生物统计学-期中测试 Mid-term Exam

2023.11

1. 假设零假设成立，以下哪个陈述是正确的？ [5 分]

（1） 具有更大样本的研究相较于具有更小样本的研究更有可能获得 P<0.05 的结果。

（2） 具有更大样本的研究相较于具有更小样本的研究更不太可能获得 P<0.05 的结果。

（3） 具有更大样本的研究相较于具有更小样本的研究同样可能获得 P<0.05 的结果。

陈述（3）正确。

2. 对于以下每个研究，请说明哪个是解释变量，哪个是响应变量，以及说明这个研究是观测性的(observational)还是实验性的(experimental)。 [10 分]

（1） 林业研究人员想比较高海拔地区生长的树木与低海拔地区生长的树木生长速度。他们使用从天然森林中砍伐的一组树木的树环之间的间距来测量生长速度。

（2）一种自然界中的蜘蛛的雌性经常吃掉试图与它们交配的雄性。研究人员进行了一项研究，将一组雄性蜘蛛进行切除一条腿的处理（使其更脆弱并容易被捕食），而另一组蜘蛛雄性则保持完整（未受伤），进而研究这两组雄性蜘蛛在交配期间的生存情况是否有差异。

（1）解释变量：海拔高度 [2 分]；响应变量：树木生长速度 [2 分]；研究类型：观测性研究 [1 分]

（2）解释变量：腿的状态（被切除或未受损伤 [2 分]； 响应变量：雄性蜘蛛在交配期间的生存情况 [2 分]；研究类型：实验性研究 [1 分]

3. 我们从人类基因组中得到一个随机抽样的 100 个基因的样本（如 Lecture04-uncertainty_and_probability 第 10 页），基因片段长度的中位数（median）为 2640. 5 个核苷酸请指出以下各种说法正确还是错误。 [9 分]

（1）所有人类基因的中位基因长度为 2640.5 核苷酸。

（2）所有人类基因的中位基因长度估计为 2640.5 核苷酸。

（3）样本中位数有具有标准误的抽样分布。

（1）假-错误；（2）真-正确；（3）真-正确； [各 3 分]

4. HIV 病毒与细菌和多细胞生命相比，具有较高的突变率。其中许多突变对病毒不利，导致其复制速度较慢。研究人员据此构建了一个庞大的已知 HIV 突变体及其复制速率的数据库。我们使用计算机从该数据库中随机抽取了 100 个突变体，计算了样本的复制速率的均值和中位数。我们多次重复了该过程，每次都计算了样本的中位数和均值。样本均值分布（均值的抽样分布）的标准差为 0.00073；样本中位数分布（中位数的抽样分布）的标准差为 0.0224。 [10 分]

（1） 复制速率均值的标准误是多少？

（2） 复制速率中位数的标准误是多少？

（3） 如果从数据库中随机抽取 100 个突变体的样本数据，哪个中心趋势测量值可能更精确，均值还是中位数？请解释你的答案。

（1）均值复制速率的标准误是 0.00073。 [2 分]

（2）中位数复制速率的标准误是 0.0224。 [2 分]

（3）如果从数据库中随机抽取 100 个数据突变体的样本，均值可能会更精确。这是因为标准误是测量数据样本的离散程度的度量，标准误较小表示数据点较接近均值（更精确）。在这种情况下，均值的标准误远远小于中位数的标准误，表明均值更具精确性，更能代表数据的中心趋势。 [6 分]

5. 一项在上海市闵行区进行的研究中，研究者通过户籍登记确定了 500 个家庭的随机样本。他们共发出了 500 份问卷，针对每个家庭的一名成年人调查了其对共享单车的态度。其中 80 份问卷被填写并寄回给研究人员。请问： [10 分]

（1）那 80 个寄回问卷的家庭是否可以被视为家庭的随机样本？请解释。

（2）会影响调查结果的偏差类型是什么？

（1）80 个寄回的问卷不能被视为家庭的随机样本。原因是这些问卷的回应者是自愿参与的，只有那些有兴趣或强烈意见的人才会回应问卷，而其他人可能选择不回应。[5 分]

（2）自愿性偏差（voluntary bias）。【自愿性偏差是因为只有感兴趣或有强烈意见的人回应了问卷，而其他人没有回应，从而导致结果不具有代表性。这样的偏差可能使研究的结果不准确或倾向于某种特定类型的受访者。】 [5 分]

6. 弗朗西斯·高尔顿（1894 年）提供了以下关于至少飞行 90 英里(miles)的 3207 只信鸽的飞行速度的数据（见下图；横坐标(x)为速度，单位是码/分钟（yards/min），纵坐标(y)为信鸽的频数，单位通常为“个”，故省略）。请回答： [15 分]

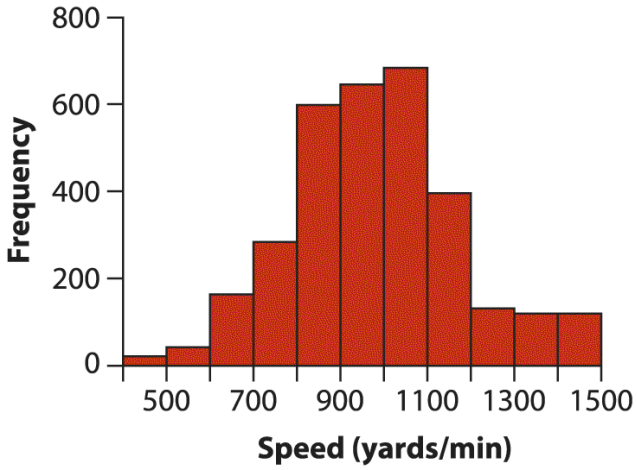
（1）这是什么类型的图表？

（2）检查图表并用肉眼直观地估计均值（mean, 取到最接近的 100 码/分钟，如 1000 码/分钟，而不用精确到 1030 码/分钟）。解释如何获得这个估值。

（3）检查图表并用肉眼直观地估计中位数的估值（median, 取到最接近的 100 码/分钟）。解释如何获得这个估值。

（4）检查图表并用肉眼直观地估计众数（mode, 出现频数最多的值）的大致值或区间（取到最接近的 100 码/分钟）。解释如何获得这个估计。

（5）检查图表并用肉眼直观地估计标准差的大致值（standard deviation, SD, 取到最接近的 50 码/分钟）。解释如何获得这个估计（提示：95%的观测值范围大致等于 mean ± 2SD）。



（1）这是一个直方图。 [2 分]

（2）均值：大约 1000 码/分钟[2 分]。频率分布相对对称，因此均值应位于中间附近[1 分]。

（3）中位数：大约 900 码/分钟[2 分]。频率分布相对对称，因此中位数应位于中间附近，接近均值[1 分]。

（4）众数：1000-1100 码/分钟[2 分]。频率分布中出现最频繁的区间[1 分]。

（5）标准差（s）：大约 200 码/分钟[2 分]。基于以下事实，如果分布大致呈钟形（正态分布），那么约 95%的观测值将落在均值减去 2s 和均值加 2s 之间。从直方图上我们观察到 600 到 1400 码/分钟应该包括约 95%的频率分布，所以（1400-600）/4 = 200 码/分钟。这是一个非常粗略的计算！ [2 分]

7. 假设孟德尔的一个实验中有 1600 株豌豆植株，其中有 900 株是高植株并且具有绿色的豆荚，300 株是高植株并且具有黄色的豆荚，300 株是矮植株并且具有绿色的豆荚，还有 100 株是矮植株并且具有黄色的豆荚。 [17 分]

（1）对于这批植株，"高"和"绿色豆荚"是互斥的特征吗？

（2）对于这批植株，"高"和"绿色豆荚"是独立的特征吗？（提示：可先分别计算出"高"和"绿"各自发生的概率）

（1）不是的，因为有一些植株既高又具有绿色的豆荚，所以 "高" 和 "绿色豆荚" 不是互斥的。 [5 分]

（2）1200/1600 是高的，1200/1600 具有绿色豆荚[4 分]。如果它们是独立的，使用概率的乘法公式，那么 Pr[高和绿色] = Pr[高] × Pr[绿色] = 3/4 × 3/4 = 9/16[4 分]，等于说在 1600 株中有 900 株。实际上，在 1600 中有 900，因此看起来绿色和高度是独立的。 [4 分]

8. 一项临床试验的目的是测试一种新疗法是否会影响患有严重疾病的患者的康复率。结果为零假设 "H_0: 治疗没有效果" 被拒绝，P 值为 0.04。研究人员使用了显著性水平 α=0.05。请说明以下各项结论是否正确。如果不正确，请解释原因。 [24 分]

（1） 治疗只有很小的效应。

（2） 治疗有一些效应。

（3） 发生第一类错误的概率为 0.04。

（4） 发生第二类错误的概率为 0.04。

（5） 如果显著性水平设置为 α=0.01 而不是 0.05，零假设就不会被拒绝。

（1）不正确[3 分]。P 值不提供效应的大小[3 分]。

（2）正确。[3 分]

（3）不正确[3 分]。第一类错误的概率由事先决定的显著性水平 0.05 来确定[3 分]。

（4）不正确[3 分]。第二类错误的概率取决于效应大小，而效应大小是未知的[3 分]。

（5）正确。 [3 分]