Statistical Interference | Caltech IDS/ACM/CS 157
Set 5 | May 2022

Dallas Taylor

1 Problem 1

Let $X_1, ..., X_n \sim \mathcal{U}[0, \theta]$, and suppose we want to test $H_0 : \theta = \frac{1}{2}$ versus $H_1 : \theta > \frac{1}{2}$. It seems natural to reject H_0 if $X_{(n)} = \max\{X_1, ..., X_n\}$ is large. So let's use a test with the rejection region of the following natural form: $\mathcal{R} = \{X : X_{(n)} > c\}$

Problem A: Find the power function of this test.

Solution A: *We find the power function of this test to be the following:*

$$\begin{aligned}
\beta(\theta) &= \begin{cases} \mathbb{P}(X \in \mathcal{R} \mid \theta = \frac{1}{2}), & \theta = \frac{1}{2} \\ \mathbb{P}(X \in \mathcal{R} \mid \theta > \frac{1}{2}), & \theta > \frac{1}{2} \end{cases} \\
\beta(\theta) &= \begin{cases} \mathbb{P}(X_{(n)} > c \mid \theta = \frac{1}{2}), & \theta = \frac{1}{2} \\ \mathbb{P}(X_{(n)} > c \mid \theta > \frac{1}{2}), & \theta > \frac{1}{2} \end{cases} \\
\beta(\theta) &= 1 - \left(\frac{c}{\theta}\right)^n
\end{aligned}$$

For the above, we assume that $0 \leq c \leq \theta$, however we can clearly determine what would occur if c was not in this range by using the definition of the uniform distribution. If $c < 0$, then we would clearly have that $\mathbb{P}(X_{(n)} > c) = 1$, and if $c > \theta$, then we would clearly have that $\mathbb{P}(X_{(n)} > c) = 0$.

Problem B: What choice of c will make the size of the test α?

Solution B: *We can find that the size of the test becomes α when we have the following:*

$$\begin{aligned}
\alpha &= \sup_{\theta = \frac{1}{2}} \beta(\theta) \\
\alpha &= \beta(\frac{1}{2}) \\
\alpha &= 1 - (2c)^n \\
\therefore c &= \frac{(1-\alpha)^{1/n}}{2}
\end{aligned}$$

Dallas Taylor

Problem C: If a sample size $n = 20$ and $X_{(n)} = 0.48$, what is the p-value?

Solution C: *We find the p-value to be the following:*

$$\begin{aligned} p(x) &= \inf_{\alpha \in (0,1)} \{\alpha : X \in \mathcal{R}_\alpha\} \\ p(x) &= \inf_{\alpha \in (0,1)} \{\alpha : X_{(n)} > \frac{(1-\alpha)^{1/n}}{2}\} \end{aligned}$$

Therefore we want to find when the following is satisfied (as we are minimizing α):

$$\begin{aligned} \frac{(1-\alpha)^{1/n}}{2} &= X_{(n)} \\ (1-\alpha) &= (2X_{(n)})^n \\ \alpha &= 1 - (2X_{(n)})^n \\ \therefore\ \alpha &= 1 - (2(0.48))^{20} = 0.558 \end{aligned}$$

Dallas Taylor

2 Problem 2

Let $X_1, ..., X_n \sim \mathcal{N}(\mu, \sigma^2)$, where $\sigma^2 = 1$. Suppose we want to test $H_0 : \mu = 0$ versus $H_1 : \mu = 1$. Consider a test with the rejection region $R = \{X : \bar{X}_n > c\}$.

Problem A: Construct a test of size α.

Solution A: *We can first find the power function of the test to be:*

$$\begin{aligned}
\beta(\mu) &= \mathbb{P}(\bar{X}_n > c \mid \mu) \\
\beta(\mu) &= \mathbb{P}\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} > \frac{\sqrt{n}(c-\mu)}{\sigma}\right) \\
\beta(\mu) &= 1 - \Phi\left(\sqrt{n}(c-\mu)\right)
\end{aligned}$$

We know that $\alpha = \beta(0)$ from the monotonicity of β. Therefore, we find our equation for c to be the following:

$$c = \frac{\Phi^{-1}(1-\alpha)}{\sqrt{n}}$$

Problem B: Find its power under H_1.

Solution B:

$$\begin{aligned}
\beta(1) &= \mathbb{P}(\bar{X}_n > c \mid \mu = 1) = 1 - \Phi\left(\sqrt{n}(c-1)\right) \\
\beta(1) &= 1 - \Phi\left(\sqrt{n}\left(\frac{\Phi^{-1}(1-\alpha)}{\sqrt{n}} - 1\right)\right) \\
\beta(1) &= 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \sqrt{n}\right)
\end{aligned}$$

Problem C: Find the limit of the power function as $n \to \infty$.

Solution C: *Let us first consider the general power function found in Problem A:*

$$\lim_{n\to\infty} \beta(\mu) = \lim_{n\to\infty} 1 - \Phi(\Phi^{-1}(1-\alpha) - \mu\sqrt{n}) = 1 - \Phi(-\infty) = 1$$

We can see that the limit above holds for $\mu > 0$, such as $\mu = 1$ (Problem B). However, when $\mu = 0$, $\lim_{n\to\infty} \beta(0) = \alpha$.

Dallas Taylor

3 Problem 3

The Poisson distribution is often used for modeling the number of times a certain event occurs in an interval of time or space. Let $X_1, ..., X_n \sim \text{Poisson}(\lambda)$. Suppose we want to test

$$H_0 : \lambda = \lambda_0 \;\text{ versus }\; H_1 : \lambda \neq \lambda_0,$$

where $\lambda_0 > 0$ is some constant.

Problem A: Construct the size α Wald test. To estimate λ, use the maximum likelihood method.

Solution A: *We first want to find the Wald statistic and thus our size α Wald Test:*

$$\begin{aligned} W &= \left|\frac{\hat{\lambda}_{MLE} - \lambda_0}{\hat{se}}\right| \\ \therefore W &= \left|\frac{\hat{\lambda}_{MLE} - \lambda_0}{\hat{se}}\right| > z_{1-\frac{\alpha}{2}} \end{aligned}$$

We can find $\hat{\lambda}_{MLE}$ to be found by the following:

$$\begin{aligned} \mathcal{L}(\lambda \mid X_1, ..., X_n) &= \prod_{i=1}^{n} f(X_i; \lambda) \\ \mathcal{L}(\lambda \mid X_1, ..., X_n) &= \prod_{i=1}^{n} \frac{\lambda^{X_i}}{X_i} e^{-\lambda} = e^{-n\lambda} \prod_{i=1}^{n} \frac{\lambda^{X_i}}{X_i} \\ \mathcal{L}(\lambda \mid X_1, ..., X_n) &= e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^{n} X_i}}{\prod_{i=1}^{n} X_i} \\ \therefore \frac{d}{d\lambda}\mathcal{L} &= -ne^{-n\lambda} \frac{\lambda^{\sum_{i=1}^{n} X_i}}{\prod_{i=1}^{n} X_i} + e^{-n\lambda} \frac{\sum_{i=1}^{n} X_i \lambda^{-1+\sum_{i=1}^{n} X_i}}{\prod_{i=1}^{n} X_i} \\ \frac{d}{d\lambda}\mathcal{L} &= e^{-n\lambda} \frac{\lambda^{-1+\sum_{i=1}^{n} X_i}}{\prod_{i=1}^{n} X_i} \left(-n\lambda + \sum_{i=1}^{n} X_i\right) \\ \therefore 0 &= e^{-n\hat{\lambda}_{MLE}} \frac{\hat{\lambda}_{MLE}^{-1+\sum_{i=1}^{n} X_i}}{\prod_{i=1}^{n} X_i} \left(-n\hat{\lambda}_{MLE} + \sum_{i=1}^{n} X_i\right) \\ 0 &= -n\hat{\lambda}_{MLE} + \sum_{i=1}^{n} X_i \\ \hat{\lambda}_{MLE} &= \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}_n \end{aligned}$$

We verify that this our desired maximum by examining the second derivative of $\mathcal{L}$:

$$\begin{aligned}
\frac{d^2}{d\lambda^2}\mathcal{L} &= e^{-n\lambda}\left(\frac{\lambda^{-2+\sum_{i=1}^n X_i}}{\prod_{i=1}^n X_i}\right)\left((n\lambda - \sum_{i=1}^n X_i)^2 - \sum_{i=1}^n X_i\right) \\
\therefore \frac{d^2}{d\lambda^2}\mathcal{L}(\hat{\lambda}_{MLE}) &= e^{-n\bar{X}_n}\left(\frac{\bar{X}_n^{-2+\sum_{i=1}^n X_i}}{\prod_{i=1}^n X_i}\right)\left((n\bar{X}_n - \sum_{i=1}^n X_i)^2 - \sum_{i=1}^n X_i\right) \\
\frac{d^2}{d\lambda^2}\mathcal{L}(\hat{\lambda}_{MLE}) &= e^{-n\bar{X}_n}\left(\frac{\bar{X}_n^{-2+\sum_{i=1}^n X_i}}{\prod_{i=1}^n X_i}\right)\left(-\sum_{i=1}^n X_i\right) \\
\frac{d^2}{d\lambda^2}\mathcal{L}(\hat{\lambda}_{MLE}) &= (+)(+)(-) < 0
\end{aligned}$$

We clearly have that $\mathcal{L}$ is negative at $\hat{\lambda}_{MLE}$ since all X_i must be positive for a Poisson distribution.

Furthermore, we can determine that $\hat{se}$ for our test is found using asymptotic normality (Properties of MLEs). We also know that the Fisher Information for the Poisson distribution is given to be $1/\lambda$, and thus we produce the following:

$$\begin{aligned}
\hat{se} &= \frac{1}{\sqrt{nI(\hat{\lambda}_{MLE})}} = \frac{1}{\sqrt{\frac{n}{\hat{\lambda}_{MLE}}}} = \sqrt{\frac{\bar{X}_n}{n}} \\
\therefore W &= \left|\frac{\bar{X}_n - \lambda_0}{\sqrt{\bar{X}_n/n}}\right| > z_{1-\alpha/2}
\end{aligned}$$

Problem B:

Solution B: *See attached scripts.*

4 Problem 4

Solution : *See attached scripts.*

5 Problem 5

Solution : *See attached scripts.*

IDS/ACM 157 PS5 MatLab - Problem 3

Part b

Set $\lambda_0 = 1, n = 20, \alpha = 0.05$. Simulate the data $X_1, \ldots, X_n \sim \text{Poisson}(\lambda_0)$ and perform the Wald test constructed in part (a). Repeat $m = 10^4$ times and report the estimated type I error rate. Write a script that implements the task in part (b). Write you results and conclusions as comments in the script. Hint: the following MATLAB function can be useful: poissrnd.

```
lambda0 = 1;
n = 20;
alpha = 0.05;
M = 10^4;
z_alpha = 1.96;

data = poissrnd(lambda0,n,1);
X_bar = mean(data);
W = abs((X_bar - lambda0)/sqrt(X_bar/n));
disp('Wald statistic:'); disp(W);
```

```
Wald statistic:
    0.2294
```

```
disp('Reject H0?'); disp(W > z_alpha); % Note that 0 is false and 1 is true
```

```
Reject H0?
   0
```

```
t1_rejects = 0;
for m = 1:M
    data_m = poissrnd(lambda0,n,1);
    X_bar_m = mean(data_m);
    W_m = abs((X_bar_m - lambda0)/sqrt(X_bar_m/n));
    if W_m > z_alpha
        t1_rejects = t1_rejects + 1;
    end
end
t1_rejects = t1_rejects / M;
disp('Estimated Type I Error Rate:'); disp(t1_rejects);
```

```
Estimated Type I Error Rate:
    0.0517
```

Discussion

We can clearly see that we have an extremely low Type I Error Rate. Thus, we can determine that H_0 is not frequently rejected by our test.

IDS/ACM 157 PS5 MatLab - Problem 4

A geologist wonders if the surface soil pH levels at two different locations of a certain desert site are similar. The scientist obtained the following pH levels at randomly selected points within each of the two locations.

Location A : $7.58,\ 8.52,\ 8.01,\ 7.99,\ 7.93,\ 7.89,\ 7.85,\ 7.82,\ 7.80$

Location B : $7.85,\ 7.73,\ 8.53,\ 7.40,\ 7.35,\ 7.30,\ 7.27,\ 7.27,\ 7.23$

Does the data suggest that the true mean soil pH values differ for the two location? Write a script that implements this task. Write you results and conclusions as comments in the script.

```
A = [7.58,8.52,8.01,7.99,7.93,7.89,7.85,7.82,7.80];
B = [7.85,7.73,8.53,7.40,7.35,7.30,7.27,7.27,7.23];
X_bar_A = mean(A);
X_bar_B = mean(B);
n = length(A);
disp('Mean pH at location A:'); disp(X_bar_A);
```

```
Mean pH at location A:
    7.9322
```

```
disp('Mean pH at location B:'); disp(X_bar_B);
```

```
Mean pH at location B:
    7.5478
```

```
K = 10^5;
s_obs = abs(X_bar_A - X_bar_B);
Z = [A B];

p_val = 0;
for k = 1:K
    Z_pi = Z(:,randperm(2*n));
    X_bar_A_k = mean(Z_pi(1:n));
    X_bar_B_k = mean(Z_pi(n+1:2*n));
    s_pi = abs(X_bar_A_k - X_bar_B_k);
    if s_pi > s_obs
        p_val = p_val + 1;
    end
end
p_val = p_val / K;
disp('Estimated p-value:'); disp(p_val);
```

```
Estimated p-value:
    0.0337
```

Discussion

The data produces a p-value of ~ 0.035. Thus, we can determine that there is a probability of ~ 0.035 that we observe a statistic more extreme than s_{obs}, which is evidence favoring the rejection of our null hypothesis H_0. Thus, we are able to determine that there is a strong likelihood that the mean soil pH values are different between location A and location B. Specifically such that the mean of the pH of the soil at location A is higher than that of location B.

IDS/ACM 157 PS5 MatLab - Problem 5

In 1861, ten essays appeared in the *New Orleans Daily Crescent*. They were signed "Quintus Curtius Snodgrass". Some people suspected they were actually written by Mark Twain. To investigate this, we will consider the proportion of the three-letter words found in an author's work.

In eight Twain essays, the proportions are:

$0.225, 0.262, 0.217, 0.240, 0.230, 0.229, 0.235, 0.217.$

In ten Snodgrass essays, the proportions are:

$0.209, 0.205, 0.196, 0.210, 0.202, 0.207, 0.224, 0.223, 0.220, 0.201.$

Part a

Perform the Wald test for equality of the means and report the p-value. What do you conclude?

```
T = [0.225,0.262,0.217,0.240,0.230,0.229,0.235,0.217];
S = [0.209,0.205,0.196,0.210,0.202,0.207,0.224,0.223,0.220,0.201];
X_bar_T = mean(T);
X_bar_S = mean(S);
var_T = var(T);
var_S = var(S);
n_T = 8;
n_S = 10;
disp('Mean Twain proportions:'); disp(X_bar_T);
```

```
Mean Twain proportions:
    0.2319
```

```
disp('Mean Snodgrass proportions:'); disp(X_bar_S);
```

```
Mean Snodgrass proportions:
    0.2097
```

```
W = abs((X_bar_T - X_bar_S)/sqrt(var_T/n_T + var_S/n_S));
p_val = 2 * normcdf(-1 * W);
disp('p-value from Wald test:'); disp(p_val);
```

```
p-value from Wald test:
   2.1260e-04
```

Discussion

We can see that we have a p-value of $\sim 2.13 \times 10^{-4}$, and thus we clearly have evidence that favors rejection of the null hypothesis H_0. Thus, we can determine that it is very likely that Twain is not the author of the Snodgrass essays.

Part b

To avoid large sample normality assumptions, perform the permutation test. What is your conclusion?

```
K = 10^5;
```

```
s_obs = abs(X_bar_T - X_bar_S);
Z = [T S];

p_val = 0;
for k = 1:K
    Z_pi = Z(:,randperm(n_T + n_S));
    X_bar_T_k = mean(Z_pi(1:n_T));
    X_bar_S_k = mean(Z_pi(n_T+1:n_T + n_S));
    s_pi = abs(X_bar_T_k - X_bar_S_k);
    if s_pi > s_obs
        p_val = p_val + 1;
    end
end
p_val = p_val / K;
disp('Estimated p-value:'); disp(p_val);
```

```
Estimated p-value:
   7.5000e-04
```

Discussion

We can see that we have an estimated p-value of $\sim 7.5 \times 10^{-4}$, which is even closer to 0 than the p-value calculated in Part a. Thus, we can determine that we still have strong evidence in favor of rejecting H_0, even when avoiding large sample normality assumptions. Thus, there is a strong likelihood that Twain did not write the Snodgrass essays once again.