

Homework 5 — Solutions

STAT 515 · Chapters 4.9–4.13 (Lectures 8–10) · 90 points

Problem 1: Assessing Normality and the Normal Q–Q Plot (20 points)

Data (already sorted): 18, 21, 23, 24, 26, 26, 28, 30, 33, 37, $n = 10$.

(a) $\sum x_i = 266$, so

$$\bar{x} = \frac{266}{10} = 26.6 \text{ minutes}, \quad s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{288.4}{9}} = \sqrt{32.04} = 5.66 \text{ minutes}.$$

(b)

Interval	Endpoints	Values inside	Count	Percent	Empirical Rule
$\bar{x} \pm s$	(20.94, 32.26)	21, 23, 24, 26, 26, 28, 30	7	70%	$\approx 68\%$
$\bar{x} \pm 2s$	(15.28, 37.92)	all ten values	10	100%	$\approx 95\%$

The observed 70% matches the Empirical Rule's 68% very well. The 100% at two standard deviations is a little high, but with only $n = 10$ observations the possible percentages jump in steps of 10%, so this is not evidence against normality.

(c) Using Q_p at position $h = (n+1)p$:

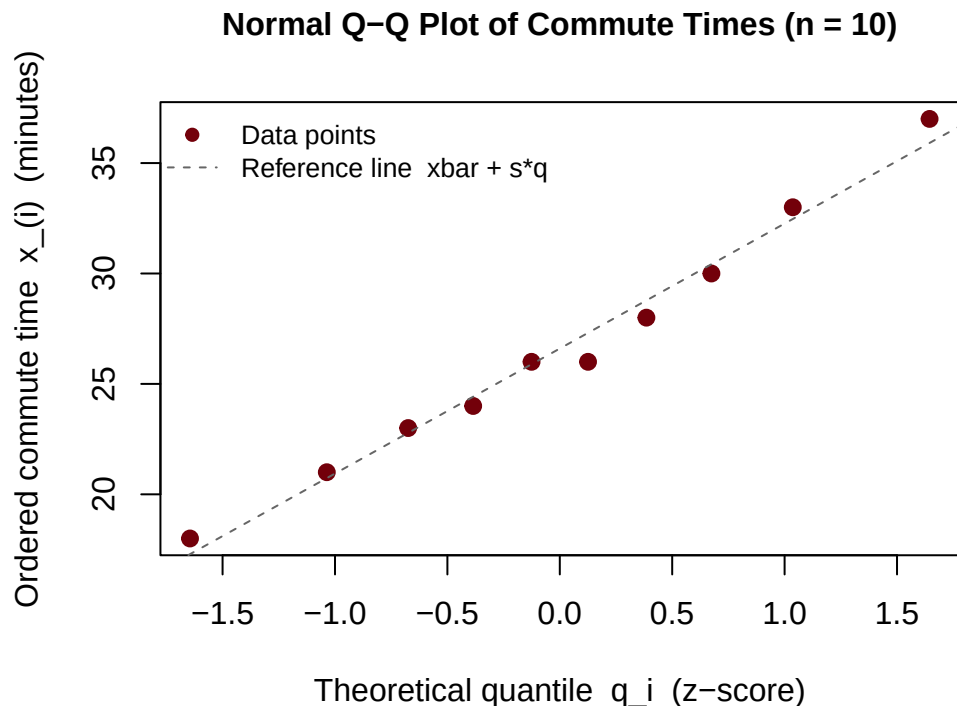
- Q_1 : $h = 11(0.25) = 2.75 \Rightarrow Q_1 = x_{(2)} + 0.75(x_{(3)} - x_{(2)}) = 21 + 0.75(2) = 22.5$
- Q_3 : $h = 11(0.75) = 8.25 \Rightarrow Q_3 = x_{(8)} + 0.25(x_{(9)} - x_{(8)}) = 30 + 0.25(3) = 30.75$

$$\text{IQR} = 30.75 - 22.5 = 8.25, \quad \frac{\text{IQR}}{s} = \frac{8.25}{5.66} = 1.46.$$

This is reasonably close to the benchmark of 1.35 expected for normal data (it is about 8% high), and far from values that would signal heavy tails (much larger than 1.35) or light tails.

(d) The ordered values paired with their theoretical quantiles:

i	1	2	3	4	5	6	7	8	9	10
q_i	-1.645	-1.036	-0.674	-0.385	-0.126	0.126	0.385	0.674	1.036	1.645
$x_{(i)}$	18	21	23	24	26	26	28	30	33	37



The reference line is $x = \bar{x} + s q = 26.6 + 5.66 q$. The points fall very close to it (the correlation between q_i and $x_{(i)}$ is 0.992); there is no systematic curvature or S-shape.

(e) Yes. All three checks agree: the Empirical Rule percentages are close to 68%/95%, $IQR/s = 1.46 \approx 1.35$, and the Q-Q plot is nearly a straight line with no systematic curvature. Treating these commute times as a sample from a normal distribution is reasonable. (The largest value, 37, sits slightly above the line, hinting at very mild right skew, but with $n = 10$ that much wiggle is expected from sampling variability alone.)

(f) R output:

```
> mean(commute)
[1] 26.6
> sd(commute)
[1] 5.660781
> quantile(commute, c(0.25, 0.75), type = 6)
 25%   75%
22.50 30.75
> IQR(commute, type = 6) / sd(commute)
[1] 1.457396
```

`qqnorm(commute)`; `qqline(commute)` produces the plot above.

Problem 2: Normal Approximation to the Binomial (15 points)

$X \sim \text{binomial}(n = 900, p = 0.20)$.

(a)

$$\mu = np = 900(0.20) = 180, \quad \sigma = \sqrt{np(1-p)} = \sqrt{900(0.20)(0.80)} = \sqrt{144} = 12.$$

The approximation is appropriate because both $np = 180$ and $n(1-p) = 720$ are much larger than 15, so the binomial histogram is close to bell-shaped. Hence $X \dot{\sim} N(\mu = 180, \sigma = 12)$.

(b) Standardize using $\mu = 180$ and $\sigma = 12$:

$$P(X \leq 165) \approx P\left(Z \leq \frac{165 - 180}{12}\right) = P(Z \leq -1.25) = 0.1056.$$

(c)

$$\begin{aligned} P(165 \leq X \leq 195) &\approx P\left(\frac{165 - 180}{12} \leq Z \leq \frac{195 - 180}{12}\right) = P(-1.25 \leq Z \leq 1.25) \\ &= 0.8944 - 0.1056 = 0.7888. \end{aligned}$$

(d) Exact binomial values in R:

```
> pbinom(165, 900, 0.2)
[1] 0.1126878
> pbinom(195, 900, 0.2) - pbinom(164, 900, 0.2)
[1] 0.8035798
```

	Normal approximation	Exact binomial	Difference
$P(X \leq 165)$	0.1056	0.1127	0.0071
$P(165 \leq X \leq 195)$	0.7888	0.8036	0.0148

The approximation is good — both answers are within about 0.015 of the exact binomial, and either would lead to the same practical conclusion. The small gap arises because a smooth normal curve is being used to describe a discrete count; the agreement improves as n grows.

Problem 3: The Exponential Distribution (15 points)

(a) Since arrivals follow a Poisson process with $\lambda = 3$ per hour, the waiting time until the next arrival is exponential with

$$\theta = \frac{1}{\lambda} = \frac{1}{3} \text{ hour}, \quad Y \sim \text{Exponential}(\theta = \frac{1}{3}).$$

$$\mu_Y = \theta = \frac{1}{3} \text{ hr} = 20 \text{ minutes}, \quad \sigma_Y = \theta = \frac{1}{3} \text{ hr} = 20 \text{ minutes}.$$

(For the exponential the mean and standard deviation are always equal.)

(b) 30 minutes = 0.5 hour:

$$P(Y \geq 0.5) = e^{-0.5/(1/3)} = e^{-1.5} = 0.2231.$$

(c) 10 minutes = 1/6 hour:

$$P(Y \leq \frac{1}{6}) = 1 - e^{-(1/6)/(1/3)} = 1 - e^{-0.5} = 1 - 0.6065 = 0.3935.$$

(d) Set the cumulative probability equal to 0.90:

$$1 - e^{-a/\theta} = 0.90 \implies e^{-3a} = 0.10 \implies -3a = \ln(0.10) \implies a = \frac{\ln 10}{3} = \frac{2.3026}{3} = 0.7675 \text{ hour}.$$

That is $0.7675 \times 60 = 46.1$ minutes. (In R: `qexp(0.9, rate = 3)`.)

(e) They are two descriptions of the same process. If the *number* of events in a fixed time period is Poisson with rate λ per unit time, then the *time between* consecutive events is exponential with mean $\theta = 1/\lambda$. A high arrival rate (large λ) means short typical gaps (small θ), and vice versa.

Problem 4: Sums of Independent Normals vs. the Sample Mean (15 points)

$X_i \sim N(\mu = 340, \sigma = 4)$, independent, $n = 12$.

(a) A sum of independent normals is normal, with the means and the *variances* adding:

$$\begin{aligned} E(T) &= \sum \mu_i = 12(340) = 4080 \text{ g}, & \text{Var}(T) &= \sum \sigma_i^2 = 12(4^2) = 192, \\ \sigma_T &= \sqrt{192} = 13.86 \text{ g}, & T &\sim N(4080, 13.86^2). \end{aligned}$$

Note that the standard deviations do **not** add ($12 \times 4 = 48$ is wrong).

(b)

$$P(T < 4060) = P\left(Z < \frac{4060 - 4080}{13.86}\right) = P(Z < -1.44) = 0.0749 \quad (\text{exact } z = -1.443 \Rightarrow 0.0745).$$

(c) $\bar{X} \sim N\left(340, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$ exactly, since the population is normal, with

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{12}} = \frac{4}{3.464} = 1.15 \text{ g}.$$

(d)

$$P(\bar{X} > 342) = P\left(Z > \frac{342 - 340}{1.1547}\right) = P(Z > 1.73) = 1 - 0.9582 = 0.0418 \quad (\text{exact } z = 1.732 \Rightarrow 0.0416).$$

(e) $\sigma = 4$ g describes how much *one individual bag* varies about 340 g. $\sigma_{\bar{X}} = 1.15$ g describes how much the *average of 12 bags* varies about 340 g. Averaging lets high and low bags offset one another, so the average is much more stable than a single bag. The reduction factor is $\sqrt{n} = \sqrt{12} = 3.46$, and indeed $4/3.46 = 1.15$.

Problem 5: The Central Limit Theorem (15 points)

Population: right-skewed, $\mu = 4.5$ min, $\sigma = 2.4$ min; $n = 36$.

(a) Not *exactly* normal — exact normality of $\bar{X}$ requires a normal population, and this population is strongly right-skewed. But by the **Central Limit Theorem**, for a random sample of size n from *any* population with mean μ and finite standard deviation σ , the sampling distribution of $\bar{X}$ is approximately

$$\bar{X} \sim N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right) \quad \text{when } n \text{ is large.}$$

Here $n = 36 \geq 30$ (and > 25), so the condition is satisfied and $\bar{X}$ is **approximately normal**.

(b)

$$\mu_{\bar{X}} = \mu = 4.5 \text{ min}, \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2.4}{\sqrt{36}} = \frac{2.4}{6} = 0.40 \text{ min.}$$

(c)

$$P(\bar{X} > 5.0) = P\left(Z > \frac{5.0 - 4.5}{0.40}\right) = P(Z > 1.25) = 1 - 0.8944 = 0.1056.$$

(d)

$$\begin{aligned} P(4.2 < \bar{X} < 5.0) &= P\left(\frac{4.2 - 4.5}{0.40} < Z < \frac{5.0 - 4.5}{0.40}\right) = P(-0.75 < Z < 1.25) \\ &= 0.8944 - 0.2266 = 0.6678. \end{aligned}$$

(e) With $n = 144$:

$$\sigma_{\bar{X}} = \frac{2.4}{\sqrt{144}} = \frac{2.4}{12} = 0.20 \text{ min.}$$

The sample size increased by a factor of 4 ($36 \rightarrow 144$) while the standard error decreased by a factor of 2 ($0.40 \rightarrow 0.20$). In general $\sigma_{\bar{X}} = \sigma/\sqrt{n}$, so the standard error is inversely proportional to $\sqrt{n}$: **to cut the standard error in half you must quadruple the sample size.**

Problem 6: Sampling Distribution of a Sample Proportion (10 points)

$p = 0.60$, $n = 400$.

(a) Large-sample check: $np = 400(0.60) = 240 \geq 15$ and $n(1 - p) = 400(0.40) = 160 \geq 15$, so the normal approximation applies.

$$\mu_{\hat{p}} = p = 0.60, \quad \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.60)(0.40)}{400}} = \sqrt{0.0006} = 0.0245,$$

$$\hat{p} \sim N(0.60, 0.0245^2).$$

(b)

$$P(\hat{p} > 0.65) = P\left(Z > \frac{0.65 - 0.60}{0.0245}\right) = P(Z > 2.04) = 1 - 0.9793 = 0.0207.$$

(c)

$$\begin{aligned} P(0.57 < \hat{p} < 0.63) &= P\left(\frac{0.57 - 0.60}{0.0245} < Z < \frac{0.63 - 0.60}{0.0245}\right) = P(-1.22 < Z < 1.22) \\ &= 0.8888 - 0.1112 = 0.7776 \quad (\text{exact } z = \pm 1.2247 \Rightarrow 0.7793). \end{aligned}$$

So roughly 78% of samples of 400 students would give a sample proportion within 3 percentage points of the true 60%.

Suggested Point Allocation (90 points)

Problem	Topic	Points	Breakdown
1	Normality & Q-Q plot	20	(a) 3, (b) 4, (c) 4, (d) 4, (e) 2, (f) 3
2	Normal approx. to binomial	15	(a) 4, (b) 4, (c) 4, (d) 3
3	Exponential distribution	15	(a) 4, (b) 3, (c) 3, (d) 3, (e) 2
4	Sums of normals vs. mean	15	(a) 4, (b) 3, (c) 3, (d) 3, (e) 2
5	Central Limit Theorem	15	(a) 4, (b) 3, (c) 3, (d) 3, (e) 2
6	Sampling distribution of $\hat{p}$	10	(a) 4, (b) 3, (c) 3

Common errors to watch for. Problem 2: using $\sigma = np(1-p) = 144$ instead of $\sqrt{np(1-p)} = 12$. Problem 3: mixing hours and minutes, or using $\lambda = 3$ where $\theta = 1/3$ belongs. Problem 4: adding standard deviations instead of variances in part (a), or confusing σ_T with $\sigma_{\bar{X}}$. Problem 5(a): claiming $\bar{X}$ is exactly normal. Problem 6: using $\sigma = \sqrt{pq}$ without dividing by n .