Transforamtions and Combinations of Random Variables

Stat 241

Creating new random variables from old

Transformations and combinations allow us to create new random variables from old.

Examples

- Toss a pair of fair dice. Let X be the result on the first die and Y the result on the second die. Then $S = X + Y$ is the sum of the two dice and $P = XY$ is the product.
- Let F be the temperature of a randomly selected object in degrees Fahrenheit. Then $C = 5/9(F - 32)$ is the object's temperature in degrees Celsius.
- A random sample of n adult females is chosen. Let X_i be the height (in inches) of the ith person in the sample. Then $\overline{X} = \frac{X_1+X_2+\cdots X_n}{n}$, is the sample mean.

Important Rules for Combinations and Transformations

General Question: If X is the result of combining two or more known random variables or of transforming a single random variable, what can we know about the distribution of X?

Rules for transforming and combining random variables

0. $\mathrm{Var}(X) = \mathrm{E}(X^2)\,\check{}\,\mathrm{E}(X)^2$

- Not really about transformation and combinations, but useful to remember.

1a. $\mathrm{E}(X + b) = E(X) + b$ and $\mathrm{Var}(X + b) = \mathrm{Var}(X)$.

1b. $\mathrm{E}(aX) = aE(X)$ and $\mathrm{Var}(aX) = a^2\,\mathrm{Var}(X)$.

1. $\mathrm{E}(aX + b) = a\,\mathrm{E}(X) + b$

- The expected value of a linear transformation is the linear transformation of the expected value.

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

- The expected value of a sum is the sum of the expected values.

3. If X and Y are independent, then $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$.

- The expected product ist the procuct of the expected values – provided the random variables are independent.

4. If X and Y are independent, then $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

- The variance of a sum is the sum of the variances – provided the random variables are independent.

These rules are not too difficult to prove, but we will focus mainly an how to use the rules.

Here are two example proofs (for the continuous case). We will omit the limits of integration to focus attention on the use of simple rule for integration.

$$\mathrm{E}(X+b) = \int (x+b) f(x)\, dx = \int x f(x)\, dx + b \int f(x)\, dx = \mathrm{E}(X) + b$$

$$\mathrm{E}(aX) = \int a x f(x)\, dx = a \int x f(x)\, dx = a\,\mathrm{E}(X)$$

The rules involving more than one random variable require double integrals, but they are also straightforward. The rules for discrete random variables involve sums (and double sums) instead of integrals.

Independent Random Variables

X and Y are **independent random variables** if the distribution of X is the same for each value of Y and vice versa. This is equivalent to saying that

$$\mathrm{P}(X \leq x \,\mathrm{and}\, Y \leq y) = \mathrm{P}(X \leq x) \cdot \mathrm{P}(Y \leq y)$$

for all x and y.

Independence Examples

- If a pair of fair dice are tossed, X is the value of the first die, and Y is the value of the second, then X and Y are independent.
- If X and Y are the height and weight of a randomly selected person, then X and Y are not independent. (The distribution of weights is different for taller people compared to the distribution for shorter people.)

Examples

In the examples below we will compare simulations to the rules above.

Two Dice

Let X and Y be the values on two fair dice. Look at the sum and product: $X + Y$ and XY.

```
TwoDice <-
  tibble(
    die1 = resample(1:6, 10000),
    die2 = resample(1:6, 10000),
    S = die1 + die2,
    P = die1 * die2
  )
```

```
mean(~ die1, data = TwoDice)
```

```
## [1] 3.4908
```

```
mean(~ die2, data = TwoDice)
```

```
## [1] 3.5356
```

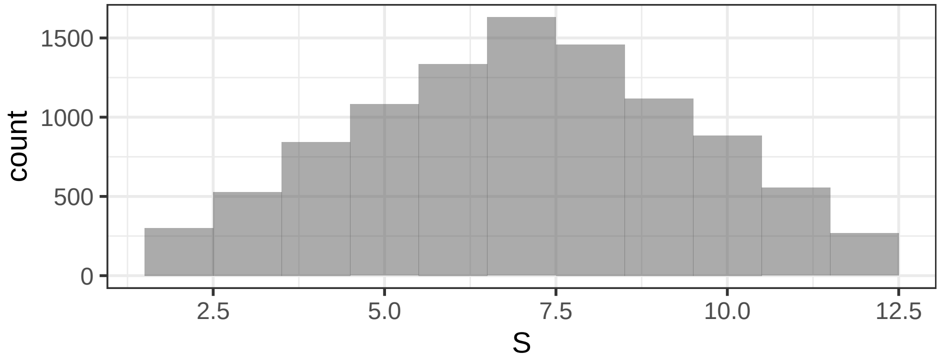
```
mean(~ S, data = TwoDice)
```

```
## [1] 7.0264
```

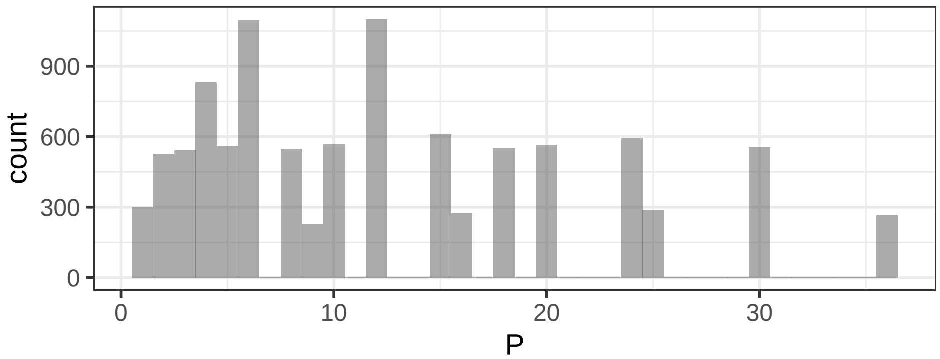
```
mean(~ P, data = TwoDice)
```

```
## [1] 12.3483
```

```
gf_histogram(~ S, data = TwoDice, binwidth = 1)
```




```
gf_histogram(~ P, data = TwoDice, binwidth = 1)
```



Uniform

Let X and Y be independent random variables that have the uniform distribution on [0,1]. Again, let's look at the sum and product.

```
TwoUnif <-
  tibble(
    X = runif(10000, 0, 1),
    Y = runif(10000, 0, 1),
    S = X + Y,
    P = X * Y
  )
```

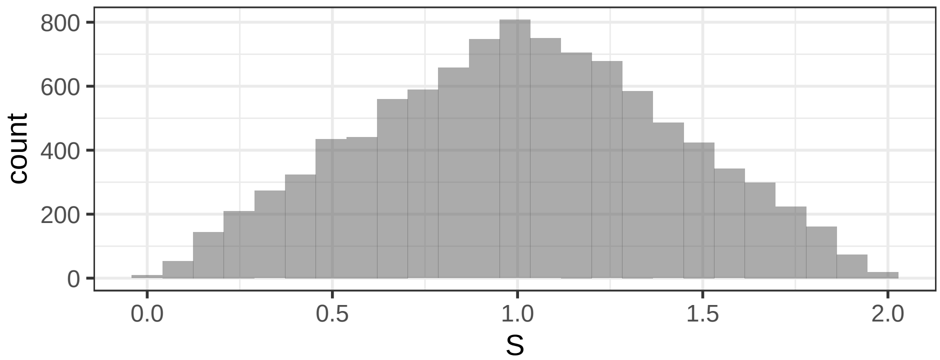
```
mean(~ S, data = TwoUnif)
```

```
## [1] 1.006544
```

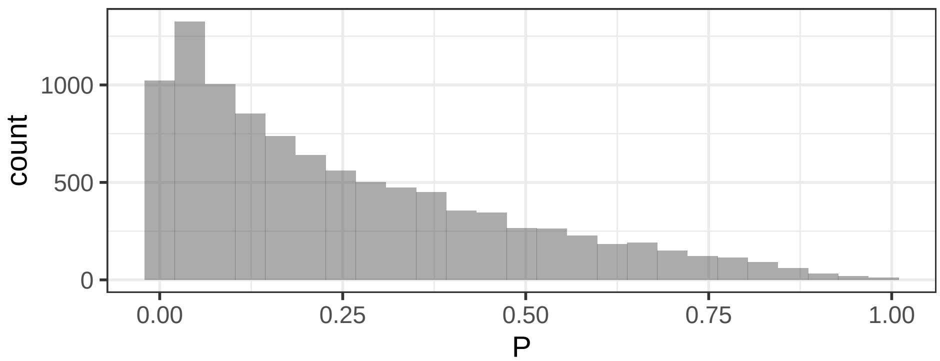
```
mean(~ P, data = TwoUnif)
```

```
## [1] 0.2525068
```

```
gf_histogram(~ S, data = TwoUnif)
```




```
gf_histogram(~ P, data = TwoUnif)
```



Build your own examples

You can do a similar thing with any distributions you like. It even works if X and Y have different distributions!

Linear Combinations

We can combine the rules above to build two more rules.

Rules for transforming and combining random variables (continued)

5. $\mathrm{E}(a_1X_1 + a_2X_2 + \cdots a_nX_n) = a_1\,\mathrm{E}(X_1) + a_2\,\mathrm{E}(X_2) + \cdots a_n\,\mathrm{E}(X_n)$

- The expected value of a linear combination is the linear combination of the expected values.

6. If $X_1, X_2, \ldots, X_n$ are indepenent, then

$$\mathrm{Var}(a_1X_1 + a_2X_2 + \cdots a_nX_n) = a_1^2\,\mathrm{Var}(X_1) + a_2^2\,\mathrm{Var}(X_2) + \cdots a_n^2\,\mathrm{Var}(X_n)$$

- Note the squaring.
- We can write this in terms of standard deviation if we like

$$\mathrm{SD}(a_1X_1 + a_2X_2 + \cdots a_nX_n) = \sqrt{a_1^2\,\mathrm{SD}(X_1)^2 + a_2^2\,\mathrm{SD}(X_2)^2 + \cdots a_n^2\,\mathrm{SD}(X_n)^2}$$

- This is sometimes called the Pythagorean identity for standard deviation. The independence assumption is analogous to the assumption that the triangle has a right angle.

Normal Distributions are special

Fact 1 Any linear transformation of a normal random variable is normal.

Fact 2 Any linear combination of *independent* normal random variables is normal

Example
If $X \sim \mathsf{Norm}(1, 1)$, $Y \sim \mathsf{Norm}(-1, 2)$, $W \sim \mathsf{Norm}(5, 4)$, and X, Y, W are independent, find the distribution of $C = 2X + 3Y + W$.

By Fact 2, C will be normal. We just need to work out the mean and variance.

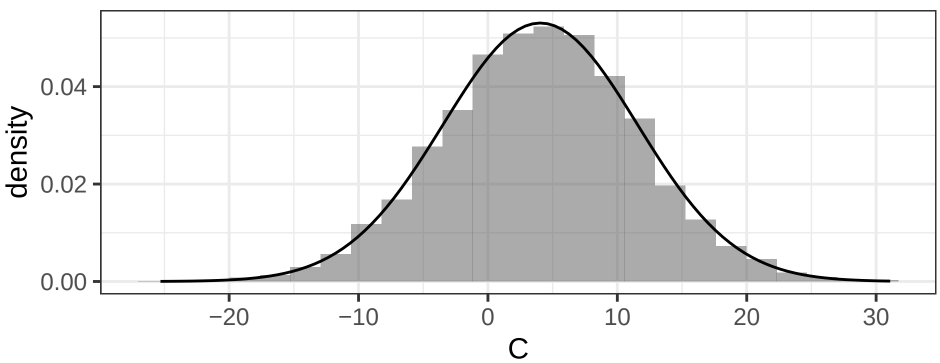
- $\mathrm{E}(C) = 2 \cdot 1 + 3 \cdot (-1) + 5 = 4$
- $\mathrm{Var}(C) = 4 \cdot 1 + 9 \cdot 4 + \cdot 16 = 56$
- So $C \sim \mathsf{Norm}(4, 20.98)$

Let's compare to some simulations

```
NormalExample <-
  tibble(
    X = rnorm(10000, 1, 1),
    Y = rnorm(10000, -1, 2),
    W = rnorm(10000, 5, 4),
    C = 2 * X + 3 * Y + W
  )
df_stats(~ C, data = NormalExample, mean, sd)
```

```
##   response     mean       sd
## 1        C 4.045686 7.519876
```

```
gf_dhistogram( ~ C, data = NormalExample) %>%
  gf_fitdistr(~ C, data = NormalExample, "norm")
```



We used `gf_fitdistr()` to overlay a normal density curve on our density histogram above. What is that doing?

It is based on a function called `fitdistr()` which (as you can probably guess) fits distributions to data. Basically, `fitdistr()` is looking for the best fitting distribution in a given family for a given data set. `fitdistr()` is a little bit awkward to use becuase

- it doesn't use the formula interface we are used to,
- it sometimes uses different abbreviations for the distributions than we are used to.

But otherwise it is simple to use. Here it tells us what it thinks are the best fitting normal distributions in our previous example.

```
fitdistr(NormalExample$X, "normal")
```

```
##       mean           sd
##   0.991422114   1.007772233
##  (0.010077722) (0.007126026)
```

```
fitdistr(NormalExample$Y, "normal")
```

```
##       mean          sd
##   -0.97969889    1.99990948
##  ( 0.01999909) ( 0.01414150)
```

```r
fitdistr(NormalExample$W, "normal")
```

```
##       mean          sd
##   5.00193842   3.99621302
##  (0.03996213) (0.02825749)
```

```r
fitdistr(NormalExample$C, "normal")  # this is the one that was added to our plot
```

```
##       mean          sd
##   4.04568598   7.51949978
##  (0.07519500) (0.05317089)
```

How does `fitdistr()` work?

There are two important ways to fit distributions to data:

1. Maximum likelihood

- Choose the parameter values (mean and sd for a normal distribution) that would make the data more likely to occur than any other values. This is an optimization problem.

2. Method of moments

- Choose the parameter values that make the mean or mean and standard deviation of the data equal to the mean and standard deviation of the distribution.
  - for a normal distribution, this is very easy since the parameters are just the mean and standard deviation. For other distributions, we may have to solve a system of equations. Sometimes this is easy, sometimes not so easy.

`fitdistr()` uses the maximum likelihood method – numerical optimization. But in the case of a normal distrbution, both methods give the same result (up to some round off for the numerical optimization).

```r
fitdistr(NormalExample$C, "normal")  # this is the one that was added to our plot
```

```
##       mean          sd
##   4.04568598   7.51949978
##  (0.07519500) (0.05317089)
```

```r
# method of moments
df_stats(~ C, data = NormalExample, mean, sd)
```

```
##   response     mean       sd
## 1        C 4.045686 7.519876
```

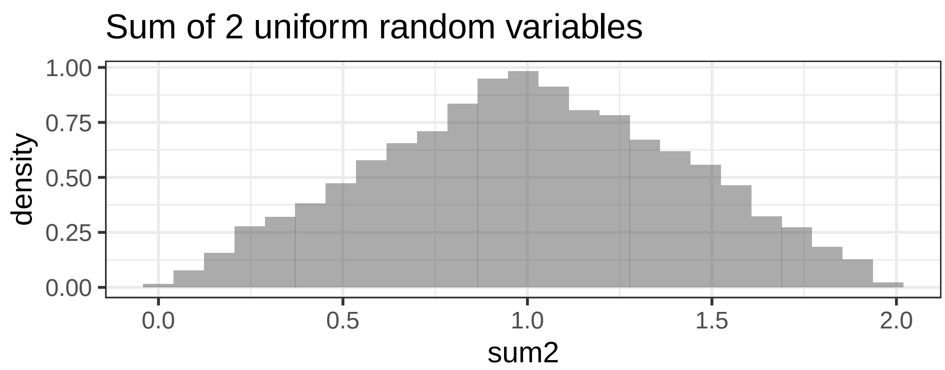
`fitdistr()` knows about several other families of distributions as well. See `?fitdistr` for the list and the names `fitdistr()` uses. Note that or some distributions, you will need to provide a reasonable starting guess for the parameters values.

Fact 3 If $X_1, X_2, \dots X_n$ are independent random variables that have the same distribution, then the sum will be approximately normal, no matter what distribution X_i has, provided n is "large enough".
The approximation gets better and better as n increases.

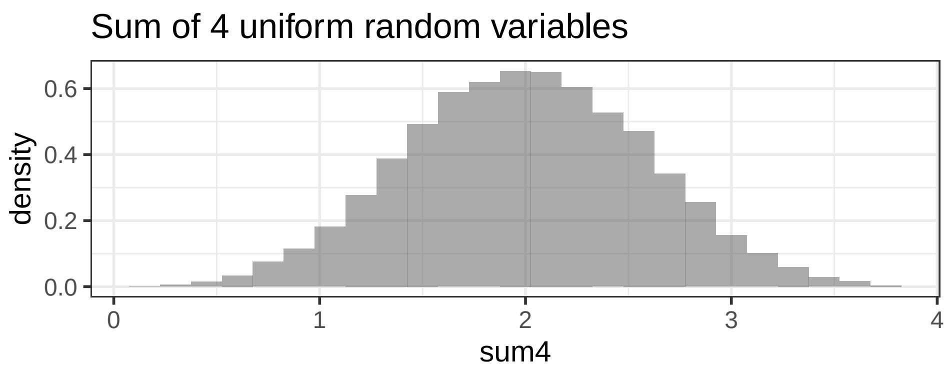
Simulation

```r
sim1 <- runif(10000, 0, 1)
sim2 <- runif(10000, 0, 1)
sim3 <- runif(10000, 0, 1)
```

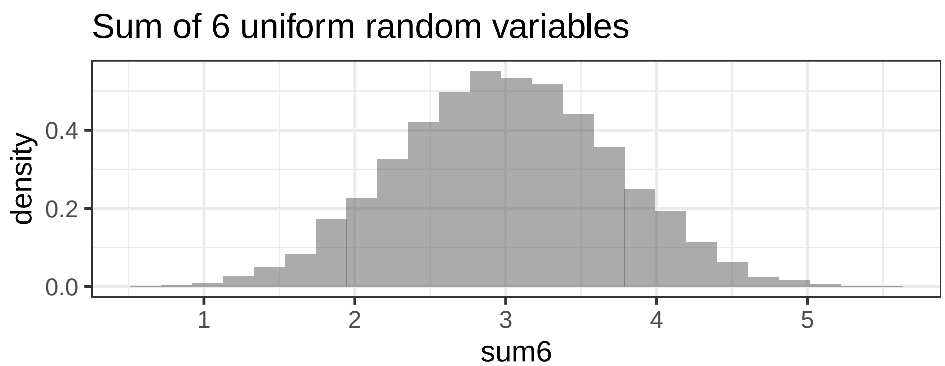
```r
sim4 <- runif(10000, 0, 1)
sim5 <- runif(10000, 0, 1)
sim6 <- runif(10000, 0, 1)
sum2 <- sim1 + sim2
sum4 <- sim1 + sim2 + sim3 + sim4
sum6 <- sim1 + sim2 + sim3 + sim4 + sim5 + sim6
gf_dhistogram( ~ sum2, title = "Sum of 2 uniform random variables")
```




```r
gf_dhistogram( ~ sum4, title = "Sum of 4 uniform random variables")
```



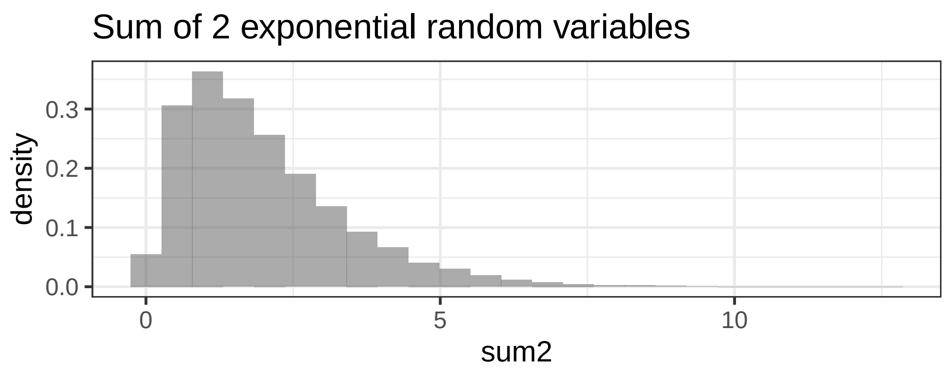

```r
gf_dhistogram( ~ sum6, title = "Sum of 6 uniform random variables")
```



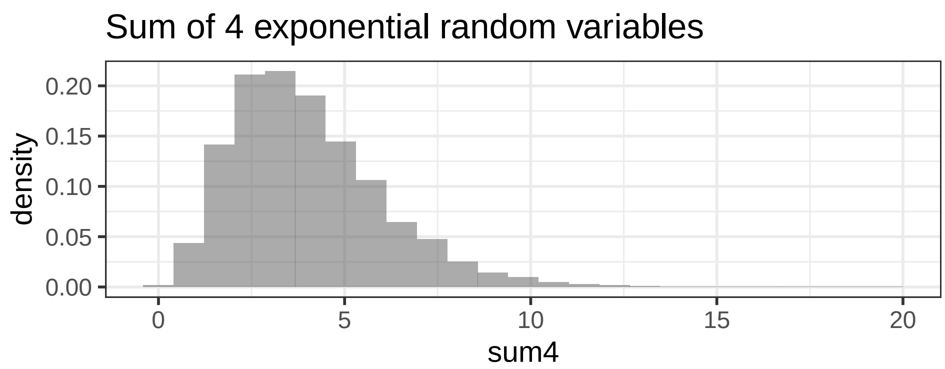
It takes longer for an exponential distribution, but it still converges to normal pretty quickly.

```r
sim1 <- rexp(10000, 1)
sim2 <- rexp(10000, 1)
sim3 <- rexp(10000, 1)
sim4 <- rexp(10000, 1)
sim5 <- rexp(10000, 1)
sim6 <- rexp(10000, 1)
```

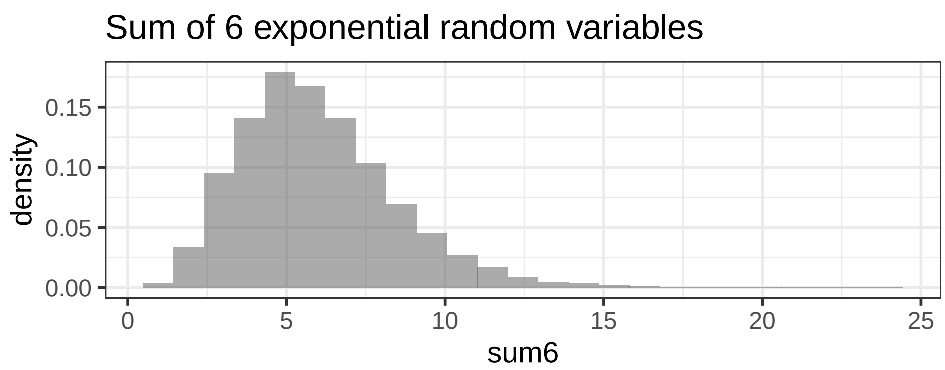
```
sim7 <- rexp(10000, 1)
sim8 <- rexp(10000, 1)
sum2 <- sim1 + sim2
sum4 <- sim1 + sim2 + sim3 + sim4
sum6 <- sim1 + sim2 + sim3 + sim4 + sim5 + sim6
sum8 <- sim1 + sim2 + sim3 + sim4 + sim5 + sim6 + sim7 + sim8
gf_dhistogram( ~ sum2, title = "Sum of 2 exponential random variables")
```



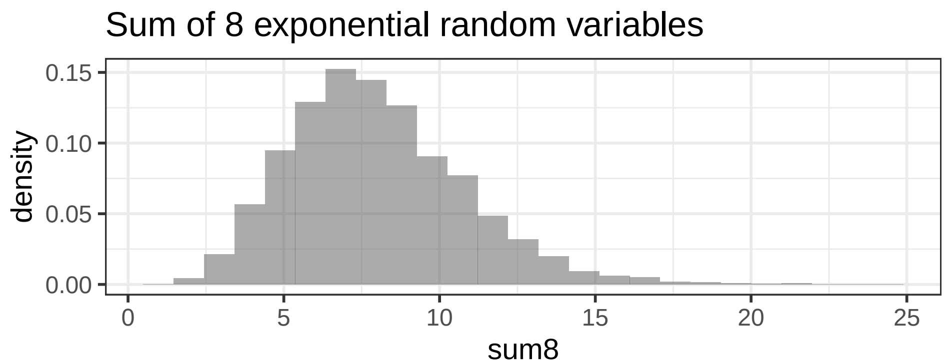

```
gf_dhistogram( ~ sum4, title = "Sum of 4 exponential random variables")
```




```
gf_dhistogram( ~ sum6, title = "Sum of 6 exponential random variables")
```




```
gf_dhistogram( ~ sum8, title = "Sum of 8 exponential random variables")
```



What does this have to do with statistics?

Often we are interested the mean of something. The mean can be written as

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \cdots X_n}{n}$$

If each X_i is randomly selected from the same population, then the numerator is a sum of independent and identially distributed (iid) random variables, so...

1. Our rules tell us the mean and variance (and standard deviation) of $\overline{X}$.
2. $\overline{X}$ will be approximately normal, provided our sample is large enough.
3. This means we can approximate probability about $\overline{X}$, no matter what distribution the X_i's come from!

Some Practice

1. If each X_i has a mean of μ and a standard deviation of σ, and the X_i are independent, fill in the question marks below:

$$\overline{X} \approx \mathsf{Norm}(?, ?)$$

This is arguably the most important result in all of statistics and is referred to as the **Central Limit Theorem**.

2. If X and Y are independent random variables, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$. What about $\mathrm{Var}(X - Y)$? Your first guess might be that $\mathrm{Var}(X - Y) = \mathrm{Var}(X)\,\breve{}\,\mathrm{Var}(Y)$. But this cannot be true, since if $\mathrm{Var}(Y) > \mathrm{Var}(X)$ we would have $\mathrm{Var}(X - Y) < 0$, but variance is always non-negative.

What is the correct rule? [Hint: use the rules we have.]

$$\mathrm{Var}(X - Y) = ??$$

3. Suppose $X \sim \mathsf{Norm}(10, 3)$, $Y \sim \mathsf{Norm}(6, 2)$, and X and Y are independent.

a. What is the distribution of $2X\breve{}Y$?
b. What is the distribution of $X + Y$?
c. What is $\mathrm{P}(X \leq 4)$?
d. What is $\mathrm{P}(Y \geq 2)$?
e. What is $\mathrm{P}(1 \leq 2X - Y \geq 4)$?
f. What is $\mathrm{P}(X - Y \leq 4)$?
g. What is $\mathrm{P}(X \leq Y)$? (Hint: $\mathrm{P}(X \leq Y) = \mathrm{P}(X - Y \leq 0)$. Use the distribution of $X - Y$.)

4. Suppose X is a random variable with mean = 6 and sd = 2, Y is a random variable with mean = -1 and sd = 2, and X and Y are independent.

a. What are the mean and standard deviation of $2X - Y$?

b. What are the mean and standard deviation of $3X + 4Y$?

5. Let X and Y be independent $\mathsf{Gamma}(shape = 3, rate = 4)$ random variables.
Let $S = X + Y$ and $D = X - Y$. Is a gamma distribution a good fit for S? Use a simulation with 10000 repetitions to test this. For each (sum and product):

- Use `gf_fitdistr()` to compare your histogram to the best fitting Gamma distribution.
- If it fits well, use `fitdistr()` to get the shape and rate parameters for the best fit.

6. So, it looks like the sum of two independent gamma random variables with the same shape and scale also has a gamma distribution.

We would not expect the difference to have a gamma distribution, since values or D can be negative and a gamma RV cannot be negative. Use simulations to see whehter it looks like D is normal.

7. Suppose $X \sim \mathsf{Gamma}(3, 4)$ and $W \sim \mathsf{Gamma}(1, 5)$ and W is independent of X. Use simulations to check whether it is reasonable to conclude that $X + W$ has a gamma distribution.

8. Another important distribution in applications is the Chi-squared distribution. The Chi-squared distribution is a special case of the Gamma distribution. If n is a positive integer, then X has a Chi-square distribution with n degrees of freedom if X has a Gamma distribution with shape = $n/2$ and rate = $1/2$, i.e., $\mathsf{Chisq}(n) = \mathsf{Gamma}(n/2, 1/2)$. The abbreviation for Chi-squared in R is `chisq`.

Let Z be the standard normal random variable; i.e., $Z \sim Norm(0, 1)$. What kind of distribution does Z^2 have? One of the following is true about Z^2. * It has a normal distribution or * It has a Chi-squared distribution with 1 degree of freedom. Use a simulations to determine which of these two alternatives is correct.