Question 5.

cf.

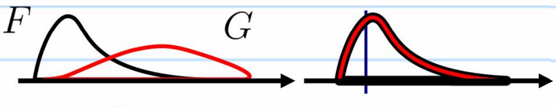
- In the materials taught before, we usually assume, in the statistical modeling, that the data follows a *particular* joint distribution which contains some unknown parameters of *finite* dimension. — e.g., normal with 3 parameters μ_X, μ_Y, σ^2
- The statistical inferences, estimation and testing, are then based on a formulation of these parameters. → e.g., $\Delta = \mu_X - \mu_Y$

Q: What if we do not have any knowledge about the *particular* form of the joint distribution of data?

Consider the problem of 2-sample comparison.

F, G: any continous distribution

- Let Ω be the collection of *all* continuous distributions
- Only assume that $F, G \in \Omega$ — Recall. LNp.2 case (a)
- Thus, the statistical model is:



parameter space/ model space

$$\left.\begin{array}{llll} \text{1st sample:} & X_1, \ldots, X_n & \sim & \text{i.i.d. from } F \\ \text{2nd sample:} & Y_1, \ldots, Y_m & \sim & \text{i.i.d. from } G \end{array}\right\} \Leftarrow \text{independent} \qquad (\square)$$

- This model contains parameters of *infinitely many* dimension because

$$\dim(\Omega) = \infty \quad (\textbf{why}?)$$

— pdf: $\int_{-\infty}^{\infty} f(x)dx = 1$
— cdf: nondecreasing F
— mgf (chf): $M(t) \Rightarrow$ kth moment $= M^{(k)}(0)$, $k = 1, 2, \cdots$

check LNp.5 case (a)

- Under this model, a 2-sample comparison examines the null and alternative hypotheses: $H_0: F = G$ vs. $H_A: F \neq G$. ← no difference

Definition 2 (nonparametric models and nonparametric methods)

such as normal, exponential, Poisson, ...

- Nonparametric models do not assume any *particular* distributional form. Nonparametric models can be viewed as having *infinitely many* parameters. ($\overset{\text{cf.}}{\longleftrightarrow}$ parametric models: parameters are of *finite* dimension) cf.
- Statistical methods developed under nonparametric models are called *nonparametric methods*.

The methods can work for data from any distribution

Q: What statistics (transformation) often appear in nonparametric method?

Review 3 (order statistics and ranks)

empirical cdf (TBp.378)

- Let $X_1, X_2, \ldots, X_n$ be random variables. We sort the X_i's and denote by $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ the *order statistics*. Using the notation, ← 順序統計量

$$\begin{array}{lcll} X_{(1)} & = & \min(X_1, X_2, \ldots, X_n) & \text{is the } \textit{minimum}, \\ X_{(n)} & = & \max(X_1, X_2, \ldots, X_n) & \text{is the } \textit{maximum}. \end{array}$$

- TBp.105-107
- ⊠, CH1-6, p.34-37

percentile — median is a function of order statistics

true cdf, y_n, $X_{(1)} X_{(2)} \cdots X_{(n)}$

- Let $R(X_1, X_2, \ldots, X_n) = (R_1, R_2, \ldots, R_n)$ such that $X_i = X_{(R_i)}$, $i = 1, \ldots, n$. Then, $(R_1, R_2, \ldots, R_n)$ is called the *ranks* of $X_1, X_2, \ldots, X_n$. Notice that

# of observations $\le X_i$ → $$R_i = \sum_{j=1}^{n} \delta(X_i - X_j), \quad \text{where } \delta(t) = \begin{cases} 1, & \text{if } t \ge 0, \\ 0, & \text{if } t < 0. \end{cases}$$

complete information in $X_1, \cdots, X_n$

transformation g

$$\begin{array}{lcccccc} \text{data:} & X_4 & X_2 & X_3 & X_6 & X_1 & X_5 \\ \text{order statistics:} & X_{(1)} & X_{(2)} & X_{(3)} & X_{(4)} & X_{(5)} & X_{(6)} \\ \text{ranks:} & R_4=1 & R_2=2 & R_3=3 & R_6=4 & R_1=5 & R_5=6 \end{array} \quad \mathbb{R}$$

g^{-1} exist? NO.

1-sample data $\underline{X}$ $\underline{T}$ $\underline{T} = g(\underline{X})$

Theorem 8 (sufficient and complete statistics for nonparametric models)

Let $X_1, \ldots, X_n$ be i.i.d. from F, where $F \in \Omega$. ← dim = ∞

Then, $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ is sufficient and complete. → removing information in $(R_1, \ldots, R_n)$ (Why? ∵ i.i.d.)

ancillary statistics

$(\overset{\text{cf.}}{\longleftrightarrow}$ $X_1, \ldots, X_n \sim$ i.i.d. $N(\mu, \sigma^2)$ $\Rightarrow$ $(\overline{X}, s_X^2)$ is sufficient and complete.)

cf. dim = 2

Proof. Denote the pdf of F by f. The joint pdf of $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ is ← TBp.106

$$f_{X_{(1)}, X_{(2)}, \ldots, X_{(n)}}(x_1, x_2, \ldots, x_n) = n! \times f(x_1) f(x_2) \cdots f(x_n),$$

joint pdf of $\underline{X}$ is of the form, but $(x_1, \cdots, x_n) \in \mathbb{R}^n$

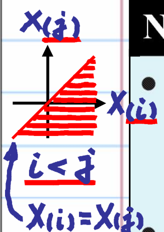
for $x_1 < x_2 < \cdots < x_n$ and zero, otherwise.

$X_1 \leftarrow X_{(1)}$, $X_n \leftarrow X_{(n)}$, $(x_1 \cdots x_n)$

The proof of sufficiency follows from the fact that the conditional probability of $X_1, \ldots, X_n$ given $X_{(1)}, \ldots, X_{(n)}$ is $\dfrac{1}{n!}$, which is irrelevant to F. ← TBp.305

joint pdf of $\underline{X}$ = joint pdf $(\underline{X}, \underline{T})$ / joint pdf of $\underline{T}$

The proof of completeness is omitted (out of the scope of this course).



Note 7 (Some notes about order statistics and ranks)

- Order statistics and ranks are defined precisely, i.e., **no ties**, under the condition $P(X_i = X_j) = 0$, $i \neq j$ (**Note.** this condition holds when $X_1, \ldots, X_n \sim$ i.i.d. from F and F is a continuous distribution).
- Under Ω, the dimension of data (i.e., n) cannot be reduced without losing the information about F ($\in \Omega$). combine information
- Under 1-sample model, ranks + order statistics = complete data
- Order statistics are intuitive estimator of quantiles, e.g., median. $\xrightarrow{e}$ 50% quantile

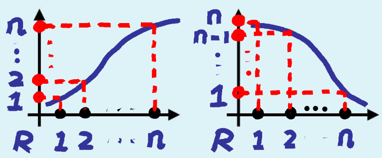
percentile

- Ranks are invariant under any monotonic transformation of data, i.e.,

$$R(X_1, \ldots, X_n) = R(H(X_1), \ldots, H(X_n)),$$

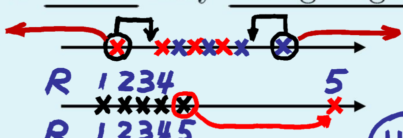
if H is a monotone increasing function and

$$R(X_1, \ldots, X_n) = (n+1) - R(H(X_1), \ldots, H(X_n)),$$



if H is a monotone decreasing function. ($\overset{\text{cf.}}{\longleftrightarrow}$ z- or t-tests may change significantly under monotonic transformations of data).

$(X_1, \cdots, X_n) \to (R_1, \cdots, R_n)$



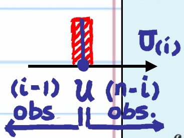
- Replacing the data by their ranks also has the effect of moderating the influence of outliers.

check graphs in LNp.25

- Many nonparametric methods are based on order statistics and/or ranks. $1/1$ nondecreasing
- **Q**: Why are many nonparametric methods based on replacement of the data by ranks? What information of data are contained in their ranks?

TBp.63 (exercise) – **Recall.** Let $X_1, \ldots, X_n$ be i.i.d. from a continuous cdf F, and let $U_i = F(X_i)$, $i = 1, \ldots, n$. Then, $U_1, \ldots, U_n$ are i.i.d. from $U(0,1)$.

TBp.105 (exercise) – **Recall.** If $U_1, \ldots, U_n \sim$ i.i.d. $U(0,1)$, the pdf of the ith-order statistic $U_{(i)}$ is



$$f_{U_{(i)}}(u) = \frac{n!}{(i-1)!(n-i)!} u^{i-1} (1-u)^{n-i},$$

Beta$(i, n-i+1)$

for $0 < u < 1$ and zero, otherwise.

Note that $E(U_{(i)}) = i/(n+1)$.