STAT 581 - HW #1

Alex Towell (atowell@siue.edu)

```
library("readxl")

# read from directory the markdown file is located in
h1.data <- read_excel("./handout1data.xlsx")
h1.data$method <- as.factor(h1.data$method)

h1.time.dry <- na.omit(h1.data$time[h1.data$method=='d'])
h1.time.wet <- na.omit(h1.data$time[h1.data$method!='d'])
```

An experiment is designed to investigate whether the time to drill holes in rock holes using wet or dry drilling. A completely randomized design (CRD) is used. Each method is replicated on 12 rocks. The drilling times (in 1 / 100 minutes) are observed to be

dry	727	965	904	987	847	918	814	750	804	989	902	939
wet	607	549	762	665	588	798	704	772	780	599	603	699

Q1

State the hypotheses of interest. Provide an interpretation, stated in the context of the problem.

The hypothesis of interest is whether drilling time is effected by the choice of wet or dry drilling. We may formulate this as a hypothesis test of the form

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2 \,,$$

where μ_1 and μ_2 are respectively the expected drilling time for dry and wet drilling.

If H_0 is true, the drilling method has *no* effect on drilling time. If H_A is true, the choice does have an effect on drilling time.

Q2

Give two advantages of the completely randomized design.

Two advantages of the completely randomized design (CRD) are:

1. It is simple to implement and analyze.
2. It accounts for experimental unit variance, namely by controlling for the type I error,

$$\alpha = \Pr(\text{decide } H_A | H_0 \text{ true}),$$

where α is the probability of a type I error. It may also be thought of as the type I *error rate*, or *false positive rate*.

Q3

Give two disadvantages of the completely randomized design.

Two disadvantages of the completely randomized design (CRD) are:

1. It does not account for type II error,

$$\beta = \Pr(\text{decide } H_0 | H_A \text{ true}).$$

where β is the probability of a type II error.

Additional remarks: Since we may denote a type I error a false positive, it may be appropriate to denote a type II error a false negative, which implies the convention $H_0 \equiv$ negative and $H_A \equiv$ positive. Thus, β may be thought of as the false negative rate.

Consider the unrelated *Bloom filter*, which is a probabilistic data structure that models sets in which membership queries have a controllable false positive rate ϵ and a false negative rate 0. If we take the complement of an object representing such a Bloom filter, then membership queries on its complement have a false positive rate 0 and a false negative rate ϵ.

2. It does not adjust for differences in experimental units nor differences in any other factors.

Q4

Compute the sample means, the sample standard deviations, and the pooled sample standard deviation.

```
time.mean <- round(c(mean(h1.time.dry),
                     mean(h1.time.wet)),digits=3)
time.var <- round(c(var(h1.time.dry),
                    var(h1.time.wet)),digits=3)
time.sd <- round(sqrt(time.var),digits=3)
n <- c(length(h1.time.dry),
       length(h1.time.wet))
time.sd.pooled <-
  round(sqrt(((n[1] - 1) * time.var[1] + (n[2] - 1) * time.var[2]) / (sum(n)-2)),
        digits=3)

c("y.bar.1" = time.mean[1],
  "y.bar.2" = time.mean[2],
  "s.1" = time.sd[1],
  "s.2" = time.sd[2],
  "s.p" = time.sd.pooled)
```

```
## y.bar.1 y.bar.2     s.1     s.2     s.p
## 878.833 677.167  89.470  87.189  88.337
```

We see that the sample means are

$$(\bar{y}_1, \bar{y}_2) = (878.833, 677.167),$$

the standard deviations are

$$(s_1, s_2) = (89.47, 87.189),$$

and the pooled standard deviation is

$$s_p = 88.337.$$

Q5

Compute the t_0 statistic, the critical point for a level $\alpha = .05$ test, and the p-value.

The test statistic is defined as

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

which is a drawn from a t-student distribution with $n_1 + n_2 - 2$ degrees of freedom under the null model (reference distribution).

> Under the null model, $\bar{y}_1$ and $\bar{y}_2$ are the means of two samples respectively of size n_1 and n_2 whose elements are drawn from the same normal distribution.

Denote this reference distribution by T. Under the null model, t_0 is a realization of T. Our general approach is to compute the test statistic t_0 for the given samples and say it is compatible with the null model if it obtains a value that is not too extreme under T, e.g., a p-value that is less than some controllable parameter α.

The critical point (the minimum value that is considered to be too extreme) is given by solving for t^* in

$$\Pr\{-|t^*| < T < |t^*|\} = \alpha,$$

which has the solution $t^* = t_{\mathrm{df}=n_1+n_2-2}(\alpha/2)$, and

$$\begin{aligned} p\text{-value} &= \Pr\{T \notin [-|t_0|, |t_0|]\} \\ &= 2\Pr\{T > |t_0|\} \\ &= 2(1 - F_T(|t_0|)), \end{aligned}$$

where F_T is the cdf of T.

We compute these results in R with:

```
se <- time.sd.pooled*sqrt(1/n[1]+1/n[2])
t.0 <- round((time.mean[1]-time.mean[2])/se,digits=3)
alpha <- 0.05
t.star = round(qt(alpha/2,lower.tail=F, df=n[1]+n[2]-2),digits=3)
p.value = 2*pt(abs(t.0),df=n[1]+n[2]-2,lower.tail=F)

c("t.0"=t.0,"t.*"=t.star,"p.value"=p.value)
```

```
##          t.0          t.*      p.value
## 5.592000e+00 2.074000e+00 1.272907e-05
```

We see that $t_0 = 5.592$, the critical point t^* for $\alpha = 0.05$ is 2.074, and the p-value is less than .001.

Q6

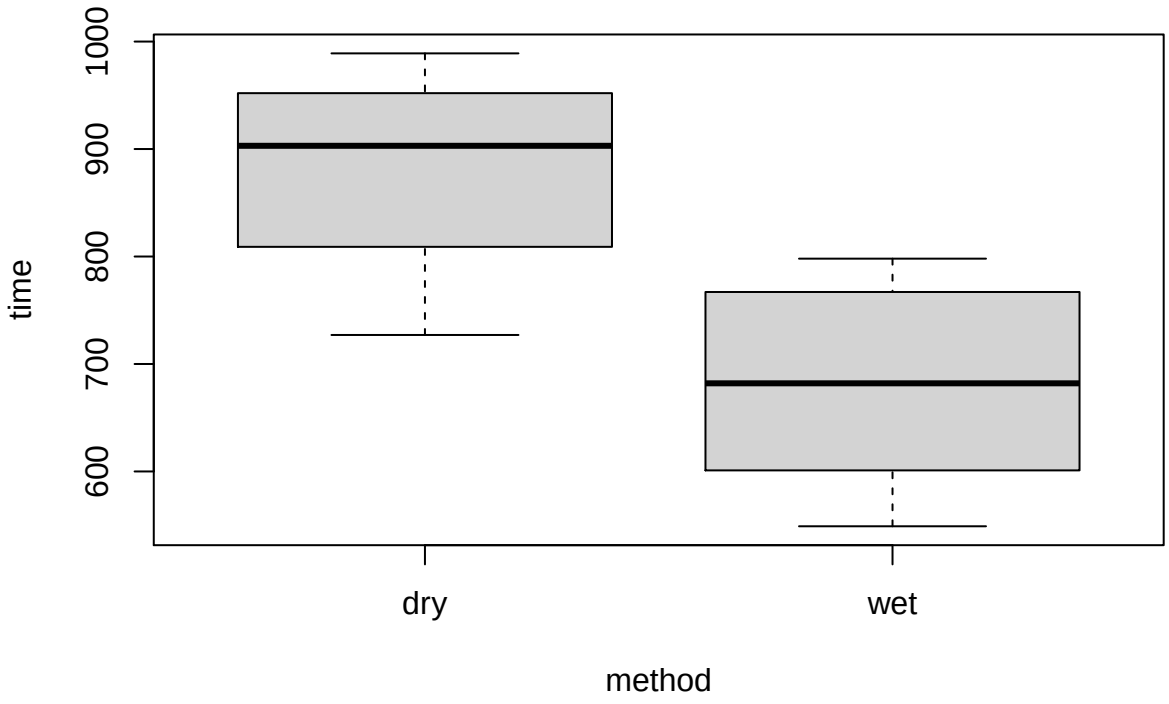
Provide an interpretation, stated in the context of the problem.

The experiment finds that the wet drilling method leads to reduced drilling times compared to the dry drilling method.

Q7

Create a boxplot as a graphical display of the data.

```r
method <- na.omit(h1.data$method)
times <- na.omit(h1.data$time)
boxplot(times~method,names=c("dry","wet"),xlab="method",ylab="time")
```



Q8

Refering to the boxplot, is it true that all dry drilling times exceed all wet drilling times? In what sense can the experiment find that dry drilling takes longer than wet drilling?

No, not all dry drilling times exceed all wet drilling times.

However, the boxplot does show that the spread of the drilling times area approximately the same between the two methods (providing justification for assuming the two samples have the same variance) and the mean drilling times are significantly different. As a result, most of the time, the wet drilling time is going to be less than the dry drilling time.

For instance, we see that the mean wet drilling time is less than the best dry drilling time and we also see that the 25-th percentile of the dry method's times is a bit larger than the wet method's longest drilling time. Most tellingly, we see that around 67% of the time, the wet drilling time is less than the smallest dry drilling time.