Midterm Review

Table of contents

Chapter 3: Probability

Normal distribution

$X \sim \mathrm{N}(8, 3)$

$\mathrm{P}(X < 10)$

```
dfs <- tibble(
  x=seq(-5, 25, length=500),
  `f(x)` = dnorm(x,  mean=8, sd=3)
  )

gp <- ggplot(dfs) +
  aes(x = x, y = `f(x)`) +
  geom_area(alpha = 0.4)

gp +
  geom_area(
    data = dfs |>
      filter(x < 10),
    fill="coral1"
    )
```

```
pnorm(q=10, mean=8, sd=3)
```

```
[1] 0.7475075
```

Poisson Distribution The number of parasites in a host is a Poisson random variable with mean 3.5. What is the probability that there will be at least one parasite in the host?

$Z \sim \mathrm{P}(\lambda = 3.5)$

```
1 - dpois(x=0, lambda=3.5)
```

```
[1] 0.9698026
```

Binomial Distribution

A microbiologist conducts an experiment to create a recombinant strain of bacteria that is resistant to penicillin. She plates out the bacteria on a plate, and picks out 10 colonies. She knows that the probability of successfully creating a recombinant is 0.15. Given $X \sim \mathrm{Bin}(n = 10, p = 0.15)$, what is $P(x > 0)$

```
1 - dbinom(x=0, size=10, prob=0.15)
```

```
[1] 0.8031256
```

Chapter 4: Statistical Inference

```
url1 <- "http://www.massey.ac.nz/~anhsmith/data/rangitikei.RData"
download.file(url = url1, destfile = "rangitikei.RData")
load("rangitikei.RData")
```

```
# ggplot style
library(ggplot2)
p1 <- ggplot(rangitikei, mapping = aes(people))+
  geom_histogram(aes(y=..density..), bins=10)

p1 + stat_function(fun = dnorm, args=list(mean=mean(rangitikei$people), sd=sd(rangitikei$p
```

```r
# Old style
qqnorm(rangitikei$people)
qqline(rangitikei$people)
```

```r
# ggplot style
ggplot(rangitikei, aes(sample=people))+stat_qq()+stat_qq_line()
```

```r
# test for normality
shapiro.test(rangitikei$people)
```

```
    Shapiro-Wilk normality test

data:  rangitikei$people
W = 0.65346, p-value = 1.382e-07
```

```r
ks.test(rangitikei$people, "pnorm")  # ties
```

```
    Asymptotic one-sample Kolmogorov-Smirnov test

data:  rangitikei$people
D = 0.99997, p-value < 2.2e-16
alternative hypothesis: two-sided
```

What type of test does each of the following code chucks specify? State a null and alternative hypothesis for each. Interpret results.

```r
t.test(rangitikei$people, mu=100)
```

```
    One Sample t-test

data:  rangitikei$people
t = -1.8824, df = 32, p-value = 0.0689
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
  41.1334 102.3211
sample estimates:
mean of x
 71.72727
```

```r
# Null true mean of people is = 100
# alternative is the true mean of people is not = 100
```

```r
t.test(rangitikei$people, mu=100, alternative="greater")
```

```
    One Sample t-test

data:  rangitikei$people
t = -1.8824, df = 32, p-value = 0.9655
alternative hypothesis: true mean is greater than 100
95 percent confidence interval:
 46.28578      Inf
sample estimates:
mean of x
 71.72727
```

```r
# Null true mean of people is less than or equal to 100
# alternative is the true mean of people is greater than 100
```

```r
t.test(rangitikei$people~factor(rangitikei$time))
```

```
    Welch Two Sample t-test

data:  rangitikei$people by factor(rangitikei$time)
t = -3.1677, df = 30.523, p-value = 0.003478
alternative hypothesis: true difference in means between group 1 and group 2 is not equal to
95 percent confidence interval:
 -102.28710  -22.13049
sample estimates:
mean in group 1 mean in group 2
       22.71429        84.92308
```

```r
t.test(people~factor(time), data = rangitikei, var.equal = TRUE)
```

```
    Two Sample t-test

data:  people by factor(time)
t = -1.7466, df = 31, p-value = 0.0906
```

```
alternative hypothesis: true difference in means between group 1 and group 2 is not equal to
95 percent confidence interval:
 -134.84849   10.43091
sample estimates:
mean in group 1 mean in group 2
       22.71429        84.92308
```

Transformations

According to this boxcox figure what is the most appropriate transformation to try first.

```
library(lindia)
gg_boxcox(lm(rangitikei$people~1))
```

How would you perform this transformation for a ttest?

```
rep_sq_people <- (-1-sqrt(rangitikei$people))
t.test(rep_sq_people, mu=100, alternative="greater")
```

```
    One Sample t-test

data:  rep_sq_people
t = -156.53, df = 32, p-value = 1
alternative hypothesis: true mean is greater than 100
95 percent confidence interval:
 -9.680899       Inf
sample estimates:
mean of x
  -8.5067
```

```
# Null true mean of people is less than or equal to 100
# alternative is the true mean of people is greater than 100
```

Chapter 5: Tabulated Counts

Exercise 5.2

The genetic information of an organism is stored in its Deoxyribonucleic acid (DNA). DNA is a double stranded helix made up of four different nucleotides. These nucleotides differ in which of the four bases Adenine (A), Guanine (G), Cytosine (C), or Thymine (T) they contain. A simple pattern that we may want to detect in a DNA sequence is that of the nucleotide at position `i+1` based on the nucleotide at position `i`. The nucleotide positional data collected by a researcher in a particular case is given in the following table:

`i\(i+1)`	A	C	G	T
A	622	316	328	536
C	428	262	204	306
G	354	294	174	266
T	396	330	382	648

Perform a test of association and then obtain the symmetric plot.

```r
tabledata <- data.frame(
  A = c(622, 428, 354, 396),
  C = c(316, 262, 294, 330),
  G = c(328, 204, 174, 382),
  T = c(536, 306, 266, 648),
  row.names = c("A", "C", "G", "T")
  )
```

```r
chisq.test(tabledata)$exp
chisq.test(tabledata)
chisq.test(tabledata, simulate.p.value = T)

library(MASS)
corresp(tabledata)
plot(corresp(tabledata, nf=2))
abline(v=0)
abline(h=0)
```

```r
#or
library(FactoMineR)
CA(tabledata)
```

Test questions

21

n=5

mean = 158

var = 20

```r
#95% ci for the mean
n <- 5
mu <- 158
sd <- sqrt(20)

SE <- sd / sqrt(n)
SE
```

```
[1] 2
```

$$\bar{x} \pm t \times \text{se}$$

The t value is the 0.975 quantile of the t distribution with the degrees of freedom given by $n-1$.

```r
qt(p = 0.975, df = n - 1)
```

```
[1] 2.776445
```

```r
mu + (2.77*SE)
```

```
[1] 163.54
```

```
mu - (2.77*SE)
```

```
[1] 152.46
```

```
mu + (qt(p = 0.975, df = n - 1)*SE)
```

```
[1] 163.5529
```

```
n = 5
s = sqrt(12)

se = s / sqrt(n)
se
```

```
[1] 1.549193
```

$$\bar{x} \pm t \times \text{se}$$

The t value is the 0.975 quantile of the t distribution with the degrees of freedom given by $n - 1$.

```
qt(p = 0.975, df = n - 1)
```

```
[1] 2.776445
```

So, the sample mean and confidence interval is:

```
( xbar <- 161 )
```

```
[1] 161
```

```
( xbar - qt(p = 0.975, df = n - 1) * se )
```

```
[1] 156.6987
```

```r
( xbar + qt(p = 0.975, df = n - 1) * se )
```

```
[1] 165.3013
```

22 Two fertilizers

n=23 stand n=12, mu=101, var=18 new n=11, mu=124, var=14

pooled variance

```r
# w1= (n1-1)/
```

23

```r
new_fert <- c(124.8, 118.5, 128.8, 117.8, 124.2, 122.3, 114.5, 120.7, 123.9, 119.1, 121.5)
stand_fert <- c(106.4, 105.5, 103.8, 97.7, 96.5, 91.4, 97.7, 99.6, 97, 92.3, 103.9, 102)

mean(new_fert - stand_fert)
```

```
[1] 22.25833
```

```r
mean(new_fert) - mean(stand_fert)
```

```
[1] 21.9803
```

```r
# y+t x S/sqrt(n)
#
# n=11
```