

I will post HW1 later this week.

I will post solutions
one week from now

Discrete Probability Distributions

Random Variables

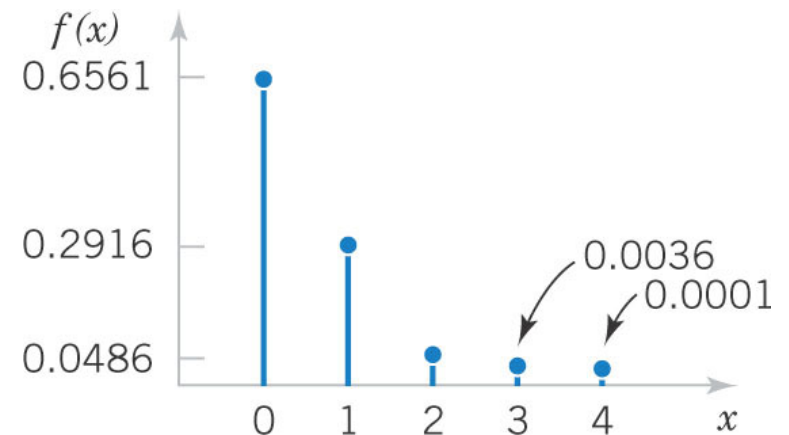
- A variable that associates a number with the outcome of a **random experiment** is called a **random variable**.
- Notation: **random variable** is denoted by an uppercase letter, such as ***X***. After the experiment is conducted, the **measured value** is denoted by a **lowercase letter**, such a ***x***. Both *X* and *x* are shown in italics, e.g., ***P(X=x)***.

Continuous & Discrete Random Variables

- A **discrete random variable** is usually an integer number
 - N - the number of p53 proteins in a cell
 - D - the number of nucleotides different between two gene sequences of length L
- A **continuous random variable** is a real number
 - $C=N/V$ – the concentration of p53 protein in a cell of volume V
 - Percentage $(D/L)*100\%$ of different nucleotides in gene sequences of different lengths L
(depending on the set of L 's it may be discrete but dense)

Probability Mass Function (PMF)

- I want to **compare all 4-mers** in a pair of human genomes
- **X – random variable:** the number of nucleotide differences in a given 4-mer
- **Probability Mass Function:** $f(x)$ or $P(X=x)$ – the probability that the # of SNPs is **exactly equal to x**



Probability Mass Function for the # of mismatches in 4-mers

$P(X=0) =$	0.6561
$P(X=1) =$	0.2916
$P(X=2) =$	0.0486
$P(X=3) =$	0.0036
$P(X=4) =$	0.0001
$\sum_x P(X=x) =$	1.0000

Cumulative Distribution Function (CDF)

	x	$P(X=x)$	CDF: $P(X \leq x)$	CCDF: $P(X > x)$
	-1	0.0000	0.0000	1.0000
	0	0.6561	0.6561	0.3439
	1	0.2916	0.9477	0.0523
	2	0.0486	0.9963	0.0037
	3	0.0036	0.9999	0.0001
	4	0.0001	1.0000	0.0000

Cumulative Distribution Function CDF: $F(x) = P(X \leq x)$

Example:

$$F(3) = P(X \leq 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3) = 0.9999$$

Complementary Cumulative Distribution Function
(tail distribution) or CCDF: $F_{>}(x) = P(X > x)$

$$\text{Example: } F_{>}(0) = P(X > 0) = 1 - P(X \leq 0) = 1 - 0.6561 = 0.3439$$

Mean or Expected Value of X

The **mean** or **expected value** of the discrete random variable X, denoted as μ or $E(X)$, is

$$\mu = E(X) = \sum_x x \cdot P(X = x) = \sum_x x \cdot f(x)$$

- **The mean** = the weighted average of all possible values of X. It represents its “center of mass”
- The **mean** may, or may not, be an **allowed value of X**
- It is also called the **arithmetic mean** (to distinguish from e.g. the **geometric mean** discussed later)
- **Mean** may be infinite if X any integer and tail $P(X=x) > c/x^2$

Outcomes of 6 random experiments

0, 1, 0, 0, 2, 1

$$\text{Mean} = \frac{0 + 1 + 0 + 0 + 2 + 1}{6} =$$

$$= \frac{3 \times 0 + 2 \times 1 + 1 \times 2}{6} =$$

$$= 0 \times \frac{3}{6} + 1 \times \frac{2}{6} + 2 \times \frac{1}{6} = \sum_{x=0}^2 x P(X=x)$$

- $E[X] = \sum_x x \cdot P(X=x)$

- $E[X^2] = \sum_x x^2 \cdot P(X=x)$

- $E[a \cdot X + b \cdot X^2] = \sum (ax + bx^2) \cdot$
 $\cdot P(X=x) = a \cdot \sum x P(X=x) +$
 $+ b \sum x^2 P(X=x)$

- $E[e^X] = \sum e^x P(X=x)$

Variance $V(X)$: Square
of a typical deviation from
the mean $\mu = E(X)$

$V(X) = \sigma^2$, where σ is called
standard deviation

$$\begin{aligned}\sigma^2 &= V(X) = E((X - \mu)^2) = \\&= E(X^2 - 2\mu X + \mu^2) = E(X^2) - \\&- 2\mu E(X) + \mu^2 = E(X^2) - 2\mu^2 + \mu^2 = \\&= E(X^2) - \mu^2 = E(X^2) - (E(X))^2\end{aligned}$$

Variance of a Random Variable

If X is a discrete random variable with probability mass function $f(x)$,

$$E[h(X)] = \sum_x h(x) \cdot P(X = x) = \sum_x h(x) f(x) \quad (3-4)$$

If $h(x) = (X - \mu)^2$, then its expectation, $V(x)$, is the **variance of X** .

$\sigma = \sqrt{V(x)}$, is called **standard deviation of X**

$\sigma^2 = V(X) = \sum_x (x - \mu)^2 f(x)$ is the **definitional** formula

$$= \sum_x (x^2 - 2\mu x + \mu^2) f(x)$$

$$= \sum_x x^2 f(x) - 2\mu \sum_x x f(x) + \mu^2 \sum_x f(x)$$

$$= \sum_x x^2 f(x) - 2\mu^2 + \mu^2$$

$$= \sum_x x^2 f(x) - \mu^2 \text{ is the } \mathbf{computational} \text{ formula}$$

Variance can be infinite
if X can be any integer
and tail of $P(X=x) \geq c/x^3$

Skewness of a random variable

- Want to quantify **how asymmetric** is the **distribution around the mean**?
- Need any **odd moment**: $E[(X-\mu)^{2n+1}]$
- **Cannot** do it with the **first moment**: $E[X-\mu]=0$
- Normalized 3-rd moment is **skewness**:
 $\gamma_1 = E[(X-\mu)^3/\sigma^3]$
- Skewness **can be infinite** if X takes unbounded positive integer values and the tail $P(X=x) \geq c/x^4$ for large x

Geometric mean of a random variable

- Useful for **very broad distributions** (many orders of magnitude)?
- Mean may be dominated by **very unlikely** but **very large events**. Think of a **lottery**
- **Exponent of the mean of $\log X$:**
Geometric mean = $\exp(E[\log X])$
- Geometric mean usually **is not infinite**

Summary: Parameters of a Probability Distribution

- **Probability Mass Function (PMF):** $f(x) = \text{Prob}(X=x)$
- **Cumulative Distribution Function (CDF):** $F(x) = \text{Prob}(X \leq x)$
- **Complementary Cumulative Distribution Function (CCDF):** $F_{>}(x) = \text{Prob}(X > x)$
- The **mean**, $\mu = E[X]$, is a measure of the **center of mass of a random variable**
- The **variance**, $V(X) = E[(X - \mu)^2]$, is a measure of the **dispersion** of a random variable **around its mean**
- The **standard deviation**, $\sigma = [V(X)]^{1/2}$, is another measure of the **dispersion** around mean. Has the same units as X
- The **skewness**, $\gamma_1 = E[(X - \mu)^3 / \sigma^3]$, a measure of asymmetry around mean
- The **geometric mean**, $\exp(E[\log X])$ is useful for very broad distributions

A gallery of useful discrete probability distributions

Discrete Uniform Distribution

- Simplest discrete distribution.
- The random variable X assumes only a finite number of values, each with equal probability.
- A random variable X has a discrete uniform distribution if each of the n values in its range, say $x_1, x_2, \dots, x_n$, has equal probability.

$$f(x_i) = 1/n$$

Uniform Distribution of Consecutive Integers

- Let X be a discrete uniform random variable all integers from a to b (inclusive). There are $b - a + 1$ integers. Therefore each one gets:

$$f(x) = 1/(b-a+1)$$

- Its measures are:

$$\mu = E(x) = (b+a)/2$$

$$\sigma^2 = V(x) = [(b-a+1)^2-1]/12$$

Note that the mean is the midpoint of a & b .

A random variable X has the same probability for integer numbers

$$x = 1:10$$

What is the behavior of its **Probability Mass Function (PMF): $P(X=x)$** ?

- A. does not change with $x=1:10$
- B. linearly increases with $x=1:10$
- C. linearly decreases with $x=1:10$
- D. is a quadratic function of $x=1:10$

Get your i-clickers

A random variable X has the same probability for integer numbers

$$x = 1:10$$

What is the behavior of its Cumulative Distribution Function (CDF): $P(X \leq x)$?

A. does not change with $x=1:10$

B. linearly increases with $x=1:10$

C. linearly decreases with $x=1:10$

D. is a quadratic function of $x=1:10$

Get your i-clickers

A random variable X has the same probability for integer numbers

$$x = 1:10$$

What is its mean value?

A. 0.5

B. 5.5

C. 5

D. 0.1

Get your i-clickers

A random variable X has the same probability for integer numbers

$$x = 1:10$$

What is its **skewness**?

A. 0.5

B. 1

C. 0

D. 0.1

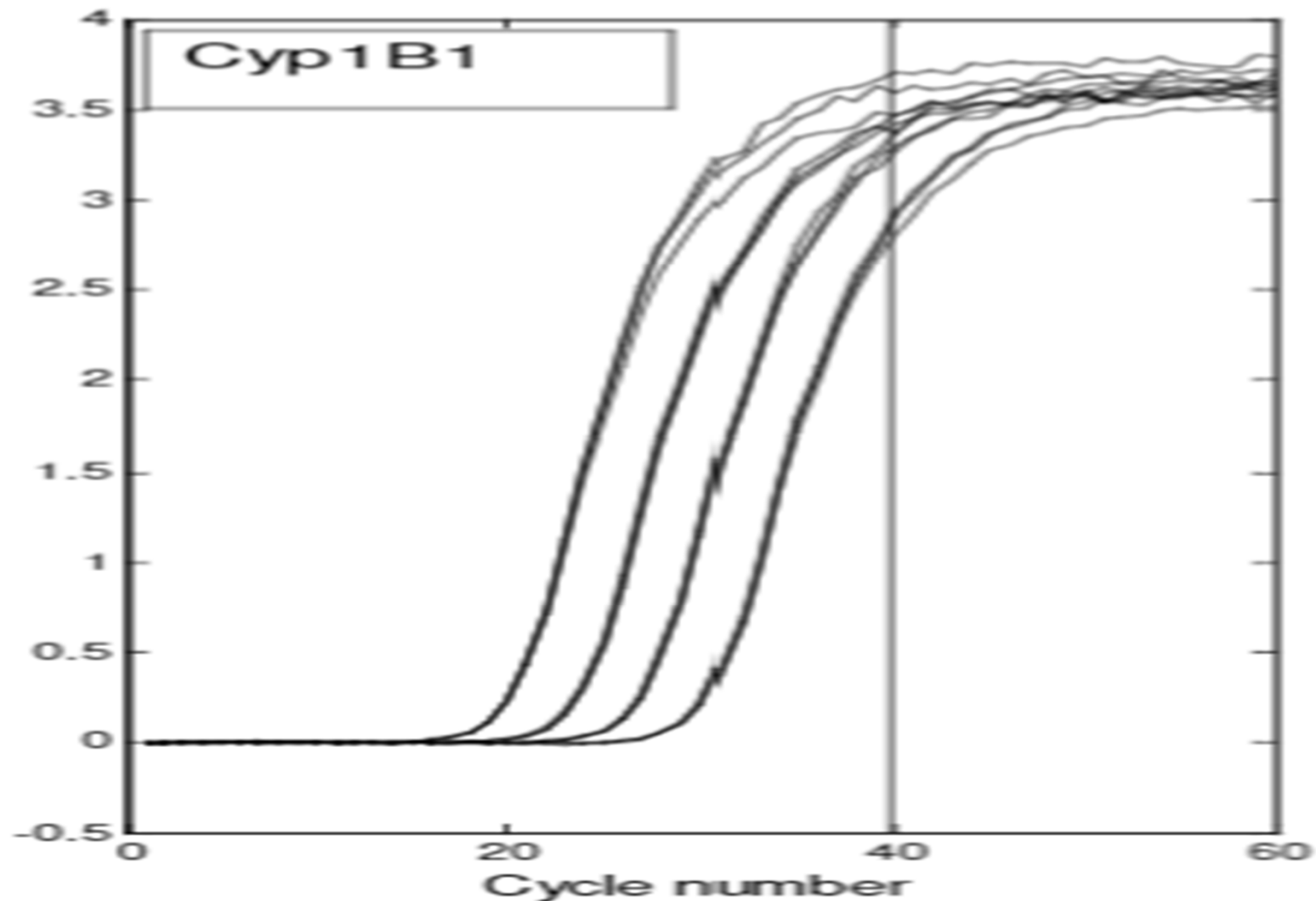
Get your i-clickers

An example of the uniform
distribution

Cycle threshold (Ct) value in
COVID-19 infection

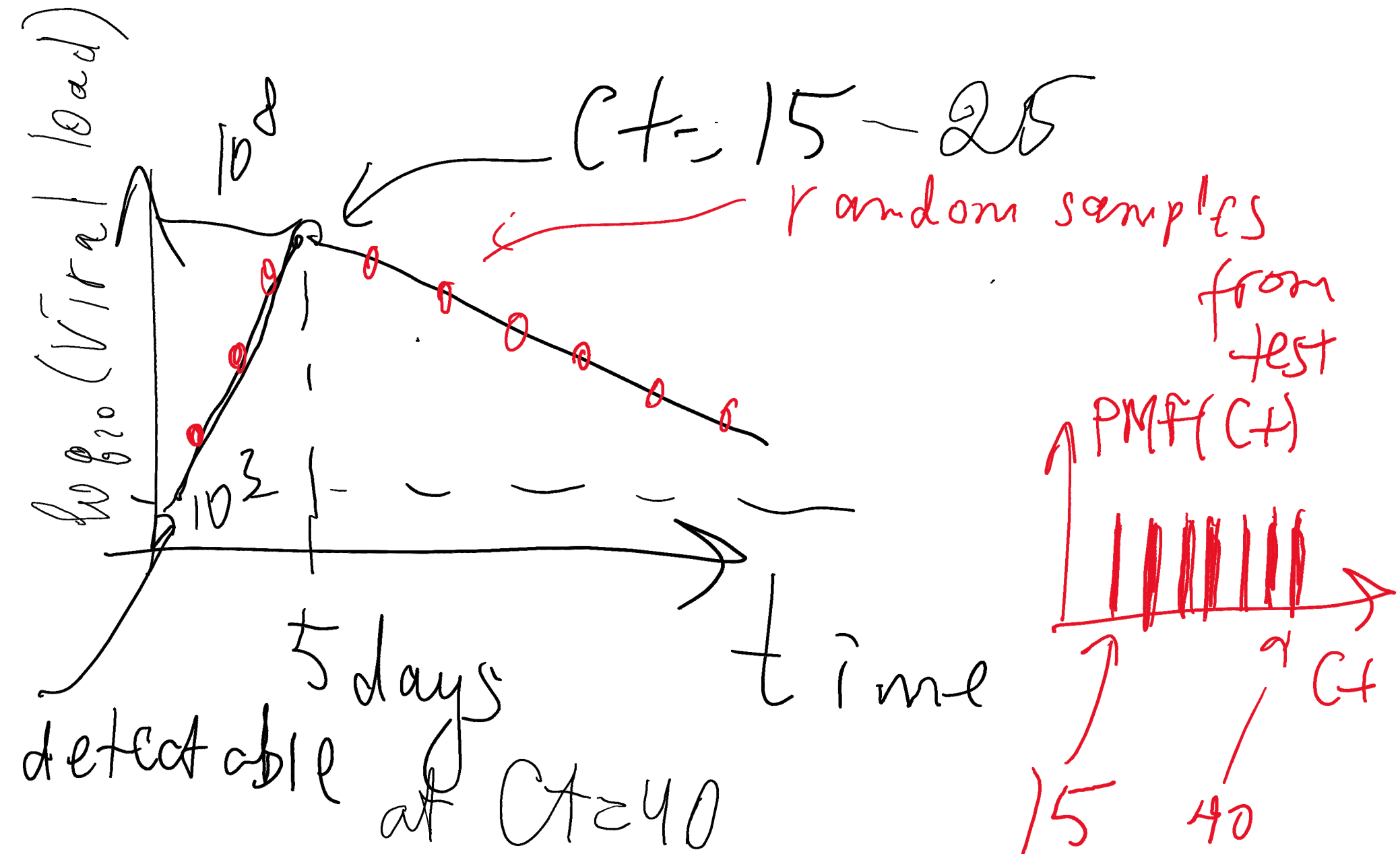
What is the Ct value of a PCR test?

$Ct = \text{const} - \log_2(\text{viral DNA concentration})$



Why Ct distribution should be uniform?

Why Ct distribution should be uniform?



Examples of uniform distribution: Ct value of a PCR test for a virus

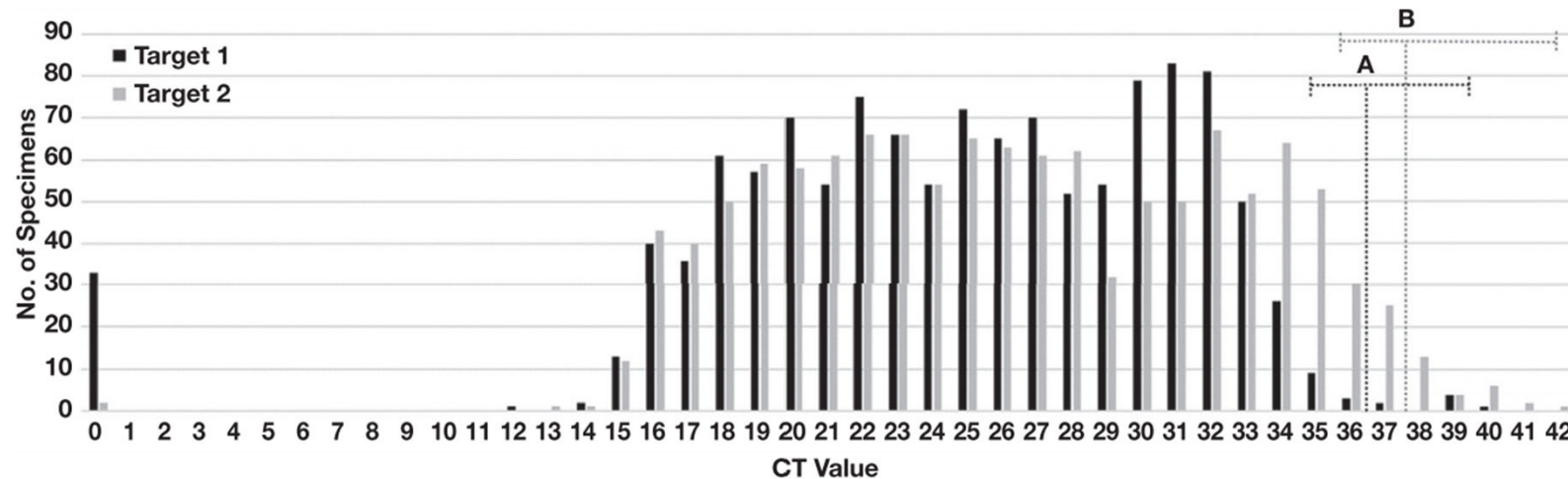


Figure 3 Distribution of cycle threshold (CT) values. The total number of specimens with indicated CT values for Target 1 and 2 are plotted. The estimated limit of detection for (A) Target 1 and (B) Target 2 are indicated by vertical dotted lines. Horizontal dotted lines encompass specimens with CT values less than 3× the LoD for which sensitivity of detection may be less than 100%. This included 19/1,180 (1.6%) reported CT values for Target 1 and 81/1,211 (6.7%) reported CT values for Target 2. Specimens with Target 1 or 2 reported as “not detected” are denoted as a CT value of “0.”

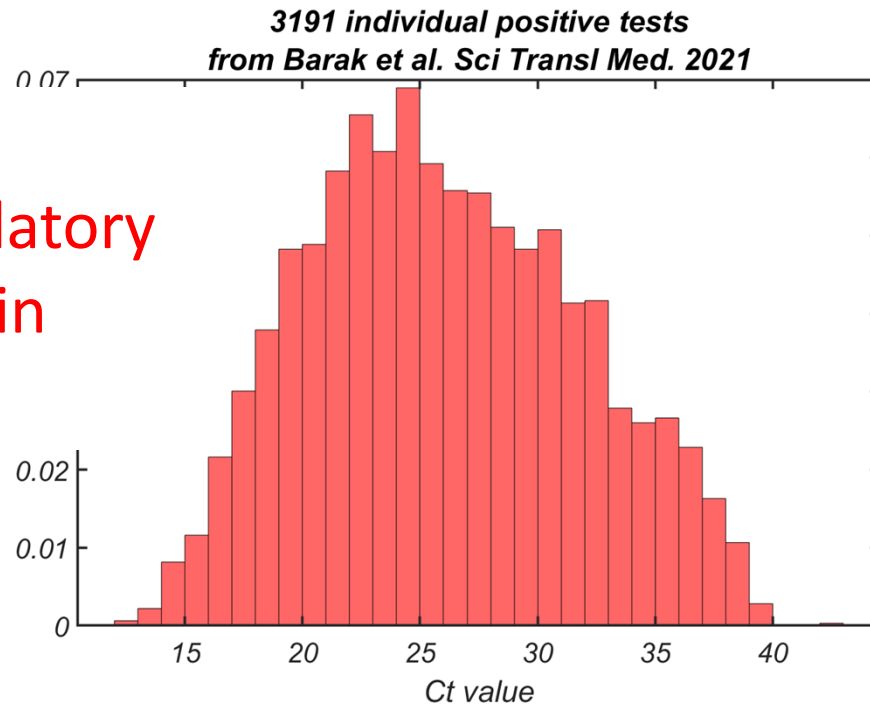
Distribution of SARS-CoV-2 PCR Cycle Threshold Values Provide Practical Insight Into Overall and Target-Specific Sensitivity Among Symptomatic Patients

Blake W Buchan, PhD, Jessica S Hoff, PhD, Cameron G Gmehlin, Adriana Perez, Matthew L Faron, PhD, L Silvia Munoz-Price, MD, PhD, Nathan A Ledebor, PhD *American Journal of Clinical Pathology*, Volume 154, Issue 4, 1 October 2020,

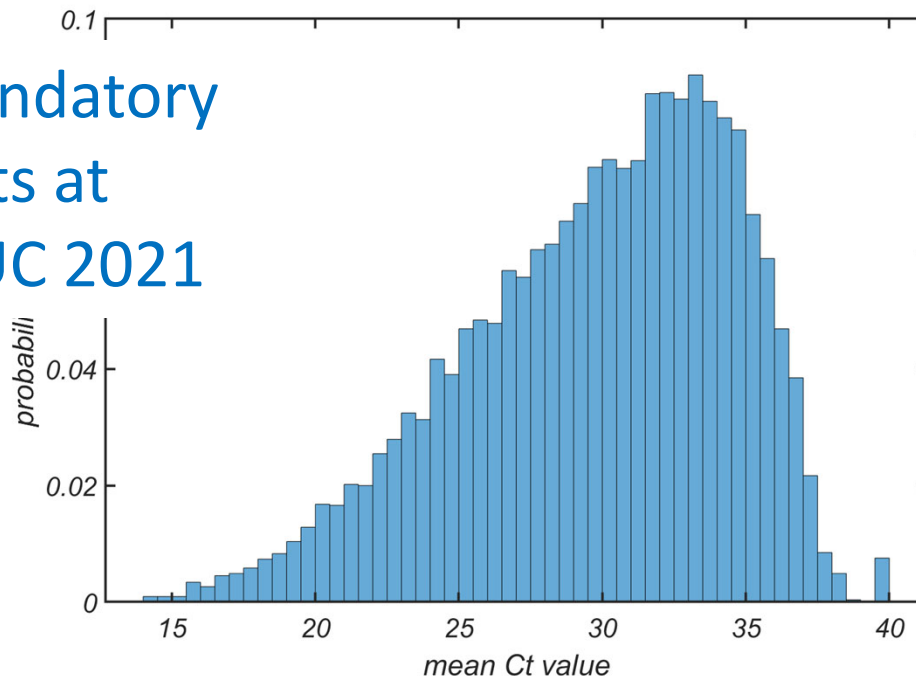
<https://academic.oup.com/ajcp/article/154/4/479/5873820>

Why should we care?

Non-mandatory tests in Israel



Mandatory tests at UIUC 2021



- High Ct value means we identified the infected individual early, hopefully before transmission to others
- When testing is mandatory, and people are tested frequently – the mean Ct value is shifted towards higher values