Probability and Inference

Review

- A **random variable** is like drawing a numbered ticket from a hat.
- We can define a ***discrete*** **random variable** with a **probability *mass* function (pmf)** $f(x)$, so that $\Pr(X = x) = f(x)$.
- We can define a ***continuous*** **random variable** with a **probability *density* function (pmf)** $f(x)$, so that $\Pr(a \leq X \leq b) = \int_a^b f(x)dx$.
- Or we can define a random variable with a **cumulative distribution function** $F(x)$, so that $\Pr(X \leq x) = F(x)$.

Essentially, the pmf/pdf/cdf tells us how likely we are to draw certain numbers from the hat.

Definition 1

*The **expected value** $E(X)$ (or "**mean**") of a random variable X is defined as follows:*

- *For a discrete random variable X, $E(X) = \sum_{\text{all } x} xf(x)$.* [a]
- *For a continuous random variable X, $E(X) = \int_{-\infty}^{\infty} xf(x)dx$.*[b]

[a] We need the additional restriction that the sum over just the *positive* values or the sum over just the *negative* values is finite. If *both* sums are infinite, then the expected value is not defined. This restriction holds for all bounded discrete random variables and most unbounded discrete random variables.

[b] For continous random variables, we need either the integral over just the positive values or just the negative values to be finite, else the expected value is not defined. This holds for all bounded random variables and most unbounded random variables.

We should think of the expected value of a random variable as a "hypothetical, long-run average" if we sampled from the distribution over, and over, and over again and took the average of those samples.

Theorem 1 (Law of the Unconscious Statistician)

Suppose a random variable X with pdf or pmf $f(x)$. Then $E\left[g(X)\right] = \int_{-\infty}^{\infty} g(x)f(x)dx$ or $E\left[g(X)\right] = \sum_x g(x)f(x)$, respectively.

Example 1 (Expectation of a Draw from a Hat)

I will draw one ticket from a hat that contains four tickets numbered -4, 0, 6, and 6. Find $E(X)$.

We can see the pmf is a stepwise function with $\Pr(X = -4) = \Pr(X = 0) = 0.25$ and $\Pr(X = 6) = 0.50$ (and 0 otherwise). Then we have $E(X) = (-4 \times 0.25) + (0 \times 0.25) + (6 \times 0.50) = -1 + 0 + 3 = 2.$

Exercise 1

Suppose $X \sim \text{Bernoulli}(\pi)$. Find $E(X)$.

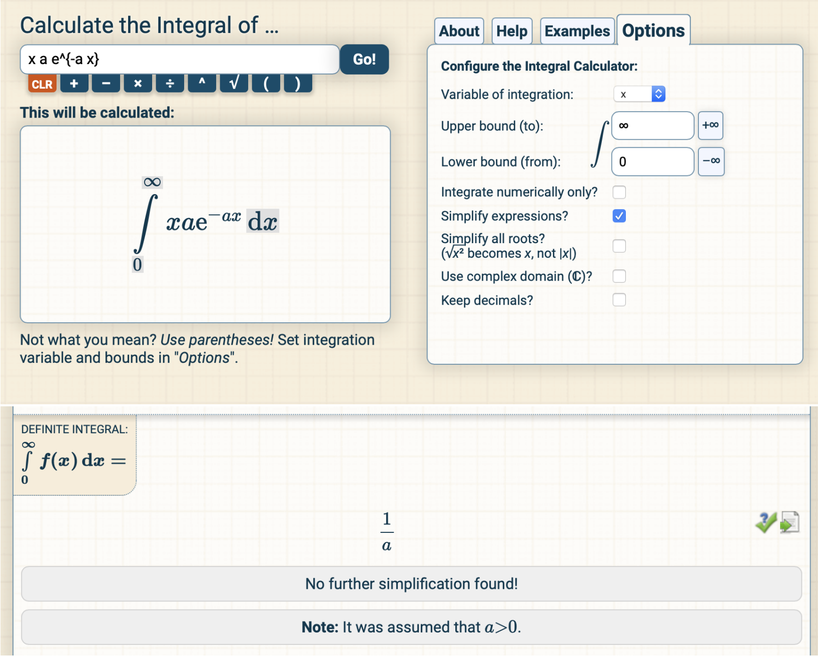
Example 2 (Expectation of an Exponential Random Variable)

An exponential random variable X has pdf $f(x) = \lambda e^{-\lambda x}$. Find $E(X)$.

$$\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} xf(x)dx \text{ (def. of } E\text{)} \\
&= \int_{0}^{\infty} xf(x)dx \text{ ((}f(x) > 0 \text{ for } x > 0\text{)} \\
&= \int_{0}^{\infty} x\lambda e^{-\lambda x}dx \text{ (just fill in } f\text{)} \\
&= \lim_{r\to\infty} \left[-\frac{(\lambda x + 1)e^{-\lambda x}}{\lambda}\right]\Big|_{0}^{\infty} \text{ (integral-calculator.com)} \\
&= (0) - \left(-\frac{1}{\lambda}\right) \\
&= \frac{1}{\lambda}
\end{aligned}$$

We can verify the definite integral with integral-calculator.com.

We can verify the definite integral with integral-calculator.com.



Calculate the Integral of …

x a e^{-a x}

This will be calculated:

$$\int_{0}^{\infty} x a e^{-ax} \, dx$$

Not what you mean? *Use parentheses!* Set integration variable and bounds in "*Options*".

Configure the Integral Calculator:

Variable of integration: x

Upper bound (to): ∞

Lower bound (from): 0

Integrate numerically only? □

Simplify expressions? ☑

Simplify all roots? ($\sqrt{x^2}$ becomes *x*, not |*x*|) □

Use complex domain (ℂ)? □

Keep decimals? □

DEFINITE INTEGRAL:

$$\int_{0}^{\infty} f(x)\,dx =$$

$$\frac{1}{a}$$

No further simplification found!

Note: It was assumed that $a>0$.

Theorem 2 (Some Properties of Expectations)

Suppose random variables X and Y so that $E(X)$ and $E(Y)$ exist. Suppose constants a and b. Then the following results hold.

1. $E(aX + b) = aE(X) + b$
   1.1 $E(b) = b$.
   1.2 $E(X + b) = E(X) + b$.
   1.3 $E(aX) = aE(X)$.
2. $E(X + Y) = E(X) + E(Y)$.
3. $E(XY) = E(X)E(Y)$ *if X and Y are independent.*
4. $E\left[g(X)\right] \geq g\left[E(X)\right]$ *for a convex function g.*[a] *(Jensen's Inequality.)*

[a] Key takeaway: $E\left(X^2\right) \neq E(X)^2$. For *strictly* convex g and *nondegenerate* X (i.e., X is not a constant), the inequality is strict as well. For concave g, the inequality flips, as you would expect.

Exercise 2

Suppose three independent random variables W, X, and Y so that $E(W) = 1$, $E(X) = -2$, and $E(Y) = 14$. Find $E\left[5W(17 + 2X - 4Y)\right]$.

Definition 2

Suppose a random variable X with finite mean $E(X) = \mu$. We define the variance of X as $V(X) = E\left[(X-\mu)^2\right]$. If X has an infinite or not-existent mean, the we say $V(X)$ does not exist.

Notes

- Some authors denote $V(X)$ as $\text{Var}(X)$ or σ^2.
- The *standard deviation* of *SD* equals $\sqrt{V(X)} = SD(X) = \sigma$ (if $V(X)$ exists).
- We should think of the variance (or SD) of a random variable as a "hypothetical, long-run variance (or SD)" if we sampled from the distribution over, and over, and over again and took the variance (or SD) of that distribution.

In practice, the formula below makes computing a variance a little easier.

Theorem 3 (Easier Method to Calculate the Variance)

For random variable X, $V(X) = E\left(X^2\right) - \mu^2$.

Exercise 3

Prove Theorem 3. Hint: Use algebra and the rules for manipulating expectations.

Example 3 (Variance of of a Draw from a Hat)

I will draw one ticket from a hat that contains tickets numbered -4, 0, 6, and 6. Find $V(X)$.

From before:

1. The pmf is a stepwise function with $\Pr(X = -4) = \Pr(X = 0) = 0.25$ and $\Pr(X = 6) = 0.50$ (and 0 otherwise).
2. $E(X) = 2$.

Then

$$\begin{aligned} V(X) &= E\left(X^2\right) - \mu^2 \\ &= [((-4)^2 \times 0.25) + (0^2 \times 0.25) + (6^2 \times 0.50)] - 2^2 \\ &= (4 + 0 + 18) - 4 = 18 \end{aligned}$$

We can confirm our $E(X) = 2$ and $V(X) = 18$ with a quick simulation.

```r
box <- c(-4, 0, 6, 6)
s <- sample(box, size = 100000, replace = TRUE)

mean(s)
```

```
## [1] 1.99052
```

```r
var(s)
```

```
## [1] 17.96193
```

Exercise 4

Suppose $X \sim \text{Bernoulli}(\pi)$. Find $V(X)$.

Example 4 (Variance of an Exponential Random Variable)

An exponential random variable X has pdf $f(x) = \lambda e^{-\lambda x}$. Find $V(X)$.

Recall that $V(X) = E\left(X^2\right) - \mu^2$. We already found that $\mu = \frac{1}{\lambda}$. We just need $E\left(X^2\right)$. By the law of the unconscious statistician, $E\left(X^2\right) = \int_0^\infty x^2 \lambda e^{-\lambda x}$.

Make integral-calculator.com go brrrrrr… $E\left(X^2\right) = \frac{2}{\lambda^2}$.

Then $V(X) = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$.

We can confirm our $E(X) = \frac{1}{\lambda}$ and $V(X) = \frac{1}{\lambda^2}$ with a quick simulation.

```r
lambda <- 1/10
s <- rexp(100000, rate = lambda)

1/lambda
```

```
## [1] 10
```

```r
mean(s)
```

```
## [1] 9.993483
```

```r
1/(lambda^2)
```

```
## [1] 100
```

```r
var(s)
```

```
## [1] 100.3316
```

We can confirm our $E(X) = \frac{1}{\lambda}$ and $V(X) = \frac{1}{\lambda^2}$ with a quick simulation.

```r
lambda <- 3
s <- rexp(100000, rate = lambda)

1/lambda
```

```
## [1] 0.3333333
```

```r
mean(s)
```

```
## [1] 0.3336857
```

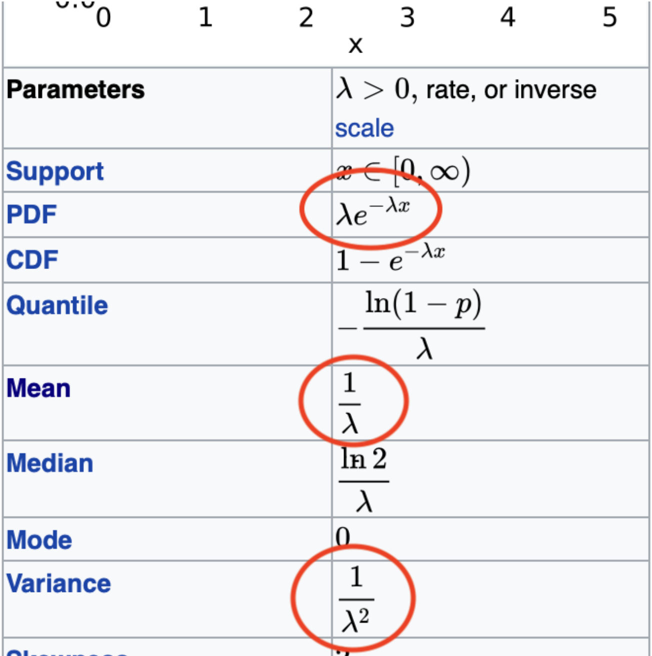
```r
1/(lambda^2)
```

```
## [1] 0.1111111
```

```r
var(s)
```

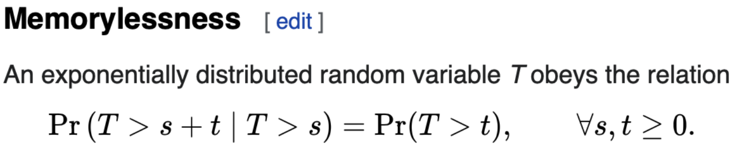
```
## [1] 0.1115846
```

Wikipedia!



Parameters	$\lambda > 0$, rate, or inverse scale
Support	$x \in [0, \infty)$
PDF	$\lambda e^{-\lambda x}$
CDF	$1 - e^{-\lambda x}$
Quantile	$-\frac{\ln(1-p)}{\lambda}$
Mean	$\frac{1}{\lambda}$
Median	$\frac{\ln 2}{\lambda}$
Mode	0
Variance	$\frac{1}{\lambda^2}$

Wikipedia!



Memorylessness [edit]

An exponentially distributed random variable *T* obeys the relation

$$\Pr\left(T > s + t \mid T > s\right) = \Pr(T > t), \qquad \forall s, t \geq 0.$$

Theorem 4 (Some Properties of Variances)

Suppose random variables X and Y so that $V(X)$ and $V(Y)$ exist. Suppose constants a and b. Then the following results hold.

1. $V(aX+b)=a^2E(X)$
   1.1 $V(b)=0.$
   1.2 $V(X+b)=V(X).$
   1.3 $V(aX)=a^2E(X).$
2. $V(X+Y)=V(X)+V(Y)$ *if* X *and* Y *are indepedent.*
3. $V(X+Y)=V(X)+V(Y)+2Cov(X,Y).$
4. $V(aX+bY)=a^2V(X)+b*2V(Y)+2abCov(X,Y).$
5. $V(aX-bY)=a^2V(X)+b*2V(Y)-2abCov(X,Y).$

Multivariate Distributions, Briefly

Sometimes, we have two discrete random variables X and Y that we wish to model jointly. In this case, we can use a *joint pmf* $f(x, y) = \Pr(X = x \text{ and } Y = y)$. This easily (and intuitively, I think) generalizes to three or more random variables.

Similarly, we can use a joint pdf $f(x, y)$ to model two continuous random variables X and Y, where $\Pr[(X, Y) \in A] = \int_A \int f(x, y) dydx$.

Rather than integrating to find the area under a curve, we're integrating to find the area under a surface.

Multivariate Distributions, Briefly

Our usual, univariate random variables give us a single number or “scalar” for each draw. Bivariate and multivariate random variables gives us a vector with two and n values, respectively.

```r
library(mvtnorm)

# mean vector
# E(X1) = -2; E(X2) = 3
mu <- c(-2, 15); mu
```

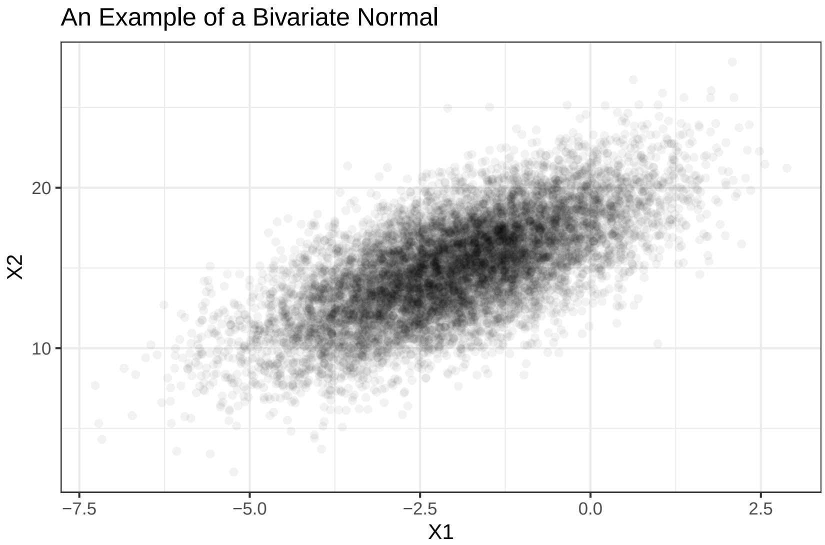
```
## [1] -2 15
```

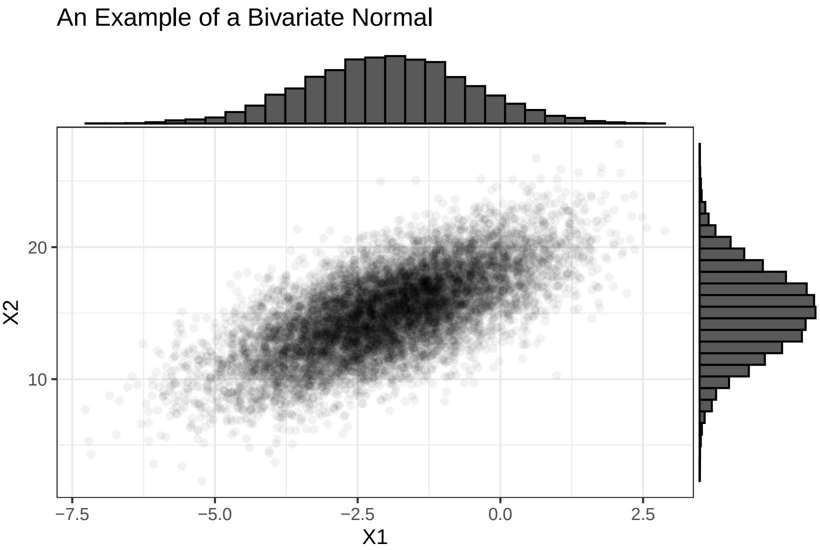
```r
# variance matrix // covaraince matrix
# V(X1) = 2.0; V(X2) = 10; COV(X1, X2) = 3
Sigma <- matrix(c(2, 3, 3, 10), nrow = 2, ncol = 2); Sigma
```

```
##      [,1] [,2]
## [1,]    2    3
## [2,]    3   10
```

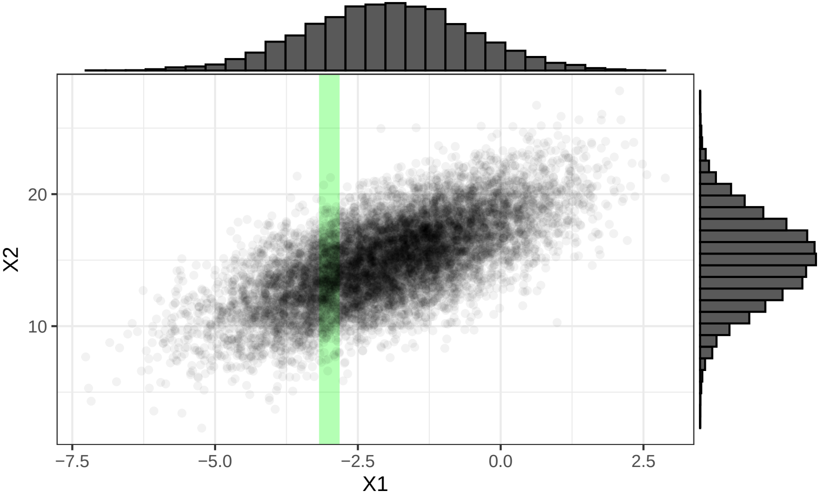
```r
# draws from MNV(mu, Sigma)
draws <- rmvnorm(10000, mean = mu, sigma = Sigma)
head(draws)  # show first 6 rows
```

```
##           [,1]      [,2]
## [1,] -3.507402 11.009894
## [2,] -2.543800 18.337950
## [3,] -3.472303 14.228810
## [4,] -1.345344 15.448436
## [5,] -3.340251 12.211119
## [6,] -4.357936  9.102372
```

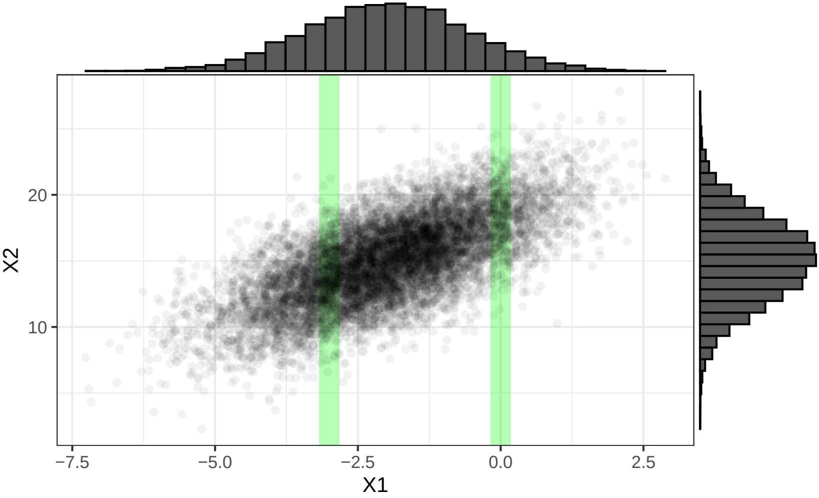

An Example of a Bivariate Normal
X2
20
10
X1
−7.5
−5.0
−2.5
0.0
2.5


An Example of a Bivariate Normal
X2
20
10
X1
−7.5
−5.0
−2.5
0.0
2.5

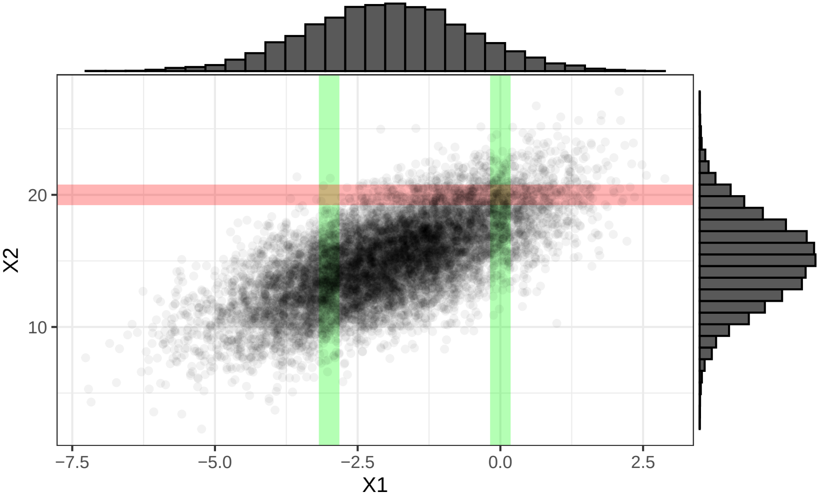
An Example of a Bivariate Normal



An Example of a Bivariate Normal



An Example of a Bivariate Normal



Multivariate Distributions, Briefly

For a bivariate pmf or pdf $f(x, y)$:

The **marginal distribution** of X is $f_X(x) = \sum_y f(x, y)$ or $f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy$. (Similar for Y.)

The **conditional distribution** of X given Y is $f_{X|Y}(x \mid y) = \frac{f(x,y)}{f_Y(y)}$. (Works similarly for Y given X.)

Notice that the fs quickly start to take on multiple meanings with joint, marginal, and conditional distributions floating around, but the context/notation usually makes it clear.

Multivariate Distributions, Briefly

X and Y are independent iff $f(x, y) = f_X(x) f_Y(x)$.

The **covariance** of X and Y (analogous to the variance) is $\mathrm{Cov}(X, Y) = E\left[(X - \mu_X)(Y - \mu_Y)\right]$.

The **correlation** of X and Y is $\rho(X, Y) = \frac{\mathrm{Cov}(X,Y)}{V(X)V(Y)}$.

Multivariaet Distributions, Briefly

For a bivariate normal distribution, the marginal and conditional distributions have really easy formulas.

Law of Large Numbers

Definition 3 (Convergence in Probability)

A sequence of random variables $Z_1, Z_2, ..., Z_n$ *converges in probability to c if* $\lim_{n\to\infty} \Pr(|Z_n - c| < \epsilon) = 1$ *for all* $\epsilon > 0$*. We write "*Z_n *converges in probability to c" as* $Z_n \xrightarrow{p} \mu$*.*

Definition 4 (Independent and Identical Distributed)

A sequence of random variables $Z_1, Z_2, ..., Z_n$ *with pdfs or pmfs* $g_1, g_2, ..., g_n$ *are independent and identically distributed (i.i.d.) if and only if two conditions hold. First, they are mutually independent, so that the joint distribution* $g(z_1, z_2, ..., z_n)$ *equals the product of the marginal distributions* $\prod_{i=1}^{n} g_{z_i}(z_i)$*. Second, they are identical, so that each pdf or pmf is the same function* $g_i = g$ *for* $i \in \{1, 2, ..., n\}$*.*

Theorem 5 ((Weak) Law of Large Numbers)

Suppose a sequence of i.i.d. random variables $X_1, X_2, ..., X_n$ *are each an i.i.d. random sample from a distribution with expected value* μ *and finite variance* σ^2*. If* $\overline{X}_n$ *denotes the average of the n samples, then* $\overline{X}_n \xrightarrow{p} \mu$*.*

Law of Large Numbers

Here's the intution: Choose any error tolerance you like. There is a random sample large enough that the average of the sample will, *for sure*, fall inside the tolerance.

Using Simulation to Compute $E(X)$

We can use the Law of Large Numbers to compute $E(X)$–we just take a "large" number of samples from the distribution $f(x)$ and take the average of those draws.

The code below shows this for $X \sim$ exponential(3).

```r
rate <- 3
1/rate  # analytical expected values
```

```
## [1] 0.3333333
```

```r
x <- rexp(100000, rate = rate)  # large number of sims
mean(x)  # avg of simulations
```

```
## [1] 0.332187
```

Exercise 5

Use `draws <- rexp(100000, rate = 0.1)` to take a large number of draws from exponential(0.1). Then compute the average-of-the-squares `mean(draws^2)` and the square-of-the-average `mean(draws^2`. Are these the same or different? Connect this simulation result to Jensen's inequality.

Central Limit Theorem

Definition 5 (Convergence in Distribution)

A sequence of random variables $Z_1, Z_2, ..., Z_n$ with cdfs $G_1, G_2, ..., G_n$ converges in distribution to Z^* with cdf G^* if $\lim_{n\to\infty} G_n(z) = G^*(z)$ at all points where z is continuous. We write "Z_n converges in distribution to Z^*" as $Z_n \xrightarrow{d} Z^*$.

Theorem 6 (Central Limit Theorem)

Suppose a sequence of i.i.d. random variables $X_1, X_2, ..., X_n$ from a distribution with finite expected value μ and finite variance σ^2. Let $\overline{X}_n = avg(X_1, ..., X_n)$. Then $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$ converges in distribution to the standard normal.

In slightly different notation, $Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$ and $Z^* \sim N(0, 1)$, then $Z_n \xrightarrow{d} Z^*$.

Central Limit Theorem

We can think of the CLT in several different ways.

- $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \overset{d}{\to} N(0,1)$
- $\sqrt{n}\left(\overline{X}_n - \mu\right) \overset{d}{\to} N\left(0, \sigma^2\right)$
- $\left(\overline{X}_n - \mu\right) \overset{d}{\to} N\left(0, \frac{\sigma^2}{n}\right)$
- $\overline{X}_n \overset{d}{\to} N\left(\mu, \frac{\sigma^2}{n}\right)$
- $\frac{X_1 + X_2 + \ldots + X_n}{n} \overset{d}{\to} N(\mu, \frac{\sigma^2}{n})$
- $(X_1 + X_2 + ... + X_n) \overset{d}{\to} N(n\mu, n\sigma^2)$ (FPP!)

Implication: If we have a large number of draws and know the expected value and variance of each draw, then we know (approximately) the distribution of the average (and sum) of the draws.

Note: FPP refer to $n\mu$ as the `expected value (for the sum)'' and $\sqrt{n \sigma^2} = \sqrt{n} \sigma$ as the`standard error (SE) (for the sum)."

Illustration of the Central Limit Theorem

```r
# a large number of bernoulli(0.1) trials
avg <- numeric(10000) # a container for the 10,000 simulations

# 10,000 times, do the following:
for (i in 1:10000) {
  # take 1,000 draws from bernoulli(0.1) distributiton
  draws <- rbinom(1000, size = 1, prob = 0.1)
  # find the avg; store it
  avg[i] <- mean(draws)
}

# put in a data frame
data <- tibble(avg)
```

Illustration of the Central Limit Theorem

```r
# CLT distribution
mu <- 0.1
sigma <- sqrt(0.1*0.9)/sqrt(1000)

# plot
ggplot(data, aes(x = avg)) +
  geom_histogram(aes(y = ..density..), binwidth = 1/1000, center = 0.5) +
  geom_function(fun = dnorm, args = list(mean = mu, sd = sigma),
                color = "red", size = 1.0)
```

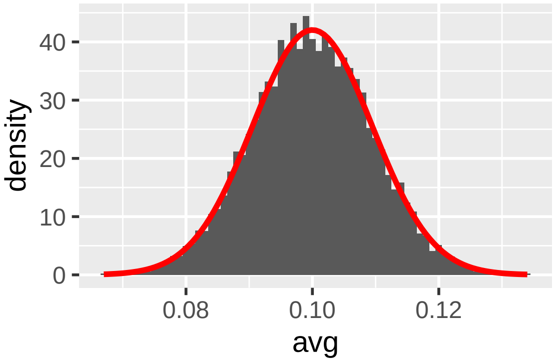


Illustration of CLT's Conv. in Prob.

To illustrate how the CLT works, let's do a little simulation with a die.

Remember this: The CLT says that the standardized sample average *converges in distribution* to the standard normal distribution as the sample size increases.

Roll a die n times. Treating sixes as 1 and not-sixes at 0. Compute the standardized sample average from the 10 rolls. Do this 10,000 times to get a good sense of the *distribution* of the standardized sample average.

This is a Bernoulli$\left(\frac{1}{6}\right)$ distribution, so we have $\mu = \frac{1}{6}$ and $\sigma = \sqrt{\frac{1}{6} \times \frac{5}{6}} \approx 0.37$.

The standarized sample average is $\dfrac{\sqrt{n}\,(\text{sample avg.} - \mu)}{\sigma}$.

Illustration of CLT's Conv. in Prob.

First, let's do it for $n = 10$ rolls of the die.

```
die <- c(0, 0, 0, 0, 0, 1)

mu <- 1/6
sigma <- sqrt((1/6)*(5/6))
n <- 10

# trial 1
s <- sample(die, size = n, replace = TRUE)
std_avg <- sqrt(n)*(mean(s) - mu/sigma); std_avg
```

```
## [1] -0.1493025
```

```
# trial 2
s <- sample(die, size = n, replace = TRUE)
std_avg <- sqrt(n)*(mean(s) - mu/sigma); std_avg
```

```
## [1] -0.4655303
```

Illustration of CLT's Conv. in Prob.

Now let's do it 10,000 times.

```
std_avgs <- numeric(10000)  # a container
for (i in 1:10000) {
  s <- sample(die, size = n, replace = TRUE)
  std_avgs[i] <- sqrt(n)*(mean(s) - mu/sigma)
}

std_avgs[1:20]
```

```
##  [1] -1.4142136 -1.0979858 -1.4142136 -1.0979858 -1.0979858 -1.09798
##  [7] -0.7817580 -1.0979858 -1.0979858 -0.7817580 -0.7817580 -0.46553
## [13] -0.7817580 -0.7817580 -1.0979858 -1.0979858 -0.4655303 -1.09798
## [19] -0.4655303 -0.7817580
```

Illustration of CLT's Conv. in Prob.

```r
qplot(std_avgs)
```

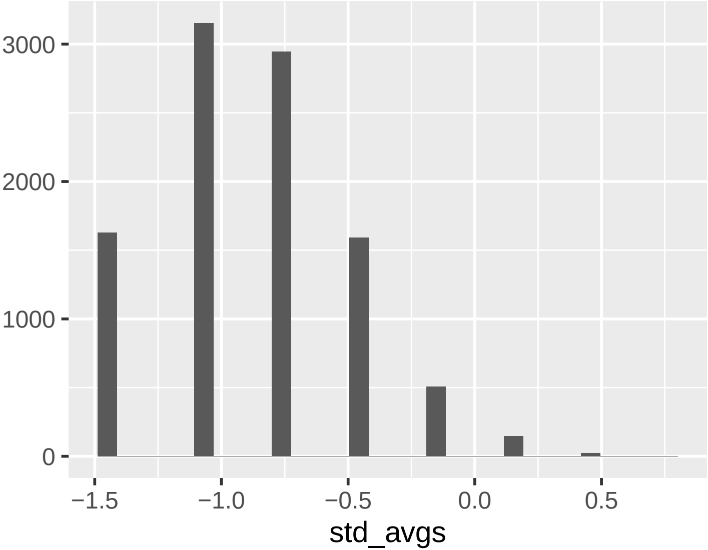


Illustration of CLT's Conv. in Prob.

Now I repeat that for different samples sizes than 10.

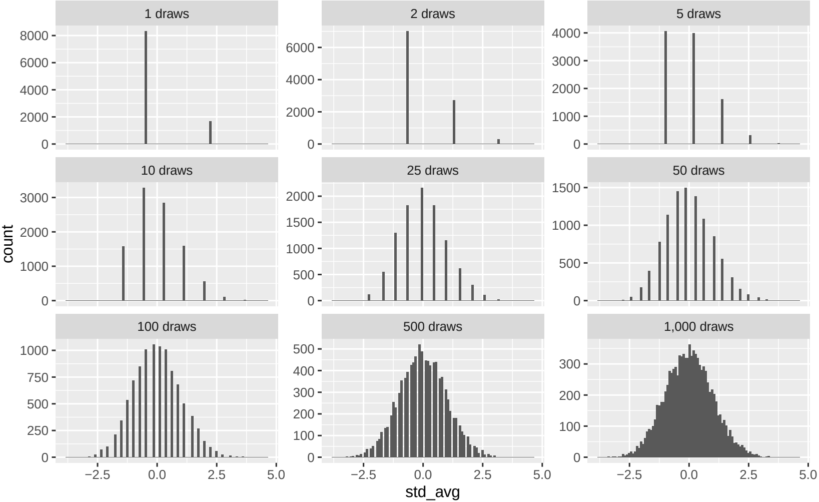
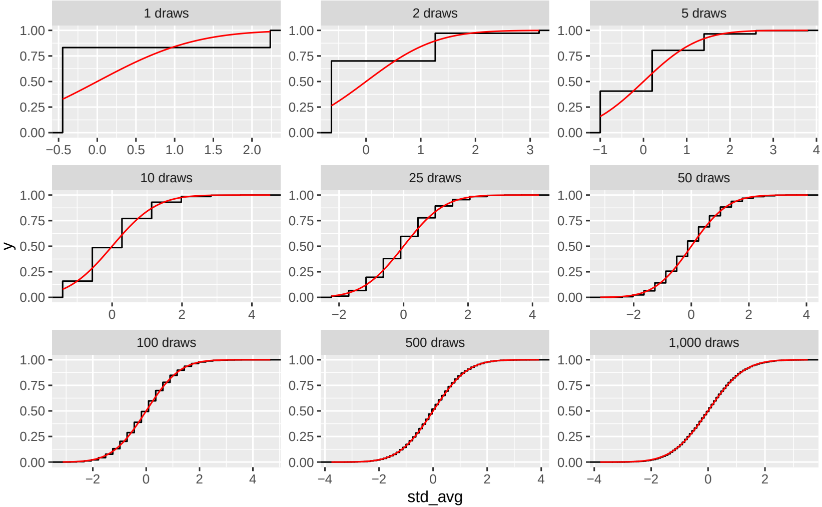


Illustration of CLT's Conv. in Prob.

It's a little easier to see the convergence if we compare the emprical cdf of the standardized sample averages to the standard normal cdf.


1 draws
2 draws
5 draws
10 draws
25 draws
50 draws
100 draws
500 draws
1,000 draws
y
std_avg