# Lecture 4: Statistical functionals and Influence functions

Pratheepa Jeganathan

04/10/2019

# Robustness

# Properties of estimators

- Measures of robustness
  - efficiency
  - influence
  - breakdown
- Asymptotic relative efficiency **HWC** Chapter 3.11
- Consider influence and breakdown

# Sensitivity to gross errors

- Sensitivity curve: function of observations.
- Let $\boldsymbol{z}_n = (z_1, \cdots, z_n)^T$ drawn from cdf $F$ and $\theta$ is the location parameter.
- Let $\hat{\theta}$ is an estimator of $\theta$.
- Add an outlier observation $z$ to $\boldsymbol{z}_n$, $\boldsymbol{z}_{n+1} = (z_1, \cdots, z_n, z)^T$.
- The sensitivity curve of an estimator $\hat{\theta}$ is

$$S\left(z; \hat{\theta}\right) = \frac{\hat{\theta}_{n+1} - \hat{\theta}_n}{1/(n+1)}. \tag{1}$$

# Sensitivity to gross errors (examples)

```r
z_n = c(1.85, 2.35, -3.85, -5.25, -0.15,
  2.15, 0.15, -0.25, -0.55, 2.65)
mean(z_n)
```

```
## [1] -0.09
```

```r
median(z_n)
```

```
## [1] 0
```

```r
library(ICSNP)
hl.loc(z_n)
```

```
## [1] 0
```

# Example (Sensitivity curve for mean)

```
z_n_plus_1_df = data.frame(z_n_plus_1 = seq(-20, 20,
  by = 1))

sensitivity <- function(theta_n_plus_1, theta_n, n){
  (theta_n_plus_1- theta_n)*(n+1)
}

mean_z_n_plus_1 = apply(z_n_plus_1_df, 1, function(x){
  x = c(z_n,x)
  mean(x)
})

sensitivity_mean = sensitivity(mean_z_n_plus_1,
  mean(z_n), length(z_n))
```

# Example (Sensitivity curve for median)

```r
median_z_n_plus_1 = apply(z_n_plus_1_df, 1, function(x){
  x = c(z_n,x)
  median(x)
})

sensitivity_median = sensitivity(median_z_n_plus_1,
  mean(z_n), length(z_n))
```

# Example (Sensitivity curve for HL)

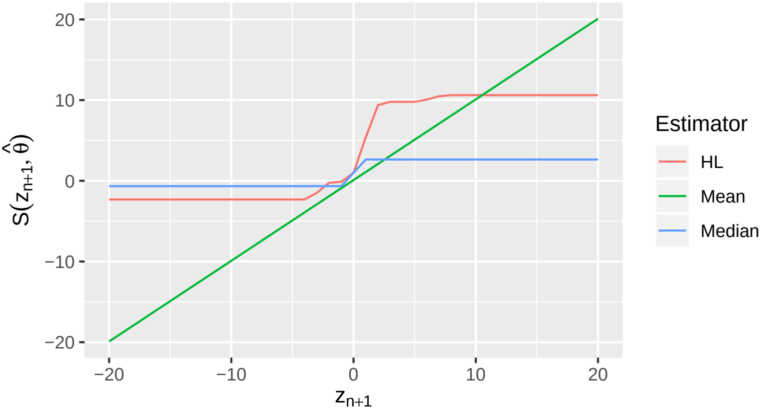
```r
HL_z_n_plus_1 = apply(z_n_plus_1_df, 1, function(x){
  x = c(z_n,x)
  hl.loc(x)
})
sensitivity_HL = sensitivity(HL_z_n_plus_1,
  mean(z_n), length(z_n))
```

# Example (Sensitivity curve)

```
library(tidyr)
library(ggplot2)
df = data.frame(z_n_plus_1 = z_n_plus_1_df$z_n_plus_1,
  sensitivity_mean = sensitivity_mean,
  sensitivity_median = sensitivity_median,
  sensitivity_HL = sensitivity_HL)
df_long = gather(df, key = "estimator",
  value = "value", -z_n_plus_1)
df_long$estimator = factor(df_long$estimator)
ggplot(data = df_long) +
  geom_line(aes(x = z_n_plus_1,
    y = value, group =estimator, color = estimator)) +
  xlab(bquote(z[n+1]))+
  scale_color_discrete(name = "Estimator",
    labels = c("HL", "Mean", "Median")) + ylab(bquote(S(z[n
```

# Example (Sensitivity curve)

- Mean: unbounded.
- Median and Hodges–Lehmann: bounded.



# Statistical functionals

- Statistical inference involves estimating some aspects of a cdf $F$ on the basis of a random sample drawn from $F$.
- Statistical functional $T(F)$: any function of $F$.
  - Let $Z_1, \cdots, Z_n \sim F$, where $F(z) = P(Z \leq z)$, define $\theta = T(F)$.
- Examples:
  - Mean: $T(F) = \int z dF(z)$.
  - Median: $T(F) = F^{-1}(1/2)$.
  - HL: $T(F) = (1/2)\{F * F\}^{-1}(1/2)$, where $*$ denotes convolution.

# Estimating statistical functionals

- Estimator of $F$: empirical CDF $\hat{F}(z) = \frac{\#\{z_i \leq z\}}{n}$.
- Plug-in principal: plug-in estimator of $T(F)$ is $T\left(\hat{F}\right)$ - (summary statistic).
- Plug-in principal is good when there is information about $F$ only through sample $\boldsymbol{z}$ (not from the model).

# Influence functions

- Influence function
  - Measures rate of change of $T(F)$ under small contamination at $z$ (kind of derivative).
  - Indicates statistical accuracy of a statistic (if influence function is bounded - robustness).
  - Useful for computing the approximate standard error of plug-in estimate $T\left(\hat{F}\right)$ (standard deviation of a summary statistic).
- Gateaux derivative of $T$ at $F$ in the direction $G$

$$L(G) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon G) - T(F)}{\epsilon} \tag{2}$$

# Influence functions

- If $G = \delta_z$ is a point mass at $z$

$$L(z) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon\delta_z) - T(F)}{\epsilon}. \tag{3}$$

- $L(z)$ is the influence function.
- Empirical influence function/plug-in estimator for $L(z)$

$$\hat{L}(z) = \lim_{\epsilon \to 0} \frac{T\left((1-\epsilon)\hat{F} + \epsilon\delta_z\right) - T\left(\hat{F}\right)}{\epsilon}. \tag{4}$$

# Examples (influence functions)

- The influence function for our estimators are (up to constant of proportionality and center)
  - Mean: $z$
  - Median: $\operatorname{sign}(z)$
  - HL: $F(z) - .5$
- Mean is not robust, but median and HL are robust

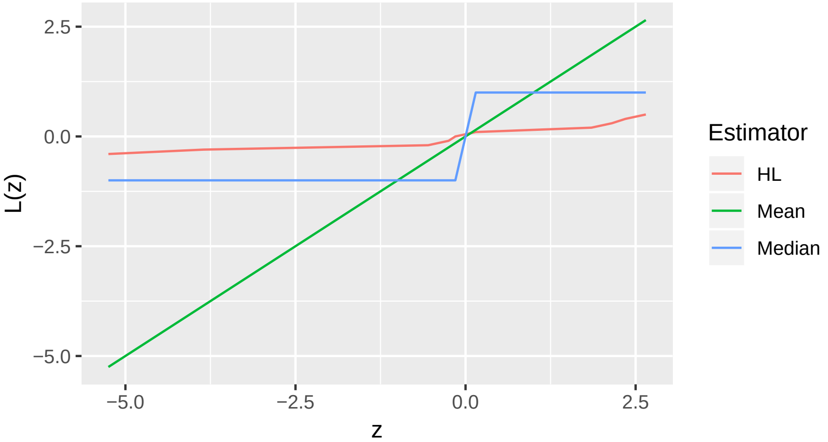
# Example (influence curves)

```
influence_mean = z_n
influence_median = sign(z_n)
z_n_df = data.frame(z_n = z_n)
influence_HL = apply(z_n_df, 1, function(x){
  mean(z_n <= x) -.5
})
```

# Example (influence curves)

```
df_inf = data.frame(z = z_n,
  influence_mean = influence_mean,
  influence_median = influence_median,
  influence_HL = influence_HL)
df_inf_long = gather(df_inf, key = "estimator",
  value = "value", -z)
df_inf_long$estimator = factor(df_inf_long$estimator)
ggplot(data = df_inf_long) +
  geom_line(aes(x = z,
    y = value, group =estimator, color = estimator)) +
  xlab("z")+
  scale_color_discrete(name = "Estimator",
    labels = c("HL", "Mean", "Median")) +
  ylab("L(z)")
```

# Example (influence curves)



# Standard error of a plug-in estimator

- If $T(F) = \int a(z)\, dF(z)$, a linear functional
  - $L(z) = a(z) - T(F)$.
  - $\mathbb{E}(L(z)) = 0$.
  - $\tau^2 = \int L(z)^2\, dF(z) = \int (a(z) - T(F))^2\, dF(z)$.
  - $\hat{\tau}^2 = \frac{1}{n}\sum_{i=1}^n \left(a(Z_i) - T\left(\hat{F}\right)\right)^2$.
  - $\text{se}^2\left(T\left(\hat{F}\right)\right) = \frac{\hat{\tau}^2}{n}$.

# Example (Standard error of a plug-in estimator)

- $\theta = T\left(F\right) = \int z dF\left(z\right)$.
- $\hat{\theta} = T\left(\hat{F}\right) = \int z d\hat{F}\left(z\right) = \frac{\sum_{i=1}^{n} Z_i}{n} = \bar{Z}$.
- $L\left(z\right) = z - \int z dF\left(z\right)$.
- $\hat{L}\left(z\right) = z - \bar{Z}$.
- $\mathrm{se}^2\left(T\left(\hat{F}\right)\right) = \frac{n^{-1}\sum_{i=1}^{n}\left(Z_i - \bar{Z}\right)^2}{n}$.

# Breakdown point of an estimator

- Reference: Following notes from this link.
- Suppose we contaminate $n - m$ points in our sample

$$z_n^* = \left(z_1, \cdots, z_m, z_{m+1}^*, \cdots, z_n^*\right)$$

.

- Consider $z_{m+1}^*, \cdots, z_n^*$ are very large (close to $\infty$).
- Breakdown point: the smallest value $n - m$ so that $\hat{\theta}\left(z_n^*\right)$ is bad.
- Finite sample breakdown point : a function of sample size $\dfrac{n-m}{n}$.
- Asymptotic breakdown point: (single number) the limit of the finite sample breakdown point as $n \to \infty$.

# Examples (breakdown point)

- Sample mean
  - Finite sample BP: $\frac{1}{n}$.
  - Asymptotic BP: 0.
- Sample median
  - Finite sample BP: $\left[\frac{n-1}{2n}\right]$, $[u]$ is the largest integer less than or equal to $u$.
  - Asymptotic BP: .5.
- HL
  - Finite sample BP: Read this notes.
  - Asymptotic BP: $\approx 0.29$.

# References

# References for this lecture

**W** Chapter 2 (Statistical functionals and influence functions)

**ET** Chapter 4, 5, 21.3 (Statistical functionals and influence functions)

**KM** Chapter 3.5 (R codes for sensitivity, breakdown, influence)

**HWC** Chapter 3.2 page 57, comment 16 (sensitivity to gross errors-HL)

**HWC** Chapter 3.5 page 77, comment 40 (sensitivity to gross errors-median)