18.05 Problem Set 8, Spring 2022 Solutions

Problem 1. (35: 5,5,5,5,5,5,5 pts.) **(Confident coin: III)**
When spun on edge 250 times a certain coin came up heads 140 times and tails 110. We can make the statement: 'if the coin is unbiased then the probability of getting a result at least this extreme is 7%.'

(a) *Let θ be the probability of coming up heads. Consider the null hypothesis that the coin is fair, $H_0 : \theta = 0.5$. Carefully explain how the 7% figure arises. What term describes this value in NHST? Does it correspond to a one-sided or two-sided test?*

Solution: H_0: $\theta = 0.5$

Test statistic: $x =$ number of heads in 250 spins.

Data: $x = 140$.

The probability of geting a result at least as extreme as seen is the $\boxed{p\text{-value.}}$

We'll show how to compute this value and decide if the test is one or two-sided at the same time.

A one-sided test would have alternative hypothesis H_A: $\theta > 0.5$. In this case, data at least as extreme means $x \geq 140$. Using R we compute the one-sided p-value:

$$p = P(x \geq 140 \,|\, H_0) = \texttt{1 - pbinom(139, 250, 0.5)} = 0.03321$$

This is not the 7% value we were looking for. So let's consider the two-sided test.

A two-sided test would have alternate hypothesis H_A: $\theta \neq 0.5$. Since the null distribution, binomial(250,0.5), is symmetric around 0, each tail in the rejection region will have probability $\alpha/2$ and the two-sided p-value is computed by doubling the smaller of the one-sided p-values. We computed the right tail p-value just above. This is the smaller of the two p-values so our two-sided p-value is $2 \times 0.03321 = 0.06642$. This rounds to 0.07, so the figure of 7% is the two-sided p-value.

Note: We could have used the normal approximation binomial$(250, 0.5) \approx$ N$(125, 250/4)$, and the z-statistic $z = \frac{x-125}{\sqrt{250/4}} \approx$ N$(0, 1)$. In this case, our p-values would be: one-sided: $p = P(z \geq \frac{15}{\sqrt{250/4}}) \approx 0.02889$ and two-sided: $p = P(|z| \geq \frac{15}{\sqrt{250/4}}) \approx 0.05778$.

(b) *Would you reject H_0 at a significance level of $\alpha = 0.1$? What about $\alpha = 0.05$? (For this problem assume the test has the same sidedness as the one used to get the 7% p-value in part (a).)*

Solution: As instructed, we use a two-sided rejection region as in part (a). The exact p-value was $p = 0.066$. Since $0.05 < p < 0.1$ we reject H_0 at significance $\alpha = 0.1$ and don't reject at $\alpha = 0.05$.
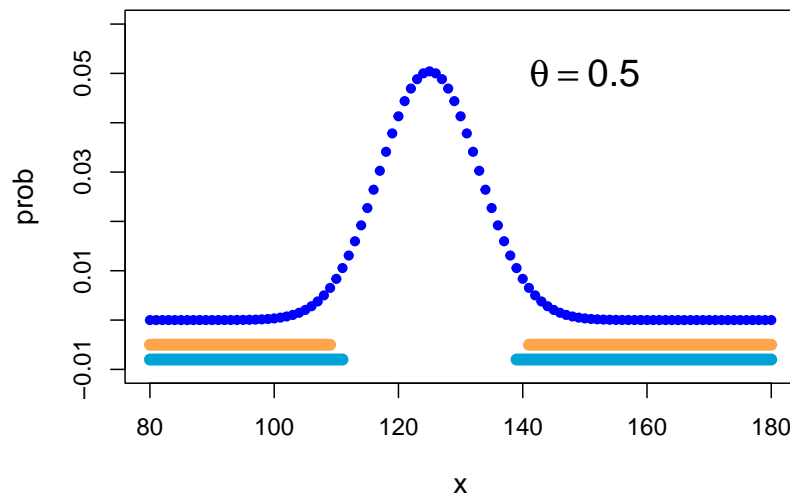
The figure shows the null distribution, the $\alpha = 0.1$ rejection region (blue-green) and the $\alpha = 0.05$ rejection region (orange). Notice that the data $x = 140$ is in the 0.1 rejection region but not the 0.05 rejection region.

(c) *How many heads would you need to have observed out of 250 spins to reject at a significance of $\alpha = 0.01$? (Again assume the same sidedness as the test used to get the 7% p-value in part (a).)*

Solution: Again, we use two-sided rejection region. The problem asks us to find the rejection region for $\alpha = 0.01$. We use R to find the endpoints for the rejection region (called critical values):

$$\text{critical\_value\_left = qbinom}(0.005, 250, 0.5) - 1 = 104$$
$$\text{critical\_value\_right = qbinom}(0.995, 250, 0.5) + 1 = 146$$

Note: we added or subtracted one to the value returned by `qbinom`. For a discrete distribution like the binomial there is not an exact critical value. So `qbinom(x, n, p)` returns the smallest integer with more than x probability in its left tail. Since the rejection region must have at most $\alpha/2$ in either tail we have to move the R answer by one towards the tail.

Conclusion: we reject for greater than or equal to 146 heads or less than or equal to 104 heads.

(d) *(i) Fix the significance level at $\alpha = 0.05$. Compute and compare the power of the test for values of the alternative hypothesis $\theta = 0.55$ and $\theta = 0.6$? (Again assume the same sidedness as in part (a). Here, H_A is composite, so the power is different for different values of the hypothesis.)*

(ii) Sketch the pmf of each hypothesis and use it to explain the change in power observed in part (i).

Solution: (i) Again we use a two-sided rejection region. For $\alpha = 0.05$ the rejection region is given by the critical values

$$\text{critical\_value\_left = qbinom}(0.025, 250, 0.5) - 1 = 109$$
$$\text{critical\_value\_right = qbinom}(0.975, 250, 0.5) + 1 = 141$$

power when $\theta = 0.55 = P(\text{reject} \mid \theta = 0.55)$

$$= P(x \leq 109 \text{ or } x \geq 141 \mid \theta = 0.55)$$

$$= \texttt{sum(dbinom(0:109, 250, 0.55))} + \texttt{sum(dbinom(141:250, 250,0.55))} = 0.35237$$

Likewise

power when $\theta = 0.6 = P(\text{reject} \mid \theta = 0.6)$

$$= P(x \leq 109 \text{ or } x \geq 141 \mid \theta = 0.6)$$

$$= \texttt{sum(dbinom(0:109, 250, 0.6))} + \texttt{sum(dbinom(141:250, 250,0.6))} = 0.88963$$

Note: We could have used `pbinom`. Doing the sums with `dbinom` was an easy way of avoiding off-by-one errors with `pbinom`.

(ii) The two plots below show the null distribution and the distribution of H_A for $\theta = 0.55$ and $\theta = 0.6$ The blue line below the graphs shows the rejection region. The greater power when $\theta = 0.6$ is explained by its greater separation from H_0. Most of the probability of $p(x \mid \theta = 0.6)$ is over the right side of the rejection region.
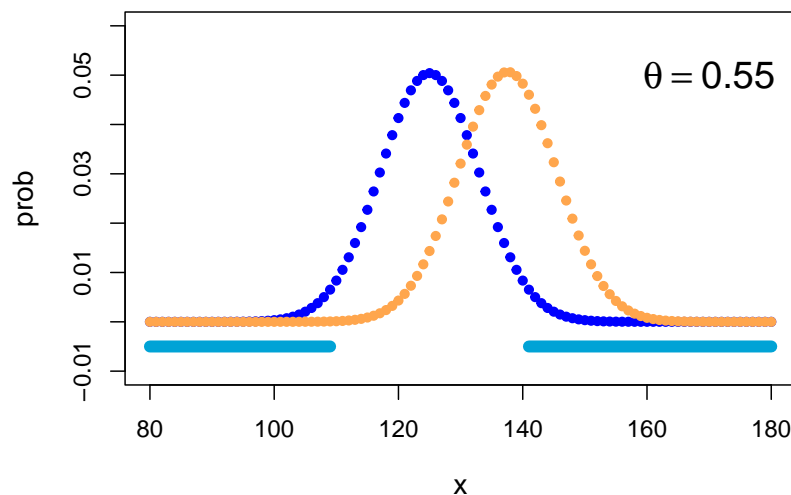
Distributions of H_0 and H_A with $\theta = 0.55$
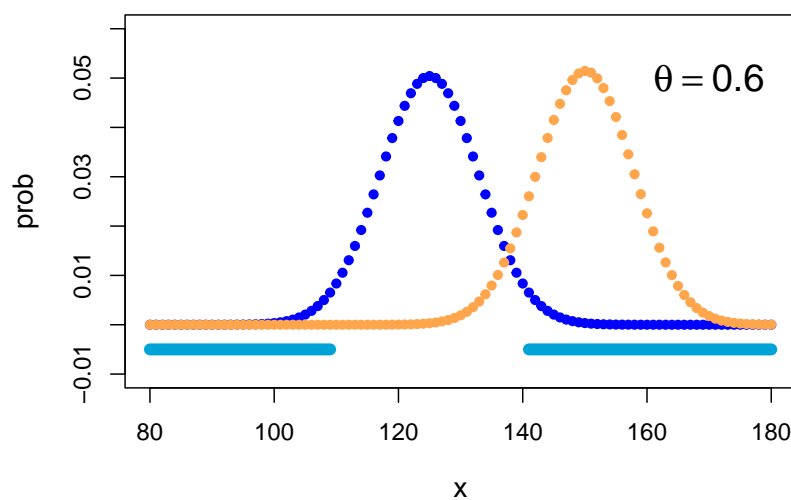
Distributions of H_0 and H_A with $\theta = 0.6$

(e) *(i) Again fix $\alpha = 0.05$. What is the* smallest *number of spins necessary for the test to have a power of 0.9 when $H_A : \theta = 0.55$? (Again, use the sidedness from part (a).)*

(ii) As in part (d), draw sketches and explain how they illustrate the change in power.

**Solution:** The answer is $n = 1055$ with H_A giving a power of 0.9003.

To get this we need to compute the power for various values of n. The steps for each n are:

1. Find the rejection region.
2. Compute the power.

Here is the R-code for one value of n. We ran this code in a loop to check through all values of n until we found the first with power $= 0.9$. The full code is in mit18_05_s22_ps8_sol.r, which is posted in the usual place.

```
theta = 0.55
n = 300;
# Find critical values for rejection region (based on theta=0.5)
critical_value_left = qbinom(0.025,n,0.5) - 1;
critical_value_right = qbinom(0.975,n,0.5) + 1;
rejection_region = c(0:critical_value_left, critical_value_right:n)
power = sum(dbinom(rejection_region, n, theta))
print(power)
```

See the two plots with part (d): power increases as n increases because the distributions become more separated.
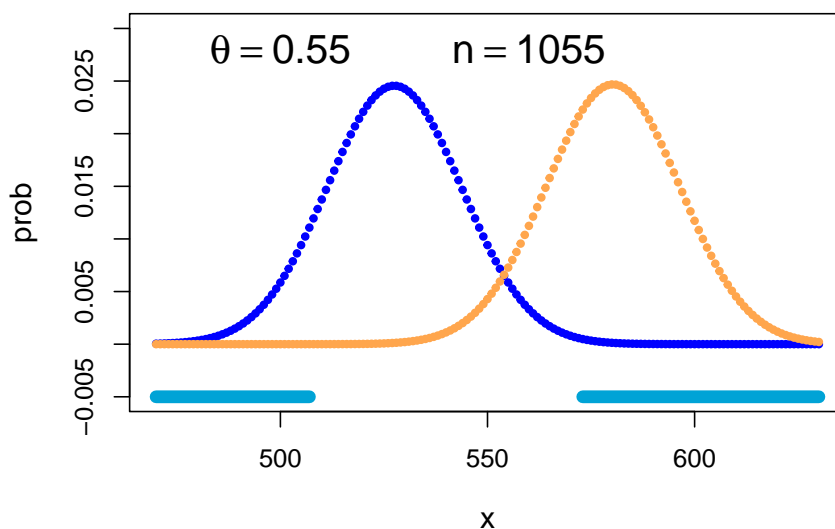
Plot for $n = 1055$ of the H_0 and $H_A : \theta = 0.55$ distributions. The blue lines show the rejection region.

An alternative approach approximating the exact answer with normal distributions is given at the end of these solutions.

(f) *Let $H_A : \theta = 0.55$. Suppose we have only two hypotheses H_0 and H_A, and a flat prior $P(H_0) = P(H_A) = 0.5$. What is the posterior probability of H_A given the data? (In this part H_A is different from in the previous parts; it consists of one specific value of θ.)*

Solution: We use the usual Bayesian update table.

Hypothesis	prior	likelihood	posterior
$\theta = 0.5$	$1/2$	$c_1(0.5)^{250}$	$c_2(0.5)^{250} = 0.14757$
$\theta = 0.55$	$1/2$	$c_1(0.55)^{140}(0.45)^{110}$	$c_2(0.55)^{140}(0.45)^{110} = 0.85243$

The normalizing factor $c_2 = \dfrac{1}{(0.5)^{250} + (0.55)^{140}(0.45)^{110}}$.

The posterior probability that $\theta = 0.55$ is 0.85.

(g) *Given the data, what probability would you personally place on the coin being biased toward heads? Why? There is no one right answer, we are simply interested in your thinking.*

Solution: This is a question about the probability of hypotheses. That's a Bayesian, not a frequentist, idea. I'd go with 90% probability the coin is biased towards heads. My Bayesian analysis is as follows.

If we used the Beta$(1, 1)$ (flat) prior on θ in [0,1]. Then the posterior for θ is a Beta$(141, 111)$ distribution. With this posterior

$$P(\theta > 0.5 \,|\, \text{data}) = 1 - \texttt{pbeta}(0.5, 141, 111) \approx 0.97.$$

This is 97%. Likewise, $P(\theta > 0.53 \,|\, \text{data}) = 83\%$.

Strictly speaking bias means $\theta > 0.5$. Leaving a small margin of error we can require $\theta > 0.53$. My answer is between these two probabilities.

Problem 2. (15: 10,5 pts.) **Polygraph analogy.**
In an experiment on the accuracy of polygraph tests, 140 people were instructed to tell the truth and 140 people were instructed to lie. Testers used a polygraph to guess whether or not each person was lying. By analogy, let's say H_0 corresponds to the testee telling the truth and H_A corresponds to the testee lying.

(a) *Describe the meaning of type I and type II errors in this context, and estimate their probabilities based on the table.*

	Testee is truthful	*Testee is lying*
Tester thinks testee is truthful	*131*	*15*
Tester thinks testee is lying	*9*	*125*

Solution: Type I error is rejecting the null-hypothesis when it is indeed true. This corresponds to thinking someone is lying when they are in fact being truthful. The experiment had $\frac{9}{140}$ type I errors. This is our estimate of the probability of a type I error.

Type II error is not rejecting the null-hypothesis when it is indeed false. This corresponds to thinking someone is telling the truth when they are in fact lying. Based on the data our estimate of the probability of a Type II error is $\frac{15}{140}$.

(b) *In NHST, what relationships exist between the terms significance level, power, type I error, and type II error?*

Solution: Significance $= P(\text{type I error}) = P(\text{reject } H_0 \,|\, H_0)$.
Power $= 1 - P(\text{type II error}) = P(\text{reject } H_0 \,|\, H_A)$.

Problem 3. (25: 10,10,5 pts.) **z-test**
Suppose three radar guns are set up along a stretch of road to catch people driving over the speed limit of 40 miles per hour. Each radar gun is known to have a normal measurement error modeled on $N(0, 5^2)$. For a passing car, let $\bar{x}$ be the average of the three readings. Our default assumption for a car is that it is not speeding.

(a) *Describe the above story in the context of NHST. Are the most natural null and alternative hypotheses simple or composite?*

Solution: Let μ be the actual speed of a given driver. We are given that

$$x_i \sim N(\mu, 5^2) \implies \bar{x} \sim N(\mu, 5^2/3).$$

The most natural hypotheses are:

H_0: the driver is not speeding, i.e. $\mu \le 40$.
H_A: the driver is speeding, i.e. $\mu > 40$.

Both are composite.

Note: we will work with H_0: $\mu = 40$, which is simple.

(b) *The police would like to set a threshold on $\bar{x}$ for issuing tickets so that no more than 4% of law abiding, non-speeders are mistakenly given tickets. You should assume this means that they set the threshold conservatively, so that no more 4% of drivers going exactly 40 mph get a ticket.*

(i) Use the NHST description in part (a) to help determine what threshold should they set.

(ii) Sketch a graph illustrating your reasoning in part (i).

(iii) What is the probability that a person getting a ticket was not speeding?

(iv) Suppose word gets out about the speed trap and no one speeds along it anymore. What percentage of tickets are given in error?

Solution: (i) Giving a ticket to a non-speeder is a type I error (rejecting H_0 when it is true). H_0 is composite, but we can do all our computations with the most extreme value $\mu = 40$ because the one-sided rejection region will have its largest significance level when $\mu = 40$.
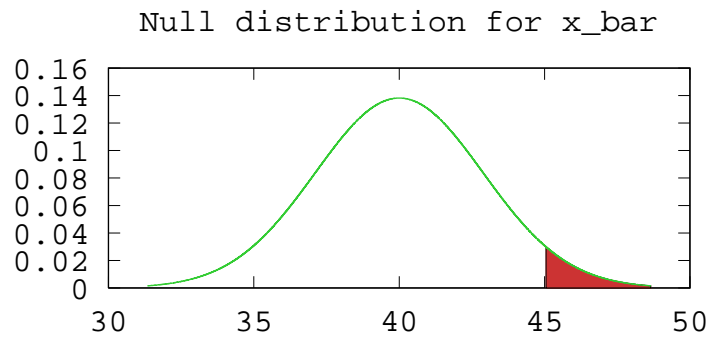
So the null distribution is $\bar{x} \sim N(40, 5^2/3)$. The critical value is

$$c_{0.04} = \texttt{qnorm(0.96,40,5/sqrt(3))} = 45.054$$

(Equivalently $c_{0.04} = 40 + z_{0.04}\frac{5}{\sqrt{3}} = 45.054$.)

That is, they should issue a ticket if the average of the three guns is more than 45.054.

(ii) Here is a plot of the null distribution $N(40, 5^2/3)$. The rejection region with probability of 0.04 is shown.

Null distribution for x_bar

(iii) (See (iv).) We don't know this probability. To find it out would require a prior probability that a random driver is speeding.

(iv) If no one is speeding then 100% of tickets are given in error.

(c) *What is the power of this test with the alternative hypothesis that the car is traveling at 45 miles per hour? How many cameras are needed to achieve a power of* 0.9 *with* $\alpha = 0.04$*?*

Solution: Power $= P(\text{rejection} \,|\, H_A)$. So to find the power we first must find the rejection region. For $n = 3$ this was done in part (b): rejection region $= [45.054, \infty)$. So

$$\text{power } = P(\text{rejection} \,|\, \mu = 45) = 1 - \texttt{pnorm(45.054, 45, 5/sqrt(3))} = \boxed{0.493}$$

With n cameras (guns) let's write $\bar{x}_n$ for the sample mean. The null distribution is

$$\bar{x}_n \sim \text{N}(40, 5^2/n)$$

The critical value, i.e. the left endpoint of the rejection region and power depends on n.

We can use R to compute: The code for computing the power when $n = 3$ is shown below. You only have to change the first line to compute power for different values of n.

```
n = 3
mu = 40
sigma = 5/sqrt(n)
alpha = 0.04
xcrit = qnorm(1-alpha, mu, sigma)
power = 1-pnorm(xcrit, 45, sigma)
```

We can now use R to compute power for increasing values of n and find the first value of n gives power more than 0.9. We find that $\boxed{n = 10}$.

Alternative, algebraic approach.

In order to do the computations algebraically we need to write everything in terms of standard normal values.

$$c_{0.04} = \texttt{qnorm(0.96, 40, 5/}\sqrt{n}\texttt{)} = 40 + z_{0.04}\frac{5}{\sqrt{n}}$$

where $z_{0.04}$ is the standard normal critical value

$$z_{0.04} = \texttt{qnorm(0.96, 0, 1)} = 1.751.$$

We want
$$\text{power} = P(\overline{x} \geq c_{0.04} \mid \mu = 45) = 0.9$$

Standardizing and doing some algebra we get
$$P\left(\frac{\overline{x} - 45}{5/\sqrt{n}} \geq \frac{c_{0.04} - 45}{5/\sqrt{n}}\right) = 0.9 \quad \Rightarrow \quad P\left(z \geq \frac{-5}{5/\sqrt{n}} + z_{0.04}\right) = 0.9$$

Thus $\dfrac{-5}{5/\sqrt{n}} + z_{0.04} = z_{0.9}$. We get
$$n = (z_{0.04} - z_{0.9})^2 = (1.7507 - (-1.2816))^2 = 9.1945.$$

Setting n to be the next biggest integer we get $n = 10$.

Problem 4. (10: 5,5 pts.) *One generates a number x from a uniform distribution on the interval $[0, \theta]$. One decides to test $H_0 : \theta = 2$ against $H_A : \theta \neq 2$ by rejecting H_0 if $x \leq 0.1$ or $x \geq 1.9$.*

(a) *Compute the probability of a type I error.*

Solution: $P(\text{type I}) = P(\text{reject } H_0 \mid H_0) = P(x \leq 0.1 \text{ or } x \geq 1.9 \mid \theta = 2) = \dfrac{0.2}{2} = 0.1.$

(b) *Compute the probability of a type II error if the true value of θ is 2.5.*

Solution: $P(\text{type II}) = P(\text{don't reject } H_0 \mid \theta = 2.5) = P(0.1 < x < 1.9 \mid \theta = 2.5) = \frac{1.8}{2.5} = 0.72.$

Problem 5. (Optional – for learning and fun. 0 pts.)
Give a careful, **accurate and precise** *explanation of the following XKCDs.*

(a) *https://xkcd.com/1132/*

Solution: We have hypotheses $H_0 = $ 'the sun is okay' and $H_A = $ 'the sun has gone nova'

The data is either 'yes' or 'no' from the detector.

We have the following likelihood table.

data	yes	no
$p(\text{data} \mid H_0)$	1/36	35/36
$p(\text{data} \mid H_A)$	35/36	1/36

The frequentist chooses the rejection region 'yes', which has significance 1/36. (Note: 1/36 is really $0.02777\ldots \approx 0.028$ not 0.027.)

The experimental data is 'yes', which is in the rejection region, so the frequentist correctly rejects H_0 in favor of H_A.

The Bayesian views this as silly, since, from their perspective, the posterior odds that the sun has gone nova are

$$\text{prior odds} \times \text{likelihood ratio} = \text{prior odds} \times \frac{p(\text{yes} \mid H_A)}{p(\text{yes} \mid H_0)} = \text{prior odds} \times \frac{35}{1}.$$

If we conservatively put the prior odds at $1/10^8$ then the posterior odds are still very small.

Besides, if the sun has gone nova, losing the bet is the least of the Bayesian's problem.

(b) *https://xkcd.com/882/*

Solution: The comic is pointing out the flaw of multiple testing or what's sometimes called data mining. (The bad type of data mining, there is also a good type.) A significance level of 0.05 means that in 20 experiments where H_0 is true we'd expect to reject it once. The scientists test 20 colors. So even if no jelly bean color causes cancer there is a high probability (well, 64%) that one of the tests will produce a test statistic in the rejection region.

The fix is to plan on doing n tests and set the significance level for any one test to α/n. Then, assuming H_0 is true for all the tests, the probability that at least one of them will reject is roughly $n \cdot \alpha/n = \alpha$. This is called the Bonferroni correction. (Actually, because of the possibility of multiple rejections the probabilitiy at least one will reject is less than or equal to α.)

Normal approximation approach to problem 1(e).

Here is the alternative approach to 1(e) mentioned at the end of the solution to that problem.

We use the normal approximation binomial$(n, \theta) \approx \mathrm{N}(n\theta, n\theta(1 - \theta))$. The normal approximation introduces some error, but we can compute n directly instead of needing to search through a sequence of possible values.

The algebra is a little involved and it requires that we keep all our definitions straight. I find it easiest to work with symbols and substitute values at the end.

We have the null and alternative values of θ are $\theta_0 = 0.5$ and $\theta_A = 0.55$.

The null distribution is $N(n\theta_0, n\theta_0(1 - \theta_0))$.

The alternative distribution is $N(n\theta_A, n\theta_A(1 - \theta_A))$.

The left critical value for the rejection region is $x_{0.975}$. This can be computed in terms of the standard normal critical value $z_{0.975}$. Similarly for the right critical value $x_{0.025}$.

$$x_{0.025} = n\theta_0 + \sqrt{n\theta_0(1 - \theta_0)}\, z_{0.025}$$
$$x_{0.975} = n\theta_0 + \sqrt{n\theta_0(1 - \theta_0)}\, z_{0.975}.$$

Power is the probability of our test statistic x is in the rejection region assuming H_A. That is,
$$\text{power} = P(x \le x_{0.975} \mid \theta = \theta_A) + P(x > x_{0.025} \mid \theta = \theta_A)$$

For n large the left hand probability is essentially 0. So we can ignore it. Thus, to get power $= 0.9$ we need
$$P(x > x_{0.025} \mid \theta = \theta_A) = 0.9. \tag{1}$$

Let $c_{0.9}$ be the critical value for the *alternative distribution.* By the usual linearity of normal distributions we have
$$c_{0.9} = n\theta_A + \sqrt{n\theta_A(1 - \theta_A)}\, z_{0.9},$$

where $z_{0.9}$ is the standard normal critical point.

Equation 1 can be restated as $x_{0.025} = c_{0.9}$. Now, rewriting both of these values in terms of n, θ_0 and θ_A, we get
$$n\theta_A + \sqrt{n\theta_A(1 - \theta_A)}\, z_{0.9} = n\theta_0 + \sqrt{n\theta_0(1 - \theta_0)}\, z_{0.025}$$

This is easily solved for n:

$$\sqrt{n} = \frac{\sqrt{\theta_0(1-\theta_0)}\, z_{0.025} - \sqrt{\theta_A(1-\theta_A)}\, z_{0.9}}{\theta_A - \theta_0}.$$

We used R to compute $z_{0.025} = 1.96$ and $z_{0.9} = -1.28$. Putting this in the formula above, with $\theta_0 = 0.5$ and $\theta_A = 0.55$ we get $\boxed{n = 1046.582 \approx 1047.}$ This is not too far off from the exact answer of 1055 found in problem 1(e).

MIT OpenCourseWare

https://ocw.mit.edu

18.05 Introduction to Probability and Statistics

Spring 2022