6.4 Inference for matched pairs and two-sample data

An important type of application of confidence interval estimation and significance testing is when we either have *paired data* or *two-sample* data.

6.4.1 Matched pairs

Recall,

paired data is bivariate responses that consist of two determinations of basically the same characteristic. (Ch. 1).

Examples:

Practice SAT Scores before and after a prep course

Severity of a disease before and after treatment.

Leading-edge and trailing-edge measurement of each workpiece in a sample.

Bug bites on right arm and bug bites on left arm. (one has repellent).

One simple method of investigating the possibility of a consistent difference between paired data is to

1. Reduce the two paired measurements on each object to a single difference between them.
2. Methods of confidence interval estimation and significance testing applied to the differences.

(use the Normal or t distributions when appropriate).

Example 6.17 (Fuel economy). Twelve cars were equipped with radial tires and driven over a test course. Then the same twelve cars (with the same drivers) were equipped with regular belted tires and driven over the same course. After each run, the cars gas economy (in km/l) was measured. Using significance level $\alpha = 0.05$ and the method of critical values, test for a difference in fuel economy between the radial tires and belted tires. Construct a 95% confidence interval for true mean difference due to tire type.

car	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	11.0	12.0
radial	4.2	4.7	6.6	7.0	6.7	4.5	5.7	6.0	7.4	4.9	6.1	5.2
belted	4.1	4.9	6.2	6.9	6.8	4.4	5.7	5.8	6.9	4.7	6.0	4.9
Differences	0.1	-.2	.4	.1	.1	.1	0	.2	.5	.2	.1	.3

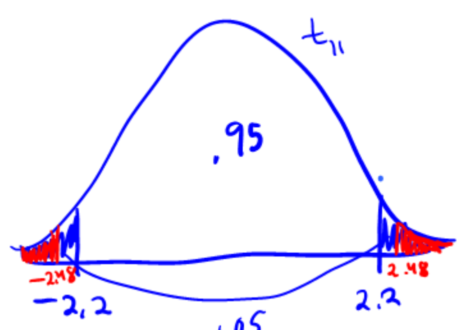
$n=12$, $\bar{d} = 0.142$ $S_d = 0.198$

① $H_0: \mu_d = 0$ $H_A: \mu_d \neq 0$ where μ_d = true mean of the difference between radial and belted tire fuel economy.

② $\alpha = 0.05$

③ I will use the test statistic $K = \dfrac{\bar{d} - 0}{S_d/\sqrt{n}}$ which has t_{n-1} distribution assuming
- H_0 true
- $d_1, \dots, d_{12}$ are iid draws from $N(\mu_d, \sigma_d^2)$. [we probably want to look at a QQ plot]

④ $K = \dfrac{0.142 - 0}{0.198/\sqrt{12}} = 2.48 >$ $t_{11, 1-\alpha/2} = t_{11, .975} = 2.2$



⑤ With $K = 2.48 > 2.2 = t_{11, .975} \Rightarrow$ p-value is $< .05 = \alpha \Rightarrow$ we reject H_0.

⑥ There is enough evidence to conclude that the fuel economy differs between radial and belted tires.

Two sided 95% CI for the true mean fuel economy difference is

$$\left(\bar{d} - t_{11,1-\alpha/2}\frac{s_d}{\sqrt{n}},\ \bar{d} + t_{11,1-\alpha/2}\frac{s_d}{\sqrt{n}}\right) = \left(.142 - t_{11,.975}\frac{0.198}{\sqrt{12}},\ .142 + t_{11,.975}\frac{0.198}{\sqrt{12}}\right)$$

$$= \left(.142 - 2.2\frac{0.198}{\sqrt{12}},\ .142 + 2.2\frac{0.198}{\sqrt{12}}\right)$$

$$= (0.0166,\ 0.2674)$$

We are 95% confident that for the car type studied, radial tires get between 0.0166 km/l and 0.2674 km/l more in fuel economy than belted tires on average.

Example 6.18 (End-cut router)**.** Consider the operation of an end-cut router in the manufacture of a company's wood product. Both a leading-edge and a trailing-edge measurement were made on each wooden piece to come off the router. Is the leading-edge measurement different from the trailing-edge measurement for a typical wood piece? Do a hypothesis test at $\alpha = 0.05$ to find out. Make a two-sided 95% confidence interval for the true mean of the difference between the measurements.

piece	1.000	2.000	3.000	4.000	5.000
leading_edge	0.168	0.170	0.165	0.165	0.170
trailing_edge	0.169	0.168	0.168	0.168	0.169
differences	-.001	.002	-.003	-.003	.001

$n=5$

$\bar{d} = -8\times10^{-4}$

$s_d = 0.0023 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2}$

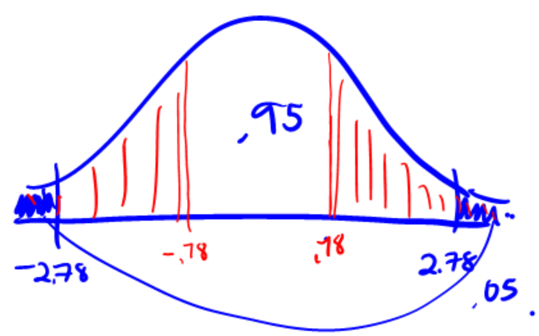
① $H_0: \mu_d = 0$, $H_A: \mu_d \neq 0$

② $\alpha = .05$

③ Since σ_d is unknown and $n=5<25$, we use $K = \dfrac{\bar{d}-0}{s_d/\sqrt{n}}$ and assume $d_1, \ldots, d_5 \overset{iid}{\sim} N(\mu_d, \sigma_d^2)$. Then, if H_0 holds, $K \sim t_{n-1} = t_4$

④ $K = \dfrac{-8\times10^{-4} - 0}{0.0023/\sqrt{5}} = -0.78$

$t_{4,1-\alpha/2} = t_{4,.975} = 2.78$



⑤ Since $|K| = .78 < t_{4,.975} = 2.78$, $\Rightarrow$ the p-value is greater than .05

So, we fail to reject H_0.

⑥ There is not enough evidence to conclude that the leading-edge measurements differ significantly from the trailing-edge measurements.

Two-sided 95% CI for μ_d:

$$\left(\bar{d} - t_{4,1-\alpha/2}\frac{s_d}{\sqrt{n}}, \bar{d} + t_{4,1-\alpha/2}\frac{s_d}{\sqrt{n}}\right)$$

$$= \left(-8\times 10^{-4} - 2.78\frac{0.0023}{\sqrt{5}}, -8\times 10^{-4} + 2.78\frac{0.0023}{\sqrt{5}}\right)$$

$$= (-.00358, .00198)$$

We are 95% confident that the true mean difference between leading-edge and trailing-edge measurements is between $-.00358$ in and $.00198$ in.

6.4.2 Two-sample data

Paired differences provide inference methods of a special kind for comparison. Methods that can be used to compare two means where two different *unrelated* samples will be discussed next.

Examples:

SAT Scores of high school A vs. high school B.

Severity of a disease in men vs. women.

heights of New Zealanders vs. heights of Ethiopians.

Coefficients of friction after wear of sandpaper A vs. B.

Notation:

Sample	1	2
sample size	n_1	n_2
true mean	μ_1	μ_2
sample mean	$\bar{x}_1$	$\bar{x}_2$
True variance	σ_1^2	σ_2^2
Sample variance	s_1^2	s_2^2

6.4.2.1 Large samples ($n_1 \geq 25, n_2 \geq 25$)

The difference in sample means $\overline{x}_1 - \overline{x}_2$ is a natural statistic to use in comparing μ_1 and μ_2.

If σ_1 and σ_2 are **known**, then Proposition 5.1 tells us

$E\overline{X}_1 = \mu_1 \quad Var\,\overline{X}_1 = \frac{\sigma_1^2}{n_1}, \quad E\overline{X}_2 = \mu_2 \quad Var\,\overline{X}_2 = \frac{\sigma_2^2}{n_2}$ by prop 5.1.

$\overline{X}_1, \overline{X}_2$ are independent because samples are unrelated.

$$\mathrm{E}(\overline{X}_1 - \overline{X}_2) = E\overline{X}_1 - E\overline{X}_2 \quad \text{by prop 5.1.}$$

$$= \mu_1 - \mu_2,$$

$$\mathrm{Var}(\overline{X}_1 - \overline{X}_2) = Var\,\overline{X}_1 + (-1)^2 Var\,\overline{X}_2 \quad \text{by prop. 5.1}$$

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

If, in addition, n_1 and n_2 are large, ($n_1 \geq 25$, $n_2 \geq 25$),

$$\overline{X}_1 \dot\sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \text{ independent of } \overline{X}_2 \dot\sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right) \text{ by CLT}$$

> If we have two independent normal random variable X, Y then $aX + bY$ is also normal.

So then $\overline{X}_1 - \overline{X}_2$ is approximately normal and

$$\boxed{\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}} \dot\sim N(0,1).$$

So, if we want to test $\mathrm{H}_0 : \mu_1 - \mu_2 = \#$ with some alternative hypothesis, σ_1 and σ_2 are known, and $n_1 \geq 25, n_2 \geq 25$, then we use the statistic

$$K = \frac{(\bar{x}_1 - \bar{x}_2) - \#}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

which has approximately a $N(0,1)$ distribution if

1. H_0 is true
2. The sample 1 points are iid with mean μ_1 and variance σ_1^2, and the sample 2 points are iid with mean μ_2 and variance σ_2^2. and sample 1 is independent of sample 2.

The confidence intervals (2-sided, 1-sided upper, and 1-sided lower, respectively) for $\mu_1 - \mu_2$ are:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\left(-\infty,\ (\bar{x}_1 - \bar{x}_2) + z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

$$\left((\bar{x}_1 - \bar{x}_2) - z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ \infty\right)$$

If σ_1 and σ_2 are **unknown**, and $n_1 \geq 25, n_2 \geq 25$, then we use the statistic

$$K = \frac{(\bar{x}_1 - \bar{x}_2) - \#}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

and confidence intervals (2-sided, 1-sided upper, and 1-sided lower, respectively) for $\mu_1 - \mu_2$:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{1-\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\left(-\infty, \ (\bar{x}_1 - \bar{x}_2) + z_{1-\alpha} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

$$\left((\bar{x}_1 - \bar{x}_2) - z_{1-\alpha} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \ \infty \right)$$

Example 6.19 (Anchor bolts). An experiment carried out to study various characteristics of anchor bolts resulted in 78 observations on shear strength (kip) of 3/8-in. diameter bolts and 88 observations on strength of 1/2-in. diameter bolts. Let Sample 1 be the 1/2 in diameter bolts and Sample 2 be the 3/8 indiameter bolts. Using a significance level of $\alpha = 0.01$, find out if the 1/2 in bolts are more than 2 kip stronger (in shear strength) than the 3/8 in bolts. Calculate and interpret the appropriate 99% confidence interval to support the analysis.

- $n_1 = 88, n_2 = 78$
- $\overline{x}_1 = 7.14, \overline{x}_2 = 4.25$
- $s_1 = 1.68, s_2 = 1.3$

① $H_0: \mu_1 - \mu_2 = 2$, $H_A: \mu_1 - \mu_2 > 2$

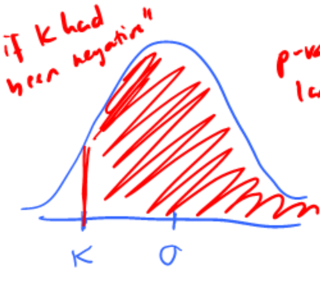
② $\alpha = 0.01$

③ The test statistic is $K = \dfrac{(\bar{x}_1 - \bar{x}_2) - 2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

If we assume Sample 1 is drawn iid w/ mean μ_1 and variance σ_1^2, independently of Sample 2, which is also drawn iid w/ mean μ_2 and variance σ_2^2. And if H_0 is true, then since $n_1 = 88 \geq 25$ and $n_2 = 78 \geq 25$, $K \dot{\sim} N(0,1)$

④ $K = \dfrac{(7.14 - 4.25) - 2}{\sqrt{\frac{(1.68)^2}{88} + \frac{(1.3)^2}{78}}} = 3.84$

"if K had been negative"



p-value would be large, fail to reject

"more extreme"

$$\text{p-value} = P(Z \geq K) = 1 - P(Z \leq K) = 1 - P(Z \leq 3.84)$$
$$= 1 - \Phi(3.84)$$
$$\approx 1 - 1 \approx 0.$$

⑤ With p-value $\approx 0 << \alpha = .01$, we reject H_0!

⑥ There is overwhelming evidence that the $\frac{1}{2}$ in. anchor bolts are more than 2 kip stronger in shear strength than the 3/8 in bolts on average.

99% lower confidence interval

$$\left((\bar{x}_1-\bar{x}_2) - z_{1-\alpha}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}},\ \infty\right) = \left((7.14-4.25) - z_{.99}\sqrt{\frac{1.68^2}{88}+\frac{1.3^2}{78}},\infty\right)$$

$$=(2.89 - 2.33\cdot 0.232,\ \infty)$$

$$=(2.35,\infty).$$

We are 99% confident that the true mean shear strength of the $\frac{1}{2}$ in. anchor bolts is at least 2.35 kip stronger than the true mean shear strength of the 3/8 in. anchor bolts.

6.4.2.2 Small samples

If $n_1 < 25$ or $n_2 < 25$, then we need some **other assumptions** to hold in order to complete inference on two-sample data.

Need independent data and iid Normally distributed

AND

(new) 4 $\sigma_1^2 \approx \sigma_2^2$ 4

A test statistic to test $\text{H}_0 : \mu_1 - \mu_2 = \#$ against some alternative is $K = \dfrac{(\bar{x}_1 - \bar{x}_2) - \#}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

where $S_p^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$ is the "pooled sample variance"

We need this so we know the distn of K

Also assuming 1 H_0 is true, 2 The sample 1 points are iid $N(\mu_1, \sigma_1^2)$, the sample 2 points are iid $N(\mu_2, \sigma_2^2)$, 3 and the sample 1 points are independent of the sample 2 points.

Then $K \sim t_{n_1+n_2-2}$

$1 - \alpha$ confidence intervals (2-sided, 1-sided upper, and 1-sided lower, respectively) for $\mu_1 - \mu_2$ under these assumptions are of the form:

$\nu = n_1 + n_2 - 2$

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\nu, 1-\alpha/2}\, S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

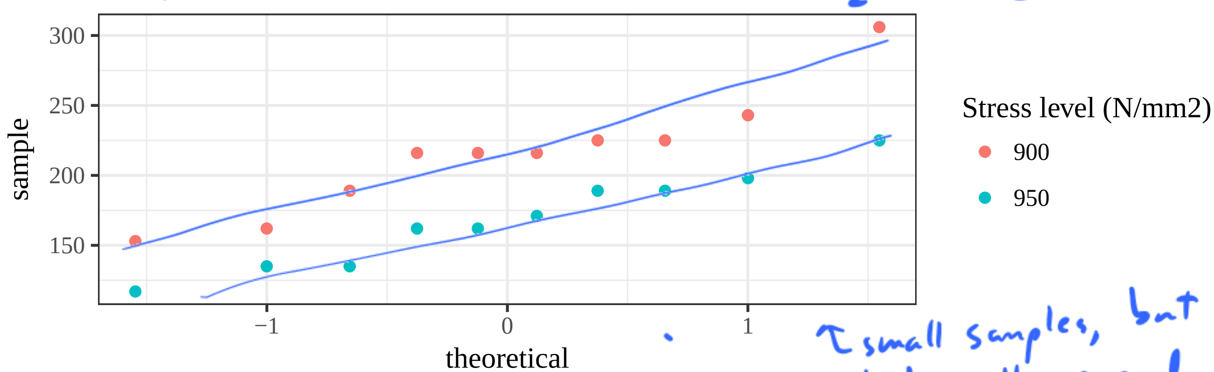
$$\left(-\infty,\ (\bar{x}_1 - \bar{x}_2) + t_{\nu, 1-\alpha}\, S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

$$\left((\bar{x}_1 - \bar{x}_2) - t_{\nu, 1-\alpha}\, S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ \infty\right)$$

Example 6.20 (Springs). The data of W. Armstrong on spring lifetimes (appearing in the book by Cox and Oakes) not only concern spring longevity at a 950 N/mm^2 stress level but also longevity at a 900 N/mm^2 stress level. Let sample 1 be the 900 N/mm^2 stress group and sample 2 be the 950 N/mm^2 stress group. Let's do a hypothesis test to see if the sample 1 springs lasted significantly longer than the sample 2 springs.

900 N/mm2 Stress $n_1 = 10$	950 N/mm2 Stress $n_2 = 10$
216, 162, 153, 216, 225, 216, 306, 225, 243, 189	225, 171, 198, 189, 189, 135, 162, 135, 117, 162

$\bar{x}_1 = 215.1$ $\bar{x}_2 = 168.3$



↑ small samples, but data look pretty normal.

Figure 1: Normal plots of spring lifetimes under two different levels of stress.

① $H_0: \mu_1 - \mu_2 = 0$, $H_A: \mu_1 - \mu_2 > 0$

② $\alpha = 0.05$

③ The test statistic is $K = \dfrac{(\bar{x}_1 - \bar{x}_2) - 0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$. If we assume ① H_0 is true, ② sample 1 is iid $N(\mu_1, \sigma_1^2)$, sample 2 is iid $N(\mu_2, \sigma_2^2)$, ③ sample 1 is indep of sample 2, and ④ $\sigma_1^2 \approx \sigma_2^2$, then $K \sim t_{n_1+n_2-2} = t_{10+10-2} = t_{18}$

④ $S_1 = \sqrt{\frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2} = 42.9$ $S_2 = \sqrt{\frac{1}{n_2-1}\sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2} = 33.1$

$s_1^2 = 1840.41$ and $s_2^2 = 1095.61 \Rightarrow S_p = \sqrt{\dfrac{(10-1)1840.41 + (10-1)1095.61}{10+10-2}} = 38.3$

$K = \dfrac{(215.1 - 168.3) - 0}{38.3\sqrt{\frac{1}{10} + \frac{1}{10}}} = 2.7$

$t_{18, 1-\alpha} = t_{18, .95} = 1.73$ p-value $< .05$

⑤ Since $K > t_{18,.95}$ $\Rightarrow$ p-value $< \alpha = .05$ $\Rightarrow$ reject H_0.

⑥ There is enough evidence to conclude that springs on average last longer if subjected to $900\ N/mm^2$ of stress than $950\ N/mm^2$ of stress.

Example 6.21 (Stopping distance)**.** Suppose μ_1 and μ_2 are true mean stopping distances (in meters) at 50 mph for cars of a certain type equipped with two different types of breaking systems. Suppose $n_1 = n_2 = 6$, $\overline{x}_1 = 115.7$, $\overline{x}_2 = 129.3$, $s_1 = 5.08$, and $s_2 = 5.38$. Use significance level $\alpha = 0.01$ to test $\mathrm{H}_0 : \mu_1 - \mu_2 = -10$ vs. $\mathrm{H}_A : \mu_1 - \mu_2 < -10$. Construct a 2-sided 99

6.5 Prediction intervals

Methods of confidence interval estimation andd hypothesis testing concern the problem of reasoning from sample information to statements about underlying *parameters* of the data generation (such as μ).

Sometimes it is useful to not make a statement about a parameter value, but create bounds on other *individual values* generated by the process.

Question:

> How can we use our data $x_1, \dots, x_n$ to create an interval likely to contain one additional (as yet unobserved) value x_{n+1} from the same data generating mechanism?

Let $X_1, \dots, X_n$ be iid Normal random variables with

$$\text{E}(X_i) = \mu \text{ for all } i = 1, \dots, n$$

$$\text{Var}(X_i) = \sigma^2 \text{ for all } i = 1, \dots, n$$

Then, for $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$ exact (not from CLT)

Let X_{n+1} be an additional observation from the same data generating mechanism.

i.e. X_{n+1} is also $N(\mu, \sigma^2)$

AND

X_{n+1} is independent of $X_1, \dots, X_n \Rightarrow$ indep of $\bar{X}_n$

prop 5.1

$$\mathrm{E}(\overline{X}_n - X_{n+1}) = E(\overline{X}_n) - E(X_{n+1})$$

$\overline{X}_n$, X_{n+1} indep'ndt

$$= \mu - \mu = 0$$

$$\mathrm{Var}(\overline{X}_n - X_{n+1}) = Var(\overline{X}_n) + (-1)^2 Var(X_{n+1})$$

$\overline{X}_n$, X_{n+1} indep.

$$= \frac{\sigma^2}{n} + \sigma^2$$

$$= \left(1 + \frac{1}{n}\right)\sigma^2$$

So,

$$Z = \frac{\overline{X}_n - X_{n+1}}{\sigma\sqrt{1 + \frac{1}{n}}} \sim N(0,1).$$

Generally, σ is unknown, so replace σ by s, and

$$T = \frac{\bar{X}_n - X_{n+1}}{s\sqrt{1+\frac{1}{n}}} \sim t_{n-1}$$

two-sided

Then, $1-\alpha$ **Prediction intervals** for X_{n+1} are

$$\left(\bar{x}_n - t_{n-1, 1-\frac{\alpha}{2}}\, s\sqrt{1+\frac{1}{n}},\ \bar{x} + t_{n-1, 1-\frac{\alpha}{2}}\, s\sqrt{1+\frac{1}{n}}\right)$$