STAT 135, Concepts of Statistics

Helmut Pitters

Comparing two populations - independent samples

Department of Statistics
University of California, Berkeley

April 12, 2017

Comparing two populations.

New treatments[1] are proposed continually:

- treatments for breast cancer, multiple sclerosis
- drugs,
- surgical techniques,
- medicine for pain relief,
- fertilizers,
- teaching methods,
- etc.

as well as new formulas for

- making bread,
- detergents,
- alloys,
- etc.

to improve some aspect of live (increase productivity/profit, reduce pain, decrease waiting time, enhance taste, etc.)

[1] Treatment being understood in a very broad sense of the word.

Comparing two populations.

Before spending money on a new treatment we surely want to convince ourselves of its efficacy.

Want to *assess effects of treatment* by comparing population of *responses* of treated individuals (*treatment group*) with population of untreated individuals (*control group*). Usually, can only compare samples of different populations (examining entire population is unfeasible).

Comparing two populations.

Remark

Won't consider categorial data (e.g. little, moderate, complete relief from pain).
Numerical observations (e.g. blood pressure, yield, waiting time) will differ from individual to individual. Treatment may increase some responses and decrease others. However, we are interested in the *average response*.

Comparing two populations.

Controlled experiments vs. observational studies.
Experiments are studies where a researcher assigns treatments to cases. When this assignment is done in a random fashion, it is called a *randomized experiment*.
A study in which the researcher collects data without interfering with how the data arise is called an *observational study*. From observational studies one can possibly infer association, but not causation.
Not always possible to have randomized controls: e.g. in order to asses effect of long-term smoking on health one cannot possibly ask people to smoke for study purposes.

Summarizing data.

I Comparing two populations via independent samples

Comparing two populations.

Example (Age of customers)

Company with department stores in Atlanta, Georgia has stores in inner city and in suburban shopping centers. Customers in different shops are sampled randomly and their age is recorded.

Table: Age of customers in inner city and suburban stores.

store type	sample size	sample mean (age)	sample SD (age)
inner city	$n_1 = 60$	$\bar{x}_1 = 40$yrs	$s_1 = 9$yrs
suburban	$n_2 = 80$	$\bar{x}_2 = 35$yrs	$s_2 = 10$yrs

Table: Two types of errors in hypothesis testing.

Question. Are customers shopping in inner city stores older than customers in suburban stores?
How could one answer such a question?

Comparing two populations.

Consider n_1 independent samples

$$X_1, \ldots, X_{n_1} \sim \mathcal{N}(\mu_X, \sigma_X^2)$$

from the treatment population, and n_2 independent samples

$$Y_1, \ldots, Y_{n_2} \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

from the population of controls.

As a measure for treatment effect we would like to know average difference

$$\mu_X - \mu_Y$$

of treatments and controls.
As an estimator for this difference we use

$$\bar{X}_{n_1} - \bar{Y}_{n_2}.$$

Comparing two populations.

Since $\bar{X}_{n_1} - \bar{Y}_{n_2}$ is a linear combination of independent normals, we find

$$\bar{X}_{n_1} - \bar{Y}_{n_2} \sim \mathcal{N}(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}).$$

Suppose σ_X^2, σ_Y^2 are known. Then the statistic

$$Z := \frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}} \sim \mathcal{N}(0, 1)$$

follows a standard normal distribution.
In particular, this allows us to work out that

$$\bar{X}_{n_1} - \bar{Y}_{n_2} \pm z(\frac{\alpha}{2})\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}$$

is a $(1-\alpha)$ confidence intervals for $\mu_X - \mu_Y$.[2]

[2]We denote by $z(\alpha)$ the (100α)th percentile of the standard normal distribution, i.e. $\Phi(z(\alpha)) = \alpha$.

Comparing two populations.

Example (Age of customers)

We may not be willing to model ages of customers by normal distribution. However, since sample sizes

$$n_1 = 60 \qquad n_2 = 80$$

are large, can approximate $\bar{X}_{n_1}$ and $\bar{Y}_{n_2}$ by normal distributions (due to CLT).
Using s_1, s_2 as approximations for σ_X, σ_Y we find the 90% confidence interval for $\mu_x - \mu_Y$ to be

$$40 - 35 \pm z(0.05)\sqrt{\frac{9^2}{60} + \frac{10^2}{80}} = 5 \pm 1.645 \times 1.61 = 5 \pm 2.65.$$

Comparing two populations.

If sample sizes are small[3] a normal approximation is not accurate in general, and one has to work harder in order to find the distribution of $X_{n_1} - Y_{n_2}$.
We're going to use the *pooled sample variance*

$$s_p^2 := \frac{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2}{n_1 + n_2 - 2},$$

as an estimator for the variance of $X_{n_1} - Y_{n_2}$. Here

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2$$

denotes the sample variance of the Y's and s_Y^2 is defined in complete analogy.

[3] I.e. smaller than 30.

Comparing two populations.

Notice that

$$s_p^2 = \frac{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2}{n_1 + n_2 - 2},$$

is a weighted sum of the variances, where each variance is weighed according to the corresponding sample size.

Remark

Imagine sampling from two populations, e.g.

$$n_1 = 1000, s_1^2 = 50 \qquad n_2 = 2, s_2^2 = 1.$$

Surely, the overall variance

$$\bar{X}_{1000} - \bar{Y}_2$$

will be much closer to $s_1^2 = 50$ than to $s_2^2 = 1$, as the contribution in variance from $\bar{Y}_2$ is comparatively negligible.

Comparing two populations.

With the methods we developed in the first chapter (distributions derived from the normal distribution), it is not hard to show the following

Fact

Provided the variances $\sigma_X^2 = \sigma_Y^2$ in the treatment and control populations are equal, the statisic

$$t := \frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - (\mu_X - \mu_Y)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

follows Student's t distribution with $n_1 + n_2 - 2$ df.

As before, this implies that we can work out confidence intervals for $\mu_X - \mu_Y$

Comparing two populations.

Example (Household incomes of neighborhoods)

Urban planning department interested in difference between average household income for two neighborhoods. The table summarizes data collected from randomly sampled households.

	neighborhood X	neighborhood Y
sample size	$n_1 = 8$	$n_2 = 12$
s. mean (income)	$\bar{x}_1 = \$75,000$	$\bar{x}_2 = \$82,000$
s. SD	$s_X = \$2,000$	$s_Y = \$1,800$

Table: Household incomes in two neighborhoods.

Question: Construct a 95% confidence interval for $\mu_X - \mu_Y$.

Comparing two populations.

Example (Household incomes of neighborhoods)

We assume that incomes are normally distributed[4] with equal variances $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ for both neighborhoods. Since

$$t_{18}(0.025) = 2.101 \qquad s_p^2 = \frac{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2}{n_1 + n_2 - 2} = 1,880.31$$

we find a 95% confidence interval for $\mu_X - \mu_Y$ as

$$\begin{aligned} &\bar{x}_1 - \bar{x}_2 \pm t_{18}(0.025) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\ &\quad = -7,000 \pm 2.101 \times 43.36 \times 0.456 \\ &\quad = -7.000 \pm 41.54 \end{aligned}$$

[4] We cannot apply CLT here due to small sample sizes.

Comparing two populations.

Hypothesis tests. Since we know that the statistic t follows Student's t distribution, we can carry out hypothesis tests.
Only want to introduce new treatment if there is strong evidence that it has an effect. In other words, unless there is strong evidence in the data that treatment has an effect, we assume null hypothesis

$$H_0 \colon \mu_X = \mu_Y$$

that on average the treatment has no effect.
When comparing two populations, alternatives usually are of the form

$$\begin{cases} \mu_X \neq \mu_Y & \text{(two-sided alternative)} \\ \mu_X > \mu_Y & \text{(one-sided alternative)} \\ \mu_X < \mu_Y & \text{"} \qquad \text{"}. \end{cases}$$

Comparing two populations.

Hypothesis tests. Suppose we study the two-sided alternative

$$H_A \colon \mu_X \neq \mu_Y .$$

Put differently, we reject H_0 for large values of $|t|$.
Under H_0 the t statistic

$$t = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

follows Student's t distribution with $n_1 + n_2 - 2$ df.
Consequently, the decision rule

$$\text{reject } H_0 \text{ if } |t| > t_{n_1+n_2-2}(\frac{\alpha}{2})$$

yields a test at level α.[5]
Question: Test whether the mean time required until pain relief is the same for both medicines at level $\alpha = 0.5$.

[5] We denote by $t_n(\alpha)$ the $(100\alpha)th$ percentile of Student's t distribution with n df.

Comparing two populations.

Example (Medicine for pain-relief)

Effectiveness of two medicines for pain-relief are compared in a medical research study. $n = 473$ patients were randomly assigned to one of two groups. Medicine 1 was prescribed to the patients of group 1, medicine 2 was prescribed to patients in the other group. Experimenters recorded the time required to receive pain relief. These data are summarized in the table.

	group 1	group 2
sample size	$n_1 = 248$	$n_2 = 225$
sample mean (time)	$\bar{x}_1 = 24.8m$	$\bar{x}_2 = 26.1m$
sample SD	$s_X = 3.3m$	$s_Y = 4.2m$

Table: Summary of times (in minutes) required to receive pain relief.

Test whether the average time until pain relief is the same for both medicines.

Comparing two populations.

Example (Medicine for pain-relief)

Want to test null hypothesis

$$H_0 \colon \mu_1 = \mu_2$$

versus alternative

$$H_A \colon \mu_1 \neq \mu_2,$$

where

$$\mu_i := \text{mean time until pain relief for medicine } i.$$

Comparing two populations.

Example (Medicine for pain-relief)

Decision rule for the two-sided test at level $\alpha = 0.5$ is

$$\text{accept } H_0 \text{ if} \quad t_{n_1+n_2-2}(\frac{\alpha}{2}) \leq t \leq t_{n_1+n_2-2}(1-\frac{\alpha}{2}).$$

Since

$$t_{n_1+n_2-2}(\frac{\alpha}{2}) = t_{471}(0.025) \approx -1.97$$

and the t distribution is symmetric, we have

$$t_{471}(0.975) \approx 1.97$$

Comparing two populations.

Example (Medicine for pain-relief)

We reject H_0 at level $\alpha = 0.5$, since

$$s_p^2 = \frac{(n_1-1)s_X^2 + (n_2-1)s_Y^2}{n_1+n_2-2} = \frac{247(3.3m)^2 + 224(4.2m)^2}{248+225-2}$$
$$= \frac{6641.2m^2}{471} \approx 14.1m^2,$$
$$\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = \sqrt{\frac{1}{248}+\frac{1}{225}} \approx 0.092$$

and we obtain for the t statistic (under the null)

$$t = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2}}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{24.8-26.1}{3.75 \times 0.092} \approx -3.77 < -1.97.$$

Comparing two populations.

Remark

Often the requirement $\sigma_X^2 = \sigma_Y^2$ that the variances in treatment and control population be equal is not met. In this case the variance of $\bar{X}_{n_1} - \bar{Y}_{n_2}$ could be estimated by the corresponding sample variance

$$\frac{s_X^2}{n} + \frac{s_Y^2}{n}.$$

However, as a rule of thumb, the statistic

$$\frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$$

no longer exactly follows Student's t distribution, but it can be approximated by Student's t distribution with df the integer nearest to

$$\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2 \Bigg/ \left(\frac{s_X^2}{n^2(n-1)} + \frac{s_Y^2}{m^2(m-1)}\right).$$

Comparing two populations.

Nonparametric tests

Comparing two populations. Wilcoxon rank-sum test.

A nonparametric statistical method does not assume the data $x_1, \ldots, x_n$ to be observations of random variables $X_1, \ldots, X_n$ whose joint distribution stems from a specific parametric family of distributions

$$P_\theta, \theta \in \Theta.$$

Do you know an example of a nonparametric method?

Comparing two populations. Wilcoxon rank-sum test.

Evidently, a nonparametric method is per se more widely applicable (there is no restriction on the distribution that gives rise to the data) than a parametric method.
However, one often pays a prize for this generality, e.g.

- if information about the data generating mechanism is known, one may lose power, and
- derivations are analytically more involved.

Comparing two populations. Wilcoxon rank-sum test.

So far we compared two populations (treatments and controls) by comparing independent random (unpaired) samples form these populations. The hypothesis that a treatment has no effect was formalized as

$$H_0 \colon \mu_X = \mu_Y .$$

Moreover, we assumed

(1) both populations to be normally distributed
(in particular $X_i \sim \mathcal{N}(\mu_X, \sigma_X^2)$, $Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$)

and, in the case of small sample size, we additionally assumed

(2) $\sigma_X^2 = \sigma_Y^2$.

Comparing two populations. Wilcoxon rank-sum test.

Wilcoxon rank-sum test (proposed by Wilcoxon in 1945) does not make any of these assumptions.

Wilcoxon's insight: Instead of the observations study their *ranks*.

(Also, makes test less sensitive to outliers.)

Null hypothesis (treatment has no effect)

$$H_0\text{: the two populations are identical}$$

implies that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ have the same joint distribution.

Alternative hypothesis

$$H_A\text{: the two populations are not identical.}$$

More precisely, if F/G denote the cdfs of treatment/control population

$$H_A\text{: } \begin{cases} F(x) \leq G(x) \text{ for all x} & \text{(one-sided)} \\ F(x) \leq G(x) \text{ for all x, or } F(x) \geq G(x) \text{ for all x.} & \text{(two-sided)} \end{cases}$$

Comparing two populations. Wilcoxon rank-sum test.

This time we do not construct the test statistic from

$$\bar{X}_{n_1} \qquad \text{and} \qquad \bar{Y}_{n_2}.$$

Instead we study the pooled observations

$$X_1, X_2, \ldots, X_{n_1}, Y_1, Y_2, \ldots, Y_{n_2}$$

which we denote by

$$Z_1, Z_2, \ldots, Z_{n_1+n_2}.$$

Consider the order statistics

$$Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(n_1+n_2)}.$$

For simplicity, assume there are no ties,
i.e. $Z_{(1)} < Z_{(2)} < \cdots < Z_{(n_1+n_2)}$.

Comparing two populations. Wilcoxon rank-sum test.

The *rank*[6] of observation X_i among the pooled observations is

$$\mathsf{rank}(X_i) := j \quad \text{if } X_i = Z_{(j)}.$$

The rank sum of the sample from the first, respectively second, population is

$$R_1 := \sum_{i=1}^{n_1} \mathsf{rank}(X_i) \quad \text{respectively} \quad R_2 := \sum_{i=1}^{n_2} \mathsf{rank}(Y_i).$$

In particular,

$$R_1 + R_2 = \sum_{i=1}^{n_1+n_2} i = \frac{1}{2}(n_1+n_2)(n_1+n_2+1).$$

[6] If observations $z_{(i)} = z_{(i+1)} = \ldots = z_{(i+j)}$ are ties, set their ranks equal to their average rank.

Comparing two populations. Wilcoxon rank-sum test.

Example

Consider independent samples

$$5, 10 \qquad 2, 7, 9$$

of sizes $n_1 = 2$ and $n_2 = 3$ from two different populations.
Ordered pooled observations and their ranks are given in the table.

2	5	7	9	10
1	2	3	4	5

Rank sums

$$R_1 = 7, \qquad R_2 = 8.$$

Comparing two populations. Wilcoxon rank-sum test.

Under the null hypothesis, observation X_1 will be smaller or larger than Y_1 with equal probability, i.e.[7]

$$\mathbb{P}\{X_1 < Y_1\} = \mathbb{P}\{X_1 > Y_1\} = \frac{1}{2},$$

and this is true for any two observations X_i, Y_j, i.e.

$$\mathbb{P}\{X_i < Y_j\} = \mathbb{P}\{X_i > Y_j\} = \frac{1}{2}.$$

What is more, we will see any particular assignment of rankings to the Xs (respectively Ys) with equal probability. The total number of ways to assign rankings (numbers between 1 and $n_1 + n_2$) to the Xs is $\binom{n_1+n_2}{n_1}$.

[7] We assume there are no ties, e.g. the observations could be drawn from continuous populations.

Comparing two populations. Wilcoxon rank-sum test.

Example (Balances in bank accounts)

$n = 22$ bank accounts are sampled randomly from two different branches of a bank. The records of balances are given in the table.

branch 1		branch 2	
acc. balance	rank	acc. balance	rank
1095	20	885	7
955	14	850	4
1200	22	915	8
1195	21	950	12.5
925	9	800	2
950	12.5	750	1
805	3	865	5
945	11	1000	16
875	6	1050	18
1055	19	935	10
1025	17		
975	15		

Comparing two populations. Wilcoxon rank-sum test.

Example (Balances in bank accounts)

The sum of ranks for each sample are

$$R_1 = 169.5, \qquad R_2 = 83.5.$$

There were $n_1 = 12$ bank accounts sampled from branch 1. The minimal value for the sum of ranks for a sample of this size is

$$R_1 = 1 + 2 + \cdots + n_1 = \frac{1}{2} n_1 (n_1 + 1) = 78,$$

corresponding to the fact that all observations from the first population are smaller than any observation in the second population.
R_1 close to 78 implies that branch 1 has smaller account balances, and contradicts H_0.

Comparing two populations. Wilcoxon rank-sum test.

Example (Balances in bank accounts)

The maximal value for the sum of ranks for a sample of size $n_1 = 12$ where the second sample is of size $n_2 = 10$ is

$$R_1 = (n_2+1)+(n_2+2)+\cdots+(n_2+n_1) = n_1 n_2 + \frac{1}{2} n_1(n_1+1) = 198,$$

corresponding to the fact that all observations from the first population are greater than any observation in the second population.
R_1 close to 198 implies that branch 1 has greater account balances, and also contradicts H_0.

Comparing two populations. Wilcoxon rank-sum test.

Example (Balances in bank accounts)

Under the null the distribution R_1 is symmetric, and we expect R_1 to be close to the average

$$\frac{n_1 n_2 + n_1(n_1+1)}{2} = \frac{n_1(n_1+n_2+1)}{2} = 138.$$

of its minimal and maximal value.
Recall that

$$R_1 = 169.5 > 138,$$

and we might suspect that branch 1 has greater account balances.
Or could this deviation be just due to chance?
—In order to answer this question, need to know null distribution of R_1.

Comparing two populations. Wilcoxon rank-sum test.

Fact

Consider n random samples, n_1 from treatment population, n_2 from population of controls. If R_1 denotes the rank sum of the treatment group, then under the null hypothesis

$$\mathbb{E}R_1 = \frac{n_1(n_1+n_2+1)}{2}$$
$$\mathrm{Var}(R_1) = \frac{n_1 n_2(n_1+n_2+1)}{12}.$$

Notice that because of symmetry this fact directly yields mean and variance of R_2 (interchange n_1 and n_2 in the formulas).

Comparing two populations. Wilcoxon rank-sum test.

In practice, instead of R_1 often a function thereof is used as test statistic
(in order to exploit symmetries of distribution of R_1).
Namely,

1. Consider the pooled observations
   $Z := (X_1, \ldots, X_n, Y_1, \ldots, Y_n)$.
2. Let n^* denote the size of the smaller sample.
3. Calculate $R :=$ sum of ranks in Z of smaller sample.
4. Set $R' := n^*(n_1 + n_2 + 1) - R$.
5. Set $R^* := \min(R, R')$.
6. Critical values for R^* are tabulated. If R^* is too small, reject H_0 that populations are identical.

Comparing two populations. Wilcoxon rank-sum test.

Example (Balances in bank accounts)

We find

2. $n^* = 10$,
3. $R = R_2 = 83.5$,
4. $R' := n^*(n_1 + n_2 + 1) - R = 10(12 + 10 + 1) - 83.5 = 146.5$.
5. $R^* := \min(R, R') = 83.5$.
6. From Table 8 in Appendix B of Rice's textbook find 84 as critical value for R^* ($n_1 = 12$, $n_2 = 10$) for a two-sided test at level $\alpha = 0.05$. Since $R^* = 83.5 < 84$, reject H_0.[8]

This means that, at significance level $\alpha = 0.05$, the differences in account balances we see in the data cannot be explained by chance alone. They are due to the fact that balances in the two branches[9] are indeed different.

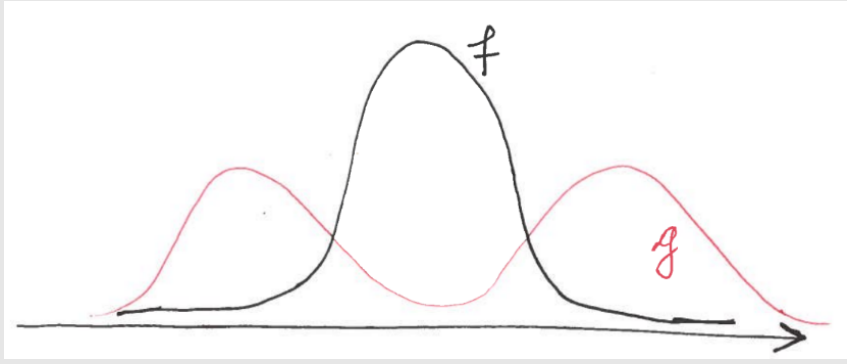
[8] Critical value for significance level $\alpha = 0.01$ is 76, so Wilcoxon does not reject at this level.

[9] More precisely: the distributions of balances in the bank accounts differ.

Comparing two populations. Wilcoxon rank-sum test.

Example (A cautionary example)

(X_i) i.i.d. $\sim F$ with density f $\quad$ (Y_i) i.i.d. $\sim G$ with density g



Densities f, g symmetric about their common mean
$\mu = \int_{-\infty}^{\infty} tf(t)dt = \int_{-\infty}^{\infty} tg(t)dt.$
Under H_0 expect R_1 to be close to its mean $n_1(n_1 + n_2 + 1)/2$ supporting

$$H_0 \colon F = G.$$

But, we *know* that $F \neq G$ by construction! What went wrong?

Comparing two populations. Mann-Whitney test.

In 1947 Mann and Whitney proposed a different test based on ranks, which (at first sight) is not based on the rank-sum statistic. This test occurs often in the literature.
It turns out, however, that Mann and Whitney's U statistic is a simple function of the rank sum R_2, and the two tests are therefore equivalent.

Comparing two populations. Mann-Whitney test.

Suppose the data are modeled as draws from two populations with

$$F \text{ cdf of treatment population}$$
$$G \text{ cdf of control population.}$$

As before, we are interested in testing

$$H_0 \colon F = G.$$

Instead of ranking the observations, consider the probability

$$\pi := \mathbb{P}\{X < Y\}$$

that a sample X from the treatment population is smaller than a sample Y from the control population. Under H_0 we have $\pi = \frac{1}{2}$.

Comparing two populations. Mann-Whitney test.

A good estimator for

$$\pi := \mathbb{P}\{X < Y\}$$

should be the relative frequency of the pairs (X_i, Y_j) of observations such that $X_i < Y_j$. Since there are $n_1 n_2$ possible pairs in total, the estimator is

$$\hat{\pi} := \frac{\#\{(X_i, Y_j)\colon X_i < Y_j\}}{n_1 n_2} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathbf{1}_{\{X_i < Y_j\}}$$

$$= \frac{1}{n_1 n_2} \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} \mathbf{1}_{\{X_{(i)} < Y_{(j)}\}}.$$

Notice that

$$\sum_{i=1}^{n_1} \mathbf{1}_{\{X_{(i)} < Y_{(j)}\}} = \text{number of } X\text{s less than } Y_{(j)} = \text{rank}(Y_{(j)}) - j,$$

where the $-j$ accounts for $Y_{(1)}, \ldots, Y_{(j)}$.

Comparing two populations. Mann-Whitney test.

Plugging

$$\sum_{i=1}^{n_1} \mathbf{1}_{\{X_{(i)}<Y_{(j)}\}} = \mathsf{rank}(Y_{(j)}) - j$$

into the formula for $\hat{\pi}$ yields

$$\hat{\pi} = \frac{1}{n_1 n_2} \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} \mathbf{1}_{\{X_{(i)}<Y_{(j)}\}} = \frac{1}{n_1 n_2} \sum_{j=1}^{n_2} (\mathsf{rank}(Y_{(j)}) - j)$$

$$= \frac{1}{n_1 n_2} \left(\sum_{j=1}^{n_2} \mathsf{rank}(Y_{(j)}) - \sum_{j=1}^{n_2} j \right) = \frac{1}{n_1 n_2} (R_2 - \frac{n_2(n_2+1)}{2}),$$

where R_2 is the sum of ranks of the sample from the control population.

Comparing two populations. Mann-Whitney test.

The Mann-Whitney U statistic is defined as

$$U_Y := n_1 n_2 \hat{\pi} = \#\{(X_i, Y_j) \colon X_i < Y_j\} = R_2 - \frac{n_2(n_2+1)}{2}.$$

Corollary

From the previous fact on mean and variance of R_1 (respectively R_2) we find

$$\mathbb{E}U_Y = \frac{n_1 n_2}{2}, \qquad \mathrm{Var}(U_Y) = \mathrm{Var}(R_2) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{2}.$$

Ex: prove these statements.

Comparing two populations. Mann-Whitney test.

It can be shown that as n_1, $n_2 \to \infty$, the null distribution of the Mann-Whitney statistic U_Y converges to a normal distribution, i.e.

$$\frac{U_Y - \mathbb{E}U_Y}{\sqrt{\mathrm{Var}(U_Y)}} \to N,$$

under the null, where $N \sim \mathcal{N}(0, 1)$.

[Heuristics: Normal approximation to binomial distribution]

In practice, the normal approximation is already used for sample sizes n_1, n_2 greater than 10.

Comparing two populations. Mann-Whitney test.

Example (Balances in bank accounts)

For our data we find

$$U_Y = 91.5$$

and a p-value of 0.041 (e.g. using **wilcox.test** in R).

As with the Wilcoxon rank-sum test the Mann-Whitney test rejects H_0 at significance level $\alpha = 0.05$, but not at significance level $\alpha = 0.01$

Comparing two populations. Mann-Whitney test.

Bootstrapping $\hat{\pi}$. We saw the importance of the estimator

$$\hat{\pi} := \frac{\#\{(X_i, Y_j) \colon X_i < Y_j\}}{n_1 n_2}$$

of $\pi := \mathbb{P}\{X < Y\}$ for the Mann-Whitney test.

To bootstrap $\hat{\pi}$,

1. Approximate the
   - cdf F of the treatment population by the ecdf F_{n_1},
   - cdf G of the control population by the ecdf G_{n_2}.
2. Take independent random samples
   - $x_1^*, x_2^*, \ldots, x_{n_1}^*$ from F_{n_1}
   - $y_1^*, y_2^*, \ldots, y_{n_2}^*$ from G_{n_2},

   and use them to compute the corresponding bootstrap sample

$$\hat{\pi}_1^* = \frac{\#\{(x_i^*, y_j^*) \colon x_i < y_j\}}{n_1 n_2}.$$

   Repeat B times to obtain simulated samples $\hat{\pi}_1^*, \hat{\pi}_2^*, \ldots, \hat{\pi}_B^*$.
3. Can now either study distribution of $\hat{\pi}_1^*, \hat{\pi}_2^*, \ldots, \hat{\pi}_B^*$, or compute any of its properties.