Power

T.Scofield

Zener cards and ESP

Suppose we are testing whether a person has ESP, measuring this by the success rate at identifying the correct images on 100 Zener cards.

Following steps described in the text, we first set our hypotheses.

Step 1

Given there are 5 different images, the "you don't have ESP until you prove otherwise" hypothesis test would naturally set

$$\mathbf{H}_0 : \pi = 0.2 \qquad \text{against} \qquad \mathbf{H}_a : \pi > 0.2$$

Step 2

Now we must gather data. We recognize that any person undergoing this task has, potentially, an infinitude of "intuitive guesses" he can make about the images being seen on these cards. When a researcher finds a single hour to sit down with the participant and record a set of 100 actual "guesses", the result will be a **sample** from that larger population. Results under various conditions and different days are sure to vary, and all one ask is that the sample be as free from bias as possible: the participant gets a good night sleep in advance, the researcher puts no "tone" in her voice nor exhibits visual cues, etc. If testing will happen over numerous days, one tries to make conditions as similar as possible on each test day.

Now let us suppose that our participant correctly identifies 27 of the 100 Zener cards. There are several test statistics one might consider, such as the **count** of success, 27, or the **proportion** of successes, 0.27. In anticipation of Step 3, we will use the count $X = 27$. This is because $X \sim \text{Binom}(100, \pi)$; it has a distribution that, when we supply the value of π from the null hypothesis, is one whose distributional function values (those of the pmf and cdf) are known.

Step 3

We use the binomial distribution with parameters $n = 100$, $\pi = 0.2$ to calculate the likelihood, under the null hypothesis, of obtaining a value at least as extreme as $X = 27$. Given our alternative hypothesis is one-sided, anticipating $\pi > 0.2$, we are talking about

$$\Pr(X \geq 27).$$

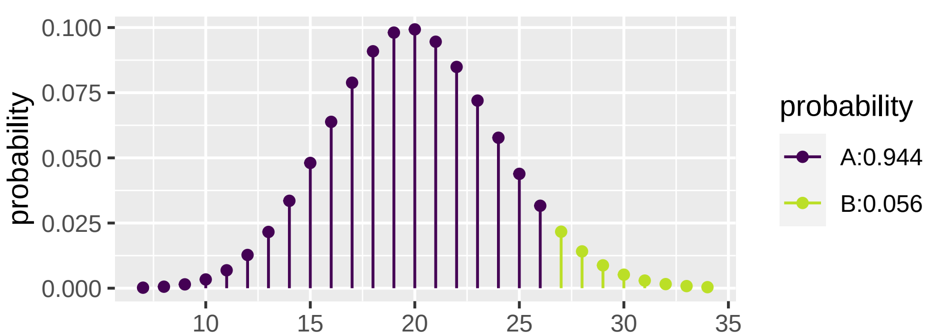
This probability is called the P-value. It can also be computed directly using `pbinom()`:

```
1 - pbinom(26, 100, 0.2)
```

```
[1] 0.05583272
```

It may also be pictured as the sum of values for the pmf which are shaded green.

```
xpbinom(26, 100, 0.2)
```




```
[1] 0.9441673
```

Results, counts X, at least this large occur only about 5.58-percent of the time when someone is purely guessing (i.e., has no intrinsic advantage, which includes, but may not be limited to, blindly guessing).

Alternate Step 3

One may argue that doing poorly—inordinately so—is also evidence of ESP. (Can you make such an argument?) This would correspond to our employing a 2-sided alternative hypothesis

$$\mathbf{H}_a : \pi \neq 0.5.$$

While, in many cases, this roughly doubles our P-value from what it was with a 1-sided alternative, that is not always the case. A more careful assessment takes the P-value as the sum of probabilities for all choices of X that are not more likely than $X = 27$.

```r
probAt27 = dbinom(27, 100, 0.2)
allPMFvals = dbinom(0:100, 100, 0.2)
Pval = sum(allPMFvals[allPMFvals <= probAt27]); Pval
```

```
[1] 0.102745
```

Step 4

If a conclusion is required, we will generally have some threshold α, known as a **significance level**, on which to base our decision. Taking $\alpha = 0.05$, we would **fail to reject the null hypothesis**, since $P > \alpha$.

Using `binom.test()`

R has a command that hides this work from us, `binom.test()`. To get the same P-value as our test with 1-sided alternative $\mathbf{H}_a : \pi > 0.2$, we do

```r
binom.test(27,100,.2, alternative="greater")
```

```
data:  27 out of 100
number of successes = 27, number of trials = 100, p-value = 0.05583
alternative hypothesis: true probability of success is greater than 0.2
95 percent confidence interval:
 0.1979249 1.0000000
sample estimates:
probability of success
                  0.27
```

Removing the `alternative=` switch will result in the P-value we obtained using a 2-sided alternative.

Rejection and non-rejection regions

When, prior to the other steps of hypothesis testing, we select the significance level α, we are dictating how we will respond in Step 4, once we know the P-value. We might take this one level beyond, and pre-determine how we will respond once we see X.

In a test with a 2-sided alternative and $\alpha = 0.05$, we can think in terms of neither tail probability exceeding 0.025. The `qbinom()` can tell us the X that best corresponds to the 0.025-quantile.

```r
qbinom(0.025, 100, .2)
```

```
[1] 12
```

```r
pbinom(12, 100, .2)
```

```
[1] 0.02532875
```

We see that Prob($X \leq 12$) is very close to 0.025. On the other end,

```r
qbinom(0.975, 100, .2)
```

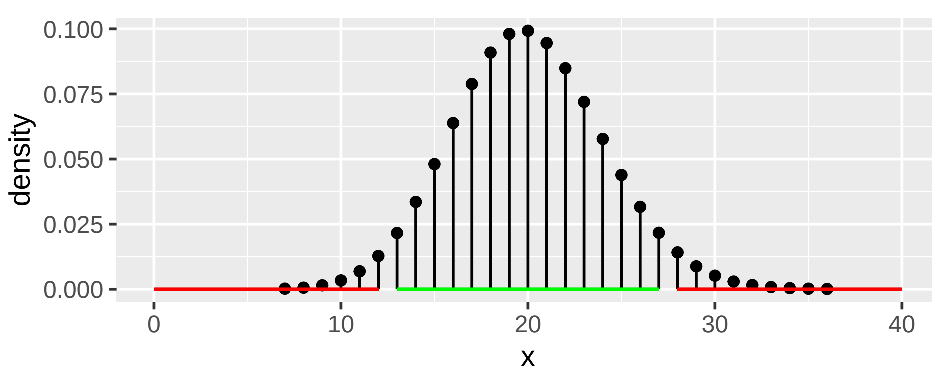
```
[1] 28
```

```r
pbinom(28, 100, .2)
```

```
[1] 0.9799798
```

Thus, without knowing any data just yet, we can determined that rejection of the null hypothesis will occur whenever the count $X \leq 12$ or $X \geq 28$. We can call this the **rejection region**; I have colored it red in the picture. These boundaries for the rejection region are called **critical values**. When $13 \leq X \leq 27$, we will fail to reject the null; that part is colored green.

```r
plot.1 <- gf_dist("binom", params=c(100,0.2)) |>
  gf_segment(c(0)+c(0)~c(13)+c(27), color="green") |>
  gf_segment(c(0)+c(0)~c(0)+c(12), color="red") |>
  gf_segment(c(0)+c(0)~c(28)+c(40), color="red")
plot.1
```



These regions have been drawn taking into account the distributional family (binomial), the parameters, and α.

Power

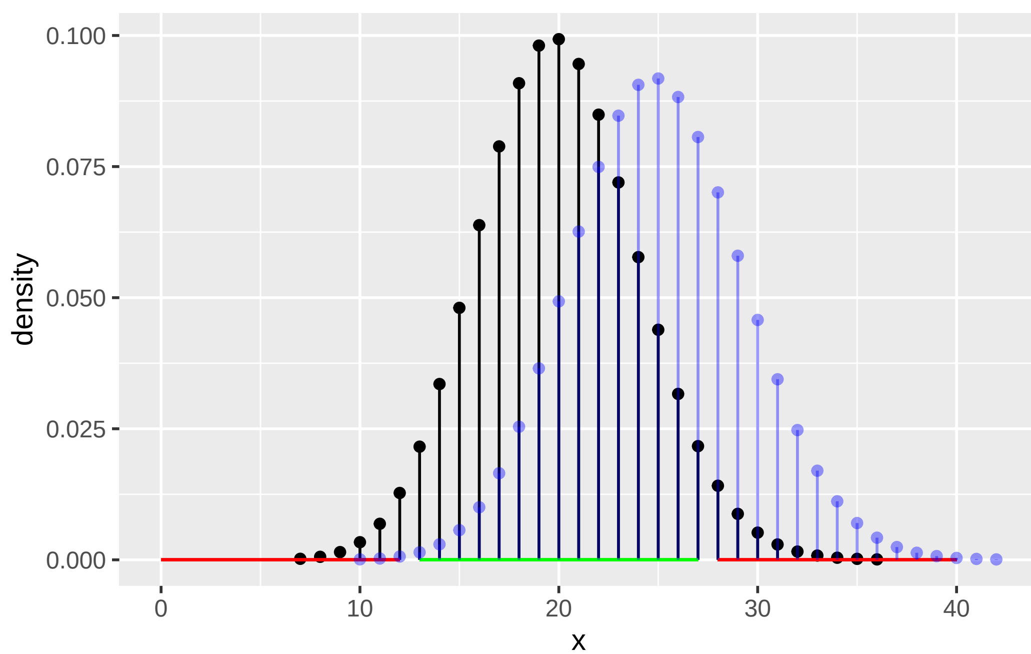
Power may be interpreted as the likelihood that a false null hypothesis will be *judged* as false. The book correctly indicates this likelihood depends upon how far the real value of π is from the one we propose in the null hypothesis. Here are two illustrations.

Case: π is really equal to 0.25

This would represent a person who is slightly better at getting the image on the Zener card correct than one would expect from blind guessing. Let us continue to take $n = 100$ and $\alpha = 0.05$ so that, under the null hypothesis, critical valuse are still $X = 12$ and $X = 28$.

The graph below depicts the null distribution, and then overlays (in blue) the distribution when $X \sim \text{Binom}(100, 0.25)$, valid for a person will this elevated ability.

```
plot.1 |> gf_dist("binom", params=c(100, 0.25), color="blue", alpha = 0.4)
```



To assess power, we compute $\Pr(X \leq 12 \text{ or } X \geq 28 \mid X \sim \text{Binom}(100, 0.25))$, a conditional probability evaluated as

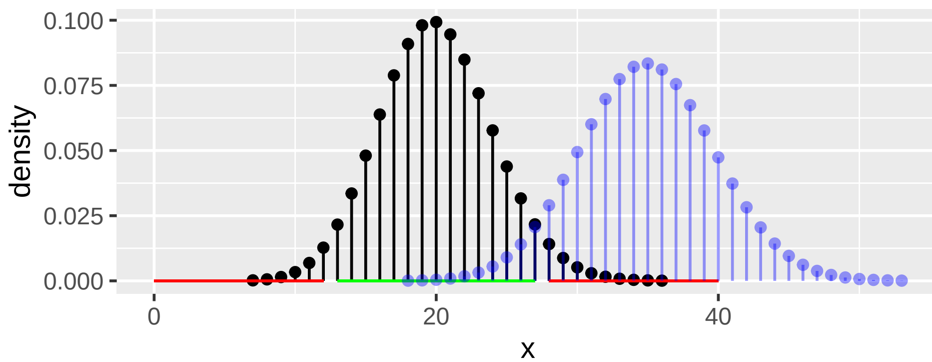
```
sum(dbinom(c(0:12, 28:100), 100, 0.25))
```

```
[1] 0.2786469
```

Case: π is really equal to 0.35

You might instinctively, under this scenario, power is increased over that when the participant has a 25% success rate—that is, our participant with success rate further from 0.2 is more likely to produce a value of X in the rejection region. A picture and calculation similar to the one above but adapted to this situation follow.

```
plot.1 |> gf_dist("binom", params=c(100, 0.35), color="blue", alpha = 0.4)
```



Power, now with a true success rate of 35%, is $\Pr(X \leq 12 \text{ or } X \geq 28 \mid X \sim \text{Binom}(100, 0.35))$:

```
sum(dbinom(c(0:12, 28:100), 100, 0.35))
```

```
[1] 0.9441925
```

Plotting power as a function of the actual success rate, holding $\alpha = 0.05$

```
power.fn <- makeFun(pbinom(12,100,x) + 1 - pbinom(27,100,x) ~ x)    # probability of rejection
gf_fun(power.fn(x) ~ x, xlim=c(0,1))
```