18.650. Fundamentals of Statistics
Fall 2023. Problem Set 3

Due Monday, November 13

Problem 1

Let $X_1, \ldots, X_n$ be an i.i.d sample of $N(0,1)$ and recall that the CDF of $N(0,1)$ is denoted by Φ. Let $X_1^*, \ldots, X_m^*$ be a bootstrap sample from the X_is. Denote by $\hat{F}_n$ and $\hat{F}_m^*$ their respective empirical CDF. Hereafter, we fix $x \in \mathbb{R}$.

1. Compute the probability that two of the X_is are equal.
   Since the X_is are drawn from a continuous distribution, the probability that any two are equal is $\boxed{0}$.
2. Compute the probability that $X_1^* = X_2^*$.
   By the tower property of conditional expectation, we have
   $$\mathbb{P}(X_1^* = X_2^*) = \mathbb{E}[\mathbb{P}(X_1^* = X_2^* \mid X_1, \ldots, X_n)].$$
   Since the probability that any two of the X_i's are equal to each other is zero, we can assume the X_i are all distinct when computing the inner conditional probability. Therefore, the only way that $X_1^* = X_2^*$ is if both equal the same X_j. So we get
   $$\mathbb{P}(X_1^* = X_2^* \mid X_1, \ldots, X_n) = \sum_{j=1}^{n} \mathbb{P}(X_1^* = X_j, X_2^* = X_j \mid X_1, \ldots, X_n) = \sum_{j=1}^{n} \frac{1}{n} \times \frac{1}{n} = \frac{1}{n}.$$
   Here we used independence of X_1^* and X_2^*, and the fact that each of them is uniformly distributed. Since $\mathbb{P}(X_1^* = X_2^* \mid X_1, \ldots, X_n) = 1/n$ for each fixed $X_1, \ldots, X_n$, the expectation over $X_1, \ldots, X_n$ is also $\boxed{1/n}$
3. Compute the conditional probability $\mathbb{P}(X_1^* \leq x | X_1, \ldots, X_n)$.
   This is equal to $\boxed{\hat{F}_n(x)}$, as we sample according to the empirical CDF.
4. Compute the unconditional probability $\mathbb{P}(X_1^* \leq x)$.
   One way to compute this is the following, noting that X_1^* must equal one of the X_i's, which are i.i.d. standard normal random variables:
   $$\mathbb{P}(X_1^* \leq x) = \sum_{i=1}^{n} \mathbb{P}(X_1^* \leq x \mid X_1^* = X_i) \cdot \mathbb{P}(X_1^* = X_i) = \sum_{i=1}^{n} \frac{\mathbb{P}(X_i \leq x)}{n} = \boxed{\Phi(x)}.$$

5. Write
$$\hat{F}_m^*(x) = \frac{1}{m}\sum_{i=1}^{m} B_i$$
as an average of Bernoulli random variables $B_1, \ldots, B_m$. What is the parameter of B_i?
The parameters are all $\boxed{\hat{F}_n(x)}$, as that is the probability that $X_i^* \le x$.

6. Compute the conditional distribution of $m\hat{F}_m^*(x)$ given $X_1, \ldots, X_n$.
Since $\hat{F}_m^*(x)$ is the average of Bernoulli random variables, $m\hat{F}_m^*(x)$ is thus a binomial random variable $\boxed{\text{Binomial}(m, \hat{F}_n(x))}$.

7. Compute the conditional expectation $\mathbb{E}[\hat{F}_m^*(x)|X_1, \ldots, X_n]$.
From the above, along with division by m, the expectation is thus $\boxed{\hat{F}_n(x)}$.

8. Compute the unconditional expectation $\mathbb{E}[\hat{F}_m^*(x)]$.
Note that the expected value of $\hat{F}_n(x)$ is $\Phi(x)$, as we can similarly write $\hat{F}_n(x)$ as the average of Bernoulli random variables with parameter $\Phi(x)$. Thus, by the law of total expectation,
$$\mathbb{E}[\hat{F}_m^*(x)] = \mathbb{E}[\mathbb{E}[\hat{F}_m^*(x)|X_1, \ldots, X_n]] = \mathbb{E}[\hat{F}_n(x)] = \boxed{\Phi(x)}.$$

9. For $i \neq j$, compute the (unconditional) covariance $\text{cov}(B_i, B_j)$ of two Bernoullis above. Are they independent?
Note that $\mathbb{E}[B_i] = \mathbb{E}[\mathbb{E}[B_i \mid X_1, \ldots, X_n]] = \mathbb{E}[\hat{F}_n(x)] = \Phi(x)$ for all i, and
$$\mathbb{E}[B_iB_j] = \mathbb{E}[\mathbb{E}[B_iB_j \mid X_1, \ldots, X_n]] = \mathbb{E}[\mathbb{P}(X_i^* = X_j^* = 1 \mid X_1, \ldots, X_n)]$$
$$= \mathbb{E}[(\hat{F}_n(x))^2] = \mathbb{E}[(\hat{F}_n(x))]^2 + \mathbb{V}(\hat{F}_n(x)) = \Phi(x)^2 + \frac{\Phi(x)(1-\Phi(x))}{n}$$
for all $i \neq j$ (again using the fact that $\hat{F}_n(x)$ is the average of Bernoulli random variables), so the covariance is
$$\mathbb{E}[B_iB_j] - \mathbb{E}[B_i]\mathbb{E}[B_j] = \left(\Phi(x)^2 + \frac{\Phi(x)(1-\Phi(x))}{n}\right) - \Phi(x)^2 = \boxed{\frac{\Phi(x)(1-\Phi(x))}{n}}.$$
Thus, they are not independent.

10. Compute the (unconditional) variance $\mathbb{V}(\hat{F}_m^*(x))$. Compare it with the variance $\mathbb{V}(\hat{F}_n(x))$.
Note that $\mathbb{E}[B_i^2] = \mathbb{E}[B_i] = \Phi(x)$, as each B_i only takes on either 0 or 1, so $\mathbb{V}(B_i) = \Phi(x)(1-\Phi(x))$. We also had that $\text{cov}(B_i, B_j) = \frac{\Phi(x)(1-\Phi(x))}{n}$ from before.

Thus,

$$\mathbb{V}\left(\sum_{i=1}^{m} B_i\right) = \sum_{i=1}^{m} \mathbb{V}(B_i) + \sum_{\substack{1\le i,j\le m \\ i\neq j}} \operatorname{cov}(B_i, B_j) = \Phi(x)(1-\Phi(x))\left(m + \frac{m(m-1)}{n}\right);$$

dividing by m yields $\mathbb{V}(\hat{F}^*_m(x)) = \boxed{\Phi(x)(1-\Phi(x))\left(\frac{n+m-1}{nm}\right)}$. This is always at least $\mathbb{V}(\hat{F}_n(x))$, as $n+m-1 \geq m$.

Problem 2

Let $X_1, \ldots, X_n$ be i.i.d $\mathsf{Unif}[0,\theta]$ for some unknown $\theta > 0$. We want to test

$$H_0 : \theta = 1 \quad \text{vs.} \quad \theta > 1$$

1. Compute the maximum likelihood estimator $\hat{\theta}$ of θ.
   Solution: Note that the likelihood is

   $$L_n(X_1, \ldots, X_n; \theta) = \frac{1}{\theta^n} \mathbb{1}(X_1, X_2, \ldots, X_n \in [0,\theta]).$$

   As the derivative of $\frac{1}{\theta^n}$ is negative, we wish to minimize θ, but we have the constraint that $\theta \geq \max_i X_i$ (otherwise we would have some $X_i \geq \theta$ which causes the indicator function, and thus the likelihood, to be 0). Thus, our MLE is $\hat{\theta} = \boxed{\max_i X_i}$.
2. Why can't we use the Wald test here?
   Solution: The support of $\mathsf{Unif}[0,\theta]$ depends on θ, so the MLE is not asymptotically normal. Thus, we can't use the Wald test, which requires asymptotic normality.
3. Consider a rejection of region of the form $\hat{\theta} > c_\alpha$. What choice of c_α yields a test of size $\alpha \in (0,1)$?
   Solution: We use the fact that, under the null hypothesis, and assuming $0 < c_\alpha < 1$, the probability

   $$\mathbb{P}(\hat{\theta} > c_\alpha) = \mathbb{P}(\max_i X_i > c_\alpha) = 1 - \mathbb{P}(\max_i X_i \leq c_\alpha) = 1 - c_\alpha^n;$$

   since we want size α, we need $1 - c_\alpha^n = \alpha \implies \boxed{c_\alpha = \sqrt[n]{1-\alpha}}$.

4. We observed $\hat{\theta} = 0.97$ with $n = 20$ observations. What is the p-value? How much evidence is there against H_0?
Solution: Here, the p-value satisfies

$$p = \mathbb{P}(\hat{\theta} \geq 0.97) = 1 - \mathbb{P}(\max_i X_i < 0.97) = \boxed{1 - 0.97^{20} \approx 0.456};$$

there is little or no evidence against H_0.

5. We observed $\hat{\theta} = 1.03$ with $n = 20$ observations. What is the p-value? How much evidence is there against H_0?
Solution: Here, note that $\max_i X_i \leq 1$ with probability 1 under H_0, so observing $\hat{\theta} = 1.03 > 1$ means that our p-value is $\boxed{0}$, and there is very strong evidence against H_0.

Problem 3

Let $X_1, \ldots, X_n$ be n i.i.d. $N(\theta, 1)$ for some unknown θ. We want to test

$$H_0 : \theta = 0 \quad \text{vs.} \quad \theta \neq 0$$

1. Assume that the Wald test based on $\bar{X}_n$ is chosen to have size α and let $\beta(\theta)$ denote the power function. For each $\theta \in \mathbb{R}$, compute

$$\lim_{n\to\infty} \beta(\theta)$$

Solution: Note that $\sqrt{n}(\overline{X_n} - \theta) \sim \mathcal{N}(0, 1)$, as the sum of Gaussians is a Gaussian, so our test statistic, using θ under H_0, would be $T_n = |\sqrt{n} \cdot \overline{X_n}|$, and our test would be $\mathbf{1}(T_n > q_{\alpha/2})$.
Now, for $\theta = 0$, since $\sqrt{n} \cdot \overline{X_n} \sim \mathcal{N}(0, 1)$, the probability of rejecting would be the probability that a standard normal random variable Z has absolute value at least $q_{\alpha/2}$, or α. Thus, $\boxed{\beta(0) = \alpha}$ regardless of the value of n.
Otherwise, since $\sqrt{n} \cdot (\overline{X_n} - \theta) \to \mathcal{N}(0, 1)$, we have

$$\sqrt{n}\overline{X_n} \approx N(\theta\sqrt{n}, 1);$$

since $\theta\sqrt{n} \to \pm\infty$ as $n \to \infty$, the probability that $|T_n| > q_\alpha$ goes to 1 as $n \to \infty$, so

$$\boxed{\lim_{n\to\infty} \beta(\theta) = 1} \qquad \text{for } \theta \neq 0.$$

2. We observe $\bar{X}_n = 0.24$ for $n = 80$ observations. What is the p-value of the Wald test?
Solution: From before, the test statistic would be

$$T_{80} = \sqrt{80} \cdot 0.24 \approx 2.14663;$$

we get a p-value of $\boxed{0.032}$ (via consulting a table or calculator).

Problem 4

We want to test if juries are representative of the racial distribution in a given county. To that end, we collected data and recorded the race of 275 randomly selected jurors in this county. The following table represents both the data collected and the true proportion of each race in this county.

Race	White	Black	Hispanic	Other	Total
# jurors	205	26	25	19	275
proportion in county	0.72	0.07	0.12	0.09	1

Choose an appropriate test to verify that race representation is adequate in this county (set up hypotheses, test statistic, and compute the p-value).

Solution: We'll use Pearson's χ^2 test. Let $p = (p(\text{White}), p(\text{Black}), p(\text{Hispanic}), p(\text{Other}))$ denote the true proportions associated with the jury selection, and $p_0 = (0.72, 0.07, 0.12, 0.09)$ denote the proportions in the county. We wish to test

$$H_0 : p = p_0, \quad H_1 : p \neq p_0.$$

Note that under H_0, the expected numbers of jurors under each category are $0.72n$, $0.07n$, $0.12n$, and $0.09n$, so the test statistic is

$$T_n = \frac{(O_1 - 0.72n)^2}{0.72n} + \frac{(O_2 - 0.07n)^2}{0.07n} + \frac{(O_3 - 0.12n)^2}{0.12n} + \frac{(O_4 - 0.09n)^2}{0.09n};$$

using $n = 275$ and our observed values, we get

$$T_{275} = \frac{(205 - 198)^2}{198} + \frac{(26 - 19.25)^2}{19.25} + \frac{(25 - 33)^2}{33} + \frac{(19 - 24.75)^2}{24.75} \approx 5.8896.$$

Since $T_n \to \chi_3^2$ in distribution (there are 4 categories), we get a p-value of around $\boxed{0.1171}$.

Problem 5

A large study consists in testing the ability of four TAs (Suzi, Donita, Jennifer, and Dee) to help students get an A in a statistics class.

Students were assigned at random into 5 groups (one without any TA help and four TA groups). We recorded the number of students in each group who received an A.

Compound	NoTA	Suzi	Donita	Jennifer	Demetra
# students	80	75	85	70	75
# As	55	59	42	60	49

1. For each of the four TAs, test at level 10% if they improve the probability of getting an A?

   Solution: Let $i = 0$ correspond to No TA, $i = 1$ correspond to Suzi, $i = 2$ to Donita, $i = 3$ to Jennifer, and $i = 4$ to Demetra. Note that our statistical model is Bernoulli (as each student either gets an A or does not get an A), with parameters p_i for $0 \leq i \leq 4$. For each $1 \leq i \leq 4$, we wish to test

$$H_0 : p_i \leq p_0 \quad \text{vs.} \quad H_1 : p_i > p_0,$$

   where p_i denotes the true proportion of the students assigned to TA i who get an A. In particular, for each i, we observe a number of students n_i, and an estimated proportion $\hat{p}_i = \frac{(\#\text{ As})}{n_i}$. Using the fact that

$$(\hat{p}_i - p_i) \approx \mathcal{N}(0, p_i(1-p_i)/n_i) \text{ in distribution for all } 0 \leq i \leq 4$$

$$\implies ((\hat{p}_i - \hat{p}_0) - (p_i - p_0)) \approx \mathcal{N}(0, \frac{p_i(1-p_i)}{n_i} + \frac{p_0(1-p_0)}{n_0}) \text{ in distribution for } 1 \leq i \leq 4,$$

   our test statistic would be

$$T^{(i)} = \frac{\hat{p}_i - \hat{p}_0}{\sqrt{(\hat{p}_i(1-\hat{p}_i)/n_i) + (\hat{p}_0(1-\hat{p}_0)/n_0)}},$$

   via the plug-in method and using $p_i - p_0 \leq 0$ under the null hypothesis. The test statistics themselves converge to $\mathcal{N}(0,1)$ in distribution.

   In particular, we get the following test statistic values and p-values (noting that p-values are calculated via $p^{(i)} = 1 - \Phi(T^{(i)})$):

$$T^{(1)} = \frac{\frac{59}{75} - \frac{55}{80}}{\sqrt{\left(\left(\frac{59}{75}\right)\left(\frac{16}{75}\right)/75\right) + \left(\left(\frac{55}{80}\right)\left(\frac{25}{80}\right)/80\right)}} \approx 1.413 \implies p^{(1)} \approx 0.0788;$$

$$T^{(2)} = \frac{\frac{42}{85} - \frac{55}{80}}{\sqrt{\left(\left(\frac{42}{85}\right)\left(\frac{43}{85}\right)/85\right) + \left(\left(\frac{55}{80}\right)\left(\frac{25}{80}\right)/80\right)}} \approx -2.578 \implies p^{(2)} \approx 0.995;$$

$$T^{(3)} = \frac{\frac{60}{70} - \frac{55}{80}}{\sqrt{\left(\left(\frac{60}{70}\right)\left(\frac{10}{70}\right)/70\right) + \left(\left(\frac{55}{80}\right)\left(\frac{25}{80}\right)/80\right)}} \approx 2.547 \implies p^{(3)} \approx 0.0054;$$

$$T^{(4)} = \frac{\frac{49}{75} - \frac{55}{80}}{\sqrt{\left(\left(\frac{49}{75}\right)\left(\frac{26}{75}\right)/75\right) + \left(\left(\frac{55}{80}\right)\left(\frac{25}{80}\right)/80\right)}} \approx -0.452 \implies p^{(4)} \approx 0.6744.$$

In particular, we only reject the tests corresponding to $\boxed{i = 1, 3}$.

2. How do your conclusions change when you use the Bonferroni correction?
Solution: Here, if we use the Bonferroni correction, we change the level of each individual test to $\frac{0.10}{4} = 0.025$. Now the only test with a p-value below this level is the test corresponding to $i = 3$. Our $\boxed{\text{conclusion changes from reject } i = 1, 3 \text{ to reject } i = 3 \text{ only}}$.

3. How do your conclusions change when you use the FDR method?
Solution: Here, if we order the p-values, we get the following:

$$P^{(1)} = 0.0054, P^{(2)} = 0.0788, P^{(3)} = 0.6744, P^{(4)} = 0.995$$

(note that $P^{(i)}$ doesn't necessarily correspond to the p-value of the ith test; this is a relabeling of the p-values).
Note that

$$0.0054 < \frac{1 \cdot 0.10}{4}, 0.0788 > \frac{2 \cdot 0.10}{4}, 0.6744 > \frac{3 \cdot 0.10}{4}, 0.995 > \frac{4 \cdot 0.10}{4},$$

so the highest-indexed p-value such that $P^{(i)} \leq \frac{0.10i}{4}$ is $P^{(1)}$, which corresponds to test 3. So our conclusion is to $\boxed{\text{reject } i = 3 \text{ only}}$. (The numbers are slightly different if the test is run for dependent p-values, but the conclusions still shouldn't change).

Problem 6

The lifetime (in thousands of hours) X of a random lightbulb has pdf

$$g(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

for some unknown $\lambda > 0$.

We collected $n = 33$ independent lightbulbs at random and record their lifetime $X_1, \ldots, X_n$ (independent copies of X). We find that $\bar{X}_n = 42.6$ thousand hours.

1. Consider the prior distribution $\lambda \sim \mathsf{Exp}(1)$. Compute the maximum a posteriori (this is the maximizer of the posterior) estimator of λ.
   Solution: Note that

$$\begin{aligned}\pi(\lambda \mid X_1, X_2, \ldots, X_n) &\propto \pi(\lambda) L(X_1, X_2, \ldots, X_n \mid \lambda) \\ &\propto (e^{-\lambda})(\lambda^n e^{-\lambda \sum X_i}) \\ &\propto \lambda^n e^{-\lambda(1+\sum X_i)}.\end{aligned}$$

   The value of λ that maximizes this is the value that makes the derivative of the logarithm of $\pi(\lambda \mid X_1, X_2, \ldots, X_{33})$ equal to 0; in other words,

$$\frac{n}{\hat{\lambda}} - (1 + \sum X_i) = 0 \implies \boxed{\hat{\lambda} = \frac{1}{\overline{X_n} + \frac{1}{n}} = \frac{165}{7034} \approx 0.02345749}.$$

2. Consider the prior distribution

$$\lambda = \begin{cases} 1 & \text{with probability } .3 \\ 2 & \text{with probability } .7 \end{cases}$$

   Compute the maximum a posteriori estimator of λ.
   Solution: Note that

$$\begin{aligned}\frac{\pi(\lambda = 1 \mid X_1, X_2, \ldots, X_n)}{\pi(\lambda = 2 \mid X_1, X_2, \ldots, X_n)} &= \frac{\pi(\lambda = 1) L(X_1, X_2, \ldots, X_n \mid \lambda = 1)}{\pi(\lambda = 2) L(X_1, X_2, \ldots, X_n \mid \lambda = 2)} \\ &= \frac{0.3 e^{-\sum X_i}}{0.7(2^{33} e^{-2\sum X_i})} > 1\end{aligned}$$

   Thus, the maximum a posteriori is $\boxed{\hat{\lambda} = 1}$.

Problem 7 Evaluating Model Performance with Bootstrapping

You're a data scientist at TechX, a company specializing in machine learning solutions. Your team has recently developed a new classification algorithm, SkyNet, to predict customer opinions based on historical user activity. You've trained SkyNet on a dataset of 1,000 customers, with each customer labeled as either "satisfied" or "unsatisfied." Before deploying SkyNet, you want to obtain a robust estimate of its classification accuracy. Having studied 18.650, you recall the bootstrap method and decide to employ it to assess the model's performance.

1. Explain the bootstrap method in the context of evaluating the classification accuracy of SkyNet.

   The bootstrap method involves randomly sampling from the original dataset with replacement to create many "resampled" datasets. For evaluating the classification accuracy of SkyNet, each bootstrap sample would be used to test the model, and an accuracy score would be recorded for each sample. This process provides an empirical distribution of the accuracy scores, offering insights into the variability and uncertainty of SkyNet's performance.

2. Describe the steps you would take to perform bootstrap resampling to estimate the distribution of classification accuracy for SkyNet. How many resamples would you recommend and why?

   Steps to perform bootstrap resampling:

   1. Draw a random sample of size 1000 from the dataset with replacement.
   2. Use the drawn sample as a test set and predict the labels using SkyNet.
   3. Compute the classification accuracy for this bootstrap sample.
   4. Repeat steps 1-3 for a specified number of bootstrap iterations.
   5. Collect all the accuracy scores to form an empirical distribution of classification accuracy.

   The key point here is step 4. If we just do steps 1-3 without step 4 we won't get a real picture of SkyNet's performance. The number of resamples or bootstrap iterations is a matter of balance between computational cost and the precision of the estimates. A reasonable recommendation is to use at least 100 bootstrap iterations (any number greater than 1 with sufficient explanation would be accepted), but in practice, one might use even more to ensure stable estimates. The goal is to have enough resamples such that the distribution of accuracy scores stabilizes and provides consistent estimates.

Problem 8 Efficacy of a New Drug

PharmaX, a leading pharmaceutical company, has developed a new drug, MedX, intended to reduce high blood pressure. In a randomized controlled trial, 500 patients were administered MedX, while another 500 patients were given a placebo. After six months, 300 patients from the MedX group exhibited normalized blood pressure levels, while 250 patients from the placebo group had normalized blood pressure. PharmaX believes that MedX is more effective than the placebo and wants to validate this statistically. As an 18.650 student, you've been consulted to provide a statistical assessment of the drug's efficacy.

1. Based on the data provided, establish a statistical model for the proportion of patients with normalized blood pressure in both the MedX and placebo groups. Describe the parameters of your model.

   Let p_M represent the proportion of patients with normalized blood pressure in the MedX group and p_P represent the proportion in the placebo group. The sample proportions are given by:

$$\hat{p}_M = \frac{300}{500}$$
$$\hat{p}_P = \frac{250}{500}$$

   We can model the number of patients with normalized blood pressure in each group as binomially distributed, since each patient either has or does not have normalized blood pressure. Thus:

$$X_M \sim Binomial(500, p_M)$$
$$X_P \sim Binomial(500, p_P)$$

   Where X_M and X_P are the number of successes (normalized blood pressure) in the MedX and placebo groups, respectively.

2. Propose a test to assess if MedX has a significantly higher proportion of patients with normalized blood pressure than the placebo. Clearly state the null and alternative hypotheses for this test and come to conclusion justifying the choices that you make along the way.

   The Wald test can be used to test the difference between two proportions. For our scenario: Null Hypothesis, H_0: $p_M = p_P$ (MedX does not have a significantly different effect than the placebo) Alternative Hypothesis, H_a: $p_M > p_P$ (MedX has a higher proportion of patients with normalized blood pressure than the placebo)

Problem 9 Analyzing Streaming Service Playlists

Xpotify, a prominent music streaming service is analyzing user engagement with two different playlist generation algorithms: Algorithm A, which recommends songs based on a user's listening history, and Algorithm B, which recommends songs based on trending patterns and social media activity. They've rolled out Algorithm A to a subset of 1000 users and Algorithm B to another subset of 1000 users. After one month, they measured the average number of songs listened to per user from the generated playlists. Preliminary analyses suggest that the number of songs listened to for both algorithms may be normally distributed. As an 18.650 student, the streaming service has hired you to delve deeper into the investigation.

1. Model the number of songs listened to for both algorithms as random variables and specify their potential distributions based on the preliminary observations. Explain your choices.

Let's denote the number of songs listened to for Algorithm A as S_A and for Algorithm B as S_B. Given the preliminary observations, we can model both S_A and S_B as normally distributed random variables. Specifically, $S_A \sim \mathcal{N}(\mu_A, \sigma_A^2)$ and $S_B \sim \mathcal{N}(\mu_B, \sigma_B^2)$, where μ_A and μ_B are the average number of songs listened to for each algorithm, and σ_A^2 and σ_B^2 are their respective variances. The normal distribution is chosen based on the Central Limit Theorem, which suggests that the average of a large number of independent and identically distributed random variables (like each time a song is being listened to) will be approximately normally distributed. Notice that we model S_A S_B with different parameters for the mean and variance because we have no reason to believe that they will share these values.

2. Propose a statistical test to check if there's a significant difference in the average number of songs listened to between the two algorithms. Clearly define the null and alternative hypotheses for this test.

   To compare the means of two independent samples, we can use the two-sample t-test. We choose this test because when the data are assumed to be normal and the sample size is small, it is common instead to use the (non-asymptotic) t-test. (Other tests will also be accepted if the choice is explained well) The null and alternative hypotheses are:

   $H_0 : \mu_A = \mu_B$

   (There is no difference in the average number of songs listened to between the two algorithms)

   $H_a : \mu_A \neq \mu_B$

   (There is a difference in the average number of songs listened to between the two algorithms)

Hedge fund interview question

In every PSet, we have an additional question taken from a hedge fund interview. This question is not mandatory and does not hold any point but you are welcome to give it a shot.

Problem 10 (Source: Two Sigma)

Suppose a survey was conducted on 100 people to collect information on their phone numbers. Observations found that the number of people whose phone numbers ended in 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9 were 14, 10, 6, 17, 6, 9, 6, 17, 11, and 4, respectively. Would you conclude that the distribution of last digit is uniformly random?

Solution: We can use several different tests, but let's conduct a Pearson's χ^2 test to analyze our hypothesis. Our $p_0 = (0.14, 0.1, 0.06, 0.17, 0.06, 0.09, 0.06, 0.17, 0.11, 0.04)$ and $p = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1)$. We want to test

$$H_0 : p = p_0, \quad H_1 : p \neq p_0.$$

Our test statistic is thus

$$\begin{aligned}\chi^2 &= \frac{(14-10)^2}{10} + \frac{(10-10)^2}{10} + \frac{(6-10)^2}{10} + \frac{(17-10)^2}{10} + \frac{(6-10)^2}{10} \\ &\quad + \frac{(9-10)^2}{10} + \frac{(6-10)^2}{10} + \frac{(17-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(4-10)^2}{10} = 20.\end{aligned}$$

Our p-value is consequently $\mathbb{P}[\chi^2_9 > 20] = 0.01791$, which is strong evidence against H_0.

Chi-square probability table

The table lists the quantiles x such that $P(X \geq x) = \alpha$ where $X \sim \chi^2_k$ for several values of α $(0.3, 0.2, \dots)$.

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
	8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
	9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88
	10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59
	11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26
	12	14.01	15.81	18.55	21.03	24.05	26.22	28.30	32.91
	13	15.12	16.98	19.81	22.36	25.47	27.69	29.82	34.53
	14	16.22	18.15	21.06	23.68	26.87	29.14	31.32	36.12
	15	17.32	19.31	22.31	25.00	28.26	30.58	32.80	37.70
	16	18.42	20.47	23.54	26.30	29.63	32.00	34.27	39.25
	17	19.51	21.61	24.77	27.59	31.00	33.41	35.72	40.79
	18	20.60	22.76	25.99	28.87	32.35	34.81	37.16	42.31
	19	21.69	23.90	27.20	30.14	33.69	36.19	38.58	43.82
	20	22.77	25.04	28.41	31.41	35.02	37.57	40.00	45.31
	25	28.17	30.68	34.38	37.65	41.57	44.31	46.93	52.62
	30	33.53	36.25	40.26	43.77	47.96	50.89	53.67	59.70
	40	44.16	47.27	51.81	55.76	60.44	63.69	66.77	73.40
	50	54.72	58.16	63.17	67.50	72.61	76.15	79.49	86.66

	Second decimal place of Z									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

The table lists $P(Z \leq z)$ where $Z \sim N(0,1)$ for positive values of z.