Tutorial 2 Solutions

STAT 3013/40278027

---

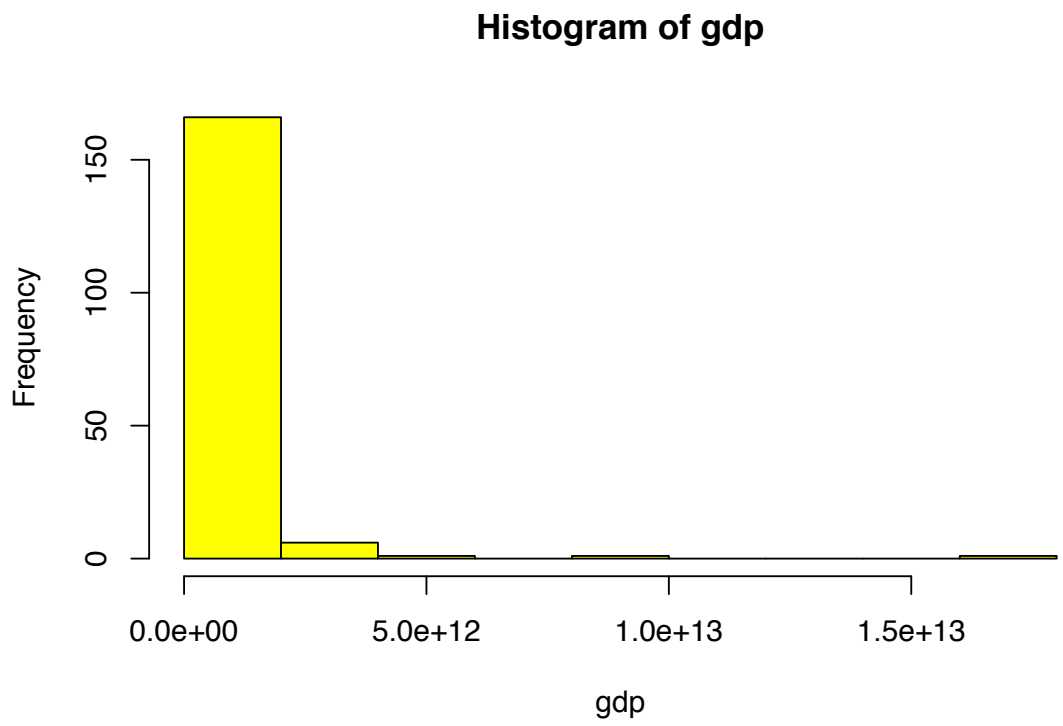
1. SI: 2.1, 2.2, 2.4. **Ans.** See the handwritten pages.
2. Consider the following:
   a. Visually display the data and discuss. Try taking the natural log of the data (when statisticians say "log" they mean natural log).
   b. Compute a six number summary of the data.
   c. Based on the "box plot rule", determine if there are any outliers. Which countries are outliers? To use the rule examine the following: Are any values in the data below the 1st Quartile - 1.5 IQR? Are any values in the data above the 3rd Quartile + 1.5 IQR? IQR is the inter-quartile range.
   d. Let $Y = \log(\text{GDP})$. Suppose $Y \sim \text{normal}(\mu, \sigma^2)$. What is your best guess for μ and σ^2 as functions of Y (call these T_1 and T_2)? What are the means (expected values) of T_1 and T_2?

Ans. First let's load in the data. Note that I removed the missing values (NA's). Missing data is an extensive and important topic in statistics; as such be careful about removing missing data. At the very least, discuss your full sample of data and then which cases were removed due to missing data.

```
## GDP
gdp2013 <- read.table("gdp2013.txt", header=TRUE)
D <- gdp2013
D <- na.omit(D)
gdp <- D$Y2013
```

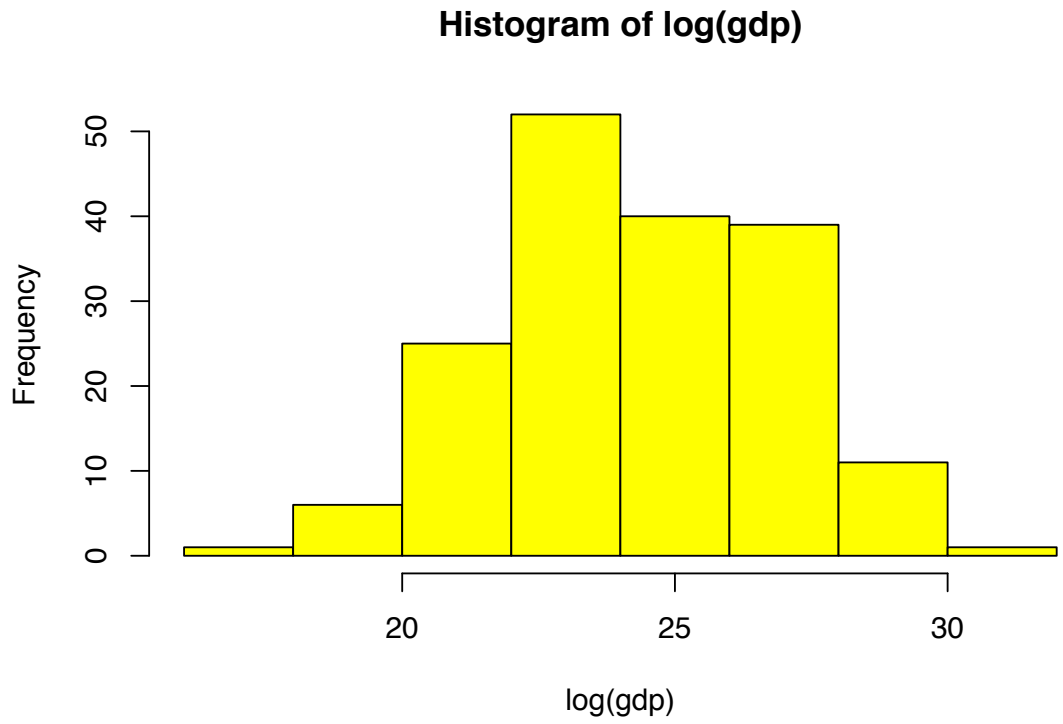
The histogram is unimodal, right skewed and appears to have some outliers.

```
## visual display
hist(gdp, col="yellow")
```



If we take the log of the data it is far more symmetric.

```
## visual display
hist(log(gdp), col="yellow")
```



Let's get a six number summary (using the logged data)

```
log.gdp <- log(gdp)
summary(log.gdp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.46   22.73   24.15   24.23   26.13   30.45
```

Let's get the quartiles and the IQR.

```
##
Q <- quantile(log.gdp, prob=c(0.25, 0.5, 0.75), type=6)
Q
```

```
##      25%      50%      75%
## 22.71874 24.14522 26.13676
```

```
IQR <- unname(Q[3]-Q[1])
IQR
```

```
## [1] 3.418021
```

Let's identify outliers based on the box plot method.

```
##
high.outliers <- log.gdp[log.gdp > Q[3] + 1.5*IQR]
low.outliers <- log.gdp[log.gdp < Q[1] - 1.5*IQR]

high.outliers
```

```
## numeric(0)
```

```
low.outliers
```

```
## [1] 17.45664
```

```
# let's find the countries that have high outliers
D[log.gdp >= min(high.outliers),1]
```

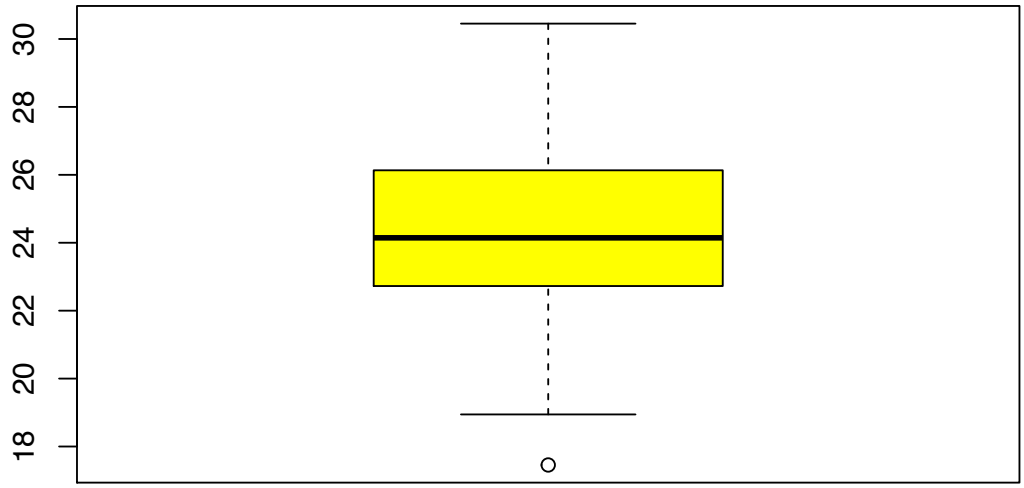
```
## Warning in min(high.outliers): no non-missing arguments to min; returning
## Inf
```

```
## factor(0)
## 214 Levels: Afghanistan Albania Algeria American Samoa Andorra ... Zimbabwe
```

```
# let's find the countries that have high outliers
D[log.gdp <= max(low.outliers),1]
```

```
## [1] Tuvalu
## 214 Levels: Afghanistan Albania Algeria American Samoa Andorra ... Zimbabwe
```

```
# let's make a boxplot.
boxplot(log.gdp, col="yellow")
```



- If we assume the data $Y_1, \ldots, Y_n \sim iid\ n(\mu, \sigma^2)$ then reasonable guesses for the population mean and variance are the sample mean and variances. So we have:

$$\bar{y} = 24.2263843 = \hat{\mu}$$

$$S^2 = 6.0934187 = \hat{\sigma}^2$$

A nice property of these estimators is that $E[\hat{\mu}] = E[\bar{Y}] = \mu$ and $E[\hat{\sigma}^2] = E[S^2] = \sigma^2$. We proved the latter in lecture and for fun let's do the former:

$$\begin{aligned} E[\hat{\mu}] = E[\bar{Y}] &= E[(1/n)(Y_1 + \cdots + Y_n)] \\ &= (1/n)E[(Y_1 + \cdots + Y_n)] \\ &= (1/n)(E[Y_1] + \cdots + E[Y_n]) \\ &= (1/n)nE[Y_1] = (1/n)n\mu = \mu \end{aligned}$$

GJJ Q 2.1.) $X_1, \ldots, X_n \overset{iid}{\sim} Unif(0, \theta)$

$$f(x) = \frac{1}{\theta - 0} = \frac{1}{\theta}$$

$Z \sim Unif(a, b)$

$f(z) = \frac{1}{b-a}$

See the table of distributions

a.) Find the density of $x_{(n)} = \max(x_1, \ldots, x_n)$

- Let $Y = \max(x_1, \ldots, x_n)$
- $P(Y \le y) = P(x_{(n)} \le y)$ — If the largest x_i is less than y then all xs are less than y.

$$= P(x_1 \le y, x_2 \le y, \ldots, x_n \le y)$$

$$= P(x_1 \le y) \times \cdots \times P(x_n \le y)$$ B/c of independence

$$= [\underbrace{P(x_i \le y)}_{CDF}]^n$$ B/c identically distributed

$$F(y) = P(x_i \le y) = \int_0^y \frac{1}{\theta}\, dx = \frac{y}{\theta}$$

$$\therefore P(x_{(n)} \le y) = \left[\frac{y}{\theta}\right]^n \quad \therefore f(y) = \frac{n y^{n-1}}{\theta^n}$$

b.) If $MSE(T(\underset{\sim}{x})) \to 0$ as $n \to \infty$, then $T(\underset{\sim}{x})$ is a Consistent estimator.

- $T_1(\underset{\sim}{x}) = 2\bar{x}$

$$\Rightarrow E(2\bar{x}) = 2E\left(\frac{1}{n}(x_1 + \cdots + x_n)\right) = \frac{2}{n} E(x_1 + \cdots + x_n)$$

$$= \frac{2}{n} \sum_{i=1}^{n} E(x) = \frac{2}{n} n \left[\frac{0+\theta}{2}\right] = \theta \quad \therefore \text{unbiased}$$

$$\Rightarrow V(2\bar{x}) = 4V\left(\frac{1}{n}(x_1 + \cdots + x_n)\right) = \frac{4}{n^2} V(x_1 + \cdots + x_n)$$

Independence means there are no Covariance terms

$$= \frac{4}{n^2} n V(x) = \frac{4}{n}\left[\frac{(\theta - 0)^2}{12}\right] = \frac{\theta^2}{3n} \to 0 \text{ as } n \to \infty$$

$$\Rightarrow MSE(2\bar{X}) = V(2\bar{X}) + \left[Bias(2\bar{X})\right]^2$$

$$= \frac{\theta^2}{3n} + 0^2 \rightarrow 0 \text{ as } n \rightarrow \infty$$

$\therefore 2\bar{X}$ is a consistent estimator of θ

- $T_2(\underset{\sim}{X}) = \dfrac{(n+1)\,X_{(n)}}{n}$

$$\Rightarrow E\left(\frac{(n+1)\,X_{(n)}}{n}\right) = \frac{(n+1)}{n} E(X_{(n)})$$

$$\Rightarrow E(X_{(n)}) = \int_0^\theta x \, \frac{n x^{n-1}}{\theta^n} dx = \frac{n}{\theta^n} \int_0^\theta x^n \, dx$$

$$= \frac{n}{\theta^n} \left[\frac{1}{n+1} x^{n+1}\right]\Big|_0^\theta = \frac{n}{n+1}\theta$$

$$\therefore E(T_2) = \frac{(n+1)}{n} \frac{n}{(n+1)} \theta = \theta \quad \therefore \text{unbiased}$$

$$\Rightarrow V(T_2(\underset{\sim}{X})) = \frac{(n+1)^2}{n^2} V(X_{(n)}) \quad \left| V(x) = E(x^2) - [E(x)]^2 \right.$$

$$\Rightarrow E(X_{(n)}^2) = \int_0^\theta x^2 \, \frac{n x^{n-1}}{\theta^n} dx = \frac{n}{\theta^n} \int_0^\theta x^{n+1} dx$$

$$= \frac{n}{\theta^n} \frac{1}{n+2} x^{n+2}\Big|_0^\theta = \frac{n}{\theta^n} \frac{\theta^{n+2}}{n+2} = \frac{n}{n+2}\theta^2$$

$$\Rightarrow V(T_2) = E(T_2^2) - \left[E(T_2)\right]^2$$

$$= \frac{(n+1)^2}{n(n+2)}\theta^2 - \theta^2 = \frac{1}{n(n+2)}\theta^2$$

$$\therefore MSE(T_2) = V(T_2) + \left[Bias(T_2)\right]^2$$

$$= \frac{1}{n(n+2)}\theta^2 + 0^2 \rightarrow 0 \text{ as } n \rightarrow \infty$$

$\therefore T_2$ is a consistent estimator of θ

- Which estimator should we choose?
  - Both are unbiased and consistent
  - ∴ Let's compare based on their variances, which in this case is the same as comparing their MSEs.

$$V(T_1) = \frac{\theta^2}{3n} \quad \text{vs} \quad V(T_2) = \frac{\theta^2}{n(n+2)}$$

$$V(T_2) - V(T_1) = \frac{\theta^2}{n(n+2)} - \frac{\theta^2}{3n}$$

$$= \left[\frac{1}{n(n+2)} - \frac{1}{3n}\right]\theta^2$$

$= 0$ for $n = 1$

< 0 for $n \geq 2$.

$V(T_2) < V(T_1)$ for $n \geq 2$ ∴ Choose T_2.

GJJ Q 2.2) $\hat{\Theta}_1, \hat{\Theta}_2$ are independent and unbiased estimators of Θ.

$V(\hat{\Theta}_1) = \sigma_1^2$, $V(\hat{\Theta}_2) = \sigma_2^2$

- Consider: $\hat{\Theta} = k_1 \hat{\Theta}_1 + k_2 \hat{\Theta}_2$

→ we want an estimator which is unbiased but has the smallest variance.

1.) $E(\hat{\Theta}) = k_1 E(\hat{\Theta}_1) + k_2 E(\hat{\Theta}_2) = \Theta$

$= k_1 \Theta + k_2 \Theta = \Theta$

$\therefore \; k_1 + k_2 = 1$

2.) Now $V(\hat{\Theta}) = k_1^2 V(\hat{\Theta}_1) + k_2^2 V(\hat{\Theta}_2)$

$= k_1^2 \sigma_1^2 + k_2^2 \sigma_2^2$

$\Rightarrow \min_{k_1, k_2} \; k_1^2 \sigma_1^2 + k_2^2 \sigma_2^2 \quad$ s.t. (such that) $\; k_1 + k_2 = 1$

$\Rightarrow \min_{k_1} \; k_1^2 \sigma_1^2 + (1-k_1)^2 \sigma_2^2$

$\frac{d}{dk_1} = 2 k_1 \sigma_1^2 - 2(1-k_1)\sigma_2^2 = 0$

$\Rightarrow k_1 \sigma_1^2 - \sigma_2^2 + k_1 \sigma_2^2 = 0$

$k_1 (\sigma_1^2 + \sigma_2^2) = \sigma_2^2$

$\therefore \; k_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$ & $k_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}$

Do we have a minimum?

$\frac{d^2}{dk_1^2} = 2\sigma_1^2 + 2\sigma_2^2 > 0$ ∴ we found the minimum.

GJJ Q 2.4) Recall Def 2.2] $\hat{\theta}$ is weakly consistent if $P(|\hat{\theta}-\theta|>\varepsilon) \to 0$ as $n \to \infty$.

→ We can show: If $V(\hat{\theta}) \to 0$ & $Bias[\hat{\theta}] \to 0$

$\therefore MSE(\hat{\theta}) \to 0$ as $n \to \infty$

then $\hat{\theta}$ is a consistent estimator of θ.

→ $P(\hat{\theta}=\theta) = \frac{(n-1)}{n}$

$P(\hat{\theta}=\theta+n) = \frac{1}{n}$

} We only have two options for $\hat{\theta}$

- Let's consider the direct approach:

$$\lim_{n\to\infty} P(|\hat{\theta}-\theta|>\varepsilon) = \lim_{n\to\infty} P(\hat{\theta}=\theta+n)$$

$$= \lim_{n\to\infty} \frac{1}{n} = 0$$

$\therefore \hat{\theta}$ is consistent estimator

- Note: $E(\hat{\theta}) = \theta\, P(\hat{\theta}=\theta) + (\theta+n)\, P(\hat{\theta}=\theta+n)$

Possibility Probability

$$= \theta\,\frac{(n-1)}{n} + (\theta+n)\,\frac{1}{n}$$

$$= \theta+1 \not\to 0 \text{ as } n \to \infty \quad \therefore Bias \not\to 0.$$