EXAM SOLUTION

1 Winter games [19 pt]

a) **[1 pt]** $p_1^5 + \binom{5}{4} p_1^4 * (1-p_1) + \binom{5}{3} p_1^3 * (1-p_1)^2 = 0.328 + 5*0.41*0.2 + 10*0.512*0.04 = 0.328 + 0.41 + 0.205 = 0.943$

b) **[2 pt]** 4 see above. It makes sense as the expectation value of the Binomial distribution is np.

c) **[7 pt]**

$X =$	0	1	2	3	4	5
$Y =$						
0						
1	$P_{X=0}P_{Y=1}$					
2	$P_{X=0}P_{Y=2}$	$P_{X=1}P_{Y=2}$				
3	$P_{X=0}P_{Y=3}$	$P_{X=1}P_{Y=3}$	$P_{X=2}P_{Y=3}$			
4	$P_{X=0}P_{Y=4}$	$P_{X=1}P_{Y=4}$	$P_{X=2}P_{Y=4}$	$P_{X=3}P_{Y=4}$		
5	$P_{X=0}P_{Y=5}$	$P_{X=1}P_{Y=5}$	$P_{X=2}P_{Y=5}$	$P_{X=3}P_{Y=5}$	$P_{X=4}P_{Y=5}$	

$P(Y > X) = P_{Y=5}(P_{X=4} + P_{X=3} + P_{X=2}) + P_{Y=4}(P_{X=3} + P_{X=2}) + P_{Y=3}P_{X=2} = 0.17*(0.41 + 0.20 + 0.05) + 0.36*(0.20 + 0.05) + 0.31*0.05 = 0.17*0.67 + 0.36*0.25 + 0.31*0.05 = 0.22$

d) **[2 pt]**

[1 pt] χ^2 test.

[1 pt] $H_0 : p_1 = 0.8$; $H_1 : p_1 \neq 0.8$.

e) **[10 pt]**

All terms in E with 0.01^2 and lower are ignored. That leaves:

[2 pt]
$E(0) = 10 * \binom{5}{0} * p_1^0(1-p_1)^5 = 10 * p_1^0(1-p_1)^5 = 0.0032$
$E(1) = 10 * \binom{5}{1} * p_1^1(1-p_1)^4 = 10*5*p_1^1(1-p_1)^4 = 0.064$
$E(2) = 10 * \binom{5}{2} * p_1^2(1-p_1)^3 = 10*10*p_1^2(1-p_1)^3 = 0.512$
$E(3) = 10 * \binom{5}{3} * p_1^3(1-p_1)^2 = 10*10*p_1^3(1-p_1)^2 = 2.05$
$E(4) = 10 * \binom{5}{4} * p_1^4(1-p_1)^1 = 10*5*p_1^4(1-p_1)^1 = 4.1$
$E(5) = 10 * \binom{5}{5} * p_1^5(1-p_1)^0 = 10*p_1^5(1-p_1)^0 = 3.28$

[2 pt] We compute the χ^2 Statistic as $\chi^2 = \sum_i (O_i - E_i)^2/E_i = (0.0032-0)^2/0.0032 + (0-0.064)^2/0.064 + (2-0.512)^2/0.512 + (3-2.05)^2/2.05 + (2-4.1)^2/4.1 + (3-3.28)^2/3.28 = 12.94$.

[3 pt] For 6 possible outcomes, we have $df = 5$ degrees of freedom. We look up in the table $q\chi_5^2(0.95) = 11.07$

[3 pt] As $\chi^2 > q\chi_5^2(0.95)$, we reject the null hypothesis. This data is incompatible with Susy shooting at her old $p_1 = 0.8$. The main deviation comes from low values of X, hence she actually got worse, not better. Her trainer is correct.

2 From the food lab [pt]

a) **[3 pt]** $\sum_i (X_i - 10)^2 = \sum_i X_i^2 - 20\sum_i X_i + \sum_i 100$
$\sum_i (X_i - 10)^2 = \sum_i X_i^2 - 20 * 40\bar{X} + 4000$
$\bar{X} = 1/800 * (\sum_i X_i^2 - \sum_i (X_i - 10)^2 + 4000) = 4.5225$

b) **[3 pt]**

[1 pt] We have $n = 40$ datapoints: $n * 0.18 = 7.2$ non integer. $q_{0.18} = X(8) = 2.7$

All other points are integer:

[0.66 pt] $q_{0.25} = 1/2(X(10) + X(11)) = 2.95$

[0.66 pt] $q_{0.5} = 1/2(X(20) + X(21)) = 4.35$

[0.66 pt] $q_{0.75} = 1/2(X(30) + X(31)) = 5.8$

c) **[5 pt]**

[2 pt] boxplot well drawn to proportions.

[0.5 pt] IQR=$q_{0.75} - q_{0.25} = 2.85$

[0.5 pt] box between $q_{0.75}$ and $q_{0.25}$

[0.5 pt] candlesticks to $q_{0.75} + 1.5IQR = 10.08$ and $q_{0.25} - 1.5 * IQR = -1.33/0$

[0.5 pt] no formal outliers

[1 pt] notice: quite symmetric distribution, median close to mean.

d) **[5 pt]**

[1 pt] recompute mean without datapoint: $\bar{X}_{new} = 1/39(40\bar{X} - X_{40}) = 4.4$

[1 pt] $n - 1 = 38$ in estimator for σ

[3 pt] $\sigma^2 = \frac{1}{38}\sum_{i=1}^{39}(x_i - \bar{x})^2 = \frac{1}{38}(\sum_{i=1}^{39} x_i^2 - 39\bar{X}_{new})$

$\sum_{i=1}^{39} x_i^2 = 963.83 - 9.2^2 = 879.19$

$\sigma^2 = 1/38(\sum_i^{40} x_i^2 - 84.64 - 39\bar{X}_{new}^2) = 3.24$, giving $\sigma = \sqrt{3.24} = 1.8$

[-1 pt] for confusing σ and σ^2

e) **[5 pt]**

[4 pt] $P(X > 8) = 1 - P(X < 8) = 1 - P(Z < (8 - 4.4)/1.8) = 1 - P(Z < 2) = 1 - 0.9772 = 2.3$

[1 pt] This exceeds the required acceptance probability of 1%, hence we cannot yet sell the product.

f) **[2 pt]** Except for the removed wrong measurement, not a single actual measurement exceeded the threshold. This is compatible with the data, as $E[X > 8] = nP(X > 8) = 0.9$. So on average, we do not expect to find yet an actual measurement above 8, yet the distribution of our data still lets us worry that products above 8% may be too likely. More datapoints are needed to be more certain.

3 Clean Power [27 pt]

a) **[2 pt]**

[1 pt] $\bar{X}_1 = 7.83$

[1 pt] $s_1 = 1.47$

b) **[2 pt]**

[1 pt] H_0: $\mu_1 \leq \mu_2$ $H_1 : \mu_1 > \mu_2$.

[1 pt] We do a one sided test as we want to know if our product beats regular panels. We do not care if it is much worse.

c) **[8 pt]**

[1 pt] We are not given a known variance, hence we have to perform a Welch test.

[2 pt] Welch's T-statistic is computed as $T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2+s_2^2}}\sqrt{6} = 1.872$.

[3 pt] degrees of freedom: $a = (\frac{s_1^2}{n} + \frac{s_2^2}{n})^2 = 1.79$ and $b = \frac{1}{2n-1}((\frac{s_1^2}{n})^2 + (\frac{s_2^2}{n})^2) = 0.217$. Then $df = round(a/b) = round(8.24) \approx 8$.

[1.5 pt] In this one-sided test, large values of T speak against H_0. $qt(p = 0.95, df = 8) = 1.86$.

[0.2 pt] As $T > qt(p = 0.95, df = 8)$, we do have (barely) statistically significant evidence that your green panels are more appealing than the standard ones.

d) **[7 pt]**

	df [1 pt]	SS [4 pt]	MS [1 pt]	F [1 pt]
Model	2	$6 * \sum_{i=1}^{3}(\bar{X}_i - \bar{X})^2 = 25$	12.5	F=MST/MSE=8.72
Error	15	$5 * \sum_{i=1}^{3} s_i^2 = 21.5$	1.43	
Total	17	46.5		

e) **[4 pt]**

The F-statistic of this model is 8.72.

[1 pt] We have $\nu_1 = 2$, $\nu_2 = 15$ degrees of freedom.

[1 pt] Table: $qF_{2,15}(p = 0.95) = 3.682$.

[2 pt] Our F falls well inside the critical region. We reject the null Hypothesis. As the mean of our product is well above those of the others, we have strong evidence that our product is significantly better than the other technologies.

f) **[4 pt]**

[1 pt] The T-test tells us that our product is slightly better than the regular solar cells. However, it is marginal, and more data is needed to be more certain.

[1 pt] The difference to the other technologies is quite striking. Clearly our product is superior to them, but as it is so similar to the normal solar cells, likely even the normal cells would beat the other technologies.

4 Climate change modeling[19 pt]

a) **[6 pt]**

[1 pt] (just writing down) $SSE = \sum_i(\hat{C}(T_i) - C_i)^2 = \sum_i(AT_i - U_i)^2$

[1 pt] (seeing concept) We find the coefficient A by minimizing the SSE: $d/dASSE = 0$

[4 pt] $\sum_i(AT_i - C_i)T_i = 0$ -> $A = \sum_i T_iC_i / \sum_i T_i^2 = 2.45$

b) **[5 pt]**

[1 pt] Exactly the same as in a), just replace T_i by T_i^2.

[4 pt] $B = \sum_i T_i^2 U_i / \sum_i T_i^4 = 0.985$

c) **[1 pt]** no. These models are independent, one is not a special case of the other (=nested).

d) **[7 pt]**

[3 pt] $SSE^{(1)} = \sum_i (AT_i - C_i)^2 = 14.23$

[3 pt] $SSE^{(2)} = \sum_i (BT_i^2 - C_i)^2 = 3.26$

[1 pt] The second model explains much more of the variance at the same number of parameters (1). It is the better model to the data.