PSYCGR01: Statistics
Introduction worksheet

Maarten Speekenbrink *UCL Experimental Psychology*
October, 2016

The weekly exercises are divided into two sections. The first section contains general/theoretical questions, which you can answer at home (or any other place you like to work). Some questions may require hand calculation. The questions in the second section are more appropriate for doing in the lab class using a computer program (e.g., SPSS, or possibly R). To answer both sections, you will find it useful to do some reading. Questions marked with an asterisk (∗) are more advanced and optional.

1 General exercises

The Normal distribution has some properties which are handy to know. In the following, assume that a variable Y_1 has a Normal distribution with mean μ_1 and standard deviation σ_1. A second variable Y_2 has a Normal distribution with mean μ_2 and standard deviation σ_2. In shorthand notation:

$$\begin{aligned} Y_1 &\sim N(\mu_1, \sigma_1) \\ Y_2 &\sim N(\mu_2, \sigma_2) \end{aligned}$$

Sum of Normal variables Let Y_S be the sum of two Normally distributed variables:

$$Y_S = Y_1 + Y_2$$

Then Y_S follows a Normal distribution with mean $\mu_S = \mu_1 + \mu_2$, variance $\sigma_S^2 = \sigma_1^2 + \sigma_2^2$ and standard deviation $\sigma_S = \sqrt{\sigma_1^2 + \sigma_2^2}$. In shorthand notation:

$$\begin{aligned} Y_S &= Y_1 + Y_2 & Y_S &\sim N(\mu_S, \sigma_S) \\ \mu_S &= \mu_1 + \mu_2 & \sigma_S &= \sqrt{\sigma_1^2 + \sigma_2^2} \end{aligned}$$

Linear transformation Let Y_T be a *linear transformation* of variable Y_1:

$$Y_T = \beta_0 + \beta_1 \times Y_1$$

(i.e., Y_T is computed by multiplying Y_1 by a constant β_1 and then adding another constant β_0). Then Y_T has a Normal distribution with mean $\mu_T = \beta_0 + \beta_1 \times \mu_1$, variance $\sigma_T^2 = \beta_1^2\sigma_1^2$ and standard deviation $\sigma_T = |\beta_1|\sigma_1$. In shorthand notation:

$$
\begin{aligned}
Y_T &= \beta_0 + \beta_1 \times Y_1 \qquad & Y_T &\sim N(\mu_T, \sigma_T) \\
\mu_T &= \beta_0 + \beta_1 \times \mu_1 \qquad & \sigma_T &= |\beta_1| \times \sigma_1
\end{aligned}
$$

As an example, suppose a fast-food chain is well known for its "mega-megalicious" burger. The weight of each burger is Normally distributed with a mean of 1.2 pounds, and a standard deviation of .16 pounds. We can work out the distribution of the weight in kilograms by first noting that 1 kilogram = 2.2046 pounds and conversely, that one pound is 1/2.2046 = 0.4536 kilograms. Weight in kilograms is thus a linear transformation of weight in pounds:

$$\text{kg} = 0 + 0.4536 \times \text{lbs}$$

The distribution of the weight in kilograms is then a Normal distribution with mean $0.4536 \times 1.2 = 0.5443$ kilograms, and standard deviation $|0.4536| \times .16 = 0.0725$ kilograms.

As a second example, we can derive the sampling distribution of the mean $\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$ when the samples Y_i are from the same Normal distribution with mean μ and standard deviation σ. Let's start by working out the distribution of the sum of 2 Normally distributed variables, Y_1 and Y_2, both with mean μ and standard deviation σ. This sum follows a normal distribution with mean $\mu + \mu = 2\mu$ and standard deviation $\sqrt{\sigma^2 + \sigma^2} = \sqrt{2\sigma^2} = \sqrt{2}\sigma$. This generalizes easily to the sum of n Normally distributed variables, $Y_S = \sum_{i=1}^{n} Y_i$, which is again Normally distributed, now with a mean

$$\mu_S = n \times \mu$$

and standard deviation

$$\sigma_S = \sqrt{n} \times \sigma$$

Now we need to work out the distribution if we take the (sample) mean

$$
\begin{aligned}
\overline{Y} &= \frac{Y_1 + Y_2 + \ldots + Y_n}{n} \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right) \\
&= \frac{1}{n} Y_S
\end{aligned}
$$

This shows that the mean is a linear transformation of the sum with $\beta_0 = 0$ and $\beta_1 = \frac{1}{n}$. We've already worked out the distribution of the sum, Y_S. So the distribution of the sample mean $\overline{Y}$ is a Normal distribution with mean

$$
\begin{aligned}
\mu_{\overline{Y}} &= \beta_0 + \beta_1 \mu_S \\
&= 0 + \frac{1}{n} \times (n \times \mu) \\
&= \frac{n}{n} \times \mu \\
&= \mu
\end{aligned}
$$

and standard deviation

$$\begin{aligned}
\sigma_{\overline{Y}} &= |\beta_1|\sigma_S \\
&= \left|\frac{1}{n}\right| (\sqrt{n}\sigma) \\
&= \frac{\sqrt{n}}{n}\sigma \\
&= \frac{1}{\sqrt{n}}\sigma \\
&= \frac{\sigma}{\sqrt{n}}
\end{aligned}$$

(If you're wondering how we got from line 3 to line 4 above, you should remember that you can multiply or divide the top and bottom part of a fraction by the same number without affecting the result, so $\frac{\sqrt{n}}{n} = \frac{\sqrt{n}/\sqrt{n}}{n/\sqrt{n}} = \frac{1}{\sqrt{n}}$.)

1. Use the two properties of the Normal distribution to answer the following questions:

   (a) Suppose the temperature in degrees Celcius on 3 October follows a Normal distribution with mean 17 and standard deviation 2. What is the distribution of the temperature in degrees Fahrenheit on 3 October (note: $°\text{F} = 32 + \frac{9}{5} \times °\text{C}$)?
   Answer: Fahrenheit is a linear transformation of Celcius, with $\beta_0 = 32$ and $\beta_1 = \frac{9}{5}$, so the distribution of temperature in Fahrenheit is also Normal, but with mean

$$\begin{aligned}
\mu_F &= \beta_0 + \beta_1 \times \mu_C \\
&= 32 + \frac{9}{5} \times 17 \\
&= 62.6
\end{aligned}$$

   and standard deviation

$$\begin{aligned}
\sigma_F &= |\beta_1| \times \sigma_C \\
&= \left|\frac{9}{5}\right| \times 2 \\
&= \frac{18}{5} \\
&= 3.6.
\end{aligned}$$

   (b) Suppose we predict the temperature on 3 October to be 16 degrees Celcius. What is the distribution of the prediction error $e_i = \texttt{temperature}_i - 16$?
   Answer: The prediction error is a linear transformation of the temperature, with $\beta_0 = -16$ and $\beta_1 = 1$, so it also has a Normal distribution with mean

$$\begin{aligned}
\mu_e &= \beta_0 + \beta_1 \times \mu_C \\
&= (-16) + 1 \times 17 \\
&= 1
\end{aligned}$$

and standard deviation

$$\begin{aligned}\sigma_e &= |\beta_1| \times \sigma_C \\ &= 1 \times 2 \\ &= 2\end{aligned}$$

2. Direct a web browser to `http://watch.psychol.ucl.ac.uk:3838/shiny_apps/PSYCGR01/NormalSampling`. This will open an online program which allows you to draw (repeated) samples from a Normal distribution.

    (a) Draw 10 samples (each consisting of 10 observations). For each sample, write down the sample mean and median. Is the mean or the median generally closer to the population mean/median?

    Answer: Generally, the mean should be closer to the population mean/median. However, with random samples, this will not always be the case.

    (b) Repeat the above, but now with samples of 100 observations each.

    Answer: Samples of 100 observations should provide means and medians which are closer to the population mean/median than from samples of 10 observations. But for these samples, the sample mean should still be closer to the population mean/median than the sample median. With random samples, this will not always be the case.

3. Direct a web browser to `http://watch.psychol.ucl.ac.uk:3838/shiny_apps/PSYCGR01/Distributions`. This will open an online program which allows you to compute probabilities for a number of distributions.

    (a) For a variable Y, which follows a Normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$, what is $P(Y \geq 1.96)$, the probability that Y is greater than or equal to 1.96? What is $P(Y \leq -1.96)$, the probability that Y is less than or equal to -1.96? What is $P(Y \leq -1.96 \text{ or } Y \geq 1.96)$, the probability that Y is either less than or equal to -1.96, or greater than or equal to 1.96? What is $P(-1.96 \leq Y \leq 1.96)$, the probability that Y greater than or equal to -1.96 and less than or equal to 1.96?

    Answer: The first probability is $P(Y \geq 1.96) = 0.024998 \approx .025$

    The second probability is $P(Y \leq -1.96) = 0.024998 \approx .025$. This illustrates that the Normal distribution is symmetric; the probability of obtaining a value higher than $\mu + x$ is equal to the probability of obtaining a value smaller than $\mu - x$ (where x can take any value).

    To obtain the third probability, you should realise that $P(Y \leq -1.96 \text{ or } Y \geq 1.96) = P(Y \leq -1.96) + P(Y \geq 1.96) = 0.024998 + 0.024998 = 0.0499958 \approx 0.05$.

    The final probability can be computed as either $P(-1.96 \leq Y \leq 1.96) = P(Y \leq 1.96) - P(Y \leq -1.96) = 0.975002 - 0.024998 \approx 0.95$, or $P(-1.96 \leq Y \leq 1.96) = 1 - P(Y \leq -1.96 \text{ or } Y \geq 1.96) = 1 - 0.0499958 \approx 0.95$.

(b) For a variable Y, which follows a Normal distribution with mean $\mu = 10$ and standard deviation $\sigma = 1$, what is $P(Y \geq 11.96)$, the probability that Y is greater than or equal to 11.96?

Answer: The probability is $P(Y \geq 11.96) = 0.024998$, which is the same as the first probability computed above. In both cases, we are computing $P(Y \geq \mu + 1.96\sigma)$.

(c) For a variable Y, which follows a Normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 10$, what is $P(Y \geq 19.6)$, the probability that Y is greater than or equal to 19.6?

Answer: The probability is $P(Y \geq 19.6) = 0.024998$, which is the same as the probability above. Again, we are computing $P(Y \geq \mu + 1.96\sigma)$. The main thing to take away is that for the Normal distribution, these probabilities are completely determined in terms of “standard deviations from the mean”.

2 Computer exercises

1. The datafile `IQscores.sav` contains IQ scores of 100 fictitious participants in an experiment. Let’s assume the participants are a random sample from the general population. So, we assume the IQ scores (Y_i) can be modelled as MODEL C:

$$Y_i = 100 + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma) \tag{1}$$

(so ϵ_i is a normally distributed variable with mean 0 and standard deviation σ).

(a) What are the mean, mode and median values of IQ according to MODEL C?

Answer: MODEL C states that IQ is Normally distributed variable with a mean of $\mu = 100 + 0$ (you can think of it as a linear transformation of ϵ with $\beta_0 = 100$). In a Normal distribution, the mean, median, and mode are all the same, so 100.

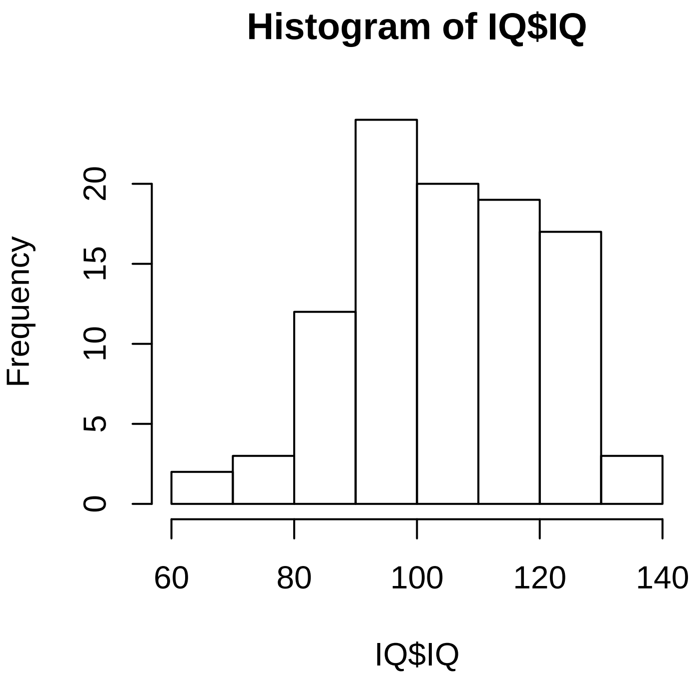
(b) Compute the total number of errors, the sum of absolute errors, and the Sum of Squared Errors (SSE) from MODEL C.

```
> library(foreign)
> IQ <- read.spss("~/MEGA/MSc Statistics/data/IQscores.sav")
> # total number of errors:
> sum( (IQ$IQ - 100) != 0)

[1] 99

> # sum of absolute errors:
> sum(abs(IQ$IQ - 100))

[1] 1342

> # sum of squared errors
> sum((IQ$IQ - 100)^2)

[1] 25988
```

Answer: Total number of errors is 99; the sum of absolute errors is 1342, and the SSE is 25,988.

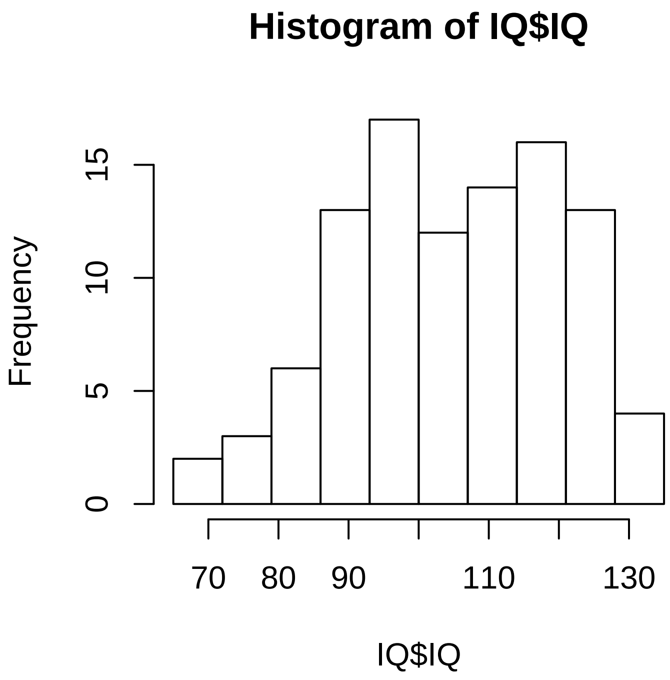
(c) MODEL C is a theoretical model of the population distribution. The sample data may, or may not be, actually drawn from this theoretical distribution. Create a histogram (frequency distribution) of the sample data.

```
> hist(IQ$IQ)
```



(d) Create a histogram (frequency distribution) with the data grouped into 10 intervals. Visually compare this graph to the previous one. Does one look more like what you would expect from MODEL C?

```
> hist(IQ$IQ,breaks=seq(min(IQ$IQ),max(IQ$IQ),length=10+1))
```



Answer: The main thing to notice here is that the visual appearance of a histogram depends on the number of bins used. Generally, a histogram looks “rougher”, but more detailed, when more bins are used. A smaller number of bins generally looks “smoother”. In this case, neither graph looks exactly like a Normal distribution. Note that, due to sampling variation, such deviations from a Normal distribution are to be expected.

(e) What are the sample mean, mode and median? What is the standard deviation and variance in the sample?

```
> # sample mean
> mean(IQ$IQ)

[1] 105.4

> # sample median
> median(IQ$IQ)

[1] 106

> # sample mode
> # first deine the function
> Mode <- function(x) {
+   ux <- unique(x)
+   nx <- tabulate(match(x, ux))
+   maxid <- which(nx == max(nx))
+   if(length(maxid) > 1) {
```

```
+     warning("There are multiple modes, only the first value is returned, o
+   }
+   ux[maxid[1]]
+ }
> # now compute the mode
> Mode(IQ$IQ)

Warning in Mode(IQ$IQ): There are multiple modes, only the first value
is returned, other modes are at:  98, 122

[1] 109

> # variance
> var(IQ$IQ)

[1] 233.0505

> # standard deviation
> sd(IQ$IQ)

[1] 15.26599
```

Answer: The sample mean is $\overline{Y} = 105.4$, the sample median is 106, and there are three sample modes, at 98, 109 and 122. The sample variance is 233.05 and the standard deviation is 15.27.

(f) Compute the number of errors, using the sample mode. Compare this value to the one for MODEL C. Which is smaller? Can you explain this result?

```
> sum( (IQ$IQ - Mode(IQ$IQ)) != 0)

Warning in Mode(IQ$IQ): There are multiple modes, only the first value
is returned, other modes are at:  98, 122

[1] 95
```

Answer: There are 5 observations for each mode, so the total number of errors, regardless of which one you take, is 100-5 = 95. The total number of errors is reduced compared to that for MODEL C. This is because using the sample mode for your predictions minimizes the total number of errors in the sample.

(g) Compute the sum of absolute errors, using the sample median as the prediction. Compare this value to the one for MODEL C. Which is smaller? Can you explain this result?

```
> sum(abs(IQ$IQ - median(IQ$IQ)))

[1] 1260
```

Answer: The sum of absolute errors is 1260. The sum of absolute errors is reduced compared to that for MODEL C. This is because using the sample median for your predictions minimizes the sum of absolute errors in the sample.

(h) Compute the Sum of Squared Errors (SSE), using the sample mean as the prediction. Compare this value to the one for MODEL C. Which is smaller? Can you explain this result?

```
> sum((IQ$IQ - mean(IQ$IQ))^2)

[1] 23072
```

Answer: The SSE is 23,072. The SSE is reduced compared to that for MODEL C. This is because using the sample mean for your predictions minimizes the SSE in the sample.