MA677_Final

Sky Liu

5/7/2019

Statistics and the Law

```
# read data
acorn<-read.csv("acorn.csv")
#calculate effect size
ef<-cohen.d(acorn$MIN,acorn$WHITE)
ef$estimate
```

```
## [1] 1.98
```

```
# n = 20
pwr.t.test(
n = dim(acorn)[1], ef$estimate,
sig.level = 0.05, power = NULL, type = c("two.sample")
)
```

```
##
##      Two-sample t test power calculation
##
##               n = 20
##               d = 1.98
##       sig.level = 0.05
##           power = 1
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
# when power = 0.95, n = ?
pwr.t.test(
  n = NULL, ef$estimate,
  sig.level = 0.05, power = 0.95, type = c("two.sample")
  )
```

```
##
##      Two-sample t test power calculation
##
##               n = 7.75
##               d = 1.98
##       sig.level = 0.05
##           power = 0.95
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
# t-test
t.test(acorn$MIN, acorn$WHITE)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  acorn$MIN and acorn$WHITE
## t = 6, df = 30, p-value = 0.0000006
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  14.3 28.2
## sample estimates:
## mean of x mean of y
##      36.9      15.6
```

We used two sample t-test to explore the discrimination to warrant corrective action. The power of the t-test is 1 and the p value is small enough to reject the null hypothesis of no discrimination.

To verify if the data has enough evidence, we found that when n >= 8, the power of t-test would be greater than 0.95. Therefore, given we have n = 20 in this data, it is a sufficient evidence.

Comparing Suppliers

```r
Dead_Bird=c(12,8,21);Display_Art=c(23,12,30);Flying_Art=c(89,62,119)
orithopter <- as.data.frame(cbind(Dead_Bird,Display_Art,Flying_Art))
rownames(orithopter) <- c("Area 51","BDV","Giffen")
chisq.test(orithopter)
```

```
##
##  Pearson's Chi-squared test
##
## data:  orithopter
## X-squared = 1, df = 4, p-value = 0.9
```

We use chi square test to test null hypothesis that all three schools produce the same quality ornithopters. The result of Chi square shows p-value equaling 0.9, which is larger than the siginificance level 0.05. Therefore, we reject the null hypothesis and conclude that not all three schools produce the same quality ornithopters.

How deadly are sharks?

```r
shark<-read.csv("sharkattack.csv")



#Fatal
USshark_fatal <- shark %>%
  filter(Country == "United States" ) %>%
  filter(Fatal == "Y" | Fatal == "N")%>%
  mutate(Fatal_code = ifelse(Fatal == "Y", 1, 0))

AUshark_fatal <- shark %>%
  filter(Country == "Australia" ) %>%
  filter(Fatal == "Y" | Fatal == "N")%>%
  mutate(Fatal_code = ifelse(Fatal == "Y", 1, 0))
```

```
#calculate effect size
ef<-cohen.d(AUshark_fatal$Fatal_code,USshark_fatal$Fatal_code)
ef$estimate
```

```
## [1] 0.432
```

```
# n = 20
pwr.t.test(
  n = dim(USshark_fatal)[1], ef$estimate,
  sig.level = 0.05, power = NULL, type = c("two.sample"),
  alternative = "greater"
  )
```

```
##
##      Two-sample t test power calculation
##
##               n = 2012
##               d = 0.432
##       sig.level = 0.05
##           power = 1
##     alternative = greater
##
## NOTE: n is number in *each* group
```

```
# proportion z-test
prop.test( x = c(
  sum(AUshark_fatal$Fatal_code == 1),
  sum(USshark_fatal$Fatal_code == 1) ),
n = c(dim(AUshark_fatal)[1], dim(USshark_fatal)[1]),
  alternative = "greater"
)
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(sum(AUshark_fatal$Fatal_code == 1), sum(USshark_fatal$Fatal_code ==  out of c(dim(AUshark_fa
## X-squared = 100, df = 1, p-value <0.0000000000000002
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.133 1.000
## sample estimates:
## prop 1 prop 2
##  0.266  0.108
```

By calculating the effect size and power analysis, we obtained the power being 1. Then, by performing a two sample test for null hypothesis that AU sharks are less or equally fatal than US shark. We obtained a p-value small enough to reject the null hypothesis.

Power Analysis

As discussed in Cohen's *Statistical Power Analysis for the Behavioral Sciences*, the equal difference between two P value can be detected with different power, thus P cannot provide a scale of equal units of detectability. Therefore, the difference between P's is not an appropriate index of effect size.

Arcsine transformation of P value solves this problem. By taking $\phi = 2arcsine\sqrt{P}$, equal difference between ϕs can be detected. Unlike, $P_1 - P_2$, the value of $h = \phi_1 - \phi_2$ does not depend of value of ϕs and where it falls in its possibility range. Therefore, we can use to represent effect size index for a difference in proportion.

Estimators

Exponential

$$f(x_i;\lambda) = \lambda e^{-\lambda}x$$

mm.

$$E(x) = \int_0^\infty x\lambda e^{-\lambda x}\,dx$$

$$= \lambda\int_0^\infty x e^{-\lambda x}\,dx$$

$$= \frac{1}{\lambda}$$

$$\Rightarrow \hat{x} = \frac{1}{\lambda} \quad \Rightarrow \hat{\lambda} = \frac{1}{\bar{x}}$$

MLE

$$\mathcal{L}(\lambda; x_1 \dots x_n) = f(x_1)f(x_2)\dots f(x_n)$$

$$= \lambda^n e^{-\lambda \Sigma x_i}$$

Taking log : $\ell(\lambda; x_1 \dots x_n)$

$$= n\log(\lambda) - \lambda\Sigma x_i$$

Taking derivative :

$$\frac{d\ell}{d\lambda} = \frac{n}{\lambda} - \Sigma x_i = 0$$

$$\Rightarrow \frac{n}{\lambda} = \Sigma x_i$$

$$\Rightarrow \hat{\lambda} = \frac{n}{\Sigma x_i} = \frac{1}{\bar{x}}$$

A new distribution

$$f(x) = \begin{cases} (1-\theta) + 2\theta x & 0 < x < 1 \\ 0 & o.w. \end{cases}$$

MM

$$E(x) = \int_0^1 x(1-\theta)+2\theta x)\,dx$$

$$= (1-\theta)\int_0^1 x\,dx + \int_0^1 2\theta x^2\,dx$$

$$= \frac{1}{2} - \frac{1}{2}\theta + \frac{2}{3}\theta$$

$$= \frac{1}{2} + \frac{1}{6}\theta$$

$$\Rightarrow \quad \bar{x} = \frac{1}{2} + \frac{1}{6}\theta$$

$$\Rightarrow \hat{\theta} = 6\bar{x} - 3$$

MLE

$$L(\theta; x_1 \dots x_n) = \prod [(1-\theta) + 2\theta x_i]$$

Taking log $\quad \ell(\theta; x_1 \dots x_n)$

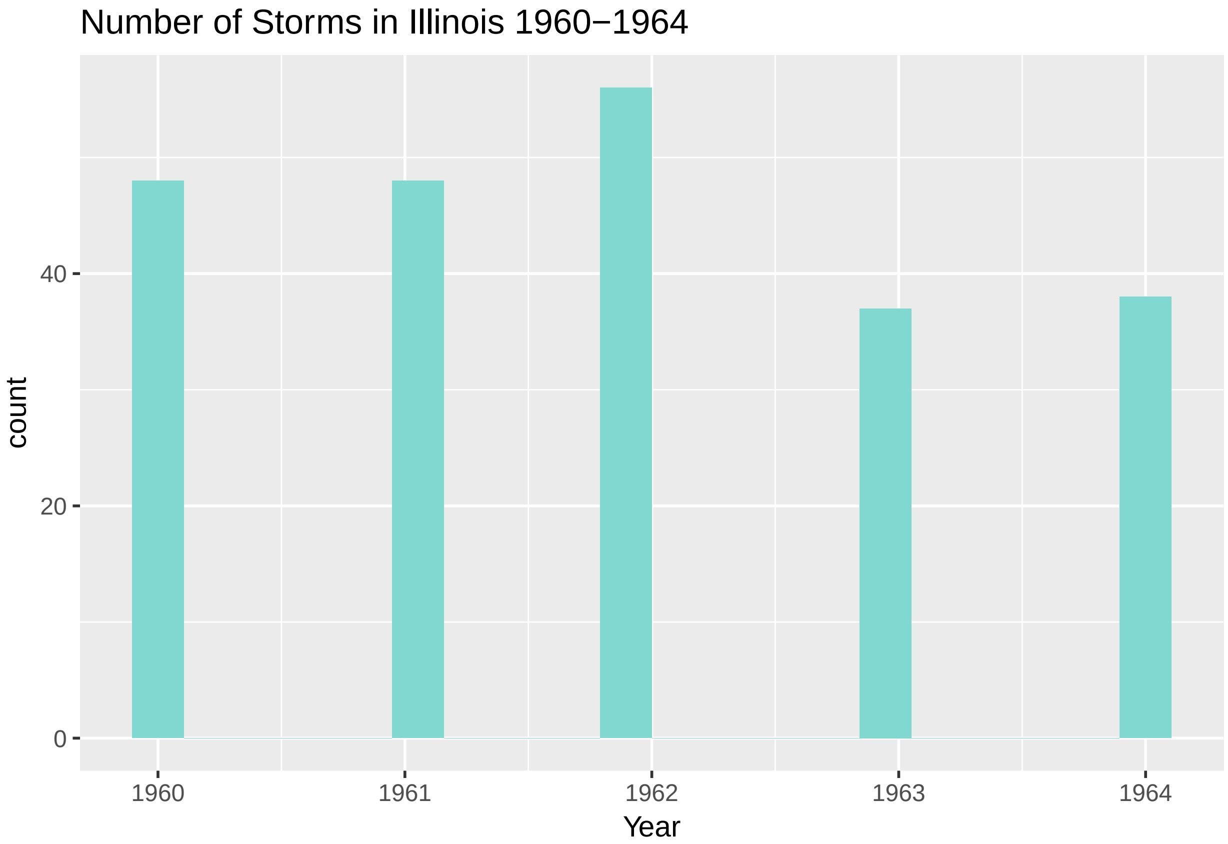
$$= \sum \ln[(1-\theta) + 2\theta x_i]$$

Taking derivative $\quad \frac{d\ell}{d\theta} = \sum \frac{2x_i - 1}{1-\theta+2\theta x_i} = 0$

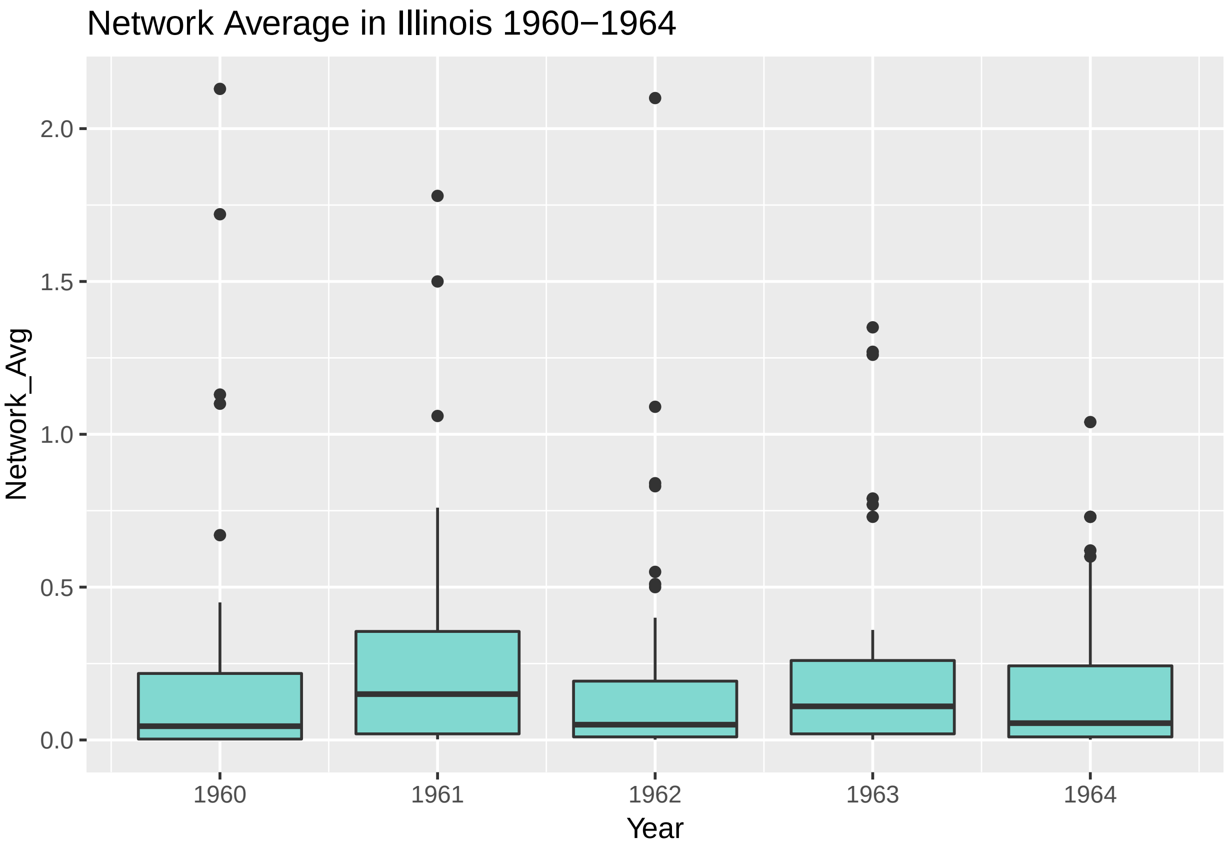
$\hat{\theta}$ = the solution of the equation above.

Rain in Southern Illinois

```
ggplot(data=ill,aes(x = Year)) +
  geom_histogram(bins = 20, fill = "#81D8D0")  +
  labs( title = "Number of Storms in Illinois 1960-1964"
)
```



```
ggplot(data=ill,aes(x = Year,y=Network_Avg,group=Year)) +
  geom_boxplot(fill = "#81D8D0")+
  labs( title = "Network Average in Illinois 1960-1964"
)
```



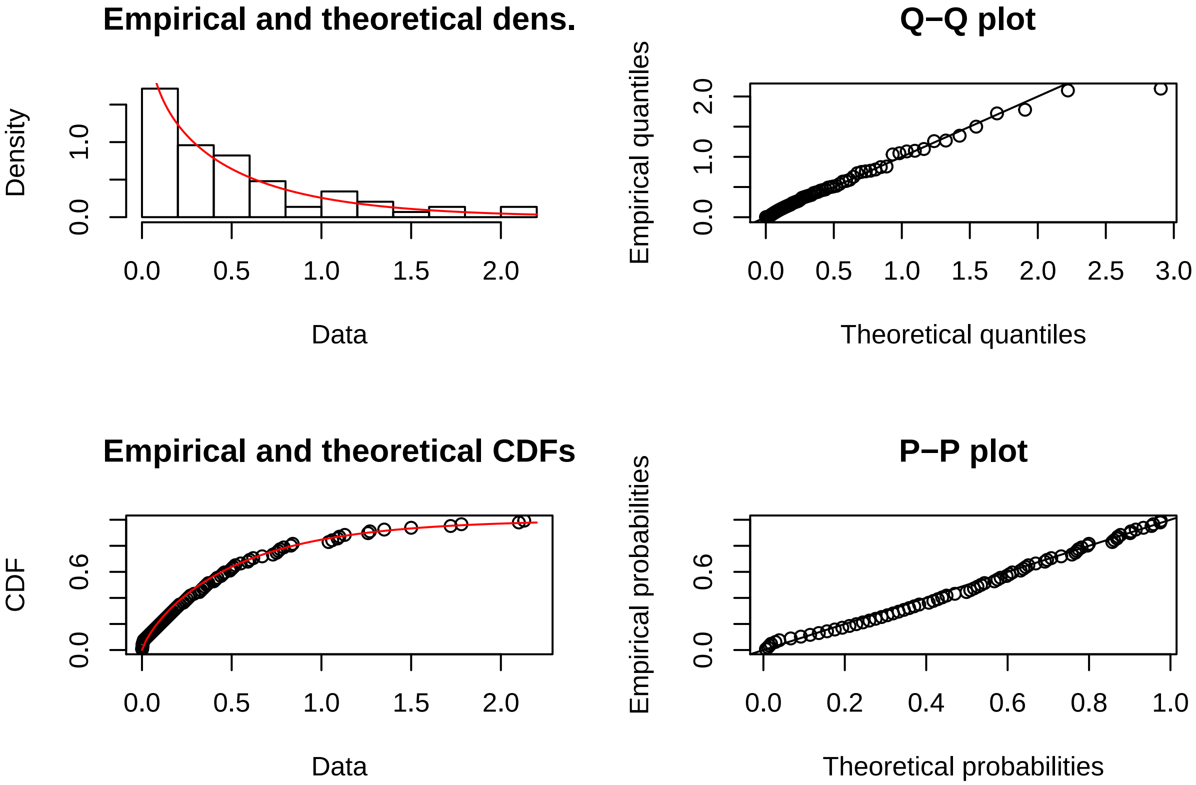
From the two plots above we could see that 1962 had more storms than the other years while it's average raingage network was not high amoung the five years. 1961 was the year with the highest average raingage network.

```
ill_reshaped <- dcast(ill, Network_Avg ~ Year)
```

```
## Using 'Year' as value column. Use 'value.var' to override
```

```
## Aggregate function missing, defaulting to 'length'
```

```
fitgamma <- fitdist(ill_reshaped$Network_Avg, "gamma")
plot(fitgamma)
```



From the plots above, we found that gamma distribution is a good fit to rainfell data

```r
set.seed(20190507)
mm <- fitdist(ill_reshaped$Network_Avg, "gamma", method = "mme")
bs_mm <- bootdist(mm)
summary(bs_mm)
```

```
## Parametric bootstrap medians and 95% percentile CI
##       Median 2.5% 97.5%
## shape   1.11 0.71  1.60
## rate    2.19 1.35  3.29
```

```r
mle <- fitdist(ill_reshaped$Network_Avg, "gamma", method = "mle")
bs_mle <- bootdist(mle)
summary(bs_mle)
```

```
## Parametric bootstrap medians and 95% percentile CI
##       Median  2.5% 97.5%
## shape  0.799 0.619  1.08
## rate   1.584 1.085  2.33
```

The estimate using bootstrap through MM method has confidence interval (0.71,1.60).

The estimate using bootstrap through MLE method has narrower confidence interval (0.622,1.07).

Thus, I would prefer using MLE

Analysis of decision theory article

Let δ be the probability a patient being allocated with treatment B $1-\delta$ to treatment A. $\delta \in [0,1]$

Let y be a response function from treatments. to outcomes.

Let P denote the distribution of treatment responses.

We define a function

$$U(\delta, P) = E[y(A)](1-\delta) + E[y(B)]\delta$$

to express addictive welfare.

The goal is to maximize U by selecting δ.

Let $a = E[y(A)]$, $b = E[y(B)]$ be mean outcomes, we have

$$U(\delta, P) = a(1-\delta) + b \cdot \delta = a + (\beta - a)\delta.$$

Let Q be samplig distribution, Ψ as sample space. we obtain

$$U(\delta, P, \psi) = a + (\beta - a)\delta(\psi)$$

The expected value in state s is

$$W(\delta, \beta_s, Q_s) = a_s + (\beta_s - a_s) E_s[\delta(\psi)]$$

where $E_s[\delta(\psi)] = \int_\Psi \delta(\psi) dQ_s(\psi)$

n is distributed Binomal $B[\beta, N]$. Thus,

$$E[\delta(n)] = \sum_0^N \delta(i) f(n=i, \beta, N)$$

where $f(n=i, \beta, N)$

$= N![i!(N-i)!]^{-1} \beta^i (1-\beta)^{N-i}$

is the Prob(sussess)

Hence, δ is admissible iff

$$\begin{aligned} \delta(n) &= 0 & n &< n_0 \\ \delta(n) &= \lambda & n &= n_0 \\ \delta(n) &= 1 & n &> n_0 \end{aligned}$$

for some $n_0 \in [0, N]$. $\lambda \in [0,1]$

Let $(\beta_s, s \in S) = (0,1)$ with parameter (c,d)

We obtain posterior mean for β as $\frac{c+n}{c+d+N}$

By Bayes rule, we obtain

$$\begin{aligned} \delta(n) &= 0 & \frac{c+n}{c+d+N} &< a \\ \delta(n) &= \lambda & \frac{c+n}{c+d+N} &= a \\ \delta(n) &= 1 & \frac{c+n}{c+d+N} &= a \end{aligned}$$