

# Math 2311

Bekki George – [bekki@math.uh.edu](mailto:bekki@math.uh.edu)

Office Hours: MW 11am to 12:45pm in 639 PGH

★ Online Thursdays 4-5:30pm ★

And by appointment

Class webpage: <http://www.math.uh.edu/~bekki/Math2311.html>

What is the difference between a discrete random variable and a continuous random variable?

↑  
Countable

can take on any value in an interval

How do we know when we are working with  $\bar{X}$  vs.  $X$ ?

↑  
mean  
average  
st. dev. =  $\frac{\sigma}{\sqrt{n}}$

What is  $\hat{p}$ ? The statistic  $\hat{p}$  is an estimate of the parameter  $p$ .  $\hat{p} = X/n$

$$\mu = p$$

Since values of  $X$  and  $\hat{p}$  vary in repeated samples, they are both random variables.  $X$  will follow a binomial distribution (provided that the population is more than 10 times the sample) BUT  $\hat{p}$  does not. The sampling distribution of  $\hat{p}$  will follow the normal

distribution with mean  $p$  and standard deviation  $\sqrt{\frac{p(1-p)}{n}}$  (provided our conditions are met - see last week's notes)

$$\sigma$$

Some more practice:

$$\mu = p \quad \sigma = \sqrt{\frac{p(1-p)}{750}}$$

1. Suppose that you are interested in the proportion of all registered voters who intend to vote in the next election, call this (parameter) proportion  $p$ . In practice, you would never know  $p$ , but you could estimate it based on a sample proportion. Assume for now, though, that you know  $p = .6$ , and suppose that you plan to interview a simple random sample of  $n = 750$  registered voters and to ask each whether she or he intends to vote.

$$\mu = .6 \quad \sigma = \sqrt{\frac{.6(1-.6)}{750}} = .0179$$

a.) What does the Central Limit Theorem say about the sampling distribution of this sample proportion?

$$750(.6) = 450$$
$$750(1-.6) = 300$$

$$\left. \begin{array}{l} np \geq 10 \\ n(1-p) \geq 10 \end{array} \right\} \Rightarrow \text{use normal distr. } N(.6, .0179)$$

b.) Use this result and the table of standard normal probabilities to find the probability that the sample proportion intending to vote would fall within .029 of  $p$ .

$$P(.6 - .029 \leq \hat{p} \leq .6 + .029)$$

$$P(.571 \leq \hat{p} \leq .629) = .895$$

- normal cdf (.571, .629, .6, .0179)

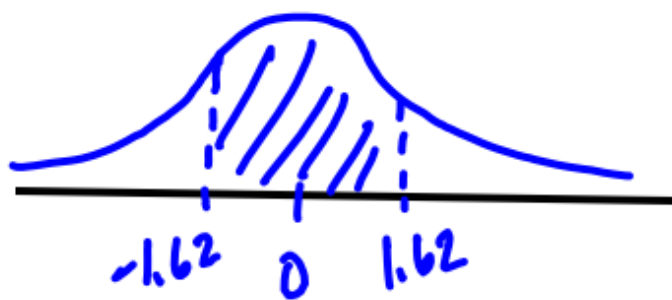
$$pnorm(.629, .6, .0179) - pnorm(.571, .6, .0179)$$

If I wanted to use the table, we need z-scores

$$z = \frac{x - \mu}{\sigma}$$

$$z_1 = \frac{.571 - .6}{.0179}$$
$$= -1.62$$

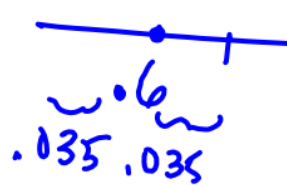
$$z_2 = \frac{.629 - .6}{.0179}$$
$$= 1.62$$



$$P(z < -1.62) = .0526 \quad P(z < 1.62) = .9474$$
$$.9474 - .0526 = .8948$$

$$\mu = .6 \quad \sigma = \sqrt{\frac{p(1-p)}{n}} = .0179$$


c.) Calculate the probability that the sample proportion intending to vote would fall within .035 of  $p$ .



$$P(.6 - .035 < \hat{p} < .6 + .035)$$

$$P(.565 < \hat{p} < .635) = .9495$$

d.) Calculate the probability that the sample proportion intending to vote would fall within .046 of  $p$ .



$$P(.554 < \hat{p} < .646) = .9898$$

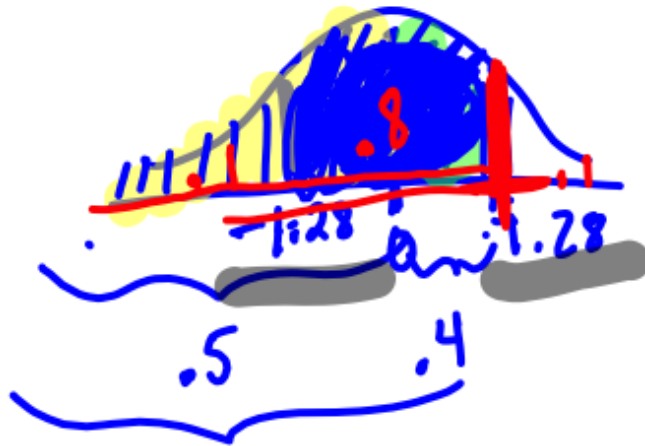
e.) Find the value  $k$  such that the probability of the sample proportion falling between  $.6 - k$  and  $.6 + k$  would equal .8

$$1.28 (.0179) = \boxed{.0229}$$

check:

$$P(.6 - .0229 < \hat{p} < .6 + .0229) = .7992$$

$$P(-c < Z < c) = .8$$



Z-score tells  
how many  
standard deviations  
you are from  
mean

$$\text{InvNorm}(.9) = 1.28$$

1.28 stand deviations from mean

InvNorm or  $z_{\text{norm}}$  use prob. for  $<$

2. <sup>← P</sup> 43% of the voters in the 1992 Presidential election voted for Bill Clinton. Suppose that you take a simple random sample of 500 voters from this population.

a.) Is 43% a parameter or a statistic?

↑  $\hat{p}$

$$n = 500$$

$$\mu = .43$$

$$\sigma = \sqrt{\frac{.43(1-.43)}{500}} = .022$$

b.) Determine the probability that the sample proportion of Clinton voters turns out to be less than 40%.

$$\begin{aligned} P(\hat{p} < .4) &= \text{pnorm}(.4, .43, .022) \\ &= \text{normalcdf}(-1000000, .4, .43, .022) \\ &= \cancel{.0863} \cdot 0863 \end{aligned}$$

c.) Determine the probability that the sample proportion of Clinton voters exceeds 50%.

$$\begin{aligned} 7.32 \times 10^{-4} \\ = .000732 \end{aligned}$$

$$\begin{aligned} P(\hat{p} > .5) &= \text{normalcdf}(.5, 10000, .43, .022) \\ &= 1 - \text{pnorm}(.5, .43, .022) \end{aligned}$$

d.) Determine the probability that the sample proportion of Clinton voters falls between 40% and 46%.

$$P(.4 < \hat{p} < .46) = .8273$$

e.) Determine the probability that the sample proportion of Clinton voters falls between 37% and 49%.

$$P(.37 < \hat{p} < .49) = .9936$$

~~\*~~ if  $n = 1500 \Rightarrow \sigma = .0128$   
 $P(.37 < \hat{p} < .49) = .999997$

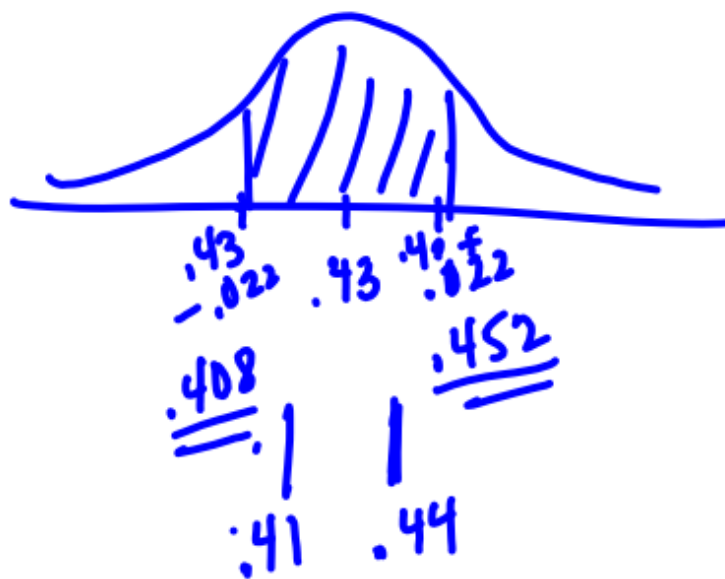
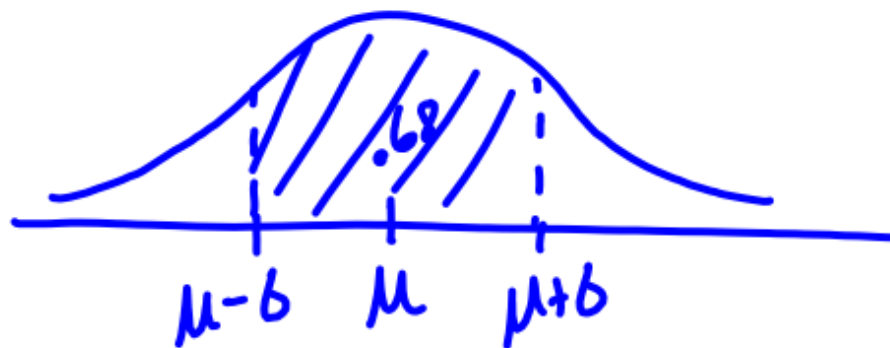
f.) Determine the probability that the sample proportion of Clinton voters falls between 43% and 89%.

$$P(.43 < \hat{p} < .89) = .5$$

g.) Without doing the actual calculations, indicate how your answers to b-f would change (get smaller, get larger, or stay the same) if the sample size were 1500 instead of 500.

increasing sample size increases probability  
on an interval  $P(- < \hat{p} < -)$



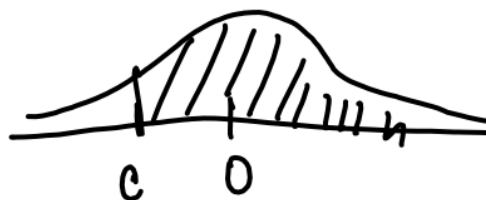


$$\mu = .43 \quad \sigma = .022$$

## Popper 09

1. Find  $c$  such that  $P(Z > c) = 0.8790$

- a. -0.19
- b. 1.17
- c. -1.17
- d. 0.19
- e. none of these



invNorm or gnorm (1 - .8790)

2. Find  $c$  such that  $P(-c < Z < c) = 0.8790$

- a. 1.55
- b. 1.17
- c. 0.81
- d. 1.21
- e. none of these



invNorm(.0605 + .8790)

$$1 - .8790 = .121$$
$$\text{each "tail"} = \frac{.121}{2} = .0605$$

## Section 5.1 - Bivariate data

**Bivariate data** is data for two different variables (usually related in some way).

Variables are classified as response variables and explanatory variables. A **response variable** (dependent) measures the outcome of a study. An **explanatory variable** (independent) attempts to explain the observed outcomes. Algebraically speaking, your explanatory variable is your "x" and the response variable is your "y".

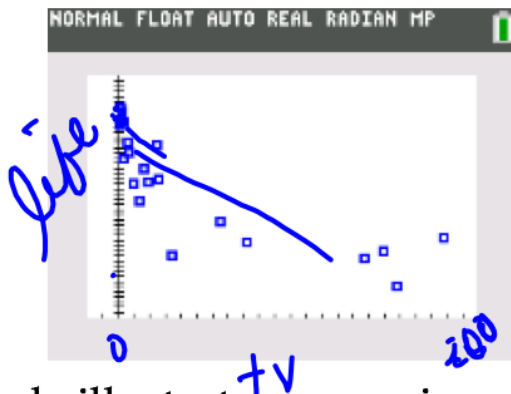
Once the explanatory and response variables are identified, we can display the association between the two using a scatterplot.

Example: Televisions and Life Expectancy  $y$   $x$

Country	Life Exp.	Per TV
Angola	44	200
Australia	76.5	2
Cambodia	49.5	177
Canada	76.5	1.7
China	70	8
Egypt	60.5	15
France	78	2.6
Haiti	53.5	234
Iraq	67	18
Japan	79	1.8
Madagascar	52.5	92
Mexico	72	6.6
Morocco	64.5	21
Pakistan	56.5	73
Russia	69	3.2
S. Africa	64	11
Sri Lanka	71.5	28
Uganda	51	191
U.K.	76	3
U.S.	75.5	1.3
Vietnam	65	29
Yemen	50	38

a) Which of the countries listed has the fewest people per television set? Which has the most? What are those numbers?

b) Use the calculator to produce a scatter plot. Does there appear to be an association?



bigger "x" meant fewer tvs per person

This example illustrates a very important distinction between association and causation. Two variables may be strongly associated without a cause-and-effect relationship existing between them. Often the explanation is that both variables are related to a third variable not being measured; this variable is often called a *lurking* or *confounding* variable.

c) In this case, suggest a confounding variable that is associated with both a country's life expectancy and the prevalence of televisions in the country.

## Summary for creating a scatter plot:

### R-Studio:

1. Choose variable names.
2. Enter the lists in R:  
> xListName=c(...)  
> yListName=c(...)
3. Now use the plot command:  
> plot(xListName, yListName)

$tvs = c( \dots )$   
 $life = c( \dots )$

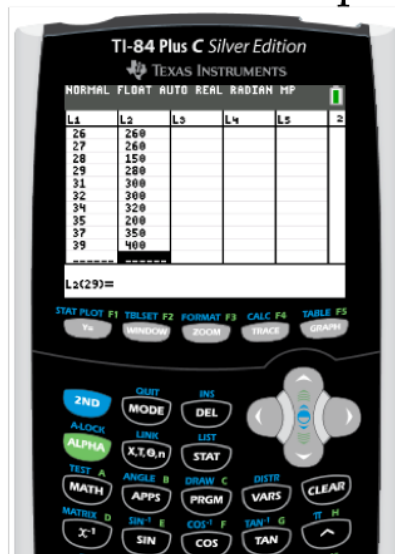
$plot(tvs, life)$

### TI-83/84:

1. Go to STAT - EDIT

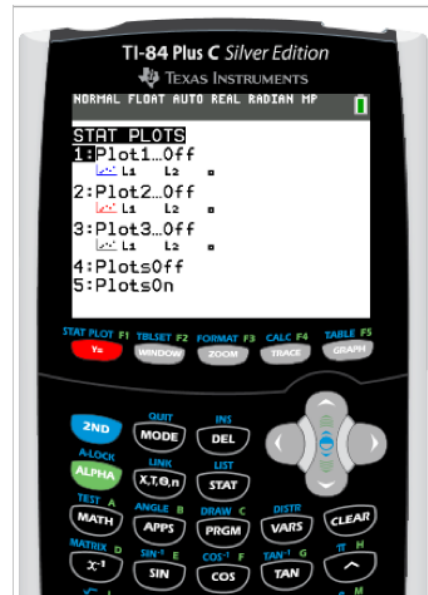


2. Enter the explanatory variable under L1 and the response under L2

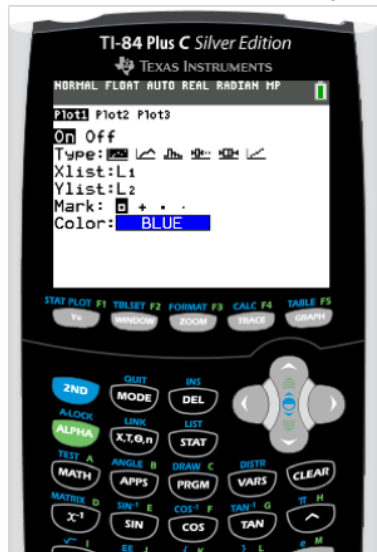


3. Choose 2ND - QUIT to go back to home screen.

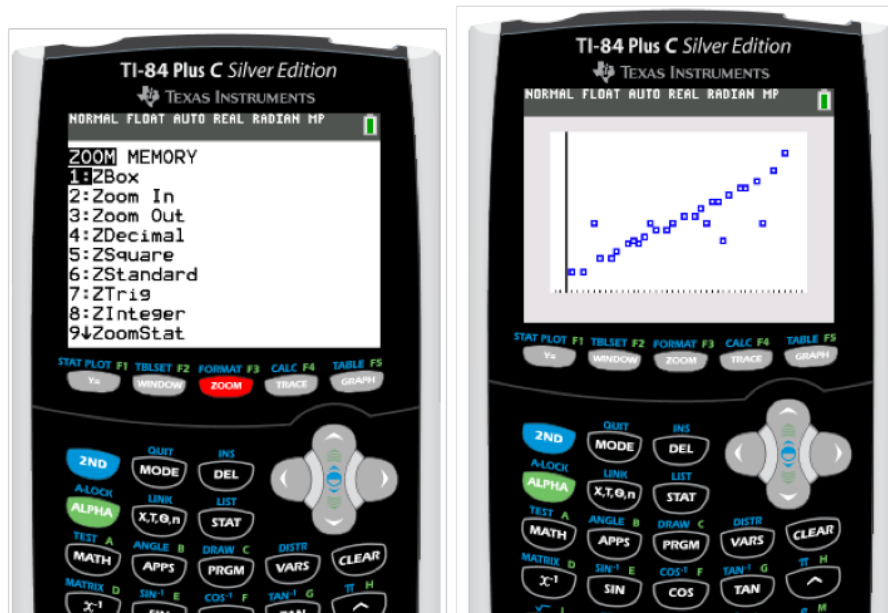
4. Go to 2ND - Y= and select Plot1



5. Choose ON, select the first type and make sure your Xlist is L1 and the Ylist is L2



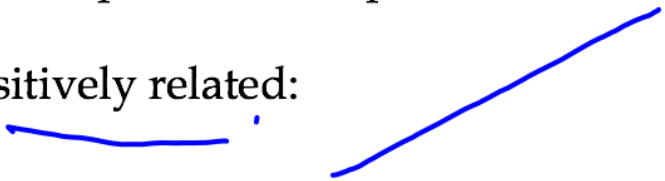
6. Choose GRAPH then ZOOM and ZoomStat



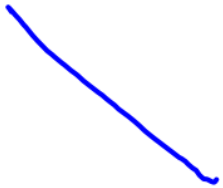


To interpret a scatter plot we will look at the direction, form and strength.

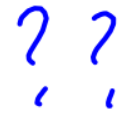
Positively related:



Negatively related:



Linear relationship:



## Popper 09

3. A positive slope indicates that large values of  $x$  are associated with \_\_\_\_\_ values of  $y$ .

- a. Large
- b. Small
- c. Cannot be determined from this information

4. In regression analysis, the independent variable is also called the \_\_\_\_\_ variable.

- a. Explanatory
- b. Response
- c. Correlation
- d. Expected
- e. None of these

5. Choose A