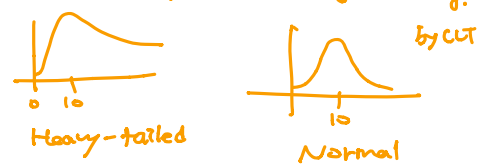
* Announcement
① Quiz 8 : Thu (11/5) 9:00 AM
~ Sat (11/7) 9:00 AM

Ch8
CLT : If $X_1, \dots, X_n$ from pop. w/ mean μ, SD σ,
$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, $S_n \sim N(n\mu, n\sigma^2)$
Standard normal cdf: $\Phi(z) = P(Z \le z)$ for $Z \sim N(0,1)$
Symmetry: $\Phi(-z) = 1 - \Phi(z)$
pop. has mean = 10, SD = 50, Non-neg. → Histogram of pop. → Histogram of avg. by CLT



STAT 88: Lecture 30

Contents

Warm up: (Exercise 9.5.12)

n = 400

12. A survey organization takes a simple random sample of 400 adults in a city. The annual incomes of the sampled people have an average of 68,000 dollars and an SD of 40,000 dollars.

$\bar{x} = 68{,}000$

$\sigma \approx s = 40{,}000$

a) Fill in the blank with one of the words "sample" or "city".

95% CI = $(\bar{x} \pm 2 \cdot \frac{\sigma}{\sqrt{n}}) = (68{,}000 \pm 2 \cdot \frac{40{,}000}{\sqrt{400}}) = (68{,}000 \pm 4000)$

The interval "68,000 dollars ± 4,000 dollars" is an approximate 95% confidence interval for the average annual income of adults in the ___City___.

b) Pick all of the correct options and justify your choices. More than one option may be correct.

The normal curve used in the construction of the confidence interval in Part **a** is the distribution of $\bar{x}$

(i) the incomes of the adults in the city X_1 from pop.

(ii) the incomes of the adults in the sample x

(iii) the averages of all possible simple random samples of 400 adults from the city

(iv) probabilities for how the average of a simple random sample of 400 adults from the city could come out

c) True or false (explain):

population mean. = annual avg incomes

The incomes of ~~approximately~~ 95% of the adults in the city are in the range 68,000 dollars ± 4,000 dollars.

False.

d) Fill in the blanks with the best choices you can make from the following set. You are welcome to use the same entry more than once.

- the average income of adults in the city
- the average income of adults in the sample
- 68,000 dollars
- 40,000 dollars
- 2,000 dollars

If you draw one adult at random from the city, that person's income has expectation equal to X_1 68,000 and SD approximately equal to 40,000.

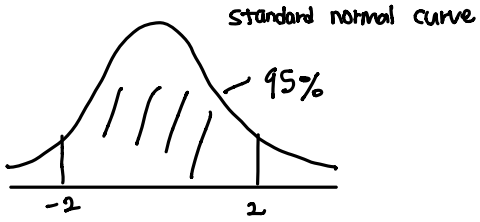
e) Fill in the blanks with the best choices you can make from the same set as in the previous part. Again you are welcome to use the same entry more than once.

If you draw a simple random sample of 400 adults from the city, the average income of the sampled adults has expectation equal to $\bar{X}$ 68,000 and SD approximately equal to 2,000.

Last time

Confidence interval for μ:

$X_1, \ldots, X_n$ are i.i.d. with mean μ and SD σ. Sampling distribution of $(\bar{X} - \mu)/\frac{\sigma}{\sqrt{n}}$:



We have

$$P\left(\bar{X} - 2\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 2\frac{\sigma}{\sqrt{n}}\right) \approx 0.95.$$

A 95% CI has a 95% chance of containing μ. Your particular 95% CI either does or does not contain μ since it is a fixed interval and μ is fixed.

Computing a 95% CI for μ_0 is equivalent to conducting a level 5% hypothesis testing for $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$.

95% CI

$$\mu_0 \text{ lies in } \bar{X} \pm 2\frac{\sigma}{\sqrt{n}}$$

$$\iff \bar{X} \text{ lies in } \mu_0 \pm 2\frac{\sigma}{\sqrt{n}}.$$

Acceptance region at level 5%

(Only if you are interested)

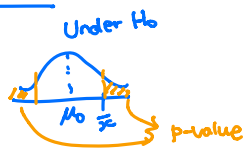
Acceptance region for $H_0: \mu = \mu_0$ vs $H_A: \mu \neq \mu_0$

Test statistic = $\bar{x}$, obs-value = $\bar{x}$

p-value $= 2 \cdot P(\bar{x} - \mu_0 \geq |\bar{x} - \mu_0|)$

$= 2P\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq \left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right|\right)$

$= 2 \cdot (1 - \Phi(\left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right|))$



Notice that $\bar{X}$ is in the accentance region of H_0 at level 5%, exactly when μ_0 is in your 95% CI.

Accept H_0 iff

p-value > 0.05

$\iff 2 \cdot (1 - \Phi(\left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right|)) > 0.05$

$\iff \Phi(\left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right|) < 0.975$

$\iff \left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right| < \Phi^{-1}(0.975) = 2$

$\iff \bar{x}$ lies in $\mu_0 \pm 2\frac{\sigma}{\sqrt{n}}$

Example: Suppose an observed instance of $\bar{X} = 27.23$, and $\sigma = 5.8$, and $n = 1174$. Test the hypothesis

$$H_0 : \mu = 27.4 \text{ vs } H_A : \mu \neq 27.4$$

at level 5%.

95% CI: $(\bar{x} \pm 2\frac{\sigma}{\sqrt{n}})$

$= (27.23 \pm 2 \cdot \frac{5.8}{\sqrt{1174}})$

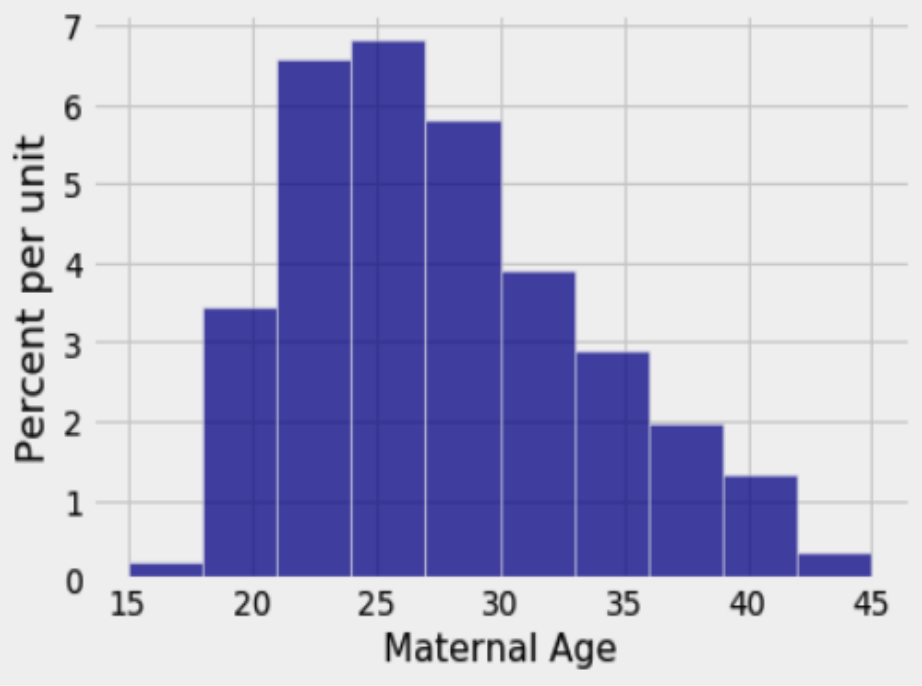
$= (26.59, 27.57)$

27.4 is in your CI, so we accept H_0 at level 5%.

9.4. Confidence Intervals: Interpretation (Continued)

Comparison with the Bootstrap The interpretation of CI is the same as in Data 8.

Example: Here is a distribution of 1174 maternal ages (years) from a random sample.



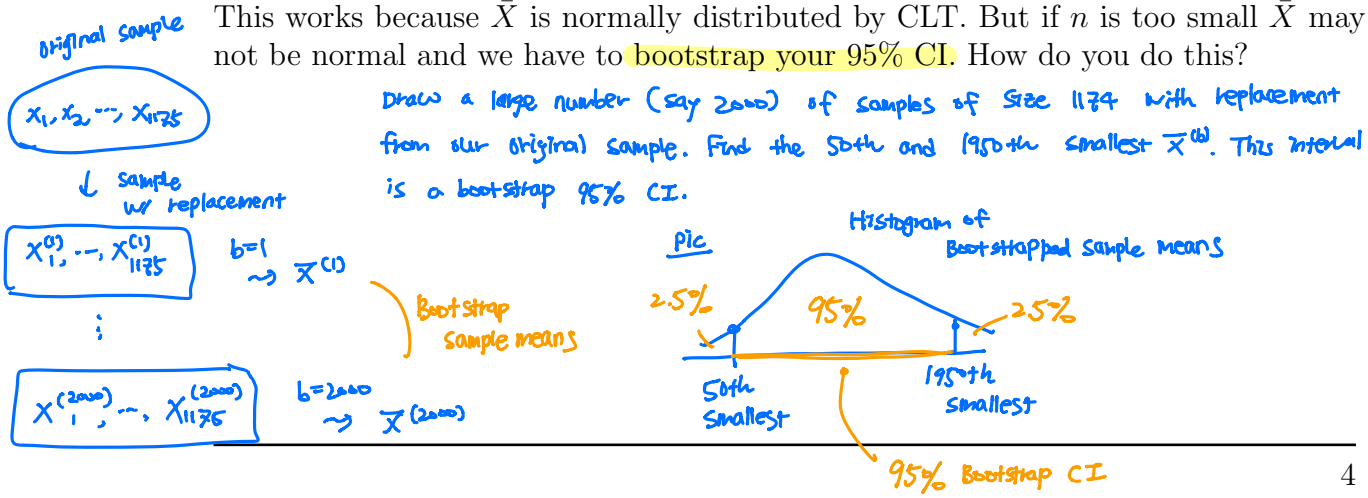
$\approx \bar{x}$

The sample mean is about 27.23 years and the sample SD is about 5.8 years. Find the approximate 95% CI of μ and interpret.

(26.89, 27.57)

↝ There is a 95% chance that a CI contains μ, but my CI either does or does not contain μ

This works because $\bar{X}$ is normally distributed by CLT. But if n is too small $\bar{X}$ may not be normal and we have to bootstrap your 95% CI. How do you do this?



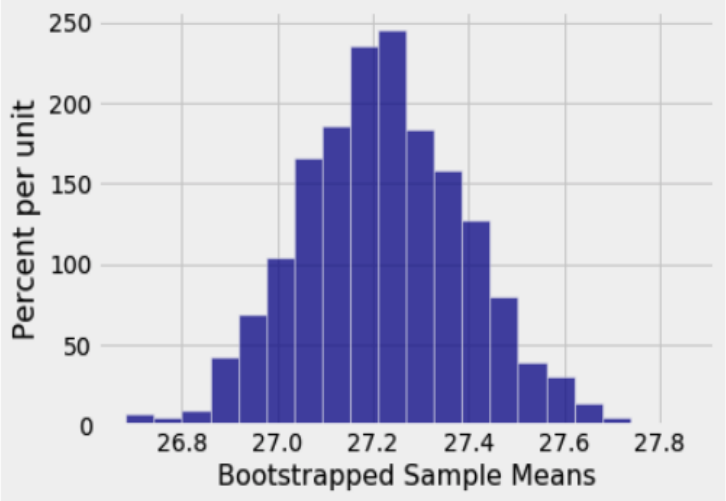
```
def one_resampled_mean():
    return np.average(births.sample().column('Maternal Age'))
```

We then called this function repeatedly to create an array of 2,000 bootstrap means:

```
means = make_array()

for i in np.arange(2000):
    means = np.append(means, one_resampled_mean())

Table().with_column('Bootstrapped Sample Means', means).hist(0, bins=2
```



Finally, we found the "middle 95%" of the bootstrapped means. That was our empirical bootstrap 95% confidence interval for the population mean.

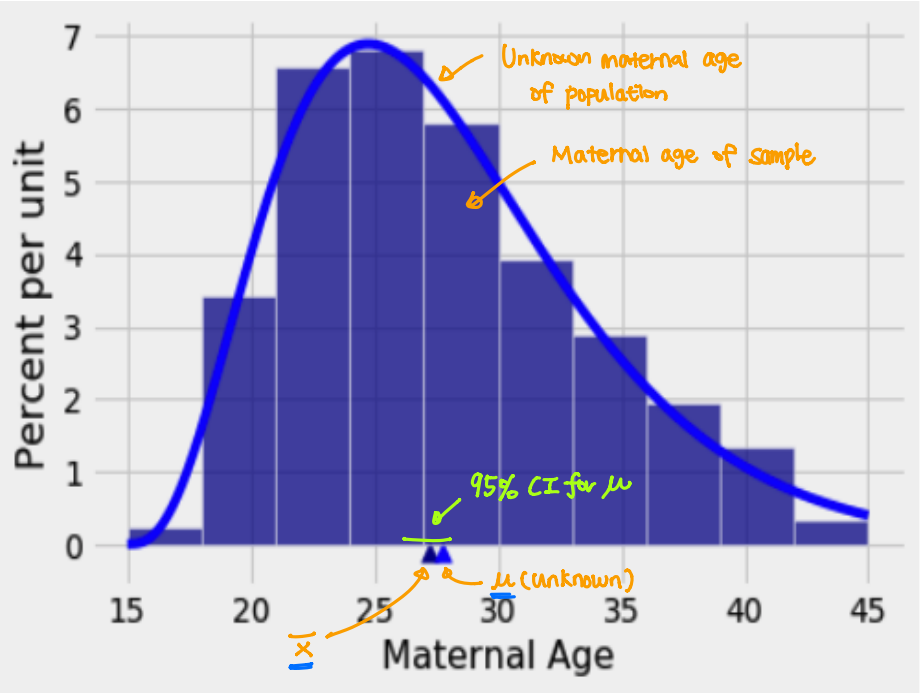
```
left = percentile(2.5, means)
right = percentile(97.5, means)
left, right
```

```
(26.89182282793867, 27.572402044293014)
```

close to (26.69, 27.57)

What the Confidence Interval Measures

CI is an interval of estimates of μ:



$\bar{X}$ is close to μ. On average it is $\text{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ away from μ. Is there a 95% chance that maternal ages are between $(26.89, 27.57)$?

no, maternal ages are in a much wider range of values

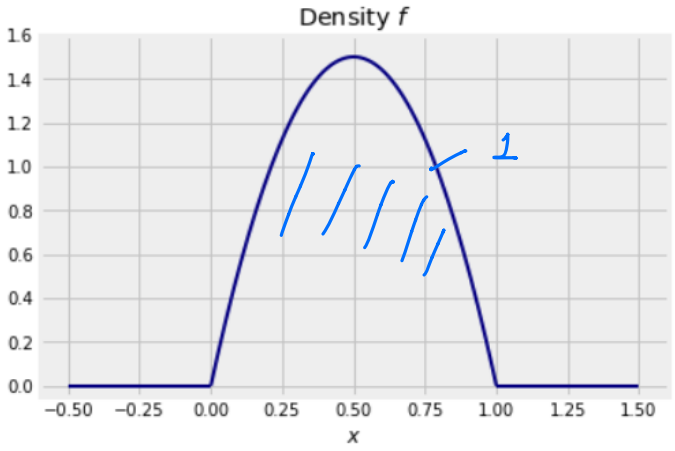
10.1. Density

A density, f, is a nonnegative function such that

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

Example: Let

$$f(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 6x(1-x) & \text{if } 0 < x < 1 \\ 0 & \text{if } x \geq 1 \end{cases}$$



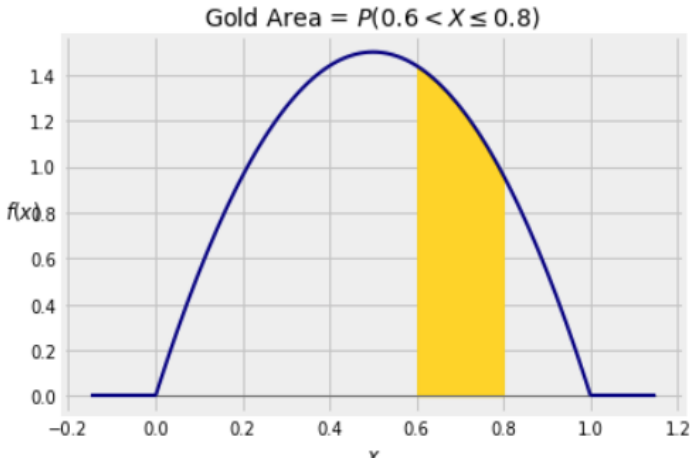
Density is Not the Same as Probability

$f(x)$ can be > 1 (e.g. $f(0.5) = 1.5$ in the example above) so $f(x)$ is not a probability.

Areas are Probabilities

A random variable X is said to have density f if for every pair $a < b$,

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



Here

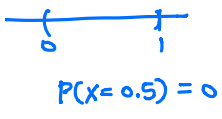
$$P(0.6 < X \leq 0.8) = \int_{0.6}^{0.8} 6x(1-x)dx.$$

No Probability at Any Single Point

A wonderful aspect of a random variable that has a density, like the random variable X above, is that there is no chance of hitting a possible value exactly. In fact,

$$P(X = b) = \int_b^b f(x)dx = 0.$$

This implies that we can ignore endpoints:



$$\begin{aligned} P(a < X \leq b) &= P(a < X < b) \\ &= P(a \leq X < b) \\ &= P(a \leq X \leq b). \end{aligned}$$

Cumulative Distribution Function (CDF)

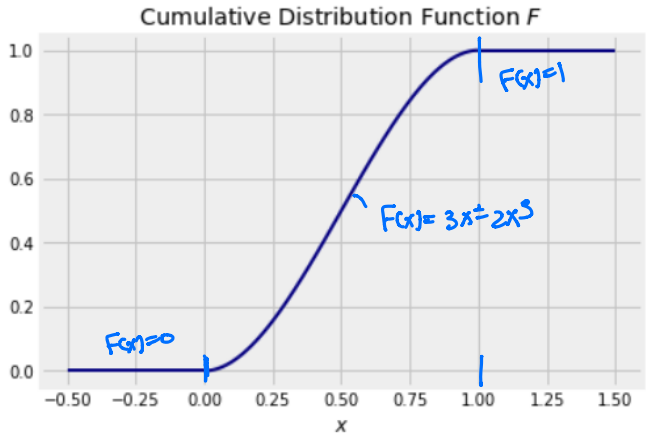
Discrete case
$F(x) = \sum_{s=-\infty}^{x} P(X=s)$

The CDF of X is

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(s)ds.$$

Example: For $f(x) = 6x(1-x)$ for $0 < x < 1$,

$$F(x) = \int_{-\infty}^{x} f(s)ds = \int_{0}^{x} f(s)ds = \int_{0}^{x} 6s(1-s)ds = 3x^2 - 2x^3.$$



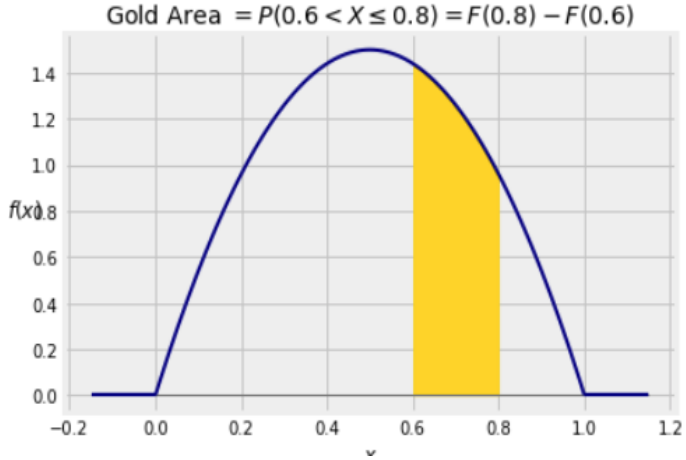
As before, the cdf can be used to find probabilities of intervals. For every pair $a < b$,

$$P(a < X \le b) = F(b) - F(a).$$

$= \int_a^b f(x)dx = \int_{-\infty}^{b} f(x)\,dx - \int_{-\infty}^{a} f(x)\,dx = F(b) - F(a)$

Example: Since $F(x) = 3x^2 - 2x^3$ for $0 < x < 1$,

$$P(0.6 < X \le 0.8) = F(0.8) - F(0.6) = 0.248.$$



By the Fundamental Theorem of Calculus, the density and cdf can be derived from each other:

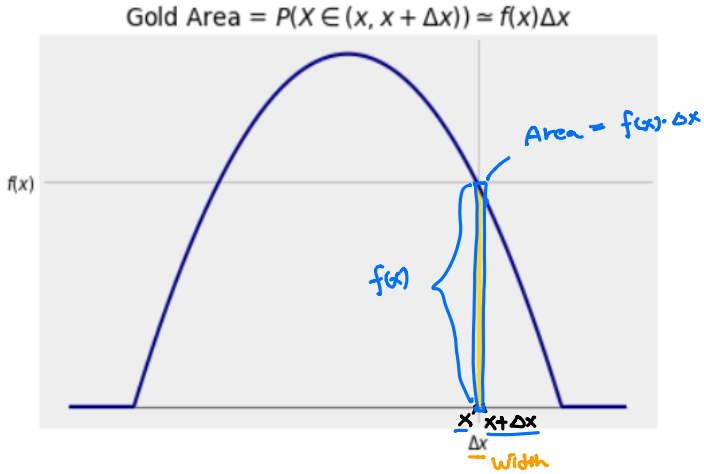
$$F(x) = \int_{-\infty}^{x} f(s)ds \quad \text{and} \quad f(x) = \frac{d}{dx}F(x).$$

The Meaning of Density

Consider a very small width $\Delta x > 0$. Although $P(X = x) = 0$, $P(x \le X \le x + \Delta x)$ is positive as long as $f(x) > 0$. Note that

$$\begin{aligned} P(x \le X \le x + \Delta x) &= \int_{x}^{x+\Delta x} f(s)ds \\ &\approx \int_{x}^{x+\Delta x} f(x)ds \\ &= f(x)\Delta x. \end{aligned}$$

↓ On $s \in [x, x+\Delta x]$, $f(s) \approx f(x)$



Thus

$$f(x) \approx \frac{P(x \le X \le x + \Delta x)}{\Delta x}.$$

Probability per unit length

In other words, f measures the chance that X is in a tiny interval near x, relative to the width of the interval.

10.2. Expectation and Variance

Discrete case

If a random variable X has density f,

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx. \qquad E(X) = \sum_{x=-\infty}^{\infty} x P(X=x)$$

For any function g, we have

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f(x) dx. \qquad E(g(X)) = \sum_{x=-\infty}^{\infty} g(x) P(X=x)$$

In particular, this shows

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx. \qquad E(X^2) = \sum_{x=-\infty}^{\infty} x^2 P(X=x)$$

The variance and SD are then given by

$$\mathrm{Var}(X) = E(X^2) - (EX)^2 \text{ and } \mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}.$$

Properties of expectation and variance are the same as discrete case.

- $E(aX + b) = aE(X) + b$ and $\mathrm{SD}(aX + b) = |a|\mathrm{SD}(X)$.
- $E(X + Y) = E(X) + E(Y)$.
- If X and Y are independent,

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B).$$

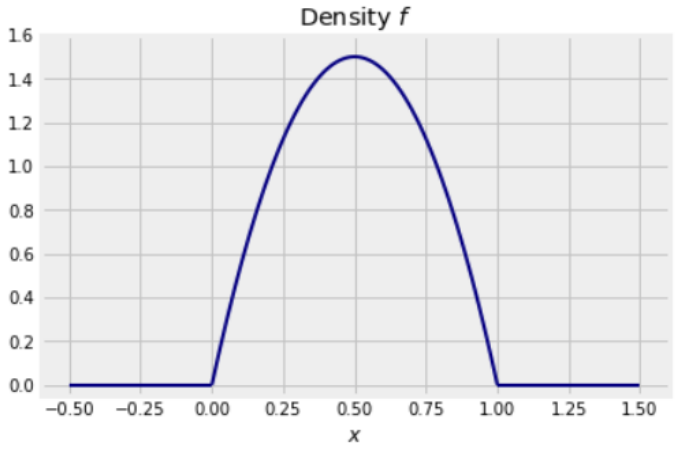
- If X and Y are independent,

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

The Central Limit Theorem holds too: if $X_1, \ldots, X_n$ are i.i.d. then for large n the distributions of $S_n = \sum_{i=1}^{n} X_i$ and $\bar{X} = S_n/n$ are approximately normal.

Example: Let

$$f(x) = \begin{cases} 0 & \text{if } x \le 0 \\ 6x(1-x) & \text{if } 0 < x < 1 \\ 0 & \text{if } x \ge 1 \end{cases}$$



Find $E(X)$ and $\mathrm{Var}(X)$.

$E(X) = 0.5$ (by symmetry)

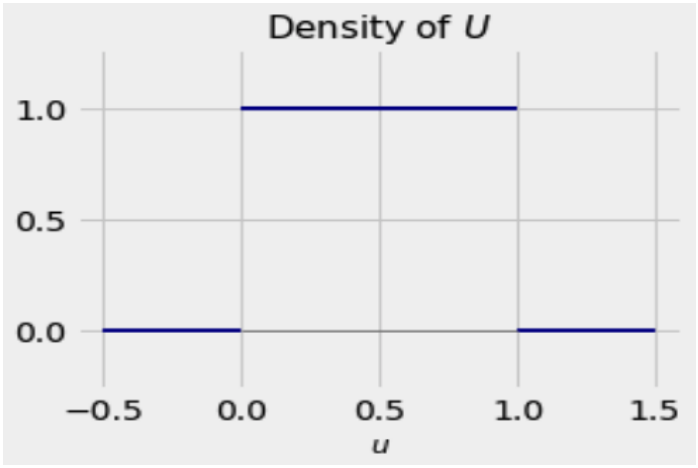
$E(X^2) = \int_0^1 x^2 f(x)\,dx = \int_0^1 x^2 6x(1-x)\,dx = 0.3$

$\rightsquigarrow \mathrm{Var}(X) = E(X^2) - (EX)^2$
$= 0.3 - 0.5^2$
$= 0.05$

Uniform$(0,1)$ **Distribution**

A random variable U has the uniform distribution on the unit interval $(0,1)$ if

$$f(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$



For $0 < u_1 < u_2 < 1$, what is $P(u_1 < U < u_2)$?

Find and sketch the cdf of f, $F(x)$:

Find $E(U)$ and $\text{Var}(U)$.

Uniform(a, b) **Distribution**

For $a < b$, the random variable X has the uniform distribution on the interval (a, b) if

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

For $a < x_1 < x_2 < b$, what is $P(x_1 < X < x_2)$?

What function $X = g(U)$ stretches and shifts?

Find $E(X)$ and $\text{Var}(X)$.

Example: (Exercise 10.5.2) A class starts at 3:10 p.m. Seven students in the class arrive at random times $T_1, T_2, \ldots, T_7$ that are i.i.d. with the uniform distribution on the interval 3:07 to 3:12.

(a) Find $E(T_1)$.

(b) What is the chance that all seven students arrive before 3:10?

(c) Let $X = \max(T_1, T_2, \ldots, T_7)$ be the time when the last of the seven students arrives. Find the cdf of X.