Problem Sheet 3

Questions

1. The size of a standard croquet ball is 3 5/8 inches[1]. The height of a croquet hoop is 3 3/4 inches. If a not very good croquet-ball making machine makes croquet balls whose mean matches the standard and with standard deviation 1/8 inch, what is the chance it will make a ball too large to fit through the hoop. You can write the solution in terms of the error function.

   Solution: So

   $$z = \frac{x - \mu}{\sqrt{2}\sigma} \tag{1}$$

   so for $x_1 = 3.75$ in, we have

   $$z_1 = \frac{1/8}{\sqrt{2}/8} = \frac{1}{\sqrt{2}} \tag{2}$$

   Any height bigger than this will not fit, so $z_2 = \infty$ and $\text{erf}\infty = 1$ so

   $$\text{Prob}(x > 3.75) = \frac{1}{2}[1 - \text{erf}(1/\sqrt{2})] \approx 0.16 \tag{3}$$

   where the 0.16 is given for interest, it wasn't expected as part of the answer.

2. This will look like a long question but it is almsot all background and the question is not too bad when you actually read through it. In particle physics when a collider is being used to find a new particle like the Higgs boson or the top squark scientists don't detect the sought after particle directly since it usually decays almost straight away, instead they detect the more common particles that particle will decay into, for example, a Higgs boson can decay in to two photons and these can be detected. Roughly speaking scientists count these events. However, the whole situation is very messy and there will always be some events even if the particle doesn't exist at the energy being examined. The amount of these background events will fluctuate from experiment to experiment, typically like a Gaußian. The scientific team is allowed to claim they have discovered the particle if the number of events they measure is more than five standard deviations above what would be expected if the particle didn't exist. What is the probability of this 'discovery' happening by chance?

   Solution: So we are interested in the probability of a results bigger than $\mu + 5\sigma$. Now

   $$z_1 = \frac{\mu + 5\sigma - \mu}{\sqrt{2}\sigma} = \frac{5}{\sqrt{2}} \tag{4}$$

   and

   $$\text{Prob}(x > \mu + 5\sigma) = \frac{1}{2}[1 - \text{erf}(5/\sqrt{(2)})] \tag{5}$$

   which is about one chance in 3.5 million.

[1] Everything in croquet is measured in old timey units

Extra question

- Another useful distribution is the exponential distribution:

$$p(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

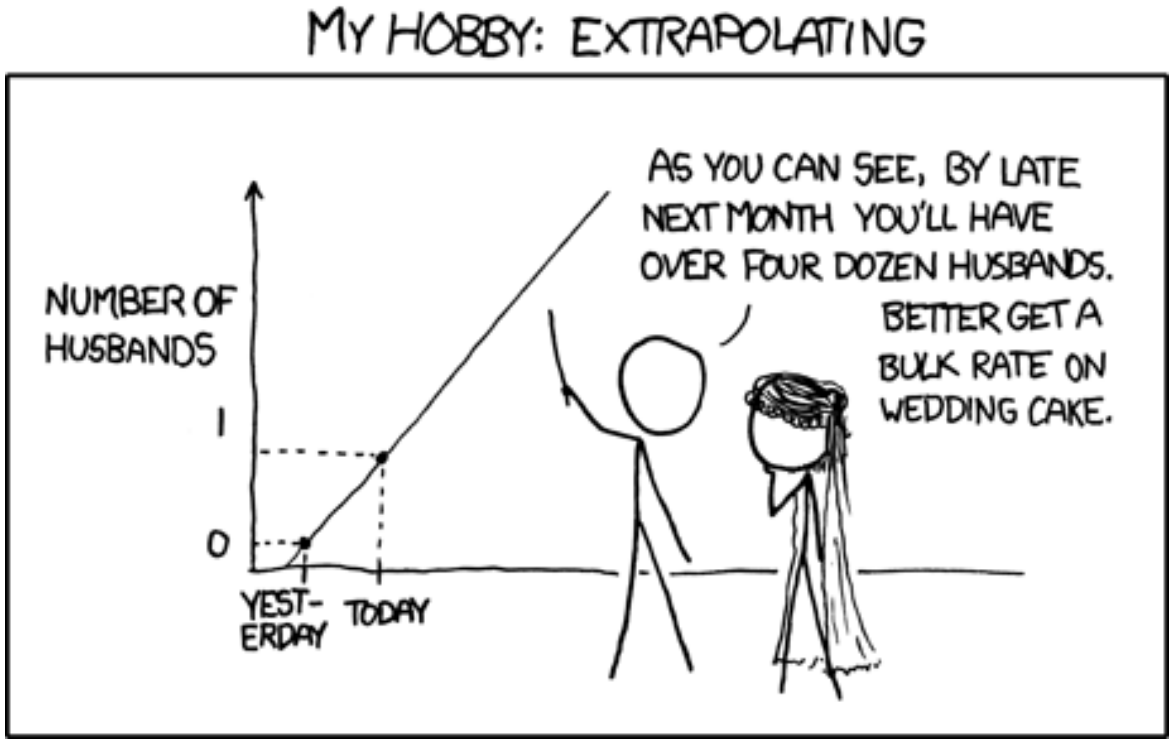
What is the probability $\text{Prob}(x_1 < x < x_2)$ where x_1 and x_2 are both positive.

Solution:

$$\text{Prob}(x_1 < x < x_2) = \lambda \int_{x_1}^{x_2} e^{-\lambda y} dy = e^{-\lambda y} \Big]_{x_1}^{x_2} = e^{-\lambda x_1} - e^{-\lambda x_2} \tag{6}$$

Problem sheet 3

Useful regressions



A couple is planning a road trip for their honeymoon. They want to drive all over Ireland in a camping car. They are planning a 1000km round trip and they wonder how much they should allocate for the fuel. The couple thinks there must be a way to estimate the amount of money needed, based on the distance they are going to travel. What they want to figure out is "If we drive for 1000km, how much will we pay for gas?"

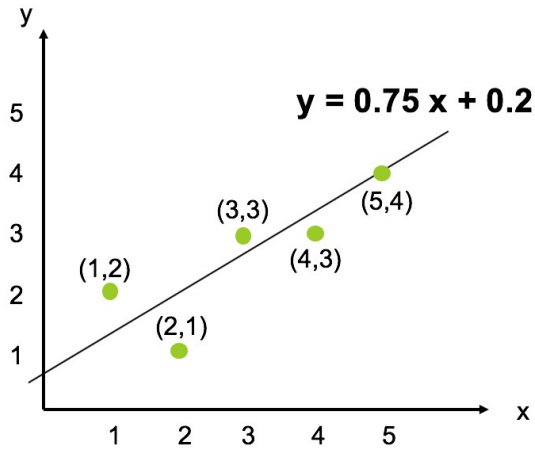
Fortunately, they have been laboriously tracking their car's efficiency for the last year and have made the following graph which seems to show a connection between km driven and fuel paid. They now want to use the data they have been collecting so far and use it to **predict** how much they are going to spend. The idea is that you can make estimated guesses about the future based on data from the past — the data points they have been laboriously logging. To do that they need a mathematical **model** that describes the relationship between km driven and money spent to fill the tank.

Concretely By using linear regression on their log, they can fit a line and find its equation. From this equation they can then predict how much they will pay for gas for their 1000km round trip. Tada!

Questions

Four questions, each worth two marks with two marks for attendance.

Let's try to compute some goodness of fit but I have used some simpler data than our newly wed example to make the computation faster on paper. Below is a graph representing five observations and a regression line.



Q1. Using the equation of the line, compute the goodness of fit using the standard error of the estimate. You can do this by filling in the following table.

x	Actual y	Estimated $\hat{y}$	$\hat{y}$ - y	$(\hat{y} - y)^2$
1	2	1	-1	1
2	1	1.75	0.75	0.5625
3	3	2.5	0.5	0.25
4	3	3.25	0.25	0.0625
5	4	4	0	0

$\sum (\hat{y} - y)^2 / (N-2) = 1.875 / (5-2) = 0.625$

Q2. Using the equation of the line, compute the goodness of fit using the R square. The man of y is (2+1+3+3+4)/5= 2.6. You can do this by filling in the following table.

x	Actual y	y – mean(y)	$(y - \text{mean}(y))^2$	Estimated $\hat{y}$	$\hat{y}$ - mean(y)	$(\hat{y} - \text{mean}(y))^2$
1	2	-0.6	0.36	1	-1.6	2.56
2	1	-1.6	2.56	1.75	-0.85	0.7225
3	3	0.4	0.16	2.5	-0.1	0.01
4	3	0.4	0.16	3.25	0.65	0.4225
5	4	1.4	1.96	4	1.4	1.96

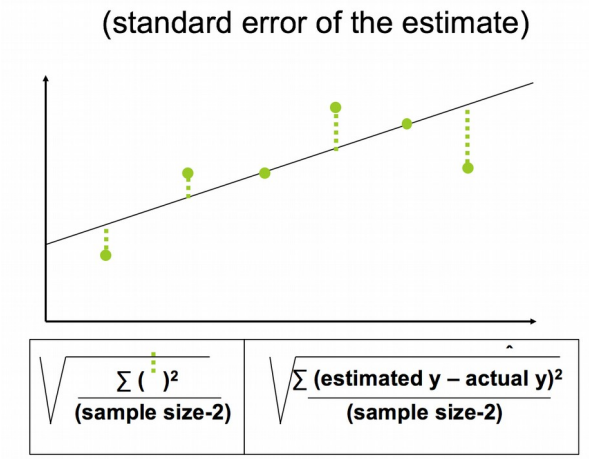
$\Sigma (\hat{y} - \text{mean}(y))^2 = 5.675$
$\Sigma (y - \text{mean}(y))^2 = 5.2$

$R^2 = 5.675 / 5.2 = 1.09.$

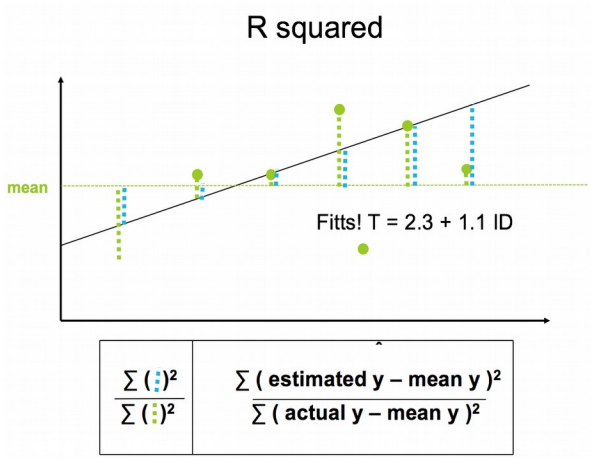
Note that having a R > 1 is not usual and is probably the fact that we have a very small sample size.

Useful

Here are two reminders to compute the goodness of fit.



$$\sqrt{\frac{\sum (\quad)^2}{\text{(sample size-2)}}} \qquad \sqrt{\frac{\sum (\text{estimated y} - \text{actual y})^2}{\text{(sample size-2)}}}$$



$$\frac{\sum (\quad)^2}{\sum (\quad)^2} \qquad \frac{\sum (\text{estimated y} - \text{mean y})^2}{\sum (\text{actual y} - \text{mean y})^2}$$