

# ENV 710: Lecture 4

---

inference

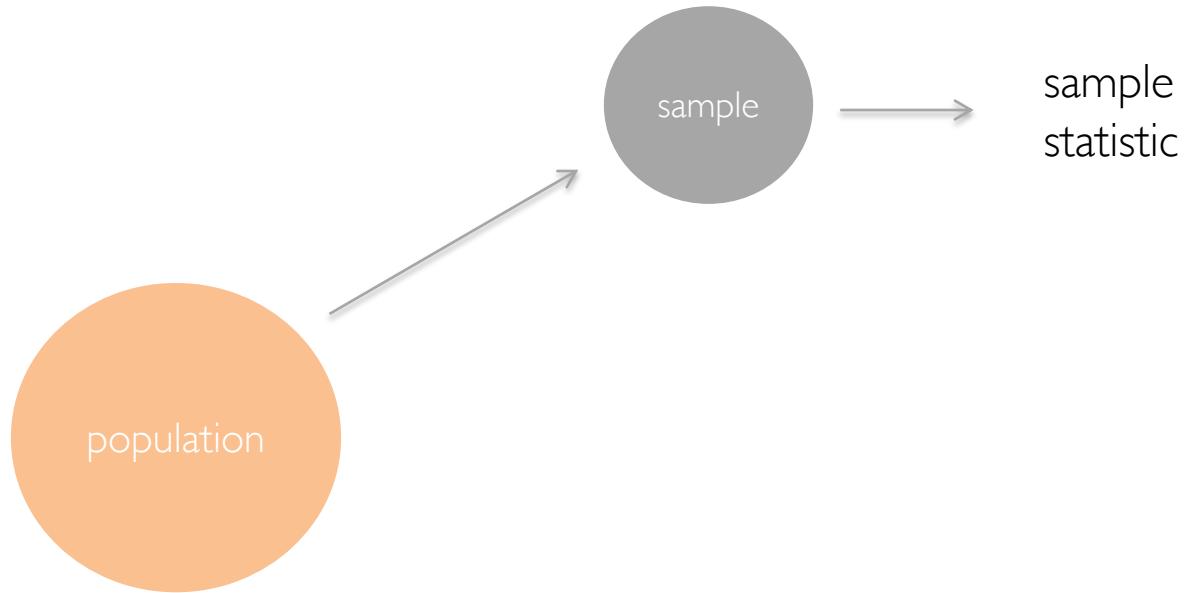
**statistical inference**

**sampling**

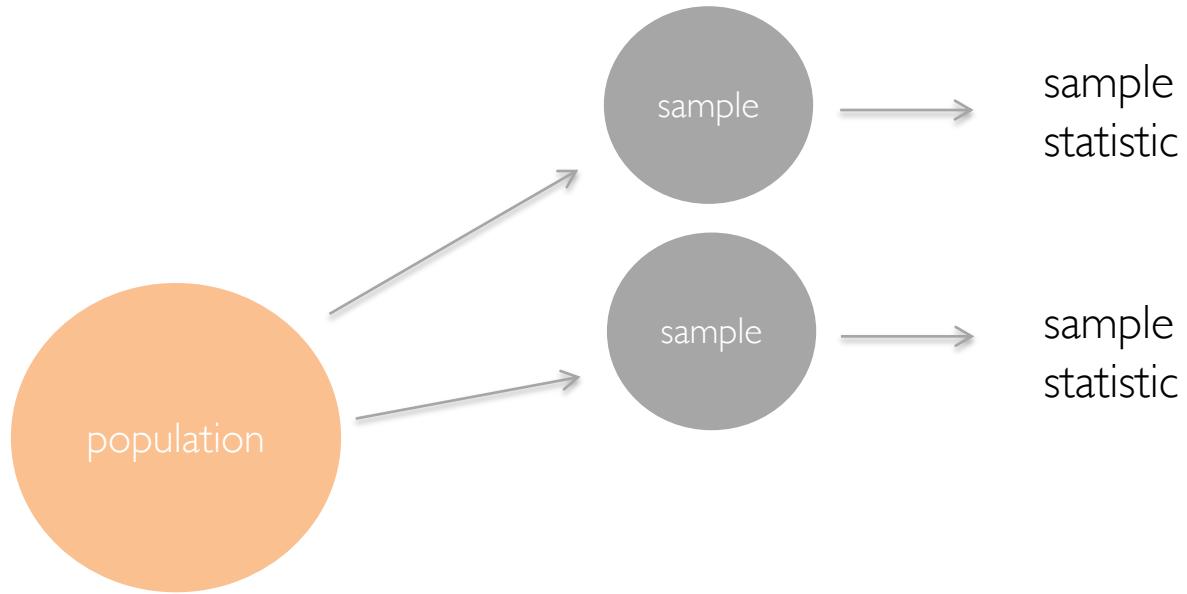
# learning goals

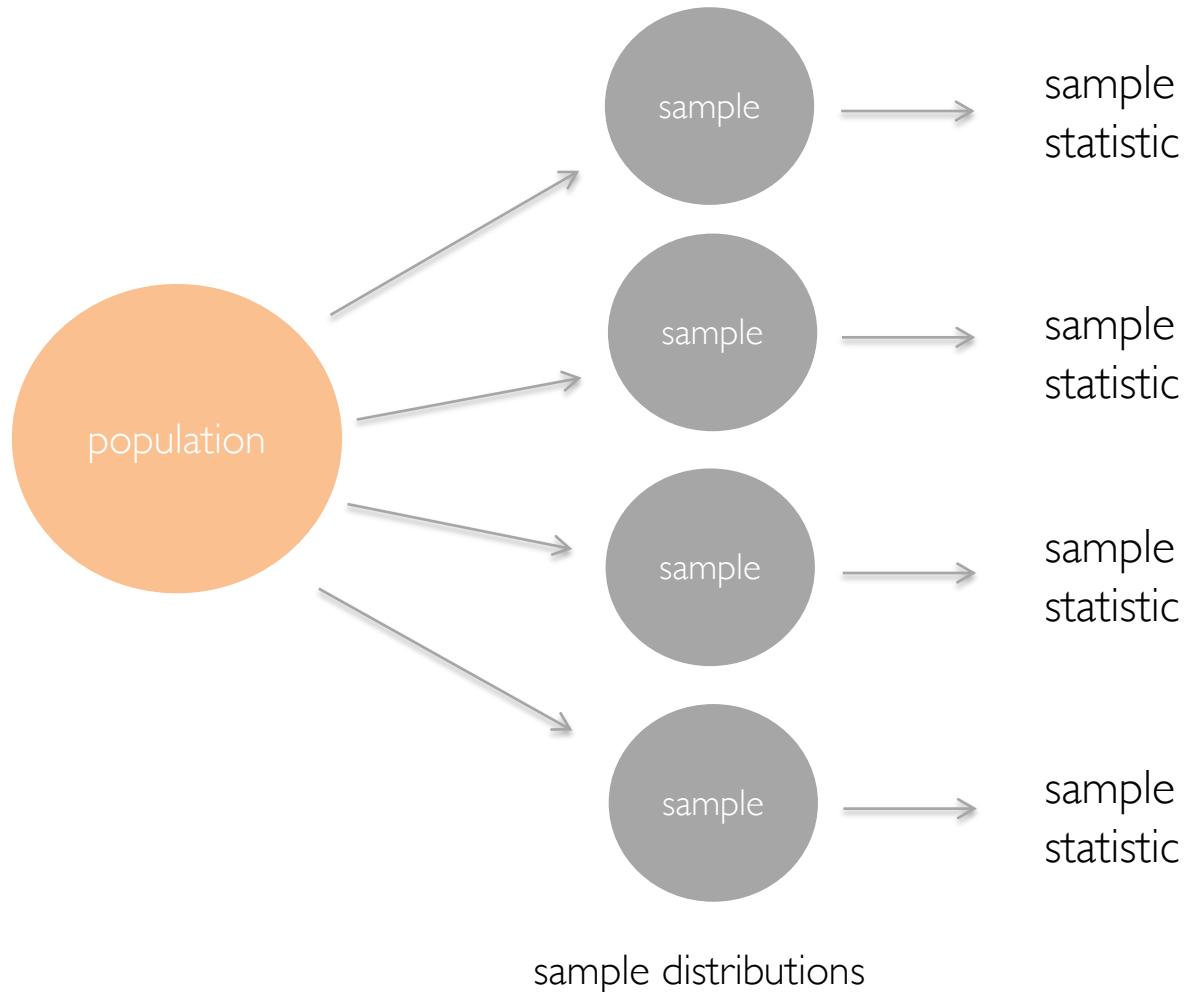
- what is sampling and why do we do it?
- understand concepts such as IID and sampling distributions
- know when to use standard error versus standard deviation

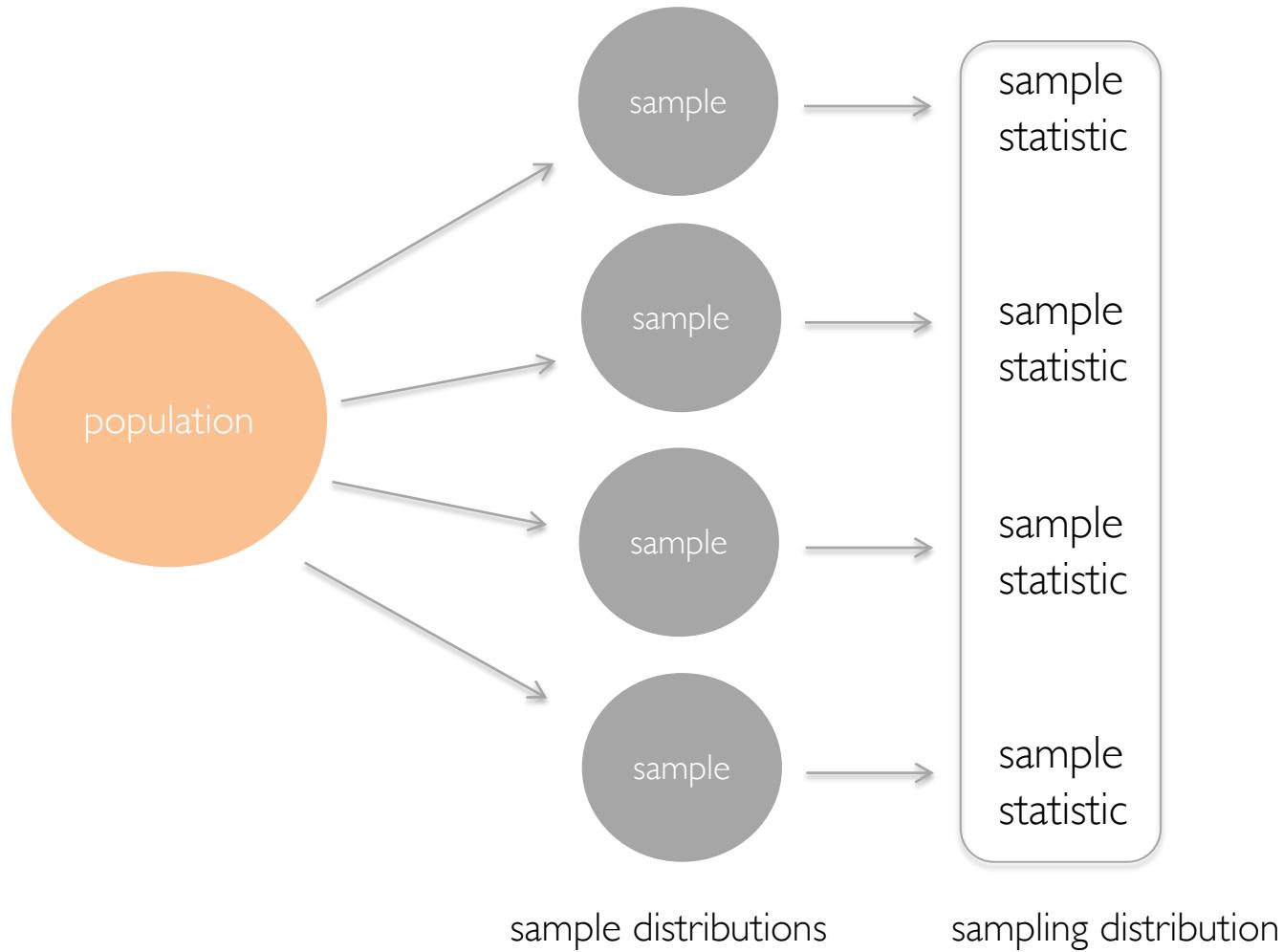
stuff  
you  
should  
know

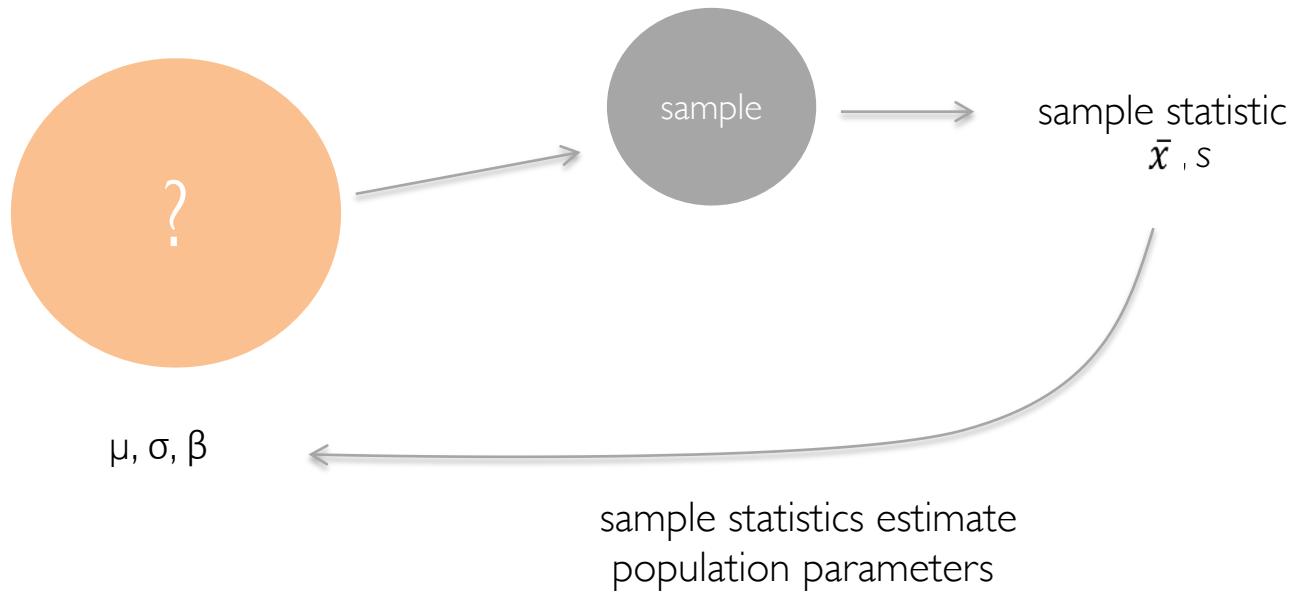


IID  
independent and  
identically distributed

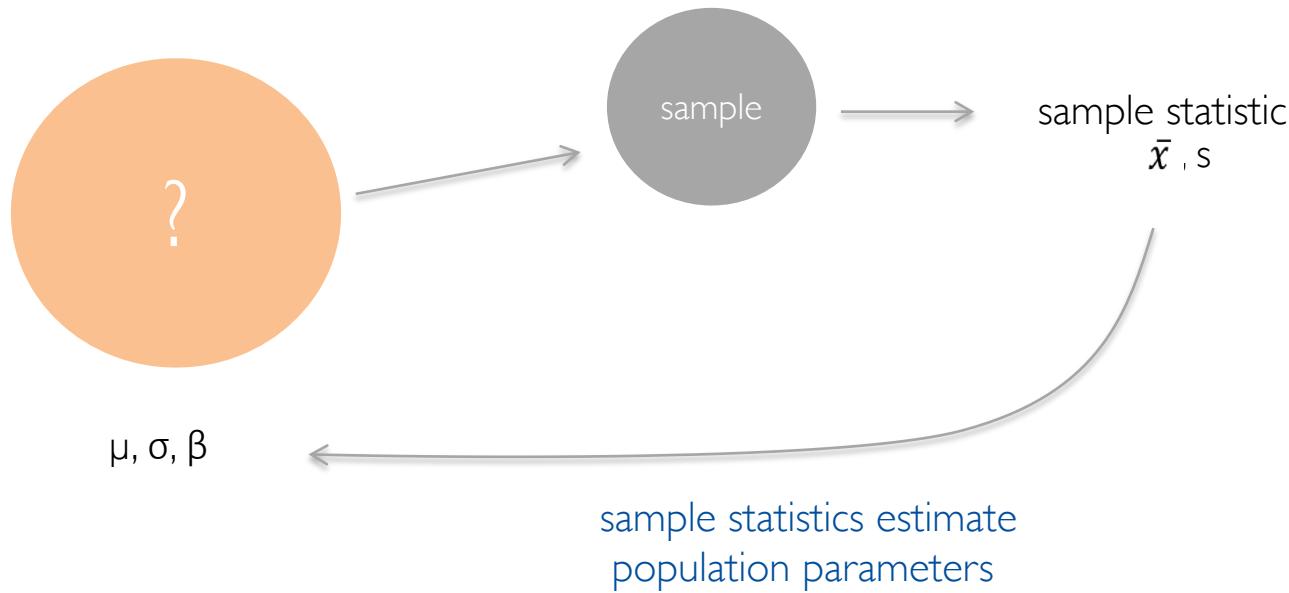






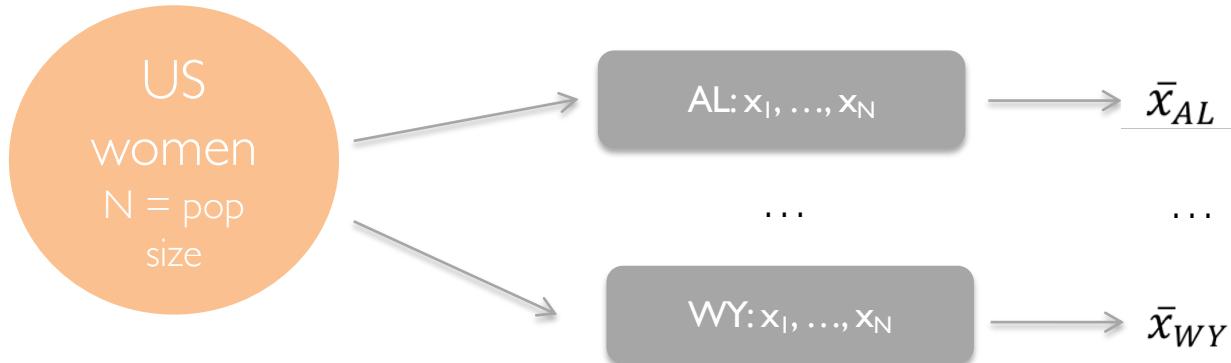


- sample statistics are random variables that vary from sample to sample
- $\bar{x}$  is an estimator of  $\mu$



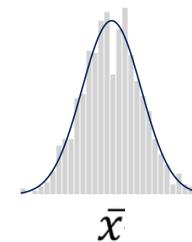
estimators should be unbiased and efficient

- an estimator is unbiased if its expected value equals the population value, e.g.  $E(\bar{X}) = \mu$
- an efficient estimator needs relatively few samples to achieve a given performance



$$\mu = \frac{x_1 + x_2 + \dots + x_N}{N}$$

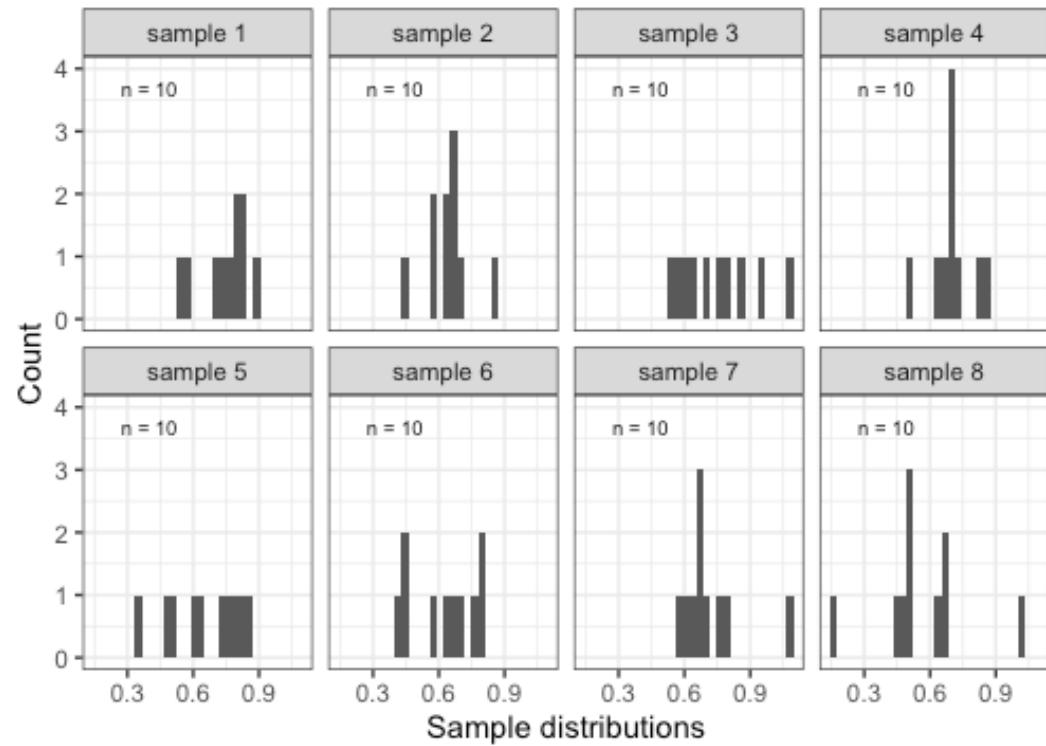
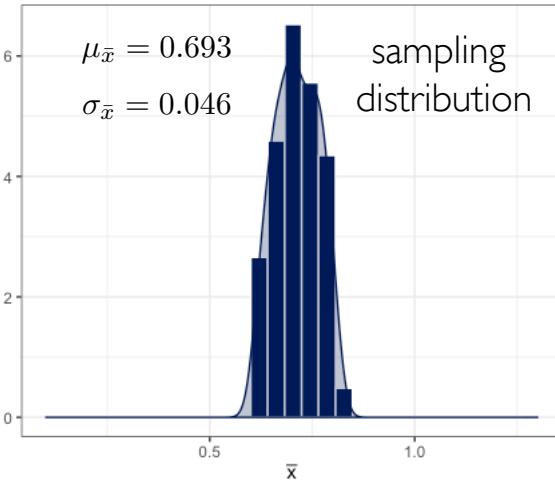
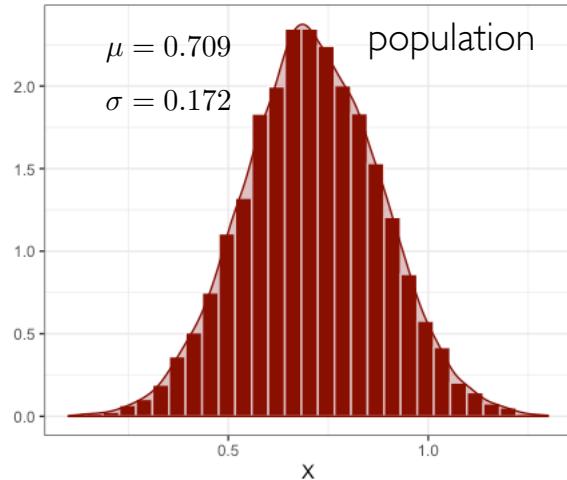
$$\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N}}$$

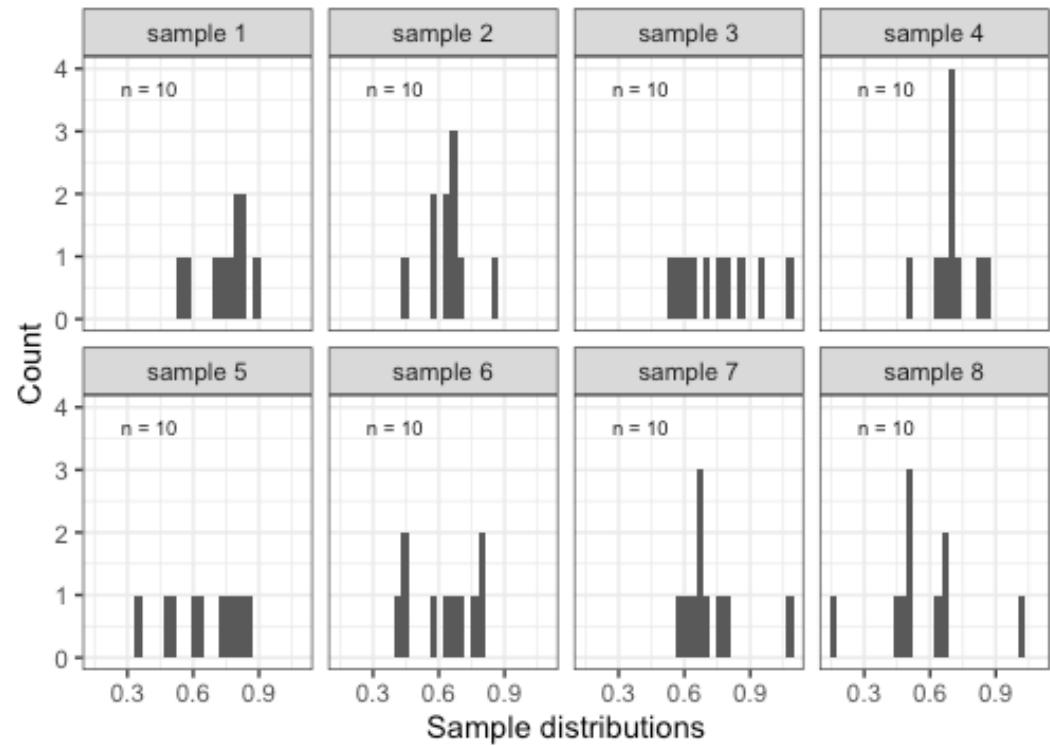
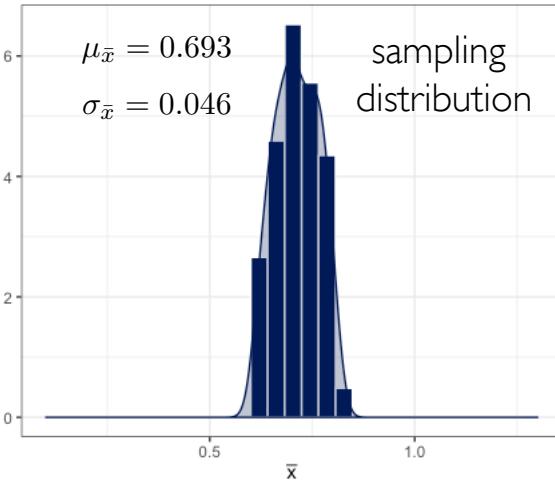
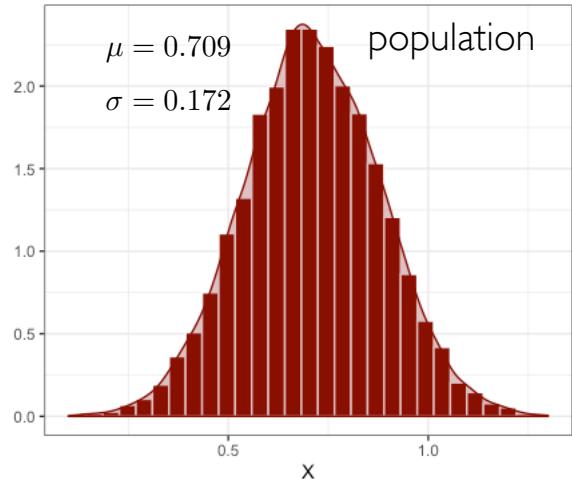


sampling distribution

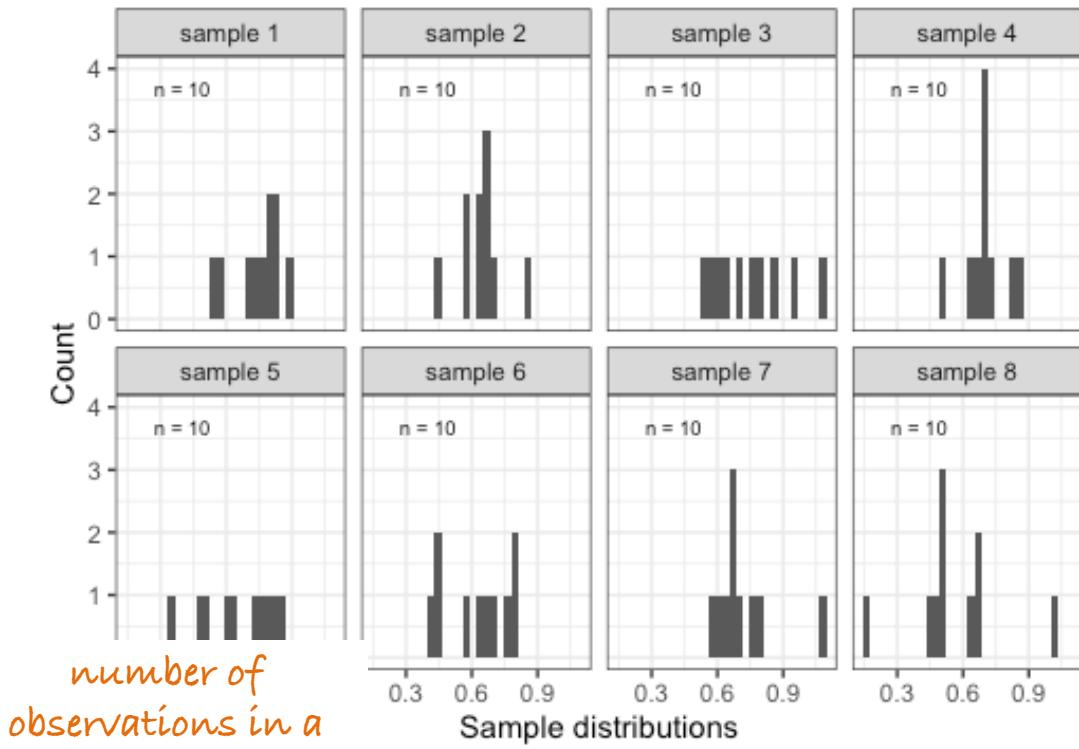
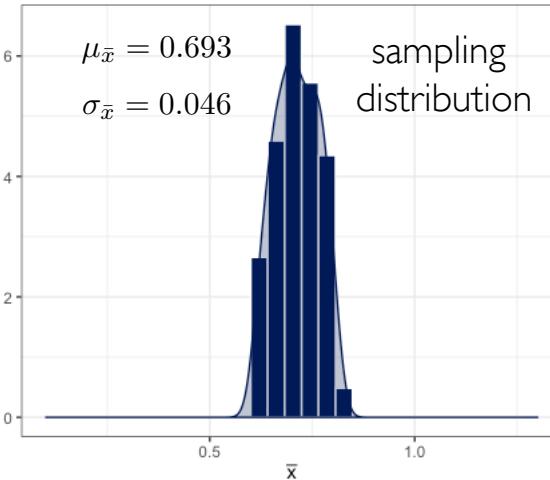
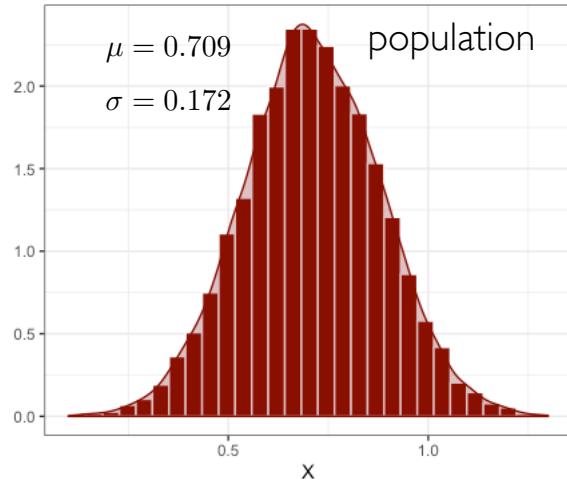
$$\mu(\bar{x}) \sim \mu$$

$$\sigma(\bar{x}) < \sigma$$



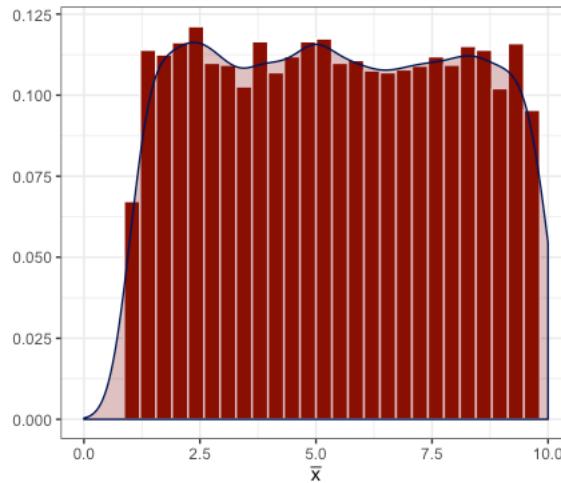


$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.172}{\sqrt{10}} = 0.054$$



number of  
observations in a  
single sample

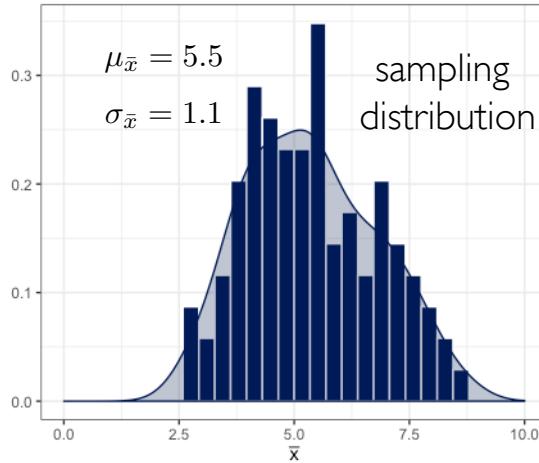
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = 0.054$$



population  
 $X \sim U(1, 10)$

$$\mu = 5.4$$

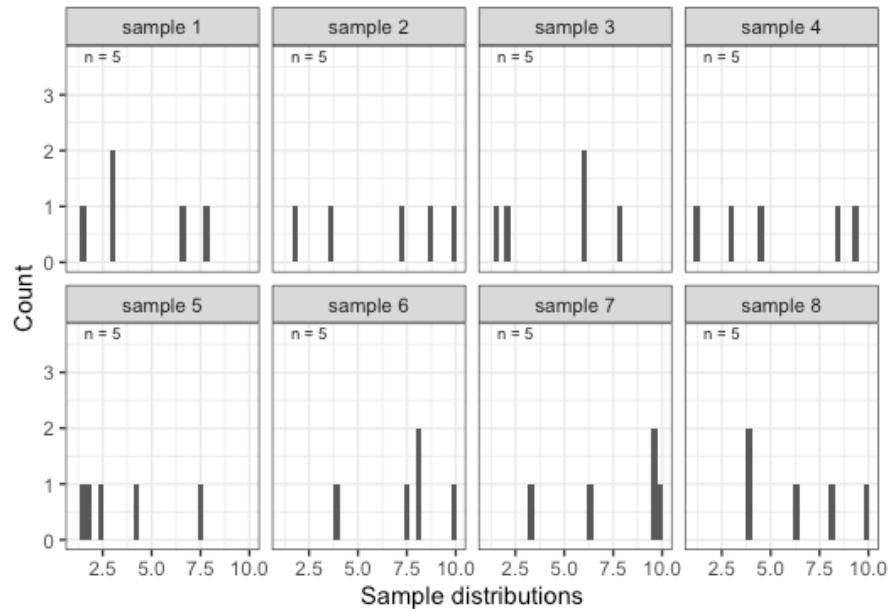
$$\sigma = 2.6$$



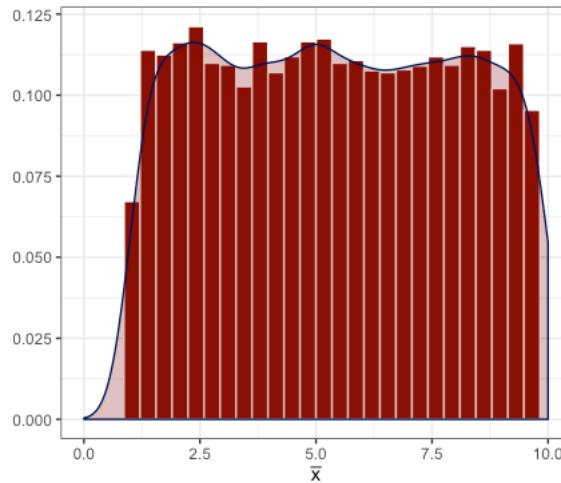
$$\mu_{\bar{x}} = 5.5$$

$$\sigma_{\bar{x}} = 1.1$$

sampling  
distribution



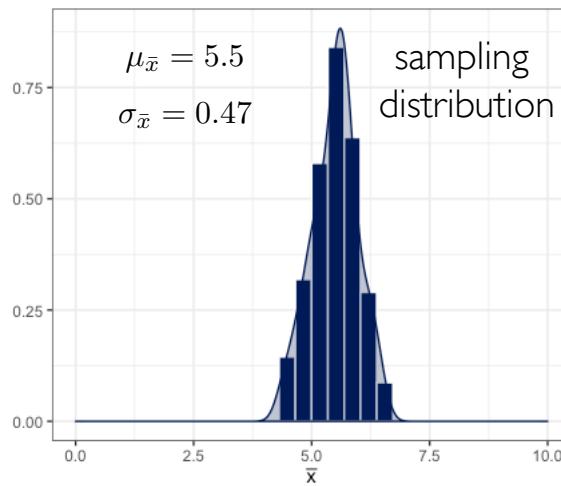
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.6}{\sqrt{5}} = 1.2$$



population  
 $X \sim U(1, 10)$

$$\mu = 5.4$$

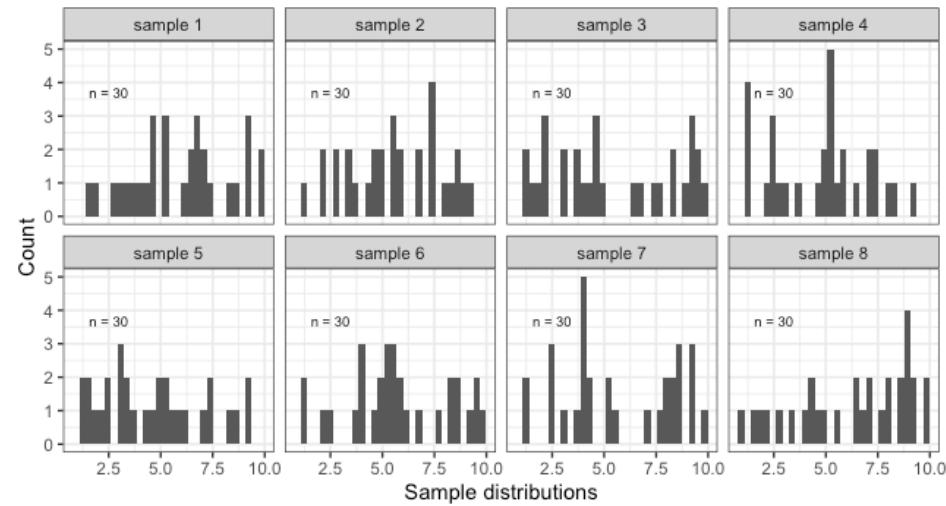
$$\sigma = 2.6$$



$$\mu_{\bar{x}} = 5.5$$

sampling  
distribution

$$\sigma_{\bar{x}} = 0.47$$



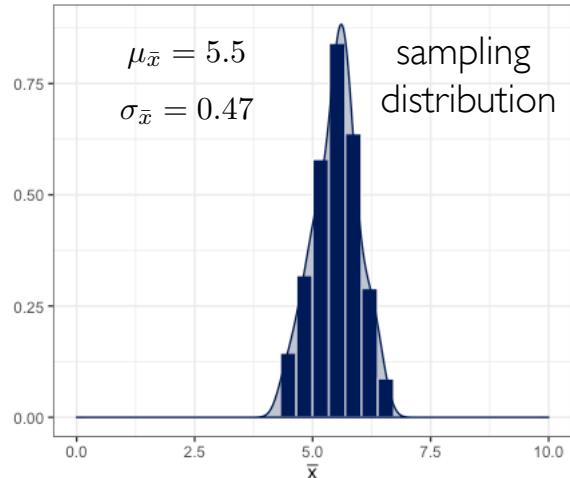
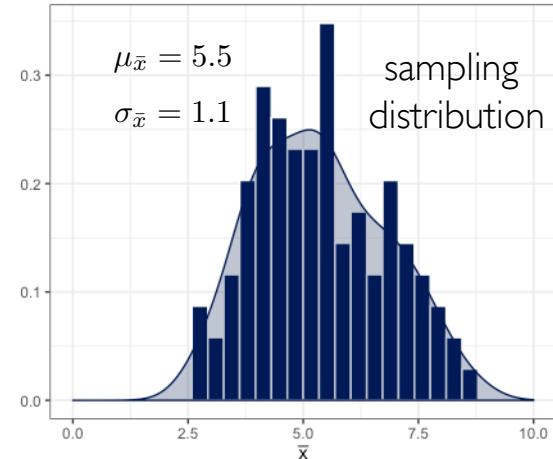
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.6}{\sqrt{30}} = 0.47$$

# note on standard error

- SE or SEM (standard error of the mean) is the standard deviation of the sampling distribution of the sample means, where  $s$  is the sample standard deviation and  $n$  is the sample size

$$se = \frac{s}{\sqrt{n}}$$

- SE is a measure of the precision of the sampling mean
- SE demonstrates how close the sample mean is to the population mean



# standard error vs. standard deviation

## standard deviation

- measure of variability in a population
- standard deviation of a sample,  $s$ , estimates the variability in a population

## standard error

- measure of how precisely a sample statistic estimates a population parameter
- SE is an inferential measure describing uncertainty in your measured effects and depends heavily on sample size

**inference**

**central limit theorem**

# learning goals

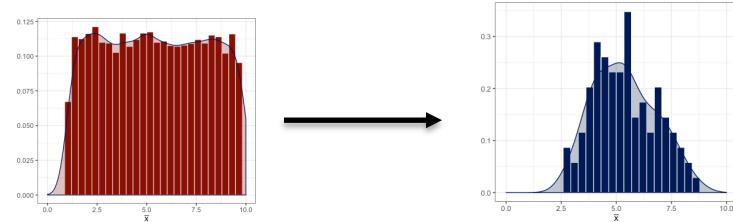
- what is the Central Limit Theorem (CLT)?
- what are the conditions for the CLT?
- are the three tenets of the CLT?
- why is the CLT an earth shattering, big deal?

stuff  
you  
should  
know

# Central Limit Theorem (CLT)

the distribution of sample statistics is nearly normal, centered at the population mean, and with a standard deviation equal to the population standard deviation divided by the square root of the sample size

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$



# Conditions

CLT describes the “population of the means” created from sets of independent, identically distributed random variables

*iid*

1. **independence:** sampled observations must be independent
  - random sample/assignment
  - if sampling without replacement,  $n < 10\%$  of the population
2. **sample size/skew:** either the population is normal, or if the distribution is skewed, the sample size is large ( $n > 30$ )

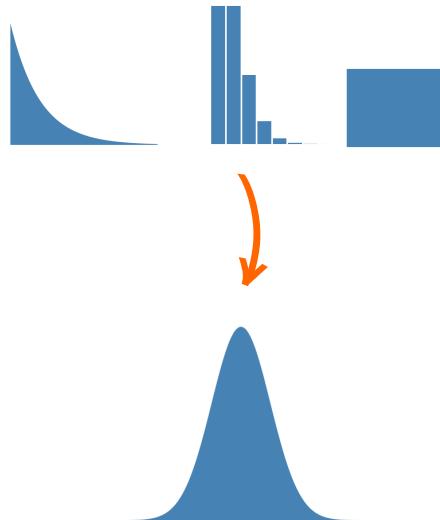
# regardless of the parent population...

- ① the mean of the population of means is always equal to the mean of the parent population
- ② the standard deviation of the population of means is always equal to the standard deviation of the parent population,  $\sigma$ , divided by the square root of the sample size
- ③ the distribution of means will approximate a normal distribution as the sample size,  $n$ , of samples increases

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

# this is a big deal...

as sample size gets large, even if you start with a non-normal distribution, the sampling distribution approaches a normal distribution



simulation

<https://jrp-projects.shinyapps.io/cltapp/>

# why do we care...?

with data that are approximately normally distributed, it is relatively easy to calculate

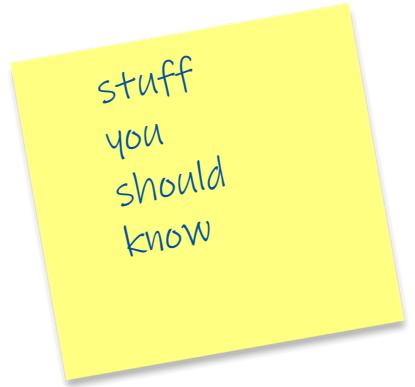
- probabilities
- confidence intervals
- conduct statistical inference and hypothesis testing

**inference**

**checking for normality**

# learning goals

- why do you need to check the distribution of your data?
- how do you check the fit of a distribution to the data?
- what is a quantile-quantile plot?
- how do you test for the normality of your data?



# choice of distribution

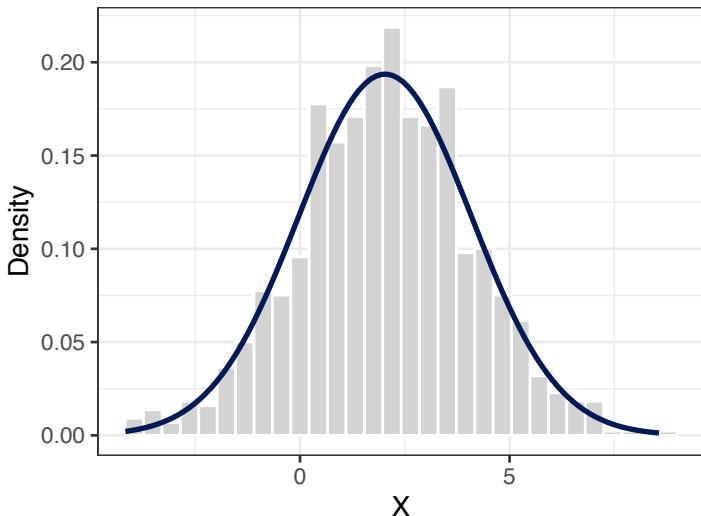
do the data meet the assumptions of the underlying distribution or statistical model?

- fitting a statistical distribution to data is *part art and part science*
- maintain a balance between getting a good distributional fit and preserving ease of estimation
- may settle for a distribution that less completely fits the data because estimating the parameters is easier

# choice of distribution

1. are the data continuous or discrete?
2. how symmetric are the data?
  - symmetric: normal, logistic, Cauchy
  - positively skewed: log-normal, gamma, Weibull
  - negatively skewed: beta
3. are there upper or lower limits to the data?
  - lognormal – values never less than 0
4. how likely are extreme values?
  - logistic, cauchy

# choice of distribution

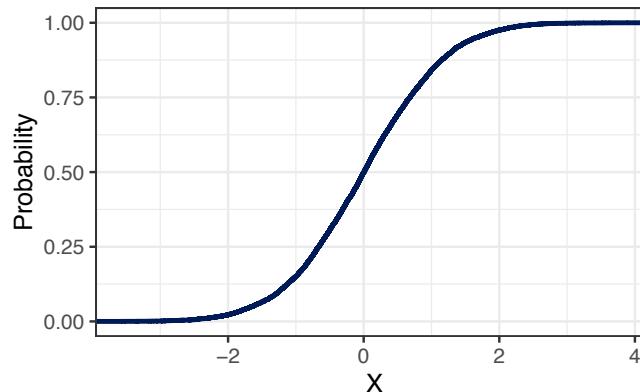
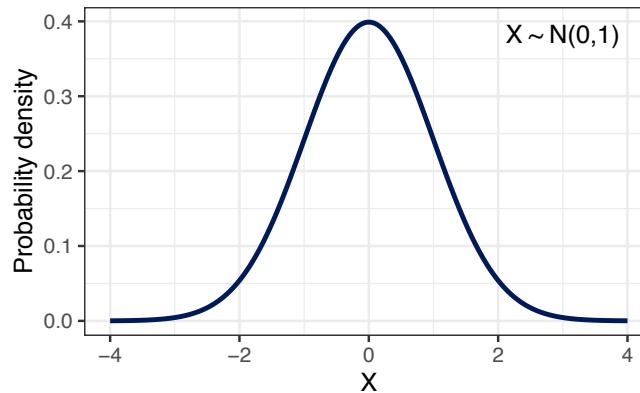


1. compare a histogram of observed data to the fitted data
2. compute the moments of the data distribution ( $\mu, \sigma, \text{skewness}, \text{kurtosis}$ )
3. compare the data to the cumulative distribution function (CDF) to test whether the fitted distribution fits the data

# cumulative distribution function

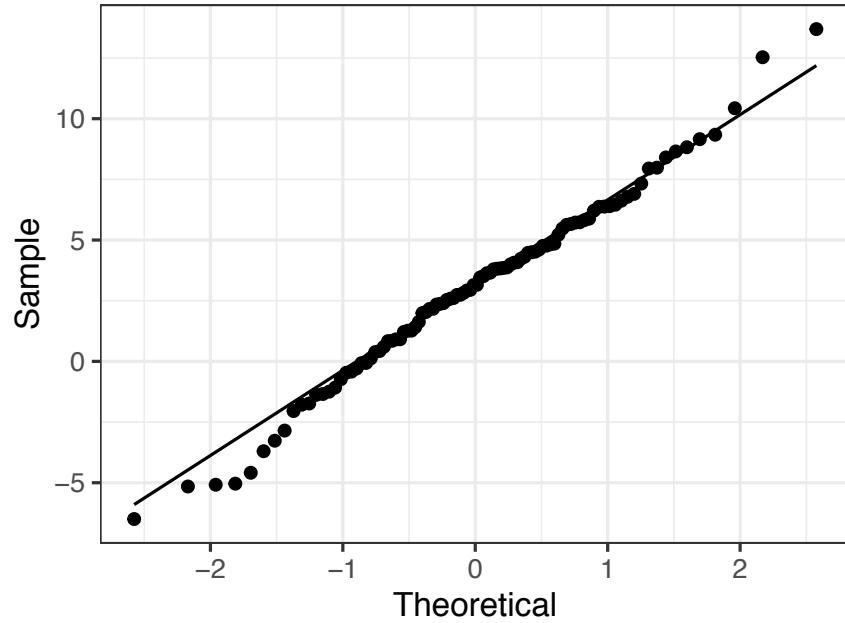
“area in so far function”

cdf is the probability  
that a random  
variable  $X$  will be  
found at a value less  
than or equal to  $X$



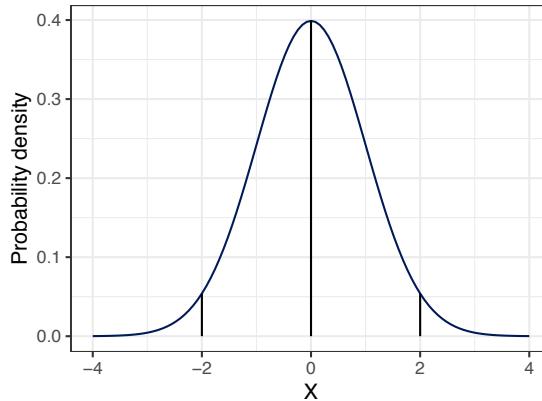
# quantile-quantile (Q-Q) plot

- graphical method for comparing 2 probability distributions by plotting their quantiles against each other
- graphical assessment of ‘goodness of fit’
- what is a quantile?

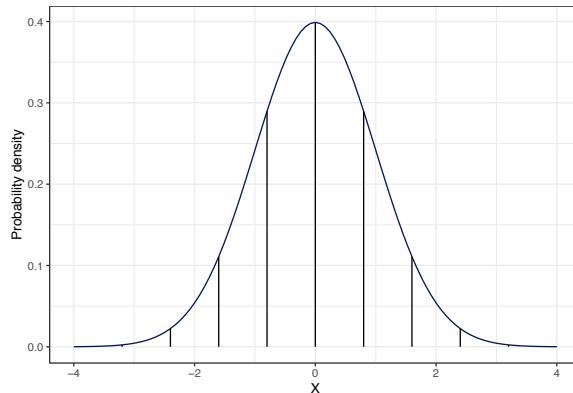


# quantiles

- **quantiles** are cut points dividing the range of a probability distribution into intervals with equal probabilities
- there are one fewer quantiles than the number of groups created
- **quartiles** are the three cut points that divide a dataset into four equal-size intervals
- 10 quantiles are **deciles**



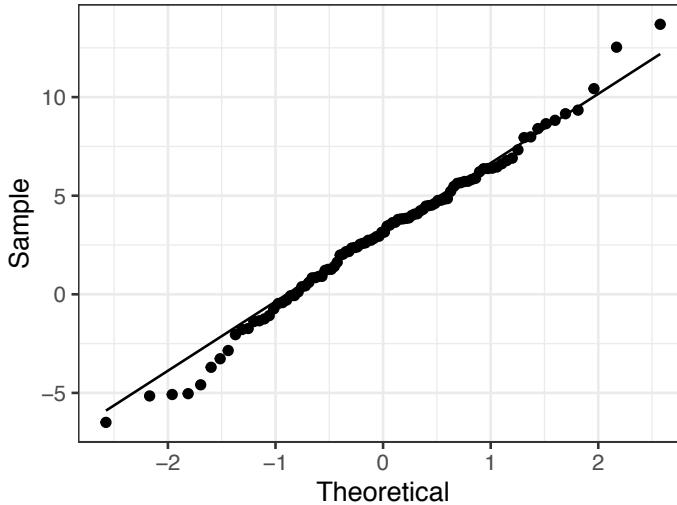
```
x.vals <- seq(-4, 4, length = 1000)  
quantile(x.vals)
```



```
quantile(x.vals, c(seq(0, 1, by = 0.1)))
```

# quantile-quantile (Q-Q) plot

- does the dataset match a theoretical distribution (e.g. normal)?
- do two datasets come from populations with a common distribution?
- one-to-one relationship between the data and theoretical quantiles, suggests the data follow a nearly normal distribution
- the closer the points are to a perfect straight line, the more confidence we have that they follow a normal distribution



# testing for normality

## Shapiro-Wilk test

- $H_0$  that a sample comes from a normal distribution
- $H_a$  that a sample does not come from a normal distribution

```
shapiro.test()
```

```
set.seed(1001)
x <- rnorm(20, mean = 0, sd = 1)
shapiro.test(x)
```

```
Shapiro-Wilk normality test
W = 0.9818, p-value = 0.9554
```

# testing for normality

but is this a good idea...?

- p-value of 0.05 rejects the  $H_0$  that the data come from a normal distribution
- therefore,  $p \leq 0.05$  will rarely happen, leading to the acceptance of  $H_0$  most of the time

```
set.seed(450)
x <- runif(50, min=2, max=4)
shapiro.test(x)
```

```
Shapiro-Wilk normality test
W = 0.9601, p-value = 0.08995
```



SW tests that the data are not normally distributed, rather than testing that they are normally distributed

# testing for normality

## Kolmogorov-Smirnov Test

- compare a sample to a reference probability distribution
- $H_0$ :  $x$  comes from the probability distribution

```
ks.test(x, "pnorm", mean, sd)
```

```
set.seed(999)
s1 <- rnorm(30, mean=0, sd=1)
ks.test(x=s1, "pnorm", mean=0, sd=1)
```

One-sample Kolmogorov-Smirnov test

```
data: s1
D = 0.21188, p-value = 0.1164
alternative hypothesis: two-sided
```

```
set.seed(999)
s2 <- rnorm(30, mean=2, sd=3)
ks.test(x=s2, "pnorm", 0, 1)
```

One-sample Kolmogorov-Smirnov test

```
data: s2
D = 0.41961, p-value = 2.656e-05
alternative hypothesis: two-sided
```

# testing for normality

## Kolmogorov-Smirnov Test

- compare whether two samples are from the same distribution
- $H_0: x_1$  and  $x_2$  come from the same distribution

```
ks.test(x = sample1, y = sample2)
```

```
set.seed(1001)
s.norm <- rnorm(n=20, mean=0, sd=1)
```

```
set.seed(1001)
s.pois <- rpois(n=20, lambda=0.7)

ks.test(x=s.norm, y=s.pois)
```

```
Two-sample Kolmogorov-Smirnov test
D = 0.55, p-value = 0.004716
alternative hypothesis: two-sided
```

Post your questions to be  
answered during lecture