MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.s077 Spring 2018

Problem Set 4 due Thursday 3/15, in class

Problem 1. Different tests may be not that different.

1. We recall that the G-test is a good testbed for testing whether the data has a uniform distribution or not, due to the fact that, under the scenario involving a uniform distribution; the KL-divergence takes a particularly simple form. Coming back to the problem, we first express the G-statistics, via,

$$G = 2n\sum_{i=0}^{1}\hat{P}(i)\log\frac{\hat{P}(i)}{P_o(i)} = 20d\left(p||\frac{1}{2}\right);$$

where, $d(p||q) = -p\log(p/q) - (1-p)\log((1-p)/(1-q))$ is the binary divergence (can also be perceived as a divergence, between two Bernoulli distributions); and, p above is simply, $\frac{1}{n}\sum_{i=1}^{n} X_i$; in other words, the (empirical) fraction of ones. Hence, we are demanding a threshold α, such that,

$$\mathbb{P}\left(20d\left(p\middle|\middle|\frac{1}{2}\right) > \alpha\right) = 0.109375.$$

Next, we rearrange the item above, and note that, the probability above is equivalent to,

$$\mathbb{P}\left(h(p) < 1 - \frac{\alpha}{20}\right) = 0.109375,$$

where, $h(p) = p\log p + (1-p)\log(1-p)$ is the binary entropy function. We next note that, for a Binomial distribution $X \sim \mathrm{Bin}(10, 1/2)$;

$$\mathbb{P}(X > 7) = \mathbb{P}(X < 3) = \frac{0.109375}{2}.$$

Hence, the event above is precisely having, $p \in (0, 0.3) \cup (0,7,1)$. Finally, using the fact that, $h(\cdot)$ is symmetric around $1/2$; and is increasing on $[0, 1/2]$; and, decreasing on $[1/2, 1]$; the threshold α simply corresponds to,

$$\alpha = 20(1 - h(0.3)) \approx 2.37418.$$

2. Throughout, let, $S_n = \sum_{i=1}^{n} X_i$. Since, X_i's are i.i.d. Bernoulli trials; S_n is a binomial, with parameters, $(10, 1/2)$.
Next, we write the Z-statistics; which is,

$$Z = \frac{\sum_{i=1}^{n}(X_i - \mu_0)}{\sqrt{n\sigma_0^2}} = \frac{S_n - 5}{\sqrt{n}/2},$$

where, under H_0; X_i is Bernoulli, with mean, $\mu_0 = 1/2$; hence, $\sigma_0^2 = 1/4$. Equipped with this, we are demanding that, the significance of the test is, 0.109375, that is,

$$\mathbb{P}\left(|S_n - 5| > \frac{t\sqrt{10}}{2}\right) = 0.109375,$$

which implies,

$$\mathbb{P}\left(S_n > 5 + \frac{t\sqrt{10}}{2}\right) + \mathbb{P}\left(S_n < 5 - \frac{t\sqrt{10}}{2}\right) = 0.109375.$$

Now, the puncline is, for a $X \sim \mathrm{Bin}(10, 1/2)$; $\mathbb{P}(X > 7) = \mathbb{P}(X < 3) = \frac{0.109375}{2}$. Hence, letting,

$$\frac{t\sqrt{10}}{2} = 2 \implies t = \frac{4}{\sqrt{10}} \approx 1.265,$$

gives the desired result.
Side note: Since, S_n is a random variable, attaining integer values; we are free to select a threshold $\frac{t\sqrt{10}}{2}$; with,

$$\lfloor \frac{t\sqrt{10}}{2} \rfloor = 2,$$

where, $\lfloor \cdot \rfloor$ is a function, outputting the smallest integer, that is less than or equal to its input. Note that, for any such threshold (with, $2 + \epsilon$ of threshold, where, $0 \leq \epsilon < 1$),

$$\mathbb{P}(S_n > 7 + \epsilon) = \mathbb{P}(S_n \geq 8) = \mathbb{P}(S_n > 7),$$

and similarly,

$$\mathbb{P}(S_n < 5 - \epsilon) = \mathbb{P}(S_n < 3 - \epsilon) = \mathbb{P}(S_n \leq 2) = \mathbb{P}(S_n < 3),$$

giving us the same power for the test.

3. Let, t_1 and t_2 be the thresholds, for the GLRT and z-test, respectively. The rejection region, involving GLRT, can be written -using the fact that, $d(\hat{\mu}||1/2) = 1 - h(\hat{\mu})$; where these quantities have been introduced previously,

$$\begin{aligned}\{2nd(\hat{\mu}||1/2) \geq t_1\} &= \left\{1 - h(\hat{\mu}) \geq \frac{t_1}{2n}\right\} \\ &= \left\{h(\hat{\mu}) \leq 1 - \frac{t_1}{2n}\right\}.\end{aligned}$$

Let, $t \in [0, 1/2]$ be chosen, such that, $h(t) = -t\log t - (1-t)\log(1-t) = 1 - \frac{t_1}{2n}$; where, t_1 is the threshold defined above (here, we have used the fact that, $h(\cdot) : [0, 1/2] \to [0, 1]$ is a one-to-one, and onto function; hence it admits an inverse). The decision region obtained above, corresponds to,

$$
\begin{aligned}
&\{\hat{\mu} \leq t\} \cup \{\hat{\mu} \geq 1 - t\} \\
&= \left\{\hat{\mu} - \frac{1}{2} \leq t - \frac{1}{2}\right\} \cup \left\{\hat{\mu} - \frac{1}{2} \geq \frac{1}{2} - t\right\} \\
&= \left\{\left|\hat{\mu} - \frac{1}{2}\right| \geq \frac{1}{2} - t\right\},
\end{aligned}
$$

thus, the GLRT corresponds to testing the deviation of sample mean $\hat{\mu}$ from the mean under null hypothesis ($\mu_0 = 1/2$); across a threshold.
For the second test, the Z statistic can be written as,

$$
Z = \frac{\sum_{i=1}^{n} X_i - \frac{n}{2}}{\sqrt{n}/2} = \frac{S_n - n/2}{\sqrt{n}/2}.
$$

$$
\begin{aligned}
\{|Z| \geq t_2\} &= \left\{|S_n - n/2| \geq \frac{\sqrt{n}t_2}{2}\right\} \\
&= \left\{S_n \geq \frac{n}{2} + \frac{\sqrt{n}t_2}{2}\right\} \bigcup \left\{S_n \leq \frac{n}{2} - \frac{\sqrt{n}t_2}{2}\right\} \\
&= \left\{\hat{\mu} \geq \frac{1}{2} + \frac{t_2}{2\sqrt{n}}\right\} \bigcup \left\{\hat{\mu} \leq \frac{1}{2} - \frac{t_2}{2\sqrt{n}}\right\} \\
&= \left\{\left|\hat{\mu} - \frac{1}{2}\right| \geq \frac{t_2}{2\sqrt{n}}\right\}
\end{aligned}
$$

which is, not surprisingly, has a symmetry center of $1/2$.Hence, similar to GLRT; the $z-$test also corresponds to testing the deviation of sample mean $\hat{\mu}$ from the mean under null hypothesis ($\mu_0 = 1/2$); across a threshold. Therefore, both tests are equivalent.

Problem 2. Human sex ratio.

1. $n_{boys}/n_{tot} \approx 0.516275$. Let's construct the $z-$statistic. Note that, $\sigma_0^2 = \frac{1}{4}$; and, $n = 938223$. With these parameters, the $z-$statistic evaluates,

$$Z = \frac{\sum_{i=1}^{n}(X_i - 1/2)}{\sqrt{938223}/2} = \frac{484382 - 938223 \cdot 0.5}{\sqrt{938223}/2} \approx 31.53.$$

   It is very unlikely to see standard normal, deviated this much from its mean. Hence, we can declare that we rejected the null.

2. For this correction, the $z-$statistic evaulates,

$$Z = \frac{484382 - 0.515 \cdot 938223}{\sqrt{938223} \cdot \sqrt{0.515 - 0.515^2}} \approx 2.47.$$

   The $p-$value for this, is,

$$\mathbb{P}(|Z(X)| \geq 2.47|H) \approx \mathbb{P}(|Z| \geq 2.47) \approx 0.0135,$$

   where, Z is standard normal.

3. Begin by noticing, $\mathbb{E}[B_i] = n_i\pi_i$; and therefore, $\mathbb{E}[(B_i - n_i\pi_i)^2]$ is just the variance of B_i. Next, if n_i is the total number of borns in year i; we have,

$$\sum_{i=1}^{82} n_i = n_{tot} = 938223.$$

   Using these, we are ready to calculate the quantity being asked. We start with the expectation of $\frac{1}{n_{tot}}\sum_{i=1}^{82}(B_i - n_i\pi_i)^2$ under null hypothesis (namely, $\pi_i = 18/35$, for $i = 1, 2, \ldots, 82$).

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n_{tot}}\sum_{i=1}^{82}(B_i - n_i\pi_i)^2\right] &= \frac{1}{n_{tot}}\sum_{i=1}^{82}\mathbb{E}[(B_i - (\mathbb{E}[B_i])^2)] \\
&= \frac{1}{n_{tot}}\sum_{i=1}^{82}\operatorname{var}(B_i) \\
&= \frac{1}{n_{tot}}\sum_{i=1}^{82} n_i\pi_i(1 - \pi_i) \\
&= \frac{\sum_{i=1}^{82} n_i}{n_{tot}}\pi_i(1 - \pi_i) \\
&= \pi_i(1 - \pi_i) \approx 0.25.
\end{aligned}$$

Next, we compute the observed value of $\frac{1}{n_{tot}}\sum_{i=1}^{82}(B_i - n_i\pi_i)^2$. Note that, the observed value of B_i is b_i. We keep $\pi_i = \pi = 18/35$ throughout, and get

$$\begin{aligned}\frac{1}{n_{tot}}\sum_{i=1}^{82}(b_i - n_i\pi)^2 &= \frac{1}{n_{tot}}\left(\sum_{i=1}^{82} b_i^2 - \sum_{i=1}^{82} 2n_ib_i\pi + \sum_{i=1}^{82} n_i^2\pi^2\right)\\ &\approx 0.522644\end{aligned}$$

showing that, the variability is roughly double what we would expect.

Problem 3. Problem 3. London vs. country. We will use the formulation, as suggested by the hint. Let, μ_L be the probability that a baby born in London, is a girl; and, let, μ_R be the probability that, a baby born in Romsey is a girl. We have two hypotheses:

$$H : \mu_L = \mu_R \quad \text{and} \quad K : \mu_L < \mu_R.$$

Now, think of each birth as a sample from Bernoulli distribution. From London, we get iid samples of X_i; and these are 1 (meaning, a baby is girl), with probability μ_L. Similarly, from Romsey, we get iid samples of Y_i; and these are 1 (indicating a baby girl), with probability μ_R.
To construct the statistics for the two sample t-test (in general case), we begin by getting the empirical means (for girls):

$$\bar{X}_L = \frac{130866}{139782 + 130866} \approx 0.4835284 \quad \text{and} \quad \bar{Y}_R = \frac{3083}{3083 + 3256} \approx 0.4863543.$$

Next, estimating variances. We have selected a convention that, X_i's are from London; Y_i's from Romsey; and those are 1, if a baby is a girl.

$$\widehat{\sigma_L^2} = \frac{130866(1 - 0.4835284)^2 + 139782(0.4835284)^2}{270647} \approx 0.2497296.$$

$$\widehat{\sigma_R^2} = \frac{3083(1 - 0.4863543)^2 + 3256(0.4863543)^2}{6338} \approx 0.24981379.$$

Hence, the test statistics T turns out to be,

$$\begin{aligned} T &= \frac{\bar{Y}_R - \bar{X}_L}{\sqrt{\frac{\widehat{\sigma_L^2}}{270648} + \frac{\widehat{\sigma_R^2}}{6339}}} \\ &\approx \frac{\bar{Y}_R - \bar{X}_L}{\sqrt{\frac{\widehat{\sigma_R^2}}{6339}}} \\ &= \frac{0.4863543 - 0.4835284}{\sqrt{\frac{0.24981379}{6339}}} \\ &\approx 0.4501. \end{aligned}$$

Hence, the $p-$value, that is, $\mathbb{P}(T \geq t_{obs}|H)$ is obtained through, $1-\Phi^{-1}(0.4501) \approx 0.32 = 32\%$, where, $\Phi(x)$ is the probability, $\mathbb{P}(Z \leq x)$, where, Z is the standard normal random variable (here, we consider only one portion of the tail, as the hypothesis K is, $\mu_L < \mu_R$). Therefore, the result is not significant enough to reject null, and declare that, London is more apt.

Problem 5. Publication bias in "hot" fields.

1. To compute PPV, we will use Bayes' rule.

$$\begin{aligned}\mathbb{P}(H_1|R) &= \frac{\mathbb{P}(R|H_1)\mathbb{P}(H_1)}{\mathbb{P}(R)} \\ &= \frac{\mathbb{P}(R|H_1)\mathbb{P}(H_1)}{\mathbb{P}(R|H_0)\mathbb{P}(H_0)+\mathbb{P}(R|H_1)\mathbb{P}(H_1)} \\ &= \frac{\pi_1\beta}{\pi_1\beta+\pi_0\alpha},\end{aligned}$$

which evaluates as, 0.001996.
What actually matters (in the realm of a discovery) is PPV, rather than $p-$value (namely, $\mathbb{P}(\text{discovery}|H)$).

2. For this case, we compute PPV as follows.

$$\begin{aligned}\text{PPV} &= \mathbb{P}(H_1|R_m) \\ &= \frac{\mathbb{P}(R_m|H_1)\mathbb{P}(H_1)}{\mathbb{P}(R_m)} \\ &= \frac{\mathbb{P}(R_m|H_1)\mathbb{P}(H_1)}{\mathbb{P}(R_m|H_0)\mathbb{P}(H_0)+\mathbb{P}(R_m|H_1)\mathbb{P}(H_1)}.\end{aligned}$$

Now, the probabilities, $\mathbb{P}(R_m|H_1)$ are not very clear. For those, we do the following.

$$\mathbb{P}(R_m|H_1) = 1-\mathbb{P}\left(\bigcap_i R_m^c \middle| H_1\right) = 1-\mathbb{P}(R^c|H_1)^m = 1-(1-\mathbb{P}(R|H_1))^m = 1-(1-\beta)^m.$$

Similarly, $\mathbb{P}(R_m|H_0) = 1-(1-\alpha)^m$. Combining these two, we arrive at,

$$\text{PPV} = \frac{\pi_1\cdot(1-(1-\beta)^m)}{\pi_1\cdot(1-(1-\beta)^m)+\pi_0\cdot(1-(1-\alpha)^m)}.$$

Above, we used the independence, to compute the probability,

$$\mathbb{P}\left(\bigcup_i E_i\right) = 1-\mathbb{P}\left(\bigcap_i E_i^c\right) = 1-\prod_i(1-\mathbb{P}(E_i));$$

where, the first equality is De Morgan's law; the second equality is independence.

3. Inserting the values, we get, 0.000249; which roughly only doubled the prior odds! This is despite carrying out 10 expensive experiments and using a 95% significance threshold!