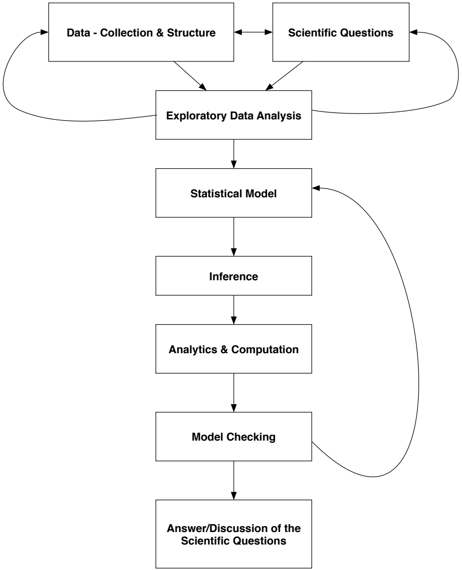
Statistical Inference

Lecture 01b

ANU - RSFAS

Last Updated: Wed Feb 21 10:56:19 2018

Thoughts on Statistics & Science - Example



Macroeconomics

- Scientific Question/Theory
  - What impacts the total production in a country (Y)?
    - Perhaps labor (L), capital (K), productivity (A).
    - $Y = h(L, K, A)$.
    - Cobb-Douglas production function: $Y = AL^{\beta}K^{\alpha}$
- What data are available (http://data.worldbank.org)?
  - GDP, Population, Labor Force, ...
- Let's start simple with GDP & Labor Force for 2013.

```
gdp <- read.csv("gdp2013.csv")
labor <- read.csv("labor2013.csv")
Data <- merge(gdp, labor, by=c("Country.Name",
                               "Country.Code"))
```

Data

```r
head(Data)
```

```
##      Country.Name Country.Code      X2013.x  X2013.y
## 1     Afghanistan          AFG  20309671015  7811221
## 2         Albania          ALB  12923240278  1212997
## 3         Algeria          DZA 210183000000 12431290
## 4  American Samoa          ASM           NA       NA
## 5         Andorra          AND           NA       NA
## 6          Angola          AGO 124178000000  7890692
```

```r
names(Data)[3:4] <- c("gdp", "labor")
names(Data)
```

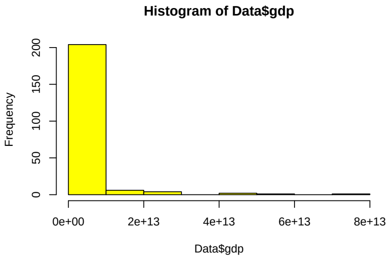
```
## [1] "Country.Name" "Country.Code" "gdp"          "labor"
```

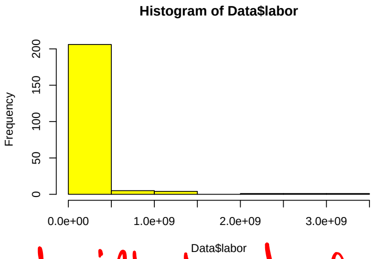
EDA

```r
par(mfrow=c(2,2))
hist(Data$gdp, col="yellow")
hist(Data$labor, col="yellow")
plot(Data$labor, Data$gdp)
```

```r
par(mfrow=c(2,2))
hist(log(Data$gdp), col="yellow")
hist(log(Data$labor), col="yellow")
plot(log(Data$labor), log(Data$gdp))
```

EDA

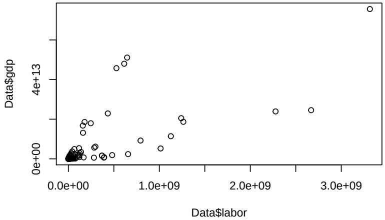




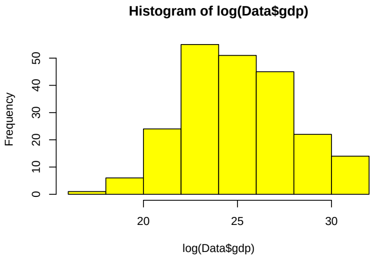
extremely right-skewed & positive

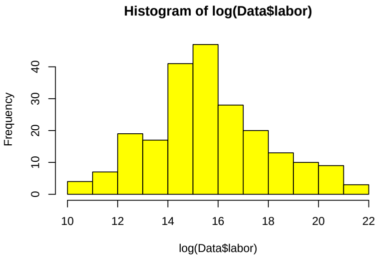
transformation

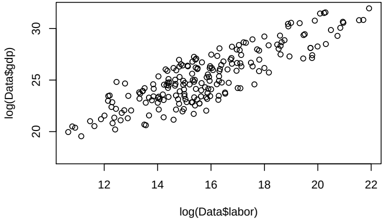
log



EDA


Histogram of log(Data$gdp)
Frequency
0
10
20
30
40
50
20
25
30
log(Data$gdp)


Histogram of log(Data$labor)
Frequency
0
10
20
30
40
10
12
14
16
18
20
22
log(Data$labor)


log(Data$gdp)
20
25
30
12
14
16
18
20
22
log(Data$labor)

Statistical Model

- Simple linear regression model:

3 parameters: intercept, slope & errors

$$\log(\text{GDP})_i = \beta_0 + \beta_1 \log(\text{labor})_i + \epsilon_i$$
$$\epsilon_1, \ldots, \epsilon_n \sim \text{normal}(0, \sigma^2)$$

- Hmmmm . . . seems to fit nicely with the economic theory:

$$Y = AL^{\beta}K^{\alpha}$$
$$log(Y) = \underbrace{log(A)}_{\beta_0} + \underbrace{\beta}_{\beta_1} log(L) + \cancel{\alpha log(K)}$$

Estimation of the Parameters and Computation

LS : a way/method , a tool.

- $\theta = \{\beta_0, \beta_1, \sigma^2\}$
- Many ways to proceed for inference. In regression class you learned about least-squares estimation but we can also consider maximum likelihood, Bayesian, ... other tools
- You will hear people say "I fit a least-squares model" or "I have a least-squares model". This is incorrect!! They have a model and used least-squares to estimate the parameters!!

$$min_{\beta_0,\beta_1} \sum_{i=1}^{n} (\underbrace{\log(\overset{Y}{\text{GDP}})}_{} - [\underbrace{\overset{E(Y)}{\beta_0 + \beta_1 \log(\text{labor})}}])^2$$

minimize the distance b/w random variable Y & mean of Y

- Computation/analytics is the actual mechanism to determine the minimum.

Estimation of the parameters and Computation

- Let's estimate the parameters in R (via least-squares):

```
mod <- lm(log(gdp) ~ log(labor), data=Data)
summary(mod)
```

Estimation of the Parameters and Computation

```
##
## Call:
## lm(formula = log(gdp) ~ log(labor), data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2597 -1.0684  0.0685  0.9935  2.8452
##
## Coefficients:
```

$\hat{\beta}_0$

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.66902    0.66436   14.55   <2e-16 ***
## log(labor)   0.98753    0.04165   23.71   <2e-16 ***
```

$\hat{\beta}_1$

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.4 on 204 degrees of freedom
##   (42 observations deleted due to missingness)
## Multiple R-squared:  0.7338, Adjusted R-squared:  0.7325
## F-statistic: 562.2 on 1 and 204 DF,  p-value: < 2.2e-16
```

Model Checking

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i; \quad \epsilon_i \overset{\text{iid}}{\sim} \text{normal}(0, \sigma^2)$$

- Residual analyses:
  - Plot $\hat{\epsilon}$ against $x \Rightarrow$ any odd patterns of outliers
  - Plot a histogram or QQ plot of $\hat{\epsilon} \Rightarrow$ examine normality of the residuals.
  - Michael Ward and Kristian Gleditsch suggest that GDP (along with many national level data) are not independent but spatially dependent (this also can be examined via residual analyses).

Michael Ward and Kristian Skrede Gleditsch. 2008. Spatial Regression Models. Thousand Oaks, CA: Sage.

- What type of sample did I take? It is pretty clear I have a finite population. Actually a Bayesian paradigm has nice interpretation to this question. More to come . . .

Answering the Scientific Questions

- From the results of the statistical analysis we can say:

"If we observe an increase in the log of labor by one unit then we predict that the log of GDP will increase by 0.9875." Here we have a point estimate (single best guess).

- We can also add a numerical uncertainty statement (interval estimate) for that prediction! More to come ...
- What does "observe" mean in the above? Do we have observational or experimental data?

Generating Random Variables

- In many situations it will be useful to be able to generate samples from a distribution and examine functions (i.e. statistics) of those simulated data (frauta = simulated [fradulent data])
- Given $X_1, \ldots, X_N \sim f(x;\theta)$, we will generate random samples of X to learn about their behavior, as well as $h(X)$.
- If we generate independent samples, then this is termed Monte Carlo analysis.
- Monte Carlo integration:
  - Many quantities of statistical analyses can be expressed as the expectation of a function of a random variable $E[h(X)]$.
  - Let $f(X|\theta)$ denote the density of X
  - Let μ denote the expectation of $h(X)$.
  - Then when an iid sample $X_1, \ldots, X_n$ is obtained from $f(X;\theta)$, we can approximate μ by a sample average:

$$\hat{\mu}_{MC} = \frac{1}{n}\sum_{i=1}^{n} h(X_i) \to \int h(x)f(x)dx = \mu$$

Monte Carlo Integration

- We can approximate σ^2 similarly:

$$\hat{\sigma}^2_{MC} = \frac{1}{n-1} \sum_{i=1}^{n} \left(h(X_i) - \hat{\mu}_{MC}\right)^2 \to \sigma^2$$

- **These results are based on the Law of Large Numbers.**

Monte Carlo Integration

Example (Exponential lifetime):

- Suppose that a particular electrical component can be modeled with an exponential ($\beta = 50$) lifetime.

$$f(x;\beta) = \frac{1}{\beta} exp(-x/\beta)$$

- The manufacturer is interested in determining the probability that, out of $c = 100$ components, at least $t = 35$ of them will last $h = 45$ hours.

- We can first consider the analytical solution. The probability that a single component last at least $h = 45$ is:

$$p_1 = \int_{45}^{\infty} \frac{1}{50} exp(-x/50) dx = 1/exp(45/50) \approx 0.4066$$

```r
set.seed(1001)
n <- 20000
x <- rexp(n, 1/50)
x[1:5]
```

```
## [1]  14.30061  34.86863 118.94469  42.38883  26.53421
```

```r
mean(x)
```

```
## [1] 50.22228
```

```r
p1 <- length(x[x>=45])/n
p1
```

```
## [1] 0.4082
```

```r
mean(x>=45)
```

```
## [1] 0.4082
```

$$
\begin{aligned}
p_2 &= P(\text{at least } t = 35 \text{ components last at least } h = 45 \text{ hours}) \\
&= \sum_{t=35}^{100} \binom{100}{t} p_1^t (1 - p_1)^{100-t}
\end{aligned}
$$

```
1-pbinom(34, 100, 0.4066)
```

```
## [1] 0.895889
```

How about at least 90 out of 100 last at least 45 hours?

```
1-pbinom(89, 100, 0.4066)
```

```
## [1] 0
```

Full Monte Carlo Solution

- For $j = 1, \ldots, n$:
  1. Generate $X_1, \ldots, X_{c=100} \overset{\text{iid}}{\sim} \text{exponential}(\beta = 50)$.
  2. Set $Y_j = 1$ if at least $t = 35$ X_is are $\geq h = 45$; otherwise set $Y_j = 0$.

Then, because $Y_j \sim \text{Bernoulli}(p_2)$ and $E[Y_j] = p_2$,

$$\frac{1}{n} \sum_{j=1}^{n} Y_j \to p_2 \text{ as } n \to \infty$$

```r
set.seed(1001)
n <- 10000

y <- rep(0, n) ## storage
for(i in 1:n){
  x <-rexp(100, 1/50)
  if(length(x[x>=45])>=35){
  y[i] <- 1}
  }

mean(y)
```

```
## [1] 0.8949
```

We can see that being able to generate random values from various probability distributions can be quite useful!

Generating Random Samples

- There are a number of approaches to the generation of random variables.
- Let's start by considering the simplest approach, the probability inverse transform.
- For this approach (and actually most every approach I can think of) we assume that we are able to generate:

$$\boxed{U_1, \ldots U_m \overset{\text{iid}}{\sim} \text{uniform}(0, 1)}$$

** Tutorial 0 (Probability Inverse Transform):**

- Let X have a continuous cdf $F_X(x)$.
- Define the random variable $Y = F_X(x)$.
- Then Y is uniformly distributed on (0,1). $P(Y \leq y) = y \quad 0 < y < 1$.

Proof:

$$
\begin{aligned}
P(Y \leq y) &= P(F_X(x) \leq y) \\
&= P(F_X^{-1}[F_X(x)] \leq F_X^{-1}[y]) \\
&= P(X \leq F_X^{-1}[y]) \\
&= F_X(F_X^{-1}[y]) = y
\end{aligned}
$$

Note: If F_X is flat in a region then it may be that $F_X^{-1}[F_X(x)] \neq x$

- Let $x \in [x_1, x_2] \Rightarrow F_X^{-1}[F_X(x)] = x_1$ for any x in the interval.
- However, $P(X \leq x) = P(X \leq x_1)$.
- Generally we just define $F_X^{-1}(y) = \inf\{x | F(x) \geq y\}$

- Simply: $X = F_X^{-1}(U)$ has the distribution F_X.
- Consider $X \sim \mathrm{exponential}(\beta = 2)$:

$$F_X(c) = \int_0^c \frac{1}{\beta} exp(-x/\beta) dx = 1 - exp(-c/\beta)$$

$$U = F_X(X) = 1 - exp(-X/\beta)$$

$$\begin{aligned} U &= F_X(X) = 1 - exp(-X/\beta) \\ 1 - U &= exp(-X/\beta) \\ log(1 - U) &= -X/\beta \\ -\beta log(1 - U) &= X = F_X^{-1}(U) \end{aligned}$$

```r
set.seed(1001)
u <- runif(10000, 0, 1)
x <- - 2*log(1-u)

mean(x)
```

$\beta = 2$

```
## [1] 1.989107
```

```r
var(x)
```

```
## [1] 3.915259
```

- $E[X] = \beta = 2, \; V(X) = \beta^2 = 4$

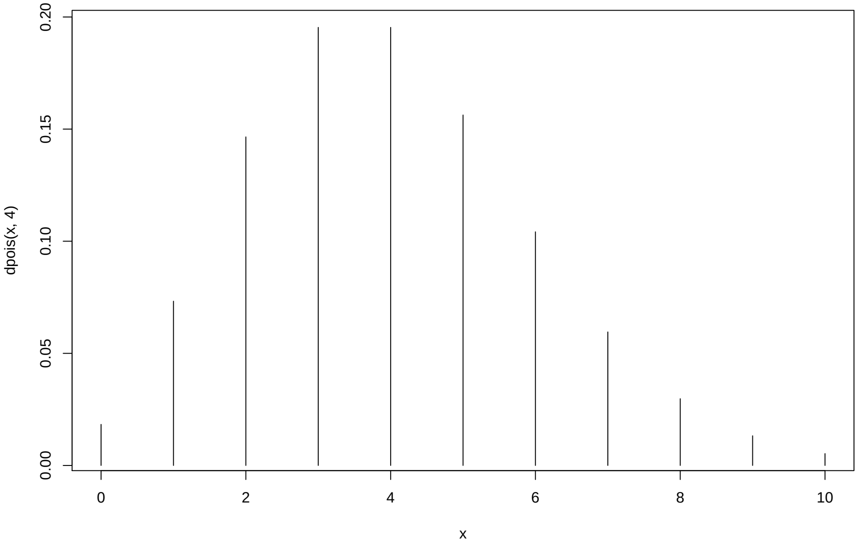
Distributions in R

- Consider $X \sim \mathrm{Poisson}(\lambda = 4)$ (density):
- $P(X = 2)$ use 'd':

```r
dpois(2, 4)
```

```
## [1] 0.1465251
```

```r
x <- 0:10
plot(x, dpois(x, 4), type="h")
```



Distributions in R

- $P(X \leq 2)$ use 'p' (probability):

```r
ppois(2, 4)
```

```
## [1] 0.2381033
```

- $P(X \leq x^*) = 0.25$, to find x^* use 'q' (quantile):

```r
qpois(0.25, 4)
```

```
## [1] 3
```

Distributions in R

- Remember the quantile must achieve the specified probability:

```r
ppois(2, 4)
```

```
## [1] 0.2381033
```

```r
ppois(3, 4)
```

```
## [1] 0.4334701
```

- So $x^* = 3$

Distributions in R

- To generate random values use 'r'. Let's generate $n = 10$ random values:

```r
rpois(10, 4)
```

```
##  [1] 2 6 7 3 1 7 1 5 4 4
```