Stat 3005 – Nonparametric Statistics

Assignment 2

Due: 11 October (Friday) at 1800

Fall 2024

Exercise 2.1 (★☆☆ — Universal law of urination (50%))**.** A study shows that nearly all mammals take the approximately same amount of time to urinate. The research entitled "Duration of urination does not change with body size" was published in Proceedings of the National Academy of Sciences (PNAS) of the United States of America. An image reported in the paper is shown below:

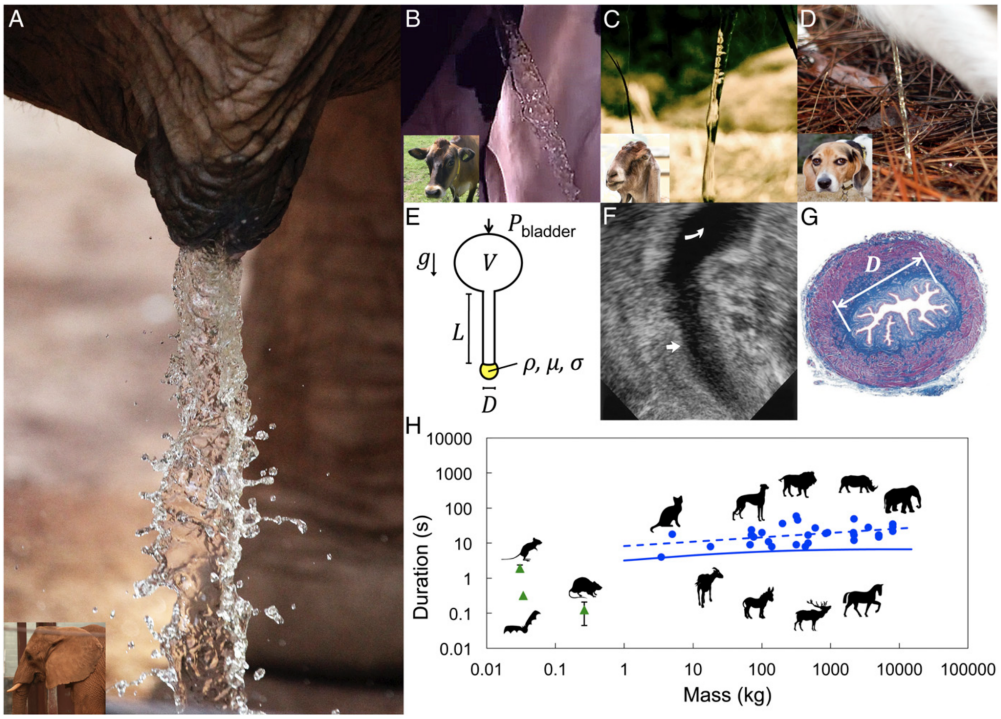


Fig. 1. Jetting urination by large animals, including (*A*) elephant, (*B*) cow, (*C*) goat, and (*D*) dog. *Inset* of cow is reprinted from the public domain and cited in *SI Appendix*. (*E*) Schematic of the urinary system. (*F*) Ultrasound image of the bladder and urethra of a female human. The straight arrow indicates the urethra, and the curved arrow indicates the bladder. Reproduced with permission from ref. 20, (Copyright 2005, Radiological Society of North America). (*G*) Transverse histological sections of the urethra from a female pig. Reproduced with permission from ref. 9, (Copyright 2001, Elsevier). (*H*) The relationship between body mass and urination time.

This has been awarded one of the 2015 Ig Nobel prizes at Harvard University. You may read the news and watch the winner speech in Guardian News.

1. Consider the dataset `urination.csv` obtained from the article "Duration of urination does not change with body size". (Note: We randomly selected one record if the same animal was observed more than one time.) It contains the urination duration of mammals above 3 kg. Denote the median urination time of a mammal above 3 kg by θ. We are interested in testing $H_0 : \theta = 21$ against $H_1 : \theta \neq 21$.
   (a) (10%) Perform an appropriate nonparametric test, and state your assumptions.
   (b) (10%) Briefly explain your findings ($\lesssim$ 100 words) in layman terms. Your answers should be understandable and could be appreciated by people without formal training in statistics.
2. Ben does not believe that all mammals above 3 kg urinate the same period of time. He suspects that the heavier the mammal, the longer is the urination duration. Divide the dataset into two samples. The sample 1 contains all mammals with weight less than 500 kg, whereas the sample 2 contains the remaining mammals. Denote the median urination time of the mammals in sample j by θ_j for $j = 1, 2$.
   (a) (10%) Write down the null and alternative hypotheses.
   (b) (10%) Perform the two-sample t-test and a nonparametric test. State your assumptions. Do they give the same results? Which one do you trust? Why? [Updated]
   (c) (10%) Write a short report ($\lesssim$ 100 words) to summarize your finding. Your answers should be in a style that is suitable for publication in a scientific journal, e.g., PNAS.

Hints: See Remark 2.1. Don't read the hints unless you have no ideas and have tried for more than 15 mins.

Exercise 2.2 (★★☆ — Theory of rank-type test (50%)). Let the data $X_1, \ldots, X_n$ be generated as follows:

$$X_i = \begin{cases} \mu_1 + e_i & \text{if } i = 1, \ldots, n_1; \\ \mu_2 + e_i & \text{if } i = n_1 + 1, \ldots, n_1 + n_2; \\ \mu_3 + e_i & \text{if } i = n_1 + n_2 + 1, \ldots, n_1 + n_2 + n_3, \end{cases}$$

where $\mu_1, \mu_2, \mu_3 \in \mathbb{R}$ are constants, $n = n_1 + n_2 + n_3$, and $e_1, \ldots, e_n$ are IID continuous RVs. Suppose $p_j \equiv n_j/n \to \pi_j \in (0, 1)$ for $j = 1, 2, 3$. Jojo suspected that

$$\mu_1 < \mu_2 < \mu_3.$$

She proposed to use the following test statistic:

$$T = \frac{1}{n+1} \left\{ \left(\sum_{i=1}^{n_1} R_i \right) - \left(\sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} R_i \right) \right\}.$$

1. (10%) Formally state the null and alternative hypotheses for testing her claim.
2. (10%) Derive $\mu_T := \mathsf{E}(T)$ and $\sigma_T^2 := \mathsf{Var}(T)$ under your null hypothesis.
3. (10%) Derive the limiting null distribution of T after proper standardization.
4. (10%) Argue that T is a sensible test statistic, and state the testing procedure. (Use $\lesssim 100$ words.)
5. (10%) Let $\mu_1 = \mu_2 = \mu_3 = 3005$, and $e_1, \ldots, e_n \overset{\text{IID}}{\sim} \text{Unif}(-2, 2)$. Use simulation to compute the exact 80%, 90%, 95%, 99% and 99.9% quantiles of t when $n_1 \in \{10, 20, 30\}$, $n_2 = 2n_1$, and $n_3 = 3n_1$. Compare them with the asymptotic quantiles. Report your results below. Briefly comment with $\lesssim 100$ words.

	80%	90%	95%	99%	99.9%
$n_1 = 10$					
$n_1 = 20$					
$n_1 = 30$					
$n_1 \to \infty$					

Hints: See Remark 2.2. Don't read the hints unless you have no ideas and have tried for more than 15 mins.

Exercise 2.3 (★★★ — ✂ Point estimator based on the signed rank test (30%)). A test can be generalized to a point estimator. We illustrate it by using the signed rank test. Let $X_1, \ldots, X_n$ be IID and continuous RVs that are symmetric about θ. Recall that the signed rank test statistic for testing $H_0 : \theta = \theta_0$ is

$$T(\theta_0) = \frac{1}{n+1} \sum_{i=1}^{n} \text{sign}(X_i - \theta_0) \times \text{rank}(|X_i - \theta_0|).$$

We view $\theta \mapsto T(\theta)$ as a function of θ. By the principle of (generalized) method of moment:

$$\boxed{\mathsf{E}\{T(\theta)\} = 0} \quad \Rightarrow \quad \boxed{\text{We may estimate } \theta \text{ by } \widehat{\theta} \text{ such that } T(\widehat{\theta}) \approx 0}.$$

So, the estimator of θ implied by the signed rank test is $\widehat{\theta} = \arg\min_\theta |T(\theta) - 0|$. Unfortunately, computing $\widehat{\theta}$ is non-trivial. In this example, a simple formula for $\widehat{\theta}$ is derived. Let $W_{i,j} = (X_i + X_j)/2$ for $1 \le j \le i \le n$. The RVs $W_{i,j}$'s are called the Walsh averages. Denote

$$V(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{i} \mathbb{1}(W_{i,j} \ge \theta). \tag{2.1}$$

1. (10%) Use no more than ~ 20 words to explain the meaning of $V(\theta)$.
2. (10%) Prove that
$$V(\theta) = \frac{n+1}{2} \left\{ T(\theta) + \frac{n}{2} \right\}.$$
3. (10%) Prove that $\widehat{\theta} = \arg\min_\theta |V(\theta) - N/2|$, where $N = n(n+1)/2$. Hence, or otherwise, argue that one possible solution is given by
$$\widehat{\theta} = \begin{cases} W_{((N+1)/2)} & \text{if } N \text{ is odd;} \\ \{W_{(N/2)} + W_{(N/2+1)}\}/2 & \text{if } N \text{ is even,} \end{cases} \tag{2.2}$$
where $W_{(1)} \le W_{(2)} \le \cdots \le W_{(N)}$ are the N sorted values of $\{W_{i,j} : 1 \le j \le i \le n\}$.
4. Use the above result to find a point estimate for θ in Example 2.1.

Hints: See Remark 2.3. Don't read the hints unless you have no ideas and have tried for more than 15 mins.

Exercise 2.4 (★★☆ — ✂ Confidence interval based on the signed rank test)**.** A test can be generalized to an interval estimator. We illustrate it by using the signed rank test again. The notations and assumptions in Exercise 2.3 are adopted. By the duality of confidence interval (CI) and hypothesis test,

$$\boxed{T^2(\theta_0) \le c_{1-\alpha}} \quad \Rightarrow \quad \boxed{H_0 : \theta = \theta_0 \text{ is not rejected at level } \alpha} \quad \Rightarrow \quad \boxed{\theta_0 \text{ is in the } 100(1-\alpha)\% \text{ CI of } \theta},$$

where $\alpha \in (0, 1/2)$, c_q is the $100q\%$ quantile of $T^2(\theta_0)$ when $\theta = \theta_0$. Hence, the $100(1-\alpha)\%$ CI of θ implied by the signed rank test is given by $\widehat{I} = \left\{\theta : T^2(\theta) \le c_{1-\alpha}\right\}$. Unfortunately, computing $\widehat{I}$ is non-trivial (again). In this example, a simple formula for $\widehat{I}$ is derived.

1. Let $N_\alpha^{\pm} = n(n+1)/4 \pm (n+1)c_{1-\alpha}^{1/2}/2$ Prove that
$$\left\{\theta : T^2(\theta) \le c_{1-\alpha}\right\} = \left\{\theta : N_\alpha^- \le V(\theta) \le N_\alpha^+\right\}. \tag{2.3}$$
2. Denote $W_{(0)} = -\infty$. Prove that
$$\left\{\theta : N_\alpha^- \le V(\theta) \le N_\alpha^+\right\} = \left\{\theta : W_{(N-N_\alpha^+)} < \theta \le W_{(N-N_\alpha^- +1)}\right\}. \tag{2.4}$$
Hence, or otherwise, argue that $\widehat{I} = \left(W_{(N-N_\alpha^+)}, W_{(N-N_\alpha^- +1)}\right]$ satisfies $\mathsf{P}(\theta \in \widehat{I}) \ge 1-\alpha$ and, thus, is a $100(1-\alpha)\%$ CI of θ.
3. Use the above result to find a 99.5% confidence interval for θ in Exercise 2.1.

Hints: See Remark 2.4. Don't read the hints unless you have no ideas and have tried for more than 15 mins.

Remark 2.1 (Hints for Exercise 2.1)**.** In this exercise, as long as it is appropriate, you may choose to

- use exact or asymptotic reference null distribution,
- perform or not perform continuity correction, and
- adjust or not adjust for ties.

1. (a) You can easily download the dataset as follows.

```
id = "1-RP9fZbyi6YIiYBi4g5zCoSGWGCA0-0L"
data = read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download",id))
head(data)
mass = data[,3]
duration = data[,4]
```

You may use the R built-in function or a self-written function to perform the tests. (I suggest to try both as an exercise). You may find the assumptions from the summary table in Section 3.5.

(b) Technical statistical terms should be avoided.

2. (a) The alternative hypothesis is the statement that you suspect true.

(b) You may use the built-in function `t.test` to perform the t-test; see `help(t.test)` for details. Alternatively, you may write the t-test by yourself.

(c) Open ended question. You should use a test that assumes the most appropriate assumptions according to the problem. You may trust both or none of the tests. Explain clearly.

(c) The following analysis report is copied from a recent PNAS article entitled "A data-driven approach to identify risk profiles and protective drugs in COVID-19".

> *"The study population consisted of 576 hospitalized eligible patients admitted between March 1, 2020 and May 1, 2020. ... Crude and adjusted hazard rates were estimated by univariate and multivariate Cox regression analysis in order to identify significant predictors of in-hospital death. ... Of the 576 patients, 436 had complete records on all variables and were therefore included in a multivariate model. ... [T]he effects of the two main classes of RAASi, angiotensin converting enzyme inhibitor (ACEi) and angiotensin II receptor blocker (ARB), have been also estimated. Both ACEi (HR, 0.45; 95% CI, 0.20 to 0.99; P = 0.0474) and ARB (HR, 0.28; 95% CI, 0.13 to 0.61; P = 0.0011) were shown to have a significant protective effect."*

Remark 2.2 (Hints for Exercise 2.2)**.** This question concerns a 3-sample location test. We regard the first n_1 observations as sample 1, the next n_2 observations as sample 2, and the last n_3 observation as sample 3:

$$\underbrace{X_1,\ldots,X_{n_1}}_{\text{Sample 1 of size } n_1},\quad \underbrace{X_{n_1+1},\ldots,X_{n_1+n_2}}_{\text{Sample 2 of size } n_2},\quad \underbrace{X_{n_1+n_2+1},\ldots,X_{n_1+n_2+n_3}}_{\text{Sample 3 of size } n_3}.$$

1. The alternative hypothesis is a statement that you suspect true.
2. You may need to show that if $c(i)=\mathbb{1}(i\le n_1)-\mathbb{1}(i\ge n_1+n_2+1)$, then
$$\bar{c}\equiv\frac{1}{n}\sum_{i=1}^{n}c(i)=p_1-p_3,$$
where $p_j=n_j/n$ is the proportion of the jth sample. Prove that
$$\mu_T=\frac{n(p_1-p_3)}{2}\qquad\text{and}\qquad\sigma_T^2=\frac{n^2\{p_1+p_3-(p_1-p_3)^2\}}{12(n+1)}.$$
You may use WolframAlpha to symbolically compute the above summations and simplify expressions.
3. See Example 3.8 and other related examples. After properly standardizing T, we have $t=\cdots$. You may use without proof that
$$\max_{1\le i\le n}\{c(i)-\bar{c}\}^2=\max\left\{(p_1-p_3)^2,(p_1-p_3-1)^2,(p_1-p_3+1)^2\right\}\le 4.$$
4. Under H_1, the test statistic t is more likely to be ... We reject ... iff ...
5. See Example 3.17. The following R codes may be useful:

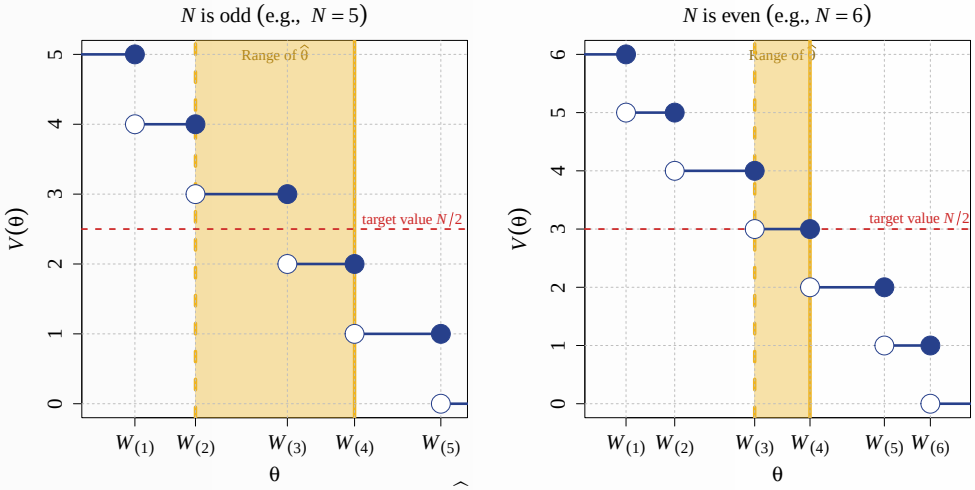
```r
test.3sample = function(x1,x2,x3){
    x = c(x1,x2,x3)
    nj = c(length(x1),length(x2),length(x3))
    n = sum(nj)
    p = nj/sum(n)
    R = rank(x)
    T = ...   # Complete this line
    muT = n/2*(p[1]-p[3])
    varT = n^2/12/(n+1)*(p[1]+p[3]-(p[1]-p[3])^2)
    t = ...   # Complete this line
    t
}
x1=rnorm(20)+2;   x2=rnorm(50)+4;   x3=rnorm(40)+5;  test.3sample(x1,x2,x3)
```

Takeaway: This question trains your ability to create a sensible rank-type test for new statistical problems.

Remark 2.3 (Hints for Exercise 2.3)**.**

1. $V(\theta)$ is the number of ...
2. The proof can be done in several steps.
   - **Step 1: Rewrite $V(\theta)$.** Let $Z_i=X_i-\theta$ for $i=1,\ldots,n$. Arrange them to $Z_{[1]},\ldots,Z_{[n]}$ according to their absolute values: $|Z_{[1]}|\le\cdots\le|Z_{[n]}|$. Note that $Z_{[i]}$ is different from $Z_{(i)}$ or $|Z_{(i)}|$. Then prove that
   $$V(\theta)=\sum_{i=1}^{n}V_i(\theta),\qquad\text{where}\qquad V_i(\theta)=\sum_{j=1}^{i}\mathbb{1}(Z_{[i]}+Z_{[j]}\ge 0).$$
   - **Step 2: Simplify $V_i(\theta)$.** Noting that $Z_{[i]}+Z_{[j]}\ge 0$ iff $Z_{[i]}\ge 0$. Argue that
   $$V_i(\theta)=i\mathbb{1}(Z_{[i]}\ge 0).$$
   - **Step 3: Introduce absolute ranks.** Argue that
   $$\sum_{i=1}^{n}i\mathbb{1}(Z_{[i]}\ge 0)=\sum_{i=1}^{n}\text{rank}(|Z_i|)\times\mathbb{1}(Z_i\ge 0).$$
   - **Step 4: Relate $V(\theta)$ with $T(\theta)$.** Using the identity $\text{sign}(Z_i)\equiv 2\mathbb{1}(Z_i\ge 0)-1$, prove that $V(\theta)$ can be written as a function of
   $$T(\theta)=\frac{1}{n+1}\sum_{i=1}^{n}\text{sign}(Z_i)\times\text{rank}(|Z_i|).$$

3. The graph of $\theta \mapsto V(\theta)$ looks like the following example:



Hence, we can argue that the target estimator $\widehat{\theta}$ satisfies that

- $\widehat{\theta} \in \left(W_{((N-1)/2)}, W_{((N+3)/2)}\right]$ if N is odd; and
- $\widehat{\theta} \in \left(W_{(N/2)}, W_{(N/2+1)}\right]$ if N is even.

The estimator $\widehat{\theta}$ is called the **Hodges–Lehmann estimator**. The form (2.2) is typically used in the literature. Generally, it is an estimator of the pseudo-median of F (i.e., $\vartheta := \mathsf{Med}(X_1 + X_2)/2$, where $X_1, X_2 \overset{\text{IID}}{\sim} F$) instead of the median of F (i.e., $\theta := \mathsf{Med}(X_1)$). Note that $\vartheta = \theta$ iff F is a symmetric distribution.

Remark 2.4 (Hints for Exercise 2.4).

1. Use Exercise 2.3 (2).
2. Note that (i) $\theta \mapsto V(\theta)$ is a decreasing function and left continuous, and (ii) $N_\alpha^{\pm}$ are integers. So, $V(\theta) \le M$ iff $\theta > \ldots$. (Please pay attention to the inequality signs $\le, <, \ge, >$. Drawing a graph of $\theta \mapsto V(\theta)$ is helpful.) Then using (2.3) and (2.4), you can prove that [Updated]

$$1 - \alpha \le \mathsf{P}\left\{\theta : T^2(\theta) \le c_{1-\alpha}\right\} = \cdots = \mathsf{P}(\theta \in \widehat{I}).$$

The take-away of this assignment is that the signed rank statistic can be used to perform (i) point estimation, (ii) interval estimation, and (iii) hypothesis test.

Inference based on signed rank statistic

Let $X_1, \ldots, X_n$ be IID continuous RVs that are symmetric about θ.

- The point-estimate of θ is

$$\widehat{\theta} = \begin{cases} W_{((N+1)/2)} & \text{if } N \text{ is odd;} \\ \{W_{(N/2)} + W_{(N/2+1)}\}/2 & \text{if } N \text{ is even,} \end{cases}$$

  where $W_{(k)}$ is the kth smallest value of $\{(X_i + X_j)/2 : 1 \le j \le i \le n\}$ for $1 \le k \le N$, and $N = n(n+1)/2$.
- The $100(1-\alpha)\%$ confidence interval of θ is

$$\widehat{I} = \left(W_{(N - N_\alpha^+)}, W_{(N - N_\alpha^- + 1)}\right],$$

  where $N_\alpha^{\pm} = n(n+1)/4 \pm (n+1)c_{1-\alpha}^{1/2}/2$, and $c_{1-\alpha} \approx \texttt{qchisq}(1-\alpha, 1) \times n(2n+1)/\{6(n+1)\}$.
- The p-value for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \ne \theta_0$ is

$$\widehat{p} \approx 1 - \texttt{pchisq}\left\{\frac{6}{n(2n+1)(n+1)}\left(\sum_{i=1}^{n} \text{sign}(X_i - \theta_0) \times \text{rank}(|X_i - \theta_0|)\right)^2, 1\right\}.$$

Instructions: Please follow the instructions stated on the last page of assignment 1.