Probability Distributions

Statistics

Populations

- Forming judgments about a population *P* based on a sample from *P* is called *statistical inference*.
- A random number drawn from a population is called a *random variable* (r.v.).
- Prior to being observed, a random variable can be any value in the population.
- Once observed, the value of the random variable is known.
- Not all values (or range of values) of a population are equally likely.
- We are interested in the distribution of the random variable

Probability Distributions

- Consider an experiment: toss a coin
  - Coin comes up either a head or a tail
- Another experiment: throw a dice
  - Either a 1, 2, 3, 4, 5, 6 will come up

- With each of the outcomes there is a probability associated
  - Coin Toss: probability of 0.5 for either a head or a tail
  - Throw of Dice: probability of 1/6 for each of 1, 2, 3, 4, 5, 6

Random Variables

- A Probability Distribution assigns a probability to each of the possible outcomes of a random experiment

- Constant: the value does not change
- Variable: the value can change

- Random variable: a variable whose value depends on chance, it is random (stochastic variable)

Sampling

- random.choice() generates observations of a discrete random variable from the numpy package.
- numpy.random.choice(a, size= n, replace=True, p= p1) will choose *n* values in array *a* according to probabilities in vector *p1*.

Sample example in python

```
import numpy as np
list1 = [0,1,2]
p1=[1/4,1/2,1/4]
print( np.random.choice(list1, size=3, replace=False, p=p1))
[1 0 2]
```

Discrete Vs Continuous Random Variables

- *A **discrete random variable*** has a countable number of possible values.
- Examples: The sum of two dice.

  Number of students in a class room

  Number of shots scored by basket player
- A ***continuous random variable*** is a random variable where the data can take infinitely many values.
- Examples: The time required to run a mile

  The total sugar present in an orange

  The height of a nurse

Discrete Random Variables

- For a discrete random variable, *X*, and we can find the probability it will be equal to a value *k,* P(*X* = *k*).

- The *population mean* of a discrete random variable, *X,* is the expected value of *X:*

$$\mu = E(X) = \sum_k k\, P(X = k)$$

- The *population standard deviation* (denoted σ) is the square root of the variance:

$$\sigma^2 = VAR(X) = E((X - \mu)^2).$$

Continuous random variables

- For a continuous random variable, X, and k a value in the range of X, $P(X = k)$ makes no sense because the range has too many values.
- Instead we specify probabilities as the chance that X is in some interval.
- E.g. $P(a < X \leq b)$ is the chance that X is more than a but less or equal to b.

Continuous random variables

- These probabilities are given in terms of an area under a specific curve.
- Such a curve $f(x)$ is the density function of X, i.e.

$P(a < X \leq b) = \int_a^b f(x)dx.$

Useful properties:

- $P(a < X \leq b) = P(X \leq b) - P(X \leq a)$, and
- $P(X \leq b) = 1 - P(X > b)$.

Density function

- The *probability mass function* (p.m.f) of a discrete random variable X is $f(k) = P(X = k)$.
- The *cumulative distribution function* (c.d.f) is $F(b) = P(X \leq b) = \sum_{k \leq b} P(X = k)$

Density function

- The *probability density function* (p.d.f.) of a continuous random variable *X* with domain *S* is an integrable function *f*(*x*) satisfying the following:

1. *f*(*x*) is positive everywhere in the domain *S*, that is, *f*(*x*) > 0, for all *x* in *S*
2. The area under the curve *f*(*x*) in the domain *S* is 1, that is:

$$\int_S f(x)dx = 1$$

3. If *f*(*x*) is the p.d.f. of *X*, then the probability that *X* belongs to *A*, where *A* is some interval (subset of S), is given by the integral of *f*(*x*) over that interval, that is:

$$P(X \in A) = \int_A f(x)dx$$

Sampling

- To perform statistical inference about a parent population we need a sample from the population.
- A *sample* is a sequence of random variables $X_1, X_2, \ldots, X_n$.
- A sequence is *identically distributed* if every random variable has the same distribution.

Sampling

- A sequence is *independent* if knowing the value of any random variable in the list does not give information about the outcome of the others.
- A *random sample* is a sequence that is both identically distributed and independent (denoted *i.i.d.*).
- Example: Toss a coin n times, let X_i be 1 if the ith toss is heads (0 otherwise), then $X_1, X_2, \ldots, X_n$ is an i.i.d sequence.

Sampling

- If we randomly select a sample from a finite population, the values will be independent if we sample with replacement.
- To produce an i.i.d sample using random.choice(), we need to specify replace=True (default is True in numpy).

Example: roll of a dice

```
# roll a dice 10 times
dice1 = [1, 2, 3, 4, 5, 6]
print(np.random.choice(dice1, 10))
[5 3 3 1 2 5 4 1 4 5]



# sum of two dice rolled 10 times
sum1=np.random.choice(dice1, 10)+np.random.choice(dice1, 10)
print(sum1)
[12 8 9 9 9 6 3 9 6 7]
```

Example: opinion poll

- An opinion poll tries to estimate the proportion of the population that votes yes on a given issue.
- It is a random sample from the target population if each person polled is randomly chosen with replacement.
- Example: suppose that in a target population of 10000 people, 6200 will vote yes.
- A sample of size 10 can be generated by:

```
pop=np.repeat([0,1], [3800,6200])

print(np.random.choice(pop, 10))
[0 1 1 1 0 1 1 0 1 1]
```

Sampling distribution

- A *statistic* is a value derived from a random sample.
- E.g. sample mean $\bar{X} = (X_1 + X_2 + \ldots X_n)/n$ and the sample median.
- Since a statistic depends on a random sample, it is a random variable.
- The distribution of a statistic is called its *sampling distribution*.
- The distribution of a statistic can be quite complicated; however for many common statistics, properties of the sampling distributions are known (with respect to the population parameters).

Recap on important terminology

- Probability Distribution
- Random Variable
- Pdf/pmf
- cdf
- i.i.d. – independently and identically distributed
- Sampling distribution

Families of distributions

- In statistics there are several standard families of distributions (e.g. normal, uniform, Bernoulli,. . .)
- Each family is described by a parameter characterizing the distribution.
- We can use scipy.stats to return the p.d.f., the c.d.f, the quantiles and a random sample of the specified distribution.

Functions for Probability Distributions

- rvs: Random Variates
- pdf: Probability Density Function
- cdf: Cumulative Distribution Function
- ppf: Percent Point Function (Inverse of CDF)
- sf: Survival Function (1-CDF)
- isf: Inverse Survival Function (Inverse of SF)
- stats: Return mean, variance, (Fisher’s) skew, or (Fisher’s) kurtosis
- moment: non-central moments of the distribution

Parameters

- The parameters of a distribution define the distribution – determine its shape
- Change the values of the parameters and the distribution changes
- Distributions are defined by a number of parameters

Bernoulli distribution

- A Bernoulli random variable X has only two values: 0 or 1.
  The distribution of X is characterized by $p = P(X = 1)$.
- This distribution is denoted Bernoulli(p).
  Ex: toss a coin, let $X = 1$ if a heads occurs. Then $p = 1/2$ corresponds to a fair coin.
- A sequence of coin tosses is an i.i.d sequence, called a sequence of Bernoulli trials.

Bernoulli distribution

- A Bernoulli random variable has mean $\mu = p$ and variance $\sigma^2 = p(1-p)$.
- np.random.choice() can be used to generate random Bernoulli trials.

Binomial distribution – Discrete Distribution

- A *binomial random variable* counts the number of successes (i.e. 1s) in n Bernoulli trials.
- Parameters are the number of trials n and success probability p.
- This distribution is denoted Binomial(n, p).
- Range of X is 0,1,2,...,n.

Binomial Distribution

- Gives the probability for the number of successes in a sequence of n independent yes/no experiments
- Each of the individual experiments has a probability p of success
- Only two possible outcomes: success and failure

Binomial distribution

- The distribution of a binomial random variable *X* is

$$\mathrm{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where $\binom{n}{k}$ is the binomial coefficient

$$\binom{n}{k} = \frac{n!}{(n-k)!\,k!}$$

- $\binom{n}{k}$ is the number of ways k objects can be chosen from n distinct objects.

Binomial distribution

- The mean (expected value) of a Binomial(n, p) random variable is $\mu = np$.
- The standard deviation is $\sigma = \sqrt{np(1-p)}$
- In scipy.stats the family name is binom. Parameters are n= for *number of trials* and p= for *probability of success*.

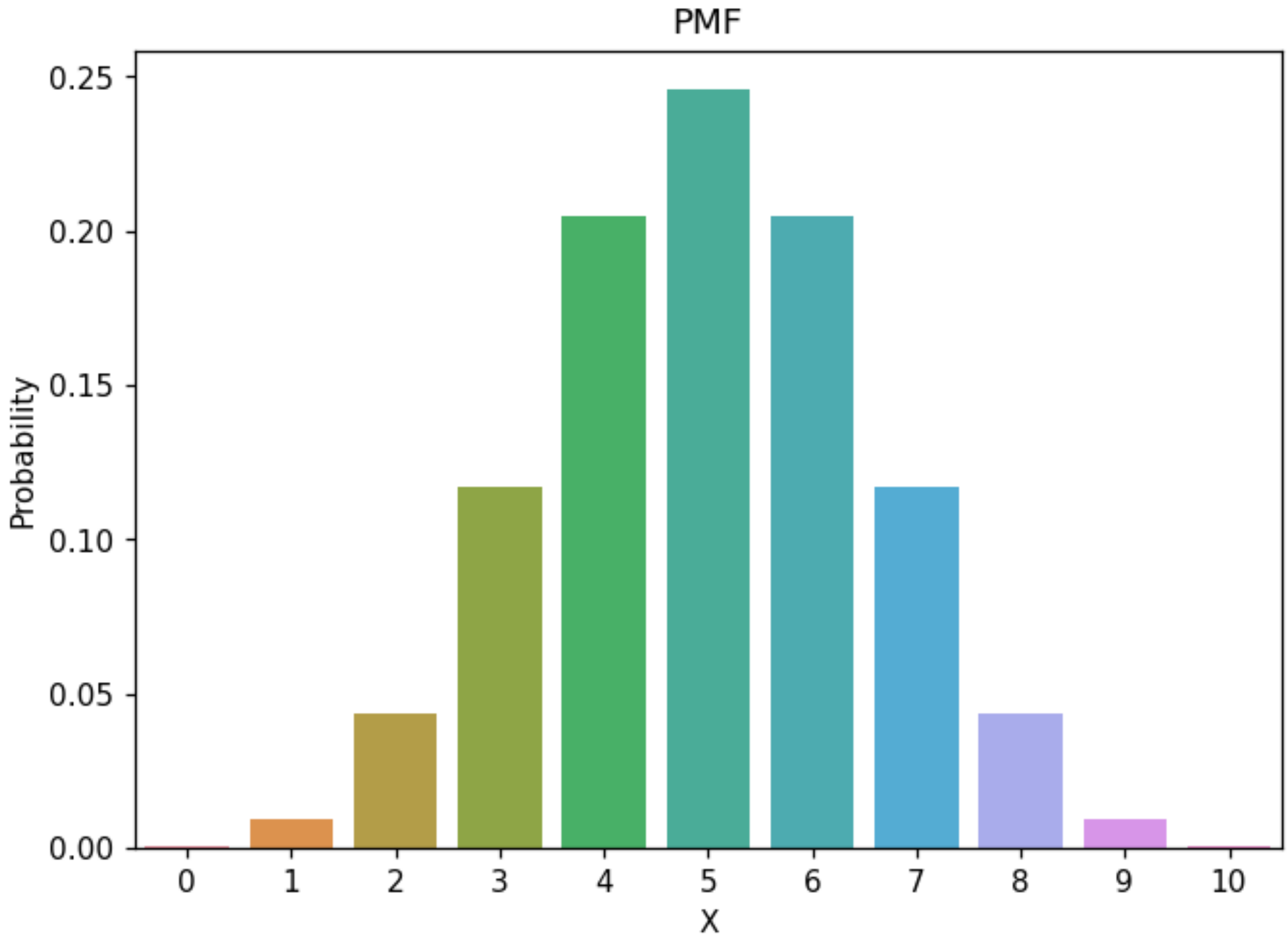
	PURPOSE	SYNTAX	EXAMPLE
rvs	Generates random number variates	binom.rvs(n,p, loc=0, size=1, random_state=None)	binom.rvs(n=20, p=0.25, size=1000) Generates 1000 number of success from 20 Bernoulli trials with probability of success is 0.25
pmf	Probability Mass Function (PMF)	binom.pmf(k, n, p, loc=0)	binom.pmf(5,n=20,p=0.25) Gives the probability of having 5 success out of 20 Bernoulli trials with probability of success is 0.25
cdf	Cumulative Distribution Function (CDF)	binom.cdf(k, n, p, loc=0)	binom.cdf(5, n=20,p=0.25) Gives the cumulative probability of having 5 or less number of success out of 20 Bernoulli trials with probability of success is 0.25
ppf	Quantile Function – inverse of CDF	binom.ppf(q, n, p, loc=0)	binom.ppf(0.5,20,0.25) Gives the smallest value at which the CDF of the binomial distribution is 0.5 out of 20 Bernoulli trials with probability of success is 0.25

Example: Binomial Distribution

- Toss a coin ten times. Let *X* be the number of heads. *X* has distribution Binomial(10, 1/2).

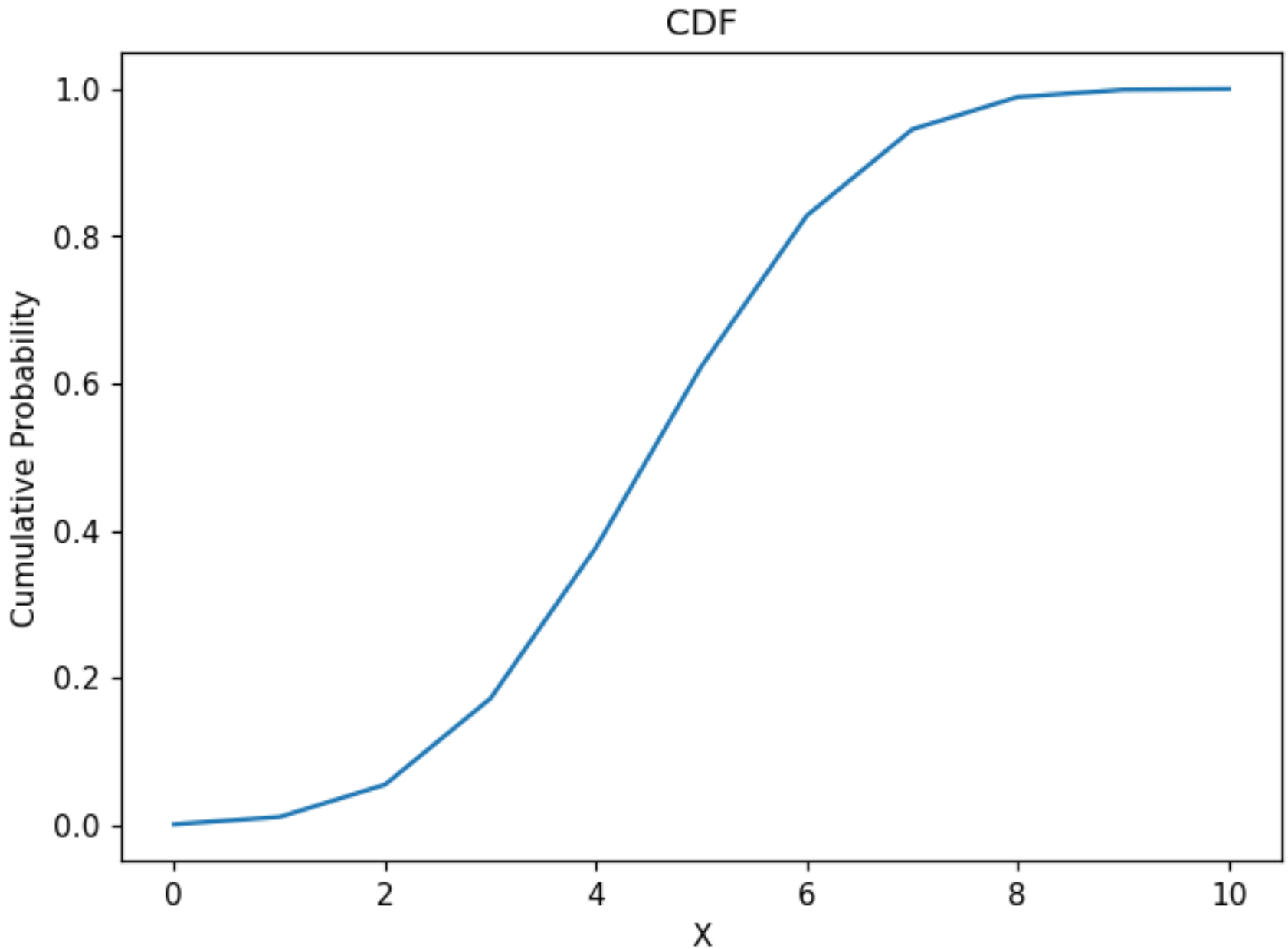
```
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import binom
```

```
n = 10
p=0.5
x = binom.pmf(np.arange(0,11,1), n, p)
sns.barplot(x=np.arange(0,11,1), y=x)
plt.xlabel('X')
plt.ylabel('Probability')
plt.title('PMF')
```


PMF
Probability
0.25
0.20
0.15
0.10
0.05
0.00
0
1
2
3
4
5
6
7
8
9
10
X

Example: Binomial Distribution

```
y= binom.cdf(np.arange(0,11,1), n, p)
plt.plot(y)
plt.xlabel('X')
plt.ylabel('Cumulative Probability')
plt.title('CDF')
```


CDF
Cumulative Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
X

Example: Binomial Distribution

- What is the probability of seeing five heads?

```
binom.pmf(5,n=10,p=0.5)
0.24609375000000025
```

- What is the probability of seeing 7 heads or more? (we use $\Pr(X > 6) = 1 - \Pr(X \leq 6)$)

```
1-binom.cdf(6,n=10,p=0.5)
0.171875
```

Example: Binomial Distribution

- The quantile is defined as the smallest value x such that $F(x) \geq p$, where F is the distribution function.

```
binom.ppf(0.25, n=10,p=0.5)
4.0
```

```
#Expected value is np
# variance is(np(1 − p))
binom.stats(n=10,p=0.5)
(array(5.), array(2.5))
```

Example: Binomial Distribution

- Blood groups: B, O, A, AB

- Probability of an individual having blood group B = 0.08
- Probability of an individual not having blood group B, being one of O, A, AB = 1- 0.08 = 0.92

- Two random, unrelated individuals
  - What is the probability neither have blood group B?
  - What is the probability one has blood group B?
  - What is the probability both have blood group B?

D. Altman, Practical Statistics for Medical Research

Binomial Distribution

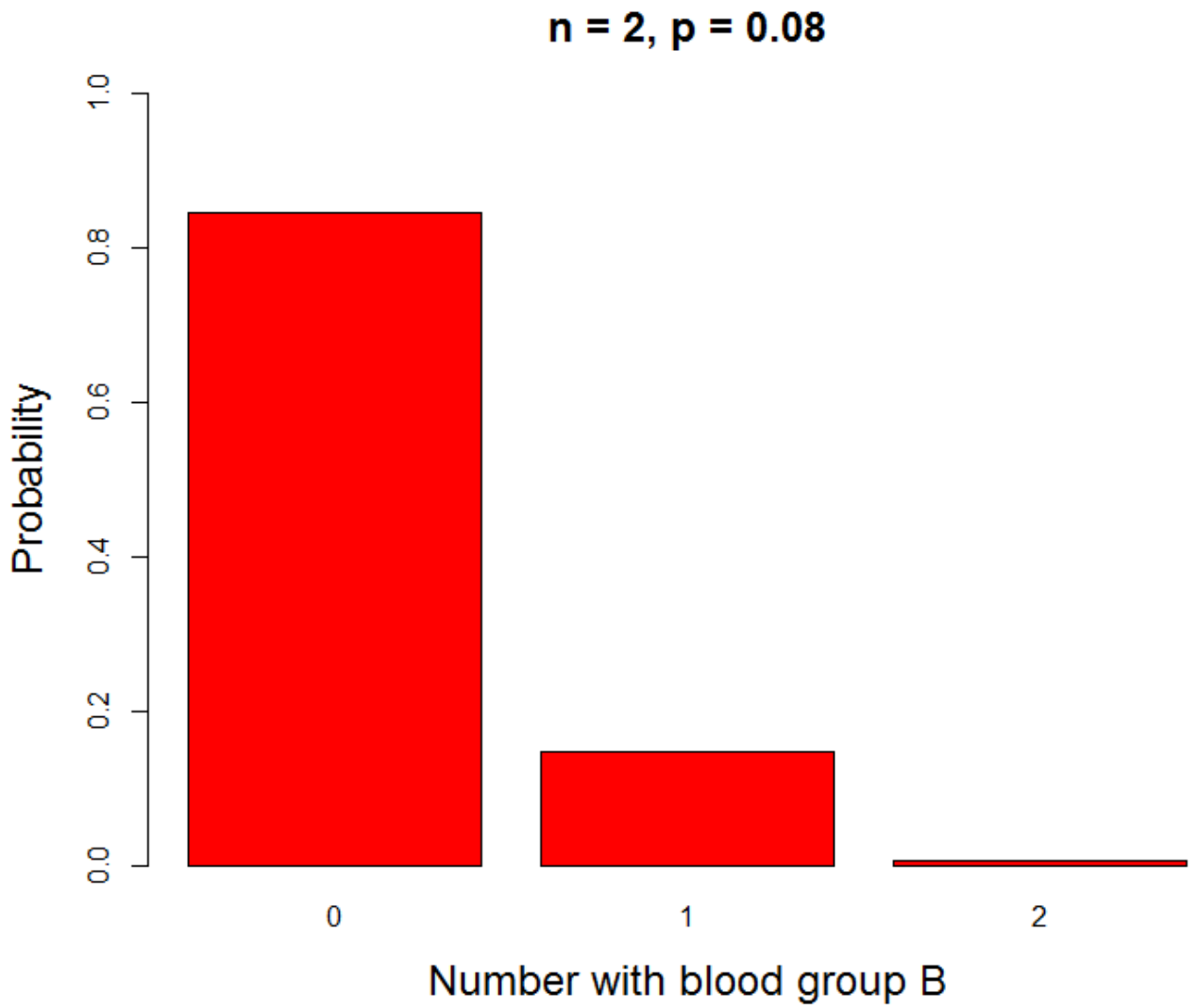
- Are the assumptions of the binomial distribution satisfied?

- Only two possible outcomes:
  - Blood group B
  - Not blood group B (O, A, AB)

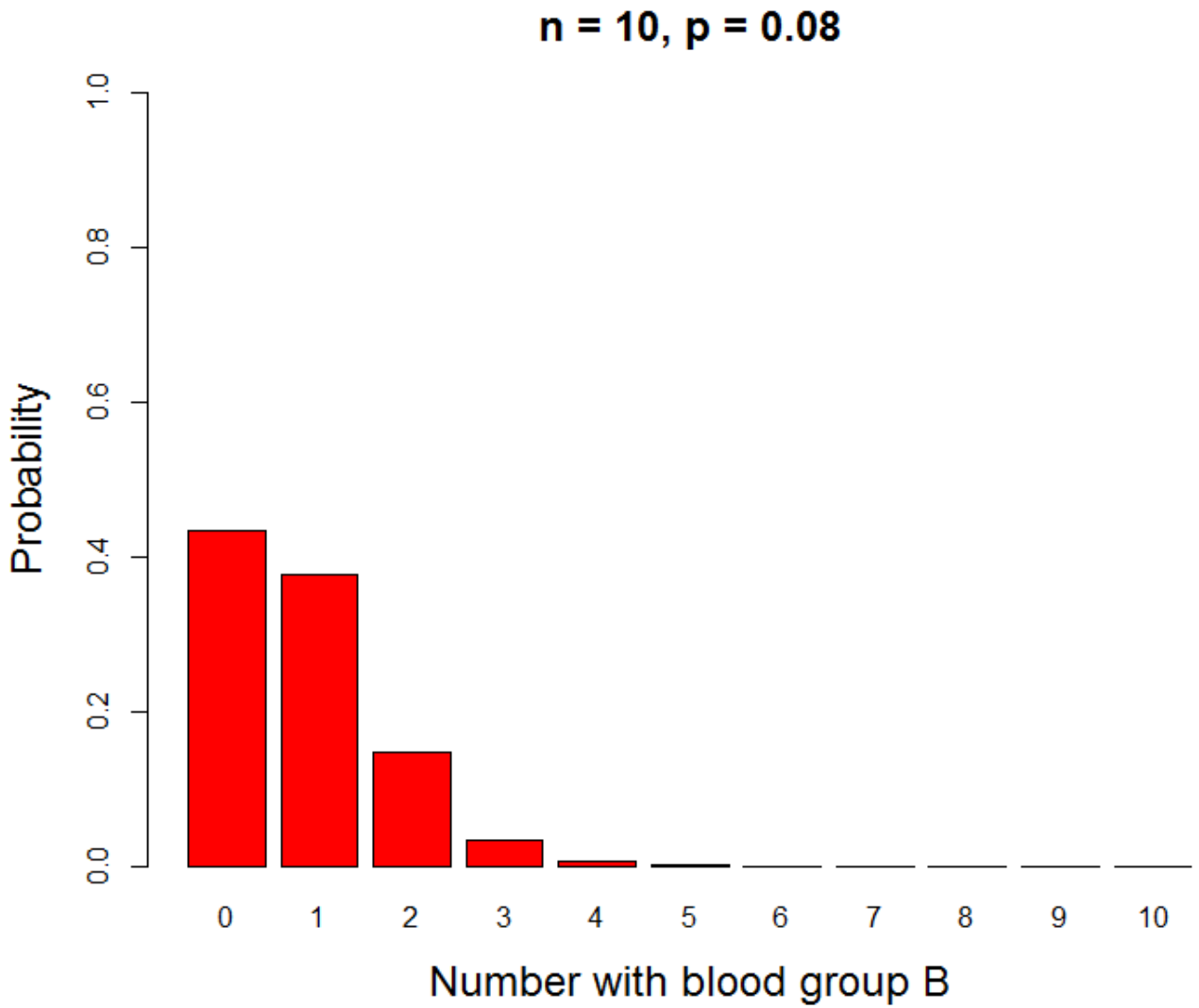
- The individuals are unrelated – independence

- The probability of each person having blood group B does not change from person to person($p = 0.08$)

Binomial Distribution



Binomial Distribution

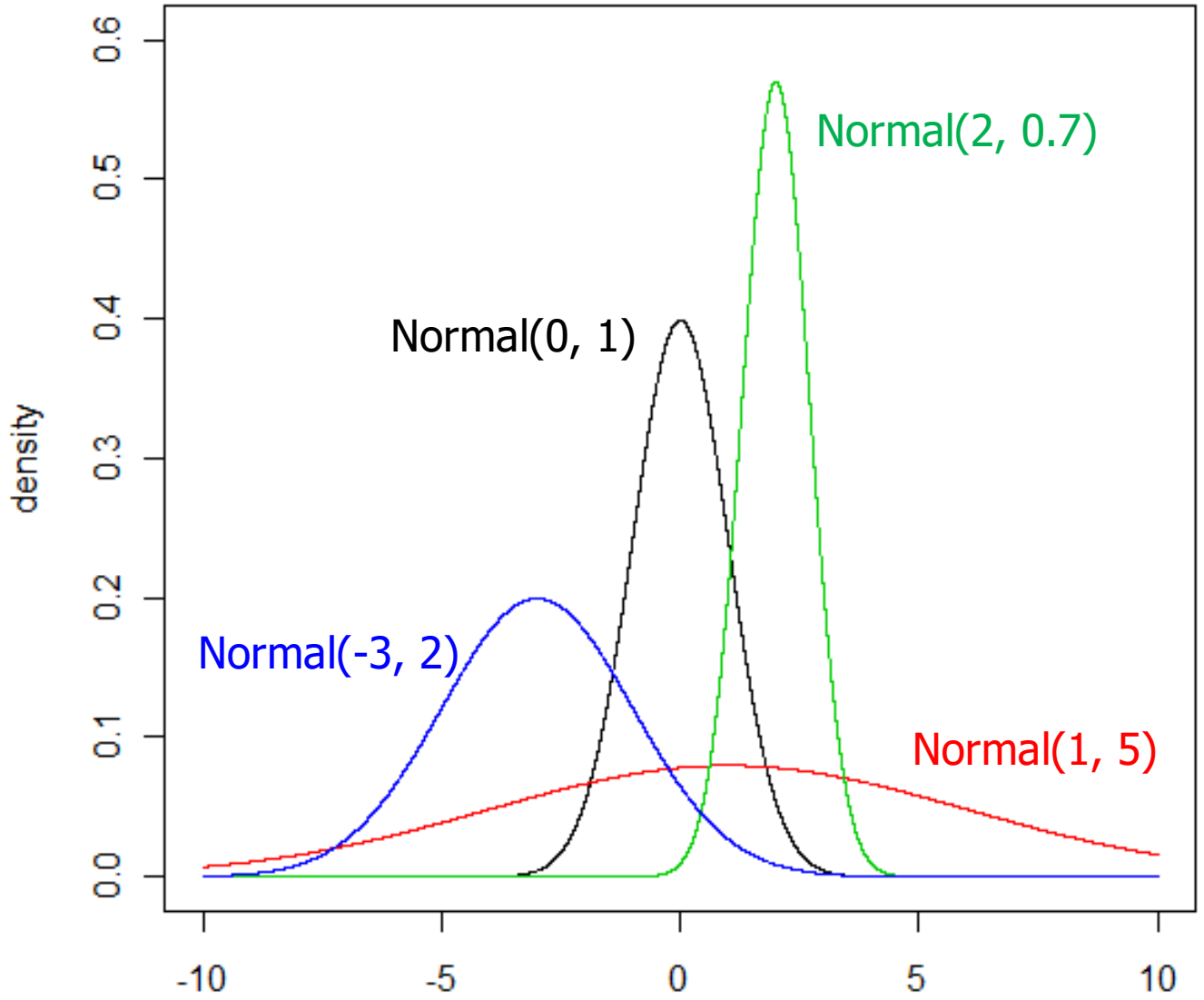


Normal distribution

- The normal distribution (Gaussian or bell-shaped distribution) is a continuous distribution that describes many populations in nature.
- It is also used to describe the sampling distribution of certain statistics.
- The density function is:

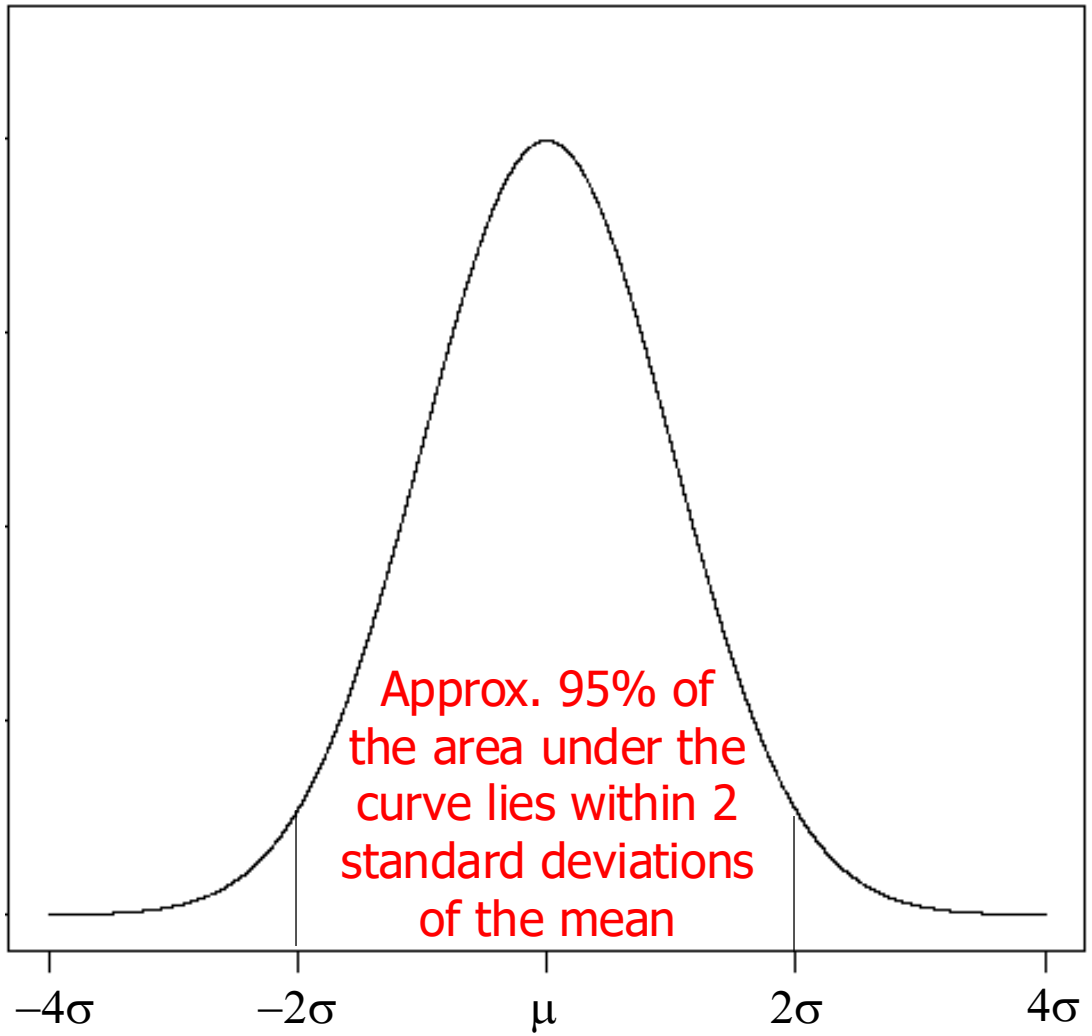
$$f(x|\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}\,e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- The two parameters are the mean μ and the standard deviation σ .



- The mean μ is the point of symmetry.
- The standard deviation σ controls the spread of the curve.

Normal Distribution



Standard Normal Distribution

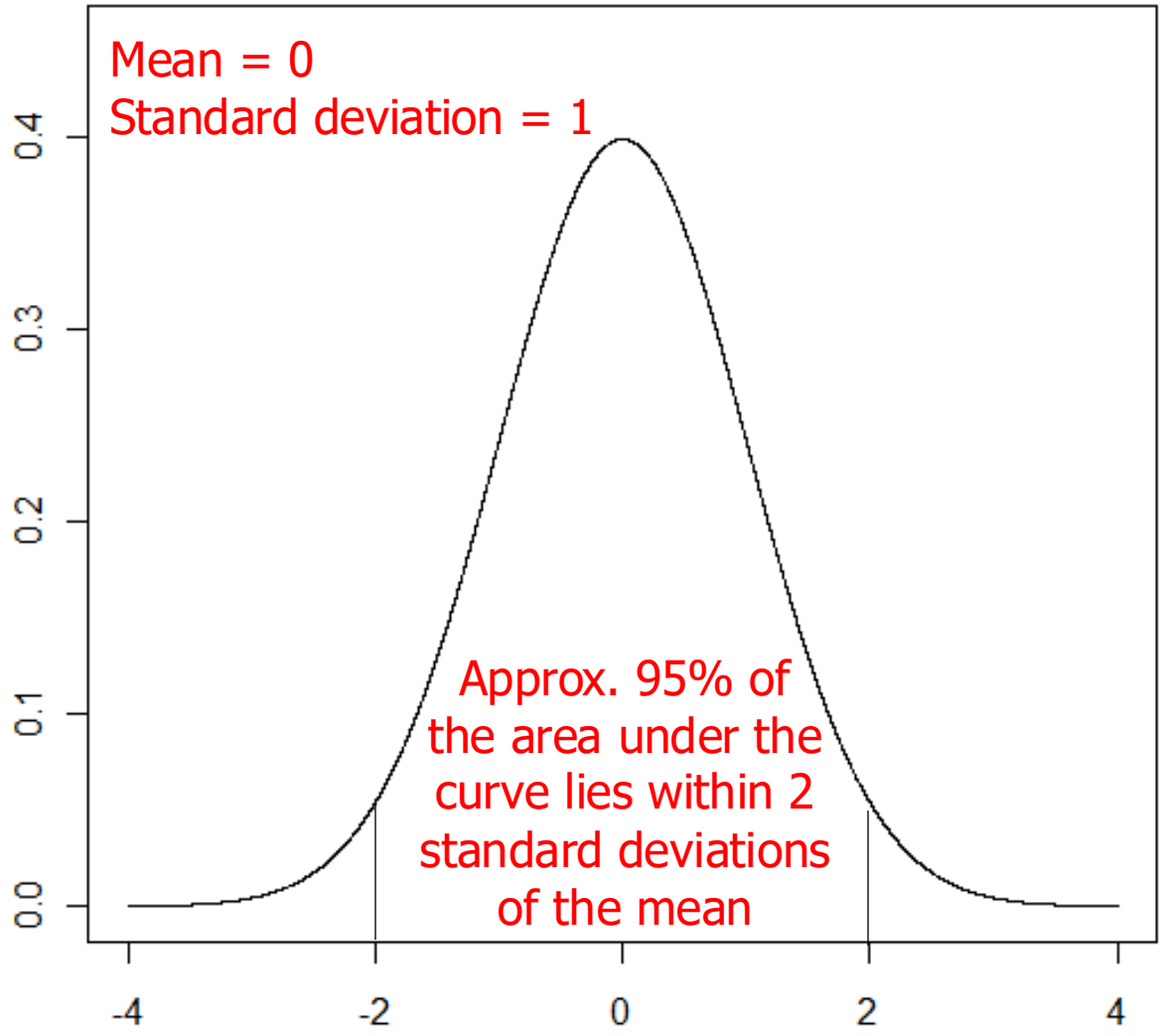
- The area under a part of the curve gives a particular probability

- To find out the area/probability we use the
  Standard Normal Distribution
  (mean = 0, standard deviation = 1)
  and look up the area in tables or use a computer

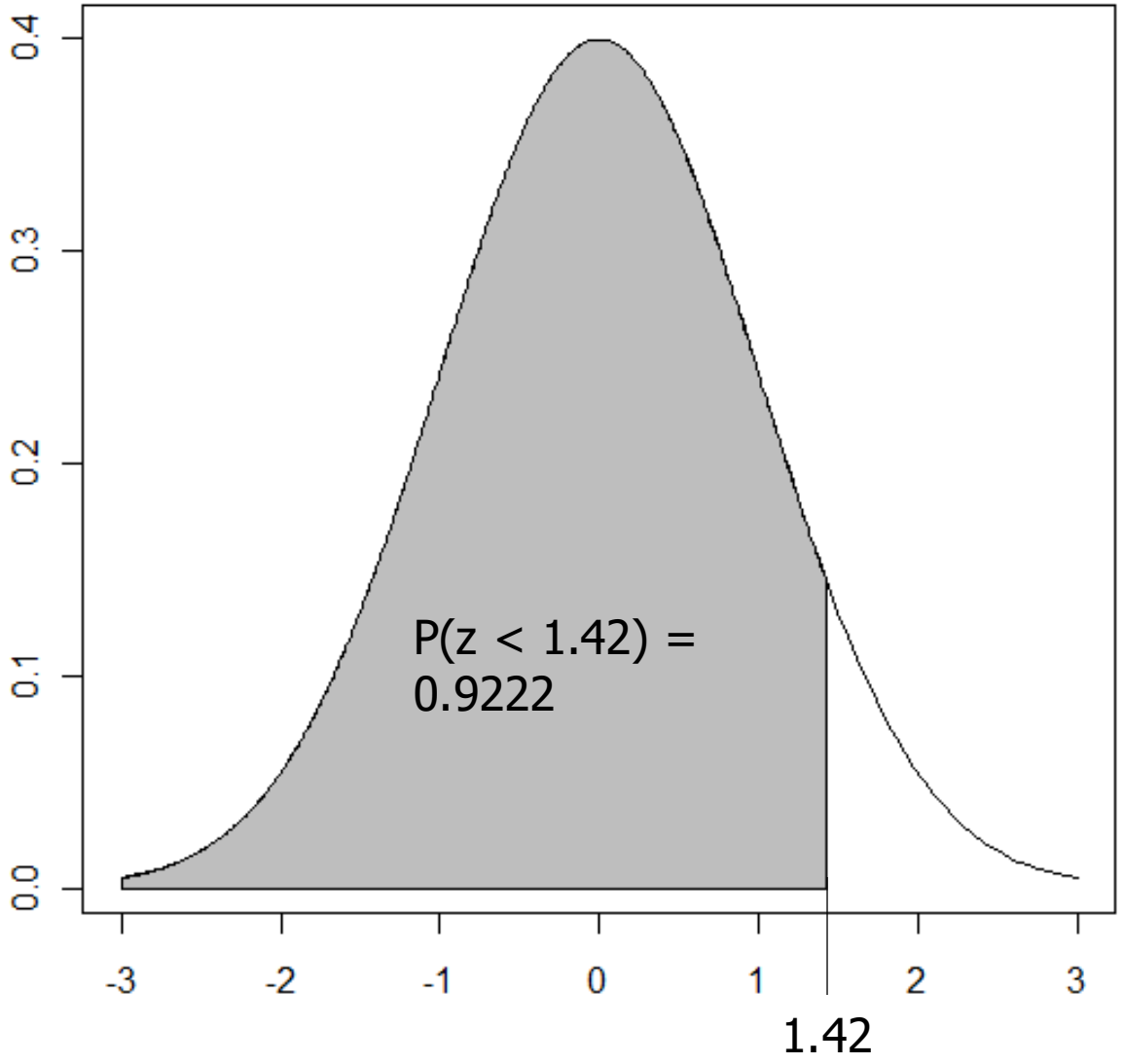
- z has a standard Normal distribution*:*

$$z \sim N(0, 1)$$

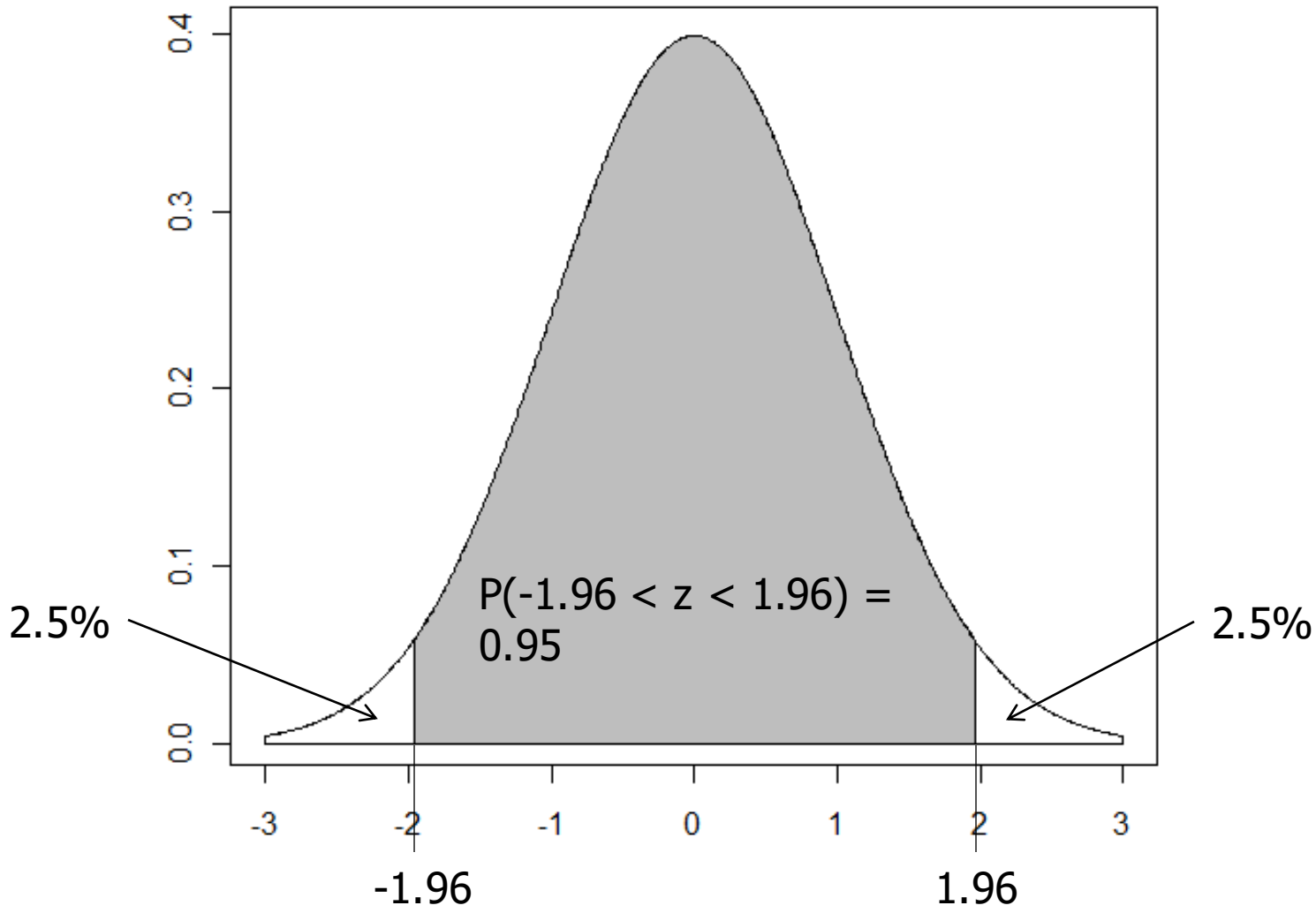
Standard Normal



Area Under Standard Normal



Area Under Standard Normal



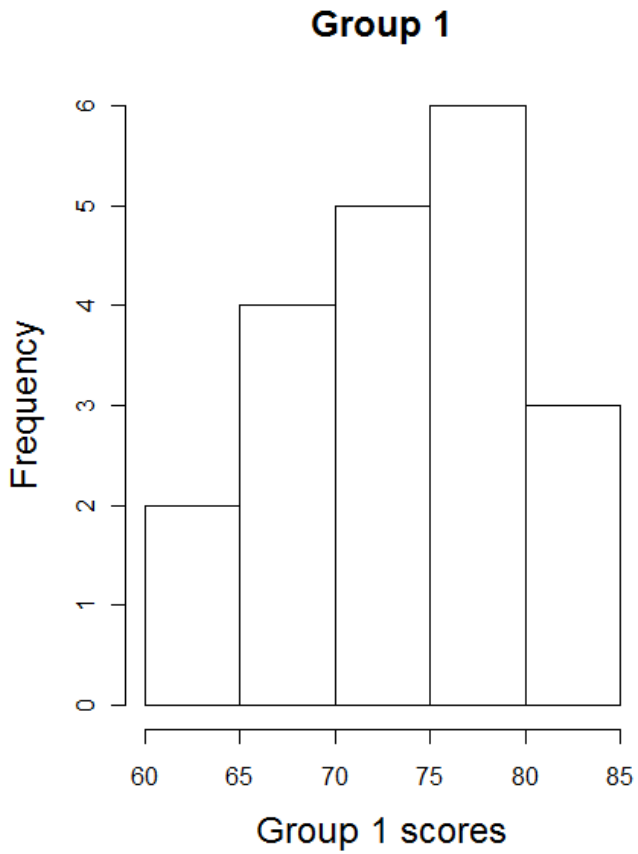
Normal Distribution Example

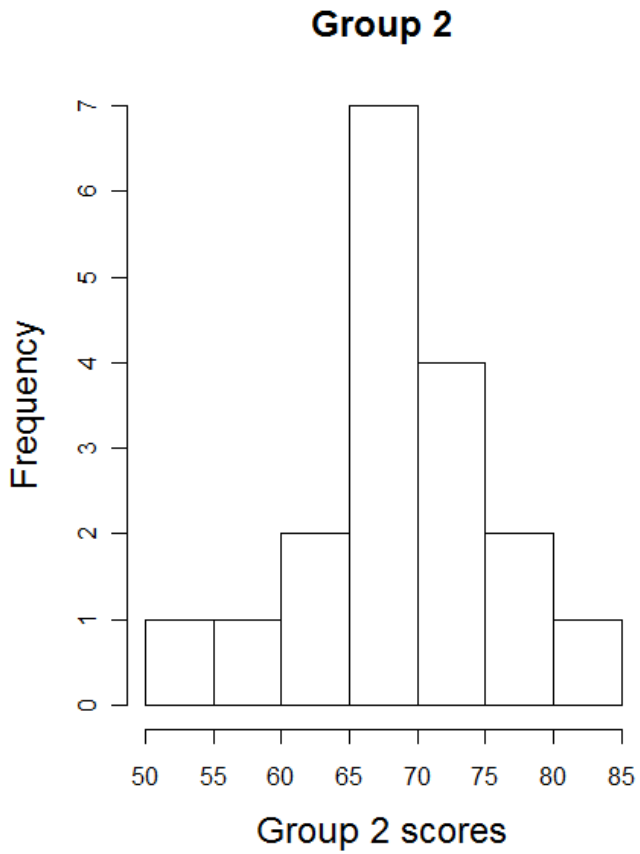
- Here are some data from 2 class group test scores

Sample ID	Group	Score
01	1	71.2
02	2	68.0
03	1	73.6
04	2	75.6
05	1	62.3
06	2	74.5
04	1	75.4
05	2	65.9
06	1	74.9
.	.	.
.	.	.
.	.	.

Normal Distribution Example

- And the distribution of these scores for each group





The Standard Normal distribution

- Any Normal distribution can be converted to a standard Normal by subtracting the mean and dividing by the standard deviation

$$X \sim N(\mu, \sigma)$$

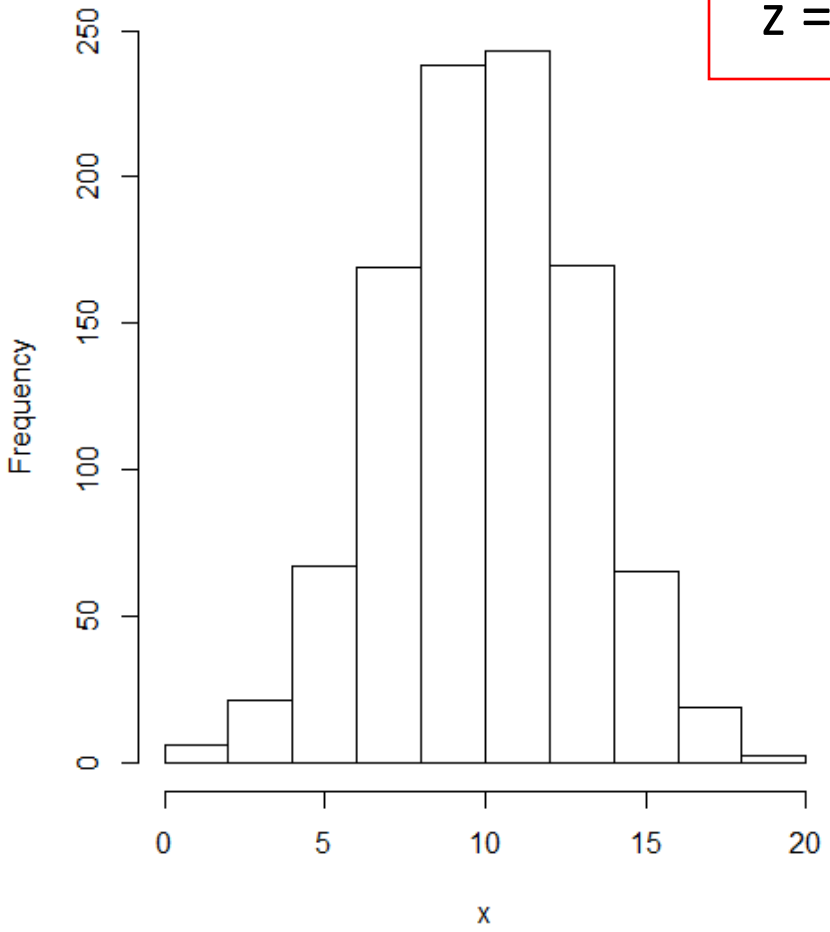
$$z = \frac{X - \mu}{\sigma}$$

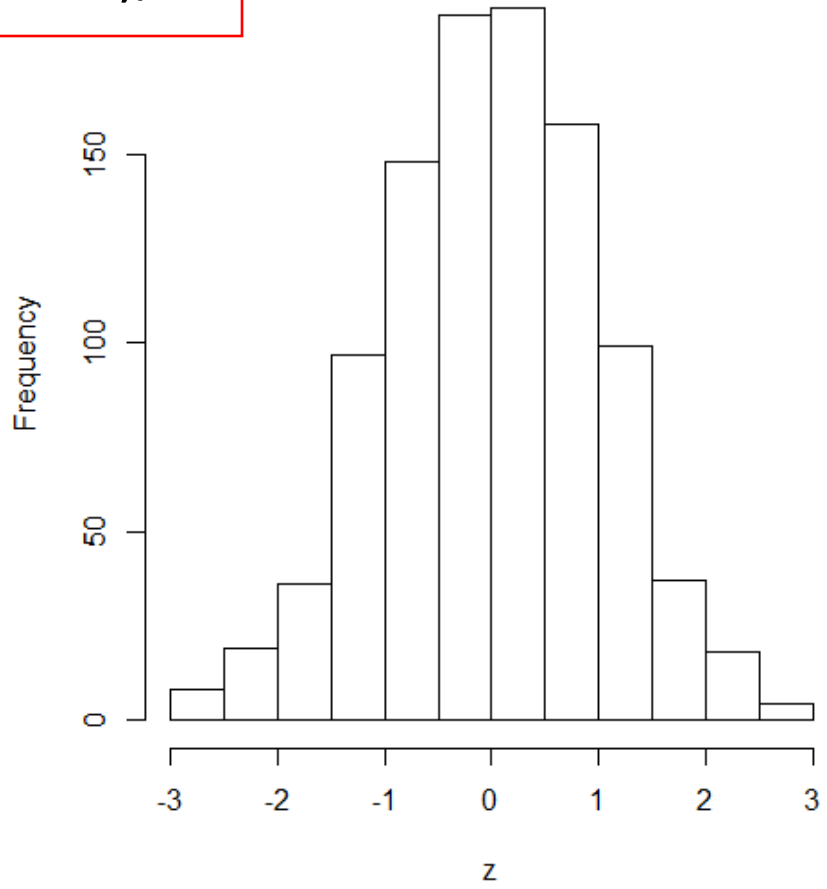
The Standard Normal distribution

X ~ Normal(10, 3)

Z ~ Normal(0, 1)

$$z = (x - 10)/3$$





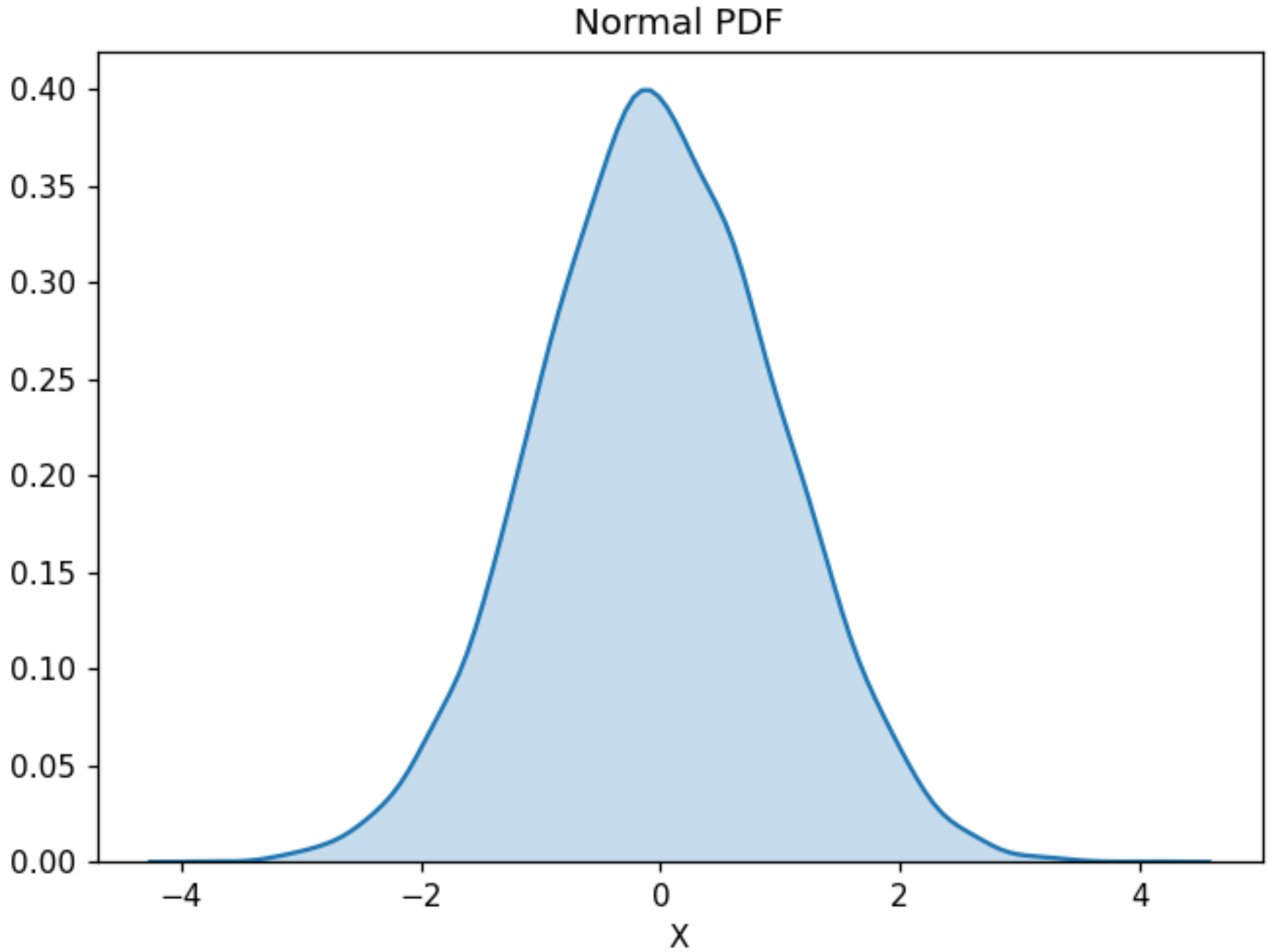
	PURPOSE	SYNTAX	EXAMPLE
rvs	Generates random number variates	norm.rvs(loc=0, scale=1, size=1, random_state=None)	norm.rvs(size=1000, loc=3, scale=0.25) Generates 1000 numbers from a normal with mean 3 and sd=.25
pdf	Probability Density Function (PDF)	norm.pdf(x,loc=0, scale=1)	norm.pdf(0, loc=0, scale=.5) Gives the density (height of the PDF) of the normal with mean=0 and sd=.5.
cdf	Cumulative Distribution Function (CDF)	norm.cdf(x, loc=0, scale=1)	norm.cdf(1.96, loc=0, scale=1) Gives the area under the standard normal curve to the left of 1.96, i.e. ~0.975
ppf	Quantile Function – inverse of CDF	norm.ppf(q, loc=0, scale=1)	norm.ppf(0.975, loc=0, scale=1) Gives the value at which the CDF of the standard normal is .975, i.e. ~1.96

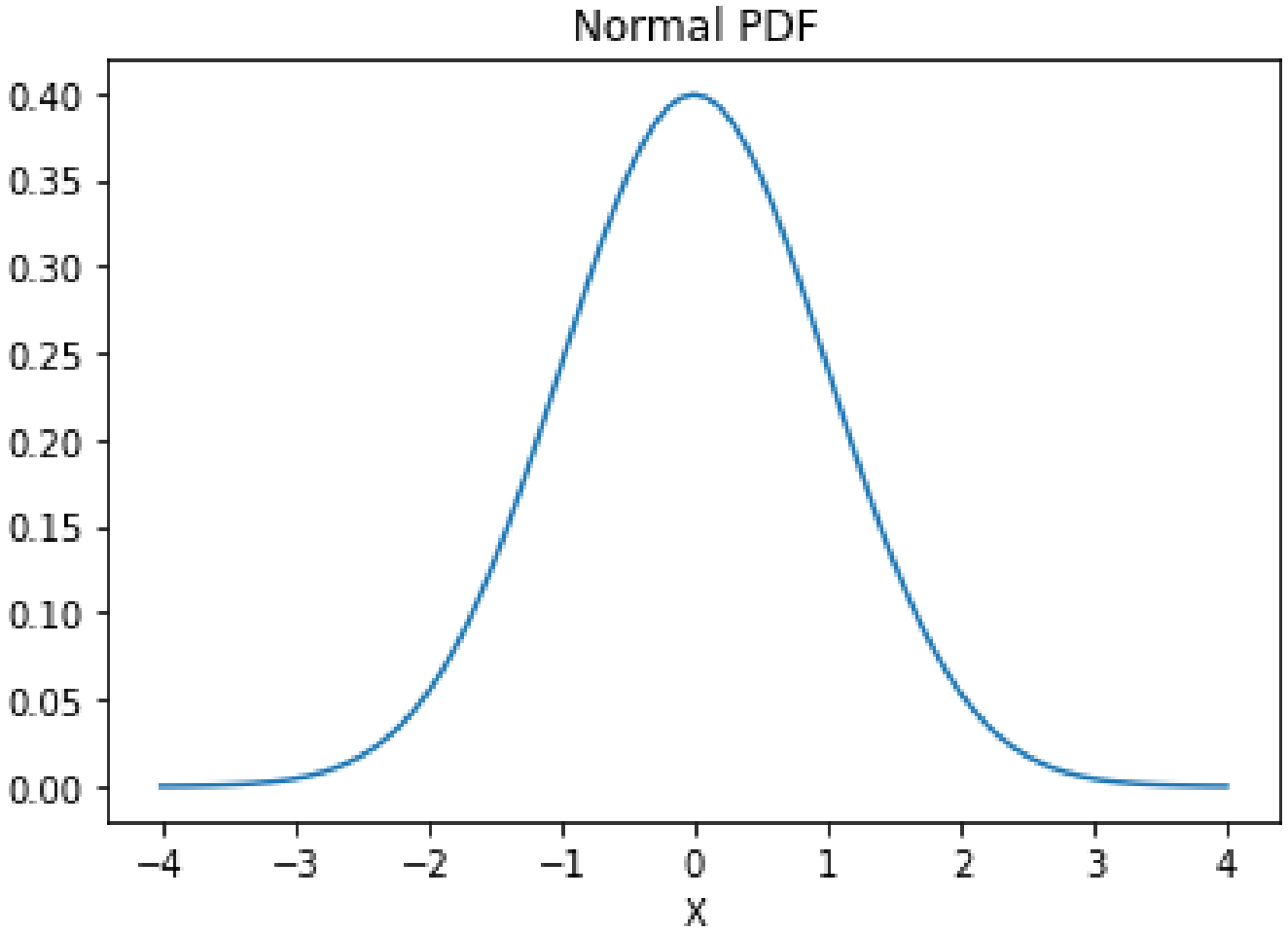
Examples with Normal Distribution

- Assume a normally distributed random variable *X*.
- What is the probability that *X* falls 1 standard deviation (or 2 or 3 standard deviations, respectively) from the mean?

```
from scipy.stats import norm

X=norm.rvs(size=10000)
#plot density curve
sns.kdeplot(X, fill=True)
plt.xlabel('X')
plt.title('Normal PDF')
plt.show()
```

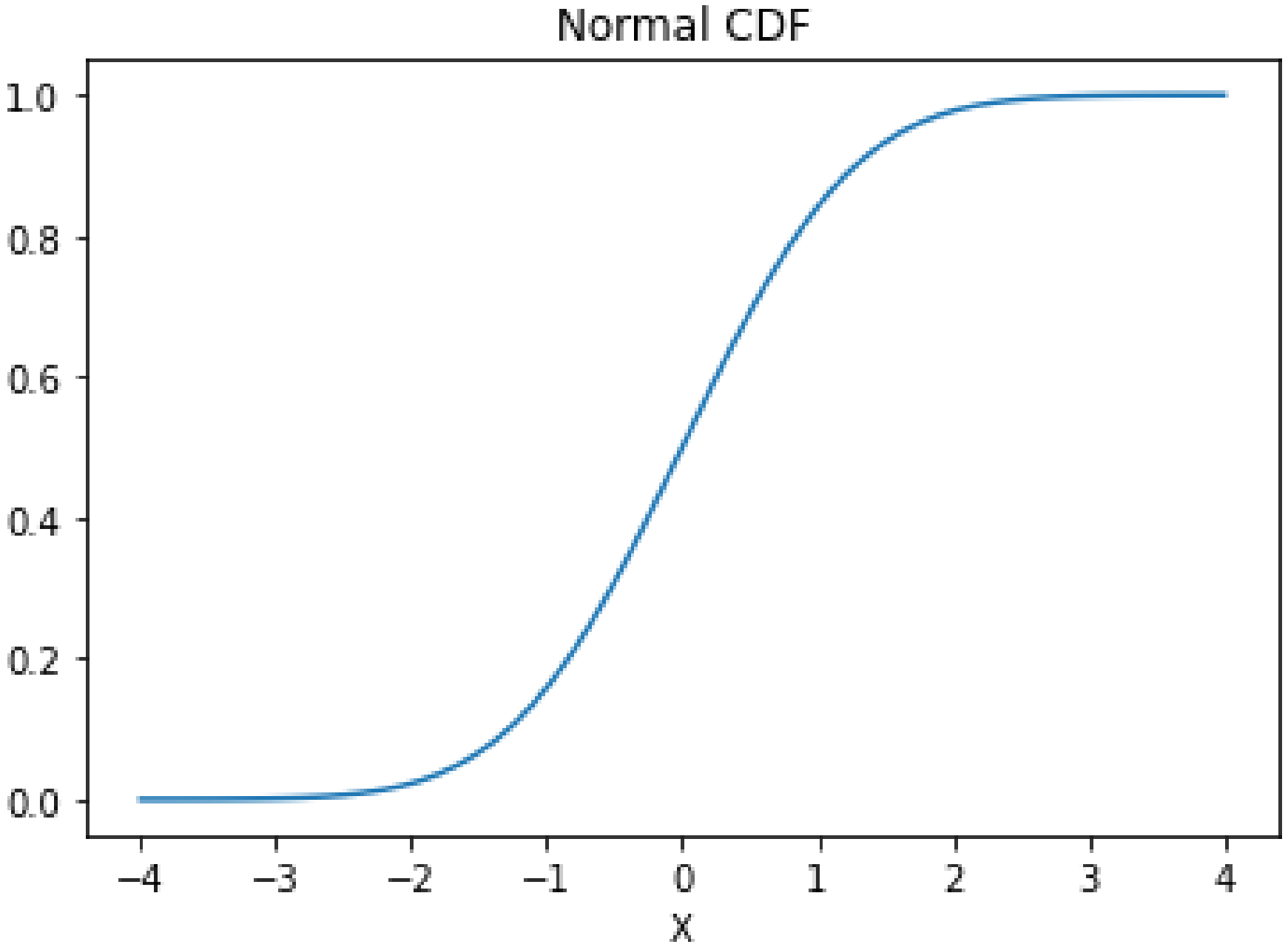

Normal PDF
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−2
0
2
4
X




```
#Exact PDF
X1=np.arange(-4,4.1,0.1)
norm.pdf(X1)
sns.lineplot( x=X1, y=norm.pdf(X1))
plt.xlabel('X')
plt.title('Normal PDF')
plt.show()
```

Examples with Normal Distribution

```
X1=np.arange(-4,4.1,0.1)
sns.lineplot( x=X1, y=norm.cdf(X1))
plt.xlabel('X')
plt.title('Normal CDF')
plt.show()
```


Normal CDF
1.0
0.8
0.6
0.4
0.2
0.0
−4
−3
−2
−1
0
1
2
3
4
X

Normal distribution Rule of thumb

- What is the probability that *X* falls 1 standard deviation from the mean?

```
norm.cdf(1)-norm.cdf(-1)
 0.6826894921370859
```

- What is the probability that *X* falls 2 standard deviation from the mean?

```
norm.cdf(2)-norm.cdf(-2)
 0.9544997361036416
```

- What is the probability that *X* falls 3 standard deviation from the mean?

```
norm.cdf(3)-norm.cdf(-3)
0.9973002039367398
```

Normal distribution Rule of thumb

- We see there's a 68% chance X is within one SD from its mean.
- We see there's a 95% chance X is within two SD from its mean.
- We see there's a 99.7% chance X is within three SD from its mean.

Normal distribution (z score) example

Scores on an aptitude test for programmers are Normally distributed with a mean of 78 and standard deviation of 4. The test is used for screening applicants for job interviews.

1. What percentage of applicants will score more than 83 on the test?

2. What percentage of applicants will score between 69 and 81 on the test?

Normal distribution (z score) example

1. What percentage of applicants will score more than 83 on the test?

$$X = 83, \mu = 78, \sigma = 4$$

$$z = \frac{X - \mu}{\sigma}$$

$$P(X > 83) = P\left(z > \frac{83 - 78}{4}\right)$$

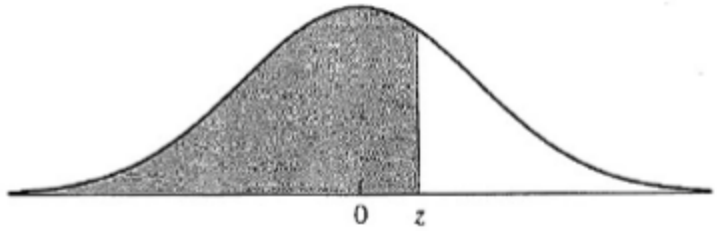
Normal distribution (z score) example

1. What percentage of applicants will score more than 83 on the test?

$$P\left(z > \frac{83-78}{4}\right) = P(z > 1.25) = 1 - P(z < 1.25)$$

How did we find $P(z < 1.25) = 0.8944$?

TABLE A.2 Cumulative normal distribution (continued)



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857

Normal distribution (z score) example

1. What percentage of applicants will score more than 83 on the test?

$$P\left(z > \frac{83-78}{4}\right) = P(z > 1.25) = 1 - P(z < 1.25)$$

$$= 1 - 0.8944 = 0.1056$$

Part 1 using python

```
1 - norm.cdf(83, 78, 4)
0.1056498
```

Normal distribution (z score) example

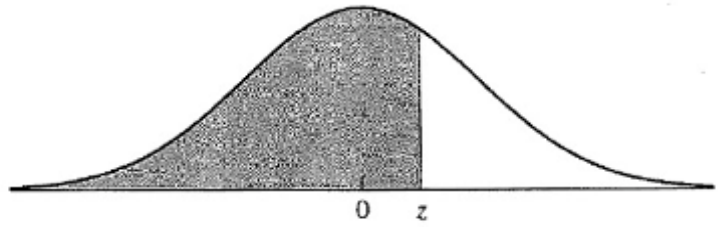
2. What percentage of applicants will score between 69 and 81 on the test?

$$P(69 < X < 81) = P\left(\frac{69-78}{4} < z < \frac{81-78}{4}\right)$$

$$= P(-2.25 < z < 0.75)$$

$$= P(z < 0.75) - P(z < -2.25)$$

TABLE A.2 Cumulative normal distribution (continued)



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952

Normal distribution (z score) example

2. What percentage of applicants will score between 69 and 81 on the test?

$$P(69 < X < 81) = P\left(\frac{69-78}{4} < z < \frac{81-78}{4}\right)$$

$$= P(-2.25 < z < 0.75)$$

$$= P(z < 0.75) - P(z < -2.25)$$

$$= 0.7734 - (1 - 0.9878) = 0.7612$$

Part 2 using python

```
norm.cdf(81, 78, 4)-norm.cdf(69, 78, 4)
0.7611482
```

Normal distribution example

- The heights of adult males within an ethnic class follows a normal distribution.
- Suppose the mean is 1.78 metres and a standard deviation of 0.07 m.

1. What percentage of males are taller than 2 metres?
2. What does it predict for the tallest male on the planet?

Normal distribution example

```
mu=1.78
sigma=0.07
1-norm.cdf(2,loc=mu, scale=sigma)
0.00083653736107613 95
```

1. That is fewer that 1% are 2 meters tall.

Normal distribution

```
#Assuming there are 2.5 billion males
#the tallest is in the top
#1/(2.5 billion) quantile

p=1-1/2500000000
norm.ppf(p,loc=mu, scale=sigma)
2.210144930565437
```

2. For extreme values the model does not fit.

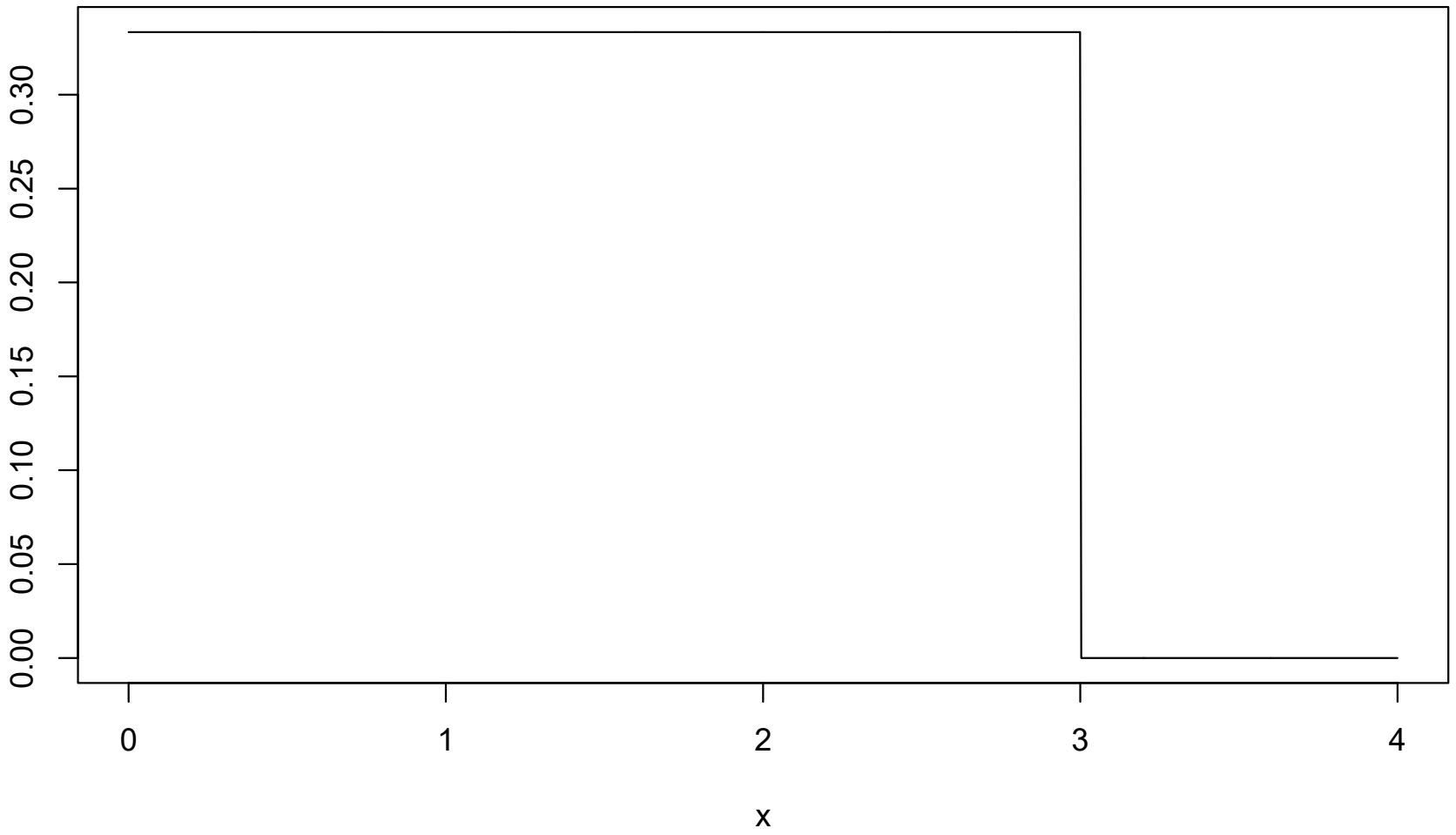
Other distributions

- Some populations are not well described by a normal distribution, e.g. multimodal, non symmetric or fat tails distributions.
- Other families of distribution have been defined, including:
- The uniform distribution (every value is equally likely). Name in scipy.stats: uniform.
- The exponential distribution. Name in scipy.stats: expon.
- The Poisson distribution (# events occurring in a fixed time period). Name in scipy.stats: poisson.
- The lognormal distribution (i.e. its logarithm is normal) heavily skewed, e.g. income distribution. Name in scipy.stats: lognorm.

Uniform distribution

- Useful for populations with no preferred value over their range.
- The density is constant over [a, b].
- a is the min and b is the max

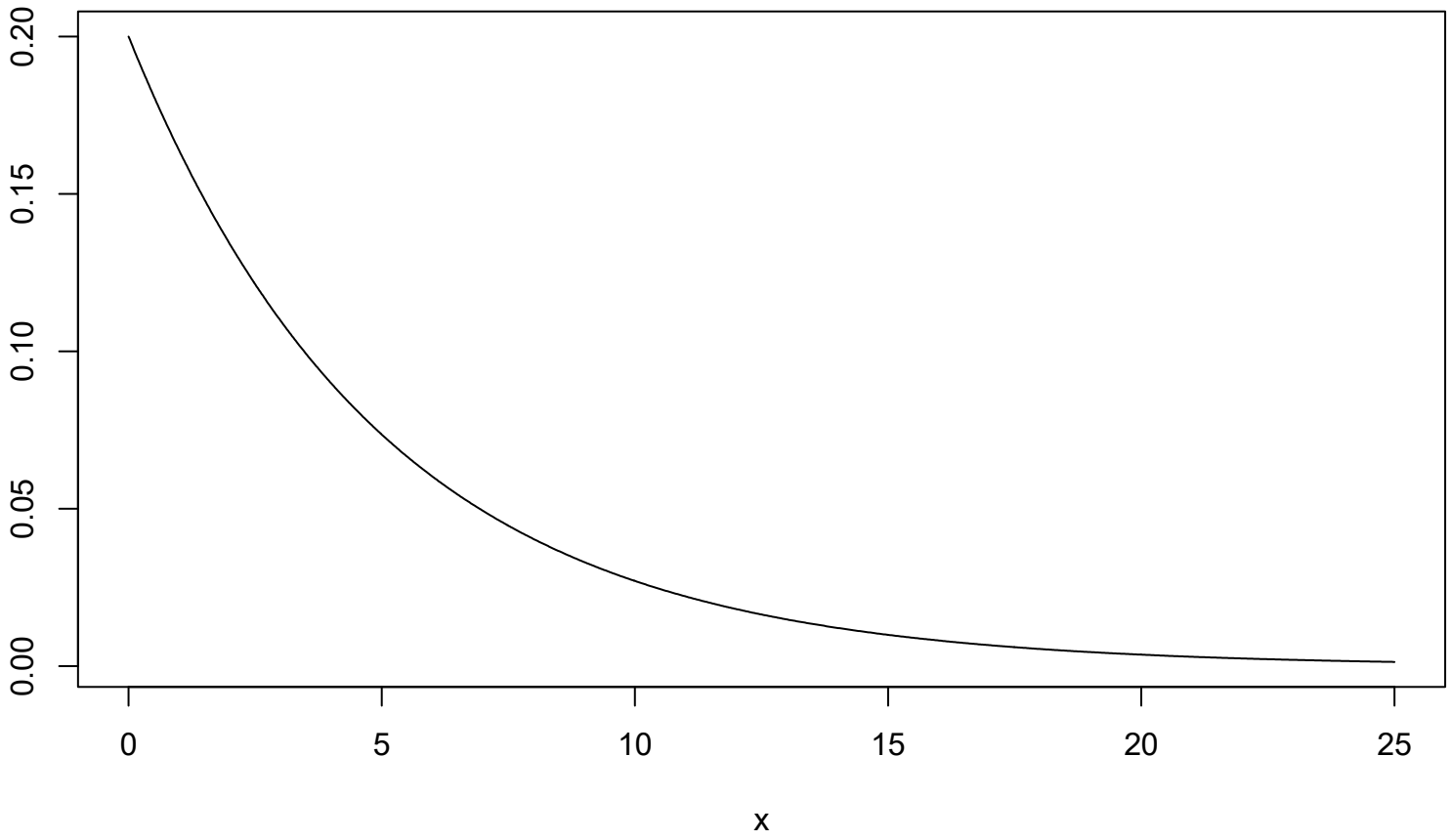
Density of Uniform(0,3)



Exponential distribution

- It is a skewed distribution.
- E.g. include lifetime of light bulbs.
- Density is $f(x \mid a) = ae^{-ax}$.
- Mean is $1/a$, SD is $1/a$.

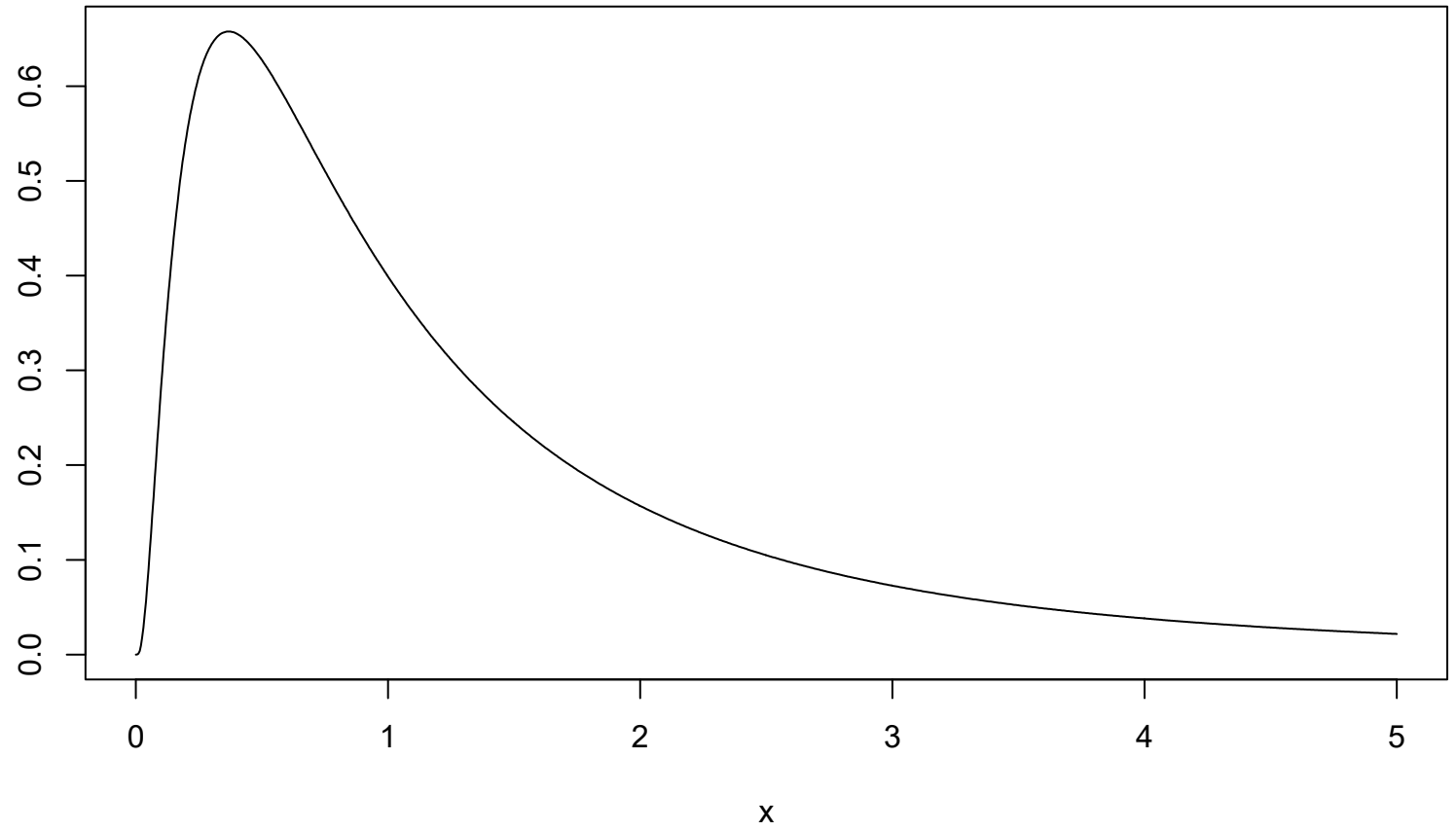
Density Expon(1/5)



Lognormal distribution

- It is a heavily skewed distribution.
- E.g. include stock market returns, income distributions etc.
- If X is lognormally distributed then $\log(X)$ is normally distributed.

Density lognorm(0,1)



Poisson distribution

- The Poisson distribution is a discrete probability distribution.
- It is the distribution for events occurring in a fixed interval of time for events occurring with a known average rate (λ).

$$\mathrm{P}(X = k) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

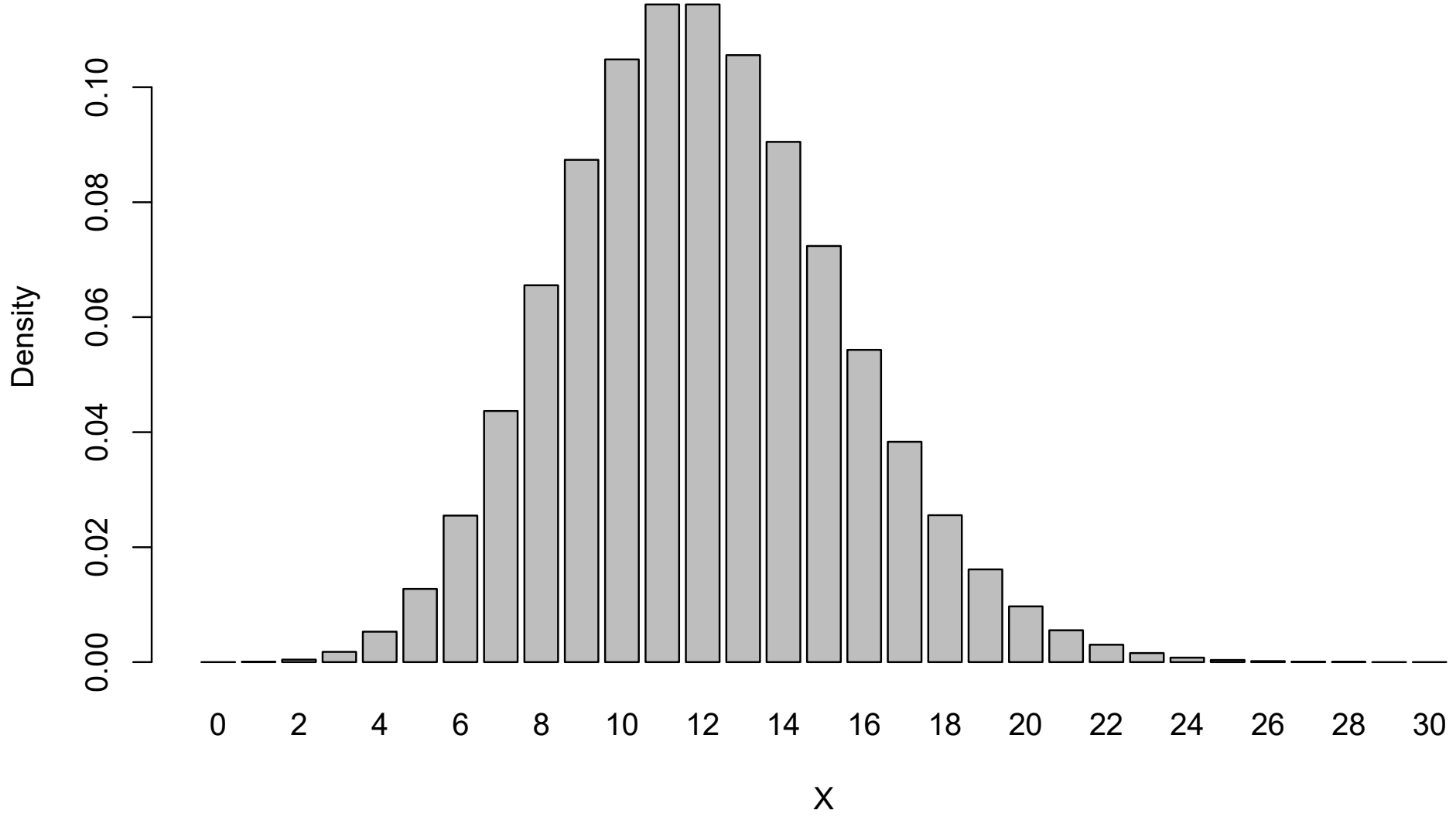
E.g telephone calls arriving in a system.

- E.g. buses arriving at a bus stop.

Example Poisson Distribution

- If there are twelve cars crossing a bridge per minute on average, find the probability of having seventeen or more cars crossing the bridge in a particular minute.

PMF


Density
0.00
0.02
0.04
0.06
0.08
0.10
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
X

Example Poisson Distribution

- The probability of having *sixteen or less* cars crossing the bridge in a particular minute is

```
poisson.cdf(16,mu=12)
0.89871
```

- Hence the probability of having seventeen or more cars crossing the bridge in a minute is

```
1- poisson.cdf(16,mu=12)
0.10129
```

Poisson distribution example

Suppose a space enthusiast says that at night we expect to see one meteor every 12 minutes.

1. What distribution can we use to model the number of meteor sightings? Find the parameter of this distribution.
2. What is the probability of seeing exactly 3 meteors in one hour?
3. What is the probability of seeing more than 3 meteors in one hour?

Poisson distribution example

1. What distribution can we use to model the number of meteor sightings? Find the parameter of this distribution.

This can be modelled by a Poisson distribution with rate $\lambda = 5$ per hour, since $\frac{1\ meteor}{12\ minutes} * 60\ minutes = 5\ meteors\ expected\ per\ hour$.

Poisson distribution is used for count data (number of events occurring in a fixed period of time).

Poisson distribution example

2. What is the probability of seeing exactly 3 meteors in one hour?

$$P(X = k) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

$$P(X = 3) = \frac{e^{-5}5^{3}}{3!} = 0.1404$$

Part 2 using python

```
poisson.pmf(3,mu=5)
0.1403739
```

Poisson distribution example

3. What is the probability of seeing more than 3 meteors in one hour?

$$P(X > 3) = 1 - P(X \leq 3)$$

$$P(X > 3) = 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)]$$

$$P(X > 3) = 1 - [\frac{e^{-5}5^0}{0!} + \frac{e^{-5}5^1}{1!} + \frac{e^{-5}5^2}{2!} + \frac{e^{-5}5^3}{3!}]$$

$$P(X > 3) = 1 - [0.0067 + 0.0337 + 0.0842 + 0.1404] = 0.735$$

Part 3 using python

```
1 - poisson.cdf(3,mu=5)
0.7349741
```

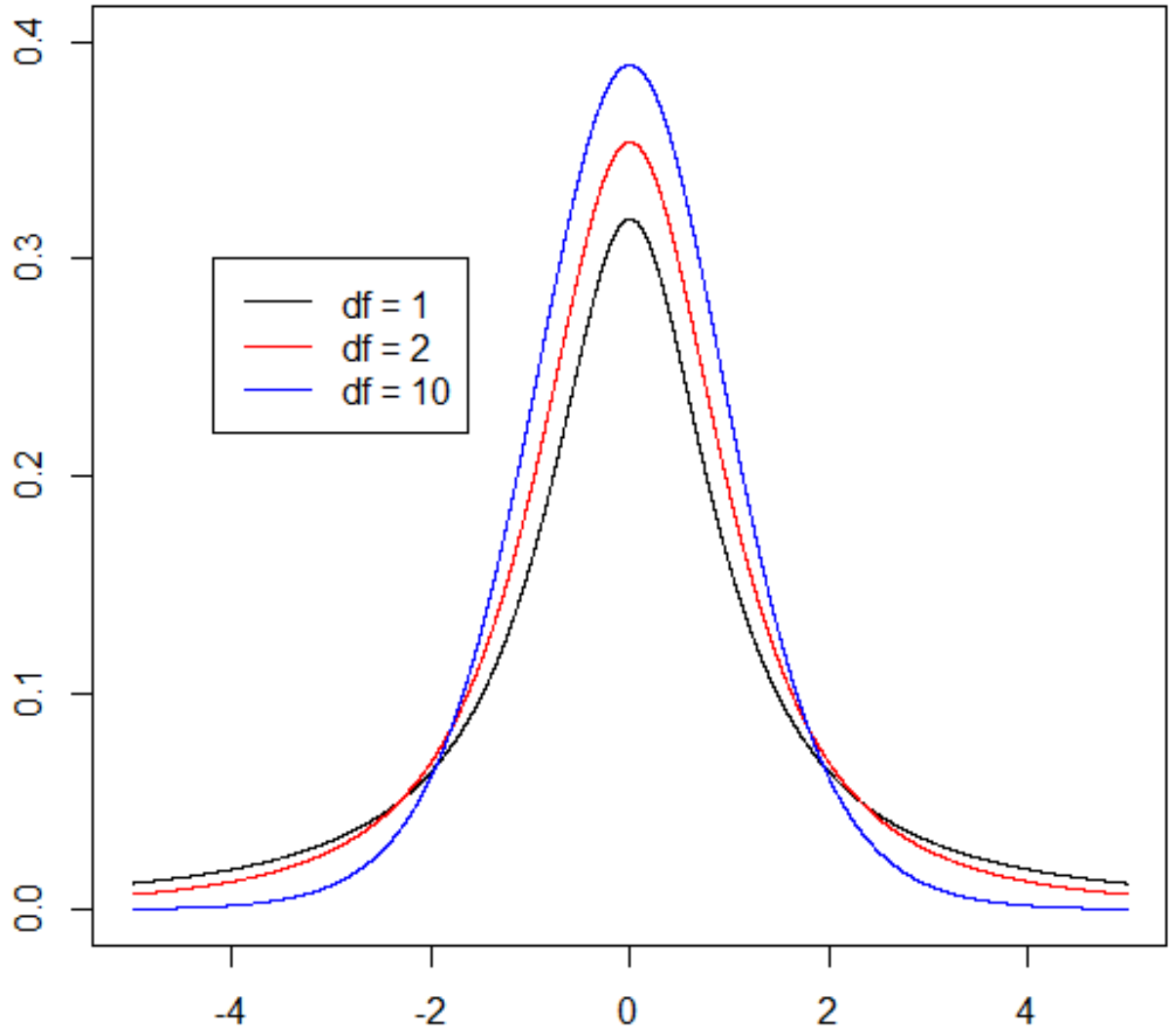
Sampling distribution

- Sampling distributions are often of the 3 types: t-distribution, F-distribution, chi-squared distribution.
- The parameters are called "degrees of freedom". They are related to the sampling size.

Student’s t-distribution

- The Student’s t-distribution is another symmetric continuous probability distribution
- This distribution is very similar to the Normal Distribution but has heavier tails
- Appear in many statistical tests when the sample size is relatively small
- Has one parameter: degrees of freedom(df)

Student’s t-distribution



Law of large numbers

- When the sample $X_1, X_2, \ldots, X_n$ is drawn from Normal(μ, σ), the sample mean $\bar{X}$ is Normal(μ, $\sigma/\sqrt{n}$).
- The centre stays the same, but as *n* grows, the spread decreases, i.e. with greater probability the value of $\bar{X}$ is close to the mean μ.
- The sample average concentrating on the mean is called the ***law of large number***.

Law of large numbers example

- Suppose adult males heights are Normal(1.78, 0.07).
- The probability that the average of a random sample of 25 males is between 1.77 and 1.79 is:

```
mu=1.78
sigma=0.07
n=25
nsigma=0.07/math.sqrt(n)
norm.cdf(1.79,loc=mu, scale=nsigma)-norm.cdf(1.77,loc=mu, scale=nsigma)
0.5249494759460474
```

- Compare this with the probability for a single person:

```
norm.cdf(1.79,loc=mu, scale=sigma)-norm.cdf(1.77,loc=mu, scale=sigma)
0.11359699363293646
```

Central limit theorem

- When the population is not normal, we use the central limit theorem (CLT).
- CLT says that for *any* population (i.e. including non-normally distributed) with mean μ and standard deviation σ, $\bar{X}$ is asymptotically normal, i.e. for *n* large enough, $\bar{X}$ is Normal(μ, σ/$\sqrt{n}$).
- Again: as *n* gets bigger, the sampling distribution of $\bar{X}$ becomes more and more normal.
- Because real-world data is often the balanced sum of many unobserved random events, CLT partially explains the prevalence of the normal probability distribution in nature.