# P-Value Examples

John Hawkins

February 24, 2014

The activity for last Friday was to revisit the examples of experiments where we might wish to use p-values. Below are the four example questions, along with the exact question put on the slide Friday to describe it. These questions were intentionally left vague to let you decide what your model assumptions are and then to answer the question according to your assumptions, as one would do in real life. Hence, for each of the questions, I will have to specify what modeling assumptions I am making.

## Pride of Lions

**Question:** Does one pride of lions have a different sex ratio than others?
**Solution:** First, we have to recognize that what the problem appears to really be asking is not whether the sex ratio is different, but whether the probability that an i.i.d. born child is, say, male is different. That is, we are assuming that each child born is a Bernoulli trial in gender, and hence we expect a binomial distribution of males for some probability $m$ of getting a male child. I will solve this problem with two different sets of assumptions, depending on how we interpret our knowledge of the "other" prides.

We define the variables, $m_{pride}$ and $m_{others}$, the Bernoulli trial probability of getting a male child in our pride or other prides, respectively. We correspondingly label the number of males and total number of lions in our pride and other prides $n_{pride}$, $N_{pride}$, $n_{others}$, and $N_{others}$ respectively.

1. First, let's assume that our knowledge of the "other" prides of lions is perfect, so we know the value of $m_{others}$ *exactly*. This assumption is equivalent to assuming that we have observed infinitely many children born to the "other" prides. And the method of evaluation is equivalent to evaluating whether a coin is fair. The null hypothesis is:

$$H_0 : \ m_{pride} = m_{others}.$$

Let's make up some example data. Let's say 40% of the lions in the other prides are male, and 6 of 30 lions in our pride are male. Figure 1 shows the distribution of $n_{pride}$ given $N_{pride} = 30$ lions under the null hypothesis. We want to know whether the rate is different, either higher or lower. That is, we consider both differences extreme, so we will perform a two-tailed test. We interpret "as extreme or more extreme" as the distance from the mean:

$$\mathbb{E}[n_{pride}] = N_{pride} m_{others} = (30)(0.4) = 12.$$

Hence, our formula for the p-value given $n_{pride}$ is given by

$$p\text{-}value = \sum_{\{i \in \{0,1,\ldots,30\} : |12-i| \geq |12-n_{pride}|\}} \binom{30}{i} (m_{others})^i \ (1 - m_{others})^{(30-i)}$$

The critical region for a significance level of 0.05 is highlighted in Figure 1, and the observed value of $n_{pride} = 6$ is shown as a dashed line. The observed value has a p-value of $0.038 < 0.05$,

so it is in the critical region. We therefore reject the null hypothesis and declare that our pride of lions does indeed have a different probability of male children.

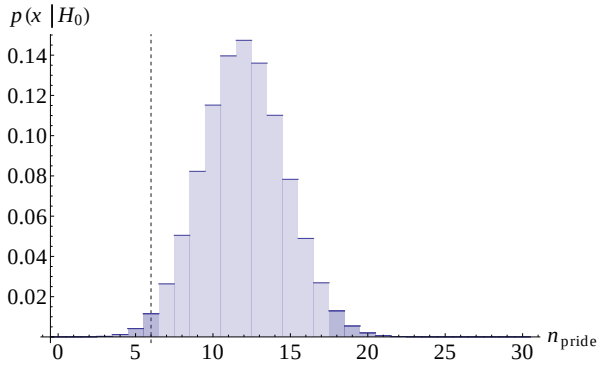


Figure 1: Distribution of $n_{pride}$ under the null hypothesis, assuming perfect knowledge of $m_{others}$. Critical region highlighted.

2. Now, let's move on to the other set of assumptions we might make. Before starting, I want to say that we would not have expected this rigorous of an argument to be made during the time constraints of the class period. However, you should be able to understand it. Furthermore, I think there is a useful consideration of possible test statistics here that will be worth your while.

   So, this time let's not assume we have perfect knowledge of $m_{others}$. We still have the same null hypothesis:

   $$H_0 : \ m_{pride} = m_{others}.$$

   However, we must work a little harder to evaluate it. Let's continue with the above example, but now instead of using $m_{others} = 0.4$ exactly, we use as our data that $n_{others} = 40$ and $N_{others} = 100$.

   Now, we wish to evaluate whether $m_{pride} = m_{others}$, but we do not have a value of $m_{others}$ to test against. This is why these assumptions are a bit trickier. What to do? There are a couple answers to this question. Later in the course we will mention something called posterior-predictive p-values. But for now, a slightly more conservative and slightly simpler solution: We happen to be able to find the value of $m_{others}$ which will make the p-value the largest. That value is precisely the MAP estimate of $m_{others}$ assuming $m_{others} = m_{pride}$. If this largest of all p-values is still too small, we can conservatively reject the null hypothesis. That is, if we assume the value of $m_{others}$ that makes the observed data most likely (the MAP estimate under the null hypothesis), and the observed data is still extremely unlikely (small p-value), then we can conservatively reject the null hypothesis. Think about this for a moment to make sure you understand the argument. Why do we not need multiple-hypothesis correction? (Hint: What does independence have to do with it?)

   Figure 2 shows the posterior of $m_{others}$ assuming $m_{others} = m_{pride}$. The MAP estimate is $23/65 \approx 0.35$, shown as a dashed line. So now, we need to choose a test statistic. We can choose anything we want. We might think we want to use

   $$\frac{n_{pride}/N_{pride}}{n_{others}/N_{others}}$$

   because then our MAP estimate is simple:

   $$\frac{N_{pride}m_{others}/N_{pride}}{N_{others}m_{others}/N_{others}} = 1.$$

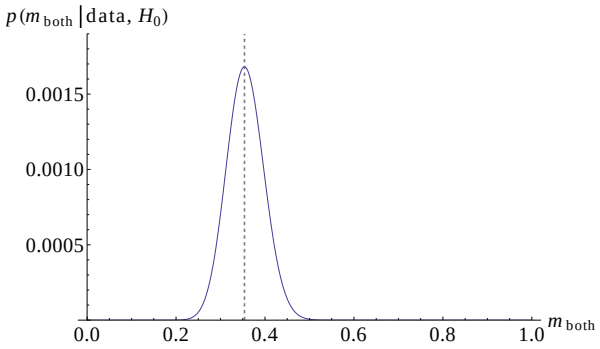


Figure 2: Posterior distribution of $m_{others}$ assuming $m_{others} = m_{pride}$. The MAP estimate is $m_{others} = 23/65$.

This could be done, but the distribution turns out to be messy due to discreteness effects. Specifically, after dividing there are many possible non-integer values with small probabilities scattered around everywhere. So here, we will use the following test statistic:

$$X = n_{others} - n_{pride}$$

which, under the null hypothesis, has a MAP estimate of

$$N_{others}m_{others} - N_{pride}m_{others} = (N_{others} - N_{pride})m_{others} = (100 - 30)\frac{23}{65} \approx 24.8.$$

Now, since we chose the difference $n_{others} - n_{pride}$ as the test statistic, the only possible values are still integer values, giving an easily tractable pdf:

$$pdf_X(x) = \sum_{\{j-i=x\}} \binom{30}{i} m^i (1-m)^{(30-i)} \binom{100}{j} m^j (1-m)^{(100-j)}$$

$$= \sum_{i=\max\{0,-x\}}^{\min\{30,100-x\}} \binom{30}{i} \binom{100}{x+i} m^{2i+x} (1-m)^{130-(2i+x)}$$

where the max and min starting and stopping values of $i$ ensure that $0 \leq i \leq 30$ and $0 \leq j \leq 100$.

We now can consider our observed data: $n_{others} - n_{pride} = 40 - 6 = 34$. Figure 3 shows the distribution of our test statistic under the null hypothesis over all possible values from -30 to 100. We see that our data is NOT in the critical region. Indeed, we calculate a p-value of $0.10 > 0.05$. Therefore, we cannot reject the null hypothesis.

One a small side note, if you look carefully, you can see a dashed line over the plot in Figure 3. That dashed line is the standard gaussian approximation for this distribution. We can see that it is indeed a very good fit. In fact, for a two-tailed test, one could argue that the gaussian is a better distribution to use since it is smooth, while the exact distribution has significant discreteness effects. If I remember correctly, Bill will discuss that a bit in one of the lectures concerning contingency tables.

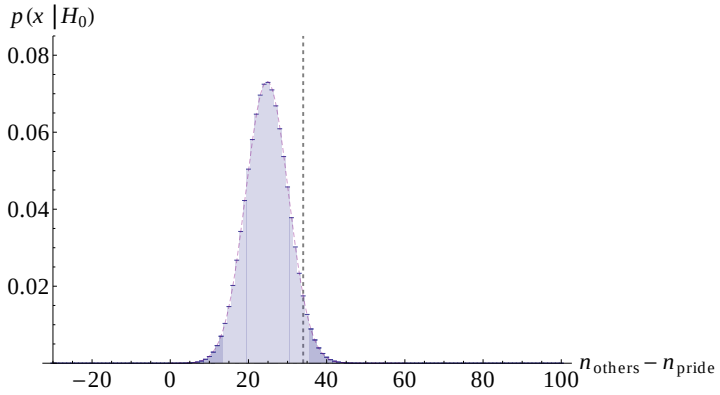


Figure 3: Distribution of $n_{others} - n_{pride}$ under the null hypothesis. Critical region highlighted. Observed value at dashed line.

## Lottery

**Question:** Are the odds given for winning a lottery accurate?
**Solution:** I don't remember exactly the proposed experiment for this problem, but I seem to remember it was very similar to the ORF problem. So I believe it was something like this:

> The lottery company states that 1/3 of all tickets are winners. We decide to buy tickets until we win. We do not win until the 10th ticket. Is the lottery company over-estimating the proportion of winning tickets?

Thus, we expect the number of tickets bought to be a random variable with a geometric distribution, $p = 1/3$:

$$pdf_X(x) = (1-p)^{(x-1)}p.$$

We only consider data extreme on the right-tail, so we perform a one-tailed test. The formula for the p-value is thus

$$p\text{-}value(x) = \sum_{i=x}^{\infty}(1-p)^{(i-1)}p = (1-p)^x = \left(\frac{2}{3}\right)^x.$$

For our data of $x = 10$, we therefore have a p-value of $(2/3)^10 = 0.017 < 0.05$, so we can reject the null hypothesis at the 0.05 significance level.

## GPS Coordinates

**Question:** Are the GPS coordinates given by a device accurate?
**Solution:** The proposed data collection scheme was as follows: Place the GPS device in a known location, then collect data points from the GPS and compare with the known true location.

So, we have a basic setup. We must now answer what exactly it is we mean by "accurate". I believe after questioning, the team said there was a specified variance, and we wish to test if the data collected is at or better than that variance. Let's formalize that:

> A GPS manufacturer claims their device has a given spherically symmetric variance $\sigma^2$. We place our device at a precisely known location and collect data points. Is the variance larger than advertised?

Placing the origin of our coordinate system at the known true location of the device, we thus have the following null hypothesis:

$$H_0: \text{ the data is a spherically symmetric gaussian with } \mu = 0, \text{ and given } \sigma^2.$$

Now, we have to remember a couple things about multivariate gaussians which we haven't learned in class yet. Makes it trickier to remember when we haven't seen it yet, but we're pretty awesome, so we've got this. (We will learn about the multivariate normal distribution next Wednesday, anyway, so now's as good a time as any.)

A random variable $X$ with a multivariate gaussian distribution and mean zero has a pdf given by

$$pdf_X(\boldsymbol{x}) = \frac{1}{(2\pi)^{(d/2)}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}\right),$$

Since we are considering only a spherically symmetric gaussian, this simplifies significantly. Specifically, we assume that the covariance matrix $\Sigma$ is given by

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix},$$

so our pdf is given by

$$pdf_X(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi\sigma^2)^3}} \exp\left(-\frac{r^2}{2\sigma^2}\right),$$

where $r^2 = \boldsymbol{x}^T\boldsymbol{x}$ is the square of the distance from the origin.

What shall we choose as our test statistic? Most people in the class decided to use the average distance from the known value to the observed value as the test statistic. That is, our test statistic is the random variable $Y$ given by

$$Y = \frac{1}{n}\sum_{i=1}^{n} r_i.$$

This is a fine test statistic, and certainly the most immediately intuitive thing to use. However, it is not commonly used because there is standard, easier statistic to work with, but let's think about it for a moment. How do you expect the average distance to be distributed under the null hypothesis of a gaussian distribution of points? Several people immediately thought it would be gaussian. This is clearly not the case, though, since distance cannot be negative. In one dimension the correct distribution for a single data point would simply be a folded gaussian (which is exactly what you think it is), but in higher dimensions it is a bit more complicated. If you integrate the pdf above (with correct spherical coordinate correction factors), we get that the pdf of the distance of a single spherically symmetric gaussian draw is given by

$$pdf(\boldsymbol{x}) = \sqrt{\frac{2}{\pi\sigma^2}}\left(\frac{r}{\sigma}\right)^2 e^{-\frac{1}{2}\left(\frac{r}{\sigma}\right)^2}.$$

This turns out to be the pdf of the $\chi$ distribution. Unsurprisingly, the $\chi$ statistic is given simply by

$$\chi = \frac{r}{\sigma}.$$

But what is the distribution of $Y$, the average of many $\chi$-like statistics? I do not know the answer to that, nor do I need to because there is a more straight-forward statistic.

And what is this standard test statistic? It is simply the $\chi^2$ statistic! The reason this is much easier to use is because the $\chi$ distribution implicity has a square root: $r = \sqrt{\sum_i x_i^2}$. Summing these up is awkward to work with. However, the $\chi^2$ has no such problem. And in fact, we can sum up the

$\chi^2$ statistic of many data points, and it is still $\chi^2$-distributed (with a different number of degrees of freedom)! This makes it very convenient to use. For our data, since $\mu = 0$, the $\chi^2$ statistic is given by

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{r_i}{\sigma}\right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} \sum_{j=1}^{3} x_{ij}^2.$$

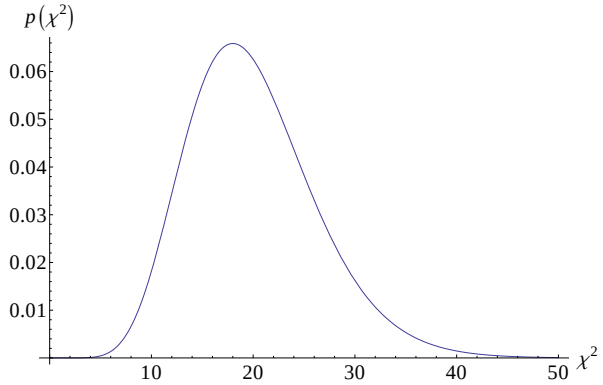


Figure 4: A $\chi^2$ distribution with degrees of freedom $\nu = 20$.

We have seen the distribution of the $\chi^2$ statistic of normal data before. This distribution was understandably–but perhaps confusingly–called the $\chi^2$ distribution. We will prove in a lecture next week that for a gaussian distribution, the $\chi^2$ statistic has a $\chi^2$ distribution:

$$pdf_{\chi^2}(x) = \frac{1}{2^{\nu/2}\Gamma(\frac{\nu}{2})} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}},$$

where $\nu$ represents the degrees of freedom of the $\chi^2$ statistic. In a couple weeks we will revisit the idea of degrees of freedom. For now, we leave it as is, noting that we will consider data to be extreme only at the right tail of the distribution, so we will have a one-tailed test. If the area under the $\chi^2$ distribution curve to the right of the observed value is less than 0.05, we can reject the null hypothesis at a significance level of 0.05. Of course, GPS coordinates fall into the world of physics, so 0.05 is not gonna cut it. We would need something more along the lines of so-called $3\sigma$, which is about 0.0026. Figure 4 shows an example $\chi^2$ distribution with $\nu = 20$. Take a look and convince yourself that's what you would expect the $\chi^2$ statistic above to look like with normally-distributed data.

## Traffic lights

**Question:** Are traffic lights lasting as long as their quoted time to failure?
**Solution:** Lastly, we consider the traffic lights. The exact problem statement as I remember it:

> A manufacturer claims that the time to failure of their traffic lights has an exponential distribution with mean of 100 days. We have observed that with our traffic lights the average time to failure is 150 days. Is the manufacturer underestimating the mean time?

This is a nice, tidy example, but just complicated enough to be interesting. Figure 5 shows the claimed exponential time to failure. This is not what our data will look like, however. Our test statistic is the *average* time to failure, and, obviously, we expect the average value to approach a

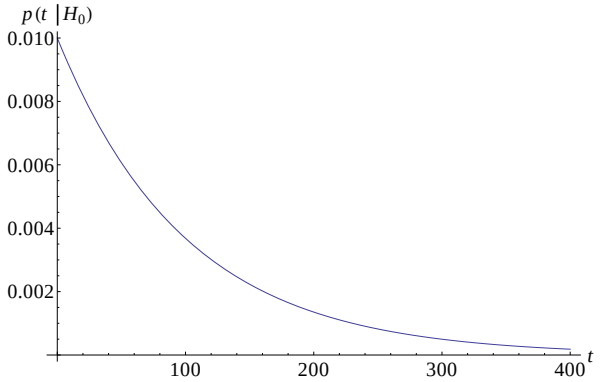


Figure 5: An exponential distribution with $\lambda = 1/100$, i.e. $\mu = 100$.

normal distribution as we increase the amount of data. Let $X_i$ be the random variable for the time to failure of the $i$-th traffic light. Our test statistic, $Y$, the average time to failure, is given by

$$Y = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Now, our null hypothesis is that the $X_i$ are i.i.d. exponential with $\lambda = 1/100$, i.e. $\mu = 100$. Now, while we could use the CLT for large $n$, in this case we happen to know the exact distribution. See the class activity for Friday February 7 (lost to a snow day) for an exercise showing that the sum of i.i.d. exponentially distributed random variables is gamma distributed. (It's pretty cool. Check it out if you have time.) Hence,

$$pdf_Y(x) = n\lambda^n \frac{(nx)^{n-1}}{(n-1)!} e^{-\lambda(nx)}$$

Figure 6 shows the expected distribution for $n = 30$. We see that if our observed average is 150, there is very little area under the curve. The corresponding p-value is 0.007, so we can reject the null hypothesis at the 0.01 significance level.

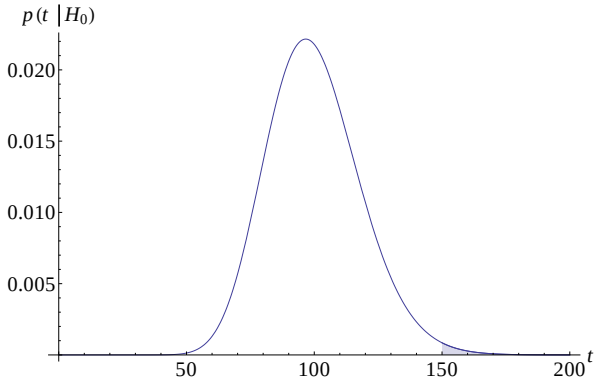


Figure 6: The expected distribution of $Y$ given 30 data points.