Homework 3

(due on Thursday, Feb 1, 2018) When solving the problems below as well as future homework problems, please give detailed derivations and arguments in order to receive credit. In your solution do not forget to include your name and the homework number. Please staple your pages together.

1. Consider i.i.d. observations $X_1, ..., X_n \sim N(\mu, 1)$. The hypothesis test problem is $H_0 : \mu \leq 0$ against $H_1 : \mu > 0$. In the class, we derived that the rejection event should be $\{\sqrt{n}\bar{X} > z_{1-\alpha}\}$. Thus, the power function is $g(\mu) = \mathbb{P}_\mu(\sqrt{n}\bar{X} > z_{1-\alpha})$. Prove that $g(\mu)$ is an increasing function by taking derivative.

   $\sqrt{n}(\bar{X} - \mu) \sim N(0, 1)$

   $$\begin{aligned} g(\mu) &= P_\mu(\sqrt{n}\bar{X} > z) \\ &= 1 - P_\mu(\sqrt{n}\bar{X} \leq z) \\ &= 1 - P_\mu(\sqrt{n}\bar{X} - \sqrt{n}\mu \leq z - \sqrt{n}\mu) \\ &= 1 - P_\mu(Z \leq z - \sqrt{n}\mu) \\ g'(\mu) &= f(z - \sqrt{n}\mu)\sqrt{n} > 0 \end{aligned}$$

   Therefore, $g(\mu)$ is an increasing function of μ.

2. Consider independent observations $X_1, ..., X_n \sim N(\mu_1, 1)$ and $Y_1, ..., Y_m \sim N(\mu_2, 1)$. The hypothesis test problem is $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$. This is called a two-sample test problem.

   (a) Since we want to know whether μ_1 and μ_2 are equal or not, it is natural to consider the summary statistic $\bar{X} - \bar{Y} = \frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{m}\sum_{j=1}^m Y_j$. What is the distribution of this statistic under H_0?

   $$\frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu_1, \frac{1}{n}) \quad \frac{1}{m}\sum_{j=1}^{n} Y_j \sim N(\mu_2, \frac{1}{m})$$

   $$\bar{X} - \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{m}\sum_{j=1}^{m} Y_j \sim N(\mu_1 - \mu_2, \frac{1}{n} + \frac{1}{m}) \quad \text{X, Y independent}$$

   Under H_0:

   $$\bar{X} - \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{m}\sum_{j=1}^{m} Y_j \sim N(0, \frac{1}{n} + \frac{1}{m})$$

   (b) Construct a rejection region of Type-I error α.

   $$\alpha = P(|\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{1}{n} + \frac{1}{m}}}| > t)$$

   $$t = z_{1-\frac{\alpha}{2}}$$

The rejection region is $\{|\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{1}{n}+\frac{1}{m}}}| > z_{1-\frac{\alpha}{2}}\}$

(c) Derive the condition under which the power tends to 1.

$$
\begin{aligned}
\text{Power} &= P_\mu(|\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{1}{n}+\frac{1}{m}}}| > z_{1-\frac{\alpha}{2}}) \\
&= P_\mu(|\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{1}{n}+\frac{1}{m}}} + \frac{\mu_1-\mu_2}{\sqrt{\frac{1}{n}+\frac{1}{m}}}| > z_{1-\frac{\alpha}{2}}) \\
&= P_\mu(|N(0,1) + \frac{\mu_1-\mu_2}{\sqrt{\frac{1}{n}+\frac{1}{m}}}| > z_{1-\frac{\alpha}{2}}) \\
&= P_\mu(N(0,1) > z_{1-\frac{\alpha}{2}} - \frac{\mu_1-\mu_2}{\sqrt{\frac{1}{n}+\frac{1}{m}}} \; or \; N(0,1) < -z_{1-\frac{\alpha}{2}} - \frac{\mu_1-\mu_2}{\sqrt{\frac{1}{n}+\frac{1}{m}}})
\end{aligned}
$$

The power tends to 1 if $|\frac{\mu_1-\mu_2}{\sqrt{\frac{1}{n}+\frac{1}{m}}}| \to \infty$.

(d) For the testing procedure with $\alpha = 0.05$, derive the condition under which the power is at least 0.95. Then, suppose $n = 20$ and $m = 30$, what is the smallest gap $|\mu_1 - \mu_2|$ that allows you to reject the null correctly with probability at least 0.95?

$$
\text{Power} = P_\mu(N(0,1) > 1.96 - \frac{\mu_1-\mu_2}{\sqrt{\frac{1}{n}+\frac{1}{m}}} \; or \; N(0,1) < -1.96 - \frac{\mu_1-\mu_2}{\sqrt{\frac{1}{n}+\frac{1}{m}}}) \geq 0.95
$$

$|\frac{\mu_1-\mu_2}{\sqrt{\frac{1}{n}+\frac{1}{m}}}| \geq 1.96 + 1.64$

With n=20, m=30,

$$
\begin{aligned}
|\frac{\mu_1-\mu_2}{\sqrt{\frac{1}{20}+\frac{1}{30}}}| &\geq 3.6 \\
|\mu_1-\mu_2| &\geq 1.039
\end{aligned}
$$

The smallest gap is 1.039.

3. Consider i.i.d. observations $X_1, ..., X_n \sim$Bernoulli(p). The hypothesis test problem is $H_0 : p = \frac{1}{2}$ against $H_1 : p \neq \frac{1}{2}$.

(a) Find the MLE of p.

From HW2, $\hat{p} = \bar{X}$.

(b) Use the MLE as your summary statistic and derive its asymptotic distribution under H_0.

$$\frac{\sqrt{n}(\hat{p}-p)}{\sqrt{p(1-p)}} \sim N(0,1)$$

Under H_0:

$$2\sqrt{n}(\hat{p}-\frac{1}{2}) \sim N(0,1)$$

(c) Construct an α-level test.
Reject H_0 if $\{2\sqrt{n}\hat{p}-\sqrt{n} \geq z_{1-\frac{\alpha}{2}}\}$

(d) For a more general problem $H_0: p = p_0$ vs $H_1: p \neq p_0$, do (b) and (c) again.
Under H_0:

$$\frac{\sqrt{n}(\hat{p}-p_0)}{\sqrt{p_0(1-p_0)}} \sim N(0,1)$$

The rejection region is $\{|\frac{\sqrt{n}(\hat{p}-p_0)}{\sqrt{p_0(1-p_0)}}| \geq z_{1-\frac{\alpha}{2}}\}$

(e) Can you invert your test into a confidence interval of p_0? What is the name of this confidence interval?
Similar as HW2 Q1d,

$$|\frac{\sqrt{n}(\hat{p}-p_0)}{\sqrt{p_0(1-p_0)}}| \leq z_{1-\frac{\alpha}{2}}$$

$$(\hat{p}-p_0)^2 \leq \frac{p_0(1-p_0)}{n} z^2_{1-\frac{\alpha}{2}}$$

$$f(p_0) = (1+\frac{z^2_{1-\frac{\alpha}{2}}}{n})p_0^2 - (2\hat{p}+\frac{z^2_{1-\frac{\alpha}{2}}}{n})p_0 + \hat{p}^2 \leq 0$$

$(1-\alpha)$ CI is $[\frac{\hat{p}+\frac{z^2_{1-\frac{\alpha}{2}}}{2n} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{z^2_{1-\frac{\alpha}{2}}}{4n}}}{1+\frac{z^2_{1-\frac{\alpha}{2}}}{n}}, \frac{\hat{p}+\frac{z^2_{1-\frac{\alpha}{2}}}{2n} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{z^2_{1-\frac{\alpha}{2}}}{4n}}}{1+\frac{z^2_{1-\frac{\alpha}{2}}}{n}}]$
Wilson confidence interval.
Note: Wald is not the equivalent CI.

4. A very general way to construct a summary statistic is by study the the likelihood ratio. Consider i.i.d. observations $X_1, ..., X_n \sim N(\mu, 1)$. The test problem is $H_0: \mu \leq 0$ against $H_1: \mu > 0$. The notation $\phi_\mu(x)$ stands for the density function of $N(\mu, 1)$.

(a) Take some $\mu_0 \leq 0$ (from null) and some $\mu_1 > 0$ (from alternative). The idea of likelihood ratio test is to compare the likelihood functions $\prod_{i=1}^n \phi_{\mu_0}(X_i)$ and $\prod_{i=1}^n \phi_{\mu_1}(X_i)$. Show that the likelihood ratio $\frac{\prod_{i=1}^n \phi_{\mu_1}(X_i)}{\prod_{i=1}^n \phi_{\mu_0}(X_i)}$ is an increasing function of $\bar{X}$.

$$\frac{\prod_{i=1}^n \phi_{\mu_1}(X_i)}{\prod_{i=1}^n \phi_{\mu_0}(X_i)} = \frac{\exp(-\frac{1}{2}\sum_{i=1}^n (X_i-\mu_1)^2)}{\exp(-\frac{1}{2}\sum_{i=1}^n (X_i-\mu_0)^2)}$$
$$= \exp(-\frac{1}{2}\sum_{i=1}^n (X_i-\mu_1)^2 + \frac{1}{2}\sum_{i=1}^n (X_i-\mu_0)^2)$$
$$= \exp(n\bar{X}(\mu_1-\mu_0) - \frac{1}{2}(\mu_1^2-\mu_0^2))$$

$\mu_1 - \mu_0 > 0 \Rightarrow$ the likelihood ratio $\frac{\prod_{i=1}^n \phi_{\mu_1}(X_i)}{\prod_{i=1}^n \phi_{\mu_0}(X_i)}$ is an increasing function of $\bar{X}$.

(b) The likelihood ratio test is to reject H_0 if $\frac{\prod_{i=1}^n \phi_{\mu_1}(X_i)}{\prod_{i=1}^n \phi_{\mu_0}(X_i)} > C$ for some threshold $C > 0$. Explain why this is a good idea?

The likelihood for a model means the probability of the data under the model. If $\frac{\prod_{i=1}^n \phi_{\mu_1}(X_i)}{\prod_{i=1}^n \phi_{\mu_0}(X_i)} > C$, the probability of the data under the model with μ_1 is larger than C times the probability of the data under the model with μ_0. The data is more likely be generated by model with μ_1.

Note:

Here, we have continuous data. The likelihood ratio is

$$LR = \frac{\prod_{i=1}^n P_{\mu_1}(X \in [X_i-\epsilon, X_i+\epsilon])}{\prod_{i=1}^n P_{\mu_0}(X \in [X_i-\epsilon, X_i+\epsilon])} \approx \frac{\prod_{i=1}^n 2\epsilon\phi_{\mu_1}(X_i)}{\prod_{i=1}^n 2\epsilon\phi_{\mu_0}(X_i)} = \frac{\prod_{i=1}^n \phi_{\mu_1}(X_i)}{\prod_{i=1}^n \phi_{\mu_0}(X_i)}$$

When we do the hypothesis testing, the null hypothesis is often stated as $\mu \in \Theta_0$. The alternative hypothesis is $\mu \in \Theta_0^C$. The likelihood ratio can be written as

$$LR = \frac{\sup_{\mu\in\Theta_0^C} \prod_{i=1}^n \phi_\mu(X_i)}{\sup_{\mu\in\Theta_0} \prod_{i=1}^n \phi_\mu(X_i)}$$

This likelihood ratio depends only on the minimal sufficient statistic. The maximum likelihood estimator μ (in Θ_0, or in Θ_0^C) is a function of minimal sufficient statistic.

(c) Show that $\frac{\prod_{i=1}^n \phi_{\mu_1}(X_i)}{\prod_{i=1}^n \phi_{\mu_0}(X_i)} > C$ is equivalent to $\sqrt{n}\bar{X} > t$ for some t?

$$\frac{\prod_{i=1}^n \phi_{\mu_1}(X_i)}{\prod_{i=1}^n \phi_{\mu_0}(X_i)} = \exp(n\bar{X}(\mu_1-\mu_0) - \frac{1}{2}(\mu_1^2-\mu_0^2)) > C$$
$$n\bar{X}(\mu_1-\mu_0) - \frac{1}{2}(\mu_1^2-\mu_0^2) > \log C$$
$$\sqrt{n}\bar{X} > \frac{1}{\sqrt{n}(\mu_1-\mu_0)}(\log C + \frac{1}{2}(\mu_1^2-\mu_0^2)) = t$$

(d) What is an α-level likelihood ratio test?

Reject H_0 if $\{\frac{\prod_{i=1}^n \phi_{\mu_1}(X_i)}{\prod_{i=1}^n \phi_{\mu_0}(X_i)} > \exp(z_{1-\frac{\alpha}{2}}\sqrt{n}(\mu_1-\mu_0) - \frac{1}{2}(\mu_1^2-\mu_0^2))\}$

5. Consider i.i.d. observations $X_1, ..., X_n \sim$Poisson(λ). Let $p_\lambda(k)$ be the probability mass function of Poisson(λ). Show that the likelihood ratio $\frac{\prod_{i=1}^n p_{\lambda_1}(X_i)}{\prod_{i=1}^n p_{\lambda_0}(X_i)}$ is a monotone function of $\bar{X}$.

$$\frac{\prod_{i=1}^n p_{\lambda_1}(X_i)}{\prod_{i=1}^n p_{\lambda_0}(X_i)} = \frac{\lambda_1^{\sum_{i=1}^n X_i}}{\lambda_0^{\sum_{i=1}^n X_i}} = \frac{\lambda_1}{\lambda_0}^{n\bar{X}}$$

which is a monotone function of $\bar{X}$.

6. Consider i.i.d. observations $X_1, ..., X_n \sim N(\mu, 1)$. The testing problem is $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$. An 0.05-level test is $\{|\sqrt{n}\bar{X}| > z_{0.975}\}$.

   (a) If you observe $|\sqrt{n}\bar{X}| = 10$, will you reject the null hypothesis? What is the p-value?

   Yes.

$$p = P(|N(0,1) >= 10|) \approx 0$$

   (b) If you observe $|\sqrt{n}\bar{X}| = 1$, will you reject the null hypothesis? What is the p-value?

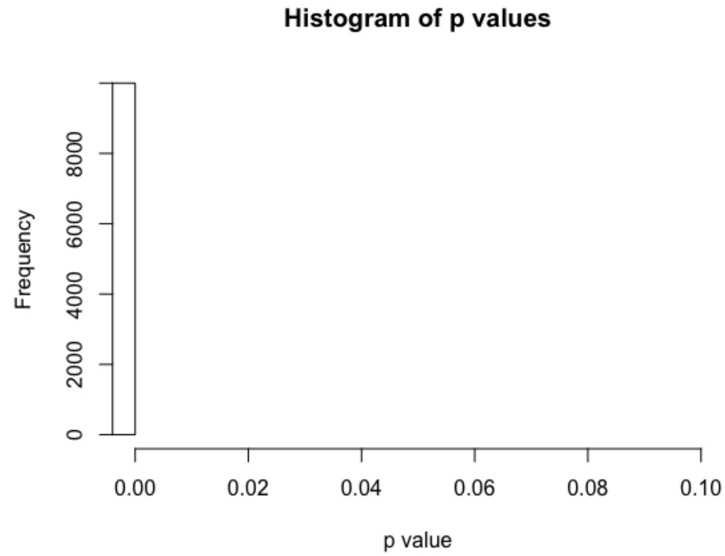
   No. $p = P(|N(0,1) >= 1|) \approx 0.3173$

   (c) Simulate i.i.d. observations from $N(50, 1)$ with $n = 100$. Calculate your p-value.

```
> n=100
> x = rnorm(n, 50, 1)
> pnorm(-abs(sqrt(n) * mean(x)))*2
[1] 0
```

   The p value is 0.

   (d) Repeat (c) for 10000 times, and then you have 10000 p-values. Plot the histogram of those 10000 numbers.



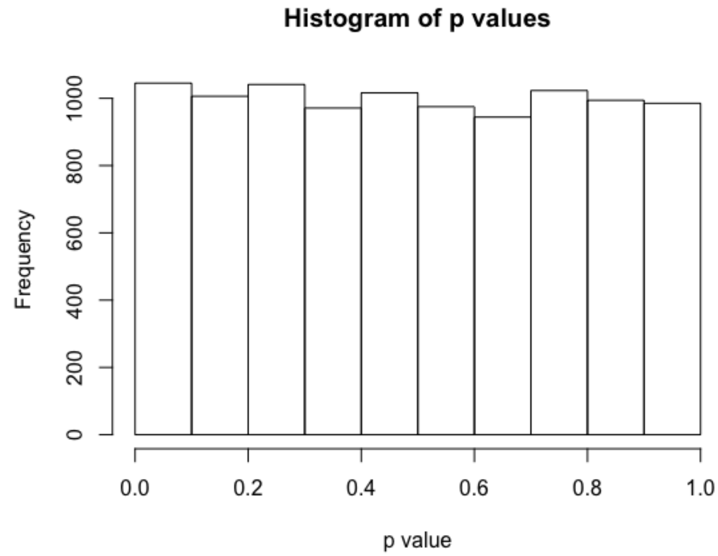

```
p.vec = numeric(10000)
for(i in 1:10000){
n=100
x = rnorm(n, 50, 1)
p.vec[i] = pnorm(-abs(sqrt(n) * mean(x)))*2
}
hist(p.vec, xlim=c(0,0.1), breaks=10, xlab='p value', main = '
    Histogram of p values')
```

(e) Simulate i.i.d. observations from $N(0,1)$ with $n = 100$. Calculate your p-value.

```
> n=100
> x = rnorm(n, 0, 1)
> pnorm(-abs(sqrt(n) * mean(x)))*2
[1] 0.2347
```

The p value is 0.2347.

(f) Repeat (e) for 10000 times, and then you have 10000 p-values. Plot the histogram of those 10000 numbers. How many numbers are below 0.05? Why?



There are 530 p values are below 0.05. Because of randomness, the generated data could have $|\sqrt{n}\bar{X}| > 1.96$ with probability 0.05. So we have around 5% p values that are less than 0.05.

```
p.vec = numeric(10000)
for(i in 1:10000){
n=100
x = rnorm(n, 0, 1)
p.vec[i] = pnorm(-abs(sqrt(n) * mean(x)))*2
}
hist(p.vec, xlim=c(0,1), breaks=10, xlab='p value', main = '
    Histogram of p values')
```

(g) Discuss the two plots you get in (d) and (f).

The plot in f shows that if the null hypothesis is true, the p value is uniformly distributed in (0,1). If the null hypothesis is not true (plot in d), the distribution of p value is right-skewed.

(h) If you have a slow computer, you may run the above experiments for 1000 times instead of 10000 times.