# Quick recap

## Population mean and variance

How can we think of the population mean of a random variable?

The population mean is the

**long run average**

value of a random variable.

So if we played the lottery again, and again, and again, infinitely many times, the average amount we win in the long run is -$1.20.

In an experiment with binomially distributed outcomes and parameters n and p, the long-run average number of successes is n*p.

## Population mean and variance

How can we think of the population variance of a random variable?

The population variance is the

**long run variance**

value of a random variable.

## Adding means and variances

Suppose that the graduate student’s experimental runs are independent, and the probability of success on the first experiment was 0.75, the probability of success on the second experiment was 0.75, and the probability of success on the third experiment was 0.75.

What's the long run mean number of successes?

## Adding means and variances

Suppose that the graduate student's experimental runs are independent, and the probability of success on the first experiment was 0.75, the probability of success on the second experiment was 0.75, and the probability of success on the third experiment was 0.75.

What's the long run mean number of successes?

Mean # successes over 3 experiments

= mean number of successes in 1$^{st}$ experiment +

mean number of successes in 2$^{nd}$ experiment +

mean number of successes in 3$^{st}$ experiment

= 0.75 + 0.75 + 0.75

= 2.25

## Adding means and variances

Suppose that the graduate student's experimental runs are independent, and the probability of success on the first experiment was 0.75, the probability of success on the second experiment was 0.75, and the probability of success on the third experiment was 0.75.

What's the long run mean number of successes?

Mean # successes over 3 experiments

= mean number of successes in 1$^{st}$ experiment +

mean number of successes in 2$^{nd}$ experiment +

mean number of successes in 3$^{st}$ experiment

= 0.75 + 0.75 + 0.75

= 2.25

**Assumes that experiments are independent!**

## Quiz: Adding means and variances

Suppose that the graduate student's experimental runs were still independent, but the probability of success on the first experiment was 0.2, the probability of success on the second experiment was 0.5, and the probability of success on the third experiment was 0.9.

What is the long-run mean # successes?

## Quiz: Adding means and variances

Suppose that the graduate student’s experimental runs were still independent, but the probability of success on the first experiment was 0.2, the probability of success on the second experiment was 0.5, and the probability of success on the third experiment was 0.9.

Long-run mean # successes

$= 0.2 + 0.5 + 0.9$

$= 1.6$

**This means that if we took the average over lots and lots of graduate students with these success probabilities, the average number of successes over all graduate students would be 1.6**

# Probability Distributions

# II

## Multinomial Distribution

The Binomial distribution was a great model for cases where we have independent trials and 2 possibilities: success and failure

What about when we have 3 different categories?

4 different categories?

20 different categories?

The multinomial distribution is a model for categorical data where we have multiple categories, independent runs and the same probabilities of observing each category.

## Multinomial Distribution - Motivation

Consider a heterozygous allele for a recessive trait. We have 3 categories

- Unaffected (AA)
- Affected (aa)
- Carrier (Aa)

If both parents are carriers, what's the probability that of their 4 offspring

1. One will be unaffected, two will be affected and one will be a carrier?
2. All of their offspring will be carriers?
3. Two of their offspring will be affected and two will be carriers?

## Multinomial Distribution - Motivation

Suppose we have J categories

Let $Y_i$ be the number of times we observe category i

For example, let

- i = 1 refer to unaffected (AA)
- i = 2 refer to carrier (Aa)
- i = 3 refer to affected (aa)

Then

**"0 unaffected, 3 carriers, 1 affected,"**

corresponds to

$$\mathbf{Y_1 = 0, Y_2 = 3, Y_3 = 1}$$

# Multinomial Probabilities

What is 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?

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

## Multinomial Probabilities

**Assumptions**:

1) J possible outcomes in a given trial.

2) The probability of each category $p_j$, is the same from trial to 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.

| | | | | | | |
|---|---|---|---|---|---|---|
| $k_1$ | $k_2$ | $k_3$ | $k_4$ | $\cdots$ | $k_{J-1}$ | $k_J$ |

$n = \Sigma_j k_j$ is the total number of trials.

## Multinomial Probabilities - Examples

Returning to the original questions:

**$Q_1$**: Of $n$=4 offspring, one will be unaffected, two will be affected and one will be a carrier?

**Solution:** For a given child, the probabilities of the three outcomes are:

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

We have

$$
\begin{aligned}
&Pr(Y_1 = 1, Y_2 = 2, Y_3 = 1) \\
&\quad = \frac{4!}{1!2!1!} p_1^1 p_2^2 p_3^2 \\
&\quad = \frac{4 \times 3 \times 2 \times 1}{2 \times 1} \left(\frac{1}{4}\right)^1 \left(\frac{1}{2}\right)^2 \left(\frac{1}{4}\right)^1 \\
&\quad = 0.1875
\end{aligned}
$$

## Quiz

**Suppose the couple decided they wanted 3 children**

1. What is the probability that all three offspring will be carriers?

2. What is the probability that exactly two offspring will be affected and one a carrier?

## Solutions

**Q₂**: What is the probability that all three offspring will be carriers?

$$P(Y_1 = 0, Y_2 = 3, Y_3 = 0) = \frac{3!}{0!3!0!} p_1^0 p_2^3 p_3^0$$
$$= \frac{(3)(2)(1)}{(3)(2)(1)} \left(\frac{1}{4}\right)^0 \left(\frac{1}{2}\right)^3 \left(\frac{1}{4}\right)^0$$
$$= \frac{1}{8} = 0.125.$$

**Q₃**: What is the probability that exactly two offspring will be affected and one a carrier?

$$P(Y_1 = 0, Y_2 = 1, Y_3 = 2) = \frac{3!}{0!1!2!} p_1^0 p_2^1 p_3^2$$
$$= \frac{(3)(2)(1)}{(2)(1)} \left(\frac{1}{4}\right)^0 \left(\frac{1}{2}\right)^1 \left(\frac{1}{4}\right)^2$$
$$= \frac{3}{32} = 0.09375.$$

## Example - Mean and Variance

It turns out that the (marginal) outcomes of the multinomial distribution are binomial. We can immediately obtain the means for each outcome (i.e., the j^th cell)

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:

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

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

COVARIANCE:

$$Cov[k_j, k_{j'}] = -np_j p_{j'}$$

## Multinomial Distribution Summary

**Multinomial**

1. Discrete, bounded
2. Parameters - $n$, $p_1$, $p_2$, ..., $p_J$
3. Sum of $n$ independent outcomes
4. Extends binomial distribution
5. Polytomous regression, contingency tables

# Continuous Distributions

## Continuous Distributions

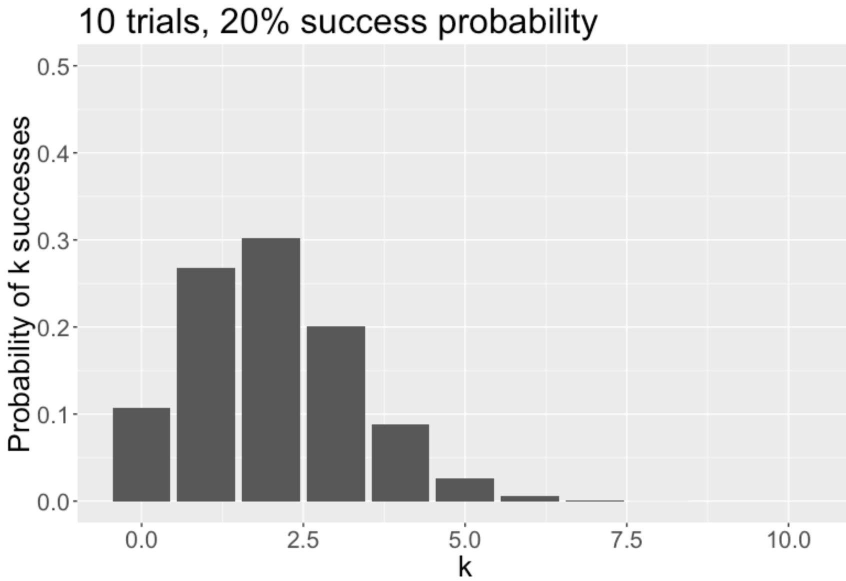
For measurements like height and weight which can be measured with arbitrary precision, it does not make sense to talk about the probability of any single value. Instead we talk about the probability for an **interval**.

$$P[\text{weight} = 70.000\text{kg}] \approx 0$$

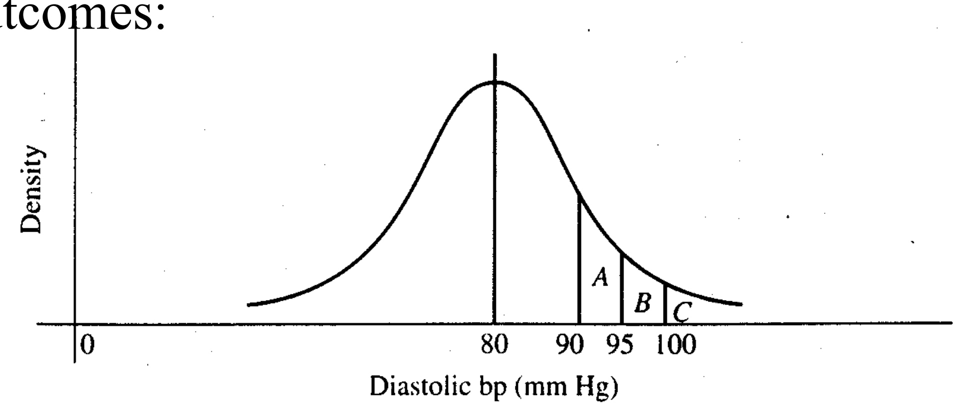
$$P[69.0\text{kg} \leq \text{weight} \leq 71.0\text{kg}] = 0.08$$

For discrete random variables we had a probability mass function to give us the probability of each possible value. For continuous random variables we use a **probability density function** to tell us about the probability of obtaining a value within some interval.

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




With continuous probability distributions, we can 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, typically denoted 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 1.0.

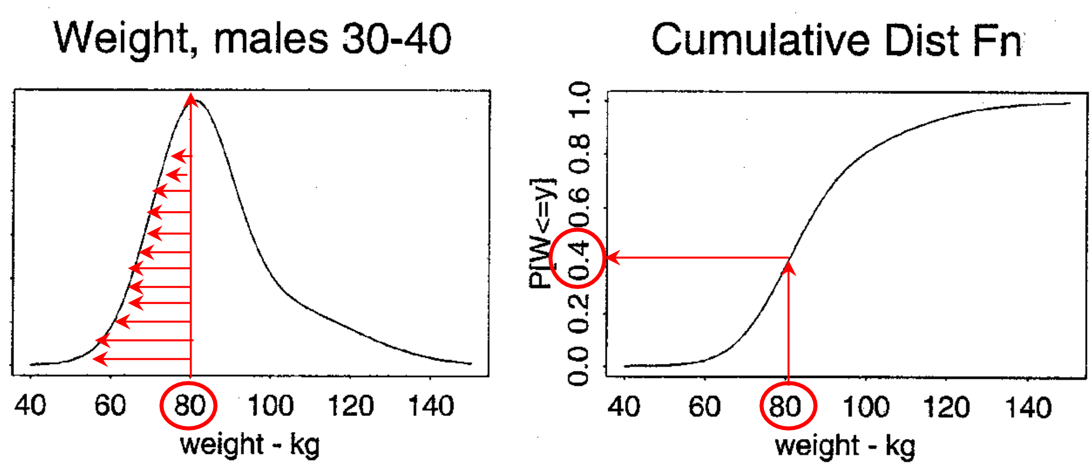
$$\int f(x)dx = 1.0$$

## Cumulative distribution function

The cumulative distribution function, F(t), tells us the total probability less than some value t.

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

This is analogous to the cumulative relative frequency.




Prob[wgt < 80] = 0.40

*Area under the curve*

## Normal Distribution

- A ~~common~~ probability model for continuous data
- Bell-shaped curve

$\Rightarrow$ takes values between $-\infty$ and $+\infty$

$\Rightarrow$ symmetric about mean

$\Rightarrow$ mean=median=mode

- Examples (common but questionable!)

  birthweights

  blood pressure

  CD4 cell counts (perhaps transformed)

**The normal distribution is more useful as a derived distribution, as we will see when we talk about the central limit theorem…**

## Normal Distribution

Specifying the mean and variance of a normal distribution completely determines the probability distribution function and, therefore, all probabilities.

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

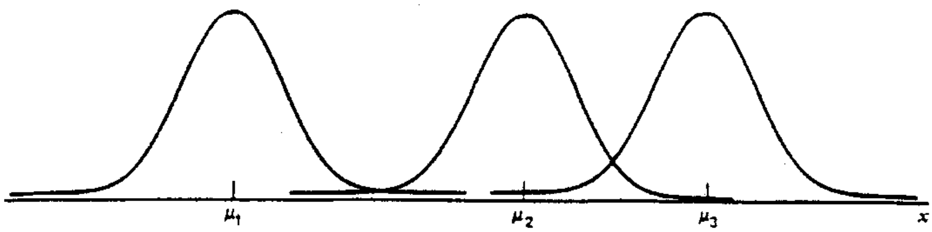
$\pi \approx 3.14$ (a constant)

Notice that the normal distribution has two parameters:

$\mu$ = the mean of X

$\sigma$ = the standard deviation of X

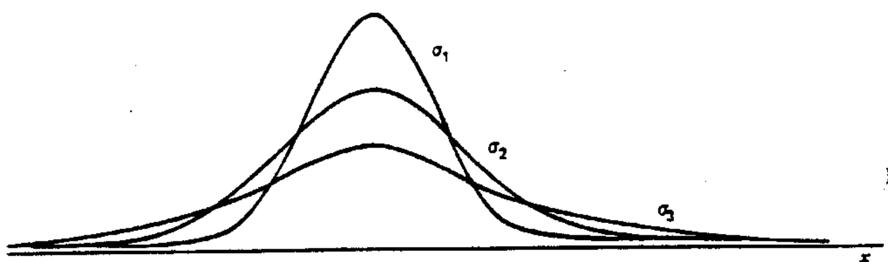
We write $X \sim N(\mu, \sigma^2)$. The **standard normal** distribution is a special case where $\mu = 0$ and $\sigma = 1$.




$\mu_1 < \mu_2 < \mu_3$

**FIGURE 3.6.3**

**Three Normal Distributions with Different Means**




$\sigma_1 < \sigma_2 < \sigma_3$

**FIGURE 3.6.4**

**Three Normal Distributions with Different Standard Deviations**

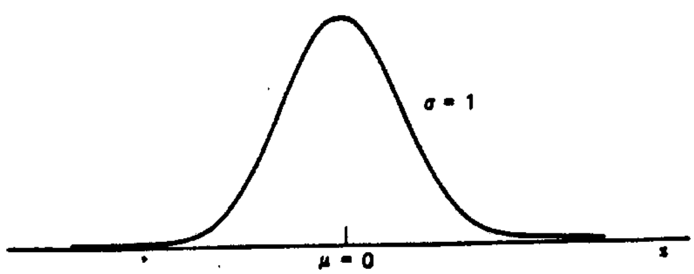
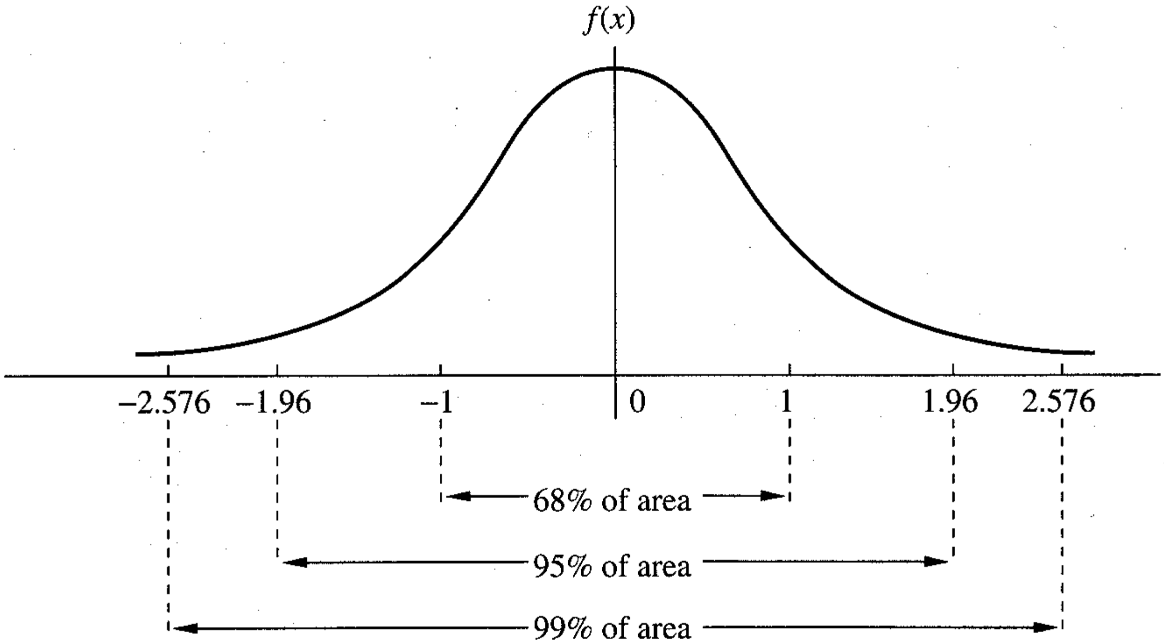



**FIGURE 3.6.5**

**The Unit Normal Distribution**


f(x)
−2.576
−1.96
−1
0
1
1.96
2.576
68% of area
95% of area
99% of area

## Normal Distribution - Calculating Probabilities

Example: Rosner 5.20

Serum cholesterol is approximately normally distributed with mean 219 mg/mL and standard deviation 50 mg/mL. If the clinically desirable range is < 200 mg/mL, then what proportion of the population falls in this range?

X = serum cholesterol in an individual.

$\mu =$

$\sigma =$

$$P[x < 200] = \int_{-\infty}^{200} \frac{1}{50\sqrt{2\pi}} \exp\left( -\frac{1}{2} \frac{(x-219)^2}{50^2} \right) dx$$

negative values for cholesterol - huh?

## Standard Normal Distribution - Calculating Probabilities

First, let's consider the **standard normal** - N(0,1). We will usually use Z to denote a random variable with a standard normal distribution. The density 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:

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

Any computing software can give you the values of f(z) and Φ(z)

# Standard Normal Distribution - Calculating Probabilities

Google

cdf normal distribution calculator

All Videos Images News Shopping More Settings Tools

About 69,900 results (0.84 seconds)

Free Cumulative Distribution Function (CDF) Calculator for the Normal ...
www.danielsoper.com/statcalc/calculator.aspx?id=53
This **calculator** will compute the **cumulative distribution function** (**CDF**) for the **normal distribution** (i.e., the area under the **normal distribution** from negative infinity ...

## Cumulative Distribution Function (CDF) Calculator for the Normal Distribution

This calculator will compute the cumulative distribution function (CDF) for the normal distribution (i.e., the area under the normal distribution from negative infinity to x), given the upper limit of integration x, the mean, and the standard deviation.

Please enter the necessary parameter values, and then click 'Calculate'.

**Mean:** 0

**Standard deviation:** 1

**x:** 0.5

**Calculate!**

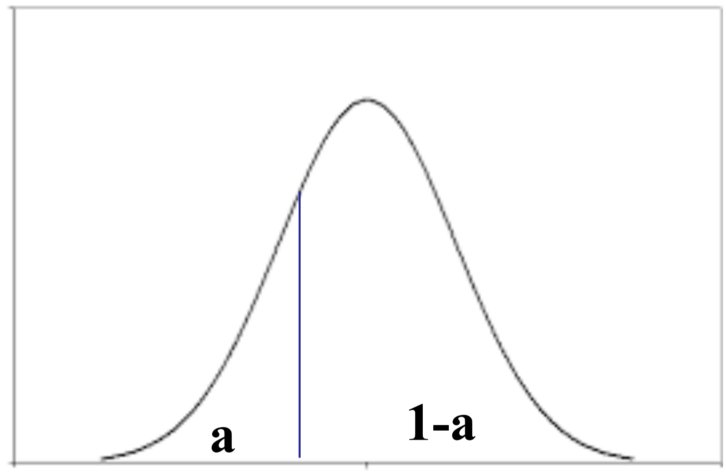
**Cumulative distribution function: 0.69146246**

$$\boxed{Pr(Z \leq 0.5) = 0.69146}$$

## Facts about probability distributions

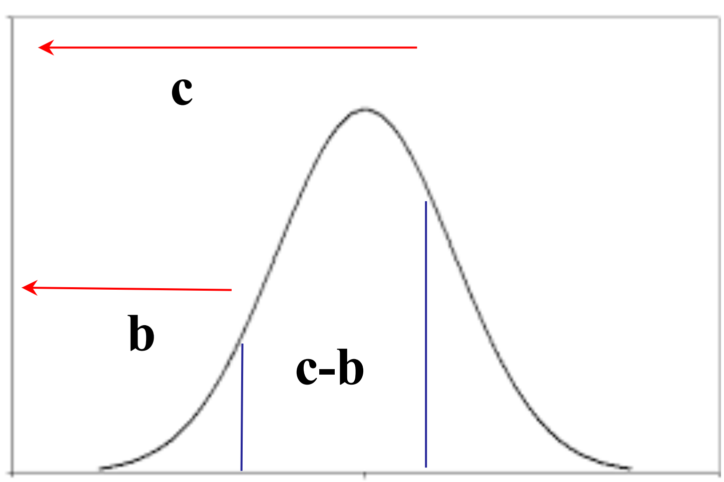
$P[Z \leq z] = a$

$\Rightarrow P[Z > z] = 1\text{-}a$




$P[Z \leq x] = b, P[Z \leq y] = c$

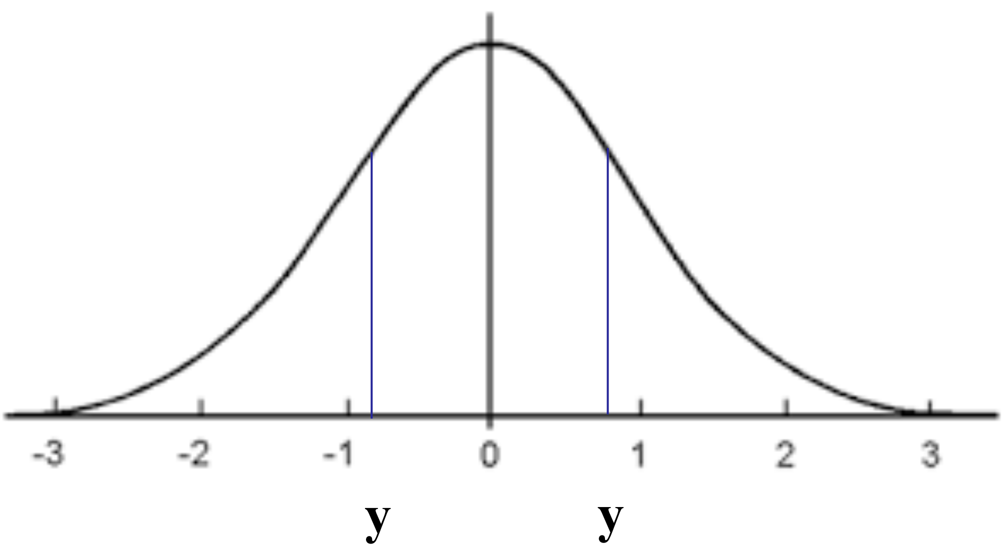
$\Rightarrow Pr[x < Z \leq y] = c\text{-}b$



## Facts about the standard normal distribution

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

$$\Pr[Z \leq -y] = \Pr[Z \geq y]$$



## Quiz: 3 minutes

Google “cdf normal distribution calculator” and find the following if $Z \sim N(0,1)$

$P[Z \leq 1.65] =$

$P[Z \geq 0.5] =$

$P[-1.96 \leq Z \leq 1.96] =$

$P[-0.5 \leq Z \leq 2.0] =$

## Solutions to Quiz

$P[Z \leq 1.65] = 0.9505$

$P[Z \geq 0.5] = 1 - 0.6915 = 0.3085$

$P[-1.96 \leq Z \leq 1.96] = 0.975 - 0.025 = 0.95$

$P[-0.5 \leq Z \leq 2.0] = 0.9772 - 0.3085 = 0.6687$

## Converting to Standard Normal

This solves the problem for the N(0,1) case.
Do we need a special table for every $(\mu,\sigma)$?

No!

Define: $X = \mu + \sigma Z$ where $Z \sim N(0,1)$

1. $E(X) = \mu + \sigma E(Z) = \mu$
2. $V(X) = \sigma^2 V(Z) = \sigma^2$.
3. <u>X is normally distributed!</u>

**Linear functions of normal RV's are also normal.**

If $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$
then

$$Y \sim N(a\mu + b, a^2\sigma^2)$$

## Converting to Standard Normal

How can we convert a $N(\mu,\sigma^2)$ to a standard normal?

**Standardize:**

$$Z = \frac{X - \mu}{\sigma}$$

What is the mean and variance of Z?

1. $E(Z) = (1/ \sigma)E(X - \mu) = 0$
2. $V(Z) = (1/ \sigma^2)V(X) = 1$

## Normal Distribution - Calculating Probabilities

Return to cholesterol example (Rosner 5.20)

Serum cholesterol is approximately normally distributed with mean 219 mg/mL and standard deviation 50 mg/mL. If the clinically desirable range is < 200 mg/mL, then what proportion of the population falls in this range?

$$
\begin{aligned}
P(X < 200) &= P\left(\frac{X-\mu}{\sigma} < \frac{200-219}{50}\right) \\
&= P\left(Z < \frac{200-219}{50}\right) \\
&= P(Z < -0.38) \\
&= P(Z > 0.38) \\
&= 0.3520.
\end{aligned}
$$

## Normal Approximation to Binomial

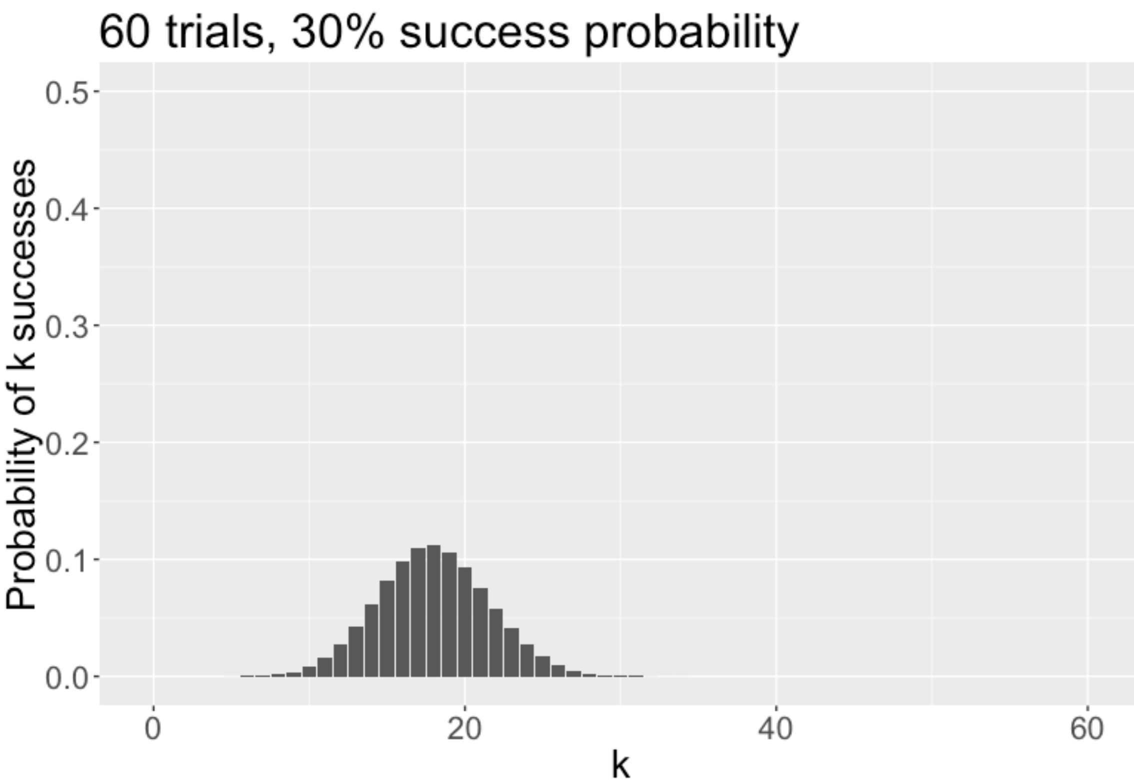
<u>Example</u>

Suppose the prevalence of HPV in women 18 - 22 years old is 0.30. What is the probability that in a sample of 60 women from this population 9 or fewer would be infected?

## Normal Approximation to Binomial

Example

Suppose the prevalence of HPV in women 18 - 22 years old is 0.30. What is the probability that in a sample of 60 women from this population 9 or fewer would be infected?



# Normal Approximation to Binomial

**Binomial**

- When **np(1-p)** is “large enough” ($\geq 10$) the normal may be used to approximate the binomial.
- $X \sim Bin(n,p)$

  $E(X) = np$

  $V(X) = np(1-p)$
- X is approximately $N(np, np(1-p))$

# Normal Approximation to Binomial

<u>Example</u>

Suppose the prevalence of HPV in women 18 - 22 years old is 0.30. What is the probability that in a sample of 60 women from this population that 9 or less would be infected?

<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

### Cumulative Distribution Function (CDF) Calculator for the Normal Distribution

This calculator will compute the cumulative distribution function (CDF) for the normal distribution (i.e., the area under the normal distribution from negative infinity to x), given the upper limit of integration x, the mean, and the standard deviation.

Please enter the necessary parameter values, and then click 'Calculate'.

**Mean:** 18

**Standard deviation:** 3.54964

**x:** 9.5

**Calculate!**

**Cumulative distribution function:** 0.00831905