SFU

STAT 330
Tutorial 11

Mengqi (Molly) Cen
Department of Statistics & Actuarial Science

Nov 30th/ Dec 2nd

Outline

Nonparametric and Robust Statistics

- Sign Test
- Signed-Rank Test
- MWW Test
- Kendall's τ

Bayesian Statistics

- Prior and Posterior Distribution
- Bayesian Point Estimation

Sign Test

Let X_i denote the length, in centimeters, of a randomly selected pygmy sunfish, $i = 1, 2, \ldots 10$. If we obtain the following data set:

5.0 3.9 5.2 5.5 2.8 6.1 6.4 2.6 1.7 4.3

can we conclude that the median length of pygmy sunfish differs significantly from 3.7 centimeters?

Exact Approach:

We are interested in testing $H_0: m = 3.7$ vs $H_1: m \neq 3.7$.

First, we calculate $X_i - m$ for $i = 1, \ldots, 10$.

We get: 1.3, 0.2, 1.5, 1.8, -0.9,

2.4, 2.7, -1.1, -2.0, 0.6.

By the sign test, $S = 7 \sim B(10, 1/2)$ under H_0.

From the binomial table ($n = 10, x = 7, p = 0.5$),

$P(H_0 \text{ is true}) = 0.117 \times 2 = 0.234 > 0.05$.

Thus, we can not conclude that the median length of pygmy sunfish differs significantly from 3.7 centimeters.

Approximate Approach:

If $n >> 1$,

$$\frac{S(\theta_0) - n/2}{\sqrt{n/4}} \sim N(0,1) \text{ approximately.}$$

Signed-Rank Test

Let X_i denote the length, in centimeters, of a randomly selected pygmy sunfish, $i = 1, 2, \ldots 10$. If we obtain the following data set:

5.0 3.9 5.2 5.5 2.8 6.1 6.4 2.6 1.7 4.3

can we conclude that the median length of pygmy sunfish differs significantly from 3.7 centimeters?

Exact Approach :

We are interested in testing $H_0: m = 3.7$ vs $H_1: m \neq 3.7$.

First, we calculate $X_i - m$ for $i = 1, \ldots, 10$.

We get: 1.3, 0.2, 1.5, 1.8, -0.9,
2.4, 2.7, -1.1, -2.0, 0.6.

Then, we calculate $|X_i - m|$ for $i = 1, \ldots, 10$.

We get: 1.3, 0.2, 1.5, 1.8, 0.9,
2.4, 2.7, 1.1, 2.0, 0.6.

The rank is: 5, 1, 6, 7, 3,
9, 10, 4, 8, 2.

$T = 5\times1 + 1\times1 + 6\times1 + 7\times1 + 3\times0 + 9\times1 + 10\times1 + 4\times0 + 8\times0 + 2\times1 = 40$

From the distribution table ($n = 10, w_2^* = 40$)

$P(H_0 \text{ is true}) = 0.116 \times 2 = 0.232 > 0.05$.

Thus, we can not conclude that the median length of pygmy sunfish differs significantly from 3.7 centimeters.

Approximate Approach :

If $n >> 1$,

$$\frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \sim N(0,1) \text{ approximately.}$$

MWW Test

The following data shows the age at diagnosis of type II diabetes in young adults. Is the age at diagnosis different for males and females?

Males:	19 22 16 29 24
Females:	20 11 17 12

Exact Approach:

We are interested in $H_0: \Delta = 0$, $H_1: \Delta \neq 0$.

for $i = 1, \ldots, 5$, $j = 1, \ldots, 4$.

First, we rank 9 observations: 11, 12, 16, 17, 19, 20, 22, 24.

The rank of females are 6, 1, 4, 2.

Thus, $W = 6 + 1 + 4 + 2 = 13$.

From the distribution table ($n_1 = 5, n_2 = 4$),

p value is between 0.1 and 0.2.

Thus, we can not conclude the age at diagnosis different for males and females.

Approximate Approach:

If $n >> 1$,

$$\frac{W - n_1(n+1)/2}{\sqrt{n_1 n_2 (n+1)/12}} \sim N(0,1) \text{ approximately.}$$

Kendall's τ

	1	2	3	4	5	6	7	8	9	10	11	12	13
X_i	277	169	157	139	108	213	232	229	114	232	161	149	128
Y_i	256	118	137	144	146	221	184	188	97	231	114	187	230

Calculate Kendall's τ.

```
x = c(277,169,157,139,108,213,232,229,114,232,161,149,128)
y = c(256,118,137,144,146,221,184,188,97,231,114,187,230)
n = 13
Qmat = matrix(0,n-1,n-1)
colnames(Qmat) = 1:(n-1)
rownames(Qmat) = 2:n
for(i in 1:(n-1)){
  for(j in (i+1):n){
    qval = (y[j]-y[i])*(x[j]-x[i])
    if(qval>0){
      Qmat[j-1,i] = 1
    } else if(qval<0){
      Qmat[j-1,i] = -1
    }
  }
}
Qmat
K=sum(Qmat)
tauhat=K/(n*(n-1)/2)
tauhat
z=K/(sqrt((2*(2*n+5))/(9*n*(n-1))))
2*pnorm(-abs(z))
```

```
> Qmat
   1  2  3  4  5  6  7  8 9 10 11 12
2  1  0  0  0  0  0  0  0 0  0  0  0
3  1 -1  0  0  0  0  0  0 0  0  0  0
4  1 -1 -1  0  0  0  0  0 0  0  0  0
5  1 -1 -1 -1  0  0  0  0 0  0  0  0
6  1  1  1  1  1  0  0  0 0  0  0  0
7  1  1  1  1  1 -1  0  0 0  0  0  0
8  1  1  1  1  1 -1 -1  0 0  0  0  0
9  1  1  1  1 -1  1  1  1 0  0  0  0
10 1  1  1  1  1  1  0  1 1  0  0  0
11 1  1 -1 -1 -1  1  1  1 1  1  0  0
12 1 -1 -1  1  1  1 -1  1 1  1 -1  0
13 1 -1 -1 -1  1 -1 -1 -1 1  1 -1 -1
> tauhat
[1] 0.3461538
> 2*pnorm(-abs(z))
[1] 0
```

$H_0: \tau = 0$

$H_1: \tau \neq 0$

The p-values = 0 shows strong evidence to reject the hypothesis of the independence of X and Y.

Prior and Posterior Distribution

Let $X_1, ..., X_n \sim \text{Uniform}(0, \theta)$. Let $f(\theta) \propto 1/\theta$. Find the posterior density.

$$f(x_1, ..., x_n | \theta) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}(\theta \geq x_{(n)})$$

$$= \theta^{-n} \mathbb{1}(\theta \geq x_{(n)})$$

$$\propto \theta^{-[(n-1)+1]} (n-1) x_{(n)}^{n-1} \mathbb{1}(x_{(n)} \leq \theta)$$

which is Pareto$(n-1, x_{(n)})$ distribution.

The posterior density of θ given $x_1, ..., x_n$ is

$$f(\theta | x) \propto f(x | \theta) \pi(\theta)$$

$$\propto \theta^{-[(n-1)+1]} (n-1) x_{(n)}^{n-1} \mathbb{1}(x_{(n)} \leq \theta) \theta^{-1}$$

$$= \theta^{-(n+1)} (n-1) x_{(n)}^{n-1} \mathbb{1}(x_{(n)} \leq \theta)$$

$$\propto \theta^{-(n+1)} n x_{(n)}^{n} \mathbb{1}(x_{(n)} \leq \theta)$$

which is Pareto$(n, x_{(n)})$ distribution.

Bayesian Point Estimation

Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$. Let $\lambda \sim \text{Gamma}(\alpha, \beta)$ be the prior.

Show that the posterior is also a Gamma. Find the posterior mean.

$$\pi(\lambda \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$f(x_1, \ldots, x_n \mid \lambda) = \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} e^{-n\lambda}$$

Thus, the posterior distribution is

$$\begin{aligned} f(\lambda \mid x_1, \ldots, x_n) &\propto \lambda^{\sum_{i=1}^n x_i} e^{-n\lambda} \lambda^{\alpha-1} e^{-\beta\lambda} \\ &= \lambda^{\sum_{i=1}^n x_i + \alpha - 1} e^{-n\lambda - \beta\lambda} \\ &= \lambda^{\sum_{i=1}^n x_i + \alpha - 1} e^{-(n+\beta)\lambda} \end{aligned}$$

which is Gamma $\left(\sum_{i=1}^n x_i + \alpha,\ n+\beta\right)$ distribution

The posterior mean $E[\lambda \mid x_1, \ldots, x_n] = \dfrac{\sum_{i=1}^n x_i + \alpha}{n+\beta}$.

If the squared-error loss function is used, this posterior mean is the Bayes estimate.

Questions