9.6 Two Sample t-test

We will look at an example comparing means of two Normal Distributions. We will explore some of the properties, investigate power, and consider whether this is a LRT.

Example: Cloud Seeding (Example 8.3.1 & 9.6.1)

Comparing log rainfall over the two groups

```
favstats(lograin~group,data=clouds)
```

```
##      group      min       Q1   median       Q3      max     mean       sd
## 1   Seeded 1.410987 4.581480 5.396406 6.000699 7.917755 5.134187 1.599514
## 2 Unseeded 0.000000 3.211421 3.786259 5.069278 7.092241 3.990406 1.641847
##    n missing
## 1 26       0
## 2 26       0
```

Is there evidence of greater rainfall in the seeded group compared to the unseeded group?

Is there a significant difference in the average amount of rainfall?

- hypothesis test/p-value
- Confidence Intervals
- Interpretations/Conclusions (Frequentist perspective)
- Assumptions

Set-up

$X = (X_1, ..., X_m) \sim N(\mu_1, \sigma^2); Y = (Y_1, ..., Y_n) \sim N(\mu_2, \sigma^2)$

$H_0 : \mu_1 \leq \mu_2; H_1 : \mu_1 > \mu_2$

Note: σ^2 unknow and also the same in each group (equal variance assumption)

$$\text{test statistic} = \frac{\text{estimate - hypothesized value}}{\text{standard error of the estimate}}$$

$$U = \frac{(\bar{X}_m - \bar{Y}_n) - 0}{SE_{(\bar{X}_m - \bar{Y}_n)}}$$

Standard Error

$$
\begin{aligned}
Var(\bar{X}_m - \bar{Y}_n) &= Var(\bar{X}_m) + Var(\bar{Y}_n) \\
&= \frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n} \\
sd(\bar{X}_m - \bar{Y}_n) &= \sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}
\end{aligned}
$$

If $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ as we've assumed, then our best estimate of σ^2 is

$$s_p^2 = \frac{(m-1)s_x^2 + (m-1)s_y^2}{m+n-2} = \frac{\sum(X_i - \bar{X}_m)^2 + \sum(Y_i - \bar{Y}_n)^2}{(m-1)+(n-1)}$$

So $SE(\bar{X}_m - \bar{Y}_n) = \sqrt{\frac{s_p^2}{m} + \frac{s_p^2}{n}} = s_p\sqrt{\frac{1}{m} + \frac{1}{n}}$

Two Sample Test Statistic

$$U = \frac{(\bar{X}_m - \bar{Y}_n)}{s_p\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

Claim: U follows a $t_{(m+n-2)}$ distribution when H_0 is true $(\mu_1 = \mu_2)$

$$U = \frac{\frac{(\bar{X}_m - \bar{Y}_n)}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}}}{\sqrt{s_p^2/\sigma^2}}$$

Numerator is $Z \sim N(0,1)$ as desired. Denominator can be written as:

$$\sqrt{\frac{1}{\sigma^2}\frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2}} = \sqrt{\frac{\frac{(m-1)s_x^2}{\sigma^2} + \frac{(n-1)s_y^2}{\sigma^2}}{(m+n-2)}}$$

Distribution of U

$\frac{(m-1)s_x^2}{\sigma^2} \sim \chi^2_{(m-1)}$; $\frac{(n-1)s_y^2}{\sigma^2} \sim \chi^2_{(n-1)}$

So

$$U = \frac{Z}{\sqrt{\chi^2_{(m+n-2)}/(m+n-2)}} \sim t_{(m+n-2)}$$

Further

- Reject H_0 if $U > T^{-1}_{(m+n-2)}(1-\alpha_0)$.
- pvalue $= Pr(U > T_{(m+n-2)} | \mu_1 = \mu_2)$
- $(1-\alpha_0)$CI for $\mu_1 - \mu_2$

$$\bar{X}_m - \bar{Y}_n \pm T^{-1}_{(m+n-2)}(1-\alpha_0/2)s_p\sqrt{\frac{1}{m}+\frac{1}{n}}$$

Using R

```r
m <- 26
n <- 26
xbar <- mean(clouds$lograin[clouds$group=="Seeded"])
sxsq <- (m-1)*var(clouds$lograin[clouds$group=="Seeded"])
ybar <- mean(clouds$lograin[clouds$group=="Unseeded"])
sysq <- (n-1)*var(clouds$lograin[clouds$group=="Unseeded"])
sp2<- (sxsq+sysq)/(m+n-2)
se <- sqrt(sp2)*sqrt(1/n+1/m)
U=(xbar-ybar)/se;U
```

```
## [1] 2.544369
```

```r
t.test(lograin~group,var.equal=TRUE,data=clouds)
```

```
##
##  Two Sample t-test
##
## data:  lograin by group
## t = 2.5444, df = 50, p-value = 0.01408
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.240865 2.046697
## sample estimates:
##   mean in group Seeded mean in group Unseeded
##               5.134187               3.990406
```

What is the power function like?

```r
qt(0.95,df=50)
```

```
## [1] 1.675905
```

Reject H_0 if U>1.67. (We observed U=2.54 so we reject H_0.)

$$
\begin{aligned}
\pi(\mu_1, \mu_2, \sigma^2|\delta) &= Pr(\text{Reject} H_0|\mu_1, \mu_2) \\
&= Pr(U > c|\mu 1, \mu 2) \\
&= Pr\left(\frac{(\bar{X}_m - \bar{Y}_n) - 0}{s_p\sqrt{1/m + 1/n}} > c|\mu_1, \mu_2\right) \\
&= T_{(m_+ n-2)(c|\psi)}
\end{aligned}
$$

Where $\psi = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma^2}{m} + \frac{\sigma^2}{n}}}$

Consider power when $\mu 1 - \mu 2 = \Delta$; or a function of $\sigma : \mu_1 = \mu_2 - \sigma$

```r
psi=sqrt(1/m+1/n)
1-pt(1.67,df=50,ncp=psi)
```

```
## [1] 0.0860369
```

Two-sample t-test with equal variance

- The two sample t-test is fairly robust to violations of normality and equal variances.
- But it's not so good when n and m are small and there is a hight degree of skewness
- The Two-sample t-test is a likelihood ratio test when the variances are equal (p592-3).

What happens if we relax the assumption of equal variance?

$X = (X_1, ..., X_m) \sim N(\mu_1, \sigma_1^2); Y = (Y_1, ..., Y_n) \sim N(\mu_2, \sigma_2^2)$

- If we the variance of one group is a known ratio of the variance of the other group ($\sigma_2^2 = k\sigma_1^2$) then U has the same t-distribution.
- If the variances are unequal the Likelihood Ratio Test Statistic has no known distribution. This is known as the *Behrens-Fisher problem*.

What do we do?

When variances are not equal (and not a known ratio of each other) we can

1. Always use the Welch test which uses approximate distributions.
2. Conduct a hypothesis test

$$H_0 : \sigma_1^2 = \sigma_2^2; H_1 : \sigma_1^2 \neq \sigma_2^2$$

But for this test we need an F-distribution.

9.7 F Distributions

If $Y \sim \chi^2_m$ and $W \sim \chi^2_n$ where Y and W are independent then X follows an F distribution with m and n degrees of freedom. (Def 9.7.1)

$$X = \frac{Y/m}{W/n} \sim F(m, n)$$

Details

$$X = (X_1, ..., X_m) \sim N(\mu_1, \sigma_1^2); Y = (Y_1, ..., Y_n) \sim N(\mu_2, \sigma_2^2)$$

$$\frac{(m-1)s_x^2}{\sigma_1^2} = \frac{\sum(X_i - \bar{X}_m)^2}{\sigma_1^2} \sim \chi^2_{(m-1)}$$

Similarly, $\frac{(n-1)s_y^2}{\sigma_2^2} \sim \chi^2_{(n-1)}$

So

$$V = \frac{\frac{\sum(X_i - \bar{X}_m)^2}{\sigma_1^2}/(m-1)}{\frac{\sum(Y_i - \bar{Y}_n)^2}{\sigma_2^2}/(n-1)} \sim F_{(m-1),(n-1)}$$

Cloud Seeding Example

Under the Null: $\sigma_1^2 = \sigma_2^2$

So

$$V = \frac{\frac{\sum (X_i - \bar{X}_m)^2}{\sigma_1^2}/(m-1)}{\frac{\sum (Y_i - \bar{Y}_n)^2}{\sigma_2^2}/(n-1)} = \frac{s_x^2}{s_y^2} \sim F_{(m-1),(n-1)}$$

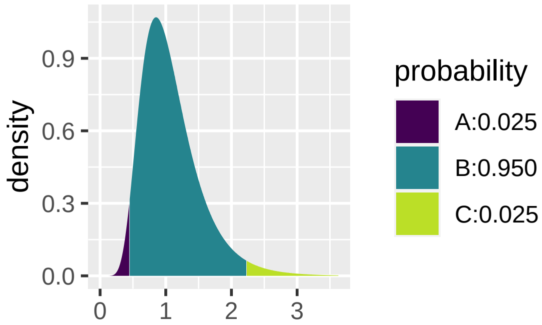
Reject H_0 if $V < c_1 = F_{(m-1),(n-1)}(\alpha_0/2)$ or $V > c_2 = F_{(m-1),(n-1)}(1-\alpha_0/2)$

The F distribution in R

```
varx <- var(clouds$lograin[clouds$group=="Seeded"])
vary <- var(clouds$lograin[clouds$group=="Unseeded"])
varx/vary
```

```
## [1] 0.9490963
```

```
xqf(c(.025,.975),df1=n-1,df2=m-1)
```




```
## [1] 0.4483698 2.2303021
```

Pvalue:

```
pf(.949,df1=25,df2=25)
```

```
## [1] 0.4484584
```

No significant differnece of unequal variances.

Properties of the F-test

- Confidence Intervals for comparing variances (ratios)
- If $X \sim F_{m-1,n-1} \frac{1}{X} \sim F_{n-1,m-1}$
- If $U \sim t_n then U^2 \sim F_{1,n}$
- Power function