# Biostatistics 201A Fall 2021 Homework 1

Lillian Chen

10/18/2021

1. Assume Z is a standard normal random variable with mean 0 and variance 1.

(a) $P(Z \geq -0.5) =?$

Solution: $P(Z \geq -0.5) = P(Z < 0.5) \approx 0.69$

```
pnorm(0.5)
```

```
## [1] 0.6914625
```

```
pnorm(-0.5, lower.tail = F)
```

```
## [1] 0.6914625
```

(b) $P(Z <?) = 0.20$

Solution: The Z-score that corresponds to a probability of 0.20 under the standard normal distribution is $Z \approx -0.84$.

```
qnorm(0.20)
```

```
## [1] -0.8416212
```

```
qnorm(0.80, lower.tail = F)
```

```
## [1] -0.8416212
```

(c) $P(-2.0 < Z < 2.0) =?$

Solution: $P(-2.0 < Z < 2.0) = P(Z < 2.0) - P(Z < -2.0) =\approx 0.954$

```
pnorm(2.0) - pnorm(-2.0)
```

```
## [1] 0.9544997
```

(d) For what value of $d$ is it true that $P(Z < -1.5) = P(Z > d)$?

Solution: We know that the standard normal curve is unimodal and symmetric around the mode, so we can say that $P(Z < -1.5) = P(Z > d)$ is true for value of $d = 1.5$.

```r
pnorm(-1.5)
```

```
## [1] 0.0668072
```

```r
lhs <- pnorm(-1.5)
qnorm(lhs, lower.tail = F)
```

```
## [1] 1.5
```

2. Calculate the following:

(a) For a vaccine that prevents disease in 99% of the people who receive it, calculate the probability that among 1,000 people receiving the vaccine, there will be 3 or fewer people who experience the disease.

Solution: We use the cumulative probability function for the binomial distribution to find the probability of 3 or fewer individuals who experience the disease in a group of 1,000 people that receive the vaccine, which comes out to approximately 1.01%.

```r
pbinom(3, 1000, 0.01)
```

```
## [1] 0.01007265
```

(b) What is the smallest number of independent tosses of a fair coin, n, such that the probability of obtaining either all heads or no heads is less than 0.05?

Solution: Let n represent the number of independent tosses for a fair coin. The probability of heads of a fair coin is 0.5, and probability of tails of a fair coin is 0.5.

$$P(\text{either all heads or no heads}) < 0.05$$
$$2 * P(\text{all heads}) < 0.05$$
$$P(\text{all heads}) < 0.025$$
$$0.5^n < 0.025$$
$$\ln(0.5^n) < \ln(0.025)$$
$$n \ln(0.5) < \ln(0.025)$$
$$n > 5.32$$

The smallest number of independent tosses for a fair coin such that the probability of obtaining either all heads or no heads is less than 0.05 is $n = 6$.

3. Assume that heights in the U.S. population are normally distributed with mean 70 inches and standard deviation 4 inches.

(a) Suppose we repeatedly take samples of size 20 from the population and calculate the sample mean height for each. What is the distribution of these sample means, and how does the standard deviation of this distribution relate to the standard deviation of heights in the population?

Solution: Since the distribution of a sample from a normally distributed population is given by $N(\mu, \sigma^2/n)$, the distribution of these sample mean heights is $\bar{x} \sim N(70, 0.8)$ $(\sigma^2/n = 4^2/20)$. The standard deviation of this distribution is scaled by $1/sqrt(n)$ (and thus smaller) compared to that of the standard deviation of heights in the population distribution. Additionally, the standard deviation of this distribution of sample mean heights is equivalent to the standard error of the standard deviation of heights in the population.

(b) Suppose we sample 30 people from an island and are interested to know whether the average height is significantly different from the average height in the U.S. Suppose further that the sample mean of the island individuals is 68.1 inches. Set up null and alternative hypotheses for this scenario, propose a test statistic, calculate a p-value, and comment on whether the result appears significant at the $\alpha = 0.05$ level.

Solution:

$H_0 : \mu_{island} = \mu_{US}$ $H_A : \mu_{island} \neq \mu_{US}$

$Z^* = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{68.1-70}{4/\sqrt{20}} = -2.12$

$p - value = 2 * P(Z \geq |-2.12|) = 2 * 0.017 = 0.034$

For a two-tailed hypothesis test, the p-value is 0.034.

Since the p-value falls below $\alpha = 0.05$, the average height from the island sample is significantly different from the average height of the U.S. population ($p = 0.034$).

```
sigma <- 4
z <- (68.1-70)/(sigma/sqrt(20))
(pvalue <- 2*pnorm(z))
```

```
## [1] 0.03364803
```

(c) Following further on the example from part (b), suppose the population standard deviation is not known, but we calculate the sample variance $s^2$ from the sample of size 30 to be 15.0 (where, as before, the sample mean is still 68.1 inches). Again set up null and alternative hypotheses to consider whether the average height of the island population is significantly different from the average height in the U.S., propose a test statistic, calculate a p-value, and comment on whether the result appears significant at the $\alpha = 0.05$ level.

Solution: Since the population standard deviation is unknown, we use a t-test for our test statistic and p-value calculation.

$H_0 : \mu_{island} = \mu_{US}$ $H_A : \mu_{island} \neq \mu_{US}$

$t^* = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{68.1-70}{\sqrt{15}/\sqrt{30}} = -0.491$

$p = 2 * P(t \geq |-2.687|) = 0.012$

For a two-tailed hypothesis test, the p-value is 0.012.

Since the p-value falls below $\alpha = 0.05$, the average height from the island sample is significantly different from the average height of the U.S. population ($p = 0.012$).

```
s <- sqrt(15.0)
t <- (68.1-70)/(s/sqrt(30))
(pvalue <- 2*pt(t, df=29))
```

```
## [1] 0.01181097
```

(d) Following further on the scenario in part (c), suppose we are interested to learn whether the average height of the island population is significantly different from the average height of visitors to the island airport, where a sample of size 20 visitors yields a sample mean of 70.1 inches and a sample variance of 16.2 inches. Set up null and alternative hypotheses to consider whether the average height of the island population is significantly different from the average height of island visitors, propose a test statistic, calculate a p-value, and comment on whether the result appears significant at the $\alpha = 0.05$ level.

Solution: Since this scenario is comparing two samples (where the sample of the island population is of size 30 with sample mean of 68.1 inches and sample variance of 15.0 inches, and the sample of the visitors to the island airport is of size 20 with sample mean of 70.1 inches and sample variance of 16.2 inches), we set up a two sample t test for two independent populations where the population variance is unknown for both. We assume equal variances since $s^2_{visit}/s^2_{island} = 16.2/15 = 1.08 < 3$, and thus use a pooled sample standard deviation $s_p$ for calculation of the test statistic.

$$H_0 : \mu_{island} = \mu_{visit}$$
$$H_A : \mu_{island} \neq \mu_{visit}$$

$$TS = \frac{(\bar{x}_{island} - \bar{x}_{visit}) - (\mu_{island} - \mu_{visit})_0}{s_p\sqrt{1/n_{island} + 1/n_{visit}}}$$
$$s_p = \sqrt{\frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}}$$

$$TS = \frac{(68.1 - 70.1) - 0}{20.04 * \sqrt{1/30 + 1/20}}$$
$$p = 2 * P(t \geq |-0.346|) = 0.731$$

Since the p-value does not fall below $\alpha = 0.05$, the average height from the island sample is not significantly different from the average height of the visitor sample ($p = 0.731$).

```
n <- 30
xbar1 <- 68.1
s1 <- 15
m <- 20
xbar2 <- 70.1
s2 <- 16.2
s_p <- sqrt(((n-1)*s1^2 + (m-1)*s2^2)/n+m-2)

ts <- (xbar1 - xbar2 - 0)/(s_p*sqrt(1/n + 1/m))

2*pt(ts, df=n+m-2)
```

```
## [1] 0.7311003
```

4. Assuming a 99% confidence interval for $(\mu_1 - \mu_2)$ is 4.8 to 9.2, comment on whether each of the following conclusions should be supported or not, and explain your reasoning.

(a) Do not reject $H_0 : \mu_1 = \mu_2$ at the $\alpha = 0.05$ level if the alternative is $H_A : \mu_1 \neq \mu_2$.
(b) Reject $H_0 : \mu_1 = \mu_2$ at the $\alpha = 0.01$ level if the alternative is $H_A : \mu_1 \neq \mu_2$
(c) Reject $H_0 : \mu_1 = \mu_2$ at the $\alpha = 0.01$ level if the alternative is $H_A : \mu_1 < \mu_2$.

(d) Do not reject $H_0 : \mu_1 = \mu_2$ at the $\alpha = 0.01$ level if the alternative is $H_A : \mu_1 \neq \mu_2$.
(e) Do not reject $H_0 : \mu_1 = \mu_2 + 3$ at the $\alpha = 0.01$ level if the alternative is $H_A : \mu_1 \neq \mu_2 + 3$

Solution:

The form of a confidence interval for $(\mu_1 - \mu_2)$ is $(\mu_1 - \mu_2) + / - t_{n+m-2}$. Thus, the CI is $7 + / - 2.2$.

a) do not support: The null should be rejected because the 99% confidence interval (and thus the 95% confidence interval by that logic) for the difference does not contain 0.
b) support: The null should be rejected because the 99% confidence interval does not contain 0, demonstrating a mean difference not equal to 0.
c) do not support: The null should not be rejected because the 99% confidence interval does not fall in the rejection region of the alternative, which includes negative values for the mean difference.
d) do not support: The null should be rejected because the 99% confidence interval for the mean difference does not contain 0.
e) do not support support: The null should be rejected because the 99% confidence interval for the mean difference does not contain 3.

5. Suppose that $\bar{X_1} = 125.2$ and $\bar{X_2} = 125.4$ are the mean systolic blood pressures for two samples of workers from different plants in the same industry. Suppose further that a test of $H_0 : \mu_1 = \mu_2$ using these samples is rejected at the $\alpha = 0.001$ level. Referring to available information as appropriate, critique each of the following possible conclusions, explaining your reasoning:

(a) There is a meaningful difference (clinically speaking) in population means but not a statistically significant difference.
(b) There difference in population means is both statistically and clinically significant.
(c) There is a statistically significant difference but not a clinically significant difference in population means.
(d) There is neither a statistically significant difference nor a clinically significant difference in population means.
(e) The sample sizes in the two groups must have been small.

Solution:

a) We cannot say that there is a clinically meaningful difference because we do not know what constitutes a meaningful difference in mean systolic blood pressures. Additionally, the test illustrated a statistically significant difference. So, a) is wrong on both counts.

b) The difference in population means is statistically significant due to the rejection of the null at $\alpha = 0.001$, however, we cannot comment on the clinical significance.

c) This conclusion is the most appropriate conclusion given what we know - the statistical test shows a statistically significant difference. We do not know about clinical significance.

d) This is incorrect since the null was rejected at the $\alpha = 0.001$ level.

e) We do not know the sample sizes and thus cannot say that the sample sizes must have been small.

6. In a study of "self-efficacy" (confidence in one's capability to perform a task) pertaining to exercise, subjects were randomly assigned to one of three groups. Group A received a one-time coaching session, treadmill exercise testing, and a personal trainer three times a week for 4 weeks. Group B received only the coaching session and treadmill exercise testing. Group C received an information brochure only. Self-efficacy was measured based on the responses to a series of questionnaire items. The following self-efficacy scores were observed after four weeks:

Group A: 156, 119, 100, 170, 130, 154 Group B: 132, 105, 144, 136, 132, 159 Group C: 110, 101, 124, 106, 113, 94

(a) Perform an analysis of variance, using $\alpha = 0.05$ as a significance level.

Solution: Because the p-value for this one-way ANOVA is $p = 0.0316$, we can say that we reject the null that $H_0 : \mu_A = \mu_B = \mu_C$ and conclude that at least one of the group means is different from each other.

```
q6a <- data.frame(score = c(156, 119, 100, 170, 130, 154,
                            132, 105, 144, 136, 132, 159,
                            110, 101, 124, 106, 113, 94),
                  group = c('A', 'A', 'A', 'A', 'A', 'A',
                            'B', 'B', 'B', 'B', 'B', 'B',
                            'C', 'C', 'C', 'C', 'C', 'C'))

anova <- aov(score ~ as.factor(group), data = q6a)
summary(anova)
```

```
##                  Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(group)  2   3267  1633.4   4.389 0.0316 *
## Residuals        15   5582   372.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b) Suppose Groups A and B are thought of as "active treatment" and Group C is thought of as a "control" treatment. Provide an estimate of the mean difference between active treatment and control treatment.

Solution:

The point estimate of the mean difference between active treatment and control treatment is 28.42 points, with the active treatment having mean score than the control treatment.

```
q6b <- data.frame(score = c(156, 119, 100, 170, 130, 154,
                            132, 105, 144, 136, 132, 159,
                            110, 101, 124, 106, 113, 94),
                  treatment = c(rep("active", 12),
                                rep("control", 6)))
q6b %>%
  slice_head(n=12) %>%
  summarise(mu_active = mean(score))
```

```
##   mu_active
## 1  136.4167
```

```
mu_active <- 136.42

q6b %>%
  slice_tail(n=6) %>%
  summarise(mu_ctrl = mean(score))
```

```
##   mu_ctrl
## 1     108
```

```
mu_ctrl <- 108

meandiff <- mu_active - mu_ctrl

anova <- aov(score ~ as.factor(treatment), data = q6b)
summary(anova)
```

```
##                       Df Sum Sq Mean Sq F value  Pr(>F)
## as.factor(treatment)  1   3230    3230   9.198 0.00792 **
## Residuals            16   5619     351
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(c) If one was to conduct pairwise comparisons between all pairs of group means using a Bonferroni correction, what significance level would be used for each test to ensure that the "experiment-wise error rate" (i.e., the probability of at least one false finding of significance) did not exceed 0.05?

Solution: Using a Bonferroni correction means that the individual confidence levels for preplanned comparisons of pairs of groups means should be set at $\alpha/m$ where $m$ represents the number of preplanned comparisons. We have 3 preplanned comparisons in this scenario, so the significance level used for each test to ensure that the experiment-wise errror rate does not exceed 0.05 is $0.05/3 = 0.0167$.

```
(bonferroni.sig <- 0.05/3)
```

```
## [1] 0.01666667
```

7. In a study of respiratory function among individuals depending on smoking status (to be discussed in class), consider the following statistical summaries that emerged from available data: Smoking status n Mean Std. Deviation Never 21 82.143 30.436 Former 44 84.250 29.298 Current 7 114.429 31.900

(a) Using connections between group means and the overall mean, between group means and the between-group sum of squares, and between the formula for the standard deviation and the within-group sum of squares, construct an ANOVA table based on the available information and carry out a test of significance at the $\alpha = 0.05$ level.

```
options(knitr.kable.NA = '')
k <- 3
N <- 21+44+7
df_b <- k-1
df_w <- N-k
df_tot <- N-1
grandmean <- (21*82.143 + 44*84.250 + 7*114.429)/N
mean11 <- 82.143
mean12 <- 84.250
mean13 <- 114.250
ssb <- 21*(82.143-grandmean)^2 + 44*(84.250-grandmean)^2 + 7*(114.429-grandmean)^2
ssw <- 20*(30.436)^2 + 43*(29.298)^2 + 6*(31.900)^2
msb <- ssb/df_b
msw <- ssw/df_w
Fstat <- msb/msw
p <- 1 - pf(Fstat, k-1, N-k)
```

```
q7a <- data.frame(Source = c("Between groups (Treatment)",
                             "Within groups (Error)",
                             "Total"),
                  df = c(k-1, N-k, N-1),
                  SS = c(ssb, ssw, ssb+ssw),
                  MS = c(msb, msw, (ssb+ssw)/(N-1)),
                  F = c(Fstat, NA, NA),
                  p = c(p, NA, NA))

kbl(q7a, booktabs = T,
    caption = "ANOVA Table for Data of Respiratory Function Among Individuals with Different Smoking Sta
 kable_styling(latex_options = c("hold_position"))
```

Table 1: ANOVA Table for Data of Respiratory Function Among Individuals with Different Smoking Status

| Source | df | SS | MS | F | p |
|---|---|---|---|---|---|
| Between groups (Treatment) | 2 | 6081.258 | 3040.6288 | 3.409071 | 0.0387365 |
| Within groups (Error) | 69 | 61542.692 | 891.9231 | | |
| Total | 71 | 67623.950 | 952.4500 | | |

(b) In contrast to the study described in Problem 6, smoking status was not randomized in this study. Describe a possible objection to inferring that differences in respiratory function across smoking-status groups is due to the individual's smoking experience. Also offer a possible rejoinder to such an objection, and based on all of the knowledge you have accumulated prior to entering this course, comment on whether you would be inclined to attribute any significant differences in respiratory function to differences in smoking status.

Solution: We cannot accurately infer that differences in respiratory function across smoking-status groups is due to the individual's smoking experience because there may be innate group differences / possible confounders that would normally be minimized with randomization present in the study. As such, these 3 groups may not be comparable to each other since we do not know if they have the same sample characteristics for possible confounders like age, sex, weight, frequency of smoking, etc, all of which may contribute to respiratory lung function. A possible rejoinder would be that the ANOVA test is good for studies that do not have full randomization and not randomizing smoking status is also the ethical thing to do in study design. Based on my knowledge, I would be inclined to attribute significant differences in respiratory function to differences in smoking status.