1. Recall that Y has an exponential distribution with parameter $\lambda > 0$ if its probability density function (pdf) is given by

$$f_Y(y) = \begin{cases} \lambda e^{-\lambda y}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

The cumulative distribution function (cdf) is

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ 1 - e^{-\lambda y}, & y > 0. \end{cases}$$

(a) If $\lambda = 0.25$, find $P(Y \leq 1)$. Show your work.

(b) If $\lambda = 0.25$, find the median of the distribution of Y. Show your work.

(c) An electrical system consists of four components placed in parallel. Recall that a parallel system remains functional if at least one of the components is functional (in other words, a parallel system fails if all of the components fail). If

- the first two component lifetimes (in years) follow an exponential distribution with $\lambda = 0.25$
- the last two component lifetimes (in years) follow an exponential distribution with $\lambda = 0.50$,

find the probability that the system remains functional for at least 10 years. Assume that the four components are independent.

2. Recall that if Y counts the number of successes in n Bernoulli trials, then Y follows a binomial distribution, written $Y \sim b(n,p)$, where p is the probability of success. The probability mass function of Y is

$$p_Y(y) = \begin{cases} \binom{n}{y} p^y (1-p)^{n-y}, & y = 0, 1, 2, ..., n \\ 0, & \text{otherwise.} \end{cases}$$

Recall that

$$\begin{aligned} E(Y) &= np \\ \text{var}(Y) &= np(1-p). \end{aligned}$$

Concrete blocks are tested and it is found that, on average, 7 percent of blocks do not meet specifications (i.e., 7 percent are "defective."). Let Y count the number of defective blocks out of 20 tested at a local plant.

(a) State the Bernoulli trial assumptions in the context of this problem (there are 3 of them; talk in terms of concrete blocks). Assume these assumptions hold for parts (b) and (c) below.

(b) What is the probability that there are at most 2 defective blocks (out of 20)?

(c) The cost of disposing defective blocks (in dollars) is described by the function

$$C = 50 + 4Y + 2Y^2.$$

Find the expected cost.

3. A civil engineer monitors water quality by measuring the amount of suspended solids in river water. She observes a random sample of $n = 40$ water samples taken from a very large river (each water sample is 1 liter) and records Y, the number of suspended solids in each one. Here are the data:

37	37	28	22	23	31	34	32	35	33
26	36	23	35	26	31	30	27	31	31
31	34	38	33	26	22	33	22	33	37
38	32	39	29	40	32	34	30	32	25

(a) The population distribution for Y is unknown. Describe the sampling distribution of the sample mean $\overline{Y}$. Explain why your answer is correct. *Hint:* What "theorem" are you using?

(b) I used R to calculate the sample mean and sample variance of the 40 solid counts above (`no.solids`):

```
> mean(no.solids)
[1] 31.2
> var(no.solids)
[1] 24.8
```

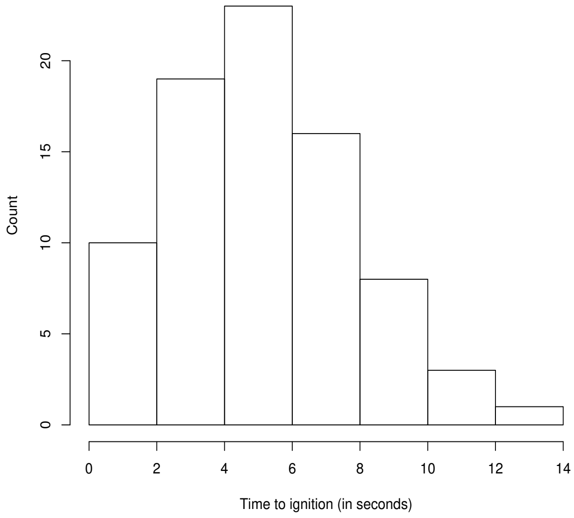
What do these quantities estimate? You can explain in words or using statistical symbols (define what the statistical symbols mean).

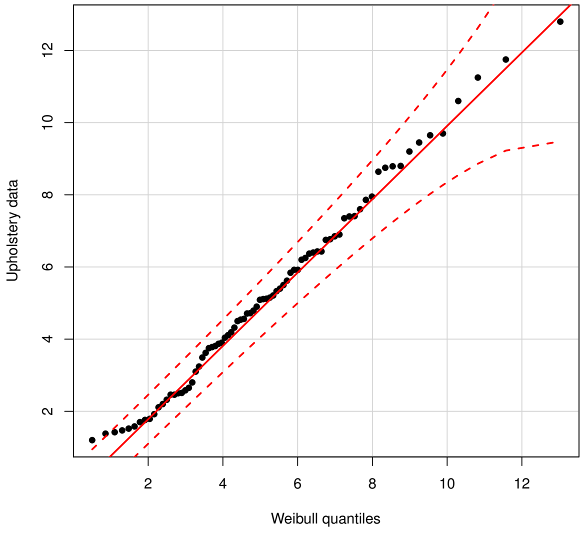
(c) Provide an estimate of the standard error of the sample mean.

4. In a fire safety study, engineers collected a random sample of $n = 80$ upholstery materials and on each material recorded Y, the ignition time when exposed to a flame (in seconds). Here are the data:

2.58	2.51	4.04	6.43	1.58	4.32	2.20	4.19
4.79	6.20	1.52	1.38	3.87	4.54	5.12	5.15
5.50	5.92	4.56	2.46	6.90	1.47	2.11	2.32
6.75	5.84	8.80	7.40	4.72	3.62	2.46	8.75
2.65	7.86	4.71	6.25	9.45	12.80	1.42	1.92
7.60	8.79	5.92	9.65	5.09	4.11	6.37	5.40
11.25	3.90	5.33	8.64	7.41	7.95	10.60	3.81
3.78	3.75	3.10	6.43	1.70	6.40	3.24	1.79
4.90	3.49	6.77	5.62	9.70	5.11	4.50	2.50
5.21	1.76	9.20	1.20	6.85	2.80	7.35	11.75

Here is a histogram of the data and a qq-plot under a Weibull model assumption.





(a) What do you think of the Weibull distribution as a population model for the ignition times?

(b) To estimate the population mean ignition time μ with a 95 percent confidence interval, one engineer used

$$\overline{y} \pm t_{79,0.025}\frac{s}{\sqrt{n}},$$

where $\overline{y}$ and s are the sample mean and sample standard deviation, respectively. In R,

```
 t.test(ignition.times,conf.level=0.95)$conf.int
[1] 4.61 5.82
```

Interpret precisely what this interval tells us.

(c) Another engineer looks at the confidence interval in part (b) and says,

> "I'm confused. Only 13 of the 80 times are between 4.61 and 5.82 seconds. The fraction 13/80 is nowhere close to 0.95."

What would you tell him? Do not waste time verifying that 13/80 times are between 4.61 and 5.82.

(d) Suppose the engineers wanted to calculate a confidence interval for the population variance σ^2 using the formula

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}} \right),$$

where s^2 is the sample variance. Explain why they might use caution in doing this.

5. I found this excerpt on a web site blog for engineering statistics:

> "Got right-skewed data? Weibull can model that. Left-skewed data? Sure, that's cool with Weibull. Symmetric data? Weibull's up for it. That flexibility is why engineers use the Weibull distribution to evaluate the reliability and material strengths of everything from vacuum tubes and capacitors to ball bearings and relays."

Recall that the pdf of $T \sim \text{Weibull}(\beta, \eta)$ is

$$f_T(t) = \begin{cases} \dfrac{\beta}{\eta}\left(\dfrac{t}{\eta}\right)^{\beta-1} e^{-(t/\eta)^{\beta}}, & t > 0 \\ 0, & \text{otherwise.} \end{cases}$$

The parameters in this distribution recall are

$$\begin{aligned} \beta &= \textbf{shape} \text{ parameter} \\ \eta &= \textbf{scale} \text{ parameter.} \end{aligned}$$

Recall that the hazard function of T is

$$h_T(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1}.$$

(a) Suppose 5 years from now, you are working for General Electric. Your co-worker asks you to perform a Weibull analysis for the time until part failure (for a critical part). However, when you look at the data, it is clear that a Weibull distribution is not a good fit and that another model (e.g., lognormal) is much better. However, your co-worker has never heard of this other distribution, so he is not interested in it and asks you to proceed with a Weibull analysis. Write a short comment to him explaining the potential dangers of doing this.

(b) On another occasion, your co-worker asks you to do a Weibull analysis for another critical part. This time, the Weibull model looks to be a good fit. After fitting the model, you find these maximum likelihood estimates:

```
> fitdistr(data,densfun="weibull")
              shape         scale
Estimate      1.188        12.922
Std Error    (0.350)       (2.008)
```

The quantities in parentheses are standard errors of the estimates. A 95 percent confidence interval for β can be formed by calculating

$$\texttt{Estimate} \pm 1.96(\texttt{Std Error}).$$

Calculate this interval for β and interpret it.

(c) Your co-worker looks at the Weibull estimate of 1.188 for the shape parameter β and proclaims,

> "I knew β would be larger than 1. This population of parts is getting weaker over time."

How would you respond?

6. From January 1, 2009 to December 31, 2009, there were a total of 615 babies admitted to the neonatal intensive care unit (NICU) at Richland Hospital in Columbia, SC.

(a) My colleagues want to know if the population mean birth weight (in grams) of babies whose mothers used some type of illegal drug (Population 1) differed from the population mean birth weight (in grams) of babies whose mothers did not use illegal drugs (Population 2).

- Among the 615 babies, there were $n_1 = 65$ mothers who used drugs and $n_2 = 550$ who did not.
- Treat the babies from these mothers as independent random samples from larger populations.

I used R to write a 95 percent confidence interval for

$$\mu_1 - \mu_2 = \text{difference of population mean birth weights}$$

to be $(-399.6, -37.8)$ grams. Assuming that all necessary statistical assumptions apply (they do reasonably well), interpret what this interval suggests about the effect of maternal drug use.

(b) Could the research question in part (a) be examined by using a matched pairs design from the same 615 mothers? If you believe it could, suggest how. If not, then explain why not.

(c) My colleagues also want to know if the population proportion of babies who developed necrotizing enterocolitis (NEC) was different across the two groups. NEC occurs when the lining of the intestinal wall dies and the tissue falls off (it is very bad). Here is a 2×2 table that summarizes the NEC frequencies by drug use:

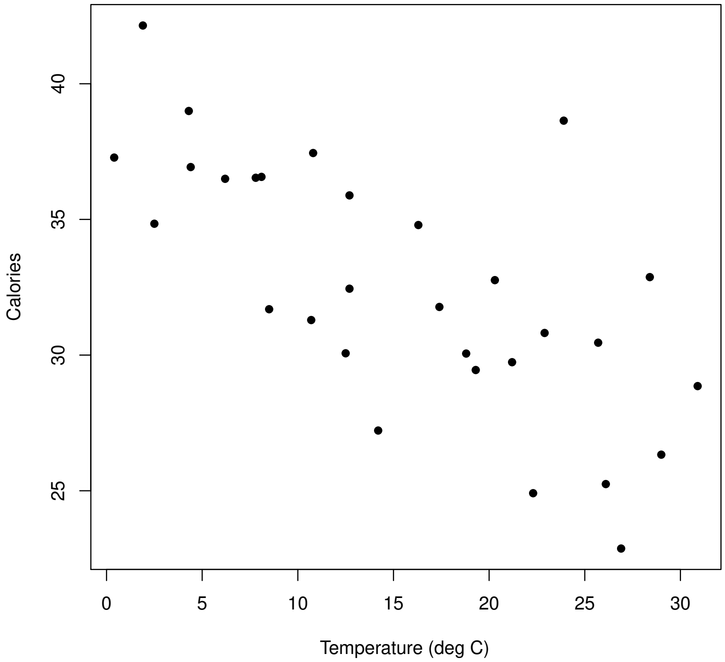
		NEC positive	NEC negative	
Mother's drug use	YES	8	57	$n_1 = 65$
	NO	37	513	$n_2 = 550$

Continue to treat the babies from these mothers as independent random samples from larger populations. Write a 95 percent confidence interval for the difference of the population proportions of NEC positive babies for the two groups (mother uses drugs/does not). Recall that this interval is given by

$$(\widehat{p}_1 - \widehat{p}_2) \pm 1.96\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}.$$

Interpret the interval clearly. Does maternal drug use influence the probability of NEC development?

7. An ornithologist wants to relate the amount of energy used (Y, measured in calories) to temperature (x, measured in deg C) in a certain species of birds. A random sample of 30 birds was available. Each bird was subjected to a specific temperature and the energy use was recorded for each bird. Here is a scatterplot of the data:



The ornithologist wants to use a simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

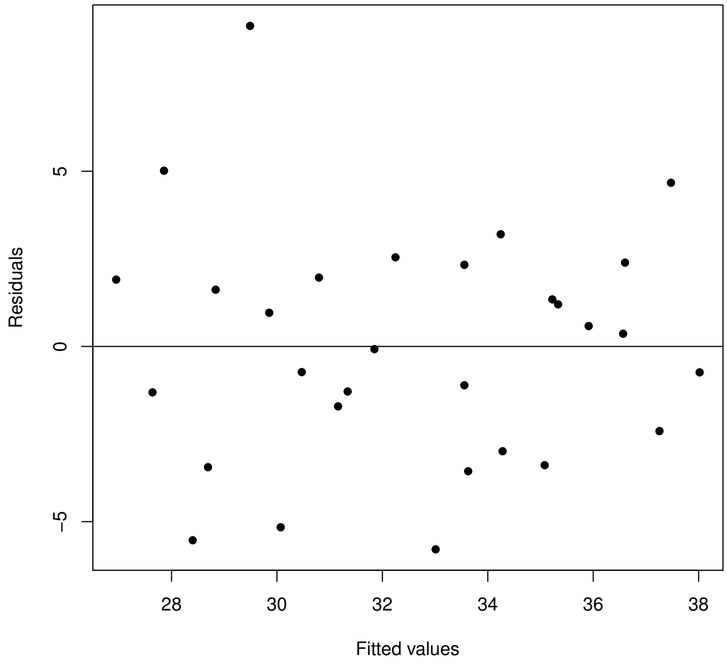
to describe this relationship for the population of birds. Here is the R output from fitting this model to the data above:

```
> fit = lm(calories~temp)
> summary(fit)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.16370    1.25781  30.341  < 2e-16 ***
temp        -0.36289    0.07027  -5.165 1.77e-05 ***

Residual standard error: 3.399 on 28 degrees of freedom
Multiple R-squared:  0.487
```

(a) Is there evidence that temperature and energy use are linearly related in the population? Explain why or why not.

(b) Interpret what $R^2 = 0.487$ means in words. Explain it so the ornithologist can understand.

(c) Here is the residual plot from the fitted model:



Describe what this plot reveals about the model fit.

(d) The ornithologist would like to estimate the mean energy use for the population of birds subjected to 15 deg C. Does he want to use a confidence interval or prediction interval? Explain.

8. Rayon whiteness is an important factor for scientists dealing with fabric quality. A random sample of $n = 16$ rayon specimens is available. On each specimen, the response Y (a numerical measure of rayon whiteness) is measured. The following six independent variables are also measured:

$$\begin{aligned} x_1 &= \text{acid bath temperature (deg C)} \\ x_2 &= \text{cascade acid concentration (percentage)} \\ x_3 &= \text{water temperature (deg C)} \\ x_4 &= \text{sulfide concentration (percentage)} \\ x_5 &= \text{amount of chlorine bleach (lb/min)} \\ x_6 &= \text{blanket finish temperature (deg C).} \end{aligned}$$

Chemists would like to model the relationship between Y and the six independent variables using the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \epsilon_i,$$

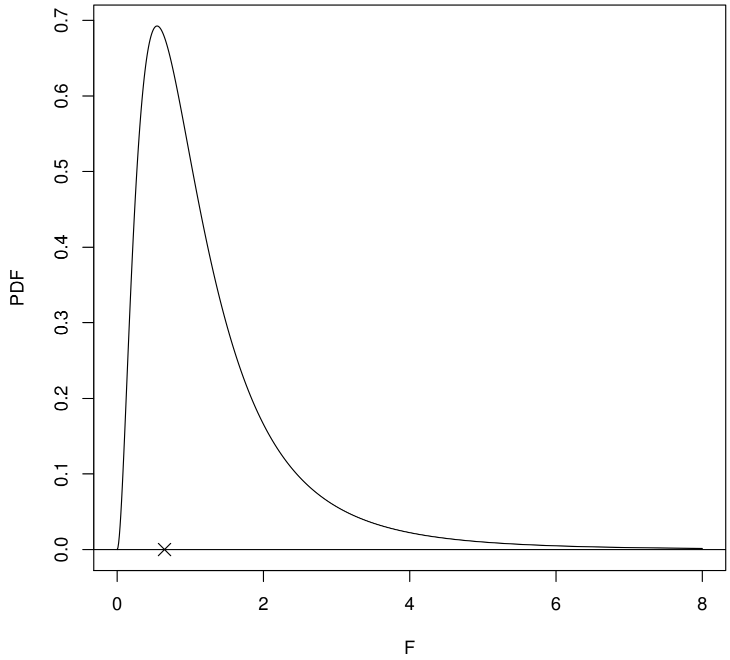
for $i = 1, 2, ..., 16$. I fit this model to the data in R (data not shown). Here is the output:

```
> fit = lm(whiteness ~ acid.temp + cascade.conc + water.temp + sulfide.conc
    + chlorine + blanket.temp)
> summary(fit)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -172.6744   179.2521  -0.963    0.361
acid.temp      -1.2619     1.5706  -0.803    0.442
cascade.conc  -27.9731    47.1617  -0.593    0.568
water.temp      2.1838     1.5819   1.381    0.201
sulfide.conc    1.3639   236.4142   0.006    0.996
chlorine      136.1482   119.0395   1.144    0.282
blanket.temp    1.0401     0.7725   1.346    0.211

Residual standard error: 19.23 on 9 degrees of freedom
Multiple R-squared:  0.3015
F-statistic: 0.6475 on 6 and 9 DF,  p-value: 0.6927
```

(a) Is rayon whiteness linearly related to acid temperature (x_1) in the population after adjusting for the effects of the other variables? Use the output above to answer this question. Note that the p-value = 0.442 is used to test

$$\begin{gathered} H_0 : \beta_1 = 0 \\ \text{versus} \\ H_1 : \beta_1 \neq 0. \end{gathered}$$

(b) I plotted the value of $F = 0.6475$ (see output) on the $F(6,9)$ pdf below:



This F statistic is used to test two hypotheses: H_0 and H_1. Write out what these hypotheses are. You can do this using notation (that you clearly define) or you can write this out in words.

(c) We have learned that values of $F \approx 1$ are generally consistent with the H_0 you specified in part (b). Values of $F > 1$ (a lot larger than 1) are generally consistent with H_1. What about values of $F < 1$? My colleague Ron Christensen (University of New Mexico) has described how F tends to be smaller than 1 when the multiple linear regression assumptions are violated. State what these assumptions are for the errors ϵ_i (there are four of them).

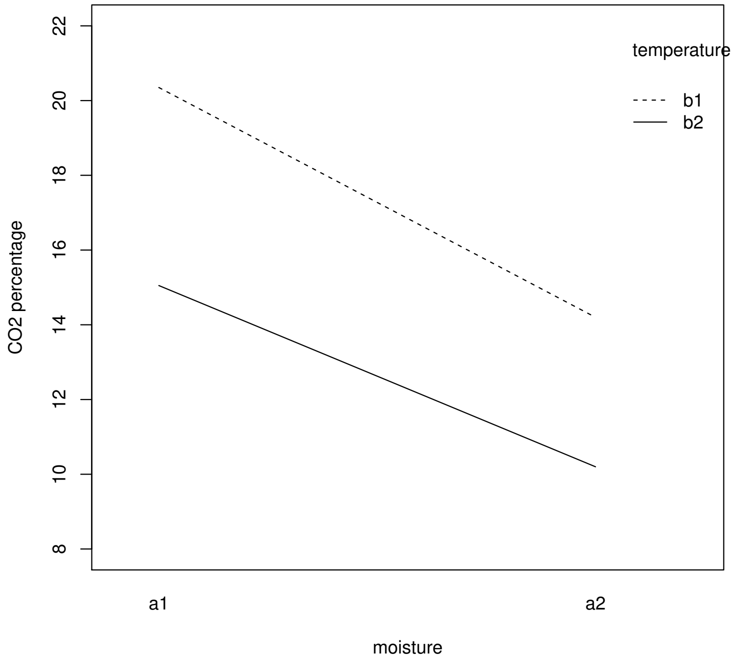
9. Researchers carried out a 2 × 2 factorial experiment to examine how much carbon dioxide concentration is produced when pine wood is burned (Y, measured as a percent) and how this concentration is related to the two factors:

Factor A: Moisture
Factor B: Temperature.

The levels of Factor A are $a_1 = 0$ (no moisture present) and $a_2 = 22$ percent (moisture present). The levels of Factor B are $b_1 = 1100$ deg K and $b_2 = 1500$ deg K. There were 2 replicates at each treatment combination, resulting in the following 8 CO_2 measurements.

Moisture (A)	Temperature (B)	Replication 1	Replication 2
0	1100	20.3	20.4
22	1100	13.6	14.8
0	1500	15.0	15.1
22	1500	9.7	10.7

I used R to construct the moisture-temperature interaction plot and also to calculate the analysis of variance table appropriate for the 2 × 2 factorial treatment structure.



```
> anova(fit)
Analysis of Variance Table
Response: co2
                     Df Sum Sq Mean Sq F value    Pr(>F)
moisture              1 60.500  60.500 196.748 0.0001499 ***
temperature           1 43.245  43.245 140.634 0.0002895 ***
moisture:temperature  1  0.845   0.845   2.748 0.1727182
Residuals             4  1.230   0.307
```

(a) One researcher is convinced that there is a population level interaction between moisture and temperature. Provide reasoning to support or refute this assertion.

(b) Which of the main effects are significant in explaining CO_2 percentage: moisture, temperature, or both? Explain.

(c) Instead of analyzing the data as data from a 2×2 factorial experiment (to produce the ANOVA table on the last page), a less-informed researcher analyzed the data as data from a one-way classification with $t = 4$ treatments: a_1b_1, a_1b_2, a_2b_1, and a_2b_2. In this less-informed analysis, what would the treatment sum of squares be?

(d) In the analysis described in part (c), calculate the F statistic to test the equality of population means across the four treatment groups (a_1b_1, a_1b_2, a_2b_1, and a_2b_2). Recall that

$$F = \frac{\text{MS}_{trt}}{\text{MS}_{res}}.$$