

2013 Autumn STAT 22000 Session 02 Midterm Solutions

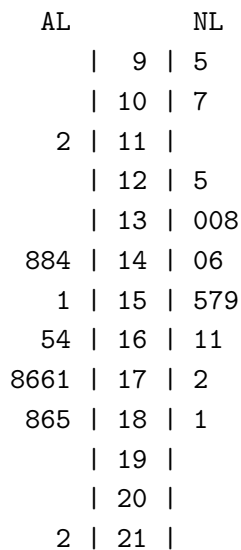
1. [MLB 2013] [6 points]

The Major League Baseball consisting of 15 teams from each of the two leagues: the American League (AL) and the National League (NL). The main difference between the two leagues is that pitchers take at bats in the National League but not in the American League. The table below lists the total number of homeruns hit during the 2013 season (except playoffs) for each of the 30 teams.

Team	League	Homeruns	Team	League	Homeruns	Team	League	Homeruns
ARI	NL	130	HOU	AL	148	PHI	NL	140
ATL	NL	181	KCR	AL	112	PIT	NL	161
BAL	AL	212	LAA	AL	164	SDP	NL	146
BOS	AL	178	LAD	NL	138	SEA	AL	188
CHC	NL	172	MIA	NL	95	SFG	NL	107
CHW	AL	148	MIL	NL	157	STL	NL	125
CIN	NL	155	MIN	AL	151	TBR	AL	165
CLE	AL	171	NYM	NL	130	TEX	AL	176
COL	NL	159	NYY	AL	144	TOR	AL	185
DET	AL	176	OAK	AL	186	WSN	NL	161

- (a) [5 points] Make a back-to-back stemplot of the numbers of homeruns for the two leagues.

Answer:



- (b) [1 point] Does it appear that teams in the American League have more homeruns?

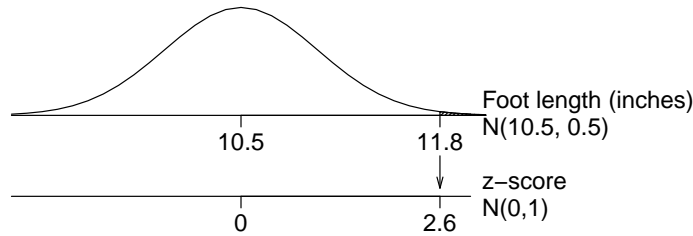
Answer: Yes, the peak of the stem plot is around 170-180 homerun for the American League, and around 150 for the National League. This is because pitchers have to bat in the National League, but not in the American League. A pitcher in the American League has a designated hitter to bat in his place. Such designated hitters are usually strong batters that bring up the total numbers of home runs.

2. [Foot Length] [8 points]

The average foot length for male adults in the U.S. is about 10.5 inches with a standard deviation of 0.5 inch. A histogram for the foot lengths is very close to the normal curve.

- (a) [4 points] What percentage of male adults in the U.S. have feet longer than 11.8 inches?

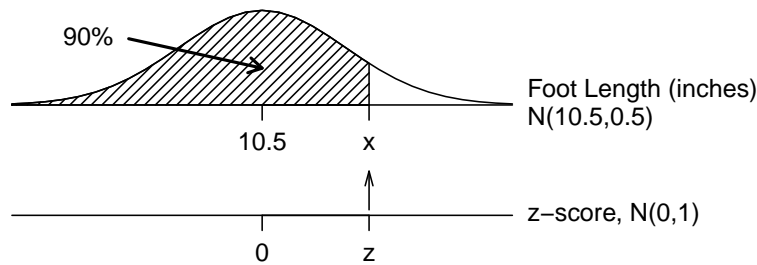
Answer: The percentage is the area under the foot lengths histogram to the right of 11.8, and histogram follows the normal curve centered at 10.5 with SD = 0.5, so the percentage is roughly the area of the shaded region in the figure below.

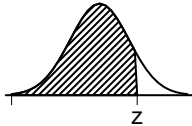


A foot length of 11.8 inches is $\frac{11.8 - 10.5}{0.5} = 2.6$ in z -score. The area corresponds to $z = 2.6$ is about 0.9953. The area to the right of 2.6 under the standard normal curve is thus roughly $1 - 0.9953 = 0.0047 = 0.47\%$.

- (b) [4 points] If a guy has foot length at the 90th percentile, how long are his feet?

Answer: To find the foot length x corresponding to the 90th percentile, let $z = (x - 10.5)/0.5$ be the z -score of x



We thus want the z such that  = 90%, We can then look up the normal table

for the z value corresponds to an area of 0.9000, which is about 1.28 (between 1.28 and 1.29). Thus the 90th percentile x is about $10.5 + 0.5z = 10.5 + 0.5 \times 1.28 = 11.14$ in. (between 11.14 in. and 11.145 in.)

3. [Hollywood Movies I] [16 points]

Opening weekend box office revenue is an important source of income to the movie industry and a crucial preliminary indicator of long-run profitability of a motion picture. The following are some numerical summaries for the Opening-Weekend Revenues (in millions of dollars) for the 136 Hollywood movies in 2011.

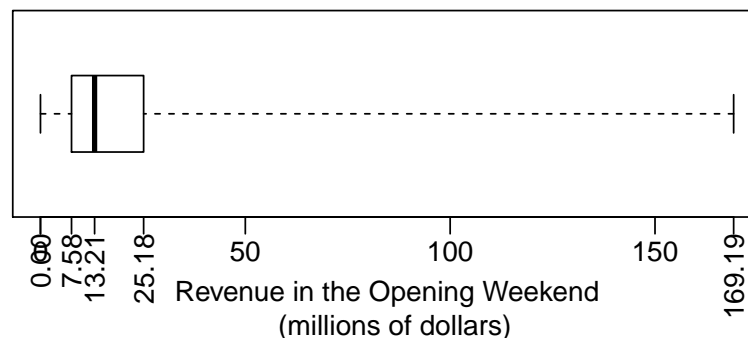
```
> summary(OpeningWeekend)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.00   7.71   13.10   20.34   25.00  169.20
> sd(OpeningWeekend)
[1] 24.80566
```

- (a) [2pts] Is the movie with the highest Opening-Weekend Revenue, \$169.20 million, an outlier based on the $1.5 \times IQR$ rule? (In fact, that movie was “*Harry Potter and the Deathly Hallows, Part 2*”).

Answer: A value is an outlier if it is above $Q3 + 1.5IQR = Q3 + 1.5 \times (Q3 - Q1) = 25.00 + (25.00 - 7.71) \times 1.5 = 50.935$. The maximum 169.20 is clearly above that and hence is an outlier.

- (b) [4pts] Sketch a boxplot (not the modified boxplot) of the Opening-Weekend Revenues. Draw the plot to scale. Clearly label the revenue values along the axis for each element of the boxplot.

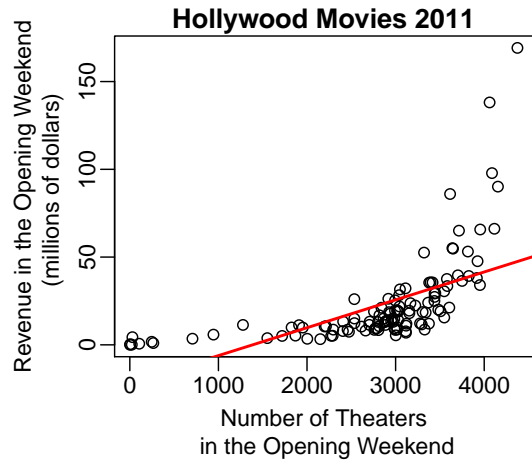
Answer:



- (c) [2pts] Give one reason why the five-number summary is a better numerical summary for the Opening-Weekend Revenue than the mean and the standard deviation.

Answer: We can see from a five-number summary if two tails of a distribution differ in length (by comparing distance of the median to $Q1$ and $Q3$ or to minimum and maximum.), but not from a mean + SD summary. The five-number summary can tell us the distribution of Opening-Weekend Revenue is highly right-skewed, but the mean + SD summary tells us nothing about the skewness.

One way to increase the opening weekend box office revenue is to increase the number of theaters showing the movie. The scatter plot below shows the revenue of 136 Hollywood movies in 2011 versus the number of theaters showing them in the opening weekend. Their correlation coefficient is $r = 0.589$.



(d) [2pts] Describe the relationship (form, direction, strength) between the two variables based on the scatter plot.

Answer: The plot shows a (moderate) positive nonlinear relationship.

(e) [2pts] Why is the correlation coefficient r NOT appropriate for measuring the strength of association between the two variables in the scatterplot?

Answer: The correlation coefficient r only measures the strength of linear relationship, but not non-linear relationship.

(f) [2pts] If John, a film producer, wants to predict the Opening-Weekend Revenue of a movie shown in 3000 theaters in the opening weekend using simple linear regression, his prediction will _____ the actual Opening-Weekend Revenue. The blank should be

- (i) tend to underestimate
- (ii) tend to overestimate
- (iii) about equally likely to be above or below

Circle one and explain briefly.

Answer: **(ii) tend to overestimate.** The regression line is roughly the red line on the scatter plot. We can see most of the dots around 3000 are below the regression line, which means linear regression tends to overestimate the actual Opening-Weekend Revenue for movies shown in 3000 theaters in the opening weekend.

(g) [2pts] Which of the following words best describes the shape of the histogram of the “number of theaters during the opening weekend”?

- (i) right-skewed
- (ii) symmetric
- (iii) left-skewed
- (iv) bimodal
- (v) uncertain, not enough information

Circle one. No explanation is required.

Answer: (iii) left-skewed. The density of dots are highest around 3000 theaters. The minimum number of theaters is about 0 and the maximum is around 4500. So the left tail of the histogram is longer than the right-tail.

4. [Hollywood Movies II] [18 points]

Opening weekend box office revenue is an important source of income to the movie industry and a crucial preliminary indicator of long-run profitability of a motion picture. Here are some R outputs as well as a scatter plot for the Opening-Weekend Revenue and World Gross Revenue for the 136 Hollywood movies in 2011.

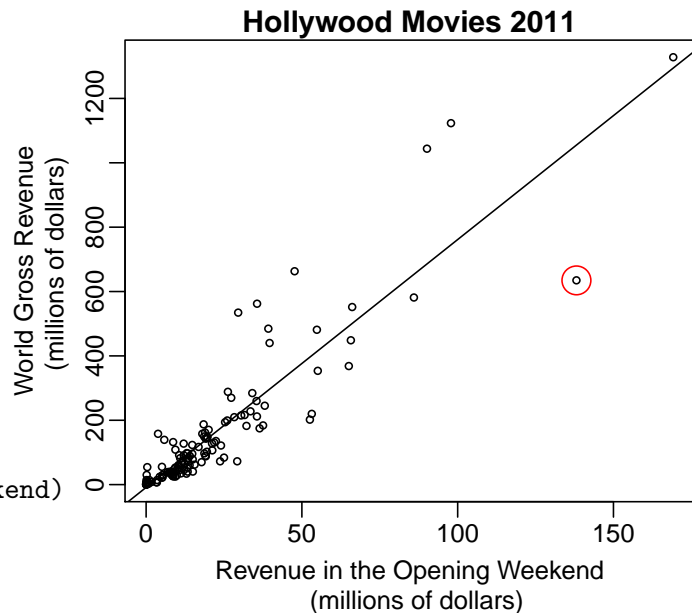
```
> mean(OpeningWeekend)
[1] 20.50153
> sd(OpeningWeekend)
[1] 24.96158
> mean(WorldGross)
[1] 149.217
> sd(WorldGross)
[1] 212.7556
> cor(OpeningWeekend, WorldGross)
[1] 0.9035918
> lm(WorldGross ~ OpeningWeekend)
```

Call:

```
lm(formula = WorldGross ~ OpeningWeekend)
```

Coefficients:

(Intercept)	OpeningWeekend
-8.678	7.702



- (a) [2 points] Write down the equation of the regression line for predicting the World Gross Revenue of a movie from its Opening-Weekend Revenue.

Answer: From the R output of the command `lm(WorldGross ~ OpeningWeekend)` above, we get

$$\begin{aligned} &\text{Predicted World Gross Revenue} \\ &= -8.678 \text{ millions of dollars} + 7.703 \times (\text{Opening-Weekend Revenue in millions of dollars}) \end{aligned}$$

- (b) [3 points] The movie “*Harry Potter and the Deathly Hallows, Part 2* had an Opening-Weekend Revenue of \$169.19 millions and a World Gross Revenue of \$1328.111 millions. Using the regression line, what is its predicted World Gross Revenue? What is the residual (prediction error)?

Answer: Substituting Opening-Weekend Revenue = 169.19 to the equation in part (a) we get the predicted World Gross Revenue to be

$$-8.678 + 7.703 \times 169.19 = 1294.593(\text{millions of dollars}),$$

and the residual is

$$\text{actual} - \text{predicted} = 1328.111 - 1294.593 = 33.518(\text{millions of dollars}).$$

- (c) [1 point] On the scatter plot, circle the dot has the largest NEGATIVE residual.

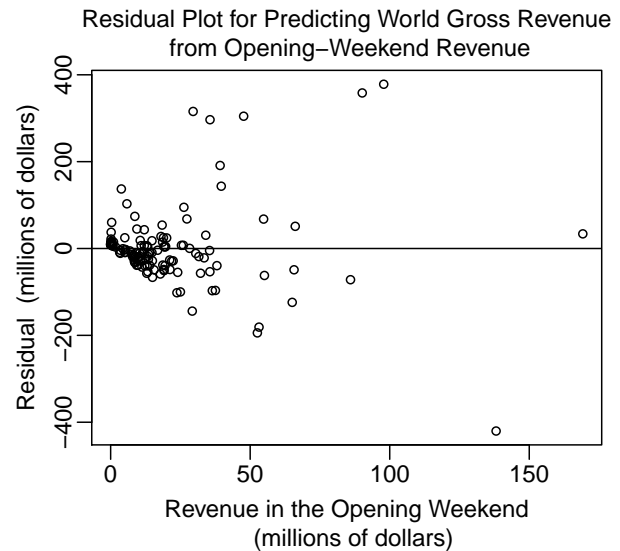
Answer: See the plot. It’s the point with a red circle around it. The dot with the largest negative residual is the dot with the largest vertical distance to the line.

- (d) [2 points] What fraction of the variance of the World Gross Revenue was explained by the least square regression on the Opening-Weekend Revenue?

Answer: $r^2 = (0.9035918)^2 = 0.8164781 \approx 81.65\%$. See p.117 of the textbook.

(e)

[6 points] The plot on the right is the residual plot for predicting the World Gross Revenue from the Opening-Weekend Revenues using simple linear regression. For each of the four statements below, determine whether it is TRUE or FALSE based on facts about residuals and the residual plot on the right.



_____ The prediction for World Gross Revenue is more accurate for movies with larger Opening-Weekend Revenues.

_____ The average of the residuals is about 0, though may not be exactly 0.

_____ The standard deviation of residuals is smaller than the standard deviation for the World Gross Revenues.

_____ On the scatter plot (not the residual plot), the residual of a point is the shortest (signed) distance from that point to the regression line.

Answer:

- False. The size of residuals increase with Opening-Weekend Revenues.
- False. The average of the residuals is always 0 in linear regression.
- True. The SD of residuals is smaller than the SD of the response (by a factor of $\sqrt{1 - r^2}$)
- False. The residual of a point is the vertical (signed) distance, not the shortest distance, from that point to the regression line.

- (f) [4 points] Write down the equation of the regression line for predicting the Opening-Weekend Revenue from the World Gross Revenues. (Note this line is different from the one in part (a).)

Answer: Here x = World Gross Revenues, y = the Opening-Weekend Revenue. The slope of the regression line is $r \times SD_y / SD_x = 0.9035918 \times 24.96158 / 212.7556 = 0.106014$. The intercept is

$$\bar{y} - (\text{slope})\bar{x} = 20.50153 - 0.106014 \times 149.217 = 4.682439.$$

The regression line is

$$\begin{aligned} & \text{Predicted Opening-Weekend Revenue} \\ & = 4.682439 \text{ millions of dollars} + 0.106014 \times (\text{World Gross Revenue in millions of dollars}) \end{aligned}$$

5. [Texting while Driving] [6 points]

To investigate whether or not sending text messages while driving impacts driving ability, we have 100 participants (50 men and 50 women) drive an obstacle course under two conditions:

- (1) No texting while driving, and
- (2) Sending five text messages while driving.

We measure the accuracy the subjects drove the obstacle course from a scale of 1 to 10 (1 = poor and 10 = excellent).

- (a) [3 points] What is the advantage of using a block design over the completely randomized design for this study? What should we block on?

Answer: The completely randomized design may end up more men in one condition than the other. Men and women on average have different driving skill and driving style. (The difference is considerable that auto insurance company even charge men and women with different rates.) So gender may confound with the effect of texting or not. By blocking on gender, we can ensure 25 men and 25 women drive in each of the two conditions. Some other possible blocking factors include driving experience (number of years people got their licence) and age.

- (b) [2 points] Briefly describe how to assign the 100 participants to the two groups using a randomized block design?

Answer: Here we block on gender. Randomly assign 25 of the 50 men to the first condition and the rest to the second. Assign the 50 women in the same way.

- (c) [1 point] Is this study blinded? Explain in one sentence.

Answer: Not blinded. Clearly the participants know whether they are suppose to text.

6. [Lie Detector] [12 points]

To reduce theft among employees, a company requires all of its employees to take a lie-detector test. The test asks if the employee has ever stolen from the company.

The test has been proven to correctly identify guilty employees 90% of the time and innocent employees pass the test at a rate of 96%.

All of the employees have to complete the test. Suppose that 5% of the employees are actually guilty (they steal from the company).

- (a) [3 points] What percentage of employees will fail the test?

Answer:

$$\begin{aligned} P(\text{guilty}) &= 0.05, P(\text{innocent}) = 0.95, \\ P(\text{fail the test}|\text{guilty}) &= 0.9, \\ P(\text{fail the test}|\text{innocent}) &= 0.04, \end{aligned}$$

By the rule of total probability,

$$\begin{aligned} P(\text{fail the test}) &= P(\text{fail the test}|\text{guilty})P(\text{guilty}) + P(\text{fail the test}|\text{innocent})P(\text{innocent}) \\ &= 0.9 \times 0.05 + 0.04 \times 0.95 = 0.083 \end{aligned}$$

- (b) [5 points] What percentage of those employees fail the test will actually be innocent? Why it is not fair to fire all employees who fail the test?

Answer:

$$\begin{aligned} P(\text{innocent}|\text{fail the test}) &= \frac{P(\text{innocent and fail the test})}{P(\text{fail the test})} \\ &= \frac{P(\text{fail the test}|\text{innocent})P(\text{innocent})}{P(\text{fail the test})} \\ &= \frac{0.04 \times 0.95}{0.083} = \frac{38}{83} = 0.4578 \end{aligned}$$

Here $P(\text{fail the test}) = 0.083$ comes from part (a).

- (c) [4 points] If those who fail the lie-detector test are given an independent second test, what percentage of those employees who fail both tests will actually be innocent?

Answer: By the Bayes' rule

$$\begin{aligned} &P(\text{innocent}|\text{fail both tests}) \\ &= \frac{P(\text{fail both tests}|\text{innocent})P(\text{innocent})}{P(\text{fail both tests}|\text{innocent})P(\text{innocent}) + P(\text{fail both tests}|\text{guilty})P(\text{guilty})} \\ &= \frac{0.04 \times 0.04 \times 0.95}{0.04 \times 0.04 \times 0.95 + 0.9 \times 0.9 \times 0.05} = 0.03617 \end{aligned}$$

7. [Laptop Warranty] [31 points]

Suppose a small company buys five laptops of the same model. Because the company is too small to hire IT technicians, whenever a laptop breaks down, the company simply replaces it with a new one, which costs \$1,000. From past experience, a laptop will be broken in 3 years with probability 0.2. Suppose the breakdown of one laptop is independent of the breakdown of the rest.

- (a) [3 points] What is the probability that at least one of the 5 laptops break down in 3 years?

Answer: By complementation rule

$$\begin{aligned} P(\text{at least one break down}) &= 1 - P(\text{none of the five breaks down}) \\ &= 1 - (0.8)^5 = 0.67232 \end{aligned}$$

- (b) [3 points] What is the probability that EXACTLY 2 of the 5 laptops break down in 3 years?

Answer: As the laptops breakdown independently, the number N of laptops breakdown in 3 years has a binomial distribution

$$N \sim B(n = 5, p = 0.2).$$

Using Binomial formula $P(N = k) = \frac{n!}{k!(n - k)!} p^k (1 - p)^{n - k}$, for $k = 2$

$$\begin{aligned} P(N = 2) &= \frac{5!}{2!3!} (0.2)^2 (0.8)^3 = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 3 \cdot 2 \cdot 1} \times 0.02048 \\ &= 10 \times 0.02048 = 0.2048. \end{aligned}$$

- (c) [3 points] Suppose the manufacturer sells a three-year warranty, which promises to fix any problem incurred or replace it with a new one. The warranty charges \$99 per laptop. Should the company purchase the warranty?

Answer: Yes, it should. For the company, the expected maintenance cost for a laptop is $\$1000 \times 0.2 = \200 . The warranty charges \$99 per laptop only. It is a good deal.

For simplicity, suppose only two parts in a laptop can cause its breakdown — the screen and the motherboard, which happen in 3 years with probability 0.1 and 0.1, and cost \$100 and \$300 to fix, respectively. Suppose the laptop can breakdown at most once in 3 years. Furthermore, suppose the breakdown of parts in one laptop is independent of the breakdown of others.

- (d) [2 points] Based on the assumptions above, are the breakdown of the motherboard and the breakdown of the screen independent? Explain briefly.

Answer: Not independent. When the motherboard breaks down, the screen will not break down.

$$P(\text{screen breaks down} | \text{motherboard breaks down}) = 0 \neq P(\text{screen breaks down}) = 0.1.$$

- (e) [4 points] Let X be the random variable for the potential cost of the manufacturer to cover one laptop. What are the possible values that X may take? Write down the probability distribution of X .

Answer:

Value of X	\$0	\$100	\$300
Probability	0.8	0.1	0.1

- (f) [4 points] Find the mean and the variance of X .

Answer:

$$\begin{aligned} \mu &= E(X) = 0 \cdot 0.8 + 100 \cdot 0.1 + 300 \cdot 0.1 = 40 \\ \sigma^2 &= \text{Var}(X) = 0^2 \cdot 0.8 + 100^2 \cdot 0.1 + 300^2 \cdot 0.1 - 40^2 = 0 + 1000 + 9000 - 1600 = 8400 \end{aligned}$$

- (g) [4 points] Suppose the manufacturer sold 1000 warranties (to cover 1000 laptops). What are the expected value (i.e., the mean) and the standard deviation of the total cost to cover these 1000 laptops?

Answer: Let X_i be the cost to cover the i th laptop. Then the total cost to cover these 1000 laptops is $S = X_1 + X_2 + \cdots + X_{1000}$. As the laptops breakdown independently, the costs $X_1, X_2, \dots, X_{1000}$ are i.i.d.

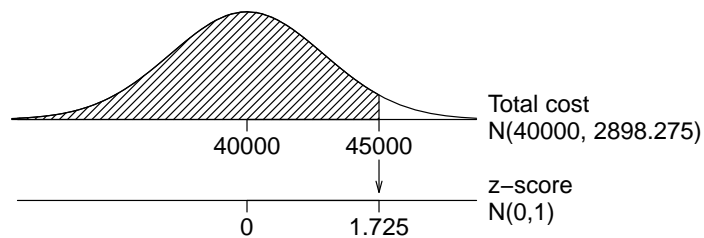
$$\begin{aligned} E(S) &= 1000 \times E(X_1) = \$40 \times 1000 = \$40000 \\ \text{Var}(S) &= 1000 \times \text{Var}(X_1) = 1000 \times 8400 = 8400000 \\ \text{SD}(S) &= \sqrt{\text{Var}(S)} = \sqrt{8400000} = \$2898.275 \end{aligned}$$

- (h) [4 points] What is the probability that total cost for cover these 1000 laptops is less than \$45,000¹?

Answer: As $n = 1000$ is large, by CLT and part (g), S is approximately normal

$$X \sim N(40000, 2898.275).$$

Thus $P(X > 45000)$ is roughly the area of the shaded region in the figure below.



The z -score of 45000 is $\frac{45000 - 40000}{2898.275} \approx 1.725$. The area corresponds to $z = 1.725$ is about 0.9577, which is the probability we are looking for.

¹Recall the manufacturer will have a revenue of $\$99 \times 1000 = 99,000$ by selling 1000 warranties. If the total cost is less than \$45,000, the manufacturer will have a net profit of $\$99,000 - \$45,000 = \$44,000$.

- (i) [4 points] What is the probability that among these 1000 covered laptops, more than 220 of them will breakdown in 3 years?

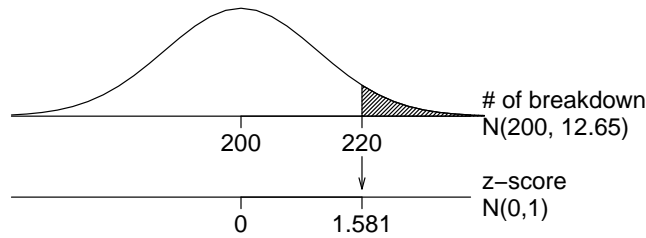
Answer: As the laptops breakdown independently, the number X of laptops breakdown in 3 years has a binomial distribution

$$X \sim B(n = 1000, p = 0.2).$$

As $n = 1000$ is large, by CLT, X is approximately normal

$$X \sim N(np, \sqrt{np(1-p)}) = N(1000 \times 0.2, \sqrt{1000 \times 0.2 \times 0.8}) = N(200, 12.65).$$

Thus $P(X > 220)$ is roughly the area of the shaded region in the figure below.



The z -score of 220 is $\frac{220 - 200}{12.65} \approx 1.58$. The area corresponds to $z = 1.58$ is about 0.9429. The probability is thus roughly $1 - 0.9429 = 0.0570 = 5.7\%$.

8. [Pets Survey] [4 points]

The state of Illinois wants to know about the fertility of owned dogs in Illinois. They select a simple random sample of 50 households from each county in the state and ask how many dogs they have, and are they spayed or neutered.

- (a) [2 points] What type of sampling method was used to collect the data?

- (i) systematic county sampling (ii) stratified sampling
 (iii) multistage cluster sampling (iv) simple random sampling

Circle one. No explanation is required.

Answer: **(ii) stratified sampling**

- (b) [2 points] Fill in the blank. The average number of dogs per household for the Cook County in Illinois is a _____.

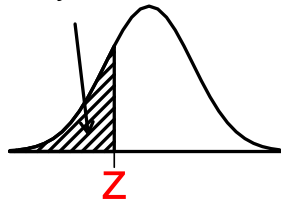
- (i) population (ii) sample (iii) parameter (iv) statistics

Circle one. No explanation is required.

Answer: **(iii) parameter**

.....END OF EXAM

table entry = shaded area



Standard Normal Probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

