Bush 631-603: Quantitative Methods

Lecture 10 (03.28.2023): Probability vol. II

Rotem Dvir

The Bush school of Government and Public Policy

Texas A&M University

Spring 2023

What is today's plan?

- Calculating uncertainty: **probability**
- How probability is linked to our data.
- Random sample sums, means and their uncertainty.
- Large samples/data and their benefits for our analysis.
- R Tech: data management with tidyverse.
- R work: table(), loops, simulations, plots.

We have findings!!!

- Data patters are systematic? Or noise?
- Our estimates $\rightarrow$ real relationship or random?

PROBABILITY:

- Set of tools to measure uncertainty in world (and our data).
- Method to formalize uncertainty or chance variation.
- Define odds for all possible outcomes.

Probability theory

Calculate probability of event:

$$P(A) = \frac{Elements(A)}{Elements(\Omega)}$$

Example: coin toss x 3

Get an least two heads?

Sample space (Ω): $\{$HHH,HHT,HTH,HTT,THH.THT,TTH,TTT$\}$.

Event A: $\{$HHH,HHT,HTH,THH$\}$.

Probability: $P(A) = \frac{4}{8} = 0.5$

Conditional probability

- We know event B occurred, what is the probability of event A?

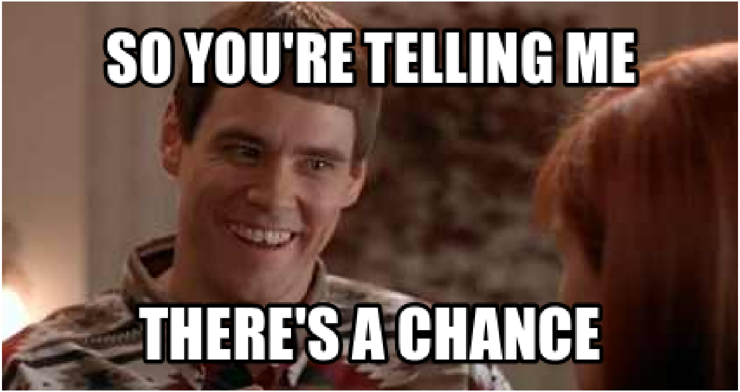
$$P(A|B) = \frac{P(A\&B)}{P(B)}$$

- Conditioning information matters:
  - Twins.
  - Monty hall problem (why switching is good..)

Independence

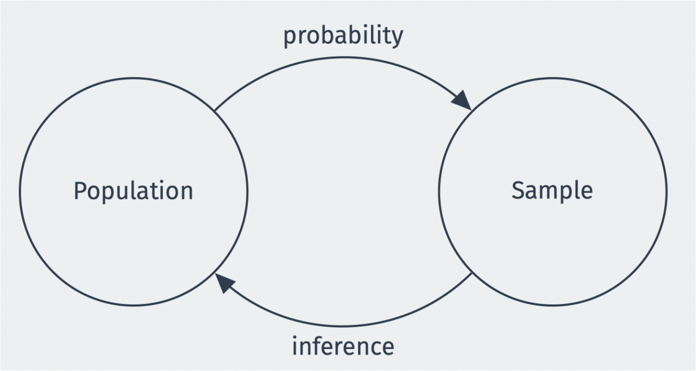
- Events are not related.
- Knowing the A occurred does not affect the probability of B occurring.
- Marginal probability of B (knowing A occurred) remains P(B).
- Formally:
  - $P(A\&B) = P(A) * P(B)$
  - $P(A|B) = P(A)$
  - $P(B|A) = P(B)$

Study probabilites



Study probability

- Foundations for estimating quantities we care about.
- Making inferences from data to population

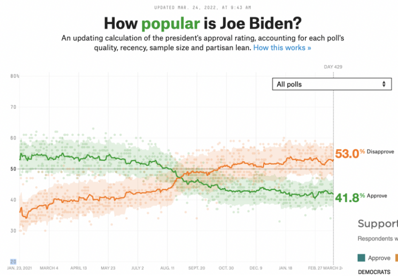


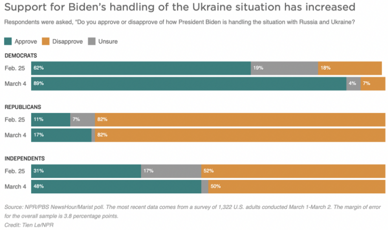
How did we get the data?

- Learn about the process that ‘generated our data’
- The role of uncertainty in this process

Approval data

How popular is president Biden?





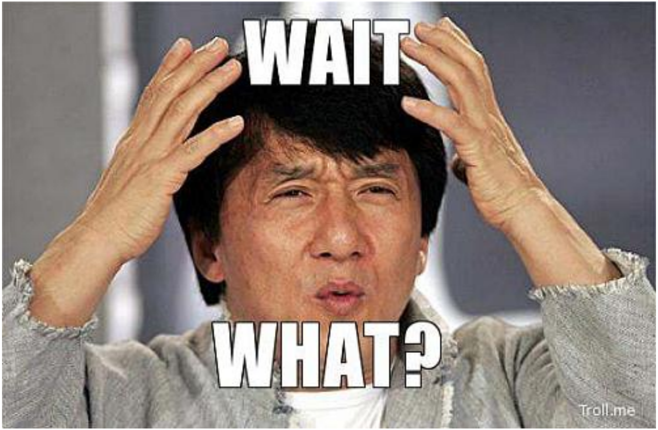
Random variables

- President's approval $\rightarrow$ public samples.
- Using probability to infer from sample to US population.
- The challenge: How to "draw" a Biden supporter?

$$\Downarrow$$

Use random variables to map outcomes to numbers

Random draws...



- Draw people???
- Random selection of values.

Random draws of. . . states

THE WORLD BANK
IBRD · IDA

Data > Data Catalog > World Development Indicators > Tables > 6.11

6.11 World Development Indicators: Aid dependency

Show Metadata Links

	Net official development assistance				Aid dependency ratios		
	Total	per capita	Grants	Technical cooperation	Net official development assistance	Net official development assistance	Net official development assistance
	$ millions	$	$ millions	$ millions	% of GNI	% of gross capital formation	% of imports of goods, services and primary income
	2019	2019	2019	2019	2019	2019	2019
Afghanistan	4,284	113	3,915	300	21.9	..	57.8
Albania	28	10	150	103	0.2	..	0.4
Algeria	176	4	98	160	0.1	0.2	..
American Samoa	..	..	..	..	..	..	..
Andorra	..	..	..	..	..	..	..
Angola	50	2	158	47	0.1	..	0.2
Antigua and Barbuda	27	283	23	1	1.7	..	..
Argentina	18	0	49	43	0.0	0.0	0.0
Armenia	420	142	109	47	3.0	17.6	5.1

- Our objective: study regime type and extent of aid.
- Regimes: dictators, democracies, semi-democracies, etc.
- Draw regimes at-random and test causal mechanism.

Random draws, why?

Randomization:

- RCT: average all pre-treatment factors.
- RCT: strong causal explanation.

- Observational: reduce **selection bias**.
  - Allow expectations to be refuted.

We generate estimates, but with uncertainty

Numbers and Aggies example

Aggies in the NFL: position groups and conferences

```
skillposition <- subset(Ags, subset = (Group == "OF" | Group == "DF"))
head(skillposition)
```

```
## # A tibble: 6 x 5
##   Player             Team                  Position Group Conference
##   <chr>              <chr>                 <chr>    <chr> <chr>
## 1 Christian Kirk     Jacksonville Jaguars  WR       OF    NFC
## 2 Jake Matthews      Atlanta Falcons       OT       OF    NFC
## 3 Otaro Alaka        Baltimore Ravens      LB       DF    AFC
## 4 Justin Madubuike Baltimore Ravens      DT       DF    AFC
## 5 Tyrel Dodson       Buffalo Bills         LB       DF    AFC
## 6 Germain Ifedi      Chicago Bears         OG       OF    NFC
```

Random variables and Aggs

```
## # A tibble: 2 x 2
##   Group     n
##   <chr> <int>
## 1 DF       12
## 2 OF       22
```

- Choose one at-random.
- Define **random variable**:
  - $X = 1$ if selected Aggie plays Offense, $X = 0$ otherwise.
- Why *random*?
- Before we draw an Aggie, uncertainty about the value of X.
- Direct link to probability:
  - $P(X = 1) = \text{P(Draw Offense)} = \frac{22}{34} = 64.7\%$

Random variables

- Classified by construction and shape

BERNOULLI

- r.v. X follows a **bernoulli distribution** with probability p if:
  - X takes one of two values only (0,1).

- $P(X = 1) = p$
  - $P(X = 0) = 1 - p$

- Fits a binary indicator
- Describes **any** potential variable with a probability that $X = 1$.

Random variables

- Why?
  - The uncertainty of our estimates.
  - Figure the uncertainty of quantities as sample means or sums.
- Aggies data: drawing **two** players (with replacement):
  - $X_1 = 1$ if Aggie is Offense, $X_1 = 0$ otherwise.
  - $X_2 = 1$ if Aggie is Offense, $X_2 = 0$ otherwise.
- Define new r.v $\rightarrow S = X_1 + X_2$
- Data is the sum of all potential X_1, X_2.
- What are the values of S?

Random variables to probabilities

- Map S values to probabilities
- Always draw 2 Aggs.
- Sample space (Ω) = $\{$OF-OF; OF-DF; DF-OF; DF-DF$\}$.
- k $\rightarrow$ Values of S (0, 1, 2).
- $P(S = k)$?
- $P(S = k) = P(Ag_1 + Ag_2) = P(Ag_1) * P(Ag_2)$
- Why? Addition rule for mutually exclusive events.

Random variables to probabilities

```r
prob_off <- 22/34
prob_def <- 12/34

# Offense:Offense (OF-OF)
prob_off * prob_off
```

```
## [1] 0.4186851
```

```r
# Offense:Defense (OF-DF)
prob_off * prob_def
```

```
## [1] 0.2283737
```

```r
# Offense:Defense (DF-OF)
prob_def * prob_off
```

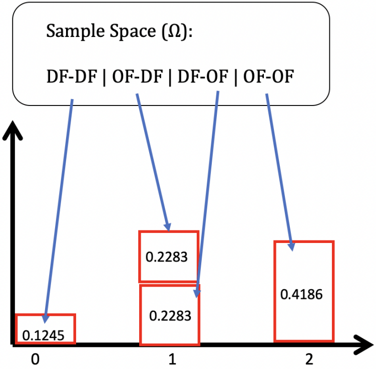
```
## [1] 0.2283737
```

```r
# Defense:Defense (DF-DF)
prob_def * prob_def
```

```
## [1] 0.1245675
```

Mapping draws to probabilities

Plotting probabilities of separate draws



Outcome	S	Probability
OF-OF	0	0.1245
OF-DF	1	0.2283
DF-OF	1	0.2283
OF-OF	2	0.4186

k	P(S = k)
0	0.1245
1	0.4567
2	0.4186

Binomial Distribution

- X is r.v. taking any value between 0 and n.
- Coin flips: number of heads with probability p in n independent flips.
- Aggs: S = number of OF when we draw **2 players** (n=2; P=0.4186).

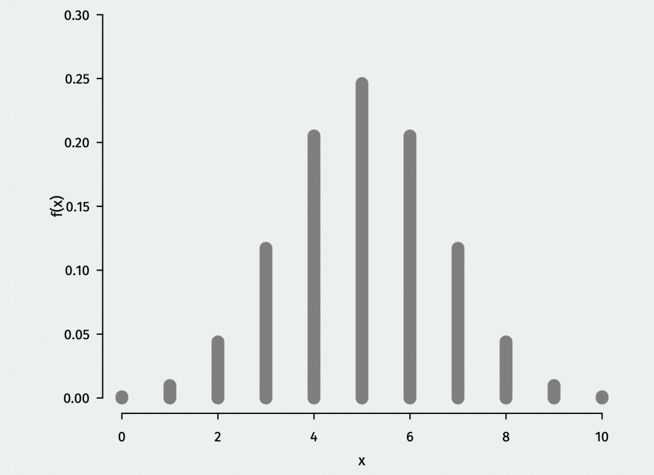
Probability Mass Function (PMF):

- Evaluates probability of *any possible value* of these random variables.

$$P(X = k) = \binom{n}{k} * p^k * (1-p)^{n-k}$$
$$\binom{n}{k} = \frac{n!}{(k!(n-k)!)}$$

Binomial distribution

- X = number of heads in multiple coin flip trails
- $P = f(x) = 0.5$; $n = 10$



Binomial random variable

- Larger sample, more draws, same probability
- How many OF players?

```
# Possible number of Offensive players of 500
rbinom(n=3, size = 500, prob = 0.647)
```

```
## [1] 314 318 319
```

- Simulation

```
sims <- 10000
draws <- rbinom(sims, size = 500, prob = 0.647)
head(draws, n=12)
```

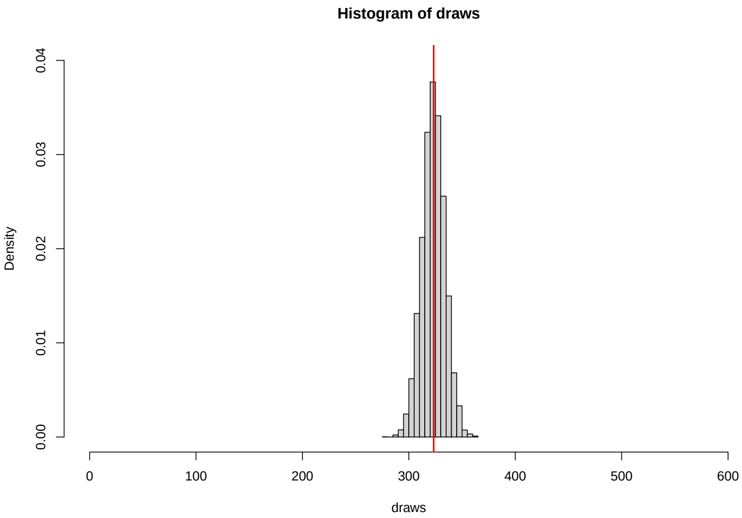
```
##  [1] 298 317 336 311 331 332 319 312 333 334 300 308
```

```
mean(draws)
```

```
## [1] 323.5901
```

Plotting our sims

```
# Histogram of draws
hist(draws, freq = FALSE, xlim = c(0, 600), ylim = c(0, 0.04))
abline(v = 323.3, col = "red", lwd = 2)
```



Simulating Congress calls

- Lobbying firm: gender balance of calls to senators
- Total number of calls = 1000, random selection (with replacement)
- How many calls to women senators?

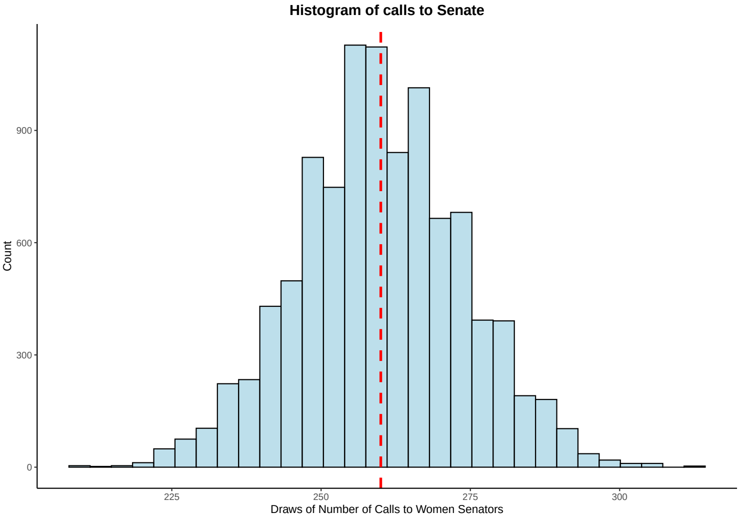
```
# Simulate calls (p=0.26)

sims2 <- 10000
draws2 <- rbinom(sims, size = 1000, prob = 0.26)
mean(draws2)
```

```
## [1] 259.9875
```

```
head(draws2, n=12)
```

```
##  [1] 263 230 293 235 257 278 258 263 271 269 276 276
```

Plotting Senate calls simulation



Probability distributions

- Describe the uncertainty of random variables
- We learn of the population after analyzing the sample

Probability distributions

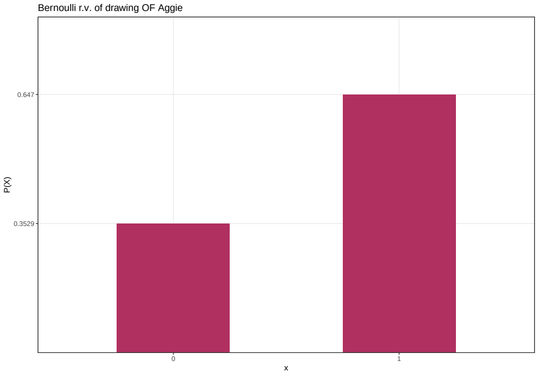
- Multiple ways to represent the distribution.
- Type of r.v. $\rightarrow$ which distribution we face.

- Two general classes:
  - Discrete: X takes finite number of values (heads in n coin flips, battle deaths in civil wars).
  - Continuous: X takes any real value (GDP/cap, how long do you spend time on Tik-Tok?)

Discrete PMF

- Barplot to illustrate probabilities (share of each possible value)
- Bernoulli r.v.: using the Ags data (OF or DF?)

```r
plot.dat <- data.frame(x = c("0", "1"), y = c("0.3529", "0.647"))
ggplot(plot.dat, aes(x,y)) +
  geom_bar(stat = "identity", width = 0.5, fill = "maroon") + ylab("P(X)") +
  ggtitle("Bernoulli r.v. of drawing OF Aggie") + theme_bw()
```



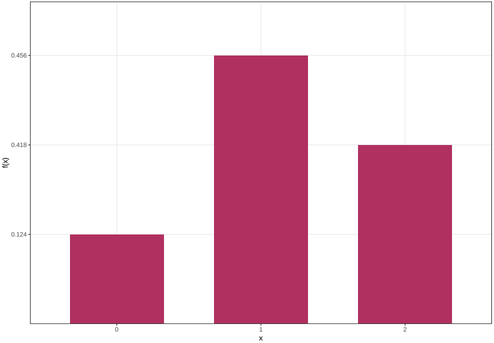
Binomial PMF

- Illustrate probabilities of 3 values (r.v. X)

```
dbinom(x = c(0,1,2), size = 2, prob = 22/34)
```

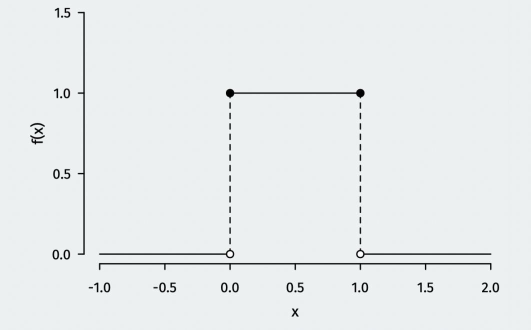
```
## [1] 0.1245675 0.4567474 0.4186851
```

```
plot.dat2 <- data.frame(x = c("0", "1", "2"), y = c("0.124", "0.456", "0.418"))
ggplot(plot.dat2, aes(x,y)) +
  geom_bar(stat = "identity", width = 0.65, fill = "maroon") + ylab("f(x)") +
  theme_bw()
```



Continuous random variables

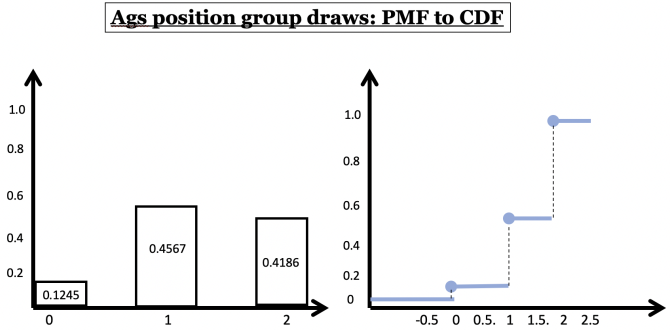
- **Probability density function (PDF)**.
- Describe probability 'around' a given point.
- An 'infinite' histogram $\rightarrow$ many bins (looks smooth).
- Probability of interval $=$ area under curve.



Random variable distributions

Cumulative distribution function (CDF).

- Common to discrete or continuous random variables.
- Describe the probability that some r.v. will be less or equal to some k.



Well, lets


Okay fine, you're right. Let's take a
break, let's cool off. Okay, let's get some
frozen yogurt or something.

R Tech

- Data management with tidyverse package.
- Functions:
  - Subsets $\rightarrow$ use filter().
  - Add variables $\rightarrow$ use mutate().
  - Organize data $\rightarrow$ use select(), arrange().
  - Group data and calculate summary stats (mean, median):
    - Use summarise() to create new sub-set for stats.
    - Create your preferred stats: mean(),median(),max()...

Using r.v. distributions

- How to use probability distributions?
  - Mean: center of our distribution.
  - Variance/Standard deviation: the 'spread' around the center.

- Mean & Variance $\rightarrow$ *Population parameters* (unknown).
- Use our sample (data) to learn about both parameters.

Means & Expectations

Calculate the average: $\{1,1,1,3,4,4,5,5\}$

1. Common: sum all objects & divide by number of objects.

$$\frac{1+1+1+3+4+4+5+5}{8} = 3$$

2. Frequency weights: multiply each value by its frequency in the sample.

$$1 * \frac{3}{8} + 3 * \frac{1}{8} + 4 * \frac{2}{8} + 5 * \frac{2}{8} = 3$$

- Use the frequency weights approach to create the mean of r.v.s.

Expectation

- Expectation ($E[X]$) for the mean of r.v. X.

$$E[X] = \sum_{j=1}^{k} *x_j * P(X = x_j)$$

- The weighted average of the values of the r.v weighted by the probability of each value.

Expectation

- What is $E[X]$?
- Let X be the age for randomly selected individual.
- $E[X] \rightarrow$ average age in the *population*.
- $E[X]$: the link of the sample and population means.
- $E[X]$ properties:
  - $E[a] = a$ (constant).
  - $E[aX] = a * E[X]$ (scale for mean).
  - $E[aX + bY] = a * E[X] + b * E[Y]$ (mean of two values).

Variance

- The 'spread' of the distribution.

$$V[X] = E[(X - E[X])^2]$$

- Weighted avg. of squared distance if each observation from mean.
- Larger deviations $\rightarrow$ larger variance.
- If X be the age for randomly selected individual.
- $V[X]$ $\rightarrow$ spread of ages in *population*.

Variance

- $SD(X) = \sqrt{V[X]}$: allows to make comparison in data.
- $V[X]$ properties:
  - $V[c] = 0$ (constant).
  - $V[aX + c] = a^2 * V[X]$ (scale distribution).
  - $V[X + Y] \neq V[X] + V[Y]$ (unless X & Y are independent).

Sums, means and random variables

- Let X_1 and X_2 be two r.v.s
- Then, $X_1 + X_2$ is also r.v.
- Mean: $E[X_1 + X_2]$; Variance: $V[X_1 + X_2]$

- We 'draw' two global leaders and assign X_1, X_2 as their ages.
- **Sample mean** $\rightarrow$ also a r.v.

$$\bar{X} = \frac{X_1 + X_2}{2}$$

- Uncertainty due to possibility of 'drawing' other leaders.

Global leaders data

- Data: personal characteristics of leaders (Horowitz 2015)

```
head(age.lead, n=9)
```

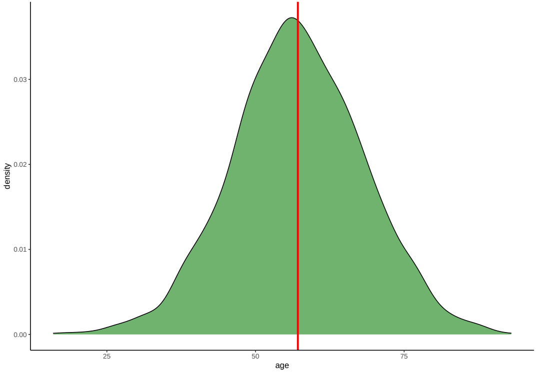
```
## # A tibble: 9 x 4
##   idacr  year leader           age
##   <chr> <dbl> <chr>          <dbl>
## 1 USA    1877 Grant             55
## 2 USA    1881 Hayes             59
## 3 USA    1881 Garfield          50
## 4 USA    1885 C. Arthur         56
## 5 USA    1889 Cleveland         52
## 6 USA    1893 Harrison          60
## 7 USA    1897 Cleveland         60
## 8 USA    1901 McKinley          58
## 9 USA    1909 Roosevelt, T.     51
```

Full sample means

```
# mean of sample
mean(age.lead$age, na.rm = T)
```

```
## [1] 57.122
```

```
# Plot distribution of all leaders in data
ggplot(age.lead, aes(x=age)) +
  geom_density(fill="forestgreen",alpha=0.65) +
  geom_vline(aes(xintercept = mean(age.lead$age, na.rm = T)),size=1.2,color="red") +
  theme_classic()
```



Distributions of sums & means

- 'Draw' two leaders, calculate sum and mean of age.

Drawing leaders at-random

	X_1	X_2	$X_1 + X_2$	Mean X
Draw 1	51 (Teddy R.)	69 (H.W.Bush)	120	60
Draw 2	55 (Rubio-MEX)	42 (Pardo – ECU)	97	48.5
Draw 3	69 (Chirac-FRN)	61 (Brandt-GFR)	130	65
Draw 4	38 (Delvina-ALB)	39 (Doe-LBR)	78	38.5
...	...	...	...	...

Distribution of sum

Distribution of mean

Independent and identical r.v.s

- $X_1 \ldots X_n$ are iid r.v.s.
- Random sample of n respondents on a survey question.
- **Identically distributed**: distribution of X_i is same for all i
  - $E(X_1) = E(X_2) = \ldots = E(X_n) = \mu$
  - $V(X_1) = V(X_2) = \ldots = V(X_n) = \sigma^2$
- Key insights of iid properties:
  - Sample mean = population mean (on average).
  - Variance $\leftarrow$ population variance and sample size.
  - SD of sample $\rightarrow$ *standard error*

$$SE = \sqrt{V[\bar{X}_n]} = \frac{\sigma}{\sqrt{n}}$$

Large samples: Global leaders

- We ‘draw’ two samples of global leaders
- Assign X_1, X_2 as their ages.
- Uncertainty of our data - leaders change each draw.

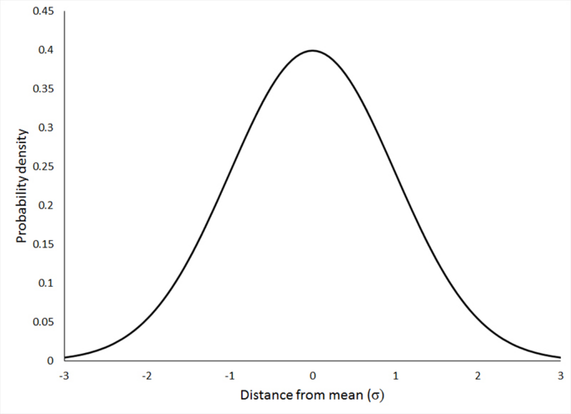
- What happens to our means when the sample size increases?

Large samples

LAW OF LARGE NUMBERS

- $X_1 ... X_n$ is iid with mean μ and variance σ^2.
- As n $\uparrow$, $\bar{x} \to \mu$.
- $P(\bar{x}) \to \mu$ increases as n get larger.

The Normal distribution



$$X \sim N(\mu, \sigma^2)$$

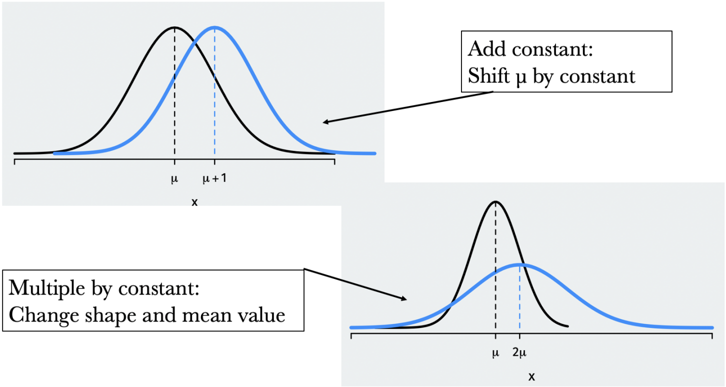
- Mean/expected value $= \mu$
- Variance $= \sigma^2$

The Normal distribution

- A "Bell-shaped" PDF
- Important properties:
  - Any r.v. is more likely to be in center than tails.
  - *Unimodal*: single peak, at the mean value.
  - Symmetric around the mean: equal probabilities.
  - Everywhere positive (tails 'stretch' to infinity).
- **Standard normal distribution**: mean $= 0$, SD $= 1$.
- Standard normal variable $\rightarrow$ *z-score*: $Z = \frac{X - \mu}{\sigma}$

The Normal distribution

- Transforming the normal distribution:



Central limit theorem

- Let X_i be r.v. which is iid and normally distributed.
- $\bar{X}$: also normally distributed in **large samples**.

Sample mean tend to be normally distributed as samples get large

- Extends the application of r.v. in large samples. How?
  - Value approaches μ and normally distributed.
  - Better approximation of population mean value.
  - Sample mean is normally distributed, regardless of the distribution of each X (r.v.).

Simulating larger sample (CLT)

- Draw at-random 1000 leaders from data.
- Calculate and save sample mean multiple times (use a loop)

```
sim.lead <- 1000
all.mn <- rep(NA, sim.lead)

for (i in 1:sim.lead){
  lead.draw <- sample_n(age.lead, 1000)
  all.mn[i] <- mean(lead.draw$age, na.rm = T)
}

head(all.mn)
```

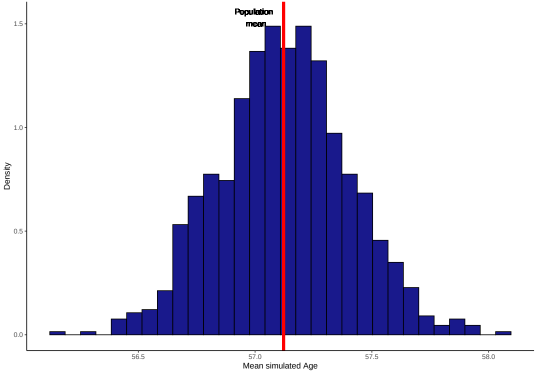
```
## [1] 57.04606 57.26931 57.20081 57.52178 57.28788 57.26289
```

```
mean(all.mn, na.rm = T)
```

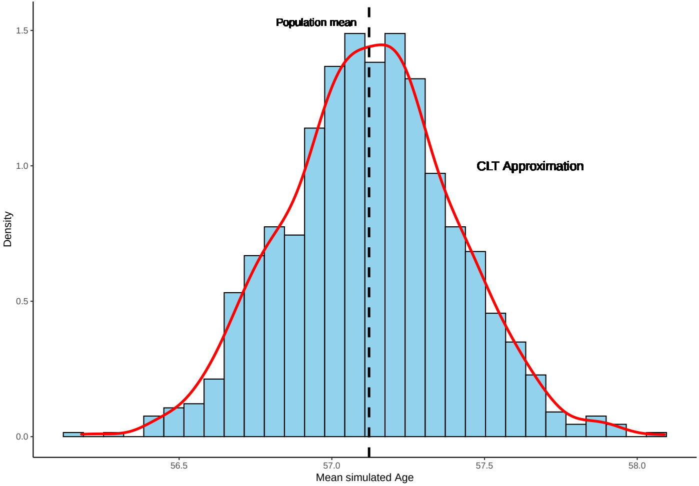
```
## [1] 57.1238
```

Plotting the simulated data

```
# Save vector in data frame and plot (add 'population' mean)
d <- data.frame(x = all.mn)
ggplot(d, aes(x)) +
  geom_histogram(aes(y = stat(density)),fill="navyblue", color="black", alpha=0.9) +
  xlab("Mean simulated Age") + ylab("Density") +
  geom_vline(xintercept = 57.122, color = "red", size = 2) +
  geom_text(aes(x = 57, y = 1.53, label = "Population \n mean")) +
  theme_classic()
```

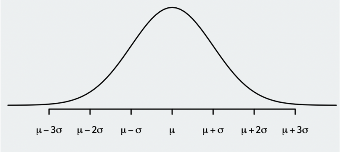


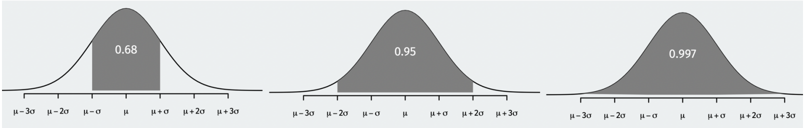
Plotting the simulated data



Empirical rule for normal distribution

If $X \sim N(\mu, \sigma^2)$, then:





68% of dis. → 1 SD of mean

95% of dis. → 2 SD of mean

99% of dis. → 3 SD of mean

Empirical rule in R

```r
# Values
pnorm(1) - pnorm(-1)
```

```
## [1] 0.6826895
```

```r
pnorm(2) - pnorm(-2)
```

```
## [1] 0.9544997
```

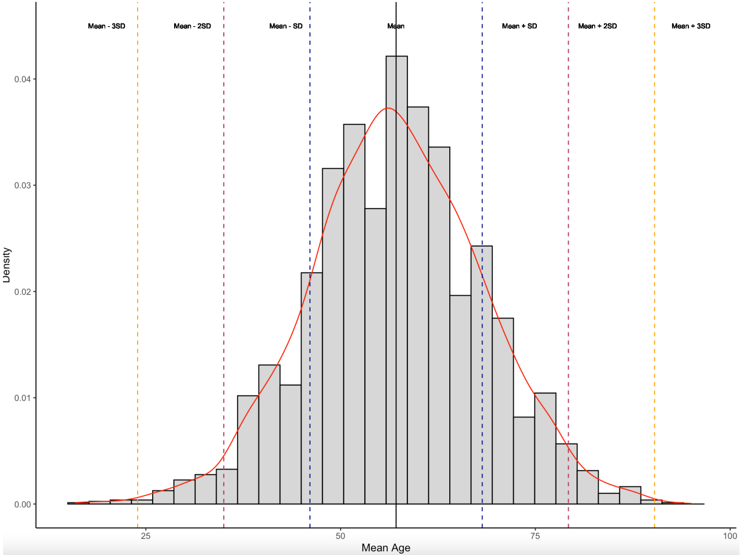
```r
# Use the leader data
mu <- mean(Leader$age, na.rm = T)
sig <- sd(Leader$age, na.rm = T)

pnorm(mu+sig, mean = mu, sd = sig) - pnorm(mu-sig, mean = mu, sd = sig)
```

```
## [1] 0.6826895
```

```r
pnorm(mu+2*sig, mean = mu, sd = sig) - pnorm(mu-2*sig, mean = mu, sd = sig)
```

```
## [1] 0.9544997
```

Leaders age: normal distribution “break-down”



Wrapping up week 10

Summary:

- Probability and uncertainty.
- Mapping probability of events to random variables.
- Linking r.v. to our data - random selection of values.
- Sums and means of random sample.
- Probability distributions (Bernoulli, Binomial, etc.).
- Large samples and their benefits.
- CLT / Law of large numbers.
- The normal distribution.

Research Proposal by Midnight Friday (3/31)!!