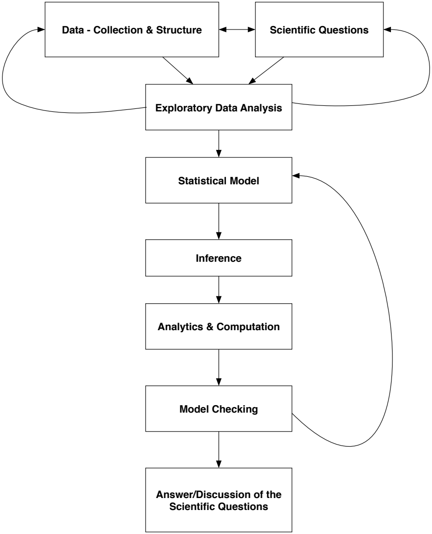
# Statistical Inference

## Lecture 01b

ANU - RSFAS

Last Updated: Tue Feb 21 16:53:35 2017

# Thoughts on Statistics & Science



# Statistical Inference

- The primary subject of statistical inference is drawing conclusions about some aspect of a population of persons or objects based on a set of quantitative observations randomly gathered from that population.
- We will be interested in estimating or testing some numerical characteristic(s) of a population.
- To formalize our task, we will be interested in probability models: a collection, or family, of related probability distributions, one of which is believed to fully characterize the population or process from which a set of observed data values arose.

# Sampling

**Def:** $X_1, \ldots, X_n$ are called a random sample of size $n$ from the population $f(x)$ if:

- $X_1, \ldots, X_n$ are mutually independent random variables and the marginal probability density function (pdf) or probability mass function (pmf) of each $X_i$ is the same function $f(x)$.
  - mutual independence: No $X$ has an effect or relationship with any other $X$s.
- We can denote this by:

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x)$$

where iid stands for independent and identically distributed.

## Sampling

- If
$$X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} f(x)$$
then the joint distribution of $X_1, \ldots, X_n$ can be written as:

$$f(x_1, \ldots, x_n) = f(x_1)f(x_2) \times \cdots \times f(x_n) = \prod_{i=1}^{n} f(x_i)$$

Note: The ‘f’s are the same functions.

# Sampling

- If the population pdf or pmf (I will likely just start to say 'density function' for both) is a member of parametric family given by $f(x|\theta)$ then if:

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$$

then:

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

# Sampling

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

The conditioning here suggests that $\theta$ is also a random variable.

- In the frequentist paradigm $\theta$ is a fixed but unknown constant. Thus it is not random and many times you will see:

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

- In the Bayesian paradigm $\theta$ is random and using the conditional bar is essential.

# Sampling

- Note: $\theta$ is THE general symbol for a parameter.
- Note: $\theta$ may be a vector. Some texts follow the convention that $\boldsymbol{\theta}$ should be in bold if it is a vector others do not and expect you understand based on the context or outlined definitions.
- Eg. $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{normal}(x|\theta = \{\mu, \sigma^2\})$
- Eg. $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{exponential}(x|\theta)$

# Sampling

- In the context outlined, the random samples are generally based on sampling from an infinite population.
- If however we are sampling from a finite population the definition may or may not hold.
  - With a finite population we have a finite set of numbers:

$$\{x_1, x_2, \ldots, x_N\}$$

  - sampling with replacement (Definition is met)
  - sampling without replacement (Definition is not met - independence is violated)

# Sampling

- Eg. Suppose we are sampling without replacement:
  1. First draw: Each of the $N$ values has $1/N$ chance of being selected. We choose and find $X_1 = x_1$. We do not allow $x_1$ to be drawn again. This is without replacement.
  2. Second draw: Now each of the $N-1$ values has a chance of $1/(N-1)$ of being selected. We choose and find $X_2 = x_2$.
  3. We continue sampling in this manner until $n$ (small $n$).
- Why is independence violated?

Note: If $A$ and $B$ are two independent events then:

$$P(A|B) = P(A)$$

Thus knowing that $B$ has occurred has no effect on the probability of $A$.

# Sampling

Let $z$ and $y$ be elements in $\{x_1, \ldots, x_N\}$.

$$P(X_2 = y | X_1 = y) = 0$$

$$P(X_2 = z | X_1 = y) = 1/(N-1)$$

The probability distribution for $X_2$ depends on $X_1$! Thus $X_1$ affects $X_2$.

- OK, let's compare this to the marginal distribution for $X_2$ to make the point more clear:

$$\begin{aligned} P(X_2 = z) &= P(X_2 = z | X_1 = x_1) P(X_1 = x_1) \\ &\quad + P(X_2 = z | X_1 = x_2) P(X_1 = x_2) \\ &\quad + \cdots + P(X_2 = z | X_1 = x_N) P(X_1 = x_N) \end{aligned}$$

## Sampling

- For one of these, say the second, $z = x_2$ then $P(X_2 = z|X_1 = x_2) = P(X_2 = z|X_1 = z) = 0$.

$$\begin{aligned} P(X_2 = z) &= P(X_2 = z|X_1 = x_1)P(X_1 = x_1) + 0 \times P(X_1 = x_2) \\ &\quad + \cdots + P(X_2 = z|X_1 = x_N)P(X_1 = x_N) \\ &= \left(\frac{1}{N-1}\right)\left(\frac{1}{N}\right) + 0 + \cdots + \left(\frac{1}{N-1}\right)\left(\frac{1}{N}\right) \\ &= (N-1)\left(\frac{1}{N-1}\right)\left(\frac{1}{N}\right) = \frac{1}{N} \end{aligned}$$

So $P(X_2 = z) \neq P(X_2 = z|X_1)$.

- If $N$ is large compared to $n$ then we say that $X_1, \ldots, X_n$ are nearly independent and we use approximate probability calculations based on independence.

# Functions of Samples - Statistics

**Def:**

- Let $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} f(x)$
- Let $T(x_1, \ldots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, \ldots, X_n)$.
- Then the random variable $Y = T(X_1, \ldots, X_n)$ is called a statistic. The probability distribution of $Y$ is called the sampling distribution of $Y$.

# Properties of Statistics (mean and variance)

**Lemma:** Let $X_1, \ldots, X_n$ be a random sample from a population and let $g(x)$ be a function such that $E[g(X_1)]$ and $Var[g(X_1)]$ exist, then

$$E\left(\sum_{i=1}^{n} g(X_i)\right) = E\left(g(X_1) + \cdots + g(X_n)\right)$$

$$E[g(X_1)] + \cdots + E[g(X_n)] = nE[g(X_1)]$$

- Based on direct application of the variance operator

$$\begin{aligned} V\left(\sum_{i=1}^{n} g(X_i)\right) &= V\left(g(X_1) + \cdots + g(X_n)\right) \\ &= V[g(X_1)] + \cdots + V[g(X_n)] \\ &+ 2\sum \sum_{1 \le i < j \le n} Cov(g(X_i), g(X_j)) \\ &= V[g(X_1)] + \cdots + V[g(X_n)] + 0 = nV[g(X_1)] \end{aligned}$$

# Properties of Statistics (mean and variance)

- Based on the definition of the variance
- Note: $V(X_1) = E(X_1 - E(X_1))^2$

$$
\begin{aligned}
V\left(\sum_{i=1}^{n} g(X_i)\right) &= E\left(\sum_{i=1}^{n} g(X_i) - E\left(\sum_{i=1}^{n} g(X_i)\right)\right)^2 \\
&= E\left(\sum_{i=1}^{n} g(X_i) - \sum_{i=1}^{n} E\left(g(X_i)\right)\right)^2 \\
&= E\left[\sum_{i=1}^{n} \left[g(X_i) - E\left(g(X_i)\right)\right]\right]^2
\end{aligned}
$$

- The first $n$ terms are of the form $E\left[g(X_i) - E(g(X_i))\right]^2 = V[g(X_i)]$

- The next $n(n-1)$ terms are of the form:

$$E\left[(g(X_i) - E(g(X_i)))\,(g(X_j) - E(g(X_j)))\right] = Cov(g(X_i), g(X_j))$$
$$= 0$$

$$E\left[\sum_{i=1}^{n}[g(X_i) - E(g(X_i))]\right]^2 = \sum_{i=1}^{n} E[g(X_i) - E(g(X_i))]^2$$
$$= \sum_{i=1}^{n} V(g(X_i)) = nV(g(X_1))$$

# Some Famous Summary Statistics

- Sample mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Sample variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- Sample standard deviation: $S = \sqrt{S^2}$
- Why do we like the sample mean? It minimizes the square distance between it and each data point:

$$min_a \sum_{i=1}^{n}(x_i - a)^2 \;\Rightarrow \hat{a} = \bar{x}$$

# Porperties of Famous Summary Statistics

**Theorem:** Let $X_1, \ldots, X_n$ be a random sample from a population with mean $\mu$ and variance $\sigma^2 < \infty$. Then

- $E[\bar{X}] = \mu$
- $V[\bar{X}] = \sigma^2/n$
- $E[S^2] = \sigma^2$

**Proof** of the last two:

$$
\begin{aligned}
V[\bar{X}] &= V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2} V\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\left(V(X_1) + V(X_2) + \cdots + V(X_n) + 2\sum\sum_{1\le i<j\le n} Cov(X_i, X_j)\right) \\
&= \frac{1}{n^2}\left(V(X_1) + V(X_2) + \cdots + V(X_n) + 0\right) = \sigma^2/n
\end{aligned}
$$

$$
\begin{aligned}
E[S^2] &= E\left(\frac{1}{n-1}\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right]\right) \\
&= E\left(\frac{1}{n-1}\left[\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right]\right) \quad = \frac{1}{n-1}\left(nE(X_i^2)-nE(\bar{X}^2)\right)
\end{aligned}
$$

Note:

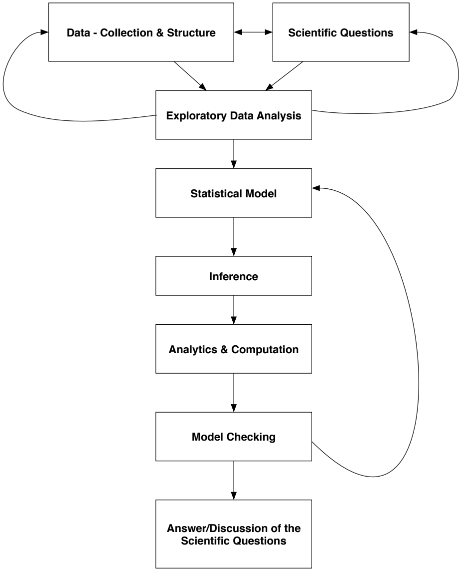
$$
\begin{aligned}
V(X) &= E(X-E(X))^2 \\
&= E(X^2-2XE(X)+E(X)^2) \\
&= E(X^2)-2E(X)E(X)+E(X)^2 \\
&= E(X^2)-2E(X)^2+E(X)^2 \\
&= E(X^2)-E(X)^2
\end{aligned}
$$

$$E(X^2) = V(X) + E(X)^2 = \sigma^2 + \mu^2$$

$$E(\bar{X}^2) = V(\bar{X}) + E(\bar{X})^2 = \sigma^2/n + \mu^2$$

$$E[S^2] = \frac{1}{n-1}\Big(n(\sigma^2 + \mu^2) - n(\sigma^2/n + \mu^2)\Big) = \sigma^2$$

# 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)
```

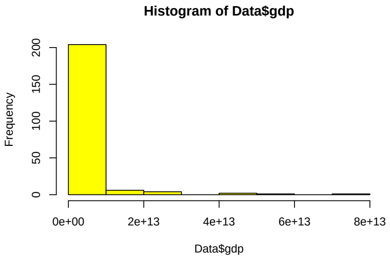
```
## [1] "Country.Name" "Country.Code" "gdp"          "labor"
```

# EDA

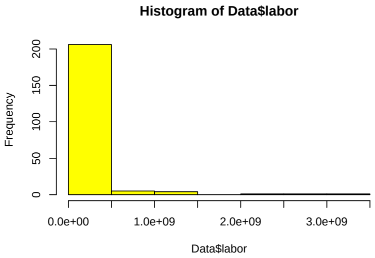
```r
par(mfrow=c(2,2))
hist(Data$gdp, col="yellow")
hist(Data$labor, col="yellow")
plot(Data$labor, Data$gdp)
```

```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))
```

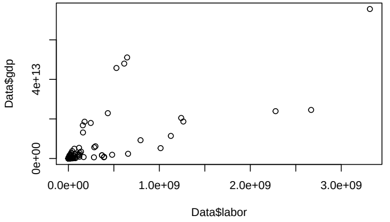
# EDA



Histogram of Data$gdp
Frequency
0
50
100
150
200
0e+00
2e+13
4e+13
6e+13
8e+13
Data$gdp

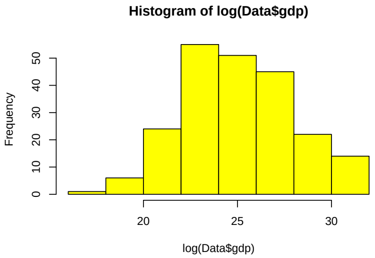



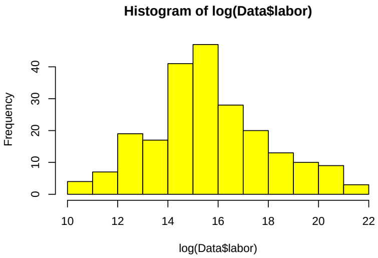
Histogram of Data$labor
Frequency
0
50
100
150
200
0.0e+00
1.0e+09
2.0e+09
3.0e+09
Data$labor

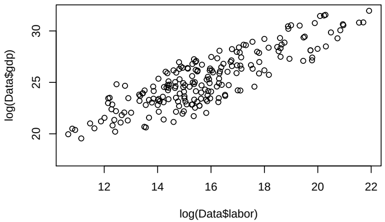



Data$gdp
0e+00
4e+13
0.0e+00
1.0e+09
2.0e+09
3.0e+09
Data$labor

# EDA


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


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


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

# Statistical Model

- Simple linear regression model:

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

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

$$Y = AL^{\beta}K^{\alpha}$$
$$log(Y) = log(A) + \beta log(L) + \alpha log(K)$$

## Estimation of the Parameters and Computation

- $\theta = \{\beta_0, \beta_1, \sigma^2\}$
- Many ways to proceed for inference. In regression class you learned about least-squares estimation but we can also consider maximum likelihood, Bayesian, . . . .
- 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!!

$$min_{\beta_0,\beta_1} \sum_{i=1}^{n} (\log(\mathrm{GDP}) - [\beta_0 + \beta_1 \log(\mathrm{labor})])^2$$

- 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:
##             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 ***
## ---
## 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?