1. (a) The three Bernoulli trial assumptions are

- each radar screen either detects an incoming missile or it doesn't (two outcomes)
- each radar screen operates independently
- the probability a radar screen detects an incoming missile ($p = 0.8$) is the same for all 3 radar screens.

Under these three assumptions, the random variable $X \sim b(n = 3, p = 0.8)$.
(b) We want

$$P(X = 0) = \binom{3}{0}(0.8)^0(0.2)^3 = (0.2)^3 = 0.008.$$

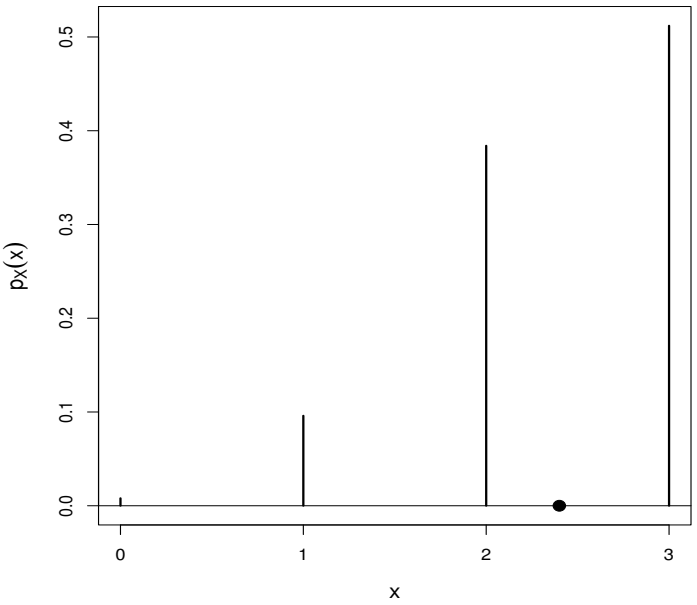
(c) Let's calculate $p_X(x) = P(X = x)$ for all possible values of x:

$$\begin{aligned} P(X = 0) &= \binom{3}{0}(0.8)^0(0.2)^3 = (0.2)^3 = 0.008 \\ P(X = 1) &= \binom{3}{1}(0.8)^1(0.2)^2 = 3(0.8)(0.2)^2 = 0.096 \\ P(X = 2) &= \binom{3}{2}(0.8)^2(0.2)^1 = 3(0.8)^2(0.2) = 0.384 \\ P(X = 3) &= \binom{3}{3}(0.8)^3(0.2)^0 = (0.8)^3 = 0.512. \end{aligned}$$

The mean of X is

$$E(X) = \sum_{\text{all } x} x p_X(x) = 0(0.008) + 1(0.096) + 2(0.384) + 3(0.512) = 2.4.$$

You could also use the fact that $E(X) = np = 3(0.8) = 2.4$. Here is the pmf of X:



The mean of X, $E(X) = 2.4$, is shown by using a dark circle.

2. (a) For injured workers, the two relevant events are

$$\begin{aligned} A &= \{\text{admitted to the hospital for treatment}\} \\ B &= \{\text{back on the job the next day}\}. \end{aligned}$$

We are given $P(A) = 0.10$, $P(B) = 0.15$, and $P(A \cap B) = 0.02$.
(b) We want the conditional probability

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.02}{0.10} = 0.2.$$

(c) We want to calculate $P(A' \cap B')$. By DeMorgan's Law, we have

$$A' \cap B' = (A \cup B)'.$$

Therefore,

$$\begin{aligned} P(A' \cap B') = P((A \cup B)') &= 1 - P(A \cup B) \\ &= 1 - [P(A) + P(B) - P(A \cap B)] \;=\; 1 - 0.23 = 0.77. \end{aligned}$$

Therefore, 77% of all injured workers will neither be admitted to the hospital for treatment nor be back on the job the next day.
(d) The manager believes that A and B are independent events. We know this is not true because

$$0.15 = P(B) \neq P(B|A) = 0.2,$$

from part (b). That is, the occurrence of A has changed how we assign probability to the event B. You could also show this by using the definition of independence:

$$0.02 = P(A \cap B) \neq P(A)P(B) = 0.10(0.15) = 0.015.$$

This also shows A and B are not independent.

3. (a) The lifetime T has an exponential distribution with

$$\lambda = \frac{1}{5} = 0.2.$$

One could calculate

$$P(T > 4) = \int_4^\infty 0.2e^{-0.2t}dt \qquad \text{or} \qquad 1 - P(T \leq 4) = 1 - \underbrace{\int_0^4 0.2e^{-0.2t}dt}_{= F_T(4)}.$$

We have

$$1 - P(T \leq 4) = 1 - F_T(4) = 1 - [1 - e^{-0.2(4)}] = e^{-0.8} \approx 0.449.$$

(b) We want

$$P(T > 8|T > 4).$$

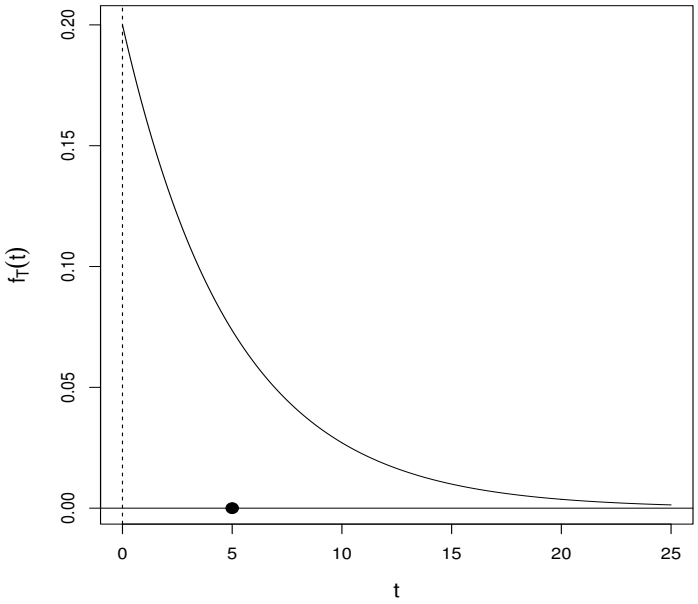
If you remember that the exponential distribution obeys the memoryless property, you know right away that

$$P(T > 8|T > 4) = P(T > 4) \approx 0.449,$$

as in part (a). If you forgot this, you could use the definition of conditional probability to get the answer:

$$\begin{aligned} P(T > 8|T > 4) &= \frac{P(T > 8 \text{ and } T > 4)}{P(T > 4)} \\ &= \frac{P(T > 8)}{P(T > 4)} \\ &= \frac{1 - F_T(8)}{1 - F_T(4)} = \frac{1 - [1 - e^{-0.2(8)}]}{1 - [1 - e^{-0.2(4)}]} = \frac{e^{-1.6}}{e^{-0.8}} = e^{-0.8} \approx 0.449. \end{aligned}$$

(c) The pdf of T looks like



You could argue the mean $E(T)$ will be larger than the median $\phi_{0.5}$ because the pdf is heavily skewed to the right (high) side. This will increase the balance point $E(T)$ when compared to the median. You could also calculate both:

Mean:

$$E(T) = \frac{1}{0.2} = 5.$$

Median: We are left to solve $F_T(\phi_{0.5}) = 0.5$ for $\phi_{0.5}$. We have

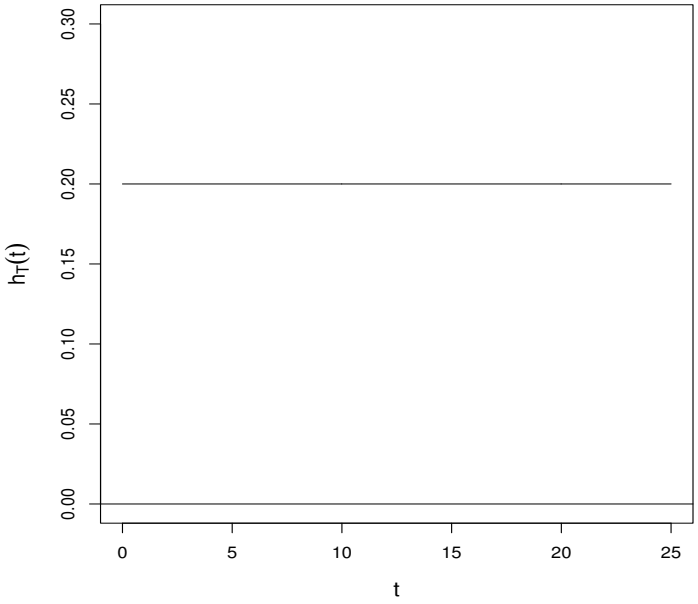
$$\begin{aligned} F_T(\phi_{0.5}) = 0.5 &\implies 1 - e^{-0.2\phi_{0.5}} = 0.5 \\ &\implies e^{-0.2\phi_{0.5}} = 0.5 \\ &\implies -0.2\phi_{0.5} = \ln(0.5) \implies \phi_{0.5} = -\frac{\ln(0.5)}{0.2} \approx 3.47. \end{aligned}$$

(d) We know the exponential distribution is another name for the Weibull distribution when the shape parameter $\beta = 1$. Therefore, the hazard function $h_T(t)$ is a constant function of t. This means the population of electronic components neither weakens nor strengthens over time. In other words, the components fail completely at random.

If you forgot the exponential-Weibull relationship, you could derive the hazard function using

$$h_T(t) = \frac{f_T(t)}{S_T(t)} = \frac{f_T(t)}{1 - F_T(t)} = \frac{0.2e^{-0.2t}}{1 - [1 - e^{-0.2t}]} = \frac{0.2e^{-0.2t}}{e^{-0.2t}} = 0.2.$$

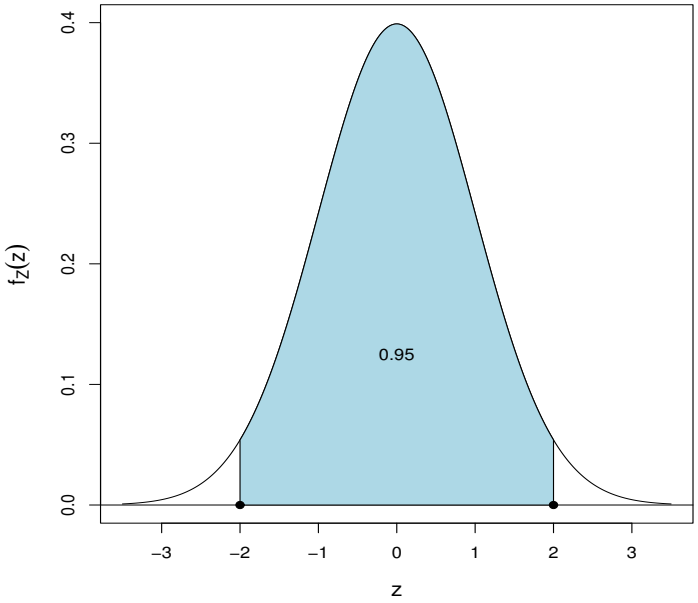
Therefore, the hazard function is a constant function of t, as shown below:



4. (a) Following the hint, we have

$$P(X \leq 600) = P(U \leq \ln 600) = P\left(Z \leq \frac{\ln 600 - 5.4}{0.5}\right) \approx P(Z \leq 2).$$

From the 68-95-99.7% Rule, we know about 95% percent of the values of Z will be between -2 and 2, that is,



Therefore,

$$P(Z \leq 2) \approx 0.95 + 0.025 = 0.975.$$

That is, approximately 97.5% of the population will have triglyceride levels less than 600 mg/dL.

(b) Not at all. The qq plot shows excellent agreement between the observed triglyceride levels and the lognormal quantiles. Based on this sample, the lognormal distribution appears to be an excellent choice for the population distribution.
(c) In short, the standard errors measure how variable the (maximum likelihood) estimates are. That is, the estimates 2.04 and 289.26 are calculated from the specific sample of 250 patients drawn from a large population. Different samples of 250 patients will produce different estimates of β and η. The standard errors measure how variable the estimates would be from sample to sample.
(d) There is a noticeable systematic departure between the observed triglyceride levels and the Weibull quantiles starting around 450 mg/dL and continuing for larger triglyceride levels (the larger the levels, the larger the departure). The Weibull distribution may not be a bad choice for the population distribution, but the lognormal distribution appears to be a better choice. It does a better job describing the data, especially when the triglyceride levels are larger.

5. A reasonable answer is "all adult residents in Arizona." There are other reasonable answers too such as, "all Arizona drivers aged 18 and older."
(b) The sample proportion is

$$\widehat{p} = \frac{385}{500} = 0.77.$$

Therefore, a 90% confidence interval for the population proportion p is

$$0.77 \pm 1.65\sqrt{\frac{0.77(0.23)}{500}} \longrightarrow 0.77 \pm 0.031 \longrightarrow (0.739, 0.801).$$

We are 90% confident the population proportion of all adult residents in Arizona who favor increasing the speed limit is between 0.739 and 0.801.
(c) The margin of error is

$$1.65\sqrt{\frac{0.77(0.23)}{500}} \approx 0.031.$$

(d) To reduce the margin of error, you could

- increase the sample size n. A 90% confidence interval with a larger sample size than 500 will reduce the margin of error.
- decrease the confidence level $100(1-\alpha)\%$. Confidence intervals with lower confidence levels will be shorter, so the margin of error will be lower.

6. (a) This is a matched pairs experiment because the researchers measured each plot under two experimental conditions: preburn and postburn. The samples are dependent because the two calcium levels are measured on the same plot. This means we would expect preburn and postburn measurements on the same plot to be "more alike" than if we were to measure preburn on one plot and postburn on a different plot.
(b) We are 95% confident the population mean difference $\Delta = \mu_1 - \mu_2$ is between 30.5 and 50.6 kg/plot. This confidence interval includes only positive values, which is consistent with $\mu_1 > \mu_2$. This suggests, at the 95% confidence level, the fire burn decreases the (population) mean calcium level in the population of all plots.
(c) A matched pairs design is preferred over a two-independent sample design because you eliminate systematic sources of variability that make 2 different plots structurally different.

- For example, suppose you measured a preburn calcium level on one plot and a postburn calcium level on a different plot (as one would do with 2 independent samples). Here you would get to see the difference between the preburn and postburn levels, but this difference also incorporates all of the variability that makes the 2 plots different from each other (e.g., differences in soil composition, water levels, different levels of other chemicals, etc.).
- When you measure both preburn and postburn on the same plot, the difference is calculated on the same plot, so all of the extra sources of variability, such as those listed above, are removed. This is why the matched pairs analysis is more precise and hence provides a shorter confidence interval for $\Delta = \mu_1 - \mu_2$.

7. (a) Under the simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon \quad \Longleftrightarrow \quad E(Y) = \beta_0 + \beta_1 x,$$

the parameter β_0 is the population mean price $E(Y)$ when $x = 0$, that is, the mean price for all homes which are 0 ft^2 in size. Obviously, this makes no sense. The confidence interval for β_0 therefore has no practical interpretation.
(b) Under the simple linear regression model (above), the parameter β_1 describes the linear relationship between the population mean house price $E(Y)$ and the house size x. If the confidence interval included "0," then we could not conclude that mean price and house size were linearly related in the population. However, that's not what we see here. The confidence interval $(47.66, 67.18)$ excludes "0," indicating a positive linear relationship between mean price $E(Y)$ and house size for all houses in the population.
(c) The confidence interval

$$(205.36, 216.86)$$

is interpreted as follows: "Among all houses in the population which are 2000 ft^2, we are 95% confident the mean price $E(Y)$ is between \$205,360 and \$216,860."

The prediction interval

$$(167.41, 254.82)$$

is interpreted as follows: "For a single house which is 2000 ft^2, the probability its price is between \$167,410 and \$254,820 is 0.95."
(d) If the residual plot displayed a totally random appearance, this would suggest the simple linear regression model is appropriate for the population of all houses and the underlying assumptions are satisfactory. However, we do see a "fanning out" appearance in the plot, which suggests variation levels in house price may be increasing as the house size increases. This pattern we see points to a violation in the constant variance assumption.

8. (a) The (least-squares) estimate of $E(Y)$ when $x_1 = 50$, $x_2 = 76$, and $x_3 = 29.30$ is

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3 = -3.5078 - 0.0026(50) + 0.0008(76) + 0.1542(29.30) \approx 0.941 \text{ ppm}.$$

(b) The F statistic is used to test

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

versus

$$H_1 : \text{at least one of } \beta_1, \beta_2, \beta_3 \text{ is nonzero.}$$

That is, the null hypothesis H_0 says that none of the independent variables (humidity, temperature, and pressure) are linearly related to the mean nitrous oxide emission $E(Y)$ in the population of all diesel trucks in Connecticut. The alternative hypothesis H_1 says that at least one of these variables is linearly related to the mean nitrous oxide emission $E(Y)$.

From the ANOVA table, we see F is large ($F = 21.395$), much larger than one would expect if H_0 was true. Remember that if H_0 was true, we would expect F to be around 1 since MS_R and MS_E estimate the same thing under H_0. We have strong evidence that H_1 is true.

You could also argue this by noting the p-value (0.000007609) is very small. The smaller the p-value, the more evidence we have against H_0. This very small p-value corresponds to having strong evidence H_1 is true. That is, at least one of the independent variables is linearly related to $E(Y)$ in the population.
(c) The coefficient of determination R^2 is

$$R^2 = \frac{\text{SS}_\text{R}}{\text{SS}_\text{TOT}} = \frac{0.202501}{0.252980} \approx 0.800.$$

This means that approximately 80% of the variation in the nitrous oxide emission data is explained by the linear relationship with humidity, temperature, and pressure.
(d) Because the confidence interval for β_3 includes "0," we do not have evidence (at the 95% confidence level) to conclude barometric pressure (x_3) and the mean nitrous oxygen content $E(Y)$ are linearly related in the population of all diesel trucks in Connecticut, after adjusting for the contributions (effects) of humidity (x_1) and temperature (x_2). This means $\text{SS}(x_3|x_1, x_2)$ will be "small."

9. (a) The corresponding p-values inform us as to whether the associated F statistics are "large" or "small" at the $\alpha = 0.05$ significance level:

- The main effect of coating is significant at the $\alpha = 0.05$ significance level (p-value = 0.0381).
- The main effect of humidity is not significant at the $\alpha = 0.05$ significance level (p-value = 0.0899).
- The interaction effect of coating and humidity is not significant at the $\alpha = 0.05$ significance level (p-value = 0.9876).

(b) The interaction plot of coating and humidity would display nearly perfectly parallel lines because the interaction effect of coating and humidity is nowhere close to being significant. The interaction F statistic (p-value) is very small (large).
(c) Here is the ANOVA table for a one-way classification analysis:

```
Analysis of Variance Table

Response: Cycles
          Df  Sum Sq   Mean Sq F value
Treatment  3 1240979  413659.7   2.702
Residuals 20 3062239  153112.0
Total     23 4303218
```