STAT 135 CONCEPTS OF STATISTICS
HOMEWORK 1

Assigned July 2, 2021, due July 9, 2021

This homework pertains to materials covered in Lecture 1 and 2. The assignment can be typed or handwritten, with your name on the document, and with properly labeled computer output for those problems that require it. To obtain full credit, please write clearly and show your reasoning. If you choose to collaborate, the write-up should be your own. Please show your work! Upload the file to the Week 1 Assignment on bCourses.

Problem 1. Let $x_1, x_2, \ldots, x_N$ be a list of numbers with mean μ and standard deviation σ (the square root of the variance σ^2). Assume $a_1, \ldots, a_M$ are the distinct values of that list with frequencies $n_1, \ldots, n_M$ respectively. Show that

$$\sigma^2 = \sum_{i=1}^{M} \frac{n_i a_i^2}{N} - \mu^2.$$

Solution. On Page 19 of Lecture 1 slides, we showed

$$\begin{aligned}\sigma^2 &= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2 \\ &\overset{\text{use the frequencies of each } x_i}{=} \frac{1}{N}\sum_{i=1}^{M} n_i a_i^2 - \mu^2 \\ &= \sum_{i=1}^{M} \frac{n_i a_i^2}{N} - \mu^2.\end{aligned}$$

Problem 2. During Lecture 1, we saw two samples randomly drawn from the entire population of Youtube users with two different mobile app designs. For each sample, two variables were collected - Session view duration (hours) and Internet speed (Mbps). The scatter plot of the samples is shown again in Figure 1.

(1) This data set may include multiple sessions from the same user during the testing period. Are the samples from Design A (blue points) i.i.d? If not, explain how the view hours of different sessions from the same user might be correlated with each other.

(2) A data analyst intern did not fully randomize the sampling of the users: Design A was mostly shown to accounts registered in rural areas, while Design B was mostly shown to accounts registered in the Bay area. As a result, he got more samples in red with higher

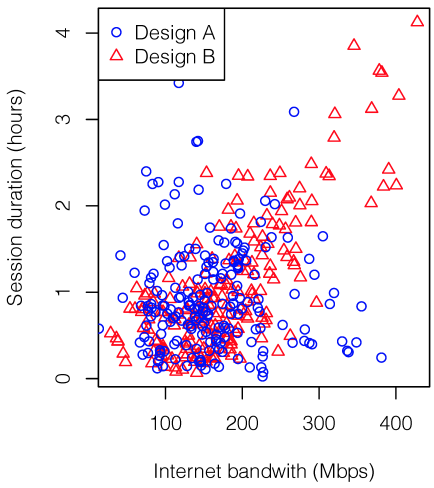

FIGURE 1. Samples from A/B testing

Internet speed than samples in blue. Is it still appropriate to analyze this data set to decide which design is better?

Solution.

(1) If multiple sessions from the same users are included in the sample, then the i.i.d assumption can not be assumed because there is high correlation in view hours among these sessions. Different users have different user behaviors. For example, elderly users may typically spend less time on Youtube than younger users. People with better Internet may have longer session duration, etc.
(2) As we see in Figure 1, higher Internet bandwidth is associated with longer session durations. If Design B included more users with faster Internet, Design B will seem to have better user retention artificially.

Problem 3. A class has two sections. Students in Section 1 have an average score of 80 with an SD of 10. Students in Section 2 have an average score of 87 with an SD of 10.

(1) If possible, say whether the SD of the scores of all the students in the class is
   (a) less than 10
   (b) equal to 10
   (c) greater than 10

   Explain your choice. If it is not possible to make the choice with the information given, explain why not.

(2) Suppose section 1 has 30 students and section 2 has 20. Find the SD of the scores of all the students in the class.

Solution.

(1) (c) greater than 10.

Because the section averages are different, the whole class will have somewhat more variability than each section. After all, the data now spread from the lowest scores in the weaker section to the highest scores in the strongest section.

(2) Denote the scores from section 1 and 2 by $x_1, \ldots, x_{30}$ and $y_1, \ldots, y_{20}$ respectively. Then

$$\sum_{i=1}^{30} x_i^2 = 30(\sigma_1^2 + \mu_1^2) = 30(10^2 + 80^2) = 195000,$$
$$\sum_{i=1}^{20} y_i^2 = 20(\sigma_2^2 + \mu_2^2) = 20(10^2 + 87^2) = 153380.$$

The mean score for all students in both sections should be

$$\mu = \frac{N_1\mu_1 + N_2\mu_2}{N_1 + N_2} = \frac{30 * 80 + 20 * 87}{30 + 20} = 82.8.$$

Thus the SD of all the scores can be calculated as the following:

$$\begin{aligned}\sigma^2 &= (x_1^2 + \ldots + x_{30}^2 + y_1^2 + \ldots + y_{20}^2)/(N_1 + N_2) - \mu^2 \\ &= \frac{195000 + 153380}{30 + 20} - 82.8^2 = 111.76, \\ SD &= \sqrt{\sigma^2} = \sqrt{111.76} = 10.572.\end{aligned}$$

Problem 4. Let X have the distribution given below, in which $0 < \theta < 1/6$:

value	1	2	3	4
probability	2θ	θ	3θ	$1 - 6\theta$

Let $X_1, \ldots, X_n$ be i.i.d samples drawn from the distribution above, and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be the sample mean.

(1) Classify each of the following quantities as a random variable or a real number.
   (a) X
   (b) $E(X)$
   (c) $X_1 * X_n$
   (d) $\bar{X}_n$

(2) Decide whether each of the following statements is true or false. Justify your answer.
   (a) $\bar{X}_n = E(X)$ for any sample size $n \geq 1$.
   (b) $E(\bar{X}_n) = E(X)$ for any sample size $n \geq 1$.
   (c) If n is sufficiently large, $\bar{X}_n$ is likely to be exactly equal to $E(X)$.
   (d) If n is sufficiently large, $\bar{X}_n$ can be very close to $E(X)$.

(3) If $n = 2$, write out the joint distribution of the two independent samples (X_1, X_2). Also, calculate $P\left(\frac{X_1+X_2}{2} \geq 3\right)$.
(4) Can you come up with an unbiased estimator of the population parameter θ?

Solution.

(1) (a) X is a random variable;
(b) $E(X)$ is a real number;
(c) $X_1 * X_n$ is a random variable;
(d) $\bar{X}_n$ is a random variable.

(2) (a) False. $\bar{X}_n$ is a random variable while $E(X)$ is a real number.
(b) True. This is proven in Theorem A of Lecture 1;
(c) False. By Central Limit Theorem, $\bar{X}_n$ converges to a Normal variable. No matter how large the sample size is, we always have $P(\bar{X}_n = E(X)) = 0$.
(d) True. By Central Limit Theorem, $\bar{X}_n \approx N(E(X), \sigma^2/n)$. As $n \to \infty$, the variance $\sigma^2/n \to 0$ and thus $\bar{X}_n$ will become highly concentrated around $E(X)$ as n becomes sufficiently large.

(3) The probability mass function for (X_1, X_2) is

(X_1, X_2)	(1,1)	(1,2)	(1,3)	(1,4)	(2,1)	(2,2)	(2,3)	(2,4)
Probability	$4\theta^2$	$2\theta^2$	$6\theta^2$	$2\theta - 12\theta^2$	$2\theta^2$	θ^2	$3\theta^2$	$\theta - 6\theta^2$
(X_1, X_2)	(3,1)	(3,2)	(3,3)	(3,4)	(4,1)	(4,2)	(4,3)	(4,4)
Probability	$6\theta^2$	$3\theta^2$	$9\theta^2$	$3\theta - 18\theta^2$	$2\theta - 12\theta^2$	$\theta - 6\theta^2$	$3\theta - 18\theta^2$	$(1 - 6\theta)^2$

From that, we can calculate

$$\begin{aligned} P\left(\frac{X_1 + X_2}{2} \geq 3\right) &= P(X_1 + X_2 \geq 6) \\ &= P\Big((X_1, X_2) = (2,4) \text{ or } (3,3) \text{ or } (3,4) \text{ or } (4,2) \text{ or } (4,3) \text{ or } (4,4)\Big) \\ &= \theta - 6\theta^2 + 9\theta^2 + 3\theta - 18\theta^2 + \theta - 6\theta^2 + 3\theta - 18\theta^2 + (1 - 6\theta)^2 \\ &= 1 - 4\theta - 3\theta^2. \end{aligned}$$

(4) First we write out the population mean:

$$\mu = 1 * (2\theta) + 2 * \theta + 3 * (3\theta) + 4 * (1 - 6\theta) = 4 - 11\theta.$$

Thus we could solve for the estimator of θ

$$\bar{X}_n = 4 - 11\hat{\theta}_n,$$

which gives $\hat{\theta}_n = \frac{4 - \bar{X}_n}{11}$. It can be easily verified that $\hat{\theta}_n$ is an unbiased estimator of θ.

Problem 5. Suppose we play the following guessing game. You pick a number from the list $x_1, x_2, \ldots, x_N$, and ask me to guess what it is. Each number from the list has an equal chance of being picked. My strategy is to guess that the value is some constant c no matter what you pick. Thus

if x_i is what you pick, the amount of error that I make with my strategy is $x_i - c$. Define the mean squared error of my strategy to be

$$mse_c = \frac{1}{N}\sum_{i=1}^{N}(x_i - c)^2.$$

Show that the value of c that minimizes mse_c is $c = \mu$, and that $mse_\mu = \sigma^2$.

Solution. The mean squared error can be re-written as

$$\begin{aligned} mse_c &= \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu + \mu - c)^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}\left\{(x_i-\mu)^2 + (\mu - c)^2 + 2(x_i - \mu)(\mu - c)\right\} \\ &= \underbrace{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}_{=\sigma^2} + \frac{1}{N}\sum_{i=1}^{N}(\mu - c)^2 + \frac{2(\mu-c)}{N}\underbrace{\sum_{i=1}^{N}(x_i - \mu)}_{=0} \\ &= \sigma^2 + (\mu - c)^2 \geq \sigma^2, \end{aligned} \tag{1}$$

in which the lower bound is only attained when $\mu = c$ and $mse_\mu = \sigma^2$.

Problem 6. Suppose $X_1, \ldots, X_n$ are i.i.d samples from a population $f(x)$ with mean μ and variance σ^2. Denote $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ as the sample mean.

(1) Prove that

$$\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

is an unbiased estimator of σ^2.

(2) Denote $\hat{\sigma}_n = \sqrt{\hat{\sigma}_n^2}$. Then

$$\hat{\sigma}_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}.$$

Is $\hat{\sigma}_n$ an unbiased estimator of σ? If not, does $\hat{\sigma}_n$ tend to underestimate or overestimate σ?
(*Hint*: Use the formula $\mathrm{Var}(\hat{\sigma}_n) = E(\hat{\sigma}_n^2) - [E(\hat{\sigma}_n)]^2$.)

Solution.

(1) In Theorem B of Lecture 2, we showed that

$$E\left\{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right\} = \frac{n-1}{n}\sigma^2.$$

Therefore,

$$E(\hat{\sigma}_n^2) = E\left\{\frac{n}{n-1} \times \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right\} = \frac{n}{n-1} \times \frac{n-1}{n}\sigma^2 = \sigma^2.$$

Hence $\hat{\sigma}_n^2$ is an unbiased estimator of σ^2. □
(*Proving from scratch without using Theorem B is also welcome.*)

(2) By the formula in the hint,

$$[E(\hat{\sigma}_n)]^2 = E(\hat{\sigma}_n^2) - \mathrm{Var}(\hat{\sigma}_n) = \sigma^2 - \mathrm{Var}(\hat{\sigma}_n),$$

from which we get

$$E(\hat{\sigma}_n) = \sqrt{\sigma^2 - \mathrm{Var}(\hat{\sigma}_n)} \leq \sigma.$$

The last inequality comes from the fact $\mathrm{Var}(\hat{\sigma}_n) \geq 0$. Therefore, $\hat{\sigma}_n$ tend to underestimate σ.

Problem 7. Two surveys were independently conducted to estimate a population mean μ. Suppose $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are the two samples obtained that are both i.i.d from the same population. Denote $\bar{X}_n$ and $\bar{Y}_m$ as the sample means. For some real numbers α and β, the two sample means can be combined to give a better estimator:

$$\hat{\mu}_{m+n} = \alpha\bar{X}_n + \beta\bar{Y}_m.$$

(1) Find the conditions on α and β that make the combined estimate unbiased.
(2) What choice of α and β minimizes the variance of $\hat{\mu}_{m+n}$, subject to the condition of unbiasedness?

Solution.

(1) Since

$$E(\hat{\mu}_{m+n}) = \alpha E(\bar{X}_n) + \beta E(\bar{Y}_m) = (\alpha + \beta)\mu,$$

We have to make $\alpha + \beta = 1$ in order for $\hat{\mu}_{m+n}$ to be an unbiased estimator for μ.

(2) While $\alpha + \beta = 1$,

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_{m+n}) &= \alpha^2\mathrm{Var}(\bar{X}_n) + \beta^2\mathrm{Var}(\bar{Y}_m) = \frac{\alpha^2\sigma^2}{n} + \frac{\beta^2\sigma^2}{m} \\
&= \sigma^2\left\{\frac{\alpha^2}{n} + \frac{(1-\alpha)^2}{m}\right\} \\
&= \sigma^2\left\{\left(\frac{1}{n} + \frac{1}{m}\right)\left(\alpha - \frac{n}{m+n}\right)^2 + \frac{1}{m} - \frac{n}{m(m+n)}\right\} \\
&\geq \sigma^2\left\{\frac{1}{m} - \frac{n}{m(m+n)}\right\} = \frac{\sigma^2}{m+n},
\end{aligned}$$

in which the minimum is only attained when $\alpha = \frac{n}{m+n}$. Therefore, the choice of α and β should be

$$\begin{cases} \alpha &= \frac{n}{m+n}, \\ \beta &= \frac{m}{m+n}. \end{cases}$$

Problem 8. A sample of size $n = 200$ is taken from a population that has a proportion $p = 1/2$. Denote $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ as the sample proportion.

(1) Find δ such that $P(|\hat{p}_n - p| \geq \delta) = 0.025$.
(2) If, in the sample, $\hat{p}_n = 0.25$, will the 95% confidence interval for p contain the true value of p?

Solution.

(1) By Theorem C of Lecture 2,

$$\begin{aligned} P(|\hat{p}_n - p| \geq \delta) &= 1 - P(|\hat{p}_n - p| < \delta) \\ &= 2 - 2\Phi(\sqrt{n}\delta/\sigma). \end{aligned}$$

The population standard deviation is $\sigma = \sqrt{p(1-p)} = 1/2$. Therefore we need to solve the following formula for δ

$$2 - 2\Phi(\sqrt{200}\delta/0.5) = 0.025,$$

that is, $\Phi(\sqrt{200}\delta/0.5) = 1 - 0.025/2$. By definition of z_α, we know

$$\sqrt{200}\delta/0.5 = z_{0.025/2}.$$

Thus, $\delta = 0.5 * z_{0.025/2}/\sqrt{200} = 0.5 * 2.241/\sqrt{200} = 0.079$.
(*Note* $z_{0.025/2}$ *can be obtained in* `R` *using* `qnorm(0.025/2, lower.tail=FALSE)`.)

(2) By Theorem D of Lecture 2, the 95% CI should be

$$\begin{aligned} \hat{p}_n \pm \frac{z_{0.05/2}\sigma}{\sqrt{n}} &= 0.25 \pm \frac{1.96 * 0.5}{\sqrt{200}} \\ &= 0.25 \pm 0.0692 = [0.1808, 0.3192], \end{aligned}$$

which does not contain the true value of p.

Problem 9. A coin lands heads with probability p. It is tossed 400 times. Compute the bootstrap 95% confidence interval for p and compare with the corresponding interval which uses a conservative estimate for the standard error, when the observed number of heads is

(1) 280,
(2) 150.

Solution.

(1) In this case, $\hat{p}_n = 280/400 = 0.7$.
Conservative estimate of 95% CI:

$$\begin{aligned} \hat{p}_n \pm \frac{z_{0.05/2} * 0.5}{\sqrt{n}} &= 0.7 \pm \frac{1.96 * 0.5}{\sqrt{400}} \\ &= 0.7 \pm 0.0480 = [0.652, 0.748]. \end{aligned}$$

Bootstrap estimate of 95% CI: The sample SD is $\hat{\sigma}_n = \sqrt{\hat{p}_n(1-\hat{p}_n)} = 0.4583$. Thus, we have

$$\hat{p}_n \pm \frac{z_{0.05/2} * 0.4583}{\sqrt{n}} = 0.7 \pm \frac{1.96 * 0.4583}{\sqrt{400}}$$
$$= 0.7 \pm 0.0449 = [0.6551, 0.7449].$$

(2) In this case, $\hat{p}_n = 150/400 = 0.375$.
Conservative estimate of 95% CI:

$$\hat{p}_n \pm \frac{z_{0.05/2} * 0.5}{\sqrt{n}} = 0.375 \pm \frac{1.96 * 0.5}{\sqrt{400}}$$
$$= 0.375 \pm 0.0480 = [0.327, 0.423].$$

Bootstrap estimate of 95% CI: The sample SD is $\hat{\sigma}_n = \sqrt{\hat{p}_n(1-\hat{p}_n)} = 0.4841$. Thus, we have

$$\hat{p}_n \pm \frac{z_{0.05/2} * 0.4841}{\sqrt{n}} = 0.375 \pm \frac{1.96 * 0.4841}{\sqrt{400}}$$
$$= 0.375 \pm .0474 = [0.3276, 0.4224].$$

Problem 10. Given a standard Normal random variable $Z \sim N(0,1)$, define $T = Z^2$. The new T is called a χ^2 distributed random variable. The probability density function of T can be computed as

$$f(t) = \begin{cases} \frac{1}{\sqrt{2\pi t}} e^{-\frac{t}{2}}, & \text{if } t \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

If we compare this with the general form of an Gamma(α, β) density function (see this wikipedia page), we will recognize that the χ^2 probability density function is, in fact, a Gamma$(\frac{1}{2}, \frac{1}{2})$.

(1) Compute $E(T)$ and $\mathrm{Var}(T)$ using $f(t)$.
(2) If $Z_1, Z_2, \ldots, Z_n$ are i.i.d standard normal random variables, define $T_n = Z_1^2 + Z_2^2 + \ldots + Z_n^2$. Then we call T_n a χ^2_n random variable (*chi-square with n degrees of freedom*). What are $E(T_n)$ and $\mathrm{Var}(T_n)$?
(3) Prove that the probability density function of T_n is the same as Gamma$(\frac{n}{2}, \frac{1}{2})$, or equivalently $T_n \sim$ Gamma$(\frac{n}{2}, \frac{1}{2})$.
(*Hint*: If $Y_1 \sim$ Gamma(α_1, β), $Y_2 \sim$ Gamma(α_2, β), and Y_1 and Y_2 are independent, then $Y_1 + Y_2 \sim$ Gamma$(\alpha_1 + \alpha_2, \beta)$.)
(4) Given two independent random variables $T_n \sim \chi^2_n$ and $S_m \sim \chi^2_m$, prove $T_n + S_m \sim \chi^2_{m+n}$.
(5) Use the central limit theorem to derive a good approximate normal distribution for T_n.

Solution.

(1) By definition of expectations,

$$\begin{aligned} E(T) &= \int_0^\infty \frac{t}{\sqrt{2\pi t}} e^{-\frac{t}{2}} dt = \int_0^\infty \frac{\sqrt{2}t}{\sqrt{\pi}} e^{-\frac{t}{2}} d\sqrt{t} \\ &\overset{s=\sqrt{t}}{=} \int_0^\infty \frac{\sqrt{2}s^2}{\sqrt{\pi}} e^{-\frac{s^2}{2}} ds \qquad (3) \\ &\overset{\text{Integration by parts}}{=} -\sqrt{\frac{2}{\pi}} s e^{-\frac{s^2}{2}} \Big|_0^\infty + \int_0^\infty \sqrt{\frac{2}{\pi}} e^{-\frac{s^2}{2}} ds \\ &= 0 + \int_{-\infty}^\infty \sqrt{\frac{1}{2\pi}} e^{-\frac{s^2}{2}} ds = 1, \end{aligned}$$

and

$$\begin{aligned} E(T^2) &= \int_0^\infty \frac{t^2}{\sqrt{2\pi t}} e^{-\frac{t}{2}} dt = \int_0^\infty \frac{\sqrt{2}t^2}{\sqrt{\pi}} e^{-\frac{t}{2}} d\sqrt{t} \\ &\overset{s=\sqrt{t}}{=} \int_0^\infty \frac{\sqrt{2}s^4}{\sqrt{\pi}} e^{-\frac{s^2}{2}} ds \\ &\overset{\text{Integration by parts}}{=} -\sqrt{\frac{2}{\pi}} s^3 e^{-\frac{s^2}{2}} \Big|_0^\infty + 3\int_0^\infty \sqrt{\frac{2}{\pi}} s^2 e^{-\frac{s^2}{2}} ds \\ &\overset{\text{Equation (3)}}{=} 0 + 3E(T) = 3. \end{aligned}$$

Therefore, $\mathrm{Var}(T) = E(T^2) - [E(T)]^2 = 2$.

(2) By the results from (1) and the independence among $Z_1, Z_2, \ldots, Z_n$, we have

$$\begin{aligned} E(T_n) &= E(Z_1^2) + E(Z_2^2) + \ldots + E(Z_n^2) = n, \\ \mathrm{Var}(T_n) &= \mathrm{Var}(Z_1^2) + \mathrm{Var}(Z_2^2) + \ldots + \mathrm{Var}(Z_n^2) = 2n. \end{aligned}$$

(3) Since each $Z_i^2 \sim \mathrm{Gamma}(1/2, 1/2)$, $i = 1, \ldots, n$, and they are independently, we know by mathematical induction of the hint that

$$T_n \sim \mathrm{Gamma}(1/2, 1/2) + \ldots + \mathrm{Gamma}(1/2, 1/2) = \mathrm{Gamma}(n/2, 1/2).$$

(4) We know from (3) that $T_n \sim \mathrm{Gamma}(n/2, 1/2)$ and $S_m \sim \mathrm{Gamma}(m/2, 1/2)$. Therefore,

$$T_n + S_m \sim \mathrm{Gamma}((m+n)/2, 1/2) = \chi^2_{m+n}.$$

(5) We already know that $E(T_n/n) = 1$ and $\mathrm{Var}(T_n/n) = 2/n$. By central limit theorem,

$$\sqrt{n}\left(\frac{T_n}{n} - 1\right) \xrightarrow{d} N(0, 2),$$

which in turn gives

$$T_n \approx N(n, 2n).$$