PS6 Lecture 5

Daniel Lim

University of California, Los Angeles

4/14/2014

Agenda

- t-distribution
- test of differences in mean
- New data set: GDP/capita
- Transforming variables

Road so Far

1. Learned stats essentials: summary statistics, histograms, boxplots, density
2. Studied normal distribution in theory
3. Applied (1) and (2) to study turnout data

However, normal distribution is not the only one you'll encounter while doing data analysis.

A few other univariate distributions

discrete

poisson: count data parameterized by single mean/variance (e.g. # of emails per unit time)

binomial: number of successes in a sequence of N independent yes/no experiments (e.g. coin flips)

uniform: equal probability of getting each of N values (e.g. roll of a fair die, single fair coin flip)

A few other univariate distributions

continuous

uniform: equal probability of getting any value within its range (e.g. probability)

beta: rate of success given past successes and failures (e.g. coin that is unknown whether fair or not)

student's T: standard normal with heavier tails (e.g. black swan events)

The t-distribution

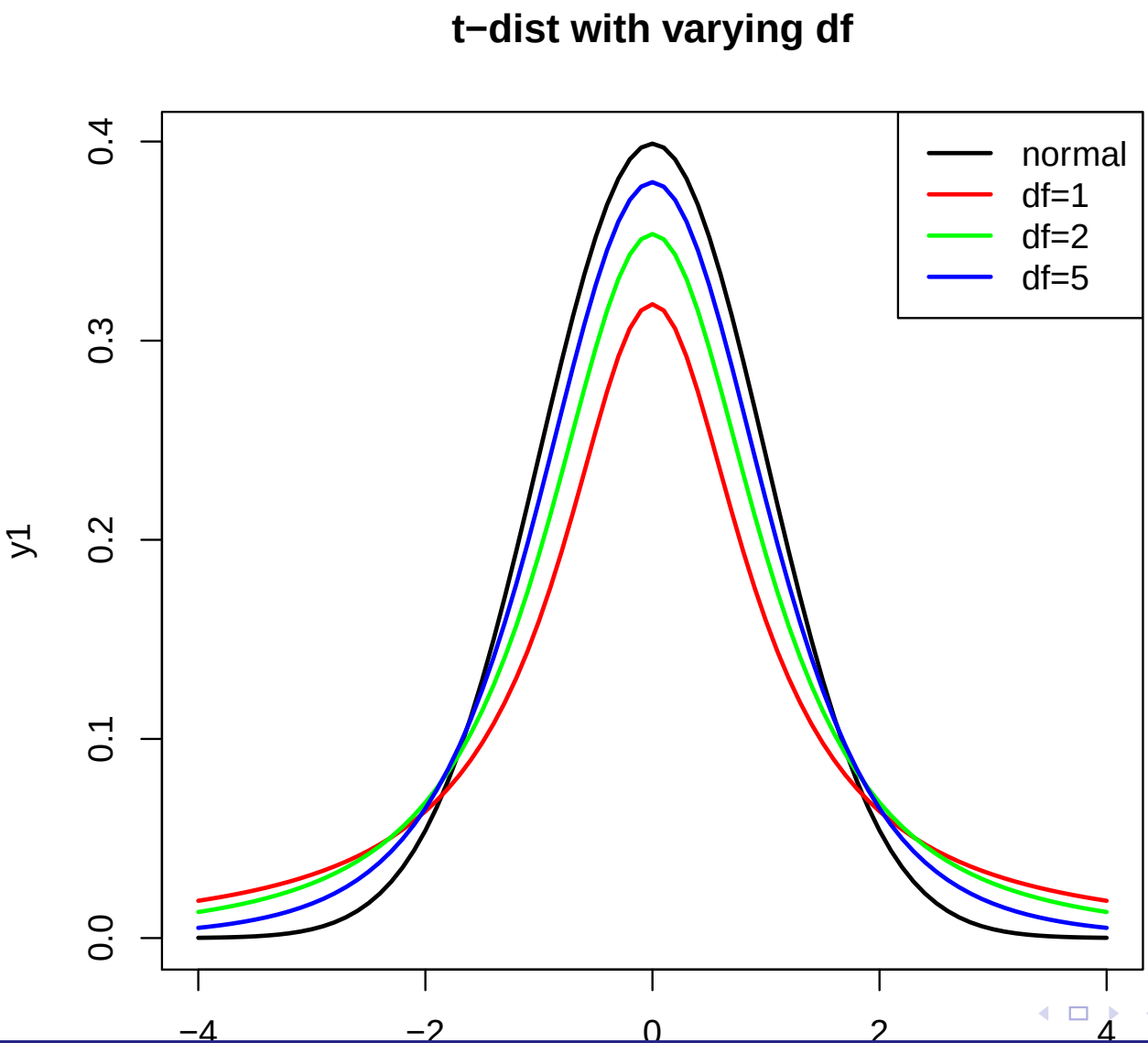
We're going to look a bit more closely at the t-distribution because it is used for a common statistical test that we will find useful.

Characteristics:

- single parameter: degrees of freedom (df)
- looks normal, but has thicker tails
- approaches normality as $df \to \infty$

The following code lets us visually compare the PDFs of t distributions with varying df with that of the normal

```
> x = seq(-4, 4, 0.1)
> y1 = dnorm(x, 0, 1)
> y2 = dt(x, 1)
> y3 = dt(x, 2)
> y4 = dt(x, 5)
> cols=c('black', 'red', 'green', 'blue')
> plot(x, y1, col=cols[1], type='l',
+  lwd=2, main='t-dist with varying df')
> lines(x, y2, col=cols[2], lwd=2)
> lines(x, y3, col=cols[3], lwd=2)
> lines(x, y4, col=cols[4], lwd=2)
> legend('topright', c('normal', 'df=1', 'df=2', 'df=5'),
+  col=cols, lwd=rep(2, 5))
```


t−dist with varying df
normal
df=1
df=2
df=5
y1
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4

Our Specific Question

Where we left off: Is turnout in Southern states different from that of non-Southern states in a way that substantively matters?

Previous approach:

1. visually compare theoretical and empirical density of data
2. give hand-wavy justification

Our Specific Question

Where we left off: Is turnout in Southern states different from that of non-Southern states in a way that substantively matters?

Previous approach:

1. visually compare theoretical and empirical density of data
2. give hand-wavy justification

New approach:

1. statistically test hypothesis that they are different
2. make an actual probabilistic statement

Question in More General Terms

More generally, given 2 normally distributed samples with...

- parameters $\mu_1, \mu_2, \sigma_1, \sigma_2$
- sample sizes N_1, N_2

are these samples from the same population?

Test of difference in means

To answer the question, we will use a **2 tailed t-test** (AKA test of difference in means)

Given some hypothesis we want to test...

1. if some basic assumptions met...
2. probability of hypothesis being supported is t-distributed...
3. ...allowing us to make probabilistic statement about hypothesis

Test of difference in means

Step 1: define null and alternate hypotheses

null: the samples are drawn from populations with the same mean

alternate: they are from different populations

Test of difference in means

Step 2: calculate t, df, $P_{|t|,\nu}$ and the p-value for the test

- $P_{|t|,\nu}$ is the value of the CDF at $abs(t)$ for a t-dist. $df = \nu$
- The *p-value* for the null hypothesis is $2(1 - P_{|t|,\nu})$

The p-value is the probability of obtaining the test statistic (t) assuming the null hypothesis is true.

Test of difference in means

Step 3: draw conclusion about hypothesis

- If calculated p-value lower than α, the significance level of the test[1], "reject the (null) hypothesis that the samples are from a population with the same mean at the α level."
- Else, data are insufficient to distinguish the population of one sample from the other.

[1] typically, 0.05

To do the calculations...

1. calculate:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \qquad \nu = \frac{(\sigma_1^2 + \sigma_2^2)^2}{\frac{\sigma_1^4}{N_1 - 1} + \frac{\sigma_2^4}{N_2 - 1}}$$

2. calculate $P_{|t|,\nu}$. In R, 'pt(abs(t), ν)'
3. calculate p-value: $2(1 - P_{|t|,\nu})$
4. If p-value $< \alpha$, reject null hypothesis

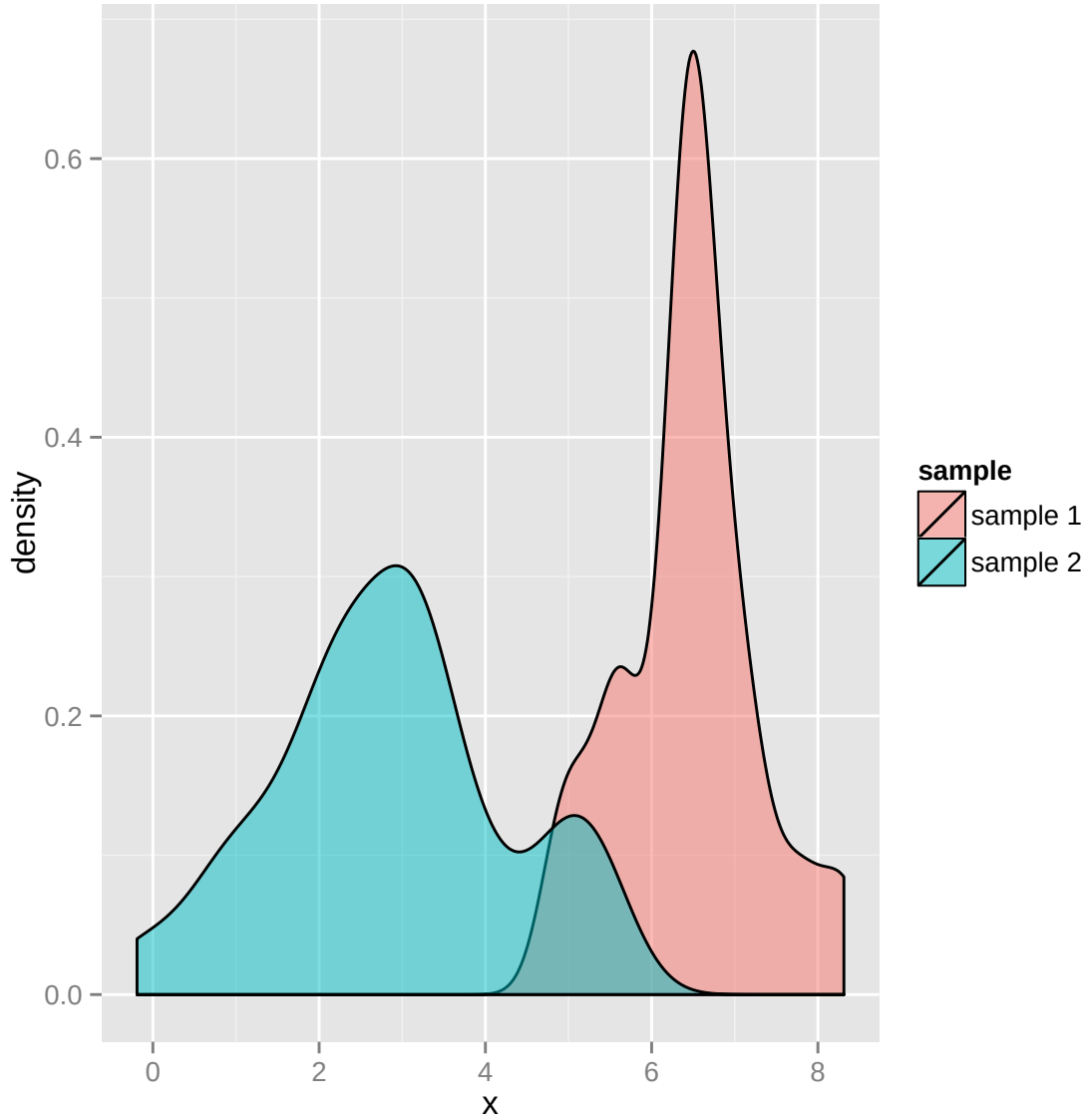
Example 1

Suppose we have 2 samples with the following params:

$$\mu_1 = 6, \sigma_1 = 1, N_1 = 20$$

$$\mu_2 = 3, \sigma_2 = 2, N_2 = 30$$

```
> mu1 = 6; sig1 = 1; n1 = 20
> mu2 = 3; sig2 = 2; n2 = 30
```


density
0.6
0.4
0.2
0.0
0
2
4
6
8
x
sample
sample 1
sample 2

Example 1

Can we reject the null hypothesis that these samples are drawn from populations with the same mean?

Complete the test by running the following code:

```
> T = (mu1 - mu2)/sqrt(sig1^2/n1 + sig2^2/n2)
> nu = (sig1^2 + sig2^2)^2/(sig1^4/(n1-1)+sig2^4/(n2-1))
> pval = 2*(1-pt(abs(T), nu))
```

Example 1

And the result...

```
> pval
[1] 1.540164e-08
```

Since 'pval' is less than 0.05, we can say that these two samples are probably not drawn from the same population.

A note on phrasing

These are accurate and essentially equivalent statements.

- "if I were to redo this test over and over again with new samples, $1 - \alpha\%$ of those would show [some number] to be non-zero."
- "Because the p-value is below α, I reject the null hypothesis at the α level."

A note on phrasing

These are accurate and essentially equivalent statements.

- "if I were to redo this test over and over again with new samples, $1 - \alpha\%$ of those would show [some number] to be non-zero."
- "Because the p-value is below α, I reject the null hypothesis at the α level."

This statement is NOT the same, nor correct.

- "Because the p-value is below α, the alternate hypothesis **is/must** be true"

A note on phrasing

So what's the logic behind the phrasing?

- These tests are set up to tell us what CANNOT be.
- They do not directly tell us what MUST be.
- We get to what is LIKELY to be true by eliminating as many possible alternate explanations as possible.

A note on phrasing

So what's the logic behind the phrasing?

- These tests are set up to tell us what CANNOT be.
- They do not directly tell us what MUST be.
- We get to what is LIKELY to be true by eliminating as many possible alternate explanations as possible.

This is not the way we are accustomed to thinking nor the easiest way to make arguments. However, given the tools we do have, you must understand these disctinctions.

Shortcomings of example 1

If you play around with example 1, what you'd eventually find is that it's not a terribly interesting test.

- the null hypothesis is that samples are from populations with the *exact same mean*
- even if population means were different by a tiny amount, the test *could* give you a statistically significant result

Shortcomings of example 1

- *More interesting:* 'sample 1 is from a population that is at least X units different from that of sample 2.'
- We can modify the test slightly to test whether the means of the populations that the samples are drawn from differ by at least Δ

(new) null hypothesis: the 2 samples are drawn from populations with means that are at least Δ apart.

The only difference is in the numerator of T:

$$t = \frac{\bar{X}_1 - \bar{X}_2 - \Delta}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$

Everything else is the same.

Example 2

Let's see whether the two samples from example 1 are drawn from populations with a difference in means of at least 2.5:

```
> delta = 2.5
> T = (mu1 - mu2 - delta)/sqrt(sig1^2/n1 + sig2^2/n2)
> nu = (sig1^2 + sig2^2)^2/(sig1^4/(n1-1)+sig2^4/(n2-1))
> pval = 2*(1-pt(abs(T), nu))
```

Example 2

Resulting p-value:

```
> pval
```

```
[1] 0.2495951
```

- because we generated these fake data, we **know** samples drawn from populations with means 3 apart ($\mu_1 = 6, \mu_2 = 3$)
- however, test **suggests** we *cannot* tell whether they are from populations with means even 2.5 apart
- due to randomness of the data and the small sample sizes

How might we analyze turnout data with this test?

One way:

- calculate Δ_{max} across time
- see how that figure changes across time
- What would we say if. . .
  - . . . increasing?
  - . . . decreasing?
  - . . . stable?

We can answer many social science questions using this test.

- How has the distribution of income changed?
- Is group A of similar partisanship as group B?
- What is the effect of policy X on characteristic Y?

Test of differences in means is not always the best way to answer such questions, but it certainly provides one way.

New Dataset: GDP/capita

Let's introduce a new data set.

GDP per capita by country, 1966-1997

- Download 'World_GDP_1966_97.csv' from the course website.
- Refer to lecture notes 1 for loading instructions.

New Dataset: GDP/capita

In case you're feeling lazy...

```
> library(foreign)
> f = 'c:/users/daniel/dropbox/ps6/data/World_GDP_1966_97.csv'
> gdp = read.csv(f, stringsAsFactors = F)
```

Once loaded, you should do all the basic exercises from lecture 1 to acquaint yourself with these data.

Shape of data

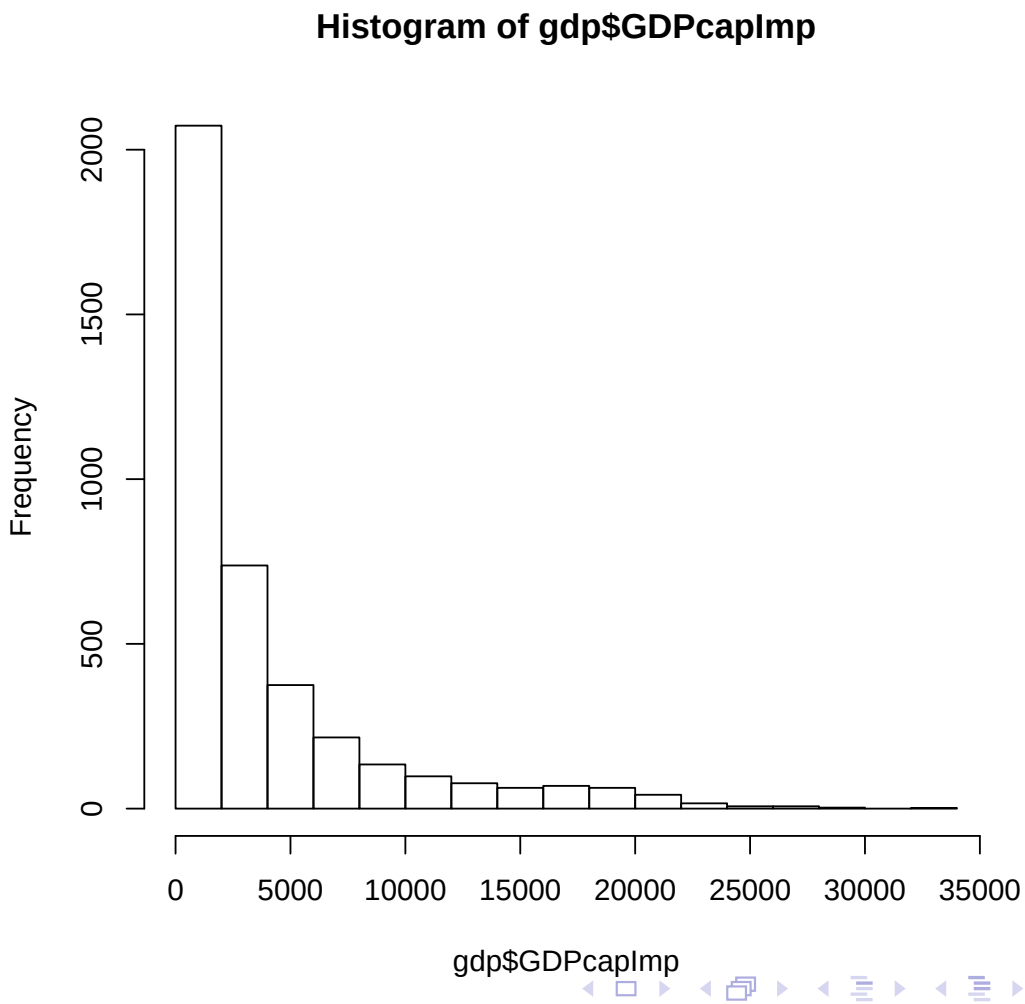
One of the first things you will notice is that the 'gdp' data have a different shape from the 'turnout' data.

- 'turnout' has 1 row for each state, with 1 column for each election year
- 'gdp' has N rows each country, where N is the number of years. Each row represents a country-year.

'turnout' is in **wide form**, while 'gdp' is in **long form**. Each configuration has pros/cons, which we'll encounter.

```
> hist(gdp$GDPcapImp)
```

Let's take a look at the column 'GDPcapImp'. These are GDP/cap vals for all countries '66-'97.



Skew

Such data are described as **skewed**

- We say the data are skewed to the side with the longer tail
- 'GDPcapImp' is skewed to the right

The histogram is not as helpful as it can be because most of the observations are clustered on the left, but we can't see what's going on there.

The log transform

In many cases, we can fix **right skew** through the use of a **log transform**

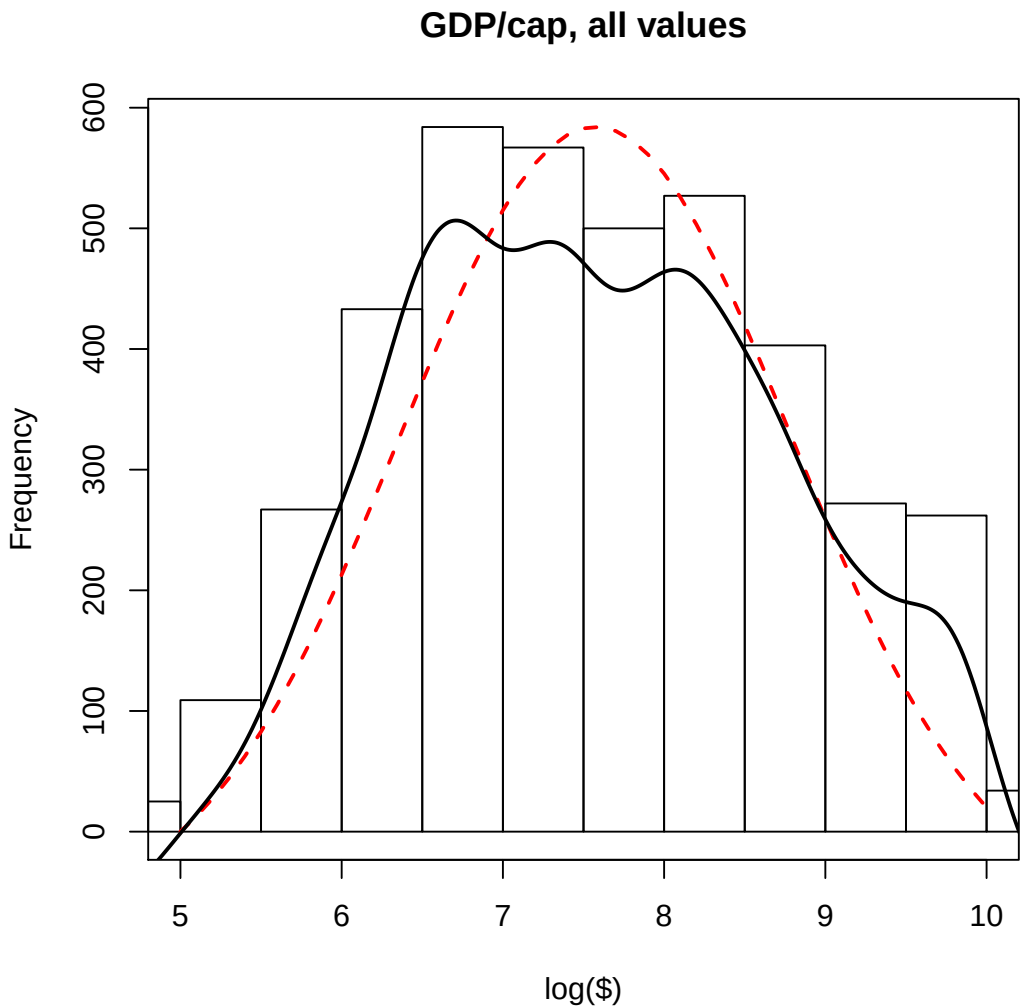
- Take the natural log of the data.
- Has the effect of decompressing smaller numbers and compressing larger ones.
- If resulting data are approx. normal, we say the original data are **log-normal**

The log transform

We'll transform 'GDPcapImp', then take a look at it.

```
> xt = log(gdp$GDPcapImp)
> xlims = round( range(xt), 0)
> xs = seq(xlims[1], xlims[2], 0.1)
> ys = dnorm(xs, mean(xt), sd(xt))
> hist(xt, xlim=xlims, main='GDP/cap, all values', xlab='log($)')
> par(new=T)
> plot(xs, ys, xlim=xlims, type='l', xaxt='n', yaxt='n',
+  main='', xlab='', ylab='', col='red', lwd=2, lty=2)
> lines(density(xt), col='black', lwd=2, lty=1 )
```

The log transform



Looks pretty normal. Depending on our purpose, it's probably safe to call these data log-normal.

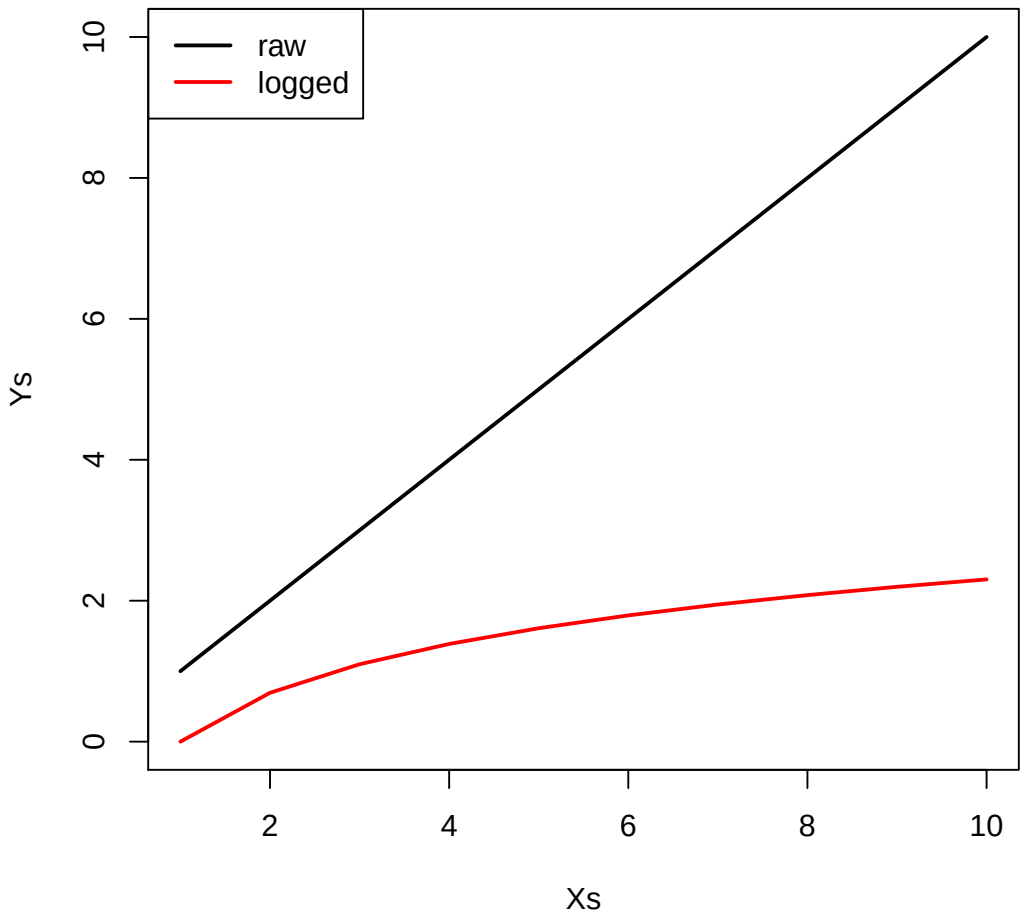
The log transform

Why does the log transform work?

Let's experiment with some easy numbers.

```
> Xs = 1:10
> LXs = log(Xs)
> plot(Xs, Xs, type='l', ylim=c(0, 10), ylab='Ys', lwd=2)
> lines(Xs, LXs, col='red', lwd=2)
> legend('topleft', c('raw', 'logged'),
+  col=c('black', 'red'), lwd=c(2, 2))
```

The log transform



Numbers that were small are shrunk less than numbers that were large

So, we get a decompressing/compressing effect.

The log transform

Looks like it can make our graphs look normal – so what?

2 major scenarios in which to use the log transform

- examine percent change
- reduce leverage of particular observations in statistical regression

We'll talk about point 2 in lectures re: regression. For now, let's look at percent change.

What do log units mean on a normal scale?

To illustrate: define a vector 1:4, and say these are the on the log scale. To get normal scale numbers, we take their exponent (because $exp(log(x)) = x$).

```
> logXs = 1:4
> Xs = exp(logXs)
> Xs

[1]  2.718282  7.389056 20.085537 54.598150

> log(Xs)

[1] 1 2 3 4
```

Log units transform back to normal units in a nonlinear form.

Log scale: difference between all elements is simply 1

- $2 - 1 = 1$
- $3 - 2 = 1$
- $4 - 3 = 1$

Regular scale: difference between elements changes

- $7.39 - 2.72 = 4.67$
- $20.09 - 7.39 = 12.7$
- $54.6 - 20.09 = 34.51$

No obvious pattern to increases of 4.67, 12.7, 34.51. However, there is more than meets the eye.

```
> (Xs[2] - Xs[1])/Xs[1]

[1] 1.718282

> (Xs[3] - Xs[2])/Xs[2]

[1] 1.718282

> (Xs[4] - Xs[3])/Xs[3]

[1] 1.718282
```

Percent change remains constant!

The log scale can be interpreted as a measure of **percent change**

- a change of 1 log unit is the same percent change whether moving from 1 to 2, or 101 to 102.
- magnitude matters (obviously) – moving from 0.1 to 0.2 is the same % change as 4.5 to 4.6, but not as 1 to 2.
- e.g. a country with log(GDP/cap) of 8 (i.e. \$2980.96) is the same percent bigger than one with 7 (i.e. \$1096.63), as 9 (i.e. \$8103.08) versus 8.

A few more thoughts on transformation

- Can't log numbers less than 0. If raw data don't meet that criteria, you can modify them so that they do.
- For data that are **left-skewed**, unskew by taking the square root (instead of logging)
- When you want to describe data in writing, you still need to use the original scale (i.e. log($) is not an intuitive unit)