10. t 分布を利用した母平均の推定

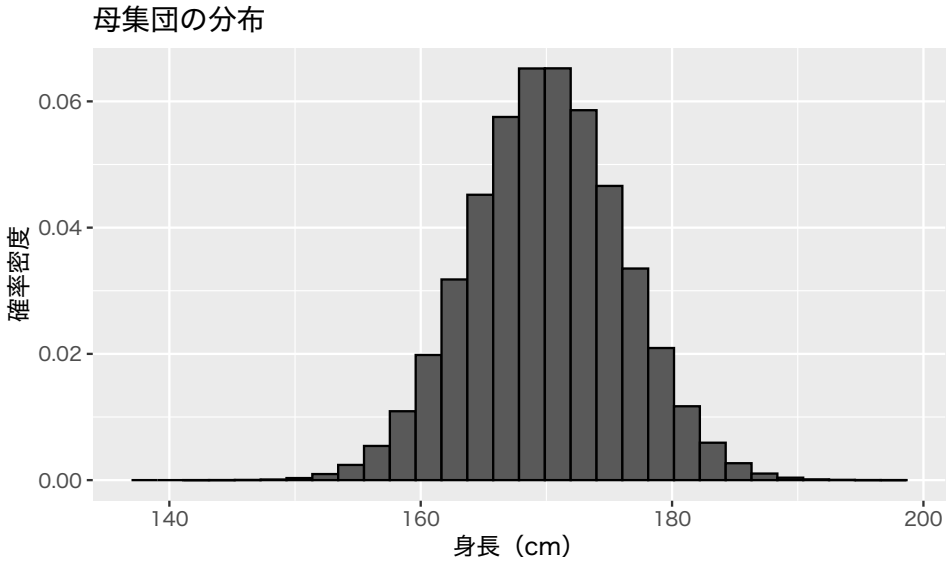
honocat

2025-12-22

母集団を定義する

成人男性の慎重に興味があるとする。母集団の人口を 100 万人、母平均を 170cm、母標準偏差を約 6cm に設定する。

```
pop <- rnorm(1e6, mean = 170, sd = 6)
pop_height <- ggplot(tibble(pop),
                     aes(x = pop,
                         y = after_stat(density))) +
  geom_histogram(color = 'black') +
  labs(x = '身長(cm) ',
       y = '確率密度',
       title = '母集団の分布')
plot(pop_height)
```



この母集団の身長の平均(母平均)は、

```
mean(pop)
```

```
[1] 169.9937
```

であり、身長の標準偏差は、

```
sd(pop)
```

```
[1] 5.999403
```

である。

標本を抽出して母平均を推定する

標本抽出と母平均の推定のシミュレーション

100 万人の母集団で全員を調べるのではなく、10 人だけ標本として抜き出して母平均を推定することを考える。シミュレーションで標本抽出を 1 万回繰り返し、それぞれの標本で標本平均を計算しよう。

```
N <- 10
n_sims <- 1e4
means <- rep(NA, 1e4)
```

後で使うので、不偏分散の平方根も保存できるようにする。

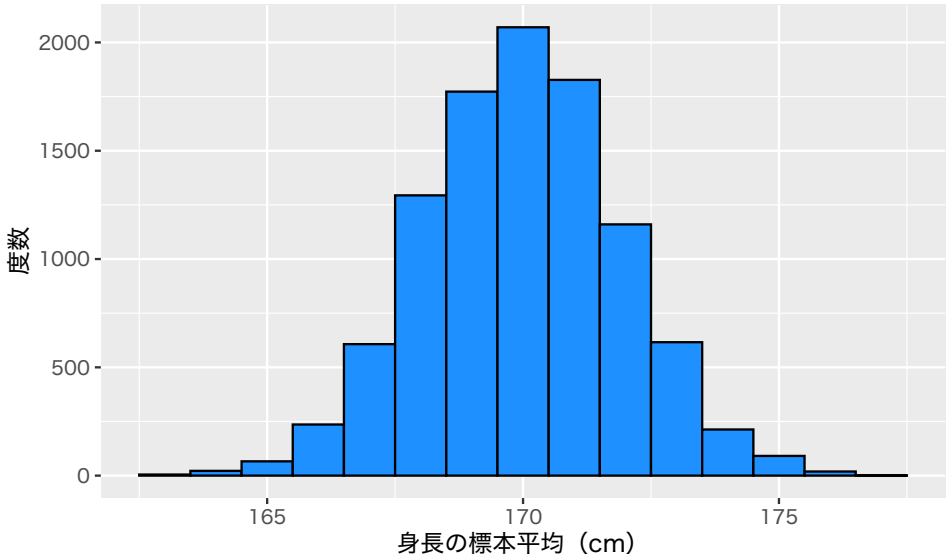
```
u <- rep(NA, 1e4)
```

母集団 `pop` から標本サイズ `N = 10` の標本を抽出する作業を `n_sims` = 10,000 回繰り返し、それぞれで標本平均と不偏分散の平方根を計算する。

```
for (i in 1 : n_sims) {
  smpl <- sample(pop, size = N, replace = FALSE)
  means[i] <- mean(smpl)
  u[i] <- sd(smpl)
}
```

標本平均の標本分布を確認する。

```r
df_sim <- tibble(mean = means,
                 sd   = u)
h1 <- ggplot(df_sim, aes(x = mean)) +
  geom_histogram(binwidth = 1,
                 color    = 'black',
                 fill     = 'dodgerblue') +
  labs(x = '身長の標本平均(cm)',
       y = '度数')
plot(h1)
```



得られた標本平均を標準化して、標準正規分布と比べてみる。

■母分散 σ^2 を知っているとき

母分散 σ^2（あるいは母標準偏差 σ）を知っているという特殊な場合について考える。このとき、標本平均 $\bar{x}$ は、以下の式で標準化(standardize)できる。

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{N}}}$$

標本平均は不偏推定量なので、

$$\mathrm{E}[\bar{x}] = \mu$$

となる。そこで、μ は、

```
(mu <- mean(means))
```

```
[1] 169.9808
```

とする。

また、仮定により σ は知っているので母集団の σ を使う。

```
(sigma <- sd(pop))
```

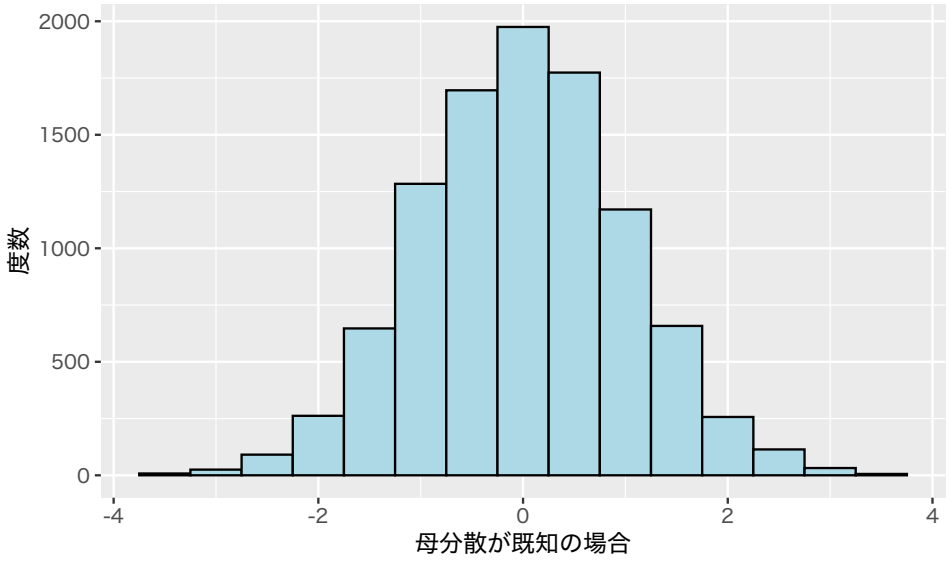
```
[1] 5.999403
```

これらの値を使うと、標準化された標本平均 z は、

```
z <- (means - mu) / (sigma / sqrt(N))
```

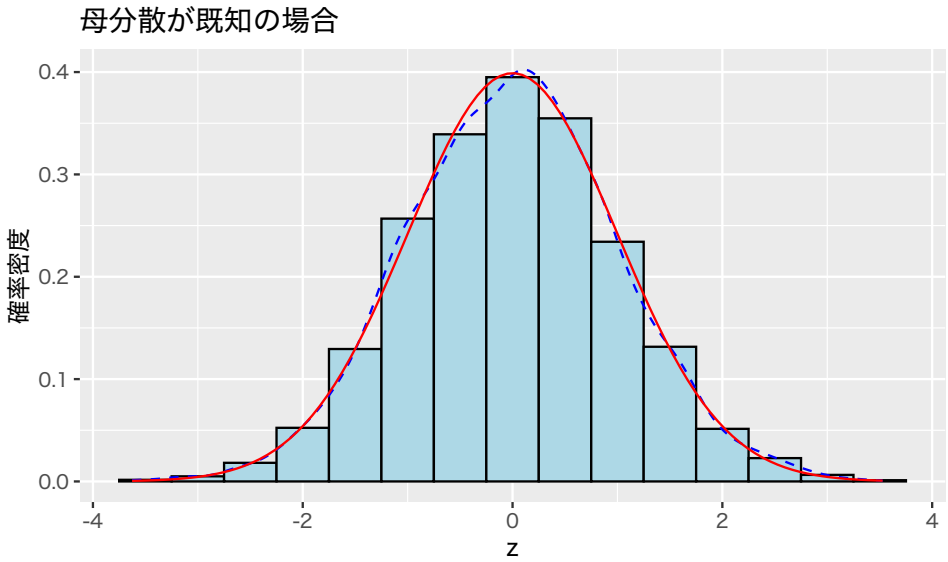
となる。分布を確認する。

```
df_sim$z <- z
h2 <- ggplot(df_sim, aes(x = z)) +
  geom_histogram(binwidth = 0.5,
                 color    = 'black',
                 fill     = 'lightblue') +
  labs(x = '母分散が既知の場合',
       y = '度数')
plot(h2)
```



geom_density() を使って z の確率密度曲線(青い点線)を描き、標準正規分布の確率密度曲線(赤い実線)と比べてみる。

```
h3 <- ggplot(df_sim) +
  geom_histogram(aes(x = z,
                     y = after_stat(density)),
                 binwidth = 0.5,
                 color    = 'black',
                 fill     = 'lightblue') +
  geom_density(aes(x = z,
                   y = after_stat(density)),
               color    = 'blue',
               linetype = 'dashed') +
  stat_function(fun         = dnorm,
                inherit.aes = FALSE,
                color       = 'red') +
  labs(y = '確率密度',
       title = '母分散が既知の場合')
plot(h3)
```



2 つの確率密度曲線はほぼ一致している。

この例のように、母分散を知っているとき、標本平均を標準化した z の分布は標準正規分布に従う。したがって、私たちは区間推定に標準正規分布を利用することができる。

■母分散 σ^2 を知らないとき

しかし、通常私たちは標本しか調べられないので母分散を知らない。

母分散を知らないとき、先程と同じ標準化はできない。なぜなら、上で使った標準化の式には、σ が出てくるが、その値を知らないからだ。そこで、母標準偏差の推定値として、不偏分散の平方根 u を使う。

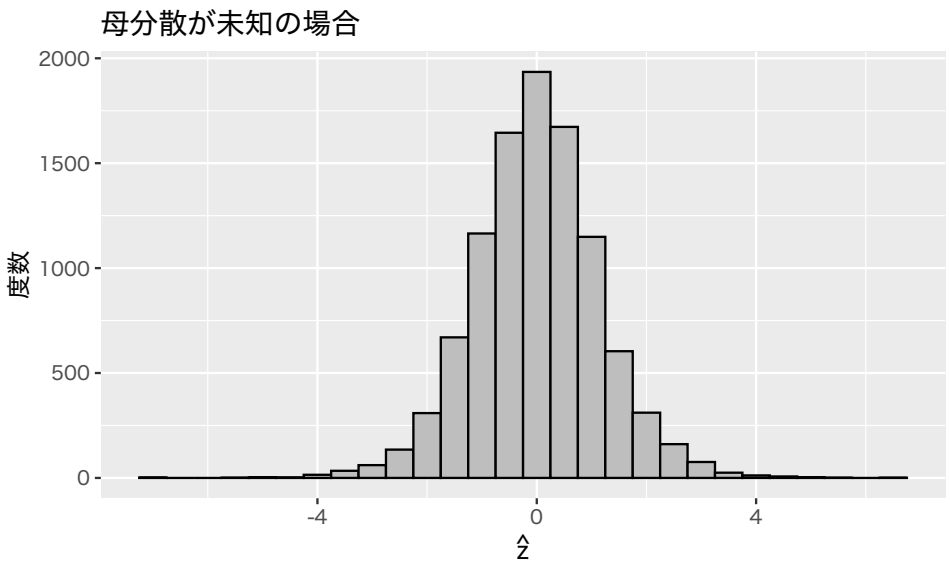
この u を使い、標本平均 $\bar{x}$ は、以下の式で標準化する。

$$\hat{z} = \frac{\bar{x} - \mu}{\frac{u}{\sqrt{N}}}$$

```
z_hat <- (means - mu) / (u / sqrt(N))
```

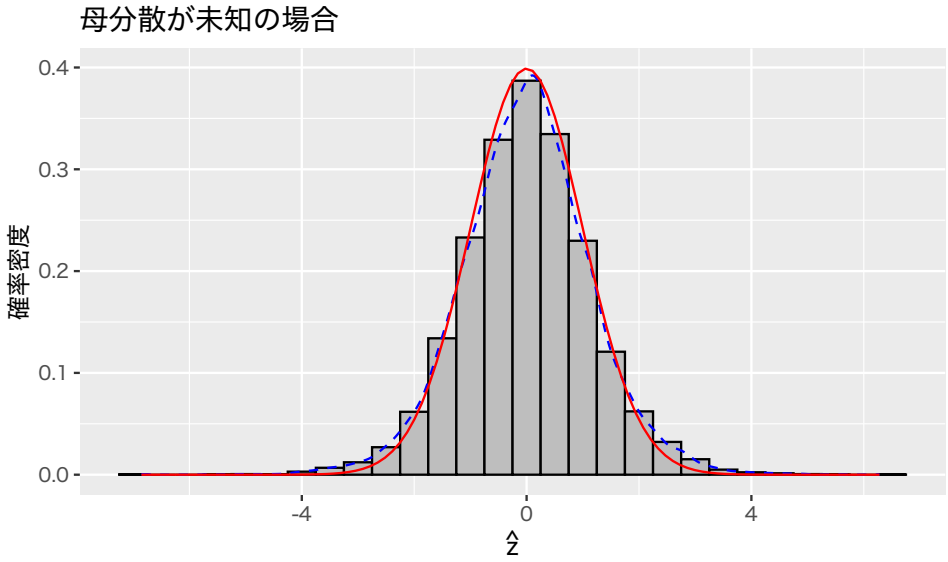
この $\hat{z}$ の分布を確認する。

```
df_sim$z_hat <- z_hat
h4 <- ggplot(df_sim, aes(x = z_hat)) +
  geom_histogram(binwidth = 0.5,
                 color    = 'black',
                 fill     = 'gray') +
  labs(x = expression(hat(z)),
       y = '度数',
       title = '母分散が未知の場合')
plot(h4)
```



標準正規分布に似ているように見えるが、どうだろうか。

```
h5 <- ggplot(df_sim) +
  geom_histogram(aes(x = z_hat,
                     y = after_stat(density)),
                 binwidth = 0.5,
                 color    = 'black',
                 fill     = 'gray') +
  geom_density(aes(x = z_hat,
                   y = after_stat(density)),
               color    = 'blue',
               linetype = 'dashed') +
  stat_function(fun         = dnorm,
                inherit.aes = FALSE,
                color       = 'red') +
  labs(x = expression(hat(z)),
       y = '確率密度',
       title = '母分散が未知の場合')
plot(h5)
```



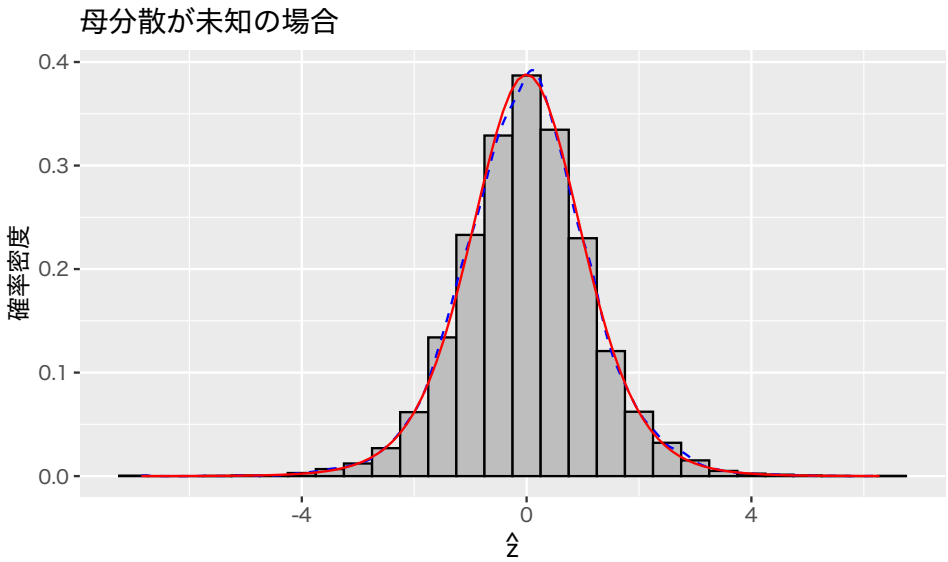
2 つの確率密度曲線は少しズレている。2 つの確率密度は、0（付近）で最大値をとるという点で同じである。しかし、分布のばらつきが違う。平均値付近を比べると、$\hat{z}$ の分布の方が確率密度が低くなっている。代わりに、分布の両裾を比べると、$\hat{z}$ の分布の方が、確率密度が高い。言い換えると、$\hat{z}$ の分布は、標準正規分布よりも裾が厚い（重い）分布になっている。

この例のように、母分散を知らないとき、標本平均を標準化した $\hat{z}$ の分布は標準正規分布に従わない。したがって、私たちは区間推定に標準正規分布を利用することができない。

実は、$\hat{z}$ の分布は自由度 $N-1$ の t 分布に従っている。試しに、（シミュレーションで使ったサンプル N は 10

なので)自由度 9 の t 分布の確率密度曲線を重ねてみる。

```
h6 <- ggplot(df_sim) +
  geom_histogram(aes(x = z_hat,
                     y = after_stat(density)),
                 binwidth = 0.5,
                 color    = 'black',
                 fill     = 'gray') +
  geom_density(aes(x = z_hat,
                   y = after_stat(density)),
               color    = 'blue',
               linetype = 'dashed') +
  stat_function(fun          = dt,
                args         = list(df = 9),
                inherit.aes = FALSE,
                color        = 'red') +
  labs(x = expression(hat(z)),
       y = '確率密度',
       title = '母分散が未知の場合')
plot(h6)
```



このように確率密度曲線がほぼ一致する。

■* 分布の比べ方

上の例では、確率密度曲線の重ね書きすることで分布を比較した。申込し厳密に分布を比べたいとき、特に 2 つの分布が同じ分布と言えるかどうか確かめたいときには、Q-Q プロット(quantile-quantile plot)という図を

使う。

この図では、分布を確かめる対象となるシミュレーションで得た変数の分位点(quantile)を縦軸に、比較対象の(基準となる)分布の分位点を横軸に取る。

分位点とは、簡単に言うと、「確率分布で下から a% に相当する値はいくつか」という値である。標準正規分布では、2.5% の分位点は -1.96、50% の分位点は 0、97.5% の分位点は 1.96 である。

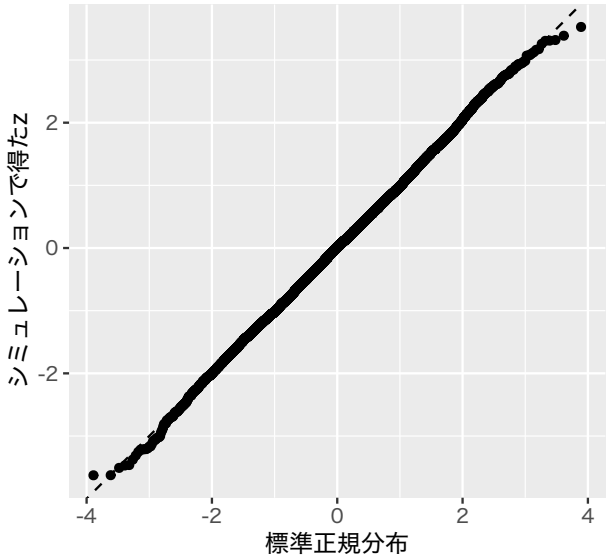
```
quantile(z, probs = c(0.025, 0.5, 0.975))
```

```
        2.5%          50%        97.5%
-1.956331145  0.009807534  1.996546090
```

2 つの分布がもし完全に一致するなら、2 つの分布の a% の分位点は、a がどんな値であっても等しいはずである。したがって、片方の分布の分位点を y、もう一方の分布の分位点を x とすれば、2 つの分布が等しいときには $x = y$ になるはず。この性質を利用し、Q-Q プロット上の点が 45 度線の上にあるかどうかを調べることで、分布を比較する。

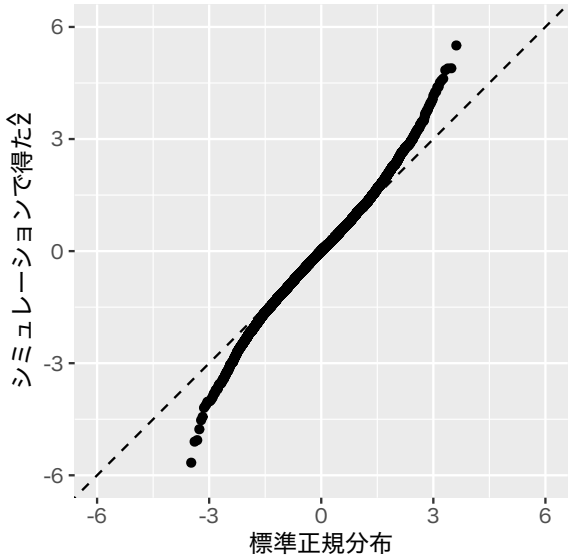
まず、z (母分散を知っているときに、標本平均を標準化したもの)と標準正規分布を比べてみる。`stat_qq()` で作れる。

```
qq1 <- ggplot(df_sim, aes(sample = z)) +
  geom_abline(intercept = 0,
              slope     = 1,
              linetype  = 'dashed') +
  stat_qq(distribution = qnorm) +
  coord_fixed(ratio = 1) +
  labs(x = '標準正規分布',
       y = 'シミュレーションで得た z')
plot(qq1)
```



次に、$\hat{z}$（母分散を知らないときに、標本平均を標準化したもの）と標準正規分布を比べる。

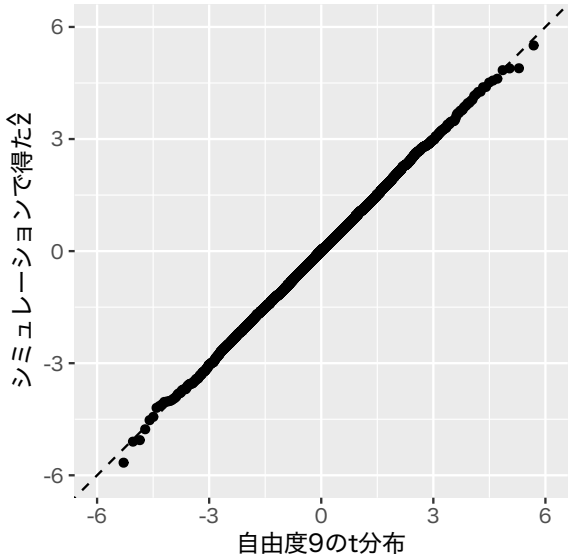
```
qq2 <- ggplot(df_sim, aes(sample = z_hat)) +
  geom_abline(intercept = 0,
              slope     = 1,
              linetype  = 'dashed') +
  stat_qq(distribution = qnorm) +
  xlim(-6, 6) +
  ylim(-6, 6) +
  coord_fixed(ratio = 1) +
  labs(x = '標準正規分布',
       y = expression(paste('シミュレーションで得た', hat(z))))
plot(qq2)
```



先ほどとは異なり、点が 45 度線から大きくズレている。ここから $\hat{z}$ は標準正規分布に従わないことがはっきりとわかる。特に裾で違う。

最後に $\hat{z}$（母分散を知らないときに標本平均を標準化したもの）と自由度 9 の t 分布（`qt(df = 9)`）を比べてみる。

```
qq3 <- ggplot(df_sim, aes(sample = z_hat)) +
  geom_abline(intercept = 0,
              slope     = 1,
              linetype  = 'dashed') +
  stat_qq(distribution = qt, dparams = 9) +
  xlim(-6, 6) +
  ylim(-6, 6) +
  coord_fixed(ratio = 1) +
  labs(x = '自由度 9 の t 分布',
       y = expression(paste('シミュレーションで得た', hat(z))))
plot(qq3)
```



やはり多少のずれはあるものの、ほとんどの点が 45 度線上に乗っている事がわかる。よって $\hat{z}$ は自由度 9 の t 分布に従っていると言えそう。

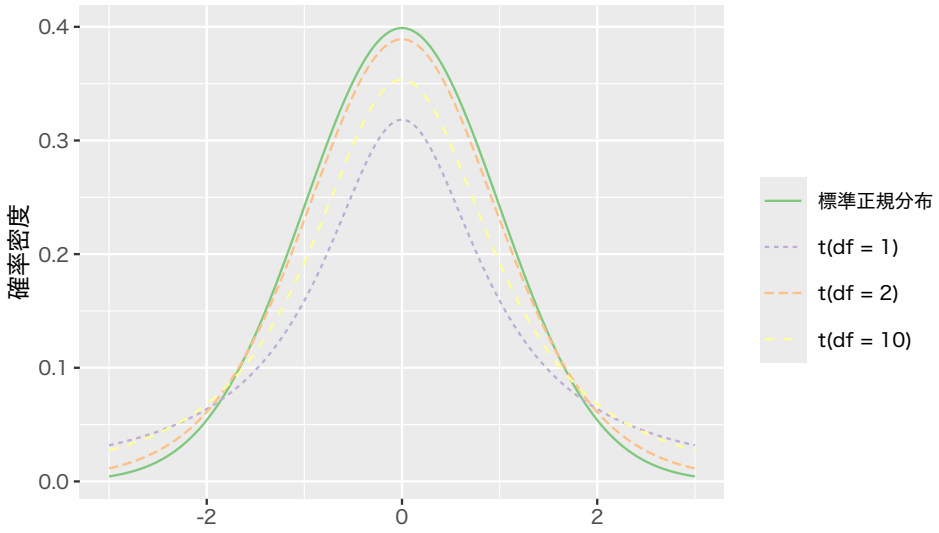
t 分布を理解する

t 分布は、自由度 $df > 0$ によってその形を変える。例えば、自由度 1, 2, 10 の t 分布を比較すると、

```
x1 <- seq(-3, 3, length = 1000)
stdn <- dnorm(x1, mean = 0, sd = 1)
t1 <- dt(x1, df = 1)
t2 <- dt(x1, df = 2)
t10 <- dt(x1, df = 10)
df_t <- tibble(x = rep(x1, 4),
               t = c(stdn, t1, t2, t10),
               group = rep(c('stdn', 't1', 't2', 't10'),
                           rep(1000, 4)))
glimpse(df_t)
```

```
Rows: 4,000
Columns: 3
$ x     <dbl> -3.000000, -2.993994, -2.987988, -2.981982, -2.975976, -2.969970~
$ t     <dbl> 0.004431848, 0.004512344, 0.004594136, 0.004677241, 0.004761679,~
$ group <chr> "stdn", "stdn", "stdn", "stdn", "stdn", "stdn", "stdn", "stdn", ~
```

```
le_labels <- c('標準正規分布', 't(df = 1)', 't(df = 2)', 't(df = 10)')
plt_t <- ggplot(df_t, aes(x = x, y = t,
                          color    = group,
                          linetype = group)) +
  geom_line() +
  scale_color_brewer(palette = 'Accent',
                     name    = '',
                     labels  = le_labels) +
  scale_linetype_discrete(name   = '',
                          labels = le_labels) +
  labs(x = '', y = '確率密度')
plot(plt_t)
```



t 分布の特徴として、

- 0 を中心として左右対称
- 標準正規分布より山の頂上が低く、裾が厚い
- 自由度が大きくなるほど標準正規分布に近づく

という点が挙げられる。

t 分布を利用した区間推定

t 分布を使った区間推定の方法も、基本的には標準正規分布を使った推定方法と同じである。身長 h の点推定値を $\bar{h}$ とすると、以下のように定義される信頼区間(confidence interval; CI)を区間推定に使う。

$$[\bar{h} - t_{N-1,p} \cdot \mathrm{SE}, \bar{h} + t_{N-1,p} \cdot \mathrm{SE}]$$

標準正規分布で Q と表していた値(`qnorm()`)を $t_{N-1,p}$（`qt()`)に変えただけ。

母集団から $N = 10$ の標本を 1 つ取り出して、身長の母平均を推定してみる。

```
smpl_1 <- sample(pop, size = 10, replace = FALSE)
```

身長 h の母平均の点推定値 $\bar{h}$ は、

```
(h_bar <- mean(smpl_1))
```

```
[1] 172.1598
```

である。また、標準誤差は、

$$\mathrm{SE} = \frac{u}{\sqrt{N}}$$

だから、

```
(se <- sd(smpl_1) / sqrt(10))
```

```
[1] 1.312271
```

である。

ここで、標本サイズ $N = 10$ だから、区間推定を行うには自由度 $N - 1 = 9$ の t 分布を利用する。95% 信頼区間を求めたいとすると、t 分布の下 2.5% 分と、上 2.5% 分を除外したい。そのために必要なのが、$t_{9,0.025}$（または、$-t_{9,0.025}$)の値である。これを `qt()` で求める。

```
qt(df = 9, p = 0.025) # qt(df = 9, p = 0.025, lower.tail = FALSE)
```

```
[1] -2.262157
```

これらの値を使うと、母平均の 95% 信頼区間を求められる。

```
(lb <- h_bar + qt(df = 9, p = 0.025) * se)
```

```
[1] 169.1913
```

```r
(ub <- h_bar + qt(df = 9, p = 0.975) * se)
```

```
[1] 175.1284
```

よって、母平均の 95% 信頼区間は、[169, 175] である。

ちなみに、標準正規分布を使って 95% 信頼区間を求めると、t 分布を使って求めた信頼区間よりも短くなる。つまり、標準正規分布を使うと、不確実性を低く見積もり、「自信過剰な」信頼区間を出してしまう。結果として、標本抽出を繰り返しても 95% 信頼区間が正解を出す確率が 95% よりも低くなってしまう。

t 分布を使って区間推定を行う関数。

```r
get_ci <- function(x, level = 0.95) {
  ## t 分布を利用して母平均の 95% 信頼区間を求める関数
  ## 引数: x = 標本(観測値のベクトル)
  ##       level = 信頼度
  N <- length(x)
  x_bar <- mean(x)
  se <- sd(x) / sqrt(N)
  t <- qt(df = N - 1, p = (1 - level) / 2, lower.tail = FALSE)
  lb <- x_bar - t * se
  ub <- x_bar + t * se
  confint <- c(lb, ub)
  names(confint) <- c(str_c(level*100, "%CI の下限値"),
                      str_c(level*100, "%CI の上限値"))
  return(confint)
}
```

この関数を使ってみる。標本 1（smpl_1)から得られる母平均の 95% 信頼区間は、

```r
get_ci(smpl_1)
```

```
95%CI の下限値 95%CI の上限値
      169.1913       175.1284
```

となり、先程と同じ結果が得られた。