USC Department Of Translational Genomics
Keck Medicine of USC

# TRGN 599: Applied Data Science and Bioinformatics

## UNIT II. Descriptive Statistics

## Week 5 - Lecture 2

**Enrique I. Velazquez Villarreal, M.D., Ph.D., M.P.H., M.S.** | **Assistant Professor**

Dept. of Translational Genomics

USC | Keck School of Medicine | Norris Comprehensive Cancer Center

Leader of the USC Bioinformatics Core – *USC CaRE2 Health Equity Center*

**David W. Craig, Ph.D.** | **Professor and Vice Chair**

Dept. of Translational Genomics

USC | Keck School of Medicine | Norris Comprehensive Cancer Center

Co-Director, Institute of Translational Genomics

# Topics

- Prevalence, incidence. Sensitivity, specificity, analytical validity.

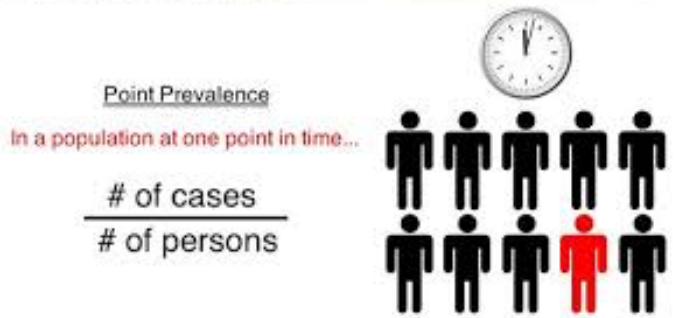
# Prevalence

- Prevalence is the proportion of a population who have a specific characteristic in a given time period.
- To estimate prevalence, researchers randomly select a sample (smaller group) from the entire population they want to describe.
- Using random selection methods increases the chances that the characteristics of the sample will be representative of (similar to) the characteristics of the population.

$$\text{Prevalence} = \frac{\text{\# of people in sample with characteristic}}{\text{Total \# of people in sample}}$$

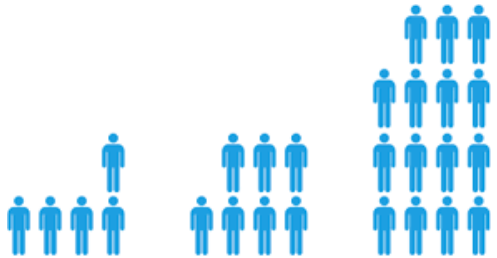
# Prevalence

- For a representative sample, prevalence is the number of people in the sample with the characteristic of interest, divided by the total number of people in the sample.
- To ensure a selected sample is representative of an entire population, statistical ‘weights’ may be applied. Weighting the sample mathematically adjusts the sample characteristics to match with the target population.



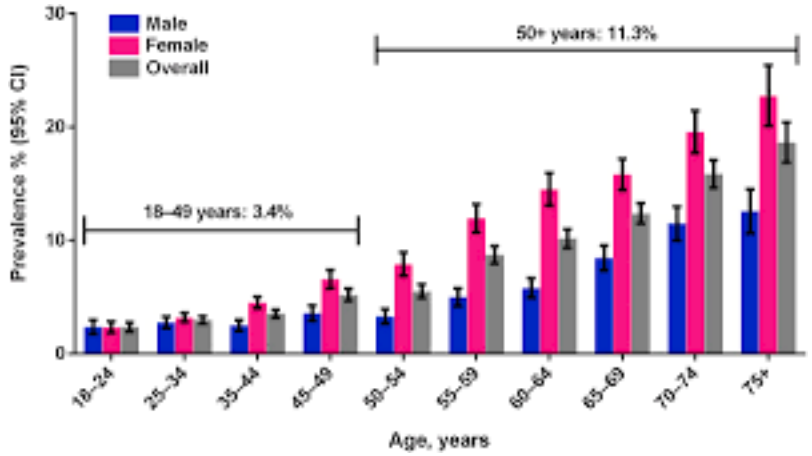
# Prevalence

- Prevalence may be reported as a percentage (5%, or 5 people out of 100), or as the number of cases per 10,000 or 100,000 people. The way prevalence is reported depends on how common the characteristic is in the population.
- There are several ways to measure and report prevalence depending on the timeframe of the estimate.



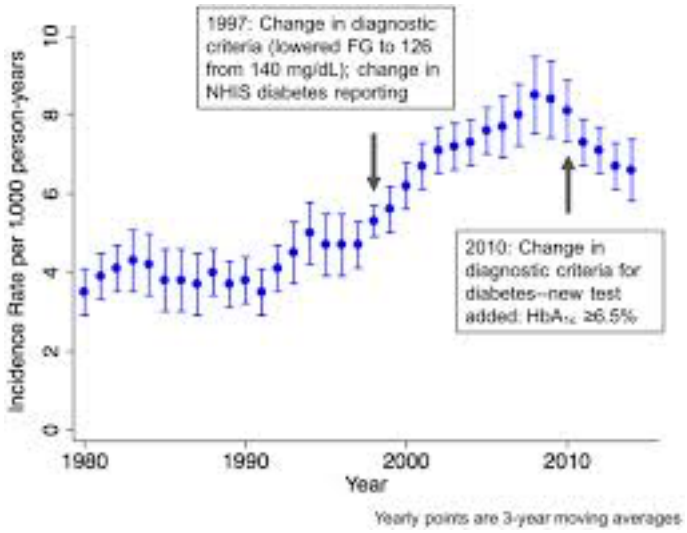
# Prevalence

- Point prevalence is the proportion of a population that has the characteristic at a specific point in time.
- Period prevalence is the proportion of a population that has the characteristic at any point during a given time period of interest. “Past 12 months” is a commonly used period.
- Lifetime prevalence is the proportion of a population who, at some point in life has ever had the characteristic.



# Incidence

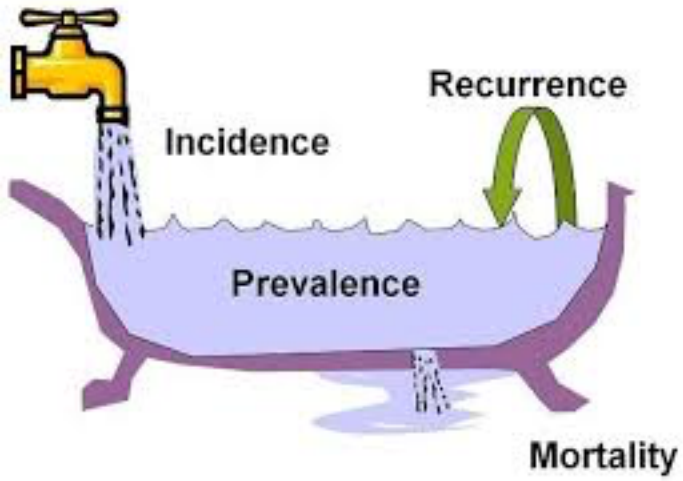
- Incidence is a measure of the number of new cases of a characteristic that develop in a population in a specified time period;
  - whereas prevalence is the proportion of a population who have a specific characteristic in a given time period, regardless of when they first developed the characteristic.



# Incidence

- Researchers may study incident (new) cases of illnesses to help identify causes and prevent additional cases.

- Incidence is often reported for infectious diseases.



# Calculating Incidence in R

# TRGN599_Week_5_Lecture_2

*Enrique I. Velazquez Villarreal, MD, PhD, MPH, MS*

*2/6/2019*

## Calculating Incidence

```
#install.packages("AER")
library(AER)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: lmtest
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
## Loading required package: survival
```

# Calculating Incidence in R

```
data(Fatalities)

?Fatalities

str(Fatalities)
```

```
## 'data.frame':    336 obs. of  34 variables:
##  $ state       : Factor w/ 48 levels "al","az","ar",..: 1 1 1 1 1 1 1 2 2 2 ...
##  $ year        : Factor w/ 7 levels "1982","1983",..: 1 2 3 4 5 6 7 1 2 3 ...
##  $ spirits     : num  1.37 1.36 1.32 1.28 1.23 ...
##  $ unemp       : num  14.4 13.7 11.1 8.9 9.8 ...
##  $ income      : num  10544 10733 11109 11333 11662 ...
##  $ emppop      : num  50.7 52.1 54.2 55.3 56.5 ...
##  $ beertax     : num  1.54 1.79 1.71 1.65 1.61 ...
##  $ baptist     : num  30.4 30.3 30.3 30.3 30.3 ...
##  $ mormon      : num  0.328 0.343 0.359 0.376 0.393 ...
##  $ drinkage    : num  19 19 19 19.7 21 ...
##  $ dry         : num  25 23 24 23.6 23.5 ...
##  $ youngdrivers: num  0.212 0.211 0.211 0.211 0.213 ...
##  $ miles       : num  7234 7836 8263 8727 8953 ...
##  $ breath      : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ jail        : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 2 2 ...
##  $ service     : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 2 2 ...
##  $ fatal       : int  839 930 932 882 1081 1110 1023 724 675 869 ...
##  $ nfatal      : int  146 154 165 146 172 181 139 131 112 149 ...
##  $ sfatal      : int  99 98 94 98 119 114 89 76 60 81 ...
##  $ fatal1517   : int  53 71 49 66 82 94 66 40 40 51 ...
##  $ nfatal1517  : int  9 8 7 9 10 11 8 7 7 8 ...
##  $ fatal1820   : int  99 108 103 100 120 127 105 81 83 118 ...
##  $ nfatal1820  : int  34 26 25 23 23 31 24 16 19 34 ...
##  $ fatal2124   : int  120 124 118 114 119 138 123 96 80 123 ...
##  $ nfatal2124  : int  32 35 34 45 29 30 25 36 17 33 ...
##  $ afatal      : num  309 342 305 277 361 ...
##  $ pop         : num  3942002 3960008 3988992 4021008 4049994 ...
##  $ pop1517     : num  209000 202000 197000 195000 204000 ...
##  $ pop1820     : num  221553 219125 216724 214349 212000 ...
##  $ pop2124     : num  290000 290000 288000 284000 263000 ...
##  $ milestot    : num  28516 31032 32961 35091 36259 ...
##  $ unempus     : num  9.7 9.6 7.5 7.2 7 ...
##  $ emppopus    : num  57.8 57.9 59.5 60.1 60.7 ...
##  $ gsp         : num  -0.0221 0.0466 0.0628 0.0275 0.0321 ...
```

# Calculating Incidence in R

```
# Calculating incidence
# Using state and year variables
(table_of_deaths<-with(Fatalities,tapply(fatal, list(state,
    year), sum)))
```

```
##    1982 1983 1984 1985 1986 1987 1988
## al  839  930  932  882 1081 1110 1023
## az  724  675  869  893 1007  937  944
## ar  550  557  525  534  603  639  610
## ca 4615 4573 5020 4960 5253 5504 5390
## co  668  646  608  579  603  591  497
## ct  515  438  469  448  450  449  484
## de  122  110  130  104  136  146  160
## fl 2653 2686 2814 2832 2830 2839 3078
## ga 1229 1296 1410 1361 1530 1599 1653
## id  256  263  242  255  258  262  257
## il 1651 1526 1547 1534 1596 1660 1837
## in  961 1016  925  974 1038 1055 1101
## ia  480  514  420  474  441  491  557
## ks  498  411  510  486  500  491  483
## ky  822  778  754  712  805  844  838
## la 1091  933  961  931  932  827  925
## me  166  224  232  206  214  232  255
## md  640  656  643  729  784  814  782
## ma  659  651  666  742  752  689  725
## mi 1392 1314 1531 1545 1605 1597 1704
## mn  571  555  582  608  571  530  612
## ms  730  715  679  662  771  756  722
## mo  890  911  967  931 1129 1044 1103
## mt  254  286  238  223  222  234  198
## ne  261  255  285  237  290  297  261
## nv  280  253  249  259  233  262  286
## nh  173  191  192  191  172  179  166
## nj 1061  932  922  964 1039 1023 1051
## nm  577  531  497  535  499  568  487
## ny 2162 2077 2060 2006 2122 2333 2255
## nc 1303 1234 1450 1482 1647 1584 1573
## nd  148  116  100   90  100  101  104
## oh 1607 1582 1646 1646 1673 1772 1763
## ok 1054  848  797  744  699  597  634
## or  518  550  572  559  619  620  677
## pa 1819 1721 1727 1771 1894 1987 1931
## ri  105  100   79  109  124  113  125
## sc  730  844  916  951 1059 1086 1034
## sd  148  175  143  130  134  134  147
## tn 1055 1037 1095 1101 1230 1248 1266
## tx 4213 3823 3912 3678 3567 3261 3393
## ut  295  283  315  303  313  296  297
## vt  107   94  114  115  109  119  129
## va  881  901 1013  976 1126 1021 1071
## wa  748  698  746  744  703  780  778
## wv  450  425  438  420  440  471  460
## wi  770  725  822  744  747  797  807
## wy  201  173  157  152  168  129  155
```

# Calculating Incidence in R

```
(table_of_exp<-with(Fatalities,tapply(milestot, list(state,
    year), sum)))
```

```
##      1982   1983   1984   1985   1986   1987   1988
## al  28516  31032  32961  35091  36259  37426  39684
## az  19729  19611  20613  21580  26655  31729  34247
## ar  16630  16684  16621  17112  17709  18306  19219
## ca 169999 182652 196537 207600 216951 226301 241575
## co  23786  24109  24588  26146  26557  26968  27665
## ct  20138  20630  21076  22152  24464  26775  26062
## de   4591   4886   5138   5365   5726   6086   6404
## fl  79498  81776  85475  88056  90848  93639 105319
## ga  48731  48837  50486  53713  57003  60293  62262
## id   7857   8287   7768   7710   7915   8119   8127
## il  65385  67370  69910  70844  73300  75756  78483
## in  39203  39837  41074  40782  42452  44122  51124
## ia  19341  19661  20497  20191  20500  20808  21907
## ks  17658  18153  18717  19275  19918  20561  21161
## ky  25627  26719  27951  28520  29285  30320  31614
## la  26902  27573  31588  33365  31982  30599  34682
## me   7649   7924   9345   9277  10021  10766  11401
## md  28920  30618  31702  33337  34915  36493  37498
## ma  36666  37541  38537  39696  41001  42305  43334
## mi  61200  60855  63470  67402  71554  75706  77899
## mn  29176  31063  31826  32688  33928  35167  36447
## ms  17146  17802  18442  19130  19652  20173  22043
## mo  35003  36543  38535  39284  41332  43379  45570
## mt   6669   7181   7386   7572   7823   8074   8138
## ne  11435  11534  41968  12054  12573  13091  13407
## nv   6413   6872   7332   7566   7981   8396   8989
## nh   6971   7181   7294   7538   8353   9167   9507
## nj  51802  52217  52312  53108  55090  57071  58671
## nm  11850  11678  12432  13269  14193  15116  15283
## ny  80484  83783  87268  90518  94260  98002 103692
## nc  43100  45038  48182  49923  52262  54600  57943
## nd   5252   5363   5377   5387   5534   5681   5765
## oh  71751  73214  74895  75549  77353  79157  81990
## ok  30011  29565  30981  31181  31394  31606  32388
## or  19384  20557  20943  21458  22395  23332  25204
## pa  71313  72302  74297  75428  77027  78626  81238
## ri   5908   6014   5300   5823   5913   6003   5853
## sc  24222  24977  25971  26677  28451  30224  31759
## sd   6361   6317   6401   6277   6243   6209   6634
## tn  34793  36261  36523  36258  39192  42126  44193
## tx 125218 131883 137737 143263 147225 151186 156458
## ut  10925  11221  11661  12037  12358  12679  13263
## vt   3993   4151   4403   4688   4864   5039   5553
## va  41430  42299  44527  47928  51381  54834  57453
## wa  31258  36144  34248  34375  36448  38520  41813
## wv  10932  11696  12671  12664  13203  13742  13884
## wi  32794  34106  35367  36679  38438  40196  42458
## wy   5281   5059   5127   5401   5384   5367   5658
```

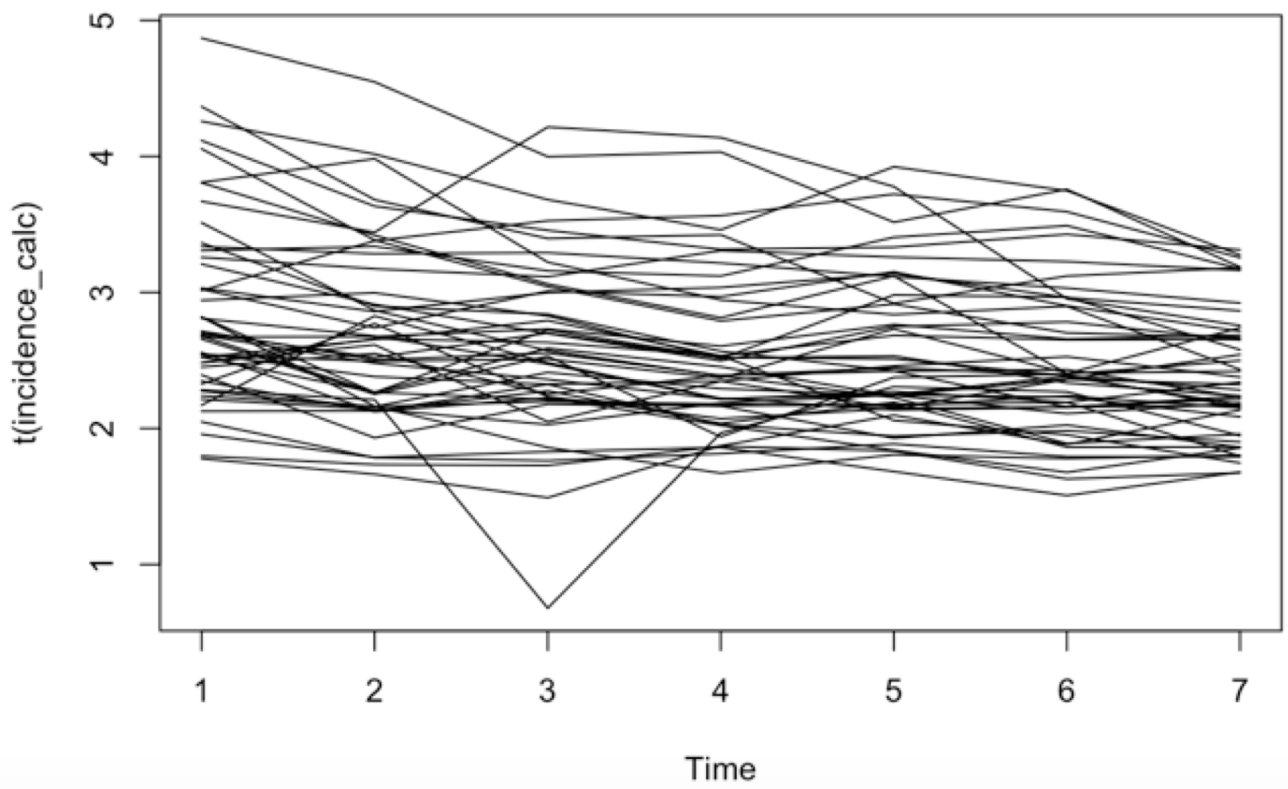
# Calculating Incidence in R

```
incidence_calc<-table_of_deaths/table_of_exp*100
incidence_calc
```

```
##        1982     1983      1984     1985     1986     1987     1988
## al 2.942208 2.996906 2.8275841 2.513465 2.981329 2.965853 2.577865
## az 3.669725 3.441946 4.2157862 4.138091 3.777903 2.953134 2.756446
## ar 3.307276 3.338528 3.1586547 3.120617 3.405048 3.490659 3.173942
## ca 2.714722 2.503668 2.5542264 2.389210 2.421284 2.432159 2.231191
## co 2.808375 2.679497 2.4727509 2.214488 2.270588 2.191486 1.796494
## ct 2.557354 2.123122 2.2252799 2.022391 1.839438 1.676937 1.857110
## de 2.657373 2.251330 2.5301674 1.938490 2.375131 2.398948 2.498438
## fl 3.337191 3.284582 3.2921907 3.216135 3.115093 3.031856 2.922549
## ga 2.522008 2.653726 2.7928535 2.533837 2.684069 2.652049 2.654910
## id 3.258241 3.173645 3.1153450 3.307393 3.259634 3.226998 3.162299
## il 2.525044 2.265103 2.2128451 2.165321 2.177353 2.191246 2.340634
## in 2.451343 2.550393 2.2520329 2.388309 2.445114 2.391097 2.153587
## ia 2.481774 2.614313 2.0490804 2.347581 2.151220 2.359669 2.542566
## ks 2.820251 2.264089 2.7247956 2.521401 2.510292 2.388016 2.282501
## ky 3.207555 2.911786 2.6975779 2.496494 2.748848 2.783641 2.650724
## la 4.055461 3.383745 3.0422945 2.790349 2.914139 2.702703 2.667090
## me 2.170218 2.826855 2.4826110 2.220545 2.135515 2.154932 2.236646
## md 2.213001 2.142531 2.0282632 2.186759 2.245453 2.230565 2.085445
## ma 1.797305 1.734104 1.7282093 1.869206 1.834102 1.628649 1.673051
## mi 2.274510 2.159231 2.4121632 2.292217 2.243061 2.109476 2.187448
## mn 1.957088 1.786692 1.8286935 1.860010 1.682976 1.507095 1.679151
## ms 4.257553 4.016403 3.6818133 3.460533 3.923265 3.747583 3.275416
## mo 2.542639 2.492954 2.5094070 2.369922 2.731540 2.406694 2.420452
## mt 3.808667 3.982732 3.2223125 2.945061 2.837786 2.898192 2.433030
## ne 2.282466 2.210855 0.6790888 1.966152 2.306530 2.268734 1.946744
## nv 4.366131 3.681607 3.3960720 3.423209 2.919434 3.120534 3.181666
## nh 2.481710 2.659797 2.6323005 2.533829 2.059140 1.952656 1.746082
## nj 2.048183 1.784859 1.7625019 1.815169 1.886005 1.792504 1.791345
## nm 4.869198 4.547011 3.9977477 4.031954 3.515818 3.757608 3.186547
## ny 2.686248 2.479023 2.3605445 2.216134 2.251220 2.380564 2.174710
## nc 3.023202 2.739909 3.0094226 2.968572 3.151429 2.901099 2.714737
## nd 2.817974 2.162968 1.8597731 1.670689 1.807011 1.777856 1.803990
## oh 2.239690 2.160789 2.1977435 2.178718 2.162812 2.238589 2.150262
## ok 3.512046 2.868256 2.5725445 2.386069 2.226540 1.888882 1.957515
## or 2.672307 2.675488 2.7312228 2.605089 2.764010 2.657295 2.686082
## pa 2.550727 2.380294 2.3244546 2.347934 2.458878 2.527154 2.376966
## ri 1.777251 1.662787 1.4905660 1.871887 2.097074 1.882392 2.135657
## sc 3.013789 3.379109 3.5270109 3.564869 3.722189 3.593171 3.255770
## sd 2.326678 2.770302 2.2340259 2.071053 2.146404 2.158158 2.215858
## tn 3.032219 2.859822 2.9981108 3.036571 3.138396 2.962541 2.864707
## tx 3.364532 2.898781 2.8401954 2.567306 2.422822 2.156946 2.168633
## ut 2.700229 2.522057 2.7013121 2.517239 2.532772 2.334569 2.239312
## vt 2.679689 2.264515 2.5891438 2.453072 2.240954 2.361580 2.323069
## va 2.126478 2.130074 2.2750241 2.036388 2.191472 1.861983 1.864132
## wa 2.392987 1.931164 2.1782294 2.164364 1.928775 2.024922 1.860665
## wv 4.116356 3.633721 3.4567122 3.316488 3.332576 3.427449 3.313166
## wi 2.347990 2.125726 2.3242005 2.028409 1.943389 1.982784 1.900702
## wy 3.806097 3.419648 3.0622196 2.814294 3.120357 2.403577 2.739484
```

# Calculating Incidence in R

```
# Ploting and Transposing data
plot.ts(t(incidence_calc), plot.type="single")
```



# Calculating Incidence in R – R Markdown

~/Documents/R_working_directory/Rmarkdown_TRGN599_Week_5_Lecture_2.html

## TRGN599_Week_5_Lecture_2

*Enrique I. Velazquez Villarreal, MD, PhD, MPH, MS*

*2/6/2019*

### Calculating Incidence

```
#install.packages("AER")
library(AER)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: lmtest
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
## Loading required package: survival
```

```
data(Fatalities)

?Fatalities

str(Fatalities)
```

```
## 'data.frame':    336 obs. of  34 variables:
##  $ state       : Factor w/ 48 levels "al","az","ar",..: 1 1 1 1 1 1 1 2 2 2 ...
##  $ year        : Factor w/ 7 levels "1982","1983",..: 1 2 3 4 5 6 7 1 2 3 ...
##  $ spirits     : num  1.37 1.36 1.32 1.28 1.23 ...
##  $ unemp       : num  14.4 13.7 11.1 8.9 9.8 ...
##  $ income      : num  10544 10733 11109 11333 11662 ...
##  $ emppop      : num  50.7 52.1 54.2 55.3 56.5 ...
##  $ beertax     : num  1.54 1.79 1.71 1.65 1.61 ...
##  $ baptist     : num  30.4 30.3 30.3 30.3 30.3 ...
##  $ mormon      : num  0.328 0.343 0.359 0.376 0.393 ...
##  $ drinkage    : num  19 19 19 19.7 21 ...
##  $ dry         : num  25 23 24 23.6 23.5 ...
##  $ youngdrivers: num  0.212 0.211 0.211 0.211 0.213 ...
##  $ miles       : num  7234 7836 8263 8727 8953 ...
##  $ breath      : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ jail        : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 2 2 ...
##  $ service     : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 2 2 ...
##  $ fatal       : int  839 930 932 882 1081 1110 1023 724 675 869 ...
##  $ nfatal      : int  146 154 165 146 172 181 139 131 112 149 ...
##  $ sfatal      : int  99 98 94 98 119 114 89 76 60 81 ...
##  $ fatal1517   : int  53 71 49 66 82 94 66 40 40 51 ...
##  $ nfatal1517  : int  9 8 7 9 10 11 8 7 7 8 ...
##  $ fatal1820   : int  99 108 103 100 120 127 105 81 83 118 ...
##  $ nfatal1820  : int  34 26 25 23 23 31 24 16 19 34 ...
##  $ fatal2124   : int  120 124 118 114 119 138 123 96 80 123 ...
##  $ nfatal2124  : int  32 35 34 45 29 30 25 36 17 33 ...
##  $ afatal      : num  309 342 305 277 361 ...
##  $ pop         : num  3942002 3960008 3988992 4021008 4049994 ...
##  $ pop1517     : num  209000 202000 197000 195000 204000 ...
##  $ pop1820     : num  221553 219125 216724 214349 212000 ...
##  $ pop2124     : num  290000 290000 288000 284000 263000 ...
##  $ milestot    : num  28516 31032 32961 35091 36259 ...
##  $ unempus     : num  9.7 9.6 7.5 7.2 7 ...
##  $ emppopus    : num  57.8 57.9 59.5 60.1 60.7 ...
```

# Sensitivity and Specificity

- Sensitivity and specificity are statistical measures of the performance of a binary classification test or classification function:
  - Sensitivity (true positive rate)
    - measures the proportion of actual positives that are correctly identified as such.
    - Example: the percentage of sick people who are correctly identified as having the condition.

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

$$= \frac{\text{number of true positives}}{\text{total number of sick individuals in population}}$$

# Sensitivity and Specificity

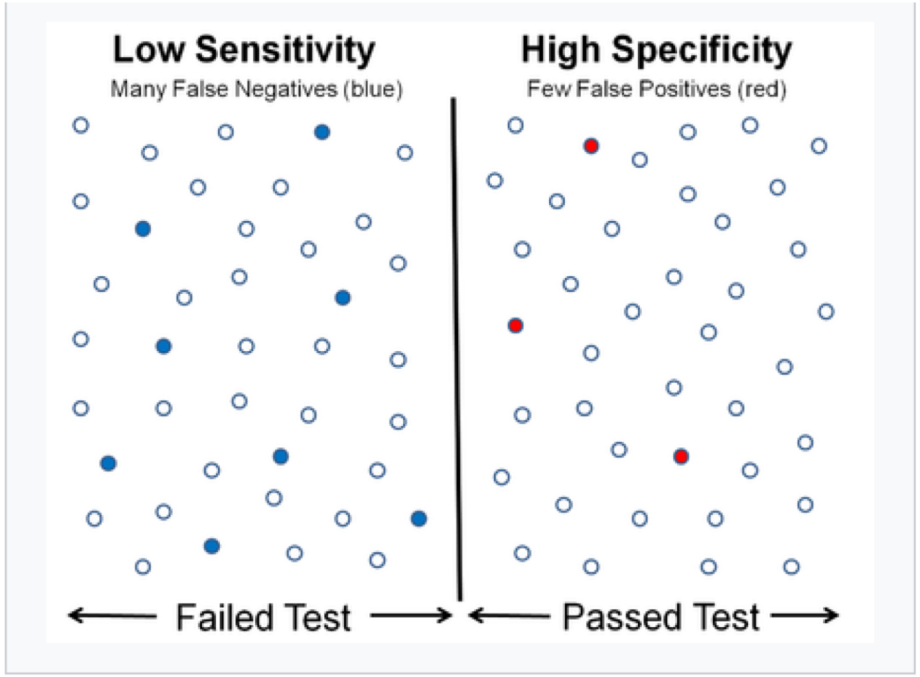
- Specificity (true negative rate)
  - measures the proportion of actual negatives that are correctly identified as such
  - Example: the percentage of healthy people who are correctly identified as not having the condition.

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

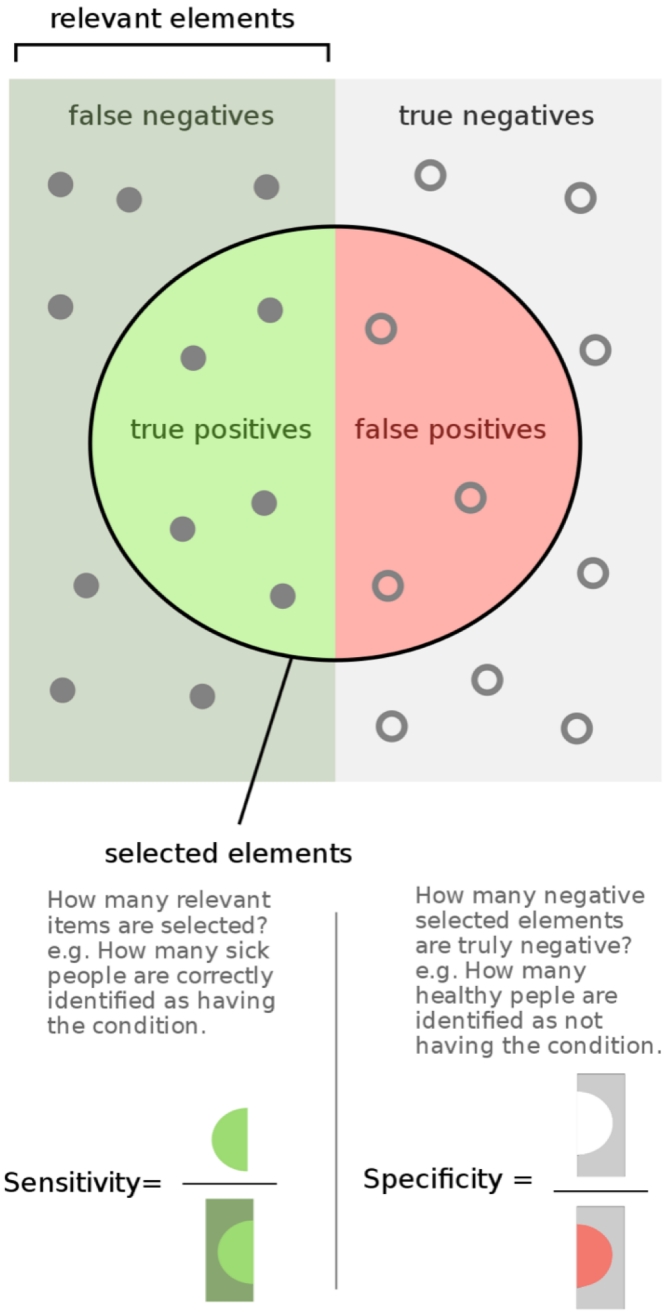
$$= \frac{\text{number of true negatives}}{\text{total number of well individuals in population}}$$

# Sensitivity and Specificity


High sensitivity and low specificity

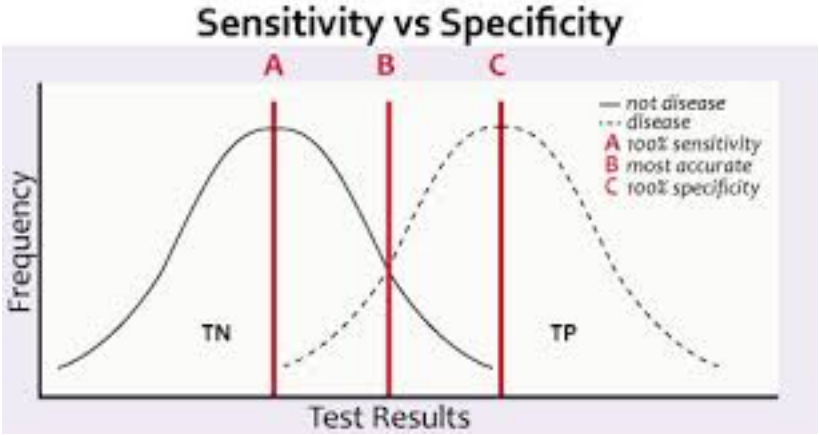



Low sensitivity and high specificity



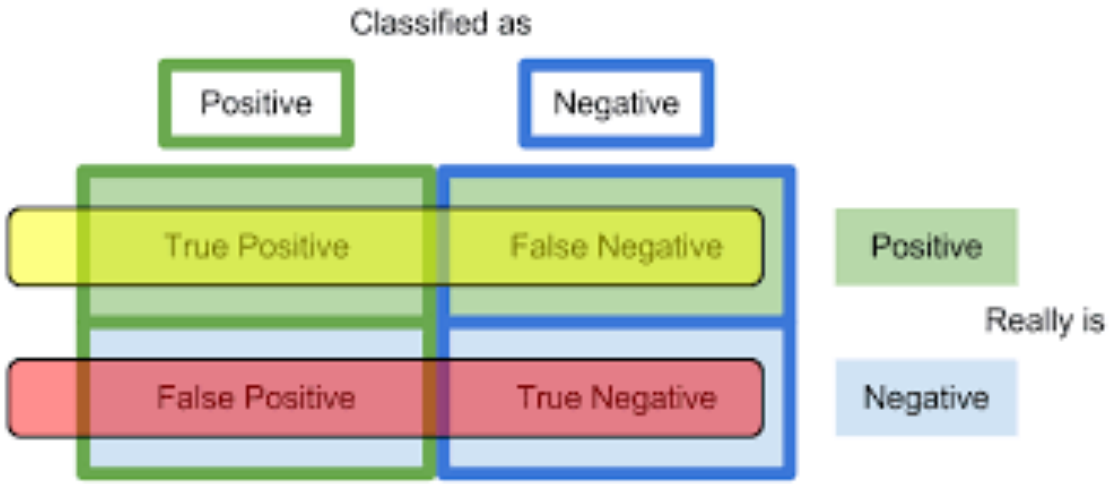
# Sensitivity and Specificity

- Application to screening study:
  - Imagine a study evaluating a new test that screens people for a disease.
  - Each person taking the test either has or does not have the disease.
  - The test outcome can be positive (classifying the person as having the disease) or negative (classifying the person as not having the disease).
  - The test results for each subject may or may not match the subject's actual status.
    - True positive: Sick people correctly identified as sick
    - False positive: Healthy people incorrectly identified as sick
    - True negative: Healthy people correctly identified as healthy
    - False negative: Sick people incorrectly identified as healthy



# Sensitivity and Specificity

- If Positive = identified and negative = rejected:
  - True positive = correctly identified
  - False positive = incorrectly identified
  - True negative = correctly rejected
  - False negative = incorrectly rejected



# Confusion Matrix

- If we consider a group with P positive instances and N negative instances of some condition:
  - The four outcomes can be formulated in a 2×2 contingency table or confusion matrix.

| | | **True condition** | | | |
|---|---|---|---|---|---|
| | Total population | Condition positive | Condition negative | Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$ | Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$ |
| **Predicted condition** | Predicted condition positive | **True positive**, Power | **False positive**, Type I error | Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$ | False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$ |
| | Predicted condition negative | **False negative**, Type II error | **True negative** | False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$ | Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$ |
| | | True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$ | False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$ | Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$ | Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$ |
| | | False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$ | Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$ | Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$ | $F_1$ score $= \frac{1}{\frac{1}{2}\left(\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}\right)}$ |

# Terminology and Derivations from Confusion Matrix

**condition positive (P)**

the number of real positive cases in the data

**condition negative (N)**

the number of real negative cases in the data

**true positive (TP)**

eqv. with hit

**true negative (TN)**

eqv. with correct rejection

**false positive (FP)**

eqv. with false alarm, Type I error

**false negative (FN)**

eqv. with miss, Type II error

**sensitivity, recall, hit rate, or true positive rate (TPR)**

$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{FNR}$$

**specificity, selectivity or true negative rate (TNR)**

$$\mathrm{TNR} = \frac{\mathrm{TN}}{N} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} = 1 - \mathrm{FPR}$$

**precision or positive predictive value (PPV)**

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

**negative predictive value (NPV)**

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}$$

**miss rate or false negative rate (FNR)**

$$\mathrm{FNR} = \frac{\mathrm{FN}}{P} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}} = 1 - \mathrm{TPR}$$

**fall-out or false positive rate (FPR)**

$$\mathrm{FPR} = \frac{\mathrm{FP}}{N} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = 1 - \mathrm{TNR}$$

**false discovery rate (FDR)**

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}} = 1 - \mathrm{PPV}$$

**false omission rate (FOR)**

$$\mathrm{FOR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TN}} = 1 - \mathrm{NPV}$$

**accuracy (ACC)**

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{P + N} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

**F1 score**

is the harmonic mean of precision and sensitivity

$$F_1 = 2 \cdot \frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

**Matthews correlation coefficient (MCC)**

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

**Informedness or Bookmaker Informedness (BM)**

$$\mathrm{BM} = \mathrm{TPR} + \mathrm{TNR} - 1$$

**Markedness (MK)**

$$\mathrm{MK} = \mathrm{PPV} + \mathrm{NPV} - 1$$

*Sources: Fawcett (2006), Powers (2011), and Ting (2011)* [2] [3] [4]

# Sensitivity and Specificity

~/Documents/R_working_directory/Rmarkdown_TRGN599_Week_5_Lecture_2_1.html

## TRGN599_Week_5_Lecture_2_1

*Enrique I. Velazquez Villarreal, MD, PhD, MPH, MS*

*2/6/2019*

### Calculating Sensitivity, Specificity and Predictive Value of a diagnostic.

### Steps will be explained in detail during class.

### From epiR package:

### Scott et al. 2008, Table 1:

### A new diagnostic test was trialled on 1586 patients.

### Of 744 patients that were disease positive, 670 tested positive.

### Of 842 patients that were disease negative, 640 tested negative.

### What is the likeliood ratio of a positive test?

### What is the number needed to diagnose?

# Sensitivity and Specificity

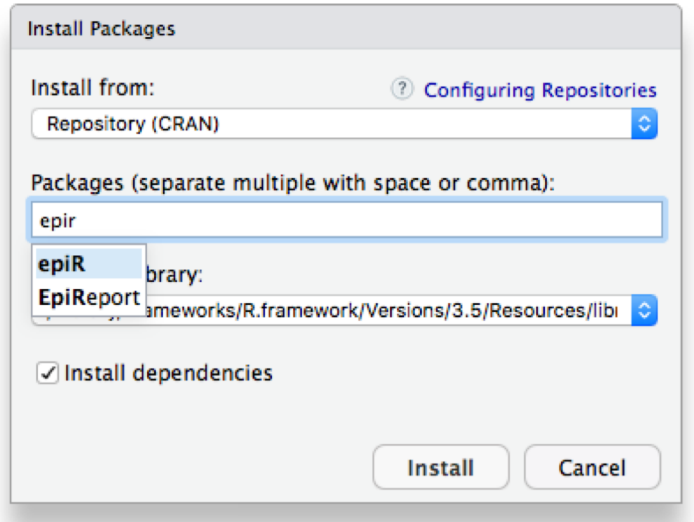
```
# Installing epiR
# install.packages("epiR")

library(epiR)
```

```
## Loading required package: survival
```

```
## Package epiR 0.9-99 is loaded
```

```
## Type help(epi.about) for summary information
```

```
##
```




```
TRGN599_dat1 <- as.table(matrix(c(670,202,74,640), nrow = 2, byrow = TRUE))
colnames(TRGN599_dat1) <- c("Dis+","Dis-")
rownames(TRGN599_dat1) <- c("Test+","Test-")
TRGN599_Eval <- epi.tests(TRGN599_dat1, conf.level = 0.95)
print(TRGN599_Eval); summary(TRGN599_Eval)
```

```
##           Outcome +    Outcome -      Total
## Test +          670          202        872
## Test -           74          640        714
## Total           744          842       1586
##
## Point estimates and 95 % CIs:
## -------------------------------------------------------
## Apparent prevalence                   0.55 (0.52, 0.57)
## True prevalence                       0.47 (0.44, 0.49)
## Sensitivity                           0.90 (0.88, 0.92)
## Specificity                           0.76 (0.73, 0.79)
## Positive predictive value             0.77 (0.74, 0.80)
## Negative predictive value             0.90 (0.87, 0.92)
## Positive likelihood ratio             3.75 (3.32, 4.24)
## Negative likelihood ratio             0.13 (0.11, 0.16)
## -------------------------------------------------------
```

# Sensitivity and Specificity

~/Documents/R_working_directory/Rmarkdown_TRGN599_Week_5_Lecture_2_1.html

## TRGN599_Week_5_Lecture_2_1

*Enrique I. Velazquez Villarreal, MD, PhD, MPH, MS*

*2/6/2019*

## Calculating Sensitivity, Specificity and Predictive Value of a diagnostic.

## Steps will be explained in detail during class.

## From epiR package:

## Scott et al. 2008, Table 1:

## A new diagnostic test was trialled on 1586 patients.

## Of 744 patients that were disease positive, 670 tested positive.

## Of 842 patients that were disease negative, 640 tested negative.

## What is the likeliood ratio of a positive test?

## What is the number needed to diagnose?

```
# Installing epiR
# install.packages("epiR")

library(epiR)
```

```
## Loading required package: survival
```

```
## Package epiR 0.9-99 is loaded
```

```
## Type help(epi.about) for summary information
```

```
##
```

```
TRGN599_dat1 <- as.table(matrix(c(670,202,74,640), nrow = 2, byrow = TRUE))
colnames(TRGN599_dat1) <- c("Dis+","Dis-")
rownames(TRGN599_dat1) <- c("Test+","Test-")
TRGN599_Eval <- epi.tests(TRGN599_dat1, conf.level = 0.95)
print(TRGN599_Eval); summary(TRGN599_Eval)
```

```
##          Outcome +    Outcome -      Total
## Test +         670          202        872
## Test -          74          640        714
## Total          744          842       1586
##
## Point estimates and 95 % CIs:
## -------------------------------------------------------------
## Apparent prevalence                   0.55 (0.52, 0.57)
## True prevalence                       0.47 (0.44, 0.49)
## Sensitivity                           0.90 (0.88, 0.92)
## Specificity                           0.76 (0.73, 0.79)
## Positive predictive value             0.77 (0.74, 0.80)
## Negative predictive value             0.90 (0.87, 0.92)
## Positive likelihood ratio             3.75 (3.32, 4.24)
## Negative likelihood ratio             0.13 (0.11, 0.16)
## -------------------------------------------------------------
```