Unit 4b

Probability Distributions in R

As an example in R we will work with the normal distribution

In my opinion, alot of things are ***easier*** in R

`dnorm(x,mean = 0,sd = 1)` gives prob *densities* over X

`rnorm(n,mean = 0,sd = 1)` gives you n samples

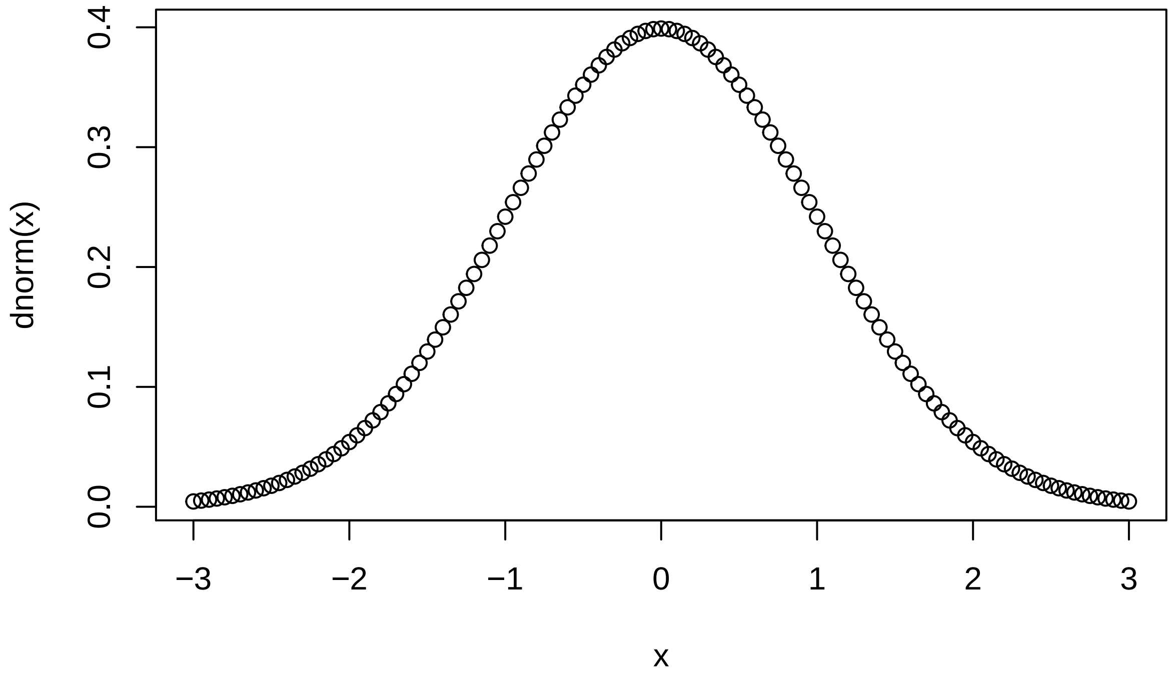
`pnorm(x,mean = 0,sd = 1)` gives you the cdf of $F(X)$

`qnorm(p,mean = 0,sd = 1)` gives you quantiles $F^{-1}(p)$

If you don't know what quantiles are, or you don't understand what $F^{-1}(p)$ means, don't worry! We will cover this topic separately later in the course (Q-Q plots)

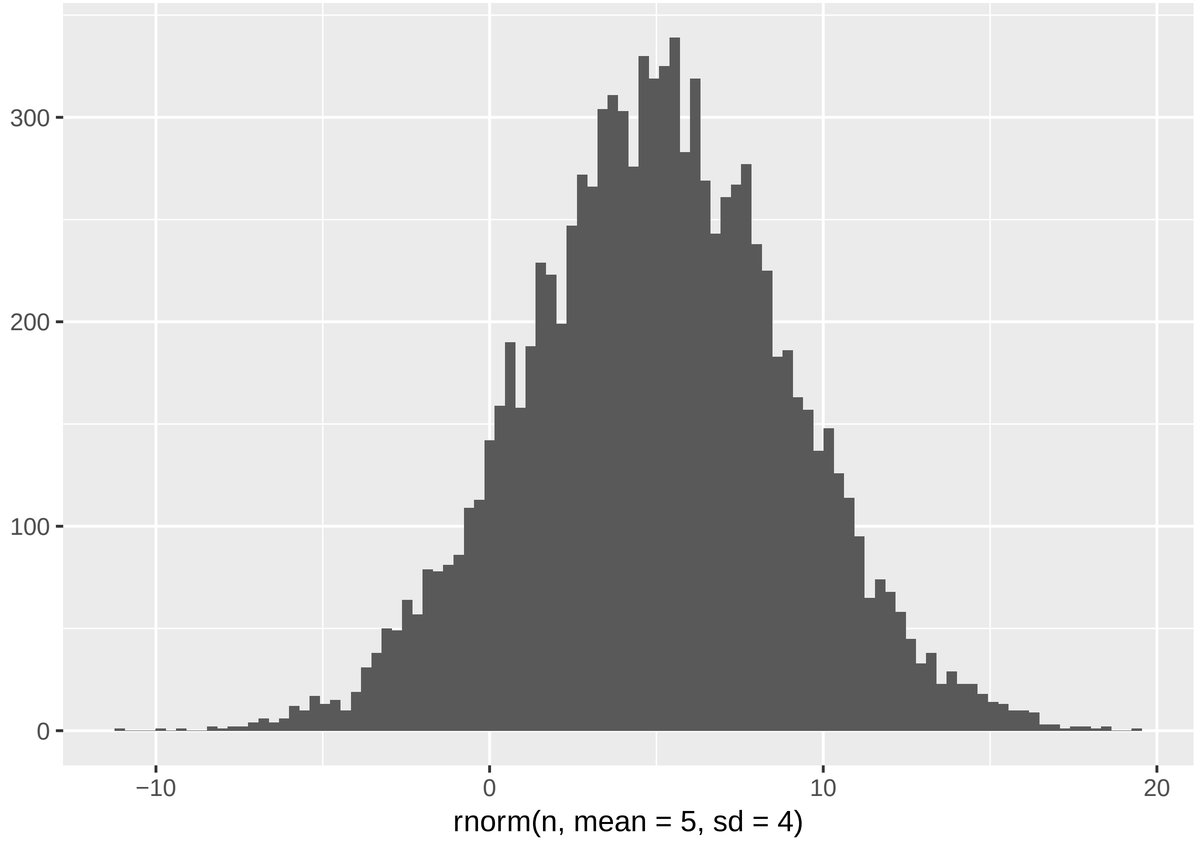
`dnorm` - get the probabilities

```r
x = seq(-3,3,by=0.05)   # choose the range of X
plot(x,dnorm(x))       # the prob densities P(X)
```



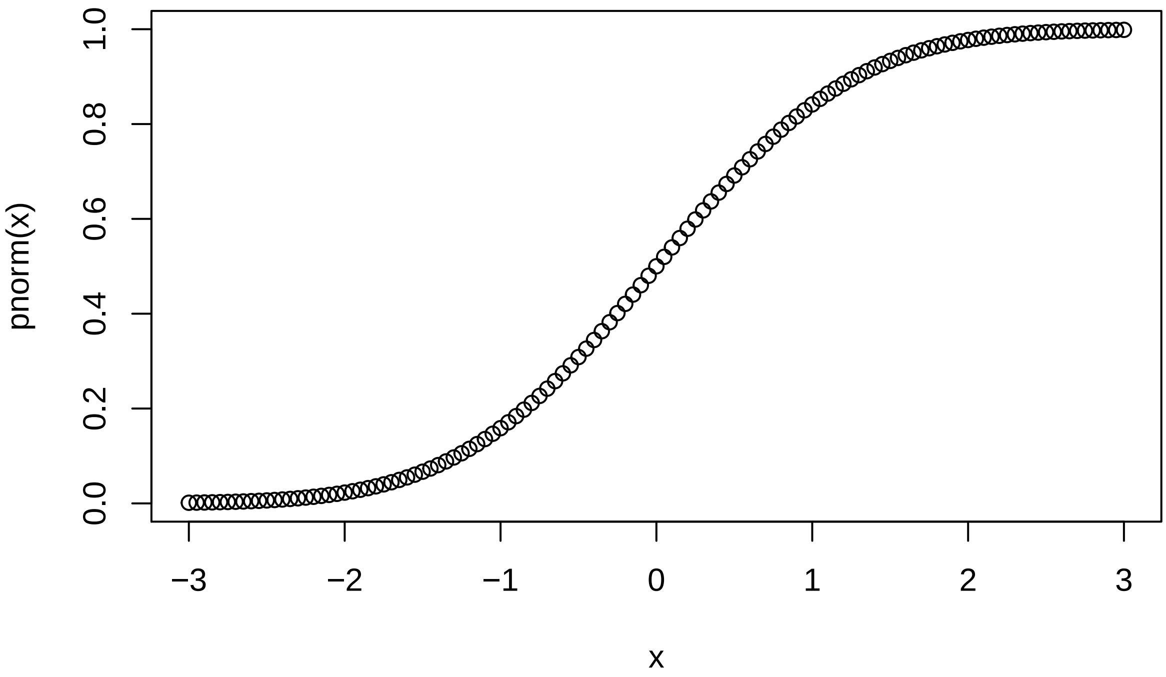
rnorm - get n samples

```
n = 10000
qplot(rnorm(n,mean=5,sd=4),geom='histogram',bins=100)
```



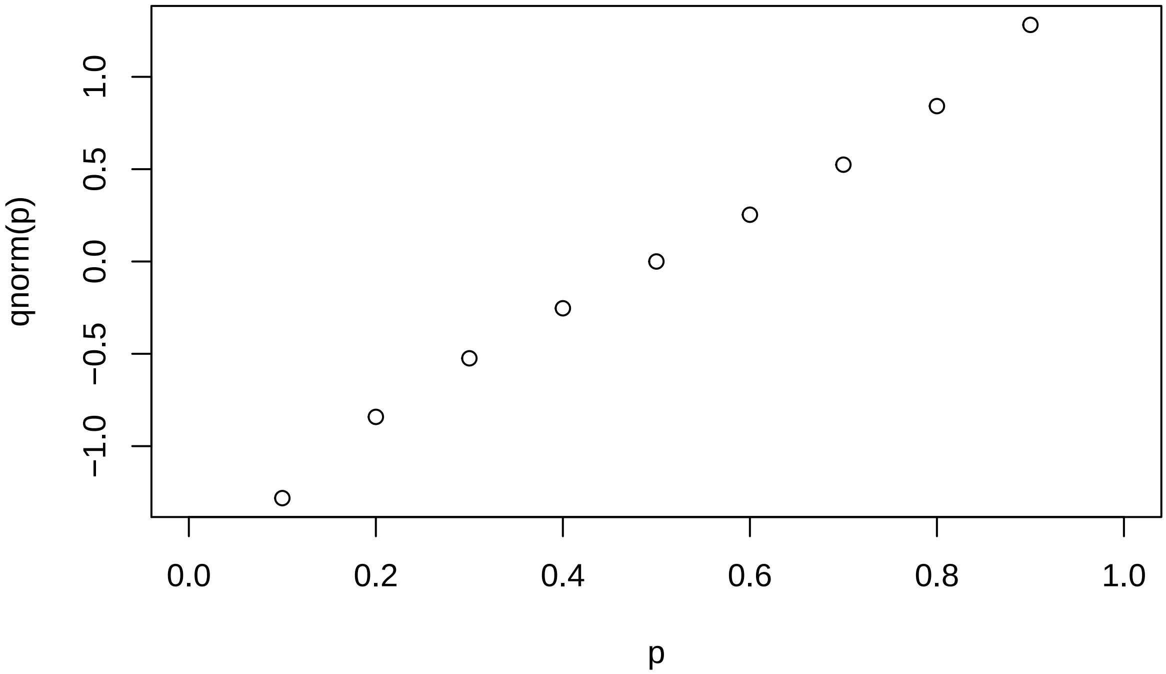
pnorm - get $F(X)$

```
x = seq(-3,3,by=0.05)   # choose a range x
plot(x,pnorm(x))       # this is the probability that X>x
```



`qnorm` - get quantiles, $F^{-1}(p)$

```r
p = seq(0,1,by=0.1)    # deciles probabilities
plot(p,qnorm(p))       # these are the quantiles - deciles
```



z score

Given an observation or data point x , the z-score,

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

is the number of standard deviations the *raw score*, x, is from the mean

also called: z-values, normal scores, and standardized variables.

Since things tend to be normally distributed, the z-score gives you a sense of how usual or unusual the data is (even if data is not exactly normally distributed you can still calculate this if you know the mean and standard deviation of the population)

Example calculating Z-scores

The table shows the mean and standard deviation for total scores on the SAT and ACT. The distribution of SAT and ACT scores are both nearly normal.Suppose Ann scored 1300 on her SAT and Tom scored 24 on his ACT. Who performed better?

	SAT	ACT
Mean	1100	21
SD	200	6

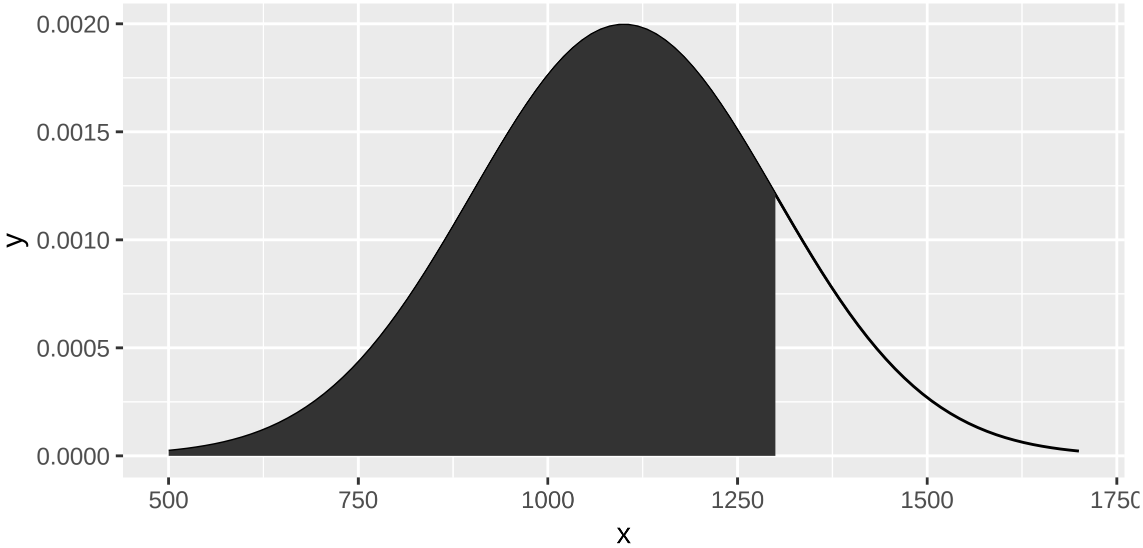
$z_{\text{Ann}} = \frac{1300-1100}{200} = 1$

$z_{\text{Tom}} = \frac{24-21}{6} = 0.5$

Ann scored 1 standard deviation above the mean, while Tom scored 0.5 sds above the mean. **Conclusion** Ann did better

Finding Tail Areas

How many people have an SAT score below Ann's 1300?



The area under the curve can be obtained by `pnorm(z)`

```
pnorm(1) # This is the same as Ann's percentile
```

```
## [1] 0.8413447
```

Summary on Probability Distributions

Here we explored R's capabilities with the normal distribution but R has many more distributions built in, for example, the exponential distribution you can run all of the same functions by replacing `dnorm` with `dexp`, `pnorm` with `pexp` etc.

Univariate probability distributions give you a simple model for how often something happens.

For the **next step** we will move forward to look at how **statistics** like the sample mean, $\overline{X}$ are distributed. We already saw that $\overline{X}$ is normally distributed with standard deviation given by S.E.M. $= \frac{\sigma}{\sqrt{n}}$. It turns out that there is more to this, we will move on to t distributions, bootstrap methods on real data and inference (hypothesis testing) in the next two units. But first…

Lab Unit 4

Exercise 1

Edward earned a 1030 on his SAT. Referring to the table on slide 8, what is his percentile?

Exercise 2

Suppose SAT scores are distributed $N(\mu = 1100; \sigma = 200)$

What percent of SAT takers get between 1100 and 1400?

HINT: It helps to draw a picture of the distribution and tail regions.

Exercise 3

Suppose men's heights scores are distributed $N(\mu = 70; \sigma = 3.3)$ in inches.

If Erik's height is at the 40th percentile. How tall is he?

Hint: you can use

```
qnorm(0.4)    # returns value of Z- corresponding to 40th percentile
```

```
## [1] -0.2533471
```

Now use this value in the Z score formula to solve for the raw score x which is Erik's height.

Exercise 4: Poisson

For this problem **Refer to the Python-Jupyter notebook** for formulas etc.

Suppose that calls arrive to a call center at an average rate of 3 per minute and can be described as a poisson process. What is the probability of getting 6 calls in 2 minutes? Plug numbers into the formula to get an answer.

Check your result by examining the poisson distribution with rate $\lambda = 3$ calls per minute and the time interval of interest will be $t = 2$ minutes $\alpha = \lambda t = 3 \cdot 2 = 6$ examine the vertical axis probability corresponding to $X = 6$ does the number in the plot agree with your calculation?

Exercise 5

For this problem **Refer to the Python-Jupyter notebook** for formulas etc.

Suppose a car insurer has determined that 88% of its drivers will not exceed their deductible in a given year.

If someone at the company were to randomly draw driver files until they found one that had not exceeded their deductible, what is the expected number of drivers the insurance employee must check?

What is the standard deviation of the number of driver files that must be drawn?

Hint: You could use the formulas from the geometric distribution

Exercise 6: Field Assignment Due July 9

In order to understand the connection between exponential distributions and independent processes such as arrival times you will conduct the following experiment with real data

- collect at least 50 interarrival time data from a real physical location such as library, fast food restaurant
- You may work in groups of 2 if you can find a partner
- submit a histogram of your interarrival time data
- You must calculate the sample mean and sample population standard deviation
- Also include details about location, date, time start collection, time end collection data

- you will need this data to do hypothesis testing in later units