# Non-parametric tests

12.2, 12.4 of Ross Textbook

# Non-parametric tests

- We make no assumptions of the form of the distribution function unlike previous cases where we assume Normal or Binomial.
- Generically denote distribution as F(X). Note form of F(X) is not known
- Possible hypothesis that can be tested in such cases
  - What is the median of F(X)?
    - Sign test
  - Is the distribution around the median similar
    - Sign rank test
  - Given samples of two distributions: Are they likely to be from the same or different distributions?
    - Two sample test

# Sign test

Let $X_1, \ldots, X_n$ denote a sample from a continuous distribution $F$ and suppose that we are interested in testing the hypothesis that the median of $F$, call it $m$, is equal to a specified value $m_0$. That is, consider a test of

$$H_0 : m = m_0 \qquad \text{versus} \qquad H_1 : m \neq m_0$$

where $m$ is such that $F(m) = .5$.

$\hat{m}$ = median of $(X_1, \ldots X_n)$

$T$ = # of $X_i$-s less than $m_0$

$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \leq x)$

$P_{H_0}(T) \sim \text{Binomial}(n, \frac{1}{2})$ $T = \hat{F}(m_0)$

# T-statistic

- T = sum of the sign of $m_0 - X_i$
  - $T = \sum_i I_i$ $\quad I_i = \begin{cases} 1 & \text{if } X_i < m_0 \\ 0 & \text{if } X_i \geq m_0 \end{cases}$ $\quad P(I_i) \sim \text{Bernoulli}\left(\frac{1}{2}\right)$
- What is the distribution of T under the null hypothesis?

$$P_{H_0}(T) \sim \text{Binomial}\left(n, \frac{1}{2}\right)$$

$x_1 \quad x_2 \quad - \quad \cdot \quad x_{n/2} \quad x_{n/2+1} \quad - \quad - \quad \cdot \quad x_n$

$$T = \sum_{j=1}^{n/2} \left( |x_j - m_0| + |x_{n-j} - m_0| \right) \qquad P(T)$$

**Example 12.2.b.** A financial institution has decided to open an office in a certain community if it can be established that the median annual income of families in the community is greater than \$90,000. To obtain information, a random sample of 80 families was chosen, and the family incomes determined. If 28 of these families had annual incomes below and 52 had annual incomes above \$90,000, is this significant enough to establish, say, at the 5 percent level of significance, that the median annual income in the community is greater than \$90,000?

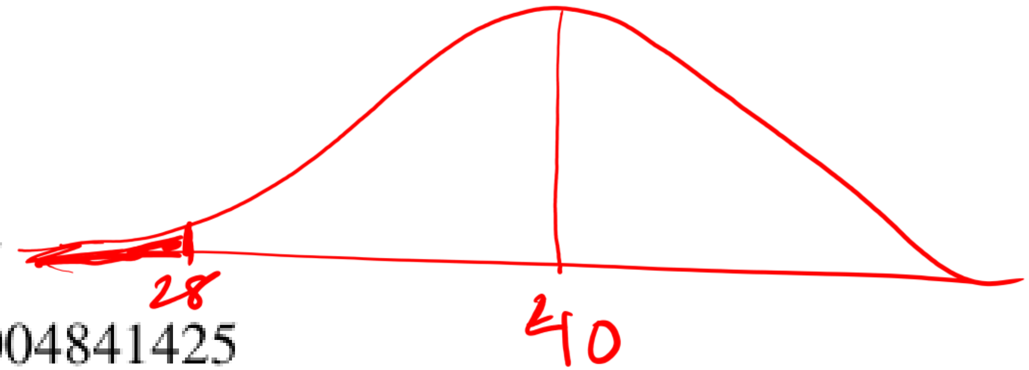
$m_0 = 90k$

$n = 80$

$T = 28$

**Solution.** We need to see if the data are sufficient to enable us to reject the null hypothesis when testing

$$H_0 : m \leq 90 \qquad \text{versus} \qquad H_1 : m > 90$$




$$p\text{-value} = P(\text{Bin}(80, 1/2) \leq 28) = \text{pbinom}(28, 80, 1/2) = 0.004841425$$

and so the null hypothesis that the median income is less than or equal to \$90,000 is rejected. ■

# Signed rank test

Given n sample $X_1, \ldots, X_n$ from unknown distribution F, we are interested in the hypothesis that F is symmetric about a given median $m_0$, that is,

- $\mathrm{H_0}$: $\mathrm{P}(X > m_0 + a) = \mathrm{P}(\mathrm{X} < \mathrm{m_0} - \mathrm{a})$, for all a

Let $Y_i = X_i - m_0$, $i = 1, \ldots, n$ and rank (that is, order) the absolute values $|Y_1|, |Y_2|, \ldots, |Y_n|$. Set, for $j = 1, \ldots, n$,

$$I_j = \begin{cases} 1 & \text{if the } j\text{th smallest value comes from a data value that is smaller than } m_0 \\ 0 & \text{otherwise} \end{cases}$$

Test statistic

$$T = \sum_{j=1}^{n} j I_j$$

**Example 12.3.a.** If $n = 4$, $m_0 = 2$, and the data values are $X_1 = 4.2$, $X_2 = 1.8$, $X_3 = 5.3$, $X_4 = 1.7$, then the rankings of $|X_i - 2|$ are .2, .3, 2.2, 3.3. Since the first of these values — namely, .2 — comes from the data point $X_2$, which is less than 2, it follows that $I_1 = 1$. Similarly, $I_2 = 1$, and $I_3$ and $I_4$ equal 0. Hence, the value of the test statistic is $T = 1 + 2 = 3$. ■

| | | | | |
|---|---|---|---|---|
| $x_i$ | 4.2 | 1.8 | 5.3 | 1.7 |
| | $-2$ | $-2$ | $-2$ | $-2$ |
| $y_i$ | 2.2 | $-0.2$ | 3.3 | $-0.3$ |
| $\|y_i\|$ | 0.2 | 0.3 | 2.2 | 3.3 |
| | 1 | 1 | 0 | 0 |

$P_{H_0}(T \le 3)$

# Distribution of test statistic $P_{H_0}(T)$ under the null hypothesis?

Expected value and variance of T under $H_0$

Probability that the $j^{th}$ smallest absolute difference is from an $x_k < m_0$.

$P\{I_j = 1\} = \frac{1}{2} = P\{I_j = 0\}, \quad j = 1, \ldots, n$ $\qquad E[I_j] = \frac{1}{2}, \qquad Var(I_j) = \frac{1}{4}$

Hence, we can conclude that under $H_0$,

$$E[T] = E\left[\sum_{j=1}^{n} jI_j\right] = \sum_{j=1}^{n} \frac{j}{2} = \frac{n(n+1)}{4}$$

$$\text{Var}(T) = \text{Var}\left(\sum_{j=1}^{n} jI_j\right) = \sum_{j=1}^{n} j^2 \text{Var}(I_j) = \sum_{j=1}^{n} \frac{j^2}{4} = \frac{n(n+1)(2n+1)}{24}$$

$P_{H_0}(T)$ = approximately normal for large n with mean and variance as above. But we can do better..

# An exact computation of probability $P_{H_0}(T)$ recursively

$$
\begin{aligned}
P_k(i) &= P_{H_0}\left\{\sum_{j=1}^{k} jI_j \le i\right\} \\
&= P_{H_0}\left\{\sum_{j=1}^{k} jI_j \le i \,|\, I_k = 1\right\} P_{H_0}\{I_k = 1\} \\
&\quad + P_{H_0}\left\{\sum_{j=1}^{k} jI_j \le i \,|\, I_k = 0\right\} P_{H_0}\{I_k = 0\} \\
&= P_{H_0}\left\{\sum_{j=1}^{k-1} jI_j \le i - k \,|\, I_k = 1\right\} P_{H_0}\{I_k = 1\} \\
&\quad + P_{H_0}\left\{\sum_{j=1}^{k-1} jI_j \le i \,|\, I_k = 0\right\} P_{H_0}\{I_k = 0\} \\
&= P_{H_0}\left\{\sum_{j=1}^{k-1} jI_j \le i - k\right\} P_{H_0}\{I_k = 1\} + P_{H_0}\left\{\sum_{j=1}^{k-1} jI_j \le i\right\} P_{H_0}\{I_k = 0\}
\end{aligned}
$$

$P_{H_0}(T \le i) = P_{H_0}($

# Continued..

$$P_{H_0}\{I_k = 1\} = P_{H_0}\{I_k = 0\} = \frac{1}{2}$$

we see that

$$P_k(i) = \frac{1}{2} P_{k-1}(i-k) + \frac{1}{2} P_{k-1}(i)$$

$$P_k(i) = P_{k-1}(i-k) P\{I_k = 1) + P_{k-1}(i) P\{I_k = 0)$$

Base Case:

$$P_1(i) = \begin{cases} 0 & i < 0 \\ \frac{1}{2} & i = 0 \\ 1 & i \geq 1 \end{cases}$$

$$P(I_1 < 0) = 0$$

$$P(I_1 \leq 0) = \frac{1}{2}$$

$$P(I \leq 1) = 1$$

$$P_1(i) = P_{H_0}(T \leq i)$$

# Example

Compute: $P_4(3)$

$$= P_{H_0}\left(\sum_{j=1}^{4} j I_j \le 3\right) = \frac{1}{2} P_3(-1) + \frac{1}{2} P_3(3)$$

$$= 0 + \frac{1}{2}\left[ P_2(0) + P_2(3) \right]$$

$$\frac{1}{2} \cdot \frac{1}{2}\left[ P_1(-2) + P_1(0) + P_1(1) + P_1(3) \right]$$

# HW

- How to extend paired-t-test to the non-parametric case?

# Are two distributions equal?

- Let F and G be two continuous distributions of unknown form
- Given
  - n samples $X_1, \dots, X_n$ from F
  - m samples $Y_1, \dots, Y_m$ from G
- Null hypothesis: $H_0: F = G$
- Test is called: Rank-sum test, Mann-Whitney test, Wilcoxon test

Rank order the n+m items.

$$R_i = \text{rank of the data value } X_i$$

Test statistic:

$$T = \sum_{i=1}^{n} R_i$$

**Example 12.4.a.** An experiment designed to compare two treatments against corrosion yielded the following data in pieces of wire subjected to the two treatments.

| | |
|---|---|
| Treatment 1 | 65.2, 67.1, 69.4, 78.2, 74, 80.3 |
| Treatment 2 | 59.4, 72.1, 68, 66.2, 58.5 |

(The data represent the maximum depth of pits in units of one thousandth of an inch.) The ordered values are 58.5, 59.4, 65.2*, 66.2, 67.1*, 68, 69.4*, 72.1, 74*, 78.2*, 80.3* with an asterisk noting that the data value was from sample 1. Hence, the value of the test statistic is $T = 3 + 5 + 7 + 9 + 10 + 11 = 45$. ■

# Distribution of test-statistic under the null hypothesis $P_{H_0}(T)$

Self-study.

- Again we will compute recursively.
- Let $P(n, m, t) = P_{H_0}(T \leq t)$

Either the last item in the rank is one of the N $X_i s$, or it is one of the M $Y_j s$. Under the null hypothesis, this probability:

$$P(N, M, K) = \frac{N}{N+M} P(N-1, M, K-N-M) + \frac{M}{N+M} P(N, M-1, K)$$

Starting with the boundary condition

$$P(1, 0, K) = \begin{cases} 0 & K \leq 0 \\ 1 & K > 0 \end{cases}, \qquad P(0, 1, K) = \begin{cases} 0 & K < 0 \\ 1 & K \geq 0 \end{cases}$$

**Example 12.4.b.** Suppose we wanted to determine $P(2, 1, 3)$. We use Equation (12.4.3) as follows:

$$P(2, 1, 3) = \tfrac{2}{3} P(1, 1, 0) + \tfrac{1}{3} P(2, 0, 3)$$

and

$$\begin{aligned}
P(1, 1, 0) &= \tfrac{1}{2} P(0, 1, -2) + \tfrac{1}{2} P(1, 0, 0) = 0 \\
P(2, 0, 3) &= P(1, 0, 1) \\
&= P(0, 0, 0) = 1
\end{aligned}$$

**Example 12.4.a.** An experiment designed to compare two treatments against corrosion yielded the following data in pieces of wire subjected to the two treatments.

| | |
|---|---|
| Treatment 1 | 65.2, 67.1, 69.4, 78.2, 74, 80.3 |
| Treatment 2 | 59.4, 72.1, 68, 66.2, 58.5 |

(The data represent the maximum depth of pits in units of one thousandth of an inch.) The ordered values are 58.5, 59.4, 65.2*, 66.2, 67.1*, 68, 69.4*, 72.1, 74*, 78.2*, 80.3* with an asterisk noting that the data value was from sample 1. Hence, the value of the test statistic is $T = 3 + 5 + 7 + 9 + 10 + 11 = 45$. ■

$$P(6,5,45) = \frac{6}{11}P(5,5,34) + \frac{5}{11}P(6,4,45) = \cdots$$

**Wilcoxon rank sum test**

data: $x$ and $y$
$W = 24$, p-value = 0.1255

# Errors in Hypothesis testing

- Type-I error: Rejecting $H_0$ even when $H_0$ is true.
  - The probability with which it happens is called significant level $\alpha$
- Type-II error: Accepting $H_0$ when it is false

# Summary of hypothesis testing

- Follow this framework:
  - Formulate null and alternative hypothesis
  - Collect data
  - Decide on test statistic
  - Identify distribution of test statistic under null hypothesis
  - Apply p-value or critical region test to accept or reject null hypothesis
- We applied this framework on
  - Mean of Gaussian with unknown variance is $\mu_0$
  - Are means of two normal distributions with shared unknown variance same?
  - Difference in means of two normal with unknown variance from paired observations

# Summary..

- Parameter p of Bernoulli is $p_0$

- Non-parametric tests
  - Median is a given value
  - Distribution is symmetric around a median
  - Are two distributions equal

# Topics not covered.

- Goodness of fit tests
- Test on sequences