STAT500 S2 2018
Applied Statistics
Assignment

Aimereh

Violet coloured text are orignal answers. # You will recall the standard Gaussian distribution from lectures and the R functions rnorm() etc.

a. What is the probability that a standard Gaussian random variable exceeds 1.3?

```
# 1-pnorm(1.3)
pnorm(1.3, mean = 0, sd = 1, lower.tail = FALSE, log.p = FALSE)
```

```
## [1] 0.09680048
```

b. Verify your answer using random sampling.

```
table(rnorm(1e6) > 1.3)
```

```
##
##   FALSE    TRUE
## 902993   97007
```

c. What value is exceeded 15% of the time when sampling from a standard Gaussian distribution?

```
qnorm(0.85)
```

```
## [1] 1.036433
```

```
1-pnorm(1.036433)
```

```
## [1] 0.1500001
```

d. [hard] You will recall dnorm(), qnorm(), pnorm(), and rnorm() from lectures. You will also recall the Cauchy distribution, which has density function dcauchy(). Answer parts (a), (b) and (c) above but using the Cauchy distribution instead of the Gaussian.

```
# 1-pcauchy(1.3)
pcauchy(1.3, location = 0, scale = 1, lower.tail = FALSE, log.p = FALSE)
```

```
## [1] 0.2087144
```

```r
table(rcauchy(1e6)>1.3)
```

```
##
##  FALSE   TRUE
## 791019 208981
```

```r
qcauchy(0.85)
```

```
## [1] 1.962611
```

```r
1-pcauchy(1.962611)
```

```
## [1] 0.15
```

Question 1

A particular AUT statistician catches the bus to work every weekday (five days). He is interested in whether or not the bus is late. The Bus being late is thus a Bernoulli trial; consider a late bus to be a "success". The statistician has determined that 10% of buses are late on average.

a. If each bus's status is independent of the others, what distribution specifies the total number of late buses in a given week?

- Binormal distribution using rbinom() to represent the status of each bus is independent.
- The value **5** will simulate every weekday (five days).
- The value **1** will simulate each bus per day
- The value **0.10** will give the probability of success.
- Bernoulli trials amounts to an experiment with two possible outcomes of *"success"* and *"failure"*.
  - The value **0** will represents *"not late"*
  - The value **1** will represents *"late"*

```r
# Set the seed of R's random number generator
# Useful for recreating simulations or random objects to be reproduced
set.seed(2018)
rbinom(5,1,0.10)
```

```
## [1] 0 0 0 0 0
```

b. Using dbinom() or otherwise, what is the probability that exactly two buses are late in a given week?

```r
# (factorial(5)/(factorial(2)*factorial(3)))* (0.1^2 * (1-0.1)^(5-2))
dbinom(2,5,0.1)
```

```
## [1] 0.0729
```

c. A new company takes over the running of the bus route. It is suspected that this new company is less punctual (ie, more late buses) than the old one. State a sensible null hypthoesis for considering the new company's performance.

H_0 = Buses of the new company that takes over the running of the bus route are less punctual.

The relationship between the new bus company and the old bus company shows no reliable supporting evidences of the two company performances in terms that the new bus company is less punctual than the old bus company

d. Is a one-sideed or two-sided test appropriate? Justify your answers.

A two-sided test is appropriate to determine if there is a difference comparing the new company performance against the old company performance in the punctuality of buses arrival. The two-sided test uses both positive or negative differences.

Both method of test is appropriate. Two-sided test would be suitable because it would compare two variables, in this case comparing the new bus company and the old bus company punctuality in arriving on time. A one-sided test will only be applicable if the aim is to see the improvement on punctuality so therefore it would not measure the effectiveness relationship between a new bus company against an old bus company punctuality.

e. State the precise definition of p-value; and calculate the p-value for the observation that three buses are late. Is this significant?

Definition: **The p value is the probability, if the null hypothesis is true, of observing data or an observation more extreme.**

```r
binom.test(3, 5, 0.1, alternative="two.sided")
```

```
##
##  Exact binomial test
##
## data:  3 and 5
## number of successes = 3, number of trials = 5, p-value = 0.00856
## alternative hypothesis: true probability of success is not equal to 0.1
## 95 percent confidence interval:
##  0.1466328 0.9472550
## sample estimates:
## probability of success
##                    0.6
```

Question 2

A zoologist in Auckland studies kiwi behavior and catches kiwi in two locations, A and B. He is interested in the weights of kiwi birds in the two locations; he weighs the kiwi and records their weight before releasing them. The dataset is as follows:

- kiwi_A <- c(1.10, 1.50, 1.10, 1.25, 1.25, 1.34, 1.53, 1.82, 1.30)
- kiwi_B <- c(1.11, 1.25, 1.02, 1.12, 1.00, 0.94, 1.18, 1.02, 1.66)

The scientist does not know which of the sites is likely to have heavier kiwi.

a. State a sensible null hypothesis.

H_0 = The weight of kiwi from site A and site B are different.

The weights between Kiwis at location A against Kiwis at location B have no relations.

b. Is a one-sided test or a two-sided test needed? Why?

A two-sided test is appropriate to compare the weights of the Kiwis at site A wither they are heavier or lighter than the Kiwis at site B since the scientist does not know which site will likely have the heavier kiwis.

c. Perform a t-test on the dataset and interpret.

```
kiwi_A <- c(1.10, 1.50, 1.10, 1.25, 1.25, 1.34, 1.53, 1.82, 1.30)
kiwi_B <- c(1.11, 1.25, 1.02, 1.12, 1.00, 0.94, 1.18, 1.02, 1.66)

t.test(kiwi_A, kiwi_B)
```

```
##
##  Welch Two Sample t-test
##
## data:  kiwi_A and kiwi_B
## t = 1.9962, df = 15.938, p-value = 0.06329
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01308674  0.43308674
## sample estimates:
## mean of x mean of y
##  1.354444  1.144444
```

Question 3

A farmer grows trees for their wood and measures the girth of 30 trees in his orchard. His data is as follows:

- girth <-c(21.1, 21.8, 22.4, 26.7, 27.2, 27.4, 27.9, 27.9, 28.2, 28.4, 28.7, 29, 29, 29.7, 30.5, 32.8, 32.8, 33.8, 34.8, 35.1, 35.6, 36.1, 36.8, 40.6, 41.4, 43.9, 44.5, 45.5, 45.7, 45.7, 52.3)
- volume <- c(0.38, 0.38, 0.38, 0.61, 0.7, 0.73, 0.58, 0.67, 0.84, 0.74, 0.9, 0.78, 0.79, 0.79, 0.71, 0.82, 1.25, 1.01, 0.95, 0.92, 1.28, 1.17, 1.34, 1.42, 1.58, 2.05, 2.06, 2.16, 1.91, 1.89, 2.85)

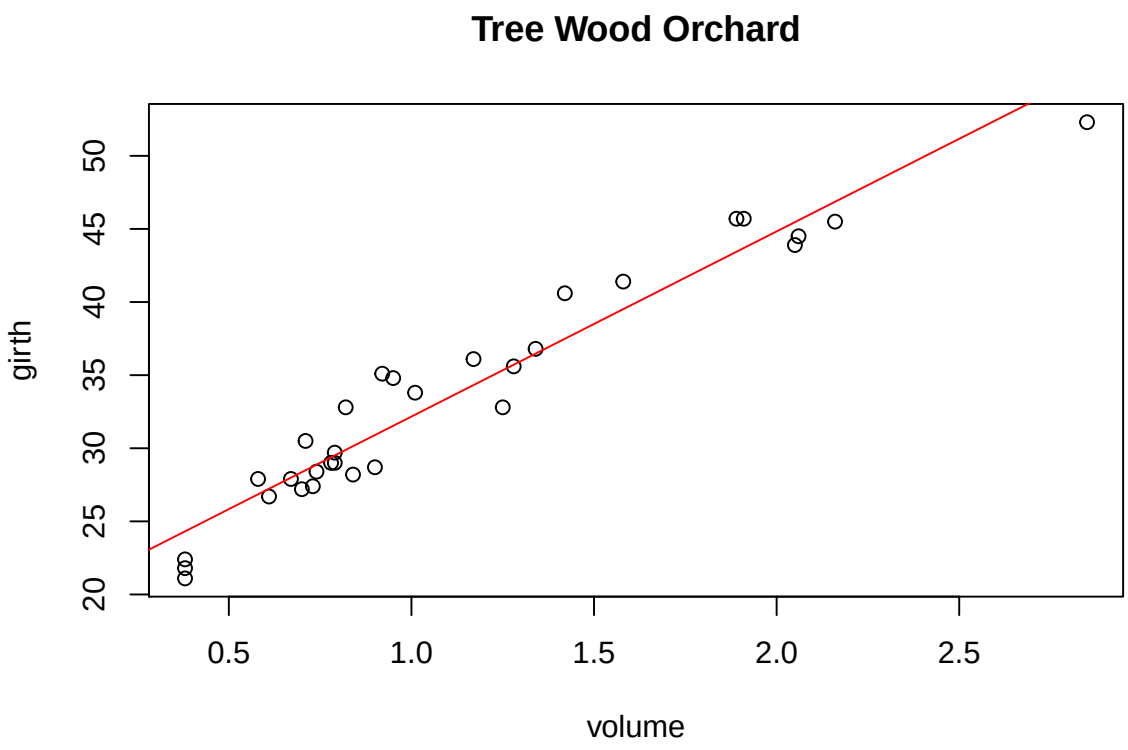
Variable "volume" represents the volume of the wood in a tree, measured in cubic meters, and "girth" represents the circumference of the tree's trunk at chest height, measured in cm. The farmer wants to predict the volume of wood in a tree as a function of its girth

a. Using R idiom such as lm(x~y) or summary(lm(x~y)), perform a linear regression on this dataset.

```
girth <-c(21.1, 21.8, 22.4, 26.7, 27.2,
          27.4, 27.9, 27.9, 28.2, 28.4,
          28.7, 29, 29, 29.7, 30.5, 32.8,
          32.8, 33.8, 34.8, 35.1, 35.6, 36.1,
          36.8, 40.6, 41.4, 43.9, 44.5, 45.5,
          45.7, 45.7, 52.3)

volume <- c(0.38, 0.38, 0.38, 0.61, 0.7,
            0.73, 0.58, 0.67, 0.84, 0.74,
            0.9, 0.78, 0.79, 0.79, 0.71,
            0.82, 1.25, 1.01, 0.95, 0.92,
            1.28, 1.17, 1.34, 1.42, 1.58,
            2.05, 2.06, 2.16, 1.91, 1.89,
            2.85)

plot(girth~volume, main="Tree Wood Orchard")
wood.lm = lm(girth~volume)
abline(wood.lm, col="red")
```




```
summary(wood.lm)
```

```
##
## Call:
## lm(formula = girth ~ volume)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2966 -1.4618 -0.3817  1.8331  3.9453
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.5036     0.7855   24.83   <2e-16 ***
## volume       12.6642     0.6197   20.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.065 on 29 degrees of freedom
## Multiple R-squared:  0.9351, Adjusted R-squared:  0.9328
## F-statistic: 417.7 on 1 and 29 DF,  p-value: < 2.2e-16
```

b. State a sensible null hypothesis for this situation, and say whether you reject it and why?

H_0 = The girth of the tree (circumference of the tree's trunk) does not influences the volume of tree wood (cubic meters).

The p-value is less than 0.05 therefore we rejected the null hypothesis as there is a strong relationship between girth and volume.

Unable to identify a relationship between girth and volume. I reject this hypothesis because it stated that the girth represents the circumference of the tree truck. The mathematical equation to find the volume of anything suggests that the bigger the circumference is and the higher the height is the volume increases.

c. Interpret coefficient of girth in the model.

The intercept value is **19.5036** which represent the average girth of a tree with a volume value of **0.6** shown by the summary above.

d. [hard] Interpret the intercept in the model. Does this interpretation make sense? If there is a problem with this interpretation, suggest a way to improve the model to remove this problem.

There is insufficient amount of data available to intercept accurate outcomes.

A Appendix: R Environment

```
sessionInfo()
```

```
## R version 4.0.4 (2021-02-15)
## Platform: i386-w64-mingw32/i386 (32-bit)
## Running under: Windows 10 x64 (build 19041)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_New Zealand.1252  LC_CTYPE=English_New Zealand.1252
## [3] LC_MONETARY=English_New Zealand.1252 LC_NUMERIC=C
## [5] LC_TIME=English_New Zealand.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] MASS_7.3-53     xtable_1.8-4    forcats_0.5.1   stringr_1.4.0
##  [5] dplyr_1.0.5     purrr_0.3.4     readr_1.4.0     tidyr_1.1.3
##  [9] tibble_3.0.6    ggplot2_3.3.3   tidyverse_1.3.0 knitr_1.31
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.1.0  xfun_0.21         haven_2.3.1       colorspace_2.0-0
##  [5] vctrs_0.3.6       generics_0.1.0    htmltools_0.5.1.1 yaml_2.2.1
##  [9] utf8_1.1.4        rlang_0.4.10      pillar_1.5.0      glue_1.4.2
## [13] withr_2.4.1       DBI_1.1.1         dbplyr_2.1.0      modelr_0.1.8
## [17] readxl_1.3.1      lifecycle_1.0.0   munsell_0.5.0     gtable_0.3.0
## [21] cellranger_1.1.0  rvest_1.0.0       evaluate_0.14     fansi_0.4.2
## [25] highr_0.8         broom_0.7.5       Rcpp_1.0.6        scales_1.1.1
## [29] backports_1.2.1   jsonlite_1.7.2    fs_1.5.0          hms_1.0.0
## [33] digest_0.6.27     stringi_1.5.3     grid_4.0.4        cli_2.3.1
## [37] tools_4.0.4       magrittr_2.0.1    crayon_1.4.1      pkgconfig_2.0.3
## [41] ellipsis_0.3.1    xml2_1.3.2        reprex_1.0.0      lubridate_1.7.10
## [45] assertthat_0.2.1  rmarkdown_2.7     httr_1.4.2        rstudioapi_0.13
## [49] R6_2.5.0          compiler_4.0.4
```