Homework 6

Due Mar 19 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. (Road renovation) A small town decides to renovate a 10-mile road in a small town. According to certain reports, there is a specific 2.5-mile section which is very dangerous. Completely renovating that section would use up all of the budget, so the engineer in charge of the renovation wants to make sure that this is a priority. She decides to perform a hypothesis test using the next 4 accidents reported on the road. Her null hypothesis is that the accidents are independent and uniformly distributed on the 10-mile road, which would imply that the *dangerous* section is not that dangerous. The test statistic is the number of accidents that occur in the dangerous 2.5-mile section. The significance level is set to $\alpha := 0.05$.

   (a) Derive the p value function of the test for all possible values of the test statistic.

   - - the p value is the likelihood of observation a test stat that is larger than or equal to what we saw under the null hypothesis
   - call $\tilde{p}_i$ a Bernoulli rv representing if the ith accident occurred within the 2.5 mile section
   - under our null we think that $\tilde{p}_i \sim \mathcal{U}(0, 10)$
   - our test stat $\tilde{t} = \tilde{p}_1 + \tilde{p}_2 + \tilde{p}_3 + \tilde{p}_4$
   - under our null we can think of the likelihood of each car crashing within the zone as independent, identically distributed Bernoulli's thus we can think of there sum ($\tilde{t}$) as a distribute according to a binomial with parameters n=4 and θ, where θ is the likelihood of crashing in the 2.5 mile danger zone
   - we know that this implies that our p-value function $pv(t) = P(\tilde{t}_{\theta_{null}} \geq t) = \Sigma_{i=t}^{n} \binom{n}{i} (\theta_{null})^i (1 - \theta_{null})^{n-i}$
   - we know that the stretch of rode we are interested in is 2.5 miles out of a total of 10 miles. thus under the null that a crash is equally likely to happen anywhere the likelihood of crashing in that zone is $\theta_{null} = \frac{2.5}{10} = \frac{1}{4}$

- so plugging in we finally get

$$pv(t) = P(\tilde{t}_{\theta_{null}} \geq t) = \Sigma_{i=t}^{n} \binom{n}{i} (\theta_{null})^i (1-\theta_{null})^{n-i} = \Sigma_{i=t}^{4} \binom{4}{i} (\frac{1}{4})^i (\frac{3}{4})^{4-i}$$

(b) What is the probability of a false positive?

- the likelihood of a false positive as we showed in class is equal to αso in this case it is 5%

(c) Let θ denote the probability that an accident occurs in the *dangerous* section. Plot the power function of the test as a function of θ under the assumption that accidents occur independently.

- the power function is a function that maps any given effect size θ to our likely hood of rejecting the null hypothesis. ie

$$pow(\theta) = P(pv(t_\theta) \leq \alpha)$$

- so first things first we need to find our $\tau_{threshold}$ which is the minimum test stat value at which we will reject the null hypothesis for a given α that is

$$\tau_{threshold} = min_{1 \leq \tau \leq n}\{\tau : P(\tilde{t} \geq \tau) \leq \alpha\} = min_{1 \leq \tau \leq n}\{\tau : \Sigma_{i=t}^{n} \binom{n}{i} (\theta)^i (1-\theta)^{n-i} \leq \alpha\}$$

$$= min_{1 \leq \tau \leq n}\{\tau : \Sigma_{i=t}^{4} \binom{4}{i} (\frac{1}{4})^i (\frac{3}{4})^{4-i} \leq \alpha\}$$

- for a fixed value of $n = 4$ and $\alpha = .05$ we can compute $pv(3) = 0.050781253$, but a bit above so i am going with 4.
- so we can write our power function as for any θ as

$$pow(\theta) = P(pv(t_\theta) \leq \alpha) = P(t_\theta \leq \tau_{threshold}) = \Sigma_{i=\tau_{threshold}}^{4} \binom{4}{i} (\theta)^i (1-\theta)^{4-i}$$

$$= \binom{4}{4} (\theta)^4 (1-\theta)^0 = (\theta)^4$$

- above is a graph of the power function

(d) What is the minimum value of θ for which the probability of a true positive is at least 50%?

- recall that as we showed above $pow(\theta) = \theta^4$ thus we can find $\theta_{.5}$ such that $pow(\theta_{.5}) = .5 = (\theta_{.5})^4$ meaning that $\theta_{.05} = \sqrt[4]{0.5}$

2. (Computer component) A computer manufacturer wants to make sure that a certain component will last on average more than a year. They decide to apply a hypothesis test, where the null hypothesis is that the average duration is less than a year. The data correspond to n instances of the component, which can be assumed to be independent. The test statistic is the minimum duration of the n instances. If the time until the component fails is modeled using an exponential distribution, what is the power function of the test as a function of the exponential parameter λ and the significance level α? What is the power at $\lambda = 1$ and what is its limit as $\lambda \to 0$?

- call $\tilde{c}_i$ the time until the ith component fails. we know $\tilde{c}_i \sim exp(\lambda)$
- further we know that the test stat is the minimum fail time of the n computers. thus $\tilde{t_\lambda} = min(\tilde{c}_1 ... \tilde{c}_n)$
- so we can see that $P(\tilde{t_\lambda} \geq t) = P(\tilde{c}_1, ... \tilde{c}_n \geq n)$ further given that the components are indecent this can be expressed as $P(\tilde{t_\lambda} \geq t) = P(\tilde{c}_1, ... \tilde{c}_n \geq t) = P(\tilde{c}_1 \geq t) * .. * P(\tilde{c}_n \leq t) = e^{-\lambda t n}$
- thus we can see that $\tilde{t_\lambda} \sim exp(\lambda n)$
- so if we think of t in terms of continuous quantities of days we have λ represents the mean number of times a component breaks down within an interval. so if we think of the interval in years our null hypothesis is that $\lambda \in \lambda_{null} = (0, 1)$
- so we first must look at our p-value function $sup_{\lambda \in (0,1)} pv(t) = sup_{\lambda \in (0,1)} P(\tilde{t_\lambda} \geq t) = sup_{\lambda \in (0,1)} 1 - F_{\tilde{t_\lambda}}(t) = sup_{\lambda \in (0,1)} 1 - 1 + e^{-\lambda n t} = sup_{\lambda \in (0,1)} e^{-\lambda n t}$
- that function is decreasing in λ so we would pick λ as small as possible
- now we need to find our $\tau_{thresohold}$
- and further with out making any specific assumptions about λ we can solve for our $\tau_{threshold}$ as $\tau_{threshold} = min_{0 \leq \tau} \{\tau : P(\tilde{t}_\theta \geq \tau) \leq \alpha\} = min_{0 \leq \tau} \{\tau : 1 - F_{\tilde{t}_\theta}(t) \leq \alpha\} = min_{0 \leq \tau} \{\tau : e^{-\lambda t} \leq \alpha\} = min_{0 \leq \tau} \{\tau : -\lambda t \leq log(\alpha)\} = min_{0 \leq \tau} \{\tau : t \geq \frac{-log(\alpha)}{\lambda}\}$
- thus we can express our power function as a function of λ, α as $pow(\lambda, \alpha) = P(pv(t_\lambda) \leq \alpha) = P(t_\lambda \geq \tau_{threshold}(\alpha, \lambda)) = 1 - F_{\tilde{t}_\lambda(\tau_{threshold}}(\lambda, \alpha)) = e^{-\lambda \tau_{threshold}(\alpha, \lambda)} = e^{-\lambda * min_{0 \leq \tau} \{\tau : t \geq \frac{-log(\alpha)}{\lambda}\}}$
- at $\lambda = 1$ this becomes $pow(\lambda, \alpha) = P(pv(t_\lambda) \leq \alpha) = P(t_\lambda \geq \tau_{threshold}(\alpha, \lambda)) = 1 - F_{\tilde{t}_\lambda(\tau_{threshold}}(\lambda, \alpha)) = e^{-\lambda \tau_{threshold}(\alpha, \lambda)} = e^{* min_{0 \leq \tau} \{\tau : t(\alpha)\}} = e^{log(\alpha)}$
- which looks like this. but that does not work as a proability it is lower bounded by 1.

3. (Tom Brady and hurricanes) The table shows in what years between 2001 and 2020 Tom Brady won the Super Bowl (top row) and there was at least one Category 5 hurricane in the North Atlantic Ocean (bottom row).

Year	02	03	04	05	06	07	08	09	10	11
Brady wins										
Hurricane										

Year	12	13	14	15	16	17	18	19	20	21
Brady wins										
Hurricane										

(a) Compute the p value of a one-tailed two-sample z test, where the null hypothesis is that hurricanes have the same distribution when Brady wins and when he doesn't.

- we were not explicitly given a test stat but lets call $\tilde{y}_i$ a Bernoulli variable representing if there was hurricane that year and have our T stat be the difrence in average number of huricanes over the years tome brady wins and did not win the super bowl that is $\tilde{t} = \frac{1}{n_A}\Sigma_{i\in A}\tilde{y}_i - \frac{1}{n_B}\Sigma_{i\in B}\tilde{y}_i$
- we estimate θ_{null} as $\theta_{null} = \frac{\text{number of times there is a hurricane}}{\text{number of years}} = \frac{8}{20} = \frac{2}{5}$
- then we can estimate the $\sigma_{null} = \sqrt{\theta_{null}(1-\theta_{null})(\frac{1}{n_A}+\frac{1}{n_B})}$ where n_A is the number of times Brady won a super bowl (7) and n_B is the number of times Brady did not win a super brawl (13)thus $\sigma_{null} = \sqrt{\theta_{null}(1-\theta_{null})(\frac{1}{n_A}+\frac{1}{n_B})} = \sqrt{\frac{2}{5}\frac{3}{5}(\frac{1}{20}} = 0.1095$
- so our $\tilde{t}_{tail-1}.109\tilde{z}$
- thus we can find $Pv(t) = P(\tilde{t}_{1-tail} \geq t) = P(\tilde{z} \geq \frac{t}{0.109})$
- so we can find the test stat of our data $t_{data} = \frac{4}{7} - \frac{4}{13} \approx .263$
- so thus we can see that $pv(t_{data}) = P(\tilde{t} \geq t_{data}) = P(\tilde{z} \geq \frac{.263}{.109}) = .008$

(b) If the p value had been extremely small, would this be convincing evidence that hurricanes occur more often when Brady wins? Justify your answer.

- no this would not convince me of that conclusion as there is just not that much data avalible, and tehse two events are clearly unrelated

4. (Disease prevalence) Doctors would like to research the prevalence of a disease of a certain population. They collected some patients records that can be interpreted as samples from that population. The table in `ehr.csv` records the diagnosis (Dx). Specifically, they choose the two null hypotheses below. For each null hypothesis, choose a hypothesis test and a test statistic and compute the corresponding p-value.

   (a) The prevalence of this disease is greater than 0.3.

   - they are interested in the prevalence of the disease.
   - call θ the prevalence of the disease
   - the null is in this case is $\theta \in \Theta_{null} = (.3, 1]$
   - so let $\tilde{x}_i$ be a Bernoulli rv with parameter θ which takes value one if a person has the disease, which we assume to be iid.
   - so lets have our test stat under so θ be $\tilde{t}_\theta = \frac{1}{n}\Sigma_{i=1}^{n}\tilde{x}_i$
   - this is a sum of iid random variables and we know we have 7212 samples in our data thus it is reasonable to assume our data is approximately Gaussian that is $\tilde{t}_\theta \sim \mathcal{N}(\theta_{null}, \frac{\theta_{null}(1-\theta_{null}}{n}))$ for the sake of notation i am going to write that as $\mu_{null}, \sigma null$
   - so thus our p value function is $pv(t) = sup_{\theta\in\Theta_{null}} P(\tilde{t}_\theta \geq t) = sup_{\theta\in\Theta_{null}} P(\tilde{t}_\theta \geq t) = P(\tilde{t}_{.3} \geq t) = \int_{i=t}^{\infty} f_{\tilde{t}_{.3}}(t)dt$
   - we can find the mean prevalence in the population as $t_{data} = 0.3068071537501733$
   - and find that $pv(t_{data}) = 0.10355045769142213$

   (b) Men and women have the same prevalence.

   - call θ the prevalence of the disease $solet_i$ be a Bernoulli rv with parameter θ which takes value one if a person has the disease, which we assume to be iid.
   - so let us break our data in two groups. group A consisting of the men in the sample, and group b consisting of the women in the sample
   - so lets do a two sided test and have our test stat under θ be $\tilde{t}_\theta = |\frac{1}{An}\Sigma_{i\in A}\tilde{x}_i - \frac{1}{An}\Sigma_{i\in A}\tilde{x}_i|$
   - this is a sum of scaled iid random variables and we know we have 7212 samples in our data thus it is reasonable to assume our data is approximately Gaussian that is $\tilde{t}_\theta \sim \mathcal{N}(0, \theta_{null}(1-\theta_{null})(\frac{1}{n_A} + \frac{1}{n_B}))$
   - from the data we can find $\theta_{null} = \frac{\text{number of people with the disease}}{n} = 0.3068071537501733$
   - in our data the two sided test stat was $t_{data} = 0.23929310619328134$
   - and so our p-value is $pv(t_{data}) = P(\tilde{t}_{null} \leq t_{data}) = \int_{t=t_{data}}^{\infty} f_{\tilde{t}_{\theta_{null}}}(t)dt \approx 0$