Math 181B Worksheet Week 1

Information:

- Jue (Marquis) Hou.
- Office Hours: Monday 2pm - 3pm, APM 6442; Wednesday 1:30pm - 2:30 pm, APM 6442.

Overview of General Ideas from 181A:

- Basic Elements in Statistics:
  - Data: the random variables you can see.
  - Model: a family of distributions for the data indexed by the parameters.
  - Parameter: the fixed quantities you cannot see but determines the distribution of the data.
- Estimation: best guess about the parameter.
  - Point estimation: MLE & MoM.
    Question: Why likelihood and moments?
  - Variance of Point estimation: exact and asymptotic.
- Hypothesis Testing: binary decision on the parameter.
  - Certain family: Normal and Binomial.
  - Generalized Likelihood Ratio.
  - (Uniformly) Most Powerful Test.

Distributions and Their Connections:

- Normal $N(\mu, \sigma^2)$: the limit of ... everything!
- Chi-square $\chi^2_{df}$: for an integer valued $df$, if $Z_1, \ldots, Z_{df} \sim iid N(0,1)$, $Z_1^2 + \cdots + Z_{df}^2 \sim \chi^2_{df}$.
- Student’s t-distribution $t_{df}$: if $Z \sim N(0,1)$, $V \sim \chi^2_{df}$ and $Z$ independent of $V$, $Z/\sqrt{V/df} \sim t_{df}$.
- Snedecor’s F-distribution $F_{df1,df2}$: if $V_1 \sim \chi^2_{df_1}$, $V_2 \sim \chi^2_{df_2}$ and $V_1$ independent of $V_2$, $(V_1/df_1)/(V_2/df_2) \sim F_{df1,df2}$.

The Wald Statistic:  The fundamental idea for asymptotic confidence intervals and hypothesis testing.

- Question: What is a statistic?
• Suppose \( W \) is a statistic whose mean is roughly the parameter of interest, \( \mathbb{E}(W) \rightsquigarrow \theta \). Denote the variance of \( W \) as \( \sigma^2 \), which can be estimated consistently by \( s^2 \) if unknown. The Wald statistic is 
\[
(W - \theta)/s \rightsquigarrow N(0, 1).
\]
• Confidence interval \((1 - \alpha) \times 100\%:\)
\[
[W - Z_{\alpha/2}s, W + Z_{\alpha/2}s].
\]
• Test statistic under \( H_0 : \theta = \theta_0 \):
\[
(W - \theta)/s_H^0 \rightsquigarrow N(0, 1).
\]
• Examples:
  1. One sample test for Normal mean. Data and model: \( X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2) \).
     Hypotheses: \( H_0 : \mu = \mu_0 \) vs. \( H_1 : \text{one of } \mu < \mu_0, \mu > \mu_0 \) or \( \mu \neq \mu_0 \).
     Set \( W = \bar{X} \) because \( \mathbb{E}(\bar{X}) = \mu \). The variance of \( W \) is estimated by \( s^2/n \), where \( s^2 \) is the sample variance if \( X_1, \ldots, X_n \).
  2. Two (unpaired) samples test for Normal mean. Data and model: \( X_1, \ldots, X_n \overset{iid}{\sim} N(\mu_x, \sigma_x^2) \) and \( Y_1, \ldots, Y_m \overset{iid}{\sim} N(\mu_y, \sigma_y^2) \). Hypotheses: \( H_0 : \mu_x = \mu_y \) vs. \( H_1 : \text{one of } \mu_x < \mu_y, \mu_x > \mu_y \) or \( \mu_x \neq \mu_y \).
     The parameter of interest is \( \mu_x - \mu_y \). Set \( W = \bar{X} - \bar{Y} \) because \( \mathbb{E}(\bar{X} - \bar{Y}) = \mu_x - \mu_y \).
     The variance of \( W \) is estimated by \( s_x^2/n + s_y^2/n \), where \( s_x^2 \) and \( s_y^2 \) are the estimators for \( \sigma_x^2 \) and \( \sigma_y^2 \), respectively.

Robustness of t-test when normality is violated:

• Question: How do you know that your data is normally distributed?
• Experiment: use two sample t-test on various distributions.

```r
N = 10000  # Repeats
m=n=2      # Should get t_2 if we assume equal variance
# We compare the sample distribution of t-statistic
# under null (same mean between groups)
# with data simulated from Normal, Exp(1), Unif(0,1) and t_1
t.norm = t.exp = t.unif = t.t1 = rep(0,N)
for(i in 1:N)
{
  x=rnorm(n); y=rnorm(m);
}
```
```r
t.norm[i] = t.test(x,y, var.equal = TRUE)$statistic
x=rexp(n);y=rexp(m);
t.exp[i] = t.test(x,y, var.equal = TRUE)$statistic
x=runif(n);y=runif(m);
t.unif[i] = t.test(x,y, var.equal = TRUE)$statistic
x=rt(n,1);y=rt(m,1);
t.t1[i] = t.test(x,y, var.equal = TRUE)$statistic
}
curve(dt(x,2),from = qt(0.02,2), to = qt(0.98,2),
    ylim = c(0,0.6))
lines(density(t.norm,from = qt(0.02,2), to = qt(0.98,2)),
    col = 2, lty = 2)
lines(density(t.exp,from = qt(0.02,2), to = qt(0.98,2)),
    col = 3, lty = 2)
lines(density(t.unif,from = qt(0.02,2), to = qt(0.98,2)),
    col = 4, lty = 2)
lines(density(t.t1,from = qt(0.02,2), to = qt(0.98,2)),
    col = 5, lty = 2)

legend("topleft",c("Density of t_2", "t-stat under Normal"
    ,"t-stat under Exp", "t-stat under Unif"
    ,"t-stat under t_1"),
    col = 1:5, lty = c(1,2,2,2,2))
```

---

\[ N = 10000 \]  # Repeats
\[ m=n=10 \]  # Should get \( t_{18} \) if we assume equal variance
# We compare the sample distribution of t-statistic
# under null (same mean between groups)
# with data simulated from Normal, Exp(1), Unif(0,1) and t_1

t.norm = t.exp = t.unif = t.t1 = rep(0,N)
for(i in 1:N)
{
x=rnorm(n);y=rnorm(m);
t.norm[i] = t.test(x,y,var.equal = TRUE)$statistic
x=rexp(n);y=rexp(m);
t.exp[i] = t.test(x,y,var.equal = TRUE)$statistic
x=runif(n);y=runif(m);
t.unif[i] = t.test(x,y,var.equal = TRUE)$statistic
x=rt(n,1);y=rt(m,1);
t.t1[i] = t.test(x,y,var.equal = TRUE)$statistic
}

curve(dt(x,18),from = qt(0.02,18), to = qt(0.98,18),
   ylim = c(0,0.6))
lines(density(t.norm,from = qt(0.02,18), to = qt(0.98,18)),
   col = 2, lty = 2)
lines(density(t.exp,from = qt(0.02,18), to = qt(0.98,18)),
   col = 3, lty = 2)
lines(density(t.unif,from = qt(0.02,18), to = qt(0.98,18)),
   col = 4, lty = 2)
lines(density(t.t1,from = qt(0.02,18), to = qt(0.98,18)),
   col = 5, lty = 2)

legend("topleft",c("Density of t_18", "t-stat under Normal"
   , "t-stat under Exp", "t-stat under Unif"
   , "t-stat under t_1"),
   col = 1:5, lty = c(1,2,2,2,2))
Density of $t_{18}$

$t$-stat under Normal

$t$-stat under Exp

$t$-stat under Unif

$t$-stat under $t_1$