This exam has 25 questions of 4 points each. All answers have been rounded-off so if your calculated answer differs from the given options slightly, choose the closest answer.

You may bring along with you a 3x5 notecard and a statistical calculator.

Throughout the exam, remember to keep breathing.

---

**Q1** A 100 WUSTL students were surveyed to determine their satisfaction with housing allocated to them in 2010. Satisfaction was recorded on a scale of 1 to 3; with 1 as unsatisfied, 2 neutral and 3 satisfied.

At the end of the survey, it was observed that most of the freshmen girls were satisfied while on the whole sophomore boys were unhappy with their allocated housing. Also, the average age of students who were satisfied with their housing was 21.

What is the least number of variables needed in the survey to form such conclusions?

A) 100 Categorical variables
B) 3 Categorical variables
C) 4 Categorical variables, 1 Quantitative variable
D) 3 Categorical variables, 1 Quantitative variable **[Gender, satisfaction, college-year; Age]**
E) 2 Categorical variables, 2 Quantitative variables
F) 2 Categorical variables, 1 Quantitative variable
Q2. A poll was taken of food that people ate in restaurants on the Loop and the percentages plotted in the following pie-chart. What seems wrong with it?

![Pie Chart]

- A) Nothing, the pie-chart is fine.
- B) Area principle is followed but percentages adding up to 100% seems wrong. (**)
- C) Area principle is followed and percentages adding up to 100% seems correct.
- D) Area principle is violated and percentages adding up to 100% seems wrong.
- E) Area principle is violated but percentages adding up to 100% seems correct.
- F) Variable is quantitative so a histogram should have been used, not pie-chart.

Q3. 300 Spartans went to fight a war with Xerxes. All the Spartans died, the reasons for their death have been listed below:

<table>
<thead>
<tr>
<th>Number of Spartans</th>
<th>Death by sword-wounds</th>
<th>Death by arrow-wounds</th>
<th>Death by indigestion</th>
<th>Death by steroid overdose</th>
<th>Death by chocolate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>85</td>
<td>113</td>
<td>24</td>
<td>78</td>
<td>0</td>
</tr>
</tbody>
</table>

What percentage of the Spartans died because of wounds suffered in battle?

- A) 50%
- B) 58%
- C) 60%
- D) 66%**
- E) 70%
- F) 72%
Q4. A company must decide which of the two delivery services they will contract with. During a recent trial period they shipped numerous packages with each service and have kept track of how often deliveries did not arrive on time. From the below given data, which delivery service should the company choose and why?

<table>
<thead>
<tr>
<th>Delivery Service</th>
<th>Type of Service</th>
<th>Number of Deliveries</th>
<th>Number of Late packages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pack Rats</td>
<td>Regular</td>
<td>100</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Overnight</td>
<td>400</td>
<td>28</td>
</tr>
<tr>
<td>Boxes R Us</td>
<td>Regular</td>
<td>400</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Overnight</td>
<td>100</td>
<td>16</td>
</tr>
</tbody>
</table>

A) Pack Rats because their overall percentage of late deliveries is lesser.
B) Boxes R Us because their overall percentage of late deliveries is lesser.
C) Pack Rats because they have better percentages for both regular and overnight deliveries
   **[This is a case of Simpson’s Paradox where overall percentages are misleading]**
D) Boxes R Us because they have better percentages for both regular and overnight deliveries
E) Both are equally good, so either is fine.
F) Insufficient data to decide

Q5. Jack and Jill played a game of running up a hill and tumbling down. Whoever tumbled farther was considered the winner. They played this game on two hills called Olympus and Everest. Their victories are summarised in the following table:

<table>
<thead>
<tr>
<th></th>
<th>Olympus</th>
<th>Everest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jack</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>Jill</td>
<td>13</td>
<td>12</td>
</tr>
</tbody>
</table>

What percentage of Jack’s victories was on Olympus?

A) 10%
B) 20%
C) 30%
D) 40%** [10/25*100]
E) 43.48%
F) 45%
Q6. A survey was taken of 57 students for reaction to the smoking-ban on campus. The following data was tabulated:

<table>
<thead>
<tr>
<th></th>
<th>Smoker</th>
<th>Non-Smoker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td>Female</td>
<td>10</td>
<td>20</td>
</tr>
</tbody>
</table>

If there are 10 male students who oppose the ban and 10 females who support it, what percentage of female students oppose the ban?

A) 10%
B) 20%
C) 40%
D) 50%
E) 60%**[18/30*100; Look at the table of Gender-Opinion and fill in the gaps using marginal distributions]
F) Insufficient data do decide

Q7. What is the interquartile range of the following distribution:

X: 126, 127, 128, 129, 130, ..., 222, 223, 224, 225 (integers from 126 through 225)

A) 45
B) 50**[Q1=150.5, Q3=200.5]
C) 55
D) 25
E) 25.5
F) 100
Q8. John lived in a city known for being particularly windy. He recorded the wind-speeds for 50 days in summer as well as in winter and plotted their histograms. However, he forgot to write the values on the x-axis. If we knew that each bin for the bars in summer and in winter was of the same width (i.e., they both had the same scale), what can be said of the standard-deviation of the windspeeds?

A) Standard-deviation is more in winter than in summer**
B) Standard deviation is more in summer than in winter
C) Standard deviation is the same in summer and winter
D) Standard deviation is more in winter than in summer only if they have the same mean
E) Standard deviation is more in summer than in winter only if they have the same mean
F) Without knowing the actual values of the data, we can’t compare standard deviations
Q9. Following is a stem-and-leaf display for a quantitative variable:

```
8 | 8 2 6 3 4 5
7 | 2 2 3 4 3 4 2
6 | 2 4 2 3 8
5 | 4 2 0
4 | 3 3
3 | 1
2 | 0
1 | 1
0 |
```

What can be said about the modes of this variable?

A) Distribution is unimodal**
B) Distribution is bimodal
C) Distribution is multimodal
D) Distribution is uniform
E) Depends on the units used
F) Insufficient data to decide

Q10. Is the mean or the median a better description of the centre for the above distribution?

A) Insufficient data to decide
B) The mean
C) The median** [as it is skewed]
D) Both are equally good
E) Neither
F) It depends on the units used

Q11. Suppose a quantitative variable has a unimodal distribution with no outliers. Let a, b, c, d, e be its 5-number summary with \( a < b < c < d < e \). If \( m \) is the mean and \( m \) is much larger than \( c \), then what can be said about the shape of the distribution?

A) Most likely, the distribution is symmetric about the median
B) Most likely, the distribution is symmetric about the mean
C) Most likely, the distribution is skewed to the right ** [as mean is more than median]
D) Most likely, the distribution is skewed to the left
E) Most likely, the distribution is uniform
F) Most likely, the distribution is a dog
**Q12.** The following were the number of sick-days taken by the employees of a company in a year:

9, 2, 10, 8, 3, 5, 3, 1, 2, 32, 6, 2, 9

What is a good measure of center and spread for this distribution?

A) Mean and Standard deviation,
B) Median and IQR**
C) Mean and IQR
D) Median and Standard deviation
E) Mean and Median
F) IQR and Standard deviation

**Q13.** Wayne Gretzky scored more points than anyone who ever played professional hockey. Here are the number of games Gretzky played during each season:

79, 80, 80, 80, 74, 80, 74, 79, 64, 78, 73, 78, 74, 45, 81, 48, 80, 82, 82, 70

What is the mean and standard deviation of the number of games played by Gretzky in a season?

A) 68.43 and 8.40
B) 69.86 and 9.34
C) 69.86 and 10.80
D) 74.35 and 10.53**
E) 74.35 and 10.94
F) 76.35 and 10.94

**Q14.** If a quantitative variable takes the following values: 1, 1, 2, 3, 3. Then what can be said about the whiskers of its boxplot:

A) It has no whiskers**[as Q1=min, Q2=max]
B) Top whisker extends to 6, bottom whisker extends to -2
C) Top whisker extends to 5, bottom whisker extends to -1
D) Top whisker extends to 6, bottom whisker extends to 0
E) Top whisker extends to 5, bottom whisker extends to 0
F) Top whisker extends to 4, bottom whisker extends to 0
Q15. The average number of hours undergraduate students sleep at night is surveyed. The survey is divided into two groups, one during the holidays at their homes and the other during term in their dorms. The two groups give the following boxplots:

The following conclusions were made from this box-plot:
(i) The maximum sleeping-duration overall was 12 hours
(ii) The distribution when at home is skewed to the right
(iii) The distribution when in dorms is skewed to the right
(iv) The minimum sleeping-duration when at home was 6 hours

Which of the following is correct?

A) All (i), (ii), (iii) and (iv) are true.
B) Only (ii) is true
C) Only (iii) is true
D) Only (iii) and (iv) are true
E) Only (ii) and (iv) are true**
F) Only (iii) is false
Q16. A computer was used to record the number of faulty components in an aircraft engine, under repeated runs of a new prototype. A mean of 100 with a standard deviation of 10 was obtained.

It was later realised that due to a programming bug a trailing zero was attached to every count, so that 1 faulty component was recorded as 10, 5 was recorded as 50, 13 was recorded as 130 and so on.

Without repeating the experiment, can you determine the correct mean and standard deviation?

A) No  
B) Yes, the correct mean is 1 and standard deviation 1  
C) Yes, the correct mean is 10 and standard deviation 1**  
D) Yes, the correct mean is 10 and standard deviation 10  
E) Yes, the correct mean is 100 and standard deviation 100  
F) Yes, the correct mean is 1000 and standard deviation 100

Q17. If the mean score on a statistics exam was 80 with a standard deviation of 10, what would be the z-score of a student be who had a grade of 55 ?

A) -3.0  
B) -2.5**  
C) -2  
D) -1.5  
E) 0  
F) 2.5

Q18. Assume that the weights of people have a mean of 180 lbs with a standard deviation of 40 lbs. Any-one weighing more than 260 lbs is considered to be obese. If the weights of Martians had a mean of 1000lbs with a standard deviation of 100 lbs, by a similar consideration what must a Martian weigh in order be considered obese?

A) More than 260 lbs  
B) More than 1100 lbs  
C) More than 1200 lbs** (2sigma more than mean)  
D) More than 1260 lbs  
E) More than 1300 lbs  
F) Insufficient data to decide
Q19. The heights of ten-thousand male students were measured by a basketball coach and 15 students were found to have a height of more than 8 feet, while the mean height was found to be 6 feet. If only the highest 2.5% were to be considered for the team, what is the minimum height the coach set in order to qualify for the team?

A) 6.25 feet  
B) 6.4 feet  
C) 6.65 feet  
D) 7 feet  
E) 7.33 feet** [3sigma=(8-6) by the 68-95-99.7% rule, so we want 6 +2sigma]  
F) 7.8 feet

Q20. If weights of girls at birth had a mean of 7.5 lbs and a standard deviation of 1 lbs, what percentage of girls weigh more than 9lbs?

A) 32%  
B) 16%  
C) 8.27%  
D) 6.68%**  
E) 2.42%  
F) 1.25%

Q21. A survey taken in St Louis, in 2000, collected the following data: There were 2,700,000 residents of which 100,000 were under-age (below 18). If the standard deviation of the ages was 12.1 what is the mean age of a resident of St Louis?

A) Around 20 years  
B) Around 25 years  
C) Around 30 years  
D) Around 35 years  
E) Around 40 years**  
F) Around 45 years
Q22. From the following scatter-plot, which of the following values might be its correlation?

![Scatter-plot image]

a) 0.983  
b) 0.675  
c) 0.007**  
d) -0.896  
e) -1  
f) Depends on the units used

Q23. A survey was conducted in the US and other Western European countries to determine the percentage of teenagers who had used marijuana and other drugs. The results are summarised in this table:

<table>
<thead>
<tr>
<th></th>
<th>Czech</th>
<th>Denmark</th>
<th>England</th>
<th>Finland</th>
<th>Italy</th>
<th>Norway</th>
<th>Portugal</th>
<th>Spain</th>
<th>USA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marijuana</td>
<td>22</td>
<td>17</td>
<td>40</td>
<td>5</td>
<td>19</td>
<td>6</td>
<td>7</td>
<td>53</td>
<td>34</td>
</tr>
<tr>
<td>Other drugs</td>
<td>4</td>
<td>3</td>
<td>21</td>
<td>1</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>31</td>
<td>24</td>
</tr>
</tbody>
</table>

What is the correlation between percentage of teenagers who used marijuana and other drugs?

A) -0.89  
B) -0.92  
C) 0.03  
D) 0.75  
E) 0.88  
F) 0.94**
Q24. From the above study the following conclusions were drawn:

i) Marijuana usage leads teens to use of other drugs as well
ii) There is a strong association between marijuana use and use of other drugs
iii) There is no association between marijuana use and use of other drugs
iv) There is a negative association between marijuana use and use of other drugs
v) Marijuana usage discourages them to take up other drugs

Which of the following is true:

A) (i) and (ii)
B) Only (i)
C) Only (ii)**
D) (i) and (iii)
E) (iv) and (v)
F) Only (iv)

Q25. For the following scatter-plot which do you think is a better measure to study association between these two variables:

a) Coefficient of correlation
b) Kendall’s Tau**
c) Both should give same strengths of association
d) It depends on the units used
e) Whichever is closer to zero
f) Insufficient data to decide

All the best!