STAT234: Lecture 8 - Two groups comparison

Kushal K. Dey

Today's plan

- Comparing two distributions

Today's plan

- Comparing two distributions
- Two sample t-test for proportions

Today's plan

- Comparing two distributions
- Two sample t-test for proportions
- Two sample t-test for means

Today's plan

- Comparing two distributions
- Two sample t-test for proportions
- Two sample t-test for means
- Matched pair t-test vs Pooled t- test

Today's plan

- Comparing two distributions
- Two sample t-test for proportions
- Two sample t-test for means
- Matched pair t-test vs Pooled t- test

More importantly !

More importantly !

Compare Midterm and Finals

What are the concerns?

- What is the difficulty level of the final paper?
- Did good in homeworks and the exams. Do you think I can solve all the problems?
- Is it going to be more difficult?

Looking at previous year's data

Table:

Student's name	Midterm Score	Final Score
S	92	52
A	36	49
C	32	40
H	96	75
I	87	86
N	72	74

Formulate the problem

- X - pre-final scores, Y - final scores
- Need to Compare them
- Consider $X_1, X_2, \ldots, X_n \sim N(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_n \sim N(\mu_2, \sigma_2^2)$
- We need to test whether μ_1 and μ_2 are same.

Formulate the problem

- X - pre-final scores, Y - final scores
- Need to Compare them
- Consider $X_1, X_2, \ldots, X_n \sim N(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_n \sim N(\mu_2, \sigma_2^2)$
- We need to test whether μ_1 and μ_2 are same.

What is the null hypothesis? Alternate? What are the assumptions?

Solve the problem

Test $H_0 : \mu_1 - \mu_2 = 0$ against $H_1 : \mu_1 - \mu_2 > (\neq) 0$

- What are the assumptions?
- What should the test statistics be?
- What is the distribution of the test statistics?

Assumptions

- $X_1, X_2, \ldots, X_n \sim N(\mu_1, \sigma_1^2)$
- $Y_1, Y_2, \ldots, Y_n \sim N(\mu_2, \sigma_2^2)$
- X_i, Y_i are not independent. Why?

Assumptions

- $X_1, X_2, \ldots, X_n \sim N(\mu_1, \sigma_1^2)$
- $Y_1, Y_2, \ldots, Y_n \sim N(\mu_2, \sigma_2^2)$
- X_i, Y_i are not independent. Why? Let the correlation be ρ (population parameter, same for all pairs)
- So, $cov(X_i, Y_i) = \rho\sigma_1\sigma_2$

Statistical solution

Test $H_0 : \mu_1 - \mu_2 = 0$ against $H_1 : \mu_1 - \mu_2 > (\neq)0$

- $E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$
- $var(\bar{X} - \bar{Y}) = \frac{1}{n}(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)$
- What is the distribution of $\bar{X} - \bar{Y}$?
- Estimate of the variance?

Statistical solution

- Define $W_i = X_i - Y_i$
- Then, we get

$$\begin{aligned} E(W_i) &= \mu_1 - \mu_2 = \mu \qquad \text{and} && (1) \\ var(W_i) &= \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2 = \sigma^2 \qquad \text{(say)} && (2) \end{aligned}$$

- Also note that $\bar{W} = \bar{X} - \bar{Y}$

The reformulated problem

$W_1, W_2, \ldots, W_n$ are i.i.d. random variables with mean μ and unknown variance σ^2. Test

$$H_0 : \mu_W = 0 \qquad \text{against} \qquad H_1 : \mu_W > (\neq) 0$$

The reformulated problem

$W_1, W_2, \ldots, W_n$ are i.i.d. random variables with mean μ and unknown variance σ^2. Test

$$H_0 : \mu_W = 0 \qquad \text{against} \qquad H_1 : \mu_W > (\neq) 0$$

You know it!

The reformulated problem

$W_1, W_2, \ldots, W_n$ are i.i.d. random variables with mean μ and unknown variance σ^2. Test

$$H_0 : \mu_W = 0 \qquad \text{against} \qquad H_1 : \mu_W > (\neq) 0$$

You know it!

That is simply a one sample t-test. $H_0 : \mu = 0$ against proper alternative. It is called Matched pair t-test.

Since $W_i = X_i - Y_i$, and X_i and Y_i are normal random variables, hence W_i also normal.

Since $W_i = X_i - Y_i$, and X_i and Y_i are normal random variables, hence W_i also normal.

$$W_i \sim N(\mu_W, \sigma_W^2)$$

Since $W_i = X_i - Y_i$, and X_i and Y_i are normal random variables, hence W_i also normal.

$$W_i \sim N(\mu_W, \sigma_W^2)$$

$$\mu_W = \mu_1 - \mu_2$$

Since $W_i = X_i - Y_i$, and X_i and Y_i are normal random variables, hence W_i also normal.

$$W_i \sim N(\mu_W, \sigma_W^2)$$

$$\mu_W = \mu_1 - \mu_2$$

$$\sigma_W^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2 = \sigma^2$$

Since $W_i = X_i - Y_i$, and X_i and Y_i are normal random variables, hence W_i also normal.

$$W_i \sim N(\mu_W, \sigma_W^2)$$

$$\mu_W = \mu_1 - \mu_2$$

$$\sigma_W^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2 = \sigma^2$$

Under H_0, we get

$$W_1, W_2, \cdots, W_n \sim N(0, \sigma_W^2)$$

Under H_0,

$$\bar{W} \sim N\left(0, \frac{\sigma_W^2}{n}\right)$$

$$\frac{\sqrt{n}\bar{W}}{\sigma_W} \sim N(0,1)$$

We do not know σ_W, so we replace it by s_W,

$$s_W^2 := \frac{1}{n-1}\sum_{i=1}^{n}\left(W_i - \bar{W}\right)^2$$

What is the distribution of

$$\frac{\sqrt{n}\bar{W}}{s_W}$$

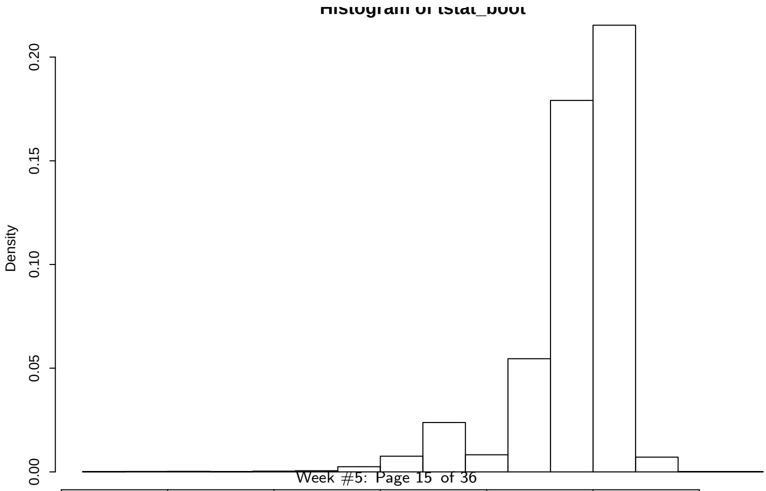
$$\frac{\sqrt{n}\bar{W}}{s_W} \sim t_{n-1}$$

Compute the observed value $\bar{w}$ pf $\bar{W}$ and s_w- realization of s_W. From the data we presented,

```
set.seed(100)
x <- c(92, 36, 32, 96, 87, 72)
y <- c(32, 49, 40, 75, 86, 74)
W <- x-y
Wbar <- mean(W); sW <- sd(W)
tstat = sqrt(length(W))*Wbar/sW
```

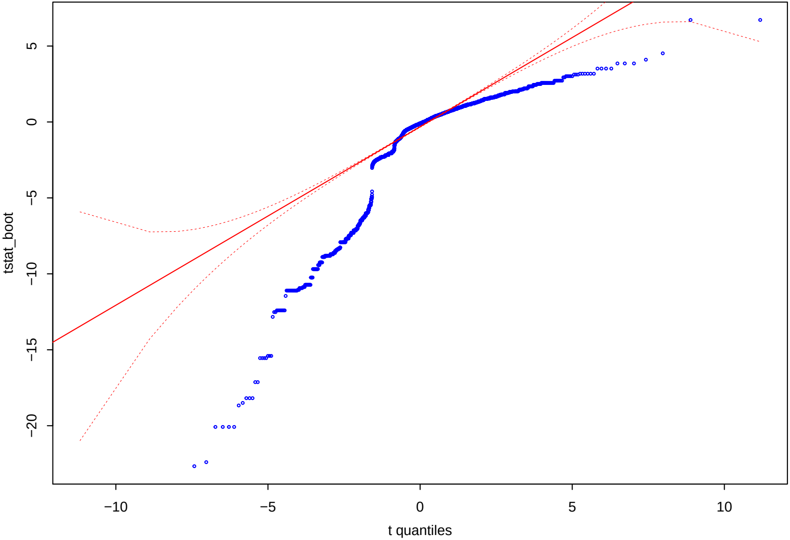
Resampling distribution

```
W_new <- W - Wbar + 0;
bootstapW <- replicate(10000, sample(W_new, replace=TRUE))
tstat_boot <- sqrt(length(W))*
              (apply(bootstapW, 2, mean)/apply(bootstapW,2, sd))
hist(tstat_boot, freq=FALSE)
```



Resampling distribution

```
library(gap)
qqfun(tstat_boot, "t", df=length(W)-1)
```



p-value

```
tstat = sqrt(length(W))*Wbar/sW
pval_t = 1 - pt(tstat,df=length(W)-1)

pval_boot <- length(which(tstat_boot > tstat))/length(tstat_boot

pval_t

[1] 0.2082

pval_boot

[1] 0.1397
```

A more complicated scenario

Suppose you sampled 30 students' scores for the midterm exam and for the 2012 exam, another person sampled 26 students' scores for the end term.

A more complicated scenario

Suppose you sampled 30 students' scores for the midterm exam and for the 2012 exam, another person sampled 26 students' scores for the end term.

You do not know if the 26 students are a different set from 30 students or not. How do you go about analyzing the data then?

A more complicated scenario

Suppose you sampled 30 students' scores for the midterm exam and for the 2012 exam, another person sampled 26 students' scores for the end term.

You do not know if the 26 students are a different set from 30 students or not. How do you go about analyzing the data then?

Then we assume if X_i denote midterm score and Y_i the final score,

- $X_1, X_2, \ldots, X_{30} \sim N(\mu_1, \sigma_1^2)$
- $Y_1, Y_2, \ldots, Y_{26} \sim N(\mu_2, \sigma_2^2)$

W assume that X_i's and Y_i's are independent. So, $\rho = cor(X, Y) = 0$.

Statistical solution

Test $H_0 : \mu_1 - \mu_2 = 0$ against $H_1 : \mu_1 - \mu_2 > (\neq) 0$

Statistical solution

Test $H_0 : \mu_1 - \mu_2 = 0$ against $H_1 : \mu_1 - \mu_2 > (\neq) 0$

- So, our test statistics should be $\bar{X} - \bar{Y}$

Statistical solution

Test $H_0 : \mu_1 - \mu_2 = 0$ against $H_1 : \mu_1 - \mu_2 > (\neq) 0$

- So, our test statistics should be $\bar{X} - \bar{Y}$
- $E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$

Statistical solution

Test $H_0 : \mu_1 - \mu_2 = 0$ against $H_1 : \mu_1 - \mu_2 > (\neq)0$

- So, our test statistics should be $\bar{X} - \bar{Y}$
- $E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$
- $var(\bar{X} - \bar{Y}) = (\sigma_1^2/30 + \sigma_2^2/26)$

Statistical solution

Test $H_0 : \mu_1 - \mu_2 = 0$ against $H_1 : \mu_1 - \mu_2 > (\neq)0$

- So, our test statistics should be $\bar{X} - \bar{Y}$
- $E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$
- $var(\bar{X} - \bar{Y}) = (\sigma_1^2/30 + \sigma_2^2/26)$
- Since we assume independence, the distribution of $\bar{X} - \bar{Y}$ is $N(\mu_1 - \mu_2, \sigma_1^2/30 + \sigma_2^2/26)$

Statistical solution

Test $H_0 : \mu_1 - \mu_2 = 0$ against $H_1 : \mu_1 - \mu_2 > (\neq)0$

- So, our test statistics should be $\bar{X} - \bar{Y}$
- $E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$
- $var(\bar{X} - \bar{Y}) = (\sigma_1^2/30 + \sigma_2^2/26)$
- Since we assume independence, the distribution of $\bar{X} - \bar{Y}$ is $N(\mu_1 - \mu_2, \sigma_1^2/30 + \sigma_2^2/26)$
- Estimate of the variance: $(s_1^2/30 + s_2^2/26)$

Statistical solution

Test $H_0 : \mu_1 - \mu_2 = 0$ against $H_1 : \mu_1 - \mu_2 > (\neq)0$

- So, our test statistics should be $\bar{X} - \bar{Y}$
- $E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$
- $var(\bar{X} - \bar{Y}) = (\sigma_1^2/30 + \sigma_2^2/26)$
- Since we assume independence, the distribution of $\bar{X} - \bar{Y}$ is $N(\mu_1 - \mu_2, \sigma_1^2/30 + \sigma_2^2/26)$
- Estimate of the variance: $(s_1^2/30 + s_2^2/26)$
- What is the distribution of the above estimate?

More complicated scenario

Consider the following set-up.

- $X_1, X_2, \ldots, X_{30} \sim N(\mu_1, \sigma_1^2)$
- $Y_1, Y_2, \ldots, Y_{26} \sim N(\mu_2, \sigma_2^2)$

More complicated scenario

Consider the following set-up.

- $X_1, X_2, \ldots, X_{30} \sim N(\mu_1, \sigma_1^2)$
- $Y_1, Y_2, \ldots, Y_{26} \sim N(\mu_2, \sigma_2^2)$
- But we assume that X_i's and Y_i's are independent. Why?

More complicated scenario

Consider the following set-up.

- $X_1, X_2, \ldots, X_{30} \sim N(\mu_1, \sigma_1^2)$
- $Y_1, Y_2, \ldots, Y_{26} \sim N(\mu_2, \sigma_2^2)$
- But we assume that X_i's and Y_i's are independent. Why?So, $\rho = 0$.

How to deal with this? - We will consider two different scenarios.

More complicated scenario

Consider the following set-up.

- $X_1, X_2, \ldots, X_{30} \sim N(\mu_1, \sigma_1^2)$
- $Y_1, Y_2, \ldots, Y_{26} \sim N(\mu_2, \sigma_2^2)$
- But we assume that X_i's and Y_i's are independent. Why?So, $\rho = 0$.

How to deal with this? - We will consider two different scenarios.

- two variances are equal
- variances are not equal

Set-up

We have $X_1, X_2, \ldots, X_n \sim N(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_m \sim N(\mu_2, \sigma_2^2)$

Set-up

We have $X_1, X_2, \ldots, X_n \sim N(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_m \sim N(\mu_2, \sigma_2^2)$

- Test $H_0 : \mu_1 - \mu_2 = 0$ against $H_0 : \mu_1 - \mu_2 \neq 0$

Set-up

We have $X_1, X_2, \ldots, X_n \sim N(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_m \sim N(\mu_2, \sigma_2^2)$

- Test $H_0 : \mu_1 - \mu_2 = 0$ against $H_0 : \mu_1 - \mu_2 \neq 0$
- σ_1, σ_2 are unknown

Set-up

We have $X_1, X_2, \ldots, X_n \sim N(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_m \sim N(\mu_2, \sigma_2^2)$

- Test $H_0 : \mu_1 - \mu_2 = 0$ against $H_0 : \mu_1 - \mu_2 \neq 0$
- σ_1, σ_2 are unknown
- You know that an estimate of $\mu_1 - \mu_2$ is $\bar{X} - \bar{Y}$. So, we should try to find out the distribution of this.

Set-up

We have $X_1, X_2, \ldots, X_n \sim N(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_m \sim N(\mu_2, \sigma_2^2)$

- Test $H_0 : \mu_1 - \mu_2 = 0$ against $H_0 : \mu_1 - \mu_2 \neq 0$
- σ_1, σ_2 are unknown
- You know that an estimate of $\mu_1 - \mu_2$ is $\bar{X} - \bar{Y}$. So, we should try to find out the distribution of this.
- The distribution is

$$\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \sigma_1^2/n + \sigma_2^2/m)$$

i.e.

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \sim N(0, 1)$$

Two sample test with equal variance

Suppose, we have reasons to believe that the variances are equal i.e. $\sigma_1 = \sigma_2 = \sigma$. Then, we have

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n + 1/m}} \sim N(0,1)$$

Two sample test with equal variance

Suppose, we have reasons to believe that the variances are equal i.e. $\sigma_1 = \sigma_2 = \sigma$. Then, we have

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n + 1/m}} \sim N(0, 1)$$

- Clearly, we have to estimate σ^2.

Two sample test with equal variance

Suppose, we have reasons to believe that the variances are equal i.e. $\sigma_1 = \sigma_2 = \sigma$. Then, we have

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n + 1/m}} \sim N(0, 1)$$

- Clearly, we have to estimate σ^2.
- Recall that both s_X^2 and s_Y^2 are unbiased estimates of σ^2. We should combine them.

Two sample test with equal variance

Suppose, we have reasons to believe that the variances are equal i.e. $\sigma_1 = \sigma_2 = \sigma$. Then, we have

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n + 1/m}} \sim N(0,1)$$

- Clearly, we have to estimate σ^2.
- Recall that both s_X^2 and s_Y^2 are unbiased estimates of σ^2. We should combine them.

$$\hat{\sigma^2} := \frac{(n-1)s_X^2 + (m-1)s_Y^2}{m+n-2}$$

Two sample test with equal variance (contd)

Test $H_0 : \mu_1 - \mu_2 = 0$ against $H_0 : \mu_1 - \mu_2 \neq 0$ and we have

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n + 1/m}} \sim N(0, 1)$$

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\hat{\sigma}\sqrt{1/n + 1/m}} \sim t_{n+m-2}$$

Two sample test with equal variance (contd)

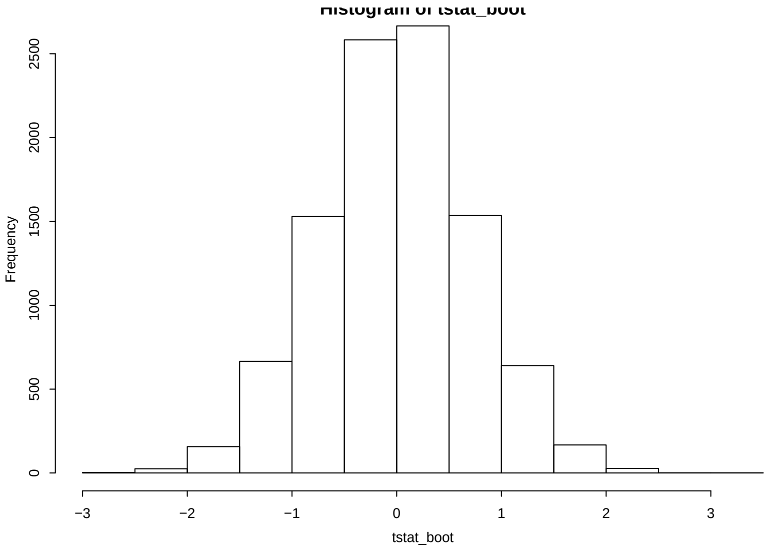
Test $H_0 : \mu_1 - \mu_2 = 0$ against $H_0 : \mu_1 - \mu_2 \neq 0$ and we have

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n + 1/m}} \sim N(0,1)$$

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\hat{\sigma}\sqrt{1/n + 1/m}} \sim t_{n+m-2}$$

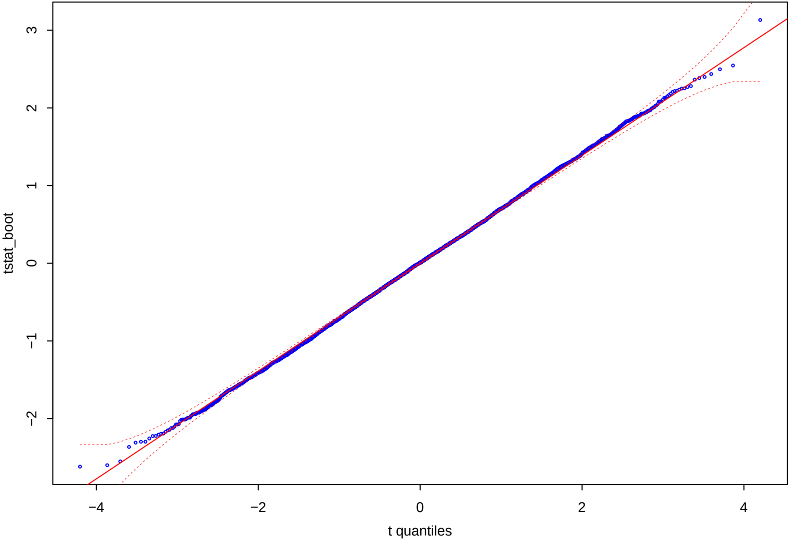
```
x <- rnorm(30, 67, 10);
y <- rnorm(26, 56, 10);

meanW <- mean(x) - mean(y)
sX2 <- var(x); sY2 <- var(y);
s2_pool <- ((length(x)-1)*sX2
            + (length(y)-1)*sY2)/(length(x)+length(y)-2);

tstat <- meanW/ (sqrt(s2_pool)*(sqrt(1/(length(x)) + 1/(length(y
```

```
xboot <- replicate(10000, sample(c(x,y), length(x), replace = FA
yboot <- replicate(10000, sample(c(x,y), length(y), replace = FA

mean_wboot <- colMeans(xboot)  - colMeans(yboot)
sx_boot2 <- apply(xboot, 2, var)
sy_boot2 <- apply(yboot, 2, var)
s_boot2 <- ((length(x)-1)*sx_boot2 + (length(y)-1)*sy_boot2)/(le

tstat_boot <- mean_wboot/(sqrt(s_boot2)*(sqrt(1/(length(x)) + 1/
```


Histogram of tstat_boot
Frequency
0
500
1000
1500
2000
2500
−3
−2
−1
0
1
2
3
tstat_boot

qqplot with t(54)

```
library(gap)
qqfun(tstat_boot, "t", df=length(x)+length(y)-2)
```



p-value

```
tstat <- meanW/ (sqrt(s2_pool)*(sqrt(1/(length(x)) + 1/(length(y
pval_t = 1 - pt(tstat,df=length(x)+length(y)-1)

pval_boot <- length(which(tstat_boot > tstat))/length(tstat_boot

pval_t

[1] 0.0002026

pval_boot

[1] 0
```

Two sample test with equal variance (contd)

Test $H_0 : \mu_1 - \mu_2 = 0$ against $H_0 : \mu_1 - \mu_2 \neq 0$ and our test statistic is

$$
\begin{aligned}
T &= \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n + 1/m}} \cdot \sqrt{\frac{\sigma^2(n+m-2)}{(n-1)s_X^2 + (m-1)s_Y^2}} \\
&= \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_{XY}\sqrt{1/n + 1/m}} \text{ where } S_{XY}^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}
\end{aligned}
$$

Two sample test with equal variance (contd)

Test $H_0 : \mu_1 - \mu_2 = 0$ against $H_0 : \mu_1 - \mu_2 \neq 0$ and our test statistic is

$$\begin{aligned} T &= \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n + 1/m}} \cdot \sqrt{\frac{\sigma^2(n+m-2)}{(n-1)s_X^2 + (m-1)s_Y^2}} \\ &= \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_{XY}\sqrt{1/n + 1/m}} \text{ where } S_{XY}^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2} \end{aligned}$$

- $S_{XY}^2 =$ Pooled variance
- For this problem, p-value will be $2P(T \geq |t|)$ where t is the observed value of the statistic and $T \sim t_{n+m-2}$

2- sample test (unequal variance) - Behren-Fisher problem

Suppose, we have reasons to believe that the variances are not equal i.e. $\sigma_1 \neq \sigma_2$. Then, we have

$$\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \sigma_1^2/n + \sigma_2^2/m)$$

- So, the two population variances must be estimated separately.
- Recall that s_X^2 and s_Y^2 are unbiased estimates of σ_1^2, σ_2^2. We should combine them.
- The estimate is

$$S^2_{\bar{X}-\bar{Y}} = \frac{s_X^2}{n} + \frac{s_Y^2}{m}$$

- The t statistic to test our hypothesis is:

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_{\bar{X}-\bar{Y}}}$$

Welch's t test

Test $H_0 : \mu_1 - \mu_2 = 0$ against $H_0 : \mu_1 - \mu_2 \neq 0$ when variances are not equal and we have the t statistic as

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_{\bar{X}-\bar{Y}}}$$

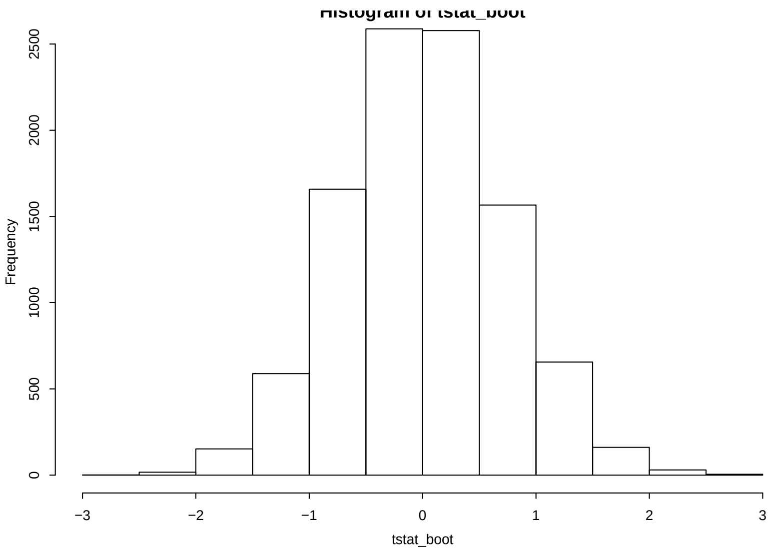
This is approximately t distributed with

$$df = \frac{(s_X^2/n + s_Y^2/m)^2}{(s_X^2/n)^2/(n-1) + (s_Y^2/m)^2/(m-1)}$$

This is known as Welch-Satterthwaite equation.

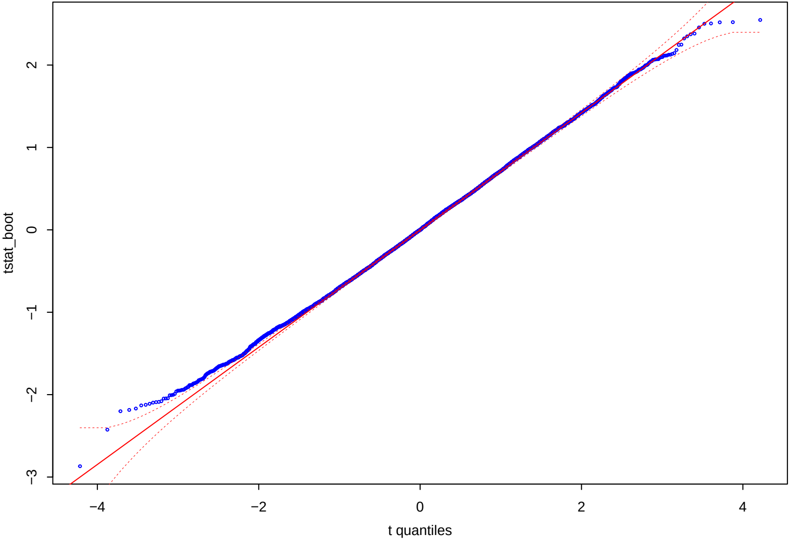
```
x <- rnorm(30, 67, 30);
y <- rnorm(26, 26, 26);

meanW <- mean(x) - mean(y)
sX2 <- var(x); sY2 <- var(y);
s2_pool <- sX2/(length(x)) + sY2/(length(y));

tstat <- meanW/ (sqrt(s2_pool));

df <- (sX2/length(x) + sY2/length(y))^2/
  (sX2^2/(length(x)^2*(length(x)-1)) + sY2^2/(length(y)^2*(lengt
```

```
xboot <- replicate(10000, sample(c(x,y), length(x),
                                 replace = FALSE));
yboot <- replicate(10000, sample(c(x,y), length(y),
                                 replace = FALSE));

mean_wboot <- colMeans(xboot)  - colMeans(yboot)
sx_boot2 <- apply(xboot, 2, var)
sy_boot2 <- apply(yboot, 2, var)
s_boot2 <- sx_boot2/(length(x)) + sy_boot2/(length(y));

tstat_boot <- mean_wboot/(sqrt(s_boot2))
```


Histogram of tstat_boot
Frequency
0
500
1000
1500
2000
2500
−3
−2
−1
0
1
2
3
tstat_boot

qqplot with t(51.9)

```
library(gap)
qqfun(tstat_boot, "t", df=51.9)
```



How to know when to assume $\sigma_1 = \sigma_2 = \sigma$.

How to know when to assume $\sigma_1 = \sigma_2 = \sigma$.

The rule is - is the ratio of sample variances less than 2.

How to know when to assume $\sigma_1 = \sigma_2 = \sigma$.

The rule is - is the ratio of sample variances less than 2.

if $\frac{largest s_X^2, s_Y^2}{smallest s_X^2, s_Y^2} < 2$, then assume $\sigma_X = \sigma_Y = \sigma$.

How to know when to assume $\sigma_1 = \sigma_2 = \sigma$.

The rule is - is the ratio of sample variances less than 2.

if $\frac{largest s_X^2, s_Y^2}{smallest s_X^2, s_Y^2} < 2$, then assume $\sigma_X = \sigma_Y = \sigma$.

When we hold this assumption to be true, we do the t-test for same σ, else the Welch method t-test.