PS6 Lecture 4

Daniel Lim

University of California, Los Angeles

8/14/2013

Agenda

- More normal distribution
- Random variables, population and samples
- CLT
- Yet more normal distribution

Motivating Question

Question: Is turnout in 1972 distributed normally?

To answer this question...

- generate a *theoretical* normal density plot using the mean and SD from the turnout data.
- generate an *empirical* density plot using the turnout data.
- visually compare the two to see whether they look the same

Empirics versus theory

theoretical: based on theory. e.g. when working with the normal distribution, the density one gets by plugging x values into the normal density function $\left(\frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right)$

empirical: based on data/evidence. e.g. saying something about the turnout data only based on the data, without making any assumptions about whether it is normal or not.

What is the rationale of this approach?

If the turnout data are basically normal...

- a *theoretical* normal density curve generated using the mean and SD of the turnout data should be similar to the *empirical* density curve fit to the data
- there are more sophisticated ways of seeing whether some data are basically normal, but this is good enough for our purposes.

Generating the theoretical density

First, find the mean and SD of the data:

```
> mu1972 = mean(t1972)
> sig1972 = sd(t1972)
```

Define the turnout (X) values we will plot – e.g. 30 - 80%

```
> Xs = 30:80
```

Use above to generate theoretical density Y values.

```
> Ys.theory = dnorm(Xs, mu1972, sig1972)
```

Generating the empirical density

We want to compare the theoretical normal density with the empirical density. To generate latter, use the function **density**

```
> Pts.emp = density(t1972)
```

- We are basically fitting a curve to the data, based *only* on the data, with no assumptions about whether it's normal
- There's a lot going on under the hood when one calls 'density,' but we don't need to worry about that.

Plotting the empirical and theoretical densities

To plot the two densities on the same graph, first plot the theoretical density, then add the empirical density to the same plot.

```
> plot(Xs, Ys.theory, type='l')
> lines(Pts.emp, col='red')
```

The lines command

lines(x, y, ...) adds lines to an existing plot.

- takes most of the same arguments as 'plot'
- calling lines without first calling 'plot' will result in an error.

Normally, both 'plot' and 'lines' require vectors of X and Y coordinates for args 1 and 2. However, the return value from 'density' can be plugged directly into 'plot' and 'lines'

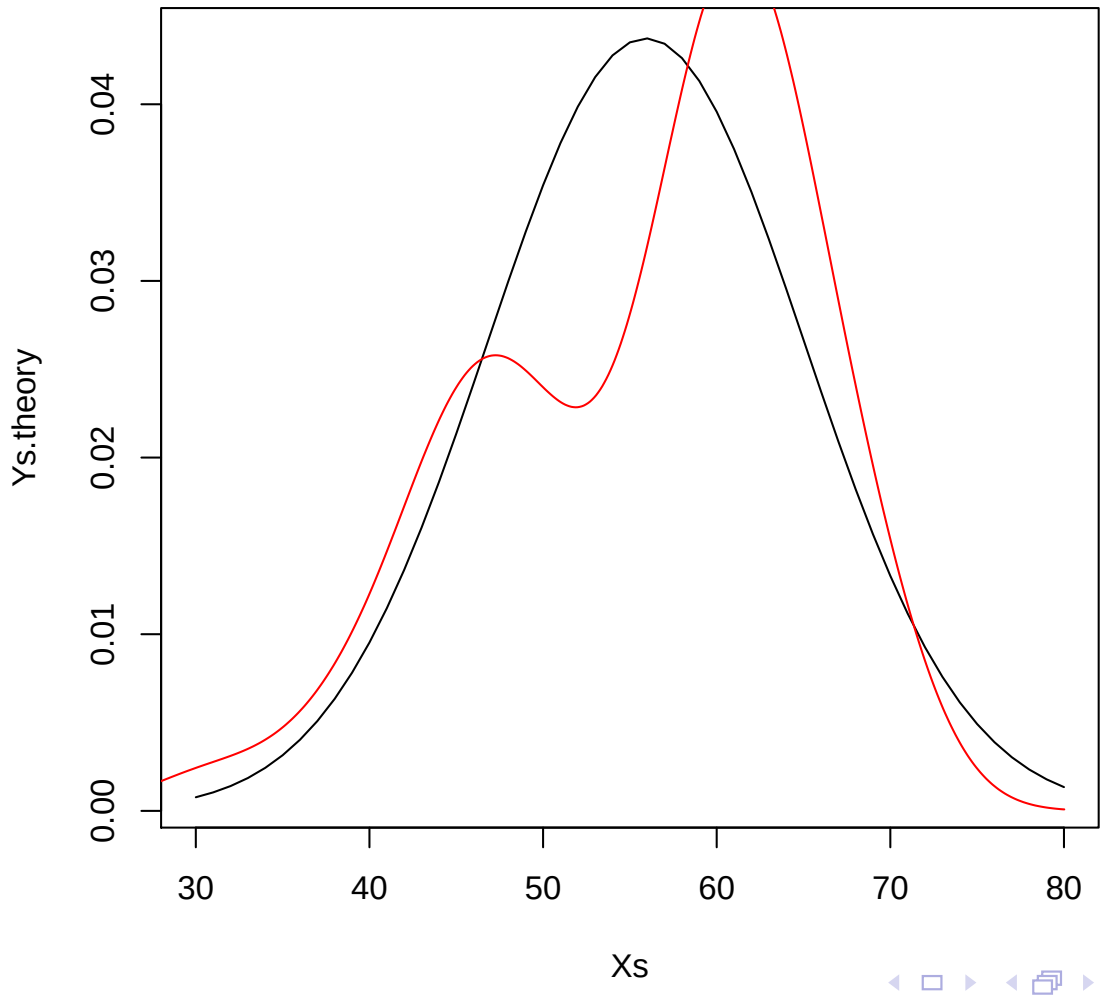
To see why this works, try calling 'str' on 'Pts.emp'

```
> str(Pts.emp)

List of 7
 $ x         : num [1:512] 19.6 19.7 19.8 19.9 20.1 ...
 $ y         : num [1:512] 2.35e-05 2.58e-05 2.83e-05 3.11e-05 3.41e-05
 $ bw        : num 3.74
 $ n         : int 51
 $ call      : language density.default(x = t1972)
 $ data.name: chr "t1972"
 $ has.na    : logi FALSE
 - attr(*, "class")= chr "density"
```

'density' returns a data.frame containing x and y columns. 'plot' has been programmed to recognize this.

Using the normal density plot to examine turnout



Prettying the plot

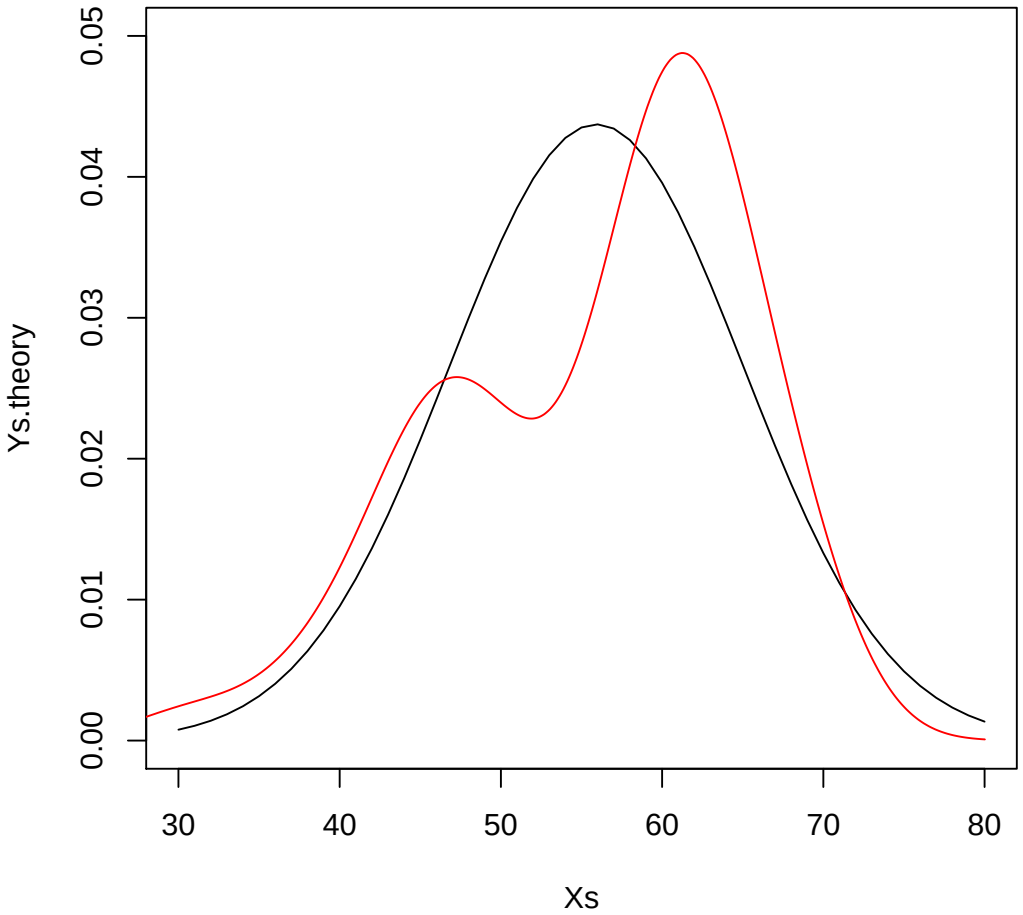
- Parts of the plot are getting cut off.
- Only the empirical curve is cut off. This is because the 'plot' was called using the theoretical data – the plot fits *that* data perfectly.
- Zoom in or out using the 'xlim' and 'ylim' arguments of plot.

Prettying the plot

We'll rerun the plot with ylim values so nothing gets cut off.

```
> plot(Xs, Ys.theory, type='l', ylim=c(0, 0.05))
> lines(Pts.emp, col='red')
```

- 'ylim' defines the min and max of the Y-axis. 'xlim' does the same for the X-axis.
- We pass it 2 element vectors, e.g. . . . ylim=c(0, 0.05) . . .



Voila. No more truncation.

What do we think about spread and centrality?

Do we think the raw 1972 data are normal?

Yes and no.

No, because. . .

- The curves do not overlap perfectly for turnout between 45% and 55%.
- We know that there turnout from different regions is clustered. The discrepancy IS NOT just due to chance.

Yes and no.

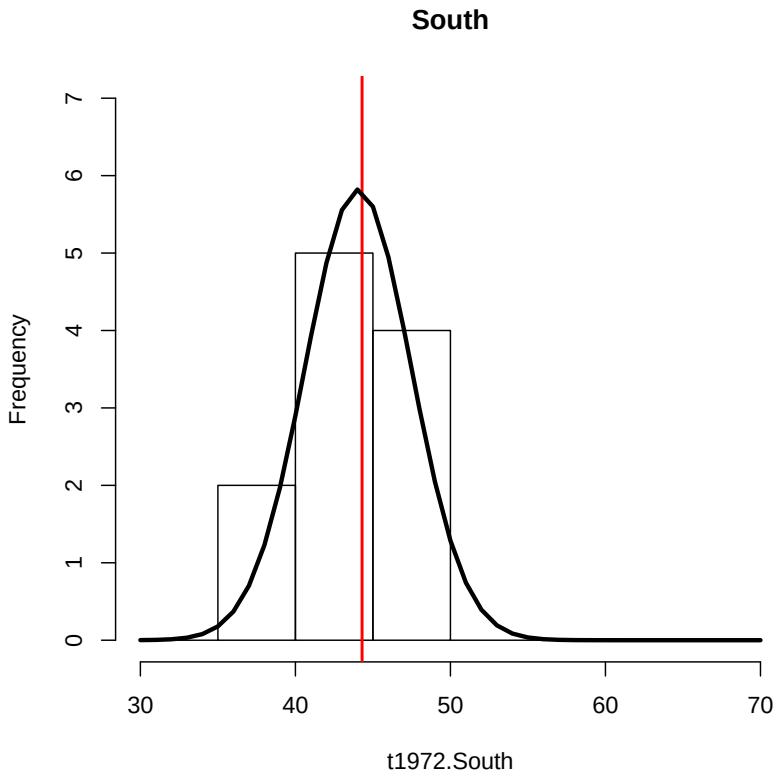
No, because. . .

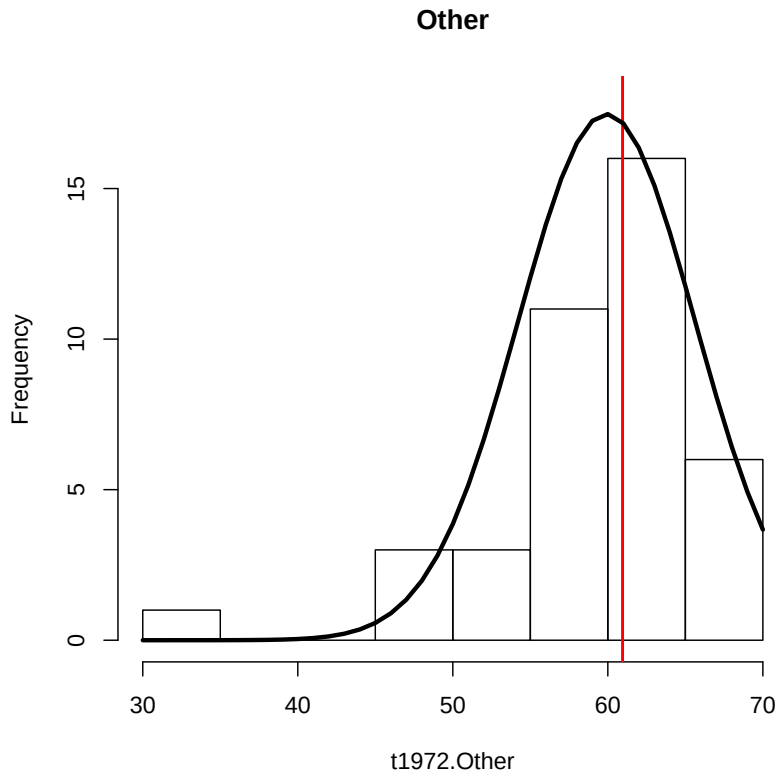
- The curves do not overlap perfectly for turnout between 45% and 55%.
- We know that there turnout from different regions is clustered. The discrepancy IS NOT just due to chance.

Yes, because. . .

- Patterns like this can arise just due to randomness in the data, esp. if there are only a few observations.
- Depending on the purpose of your analysis, a fit like this can be good-enough (approximately normal).

We've so far looked at the empirical versus theoretical density for overall turnout in 1972. Let's do the same by region – South and non-South.



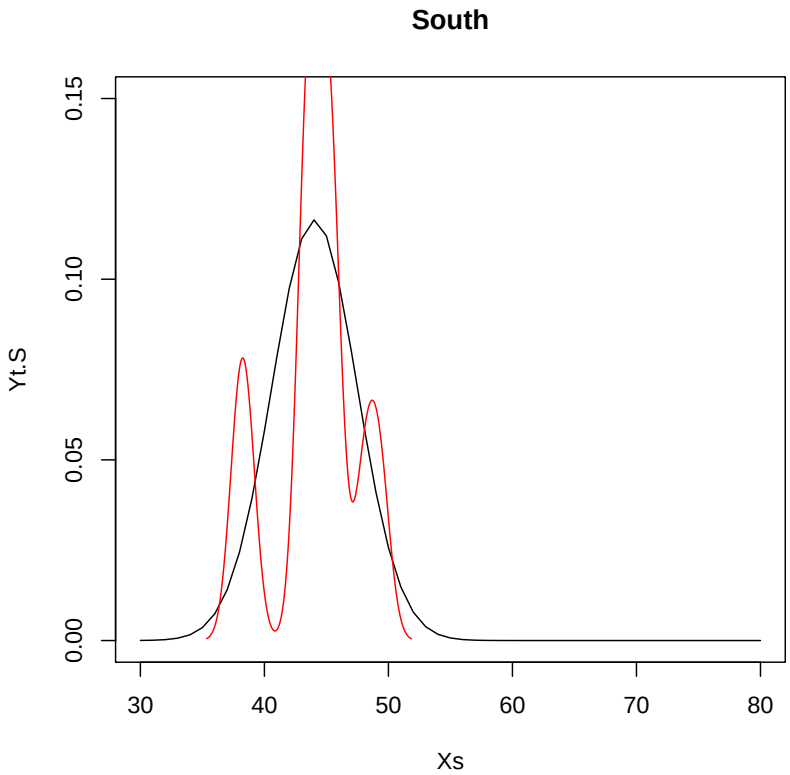


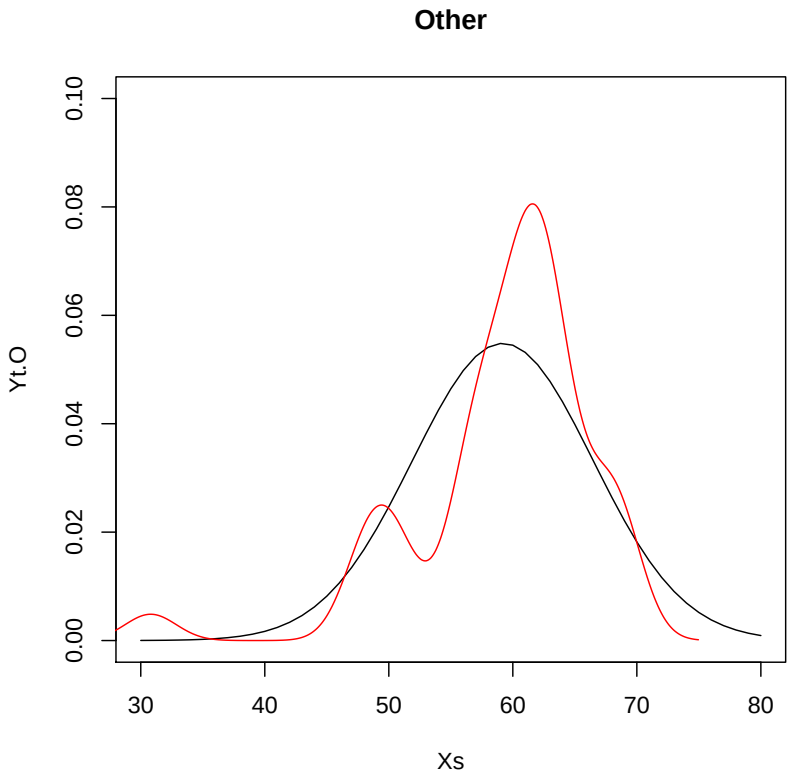
Just because the histograms look normal, will the empirical densities perfectly fit the theoretical densities?

```
> mu1972.S = mean(t1972.South)
> sig1972.S = sd(t1972.South)
> Yt.S = dnorm(Xs, mu1972.S, sig1972.S)
> Pe.S = density(t1972.South)
> plot(Xs, Yt.S, type='l', ylim=c(0, 0.15), main='South')
> lines(Pe.S, col='red')

> mu1972.O = mean(t1972.Other)
> sig1972.O = sd(t1972.Other)
> Yt.O = dnorm(Xs, mu1972.O, sig1972.O)
> Pe.O = density(t1972.Other)
> plot(Xs, Yt.O, type='l', ylim=c(0, 0.1), main='Other')
> lines(Pe.O, col='red')
```

Nope.


South
Yt.S
0.00
0.05
0.10
0.15
30
40
50
60
70
80
Xs


Other
Yt.O
0.00
0.02
0.04
0.06
0.08
0.10
30
40
50
60
70
80
Xs

The problem is the small number of observations.

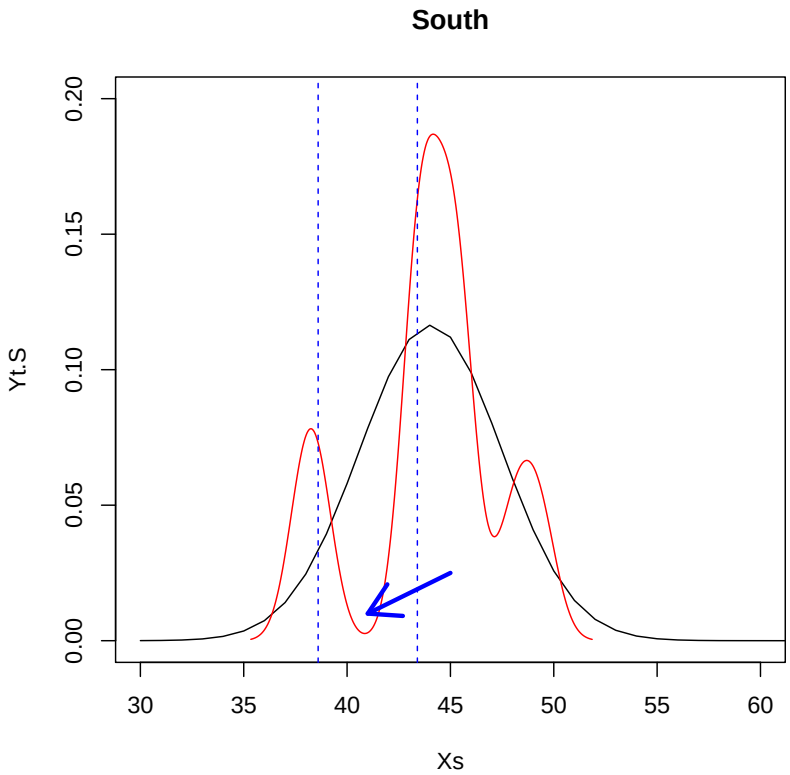
There are only 11 states in the South.

```
> sortedS = sort(t1972.South)
> sortedS
```

```
 [1] 37.9 38.6 43.4 43.4 43.6 44.3 45.0 45.4 45.5 48.1 49.3
```

See that gap between 38.6 and 43.4?

Empirical density with few observations



When we call 'density' the algorithm just assumes that the probability of the underlying data falling between 38.6 and 43.4 equals 0 (since there are no observations there). Hence, we see bumpy curves.

- Because we have so few observations, the graphs are ugly
- Nonetheless, even these ugly graphs strongly suggest that turnout overall, as well as by region, is well approximated by normal distributions
- Our case can be made stronger by turning to **qualitative** reasons to believe in normality of these data.

Two major questions:

1. What (if any) **qualitative** reasons are there to believe turnout is normally distributed?

2. Why do we *want* data to be normally distributed?

Question 1

I argue that the turnout data are basically normal because many similar phenomena are also normally distributed

Examples:

- standardized test scores
- height
- number of heads flipping multiple coins
- additive phenomena in general

Governed by the **central limit theorem**

The central limit theorem (CLT)

Definition: the mean of many *independent and identically distributed* (iid) random variables (RVs) approaches normality as sample size increases.

At least 2 concepts we need to introduce to understand the CLT

- random variables
- independence

Random Variables

A random variable (RV) is a variable that **can take on all possible values** for some underlying data-generating process

- e.g.: We believe there is a data-generating process that makes turnout in 1972 distribute normally around 55.93 with SD 9.12
- We would use a RV to represent ALL of the values that state turnout **could** have taken.

Random Variables

We can represent turnout as a random variable:

$$X \sim N(55.93, 9.12)$$

where X is the RV representing turnout.

Random Variables

The data that we have are **random** manifestations of a theoretical distribution of values that **could** have been obtained

- e.g. the first value of 'turnout' is 43.4 but it could just as easily have been 43.5.
- We use random variables to signify this element of chance.

Random Variables

- An RV describes where some data come from (i.e. the underlying distribution), but *are not* the data themselves
  - If it helps, think of data as a sample taken from a theoretical population that is described by a RV.
- RVs are written with capital letters, data with lower case
- Can be **continuous** or **discrete**

Continuous vs. discrete RVs

Continuous

- can take on any value within its range (i.e. infinite precision)
- represented by a probability density function (PDF)
- e.g. percent turnout

Continuous vs. discrete RVs

Continuous

- can take on any value within its range (i.e. infinite precision)
- represented by a probability density function (PDF)
- e.g. percent turnout

Discrete

- can only take a finite number of values (e.g. integers)
- represented by a probability mass function (PMF)
- e.g number of bills proposed in House

We're not going to worry too much about the distinction for now.

Population and Sample

Random variables are closely related to 2 important concepts

Population: All possible values that a phenomenon (i.e. something we'd represent with an RV) can take

Sample: A subset of a population. Data/observations are usually samples.

Population

The concept of population seems simple enough. . . but there are a couple of ways to think about it.

naive: population is all units we can take measurements for

- e.g. since we have turnout values for all 50 states, we have measured the entire population
- we don't have problems associated with samples

Population

realistic: population is all possible measurements that could have been made

- each measurement could have been made at an infinite number of other moments, by other measurers, subject to small variation in the underlying phenomenon, etc.
- sampling error always exists

Samples

Why do we sample?

- Cost – how much would it cost to administer a 30 minute survey to all 316M residents of US?
- Tractability – if we could get complete data, how hard would it be to manipulate and analyze all 316M observations?

Samples

Ideally, we want a **representative** sample – a sample that mirrors most important characteristics of the parent population

- **random measurement** is a widely used way of achieving a representative sample.
- there are situations where random measurement is not practical or desirable (e.g. to sample small subpopulations) – but this is a topic for a more advanced class.

Samples

Nonrandom sampling *can* lead to unrepresentative samples.

What kinds of biases might result from the following sampling schemes?

- political survey only within CA
- internet survey

Samples

Even with perfectly executed random sampling, there can/will be differences between statistics of the sample, and true population values

Luck of the draw. . .

- causes sample variance to be greater than pop. variance
- leads to **sampling error** – differences between population values and sample values

Std. error versus std. deviation

Finally, let's tighten up the vocab we use re: spread

std. deviation: spread in data

std. error: spread in statistics calculated from that data – e.g. means, medians, std. deviation, etc. Consequence of sampling.

Intro to Independence

Definition:

$$Pr(A \cap B) = Pr(A)Pr(B)$$

If 2 RVs/sets of data are independent, the value taken by 1 is not at all affected by the value of the other.

e.g. flipping 2 coins – Pr coin 1 is H is not affected by whether coin 2 is H or T.

We'll come back to this when we examine bivariate relationships.

OK, back to the CLT.

- Say we have 100 RVs ($X_1 - X_{100}$) representing 100 identical coins. We use coins because it's clear that they are independent.
- We flip each 100 times, and take the mean for each coin, giving us $\bar{X}_i$ for i from 1 to 100. The vector of $\bar{X}_i$s we'll call Y.

CLT says that if you histogram or take the density of Y, it will look approximately normal, and that as more and more coins are added, the more normal it will look.

We can try this in R using 'rbinom'.

'rbinom' is basically a coin-flipping function

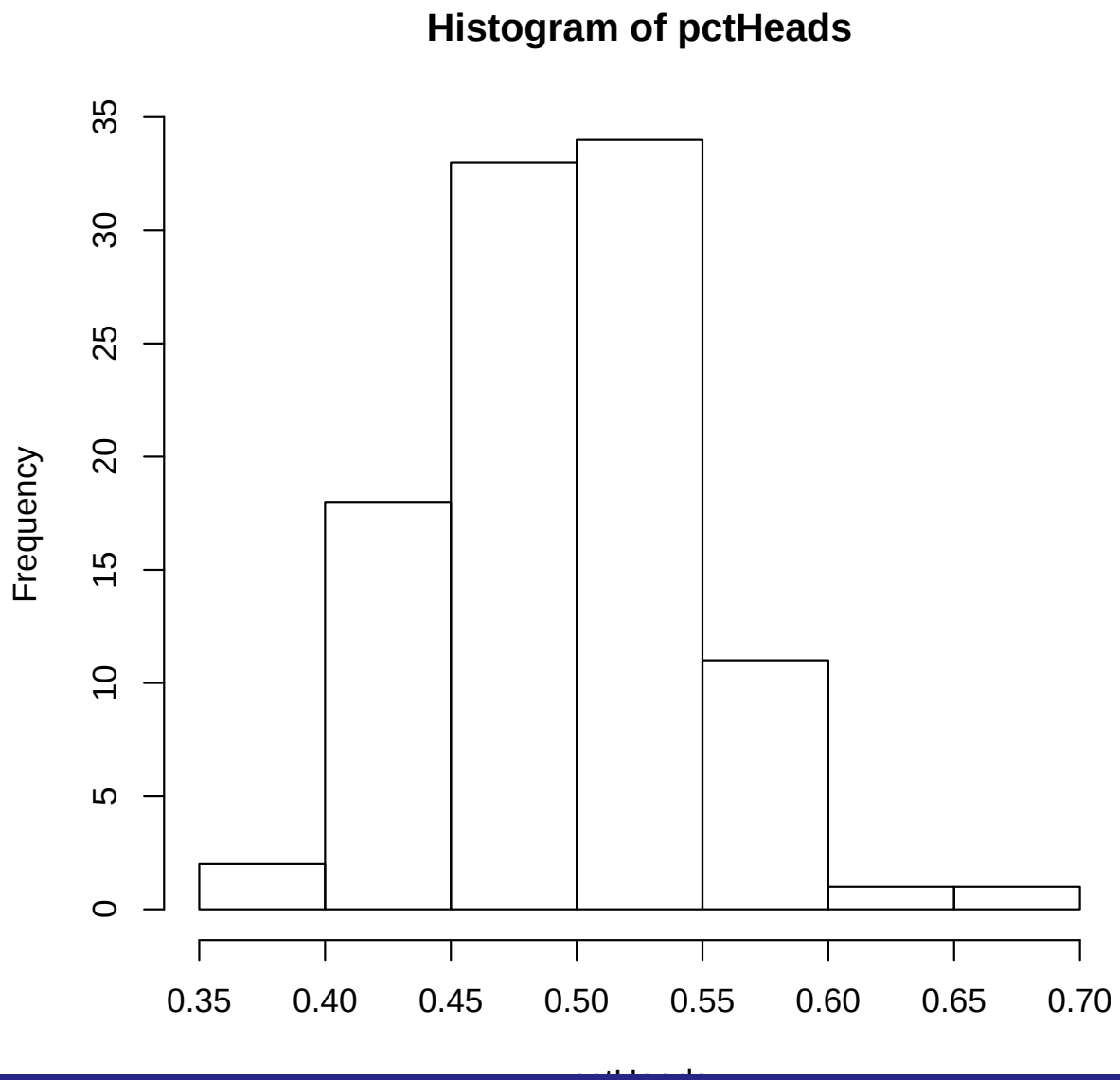
- arg 1: how many coins to flip
- arg 2: how many times to flip each coin
- arg 3: the probability of heads on each coin

Returns the number of heads for each coin that was flipped.

To demonstrate the CLT, we want to...

1. flip 100 fair (50% heads) coins, 100 times
2. divide each number of heads by the total number of flips to get % heads
3. look at the distribution

```
> numFlips = 100
> numHeads = rbinom(100, numFlips, 0.5)
> pctHeads = numHeads / numFlips
> hist(pctHeads)
```



Looks pretty normal!

As you add more 'coins', the more normal this will look.

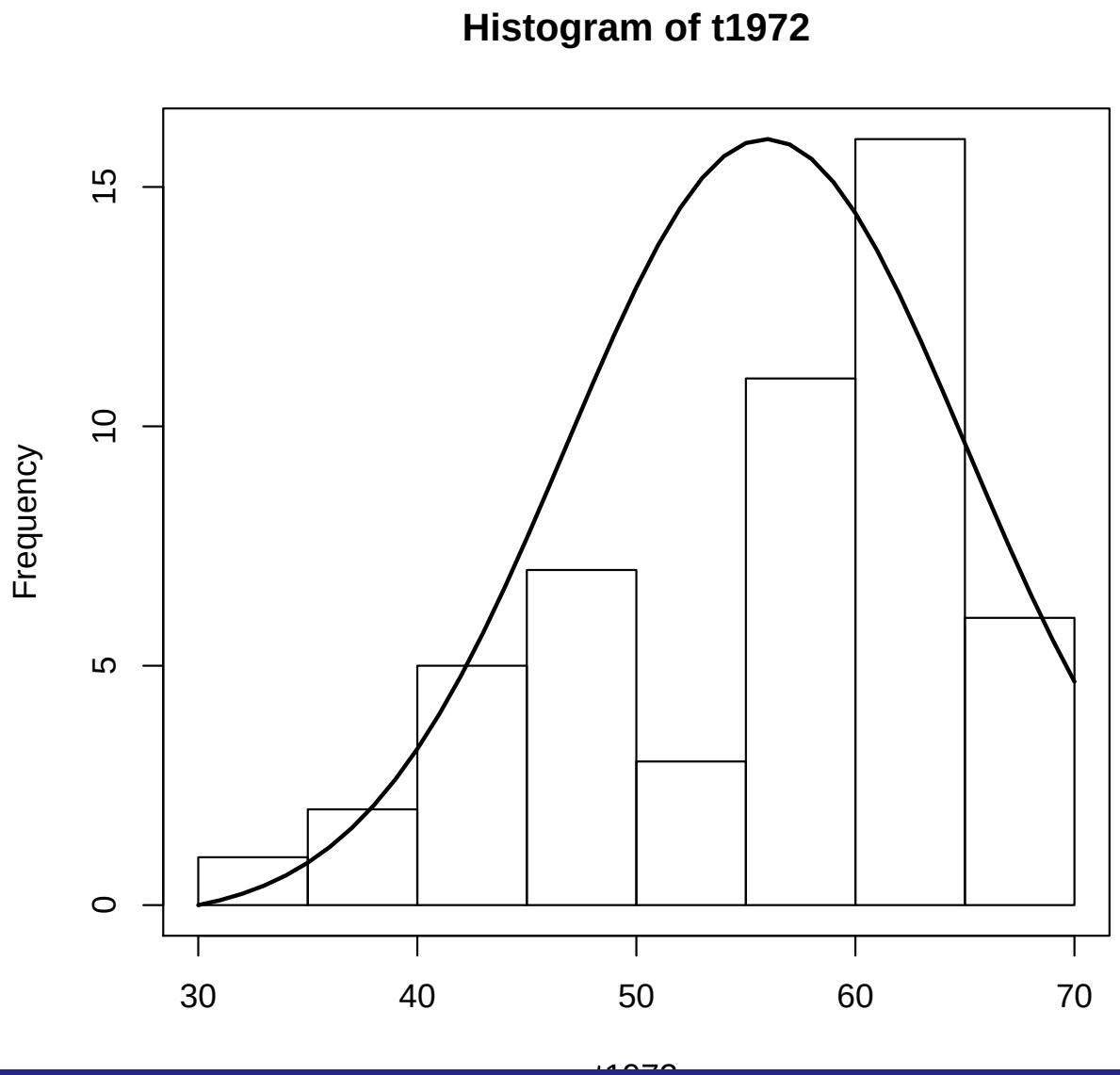
How does this help us understand the turnout data?

X_i: coin : H/T :: person : vote/no-vote

$\bar{X}_i$: % turnout by state, for states i

Y 'turnout' data are basically draws from Y

To the extent that all voters, regardless of state, are similar 'coins', we expect state-by-state 'turnout' for the entire country to look fairly normal.



OK – so voters by state are maybe not all THAT similar.

Why do I keep arguing that the South is separate from the rest?

"Throughout the first half of the twentieth century. . . turnout in nonsouthern states varied from 55 to 70 percent – more than double the southern figures,' The disfranchisement of blacks in the South after Reconstruction obviously contributed to low southern turnout. . . Low voting turnout was an important characteristic of the white electorate, too. . . many citizens lost interest in politics and disfranchised themselves because of lack of party competition in the South" (Cassel 1979, p.907)

Thus, we have a qualitative reason to think that the Pr(vote) for Southern persons is lower than that for non-Southerners

- In statistical terms, just Southern persons are closer to "identically distributed" then all Americans.
- Resultantly, each regional subset looks fairly normal, while all Americans pooled together (regardless of region) are less so
- The more similar the underlying persons, the less we should see divergence between the two groups of voters.

An aside on the process of discovery

Our reasoning process so far: We...

- noticed a pattern in data
- came up with a qualitative explanation (i.e. a theory)

An aside on the process of discovery

Now what?

- question or accept the qualitative exlanation based on how believeable it is – this is a subjective judgement
- test the explanation. If it were true, what else would have to be true? Do we observe these "what else"s?

This is the scientific method in a nutshell. Hence, we can call poli-sci and similar disciplines social *sciences*.

Question 2

Why do we like things that are normally distributed?

It's very well understood and easy to work with.

- many phenomena are approx. normally distributed
- easy and inutitive parameterization
- 68–95–99.7 rule

Easy parameterization

We saw in lecture 3 that the normal dist. is parameterized by mean μ and SD σ – easy values to think about and estimate.

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

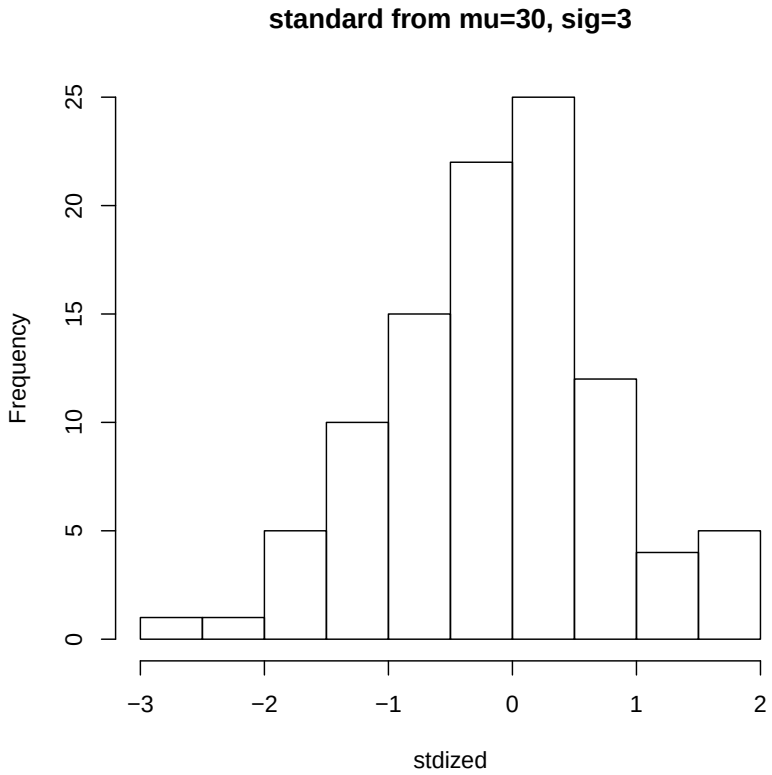
Another result of this parameterization is that it's easy to go from any normal distribution to the standard normal and vice versa:

Non-standard to standard: subtract μ, then divide by σ

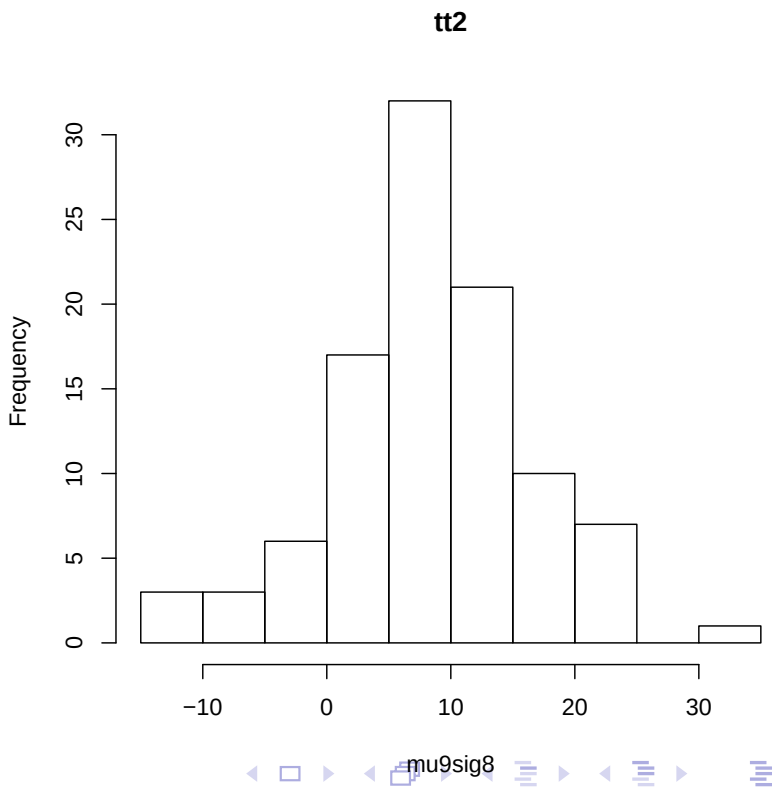
Standard to non-standard: multiply by σ, then add μ

Fitting a normal distribution

```
> mu30sig3 = rnorm(100, 30, 3)
> stdized = (mu30sig3 - 30)/3
> tt1 = 'standard from mu=30, sig=3'
> hist(stdized, main=tt1)
```



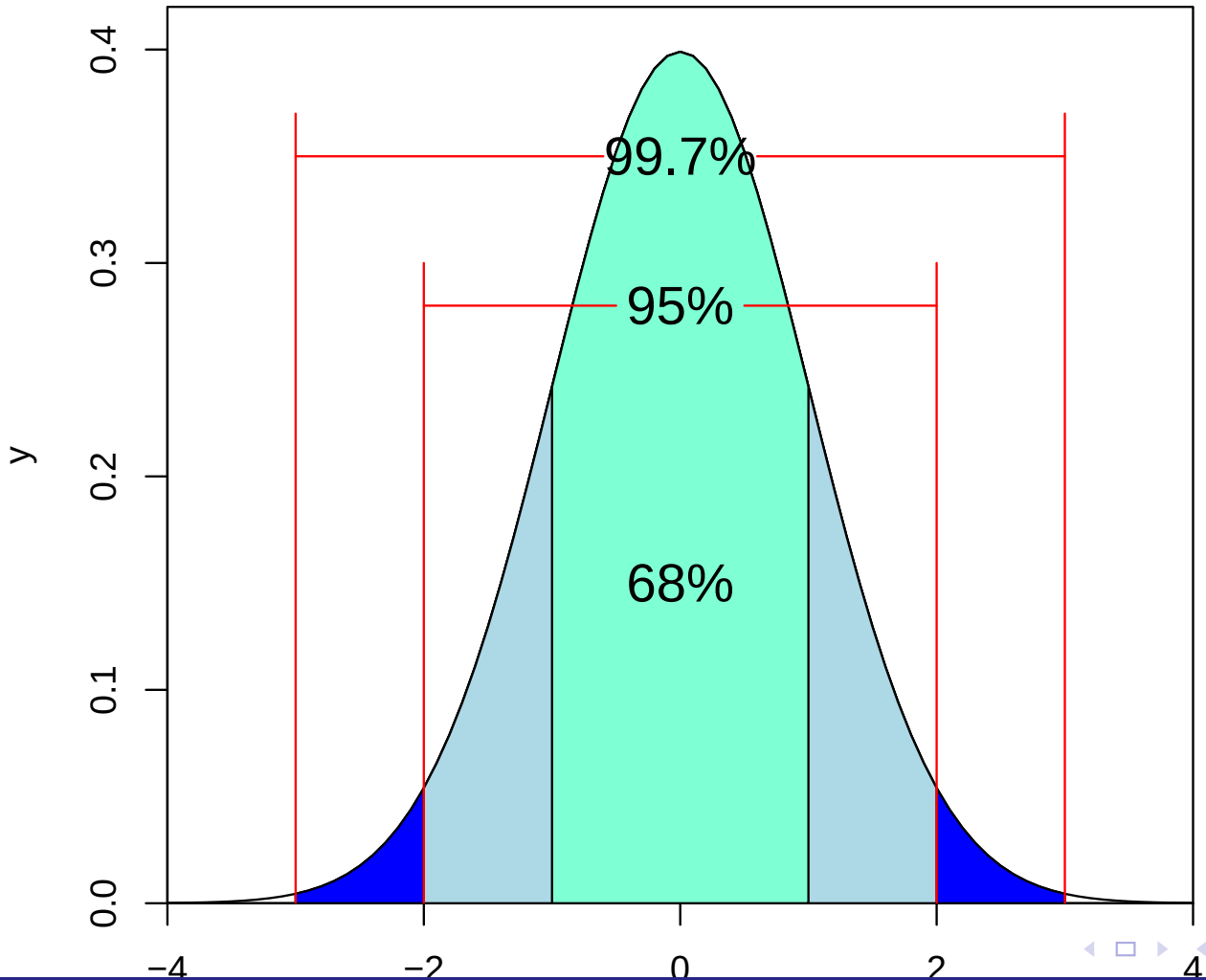

```
> standard = rnorm(100, 0, 1)
> mu9sig8 = (standard*8)+9
> tt2 = 'mu=9, sig=8 from standard'
> hist(mu9sig8, main='tt2')
```



68–95–99.7 rule

For any data that are normally distributed...

- 68% of observations lie within 1 SD of the mean.
- 95% of observations lie within 2 SD of the mean.
- 99.7% of observations lie within 3 SD of the mean.


y
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
99.7%
95%
68%

Why is the 68–95–99.7 rule (and the underlying idea of the PDF/CDF) so useful?

Once we determine that something is more or less normal, we can make statements about likely values!

- If we believe Southern turnout to be normally distributed with mean 44.05 & SD 3.43, we can say that 95% of Southern states will have turnouts in the range 37.19 to 50.91.
- Shown a state with unknown region but turnout 57.77 (i.e. $+4\sigma$), we would say that it is unlikely to be Southern (though not impossible).