Numerical Summaries of center and spread of quantitative variables

David Gerard

2017-09-18

Learning Objectives

- Mean/median.
- Standard deviation/median absolute deviation.
- Sections 1.6.2, 1.6.4, in DBC.

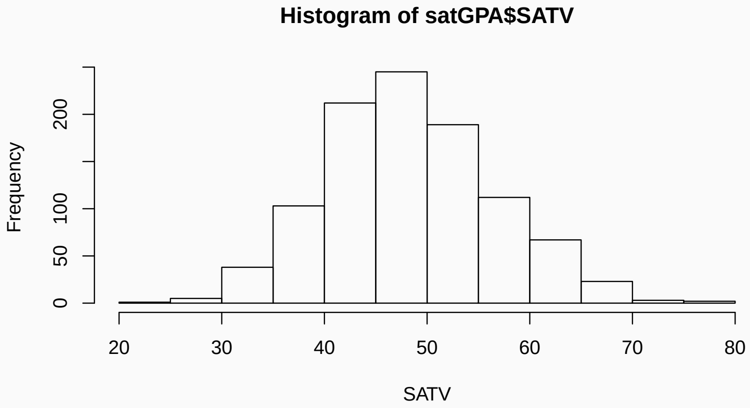
Numerical Summaries

- Sometimes it is inconvenient to provide a graphical summary of your data.
- An alternative is to provide *numerical* summaries of data.
- Summarizing the data numerically can also provide insights into distributions.

Measures of Center

Where is the distribution's "center"?

```
library(tidyverse)
data(satGPA, package = "openintro")
hist(satGPA$SATV, breaks = 15, xlab = "SATV")
```



The mean

One measure of center is the mean.

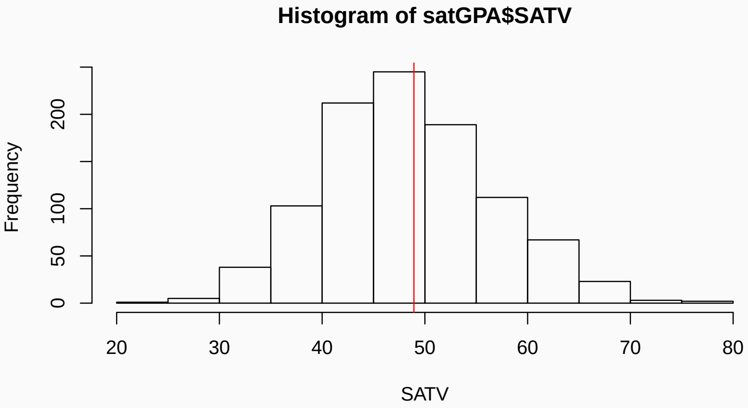
mean

To find the mean (or average) $\bar{x}$ of a set of observations, add their values and divide by the number of observations. If the n observations are $x_1, x_2, \ldots, x_n$, their mean is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

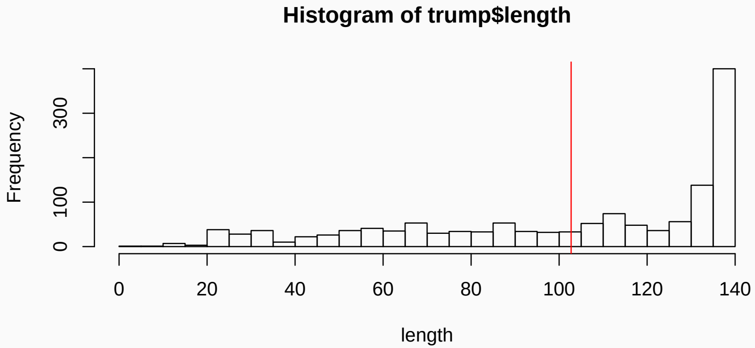
Mean makes sense here

```
xbar <- mean(satGPA$SATV)
hist(satGPA$SATV, breaks = 15, xlab = "SATV")
abline(v = xbar, col = "red")
```



But what about here?

```r
trump <- read.csv("../../data/trump.csv")
xbar <- mean(trump$length)
hist(trump$length, breaks = 30, xlab = "length")
abline(v = xbar, col = "red")
```



Why does this happen?

- The skew is pulling the mean to the right.
- This is because the mean can be interpreted as the "center of mass" of the distribution.
- The mean is not a "typical" value of of the length of a tweet.

The mean is not robust to extreme observations.

```r
mean(c(1, 2, 2, 3, 3))
```

```
[1] 2.2
```

```r
mean(c(1, 2, 2, 3, 10))
```

```
[1] 3.6
```

```r
mean(c(1, 2, 2, 3, 20))
```

```
[1] 5.6
```

```r
mean(c(1, 2, 2, 3, 100))
```

```
[1] 21.6
```

Another measure of center: The Median

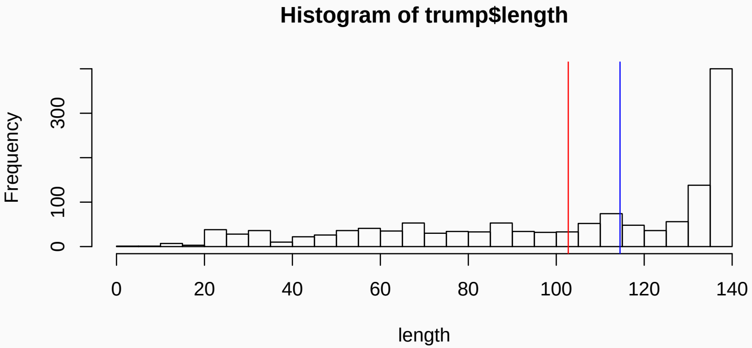
Median

The median is the midpoint of a distribution. Half of the observations are smaller than the median and the other half are larger than the median. Here is the rule for finding the median:

1. Arrange all of the observations in order of size, from smallest to largest.
2. If the number of observations n is odd, the median M is the center observation in the ordered list. Find the location of the median by counting $(n+1)/2$ observations up from the bottom of the list.
3. If the number of observations n is even, the median M is the mean of the two center observations in the ordered list.

Trump's Tweets

```
M <- median(trump$length)
hist(trump$length, breaks = 30, xlab = "length")
abline(v = xbar, col = "red")
abline(v = M, col = "blue")
```



The median is robust to extreme observations.

```
median(c(1, 2, 2, 3, 3))
```

```
[1] 2
```

```
median(c(1, 2, 2, 3, 10))
```

```
[1] 2
```

```
median(c(1, 2, 2, 3, 20))
```

```
[1] 2
```

```
median(c(1, 2, 2, 3, 100))
```

```
[1] 2
```

Exercise

Find the mean and median of the following numbers:

$$6, 3, 2, 3, 3, 7$$

Are centers enough to describe a distribution?

https://youtu.be/4B2xOvKFFz4

Measures of Spread

Deviations

- A measure of center is nice, but how do we describe variability of the points from the center?
- Idea: Use the deviations from a measure of center $(x_i - w)$.
- Can we use the average of the deviations from the mean $(w = \bar{x})$?

First proof

- **prove:** $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ for **any** sample.
- A proof is a "paragraph" of mathematical "sentences"
- Write "sentences" in order to make logical sense to the reader,
- Your proof is your personal argument as to why a claim must be true.
- **Requirement** (for completeness and clarity for the reader):
  Justify each step ("sentence") requiring statistics knowledge.
  Tell the reader what statistical concept you are using.
  ...like requiring you cite prior work you rely on in your writing
- Make liberal use of results already proven in the course.
  Just tell the reader what result you are using.

First proof: with a little too much detail

- **prove:** $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ for **any** sample.

Proof.

$$\sum(x_i - \bar{x}) = \sum x_i - \sum \bar{x} \text{ (associative property)}$$

$\square$

First proof: with a little too much detail

- **prove:** $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ for **any** sample.

Proof.

$$\begin{aligned} \sum(x_i - \bar{x}) &= \sum x_i - \sum \bar{x} \text{ (associative property)} \\ &= \sum x_i - n\bar{x} \text{ (summing up } n \text{ identical things)} \end{aligned}$$

$\square$

First proof: with a little too much detail

- **prove:** $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ for **any** sample.

Proof.

$$\begin{aligned}\sum(x_i - \bar{x}) &= \sum x_i - \sum \bar{x} \text{ (associative property)}\\ &= \sum x_i - n\bar{x} \text{ (summing up } n \text{ identical things)}\\ &= \sum x_i - n\frac{1}{n}\sum x_i \text{ (definition of } \bar{x}\text{)}\end{aligned}$$

□

First proof: with a little too much detail

- **prove:** $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ for **any** sample.

Proof.

$$\begin{aligned}\sum(x_i - \bar{x}) &= \sum x_i - \sum \bar{x} \text{ (associative property)}\\ &= \sum x_i - n\bar{x} \text{ (summing up } n \text{ identical things)}\\ &= \sum x_i - n\frac{1}{n}\sum x_i \text{ (definition of } \bar{x}\text{)}\\ &= \sum x_i - \sum x_i \text{ (}n\text{'s cancel)}\end{aligned}$$

$\square$

First proof: with a little too much detail

- **prove:** $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ for **any** sample.

Proof.

$$\begin{aligned}
\sum(x_i - \bar{x}) &= \sum x_i - \sum \bar{x} \text{ (associative property)} \\
&= \sum x_i - n\bar{x} \text{ (summing up } n \text{ identical things)} \\
&= \sum x_i - n\frac{1}{n}\sum x_i \text{ (definition of } \bar{x}) \\
&= \sum x_i - \sum x_i \text{ (} n\text{'s cancel)} \\
&= 0.
\end{aligned}$$

$\square$

Squared deviations

- Cool! We just made our first proof.
- But this means that the average deviation is not a good measure of spread:

$$\frac{1}{n}\sum(x_i - \bar{x}) = 0 \text{ for } \textbf{any} \text{ sample!}$$

- What about the average of the squared deviations?

What about the average of the squared deviations?

variance

The variance s^2 of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of n observations $x_1, x_2, \ldots, x_n$ is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

The standard deviation s is the square root of the variance s^2:

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

What about the average of the absolute deviations?

MAD

The median absolute deviation, or MAD, of a set of observations is the average of the absolute value of the deviations of the observations from their median. In symbols, the MAD of n observations $x_1, x_2, \ldots, x_n$ is

$$MAD = \frac{1}{n} \sum_{i=1}^{n} |x_i - M|,$$

where M is the median of $x_1, \ldots, x_n$.

What is so special about the median? i

Let

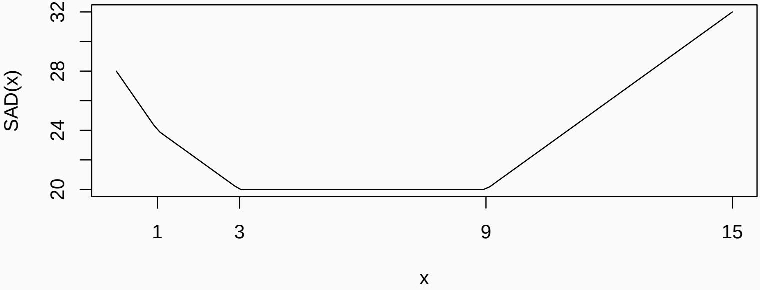
$$SAD(w) = \sum_{i=1}^{n} |x_i - w|$$

Consider the data $x_1 = 9, \;\; x_2 = 3, \;\; x_3 = 15, \;\; x_4 = 1$

What does the $SAD(w)$ function look like for these data?

```
SAD <- function(w) { sum( abs(x-w) ) }
```

What is so special about the median? ii

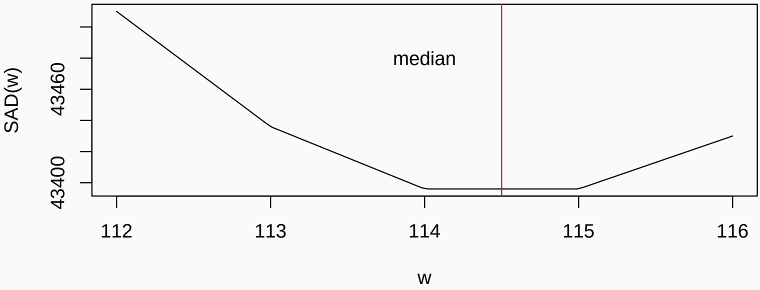


Where is the function $SAD(w)$ smallest (minimized)?

Trump's Twitter

OK back to looking at the data: Twitter length data from `trump`.

What does the $SAD(w)$ function look like for these data?



Where is the function $SAD(w)$ smallest (minimized)?

What's so special about the average?

Let $SSD(w) = \sum(x_i - w)^2$.

Consider again the data $x_1 = 9,\ \ x_2 = 3,\ \ x_3 = 15,\ \ x_4 = 1$ What is the $SSD(w)$ function for these data?

$$\sum_{i=1}^{4}(x_i - w)^2$$

What's so special about the average?

Let $SSD(w) = \sum(x_i - w)^2$.

Consider again the data $x_1 = 9, \;\; x_2 = 3, \;\; x_3 = 15, \;\; x_4 = 1$ What is the $SSD(w)$ function for these data?

$$\sum_{i=1}^{4}(x_i - w)^2 = (9 - w)^2 + (3 - w)^2 + (15 - w)^2 + (1 - w)^2$$

What's so special about the average?

Let $SSD(w) = \sum(x_i - w)^2$.

Consider again the data $x_1 = 9, \;\; x_2 = 3, \;\; x_3 = 15, \;\; x_4 = 1$ What is the $SSD(w)$ function for these data?

$$
\begin{aligned}
\sum_{i=1}^{4}(x_i - w)^2 &= (9 - w)^2 + (3 - w)^2 + (15 - w)^2 + (1 - w)^2 \\
&= (81 - 29w + w^2) + 9 - 23w + w^2) \\
&\quad + (225 - 215w + w^2) + (1 - 21w + w^2)
\end{aligned}
$$

What's so special about the average?

Let $SSD(w) = \sum (x_i - w)^2$.

Consider again the data $x_1 = 9, \;\; x_2 = 3, \;\; x_3 = 15, \;\; x_4 = 1$ What is the $SSD(w)$ function for these data?

$$\begin{aligned}
\sum_{i=1}^{4} (x_i - w)^2 &= (9 - w)^2 + (3 - w)^2 + (15 - w)^2 + (1 - w)^2 \\
&= (81 - 29w + w^2) + 9 - 23w + w^2) \\
&\quad + (225 - 215w + w^2) + (1 - 21w + w^2) \\
&= 4w^2 - 56w + 316
\end{aligned}$$

What's so special about the average?

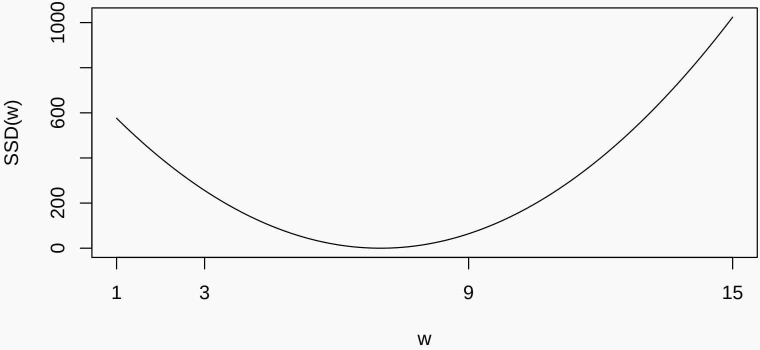
Let $SSD(w) = \sum (x_i - w)^2$.

Consider again the data $x_1 = 9, \;\; x_2 = 3, \;\; x_3 = 15, \;\; x_4 = 1$ What is the $SSD(w)$ function for these data?

$$
\begin{aligned}
\sum_{i=1}^{4} (x_i - w)^2 &= (9 - w)^2 + (3 - w)^2 + (15 - w)^2 + (1 - w)^2 \\
&= (81 - 29w + w^2) + 9 - 23w + w^2) \\
&\quad + (225 - 215w + w^2) + (1 - 21w + w^2) \\
&= 4w^2 - 56w + 316
\end{aligned}
$$

So, as ugly as $\displaystyle\sum_{i=1}^{n} (x_i - w)^2$ originally looks it's just a smooth quadratic function (convex).

What does the $SSD(w)$ function look like?



In this case $SSD(w) = 4w^2 - 56w + 316$

What value of w minimizes $SSD(w)$?

What value w minimizes $SSD(w) = 4w^2 - 56w + 316$?

$$\frac{d}{dw}SSD(w) = \frac{d}{dw}\left[4w^2 - 56w + 316\right]$$
$$= 2(4)w - 56 + 0 \;=\; 8w - 56$$

Set the derivative $= 0$ and solve for w.

$$8w - 56 = 0 \quad \Longrightarrow \quad w = \frac{56}{8} = 7$$

```
mean(x)

[1] 7
```

Check second derivative condition, etc...

General Data

What value of w minimizes $SSD(w)$ for any $x_1, x_2, \ldots, x_n$?

Minimize

$$f(w) = SSD(w) = \sum (x_i - w)^2$$

So

Second derivative $= 2n > 0$, so min (convex so global min).

General Data

What value of w minimizes $SSD(w)$ for any $x_1, x_2, \ldots, x_n$?

Minimize

$$\begin{aligned} f(w) = SSD(w) &= \sum (x_i - w)^2 \\ &= \sum (x_i^2 - 2wx_i + w^2) \end{aligned}$$

So

Second derivative $= 2n > 0$, so min (convex so global min).

General Data

What value of w minimizes $SSD(w)$ for any $x_1, x_2, \ldots, x_n$?

Minimize

$$\begin{aligned} f(w) = SSD(w) &= \sum (x_i - w)^2 \\ &= \sum (x_i^2 - 2wx_i + w^2) \\ &= \sum x_i^2 - 2w \sum x_i + \sum w^2 \end{aligned}$$

So

Second derivative $= 2n > 0$, so min (convex so global min).

General Data

What value of w minimizes $SSD(w)$ for any $x_1, x_2, \ldots, x_n$?

Minimize

$$\begin{aligned} f(w) = SSD(w) &= \sum (x_i - w)^2 \\ &= \sum (x_i^2 - 2wx_i + w^2) \\ &= \sum x_i^2 - 2w \sum x_i + \sum w^2 \\ &= \sum x_i^2 - 2w \sum x_i + nw^2 \end{aligned}$$

So

Second derivative $= 2n > 0$, so min (convex so global min).

General Data

What value of w minimizes $SSD(w)$ for any $x_1, x_2, \ldots, x_n$?

Minimize

$$
\begin{aligned}
f(w) = SSD(w) = \sum (x_i - w)^2 \\
&= \sum (x_i^2 - 2wx_i + w^2) \\
&= \sum x_i^2 - 2w \sum x_i + \sum w^2 \\
&= \sum x_i^2 - 2w \sum x_i + nw^2
\end{aligned}
$$

So

$$\frac{d}{dw} f(w) = -2 \sum x_i + 2nw \overset{set}{=} 0 \Rightarrow w = \frac{1}{n} \sum x_i = \bar{x}$$

Second derivative $= 2n > 0$, so min (convex so global min).

The point

- The mean minimizes the sum (and mean) of squared deviations.
- So the variance (and standard deviation) makes sense as a measure of spread from the mean.
- There are other (better) reasons to use the standard deviation as a measure of spread from the mean (more on this later).
- The median minimizes the sum (and mean) of absolute deviations.
- So the MAD makes sense as a measure of spread from the median.
- Caution: R's `mad()` function isn't quite the mean of absolute deviations. Multiplies this by a constant for theoretical reasons.

The standard deviation is not robust to extreme observations.

```r
sd(c(1, 2, 2, 3, 3))
```

```
[1] 0.8367
```

```r
sd(c(1, 2, 2, 3, 10))
```

```
[1] 3.647
```

```r
sd(c(1, 2, 2, 3, 20))
```

```
[1] 8.081
```

```r
sd(c(1, 2, 2, 3, 100))
```

```
[1] 43.83
```

The MAD is robust to extreme observations.

```
mad(c(1, 2, 2, 3, 3))
```

```
[1] 1.483
```

```
mad(c(1, 2, 2, 3, 10))
```

```
[1] 1.483
```

```
mad(c(1, 2, 2, 3, 20))
```

```
[1] 1.483
```

```
mad(c(1, 2, 2, 3, 100))
```

```
[1] 1.483
```

When to use each?

- Use the standard deviation for reasonably symmetric distributions without any extreme observations.
- Use the MAD as a robust version of SD (also for symmetric distributions), can accomodate a couple extreme observations.

Linear transformations

Linear Transformations

- Sometimes, we want to analyze data in different units.
- Temperature: Celsius $= \frac{5}{9}$(Fahrenheit - 32)
- Curve: exam $=$ score $+$ (0.25)(100 - score) (This curve adds back 25% of exam points missed).
- Standardized Score: $z_i = \frac{x_i - \bar{x}}{s}$.
- **Claim** All three are examples of linear transformations: $y = a + bx$.

Relationships (without proof)

- Let $y_i = a + bx_i$ for $i = 1, 2, \ldots, n$.
- **Claim**: $\bar{y} = a + b\bar{x}$.
- **Claim**: $\text{median}(y_1, \ldots, y_n) = a + b\ \text{median}(x_1, \ldots, x_n)$
- **Claim**: $\text{SD}(y) = |b|\text{SD}(x)$
- **Claim**: $\text{MAD}(y) = |b|\text{MAD}(x)$

Proof of first claim (with too much detail)

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \text{ (definition of } \bar{y}\text{)}$$

Proof of first claim (with too much detail)

$$\begin{aligned}\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \text{ (definition of } \bar{y}\text{)}\\ &= \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) \text{ (definition of } y_i\text{)}\end{aligned}$$

Proof of first claim (with too much detail)

$$
\begin{aligned}
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \text{ (definition of } \bar{y}\text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) \text{ (definition of } y_i\text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n} a + b\frac{1}{n}\sum_{i=1}^{n} x_i \text{ (associative property)}
\end{aligned}
$$

Proof of first claim (with too much detail)

$$
\begin{aligned}
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \text{ (definition of } \bar{y}\text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) \text{ (definition of } y_i\text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n} a + b\frac{1}{n}\sum_{i=1}^{n} x_i \text{ (associative property)} \\
&= \frac{1}{n}\sum_{i=1}^{n} a + b\bar{x} \text{ (definition of } \bar{x}\text{)}
\end{aligned}
$$

Proof of first claim (with too much detail)

$$
\begin{aligned}
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \text{ (definition of } \bar{y}\text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) \text{ (definition of } y_i\text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n} a + b\frac{1}{n}\sum_{i=1}^{n} x_i \text{ (associative property)} \\
&= \frac{1}{n}\sum_{i=1}^{n} a + b\bar{x} \text{ (definition of } \bar{x}\text{)} \\
&= \frac{1}{n} na + b\bar{x} \text{ (} n \text{ summations of } a\text{)}
\end{aligned}
$$

Proof of first claim (with too much detail)

$$
\begin{aligned}
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \text{ (definition of } \bar{y}\text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) \text{ (definition of } y_i\text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n} a + b\frac{1}{n}\sum_{i=1}^{n} x_i \text{ (associative property)} \\
&= \frac{1}{n}\sum_{i=1}^{n} a + b\bar{x} \text{ (definition of } \bar{x}\text{)} \\
&= \frac{1}{n} na + b\bar{x} \text{ (} n \text{ summations of } a\text{)} \\
&= a + b\bar{x} \text{ (} n\text{'s cancel)}
\end{aligned}
$$