To do the homework, you’ll have to know how the Weibull distribution is parameterized in R, which might not be very clear from the documentation in help(pweibull).

The function has pweibull() has three parameters, one for the value of the variable (i.e., \( x \) in \( P(X \leq x) \)), one for the shape parameter, and one for the scale parameter. The shape parameter will determine how skewed the distribution, and the scale parameter has a big effect on the mean, although both parameters have some effect on the mean.
We know that the Weibull reduces to the exponential when $\alpha = 1$. This is the shape parameter, which is the second parameter of the `pweibull()` function. Thus, when $\alpha = 1$, we should get different exponential distributions.

The next question is how the scale parameter, $\lambda$ is treated in R. If $\alpha = 1$ and $\lambda = 10$, then using the notation, we should get exponential random variables with rate 10, and the average of several should be close to 0.1. In R, you have to determine if the scale parameter should be $\lambda$ or $1/\lambda$. You could do this by generating several Weibull random variables with the scale parameter close to 10 and seeing what it is doing. Or you could use the density evaluated at 0, since this should be $10e^{-10\cdot0} = 10$ for an exponential with rate 10. Thus, you could use `rweibull()` to figure this out or `dweibull()`.
Adding legends to plots

For the homework, it would be good to add a legend to make your plots more readable.
Adding legends to plots

```r
x <- 1:15
plot(x,1/x,xlab="x",ylab="f(x)",ylim=c(0,1),cex.lab=1.3,cex.axis=1.4)
points(x,1/x^2,type="l",col="blue",cex=3,pch=15) #pch ignored
points(x,1/x^3,type="l",col="orange",cex=3,pch=16) #because line
legend(5,1,legend=c("1/x","1/x^2","1/x^3"),fill=c("black","blue","orange"),
cex=1.6,bty="o",bg="grey90")
```
Adding points and lines to a plot

x <- 1:15
plot(x,1/x,xlab="x",ylab="f(x)",ylim=c(0,1),cex.lab=1.3,cex.axis=1.4)
points(x,1/x^2,type="l",col="blue",cex=3,pch=15) #pch ignored
points(x,1/x^3,type="l",col="orange",cex=3,pch=16) #because line
points(x,1/x^2,col="blue",cex=3,pch=15) #pch ignored
points(x,1/x^3,col="orange",cex=3,pch=16) #because line
points(x,1/x,cex=3)
legend(5,1,legend=c("1/x","1/x^2","1/x^3"),fill=c("black","blue","orange"),
cex=1.6,bty="o",bg="grey90")
Try copying and pasting the previous code into R and modifying bits and pieces. It might be easiest to copy and paste the code into a file called `plot.r` or something similar and type `source(file = "plot.r")`. The code should be in the same directory, or you can try an environment like R Studio to have the code nearby.

Try `help(plot)`, `help(plot.default)`, and `help(legend)` for more plotting options. You can also try to search for “R graphics gallery” online to see examples of R graphics with source code that you can modify.
Mean residual life

The mean residual life is the expected remaining lifetime given that someone has survived up to time $x$. This is

$$mrl(x) = E(X - x | X > x)$$

You can also think of this as the area under the survival curve to the right of $x$ divided by $S(x)$. (See next slide)

For conditional expectations generally, we have

$$E(X|X > x) = \sum_{k>x} kP(X = k | X > x) = \frac{\sum_{k>x} kP(X = k, X > x)}{P(X > x)} = \frac{\sum_{k>x} kP(X = k)}{P(X > x)}$$

The analogous expression for a continuous random variable is

$$E(X|X > x) = \frac{\int_x^\infty uf(u) du}{P(X > x)}$$
Mean residual life

Note that \( \text{mrl}(0) = E(X - 0|X > 0) = E(X) \) if \( X \) is a positive random variable. Generally,

\[
E(X - x|X > x) = \frac{\int_{x}^{\infty} (t - x)f(t) \, dt}{P(X > x)}
\]

You can use integration by parts to show that the numerator is \( \int_{x}^{\infty} S(t) \, dt \) (I’m using \( t \) instead of \( u \) because I use \( u \) for integration by parts). Plugging in \( x = 0 \) gives us that

\[
E(X) = \int_{0}^{\infty} S(x) \, dx
\]
Integration by parts

You should verify for yourself that the integration by parts works out. Use the formula

\[ \int u \, dv = uv - \int v \, du \]

where \( u \) and \( v \) are functions of \( t \). Let

\[ dv = f(t) \, dt \]
\[ v = -S(t) \]
\[ u = t - x \]
\[ du = dt \]
The memoryless property of the exponential says that

\[ P(X > x + z | X > x) = P(X > x) \]

If we apply this to the mean residual life, this means that

\[ E(X - x | X > x) = E(X) = 1/\lambda \]

so the mean residual life is constant.
The $p$th quantile is the value $x_p$ such that

$$S(x_p) = 1 - p$$

or equivalently that $F(x_p) = p$. This is implemented in R using functions such as `qexp()`, `qweibull`, etc.

As an example, the median of a distribution is the value $x_m$ such that $F(x_m) = S(x_m) = 0.5$, and this is found in R using, for example `qexp(.5, rate=3)` (median of an exponential with rate 3). The 99th percentile is found using `qexp(.99, rate=3)`.
This can sometimes be found analytically. For example the $p$th quantile for an exponential satisfies

$$S(x_p) = 1 - p \Rightarrow e^{-\lambda x} = 1 - p \Rightarrow -\lambda x = \log(1 - p) \Rightarrow x = \frac{-\log(1 - p)}{\lambda}$$

For the median, this is $\frac{-\log(1/2)}{\lambda}$. Recalling properties of logarithms, that $\log x^a = a \log x$, we have $x_m = \frac{\log 2}{\lambda}$ for the median of an exponential distribution.
Other distributions

We’ll look at both parametric and nonparametric models for survival data, but parametric models are more common. Knowing which model best fits the data can help understand the data better than a non-parametric model (Table, p. 38)

Common models include:

- exponential
- Weibull
- gamma and generalized gamma
- lognormal
- log logistic
- exponential power
- normal
- Gompertz
- inverse Gaussian
- Pareto
Chapter 2 gives an overview of some model-based techniques such as regression and models for competing risks. We’ll move on to Chapter 3 to discuss censoring in more detail.
Censoring

Generally, censoring occurs when the value for a time is only known within an interval. We can distinguish right-censoring, where we know that $x_i > x$, left-censoring, where we know that $x_i < x$, and interval censoring, where we know that $a < x_i < b$.

Right-censoring is the kind of censoring we talked about the first week of class, where a patient drops out of study perhaps due to moving or unrelated death.

Left-censoring could occur if you wanted to model the time of onset of a disease to death, where the time of death is known but the time of onset is estimated. Another example might be length of pregnancy, where the exact start of the pregnancy is only estimated.

Interval censoring occurs when both left- and right-censoring occur.
Right censoring

We’ll start looking at Type I censoring. In the simplest case, we think of the censoring time as fixed at the end of the study, with all patients entering the study at the same time.

Figure 3.1 Example of Type I censoring
Right censoring

The notation used here is that the observed time is either $X_i$ or $C_{ri}$ depending on whether the $i$th observation is observed or censored. $C_r$ is used to stand for right-censoring.

Type I censoring can be generalized so that different individuals have different start times.
Figure 3.3  Generalized Type I censoring when each individual has a different starting time
Right censoring

This can be dealt with by treating each patient as starting at time 0 but with individualized fixed censoring times.

Figure 3.4 Generalized Type I censoring for the four individuals in Figure 3.3 with each individual starting time backed up to 0. $T_1 = X_1$ (death time for first individual) ($\delta_1 = 1$); $T_2 = C_{r2}$ (right censored time for second individual) ($\delta_2 = 0$); $T_3 = X_3$ (death time for third individual) ($\delta_3 = 1$); $T_4 = C_{r4}$ (right censored time for fourth individual) ($\delta_4 = 0$).
Another complication is to have multiple planned fixed censoring events. In an animal study, only a subset of animals might be followed for a longer period of time in order to reduce costs of the study.
In Type II right-censoring, a study continues until a fixed number $r$ of individuals experiences the event, at which point the study is terminated. For example, if $r = n/2$, then this will tell you the median time to failure without having to wait for all items to fail. This can be used in animal studies (again to save money) or for testing equipment (median time to failure for lightbulbs).

In this type of censoring, if you think of the smallest failure times, $X_{(1)}$, $X_{(2)}$, etc., then these are called order statistics. Here we observe the $r$ smallest failure times, but not the failure times of the remaining $n - r$ times. The smallest $r$ failure times are the order statistics $X_{(1)}, \ldots, X_{(r)}$, and these can be used to estimate the distribution of $X$ without observing all data using maximum likelihood and theoretical results for order statistics.
To reduce costs, you can also generalize type II censoring so that after $r$ items have failed, a subset of the remaining $n - r$ items is followed until $r_1$ of these fail. The process can be reiterated through multiple rounds.
Competing risks censoring

Here accidental death or other events might remove a subject at a random time from a study, which is different from censoring due to the study ending. This generally requires independence between the random censoring time and time to failure.

Most studies have a combination of random (competing risks) censoring and type I censoring.
Representing censoring mathematically

Mathematically, it is convenient to think of the event as occurring at some time \( X \), whether or not it is observed, and to let another random variable, say \( \delta \) indicate whether or not censoring occurs. If censoring occurs at a fixed time, \( C_r \), then let \( T = \min(X, C_r) \). Then we observe \((T, \delta)\) instead of \( X \). That is we observe a time \( T \), and we know whether or not censoring has occurred, but we don’t directly observe \( X \). Here, you can let \( \delta = 0 \) if censoring occurs and otherwise \( \delta = 1 \) (or vice versa).

Note that in this notation, \( C_r \) is not random if the censoring time is fixed, even though we used a capital letter. (I’m following the notation in the book.)
Here we let $C_l$ be a left-censoring time, meaning that the event occurred at some time $X < C_l$, where the exact value for $X$ is unknown. Observations can again be represented as pairs $(T, \varepsilon)$ where $\varepsilon = 1$ if the observation is not left-censored, and otherwise $\varepsilon = 0$, and $T = \max(X, C_l)$ (as opposed to using the minimum for right-censoring).

Examples of left-censoring include

- determining when a child learns a task after enrolling in a study, where some of the children already know the task when first enrolled
- asking when a patient first had a certain symptom before some event, where the patient knows they’ve had the symptom since at least a certain number of months ago
- wanting to measure time of onset of disease to an event, where we only know the time of the first diagnosis
Truncation

Truncation seems similar to censoring, but is not quite the same. The idea is that certain individuals are observed only when their time comes within a certain window.

An example is a survival analysis where patients receive Social Security. Since there is a minimum age for Social Security recipients, the analysis only includes patients that are at least old enough to receive social security. In this example, the data is left-truncated and we must use conditional probabilities — probabilities of survival given that the patient is at least old enough to receive Social Security.

Right-truncation might occur if you only observe deaths, even though there might be cases where deaths did not occur, but you do not know how many of these occurred.

Mathematically, the main difference between censoring and truncation is that truncation requires using conditional probabilities.
Likelihoods for censored data

To model survival data, we consider the probability of the data (or joint density of the data) as a function of the parameters in the model. The idea is then to find the particular values of the parameters that best the data, meaning that they make the data more probable, or less surprising.

The technique is called **maximum likelihood**, which is worth reviewing.
Suppose I flip a coin 100 times, and I observe heads 90 times. What is the best explanation for this data?
Maximum Likelihood

If the coin was fair, it would be very surprising to observe 90 heads out of 100. If the coin was biased, then it would not be as surprising. There is still some probability of observing tails, so we wouldn’t think that the coin always came out heads. If there is some probability of the coin coming out heads, then $P(H) = .9$ is the best guess, and this is the maximum likelihood estimate.

We can think of a maximum likelihood estimate as an inference to the best explanation of the data.
For the coin flipping example, we could also consider all possible parameters $p$ for the probability of heads, and compute the probability of the data as a function of $p$. Then we would have

$$L(p) = P(90 \text{ heads}) = \binom{100}{90} p^{90}(1 - p)^{10}$$

Recall that $\binom{a}{b} = \frac{a!}{b!(a-b)!}$ and that $a! = 1 \times 2 \times \cdots \times a$, with $0! = 1$. Also recall that $\binom{a}{b} = \binom{a}{a-b}$, so $\binom{100}{90} = \binom{100}{10}$. If we plot this function, it will have a maximum at $p = 0.9$. 
To plot the likelihood in R, compute the probability as a function of $p$. We might want to try all values of $p$ from 0 to 1 in increments of 0.01, so

```r
> p <- seq(0,1,.01)
> L <- choose(100,90)*p^90*(1-p)^10
> plot(p,L,type="l",xlab="p",ylab="Likelihood",cex.lab=1.3,
cex.axis=1.4,main="Likelihood for coin flipping",cex.main=1.4)
```
The likelihood function

Likelihood for coin flipping

![Graph showing the likelihood function for coin flipping. The x-axis represents the probability (p) ranging from 0.0 to 1.0, and the y-axis represents the likelihood ranging from 0.00 to 0.12.]
In many cases, the likelihood function must be maximized numerically in the computer. We often work with the (natural) logarithm of the likelihood rather than the likelihood function itself. There are two main reasons for this:
Maximum Likelihood

We often work with the log-likelihood because

- Probabilities (or density values) of any particular observed data are typically very small for large numbers of observations. We are not so much interested in the absolute probabilities of the data as the relative probabilities for different parameters. The probability of observing exactly 9000 heads in 10000 flips is very small even for $p = .9$, but is relatively speaking much larger for $p = .9$ than for $p = .5$. Logarithms help numerically in the computer to deal with very small numbers. You can easily work with cases that have probabilities (or densities) less than $10^{-10000}$, which would be too small to represent directly in the computer. But their logarithms can be represented in the computer.

- Mathematically, probabilities and likelihoods are often products, and since the log of a product is the sum of the logs, the logarithm is often easier to work with mathematically. To maximize a function, you can maximize the log of the function rather than the function itself, and maximizing a sum is easier than maximizing a product.
The log-likelihood function

Log Likelihood for coin flipping

![Graph showing the log likelihood function for coin flipping. The x-axis represents the probability (p) and the y-axis represents the log likelihood. The graph illustrates the likelihood of observing different outcomes for a coin flip as the probability of heads (p) varies.]
Maximum Likelihood

To illustrate the point about numerical likelihoods, consider the coin flipping example with 10,000 coin flips, 9000 of which are heads. If I try this in R, I get

```r
> L <- choose(10000,9000)*p^9000*(1-p)^1000
> L

[1] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN
[19] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN
[37] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN
[55] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN
[73] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN
[91] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN

> L <- p^9000*(1-p)^1000
> L

[1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```
The numbers are too big for R to work with directly. Taking logarithms directly also doesn’t help

```r
> log(choose(10000,9000))
[1] Inf
> L <- log(p^9000*(1-p)^1000)
> L
```

```
[1] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
[16] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
[31] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
[46] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
[61] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
[76] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
[91] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
```
The problem is that we are trying to take logarithms of things that are already undefined. Instead, we need to manipulate the probabilities with logarithms first before trying to evaluate these large exponents and binomial coefficients.

The log-likelihood is

\[
\log L(p) = \log \left( \binom{10000}{9000} p^{9000} (1 - p)^{1000} \right) \\
= \log \left( \binom{10000}{9000} \right) + 9000 \log(p) + 1000 \log(1 - p)
\]

An important point to realize is that \( \log \left( \binom{10000}{9000} \right) \) doesn’t depend on \( p \), so maximizing \( \log L(p) \) is equivalent is to maximizing \( 9000 \log(p) + 1000 \log(1 - p) \). This way, we don’t have to evaluate this very large binomial coefficient.
Log-Likelihoods

When the likelihood is multiplied by a constant that doesn’t depend on the parameter, we sometimes ignore the constant. Thus, we might write

\[ L(p) \propto p^{9000}(1 - p)^{1000} \]

or even just drop the constant altogether. So sometimes, you’ll see

\[ L(p) = p^{9000}(1 - p)^{1000} \]

even though this isn’t the probability of the data. The constant changes the scale on the \( y \)-axis, but doesn’t change the shape of the curve or the value on the \( p \) (horizontal) axis where the maximum occurs.

Now we can plot and evaluate \( 9000 \log(p) + 1000 \log(1 - p) \) in R even though we can’t evaluate \( p^{9000}(1 - p)^{1000} \) directly (even though they are mathematically equivalent).
The log-likelihood function

![Graph of the log-likelihood function.](image)
Maximizing the likelihood function

In some cases, we can maximize the likelihood function analytically, usually using calculus techniques. For the binomial case, we can take the derivative of the likelihood or log-likelihood function and set it equal to 0 to find the maximum.

\[
\frac{d}{dp} \log L(p) = \frac{d}{dp} \left\{ \log \binom{n}{k} + k \log p + (n - k) \log(1 - p) \right\} = 0
\]

\[
\Rightarrow 0 + \frac{k}{p} - \frac{n - k}{1 - p} = 0
\]

\[
\Rightarrow (1 - p)k = p(n - k)
\]

\[
\Rightarrow k - kp - np + kp = 0
\]

\[
\Rightarrow k = np
\]

\[
\Rightarrow p = \frac{k}{n}
\]
Since \( p = \frac{k}{n} \), the proportion of successes, maximizes \( \log L(p) \), and therefore the likelihood as well, the maximum likelihood estimator for \( p \) is \( \hat{p} = \frac{k}{n} \). We say estimator for the general function that works for any data, and estimate for a particular value like \( \hat{p} = 0.9 \).
Maximum likelihood for the exponential

Suppose you have 3 lightbulbs that last 700, 500, and 1100 hours. Assuming that their lifetimes are exponentially distributed with rate $\lambda$, what is the maximum likelihood estimate of $\lambda$?
Likelihoods with censoring and truncation

For survival analysis, we need likelihood functions that incorporate censoring. A general framework is to have separate densities and probabilities for cases of complete observations, censored observations, and truncated observations. Assuming that all observations are independent, we can write the likelihood as the product of densities and probabilities from all of these cases.
Likelihoods with censoring and truncation

In the most general set up, you can allow different types of functions:

\[ f(x) \] exact lifetimes/death times

\[ S(C_r) \] right-censored observations

\[ 1 - S(C_l) \] left-censored observations

\[ S(L) - S(R) \] interval-censored observations

\[ \frac{f(x)}{S(Y_L)} \] left-truncated observations

\[ \frac{f(x)}{1 - S(Y_R)} \] right-truncated observations

\[ \frac{f(x)}{S(Y_L) - S(Y_R)} \] interval-truncated observations
Likelihoods with censoring and truncation

For censored (but not truncated) data, the overall likelihood is

\[
\prod_{i \in D} f(x_i) \prod_{i \in R} S(C_{ri}) \prod_{i \in L} S(C_{li}) \prod_{i \in I} [S(L_i) - S(R_i)]
\]

where \(D\) is the set of death times, \(R\) is the set of right-censored observations, \(L\) is the set of left-censored observations, and \(I\) is the set of right-censored observations.