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 $x_1, x_2, \ldots, x_n$, 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

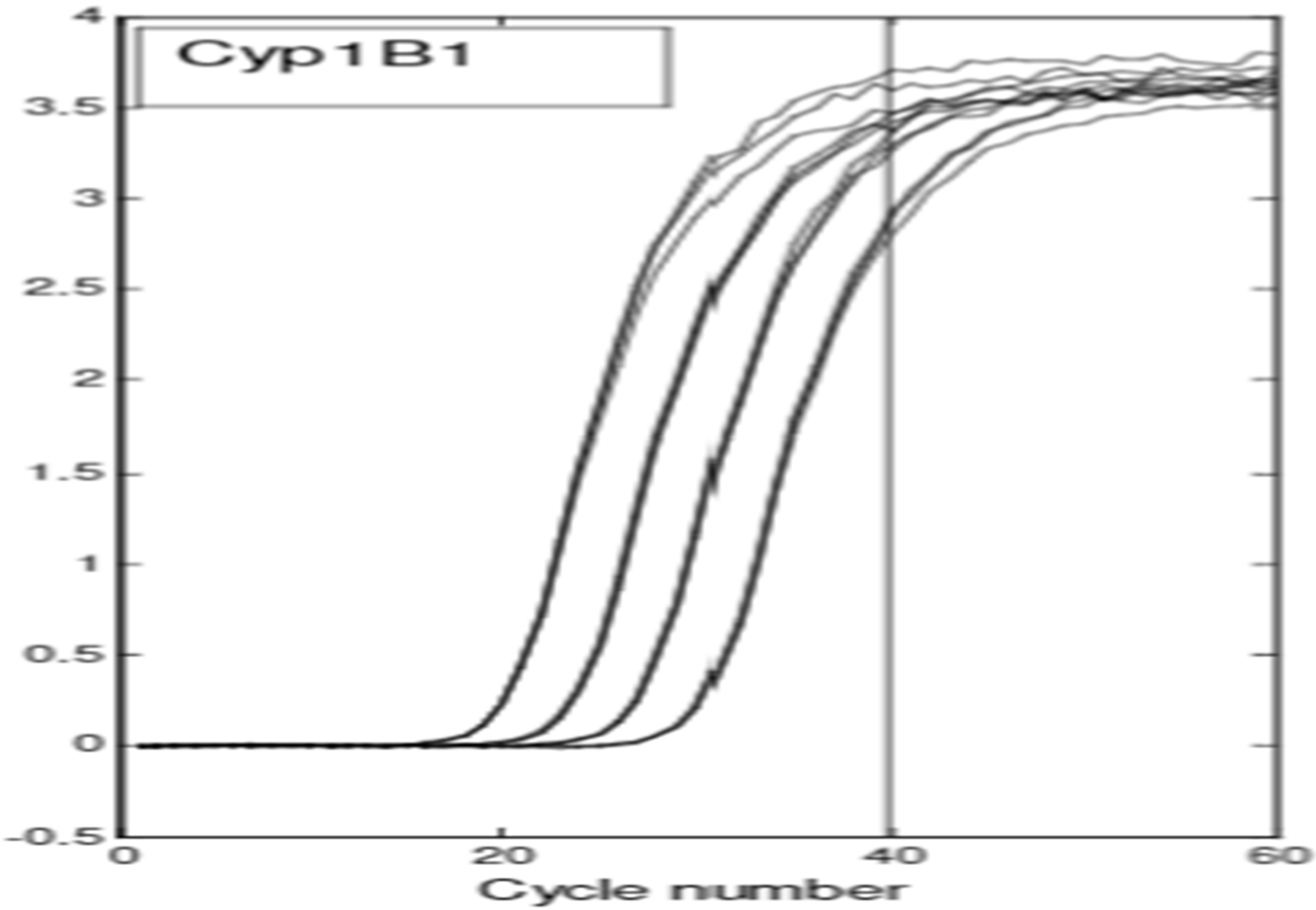
Get your i-clickers

An example of the uniform distribution

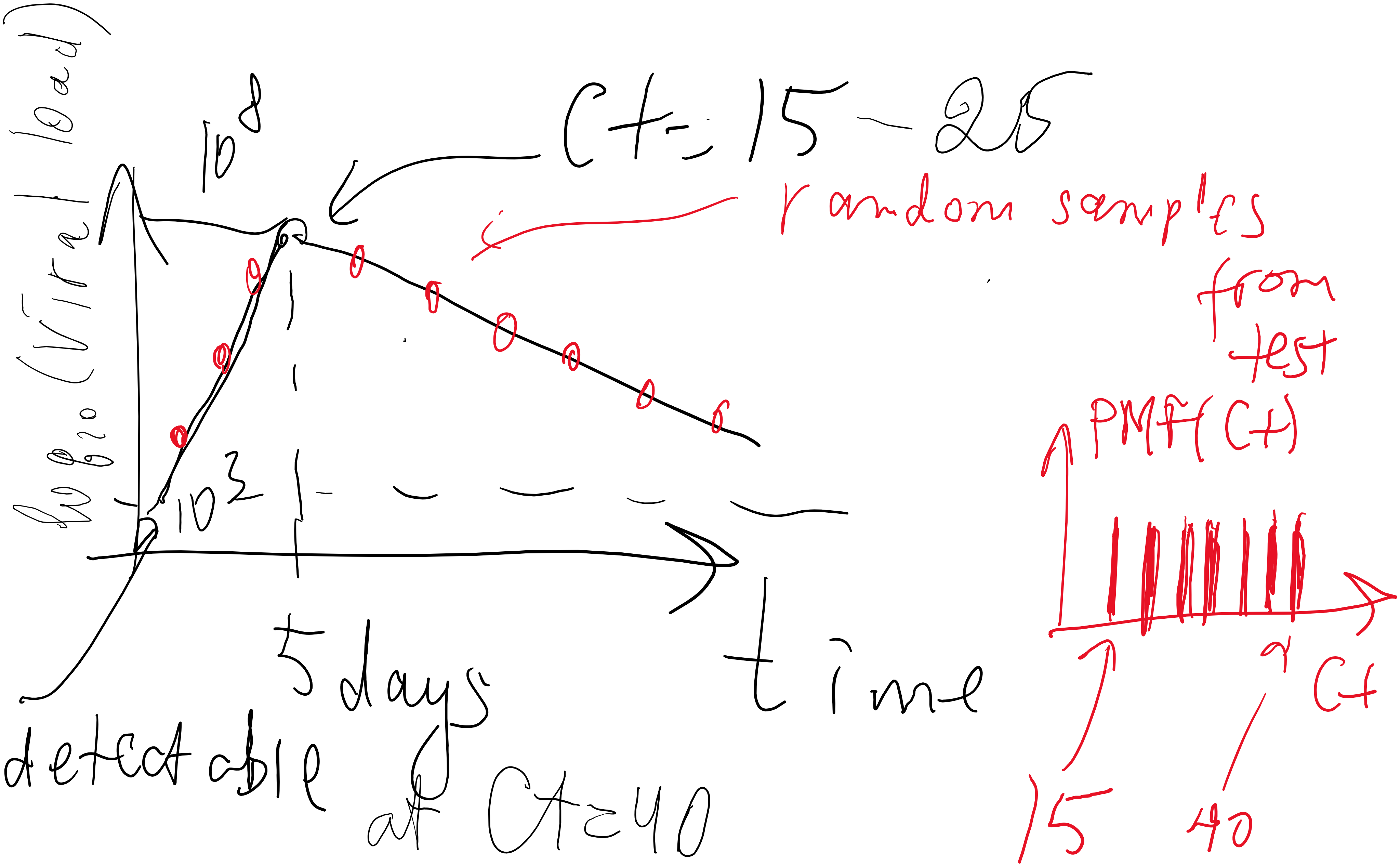
Cycle threshold (Ct) value in COVID-19 infection

What is the Ct value of a PCR test?

Ct = const – log2(viral DNA concentration)



Why Ct distribution should it be uniform?



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

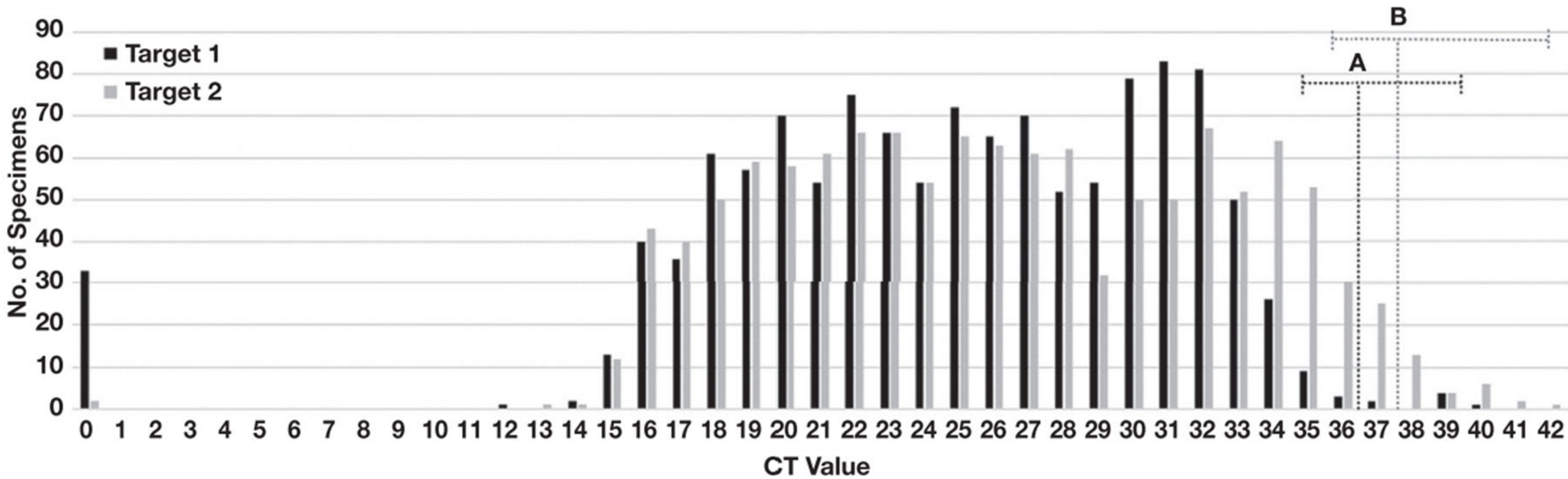
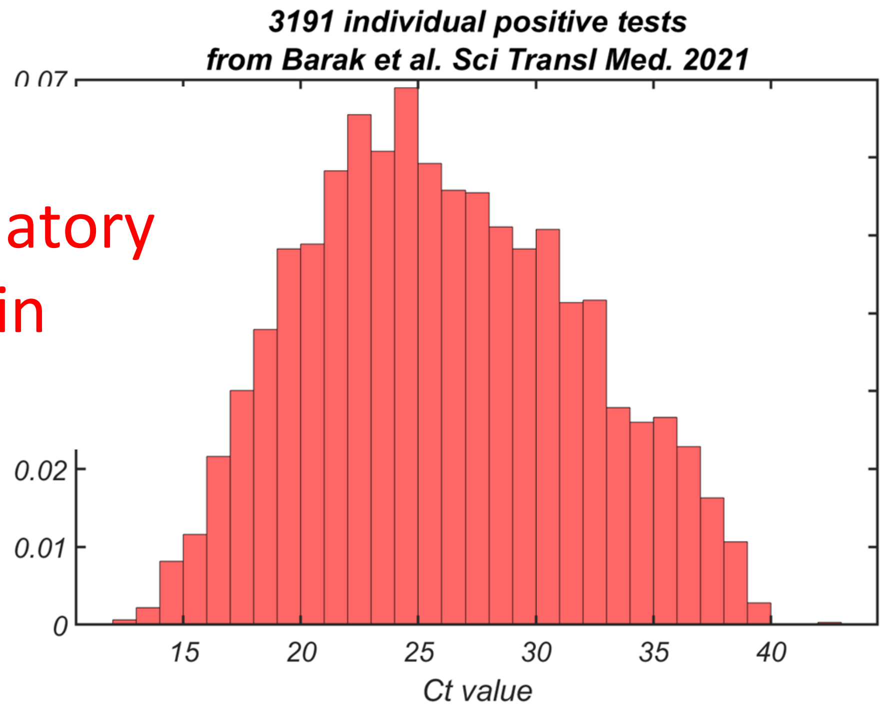


Figure 3 Distribution of cycle threshold (CT) values. The total number of specimens with indicated CT values for Target 1 and 2 are plotted. The estimated limit of detection for (A) Target 1 and (B) Target 2 are indicated by vertical dotted lines. Horizontal dotted lines encompass specimens with CT values 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."

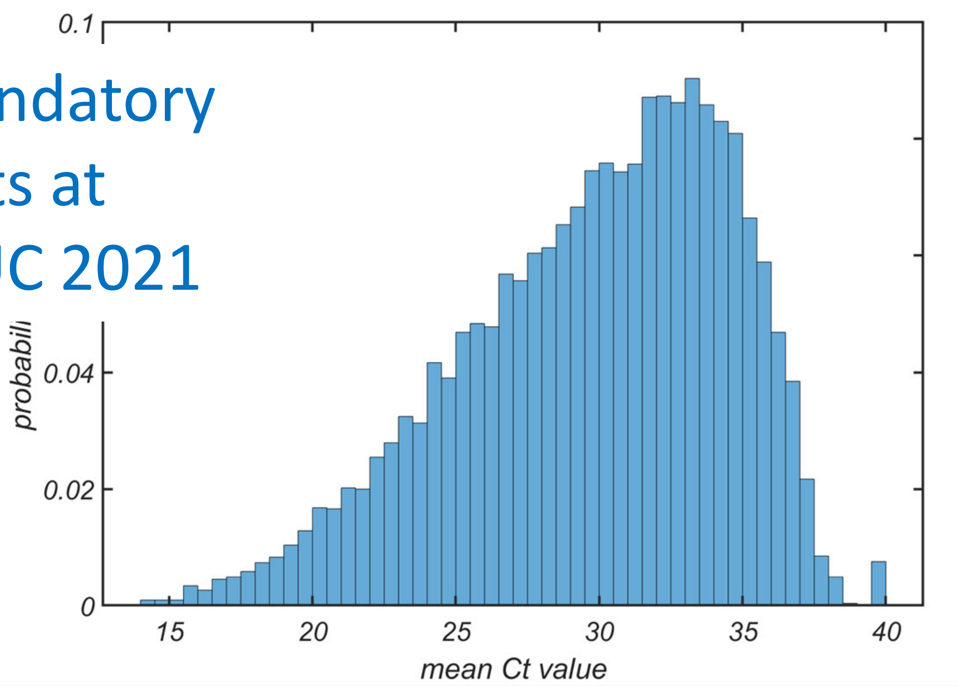
Distribution of SARS-CoV-2 PCR Cycle Threshold Values Provide Practical Insight Into Overall and Target-Specific Sensitivity Among Symptomatic Patients
Blake W Buchan, PhD, Jessica S Hoff, PhD, Cameron G Gmehlin, Adriana Perez, Matthew L Faron, PhD, L Silvia Munoz-Price, MD, PhD, Nathan A Ledeboer, PhD *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 – Ct value is skewed towards high values

Matlab exercise: Uniform distribution

- Generate a sample of size 100,000 for uniform random variable X taking values 1,2,3,...10
- Plot the approximation to the probability mass function based on this sample
- Calculate mean and variance of this sample and compare it to infinite sample predictions: $E[X]=(a+b)/2$ and $V[X]=((a-b+1)^2-1)/12$

Matlab template: Uniform distribution

- **b=10; a=1; % b= upper bound; a= lower bound (inclusive)'**
- **Stats=100000; % sample size to generate**
- **r1=rand(Stats,1);**
- **r2=floor(??*r1)+??;**
- **mean(r2)**
- **var(r2)**
- **std(r2)**
- **[hy,hx]=hist(r2, 1:10); % hist generates histogram in bins 1,2,3...,10**
- **% hy - number of counts in each bin; hx - coordinates of bins**
- **p_f=hy./??; % normalize counts to add up to 1**
- **figure; plot(??,p_f, 'ko-'); ylim([0, max(p_f)+0.01]); % plot the PMF**

Matlab exercise: Uniform distribution

- **b=10; a=1; % b= upper bound; a= lower bound (inclusive)'**
- **Stats=100000; % sample size to generate**
- **r1=rand(Stats,1);**
- **r2=floor(b*r1)+a;**
- **mean(r2)**
- **var(r2)**
- **std(r2)**
- **[hy,hx]=hist(r2, 1:10); % hist generates histogram in bins 1,2,3...,10**
- **% hy - number of counts in each bin; hx - coordinates of bins**
- **p_f=hy./sum(hy); % normalize counts to add up to 1**
- **figure; plot(hx,p_f, 'ko-'); ylim([0, max(p_f)+0.01]); % plot the PMF**

Bernoulli distribution

The simplest non-uniform distribution

p – probability of success (1)

1-p – probability of failure (0)

$$f(x) = P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

Jacob Bernoulli
(1654-1705)
Swiss mathematician (Basel)

- Law of large numbers
- Mathematical constant e=2.718…

Bernoulli distribution

$$f(x) = P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

$$E(X) = 0 \times P(X = 0) + 1 \times P(X = 1) = 0(1 - p) + 1(p) = p$$

$$\text{Var}(X) = E(X^2) - (EX)^2 = [0^2(1 - p) + 1^2(p)] - p^2 = p - p^2 = p(1 - p)$$