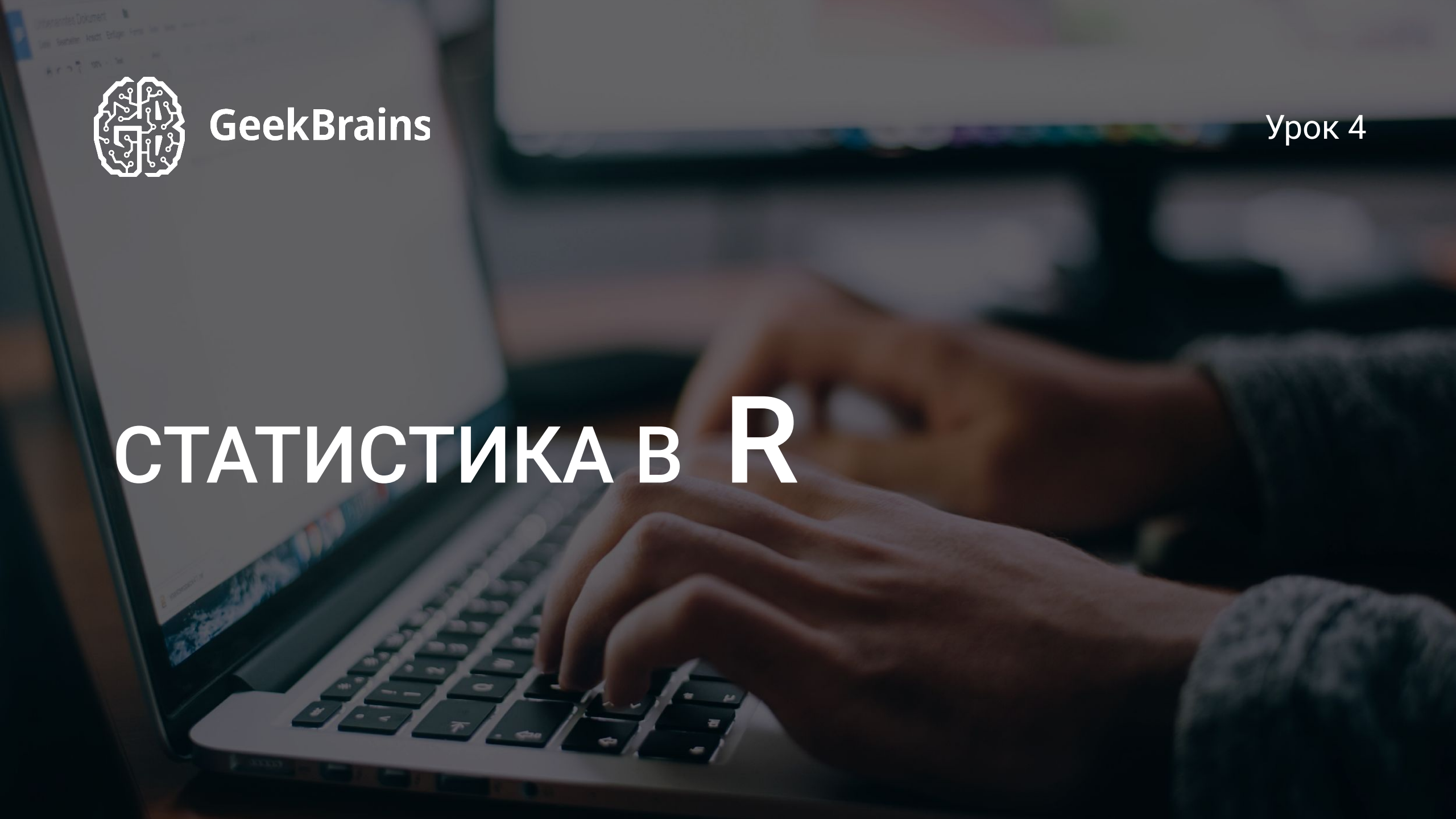
R

Вебинары



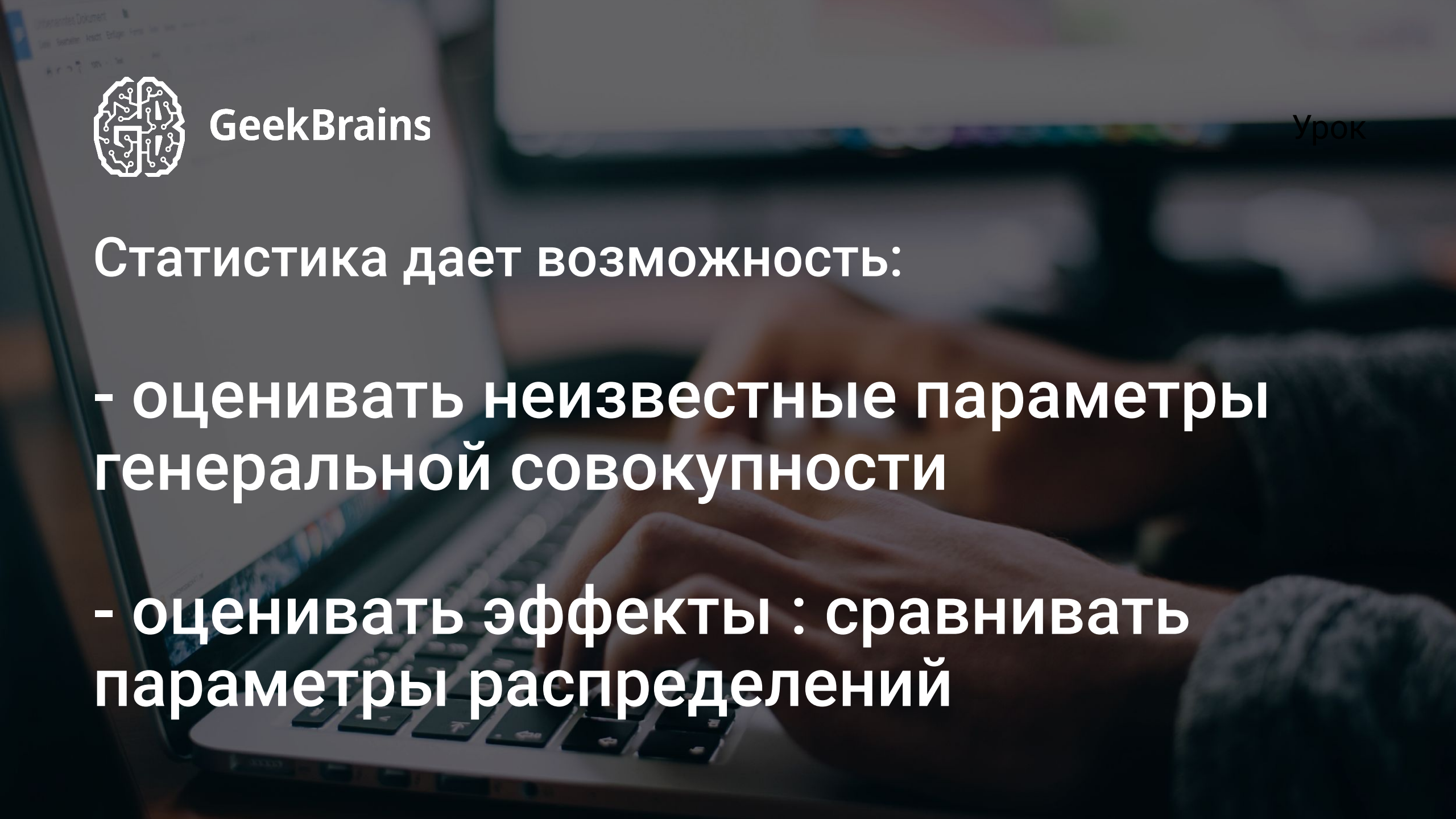
GeekBrains

Урок 4

СТАТИСТИКА В R

На этом уроке мы изучим:

1. Построение доверительных интервалов
2. Мощность теста
3. Тестирование гипотезы
3.1 Z, t критерии в R
3.2 Одновыборочный тест
3.3 Двухвыборочный тест с независимыми выборками
3.4 Двухвыборочный тест с зависимыми выборками
3.5 Интерпретация результатов



GeekBrains

Урок

Статистика дает возможность:

- оценивать неизвестные параметры генеральной совокупности

- оценивать эффекты : сравнивать параметры распределений

Для оценки параметров пользуются доверительным интервалом

Для оценки эффекта используют тестирование гипотезы

Доверительный интервал

Z – критерий для построения доверительного интервала

1. Критерий Z является более предпочтительным в том случае, если известно стандартное отклонение генеральной совокупности
2. Имеется достаточно большой объем выборки
3. Данные хорошо приближены к нормальному распределению

Задача : С помощью 95% доверительного интервала оценить среднее арифметическое нормально распределенной генеральной совокупности, зная что ее стандартное отклонение равно 3 , а объем выборки равен 50

Известно:

Sd= 3

N = 50

α =5%

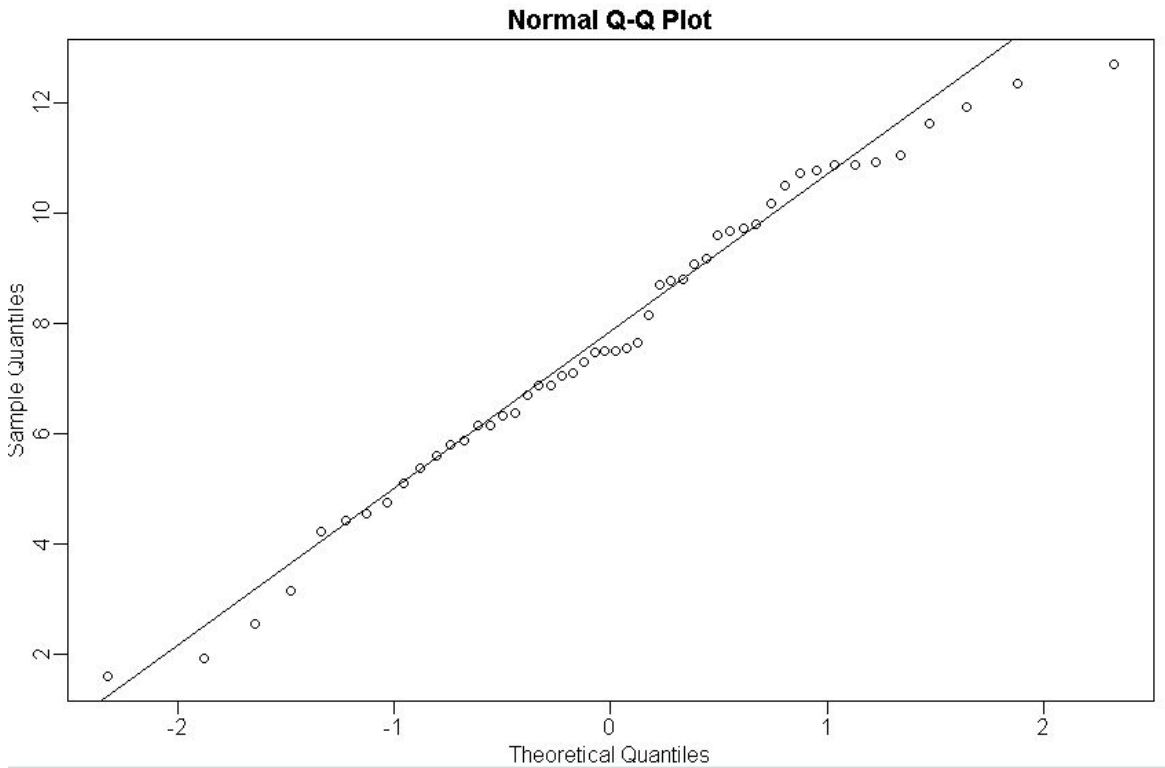
```
set.seed(4)
samp<-rnorm(50,7,3)
samp
```

```
> set.seed(4)
> samp<-rnorm(50,7,3)
> samp
 [1]  7.650265  5.372522  9.673434  8.787942 11.906854  9.067826  3.156260  6.360566 12.689620 12.330590  8.699813
[12]  7.047158  8.149172  6.864589  7.103056  7.507080 10.495081  6.867388  6.698895  6.149666 11.622445  7.495507
[23] 10.922867 10.864771  8.778691  6.151169 10.767652  9.729517  4.215916 10.720543  7.460393 10.155798  4.737366
[34]  2.553433  9.583396  5.786441  6.317784  9.802289  5.602312  5.087370 11.031126  7.544606 10.877537  1.935854
[45]  4.537019  4.413562  7.296531  5.873035  9.171712  1.607854
```

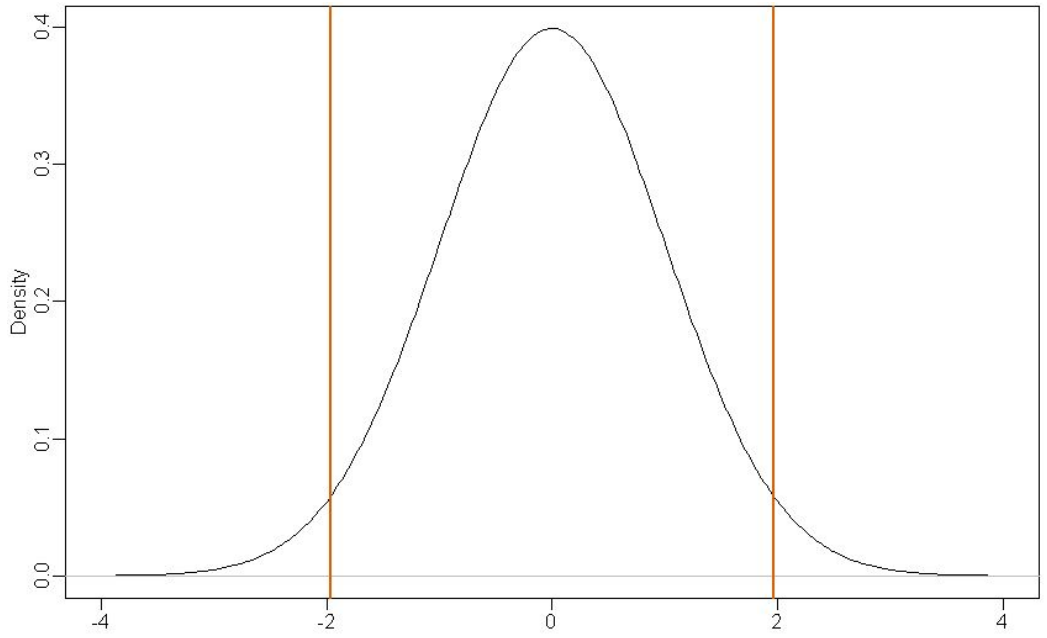
1) Определить статистик для оценки интервала.

ЦПТ хорошо работает от 30 и выше, известно стандартное отклонение генеральной совокупности и данные хорошо приближены к нормальному распределению, измерения независимы, поэтому используем критерий Z

```
qqnorm(samp)
qqline(samp)
```



```
> Z<-qnorm(0.975)  # 95% CI
> Z
[1] 1.959964
> SE<-3/sqrt(50)
> SE
[1] 0.4242641
> lolv<- mean(samp)- Z*SE
> uplv<-mean(samp)+ Z*SE
> CI<- c(lolv,uplv)
> CI
[1] 6.872903 8.535988
```



Генерируя выборку, мы представляли, будто не знаем среднее арифметическое.
Теперь мы определили доверительный интервал и можем себя проверить

```
> 7>= CI[1] & 7 <= CI[2]
[1] TRUE
```

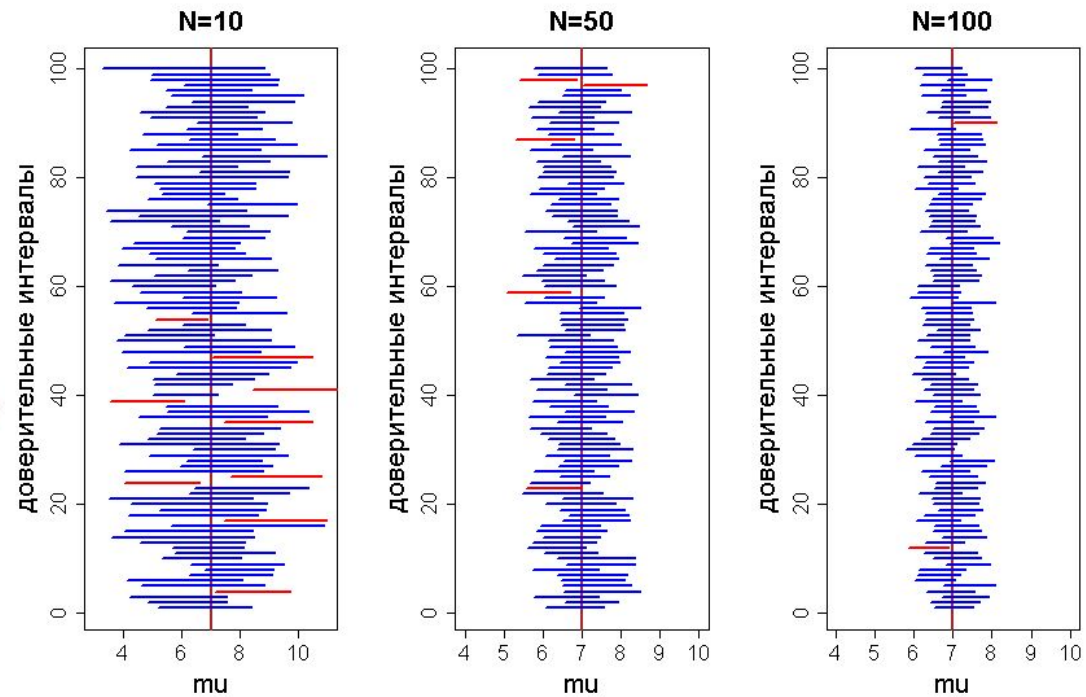
t- критерий для построения доверительного интервала

1. Неизвестно стандартное отклонение генеральной совокупности
2. Выборка небольшого объема
3. Соблюдается условие нормальности и независимости наблюдений для первых двух пунктов

Идея 95 % интервала состоит в том, что 95 % интервалов должны захватывать истинное среднее арифметическое генеральной совокупности

Сравним как работает ЦПТ на выборках разных объемов

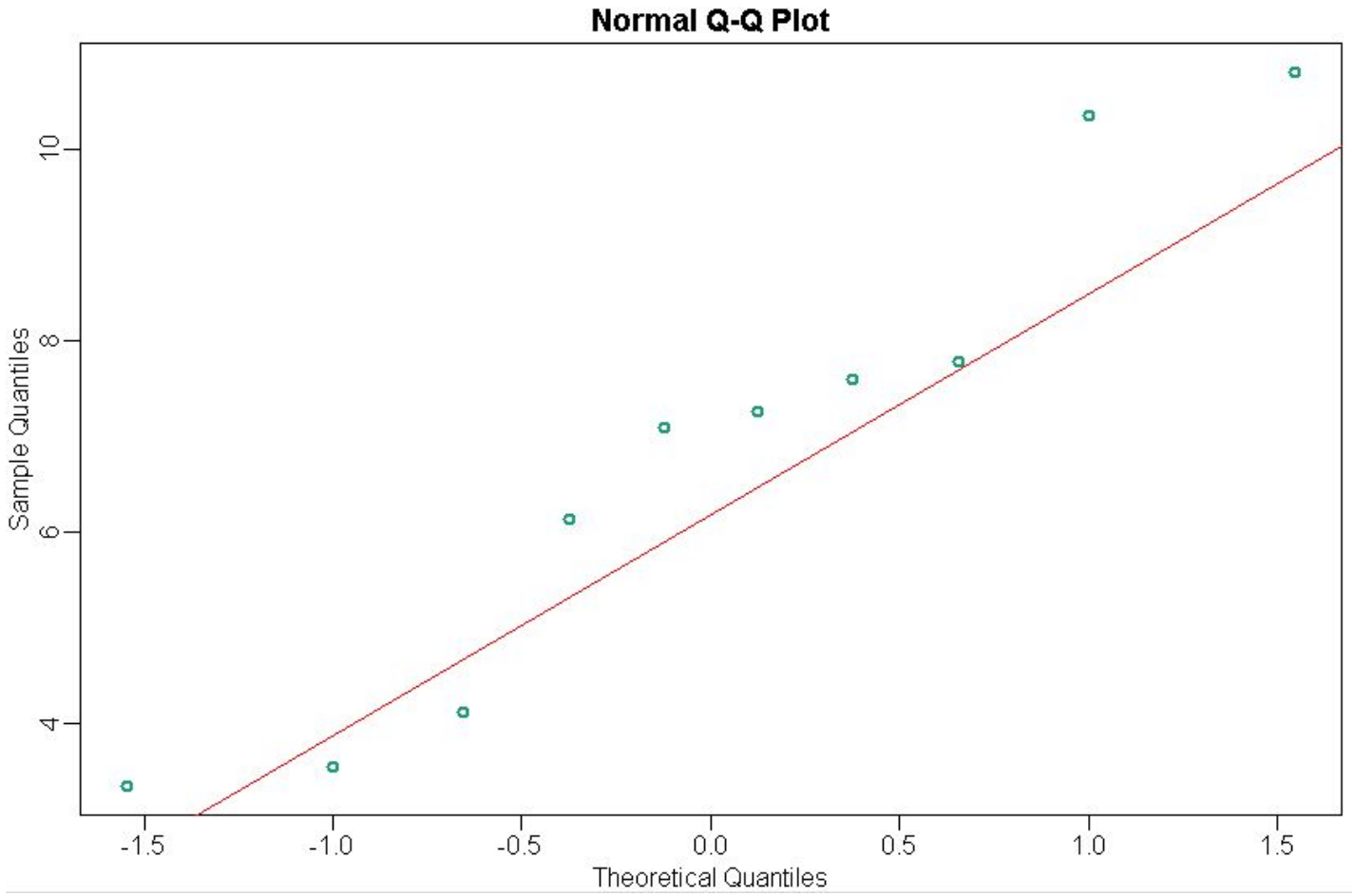
```r
bigpar(1,3)
set.seed(3)
plot(7 +c(-4,4),c(1,1),type="n",
     xlab="mu",ylab="доверительные интервалы", ylim = c(1,100), main= "N=10")  # for N=10
abline(v=7, col= "brown", lwd=2)
for (i in 1:100) {
  sam <- rnorm(10,7,3)
  SE <- sd(sam)/sqrt(10)
  CI <- c(mean(sam)-Z*SE, mean(sam)+Z*SE)
  catch <-
    7>=CI[1] &7<=CI[2]
  7 >= CI[1] & 7 <= CI[2]
  color <- ifelse(catch,"blue","red")
  lines(CI, c(i,i),col=color, lwd=2)
}

set.seed(3)

plot(7 +c(-3,3),c(1,1),type="n",
     xlab="mu",ylab="доверительные интервалы", ylim = c(1,100), main =  "N=50") #
abline(v=7, col= "brown", lwd=2)
for (i in 1:100) {
  sam <- rnorm(50,7,3)
  SE <- sd(sam)/sqrt(50)
  CI <- c(mean(sam)-Z*SE, mean(sam)+Z*SE)
  catch <-
    7>=CI[1] &7<=CI[2]
    7 >= CI[1] & 7 <= CI[2]
  color <- ifelse(catch,"blue","red")
  lines(CI, c(i,i),col=color, lwd=2)
}
set.seed(3)

plot(7 +c(-3,3),c(1,1),type="n",
     xlab="mu",ylab="доверительные интервалы", ylim = c(1,100), main = "N=100") # for N=100
abline(v=7, col= "brown", lwd=2)
for (i in 1:100) {
  sam <- rnorm(100,7,3)
  SE <- sd(sam)/sqrt(100)
  CI <- c(mean(sam)-Z*SE, mean(sam)+Z*SE)
  catch <-
    7>=CI[1] &7<=CI[2]
  7 >= CI[1] & 7 <= CI[2]
  color <- ifelse(catch,"blue","red")
  lines(CI, c(i,i),col=color, lwd=2)
}
```



Оставим теперь тот же set.seed(3) , но теперь для выборки объемом 10 будем использовать распределение Стьюдента t

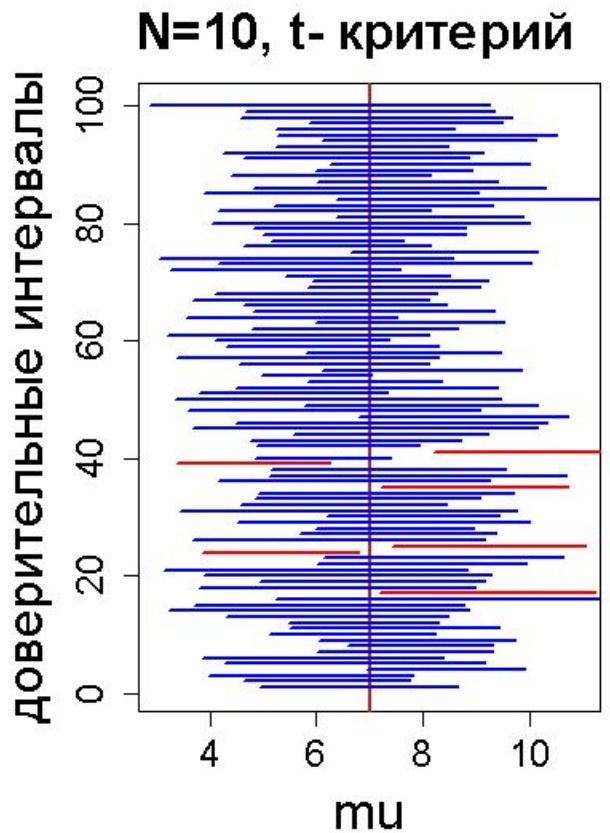
Если объем выборки меньше 15, требуется ,
чтобы данные были приближены к нормальному распределению. Проверим это требование с помощью qq- графика

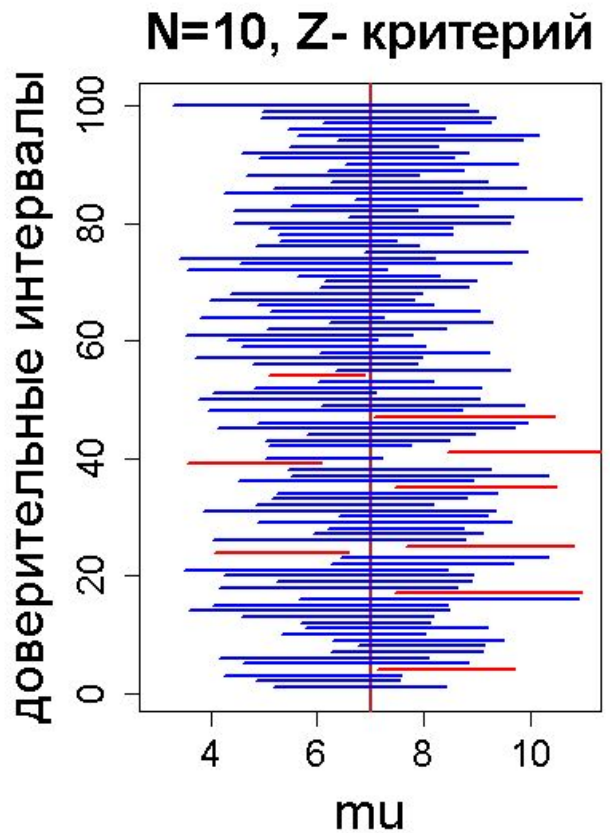
```
 97  set.seed(3)
 98
 99  sam <- rnorm(10,7,3)
100  mypar(1,1)
101  qqnorm(sam, col=1, lwd=2)
102  qqline(sam, col="red")
```



Сравним интервалы для t и z критериев при небольших выборках

```r
bigpar(1,2)
t<- qt(0.975,9)
set.seed(3)
plot(7 +c(-4,4),c(1,1),type="n",
     xlab="mu",ylab="доверительные интервалы",
     ylim = c(1,100), main = "N=10, t- критерий")  # for N=10, t
abline(v=7, col= "brown", lwd=2)
for (i in 1:100) {
  sam <- rnorm(10,7,3)
  SE <- sd(sam)/sqrt(10)
  CI <- c(mean(sam)-t*SE, mean(sam)+t*SE)
  catch <-
    7>=CI[1] &7<=CI[2]
  7 >= CI[1] & 7 <= CI[2]
  color <- ifelse(catch,"blue","red")
  lines(CI, c(i,i),col=color, lwd=2)
}

set.seed(3)
plot(7 +c(-4,4),c(1,1),type="n",
     xlab="mu",ylab="доверительные интервалы",
     ylim = c(1,100), main= "N=10, Z- критерий")  # for N=10, Z
abline(v=7, col= "brown", lwd=2)
for (i in 1:100) {
  sam <- rnorm(10,7,3)
  SE <- sd(sam)/sqrt(10)
  CI <- c(mean(sam)-Z*SE, mean(sam)+Z*SE)
  catch <-
    7>=CI[1] &7<=CI[2]
  7 >= CI[1] & 7 <= CI[2]
  color <- ifelse(catch,"blue","red")
  lines(CI, c(i,i),col=color, lwd=2)
}
```






```
> Z
[1] 1.959964
> t
[1] 2.262157
> c(mean(sam)-Z*SE, mean(sam)+Z*SE)
[1] 4.037032 9.560154
> c(mean(sam)-t*SE, mean(sam)+t*SE)
[1] 3.611246 9.985940
```

Вернемся к реальному набору данных Cardiovascular Disease

```
> tidy_set <-dat %>% filter((ap_lo<200&ap_lo>20)&(ap_hi<300&ap_hi>40))
> head(tidy_set)
  id   age gender height weight ap_hi ap_lo cholesterol gluc smoke alco active cardio age_years
1  0 18393      2    168     62   110    80           1    1     0    0      1      0         50
2  1 20228      1    156     85   140    90           3    1     0    0      1      1         55
3  2 18857      1    165     64   130    70           3    1     0    0      0      1         51
4  3 17623      2    169     82   150   100           1    1     0    0      1      1         48
5  4 17474      1    156     56   100    60           1    1     0    0      0      0         47
6  8 21914      1    151     67   120    80           2    2     0    0      0      0         60
>
```


*Cardiovascular Disease https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

Перед нами стоит задача оценить среднее диастолическое давление мужчин и женщин с помощью 95% доверительного интервала

```
> head(tidy_set[tidy_set$ap_hi < tidy_set$ap_lo,])
       id   age gender height weight ap_hi ap_lo cholesterol gluc smoke alco active cardio age_years
469   681 19099      1    156     65   120   150           2    1     0    0      1      0         52
628   913 20457      2    169     68    70   110           1    1     0    0      1      0         56
2342 3356 23361      1    154    102    90   150           1    1     0    0      0      1         64
2931 4214 21957      2    182     90    80   140           3    3     0    0      1      1         60
3383 4880 19992      2    180     80    80   125           3    3     1    1      1      1         54
3556 5130 21874      1    160     83    80   120           1    1     0    0      1      0         59
```

```
12  .tidy_set<-tidy_set[tidy_set$ap_hi>tidy_set$ap_lo,]
13  dim(.tidy_set)
14  dim(tidy_set)
15  .women<-.tidy_set$ap_lo[.tidy_set$gender==1]
16  .men<-.tidy_set$ap_lo[.tidy_set$gender==2]
```

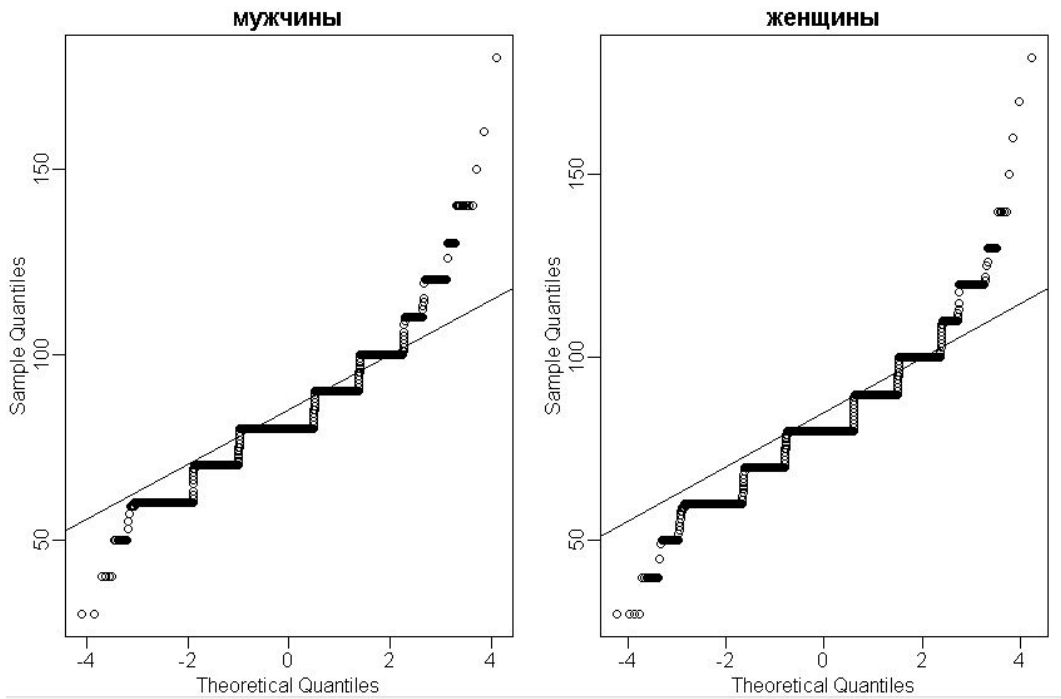
```
> dim(.tidy_set)
[1] 68678    14
> dim(tidy_set)
[1] 68781    14
```

Убедимся в предположении о нормальности

Исходя из того, что сигма неизвестна и данные приближены к нормальному распределению (хотя верхние и нижние значения лежат дальше, чем предполагалось нормальным распределением) ,

можем использовать t- критерий.

```
mypar(1,2)
qqnorm(.men, main = "мужчины")
qqline(.men)

qqnorm(.women, main =  "женщины")
qqline(.women)
```



С помощью функции summarize(), можем построить сводную таблицу, где будут подсчитаны нужные статистические значения

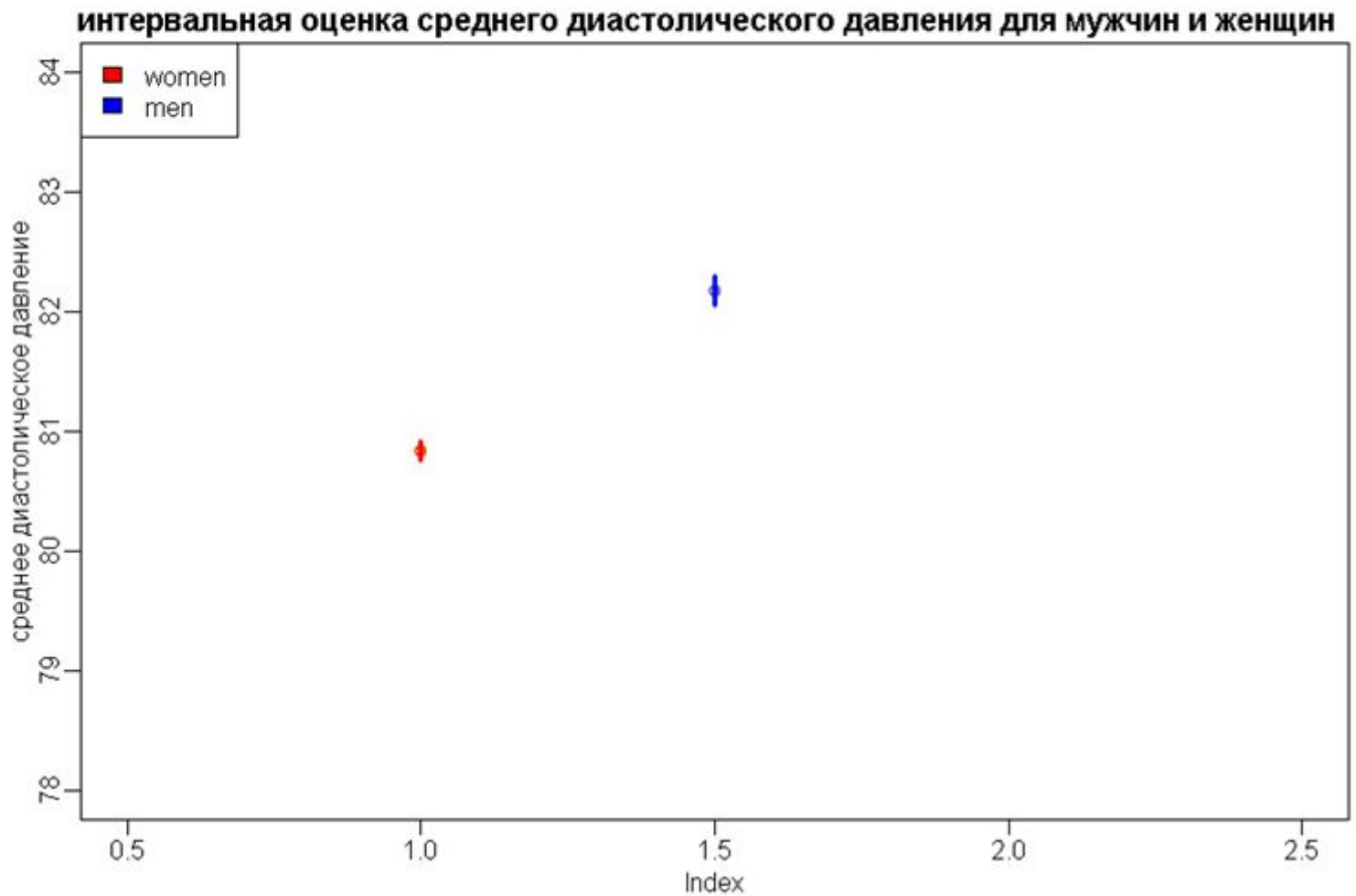
```
.tidy_set%>%group_by(gender)%>%summarise(
  mu=mean(ap_lo),
  k=qt(0.975,length(ap_lo)-1),
  se=sd(ap_lo)/sqrt(length(ap_lo)),
  lowlevel=mean(ap_lo)-k*se,
  hilevel=mean(ap_lo)+k*se)
```

```
# A tibble: 2 x 6
  gender    mu     k      se lowlevel hilevel
   <int> <dbl> <dbl>   <dbl>    <dbl>   <dbl>
1      1  80.8  1.96 0.04500     80.7    80.9
2      2  82.2  1.96 0.06011     82.1    82.3
```

```
> infer<- .tidy_set%>%group_by(gender)%>%summarise(
+   mu=mean(ap_lo),
+   k=qt(0.975,length(ap_lo)-1),
+   se=sd(ap_lo)/sqrt(length(ap_lo)),
+   lowlevel=mean(ap_lo)-k*se,
+   hilevel=mean(ap_lo)+k*se)
> infer
# A tibble: 2 x 6
  gender    mu     k      se lowlevel hilevel
   <int> <dbl> <dbl>   <dbl>    <dbl>   <dbl>
1      1  80.8  1.96 0.0450      80.7    80.9
2      2  82.2  1.96 0.0601      82.1    82.3
> ci_w<-c(infer[1,5], infer[1,6])
> ci_w<-as.numeric(c(infer[1,5], infer[1,6]))
> ci_w
[1] 80.74758 80.92404
> ci_m<-c(infer[2,5], infer[2,6])
> ci_m<- as.numeric(ci_m)
> ci_m
[1] 82.05777 82.29340
```

Изобразим графически интервальные оценки для мужчин и женщин

```
plot(mean(.women),col=2, lwd=2, xlim=c(0.5,2.5), ylim=c(78,84),ylab="среднее диастолическое давление", main=
     '"интервальная оценка среднего диастолического давления для мужчин и женщин")
interval=c(80.75 ,80.92)
lines(x=c(1,1),y=interval,col="red",lwd=3)
points(1.5,mean(men), col=3, lwd=2)
interval_1=c(82.05, 82.29)
lines(x=c(1.5,1.5),y=interval_1,col="blue",lwd=3)
legend("topleft",c("women","men"), fill=c("red","blue"))
```

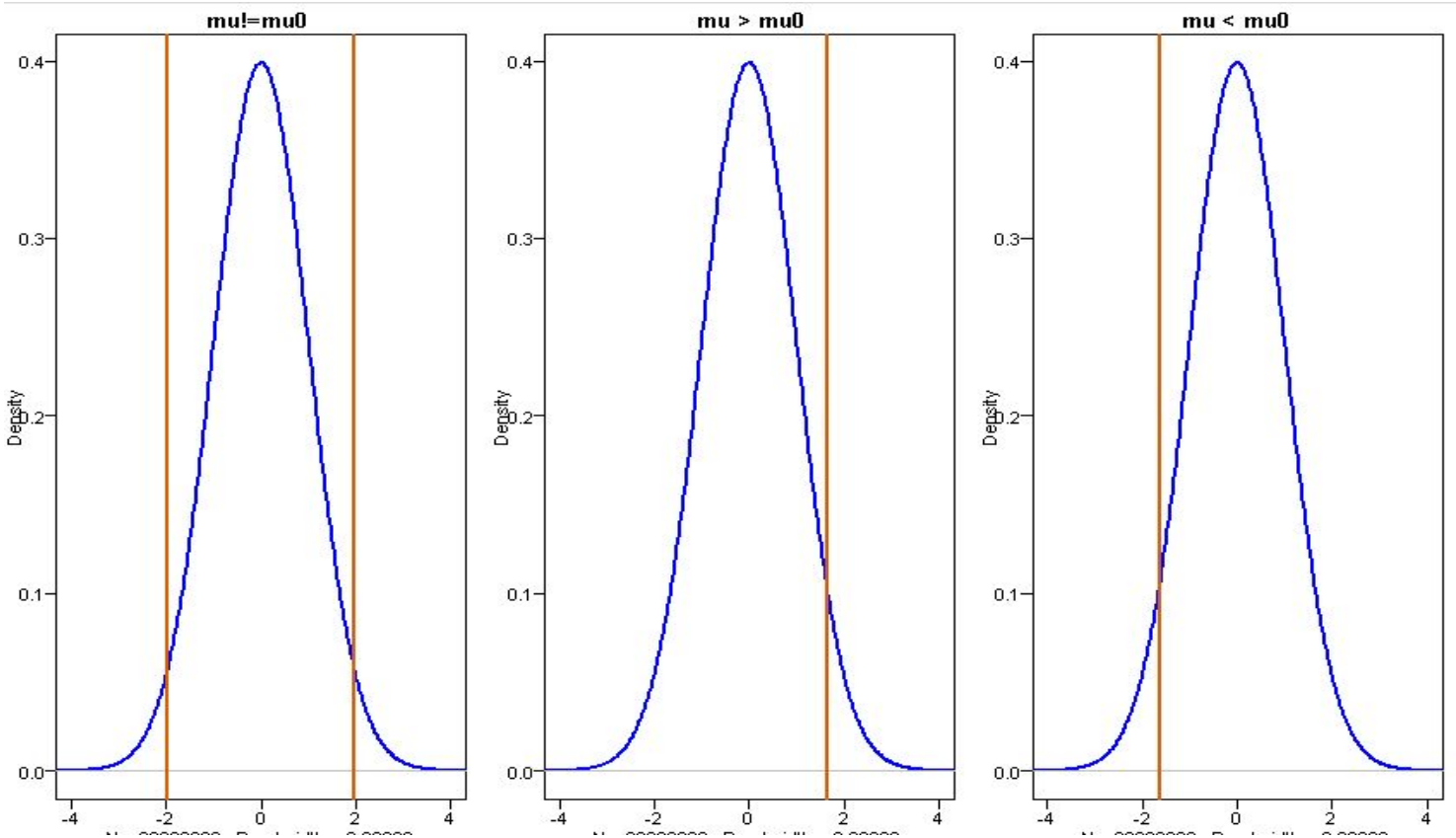


2. Тестирование гипотезы

2.1 одновыборочный t- критерий

2.2 двухвыборочный t- критерий для независимых выборок

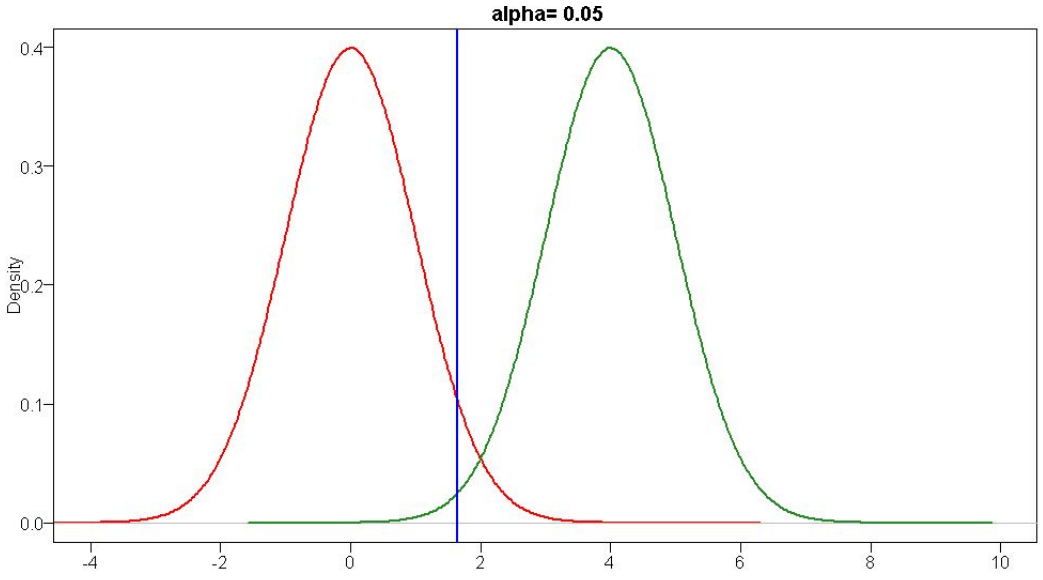
2.2 двухвыборочный t- критерий для зависимых выборок

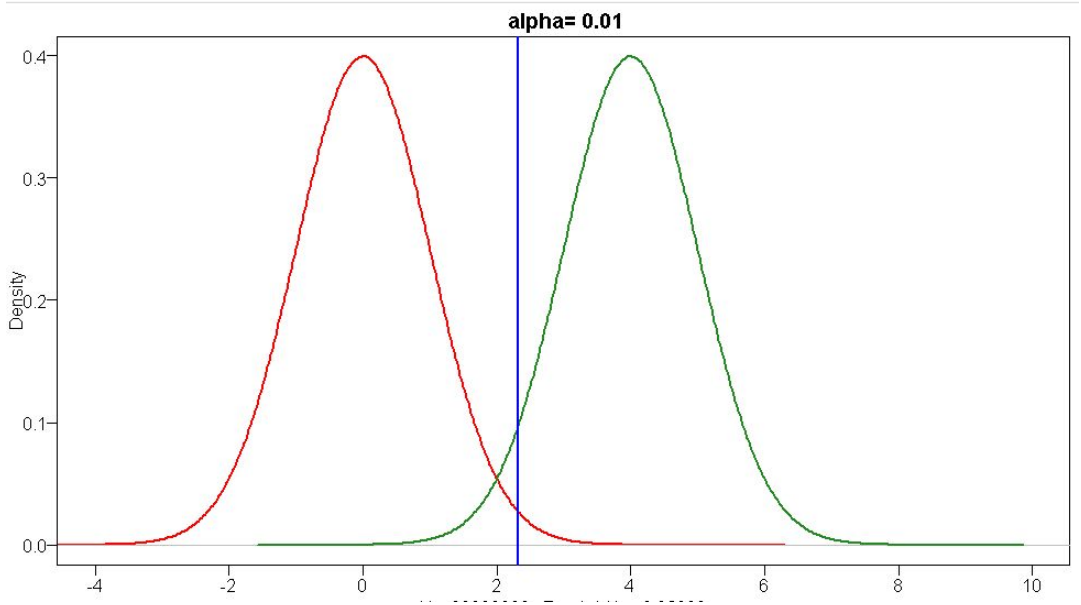


Вероятность ошибки первого рода α (0.1, 0.05, 0.01)

Вероятность ошибка второго рода β : считается, что эта вероятность не должна превышать 20%

Мощность теста (1 – β) : не менее 80%





Что влияет на мощность теста:

1 величины эффекта (разница между средними значениями);

2 объем выборки;

3 выбор уровня значимости альфа

4 разброс

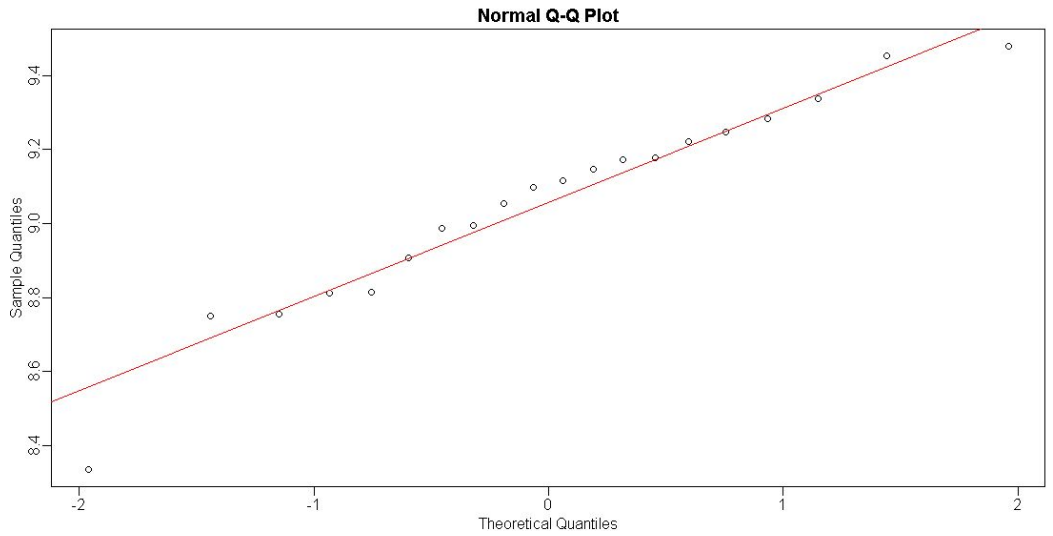
5 количество групп

Мы выбираем поставщика. Поставщик заявляет ,что он изготавливает детали размером 9 см и стандартным отклонением 0.3 см. Мы взяли 20 деталей, измерили их и получили выборку “post”. Проверить односторонним тестом, что истинное среднее не равно 9 см. (для простоты понимания и расчета сначала проведем односторонний тест, хотя правильно - провести двусторонний)

```
> post
 [1] 8.812 9.055 8.749 9.479 9.099 8.754 9.146 9.221 9.173 8.908
[11] 9.454 9.117 8.814 8.336 9.337 8.987 8.995 9.283 9.246 9.178
```

1. Убеждаемся, что наблюдения независимы
2. Если небольшой объем выборки, проверяем на нормальность данные с помощью qq - графика

```
qqnorm(post)
qqline(post, col="red")
```



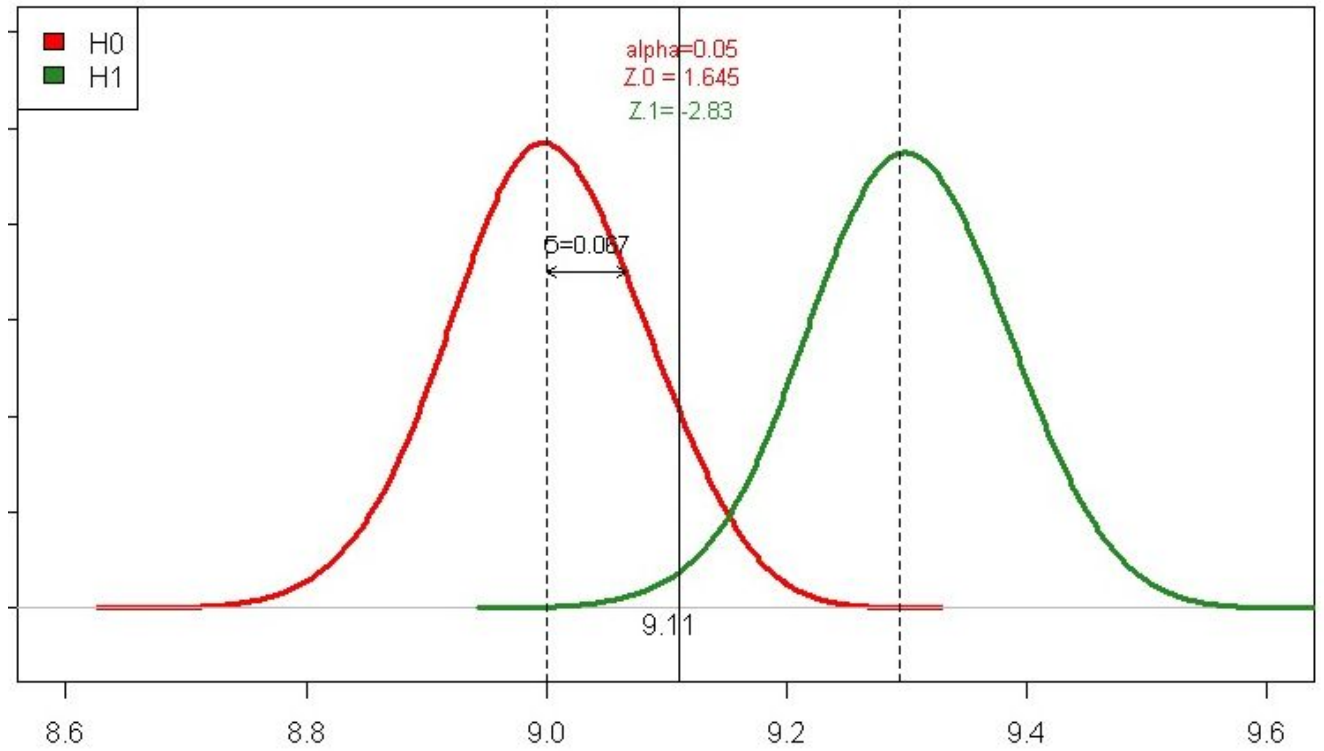
3 Установим гипотезу

H0 : mu = mu0

H1 : mu > mu0

4. Рассчитаем мощность теста:

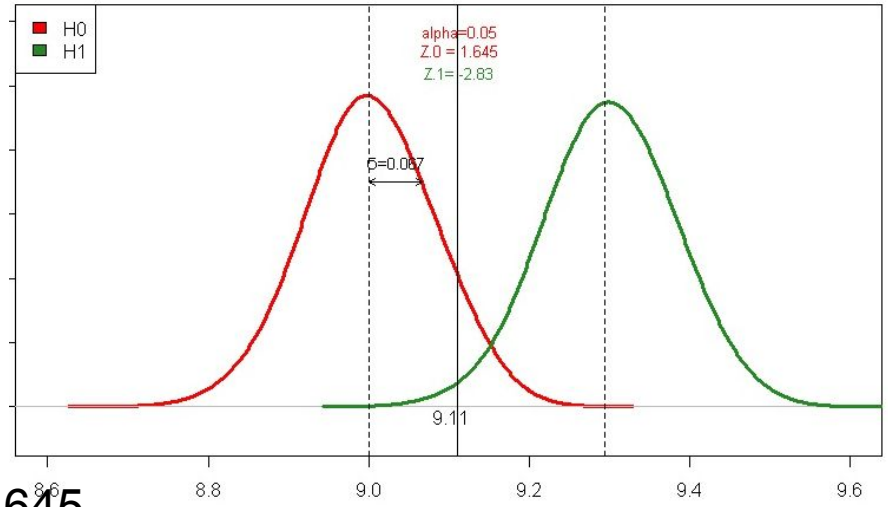
4.1. Предположим, что минимальная разница между измеренным средним и средним, заявленным производителем, которое мы хотим выявить при тестировании гипотезы 0.3 см. Известно стандартное отклонение 0.3 и альфа 0.05.



4.2 Произведем расчет мощности теста

4.2.1 Размер эффекта, значения ниже которого ,мы считаем, не имеют для нас смысла

ES (effect size) = 0.3 см



4.2.2 Рассчитываем Z.0 для альфа =0.05

```
> qnorm(0.95)
[1] 1.644854
```

4.2.3 Вычисляем значение среднего , соответствующего Z.0 =1.645

```
> sig= signif(0.3/sqrt(20),2)
> sig
[1] 0.067
> 9.0+ 1.645*0.067
[1] 9.110215
```

4.2.4 Вычислим значение Z.1 (для H1)

```
> (9.11-9.3)/0.067
[1] -2.835821
```

4.2.5 Вычисляем мощность теста

```
> 1-pnorm(-2.835)
[1] 0.9977087
```

R облегчает работу. Произведем расчет мощности теста с помощью функции power.z.test () из пакета “asbio”

```
install.packages("asbio") # устанавливаем пакет asbio
library(asbio)            # загружаем библиотеку
power.z.test(sigma=0.3, n=20, alpha = 0.05, effect = 0.3,test= "one.tail")
```

```
$sigma
[1] 0.3

$n
[1] 20

$power
[1] 0.9976528

$alpha
[1] 0.05

$effect
[1] 0.3

$test
[1] "one.tail"
```

Чтобы получить какое-то одно интересующее нас значение ,
например мощность теста

```
power.z.test(sigma=0.3, n=20, alpha = 0.05, effect = 0.3,test= "one.tail")$ power
[1] 0.9976528
```

Сравним значение, рассчитанные с помощью функции power.z.test() и вручную

```
  1-pnorm(-2.835)
[1] 0.9977087
```

Функцию power.z.test () также можно использовать, чтобы определить объем выборки для нужной мощности теста

```
 power.z.test(sigma=0.3, power=0.8, alpha = 0.05, effect = 0.3,test= "one.tail")$ n
[1] 6.182557
```

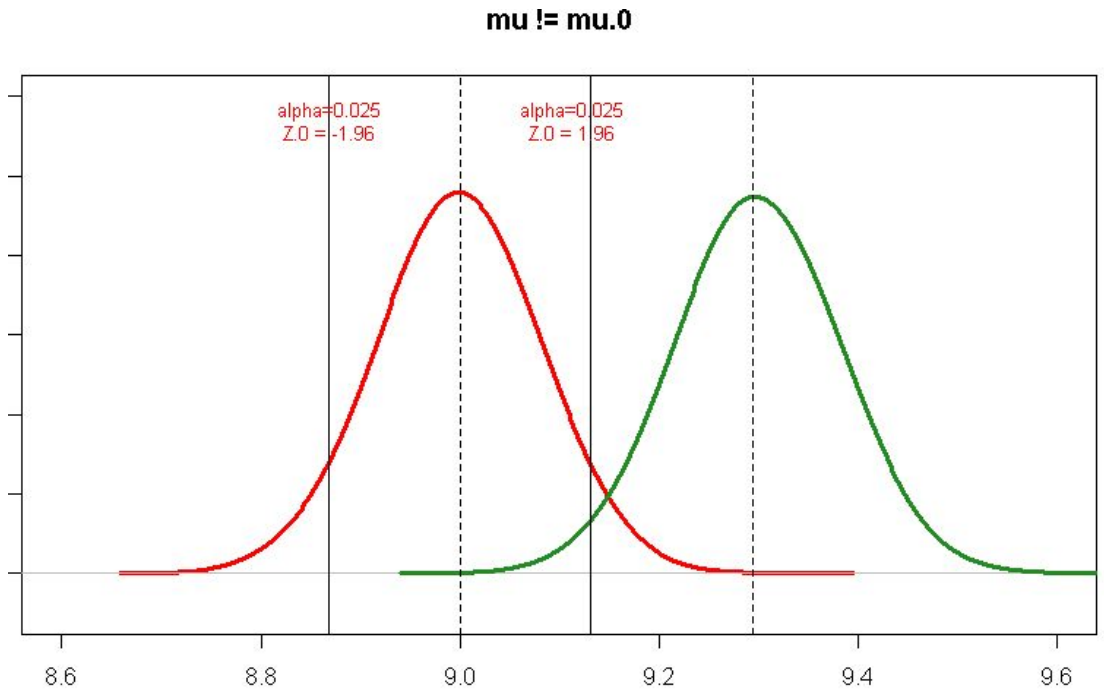
Чтобы мощность теста была не меньше 0.8 ,нужно взять выборку размером n=7. При n = 6 сила теста не будет достигать 80%

```
> power.z.test(sigma=0.3, n=6, alpha = 0.05, effect = 0.3,test= "one.tail")$ power
[1] 0.7894852
```

В этой задаче было бы правильнее проводить двусторонний тест. T.e. H1: Mu !=Mu.0

Проведем расчет мощности теста для двустороннего тестирования гипотезы с помощью функции power.z.test () У одностороннего теста мощность больше, чем у двустороннего.

```
> #проведем двусторонний тест
  power.z.test(sigma=0.3, n=20, alpha = 0.05, effect = 0.3,test= "two.tail", strict = "TRUE")
$sigma
[1] 0.3

$n
[1] 20

$power
[1] 0.9940005

$alpha
[1] 0.05

$effect
[1] 0.3

$test
[1] "two.tail"
```



5.Для тестирования гипотезы воспользуемся функцией z.test () из пакета “BSDA” (Выбираем критерий Z, поскольку он более предпочтителен при неизвестном стандартном отклонении

5.1 для начала сделаем односторонний тест

```
install.packages("BSDA") # загрузка пакета
library(BSDA) # загрузка библиотеки
z.test(post,alternative = "g",mu=9,sigma.x = 0.3)
```

```
        One-sample z-Test

data:  post
z = 0.85194, p-value = 0.1971
alternative hypothesis: true mean is greater than 9
95 percent confidence interval:
 8.94681      NA
sample estimates:
mean of x
  9.05715
```

* Рассчитаем наблюдаемое вручную и сравним со значением слева, что предоставляет функция

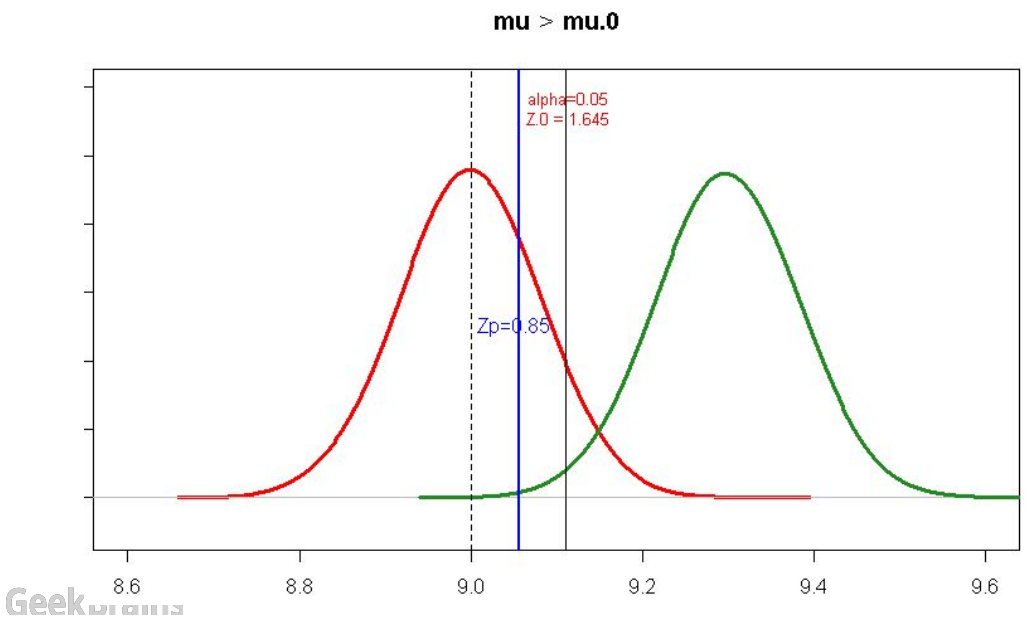
```
(mean(post)-9)*sqrt(20)/3
[1] 0.08519419
```

5.2 Поведем тест гипотезы в R только теперь **двусторонний**

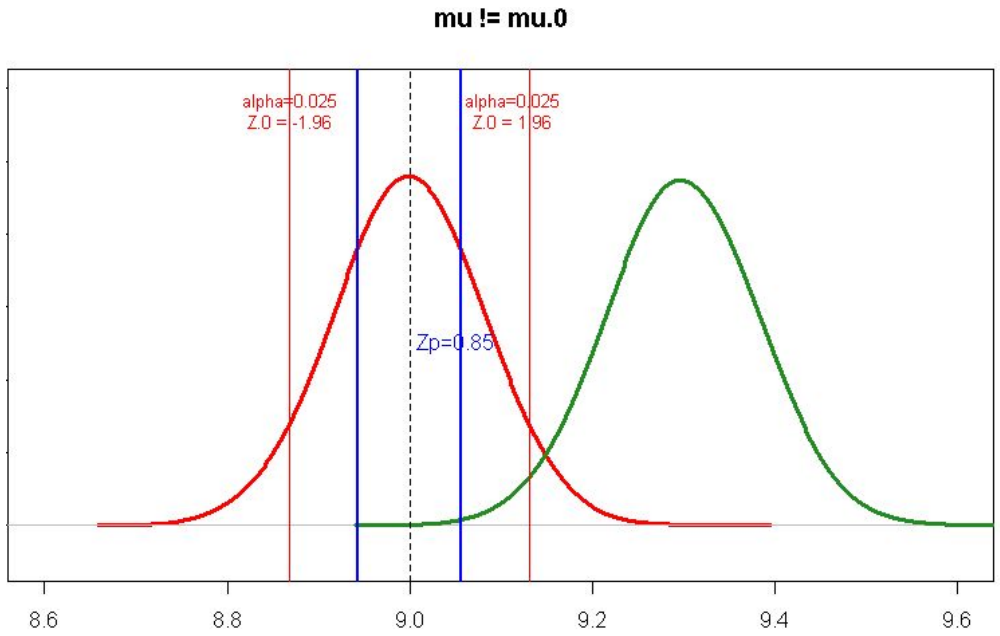
```
> z.test (post, alternative = "two.sided", mu = 9, sigma.x = 0.3)

        One-sample z-Test

data:  post
z = 0.85194, p-value = 0.3942
alternative hypothesis: true mean is not equal to 9
95 percent confidence interval:
 8.925672 9.188628
sample estimates:
mean of x
  9.05715

> 0.1971*2
[1] 0.3942
```

Сравним значения p-value для одностороннего и двустороннего теста:

```
z.test(post,alternative = "g",mu=9,sigma.x = 0.3)

        One-sample z-Test

data:  post
z = 0.85194, p-value = 0.1971
alternative hypothesis: true mean is greater than 9
95 percent confidence interval:
 8.94681      NA
sample estimates:
mean of x
  9.05715
```




```
> z.test (post, alternative = "two.sided", mu = 9, sigma.x = 0.3)

        One-sample z-Test

data:  post
z = 0.85194, p-value = 0.3942
alternative hypothesis: true mean is not equal to 9
95 percent confidence interval:
 8.925672 9.188628
sample estimates:
mean of x
  9.05715

> 0.1971*2
[1] 0.3942
```



6. Сделаем вывод *:

6.1 Для одностороннего теста:

Гипотеза H0 верна (среднее арифметическое =9 см) на уровне значимости 0.05 ,

p-value =0.1971

6.2. Для двустороннего теста :

Гипотеза H0 верна на уровне значимости 0.05

p-value =0.3942

Другой способ сообщить результаты теста – это сообщить доверительный интервал, который также посчитала функция z.test()

---

*Подробнее о выводах поговорим на примере реального набора данных

Для тестирования гипотезы с небольшими выборками(< 30) используют t-test, а также если неизвестно стандартное отклонение

T-test:

1. Одновыборочный - тест Стьюдента
2. Двухвыборочный :

2.1 Выборки с одинаковой дисперсией – тест Стьюдента

2.2 Выборки с разными дисперсиями – тест Уэлча

```
> t.test(sample(.men,20),sample(.women,20),var.equal = TRUE)

        Two Sample t-test

data:  sample(.men, 20) and sample(.women, 20)
t = 2.0153, df = 38, p-value = 0.05099
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.0239318 10.6239318
sample estimates:
mean of x mean of y
     82.0      76.7
```

```
> t.test(sample(.men,20),sample(.women,20),var.equal = FALSE)

        Welch Two Sample t-test

data:  sample(.men, 20) and sample(.women, 20)
t = 0.76252, df = 37.44, p-value = 0.4505
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.643592  8.043592
sample estimates:
mean of x mean of y
     82.5      80.3
```

$$df = n1 + n2 - 2$$

$$df = \frac{(S1^2/n1 + S2^2/n2)^2}{\frac{(S1^2/n1)^2}{n1-1} + \frac{(S2^2/n2)^2}{n2-1}}$$

t-test для независимых выборок

Сравним верхнее и нижнее диастолическое давление между мужчинами и женщинами

(.women, .men – слайд 17)

1. Посмотрим на данные : достаточно большой объем выборок, независимые измерения, независимые выборки. Стандартное отклонение неизвестно. С помощью ку-ку графика проверяем на нормальность. Все в порядке. Следовательно, однозначно используем t-критерий. Предполагаем, что выборки с разной дисперсией.
2. Формулируем нулевую и альтернативную гипотезы :

H0: mu= mu0

H1: mu != mu0

3. Устанавливаем уровень значимости α = 0.05

4. Рассчитываем мощность теста (1-β)

4.1 Для начала рассчитаем статистику d Коэна (Cohen`s d)*

разница между двумя средними, деленная на s.pool (общее)

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s} = \frac{\mu_1 - \mu_2}{s}.$$

s.pool

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

дисперсия группы

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_{1,i} - \bar{x}_1)^2$$

```
s.pool<-sqrt(((length(.women)-1)*var(.women)+(length(.men)-1)*var(.men))/
                (length(.women)+length(.men)-2))
s.pool
d<-(mean(.men)-mean(.women))/s.pool
d
```

```
[1] 0.141849
```

---

* Статистика Cohen`s d один из способов оценить размер эффекта. Показывает сколько стандартных общих отклонений между средними двух групп

С помощью функции cohen.d() из пакета “effsize” посчитаем статистику d Коэна

```
install.packages("effsize") # устанавливаем пакет
ibrary(effsize)  # загружаем библиотеку

  cohen.d(d=.men,.women)

Cohen's d

d estimate: 0.141849 (negligible)
95 percent confidence interval:
    lower     upper
0.1261364 0.1575616
  d<-cohen.d(.men,.women)$estimate
  d
[1] 0.141849
```

Cohen`s d	Размер эффекта (ES)
0.2	слабый
0.5	умеренный
0.8	сильный

Сравним значение, полученные с помощью функции cohen.d () с вычисленным вручную

```
> d<-(mean(.men)-mean(.women))/s.pool
> d
[1] 0.141849
```

4.2 Поскольку выборки разного размера используем, для расчета мощности теста функцию pwr.t2n.test () из пакета “pwr”. В противном случае можем воспользоваться pwr.t.test().

```
library("pwr")
pwr.t2n.test(n1=length(.women),n2=length(.men),d=d,sig.level = 0.05,alternative = "two.sided")
```

```
   t test power calculation

           n1 = 44735
           n2 = 23943
            d = 0.141849
    sig.level = 0.05
        power = 1
  alternative = two.sided
```

5. С помощью функции t.test, протестируем гипотезу

```
> t.test(.men,.women ,alternative = "two.sided")

        Welch Two Sample t-test

data:  .men and .women
t = 17.841, df = 49926, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.192583 1.486963
sample estimates:
mean of x mean of y
 82.17558  80.83581
```

6 . ВЫВОД

Мы получили очень маленькое значение p-value и нам следует отвергнуть нулевую гипотезу.

НО С УВЕЛИЧЕНИЕМ ВЫБОРКИ P-VALUE БУДЕТ УМЕНЬШАТЬСЯ.

И очень маленькие значения p-value не представляют уже научного интереса. Очень большие выборки позволяют обнаружить очень слабые различия, которые не несут научного смысла.

Чтобы было ясно, что мы нашли было бы правильно сообщить размер эффекта в % и его доверительный интервал, рассчитанные следующим образом:

```
> ((mean(.men)-mean(.women))/mean(.women))*100 # размер эффекта в %
[1] 1.6574
> ci<-t.test(.men,.women ,alternative = "two.sided")$conf.int
> (ci/mean(.women))*100          # доверительный интервал для ES
[1] 1.475315 1.839485
attr(,"conf.level")
[1] 0.95
```

Предположим мы хотим обнаружить сильный эффект. Cohen`s d = 0.8

Посчитаем ,сколько нужно выборок для обнаружения сильного эффекта

```
> pwr.t2n.test(n1=20,power = 0.8,d=0.8,sig.level = 0.05,alternative = "two.sided")#сколько нужно выборок,чтобы найти сильное различие

     t test power calculation

             n1 = 20
             n2 = 34.9757
              d = 0.8
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

> t.test(sample(.men,20),sample(.women,35))# не обнаружили большого эффекта

        Welch Two Sample t-test

data:  sample(.men, 20) and sample(.women, 35)
t = 1.5165, df = 45.038, p-value = 0.1364
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.012322  7.183750
sample estimates:
mean of x mean of y
 81.80000  78.71429
```

В данном случае H0 верна

Одновыборочный t.test

```
> t.test(sample(.men,20))

        One Sample t-test

data:  sample(.men, 20)
t = 50.355, df = 19, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 80.74811 87.75189
sample estimates:
mean of x
    84.25
```

Двухвыборочный t.test с зависимыми выбрками

```
t.test(x, y,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = TRUE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

Приобретенные навыки:

1. Научились находить квантили и строить доверительный интервал для оценки среднего.
2. Поняли, как работает центральная предельная теорема на выборках разных объемов и когда стоит применять t-распределение
3. Научились рассчитывать мощность теста при тестировании гипотез с известной и неизвестной сигмой
4. Познакомились со статистикой Cohen`s d
5. Научились рассчитывать объем выборки для обеспечения нужной мощности теста
6. Научились проводить тест гипотезы с известной и неизвестной сигмой в R

   6.1 Одновыборочный, двухвыборочный с зависимыми и независимыми выборками

   6.2 Узнали ,когда применяется тест Стьюдента, когда тест Уэлча

7. Научились правильно оформлять полученные результаты

   7.1. Научились рассчитывать размер эффекта и его доверительный интервал

   7.2 Поняли, как важно для формирования правильного вывода понимать, что с увеличением выборки, уменьшается p-value