01_statistical_inference - Week 1

Notes on the course
https://www.coursera.org/learn/statistical-inference

Andres FR

July 2020

WEEK 1

Intro video

> **STATISTICAL INFERENCE**: "Generating conclusions about a population from noisy data, where uncertainty must be accounted for".

Instead of simply navigating the data, we extract knowledge: statistical inference lets scientists formulate conclusions from data AND quantify the uncertainty arising from using incomplete (or bad/contaminated) data.

- "The only formal system of inference that we have". Lecture 05 04: "I would say, fundamentally, what differentiates understanding statistics from not understanding statistics is understanding variability".

We will follow the frequentist paradigm. It is popular, and serves as good foundation to expand.

Example Application: Infer causality.

> A **statistic** (singular) is a number computed from a sample of data. We use statistics to infer information about a population.

> A **probability model** connects data to a population using assumptions

---

Setup

Warning from the future: if the following line installs R 3.4, do **not** do it this way! To avoid dependency hell, find a way to install 3.5 or greater

```
sudo apt install r-base
```

The instructions here solved this for me: https://rtask.thinkr.fr/installation-of-r-3-5-on-ubuntu-18-04-lts-and-tips-for-spatial-packages/

update emacs to 26 and installed package ess https://stackoverflow.com/a/1423006

```
C-c C-e C-r     inferior-ess-r-reload-hook  (restart interpreter)
C-c C-b         ess-eval-buffer
C-c C-f         ess-eval-function
C-c C-j         ess-eval-line
C-c C-l         ess-load-file
and many others, check C-h m
```

cloned: https://github.com/bcaffo/courses This course is number 6: StatisticalInference

> **IMPORTANT**: The lectures are in the index.Rmd lecture files. Kind of markdown files that auto generate the slides. In Data Products, we'll cover how to create these sorts of slides. They also have code? INSTALL RSTUDIO:

https://rstudio.com/products/rstudio/download-server/debian-ubuntu/

```
sudo apt-get install gdebi-core
wget https://download2.rstudio.org/server/bionic/amd64/rstudio-server-1.3.1056-amd64.deb
sudo gdebi rstudio-server-1.3.1056-amd64.deb
# Running on localhost:8787, login with OS credentials
```

- Rstudio: installed all requested packages and texlive. Moved this notes to RStudio

```
Ctrl+Shift+F10: restart R session
Ctrl+Shift+K: create PDF
Ctr+Shift+Enter: Run current R chunk. Without the shift, run without displaying
{r, fig.height = 5, fig.width = 5, echo = TRUE, fig.align='center'} # plot example
eval = FALSE  # show actual code
```

Installed swirl: "swirl is a software package for the R programming language that turns the R console into an interactive learning environment. Users receive immediate feedback as they are guided through self-paced lessons in data science and R programming." For help on swirl, see https://github.com/swirldev/swirl/wiki/Coursera-FAQ

```
sudo apt-get install libcurl4-gnutls-dev
sudo apt install libssl-dev
install.packages("swirl")
packageVersion("swirl")
library(swirl)
# If you have existing variables:
ls()  # see a list of the variables in your workspace
rm(list=ls())  # clear your workspace
#
bye()  # or ESC to exit swirl. Progress will be saved
skip()  # skip current question
play()  # sandbox, swirl will ignore what you do
nxt()  # exit sandbox mode, back to swirl repl
main()  # swirl's main menu
info()  # show these options
```

Also installed:

```
# This will fail for R version 3.4
# Data Sets, Etc. for the Text ``Using R for IntroductoryStatistics'', Second Edition
install.packages("UsingR", dependencies=TRUE)
install.packages("reshape2")
install.packages("Hmisc", dependencies = T)
install.packages("stringi")
```

Also changed RStudio layout: https://www.r-bloggers.com/a-perfect-rstudio-layout/

If you'd prefer to watch the videos on YouTube, the current version of the class is here: https://www.youtube.com/playlist?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

There is also an eBook that can be found on github. The exercises may be harder but will help a lot: https://leanpub.com/LittleInferenceBook

Info on plagiarism: http://www.jhsph.edu/academics/degree-programs/master-of-public-health/current-students/JHSPH-StudentReferencing_handbook.pdf

> **SEE ALSO**: web for community-based content: http://datasciencespecialization.github.io/

Introduction to Probability

- Given a random experiment, "a probability measure is a POPULATION QUANTITY that summarizes the randomness". I.e. is a "property" of the population (e.g. a die), an NOT THE DATA. Inherently, whenever you say the word "probability", YOU ARE TALKING ABOUT A POPULATION QUANTITY (next video).

- “Probability calculus”: The ruls that a prob has to follow.
- Prob. is a function that takes a set of possible outcomes and assigns a number in [0, 1].
  - 0: impossible event, 1: sure event.
  - Union of disjoint sets, sum. Union of joint=sum minus p(intersection)
  - prob of complementary set = 1 - prob of set
  - If event A implies B, p(A) < p(B), because A is a strict subset of B

Probability mass functions

- Density/mass functions is all we need (e.g. bell curve).

> **NOTE**: “When you talk about probabilities associates with the bell curve in a norm. dist, you are talking about POPULATION QUANTITIES, and not making statements about what occurs in the data”.

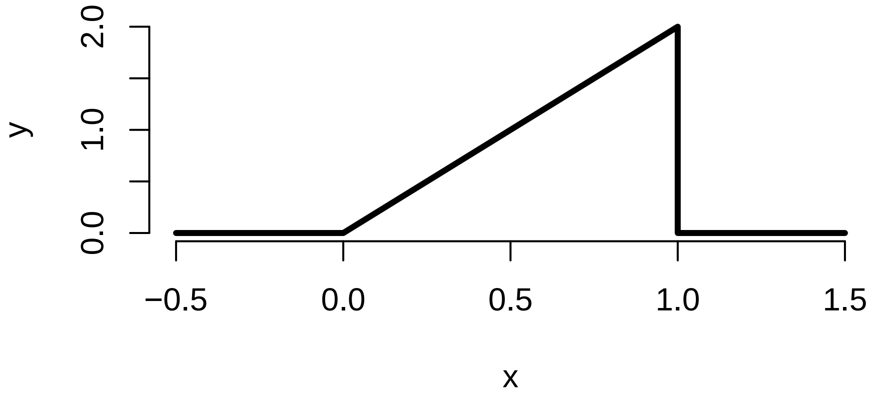
So we collect data and use it to estimate properties of the population. But first, we need to work well with population quantities

- A **random variable** is a numerical outcome of an experiment. Can be discrete or continuous.
- Example: Bernoulli dist (e.g. coin flip):
  - P(1) = theta^1 * (1-theta)^(1-1) = theta
  - P(0) = theta^0 * (1-theta)^(1-0) = 1-theta
  - We can assume that a population follows this dist and then estimate theta from the data.

Probability density functions

- “Areas under pdfs correspond to probabilities for that random variable”
- Example: ascending triangular function between 0 and 2.

```r
x <- c(-0.5, 0, 1, 1, 1.5)
y <- c(0, 0, 2, 0, 0)
plot(x, y, lwd=3, frame=FALSE, type="l")
```



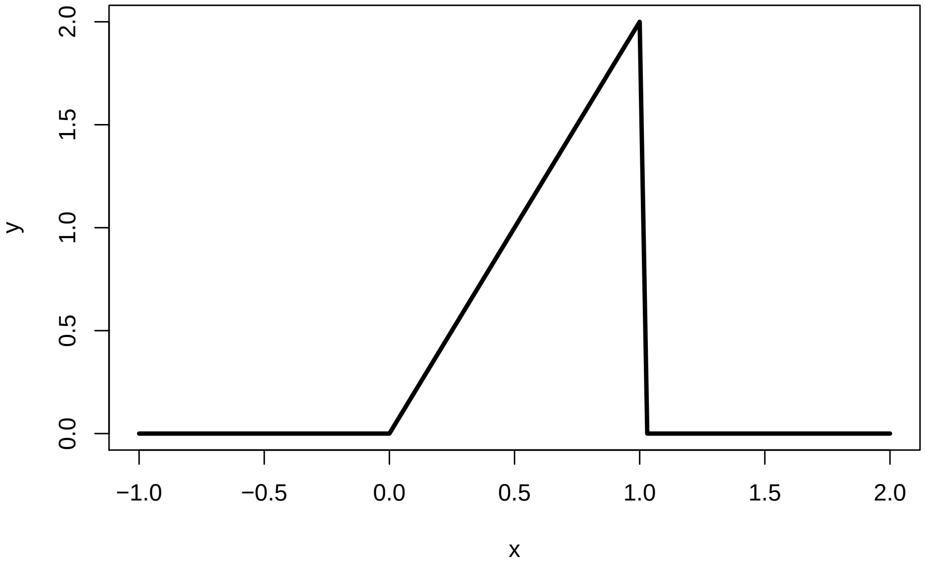
This happens to be a beta density, so the integral from 0 to 0.75 can be retrieved as follows: 2 and 1 are the parameters that

```
pbeta(0.75, 2, 1)
```

```
## [1] 0.5625
```

We can indeed plot the corresponding distribution `dbeta` (`pbeta`is the CDF):

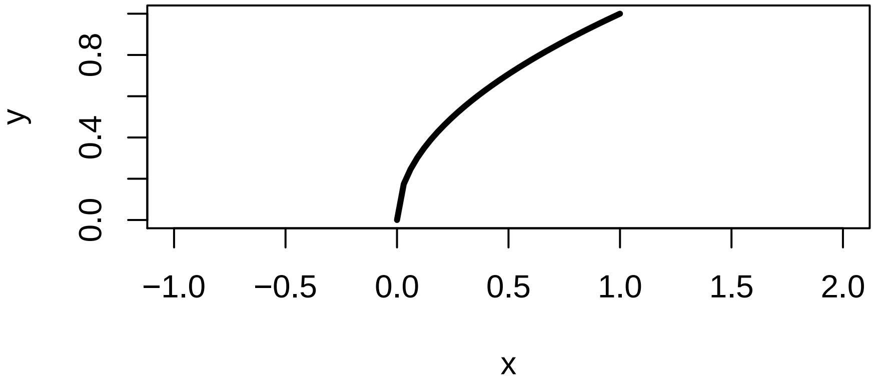
```
x <- seq(-1, 2, length=100)
y <- dbeta(x, 2, 1)
plot(x, y, type="l", lwd=3)
```



- The `p` prefix is always the **CDF** in r, also known as F, is the integral of the PDF, f. The survival function is $1 - F$.

- **QUANTILES**: The α^{th} quantile of a distribution is the point x_α at which $F(x_\alpha) = alpha$. In the CDF, we input x-values and the y raises from 0 to 1. For quantiles, we input the y-value and get the corresponding x-value. **Percentiles** are simply quantiles with α expressed as percent. The **median** is the 50th percentile.

```r
x <- seq(-1, 2, length=100)
y <- qbeta(x, 2, 1)
plot(x, y, type="l", lwd=3)
```



- In contrast to this "population median", **Sample median** doesn't require integration, just sorting samples and picking middle. But here we use both sample (**estimator**) and population (**estimand**) median.

> "A lot of assumptions are needed to connect the data (samples) to the population. A probability model is used to formally develop them"

Conditional probability

- Definition:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

If A and B are disjoint events, it turns out to be $P(A)$.

Bayes' rule

Definition:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

- Example Application: Diagnostic tests: Given:
    - `+` if test is positive, `-` if negative
    - `D` if person has disease, `d` if not

Sensitivity $= P(+|D)$ **Specificity** $= P(-|d)$

We want both to be high. Also,

prevalence of the disease $= P(D)$ **positive predictive value** $= P(D|+)$ **negative predictive value** $= P(d|-)$ Example: A test has a specificity of 99.7% and sensitivity of 98.5%. In a population with 0.1% prevalence, what is the associated positive predictive value?

$$\begin{aligned}
P(D) &= 0.001 \\
P(+|D) &= 0.997 \\
P(-|d) &= 0.985 \\
P(+) &= P(+|D)P(D) + P(+|d)P(d) \\
&= 0.997 * 0.001 + (1 - P(-|d)) \cdot (1 - P(D)) \\
&= 0.000997 + 0.015 \cdot 0.999 \approx 0.01598 \\
P(D|+) &= \frac{P(+|D) \cdot P(D)}{P(+)} \approx \frac{0.997 \cdot 0.001}{0.01598} = 0.06239 \approx 6.24\%
\end{aligned}$$

We see that the low pos. predictive value is given largely due to the low prevalence in that population. But if the subject came from a population with much more prevalence of the disease, the PPV would be much higher.

DIAGNOSTIC LIKELIHOOD RATIO: Another related and important quantity is its complementary, $P(d|+) = 1 - PPV$. If we were to divide PPV by its complementary, given the fact that both have same Bayes denominator, we get the following:

$$\frac{P(D|+)}{P(d|+)} = \frac{P(+|D) \cdot P(D)}{P(+|d) \cdot P(d)} = \frac{P(D)}{P(d)} \cdot \frac{P(+|D)}{P(+|d)}$$

The interpretation is as follows:

> “The odds of disease given a positive test result equal the prior odds of having the disease multiplied by the **diagnostic likelihood ratio for a positive test result** (or DLR_+)”. This ratio tells you how your prior ratio changes after a positive result for this specific test.

EX: for sensitivity=99.7%, specificity=98.5%,

$$\frac{P(+|D)}{P(+|d)} = \frac{0.997}{1 - P(-|d)} = \frac{0.997}{1 - 0.985} \approx 66.47$$

In other words, given +, “the hypothesis of D is 66 times more supported than not having D”.

The DLR_-, on the other hand, is:

$$\frac{P(D|-)}{P(d|-)} = \frac{P(-|D) \cdot P(D)}{P(-|d) \cdot P(d)} = \frac{P(D)}{P(d)} \cdot \frac{P(-|D)}{P(-|d)}$$

So, in our example,

$$\frac{P(-|D)}{P(-|d)} = \frac{1 - P(+|D)}{0.985} = \frac{1 - 0.997}{0.985} \approx 0.003$$

As we can see, DLR_- is NOT the inverse of DLR_+.

Independence

Definitions: A and B are independent, if

- $P(A|B) = P(A)$, for $P(B) > 0$
- $P(A \cap B) = P(A)P(B)$
- **IID**: Random variables are said to be independent, identically distributed if they all have been drawn from the same population distribution and are statistically unrelated among eachother. IID is a default statistical model for many scenarios and will be the main one for this class.

Expected Value

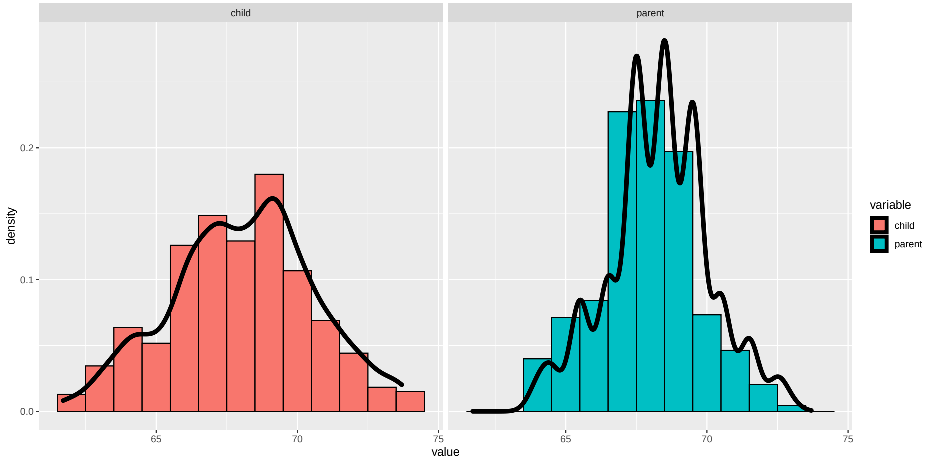
Same as sample quantiles estimate population quantiles, sample EV and Variance estimate the population's EV and Variance.

EV definition ("center of mass"): * Discrete RV X: $E[X] = \sum_x xp(x)$ * Continuous RV X: $E[X] = \int_x xp(x)dx$

It is interesting that the **sample mean** is the "center of mass" of the empirical data *if we treat each point as EQUALLY LIKELY*, so $p(x_i) = \frac{1}{n}$ in the expression: $\bar{X} = \sum_{i=1}^{n} x_i p(x_i)$

To illustrate this, we can try out different means on the interactive plot below, and see that the sample mean actually balances the sample data (by minimizing the MSE), and is a proxy for the population mean (next video).

> Note on **Interactive plots**: The `manipulate` package allows for interactive sliders, etc, but cannot be directly rendered into PDF. A practical solution is to add `eval=FALSE` to the interactive plots, so they appear on RStudio and the code appears on the PDF. Also, the option "Chunk Output in Console" must be active.



```
library(manipulate)
myHist <- function(mu){
    g <- ggplot(galton, aes(x = child))
    g <- g + geom_histogram(fill = "salmon",
      binwidth=1, aes(y = ..density..), colour = "black")
    g <- g + geom_density(size = 2)
    g <- g + geom_vline(xintercept = mu, size = 2)
    mse <- round(mean((galton$child - mu)^2), 3)
    g <- g + labs(title = paste('mu = ', mu, ' MSE = ', mse))
    g
}
manipulate(myHist(mu), mu = slider(62, 74, step = 0.5))
```

Expected values, simple examples

- Example of population mean:
  - Expected value of `Bernoulli(p)` is p
  - EV of `Unif(min, max)` is $\frac{max-min}{2}$
- EV is a property of the distribution (center of mass)

The average of RVs is itself a RV, so its associated distribution also has an EV. The center of this avg. dist. is the same as that of the original dist. In other words: The EV of the population is the same as the EV of the sample mean dist.

An **estimator** e of some parameter v is unbiased if $E[e] = v$. We can show that the EV of an iid sample mean equals the population mean via linearity of EV: Given the sample mean $\bar{X} = \frac{X_1+X_2\dots}{n}$ from the

> RV X, the EV of the sample mean is: $E[\frac{X_1+X_2...}{n}] = \frac{E[X_1]+E[X_2]...}{n} = \frac{n \cdot E[X]}{n} = E[X]$
>
> We also say (07 01) an estimator is **consistent** if it converges to what you want to estimate.

Example: If we sample from a normal distribution, and then sample from *averages* of multiple samples from that same distribution, the second sampled dist. will have much less variance, but *both will be centered around the same EV.*

More examples: Rolling a die thousands of times, shows an uniform distribution with values from 1 to 6 and EV of 3.5. If we average multiple rolls, the more rolls we average the more it approximates a normal distribution. But note that the center is always 3.5. Same with a coin: A 50/50 discrete distribution between 0 and 1 approaches a normal distribution centered at 0.5, and the more we average the more it concentrates around EV.

This is very useful because when we collect data, we only get 1 sample mean, not information about the sample mean dist. So knowing these properties can help.