Last name: First name:

Student ID #: Section:

STAT 311 Final (200 points)

Fritz Scholz

1. **(5 points)** Evaluate

$$\sum_{i=1}^{3}(2i-1) = (2-1)+(4-1)+(6-1) = 1+3+5 = 9$$

2. **(5 points)** If I give you $P(A \cap B) = 0.2$ and $P(A) = 0.9$, what is $P(A \cap B^c)$?
Since $P(A) = P(A \cap B) + P(A \cap B^c) \implies P(A \cap B^c) = 0.9 - 0.2 = 0.7$.

3. **(8 points)** If $X_1, X_2, \ldots, X_9$ are independent Bernoulli random variables with success probability $p = 0.3$, what R command would give you

$$1 - \texttt{pbinom}(4, 9, 0.3) = P(X_1 + X_2 + \ldots + X_9 \geq 5)$$

since the sum of independent Bernoulli r.v.'s is a binomial r.v.

4. **(12 points)** Explain the utility of a box plot and of several box plots side by side.
The single box plot provides information on i) the location via the sample median, ii) the spread via the iqr (distance between $\hat{q}_1$ and $\hat{q}_3$), iii) symmetry of the distribution (centering of $\hat{q}_2$ between $\hat{q}_1$ and $\hat{q}_3$, and equal length whiskers) and iv) outliers.
Side by side plots allow comparison of such characteristics across several samples.

5. **(10 points)** Let $A = \{2, 3, 4, 6, 8\}$ and $B = \{2, 3, 7, 10\}$ be two sets, contained within the sample space S consisting of the integers from 1 to 100 inclusive. Describe the set $C = (A^c \cap B) \cup (A \cap B^c)$ in terms of its elements.

$A^c \cap B = \{7, 10\}$, $A \cap B^c = \{4, 6, 8\} \implies C = (A^c \cap B) \cup (A \cap B^c) = \{7, 10, 4, 6, 8\}$

6. **(9 points)** What is meant by the term "a statistic"? Give three different examples.
A statistic is anything computed from sample data or is a function of the sample data, e.g., the sample mean $\bar{x}$, sample variance s^2, or the order statistics $x_{(1)} \leq \ldots \leq x_{(n)}$.

7. **(5 points)** Assume a bivariate normal distribution for (X, Y): Regression to the mean is strongest when $|\rho|$ is near 1. TRUE or FALSE?
The correct answer is FALSE. When $|\rho| = 1$ there is no regression to the mean μ_y and when $\rho = 0$ we have complete regression to the mean, i.e., $E(Y|X = x) = EY = \mu_y$.

8. **(12 points)** In the k-sample problem, state what is meant by the sum of squares decomposition and explain its meaning.

$$SS_T = SS_W + SS_B \quad \text{or} \quad \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{\cdot\cdot})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2$$

The variability of all observations X_{ij} around the grand average $\bar{X}_{\cdot\cdot}$, as measured by SS_T, can be split into two components, the variability of the X_{ij} around the individual group averages $\bar{X}_{i\cdot}$ summed over all groups and expressed as SS_W, and the variability of the group averages $\bar{X}_{i\cdot}$ around the grand average, expressed by SS_B.

9. **(10 points)** Explain how/why strong evidence for correlation is not necessarily the same as strong correlation.
Strong evidence for correlation refers to the fact that the hypothesis $H_0 : \rho = 0$ can be strongly rejected because we have sufficient evidence for some non-zero correlation. This can happen either for large $|r|$ or not so large $|r|$ when n is large. The latter can happen even when the actual ρ is small. Strong correlation means that the actual $|\rho|$ is close to 1 (or not close to 0) but it could well be that the sample size is so small that we don't get sufficient evidence for rejecting H_0. An example was given in class.

10. **(5 points)** For the following sample give the order statistics. $3.1, 5.4, 2.1, 6.5, 1.1$.

the order statistics are: 1.1, 2.1, 3.1, 5.4, 6.5

11. **(12 points)** In a sample $x_1, x_2, \ldots, x_{10}$ of distinct values, the three highest values are each increased by 5. What change is caused by that increase in the

(a) sample mean? it is increased by $(5 + 5 + 5)/10 = 1.5$.

(b) sample median? The sample median $(x_{(5)} + x_{(6)})/2$ is not changed.

(c) third sample quartile? $\hat{q}_3 = x_{(8)}$ changes to $x_{(8)} + 5$, an increase of 5.

12. **(5 points)** What probability $p \in [0, 1]$ is equivalent to a 0.2% chance?
It corresponds to $p = 0.002$.

13. **(15 points)** When tossing a fair penny twice we would expect the following probabilities of seeing no head (1/4), exactly one head (1/2), and two heads (1/4). After repeating such paired coin tosses 400 times we have the following observed counts: no head (80), exactly one head (220), and two heads (100). Using the Pearson X^2 test statistic for testing that the above probabilities indeed apply (i.e., we have a fair penny) evalute X^2 and give the R expression for evaluating its significance probability, providing all required numerical arguments.
The expected counts are $e_1 = 400/4 = 100$, $e_2 = 400/2 = 200$ and $e_3 = 400/4 = 100$, respectively. Thus

$$X^2 = \frac{(80 - 100)^2}{100} + \frac{(220 - 200)^2}{200} + \frac{(100 - 100)^2}{100} = \frac{400}{100} + \frac{400}{200} = 6$$

with significance probability `1-pchisq(6,3-1)=1-pchisq(6,2)`.

14. **(12 points)** Picking a number at random from $S = \{1, 2, 3, \ldots, 23, 24\}$ examine whether the following events A and B are independent. A is the event that the picked number is a multiple of 3 and B is the event that the picked number is odd. Would your answer change if $S = \{1, 2, 3, \ldots, 23, 24, 25\}$?
In $S = \{1, 2, 3, \ldots, 23, 24\}$ there are 8 multiples of 3 of which 4 are odd, and S has 12 odd numbers , thus $P(A) = 8/24 = 1/3$ and $P(B) = 12/24 = 1/2$ and

$$P(A \cap B) = \frac{4}{24} = \frac{1}{6} = \frac{1}{3} \cdot \frac{1}{2} = P(A)P(B) \implies \text{independence}$$

When $S = \{1, 2, 3, \ldots, 23, 24, 25\}$ we have

$$P(A \cap B) = \frac{4}{25} \neq \frac{8}{25} \cdot \frac{13}{25} = P(A)P(B) \implies \text{dependence}$$

15. **(15 points)** You are presented with two data situations:
a) $n = 300 = 12 \cdot 25$ and $\bar{x}_{300} = 0.56$ and
b) $n = 108 = 12 \cdot 9$ and $\bar{x}_{108} = 0.6$.
Assume that you are sampling from a Uniform$(0, a)$ distribution with mean $\mu = a/2$ and variance $\sigma^2 = a^2/12$. Which of the two presented data situations provides stronger evidence against H_0 when testing $H_0 : \mu = 0.5$ (i.e., $a = 1$) against $H_1 : \mu \neq 0.5$ (i.e., $a \neq 1$) or do they provide about the same evidence? Reason why.

[Hint: Compute $\text{var}\bar{X}_n$ under H_0 for each data situation and then the corresponding $\sigma(\bar{X}_n)$. For the latter you should wind up with $1/m$, where m is a simple integer, different for a) and b). You may appeal to a normal approximation in either case.]
(a) $\text{var}\bar{X}_n = \sigma^2(\bar{X}_n) = \sigma^2/n = 1/(n \cdot 12) = 1/(12^2 \cdot 5^2) = 1/60^2$ and thus

$$z = \frac{\bar{x}_n - 0.5}{\sigma(\bar{X}_n)} = \frac{0.56 - 0.5}{1/60} = 60 \cdot 0.06 = 3.6$$

(b) $\text{var}\bar{X}_n = 1/(n \cdot 12) = 1/(12^2 \cdot 3^2) = 1/36^2$ and thus

$$z = \frac{\bar{x}_n - 0.5}{\sigma(\bar{X}_n)} = \frac{0.6 - 0.5}{1/36} = 36 \cdot 0.1 = 3.6$$

they appear to present equally strong evidence against H_0.

16. **(15 points)** A disease has a prevalence rate of about 1/1000 in a given population. A diagnostic test for the disease can give a positive result R_+ (suggesting the disease is present) or a negative result R_- (suggesting the disease is not present). Let D represent the event that a person, randomly selected from the population, has the disease. You are told the following facts about the diagnostic test's efficacy:

(a) $P(R_-|D) = 0.2 =$ the probability that a person **with** the disease tests negative, i.e., this is the probability of a false negative..

(b) $P(R_+|D^c) = 0.1 =$ the probability that a person **without** the disease tests positive, i.e., this is the probability of a false positive.

i) What is the probability $P(R_+|D)$ of a true positive?
$P(R_+|D) = 1 - P(R_-|D) = 1 - 0.2 = 0.8$.

ii) Find $P(D|R_+)$ as the ratio of two reduced integers and roughly indicate its magnitude in decimal terms, comparing it with $P(D)$.

$$\begin{aligned} P(D|R_+) &= \frac{P(DR_+)}{P(R_+)} = \frac{P(R_+|D)P(D)}{P(R_+|D)P(D) + P(R_+|D^c)P(D^c)} \\ &= \frac{0.8\frac{1}{1000}}{0.8\frac{1}{1000} + 0.1\frac{999}{1000}} = \frac{0.8}{0.8 + 99.9} = \frac{8}{1007} \approx 0.008 \end{aligned}$$

which is about 8 times larger than $P(D)$.

17. **(8 points)** For independent random variables X and Y with $\mu_x = EX = 100$ and $\sigma_x^2 = \text{var}\, X = 9$ and $\mu_y = EY = 50$ and $\sigma_y^2 = \text{var}\, Y = 4$, what are the **mean** and **standard deviation** of $X - 2Y$?

$E(X - 2Y) = 100 - 2 \cdot 50 = 0$ and $\text{var}(X - 2Y) = 9 + 2^2 \cdot 4 = 25 \Longrightarrow \sigma(X - 2Y) = 5$

18. **(5 points)** When using a bivariate normal random sample of size $n = 100$ and testing the hypothesis $H_0 : \rho = 0$, which distribution (with specified parameters) is used to calculate significance probabilities.
The $t(n-2) = t(98)$ distribution.

19. **(5 points)** Assume that $n = 20$ pairs $(x_i, Y_i), i = 1, \ldots, 20$ are obtained according to the simple linear regression model. Which distribution (with specified parameters) is used to calculate confidence intervals for the slope parameter β_1?
The $t(n-2) = t(18)$ distribution.

20. **(15 points)** Assuming a sample of size $n = 16$ from $\mathcal{N}(\mu, \sigma^2)$ with average $\bar{x}_n = 20$ and sample standard deviation $s_n = 8$, what is the 90% confidence interval for μ when given

$$\texttt{qt}(.95, 16) = 1.746 \quad \texttt{qt}(.90, 16) = 1.337 \quad \texttt{qt}(.95, 15) = 1.753$$
$$\texttt{qt}(.95, 14) = 1.761 \quad \texttt{qt}(.90, 17) = 1.333 \quad \texttt{qt}(.90, 15) = 1.341$$

Give the answer as the two decimal numbers that form the interval.
Do you have reason to reject the hypothesis $H_0 : \mu = 23$ at level $\alpha = 0.10$ when testing it against the alternative $H_1 : \mu \neq 23$? (**Explain your reasoning.**)
The confidence interval is

$$\bar{x} \pm q_t(0.95, 15)\frac{s_n}{\sqrt{n}} = 20 \pm 1.753\frac{8}{4} = 20 \pm 3.506 = (16.494, 23.506)$$

Since this interval still contains 23, it is an acceptable hypothesis at level $\alpha = 0.10$, i.e., we should not reject H_0.

21. **(12 points)** In testing $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$ you reject H_0 when the test statistic T is too far from 0. Explain the position of the observed value t of T relative to the critical value t_{crit} for the level α test when the significance probability $\mathbf{p}(t)$ is

    (a) $\leq \alpha$
    Since $\mathbf{p}(t) = P(|T| \geq |t|) \leq \alpha = P(|T| \geq t_{\text{crit}})$ we must have that $|t| \geq t_{\text{crit}}$, i.e., t is outside the acceptance interval $(-t_{\text{crit}}, t_{\text{crit}})$.

    (b) $> \alpha$
    Since $\mathbf{p}(t) = P(|T| \geq |t|) > \alpha = P(|T| \geq t_{\text{crit}})$ we must have that $|t| < t_{\text{crit}}$, i.e., t is inside the acceptance interval $(-t_{\text{crit}}, t_{\text{crit}})$.