Suffolk County Community College

**Michael J. Grant Campus**

# Department of Mathematics

---

**Spring 2017**

# MAT 103
# Statistics I

## Final Exam: Solutions and Answers

---

**Instructor:**

Name: Alexander Kasiukov

Office: Suffolk Federal Credit Union Arena, Room A-109

Phone: (631) 851-6484

Email: kasiuka@sunysuffolk.edu

Web Site: `http://kasiukov.com`

---

**Problem 1.** According to the Himalayan Database [1] which contains 1921–2023 data on Mount Everest expeditions [2] two routes are most popular among those who want to reach the summit. Namely, among $2,306$ such expeditions recorded in the database, $1,276$ used the Southeast Ridge/South Corridor from Nepal, and 780 followed the Northeast Ridge/North Corridor from China (with the next most popular route, the South Pillar/South East Ridge, having been used by only 13 expeditions) [3].

**(1).** Assuming that the popularity of routes does not change each season, where would you look for a friend who is trying to reach the summit of Mount Everest, if you know nothing else about their location? What are the chances you would be looking for them in the right place? Round the answer to the nearest whole percent.

*Space for your solution:*

The most likely place to find them is the Southeast Ridge/South Corridor, and the probability they are using it can be estimated, based on the data given, as

$$\frac{1,276}{2,306} \approx 55\%.$$

**(2).** Would the answer change if you happen to know that your friend uses the Southeast Ridge/South Corridor or the Northeast Ridge/North Corridor?

*Space for your solution:*

The choice of where to look for them would not change — the Southeast Ridge/South Corridor is still the most likely place to find them — but now the probability they are there will be higher:

$$\frac{1,276}{1,276+780} \approx 62\%.$$

[1] `http://www.himalayandatabase.com/downloads.html` accessed May 20, 2024

[2] and covers all known attempts on 406 Himalayan peaks — including Mount Everest — from 1905 to 2023

[3] Consider these counts — as well as all other numbers appearing in this problem — to be (very close) approximations rather than the exact representation of reality, since we are routinely ignoring some border cases, like the fact that some — very few – expeditions used different routes on ascend and descent.

**(3).** Among the individual climbers who used the Southeast Ridge/South Corridor, 8,096 succeeded and 6,323 failed to reach the summit. Among those who followed the Northeast Ridge/North Corridor, the numbers were 3,505 for successes and 3,652 for the failed attempts. Find the probabilities of success for an individual climber taking each of these routes.

*Space for your solution:*

$$P\left(\begin{array}{c}\text{climber}\\ \text{reaches}\\ \text{summit}\end{array}\middle|\begin{array}{c}\text{climber uses}\\ \text{Southeast Ridge/}\\ \text{South Corridor}\end{array}\right) = \frac{8,096}{8,096+6,323} \approx 56\%$$

$$P\left(\begin{array}{c}\text{climber}\\ \text{reaches}\\ \text{summit}\end{array}\middle|\begin{array}{c}\text{climber uses}\\ \text{Northeast Ridge/}\\ \text{North Corridor}\end{array}\right) = \frac{3,505}{3,505+3,652} \approx 49\%$$

**(4).** During year 2000 season, 55 climbers attempted to reach the summit of Mount Everest via the Northeast Ridge/North Corridor. If their success rate was independent from each other and constant year-to-year, how many of them do you expect to reach the summit?

*Space for your solution:*

Based on the Law of Large Numbers, the number of successful climbs should be around

$$\mu = 55 \cdot 49\% \approx 27.$$

**(5).** In the same situation as before, find the 95% confidence interval for the number of successful climbs.

*Space for your solution:*

The number of successes $X$ has Binomial Distribution: $X \sim B(n = 55, p = 0.49)$ with mean $\mu = np \approx 26.94$, $q = 1 - p = 0.51$, and standard deviation $\sigma = \sqrt{npq} \approx 3.7$. Since both $np > 5$ and $nq > 5$, the Central Limit Theorem permits approximation of $X$ by normal distribution[a] $Y \sim \mathcal{N}(\mu = 26.94, \sigma = 3.7)$:

$$95\% = \boxed{Z \sim \mathcal{N}(\mu = 0, \sigma = 1)} = P\Big(Z \in [-z^*, z^*]\Big) = \boxed{\text{95\% critical value for } Z \text{ is } z^* = 1.96} =$$
$$= P\Big(Z \in [-1.96, 1.96]\Big) = \boxed{\text{for } n = 55,\ \mu = np \approx 26.94,\ q = .51, \text{ and } \sigma = \sqrt{npq} \approx 3.7} =$$
$$= P\Big(Y \in \big[\mu - z^*\sigma,\ \mu + z^*\sigma\big]\Big) \approx P\Big(Y \in [19.73, 34.27]\Big) \approx P\Big(X \in [20, 34]\Big).$$

[a] This can also be computed directly without normal approximation, giving $P\Big(X \in [20, 34]\Big) \approx 96\%$.

**(6).** With all assumptions of the previous two sub-problems in effect, what is the probability that 37 or more among them reach the summit? Can you make an upper estimate of the answer before computing it?

*Space for your solution:*

In view of the 95% confidence interval $[20, 34]$ found in the previous sub-problem, the probability of 37 or more climbers being successful must be lower than 2.5%.

As discussed earlier, the number of successes $X$ has Binomial Distribution: $X \sim B(n = 55, p = 0.49)$ with mean $\mu = np \approx 26.94$, $q = 0.51$, and standard deviation $\sigma = \sqrt{npq} \approx 3.7$. Since both $np > 5$ and $nq > 5$, as stated before the Central Limit Theorem permits approximation of $X$ by normal distribution $Y \sim \mathcal{N}(\mu = 26.94, \sigma = 3.7)$:

$$P(X \geq 37) = \boxed{\text{use normal approximation with continuity correction}} =$$

$$\approx P(Y \geq 36.5) = P\left(\frac{Y-\mu}{\sigma} \geq \frac{36.5-\mu}{\sigma}\right) = \boxed{\text{define } Z = \tfrac{Y-\mu}{\sigma}} =$$

$$\approx P(Z \geq 2.71) = \boxed{Z \sim \mathcal{N}(\mu = 0, \sigma = 1) \text{ is standard normal}} =$$

$$\approx 0.005 \text{ or half percent — virtually impossible.}$$

**(7).** The actual number of climbers who successfully reached the summit of Mount Everest via the Northeast Ridge/North Corridor in the year 2000 was 55 out of 55. How can you explain this apparent discrepancy with your result in the previous sub-problem?

*Space for your solution:*

Vanishingly small chance of an event with less than one half percent probability notwithstanding, this can be explained by several factors.

The main factor is the fact that the success rate does not stay constant over the years: as the climbing technology, communication, weather forecasting, organization of teams, and local infrastructure improve with time, both the individual and team success become more and more likely with each passing year.

The second factor is the fact that the success is dependent upon a particular year perhaps in the way related to the weather pattern: there are "good" and "bad" years in terms of accidents, fatalities and general failures.

The third factor is that the success of an individual climber is actually dependent upon the success of other climbers in the same team.

**(8).** Two teams attempt to climb Mount Everest: one using the South, and another — the North root. Assuming that success of the team is as likely as that of an individual, and the success of one team is independent from the success of the other, what is the probability that both teams reach the summit?

*Space for your solution:*

$$P\left(\begin{pmatrix}\text{team that uses}\\ \text{South route}\\ \text{succeeds}\end{pmatrix} \cap \begin{pmatrix}\text{team that uses}\\ \text{North route}\\ \text{succeeds}\end{pmatrix}\right) = \boxed{\text{product rule}} =$$

$$= P\begin{pmatrix}\text{team that uses}\\ \text{South route}\\ \text{succeeds}\end{pmatrix} \cdot P\left(\begin{matrix}\text{team that uses}\\ \text{North route}\\ \text{succeeds}\end{matrix}\,\middle|\,\begin{matrix}\text{team that uses}\\ \text{South route}\\ \text{succeeds}\end{matrix}\right) = \boxed{\text{independence}} =$$

$$= P\begin{pmatrix}\text{team that uses}\\ \text{South route}\\ \text{succeeds}\end{pmatrix} \cdot P\begin{pmatrix}\text{team that uses}\\ \text{North route}\\ \text{succeeds}\end{pmatrix} =$$

$$= \boxed{\text{assuming probability of individual success is the same as that for the team}} =$$

$$= 56\% \cdot 49\% \approx 27\%.$$

**(9).** What is the probability that at least one of the teams will reach the summit?

*Space for your solution:*

$$P\left(\begin{pmatrix}\text{team that uses}\\ \text{South route}\\ \text{succeeds}\end{pmatrix} \cup \begin{pmatrix}\text{team that uses}\\ \text{North route}\\ \text{succeeds}\end{pmatrix}\right) = \boxed{\text{inclusion-exclusion formula}} =$$

$$= P\begin{pmatrix}\text{team that uses}\\ \text{South route}\\ \text{succeeds}\end{pmatrix} + P\begin{pmatrix}\text{team that uses}\\ \text{North route}\\ \text{succeeds}\end{pmatrix} -$$

$$- P\left(\begin{pmatrix}\text{team that uses}\\ \text{South route}\\ \text{succeeds}\end{pmatrix} \cup \begin{pmatrix}\text{team that uses}\\ \text{North route}\\ \text{succeeds}\end{pmatrix}\right) =$$

$$= \boxed{\text{results of previous sub-problems}} =$$

$$\approx 56\% + 49\% - 27\% = 78\%.$$

**Problem 2.** This problem will introduce you to the *Simpson's Paradox.*

1314 women took part in a study [4] of thyroid disease that was conducted in 1972-1974 in Newcastle, United Kingdom. A follow-up study of the same subjects [5] took place nearly thirty years later.

**(1).** The subjects of the study were classified according to their smoking habits (current smokers at the time of the original 1970's study or those who never smoked) and according to their survival status 20 years after the original study. The outcomes are summarized in the following table:

| | **Smoker** | **Non-smoker** |
|---|---|---|
| **Dead** | 139 | 230 |
| **Alive** | 443 | 502 |

Based on this table, determine if smoking has positive or negative effect on survival.
Hint: compute and compare the conditional probabilities:

$$P(\text{Alive}|\text{Smoker})$$

$$P(\text{Alive}|\text{Non-smoker})$$

*Space for your solution:*

$$P(\text{Alive}|\text{Smoker}) = \frac{443}{139+443} \approx 76\%$$

$$P(\text{Alive}|\text{Non-smoker}) = \frac{502}{230+502} \approx 69\%$$

Thus it seems that smoking has positive effect on survival.

[4] W. M. G. Tunbridge, D. C. Evered, D. Appleton, M. Brewis, F. Clark
"The Spectrum of Thyroid Disease in a Community: The Whickham Survey",
Clinical Endocrinology, Volume 7, Issue 6, December 1977, Pages 481-493
`http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2265.1977.tb01340.x/abstract`

[5] David R. Appleton, Joyce M. French and Mark P. J. Vanderpump
"Ignoring a Covariate: An Example of Simpson's Paradox",
The American Statistician, Volume 50, Number 4, November 1996, Pages 340-341
`http://www.jstor.org/stable/2684931?seq=1#page_scan_tab_contents`

**(2).** The subjects were further classified according to their age at the time of the original study. The outcomes for women aged 18 to 64 are summarized in this table:

| **Age 18 to 64** | **Smoker** | **Non-smoker** |
|---|---|---|
| **Dead** | 97 | 65 |
| **Alive** | 436 | 474 |

Determine if smoking has positive effect on survival of women in this age group.

*Space for your solution:*

$$P(\text{Alive}|\text{Smoker aged 18 to 64}) = \frac{436}{97+436} \approx 82\%$$

$$P(\text{Alive}|\text{Non-smoker aged 18 to 64}) = \frac{474}{65+474} \approx 88\%$$

Thus it seems that smoking has negative effect on survival within the 18 to 64 years of age group.

**(3).** The outcomes for women aged 65 and above are summarized in this table:

| **Age 65 and above** | **Smoker** | **Non-smoker** |
|---|---|---|
| **Dead** | 42 | 165 |
| **Alive** | 7 | 28 |

Determine if smoking has positive effect on survival of women in this age group.

*Space for your solution:*

$$P(\text{Alive}|\text{Smoker aged 18 to 64}) = \frac{7}{42+7} \approx 14.3\%$$

$$P(\text{Alive}|\text{Non-smoker aged 18 to 64}) = \frac{28}{165+28} \approx 14.5\%$$

Thus it seems that smoking has (a very slight) negative effect on survival within the 65 years and above age group.

**(4).** What conclusion can you draw from this consideration: does smoking improve or harm survival chances? If smoking is beneficial, why it is not shown by the analysis of age groups? If smoking is harmful, why does it contradict the outcome for the combined analysis (that ignores age)?

*Space for your solution:*

The main predictor of survival over the 20 year period is age. Among the women studied, younger women (who have lower mortality) had higher percentage of smokers, and older women (who have higher mortality) had lower percentage of smokers. In the aggregated analysis, the stronger effect of age on survival masked the effect of smoking status, creating the illusion of beneficial effect of smoking on survival. The smokers in the aggregated sample died at a lower rate not because smoking was beneficial, but because smokers tended to be younger. When the two age groups are analyzed separately, the effect of smoking is isolated from the effect of age, and the negative effect of smoking on survival becomes apparent.

**Problem 3.** The FAA Aircraft Weight and Balance Handbook [6] includes different weight and balance computation procedures for different types of aircraft. Generally speaking, the small aircraft procedures require the pilot(s) and passengers to be weighted, while the large aircraft computation is permitted to use the "standard average weight" figure instead.

**(1).** What can be the basis for this disparity in weight and balance procedures used for different aircraft sizes?

*Space for your solution:*

The standard deviation of the mean for the sample size $n$ (known as the standard error of the sample mean) is

$$\frac{\sigma}{\sqrt{n}},$$

where $\sigma$ is the standard deviation of the population. As the sample size increases, the numerator does not change, but denominator becomes bigger, thus making the whole fraction smaller. This means that as the sample size increases, the sample mean becomes more and more predictable.

**(2).** The "Standard Average Weight" procedure is permitted for use in weight and balance calculation for "large cabin aircraft", namely the aircraft with 71 or more passenger seats. From 1938, the FAA uses 170 pounds as the standard average weight for passengers. [7] Assume that 170 pounds is the actual average weight of the population, the standard deviation of the weight is 27 pounds, and the weight is normally distributed. Determine the standard error of the mean $\sigma_{\bar{x}}$ in passenger weight when 9 passengers load a small plane.

*Space for your solution:*

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{27}{\sqrt{9}} = \frac{27}{3} = 9$$

**(3).** With the assumptions of previous sub-problems in effect, determine the standard error of the mean $\sigma_{\bar{x}}$ in passenger weight when 81 passengers load a large cabin aircraft.

*Space for your solution:*

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{27}{\sqrt{81}} = \frac{27}{9} = 3$$

[6] `https://www.faa.gov/sites/faa.gov/files/2023-09/Weight_Balance_Handbook.pdf` accessed August 25, 2024

[7] However, the 05/05/2019 FAA Advisory Cricular AC 120-27F `https://www.faa.gov/documentLibrary/media/Advisory_Circular/AC_120-27F.pdf` accessed August 25, 2024 charges the airlines to undertake surveys in order to update the standard average passenger weights.

**(4).** Suppose an airline wants to conduct a survey of passenger weight to determine if the standard average weight needs be increased. State the null and alternative hypothesis for this test.

*Space for your solution:*

$$H_0: \ \mu = 170$$
$$H_1: \ \mu > 170$$

**(5).** The airline measured the actual weight of randomly selected 100 passengers. In this particular sample, the average weight was 180 pounds and the sample standard deviation $\bar{\sigma}$ was 40 pounds. Calculate the test statistic for the right-tailed $t$-test.

*Space for your solution:*

$$t = \frac{\bar{x} - \mu}{\left(\frac{\bar{\sigma}}{\sqrt{n}}\right)} = \frac{180 - 170}{\left(\frac{40}{\sqrt{100}}\right)} = \frac{10}{4} = 2.5$$

**(6).** Find the $p$-value, i.e. the attained level of significance, for this test.

*Space for your solution:*

For the right-tailed $t$-test with $n - 1 = 100 - 1 = 99$ degrees of freedom we get

$$p = 0.00705 \approx \frac{1}{140}$$

**(7).** What conclusion should the airline draw from this experiment?

*Space for your solution:*

If the average passenger weight were really 170 pounds, the probability of getting these (or more extreme) results would be around 1 in 140, which is less than one percent. So, even with 1% level of significance, the airline can reject the null hypothesis.