Non-Parametric Comparisons; Paired Tests and Correlation

Non-Parametric Comparison of Location

Problem: Two independent random variables X and Y are given. Nothing is known about their distribution. Comparing means or even medians is difficult.

Approach: Compare their ***locations*** by checking if

$$P[X > Y] + \frac{1}{2}P[X = Y] \stackrel{?}{=} \frac{1}{2}.$$

If the above probability equals $1/2$, a random observation of X will be greater than or equal to a random observation of Y with probability one-half, and of course the converse is also true.

Here we will assume that X and Y are ***continuous random variables***, so we may omit $P[X = Y]$.

The Wilcoxon Rank-Sum Test

The Wilcoxon rank-sum test is used to decide whether to reject the null hypothesis

$$H_0 : P[X > Y] = \frac{1}{2} \qquad \text{or} \qquad H_0 : P[X > Y] \leq \frac{1}{2}.$$

If both X and Y follow the same distribution, possibly with different location parameter, this may be interpreted as a test comparing the medians of X and Y.

Observations of X and Y are ranked from smallest to largest. For each population, the ranks are summed independently. If $P[X > Y] = 1/2$, then the sum of ranks should be roughly the same for both populations.

To make calculations easier, it is sufficient to consider the sums of ranks of the smaller sample (if sample sizes are different).

The Wilcoxon Rank-Sum Test

22.1. Wilcoxon Rank-Sum Test. Let X and Y be two random samples following some continuous distributions.

Let $X_1, \dots, X_m$ and $Y_1, \dots, Y_n$, $m \leq n$, be random samples from X and Y and associate the rank R_i, $i = 1, \dots, m+n$, to the R_ith smallest among the $m+n$ total observations. If ties in the rank occur, the mean of the ranks is assigned to all equal values.

Then the test based on the statistic

$$W_m := \text{sum of the ranks of } X_1, \dots, X_m.$$

is called the ***Wilcoxon rank-sum test***.

We reject $H_0 \colon P[X > Y] = 1/2$ (and similarly the analogous one-sided hypotheses) at significance level α if W_m falls into the corresponding critical region.

The Wilcoxon Rank-Sum Test

The Wilcoxon rank-sum test is also called the ***Mann-Whitney U-test***. (Often, this refers to the equivalent test where all the ranks, not just those of the smaller sample, are summed.)

For large values of m ($m \geq 20$), W_m is approximately normally distributed with

$$\mathrm{E}[W_m] = \frac{m(m+n+1)}{2}, \qquad \mathrm{Var}[W_m] = \frac{mn(m+n+1)}{12}.$$

If there are many ties, the variance may be corrected by taking

$$\mathrm{Var}[W_m] = \frac{mn(m+n+1)}{12 - \sum_{\text{groups}} \frac{t^3+t}{12}}$$

where the sum is taken over all groups of t ties. However, the best way to deal with ties is still a topic of current research.

Example: Midterm Exam Scores

22.2. Example. It has been suggested that the most highly motivated JI undergraduate students do at least as well, possibly even better, than graduate students in my graduate-level mathematics courses. In the spring term of 2018, there was a significant enrolment of undergraduate students in *Vv557 Methods of Applied Maths II*. The results of the first midterm exam are taken to serve as an indication of the possible truth of this hypothesis.

The hypothesis to be tested is

$$H_0 \colon P[X_{\text{undergrad}} > X_{\text{grad}}] \leq 1/2$$

where $X_{\text{undergrad}}$ and X_{grad} are the exam scores of undergraduate and graduate students enrolled in Vv557.

(The null hypothesis can be interpreted as "undergraduate students do not do better than graduate students in the first midterm.")

The Raw Data

The following data were recorded (points out of 20 maximum in the first midterm exam):

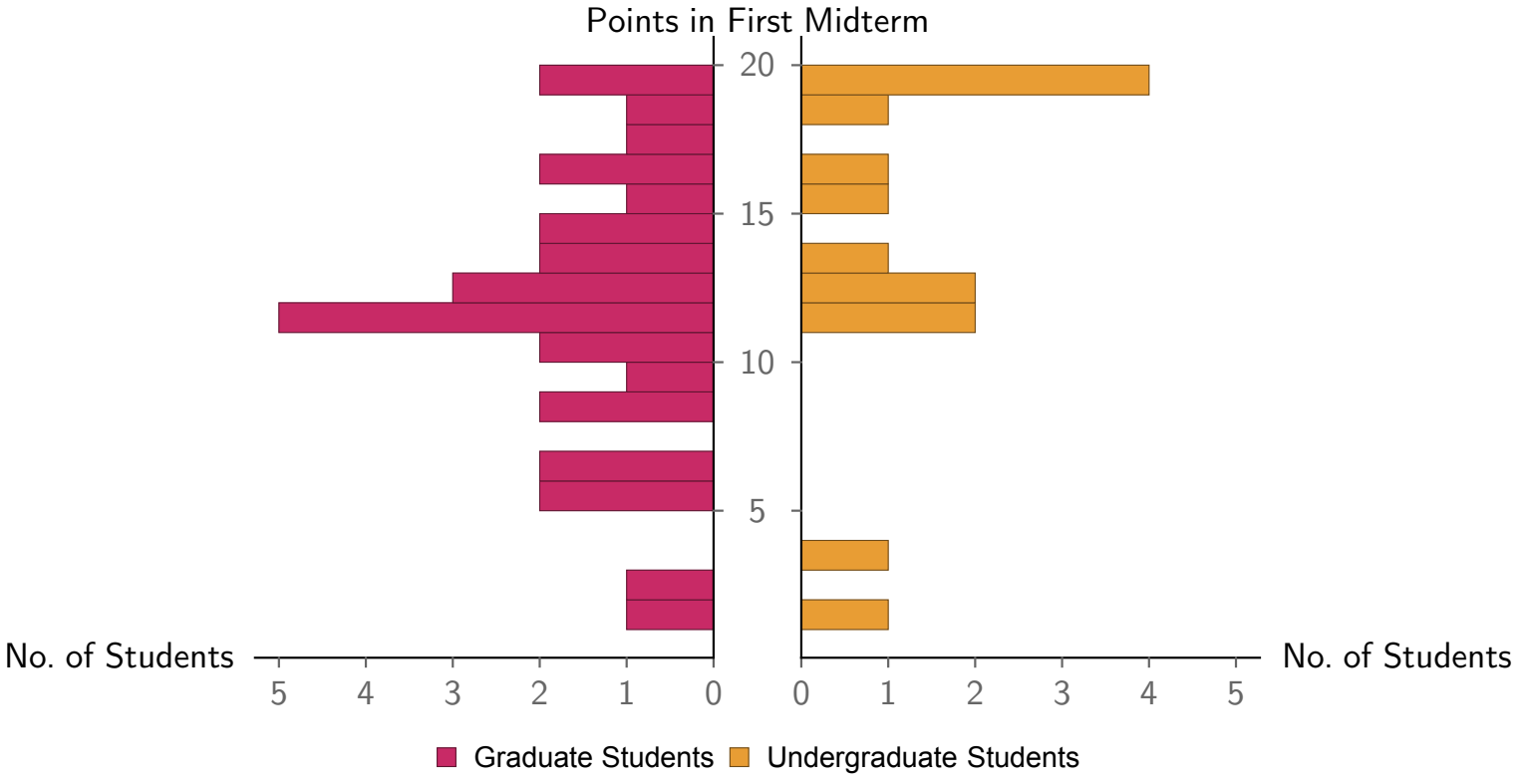
Graduate	5.5	5.5	12.75	18.75	19.25	11.25
	11.5	11.5	12.25	14.25	9.25	14.5
	13.25	8.25	16.75	10.5	6	15.25
	6.5	12.5	10.5	8.75	11.5	17
	2.75	13.25	19	16.5	11.5	1.75
Undergraduate	18.5	12.25	3	15	19.75	11.25
	11.75	19.25	12.25	19.75	16.25	13
	19.25	1.75				

The quartiles are,

Graduate :	8.75,	11.5,	14.5;
Undergraduate :	11.75,	14,	19.25.

Visualizing the Data

The double bar chart below visualizes these data:



Clearly, a non-parametric test is the best choice here.

Ranking the Data

The data are arranged in order and ranked as follows:

Student	Points	Rank	Student	Points	Rank	Student	Points	Rank
grad	1.75	1.5	grad	11.5	17.5	undergrad	15	31
undergrad	1.75	1.5	grad	11.5	17.5	grad	15.25	32
grad	2.75	3	grad	11.5	17.5	undergrad	16.25	33
undergrad	3	4	grad	11.5	17.5	grad	16.5	34
grad	5.5	5.5	undergrad	11.75	20	grad	16.75	35
grad	5.5	5.5	grad	12.25	22	grad	17	36
grad	6	7	undergrad	12.25	22	undergrad	18.5	37
grad	6.5	8	undergrad	12.25	22	grad	18.75	38
grad	8.25	9	grad	12.5	24	grad	19	39
grad	8.75	10	grad	12.75	25	grad	19.25	41
grad	9.25	10	undergrad	13	26	undergrad	19.25	41
grad	10.5	12.5	grad	13.25	27.5	undergrad	19.25	41
grad	10.5	12.5	grad	13.25	27.5	undergrad	19.75	43.5
grad	11.25	14.5	grad	14.25	29	undergrad	19.75	43.5
undergrad	11.25	14.5	grad	14.5	30			

Calculating the Test Statistic

The sum of the ranks of the undergraduate students (smaller sample size) is

$$\begin{aligned} w_{14} &= 1.5 + 4 + 14.5 + 20 + 22 + 22 + 26 \\ &\quad + 31 + 33 + 37 + 41 + 41 + 43.5 + 43.5 \\ &= 380 \end{aligned}$$

Given the large sample sizes, we use a normal approximation for the test statistic (most tables only include values for $m, n \leq 20$). We have

$$\mathrm{E}[W_{14}] = \frac{14(14 + 30 + 1)}{2} = 315,$$

$$\mathrm{Var}\, W_{14} = \frac{14 \cdot 30(14 + 30 + 1)}{12} = 1575$$

Performing the Fisher test

Therefore,

$$Z = \frac{W_m - 315}{\sqrt{1575}}$$

follows a standard normal distribution if $P[X_{\text{undergrad}} > X_{\text{grad}}] = 1/2$. The value of our test statistic is

$$z = \frac{380 - 315}{\sqrt{1575}} = 1.64.$$

Using the normal distribution table, we find a P-value of

$$P[Z \geq 1.64] = 0.0505.$$

There is possibly a small indication that undergraduate students might do better than graduate students, but the evidence is far from conclusive.

Discussion of the Wilcoxon Rank Tests

In the previous example we did not apply the correction for ties to the variance of the normal distribution - why?

Had we done so, the variance would have been ***negative*** - not good!

Could we have used an exact table of critical values? No, because no such table exist for $m, n > 20$. The Wilcoxon rank tests are ***combinatorial tests*** and P-values become increasingly hard to calculate exactly as the number of possible permutations of ranks increases with m and n.

For this reason, ***ties are problematic*** since they increase the number of possible permutations. We have presented one way to deal with ties (assigning the average rank) but this is not the only approach. This is the subject of current research!

Literature: McGee, M. ***Case for omitting tied observations in the two-sample T-test and the Wilcoxon-Mann-Whitney Test.*** PLoS One 13:7, 2018.

Paired Tests

Problem: When comparing means (or, in general, the location) of two populations ***extraneous factors*** may distort the results.

22.3. Example. Suppose we wish to study the efficacy of two different drugs in fighting a disease, Drug A and Drug B. A simple approach would be to treat 20 patients with Drug A and 20 patients with Drug B and then compare the average degree of improvement.

However, it could be that (for example) the disease affects smokers more severely than non-smokers.

If there are more smokers among the sample for Drug A than for Drug B, this could cause the improvements measured for Drug A to be less evident than for Drug B, even if overall Drug A were the better drug.

Paired Tests

Instead, a better approach is to ***pair the samples***: For every person with certain characteristics (gender, age smoker/non-smoker, etc.) administered with Drug A, a person with the same characteristics is administered Drug B.

That means that the sample sizes must be equal in both populations and every sample observation in one population is paired with a corresponding observation in the other population.

Suppose we have two populations with random variables X and Y that we wish to compare. We then define a new random variable

$$D := X - Y$$

and conduct all tests on D.

Paired T-Tests

We note that

$$\mu_D = \mathrm{E}[D] = \mathrm{E}[X - Y] = \mathrm{E}[X] - \mathrm{E}[Y] = \mu_X - \mu_Y.$$

Therefore, the hypothesis (for example)

$$H_0 \colon \mu_X = \mu_Y \qquad \text{may be replaced with} \qquad H_0 \colon \mu_D = 0.$$

We will assume that X and Y follow a ***joint bivariate normal distribution***. Then it is not hard to see that $D = X - Y$ follows a normal distribution.

We then consider a paired random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from both populations yielding a sample $D_1, \ldots, D_n$ with $D_i := X_1 - Y_i$, $i = 1, \ldots, n$.

We denote by $\overline{D}$ the sample mean and by S_D^2 the sample variance of D.

Paired T-Tests

Then

$$T_{n-1} = \frac{\overline{D} - \mu_D}{\sqrt{S_D^2/n}}$$

follows a T-distribution with $n-1$ degrees of freedom.

We may find confidence intervals for μ_D and conduct hypothesis tests as we would for any normally distributed random variable. A T-test for D is called a ***paired T-test*** for Y and Y.

Paired T-Tests

22.4. Example. In a study of the effectiveness of physical exercise in weight reduction, a group of 16 persons engaged in a prescribed program of physical exercise for one month showed the following results :

Weight before (X)	209	178	169	212	180	192	158	180
Weight after (Y)	196	171	170	207	177	190	159	180
$D = Y - X$	-13	-7	+1	-5	-3	-2	+1	0

Weight before (X)	170	153	183	165	201	179	243	144
Weight after (Y)	164	152	179	162	199	173	231	140
$D = Y - X$	-6	-1	-4	-3	-2	-6	-12	-4

We want to test at the 0.01 level of significance whether the exercise program is effective.

Paired T-Tests

We decide to test

$$H_0 : \mu_D \geq 0.$$

From the $n = 16$ data we have $\overline{D} = -4.125$ and $s_D^2 = 16.517$. The test statistic is

$$T = \frac{\overline{D}}{s_D/\sqrt{n}} = -4.06.$$

Since $t_{0.01,15} = 2.602$, we may reject H_0 at the 0.01 level of significance.

There is evidence that the physical exercise program leads to a loss of weight.

Non-Parametric Paired Test

Suppose the two independent random variables X and Y do not follow a normal distribution. Then we would like to treat $D = X - Y$ by the Wilcoxon signed-rank test.

The signed-rank test requires a random variable to have a ***symmetric*** distribution. What does that mean?

A random variable X is said to be ***symmetric about*** $a \in \mathbb{R}$ if

$$X - a \qquad \text{and} \qquad -(X - a)$$

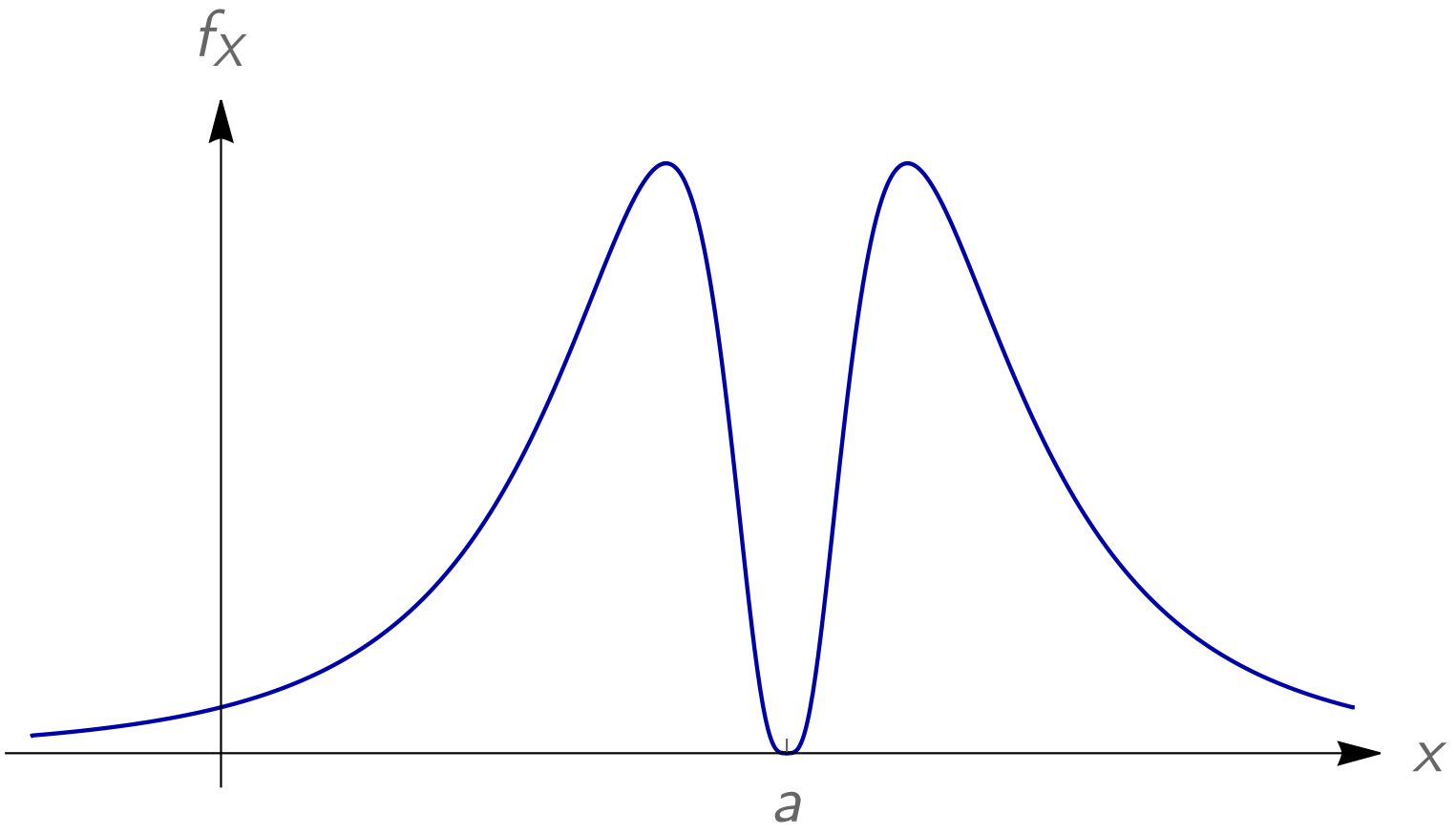
have the same distribution.

In terms of the density function f_X this means that

$$f_X(x - a) = f_X(a - x)$$

(as can be verified by applying Theorem 7.5.)

Example of the Density of a Symmetric Distribution



Properties of $D = X - Y$

Now let X and Y be two independent random variables that follow the same distribution but differ only in their location, i.e., $X' := X - \delta$ and Y are independent and identically distributed.

Then

$$P[X - Y > \delta] = P[X - \delta - Y > 0] = P[X' - Y > 0] = \frac{1}{2}$$

so δ is the median of $X - Y$.

Furthermore

$$D = X - Y = \delta + X' - Y$$

and

$$2\delta - D = \delta + Y - X'$$

have the same distribution since Y and X' are i.i.d. random variables.

Non-Parametric Paired Test

Therefore, D will be symmetric about its median δ and we can apply the Wilcoxon signed rank test to test hypotheses about δ.

Historically, Wilcoxon proposed both the rank-sum test (for pooled comparisons) and the signed-rank test (for paired comparisons) in a single publication.

Paired vs. Pooled T-Tests

Let us take another look at the test statistics for a paired T-test. We note that

$$\begin{aligned}\overline{D} &= \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} Y_i \\ &= \overline{X} - \overline{Y}\end{aligned}$$

and $\mu_D = \mu_X - \mu_Y$. This suggests that the paired and pooled T-test may actually be fairly similar.

Paired vs. Pooled T-Tests

For our comparison, let us assume that we have two populations of normally distributed random variables X and Y with equal variances σ^2.

We want to test

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

and take a paired sample of equal size n from (X, Y).

Then we could either perform a paired test or a pooled test - which is more powerful? Let us compare the test statistics:

$$T_{\text{pooled}} = \frac{\overline{X} - \overline{Y}}{\sqrt{2S_p^2/n}}, \qquad \text{critical value} = t_{\alpha/2,2n-2},$$

$$T_{\text{paired}} = \frac{\overline{X} - \overline{Y}}{\sqrt{S_D^2/n}}, \qquad \text{critical value} = t_{\alpha/2,n-1},$$

Paired vs. Pooled T-Tests

We immediately note that

the pooled test has more degrees of freedom

and so rejecting H_0 is easier - the test would be more powerful, if the test statistics were equal.

But the test statistics differ:

- In the pooled test, the denominator contains

$$2S_p^2/n \qquad \text{which estimates} \qquad 2\sigma^2/n.$$

- In the pooled test, the denominator contains

$$S_D^2/n \qquad \text{which estimates} \qquad \sigma_D^2/n = \sigma_D^2.$$

Paired vs. Pooled T-Tests

To discuss the two denominators, we will compare

$$\frac{2\sigma^2}{n} \qquad \text{with} \qquad \sigma^2_{\overline{D}}.$$

A direct calculation yields

$$\begin{aligned}
\sigma^2_{\overline{D}} &= \mathrm{Var}[\overline{D}] \\
&= \mathrm{Var}[\overline{X} - \overline{Y}] \\
&= \mathrm{Var}[\overline{X}] + \mathrm{Var}[\overline{Y}] - 2\,\mathrm{Cov}[\overline{X}, \overline{Y}] \\
&= \frac{\sigma^2}{n} + \frac{\sigma^2}{n} - 2\frac{\sigma^2}{n} \frac{\mathrm{Cov}[\overline{X}, \overline{Y}]}{\sqrt{\mathrm{Var}[\overline{X}]}\sqrt{\mathrm{Var}[\overline{Y}]}} \\
&= \frac{2\sigma^2}{n}(1 - \rho_{\overline{X}\,\overline{Y}})
\end{aligned}$$

where $\rho_{\overline{X}\,\overline{Y}}$ is the correlation coefficient of X and Y.

Correlation Coefficient of Sample Means

It is worth doing a quick calculation to verify that in our case (paired samples)

$$\rho_{\overline{X}\,\overline{Y}} = \rho_{XY}\,.$$

Since the covariance is bilinear,

$$\begin{aligned}
\mathrm{Cov}[\overline{X}, \overline{Y}] &= \mathrm{Cov}\Big[\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{j=1}^{n} Y_j\Big] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}[X_i, Y_j] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Cov}[X_i, Y_i]
\end{aligned}$$

where we have used that X_i and Y_j are independent for $i \neq j$.

Correlation Coefficient of Sample Means

Then $\mathrm{Cov}[X_i, Y_i] = \mathrm{Cov}[X, Y]$, so

$$\mathrm{Cov}[\overline{X}, \overline{Y}] = \frac{1}{n}\,\mathrm{Cov}[X, Y].$$

and, therefore,

$$\begin{aligned}
\rho_{\overline{X}\,\overline{Y}} &= \frac{\mathrm{Cov}[\overline{X}, \overline{Y}]}{\sqrt{\mathrm{Var}[\overline{X}]}\sqrt{\mathrm{Var}[\overline{Y}]}} \\
&= \frac{\frac{1}{n}\,\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{Var}[X]/n}\sqrt{\mathrm{Var}[Y]/n}} \\
&= \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{Var}[X]}\sqrt{\mathrm{Var}[Y]}} \\
&= \rho_{XY}.
\end{aligned}$$

Paired vs. Pooled T-Tests

The upshot of all this is that

$$\sigma_{\overline{D}}^2 = \frac{2\sigma^2}{n}(1 - \rho_{XY}).$$

Therefore, if $\rho_{XY} > 0$, the denominator of the paired statistic will be smaller than that of the pooled statistic, leading to a larger value of the statistic and a higher power of the test.

On the other hand, if ρ_{XY} is zero (or even negative), then pairing is intuitively unnecessary and in fact causes the test to lose power, since it is easier to reject H_0 when comparing with $t_{\alpha/2,2n-2}$ than with $t_{\alpha/2,n-1}$.

Pairing in the absence of correlation makes a test less powerful.

Estimating Correlation

Since correlation is important in deciding whether to use a paired or a pooled T-test, let us briefly discuss the estimation of ϱ.

Let us take a random sample of size n from (X, Y) as before. Then we have the natural unbiased estimators

$$\widehat{\mathrm{Var}[X]} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

$$\widehat{\mathrm{Cov}[X,Y]} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

The natural choice (method of moments!) for an estimator for the correlation coefficient is then

$$R := \hat{\rho} = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2}\sqrt{\sum(Y_i - \overline{Y})^2}}. \tag{22.1}$$

Correlation of Bivariate Normal Random Variables

Now let us suppose that (X, Y) follows a bivariate normal distribution, i.e., they have the joint density

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\varrho^2}}\, e^{-\frac{1}{2(1-\varrho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\varrho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]}$$

with $\mu_X, \mu_Y \in \mathbb{R}$, $\sigma_X, \sigma_Y > 0$ and correlation coefficient $\varrho \in (-1, 1)$.

Under this assumption, we will introduce a hypothesis test and a confidence interval for the correlation coefficient.

An important role is played by the Fisher transformation (8.3).

Hypothesis Tests for the Correlation Coefficient

It can be shown that for large n the Fisher transformation of R,

$$\frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) = \operatorname{Artanh}(R)$$

is approximately normally distributed with

$$\mu = \frac{1}{2}\ln\left(\frac{1+\varrho}{1-\varrho}\right) = \operatorname{Artanh}(\varrho), \qquad \sigma^2 = \frac{1}{n-3}.$$

We can thus test $H_0 \colon \varrho = \varrho_0$, by using the test statistic

$$\begin{aligned} Z &= \frac{\sqrt{n-3}}{2}\left(\ln\left(\frac{1+R}{1-R}\right) - \ln\left(\frac{1+\varrho_0}{1-\varrho_0}\right)\right) \\ &= \sqrt{n-3}\big(\operatorname{Artanh}(R) - \operatorname{Artanh}(\varrho_0)\big) \end{aligned} \tag{22.2}$$

Confidence Interval for the Correlation Coefficient

Furthermore, from (22.2) we can calculate a $100(1-\alpha)\%$ confidence interval for ϱ, given explicitly by

$$\left[\frac{1+R-(1-R)e^{2z_{\alpha/2}/\sqrt{n-3}}}{1+R+(1-R)e^{2z_{\alpha/2}/\sqrt{n-3}}},\ \frac{1+R-(1-R)e^{-2z_{\alpha/2}/\sqrt{n-3}}}{1+R+(1-R)e^{-2z_{\alpha/2}/\sqrt{n-3}}}\right].$$

or

$$\tanh\left(\operatorname{Artanh}(R)\pm\frac{z_{\alpha/2}}{\sqrt{n-3}}\right).$$

Correlation as a Measure of Skill vs. Luck

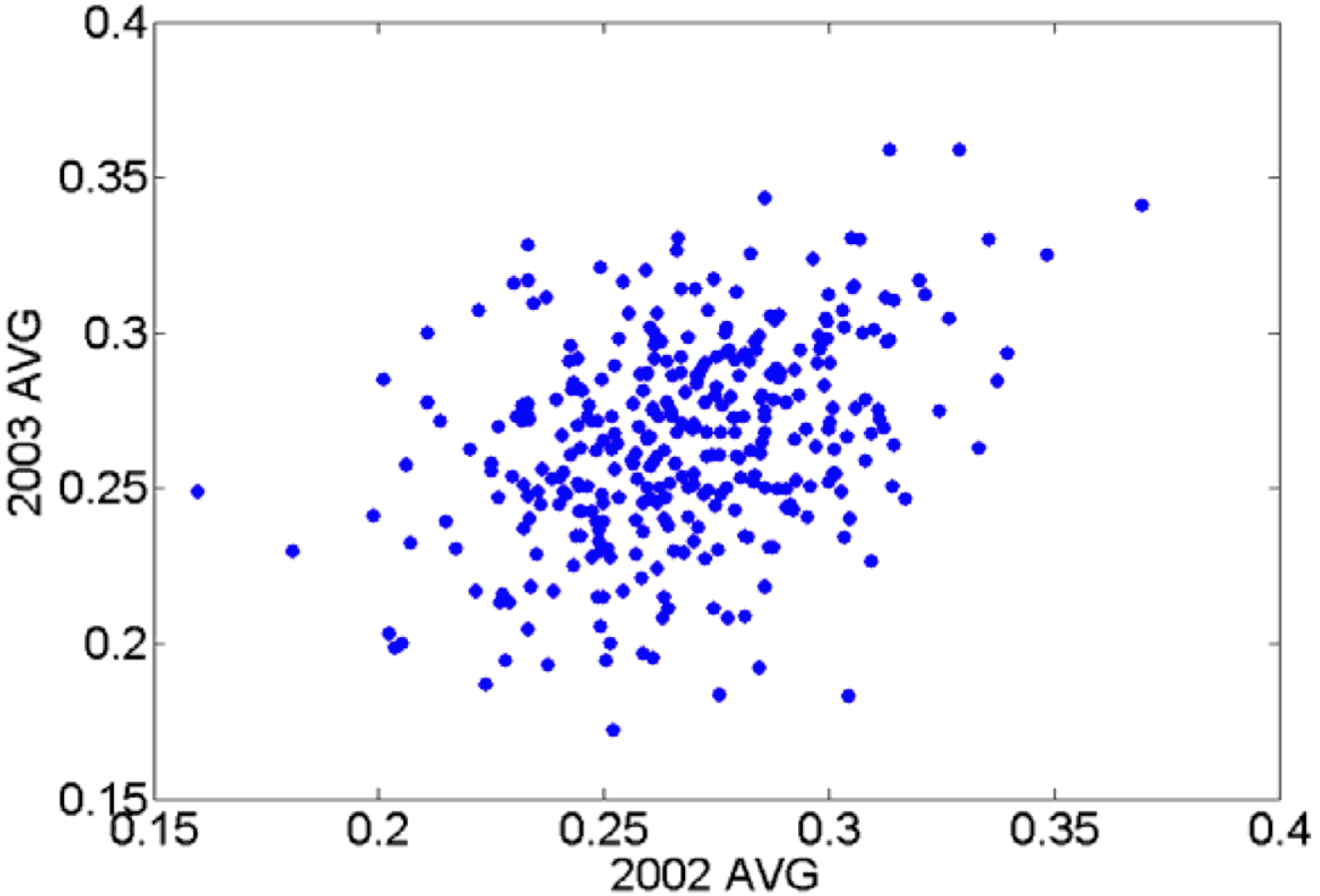
22.5. Example. The article ***A Batting Average: Does It Represent Ability or Luck?*** explores the suitability of a baseball player's "batting average" (BA for short; the number of hits divided by the number of at-bats) as a measure of skill.

The premise is the following:

- If a good BA is a matter of skill, a player's BA will have a consistent value from one year to the next. We would expect the batting average of a random player in one year to be linearly correlated to the BA in the next year.
- If a good BA is a matter of luck, a player's BA will vary from one year to the next and the BA as a random variable in a given year will be uncorrelated or even independent of the BA in another year.

We show the batting averages of all players with at least 100 at-bats in the 2002 and 2003 seasons on the next slide.

Correlation as a Measure of Skill vs. Luck

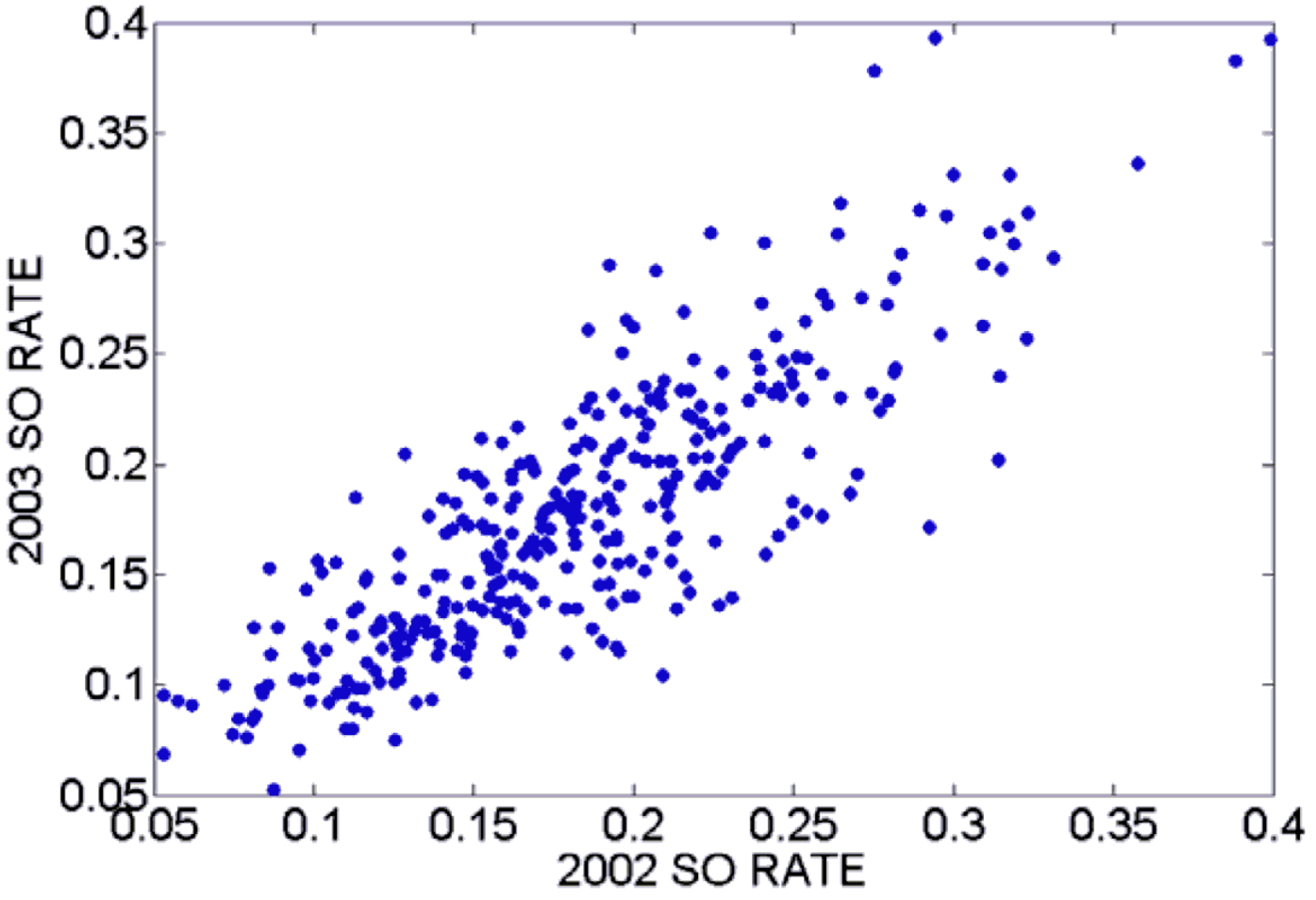


Batting averages in the 2002 and 2003 baseball seasons. J. Albert. ***A Batting Average: Does It Represent Ability or Luck?***

The article proposes that the "strikeout rate" is a better measure of skill.

Correlation as a Measure of Skill vs. Luck

Indeed, the corresponding scattergram (for the same players) seems to exhibit a stronger linear dependence from one year to the next:



Strikeout rates in the 2002 and 2003 baseball seasons. J. Albert. ***A Batting Average: Does It Represent Ability or Luck?***