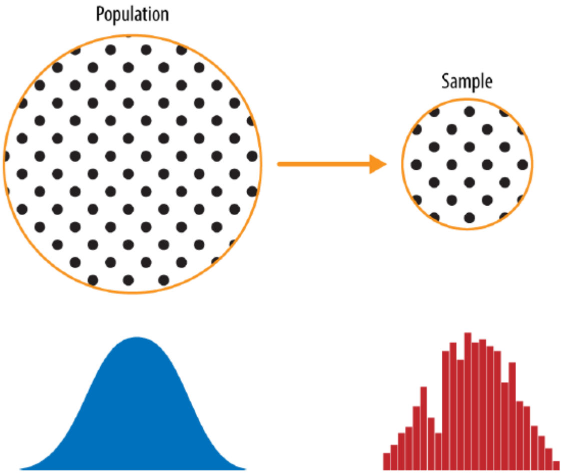
Statistical and Mathematical Methods

Statistical and Mathematical Methods for Data Science
DS5003

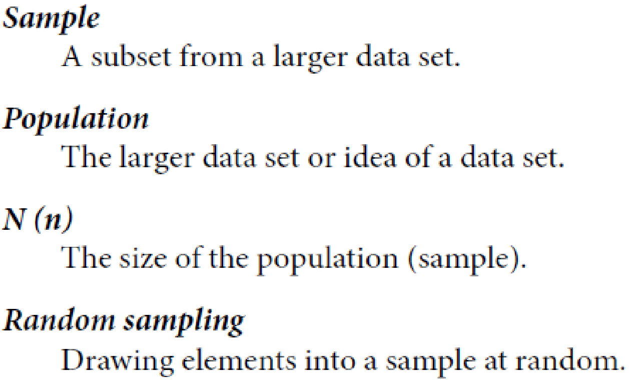
Dr. Nasir Touheed

Statistics

Sample vs Population



Random Sampling



Sample

A subset from a larger data set.

Population

The larger data set or idea of a data set.

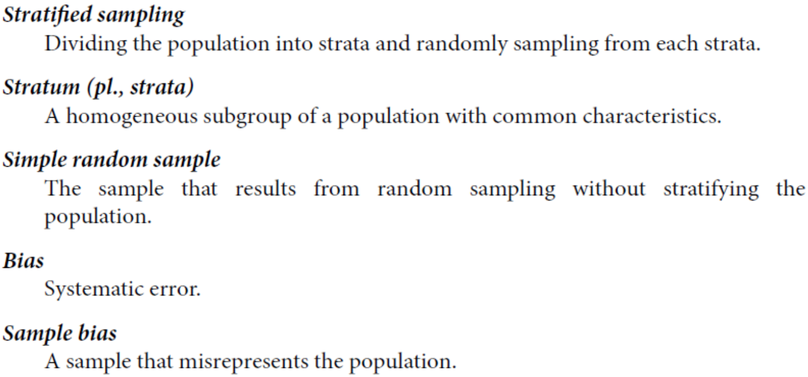
N (n)

The size of the population (sample).

Random sampling

Drawing elements into a sample at random.

Random Sampling



Stratified sampling

Dividing the population into strata and randomly sampling from each strata.

Stratum (pl., strata)

A homogeneous subgroup of a population with common characteristics.

Simple random sample

The sample that results from random sampling without stratifying the population.

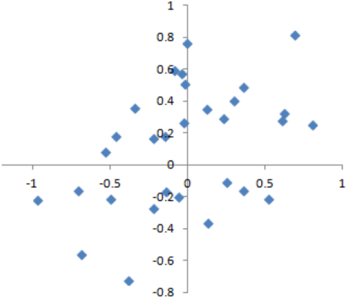
Bias

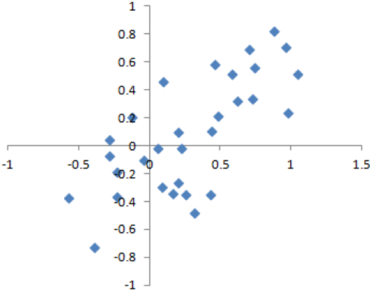
Systematic error.

Sample bias

A sample that misrepresents the population.

Random Error vs Bias Error





Bias

- Selecting a sample to represent the population fairly is actually rather difficult.
- Sampling methods that, by their nature, tend to over- or under-emphasize some characteristics of the population are said to be biased.
- Conclusions based on samples drawn from biased methods are inherently flawed.

Random Sampling

- 1936 election: Franklin Delano Roosevelt vs. Alf Landon
- Literary Digest had called the election since 1916
- Sample size: 2.4 million!
- Prediction: Roosevelt 43%
- Actual: Roosevelt: 62%
- (Literary Digest went bankrupt soon after)

What happened

- Where did the pollsters get their 10 million names?
    - Phone numbers? In 1936, the height of the depression, phones were luxuries. Selecting from a sample of only upper class citizens wouldn't be representative.
    - Driver's licenses and/or memberships in organizations such as country clubs
    - Is this representative of the entire population when the major campaign issue was the economy?

Gallop Survey

- In 1936, George Gallop used a subsample of only 3000 of the 2.4 million responses that the Literary Digest received to reproduce the wrong prediction of Landon's victory over Roosevelt.
- He then used an entirely different sample of 50,000 and predicted that Roosevelt would get 56% of the vote to Landon's 44%.
- Gallop went of to become one of the leading polling companies.

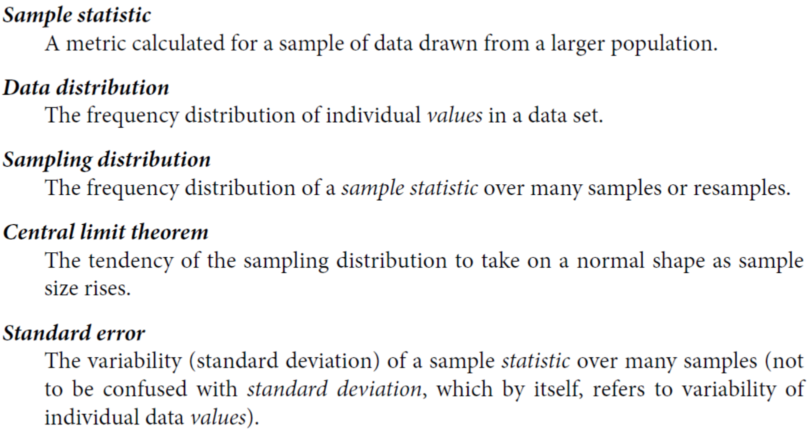
Randomize

- The best defense against bias is randomization, in which each individual is given a fair, random chance of selection.
- Randomization also protects us from the influences of all the features of our population, even one we may not have thought about. It does that by making sure that on average the sample looks like the rest of the population.
- Randomization also makes it possible to draw inferences about the population when we see only a sample.
- The fraction of the population that you've sampled does not matter
- Sample size is of key importance in the design of a sample survey because it determines the balance between how well the survey can measure the population and how much the survey costs

Random Sampling

- Even in the era of big data, random sampling remains an important arrow in the data scientist's quiver.
- Bias occurs when measurements or observations are systematically in error because they are not representative of the full population.
- Data quality is often more important than data quantity, and random sampling can reduce bias and facilitate quality improvement that would otherwise be prohibitively expensive.

Sampling Distribution



Sample statistic

A metric calculated for a sample of data drawn from a larger population.

Data distribution

The frequency distribution of individual *values* in a data set.

Sampling distribution

The frequency distribution of a *sample statistic* over many samples or resamples.

Central limit theorem

The tendency of the sampling distribution to take on a normal shape as sample size rises.

Standard error

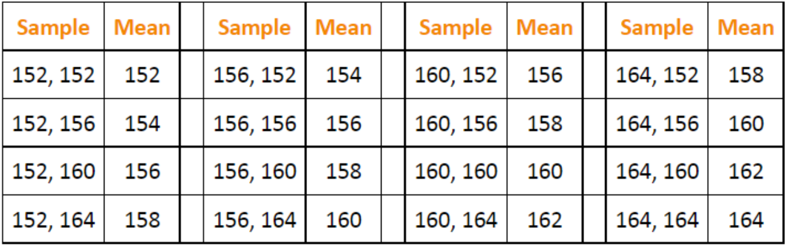
The variability (standard deviation) of a sample *statistic* over many samples (not to be confused with *standard deviation*, which by itself, refers to variability of individual data *values*).

Bias

A rowing team consists of four rowers who weigh 152, 156, 160, and 164 pounds. Find all possible random samples with replacement of size two and compute the sample mean for each one. Use them to find the probability distribution, the mean, and the standard deviation of the sample mean $\overline{X}$.

Example

A rowing team consists of four rowers who weigh 152, 156, 160, and 164 pounds. Find all possible random samples with replacement of size two and compute the sample mean for each one. Use them to find the probability distribution, the mean, and the standard deviation of the sample mean $\overline{X}$.



Sample	Mean		Sample	Mean		Sample	Mean		Sample	Mean
152, 152	152		156, 152	154		160, 152	156		164, 152	158
152, 156	154		156, 156	156		160, 156	158		164, 156	160
152, 160	156		156, 160	158		160, 160	160		164, 160	162
152, 164	158		156, 164	160		160, 164	162		164, 164	164

Example

A rowing team consists of four rowers who weigh 152, 156, 160, and 164 pounds. Find all possible random samples with replacement of size two and compute the sample mean for each one. Use them to find the probability distribution, the mean, and the standard deviation of the sample mean $\overline{X}$.

- The table shows that there are seven possible values of the sample mean $\overline{X}$.
- The value $\overline{x} = 152$ happens only one way (the rower weighing 152 pounds must be selected both times), as does the value $\overline{x} = 164$,
- the other values happen more than one way, hence are more likely to be observed than 152 and 164 are

Example

- A rowing team consists of four rowers who weigh 152, 156, 160, and 164 pounds. Find all possible random samples with replacement of size two and compute the sample mean for each one. Use them to find the probability distribution, the mean, and the standard deviation of the sample mean $\overline{X}$.
- Since the 16 samples are equally likely, we obtain the probability distribution of the sample mean just by counting:

$\overline{x}$	152	154	156	158	160	162	164
$P(\overline{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

Example

$\overline{x}$	152	154	156	158	160	162	164
$P(\overline{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

For the mean and standard deviation of discrete random variable to $\overline{X}$.

For $\mu_{\overline{X}}$ we obtain.

$$\begin{aligned}\mu_{\overline{X}} &= \Sigma\overline{x}\,P(\overline{x}) \\ &= 152\left(\tfrac{1}{16}\right) + 154\left(\tfrac{2}{16}\right) + 156\left(\tfrac{3}{16}\right) + 158\left(\tfrac{4}{16}\right) + 160\left(\tfrac{3}{16}\right) + 162\left(\tfrac{2}{16}\right) + 164\left(\tfrac{1}{16}\right) \\ &= 158\end{aligned}$$

For $\sigma_{\overline{X}}$ we first compute $\Sigma\overline{x}^2P\left(\overline{x}\right)$:

$$152^2\left(\frac{1}{16}\right) + 154^2\left(\frac{2}{16}\right) + 156^2\left(\frac{3}{16}\right) + 158^2\left(\frac{4}{16}\right) + 160^2\left(\frac{3}{16}\right) + 162^2\left(\frac{2}{16}\right) + 1$$

which is 24,974, so that

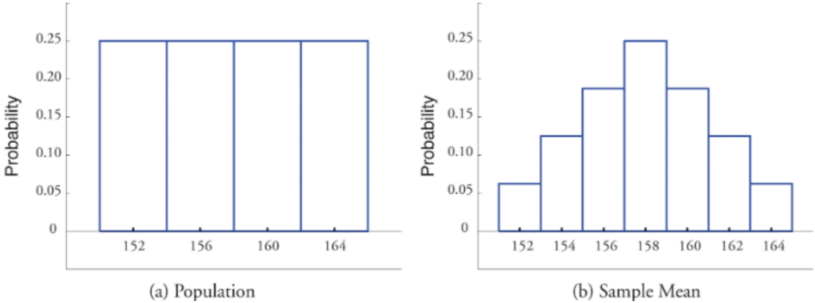
$$\sigma_{\overline{X}} = \sqrt{\Sigma\overline{x}^2P\left(\overline{x}\right) - \mu_{\overline{x}}^2} = \sqrt{24{,}974 - 158^2} = \sqrt{10}$$

The Mean and Standard Deviation of the Sample Mean

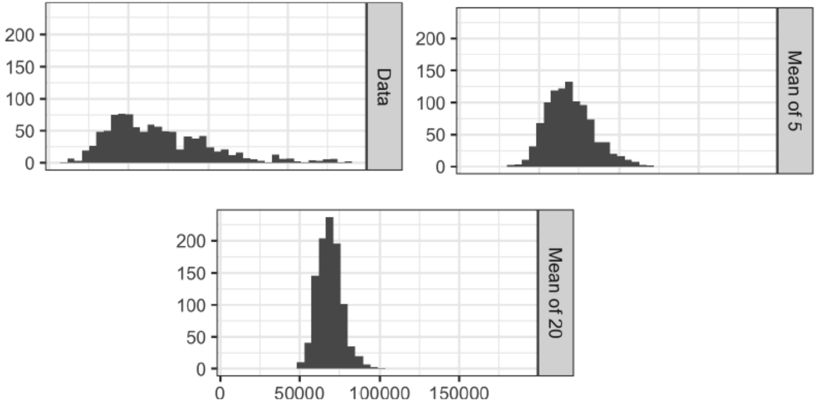
- The mean and standard deviation of the population 152,156,160,164 in the example are $\mu = 158$ and $\sigma = \sqrt{20}$
- The mean of the sample mean $\overline{X}$ that we have just computed is exactly the mean of the population.
- The standard deviation of the sample mean $\overline{X}$that we have just computed is the standard deviation of the population divided by the square root of the sample size: $\sqrt{10} = \frac{\sqrt{20}}{\sqrt{2}}$
- These relationships are not coincidences, but are illustrations of the following formulas.
  - Suppose random samples of size n are drawn from a population with mean μ and standard deviation σ.
  - The mean $\mu_{\overline{X}}$ and standard deviation $\sigma_{\overline{X}}$ of the sample mean $\overline{X}$ satisfy

$$\mu_{\overline{X}} = \mu \quad \text{and} \quad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

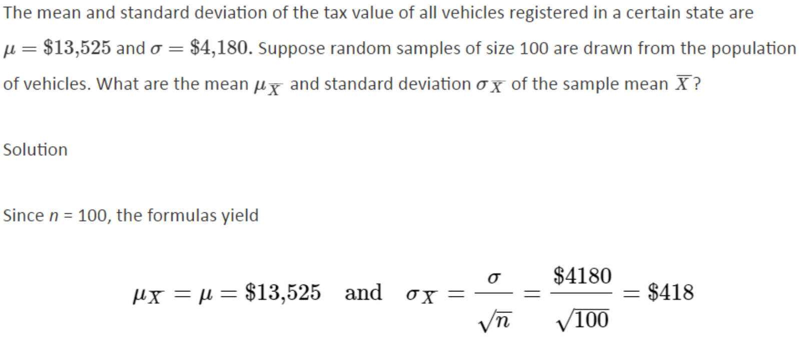
The Mean and Standard Deviation of the Sample Mean



Sampling Distribution



The Mean and Standard Deviation of the Sample Mean



The mean and standard deviation of the tax value of all vehicles registered in a certain state are $\mu = \$13{,}525$ and $\sigma = \$4{,}180$. Suppose random samples of size 100 are drawn from the population of vehicles. What are the mean $\mu_{\overline{X}}$ and standard deviation $\sigma_{\overline{X}}$ of the sample mean $\overline{X}$?

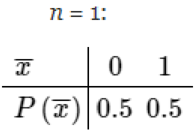
Solution

Since *n* = 100, the formulas yield

$$\mu_{\overline{X}} = \mu = \$13{,}525 \quad \text{and} \quad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{\$4180}{\sqrt{100}} = \$418$$

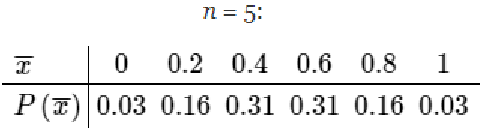
Example

- Suppose we take samples of size 1, 5, 10, or 20 from a population that consists entirely of the numbers 0 and 1, half the population 0, half 1, so that the population mean is 0.5.
- The sampling distributions are:



$n = 1$:

$\overline{x}$	0	1
$P(\overline{x})$	0.5	0.5

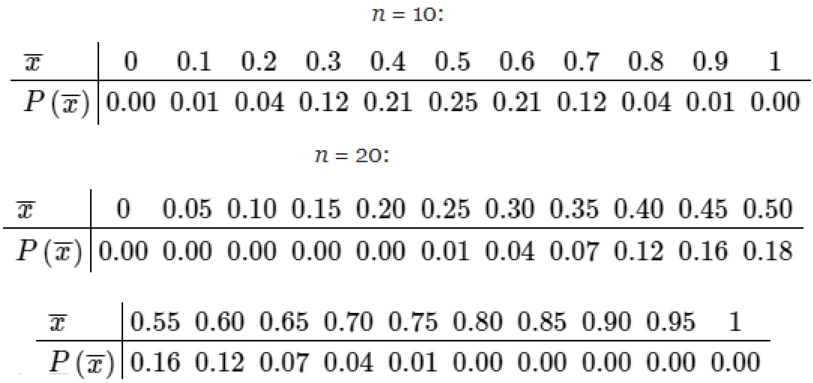


$n = 5$:

$\overline{x}$	0	0.2	0.4	0.6	0.8	1
$P(\overline{x})$	0.03	0.16	0.31	0.31	0.16	0.03

Example

The sampling distributions are:



$n = 10$:

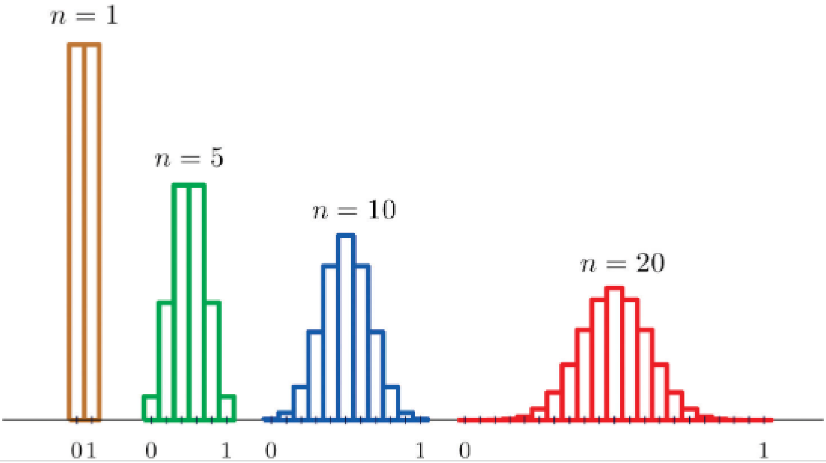
$\overline{x}$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$P(\overline{x})$	0.00	0.01	0.04	0.12	0.21	0.25	0.21	0.12	0.04	0.01	0.00

$n = 20$:

$\overline{x}$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1
$P(\overline{x})$	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.07	0.12	0.16	0.18	0.16	0.12	0.07	0.04	0.01	0.00	0.00	0.00	0.00	0.00

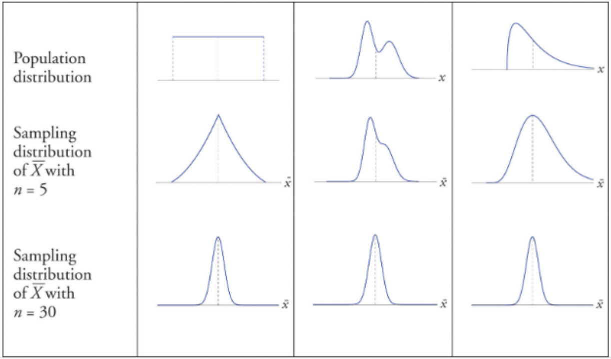
Example

The sampling distributions are:



The Central Limit Theorem

- For samples of size 30 or more, the sample mean is approximately normally distributed, with mean $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$, where n is the sample size.
- The larger the sample size, the better the approximation.



The Central Limit Theorem

- there are two separate random variables (and therefore two probability distributions) at play:
- X, the measurement of a single element selected at random from the population;
- The distribution of X is the distribution of the population, with mean the population mean μ and standard deviation the population standard deviation σ;
- $\overline{X}$, the mean of the measurements in a sample of size n; the distribution of $\overline{X}$ is its sampling distribution, with mean $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$

Continuous Distribution

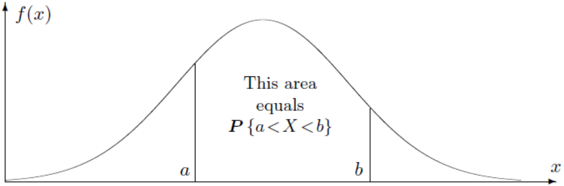
- For all continuous variables, the probability mass function (pmf) is always equal to zero,1 $P(x) = 0 \ \ \forall x$.
- we can use the cumulative distribution function (cdf) $F(x)$

$$F(x) = P\{X \leq x\} = P\{X < x\}$$

- Probability density function (pdf, density) is the derivative of the cdf, $f(x) = F'(x)$. The distribution is called continuous if it has a density.

Continuous Distributions

$$\boxed{\begin{gathered} f(x) = F'(x) \\ \boldsymbol{P}\{a < X < b\} = \int_a^b f(x)dx \end{gathered}}$$



Continuous Distributions

- The lifetime, in years, of some electronic component is a continuous random variable with the density $$f(x) = \begin{cases} \dfrac{k}{x^3} & \text{for} \quad x \geq 1 \\ 0 & \text{for} \quad x < 1. \end{cases}$$
- Find k, draw a graph of the cdf $F(x)$, and compute the probability for the lifetime to exceed 5 years

Continuous Distributions

- Find k from the condition $\int f(x)dx = 1$

$$\int_{-\infty}^{+\infty} f(x)dx = \int_{1}^{+\infty} \frac{k}{x^3}dx = -\frac{k}{2x^2}\bigg|_{x=1}^{+\infty} = \frac{k}{2} = 1.$$

$$F(x) = \int_{-\infty}^{x} f(y)dy = \int_{1}^{x} \frac{2}{y^3}dy = -\frac{1}{y^2}\bigg|_{y=1}^{x} = 1 - \frac{1}{x^2}$$

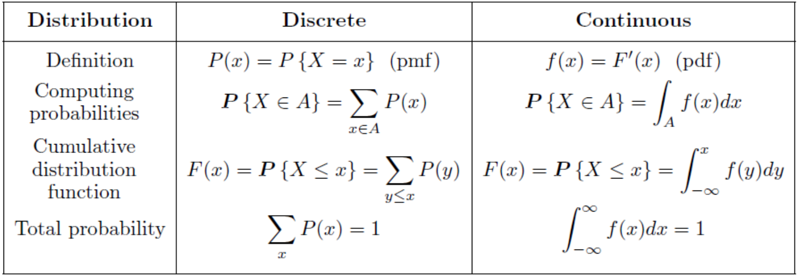
- Hence, k = 2. Integrating the density, we get the cdf
- Next, compute the probability for the lifetime to exceed 5 years,

$$P\{X > 5\} = 1 - F(5) = 1 - (1 - \frac{1}{5^2} = 0.04$$

- We can also obtain this probability by integrating the density

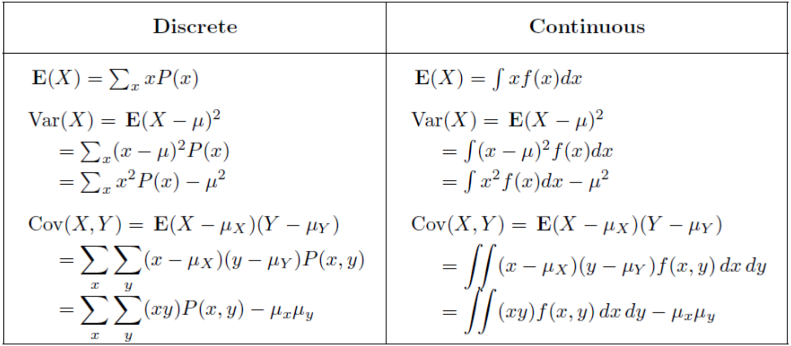
$$\boldsymbol{P}\{X > 5\} = \int_{5}^{+\infty} f(x)dx = \int_{5}^{+\infty} \frac{2}{x^3}dx = -\frac{1}{x^2}\bigg|_{x=5}^{+\infty} = \frac{1}{25} = 0.04.$$

Continuous Distributions



Distribution	Discrete	Continuous
Definition	$P(x) = P\{X = x\}$ (pmf)	$f(x) = F'(x)$ (pdf)
Computing probabilities	$\boldsymbol{P}\{X \in A\} = \sum_{x \in A} P(x)$	$\boldsymbol{P}\{X \in A\} = \int_A f(x)dx$
Cumulative distribution function	$F(x) = \boldsymbol{P}\{X \leq x\} = \sum_{y \leq x} P(y)$	$F(x) = \boldsymbol{P}\{X \leq x\} = \int_{-\infty}^{x} f(y)dy$
Total probability	$\sum_{x} P(x) = 1$	$\int_{-\infty}^{\infty} f(x)dx = 1$

Moments for discrete and continuous distributions



Discrete	Continuous
$\mathbf{E}(X) = \sum_x xP(x)$	$\mathbf{E}(X) = \int xf(x)dx$
$\text{Var}(X) = \mathbf{E}(X-\mu)^2$ $= \sum_x (x-\mu)^2 P(x)$ $= \sum_x x^2 P(x) - \mu^2$	$\text{Var}(X) = \mathbf{E}(X-\mu)^2$ $= \int (x-\mu)^2 f(x)dx$ $= \int x^2 f(x)dx - \mu^2$
$\text{Cov}(X,Y) = \mathbf{E}(X-\mu_X)(Y-\mu_Y)$ $= \sum_x \sum_y (x-\mu_X)(y-\mu_Y)P(x,y)$ $= \sum_x \sum_y (xy)P(x,y) - \mu_x\mu_y$	$\text{Cov}(X,Y) = \mathbf{E}(X-\mu_X)(Y-\mu_Y)$ $= \iint (x-\mu_X)(y-\mu_Y)f(x,y)\,dx\,dy$ $= \iint (xy)f(x,y)\,dx\,dy - \mu_x\mu_y$

$$f(x) = 2x^{-3} \text{ for } x \geq 1.$$

Its expectation equals

$$\mu = \mathbf{E}(X) = \int x\,f(x)dx = \int_1^\infty 2x^{-2}dx = -2x^{-1}\Big|_1^\infty = 2.$$

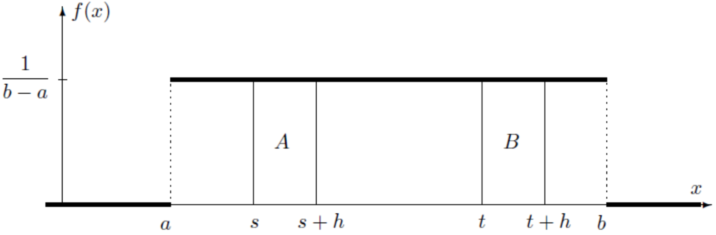
Uniform Distributions

- a random variable with any thinkable distribution can be generated from a Uniform random variable
- Uniform distribution is used in any situation when a value is picked "at random" from a given interval; that is, without any preference to lower, higher, or medium values.
- For example, locations of errors in a program, birthdays throughout a year, and many continuous random variables modulo 1, modulo 0.1, 0.01, etc., are uniformly distributed over their corresponding intervals.
- To give equal preference to all values, the Uniform distribution has a constant density On the interval (a, b), its density equals

$$f(x) = \frac{1}{b-a} a < x < b,$$

because the rectangular area below the density graph must equal 1.

Uniform Distributions



$$\boxed{\begin{aligned} (a,b) &= \text{range of values} \\ f(x) &= \frac{1}{b-a}, \quad a<x<b \\ \mathbf{E}(X) &= \frac{a+b}{2} \\ \text{Var}(X) &= \frac{(b-a)^2}{12} \end{aligned}}$$