STAT 450/460

Handout 4: Functions of random variables

Fall 2016

Chapter 6: Functions of random variables

The study of statistics requires understanding the properties of estimates obtained from random samples. For example, suppose $\{Y_1, ..., Y_n\}$ denotes a n independent observations sampled at random from a population. Of interest might be the distribution of $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$. However, to understand the distribution of $\bar{Y}$, we must understand the distribution of the sum. We can think of this as finding the distribution of $U \equiv U(Y_1, Y_2, ..., Y_n)$. The following methods are commonly used for finding distributions of functions of random variables, defined generally as U:

1. The CDF method (Section 6.3). This technique is most often used if $\{Y_1, ..., Y_n\}$ are continuous. The CDF methods finds $f_U(u)$ by finding $F_U(u)$, then differentiating.

2. The transformation method (Section 6.4). This result actually follows from the CDF method. With this method, the pdf of U is found by working directly with $f_Y(y)$.

3. The MGF method (Section 6.5). This method is often used for finding distributions of sums of random variables.

The previous 3 methods are for finding the distribution of a single function U. Often, we are interested in the joint distribution of multiple functions, say $U = g(Y_1, Y_2, ..., Y_n)$ and $V = h(Y_1, Y_2, ..., Y_n)$. This requires the Jacobian methodology of Section 6.6.

The CDF method

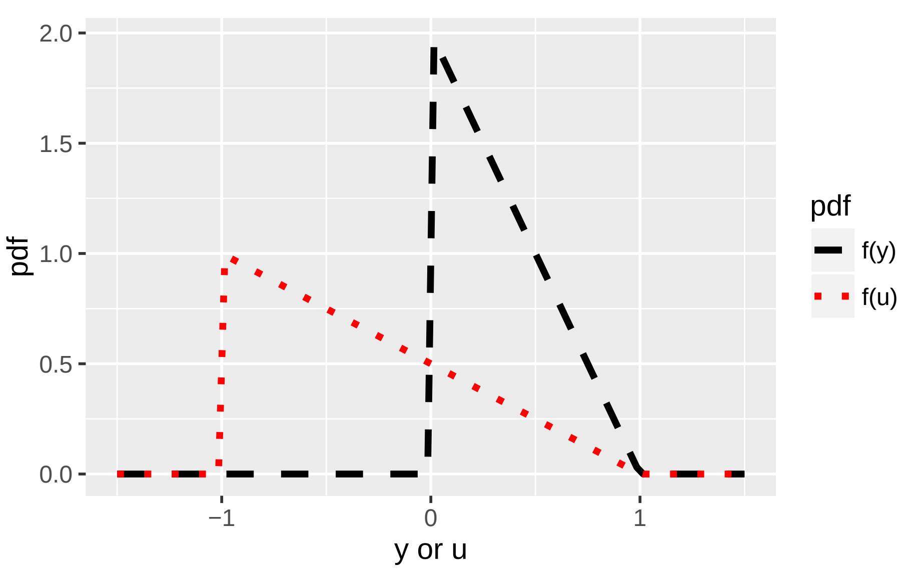
We will start with the method of distribution functions, also called the CDF method. As stated, this method depends on finding $F_U(u) = P(U \leq u)$, then differentiating. Let's consider some examples.

Example

Let Y be a random variable with the following pdf:

$$f_Y(y) = \begin{cases} 2(1-y) & 0 \leq y \leq 1 \\ 0 & otherwise \end{cases}$$

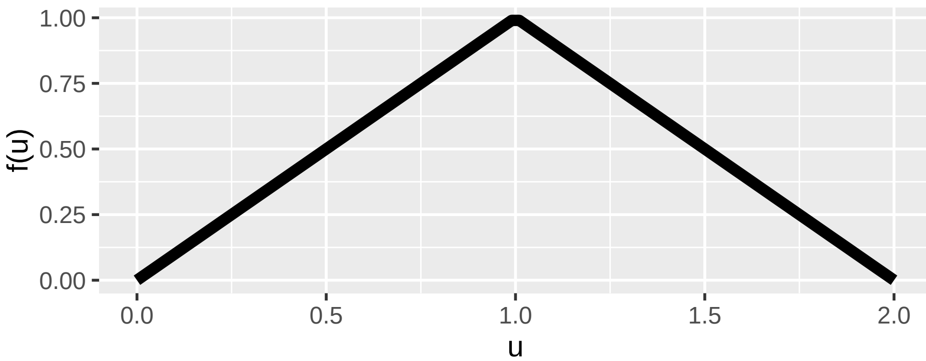
Find the distribution of $U = 2Y - 1$.



Example (Example 6.3 from book)

Let $(Y_1, Y_2) \equiv (X, Y)$ denote a single random sample of size 2 drawn independentically from the same $UNIF(0,1)$ distribution. (I.e., (Y_1, Y_2) are i.i.d $\sim UNIF(0,1)$). What is the pdf of $U = X + Y$?

Graph of $f_U(u)$:



Example: General result

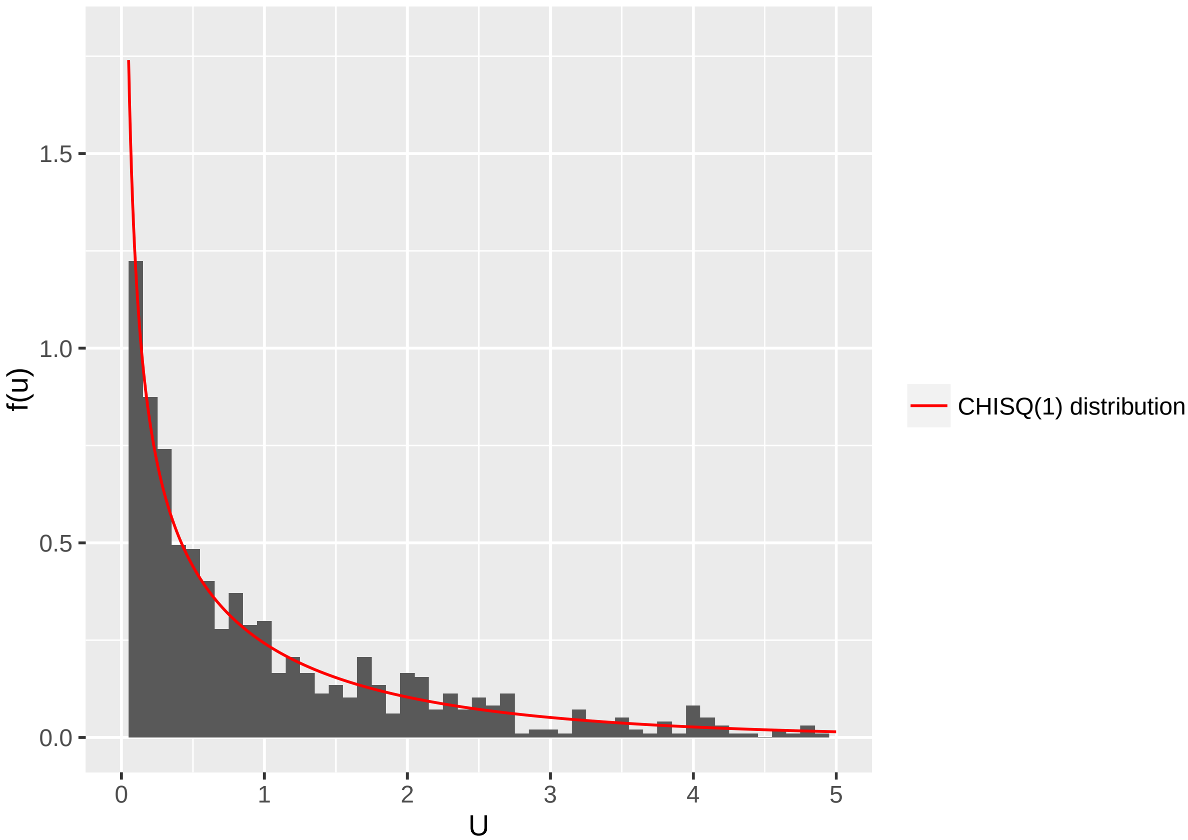
Suppose Y is a continuous random variable with pdf $f_Y(y)$ and CDF $F_Y(y)$. If $U = Y^2$, what is $F_U(u)$ and $f_U(u)$?

Example: VERY IMPORTANT application of general result

If $Z \sim N(0,1)$, what is the distribution of $U = Z^2$?

Showing this result via simulation:

```
myseq <- seq(0.05,5,l=1000)
set.seed(213642)
random.Z <- rnorm(1000)
random.U <- random.Z^2
mydata <- data.frame(Z = random.Z, U = random.U, Useq = myseq, f.u  = dchisq(myseq,df=1))
ggplot(data=mydata) + geom_histogram(aes(x = U,y=..density..),binwidth=0.1) +
    geom_line(aes(x=Useq,y=f.u,color='sdfsdf')) + xlim(c(0,5)) + ylab('f(u)') +
    scale_color_manual(name='',values='red',label='CHISQ(1) distribution')
```



Example

Let $U \sim UNIF(0,1)$. Define $Y = -\beta \ln(U)$ (where $\beta > 0$). Find the distribution of Y.

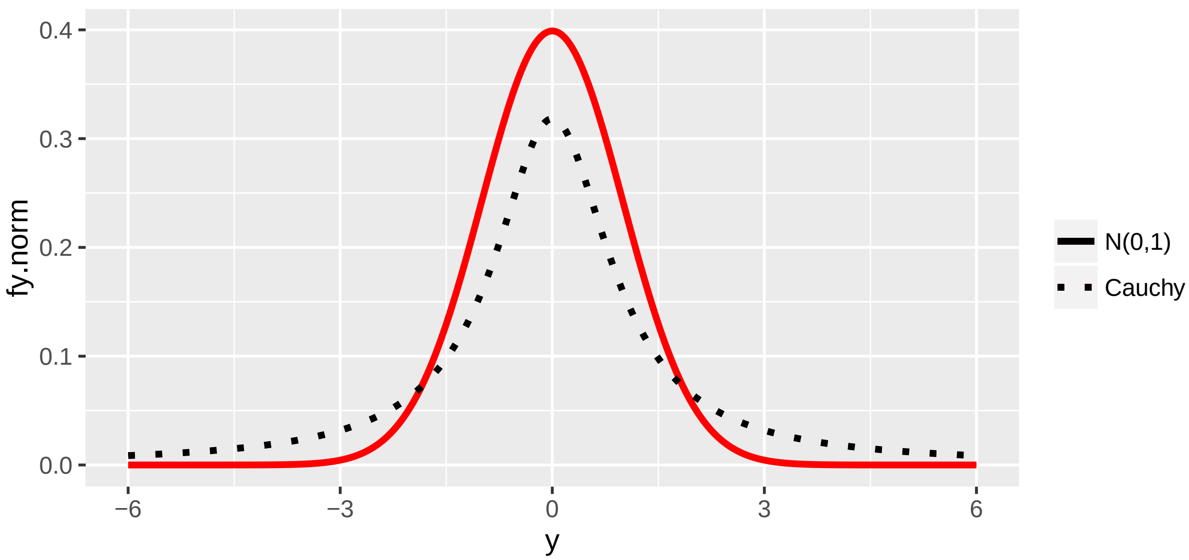
Example: General result

Suppose $U \sim UNIF(0,1)$, and define $Y = F_Y^{-1}(U)$, where $F_Y^{-1}(\cdot)$ is the inverse CDF of Y. Prove that Y will have the CDF given by $F_Y(y)$.

Example The *Cauchy distribution* is a pathological distribution often used as the "straw man" of distributions. It has the following pdf:

$$f_Y(y) = \frac{1}{\pi(1+y^2)}, \ y \in \mathcal{R}.$$

It is a symmetric bell-shaped distribution, but with much heavier tails than the $N(0,1)$. Compare:

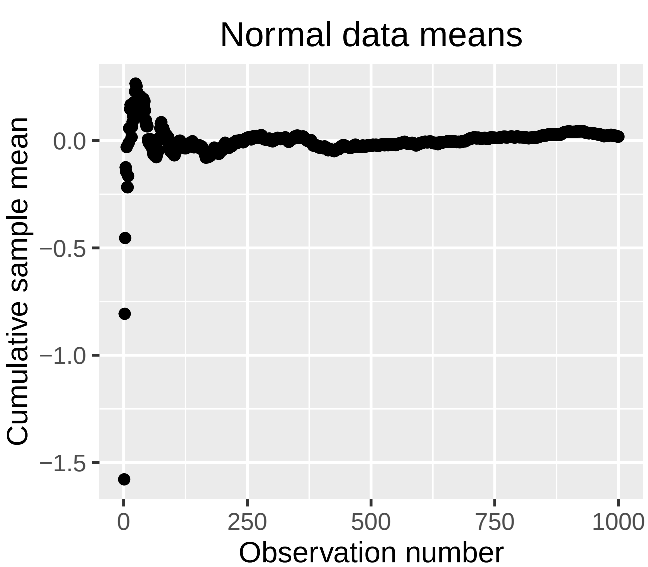


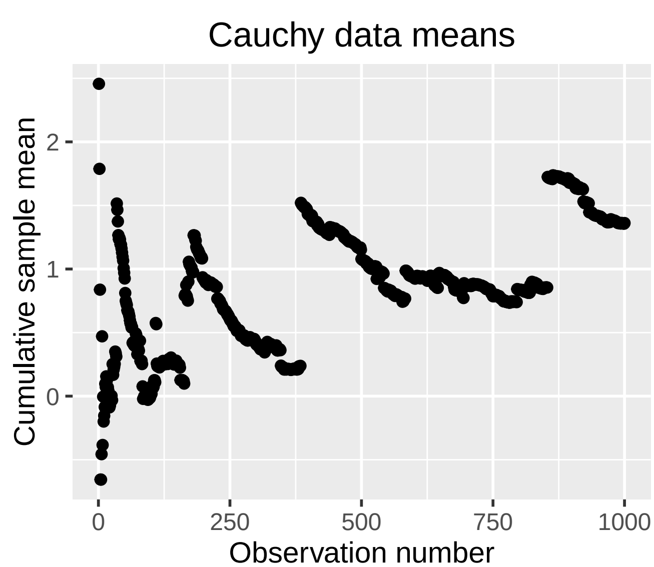
- Show that, with $U \sim UNIF(0,1)$, that $Y = \tan(\pi(U - 1/2))$ has a Cauchy distribution.

- Show that $E(Y)$ does not exist.

Simulating 1000 $N(0,1)$ and 1000 Cauchy random variables, and plotting the cumulative means:

```
set.seed(4144)
somenormaldata <-  rnorm(1000)
somecauchydata <-  rcauchy(1000)
mydata <- data.frame(obs = 1:1000, Y = somenormaldata, X = somecauchydata,
                      cummean.Y = cumsum(somenormaldata)/(1:1000),
                      cummean.X = cumsum(somecauchydata)/(1:1000))
ggplot(data = mydata) + geom_point(aes(x = obs,y = cummean.Y)) + xlab('Observation number') +
    ylab('Cumulative sample mean') + ggtitle('Normal data means')
ggplot(data = mydata) + geom_point(aes(x = obs,y = cummean.X)) + xlab('Observation number') +
    ylab('Cumulative sample mean') + ggtitle('Cauchy data means')
```

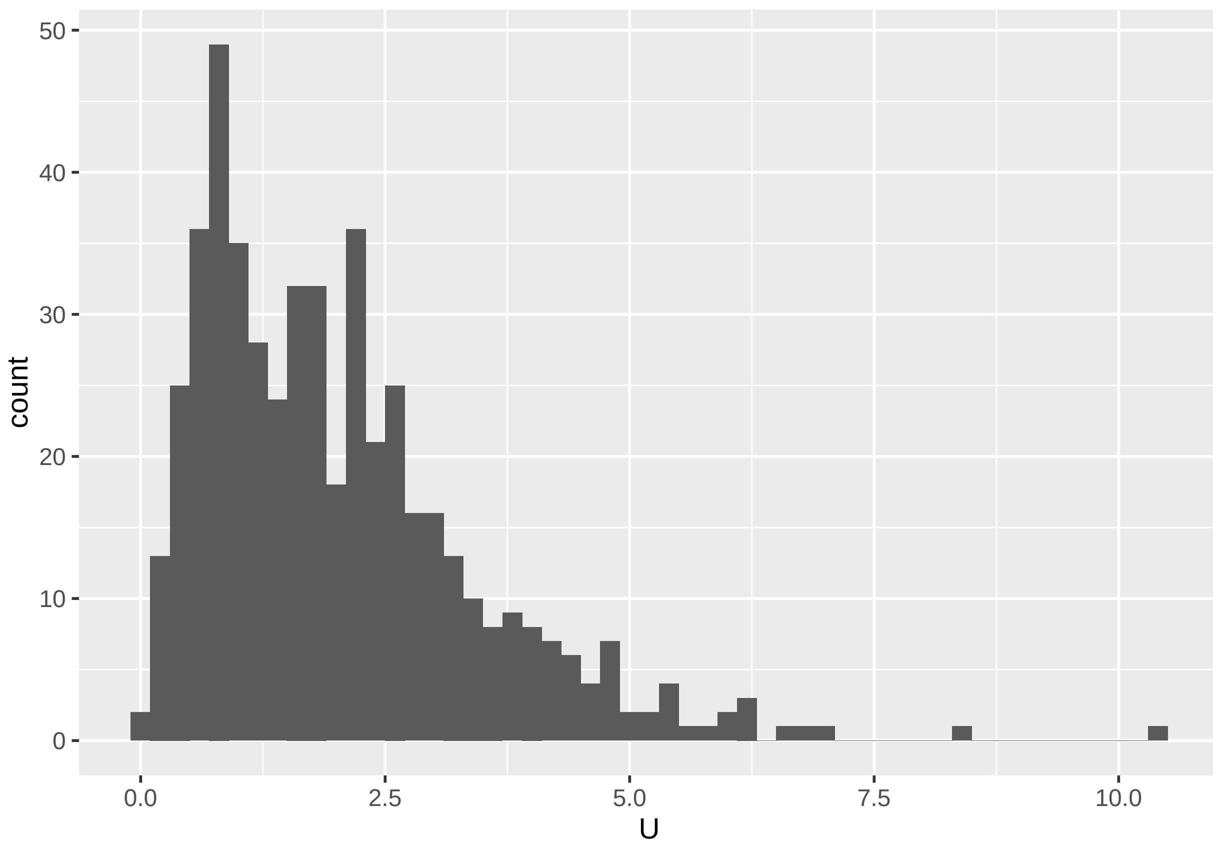




Example: Often times, the CDF method can be used to derive the distribution of a function of more than one random variable. For example, suppose X and Y are independent $UNIF(0,1)$ random variables. Find the distribution of $U = -\ln(XY)$.

Let's do some simulations to see what we might be looking for:

```
set.seed(11)
X <- runif(500)
Y <- runif(500)
mydata <- data.frame(X,Y,U = -log(X*Y))
ggplot(data=mydata) + geom_histogram(aes(x=U),binwidth=0.2)
```



Looks kind of like a Gamma! Let's verify.

The pdf method

The "pdf method" is derived from the CDF method. Let $U = g(Y)$ where $g(\cdot)$ is a strictly monotone decreasing or monotone increasing function. Suppose also that $g(\cdot)$ is differentiable. Then:

$$f_U(u) = f_Y(g^{-1}(u)) \cdot \left| \frac{d}{du} g^{-1}(u) \right|$$

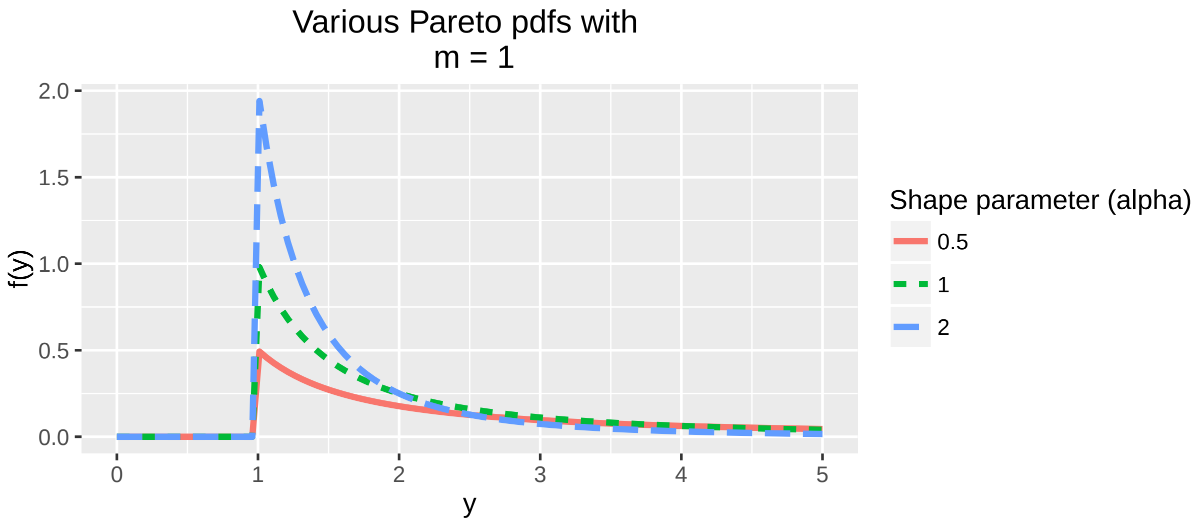
Proof:

Example

The *Pareto distribution* is named for the Italian Vilfredo Pareto. He originally used it to model income inequality: a small percent of people own the greatest amount of wealth. If X has a Pareto distribution parameterized by α and m ($X \sim Pareto(\alpha, m)$), then:

$$f_X(x) = \begin{cases} \frac{\alpha m^\alpha}{x^{\alpha+1}} & x \geq m \\ 0 & otherwise \end{cases}$$

Some pictures:



Suppose $Y \sim EXP(\lambda)$, so:

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

Show that $U = me^Y$ has a Pareto(λ, m) distribution.

Related example

Suppose $Y \sim Pareto(1, \lambda)$. Find the distribution of $U = \log(Y)$.

MGF method

The method of moment generating functions (MGFs) is another way to find the distribution of functions of random variables. It is the method used to prove **THE MOST IMPORTANT THEOREM IN STATISTICS** (arguably): the Central Limit Theorem. With the MGF method, we begin to truly see the difference between the distribution of *individual observations*, and distributions of *statistics computed from a sample.*

Recall the definition of the MGF:

1. $M_Y(t) = E(e^{tY})$

2. MGFs are unique; hence $M_{Y_1}(t) = M_{Y_2}(t)$ implies Y_1 and Y_2 have the same distribution

Point 2 is the key: If we can find the MGF of a function of random variables (say U) and compare it to well-known MGFs, we can show U follows certain well-known distributions.

Example Let $U = aY + b$ for constants $a \neq 0$ and b. Find the MGF of U.

Example The MGF method is especially important for finding the distribution of sums of random variables. Specifically, let $\{Y_1, Y_2, ..., Y_n\}$ be an i.i.d. (independent and identically distributed) vector of random variables. E.g., $Y_i \sim EXP(\beta)$ for each i. Let $U = \sum_{i=1}^{n} Y_i$. Find the MGF of U, $M_U(t)$.

Example Let $Y_i \sim POI(\lambda_i)$, with $Y_i \perp\!\!\!\perp Y_j$. Find the distribution of $U = \sum_{i=1}^{n} Y_i$. What happens if the Y_i are i.i.d. $\sim POI(\lambda)$?

Example Let Y_i be an i.i.d. sample with all $Y_i \sim EXP(\beta)$. Find the distribution of $U = \sum_{i=1}^{n} Y_i$. **Demonstrate this result with simulations.**

Example Let $Y_1, Y_2, ..., Y_n$ be an i.i.d sample drawn from a $N(\mu_i, \sigma_i^2)$ population. Find the distribution of $U = \sum_{i=1}^{n} Y_i$. **Demonstrate this result with simulations.**

Example Consider the previous example. Find the distribution of $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}U$.

Example Consider the previous example. Find the distribution of $Z = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$.

Joint transformations

Often, we are interested in the joint transformation of two or more variables. We will consider bivariate transformations only, though these concepts extend more generally.

Suppose we have two random variables X and Y with joint pdf $f_{X,Y}(x,y)$. Consider two different functions of (X,Y): $U = g(X,Y)$ and $V = h(X,Y)$. For example, we might have $g(X,Y) = X+Y$ and $h(X,Y) = \frac{X}{X+Y}$. We are then interested in the joint distribution of (U,V), $f_{U,V}(u,v)$.

By definition, this joint distribution is given by:

$$f_{U,V}(u,v) = f_{X,Y}(g^{-1}(u,v), h^{-1}(u,v))|J|,$$

as long as $|J| \neq 0$, where:

$$|J| = \begin{vmatrix} \frac{\partial g^{-1}}{\partial u} & \frac{\partial g^{-1}}{\partial v} \\ \frac{\partial h^{-1}}{\partial u} & \frac{\partial h^{-1}}{\partial v} \end{vmatrix} = \frac{\partial g^{-1}}{\partial u}\frac{\partial h^{-1}}{\partial v} - \frac{\partial h^{-1}}{\partial u}\frac{\partial g^{-1}}{\partial v}$$

Example Suppose X and Y are i.i.d $\sim EXP(\beta)$ with:

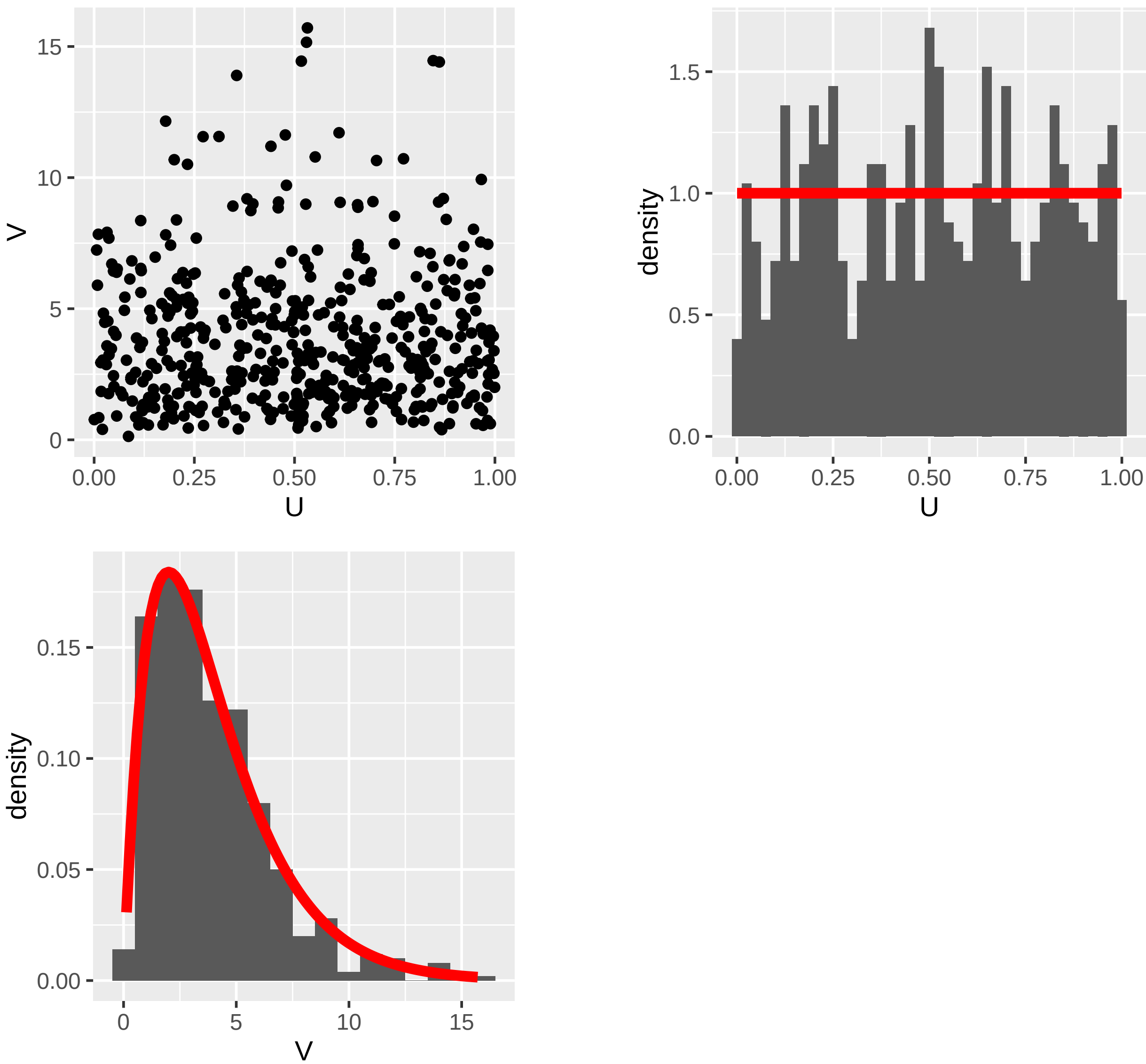
$$f_{X,Y}(x,y) = \begin{cases} \lambda^2 e^{-\lambda(x+y)} & x,y > 0 \\ 0 & otherwise \end{cases}$$

Let $U = \frac{X}{X+Y}$ and $V = X+Y$. Find $f_{U,V}(u,v)$. Also find the marginal distributions of U and V.

(previous example, continued)

Simulating this and superimposing the theoretical pdfs, with $\beta = 2$:

```
X <- rexp(500,rate=1/2)
Y <- rexp(500,rate=1/2)
U <- X/(X+Y)
V <- X+Y
seq1 <- seq(0,15,l=500)
seq2 <- seq(0,1,l=500)
mydata <- data.frame(X,Y,U,V,seq1,seq2)
ggplot(data=mydata) + geom_point(aes(x=U,y=V))
ggplot(data = mydata,aes(x=U)) +
      geom_histogram(aes(y=..density..)) +
      stat_function(fun = dunif,geom='line',col='red',size=2)
ggplot(data = mydata,aes(x=V)) +
      geom_histogram(aes(y=..density..)) +
      stat_function(fun = dgamma,geom='line',col='red',size=2,args=list(shape=2,scale=2))
```



Example Suppose X and Y are i.i.d $\sim UNIF(0,1)$. Let $U = \sqrt{-2\ln(Y)}\cos(2\pi X)$ and $V = \sqrt{-2\ln(Y)}\sin(2\pi X)$. Find the joint distribution of (U,V) as well as the marginal distributions of U and V.

Let's do some simulations to get a feel for what we might be after. Run the following code:

```
X <- runif(500,min=0,max=1)
Y <- runif(500,min=0,max=1)
U <- sqrt(-2*log(Y))*cos(2*pi*X)
V <- sqrt(-2*log(Y))*sin(2*pi*X)
seq1 <- seq(-2,2,l=500)
mydata <- data.frame(X,Y,U,V,seq1)
ggplot(data=mydata) + geom_point(aes(x=U,y=V))
ggplot(data = mydata,aes(x=U)) +
      geom_histogram(aes(y=..density..)) +
      stat_function(fun = dnorm,geom='line',col='red',size=2)
ggplot(data = mydata,aes(x=V)) +
      geom_histogram(aes(y=..density..)) +
      stat_function(fun = dnorm,geom='line',col='red',size=2)
```

(previous example, continued)

Distributions of order statistics

Consider a random sample $Y_1, Y_2, ..., Y_n$ drawn i.i.d from some population. We are often interested in distributions of *order statistics* or functions of order statistics. Let $Y_{(1)} < Y_{(2)} < ... < Y_{(n)}$ denote these order statistics; from here it is obvious that $Y_{(1)} = min(Y_1, Y_2, ..., Y_n)$ while $Y_{(n)} = max(Y_1, Y_2, ..., Y_n)$. Some additional order statistics (or functions thereof) include:

1. The median;
$$M = \begin{cases} Y_{\left(\frac{n+1}{2}\right)} & \text{if n is even} \\ \frac{Y_{\left(\frac{n}{2}\right)} + Y_{\left(\frac{n+1}{2}\right)}}{2} & \text{if n is odd} \end{cases}$$
2. The range; $R = Y_{(n)} - Y_{(1)}$
3. The k^{th} percentile; $Y_{(\lfloor nk/100 \rfloor)}$
4. The IQR; $Y_{(\lfloor 75n/100 \rfloor)} - Y_{(\lfloor 25n/100 \rfloor)}$

The joint distribution of $(Y_{(1)}, Y_{(2)}, ... Y_{(n)})$ can be found by beginning with the joint distribution of the *unordered* $(Y_1, Y_2, ..., Y_n)$. Recall that since the Y_i are i.i.d, that:

$$f_{Y_1,...,Y_n}(y_1, ..., y_n) = \begin{cases} \prod_{i=1}^{n} f_Y(y_i) & y_i \in \text{Support} \\ 0 & otherwise \end{cases}$$

However, there are $n!$ ways to order the Y_i, so the joint distribution of the *order statistics* becomes:

$$f_{Y_{(1)},...,Y_{(n)}}(y_1, ..., y_n) = \begin{cases} n! \prod_{i=1}^{n} f_Y(y_i) & y_1 < y_2 < ... < y_n \\ 0 & otherwise \end{cases}$$

Example
Suppose Y_1, Y_2, Y_3 are drawn i.i.d. from a population with pdf:

$$f_Y(y) = \begin{cases} 2y & 0 \leq y \leq 1 \\ 0 & otherwise \end{cases}$$

Find the joint distribution of (Y_1, Y_2, Y_3) and show that it integrates to 1. Also find the joint distribution of $(Y_{(1)}, Y_{(2)}, Y_{(3)})$ and show that *it* integrates to 1.

We are often also interested in the marginal distribution of specific order statistics. We will now consider the distribution of the minimum order statistic $Y_{(1)}$, the maximum $Y_{(n)}$, and the general j^{th} order statistic $Y_{(j)}$.

Find the distribution of the minimum order statistic, $Y_{(1)}$.

Find the distribution of the maximum order statistic, $Y_{(n)}$.

Find the distribution of the j^{th} order statistic, $Y_{(j)}$.

Example Suppose $Y_1, Y_2, ..., Y_n$ are i.i.d. $\sim EXP(\beta)$. Find the distribution of $Y_{(1)}$, $Y_{(n)}$, and $Y_{(j)}$.

Joint distribution of $(Y_{(i)}, Y_{(j)})$

Example Suppose $Y_1, Y_2, ..., Y_n$ are i.i.d. $\sim EXP(\beta)$. Find the distribution of $R = Y_{(n)} - Y_{(1)}$. Also verify the correct pdf by showing that $f_R(r)$ integrates to 1.

(previous example, continued)

Let's investigate this with simulation. Note that this is **very importantly** different from previous simulations. We are interested in the distribution of R across *repeated samples*, not for observations within a single sample!

```
f <- function(x,n,beta) {
  tt <- ((n-1)/beta) * exp(-x/beta)*(1-exp(-x/beta))^(n-2)
  }

many.ranges <- function(n,beta) {
  one.sample <- rexp(n,rate=1/beta)
  one.range <- max(one.sample)-min(one.sample)
  return(one.range)
}

n5 <- replicate(1000,many.ranges(n=2,beta=1))
n10 <- replicate(1000,many.ranges(n=5,beta=1))
n15 <- replicate(1000,many.ranges(n=20,beta=1))
n20 <- replicate(1000,many.ranges(n=100,beta=1))
mydata <- data.frame(Ranges = c(n5,n10,n15,n20),
                      n = rep(c(2,5,20,100),each=1000),
                      myseq=rep(seq(0,10,l=1000),4))
mydata$f.r <- f(mydata$myseq,n = mydata$n,beta = 1)
ggplot(data=mydata) + geom_histogram(aes(x=Ranges,y=..density..),binwidth = .5) +
      geom_line(aes(x = myseq,y = f.r,color=as.factor(n)),size=2)+
      facet_wrap(~n)+ scale_color_discrete(name='Size of each sample')
```