# What to do today (Nov 25, 2020)?

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

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 and Mann-Whitney-Wilcoxon Tests**
     - **4.1.4 Measures of Association**
     - **4.1.5 Robust Concepts**

   - *4.2 Bayesian Procedures (Chp 11)*

# 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 $\alpha$.
*Is the sign test efficient?*

- **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]

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

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

- **Mann-Whitney-Wilcoxon Test**.

Suppose r.v. $X \sim F(\cdot)$ and r.v. $Y \sim G(\cdot)$. Consider two-sample problem: testing on whether $X$ is **stochastically smaller** than $Y$ at the significance level of $\alpha$?

What if to consider $Y = X + \Delta$ and then $G(y) = F(y - \Delta)$?
$\Longrightarrow H_0 : \Delta = 0$ vs $H_1 : \Delta > 0$ with $\Delta = T(F_Y) - T(F_X)$.

Suppose $\{X_1, \ldots, X_{n_X}\}$ and $\{Y_1, \ldots, Y_{n_Y}\}$ are random samples from the two populations.
Consider $W = \sum_{i=1}^{n_Y} R(Y_i)$ with $R(Y_i)$ the rank of $Y_i$ in the combined sample $\{X_1, \ldots, X_{n_X}, Y_1, \ldots, Y_{n_Y}\}$ of size $n = n_X + n_Y$.

- Under $H_0$, $W$'s distribution free with a symmetric pmf and $E(W) = n_Y(n+1)/2$ and $Var(W) = n_X n_Y (n+1)/12$. Reject $H_0$ if $W > c$ with $c$ satisfying $P_{H_0}(W > c) \leq \alpha$.
- If $n >> 1$, approximately $\frac{W - n_Y(n+1)/2}{\sqrt{Var(W)}} \sim N(0,1)$.
- What if $H_1 : \Delta \neq 0$?

# 4.1.4 Measures of Association (Chp10.8)

Consider r.v.s. $X$ and $Y$ with joint distribution $F_{XY}(x, y)$.

**Goal.** to understand the association between $X$ and $Y$. *the strength of the association?*
e.g. **correlation coefficient** $\rho = Cov(X, Y)/\sqrt{Var(X)Var(Y)} =?$
Estimate $\rho$? Test on $H_0 : \rho = 0$?

Suppose $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are iid observations on $(X, Y)$: **sample correlation coefficient** (Pearson correlation coefficient)

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

*alternative measure of association?*

# 4.1.4 Measures of Association: Kendall's $\tau$

a measure of the similarity of $X$ and $Y$ in trend of taking values (the *monotonicity*)?

**Definition.** Let $(X_1, Y_1)$ and $(X_2, Y_2)$ be iid observations on $(X, Y)$. These pairs are **concordant** or **discordant** if $sgn\{(X_1 - X_2)(Y_1 - Y_2)\} = 1$ or $sgn\{(X_1 - X_2)(Y_1 - Y_2)\} = -1$, respectively.

A measure of $X$ and $Y$'s increasing vs decreasing relationship: the **Kendall's** $\tau$:

$$\tau = P\big(sgn\{(X_1-X_2)(Y_1-Y_2)\} = 1\big) - P\big(sgn\{(X_1-X_2)(Y_1-Y_2)\} = -1\big).$$

**Proposition.** If $(X_1, Y_1)$ and $(X_2, Y_2)$ be iid observations on $(X, Y)$, which follows a continuous bivariate distribution. If $X$ and $Y$ are independent, $\tau = 0$.

- If $\tau = 0$, $X$ and $Y$ are not independent in general.

Suppose $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are iid observations on $(X, Y)$.

$$K = \frac{1}{\binom{n}{2}} \sum_{i<j} sgn\{(X_i - X_j)(Y_i - Y_j)\}$$

is an unbiased estimator of the Kendall's $\tau$.

**Proposition.** $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are iid observations on $(X, Y)$ with continuous cdf $F(x, y)$. If $X$ and $Y$ are independent, $K$ is distribution free with a symmetric pmf, and

$$E(K) = 0, \quad Var(K) = \frac{2}{9}\frac{2n+5}{n(n-1)}.$$

Plus $K/\sqrt{Var(K)} \sim N(0, 1)$ approximately if $n >> 1$.

**Example 10.3** (p633) Table 10.8.1 of the textbook presents the winning times of the 1500m race and the marathon at the Olympics beginning with 1896 throuh 1980. Are the two winning times indepednent? By *"cor.test(m1500,marathon,method="kendall", exact=T)" in R*:

$$K_{obs} = 0.695 \quad p - value = 3.319e - 06.$$

$\Longrightarrow$ strong evidence against the hypothesis of the independence.

## 4.1.4 Measures of Association: Spearman's Rho

Suppose $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are iid observations on continuous bivariate $(X, Y)$: **sample correlation coefficient** (Pearson correlation coefficient)

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}.$$

**Definition.** The statistic

$$r_S = \frac{\sum_{i=1}^n (R(X_i) - \frac{n+1}{2})(R(Y_i) - \frac{n+1}{2})}{n(n^2-1)/12}$$

is called the **Spearman's rho**, where $R(X_i)$ and $R(Y_i)$ are the ranks of $X_i$ and $Y_i$ among $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, respectively.

**Proposition.** If $X$ and $Y$ are independent, $r_S$ is distribution free, symmetrically distributed with $E(r_S) = 0$ and $Var(r_S) = 1/(n-1)$. Plus, if $n >> 1$, $r_S / \sqrt{Var(r_S)} \sim N(0, 1)$ approximately.

# 4.1.5 Robust Statistics Concepts (Chp10.9): sensitivity curve

Consider $X_1, \ldots, X_n$ iid observations on $X$, which follows the location model with the location parameter $\theta$: $X = \theta + \epsilon$ with $\epsilon \sim f(\cdot)$. That is, $f_X(x) = f(x - \theta)$.

**Definition.** If $\hat{\theta} = \hat{\theta}(\mathbf{X}_n)$ is an estimator of $\theta$ with $\mathbf{X}_n = (X_1, \ldots, X_n)$ for $n \geq 1$, we call

$$S_n(x; \hat{\theta}) = \frac{\hat{\theta}(\mathbf{x}_{n+1}) - \hat{\theta}(\mathbf{x}_n)}{1/(n+1)}$$

the **sensitivity curve** of $\hat{\theta}$, where $\mathbf{x}_{n+1} = (\mathbf{x}_n, x)$.

- The sample mean $\bar{X}_n$ is an estimator of $\theta$, the solution of $\sum_{i=1}^{n}(X_i - \theta) = 0$. It's sensitivity curve is $S_n(x; \bar{X}) = x - \bar{x}_n$. $\Longrightarrow$ the sample mean is quite sensitive to the size of outliers.
- The sample median $\hat{\theta}_n$ with odd $n$, for example, is $X_{([n+1]/2)}$. It's sensitivity curve $S_n(x; \hat{\theta})$ is bounded, not much sensitive to an outlier.

# 4.1.5 Robust Statistics Concepts (Chp10.9): influence function

**Definition.** A **point-mass contamination** of the cdf $F_X(\cdot)$ at point $x$ is

$$F_{X;x,\eta}(t) = (1-\eta)F_X(t) + \eta\Delta_x(t),$$

with $\Delta_x(t) = I(x \le t)$.

**Definition.** If r.v. $X$ follows a location model with the location parameter $\theta = T(F_x)$, we call the following function the **influence function** of the estimator $\hat{\theta}_n = T(\hat{F}_{X,n})$:

$$IF(x;\hat{\theta}) = \lim_{\eta\to 0}\frac{1}{\eta}\big[T(F_{X;x,\eta}) - T(F_X)\big].$$

**Definition.** An estimator $\hat{\theta}$ is **robust** if $|IF(x;\hat{\theta})|$ is bounded $\forall x$.

## 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)*
   - **4.2 Bayesian Procedures (Chp 11)**
     - **4.2.1 Prior and Posterior Distributions**
     - **4.2.2 Bayesian Point Estimation**
     - **4.2.3 Bayesian Interval Estimation**
     - **4.2.4 Bayesian Testing Procedures**
     - *4.2.5 Additional Topics in Bayesian Statistics*