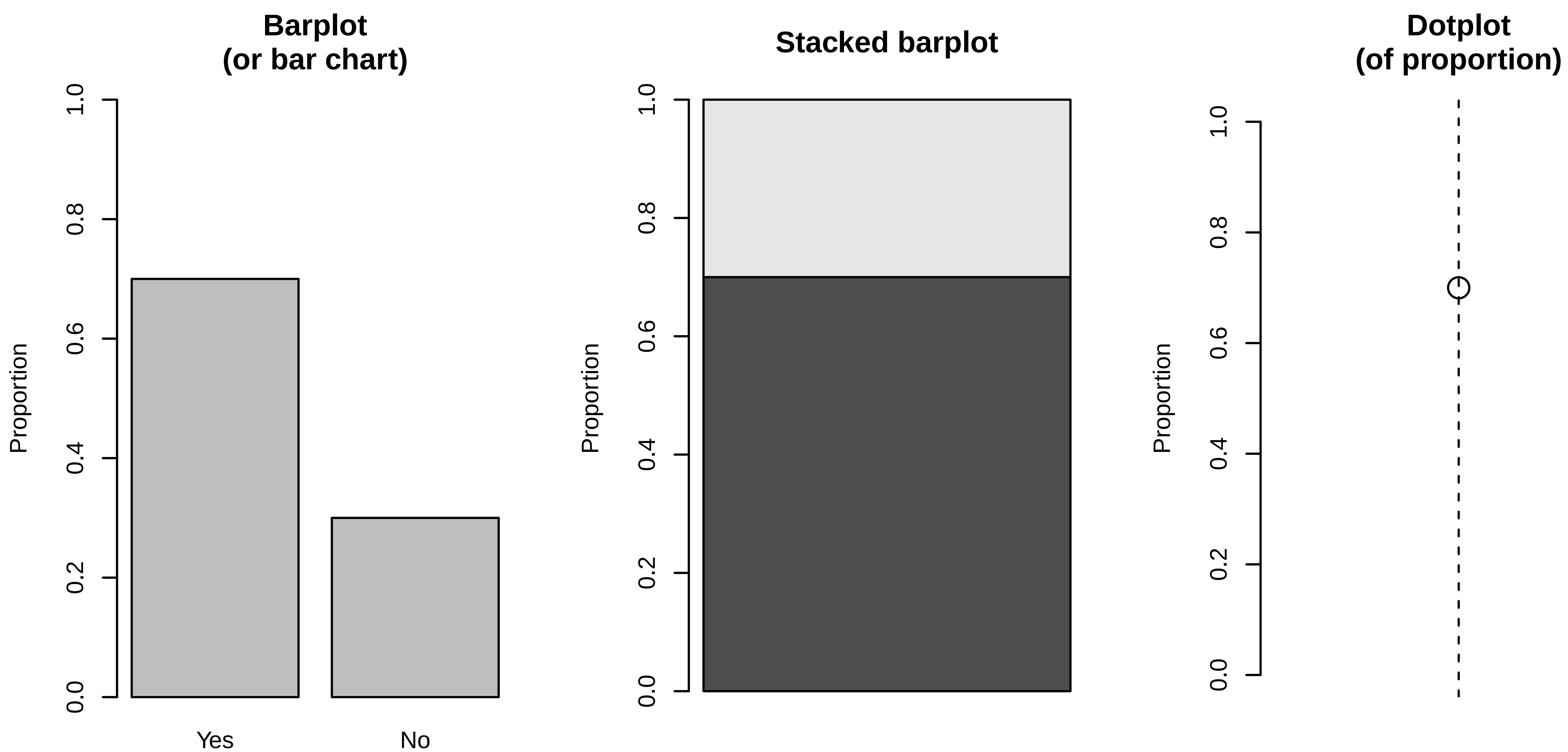
None of the approaches we have seen are great for binary (Yes/No) outcomes, e.g. death, pregnancy, hypertension.



These all show 750 Yes (coded as 1) and 250 No (coded as 0).

Binary & categorical data

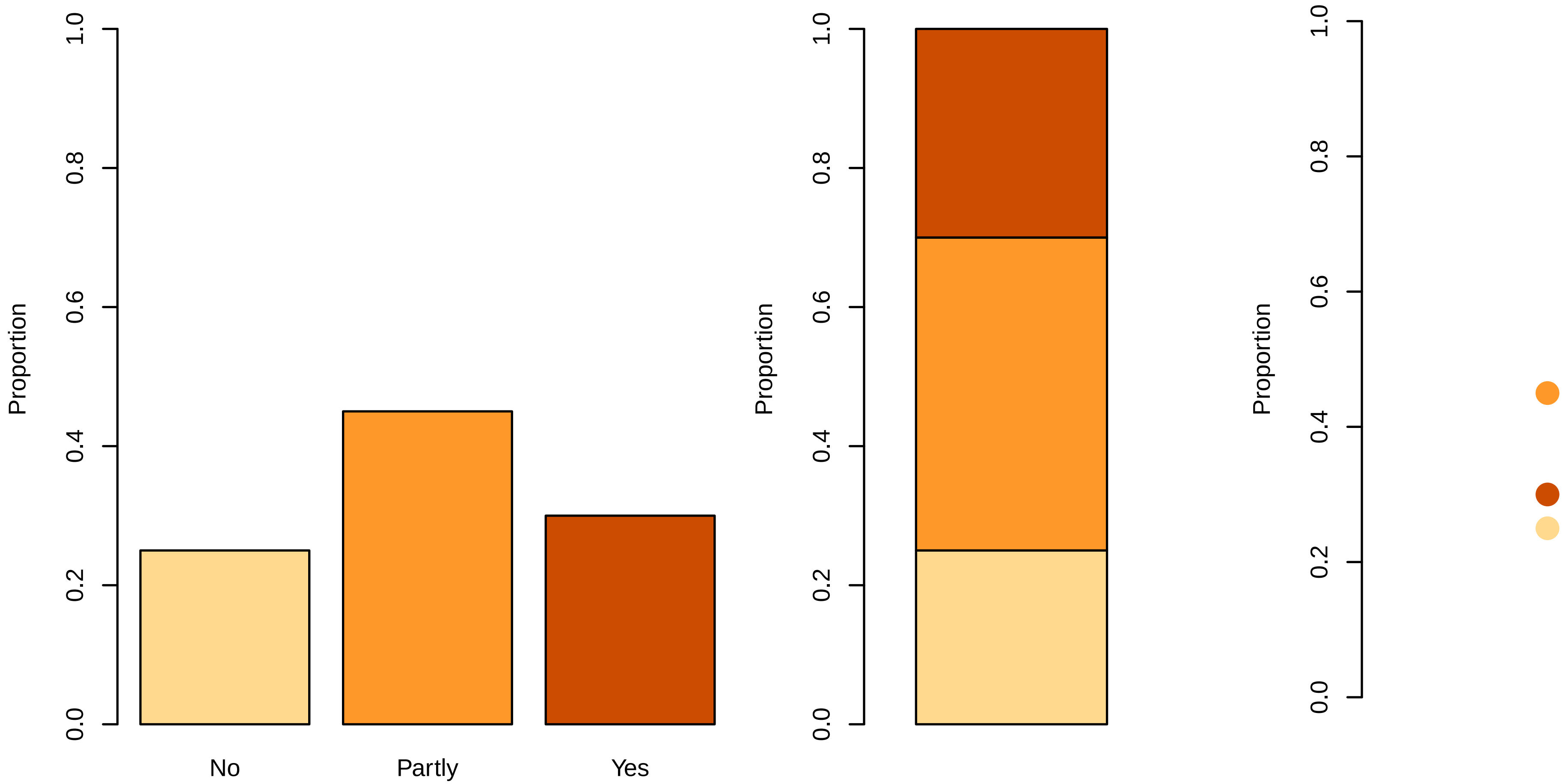
Instead, just give the percentage of 'Yes'; (somehow)



The dotchart emphasizes we've reduced the entire dataset (here n=1000) to just one number.

Binary & categorical data

For categorical data, the same ideas work;



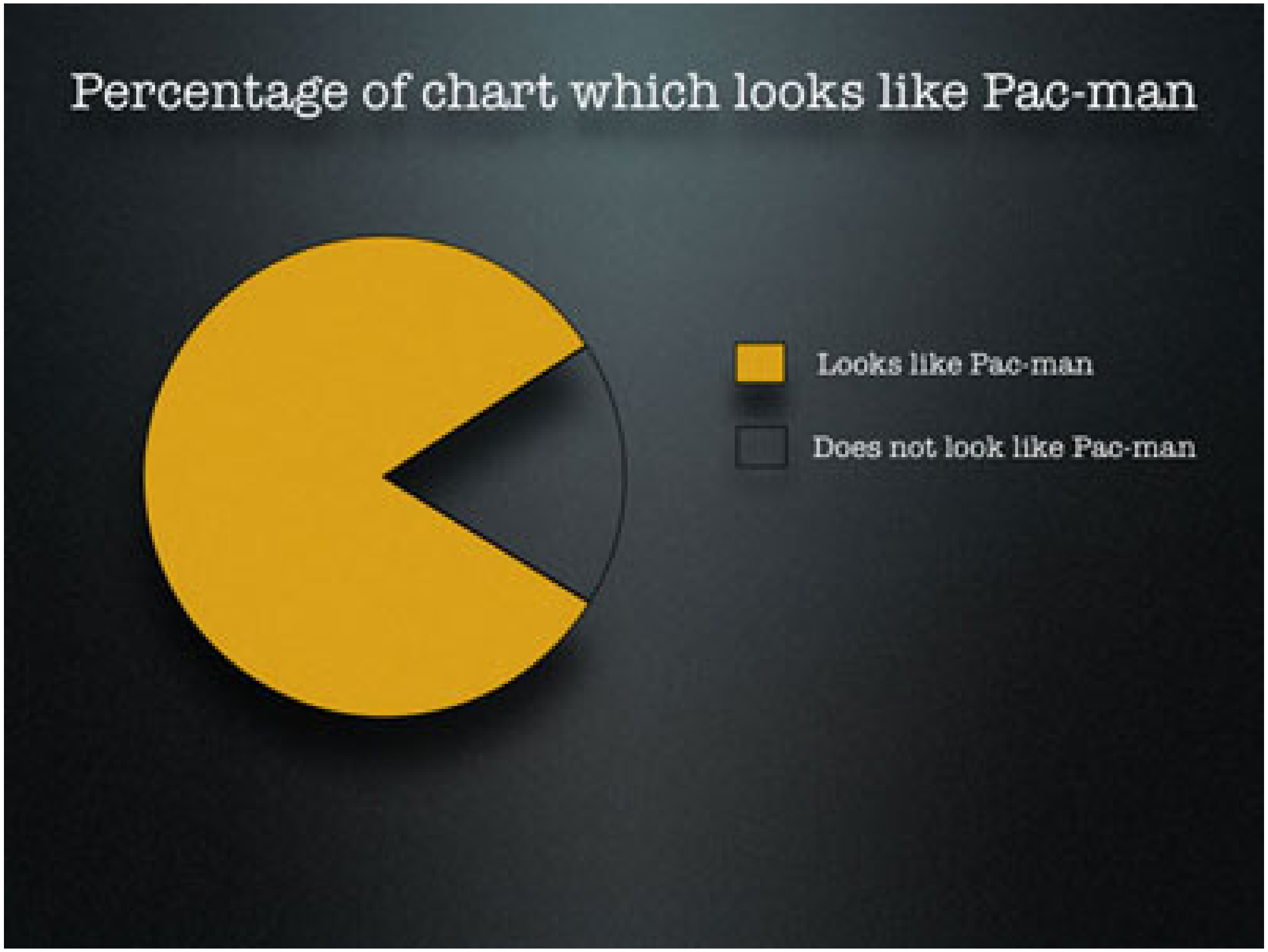
For unordered factors (e.g. hair color) ordering may not matter.
Frequently-asked-Q: *Why not use a pie chart?*

Why not use a pie chart?

Because they encode data as *angles*, not positions on a common scale – and work less well than the alternatives. But...

Why not use a pie chart?

Because they encode data as *angles*, not positions on a common scale – and work less well than the alternatives. But...



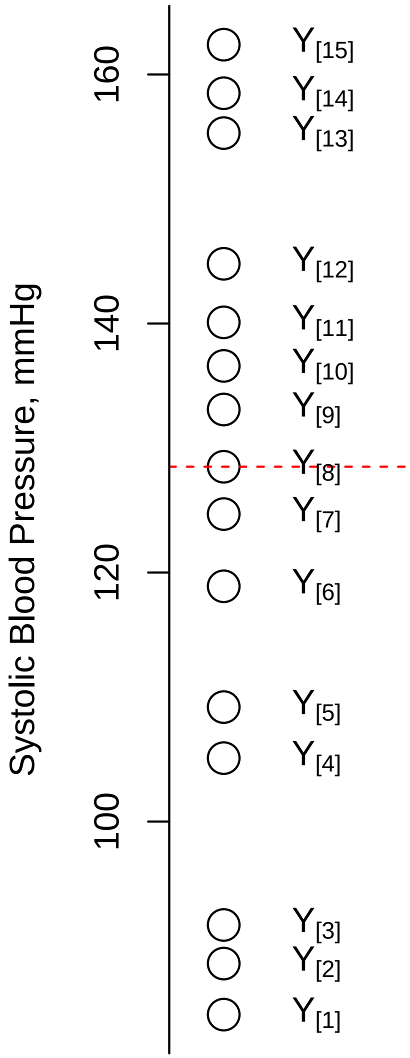
Summary

Main points

- Data summaries have graphical interpretations – often more than one interpretation
- There is no ‘right’ or ‘wrong’ summary (despite what some texts say) but they do communicate different aspects of the data
- What do you want to communicate? What is relevant to your analysis? (You must decide, the data won’t tell you)
- For larger datasets, trade off data ‘coarseness’ for clarity of message

Appendix: Math (for keen people)

First, calculating the median for our $n = 15$ example;



- Give new *ordered* labels; $Y_{[1]}, Y_{[2]}, ..., Y_{[15]}$
- Find the middle data point – we have $n = 15$, so $Y_{[8]}$ has 7 data points both above and below
- For Q 'What value is in the middle?', the median is (also) your answer

Appendix: Math (for keen people)

Now without a picture;

1. Start with n observations
   item Put them in increasing order, so
   $$Y_{[1]} < Y_{[2]} < Y_{[3]} < Y_{[4]} < \;...\; < Y_{[n-2]} < Y_{[n-1]} < Y_{[n]}$$

2. • For n odd, Median $= Y_{[(n+1)/2]}$
   • For n even, Median $= \frac{1}{2}(Y_{[n/2]} + Y_{[n/2+1]})$, i.e. the average of the two ‘middle’ values

For other quantiles, special methods (not covered here) are used when, as with n even, there is no uniquely-defined quantile.

Appendix: Math (for keen people)

For the mean:

$$\text{Mean} = \frac{Y_1 + Y_2 + ... + Y_n}{n}$$

i.e. to get the mean, add all the data points, then divide by the number you have. In 'math' notation, this is written as;

$$\text{Mean} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

... where the numerator (i.e. top part) represents the 'adding them all up' step, from 1 to n.

- Unlike the median, no need to order the data
- Also, no special treatment of n odd/even, or with ties

Appendix: Math (for keen people)

Defining measures of dispersion (spread) requires more notation;

Median absolute deviation;

$$MAD = \text{Median}\{|Y_i - \text{Median}\{Y_1, Y_2, ..., Y_n\}|\},$$

where $\text{Median}\{Y_1, Y_2, ..., Y_n\}$ means ‘take the median of all the observations $Y_1, Y_2, ..., Y_n$, and the vertical bars $|Y_i - ...|$ denote *absolute values*.

Variance and Standard Deviation;

$$\begin{aligned} \text{Variance} &= \frac{\sum_{i=1}^{n}(Y_i - \text{Mean})^2}{n} \\ \text{StdDev} &= \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \text{Mean})^2}{n}} = \sqrt{\text{Variance}} \end{aligned}$$

Note: many texts will define these with $n-1$ instead of n, in the denominator – with almost-always minor impact.

Why do that? Using $n - 1$ removes bias when using *sample* variance to estimate *population* variance.

Appendix: Math (for keen people)

Want more? The mean and median are both measures of 'location', or 'measures of central tendency'. There are many more of these, but mean & median are most commonly-used.