대표적인
연속형 확률변수

정규분포

In [1] :

```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats, integrate
from scipy.optimize import minimize_scalar

%precision 3
%matplotlib inline
```

In [2] :

```
linestyles = [ ‘ - ’ , ‘ -- ’ , ‘ : ’ ]

def E(X, g=lambda x: x):
    x_range, f = X
    def integrand(x):
        return g(x) * f(x)
    return integrate.quad(integrand, -np.inf, np.inf)[0]

def V(X, g=lambda x: x):
    x_range, f = X
    mean = E(X, g)
    def integrand(x):
        return (g(x) - mean) ** 2 * f(x)
    return integrate.quad(integrand, -np.inf, np.inf)[0]
```

정규분포

```
def check_prob(X):
    x_range, f = X
    f_min = minimize_scalar(f).fun
    assert f_min >= 0, 'density function is minus value'
    prob_sum = np.round(integrate.quad(f, -np.inf, np.inf)[0], 6)
    assert prob_sum == 1, f'sum of probability is {prob_sum}'
    print(f'expected vaue{E(X):.3f}')
    print(f'variance{V(X):.3f}')

def plot_prob(X, x_min, x_max):
    x_range, f = X
    def F(x):
        return integrate.quad(f, -np.inf, x)[0]

    xs = np.linspace(x_min, x_max, 100)

    fig = plt.figure(figsize=(10, 6))
    ax = fig.add_subplot(111)
    ax.plot(xs, [f(x) for x in xs],
            label='f(x)', color='gray')
    ax.plot(xs, [F(x) for x in xs],
            label='F(x)', ls='--', color='gray')

    ax.legend()
    plt.show()
```

```
x = "hello"

#if condition returns True, then nothing happens:
assert x == "hello"

#if condition returns False, AssertionError is raised:
assert x == "goodbye"
```

```
Traceback (most recent call last):
  File "demo_ref_keyword_assert.py", line 5, in <module>
    assert x == "goodbye"
AssertionError
```

```
x = "hello"

#if condition returns False, AssertionError is raised:
assert x == "goodbye", "x should be 'hello'"
```

```
Traceback (most recent call last):
  File "demo_ref_keyword_assert2.py", line 4, in <module>
    assert x == "goodbye", "x should be 'hello'"
AssertionError: x should be 'hello'
```

정규분포

- 평균 μ, 분산 σ^2 ~ $\mathrm{N}(\mu, \sigma^2)$

정규분포의 밀도함수

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

- 남자 고등학생의 키 ~ $\mathrm{N}(170, 5^2)$

$$P(165 \le X \le 175) = \int_{165}^{175} \frac{1}{\sqrt{2\pi} \times 5} \exp\left\{-\frac{(x-170)^2}{2 \times 5^2}\right\} dx \simeq 0.683$$

- 모의고사 점수 ~ $\mathrm{N}(70, 8^2)$

$$P(54 \le X \le 86) = \int_{54}^{86} \frac{1}{\sqrt{2\pi} \times 8} \exp\left\{-\frac{(x-70)^2}{2 \times 8^2}\right\} dx \simeq 0.954$$

정규분포

정규분포의 기댓값과 분산

$X \sim N(\mu,\ \sigma^2)$이라고 할 때

$$E(X) = \mu, \quad V(X) = \sigma^2$$

정규분포의 변환

$X \sim N(\mu,\ \sigma^2)$이라고 할 때, 임의의 실수 a, b에 대해서

$$aX + b \sim N(a\mu + b,\ a^2\sigma^2)$$

이 성립합니다.

정규분포

- $X \sim N(\mu,\ \sigma^2)$을 정규화한 $Z = \frac{X-\mu}{\sigma}$ 는 표준정규분포를 따름

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

- $Z \sim N(0,\ 1)$

In [3] :

```
def N(mu, sigma):
    x_range = [- np.inf, np.inf]
    def f(x):
        return 1 / np.sqrt(2 * np.pi * sigma**2) *\
                    np.exp(-(x-mu)**2 / (2 * sigma**2))
    return x_range, f
```

- $X \sim N(2,\ 0.5^2)$

In [4] :

```
mu, sigma = 2, 0.5
X = N(mu, sigma)
```

정규분포

- $X \sim N(2,\ 0.5^2)$

In [4] :

```
mu, sigma = 2, 0.5
X = N(mu, sigma)
```

- 기댓값과 분산

In [5] :

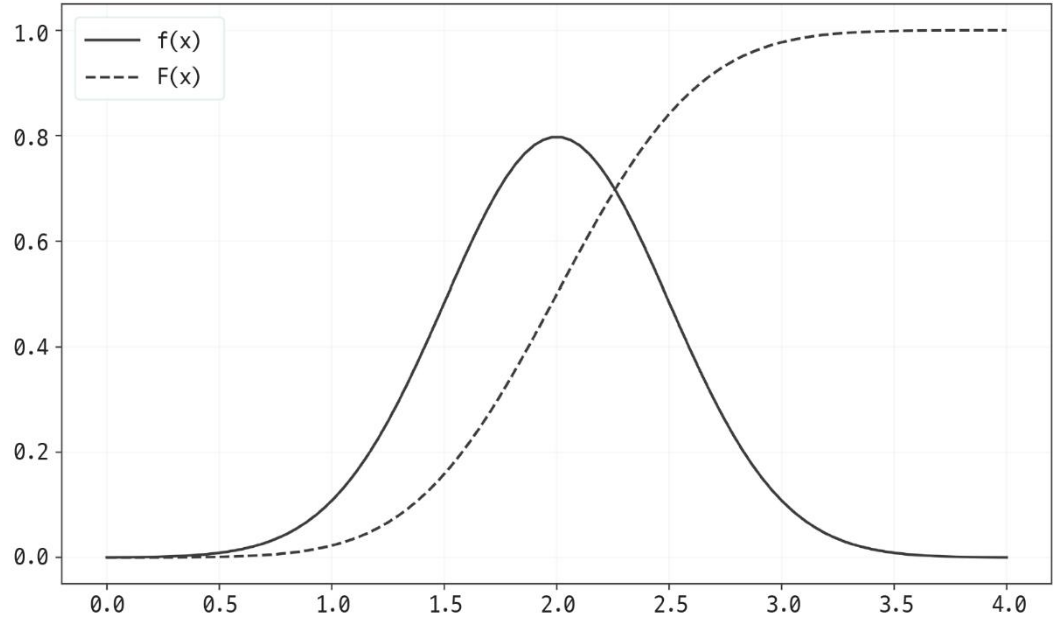
```
check_prob(X)
```

Out [5] :

```
expected vaue 2.000
variance 0.250
```

In [6] :

```
plot_prob(X, 0, 4)
```



[그림 8-1] 정규분포

정규분포

- $X \sim N(2,\ 0.5^2)$

In [7] :

```
rv = stats.norm(2, 0.5)
```

- 기댓값과 분산

In [8] :

```
rv.mean(), rv.var()
```

Out [8] :

```
(2.000, 0.250)
```

- 밀도함수

In [9] :

```
rv.pdf(2)
```

Out [9] :

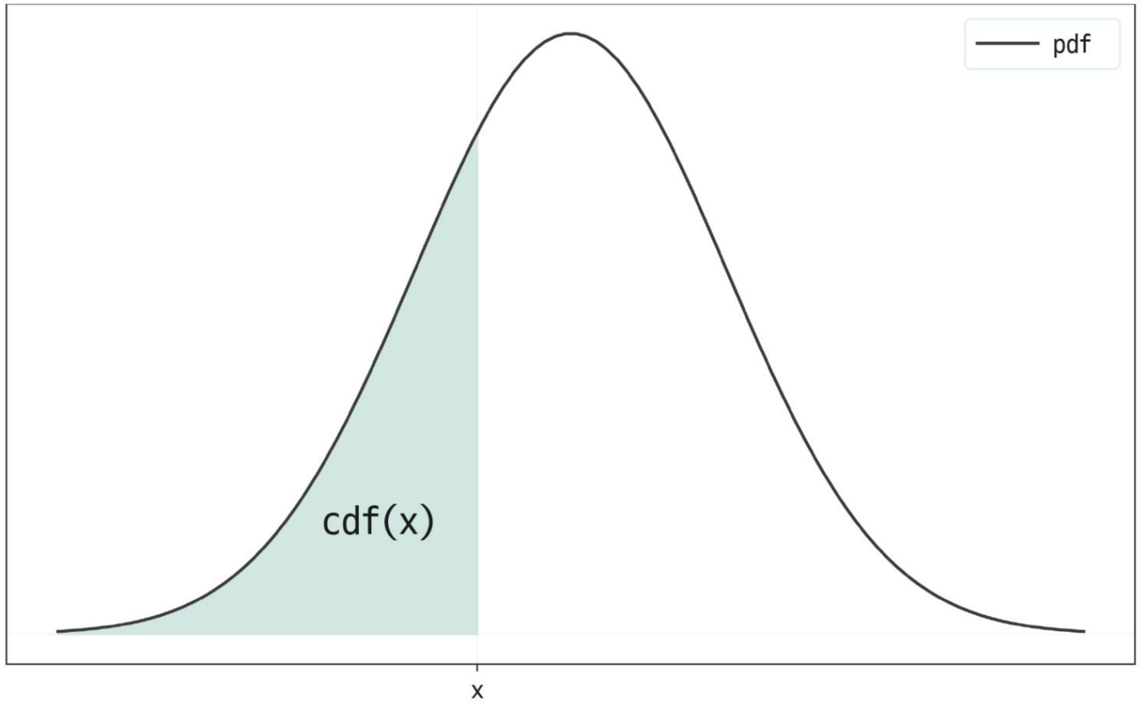
```
0.798
```

In [10] :

```
rv.pdf(1.7)
```

Out [10] :

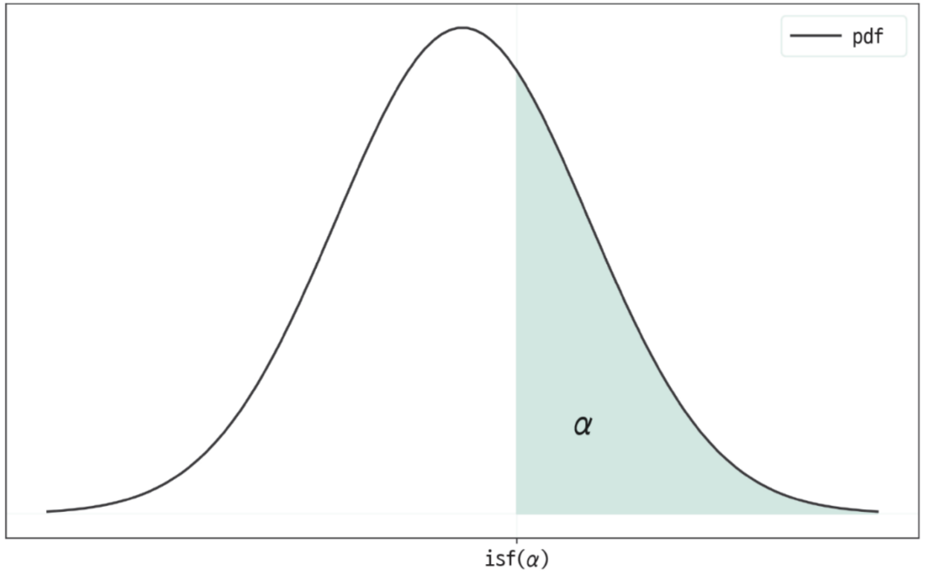
```
0.274
```



[그림 8-2] cdf 메서드

정규분포

- 상위 100α%점: z_α
  - $P(X \geq x) = \alpha$를 만족하는 x
  - $Z \sim N(0,\ 1),\ P(Z \geq z_\alpha) = \alpha$를 만족
  - $z_{1-\alpha} = -z_\alpha$



[그림 8-3] isf 메서드

- 상위 30%

In [11] :

```
rv.isf(0.3)
```

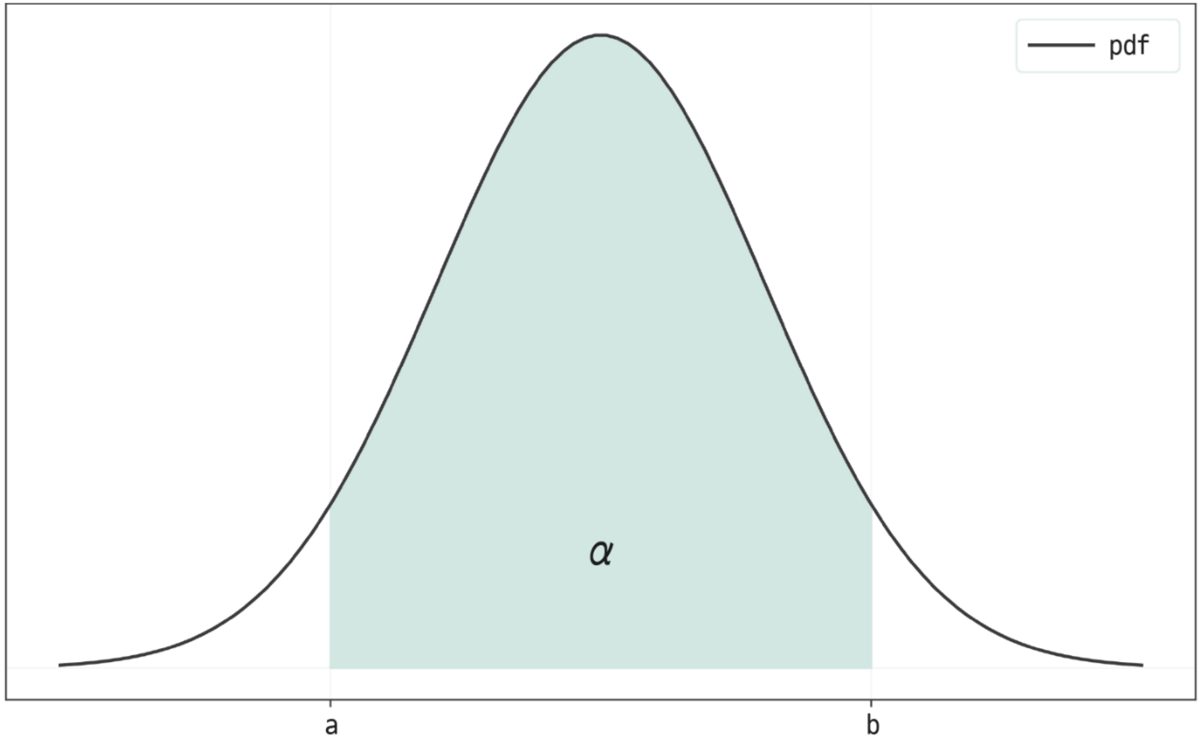
Out [11] :

```
2.262
```

정규분포

- 구간 [a, b]는 $100\alpha\%$ 구간

$$P(a \le X \le b) = \alpha \qquad P(X \le a) = P(X \ge b) = \frac{1-\alpha}{2}$$



[그림 8-4] interval 메서드

- 90% 구간

In [12] :

```
rv.interval(0.9)
```

Out [12] :

```
(1.178, 2.822)
```

In [13] :

```
rv.isf(0.95), rv.isf(0.05)
```

Out [13] :

```
(1.178, 2.822)
```

정규분포

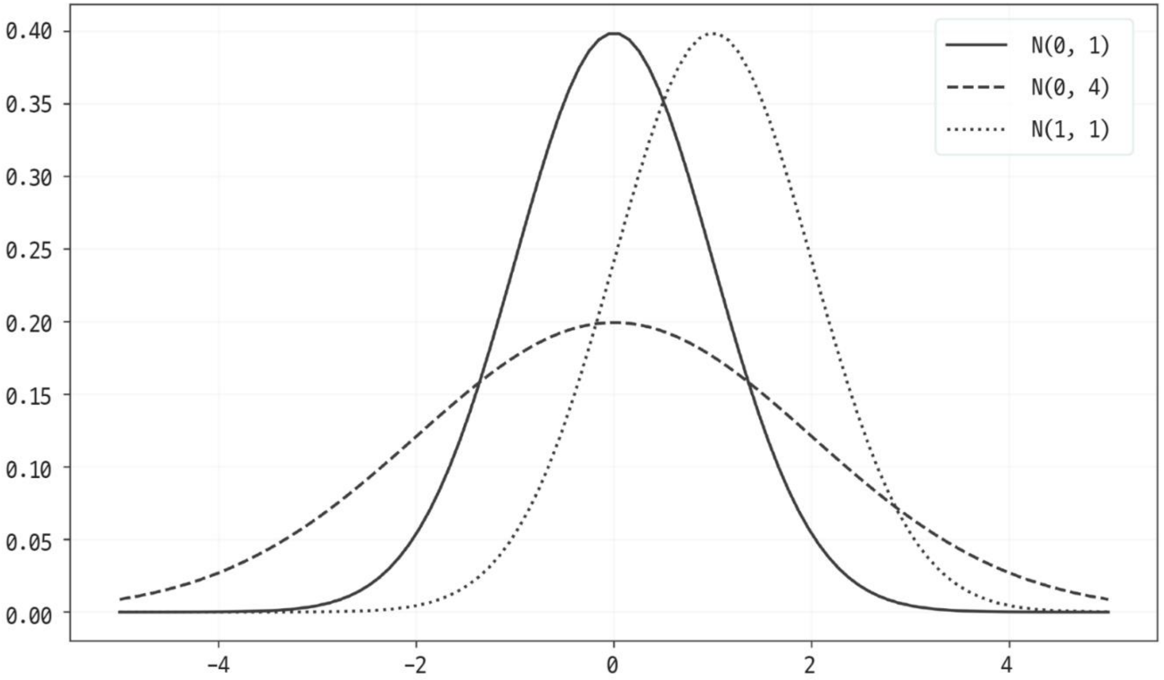
- 표준정규분포의 $100(1-\alpha)\%$ 구간 $[z_{1-\alpha/2}, z_{\alpha/2}]$.
- 예: 표준정규분포의 95% 구간 $[z_{0.975}, z_{0.025}]$

In [14] :

```
fig = plt.figure(figsize=(10, 6))
ax = fig.add_subplot(111)

xs = np.linspace(-5, 5, 100)
params = [(0, 1), (0, 2), (1, 1)]
for param, ls in zip(params, linestyles):
    mu, sigma = param
    rv = stats.norm(mu, sigma)
    ax.plot(xs, rv.pdf(xs),
            label=f'N({mu}, {sigma**2})', ls=ls, color='gray')
ax.legend()

plt.show()
```



[그림 8-5] 다양한 정규분포 SAMPLE CODE

정규분포

[표 8-1] 정규분포의 정리

파라미터	μ, σ
취할 수 있는 값	실수 전체
밀도함수	$\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$
기댓값	μ
분산	σ^2
scipy.stats	norm(μ, σ)

지수분포

- 어떤 사건이 발생하는 간격이 따르는 분포

- 파라미터 λ인 지수분포는 Ex(λ)
  - 단위시간당 평균 λ번 발생하는 사건의 발생 간격을 따르는 확률분포

지수분포의 밀도함수

$$f(x)=\begin{cases}\lambda e^{-\lambda x} & (x \geq 0) \\ 0 & (otherwise)\end{cases}$$

- 하루당 평균 2건의 교통사고가 발생하는 지역에서 3일 이내 또 교통사고가 일어날 확률

$$P(X \leq 3)=\int_0^3 2e^{-2x}dx \simeq 0.998$$

- 1시간당 평균 10번 액세스하는 사이트에서 1분 이내에 또 액세스할 확률

$$P\left(X \leq \frac{1}{60}\right)=\int_0^{60} 10e^{-10x}dx \simeq 0.154$$

지수분포

지수분포의 기댓값과 분산

$X \sim Ex(\lambda)$라고 할 때

$$E(X) = \frac{1}{\lambda}, \quad V(X) = \frac{1}{\lambda^2}$$

In [15] :

```
def Ex(lam):
    x_range = [0, np.inf]
    def f(x):
        if x >= 0:
            return lam * np.exp(-lam * x)
        else:
            return 0
    return x_range, f
```

- X ~ Ex(3)

In [16] :

```
lam = 3
X = Ex(lam)
```

In [17] :

```
check_prob(X)
```

Out [17] :

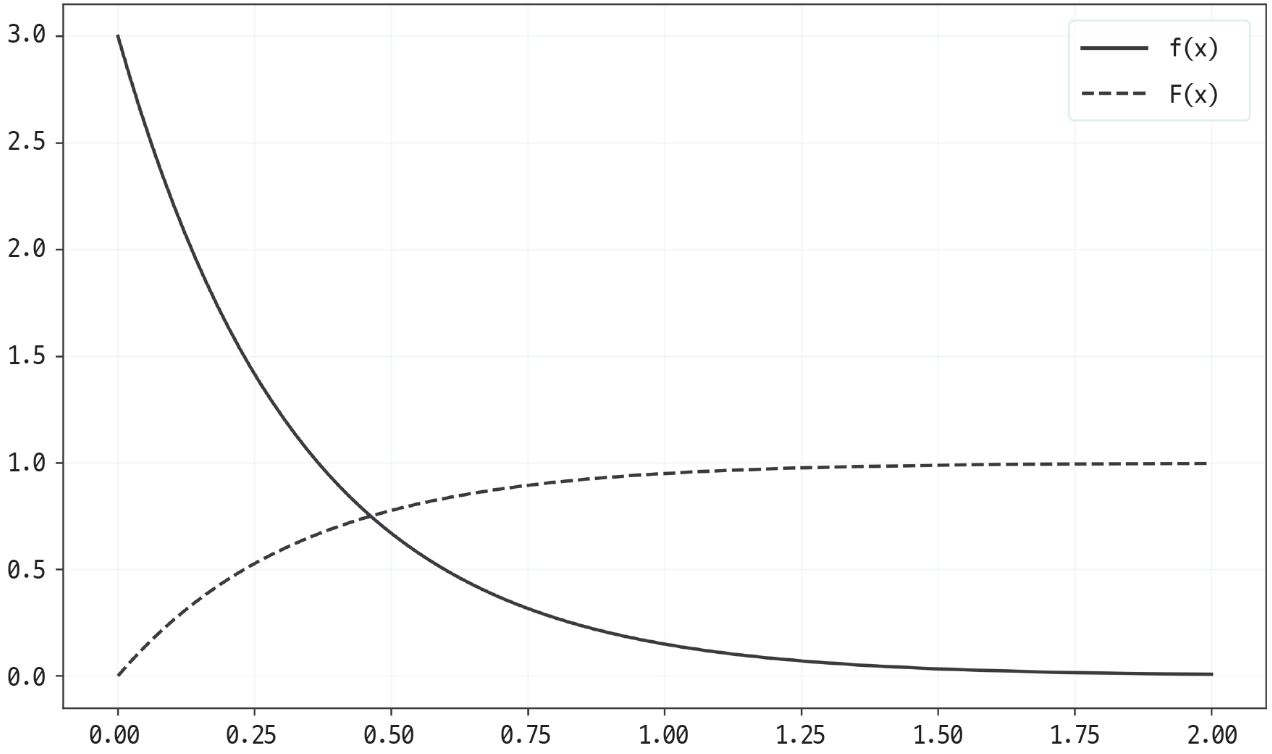
```
expected vaue 0.333
variance 0.111
```

지수분포

- 0부터 2 사이의 구간에서 밀도함수와 분포함수

In [18] :

```
plot_prob(X, 0, 2)
```



[그림 8-6] 지수분포

- scipy.stats의 expon 함수는 지수분포를 따르는 확률변수를 생성

In [19] :

```
fig = plt.figure(figsize=(10, 6))
ax = fig.add_subplot(111)

xs = np.linspace(0, 3, 100)
for lam, ls in zip([1, 2, 3], linestyles):
    rv = stats.expon(scale=1/lam)
    ax.plot(xs, rv.pdf(xs),
            label=f'lambda:{lam}', ls=ls, color='gray')
ax.legend()

plt.show()
```

지수분포

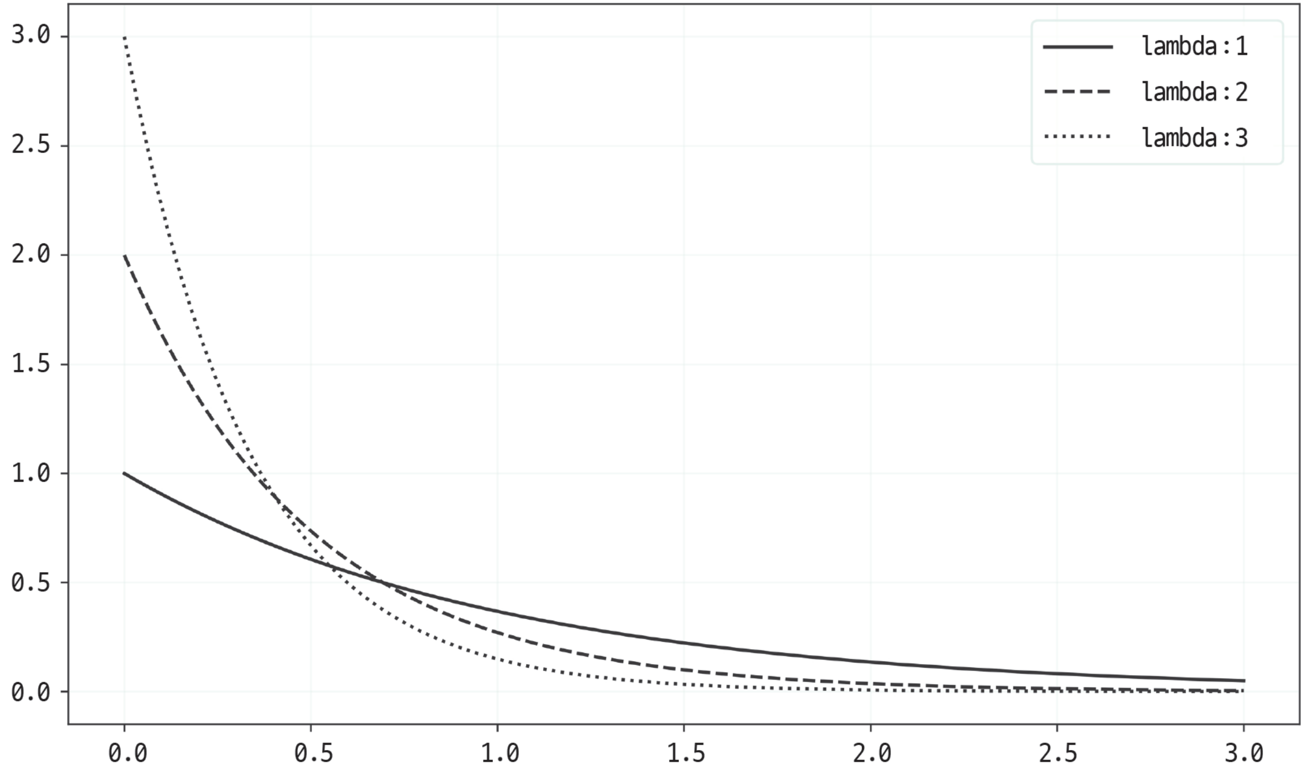
- scipy.stats의 expon 함수는 지수분포를 따르는 확률변수를 생성

In [19] :

```
fig = plt.figure(figsize=(10, 6))
ax = fig.add_subplot(111)

xs = np.linspace(0, 3, 100)
for lam, ls in zip([1, 2, 3], linestyles):
    rv = stats.expon(scale=1/lam)
    ax.plot(xs, rv.pdf(xs),
            label=f'lambda:{lam}', ls=ls, color='gray')
ax.legend()

plt.show()
```



[그림 8-7] 다양한 지수분포 SAMPLE CODE

지수분포

[표 8-2] 지수분포의 정리

파라미터	λ
취할 수 있는 값	양의 실수
밀도함수	$\lambda e^{-\lambda x}$
기댓값	$\frac{1}{\lambda}$
분산	$\frac{1}{\lambda^2}$
scipy.stats	expon(scale $=\frac{1}{\lambda}$)

카이제곱분포

- 10장 이후에 설명하는 추정과 검정에 사용하는 특수한 확률분포
- 분산의 구간추정이나 독립성 검정에서 사용

카이제곱분포

$Z_1, Z_2, \cdots, Z_n$이 서로 독립이고 $N(0, 1)$을 따르고 있을 때, 그 제곱합

$$Y = \sum_{i=1}^{n} Z_i^2$$

의 확률분포를 자유도가 n인 카이제곱분포라고 합니다.

In [20] :

```
n = 10
rv = stats.norm()
sample_size = int(1e6)
# 표준정규분포에서 표본 크기 100만으로 무작위추출한다
Zs_sample = rv.rvs((n, sample_size))
# axis=0에서 총합을 구하고, 표준정규분포의 제곱합 표본 데이터를 구한다
chi2_sample = np.sum(Zs_sample**2, axis=0)
```

- $\sum_{i=1}^{10} {Z_i}^2$에서 무작위추출한 표본크기 100만의 표본 데이터

카이제곱분포

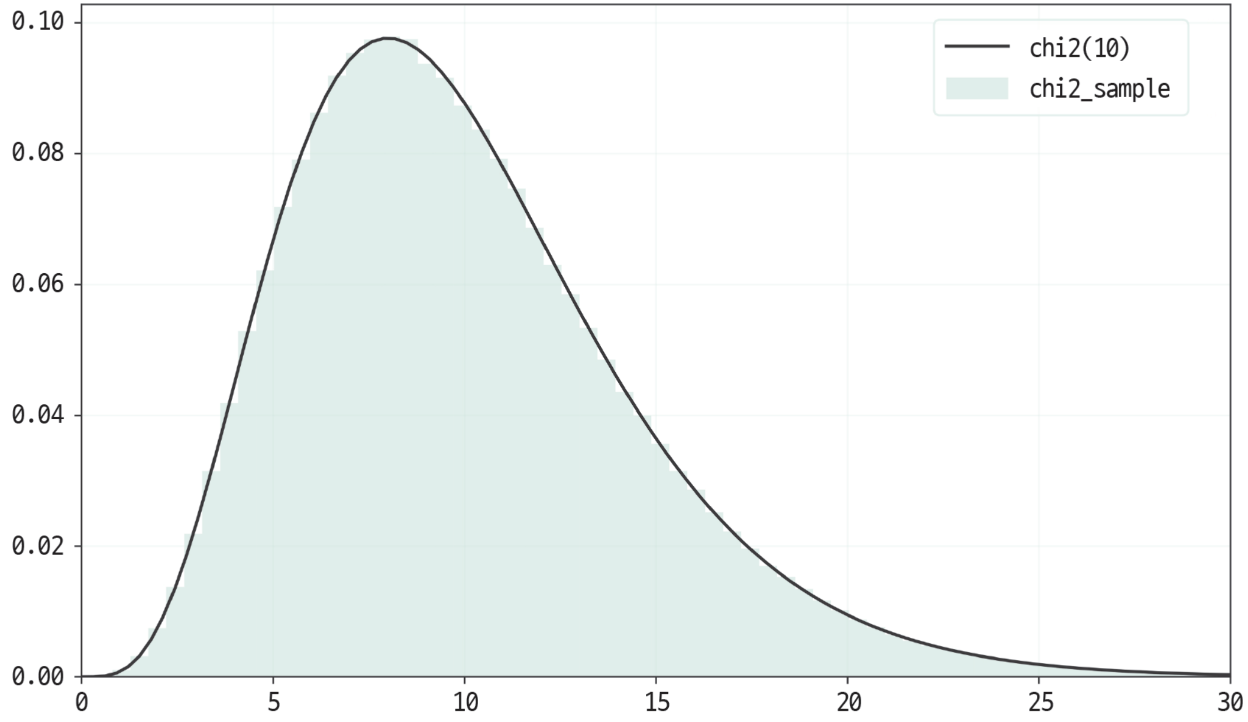
- $\sum_{i=1}^{10} Z_i{}^2$에서 무작위추출한 표본 데이터의 히스토그램과 $\chi^2(10)$의 밀도함수

In [21] :

```
fig = plt.figure(figsize=(10, 6))
ax = fig.add_subplot(111)

rv_true = stats.chi2(n)
xs = np.linspace(0, 30, 100)
ax.hist(chi2_sample, bins=100, density=True,
        alpha=0.5, label='chi2_sample')
ax.plot(xs, rv_true.pdf(xs), label=f'chi2({n})', color='gray')

ax.legend()
ax.set_xlim(0, 30)
plt.show()
```



[그림 8-8] 카이제곱분포와 표준정규분포의 관계

→ 히스토그램과 밀도함수가 일치하고 $\sum_{i=1}^{10} Z_i{}^2$가 $\chi^2(10)$이 된다.

카이제곱분포

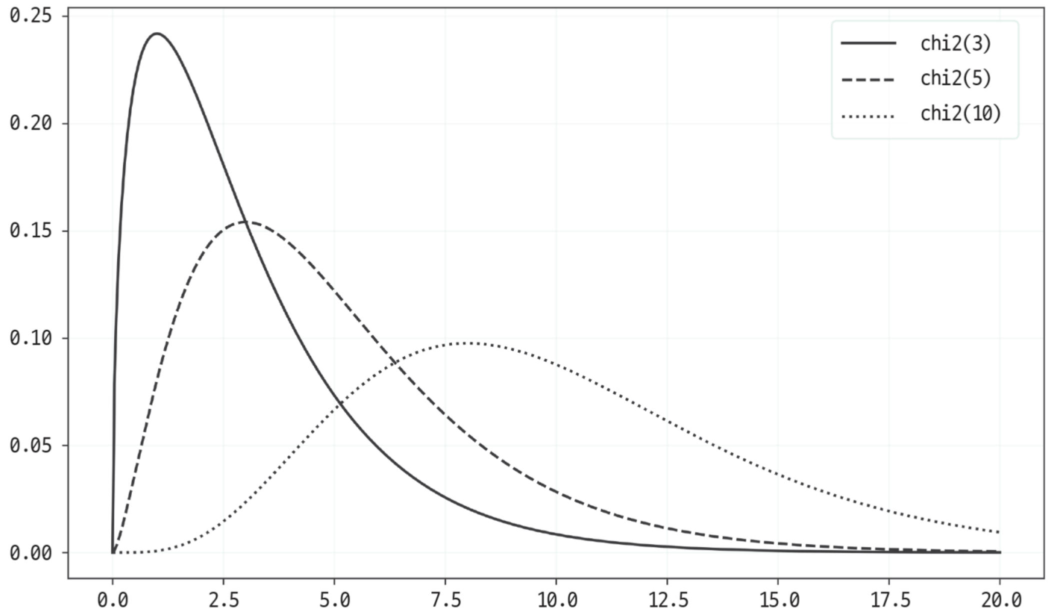
- 자유도 n에 따라 변화하는 카이제곱분포
- $n = 3, 5, 10$일 때

In [22] :

```
fig = plt.figure(figsize=(10, 6))
ax = fig.add_subplot(111)

xs = np.linspace(0, 20, 500)
for n, ls in zip([3, 5, 10], linestyles):
    rv = stats.chi2(n)
    ax.plot(xs, rv.pdf(xs),
            label=f'chi2({n})', ls=ls, color='gray')

ax.legend()
plt.show()
```



[그림 8-9] 다양한 카이제곱분포 SAMPLE CODE

- 좌우비대칭으로, 왼쪽으로 치우치고 오른쪽으로 넓어집니다.
- 자유도가 커지면 좌우대칭에 가까워집니다.
- 자유도의 값 가까이에 분포의 정점이 있습니다.

카이제곱분포

[표 8-3] 카이제곱분포의 정리

파라미터	n
취할 수 있는 값	음수가 아닌 실수
scipy.stats	chi2(n)

T분포

- 정규분포에서 모평균의 구간추정 등에 사용되는 확률분포

t 분포

확률변수 Z, Y는 서로 독립이고, Z는 표준정규분포 $N(0,1)$을, Y는 자유도가 n인 카이제곱분포 $\chi^2(n)$을 각각 따를 때,

$$t = \frac{Z}{\sqrt{Y/n}}$$

의 확률분포를 자유도가 n인 t 분포라고 합니다.

$N(\mu,\ \sigma^2/n)$

$N(0,\ 1) \qquad Z = \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}}$

$$t = \frac{\overline{X} - \mu}{\sqrt{s^2/n}}$$

$$t = \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \Big/ \sqrt{\frac{s^2}{\sigma^2}} = \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \Big/ \sqrt{\frac{(n-1)s^2}{\sigma^2} \Big/ (n-1)}$$

$$t = \frac{Z}{\sqrt{Y/n}}$$

T분포

- $Z \sim \mathrm{N}(0, 1)$과 $Y \sim \chi^2(10)$, $\frac{Z}{\sqrt{Y/10}}$에서 무작위추출

In [24] :

```
n = 10
rv1 = stats.norm()
rv2 = stats.chi2(n)

sample_size = int(1e6)
Z_sample = rv1.rvs(sample_size)
chi2_sample = rv2.rvs(sample_size)

t_sample = Z_sample / np.sqrt(chi2_sample/n)
```

- 자유도 10인 t분포 생성

- scipy.stats에서는 t분포를 따르는 확률변수를 t함수로 생성할 수 있고,인수에 자유도 지정

T분포

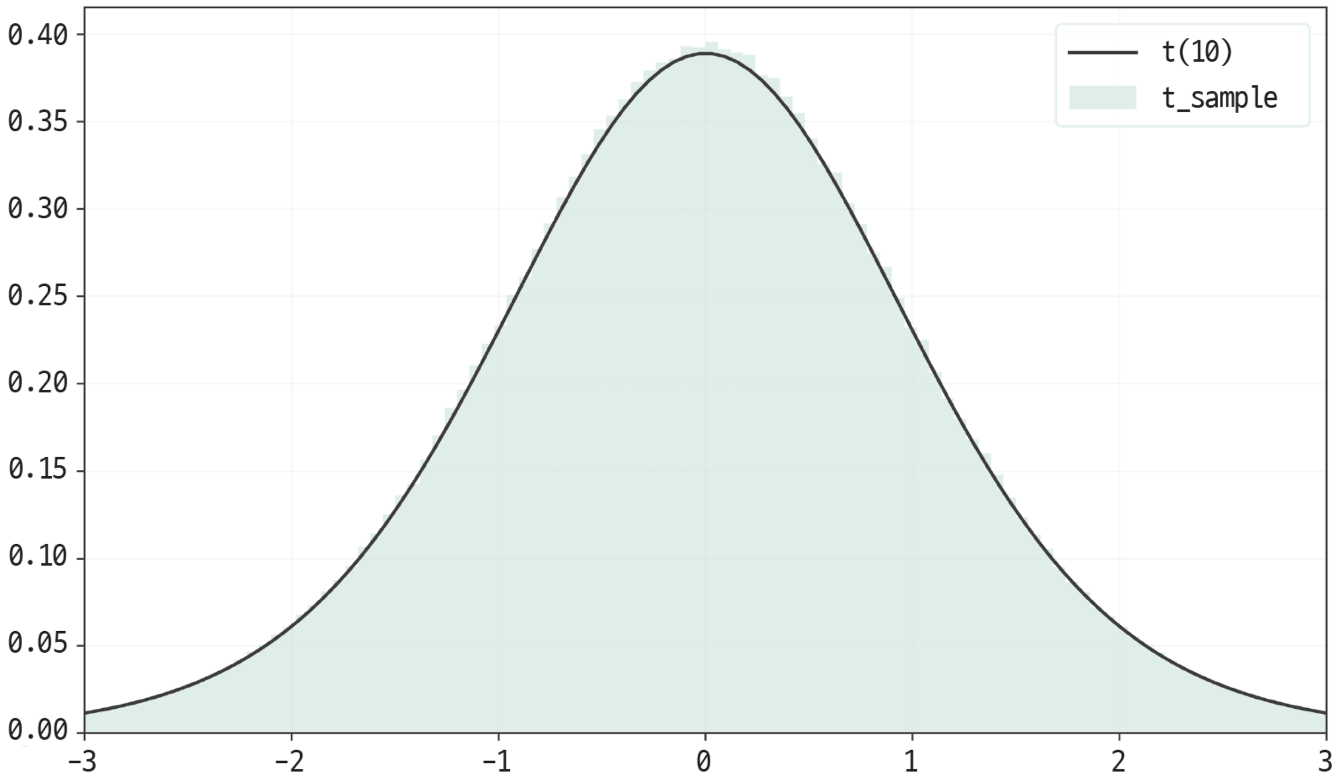
- $\frac{Z}{\sqrt{Y/10}}$에서 무작위추출한 표본 데이터의 히스토그램과 밀도함수

In [25] :

```
fig = plt.figure(figsize=(10, 6))
ax = fig.add_subplot(111)

rv = stats.t(n)
xs = np.linspace(-3, 3, 100)
ax.hist(t_sample, bins=100, range=(-3, 3),
        density=True, alpha=0.5, label='t_sample')
ax.plot(xs, rv.pdf(xs), label=f't({n})', color='gray')

ax.legend()
ax.set_xlim(-3, 3)
plt.show()
```



[그림 8-10] t 분포와 다른 분포의 관계

T분포

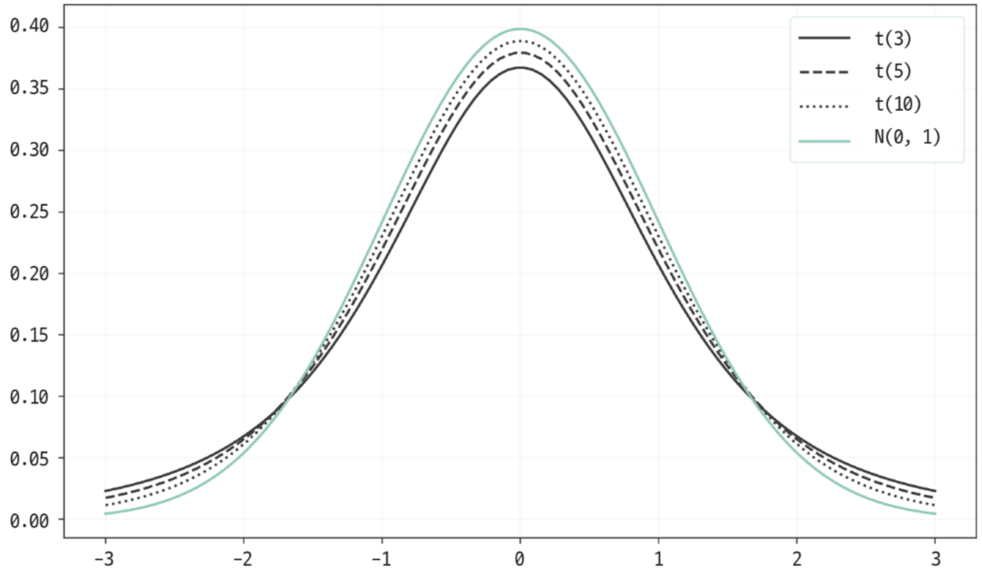
- 자유도 n에 따라 변화하는 t 분포
- $n = 3, 5, 10$일 때

In [26] :

```
fig = plt.figure(figsize=(10, 6))
ax = fig.add_subplot(111)

xs = np.linspace(-3, 3, 100)
for n, ls in zip([3, 5, 10], linestyles):
    rv = stats.t(n)
    ax.plot(xs, rv.pdf(xs),
            label=f't({n})', ls=ls, color='gray')
rv = stats.norm()
ax.plot(xs, rv.pdf(xs), label='N(0, 1)')

ax.legend()
plt.show()
```



[그림 8-11] 다양한 t 분포 SAMPLE CODE

- 좌우대칭인 분포입니다.
- 표준정규분포보다 양쪽 끝이 두껍습니다.
- 자유도가 커지면 표준정규분포에 가까워집니다.

$t_{0.05}(5)$

In [27] :

```
rv = stats.t(5)
rv.isf(0.05)
```

Out [27] :

```
2.015
```

F분포

- 분산분석 등에서 사용되는 확률분포

F 분포

확률변수 Y_1, Y_2는 서로 독립이고, 각각 $Y_1 \sim \chi^2(n_1)$, $Y_2 \sim \chi^2(n_2)$를 따를 때,

$$F = \frac{Y_1/n_1}{Y_2/n_2}$$

의 확률분포를 자유도가 n_1, n_2인 F 분포 $F(n_1, n_2)$라고 합니다.

$Y_1 \sim \chi^2(5)$와 $Y_2 \sim \chi^2(10)$을 사용하여 $\frac{Y_1/5}{Y_2/10}$에서 무작위추출을 수행

$\frac{Y_1/5}{Y_2/10}$은 정의에 따라 $F(5, 10)$

In [28] :

```
n1 = 5
n2 = 10
rv1 = stats.chi2(n1)
rv2 = stats.chi2(n2)

sample_size = int(1e6)
sample1 = rv1.rvs(sample_size)
sample2 = rv2.rvs(sample_size)

f_sample = (sample1/n1) / (sample2/n2)
```

F분포

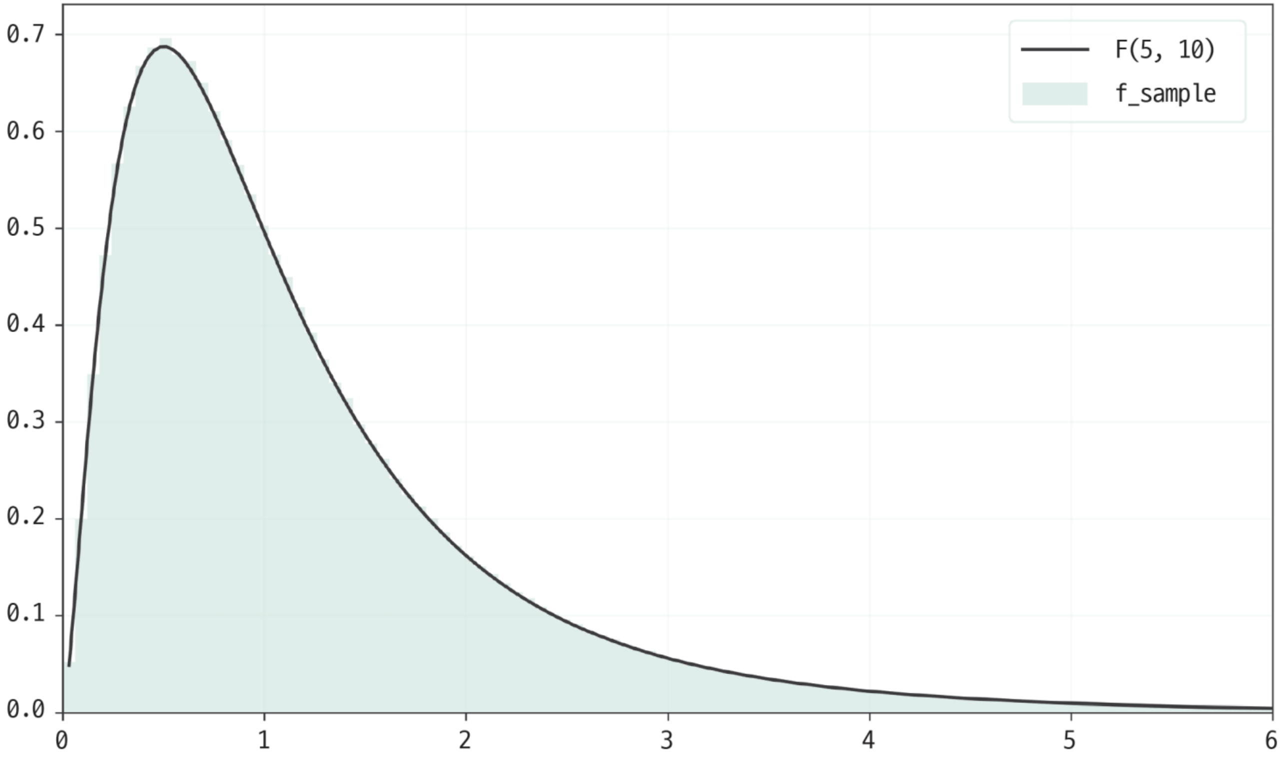
$\frac{Y_1/5}{Y_2/10}$에서 무작위추출한 표본 데이터의 히스토그램과 $F(5,10)$의 밀도함수

In [29] :

```
fig = plt.figure(figsize=(10, 6))
ax = fig.add_subplot(111)

rv = stats.f(n1, n2)
xs = np.linspace(0, 6, 200)[1:]
ax.hist(f_sample, bins=100, range=(0, 6),
        density=True, alpha=0.5, label='f_sample')
ax.plot(xs, rv.pdf(xs), label=f'F({n1}, {n2})', color='gray')

ax.legend()
ax.set_xlim(0, 6)
plt.show()
```



[그림 8-12] F 분포와 카이제곱분포의 관계

F분포

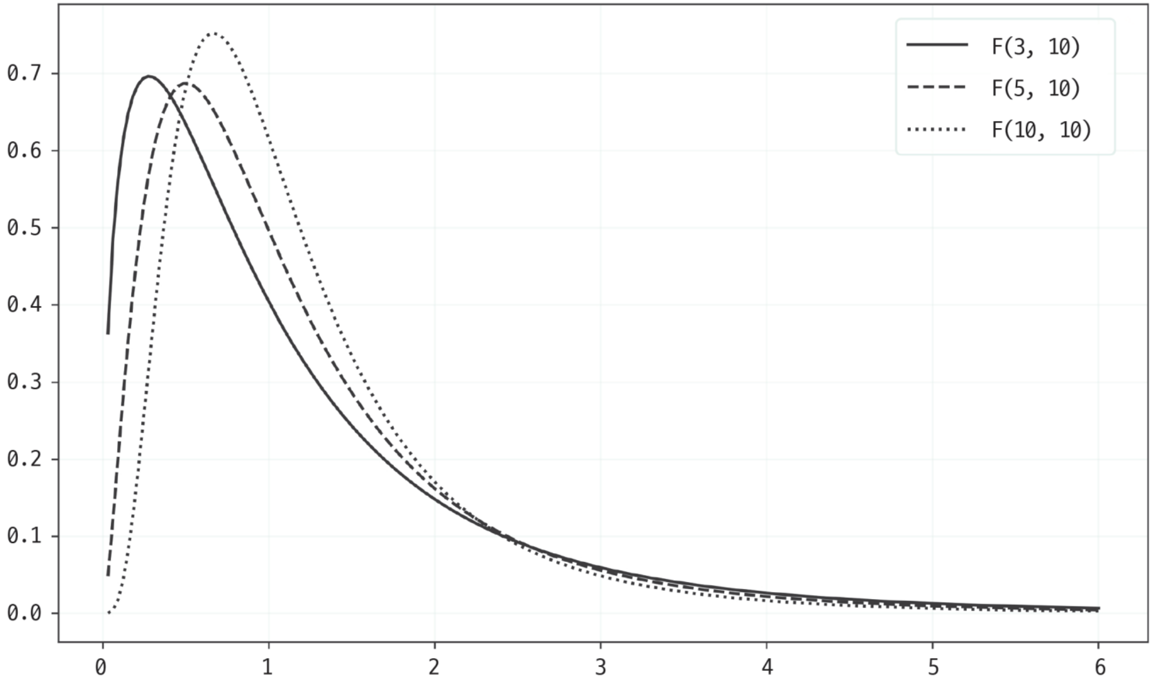
- 자유도 n에 따라 변화하는 F 분포
- $n_1 = 10$으로 고정하고, $n_1 = 3, 5, 10$일 때

In [30] :

```
fig = plt.figure(figsize=(10, 6))
ax = fig.add_subplot(111)

xs = np.linspace(0, 6, 200)[1:]
for n1, ls in zip([3, 5, 10], linestyles):
    rv = stats.f(n1, 10)
    ax.plot(xs, rv.pdf(xs),
            label=f'F({n1}, 10)', ls=ls, color='gray')

ax.legend()
plt.show()
```



[그림 8-13] 다양한 F 분포 SAMPLE CODE

- 좌우비대칭으로, 왼쪽으로 치우치고 오른쪽으로 넓어지는 분포입니다.
- 분포의 정점은 1에 가깝습니다.

F분포

[표 8-5] F 분포의 정리

파라미터	$n_1,\ n_2$
취할 수 있는 값	음수가 아닌 실수
scipy.stats	t$(n_1,\ n_2)$