

## Homework 10: Due 11/20 (Completion)

Stat061-F23

Prof Amanda Luby

---

1. Suppose you observe  $X_1, \dots, X_{10} \sim \text{Exp}(\theta)$  and plan to test  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1$ .
  - (a) Write out the GLRT test statistic (you do *not* need to find an associated probability distribution) and explain why we would reject for large values of  $\bar{X}$  when  $\theta_1 < \theta_0$  and for small values if  $\theta_1 > \theta_0$ .
  - (b) Suppose we use the test statistic  $\bar{X}$ . Note that  $\bar{X} \sim \text{Gamma}(n, \frac{\theta}{n})$ . Suppose  $\theta_0 = 1$  and  $\theta_1 = .5$ . Find a  $\alpha = .05$  rejection region for the test statistic  $\bar{X}$ .
  - (c) The distribution of  $W = 2n\theta\bar{X}$  is  $\text{Gamma}(n, 1/2)$ , which is equivalent to a  $\chi^2_{2n}$  random variable. Find a general formula for an  $\alpha = .05$  rejection region using this representation. Give the rejection region in terms of  $\bar{X}$  (that may depend on  $\theta_0$ ).
  - (d) For the test in (c), suppose we observe  $\bar{X} = 2.17$ . What values of  $\theta_0$  would *not* be rejected at  $\alpha = .05$ ?
2. A study reported in the 2011 Journal of Clinical Sleep Medicine found that the rate of teenage auto accidents was much higher in one Virginia city than in a neighboring community where school started about an hour and a quarter later. The table below reports the 2007 and 2008 teen accident rates per 1000 drivers in Virginia Beach, where public high school started at 7:20pm, and in Chesapeake, where school started at 8:45am. It also gives the average for each of the two sets of  $n = 2$  numbers.

	Virginia Beach	Chesapeake
2007	71.2	55.6
2008	65.8	46.6
-----	-----	-----
Average	68.5	51.1

- (a) First, define parameters and run a *two-sample* test for the average accident rate over the two years. Justify your choice of  $H_1$ .
  - (b) Since we have two years of data, we could also treat these data as  $n = 2$  *matched pairs* (matching on year), where each data point represents the difference between the accident rate in the early-start town compared to the late-start town. Compute these differences and carry out a *one-sample t-test* using the differences. How does your conclusion compare to the conclusion in (a)?
  - (c) Would a significant result in (a) or (b) provide good evidence that starting high school later *causes* a lower rate of accidents? Explain.
3. Frisby and Clatworthy (1975) studied the times that it takes subjects to fuse random-dot stereograms. Random-dot stereograms are pairs of images that appear at first to be random dots. After a subject looks at the pair of images from the proper distance and her eyes cross just the right amount, a recognizable object appears from the fusion of the two images. The experimenters were concerned with the extent to which prior information about the recognizable object affected the time it took to fuse the images. One group of 43 subjects was not shown a picture of the object before being

asked to fuse the images. Their average time was  $\bar{X}_1 = 8.560$  and  $s_1^2 = 2745.7$ . The second group of 35 subjects was shown a picture of the object, and their sample statistics were  $\bar{X}_2 = 5.551$  and  $s_2^2 = 783.9$ . The null hypothesis is that the mean time of the first group is no larger than the mean time of the second group, while the alternative hypothesis is that the first group takes longer.

- (a) Test the hypotheses at  $\alpha_0 = .01$ , assuming that the population variances are equal for the two groups.
  - (b) Now, test the same hypotheses assuming the population variances are not equal.
4. The 41 numbers below represent the average sulfur dioxide contents over the period 1969-1971 (micrograms/cubic meter) among 41 US cities.
- (a) Test the null hypothesis that these data come from a normal distribution
  - (b) Test the null hypothesis that these data come from a log-normal distribution

```
sulfur_dioxide = tibble(
  city = 1:41,
  average = c(10, 13, 12, 17, 56, 36, 29, 14, 10, 24, 110, 28, 17, 8, 30, 9, 47,
)

head(sulfur_dioxide)
```

```
# A tibble: 6 x 2
  city average
<int>   <dbl>
1     1     10
2     2     13
3     3     12
4     4     17
5     5     56
6     6     36
```

I'm going to use a handy R function called `cut_number` to break average into 5 equal-ish groups:

```
sulfur_dioxide %>%
  mutate(sulfur_group = cut_number(average, n = 5, right = T)) %>%
  group_by(sulfur_group) %>%
  tally()
```

```
# A tibble: 5 x 2
  sulfur_group      n
<fct>         <int>
1 [8,11]           9
2 (11,17]          8
3 (17,28]          8
4 (28,46]          8
5 (46,110]         8
```

To find the MLEs, recall we take the sample mean and variance for the normal distribution, and the mean of the logs and variance of the logs for the log-normal distribution. We then obtain:

```
mles = sulfur_dioxide %>%
  summarize(
    mu_hat = mean(average),
    sigma_sq_hat = sd(average)^2,
    log_mu_hat = mean(log(average)),
    log_sigma_sq_hat = sd(log(average))^2
  )

mles
```

```
# A tibble: 1 x 4
  mu_hat sigma_sq_hat log_mu_hat log_sigma_sq_hat
  <dbl>      <dbl>      <dbl>      <dbl>
1  29.8      551.        3.14        0.489
```

I'll then compute the probability of being in each group under both (a) the normal distribution, and (b) the log-normal distribution, using the MLEs as plug in estimators.

```
sulf_dioxide_grouped = sulfur_dioxide %>%
  mutate(sulfur_group = cut_number(average, n = 5, right = T)) %>%
  group_by(sulfur_group) %>%
  tally() %>%
  mutate(
    p_normal = c(pnorm(11, mean = mles$mu_hat, sd = sqrt(mles$sigma_sq_hat)),
                 pnorm(17, mean = mles$mu_hat, sd = sqrt(mles$sigma_sq_hat)) - pnorm(11, mean = mles$mu_hat, sd = sqrt(mles$sigma_sq_hat)),
                 pnorm(28, mean = mles$mu_hat, sd = sqrt(mles$sigma_sq_hat)) - pnorm(17, mean = mles$mu_hat, sd = sqrt(mles$sigma_sq_hat)),
                 pnorm(46, mean = mles$mu_hat, sd = sqrt(mles$sigma_sq_hat)) - pnorm(28, mean = mles$mu_hat, sd = sqrt(mles$sigma_sq_hat)),
                 1-pnorm(46, mean = mles$mu_hat, sd = sqrt(mles$sigma_sq_hat))
    ),
    p_lognormal = c(plnorm(11, mles$log_mu_hat, sqrt(mles$log_sigma_sq_hat)),
                    plnorm(17, mles$log_mu_hat, sqrt(mles$log_sigma_sq_hat)) - plnorm(11, mles$log_mu_hat, sqrt(mles$log_sigma_sq_hat)),
                    plnorm(28, mles$log_mu_hat, sqrt(mles$log_sigma_sq_hat)) - plnorm(17, mles$log_mu_hat, sqrt(mles$log_sigma_sq_hat)),
                    plnorm(46, mles$log_mu_hat, sqrt(mles$log_sigma_sq_hat)) - plnorm(28, mles$log_mu_hat, sqrt(mles$log_sigma_sq_hat)),
                    1-plnorm(46, mles$log_mu_hat, sqrt(mles$log_sigma_sq_hat))
    ),
    n_normal = p_normal*41,
    n_lognormal = p_lognormal*41
  )

sulf_dioxide_grouped
```

```
# A tibble: 5 x 6
```

	sulfur_group	n	p_normal	p_lognormal	n_normal	n_lognormal
	<fct>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
1	[8,11]	9	0.212	0.143	8.67	5.85
2	(11,17]	8	0.0812	0.185	3.33	7.59
3	(17,28]	8	0.177	0.278	7.24	11.4
4	(28,46]	8	0.286	0.230	11.7	9.44
5	(46,110]	8	0.245	0.164	10.0	6.73

Finally, run our two  $\chi^2$  tests:

```
sulf_dioxide_grouped %>%
  summarize(chi_sq= c(sum((n - n_normal)^2/n_normal), sum((n - n_lognormal)^2/n_
    p_val = 1-pchisq(chi_sq, df = 5-1-2)) %>%
  mutate(model = c("normal", "lognormal"))

# A tibble: 2 x 3
  chi_sq p_val model
  <dbl> <dbl> <chr>
1    8.24 0.0162 normal
2    3.18 0.0204 lognormal
```

So, we reject  $H_0$  that  $X \sim N(\mu, \sigma^2)$  but fail to reject  $H_0$  that  $X \sim \text{lognormal}(\mu, \sigma^2)$ .