香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

Introduction to Data Science

Lecture 12 Statistics

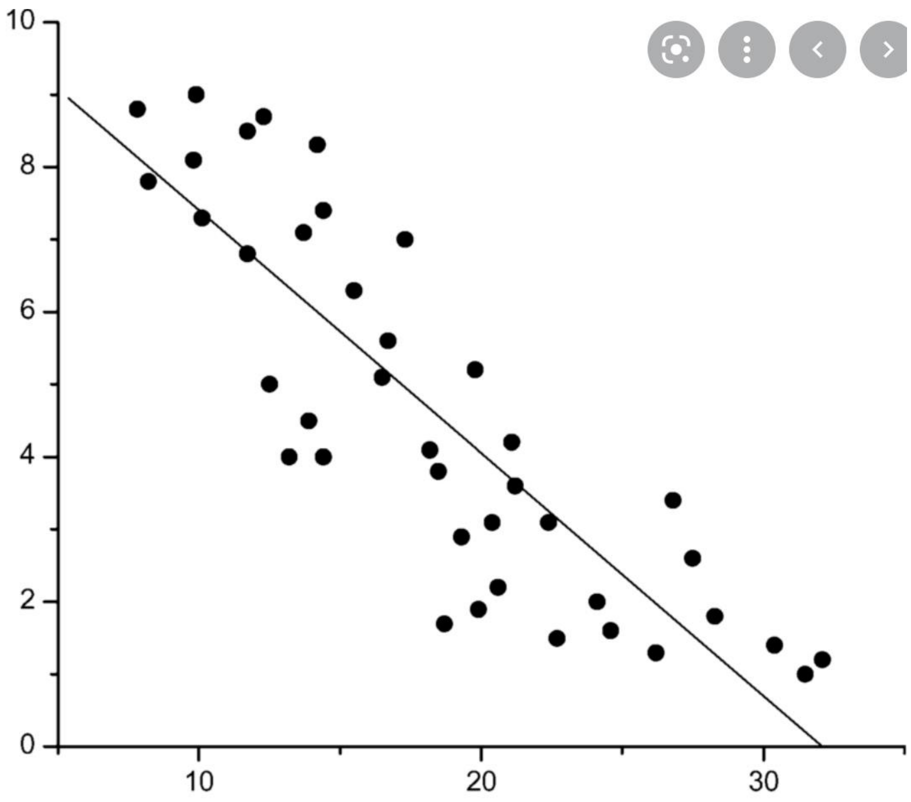
Advanced Concepts: Confidence Interval

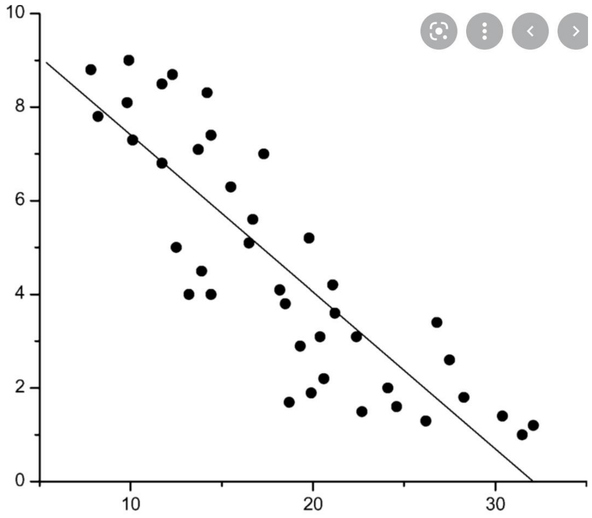
Zicheng Wang

Recap

Linear regression

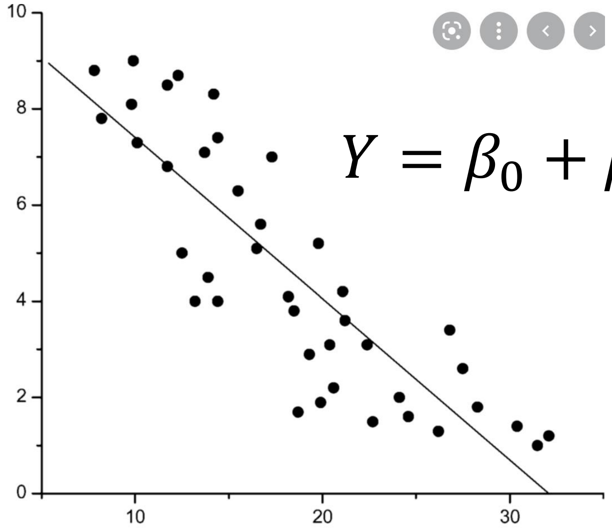
Find the relationship between X and Y





- Negative: larger x implies smaller y.
- **Question:** when x increases by a certain quantity, what’s the reduction in y?
- Use a line to approximate the relationship:
    - Regression analysis.

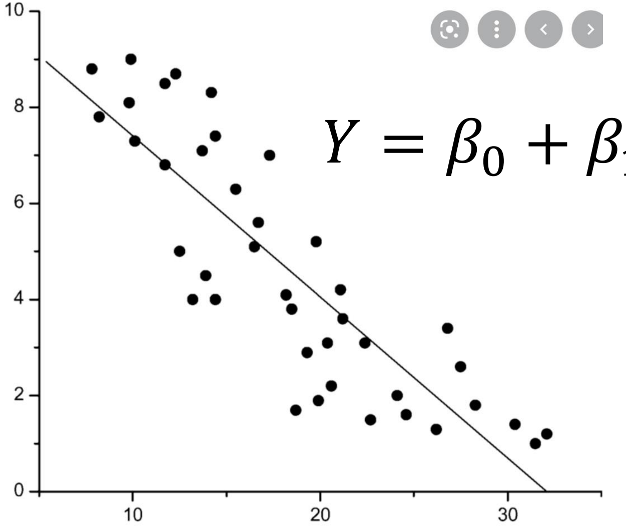
- Propose some models



$Y = \beta_0 + \beta_1 X$

- $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$
  - Given the observation of X
  - Y follows a normal distribution with mean $\beta_0 + \beta_1 X$, and variance σ^2
  - To simplify the analysis, we assume σ^2 is known

- Regression analysis: knowing β_0, β_1, σ, you can predict X given Y



$$Y = \beta_0 + \beta_1 X$$

- Propose some models

- $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$
  - Given the observation of X
  - Y follows a normal distribution with mean $\beta_0 + \beta_1 X$

- Regression analysis: knowing $\underbrace{\beta_0, \beta_1}_{\text{MLE: choose the best } \beta_0,\beta_1}, \sigma^2$, you can predict X given Y

- PDF for normal $$\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right].$$

- Samples: $(X_1, Y_1), \ldots, (X_N, Y_N)$

$$Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

- For the model with $\beta_0, \beta_1, \sigma^2$, the likelihood is

$$\frac{1}{(\sqrt{2\pi})^n \sigma^n}\exp\left[-\frac{1}{2}\frac{\Sigma_i (Y_i - \beta_1 X_i - \beta_0)^2}{\sigma^2}\right]$$

- Given σ^2, to maximize the likelihood, we only need to minimize

$$\boxed{\Sigma_i (Y_i - \beta_1 X_i - \beta_0)^2}$$

From High School:
Least square regression
最小二乘法

- Taking derivative over β_0 and β_1, we have

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0) = 0$$

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0) X_i = 0$$

- PDF for normal

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right].$$

- Samples: $(X_1, Y_1), \dots, (X_N, Y_N)$

Step 1

$$Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

- For the model with $\beta_0, \beta_1, \sigma^2$, the likelihood is

$$\frac{1}{(\sqrt{2\pi})^n \sigma^n} \exp\left[-\frac{1}{2}\frac{\Sigma_i (Y_i - \beta_1 X_i - \beta_0)^2}{\sigma^2}\right]$$

- Given σ^2, to maximize the likelihood, we only need to minimize

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0)^2$$

- Taking derivative over β_0 and β_1, we have

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0) = 0$$

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0) X_i = 0$$

Step 1

$$f_{X_i}(Y_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{Y_i-(\beta_0+\beta_1 X_i)}{\sigma}\right)^2\right]$$

$$L(\beta_0, \beta_1, \sigma^2) = f_{X_1}(Y_1) \times \cdots \times f_{X_N}(Y_N)$$
$$= \frac{1}{(\sqrt{2\pi})^n \sigma^n} \exp\left[-\frac{1}{2}\frac{\Sigma_i (Y_i-\beta_1 X_i-\beta_0)^2}{\sigma^2}\right]$$

- PDF for normal

$$\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right].$$

- Samples: $(X_1, Y_1), \ldots, (X_N, Y_N)$

$$Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

- For the model with $\beta_0, \beta_1, \sigma^2$, the likelihood is

$$\frac{1}{(\sqrt{2\pi})^n \sigma^n}\exp\left[-\frac{1}{2}\frac{\Sigma_i (Y_i - \beta_1 X_i - \beta_0)^2}{\sigma^2}\right]$$

Step 2

- Given σ^2, to maximize the likelihood, we only need to minimize

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0)^2$$

- Taking derivative over β_0 and β_1, we have

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0) = 0$$

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0) X_i = 0$$

- PDF for normal

$$\frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right].$$

- Samples: $(X_1, Y_1), \ldots, (X_N, Y_N)$

$$Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

- For the model with $\beta_0, \beta_1, \sigma^2$, the likelihood is

$$\underbrace{\boxed{\frac{1}{(\sqrt{2\pi})^n \sigma^n}}}_{\text{Constants}} \exp\left[-\boxed{\frac{1}{2}} \frac{\Sigma_i (Y_i - \beta_1 X_i - \beta_0)^2}{\boxed{\sigma^2}}\right]$$

- Given σ^2, to maximize the likelihood, we only need to minimize

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0)^2$$

- Taking derivative over β_0 and β_1, we have

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0) = 0$$

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0) X_i = 0$$

- PDF for normal

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right].$$

- Samples: $(X_1, Y_1), \ldots, (X_N, Y_N)$

$$Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

- For the model with $\beta_0, \beta_1, \sigma^2$, the likelihood is

Negative Sign

$$\frac{1}{(\sqrt{2\pi})^n \sigma^n} \exp\left[\boxed{-}\frac{1}{2}\frac{\Sigma_i (Y_i - \beta_1 X_i - \beta_0)^2}{\sigma^2}\right]$$

- Given σ^2, to maximize the likelihood, we only need to minimize

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0)^2$$

- Taking derivative over β_0 and β_1, we have

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0) = 0$$

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0) X_i = 0$$

- PDF for normal

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right].$$

- Samples: $(X_1, Y_1), \dots, (X_N, Y_N)$

$$Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

- For the model with $\beta_0, \beta_1, \sigma^2$, the likelihood is

$$\frac{1}{(\sqrt{2\pi})^n \sigma^n} \exp\left[-\frac{1}{2}\frac{\Sigma_i (Y_i - \beta_1 X_i - \beta_0)^2}{\sigma^2}\right]$$

- Given σ^2, to maximize the likelihood, we only need to minimize

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0)^2$$

- Taking derivative over β_0 and β_1, we have

Step 3

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0) = 0$$

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0) X_i = 0$$

- **First order condition**
- **Take derivative with respect to β_0 and β_1 separately**
- **Set the derivative to be equal to zero**

$$\Sigma_i(Y_i - \beta_1 X_i - \beta_0)X_i\text{=0 AND } \Sigma_i(Y_i - \beta_1 X_i - \beta_0)\text{=0}$$

Eliminate β_0 first:

$$\Sigma_i(Y_i - \beta_1 X_i - \beta_0) = 0 \;\; \text{->} \;\; \beta_0 = \frac{1}{N}\Sigma_i(Y_i - \beta_1 X_i) = \bar{Y} \text{-} \beta_1 \bar{X}$$

MLE:

$$\begin{cases} \widehat{\beta_1} = \dfrac{\Sigma_i(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma_i(X_i - \bar{X})^2} \\ \widehat{\beta_0} = \bar{Y} \text{-} \widehat{\beta_1}\bar{X} \end{cases}$$

When simple regression is invalid?

- The model we propose is not correct.

$$Y \sim N(\beta_0 + \beta_1 X, \sigma^2) \text{ or } Y - \beta_0 - \beta_1 X \sim N(0, \sigma^2)$$

Linear regression assumes that…

1. The relationship between X and Y is **linear**

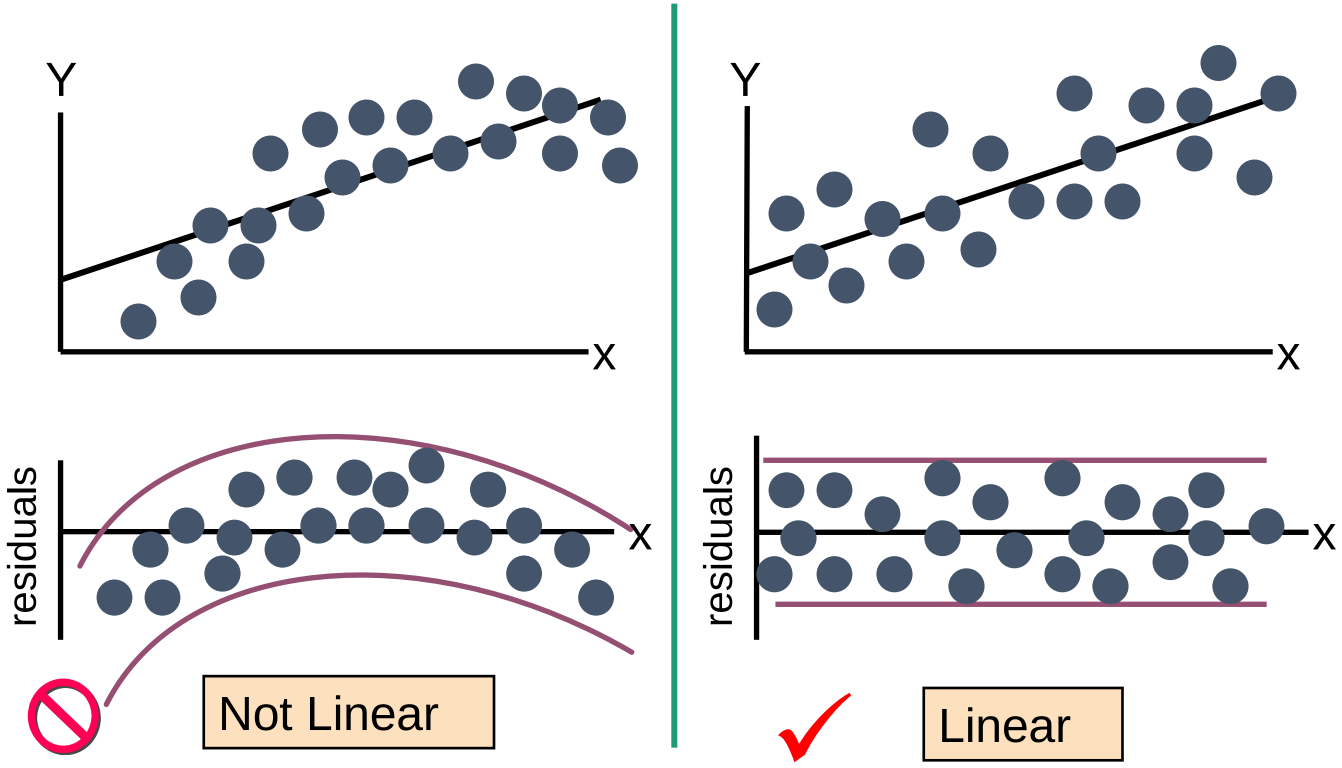
2. The variance of $Y - \beta_0 - \beta_1 X$ at every value of X is the **same** (homogeneity of variances)

Residual Analysis: check assumptions

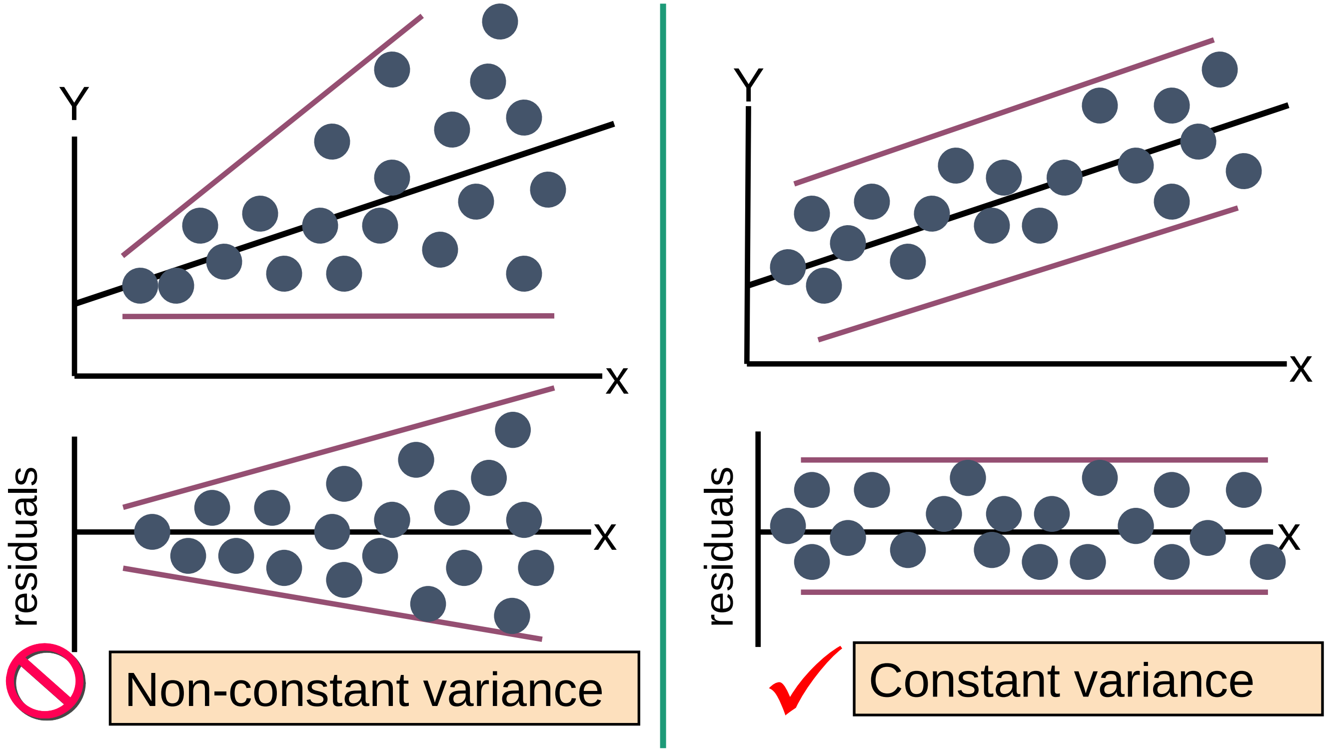
Residual: $e_i := Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i$

- Check the assumptions by examining the residuals
  - Examine for linearity assumption:
    - e_i *does not depend on* X_i
  - Evaluate constant-variance assumption:
    - variance of e_i *does not depend on* X_i
- Graphical Analysis of Residuals: Can plot residuals vs. X

Residual Analysis for Linearity



Residual Analysis for constant-variance



Advanced Concepts: Confidence Interval

Reading Materials

- Applied Statistics and Probability for Engineers, Third Edition, Douglas C. Montgomery and George C. Runger.
- Chap 8-2.1, …, 8-2.5

Experiments

Whether a drug can cure a disease: $\hat{p} = \frac{\Sigma_i X_i}{n}$ (MLE)

- Drug 1: $\hat{p}_1 = 90\%$.
- Drug 2: $\hat{p}_2 = 80\%$.

Which drug do you think is more effective?

Experiments

Whether a drug can cure a disease: $\hat{p} = \frac{\Sigma_i X_i}{n}$

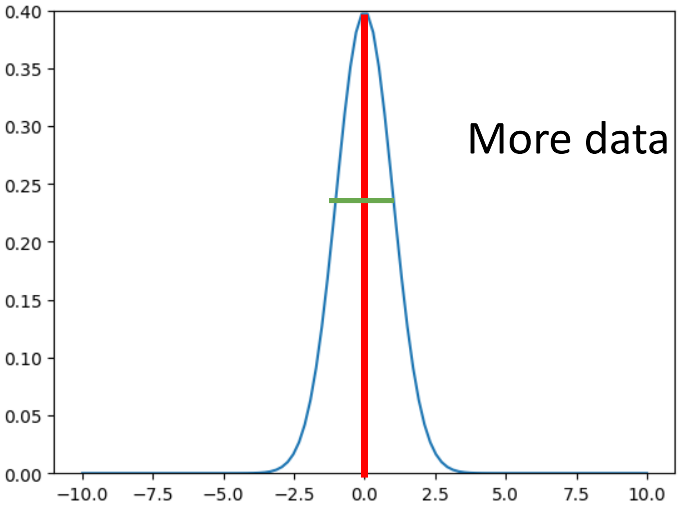
- Drug 1: $\hat{p}_1 = 90\%$. 10 experiments.
- Drug 2: $\hat{p}_2 = 80\%$. 10000 experiments.

Which drug do you think is more effective?

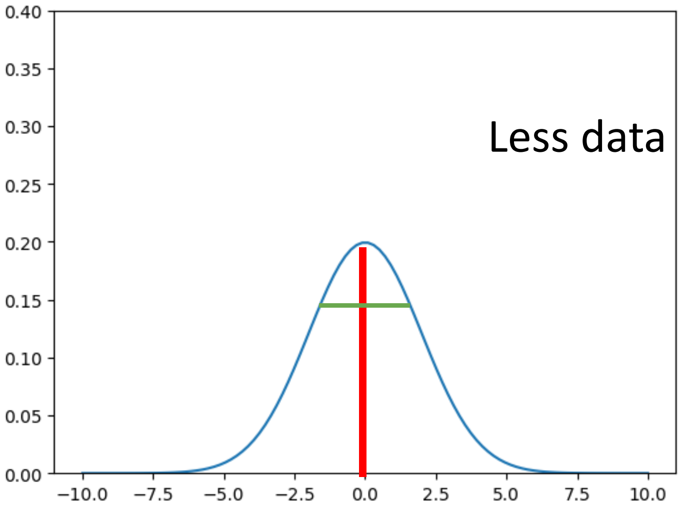
Which estimation is more reliable?

- **Number of samples can affect the accuracy!!!**

- With more data, we **believe** the estimator is closer to the true parameter.

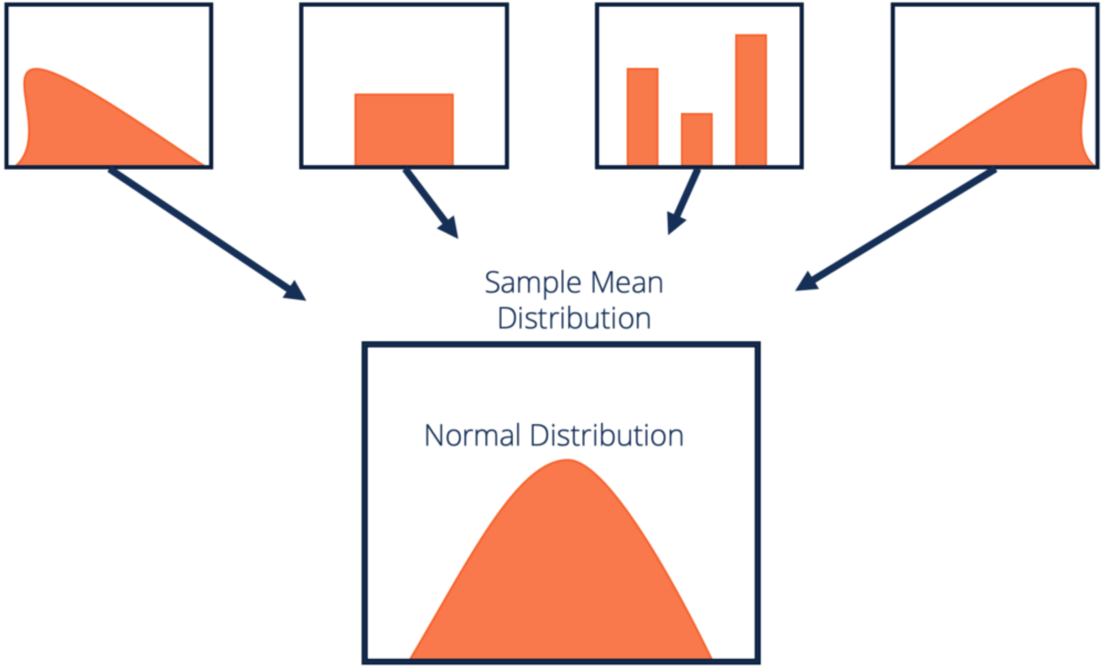


$\theta - \hat{\theta}$



$\theta - \hat{\theta}$

Central limit theorem



No matter what the true distribution is, the **sample mean** will be very close to the **normal distribution**, as long as the sample size is **large**.

Central limit theorem

$X_1, \dots, X_n$ can be non-normal **Mean: μ; Variance: σ^2**

Sample mean $$\bar{\boldsymbol{X}} = \frac{\boldsymbol{X_1} + \boldsymbol{X_2} + \cdots \boldsymbol{X_n}}{\boldsymbol{n}}$$

$$\boxed{\bar{X} \sim N\left(\boldsymbol{\mu}, \frac{\boldsymbol{\sigma}^2}{\boldsymbol{n}}\right)}$$

Or write as:

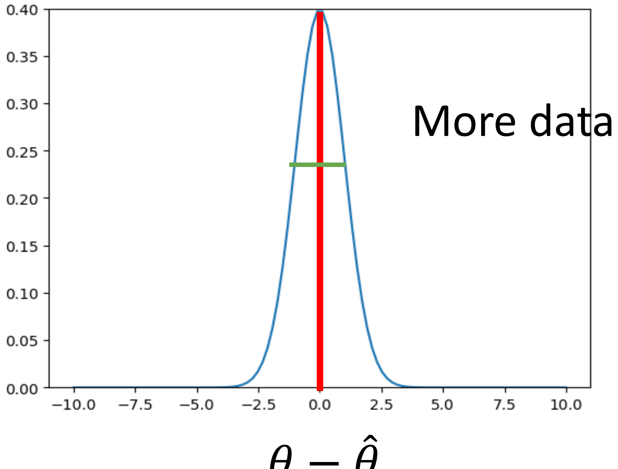
$$\boxed{\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0,1)}$$

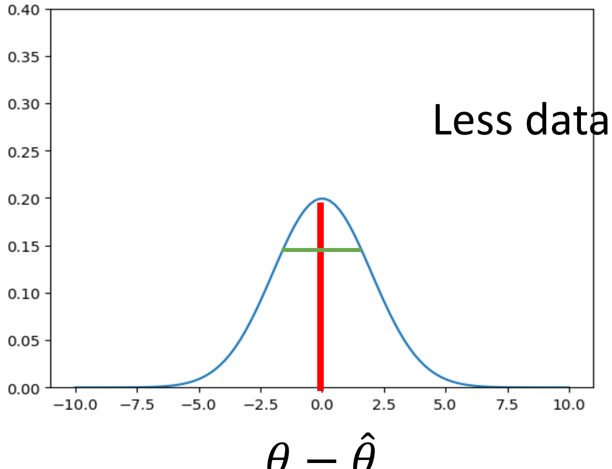
Standard Normal

Target

We will use normal distribution to show:

- with different size of data, **how close** the estimator is to the true parameter.
- With what probability, the true parameter falls in a region.





Interval Estimation – example 1

- We have data $X_1, X_2, \ldots, X_n$ that are sampled from some distribution with a **known** variance σ^2
- Their mean is $\boldsymbol{\mu}$, which we want to estimate
- We can easily give a point estimate: $\overline{\boldsymbol{X}}$ (sample mean)
- How to get an interval estimate??
  - Use Central Limit Theorem!

Interval Estimation – example 1

- $\frac{\sqrt{n}(\bar{X}-\mu)}{\boldsymbol{\sigma}} \sim \mathcal{N}\,(0,1)$
- P($a \leq \frac{\sqrt{n}(\bar{X}-\mu)}{\boldsymbol{\sigma}} \leq b$) = $\Phi(b) - \Phi(a)$
  - $\Phi(x)$: CDF of a standard normal distribution $\mathcal{N}\,(0,1)$.
- P($\bar{X} - \frac{b\,\boldsymbol{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{X} - \frac{a\,\boldsymbol{\sigma}}{\sqrt{n}}$) = $\Phi(b) - \Phi(a)$
- W.P. $\Phi(b) - \Phi(a)$, μ is within $[\bar{X} - \frac{b\,\boldsymbol{\sigma}}{\sqrt{n}}, \bar{X} - \frac{a\,\boldsymbol{\sigma}}{\sqrt{n}}]$

The best interval?

- W.P. $\Phi(b) - \Phi(a)$, μ is within $[\bar{X} - \frac{b\,\boldsymbol{\sigma}}{\sqrt{n}}, \bar{X} - \frac{a\,\boldsymbol{\sigma}}{\sqrt{n}}]$
- Fix $\Phi(b) - \Phi(a)$, there are too many **a** and **b** to choose from.
- At least, we want $\bar{X}$ to be within the interval
  - a < 0
  - b > 0

The best interval?

- W.P. $\Phi(b)-\Phi(a)$, μ is within $[\bar{X}-\frac{b\,\boldsymbol{\sigma}}{\sqrt{n}}, \bar{X}-\frac{a\,\boldsymbol{\sigma}}{\sqrt{n}}]$
- Fix $\Phi(b)-\Phi(a)$, there are too many **a** and **b** to choose from.
- μ has an **upper bound**, say U.
  - If $\bar{X}$ is too close to U, choose **a** such that $\bar{X}-\frac{a\,\boldsymbol{\sigma}}{\sqrt{n}}=\mathrm{U}$

The best interval?

- W.P. $\Phi(b) - \Phi(a)$, μ is within $[\bar{X} - \frac{b\,\boldsymbol{\sigma}}{\sqrt{n}}, \bar{X} - \frac{a\,\boldsymbol{\sigma}}{\sqrt{n}}]$
- Fix $\Phi(b) - \Phi(a)$, there are too many **a** and **b** to choose from.
- μ has a **lower bound**, say L.
    - If $\bar{X}$ is too close to L, choose **b** such that $\bar{X} - \frac{b\,\boldsymbol{\sigma}}{\sqrt{n}} = \mathrm{L}$.

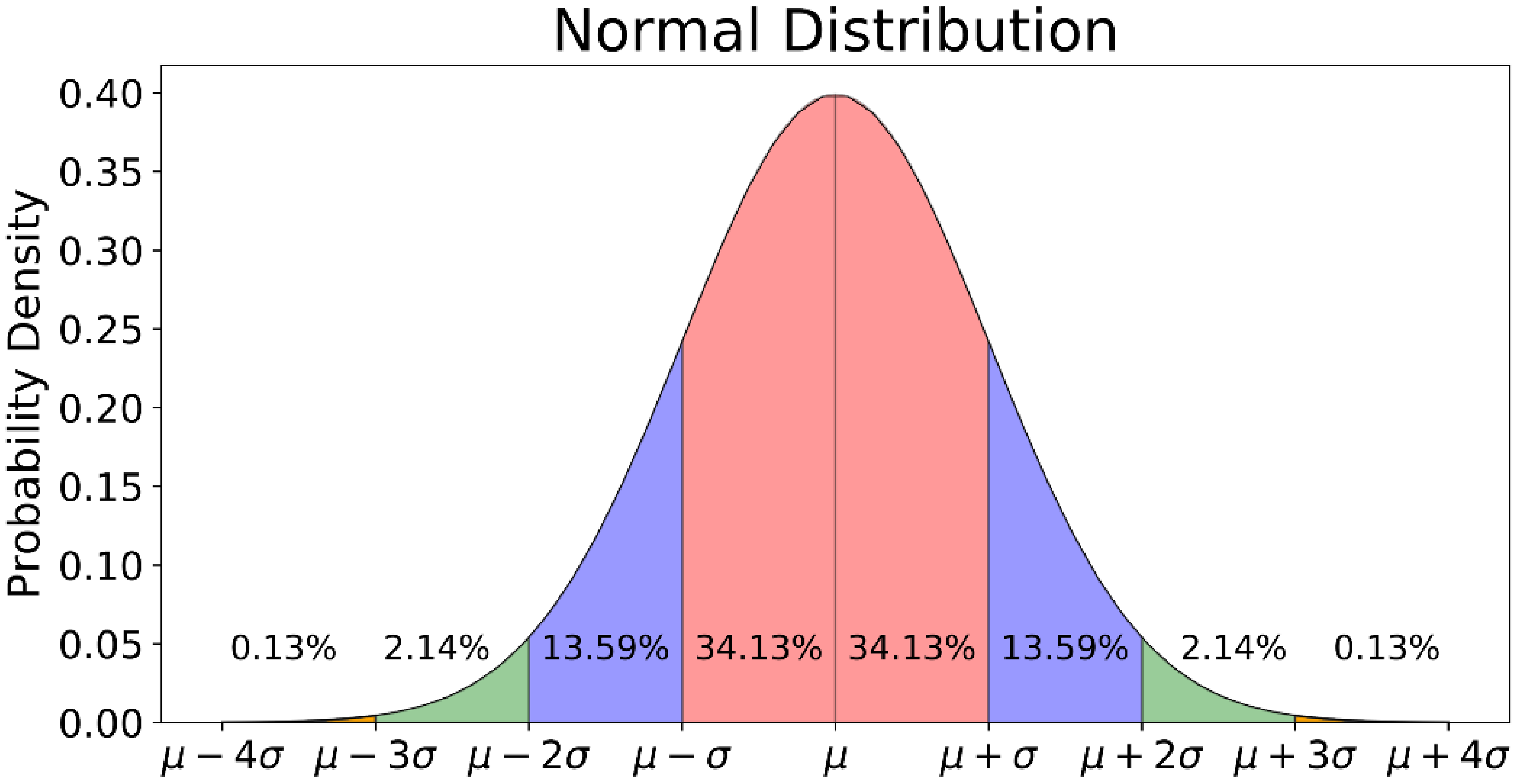
The best interval?

- W.P. $\Phi(b) - \Phi(a)$, μ is within $[\bar{X} - \frac{b\,\boldsymbol{\sigma}}{\sqrt{n}}, \bar{X} - \frac{a\,\boldsymbol{\sigma}}{\sqrt{n}}]$
- Fix $\Phi(b) - \Phi(a)$, there are too many **a** and **b** to choose from.
- μ has no **bound**.
  - Choose a and b such that b-a is minimized.

- W.P. $\Phi(b) - \Phi(a)$, μ is within $[\bar{X} - \frac{b\,\boldsymbol{\sigma}}{\sqrt{n}}, \bar{X} - \frac{a\,\boldsymbol{\sigma}}{\sqrt{n}}]$

- For **the 1- α confidence interval**, we **need**
  - $\Phi(b) - \Phi(a)$ = **1- $\boldsymbol{\alpha}$**
  - b – a is minimized.

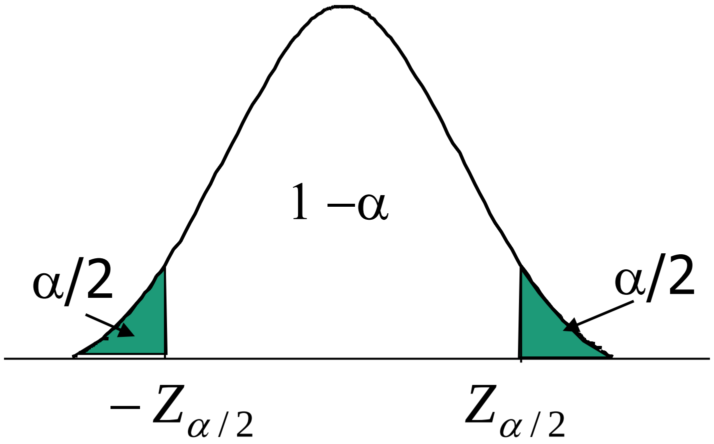
- As pdf of normal is symmetric and has a single peak, for **the 1- α confidence interval**,
  - b + a = 0.



Normal Distribution
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.13%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
0.13%
$\mu - 4\sigma$
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$
$\mu + 4\sigma$

Notation

$\boldsymbol{N(0,1)}$: **Standard Normal Distribution**



let $\boldsymbol{z_{\alpha/2}}$ be the number such that the area under the **standard normal density function** to the right of $z_{\alpha/2}$ is $\boldsymbol{\alpha/2}$.

Then if Z~N(0,1)

$$\boldsymbol{P\left(-z_{\alpha/2} \le Z \le z_{\alpha/2}\right) = 1 - \alpha}$$

Interval Estimation – example 1

- W.P. $\Phi(b) - \Phi(a)$, μ is within $[\bar{X} - \frac{b\,\boldsymbol{\sigma}}{\sqrt{n}}, \bar{X} - \frac{a\,\boldsymbol{\sigma}}{\sqrt{n}}]$
- $\boldsymbol{P\left(-z_{\alpha/2} \le Z \le z_{\alpha/2}\right) = 1 - \alpha}$
- Let b = $\boldsymbol{z_{\alpha/2}}$, a = - $\boldsymbol{z_{\alpha/2}}$. Then $\Phi(b) - \Phi(a)$ = 1- α.
- **The 1- α confidence interval** for μ : $\boldsymbol{[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]}$

Some observations

1-α Confidence Interval (CI):

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

We typically call $1 - \alpha$ as the **confidence level**.

The length of the confidence interval is affected by several factors

- As the sample size n **increases**, the length of CI **decreases**
- As the variance σ^2 **increases**, the length of CI **increases**
- As the confidence level increases (α **decreases**), the length of CI **increases**.

Interval Estimation - example

- n patients use the new drug, whether the drug can cure the disease is a Bernoulli RV
- We have data $X_1, X_2, \dots, X_n$ that are sampled from this Bernoulli distribution with unknown cure rate p to be estimated
- Clearly, the mean of Bernoulli(p) is $\boldsymbol{p}$
- We can easily give a point estimate: $\hat{\boldsymbol{p}} = \overline{\boldsymbol{X}}$ (sample cure rate)
- How to get an interval estimate??
  - Use Central Limit Theorem!

Interval Estimation - example

$$Z = \frac{\sqrt{n}(\hat{p} - p)}{\sqrt{p(1-p)}} \sim N(0,1)$$

Mean: p
Variance: $p(1-p)$
$\hat{p} = \bar{X}$ (sample cure rate)

- 1 - α confidence interval:

$$[\hat{\boldsymbol{p}} - \boldsymbol{z}_{\alpha/2}\sqrt{\frac{p(1-p)}{n}},\ \hat{\boldsymbol{p}} + \boldsymbol{z}_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}]$$

- As p is unknown, replace p above by $\hat{\boldsymbol{p}}$, 1 - α confidence interval:

$$[\hat{\boldsymbol{p}} - \boldsymbol{z}_{\alpha/2}\sqrt{\frac{\hat{\boldsymbol{p}}(1-\hat{\boldsymbol{p}})}{n}},\ \hat{\boldsymbol{p}} + \boldsymbol{z}_{\alpha/2}\sqrt{\frac{\hat{\boldsymbol{p}}(1-\hat{\boldsymbol{p}})}{n}}]$$

Experiments

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Whether a drug can cure a disease: $\hat{p} = \frac{\Sigma_i X_i}{n}$

95% Confidence Interval

- Drug 1: $\hat{p}_1 = 90\%$. 10 experiments. [71.41%, 100%]
- Drug 2: $\hat{p}_2 = 80\%$. 10000 experiments. [79.22%, 80.78%]

Confidence Statements

- Fortune Teller

“I believe the cure rate is 80%”

point

- Scientist

“I believe the cure rate is 80% plus or minus 5%”

interval