Assignment 1

1. Basic probability theory notation and terms

- probability
  - A value of the chance/risk that some event/outcome will occur.
- probability mass
  - For discrete events/outcomes, the value of the chance/risk that the exact event/outcome will occur.
- probability density
  - For continuous events/outcomes, a value of the chance/risk of the event/outcome that can be compared to other events/outcomes of the same distribution.
- probability mass function (pmf)
  - For discrete events/outcomes, the function that maps each possible event/outcome to the corresponding probability mass.
- probability density function (pdf)
  - For continuous events/outcomes, the function that maps each possible event/outcome to the corresponding probability density.
- probability distribution
  - The evaluation of probability over all possible outcomes, with properties as shape, skewness and kurtosis.
- discrete probability distribution
  - The specifict case of probability distribution for discrete events/outcomes - can be used interchangeable with pmf.
- continuous probability distribution
  - The specifict case of probability distribution for continuous events/outcomes - can be used interchangeable with pdf.
- cumulative distribution function (cdf)
  - The cumulative sum of the distribution function, given that the events/outcomes are somehow orderable. I.e. the probability that an event/outcome or less will occur.

- sampling distribution
  - probability distribution of given statistic (e.g. mean) for a (finite) random sample from the population distribution.
- observation model
  - a model of the relationship between the parameters of the model the observed events/outcomes.
- likelihood
  - the probability of observing events/outcomes given the (parametric) model.

2. Basic computer skills

In [1]:
```
%matplotlib inline

import numpy as np
from scipy.stats import beta

import matplotlib.pyplot as plt
```

Calculate α and β based on the given formula:

In [2]:
```
mean = 0.2
variance = 0.01
a = mean * (mean * (1 - mean) / variance - 1)
b = a * (1 - mean) / mean
print('alpha = {:.4f}'.format(a))
print('beta  = {:.4f}'.format(b))
```

```
alpha = 3.0000
beta  = 12.0000
```

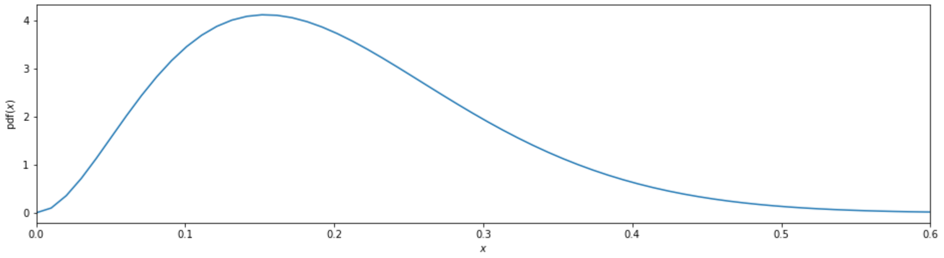
Define the distibution

In [3]:
```
dist = beta(a, b)
```

Plot the linear space between 0 and 1 in 100 steps:

In [4]:
```
x = np.linspace(0, 1, 100)
y = dist.pdf(x)
```

In [5]:
```
fig, ax = plt.subplots(figsize = (16, 4))
ax.plot(x, y)
ax.set_xlabel('$x$')
ax.set_ylabel('pdf($x$)')
ax.set_xlim(0,0.6)
plt.show()
```

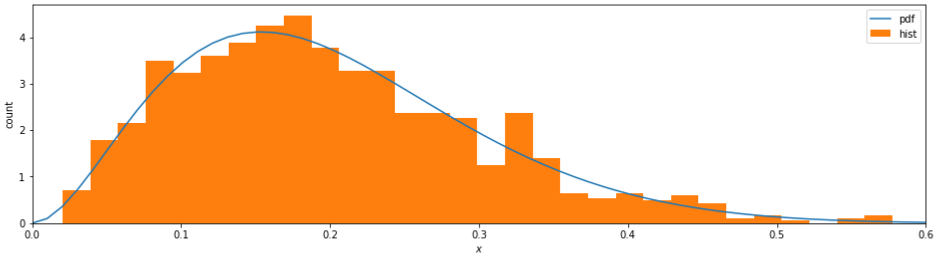


We take 1000 random samples and plot the histogram:

In [6]:
```
samples = dist.rvs(size = 1000)
```

In [7]:

```
fig, ax = plt.subplots(figsize = (16, 4))
ax.plot(x, y, label = 'pdf')
ax.hist(samples, bins= 30, normed=True, label = 'hist')
ax.set_xlabel('$x$')
ax.set_ylabel('count')
ax.set_xlim(0,0.6)
ax.legend(loc = 'upper right')
plt.show()
```



The histogram seems to follow the original distribution as expected.

We evaluate the sample mean and variance and confirm they are approx. the true orginal values used to create the distribution:

In [8]:

```
print('Sample mean = {:.4f}'.format(np.mean(samples)))
print('Sample var  = {:.4f}'.format(np.var(samples)))
```

```
Sample mean = 0.2000
Sample var  = 0.0102
```

We compute the Central 95%-interval of the drawn sample using the 2.5% and 97.5% percentiles as bounds:

In [9]:

```
lwr = np.percentile(samples, q = 2.5)
upr = np.percentile(samples, q = 97.5)
print('Central 95%-interval of sample: {:.4f} - {:.4f}'.format(lwr, upr))
```

```
Central 95%-interval of sample: 0.0502 - 0.4369
```

3. Bayes' theorem I

Given that lung cancer is rare in the general population, $p(\text{Cancer}) = 0.1\%$, we evaluate the results of the proposed screening method:

From the text we have:

$$p(\text{Positive} \mid \text{Cancer}) = 98\%$$
$$p(\text{Negative} \mid \text{No Cancer}) = 96\%$$

We easily infer:

$$p(\text{Negative} \mid \text{Cancer}) = 2\%$$
$$p(\text{Positive} \mid \text{No Cancer}) = 4\%$$

In [10]:

```
p_Cancer = .001
p_Pos_Cancer = .98
p_Neg_NoCancer = .96
p_NoCancer = 1 - p_Cancer
p_Pos_NoCancer = 1 - p_Neg_NoCancer
p_Neg_Cancer = 1 - p_Pos_Cancer
```

We calculate the propability for wrongly classify a patient, e.g.:

$$p(\text{Positive}, \text{No Cancer}) = p(\text{Positive} \mid \text{No Cancer})p(\text{No Cancer})$$
$$p(\text{Negative}, \text{Cancer}) = p(\text{Negative} \mid \text{Cancer})p(\text{Cancer})$$

In [11]:

```
p_PosNoCancer = p_Pos_NoCancer * p_NoCancer
p_NegCancer = p_Neg_Cancer * p_Cancer
print('p_PosNoCancer = {:.4%}'.format(p_PosNoCancer))
print('p_NegCancer = {:.4%}'.format(p_NegCancer))
```

```
p_PosNoCancer = 3.9960%
p_NegCancer = 0.0020%
```

The result of $p(\text{Positive},\ \text{No Cancer})$ is acceptable, given that the next step for a positively testet patient is more extensive, and thorough tests (and not immediately surgery!). It is still means that 40 out of 1000 will be sent to more thorough tests without having cancer, but this still seem much more efficient then sending all 1000 throug the extensive and expensive test.

However the $p(\text{Negative},\ \text{Cancer})$ is more concerning - this outcome can be fatal for the patient. Out of 100,000 tests, we would expect 100 to have cancer - but 2 of these will not receive the more extensive, and thorough tests.

My advice will be to improve the $p(\text{Negative},\ \text{Cancer})$ outcome before releasing the test for the market.

4. Bayes' theorem II

We have from the text:

$$p(\text{red} \mid \text{A}) = 2/7$$
$$p(\text{red} \mid \text{B}) = 4/5$$
$$p(\text{red} \mid \text{C}) = 1/4$$
$$p(\text{A}) = 40\%$$
$$p(\text{B}) = 10\%$$
$$p(\text{C}) = 50\%$$

Thus the probability of picking a red ball is:

$$p(\text{red} \mid \text{A})\ p(\text{A}) + p(\text{red} \mid \text{B})\ p(\text{B}) + p(\text{red} \mid \text{C})\ p(\text{C}) \approx 31.9\%$$

In [12]:

```
p_red_A = 2/7
p_red_B = 4/5
p_red_C = 1/4
p_A = .4
p_B = .1
p_C = .5
p_red = p_red_A * p_A + p_red_B * p_B + p_red_C * p_C
print('{:.1%}'.format(p_red))
```

```
31.9%
```

We want to calculate $p(x \mid \text{red})$ for $x \in \{\text{A},\ \text{B},\ \text{C}\}$ - we apply Bayes' theorem:

In [13]:

```
p_A_red = p_red_A * p_A / p_red
p_B_red = p_red_B * p_B / p_red
p_C_red = p_red_C * p_C / p_red
print('p_A_red = {:.1%}'.format(p_A_red))
print('p_B_red = {:.1%}'.format(p_B_red))
print('p_C_red = {:.1%}'.format(p_C_red))
```

```
p_A_red = 35.8%
p_B_red = 25.1%
p_C_red = 39.1%
```

From this we conclude that if we pick a red ball - it is most likely, that it came from box C.

5. Bayes' theorem III

We have from the text:

$$p(\text{fraternal}) = 1/125$$
$$p(\text{identical}) = 1/300$$

Since we dont know any other ways of becomming a twin we assume:

$$p(\text{twin}) = p(\text{fraternal}) + p(\text{identical})$$
$$p(\text{twin} \mid \text{fraternal}) = 1$$
$$p(\text{twin} \mid \text{identical}) = 1$$

In [14]:

```
p_fraternal = 1/125
p_identical = 1/300
p_twin = p_fraternal + p_identical
```

Since we knew Elvis were a twin, we want to calculate the following:

$$p(\text{identical} \mid \text{twin})$$

We apply Bayes' theorem and see that there were a 29% chance that Elvis and his death twin brother were identical twins:

In [16]:

```
p_identical_twin = p_identical / p_twin
print('{:.1%}'.format(p_identical_twin))
```

```
29.4%
```