STAT021 Statistical Methods II

Lecture 4 Variance

Lu Chen
Swarthmore College
9/13/2018

Review: statistical modeling

- Statistical model

$$\begin{array}{ccccc} \text{Data} & = & \text{Model} & + & \text{Error} \\ Y & = & f(X) & + & \epsilon \end{array}$$

- Purposes of statistical modeling: *making predictions, understanding relationships, assessing differences.*
- Four-step process of statistical modeling
  - CHOOSE: *exploratory data analysis*
  - FIT: *estimating parameters*
  - ASSESS: *assessing model fitting and checking assumptions*
  - USE: *making predictions, understanding relationships, assessing differences, discussing limitations*

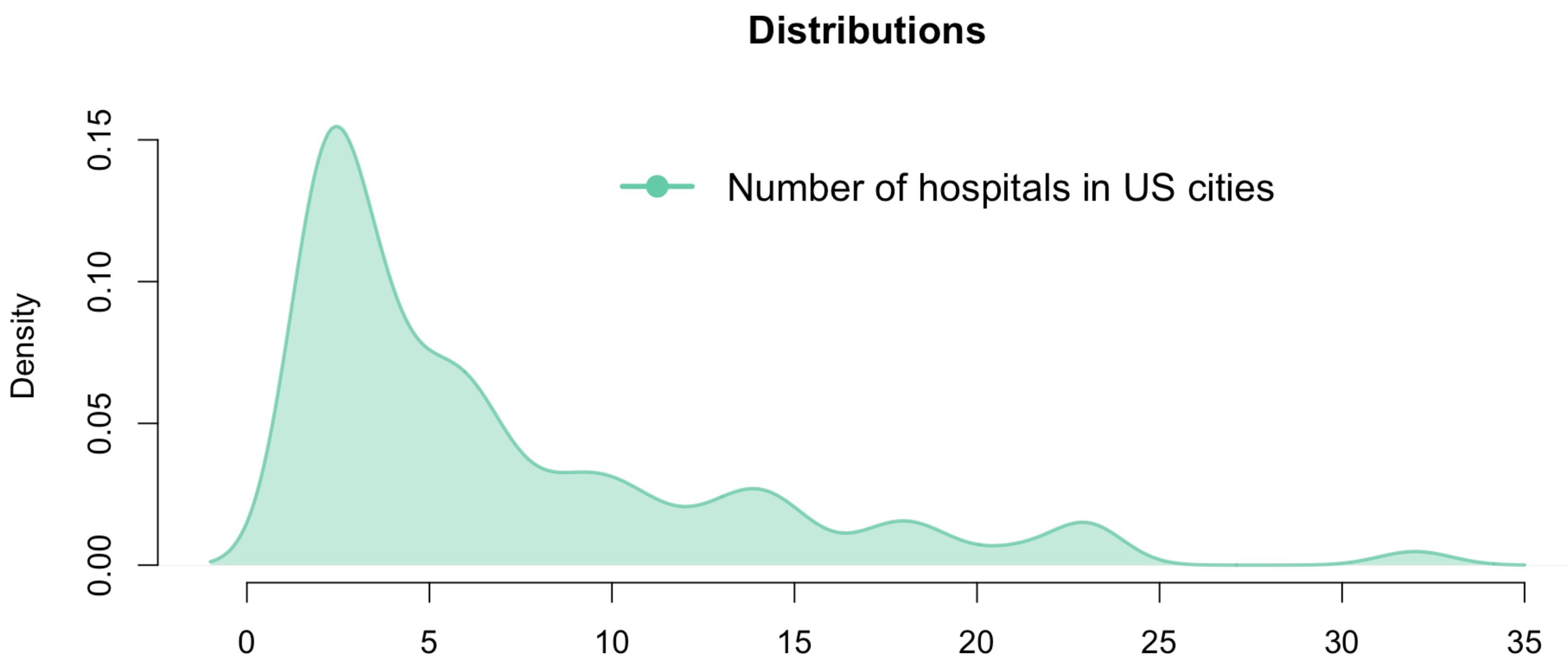
Outline

- Variability
- How to quantify variability
- Standard deviation
  - Sample standard deviation
  - Degree of freedom
- Variance
- Sampling variability of statistics
  - Definition
  - Standard error (SE)
  - Example
  - Sample size

Variability

> **Variability** is the extent to which a distribution is stretched or squeezed.

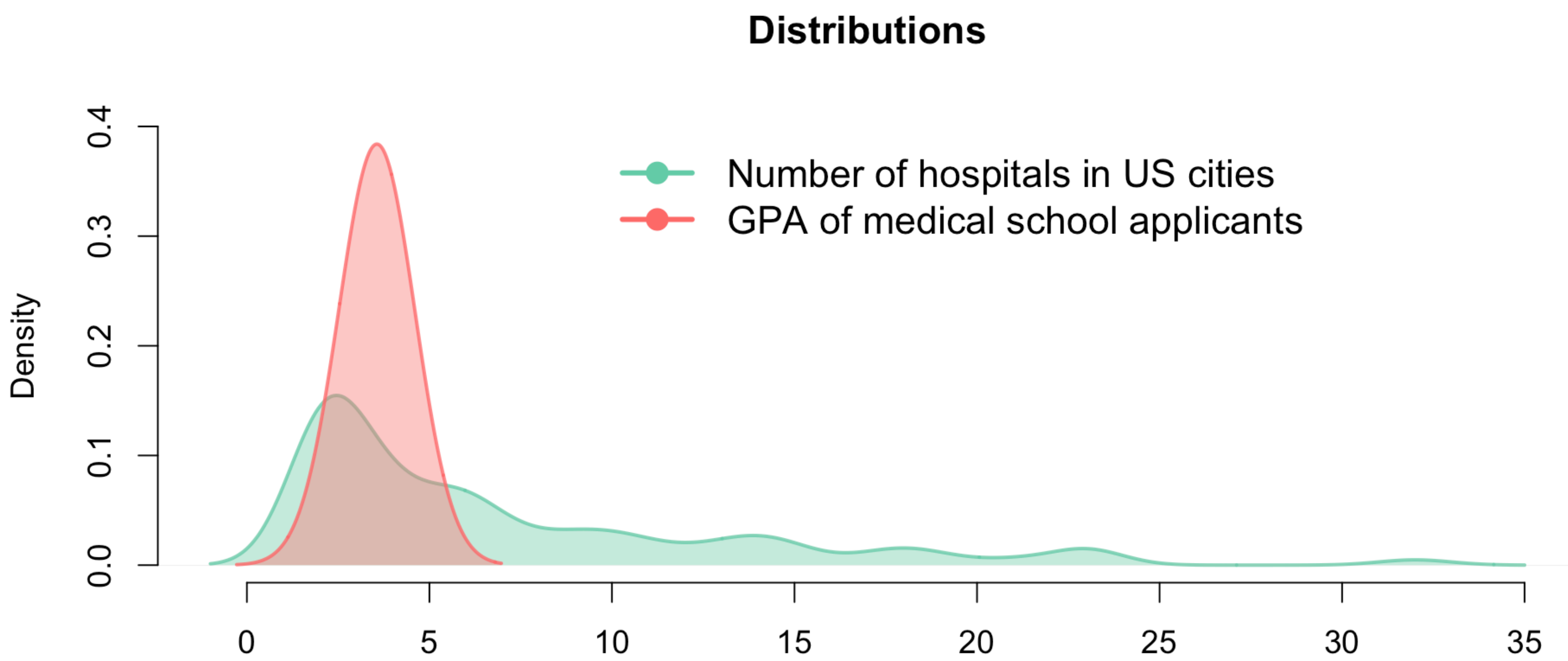
- All observed data have variability.



Variability

> **Variability** is the extent to which a distribution is stretched or squeezed.

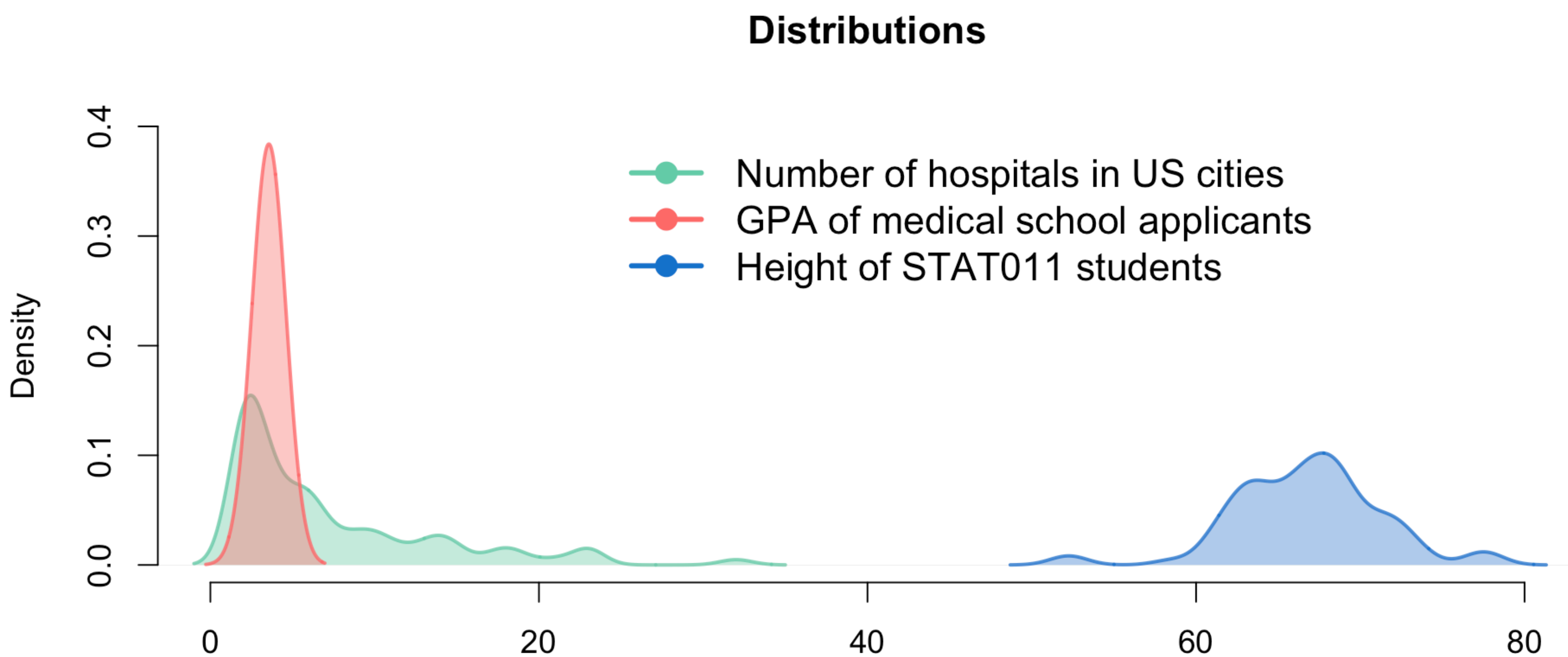
- All observed data have variability.



Variability

Variability is the extent to which a distribution is stretched or squeezed.

- All observed data have variability.



Variability

- Any exception?
- Gravitational constant
- Speed of light

Fundamental Physical Constants

Newtonian constant of gravitation G	
Value	$6.674\ 08 \times 10^{-11}\ \mathrm{m^3\ kg^{-1}\ s^{-2}}$
Standard uncertainty	$0.000\ 31 \times 10^{-11}\ \mathrm{m^3\ kg^{-1}\ s^{-2}}$
Relative standard uncertainty	4.7×10^{-5}
Concise form	$6.674\ 08(31) \times 10^{-11}\ \mathrm{m^3\ kg^{-1}\ s^{-2}}$

Fundamental Physical Constants

speed of light in vacuum c, c_0	
Value	$299\ 792\ 458\ \mathrm{m\ s^{-1}}$
Standard uncertainty	(exact)
Relative standard uncertainty	(exact)
Concise form	$299\ 792\ 458\ \mathrm{m\ s^{-1}}$

https://www.nist.gov/pml

Variability

$$
\begin{array}{ccccc}
\text{Data} & = & \text{Model} & + & \text{Error} \\
Y & = & f(X) & + & \epsilon
\end{array}
$$

- Sources of variability in data
  - Variability that can be explained by the model.
  - Variability that comes from the error term and cannot be explained by the model.
- Sources of error
  - Measurement error
    - Student A's first measurement of height is different from the second measurement
  - Random error
    - Student A's height is different from student B's height

How to quantify variability?

- Range
- Interquartile range ($Q_3 - Q_1$)
- Mean absolute difference

$$\frac{|y_1 - \mu| + |y_2 - \mu| + \cdots + |y_n - \mu|}{n}$$

  suppose μ is the population mean of $y_1, y_2, \cdots, y_n$.

- Standard deviation

$$\sqrt{\frac{(y_1 - \mu)^2 + (y_2 - \mu)^2 + \cdots + (y_n - \mu)^2}{n}}$$

Standard deviation

- Karl Pearson, 1894
- A measure that is used to quantify the amount of variation or dispersion of a set of data values.
- If $y_1, y_2, \cdots, y_n$ is a sample from a larger population, sample standard deviation is

$$s = \sqrt{\frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n-1}}$$

- If population mean μ is known, the standard deviation is

$$s = \sqrt{\frac{(y_1 - \mu)^2 + (y_2 - \mu)^2 + \cdots + (y_n - \mu)^2}{n}}$$

Standard deviation - simulation

- Let's use simulation to compare three possible ways of calculating sample standard deviation:

1. $$s_1 = \sqrt{\frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n-1}}$$

2. $$s_2 = \sqrt{\frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n}}$$

3. Suppose μ is known

$$s_3 = \sqrt{\frac{(y_1 - \mu)^2 + (y_2 - \mu)^2 + \cdots + (y_n - \mu)^2}{n}}$$

Standard deviation - simulation

```r
set.seed(10); n <- 25; s1 <- s2 <- s3 <- NULL
for(i in 1:1000){
  y <- rnorm(n) # mu = 0, sigma = 1
  s1[i] <- sd(y) # sqrt(sum((y-mean(y))^2)/(n-1))
  s2[i] <- sqrt(sum((y-mean(y))^2)/n)
  s3[i] <- sqrt(sum((y-0)^2)/n)
}
SD <- cbind(s1, s2, s3); head(SD)
```

```
##             s1        s2        s3
## [1,] 0.9424082 0.9233677 0.9779199
## [2,] 0.8023074 0.7860975 0.8648325
## [3,] 0.9687658 0.9491927 0.9497300
## [4,] 0.9938662 0.9737860 0.9882286
## [5,] 0.8745008 0.8568323 0.8688128
## [6,] 1.0652121 1.0436905 1.0480690
```

```r
colMeans(SD) # mean by columns
```

```
##        s1        s2        s3
## 0.9957356 0.9756177 0.9965334
```

Standard deviation - simulation

1. $$s_1 = \sqrt{\frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n-1}}$$

2. $$s_2 = \sqrt{\frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n}}$$

3. $$s_3 = \sqrt{\frac{(y_1 - \mu)^2 + (y_2 - \mu)^2 + \cdots + (y_n - \mu)^2}{n}}$$

- s_1 and s_3 are closer to the true population SD $\sigma = 1$ than s_2.
- s_1 is an unbiased estimator of the population SD σ.
- Sample SD s estimates the variability of the population but NOT the variability of the sample.
- $n - 1$ is the degree of freedom of s.

Standard deviation - degree of freedom

> **Degree of freedom** is the number of values in the final calculation of a statistic that are free to vary.

$$s_1 = \sqrt{\frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n - 1}}$$

$$s_3 = \sqrt{\frac{(y_1 - \mu)^2 + (y_2 - \mu)^2 + \cdots + (y_n - \mu)^2}{n}}$$

- For computing s_3, μ is known and fixed. $y_1, y_2, \cdots, y_n$ are free to vary.
- For computing s_1, $\bar{y}$ is calculated from the sample by $\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}$. The values of $y_1, y_2, \cdots, y_n$ are constrained by $\bar{y}$. Therefore, only $n - 1$ values of $y_1, y_2, \cdots, y_n$ are free to vary.
- The concept of degree of freedom will be used in ANOVA, SLR, MLR and logistic regression.

Variance - Ronald Fisher, 1918

- "The great body of available statistics show us that the deviations of a human measurement from its mean follow very closely the Normal Law of Errors, and, therefore, that the variability may be uniformly measured by the standard deviation corresponding to the square root of the mean square error."
- "When there are two independent causes of variability capable of producing in an otherwise uniform population distributions with standard deviations σ_1 and σ_2, it is found that the distribution, when both causes act together, has a standard deviation $\sqrt{\sigma_1^2 + \sigma_2^2}$."
- "It is therefore desirable in analysing the causes of variability to deal with the square of the standard deviation as the measure of variability. We shall term this quantity the **Variance**..."

Variance - Ronald Fisher, 1918

Summarizing what Fisher said:

- Variability can be measured by standard deviation
- If mean and SD of Y_1 are μ_1 and σ_1; mean and SD of Y_2 are μ_2 and σ_2; Y_1 and Y_2 are independent, and $Z = Y_1 + Y_2$, then

$$\sigma_Z = \sigma_{Y_1+Y_2} = \sqrt{\sigma_1^2 + \sigma_2^2}$$

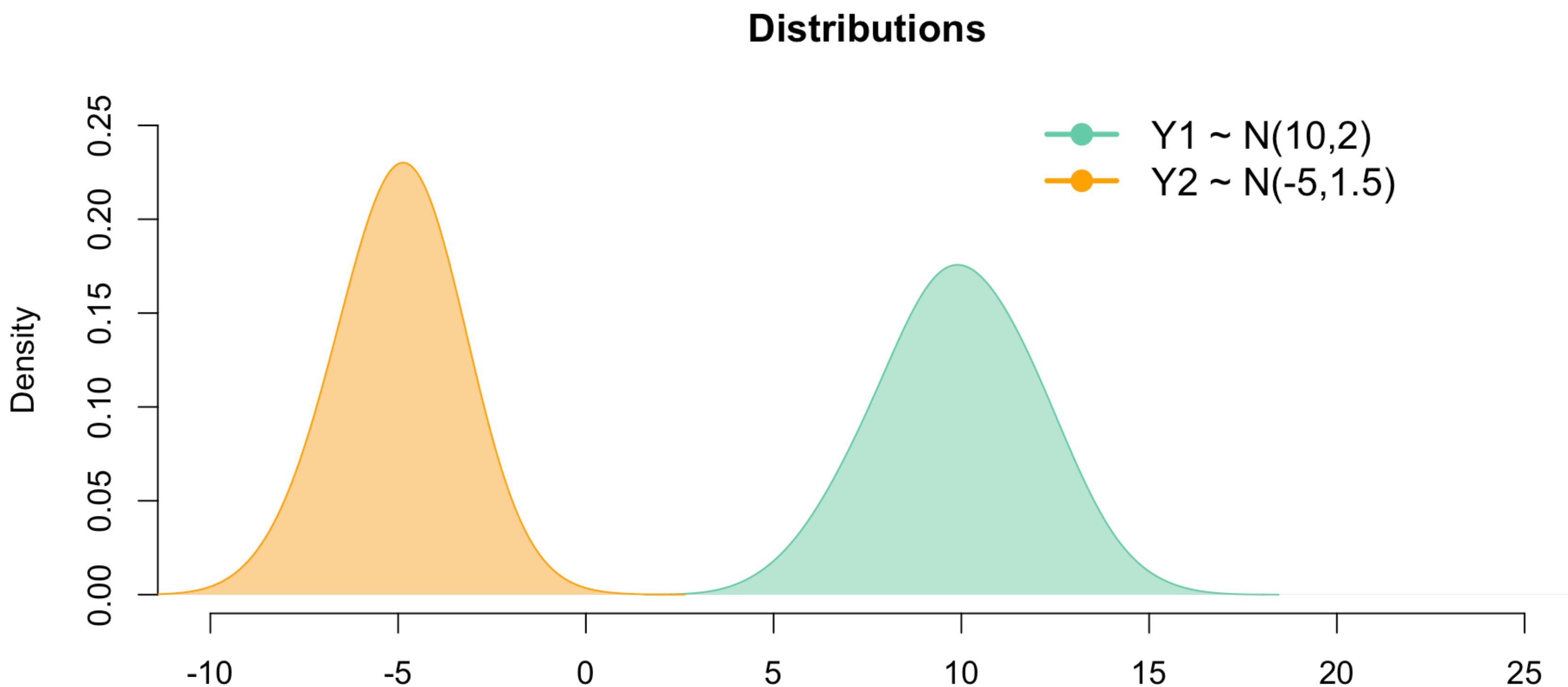
- Therefore

$$\begin{array}{ccccccc} \sigma_Z^2 & = & \sigma_{Y_1+Y_2}^2 & = & \sigma_1^2 & + & \sigma_2^2 \\ \mathrm{Var}(Z) & = & \mathrm{Var}(Y_1+Y_2) & = & \mathrm{Var}(Y_1) & + & \mathrm{Var}(Y_2) \end{array}$$

- This property of addition makes it easier to understand the sources of variability in data: variability of Z is the sum of the variability of Y_1 and Y_2.

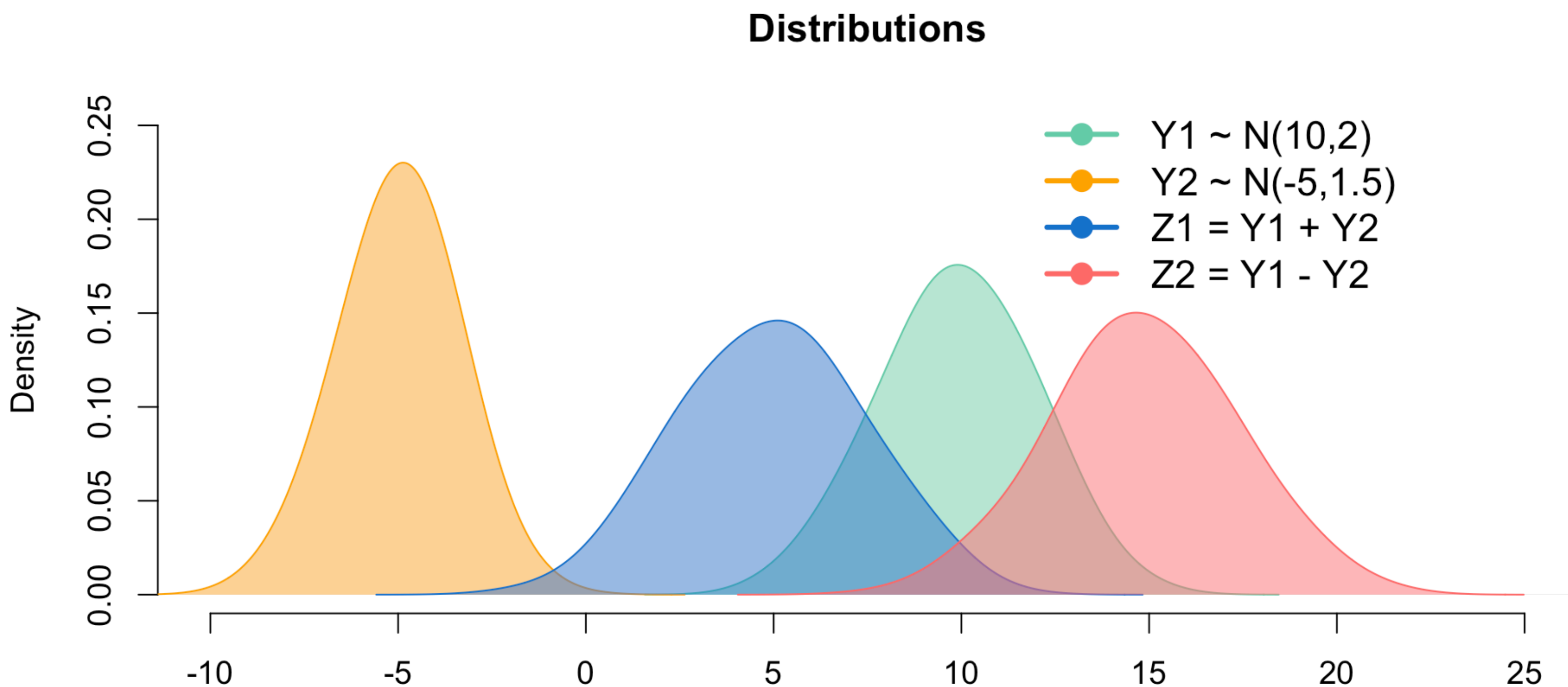
Variance - example

- $Y_1 \sim N(10, 2)$, $Y_2 \sim N(-5, 1.5)$.
- What is the distribution of $Z_1 = Y_1 + Y_2$ and $Z_2 = Y_1 - Y_2$?



Variance - example

- $Y_1 \sim N(10, 2)$, $Y_2 \sim N(-5, 1.5)$.
- $Z_1 = Y_1 + Y_2 \sim N(5, 2.5)$ and $Z_2 = Y_1 - Y_2 \sim N(15, 2.5)$



Variance

Generally, if Y_1 and Y_2 are independent,

$$\mathrm{Var}(aY_1 \pm bY_2) = a^2\mathrm{Var}(Y_1) + b^2\mathrm{Var}(Y_2)$$

- **Analysis of Variance (ANOVA)** was first published by Ronald Fisher in 1921.
- Wikipedia:
  "In the ANOVA setting, the observed variance in a particular variable is partitioned into components attributable to different sources of variation."
- "One of the attributes of ANOVA which ensured its early popularity was computational elegance. The structure of the additive model allows solution for the additive coefficients by simple algebra rather than by matrix calculations. In the era of mechanical calculators this simplicity was critical."

Sampling variability of statistics

- Y: GPA of medical school applicants.
- Suppose population mean is μ and population SD is σ.
- Sample: $y_1, y_2, \cdots, y_n$.

```
med.apply$GPA
```

```
##  [1] 3.62 3.84 3.23 3.69 3.38 3.72 3.89 3.34 3.71 3.89 3.97 3.49 3.77
## [14] 3.61 3.30 3.54 3.65 3.54 3.25 3.89 3.71 3.77 3.91 3.88 3.68 3.56
## [27] 3.44 3.58 3.40 3.82 3.62 3.09 3.89 3.70 3.24 3.86 3.54 3.40 3.87
## [40] 3.14 3.37 3.38 3.62 3.94 3.37 3.36 3.97 3.04 3.29 3.67 2.72 3.56
## [53] 3.48 2.80 3.44
```

- n = 55, $\bar{y} = 3.55$, $s_y = 0.29$
- $\bar{y}$ as the sample mean of GPA, do you think it has variability?

Sampling variability of statistics

n = 55, $\bar{y} = 3.55$, $s_y = 0.29$

- As an average of the 55 data points, it is a single value and thus would not change.
- However, as an estimate of the population mean GPA μ, since this one random sample, there is uncernty in this estimate. There IS variability in the statistic $\bar{y} = 3.55$. This is called **sampling variability**.

> **Sampling variability** is the variability of a statistic (calculated from a sample) as random sampling is repeated.

- This is in fact why we do statistical inference!

Sampling variability of statistics

Sampling variability is the key reason we do statistical inference.

- Among the 55 students, 30 were accepted by medical schools and their mean GPA was $\bar{y}_1 = 3.69$; 25 were rejected with mean GPA $\bar{y}_2 = 3.39$.
- Since there is variability in these estimates, we are uncerntain about whether $\bar{y}_1 = 3.69$ and $\bar{y}_2 = 3.39$ are different for sure or by chance.
- We need a statistical test: two-sample t test. It actually considers the variability of $\bar{y}_1$ and $\bar{y}_2$ to determine whether they are different for sure or by chance
- By CLT,

$$\bar{y} \overset{approx.}{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

SD of $\bar{y}$ is $\frac{\sigma}{\sqrt{n}}$.

Standard error (SE)

$$\text{SD}_{\bar{y}} \text{ or } \sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

- Since population SD σ is usually unknown, we use the sample SD s to estimate it.

> When the standard deviation of a statistic is estimated from the data, the result is called the **standard error (SE) of the statistic**.

$$\text{SE}_{\bar{y}} \text{ or } s_{\bar{y}} = \frac{s}{\sqrt{n}}$$

- Population SD: σ; sample SD: s.
- SD of sample mean $\bar{y}$: $\sigma/\sqrt{n}$; SE of sample mean $\bar{y}$: $s/\sqrt{n}$.

Sampling variability of statistics

Group	Size	Mean	SD
Accept	$n_1 = 30$	$\bar{y}_1 = 3.69$	$s_1 = 0.22$
Reject	$n_2 = 25$	$\bar{y}_2 = 3.39$	$s_2 = 0.27$

- We are interested in whether the two independent groups have the same mean. What is the variability of $\bar{y}_1 - \bar{y}_2$?
- By CLT,

$$\bar{y}_1 \overset{approx.}{\sim} N\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right) \text{ and } \bar{y}_2 \overset{approx.}{\sim} N\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right)$$

- $$\sigma_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{\sigma^2_{\bar{y}_1}}{n_1} + \frac{\sigma^2_{\bar{y}_2}}{n_2}}$$
- The **SE** of $\bar{y}_1 - \bar{y}_2$ is $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{0.22^2/30 + 0.27^2/25} = 0.067.$

Sampling variability of statistics

Group	Size	Mean	SD
Accept	$n_1 = 30$	$\bar{y}_1 = 3.69$	$s_1 = 0.22$
Reject	$n_2 = 25$	$\bar{y}_2 = 3.39$	$s_2 = 0.27$

The difference between the two groups is $\bar{y}_1 - \bar{y}_2 = 0.31$ with SE 0.067. Do you think the two groups have the same mean or not?

- Two sample t test. $H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$.
- $$\text{Test statistic } t = \frac{\bar{y}_1 - \bar{y}_2}{\mathrm{SE}_{\bar{y}_1 - \bar{y}_2}} = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{0.31}{0.067} = 4.6$$

The test statistic takes the ratio of the mean difference and the SE of the mean difference. If t is large, we reject H_0. If t is small, we cannot reject H_0.

Sampling variability of statistics

- Two sample t test. $H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$.
- $$t = \frac{\bar{y}_1 - \bar{y}_2}{\text{SE}_{\bar{y}_1 - \bar{y}_2}} = 4.6$$

  and $P = 3.2 \times 10^{-5} << 0.05$.
- We reject H_0 and conclude that the mean GPA is significantly different between the accepted and the rejected.

To determine whether there is difference in population means, the t test takes into account **two aspects** of the sample data:

1. The actual difference between sample means $\bar{y}_1 - \bar{y}_2$;
2. The variability of the difference between sample means $\text{SE}_{\bar{y}_1 - \bar{y}_2}$.

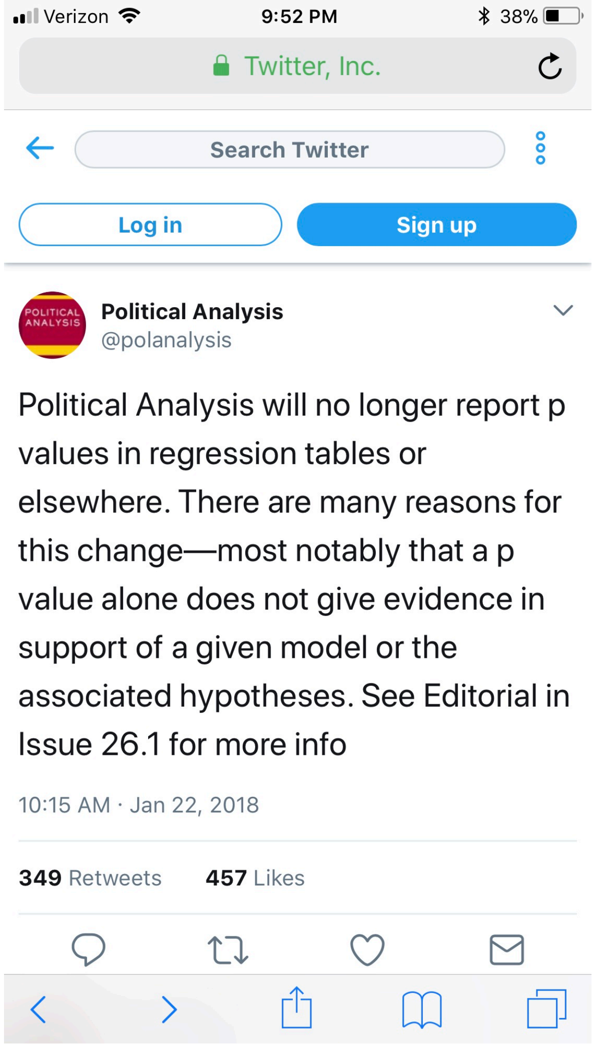
Sample size in the variability of statistics

Taking $SE_{\bar{y}} = \frac{s}{\sqrt{n}}$ or $SE_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ as an example, **variability of a statistic decreases** as

1. variability of the data decreases or
2. sample size increases.

- When sample size is very large, SE is tiny. Even if the effect size is small, we may still get large test statistic and significant P-value. Therefore, statistical significance does not always imply practical significance.
- When sample size is small, SE could be large. Even if the effect size is large, we may still get small test statistic and insignificant P-value. Therefore, statistical insignificance does not always imply practical insignificance.
- Statistical results should always be interpreted under the practical context.

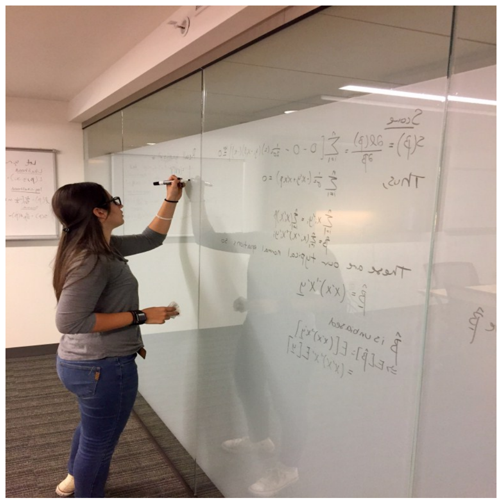
Statistical and practical significance



Summary

- Variability
- How to quantify variability
- Standard deviation (SD)
  - Sample standard deviation
  - Degree of freedom
- Variance
- Sampling variability of statistics
  - Definition
  - Standard error (SE)
  - Example
  - Sample size

βIOSTATISTICS OPEN HOUSE



Join the Graduate Group in Epidemiology and Biostatistics at the University of Pennsylvania for an engaging introduction to Penn Biostatistics. Meet our dynamic faculty and students and obtain information about our MS and PHD programs and research opportunities. Students of all levels are welcome. Students interested in applying for Fall 2019 enrollment are especially encouraged to attend.

What is Biostatistics?

bio·sta·tis·tics \ ˌbī-ō-stə-ˈtis-tiks \ : [1]statistical processes and methods applied to the collection, analysis, and interpretation of biological data and especially data relating to human biology, health, and medicine

[2] an exciting and impactful use of mathematical training, including mathematics, statistics, computer science, and engineering, to advance health

Date:

Friday, September 14, 2018

Time:

8:30a.m. to 3:00p.m.

Location:

Perelman School of Medicine,

University of Pennsylvania

Cost:

Free to attend, but advance registration is required

Contact:

Catherine Vallejo@upenn.edu

For more information or to register visit:
www.med.upenn.edu/ggeb/BiostatisticsOpenHouse2018_00.shmtl

Penn
University of Pennsylvania