Statistics 2000, Section 001, Final (300 Points)

Wednesday, May 5, 2010

Your Name: ____________________

−2 for each calculation error
−2 if no final result

Question 1: Random Variables (**30 Points**)

When figure skaters need to find a partner for "pair figure skating," it is important to find a partner who is compatible in weight. The weight of figure skaters can be modeled by a Normal distribution. Let the random variable X denote the weight of female skaters and the random variable Y denote the weight of male skaters. For female skaters (X), the mean is 110 lbs with a standard deviation of 5 lbs. For male skaters (Y), the mean is 170 lbs with a standard deviation of 10 lbs. **Show your work!**

1. (**6 Points**) What is $P(X < 100)$?

$$z\text{-score} = \frac{100-110}{5} = -2 \quad 3$$

$$P(X < 100) = P\left(Z < \frac{100-110}{5}\right) = P(Z < -2.00) = \underline{\underline{0.0228}}$$

$$= \underline{\underline{2.28\%}} \quad 3$$

2. (**6 Points**) Approximately 90% of the male skaters weigh more than how many pounds?

$z^* = -1.28$ 3 yields an approx. area of 10% to the left

$\Rightarrow$ male weight $= -1.28 \cdot 10 + 170 = \underline{\underline{157.2 \text{ lbs}}}$ 3

3. (**6 Points**) The weight of a pair of figure skaters (a male and a female) can be thought of as a new random variable. Define the random variable W as $W = X+Y$. What is the mean μ_W of this new random variable W?

$$\mu_W = \mu_{X+Y}$$
$$= \mu_X + \mu_Y$$
$$= 110 + 170$$
$$= \underline{\underline{280 \text{ lbs}}}$$

6

4. (**6 Points**) Suppose we consider the weights of the male partner and the female partner to be independent. What is the standard deviation σ_W of the random variable W?

-2 if no $\sqrt{}$

$$\sigma_W = \sigma_{X+Y}$$
$$= \sqrt{\sigma_X^2 + \sigma_Y^2}$$
$$= \sqrt{5^2 + 10^2}$$
$$= \sqrt{125}$$
$$= \underline{\underline{11.18 \text{ lbs}}}$$

6

5. (**6 Points**) It does not seem likely that the weights of the male partner and the female partner would be independent. If the correlation ρ between X and Y equals 0.77, what is the standard deviation σ_W of the random variable W?

-3 if no 3rd part

-2 if no $\sqrt{}$

$$\sigma_W = \sigma_{X+Y}$$
$$= \sqrt{\sigma_X^2 + \sigma_Y^2 + 2 \cdot \rho \cdot \sigma_X \cdot \sigma_Y}$$
$$= \sqrt{5^2 + 10^2 + 2 \cdot 0.77 \cdot 5 \cdot 10}$$
$$= \sqrt{202}$$
$$= \underline{\underline{14.21 \text{ lbs}}}$$

6

Question 2: Statistical Inference (**102 Points**)

Dichlorodiphenyltrichloroethane (thereafter DDT) is one of the most well–known synthetic pesticides. It is a chemical with a long, unique, and controversial history. First synthesized in 1874, DDT's insecticidal properties were not discovered until 1939, and it was used with great success in the second half of World War II to control malaria and typhus among civilians and troops. The Swiss chemist Paul Hermann Müller was awarded the Nobel Prize in Physiology or Medicine in 1948 "for his discovery of the high efficiency of DDT as a contact poison against several arthropods". After the war, DDT was made available for use as an agricultural insecticide, and soon its production and use skyrocketed.

Unfortunately the pesticide DDT does have a negative side. It causes tremors and convulsions if it is ingested by humans or other mammals. Researchers seek to understand how the convulsions are caused. In a randomized comparative experiment, 5 white rats poisoned with DDT were compared with a control group of 5 unpoisoned rats. Electrical measurements of nerve activity are the main clue to the nature of DDT poisoning. When a nerve is stimulated, its electrical response shows a sharp spike followed by a much smaller second spike. Researchers found that the second spike is larger in rats fed DDT than in normal rats. This observation helps biologists understand how DDT causes tremors. (This example is loosely based on D. L. Shankland, "Involvement of spinal cord and peripheral nerves in DDT-poisoning syndrome in albino rats," *Toxicology and Applied Pharmacology*, 6 (1964), pp. 197–213.) The researchers measured the amplitude of the second spike as a percentage of the first spike when a nerve in the rat's leg was stimulated.

−2 for each calculation error
−2 if no final result

For the poisoned rats the results were:

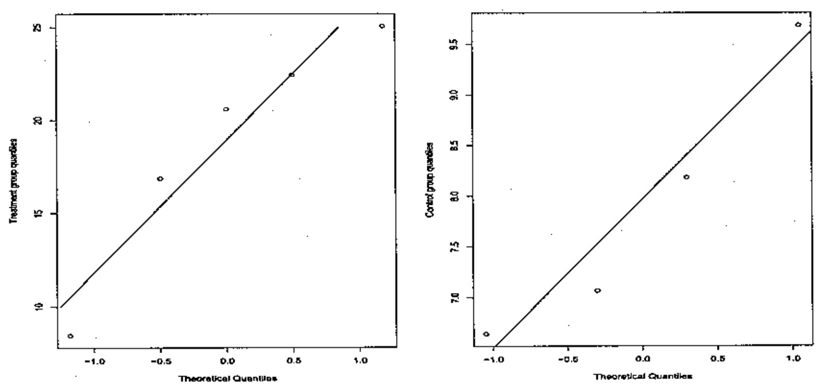
16.869, 25.050, 22.429, 8.456, 20.589. Group 1

The control group data were:

9.686, 7.064, 8.182, 6.642. Group 2

Note that one rat from the control group escaped before the experiment started.

The researchers created the normal quantile plots to investigate the distributions of data:



Part 1:

By inspecting the data we see that the difference in sample means is quite large, but in such small samples the sample mean is highly variable. A significance test can help confirm that we are seeing a real effect. *Note that the researchers did not conjecture in advance that the size of the second spike would increase or decrease in rats fed DDT.* **Show your work!**

1. **(6 Points)** First of all, indicate which test (z–, t–, χ^2–, one–sided, two–sided, one–sample, two–sample, etc.) you are going to choose, briefly justify your choice, and carefully indicate the crucial assumption(s) that should be satisfied for your test to be valid. Also indicate the ways you would check your assumptions.

two-sided, two-sample t-test, because:
1 1 1

- data roughly follow normal distributions 1
[yes: see normal quantile plots above] 1
- population standard deviations σ_1 and σ_2 are unknown 1
[yes]

2. **(6 Points)** State the appropriate null and alternative hypotheses to answer this question. Use the proper mathematical notation and symbols.

−4 if swapped

$H_0: \mu_1 = \mu_2$ 3

$H_a: \mu_1 \neq \mu_2$ 3

3. **(4 Points)** Briefly justify your choice of a one sided or two sided alternative in the previous part.

two-sided because it was not conjectured in advance that the size of the second spike would increase or decrease in rats fed DDT 4

4. **(10 Points)** Calculate the test statistic.

$\bar{x}_1 = 18.679$ 1

$\bar{x}_2 = 7.894$ 1

$s_1^2 = 41.52$ 2 $[s_1 = 6.44]$

$s_2^2 = 1.85$ 2 $[s_2 = 1.36]$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{18.679 - 7.894}{\sqrt{\frac{41.52}{5} + \frac{1.85}{4}}}$$

$$= \frac{10.785}{\sqrt{8.7665}} = \frac{10.785}{2.96} = \underline{3.64}$$ 4

5. **(6 Points)** What distribution does your statistic follow. Indicate the name of the distribution and numerical value(s) of the parameters that uniquely describe(s) the distribution of the test statistic. For example, if the test statistic follows the Binomial distribution, you need to specify the numerical value of the sample size and the probability of success, i.e., $X \sim B(n, p)$.

df = smaller of $(5-1)$ and $(4-1)$ $= 3$

3

$X \sim t$ with $k = 3$ df 3 , i.e., X has a $t(3)$ distribution

6. **(6 Points)** Determine the P–value.

$3.482 < t < 4.541$ 2

$2\times(0.02 > p > 0.01)$ 2

$\Rightarrow$ P-value between 0.02 and 0.04

2

7. **(6 Points)** If you use the usual 5% significance level, should you reject the null hypothesis? **Yes** / **No** ? Circle your answer and briefly explain why/why not.

4

Yes, we reject the null hypothesis at the 5% significance level because the P-value is less than 0.05. 2

8. **(6 Points)** If you use the 1% significance level instead, should you reject the null hypothesis? **Yes** / **No** ? Circle your answer and briefly explain why/why not.

4

No, we fail to reject the null hypothesis at the 1% significance level because the P-value is greater than 0.01. 2

9. **(4 Points)** Give a short summary of your conclusion for the usual 5% significance level, i.e., how would you explain the result to a person who does not know much about statistics?

As we reject the null hypothesis at the 5% significance level, we have some reason to believe that the two population means μ_1 (for the poisoned rats) and μ_2 (for the control group) are indeed different. 4

Part 2:

The researchers recognize that there is likely to be some error in the estimation of the difference in population means. On the basis of known characteristics of the distributions in question, and on our calculation of the standard error of the statistic we are using as an estimate, the researchers would also like to place an interval around the estimator of the difference in population means that specifies the likely range within which the true difference in population means is likely to fall. Therefore they decide to construct a confidence interval. **Show your work!**

1. **(6 Points)** First, what would be a good estimator for the unknown difference in population means? Indicate the proper mathematical symbol(s), and the numerical value(s).

$$\overline{X_d} = \overline{X_1} - \overline{X_2} \quad 4$$

$$= 18.679 - 7.894 = \underline{10.785} \quad 2$$

2. **(6 Points)** What is the standard error of the estimator? Indicate the proper mathematical symbol(s), and the numerical value(s).

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad 4$$

$$= \underline{2.96} \quad 2$$

3. **(6 Points)** Calculate the margin of error for your 95% confidence interval.

$$m = t_3^* \cdot SE \quad 4$$

$$= 3.182 \cdot 2.96 = \underline{9.419} \quad 2$$

[for t_3^* related to $t(3)$ distribution]

4. **(6 Points)** Now, construct a 95% confidence interval for the true difference in population means.

95% CI for $\mu_1 - \mu_2$:

$$10.785 \pm 9.419 = \underline{(1.366, 20.204)}$$

3 3

5. **(6 Points)** Can we be certain that the true difference in population parameters falls in this interval? **Yes** / **No**? Circle your answer.

6

[no – we are just "95% confident" that the true difference falls into this interval]

6. **(6 Points)** What critical value would you use if you wanted 90% confidence instead of 95% confidence? You do not have to calculate this interval — just indicate the value from the table that is needed.

$t_3^* = \underline{2.353}$ 6 for $C = 90\%$

7. **(6 Points)** Would the 90% confidence interval be **(i) wider** or **(ii) narrower** than your 95% confidence interval from 4. (Part 2) above? [Do NOT actually compute the 90% confidence interval — just circle the correct answer.] 6

[a lower confidence level C results in a narrower interval]

8. **(6 Points)** Suppose the alternative hypothesis in 2. (Part 1) is two-sided and $\alpha = 0.05$. Without performing any calculations, just using your confidence interval obtained in 4. (Part 2.) would you reject H_0? **Yes** / **No** / **Cannot tell?** Circle your answer. No explanation is needed.

6

[A level α (here: 5%) two-sided significance test rejects the null hypothesis when μ_0 falls outside a level $1-\alpha$ (here: 95%) CI for μ. Here $\mu_0 = \mu_1 - \mu_2$ represents the difference in the population means. The value "0" is outside the 95% CI calculated in 4. Therefore, we would reject $H_0: \mu_1 = \mu_2$, i.e., $H_0: \mu_1 - \mu_2 = 0$.]

-2 for each calculation error

Question 3: Probability and Distributions (**48 Points**) -2 if no final result

During a recent departmental dinner, there were plates with carrot cake and plates with chocolate cake randomly distributed across the tables. We can assume that the probability to obtain a piece of carrot cake is the same as the probability to obtain a piece of chocolate cake. **Show your work!** $P(\text{carrot}) = P(\text{chocolate}) = \frac{1}{2}$

1. **(6 Points)** What is the probability that for a table where 6 plates are served, all 6 plates contain carrot cake? Write your answer as a fraction and **not** as a decimal!

$$P(\text{all 6 carrot}) = \left(\frac{1}{2}\right)^6 = \frac{1}{64} \quad 6 \quad = 0.015625$$

2. **(6 Points)** What is the probability that for a table where 6 plates are served, at least 1 plate contains chocolate cake. Write your answer as a fraction and **not** as a decimal!

$$P(\text{at least 1 chocolate}) = 1 - P(\text{all 6 carrot})$$
$$= 1 - \frac{1}{64} = \frac{63}{64} \quad 6 \quad = 0.984375$$

3. **(6 Points)** There were 15 tables overall that each got 6 plates. Using the proper statistical notation, write down a probability model to describe the underlying distribution that relates to the fact that a table for which 6 plates are served, all 6 plates contain carrot cake (this would be called a success here). Properly use your answer from part 1. above as a part of your answer here. We can assume that the kitchen bakes an extremely large amount of both types of cakes and plates are randomly served at each table.

6 for any valid answer

$$\mathrm{Bin}\Big(\underset{n}{\underset{\uparrow}{15}}, \underset{p}{\underset{\uparrow}{\tfrac{1}{64}}}\Big) \quad \text{or} \quad B\Big(\underset{n}{\underset{\uparrow}{15}}, \underset{p}{\underset{\uparrow}{\tfrac{1}{64}}}\Big)$$

or: $S = \{A, B\}$, A: all 6 pieces are carrot, $P(A) = \frac{1}{64} = p$
B: not all 6 pieces are carrot, $P(B) = \frac{63}{64} = 1-p$

4. **(6 Points)** Given these 15 tables, what is the expected number of tables that would be served with carrot cake on all 6 plates.

$$\mu = n \cdot p$$
$$= 15 \cdot \frac{1}{64} = \frac{15}{64} = \underline{0.23 \text{ tables}} \quad 6,$$

[i.e., far less than a single table]

5. **(6 Points)** What is the corresponding standard deviation?

$$\sigma = \sqrt{n \cdot p \cdot (1-p)}$$
$$= \sqrt{15 \cdot \frac{1}{64} \cdot \frac{63}{64}}$$
$$= \sqrt{0.2307} \approx \underline{\underline{0.48 \text{ tables}}}$$

6

6. **(6 Points)** Can we use a Normal approximation to determine the probability that 7 to 10 of the 15 tables would be served with carrot cake on all 6 plates? **Yes** / **NO.** Circle the correct answer and explain why you chose this answer.

4

$n \cdot p = 0.23 < 10$ as determined in 4.; but it should hold that $n \cdot p \geq 10$ to use a Normal approximation 2

7. **(6 Points)** Assume that this is not just a departmental dinner, but a university-wide dinner with 1,000 tables on the Quad. The same conditions hold as before, i.e., 6 plates with either carrot cake or chocolate cake are randomly served at each table. Use a Normal approximation to determine the probability that 15 to 20 of the 1,000 tables would be served with carrot cake on all 6 plates. For the purpose of this question part, we believe that the necessary requirements hold.

$\mu_X = n \cdot p = 1{,}000 \cdot \frac{1}{64} = 15.625$ 1 $\left[\geq 10, \text{ and also } 1{,}000 \cdot \frac{63}{64} \geq 10 \right]$

$\sigma_X = \sqrt{n \cdot p \cdot (1-p)} = \sqrt{1{,}000 \cdot \frac{1}{64} \cdot \frac{63}{64}} = \sqrt{15.38} = 3.922$ 1

so, X approx. $N(15.625, 3.922)$

8. **(6 Points)** Now, do the same calculation as in the previous part, but use a Normal approximation with **continuity correction**.

$P(15 \leq X \leq 20)$ 1	$P(14.5 \leq X \leq 20.5)$ 2
$= P\left(\frac{15-15.625}{3.922} \leq Z \leq \frac{20-15.625}{3.922}\right)$	$= P\left(\frac{14.5-15.625}{3.922} \leq Z \leq \frac{20.5-15.625}{3.922}\right)$
$= P(-0.16 \leq Z \leq 1.12)$ 1	$= P(-0.29 \leq Z \leq 1.24)$ 2
$= 0.8686 - 0.4364 = \underline{\underline{0.4322}}$ 2	$= 0.8925 - 0.3859 = \underline{\underline{0.5066}}$ 2

[i.e., a difference of about 7% between the 2 calculations!]

<u>Question 4:</u> Multiple Choice Questions **(120 Points)**

Mark your answer for each multiple choice question in the table below. There is only one correct answer for each question. Each correct answer is worth 4 points.

Question	(a)	(b)	(c)	(d)	Question	(a)	(b)	(c)	(d)
1	○	○	○	●	16	○	○	○	●
2	○	●	○	○	17	○	○	●	○
3	○	●	○	○	18	○	○	●	○
4	●	○	○	○	19	○	○	●	○
5	○	○	○	●	20	○	○	○	●
6	○	○	●	○	21	○	○	○	●
7	○	○	●	○	22	○	●	○	○
8	●	○	○	○	23	○	○	●	○
9	○	○	●	○	24	●	○	○	○
10	○	●	○	○	25	○	○	●	○
11	○	○	○	●	26	○	○	○	●
12	●	○	○	○	27	○	○	○	●
13	○	○	●	○	28	○	○	●	○
14	●	○	○	○	29	○	○	●	○
15	○	○	○	●	30	●	○	○	○

Stat 2000, Final, Question 4 — Solutions

1. (d) Multistage sampling. Revisit Section 3.2 for this particular sampling design.

2. (b) If one looks only at the o's, which correspond to females, there is a clear downward trend indicating a negative association. Similarly for the +'s, which correspond to males.

3. (b) Larger samples (more tosses) give smaller spread. If the coin were tossed 100 times, the variability would go down further.

4. (a) The probability is the area under the density curve between 0 and 3/4. This region consists of two rectangles. The first is the region between 0 and 0.5, which has a height of 1.5 and a base of length 0.5, hence area 0.75. The second is the region between 0.5 and 3/4, which has a height of 0.5 and a base of length 1/4, hence area 0.125. The sum of these areas is 0.875. More mathematically, we calculate probability $=$ area $= (0.5 - 0) \cdot 1.5 + (0.75 - 0.5) \cdot 0.5 = 0.5 \cdot 1.5 + 0.25 \cdot 0.5 = 0.75 + 0.125 = 0.875$.

5. (d) Even if garlic has no effect on any of the 100 variables measured, by chance we would expect to see about 5 of them appearing to be statistically significant at the 0.05 level and about 1 of them appearing to be statistically significant at the 0.01 level. This is consistent with what was actually observed, indicating that the results may just be due to chance, not an effect of the garlic tablets.

6. (c) The variance σ^2 is $\frac{p \cdot (1-p)}{n} = \frac{0.3 \cdot 0.7}{100} = 0.0021$. So, the standard deviation σ is $\sqrt{0.0021} = 0.0458$.

7. (c) Recall that a level α two–sided significance test rejects a hypothesis $H_0 : \mu = \mu_0$ exactly when the value μ_0 falls outside a level $1 - \alpha$ confidence interval for μ. The confidence interval here suggests we should reject H_0 in favor of the two–sided alternative given at $\alpha = 0.05$, since the interval doesn't include 50. Since we reject at $\alpha = 0.05$, the P–value must be below 0.05.

8. (a) From Table D, using 19 - 1 = 18 df, we have

   ```
   P        .05      .025
   t*     1.734     2.101
   ```

   Since we exceed the upper 0.05 critical value (but do not exceed the upper 0.025 critical value), we reject the null hypothesis at $\alpha = 2 \cdot 0.05 = 0.10$, since we are carrying out a two–sided test.

9. (c) The z–score corresponding to the 98th percentile is 2.05. We need to multiply the standard deviation by 2.05 and add it to the mean, which gives $62 + 2.05 \cdot 8 = 78.40$.

10. (b) The median is the average of the third and fourth smallest observations. The third smallest must be at least -4. It will be -4 when $x \leq -4$, or else it will be larger. It can't possibly be smaller than -4. Similarly, the fourth smallest observation cannot exceed 0. So the median must be between these numbers.

11. (d) There were 3 students who earned B's who also studied less than 5 hours per week and 21 students who said they studied less than 5 hours per week. $3/21 = 1/7$.

12. (a) The sampling distribution of $\hat{p}$ is centered about p, the chance of a die showing a 2 or a 3 or a 4 or a 5 or a 6. The chance of any face of the die appearing is $1/6$, so the chance of showing a 2 or 3 or 4 or 5 or 6 is $1/6 \cdot 5 = 5/6$.

13. (c) X takes the value \$0 with probability $3/10$ (the probability of a blue ball), the value \$20 with probability $6/10$, and \$40 with probability $1/10$. The mean of X is given by $\mu_X = \$0 \cdot 3/10 + \$20 \cdot 6/10 + \$40 \cdot 1/10 = \$160/10 = \$16$.

14. (a) Probability rule 5 states that two events are independent if $P(A \text{ and } B)$ and $P(A) \cdot P(B)$ are equal. Here, $P(A) \cdot P(B) = 0.7 \cdot 0.6 = 0.42 = P(A \text{ and } B)$, so the events are independent.

15. (d) Revisit Section 2.2 related to correlation.

16. (d) We cannot say the Central Limit Theorem applies in this case because a sample size of 5 is not "large", so we cannot determine the probability with the provided information.

17. (c) When using t–procedures, we are using two estimates based on the sample — both the sample mean and the sample standard deviation. This additional uncertainty is reflected in t–multipliers being larger than z–multipliers (see Table D). Therefore, t–intervals are wider than z–intervals.

18. (c) The correct degrees of freedom are $(r-1)(c-1) = (4-1)(4-1) = 9$. From Table F, we have

```
p        .05       .025
Chi*    16.92     19.02
```

Because the value of the chi–square statistic falls between the tabled critical values of 16.92 and 19.02, the P–value must lie between 0.025 and 0.05.

19. (c) The degrees of freedom for a Chi–square test are $(r-1)(c-1) = (4-1)(5-1) = 12$.

20. (d) The z–score corresponding to the 90th percentile is 1.28. If we add 1.28 standard deviations to the mean, this gives the correct answer, i.e., $1.28 \cdot 20 + 516 = 25.6 + 516 = 541.60$.

21. (d) When we repeat an experiment many times, eventually we will get a significant result — just by chance!

22. (b) Calculate $\overline{x} \pm z^* \cdot \frac{\sigma}{\sqrt{n}} = 2.97 \pm 1.96 \cdot \frac{0.62}{\sqrt{75}} = 2.97 \pm 0.14 = (2.830, 3.110)$.

23. (c) Revisit Section 6.1 regarding the interpretation of confidence intervals.

24. (a) Calculate $5 + (-3) + (-5) + (-1) = -4$.

25. (c) Resort the date from smallest to largest:

−5	−3	−1	2	5	10

Then calculate $-3 + (-1) + 2 + 5 = 3$.

26. (d) We are summing up 6 times the same value (2), so $2+2+2+2+2+2 = 6 \cdot 2 = 12$.

27. (d) Calculate $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$.

28. (c) It is $0! = 1$ by definition.

29. (c) Calculate $\binom{5}{3} = \frac{5!}{3! \cdot (5-3)!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(3 \cdot 2 \cdot 1) \cdot (2 \cdot 1)} = 10$.

30. (a) Calculate $\binom{5}{5} = \frac{5!}{5! \cdot (5-5)!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(5 \cdot 4 \cdot 3 \cdot 2 \cdot 1) \cdot 1} = 1$.