Probability Distributions Part II

Session 3

Module 1 Probability & Statistical Inference

The Summer Institutes
DEPARTMENT OF BIOSTATISTICS
SCHOOL OF PUBLIC HEALTH
UNIVERSITY of WASHINGTON

Multinomial Distribution

Multinomial distribution generalizes beyond 2 outcomes of the binomial distribution.

For example, allows calculation of the following probabilities for n=3 offspring of parents that are heterozygote carriers of a recessive trait.

Q_1: 1 will be unaffected (AA), 1 will be affected (aa) and 1 will be a carrier (Aa).

Q_2: All 3 offspring will be carriers (Aa, Aa, Aa).

Q_3: 2 of the 3 offspring will be affected (aa) and 1 will be a carrier (Aa).

Multinomial Distribution

For each offspring, the 3 possible outcomes can be represented by:

Y_{i1} = 1 if i^{th} offpring is unaffected (AA),
= 0 otherwise

Y_{i2} = 1 if i^{th} offpring is a carrier (Aa),
= 0 otherwise

Y_{i3} = 1 if i^{th} offpring is affected (aa),
= 0 otherwise

Only one of Y_{i1}, Y_{i2}, Y_{i3} can be equal to 1 (so $Y_{i1} + Y_{i2} + Y_{i3} = 1$).

For the binomial distribution with 2 outcomes (e.g., unaffected vs. carrier/affected), there are 2^n unique outcomes in n trials. With n=3 offspring, there are $2^3 = 8$ unique outcomes.

For the multinomial distribution with 3 outcomes, the number of unique outcomes in n trials is 3^n. With 3 offspring, there are 3^3=27 unique outcomes.

Calculating Possible Outcomes

As with the binomial distribution, when order doesn't matter, the total number of possible outcomes, can be calculated using combinations.

For the multinomial distribution, the combinations are calculated as:

$$C_k^n = \frac{n!}{k_1!k_2 \ldots k_J!}$$

where k_j (j=1, 2,…, J) correspond to the totals for the different outcomes.

e.g. n=2 offspring
J=3 possible outcomes (unaffected/carrier/affected)

Offspring

1	2	Outcome
AA	AA	2 unaffected, 0 carrier, 0 affected
AA	Aa	1 unaffected, 1 carrier, 0 affected
Aa	AA	1 unaffected, 1 carrier, 0 affected
AA	aa	1 unaffected, 0 carrier, 1 affected
aa	AA	1 unaffected, 0 carrier, 1 affected
Aa	Aa	0 unaffected, 2 carrier, 0 affected
aa	Aa	0 unaffected, 1 carrier, 1 affected
Aa	aa	0 unaffected, 1 carrier, 1 affected
aa	aa	0 unaffected, 0 carrier, 2 affected

For n=2 offspring, what are the probabilities of various outcomes?

e.g. n=2, k_1=number of unaffected, k_2=number of carrier, k_3=number of affected

Offspring

1	2	Outcomes	# ways
p_1	p_1	$k_1=2, k_2=0, k_3=0$	1
p_1	p_2	$k_1=1, k_2=1, k_3=0$	2
p_2	p_1	$k_1=1, k_2=1, k_3=0$	
p_1	p_3	$k_1=1, k_2=0, k_3=1$	2
p_3	p_1	$k_1=1, k_2=0, k_3=1$	
p_2	p_2	$k_1=0, k_2=2, k_3=0$	1
p_3	p_2	$k_1=0, k_2=1, k_3=1$	2
p_2	p_3	$k_1=0, k_2=1, k_3=1$	
p_3	p_3	$k_1=0, k_2=0, k_3=2$	1

For each possible outcome, the probability $\Pr[Y_1=k_1, Y_2=k_2, Y_3=k_3]$ is $p_1^{k1} p_2^{k2} p_3^{k3}$

There are $\frac{n!}{k_1!k_2!k_3!}$ sequences for each probability.

Multinomial Probabilities

The probability that a multinomial random variable with n trials and success probabilities $p_1, p_2, \ldots, p_J$ will yield exactly $k_1, k_2, \ldots k_J$ successes is:

$$P(Y_1 = k_1, Y_2 = k_2, \ldots, Y_J = k_J) = \frac{n!}{k_1! k_2! \ldots k_J!} p_1^{k_1} p_2^{k_2} \cdots p_J^{k_J}$$

Assumptions:

1) J possible outcomes – only one can be a success, 1, in a given trial.

2) The probability of success for each possible outcome, p_j, is the same for each trial.

3) The outcome of one trial has no influence on other trials (independent trials).

4) Interest is in the (sum) total number of successes over all the trials.

Calculating a multinomial probability

Q$_1$: What is the probability that one of *n*=3 offspring will be unaffected (AA), one will be affected (aa) and one will be a carrier (Aa) (given recessive trait with carrier parents)?

Solution: For a given offspring, the probabilities of the three possible outcomes are:

p_1 = Pr[AA] = 1/4
p_2 = Pr[Aa] = 1/2
p_3 = Pr[aa] = 1/4

We have
$$\begin{aligned} \mathrm{P}(Y_1 = 1, Y_2 = 1, Y_3 = 1) &= \frac{3!}{1!1!1!} p_1^1 p_2^1 p_3^1 \\ &= \frac{(3)(2)(1)}{(1)(1)(1)} \left(\frac{1}{4}\right)^1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{4}\right)^1 \\ &= \frac{3}{16} = 0.1875. \end{aligned}$$


Paws- break time then work on exercises 1-2

Calculating the mean and variance

The marginal outcomes of the multinomial distribution are binomial.

We can obtain the means for each outcome, e.g, $Y_j = k_j$, the j^{th} outcome:

Mean:

$$E[k_j] = E\left[\sum_{i=1}^{n} Y_{ij}\right] = \sum_{i=1}^{n} E[Y_{ij}]$$

$$= \sum_{i=1}^{n} p_j = np_j$$

Variance:

$$V[k_j] = V\left[\sum_{i=1}^{n} Y_{ij}\right] = \sum_{i=1}^{n} V[Y_{ij}]$$

$$= \sum_{i=1}^{n} p_j(1-p_j) = np_j(1-p_j)$$

Multinomial Distribution Summary

1. Multinomial random variables are discrete
2. Parameters are $n, p_1, p_2, \ldots, p_J$
3. Each outcome $Y_j = k_j$ is the sum of n independent Bernoulli outcomes
4. Extends binomial distribution
5. Seen in contingency tables, polytomous regression

Continuous Distributions

Continuous Distributions

For measurements like height or weight, it does not make sense to talk about the probability of any single value.

Instead, we talk about the probability for an **interval**.

P[weight = 70.000kg] ≈ 0

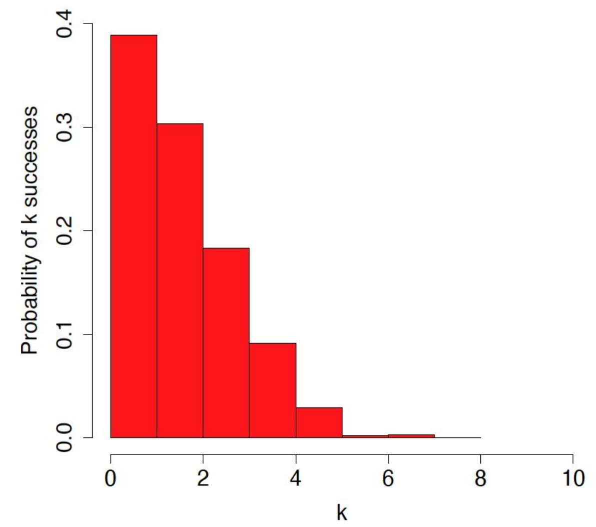
P[69.0kg ≤ weight ≤ 71.0kg] = 0.08

For discrete random variables, a probability mass function gives the probability of each possible value.

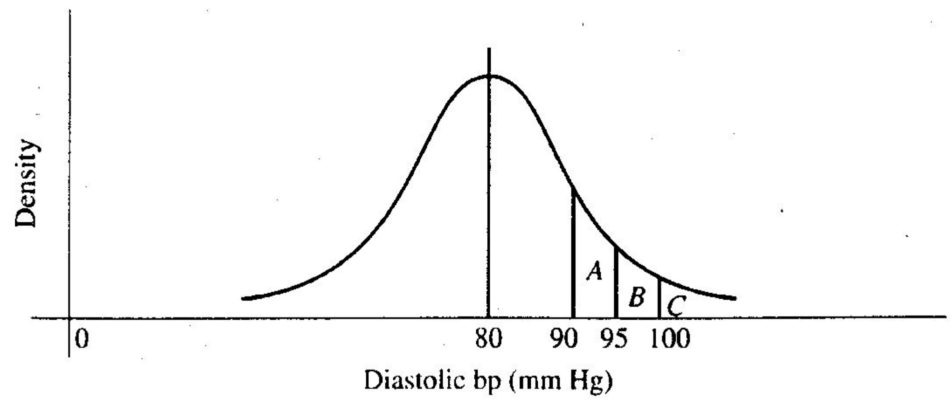
For continuous random variables, a **probability density function** to tell us about the probability of obtaining a value within an interval.

With discrete probability distributions, we can determine the probability of a single outcome, e.g.:

10 trials, 20% success probability



With continuous probability distributions, we determine the probability across a range of outcomes:



For any interval, the **area** under the curve represents the probability of obtaining a value in that interval.

Probability density function

1. A function, f(x), that gives probabilities based on the **area** under the curve.
2. $f(x) \geq 0$
3. Total area under the function f(x) is 1: $$\int f(x)dx = 1.0$$

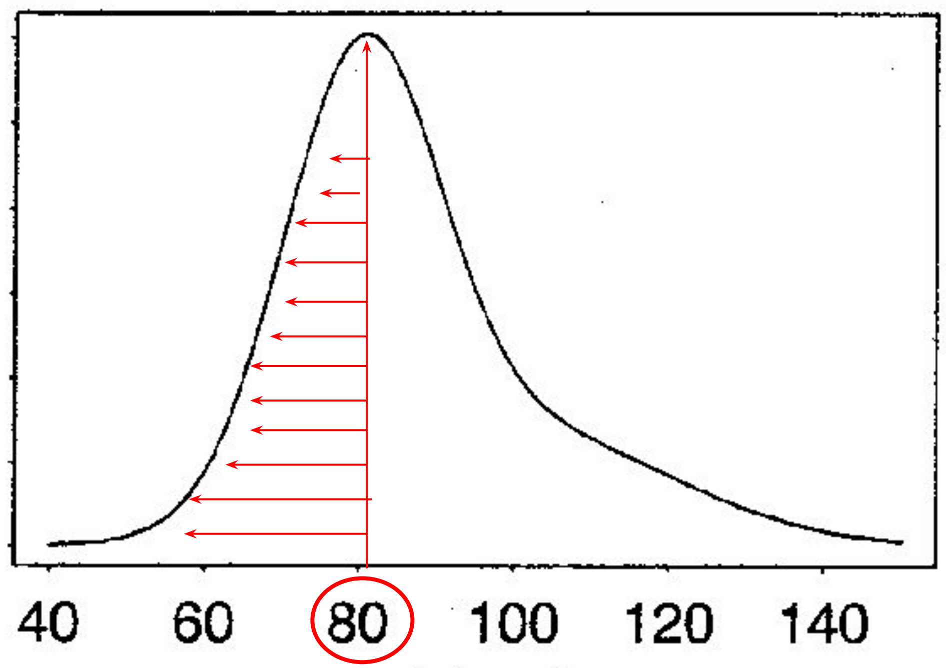
Cumulative distribution function

The **cumulative distribution function**, F(t), gives the total probability that X is less than some value t.

$F(t) = P(X \leq t)$

f(t)

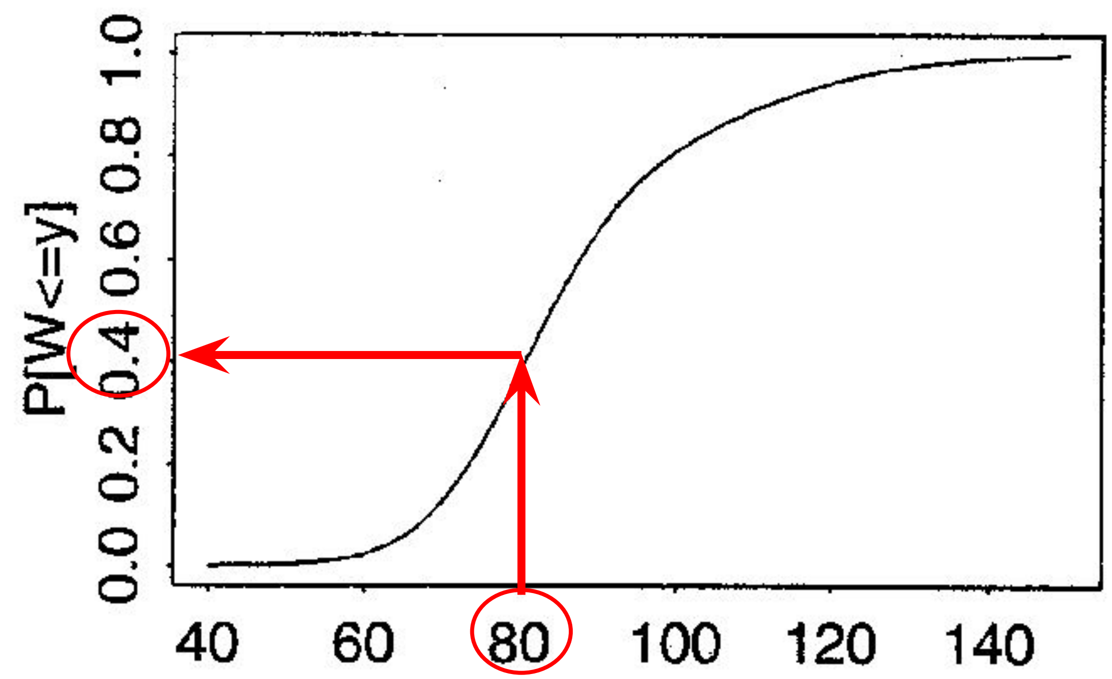
Weight in kg of males 30-40



Weight (kg)

F(t)

Cumulative Dist. Function



Weight (kg)

P(weight < 80) = 0.40
Area under the curve

P(weight < 80) = 0.40
y-value corresponding to x=80

Normal Distribution

- A well-known probability model for continuous data
- Characteristic bell-shaped curve
- Random variable values range from -∞ to + ∞
- Symmetric about mean: **mean = median = mode**

Common examples include birth weight, blood pressure, CD4 T cell counts (transformed)

The normal distribution is most useful as a derived distribution
(teaser for central limit theorem).

Normal Distribution

The mean and variance of a normal distribution completely determine the probability distribution function.

The **normal probability density function** is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right)$$

where

$\pi \approx 3.14$ (a constant)

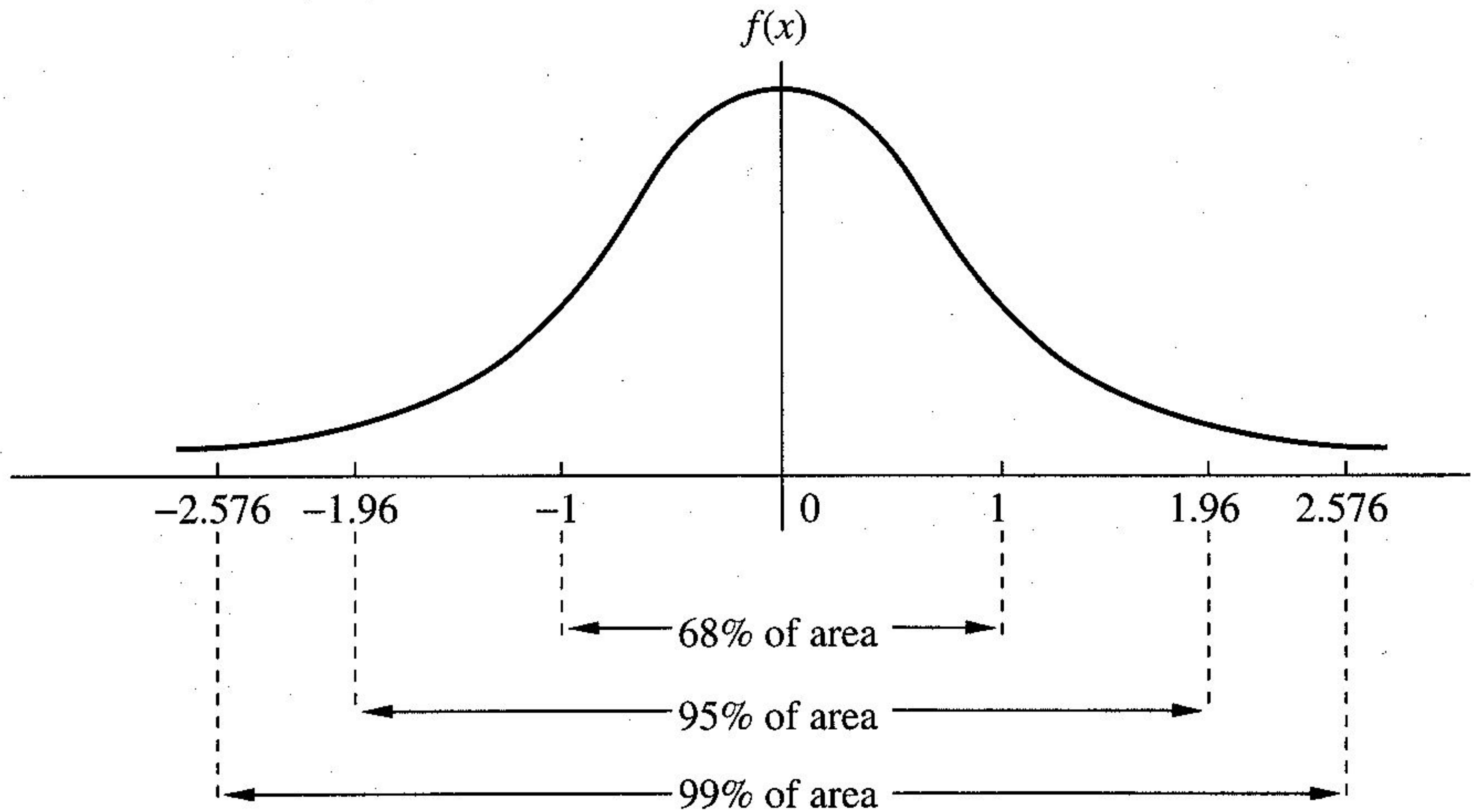
The normal distribution has two *parameters*:

μ = the mean of X

σ = the standard deviation of X

$X \sim N(\mu, \sigma^2)$: "X is normally distributed with mean μ and variance σ^2"

The standard normal distribution, N(0, 1), is a special case where $\mu = 0$ and $\sigma^2 = 1$.



Calculating Probabilities from a Standard Normal Dist'n

First, consider the **standard normal** N(0,1).

Z typically denotes a random variable with a standard normal distribution.

The probability density function of Z is:

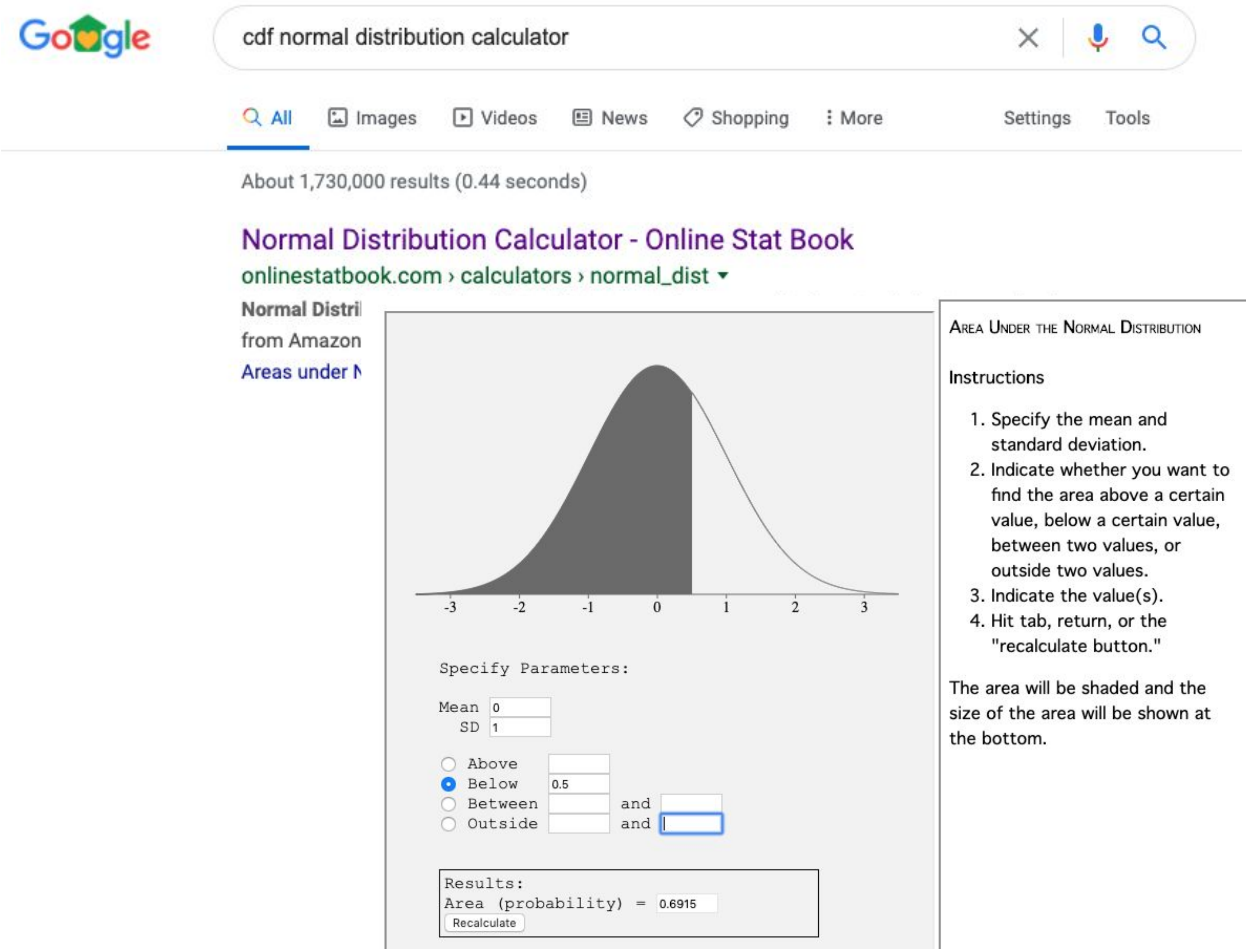
$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2 \right)$$

and the **cumulative distribution** of Z is:

$$P(Z \leq x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2 \right) dz$$

Any computing software will give the values of f(z) and Φ(x)

Online Calculators of Standard Normal Distribution Probabilities

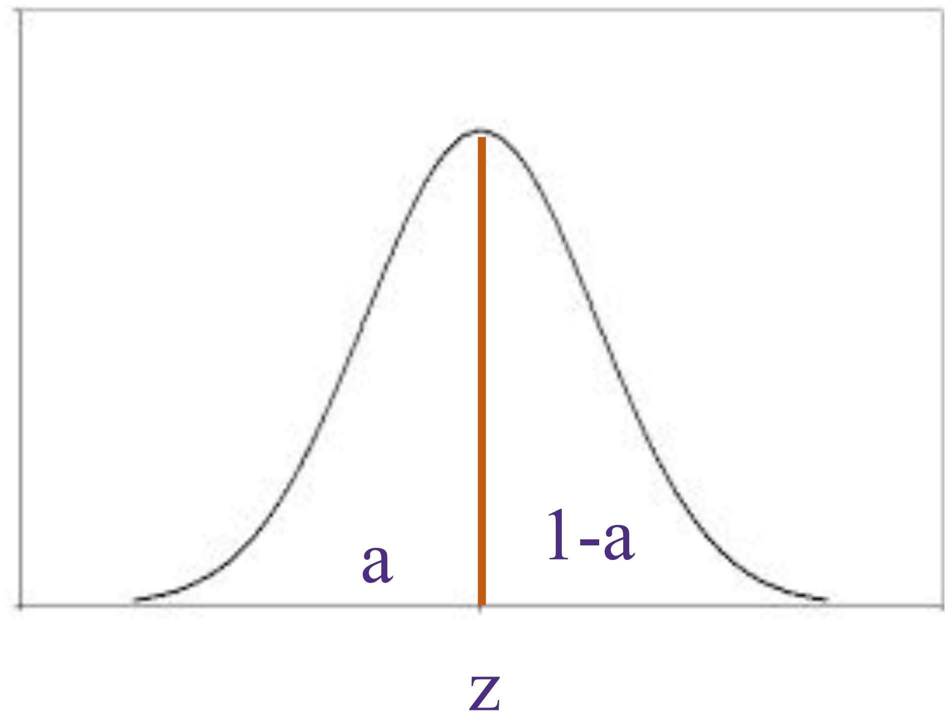


$Pr(Z \leq 0.5) = 0.6915$

Probability Distributions Calculations

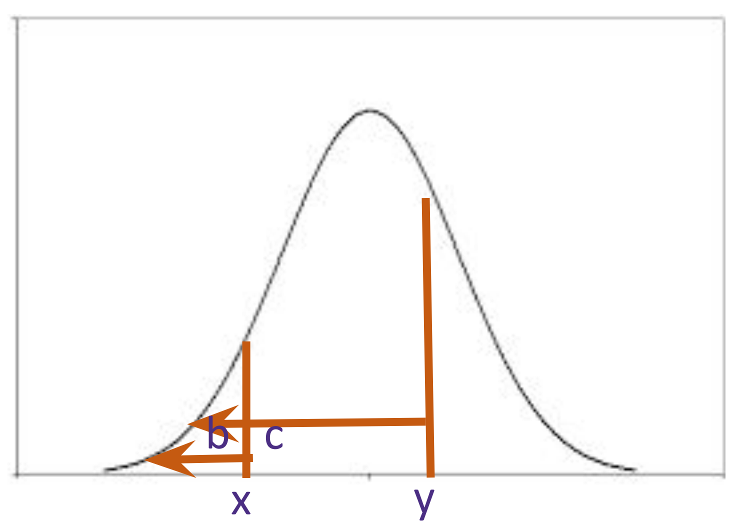
$P(Z \leq z) = a$

$P(Z > z) = 1 - a$



$P(Z \leq x) = b$, $P(Z \leq y) = c$

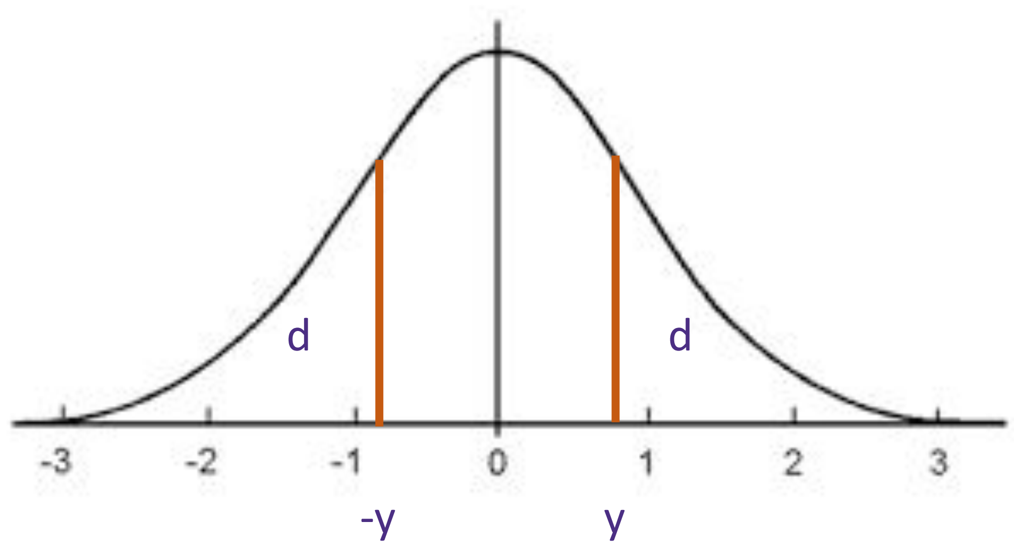
$Pr(x < Z \leq y) = c - b$



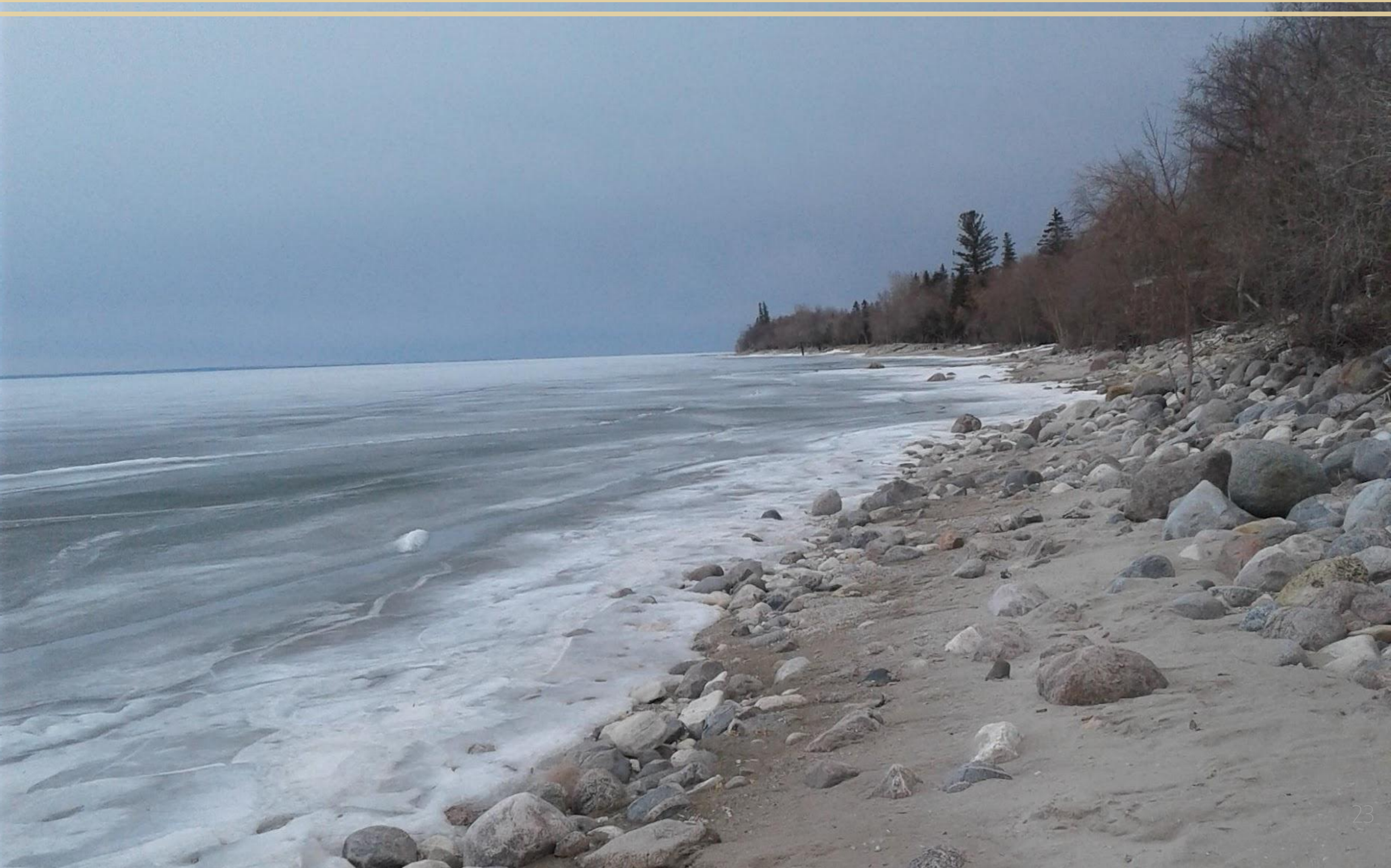
Standard Normal Distribution Calculations

Because the N(0,1) distribution is symmetric around 0,

$\Pr(Z \leq -y) = \Pr(Z \geq y) = d$



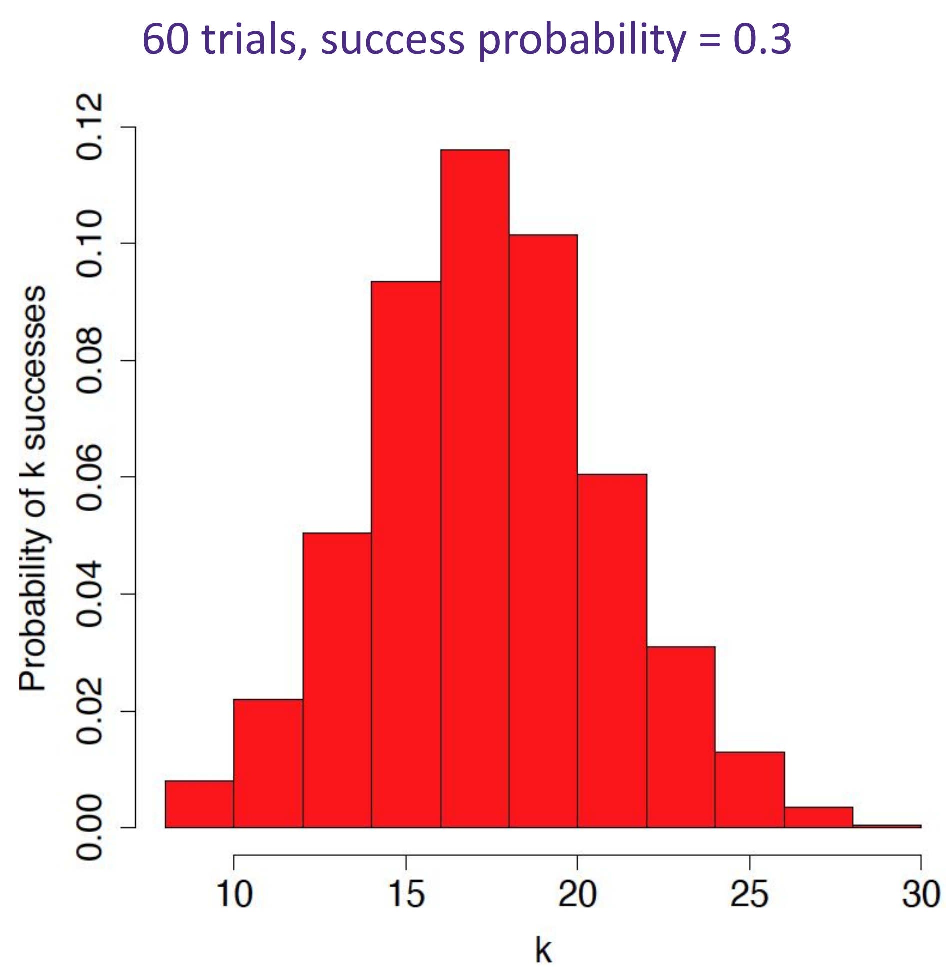
Pause- break time then work on exercises 3-5



Normal Approximation to Binomial Distribution

Example: Suppose the prevalence of HPV in 18-22 year old women is 30%.

What is the probability that 9 or fewer have HPV in a sample of 60 women from this population?



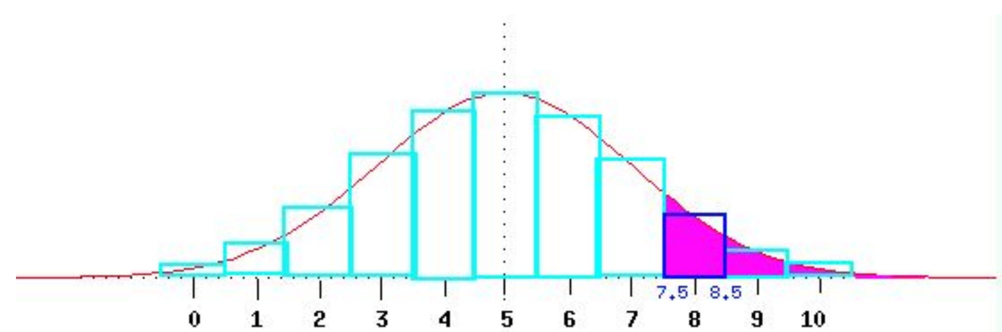
Normal Approximation to Binomial Distribution

Binomial

- When **np(1-p)** is "large" (e.g. ≥ 3), the normal distribution may be used to approximate the binomial distribution.
- X ~ bin(n,p)

  E(X) = np

  V(X) = np(1-p)
- X is approximately N(np, np(1-p))
- Apply continuity correction for discreteness:
  - P(X ≤ x) is a discrete binomial so to calculate it from a continuous normal, use P(X ≤ x + 0.5)



http://www.cs.uni.edu/~campbell/stat/prob9.html

Application of Normal Approximation to Binomial Distribution

Example:

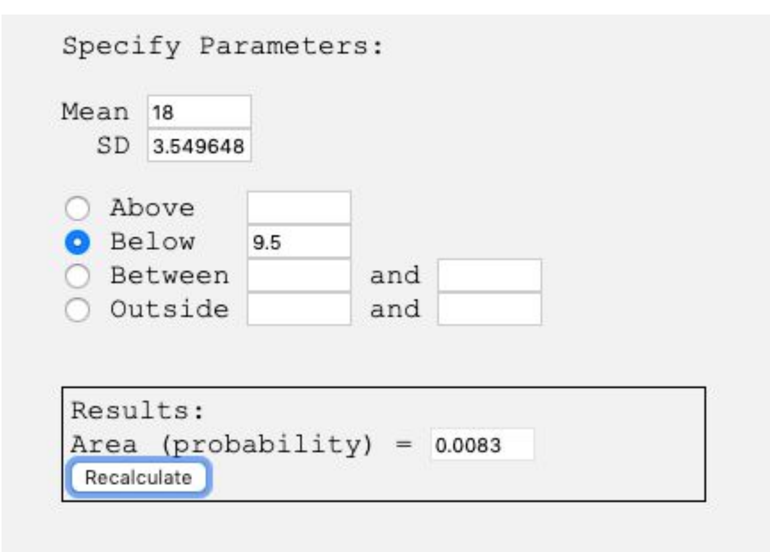
Suppose the prevalence of HPV in women 18 -22 years old is 30%. What is the probability that in a sample of 60 women from this population that 9 or less have HPV?

<u>Solution</u>

X = number infected out of 60

X ~ Binomial(n=60, p =0.3)

X close to normal distribution with mean 60*0.3=18 and variance 60*0.3*(1-0.3)=12.6



Therefore, $P(X \leq 9.5) = 0.0083$