統計学・データ分析勉強会／第２回宿題解答例

第 2 回宿題の解答例を以下に示します。いずれも手計算だけで答えを求めることができますが、ここでは R を利用して作成しています。

問１

【解答】

「カラスは黒い」 という仮説を厳密に立証するためには全てのカラスを対象とした **全数調査**が必要だが、これは現実的には **不可能**である。そこで、いま自分たちが集められるデータ（標本）の範囲のなかで、「カラスは黒い」という仮説が妥当と言えるかどうかを統計的仮説検定を使って検討する。

統計的仮説検定では、まず自分が主張したい仮説を覆すような仮説（帰無仮説 H_0）を考える。そして、自分の手持ちの標本を分析し、「そのような帰無仮説はありえないほど低い確率でしか成立しえない」、つまり「帰無仮説を棄却できる」、と言えるかどうかを吟味する。もし帰無仮説を棄却できれば、自分が主張したい仮説（対立仮説 H_1）が正しい（完全に間違っているとは言えない）ことが立証できる。

今回の問題において、帰無仮説を「**白いカラス**と黒いカラスが半々程度の割合で存在する」、対立仮説は「カラスは通常黒い」と置く。そのうえで、ランダムに選んだ多数のカラス 1000 羽を標本として集めたとして、白いカラスが 1 羽いたとする。この場合、白いカラスの割合は 0.1% に過ぎず、白いカラスが全体の半分いるという仮説のもとでは、極めて**まれ**な状況といえる。従ってここでは帰無仮説を棄却し、「カラスは黒いと考えて問題なさそう」と言うことができる。

.....

【補足】

統計的仮説検定でわざわざ帰無仮説を立てるのは、自分の主張したい仮説（対立仮説）を直接立証するより、帰無仮説を否定するほうが簡単な場合が多いからです。どんなに多数のカラスを見つけてきても「カラスは黒い」という主張は立証できませんが、「白いカラスが存在する確率はありえないほど低い」という結論を導くことは可能です。また確率論を極めて重要な判断基準にすることも統計学の大きな特徴です。

なお、実際の仮説検定において標本データが実現する確率（p 値と呼びます）を計算するためには、帰無仮説において「白いカラスは $\bullet\%$ 存在する」というように具体的な数字を盛り込む必要があります。また、帰無仮説を否定する際にどの程度の確率水準を「ありえないほど低い」と見なすのかについても、事前に設定しなければなりません。この水準のことを有意水準と呼び、通常は $1 \sim 5\%$ に設定します。

問２

コインの表が出る確率を p とおくと、帰無仮説 H_0 と対立仮説 H_1 は以下のように定義できます。

- $H_0 : p = \frac{1}{2}$
- $H_1 : p > \frac{1}{2}$

H_0 のもとで、コインを 10 回投げて 9 回以上表が出る確率を求めましょう。まず、コインは 1 回ごとに表か裏の 2 通りの出方があるので、10 回投げたときの出方の総数は $2^{10} = 1024$ 通りとなります。このうち、表がちょうど 9 回となるのは 10 通り（何回目に表が出るかで 10 通りのバリエーションがあるため）、表が 10 回となるのは 1 通り（全部表のケース）です。従って、表が 9 回以上となる確率（p 値）は、

$$(10 + 1)/1024 = 0.0107 = 1.07\%$$

となります。有意水準を 5% とすると帰無仮説 H_0 を棄却でき、「友人のコインはイカサマだ」と指摘できることになります。

.....

【補足】

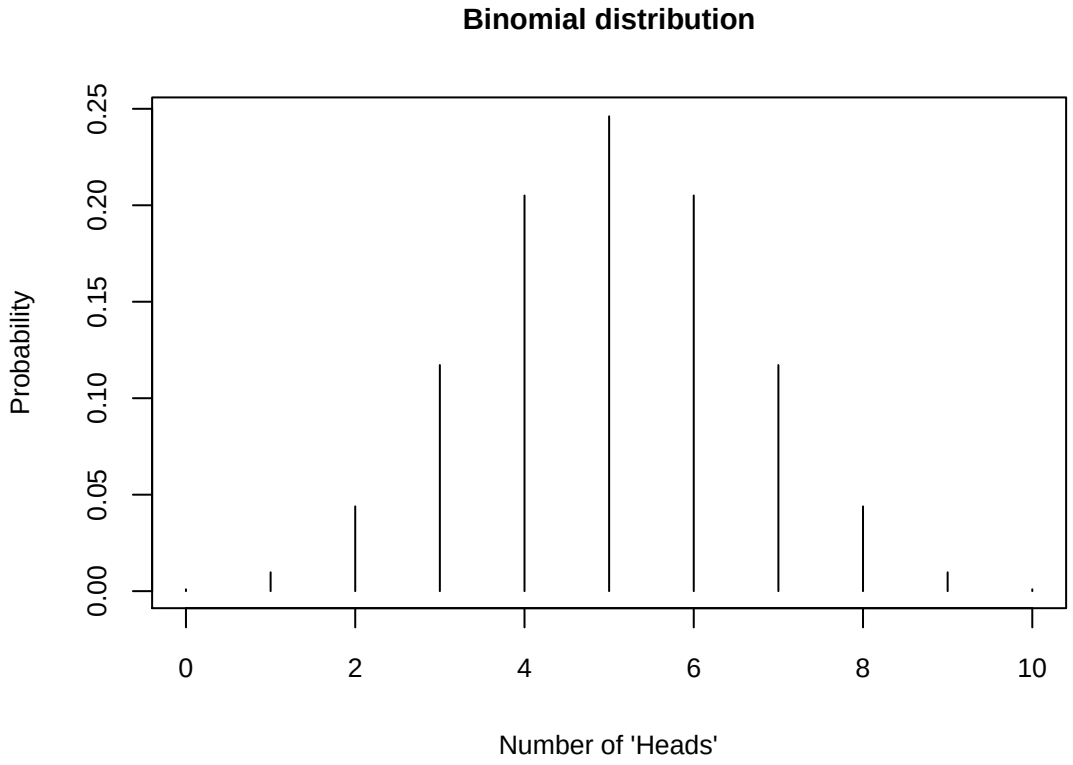
コイン投げのように、確率 p で成功する試行を n 回行ったときの成功の数（今回の問題においてはコインの表が出る数）は、「二項分布（Binomial Distribution）」という確率分布で表されます。R では、`pbinom()` 関数を使うと、二項分布において成功数が特定の値より大きくなる確率を簡単に計算できます。例えば、表が出る確率が 0.5 のコインを 10 回投げて、表が 9 回以上（つまり 8 回より多く）出る確率は以下になります。

```
pbinom(8, 10, 0.5, lower.tail=FALSE)
```

```
## [1] 0.01074219
```

先に手計算で求めた確率と一致することが確認できました。また、この場合の確率分布の全体像は以下のように表示できます。表が 9 回以上出る確率が小さいことが視覚的にも分かります。

```
plot(x=0:10, y=dbinom(0:10, 10, 0.5), type='h', #  コインを 10 回振って表が出る回数の確率分布
     main = "Binomial distribution",
     xlab='Number of \'Heads\'', ylab='Probability',
     cex.main=0.9, cex.axis=0.8, cex.lab=0.8)
```



問３

アンケート調査は、母集団（調査が関心の対象とする全体の集合）から一部を標本として取り出して平均値やばらつきなどを調べることで、**母集団について一般的に言えることを明らかにする**のが目的です。調査の際は「母集団は何か」「標本は何か」「母集団についてどのような統計量を知りたいのか」を意識するようにしましょう。

.....

まず、軽自動車を購入した 250 人に実際の購入価格を聞くアンケート調査について考えましょう。この調査では、母集団は「軽自動車を購入した全ての人（参考：業界団体の発表によると 2016 年の軽自動車の販売台数は日本全体で約 172 万台）」です。また、標本の数（n）は $n = 250$、標本の平均値（$\bar{x}$）は $\bar{x} = 1133000$、標本の標準偏差（σ）は $\sigma = 101000$ となります。

この問題で検討したいのは、上記の標本データから「軽自動車全体の平均購入価格が 110 万円超」、つまり母集団の平均値（μ）が $\mu > 1100000$ であると一般化してよいかどうかです。ここで役立つのが統計的仮説検定ですが、母集団の平均値について検討するため、特に「母平均の検定」と呼びます。まず帰無仮説 H_0 と対立仮説 H_1 を設定します。

- $H_0 : \mu = 1100000$
- $H_1 : \mu > 1100000$

H_1 の立て方から分かるように、ここでは母平均が「ある値より大きいかどうか」を調べるため「片側検定」を使います。もし H_1 が $\mu \neq 1100000$（母平均が 110 万円と言えるかどうかを調べたい）であれば「両側検定」になります。

母平均の検定は「t **検定**」を使います。まず、以下の**検定統計量（t 値）**を求めます。

$$t = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

H_0 が正しければ、t は「自由度 $n-1$ の t 分布」に従うことが理論的に知られています。$\bar{x} = 1133000$、$\mu = 1100000$、$\sigma = 101000$、$n = 250$ として t 値を計算してみると、

$$t = \frac{1133000 - 1100000}{101000 / \sqrt{250}} = 5.166$$

となります。自由度 $250 - 1 = 249$ の t 分布において、t 値が 5.166 より大きくなる確率（p 値）は、以下の `pt()` 関数で計算できます。

```
pt(5.166, 249, 0.5, lower.tail=FALSE)
```

```
## [1] 2.779066e-06
```

`2.779066e-06` は 2.779×10^{-6} の意味ですから、この確率は非常に小さいことが分かります。従って有意水準 5% のもとで帰無仮説 H_0 は棄却でき、「軽自動車の購入価格 110 万円超」という**デスクの判断は正当**だと結論できます。有意水準を 2.5% に設定しても結果は変わりません。

帰無仮説を棄却できるかどうかを調べるには、t 分布の上側確率（t 値がその値以上より大きくなる確率）が有意水準（ここでは 5%）と等しくなる t 値を求めておいて、標本データから計算した実際の t 値と比較する方法もあります。自由度 249 の t 分布で上側確率が 5% となる t 値は以下で計算できます。

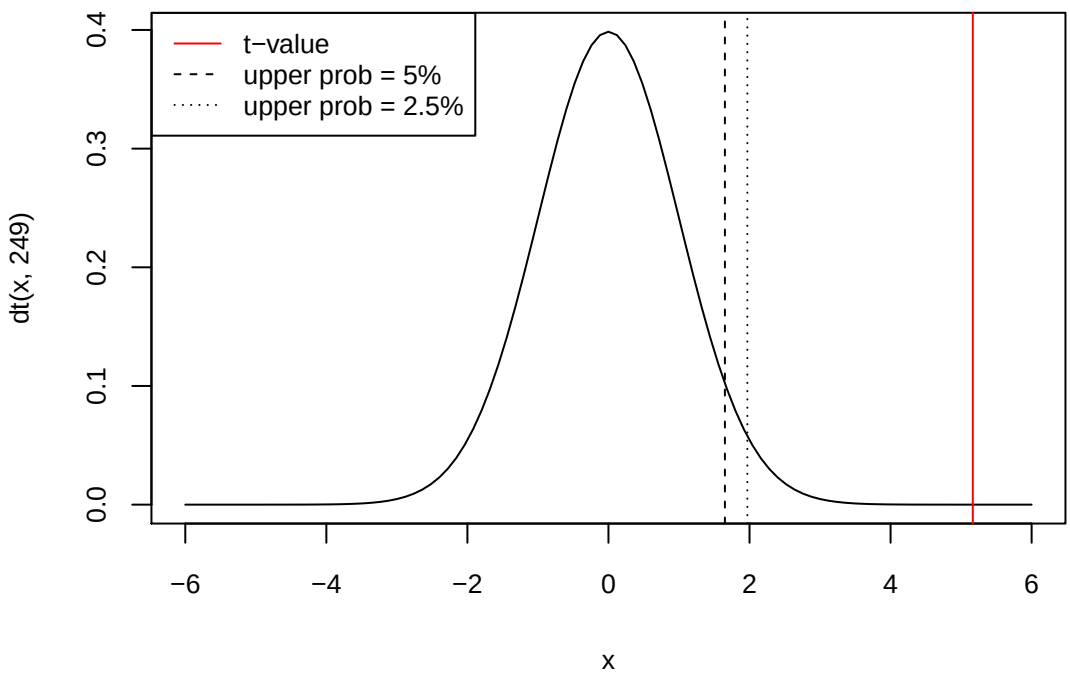
```
qt(0.05, 249, lower.tail=FALSE)
```

```
## [1] 1.650996
```

実際の t 値（5.166）は 1.650996 よりはるかに大きいですから、H_0 を棄却できます。

以上のことを視覚的に確めてみましょう。以下は自由度 249 の t 分布のグラフに、今回の標本データから計算した t 値と、上側確率が 5% および 2.5% となる t の値をそれぞれ重ねて表示したものです。

```
curve(dt(x, 249), from=-6, to=6, main="t-distribution of df=249", #  自由度 249 の t 分布
      cex.main=0.9,, cex.axis=0.8, cex.lab=0.8)
abline(v=5.166, col='red') # 標本から計算した t 値
abline(v=qt(0.05, 249, lower.tail=FALSE), lt=2)  # 上側確率 5%
abline(v=qt(0.025, 249, lower.tail=FALSE), lt=3)  # 上側確率 2.5%
label <- c("t-value", "upper prob = 5%", "upper prob = 2.5%")  # 以下は凡例の作成用
color <- c("red", "black", "black")
lt <- c(1, 2, 3)
legend("topleft", legend=label, col=color, lty=lt, cex=0.8)
```



もう 1 つ補足です。実は、t 分布は自由度（つまり標本の数）が十分に大きい（だいたい 100 以上）ときには、標準正規分布（平均 0、標準偏差 1 の正規分布）とほぼ同一の分布になります。この問題では自由度が 249 なので、t 分布の代わりに標準正規分布を使うことができます。試しに、標準正規分布で上側確率が 5% となる点を計算してみましょう。

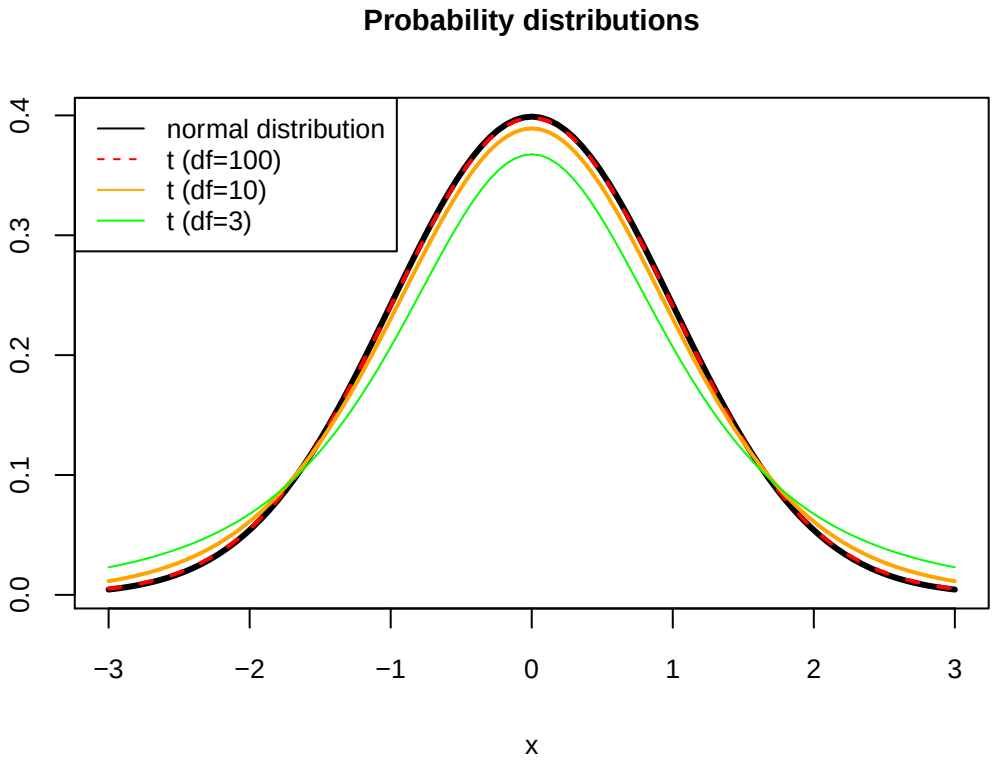
```
qnorm(0.05, lower.tail=FALSE)
```

```
## [1] 1.644854
```

このように t 分布を使って求めたのと非常に近い値が得られました。標準正規分布を使っても H_0 を棄却できることになります。

正規分布と t 分布の違いを見るため、標準正規分布および自由度の異なるいくつかの t 分布についてグラフを描いてみましょう。

```
curve(dnorm(x, mean=0, sd=1), from=-3, to=3, # 標準正規分布
      main="Probability distributions",
      ylab="",  lwd=3, cex.main=0.9,, cex.axis=0.8, cex.lab=0.8)
curve(dt(x, 100), col="red", lt=2, lwd=2, add=TRUE)  # 自由度 100 の t 分布
curve(dt(x, 10), col="orange", lwd=2, add=TRUE)  # 自由度 10 の t 分布
curve(dt(x, 3), col="green", add=TRUE)  # 自由度 3 の t 分布
label <- c("normal distribution", "t (df=100)", "t (df=10)", "t (df=3)")  # 以下は凡例の作成用
color <- c("black", "red", "orange", "green")
lt <- c(1, 2, 1, 1)
legend("topleft", legend=label, col=color, lty=lt, cex=0.8)
```



同じように、普通自動車を購入した 250 人へのアンケート調査について考えましょう。今回の母集団は「普通自動車を購入した全ての人（参考：2016 年の国内軽自動車の販売台数は日本全体で約 324 万台）」です。また、標本の数（n）は $n = 250$、標本の平均値（$\bar{x}$）は $\bar{x} = 3120000$、標本の標準偏差（σ）は $\sigma = 2050000$ となります。考えたいのは「普通乗用車の平均購入価格が 300 万円超」という主張の正当性ですので、帰無仮説（H_0）と対立仮説（H_1）はそれぞれ以下のようになります。

- $H_0 : \mu = 3000000$
- $H_1 : \mu > 3000000$

標本のデータをもとに t 値を計算します。

$$t = \frac{3120000 - 3000000}{205000/\sqrt{250}} = 0.9416$$

H_0 のもとで t 値は自由度 249（$= 250 - 1$）の t 分布に従います。t 値が 0.9416 より大きくなる確率（p 値）を計算しましょう。

```
pt(0.9416, 249, 0.5, lower.tail=FALSE)
```

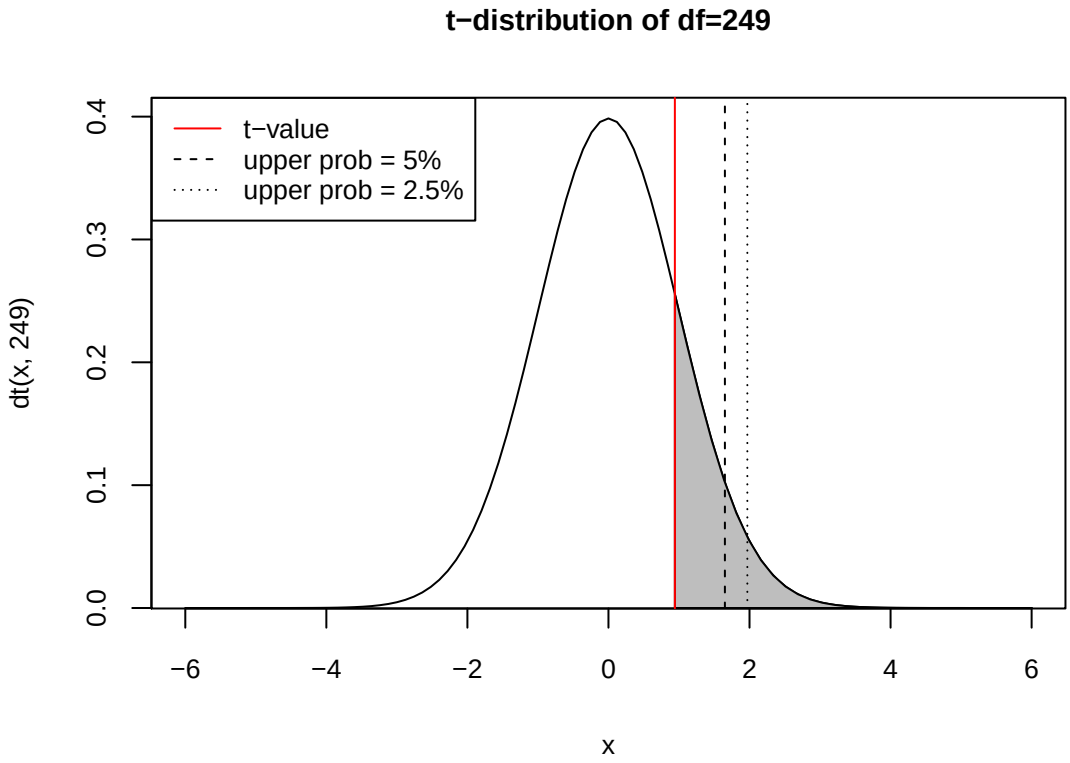
```
## [1] 0.3298732
```

今回は有意水準 5% で H_0 を棄却できない、つまり**データの判断は正しいとは言えない**ことが分かります。これは、0.9416 という値が、上側確率が 5% となる t 値（上で求めたように 1.650996）より小さいことからも結論できます。

軽自動車のケースと同様に、視覚的にも確認しておきましょう。以下のプロットで、グレーで塗りつぶした部分の面積が今回の p 値（0.3298732）に対応しています。

```
curve(dt(x, 249), from=-6, to=6, main="t-distribution of df=249", #  自由度 249 の t 分布
      ylim=c(0.015, 0.4), cex.main=0.9,, cex.axis=0.8, cex.lab=0.8)
xvals <- seq(0.9416, 6, length=30)  # 塗りつぶし用
dvals <- dt(xvals, 249)  # 塗りつぶし用
polygon(c(xvals,rev(xvals)), c(rep(0,30), rev(dvals)), col="gray") # 塗りつぶし用
abline(v=0.9416, col='red') # 標本から計算した t 値
abline(v=qt(0.05, 249, lower.tail=FALSE), lt=2)  # 上側確率 5%
```

```r
abline(v=qt(0.025, 249, lower.tail=FALSE), lt=3)  # 上側確率 2.5%
label <- c("t-value", "upper prob = 5%", "upper prob = 2.5%")  # 以下は凡例の作成用
color <- c("red", "black", "black") # 線の色
lt <- c(1, 2, 3) # 線の種類
legend("topleft", legend=label, col=color, lty=lt, cex=0.8)
```



軽自動車のアンケート調査で普通乗用車のアンケート調査で、異なる結論が得られた理由を考えてみましょう。2 つの調査結果で最も大きく違うのが標本データの標準偏差です。普通乗用車の調査はデータのばらつきが大きく、一般的な結論を導く際には注意を要するということを表しています。

問4

この問題でもやりたいことは「母平均の検定」なので、t 検定を使います。帰無仮説（H_0）と対立仮説（H_1）は次のようになります。友人は日本人の身長が「170.7cm より大きい」と主張しているので、片側検定を使います。

- $H_0 : \mu = 170.7$
- $H_1 : \mu > 170.7$

このあとの手順は基本的に問 3 と同じ（標本平均と標本標準偏差を求めて t 値を計算し、p 値を求める）ですが、ここでは R の便利な関数を使いましょう。まず、友人の身長リストを `yujin` という名前で変数（ベクトル）に格納します。

```r
yujin <- c(188, 164, 177, 190, 185, 160, 171, 188, 192, 164)
```

標本の平均値と標準偏差を確認しておきます。

```r
mean(yujin)
```

```
## [1] 177.9
```

```r
sd(yujin)
```

```
## [1] 12.26966
```

平均値が 177.9cm 以外なので、友人の主張ももっともに聞こえますが、検定でチェックしてみましょう。

問３と同様に平均値や標準偏差をもとに t 値を計算することもできますが、今回はデータセットが手元にあるので、`t.test()` 関数を使って簡単に t 検定の結果を得ることができます。

```
t.test(yujin, mu=170.7) #  母平均が 170.7 かどうかを検定したい
```

```
##
##  One Sample t-test
##
## data:  yujin
## t = 1.8557, df = 9, p-value = 0.09648
## alternative hypothesis: true mean is not equal to 170.7
## 95 percent confidence interval:
##  169.1228 186.6772
## sample estimates:
## mean of x
##     177.9
```

p 値が 0.09648 であり、有意水準 5% で H_0 を棄却できません。よって、友人の主張に反駁することができました。

【補足】

この解答例では詳しく触れませんでしたが、実は t 検定を利用するには、「母集団が正規分布であると仮定してよい」ことが前提条件として必要になります。母集団が正規分布と見なせない場合の検定については発展的なテーマとなりますので、基礎的な検定を一通り理解してから学習するのがよいでしょう。

また、統計的仮説検定には関心の対象となるデータの種類やサイズによって χ^2 検定や F 検定など様々な方法があり、それらを正しく使い分けることが重要になります。