Review
18.05 Spring 2018

Here are some board problems to finish the semester. . .

Board question: mileage

Each time it is turned off, your car reports how far you have travelled and how much gasoline you used. Here are the reports (distance,gas) from last week:

$$(0.8, 0.06),\ (1.1, 0.08),\ (0.8, 0.05), (36.2, 0.74),\ (1.1, 0.07)$$

Do a linear regression to estimate mileage.
Hint: R says that the line best fitting these points is

$$\text{gallons} = 0.019 * \text{distance} + 0.047$$

Better Hint: R says that the line with intercept zero best fitting is

$$\text{gallons} = 0.021 * \text{distance}$$

Board question: make it fit

Bivariate data:

$$(1,3),\ (2,1),\ (4,4)$$

1. Do linear regression to find the best fitting parabola.

2. Do linear regression to find the best fitting cubic.

Solutions

2. Model $\hat{y}_i = ax_i^2 + bx_i + c$.

Total squared error:

$$\begin{aligned} T &= \sum (y_i - \hat{y}_i)^2 \\ &= \sum (y_i - ax_i^2 - bx_i - c)^2 \\ &= (3 - a - b - c)^2 + (1 - 4a - 2b - c)^2 + (4 - 16a - 4b - c)^2 \end{aligned}$$

Taking the partial derivatives and setting to 0 gives the following system of simultaneous linear equations:

$$\begin{array}{llll} 273a & +73b & +21c & = 71 \\ 73a & +21b & +7c & = 21 \\ 21a & +7b & +3c & = 8 \end{array} \quad \Rightarrow \quad a = 7/6,\ b = -11/2,\ c = 22/3.$$

The least squares best fitting parabola is $y = 7x^2/6 - 11x/2 + 22/3$. All three points lie on this parabola; for example, $3 = 7/6 - 11/2 + 22/3$.

Solutions continued

3. Model $\hat{y}_i = ax_i^3 + bx_i^2 + cx_i + d$.

Total squared error:

$$\begin{aligned} T &= \sum (y_i - \hat{y}_i)^2 \\ &= \sum (y_i - ax_i^3 - bx_i^2 - cx_i - d)^2 \\ &= (3 - a - b - c - d)^2 + (1 - 8a - 4b - 2c - d)^2 \\ &\quad + (4 - 64a - 16b - 4c - d)^2. \end{aligned}$$

Setting partial derivatives equal to zero leads to a system of four equations for the four unknowns a, b, c, d, but the equations have infinitely many solutions. With only 3 data points, using a cubic model is certainly overfitting our data.

Board Question

(a) Count the number of ways to get exactly 2 heads in 10 flips of a coin.

(b) For a fair coin, what is the probability of exactly 2 heads in 10 flips?

(c) If you flip a coin 10 times and get 2 heads, should you reject the null hypothesis that the coin is fair with 95% confidence?

Start of solution

answer: (a) We have to 'choose' 2 out of 10 flips for heads: $\boxed{\binom{10}{2}.}$ One way to compute is to pick a first flip to be heads (10 choices); then pick a second flip to be heads (nine choices), for 90 choices altogether. But picking 2 and 7 is the same as picking 7 and 2: we've overcounted by a factor of 2!. So $\binom{10}{2} = (10 \cdot 9)/(2 \cdot 1) = 45$.

(b) There are 2^{10} possible outcomes from 10 flips (this is the rule of product). For a fair coin each outcome is equally probable so the probability of exactly 2 heads is

$$\frac{\binom{10}{2}}{2^{10}} = \frac{45}{1024} = 0.044,$$

or a bit less than 5%.

Solution continued

(c) What's making you want to say the coin is unfair is getting an unexpectedly extreme number of heads. The *p*-value for an experiment with ten flips is the probability of getting such an extreme result under the null hypothesis. A reasonable meaning of extreme as two heads is

zero, one, or two heads (probability $(1 + 10 + 45)/1024 = 0.0547$)

together with

ten, nine, or eight heads (probability $(1 + 10 + 45)/1024 = 0.0547$)

So the *p*-value is $112/1024 = 10.9\%$, and we don't reject.

Actually there's even less reason to reject. To get a meaningful *p*-value, you must *first* plan and design an experiment, *then* carry it out. Calculating a *p*-value after you lost five hands in a row, or after you noticed that your cultures grew better in the green test tubes, *doesn't* give a reasonable answer. (Why not?)

Board question: gourmet chocolate

The Atlas Gourmet Chocolate Company (gcc) manufactures 10 million chocolate bars each year. Before a bar is sold as a gcc bar, it is subjected to eight independent quality control tests. Three-fourths of the bars pass any one test, but passing all eight is difficult.

1. How many gcc bars should Atlas *expect* each year?

2. As production manager for the factory, would you advise Atlas to *count on* producing a million gcc bars?

3. In a recent year Atlas produced just 998,000 gcc bars. Is this evidence of possible sabotage in the factory?

Begin solution

1.

Passing eight independent tests each with success probability of 0.75 has probability $(3/4)^8 = 6561/65536 = 0.1001129$. Atlas should expect to produce $1,001,129$ gcc bars each year.

2. The number of gcc bars is a random variable following a distribution `binom(10,000,000,0.1001129)`. The variance of `binom(n,`θ`)` is $n\theta(1-\theta)$; so the variance in the number of gcc bars is

$$\text{variance} = (10,000,000) \cdot (0.1001129) \cdot (0.8998871) = 900,903.$$

Standard deviation is the square root of this number, or

$$\text{standard deviation} = 949.$$

Solution continued

A binomial distribution with large n is approximately normal (Central Limit Theorem!) so you'd expect the number of bars produced to be within two standard deviations of the average about 95% of the time. That is

production in the range 999,231–1,003,027 bars

in 95% of years. A million bars is just a bit more that one standard deviation below the mean; looking at a normal table, you'd expect to miss that target about one year in nine.

3. This production level is 3.3 standard deviations below the mean. The normal table says that should happen by chance about once in 2000 years; the one-sided p-value is 0.00048. I think there's a saboteur.

Board Question: Find the pmf

$X = \#$ of successes before the *second* failure of a sequence of independent Bernoulli(p) trials.

Describe the pmf of X.

Hint: this requires some counting.

Answer is on the next slide.

Solution

X takes values 0, 1, 2, ... The pmf is $p(n) = (n+1)p^n(1-p)^2$.

For concreteness, we'll derive this formula for $n = 3$. Let's list the outcomes with three successes before the second failure. Each must have the form

$$\_\_\ \_\_\ \_\_\ \_\_\ F$$

with three S and one F in the first four slots. So we just have to choose which of these four slots contains the F:

$$\{FSSSF, SFSSF, SSFSF, SSSFF\}$$

In other words, there are $\binom{4}{1} = 4 = 3 + 1$ such outcomes. Each of these outcomes has three S and two F, so probability $p^3(1-p)^2$. Therefore

$$p(3) = P(X = 3) = (3+1)p^3(1-p)^2.$$

The same reasoning works for general n.

Board question

I've noticed that taxis drive past 77 Mass. Ave. on the average of once every 10 minutes.

Suppose time spent waiting for a taxi is modeled by an exponential random variable

$$X \sim \text{Exponential}(1/10); \qquad f(x) = \frac{1}{10}\mathrm{e}^{-x/10}$$

(a) Sketch the pdf of this distribution

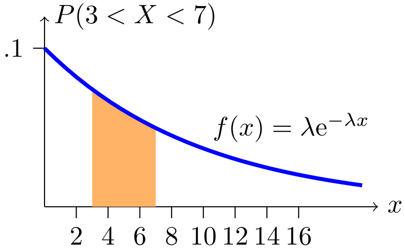
(b) Shade the region which represents the probability of waiting between 3 and 7 minutes

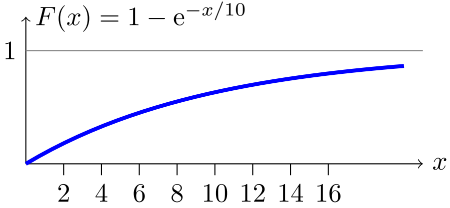
(c) Compute the probability of waiting between between 3 and 7 minutes for a taxi

(d) Compute and sketch the cdf.

Solution

Sketches for (a), (b), (d)





(c)