University of Washington

STATISTICS

Elements of Statistical Methods
2-Sample Location Problems (Ch 11)

Fritz Scholz

Spring Quarter 2010

May 18, 2010

The Basic 2-Sample Problem

It is assumed that we observe two independent random samples

$X_1, \ldots, X_{n_1}$ iid $\sim P_1$ and $Y_1, \ldots, Y_{n_2}$ iid $\sim P_2$ of continuous r.v.'s.

Let θ_1 and θ_2 be location parameters of P_1 and P_2, respectively.

For a meaningful comparison of such location parameters it makes sense to require them to be of the same type, i.e., they should both be means ($\theta_1 = \mu_1 = EX_i$ and $\theta_2 = \mu_2 = EY_j$) or both be medians ($\theta_1 = q_2(X_i)$ and $\theta_2 = q_2(Y_j)$).

$\Delta = \theta_1 - \theta_2$ measures the difference in location between P_1 and P_2, and such a difference is of main interest in the two sample problem.

The sample sizes n_1 and n_2 do not need to be the same, and there is no implied pairing between the X_i and the Y_j.

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?

Alzheimer's Disease (AD) Study

The study purpose is to investigate the performance of AD patients in a confrontation naming test relative to comparable non-AD patients.

60 mildly demented patients were selected, together with a "normal" control group of 60, more or less matched (as a group) in age and other relevant characteristics.

Each was given the Boston Naming Test (BNT): high score $=$ better performance.

1) Experimental unit is a person.

2) Experimental units belong to one of two populations:

AD patients and normal, comparable elderly persons.

3) One measurement (BNT score) per experimental unit.

4) $X_i =$ BNT score for AD patient i, and $Y_j =$ BNT score for control subject j.

$X_1, \ldots, X_{60}$ iid $\sim P_1$ and $Y_1, \ldots, Y_{n_2}$ iid $\sim P_2$,

$\Delta = \theta_1 - \theta_2 =$ parameter of interest. Test $H_0 : \Delta \geq 0$ vs $H_1 : \Delta < 0$.

Blood Pressure Medication

A drug is supposed to lower blood pressure.

$n_1 + n_2$ hypertensive patients are recruited for a double-blind study.

They are randomly divided into two groups of n_1 and n_2 patients,

the first group gets the drug the second group gets a look alike placebo.

Neither the patients nor the measuring technician know who gets what.

1) Experimental unit is a hypertensive patient.

2) Experimental unit belongs to one of two populations, a hypertensive population that gets the drug and a hypertensive population that gets the placebo.

   Randomization makes it possible to treat them as samples from the same set of hypertensive patients. Each patient could come from either population (trick).

3) Two measurements (before and after treatment) on each experimental unit.

Blood Pressure Medication (continued)

4) B_{1i} and A_{1i} are the before and after blood pressure measurements on patient i in the treatment group.

Similarly, B_{2i} and A_{2i} are the corresponding measurements for the control group.

$X_i = B_{1i} - A_{1i} =$ decrease in blood pressure for patient i in the treatment group

$Y_i = B_{2i} - A_{2i} =$ decrease in blood pressure for patient i in the control group

$X_1, \ldots, X_{n_1}$ iid $\sim P_1$ and $Y_1, \ldots, Y_{n_2}$ iid $\sim P_2$

Want to make inference about $\Delta = \theta_1 - \theta_2$. $\Delta > 0 \iff \theta_1 > \theta_2$.

To show that the drug lowers the blood pressure more than the placebo we want to test $H_0 : \Delta \leq 0$ against $H_1 : \Delta > 0$.

We reject H_0 when we have sufficient evidence for the drug's effectiveness.

Two Important 2-Sample Location Problems

1. Assume that both sampled populations are normal

$$X \sim P_1 = \mathcal{N}(\mu_1, \sigma_1^2) \qquad \text{and} \qquad Y \sim P_2 = \mathcal{N}(\mu_2, \sigma_2^2)$$

   This is referred to as the normal 2-sample location problem.

   Normal shape, with possibly different means and standard deviations.

2. The two sampled populations give rise to continuous random variables X and Y.

   We assume that the two poulations differ only in the location of their median, but are otherwise the same (same spread, same shape).

   This is referred to as the general two-sample shift problem.

   Same shape and variability, possible difference in locations (shift).

   Here the median is the natural location parameter (it always exists).

The Normal 2-Sample Location Problem

The plug-in estimator of $\Delta = \mu_1 - \mu_2$ naturally is

$$\hat{\Delta} = \hat{\mu}_1 - \hat{\mu}_2 = \bar{X} - \bar{Y}$$

It is unbiased, since

$$E\hat{\Delta} = E(\bar{X} - \bar{Y}) = E\bar{X} - E\bar{Y} = \mu_1 - \mu_2 = \Delta$$

$\hat{\Delta}$ is consistent, i.e., $\hat{\Delta} \xrightarrow{P} \Delta$ as $n_1, n_2 \longrightarrow \infty$

$\hat{\Delta}$ is asymptotically efficient, i.e., best possible within the normal model.

The Distribution of $\bar{X} - \bar{Y}$

In the context of the normal 2-sample problem we have from before

$$\bar{X} \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \qquad \text{and} \qquad \bar{Y} \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

Based on earlier results on sums of independent normal random variables

$$\Longrightarrow \quad \hat{\Delta} = \bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) = \mathcal{N}\left(\Delta, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

We address three different situations for testing and confidence intervals:

1) σ_1 and σ_2 known. Rare but serves as stepping stone, exact solution.

2) σ_1 and σ_2 unknown, but $\sigma_1 = \sigma_2 = \sigma$. Ideal but rare, exact solution.

3) σ_1 and σ_2 unknown, but not assumed equal. Most common case in practice, with good approximate solution.

Testing $H_0 : \Delta = \Delta_0$ Against $H_1 : \Delta \neq \Delta_0$ (σ_1, σ_2 Known)

Our previous treatment of the normal 1-sample problem suggests using

$$Z = \frac{\hat{\Delta} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{X} - \bar{Y} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$ as the appropriate test statistic.

Reject H_0 when $|Z| \geq q_z$, where q_z is the $1 - \alpha/2$ quantile of $\mathcal{N}(0,1)$.

When H_0 is true, then $Z \sim \mathcal{N}(0,1)$ and we get

$$P_{H_0}(|Z| \geq q_z) = \alpha$$

the desired probability of type I error. If $|z|$ denotes the observed value of $|Z|$ then the significance probability of $|z|$ is

$$\mathbf{p}(|z|) = P_{H_0}(|Z| \geq |z|) = 2\Phi(-|z|) = 2 * \texttt{pnorm}(-\texttt{abs}(\texttt{z}))$$

i.e., reject H_0 at level α when $\mathbf{p}(|z|) \leq \alpha$ or $|z| \geq q_z$.

Confidence Intervals for Δ (σ_1, σ_2 Known)

Again we can obtain $(1-\alpha)$-level confidence intervals for Δ as consisting of all values Δ_0 for which the corresponding hypothesis $H_0 = H_0(\Delta_0) : \Delta = \Delta_0$ cannot be rejected at significance level α. Or, more directly

$$\begin{aligned}
1-\alpha &= P(|Z| < q_z) = P_{\Delta_0}\left(|\hat{\Delta} - \Delta_0| < q_z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \\
&= P_{\Delta_0}\left(-q_z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \Delta_0 - \hat{\Delta} < q_z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \\
&= P_{\Delta_0}\left(\hat{\Delta} - q_z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \Delta_0 < \hat{\Delta} + q_z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)
\end{aligned}$$

with desired $(1-\alpha)$-level confidence interval $\quad \hat{\Delta} \pm q_z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

which contains Δ_0 if and only if $\mathbf{p}(|z|) > \alpha$ or $|z| < q_z$.

Example

Suppose we know $\sigma_1 = 5$ and observe $\bar{x} = 7.6$ with $n_1 = 60$ observations and have $\sigma_2 = 2.5$ and observe $\bar{y} = 5.2$ with $n_2 = 15$ observations.

For a 95% confidence interval for $\Delta = \mu_1 - \mu_2$ we compute

$$q_z = \texttt{qnorm}(.975) = 1.959964 \approx 1.96$$

$$\Longrightarrow \quad (7.6 - 5.2) \pm 1.96 \cdot \sqrt{\frac{5^2}{60} + \frac{2.5^2}{15}} = 2.4 \pm 1.79 = (0.61, 4.19)$$

Test $H_0 : \Delta = 0$ against $H_1 : \Delta \neq 0$ we find

$$z = \frac{(7.6 - 5.2) - 0}{\sqrt{5^2/60 + 2.5^2/15}} = 2.629$$

$$\mathbf{z}(|z|) = P_0(|Z| \geq |2.629|) = 2 * \texttt{pnorm}(-2.629) = 0.008563636 < 0.05$$

agreeing with the interval above not containing zero.

Estimating $\sigma^2 = \sigma_1^2 = \sigma_2^2$

We have two estimates for the common unknown variance σ^2

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2 \qquad \text{and} \qquad S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \bar{Y})^2$$

Either one could be used in the standardization of a T statistic (not efficient).

Rather use the appropriate weighted average, the pooled sample variance

$$S_P^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)} = \frac{(n_1-1)S_1^2}{(n_1-1)+(n_2-1)} + \frac{(n_2-1)S_2^2}{(n_1-1)+(n_2-1)}$$

$$\begin{aligned} ES_P^2 &= E\left(\frac{(n_1-1)S_1^2}{(n_1-1)+(n_2-1)}\right) + E\left(\frac{(n_2-1)S_2^2}{(n_1-1)+(n_2-1)}\right) \\ &= \frac{(n_1-1)ES_1^2}{(n_1-1)+(n_2-1)} + \frac{(n_2-1)ES_2^2}{(n_1-1)+(n_2-1)} \\ &= \frac{(n_1-1)\sigma^2}{(n_1-1)+(n_2-1)} + \frac{(n_2-1)\sigma^2}{(n_1-1)+(n_2-1)} = \sigma^2 \qquad \text{i.e., } S_P^2 \text{ is unbiased} \end{aligned}$$

More on S_P^2 when $\sigma_1^2 = \sigma_2^2$

S_P^2 is a consistent and efficient estimator of σ^2.

Recall

$$V_1 = \frac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi^2(n_1 - 1) \qquad \text{and} \qquad V_2 = \frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2(n_2 - 1)$$

S_1^2 and S_2^2 are independent.

Previous results about sums of independent χ^2 random variables

$$\Longrightarrow V_1 + V_2 = \frac{(n_1 + n_2 - 2)S_P^2}{\sigma^2} = \frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2(n_1 + n_1 - 2)$$

The independence of $\bar{X}$, $\bar{Y}$, S_1^2 and S_2^2 implies the independence of $\hat{\Delta}$ and S_P^2.

Standardization when $\sigma_1^2 = \sigma_2^2$

For testing $H_0 : \Delta = \Delta_0$ against $H_1 : \Delta \neq \Delta_0$ we use

$$T = \frac{\hat{\Delta} - \Delta_0}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_P^2}} = \frac{\hat{\Delta} - \Delta_0}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \sigma^2}} \frac{1}{\sqrt{S_P^2 \frac{1}{\sigma^2}}} = \frac{Z}{\sqrt{\frac{V_1+V_2}{n_1+n_2-2}}}$$

When $\Delta = \Delta_0$, then

$$Z = \frac{\hat{\Delta} - \Delta_0}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \sigma^2}} \sim \mathcal{N}(0,1) \qquad \text{and} \qquad V_1 + V_2 = \chi^2(n_1 + n_2 - 2)$$

Z and $V_1 + V_2$ are independent of each other and thus under $H_0 : \Delta = \Delta_0$

$T \sim t(n_1 + n_2 - 2)$, by definition of the Student t distribution.

Testing $H_0 : \Delta = \Delta_0$ Against $H_1 : \Delta \neq \Delta_0$ (σ_1, σ_2 Unknown)

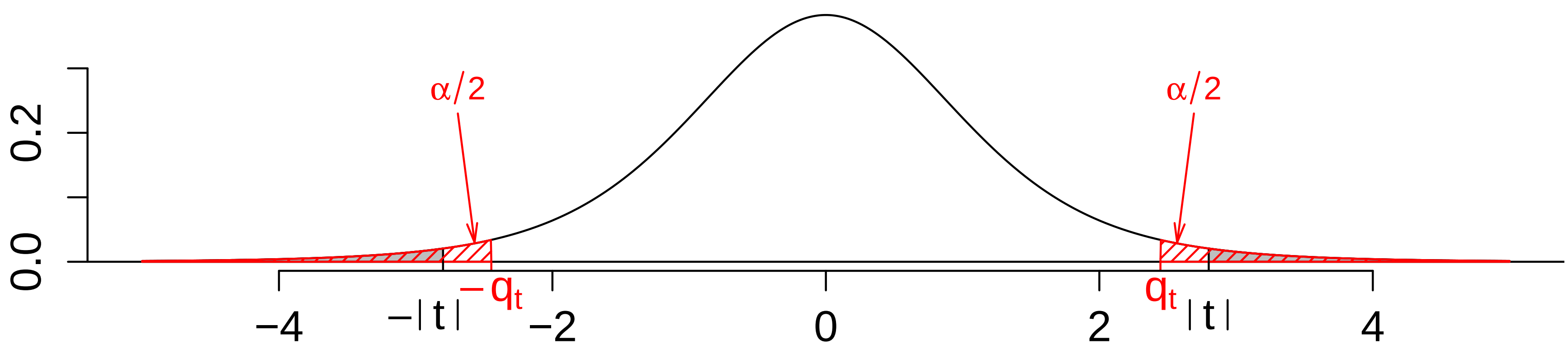
Of course, we reject H_0 when the observed value $|t|$ of $|T|$ is too large.

The significance probability of $|t|$ is

$$\mathbf{p}(|t|) = P_{H_0}(|T| \geq |t|) = 2P_{H_0}(T \leq -|t|) = 2 * \mathtt{pt}(-\mathtt{abs}(\mathtt{t}), \mathtt{n1} + \mathtt{n2} - 2)$$

Again note that $\mathbf{p}(|t|) \leq \alpha \iff |t| \geq q_t$,

where q_t is the $(1-\alpha/2)$-quantile of the $t(n_1+n_2-2)$ distribution.



Standard Error

The standard error of an estimator is its estimated standard deviation.

The standard error of $\bar{X}$ is $S/\sqrt{n}$ when σ is unknown and estimated by S.

When σ is known the standard error of $\bar{X}$ is $\sigma/\sqrt{n}$. (nothing to estimate)

The standard error of $\hat{\Delta}$ when $\sigma_1 = \sigma_2 = \sigma$ is unknown is $S_P\sqrt{1/n_1 + 1/n_2}$.

When $\sigma_1 = \sigma_2 = \sigma$ is known it is $\sigma\sqrt{1/n_1 + 1/n_2}$. (nothing to estimate)

Note that our test statistics in Z or T form always look like

$$\frac{\text{estimator} - \text{hypothesized mean of estimator}}{\text{standard error of the estimator}}$$

Confidence Interval for Δ when $\sigma_1^2 = \sigma_2^2$

Let $q_t = q_t(1-\alpha/2)$ denote the $(1-\alpha/2)$-quantile of $t(n_1+n_2-2)$.

Then for any Δ_0

$$\begin{aligned} 1-\alpha &= P_{\Delta_0}(|T| < q_t) = P_{\Delta_0}\left(\frac{|\hat{\Delta}-\Delta_0|}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right) S_P^2}} < q_t \right) \\ &= P_{\Delta_0}\left(\hat{\Delta} - q_t\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right) S_P^2} < \Delta_0 < \hat{\Delta} + q_t\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right) S_P^2} \right) \end{aligned}$$

$$\Longrightarrow \quad \hat{\Delta} \pm q_t \sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right) S_P^2} \qquad \text{(a random interval through } \hat{\Delta} \text{ and } S_P\text{)}$$

is a $(1-\alpha)$-level confidence interval for Δ.

Again it consists of all acceptable Δ_0 when testing $H_0 : \Delta = \Delta_0$ against $H_1 : \Delta \neq \Delta_0$.

Example (continued)

In our previous example instead of known σ's assume that $s_1 = 5$ and $s_2 = 2.5$.

Inspite of this discrepancy in s_1 and s_2 assume that $\sigma_1 = \sigma_2$.

$$s_P^2 = \frac{59 \cdot 5^2 + 14 \cdot 2.5^2}{59 + 14} = 21.40411$$

For a 95% confidence interval we compute $q_t = \texttt{qt}(.975, 73) = 1.992997 \approx 1.993$

$$\Longrightarrow \quad (7.6 - 5.2) \pm 1.993 \cdot \sqrt{21.40411 \cdot (1/60 + 1/15)} = 2.4 \pm 2.66 = (-0.26, 5.06)$$

For testing $H_0 : \Delta = 0$ against $H_1 : \Delta \neq 0$ we find

$$t = \frac{(7.6 - 5.2) - 0}{\sqrt{21.40411 \cdot (1/60 + 1/15)}} \approx 1.797$$

with $\quad \mathbf{p}(|t|) = P_0(|T| \geq |1.797|) = 2 * \texttt{pt}(-1.797, 73) = 0.07647185 > 0.05$

agreeing with the inference from the confidence interval.

σ_1, σ_2 Unknown, But Not Necessarily Equal

Under $H_0 : \Delta = \Delta_0$

$$Z = \frac{\hat{\Delta} - \Delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim \mathcal{N}(0,1)$$

$$\text{but} \qquad T_W = \frac{\hat{\Delta} - \Delta_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \sim \; ???$$

Welch (using Satterthwaite's approximation) argued that $T_W \approx t(\nu)$ with

$$\nu = \frac{[\sigma_1^2/n_1 + \sigma_2^2/n_2]^2}{\frac{(\sigma_1^2/n_1)^2}{n_1-1} + \frac{(\sigma_2^2/n_2)^2}{n_2-1}} \quad \text{estimated by} \quad \hat{\nu} = \frac{[s_1^2/n_1 + s_2^2/n_2]^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

provides good approximate significance probabilities and confidence intervals

when using $t(\hat{\nu})$ in the calculation of $\mathbf{p}(|t|)$ and q_t.

Fractional values of $\hat{\nu}$ present no problem in `qt` and `pt`.

This has been explored via simulation with good results in coverage

and false rejection rates for a wide range of (σ_1, σ_2)-scenarios.

Example (continued)

$$\hat{v} = \frac{[5^2/60 + 2.5^2/15]^2}{\frac{(5^2/60)^2}{60-1} + \frac{(2.5^2/15)^2}{15-1}} = 45.26027 \approx 45.26$$

For a 95% confidence interval we compute

$q_t = \texttt{qt}(.975, 45.26) = 2.013784 \approx 2.014$ and get

$$(7.6 - 5.2) \pm 2.014 \cdot \sqrt{5^2/60 + 2.5^2/15} = 2.4 \pm 1.84 = (0.56, 4.24)$$

For testing $H_0 : \Delta = 0$ against $H_1 : \Delta \neq 0$ we get

$$t_W = \frac{(7.6 - 5.2) - 0}{\sqrt{5^2/60 + 2.5^2/15}} \approx 2.629$$

$$\mathbf{p}(|t_W|) = P_0(|T_W| \geq |t_W|) \approx 2 * \texttt{pt}(-2.629, 45.26) = 0.01165687 < 0.05$$

Note the difference in results to when we assumed $\sigma_1 = \sigma_2$.

Comments on Using Student's t-Test

- if $n_1 = n_2$ then $t = t_W$ (verify by simple algebra).

- If the population variances (and hence the sample variances) tend to be approximately equal, then t and t_W tend to be approximately equal.

- If the larger sample is drawn from the population with the larger variance, then $|t|$ will tend to be less than $|t_W|$.

  All else equal, Student's t-test will give inflated significance probabilities.

- If the larger sample is drawn from the population with the smaller variance, then $|t|$ will tend to be larger than $|t_W|$.

  All else equal, Student's t-test will give understated significance probabilities.

Don't use Student's t-test, use t_W with $\hat{\nu}$ instead.

T_W Test for Large Samples

Again we can appeal to large sample results to claim

$$S_1^2 \xrightarrow{P} \sigma_1^2\,, \qquad S_2^2 \xrightarrow{P} \sigma_2^2 \qquad \text{and} \qquad T_W \approx \mathcal{N}(0,1)$$

These limiting results hold even when the sampled distributions are not normal, as long as the variances σ_1^2 and σ_2^2 exist and are finite.

Just use the $\mathcal{N}(0,1)$ distribution to calculate approximate significance probabilities for testing $H_0 : \Delta = \mu_1 - \mu_2 = \Delta_0$ against $H_1 : \Delta \neq \Delta_0$

$$P_{H_0}(|T_W| \geq |t_W|) \approx 2 * \texttt{pnorm}(-\texttt{abs}(\texttt{t}_\texttt{W}))$$

Use $q_z = \texttt{qt}(1 - \alpha/2)$ in place of q_t for $\approx (1-\alpha)$-level confidence intervals

$$\hat{\Delta} \pm q_z \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

2-Sample Shift Problem

Definition: A family of distributions $\mathcal{P} = \{P_\theta : \theta \in \mathcal{R}\}$ is a shift family iff

$$X \sim P_\theta \quad \Longrightarrow \quad X_0 = X - \theta \sim P_0 \quad \Longrightarrow \quad X = X_0 + \theta \sim P_\theta$$

Example: For fixed σ^2 the family $\{\mathcal{N}(\theta, \sigma^2) : \theta \in \mathcal{R}\}$ is a shift family.

The 2-sample location problem started with 2 normal shift families with common variance. We relaxed that to unequal variances, but kept the normality assumption.

Relax this in the other direction:

Let $\mathcal{P}_0$ denote the family of all continuous distributions on $\mathcal{R}$ with median 0.

For any $P_0 \in \mathcal{P}_0$ we take as our shift family $\mathcal{P}_{P_0} = \{P_\theta : \theta \in \mathcal{R}, P_{\theta=0} = P_0\}$.

The extra P_0 in the notation indicates which distribution shape is shifted around.

We observe two independent random samples

$$X_1, \ldots, X_{n_1} \sim P_{\theta_1} \in \mathcal{P}_{P_0} \qquad \text{and} \qquad Y_1, \ldots, Y_{n_2} \sim P_{\theta_2} \in \mathcal{P}_{P_0}$$

They arise from the same shape P_0 with possibly different shifts.

We wish to make inferences about $\Delta = \theta_1 - \theta_2$.

Testing in the 2-Sample Shift Problem

Test $H_0 : \Delta = 0$ against $H_1 : \Delta \neq 0$. The case $H_0 : \Delta = \Delta_0$ against $H_1 : \Delta \neq \Delta_0$ can be reduced to the previous case by subtracting Δ_0 from the X-sample:

$$X_i' = X_i - \Delta_0 \sim P_{\theta_1 - \Delta_0} = P_{\theta_1'}$$

$$\Delta' = \theta_1' - \theta_2 = \theta_1 - \Delta_0 - \theta_2 = 0 \Longleftrightarrow \Delta = \theta_1 - \theta_2 = \Delta_0$$

Since the distributions of the X's and Y's are continuous, they have $n_1 + n_2 = N$ distinct values with probability 1.

Rank the (assumed distinct) values in the pooled sample of N observations, e.g., X_3 gets rank 5 iff X_3 is the 5^{th} in the ordered sequence of N pooled sample values, Y_6 gets rank 1 if it is is the smallest of all N pooled values.

Let T_x be the sum of the X-sample ranks and T_y be the sum of the Y-sample ranks.

$$T_x + T_y = \sum_{k=1}^{N} k = N(N+1)/2$$

Null Distribution of the Wilcoxon Rank Sum

Under $H_0 : \Delta = 0$ both samples come from the same continuous distribution P_θ.

We may view the sampling process as follows: Select N values $Z_1, \ldots, Z_N$ iid P_θ.

Out of $Z_1, \ldots, Z_N$ randomly select n_1 as the X-sample, the remainder being the Y-sample.

All $\binom{N}{n_1}$ X-sample selections (and corresponding rank sets) are equally likely.

Some of these rank sets will give the same T_x.

By enumerating all rank sets (choices of n_1 numbers from $1, 2, \ldots, N$) it is in principle possible to get the null distribution of T_x.

Naturally, we would reject H_0 when T_x is either too high or too low.

This is the Wilcoxon Rank Sum Test.

Significance probabilities are computed from the null distribution of T_x.

Using `rank` and `combn`

```
> x <-c(9.1,8.3); y <- c(11.9,10.0,10.5,11.3)
> rank(c(x,y)) # ranking the pooled samples
[1] 2 1 6 3 4 5   # Tx = 2+1 = 3
> combn(1:6,2) # all sets of 2 taken from 1,2,...,6
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
[1,]    1    1    1    1    1    2    2    2
[2,]    2    3    4    5    6    3    4    5
     [,9] [,10] [,11] [,12] [,13] [,14] [,15]
[1,]    2     3     3     3     4     4     5
[2,]    6     4     5     6     5     6     6
> combn(1:6,2,FUN=sum) # sum of those 15 = choose(6,2) sets of 2
 [1]  3  4  5  6  7  5  6  7  8  7  8  9  9 10 11
> table(combn(1:6,2,FUN=sum)) # tabulate distict sums with frequencies
 3  4  5  6  7  8  9 10 11
 1  1  2  2  3  2  2  1  1  # null distribution is symmetric around 7
```

Significance probability is $2 \cdot 1/15 = 0.1333333 > 0.05$

$$\texttt{mean}(\texttt{abs}(\texttt{combn}(1:6,2,\texttt{FUN}=\texttt{sum})-7) >= \texttt{abs}(3-7)) = 0.1333333$$

Symmetry of T_x Null Distribution

The null distribution of T_x is symmetric around $n_1(N+1)/2 = E_0 T_x$

with variance $\mathrm{var}_0(T_x) = n_1 n_2 (N+1)^2/12$.

Let R_i and $R'_i = N+1-R_i$ be the X_i rankings in increasing and decreasing order.

$$T_x = \sum_{i=1}^{n_1} R_i \qquad \text{and} \qquad T'_x = \sum_{i=1}^{n_1} R'_i = \sum_{i=1}^{n_1} (N+1-R_i) = n_1(N+1) - T_x$$

T_x and T'_x have the same distribution, being sums of n_1 ranks randomly chosen

from $1, 2, \ldots, N$. Using $\stackrel{\mathcal{D}}{=}$ to denote equality in distribution we have

$$T_x \stackrel{\mathcal{D}}{=} T'_x = n_1(N+1) - T_x \quad \Longrightarrow \quad T_x - \frac{n_1(N+1)}{2} \stackrel{\mathcal{D}}{=} \frac{n_1(N+1)}{2} - T_x$$

$$\begin{aligned} P_0\left(T_x = \frac{n_1(N+1)}{2} + x\right) &= P_0\left(T_x - \frac{n_1(N+1)}{2} = x\right) \\ &= P_0\left(\frac{n_1(N+1)}{2} - T_x = x\right) = P_0\left(T_x = \frac{n_1(N+1)}{2} - x\right) \end{aligned}$$

The Mann-Whitney Test Statistic

Let $X_{(1)} < \ldots < X_{(n_1)}$ be the $\vec{X}$ sample in increasing order.

Let R_k denote the rank of $X_{(k)}$ in the pooled sample. Then

$$R_k = k + \#\left\{Y_j : Y_j < X_{(k)}\right\}$$

and

$$T_x = \sum_{k=1}^{n_1} R_k = \sum_{k=1}^{n_1} k + \sum_{k=1}^{n_1} \#\left\{Y_j : Y_j < X_{(k)}\right\} = \frac{n_1(n_1+1)}{2} + W_{YX}$$

where $n_1(n_1+1)/2 = 1 + \ldots + n_1$ is the smallest possible value for T_x and

$$W_{YX} = \#\left\{(X_i, Y_j) : Y_j < X_i\right\}$$ is the Mann-Whitney test statistic

Since $T_x = W_{YX} - \frac{n_1(n_1+1)}{2}$ the test can be carried out using either test statistic.

The value set of W_{YX} is 0 (no $Y_j <$ any X_i),$1, 2, \ldots, n_1 \cdot n_2$ (all $Y_j <$ all X_i).

The distribution of W_{YX} is symmetric around $n_1 \cdot n_2/2$.

Four Approaches to the Null Distribution

1. Null distribution using explicit enumeration via `combn`

2. Null distribution using the R function `pwilcox`

3. Null distribution using simulation

4. Null distribution using normal approximation

Null Distribution Using `combn`

1. Exact null distribution of T_x via `combn`, for `choose(n1+n2,n1)` not too large

$$P_0(T_x \le \texttt{tx}) = \texttt{mean}(\texttt{combn}(\texttt{n1}+\texttt{n2},\texttt{n1},\texttt{FUN}=\texttt{sum}) <= \texttt{tx})$$

$$\begin{aligned}\mathbf{p}(t_x) &= P_0\left(\left|T_x - \frac{n_1(N+1)}{2}\right| \ge \left|t_x - \frac{n_1(N+1)}{2}\right|\right) \\ &= 2*\texttt{mean}(\texttt{abs}(\texttt{combn}(\texttt{n1}+\texttt{n2},\texttt{n1},\texttt{FUN}=\texttt{sum})-\texttt{n1}*(\texttt{N}+\texttt{1})/\texttt{2}) \\ &\qquad <= -\texttt{abs}(\texttt{tx}-\texttt{n1}*(\texttt{N}+\texttt{1})/\texttt{2}))\end{aligned}$$

Null Distribution Using `pwilcox`

2. R has a function `pwilcox` that efficiently calculates

$$
\begin{aligned}
P_0(T_x \le \mathtt{tx}) &= P_0(W_{YX} \le \mathtt{tx} - \mathtt{n1} * (\mathtt{n1} + 1)/2) \\
&= \mathtt{pwilcox}(\mathtt{tx} - \mathtt{n1} * (\mathtt{n1} + 1)/2, \mathtt{n1}, \mathtt{n2})
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{p}(t_x) &= P_0\left(\left|T_x - \frac{n_1(N+1)}{2}\right| \ge \left|t_x - \frac{n_1(N+1)}{2}\right|\right) \\
&= 2P_0\left(T_x - \frac{n_1(N+1)}{2} \le -\left|t_x - \frac{n_1(N+1)}{2}\right|\right) \\
&= 2 * \mathtt{pwilcox}(-\mathtt{abs}(\mathtt{tx} - \mathtt{n1} * (\mathtt{N} + 1)/2) \\
&\qquad - \mathtt{n1} * (\mathtt{n1} + 1)/2 + \mathtt{n1} * (\mathtt{N} + 1)/2, \mathtt{n1}, \mathtt{n2})
\end{aligned}
$$

Null Distribution Using Simulation

3. It is easy to simulate random samples of `n1` ranks taken from $1, 2, \ldots, n1 + n2$

```
N <- n1+n2; Nsim <- 10000; Tx.sim <- numeric(Nsim)
for(i in 1:Nsim){ Tx.sum[i] <- sum(sample(1:N,n1))}
```

$$\mathbf{p}(t_x) = P_0\left(\left|T_x - \frac{n_1(N+1)}{2}\right| \geq \left|t_x - \frac{n_1(N+1)}{2}\right|\right)$$

$$\approx \mathtt{mean}(\mathtt{abs}(\mathtt{Tx.sim} - \mathtt{n1} * (\mathtt{N} + 1)/2) >= \mathtt{abs}(\mathtt{tx} - \mathtt{n1} * (\mathtt{N} + 1)/2))$$

The accuracy of this approximation is completely controlled by `Nsim` (LLN).

See also the function `W2.p.sim <- function(n1,n2,tx,draws=1000){...}`

given as part of shift.R on the text book web site. (`draws=Nsim`)

Null Distribution Using Normal Approximation

4. The normal approximation, good for larger sample sizes n_1 and n_2, gives

$$
\begin{aligned}
P_0(T_x \le \texttt{tx}) &= P_0\left(\frac{T_x - E_0T_x}{\sqrt{\mathrm{var}_0(T_x)}} \le \frac{\texttt{tx} - E_0T_x}{\sqrt{\mathrm{var}_0(T_x)}}\right) \approx \Phi\left(\frac{\texttt{tx} - E_0T_x}{\sqrt{\mathrm{var}_0(T_x)}}\right) \\
&= \Phi\left(\frac{\texttt{tx} - \texttt{n1}(\texttt{N}+\texttt{1})/\texttt{2}}{\sqrt{\frac{\texttt{n1n2}(\texttt{N}+\texttt{1})^{\texttt{2}}}{\texttt{12}}}}\right) \\
&= \texttt{pnorm}((\texttt{tx} - \texttt{n1} * (\texttt{N}+\texttt{1})/\texttt{2})/\texttt{sqrt}(\texttt{n1} * \texttt{n2} * (\texttt{N}+\texttt{1})\hat{}\,\texttt{2}/\texttt{12}))
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{p}(t_x) &= P_0\left(\left|T_x - \frac{n_1(N+1)}{2}\right| \ge \left|t_x - \frac{n_1(N+1)}{2}\right|\right) \\
&\approx \texttt{2} * \texttt{pnorm}(-\texttt{abs}(\texttt{tx} - \texttt{n1} * (\texttt{N}+\texttt{1})/\texttt{2})/\texttt{sqrt}(\texttt{n1} * \texttt{n2} * (\texttt{N}+\texttt{1})\hat{}\,\texttt{2}/\texttt{12}))
\end{aligned}
$$

See also `W2.p.norm <- function(n1,n2,tx) {...}`

given as part of shift.R on the text book web site.

Example

For our previous example, where we illustrated complete enumeration via `combn`,

we had $$P_0(|T_x - 7| \geq |3-7|) = \frac{2}{15} \approx 0.1333$$

Using `pwilcox` we get

```
> 2*pwilcox(-abs(3-2*7/2)-2*3/2+2*7/2,2,4)
[1] 0.1333333
```

In 5 applications of `W2.p.sim(2,4,3)` we get: 0.135, 0.114, 0.110, 0.114, 0.151,

`W2.p.norm(2,4,3)=0.1051925`, reasonably close for such small sample sizes.

The Case of Ties

When there are ties in the pooled sample we add tiny random amounts to the sample values to break the ties.

Then compute the significance probability for each such breaking of ties.

Repeat this process many times and average all these significance probabilities.

Alternatively, compute the significance probabilities for the breaking of ties least (or most) in favor of H_0 and base decisions on these two extreme cases.

We could also use complete enumeration or simulation while working with the pooled vector of midranks returned by `rz <- rank(c(x,y))` and then use `combn(rz,n1,FUN=sum)` to get all possible midrank sums.
From that get the exact significance probability, conditional on the tie pattern.
Observations tied at ranks 6, 7, 8 and 9 get midrank 7.5 each.

Example with Ties

$\vec{x}$	6.6	14.7	15.7	11.1	7.0	9.0	9.6	8.2	6.8	7.2
$\vec{y}$	4.2	3.6	2.3	2.4	13.4	1.3	2.0	2.9	8.8	3.8

Test $H_0 : \Delta = 3$ versus $H_1 : \Delta \neq 3$ at $\alpha = 0.05$. Replacing x_i by $x_i' = x_i - 3$ and assigning the pooled ranks

```
> c(x0-3,y0)
 [1]  3.6 11.7 12.7  8.1  4.0  6.0  6.6  5.2  3.8  4.2
[11]  4.2  3.6  2.3  2.4 13.4  1.3  2.0  2.9  8.8  3.8
>  rz <- rank(c(x0-3,y0)); rz
 [1]  6.0 18.0 19.0 16.0 10.0 14.0 15.0 13.0  8.5 11.5
[11] 11.5  7.0  3.0  4.0 20.0  1.0  2.0  5.0 17.0  8.5
> sort(c(x0-3,y0))
 [1]  1.3  2.0  2.3  2.4  2.9  3.6  3.6  3.8  3.8  4.0
[11]  4.2  4.2  5.2  6.0  6.6  8.1  8.8 11.7 12.7 13.4
```

We recognize three pairs of tied values $(3.6, 3.6)$, $(3.8, 3.8)$ and $(4.2, 4.2)$, receiving respective midranks $(6, 7)$ (???), $(8.5, 8.5)$ and $(11.5, 11.5)$.

What Happened to Midrank 6.5?

```
> x0[1]-3==3.6
[1] FALSE
> x0[1]-3-3.6
[1] -4.440892e-16
> y0[2]==3.6
[1] TRUE
> round(x0[1]-3,1)==3.6
[1] TRUE
> round(x0[1]-3,1)-3.6
[1] 0
> rank(c(round(x0-3,1),y0))
 [1]  6.5 18.0 19.0 16.0 10.0 14.0 15.0 13.0  8.5 11.5
[11] 11.5  6.5  3.0  4.0 20.0  1.0  2.0  5.0 17.0  8.5
```

Computer arithmetic is done in binary form and it sometimes produces surprises.

If we are aware of this we can avoid it as above by a proper rounding procedure.

Significance Probability

Using the fix we have a midrank sum of

$$t_x = \mathtt{sum(rank(c(round(x0-3,1),y0))[1:10])} = 131.5$$

Depending on how the ties are broken, t_x could have taken values as low as 130 and as high as 133.

Comparing 130 and 133 with $n_1(N+1)/2 = 105$ it would give us

$$\begin{aligned} \mathbf{p}(130) &= P_0(|T_x - 105| \geq |130-105|) = 2 \cdot P_0(T_x \leq 105-25) = 2 \cdot P_0(T_x \leq 80) \\ &= \mathtt{2*pwilcox(80-n1*(n1+1)/2,n1,n2)} = 0.06301284 \end{aligned}$$

$$\begin{aligned} \mathbf{p}(133) &= P_0(|T_x - 105| \geq |133-105|) = 2 \cdot P_0(T_x \leq 105-28) = 2 \cdot P_0(T_x \leq 77) \\ &= \mathtt{2*pwilcox(77-n1*(n1+1)/2,n1,n2)} = 0.03546299 \end{aligned}$$

Not sufficiently strong evidence against H_0 at level $\alpha = 0.05$.

Using `W2.p.sim` and `W2.p.norm`

Instead of using `pwilcox` we could, as suggested in the text, also use `W2.p.sim` or `W2.norm` to get the significance probabilities of `tx=130` and `tx=133`.

```
> W2.p.sim(10,10,130,10000)
[1] 0.0662
> W2.p.sim(10,10,133,10000)
[1] 0.037
> W2.p.norm(10,10,130)
[1] 0.0640221
> W2.p.norm(10,10,133)
[1] 0.03763531
```

Note the quality of these two approximations.

Significance Probability

Since $\texttt{choose}(20,10) = 184756$ we can dare to compute all possible midrank sums of 10 taken from the above set of 20 midranks.

We want to assess the significance probability of $t_x = 131.5$ in relation to the average of all pooled midranks. That average is again 105. Thus

$$\begin{aligned}\mathbf{p}(131.5) &= \texttt{mean}(\texttt{abs}(\texttt{combn}(\texttt{rz}, 10, \texttt{FUN} = \texttt{sum}) - 105) >= \texttt{abs}(131.5 - 105)) \\ &= 0.04523804\end{aligned}$$

This is barely significant at level $\alpha = 0.05$, i.e., we should reject $H_0 : \Delta = 3$ when testing it against $H_1 : \Delta \neq 3$.

On the other hand, averaging many simulated random breaking of ties we get

```
> W2.p.ties(x0,y0,3,100000)
[1] 0.05412
```

Point Estimation of Δ

Following our previous paradigm, we should use that value Δ_0 as estimate of Δ which makes the hypothesis $H_0 : \Delta = \Delta_0$ least rejectable.

This would be that Δ_0 for which the rank sum $T_x(\Delta_0)$ of the $X_i' = X_i - \Delta_0$ comes closest to the center $n_1(N+1)/2$ of the null distribution.

recall the equivalent form $T_x(\Delta_0) = W_{YX'} - \frac{n_1(n_1+1)}{2} = W_{Y,X-\Delta_0} - \frac{n_1(n_1+1)}{2}$

$$\begin{aligned} W_{Y,X-\Delta_0} &= \#\{(X_i', Y_j) : Y_j < X_i'\} = \#\{(X - \Delta_0, Y_j) : Y_j < X_i - \Delta_0\} \\ &= \#\{(X_i, Y_j) : \Delta_0 < X_i - Y_j\} = \frac{n_2 n_2}{2} = \text{center of the } W \text{ null distribution} \end{aligned}$$

when $\Delta_0 = \text{median}\{X_i - Y_j : i = 1, \ldots, n_1, j = 1, \ldots, n_2\}$.

$\widetilde{\Delta} = \text{median}(X_i - Y_j)$ is also known as the Hodges-Lehmann estimator of Δ.

As in interesting contrast compare it with this alternate form of $\bar{X} - \bar{Y}$

$$\hat{\Delta} = \bar{X} - \bar{Y} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i - \frac{1}{n_2} \sum_{j=1}^{n_2} Y_j = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} (X_i - Y_j)$$

Computation of $\widetilde{\Delta}$

$\widetilde{\Delta}$ can be computed by using `W2.hl`, see shift.R on the text book web site.

```
> W2.hl(x0,y0)
[1] 5.2
```

or alternatively

```
> median(outer(x0,y0,"-"))
[1] 5.2
```

where `outer` produces a matrix of all pairwise operations `"-"` on the input vectors.

```
> outer(1:3,3:2,"-")
     [,1] [,2]
[1,]   -2   -1
[2,]   -1    0
[3,]    0    1
```

Set Estimation of Δ

By the well practiced paradigm, the confidence set consists of all those Δ_0 for which we cannot reject $H_0 : \Delta = \Delta_0$ at level α when testing it against $H_1 : \Delta \neq \Delta_0$.

We accept H_0 whenever $k \leq W_{Y,X-\Delta_0} \leq n_1 n_2 - k = M - k$, where the integers $k = 1, 2, 3, \ldots \leq (M+1)/2$ govern the achievable significance levels α_k

$$\alpha_k = 2P_{\Delta_0}(W_{Y,X-\Delta_0} \leq k-1) = 2P_0(W_{YX} \leq k-1)$$

With probability 1 all the $M = n_1 \cdot n_2$ differences $D_{ij} = X_i - Y_j$ are distinct.

Denote their ordered sequence by $D_{(1)} < D_{(2)} < \ldots < D_{(M)}$ and note

$$k \leq W_{Y,X-\Delta_0} = \#\{(X_i - \Delta_0, Y_j) : Y_j < X_i - \Delta_0\} = \#\{(X_i, Y_j) : \Delta_0 < X_i - Y_j\}$$

$\Longleftrightarrow D_{(M-k+1)} > \Delta_0$. Similarly, $W_{Y,X-\Delta_0} \leq a \Longleftrightarrow D_{(M-a)} \leq \Delta_0$. Thus

(using $a = M - k$) $\quad W_{Y,X-\Delta_0} \leq M - k \Longleftrightarrow D_{(k)} \leq \Delta_0$

$$[D_{(k)}, D_{(M-k+1)}] \quad \text{is our } (1-\alpha_k)\text{-level confidence interval for } \Delta.$$

We can use exact calculation (`pwilcox`), simulation approximation or normal approximation to get the appropriate (k, α_k) combination.

Confidence Interval for Δ: Example

For the previous example of 10 and 10 obs. get a 90% confidence interval for Δ.

$k = \texttt{qwilcox}(1-0.05/2, 10, 10) = 72 =$ smallest k such $P_0(W_{YX} \leq k) \geq 0.95$.

In fact, $\texttt{pwilcox}(72, 10, 10) = 0.9553952$ and $\texttt{pwilcox}(71, 10, 10) = 0.9474388$

$$P_0(W_{YX} \geq 73) = 1 - 0.9553952 = 0.0446048 = P_0(W_{YX} \leq 27) = \alpha_{28}/2$$

$[D_{(28)}, D_{(73)}] = [3.4, 7.2]$ is a $1 - \alpha_{28} = 0.9108$ level confidence interval for Δ

$$P_0(W_{YX} \geq 72) = 1 - 0.9474388 = 0.0525612 = P_0(W_{YX} \leq 28) = \alpha_{29}/2$$

$[D_{(29)}, D_{(72)}] = [3.4, 7.0]$ is a $1 - \alpha_{29} = 0.8949$ level confidence interval for Δ

We apparently have ties among the D_{ij}, since $D_{(28)} = D_{(29)} = 3.4$.

Using `W2.ci` in Example

The text explains how to use the normal approximation for finding a range of k values such that $P_0(W_{YX} \leq k) \approx \alpha/2$ and uses simulation to get a better approximation to the actual $P_0(W_{YX} \leq k)$ for each of these k values.

This is implemented in the function `W2.ci` provided in shift.R.

```
> W2.ci(x0,y0,.1,10000)
      k Lower Upper Coverage
[1,] 27   3.2   7.3   0.9261
[2,] 28   3.4   7.2   0.9115
[3,] 29   3.4   7.0   0.8944
[4,] 30   3.6   6.9   0.8728
[5,] 31   3.7   6.9   0.8615
```

This agrees fairly well with our intervals based on exact calculations.

How to Deal With Rounding

In our example we observed ties.

Such ties often are due to rounding data that are intrinsically of continuous nature.

Rounding confines the reported data to a grid of values consisting of $0, \pm\varepsilon, \pm 2\varepsilon, \pm 3\varepsilon, \ldots$.

Let X' denote the continuous data value corresponding to the rounded value X.

$$\Longrightarrow \quad |X - X'| \le \varepsilon/2 \quad \text{and} \quad |D_{ij} - D'_{ij}| = |X_i - Y_j + X'_i - Y'_j| \le \varepsilon/2 + \varepsilon/2 = \varepsilon$$

$$\Longrightarrow \quad D_{(i)} - \varepsilon \le D'_{(i)} \le D_{(i)} + \varepsilon$$

Since our confidence procedure is correct in terms of truly continuous data, i.e.,

$P_\Delta(\Delta \in [D'_{(k)}, D'_{(M-k+1)}]) = 1 - \alpha_k$, we can view $[D_{(k)} - \varepsilon, D_{(M-k+1)} + \varepsilon]$

as a proper confidence interval for Δ with confidence level

$$P_\Delta(\Delta \in [D_{(k)} - \varepsilon, D_{(M-k+1)} + \varepsilon]) \ge P_\Delta(\Delta \in [D'_{(k)}, D'_{(M-k+1)}]) = 1 - \alpha_k$$