University of Washington

STATISTICS

# Elements of Statistical Methods
# 1-Sample Location Problems (Ch 10)

Fritz Scholz

Spring Quarter 2010

May 12, 2010

# Small or Moderate Size Samples?

The 1-sample problem assumes that we observe $X_1, \ldots, X_n \ \text{iid} \ \sim P$

We made various types of inferences about the population mean $\mu = EX_i$,
point estimates, test of hypotheses, and set estimates.

This was possible since we used the plug-in estimator $\bar{X}_n$
and because the CLT implies that approximate normality holds for $\bar{X}_n$.

This was very potent and effective since we did not need to know $P$.
However, the sample size $n$ had to be sufficiently large.

What can we do when the sample size is not so large or when we wish to
address another measure of population location, such as the median $q_2(P)$

# Sampling Awareness Questions

1. What are the experimental units, i.e., the objects being measured?

2. From what population(s) are the experimental units drawn?

3. What measurements were taken on each unit?

4. What random variables are relevant to a specific inference question?

# Paper Quality

Recycled printing paper is supposed to have a weight of 24 pounds

for 500 sheets (17" $\times$ 22") and caliper (thickness) of $0.0048$" per sheet.

We draw a sample of sheets, its thickness is measured by a micrometer.

1) Experimental unit $=$ sheet of paper.

Note: Experimental unit $\neq$ measurement unit (inch)

2) Population: all sheets produced under this process (conceptual).

3) Measurement: caliper (thickness) of the sheet (experimental unit)

4) Relevant random variable: $X_i =$ caliper of $i^{\text{th}}$ sheet.

$X_1, \ldots, X_n$ iid $\sim P$. We are interested in inference about $q_2(P)$,

e.g., test $H_0 : q_2(P) = 0.0048$ versus $H_1 : q_2(P) \neq 0.0048$.

# Blood Pressure Medication

To test the effect of a drug meant to lower blood pressure, a number of patients have their blood pressure measured before and after taking the drug for two months.

1) Experimental unit: patient

2) Population: All hypertensive patients,

but is not always clear which part of the population is sampled.

3) Measurements: Blood pressure before, $B_i$, and after, $A_i$, for the $i^{\text{th}}$ patient.

4) Relevant random variable: $X_i = B_i - A_i$ for the $i^{\text{th}}$ patient.

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$. We are interested in inference about $q_2(P)$,

e.g., test $H_0 : q_2(P) \leq 0$ versus $H_1 : q_2(P) > 0$ (drug is effective).

# Effects of Parkinson's Disease (PD)

To understand the effect of PD on speech breathing, 14 PD patients were recruited into a study, together with 14 normal control (NC) subjects.

PD patients were carefully matched with NC patients (using a variety of criteria).

1) Experimental unit: each matched (PD,NC) pair.

2) Population: All possible (PD,NC) pairs satisfying matching criteria.

3) Measurements: Two measurements, PD lung volume (D) and NC lung volume (C) for each unit.

4) Relevant random variable: $X_i = \log(D_i/C_i)$ for the $i^{\text{th}}$ unit.

The $\log$-transform tends to symmetrize right-skewed volume measurements.

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$. We are interested in inference about $q_2(P)$,

e.g., test $H_0 : q_2(P) \geq 0$ versus $H_1 : q_2(P) < 0$ (PD restricts volume).

# Three Basic Analysis Scenarios

1. $X_1, \ldots, X_n$ iid $\sim \mathcal{N}(\mu, \sigma^2)$, $\mu = q_2(P)$ because of normal symmetry.

2. $X_1, \ldots, X_n$ iid $\sim F$, a continuous cdf, not necessarily normal or symmetric.
   $q_2(P)$ is the parameter of interest.

3. $X_1, \ldots, X_n$ iid $\sim F$, some continuous cdf, symmetric around some point,
   but not necessarily normal.
   $q_2(P)$, the point of symmetry, is the parameter of interest.

For each of the above scenarios we give inference solutions
w.r.t. point estimation, testing of hypotheses, and set estimation.

# Point Estimation: Normal Case

Since $\mu = q_2(P)$, the plug-in principle provides two different point estimators, the sample mean $\bar{X}_n$ and the sample median $q_2(\hat{P})$.

Both estimators are unbiased and consistent.

Both of them have approximate normal distributions

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \approx \mathcal{N}(0,1) \quad \text{or} \quad \bar{X}_n \approx \mathcal{N}(\mu, \sigma^2/n)$$

$$\frac{q_2(\hat{P}) - \mu}{\sigma\kappa/\sqrt{n}} \approx \mathcal{N}(0,1) \quad \text{or} \quad q_2(\hat{P}) \approx \mathcal{N}(\mu, \sigma^2\kappa^2/n)$$

It turns out that the ratio of the asymptotic variances $\sigma^2/n$ and $\sigma^2\kappa^2/n$ is

$$e(\mathcal{N}) = e(q_2(\hat{P}), \bar{X}_n; \mathcal{N}) = \frac{\sigma^2/n}{\sigma^2\kappa^2/n} = \frac{1}{\kappa^2} = \frac{2}{\pi} \approx 0.64$$

called the asymptotic relative efficiency (ARE) of $q_2(\hat{P})$ relative to $\bar{X}_n$.

# The Meaning of ARE

On the previous slide, note that the ARE ratio involved $n$ in both numerator and denominator and it canceled out.

If we allow different sample sizes $n_1$ and $n_2$ so that the asymptotic variance match,i.e.,

$$\frac{\sigma^2}{n_1} = \frac{\sigma^2\kappa^2}{n_2} \qquad \text{then} \qquad \frac{n_1}{n_2} = \frac{1}{\kappa^2} = e(\mathcal{N}) \approx 0.64 \qquad \text{or} \qquad n_1 \approx 0.64 \times n_2$$

i.e., $\bar{X}_n$ requires only about $64\%$ of the sample size needed by $q_2(\hat{P})$ to have the same variability around the common target $\mu = q_2(P)$.

In that sense $\bar{X}_n$ is more efficient than $q_2(\hat{P})$.

It can be shown that for any other estimator $\hat{\mu}$ of $\mu$ we have

$e(\bar{X}_n, \hat{\mu}; \mathcal{N}) \geq 1$ yet, there exists a $\hat{\mu}_R$ such that $e(\bar{X}_n, \hat{\mu}_R; P) \leq 1$ for all $P$ that are symmetric around $q_2(P)$.

In that case we of course must have $e(\bar{X}_n, \hat{\mu}_R; \mathcal{N}) = 1$,

i.e., equally efficient in the normal case, but $\hat{\mu}_R$ is superior otherwise.

# Testing Hypotheses: Normal Case ($\sigma^2$ Known)

Suppose we wish to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$, with known $\sigma^2$.

Because of the ARE results we use $\bar{X}_n$ and would naturally again regard large values of $|\bar{X}_n - \mu_0|$ as evidence against $H_0$.

To decide whether $H_0$ should be rejected, we calculate the significance probability

$$\begin{aligned} \mathbf{p}(z_n;\mu_0) &= P_{\mu_0}(|\bar{X}_n - \mu_0| \geq |\bar{x}_n - \mu_0|) = P_{\mu_0}\left(\frac{|\bar{X}_n - \mu_0|}{\sigma/\sqrt{n}} \geq \frac{|\bar{x}_n - \mu_0|}{\sigma/\sqrt{n}}\right) \\ &= P_{\mu_0}(|Z_n| \geq |z_n|) \stackrel{*}{=} 2 \cdot \Phi(-|z_n|) \quad \text{with} \quad z_n = (\bar{x}_n - \mu_0)/(\sigma/\sqrt{n}) \end{aligned}$$

If this looks the same as before, it is, except that at $\stackrel{*}{=}$ we invoke exact normality instead of approximate normality via the CLT. Reject at level $\alpha$ when $\mathbf{p}(z_n;\mu_0) \leq \alpha$.

For $X_1, \ldots, X_n \text{ iid } \sim \mathcal{N}(\mu, \sigma^2)$ we have $X_1 + \ldots + X_n \sim \mathcal{N}(n\mu, n\sigma^2)$ and thus $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$ or $Z_n = (\bar{X}_n - \mu)/(\sigma/\sqrt{n}) \sim \mathcal{N}(0,1)$, i.e., exact normality.

# Further Motivation of the Point Estimate $\bar{x}_n$

Given that a small value of $\mathbf{p}(z_n;\mu_0)$ is strong evidence against $H_0 : \mu = \mu_0$,
we can ask: what is the weakest evidence against $H_0 : \mu = \mu_0$ or
what would make $\mu_0$ most plausible?

Since $\mathbf{p}(z_n;\mu_0) \leq 1$, clearly $\mathbf{p}(z_n;\mu_0) = 1$ would present the weakest evidence
against $H_0 : \mu = \mu_0$ and we get that exactly when $\bar{x}_n = \mu_0$ because

$$\mathbf{p}(z_n;\mu_0) = P_{\mu_0}(|\bar{X}_n - \mu_0| \geq |\bar{x}_n - \mu_0|) = 1 \quad \Longleftrightarrow \quad \bar{x}_n = \mu_0$$

Thus $\mu_0 = \bar{x}_n$ would present the most plausible value for $\mu$,
i.e., the most plausible hypothesis $H_0 : \mu = \mu_0$.

This type of reasoning later becomes useful when we have a natural test,
but don't have a natural estimate to start with.

# Test Statistic: Normal Case ($\sigma^2$ Unknown)

Again we replace the unknown $\sigma^2$ by the estimator $S_n^2$ and use as test statistic

$$T_n(\mu_0) = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$$

Rather than invoking approximate normality via the CLT, we take advantage of the normality of the sampled distribution and some distributional facts stated below.

**Theorem:** $X_1, \ldots, X_n \text{ iid } \sim \mathcal{N}(\mu, \sigma^2)$, then

$$Y = \frac{(n-1)S_n^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X}_n)^2}{\sigma^2} \sim \chi^2(n-1)$$

and $\bar{X}_n$ and $S_n^2$ (and thus $Y$) are statistically independent.

**Corollary:** $X_1, \ldots, X_n \text{ iid } \sim \mathcal{N}(\mu, \sigma^2)$, then

$$T_n(\mu) = \frac{(\bar{X}_n - \mu)/(\sigma/\sqrt{n})}{S_n/\sigma} = \frac{Z}{\sqrt{Y/(n-1)}} \sim t(n-1)$$

# Testing Hypotheses: Normal Case ($\sigma^2$ Unknown)

Sampling from a normal distribution we have under $H_{\mu_0} : \mu = \mu_0$ the exact distribution of $T_n = T_n(\mu_0) \sim t(n-1)$.

We can compute the exact significance probability as

$$\mathbf{p}(t_n;\mu_0) = P_{\mu_0}(|T_n| \geq |t_n|) = 2F_{T_n}(-|t_n|) = 2 * \texttt{pt}(-\texttt{abs}(\texttt{t.n}))$$

where $\texttt{t.n} = t_n = (\bar{x}_n - \mu_0)/(s_n/\sqrt{n})$.

Here we exploit the symmetry of the $t(n-1)$ distribution around zero.

The basic difference to our approximate approach is that we use the exact $t$-distribution rather than the approximating $\mathcal{N}(0,1)$ distribution for large $n$.

Of course, for $n$ large we have $t(n-1) \approx \mathcal{N}(0,1)$.

# Set Estimation: Normal Case ($\sigma^2$ Known)

Let $q_z$ be the $(1-\alpha/2)$-quantile of the $\mathcal{N}(0,1)$ distribution, i.e., $P(|Z| \geq q_z) = \alpha$.

Then the hypothesis $H_0 : \mu = \mu_0$ is rejected at level $\alpha$ whenever

$$\mathbf{p}(\bar{x}_n;\mu_0) = P_{\mu_0}(|\bar{X}_n - \mu_0| \geq |\bar{x}_n - \mu_0|) = P_{\mu_0}\left(\frac{|\bar{X}_n - \mu_0|}{\sigma/\sqrt{n}} \geq \frac{|\bar{x}_n - \mu_0|}{\sigma/\sqrt{n}}\right) \leq \alpha$$

i.e., when $\dfrac{|\bar{x}_n - \mu_0|}{\sigma/\sqrt{n}} \geq q_z$ or when $|\bar{x}_n - \mu_0| \geq q_z\sigma/\sqrt{n}$

Conversely, $\mu_0$ is acceptable or plausible whenever $|\bar{x}_n - \mu_0| < q_z\sigma/\sqrt{n}$ or when

$$\mu_0 \in (\bar{x}_n - q_z\sigma/\sqrt{n}\ ,\ \ \bar{x}_n + q_z\sigma/\sqrt{n})$$

Thus $(\bar{x}_n - q_z\sigma/\sqrt{n}\ ,\ \ \bar{x}_n + q_z\sigma/\sqrt{n})$ is our $(1-\alpha)$-level confidence interval for $\mu$.

$$P_\mu\left(\mu \in \left(\bar{X}_n - q_z\frac{\sigma}{\sqrt{n}}\ ,\ \ \bar{X}_n + q_z\frac{\sigma}{\sqrt{n}}\right)\right) = P_\mu\left(\frac{|\bar{X}_n - \mu|}{\sigma/\sqrt{n}} < q_z\right) = P(|Z| < q_z) = 1-\alpha$$

# Set Estimation: Normal Case ($\sigma^2$ Unknown)

The previous derivation is repeated with $\sigma$ replaced by $s_n$ and $q_z$ replaced by $q_t$, the $(1-\alpha/2)$-quantile of the $t(n-1)$-distribution, resulting in

$$(\bar{x}_n - q_t s_n/\sqrt{n}\ ,\ \ \bar{x}_n + q_t s_n/\sqrt{n})$$

as our $(1-\alpha)$-level confidence interval for $\mu$.

$$P_\mu\left(\mu \in \left(\bar{X}_n - q_t\frac{S_n}{\sqrt{n}}\ ,\ \ \bar{X}_n + q_t\frac{S_n}{\sqrt{n}}\right)\right) = P_\mu\left(\frac{|\bar{X}_n - \mu|}{S_n/\sqrt{n}} < q_t\right) = P(|T_n| < q_t) = 1-\alpha$$

Thus our random interval has the exact claimed coverage probability.

Of course, this exactness depends on the normality of the sampled population.

Simple numerical examples are given in the text (for $\sigma$ known or unknown).

# General 1-Sample Location Problem

Previously we assumed normality to get exact inference procedures for $\mu$.

Now assume $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$, with a continuous cdf $F$.

No symmetry of $F$ is assumed.

We focus on the population median $q_2(P)$ (denoted by $\theta$ for simplicity).

$\theta$ always exists and is insensitive to the tail behavior of $P$.

More importantly, it allows for simple exact inference procedures.

# General 1-Sample Location Problem: Testing Hypotheses

Test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. Distribution continuity implies $P_\theta(X = \theta) = 0$.

Thus $\quad P_{\theta_0}(X \geq \theta_0) = P_{\theta_0}(X > \theta_0) = P_{\theta_0}(X \leq \theta_0) = P_{\theta_0}(X < \theta_0) = 1/2.$

By observing which $X_i < \theta_0$ and which $X_i > \theta_0$, we turn this into a Bernoulli trials problem, with known success probability $p_0 = 0.5$ under $H_0 : \theta = \theta_0$.

Consider the sign test statistic $Y = \#\{X_i > \theta_0\} = \#\{X_i - \theta_0 > 0\}$.
Values of $Y$ far away from $n/2$ would provide strong evidence against $H_0 : \theta = \theta_0$ or $\tilde{H}_0 : P_{\theta_0}(X > \theta_0) = 0.5$ in favor of $H_1 : \theta \neq \theta_0$ or $\tilde{H}_1 : P_{\theta_0}(X > \theta_0) \neq 0.5$.

Thus we should reject $H_0$ at level $\alpha$ whenever the significance probability for the observed count $y$ is

$$\mathbf{p}(y; 0.5) = P_{0.5}(|Y - n/2| \geq |y - n/2|) \leq \alpha$$

# Calculating $\mathbf{p}(y;0.5)$ for the Sign Test

Large values of $|y-n/2| \iff$ to values $y$ too close to 0 or $n$ $(n/2=(0+n)/2)$

With $c=\min(y,n-y)$ we can express (check this with $c=y$ and $c=n-y$)

$$\left|Y-\frac{n}{2}\right| \geq \left|y-\frac{n}{2}\right| \iff \left\{Y \leq \frac{n}{2}-\left|y-\frac{n}{2}\right|\right\} \cup \left\{Y \geq \frac{n}{2}+\left|y-\frac{n}{2}\right|\right\}$$
$$\iff \{Y \leq c\} \cup \{Y \geq n-c\}$$

Since the null distribution of $Y$ is symmetric around $n/2$ we have

$$\mathbf{p}(y;0.5) = P_{0.5}(Y \leq c) + P_{0.5}(Y \geq n-c) = 2P_{0.5}(Y \leq c) = 2 * \texttt{pbinom}(\texttt{c},\texttt{n},0.5)$$

Thus we reject $H_0$ at level $\alpha$ whenever

$$\mathbf{p}(y;0.5) = 2 * \texttt{pbinom}(\texttt{min}(\texttt{y},\texttt{n}-\texttt{y}),\texttt{n},0.5) \leq \alpha$$

# Sign Test Example

With an observed sample of $n = 10$

$$\vec{x} = \{98.73,\ 97.17,\ 100.17,\ 101.26,\ 94.47,\ 96.39,\ 99.67,\ 97.77,\ 97.46,\ 97.41\}$$

should we reject $H_0 : \theta = 100$ at level $\alpha = 0.05$?

$$\mathbf{p}(y = 2; 0.5) = 2 * \texttt{pbinom}(\texttt{min}(2, 10-2), 10, 0.5) = 0.109375 > 0.05$$

we should not reject $H_0$.

Should we reject $H_0 : \theta \leq 97$ in favor of $H_1 : \theta > 97$ at level $\alpha = 0.05$?

Of course, here we should reject $H_0$ when $Y = \#\{X_i - 97 > 0\}$ is too large.

For the observed $y = 8$ we get a significance probability

$$\mathbf{p}(y = 8; 0.5) = P_{0.5}(Y \geq 8) = 1 - \texttt{pbinom}(7, 10, 0.5) = 0.0546875 > 0.05$$

and we should not reject $H_0$ at $\alpha = 0.05$.

# The Problem of Zeros

Based on the continuity assumption we had $P_\theta(X = \theta) = 0$.

However, rounding of data often creates zeros for $X_i - \theta_0$. How to treat these?

The previous data, when rounded to the nearest integer, become

$$\vec{x} = \{99,\ 97,\ 100,\ 101,\ 94,\ 96,\ 100,\ 98,\ 97,\ 97\}$$

W.r.t. $H_0 : \theta = \theta_0 = 100$ versus $H_1 : \theta \neq \theta_0$ we have two $X_i - \theta_0 = 0$

1) Average all possible significance probabilities when splitting any $X_i - \theta_0 = 0$ into $+$ or $-$, respectively. There are $2^k$ such splits, where $k = \#\{X_i - \theta_0 = 0\}$.

2) Compute $\mathbf{p}_0(y; 0.5)$ for the split most favorable to $H_0$ and $\mathbf{p}_1(y; 0.5)$ for the split most favorable to $H_1$. If $\mathbf{p}_0(y; 0.5) \leq \alpha$, reject $H_0$.

If $\mathbf{p}_1(y; 0.5) > \alpha$, don't reject $H_0$. If $\mathbf{p}_0(y; 0.5) > \alpha \geq \mathbf{p}_1(y; 0.5)$, suspend judgment.

3) Remove cases with $X_i - \theta_0 = 0$ and work with the reduced sample as before.

$$\vec{x} = \{99,\ 97,\ 100,\ 101,\ 94,\ 96,\ 100,\ 98,\ 97,\ 97\}$$

| zero splits | $y = \#\{x_i > 100\}$ | $c = \min(y, 10-y)$ | $\mathbf{p}(y; 0.5)$ |
|---|---|---|---|
| $x_3 < 100,\ x_7 < 100$ | 1 | 1 | 0.021484 |
| $x_3 < 100,\ x_7 > 100$ | 2 | 2 | 0.109375 |
| $x_3 > 100,\ x_7 < 100$ | 2 | 2 | 0.109375 |
| $x_3 > 100,\ x_7 > 100$ | 3 | 3 | 0.343750 |

We have $\mathbf{p}_1(y; 0.5) = 0.021 < 0.05 < \mathbf{p}_0(y; 0.5) = 0.34$, suspend judgment.

Only one of the four $\mathbf{p}(y; 0.5)$ is $< 0.05$.

The average of all four $\mathbf{p}(y; 0.5)$ is 0.146, don't reject $H_0$ (conservative).

For the reduced analysis of $n = 8$ cases we get

$$2 * \texttt{pbinom}(1, 8, .5) = 0.0703125$$

again not significant at $\alpha = 0.05$.

# General 1-Sample Location Problem: Point Estimation

The plug-in estimate, i.e., the median of the empirical distribution $\hat{P}$ or sample median would be the natural estimator of the population median $\theta$.

Again we can motivate this sample median as the value $\theta_0$ for which we are least inclined to reject $H_0 : \theta = \theta_0$, namely for which (in the case of even $n$)

$$\mathbf{p}(y; 0.5) = P_{\theta_0}(|Y - n/2| \geq |y - n/2|) = P_{\theta_0}(|Y - n/2| \geq 0) = 1$$

since $y - n/2 = 0$ can be achieved for any $\theta_0$ that has $n/2$ of the $x_i$ to its left and $n/2$ to its right, which defines all possible choices for the sample median, the midpoint of that interval being the preferred choice.

For odd $n$ a similar argument can be made for $\theta_0 =$ the middle sample value

The efficiency results comparing sample mean with sample median in the case of normal samples carry over to the testing situation, comparing $t$-test with sign test. The $t$-test is "as effective" as the sign test, but for $36\%$ less data.

# General 1-Sample Location Problem: Set Estimation

We use the sign test for $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ to identify all values $\theta_0$ for which the sign test cannot reject the hypothesis at level $\alpha$.

This set of $\theta_0$'s again serves as our confidence set for $\theta$ based of $\vec{x}$.

Recall that we reject $H_0$ when $y(\theta_0) = \#\{x_i > \theta_0\}$ is too large or too small, i.e., when $y(\theta_0) \geq n - c$ or $y(\theta_0) \leq c$ where $c$ is an appropriate critical value, with

$$P_{\theta_0}(Y(\theta_0) \leq c) + P_{\theta_0}(Y(\theta_0) \geq n - c) = 2P_{\theta_0}(Y(\theta_0) \leq c) = 2 * \texttt{pbinom}(\texttt{c}, \texttt{n}, 0.5) = \alpha$$

For $c = 0, 1, 2, \ldots < n/2$ only a certain, discrete set of $\alpha$ values can be achieved.

If the desired $\alpha$ is not among the achievable $\alpha$ values, we have two options:

1) take the next lower achievable $\alpha$ value (the conservative option).

2) Take the achievable $\alpha$ value closest to the target $\alpha$ (make the best of it).

Whatever option is chosen, the corresponding critical value is denoted by $c_\alpha$.

# Finding the Right Critical Value

Given a target level $\alpha = \texttt{alpha}$ we can find with $c_0 = \texttt{c.0} = \texttt{qbinom}(\texttt{alpha}/2, \texttt{n}, 0.5)$ the smallest $c$, such that

$$P_{\theta_0}(Y(\theta_0) \leq c) = \texttt{pbinom}(\texttt{c}, \texttt{n}, 0.5) \geq \alpha/2$$

By evaluating

$$\texttt{pbinom}(\texttt{c.0}, \texttt{n}, 0.5) \qquad \text{and} \qquad \texttt{pbinom}(\texttt{c.0} - 1, \texttt{n}, 0.5)$$

we can decide to take $c_\alpha = c_0 - 1$ or $c_\alpha = c_0$ depending on option 1) or 2).

For duality conversion of test to confidence interval the following definition is useful.

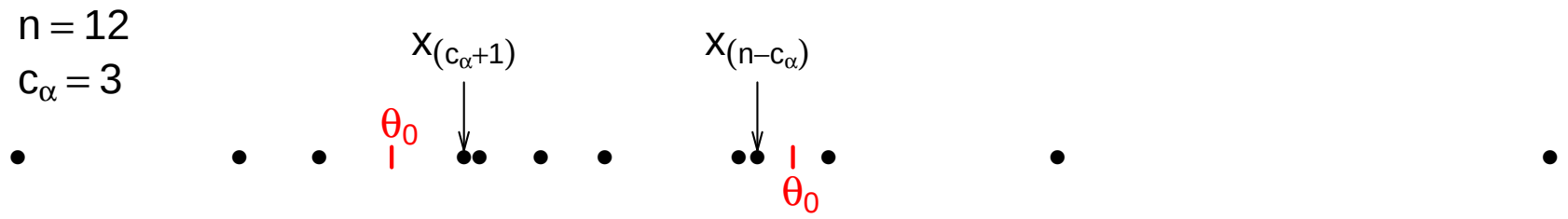
**Definition:** The order statistics of $\vec{x} = \{x_1, \ldots, x_n\}$ are any permutation of the $x_i$ that leaves them in non-decreasing order, denoted by $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.

For convenience we assume: $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$
occurring with probability 1 for continuous random variables.

# Sign Test to Confidence Interval Conversion

Note that $\theta_0 < x_{(c_\alpha+1)}$ means that at least $n - c_\alpha$ of the $x_i$ exceed $\theta_0$,

i.e. $y(\theta_0) \geq n - c_\alpha$. Then we would have to reject $H_0$ at level $\alpha$ (the achieved level).




Similarly, $\theta_0 > x_{(n-c_\alpha)}$ means that at most $c_\alpha$ of the $x_i$ exceed $\theta_0$,

i.e. $y(\theta_0) \leq c_\alpha$, in which case we also would have to reject $H_0$ at level $\alpha$.

Thus our set of acceptable $\theta_0$ is characterized by $[x_{(c_\alpha+1)},\ x_{(n-c_\alpha)}]$,

our $(1-\alpha)$-level confidence interval for $\theta$:

$$P_{\theta_0}(\theta_0 \in [X_{(c_\alpha+1)},\ X_{(n-c_\alpha)}]) = 1 - \alpha$$

$\alpha$ is the achieved level of the sign test, according to option 1) or 2).

# The Case of Ties

When there are ties among the $x_i$ due to rounding, and if rounded values are on a grid $0, \pm\varepsilon, \pm 2\varepsilon, \ldots$, then we know that each rounded value is at most $\varepsilon/2$ away from its unrounded continuous random variable source, denoted by $X_i^*$.

The previous confidence interval in terms of the continuous (distinct) order statistics $X^*_{(1)} < \ldots < X^*_{(n)}$ is

$$[X^*_{(c_\alpha+1)},\, X^*_{(n-c_\alpha)}] \subset [X_{(c_\alpha+1)} - \varepsilon/2\,,\;\; X_{(n-c_\alpha)} + \varepsilon/2]$$

$$\begin{aligned} \Longrightarrow \quad 1-\alpha &= P_{\theta_0}\left(\theta_0 \in [X^*_{(c_\alpha+1)},\, X^*_{(n-c_\alpha)}]\right) \\ &\leq P_{\theta_0}\left(\theta_0 \in [X_{(c_\alpha+1)} - \varepsilon/2\,,\;\; X_{(n-c_\alpha)} + \varepsilon/2]\right) \quad \text{conservatively} \end{aligned}$$

# Example

From our previous sample of size $n = 10$ we get the following order statistics

```
> ex10<-c(98.73, 97.17, 100.17, 101.26, 94.47,
          96.39, 99.67, 97.77, 97.46, 97.41)
> sort(ex10)
 [1]  94.47  96.39  97.17  97.41  97.46  97.77
      98.73  99.67 100.17 101.26
```

We desire a $90\%$ confidence interval for the median. `qbinom(.05,10,.5)=2` `pbinom(2,10,.5)=0.0546875` and `pbinom(1,10,.5)=0.01074219`.

Since $0.0546875$ comes a lot closer to $\alpha/2 = .05$ than $0.01074219$ it appears that $c_\alpha = 2$ is the more natural choice, corresponding to an achieved significance level $\alpha = 2 * 0.0546875 = 0.109375$, i.e., achieved confidence level $1 - \alpha = 0.890625$. The resulting confidence interval is $[x_{(2+1)}, x_{(10-2)}] = [x_{(3)}, x_{(8)}] = [97.17, 99.67]$.

# Symmetric 1-Sample Location Problem

Assume: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$ with a continuous distribution that is symmetric around some value $\theta$, which of course coincides with the median $=$ mean (if it exists).

A symmetric distribution around $\theta$ means $X_i - \theta \overset{\mathcal{D}}{=} \theta - X_i$

$$P_\theta(X_i \le \theta - x) = P_\theta(x \le \theta - X_i) = P_\theta(X_i - \theta \ge x) = P_\theta(X_i \ge \theta + x) \qquad \text{for any} \quad x \in R$$

This sample model is between our two previous assumptions: a normal sample (continuous and symmetric) and simply a sample from a continuous distribution.

Dropping the normality assumption we have a much broader application range.

First we develop an exact test for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ and, as before, use it to obtain the corresponding point estimates and confidence sets.

# Consequences of Symmetry

In considering the hypothesis $H_0 : \theta = \theta_0$ let $D_i = X_i - \theta_0$ which has a continuous distribution, symmetric around zero under $H_0$.

**Theorem:** Under $H_0$ the signs of the $D_i$ are independent of the $|D_i|$, $i = 1, \ldots, n$.

**Proof:** Since the $D_i$ are independent, this just needs to be proved for any $i$.

Let $D = D_i$, then

$$P_{\theta_0}(D > 0, |D| \leq d) = P_{\theta_0}(0 < D \leq d) = \frac{1}{2} \cdot P_{\theta_0}(|D| \leq d) = P_{\theta_0}(D > 0) \cdot P_{\theta_0}(|D| \leq d)$$

We made use of the fact that $P_{\theta_0}(D = 0) = 0$, $P_{\theta_0}(D > 0) = P_{\theta_0}(D < 0) = 1/2$

and

$$P_{\theta_0}(|D| \leq d) = P_{\theta_0}(-d \leq D < 0) + P_{\theta_0}(0 < D \leq d) = 2 \cdot P_{\theta_0}(0 < D \leq d)$$

q.e.d.

# Wilcoxon Signed Rank Test Statistic

Because of the continuous distribution assumption we have $P_{\theta_0}(D_i = 0)$ and $P_{\theta_0}(|D_i| = |D_j|) = 0$ for $i \neq j$.

Thus there is a unique ordering of the absolute differences $|D_i|$, i.e., we have the $|D_i|$ order statistics, denoted by $|D|_{(1)} < \ldots < |D|_{(n)}$, with

$$P_{\theta_0}(|D|_{(1)} < \ldots < |D|_{(n)}) = 1$$

Denote the ranks of $|D_1|, \ldots, |D_n|$ by $R_1, \ldots, R_n$ and consider the following

Wilcoxon signed rank test statistics

$$T_+(\theta_0) = T_+ = \sum_{D_i > 0} R_i \qquad \text{and} \qquad T_-(\theta_0) = T_- = \sum_{D_i < 0} R_i$$

Clearly,

$$T_+ + T_- = \sum_{i=1}^{n} R_i = \sum_{i=1}^{n} i = \underset{(n+1)}{\overset{n}{1}} + \underset{(n+1)}{\overset{(n-1)}{2}} + \ldots + \underset{(n+1)}{\overset{2}{(n-1)}} + \underset{(n+1)}{\overset{1}{n}} = \frac{n(n+1)}{2}$$

Thus it suffices to focus on $T_+$.

# Mean and Variance of $T_+$

Using Bernoulli random variables $I_{D_i>0} = 1$ for $D_i > 0$ and $I_{D_i>0} = 0$ for $D_i \leq 0$ we can rewrite

$$T_+ = \sum_{i=1}^{n} R_i \cdot I_{D_i>0}$$

Since the ranks $R_1, \ldots, R_n$ (a permutation of $1, 2, \ldots, n$) are based on the absolute values $|D_1|, \ldots, |D_n|$ and since the latter are independent of the Bernoulli random variables $I_{D_1>0}, \ldots, I_{D_n>0}$, we can view $T_+$ as having the same distribution as

$$\tilde{T}_+ = \sum_{i=1}^{n} i \cdot B_i \quad \text{with independent } B_i \sim \text{Bernoulli}(0.5),\ i = 1, \ldots, n$$

$$E_{\theta_0} T_+ = E_{\theta_0} \tilde{T}_+ = \sum_{i=1}^{n} i \cdot E_{\theta_0} B_i = \frac{1}{2} \cdot \sum_{i=1}^{n} i = \frac{1}{2} \cdot \frac{n(n+1)}{2} = \frac{n(n+1)}{4}$$

$$\text{var}_{\theta_0} T_+ = \text{var}_{\theta_0} \tilde{T}_+ = \sum_{i=1}^{n} i^2 \cdot \text{var}_{\theta_0} B_i = \frac{1}{4} \cdot \frac{n(n+1)(2n+1)}{6} = \frac{n(n+1)(2n+1)}{24}$$

# Wilcoxon Signed Rank and Sign Test Statistics

Comparing the Wilcoxon signed rank test statistic

$$T_+ = \sum_{i=1}^{n} R_i \cdot I_{D_i>0} \qquad E_{\theta_0} T_+ = n(n+1)/4 \quad \text{under } H_0$$

with the sign test statistic

$$Y = \sum_{i=1}^{n} I_{D_i>0} \qquad E_{\theta_0} Y = n/2 \quad \text{under } H_0$$

recall that we rejected $H_0 : \theta = \theta_0$ in favor of $H_1 : \theta \neq \theta_0$ whenever $Y$ was too high or too low, i.e., when $|Y - n/2|$ was too large. Similarly, we should reject $H_0$ when $|T_+ - n(n+1)/4|$ is too large.

We see that $T_+$ does more than count the number $Y$ of observations with $D_i = X_i - \theta_0 > 0$. It weights each such count by the rank $R_i$ of $|D_i|$.

The higher this weight $R_i$, the more $X_i$ differs from $\theta_0$, and the stronger is the evidence against $H_0$. $Y$ employs no such weighting (less effective).

# Symmetry of the $T_+$ Null Distribution

We consider the null distribution of $T_+$, i.e., its distribution under $H_0$.

From the representation

$$T_+ \overset{\mathcal{D}}{=} \tilde{T}_+ = \sum_{i=1}^{n} i \cdot B_i \quad \text{and the fact that} \quad \left(B_i - \frac{1}{2}\right) \overset{\mathcal{D}}{=} \left(\frac{1}{2} - B_i\right)$$

we see that

$$\tilde{T}_+ - \frac{n(n+1)}{4} = \sum_{i=1}^{n} i \cdot \left(B_i - \frac{1}{2}\right) \overset{\mathcal{D}}{=} \sum_{i=1}^{n} i \cdot \left(\frac{1}{2} - B_i\right) = \frac{n(n+1)}{4} - \tilde{T}_+$$

$\Longrightarrow$ the distribution of $\tilde{T}_+$ (and also that of $T_+$) is symmetric around $n(n+1)/4$.

# The Null Distribution of $T_+$

The null distribution of $\tilde{T}_+$ (and thus of $T_+$) is easy to obtain for small $n$, say $n = 4$.

There are $2^4 = 16$ sign patterns of length 4, all equally likely with probability $1/16$.

| $i$ | | | | | $i$ | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | $\tilde{T}_+$ | 1 | 2 | 3 | 4 | $\tilde{T}_+$ |
| $+$ | $+$ | $+$ | $+$ | 10 | $-$ | $+$ | $+$ | $+$ | 9 |
| $+$ | $+$ | $+$ | $-$ | 6 | $-$ | $+$ | $+$ | $-$ | 5 |
| $+$ | $+$ | $-$ | $+$ | 7 | $-$ | $+$ | $-$ | $+$ | 6 |
| $+$ | $+$ | $-$ | $-$ | 3 | $-$ | $+$ | $-$ | $-$ | 2 |
| $+$ | $-$ | $+$ | $+$ | 8 | $-$ | $-$ | $+$ | $+$ | 7 |
| $+$ | $-$ | $+$ | $-$ | 4 | $-$ | $-$ | $+$ | $-$ | 3 |
| $+$ | $-$ | $-$ | $+$ | 5 | $-$ | $-$ | $-$ | $+$ | 4 |
| $+$ | $-$ | $-$ | $-$ | 1 | $-$ | $-$ | $-$ | $-$ | 0 |

| $k$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $16 \cdot P_{\theta_0}(T_+ = k)$ | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 |

Note the symmetry!

# The Null Distribution of $T_+$ for Larger $n$

For $n = 20$ we already have $2^{20} = 1,048,576$ such sign patterns of length $n = 20$. Tabulate and organize their sums $\tilde{T}_+$ by frequency to get the null distribution.

R has a function `psignrank(k,n)` that calculates $P_{\theta_0}(T_+ \leq k)$ quite effectively, even for large $n$. For example for $n = 200$

```
> psignrank(8000,200)
[1] 0.006089115
> 2^200
[1] 1.606938e+60
> qsignrank(.006,200)
[1] 7996
```

`qsignrank(.006,200)` returns the smallest $k$ such that $P_{\theta_0}(T_+ \leq k) \geq 0.006$

`psignrank(7995,200)`=0.00598 & `psignrank(7996,200)`=0.0060045.

# Large Sample Approximation

Since $\tilde{T}_+ = \sum_{i=1}^n i \cdot B_i$ with $B_i$ iid $\sim \text{Bernoulli}(0.5)$

the following theorem should not surprise in view of the broader CLT.

**Theorem:** Under $H_0 : \theta = \theta_0$ the distribution of $T_+$ is approximately normal with mean $E_{\theta_0} T_+ = n(n+1)/4$ and variance $\text{var}_{\theta_0} T_+ = n(n+1)(2n+1)/24$, i.e.,

$$P_{\theta_0}\left(\frac{T_+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \le z\right) \longrightarrow \Phi(z) \quad \text{as } n \longrightarrow \infty$$

$$T_+ \approx \mathcal{N}(n(n+1)/4, n(n+1)(2n+1)/24)$$

$$P_{\theta_0}(T_+ \le t) \approx \Phi\left(\frac{t - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}\right)$$

# Simulated Null Distribution of $T_+$

It is an easy matter to simulate the null distribution of $T_+$.

Simply generate an $n$-vector of independent $\text{Bernoulli}(0.5)$ r.v.s $B_i$ via `B <- rbinom(n,1,.5)` and compute `T.plus <- sum((1:n)*B)`.

In a loop repeat this a large number of times, say `Nsim=10000`.

```
T.plus <- numeric(Nsim)
for(i in 1:Nsim){
    B <- rbinom(n,1,.5); T.plus[i] <- sum((1:n)*B)
}
```

By the LLN the vector `T.plus` of simulated $T_+$ values provides a good approximation to the $T_+$ null distribution, i.e.,

$$P_{\theta_0}(T_+ \le t) \approx \texttt{mean(T.plus <= t)} = \text{proportion of } \texttt{T.plus} \text{ values} \le t$$

# Functions Provided by M. Trosset

The text's web site `http://mypage.iu.edu/~mtrosset/StatInfeR.html` provides two functions (in `symmetric.R`) `W1.p.norm` and `W1.p.sim` that calculate

$$\mathbf{p}(t_+;\theta_0) = P_{\theta_0}(|T_+ - n(n+1)/4| \geq |t_+ - n(n+1)/4|)$$

either by normal approximation or by simulation approximation. Study the code.

Note that the approximation in the first case only becomes better as $n$ gets large, while in the second case it becomes better as the number of simulations increases.

```
> W1.p.norm(n=4,tplus=10)
[1] 0.1003482
> W1.p.sim(n=20,tplus=50,draws=10000) # draws = Nsim
[1] 0.0408
> W1.p.norm(n=20,tplus=50)
[1] 0.04188807 # the text 0.0412 is a typo
```

# The Case of Zeros and Ties

When some of $D_i = X_i - \theta_0$ are zero or when some of the absolute values $|D_i|$ are tied among each other we have sign ($\pm 0$?) and ranking problems.

There are several ways of dealing with this. The following, suggested by the text, seems easiest: Randomly perturb the $X_i$ slightly so that this situation is avoided.

Analyze the "fixed" data by the previous process and get a significance probability.

You can even figure out the worst and best case significance probabilities.

Repeat this data fixing process a few times and make a decision based on the obtained significance probabilities. Trosset provides `W1.p.ties`.

See the Example 10.11 discussion.

# Using `W1.p.ties`

```
> x0 <- c(1.5,9.7,3.9,7.6,8.0,7.3,5.0,9.7,2.3,2.3,
+         6.6,9.4,8.6,7.7,8.4,2.7,9.1,5.3,3.1,9.4)
> W1.p.ties(x=x0,theta0=5,draws=1000)
[1] 0.047
> W1.p.ties(x=x0,theta0=5,draws=1000)
[1] 0.045
> W1.p.ties(x=x0,theta0=5,draws=1000)
[1] 0.033
> W1.p.ties(x=x0,theta0=5,draws=1000)
[1] 0.032
> W1.p.ties(x=x0,theta0=5,draws=1000)
[1] 0.023
> W1.p.ties(x=x0,theta0=5,draws=1000)
[1] 0.028
```

It seems that we should reject $H_0 : \theta = 5$ at level $\alpha = 0.05$.

# Alternative Representation of $T_+$

Let $D_i = X_i - \theta_0$ in the context of testing $H_0 : \theta = \theta_0$

Consider the $\binom{n}{2} + n = n(n+1)/2$ Walsh averages $(D_i + D_j)/2$ with $i \le j$. Then

$$T_+(\theta_0) = T_+ = \text{number of positive averages } (D_i + D_j)/2 \text{ with } i \le j$$

Proof: Assume that the $D_i$ are indexed such that $0 < |D_1| < |D_2| < \ldots < |D_N|$

Note that $T_+$ does not depend on the indexing scheme for the $D_i$.

$$\sum_{i \le j} I_{[(D_i + D_j)/2 > 0]} = \sum_{i \le j} I_{[D_i + D_j > 0]} = \sum_{j=1}^{n} \sum_{i=1}^{j} I_{[D_i + D_j > 0]} \stackrel{*}{=} \sum_{j=1}^{n} j \times I_{[D_j > 0]} = T_+$$

where the indicator function $I_B$ is 1 when $B$ is true and 0 otherwise.

For $\stackrel{*}{=}$ in the above equation sequence note that

$$D_j > 0 \implies \pm D_i < D_j = |D_j| \text{ for } i < j \text{ , i.e., } D_i + D_j > 0 \text{ for all } i \le j$$
$$D_j < 0 \implies \pm D_i > D_j = -|D_j| \text{ for } i < j \text{ , i.e., } D_i + D_j < 0 \text{ for all } i \le j \qquad \square$$

# Point Estimate of $\theta$

With respect to the observed value $t_+$ of $T_+$ we just showed

$$\begin{aligned} t_+(\theta_0) = t_+ &= \#\left\{(i,j) : i \le j,\ \frac{(x_i-\theta_0)+(x_j-\theta_0)}{2} > 0\right\} \\ &= \#\left\{(i,j) : i \le j,\ \frac{x_i+x_j}{2} > \theta_0\right\} \end{aligned}$$

Since the distribution of $T_+$ is symmetric around $ET_+ = n(n+1)/4$ we would have least reason to reject $H_0 : \theta = \theta_0$ when $t_+(\theta_0) = t_+ = ET_+ = n(n+1)/4$,

i.e., when about half (i.e., $n(n+1)/4$) of the $n(n+1)/2$ Walsh averages are $> \theta_0$.

This will happen when we take

$$\theta_0 = \hat{\theta} = \text{median}_{i \le j}\left(\frac{x_i+x_j}{2}\right)$$

as point estimate for the center $\theta$ of our continuous and symmetric distribution.

# Some Comments on $\hat{\theta}$

The median of the Walsh averages is an interesting compromise between the mean $\bar{X}_n$ (suggested when sampling from a normal population) and $\text{median}(X_1, \ldots, X_n)$, when sampling from any continuous distribution.

For large $n$ the estimator $\hat{\theta}$ has an approximate normal distribution.

Its ARE relative to $\bar{X}_n$ is $3/\pi \approx .955$, when sampling from a normal population.

For any other other sampled distribution that ARE is $\geq 108/125 = .864$ and can be substantially higher than 1 (even $\infty$) in some situations.

Trosset provides `W1.walsh` in `symmetric.R` for calculating $\hat{\theta}$, e.g., for our previous sample `x0` we get `W1.walsh(x0) = 6.3`.

# Theorem about Ordered Walsh Averages

Recall our alternative representation

$$T_+(\theta_0) = \#\left\{(i,j) : i \le j,\ \frac{X_i + X_j}{2} > \theta_0\right\}$$

With probability 1 the Walsh averages $(X_i + X_j)/2$ are distinct.

Denote their $M = n(n+1)/2$ values by $A_{(1)} < \ldots < A_{(M)}$.

**Theorem:**

$$A_{(k)} \le \theta_0 \iff T_+(\theta_0) \le M - k$$

and thus

$$A_{(k)} > \theta_0 \iff T_+(\theta_0) \ge M - k + 1$$

**Proof:**

$A_{(k)} - \theta_0 \le 0 \iff$ at most $M-k$ of the differences $(X_i + X_j)/2 - \theta_0$ are $> 0$

$\iff T_+(\theta_0) \le M - k$ q.e.d.

# Confidence Intervals for $\theta$

Again we motivate the confidence set as consisting of all $\theta_0$ for which $H_0 : \theta = \theta_0$ is not rejectable at level $\alpha$ when testing against the alternative $H_1 : \theta \neq \theta_0$.

We reject $H_0$ whenever $|T_+(\theta_0) - n(n+1)/4|$ is too large, i.e.,

$$\begin{array}{rlll} & \text{when} \quad T_+(\theta_0) \leq k & \text{or} & T_+(\theta_0) \geq M-k \\ \Longleftrightarrow & \text{when} \quad A_{(M-k)} \leq \theta_0 & \text{or} & A_{(k+1)} > \theta_0 \end{array}$$

Thus we reject $H_0$ whenever $\theta_0 \notin [A_{(k+1)}, A_{(M-k)})$ with achieved significance level

$$\alpha = \alpha_k = P_{\theta_0}(T_+(\theta_0) \leq k) + P_{\theta_0}(T_+(\theta_0) \geq M-k) = 2P_{\theta_0}(T_+(\theta_0) \leq k)$$

The confidence interval for $\theta$ is $[A_{(k+1)}, A_{(M-k)})$ with achieved confidence level

$$P_{\theta_0}(\ \theta_0 \in [A_{(k+1)}, A_{(M-k)})\ ) = 1 - \alpha_k$$

Only a finite set of $\alpha$ values is achievable, corresponding to $k = 0, 1, \ldots < M/2$.

# Finding $(k, \alpha_k)$ for Given Target $\alpha$

For the given target $\alpha/2 = \texttt{alpha}/2$ find $k_\alpha = \texttt{k.alpha} = \texttt{qsignrank}(\texttt{alpha}/2, \texttt{n})$

$=$ the smallest $k$ such that $P_{\theta_0}(T_+(\theta_0) \leq k) \geq \alpha/2$.

Find $\alpha_0/2 = \texttt{psignrank}(\texttt{k.alpha} - 1, \texttt{n})$ and $\alpha_1/2 = \texttt{psignrank}(\texttt{k.alpha}, \texttt{n})$.

From $\alpha_0$ and $\alpha_1$ choose either the one closest to $\alpha$ or the one that is $\leq \alpha$.

Denote this choice by $\alpha_a$, our achievable level choice.

The corresponding $\texttt{k} = \texttt{k.alpha} - 1$ or $\texttt{k} = \texttt{k.alpha}$ denote by $k_a$.

Then $[A_{(k_a+1)}, A_{(M-k_a)})$ is the desired $(1 - \alpha_a)$-level confidence interval for $\theta$.

Since $P_{\theta_0}(A_{(i)} = x) = 0$ for any $x$, we can close the confidence interval.

# Using the Normal Approximation

Let $Z \sim \mathcal{N}(0,1)$ and $q_z = \texttt{qnorm}(1-\alpha/2)$ and solve

$$\frac{\alpha}{2} = P_{\theta_0}(T_+(\theta_0) \le k) = P_{\theta_0}(T_+(\theta_0) \le k+.5) \approx P\left(Z \le \frac{k+.5-n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}\right)$$

$$\Longrightarrow \; -q_z = \frac{k+.5-n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \quad \text{or} \quad k = \frac{n(n+1)}{4} - q_z\sqrt{\frac{n(n+1)(2n+1)}{24}} - .5$$

Since $k$ is typically not an integer, bracket it by the nearest integers $k_0$ and $k_1$.

Use the normal approximation to obtain the approximate levels associated with them and make your choice among them or some other integer $k$ nearby.

Denote your final choice by $k_a$ and the corresponding $\alpha$ by $\alpha_a$.

$[A_{(k_a+1)}, A_{(M-k_a)}]$ is the desired confidence interval for $\theta$

with approximate confidence level $(1-\alpha_a)$.

# Using Simulation

Instead of the previous two procedures we can also simulate a vector `T.vec` of $T_+(\theta_0)$ values as an approximation of the exact $T_+(\theta_0)$ null distribution.

Use the normal approximation to get us a ballpark $k$ (previous slide).
For a range of values around that $k$ use the simulated distribution of $T_+(\theta_0)$ to get estimates for $P_{\theta_0}(T_+(\theta_0) \leq k)$ via $\texttt{mean}(\texttt{T.vec} <= \texttt{k})$.
Then choose that $k$ (denoted $k_a$) for which this estimate comes closest to $\alpha/2$, possibly with the restriction $\leq \alpha/2$, if one wants to be conservative.

With that $k_a$ and corresponding $\alpha_a$ (estimated from the simulation) use $[A_{(k_a+1)}, A_{(M-k_a)}]$ as the desired confidence interval for $\theta$, with approximate confidence level $(1-\alpha_a)$

This is implemented in the provided function `W1.ci`.

# Calculating Sorted Walsh Averages

```
WalshAves <- function(x){
   Sums <- outer(x,x,"+") #creates n * n matrix of x[i]+x[j]
   w.aves <- sort(Sums[lower.tri(Sums,diag=TRUE)]/2)
   # lower.tri(A,diag=TRUE) for a matrix A creates a correponding
   # matrix with T in the lower triangle, including the diagonal,
   # and F elsewhere
   w.aves
}
```

The vector `w.aves` contains the ordered Walsh averages $A_{(i)}, i = 1, \ldots, M$

and we can use $[\texttt{w.aves}[\texttt{k}_\texttt{a} + 1], \texttt{w.aves}[\texttt{M} - \texttt{k}_\texttt{a}]]$ as our confidence interval.

# Illustration for $n = 20$ & $1 - \alpha = 0.90$

```
> qsignrank(.05,20)
[1] 61
> psignrank(61,20)
[1] 0.05269909
> psignrank(60,20)
[1] 0.0486536
> 1-2*0.0486536
[1] 0.9026928
> A=WalshAves(x0)
> A[61]
[1] 5.3
> A[150]
[1] 7.85
```

Thus $k_a = 60$ is the appropriate choice with achieved confidence level $0.9026928$ for the interval

$$[A_{(61)}, A_{(210-60)}] = [A_{(61)}, A_{(150)}]$$

same as best choice in text, Example 10.11.

However, the coverage probability $0.9083$ is slightly different since simulation was used in place of the exact calculation via `psignrank`.

For our previous sample `x0` we get

$$[A_{(61)}, A_{(150)}] = [5.3, 7.85].$$

# Using `W1.ci`

```
> W1.ci(x=x0,alpha=.1,draws=10000)
      k Lower Upper Coverage
[1,] 59  5.25  8.00   0.9182
[2,] 60  5.30  7.95   0.9117
[3,] 61  5.30  7.85   0.9041
[4,] 62  5.35  7.85   0.8945
[5,] 63  5.35  7.85   0.8858
```

with the middle $k = 61$ being the best choice.

Note: My $k_a + 1 = 61$ corresponds to Trosset's $k = 61$.