1) If two events, A and B, are *statistically independent*, is there any sense in using observations of B to forecast or predict A? Why or why not? Use conditional probabilities to support your answer.

2) Suppose you flip two coins and roll a fair die. Every "random experiment" here is independent of the others. What is the probability that the number of heads you get is greater than or equal to the number from the dice roll?

3) X can take the values $\{1, 2, 3\}$ with probabilities $\{0.6, 0.2, 0.2\}$. Compute $\mathbb{E}[X]$ and $\mathbb{V}[X]$.

4) Your friends, Rihanna and Beyonce, take turns giving you gifts for your birthday. You enjoy 30% of the gifts Rihanna gives you, and 50% of Beyonce's gifts. This year, they flip a fair coin to decide who will buy you two presents. Let $B \in \{0, 1\}$ indicate whether Beyonce bought your presents this year, and let Y be the number of gifts you enjoy.

1. Write down the joint distribution (as a table or a formula, whichever you prefer) of Y and B
2. Calculate $\mathbb{E}[Y|B = 0]$, $\mathbb{E}[Y|B = 1]$, and $\mathbb{E}[Y]$.
3. What is the distribution of the number of gifts you enjoy, *given* that Beyonce gave you the gifts $(B = 1)$.
4. You enjoyed both gifts this year, would you reject the Null Hypothesis that Rihanna chose the gifts when conducting a test of size 10%. What do you notice about this case that is different from the usual set up of Hypothesis Testing?

5) Define the following jointly distributed random variables X and Y according to the joint distribution:

$$P(x, y) = \begin{cases} 1/3 & \text{for } (x, y) = (-1, 1) \\ 1/3 & \text{for } (x, y) = (0, 0) \\ 1/3 & \text{for } (x, y) = (1, 1) \end{cases}$$

Part (a): show that $\mathbb{C}(X, Y) = 0$. Part (b): show that X and Y are not independent.

6) Beyonce runs a Lemonade stand in Houston, where her revenue every hour (in dollars) is distributed as a normal with mean 20 and variance 16.

- What is the probability that her revenue is greater than \$18?
- What is the probability that her revenue is between \$19 and \$24?
- If she sells glasses at 50c per glass, what is her expected number of glass sold in an hour? Remember that *revenue* = *price* × *items sold*. Is the number of glasses sold in an hour also normal? Explain.

CDF values for a standard normal, Z:

z	.0	0.25	0.5	0.75	1.0
$F(z)$	0.5	0.598	0.691	0.773	0.841

7) Let some data $X_1, X_2, ..., X_N$ be drawn from a normal distribution with mean μ_X and variance σ_X^2.

1. How is the sample mean, $\overline{X}_N$, distributed?
2. Define $S = X_1 + X_2$. Is S a statistic? How is S distributed?
3. Define $U = X_1 + X_2 + X_3$. Is U a statistic? How is U distributed?
4. Suppose that the observations, X_i, are not drawn from a normal distribution. What idea can we use to approximate the distribution of $\overline{X}_N$? What will this approximation be?

8) Let each observation of some dataset, $X_1, X_2, ..., X_N$, be drawn from a uniform distribution with lower bound 10 and upper bound 20. Remember that when X is a uniform distribution on the interval $[a, b]$, $\mathbb{E}[X] = (a + b)/2$ and $\mathbb{V}[X] = (b - a)^2/12$

1. For each observation X_i, what is $\mathbb{E}[X_i]$? What is $\mathbb{V}[X_i]$?
2. Based on your answer above, what is $\mathbb{E}[\overline{X}_N]$? What is $\mathbb{V}[\overline{X}_N]$?
3. If N was from a large sample, how could you approximate the distribution of $\overline{X}_N$?

9) Let $X_1, X_2, ..., X_N$ be data drawn from a normal distribution with mean 10 and standard deviation 4.

1. Define $T = X_1 + X_2 + X_3$. Is T a statistic? What is the distribution of T?
2. Write a formula for two numbers, a and b, such that $P[a \leq T \leq b] = 0.95$. Use that $P[Z \leq -1.96] = 0.025$, where Z is a standard normal (i.e. $z_{.025} = 1.96$).

10) (hard) Let $X_1, X_2, ..., X_8$ be data drawn from a normal distribution with mean 5 and standard deviation 1. Define the following:

$$T_8 = X_1 - X_2 + X_3 - X_4 + X_5 - X_6 + X_7 - X_8$$

1. Is T_8 a statistic? What is its sampling distribution?
2. Following the above logic, for a sample of size N, where N is any even number, define:

   $$T_N = \sum_{i \text{ is odd}} X_i - \sum_{i \text{ is even}} X_i = X_1 - X_2 + X_3 - ... + X_{N-1} - X_N$$

   What is the sampling distribution of T_N?

11) The length of Taylor Swift's grudges follows an exponential[1] distribution, and her average (i.e. mean) grudge lasts 5 years.

1. Exactly 3 years ago, at a *Taylor's Version* listening party, she overhears you making fun of her cats. What is the probability that she still holds a grudge against you?
2. You are going to collect a dataset consisting of each of her 1000 grudges, and the length of time each grudge lasts. Given this large sample size, what distribution will the sample mean from this dataset approximately follow? (Hint: the variance of the exponential is $\frac{1}{\lambda^2}$).

[1] The cdf of the exponential distribution is $F(x) = 1 - \exp(-\lambda x)$, and $\mathbb{E}[X] = \frac{1}{\lambda}$.

12) You are asked to estimate the population mean and variance (μ, σ^2) of workers' wages given a random sample of wages, $W_1, W_2, ..., W_N$, from the population. You have $N = 50$ observations. Suppose you get that $\overline{W} = \$10/hour$, with $s^2 = 9$.

1. Write a 95% confidence interval for μ that can be computed from this information (one of the critical values below).

2. Test the null hypothesis that $\mu = 12$ against the alternative that it is smaller than 12. Use size $\alpha = 0.05$, and use one of: $z_{0.025} = 1.96$,$z_{0.05} = 1.64$.

13 You are asked to estimate the unemployment rate, using a sample of $N = 100$ workers, who report their employment status (either employed or unemployed).

1. Describe how you would use this data to estimate the unemployment rate. Describe how you would compute a $(1 - \alpha) \times 100\%$ confidence interval.

2. You estimate a sample proportion of unemployed workers of 7.6%. Give the 95% confidence interval for the true unemployment rate.

14) Suppose you have data on the annual earnings of $N = 500$ households, $E_1, E_2, ..., E_{500}$. This is a random sample from the population. The policy group you work for deems a household to be living in poverty if their earnings fall below the poverty line, \$10,000/year. Describe how you would use this data to estimate the poverty rate (i.e. the population proportion of households living in poverty). Give an estimate as well as a formula for the $(1 - \alpha) \times 100\%$ confidence interval.

15) Calculate the bias and the variance of the following estimators.

1. $\theta = \overline{X} + a$ where a is a constant, as an estimator for μ.

2. $\theta = \overline{X} + \epsilon$ where ϵ is independent of $\overline{X}$ and $\mathbb{E}[\epsilon] = 0$, as an estimator for μ.

16) You have data from a random sample of firms in the market for widgets. In particular, you have output Y_i, and costs C_i for each firm i. Your data contains this information for a random sample of N firms. You know that every firm sells their widgets at the market price of \$2, and so profit for each firm can be defined as:

$$\Pi_i = 2Y_i - C_i$$

Assume that Y_i and C_i are each *normally distributed*, with *unknown* variance.

1. What kind of distribution will Π have? Use the parameters $\mathbb{E}[\Pi] = \mu_\pi$ and $\mathbb{V}[\Pi] = \sigma_\pi^2$ to describe this distribution.

2. How would you estimate a 95% confidence interval for μ_π?

3. Do you need to assume that Y_i and C_i are independent at any given firm i for this to work?

17) Two sample means $\overline{X}_1$ and $\overline{X}_2$ are calculated from random samples from the *same* population. The only difference is the sample size $N_1 < N_2$. Which is the more efficient estimator (i.e. which estimator has smaller variance) of the population mean, μ? Why?

18) Two sample means $\overline{X}_1$ and $\overline{X}_2$ are calculated from random samples from *different* populations with the *same* population mean, μ. The population variances are different, though, with $\sigma_1^2 < \sigma_2^2$. Which sample mean is a more efficient estimator for μ? Why?

19) Let X and Y both be *independent* random variables. Both are *normally distributed* with means μ_X, μ_Y and variances σ_X^2, σ_Y^2. Let $\overline{X}$ and $\overline{Y}$ be *sample means* of these random variables, from samples of size N_X and N_Y. Your task is to write down the *sampling distribution* of the following statistics.

1. $\overline{X}$ and $\overline{Y}$
2. $T = X + a$ where a is a constant.
3. $T = \overline{X} + Y$
4. $T = \overline{Y} + X$
5. $T = aX + bY$ where a and b are constants.
6. $T = \overline{X} - \overline{Y}$

20) Suppose we are interested in estimating the mean wage, μ, for a population. We collect a random sample of wages, $W_1, W_2, ..., W_n$ from this population. **However**, suppose that people who earn less than \$8/hour do not report their earnings, and are dropped from the sample.

1. What is the *effective* probability distribution we are drawing from, if we cannot observe wages of $w < 8$? (Hint: the answer involves a conditional distribution).
2. Given this new distribution, what is the expected value of a single (non-missing) observation W taken from the population?
3. What is the expected value of $\overline{W}$, the sample mean from this truncated sample?
4. Consider the following rule: $\mathbb{E}[X|X > c] > \mathbb{E}[X]$ for any constant c. What does this imply for $\overline{W}$ from this sample as an estimator for μ?

21) Suppose that unemployment durations (in weeks) are distributed according to an exponential (that is, the wait times until the arrival of a job offer are exponentially distributed) with a rate parameter λ. Let T be the unemployment duration of one individual. Recall that:

$$\mathbb{E}[T] = \frac{1}{\lambda}, \ \mathbb{V}[T] = \frac{1}{\lambda^2}$$

1. Let $\overline{T}$ be the sample mean from a random sample of the population of unemployed workers. Calculate $\mathbb{E}[\overline{T}]$ and $\mathbb{V}[\overline{T}]$.
2. What does the central limit theorem say about $\overline{T}$ as our sample size grows?
3. Propose an estimator for λ based on your answer to part (a).
4. Combine parts (2) and (3) to propose an approximate 95% confidence interval for λ.

22) Let p be the true proportion of UMN students who like pineapple on pizza. Let $\overline{p}$ be the fraction of students from a random, iid, sample of size N that report that they like pineapple on pizza.

1. What is the approximate sampling distribution of $\hat{p}$?
2. Construct a two-sided hypothesis test (with size α) of the null hypothesis that $p = p_0$, where p is the unknown population proportion.

23) Suppose that you have the data on some health measure, Y, for a group of treated individuals and a group of untreated individuals. Let $\overline{Y}_T$ be the sample mean of the treated, and let $\overline{Y}_N$ be the sample mean of the untreated. The *effectiveness* of the drug in the population can be written as $\mu = \mathbb{E}[Y_T] - \mathbb{E}[Y_N]$. When the population variance across the two populations is *equal, but unknown*, describe how you would construct a hypothesis test (with significance α) of the following:

1. The null that the drug has *no effect* against the alternative that it has *some effect*
2. The null that the drug has *no effect* against the alternative that it has a *positive effect*

24) Suppose that a statistic (it could be anything) is distributed such that:

$$P[-\frac{a_\alpha}{\sqrt{N}} < T - \gamma < \frac{a_\alpha}{\sqrt{N}}] = 1 - \alpha$$

Where γ is a population parameter of interest. Use this fact to:

- Write a hypothesis test, with Type I error probability α, of the Null hypothesis that $\gamma = 0$ against the alternative $\gamma \neq 0$
- Write a hypothesis test, with Type I error probability α, of the Null hypothesis that $\gamma = 2$ against the alternative $\gamma \neq 2$
- Calculate the power of your first test above to correctly reject the Null when $\gamma = 2$. Write your answer in terms of the CDF of T (i.e. in terms of a function F such that $P[T \leq t] = F(t)$)
- What happens to the power of your test as N increases? Why?