

# 18.650 – Fundamentals of Statistics

## 2. Foundations of Inference

# Goals

In this unit, we introduce a mathematical formalization of statistical modeling to make a principled sense of the **Trinity of statistical inference**.

We will make sense of the following statements:

1. Estimation:

*" $\hat{p} = \overline{R_n}$  is an estimator for the proportion  $p$  of couples that turn their head to the right"*

(side question: is 64.5% also an estimator for  $p$ ?)

2. Confidence intervals:

*" $[0.56, 0.73]$  is a 95% confidence interval for  $p$ "*

3. Hypothesis testing:

*"We find statistical evidence that more couples turn their head to the right when kissing"*

# The rationale behind statistical modeling

- ▶ Let  $X_1, \dots, X_n$  be  $n$  independent copies of  $X$ .
- ▶ The goal of statistics is to learn the distribution of  $X$ .
- ▶ If  $X \in \{0, 1\}$ , easy! It's *Bernoulli* and we only have to learn the parameter  $p$
- ▶ Can be more complicated. For example, here is a (partial) dataset with number of siblings (including self) that were collected from college students a few years back: 2, 3, 2, 4, 1, 3, 1, 1, 1, 1, 1, 2, 2, 3, 2, 2, 2, 3, 2, 1, 3, 1, 2, 3, ...
- ▶ We could make no assumption and try to learn the pmf:

$x$	1	2	3	4	5	6	$\geq 7$
$\mathbb{P}(X = x)$	$p_1$	$p_2$	$p_3$	$p_4$	$p_5$	$p_6$	$\sum_{i \geq 7} p_i$

That's 7 parameters to learn.

$$\mathbb{P}(X=x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

- ▶ Or we could assume that  $X - 1 \sim \text{Pois}(\lambda)$ . That's 1 parameter to learn!

# Statistical model

## Formal definition

Let the observed outcome of a statistical experiment be a *sample*  $X_1, \dots, X_n$  of  $n$  i.i.d. random variables in some measurable space  $E$  (usually  $E \subseteq \mathbb{R}$ ) and denote by  $\mathbb{P}$  their common distribution. A *statistical model* associated to that statistical experiment is a pair

$$(E, (\mathbb{P}_\theta)_{\theta \in \Theta}),$$

where:

- ▶  $E$  is called *sample space*
- ▶  $(\mathbb{P}_\theta)_{\theta \in \Theta}$  is a family of *probability* measures on  $E$ ;
- ▶  $\Theta$  is any set, called *parameter set*

# Parametric, nonparametric and semiparametric models

- ▶ Usually, we will assume that the statistical model is *well specified*, i.e., defined such that  $\exists \theta$  such that  $P = P_\theta$
- ▶ This particular  $\theta$  is called the *true parameter*, and is unknown: The aim of the statistical experiment is to  $\theta$ , or check it's properties when they have a special meaning ( $\theta > 2?$ ,  $\theta \neq 1/2?$ , ...)
- ▶ We often assume that  $\Theta \subseteq \mathbb{R}^d$  for some  $d \geq 1$ : The model is called *parametric*
- ▶ Sometime we could have  $\Theta$  be infinite dimensional in which case the model is called *nonparametric*
- ▶ If  $\Theta = \Theta_1 \times \Theta_2$ , where  $\Theta_1$  is finite dimensional and  $\Theta_2$  is infinite dimensional: *semiparametric* model. In these models we only care to estimate the finite dimensional parameter and the infinite dimensional one is called *nuisance parameter*. We will not cover such models in this class.

( $P, f$ )

# Examples of parametric models

1. For  $n$  Bernoulli trials:

$$\left( \{0, 1\}, (\text{Ber}(p))_{p \in (0,1)} \right).$$

2. If  $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Poiss}(\lambda)$  for some unknown  $\lambda > 0$ ,

$$\left( \mathbb{N}, (\text{Poiss}(\lambda))_{\lambda > 0} \right).$$

3. If  $X_1, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ , for some unknown  $\mu \in \mathbb{R}$  and  $\sigma^2 > 0$ :

$$\left( \mathbb{R}, (\mathcal{N}(\mu, \sigma^2))_{(\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)} \right).$$

4. If  $X_1, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}_d(\mu, I_d)$ , for some unknown  $\mu \in \mathbb{R}^d$ :

$$\left( \mathbb{R}^d, (\mathcal{N}_d(\mu, I_d))_{(\mu \in \mathbb{R}^d)} \right).$$

# Examples of nonparametric models

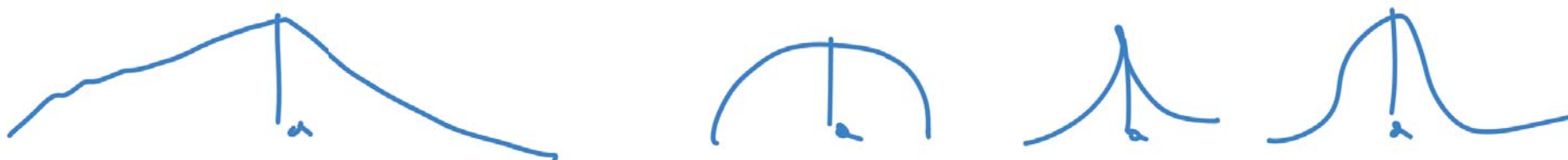
1. If  $X_1, \dots, X_n \in \mathbb{R}$  are i.i.d with unknown *unimodal*<sup>1</sup> pdf  $f$ :

$$E = \mathbb{R} \quad \Theta = \{ \text{unimodal pdf } f \}$$

$$P_\theta = P_f = \text{distribution with pdf } f$$

2. If  $X_1, \dots, X_n \in [0, 1]$  are i.i.d with unknown invertible cdf  $F$ .

$$E = [0, 1] \quad \dots$$



<sup>1</sup>Increases on  $(-\infty, a)$  and then decreases on  $(a, \infty)$  for some  $a > 0$ .

## Further examples

Sometimes we do not have simple notation to write  $(\mathbb{P}_\theta)_{\theta \in \Theta}$ , e.g.,  $(\text{Ber}(p))_{p \in (0,1)}$  and we have to be more explicit:

### 1. Linear regression model: If

$(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$  are i.i.d from the linear regression model  $Y_i = \beta^\top X_i + \varepsilon_i$   $\varepsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, 1)$  for an unknown  $\beta \in \mathbb{R}^d$  and  $X_i \sim \mathcal{N}_d(0, I_d)$  independent of  $\varepsilon_i$

$$E = \mathbb{R}^d \times \mathbb{R}$$

$$\Theta = \mathbb{R}^d$$

### 2. Cox proportional Hazard model: If

$(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$  : the conditional distribution of  $Y$  given  $X = x$  has CDF  $F$  of the form

$$F(t) = 1 - \exp \left( - \int_0^t h(u) e^{(\beta^\top x)} du \right)$$

where  $h$  is an unknown non-negative nuisance function and  $\beta \in \mathbb{R}^d$  is the parameter of interest.



# Identifiability

The parameter  $\theta$  is called *identifiable* iff the map  $\theta \in \Theta \mapsto \mathbb{P}_\theta$  is injective, i.e.,

$$\theta \neq \theta' \Rightarrow \mathbb{P}_\theta \neq \mathbb{P}_{\theta'}$$

or equivalently:

$$\mathbb{P}_\theta = \mathbb{P}_{\theta'} \Rightarrow \theta = \theta'$$

## Examples

1. In all previous examples, the parameter is identifiable.
2. If  $X_i = \mathbb{I}_{Y_i \geq 0}$  (indicator function),  $Y_1, \dots, Y_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ , for some unknown  $\mu \in \mathbb{R}$  and  $\sigma^2 > 0$ , are unobserved:  $\mu$  and  $\sigma^2$  are not identifiable (but  $\theta = \mu/\sigma$  is).

# Exercises

a) Which of the following is a statistical model?

1.  $\left(\{1\}, (\text{Ber}(p))_{p \in (0,1)}\right)$

2.  $\left(\{0, 1\}, (\text{Ber}(p))_{p \in (0.2, 0.4)}\right)$  ✓

3. Both 1 and 2

4. None of the above

← uniform distribution on  $[0, a]$

b) Let  $X_1, \dots, X_n \stackrel{iid}{\sim} \mathcal{U}([0, a])$  for some unknown  $a > 0$ . Which one of the following is the associated statistical model?

1.  $\left([0, a], (\mathcal{U}([0, a]))_{a>0}\right)$

2.  $\left(\mathbb{R}_+, (\mathcal{U}([0, a]))_{a>0}\right)$  ✓

3.  $\left(\mathbb{R}, (\mathcal{U}([0, a]))_{a>0}\right)$

4. None of the above

# Exercises

c) Let  $X_i = Y_i^2$ , where  $Y_1, \dots, Y_n \stackrel{iid}{\sim} \mathcal{U}([0, a])$ , for some unknown  $a$ , are unobserved. Is  $a$  identifiable?

1. Yes 

2. No

d) Let  $X_i = \mathbb{I}_{Y_i \geq \frac{a}{2}}$ , where  $Y_1, \dots, Y_n \stackrel{iid}{\sim} \mathcal{U}([0, a])$ , for some unknown  $a$ , are unobserved. Is  $a$  identifiable?

1. Yes

2. No 

# Estimation

# Parameter estimation

## Definitions

- ▶ **Statistic**: Any measurable<sup>2</sup> function of the sample, e.g.,  $\bar{X}_n, \max_i X_i, X_1 + \log(1 + |X_n|)$ , sample variance, etc...
- ▶ **Estimator** of  $\theta$ : Any statistic whose expression does not depend on  $\theta$
- ▶ An estimator  $\hat{\theta}_n$  of  $\theta$  is *weakly* (resp. *strongly*) *consistent* if

$$\hat{\theta}_n \xrightarrow[n \rightarrow \infty]{\mathbb{P} \text{ (resp. a.s.)}} \theta \quad (\text{w.r.t. } \mathbb{P}_\theta).$$

- ▶ An estimator  $\hat{\theta}_n$  of  $\theta$  is *asymptotically normal* if

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(\theta, \sigma^2)$$

The quantity  $\sigma^2$  is then called *asymptotic variance* of  $\hat{\theta}_n$ .

---

<sup>2</sup>Rule of thumb: if you can compute it exactly once given data, it is measurable. You may have some issues with things that are implicitly defined

# Bias of an estimator

- Bias of an estimator  $\hat{\theta}_n$  of  $\theta$ :

$$\text{bias}(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n] - \theta$$

- If  $\text{bias}(\hat{\theta}) = 0$ , we say that  $\hat{\theta}$  is *unbiased*
- Example: Assume that  $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Ber}(p)$  and consider the following estimators for  $p$ :

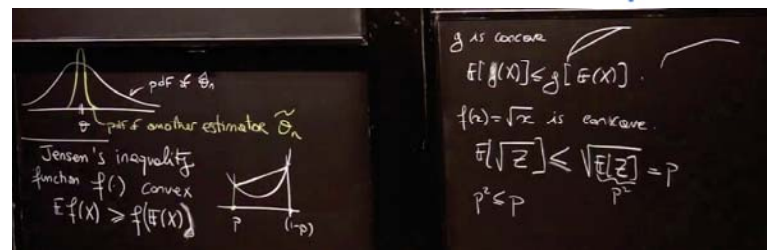
$$\text{Bias}(\hat{p}_n) = \mathbb{E}[\hat{p}_n] - p$$

- $\hat{p}_n = \bar{X}_n$ :  $\text{bias}(\hat{p}_n) = 0$

- $\hat{p}_n = X_1$ :  $\text{bias}(\hat{p}_n) = 0$

- $\hat{p}_n = \frac{X_1 + X_2}{2}$ :  $\text{bias}(\hat{p}_n) = 0$

- $\hat{p}_n = \sqrt{\mathbb{I}(X_1 = 1, X_2 = 1)}$   
 $Z \sim \text{Ber}(p^2)$



$$\hat{p}_n \sim \text{Ber}(p^2) \Rightarrow \text{bias}(\hat{p}_n) = p^2 - p$$

# Variance of an estimator

$$\text{var}(X) = E[(X - E(X))^2] = E[X^2] - (E[X])^2$$

An estimator is a random variable so we can compute its variance.

In the previous examples:

►  $\hat{p}_n = \bar{X}_n$ :  $\text{var}(\hat{p}_n) = \frac{p(1-p)}{n}$

►  $\hat{p}_n = X_1$ :  $\text{var}(\hat{p}_n) = p(1-p)$

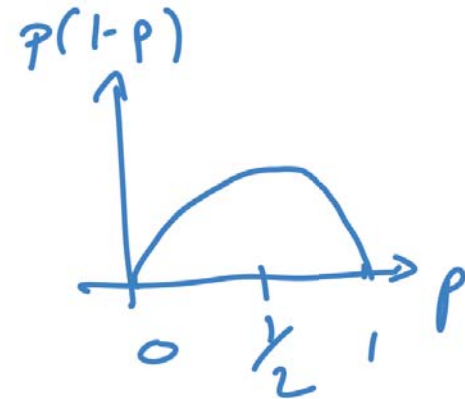
►  $\hat{p}_n = \frac{X_1 + X_2}{2}$ :  $\text{var}(\hat{p}_n) = \frac{p(1-p)}{2}$

►  $\hat{p}_n = \sqrt{\mathbb{I}(X_1 = 1, X_2 = 2)}$

$$\hat{p}_n \sim \text{Ber}(p^2)$$

$$\text{var}(\hat{p}_n) = p^2(1-p^2)$$

$$X \sim \text{Ber}(p)$$
$$\text{var}(X) = p(1-p)$$



# Quadratic risk

- ▶ We want estimators to have low bias and low variance at the same time.
- ▶ The *Risk* (or *quadratic risk*) of an estimator  $\hat{\theta}_n \in \mathbb{R}$  is

$$R(\hat{\theta}_n) = \mathbb{E} \left[ |\hat{\theta}_n - \theta|^2 \right]$$

*Handwritten derivation:*

$$\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n] + \mathbb{E}[\hat{\theta}_n] - \theta)^2] = \underbrace{\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^2]}_{\text{Var}(\hat{\theta}_n)} + \underbrace{\mathbb{E}[(\mathbb{E}[\hat{\theta}_n] - \theta)^2]}_{\text{bias}(\hat{\theta}_n)^2} + 2 \underbrace{\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])(\mathbb{E}[\hat{\theta}_n] - \theta)]}_{=0}$$

- ▶ Low quadratic risk means that both bias and variance are small:

quadratic risk =  $\text{VARIANCE} + \text{BIAS}^2$



# Exercises

Let  $X_1, X_2, \dots, X_n$  be a random sample from  $\mathcal{U}([a, a + 1])$ . Questions a), b), c) and d) are about this sample.

a) Find  $\mathbb{E}[\bar{X}_n] = a + \frac{1}{2}$

b) Is  $\bar{X}_n - \frac{1}{2}$  an unbiased estimator for  $a$ ? Yes :  $\mathbb{E}[\bar{X}_n - \frac{1}{2}] = a$

c) Find the variance of  $\bar{X}_n - \frac{1}{2}$ .  $\text{Var}(\bar{X}_n - \frac{1}{2}) = \frac{1}{12n}$

d) Find the quadratic risk of  $\bar{X}_n - \frac{1}{2}$ .

$$R(\bar{X}_n - \frac{1}{2}) = \frac{1}{12n} + 0^2 = \frac{1}{12n}$$

# Confidence intervals

# Confidence intervals

Let  $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$  be a statistical model based on observations  $X_1, \dots, X_n$ , and assume  $\Theta \subseteq \mathbb{R}$ . Let  $\alpha \in (0, 1)$ .

- *Confidence interval (C.I.) of level  $1 - \alpha$  for  $\theta$* : Any random (depending on  $X_1, \dots, X_n$ ) interval  $\mathcal{I}$  whose boundaries do not depend on  $\theta$  and such that:

$$\mathbb{P}_\theta [\mathcal{I} \ni \theta] \geq 1 - \alpha, \quad \forall \theta \in \Theta.$$

- *C.I. of asymptotic level  $1 - \alpha$  for  $\theta$* : Any random interval  $\mathcal{I}$  whose boundaries do not depend on  $\theta$  and such that:

$$\lim_{n \rightarrow \infty} \mathbb{P}_\theta [\mathcal{I} \ni \theta] \geq 1 - \alpha, \quad \forall \theta \in \Theta.$$

---

<sup>3</sup> $\mathcal{I} \ni \theta$  means that  $\mathcal{I}$  contains  $\theta$ . This notation emphasizes the randomness of  $\mathcal{I}$  but we can equivalently write  $\theta \in \mathcal{I}$ .

# A confidence interval for the kiss example

- ▶ Recall that we observe  $R_1, \dots, R_n \stackrel{iid}{\sim} \text{Ber}(p)$  for some unknown  $p \in (0, 1)$ .
- ▶ Statistical model:  $\left(\{0, 1\}, (\text{Ber}(p))_{p \in (0, 1)}\right)$ . ✓
- ▶ Recall that our estimator for  $p$  is  $\hat{p} = \bar{R}_n$ . ✓
- ▶ From CLT:

$$\sqrt{n} \frac{\bar{R}_n - p}{\sqrt{p(1-p)}} \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(0, 1)$$

This means (precisely) that:

- ▶  $\Phi(x)$ : cdf of  $\mathcal{N}(0, 1)$ ;  $\Phi_n(x)$ : cdf of  $\sqrt{n} \frac{\bar{R}_n - p}{\sqrt{p(1-p)}}$ .
- ▶ Then:  $\Phi_n(x) \approx \Phi(x)$  (CLT) when  $n$  becomes large. Hence, for all  $x > 0$ ,

$$\mathbb{P} [|\bar{R}_n - p| \geq x] \simeq 2 \left( 1 - \Phi \left( \frac{x\sqrt{n}}{\sqrt{p(1-p)}} \right) \right).$$

# Confidence interval?

- ▶ For a fixed  $\alpha \in (0, 1)$ , if  $q_{\alpha/2}$  is the  $(1 - \alpha/2)$ -quantile of  $\mathcal{N}(0, 1)$ , then with probability  $\simeq 1 - \alpha$  (if  $n$  is large enough !),

$$\bar{R}_n \in \left[ p - \frac{q_{\alpha/2} \sqrt{p(1-p)}}{\sqrt{n}}, p + \frac{q_{\alpha/2} \sqrt{p(1-p)}}{\sqrt{n}} \right].$$

- ▶ It yields

$$\lim_{n \rightarrow \infty} \mathbb{P} \left( \left[ \bar{R}_n - \frac{q_{\alpha/2} \sqrt{p(1-p)}}{\sqrt{n}}, \bar{R}_n + \frac{q_{\alpha/2} \sqrt{p(1-p)}}{\sqrt{n}} \right] \ni p \right) = 1 - \alpha$$

- ▶ But this is **not** a confidence interval because *it depends on p!*

- ▶ To fix this, there are 3 solutions.

## Solution 1: Conservative bound

- Note that no matter the (unknown) value of  $p$ ,

$$p(1 - p) \leq \frac{1}{4}$$

- Hence, roughly with probability at least  $1 - \alpha$ ,

$$\bar{R}_n \in \left[ p - \frac{q_{\alpha/2}}{2\sqrt{n}}, p + \frac{q_{\alpha/2}}{2\sqrt{n}} \right].$$

- We get the asymptotic confidence interval:

$$\mathcal{I}_{\text{conserv}} = \left[ \bar{R}_n - \frac{q_{\alpha/2}}{2\sqrt{n}}, \bar{R}_n + \frac{q_{\alpha/2}}{2\sqrt{n}} \right]$$

- Indeed

$$\lim_{n \rightarrow \infty} \mathbb{P}(\mathcal{I}_{\text{conserv}} \ni p) \geq 1 - \alpha$$

## Solution 2: Solving the (quadratic) equation for $p$

- ▶ We have the system of two inequalities in  $p$ :

$$\bar{R}_n - \frac{q_{\alpha/2} \sqrt{p(1-p)}}{\sqrt{n}} \leq p \leq \bar{R}_n + \frac{q_{\alpha/2} \sqrt{p(1-p)}}{\sqrt{n}}$$

- ▶ Each is a quadratic inequality in  $p$  of the form

$$(p - \bar{R}_n)^2 \leq \frac{q_{\alpha/2}^2 p(1-p)}{n}$$

We need to find the roots  $p_1 < p_2$  of

$$\left(1 + \frac{q_{\alpha/2}^2}{n}\right)p^2 - \left(2\bar{R}_n + \frac{q_{\alpha/2}^2}{n}\right)p + \bar{R}_n^2 = 0$$

- ▶ This leads to a new confidence interval  $\mathcal{I}_{\text{solve}} = [p_1, p_2]$  such that:

$$\lim_{n \rightarrow \infty} \mathbb{P}(\mathcal{I}_{\text{solve}} \ni p) = 1 - \alpha$$

(it's complicated to write in generic way so let us wait to have values for  $n, \alpha$  and  $\bar{R}_n$  to plug-in)

## Solution 3: plug-in

- ▶ Recall that by the LLN  $\hat{p} = \bar{R}_n \xrightarrow[n \rightarrow \infty]{\mathbb{P}, \text{a.s.}} p$
- ▶ So by Slutsky, we also have

$$\sqrt{n} \frac{\bar{R}_n - p}{\sqrt{\hat{p}(1 - \hat{p})}} \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(0, 1)$$

- ▶ This leads to a new confidence interval:

$$\mathcal{I}_{\text{plug-in}} = \left[ \bar{R}_n - \frac{q_{\alpha/2} \sqrt{\hat{p}(1 - \hat{p})}}{\sqrt{n}}, \bar{R}_n + \frac{q_{\alpha/2} \sqrt{\hat{p}(1 - \hat{p})}}{\sqrt{n}} \right]$$

such that

$$\lim_{n \rightarrow \infty} \mathbb{P}(\mathcal{I}_{\text{plug-in}} \ni p) = 1 - \alpha$$



## 95% asymptotic CI for the kiss example

Recall that in the kiss example we had  $n = 124$  and  $\bar{R}_n = 0.645$ .

Assume  $\alpha = 5\%$ .

For  $\mathcal{I}_{\text{solve}}$ , we have to find the roots of:

$$1.03p^2 - 1.32p + 0.41 = 0 \quad p_1 = 0.53, p_2 = 0.75$$

We get the following confidence intervals of asymptotic level 95%:

- ▶  $\mathcal{I}_{\text{conserv}} = [0.56, 0.73]$
- ▶  $\mathcal{I}_{\text{solve}} = [0.53, 0.75]$
- ▶  $\mathcal{I}_{\text{plug-in}} = [0.56, 0.73]$

There are many<sup>4</sup> other possibilities in softwares even ones that use the exact distribution of  $n\bar{R}_n \sim \text{Bin}(n, p)$

$$\mathcal{I}_{\text{R default}} = [0.55, 0.73]$$

---

<sup>4</sup>See R. Newcombe (1998). *Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods*.

# Exercises

a) Let  $I$ ,  $J$  be some 95% and 98% asymptotic confidence intervals (respectively) for  $p$ . Which one of the following statements is correct?

1. We always have  $I \subset J$ .
2. We always have  $J \subset I$ .
3. None of the above.

b) Find a 98% asymptotic confidence interval for  $p$ .

$1-\alpha$  C.I.

$$\bar{R}_n \pm \frac{q_{\alpha/2}}{\sqrt{n}}$$

$$\left[ 0.645 \pm \frac{2.33}{\sqrt{124}} \right]$$

$$q_{1\%} = 2.33(?)$$

(Table check!)

# Exercises

c) Consider a new experiment in which there are 150 participants, 75 turned left and 75 turned right. Which of the following is the correct answer?

- 1.  $[0, 0.5]$  is a 50% asymptotic confidence intervals for  $p$
- 2.  $[0.5, 1]$  is a 50% asymptotic confidence intervals for  $p$
- 3.  $[0.466, 0.533]$  is a 50% asymptotic confidence intervals for  $p$
- 4.  $[0.48, 0.52]$  is a 50% asymptotic confidence intervals for  $p$
- 5. both (1) and (2)
- 6. (1), (2) and (3)
- 7. (1), (2), (3) and (4)

$[0, \bar{R}_n]$  also because  
$$P(\bar{R}_n - p \geq 0) \xrightarrow{n \rightarrow \infty} \frac{1}{2}$$

$$P\left[\frac{\bar{R}_n - p}{\sqrt{p(1-p)}} \leq 0\right] \xrightarrow{n \rightarrow \infty} \frac{1}{2} = 50\%$$

$$\bar{R}_n = 0.5$$

$$P[p \geq \bar{R}_n] \xrightarrow{n \rightarrow \infty} 50\%$$

$[\bar{R}_n, 1]$  is a 50% asymptotic CI.

## Exercises

d) If  $[0.34, 0.57]$  is a 95% confidence interval for an unknown proportion  $p$ , then the probability that  $p$  is in this interval is *at least asymptotically*

1. 0.025

2. 0.05

3. 0.95

☒ 4. None of the above

e) If  $[0.34, 0.57]$  is a 95% confidence interval for an unknown proportion  $p$ , is it also a 98% confidence interval?

1. Yes

☒ 2. No

f) If  $[0.34, 0.57]$  is a 95% confidence interval for an unknown proportion  $p$ , is it also a 90% confidence interval?

☒ 1. Yes

2. No

## Another example: The T



# Statistical problem

- ▶ You observe the times (in minutes) between arrivals of the T at Kendall:  $T_1, \dots, T_n$ .
- ▶ You **assume** that these times are:
  - ▶ Mutually independent ✓
  - ▶ Exponential random variables with common parameter ✓
- ▶ You want to *estimate* the value of  $\lambda$ , based on the observed arrival times.

# Discussion of the modeling assumptions

- ▶ Mutual independence of  $T_1, \dots, T_n$ : plausible but not completely justified (often the case with independence).
- ▶  $T_1, \dots, T_n$  are exponential r.v.: **lack of memory** of the exponential distribution:

$$\mathbb{P}[T_1 > t + s | T_1 > t] = \mathbb{P}[T_1 > s], \quad \forall s, t \geq 0.$$

Also,  $T_i > 0$  almost surely!

- ▶ The exponential distributions of  $T_1, \dots, T_n$  have the same parameter: in average all the same inter-arrival time. True only for limited period (rush hour  $\neq$  11pm).

# Estimator

- Density of  $T_1$ :

$$f(t) = \lambda e^{-\lambda t}, \quad \forall t \geq 0.$$

- $\mathbb{E}[T_1] = \frac{1}{\lambda}$ .

- Hence, a natural estimate of  $\frac{1}{\lambda}$  is

$$\bar{T}_n := \frac{1}{n} \sum_{i=1}^n T_i.$$

- A natural estimator of  $\lambda$  is

$$\hat{\lambda} := \frac{1}{\bar{T}_n} \xrightarrow[n \rightarrow \infty]{\text{a.s., } P} \lambda$$

$$\mathbb{E}\left[\frac{1}{\bar{T}_n}\right] > \frac{1}{\mathbb{E}[\bar{T}_n]} = \lambda$$

$$\lambda > 0$$



# First properties

- ▶ By the LLN's,

$$\bar{T}_n \xrightarrow[n \rightarrow \infty]{\text{a.s./}\mathbb{P}} \frac{1}{\lambda} \quad \checkmark$$

- ▶ Hence,

$$\hat{\lambda} \xrightarrow[n \rightarrow \infty]{\text{a.s./}\mathbb{P}} \lambda. \quad \checkmark$$

- ▶ By the CLT,

$$\sqrt{n} \left( \bar{T}_n - \frac{1}{\lambda} \right) \xrightarrow[n \rightarrow \infty]{} \mathcal{N}(0, \lambda^{-2}).$$

- ▶ How does the CLT transfer to  $\hat{\lambda}$  ? How to find an asymptotic confidence interval for  $\lambda$  ?

# The Delta method



← this is important

Let  $(Z_n)_{n \geq 1}$  sequence of r.v. that satisfies

$$\sqrt{n}(Z_n - \theta) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(0, \sigma^2)$$

for some  $\theta \in \mathbb{R}$  and  $\sigma^2 > 0$  (the sequence  $(Z_n)_{n \geq 1}$  is said to be *asymptotically normal around  $\theta$* ).

Let  $g : \mathbb{R} \rightarrow \mathbb{R}$  be continuously differentiable at the point  $\theta$ .

Then,

- ▶  $(g(Z_n))_{n \geq 1}$  is also asymptotically normal; around  $g(\theta)$
- ▶ More precisely,

$$\sqrt{n}(g(Z_n) - g(\theta)) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(0, (g'(\theta))^2 \sigma^2).$$

# Consequence of the Delta method

►  $\sqrt{n} \left( \hat{\lambda} - \lambda \right) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(0, \lambda^2).$

► Hence, for  $\alpha \in (0, 1)$  and when  $n$  is large enough,

$$|\hat{\lambda} - \lambda| \leq \lambda \cdot \frac{q_{\alpha/2}}{\sqrt{n}}$$

with probability approximately  $1 - \alpha$ .

► Can  $\left[ \hat{\lambda} - \frac{q_{\alpha/2}\lambda}{\sqrt{n}}, \hat{\lambda} + \frac{q_{\alpha/2}\lambda}{\sqrt{n}} \right]$  be used as an asymptotic confidence interval for  $\lambda$  ?

No : depends on  $\lambda$

# Three solutions

1. The conservative bound: we have no a priori way to bound  $\lambda$
2. We can solve for  $\lambda$ :

$$|\hat{\lambda} - \lambda| \leq \frac{q_{\alpha/2}\lambda}{\sqrt{n}} \iff \lambda \left(1 - \frac{q_{\alpha/2}}{\sqrt{n}}\right) \leq \hat{\lambda} \leq \lambda \left(1 + \frac{q_{\alpha/2}}{\sqrt{n}}\right)$$
$$\iff \frac{\hat{\lambda}}{1 + \frac{q_{\alpha/2}}{\sqrt{n}}} \leq \lambda \leq \frac{\hat{\lambda}}{1 - \frac{q_{\alpha/2}}{\sqrt{n}}}$$

It yields

$$\mathcal{I}_{\text{solve}} = \left[ \hat{\lambda} \left(1 + \frac{q_{\alpha/2}}{\sqrt{n}}\right)^{-1}, \hat{\lambda} \left(1 - \frac{q_{\alpha/2}}{\sqrt{n}}\right)^{-1} \right]$$

3. Plug-in yields

$$\mathcal{I}_{\text{plug-in}} = \left[ \hat{\lambda} \left(1 - \frac{q_{\alpha/2}}{\sqrt{n}}\right), \hat{\lambda} \left(1 + \frac{q_{\alpha/2}}{\sqrt{n}}\right) \right]$$

## 95% asymptotic CI for the T example

Assume that  $n = 64$  and  $\bar{T}_n = 6.23$  and  $\alpha = 5\%$ .

We get the following confidence intervals of asymptotic level 95%:

- ▶  $\mathcal{I}_{\text{solve}} = [0.13, 0.21]$
- ▶  $\mathcal{I}_{\text{plug-in}} = [0.12, 0.20]$

# Meaning of a confidence interval

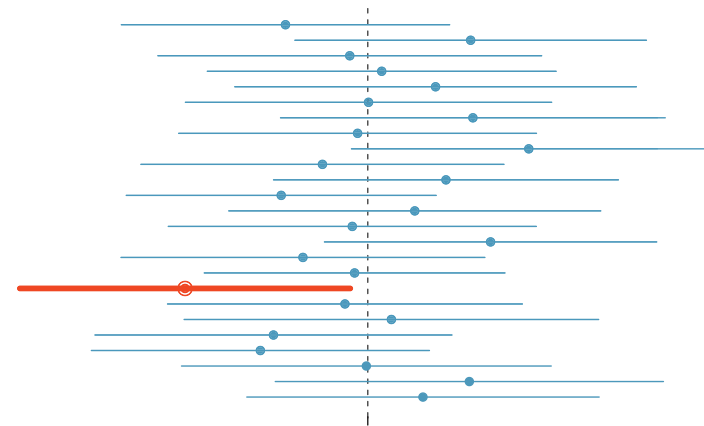
Take  $\mathcal{I}_{\text{plug-in}} = [0.12, 0.20]$  for example. What is the meaning of “ $\mathcal{I}_{\text{plug-in}}$  is a confidence intervals of asymptotic level 95%”.

Does it mean that

$$\lim_{n \rightarrow \infty} \mathbb{P}(\lambda \in [0.12, 0.20]) \geq .95?$$

No

There is a *frequentist* interpretation<sup>5</sup>:  
If we were to repeat this experiment (collect 64 observations) then  $\lambda$  would be in the resulting confidence interval about **95%** of the time (image credit: [openintro.org](http://openintro.org)).



---

<sup>5</sup>The frequentist approach is often contrasted with the Bayesian approach.

# Hypothesis testing

# How to board a plane?





# What is the fastest boarding method?

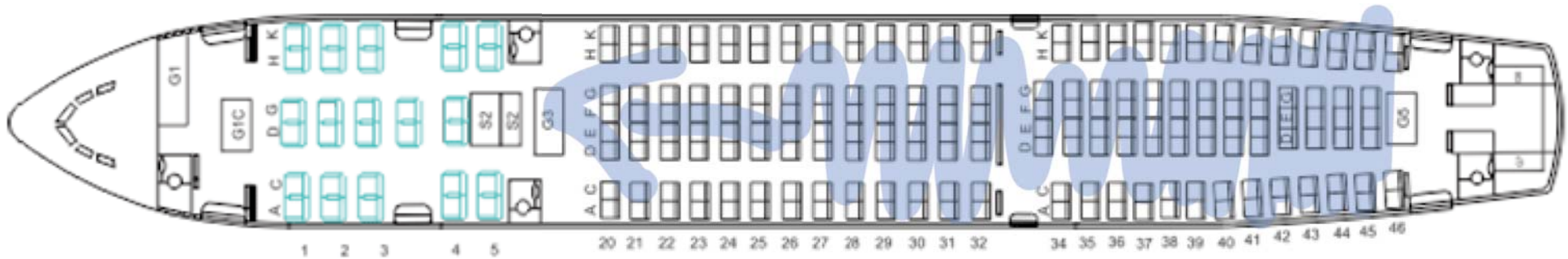
What is the fastest method to board a plane?

R2F

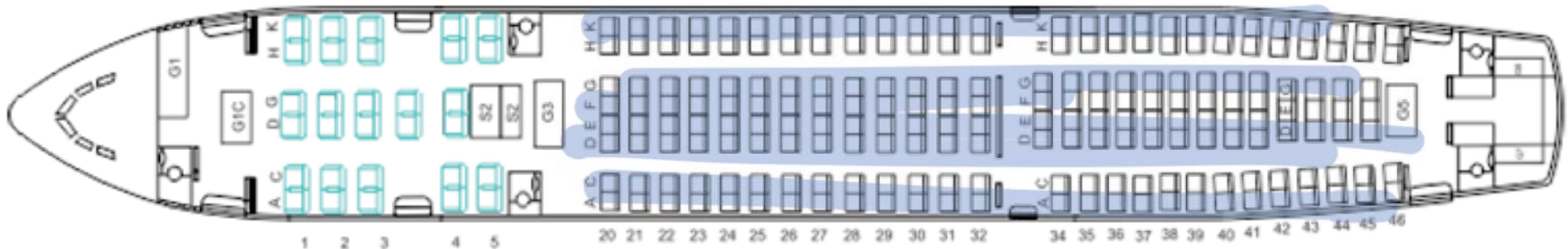
or

WilMA?

- ▶ R2F= Rear to Front



- ▶ WilMA=Window, Middle, Aisle. It is basically an OUTSIDE to INSIDE method.




# The data

We collected data from two different airlines: JetBlue (R2F) and United (WiMA).

We got the following results:

	R2F	WiMA
Average (mins)	24.2	15.9
Std. Dev (mins)	2.1	1.3
Sample size	72	56

# Model and Assumptions

- ▶ Let  $X$  (resp.  $Y$ ) denote the boarding time of a random JetBlue (resp. United) flight.
- ▶ We *assume* that  $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$  and  $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$
- ▶ Let  $n$  and  $m$  denote the JetBlue and United sample sizes respectively.
- ▶ We have  $X_1, \dots, X_n$   independent copies of  $X$  and  $Y_1, \dots, Y_m$  independent copies of  $Y$ .
- ▶ We further assume that the two samples are independent.

We want to answer the question:

Is  $\mu_1 = \mu_2$  or is  $\mu_1 > \mu_2$ ?

By making **modeling assumptions**, we have reduced the number of ways the hypothesis  $\mu_1 = \mu_2$  may be rejected. We do not allow that  $\mu_1 < \mu_2$ !

We have two samples: this is a **two-sample test**

# A first heuristic

Simple heuristic:

$$\text{"If } \bar{X}_n > \bar{Y}_m, \text{ then } \mu_1 > \mu_2\text{"}$$

This could go wrong if I randomly pick only full flights in my sample  $X_1, \dots, X_n$  and empty flights in my sample  $Y_1, \dots, Y_m$ .

Better heuristic:

$$\begin{aligned} &\text{"If} \\ &\bar{X}_n - \text{Buffer}_n > \bar{Y}_m + \text{Buffer}_m \\ &\text{then } \mu_1 > \mu_2\text{"} \end{aligned}$$

To make this intuition more precise, we need to take the size of the random fluctuations of  $\bar{X}_n$  and  $\bar{Y}_m$  into account!

# Waiting time in the ER

- ▶ The average waiting time in the Emergency Room (ER) in the US is 30 minutes according to the CDC
- ▶ Some patients claim that the new Princeton-Plainsboro hospital has a longer waiting time. Is it true?
- ▶ Here, we collect only one sample:  $X_1, \dots, X_n$  (waiting time in minutes for  $n$  random patients) with unknown expected value  $\mathbb{E}[X_1] = \mu$ .
- ▶ We want to know if  $\mu > 30$ .

This is a **one-sample test**



# Heuristic

Heuristic:

“If

$$\bar{X}_n + \text{Buffer}_n < 30$$

then conclude that

$$\mu \leq 30 \quad ”$$

# Example 1

According to a survey conducted in 2017 on 4,971 randomly sampled Americans, 32% report to get at least some of their news on Youtube. Can we conclude that at most a third of all Americans get at least some of their news on Youtube?

►  $n = 4,971$ ,  $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Ber}(p)$ ;

►  $\bar{X}_n = 0.32$

► If it was true that  $p = .33$ : By CLT,

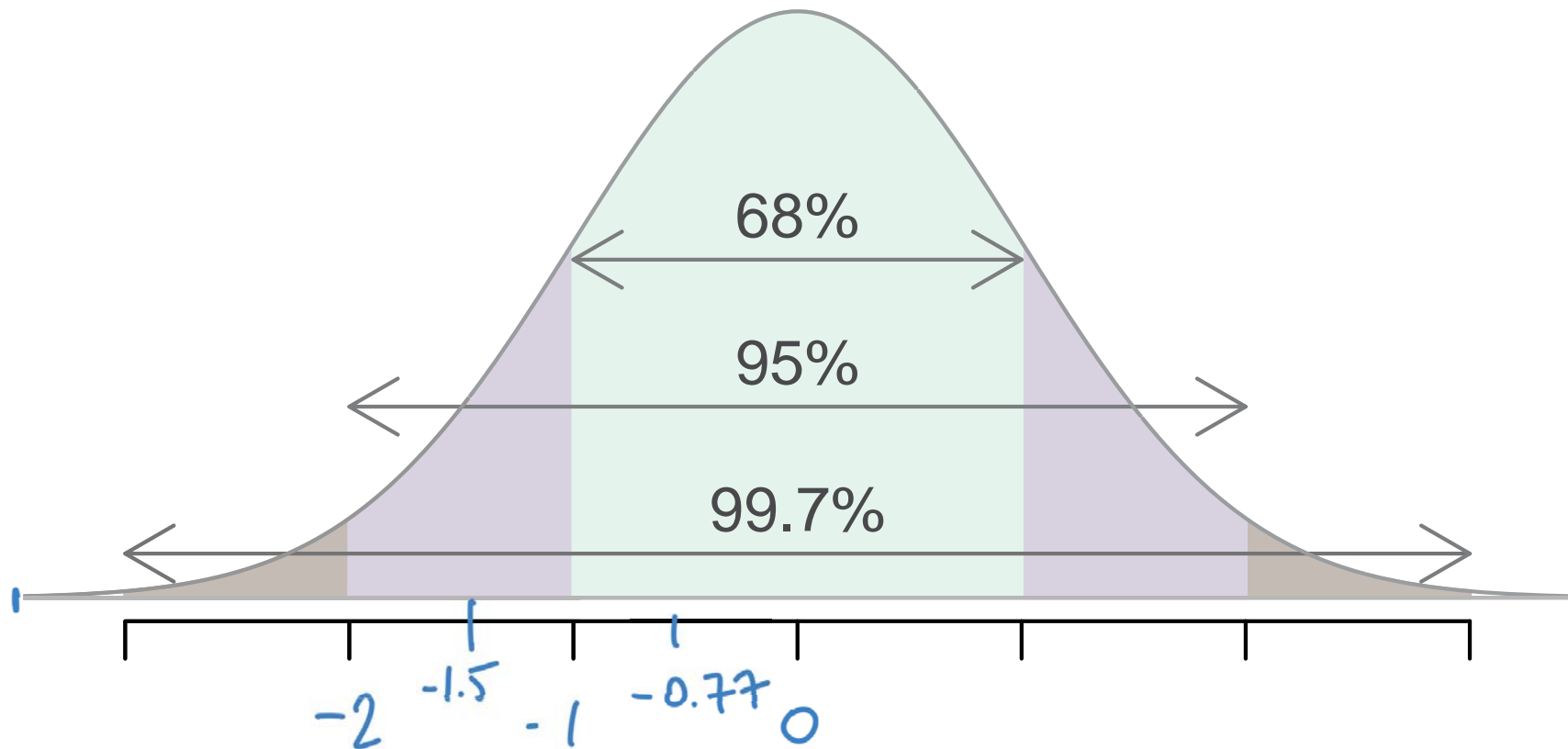
$$\begin{aligned} E[\bar{X}_n] &= .33 \\ \text{Var}[\bar{X}_n] &= \frac{.33(1-.33)}{4,971} \end{aligned}$$

$$\sqrt{n} \frac{\bar{X}_n - .33}{\sqrt{.33(1-.33)}} \approx \mathcal{N}(0, 1).$$

►  $\sqrt{n} \frac{\bar{X}_n - .33}{\sqrt{.33(1-.33)}} \approx -1.50$

► Conclusion:

# The Standard Gaussian distribution





## Example 2

**Example 2:** A coin is tossed 30 times, and Heads are obtained 13 times. Can we conclude that the coin is significantly unfair ?

►  $n = 30, X_1, \dots, X_n \stackrel{iid}{\sim} \text{Ber}(p);$

►  $\bar{X}_n = 13/30 \approx .43$

► If it was true that  $p = .5$ : By CLT,

$$\sqrt{n} \frac{\bar{X}_n - .5}{\sqrt{.5(1 - .5)}} \approx \mathcal{N}(0, 1).$$

► Our data gives  $\sqrt{n} \frac{\bar{X}_n - .5}{\sqrt{.5(1 - .5)}} \approx -0.77$

► The number  $-0.77$  is a plausible realization of a random variable  $Z \sim \mathcal{N}(0, 1)$ .

► Conclusion: ~~It is unlikely that the coin is unfair~~  
It is not unlikely that the coin is fair

At the end of the video we're presented the conclusion "It is unlikely that the coin is unfair." The professor contrast that with "The coin is likely to be fair." To me these mean exactly the same, and I think we can't conclude neither of these. I think the conclusion should actually have been "It is *not* unlikely that the coin is fair." Would the logicians among us agree?

There is no such thing as a fair coin, but a particular test may not detect that a coin is unfair. This is confusing at this point, but the later videos make clear what is meant. The hypotheses "the coin is fair" and "the coin is unfair" are not symmetric, the first is null hypothesis, the second is alternative hypothesis. The purpose of test is to reject the null hypothesis, and it is quite possible that the test can't do it; then we say "it is unlikely that the coin is unfair", using natural language. This statement makes clear sense in context of hypotheses testing, while the statements "it is likely that the coin is fair" or "it is not unlikely that the coin is fair" do not, IMO.

"It is not unlikely = It is likely (logic)"

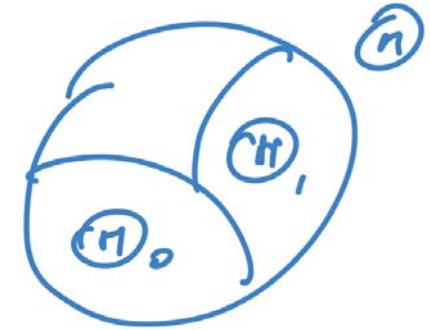
This would be true if unlikely = not likely. I argue somewhere else that that isn't the case. There is wise people, there is unwise people, but also a lot of people (the majority I would even argue) who are neither. (So 'not wise' = 'unwise' or 'neither wise nor unwise'.)

# Statistical formulation

- ▶ Consider a sample  $X_1, \dots, X_n$  of i.i.d. random variables and a statistical model  $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$ .

- ▶ Let  $\Theta_0$  and  $\Theta_1$  be disjoint subsets of  $\Theta$ .

- ▶ Consider the two hypotheses: 
$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_1 \end{cases}$$



- ▶  $H_0$  is the *null hypothesis*,  $H_1$  is the *alternative hypothesis*.
- ▶ If we believe that the true  $\theta$  is either in  $\Theta_0$  or in  $\Theta_1$ , we may want to *test  $H_0$  against  $H_1$* .
- ▶ We want to decide whether to *reject  $H_0$*  (look for evidence against  $H_0$  in the data).

# Asymmetry in the hypotheses

- ▶  $H_0$  and  $H_1$  do not play a symmetric role: the data is only used to try to disprove  $H_0$
- ▶ In particular lack of evidence, does not mean that  $H_0$  is true (“innocent until proven guilty”)

$$\psi(x) = \mathbb{I}(\psi(x) = 1)$$

- ▶ A *test* is a statistic  $\psi \in \{0, 1\}$  such that:
  - ▶ If  $\psi = 0$ ,  $H_0$  is not rejected;
  - ▶ If  $\psi = 1$ ,  $H_0$  is rejected.  $\Leftrightarrow H_1$

- ▶ Coin example:  $H_0: p = 1/2$  vs.  $H_1: p \neq 1/2$ .

- ▶  $\psi = \mathbb{I}\left\{\frac{\sqrt{n}|\bar{X}_n - \frac{1}{2}|}{\sqrt{0.5(1-0.5)}} > C\right\}$ , for some  $C > 0$ .

- ▶ How to choose the *threshold*  $C$  ?

# Errors

- ▶ Rejection region of a test  $\psi$ :

$$R_\psi = \{x \in E^n : \psi(x) = 1\}.$$

where  $(X_1, \dots, X_n)$  lives

$$\psi(x) = \mathbb{1}_{(x \in R_\psi)}$$

- ▶ Type 1 error of a test  $\psi$  (rejecting  $H_0$  when it is actually true):

$$\begin{aligned} \alpha_\psi &: \Theta_0 \rightarrow \mathbb{R} && \text{(or } [0, 1]) \\ \theta &\mapsto \mathbb{P}_\theta[\psi = 1]. \end{aligned}$$

- ▶ Type 2 error of a test  $\psi$  (not rejecting  $H_0$  although  $H_1$  is actually true):

$$\begin{aligned} \beta_\psi &: \Theta_1 \rightarrow \mathbb{R} \\ \theta &\mapsto \mathbb{P}_\theta[\psi = 0] \end{aligned}$$

- ▶ Power of a test  $\psi$ :

$$\pi_\psi = \inf_{\theta \in \Theta_1} (1 - \beta_\psi(\theta)).$$

下确界  
beta's domain

# Level, test statistic and rejection region

- ▶ A test  $\psi$  has *level*  $\alpha$  if (think  $\alpha = 5\%, 1\%, \dots$ )

$$\alpha_\psi(\theta) \leq \alpha, \quad \forall \theta \in \Theta_0.$$

- ▶ A test  $\psi$  has *asymptotic level*  $\alpha$  if

$$\lim_{n \rightarrow \infty} \alpha_\psi(\theta) \leq \alpha, \quad \forall \theta \in \Theta_0.$$

- ▶ In general, a test has the form

$$\psi = \mathbb{I}\{T_n > c\},$$

$$\psi = \mathbb{I}\{|T_n| > c\}$$
$$\psi = \mathbb{I}\{|T_n| \leq c\}$$

for some statistic  $T_n$  and threshold  $c \in \mathbb{R}$ .

- ▶  $T_n$  is called the *test statistic*. The rejection region is  $R_\psi = \{T_n > c\}$

# One-sided vs two-sided tests

We can refine the terminology when  $\theta \in \Theta \subset \mathbb{R}$  and  $H_0$  is of the form

$$H_0 : \theta = \theta_0 \quad \Leftrightarrow \quad \underbrace{\Theta_0 = \{\theta_0\}}$$

- ▶ If  $H_1 : \theta \neq \theta_0$ : **two-sided test** ✓
- ▶ If  $H_1 : \theta > \theta_0$  or  $H_1 : \theta < \theta_0$ : **one-sided test** ✓

Examples:

- ▶ Boarding method: *one sided*
- ▶ Waiting time in the ER: *one sided*
- ▶ The kiss example: *two sided*
- ▶ Fair coin: *two sided*

One or two sided tests will have different rejection regions.

# Bernoulli experiment

- ▶ Let  $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Ber}(p)$ , for some unknown  $p \in (0, 1)$ .
- ▶ We want to test:

$$H_0: p = 1/2 \text{ vs. } H_1: p \neq 1/2$$

with asymptotic level  $\alpha \in (0, 1)$ .

- ▶ Let  $T_n = \left| \sqrt{n} \frac{\hat{p}_n - 0.5}{\sqrt{.5(1 - .5)}} \right|$ , where  $\hat{p}_n$  is the MLE.
- ▶ If  $H_0$  is true, then by CLT,

$$\mathbb{P}[T_n > q_{\alpha/2}] \xrightarrow{n \rightarrow \infty} 0.05$$

- ▶ Let  $\psi_\alpha = \mathbb{I}\{T_n > q_{\alpha/2}\}$ .

# Examples

For  $\alpha = 5\%$ ,  $q_{\alpha/2} = 1.96$

## Fair coin

$H_0$  is \_\_\_\_\_ at the asymptotic level 5% by the test  $\psi_{5\%}$ .

## News on Youtube

$H_0 : p \geq 0.33$  vs.  $H_1 : p < 0.33$ . This is a \_\_\_\_\_-sided test.

We reject if:

$$\sqrt{n} \frac{\hat{p}_n - p}{\sqrt{p(1-p)}} > c$$

这个 $p_n$ 是对数据背后的真实分布的估计值 (一般是mean)  
而 $p$ 是一个你选择的参数, 用来进行假设检验  
比如0.5, 也就是和0.5来比, 原来的真实 $p_n$ 是大于0.5吗?

But what value for  $p \in \Theta_0 =$  \_\_\_\_\_ should we choose?

Type 1 error is the function  $p \mapsto \mathbb{P}_p[\psi = 1]$ . To control the level we need to find the  $p$  that **maximizes** it over  $\Theta_0$

→ no need for computations, it's clearly  $p =$  \_\_\_\_\_

$H_0$  is \_\_\_\_\_ at the asymptotic level 5% by the test  $\psi_{5\%}$ .



# p-value

## Definition

The (asymptotic) *p-value* of a test  $\psi_\alpha$  is the smallest (asymptotic) level  $\alpha$  at which  $\psi_\alpha$  rejects  $H_0$ . It is random, it depends on the sample.

## Golden rule

$\text{p-value} \leq \alpha \Leftrightarrow H_0$  is rejected by  $\psi_\alpha$ , at the (asymptotic) level  $\alpha$ .

**The smaller the p-value, the more confidently one can reject  $H_0$ .**

- ▶ Example 1:  $\text{p-value} = \mathbb{P}[|Z| > 3.21] \ll .01$ .
- ▶ Example 2:  $\text{p-value} = \mathbb{P}[|Z| > .77] \approx .44$ .

## Exercise: Cookies<sup>6</sup>

Students are asked to count the number of chocolate chips in 32 cookies for a class activity. They found that the cookies on average had 14.77 chocolate chips with a standard deviation of 4.37 chocolate chips. The packaging for these cookies claims that there are at least 20 chocolate chips per cookie. One student thinks this number is unreasonably high since the average they found is much lower. Another student claims the difference might be due to chance. What do you think (compute a p-value)?



---

<sup>6</sup>from the textbook OpenIntro Statistics

## Exercise: kiss

Recall that in the Kiss example we observed 80 out of 124 couples turning their head to the right. Formulate the statistical hypothesis problem, compute the p-value and conclude.

## Exercise : Machine learning predicts breast cancer

A vast problem in breast cancer are false positive, that is surgery performed on benign tumors. A new machine learning procedure claims to improve the state-of-the art (95% of false positive) significantly while preserving the same true positive rate (detecting malignant tumors as malignant). To verify this claim, we collected data on 297 benign tumors. The algorithm recommended to perform surgery on 206 of them.

Let  $p$  denote the proportion of benign tumors on which the algorithm prescribes surgery.

Formulate the statistical hypothesis problem, compute the p-value and conclude.

# Recap

- ▶ A statistical model is a pair of the form  $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$  where  $E$  is the sample space and  $(\mathbb{P}_\theta)_{\theta \in \Theta}$  is a family of candidate probability distributions.
- ▶ A model can be well specified and identifiable.
- ▶ The trinity of statistical inference: estimation, confidence intervals and testing
- ▶ Estimator: one value whose performance can be measured by consistency, asymptotic normality, bias, variance and quadratic risk
- ▶ Confidence intervals provide “error bars” around estimators. Their size depends on the confidence level
- ▶ Hypothesis testing: we want to ask a yes/no answer about an unknown parameter. They are characterized by hypotheses, level, power, test statistic and rejection region. Under the null hypothesis, the value of the unknown parameter becomes known (no need for plug-in).