Statistics

Readings

Intro to stats

Estimation

Continuous random variables

Asymptotics

Readings

Readings

- Appendix B:
  - B.1, “Continuous random variables”
  - B.5:
    - “The Normal distribution”
    - “The Standard Normal Distribution”
    - “Additional Properties of the Normal Distribution”
- Appendix C:
  - C.1-C.3
  - Skip C.4
  - Self-study C.5
  - Skip C.6
- Other lecture material
  - Stochastic convergence
  - Chebyshev inequality
  - ...

Intro to stats

Stats v PT

- **Probability theory** formalizes uncertainty
- **Statistics** uses probability theory, with **sampling**, to learn about the world

Sampling framework

- Population
- Model
- Data

whiteboard:3.1

Sampling model

We assume that the data was obtained by **random sampling**

- Model of how data were obtained
- Formally, we have a random sample

$$(Y_1, \cdots, Y_n)$$

if

$$Y_1, \cdots, Y_n$$

are **i**ndependent and **i**dentically **d**istributed (i.i.d)

Implications (pt)

For each $i = 1, \cdots, n$:

- Y_i is a RV
- $Y_i \perp Y_j$ for all $j \neq i$
- $E(Y_i) = E(Y_1)$
- $Var(Y_i) = Var(Y_1)$

Implications (stats)

Random sampling leads to **sampling variability** in statistical procedures.

- Sampling process is *random*
- Resulting data (random sample) is *random*
- Functions of the random sample are *random*

Estimation

Estimator

An estimator is a function of the *data*:

$$\hat{\theta} = \hat{\theta}\left(Y_1, \cdots, Y_n\right)$$

- **Random**
- Informally: guess for parameter based on data
- Cannot depend on parameter

Estimator: examples

whiteboard:3.2

Properties

- What makes a good estimator?
- Judge estimator by properties of its **sampling distribution**
  - Expectation
  - Variance
  - Later: behavior as $n \to \infty$

Property 1: Unbiasedness

An estimator $\hat{\theta}$ for a parameter θ is **unbiased** if

$$E\left(\hat{\theta}\right) = \theta$$

for each possible value of θ.

- Gets it right *on average*
  - Averaged across *what*?
- Many estimators are unbiased

Property 2: Efficiency

Relative efficiency. If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators for θ, then $\hat{\theta}_1$ is efficient relative to $\hat{\theta}_2$ when

$$\text{Var}\left(\hat{\theta}_1\right) \leq \text{Var}\left(\hat{\theta}_2\right).$$

- $\text{Var}\left(\hat{\theta}_1\right)$ is the *sampling variance* of $\hat{\theta}_1$
- smaller confidence intervals

Properties: Examples

whiteboard:3.3

Simulation study

Interested in estimating $E(X) = \theta$, where X is the outcome of a dice throw. Approach:

- Obtain a random sample $(Y_1, \cdots, Y_n)$
- Compute a few estimators:
  - $\bar{Y} = \frac{1}{n} \sum_i Y_i$, the sample mean.
  - Use the first observation: Y_1
  - Even sample mean: $\frac{1}{2n} (Y_2 + Y_4 + ... + Y_n)$
  - Estimate 3.5.

Example (ctd)

Throw 5 dice and look at outcomes of the estimators.

```
n <- 5
x <- sample(1:6,size=n,replace=TRUE)
theta_1_hat <- mean(x)
theta_2_hat <- x[1]
theta_3_hat <- mean(x[seq(from=2,by=2,to=n-n%%2)])
theta_4_hat <- 3.5
c(theta_1_hat,theta_2_hat,theta_3_hat,theta_4_hat)
```

```
## [1] 2.6 1.0 3.0 3.5
```

Example (ctd)

What does this function do?

```
dice_estimators <- function(n,S) {
  thetas <- matrix(0,nrow=S,ncol=3)
  for(s in 1:S){
    x <- sample(1:6,size=n,replace=TRUE)
    thetas[s,1] <- mean(x)
    thetas[s,2] <- x[1]
    thetas[s,3] <- mean(x[seq(from=2,by=2,to=n-n%%2)])
  }
}
```

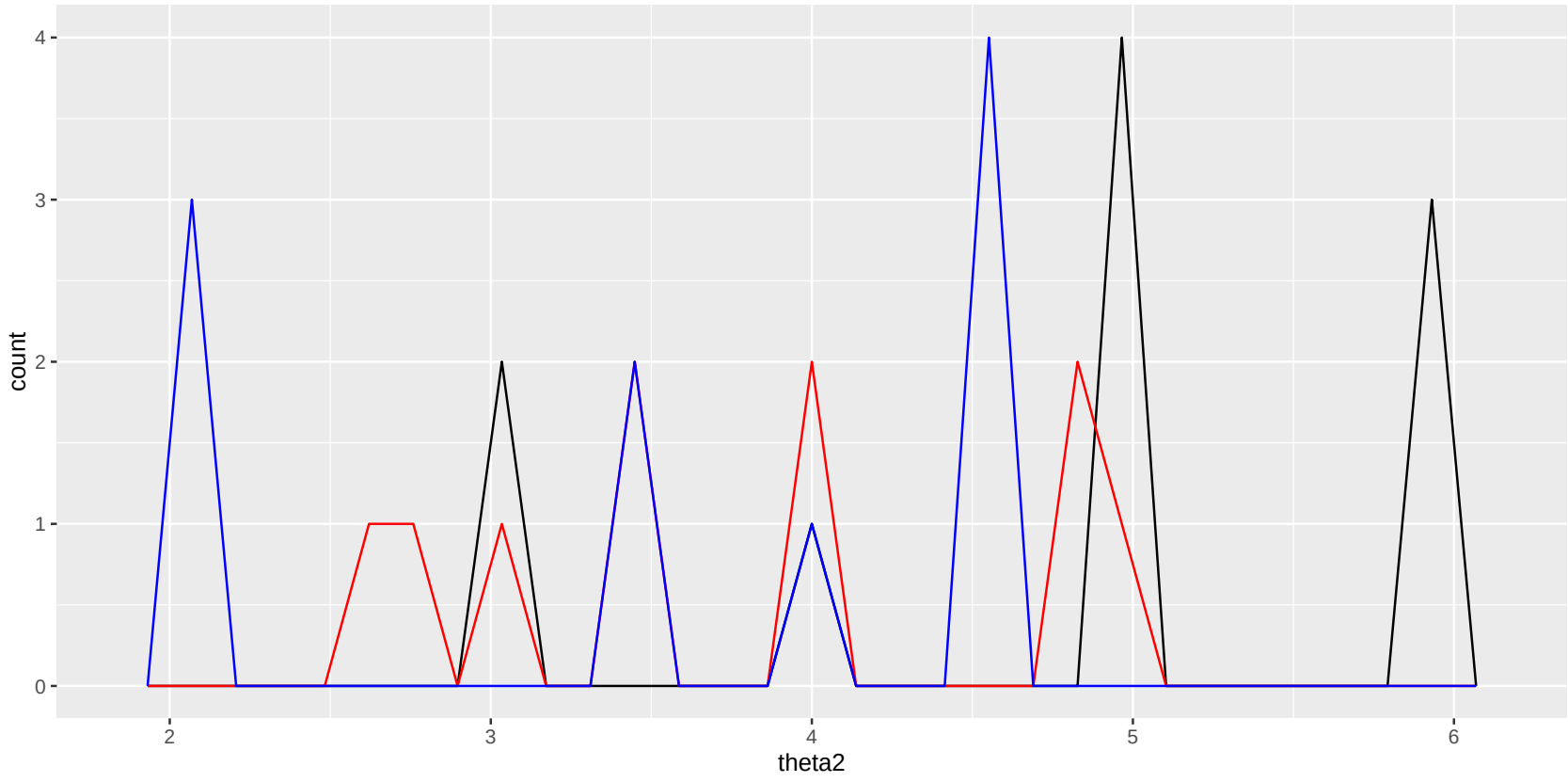
Example (plotting)

Add some plotting functionality.

```
library(reshape)
library(ggplot2)
dice_estimators <- function(n,S) {
  thetas <- matrix(0,nrow=S,ncol=3)
  for(s in 1:S){
    x <- sample(1:6,size=n,replace=TRUE)
    thetas[s,1] <- mean(x)
    thetas[s,2] <- x[1]
    thetas[s,3] <- mean(x[seq(from=2,by=2,to=n-n%%2)])
  }
  thetas <- data.frame(thetas)
  colnames(thetas) <- c("theta1","theta2","theta3")
  ggplot(data=thetas) + geom_freqpoly(aes(x=theta2)) +
              geom_freqpoly(aes(x=theta1),col="red") +
             geom_freqpoly(aes(x=theta3),col="blue")
}
```

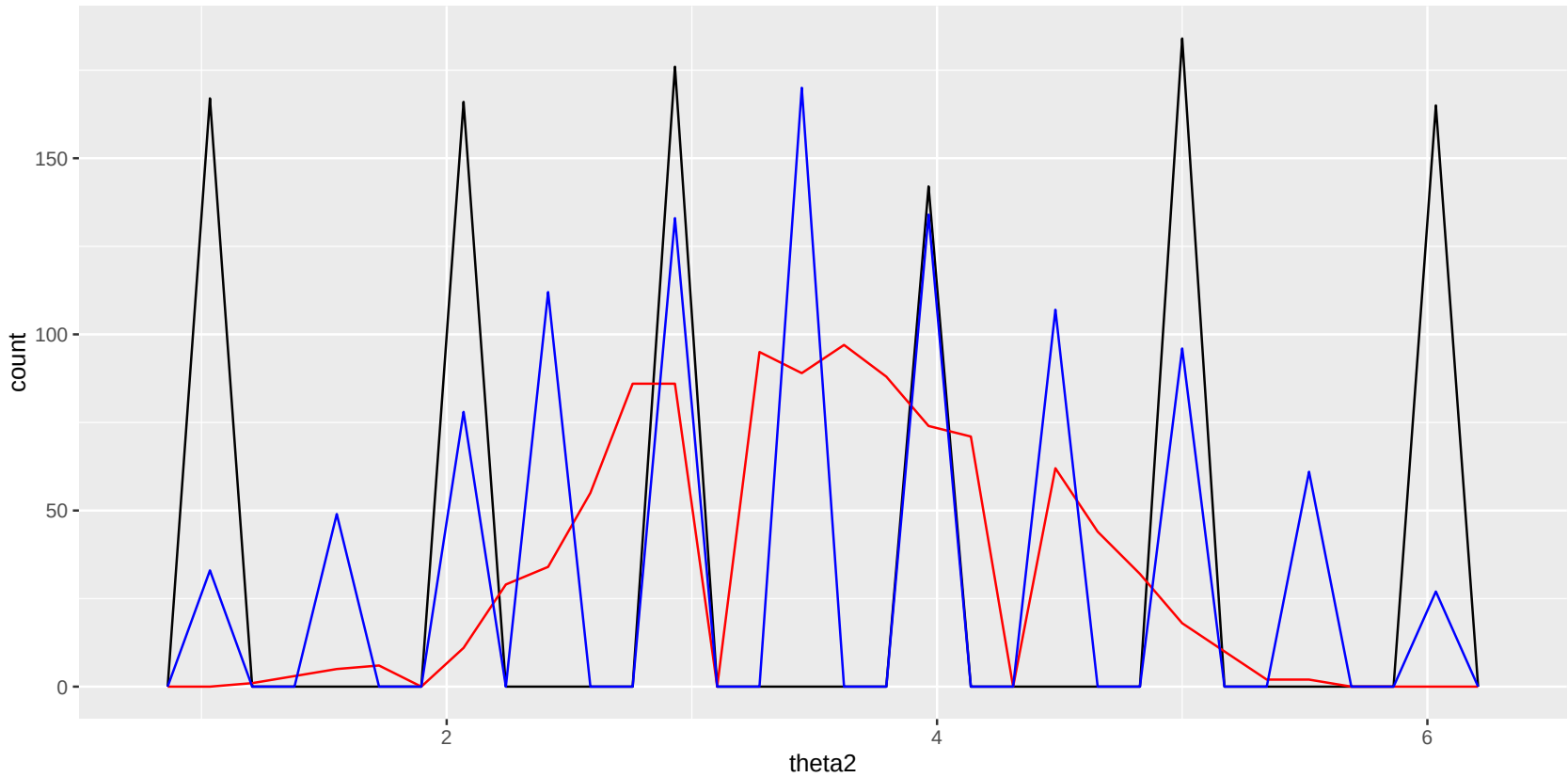
Run it

```
dice_estimators(5,10)
```



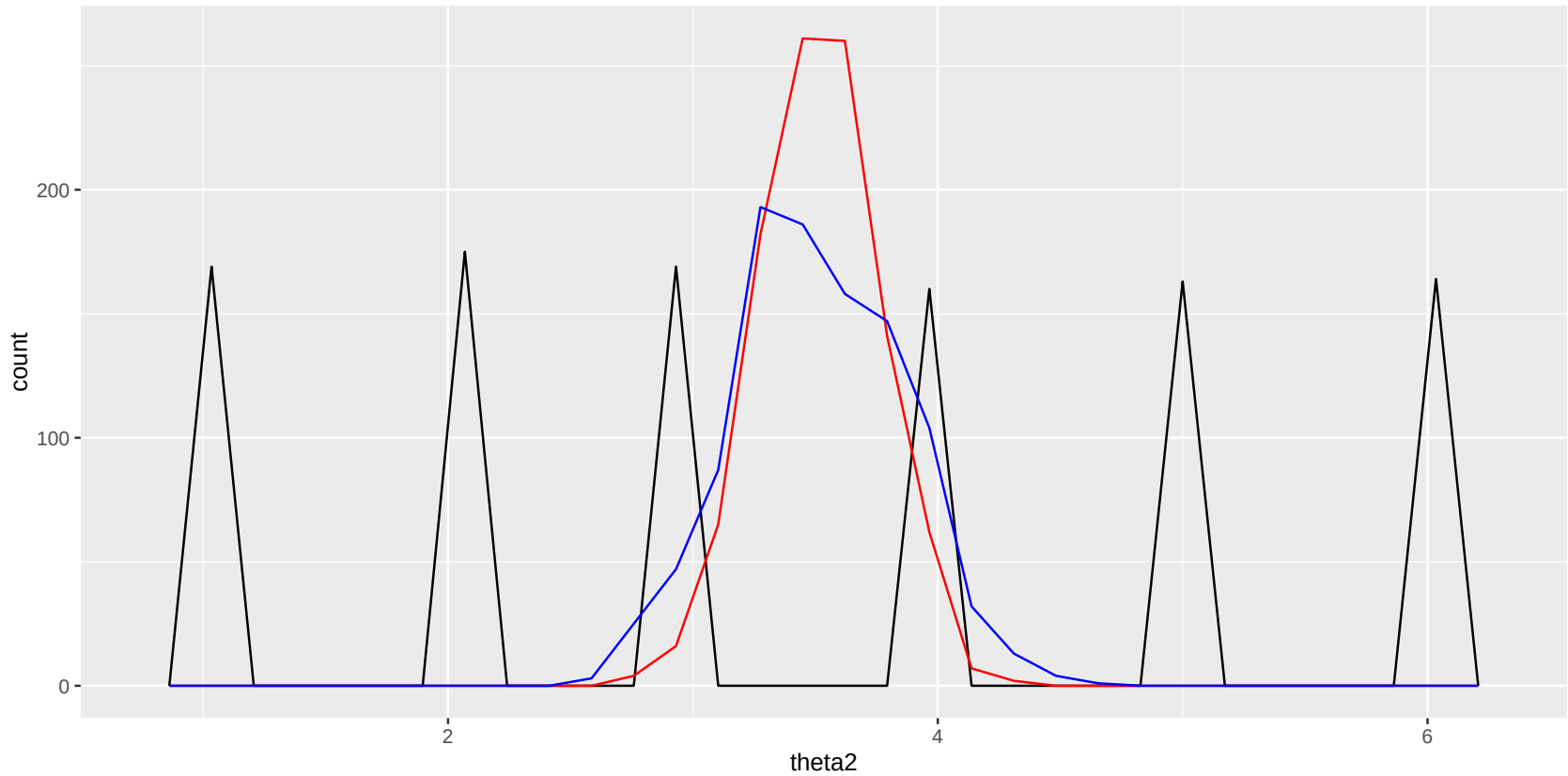
Run it

```
dice_estimators(5,1000)
```



Run it

```
dice_estimators(50,1000)
```

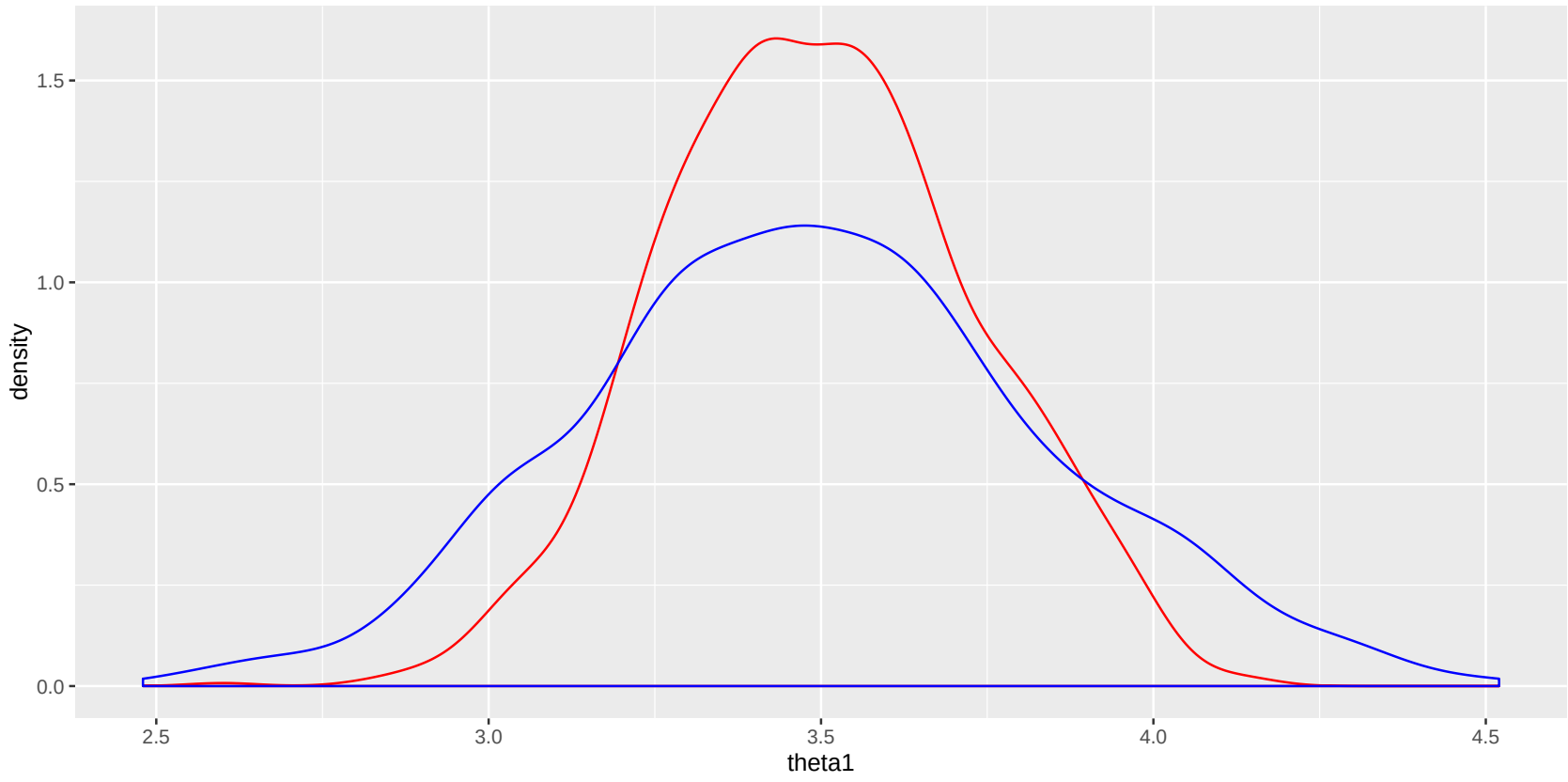


Smooth histograms

```
library(reshape)
library(ggplot2)
dice_estimators <- function(n,S) {
  thetas <- matrix(0,nrow=S,ncol=3)
  for(s in 1:S){
    x <- sample(1:6,size=n,replace=TRUE)
    thetas[s,1] <- mean(x)
    thetas[s,2] <- x[1]
    thetas[s,3] <- mean(x[seq(from=2,by=2,to=n-n%%2)])
  }
  thetas <- data.frame(thetas)
  colnames(thetas) <- c("theta1","theta2","theta3")
  ggplot(data=thetas) +
    geom_density(aes(x=theta1),col="red") +
    geom_density(aes(x=theta3),col="blue")
}
```

Run it

```
dice_estimators(50,1000)
```



Continuous random variables

Continuous v discrete

- Discrete RV: finite number of outcomes
- Some RVs are better modelled as continous
- For a continous RV X, $P(X = x) = 0$
  - Example: *guessing game*

Continuous RVs

A continous RV is described by

- its outcome space (real line, [0,1])
- its **cumulative density function**

$$F(x) = P(X \leq x)$$

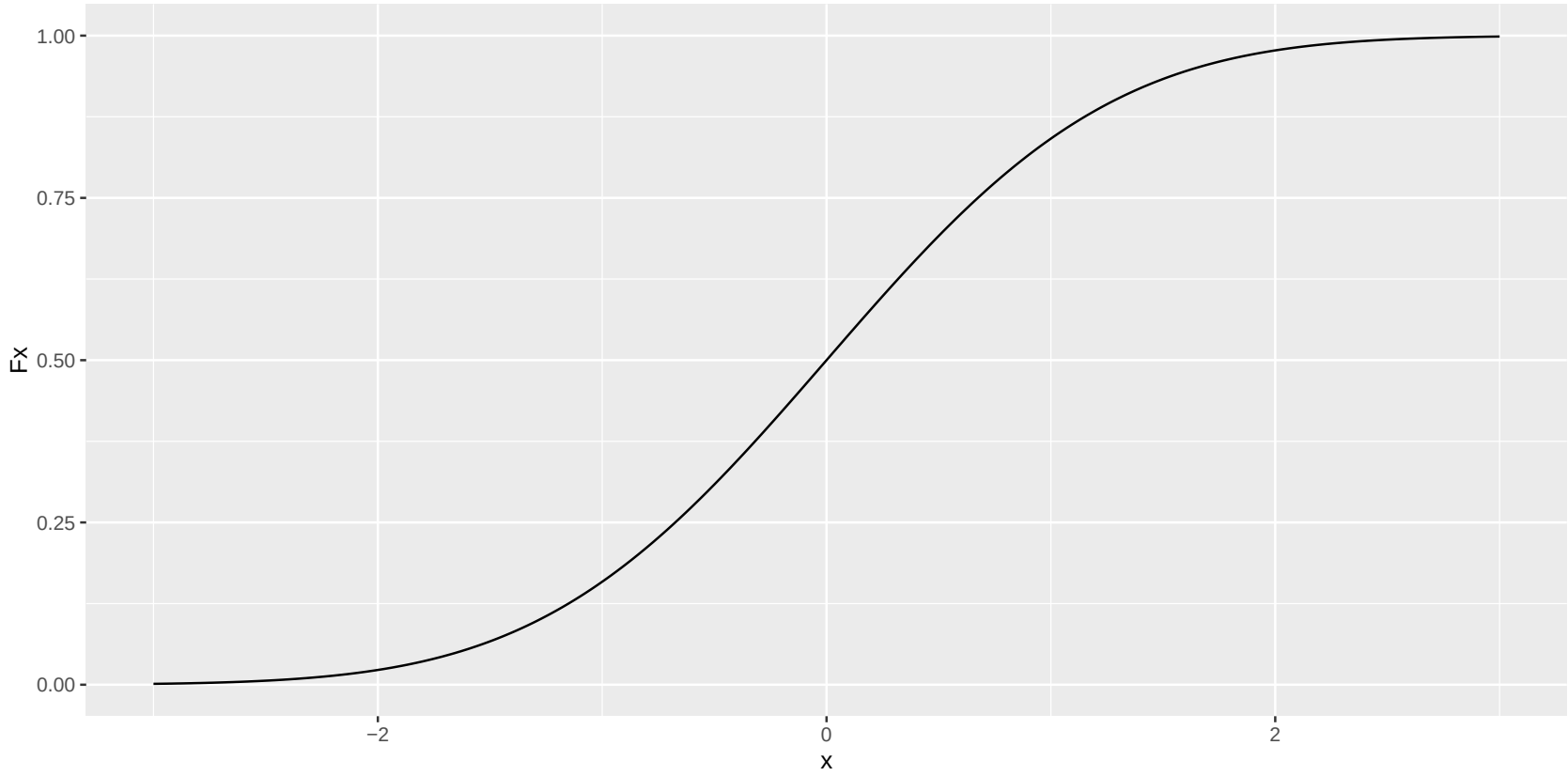
and its derivative

$$f(x) = \frac{\partial F}{\partial x},$$

the **probability density function**

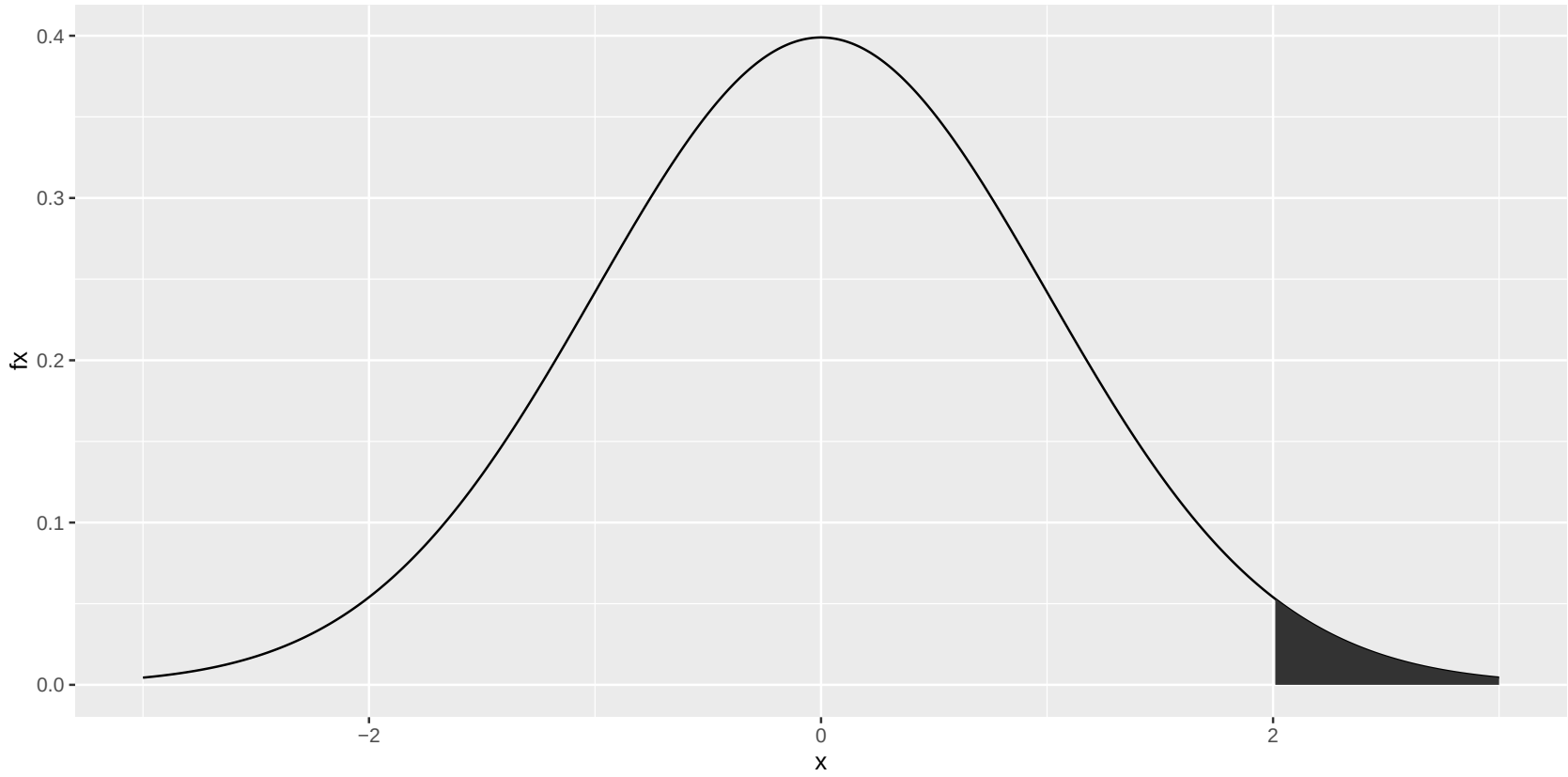
Normal RV

```
x <- seq(from=-3,to=3,by=0.01)
Fx <- pnorm(x)
qplot(x=x,y=Fx,geom="line")
```



Normal RV: pdf

```
x <- seq(from=-3,to=3,by=0.01)
fx <- dnorm(x)
qplot(x=x,y=fx,geom="line") + geom_area(aes(x=x[x>2],y=fx[x
```



Continuous RVs: summary

Intuition for probability theory concepts are the same. Continuous RVs matter for this course because:

- asymptotics
- inference (self-study)

Asymptotics

Big data

- this course: large data sets
- use asymptotics to study behavior of estimators

Asymptotics: What happens to $\hat{\theta}_n$ as $n \to \infty$?

Reminder: convergence of sequences

A deterministic sequence $\{x_i\}_{i=1}^{\infty}$ converges to its limit c if for each $\epsilon > 0$ there exists an N such that, for each $i > N$ we have $|x_i - c| < \epsilon$

Stochastic sequences

It is more challenging to think about stochastic sequences.

whiteboard:3.4

Convergence of stochastic sequences

We look at two concepts:

1. convergence in probability (*heads-up: LLN*)
2. convergence in distribution (*heads-up: CLT*)

Convergence in probability

Consider a stochastic sequences X_n. We say that X_n converges to θ **in probability** if for every $\epsilon > 0$, the *deterministic* sequence

$$P(|X_n - \theta| > \epsilon) \to 0$$

- Short notation: $\text{plim} X_n = \theta$.
- Other notation: $X_n \xrightarrow{p} \theta$
- An estimator $\hat{\theta}$ is **consistent** if its probability limit is θ

Note (advanced)

Writing out the above definition yields:

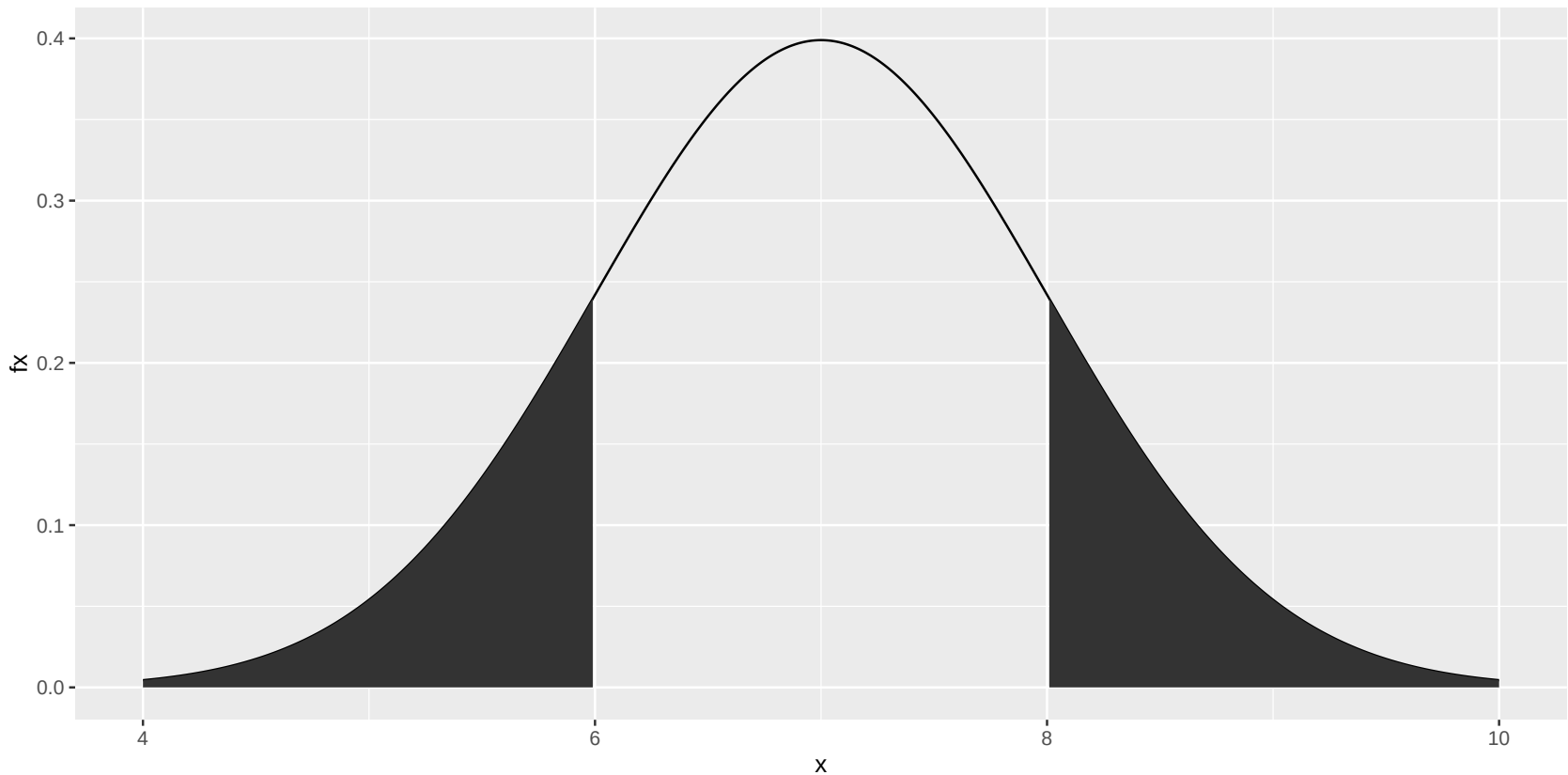
For each $\epsilon > 0$, and for each $\delta > 0$ there exists a natural number N so that, for all $n > N$,

$$P(|X_n - \theta| > \epsilon) < \delta$$

- δ plays the role of ϵ in the original definition of convergence
- ϵ is already in use within the probability

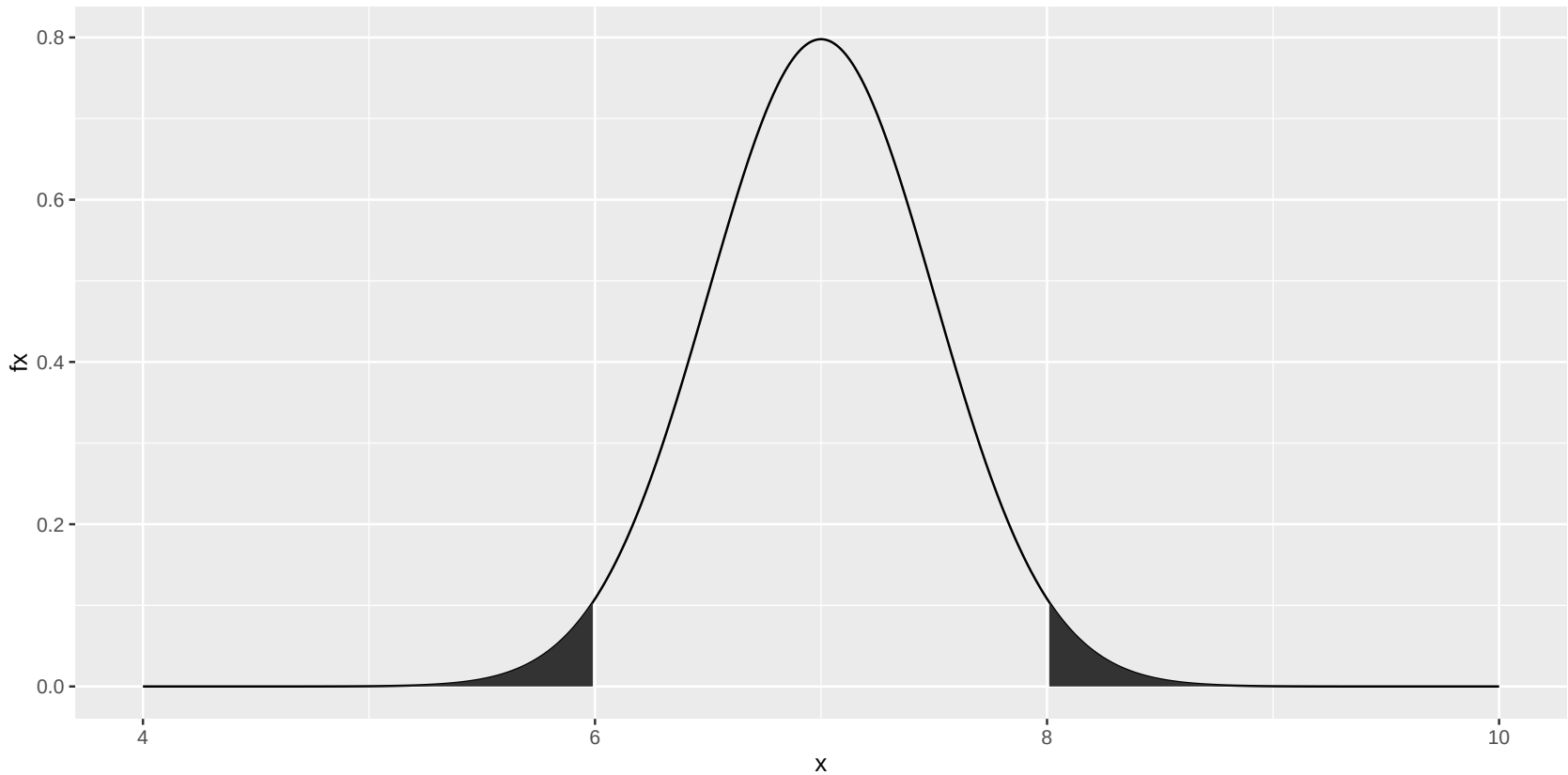
In pictures: $\epsilon = 1$

```
x <- seq(from=4,to=10,by=0.01)
fx <- dnorm(x-7)
qplot(x=x,y=fx,geom="line") + geom_area(aes(x=x[x<6],y=fx[x
                              geom_area(aes(x=x[x>8],y=fx[x
```



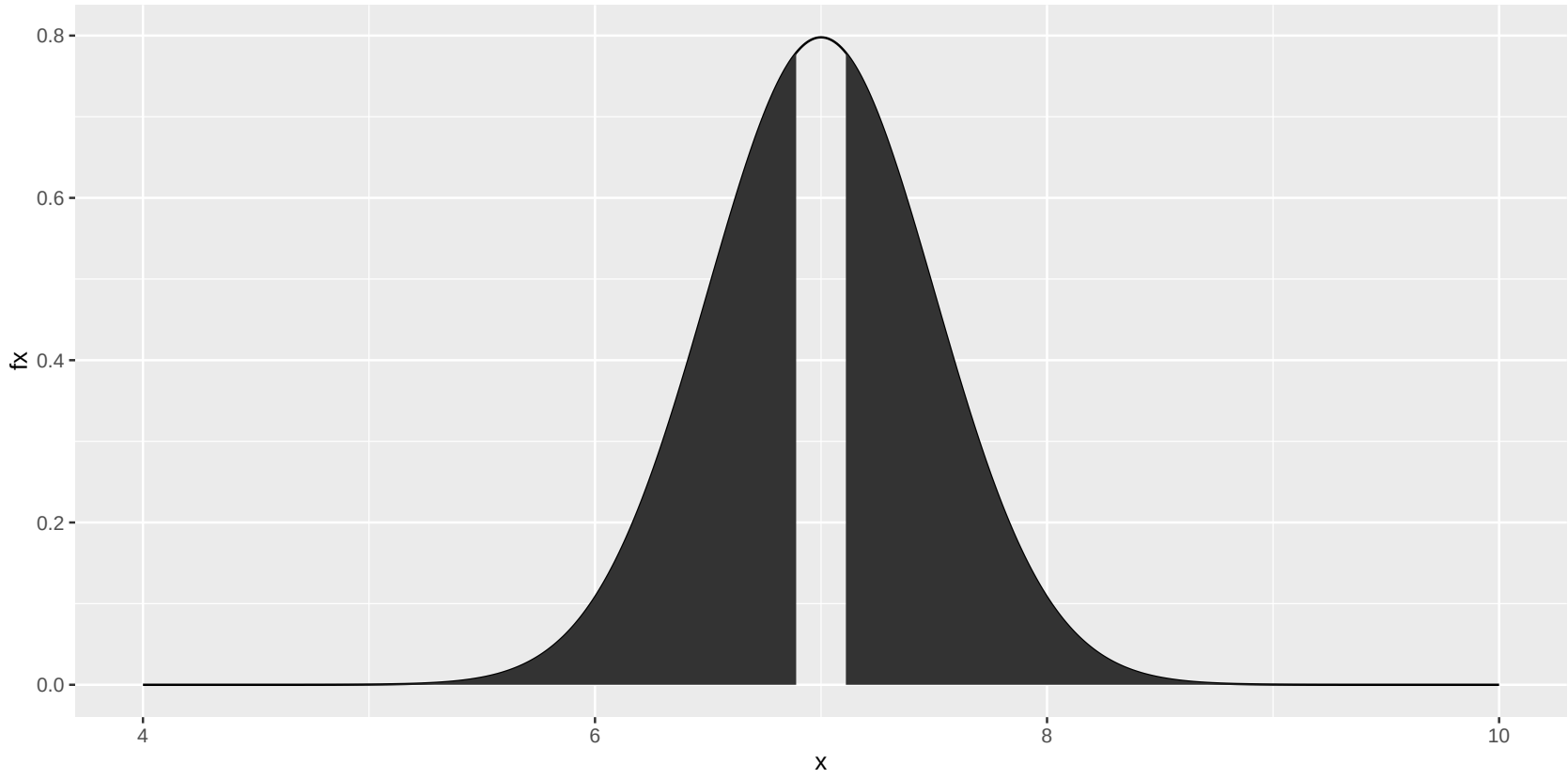
In pictures: $\epsilon = 1$

```
x <- seq(from=4,to=10,by=0.01)
fx <- 2*dnorm((x-7)*2)
qplot(x=x,y=fx,geom="line") + geom_area(aes(x=x[x<6],y=fx[x
                                geom_area(aes(x=x[x>8],y=fx[x
```



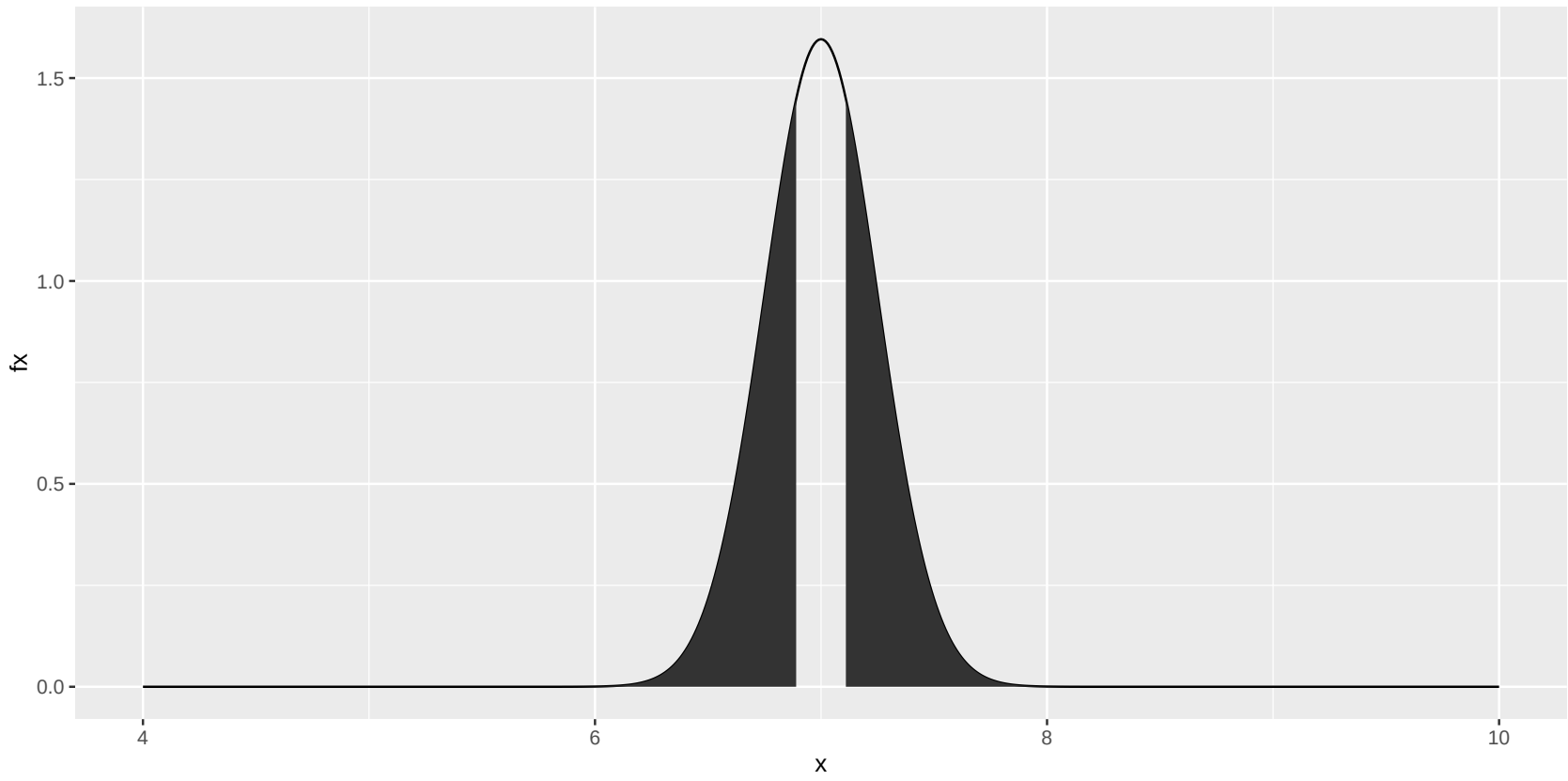
In pictures: : $\epsilon = 0.1$

```
x <- seq(from=4,to=10,by=0.01)
fx <- 2*dnorm((x-7)*2)
qplot(x=x,y=fx,geom="line") + geom_area(aes(x=x[x<6.9],y=fx
                                geom_area(aes(x=x[x>7.1],y=fx
```



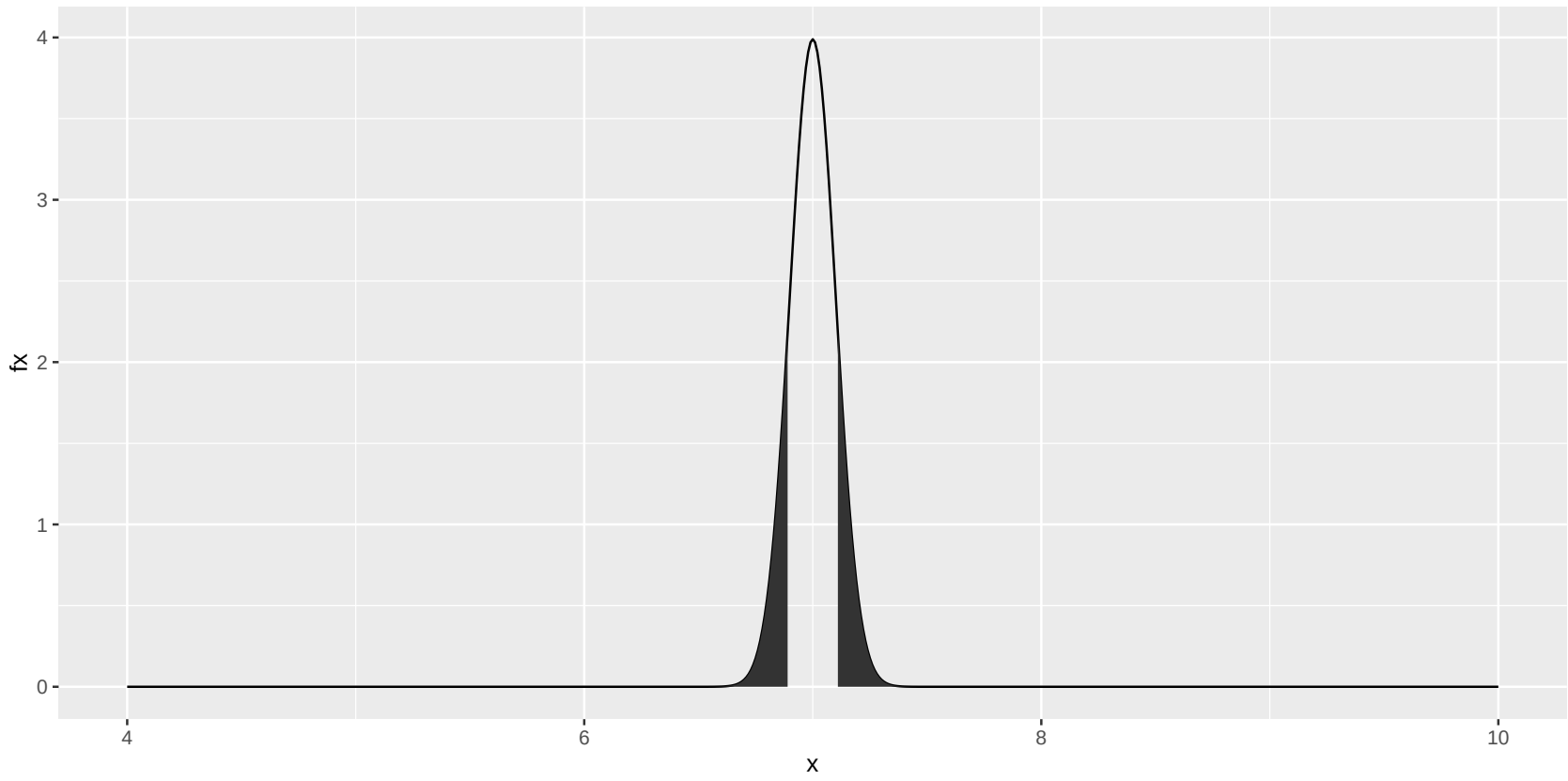
In pictures: : $\epsilon = 0.1$

```
x <- seq(from=4,to=10,by=0.01)
fx <- 4*dnorm((x-7)*4)
qplot(x=x,y=fx,geom="line") + geom_area(aes(x=x[x<6.9],y=fx
                                geom_area(aes(x=x[x>7.1],y=fx
```



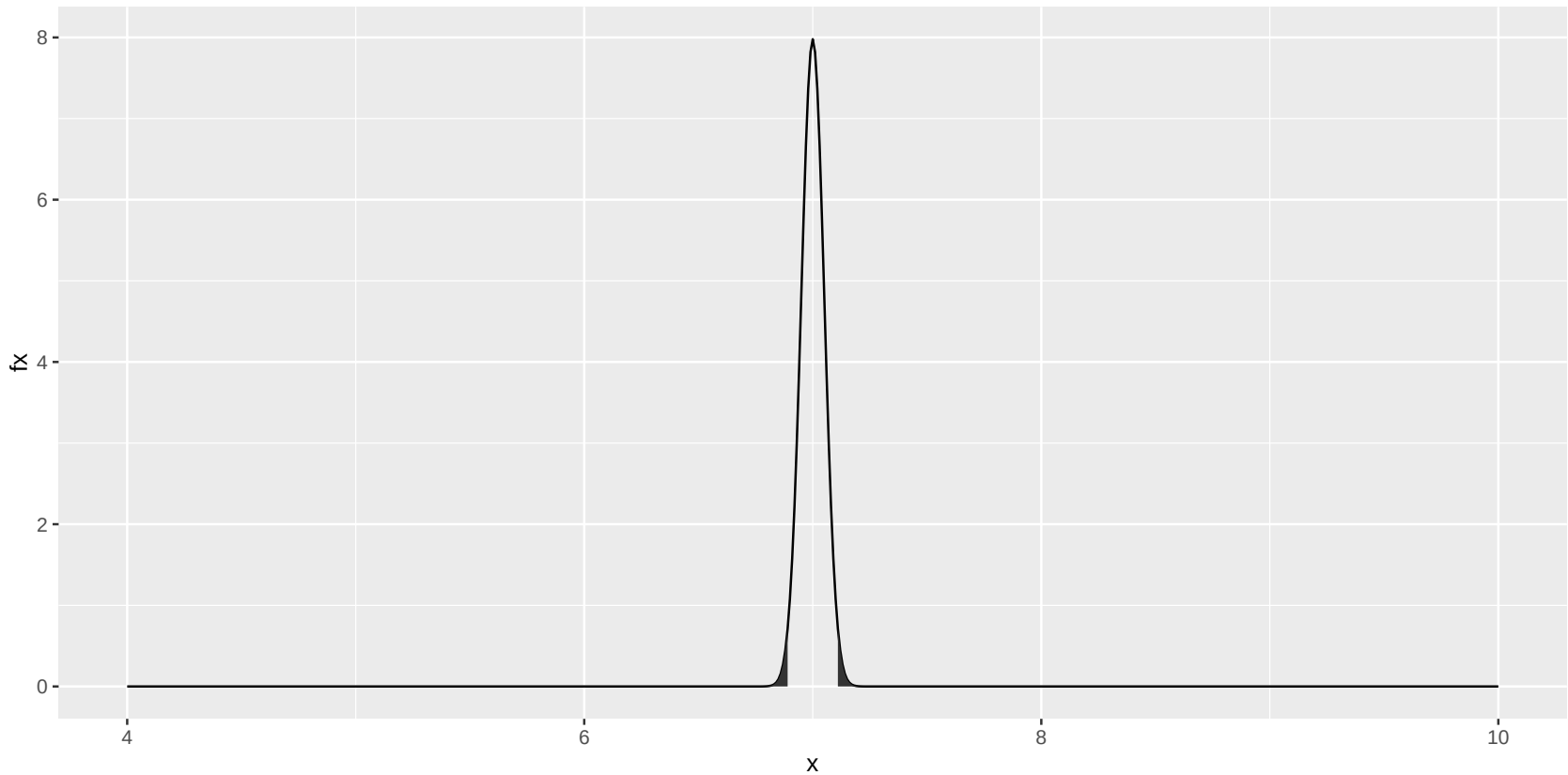
In pictures: : $\epsilon = 0.1$

```
x <- seq(from=4,to=10,by=0.01)
fx <- 10*dnorm((x-7)*10)
qplot(x=x,y=fx,geom="line") + geom_area(aes(x=x[x<6.9],y=fx
                                geom_area(aes(x=x[x>7.1],y=fx
```



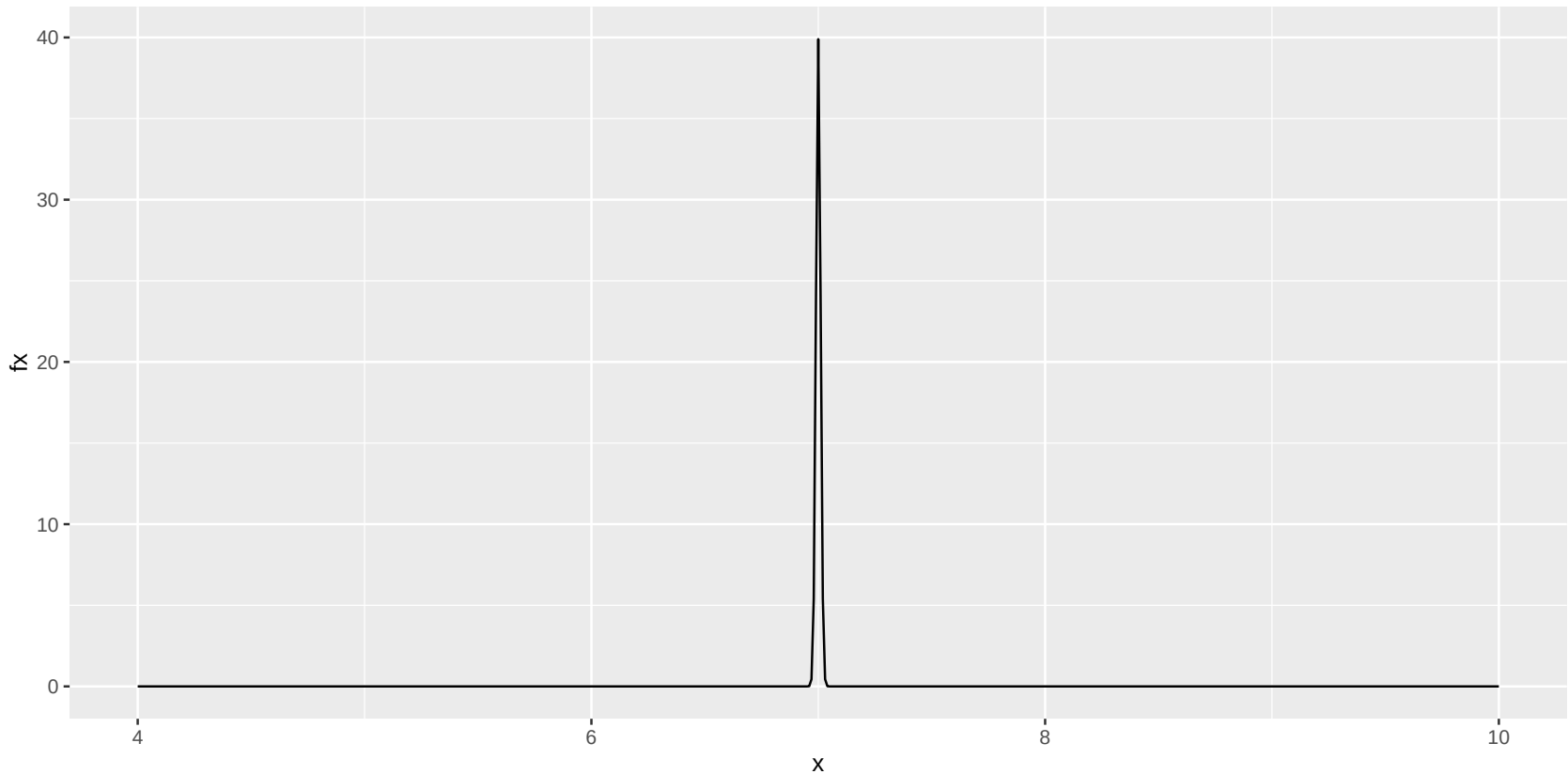
In pictures: : $\epsilon = 0.1$

```
x <- seq(from=4,to=10,by=0.01)
fx <- 20*dnorm((x-7)*20)
qplot(x=x,y=fx,geom="line") + geom_area(aes(x=x[x<6.9],y=fx
                                geom_area(aes(x=x[x>7.1],y=fx
```



In pictures: : $\epsilon = 0.1$

```
x <- seq(from=4,to=10,by=0.01)
fx <- 100*dnorm((x-7)*100)
qplot(x=x,y=fx,geom="line") + geom_area(aes(x=x[x<6.9],y=fx
                                geom_area(aes(x=x[x>7.1],y=fx
```



Chebyshev inequality

If X is a discrete RV, then

$$P(|X - E(X)| \geq \epsilon) \leq \frac{\text{Var}(X)}{\epsilon^2}.$$

Cheb: Proof sketch

The variance is bounded by

$$\begin{aligned}\mathrm{Var}(X) &= \sum_x (x - E(X))^2 P(X = x) \\ &\geq \sum_{x:|x-E(X)|>\epsilon} (x - E(X))^2 P(X = x) \\ &\geq \epsilon^2 P(|x - E(X)| > \epsilon).\end{aligned}$$

Then re-arrange. Proof for continuous RVs is very similar.

LLN

Law of large numbers. If $(Y_1, \cdots, Y_n)$ is i.i.d. with mean $E(Y_1) = \mu_Y$ and $\text{Var}(Y_1) = \sigma_Y^2 < \infty$. Then $\bar{Y}_n \xrightarrow{p} \mu_Y$

LLN: Proof sketch (advanced)

We need to show that the limit of the deterministic sequence $x_n = P(|Y_n - mu_Y| \geq \epsilon)$ is zero for all ϵ.

1. The variance of X_n is . . . ?
2. $x_n = \frac{\sigma_Y^2}{\epsilon^2} \cdot \frac{1}{n}$
3. For each ϵ, the first term is fixed, so $x_n \to 0$ as $n \to \infty$

LLN: conclusion

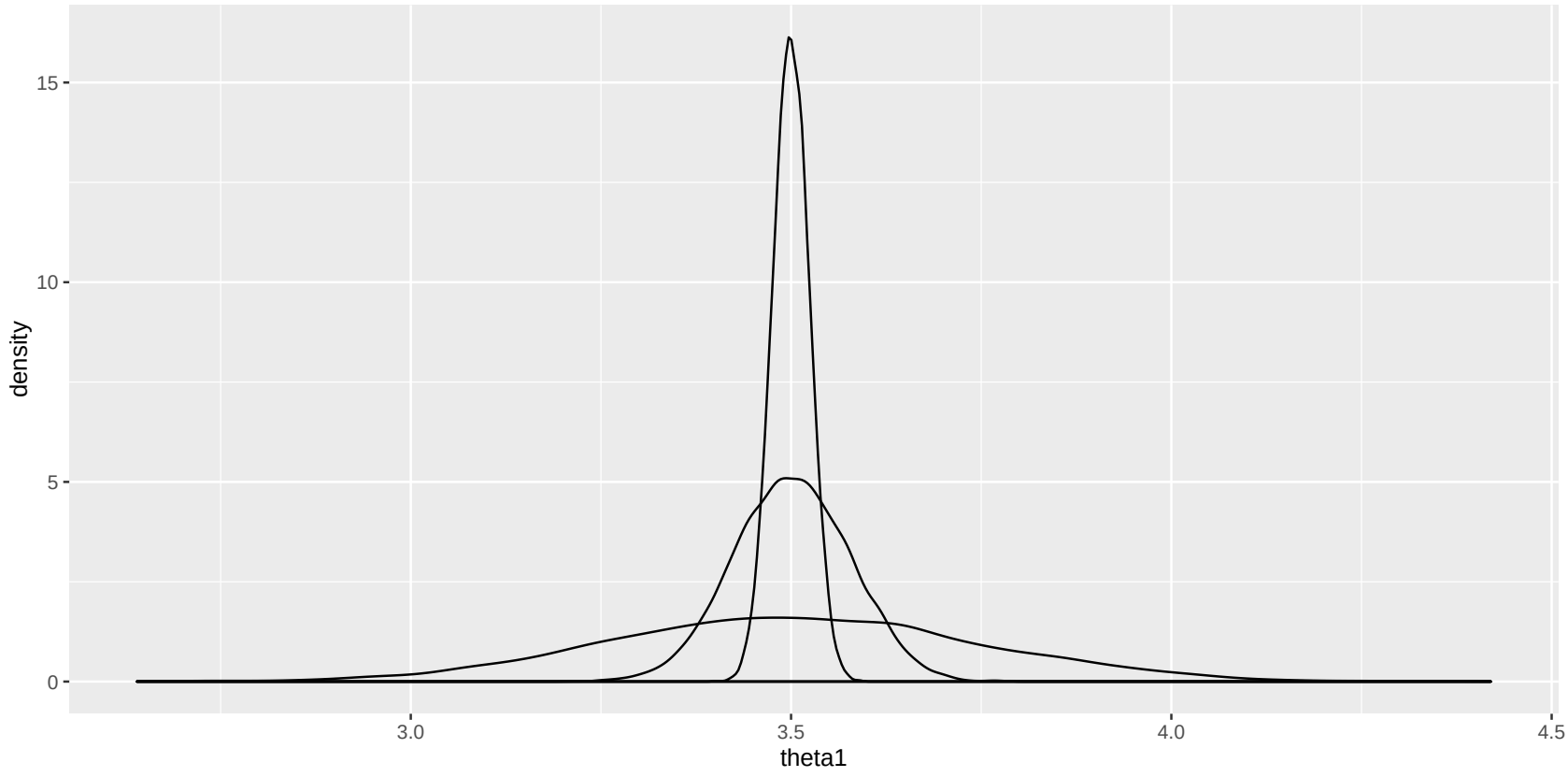
The sample mean is consistent for the expectation.

LLN: recode

```
dice_estimators <- function(nn,S) {
  thetas <- matrix(0,nrow=S,ncol=3)
  for(s in 1:S){
    x <- sample(1:6,size=nn[1],replace=TRUE)
    thetas[s,1] <- mean(x)

    x <- sample(1:6,size=nn[2],replace=TRUE)
    thetas[s,2] <- mean(x)

    x <- sample(1:6,size=nn[3],replace=TRUE)
    thetas[s,3] <- mean(x)

  }
  thetas <- data.frame(thetas)
  colnames(thetas) <- c("theta1","theta2","theta3")
  ggplot(data=thetas) + geom_density(aes(x=theta1),col="bla
    geom_density(aes(x=theta2),col="black") +
    geom_density(aes(x=theta3),col="black")
```

Run it

```
dice_estimators(c(50,500,5000),10000)
```



Consistency (1)

The LLN becomes very useful when combined with the following properties

1. plim.1: If $\hat{\theta}$ is consistent for θ, then $g\left(\hat{\theta}\right)$ is consistent for $g(\theta)$
2. plim.2: Sums, products, and ratios of consistent estimators are consistent as long as the limit is well-defined.

Consistency (1): Example

whiteboard:3.5: consistency of sample variance

Consistency (2)

If $\hat{\theta}$ is unbiased and $Var(\hat{\theta}_n) \to 0$, then $\hat{\theta}$ is consistent.

Convergence in distributions

Let X_n be a sequence of random variables with CDF $F_n(x)$. Let X be a random variable with CDF $F(x)$. The sequence X_n **converges in distribution** to X if, for each x, $F_n(x) \to F(x)$ as $n \to \infty$.

CLT

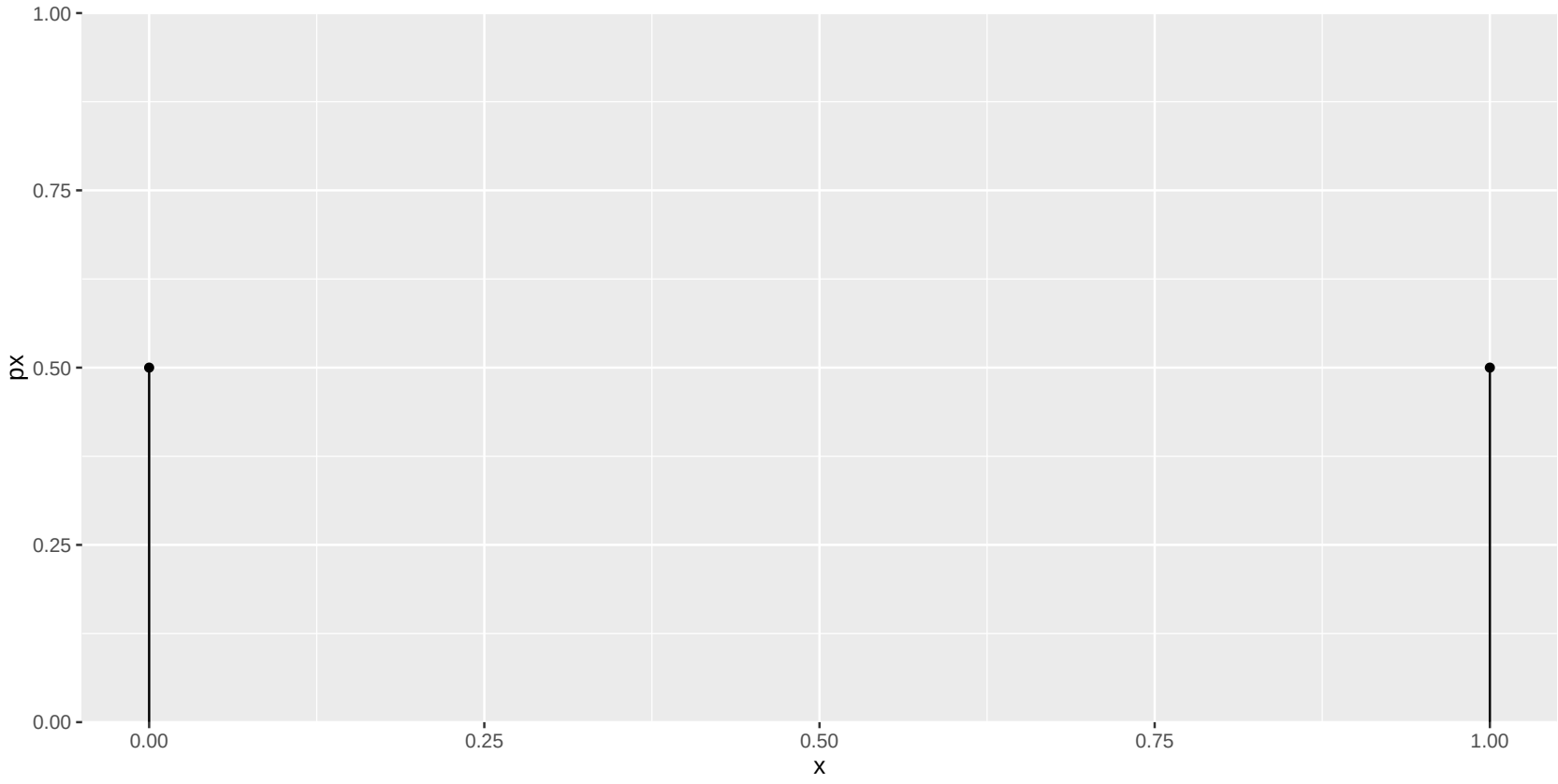
Central Limit Theorem. Let $(Y_1, \cdots, Y_n)$ be a sequence of random variables. Denote by

$$Z_i = (Y_i - E(Y_1))/\sqrt{\text{Var}(Y_1)}$$

the **standardized RV**. Then Z_i converges in distribution to a standard Normal random variable.

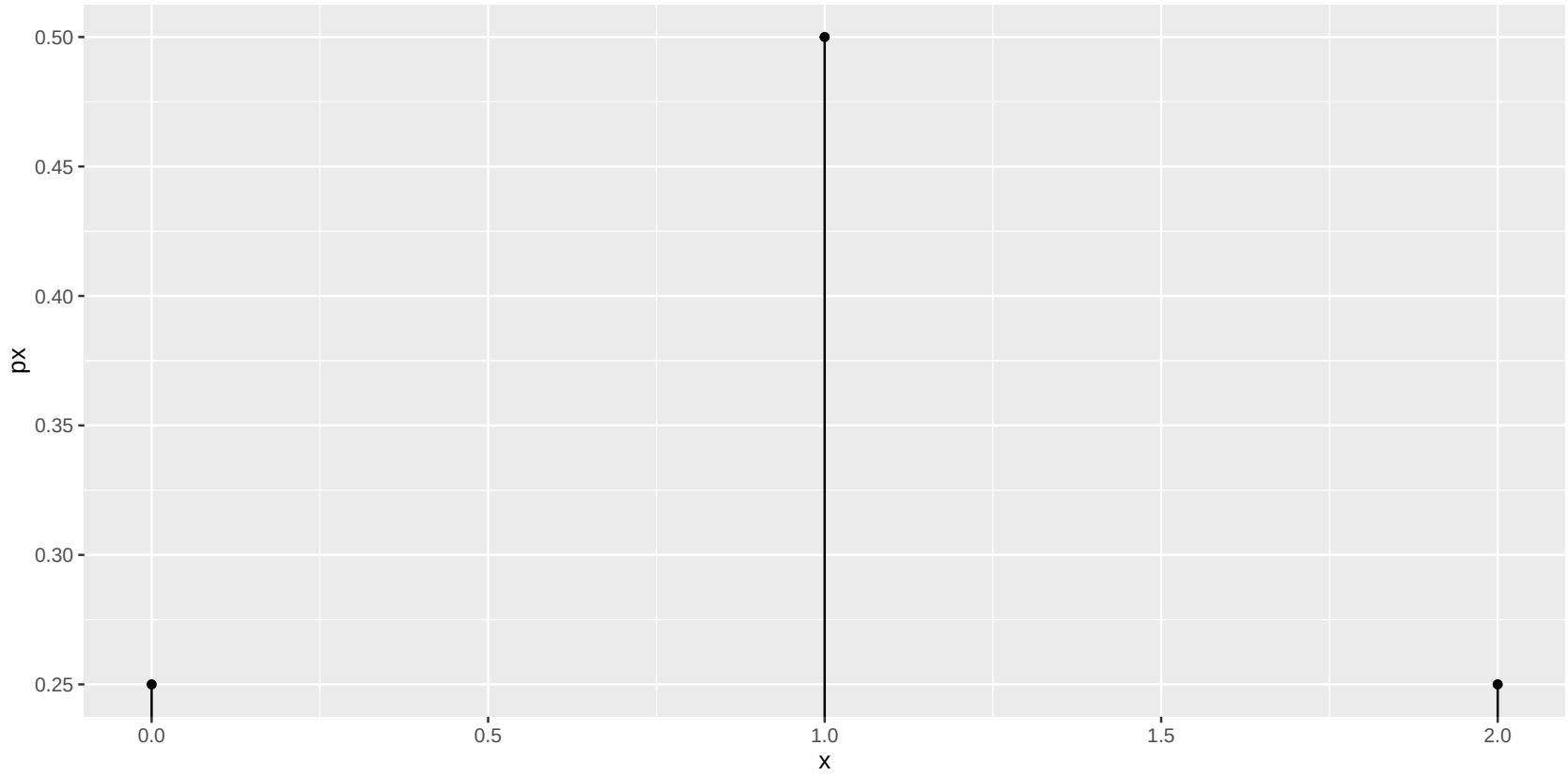
Demonstration in R

```
x <- 0:1
px <- dbinom(x,size=max(x),prob=0.5)
qplot(x=x,y=px) + geom_linerange(x=x,ymin=0,ymax=px)
```



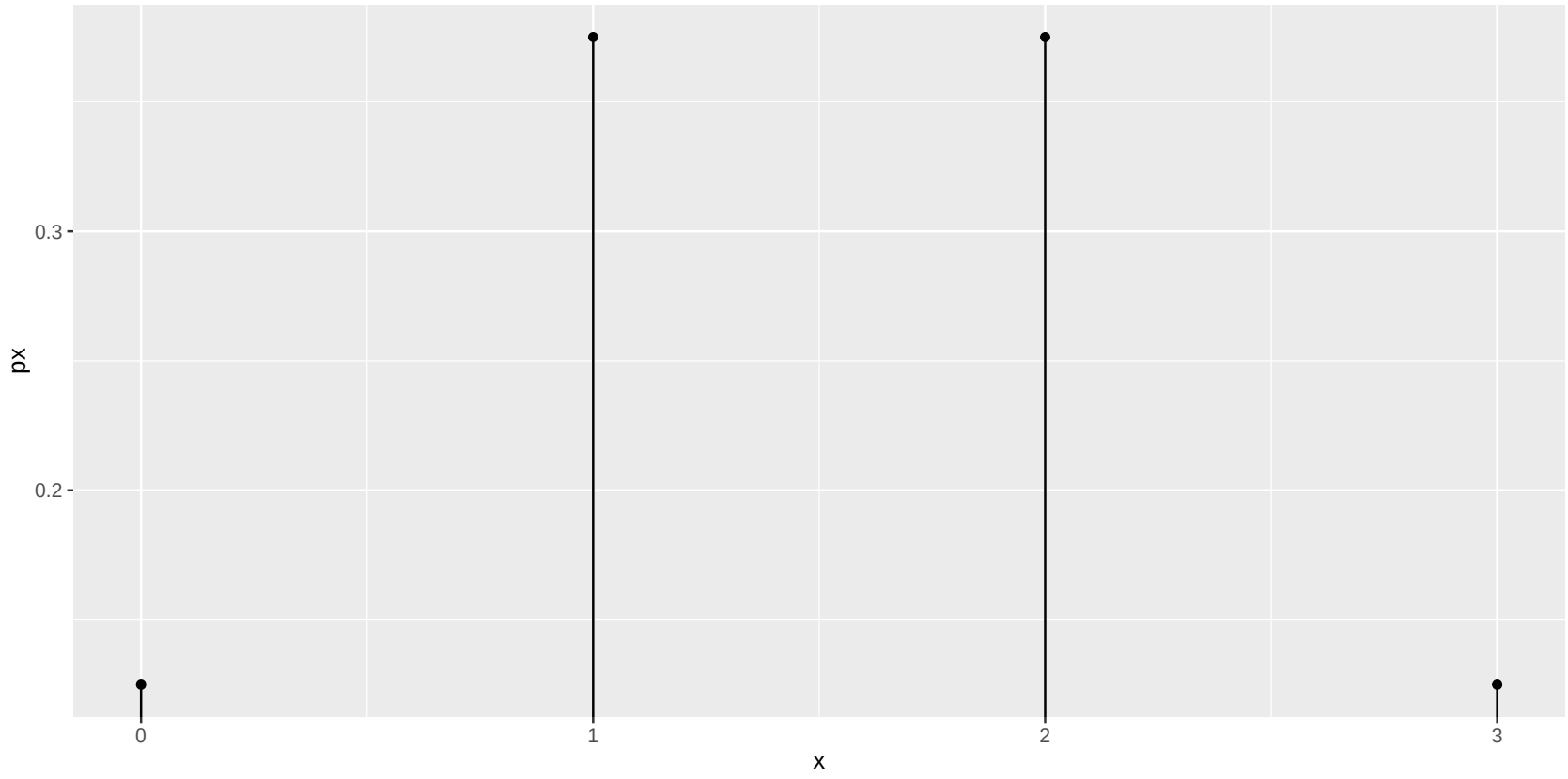
Demonstration in R

```
x <- 0:2
px <- dbinom(x,size=max(x),prob=0.5)
qplot(x=x,y=px) + geom_linerange(x=x,ymin=0,ymax=px)
```



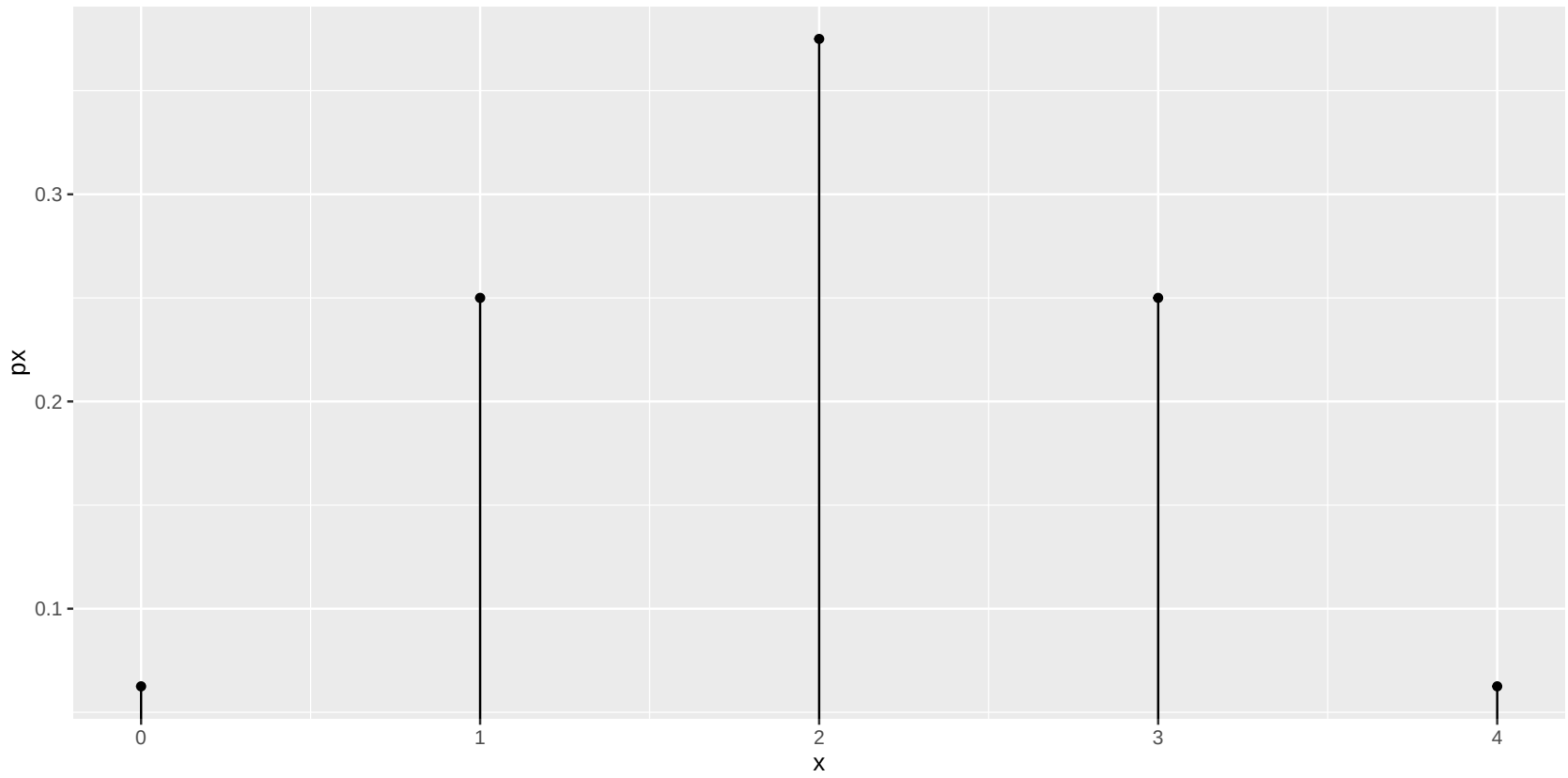
Demonstration in R

```
x <- 0:3
px <- dbinom(x,size=max(x),prob=0.5)
qplot(x=x,y=px) + geom_linerange(x=x,ymin=0,ymax=px)
```



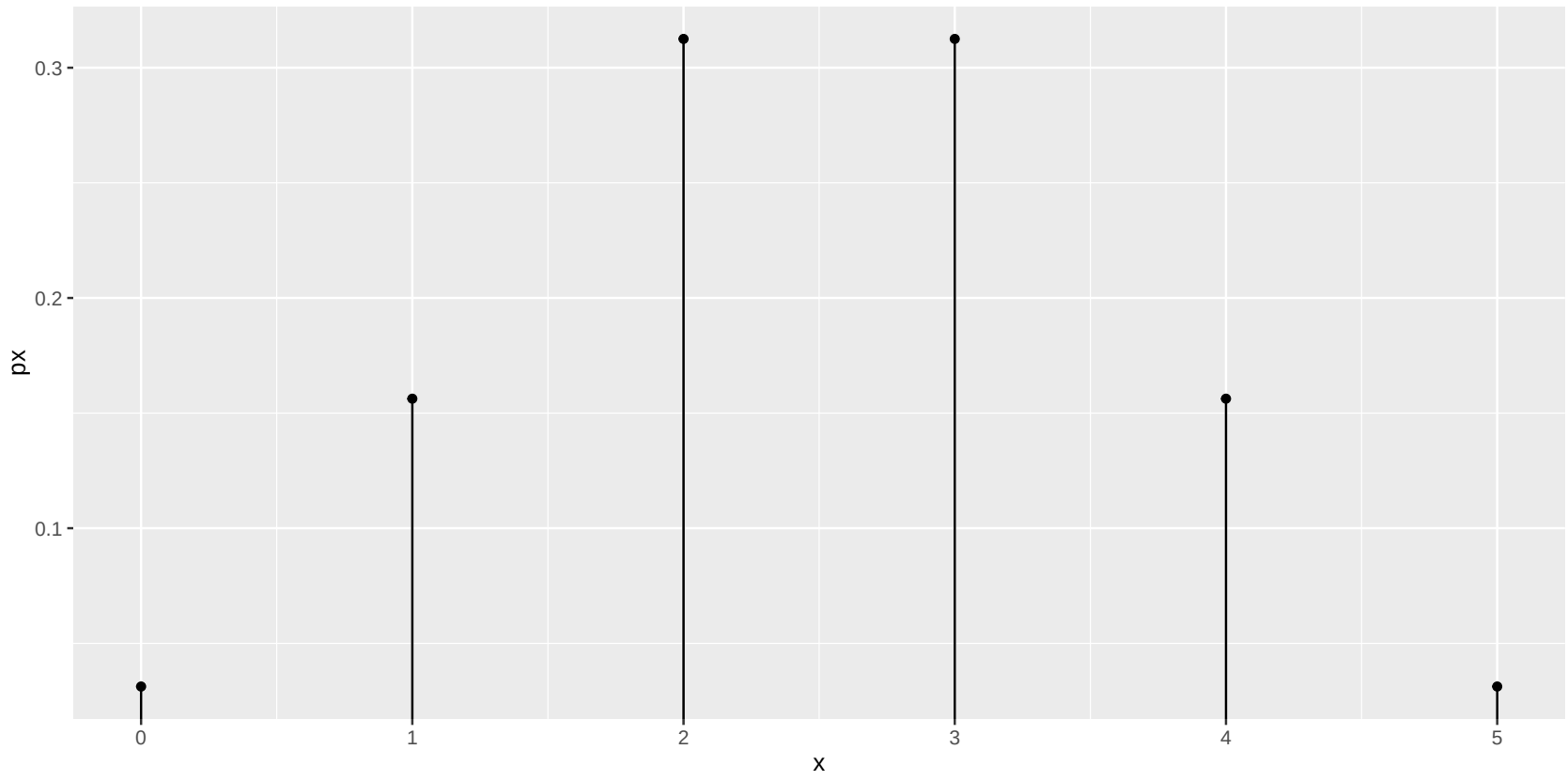
Demonstration in R

```
x <- 0:4
px <- dbinom(x,size=max(x),prob=0.5)
qplot(x=x,y=px) + geom_linerange(x=x,ymin=0,ymax=px)
```



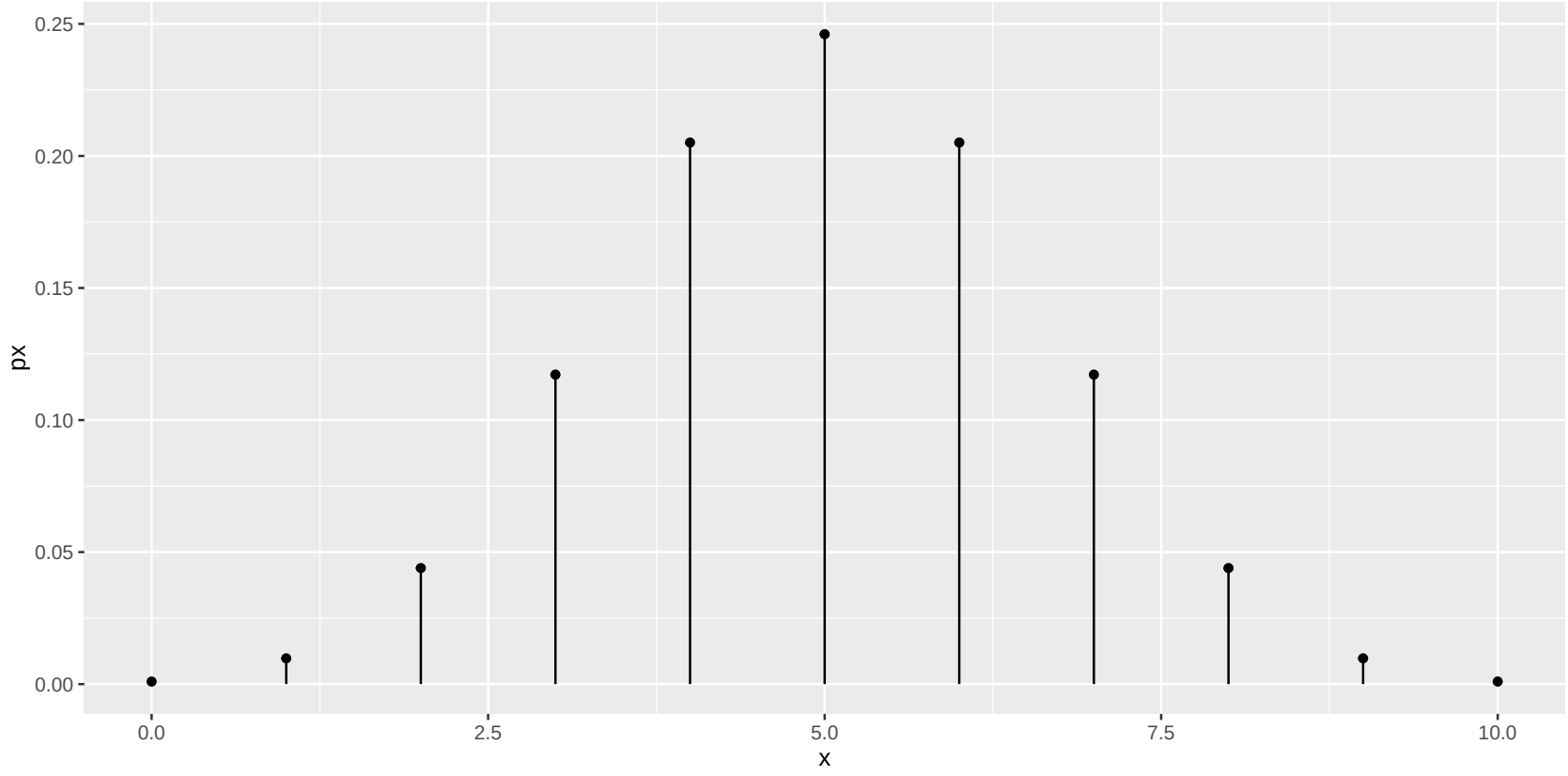
Demonstration in R

```
x <- 0:5
px <- dbinom(x,size=max(x),prob=0.5)
qplot(x=x,y=px) + geom_linerange(x=x,ymin=0,ymax=px)
```



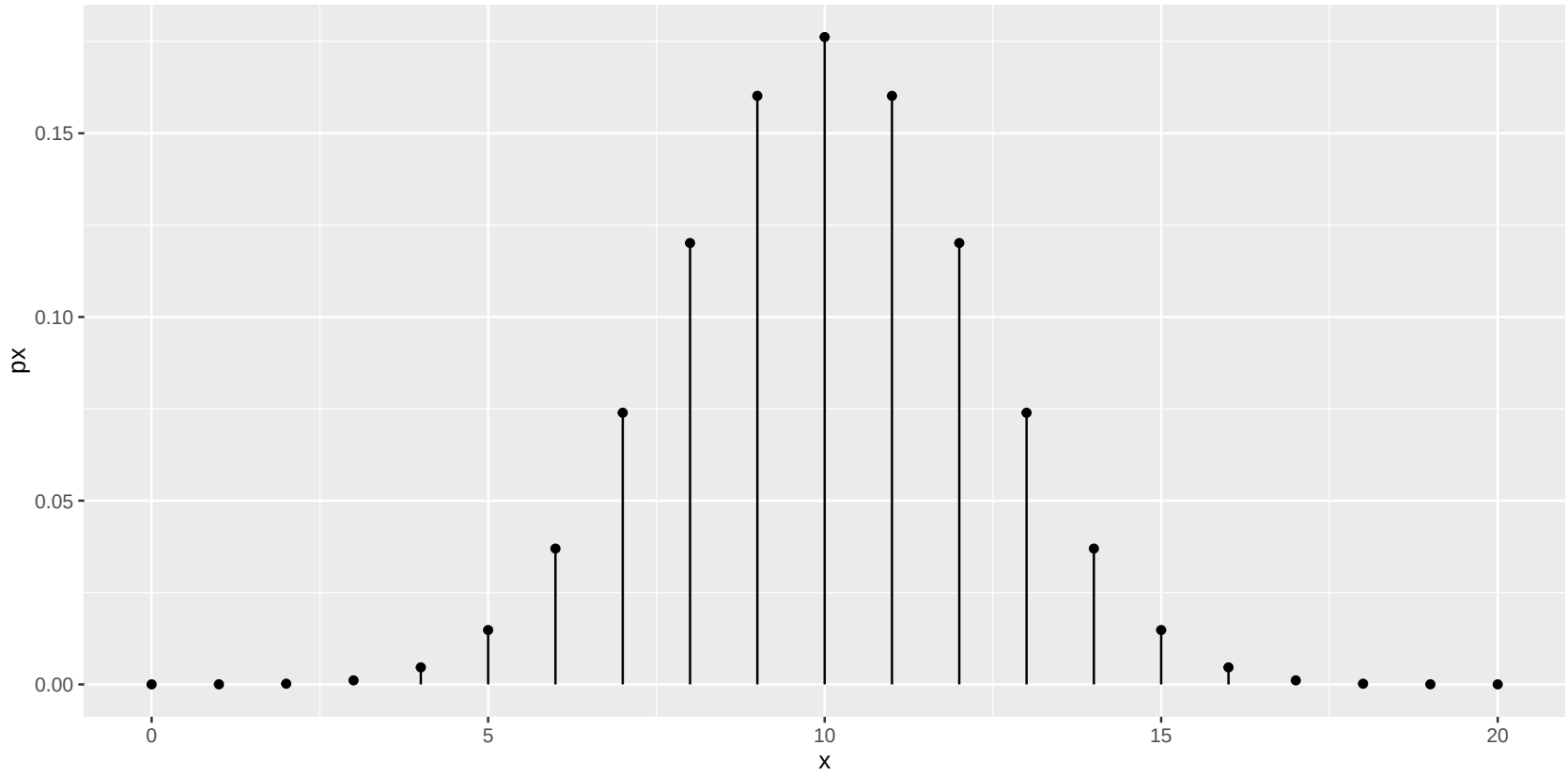
Demonstration in R

```
x <- 0:10
px <- dbinom(x,size=max(x),prob=0.5)
qplot(x=x,y=px) + geom_linerange(x=x,ymin=0,ymax=px)
```



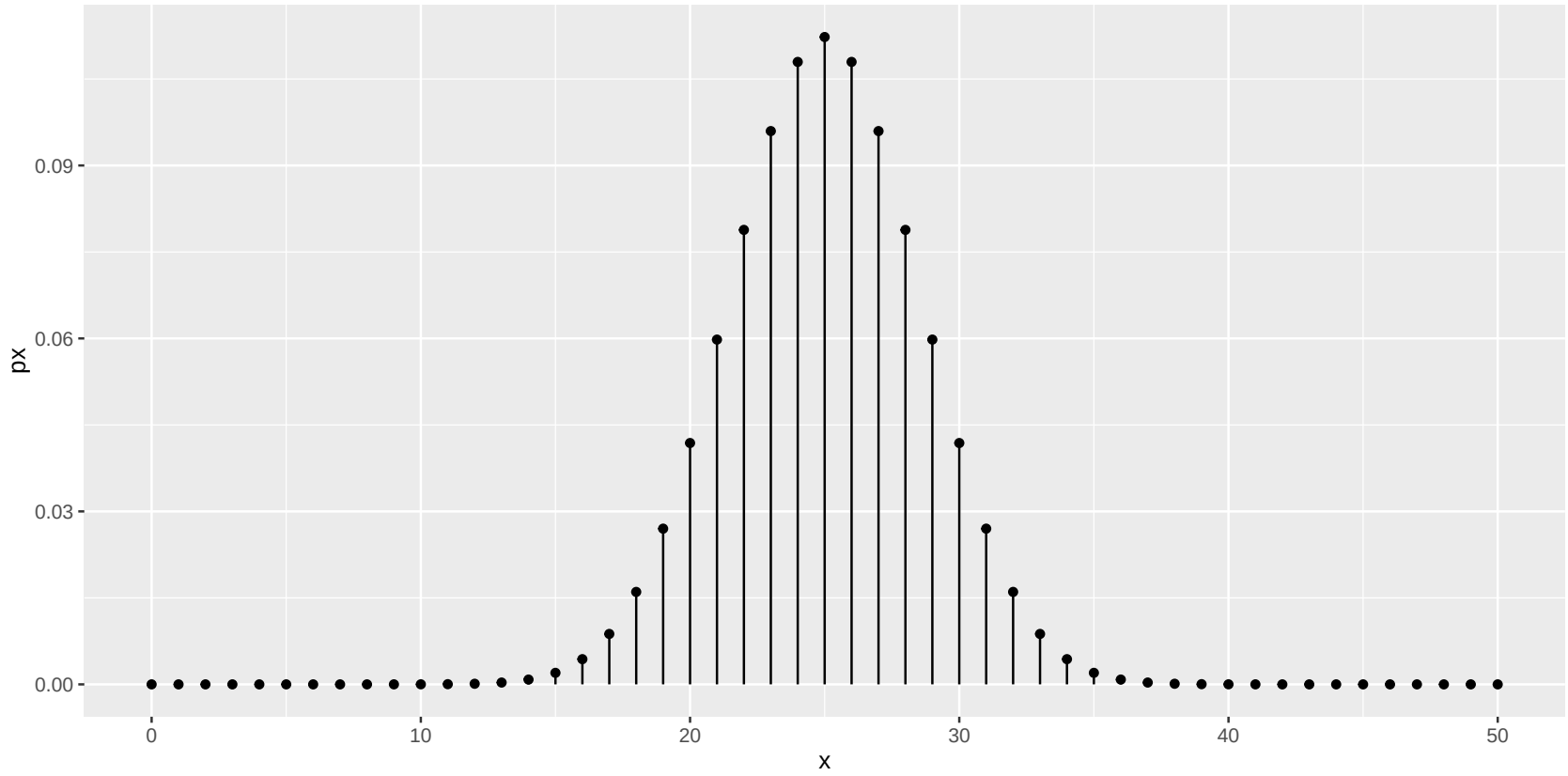
Demonstration in R

```
x <- 0:20
px <- dbinom(x,size=max(x),prob=0.5)
qplot(x=x,y=px) + geom_linerange(x=x,ymin=0,ymax=px)
```



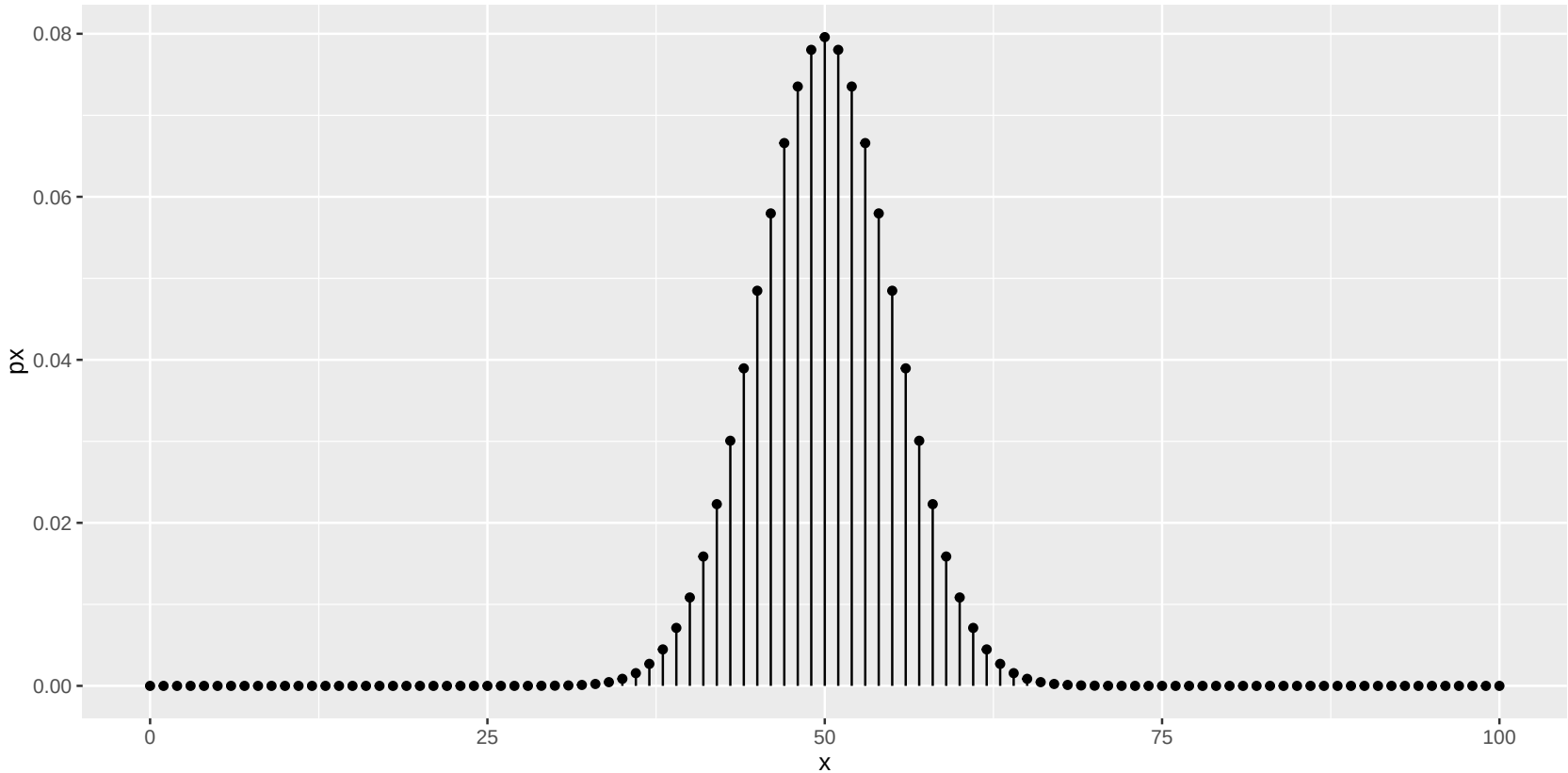
Demonstration in R

```
x <- 0:50
px <- dbinom(x,size=max(x),prob=0.5)
qplot(x=x,y=px) + geom_linerange(x=x,ymin=0,ymax=px)
```



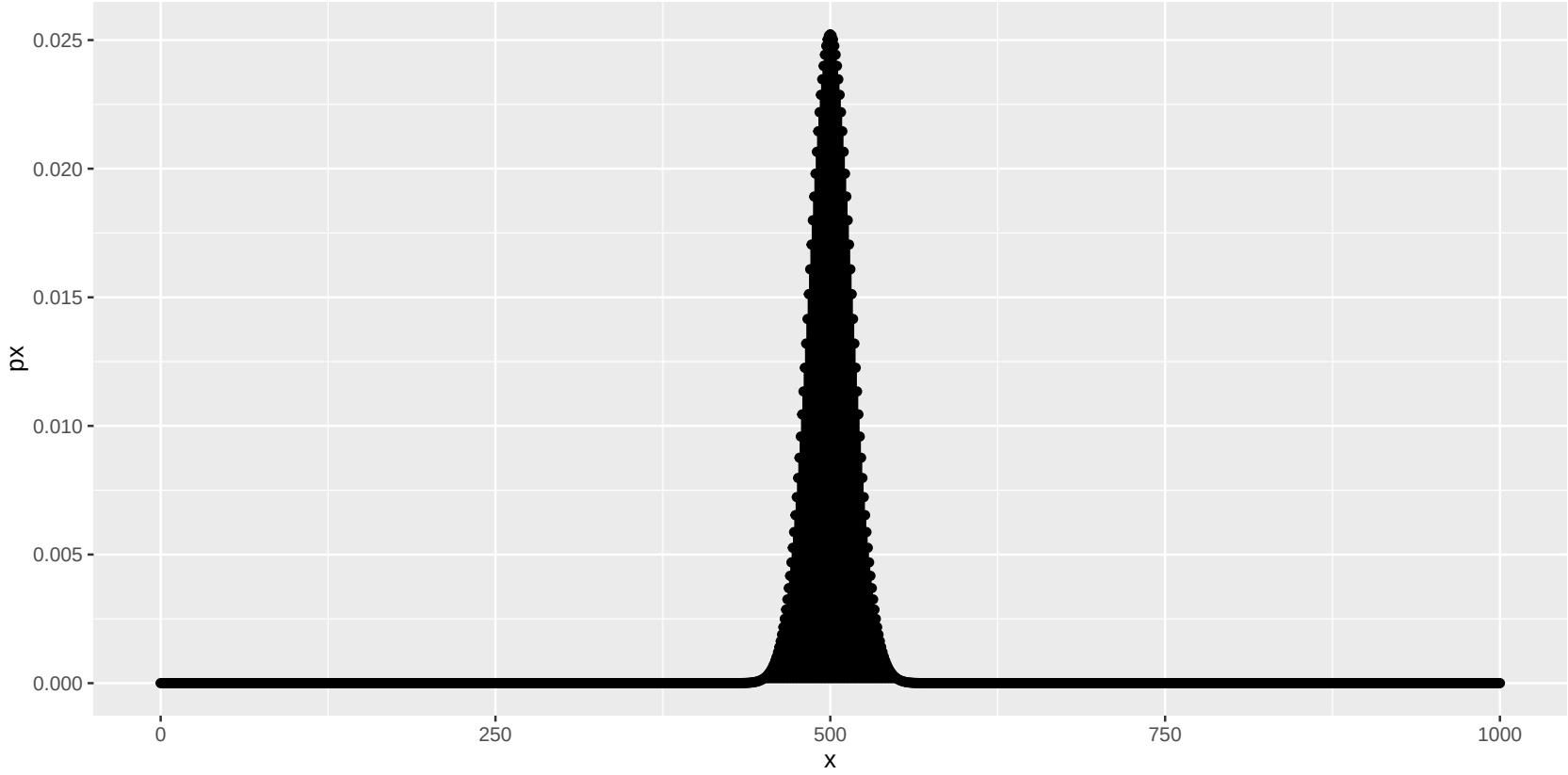
Demonstration in R

```
x <- 0:100
px <- dbinom(x,size=max(x),prob=0.5)
qplot(x=x,y=px) + geom_linerange(x=x,ymin=0,ymax=px)
```



Demonstration in R

```
x <- 0:1000
px <- dbinom(x,size=max(x),prob=0.5)
qplot(x=x,y=px) + geom_linerange(x=x,ymin=0,ymax=px)
```



CLT: Conclusion

CLT is useful because:

1. It guarantees that estimators behave nicely if you have lots of data
2. It allows you to construct confidence intervals