Statistics for Engineers Lecture 7
Two-Sample Inference

Chong Ma

Department of Statistics
University of South Carolina
chongm@email.sc.edu

March 27, 2017

Outline

Two Sample Inference

In this lecture, we discuss **two-sample inference** procedures for the following population parameters:

- The difference of two population **means** $\mu_1 - \mu_2$
- The ratio of two population **variance** σ_2^2/σ_1^2
- The difference of two population **proportion** $p_1 - p_2$

Note these quantities are population-level quantities(involving two different populations), so they are unknown. Our goal is to use sample information to estimate these quantities. For example, we wish to compare

- The population **mean** starting salaries of male and female engineers(compare μ_1 and μ_2). Is there evidence that males have a larger mean starting salary?
- The population **variance** of sound levels from two indoor swimming poopl designs(compare σ_1^2 and σ_2^2). Are the sound-level acoustics of a new design more variable than the standard design?
- The population **proportion** of defectives produced from two different suppliers (compare p_1 and p_2). Are there differences between the two suppliers?

Two Sample Inference

Setting: Suppose that we have two **independent** random samples:

$$\text{Sample 1: } Y_{11}, Y_{12}, \ldots, Y_{1n_1} \overset{i.i.d}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$$

$$\text{Sample 2: } Y_{21}, Y_{22}, \ldots, Y_{2n_2} \overset{i.i.d}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$$

Point estimators: Define the statistics

$$\text{Sample mean for sample 1: } \bar{Y}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1j}$$

$$\text{Sample mean for sample 2: } \bar{Y}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_{2j}$$

$$\text{Sample variance for sample 1: } S_1^2 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(Y_{1j} - \bar{Y}_1)^2$$

$$\text{Sample variance for sample 1: } S_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(Y_{2j} - \bar{Y}_2)^2$$

Outline

CI for $\mu_1 - \mu_2$

In the above setting, we would like to construct a $100(1-\alpha)$ confidence interval for the difference of two population means $\mu_1 - \mu_2$. However, the construction of such an interval depends on the assumptions on the population variances σ_1^2 and σ_2^2. In particular, we consider two cases:

- $\sigma_1^2 = \sigma_2^2$, that is, the two population variances are **equal**.
- $\sigma_1^2 \neq \sigma_2^2$, that is, the two population variances are **not equal**.

CI for $\mu_1 - \mu_2 (\sigma_1^2 = \sigma_2^2)$

Under the assumptions above and when $\sigma_1^2 = \sigma_2^2$, the quantity

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} \sim t_{n_1+n_2-2}$$

where $s_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$, called **pooled sample variance estimator** of the common population variance σ^2, is a weighted average of the two sample variance S_1^2 and S_2^2.

- For the sampling distribution to hold exactly, we need
  - the two random samples to be independent
  - the two population distributions to be normal
  - the two population variances are the same,i.e., $\sigma_1^2 = \sigma_2^2$.

CI for $\mu_1 - \mu_2 (\sigma_1^2 = \sigma_2^2)$

- The sampling distribution $t_{n_1+n_2-2}$ suggests that confidence interval quantiles will come from this t distribution; note that this distribution depends on the **sample sizes** from both samples.
- Because $t \sim t_{n_1+n_2-2}$, the upper quantile $t_{n_1+n_2-2,\alpha/2}$ satisfies that

$$P\left(-t_{n_1+n_2-2,\alpha/2} < \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} < t_{n_1+n_2-2,\alpha/2}\right) = 1 - \alpha$$

- A $100(1-\alpha)$ **percent confidence interval** for the difference of two population means $\mu_1 - \mu_2$ is

$$\left[(\bar{Y}_1 - \bar{Y}_2) \pm t_{n_1+n_2-2,\alpha/2}\sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}\right]$$

CI for $\mu_1 - \mu_2 (\sigma_1^2 = \sigma_2^2)$

- We see that the interval again has the same form:

$$\underbrace{\text{point estimate}}_{\bar{Y}_1 - \bar{Y}_2} \pm \underbrace{\text{quantile}}_{t_{n_1+n_2-2,\alpha/2}} \times \underbrace{\text{standard error}}_{\sqrt{S_p^2(\frac{1}{n_1}+\frac{1}{n_2})}}$$

  We interpret the interval in the same way: **"We are** $100(1-\alpha)$ **percent confident that the population mean difference** $\mu_1 - \mu_2$ **is in this interval."**

- In two-sample situations, it is usually of interest to compare population means μ_1 and μ_2:
  - If the confidence interval for $\mu_1 - \mu_2$ includes 0, this does not suggest that the population means μ_1 and μ_2 are different.
  - If the confidence interval for $\mu_1 - \mu_2$ does not include 0, this suggests that the population means μ_1 and μ_2 are different.

CI for $\mu_1 - \mu_2 (\sigma_1^2 = \sigma_2^2)$

Example. In the vicinity of a nuclear power plant, environmental engineers from the EPA would like to determine if there is a difference between the population mean weight in fish (of the same species) from two location. Independent samples are taken from each location and the following weights (in ounces) are observed:

Location 1:	21.9	18.5	12.3	...	23.0	36.8	26.6
Location 2:	21.0	19.6	14.4	...	16.5		

Construct a 90 percent confidence interval for the population mean weight difference $\mu_1 - \mu_2$, where the mean weight $\mu_1(\mu_2)$ corresponds to location 1(2). In order to use the confidence interval formula under the assumption of $\sigma_1^2 = \sigma_2^2$, we need check if this assumption is valid for the data. The following boxplots roughly indicate that this assumption is not violated. However, we will look at formal statistical inference procedures to compare two population variance soon.

CI for $\mu_1 - \mu_2 (\sigma_1^2 = \sigma_2^2)$

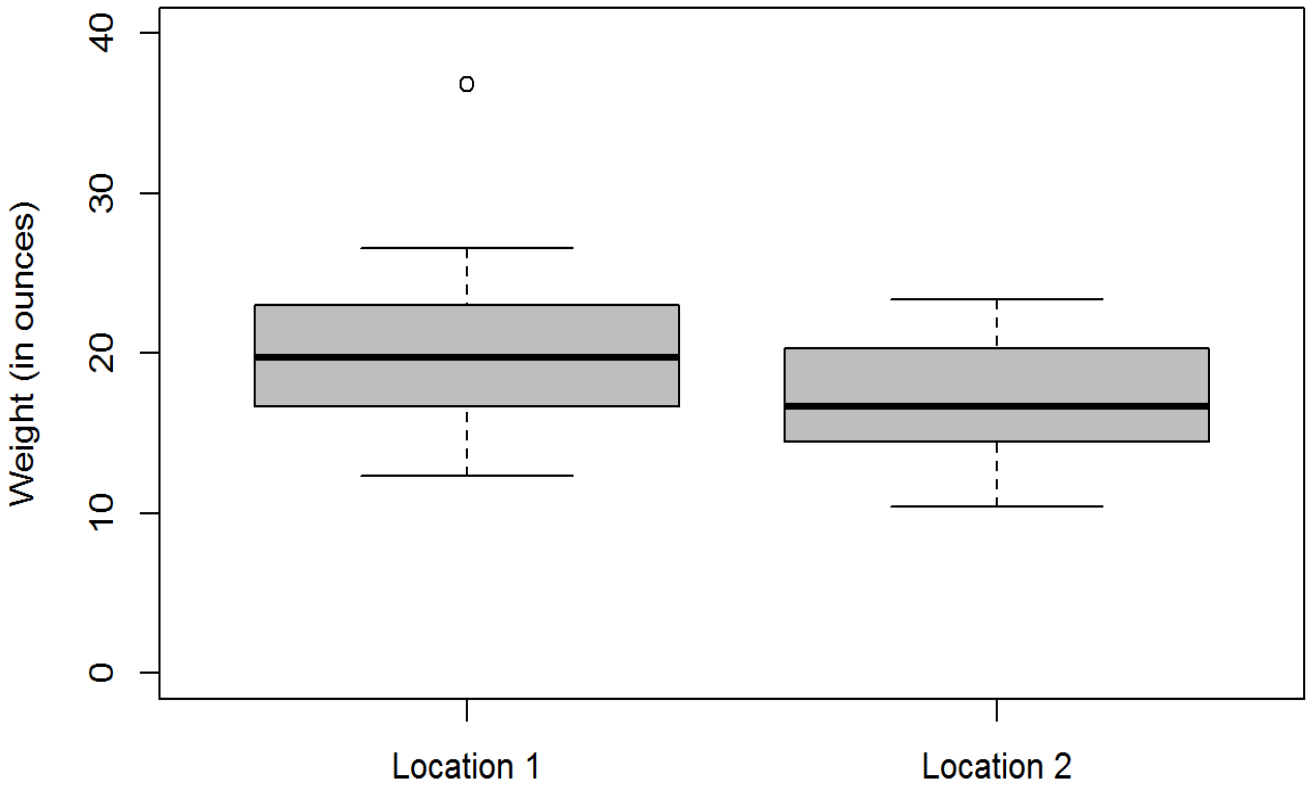

Figure 1: Boxplots of fish weight data by locations. It illustrates that the assumption of $\sigma_1^2 = \sigma_2^2$ is not apparently violated.

CI for $\mu_1 - \mu_2 (\sigma_1^2 = \sigma_2^2)$

- In R, we can use the code

  t.test(loc1,loc2,conf.level=0.9,var.equal=TRUE)

  to get the 90 percent confidence interval for $\mu_1 - \mu_2$, which is $(-0.940, 8.760)$(oz).
- **Interpretation:** We are 90 percent confident that the population mean difference $\mu_1 - \mu_2$ is between -0.940 and 8.760 oz. Note that this interval includes 0. There fore, we do not have sufficient evidence to conclude that the two population mean fish weight is different.
- Only use this confidence interval formula if there is **strong evidence** that the population variance σ_1^2 and σ_2^2 are similar. Put in another way, this confidence interval is not robust to a violation of the equal population variance assumption.
- Like the one-sample t confidence interval for a single population mean μ, this two sample t confidence interval is robust to mild departures from normality.

CI for $\mu_1 - \mu_2 (\sigma_1^2 = \sigma_2^2)$

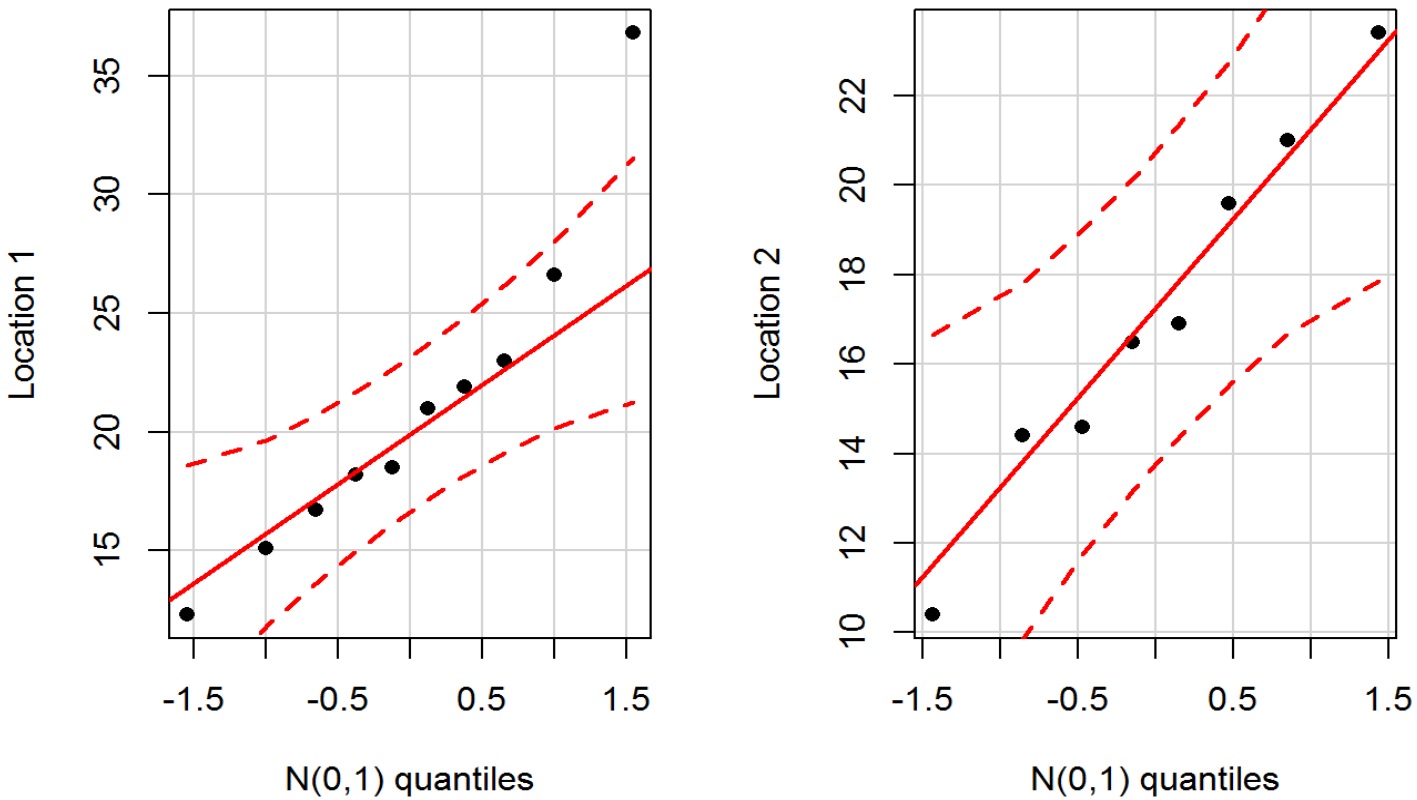

Figure 2: Normal QQ plots for fish weights in two locations.

CI for $\mu_1 - \mu_2 (\sigma_1^2 \neq \sigma_2^2)$

An approximate $100(1-\alpha)$ percent confidence interval for the difference of two population means $\mu_1 - \mu_2$ is given by

$$\left[(\bar{Y}_1 - \bar{Y}_2) \pm t_{\nu,\alpha/2} \sqrt{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})} \right]$$

where the degrees of freedom ν is calculated by

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Remarks: The interval is always approximately valid as long as

- The two samples are independent.
- The two population distributions are approximately normal.

CI for $\mu_1 - \mu_2 (\sigma_1^2 \neq \sigma_2^2)$

Example You are part of a recycling project that is examining how much paper is being discarded (not recycled) by employees at two large plants. These data are obtained on the amount of white paper thrown out per year by employees(data are in hundreds of pounds). Sampled of employees at each plant were randomly selected.(See the whole data in R tutorial.)

Plant 1:	3.01	2.58	3.04	...	2.23	1.92	3.02
Plant 2:	3.99	2.08	3.66	...	2.74	4.81	

Construct a 95 percent confidence interval for the population mean difference $\mu_1 - \mu_2$, where the mean amount discarded $\mu_1(\mu_2)$ corresponds to Plant 1(2).

Note that boxplots of the two samples suggest that the equal population variance assumption is doubtful. Therefore, we calculate the confidence interval for $\mu_1 - \mu_2$ under the non-equal variance assumption, by using the R code

t.test(plant1,plant2,conf.level=.95,var.equal=FALSE)

CI for $\mu_1 - \mu_2 (\sigma_1^2 \neq \sigma_2^2)$

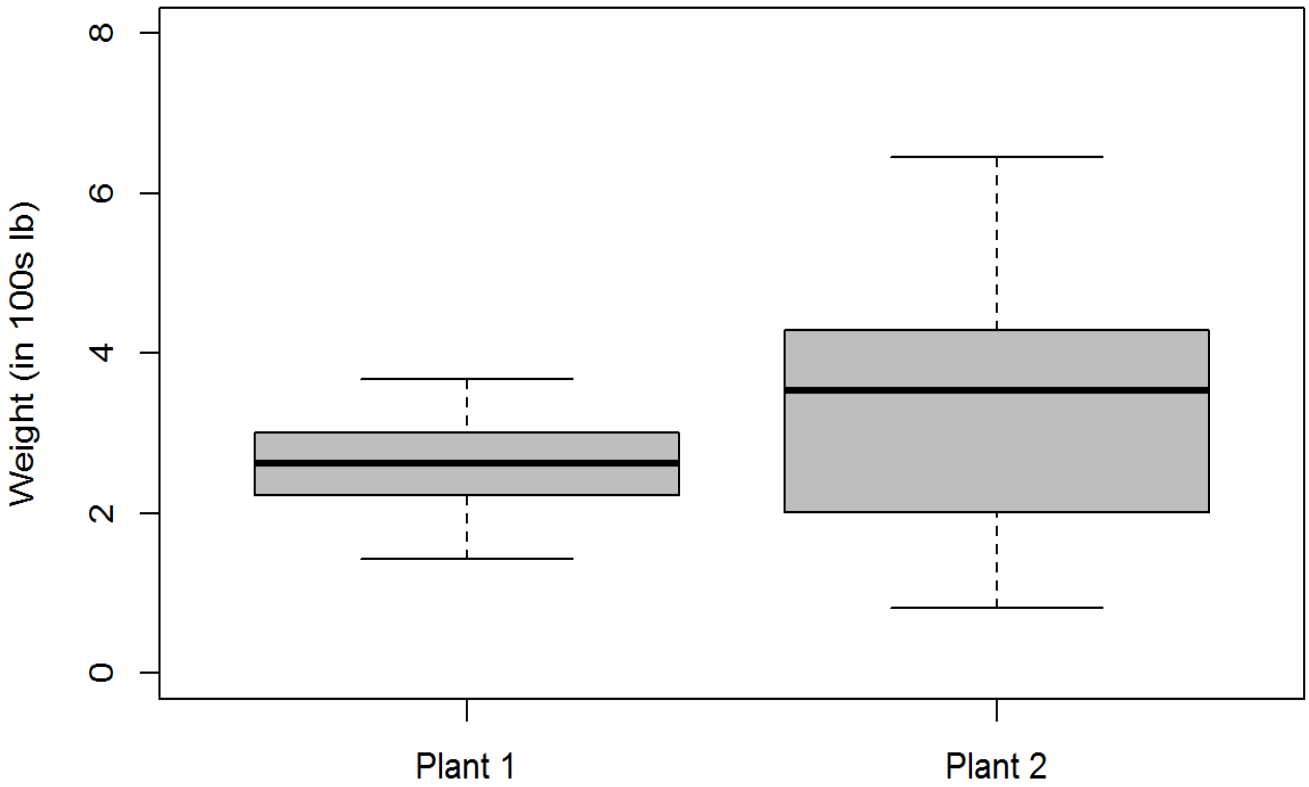

Figure 3: Boxplots of discarded white paper amount (in 100s lb) for two Plants. It indicates that the equal population variance assumptions is doubtful. The spread in the two boxplots is markedly different.

CI for $\mu_1 - \mu_2 (\sigma_1^2 \neq \sigma_2^2)$

- A 95 percent confidence interval for the population mean difference $\mu_1 - \mu_2$ is

$$(-1.461, -0.069)100\mathrm{s}(\mathrm{lbs})$$

- **Interpretation:** We are 95 percent confident that the population mean difference $\mu_1 - \mu_2$ is between -146.1 and -6.9 lbs. Because the interval does not contain 0, we have sufficient evidence to conclude that the population mean amount of discarded paper is smaller for Plant 1 than it is for Plant 2.
- The Normal QQ plots for the samples of amount of discarded paper do not reveal serious departures from normality assumption.
- When unsure about which interval to use, go with the unequal variance interval. The penalty for using it when $\sigma_1^2 = \sigma_2^2$ is much smaller than the penalty for using the equal variance interval when $\sigma_1^2 \neq \sigma_2^2$.

CI for $\mu_1 - \mu_2 (\sigma_1^2 \neq \sigma_2^2)$

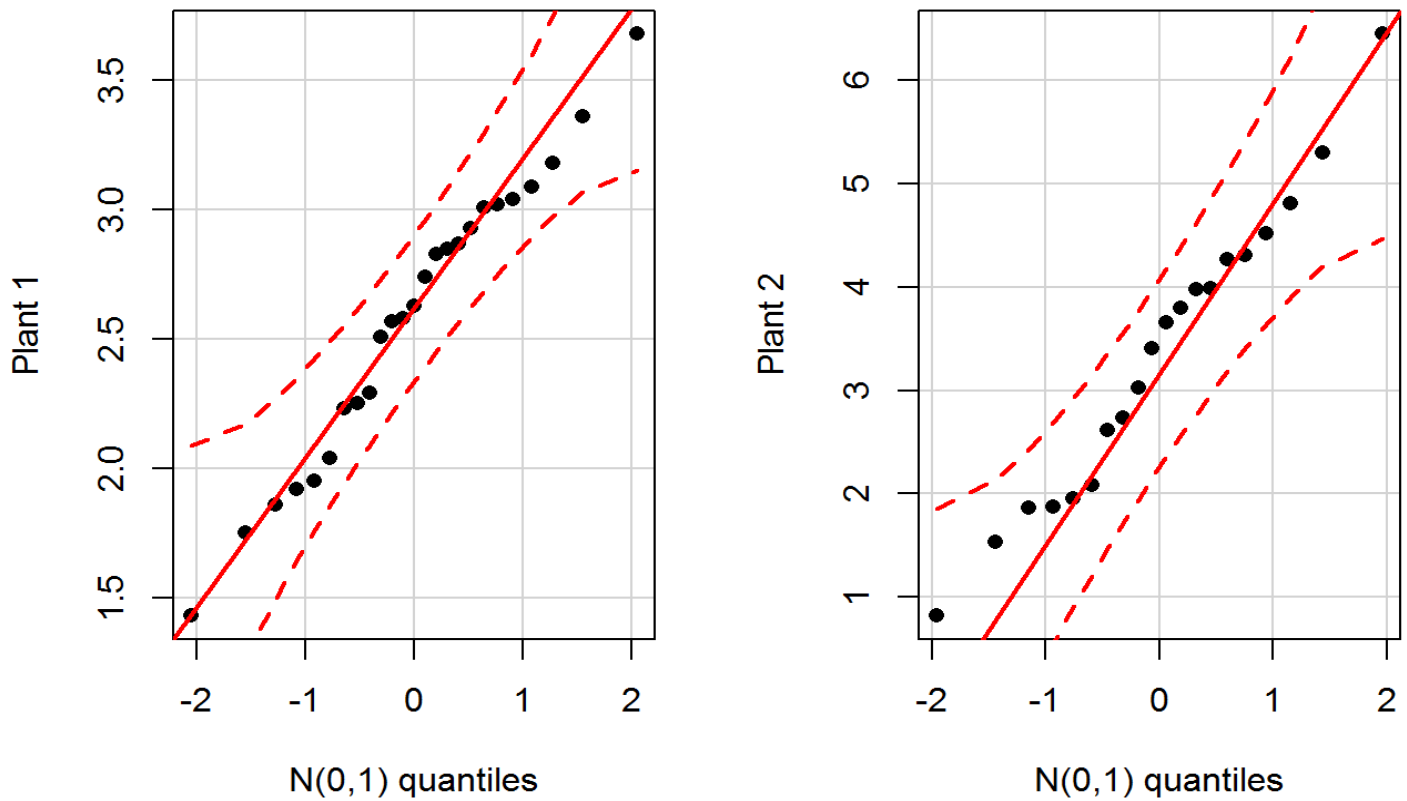

Figure 4: Normal QQ plots for the white paper data. There are no cause to question the normality assumption.

CI for $\mu_1 - \mu_2$(Matched Pairs)

A matced-pairs design is an experimental design where one obtains a pair of measurements on each individual(e.g., employee, material, machine, etc.):

- one measurement corresponds to "Treatment 1"
- the other measurement corresponds to "Treatment 2"
- Clearly, the two samples are no longer independent. Each individual contributes a response to both samples.

This type of design removes variation **among** the individuals. This allows you to compare the two treatments(e.g., before/after working environment) under more **homogeneous** conditions where only variation within individuals is present(that is, the variation arising from the difference in the two treatments).

Design	Sources of Variation
Two Independent Samples	Among Individuals + Within Individuals
Matched Pairs	Within Individuals

CI for $\mu_1 - \mu_2$(Matched Pairs)

Advantages: When you remove extra variability, this enables you to compare the two experimental conditions(treatments) more precisely. This gives you a better chance of identifying a difference between the treatment means if one really exists.

Implementation: Data from matched pairs experiments are analyzed by examining the difference in responses of the two treatments. Specially, compute

$$D_j = Y_{1j} - Y_{2j}$$

for each individual $j = 1, 2, \ldots, n$. After doing this, we have essentially created a "one sample problem", where our data are now

$$D_1, D_2, \ldots, D_n$$

the so-called **data differences**.

CI for $\mu_1 - \mu_2$(Matched Pairs)

The one sample $100(1-\alpha)$ percent confidence interval

$$\bar{D} \pm t_{n-1,\alpha/2}\frac{S_D}{\sqrt{n}}$$

is an interval estimate for the **population mean difference** $\mu_D = \mu_1 - \mu_2$ between two treatment groups, where $\bar{D}$ and S_D are the sample mean and sample standard deviation of the differences, respectively. If the two population means are the same, then $\mu_D = 0$. Therefore,

- If the confidence interval for μ_D includes 0, this does not suggest that the two population means are different.
- If the confidence interval for μ_D does not includes 0, this suggests that the two population means are different.

CI for $\mu_1 - \mu_2$(Matched Pairs)

Example. Ergonomics experts hired by a large company designed a study to determine more varied work conditions would have any impact on arm movement. The data were obtained on a random sample of $n = 26$ employees. Each observation is the amount of time, expressed as a percentage of the total time observed, during which arm elevation **was below 30 degrees**. This percentage is a surrogate for the percentage of time spent on repetitive tasks. The two measurements from each employee were obtained 18 months apart. During this 18-month period, work conditions were "changed" by the ergonomics team, and subjects were allowed to engage in a wider variety of work tasks.
Question: Does the population mean time (during which elevation is below 30 degrees) decrease after the ergonomics team changes the working conditions?

CI for $\mu_1 - \mu_2$(Matched Pairs)

Individual	Before	After	Difference
1	81.3	78.9	2.4
2	87.2	91.4	-4.2
3	86.1	78.3	7.8
...	...	...	...
24	91.1	81.8	9.3
25	97.5	91.6	5.9
26	70.0	74.2	-4.2

Table 1: Ergonomics data. Percentage of time arm elevation was less than 30 degrees. The data differences $D_j = Y_{1j} - Y_{2j}$ are listed as well. The completed data is available in R tutorial.

Analysis: Use R code **t.test(diff,conf.level=0.95)** to construct a 95 percent confidence interval for $\mu_D = \mu_1 - \mu_2$ is $(3.64, 9.89)$.

CI for $\mu_1 - \mu_2$(Matched Pairs)

Interpretation: We are 95 percent confident that the population mean difference $\mu_D = \mu_1 - \mu_2$ is between 3.64 and 9.89 percent. Since the interval does not conclude 0 and only contains positive values, it suggests that the population mean percentage of time that arm elevation is below 30 degrees is larger in the "before" condition than in the "after" condition.
Assumation: In matched paris experiment, the relevant assumptions are

1. The individuals are a random sample from the target population.
2. The **data difference** $D_1, D_2, \ldots, D_n$ are normally distributed.

Ergonomics data: A normal qq plot for the data differences is given in figure . There might be a mild departure from normality in the upper tail. However, because the one-sample t confidence interval is generally robust to these mild departures, here the slight departure does not likely affect our conclusion.

CI for $\mu_1 - \mu_2$(Matched Pairs)

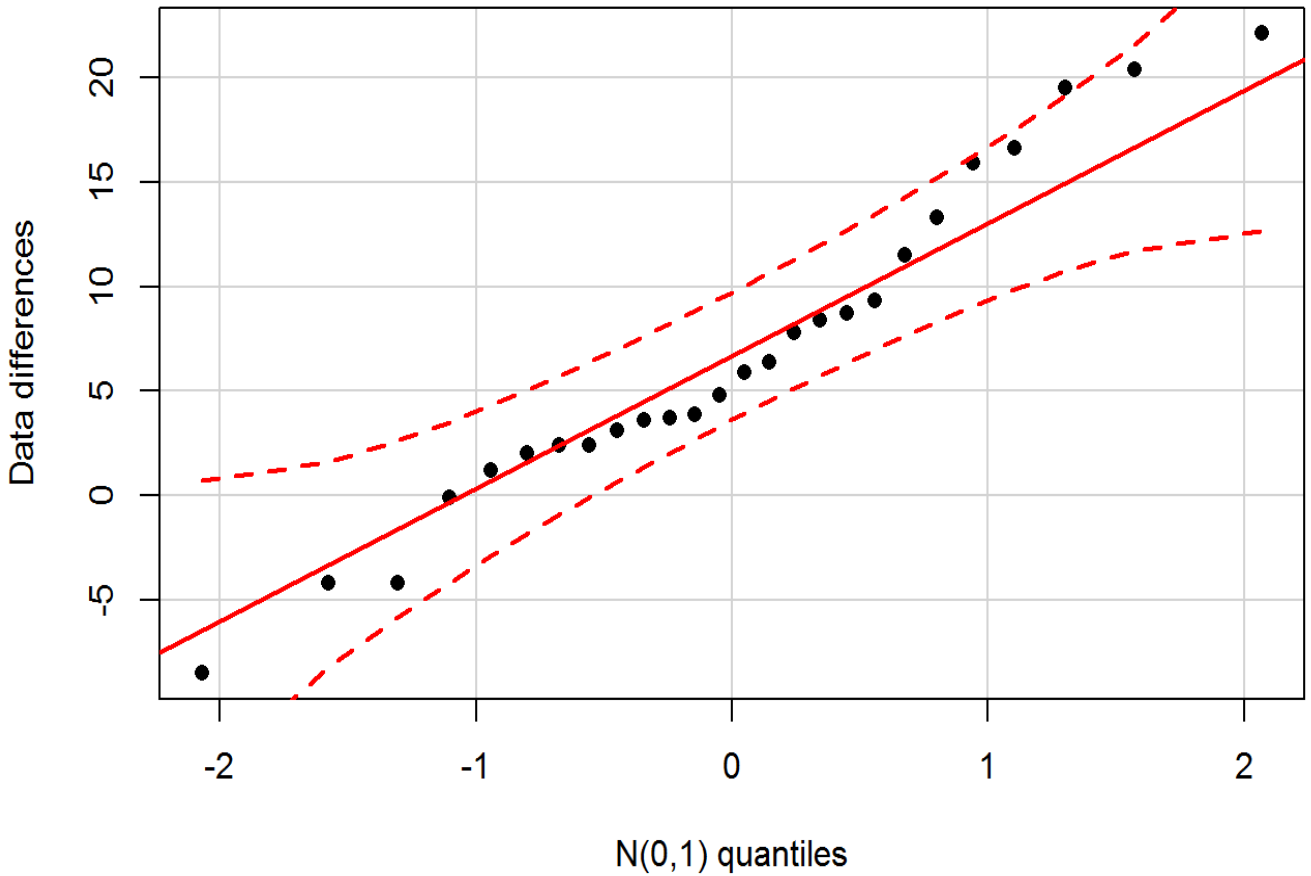

Figure 5: Normal qq plot for the ergonomics data. The observed data difference are plotted versus the theoretical quantiles from a normal distribution. The line added passes through the first and third theoretical quantiles.

Outline

CI for $\sigma_1^2 - \sigma_1^2$

Recall that when we computed a confidence interval for $\mu_1 - \mu_2$, the difference of the population means (with independent samples), we proposed two intervals:

- one interval that assumed $\sigma_1^2 = \sigma_2^2$.
- one interval that assumed $\sigma_1^2 \neq \sigma_2^2$.

We now propose a confidence interval procedure that can be used to determine which assumption is more appropriate.

Setting: Suppose that we have two **independent** random samples:

$$\text{Sample 1: } Y_{11}, Y_{12}, \ldots, Y_{1n_1} \overset{i.i.d}{\sim} \mathrm{N}(\mu_1, \sigma_1^2)$$

$$\text{Sample 2: } Y_{21}, Y_{22}, \ldots, Y_{2n_2} \overset{i.i.d}{\sim} \mathrm{N}(\mu_2, \sigma_2^2)$$

Goal: To construct a $100(1-\alpha)$ percent confidence interval for the **ratio** of population variances σ_2^2/σ_1^2.

CI for $\sigma_1^2 - \sigma_1^2$

Result: Under the setting described above,

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$$

an F distribution with (numerator) $n_1 - 1$ and (denominator) $n_2 - 1$ degrees of freedom. Some characteristics for F distribution:

- continuous, skewed right, and always positive
- indexed by two **degrees of freedom** parameters ν_1 and ν_2; these are usually integers and are related to sample sizes.
- the **mean** of an F distribution is close to 1 (regardless of the values of ν_1 and ν_2).
- The F pdf formula is complicated and is unnecessary for our purposes. We use R commands **pf(x,ν_1,ν_2)** to compute the cdf of $F_Q(x)$ and **qf(p,ν_1,ν_2)** to compute the pth quantile for $Q \sim F(\nu_1, \nu_2)$.

CI for $\sigma_1^2 - \sigma_1^2$

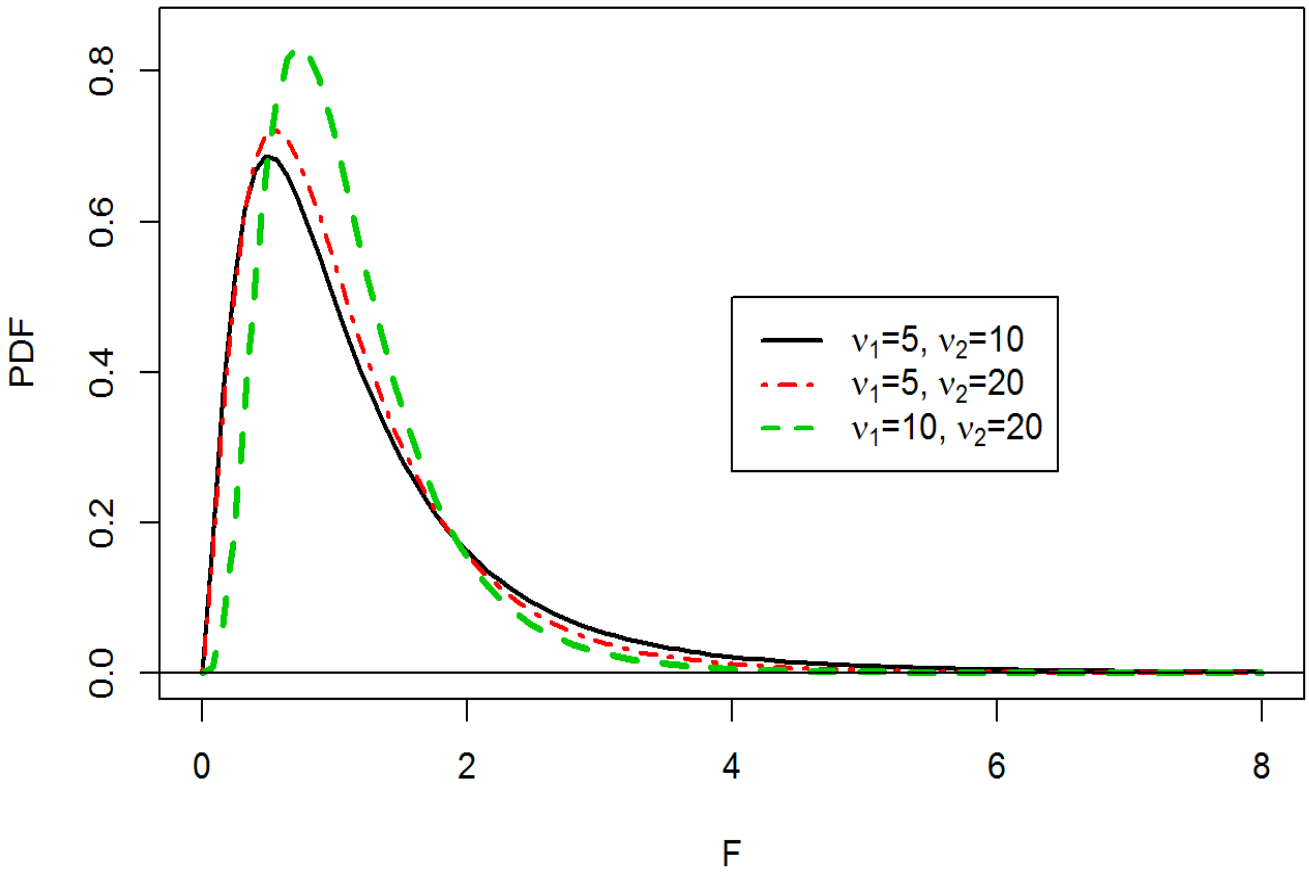

Figure 6: F distribution with various degrees of freedom.

CI for $\sigma_1^2 - \sigma_1^2$

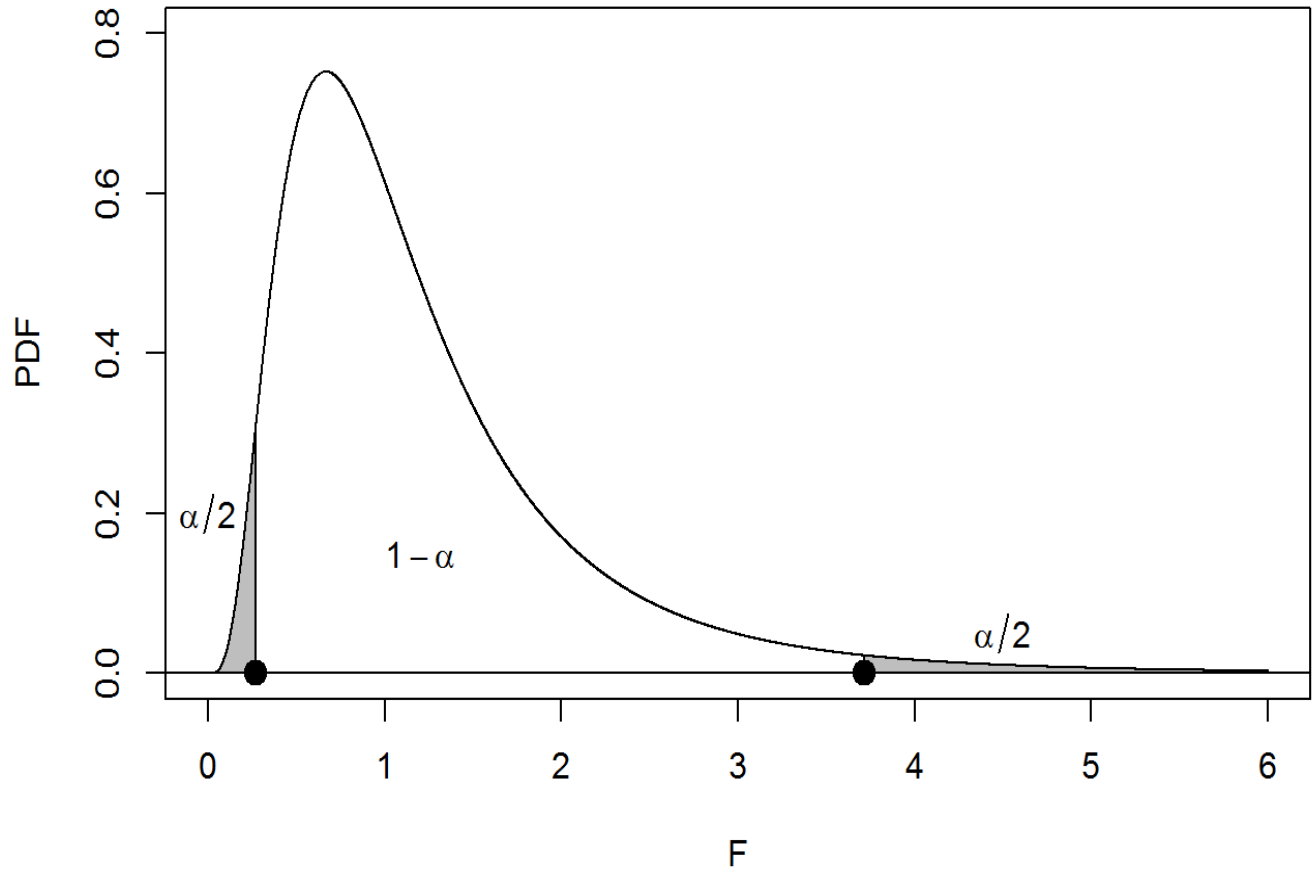

Figure 7: An F pdf with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. The upper $\alpha/2$ and lower $\alpha/2$ areas are shaded.

CI for $\sigma_1^2 - \sigma_1^2$

Denote that $F_{n_1-1,n_2-1,\alpha/2}$ is the **upper** $\alpha/2$ quantile and $F_{n_1-1,n_2-1,1-\alpha/2}$ is the **lower** $1-\alpha/2$ quantile. A $100(1-\alpha)$ percent confidence interval for the ratio of the population variances σ_2^2/σ_1^2 is

$$(\frac{S_2^2}{S_1^2}F_{n_1-1,n_2-1,1-\alpha/2}, \frac{S_2^2}{S_1^2}F_{n_1-1,n_2-1,\alpha/2})$$

Since for any value of $\alpha, 0 < \alpha < 1$, we have

$$\begin{aligned} 1-\alpha &= P(F_{n_1-1,n_2-1,\alpha/2} < \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} < F_{n_1-1,n_2-1,1-\alpha/2}) \\ &= P(\frac{S_2^2}{S_1^2}F_{n_1-1,n_2-1,1-\alpha/2} < \frac{\sigma_2^2}{\sigma_1^2} < \frac{S_2^2}{S_1^2}F_{n_1-1,n_2-1,\alpha/2}) \end{aligned}$$

CI for $\sigma_1^2 - \sigma_1^2$

Example. Two automated filling processes are used in the production of automobile paint. The target weight of each process is 128.0 fluid oz(1 gallon). There is little concern about the process population mean fill amounts (no complaints about under/overfilling on average). However, there is concern that the population variation levels between the two processes are different. To test this claim, industrial engineers took independent random samples of $n_1 = 24$ and $n_2 = 24$ gallons of paint and observed the fill amounts.

	127.75	127.87	127.86	127.92
Process 1:	...	...	...	...
	127.74	127.78	127.85	127.96
	127.90	127.90	127.74	127.93
Process 1:	...	...	...	...
	127.82	127.92	127.71	127.78

CI for $\sigma_1^2 - \sigma_1^2$

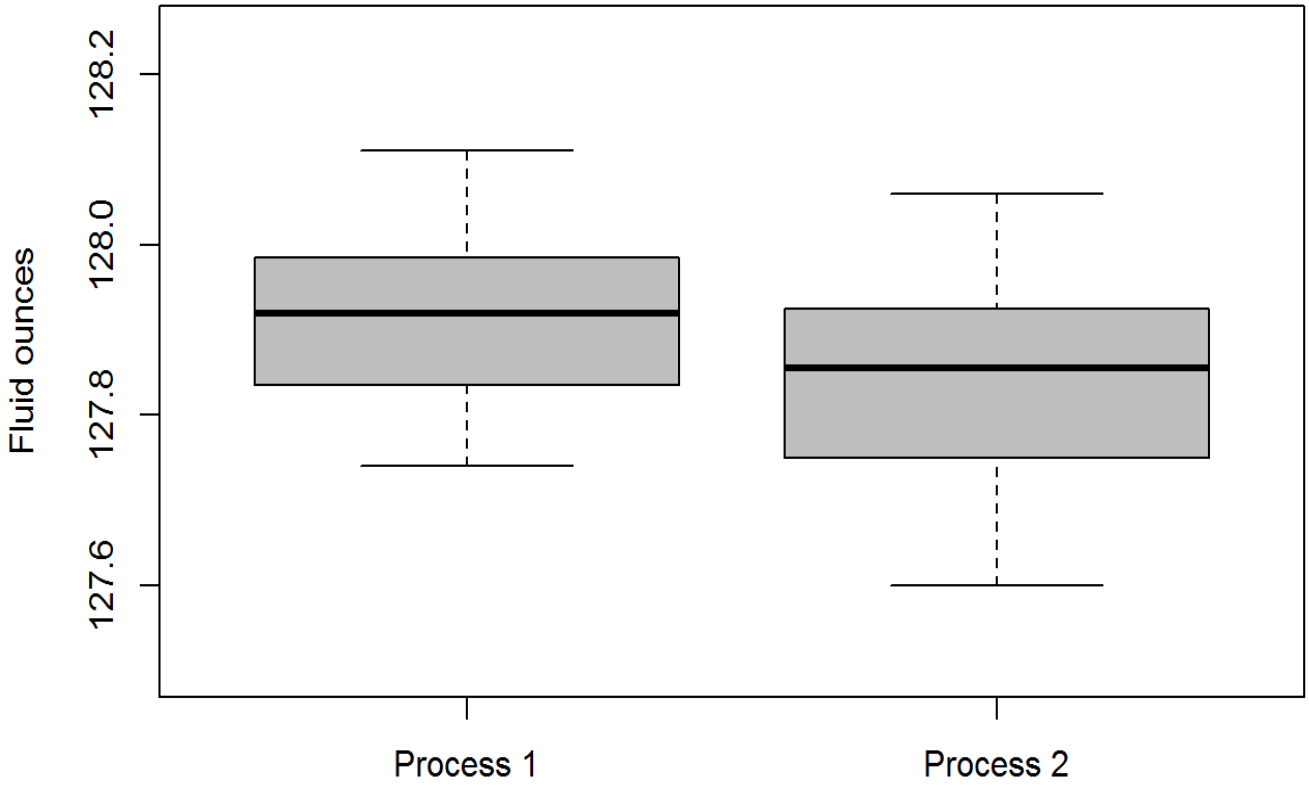

Figure 8: Boxplots of paint fill volume data.

CI for $\sigma_1^2 - \sigma_1^2$

95% CI: $(0.589, 3.145)$
Interpretation: We are 95 percent confident that the ratio of the population variance σ_2^2/σ_1^2 is between 0.589 and 3.145. Because this interval includes "1", we do not have evidence to conclude that the two population variances σ_1^2 and σ_2^2 are different.
Remarks: Like the χ^2 interval for single population variance σ^2, the two-sample F interval for the ratio of two population variances σ_2^2/σ_1^2 is **not robust** to normality departures. This is true because the sampling distribution

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$$

depends critically on the normal distribution assumption for both populations. If either underlying population distribution is non-normal (non-Gaussian), then the confidence interval formula for σ_2^2/σ_1^2 is not to be used.

CI for $\sigma_1^2 - \sigma_1^2$

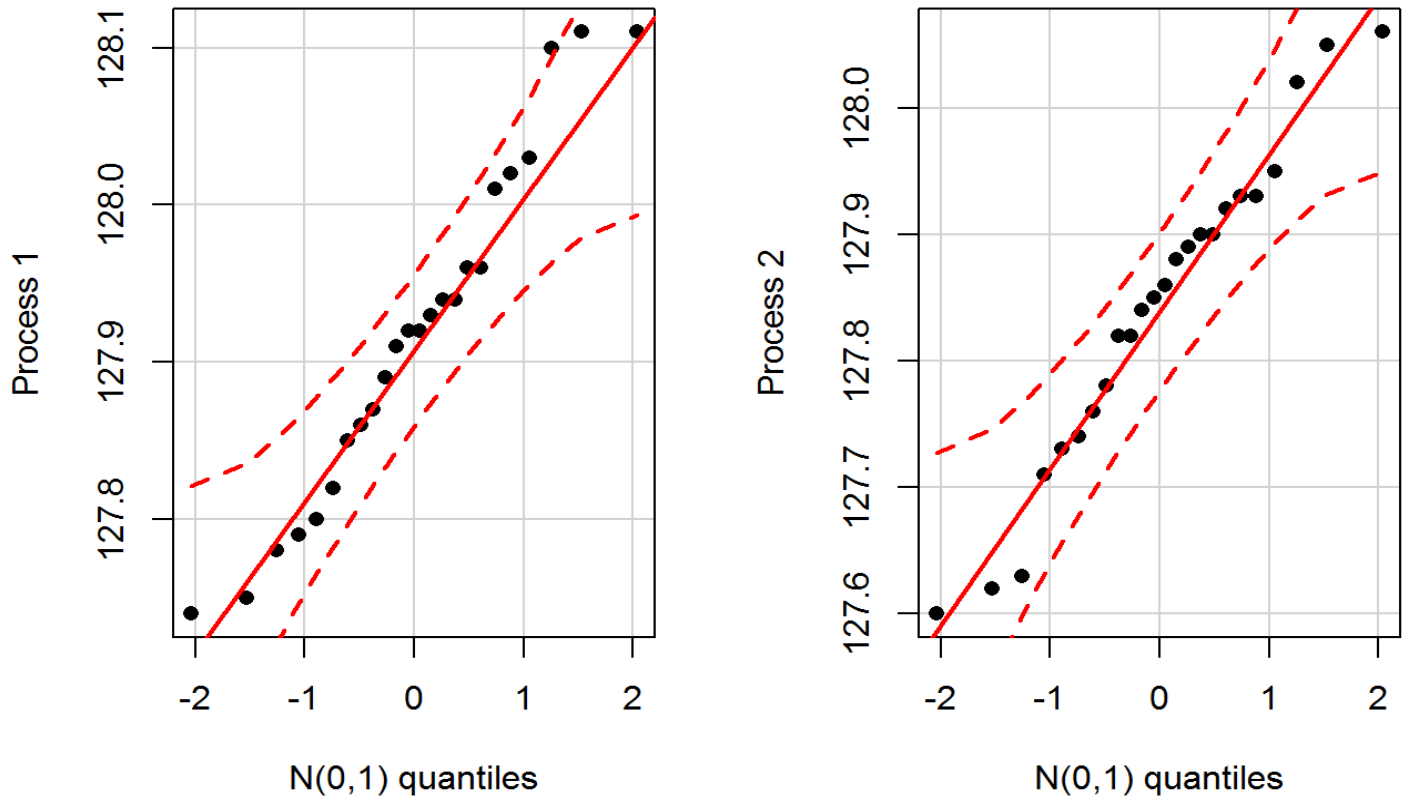

Figure 9: Normal QQ plots for the paint fill volume data. There is no major cause for concern here.

Outline

CI for $p_1 - p_2$

Setting: We now extend our confidence interval procedure for a single population proportion p to **two populations**. Define

$$p_1 = \text{population proportion in Population 1}$$
$$p_2 = \text{population proportion in Population 2}$$

For example, we might want to compare the proportion of

- defective circuit boards for two different suppliers.
- satisfied customers before and after a product design change(e.g., Facebook, etc)
- on-time payments for two classes of customers
- HIV positive for individuals in two demographics classes.

Goal: We would like to construct a $100(1-p)$ percent confidence interval for $p_1 - p_2$, the difference of two population proportions.

CI for $p_1 - p_2$

Point Estimators: We assume that there are two independent random samples of individuals (one sample for each population to be compared). Define

$$Y_1 = \text{number of "successes" in Sample 1} \sim \text{Binomial}(n_1, p_1)$$
$$Y_2 = \text{number of "successes" in Sample 2} \sim \text{Binomial}(n_2, p_2)$$

Result: We need the following sampling distribution result. When the sample sizes n_1 and n_2 are large,

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0,1)$$

The $100(1-p)$ **percent confidence interval** for $p_1 - p_2$ is

$$[(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}]\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

CI for $p_1 - p_2$

- The form of the interval is again like:

$$\underbrace{\text{point estimate}}_{\hat{p}_1 - \hat{p}_2} \pm \underbrace{\text{quantile}}_{z_{\alpha/2}} \times \underbrace{\text{standard error}}_{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

- For the Z sampling distribution to hold approximately, we need
  - The two random samples to be independent
  - The sample sizes n_1 and n_2 need to be "large"; common rules of thumb are to require

$$n_i \hat{p}_i \geq 5, n_i(1 - \hat{p}_i) \geq 5$$

Under these conditions, the Central Limit Theorem should adequately approximate the true sampling distribution of Z, thereby making the confidence interval formula above approximately valid.

CI for $p_1 - p_2$

Example. A large public health study was conducted to estimate the prevalence and to identify risk factors of hepatitis B virus (HBV) infection among Irish prisoners. Two independent samples of female ($n_1 = 82$) and male ($n_2 = 555$) prisoners were obtained from five prisons in Ireland:

- 18 out of 82 female prisoners were HBV-positive
- 28 out of 555 male prisoners were HBV-positive

Find a 95 percent confidence interval for $p_1 - p_2$, the difference in the population proportions for the two genders(Female=1; Male=2).
Analysis: There is no internal function in R to directly calculate the confidence interval for $p_1 - p_2$. Using the one I wrote, we can get the according 95% confidence interval $(0.078, 0.260)$.
Interpretation: We are 95% confident that the difference of the population proportions $p_1 - p_2$ is between 0.078 and 0.260. Because the interval does not contain 0, it suggests that the population proportion of female prisoners who are HBV positive is larger than the corresponding male population proportion.