GROUND RULES:

- Print your name at the top of this page.
- This is a closed-book and closed-notes exam.
- You may use a calculator. **Translation:** Show all of your work; use a calculator only to do final calculations and/or to check your work.
- This exam contains 9 questions. Each question is worth 10 points. This exam is worth 90 points.
- Each question contains subparts. 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.
- Any discussion or inappropriate communication between you and another examinee, as well as the appearance of any unnecessary material, will result in a very bad outcome for you.
- 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. On February 27, 2013, the City Council of Cincinnati (OH) passed an ordinance requiring photoelectric smoke detectors in all rental properties. However, over 5 years later, the city's fire department representatives are concerned that not all properties are adhering to the ordinance.

Suppose the population proportion of rental properties in Cincinnati having photoelectric smoke detectors is 0.80 (i.e., 80 percent).

(a) If a sample of 6 rental properties is selected at random, what is the probability at least 5 properties have photoelectric smoke detectors installed?

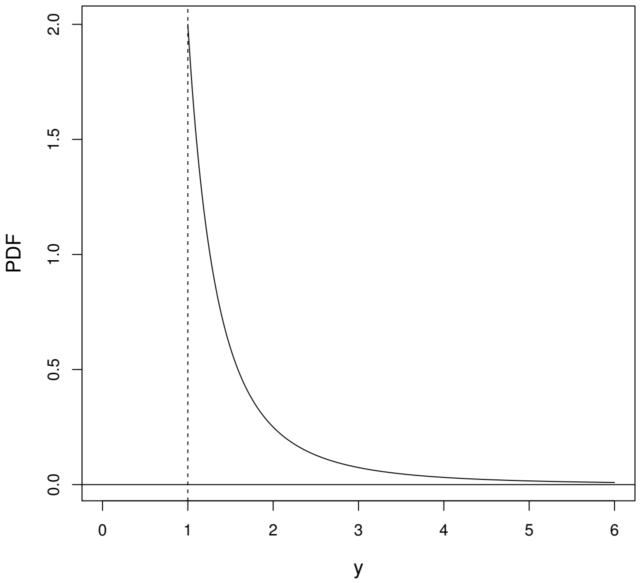
(b) What three Bernoulli trial assumptions did you make in performing the calculation in part (a)?

(c) Instead of sampling a fixed number of rental properties (e.g., like 6, etc.), suppose rental properties were inspected until the first one <u>without</u> photoelectric smoke detectors was found. Under the assumptions you outlined in part (b), what is the distribution of the number of rental properties that would be inspected? Be precise.

2. “Time headway” in traffic flow is the elapsed time between when one car completely passes a fixed point and when the next car begins to pass the same point. Let Y denote this elapsed time (in seconds) for traffic on I-77 during “normal traffic.” Traffic engineers model Y using the probability density function (pdf)

$$f_Y(y) = \begin{cases} \dfrac{2}{y^3}, & y > 1 \\ 0, & \text{otherwise.} \end{cases}$$

A graph of this pdf is shown below:



(a) For two cars randomly selected, calculate the probability the time headway Y will be larger than 3 seconds. Show all calculations.

Two additional questions are on the next page.

(b) Calculate $E(Y)$. Show all calculations. Interpret what $E(Y)$ means in words.

(c) Prepare a graph of the cumulative distribution function $F_Y(y) = P(Y \leq y)$. Use a horizontal axis range of $y = 0$ to 6 (by 1) like on the graph of the pdf (see last page). You don't have to derive the cdf formula, although you can if you want.

3. Time to event studies are common in medical applications. In many of these studies, the event of interest means "death" (e.g., from a serious disease). However, in other studies, the event is something positive. To illustrate, consider a recent study involving patients with venous ulcers (also known as leg ulcers). For one group of $n = 187$ patients, a short-stretch bandage was applied to each patient's infected leg area. The time to event measured on each patient was

$$T = \text{time (in days) until the leg ulcer was completely healed.}$$

Under a Weibull model assumption for T, I estimated the parameters β and η using maximum likelihood; here is the R output:

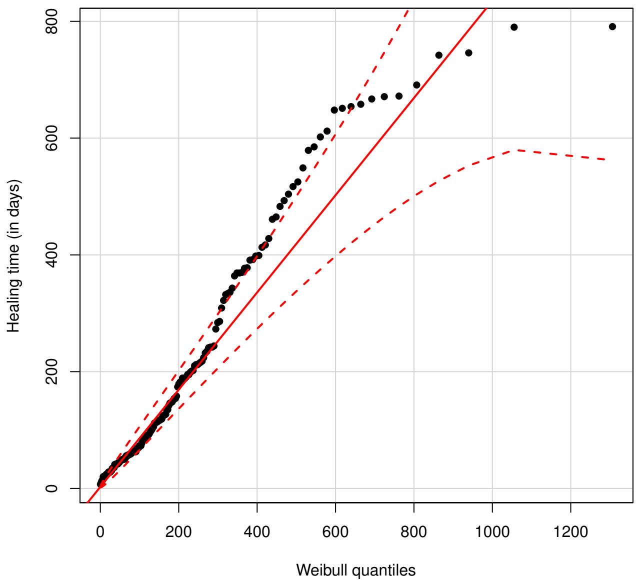
```
> fitdist(healing.times,"weibull")
Parameters:
        estimate  Std.Error
shape   0.99995     0.05601
scale 190.98720    14.79422
```

The estimates of β (shape) and η (scale) are $\widehat{\beta} \approx 1$ and $\widehat{\eta} \approx 191$, respectively.

(a) What distribution is a special case of the Weibull when the shape parameter $\beta = 1$? If β really was 1, what would this say about rate of healing in this population of patients?

(b) Under the estimated Weibull model (with $\widehat{\beta} \approx 1$ and $\widehat{\eta} \approx 191$), calculate the median healing time $\phi_{0.5}$.

Two additional questions are on the next page.



(c) What does the quantile-quantile (qq) plot above suggest about the Weibull model fit for the healing times data?

(d) Name another time-to-event distribution that might be used to model the healing times in this example.

4. A recent article in the *Magazine of Concrete Research* summarized an observational study involving the flexural strength of concrete beams. A random sample of $n = 27$ beams was tested and the strength of each beam Y (measured in MPa) was recorded. Here are the data:

5.9	7.2	7.3	6.3	8.1	6.8	7.0	7.6	6.8	6.5	7.0	6.3	7.9	9.0
8.2	8.7	7.8	9.7	7.4	7.7	9.7	7.8	7.7	11.6	11.3	11.8	10.7	

(a) Although I have not given you much information, describe what one might consider "the population" to be in this example.

(b) A 95 percent confidence interval for the population mean μ uses the formula

$$\overline{y} \pm t_{26,0.025}\frac{s}{\sqrt{27}}.$$

What part of this formula estimates the standard error of the sample mean? the margin of error?

Two additional questions are on the next page.

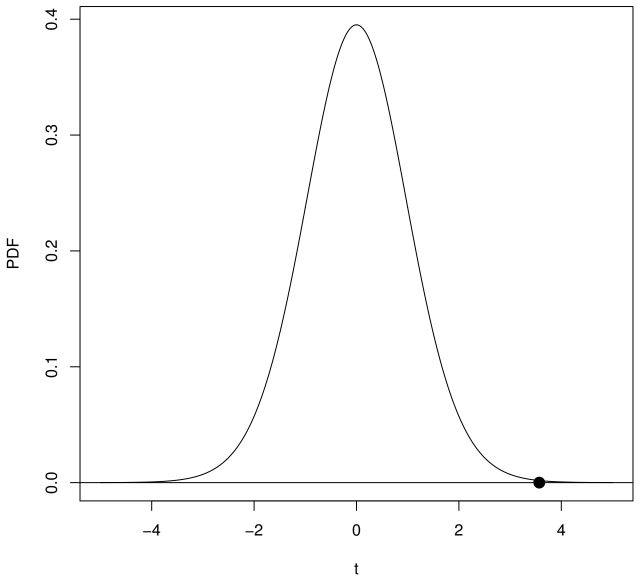
The manufacturer of this type of beam has to demonstrate to customers the population mean is 7.0 MPa. One way to do this would be to write a confidence interval for the population mean (using the data on the previous page) and then determine whether the interval contains 7.0. An equivalent way of doing this would be to calculate the value of

$$t = \frac{\overline{y} - \mu}{s/\sqrt{27}}$$

and then compare t to its sampling distribution when the population mean is 7.0.

(c) What is the sampling distribution of t when the population mean is 7.0? What assumptions are you making here?

(d) When I calculated t and plotted it on its sampling distribution as described above, here is what I observed:



Is this result ($t \approx 3.57$) most consistent with the population mean being equal to, smaller than, or larger than 7.0 MPa? Explain. Use the back of this page if necessary.

5. The NFL's Scouting Combine provides an opportunity for participants to display their professional football potential. A special teams coach at this year's event was interested in comparing the population mean punting distance (in yards) between two types of footballs:

Type 1: Air-filled footballs
Type 2: Helium-filled footballs.

To learn how the population mean punting distances might compare, he recruited $n = 15$ punters and had each of them punt each type of ball. The order in which each punter punted an air-filled football and a helium-filled football was randomized. Here were the observed punt distances (measured in yards):

Punter	Air	Helium
1	56.4	55.2
2	44.8	47.9
3	43.7	41.8
4	37.1	38.3
5	33.8	33.5
6	37.1	40.2
7	39.9	43.2
8	33.2	35.5
9	42.8	46.1
10	48.7	50.6
11	32.7	37.3
12	44.6	48.1
13	40.4	40.3
14	44.9	46.1
15	46.1	47.5

The special teams coach is good at coaching but he forgot his statistics course. Searching through an online resource, he found the words "two sample t confidence interval" and "independent samples" and then calculated a 95 percent confidence interval for $\mu_1 - \mu_2$ to be $(-6.35, 2.99)$ yards. I reproduced his analysis in R:

```
> t.test(air,helium,conf.level=0.95,var.equal=TRUE)$conf.int
[1] -6.35  2.99
```

(a) Explain what is wrong with this analysis.

Two additional questions are on the next page.

(b) To compare the population means μ_1 and μ_2 using a confidence interval, the correct analysis involves analyzing the data differences on each punter; i.e.,

$$D_i = \text{Air}_i - \text{Helium}_i,$$

for $i = 1, 2, ..., 15$. Here is this analysis in R:

```
> diff = air-helium
> t.test(diff,conf.level=0.95)$conf.int
[1] -2.72 -0.63
```

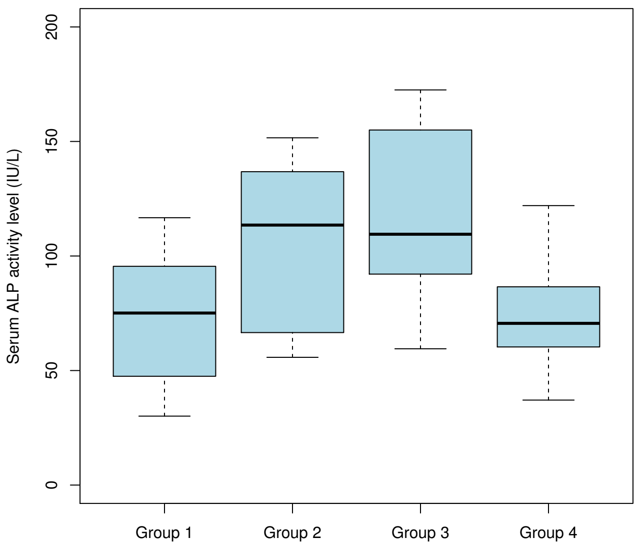
Interpret this interval and describe what this suggests about how μ_1 and μ_2 compare.

(c) Provide a statistical explanation why the confidence interval for $\mu_1 - \mu_2$ in part (b) is so much shorter than the incorrect interval in part (a).

6. An observational study was performed to compare the population mean serum alkaline phosphatase (ALP) levels in children with seizures who were receiving anticonvulsant therapy. Forty-five children were found for the study and were categorized into one of four drug groups:

Group 1:	Control (no anticonvulsant drug and/or no history of having seizures)
Group 2:	Phenobarbital
Group 3:	Carbamazepine
Group 4:	Other anticonvulsants.

Using a blood sample from each child, the serum ALP level was recorded (in IU/L, international units per liter). Here are the data shown using side-by-side boxplots:



Let μ_i denote the population mean serum ALP level for the ith group ($i = 1, 2, 3, 4$). I used R to perform an analysis of variance (ANOVA) to test

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$
versus
$$H_1 : \text{the population means } \mu_i \text{ are not all equal.}$$

Here is the output:

```
> anova(lm(alp.level~group))

Analysis of Variance Table
          Df Sum Sq Mean Sq F value   Pr(>F)
group      3  15509  5169.7  5.2435 0.003722 **
Residuals 41  40423   985.9
```

(a) Is the overall F statistic ($F \approx 5.24$) consistent with what we would expect when H_0 is true or when H_1 is true? Explain. Cite the probability value (p-value $\approx$ 0.004) in defending your decision.

(b) As a follow-up, one investigator is interested in comparing the population mean ALP level for children taking phenobarbital (Group 2) to the population mean ALP level for the control group (Group 1).

Here is the 95 percent Tukey confidence interval for this population mean difference:

```
> TukeyHSD(aov(lm(alp.level˜group)),conf.level=0.95)

Tukey multiple comparisons of means
95% family-wise confidence level
                   diff    lwr    upr  p adj
group.2-group.1  28.58  -5.15  62.33   0.12
```

The values `lwr` and `upr` are the lower and upper limits of this interval. Interpret the interval. How do the population mean ALP levels compare for these two groups?

Two additional questions are on the next page.

(c) Interestingly, had the investigator not done the overall ANOVA and just focused on the Groups 2 and 1 to begin with, a 95 percent confidence interval for $\mu_2 - \mu_1$ based on the independent sample and equal population variance assumptions would be

```
> t.test(group.2,group.1,conf.level=0.95,var.equal=TRUE)$conf.int
[1]  3.75 53.43
```

Why is this interval (and its conclusion) so different than the interval in part (b)?

(d) The same investigator in part (c) asks you the following question:

> *"If the analysis of variance procedure is designed to compare population of means, why isn't the procedure called the analysis of means?"*

How would you respond?

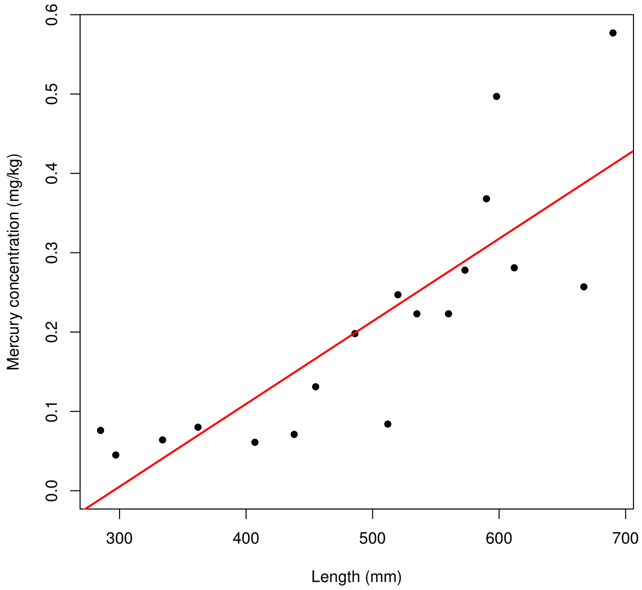
7. Mercury can accumulate in fish tissue over time which, in turn, can pose a public health risk to humans who consume fish. Researchers at the Florida Fish and Wildlife Conservation Commission recently sampled $n = 18$ scamp grouper fish from the Gulf of Mexico and measured the following two variables on each fish:

$$\begin{aligned} Y &= \text{mercury concentration (mg/kg)} \\ x &= \text{length (mm)}. \end{aligned}$$

One goal was to model Y as a function of x using simple linear regression; i.e.,

$$Y_i = \beta_0 + \beta_1 x + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Below is a scatterplot of the data for the 18 fish caught; superimposed on the scatterplot is the least-squares regression line.



(a) Although I have not given you much information, describe what one might consider "the population" to be in this example.

Three additional questions are on the next page.

(b) The least squares estimates of β_0 and β_1 are given in the output below:

```
> fit = lm(mercury˜length)
Coefficients:
(Intercept)       length
   -0.30733      0.00104
```

Ninety-five percent confidence intervals for β_0 and β_1 are given below:

```
> confint(fit,conf.level=0.95)
                2.5 %   97.5 %
(Intercept) -0.49921 -0.11546
length       0.00067  0.00142
```

Does this analysis demonstrate that mercury concentration and length are linearly related in the population? Explain.

(c) I used R to calculate $R^2 \approx 0.683$ for these data. Explain what this means and why we might use caution in interpreting it.

(d) The researchers would like to infer on the subpopulation of scamp grouper fish whose length is $x = 100$ mm. Use the model fit in part (b) to estimate the mean mercury content for this subpopulation and comment. Use the back of this page if necessary.

8. A recent article in the *Journal of Air and Waste Management Association* described an observational study in Kaohsiung City, Taiwan. The goal was to develop a multiple linear regression model to explain how the response variable

$$Y = \text{ energy content of municipal solid waste specimen}$$

was related to four independent variables

$$\begin{aligned} x_1 &= \text{plastic by weight (measured as a \% of the total weight)} \\ x_2 &= \text{paper by weight (measured as a \% of the total weight)} \\ x_3 &= \text{garbage by weight (measured as a \% of the total weight)} \\ x_4 &= \text{moisture percentage.} \end{aligned}$$

The authors of the article describe how $n = 30$ municipal solid waste specimens were available to estimate the multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The energy content Y (kcal/kg) was measured on each waste specimen after it was incinerated.

Based on the $n = 30$ waste specimens, I used R to estimate the model above using least squares, reproducing the estimates reported by the authors:

```
> fit
Coefficients:
(Intercept)      plastic        paper      garbage     moisture
   2244.923       28.925        7.644        4.297      -37.354
```

(a) The first waste specimen in the authors' data set had independent variable measurements

$$x_1 = 18.69 \quad x_2 = 15.65 \quad x_3 = 45.01 \quad x_4 = 58.21.$$

That is, about 19% of the specimen was `plastic`, 16% was `paper`, and 45% was `garbage` (so about 20% of the specimen was "other" waste). The `moisture` content of the combined specimen was measured to be about 58%.

The energy content Y for the first waste specimen was $Y = 947$. Use the least-squares estimates in the R output above to calculate the predicted value $\widehat{Y}$ and residual e for this first specimen. Show your work.

Two additional questions appear on the next two pages.

(b) Here is the sequential sums of squares breakdown from estimating the model:

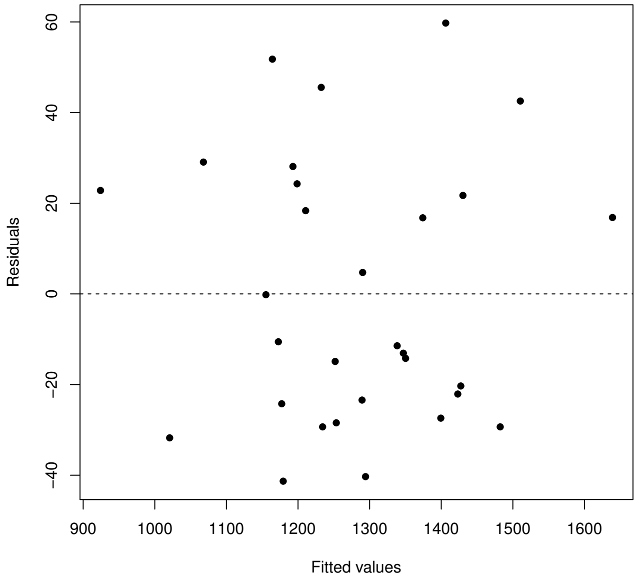
```
> fit = lm(energy ~ plastic + paper + garbage + moisture)
> anova(fit)

Analysis of Variance Table
          Df Sum Sq Mean Sq  F value    Pr(>F)
plastic    1 239735  239735 241.8709  2.31e-14
paper      1  11239   11239  11.3392   0.00245
garbage    1   2888    2888   2.9136   0.10023
moisture   1 411069  411069 414.7313  2.20e-16
Residuals 25  24779     991
```

The p-value for `garbage` is somewhat large (p-value ≈ 0.1002). However, a 95 percent confidence interval for β_3, the regression parameter attached to `garbage`, is

```
> confint(fit,conf.level=0.95)
              2.5 %  97.5 %
garbage        0.35    8.24
```

which does not contain 0. Are these analyses giving contradictory conclusions? Explain.

(c) Here is the residual plot for the least-squares regression model fit:



What does this plot say about the quality of the model fit for these data? Are there any glaring model deficiencies? Explain.

9. Civil engineers performed a 2^2 factorial experiment to investigate how the fracture toughness of an asphalt specimen (Y, measured in MPa) depends on two factors: mixture type (Factor A) and temperature (Factor B). Here are the data from the experiment:

Mixture Type (A)	Temperature (B)	Replicate 1	Replicate 2	Replicate 3
Normal	−35 deg C	15.8	15.6	14.9
Polymer added	−35 deg C	13.7	13.8	13.2
Normal	−10 deg C	13.3	13.9	12.8
Polymer added	−10 deg C	15.6	15.9	16.6

(a) Ignoring the factorial treatment structure, I analyzed the data as data from a one-way classification with four treatment groups, like we did in Chapter 9:

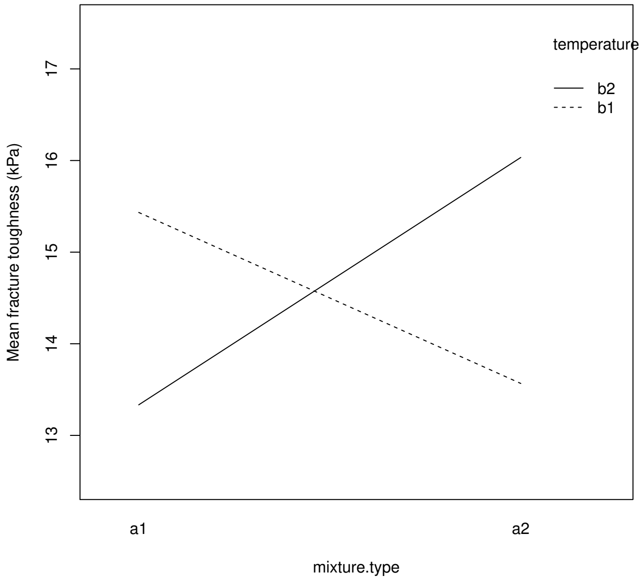
```
> anova(lm(fracture.toughness~treatment))

Analysis of Variance Table
          Df  Sum Sq Mean Sq F value    Pr(>F)
treatment  3 16.2625  5.4208  24.272 0.0002267
Residuals  8  1.7867  0.2233
```

The overall F statistic (here $F \approx 24.3$) tests which two hypotheses? You can write your answer out in words, or you can use statistical symbols. If you use symbols, define what the symbols mean.

(b) Acknowledging the factorial treatment structure, let SS_A, SS_B, and SS_{AB} denote the sums of squares for the main effect of mixture type, the main effect of temperature, and the interaction effect between mixture type and temperature, respectively. What is $\text{SS}_A + \text{SS}_B + \text{SS}_{AB}$?

Two additional questions are on the next page.

(c) Here is the interaction plot between mixture type (A) and temperature (B):



Consider the following two hypotheses:

H_0: mixture type and temperature do not interact in the population
H_1: mixture type and temperature interact in the population.

The interaction plot above makes a strong argument for which hypothesis? Explain.

(d) In addition to mixture type (A) and temperature (B), suppose the engineers wanted to include two additional factors in the experiment:

C: manufacturer (M1 and M2)
D: air void percentage (4 percent and 6 percent).

With four factors now (each with two levels), how many asphalt specimens would be needed to complete three full replications?

Binomial:

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

Geometric:

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

Negative binomial:

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

Hypergeometric:

$$p_Y(y) = \begin{cases} \dfrac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}, & y \leq r \text{ and } n - y \leq N - r \\ 0, & \text{otherwise.} \end{cases}$$

Poisson:

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

Exponential:

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

Gamma:

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

Normal (Gaussian):

$$f_Y(y) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/2\sigma^2}, & -\infty < y < \infty \\ 0, & \text{otherwise.} \end{cases}$$

Weibull:

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