What to do today (Nov 23, 2020)?

1. *Introduction*
2. *Probability and Distribution (Chp 1-3)*
3. *Essential Topics in Mathematical Statistics*

   3.1 Elementary Statistical Inferences (Chp 4)

   3.2 Consistency and Limiting Distributions (Chp 5)

   3.3 Maximum Likelihood Methods (Chp 6)

4. **Further Topics, Selected from Chp 7-11**
   - **4.1 Nonparametric and Robust Statistics (Chp 10)**
     - **4.1.1 Location Models**
     - **4.1.2 Sample Median and the Sign Test**
     - **4.1.3 Signed-Rank Test and Mann-Whitney-Wilcoxon Test**
     - *4.1.4 Measures of Association*
     - *4.1.5 Robust Concepts*
   - *4.2 Bayesian Procedures (Chp 11)*

4.1 Nonparametric and Robust Statistics (Chp 10)

Why to study nonparametrics? Why to study robust statistics?

Recall most statistical methods studied so far .. ...

- Specifying r.v. $X \sim f(\cdot;\theta)$
  - estimating θ,
  - testing on hypotheses about θ
- Specifying r.v.s. $X \sim f(\cdot;\theta)$, $Y \sim g(\cdot;\phi)$
  - estimating θ and ϕ, testing on hypotheses about θ and ϕ

What if $f(\cdot;\theta)$ and/or $g(\cdot;\phi)$ can not be confidently specified?
e.g., in medical settings, to play "safe"!
$\Longrightarrow$

- any statistical methods robust to the model assumption?
- any statistical methods not requiring to specify the population distribution(s), distribution-free procedures?

Have we ever studied anything like that?

4.1 Nonparametric and Robust Statistics (Chp 10)

Review 1.A Summary Statistics: order statistics Definition. Suppose $X_1, \ldots, X_n$ are iid observations from a continuous r.v. $X \sim f(\cdot)$ with cdf $F(\cdot)$. The **order statistics** of the random sample are $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$: $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$.

$X_{(1)} =$ the smallest value of $X_1, \ldots, X_n$,

$X_{(2)} =$ the 2nd smallest value of $X_1, \ldots, X_n$, ...,

$X_{(n)} =$ the largest value of $X_1, \ldots, X_n$.

Distribution. $X_{(k)} \sim \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1}(1-F(x))^{n-k} f(x)$ for $k = 1, \ldots, n$.

Example 10.1 Realizations of 5 iid observations $X_1, \ldots, X_5$ from a population are given in the table below.

x_1	x_2	x_3	x_4	x_5
0.62	0.98	0.31	0.81	0.53

The order statistics?

Review 1.B Summary Statistics: rank statistics

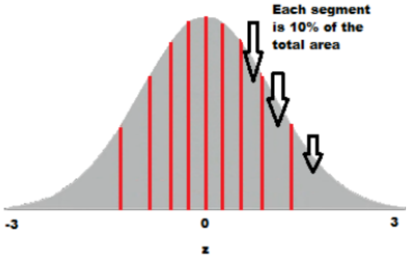
Definition. The rank of X_k, the kth observation in a random sample of size n, is r_k such that $X_k = X_{(r_k)}$, for $k = 1, \ldots, n$.

Example 10.1 (cont'd)

obs	x_1	x_2	x_3	x_4	x_5
	0.62	0.98	0.31	0.81	0.53
order stat. $x_{(r_k)}$	$x_{(3)}$	$x_{(5)}$	$x_{(1)}$	$x_{(4)}$	$x_{(2)}$
rank stat.					

Review 1.C Summary Statistics: percentiles/quantiles.

Definition. Suppose r.v. $X \sim f(\cdot)$ with a random sample $X_1, \ldots, X_n$.
(i) **Population percentitles:** π_p is the $(100p)$th percentile of the population if $P(X \leq \pi_p) = p$. That is, $\int_{-\infty}^{\pi_p} f(x)dx = p$.



(ii) **Sample percentiles:** Let $X_{(1)}, \ldots, X_{(n)}$ be the order statistics. Then $X_{(r)}$ is the $(r/n)100$th (or $(r/n+1)100$th) sample percentile.

e.g., If $p = 0.5$, the population median m is the $(100p)$th population percentile.
The order statistic $X_{(n+1/2)}$ is the sample median when n is odd; all values in between $X_{(n/2)}$ and $X_{(n/2+1)}$ are the sample median when n is even.

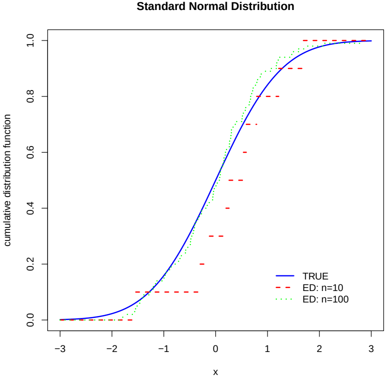
Review 2. Empirical Distribution Function:

Definition. Suppose r.v. $X \sim F(\cdot)$ with a random sample $X_1, \ldots, X_n$. Its **empirical distribution** is defined as

$$\hat{F}_n(x) = \frac{1}{n}\#\{X_i : X_i \leq x; i = 1, \ldots, n\}, \quad x \in (-\infty, \infty).$$

That is, $\hat{F}_n(x) = 0$ if $x < X_{(1)}$; k/n, if $X_{(k)} \leq x < X_{(k+1)}$ when $1 \leq k \leq n-1$; 1, if $x \geq X_{(n)}$.

- $E[\hat{F}_n(x)] = F(x)$, $Var[\hat{F}_n(x)] = F(x)[1 - F(x)]/n$ for a fixed x



4.1.1 Nonparametric and Robust Statistics: Location Models

Definition. Let X be a r.v. with cdf $F_X(\cdot)$. We call the functional $T(\cdot)$ a **location functional** if it satisfies the following:

- If $Y = X + a$ for $-\infty < a < \infty$, the r.v. Y's cdf $F_Y(\cdot)$ satisfies $T(F_Y) = T(F_X) + a$;
- if $Y = aX$ for $-\infty < a < \infty$, the r.v. Y's cdf $F_Y(\cdot)$ satisfies $T(F_Y) = aT(F_X)$.

Definition. The distribution of r.v. X is a location model if there is a location functional $T(\cdot)$, and $X = \theta_X + \epsilon$ with $\theta_X = T(F_X)$ and $T(F_\epsilon) = 0$.

A location model depends very much on the functional.

- Let $\epsilon \sim F(\cdot)$ such that $F(0) = 1/2$. If $X = \theta + \epsilon$ with $-\infty < \theta < \infty$, X follows the location model with the locational functional $T(F_X) = \theta$.

- If X is a continuous r.v. following a location model $X = \theta_X + \epsilon$ with pdf $f_X(\cdot)$, $f_X(x) = f(x - \theta_X)$ with $f(\cdot)$ the pdf of ϵ.

- If the distribution of r.v. X is symmetric about a, for any location functional $T(\cdot)$, $T(F_X) = a$.

Proof:

4.1.2 Sample Median and the Sign Test

Let $\{X_1, \ldots, X_n\}$ be a random sample following the location model: $X_i = \theta + \epsilon_i$, with $\epsilon_i \sim F(\cdot)$ i.i.d. and median 0.

- Test on $H_0 : \theta = \theta_0$ vs $H_1 : \theta > \theta_0$ at the significance level of α.
  - The location functional $T(F_X) = \theta$ is the median of $X_1, \ldots, X_n$.

Consider the **sign statistic** $S(\theta_0) = \#\{i : X_i > \theta_0\} = \sum_{i=1}^{n} sgn(X_i - \theta_0)$ with $sgn(X_i - \theta) = \mathrm{I}(X_i > \theta_0)$: under H_0,

$$S(\theta_0) \sim B(n, 1/2).$$

Reject H_0 if $S(\theta_0) \geq c$ with c the upper α quantile of $B(n, 1/2)$, i.e. $P_{H_0}\big(S(\theta_0) \geq c\big) \leq \alpha$.

Let $\{X_1, \ldots, X_n\}$ be a random sample following the location model: $X_i = \theta + \epsilon_i$, with $\epsilon_i \sim F(\cdot)$ i.i.d. and median 0.

- Test on $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ at the significance level of α.

Consider the **sign statistic** $S(\theta_0) = \#\{i : X_i > \theta_0\} = \sum_{i=1}^{n} sgn(X_i - \theta_0)$: under H_0,

$$S(\theta_0) \sim B(n, 1/2).$$

Reject H_0 if either $S(\theta_0) \geq c_1$ or $S(\theta_0) \leq c_2$ with c_1 and c_2 the upper and lower $\alpha/2$ quantile of $B(n, 1/2)$, respectively, i.e. $P_{H_0}\big(c_2 < S(\theta_0) < c_1\big) = 1 - \alpha$.

Remarks:

- The sign test is **distribution free**.
- If $n >> 1$, $\big[S(\theta_0) - (n/2)\big]\big/\sqrt{n/4} \sim N(0, 1)$ approximately.

Example 10.2

- *Study.* a type of steel beam with a compressive strength $\geq$ 50K lb/in^2?
- *Data.* $n = 25$ beams (observations). (Assume they're iid.) Among them, there were 6 beams with strength greater than $\geq$ 50K lb/in^2.
- *Hypotheses.* $H_0 : m = 50K$ vs $H_1 : m < 50K$
- *by the Sign Test.*

  (i) [the exact approach]
  $S = \sum_{i=1}^{25} S_i \sim B(25, 1/2)$ under H_0. From Table A.1, the critical value c with $\alpha = 0.01$ for $P_{H_0}(S < c) = 0.01$ is between 6 and 7.
  Since $S_{obs} = 6$, $\Longrightarrow$ inconclusive.
  (ii) [the approximate approach]

$$Z = \frac{\sum_{i=1}^{25} S_i - \frac{25}{2}}{\sqrt{25/4}}$$

  approximately under H_0. $Z_{obs} = -2.6 < -z_{0.01} = -2.33$
  $\Longrightarrow$ reject H_0.

4.1.2 Sample Median and the Sign Test

Suppose $X_1, \ldots, X_n$ are i.i.d. with median θ. Denote the sign test statistic by $S(\theta)$.

Estimation of population median θ based on the sign statistic

- **Point estimator**

Note that $\bar{X}_n = argmin \sqrt{\sum_{i=1}^{n}(X_i - \theta)^2}$. (via the Euclidean distance, i.e. the L_2-distance)

What is $\hat{\theta} = argmin \sum_{i=1}^{n} |X_i - \theta|$? (via the L_1-distance)

$\hat{\theta}$ is the solution of

$$\frac{\partial}{\partial\theta} \sum_{i=1}^{n} |X_i - \theta| = -\sum_{i=1}^{n} sgn(X_i - \theta) = 0,$$

equivalent to $S(\theta) - n/2 = 0$.

- **Confidence interval**

Choose c such that $P_\theta\big(S(\theta) \le c\big) = \alpha/2$, and thus

$$P_\theta\big(c < S(\theta) < n - c\big) = 1 - \alpha.$$

$\Longrightarrow \big\{\phi : c < S(\phi) < n - c\big\}$ is a $1 - \alpha$ CI of θ.

If $n >> 1$, $c \approx \frac{1}{2}\big[n - 1 - \sqrt{n} z_{1-\alpha/2}\big]$, where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of $N(0,1)$.

4.1.3 Signed-Rank Test and MWW Test

Let $\{X_1, \ldots, X_n\}$ be a random sample following the location model: $X_i = \theta + \epsilon_i$, with $\epsilon_i \sim F(\cdot)$ i.i.d. and median 0.
Goal: Test $H_0 : \theta = 0$ vs $H_1 : \theta > 0$ at the significance level of α.

- **Signed-Rank Test.**

$$T = \sum_{i=1}^{n} sgn(X_i) R|X_i|$$

with $R|X_i|$ the rank of $|X_i|$ among $|X_1|, \ldots, |X_n|$.
Reject H_0 if $T \geq c$ with c determined by $P_{H_0}(T \geq c) = \alpha$.

Under H_0, (i) T's distribution is symmetric and determined with $E(T) = 0$ and $Var(T) = n(n+1)(2n+1)/6$, and (ii) $T/\sqrt{Var(T)} \to N(0,1)$ in distribution as $n \to \infty$.

Example 10.2 (cont'd) *by the Signed-Rank Test.*

(i) [the exact approach]
$T = \sum_{i=1}^{25} sgn(X_i - 50)R|X_i - 50|$ with $n = 25$ – in principle, the critical value can be determined using the distribution table.

(ii) [the approximate approach]

$$Z = \frac{T - \frac{25(25+1)}{4}}{\sqrt{25(26)(50+1)/24}} \sim N(0,1)$$

approximately under H_0. $Z_{obs} = -2.78 < -z_{0.01} = -2.33$
$\Longrightarrow$ Reject H_0.

What will we study next?

1. *Introduction*

2. *Probability and Distribution (Chp 1-3)*

3. *Essential Topics in Mathematical Statistics (Chp 4-6)*

4. **Further Topics, Selected from Chp 7-11**
   - **4.1 Nonparametric and Robust Statistics (Chp 10)**
     - *4.1.1 Location Models*
     - *4.1.2 Sample Median and the Sign Test*
     - **4.1.3 Signed-Rank Test and Mann-Whitney-Wilcoxon Test**
     - **4.1.4 Measures of Association**
     - **4.1.5 Robust Concepts**

   - *4.2 Bayesian Procedures (Chp 11)*