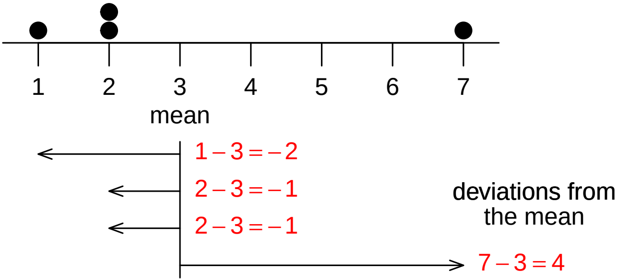
STAT 234 Lecture 15A
Standard Deviation & Sample Variance
(Section 1.4)

Yibi Huang
Department of Statistics
University of Chicago

Standard Deviation (SD) — Another Measure of Variability

To understand how *standard deviation (SD)* works, let's use a small data set $\{1, 2, 2, 7\}$ as an example.

- Each of these numbers deviates from the mean $\frac{1+2+2+7}{4} = 3$ by some amount:



Standard Deviation (Cont'd)

- How should we measure the overall size of these deviations?
- Taking their mean doesn't tell us anything about their magnitude
  - since $\sum_i (x_i - \bar{x}) = 0$
- One sensible way is take the average of their absolute values:

$$\frac{|-2| + |-1| + |-1| + |4|}{4} = 2$$

  This is called the mean absolute deviation (MAD), not the SD.
- But for a variety of reasons, statisticians prefer using the root-mean-square as a measure of overall size:

$$\sqrt{\frac{(-2)^2 + (-1)^2 + (-1)^2 + 4^2}{4}} \approx 2.35$$

  but this is still not the (sample) SD.

The formula for the (sample) *standard deviation (SD)* is

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

Why divide by $n-1$? Not n?

The formula for the (sample) *standard deviation (SD)* is

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

Why divide by $n - 1$? Not n?

- Short answer: One cannot measure variability with only ONE observation $(n = 1)$. We need at least 2.

The formula for the (sample) *standard deviation (SD)* is

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

Why divide by $n - 1$? Not n?

- Short answer: One cannot measure variability with only ONE observation $(n = 1)$. We need at least 2.
- Long answer: Dividing by n would underestimate the true (population) standard deviation. Dividing by $n - 1$ instead of n corrects some of that bias, which we'll prove shortly after

The formula for the (sample) *standard deviation (SD)* is

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

Why divide by $n - 1$? Not n?

- Short answer: One cannot measure variability with only ONE observation $(n = 1)$. We need at least 2.
- Long answer: Dividing by n would underestimate the true (population) standard deviation. Dividing by $n - 1$ instead of n corrects some of that bias, which we'll prove shortly after
- The standard deviation of $\{1, 2, 2, 7\}$ is

$$\sqrt{\frac{(-2)^2 + (-1)^2 + (-1)^2 + 4^2}{4-1}} \approx 2.71$$

(recall we get 2.35 when dividing by $n = 4$)

(Sample) Variance

The square of the (sample) standard deviation is called the *(sample) variance*, denoted as

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

which is roughly the average squared deviation from the mean.

- Note the **sample variance** for a variable in a data set is not the same as the **variance** for a random variable defined to be

$$\mathrm{Var}(X) = \mathrm{E}(X-\mu)^2 = \begin{cases} \sum_x (x-\mu)^2 p(x) & \text{if } X \text{ is discrete} \\ \int (x-\mu)^2 f(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

Shortcut Formula for Sample Variance

Just like the shortcut formula for the variance of a random variable,

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2, \quad \text{where } \mu = \mathrm{E}(X),$$

the sample variance also has its shortcut formula

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \frac{\left(\sum_{i=1}^{n} x_i^2\right) - n(\overline{x})^2}{n-1}.$$

Proof of Shortcut Formula for Sample Variance

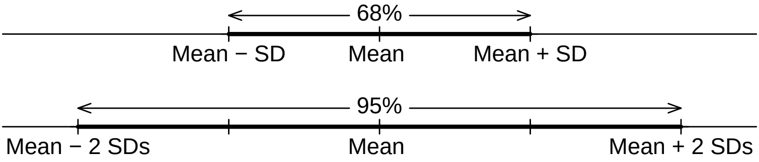
$$
\begin{aligned}
(n-1)s^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 &= \sum_{i=1}^{n}(x_i^2 - 2\overline{x}x_i + \overline{x}^2) \\
&= \left(\sum_{i=1}^{n} x_i^2\right) - 2\left(\sum_{i=1}^{n} x_i\overline{x}\right) + \underbrace{\sum_{i=1}^{n} \overline{x}^2}_{=n\overline{x}^2} \\
&= \left(\sum_{i=1}^{n} x_i^2\right) - 2\overline{x}\underbrace{\left(\sum_{i=1}^{n} x_i\right)}_{=n\overline{x}} + n\overline{x}^2 \\
&= \left(\sum_{i=1}^{n} x_i^2\right) - 2n\overline{x}^2 + n\overline{x}^2 \\
&= \left(\sum_{i=1}^{n} x_i^2\right) - n\overline{x}^2
\end{aligned}
$$

Now we have proved the shortcut formula for sample variance

$$
s^2 = \frac{\left(\sum_{i=1}^{n} x_i^2\right) - n\overline{x}^2}{n-1}.
$$

The 68% and 95% Rule of SD

- Roughly 68% of the observations will be within 1 SD away from the mean
- Roughly 95% will be with 2 SD away from the mean



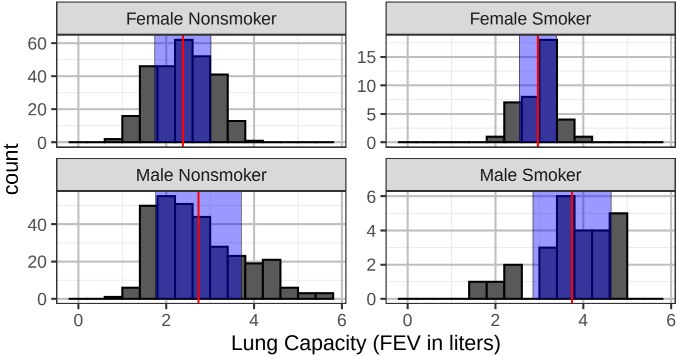
The value 68% and 95% comes from the standard normal distribution that

$$P(-1 < Z < 1) \approx 0.68 \quad \text{and} \quad P(-2 < Z < 2) \approx 0.95 \quad \text{if } Z \sim N(0, 1).$$

The 68% and 95% rule work very well for bell-shaped data, and reasonably well for unimodal and not seriously skewed data, but not for all data.

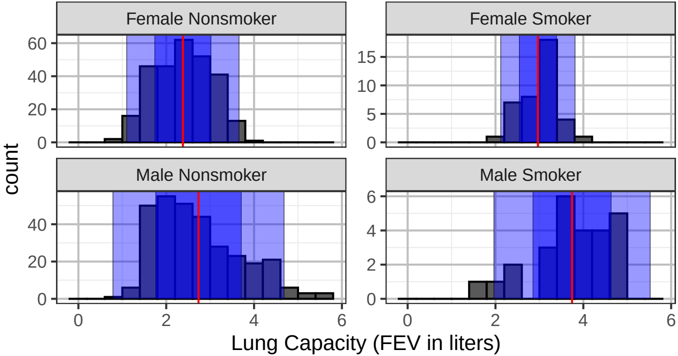
68% Rule for the FEV Data

Sex/Smoke	Mean	SD	Count	proportion within *1 SD* from mean
Female Nonsmoker	2.379	0.639	279	$170/279 \approx 60.9\%$
Female Smoker	2.966	0.423	39	$27/39 \approx 69.2\%$
Male Nonsmoker	2.734	0.974	310	$203/310 \approx 65.5\%$
Male Smoker	3.743	0.889	26	$17/26 \approx 65.4\%$



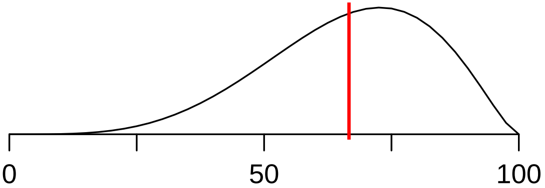
95% Rule for the FEV Data

Sex/Smoke	Mean	SD	Count	proportion within *2 SDs* from mean
Female Nonsmoker	2.379	0.639	279	$270/279 \approx 96.8\%$
Female Smoker	2.966	0.423	39	$38/39 \approx 97.4\%$
Male Nonsmoker	2.734	0.974	310	$298/310 \approx 96.1\%$
Male Smoker	3.743	0.889	26	$24/26 \approx 92.3\%$



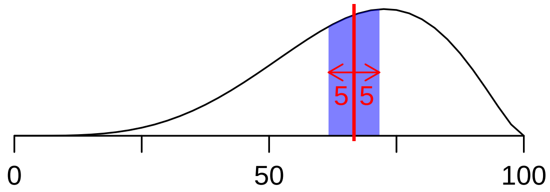
Exercise

Is the SD of the histogram below closest to 5, 15, or 50?



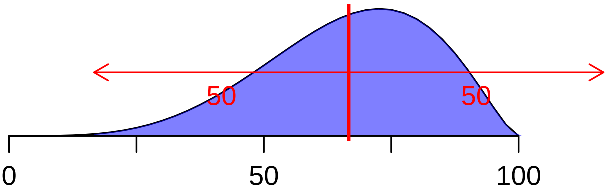
Exercise

Is the SD of the histogram below closest to 5, 15, or 50?



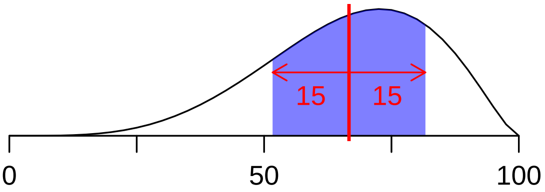
Exercise

Is the SD of the histogram below closest to 5, 15, or 50?



Exercise

Is the SD of the histogram below closest to 5, 15, or 50?



Properties of Standard Deviation (SD)

- When SD $= 0$, what do the data look like?
  - what if IQR $= 0$?

Properties of Standard Deviation (SD)

- When SD $= 0$, what do the data look like?
  - what if IQR $= 0$?
- and what if SD < 0?

Properties of Standard Deviation (SD)

- When SD $= 0$, what do the data look like?
    - what if IQR $= 0$?
- and what if SD < 0?
- SD is very sensitive to outliers.

Properties of Standard Deviation (SD)

- When SD $= 0$, what do the data look like?
  - what if IQR $= 0$?
- and what if SD < 0?
- SD is very sensitive to outliers.
- SD has the same units of measurement as the original observations, while the variances in the square of these units.

Linear Transformation and Sample SD/Variance

Sample SD/variance has the same scaling properties as the SD/variance of random variables.

If $y_i = ax_i + b$ for all $i = 1, 2, \ldots, n$, then

- $s_y^2 = a^2 s_x^2$
- $s_y = |a| s_x$

Recap: Graphical Summaries Numerical Variables

Histograms

- Shape (skewness, modality), outlier, center, spread

Box-plots

- Graphical display of the five-number summary + 1.5 IQR rule
- can reveal skewness but not modality

Recap: Numerical Summaries of Numerical Variables

Common measure of **center**:

- Mean
- Median

Common measure **spread**:

- Range: max – min
- SD
- IQR = $Q_3 - Q_1$

Center and **spread** are important summaries of a distribution. But they don't tell us about the modality and skewness of a distribution, and whether there are outliers.

Always check the histogram!