ACADGILD

Mastering Data Science

Statistics

Session 16 - Distributions and CLT

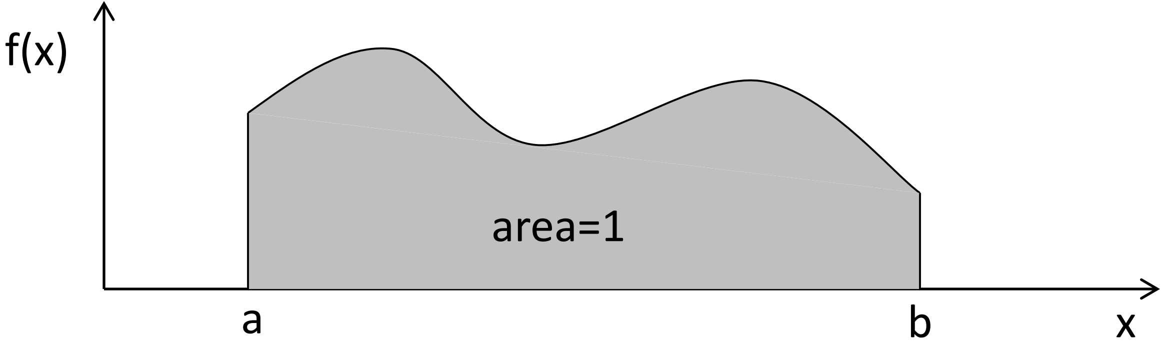
Agenda

1. Probability Function
2. Binomial Distribution
3. Binomial Random Variable
4. Poisson Distribution
5. Poisson Probability Distribution
6. The Normal Distribution
7. Standard Normal Distribution
8. Calculating Normal Probabilities
9. Using the Normal Table
10. Finding and Using the Values of Z
11. Central Limit Theorem
12. Sampling Distribution of the Sample Mean

Probability Density Function

Gives the likelihood of the value of a random variable falling within a range. A function f(x) is called a Probability Density Function over the range $a \le x \le b$ if it meets the following requirements:

1. $f(x) \ge 0$ for all x between a and b, and



2. The total area under the curve between a and b is 1.0

Copyrights© 2017, AcadGild. All rights reserved

Binomial Distribution

- The binomial distribution is the probability distribution that results from doing a “binomial experiment”. Binomial experiments have the following properties:

    1. Fixed number of trials, represented as n.
    2. Each trial has two possible outcomes, a “success” and a “failure”.
    3. P(success)=p (and thus: P(failure)=1–p), for all trials.
    4. The trials are independent, which means that the outcome of one trial does not affect the outcomes of any other trials.

Copyrights© 2017, AcadGild. All rights reserved

Binomial Random Variable

- The binomial random variable counts the number of successes in n trials of the binomial experiment. It can take on values from 0, 1, 2, …, n. Thus, its a discrete random variable.
- To calculate the probability associated with each value we use combinatory:

$$P(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

for x=0, 1, 2, …, n

Copyrights© 2017, AcadGild. All rights reserved

Binomial Distribution

- As you expect, Statisticians have developed general formulas for the mean, variance, and standard deviation of a binomial random variable. They are:

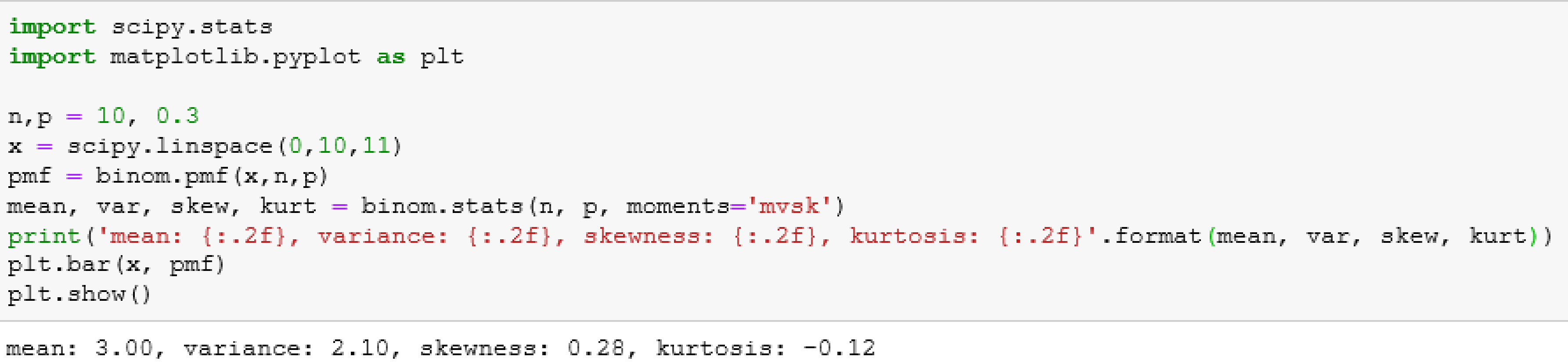
$$\mu = np$$

$$\sigma^2 = np(1-p)$$

$$\sigma = \sqrt{np(1-p)}$$

Copyrights© 2017, AcadGild. All rights reserved

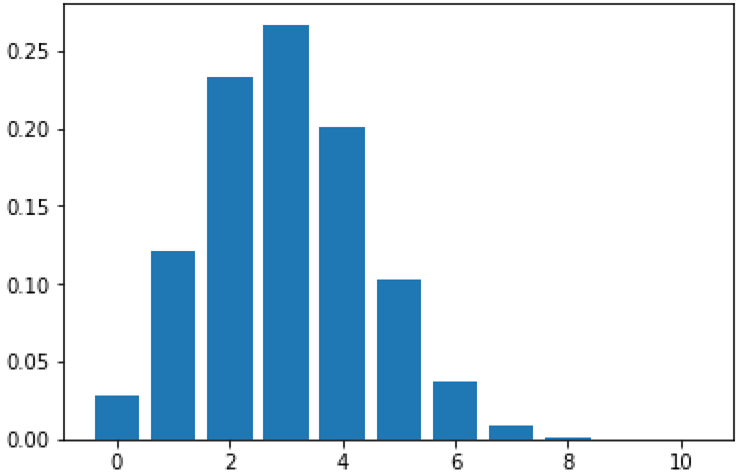
Binomial Distribution Using Python



```
import scipy.stats
import matplotlib.pyplot as plt

n,p = 10, 0.3
x = scipy.linspace(0,10,11)
pmf = binom.pmf(x,n,p)
mean, var, skew, kurt = binom.stats(n, p, moments='mvsk')
print('mean: {:.2f}, variance: {:.2f}, skewness: {:.2f}, kurtosis: {:.2f}'.format(mean, var, skew, kurt))
plt.bar(x, pmf)
plt.show()
```

```
mean: 3.00, variance: 2.10, skewness: 0.28, kurtosis: -0.12
```



[Practice Code]

Poisson Distribution

- Named after Simeon Poisson, the Poisson distribution is a discrete probability distribution and refers to the number of events (a.k.a. successes) within a specific time period or region of space.
- For example,
  - The number of cars arriving at a service station in 1 hour. (The interval of time is 1 hour)
  - The number of flaws in a bolt of cloth. (The specific region is a bolt of cloth)
  - The number of accidents in 1 day on a particular stretch of highway. (The interval is defined by both time, 1 day, and space and the particular stretch of highway.)

Copyrights© 2017, AcadGild. All rights reserved

The Poisson Experiment

Similar to binomial experiment, a Poisson experiment has four defining characteristic properties:

1. The number of successes that occur in any interval is independent of the number of successes that occur in any other interval.
2. The probability of a success in an interval is the same for all equal-size intervals
3. The probability of a success is proportional to the size of the interval.
4. The probability of more than one success in an interval approaches 0 as the interval becomes smaller.

Copyrights© 2017, AcadGild. All rights reserved

Poisson Distribution

- The Poisson random variable is the number of successes that occur in a period of time or an interval of space in a Poisson experiment.
- For example, on average, 96 trucks arrive at a border crossing every hour.
- For example, the number of typographic errors in a new textbook edition averages 1.5 per 100 pages.

🟦 Success
🟥 Time Period
🟩 Interval

Copyrights© 2017, AcadGild. All rights reserved

Poisson Probability Distribution

- The probability that a Poisson random variable assumes a value of x is given by:

$$P(x) = \frac{e^{-\mu}\mu^{x}}{x!} \quad for \quad x = 0,\ 1,\ 2,\ldots$$

where μ is the mean number of successes in the interval

e – natural logarithm base

$E(X) = V(X) = \mu$

Copyrights© 2017, AcadGild. All rights reserved

Poisson Probability Distribution

Example

- The number of typographical errors in new editions of textbooks varies considerably from book to book. After some analysis it concludes that the number of errors is Poisson distributed with a mean of 1.5 per 100 pages. The instructor randomly selects 100 pages of a new book. What is the probability that there are no typos?

That is, what is P(X=0) given that μ = 1.5?

$$P(0) = \frac{e^{-\mu}\mu^{x}}{x!} = \frac{e^{-1.5}1.5^{0}}{0!} = .2231$$

"There is about a 22% chance of finding zero errors"

Copyrights© 2017, AcadGild. All rights reserved

Poisson Distribution

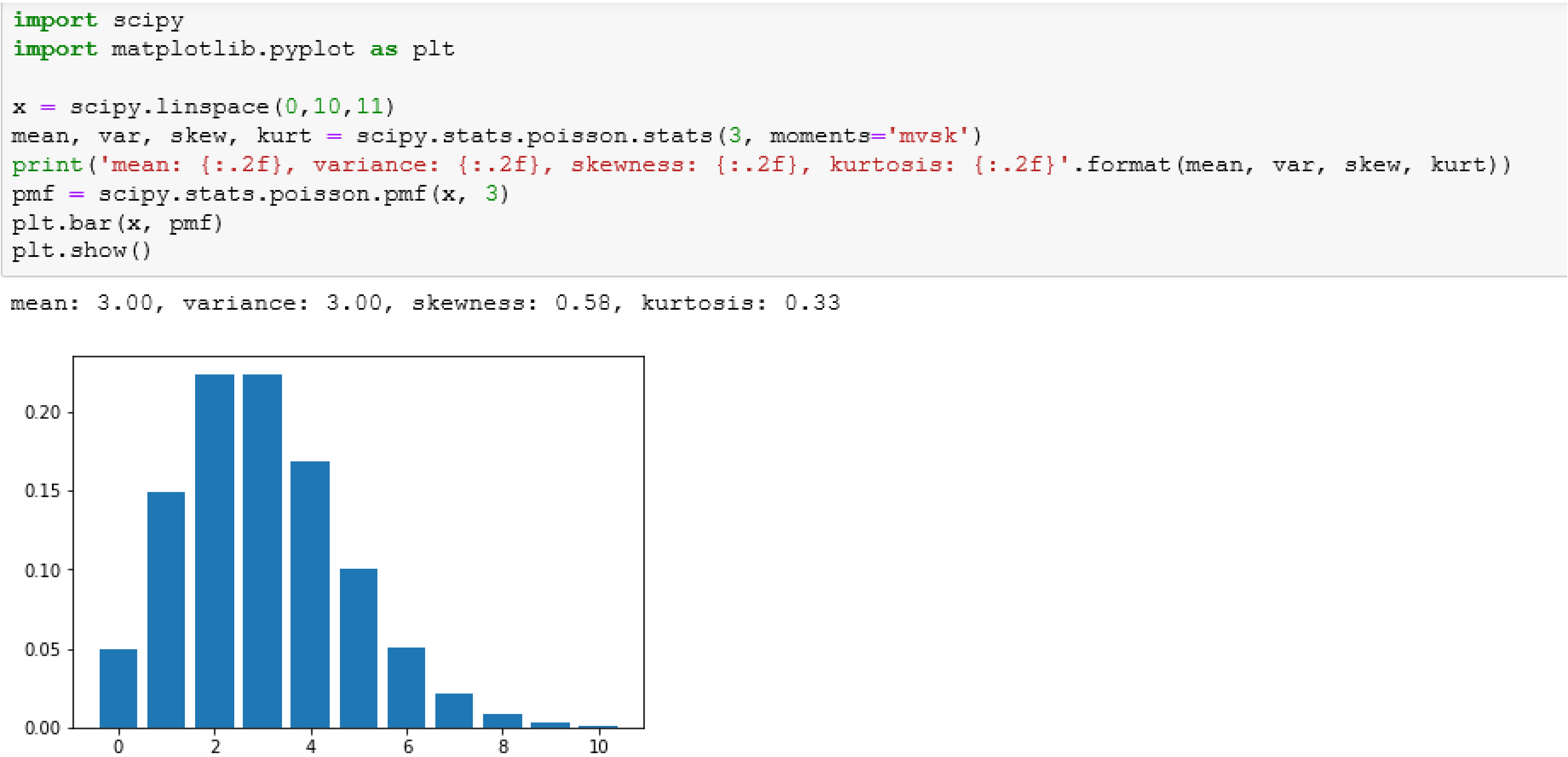
- The probability of success is proportional to the size of the interval.
- Thus knowing an error rate of 1.5 typos per 100 pages, we can determine a mean value for a 400 page book as:

μ = 1.5 (4) = 6 typos/400 pages

μ = 1.5 (4) = 6 typos / 400 pages

Copyrights© 2017, AcadGild. All rights reserved

Poisson Distribution Using Python



```
import scipy
import matplotlib.pyplot as plt

x = scipy.linspace(0,10,11)
mean, var, skew, kurt = scipy.stats.poisson.stats(3, moments='mvsk')
print('mean: {:.2f}, variance: {:.2f}, skewness: {:.2f}, kurtosis: {:.2f}'.format(mean, var, skew, kurt))
pmf = scipy.stats.poisson.pmf(x, 3)
plt.bar(x, pmf)
plt.show()
```

```
mean: 3.00, variance: 3.00, skewness: 0.58, kurtosis: 0.33
```

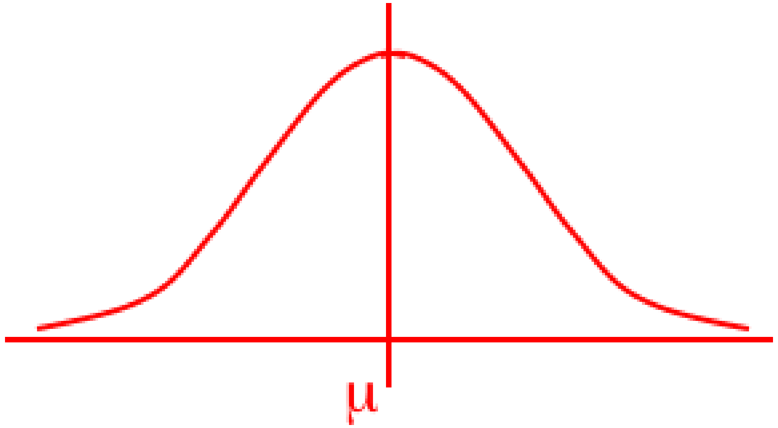
[Practice Code]

Normal Distribution

- The normal distribution is the most important of all probability distributions. The probability density function of a normal random variable is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < \infty$$

- It looks like bell shaped, symmetrical around the mean, μ



Copyrights© 2017, AcadGild. All rights reserved

Normal Distribution

- The normal distribution is completely defined by two parameters Standard Deviation and Mean.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < \infty$$

- The normal distribution is bell shaped and symmetrical about the mean.

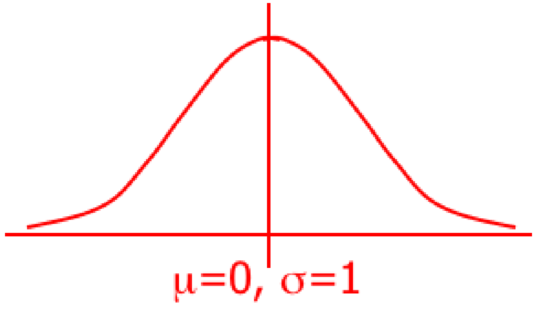
Copyrights© 2017, AcadGild. All rights reserved

Standard Normal Distribution

- A normal distribution whose mean is zero and standard deviation is 1 is called the standard normal distribution.

$$f(x) = \frac{1}{1\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-0}{1}\right)^2} \qquad -\infty < x < \infty$$

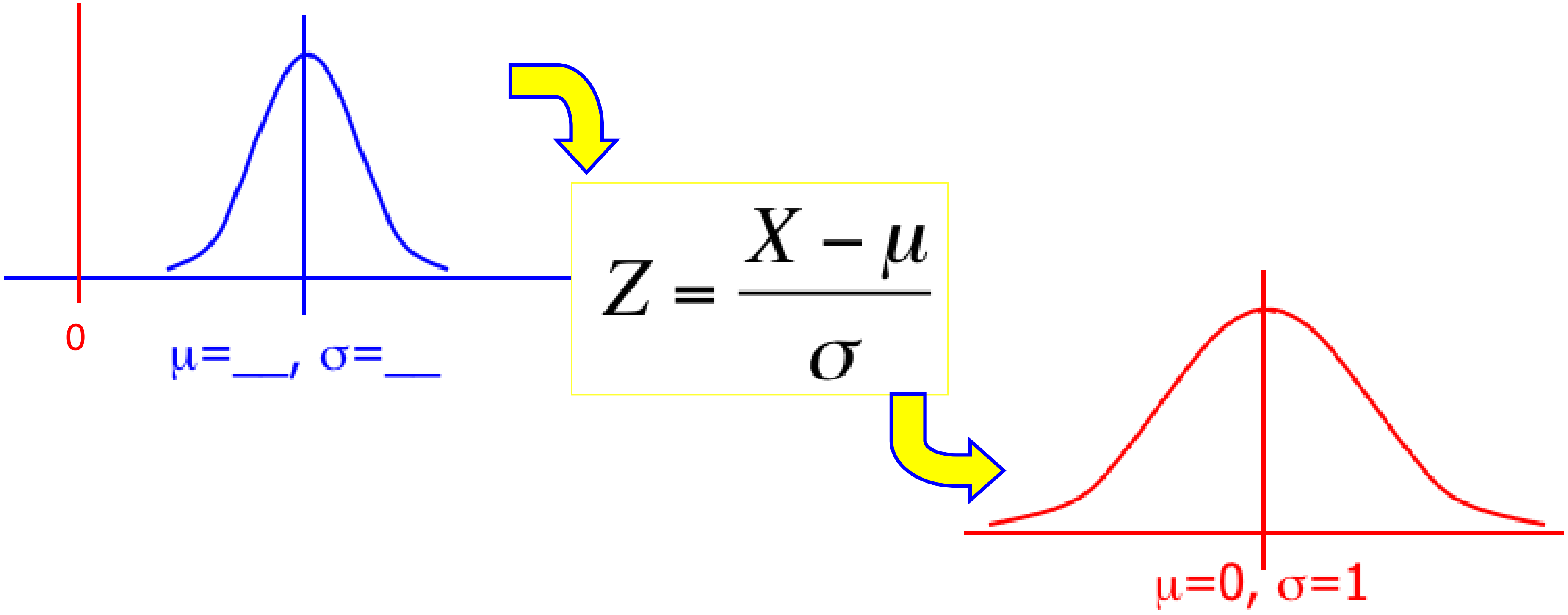
- As we shall see shortly, any normal distribution can be ***converted*** to a standard normal distribution with simple algebra. This makes calculations much easier.



Copyrights© 2017, AcadGild. All rights reserved

Calculating Normal Probabilities

- We can use the following function to convert any normal random variable to a standard normal random variable

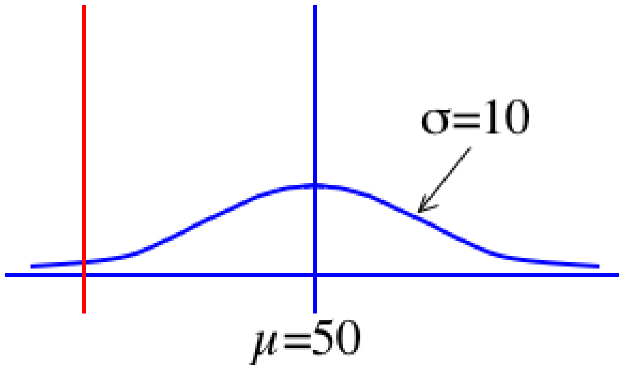


$$Z = \frac{X - \mu}{\sigma}$$

Copyrights© 2017, AcadGild. All rights reserved

Calculating Normal Probabilities

Example: The time required to build a computer is normally distributed with a mean of 50 minutes and a standard deviation of 10 minutes.



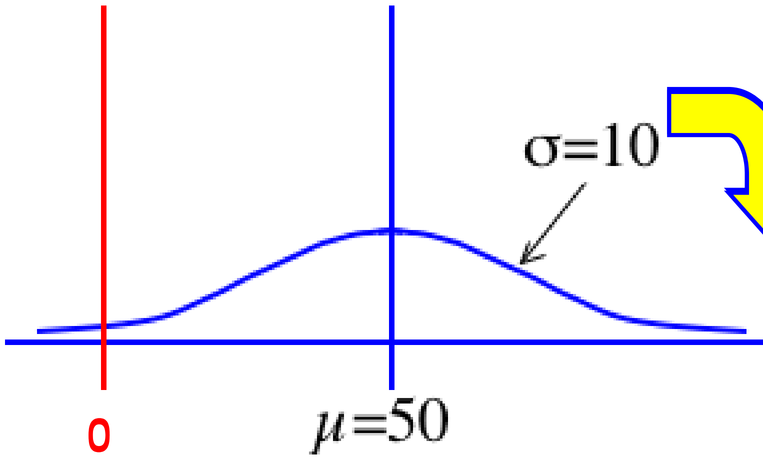
- What is the probability that a computer is assembled in a time between 45 and 60 minutes?
- Algebraically speaking, what is $P(45 < X < 60)$?

Copyrights© 2017, AcadGild. All rights reserved

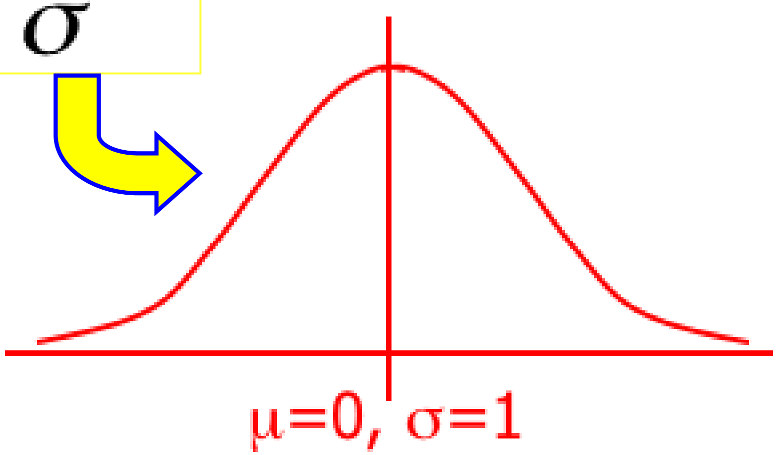
Calculating Normal Probabilities

P(45 < X < 60)?

Mean of 50 minutes and a standard deviation of 10 minutes…



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



$$P(45 < X < 60) =$$

$$P\left(\frac{45-50}{10} < \frac{X-\mu}{\sigma} < \frac{60-50}{10}\right) =$$

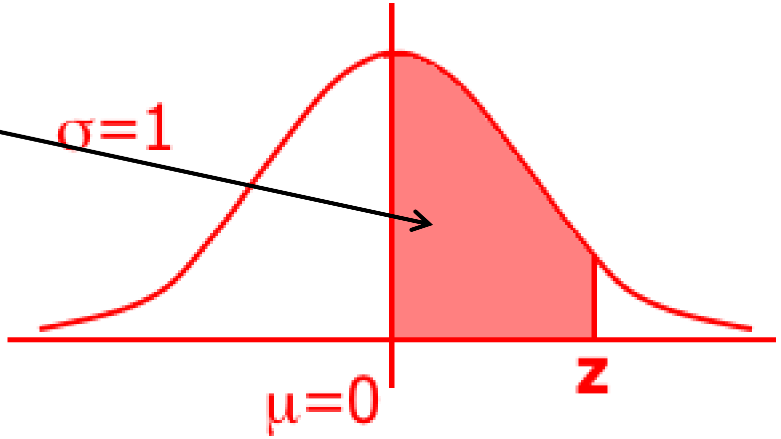
$$P(-.5 < Z < 1)$$

Copyrights© 2017, AcadGild. All rights reserved

Calculating Normal Probabilities

We can use z-table

probabilities $P(0 < Z < z)$



We can break up $P(-.5 < Z < 1)$ into:

$P(-.5 < Z < 0) + P(0 < Z < 1)$

The distribution is ***symmetric*** around zero, so we have:

$P(-.5 < Z < 0) = P(0 < Z < .5)$

Hence: $P(-.5 < Z < 1) = P(0 < Z < .5) + P(0 < Z < 1)$

Copyrights© 2017, AcadGild. All rights reserved

Calculating Normal Probabilities

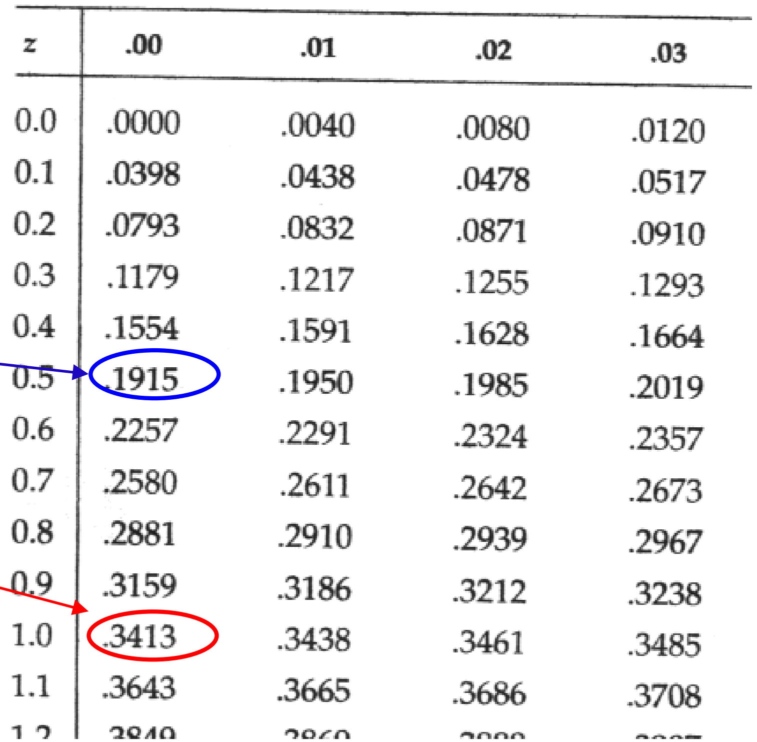
How to use z-table?

This table gives probabilities **P(0 < Z < z)**
First column = integer + first decimal
Top row = second decimal place

P(0 < Z < 0.5)

P(0 < Z < 1)

P(–.5 < Z < 1) = .1915 + .3414 = .5328



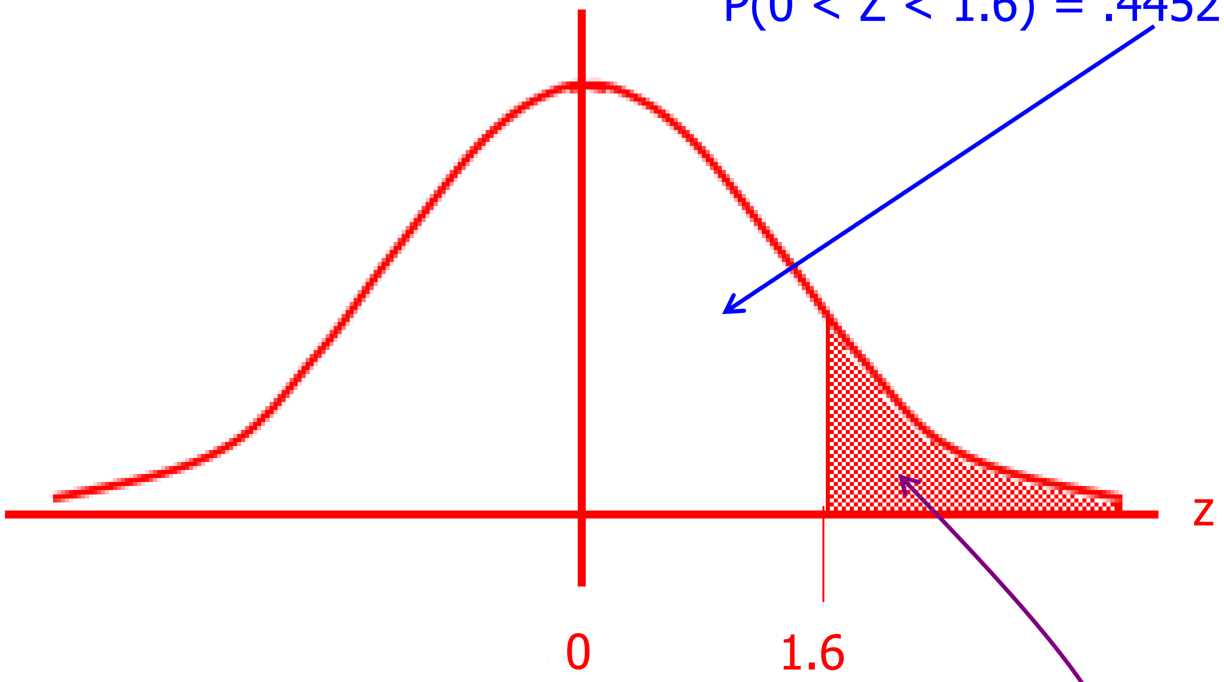
z	.00	.01	.02	.03
0.0	.0000	.0040	.0080	.0120
0.1	.0398	.0438	.0478	.0517
0.2	.0793	.0832	.0871	.0910
0.3	.1179	.1217	.1255	.1293
0.4	.1554	.1591	.1628	.1664
0.5	.1915	.1950	.1985	.2019
0.6	.2257	.2291	.2324	.2357
0.7	.2580	.2611	.2642	.2673
0.8	.2881	.2910	.2939	.2967
0.9	.3159	.3186	.3212	.3238
1.0	.3413	.3438	.3461	.3485
1.1	.3643	.3665	.3686	.3708
1.2	.3849	.3869	.3888	.3907

Copyrights© 2017, AcadGild. All rights reserved

Using the Normal Table

What is P(Z > 1.6) ?

P(0 < Z < 1.6) = .4452

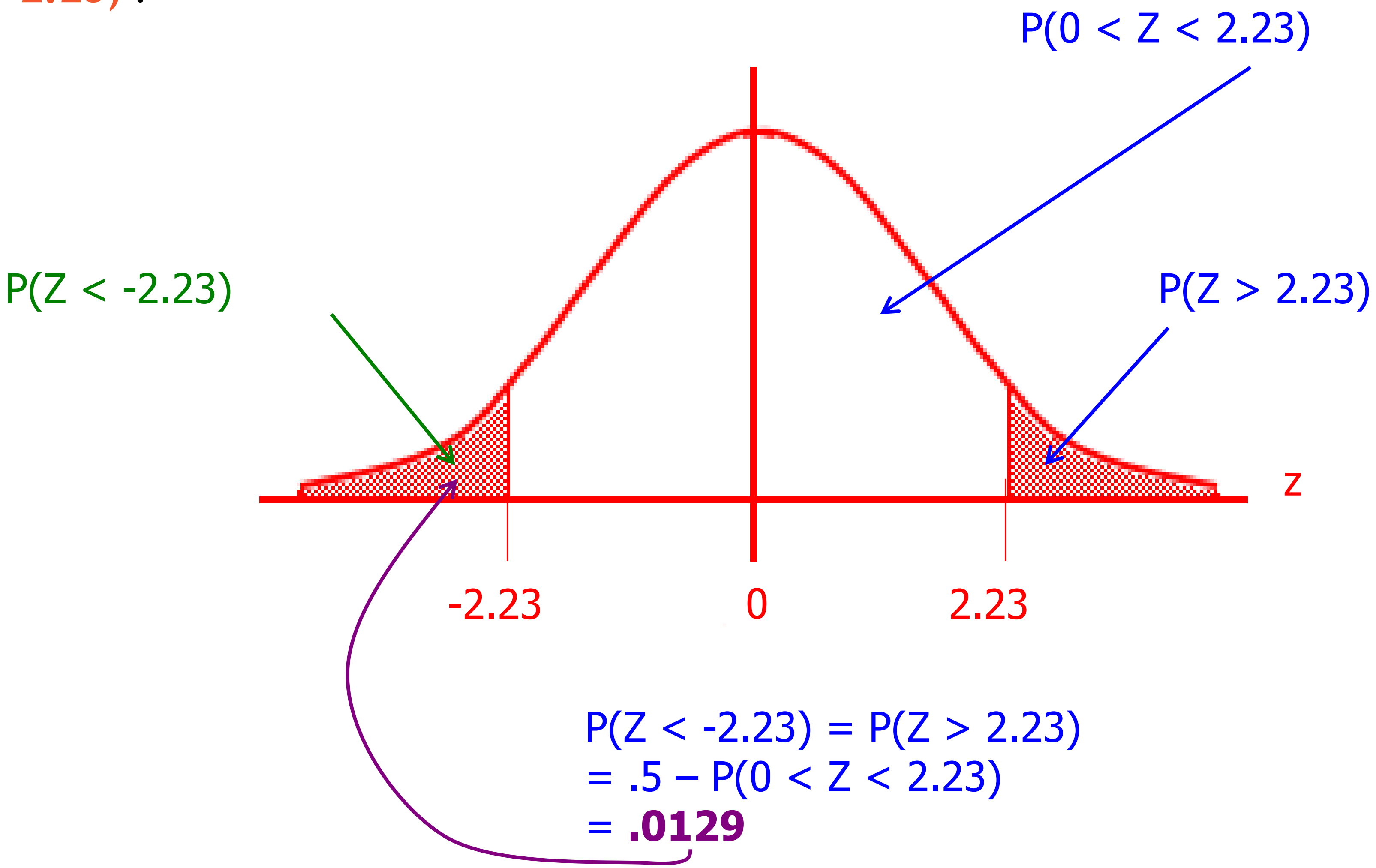


P(Z > 1.6) = .5 – P(0 < Z < 1.6)
= .5 – .4452
= **.0548**

Copyrights© 2017, AcadGild. All rights reserved

Using the Normal Table

What is P(Z < -2.23) ?

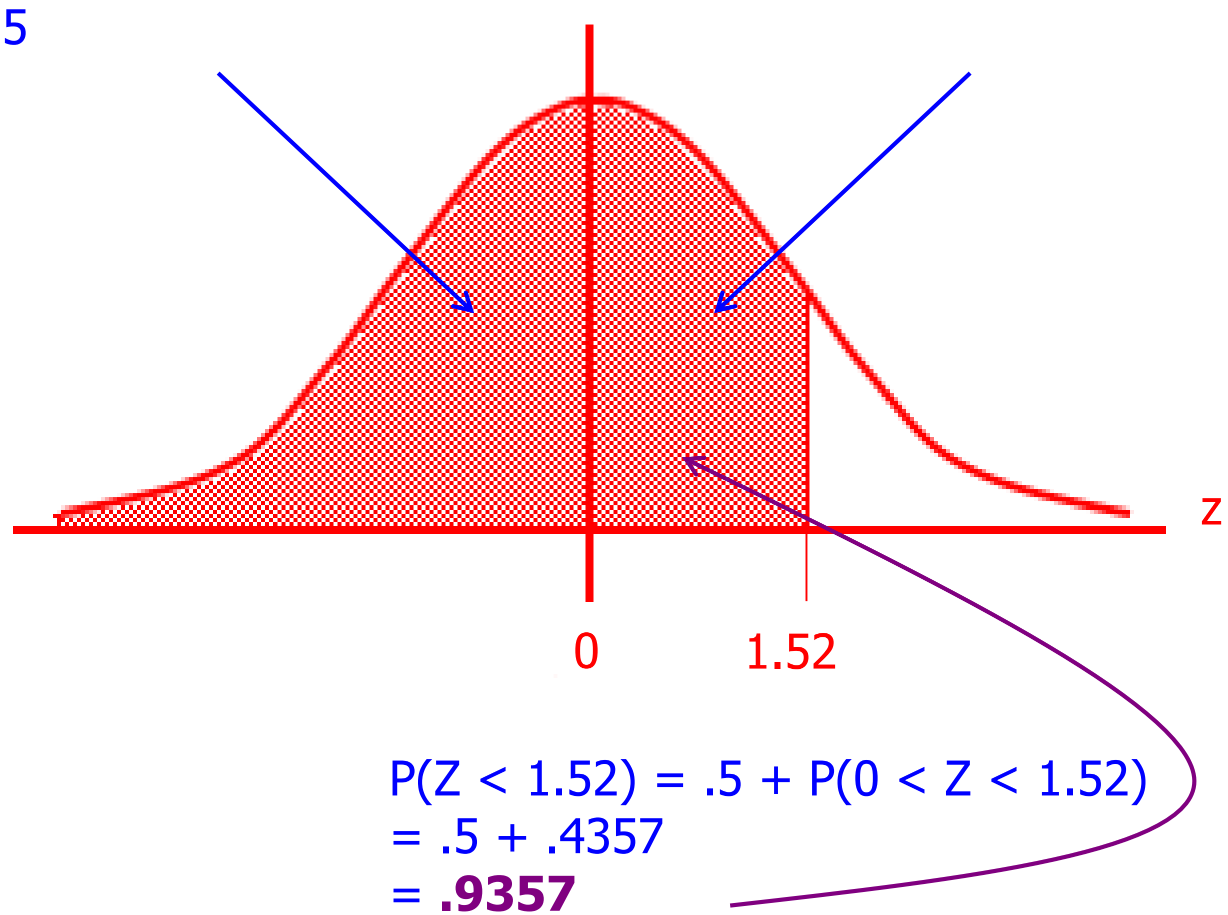


P(Z < -2.23) = P(Z > 2.23)
= .5 – P(0 < Z < 2.23)
= **.0129**

Copyrights© 2017, AcadGild. All rights reserved

Using the Normal Table

What is P(Z < 1.52) ?

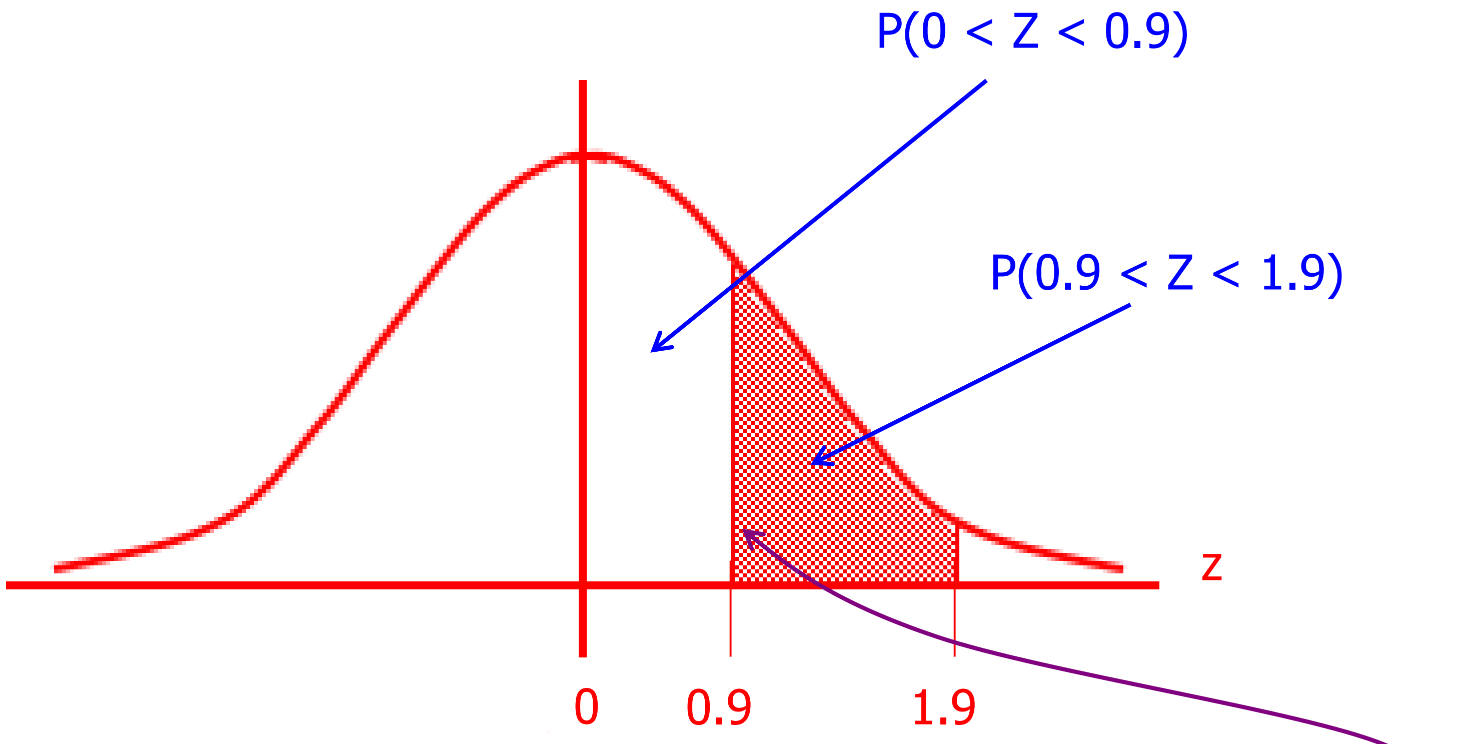


P(Z < 1.52) = .5 + P(0 < Z < 1.52)
= .5 + .4357
= **.9357**

Copyrights© 2017, AcadGild. All rights reserved

Using the Normal Table

What is P(0.9 < Z < 1.9) ?



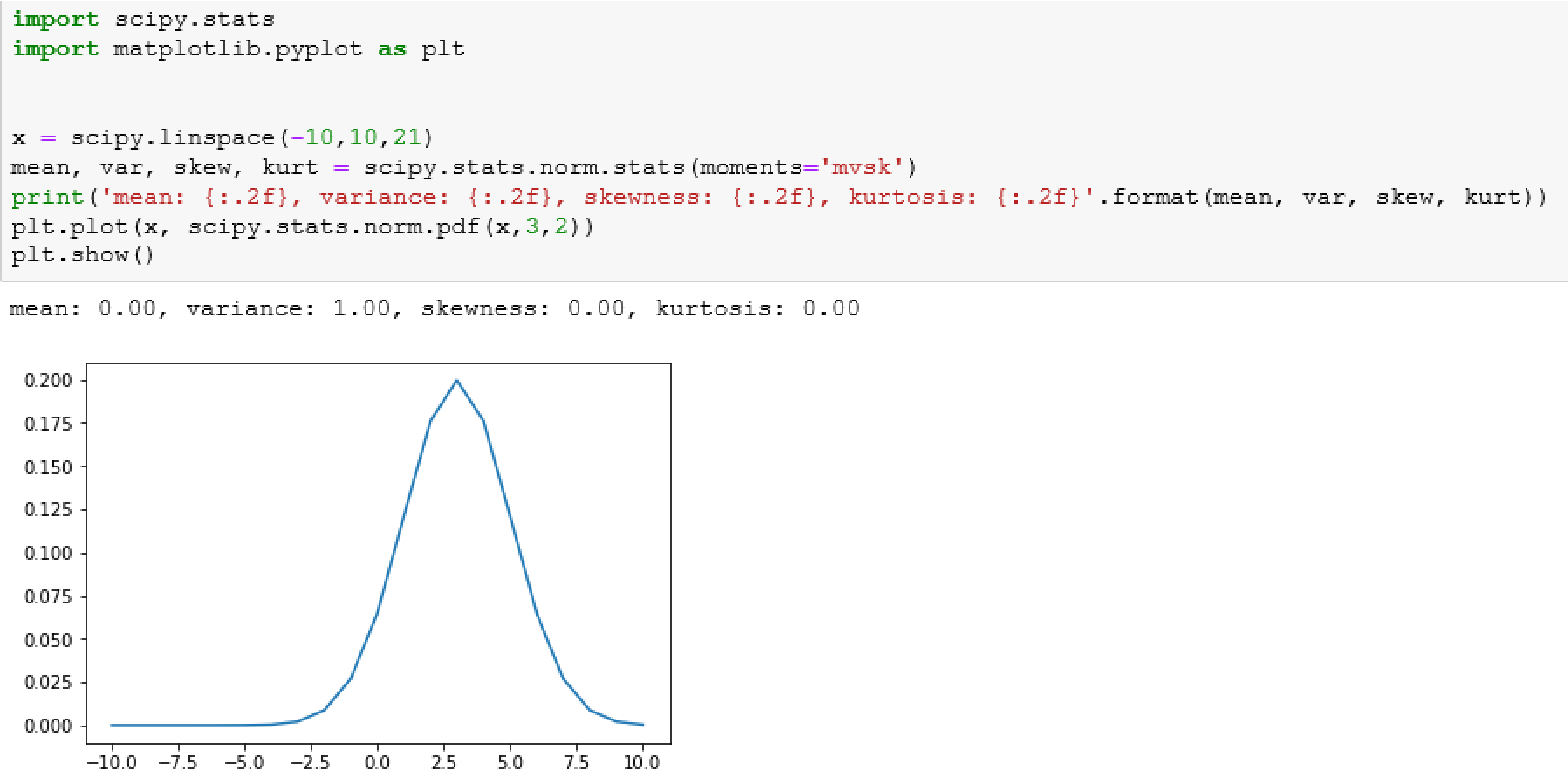
$P(0.9 < Z < 1.9) = P(0 < Z < 1.9) – P(0 < Z < 0.9)$

$=.4713 – .3159$

$=$ **.1554**

Copyrights© 2017, AcadGild. All rights reserved

Normal Distribution Using Python



```
import scipy.stats
import matplotlib.pyplot as plt


x = scipy.linspace(-10,10,21)
mean, var, skew, kurt = scipy.stats.norm.stats(moments='mvsk')
print('mean: {:.2f}, variance: {:.2f}, skewness: {:.2f}, kurtosis: {:.2f}'.format(mean, var, skew, kurt))
plt.plot(x, scipy.stats.norm.pdf(x,3,2))
plt.show()
```

```
mean: 0.00, variance: 1.00, skewness: 0.00, kurtosis: 0.00
```

Practice Code

Finding Values of Z

- The other values of Z are:
  - Z.05 = 1.645
  - Z.01 = 2.33

Copyrights© 2017, AcadGild. All rights reserved

Using the Values of Z

Because z.025 = 1.96 and - z.025= -1.96, it follows that we can state:

$P(-1.96 < Z < 1.96) = .95$

Similarly

$P(-1.645 < Z < 1.645) = .90$

Copyrights© 2017, AcadGild. All rights reserved

Central Limit Theorem

- The sampling distribution of the mean of a random sample drawn from any population is approximately normal for a sufficiently large sample size.
- The larger the sample size, the more closely the sampling distribution of ‘X’ will resemble a normal distribution.

Copyrights© 2017, AcadGild. All rights reserved

Central Limit Theorem

- If the population is normal, then ‘X’ is normally distributed for all values of n.
- If the population is not-normal, then ‘X’ is approximately normal only for larger values of n.
- In many practical situations, a sample size of 30 may be sufficiently large to allow us to use the normal distribution as an approximation for the sampling distribution of X.

Copyrights© 2017, AcadGild. All rights reserved

Sampling Distributions of the Sample Mean

1. $\mu_{\bar{x}} = \mu$

2. $\sigma_{\bar{x}}^2 = \sigma^2 / n \quad and \quad \sigma_{\bar{x}} = \sigma / \sqrt{n}$

3. If X is normal, X is normal. If X is not normal then X is approximately normal for sufficiently large sample sizes.

Note:

- The definition of “sufficiently large” depends on the extent of non-normality of X.
- For example, heavily skewed; multimodal

Copyrights© 2017, AcadGild. All rights reserved

Sampling Distributions of the Sample Mean

For Example,

The foreman of a bottling plant has observed that the amount of soda in each “32-ounce” bottle is actually a normally distributed random variable, with a mean of 32.2 ounces and a standard deviation of 0.3 ounce.

If a customer buys one bottle, what is the probability that the bottle will contain more than 32 ounces?

Copyrights© 2017, AcadGild. All rights reserved

Email us - support@acadgild.com