

## Praktikum Eksplorasi dan Visualisasi Data Pertemuan 5

### Sampel Random Sederhana dan Distribusi Peluang Normal, T, dan F

---

#### Pengambilan Sampel Random Sederhana

Sampel random sederhana adalah sampel yang diambil dari suatu populasi dimana setiap elemen mempunyai peluang yang sama untuk diambil sebagai sampel. Pengambilan sampel dibedakan menjadi dua, yaitu **dengan pengembalian** dan **tanpa pengembalian**. Apabila sampel berukuran kecil, maka pengambilan sampel dengan pengembalian akan menghasilkan sampel random yang independen, sedangkan pengambilan sampel tanpa pengembalian akan menghasilkan sampel yang dependen. Namun, apabila ukuran sampel besar, perbedaan cara pengambilan sampel dapat diabaikan. Secara umum, pengambilan sampel untuk sampel kecil dapat dibedakan menjadi dua, yaitu:

- a. Pengambilan Sampel dengan Pengembalian (Sampling with Replacement)  
Sampel berukuran kecil, dapat menghasilkan sampel random yang independen.

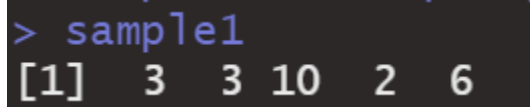
Di dalam R, fungsi yang digunakan untuk mengambil sampel adalah fungsi `sample()`. Penggunaannya dapat dilihat dari syntax di bawah:

#### Syntax :

```
# Contoh pengambilan data sampel
# definisikan populasinya terlebih dahulu
pop <- 1:10

#pengambilan sampel dengan pengembalian
set.seed(123)
sample1 <- sample(pop,size=5,replace=T)
sample1
```

#### Output :



```
> sample1
[1] 3 3 10 2 6
```

Dapat dilihat bahwa dengan pengembalian, angka 3 bisa muncul 2 kali.

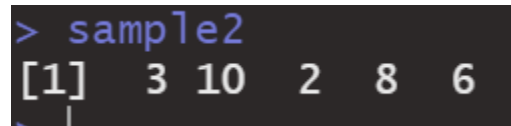
- b. Pengambilan Sampel tanpa Pengembalian (Sampling without Replacement)  
Sampel berukuran kecil, dapat menghasilkan sampel random yang dependen.

Syntax yang digunakan hampir sama dengan pengambilan sampel dengan pengembalian. Namun, perbedaannya terletak di argumen `replace`, dimana argumen yang digunakan adalah `False`. Contohnya seperti di bawah:

**Syntax :**

```
#pengambilan sampel tanpa pengembalian
set.seed(123)
sample2 <- sample(pop,size=5,replace=F)
sample2
```

**Output :**



```
> sample2
[1] 3 10 2 8 6
```

Dari output di atas, jika line tersebut dieksekusi berulang kali, angka yang dihasilkan tetap akan berbeda-beda.

## Distribusi Peluang

Metode analisis data konfirmasi membutuhkan asumsi normalitas. Ada dua macam distribusi lain yang merupakan turunan dari distribusi normal, yaitu distribusi t (student-t) dan distribusi F.

Distribusi yang akan dibahas antara lain :

- a. Distribusi Normal
- b. Distribusi t (student-t)
- c. Distribusi F

Dimana, ketiga distribusi tersebut adalah distribusi peluang variabel random kontinu yang dikenal dalam ilmu probabilitas. Maka dari itu, perlu diingat kembali sifat-sifat variabel random kontinu sebagai berikut :

1.  $P(X < x) = P(X \leq x)$
2.  $P(X > x) = P(X \geq x)$

Untuk distribusi yang simetris (distribusi normal dan t) kita bisa memakai sifat berikut :

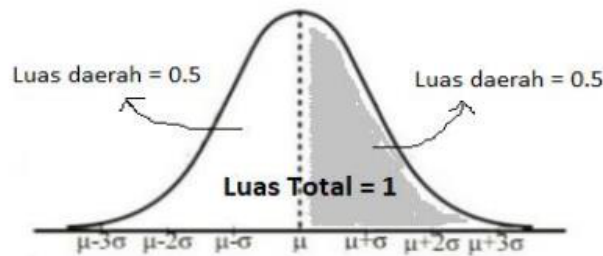
1.  $P(X \leq -x) = P(X \geq x)$

$$2. P(X \leq -x) = 1 - P(X \leq x)$$

**Di R, kita dapat mencari:**

1. Kalkulasi *probability density* (Menentukan peluang satu titik) suatu distribusi,
2. Kalkulasi *cumulative distribution* (Menentukan peluang kumulatif) suatu distribusi,
3. Kalkulasi *inverse cumulative distribution* (Menentukan invers peluang kumulatif) suatu distribusi.

## Distribusi Normal



**Ciri-ciri :**

1. Kurva distribusi berbentuk lonceng.
2. Simetris terhadap  $\mu$ .
3. Mean, median, dan modus relatif sama.
4.  $\mu$  (mean) merupakan pusat lonceng, sedangkan  $\sigma^2$  merupakan variansi/salah satu ukuran sebaran. Semakin besar  $\sigma^2$ , mengakibatkan distribusi normal semakin menyebar.

**Fungsi probabilitas dari distribusi normal**

$$P(X = x) = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

**Notasi penulisan**

Variabel random berdistribusi normal dengan mean  $\mu$  dan variansi  $\sigma^2$  dituliskan

$$X \sim N(\mu, \sigma^2)$$

## Transformasi normal ke Z

Transformasi variabel random X menjadi Z yang berdistribusi normal standard dengan mean 0 dan variansi 1, dituliskan  $Z \sim N(0,1)$

$$z = \frac{x - \mu}{\sigma}$$

## Aplikasi di R :

Untuk distribusi normal, syntax yang digunakan adalah `pnorm()`, `dnorm()`, dan `qnorm()`.

1. `pnorm()` Untuk mencari probabilitas kumulatif dari distribusi normal.  
[ $P(X \leq x)$ ]
2. `dnorm()` Untuk mencari densitas probabilitas dari distribusi normal. [ $P(X = x)$ ]
3. `qnorm()` Untuk mencari invers dari probabilitas kumulatif.  
[ $x = P^{-1}(p), 0 \leq p \leq 1$ ]

Contoh penggunaannya:

*Taylor Manifest Anxiety Scale* (TMAS) secara luas digunakan dalam penelitian psikologi. Skor TMAS mental pasien penderita depresi dianggap berdistribusi normal dengan mean 47,6 dan deviasi standar 10,3.

- a. Apabila diambil satu pasien secara random, berapa peluang skor TMAS pasien tersebut adalah 50?
- b. Berapa peluang seorang pasien mendapatkan skor kurang dari 47,6?
- c. Berapa peluang seorang pasien mendapatkan skor lebih dari 45,5?
- d. Jika dalam suatu penelitian diketahui bahwa 10% pasien yang diteliti memiliki skor TMAS tinggi, berapa nilai skor terendah untuk kategori tinggi ini?

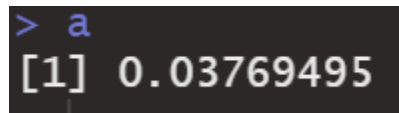
Jawaban :

- a. Peluang skor TMAS pasien tersebut adalah 50

### Syntax :

```
# Peluang skor TMAS pasien tersebut adalah 50
a = dnorm(50, mean=47.6, sd=10.3)
a
```

### Output :



```
> a
[1] 0.03769495
```

$P(X = 50) = 0,03769495$

- b. Peluang seorang pasien mendapatkan skor kurang dari 47,6

### Syntax :

```
# Peluang seorang pasien mendapatkan skor kurang dari
47,6
b = pnorm(47.6, mean=47.6, sd=10.3, lower.tail = TRUE)
```

b

**Output :**

```
> b  
[1] 0.5
```

$$P(X \leq 10,3) = 0,5$$

- c. Peluang seorang pasien mendapatkan skor lebih dari 45,5

**Syntax :**

```
# Peluang seorang pasien mendapatkan skor lebih dari  
45,5
```

```
c = 1-pnorm(45.5,mean=47.6,sd=10.3,lower.tail = TRUE)
```

```
c
```

```
# atau
```

```
c2 = pnorm(45.5,mean=47.6,sd=10.3,lower.tail = FALSE)
```

```
c2
```

**Output :**

```
> c = 1-pnorm(45.5,mean=47.6,sd=10.3,lower.tail = TRUE)  
> c  
[1] 0.5807777  
> # atau  
> c2 = pnorm(45.5,mean=47.6,sd=10.3,lower.tail = FALSE)  
> c2  
[1] 0.5807777
```

$$P(X \geq 45,5) = 1 - P(X \leq 45,5)$$

$$= 1 - 0,419222$$

$$= 0,580778.$$

Jadi, Peluang seorang pasien mendapatkan skor lebih dari 45,5 adalah 0,580778.

- d. Diketahui bahwa 10% pasien yang diteliti memiliki skor TMAS tinggi, berapa nilai skor terendah untuk kategori tinggi ini?

$$P(X \geq x) = 0,1$$

$$1 - P(X \leq x) = 0,1$$

$$P(X \leq x) = 0,9$$

**Syntax :**

```
# Inverse CDF 0.9
```

```
d = qnorm(0.1,mean=47.6,sd=10.3,lower.tail = FALSE)
```

```
d
```

```
# atau
```

```
d2 = qnorm(0.9,mean=47.6,sd=10.3,lower.tail = TRUE)
```

```
d2
```

**Output :**

```
> d = qnorm(0.1,mean=47.6,sd=10.3,lower.tail = FALSE)
> d
[1] 60.79998
> # atau
> d2 = qnorm(0.9,mean=47.6,sd=10.3,lower.tail = TRUE)
> d2
[1] 60.79998
```

Jadi, nilai skor terendah untuk kategori tinggi ini adalah 60,79998.

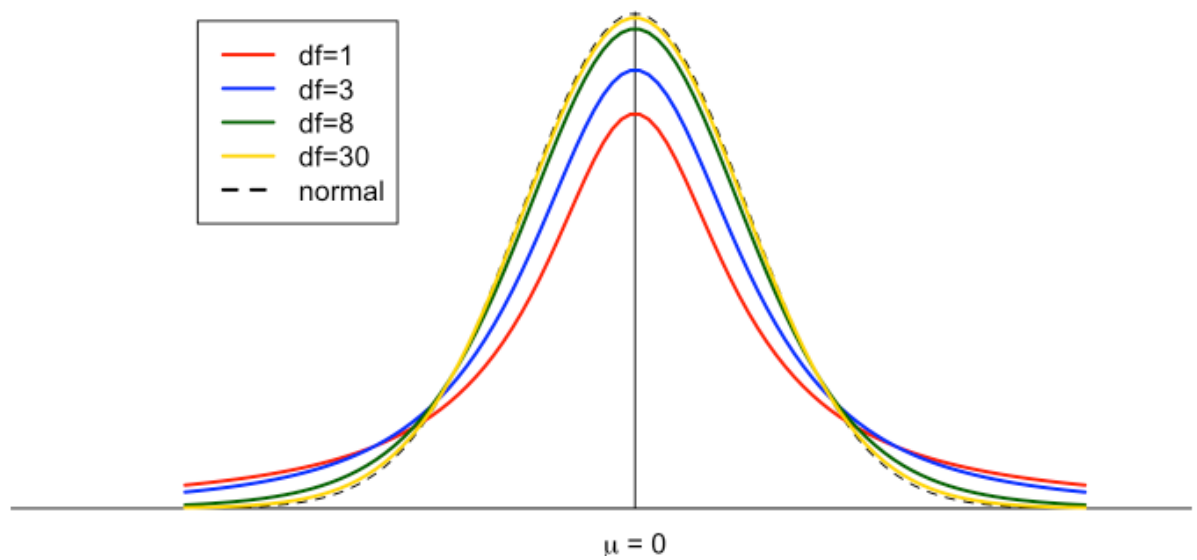
## Distribusi T

The  $t$ -distribution curve is, like the normal distribution curve, symmetric (bell shaped) about the mean. However, the  $t$ -distribution curve is flatter than the standard normal distribution curve. Consequently, the  $t$ -distribution curve has a lower height and a wider spread than the standard normal distribution.

The  $t$ -distribution has only one parameter, called the **degrees of freedom** (df). The shape of a particular  $t$ -distribution curve depends on the number of degrees of freedom. The number of degrees of freedom for a  $t$ -distribution is equal to the sample size minus one, that is,

$$df = n - 1$$

### Comparison of t-Distributions



## Distribusi T dalam Software R

### Description

Density, distribution function, quantile function and random generation for the t distribution with df degrees of freedom (and optional non-centrality parameter ncp).

### Usage

```
dt(x, df, ncp, log = FALSE)
pt(q, df, ncp, lower.tail = TRUE, log.p = FALSE)
qt(p, df, ncp, lower.tail = TRUE, log.p = FALSE)
rt(n, df, ncp)
```

### Arguments

<code>x, q</code>	vector of quantiles.
<code>p</code>	vector of probabilities.
<code>n</code>	number of observations. If <code>length(n) &gt; 1</code> , the length is taken to be the number required.
<code>df</code>	degrees of freedom ( $> 0$ , maybe non-integer). <code>df = Inf</code> is allowed.
<code>ncp</code>	non-centrality parameter <i>delta</i> ; currently except for <code>rt()</code> , only for <code>abs(ncp) &lt;= 37.62</code> . If omitted, use the central t distribution.
<code>log, log.p</code>	logical; if TRUE, probabilities p are given as $\log(p)$ .
<code>lower.tail</code>	logical; if TRUE (default), probabilities are $P[X \leq x]$ , otherwise, $P[X > x]$ .

## Probabilitas density function

```
dt(x, df, ncp, log = FALSE)
```

### Example

```
#a. calculate the probability density function at x = 1.96
with df 5000
dt(1.96, 5000)
```

```
> dt(1.96, 5000)
[1] 0.05845868
```

```
#b. calculate the probability density function at t=-4,2,0,2,4
with df=5
```

```
x<-seq(from=-4, to=4, by=2)
a<-dt(x, 5)
```

```
a
```

```
> a
[1] 0.005123727 0.065090310 0.379606690 0.065090310 0.005123727
```

## Peluang kumulatif

```
pt(q, df, ncp, lower.tail = TRUE, log.p = FALSE)
```

### Example

```
# Mencari  $P(X \leq -2)$  dengan df 5
```

```
pt(-2,5)
```

```
> pt(-2,5)  
[1] 0.05096974
```

```
# Mencari  $P(X \leq -2)$ ,  $P(X \leq -0)$ , dan  $P(X \leq 2)$  dengan df 5
```

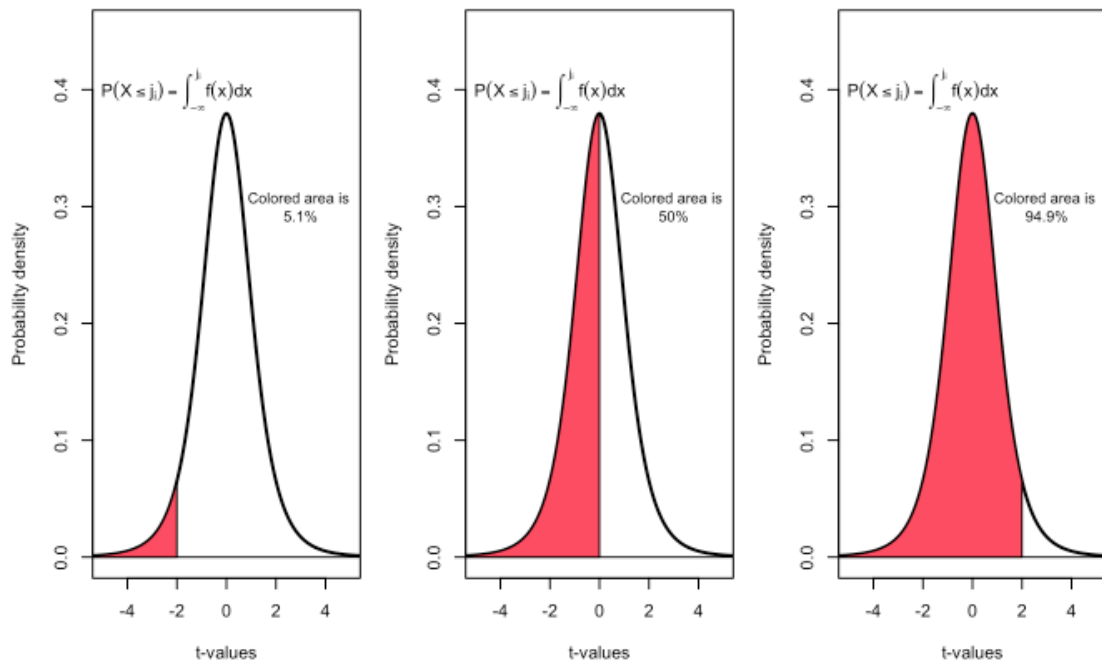
```
df <- 5
```

```
ji <- c(-2,0,2)
```

```
pt(ji, df = df, lower.tail = TRUE)
```

```
> pt(ji, df = df, lower.tail = TRUE)  
[1] 0.05096974 0.50000000 0.94903026
```

Area under the curve for  $j_i$



```
#jika lower.tail = FALSE
```

**Berarti ini nilai peluang dari???**

```
df <- 5
```

```
ki <- c(-2,0,2)
```

```
f<-pt(ki, df = df, lower.tail = FALSE)
```

```
f
```



```
> f
[1] 0.94903026 0.50000000 0.05096974
```

**Coba dipikirkan jika tiap-tiap nilai ki df nya berbeda-beda!**

### Nilai Quantile/Nilai Invers dari Probabilitas Kumulatif

```
qt(p, df, ncp, lower.tail = TRUE, log.p = FALSE)
```

#### Example

```
#Mencari nilai x jika  $P(X \leq x) = 0.05096974$  dengan df 5
qt(0.05096974, 5)
```

```
> qt(0.05096974, 5)
[1] -2
```

```
#Mencari nilai masing-masing x jika
 $P(X < x) = 0.05096974$ ,  $P(X < x) = 0.50000000$ , dan  $P(X < x) = 0.94903026$  dengan nilai
df masing-masing 5
qt(c(0.05096974, 0.50000000, 0.94903026), 5)
```

```
> qt(c(0.05096974, 0.50000000, 0.94903026), 5)
[1] -2 0 2
```

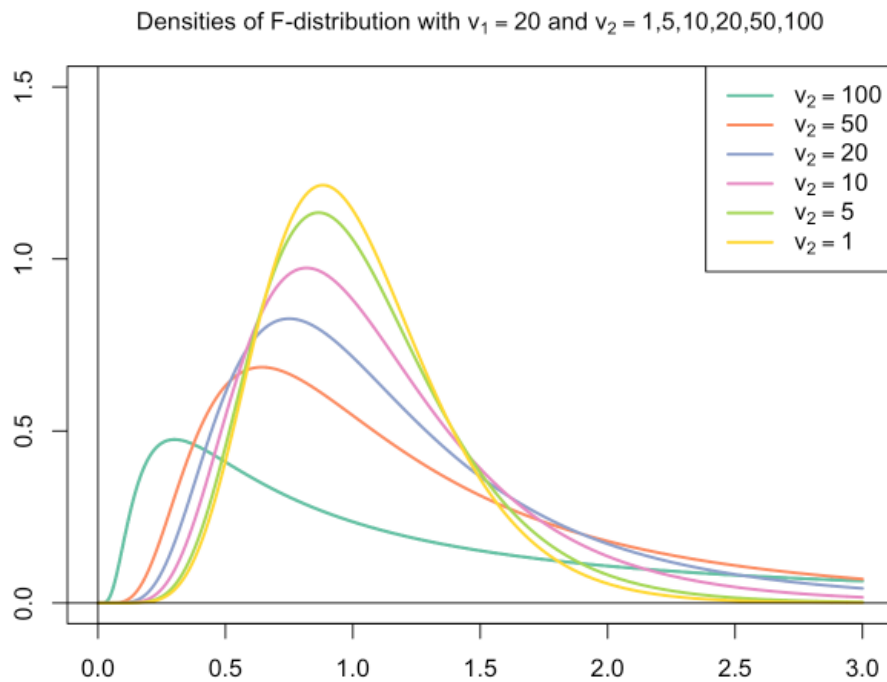
```
#
hasil<-qt(f, 5, lower.tail = FALSE)
data.frame(Nilai_probabilitas=f,
           Nilai_invers=hasil)

  Nilai_probabilitas Nilai_invers
1         0.94903026          -2
2         0.50000000           0
3         0.05096974           2
```

### Distribusi F

The Snedecor's F-distribution or the Fisher-Snedecor distribution or short the F-distribution is a continuous probability distribution with range  $[0, +\infty]$ , depending on two parameters denoted  $v_1, v_2$  (Lovric, 2011). In statistical applications,  $v_1, v_2$  are positive integers.

A F-distribution has two numbers of degrees of freedom,  $v_1$  and  $v_2$ , determining its shape. The first number of degrees of freedom  $v_1$ , called degrees of freedom of numerator and the second,  $v_2$  the degree of freedom of the denominator.



## Distribusi F dalam Software R

### Description

Density, distribution function, quantile function and random generation for the F distribution with  $df1$  and  $df2$  degrees of freedom (and optional non-centrality parameter  $ncp$ ).

### Usage

```
df(x, df1, df2, ncp, log = FALSE)
pf(q, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)
qf(p, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)
rf(n, df1, df2, ncp)
```

### Arguments

<code>x, q</code>	vector of quantiles.
<code>p</code>	vector of probabilities.
<code>n</code>	number of observations. If <code>length(n) &gt; 1</code> , the length is taken to be the number required.
<code>df1, df2</code>	degrees of freedom. <code>Inf</code> is allowed.
<code>ncp</code>	non-centrality parameter. If omitted the central F is assumed.
<code>log, log.p</code>	logical; if TRUE, probabilities $p$ are given as $\log(p)$ .
<code>lower.tail</code>	logical; if TRUE (default), probabilities are $P[X \leq x]$ , otherwise, $P[X > x]$ .

## Probabilitas density function

```
df(x, df1, df2, ncp, log = FALSE)
```

### Example

```
#Mencari P(X=2) dengan df1 = 10, df2 = 20
df(1.2, df1 = 10, df2 = 20)
> df(1.2, df1 = 10, df2 = 20)
[1] 0.5626125
```

## Peluang kumulatif

```
pf(q, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)
```

### Example

We use the pf() to calculate the area under the curve for the interval [0,1.5] and the interval [1.5,+infinity) of a F-curve with with v1=10 and v2=20. Further we ask R if the sum of the intervals [0,1.5] and [1.5,+infinity) sums up to 1

```
x = 1.5
v1 = 10
v2 = 20
```

```
# interval [0,1.5]
pf(x, df = v1, df2 = v2, lower.tail = TRUE)
> pf(x, df = v1, df2 = v2, lower.tail = TRUE)
[1] 0.7890535
```

```
# interval [1.5,+inf)
pf(x, df = v1, df2 = v2, lower.tail = FALSE)
> pf(x, df = v1, df2 = v2, lower.tail = FALSE)
[1] 0.2109465
~
```

```
#apakah jumlahnya mencapai 1?
pf(x, df = v1, df2 = v2, lower.tail = TRUE) + pf(x, df = v1, df2 = v2,
lower.tail = FALSE) == 1
[1] TRUE
```

## Nilai Quantile/Nilai Invers dari Probabilitas Kumulatif

```
qf(p, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)
```

### Example

We use the `qf()` to calculate the quantile for a given area (= probability) under the curve for a F-curve with  $v_1=10$  and  $v_2=20$  that corresponds to  $q=0.25, 0.5, 0.75$  and  $0.999$ . We set `lower.tail = TRUE` in order to get the area for the interval  $[0, q]$ .

```
#Nilai probabilitas
q <- c(0.25, 0.5, 0.75, 0.999)
#df 1
v1=10
#df2
v2=20

#Mencari nilai x dari  $P(X \leq x) = 0.25$  dengan df1 10 dan df2 20
qf(q[1], df1 = v1, df2 = v2, lower.tail = TRUE)
> qf(q[1], df1 = v1, df2 = v2, lower.tail = TRUE)
[1] 0.6563936

#
qf(q[2], df1 = v1, df2 = v2, lower.tail = TRUE)
> qf(q[2], df1 = v1, df2 = v2, lower.tail = TRUE)
[1] 0.9662639

#
qf(q[3], df1 = v1, df2 = v2, lower.tail = TRUE)
> qf(q[3], df1 = v1, df2 = v2, lower.tail = TRUE)
[1] 1.399487

#
qf(q[4], df1 = v1, df2 = v2, lower.tail = TRUE)
> qf(q[4], df1 = v1, df2 = v2, lower.tail = TRUE)
[1] 5.075246
```

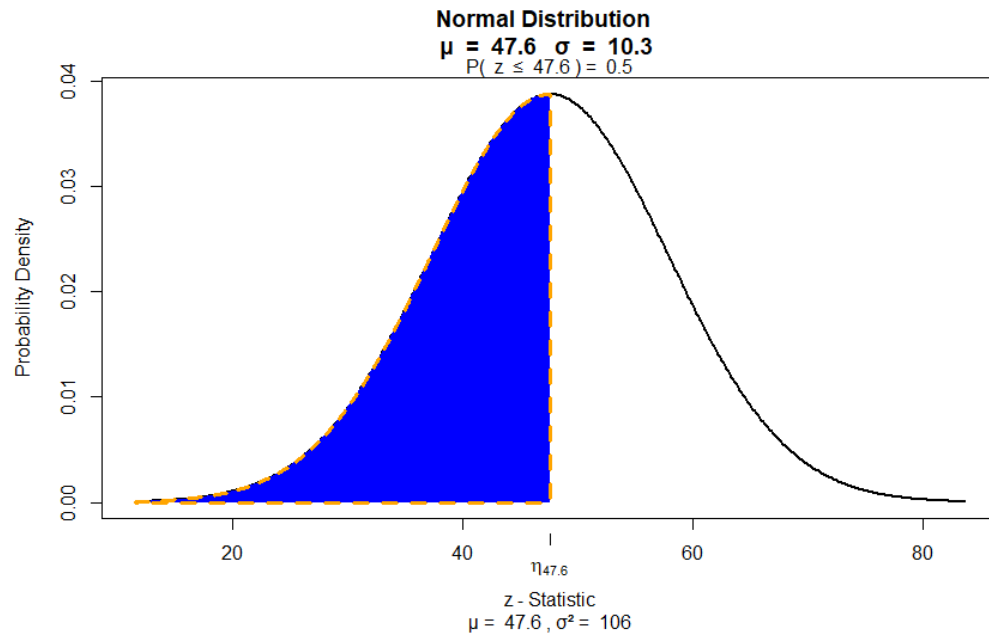
## Plot Distribusi Peluang

Diambil contoh kita ingin membuat plot dari contoh soal distribusi normal sebelumnya

- Plot peluang seorang pasien mendapatkan skor kurang dari 47,6

**Syntax :**

```
# Plot peluang seorang pasien mendapatkan skor kurang
dari 47,6
library(visualize)
visualize.norm(stat = 47.6,mu=47.6,sd=10.3,section =
"lower")
```

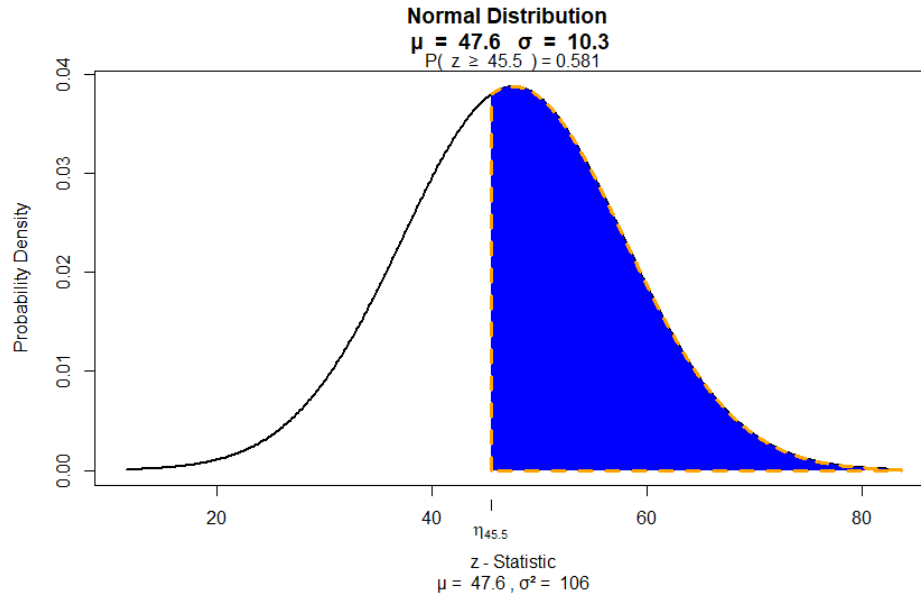
**Output :****Interpretasi :**

Berdasarkan plot peluang distribusi normal yang telah didapatkan dengan software R. didapatkan hasil bahwa peluang seorang pasien mendapatkan skor kurang dari 47,6 dengan mean 47.6 dan standar deviasi 10.3 adalah 0.5. Dengan plot probabilitas *lower section (lower tail)* seperti diatas.

- b. Plot peluang seorang pasien mendapatkan skor lebih dari 45,5

**Syntax :**

```
# Plot peluang seorang pasien mendapatkan skor lebih
dari 45,5
visualize.norm(stat = 45.5,mu=47.6,sd=10.3,section =
"upper")
```



### Interpretasi :

Berdasarkan plot peluang distribusi normal yang telah didapatkan dengan software R, didapatkan hasil bahwa peluang seorang pasien mendapatkan skor lebih dari 45,5 dengan mean 47.6 dan standar deviasi 10.3 adalah 0.581. Dengan plot probabilitas *upper section* (*upper tail*) seperti diatas.

**Note = Untuk distribusi lainnya (t, f, dan lain lain) bisa dipelajari sendiri dalam dokumentasi package “visualize” yang sudah disertakan**

### Latihan Soal :

1. Kota Hiroshi merupakan kota kecil di Negara Nagito. Kota tersebut mempunyai penduduk sebanyak 100 orang. Seorang peneliti ingin mengetahui besar gaji di kota tersebut. Dikumpulkanlah data mengenai gaji di kota tersebut pada “Pertemuan 5.xlsx”.
  - a. Carilah mean dan standar deviasi dari populasi tersebut.
  - b. Lakukan pengambilan sampel dengan pengembalian sebesar 30 sampel. Namai dengan sampel A.
  - c. Lakukan pengambilan sampel tanpa pengembalian sebesar 30 sampel. Namai dengan sampel B.
  - d. Hitung mean dan standar deviasi masing-masing sampel A dan sampel B.
  - e. Bandingkan perhitungan mean dan standar deviasi dari sampel A dan sampel B dengan populasi. Apa kesimpulan yang dapat kalian peroleh?
2. Suatu peneliti melakukan uji statistik dengan memanfaatkan distribusi F. Peneliti itu ingin mencari estimasi titik F, untuk daerah kritis 0.95 dengan nilai 5 sebagai pembilang dan 10 sebagai penyebut yang akan digunakan dalam pengujiannya, bantulah peneliti tersebut!

3. Seorang mahasiswa berkendara setiap harinya dari kos menuju kampus. Diketahui distribusi waktu berkendara mahasiswa tersebut berdistribusi normal dengan rata-rata waktu perjalanan sebesar 24 menit, dan standar deviasi 3.8 menit. Jika mahasiswa tersebut memiliki kelas pada jam 07.30 WIB dan mahasiswa tersebut baru meninggalkan kos pada pukul 07.15 WIB, hitung persentase waktu mahasiswa tersebut terlambat kuliah. (Gunakan plot distribusi peluang untuk membantu visualisasi peluangnya)
4. Diketahui tekanan darah sistolik suatu pasien di Rumah Sakit B sebagai berikut :  
183 152 178 157 114 163 144 114 178 152  
Berapa peluang tekanan darah sistolik sama dengan 145? (data berdistribusi t)