# Lecture 04 and 05: Probability Distributions

*Presented By:*
**Dr. Vinay Kulkarni**

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

# Random Variables

- **Random experiment**
  - Process of measurement or observation in which the outcome **cannot** be completely determined in advance
- **Sample space**
  - All possible outcomes of a random experiment
- **Random Variable**
  - A real-valued quantity, or numerical measure, whose value depends on the outcomes of a random experiment
  - Can be **Discrete** or **Continuous**

# Random Variable

- The probability that a Random Variable may assume a particular value is governed by a Probability Function:
  - For Discrete variables
    - **Probability Mass Function (PMF)**
  - For Continuous variables
    - **Probability Density Function (PDF)**
  - For Discrete / Continuous variables
    - **Cumulative Distribution Function**

# Random Variable and Probabilities

- Let X be the Random Variable
- Let x be one of its possible values
- Let P(x) be the probability that X=x
- Then

$$0 \leq P(x) \leq 1$$

$$\sum_{\text{all values of } x} P(x) = 1$$

# Random Variable: Expected Value

- Expected value : E(X)
  - Weighted average, of all possible values, considering their probabilities

- For Discrete random variable

$$\mathrm{E}(\mathrm{X}) = \mu_X = \sum_{all\ values\ of\ x} [x.p(x)]$$

- For Continuous random variable

$$\mathrm{E}(\mathrm{X}) = \mu_X = \int_{-\infty}^{\infty} x.f(x)dx$$

# Random Variable: Variance and Standard Dev

- Variance of a Random Variable

$$\sigma_X^2 = E[(X - \mu_X)^2] = E(X^2) - \mu_X^2$$

- Standard Deviation of a Random Variable

$$\sigma_x = +\sqrt{\sigma_X^2}$$

- Where

$$E(X^2) = \sum_{\text{all values of } x} [x^2 . p(x)], \text{ in the discrete case, and}$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 . f(x) dx, \text{ in the continuous case}$$

- Variance and Standard Deviation reflects the extent to which the Random variable is close to its mean

# Exercise

- Which of the following tables represent valid discrete probability distributions?

(a)

| X | p(x) |
|---|---|
| 1 | 0.2 |
| 2 | 0.35 |
| 3 | 0.12 |
| 4 | 0.40 |
| 5 | -0.07 |

(b)

| x | p(x) |
|---|---|
| 1 | 0.2 |
| 2 | 0.25 |
| 3 | 0.10 |
| 4 | 0.14 |
| 5 | 0.49 |

(c)

| x | p(x) |
|---|---|
| 1 | 0.2 |
| 2 | 0.25 |
| 3 | 0.10 |
| 4 | 0.15 |
| 5 | 0.30 |

- For valid distributions, calculate:
  - The expected value , variance and standard deviation

# Exercise - Answers

- E(X) = 3.10
- Variance = 2.39
- Standard Deviation = 1.54596

# Sample – to – Population

- Goal of all these definitions:
  - Given the nature of the phenomena
    - Probability distributions
  - And a sample with certain deductions
    - Measured observations
  - Predict additional properties and confidence levels
- We therefore need to
  - Understand generic phenomena
  - And the probabilistic nature of their events

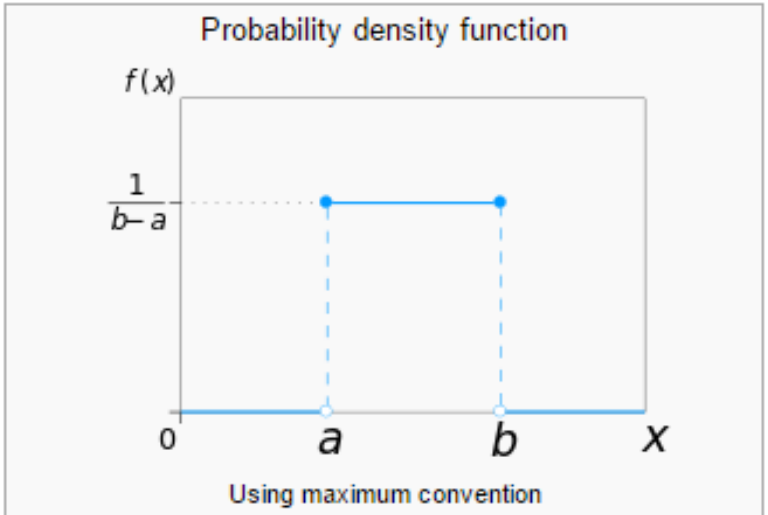
# Probability Distributions

- Binomial Distribution
- Poisson Distribution
- Students T-Distribution
- Chi-Square Distribution
- F-Distribution
- Normal Distribution
- Log normal Distribution
- Other Distributions
  - Bernoulli
  - Geometric
  - Hypergeometric
  - Multinomial
  - Exponential
  - Beta
  - Gamma

# Permutations and Combinations

- Revision of permutations and combinations

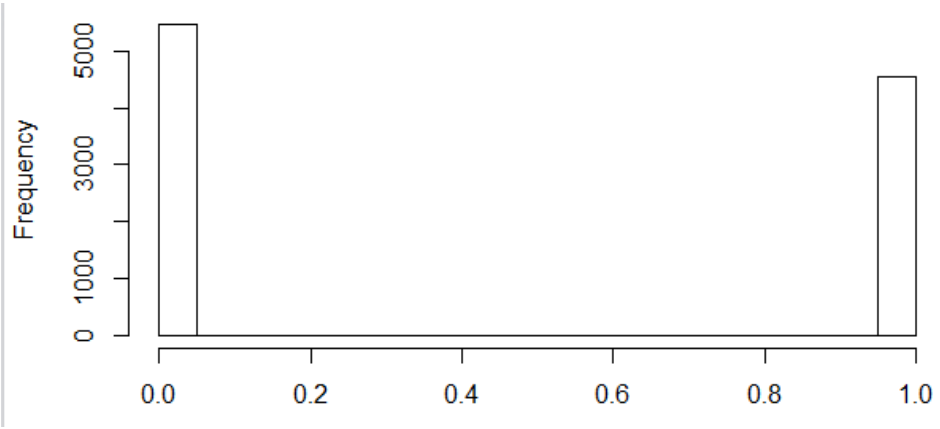
# Uniform Distribution




| | |
|---|---|
| **Notation** | $\mathcal{U}(a,b)$ or $\operatorname{unif}(a,b)$ |
| **Parameters** | $-\infty < a < b < \infty$ |
| **Support** | $x \in [a,b]$ |
| **PDF** | $\begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$ |
| **CDF** | $\begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a,b) \\ 1 & \text{for } x \geq b \end{cases}$ |
| **Mean** | $\frac{1}{2}(a+b)$ |
| **Median** | $\frac{1}{2}(a+b)$ |
| **Mode** | any value in $(a,b)$ |
| **Variance** | $\frac{1}{12}(b-a)^2$ |

# Bernoulli Trials and Bernoulli Distribution

- If an experiment has **only two outcomes**, it is known as a **Bernoulli Trial**
- p = Probability of success
- q = Probability of failure




Bernoulli Dist: p = 0.45

**Bernoulli**

| | |
|---|---|
| Parameters | $0 < p < 1, p \in \mathbb{R}$ |
| Support | $k \in \{0, 1\}$ |
| pmf | $\begin{cases} q = (1-p) & \text{for } k = 0 \\ p & \text{for } k = 1 \end{cases}$ |
| CDF | $\begin{cases} 0 & \text{for } k < 0 \\ 1-p & \text{for } 0 \leq k < 1 \\ 1 & \text{for } k \geq 1 \end{cases}$ |
| Mean | $p$ |
| Median | $\begin{cases} 0 & \text{if } q > p \\ 0.5 & \text{if } q = p \\ 1 & \text{if } q < p \end{cases}$ |
| Mode | $\begin{cases} 0 & \text{if } q > p \\ 0,1 & \text{if } q = p \\ 1 & \text{if } q < p \end{cases}$ |
| Variance | $p(1-p)(=pq)$ |

# Use of Probability Functions of “R”

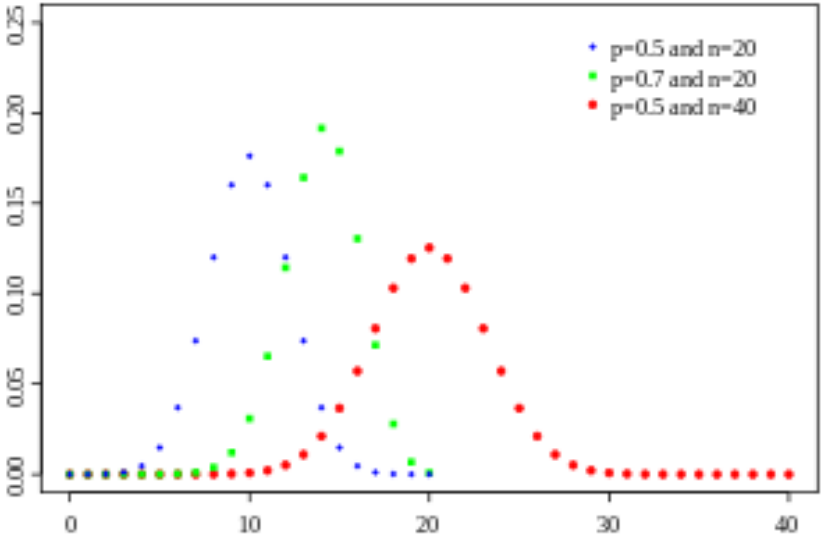
- dxxxx
  - Probability density function
  - Given a ‘number’ (quantile) this function will tell you the probability of getting that number (quantile)
- pxxxx
  - Cumulative probability distribution function
  - Given a ‘number’ (quantile) this function will tell you the cumulative probability of getting all values up to that number
- qxxxx
  - Given a cumulative probability, this function returns the ‘number’ (quantile) associated with that probability
- rxxxx
  - This function generates the required number of data points conforming to the desired probability distribution (as per the specified parameters)

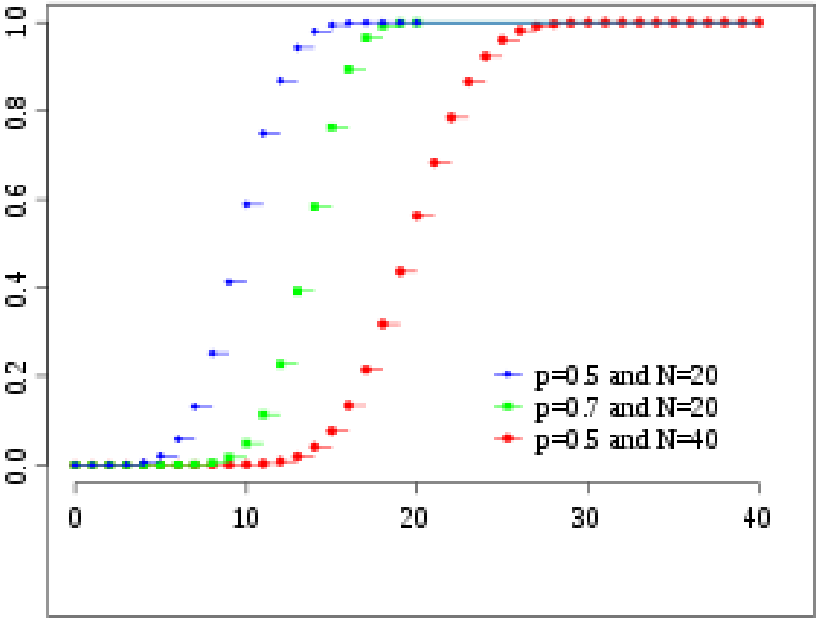
# Binomial Distribution

- If an experiment has **only two outcomes**, it is known as a **Bernoulli Trial**
- Such an experiment is said to have Binomial Probability Distribution, if:
  - There are finite, **independent**, trial
  - Probability of success / failure is constant throughout the experiment
  - We are interested in the **number** of successes 'x', regardless of how they occur
- The number of successes is given by:

$$P(X = x) = P(x) = {}_nC_x\, p^x(1-p)^{n-x} = \binom{n}{x} p^x(1-p)^{n-x},\ x = 0, 1, 2, \ldots, n$$

# Binomial Distribution




| | |
|---|---|
| **Notation** | $B(n, p)$ |
| **Parameters** | $n \in \mathbb{N}_0$ — number of trials<br>$p \in [0,1]$ — success probability in each trial |
| **Support** | $k \in \{0, \ldots, n\}$ — number of successes |
| **pmf** | $\binom{n}{k} p^k (1-p)^{n-k}$ |
| **CDF** | $I_{1-p}(n-k, 1+k)$ |
| **Mean** | $np$ |
| **Median** | $\lfloor np \rfloor$ or $\lceil np \rceil$ |
| **Mode** | $\lfloor (n+1)p \rfloor$ or $\lfloor (n+1)p \rfloor - 1$ |
| **Variance** | $np(1-p)$ |
| **Skewness** | $\frac{1-2p}{\sqrt{np(1-p)}}$ |

Src: Wikipedia

# Exercise

- Construct Binomial Probability Histograms for the following cases
  - n = 10, p = 0.2
  - n = 10, p = 0.5
  - n = 10, p = 0.8

- Hint: Use the Binomial CDF Table to derive the probability values for:
  - x = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

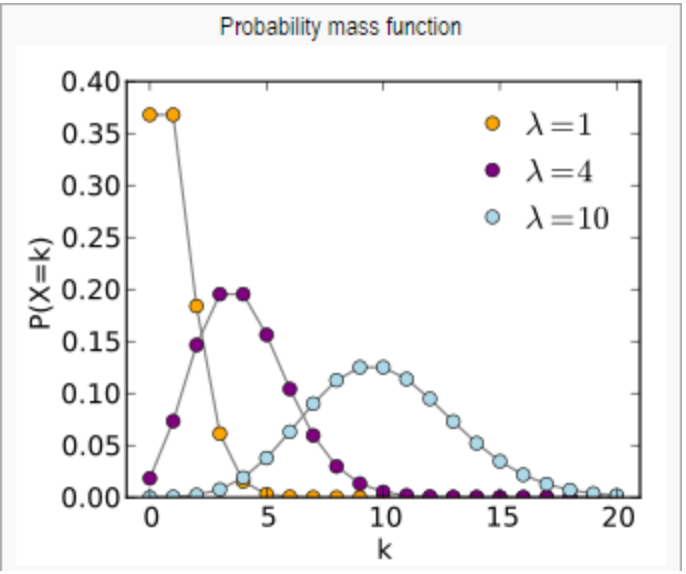
# Exercise: Binomial Distribution

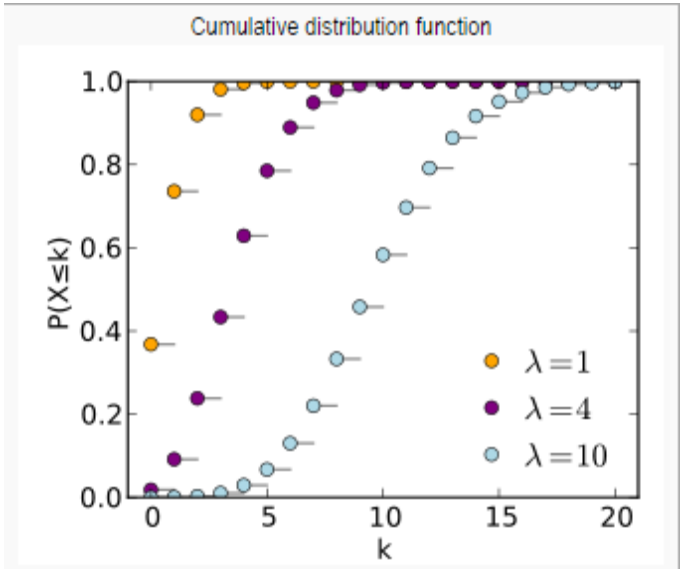
- Given:
  - From clinical trials of a vaccination serum it is observed that 2 in 10 people will develop the disease
- Now find, In a sample size of 10 people
  - What is the probability that at most 3 people will develop the disease
  - What is the probability that exactly 5 people will develop the disease

# Poisson Distribution

- In case of random phenomena
- Where events are continuous
  - Calls arriving at a switchboard during lunch
  - Accidents at an intersection between 10 and noon
  - Misprints per page in a book
- The probability that a continuous measure X will take on value x, in a given unit of measurement, is governed by Poisson Distribution

# Poisson Distribution




| | |
|---|---|
| **Notation** | $\mathrm{Pois}(\lambda)$ |
| **Parameters** | $\lambda > 0$ (real) |
| **Support** | $k \in \{0, 1, 2, 3, \ldots\}$ |
| **pmf** | $\frac{\lambda^k}{k!}e^{-\lambda}$ |
| **CDF** | $\frac{\Gamma(\lfloor k+1 \rfloor, \lambda)}{\lfloor k \rfloor !}$, or $e^{-\lambda}\sum_{i=0}^{\lfloor k \rfloor}\frac{\lambda^i}{i!}$, or $Q(\lfloor k+1 \rfloor, \lambda)$ (for $k \geq 0$, where $\Gamma(x, y)$ is the incomplete gamma function, $\lfloor k \rfloor$ is the floor function, and Q is the regularized gamma function) |
| **Mean** | $\lambda$ |
| **Median** | $\approx \lfloor \lambda + 1/3 - 0.02/\lambda \rfloor$ |
| **Mode** | $\lceil \lambda \rceil - 1, \lfloor \lambda \rfloor$ |
| **Variance** | $\lambda$ |
| **Skewness** | $\lambda^{-1/2}$ |

Src: Wikipedia

# Poisson Distribution

- Probability Mass Function (PMF)

$$P(X = x) = p(x) = \lambda^x e^{-\lambda} / x!, \qquad x = 0, 1, 2, \ldots$$
$$= \text{zero, otherwise.}$$

  - Where:
    - $\lambda$ = average number per unit measurement
    - X = Specific value of the measure
- Expected value / Mean
    - $\lambda$
- Variance
    - $\lambda$

# Exercise

Consider the number of accidents between 8 and 9 am on an intersection on Saturday. From data recorded let the mean of accidents on that intersection have a mean of 4. Hence this follows what we call a Poisson distribution with $\lambda = 4$. Find the probability that on a given Saturday, between 8 and 9 am, there will be:

a) No accident,
b) At least one accident,
c) Exactly 4 accidents.

# Geometric Distribution

- If in an experiment there are only two outcomes: Success | Failure
  - P(s) = p;  P(f) = q;
  - p + q = 1
- We are interested in:
  - Number of trials ‘x’ to get the first success
- Geometric Distribution governs this case

$$P(X = x) = p \cdot q^{x-1}, \qquad x = 1, 2, 3, \ldots$$

$\mu = 1/p$ and $\sigma^2 = (1 - p)/p^2$

# Exercise

- In a certain production process
  - 1 in every 100 items is defective
  - What is the probability that the $5^{th}$ item inspected is the first defect to be found?

# Exercise

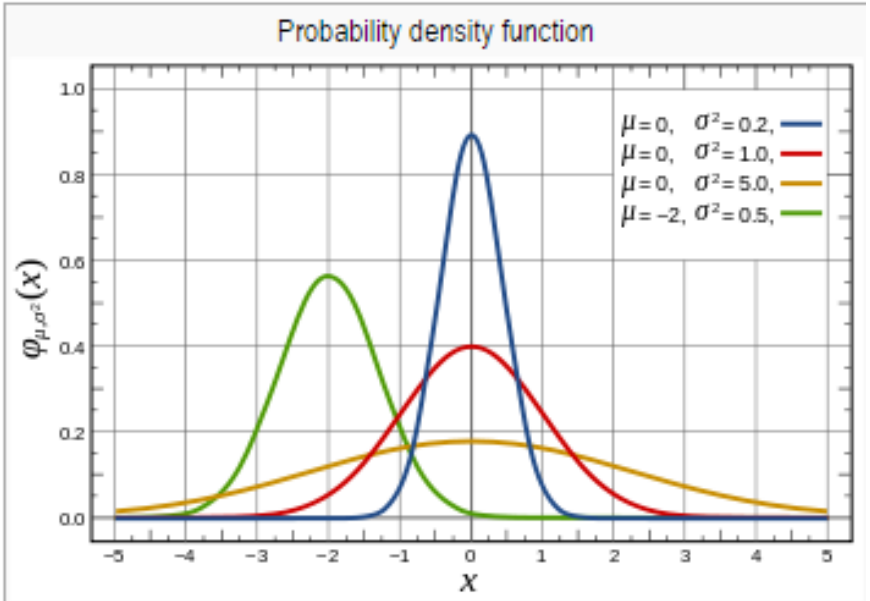
- In a certain production process
  - 1 in every 100 items is defective
  - What is the probability that the $5^{th}$ item inspected is the first defect to be found?
- In this problem:
  - $x = 5$
  - $p = 1/100 = 0.01$
  - $q = (1-p) = 0.99$
  - $P(5) = (0.01) * (0.99)^{(5-1)} = 0.0096$

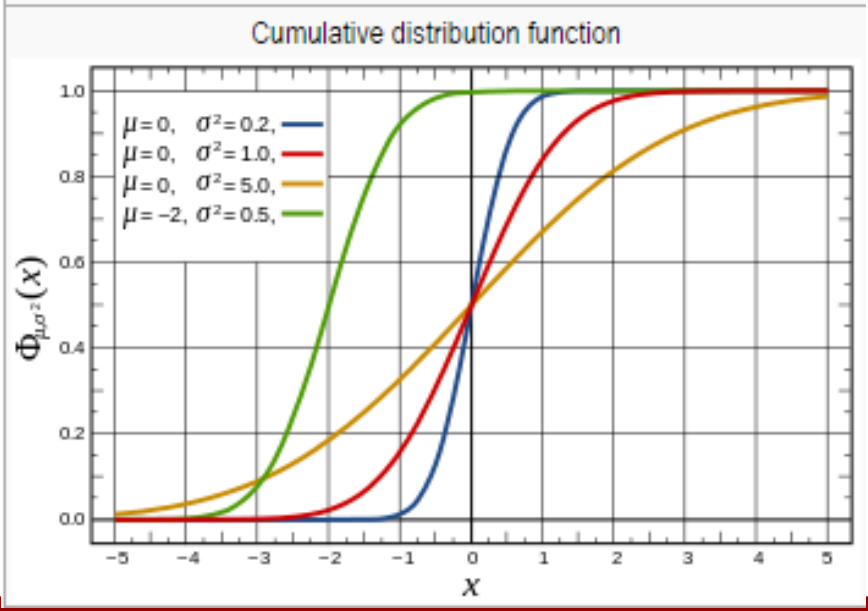
# Normal Distribution

- A very well known Continuous Probability Distribution Function (PDF)
  - Applies to many phenomena
    - Human characteristics, physical quantities and processes, errors in physical and econometric measurements
  - Provides accurate approximation to a large number of probability laws
  - Important role in statistics and inferences
- PDF

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

# Normal Distribution




| | |
|---|---|
| **Notation** | $\mathcal{N}(\mu, \sigma^2)$ |
| **Parameters** | $\mu \in \mathbb{R}$ — mean (location)<br>$\sigma^2 > 0$ — variance (squared scale) |
| **Support** | $x \in \mathbb{R}$ |
| **pdf** | $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ |
| **CDF** | $\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$ |
| **Quantile** | $\mu + \sigma\sqrt{2}\,\operatorname{erf}^{-1}(2F-1)$ |
| **Mean** | $\mu$ |
| **Median** | $\mu$ |
| **Mode** | $\mu$ |
| **Variance** | $\sigma^2$ |
| **Skewness** | 0 |
| **Ex. kurtosis** | 0 |

Src: Wikipedia

# Normal Distribution

- The Cumulative Distribution Function for Normal Distribution is available
  - In the form of a Standard Normal Distribution Table
  - Based on $Z = (x-\mu)/\sigma$
- The Standard Normal Distribution table is used in solving problems

# Example

- The weight of a certain category of watermelon follows normal distribution with mean 1.0 kg and standard deviation of 0.20 kg. Find:
  1. The probability that a watermelon weighs less than 1.5kg
  2. Probability that it weighs between 0.9kg and 1.2kg
  3. Probability that it weighs more than 1.6kg
  4. Percentage of watermelons that weigh between 0.8kg and 1.50kg
  5. Among a group of 300 watermelons how many will weigh between 0.8kg and 1.5kg?

# Answers

1. 0.9938
2. 0.5328
3. 0.0013
4. 83.51%
5. About 251

Weights of fish caught by a certain method are approximately normally distributed with mean of 4.5 lbs. and a standard deviation of 0.50 lbs.

a) What percentage of fish will weigh less than 4 lbs?
b) What percentage of the fish will weigh within one lb. of the average weight?
c) What is the chance that one fish will weigh more than 5 lbs.?

The inside diameter of a piston ring is normally distributed with mean of 4 inches and a standard deviation of 0.01 inches.

a) What percentage of the rings will have an inside diameter exceeding 4.025 inches?
b) What is the probability that a piston ring will have an inside diameter between 3.99 and 4.01 inches?
c) Below what value of the inside diameter will 15% of the rings fall?

Gauges are used to reject all components in which a certain dimension is not within the specifications of 1.5 – d and 1.5 + d. It is known that this dimension is normal distributed with mean 1.50 and standard deviation 0.2. Determine the value of d such the specifications

a) Cover 95% of the components
b) Cover 90% of the components
c) Cover 99.7% of the components