考点 18　统计与统计案例

一、选择题

1．某学校近几年来通过“书香校园”主题系列活动，倡导学生整本阅读纸质课外书籍.下面的统计图是该校 2013 年至 2018 年纸质书人均阅读量的情况，根据统计图提供的信息，下列推断不合理的是（　　）



A．从 2013 年到 2016 年，该校纸质书人均阅读量逐年增长

B．2013 年至 2018 年，该校纸质书人均阅读量的中位数是 46.7 本

C．2013 年至 2018 年，该校纸质书人均阅读量的极差是 45.3 本

D．2013 年至 2018 年，该校后三年纸质书人均阅读量总和是前三年纸质书人均阅读量总和的 2 倍

【答案】D

【解析】对于 A，根据统计图分析可知，从 2013 年到 2016 年，该校纸质书人均阅读量分别是：15.5，38.5，43.3，58.4 是逐年增长的，故 A 是合理的；

对于 B，2013 年至 2018 年，该校纸质书人均阅读量按从小到大的顺序排列为：15.5，38.5，43.3，50.1，58.4，60.8，其中位数为 $\dfrac{43.3+50.1}{2}=46.7$ 本，故 B 是合理的；

对于 C，因为最大阅读量为 60.8 本，最小阅读量为 15.5 本，所以极差为 $60.8-15.5=45.3$ 本，故 C 是合理的；

对于 D，2013 年至 2018 年，该校后三年纸质书人均阅读量总和为 $58.4+50.1+60.8=169.3$ 本，前三年纸质书人均阅读量总和为 $15.5+38.5+43.3=97.3$ 本，$169.3\neq 97.3\times 2=194.6$，　故 D 是不合理的.

故选 D.

2．设某大学的女生体重 y（单位：kg）与身高 x（单位：cm）具有线性相关关系，根据一组样本数据（x_i，y_i）（i=1，2，…，n），用最小二乘法建立的回归方程为 $\hat{y}$=0.85x-85.71，则下列结论中不正确的是（　　）

A．y 与 x 具有正的线性相关关系

B．回归直线过样本点的中心（$\overline{x}$，$\overline{y}$）

C．若该大学某女生身高增加 1cm，则其体重约增加 0.85kg

D．若该大学某女生身高为 170cm，则可断定其体重必为 58.79kg

【答案】D

【解析】

根据 y 与 x 的线性回归方程为 $y=0.85x-85.71$，则

$\hat{b}=0.85>0$，y 与 x 具有正的线性相关关系，A 正确；

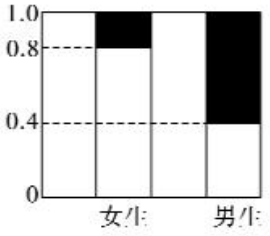
回归直线过样本点的中心（$\overline{x},\overline{y}$），B 正确；

该大学某女生身高增加 1cm，预测其体重约增加 0.85kg，C 正确；

该大学某女生身高为 170cm，预测其体重约为 $0.85\times170-85.71=58.79kg$，D 错误．

故选 D.

3．如图是调查某学校高三年级男女学生是否喜欢数学的等高条形图，阴影部分的高表示喜欢数学的频率．已知该年级男生女生各 500 名（所有学生都参加了调查），现从所有喜欢数学的同学中按分层抽样的方式抽取 32 人，则抽取的男生人数为（　　）



A．16　　B．32　　C．24　　D．8

【答案】C

【解析】由等高条形图可知：喜欢数学的女生和男生的比为 1:3，所以抽取的男生数为 24 人．

故选 C.

4．某学校为落实学生掌握社会主义核心价值观的情况，用系统抽样的方法从全校 2400 名学生中抽取 30 人进行调查.现将 2400 名学生随机地从 1～2400 编号，按编号顺序平均分成 30 组（1～80 号，81～160 号，…，2321～2400 号），若第 3 组与第 4 组抽出的号码之和为 432，则第 6 组抽到的号码是（　　）

A．416　　B．432　　C．448　　D．464

【答案】A

【解析】设第 n 组抽到的号码是 a_n，则 $\{a_n\}$ 构成以 80 为公差的等差数列，

所以 $a_3=a_1+80\times2=160+a_1$，$a_4=a_1+80\times3=240+a_1$，

所以 $a_3+a_4=2a_1+80\times5=432$，解得 $a_1=16$，

所以 $a_6=16+80\times5=416$.

故选 A

5．下列有关线性回归分析的四个命题：

①线性回归直线必过样本数据的中心点 $\left(\overline{x},\overline{y}\right)$；

②回归直线就是散点图中经过样本数据点最多的那条直线；

③当相关性系数 $r>0$ 时，两个变量正相关；

④如果两个变量的相关性越强，则相关性系数 r 就越接近于 1.

其中真命题的个数为（　　）

A．1个　　B．2个　　C．3个　　D．4个

【答案】B

【解析】①线性回归直线必过样本数据的中心点（ $\overline{x},\overline{y}$ ）,故①正确;

②回归直线在散点图中可能不经过任一样本数据点,故②错误;

③当相关性系数 $r>0$ 时,则两个变量正相关,故③正确;

④如果两个变量的相关性越强,则相关性系数 r 就越接近于 1 或-1,故④错误.

故真命题的个数为 2 个,

故选 B

二、填空题

6．回归方程 $\hat{y}=2.5\hat{x}+0.2$ 在样本 $(4,1.2)$ 处的残差为_____.

【答案】–9

【解析】由题当 x=4 时，$\hat{y}=2.5\times4+0.2=10.2$，

1.2－10.2=－9

所以回归方程 $\hat{y}=2.5\hat{x}+0.2$ 在样本 $(4,1.2)$ 处的残差为－9.

故填－9

7．为了解某班学生喜爱打篮球是否与性别有关，对该班 50 名学生进行了问卷调查，得到了如下的 2×2 列联表：

	喜爱打篮球	不喜爱打篮球	合计
男生	20	5	25
女生	10	15	25
合计	30	20	50

则在犯错误的概率不超过_____的前提下认为喜爱打篮球与性别有关.

附：$K^2=\dfrac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$.

$P(K^2\geq k_0)$	0.10	0.05	0.025	0.010	0.005	0.001
k_0	2.706	3.841	5.024	6.635	7.879	10.828

【答案】0.005

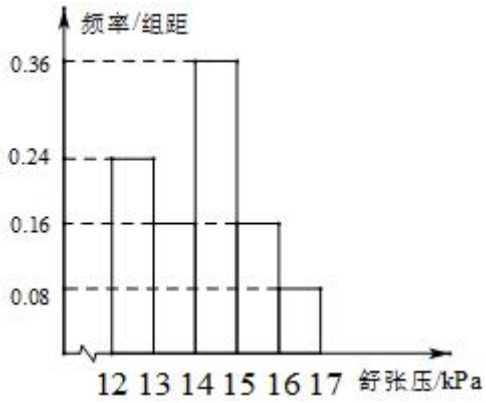
【解析】根据所给的列联表，

得到 $K^2=\dfrac{50(20\times15-10\times5)^2}{30\times20\times25\times25}=8.333>7.879$

∴至少有 99.5%的把握说明喜爱打篮球与性别有关，即在犯错误的概率不超过 0.005 的前提下认为喜爱打篮球与性别有关.

故填 0.005

8．某药厂选取若干名志愿者进行临床试验，所有志愿者的舒张压数据（单位：kPa）的分组区间为[12，13），[13，14），[14，15），[15，16），[16，17），将其按从左到右的顺序分别编号为第一组，第二组，……，第五组，如图是根据实验数据制成的频率分布直方图，已知第一组与第二组共有 20 人，则第三组中的人数为 _____.



【答案】18

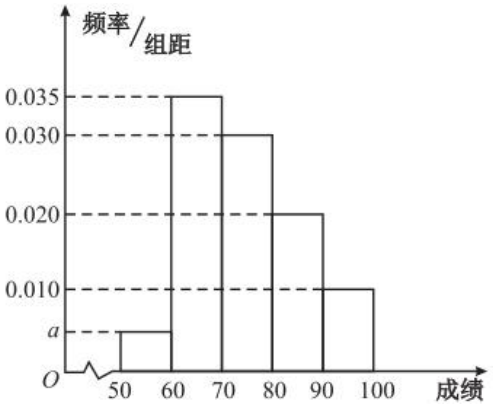
【解析】由直方图可得分布在区间第一组与第二组共有 20 人，分布在区间第一组与第二组的频率分别为 0.24，0.16，设总的人数为 n,则 $\frac{20}{n}=0.24+0.16=0.4,\therefore n=50$. 所以第 3 小组的人数为 $50\times0.36=18$ 人.

故填 18

三、解答题

9．某校两个班级 100 名学生在一次考试中的成绩的频率分布直方图如图所示，其中成绩分组区如下表：

组号	第一组	第二组	第三组	第四组	第五组
分组	$[50,60)$	$[60,70)$	$[70,80)$	$[80,90)$	$[90,100]$



（1）求频率表分布直方图中 a 的值；

（2）根据频率表分布直方图，估计这 100 名学生这次考试成绩的平均分；

（3）现用分层抽样的方法从第三、四、五组中随机抽取 6 名学生，将该样本看成一个总体，从中随机抽取 2 名，求其中恰有 1 人的分数不低于 90 分的概率.

【解析】（1）由题意得 $10a+0.01\times10+0.02\times10+0.03\times10+0.035\times10=1$，所以 $a=0.005$.

（2）由直方图分数在[50，60]的频率为 0.05，[60，70]的频率为 0.35，[70，80]的频率为 0.30，[80，90]的频率为 0.20，[90，100]的频率为 0.10，所以这 100 名学生期中考试数学成绩的平均分的估计值为：55×0.05+65×0.35+75×0.30+85×0.20+95×0.10=74.5

（3）由直方图，得：

第 3 组人数为 0.3×100=30，

第 4 组人数为 0.2×100=20 人，

第 5 组人数为 0.1×100=10 人．

所以利用分层抽样在 60 名学生中抽取 6 名学生，

每组分别为：

第 3 组：$\frac{30}{60}\times 6=3$ 人，

第 4 组：$\frac{20}{60}\times 6=2$ 人，

第 5 组：$\frac{10}{60}\times 6=1$=1 人．

所以第 3、4、5 组分别抽取 3 人、2 人、1 人．

设第 3 组的 3 位同学为 A1，A2，A3，第 4 组的 2 位同学为 B1，B2，第 5 组的 1 位同学为 C1，则从六位同学中抽两位同学有 15 种可能如下：

（A1，A2），（A1，A3），（A2，A3），（B1，B2），（A1，B1），（A1，B2），（A2，B1），（A2，B2），（A3，B1），（A3，B2），（A1，C1），（A2，C1），（A3，C1），（B1，C1），（B2，C1），其中恰有 1 人的分数不低于 90（分）的情形有：（A1，C1），（A2，C1），（A3，C1），（B1，C1），（B2，C1），共 5 种．

所以其中第 4 组的 2 位同学至少有一位同学入选的概率为 $\frac{5}{15}=\frac{1}{3}$.

10．随着资本市场的强势进入，互联网共享单车“忽如一夜春风来”，遍布了各级城市的大街小巷，为了解我市的市民对共享单车的满意度，某调查机构借助网络进行了问卷调查，并从参与调查的网友中随机抽取了 50 人进行分析．若得分低于 60 分，说明不满意，若得分不低于 60 分，说明满意，调查满意度得分情况结果用茎叶图表示如图 1．

40 岁及以下							40 岁以上						
				5	0	9	2						
	8	7	6	5	3	2	8	2					
8	8	6	5	4	3	2	7	5	5	7			
		6	5	4	2	0	6	2	4	5	6	9	
			8	8	5	2	5	0	4	5	6	7	8
					6	4	4	2	6	8	9		
					8	2	3	0	8				

（1）根据茎叶图找出 40 岁以上网友中满意度得分的众数和中位数；

（2）根据茎叶图完成下面列联表，并根据以上数据，判断是否有 99% 的把握认为满意度与年龄有关；

	满意	不满意	合计
40 岁以下			
40 岁以上			
合计			

（3）先采用分层抽样的方法从 40 岁及以下的网友中选取 7 人，再从这 7 人中随机选出 2 人，将频率视为概率，求选出的 2 人中至少有 1 人是不满意的概率．

参考格式：$K^2=\dfrac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$，其中 $n=a+b+c+d$

$P\left(K^2\geq k\right)$	0.150	0.100	0.050	0.025	0.010	0.005	0.001
k	2.072	2.706	3.841	5.024	6.635	7.879	10.828

【解析】（1）由题意可得，40 岁以上网友中满意度得分的众数为 75，中位数为 $\dfrac{57+58}{2}=57.5$．

（2）由茎叶图可得 2×2 列联表如下：

	满意	不满意	合计
40 岁以下	20	8	28
40 岁以上	10	12	22

合计	30	20	50

可知 $k^2=\dfrac{50(20\times12-10\times8)^2}{28\times22\times30\times20}\approx3.463<6.635$

所以没有 99% 的把握认为满意度与年龄有关.

（3）从所选取的 40 岁以下的网友中，采用分层抽样的方法选取 7 人，其中满意度为满意的有 5 人，

分别为 A_1，A_2，A_3，A_4，A_5，不满意的有 2 人，

分别为 B_1，B_2，所有组合的情况为 (A_1,A_2)，(A_1,A_3)，

(A_1,A_4)，(A_1,A_5)，(A_1,B_1)，(A_1,B_2)，(A_2,A_3)，

(A_2,A_4)，(A_2,A_5)，(A_2,B_1)，(A_2,B_2)，(A_3,A_4)，

(A_3,A_5)，(A_3,B_1)，(A_3,B_2)，(A_4,A_5)，(A_4,B_1)，

(A_3,B_2)，(A_5,B_1)，(A_5,B_2)，(B_1,B_2)，共有 21 种.

其中选出的 2 人中至少有 1 人是不满意的有 11 种，

故所求的概率 $P=\dfrac{11}{21}$.

一、选择题

11．已知 x,y 的取值如下表：

x	0	1，	2	3	4
y	1	1.3	3.2	5.6	8.9

若依据表中数据所画的散点图中，所有样本点 $(x_i,y_i)(i=1,2,3,4,5)$ 都在曲线 $y=\dfrac{1}{2}x^2+a$ 附近波动，则 $a=$（　　）

A．1　　B．$\dfrac{1}{2}$　　C．$\dfrac{1}{3}$　　D．$-\dfrac{1}{2}$

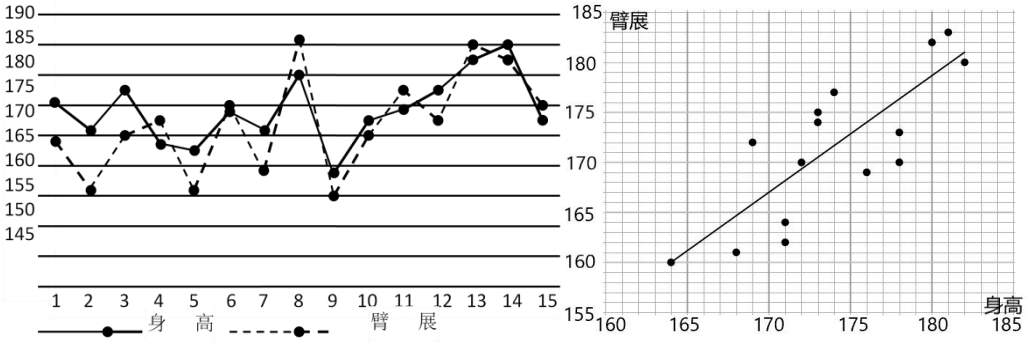
【答案】A

【解析】设 $t=x^2$，则 $\overline{t}=\dfrac{1}{5}(0+1+4+9+16)=6,\overline{y}=\dfrac{1}{5}(1+1.3+3.2+5.6+8.9)=4$，

所以点（6,4）在直线 $y=\dfrac{1}{2}t+a$ 上，求出 $a=1$，

故选 A.

12．某运动制衣品牌为了成衣尺寸更精准，现选择 15 名志愿者，对其身高和臂展进行测量（单位：厘米），左图为选取的 15 名志愿者身高与臂展的折线图，右图为身高与臂展所对应的散点图，并求得其回归方程为 $\hat{y}=1.16x-30.75$，以下结论中不正确的为（　　）



A．15 名志愿者身高的极差小于臂展的极差

B．15 名志愿者身高和臂展成正相关关系，

C．可估计身高为 190 厘米的人臂展大约为 189.65 厘米

D．身高相差 10 厘米的两人臂展都相差 11.6 厘米，

【答案】D

【解析】A，身高极差大约为 25，臂展极差大于等于 30，故正确；

B，很明显根据散点图像以及回归直线得到，身高矮臂展就会短一些，身高高一些，臂展就长一些，故正确；

C，身高为 190 厘米，代入回归方程可得到臂展估计值等于 189.65 厘米，但是不是准确值，故正确；

D，身高相差 10 厘米的两人臂展的估计值相差 11.6 厘米，但并不是准确值，回归方程上的点并不都是准确的样本点,故说法不正确.

故选 D.

二、填空题

13．已知一组数据 x_1，x_2，x_3，x_4，x_5 的方差为 2，则数据 $2x_1+3$，$2x_2+3$，$2x_3+3$，$2x_4+3$，$2x_5+3$ 的方差为_____.

【答案】8

【解析】设 $\overline{x}$，s^2 为数据 x_1，x_2，x_3，x_4，x_5 的平均数，方差，$\overline{X}$，S^2 为数据 $2x_1+3$，$2x_2+3$，$2x_3+3$，$2x_4+3$，$2x_5+3$ 的平均数，由题意可得方差为 $\frac{1}{5}\left[(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_5-\overline{x})^2\right]=2$，

所以$\left[(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_5-\overline{x})^2\right]=10$，

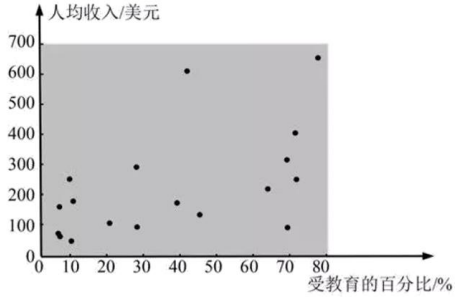
$\overline{X}=\dfrac{2(x_1+x_2+\cdots+x_5)+3\times5}{5}=2\overline{x}+3$，

$S^2=\dfrac{1}{5}\left[(2x_1+3-2\overline{x}-3)^2+(2x_2+3-2\overline{x}-3)^2+\cdots+(2x_5+3-2\overline{x}-3)^2\right]$

$=\dfrac{1}{5}\left[4(x_1-\overline{x})^2+4(x_2-\overline{x})^2+\cdots+4(x_5-\overline{x})^2\right]=\dfrac{4}{5}\times10=8$

故填 8

14．如图所示是世界 20 个地区受教育程度的人口百分比与人均收入的散点图，样本点基本集中在一个条型区域，因此两个变量呈线性相关关系．利用散点图中的数据建立的回归方程为$\hat{y}=3.193x+88.193$，若受教育的人口百分比相差 10%，则其人均收入相差_____.



【答案】31.93 美元

【解析】设所受教育百分比分别为$a\%,b\%$，且$a-b=10$

根据回归方程为$\hat{y}=3.193x+88.193$，

收入相差大约为：

$3.193\times a+88.193-(3.193\times b+88.193)=3.193\times10=31.93$，

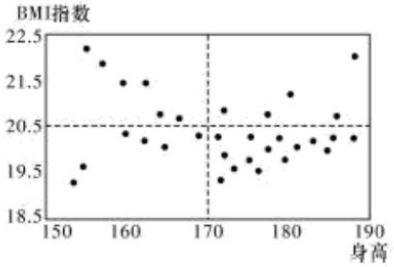
即受教育的人口百分比相差10%，则其人均收入相差约31.93美元.

故填 31.93 美元.

三、解答题

15．*BIM* 指数是用体重公斤数除以身高米数的平方得出的数字，是国际上常用的衡量人体胖瘦程度以及是否健康的一个标准.对于高中男体育特长生而言，当 *BIM* 数值大于或等于 20.5 时，我们说体重较重，当 *BIM* 数值小于 20.5 时，我们说体重较轻，身高大于或等于170cm我们说身高较高，身高小于 170cm 我们说身高较矮.

（1）已知某高中共有 32 名男体育特长生，其身高与 BMI 指数的数据如散点图，请根据所得信息，完成下述列联表，并判断是否有 95% 的把握认为男生的身高对 BMI 指数有影响.



	身高较矮	身高较高	合计
体重较轻			
体重较重			
合计			

（2）①从上述 32 名男体育特长生中随机选取 8 名，其身高和体重的数据如表所示：

编号	1	2	3	4	5	6	7	8
身高 $(\mathrm{cm})x$	166	167	160	173	178	169	158	173
体重 $(\mathrm{kg})y$	57	58	53	61	66	57	50	66

根据最小二乘法的思想与公式求得线性回归方程为 $\hat{y}=0.8x-75.9$．利用已经求得的线性回归方程，请完善下列残差表，并求 R^2（解释变量（身高）对于预报变量（体重）变化的贡献值）（保留两位有效数字）；

编号	1	2	3	4	5	6	7	8
体重（kg）y	57	58	53	61	66	57	50	66
残差 $\hat{e}$	0.1	0.3	0.9	-1.5	-0.5	-2.3	-0.5	

②通过残差分析，对于残差的最大（绝对值）的那组数据，需要确认在样本点的采集中是否有人为的错误，已知通过重新采集发现，该组数据的体重应该为 $58(\mathrm{kg})$．小明重新根据最小二乘法的思想与公式，已算出 $\hat{y}=0.675x+\hat{a}$，请在小明所算的基础上求出男体育特长生的身高与体重的线性回归方程.

参考数据：

$(0.1)^2+(0.3)^2+(0.9)^2+(-1.5)^2+(-0.5)^2+(-2.3)^2+(-0.5)^2=8.95$，$\bar{x}=168$，

$\sum_{i=1}^{8}\left(y_i-\bar{y}\right)^2=226$，$0.675\times168=113.4$，

参考公式：$R^2=1-\dfrac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}$，$\hat{b}=\dfrac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}=\dfrac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\cdot\bar{y}}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}$，$\hat{a}=\bar{y}-\hat{b}\bar{x}$，

$\hat{e}_i=y_i-\hat{b}x_i-\hat{a}$，$K^2=\dfrac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)},(n=a+b+c+d)$.

$P\left(K^2\geq0\right)$	0.10	0.05	0.01	0.005
k_0	2.706	3.811	6.635	7.879

【解析】（1）

	身高较矮	身高较高	合计
体重较轻	6	15	21
体重较重	6	5	11
合计	12	20	32

由于 $K^2=\dfrac{32(6\times5-6\times15)^2}{12\times20\times21\times11}=\dfrac{160}{77}<3<3.841$，

因此没有95%的把握认为男生的身高对 BMI 指数有影响.

（2）①对编号为 8 的数据 $\hat{e}_8=66-0.8\times173+75.9=3.5$，完成残差表如下所示：

编号	1	2	3	4	5	6	7	8
体重 $(kg)y$	57	58	53	61	66	57	50	66
残差 $\hat{e}$	0.1	0.3	0.9	−1.5	−0.5	− 2.3	−0.5	3.5

$$\sum_{i=1}^{8}\left(y_i-\hat{y}_i\right)^2=(0.1)^2+(0.3)^2+(0.9)^2+(-1.5)^2+(-0.5)^2+(-2.3)^2+(-0.5)^2+(3.5)^2=21.2$$

$$R^2=1-\frac{\sum_{i=1}^{8}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{8}\left(y_i-\overline{y}\right)^2}=1-\frac{21.2}{226}\approx 0.91\text{.}$$

所以解释变量（身高）对于预报变量（体重）变化的贡献值 R^2 约为 0.91.

②由①可知，第八组数据的体重应为 58．此时，易知，$\overline{x}=168$，$\overline{y}=57.5$，

$\hat{a}=57.5-0.675\times 168=-55.9$，

所以重新采集数据后，男体育特长生的身高与体重的线性回归方程为 $\hat{y}=0.675x-55.9$.