Example 3.6: Ear Infections
(from Bernard Rosner's Fundamentals of Biostatistics)

A common symptom of otitis media (ear infection) in young children is the prolonged presence of fluid in the middle ear. The hypothesis has been proposed that babies who are breast-fed for at least 1 month may build up some immunity against the effects of the disease. A small study of 24 pairs of babies is set up, where the babies are matched on a one-to-one basis according to age, sex, socioeconomic status, and type of medications taken. One member of the matched pair is a breast-fed baby, and the other was bottle-fed. The researchers recorded the duration (in days) of fluid in the middle ear after the first episode of otitis media. The results from the 24 pairs are shown below:

Pair	Breast-fed duration	Bottle-fed duration
1	20	18
2	11	35
3	3	7
4	24	182
5	7	6
6	28	33
7	58	223
* 8	7	7 ← tie
9	39	57
10	17	76
11	17	186
12	12	29
13	52	39
14	14	15
15	12	21
16	30	28
17	7	8
18	15	27
19	65	77
20	10	12
21	7	8
22	18	16
23	34	28
24	25	20

Here we are directly comparing how long fluid remained in the ears of similar infants, the difference being one infant was breast fed and the other was bottle fed.

Research Question: Is there a statistically significant *difference* in the duration of ear infection between the breast-fed and the bottle-fed babies?

similar ~ on the basis of age, sex, socioeconomic status, and meds taken.

Because of the matched-pairs nature of the data, the comparisons should be made WITHIN each pair of babies. Therefore, it makes sense to consider the difference between the two groups.

Pair	Breast-fed duration	Bottle-fed duration	Difference = Breast - Bottle
1	20	18	2
2	11	35	-24
3	3	7	-4
4	24	182	-158
5	7	6	1
6	28	33	-5
7	58	223	-165
8	7	7	0
9	39	57	-18
10	17	76	-59
11	17	186	-169
12	12	29	-17
13	52	39	13
14	14	15	-1
15	12	21	-9
16	30	28	2
17	7	8	-1
18	15	27	-12
19	65	77	-12
20	10	12	-2
21	7	8	-1
22	18	16	2
23	34	28	6
24	25	20	5

Dark cells
Breast did better than bottle

Lighter cells
Bottle did better than breast.

~ tie, drop!

Questions:

1. In how many pairs did the breast-fed baby do better than the bottle-fed baby?
   16 pairs
2. In how many pairs did the bottle-fed do better?
   7 pairs
3. Pair #8 is a tie. What does this mean in the context of the problem? Does this pair provide evidence for bottle-fed doing better, breast-fed doing better, or neither? Explain.
   - same length of fluid on the ears (days)
   - neither ~ i.e. no information provided so we drop them.

Recall that we are restricted to considering only two outcomes when using the binomial distribution. So, we will not include Pair #8 in our analysis. "coin landing on its edge"

Questions:

4. If the tie is removed, how many pairs do we have in the sample?

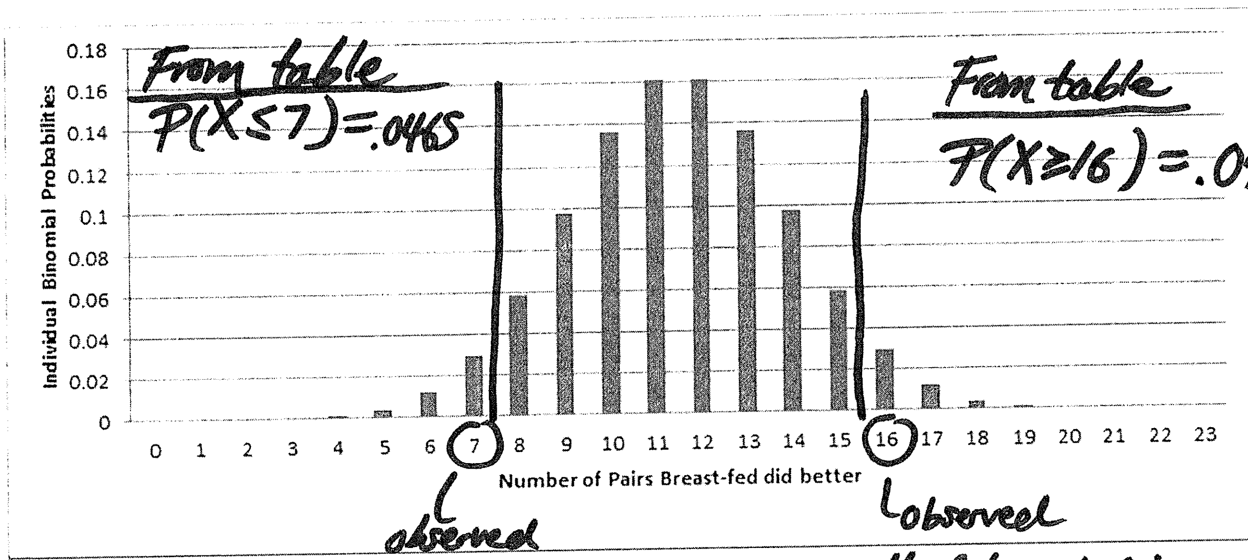
   23 pairs

5. If there is really no difference in the duration of ear infection between breast-fed and bottle-fed babies, what is the probability that the bottle-fed baby will do better than the breast-fed baby in any given pair?

   50% chance, i.e. p=.50

Now, we can use the binomial distribution to investigate this research question.

	A	B	C	D	E	F	G
1				# of pairs			
2	n =	23					
3	p =	0.5		Number of Successes, x	Individual Binomial Probabilities, P(exactly x)	Cumulative Binomial Probabilities, P(x or fewer)	P(x or more)
4				0	0.000000	0.000000	1.000000
5				1	0.000003	0.000003	1.000000
6	equally likely			2	0.000030	0.000033	0.999997
7	as to which			3	0.000211	0.000244	0.999967
8	infant in each			4	0.001056	0.001300	0.999756
9	pair will do			5	0.004011	0.005311	0.998700
10	better.			6	0.012034	0.017345	0.994689
11				7	0.029225	0.046570	0.982655
12				8	0.058450	0.105020	0.953430
13				9	0.097417	0.202436	0.894980
14				10	0.136383	0.338820	0.797564
15				11	0.161180	0.500000	0.661180
16				12	0.161180	0.661180	0.500000
17				13	0.136383	0.797564	0.338820
18				14	0.097417	0.894980	0.202436
19				15	0.058450	0.953430	0.105020
20				16	0.029225	0.982655	0.046570
21				17	0.012034	0.994689	0.017345
22				18	0.004011	0.998700	0.005311
23				19	0.001056	0.999756	0.001300
24				20	0.000211	0.999967	0.000244
25				21	0.000030	0.999997	0.000033
26				22	0.000003	1.000000	0.000003
27				23	0.000000	1.000000	0.000000



\# of bottle doing than breast

\# of breast doing better than bottle

Questions:

6. Based on the binomial distribution, was the observed value (16 pairs in which the breast-fed baby did better) unlikely to have happened by chance? What conclusion, if any, can we make regarding the research question?

As we are not specifying in our research question which infant in each pair we expect to do better, the probability of obtaining results as extreme or more extreme than those observed is

$.0465 + .0465 = .0930$ or 9.3% a chance

we do Not ~~have~~ have evidence to conclude similar breast and bottle fed infants differ in terms of the length of time they have fluid in their ears.

FORMAL HYPOTHESIS TESTING

In the previous examples, we have used the binomial distribution to make statistical inferences in problems involving a single categorical variable. Next, we will add more structure to these statistical investigations by introducing a procedure which statistician's call *hypothesis testing*.

Hypothesis testing is a procedure, based on sample evidence and probability, used to test claims regarding a population parameter. The test will measure how well our observed data agrees with a statement concerning the parameter of interest.

Before you begin a hypothesis test, you should clearly state your question of interest. For instance, let's reconsider the research question from three of our previous examples.

Example	**Research Question**
Example 3.3: Evaluating Deafness	Do these data provide statistical evidence the subject is answering incorrectly on purpose?
Example 3.4: Are women passed over for managerial training?	Is there statistical evidence for gender discrimination against females?
Example 3.5: Success rate of a new drug	Is there statistical evidence that the new drug has a higher success rate than the current one?
Example 3.6: Ear Infections	Is there a statistically significant *difference* in the duration of ear infection between the breast-fed and the bottle-fed babies?

The hypothesis test is then carried out as follows.

Step One: Writing The Null And Alternative Hypothesis

- The null hypothesis, H_o, is assumed true until evidence indicates otherwise. This usually contains statements of equality (e.g., "the probability or population proportion is equal to…")

- The alternative hypothesis, H_a, is what we are trying to show. Therefore, the question of interest is restated here in the alternative hypothesis. Also, this usually contains statements such as "the probability or population proportion is not equal to…" or "is greater than…" or "is less than…"

For our four examples, the null and alternative hypotheses are shown below.

Research Question	Hypotheses
Do these data provide statistical evidence the subject is answering incorrectly on purpose?	H_o: The subject is just guessing; that is, the probability of an incorrect guess is 50%. H_a: The subject is answering incorrectly on purpose; that is, the probability of an incorrect guess is greater than 50%.
Is there statistical evidence for gender discrimination against females?	H_o: The selection process is fair; that is, the probability a female is selected is 60%. H_a: The selection process is biased against females; that is, the probability a female is selected is less than 60%.
Is there statistical evidence that the new drug has a higher success rate than the current one?	H_o: The selection process is fair; that is, the probability a female is selected is 60%. H_a: The selection process is biased against females; that is, the probability a female is selected is less than 60%.

Is there a statistically significant *difference* in the duration of ear infection between the breast-fed and the bottle-fed babies?	H_o: There is no difference in duration of fluid between bottle- and breast-fed babies; that is, the probability the breast-fed baby in each pair did better is equal to 50%. $p = .50$ H_a: There is a difference in duration of fluid between bottle- and breast-fed babies; that is, the probability the breast-fed baby in each pair did better is different from 50%. $p \neq .50$

Step Two: Finding Either the Critical Value/Region or the p-value

Finding the correct *critical value/region* or *p-value* for a given problem depends on whether the test is ***upper-tailed, lower-tailed, or two-tailed.***

Example 3.3, Revisited: An Upper-Tailed Test

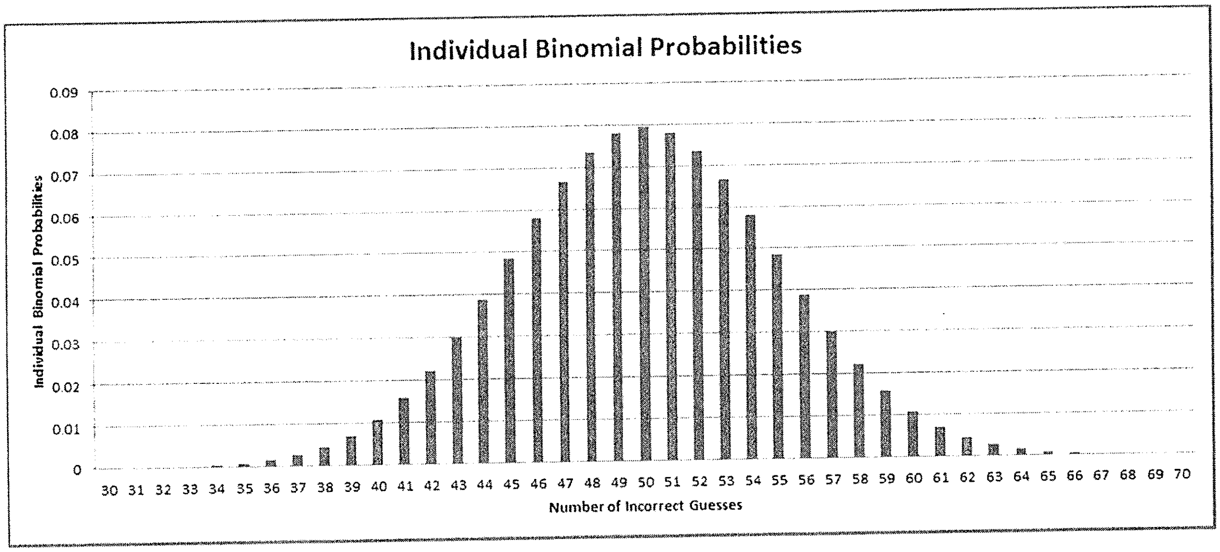
The example regarding the evaluation of deafness is an example of an *upper-tailed* test because we were trying to show that the observed number of incorrect guesses was *higher* than expected if the null hypothesis were true. Recall the subject got 64 of the 100 audio tests wrong.

To find the critical value/region or p-value for this example, we must consider a binomial distribution with n = 100 and p = .50.

n = 100
p = 0.5

	Number of Successes, x	Individual Binomial Probabilities, P(exactly x)	Cumulative Binomial Probabilities, P(x or fewer)	P(x or more)
49	45	0.048474	0.184101	0.864373
50	46	0.057958	0.242059	0.815899
51	47	0.066590	0.308650	0.757941
52	48	0.073527	0.382177	0.691350
53	49	0.078029	0.460205	0.617823
54	50	0.079589	0.539795	0.539795
55	51	0.078029	0.617823	0.460205
56	52	0.073527	0.691350	0.382177
57	53	0.066590	0.757941	0.308650
58	54	0.057958	0.815899	0.242059
59	55	0.048474	0.864373	0.184101
60	56	0.038953	0.903326	0.135627
61	57	0.030069	0.933395	0.096674
62	58	0.022292	0.955687	0.066605
63	59	0.015869	0.971556	0.044313
64	60	0.010844	0.982400	0.028444
65	61	0.007111	0.989511	0.017600
66	62	0.004473	0.993984	0.010489
67	63	0.002698	0.996681	0.006016
68	64	0.001560	0.998241	0.003319
69	65	0.000864	0.999105	0.001759
70	66	0.000458	0.999563	0.000895
71	67	0.000232	0.999796	0.000437
72	68	0.000113	0.999908	0.000204
73	69	0.000052	0.999961	0.000092
74	70	0.000023	0.999984	0.000039

Finding the Critical Value and Critical Region:



Finding the p-value:

	A	B	C	D	E	F	G
1							
2	n =	100					
3	p =	0.5		Number of Successes, x	Individual Binomial Probabilities, P(exactly x)	Cumulative Binomial Probabilities, P(x or fewer)	P(x or more)
49				45	0.048474	0.184101	0.864373
50				46	0.057958	0.242059	0.815899
51				47	0.066590	0.308650	0.757941
52				48	0.073527	0.382177	0.691350
53				49	0.078029	0.460205	0.617823
54				50	0.079589	0.539795	0.539795
55				51	0.078029	0.617823	0.460205
56				52	0.073527	0.691350	0.382177
57				53	0.066590	0.757941	0.308650
58				54	0.057958	0.815899	0.242059
59				55	0.048474	0.864373	0.184101
60				56	0.038953	0.903326	0.135627
61				57	0.030069	0.933395	0.096674
62				58	0.022292	0.955687	0.066605
63				59	0.015869	0.971556	0.044313
64				60	0.010844	0.982400	0.028444
65				61	0.007111	0.989511	0.017600
66				62	0.004473	0.993984	0.010489
67				63	0.002698	0.996681	0.006016
68				64	0.001560	0.998241	0.003319
69				65	0.000864	0.999105	0.001759
70				66	0.000458	0.999563	0.000895
71				67	0.000232	0.999796	0.000437
72				68	0.000113	0.999908	0.000204
73				69	0.000052	0.999961	0.000092
74				70	0.000023	0.999984	0.000039

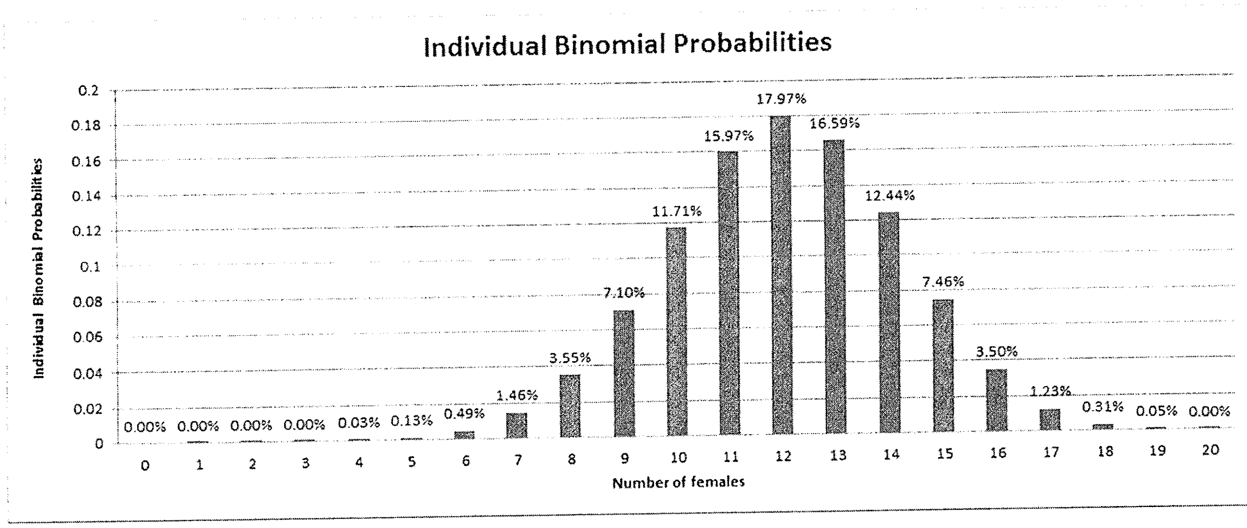
Example 3.4, Revisited: A Lower-Tailed Test

This example regarding possible discrimination against females is an example of a *lower-tailed* test because we were trying to show that the observed number of females was *lower* than expected if the null hypothesis were true. Recall that 9 of 20 individuals selected were female.

To find the critical value/region or p-value for this example, we must consider a binomial distribution with n = 20 and p = .60.

	A	B	C	D	E	F	G
1							
2	n =	20					
3	p =	0.6		Number of Successes, x	Individual Binomial Probabilities, P(exactly x)	Cumulative Binomial Probabilities, P(x or fewer)	P(x or more)
4				0	0.000000	0.000000	1.000000
5				1	0.000000	0.000000	1.000000
6				2	0.000005	0.000005	1.000000
7				3	0.000042	0.000047	0.999995
8				4	0.000270	0.000317	0.999953
9				5	0.001294	0.001612	0.999683
10				6	0.004854	0.006466	0.998388
11				7	0.014563	0.021029	0.993534
12				8	0.035497	0.056526	0.978971
13				9	0.070995	0.127521	0.943474
14				10	0.117142	0.244663	0.872479
15				11	0.159738	0.404401	0.755337
16				12	0.179706	0.584107	0.595599
17				13	0.165882	0.749989	0.415893
18				14	0.124412	0.874401	0.250011
19				15	0.074647	0.949048	0.125599
20				16	0.034991	0.984039	0.050952
21				17	0.012350	0.996389	0.015961
22				18	0.003087	0.999476	0.003611
23				19	0.000487	0.999963	0.000524
24				20	0.000037	1.000000	0.000037

Finding the Critical Value and Critical Region:



Finding the p-value:

n = 20

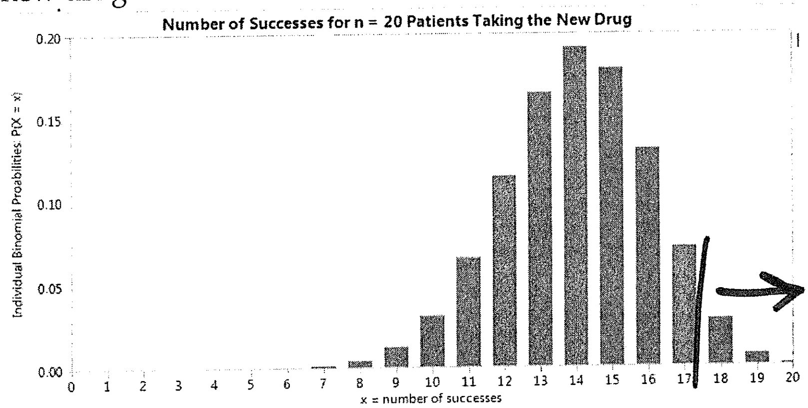
p = 0.6

Number of Successes, x	Individual Binomial Probabilities, P(exactly x)	Cumulative Binomial Probabilities, P(x or fewer)	P(x or more)
0	0.000000	0.000000	1.000000
1	0.000000	0.000000	1.000000
2	0.000005	0.000005	1.000000
3	0.000042	0.000047	0.999995
4	0.000270	0.000317	0.999953
5	0.001294	0.001612	0.999683
6	0.004854	0.006466	0.998388
7	0.014563	0.021029	0.993534
8	0.035497	0.056526	0.978971
9	0.070995	0.127521	0.943474
10	0.117142	0.244663	0.872479
11	0.159738	0.404401	0.755337
12	0.179706	0.584107	0.595599
13	0.165882	0.749989	0.415893
14	0.124412	0.874401	0.250011
15	0.074647	0.949048	0.125599
16	0.034991	0.984039	0.050952
17	0.012350	0.996389	0.015961
18	0.003087	0.999476	0.003611
19	0.000487	0.999963	0.000524
20	0.000037	1.000000	0.000037

p-value = .1275 or a 12.75% chance

Example 3.5, Revisited: An Upper-Tailed Test

The is another example of an *upper-tailed* test because we were trying to show that the number of successes for patients taking the drug *higher* than expected if the null hypothesis were true. Recall in the clinical trial the researchers found 18 patients being successfully treated with the new drug out of $n = 20$.



Finding the Critical Value and Critical Region:

A	B	C
n =	20	
p =	0.7	

D	E	F	G
		x or fewer successes	x or more successes
Number of Successes, x	Individual Binomial Probabilities, P(exactly x)	Cumulative Binomial Probabilities, P(x or fewer)	P(x or more)
0	0.000000	0.000000	1.000000
1	0.000000	0.000000	1.000000
2	0.000000	0.000000	1.000000
3	0.000001	0.000001	1.000000
4	0.000005	0.000006	0.999999
5	0.000037	0.000043	0.999994
6	0.000218	0.000261	0.999957
7	0.001018	0.001279	0.999739
8	0.003859	0.005138	0.998721
9	0.012007	0.017145	0.994862
10	0.030817	0.047962	0.982855
11	0.065370	0.113331	0.952038
12	0.114397	0.227728	0.886669
13	0.164262	0.391990	0.772272
14	0.191639	0.583629	0.608010
15	0.178863	0.762492	0.416371
16	0.130421	0.892913	0.237508
17	0.071604	0.964517	0.107087
18	0.027846	0.992363	0.035483
19	0.006839	0.999202	0.007637
20	0.000798	1.000000	0.000798

Finding the p-value: $= .0355 < .05$ @ 5%

Thus we have evidence new drug has a higher success rate, i.e. greater than 70%.

Example 3.6, Revisited: A Two-Tailed Test

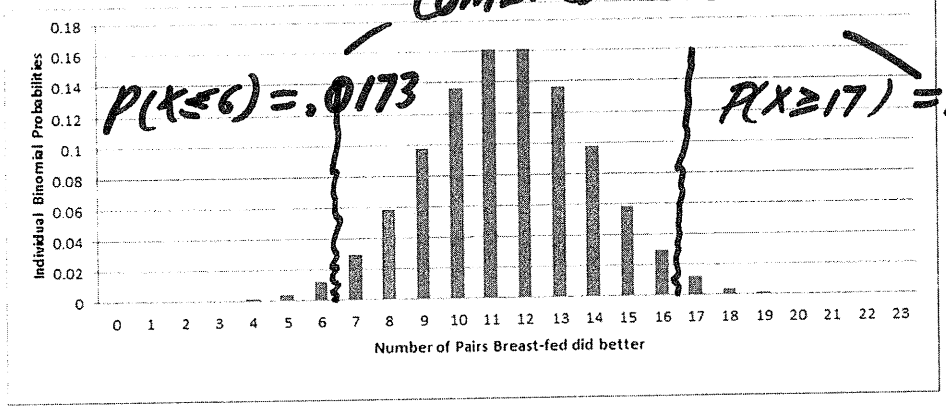
This example regarding ear infections is an example of a *two-tailed* test because we were simply looking for a *difference* between the two groups. Recall in 16 of 23 non-tied pairs the breast fed infant had shorter effusion times.

To find the critical value/region or p-value for this example, we must consider a binomial distribution with n = 23 and p = .50.

$X \leq 6$ or $X \geq 17$

Finding the Critical Value and Critical Region:

Combined area $= .0346 < .05$



Finding the p-value: $2(.0465) = P(X \leq 7) + P(X \geq 16) = .0930$

	A	B	C	D	E	F	G
1							
2	n =	23					
3	p =	0.5		Number of Successes, x	Individual Binomial Probabilities, P(exactly x)	Cumulative Binomial Probabilities, P(x or fewer)	P(x or more)
4				0	0.000000	0.000000	1.000000
5				1	0.000003	0.000003	1.000000
6				2	0.000030	0.000033	0.999997
7				3	0.000211	0.000244	0.999967
8				4	0.001056	0.001300	0.999756
9				5	0.004011	0.005311	0.998700
10				6	0.012034	crit 0.017345	0.994689
11				7	0.029225	0.046570	0.982655
12				8	0.058450	0.105020	0.953430
13				9	0.097417	0.202436	0.894980
14				10	0.136383	0.338820	0.797564
15				11	0.161180	0.500000	0.661180
16				12	0.161180	0.661180	0.500000
17				13	0.136383	0.797564	0.338820
18				14	0.097417	0.894980	0.202436
19				15	0.058450	0.953430	0.105020
20				16	0.029225	0.982655	0.046570
21				17	0.012034	0.994689	0.017345 crit
22				18	0.004011	0.998700	0.005311
23				19	0.001056	0.999756	0.001300
24				20	0.000211	0.999967	0.000244
25				21	0.000030	0.999997	0.000033
26				22	0.000003	1.000000	0.000003
27				23	0.000000	1.000000	0.000000

Step Three: Writing a Conclusion Regarding the Research Question

Critical Value/Region Method:

- If the observed value falls in the critical region, then we have evidence to support the alternative hypothesis (i.e., the research question).
- If the observed value does not fall in the critical region, then we say that we have no evidence to support the research question.

P-value Method:

- If the p-value falls below .01, we have very strong evidence to support the alternative hypothesis (i.e., the research question).
- If the p-value falls below .05, we have strong evidence to support the alternative hypothesis (i.e., the research question).
- If the p-value falls below .10 but above .05, we have "marginal" evidence to support the alternative hypothesis (i.e., the research question).
- If the p-value is above .10, we have no evidence to support the research question

Using these rules, write conclusions for each of our four examples:

Hypotheses	Conclusion
H_o: The subject is just guessing; that is, the probability of an incorrect guess is 50%. H_a: The subject is answering incorrectly on purpose; that is, the probability of an incorrect guess is greater than 50%.	observed 64 incorrect out of 100 Critical Region: $X \geq 59$ ⟹ p-value = .0033 < .01 Reject Ho
H_o: The selection process is fair; that is, the probability a female is selected is 60%. H_a: The selection process is biased against females; that is the probability a female is selected is less than 60%.	observed 9 of 20 female Critical Region: $X \leq 7$ p-value = .1275 > .10 NO EVIDENCE
H_o: The new drug is no better than the current one, i.e. the success rate of the new drug is 70%. H_a: The new drug is better than the current one, i.e. the success rate of the new drug is greater than 70%.	observed 18 out of 20 successes Critical Region: $X \geq 18$ p-value = .0355 < .05 STRONG EVIDENCE
H_o: There is no difference in duration of fluid between bottle- and breast-fed babies. H_a: There is a difference in duration of fluid between bottle- and breast-fed babies.	observed 16 pairs where breast did better 7 pairs where bottle did better CRITICAL REGION: $X \leq 6$ or $X \geq 17$ p-value = .0930 "marginal" evidence of a difference in the ear effusion times.

Deafness:

Conclusion

We have very strong evidence to conclude the patient is answering over 50% of questions incorrectly ($p = .0033$). We can logically conclude he is doing this in order to appear hearing impaired.