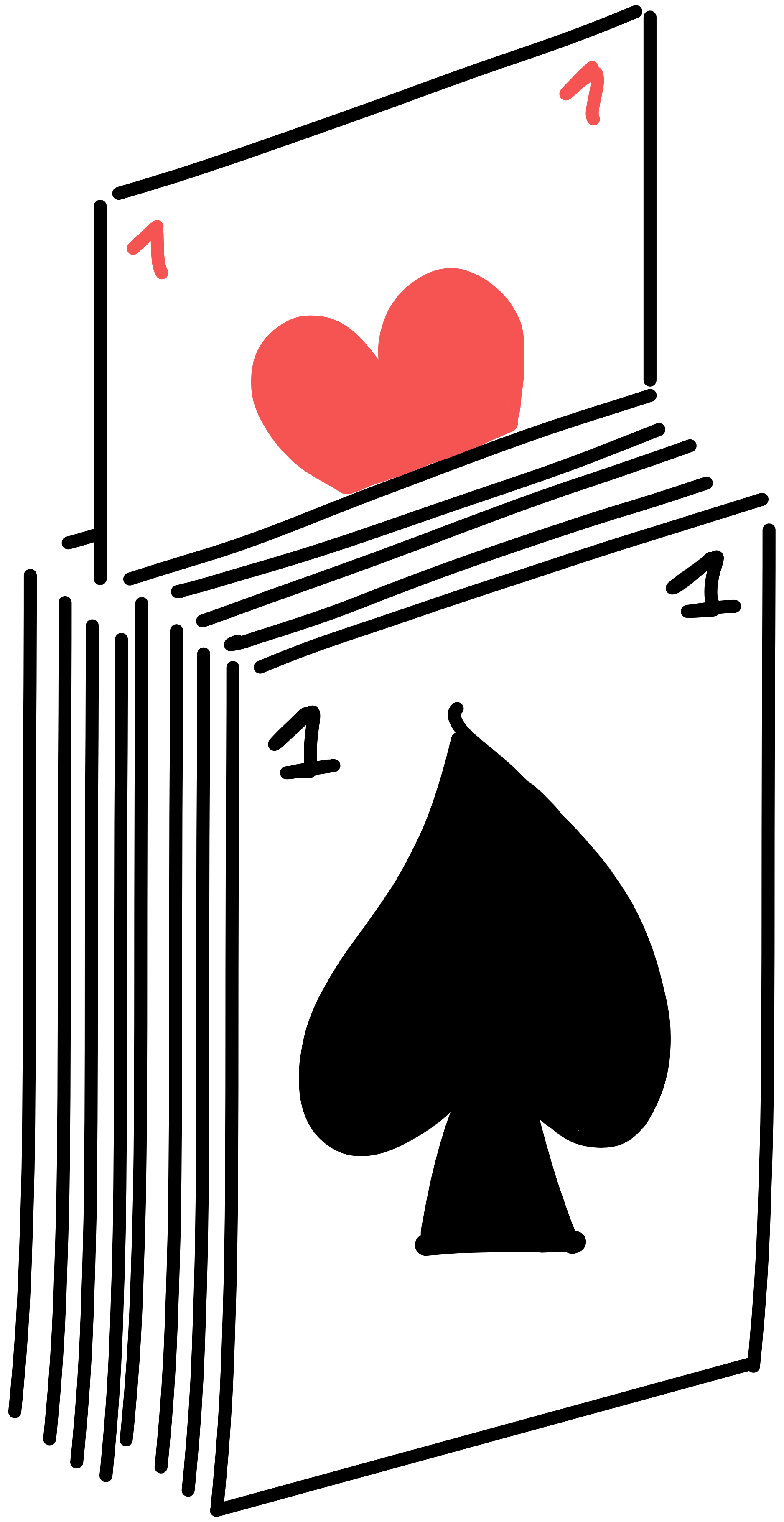
PROBABILITY & STATISTICS

TPP Math Review

Manon Revel
mrevel@mit.edu



PROBABILITY

- Probability Space
- Conditional Probability
- Random Variables
- Expectation and Variance
- Gaussian Distribution

PROBABILITY SPACE

SAMPLE SPACE, SIGMA-FIELD AND PROBABILITY MEASURE

$$(\Omega, \mathscr{F}, \mathbb{P})$$

Sample space

↳ possible outcomes

σ-field

↳ collection of events

Probability measure

↳ measures events' mass

A COIN TOSS

$$\Omega = \{H, T\} \leftarrow \text{two possible outcomes}$$

$$\mathcal{F} = \{\emptyset, \{H, T\}, \{H\}, \{T\}\}$$

$$\mathbb{P}_1 = \mathcal{U}(\{H, T\}) \qquad \mathbb{P}_2(\{H\}) = \frac{3}{4}$$

fair coin

biased coin

A COIN TOSS

$$(\Omega, \mathcal{F}, \mathbb{P}_2)$$

$$\mathbb{P}_2(\emptyset) = 0$$

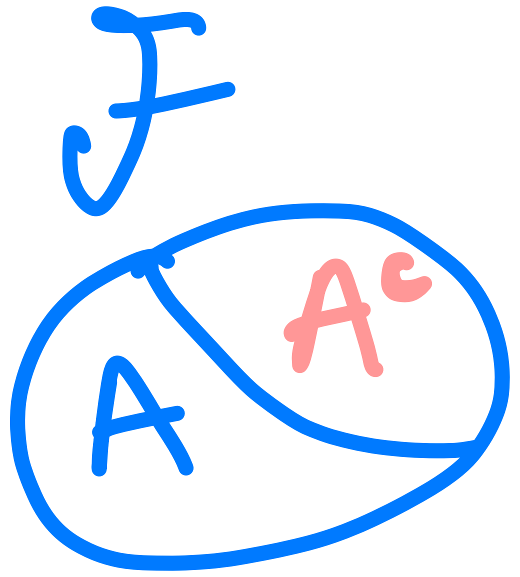
$$\mathbb{P}_2(\{T\}) = 1 - \frac{3}{4} = \frac{1}{4}$$

$$\mathbb{P}_2(\{H, T\}) = \mathbb{P}(\{H\}) + \mathbb{P}(\{T\}) = 1$$

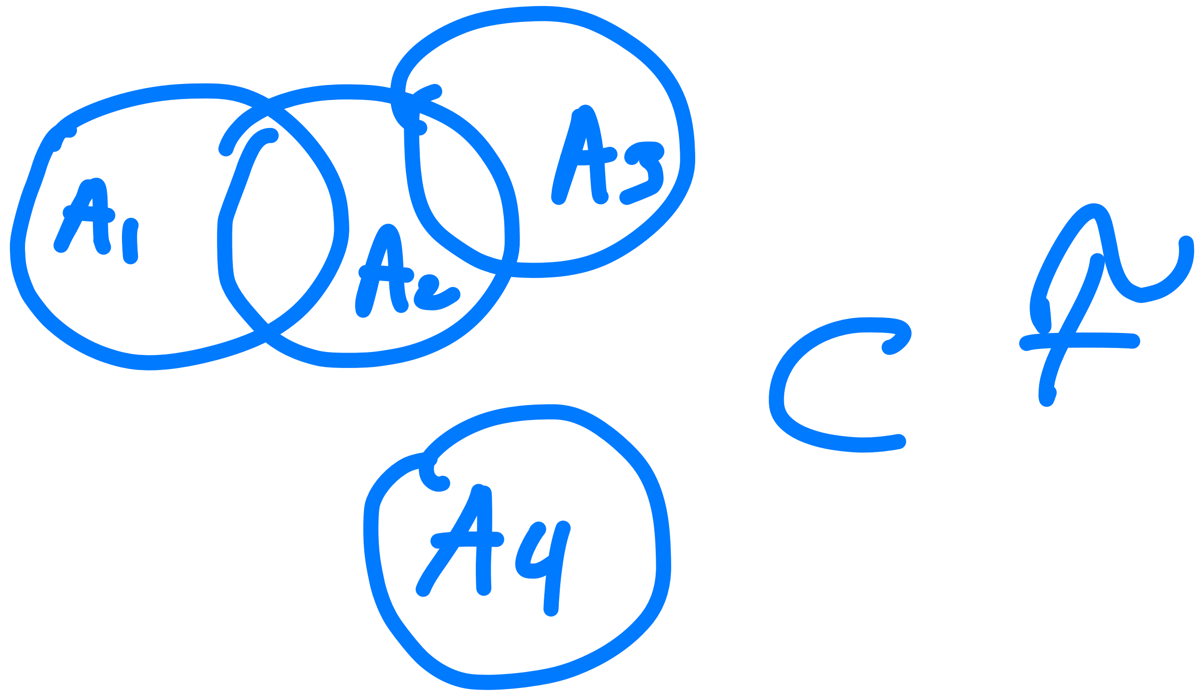
one gets H or T

EVENTS COLLECTION

- Let $\mathscr{F}$ be a collection of events, formally, a sigma field.



- $\emptyset \in \mathscr{F}$
- $A \in \mathscr{F} \implies A^c \in \mathscr{F}$
- $A_i \in \mathscr{F} \implies \cup_{i \in \mathbb{N}} A_i \in \mathscr{F}$



PROBABILITY MEASURE

- Let $\mathbb{P}$ be a probability measure.

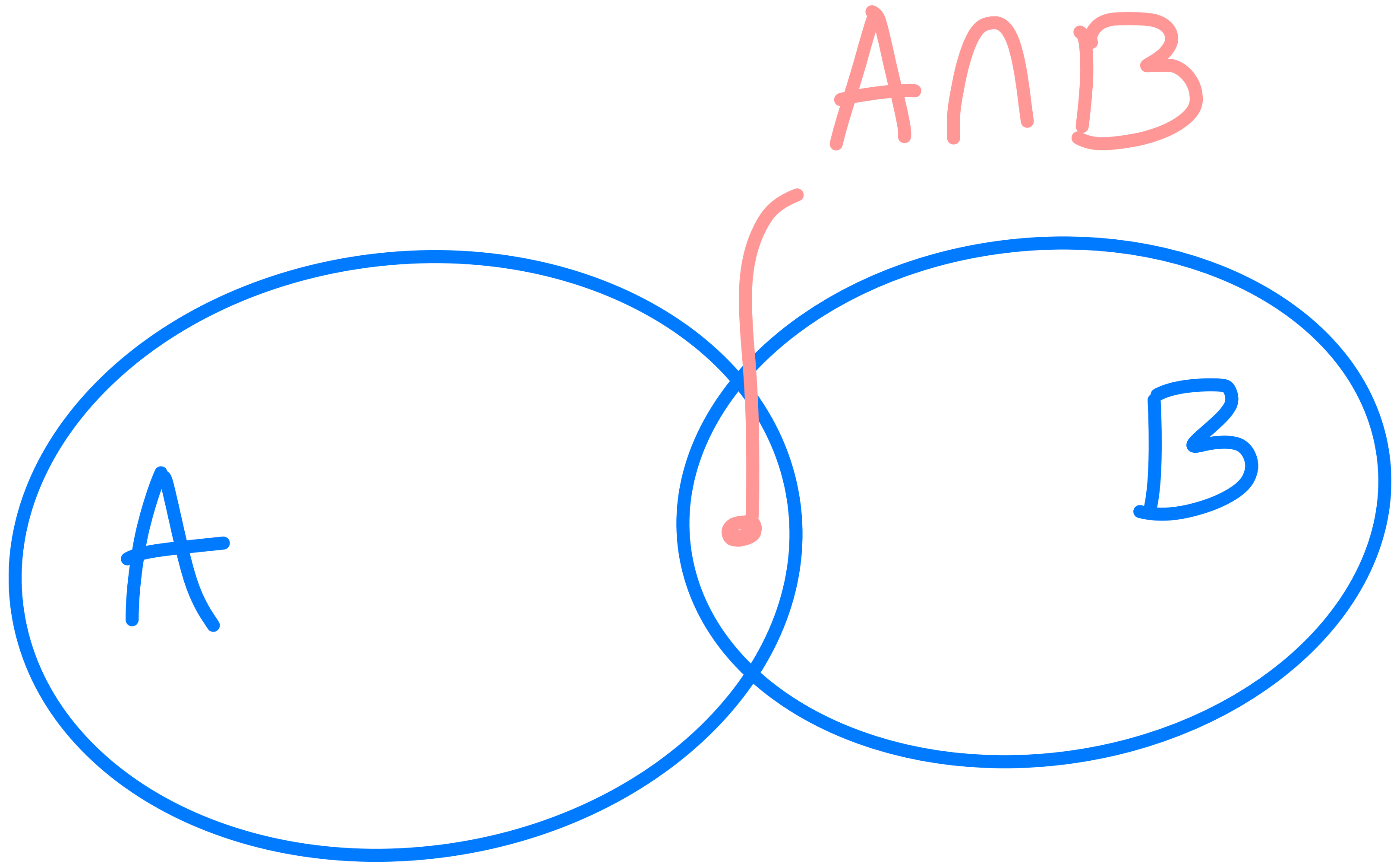
  - $\mathbb{P} : \mathscr{F} \to [0,1]$
  - $\mathbb{P}(\varnothing) = 0,\ \mathbb{P}(\Omega) = 1$

  - If the B_i are disjoint events, $\mathbb{P}(\cup_{i \in \mathbb{N}} B_i) = \sum_{i \in \mathbb{N}} \mathbb{P}(B_i)$

B_1 B_2

PROBABILITY MEASURE

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

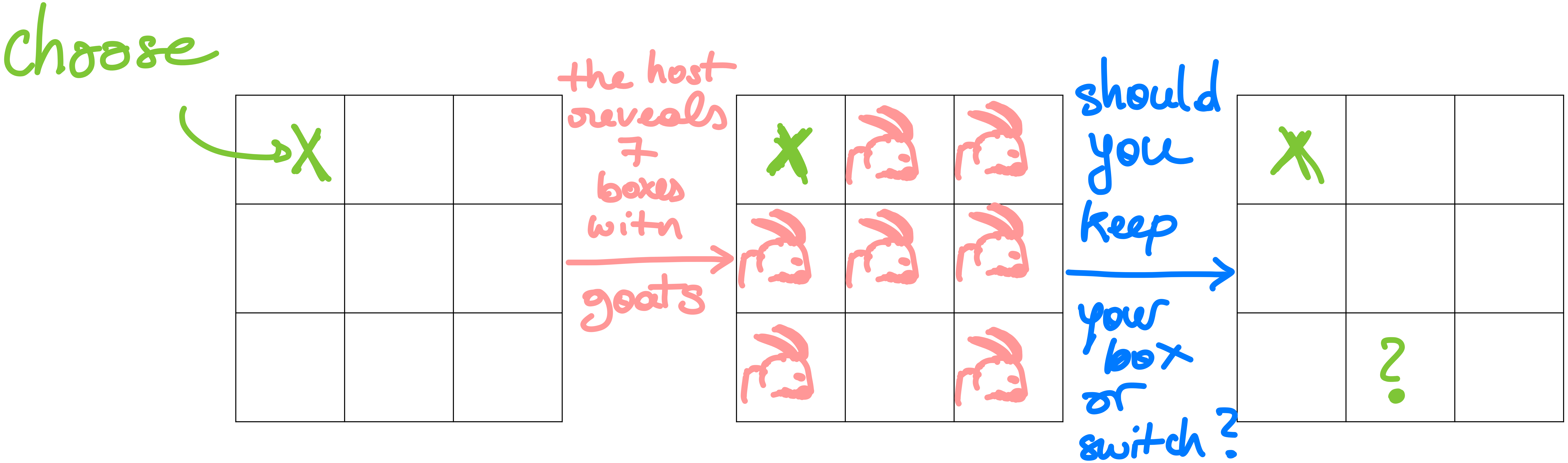


CONDITIONAL PROBABILITY

CONDITIONAL PROBABILITY

The Monty Hall Show

- You have to choose one box among 9.
- 8 boxes hide a goat. 1 box hides one million $.



CONDITIONAL PROBABILITY

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \overbrace{\frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}}^{\mathbb{P}(B \mid A)} \frac{\mathbb{P}(A)}{\mathbb{P}(B)}$$

BAYES' RULE

$$\mathbb{P}(A \mid B) = \mathbb{P}(B \mid A)\frac{\mathbb{P}(A)}{\mathbb{P}(B)}$$

INDEPENDENCE

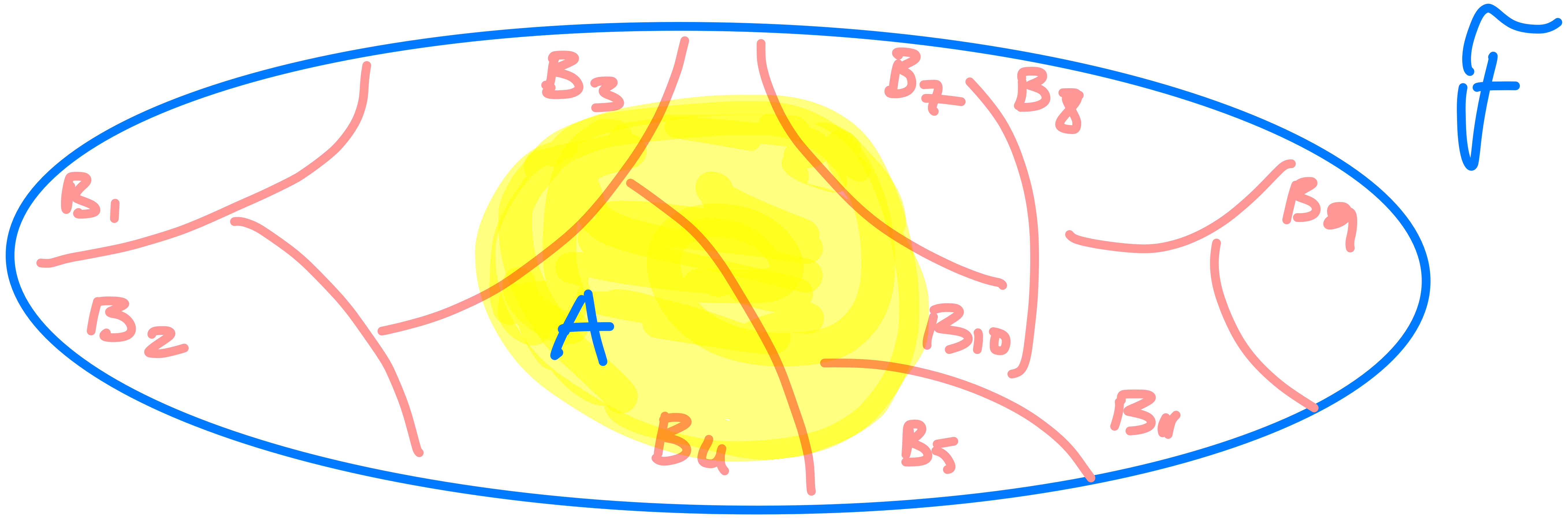
- Two events are independent if the occurence of not does not influence the occurence of the other one.

$$\mathbb{P}(A \mid B) = \mathbb{P}(A)$$

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B)$$

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(B)}$$

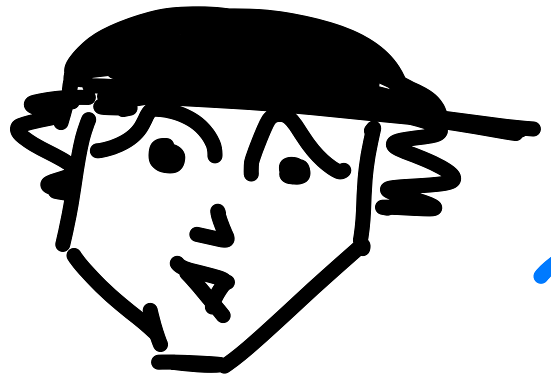
LAW OF TOTAL PROBABILITIES



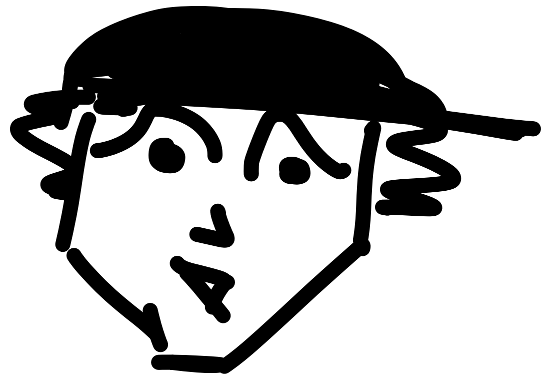
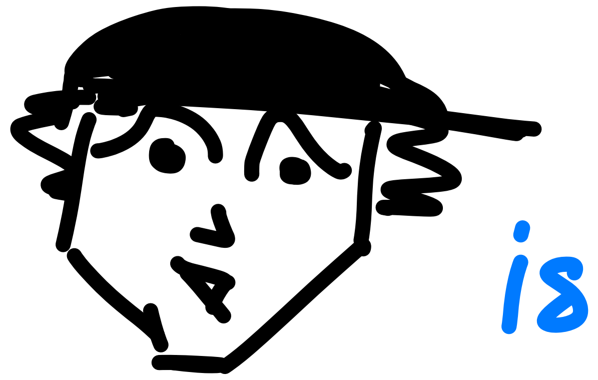
$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{n} \mathbb{P}(A \mid B_i) \times \mathbb{P}(B_i)$$

COVID TESTING

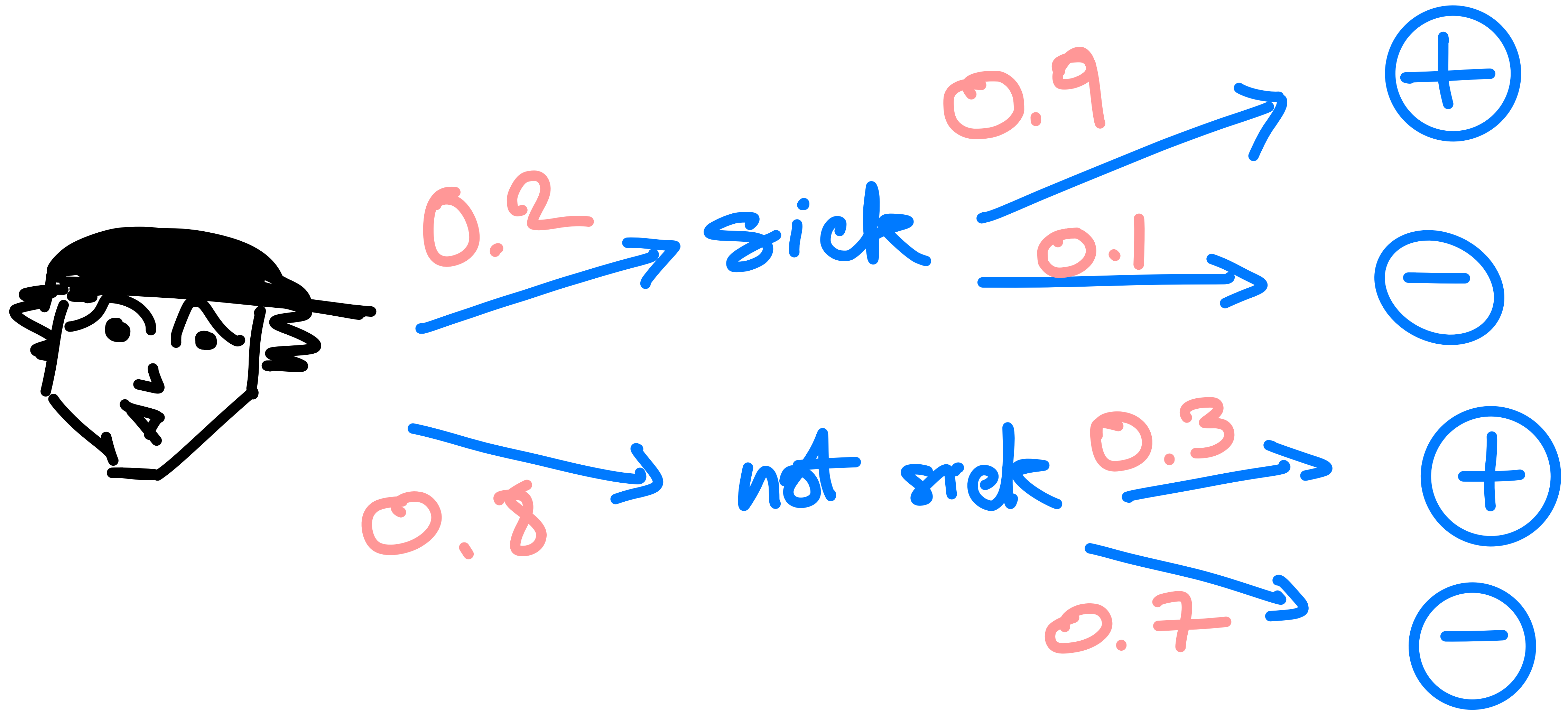
Interpreting a positive result...

 has $\frac{2}{10}$ chances of being sick...

The test gives $\frac{1}{10}$ false negative
and $\frac{1}{3}$ false positive.

 tested $+$. What is the $\mathbb{P}$ that  is sick?

COVID TESTING



$$\mathbb{P}(\text{sick} \mid \oplus) = \frac{\mathbb{P}(\oplus \mid \text{sick})\,\mathbb{P}(\text{sick})}{\mathbb{P}(\oplus)}$$

Bayes Rule

* $\mathbb{P}(\oplus \mid \text{sick}) = 0.9$ * $\mathbb{P}(\text{sick}) = 0.2$

* $\mathbb{P}(\oplus) = \mathbb{P}(\oplus \mid \text{sick})\,\mathbb{P}(\text{sick}) + \mathbb{P}(\oplus \mid \text{not sick})\,\mathbb{P}(\text{not sick})$

law of total probability

$$\mathbb{P}(\text{sick} \mid \oplus) = \frac{0.9 \times 0.2}{0.9 \times 0.2 + 0.3 \times 0.8} = \frac{18}{42} = \frac{3}{7} = 0.43$$

RANDOM VARIABLES

RANDOM VARIABLES

- Let X be a function
  - $X : \Omega \to \mathbb{R}$
  - such that $\{w \mid X(w) \leq c\} \in \mathscr{F}$

a possible outcome

ROLL TWO DICES

$\Omega = \{ \{i,j\} \in [\![1,6]\!]^2 \}$

$\mathcal{F} = \{ \emptyset, \Omega, \{1,1\}, \{1,1\}^c, \dots \} = 2^\Omega$

$\mathbb{P}(ij) = 1/36$ ← uniform probability measure

$X(\{i,j\}) = i + j$ ← sum of the dice rolls

$Y(\{i,j\}) = \max(i,j)$ ← maximum dice roll

ROLL TWO DICES

disjoint events

$\mathbb{P}(\{\omega \mid X(\omega) \leq 3\}) = \mathbb{P}(\{11, 12, 21\}) = \mathbb{P}(11) + \mathbb{P}(12) + \mathbb{P}(21)$

" $\mathbb{P}(X \leq 3)$ "

uniform probability $= 3/36$

$\mathbb{P}(\{\omega \mid Y(\omega) = 6\}) = \mathbb{P}(\{16, 26, 36, 46, 56, 66, 61, 62, 63, 64, 65\})$

" $\mathbb{P}(Y = 6)$ "

$= 11/36$ (disjoint events + UP)

ROLL TWO DICES

$\mathbb{P}(X=1)=0$

11 $\quad \mathbb{P}(X=2)=1/36$

12, 21 $\quad \mathbb{P}(X=3)=2/36$

$\vdots$

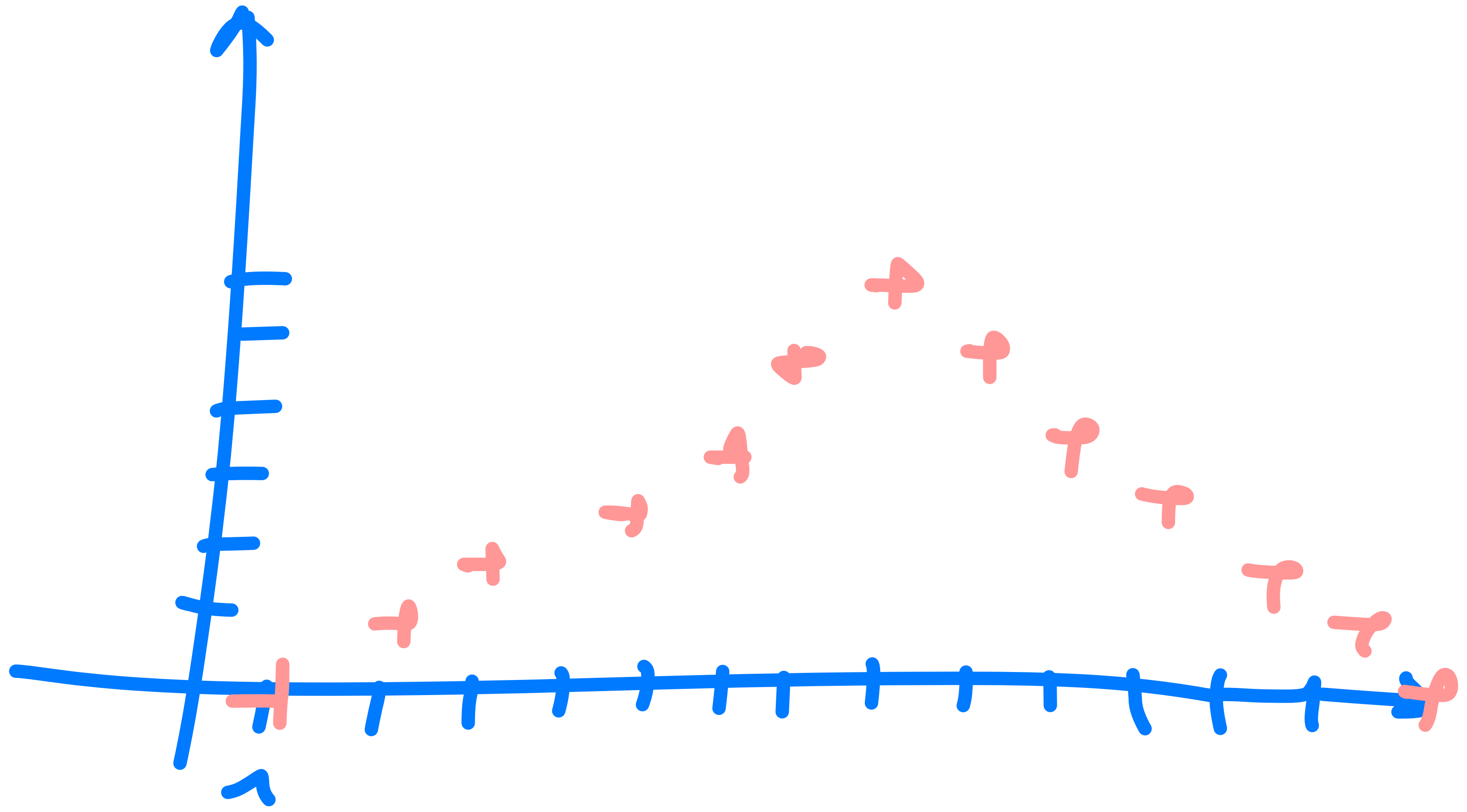
16, 25, 36 $\quad \mathbb{P}(X=7)=6/36$

$\vdots$

$\mathbb{P}(X=12)=1/36$

$$\mathbb{P}(X=k)=\frac{6-|k-7|}{36}=\mathbb{P}_X(k)$$

probability law of X



ROLL TWO DICES

$$\mathbb{P}(Y=6)=11/36$$

15 25 35 45 55

$$\mathbb{P}(Y=5)=9/36$$

$$\mathbb{P}(Y=4)=7/36$$

$$\mathbb{P}(Y=3)=5/36$$

$$\mathbb{P}(Y=2)=3/36$$

$$\mathbb{P}(Y=1)=1/36$$

$$\mathbb{P}(Y=k)=\frac{2k-1}{36}=\mathbb{P}_Y(k)$$

probability law of Y

PROBABILITY LAW

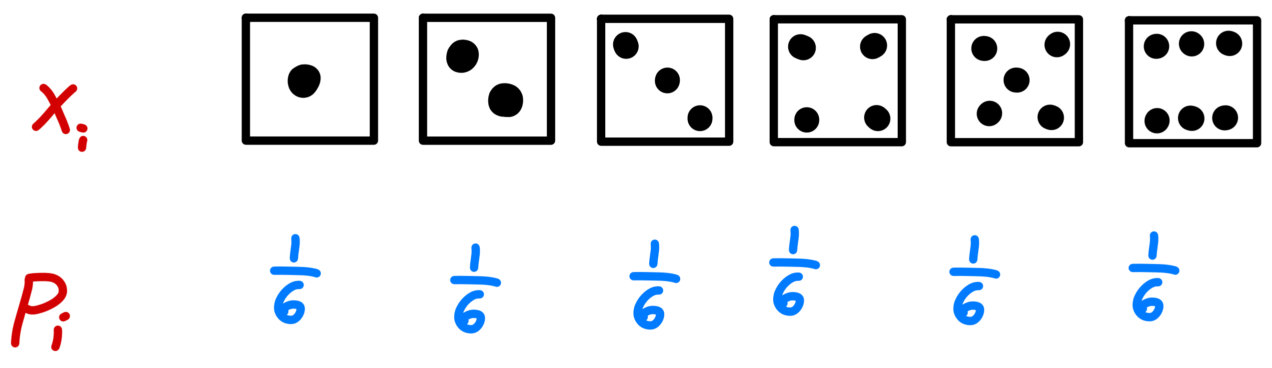
$$\mathbb{P}(\{w \,|\, X(w) = c\}) = \mathbb{P}(X = c) = \mathbb{P}_X(c)$$

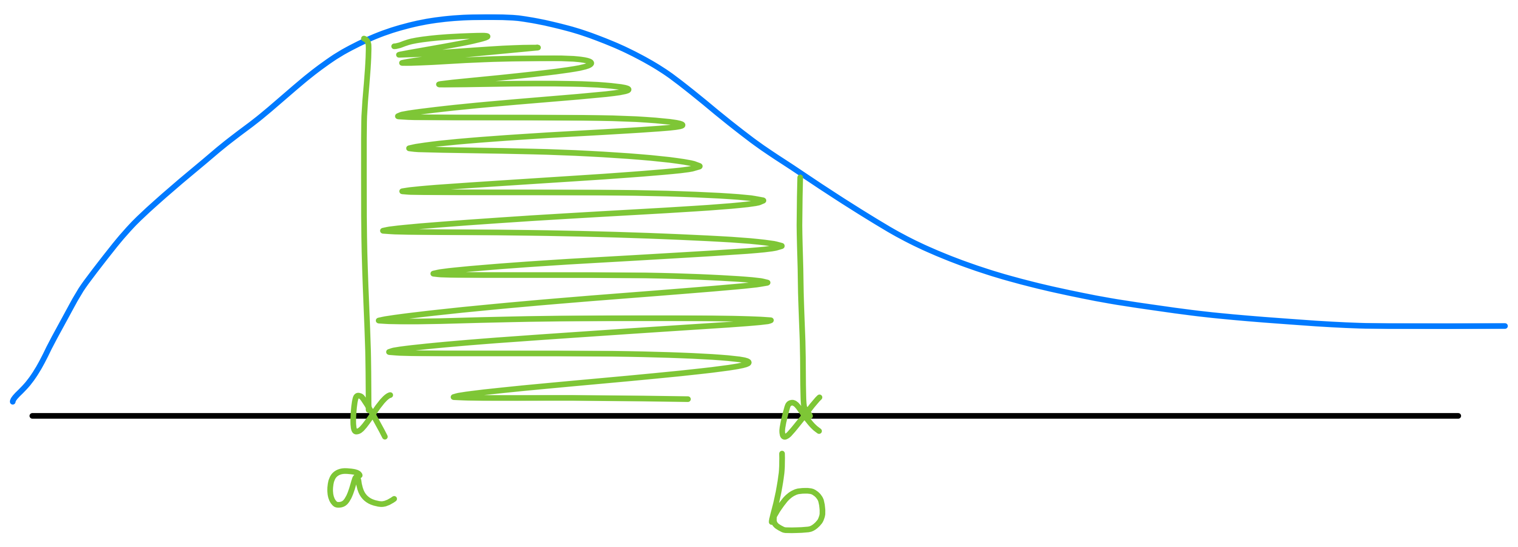
DISCRETE AND CONTINUOUS RANDOM VARIABLES

- **Discrete Random Variable** assumes a countable number of distinct values.
- **Continuous Random Variable** assumes values within intervals.

$$\mathbb{P}(X = x)$$

$$\mathbb{P}(X \leq x) = F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$





$$P(X = 1) = P(X = 2) = \ldots = \frac{1}{6}$$

$$P(a < X < b)$$

DISCRETE RANDOM VARIABLES

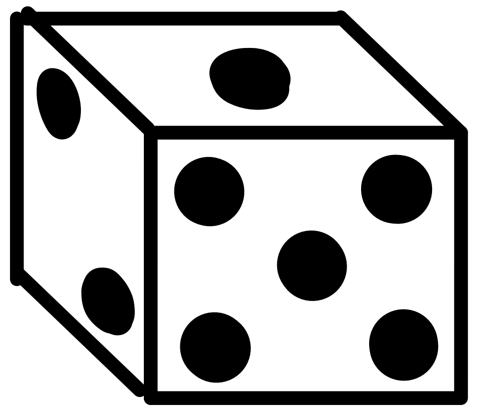
	Distribution	Sample Space	Probability Distribution
$X \sim Ber(p)$	Bernouilli	$\{0, 1\}$	$\mathbb{P}(X=1) = p$
$X \sim Bin(n, p)$	Binomial	$\{1, \dots, n\}$	$\mathbb{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$
$X \sim Poi(\lambda)$	Poisson	$\mathbb{N}$	$\mathbb{P}(X=k) = e^{-\lambda} \frac{\lambda^k}{k!}$

CONTINOUS RANDOM VARIABLES

	Distribution	Sample Space	Probability Distribution
$X \sim U(a,b)$	Uniform	$[a, b]$	$\mathbb{P}(X \le x) = \frac{x-a}{b-a} \mathbb{1}_{\{x \in [a,b]\}}$ $f_X(x) = \frac{\mathbb{1}_{\{x \in [a,b]\}}}{b-a}$
$X \sim \exp(\lambda)$	Exponential	$\mathbb{R}^+$	$\mathbb{P}(X \le x) = 1 - e^{-\lambda x}$ $f_X(x) = \lambda e^{-\lambda x}$
$X \sim \mathcal{N}(\mu, \sigma^2)$	Gaussian	$\mathbb{R}$	$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$

EXPECTATIONS AND VARIANCE

EXPECTATION



- Probability to get 1 is 1/6
- Probability to get 2 is 1/6
- …
- Probability to get 6 is 1/6

$$E[X] = 1.\frac{1}{6} + 2.\frac{1}{6} + 3.\frac{1}{6} + 4.\frac{1}{6} + 5.\frac{1}{6} + 6.\frac{1}{6} = 3.5$$

EXPECTATION

- Discrete Random Variables
- Continuous Random Variables

$$E[X] = \sum_{i}^{n} x_i P(X = x_i)$$

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

$$E[g(X)] = \sum_{i}^{n} g(x_i) P(X = x_i)$$

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

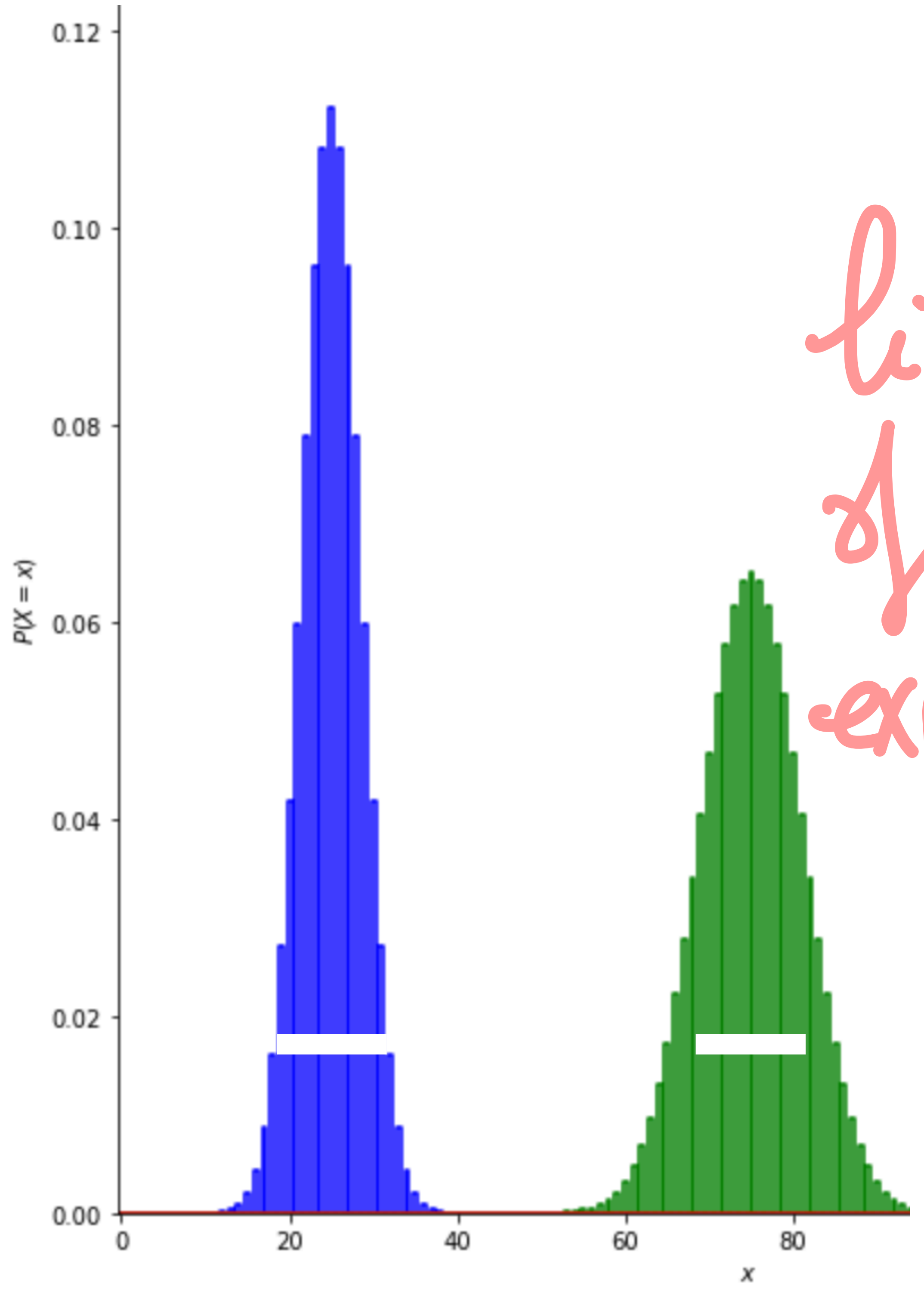
LINEARITY OF EXPECTATION

$$W = aX + bY + c$$

$$E[W] = aE[X] + bE[Y] + c$$

VARIANCE

$$Var(X) = E[(X - E[X])^2]$$



linearity of expectation

$$= E[\,X^2 - 2X\,\overbrace{E(X)}^{\text{constant}} + E(X)^2\,)$$

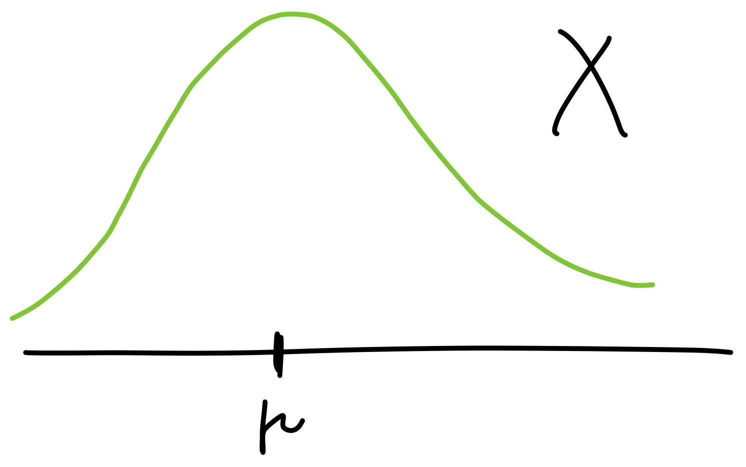
$$= E(X^2) - 2\,E(X)\,E(X) + E(X)^2$$

$$= E(X^2) - E(X)^2$$

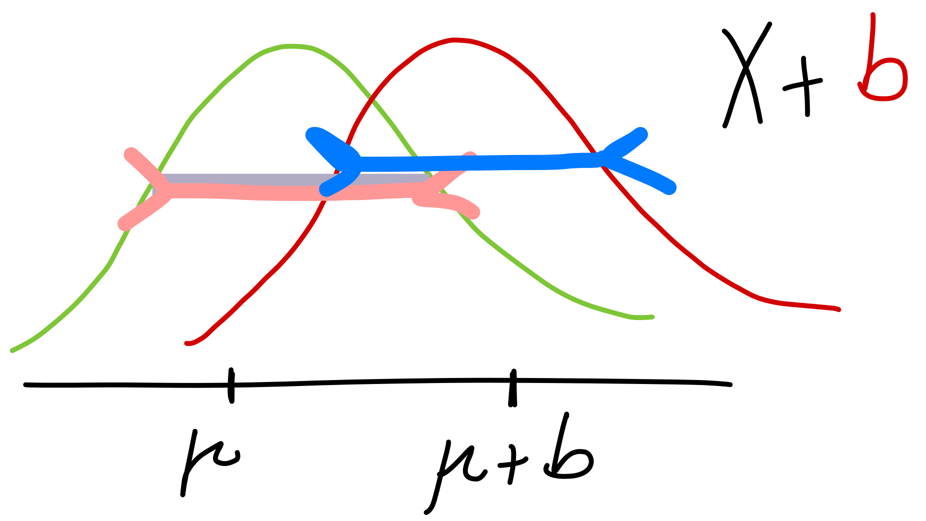
$$SD(X) = \sigma_X = \sqrt{Var(X)}$$

VARIANCE

$$W = X + b$$



$$Var(W) = Var(X)$$



VARIANCE

$$W = X + b$$

$$Var(W) = a^2 Var(X)$$

$$\sigma_W = |a| \sigma_X$$

DISCRETE RANDOM VARIABLES

Distribution	Expectation	Variance
Bernouilli	p	$p(1-p)$
Binomial	np	$np(1-p)$
Poisson	λ	λ

CONTINOUS RANDOM VARIABLES

Distribution	Expectation	Variance
Uniform	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$
Exponential	λ^{-1}	λ^{-2}
Gaussian	μ	σ^2

COVARIANCE

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$$

$$-1 < corr(X, Y) = \frac{COV(X, Y)}{\sigma_X \sigma_Y} < 1$$

COVARIANCE

$$W = aX + bY$$

$$Var(W) = a^2Var(X) + b^2Var(Y) + 2abCOV(X, Y)$$

GAUSSIAN DISTRIBUTION

NORMAL DISTRIBUTION

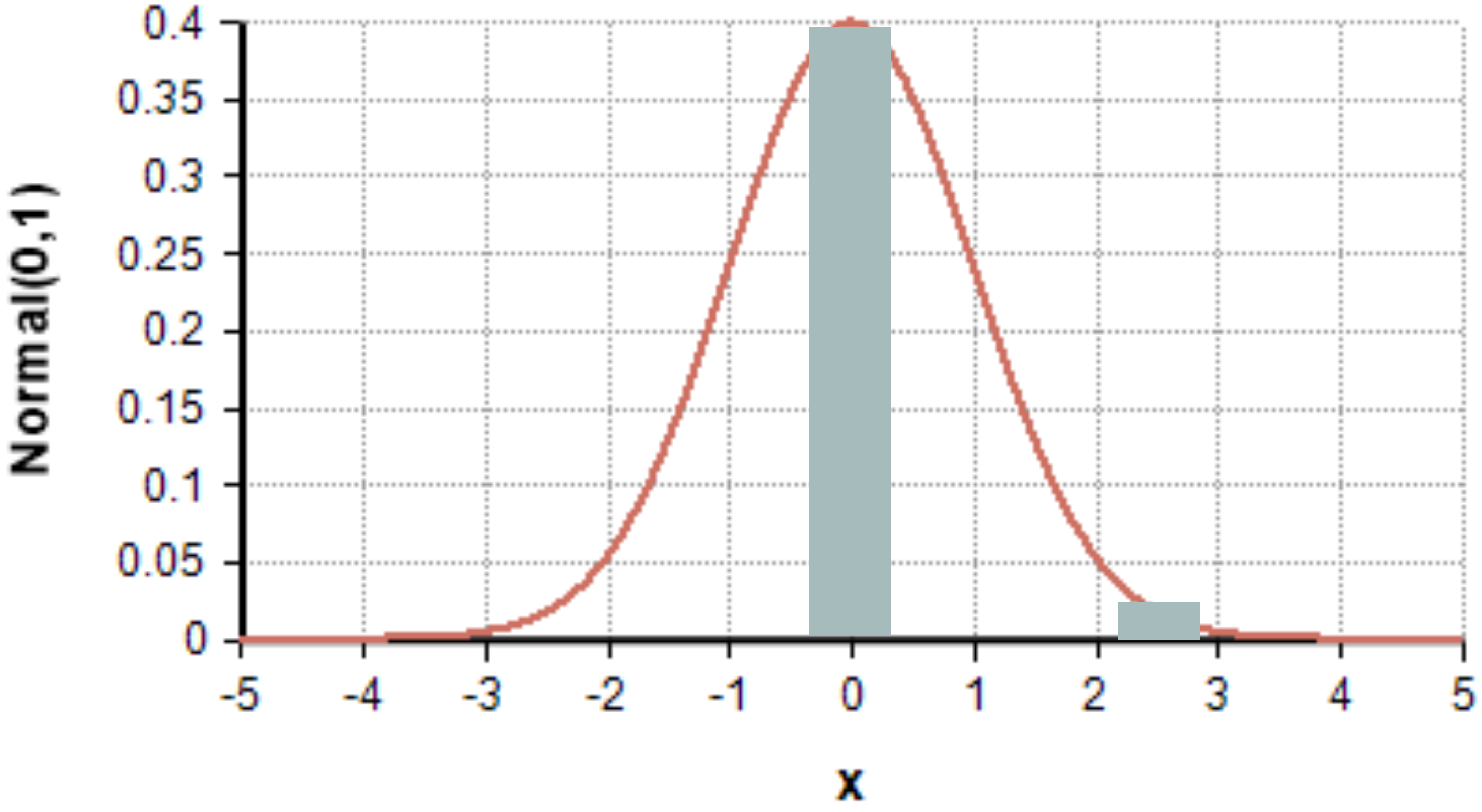
- The most frequently occurring distribution
- Symmetric. Bell-shaped curve.
- More likely to take on values close to the mean

$$\mathcal{N}(\mu, \sigma^2)$$

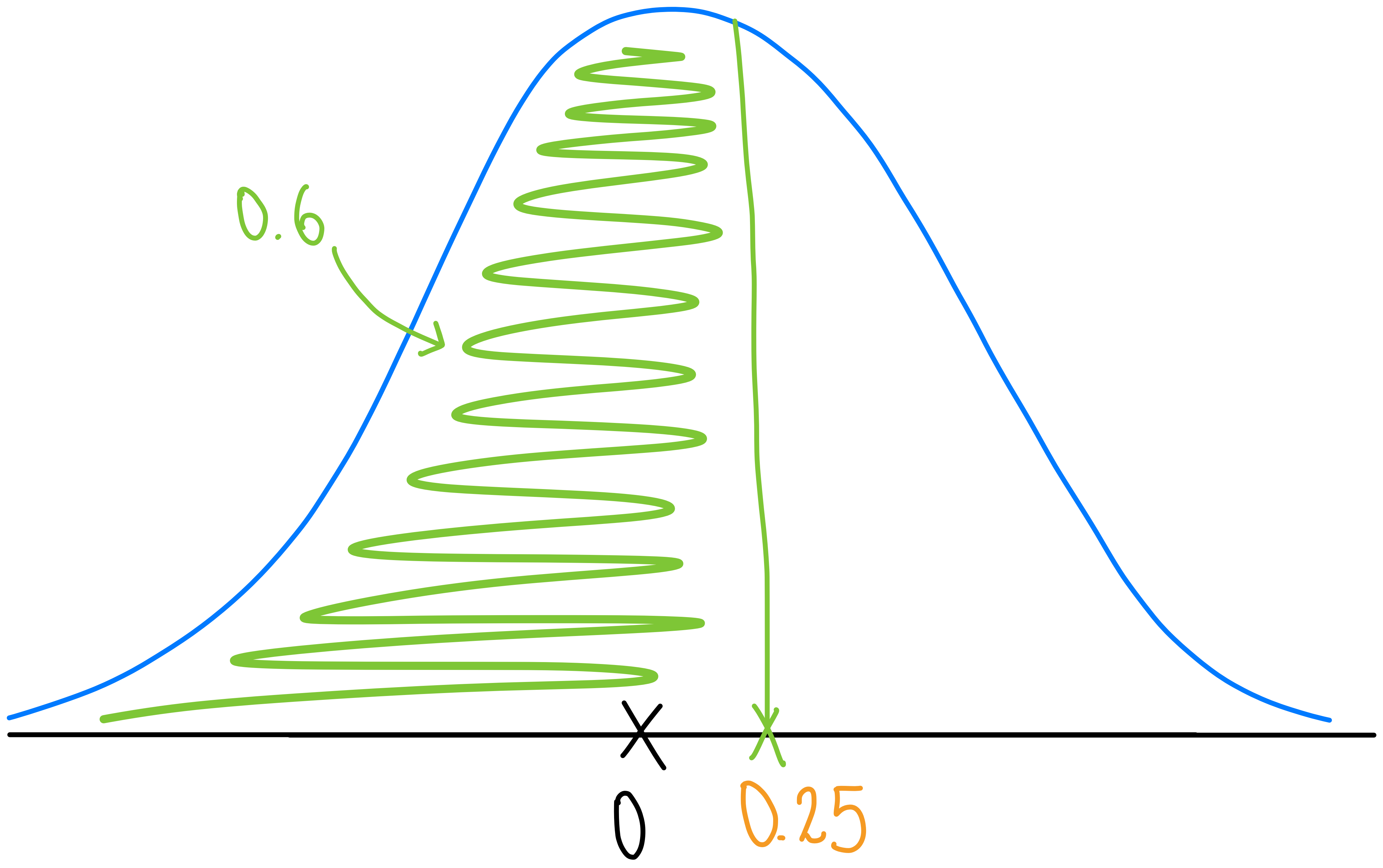
mean variance

NORMAL DISTRIBUTION

➤ Z~N(0,1)

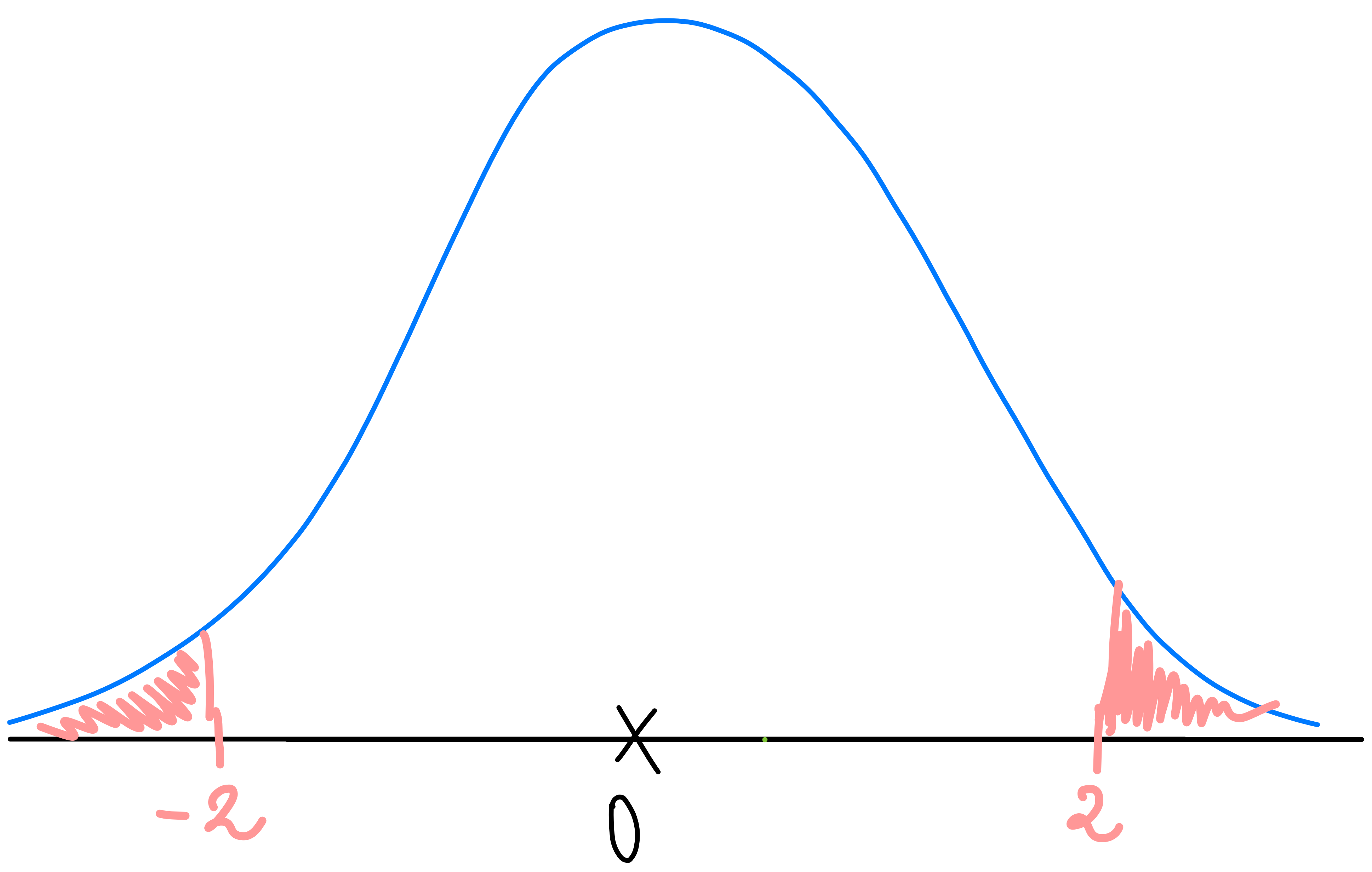


NORMAL DISTRIBUTION



$$\mathbb{P}(Z \leq 0.25) = 0.6$$

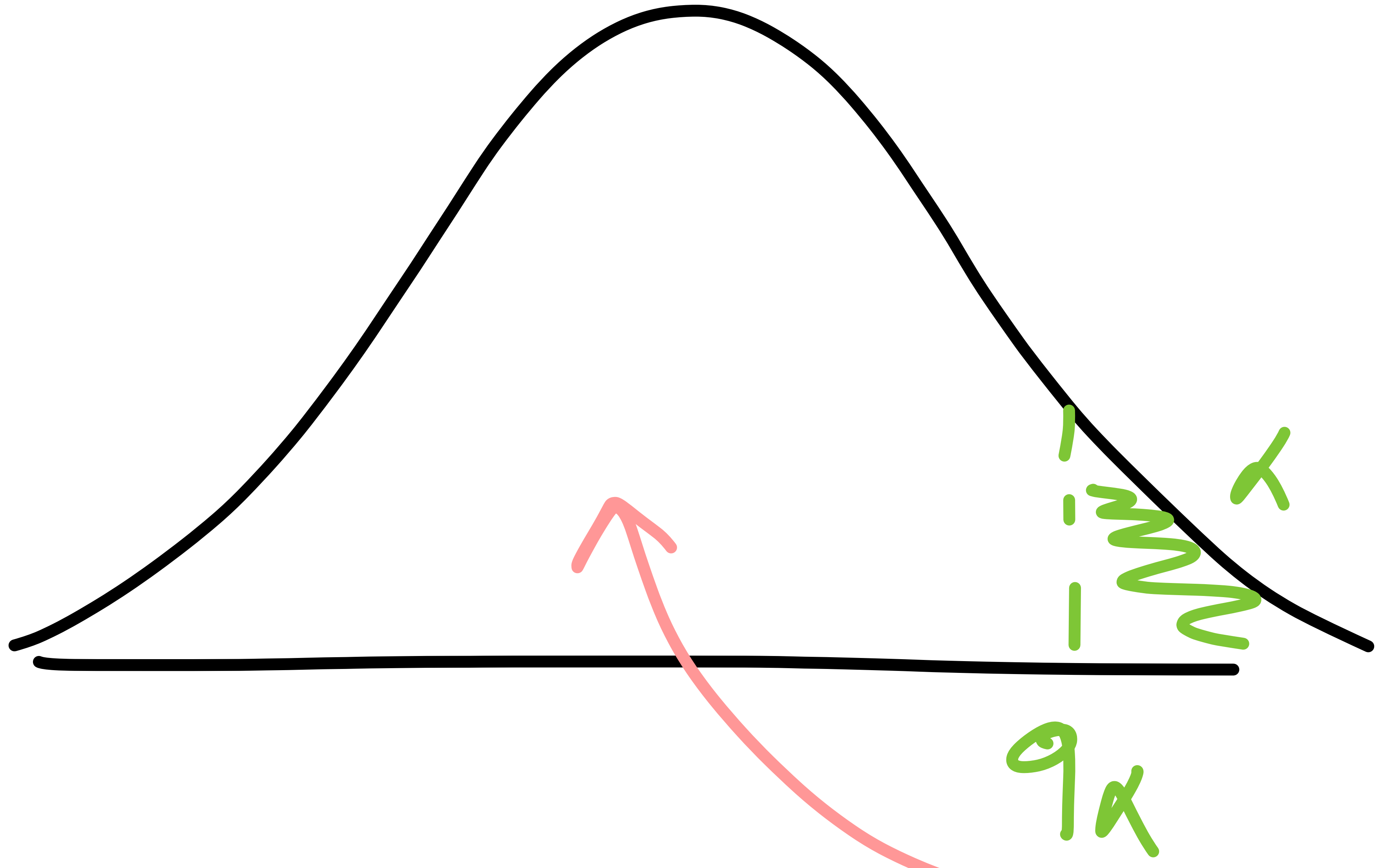
NORMAL DISTRIBUTION

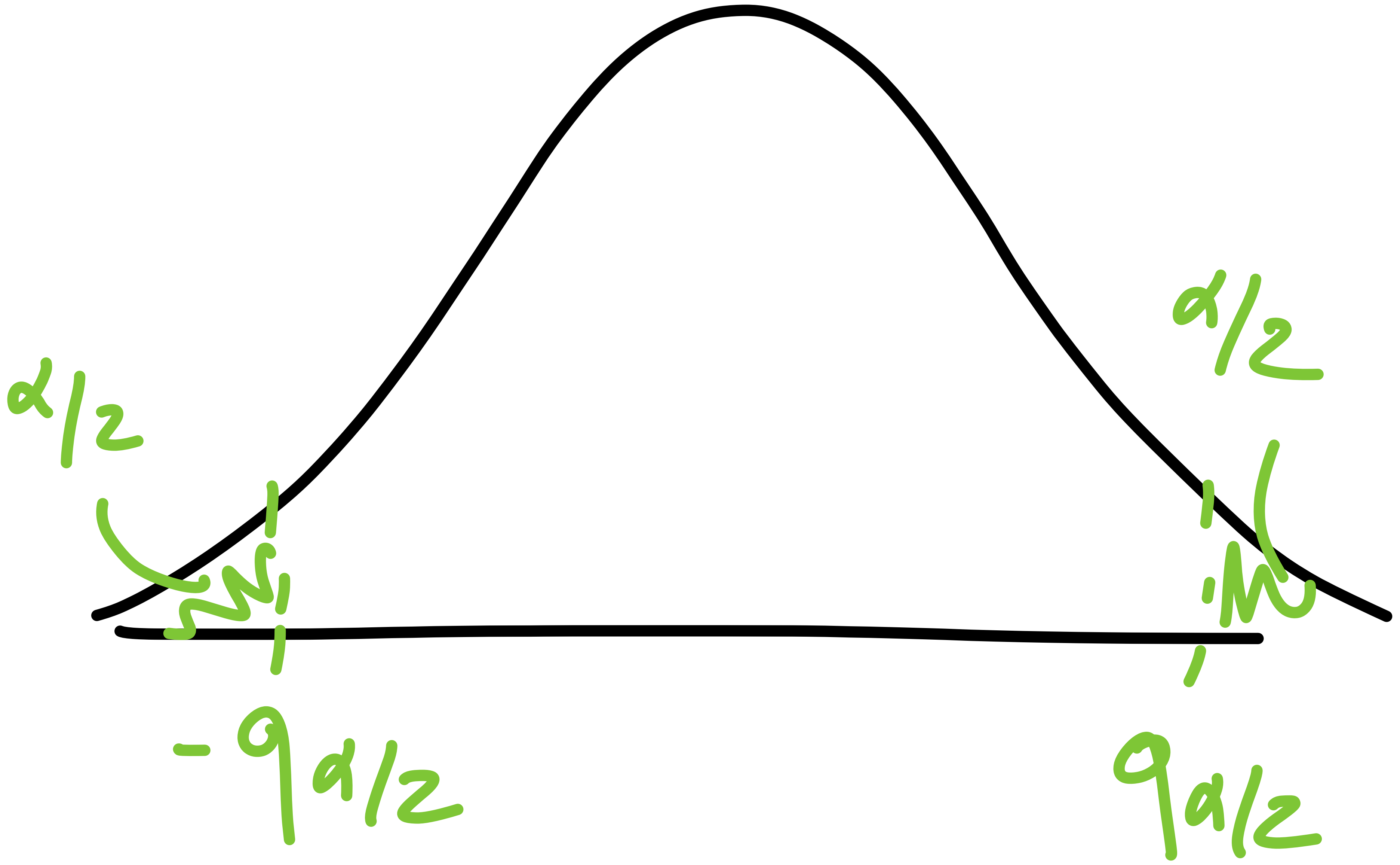


$$P(Z < -2) = P(Z > 2)$$

$$P(Z > -2) = P(Z < 2)$$

QUANTILES





$$P(Z < q_\alpha) = 1 - \alpha$$

$$P(|Z| > q_{\alpha/2}) = \alpha$$

NORMAL DISTRIBUTION

- $Z \sim N(0,1)$ $\quad\quad$ $X = \sigma Z + \mu$
- Find $E[X]$ and $Var(X)$
- $X = \sigma Z + \mu \sim N(\mu,\sigma^2)$

- $X \sim N(\mu,\sigma^2)$ $\quad\quad$ $Z = (X - \mu)/\sigma = (1/\sigma)X - \mu/\sigma$
- Find $E[Z]$ and $Var(Z)$
- $Z = (X - \mu)/\sigma \sim N(0,1)$

CENTRAL LIMIT THEOREM

- Let X_i be n iid random variables.
- Let μ and σ^2 be the expectation and variances of X.

Central Limit Theorem

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \underset{n \to \infty}{\overset{(d)}{\longrightarrow}} \mathcal{N}(0,1)$$

MULTIVARIATE GAUSSIAN

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mathbb{P}(a < X < b) = \mathbb{P}(a-\mu \leq X-\mu \leq b-\mu)$$

$$= \mathbb{P}\left(\frac{a-\mu}{\sigma} \leq \underbrace{\frac{X-\mu}{\sigma}}_{\mathcal{N}(0,1)} \leq \frac{b-\mu}{\sigma}\right)$$

MULTIVARIATE GAUSSIAN

$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ where $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$

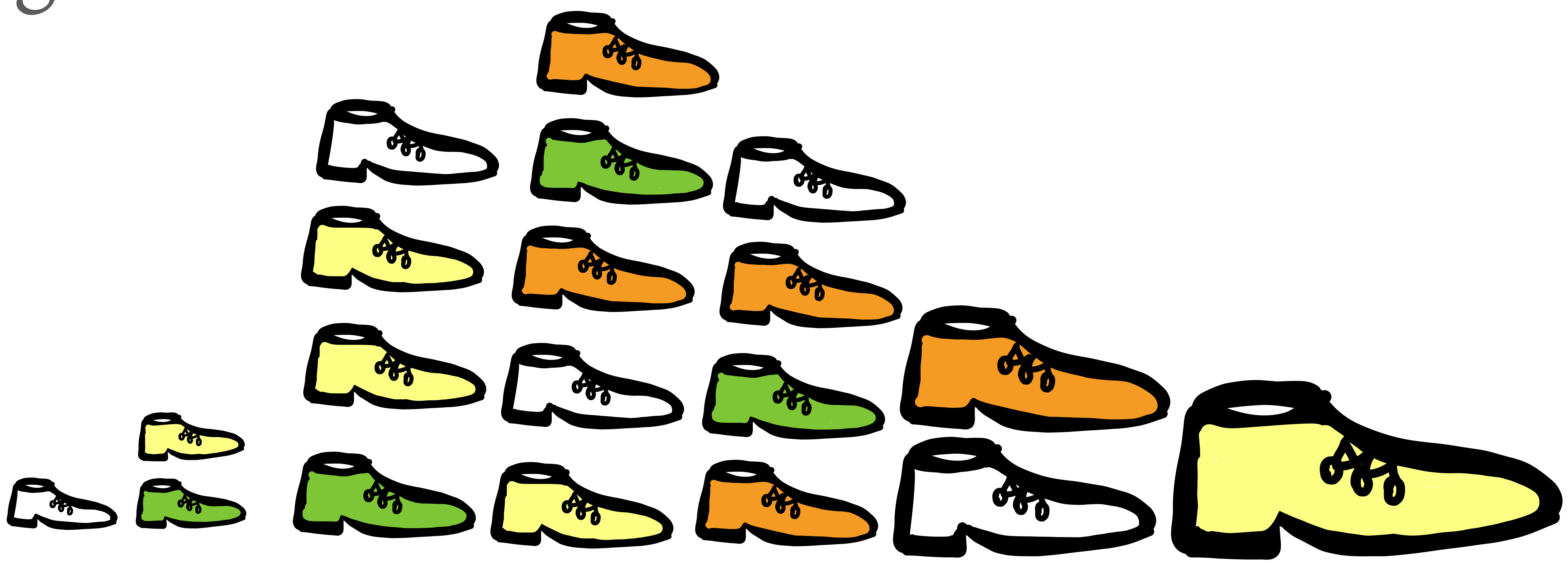
such that $\forall a \in \mathbb{R}^n$, $a^T X$ is a Gaussian.

$$\mathrm{cov}(X) = \left(\mathrm{cov}(X_i, X_j)\right)_{1 \le i,j \le n} = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{cov}(X_1, X_2) & \cdots \\ \vdots & \ddots & \\ & & \mathrm{Var}(X_n) \end{pmatrix}$$

positive-semi definite!

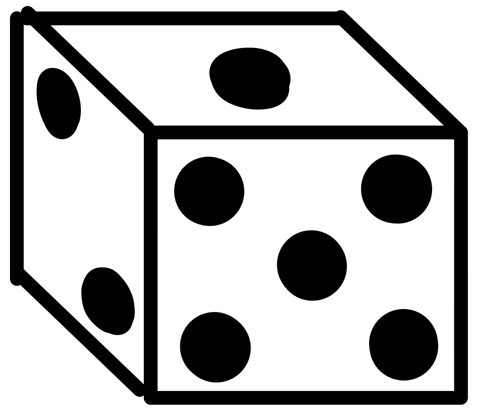
STATISTICS

- Estimation and Estimators
- Hypotheses testing
- Confidence Intervals
- Linear Regression
- Examples



ESTIMATION AND ESTIMATORS

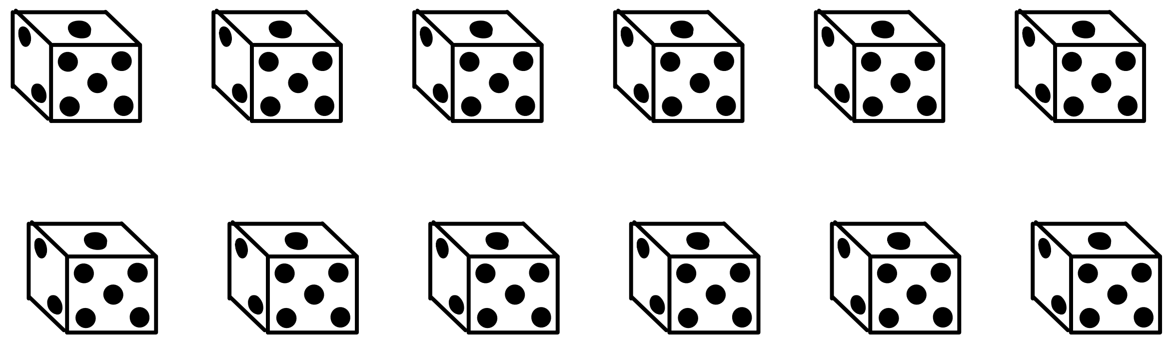
PROBABILISTIC



- Probability to get 1 is 1/6
- Probability to get 2 is 1/6
- …
- Probability to get 6 is 1/6

An underlying law governs a phenomenon. Experiments allow to uncover that law.

EMPIRICAL



- 1, 1, 2, 3, 3, 4, 4, 5, 5, 6, 6, 6

PROBABILISTIC

- Recover the parameters of the distribution, for example the expected value and the variance
- Use the estimators computed empirically
- Test whether the estimator is consistent with a prior hypothesis

ESTIMATION

EMPIRICAL

- Conduct and experiment, compute estimators of the quantities of interest.

$$E[X] = \frac{X_1 + .. + + X_n}{n}$$

$$Var(X) = \frac{(X_1 - E[X])^2 + \ldots + (X_n - E[X])^2}{n-1}$$

PROBABILISTIC

- X model a phenomenon
- X is characterized by a probability distribution
- E[X] is the expected value, computed from the probability

$$E[X] = \sum_{i}^{n} p_i x_i$$

- Var(X), COV(X,Y)… are computed from the probability

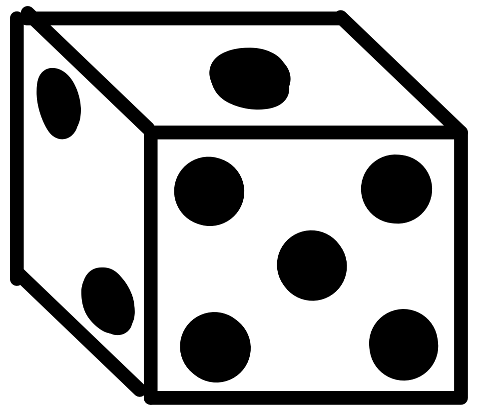
EMPIRICAL

- $X_1, X_2, \ldots, X_n$ describe an experiment
- $X_1, X_2, \ldots, X_n$ are characterised by experimental values
- E[X] is the empirical expected value, computed from the data

$$E[X] = \frac{X_1 + .. + + X_n}{n}$$

- Var(X), COV(X,Y)… are computed from the data

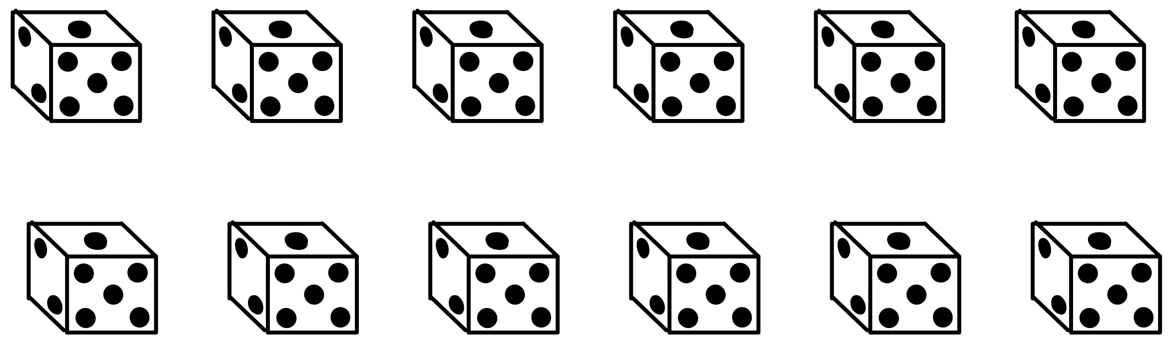
PROBABILISTIC



- Probability to get 1 is 1/6
- Probability to get 2 is 1/6
- …
- Probability to get 6 is 1/6

$$E[X] = 1.\frac{1}{6} + 2.\frac{1}{6} + 3.\frac{1}{6} + 4.\frac{1}{6} + 5.\frac{1}{6} + 6.\frac{1}{6} = 3.5$$

EMPIRICAL



- 1, 1, 2, 3, 3, 4, 4, 5, 5, 6, 6, 6

$$E[X] = \frac{1+1+2+3+3+4+4+5+5+6+6+6}{12}$$

$$E[X] = 1.\frac{1}{6} + 2.\frac{1}{12} + 3.\frac{1}{6} + 4.\frac{1}{6} + 5.\frac{1}{6} + 6.\frac{3}{12} = 3.8$$

ESTIMATORS

$$\left(\{0,1\}, (Ber(p))_{p \in [0.2,0.4]} \right)$$

- Let $X_1, X_2, \ldots, X_n$ be data points from the experiment.
- Let's define an estimator for p:

$$\hat{p}_n = \frac{\sum_{i=1}^{n} X_i}{n}$$

$\mathbb{E}(\hat{p}_n) = \mathbb{E}(X_1) = p$

$\hat{p}_n$ is unbiased

$Var(\hat{p}_n) = \frac{p(1-p)}{n}$

ESTIMATORS

$$\left(\mathbb{R}^+, (\mathbb{U}[a, a+1])_{a>0} \right)$$

- Let $X_1, X_2, \ldots, X_n$ be data points from the experiment.
- Let's define an estimator for a:

$$\hat{a}_n = \frac{\sum_{i=1}^{n} X_i}{n}$$

$\mathbb{E}(\hat{a_n}) = \mathbb{E}(X_1) = a + \frac{1}{2}$

$\hat{a_n}$ is biased

$Var(\hat{a_n}) = \frac{1}{12n}$

QUADRATIC RISK

- Estimator's bias:

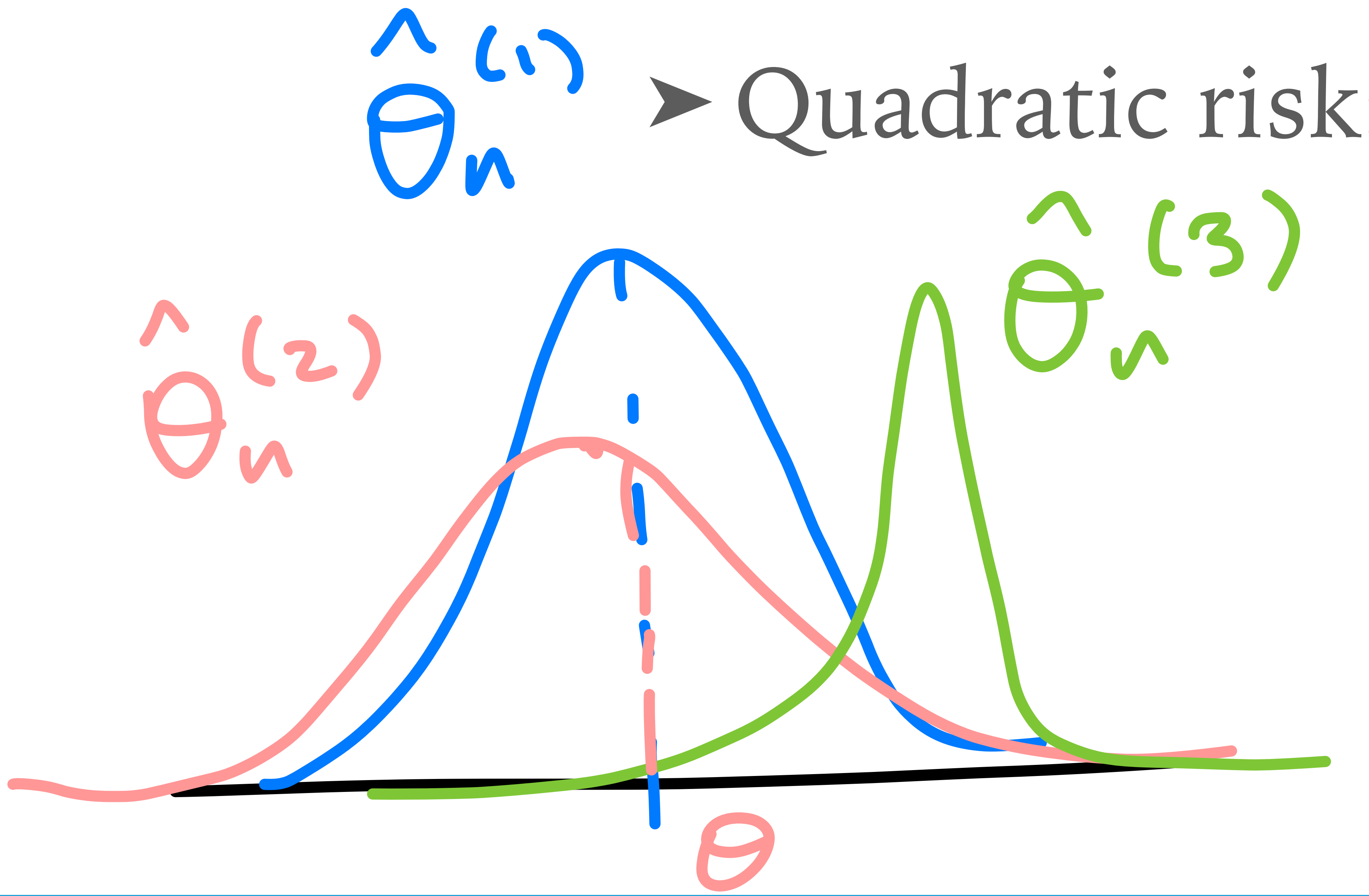
$$bias = E[\hat{\theta}_n] - \theta$$

- Estimator's variance:

$$variance = Var[\hat{\theta}_n]$$

- Quadratic risk:

$$risk = variance + bias^2$$



EXERCISE

$$X_1, X_2, \cdots, X_n \sim^{iid} Ber(p)$$

$$\hat{p}_n = \frac{\sum_{i=1}^{n} X_i}{n}$$

$\mathbb{E}(\hat{p}_n) = p$

$Var(\hat{p}_n) = \frac{p(1-p)}{n}$

$$\hat{p}_n = \frac{X_1 + X_2}{2}$$

$\mathbb{E}(\hat{p}_n) = p$

$Var(\hat{p}_n) = \frac{p(1-p)}{2}$

HYPOTHESIS TESTING

HYPOTHESIS TESTING — TATUM'S 3 POINT SHOTS

- Tatum that he scores 80% at 3 pts. No more, no less.
- Brown challenges Tatum… They collect data on Tatum shooting 3 pts.
- n = 400, $X_1, X_2, \cdots, X_n \sim^{iid} Ber(p)$

$$H_0 : p = 0.8$$

$$H_1 : p \neq 0.8$$

HYPOTHESIS TESTING — TATUM'S 3 POINT SHOTS

➤ Let's build an estimator for the test: $\hat{p}_n = \sum_{i=1}^{n} \frac{X_i}{n}$ ← unbiased estimator

➤ If H_0 is true, then, by the $\underbrace{\text{Central Limit Theorem}}$:

$$\to \sqrt{n}\,\frac{\bar{X}_n - \mathbb{E}(X_1)}{\sqrt{Var(X_1)}} \xrightarrow[n\infty]{} \mathcal{N}(0,1)$$

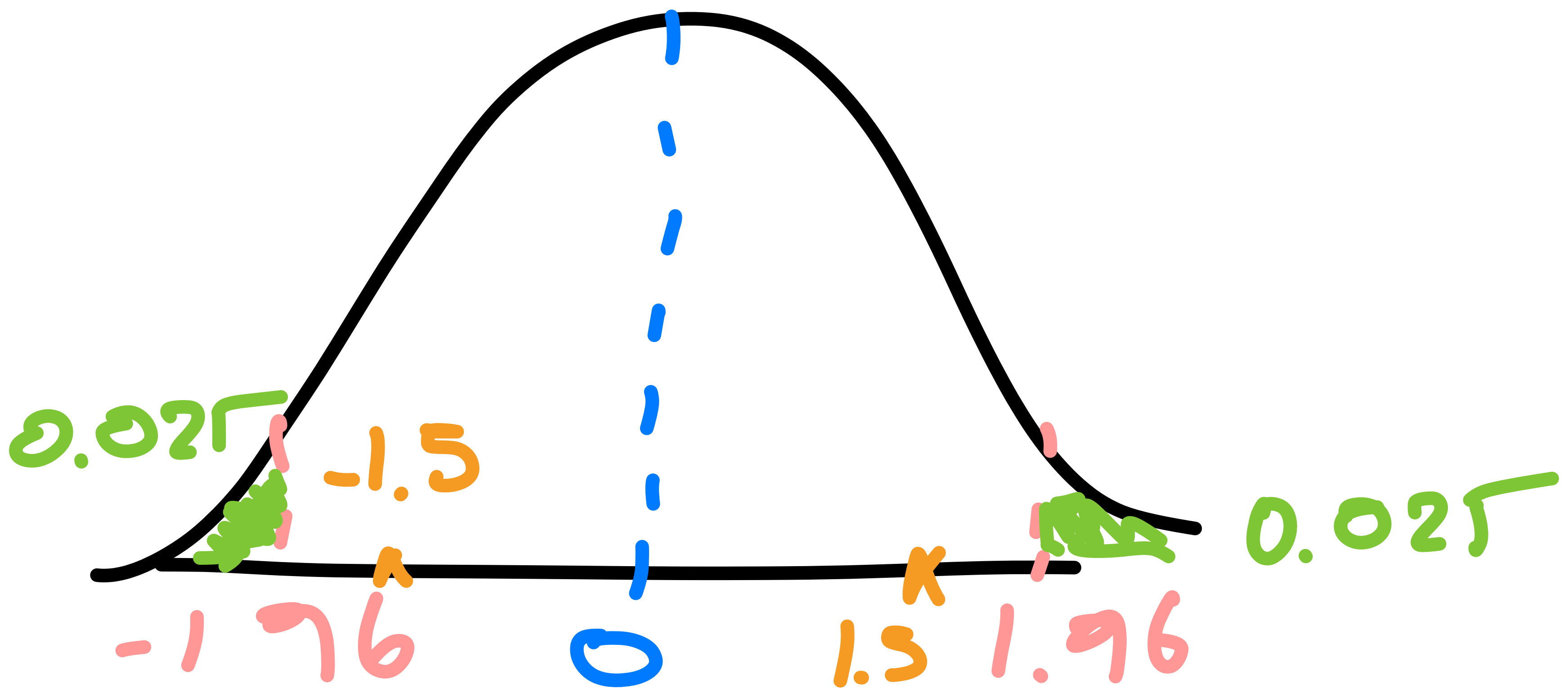
$$\sqrt{n}\,\frac{\hat{p}_n - 0.8}{\sqrt{0.8 \times 0.2}} \approx \mathcal{N}(0,1)$$

HYPOTHESIS TESTING — TATUM'S 3 POINT SHOTS

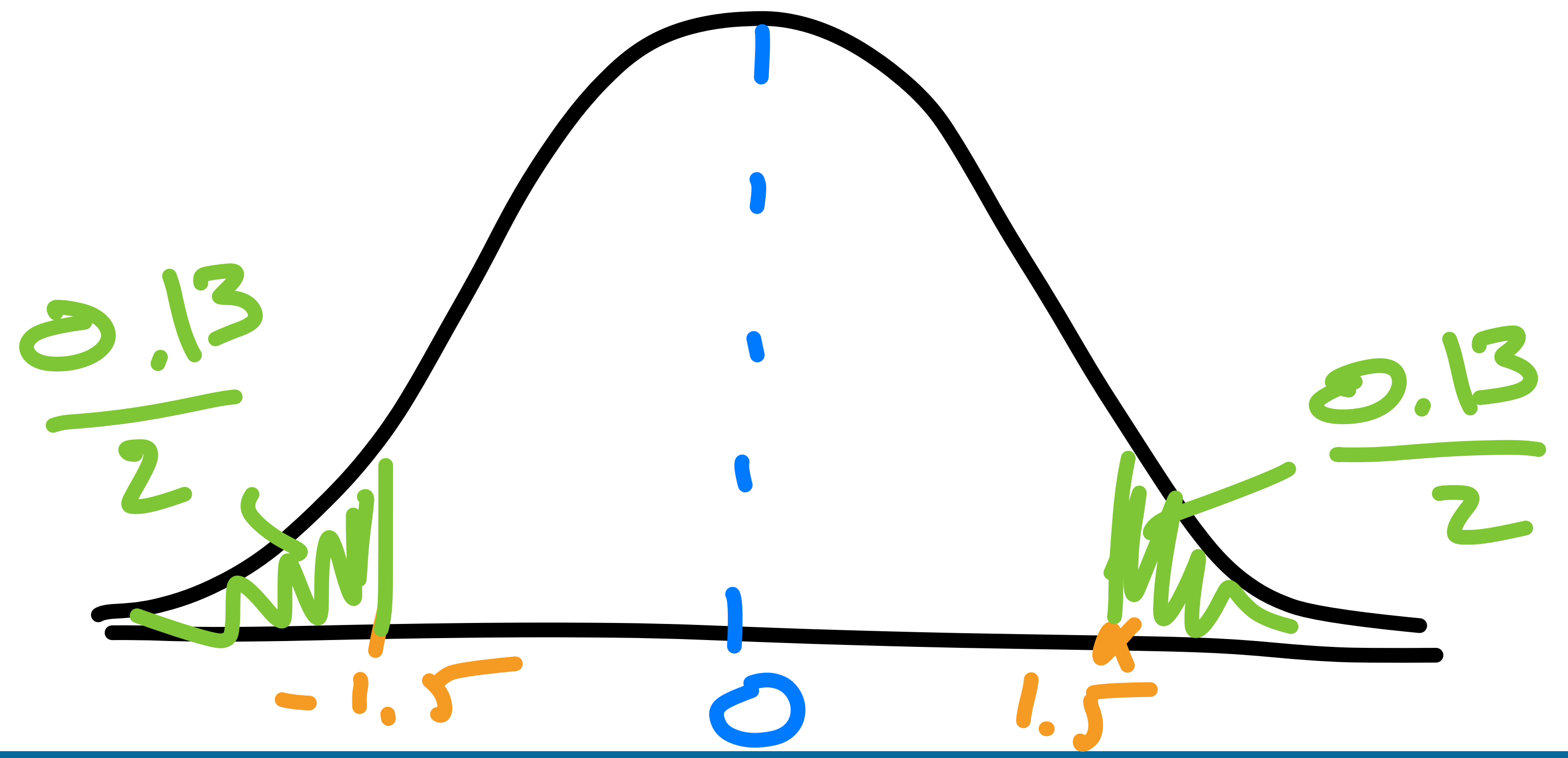
$$\sqrt{n}\frac{\hat{p}_n - 0.8}{\sqrt{0.8 \times 0.2}} = -1.5$$

- Is it a plausible realisation for a Gaussian?

$\mathbb{P}(|Z| > 1.96) = 0.05$



$\mathbb{P}(|Z| > 1.5) = 0.13$

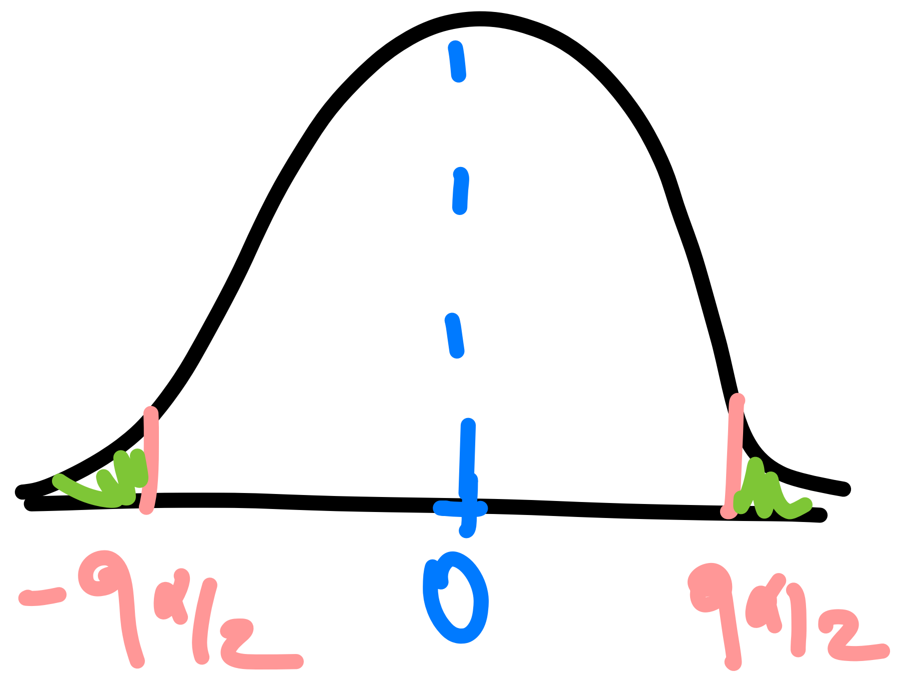


REJECTION SET

Two-sided test

$$H_0 : \theta = 0.8$$

$$H_1 : \theta \neq 0.8$$



$$\left|\underbrace{\sqrt{n}\frac{\bar{X}_n - 0.8}{\sqrt{0.8 \times 0.2}}}_{Z_n^*}\right| > q_{\alpha/2}$$

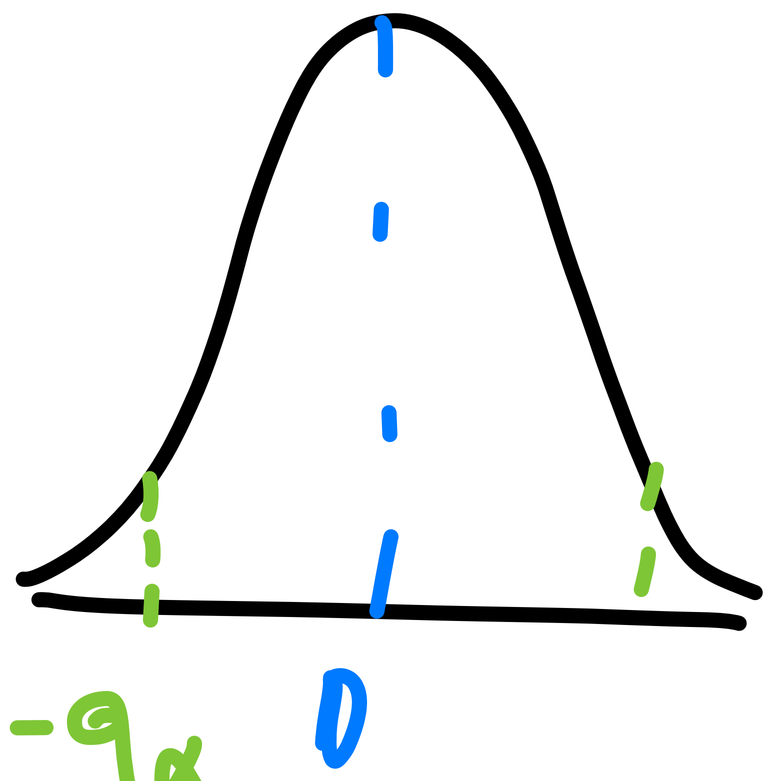
if $Z_n^* \in$ ..., Z_n^* is likely not from "H_0's distribution" => REJECT H_0

One-sided test

$$H_0 : \theta \geq 0.8$$

$$H_1 : \theta < 0.8$$

$$\underbrace{\sqrt{n}\frac{\bar{X}_n - 0.8}{\sqrt{0.8 \times 0.2}}}_{Z_n^*} < -q_{\alpha}$$



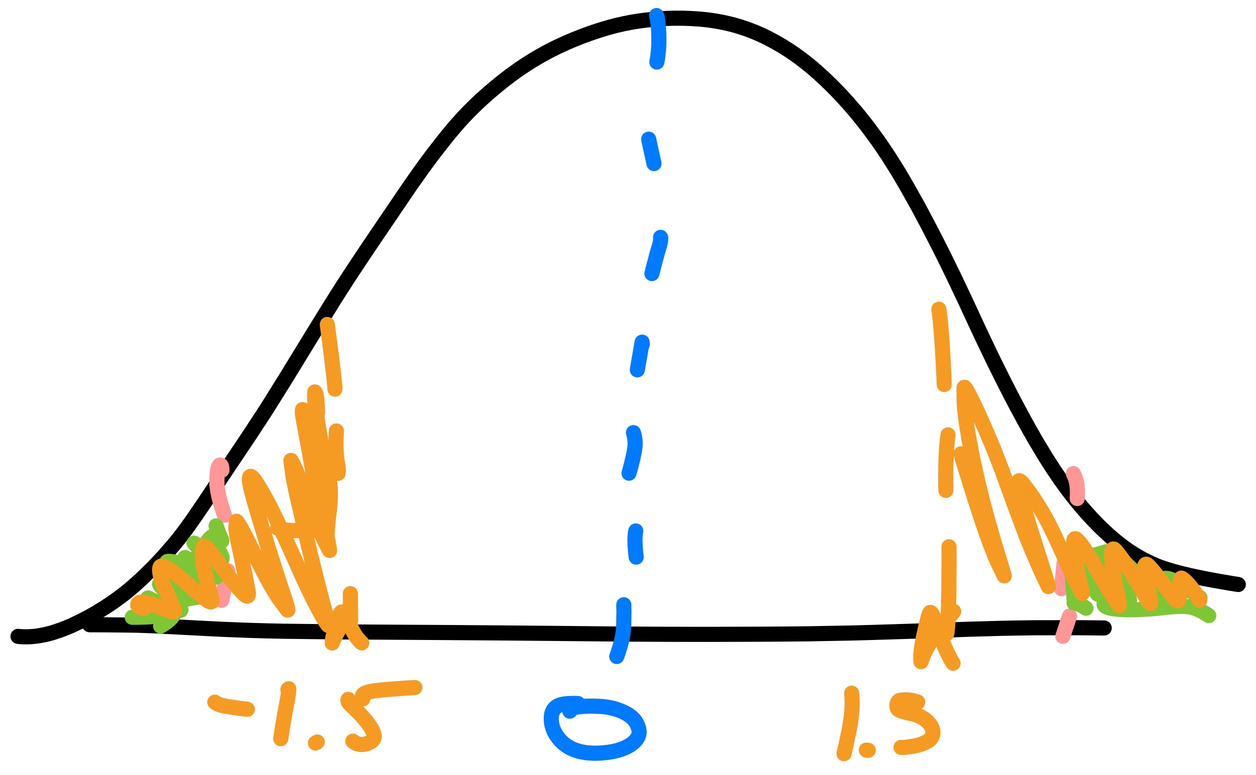
If $Z_n^* < -q_\alpha$, Z_n^* is unlikely from "H_0's distribution" } reject H_0

P-VALUES

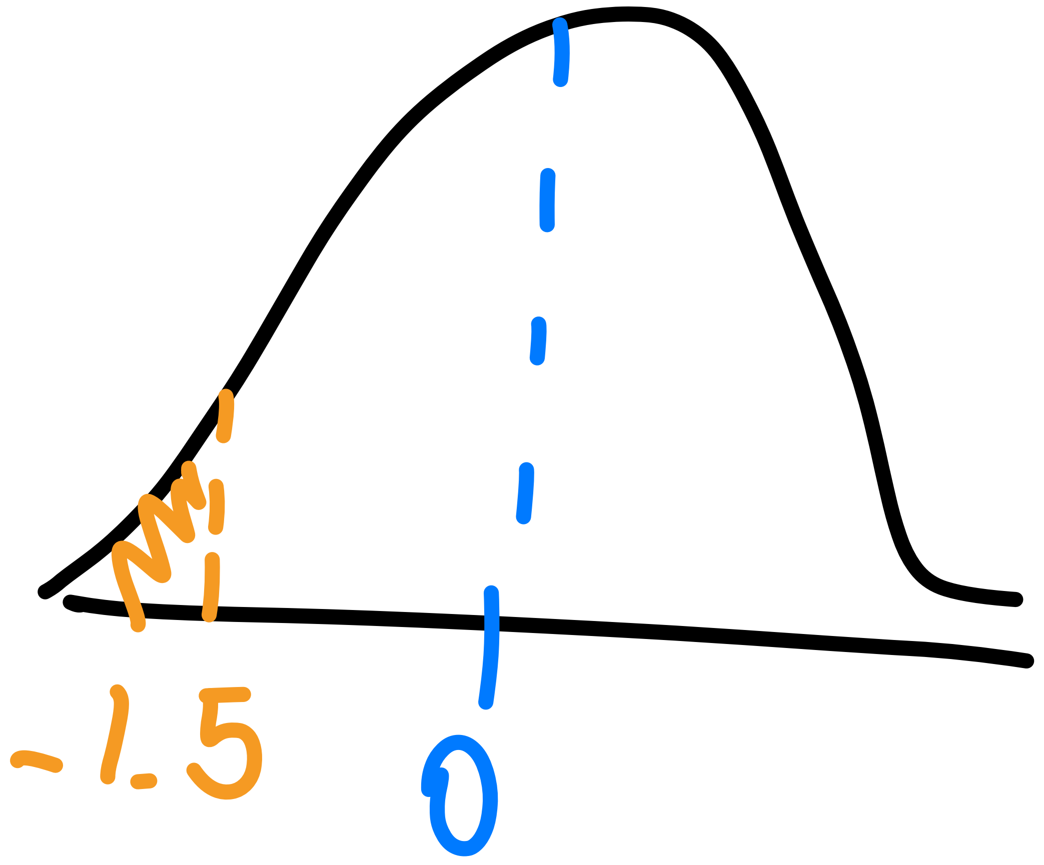
$$H_0 : \theta = 0.8$$
$$H_1 : \theta \neq 0.8$$

$$H_0 : \theta = 0.8$$
$$H_1 : \theta < 0.8$$

$\mathbb{P}(|Z| > z^*)$



$\mathbb{P}(Z < z^*)$



TYPE 1 AND TYPE 2 ERRORS

Type 1 error: Reject H_0 when it is true

Type 2 error: Accept H_0 when H_1 is true

STUDENT TEST

$$X_1, X_2, \cdots, X_n \sim^{iid} \mathcal{N}(\mu, \sigma^2)$$

$\hat{\sigma}^2$?

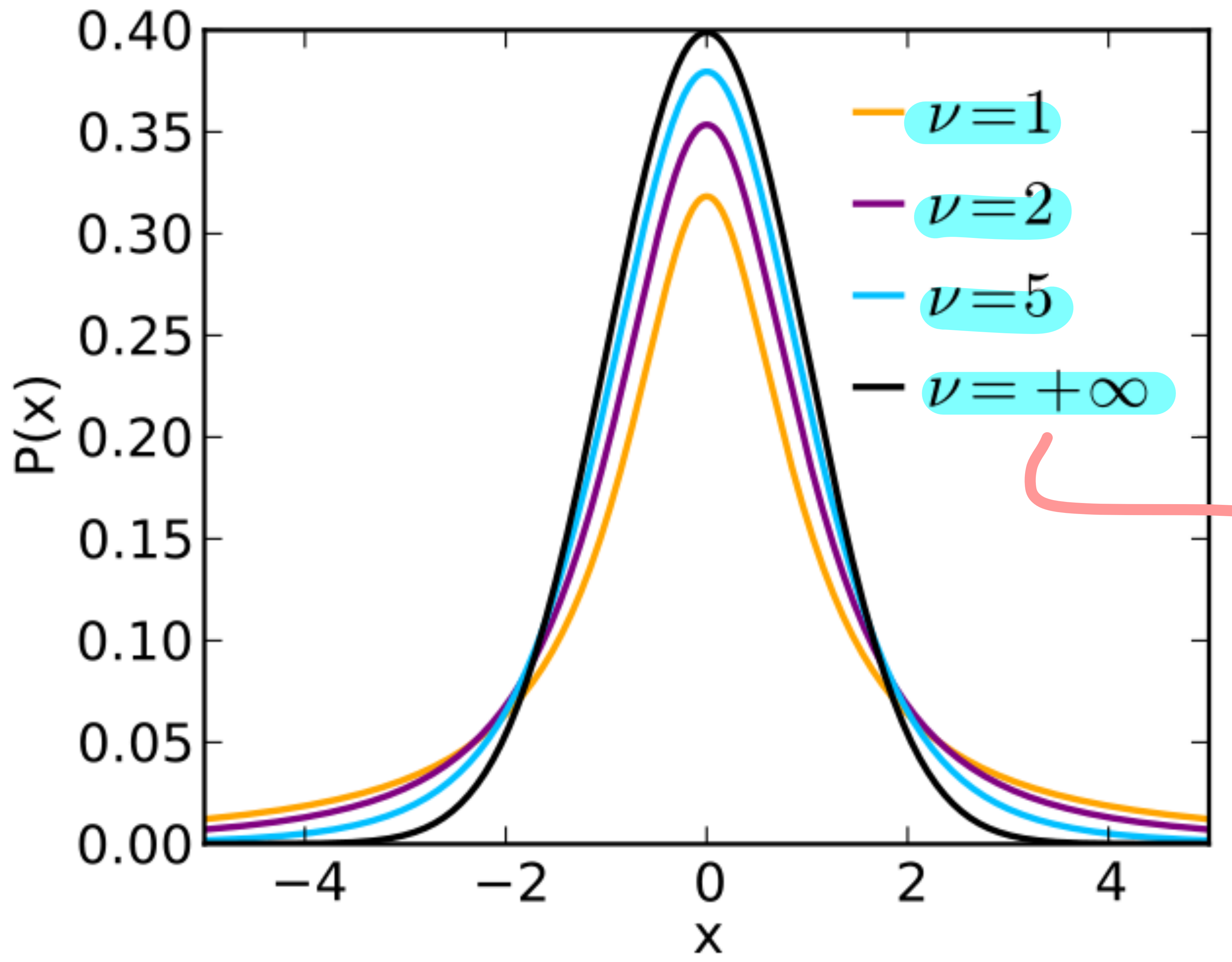
$\hat{\mu} = \overline{X_n}$

$$\hat{S}_n = \frac{\sum_{i=1}^{n} (X_i - \hat{X}_n)^2}{n-1}$$

unbiased estimator

STUDENT TEST

$$\sqrt{n}\,\frac{\hat{X}_n - \mu}{\hat{S}_n} \sim t_{n-1} \qquad \hat{S}_n = \frac{\sum_{i=1}^{n} (X_i - \hat{X}_n)^2}{n-1}$$



gaussian

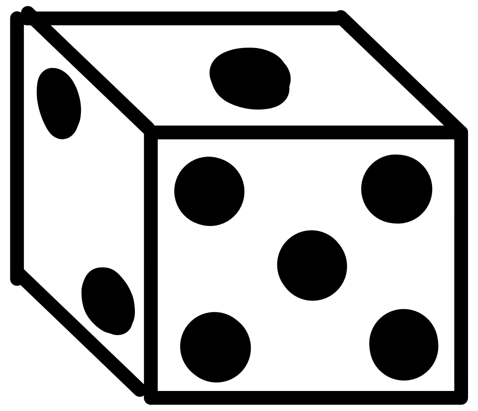
CONFIDENCE INTERVAL

CONFIDENCE INTERVALS

$$\mathbb{P}(|\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}| < 1.96) = 0.95$$

$$\mathbb{P}(\bar{X}_n - \frac{1.96\sigma}{\sqrt{n}} < \mu < \bar{X}_n + \frac{1.96\sigma}{\sqrt{n}}) = 0.95$$

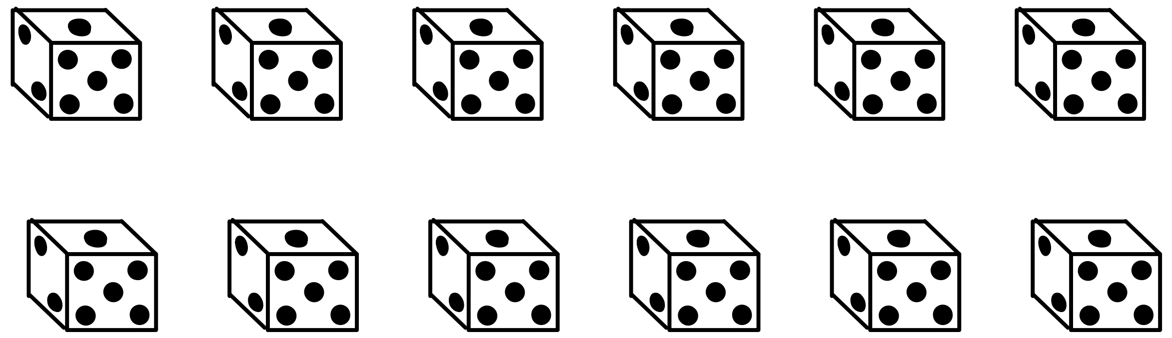
PROBABILISTIC



- Probability to get 1 is 1/6
- Probability to get 2 is 1/6
- …
- Probability to get 6 is 1/6

$$E[X] = 1.\frac{1}{6} + 2.\frac{1}{6} + 3.\frac{1}{6} + 4.\frac{1}{6} + 5.\frac{1}{6} + 6.\frac{1}{6} = 3.5$$

EMPIRICAL



- 1, 1, 2, 3, 3, 4, 4, 5, 5, 6, 6, 6

$$E[X] = \frac{1+1+2+3+3+4+4+5+5+6+6+6}{12}$$

$$E[X] = 1.\frac{1}{6} + 2.\frac{1}{12} + 3.\frac{1}{6} + 4.\frac{1}{6} + 5.\frac{1}{6} + 6.\frac{3}{12} = 3.8$$

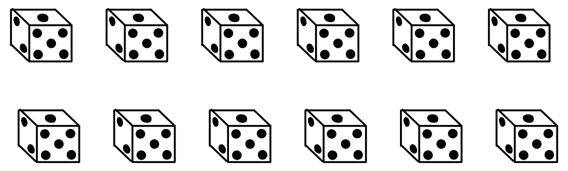
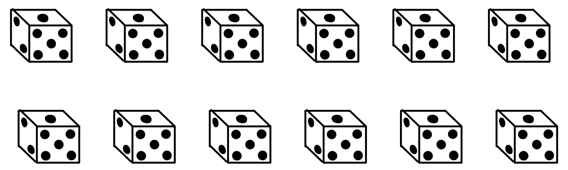
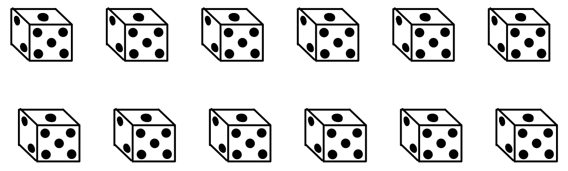
CONFIDENCE INTERVAL

- Empirical result: 3.8
- 95% Confidence Interval

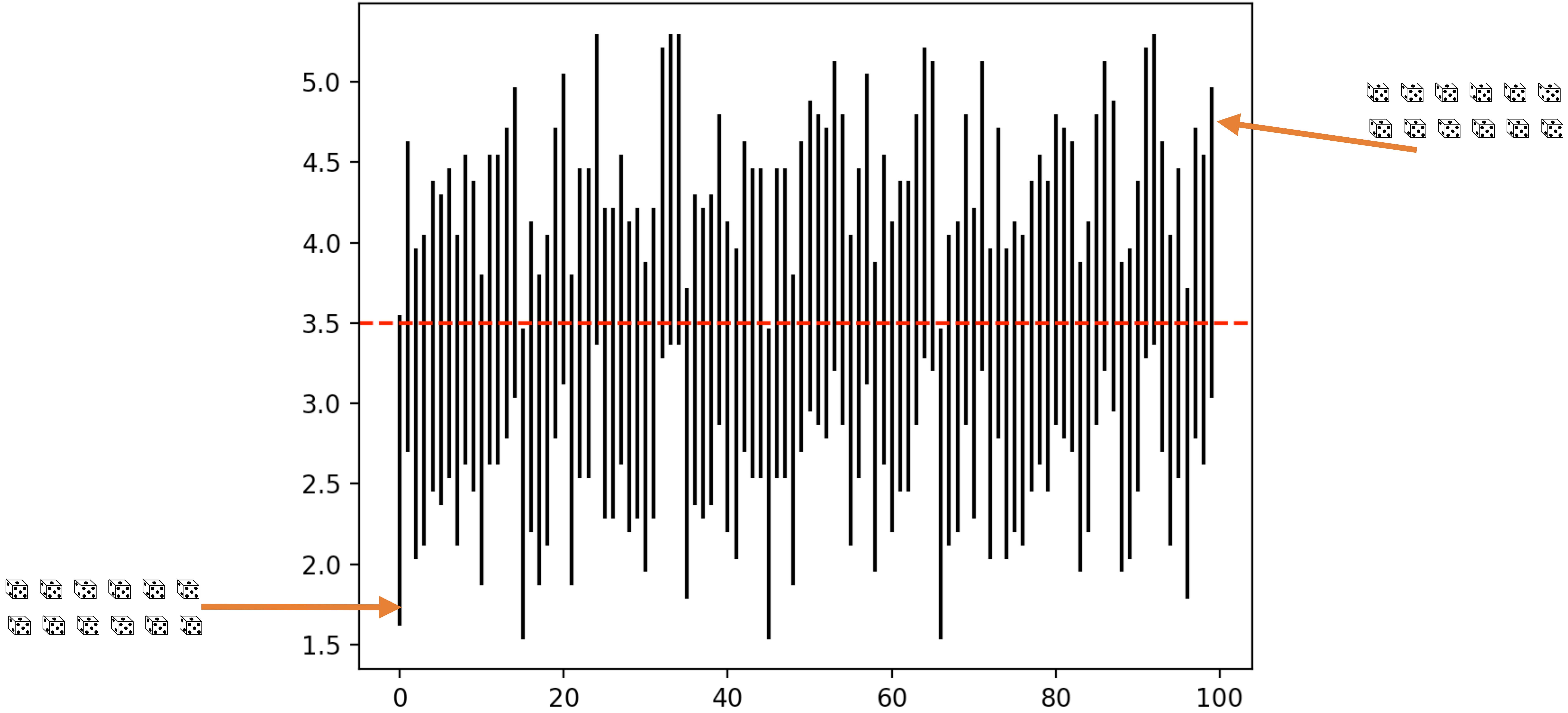
$$CI = [\bar{X}_n - \frac{1.96s}{\sqrt{n}}, \bar{X}_n + \frac{1.96s}{\sqrt{n}}]$$

- CI = [2.05, 5.35]
- What does 95% interval mean?

CONFIDENCE INTERVAL

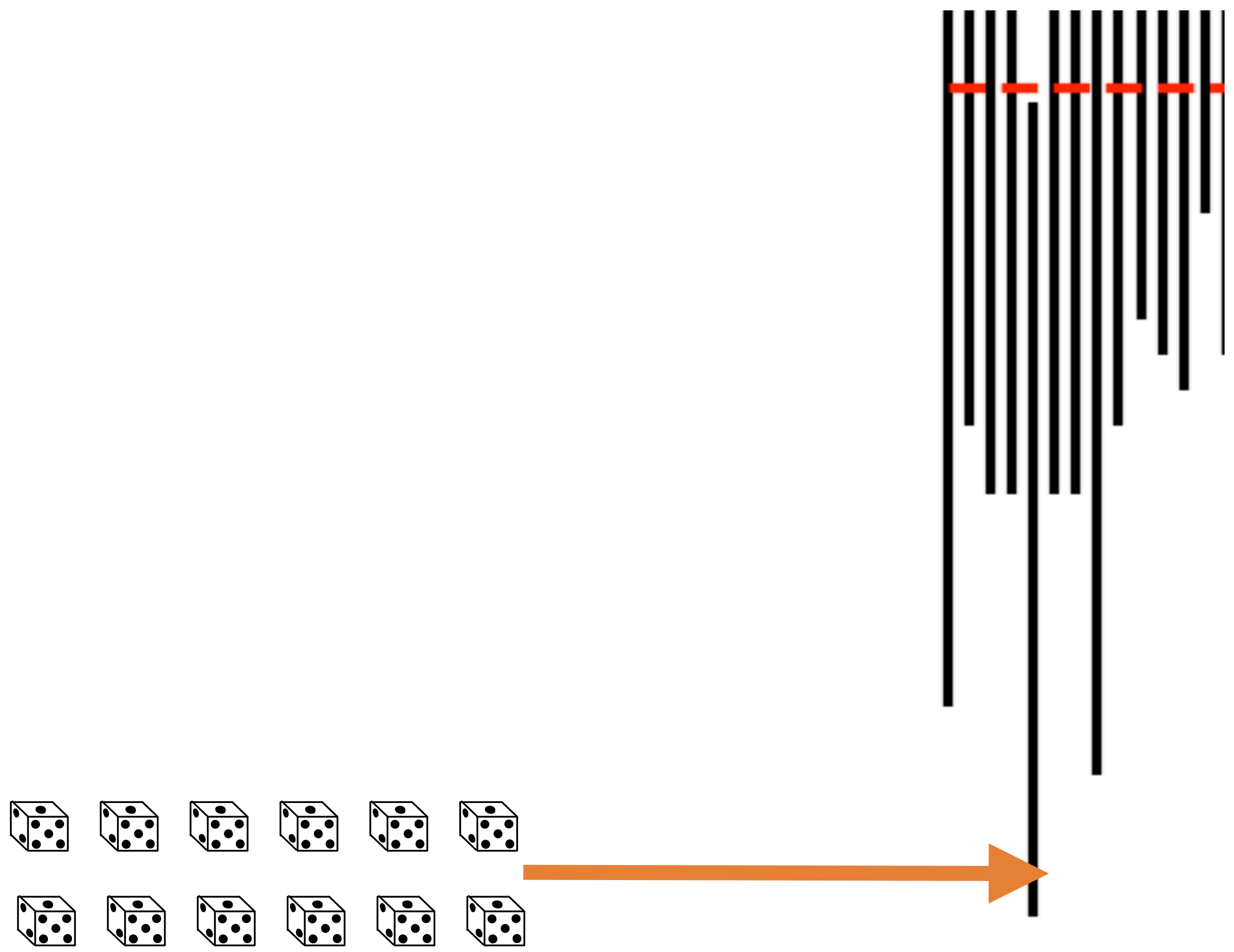
- I launch 100 times 12 dices
- 1.  I compute CI_1
- 2.  I compute CI_2
- …
- 100.  I compute CI_{100}
- My TRUE parameter shall be in at least 95 of these 100 CIs.

CONFIDENCE INTERVAL



CONFIDENCE INTERVAL

- Remember, in real life cases, we do NOT know the TRUE parameter. In fact, the point of the CI is to allow us to estimate this TRUE parameter.
- If the hypothesis is that the TRUE parameter is 3.5, it can be verified looking at the CI.
- There is a 0.05 risk to make an error.



PROBABILISTIC

- I reject the null hypothesis if my observation seems implausible, meaning the probability that it is a Gaussian is less than 5%.

EMPIRICAL

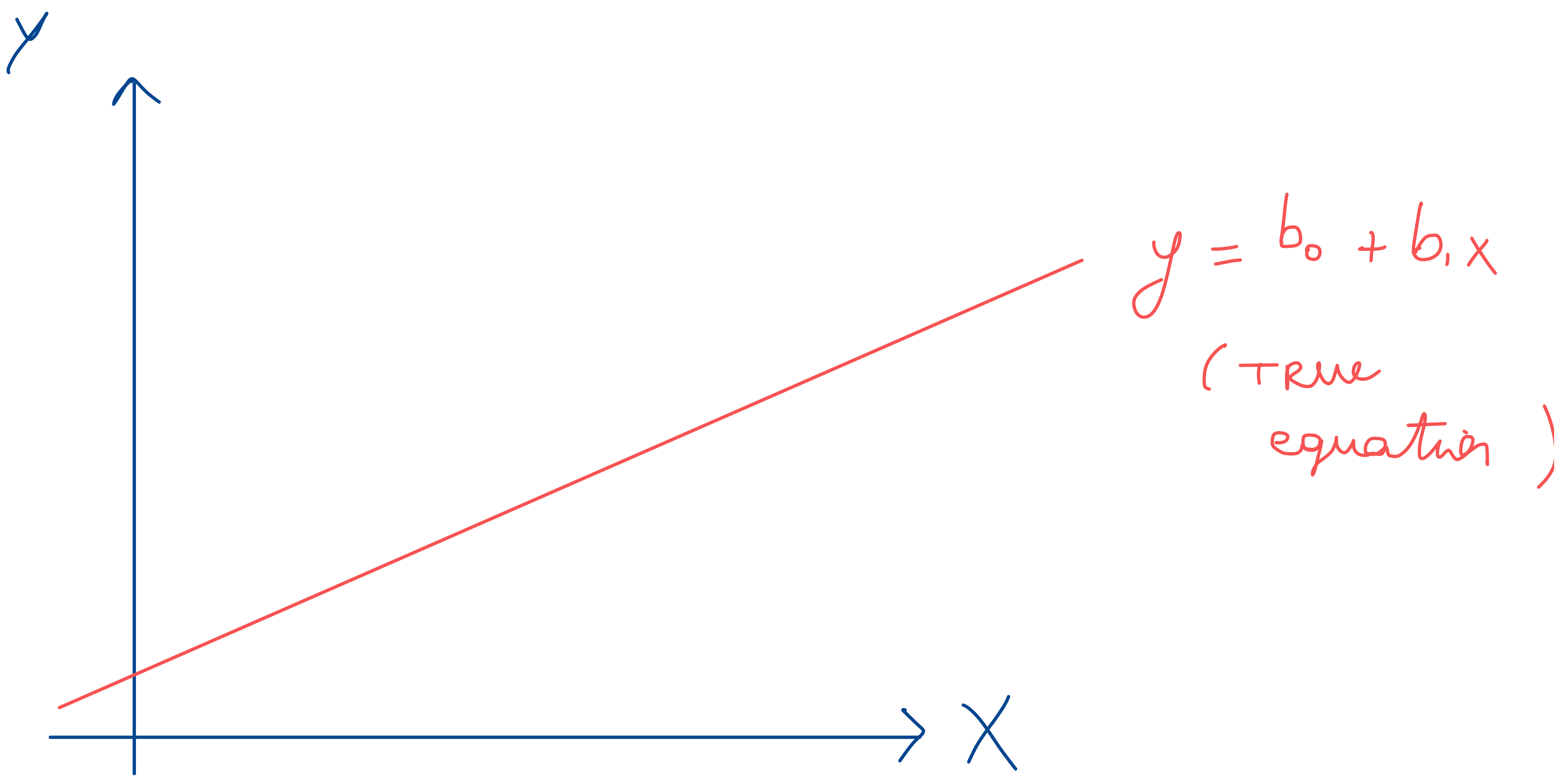
- I reject the null hypothesis if my observation does not belong to the confidence interval.

LINEAR REGRESSION

SIMPLE REGRESSION ANALYSIS

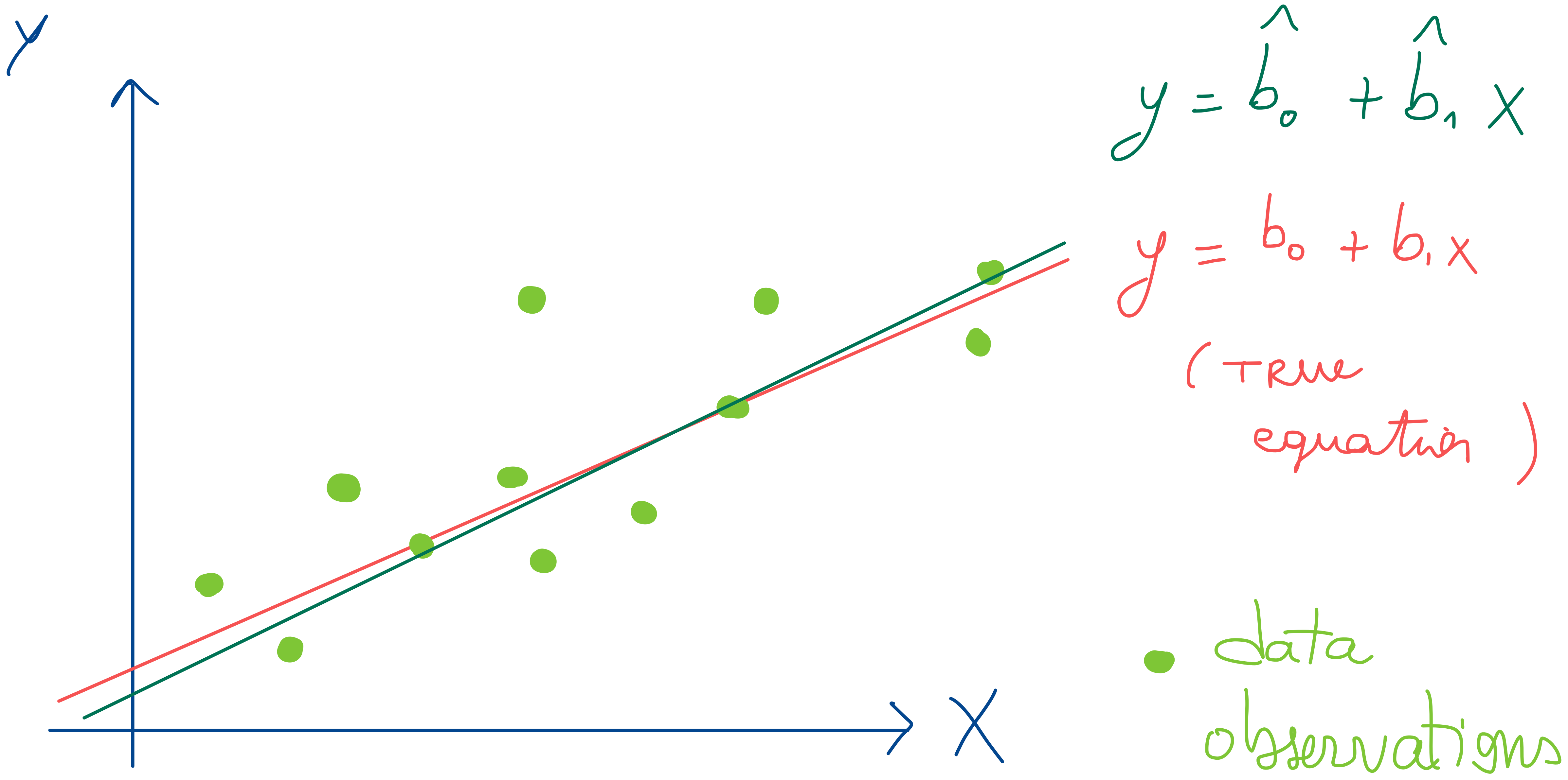
- Goal: To develop a model that relates two quantities
- x: **Independent** (explanatory) variable; quantity sometimes under managerial control
- Y: **Dependent** variable; quantity to be predicted — magnitude is determined (in large part) by x

PROBABILISTIC WORLD



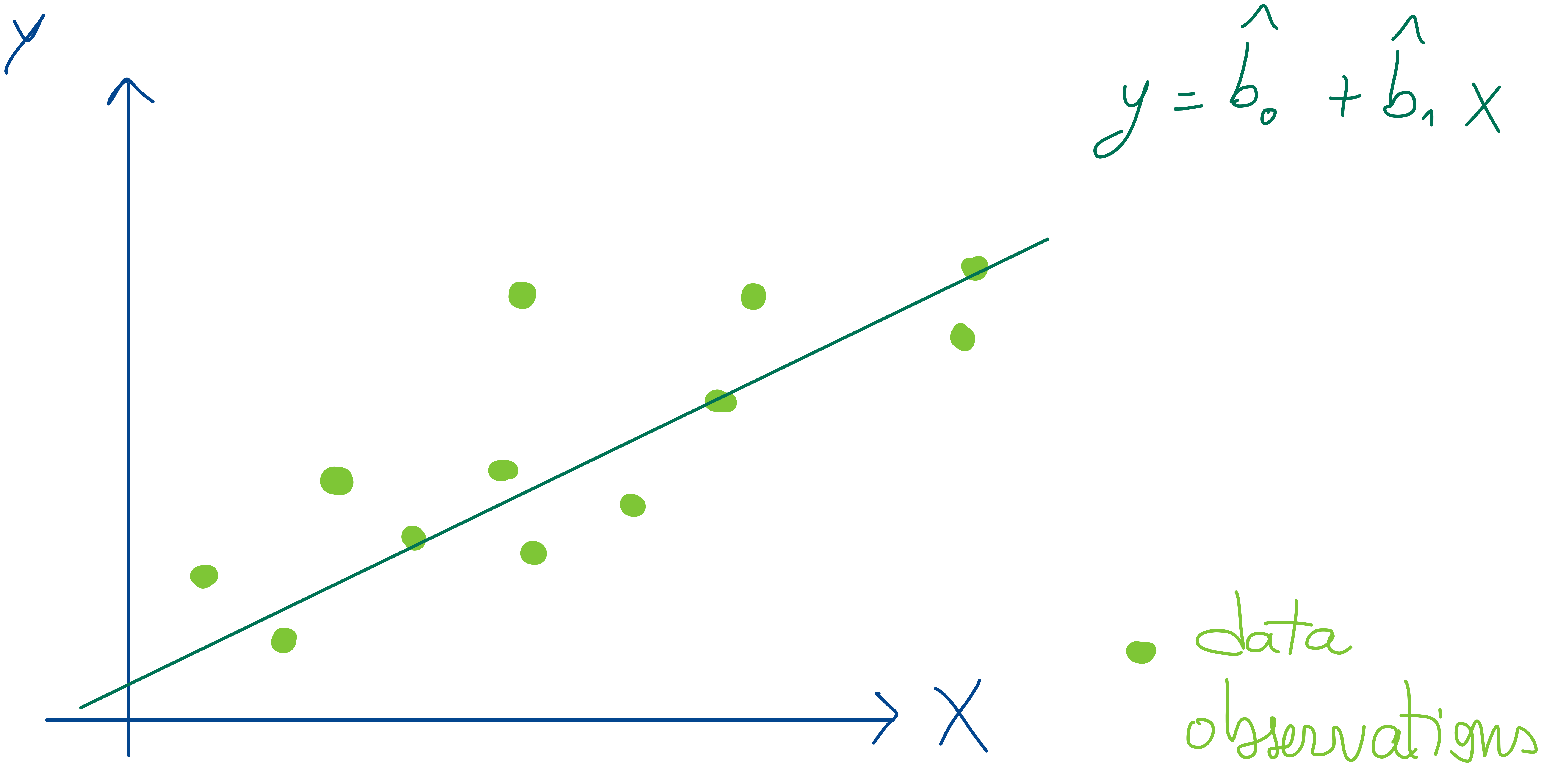
$$Y_i = b_0 + b_1 X_i + \epsilon_i$$

EMPIRICAL WORLD



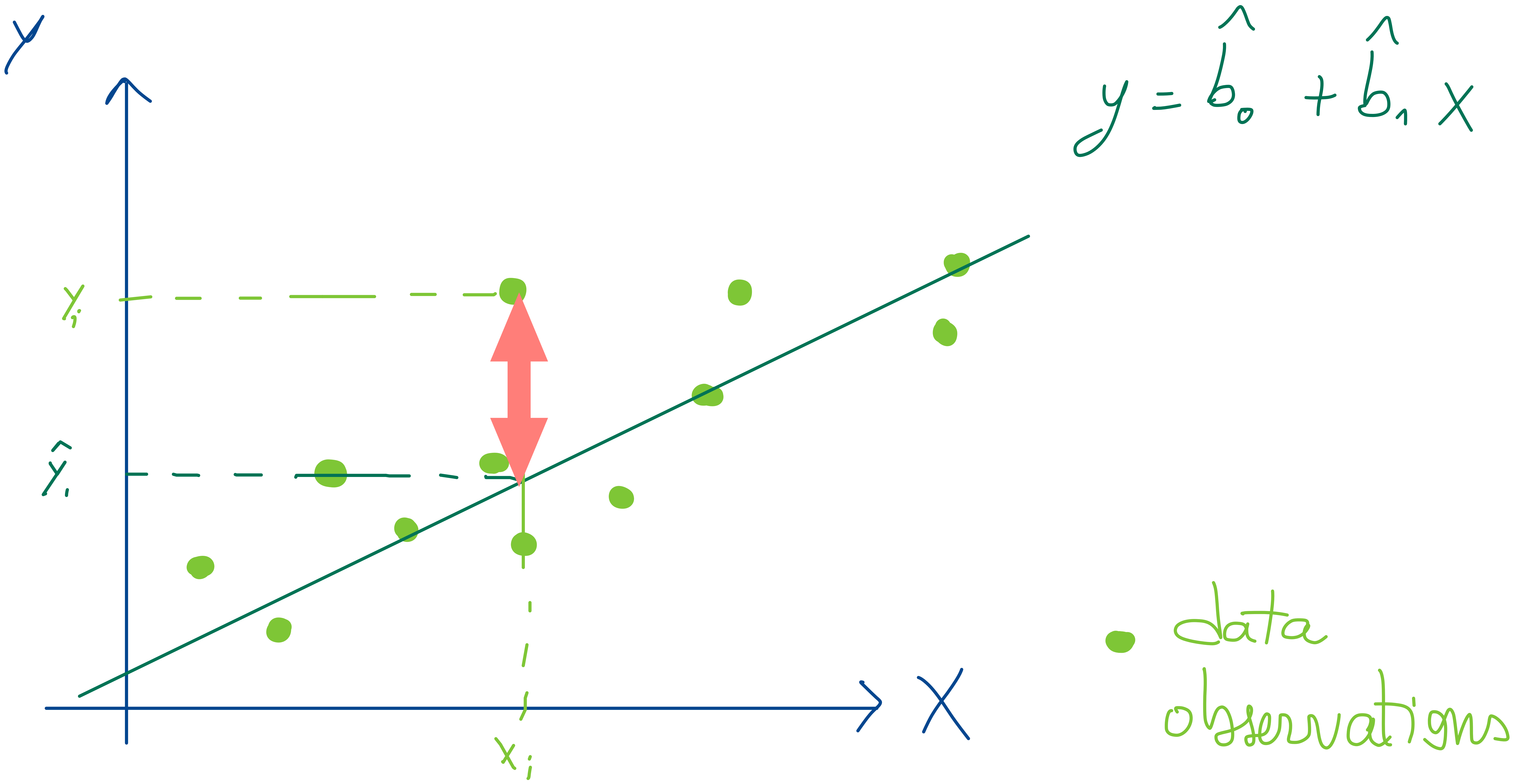
$$Y_i = \hat{b}_0 + \hat{b}_1 X_i + \hat{\epsilon}_i$$

LINEAR REGRESSION



➤ Estimated values! Hypothesis testing? Confidence interval?

PARAMETERS ESTIMATION



$$e_i = y_i - \hat{y}_i$$

PARAMETERS ESTIMATION

- Residuals

$$e_i = y_i - \hat{y}_i$$

- Sum squared of residuals

$$SSE = \sum_{i}^{n} e_i^2 = \sum_{i}^{n} (y_i - \hat{y}_i)^2$$

- Minimize SSE with b_0 and b_1
- Wait… where are b_0 and b_1?

$$\hat{y}_i = b_0 + b_1 x_i$$

PARAMETERS ESTIMATION

$$SSE = \sum_{i}^{n} (y_i - b_0 - b_1 x_i)^2$$

observed data

MODEL

$$y_i = b_0 + b_1 x_i + \epsilon_i$$

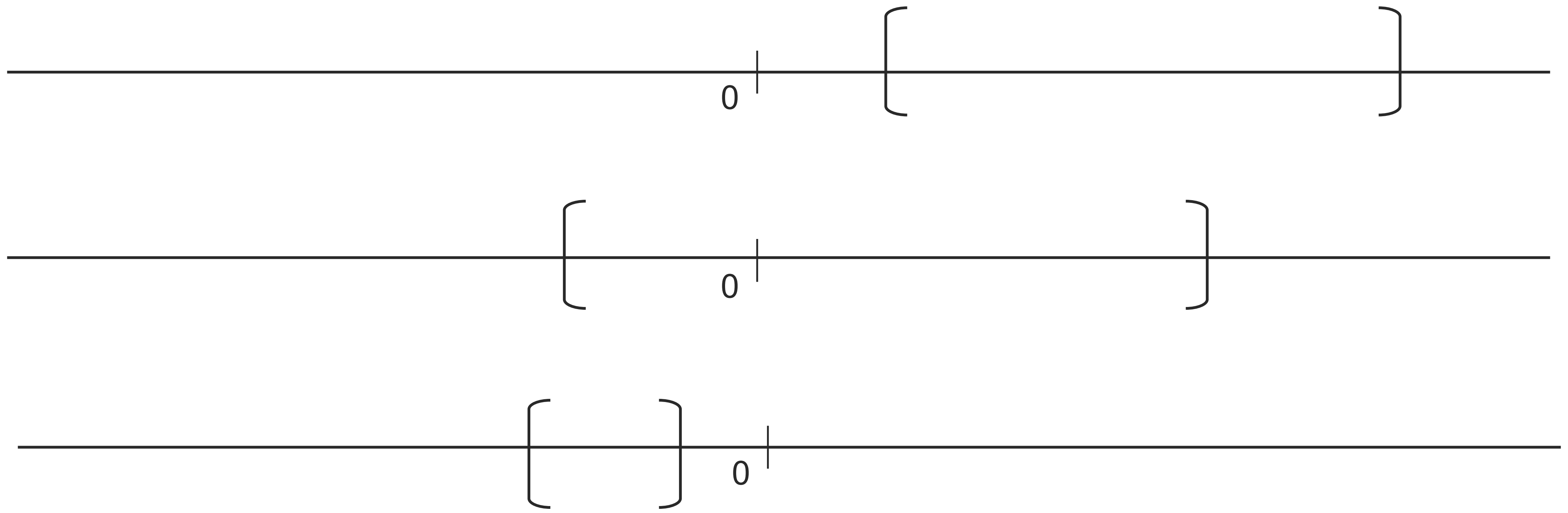
INTERPRETATION

- What does it mean for b_1 to be 0?
- Let: H0 (null hypothesis) $b_1 = 0$
- How can we reject H0 based on the data?
- Compute the Confidence Intervals!

Reject H0

Fait to Reject H0

Reject H0



EXAMPLE

CANDY BAGS

- A candy manufacturer at MIT produces Bertie Bott's Every Flavour Beans. Each jelly bag contains candies that ALL have the same flavor (hence color). Further, the manufacturer claims that each bag contains at least 49 candies.
- Among 50 candy bags, the students found that there is between 41 and 55 candies, and 47.76 on average, with a $\sigma = 4.42$
- Do you think that the manufacturer's claim is fair?

CANDY BAGS

$n = 50$, $X_1, \dots, X_n \overset{iid}{\sim} X$ with $E(X) = \mu$

$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i = 47.76$ $\sigma = 4.42$

$$\begin{cases} H_0: \mu \geq 49 \\ H_1: \mu < 49 \end{cases}$$

claim from manufacturer

one-sided test

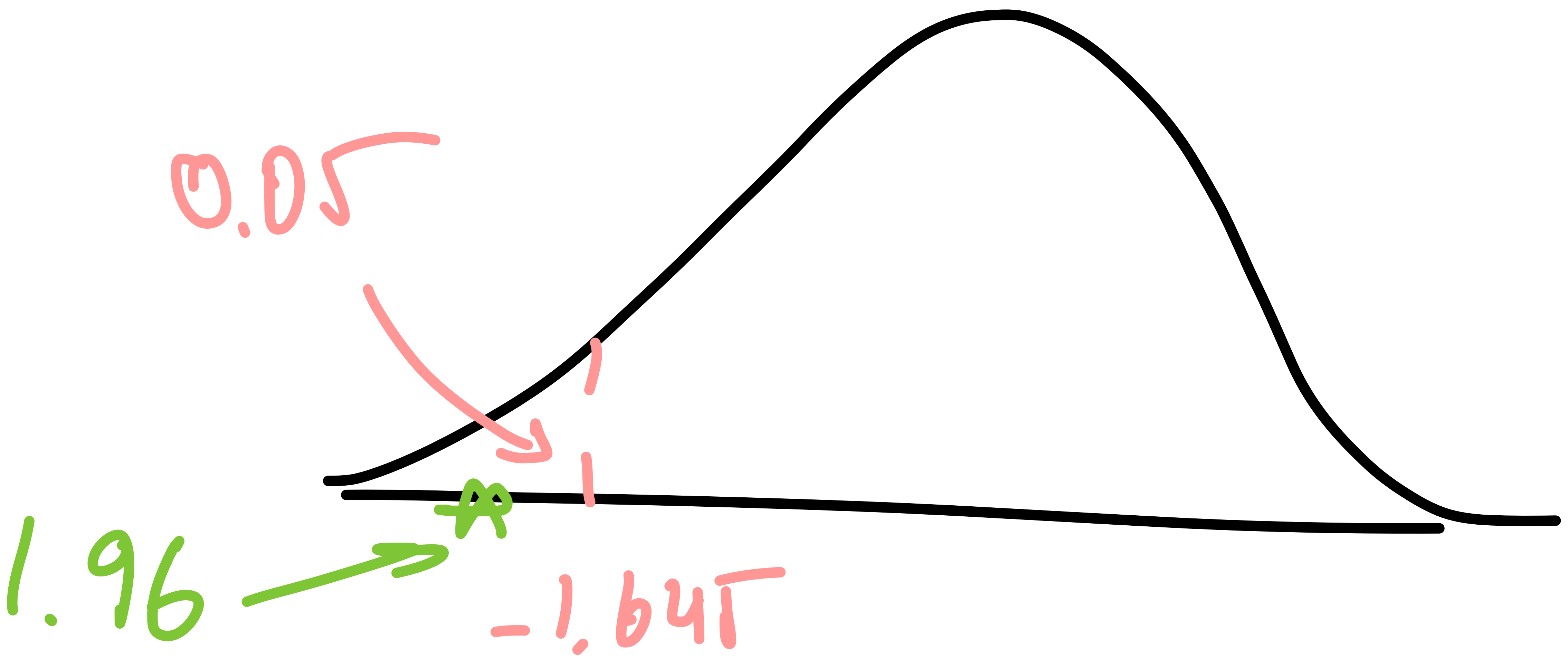
alternative possibility

CANDY BAGS

$$\underbrace{\sqrt{n}\,\frac{\bar{X}_n-\mu}{\sigma}}_{=-1.98} \sim \mathcal{N}(0,1)$$

$$\mathbb{P}(Z \le -1.98) = 0.024$$

$\mathbb{P}$ for a gaussian to be more extreme that the test value



CANDY BAGS

The probability that the test value is from the distribution induced by H_0 is less than 5%.

We then REJECT H_0.