Mathematical Statistics

Anna Janicka

Lecture X, 5.05.2022

HYPOTHESIS TESTING II:

COMPARING TESTS

Plan for Today

0. Definitions – reminder and supplement
1. Comparing tests
2. Uniformly Most Powerful Test
3. Likelihood ratio test: Neyman-Pearson Lemma
4. Examples of tests for simple hypotheses and generalizations

Definitions – reminder

We are testing H_0: $\theta \in \Theta_0$ against H_1: $\theta \in \Theta_1$

C – critical region of the test, the set of outcomes for which we reject H_0, $C = \{x \in \mathcal{X} : \delta(x) = 1\}$

The test has a **significance level** α, if for any $\theta \in \Theta_0$ we have $P_\theta(C) \leq \alpha$.

decision	In reality we have	
	H_0 true	H_0 false
reject H_0	**Type I error**	OK
do not reject H_0	OK	**Type II error**

Statistical test – example (is the coin symmetric?) $H_0: p = ½$ v $H_1: p \neq ½$ – reminder

Taking significance level $\alpha = 0.01$

We look for c such that (assuming $p = ½$)

$$P(|X - 200| > c) = 0.01$$

From the de Moivre-Laplace theorem for large n!

$$P(|X - 200| > c) \approx 2\,\Phi(-c/10), \text{ to get}$$

$$= 0.01 \text{ we need } c \approx 25.8$$

For a significance level approximately 0.01 we reject $H_0: p = ½$ when the number of tails is lower than 175 or higher than 225

$$C = \{0,1,\ldots,174\} \cup \{226, 227,\ldots, 400\}$$

Statistical test – example cont. (2). p-value

Slightly different question: what if the number of tails were 220 (T = 20)?

We have:

$$P_{½}(|X - 200| > 20) \approx 0.05$$

p-value: probability of type I error, if the value of the test statistic obtained was the critical value

So: p-value for T = 20 is approximately 0.05

p-value

p-value – probability of obtaining results *at least as extreme* as the ones obtained

(contradicting the null at least as much as those obtained)

decisions:

- p-value < α – reject the null hypothesis
- p-value ≥ α – no grounds to reject the null hypothesis

Statistical test – example cont. (3)
The choice of the alternative hypothesis

For a different alternative...

For example, we lose if tails appear *too often*.

- $H_0 : p = ½, \quad H_1 : p > ½$
- Which results would lead to rejecting H_0?
  - $X - 200 \leq c$ – do not reject H_0.
  - $X - 200 > c$ – reject H_0 in favor of H_1.

i.e. $T(x) = x - 200$

we could have
$H_0: p \leq ½$

Statistical test – example cont. (4)
The choice of the alternative hypothesis

Again, from the de Moivre – Laplace theorem:

$P_{½}(X – 200 > c) \approx 0.01$ for $c \approx 23.3$,

so for a significance level 0.01 we reject $H_0 : p = ½$ in favor of $H_1 : p > ½$ if the number of tails is at least 224

What if we got 220 tails?

p-value is equal to ≈ 0.025; do not reject H_0

Power of the test (for an alternative hypothesis)

$P_\theta(C)$ for $\theta \in \Theta_1$ – power of the test (for an alternative hypothesis)

Function of the power of a test:

$1\text{-}\beta : \Theta_1 \to [0,1]$ such that $1\text{-}\beta(\theta) = P_\theta(C)$

Usually: we look for tests with a given level of significance and the highest power possible.

Power of the test (for an alternative hypothesis)

Statistical test – example cont.
Power of the test

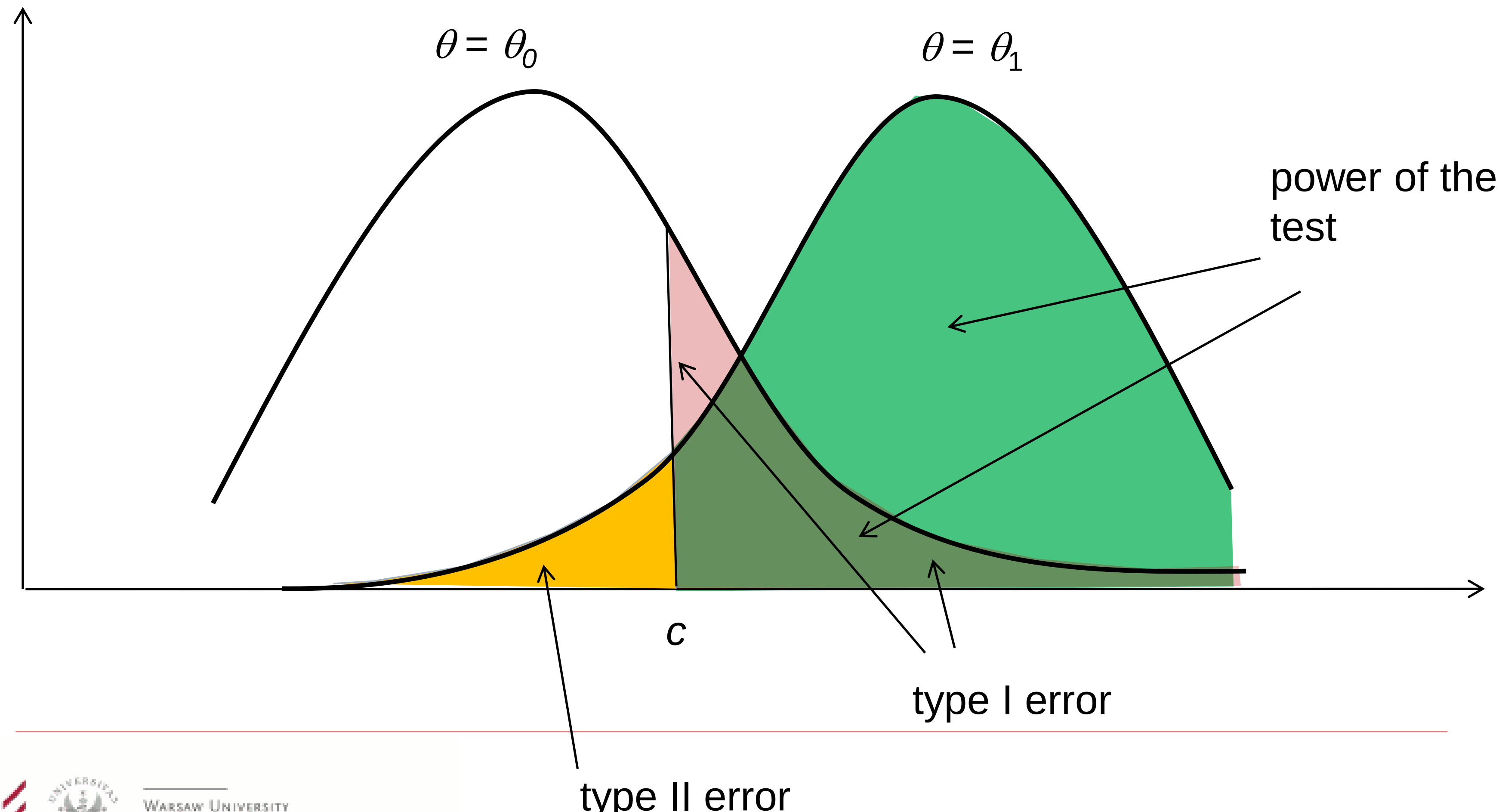
- We test $H_0 : p = ½$ against $H_1 : p = ¾$
  with: $T(x) = X – 200$, $C = \{T(x) > 23.3\}$
  (i.e. for a significance level $\alpha = 0.01$)
  Power of the test:
  $1\text{-}\beta(¾) = P(T(x) > 23.3 \mid p = ¾) = P_{¾}(X>223.3)$
  $\approx 1\text{-}\Phi((223.3\text{-}300)/5\sqrt{3}) \approx \Phi(8.85) \approx 1$
- But if $H_1 : p = 0.55$

$1\text{-}\beta(0.55) = P(T(x) > 23.3 \mid p = 0.55) \approx 1\text{-}\Phi(0.33) \approx 1\text{-}0.63 \approx 0.37$

- And if $H_1 : p = ¼$ for the same T we would get
  $1\text{-}\beta(¼) = P(T(x) > 23.3 \mid p = ¼) \approx 1\text{-}\Phi(14.23) \approx 0$

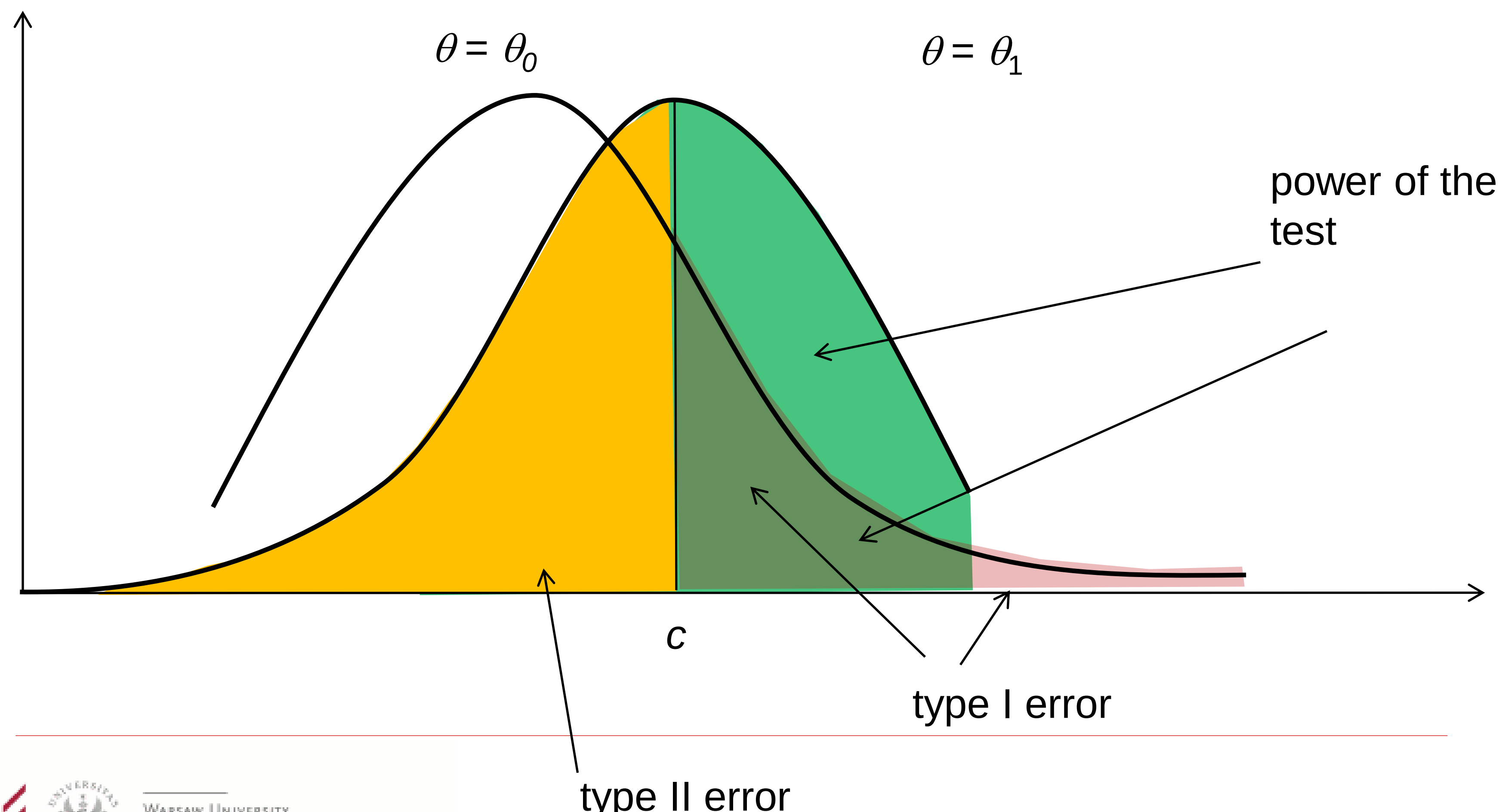
Power of the test: Graphical interpretation (1)

distributions of the test statistic T assuming that the null and alternative hypotheses are true

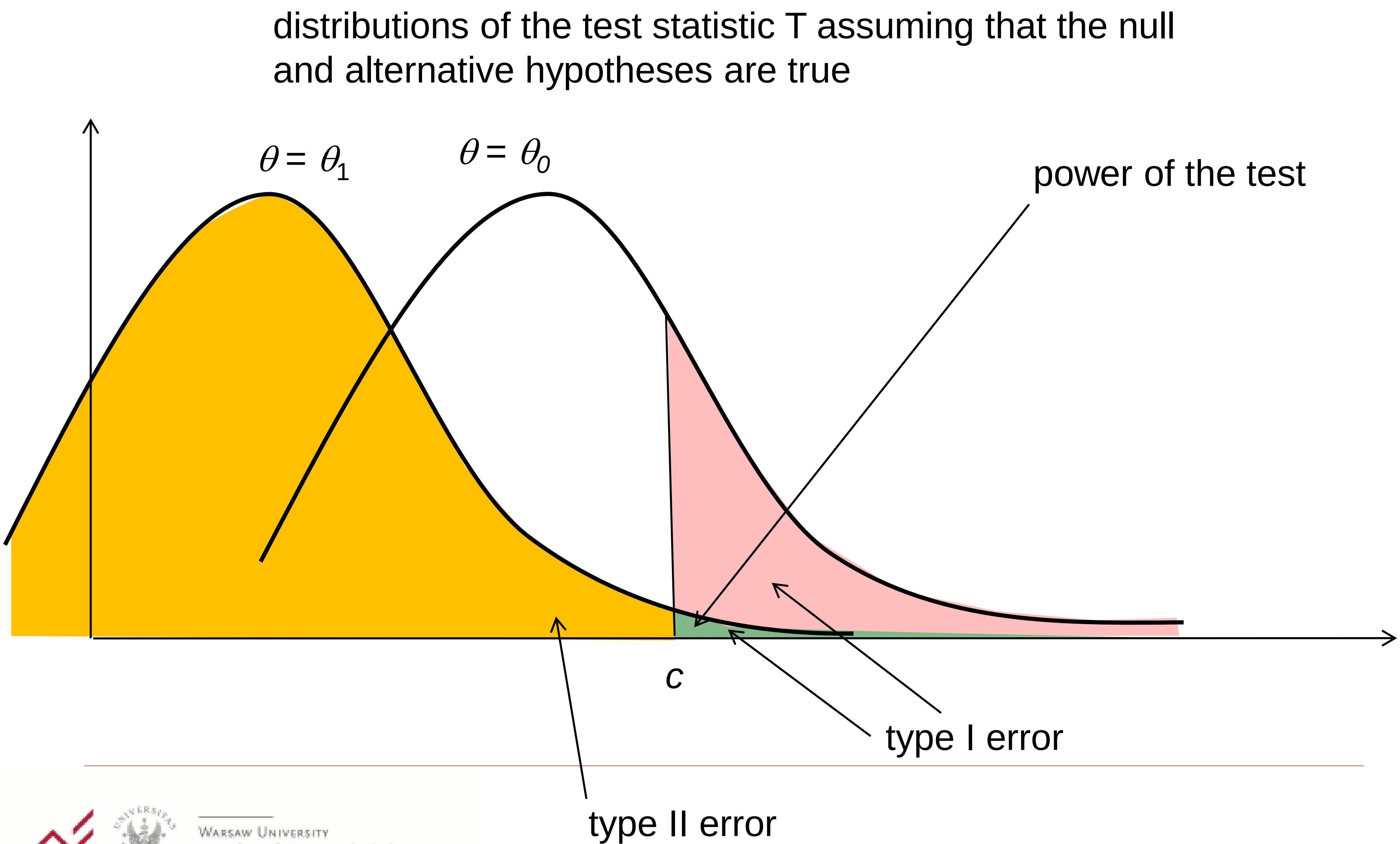


Power of the test: Graphical interpretation (2)

distributions of the test statistic T assuming that the null and alternative hypotheses are true



Power of the test: Graphical interpretation (3) – a very bad test



Sensitivity and specificity

Specificity – *true negative rate* (when in reality H_0 is not true)

Sensitivity – *true positive rate* (when in reality H_0 is true)

terms used commonly in diagnostic tests

(H_0 is having a medical condition)

Sensitivity and specificity – example

Performance of a SARS-COV-2 antigen test

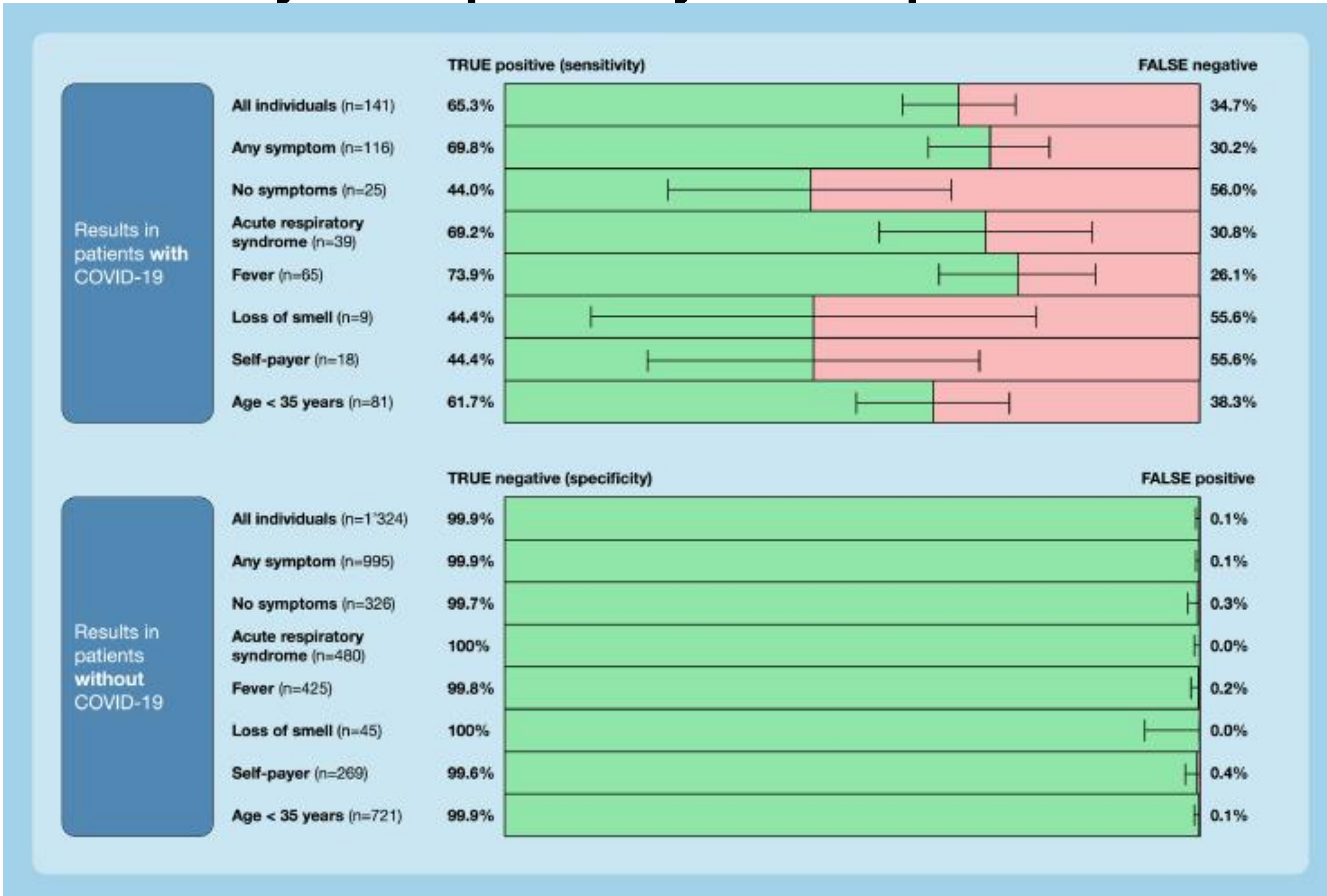
	Infected (null is true)	Not infected (null is false)	Overall nuber of cases
Positive test result	92	2 (Type II error, false positive)	94
Negative test result (reject null)	49 (Type I error, false negative)	1319	1368
Overall	141	1321	1432

Sensitivity: 92/141 = 65.3%

Specificity: 1319/1321 = 99.9%

Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8260496/

Sensitivity and specificity – example



Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8260496/

Size of a test

sometimes we also look at the **size** of a test:

$\sup_{\theta \in \Theta_0} P_\theta(C)$

then we have:

significance level = α if the size of the test does not exceed α.

Comparing tests

How do we chose the best test?

- for given null and alternative hypotheses
- for a given significance level

→ the test which is *more powerful* is better

Comparing the power of tests

$X \sim P_\theta$, $\{P_\theta : \theta \in \Theta\}$ – family of distributions

We test H_0: $\theta \in \Theta_0$ against H_1: $\theta \in \Theta_1$

such that $\Theta_0 \cap \Theta_1 = \varnothing$

with two tests with critical regions C_1 and C_2; both at significance level α.

The test with the critical region C_1 is **more powerful** than the test with critical region C_2, if

$$\forall \theta \in \Theta_1 : P_\theta(C_1) \geq P_\theta(C_2)$$

$$\text{and } \exists \theta_1 \in \Theta_1 : P_{\theta_1}(C_1) > P_{\theta_1}(C_2)$$

Uniformly most powerful test

For given H_0: $\theta \in \Theta_0$ and H_1: $\theta \in \Theta_1$:

δ^* is a **uniformly most powerful test** (UMPT) at significance level α, if

1) δ^* is a test at significance level α,
2) for any test δ at significance level α, we have, for any $\theta \in \Theta_1$:

$$P_\theta(\delta^*(X)=1) \geq P_\theta(\delta(X)=1)$$

i.e. the power of the test δ^* is not smaller than the power of any other test of the same hypotheses, for any $\theta \in \Theta_1$

if Θ_1 has one element, the word *uniformly* is redundant

Uniformly most powerful test – alternative form

For given H_0: $\theta \in \Theta_0$ and H_1: $\theta \in \Theta_1$:

A test with critical region C^* is a **uniformly most powerful test** (UMPT) at significance level α, if

1) The test with critical region C^* is a test at significance level α, i.e.

$$\text{for any } \theta \in \Theta_0: P_\theta(C^*) \leq \alpha,$$

2) for any test with critical region C at significance level α, we have for any $\theta \in \Theta_1$:

$$P_\theta(C^*) \geq P_\theta(C)$$

Testing simple hypotheses

We observe X. We want to test

H_0: $\theta = \theta_0$ against H_1: $\theta = \theta_1$.

(two simple hypotheses)

We can write it as:

H_0: $X \sim f_0$ against H_1: $X \sim f_1$,

where f_0 and f_1 are *densities* of distributions defined by θ_0 and θ_1 (i.e. P_0 and P_1)

Likelihood ratio test for simple hypotheses. Neyman-Pearson Lemma

Let $$C^* = \left\{ x \in \boldsymbol{X} : \frac{f_1(x)}{f_0(x)} > c \right\}$$

such that $P_0(C^*) = \alpha$ and $P_1(C^*) = 1 - \beta$

Then, for any $C \subseteq \boldsymbol{X}$:

if $P_0(C) \leq \alpha$, then $P_1(C) \leq 1 - \beta$.

(i.e.: the test with critical region C^* is the most powerful test for testing H_0 against H_1)

In many cases, it is easier to write the test as

$$C^* = \{x: \ln f_1(x) - \ln f_0(x) > c_1\}$$

Likelihood ratio test: we compare the likelihood ratio to a constant; if it is bad we reject H_0

Neyman-Pearson Lemma – Example 1

Normal model: X_1, X_2, ..., X_n are an IID sample from N(μ, σ^2), σ^2 is known

The most powerful test for

$$H_0\colon \mu = 0 \text{ against } H_1\colon \mu = 1.$$

$\mu_0 < \mu_1$

At significance level α :

$$C^* = \left\{(x_1, x_2, \ldots, x_n)\colon \bar{X} > {}^{u_{1-\alpha}\sigma}\!\big/_{\sqrt{n}}\right\}$$

For obs. 1.37; 0.21; 0.33; -0.45; 1.33; 0.85; 1.78; 1.21; 0.72 from N(μ, 1) we have, for $\alpha = 0.05$:

$$\bar{X} \approx 0.82 > {}^{1.645 \cdot 1}\!\big/_{\sqrt{9}} \approx 0.54$$

$\rightarrow$ we reject H_0

Neyman-Pearson Lemma – Example 1 cont.

Power of the test

$$P_1(C^*) = P\left(\bar{X} > {}^{1.645\sigma}\!/_{\sqrt{n}} \,|\, \mu = 1\right) = \ldots.$$

$$= 1 - \Phi\left(1.645 - {}^{\mu_1 \cdot \sqrt{n}}\!/_{\sigma}\right) \quad \approx 0.91$$

If we change α, μ_1, n – the power of the test....

Neyman-Pearson Lemma: Generalization of example 1

The same test is UMP for H_1: $\mu > 0$ and for

$$H_0: \mu \leq 0 \text{ against } H_1: \mu > 0$$

more generally: under additional assumptions about the family of distributions, the same test is UMP for testing

$$H_0: \mu \leq \mu_0 \text{ against } H_1: \mu > \mu_0$$

Note the change of direction in the inequality when testing

$$H_0: \mu \geq \mu_0 \text{ against } H_1: \mu < \mu_0$$

Neyman-Pearson Lemma – Example 2

Exponential model: $X_1, X_2, \ldots, X_n$ are an IID sample from an exp(λ) distribution, $n = 10$.

MP test for

$$H_0: \lambda = \tfrac{1}{2} \text{ against } H_1: \lambda = \tfrac{1}{4}.$$

At significance level $\alpha = 0.05$:

$$C^* = \left\{(x_1, x_2, \ldots, x_{10}) : \sum x_i > 31.41\right\}$$

E.g. for a sample: 2; 0.9; 1.7; 3.5; 1.9; 2.1; 3.7; 2.5; 3.4; 2.8:

$\Sigma = 24.5 \rightarrow$ no grounds for rejecting H_0.

$$\exp(\lambda) = \Gamma(1, \lambda) \qquad \Gamma(a, \lambda) + \Gamma(b, \lambda) = \Gamma(a+b, \lambda) \qquad \Gamma(n/2, 1/2) = \chi^2(n)$$

Neyman-Pearson Lemma – Example 2'

Exponential model: $X_1, X_2, \ldots, X_n$ are an IID sample from an exp(λ) distribution, $n = 10$.

MP test for

$$H_0: \lambda = \tfrac{1}{2} \text{ against } H_1: \lambda = \tfrac{3}{4}.$$

At significance level $\alpha = 0.05$:

$$C^* = \left\{ (x_1, x_2, \ldots, x_{10}) : \sum x_i < 10.85 \right\}$$

E.g. for a sample: 2; 0.9; 1.7; 3.5; 1.9; 2.1; 3.7; 2.5; 3.4; 2.8:

$\Sigma = 24.5 \rightarrow$ no grounds for rejecting H_0.

$$\exp(\lambda) = \Gamma(1, \lambda) \qquad \Gamma(a, \lambda) + \Gamma(b, \lambda) = \Gamma(a + b, \lambda) \qquad \Gamma(n/2, 1/2) = \chi^2(n)$$

Example 2 cont.

The test $$C^* = \left\{(x_1, x_2, \ldots, x_{10}) : \sum x_i > 31.41\right\}$$

is UMP for H_0: $\lambda \geq ½$ against H_1: $\lambda < ½$

The test $$C^* = \left\{(x_1, x_2, \ldots, x_{10}) : \sum x_i < 10.85\right\}$$

is UMP for H_0: $\lambda \leq ½$ against H_1: $\lambda > ½$