# Bush 631-600: Quantitative Methods

## Lecture 4 (09.20.2022): Measurement vol. I

Rotem Dvir

The Bush school of Government and Public Policy

Texas A&M University

Fall 2022

# What is today's plan?

- R Markdown
- From concepts to measures.
- Why measurement? and its challenges.
- Visualizing data: plots.
- Methods: Surveys.
- R work: summary(), NAs, barplot, histogram, boxplot.

# Working with R Markdown - Class Task

Data (BAAD v.2): 140 insurgent groups (1998-2012).

- Upload data - STATA file!!
- Use index or $ method to display:
  1. rows (155,215,235,411) and columns (group name, home base, year, age, police/military fatalities).
- Create subsets for cross-tabs and diff-in-means:
  1. 2 subsets: group operate in Iraq / Somalia
  2. Tab group size and stick strategy for Iraqi groups
  3. Diff-in-means of total fatalities b-w Iraq and Somalia
- Use prop.table() for proportions of groups with religious ideology
- Calculate the mean and median of battle deaths

# Your research interests

Topics in INTA:

- Diplomacy.
- Trade.
- Conflict / wars.
- Terrorism.
- Development.
- Foreign Aid.

# Measurement

# Measurement

Why?

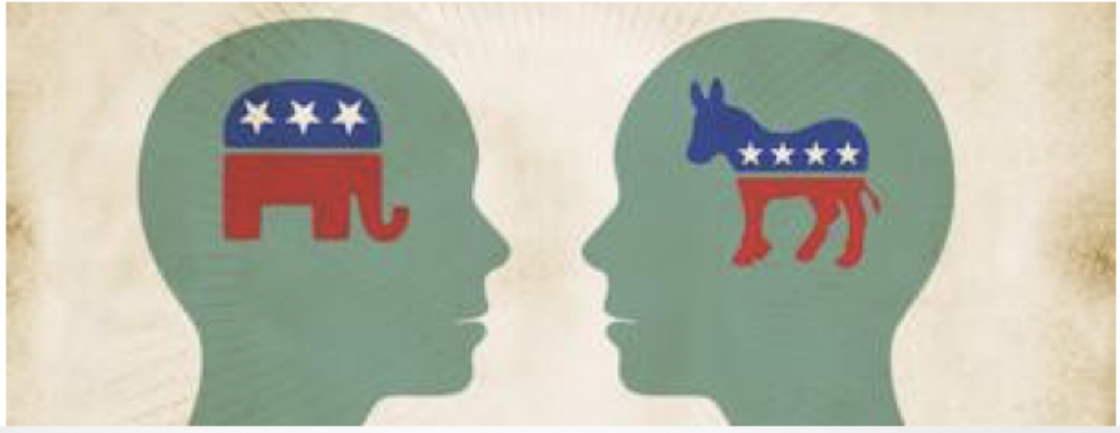
- Social science: develop and test causal theories.
- Leader background and conflict behavior.
- Minimum wage and levels of full-time employment?
- Concepts: level of unemployment, leader background, public approval.

How?

**Measures - the context of theoretical concepts**

# Measures

- Some are easy $\rightarrow$ age, place of residence.
- Others?



On a scale from 1 to 7, where 1 is extremely liberal, 7 is extremely conservative, and 4 is exactly in the middle, where would you place yourself?

| Extremely liberal 1 | 2 | 3 | In the middle 4 | 5 | 6 | Extremely conservative 7 |
|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

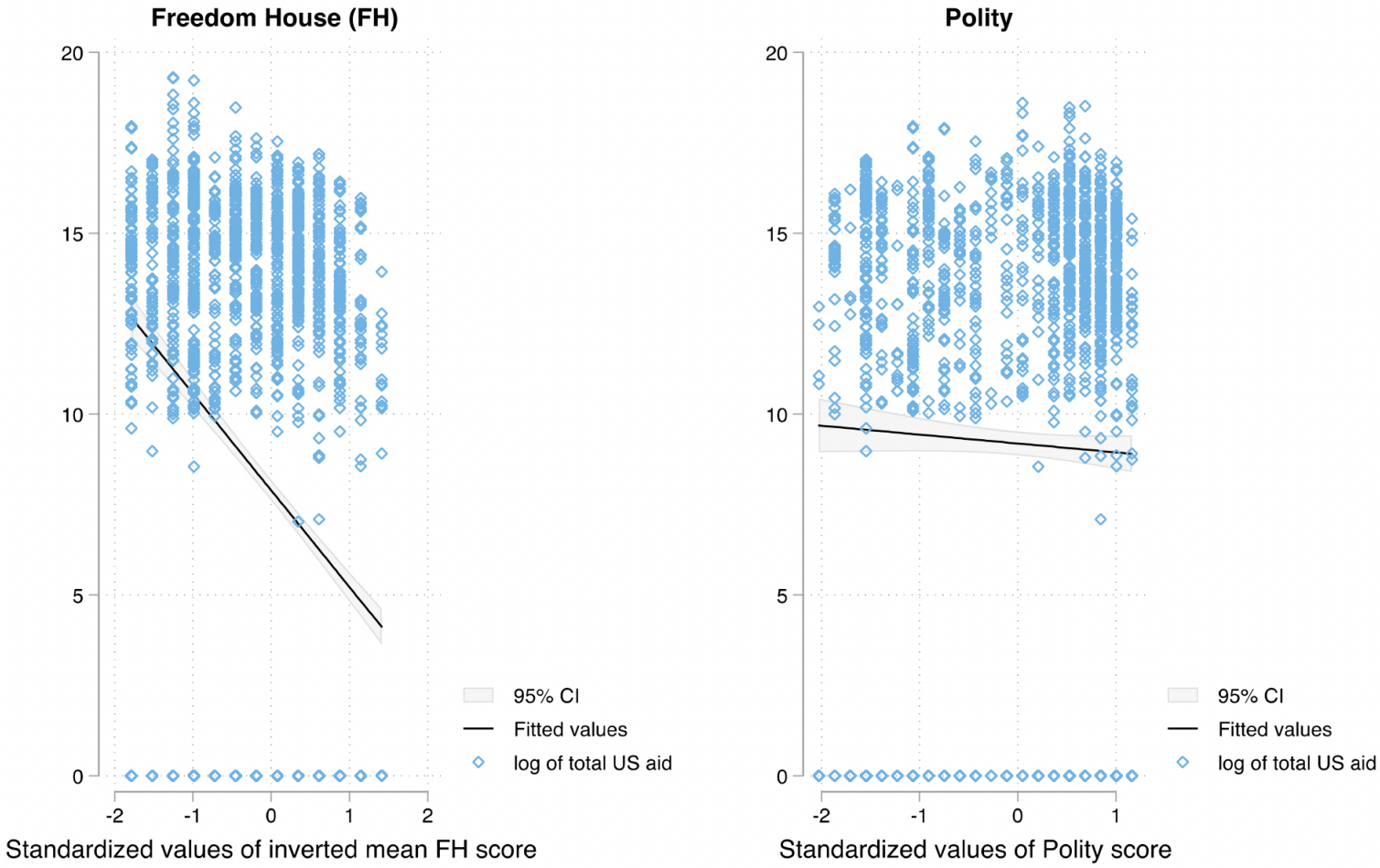
# Measuring democracy

- How do we measure 'levels' of democratic regimes?

## Two scales


# Measuring Democracies

- Does aid helps democracy promotion?



# Measuring regime types

Why the differences?

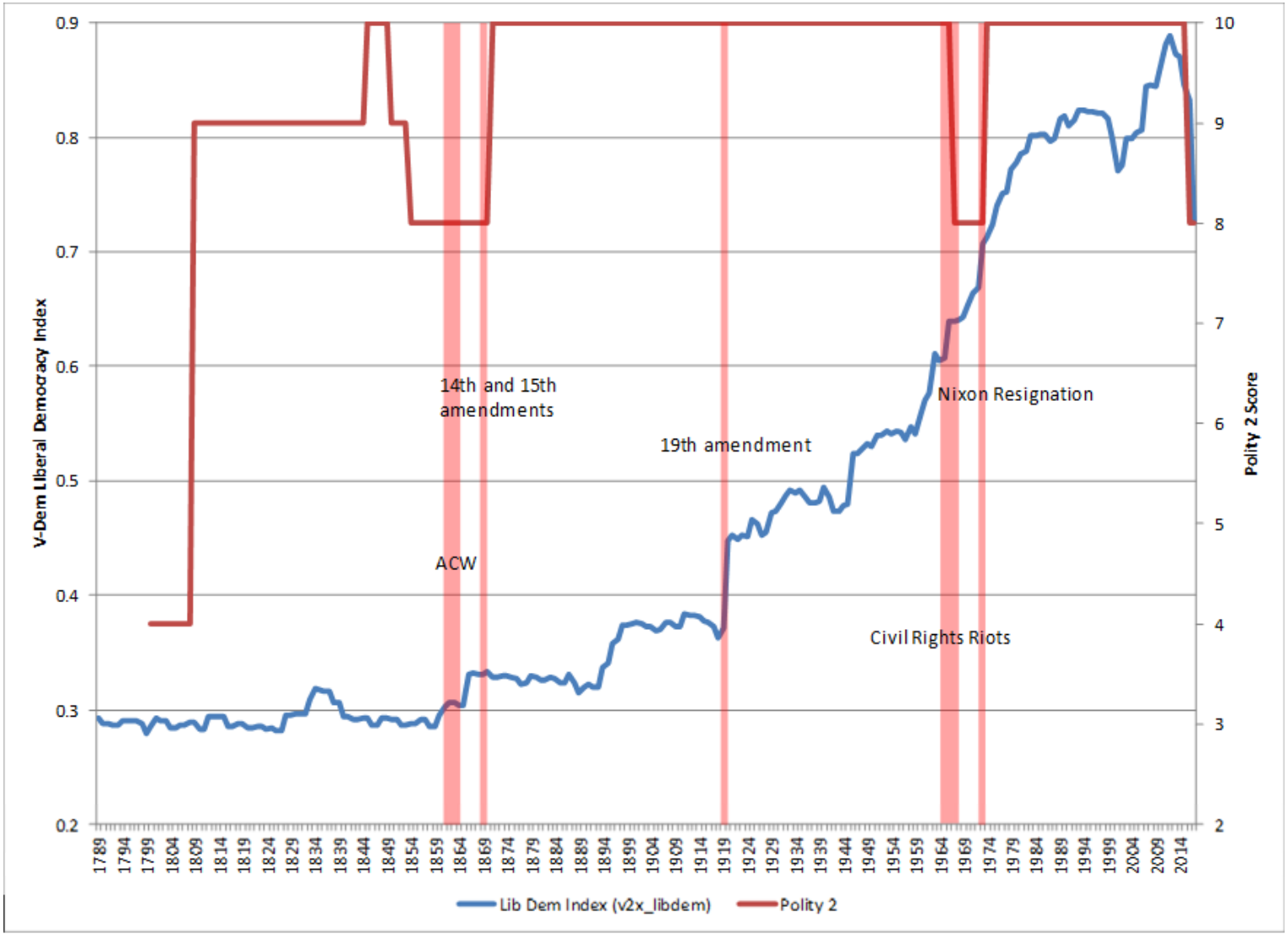
**Freedom House Scale** (Link) - personal and civil rights:

1. Political pluralism.
2. Electoral process and function of government.
3. Personal autonomy and individual rights.
4. Organizational rights.
5. Rule of law.

**Polity V Scale** (Link) - institutional features:

1. Openness and competitiveness of elections.
2. Executive constraints.
3. Regulation of participation.

# Polity V Scale: USA



# Polity V Scale

Problematic measurement:

- US & its allies
- Adversaries like Russia.
- Dynamic but inconsistent (Colgan 2019-Link).

> For one period, 1997–2003, Iran's Polity score jumped massively, by nine points. What accounts for this change? It coincides with the presidency of Khatami, a pro-Western reformer. Khatami tried to befriend the United States and reorient Iranian foreign policy. He also campaigned to make the government more accountable to the people. He did not, however, change or even seek to change the constitution or any of the key institutions or processes of the regime, saying, "there will not be a democratic regime in the true sense of the word."[9] Moreover, even his limited reform efforts failed.

# Measures & Definitions

**Operational definition**: the way we describe (represent) the relevant concept in the data (indicators/variables used).

Example: *US president approval*

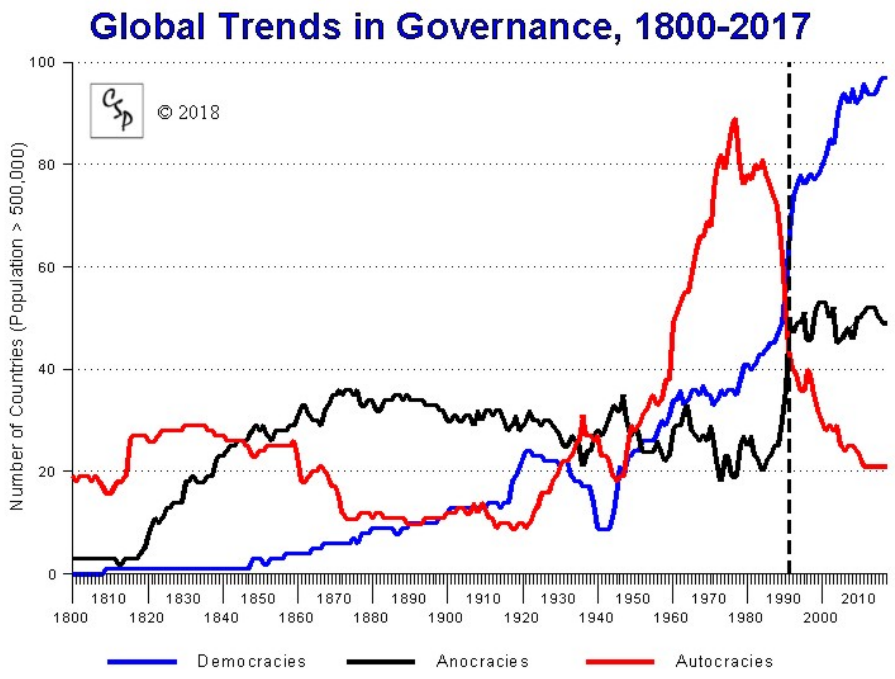
- Conceptual definition: the extent to which US adults support the actions and policies of the current US president.
- Operational definition: "On a scale from 1 to 5, where 1 is least supportive and 5 is more supportive, how much would you say you support the job that Joe Biden is doing as president?" (survey/poll item).

# Measurement Errors

The chance that there is some variation in the measures we use for our concepts.

Sources of errors:

- Data entry or respondent errors.
- Systematic Bias: US, Russia, Iran 'fluctuations'.



# Variables

- Measures of concepts.
- Many values $\rightarrow$ variables.
- Rebel background: 0,1.
- GDP/Cap; MID involvement over 5 years.
- Internationalism categories.

# Types of variables

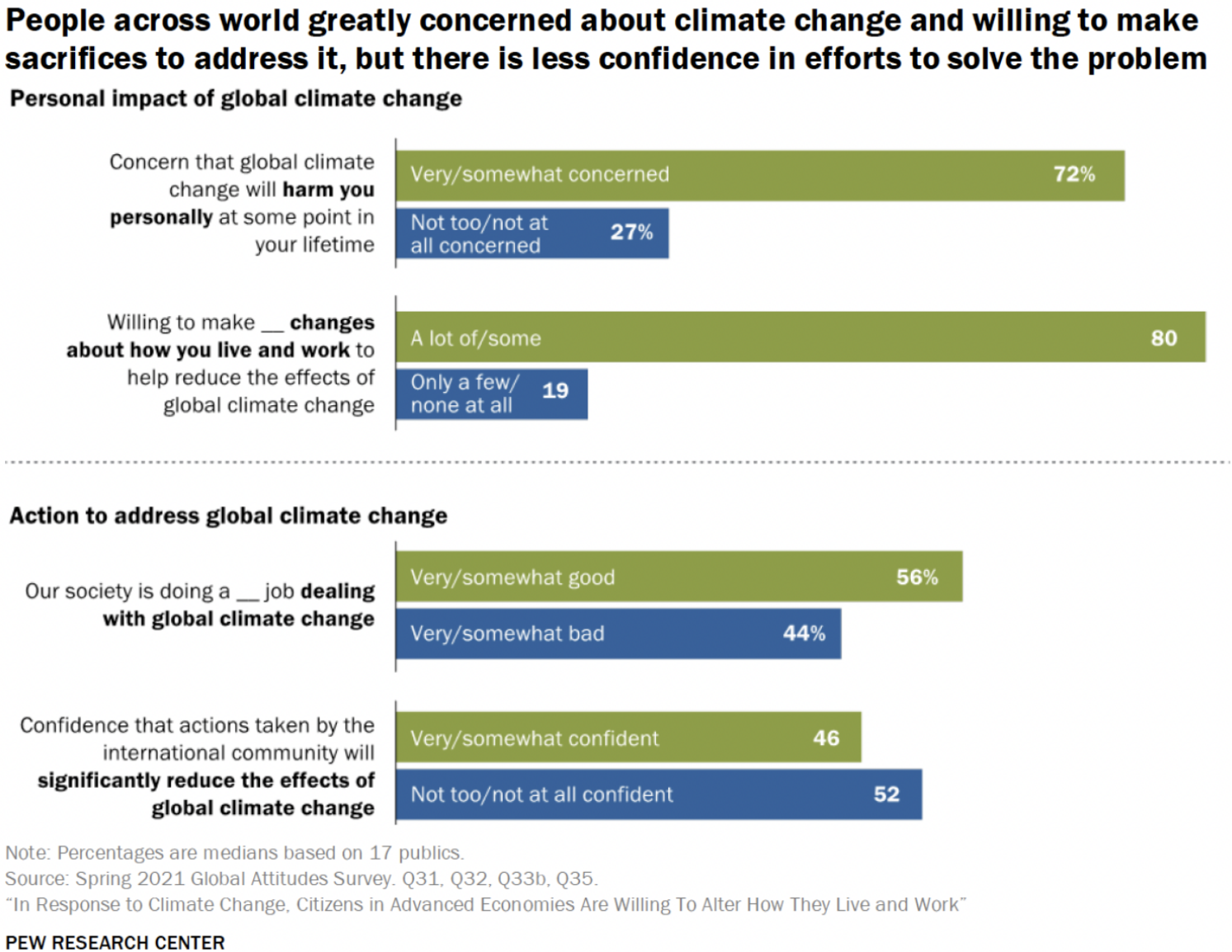
- Dependent variable (Y):
  - The explained outcome.
  - Response variable.
- Independent variable (X):
  - Determinant of the DV.
  - Explanatory / predictor variables.
  - Indicator for our main theory (explanation).
- Control variables (Z):
  - Additional predictors.
  - Pre-treatment variables.
  - Confounders.

# Fit the variables

- Oil, democracy and development (2001)

| Name | Description |
|---|---|
| `cty_name` | Country name |
| `year` | Year |
| `logGDPcp` | Logged GDP per capita |
| `regime` | A measure of a country's level of democracy: -10 (authoritarian) to 10 (democratic) |
| `oil` | Amount of oil exports as a percentage of the country's GDP |
| `metal` | Amount of non-fuel mineral exports as a percentage of the country's GDP |
| `illit` | Percentage of the population that is illiterate |
| `life` | Life expectancy in the country |

# Measurement Tools: Surveys



# Terrorism: Public Survey (2016)

- ISTPP project: national security.
- Multiple attitude measures: concern, likelihood.
- Compare types of terrorism: cyber, conventional.

| | CaseID | concern_bomb | concern_cyber | severity_bomb | severity_cyber | publicKnow_bomb | publicKnow_cyber | expertKnow_bomb | expertKnow_cyber | cas_bomb | cas_cyber |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 2 | 1 | 4 | 1 | 2 | 2 | 2 | 2 | 4 | 2 |
| 2 | 3 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 3 |
| 3 | 4 | 2 | 2 | 4 | 4 | 1 | 1 | 1 | 2 | 4 | 5 |
| 4 | 5 | 3 | 4 | 3 | 4 | 3 | 2 | 2 | 2 | 3 | 5 |
| 5 | 6 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 4 |
| 6 | 7 | 3 | 1 | 3 | 2 | 1 | 1 | 1 | 2 | 5 | 3 |
| 7 | 8 | 2 | 2 | 3 | 3 | 2 | 1 | 2 | 2 | 2 | 2 |
| 8 | 9 | 1 | 1 | 1 | 2 | 2 | 2 | 3 | 3 | 2 | 1 |
| 9 | 10 | 4 | 4 | 4 | 3 | 4 | 2 | 4 | 3 | 5 | 5 |
| 10 | 11 | 3 | 2 | 4 | 3 | 4 | 2 | 4 | 4 | 5 | 1 |
| 11 | 12 | 4 | 2 | 3 | 3 | 3 | 1 | 3 | 3 | 4 | 5 |
| 12 | 13 | 2 | 2 | 4 | 3 | 2 | 1 | 2 | 2 | 4 | 5 |
| 13 | 14 | 3 | 2 | 4 | 3 | 3 | 2 | 3 | 3 | 3 | 5 |
| 14 | 15 | 2 | 1 | 4 | 2 | 3 | 3 | 3 | 2 | 4 | 4 |

# Terrorism Survey

```
# Proportions: concerns about types of terrorism
prop.table(table(conventional = mydata$concern_bomb,
                 cyber = mydata$concern_cyber))
```

```
##              cyber
## conventional           1           2           3           4
##            1 0.100177830 0.019561352 0.003556609 0.000000000
##            2 0.065797273 0.296976882 0.074096028 0.007705987
##            3 0.012448133 0.103141672 0.148784825 0.034380557
##            4 0.002963841 0.010077060 0.036751630 0.083580320
```

# Terrorism Survey

- Individual characteristics, policy preferences.

```
# Proportions: damages from attack and respondent gender
prop.table(table(Lethality = mydata$severity_bomb,
                 Gender = mydata$PPGENDER))
```

```
##           Gender
## Lethality          0          1
##         1 0.01837582 0.01422644
##         2 0.08891523 0.05690575
##         3 0.17783047 0.18612922
##         4 0.19383521 0.26378186
```

```
# Proportions: Likelihood of attach and airport
prop.table(table(Attack_Coming = mydata$likely_bomb,
                 Airport_Checks = mydata$Pol_screenUS))
```

```
##                Airport_Checks
## Attack_Coming           1           2           3           4           5
##             1 0.016806723 0.014405762 0.045018007 0.024609844 0.020408163
##             2 0.023409364 0.054621849 0.111644658 0.105642257 0.081032413
##             3 0.014405762 0.036014406 0.086434574 0.124249700 0.091836735
##             4 0.008403361 0.007202881 0.022809124 0.036014406 0.075030012
```

# Missing data: Non-response

- Why NAs?
- Item nonresponse: Refusal to answer.
- Examples: income, national origin, religion.

- Misreporting: not true attitude.
- *Social desirability bias*.
- Problematic issues: racial prejudice, corruption, etc.

- No access to data / no data: unemployment in developing countries.
- Don't know the answer / no opinion...

# Missing data: Non-response

## NAs in our survey data

```
# Responses to item: likelihood of attack (observations 1-15)
head(mydata$likely_bomb, n = 15)
```

```
##  [1]  2  3  2  3  2  4  2  1  4  3 NA  2  4  3  3
```

```
# Responses to item: support using force (observations 1-15)
# Using logical values
head(is.na(mydata$Pol_force), n=10)
```

```
##  [1] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

# NAs in our data

- Aggregate view of missing values in data.
- Syntax: function(is.na(data$variable))

```
# Sum of NAs per variable/item
sum(is.na(mydata$likely_bomb))
```

```
## [1] 53
```

```
# Proportion of NAs per variable/item
mean(is.na(mydata$likely_bomb))
```

```
## [1] 0.03063584
```

# NAs in our data

- Proportions of NA across variables

```
prop.table(table(Attack_Coming = mydata$likely_bomb,
                 Airport_Checks = mydata$Pol_screenUS, exclude = NULL))
```

```
##                Airport_Checks
## Attack_Coming             1             2             3             4
##             1    0.0161849711 0.0138728324 0.0433526012 0.0236994220 0.01965317
##             2    0.0225433526 0.0526011561 0.1075144509 0.1017341040 0.07803468
##             3    0.0138728324 0.0346820809 0.0832369942 0.1196531792 0.08843930
##             4    0.0080924855 0.0069364162 0.0219653179 0.0346820809 0.07225433
##             <NA> 0.0005780347 0.0005780347 0.0040462428 0.0011560694 0.00173410
##                Airport_Checks
## Attack_Coming          <NA>
##             1    0.0034682081
##             2    0.0017341040
##             3    0.0005780347
##             4    0.0005780347
##             <NA> 0.0225433526
```

# Study surveys with NAs

- NAs interfere with our analysis
- Return NA value.
- Must be accounted for in selected function.

```
# Calculate mean of variable with NAs: return NA
mean(mydata$Pol_survMusl)
```

```
## [1] NA
```

```
# Calculate mean of variable with NAs: accounting for missing
mean(mydata$Pol_survMusl, na.rm = TRUE)
```

```
## [1] 2.067584
```

# Study surveys with NAs

Removing missing values

- *Listwise deletion*: remove all observation with at-least one NA.
- May substantially reduce the data.

```
# Losing observations: full dataset
nrow(mydata)
```

```
## [1] 1730
```

```
mydata.del <- na.omit(mydata)
nrow(mydata.del)
```

```
## [1] 1519
```

```
# Losing observations: single variable
length(mydata$concern_bomb)
```

```
## [1] 1730
```

```
length(na.omit(mydata$concern_bomb))
```

```
## [1] 1690
```

# Visual display of data

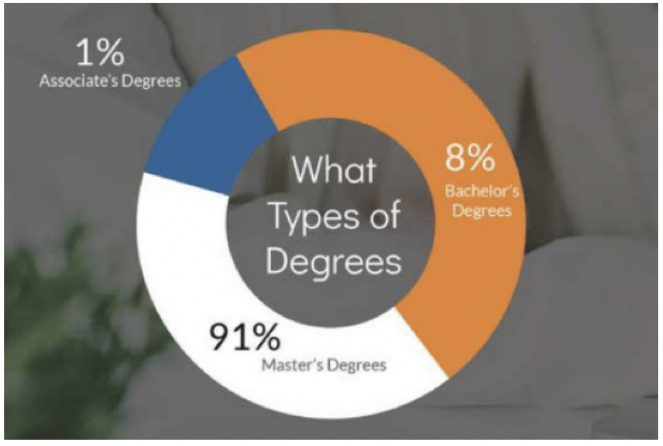
Numerical display of data:

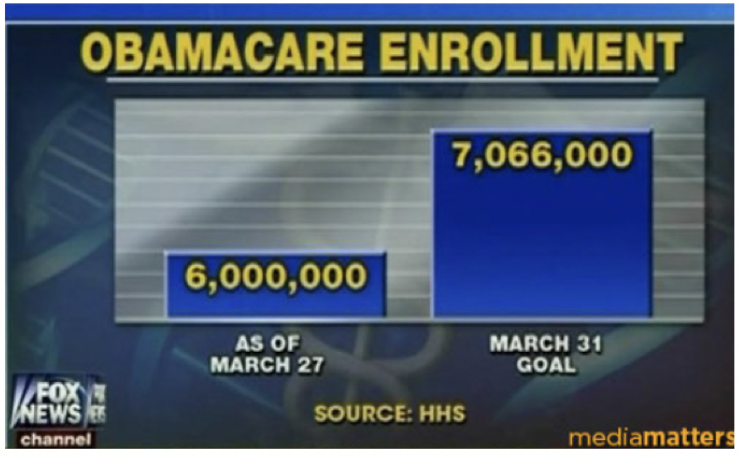
- Summaries: mean, median.
- Specific values: max, min.
- Distributions: range, SD.

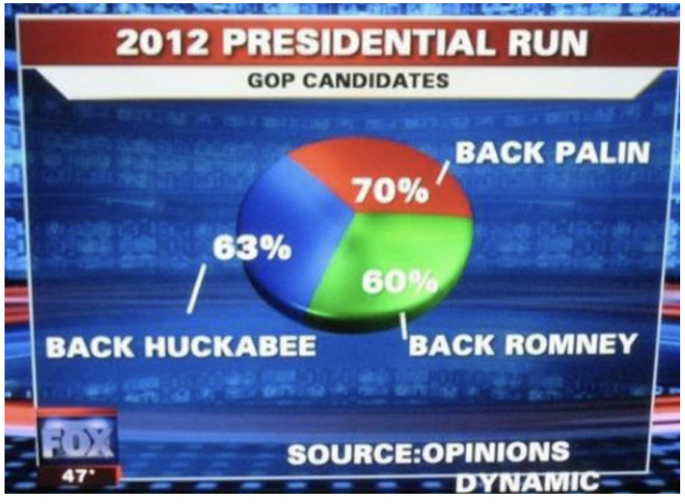
Visuals: plots, graphs

- More comprehensive.
- Highlight important elements.
- Great for presentation.
- Audience focus on important insights

# Visuals: please don't. . .

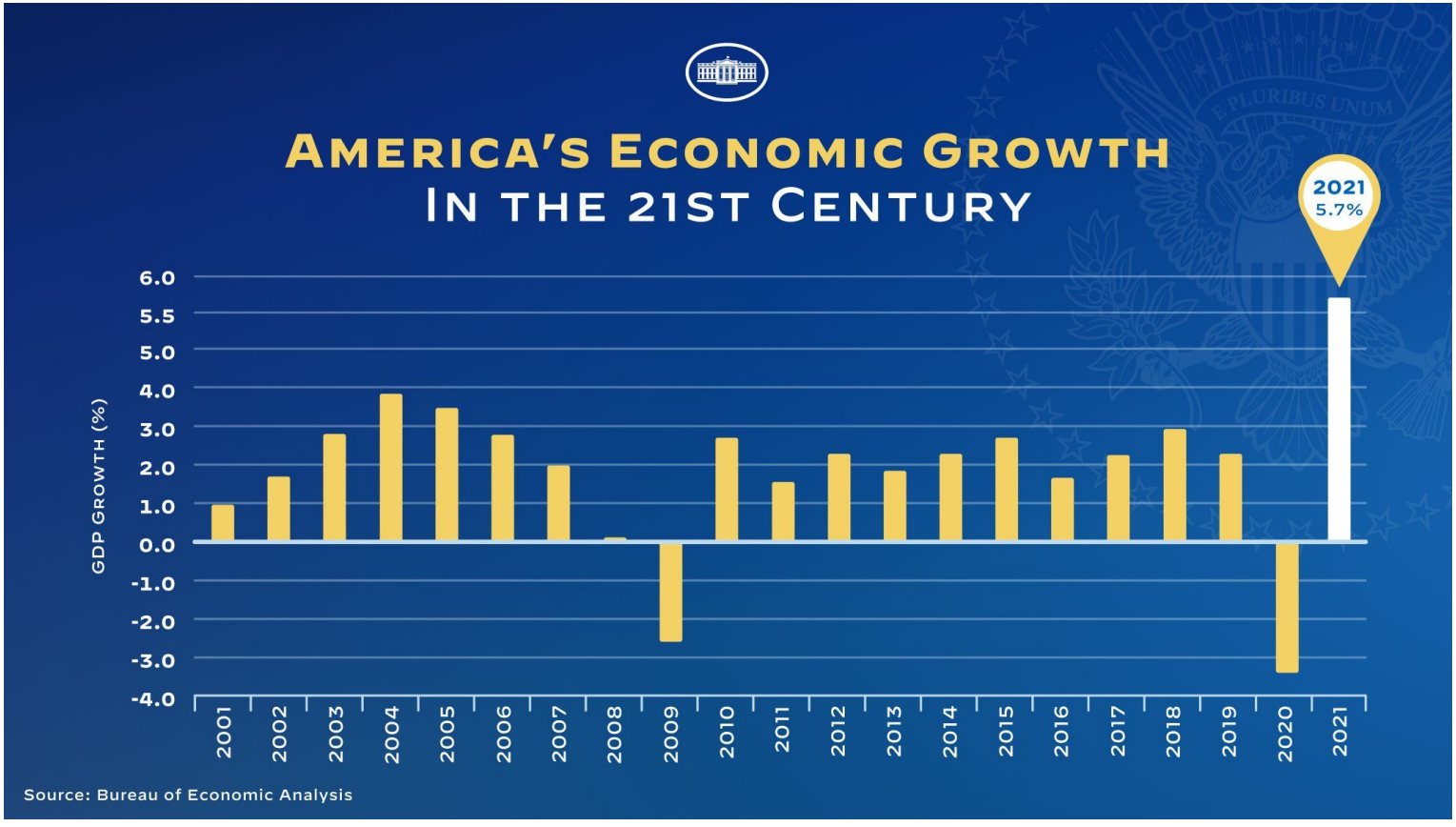



# Visuals

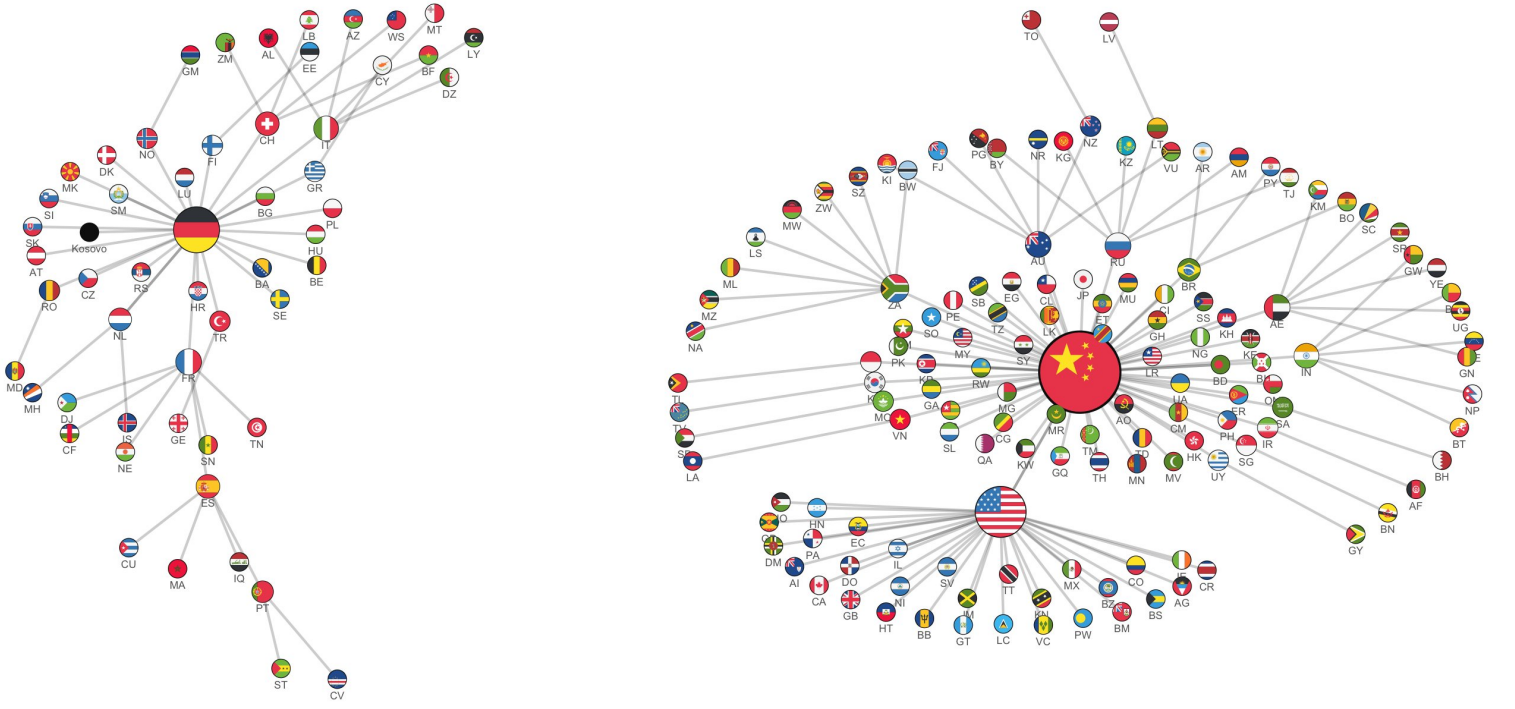
This is an official government product. . .



# Visuals: Much better

## Countries connected to their primary trading partner in 2020

Exports + imports. Data: International Monetary Fund. Flags were not available for countries in black.




@sundellviz

# Visuals

## Bar plot

- Useful for factor variables
- Shows counts and proportion for multiple categories
- How many men/women?
- Proportion of college graduates in our data?

# Visuals: INTA study

# Ethics in combat

**Sagan and Valentino (2018)**: public attitudes and ethics of war.

- Survey experiments.
- Combat scenarios $\rightarrow$ treatments.
- Support for action.

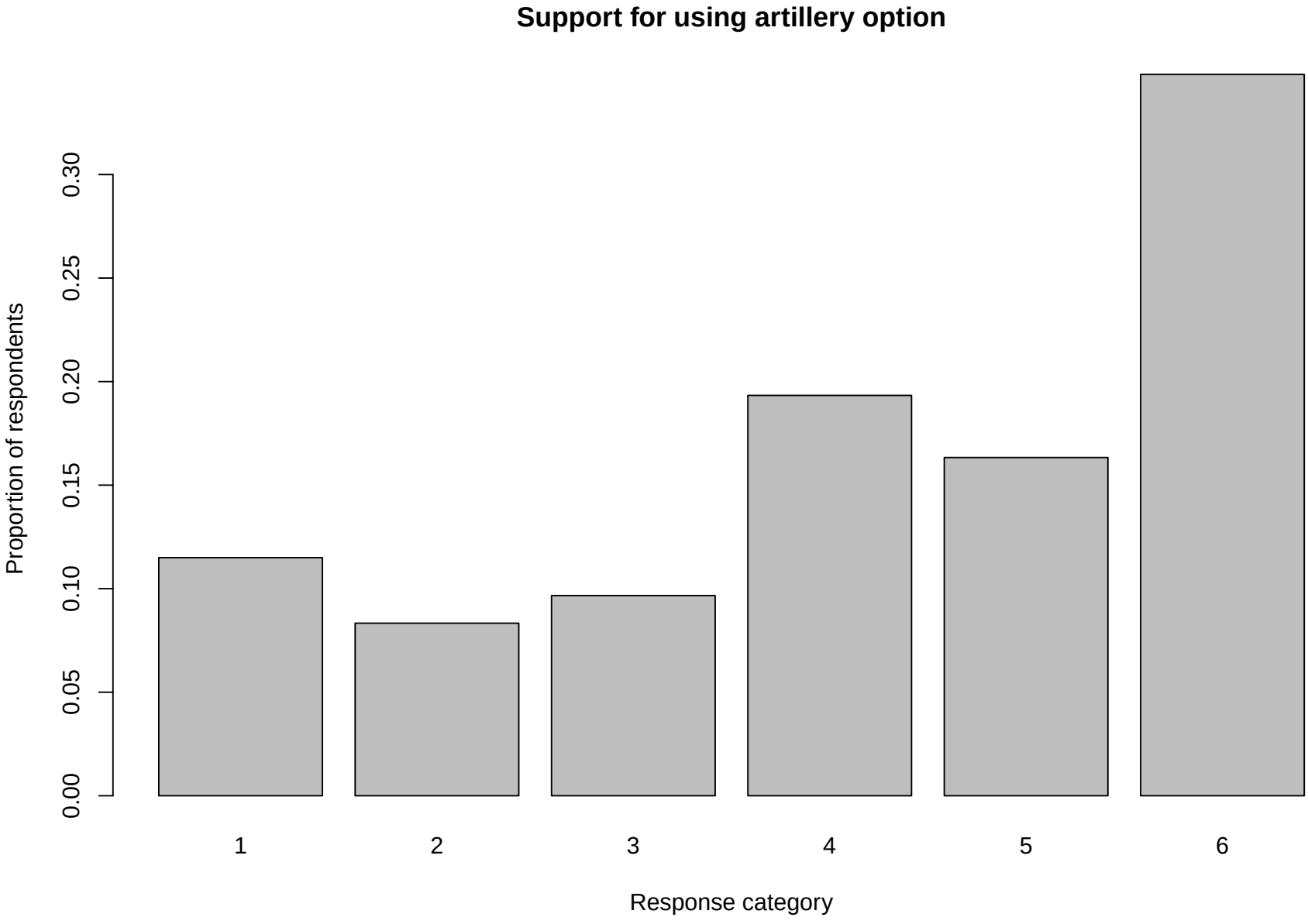
# Ethics in combat

*The virtue of proportional response*


**Iraq War (2003)**:

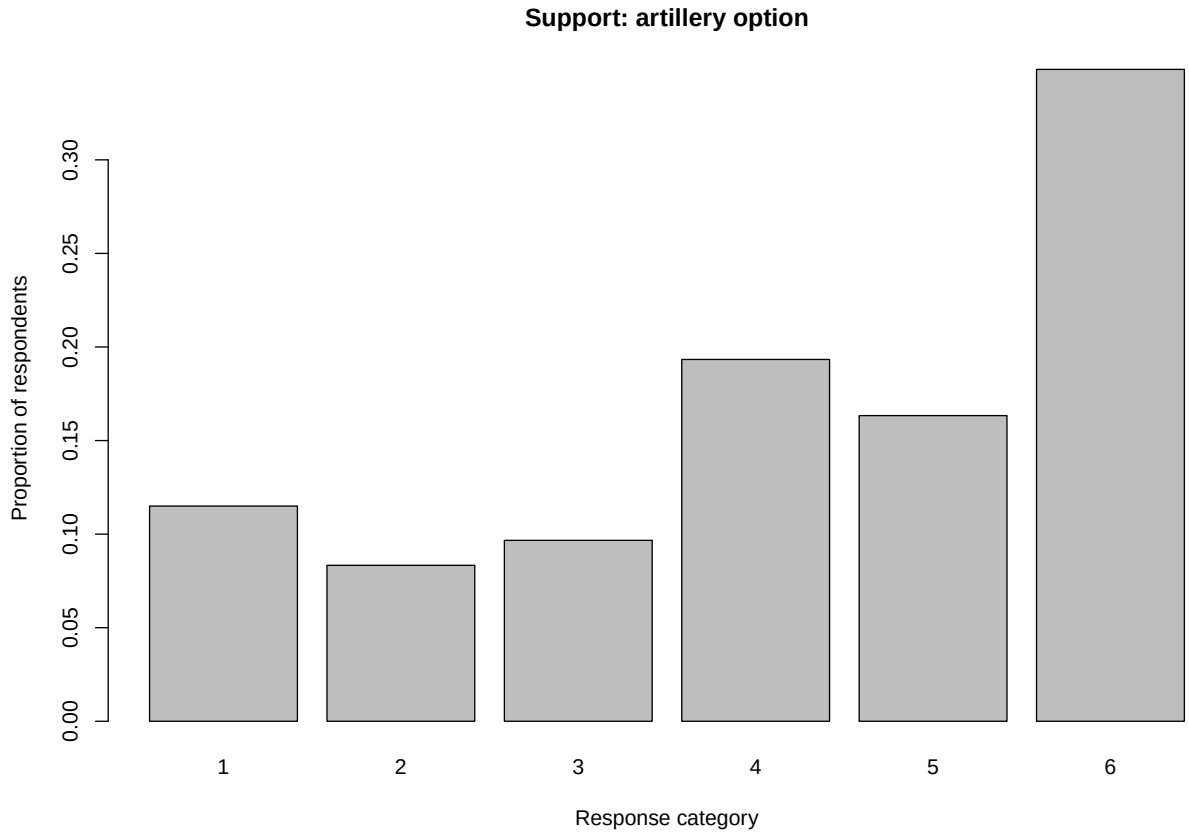
- Threshold for collateral Iraqi noncombatant deaths.
- Define "high" versus "low" value targets.
- "Due-care" principle (war in Afghanistan).

# Just war: Public attitudes



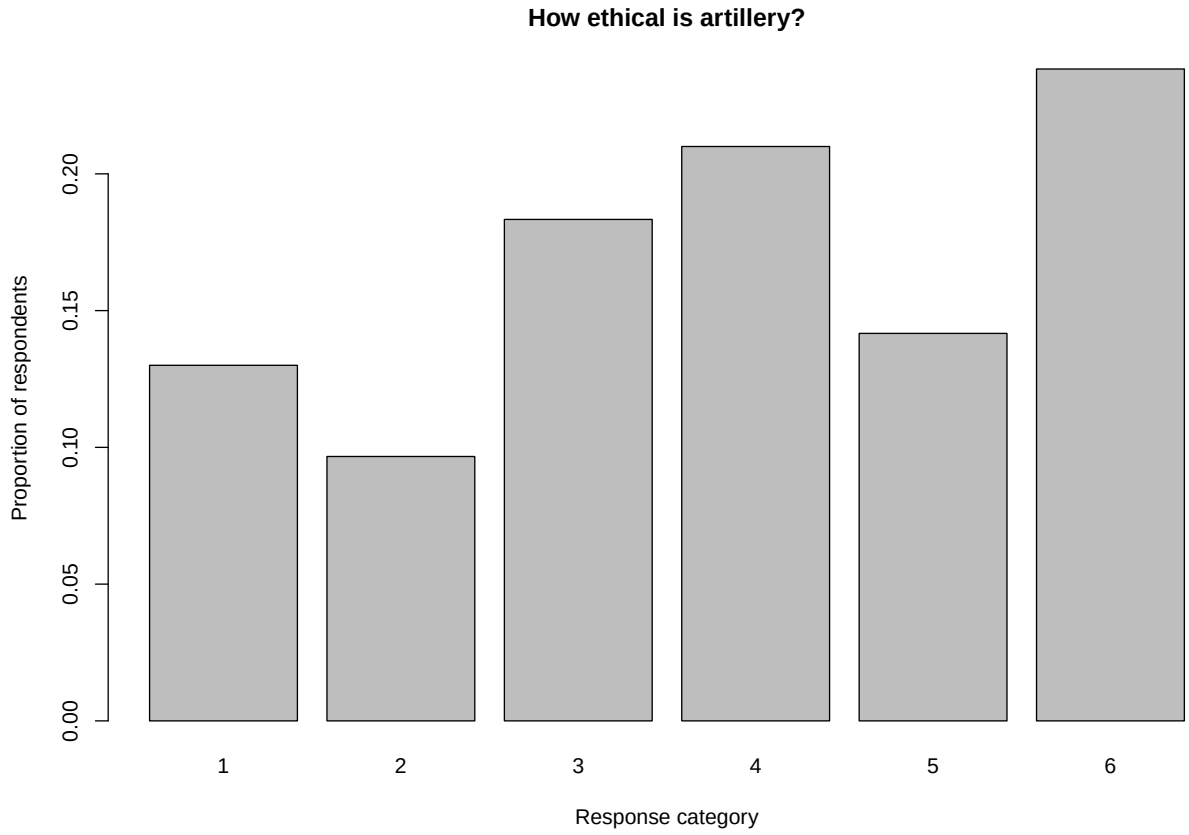
# Bar plot: Base R code

```
# Create proportions of support
artillery.tab <- prop.table(table(Support = wardata$artillery_approve,
                                  exclude = NULL))
# Create barplot
barplot(artillery.tab, main = "Support: artillery option",
        xlab = "Response category", ylab = "Proportion of respondents")
```



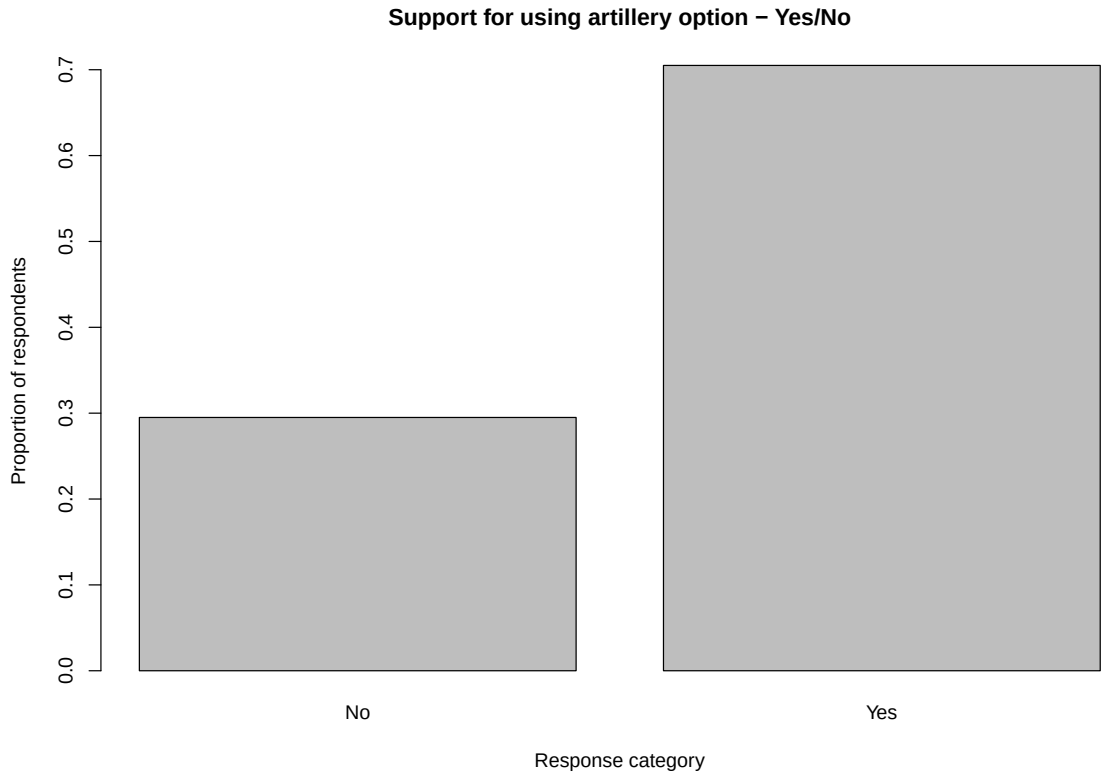
# Artillery option: Ethical?

```
artillery.ethic <- prop.table(table(Support = wardata$artillery_ethical,
                                    exclude = NULL))
barplot(artillery.ethic, main = "How ethical is artillery?",
        xlab = "Response category", ylab = "Proportion of respondents")
```



# Survey responses: Binary measure

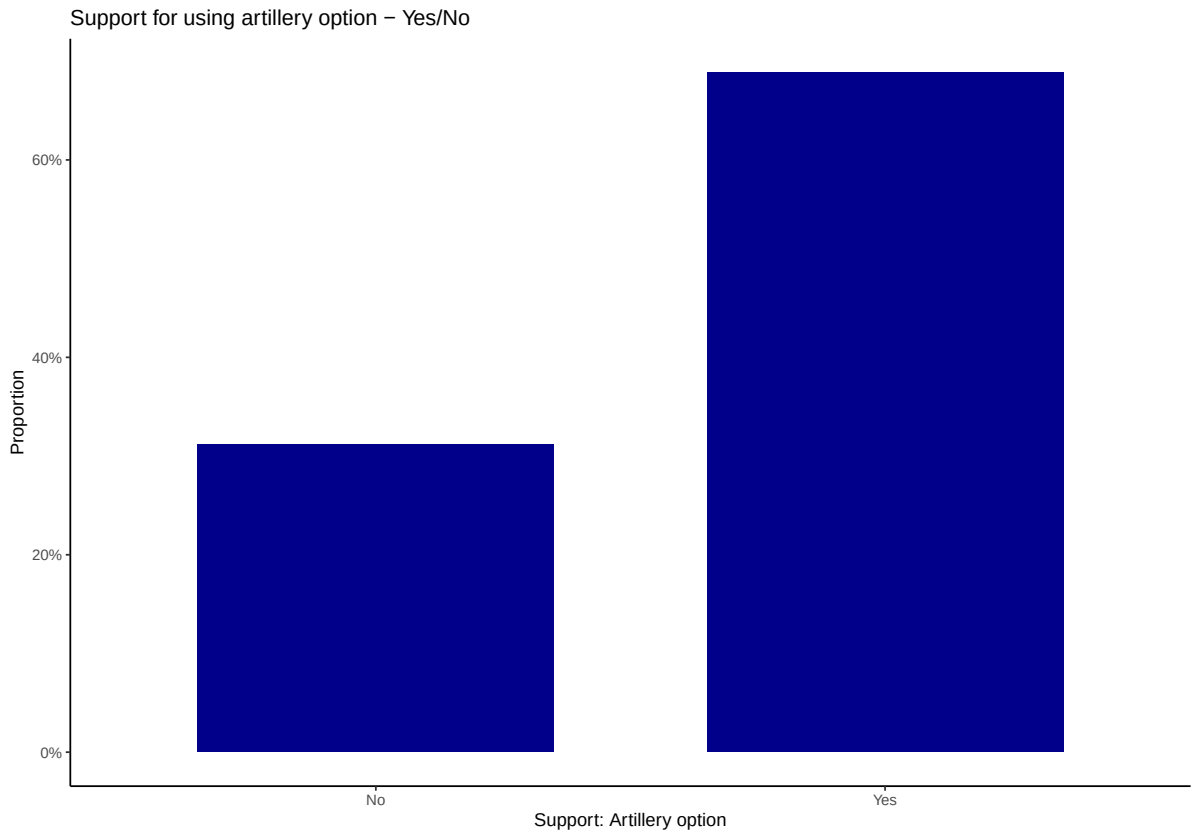
```
artillery.binary <- prop.table(table(Support_Artillery = wardata$approve_artill
                                     exclude = NULL))
barplot(artillery.binary, main = "Support for using artillery option - Yes/No",
        xlab = "Response category", ylab = "Proportion of respondents",
        names.arg = c("No","Yes"))
```



# Plotting alternative

# Bar-plot with tidyverse

```
ggplot(wardata, aes(x=factor(prefer_artillery_dummy))) +
  geom_bar(aes(y = (..count..)/sum(..count..)), width = 0.7, fill = "darkblue") +
  xlab("Support: Artillery option") + ylab("Proportion") +
  scale_y_continuous(labels=scales::percent) +
  scale_x_discrete(labels = c("0"="No","1"="Yes")) +
  ggtitle("Support for using artillery option - Yes/No") + theme_classic()
```



# Visual options

## Histogram

- Useful for numeric values.
- Plotting the distribution of variable.

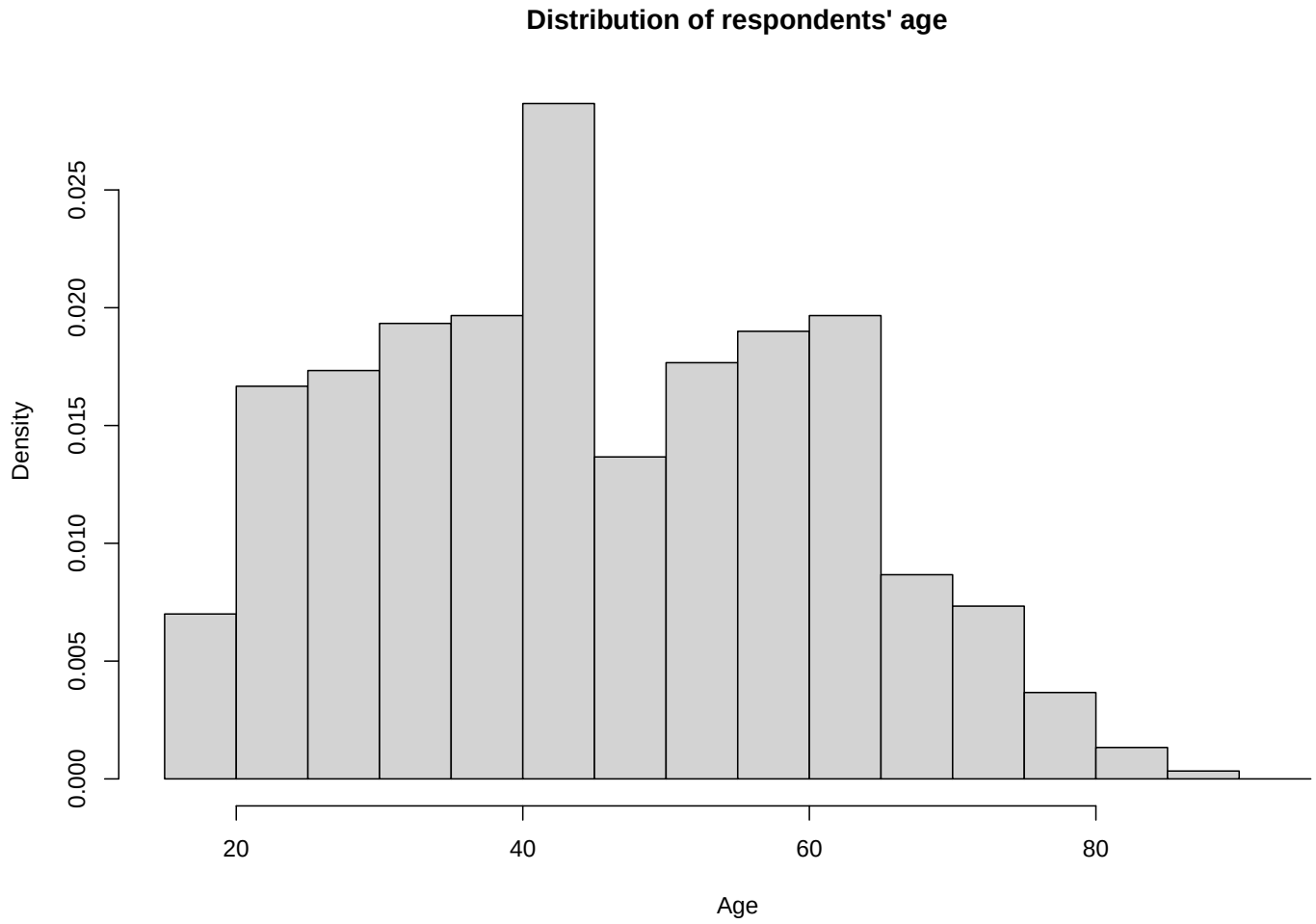
```
# Distribution of respondents' age
wardata$age <- (2014 - wardata$birthyr)
summary(wardata$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   33.00   44.00   45.39   58.00   88.00
```

- Create bins along values of interest.
- 5-year bins: [15,20), [20,25), [25,30), . . . [90,95]

# Histogram: Base R

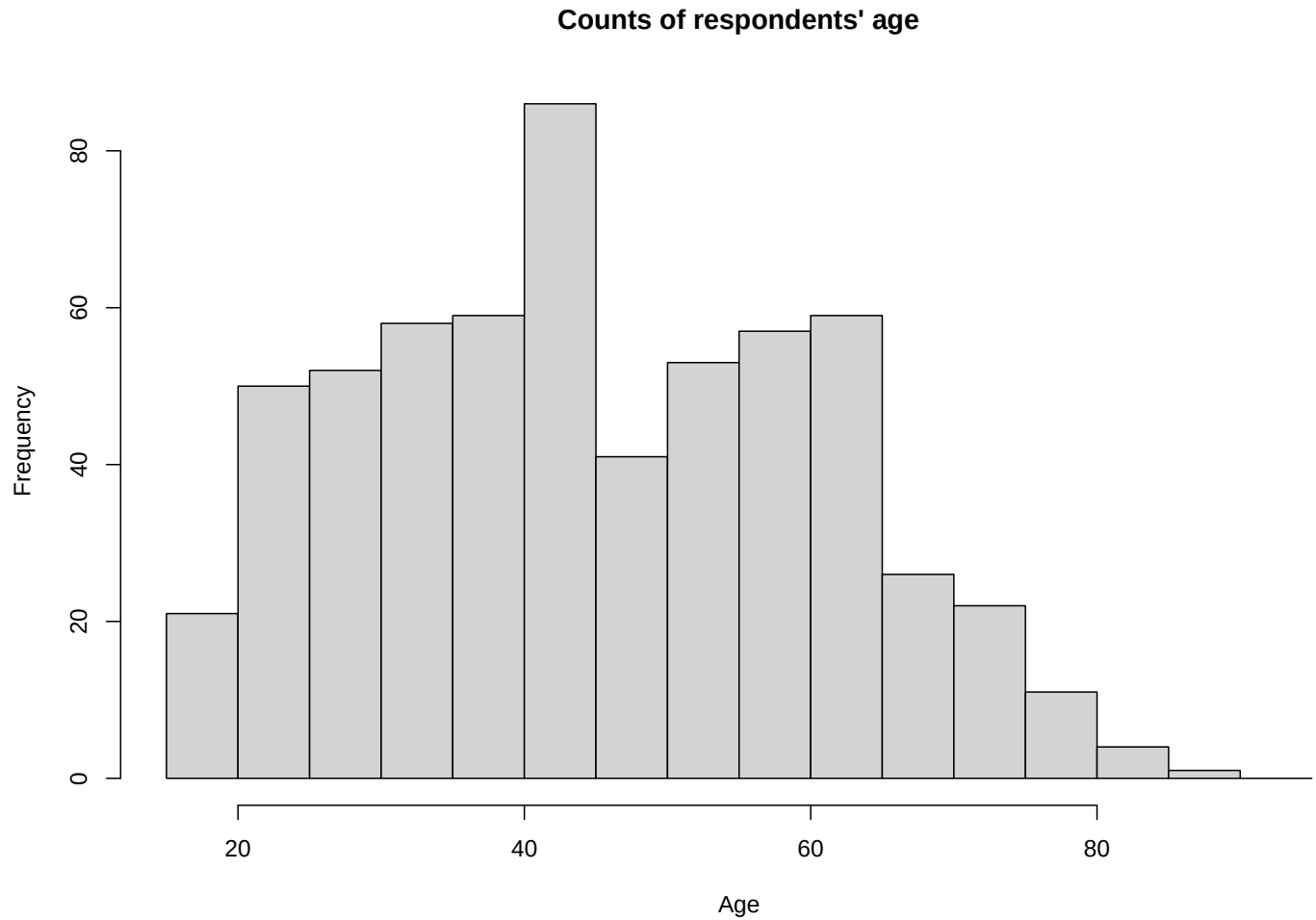
```
hist(wardata$age, freq = FALSE, breaks = seq(from = 15, to = 95, by = 5),
     xlab = "Age",
     main = "Distribution of respondents' age")
```



# Histogram

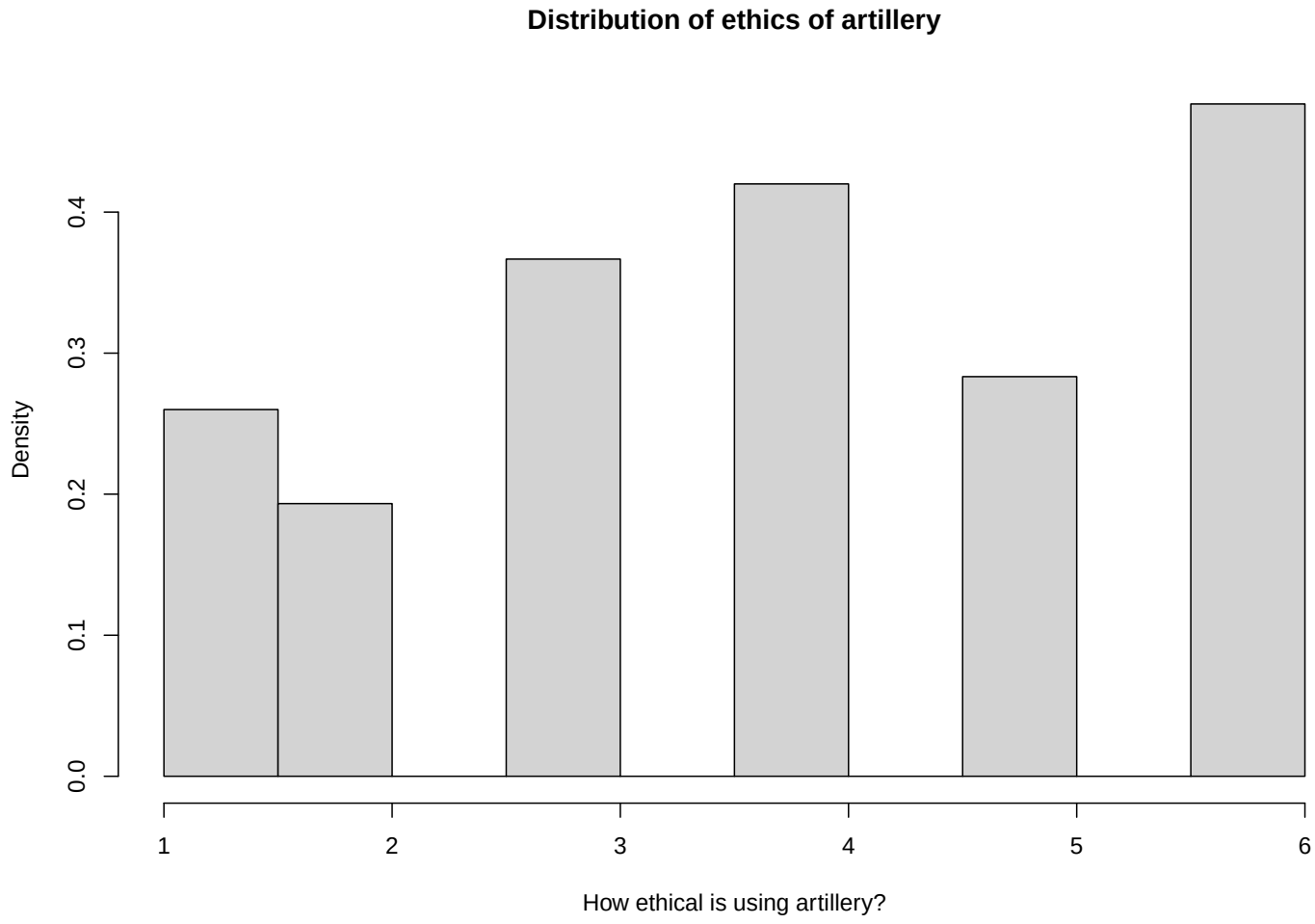
## Counts instead of density

```
hist(wardata$age, freq = TRUE, breaks = seq(from = 15, to = 95, by = 5),
     xlab = "Age",
     main = "Counts of respondents' age")
```



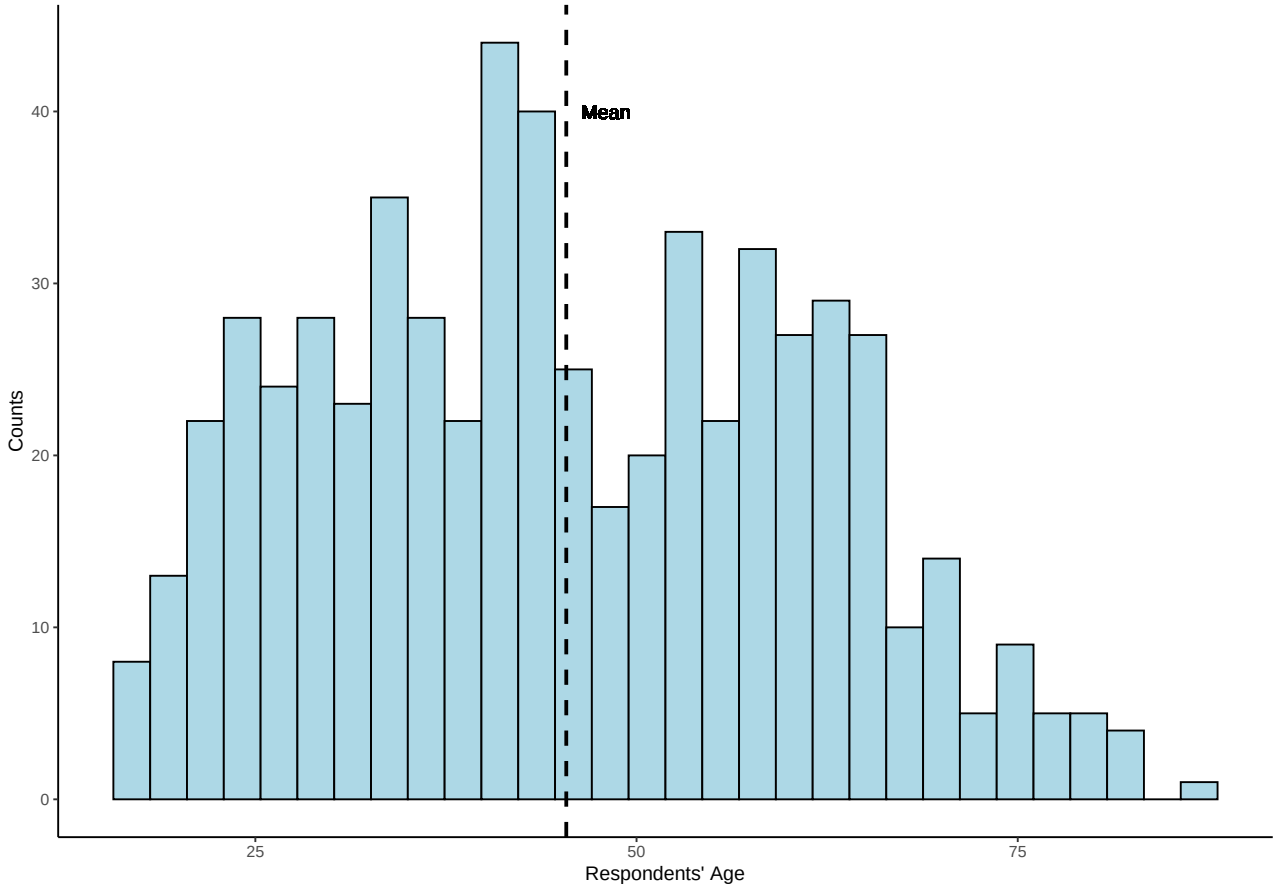
# Artillery ethical?

```
hist(wardata$artillery_ethical, freq = FALSE,
     xlab = "How ethical is using artillery?",
     main = "Distribution of ethics of artillery")
```



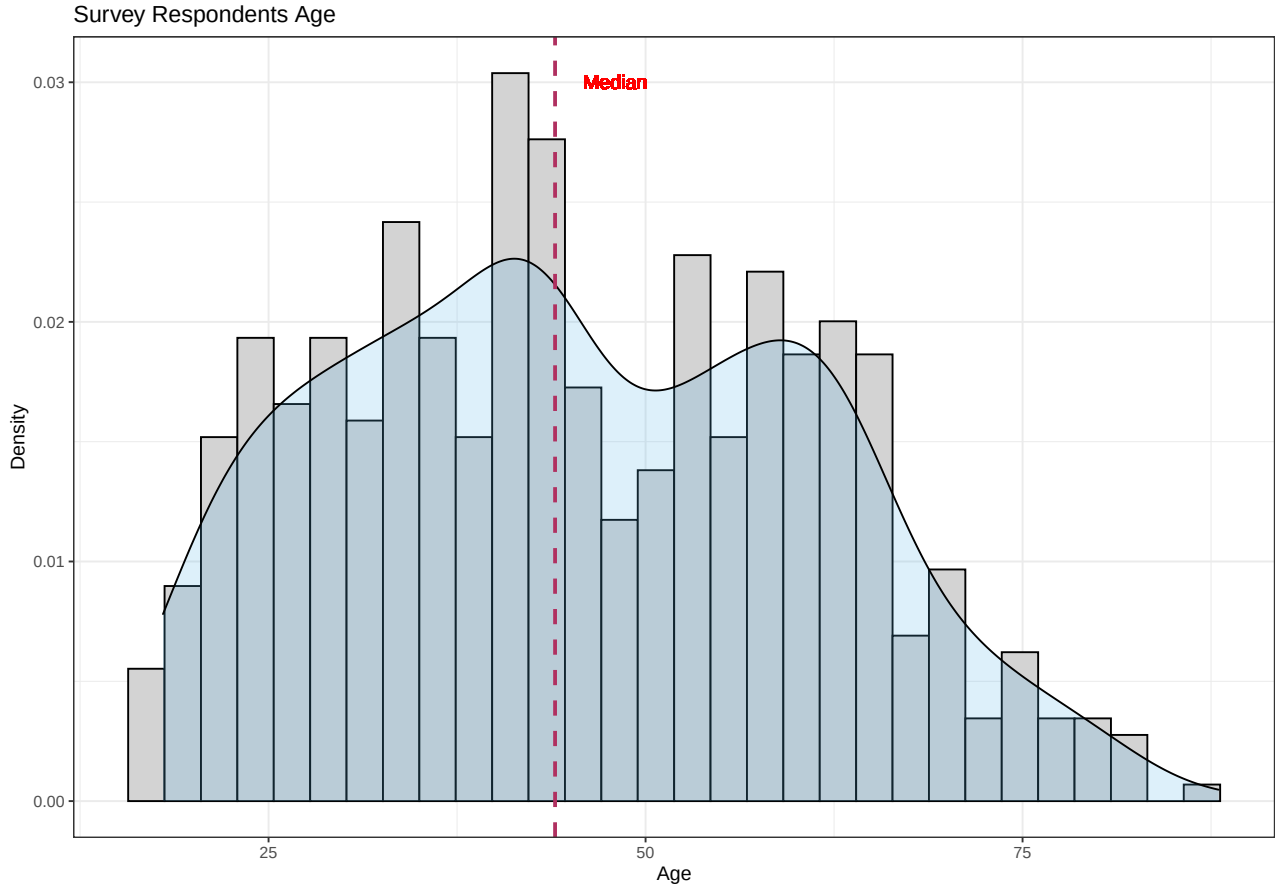
# Histogram: Tidyverse

```
ggplot(wardata, aes(x=age)) +
  geom_histogram(color="black", fill="lightblue") +
  theme_classic() + ylab("Counts") + xlab("Respondents' Age") +
  geom_vline(aes(xintercept=mean(age)),
             color="black", linetype="dashed", size=1) +
  geom_text(x = 48, y = 40, label = "Mean")
```



# Histogram: Tidyverse

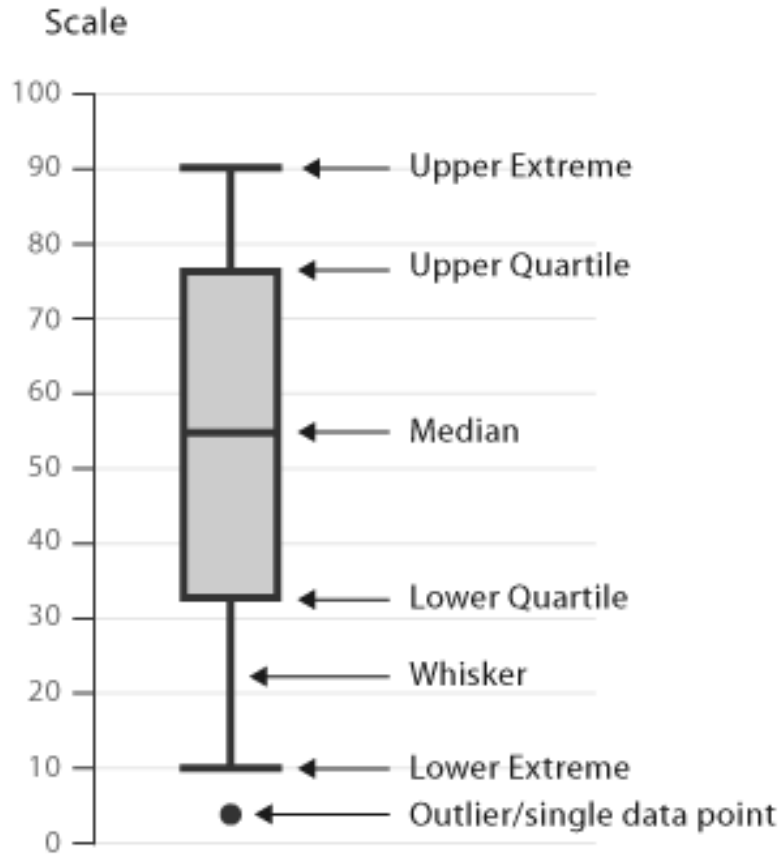
```
ggplot(wardata, aes(x=age)) +
  geom_histogram(aes(y=..density..), colour="black", fill="lightgrey")+
  geom_density(alpha=.2, fill="#56B4E9") +
  xlab("Age") + ylab("Density") + theme_bw() + ggtitle("Survey Respondents Age") +
  geom_vline(aes(xintercept=median(age)),
             color="maroon", linetype="dashed", size=1) +
  geom_text(x = 48, y = 0.03, label = "Median", col = "red")
```



# Visual Options

## Boxplot

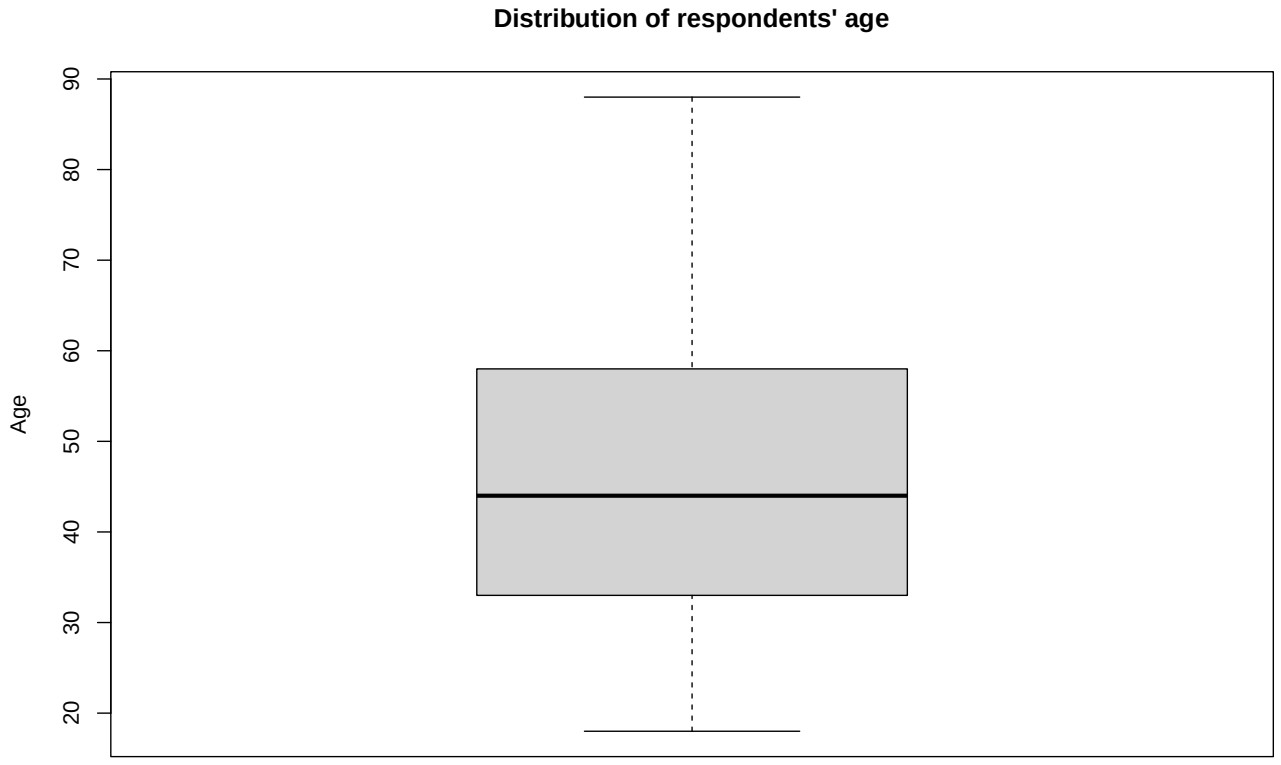
- Useful for a single variable distribution.
- Comparing multiple variables.



# Exploring variables with boxplots

## Base R: single variable

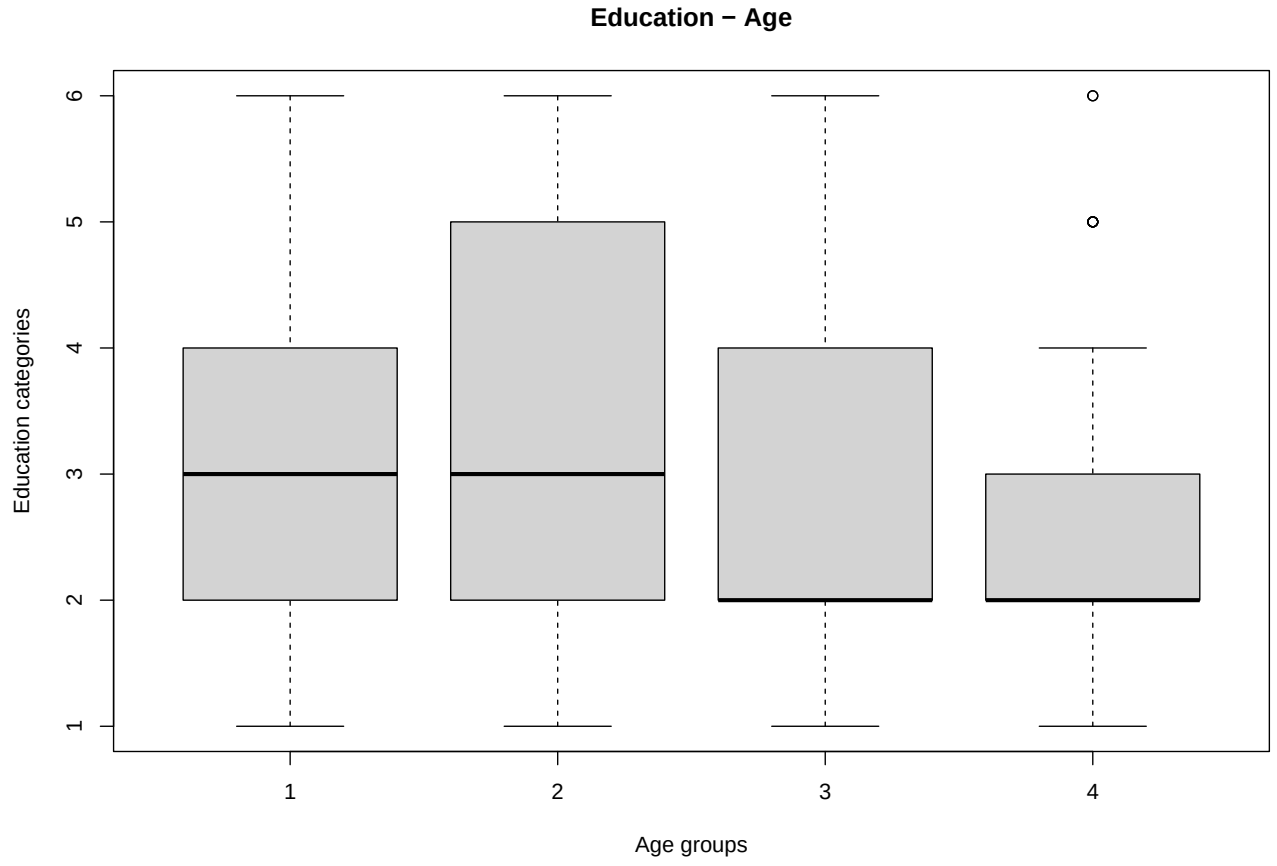
```
boxplot(wardata$age, ylab = "Age",
        main = "Distribution of respondents' age")
```



# Comparing variables: Boxplots

## Education and Age: Base R

```
boxplot(educ ~ agegroup, data = wardata, xlab = "Age groups",
        ylab = "Education categories", main = "Education - Age")
```



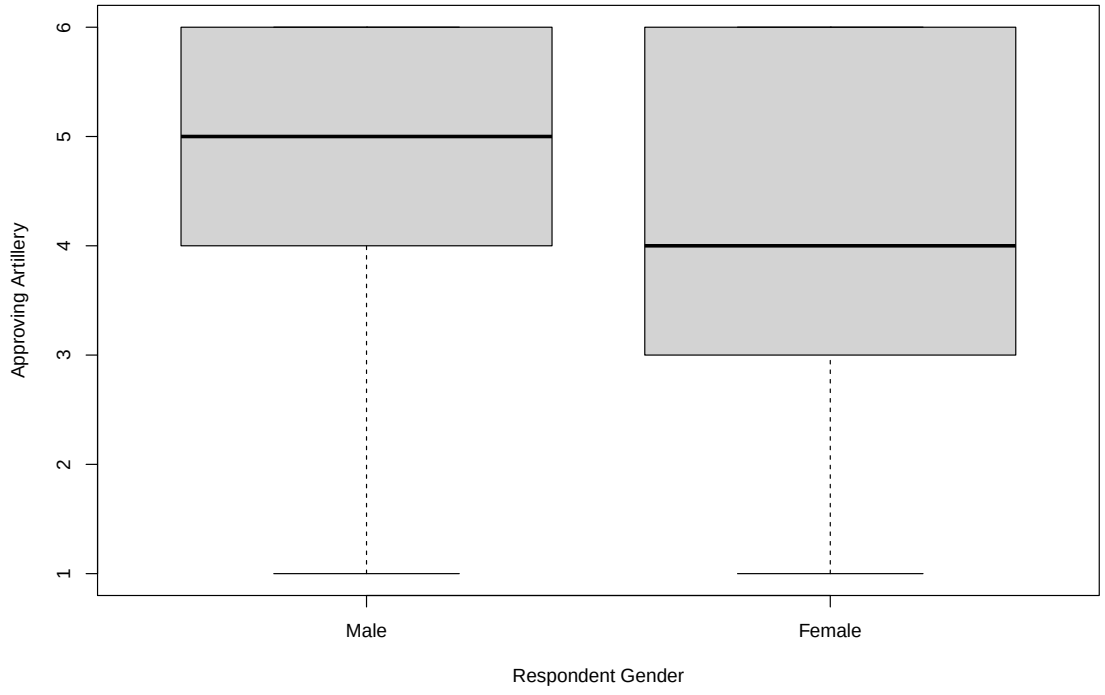
# Comparing variables: Boxplots

## Gender and using artillery

```
tapply(wardata$artillery_approve, wardata$gender, mean)
```

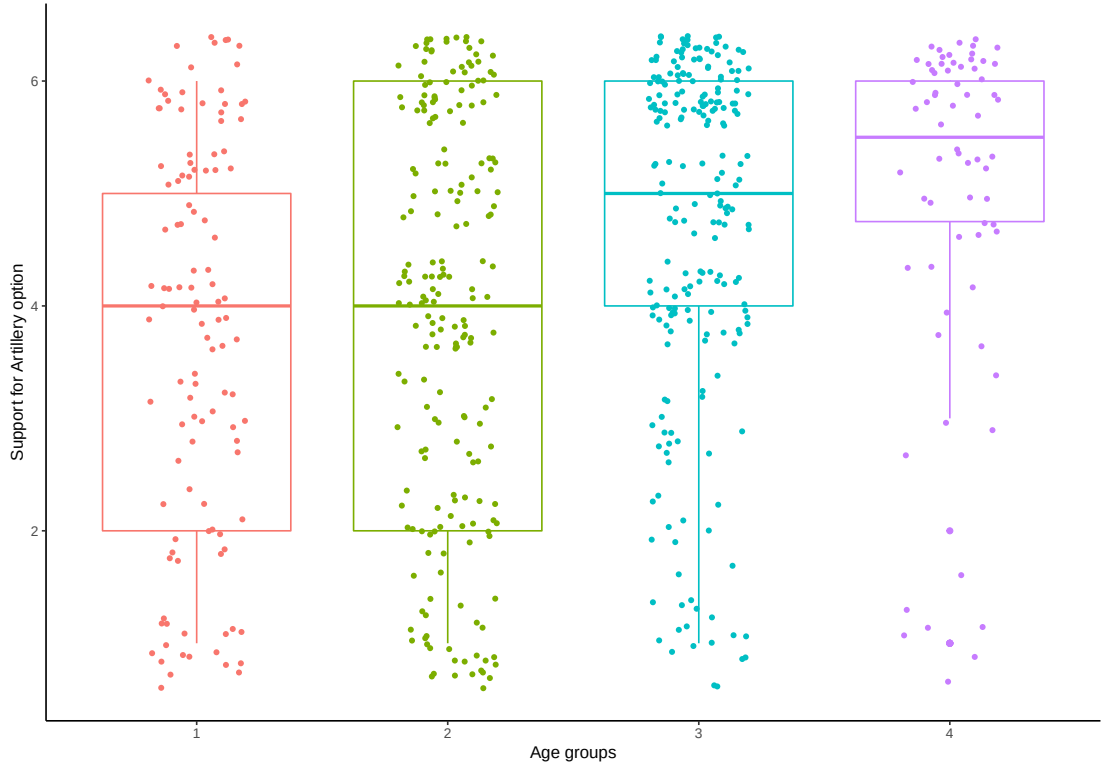
```
##        1        2
## 4.538182 4.009231
```

```
boxplot(artillery_approve ~ gender, data = wardata, xlab = "Respondent Gender",
        ylab = "Approving Artillery", names = c("Male", "Female"))
```



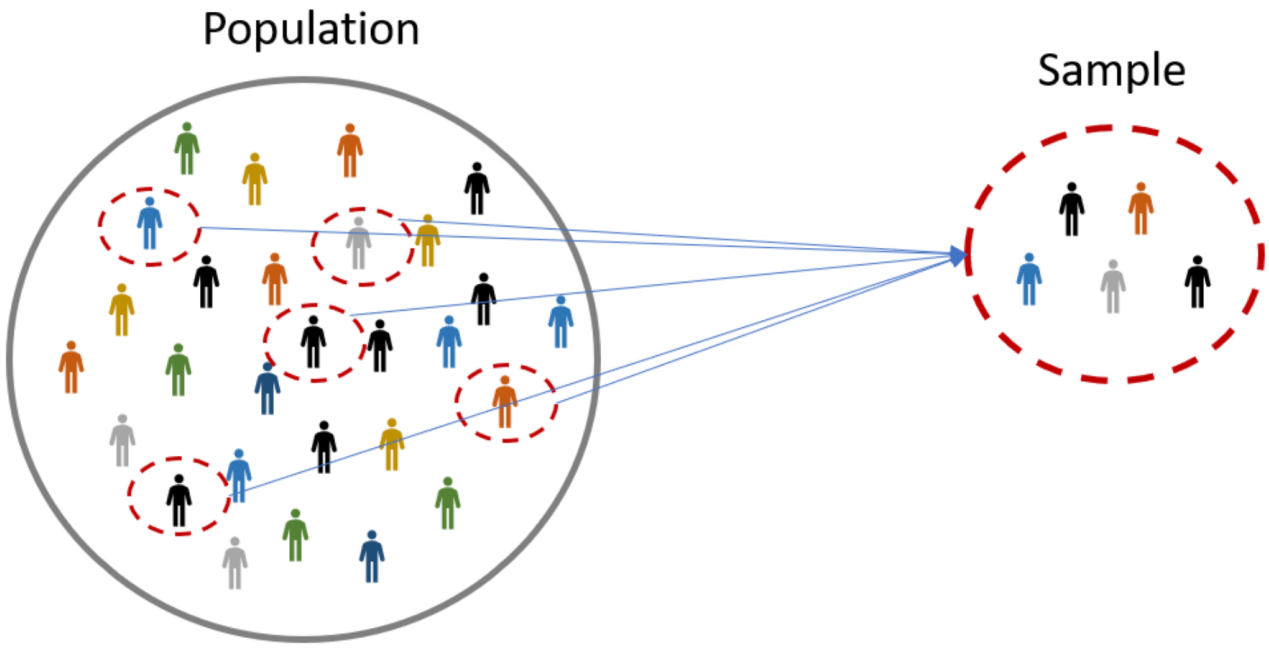
# Boxplots: Artillery option and Age (tidyverse version)

```
ggplot(wardata, aes(x=factor(agegroup), y = artillery_approve,
                    color = factor(agegroup))) +
  geom_boxplot() +
  geom_jitter(shape=16, position=position_jitter(0.2)) +
  xlab("Age groups") + ylab("Support for Artillery option") +
  theme_classic() + theme(legend.position = "none")
```



# Surveys

- Sampling and randomization.
- *Probability sampling*



# Sampling

- **Simple random sampling (SRS)**.
- *Without replacement* procedure.

# Apply SRS

- Obtain our sampling frame.
- Problems:
  - Address lists not updated.
  - Who uses land-lines?
  - Method of RDD - Random digit dialing.

*How representative is our sample?*

# Data manipulations

**Log transform**: deal with outliers (extreme large values).

Skew the analysis of the data

```
summary(afghan$population)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     4.0   239.0   450.0   746.1   823.2 35900.0
```

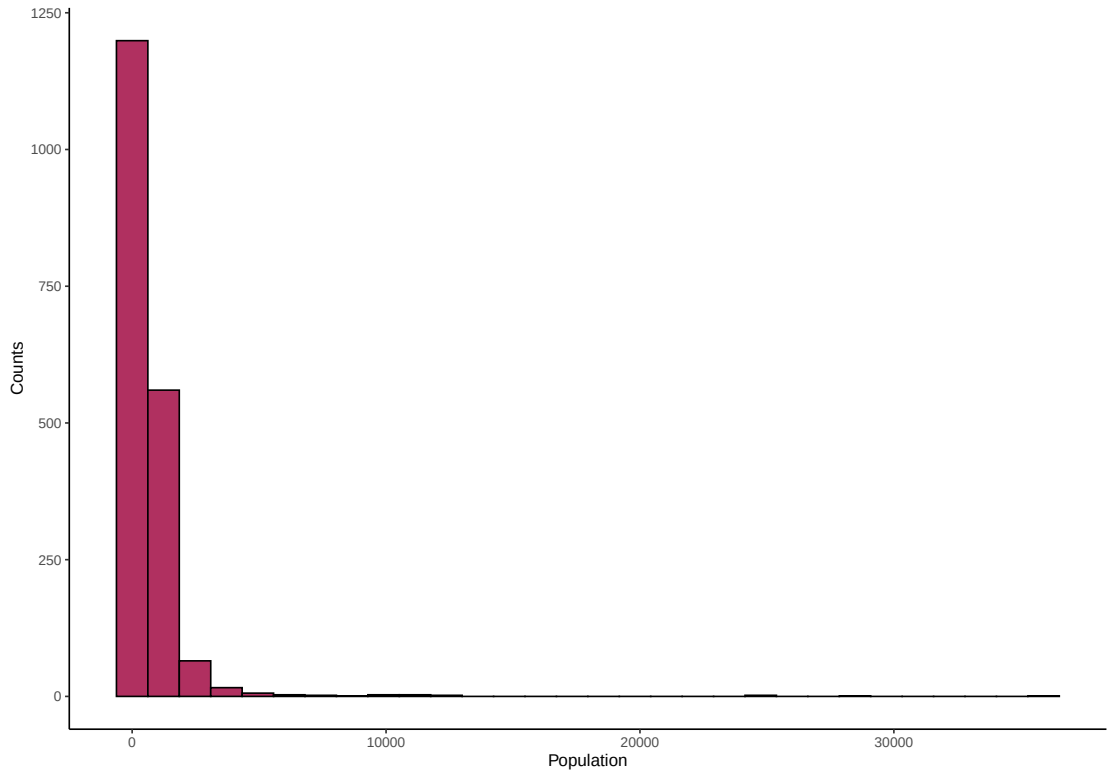
```
afghan$pop_out <- ifelse(afghan$population > 2000, 1,0)
prop.table(table(outliers = afghan$pop_out))
```

```
## outliers
##         0         1
## 0.9527897 0.0472103
```

# Outliers visual

Small number of villages with large population ( $>$ 2000)

```
ggplot(afghan, aes(x=population)) +
  geom_histogram(color="black", fill="maroon") +
  theme_classic() + ylab("Counts") + xlab("Population")
```

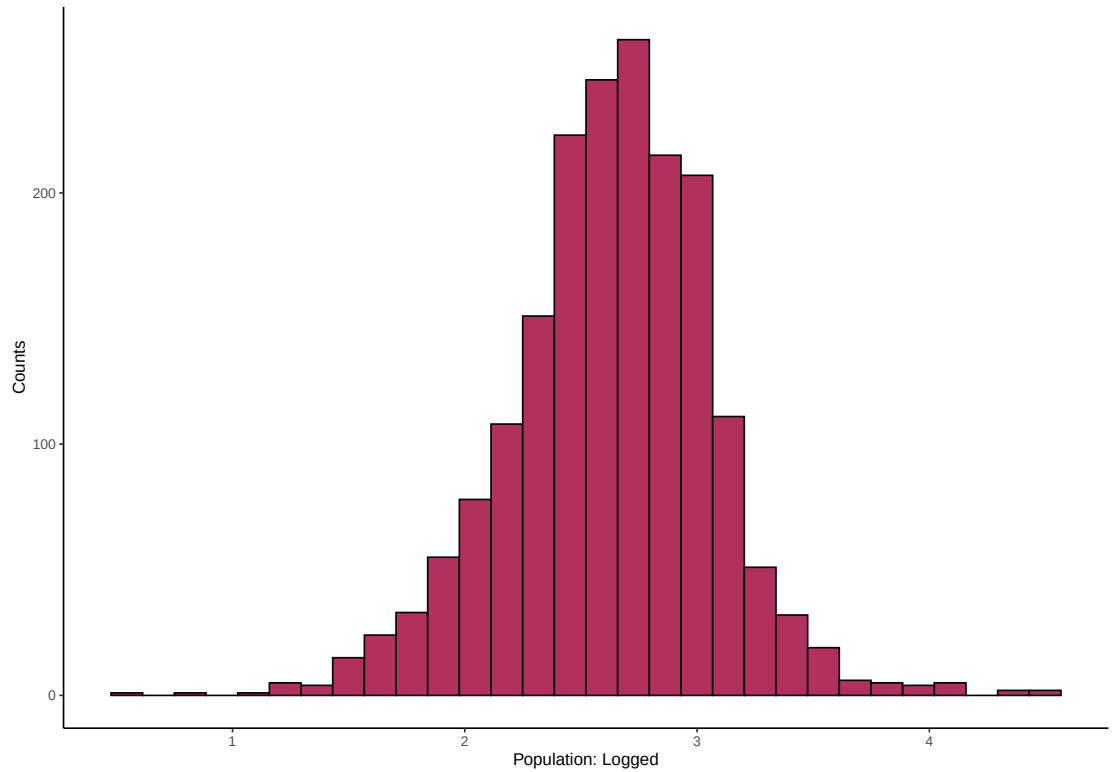


# Log transform

Use natural log to reduce outliers effect

```
afghan$pop_l <- log(afghan$population, 10)

ggplot(afghan, aes(x=pop_l)) +
  geom_histogram(color="black", fill="maroon") +
  theme_classic() + ylab("Counts") + xlab("Population: Logged")
```



# Wrapping up Week 4

Summary:

- Measurement - why? what's so important? Erros in measurement.
- Operational and conceptual definitions.
- Surveys: sampling, randomization, challenges.
- Visuals: why? what not to do? types of plots.
- R work: counting NAs, na.omit(), plots using ggplot and base R, log transform.
- R task details - **due Oct. 4th**