Homework 7

Due Mar 26 at 11 pm

Unless stated otherwise, justify any answers you give. You can work in groups, but each student must write their own solution based on their own understanding of the problem.

When uploading your homework to Gradescope you will have to select the relevant pages for each question. Please submit each problem on a separate page (i.e., 1a and 1b can be on the same page but 1 and 2 must be on different pages). We understand that this may be cumbersome but this is the best way for the grading team to grade your homework assignments and provide feedback in a timely manner. Failure to adhere to these guidelines may result in a loss of points. Note that it may take some time to select the pages for your submission. Please plan accordingly. We suggest uploading your assignment at least 30 minutes before the deadline so you will have ample time to select the correct pages for your submission. If you are using LaTeX, consider using the minted or listings packages for typesetting code.

1. (Sign test) The sign test is a nonparametric two-sample test. Given n pairs of data (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n), the null hypothesis is that the first and second entries of each pair belong to the same distribution. The test statistic is

$$t = \sum_{i=1}^{n} 1\left(y_i > x_i\right), \tag{1}$$

where $1\left(y_i > x_i\right)$ is an indicator function that is equal to 1 if $y_i > x_i$ and 0 otherwise.

   (a) If $\tilde{a}$ and $\tilde{b}$ are independent continuous random variables with the same distribution, what is the probability that $\tilde{a} > \tilde{b}$?

   - $P(\tilde{a} \geq \tilde{b}) = \int_{b\in\mathbb{R}} P(\tilde{a} \in [b, \infty], \tilde{b} \in [-\infty, b])db = \int_{b\in\mathbb{R}} P(\tilde{a} \in [b, \infty])P(\tilde{b} \in [-\infty, b])db = \int_{b\in\mathbb{R}}(1 - F_{\tilde{a}}(b))F_{\tilde{b}}(b)db = \int_{b\in\mathbb{R}}(1 - F_{\tilde{b}}(b))F_{\tilde{b}}(b)db = \frac{1}{2}$

   (b) Derive the distribution of the test statistic of the sign test under the null hypothesis that the data are independent samples from the same continuous distribution.

   - consider a random varible $\tilde{t_i} = 1(\tilde{y_i} \geq \tilde{x_i})$ we can think of $\tilde{t_i}$ as a bernoulli random variable with parameter θ
   - thus we can think of our test stat under the null as the sum of independent identically distributed bernoulli random variables and thus distributed as a binomial with parameters n, θ_{null}
   - where θ_{null} represents the $P(x_i \geq y_i)$, which under the null being $\tilde{x}, \tilde{y}$ are distributed the same, we would have $\theta_{null} = 0.5$
   - thus we know that the test stat under the null is distributed as a binomial with parameter n, .5 that is $t \sim b(n, 0.5)$

   (c) Your friend is convinced that in general the left ear of most people is longer than the right ear. You measure the ears of some of your other friends and obtain the following data (in inches).

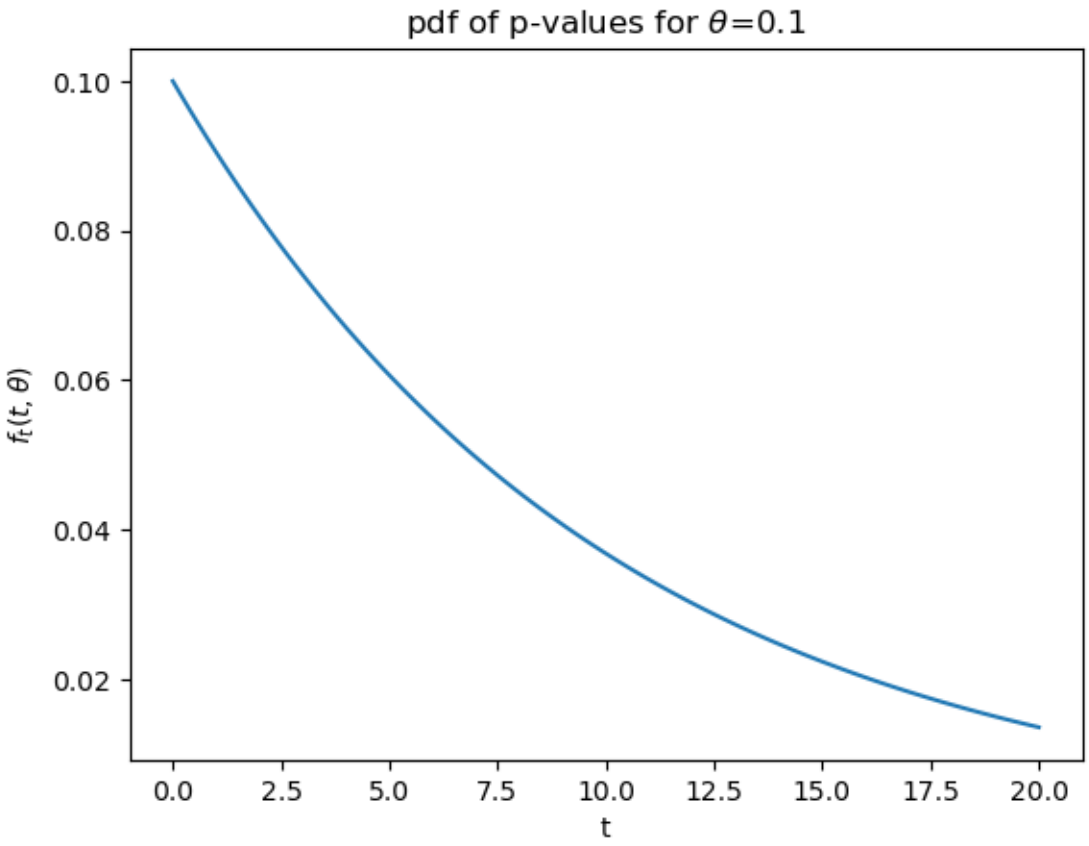
Left	2.4	2.7	3.2	2.3	2.0	2.6	3.2	2.3	2.9	2.3
Right	2.2	2.6	3.3	2.2	2.1	2.5	3.1	2.5	2.7	2.2

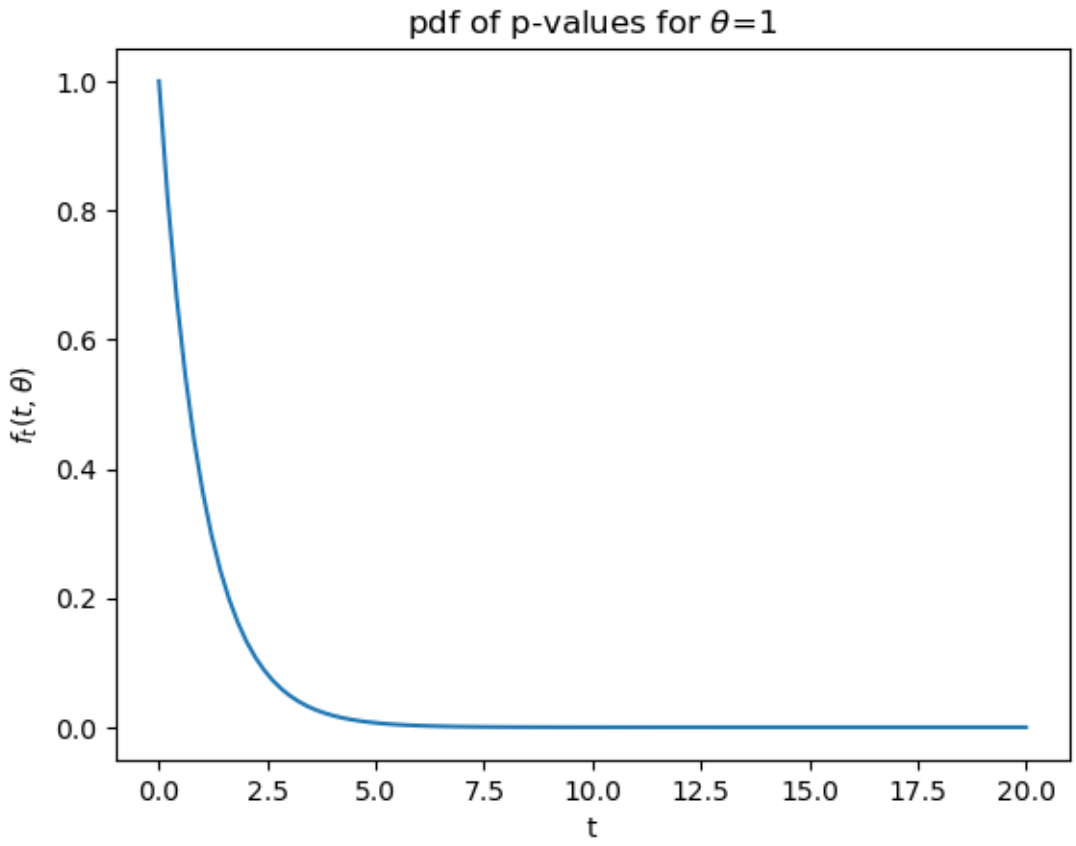
Apply the sign test to compute a p value associated to the null hypothesis that the left and right ear have the same length.

- we know our p-value is $P(t \leq \tilde{t}_{null}) = (1 - F_{\tilde{t}_{null}}(t)) = \Sigma_{i=t}^{10} \begin{pmatrix} 10 \\ i \end{pmatrix} (0.5)^i (0.5)^{10-i} = \Sigma_{i=t}^{10} \begin{pmatrix} 10 \\ i \end{pmatrix} (0.5)^{10}$
- we can see in our data $n = 10$ as there are 10 samples and $k = 7$ as in 7 cases the left ear length is greater than the right ear length
- ```
  def chose(n,k):
      factorial = lambda n: np.product([i for i in range(1, n+1)])
      return factorial(n) / (factorial(k)*factorial(n-k))
  def binomial(theta,k):
      return chose(10,k) * ((theta) ** (k)) * ((theta) ** (10-k))
  def p_val(t):
      v_binomial = np.vectorize(binomial)
      return np.sum(v_binomial(.5, np.array(range(t,11))))
  p_val(7)
  ```
- so computing this out we see that our p value ends up being 0.171875

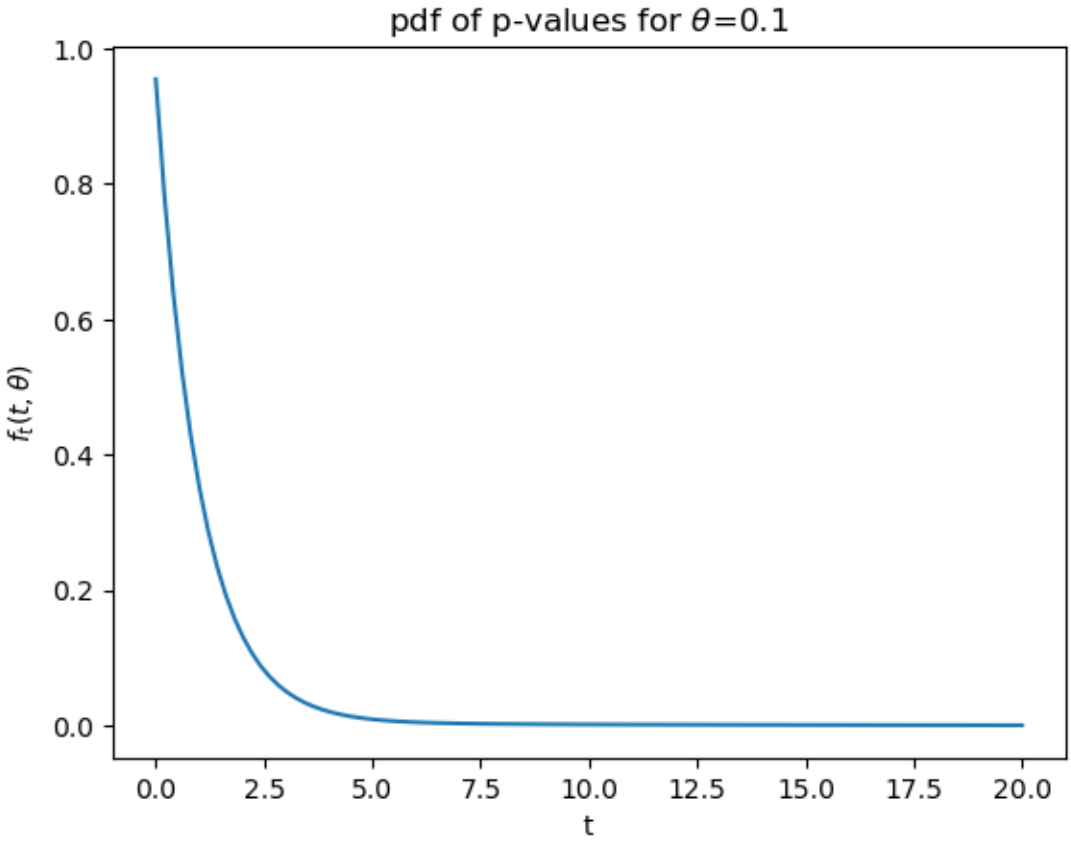
(d) Is the sign test still valid for the null hypothesis that each pair of data (x_i, y_i) are sampled from the same continuous distribution, but these distributions are different for different values of i?

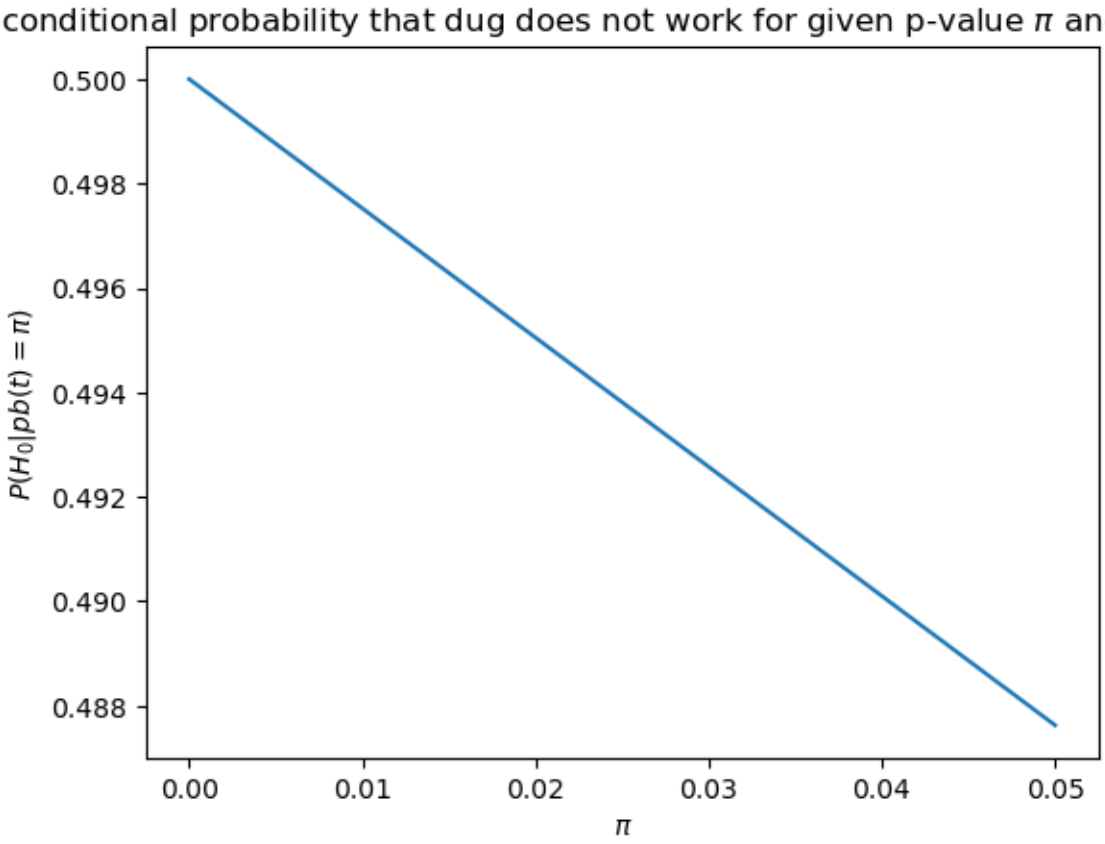
- yes it should still hold
- at each sample we are finding $P(y_i \geq x_i)$
- and regardless of how $\tilde{y}_i, \tilde{x}_i$ are distributed under the null hypothesis we assume they have the same distribution thus, $P(y_i \geq x_i) = 0.5$
- the above implies that regardless of the actual distribution $\tilde{y}_i, \tilde{x}_i$ we know $1(\tilde{x}_i \geq \tilde{y}_i) = \tilde{t}_i \sim$ bernoulli(0.5)
- which implies $\Sigma_{1=1}^{n} 1(\tilde{y}_i \geq \tilde{x}_i) = \Sigma_{1=1}^{n} \tilde{t}_i = \tilde{t} \sim b(n, 0.5)$
- and thus the sign test would still hold.
  ```

2. (Drug design) A company wants to design a drug to extend the life of patients with a certain disease. From past data, the survival time of the patients is known to be exponential with parameter 1. The company has a large number of promising drug candidates. When a candidate is effective, it modifies the survival time to be exponential with parameter θ. Otherwise it does not affect the survival time at all. In order to evaluate the candidates, each one is given to a different patient. A hypothesis test is then applied where the test statistic is the survival time of the patient. The null hypothesis is that the drug does not work and therefore the survival time is exponential with parameter 1.

   (a) Derive the pdf of the p value when the drug is effective as a function of θ. Plot it for θ equal to 1 and to 0.1.

   - call H_0 the null hypothesis and H_1 the alternative hypothesis
   - given we know the alternative hypothesis is true we can write the p value as $pv(t) = P(\tilde{t} \geq t|H_0) = P(\tilde{t} \geq t|H_1) = e^{-\theta t} \sim e(\theta)$
   - so our pdf is $\theta e^{-\theta t}$
   - ```
     def pdf(theta,x):
         return theta*np.exp(-theta*x)
     theta=.1
     x=np.linspace(0,20,100)
     y=pdf(theta, x)
     plt.plot(x,y)
     plt.xlabel("t")
     plt.ylabel("$f_{t}(t, \\theta)$")
     plt.title("pdf of p-values for $\\theta$={0}".format(theta))
     plt.show()
     theta=1
     x=np.linspace(0,20,100)
     y=pdf(theta, x)
     plt.plot(x,y)
     plt.xlabel("t")
     plt.ylabel("$f_{t}(t, \\theta)$")
     plt.title("pdf of p-values for $\\theta$={0}".format(theta))
     ```
     ```


pdf of p-values for θ=0.1
0.10
0.08
0.06
0.04
0.02
$f_t(t, \theta)$
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
t



(b) Derive the pdf of the p value if the probability that the drug is effective is 1/20 (i.e. only 5% of the candidates are effective) as a function of θ. Plot the pdf for $\theta := 0.1$.

- by the law of total probability $pv(t) = P(\tilde{t}_{null} \geq t) = P(\tilde{t}_{null} \geq t|H_0)P(H_0) + P(\tilde{t}_{null} \geq t|H_1)P(H_1)$
- then using what we derived in last question and the information in this question we can write $pv(t) = P(\tilde{t}_{null} \geq t) = \frac{19}{20}(1 - F_{\tilde{t}_{null}}(t)) + \frac{1}{20}(1 - F_{\tilde{t}_\theta}(t)) = \frac{19}{20}(e^{-t}) + \frac{1}{20}(e^{-\theta t})$
- then as we know difertiation is linear the pdf is just the sum of there darivatives with respect to θ that is $f_{pv(t)} = e^{-t} + \theta e^{-\theta t}$
- ```
  def pdf(theta,x):
      return (theta*np.exp(-theta*x)/20)+(19*np.exp(-x)/20)
  ```
  ```



-

(c) If we reject the null hypothesis, what is the conditional probability that there is a false positive and the drug candidate doesn't actually work? Derive the conditional probability as a function of the significance level α and of θ, and report it for $\alpha = 0.05$ and $\theta := 0.1$.

- first we need to solve for t_{thresh} under the null hypothiss such that $t_{thresh} := argmin_t\{t : pv(t) \leq \alpha | H_0\} = argmin_t\{t : P(\tilde{t} \geq t) \leq \alpha | H_0\} = argmin_t\{t : 1 - F_{\tilde{t}_{null}}(t) \leq \alpha\} = argmin_t\{t : e^{-(t)} \leq \alpha\} = -ln(\alpha)$
- so if we reject the null we know that $\tilde{t} \leq t_{thresh}$
- given this is the case our likelihood of a false positive is thus $P(H_0|\tilde{t} \leq t_{thresh})$ that is the likelihood the null is true given we reject the null
- we can write this with bayes therome as $P(H_0|\tilde{t} \leq t_{thresh}) = \frac{P(H_0, \tilde{t} \leq t_{thresh})}{P(\tilde{t} \leq t_{thresh})}$
- so first we know $P(H_0, \tilde{t} \leq t_{thresh})$ is the case where the null is true but we reject the null that is we have a false postive is $P(H_0)\alpha$
- now consider $P(\tilde{t} \leq t_{thresh}) = P(\tilde{t} \leq t_{thresh}|H_0)P(H_0) + P(\tilde{t} \leq t_{thresh}|H_1)P(H_1) = \alpha(P(H_0)) + (1 - F_\theta(t_{threshold}))(1 - P(H_0))$
- so this equals $\frac{P(H_0)\alpha}{\alpha P(H_0) + (1 - F_{\theta_t}(t_{threshold}))(1 - P(H_0))} = \frac{P(h_0)\alpha}{P(h_0)\alpha + P(h_a)(e^{-\theta ln(\alpha)})}$
- i am going to assume that we are supposed to use $P(H_1) = \frac{1}{2}$ since there is no other indication of what value to use
- `p=1/2`
  `alpha=.05`

```
theta=.01
def condtional(p, theta , alpha):
    numerator=p*alpha
    denominator= numerator +(1-p)*(np.exp(-theta*np.log(alpha)))
    return numerator/denominator
condtional(p, theta , alpha)
```

- doing this out i got 0.04627870362231976 as the liklyhood of the null being true given we rejet the null

- that is encouragaing since the value is around 5% which is where we expect it to be.

(d) Derive the conditional probability that the drug candidate doesn't work given that the p value equals π, as a function of π and θ. Plot the function for $0 \leq \pi \leq 0.05$ and $\theta := 0.1$. Based on your result, suggest a strategy to select candidates sequentially for follow-up testing.

- so we are looking for $P(H_0|pv(t) = \pi) = \frac{P(Pv(t)=\pi|H_0)P(H_0)}{P(Pv(t)=\pi|H_0)P(H_0)+P(Pv(t)=\pi|H_1P(H_1)} = \frac{e^{-\pi}P(H_0)}{e^{-\pi t}P(H_0)+e^{-\theta\pi}P(H_1)}$

- again we are not explicitly given a value for the likelihood of each hypothesis i am going to assume $P(H_0) = P(H_1)$ since i guess that makes the least assumptions?

- ```
  p=1/2
  alpha=.05
  theta=.01
  pi=.01
  def condtional(p, theta , alpha,pi):
      numerator=p*np.exp(-pi )
      denominator= numerator +(1-p)*(np.exp(-theta*pi))
      return numerator/denominator
  x=np.linspace(0,.05,100)
  plt.plot(x,condtional(p, theta , alpha,x))
  plt.xlabel("$\\pi$")
  plt.ylabel("$P(H_0|pb(t)=\\pi)$")
  plt.title("conditional probability that dug does not work for
  ↪  given p-value $\\pi$ and $\\theta=0.1$")
  ```
  ```

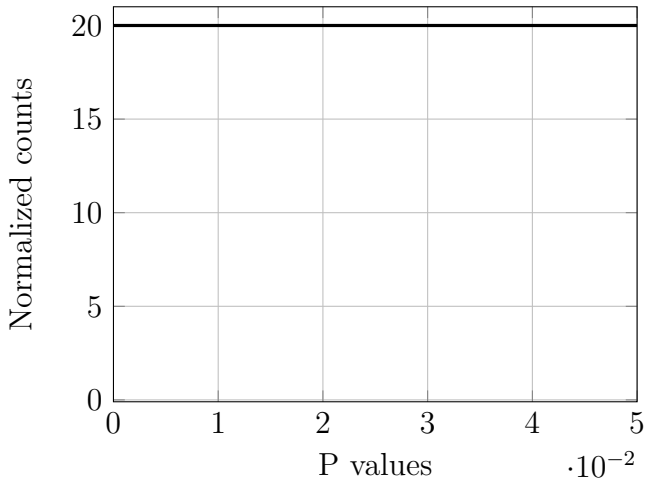

conditional probability that dug does not work for given p-value π and $\theta = 0$
$P(H_0|pb(t) = \pi)$
0.500
0.498
0.496
0.494
0.492
0.490
0.488
0.00
0.01
0.02
0.03
0.04
0.05
π

3. (P-hacking)

(a) If $\tilde{u}$ is a uniform random variable between 0 and 1, derive the conditional pdf of $\tilde{u}$ conditioned on the event $\tilde{u} \leq \alpha$ for $0 < \alpha < 1$.

- consider an arbitrary value k, $P(\tilde{u} \leq k | \tilde{u} \leq \alpha)$
- we can see that $1 = \int_{k \in \mathbb{R}} f_{\tilde{u} \leq \alpha}(k)dk = \int_{k \in -\infty}^{\alpha} f_{\tilde{u} \leq \alpha}(k) + \int_{0}^{\alpha} f_{\tilde{u} \leq \alpha}(k) + \int_{\alpha}^{\infty} f_{\tilde{u} \leq \alpha}(k)dk = \int_{0}^{\alpha} f_{\tilde{u} \leq \alpha}(k)dk = (1 - \alpha) f_{\tilde{u} \leq \alpha} = 1 \Rightarrow f_{\tilde{u} \leq \alpha} = \frac{1}{\alpha}$

(b) The histogram of the p values in the publications of a research group looks like this:



What does this suggest?

- this suggest the group is not p-hacking as there is a uniform distribution over the interval as we would expect to see.

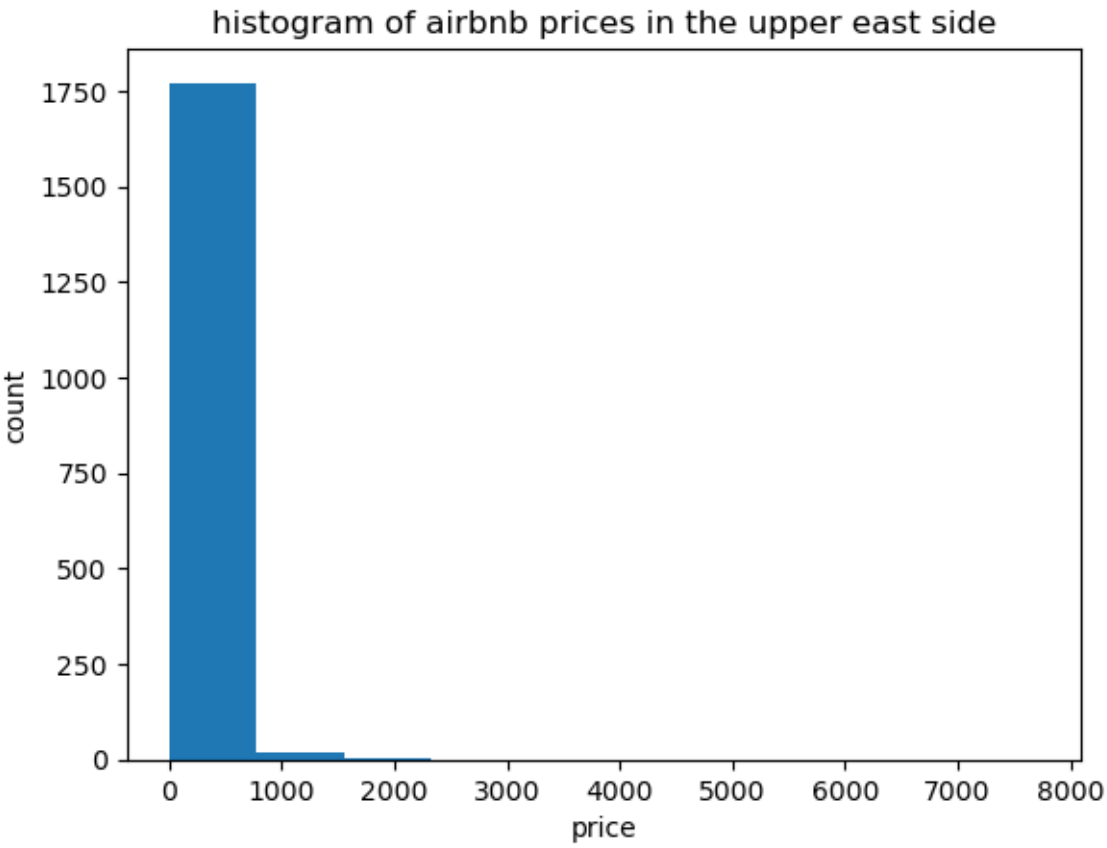
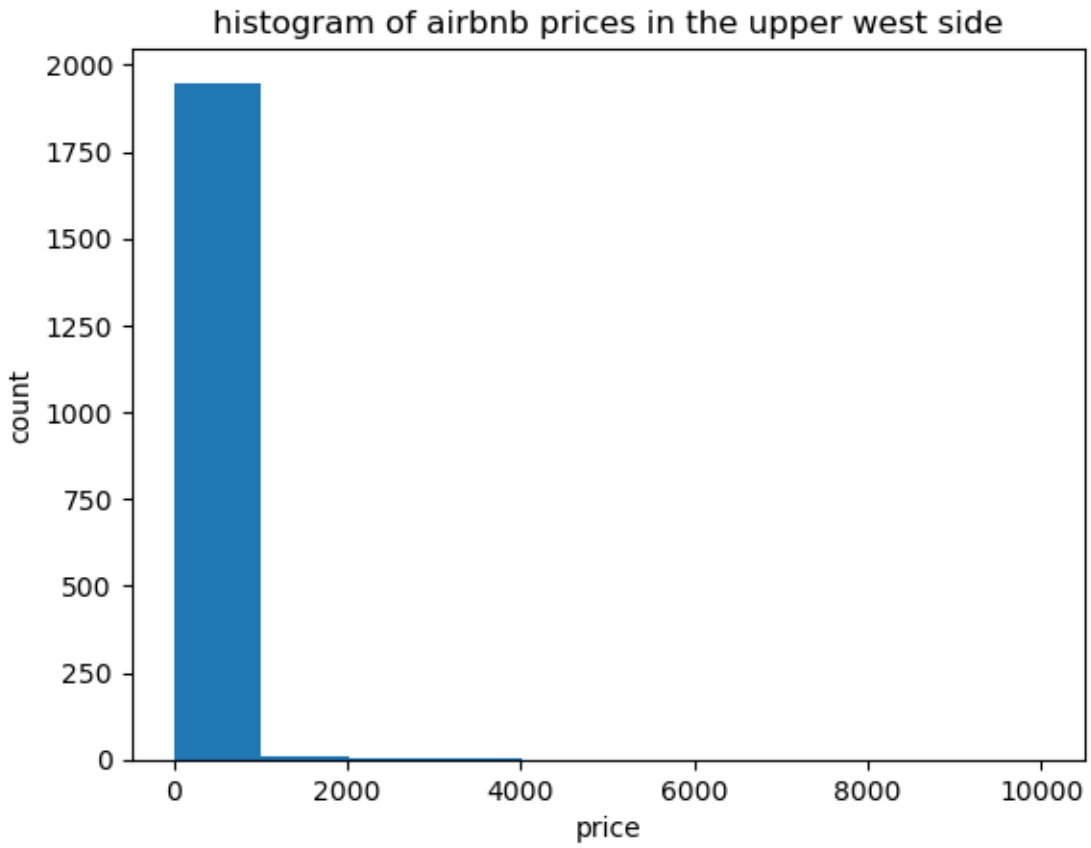
(c) If the total number of p values is 100, estimate the number of results that the research group has not published.

- assuming a uniform distribution of p values we would expect for there to be k total studies where k is $k : \frac{5}{100} * k = 100 \Rightarrow k = 2,000$
- thus the number of unpublished studies would be 1,900

4. (Airbnb Pricing) Airbnb prices oftern vary by neighborhood. In this problem, our conjecture is that prices in the Upper East Side and Upper West Side are different. Our goal is to evaluate this conjecture using hypothesis testing. The null hypothesis is that the price distribution is the same. Based on the *price* and *neighbourhood* columns in the open Airbnb dataset in `AB_NYC_2019.csv`, we conduct parametric and non-parametric hypothesis testing.

   (a) Assuming the price in the two neighbourhoods are two Gaussian random variables with known population standard deviation 240.15, apply a parametric two-sample test. Select a test statistic and compute the p-value.

   - let people in the upper east side be group a and poeple in the upper west side be group b
   - and let x_i be a random varible representing the price of data point i's Airbnb
   - let the test stat be $t = |\frac{1}{n_a}\Sigma_{i\in a}x_i - \frac{1}{n_b}\Sigma_{i\in b}x_i|$
   - i am doing a two sample test, as in our conjecture we do not specficy that one neighborhood is more expensive than the other
   - ```
     def test_stat(a,b):
         return np.abs(np.mean(a)-np.mean(b))
     observed_test_stat=test_stat(group_a, group_b)
     gaussian=stats.norm(loc=0,scale=240.15)
     two_sided_pval=2*(1-gaussian.cdf(observed_test_stat))
     two_sided_pval
     ```
   - doing this we get a p value of 0.0522

   (b) Plot the histograms of prices in the two neighbourhoods. What do you think might be problematic about the test statistic you selected for the parametric test?
     ```


histogram of airbnb prices in the upper east side
count
price
0
250
500
750
1000
1250
1500
1750
0
1000
2000
3000
4000
5000
6000
7000
8000



-
- as can be seen from this graph both are clearly not gaussian, and infact are heavily sckewed by there outliers

(c) Select another test statistic that is more robust to outliers. Apply a permutation test and compute the corresponding p-value.

- let people in the upper east side be group a and poeple in the upper west side be group b
- and let x_i be a random varible representing the price of data point i's Airbnb
- let the test stat be $t = |\text{median(group a)} - \text{median(group b)}|$
- ```
  group_a=df[df["neighbourhood"]=="Upper East Side"]["price"]
  group_b=df[df["neighbourhood"]=="Upper West Side"]["price"]
  labelless_group=np.concatenate((group_a,group_b))
  def test_stat(a,b):
      return np.abs(np.median(a)-np.median(b))
  def monte_carlo_permutation(labelless_group,n=1000):
      rng = np.random.default_rng()
      test_stats=[]
      for i in range(n):
          rng.shuffle(labelless_group)
          a=labelless_group[:len(group_a)]
          b=labelless_group[len(group_a):]
          test_stats.append(test_stat(a,b))
  ```
  ```

```
    return test_stats
n=1000
mc_test_stats=monte_carlo_permutation(labelless_group,n)
test_stat_observed=test_stat(group_a,group_b)
np.sum(mc_test_stats<=test_stat_observed)/1000
```

- running this for 1000 monte carlo simultions of permutations in the multiset i found a p-value of 0.926