Biostats 200 – Lab 2

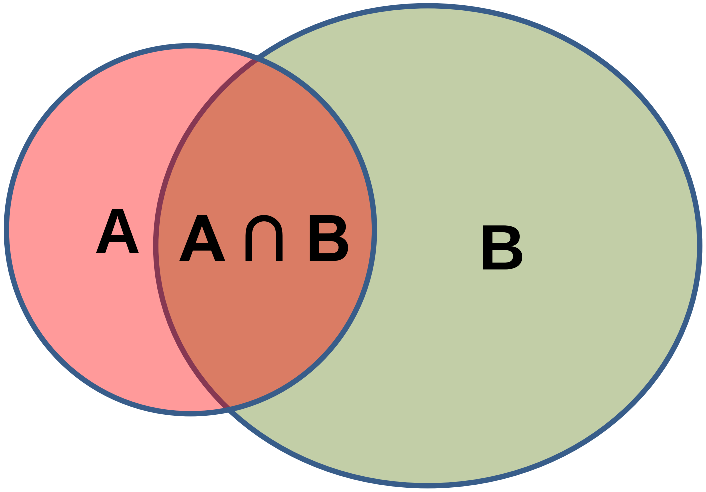
Annie Narla MD (Adapted from Dr Judy Hahn's slides)

- <u>Probability of an event</u> – relative frequency of its occurrence in a large number of trials repeated under the same conditions
- <u>Complement</u> of an event, Ā or A^C: P(A)= 1-P(Ā)

- The <u>intersection</u> of 2 events is written A ∩ B: when <u>both A and B</u> occur
- The <u>union</u> of 2 events is written A U B: when <u>either A or B or both occur</u>

Union

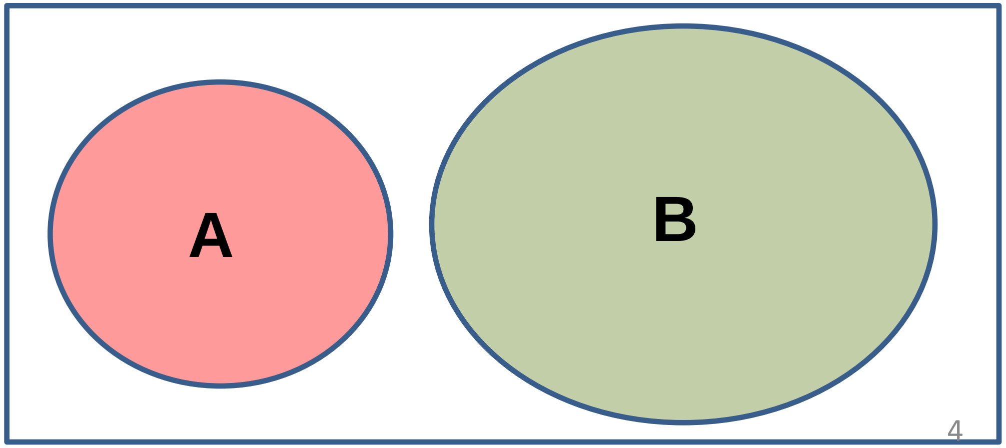
- **P(A U B) = P(A) + P(B) – P(A ∩ B)**
- The probability of A or B or both is the sum of their individual probabilities minus the probability of their intersection



Mutual exclusivity

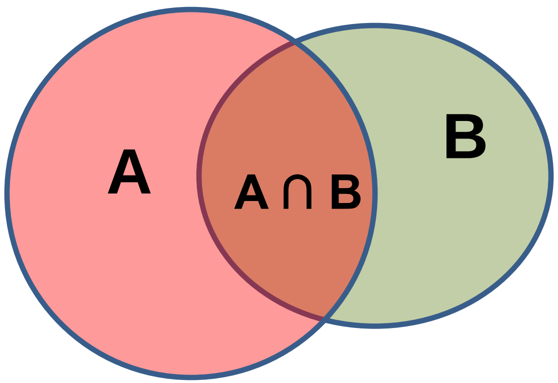
- Two events are <u>mutually exclusive</u> if they cannot occur together
  - There is no overlap area because both can't happen together
  - Therefore, for mutually exclusive events, <u>additive rule:</u>
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

  $= P(A) + P(B)$



Conditional probability

$P(B|A) = P(A \cap B) / P(A)$



It is the relative size of the probability of the intersection $A \cap B$ compared to the relative size of the probability of A occurring

Independence: If the occurrence of B does not depend on A, then by the definition of independence, $P(B|A) = P(B)$ -> $P(A \cap B) = P(A)\, P(B)$ (multiplicative rule)

Independence vs. mutual exclusivity

– Mutual exclusitivity gets us the additive rule
  – If A and B are mutually exclusive, then

  $P(A \cup B) = P(A) + P(B)$

– Independence gets us the multiplication rule
  – If A and B are independent, then

  $P(A \cap B) = P(A)*P(B)$

Remember:

-- $P(B|A) = P(A \cap B) / P(A)$

--$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Diagnostic Tests:

- $P(T^+|D^+)$ = Sensitivity (true positives)
- $P(T^-|D^-)$ = Specificity (true negatives)
- $P(T^-|D+)$ = 1-sensitivity (false negatives)
- $P(T^+|D-)$ = 1-specificity (false positives)

- BAYES THEOREM:
- $P(D+|T^+) = P(T^+|D+)*P(D+)/ [P(T^+|D+) P(D+) + P(T^+|D-) P(D-)]$
- $P(D+|T^+)$ = positive predictive value

Binomial Distribution

- The probability distribution of a discrete random variable X
- Assumptions: There are a fixed number of independent trials n, each of which results in 1 of 2 mutually exclusive outcomes. Prob of success, p, is constant for each trial.

$$P(X=x)=\binom{n}{x}p^{x}(1-p)^{n-x}$$

In Stata:
Display comb(n,k)

Binomial Distribution: stata

- For $P(X=k)$: Binomialp (n,k,p)
- For $P(X \geq k)$: Binomialtail (n,k,p)
- Mean: np
- SD: square root of $np(1-p)$

Example:

- The probability that a student is accepted to a prestigious college is 0.3. If 5 students from the same school apply, what is the probability that at most 2 are accepted?
- To solve this problem, we compute 3 individual probabilities, using the binomial formula. The sum of all these probabilities is the answer. Thus,
- $b(5, x \leq 2, 0.3) = b(5, x = 0, 0.3) + b(5, x = 1, 0.3) + b(5, x = 2, 0.3) = 0.1681 + 0.3601 + 0.3087 = 0.8369$
- =di 1-binomialtail(5,3,0.3)=0.8369

Normal Distribution

- Used for <u>continuous variables</u> that cover the entire range (whereas binomial is for <u>discrete</u> random variables)
- Gaussian, bell-shaped, unimodal, symmetric, mean=mode=median.
- <u>Standard Normal Distribution:</u> has mean $\mu = 0$ and standard deviation $\sigma = 1$
- $Z = (X - \mu)/\sigma$ is a *standard* normal random variable

Normal Distribution

- A continuous variable X can take on an infinite number of values, therefore $P(X=x)=0$
- However, we can calculate $P(X \geq x)$, which is the area under the normal curve from x to infinity

• In Stata, the left portion of the curve P(Z<z) is calculated for you.

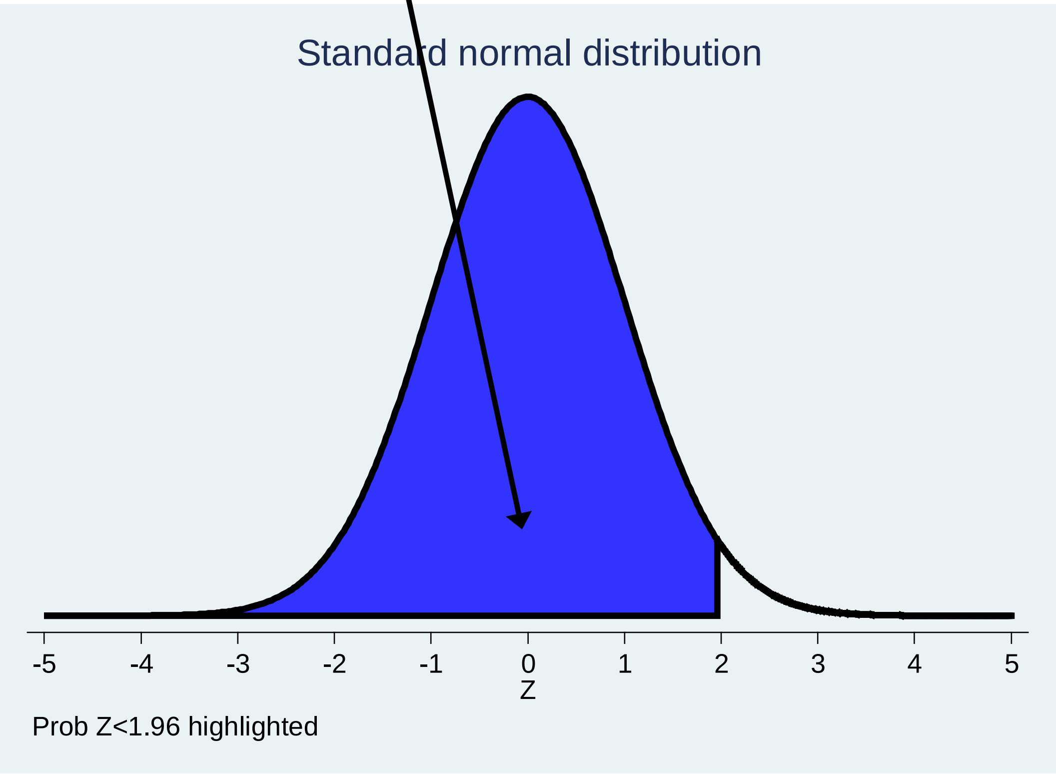
```
*** this gives you P(Z<1.96)
display normal(1.96)
.9750021
```

• If you want the right hand portion of the curve, P(Z≥z), you subtract your answer from 1

```
** this gives you P(Z<1.96)
display 1-normal(1.96)
.0249979
```

• If you want the middle:

```
 display normal(1.96)
                -normal(-1.96)
.95000421
```



Remember: book gives you Right portion of the curve P(Z>z).

Finding z values for probabilities in Stata

- To get the z value for P(Z<z) = p use

```
display invnormal(p)
```

E.g. what is the z value for P(Z<z) = 0.025

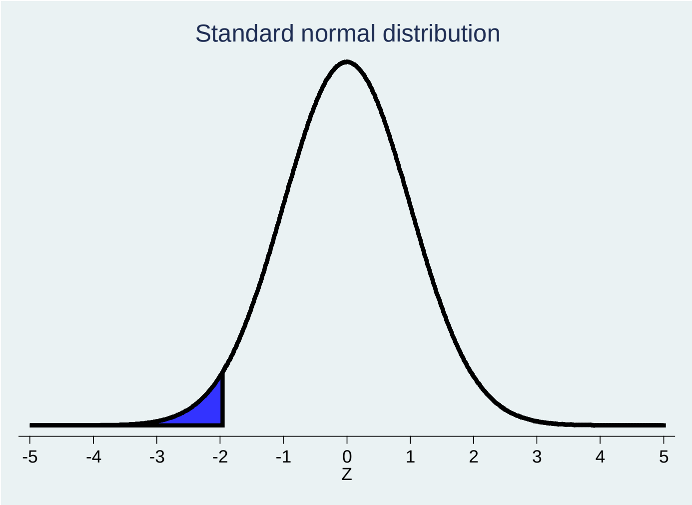
```
display invnormal(0.025)
 -1.959964
```

- To get the z value for P(Z≥z) = p use

```
display invnormal(1-p)
```

E.g. what is the z value for P(Z ≥ z) = 0.025

```
display invnormal(1-.025)
 1.959964
```

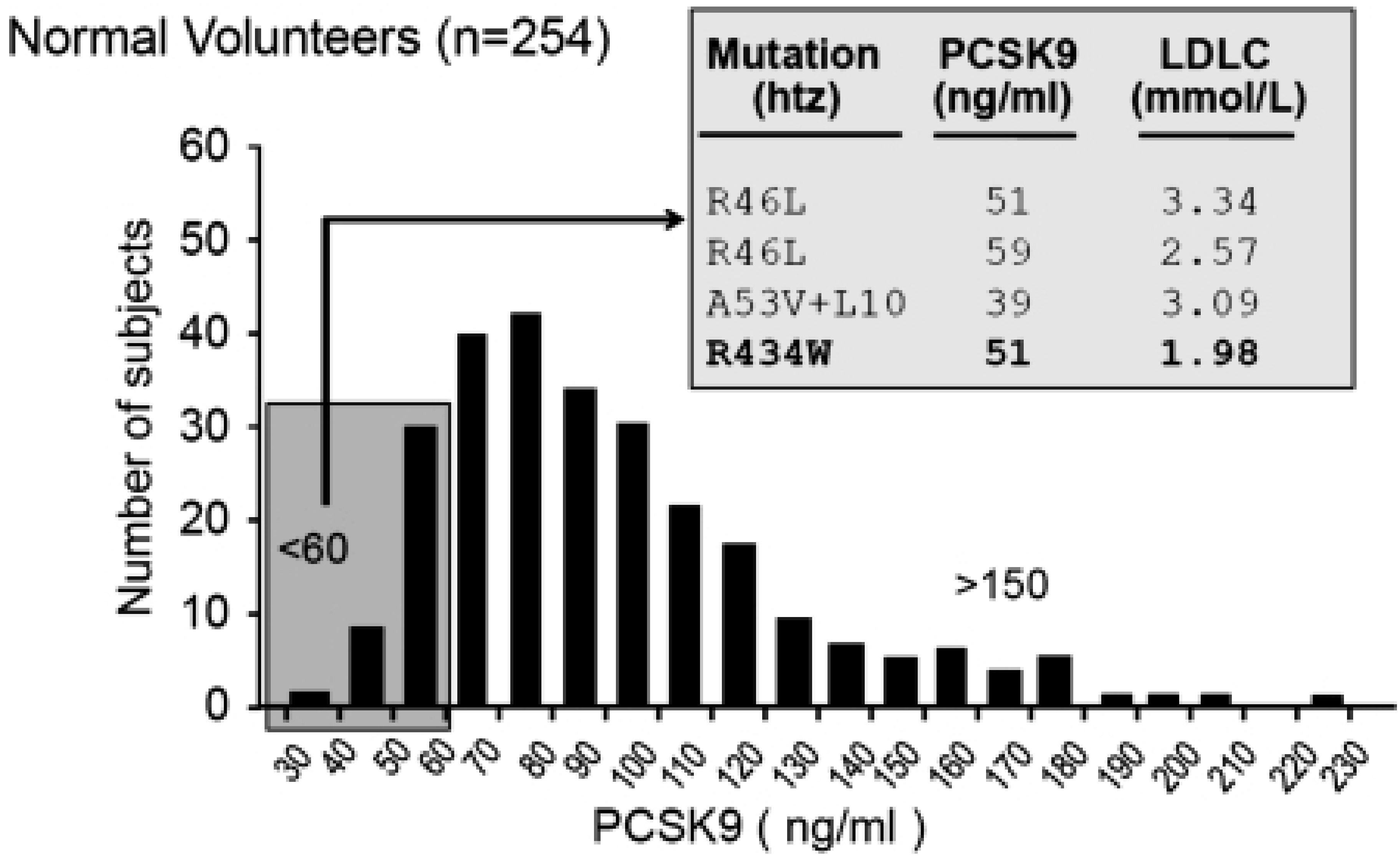


STATA:

- `Display comb(n,k)`
- `For P(X=k): Binomialp (n,k,p)`
- `For P(X≥k):Binomialtail (n,k,p)`
- the left portion of the curve P(Z<z) : `display normal(1.96)`
- right hand portion of the curve, P(Z≥z): `display 1-normal(1.96)`
- Middle: `display normal(1.96)-normal(-1.96)`
- To get the z value for P(Z<z) = p use
  `display invnormal(p)` E.g. what is the z value for P(Z<z) = 0.025: `display invnormal(0.025)`
- To get the z value for P(Z≥z) = p use
  `display invnormal(1-p)` E.g. what is the z value for P(Z ≥ z) = 0.025
  `display invnormal(1-.025)`

Example (PCSK9)

- PCSK9 now recognized as important contributor in cholesterol homeostasis and promising target for cholesterol-lowering therapy in CAD.
- PCSK9 leads to destruction of LDLR and therefore, increased plasma LDL-C levels.
- Dubuc et al found: Sequencing PCSK9 levels from individuals at the extremes of the normal PCSK9 distribution identified a new loss-of-function R434W mutation associated with lower levels of PCSK9 and LDL-C.


Normal Volunteers (n=254)
Mutation (htz)
PCSK9 (ng/ml)
LDLC (mmol/L)
R46L 51 3.34
R46L 59 2.57
A53V+L10 39 3.09
R434W 51 1.98
Number of subjects
0
10
20
30
40
50
60
<60
>150
30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230
PCSK9 (ng/ml)