Modern Inference (Statistical Learning) - University of Milano-Bicocca

# Homework 1 - Solution

# 1 Darwin's plants experiment

## 1.1 Galton's model

**TASK 1**:

1. **Analysis Plan:**

   *Answer*: If we assume that $\epsilon_i$ and $\varepsilon_i$ have a $N(0,1)$ distribution, we can use a *two-sample t test* (t = 2.4371, df = 28, p-value = 0.02141).

2. **Code:**

   *Answer*: `t.test(height ~ type, data=darwin, var.equal=TRUE)`

3. **Claim:**

   *Answer*:

   From the two-sample t test we obtain a point estimate of $\hat{\theta} = 20.19 - 17.57 = 2.62$ along with a 95% confidence interval $(0.417, 4.816)$ not including 0. The answer to G1 is there is a significant height increase - estimated on average 2.62 eighths of an inch for cross-fertilized plants - and we reject $H_0 : \theta = 0$ with $p$-value = 0.021 (two-sided alternative). The answer to G2 is that the average height increase is between $(0.417, 4.816)$ with 95% confidence.

## 1.2 Fisher's model

**TASK 2**:

1. **Analysis Plan:**

   *Answer*: If we assume that $\epsilon_i$ and $\varepsilon_i$ have a $N(0,1)$ distribution, we can use a *paired t test*, or one-sample t test for the difference (t = 2.148, df = 14, p-value = 0.0497).

2. **Code:**

   *Answer*: `differences = apply(darwin_pair[,3:2],1,diff); t.test(differences)`

3. **Claim:**

   *Answer*:

   From the one-sample t test we obtain a point estimate of $\hat{\theta} = 2.62$ (exactly as before) along with a 95% confidence interval $(0.004, 5.24)$. The answer to G1 is that although we obtain the same estimated height increase of 2.62 for cross-fertilized plants, there more uncertainty than before in concluding that it is significantly different from zero ($p$-value = 0.0497, two-sided alternative). The answer to G2 is that the average height increase is between $(0.004, 5.24)$ with 95% confidence. A long interval close to zero is not surprising given the small sample size ($n = 15$).

## 1.3 A more general model

**TASK 3**:

1. **Analysis Plan:**

   *Answer*: If we assume that $\epsilon_i$ and $\varepsilon_i$ are independent and symmetrically distributed around 0, then $D_i$ is symmetrically distributed around $\theta$ and we can test $H_0 : \theta = 0$ by a nonparametric or distribution-free test, the *sign test* for the difference (number of successes = 13, number of trials = 15, p-value = 0.007385 [two-sided], where success means that in a given pair the height of cross-fertilized > self-fertilized, and trials are the number of pairs).

2. **Code:**

   *Answer*: `binom.test(x=13, n=15, p=0.5, alternative="two.sided")`

3. **Claim:**

   *Answer*:

   The sign test provides significant evidence that $\theta \neq 0$ ($p$-value = 0.007385, two-sided alternative) with little assumptions. The probability for cross-fertilized plant of being higher than its self-fertilized counterpart is estimated 86% with a 95% confidence interval of $(0.59, 0.98)$.

4. Let $X \sim f(x)$ and $Y \sim g(y)$ be continuously distributed random variables with density functions $f$ and $g$, and assume that $f$ is symmetric around 0, $g$ is symmetric around 0, and $X$ and $Y$ are independent. Prove that $Z = X - Y$ is symmetric around 0

   *Answer*: Note that $Y$ is equal in distribution to $-Y$, so we need to prove that $X + Y$ is symmetric around 0. See https://stats.stackexchange.com/questions/95002/the-sum-of-two-symmetric-random-variables-is-symmetric for the proof.

## 2 Galieo's inclined plane experiment (1604)

**TASK 4**:

1. **Analysis Plan** Assume a Gaussian linear model $y = X\beta + \varepsilon$, where $y$ represents the `distance` and $X$ is the design matrix including the `intercept`, `time` and `time`$^2$. This represents Galileo model. Aristotle's model is a particular case with $\beta_2 = 0$. Then we can use the $F$ test for comparing nested models.

2. **Code**

```
experiment = data.frame(distance=c(33,130,298,526,824,1192,1620,2104), time=1:8)
Galileo = lm(distance ~ 1 + time + I(time^2), experiment)
Aristotle = lm(distance ~ 1 + time, experiment)
anova(Aristotle, Galileo)
```

```
Analysis of Variance Table

Model 1: distance ~ 1 + time
Model 2: distance ~ 1 + time + I(time^2)
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1      6 179415
2      5     75  1    179340 11917 1.223e-09 ***
```

3. **Claim**

   We reject Aristotle model in favour of Galileo model.