GROUND RULES:

- Print your name at the top of this page. Do not put your name on any other page.
- This is a closed-book and closed-notes exam.
- This exam contains 9 questions, each worth 10 points. This exam is worth **90 points** total.
- You may use a calculator, but this calculator cannot have internet access. You cannot use your phone as a calculator. You cannot share calculators with another student. Show all of your work; use a calculator only to do final calculations or to check your work.
- Each problem contains parts. On each part, there is opportunity for partial credit, so show all of your work and explain all of your reasoning. **Translation:** No work/no explanation means no credit.
- On any problem, you may use the back of the page if you need more space. I also have extra paper if you need it.
- Any discussion or inappropriate communication between you and another examinee, as well as the appearance of any unnecessary material, will result in a declaration of academic dishonesty. Don't risk it!
- You have 2.5 hours to complete this exam.

HONOR PLEDGE FOR THIS EXAM:

After you have finished the exam, please read the following statement and sign your name below it.

> *I promise that I did not discuss any aspect of this exam with anyone other than the instructor, that I neither gave nor received any unauthorized assistance on this exam, and that the work presented herein is entirely my own.*

1. National security requires that defense technology is able to detect incoming missiles. To make defense systems successful, multiple radar screens are used. One defense system uses 3 radar screens which operate independently. The probability any one screen will detect an incoming missile is 0.8. Define

$$X = \text{the number of screens (out of 3) which will detect an incoming missile.}$$

(a) List the three assumptions needed for X to have a binomial distribution.

(b) What is the probability an incoming missile will **not** be detected by any of the three screens?

Part (c) is on the next page.

(c) Make a graph of the probability mass function (pmf) of X and identify where $E(X)$ falls. Label both axes. Neatness counts.

2. Human resource officers at an industrial plant are conducting a study to determine how quickly injured workers are back on the job following injury. Records show the following:

- Ten percent (10%) of all injured workers are admitted to the hospital for treatment.
- Fifteen percent (15%) of all injured workers are back on the job the next day.
- Two percent (2%) of all injured workers are admitted to the hospital for treatment and are back on the job the next day.

(a) Define two events A and B and write each number above using our notation for probability. Do not define more than two events or you will make the problem too hard.

(b) Suppose an injured worker is admitted to the hospital for treatment. What is the probability he will be back on the job the next day?

Parts (c) and (d) are on the next page.

(c) What percentage of all injured workers will neither be admitted to the hospital for treatment nor be back on the job the next day?

(d) A manager believes that whether an injured worker is admitted to the hospital for treatment has no effect on whether the worker will back on the job the next day (and vice versa). Provide evidence to confirm this belief or refute it.

3. The lifetime T of an electronic component (measured in 1000s of hours) has an exponential distribution with probability density function

$$f_T(t) = \begin{cases} \frac{1}{5}e^{-t/5}, & t > 0 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Calculate $P(T > 4)$, the probability a component lasts longer than 4000 hours.

(b) Suppose a component has been in operation for 4000 hours and is still functioning. Calculate the probability this component is still functioning at 8000 hours of operation. *Hint*: This part is asking for a conditional probability.

Parts (c) and (d) are on the next page.

(c) Which is larger: the mean lifetime $E(T)$ or the median lifetime $\phi_{0.5}$? Why?

(d) Sketch a graph of the hazard function of T. Label both axes. Describe what this graph tells us.

4. The lognormal distribution is common in biomedical applications involving positive measurements such as pollution concentrations, triglyceride levels, blood pressure, parasite counts, and cancer survival times. Typically, distributions of measurements like these are not symmetric, rather, the distributions are skewed to the right.

Mathematically, we have the following relationship:

$$X \sim \text{lognormal}(\mu, \sigma^2) \iff U = \ln X \sim \mathcal{N}(\mu, \sigma^2).$$

This means we can write the cumulative distribution function (cdf) of X as

$$F_X(x) = P(X \leq x) = P(\ln X \leq \ln x) = P(U \leq \ln x) = \underbrace{F_U(\ln x) = F_Z\left(\frac{\ln x - \mu}{\sigma}\right)}_{\text{standardizing } \ln x},$$

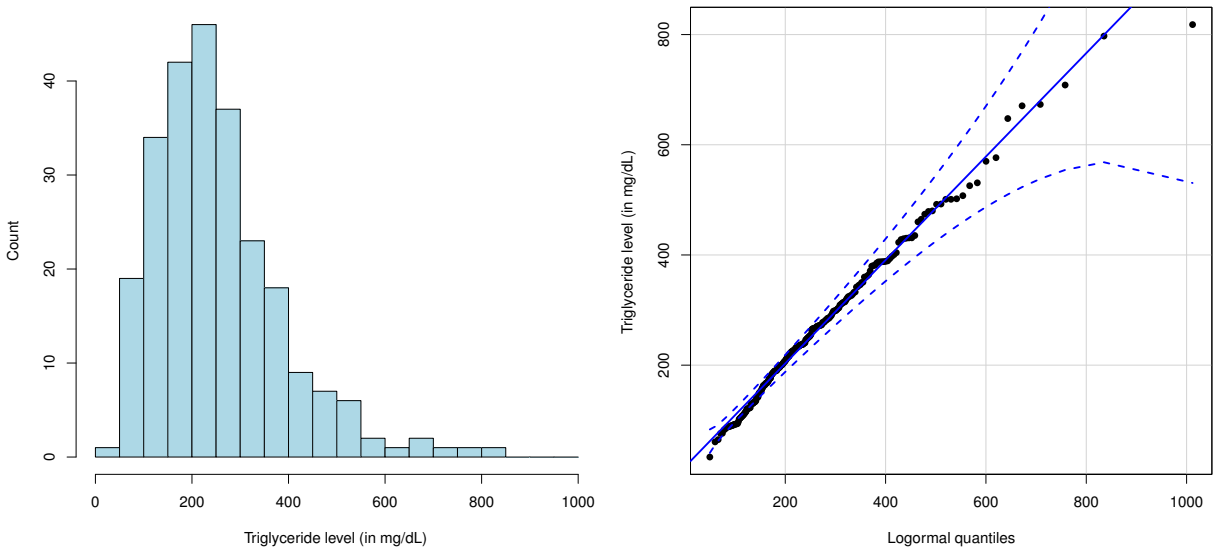
where F_Z is the cdf of a $\mathcal{N}(0,1)$ random variable. That is, lognormal probabilities can be calculated by using the normal distribution after a log-transformation.

(a) In a population of diabetic patients, suppose the distribution of a patient's triglyceride level X (measured in mg/dL) is lognormal with $\mu = 5.4$ and $\sigma^2 = 0.25$. Approximate $P(X \leq 600)$, the percentage of patients in the population whose triglyceride level is less than or equal to 600 mg/dL. *Hint:* Using the relationship above, argue that

$$P(X \leq 600) \approx P(Z \leq 2),$$

where $Z \sim \mathcal{N}(0,1)$. Now use the 68-95-99.7% Rule for normal distributions. Draw a picture.

Parts (b), (c), and (d) are on the next two pages.



(b) One researcher is concerned about whether the lognormal distribution is appropriate for triglyceride levels in this population of patients. He observes a random sample of $n = 250$ patients and measures the triglyceride level on each one. A histogram of the observed data and a quantile-quantile plot (assuming a **lognormal** population distribution) are shown above.

Should the researcher be concerned about the lognormal population distribution assumption? Why or why not?

Parts (c) and (d) are on the next page.

Another researcher has never heard of the lognormal distribution, but she has heard of the Weibull distribution and "its ability to model right skewed data in engineering applications." Using the same sample of $n = 250$ patients, she uses R to estimate a Weibull model for the population:

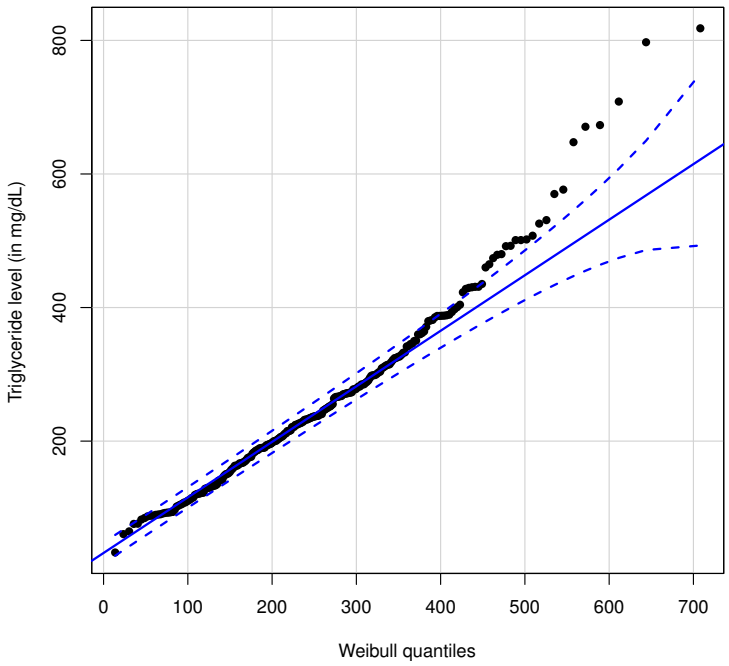
```
> fitdist(triglycerides,distr="weibull",method="mle")

      estimate  Std.Error
shape     2.04      0.094
scale   289.26      9.498
```

The maximum likelihood estimates of the Weibull shape (β) and scale (η) parameters are

$$\begin{aligned} \widehat{\beta} &= 2.04 \\ \widehat{\eta} &= 289.26. \end{aligned}$$

(c) The standard error of each estimate is shown in the output above. Describe what the standard errors measure.

(d) Here is a quantile-quantile plot of the triglyceride data assuming a **Weibull** population distribution:



Are these data well represented by a Weibull distribution? Would you prefer to use a lognormal distribution or a Weibull distribution for triglyceride levels in the population? Explain. Use the back of this page for your answers.

5. The Department of Transportation in Arizona is conducting a survey to determine residents' thoughts on the current speed limit for state and urban highways. A random sample of 500 adult residents (aged 18 and over) was selected from all over the state, and 385 were in favor of increasing the speed limit to 75 miles per hour for all highways in the state.

(a) What is the population in this example? Give a reasonable answer.

(b) Use the formula

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

to calculate a 90% confidence interval for the population proportion, that is, for the population you described in part (a). Note that $z_{0.10/2} = 1.65$ for 90% confidence. Interpret your confidence interval with a clearly written sentence.

(c) What is the margin of error associated with your interval in part (b)? I'm looking for a numerical answer here.

(d) Provide two ways one could reduce the margin of error associated with the confidence interval in part (b).

Use the back of this page if you need extra space.

6. Calcium is a vital component of soil structure, enhancing water infiltration and root penetration. In a recent study at Virginia Tech University, researchers hypothesized that fire burns may alter the calcium level present in soil. The researchers selected a tract of land in Fishburn Forest (next to the university) for a fire burn. Soil specimens were taken from 12 plots of equal area prior to the burn and analyzed for their calcium level (in kg/plot). Postburn calcium levels were then analyzed on the same 12 plots. Here are the data:

	Calcium levels (kg/plot)	
Plot	Preburn	Postburn
1	50	9
2	50	18
3	82	45
4	64	18
5	82	18
6	73	9
7	77	32
8	54	9
9	23	18
10	45	9
11	36	9
12	54	9

The goal is to make inference for $\Delta = \mu_1 - \mu_2$, where

$$\begin{aligned} \mu_1 &= \text{population mean calcium level for all plots (preburn)} \\ \mu_2 &= \text{population mean calcium level for all plots (postburn).} \end{aligned}$$

(a) Explain why this is a matched pairs design. What makes the two samples above dependent?

(b) I used R to analyze the data differences

$$d_i = \text{Preburn}_i - \text{Postburn}_i,$$

for $i = 1, 2, ..., 12$. Here is the R output for a 95% confidence interval for $\Delta = \mu_1 - \mu_2$:

```
> diff = preburn-postburn
> t.test(diff,conf.level=0.95)$conf.int
[1] 30.5  50.6
```

Interpret what the interval $(30.5, 50.6)$ means. Then discuss what effect a fire burn has on the population mean calcium levels. Use the back of this page for your answers.

Part (c) is on the next page.

(c) When I incorrectly analyzed the data on the previous page as two independent samples (e.g., 12 plots preburn and 12 separate plots postburn), I got the following results:

```
> t.test(preburn,postburn,conf.level=0.95,var.equal=FALSE)$conf.int
[1] 27.3  53.7
```

Why is this 95% confidence interval for $\Delta = \mu_1 - \mu_2$ longer than the one in part (b)? Don't just say something like, "You're analyzing the data incorrectly." Get to the root cause of why this happens.

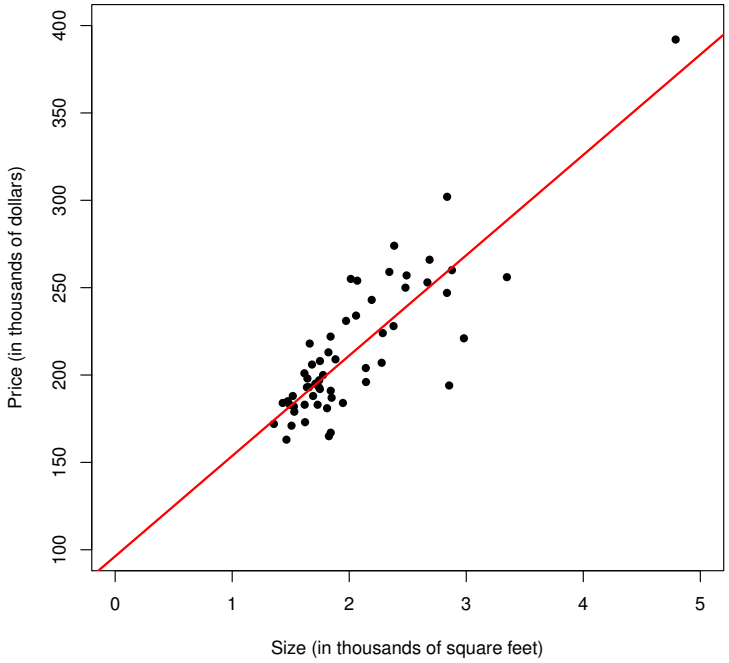
7. A realtor in Columbia, SC, is trying to model the relationship between house price and house size for middle-class neighborhoods in the Columbia metropolitan area. From a listing service, he selects a random sample of $n = 57$ houses and records

$$\begin{aligned} Y &= \text{price (in \$1000s)} \\ x &= \text{size (in 1000s ft}^2\text{)}. \end{aligned}$$

He considers the simple linear regression model

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

to describe this relationship for all houses in middle-class neighborhoods in the Columbia metropolitan area. Here is a scatterplot of the data with the least-squares regression line superimposed:



Here are the least-squares estimates of β_0 and β_1 and 95% confidence intervals for each parameter:

```
> fit = lm(price ~ size)
> fit

Coefficients:
(Intercept)        Size
      96.27       57.42

> confint(fit,conf.level=0.95)
             2.5%    97.5%
(Intercept) 75.57   116.96
size        47.66    67.18
```

Parts (a), (b), (c), and (d) are on the next two pages.

(a) A 95% confidence interval for β_0 is $(75.57, 116.96)$. Explain why this interval has no practical meaning in this application.

(b) Are house price and house size linearly related in the population of all houses in middle-class neighborhoods in the Columbia metropolitan area? Provide statistical evidence to support this or to refute it.

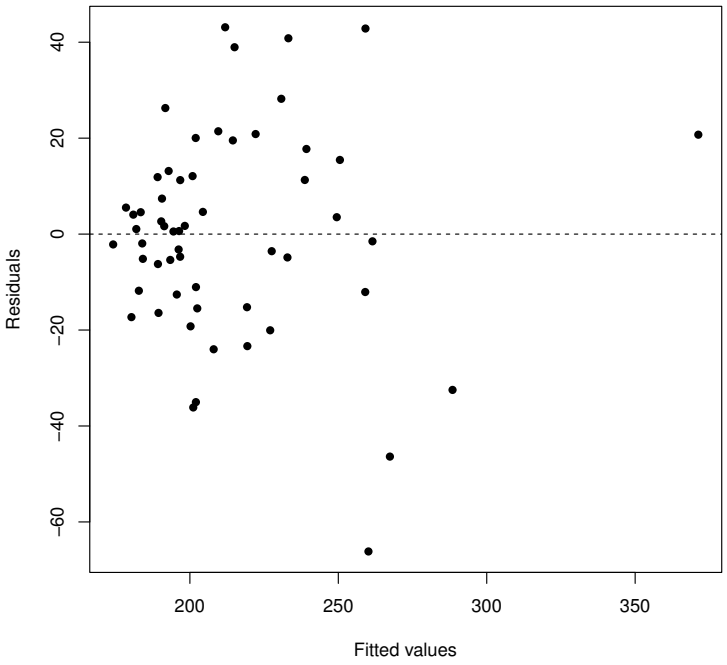
(c) I used R's `predict` function to calculate the following intervals (`lwr`, `upr`) when $x = 2.0$, that is, when the house size is 2000 square feet:

```
> predict(fit,data.frame(size=2.0),level=0.95,interval="confidence")
       fit     lwr     upr
1   211.11  205.36  216.86
> predict(fit,data.frame(size=2.0),level=0.95,interval="prediction")
       fit     lwr     upr
1   211.11  167.41  254.82
```

Interpret what each of these intervals means. Use the back of this page if necessary.

Part (d) is on the next page.

(d) The residual plot from the simple linear regression model fit is shown below:



What does this plot suggest? Do you detect any violations in the underlying assumptions for linear regression? Explain.

8. Nitrous oxide (N_2O) in diesel trucks is a popular performance upgrade that injects extra oxygen into the combustion chamber to burn excess fuel and significantly increase horsepower. In a study carried out by the Department of Motor Vehicles in Connecticut, a random sample of $n = 20$ diesel-powered trucks was observed to see how

$$Y = \text{N}_2\text{O emission (in ppm)}$$

might be influenced by the following independent variables:

$$\begin{aligned} x_1 &= \text{humidity (measured as a percentage)} \\ x_2 &= \text{temperature (in deg F)} \\ x_3 &= \text{barometric pressure (in inches Hg).} \end{aligned}$$

These four variables were measured on each of the 20 trucks, producing the data below:

Truck	Y	x_1	x_2	x_3	Truck	Y	x_1	x_2	x_3
1	0.90	72.4	76.3	29.18	11	1.07	23.2	76.8	29.38
2	0.91	41.6	70.3	29.35	12	0.94	47.4	86.6	29.35
3	0.96	34.3	77.1	29.24	13	1.10	315	76.9	29.63
4	0.89	35.1	68.0	29.27	14	1.10	10.6	86.3	29.56
5	1.00	10.7	79.0	29.78	15	1.10	11.2	86.0	29.48
6	1.10	12.9	67.4	29.39	16	0.91	73.3	76.3	29.40
7	1.15	8.3	66.8	29.69	17	0.87	73.3	77.9	29.28
8	1.03	20.1	76.9	29.48	18	0.78	96.6	78.7	29.29
9	0.77	72.2	77.7	29.09	19	0.82	107.4	86.8	29.03
10	1.07	24.0	67.7	29.60	20	0.95	54.9	70.9	29.37

The R output from estimating the multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \quad \Longleftrightarrow \quad E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

is shown below. Here are the least squares estimates:

```
fit = lm(n2o ~ humidity + temperature + pressure)

Coefficients:
(Intercept)     humidity  temperature     pressure
 -3.5078         -0.0026       0.0008       0.1542
```

Here is the ANOVA table for the multiple linear regression:

```
> anova(fit)
Analysis of Variance Table

          Df   Sum Sq  Mean Sq F value    Pr(>F)
Model      3 0.202501 0.067500  21.395 7.609e-06
Residuals 16 0.050479 0.003155
Total     19 0.252980
```

Parts (a), (b), (c), and (d) are on the next two pages.

(a) Estimate the mean amount of nitrous oxide emitted for the conditions where humidity is 50%, temperature is 76 deg F, and barometric pressure is 29.30 inches Hg.

(b) The F statistic (in the ANOVA table) is used to test two hypotheses: H_0 and H_1. Write out what these hypotheses are. You can do this using notation or you can write this out in words. Which hypothesis is more supported by the data?

(c) Calculate R^2 and interpret its value.

Part (d) is on the next page.

(d) I used R to calculate 95% confidence intervals for the population-level regression parameters β_1, β_2, and β_3:

```
> confint(fit,level=0.95)
                2.5%   97.5%
(Intercept)  -9.8778  2.8622
humidity     -0.0040 -0.0012
temperature  -0.0035  0.0051
pressure     -0.0607  0.3690
```

If we displayed the sequential sums of squares corresponding to

```
fit = lm(n2o ~ humidity + temperature + pressure)
```

in the form of

$$\text{SS}_\text{R} = \text{SS}(x_1) + \text{SS}(x_2|x_1) + \text{SS}(x_3|x_1, x_2),$$

would you expect $\text{SS}(x_3|x_1, x_2)$ to be "large" or "small." Explain.

9. Corrosion fatigue refers to the failure of metals due to the combined action of cyclic stress and a corrosive environment. This fatigue usually leads to cracking which can make the metal unusable. In an experiment involving a new type of aluminum metal for use in aerospace applications, engineers wanted to investigate the relationship between

$$Y = \text{number of cycles until failure (in 1000s)}$$

and the two factors

Factor A: coating type (Type I/Type II)
Factor B: humidity (low/high).

A 2×2 factorial treatment structure was used with 6 replications at each treatment combination (so there were $2 \times 2 \times 6 = 24$ pieces of aluminum used in total). Here are the data:

Treatment			Number of cycles until failure (in 1000s)					
Combination	Coating (A)	Humidity (B)	Rep 1	Rep 2	Rep 3	Rep 4	Rep 5	Rep 6
a_1b_1	Type I	Low	114	1236	533	1032	92	211
a_1b_2	Type I	High	78	387	130	466	107	327
a_2b_1	Type II	Low	130	841	1595	1482	529	754
a_2b_2	Type II	High	586	402	846	524	751	529

The goals of the experiment were to learn if coating type and humidity might affect the number of cycles until failure and also if these two factors interact with each other.

I calculated the ANOVA table appropriate for the 2×2 factorial treatment structure:

```
> fit = lm(cycles ~ coating + humidity + coating*humidity)
> anova(fit)

                Df  Sum Sq Mean Sq F value  Pr(>F)
coating          1  754731  754731  4.9293  0.0381
humidity         1  486211  486211  3.1755  0.0899
coating:humidity 1      37      37  0.0002  0.9876
Residuals       20 3062239  153112
```

(a) Interpret the output above. Specifically, which (if any) of the three effects are significant? Use $\alpha = 0.05$ when making your decisions. Use the back of this page if necessary.

Parts (b) and (c) are on the next page.

(b) What would you expect the interaction plot between coating type (A) and humidity (B) to look like? You don't have to construct the plot (although you can if you want). You can simply describe in words how it would look. Provide justification for your answer.

(c) Instead of analyzing the data as we did above, suppose we analyzed the data as data from a one-way classification (Chapter 9) with $t = 4$ treatment groups: a_1b_1, a_1b_2, a_2b_1, and a_2b_2. What would the one-way classification ANOVA table be? Fill in the missing pieces below:

Source	df	SS	MS	F
Treatments	_____	_____	_____	_____
Residuals	_____	_____	_____	
Total	_____	_____		

Hint: The `df` and `SS` columns can be determined from the (two-way) factorial treatment structure analysis on the preceding page. Show your calculations below for partial credit opportunity. Use the back of this page if necessary.

Binomial:

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

Geometric:

$$p_X(x) = \begin{cases} (1-p)^{x-1} p, & x = 1, 2, 3, ... \\ 0, & \text{otherwise.} \end{cases}$$

Negative binomial:

$$p_X(x) = \begin{cases} \binom{x-1}{r-1} (1-p)^{x-r} p^r, & x = r, r+1, r+2, ..., \\ 0, & \text{otherwise.} \end{cases}$$

Hypergeometric:

$$p_X(x) = \begin{cases} \dfrac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}, & x \le K \text{ and } n - x \le N - K \\ 0, & \text{otherwise.} \end{cases}$$

Poisson:

$$p_X(x) = \begin{cases} \dfrac{\lambda^x e^{-\lambda}}{x!}, & x = 0, 1, 2, ... \\ 0, & \text{otherwise.} \end{cases}$$

Exponential:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & \text{otherwise.} \end{cases} \qquad F_X(x) = \begin{cases} 0, & x \le 0 \\ 1 - e^{-\lambda x}, & x > 0. \end{cases}$$

Gamma:

$$f_X(x) = \begin{cases} \dfrac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Normal (Gaussian):

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-(x-\mu)^2/2\sigma^2}, \quad \text{for } -\infty < x < \infty.$$

Weibull:

$$f_T(t) = \begin{cases} \dfrac{\beta}{\eta}\left(\dfrac{t}{\eta}\right)^{\beta-1} \exp\left[-\left(\dfrac{t}{\eta}\right)^{\beta}\right], & t > 0 \\ 0, & \text{otherwise.} \end{cases} \qquad F_T(t) = \begin{cases} 0, & t \le 0 \\ 1 - \exp\left[-\left(\dfrac{t}{\eta}\right)^{\beta}\right], & t > 0. \end{cases}$$