

MATH 203: PRINCIPLES OF STATISTICS 1
FALL 2024

TWO SAMPLE PROBLEMS: EXAMPLES

Example 1 Two sections of a class in statistics were taught by two different methods. Scores on a standardized test are shown below.

Class A		Class B	
74	76	78	79
97	75	92	76
79	82	94	93
88	86	78	82
78	100	71	69
93	94	85	84
70			

Do the results present evidence of a difference in the effectiveness of the two methods? Test using $\alpha = 0.01$.

```
x<-c(74,76,97,75,79,82,88,86,78,100,93,94,70)
m<-length(x);xbar<-mean(x); s1<-sd(x)
y<-c(78,79,92,76,94,93,78,82,71,69,85,84)
n<-length(y); ybar<-mean(y); s2<-sd(y)
sp<-sqrt(((m-1)*s1^2+(n-1)*s2^2)/(m+n-2))
tobs<-(xbar-ybar)/(sp*sqrt(1/m+1/n))
alpha<-0.01
t.al<-qt(1-alpha/2,m+n-2)
```

Solution: Here we have $m = 13, n = 12$ and so we must assume that the samples are Normally distributed, and that the population variances are equal. We have

$$\bar{x} = 84 \quad s_1 = 9.7125349 \quad \bar{y} = 81.75 \quad s_2 = 8.2255146$$

so that $s_p = 9.0319481$. We may test the hypotheses

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

using the test statistic

$$T = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \quad \text{with observed value} \quad T_{obs} = \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}} = 0.6222908$$

Here the Rejection Region for the two-sided alternative is $\{|T| \geq t_{\alpha/2, m+n-2}\}$, where

$$t_{\alpha/2, m+n-2} = t_{0.005, 23} = 2.8073$$

so we do **not** reject H_0 in favour of H_a for $\alpha = 0.01$.

Example 2 The table below shows the observed pollution indexes of air samples in two areas of a city. Test the hypothesis that the mean pollution indexes are the same

Area A		Area B	
2.92	4.69	1.84	3.44
1.88	4.86	0.95	3.69
5.35	5.81	4.26	4.95
3.81	5.55	3.18	4.47

for the two areas. Test using $\alpha = 0.05$.

```
x<-c(2.92,4.69,1.88,4.86,5.35,5.81,3.81,5.55)
m<-length(x);xbar<-mean(x); s1<-sd(x)
y<-c(1.84,3.44,0.95,3.69,4.26,4.95,3.18,4.47)
n<-length(y); ybar<-mean(y); s2<-sd(y)
sp<-sqrt(((m-1)*s1^2+(n-1)*s2^2)/(m+n-2))
tobs<-(xbar-ybar)/(sp*sqrt(1/m+1/n))
alpha<-0.05
t.al<-qt(1-alpha/2,m+n-2)
ci<-(xbar-ybar)+c(-1,1)*t.al*sp*sqrt(1/m+1/n)
```

Solution: Here we have $m = 8$, $n = 8$ and so we must assume that the samples are Normally distributed, and that the population variances are equal. We have

$$\bar{x} = 4.35875 \quad s_1 = 1.3827554 \quad \bar{y} = 3.3475 \quad s_2 = 1.3541022$$

so that $s_p = 1.3685038$. We may test the hypotheses

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

using the test statistic

$$T = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \quad \text{with observed value} \quad T_{obs} = \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}} = 1.4778914$$

Here the Rejection Region for the two-sided alternative is $\{|T| \geq t_{\alpha/2, m+n-2}\}$, where

$$t_{\alpha/2, m+n-2} = t_{0.025, 14} = 2.1448$$

so we do **not** reject H_0 in favour of H_a for $\alpha = 0.05$.

We may construct a 95% Confidence Interval for $\mu_1 - \mu_2$ as

$$(\bar{x} - \bar{y}) \pm t_{0.025, 14} s_p \sqrt{\frac{1}{m} + \frac{1}{n}} = (-0.4563, 2.4788).$$

Example 3 A closer examination of the records of the air samples in **Example 2** reveals that each line of the data actually represents readings on the same day: 2.92 and 1.84 are from day 1, and so forth.

Area A	Area B	Difference
2.92	1.84	1.08
1.88	0.95	0.93
5.35	4.26	1.09
3.81	3.18	0.63
4.69	3.44	1.25
4.86	3.69	1.17
5.81	4.95	0.86
5.55	4.47	1.08

With this new information, test the hypothesis that the mean pollution indexes are the same for the two areas. Test using $\alpha = 0.05$.

```
d<-x-y
dbar<-mean(d)
sigd<-sd(d)
tobs<-dbar/(sigd*sqrt(1/n))
alpha<-0.05
t.al<-qt(1-alpha/2,n-1)
ci<-dbar+c(-1,1)*t.al*sigd/sqrt(n)
```

Solution: Here we have a paired sample case, with $n = 8$ and so we must assume that the differences are Normally distributed. We have

$$\bar{d} = 1.01125 \quad s_d = 0.1973711.$$

We may test the hypotheses concerning $\mu_d = \mu_1 - \mu_2$

$$H_0 : \mu_d = 0$$

$$H_a : \mu_d \neq 0$$

using the test statistic

$$T = \frac{\bar{D}}{S_d/\sqrt{n}} \quad \text{with observed value} \quad T_{obs} = \frac{\bar{d}}{s_d/\sqrt{n}} = 14.49172.$$

Here the Rejection Region for the two-sided alternative is $\{|T| \geq t_{\alpha/2, n-1}\}$, where

$$t_{\alpha/2, n-1} = t_{0.025, 7} = 2.3646$$

so we **do** reject H_0 in favour of H_a for $\alpha = 0.05$.

We may construct a 95% Confidence Interval for μ_d as

$$\bar{d} \pm t_{0.025, 7} \frac{s_d}{\sqrt{n}} = (0.8462, 1.1763).$$

Example 4 A random sample of 170 recently sold properties in a city was collected to examine the additional cost of waterfront properties. The summary of the data (in thousands of dollars) is reported in the following table.

	Sample size	Sample mean	Sample sd.
Non-Waterfront	100	219.8966	94.62715
Waterfront	70	319.9064	153.3038

Is there any evidence that waterfront properties do cost more than non-waterfront properties? Test at the $\alpha = 0.05$ level.

```
m<-100;xbar<-219.8966; s1<-94.62715
n<-70; ybar<-319.9064; s2<-153.3038
se.val<-sqrt(s1^2/m+s2^2/n)
tobs<-(xbar-ybar)/se.val
alpha<-0.05
z.al<-qnorm(1-alpha)
ci<-(xbar-ybar)+c(-1,1)*qnorm(1-alpha/2)*se.val
```

Solution: Here we have $m = 100$, $n = 70$, so we are in the large sample case. We may test the hypotheses

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 < \mu_2$$

using the test statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \quad \text{with observed value} \quad T_{obs} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} = -4.8495529$$

Here the Rejection Region for the one-sided alternative is $\{T \leq -z_\alpha\}$, where

$$z_\alpha = z_{0.05} = 1.6449$$

so we **do** reject H_0 in favour of H_a for $\alpha = 0.05$.

We may construct a 95% Confidence Interval for $\mu_1 - \mu_2$ as

$$(\bar{x} - \bar{y}) \pm z_{0.025} \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}} = (-140.4291, -59.5905)$$

where $z_{0.025} = 1.9600$.

Example 5 Eight samples of effluent from a pulp mill were each divided into ten batches. From each quantity, five randomly selected batches were subjected to a treatment process intended to remove toxic substances. Five fish of the same species were placed in each batch, and the mean number surviving in the five treated and untreated portions of each effluent quantity after five days were recorded and are given below.

Sample	1	2	3	4	5	6	7	8
Untreated	5.0	1.0	1.8	1.0	3.6	5.0	2.6	1.0
Treated	5.0	5.0	1.2	4.8	5.0	5.0	4.4	2.0
Difference	0.0	-4.0	0.6	-3.8	-1.4	0.0	-1.8	-1.0

Test to see if the treatment increased the mean number of surviving fish. Test using $\alpha = 0.01$.

```
x<-c(5.0,1.0,1.8,1.0,3.6,5.0,2.6,1.0)
y<-c(5.0,5.0,1.2,4.8,5.0,5.0,4.4,2.0)
d<-x-y
n<-length(d)
dbar<-mean(d)
sigd<-sd(d)
tobs<-dbar/(sigd*sqrt(1/n))
alpha<-0.01
t.al<-qt(1-alpha,n-1)
ci<-dbar+c(-1,1)*qt(1-alpha/2,n-1)*sigd/sqrt(n)
```

Solution: Here we have a paired sample case, with $n = 8$ and so we must assume that the differences are Normally distributed. We have

$$\bar{d} = -1.425 \quad s_d = 1.7219175.$$

We may test the hypotheses concerning $\mu_d = \mu_1 - \mu_2$

$$H_0 : \mu_d = 0$$

$$H_a : \mu_d < 0$$

using the test statistic

$$T = \frac{\bar{D}}{S_d/\sqrt{n}} \quad \text{with observed value} \quad T_{obs} = \frac{\bar{d}}{s_d/\sqrt{n}} = -2.3407095.$$

Here the Rejection Region for the two-sided alternative is $\{T \leq -t_{\alpha,n-1}\}$, where

$$t_{\alpha,n-1} = t_{0.01,7} = 2.9980$$

so we do **not** reject H_0 in favour of H_a for $\alpha = 0.01$.

We may construct a 99% Confidence Interval for μ_d as

$$\bar{d} \pm t_{0.005,7} \frac{s_d}{\sqrt{n}} = (-3.5554, 0.7054)$$

where $t_{0.005,7} = 3.4995$.

Example 6 Researchers wish to compare the average weights of blue crabs in two river basins. Based on the health of the two rivers, it is believed that the crabs in the Basin 2 will be larger, on average, than those in Basin 1. 100 blue crabs are sampled in each basin. The mean weight (in grams) of the Basin 1 crabs is 703.2g with a standard deviation of 301.5g. The mean weight of the Basin 2 crabs is 810.2g with a standard deviation of 397.3g.

Is there any evidence to support the research hypothesis ? Test with $\alpha = 0.05$.

```
m<-100;xbar<-703.2; s1<-301.5
n<-100; ybar<-810.2; s2<-397.3
se.val<-sqrt(s1^2/m+s2^2/n)
tobs<-(xbar-ybar)/se.val
alpha<-0.05
z.al<-qnorm(1-alpha)
ci<-(xbar-ybar)+c(-1,1)*qnorm(1-alpha/2)*se.val
```

Solution: Here we have $m = 100$, $n = 100$, so we are in the large sample case. We may test the hypotheses

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 < \mu_2$$

using the test statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \quad \text{with observed value} \quad T_{obs} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} = -2.1453721$$

Here the Rejection Region for the one-sided alternative is $\{T \leq -z_\alpha\}$, where

$$z_\alpha = z_{0.05} = 1.6449$$

so we **do** reject H_0 in favour of H_a for $\alpha = 0.05$.

We may construct a 95% Confidence Interval for $\mu_1 - \mu_2$ as

$$(\bar{x} - \bar{y}) \pm z_{0.025} \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}} = (-204.7528, -9.2472)$$

where $z_{0.025} = 1.9600$.

Example 7 In a test of the effectiveness of a device that is supposed to increase gasoline mileage in a particular make of US automobiles, 12 cars were run, in random order, over a prescribed course both with and without the device in random order. The mileages per gallon are given below. Test using $\alpha = 0.01$.

Car	Without Device	With Device	Difference
1	21.0	20.6	-0.4
2	30.0	29.9	-0.1
3	29.8	30.7	0.9
4	27.3	26.5	-0.8
5	27.7	26.7	-1.0
6	33.1	32.8	-0.3
7	18.8	21.7	2.9
8	26.2	28.2	2.0
9	28.0	28.9	0.9
10	18.9	19.9	1.0
11	29.3	32.4	3.1
12	21.0	22.0	1.0

Is there evidence that the device is effective? Test using $\alpha = 0.01$.

```
d<-c(-0.4,-0.1,0.9,-0.8,-1.0,-0.3,2.9,2.0,0.9,1.0,3.1,1.0)
n<-length(d)
dbar<-mean(d)
sigd<-sd(d)
tobs<-dbar/(sigd*sqrt(1/n))
alpha<-0.01
t.al<-qt(1-alpha,n-1)
ci<-dbar+c(-1,1)*qt(1-alpha/2,n-1)*sigd/sqrt(n)
```

Solution: Here we have a paired sample case, with $n = 12$ and so we must assume that the differences are Normally distributed. We have

$$\bar{d} = 0.7666667 \quad s_d = 1.3647067.$$

We may test the hypotheses concerning $\mu_d = \mu_1 - \mu_2$

$$H_0 : \mu_d = 0$$

$$H_a : \mu_d > 0$$

using the test statistic

$$T = \frac{\bar{D}}{S_d/\sqrt{n}} \quad \text{with observed value} \quad T_{obs} = \frac{\bar{d}}{s_d/\sqrt{n}} = 1.9460675.$$

Here the Rejection Region for the one-sided alternative is $\{T \geq t_{\alpha,n-1}\}$, where

$$t_{\alpha,n-1} = t_{0.01,11} = 2.7181$$

so we do **not** reject H_0 in favour of H_a for $\alpha = 0.01$.

We may construct a 99% Confidence Interval for μ_d as

$$\bar{d} \pm t_{0.005,11} \frac{s_d}{\sqrt{n}} = (-0.4569, 1.9902)$$

where $t_{0.005,11} = 3.1058$.

Example 8 A feeding test is conducted on a herd of 25 dairy cows to compare two diets, A and B. A sample of 13 cows randomly selected from the herd are fed diet A and the remaining cows are fed with diet B. From observations made over a three-week period, the average daily milk production (in litres) is recorded for each cow:

Diet A: 44 44 56 46 47 38 58 53 49 35 46 30 41
Diet B: 35 47 55 29 40 39 32 41 42 57 51 39

Is there any evidence that Diet A has a higher milk production than Diet B ? Test with $\alpha = 0.05$.

```
x<-c(44,44,56,46,47,38,58,53,49,35,46,30,41)
m<-length(x);xbar<-mean(x); s1<-sd(x)
y<-c(35,47,55,29,40,39,32,41,42,57,51,39)
n<-length(y); ybar<-mean(y); s2<-sd(y)
sp<-sqrt(((m-1)*s1^2+(n-1)*s2^2)/(m+n-2))
tobs<-(xbar-ybar)/(sp*sqrt(1/m+1/n))
alpha<-0.05
t.al<-qt(1-alpha,m+n-2)
ci<-(xbar-ybar)+c(-1,1)*qt(1-alpha/2,m+n-2)*sp*sqrt(1/m+1/n)
```

Solution: Here we have $m = 13$, $n = 12$ and so we must assume that the samples are Normally distributed, and that the population variances are equal. We have

$$\bar{x} = 45.1538462 \quad s_1 = 7.9983973 \quad \bar{y} = 42.25 \quad s_2 = 8.7399293$$

so that $s_p = 8.361252$. We may test the hypotheses

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 > \mu_2$$

using the test statistic

$$T = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \quad \text{with observed value} \quad T_{obs} = \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}} = 0.8675501$$

Here the Rejection Region for the one-sided alternative is $\{T \geq t_{\alpha, m+n-2}\}$, where

$$t_{\alpha, m+n-2} = t_{0.05, 23} = 1.7139$$

so we do **not** reject H_0 in favour of H_a for $\alpha = 0.05$.

We may construct a 95% Confidence Interval for $\mu_1 - \mu_2$ as

$$(\bar{x} - \bar{y}) \pm t_{0.025, 23} s_p \sqrt{\frac{1}{m} + \frac{1}{n}} = (-4.0203, 9.8280)$$

where $t_{0.025, 23} = 2.0687$.

Example 9 A university wants to determine whether a new instructor is successful in their in-class teaching. On one test, 44 out of 60 of Instructor A's students passed on the first try. In comparison, 57 out of 72 of more experienced Instructor B's students passed the test on the first try. Is Instructor A's pass rate worse than Instructor B's? Test using $\alpha = 0.01$.

```
x<-44;m<-60
y<-57;n<-72
p1.hat<-x/m
p2.hat<-y/n
p.hat<-(x+y)/(m+n)
se.p<-sqrt(p.hat*(1-p.hat)*(1/m+1/n))
tobs<-(p1.hat-p2.hat)/se.p
alpha<-0.01
z.al<-qnorm(1-alpha)
ci<-(p1.hat-p2.hat)+c(-1,1)*
  qnorm(1-alpha/2)*sqrt(p1.hat*(1-p1.hat)/m+p2.hat*(1-p2.hat)/n)
```

Solution: Here we have Bernoulli data with $x, n-x, y, n-y \geq 15$, which is large enough to apply the Central Limit Theorem. We have

$$\hat{p}_1 = 0.7333333 \quad \hat{p}_2 = 0.7916667$$

and the pooled estimate

$$\hat{p} = \frac{x+y}{m+n} = \frac{101}{132} = 0.7651515.$$

We may test the hypotheses

$$H_0 : p_1 = p_2$$

$$H_a : p_1 < p_2$$

using the test statistic

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p}) \left(\frac{1}{m} + \frac{1}{n} \right)}}$$

with observed value

$$T_{obs} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p}) \left(\frac{1}{m} + \frac{1}{n} \right)}} = -0.7872345$$

Here the Rejection Region for the one-sided alternative is $\{T \leq -z_\alpha\}$, where

$$z_\alpha = z_{0.01} = 2.3263$$

so we do **not** reject H_0 in favour of H_a for $\alpha = 0.01$.

We may construct a 95% Confidence Interval for $p_1 - p_2$ as

$$(\hat{p}_1 - \hat{p}_2) \pm z_{0.025} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}} = (-0.2502, 0.1336)$$

where $z_{0.025} = 1.9600$.

Example 10 In a test of the reliability of products produced by two machines, machine A produced 25 defective parts in a run of 208, while machine B produced 18 defective parts in a run of 200.

Do these results imply a difference in the reliability of these two machines? Test using $\alpha = 0.1$

```
x<-25;m<-208
y<-18;n<-200
p1.hat<-x/m
p2.hat<-y/n
p.hat<-(x+y)/(m+n)
se.p<-sqrt(p.hat*(1-p.hat)*(1/m+1/n))
tobs<-(p1.hat-p2.hat)/se.p
alpha<-0.1
z.al<-qnorm(1-alpha/2)
ci<-(p1.hat-p2.hat)+c(-1,1)*
  qnorm(1-alpha/2)*sqrt(p1.hat*(1-p1.hat)/m+p2.hat*(1-p2.hat)/n)
```

Solution: Here we have Bernoulli data with $x, n-x, y, n-y \geq 15$, which is large enough to apply the Central Limit Theorem. We have

$$\hat{p}_1 = 0.1201923 \quad \hat{p}_2 = 0.09$$

and the pooled estimate

$$\hat{p} = \frac{x+y}{m+n} = \frac{43}{408} = 0.1053922.$$

We may test the hypotheses

$$H_0 : p_1 = p_2$$

$$H_a : p_1 \neq p_2$$

using the test statistic

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p}) \left(\frac{1}{m} + \frac{1}{n} \right)}}$$

with observed value

$$T_{obs} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p}) \left(\frac{1}{m} + \frac{1}{n} \right)}} = 0.99287$$

Here the Rejection Region for the two-sided alternative is $\{|T| \geq z_{\alpha/2}\}$, where

$$z_{\alpha} = z_{0.05} = 1.6449$$

so we do **not** reject H_0 in favour of H_a for $\alpha = 0.1$.

We may construct a 90% Confidence Interval for $p_1 - p_2$ as

$$(\hat{p}_1 - \hat{p}_2) \pm z_{0.05} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}} = (-0.0196, 0.0800)$$

where $z_{0.05} = 1.6449$.