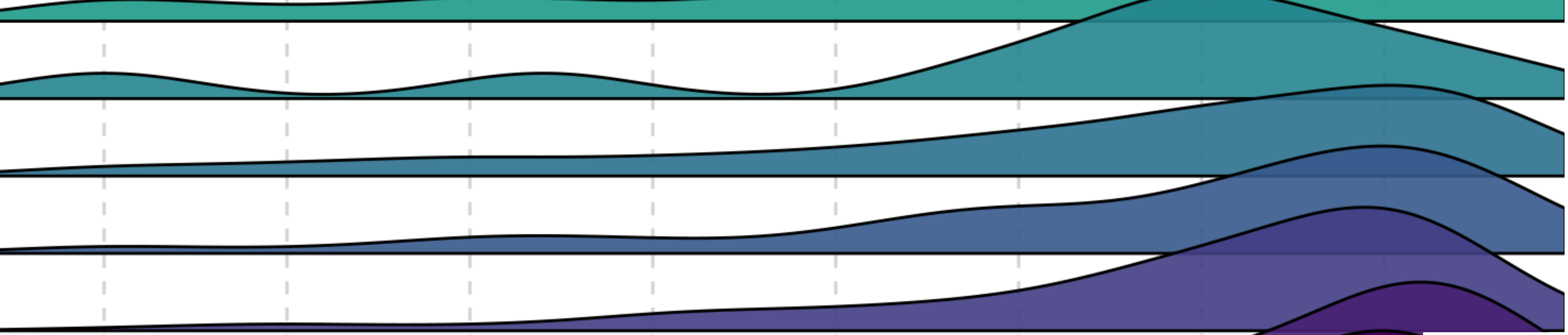


ST 117

1. Introduction

WARWICK

Lectures 8 & 9

(Week 3)

Normal distribution with data examples

Binomial approximation

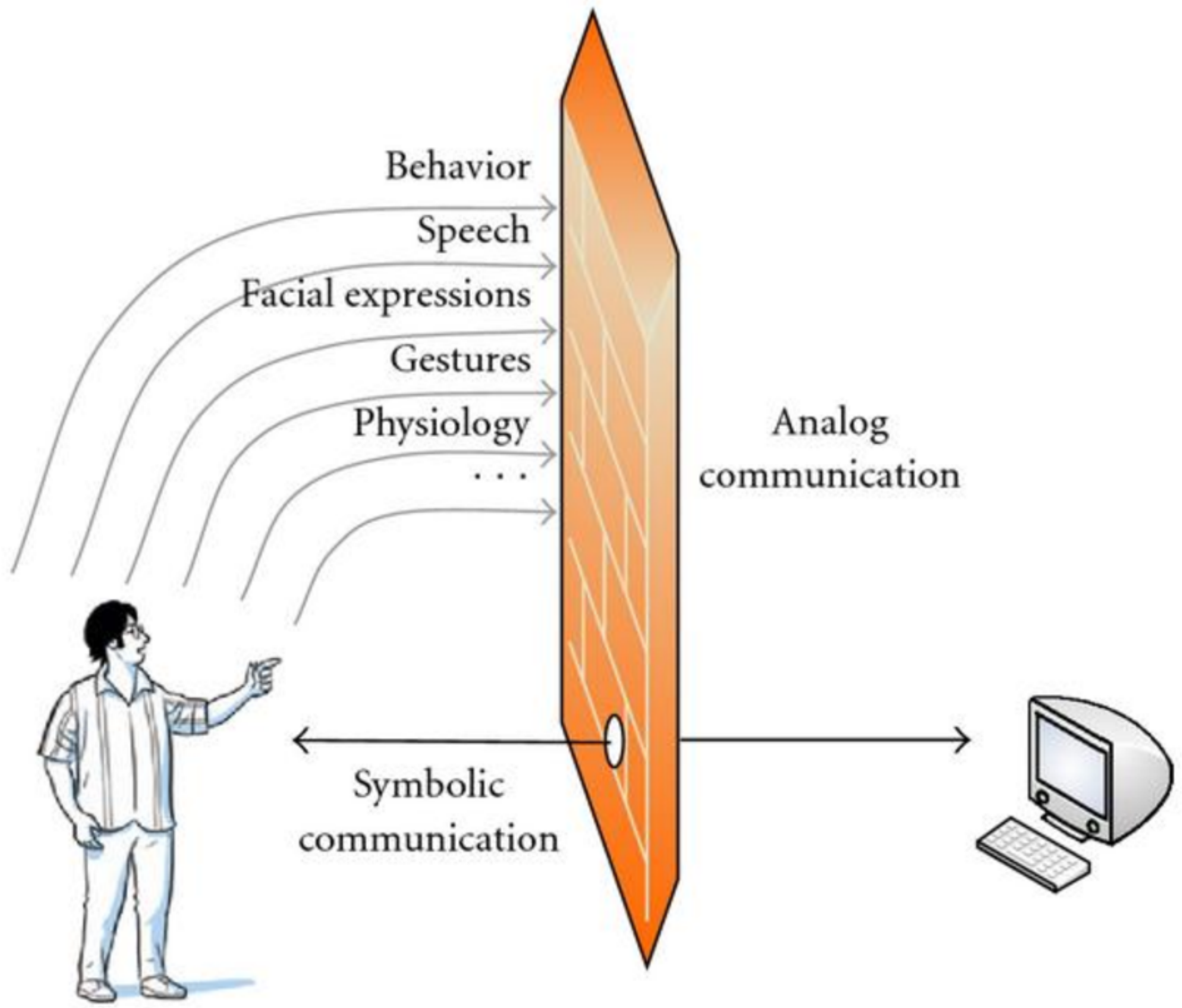
Joint distributions

Barriers to learning R

- Computers “too stupid” to understand what humans tell them
- Technical slang
- More experienced people overusing the slang
- Too much unstructured non-quality ranked information online
- It take some time and patience, in particular that beginning
- For more experienced people it may by boring/too slow

Enablers to learning R

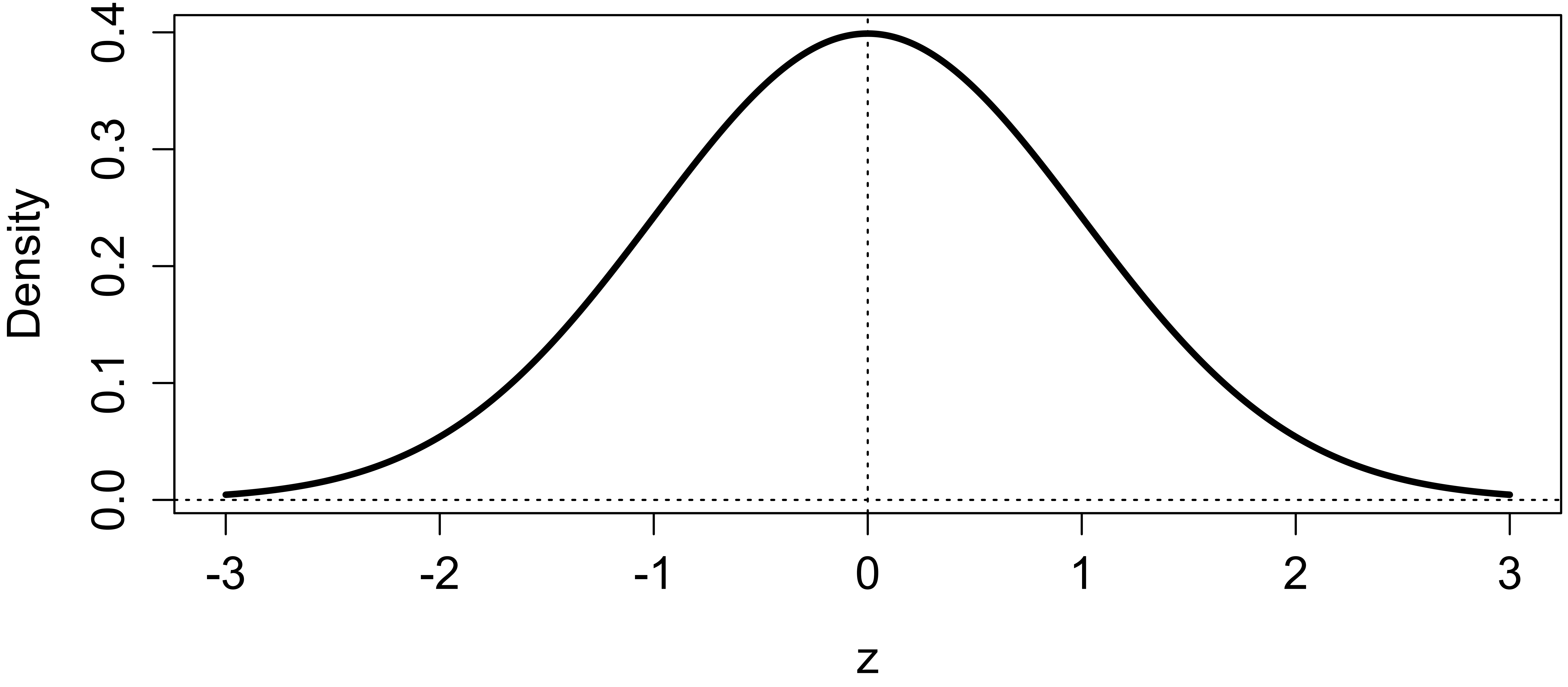
- Motivation by the exciting data analysis or simulation projects
- Online resources (cheat sheets, tutorials, videos…)
- Help files as part of base R
- Vignettes to come with R packages
- Convenient environment (RStudio)


Behavior
Speech
Facial expressions
Gestures
Physiology
. . .
Analog communication
Symbolic communication

The normal distribution

The standard normal distribution

$$\frac{1}{\sqrt{2\pi}}e^{-z^2/2}$$



The general normal distribution

A two-parameter family of distributions.

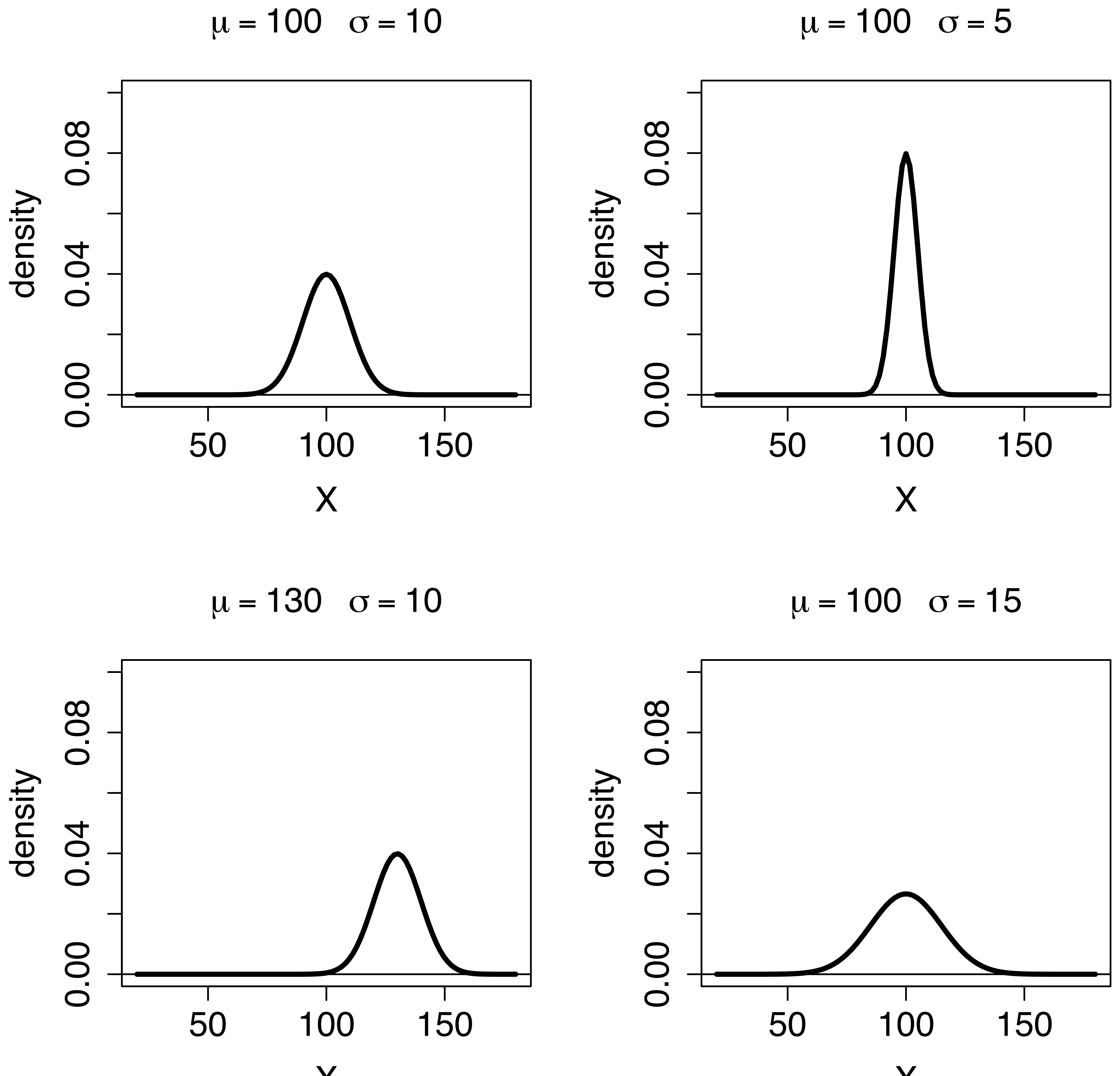
Parameters

mean μ
SD σ

Density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

The general normal distribution



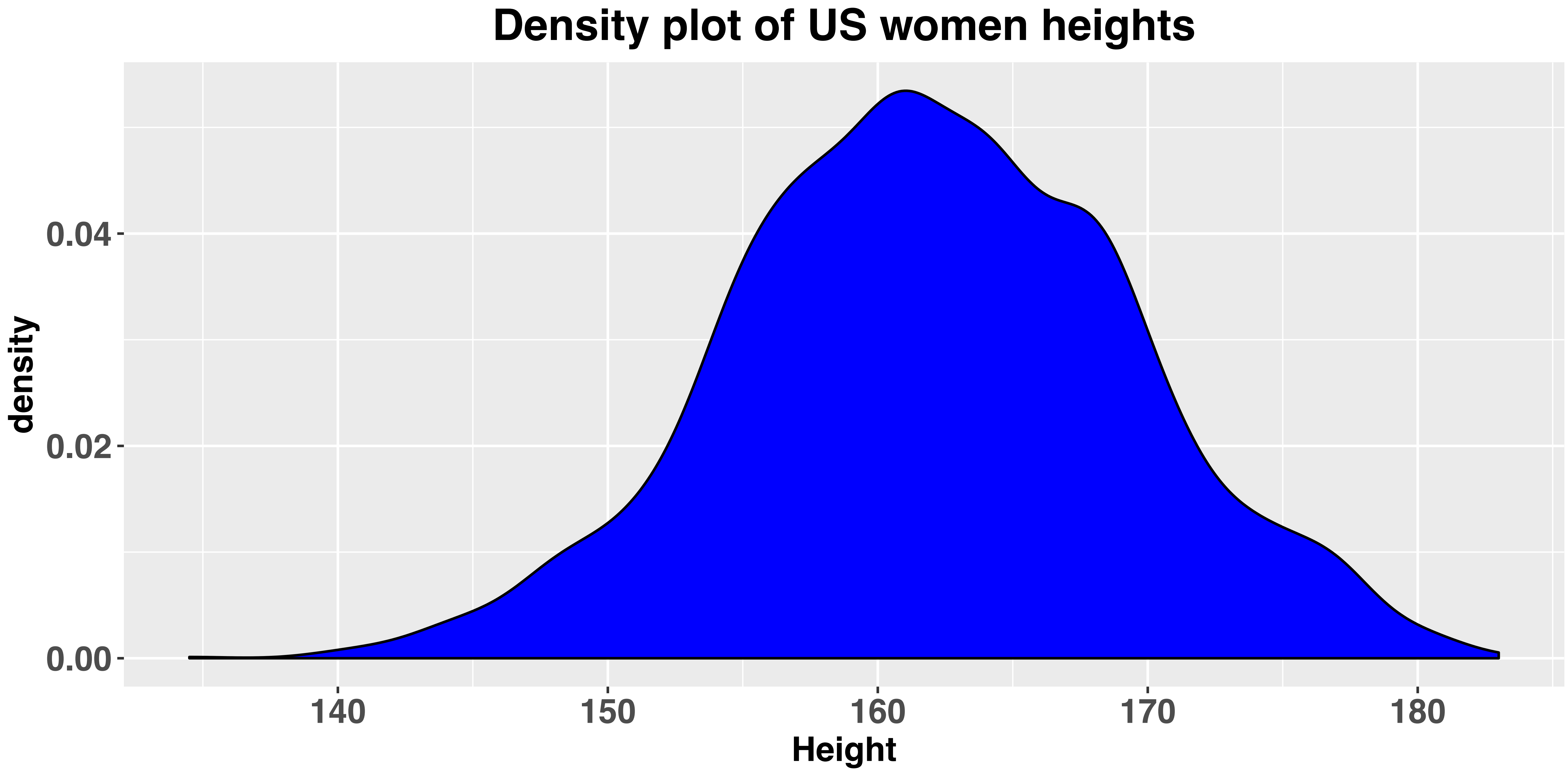
Intuitive facts about normal (Gaussian) distributions

- Symmetric and unimodal: mode=mean=median.
- Sums and differences of independent normal random variables are also normal.
- Nearly all the probability is within 3 SDs of the mean. 95% is within 2 SDs.
- Normal distributions come up A LOT. Heights and weights tend to be normal, measurement errors, blood pressures.
- ... but only approximately…
- … and not all data are normal.

Adult heights

NHANES (US National Health and Nutrition Examination Survey)

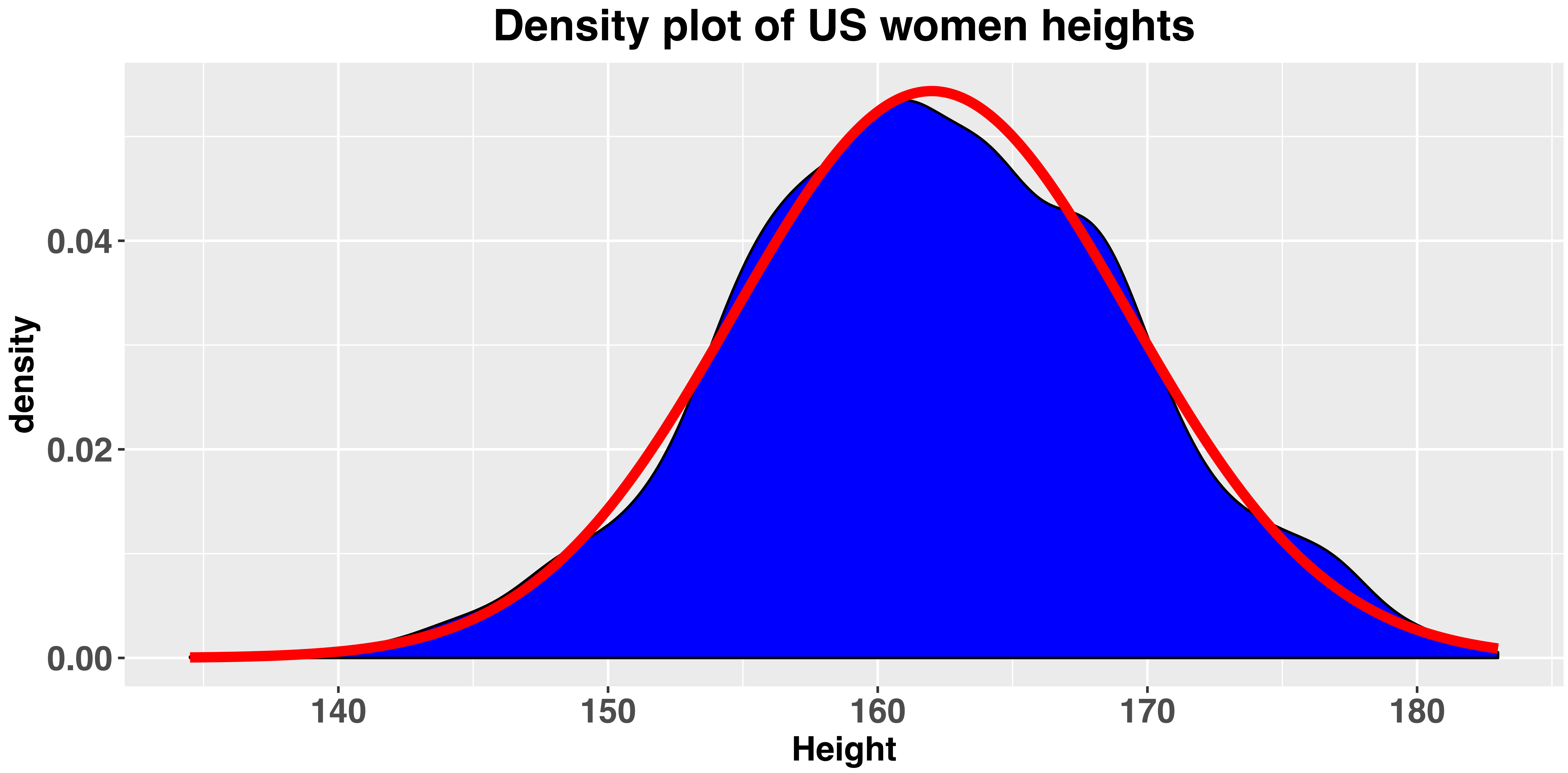
- NHANES package in R includes data on 10,000 survey participants from 2009—2012.
- Weighted to be like a simple random sample from the US population



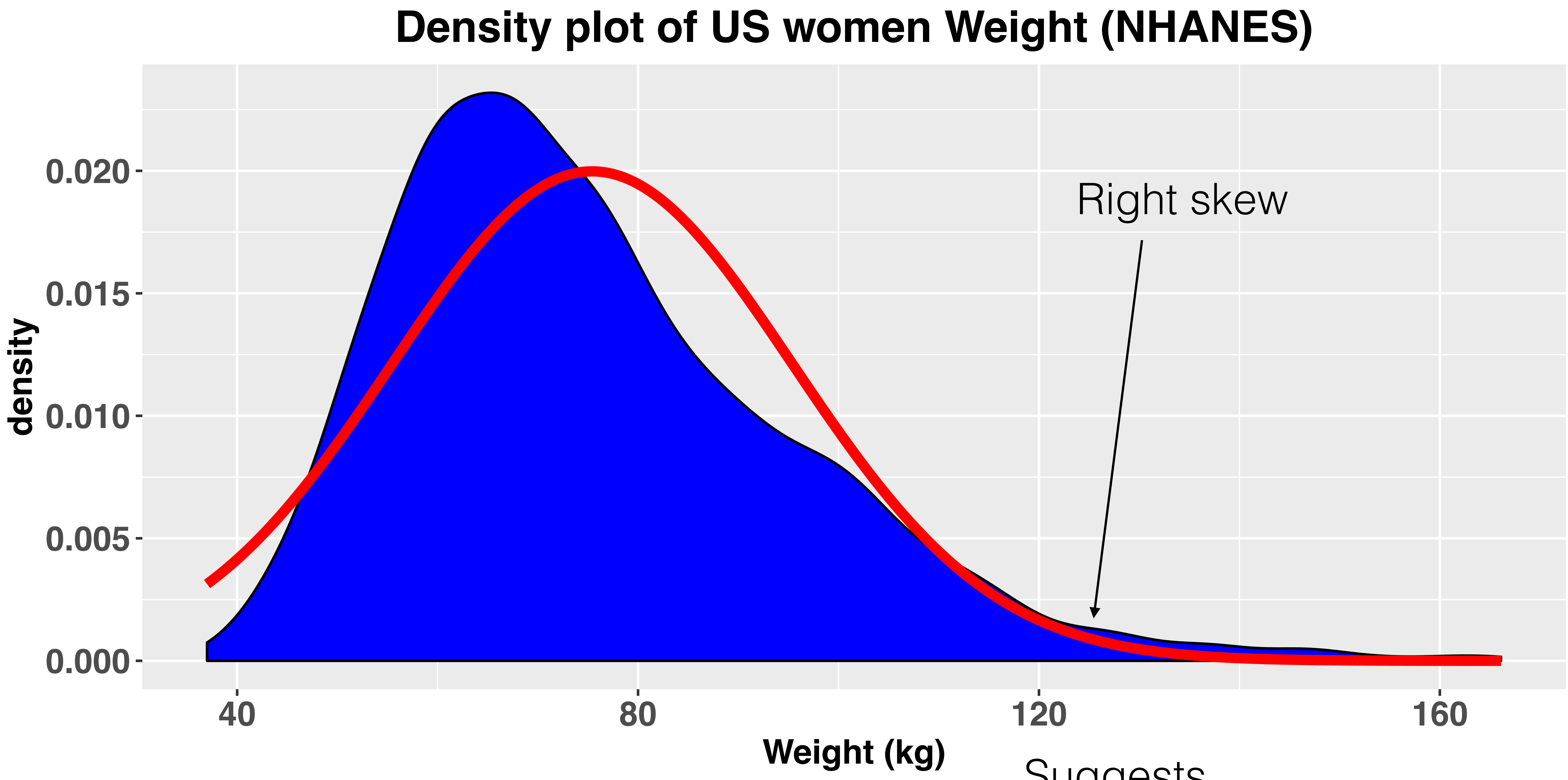
Adult heights

NHANES (US National Health and Nutrition Examination Survey)

- NHANES package in R includes data on 10,000 survey participants from 2009—2012.
- Weighted to be like a simple random sample from the US population

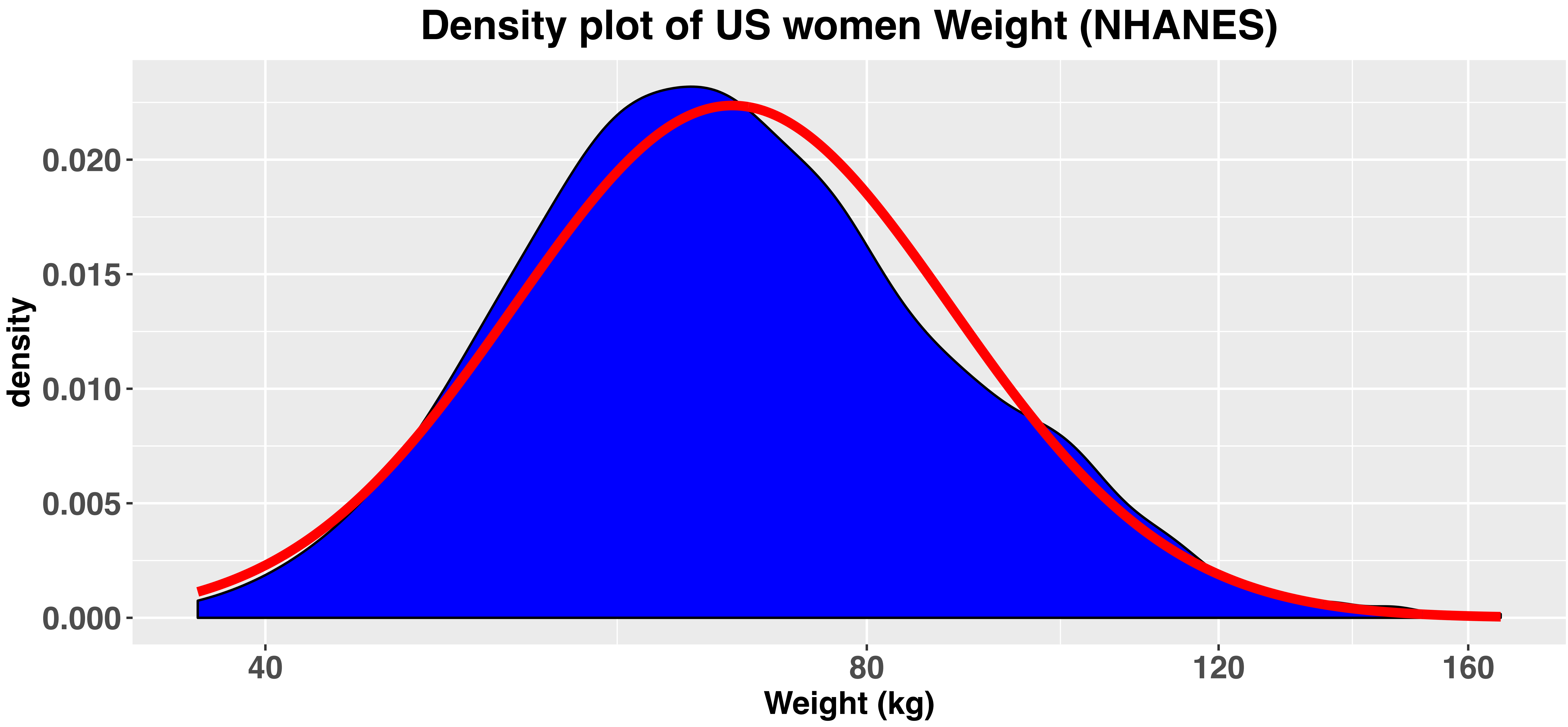


Adult weights

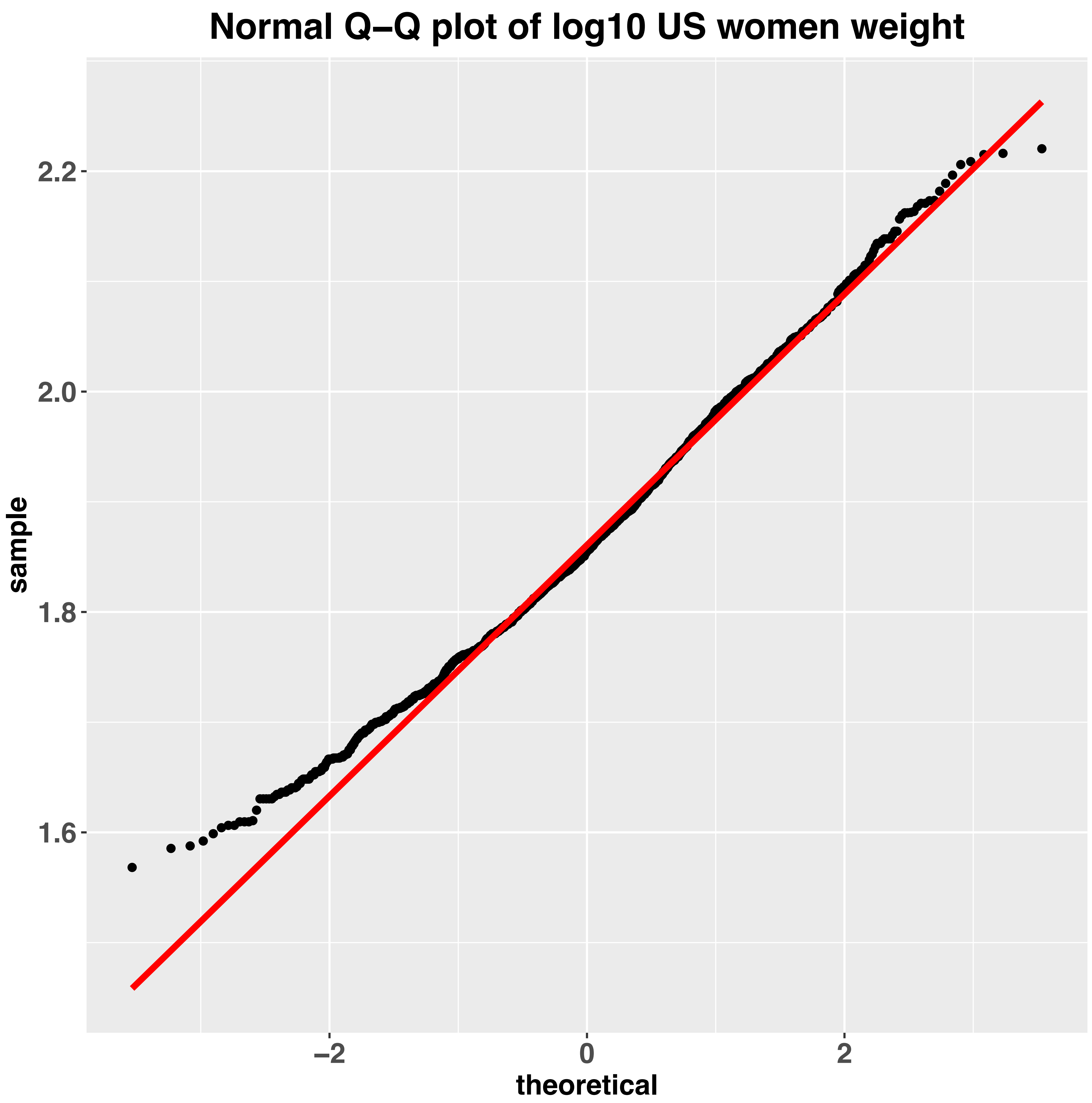


Suggests
logarithmic transformation

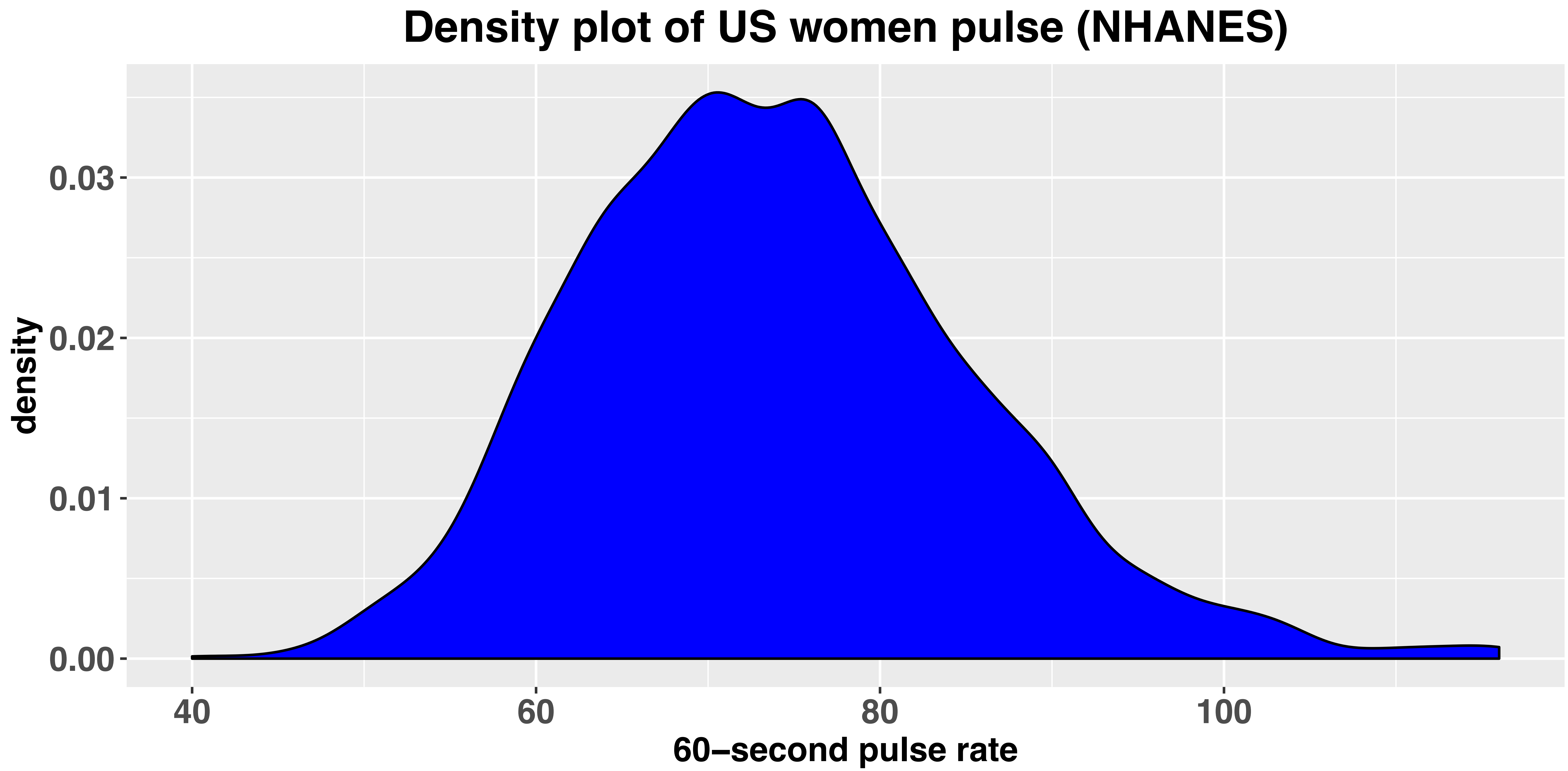
Adult weights log scale



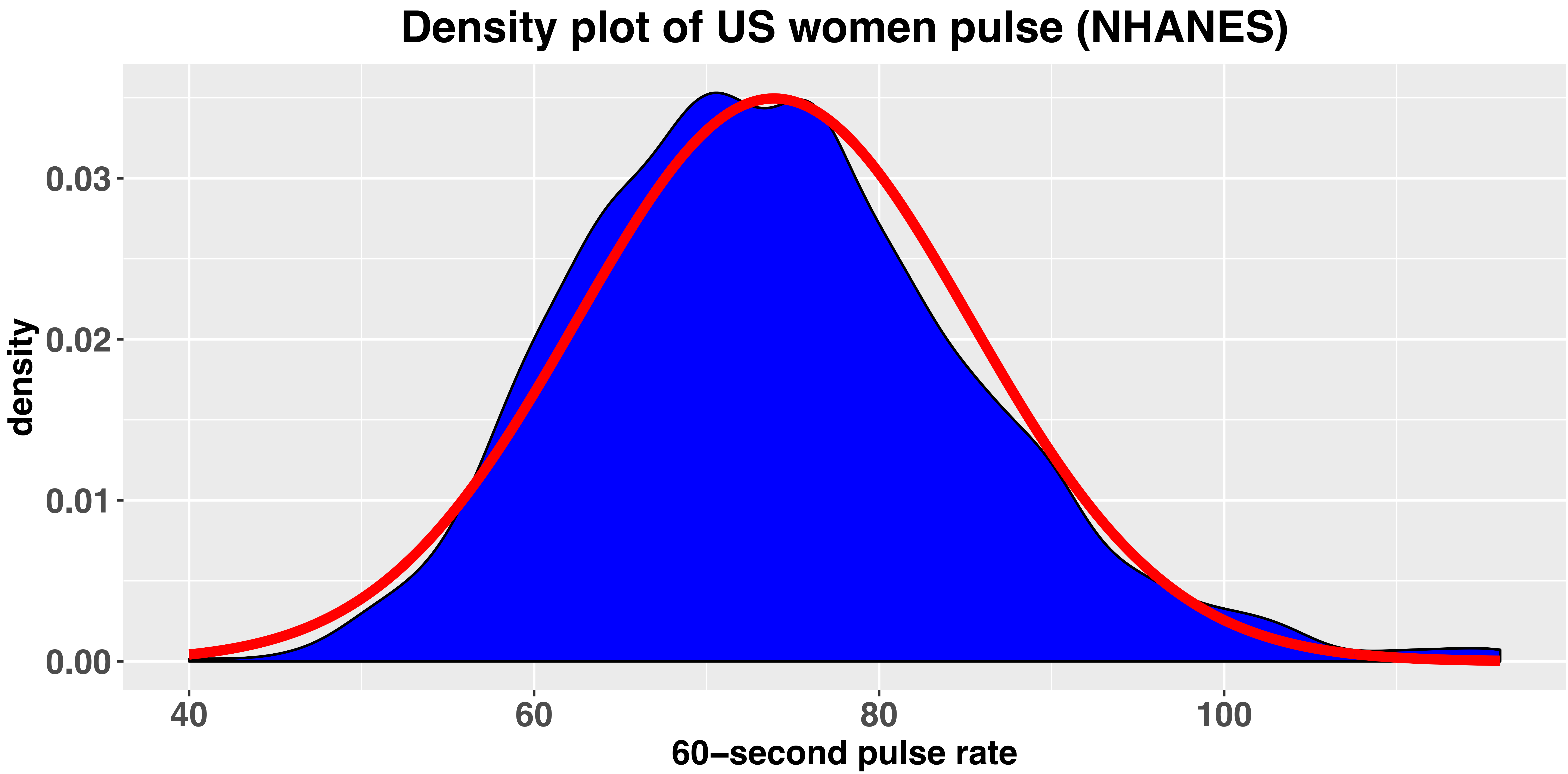
Log adult weights Q-Q plot



Pulse rate

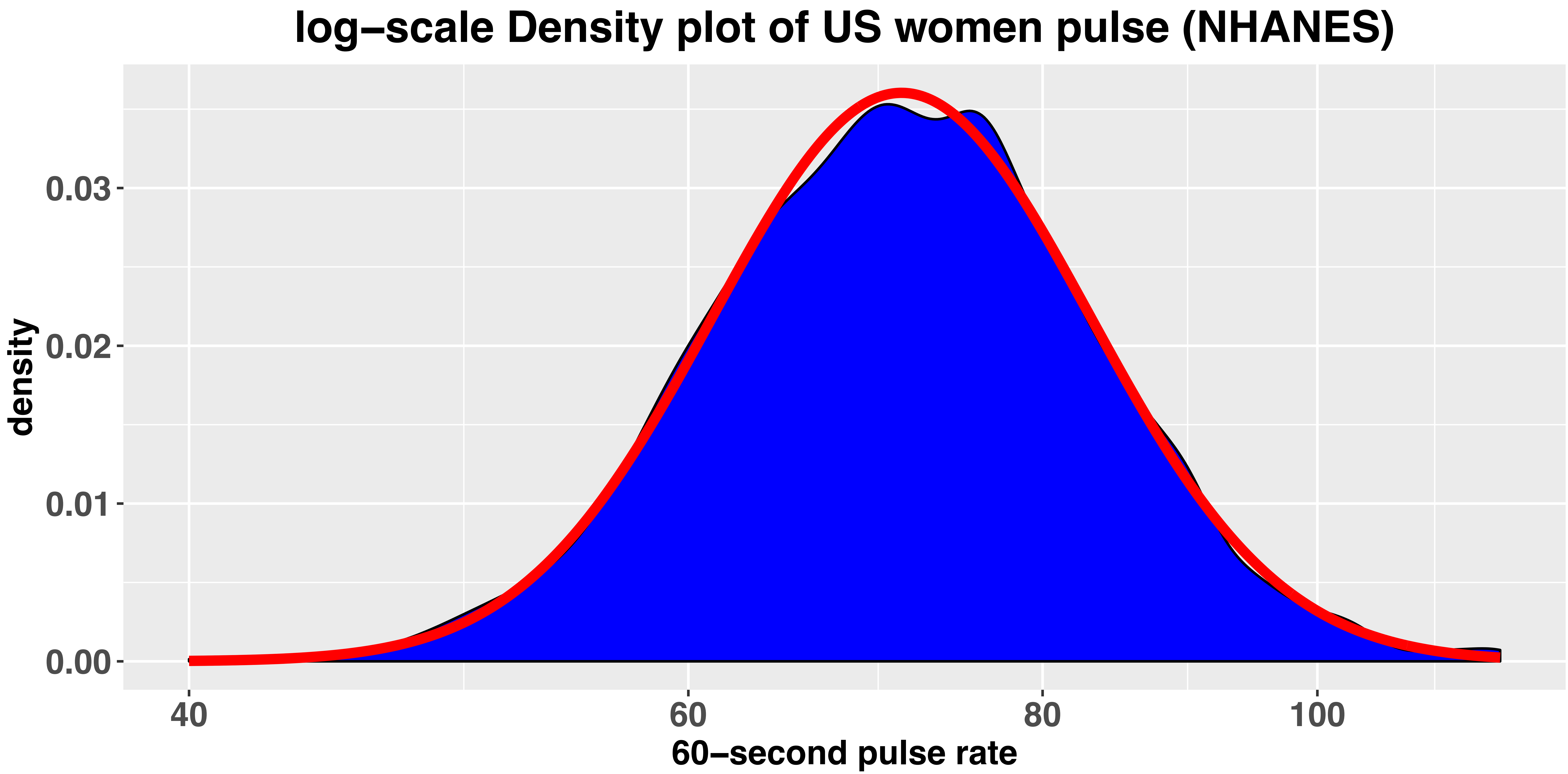

Density plot of US women pulse (NHANES)
density
0.03
0.02
0.01
0.00
40
60
80
100
60−second pulse rate

Pulse rate



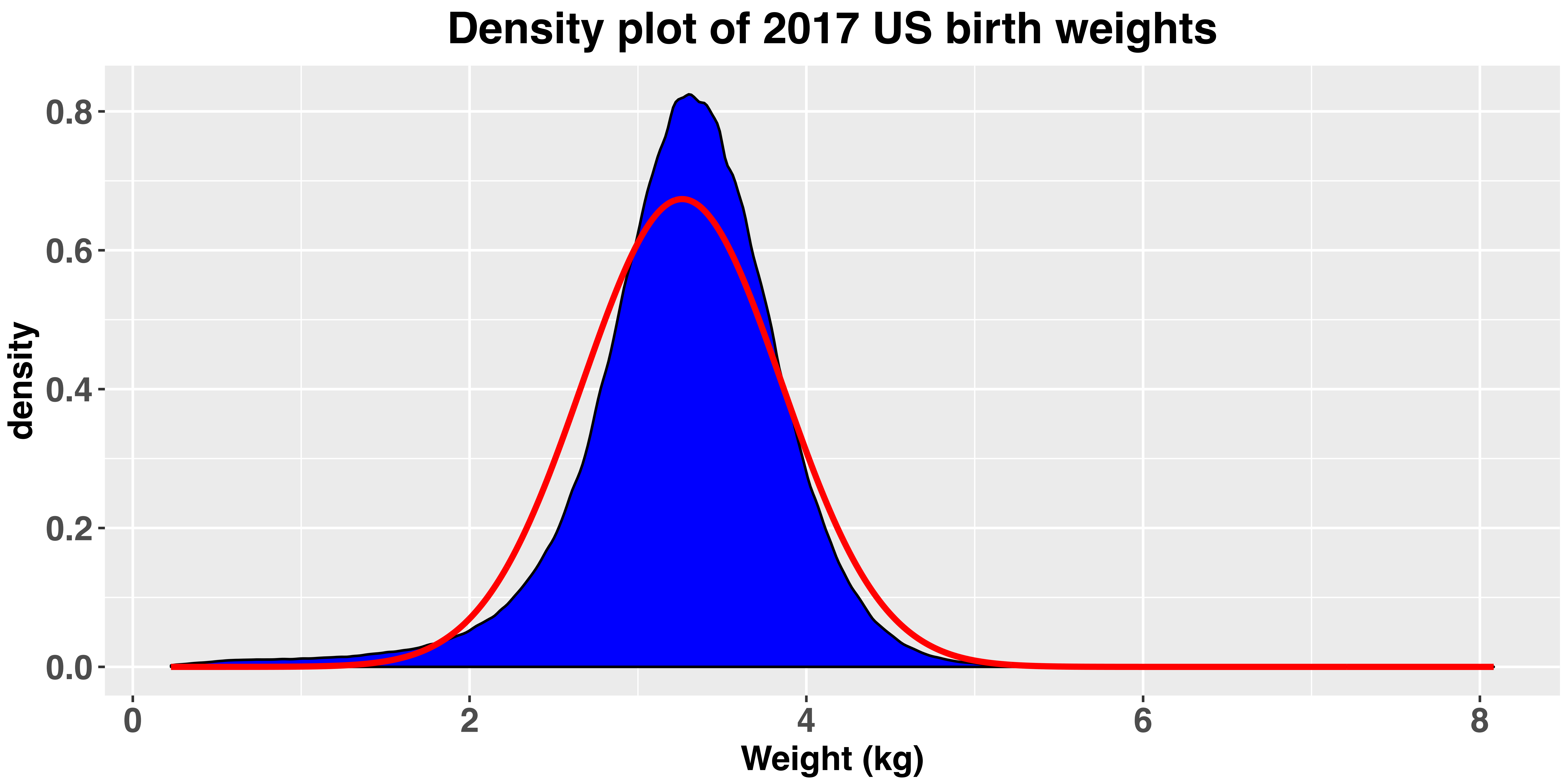
Density plot of US women pulse (NHANES)
density
0.00
0.01
0.02
0.03
40
60
80
100
60–second pulse rate

Pulse rate

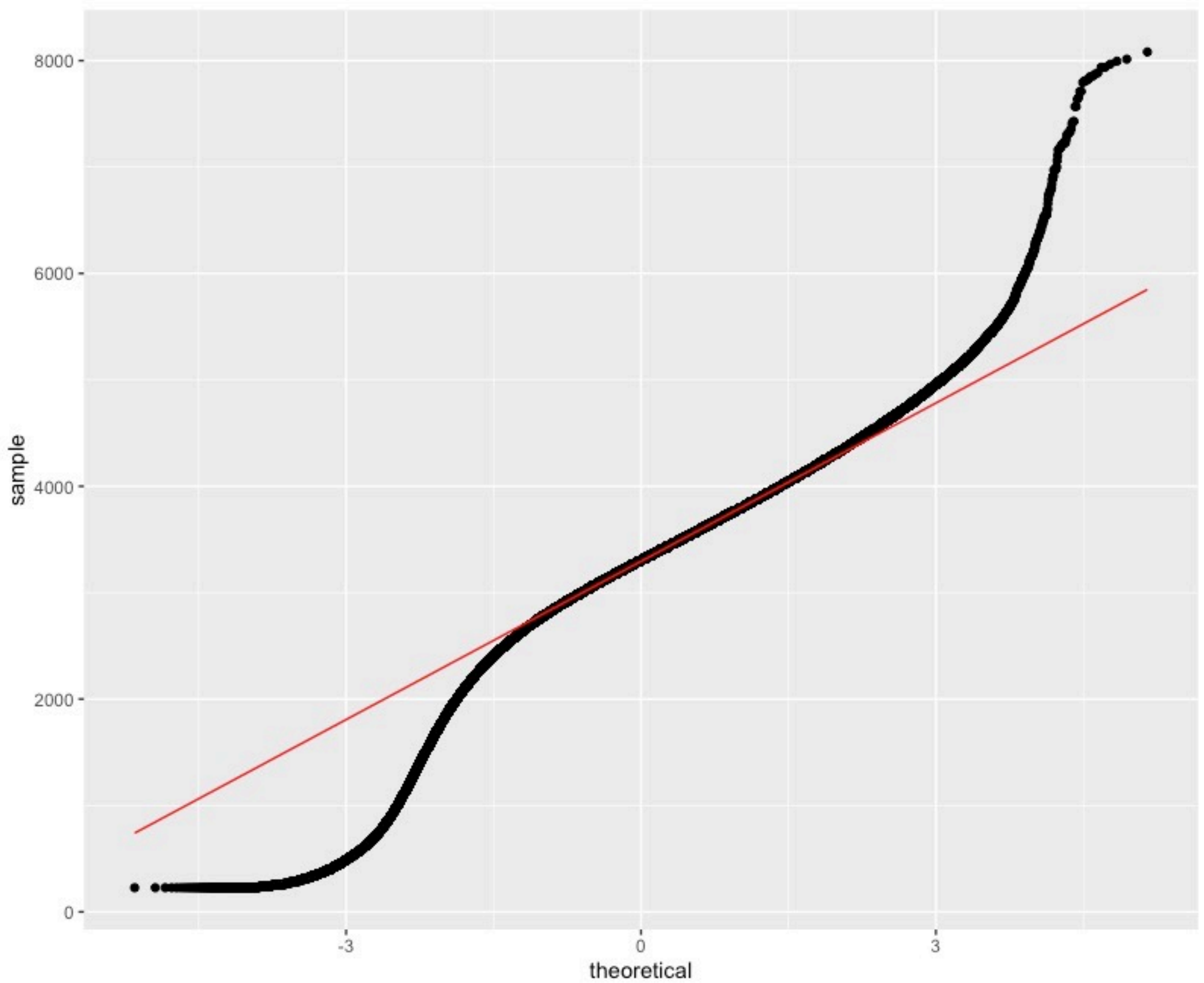


log–scale Density plot of US women pulse (NHANES)
0.03
0.02
0.01
0.00
density
40
60
80
100
60–second pulse rate

Weights of 3.9 million newborn babies



Weights of 3.9 million newborn babies



Example: Heights

Question: Given a randomly chosen US man and woman, what is the probability that the woman is taller?

Suppose the heights are normally distributed.
Which normal distributions would these be?

Men	Women
mean(heights)=1754mm	mean(heights)=1616mm
SD(heights)=75.8mm	SD(heights)=73.3mm
$\mathcal{N}(1754, 75.8^2)$	$\mathcal{N}(1616, 73.3^2)$
X = random man's height	Y = random woman's height

$$X - Y \sim \mathcal{N}(1754 - 1616, 75.8^2 + 73.3^2)$$

mean = 138mm SD = $\sqrt{75.8^2 + 73.3^2} = 105.4$mm

Example: Heights

Question: Given a randomly chosen US man and woman, what is the probability that the woman is taller?

X = random man's height Y = random woman's height

$$X - Y \sim \mathcal{N}(1754 - 1616, 75.8^2 + 73.3^2)$$

mean = 138mm SD = $\sqrt{75.8^2 + 73.3^2} = 105.4$mm

$$\mathbb{P}(X - Y < 0) = \texttt{pnorm}(0, \texttt{mean} = 138, \texttt{sd} = 105) = 0.094.$$

Alternative: Standardise.

$Z = \dfrac{\text{Height difference - 138}}{105}$ has standard normal distribution

difference $< 0 \Leftrightarrow Z < \dfrac{0 - 138}{105} = -1.31$

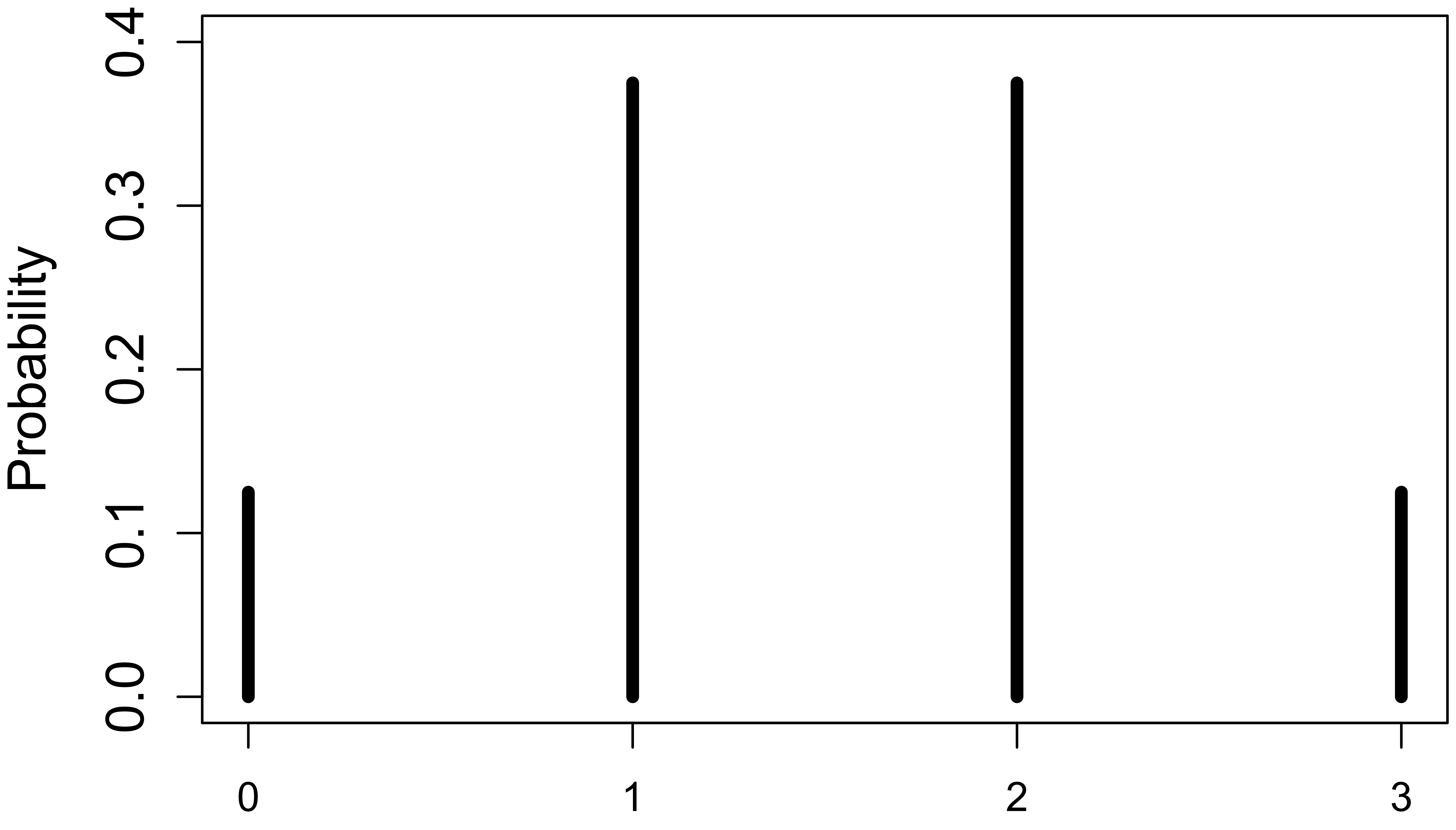
$$\mathbb{P}(Z < -1.31) = \texttt{pnorm(-1.31)} = 0.094$$

The normal approximation

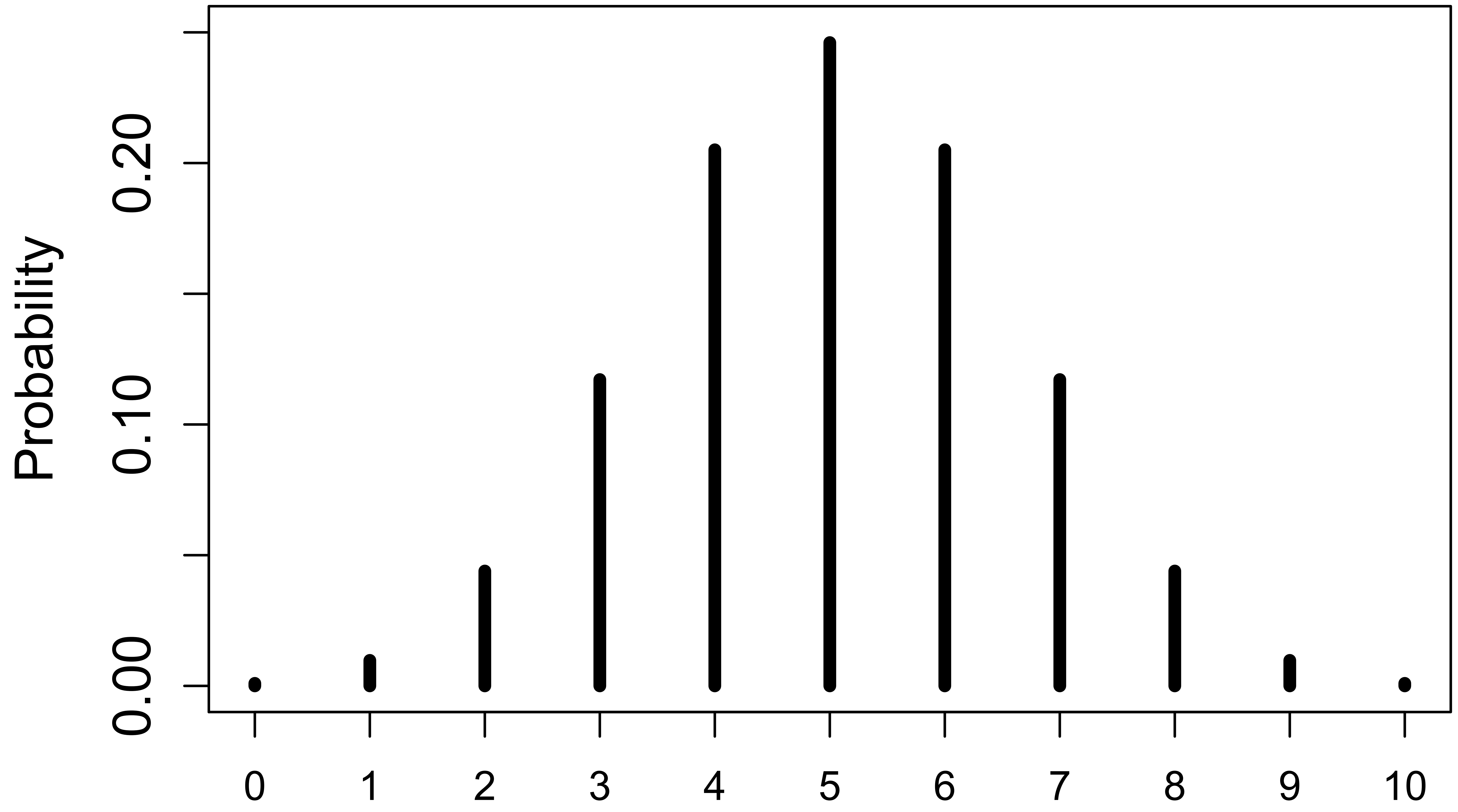
Normal approximation to the binomial

- If $X \sim \text{Bin}(n,p)$ for large n, then X is approximately normally distributed.
- Which normal distribution? We already know the mean and SD: $\mu = np$, $\sigma^2 = np(1-p)$. That's all you need to determine a normal distribution.
- How large is large? It depends on p. Rule of thumb: μ should be at least 3σ.
- What do we mean by "approximately"? $P(a < X < b)$ is close to $P(a < \mu + \sigma Z < b)$, where Z has standard normal distribution.

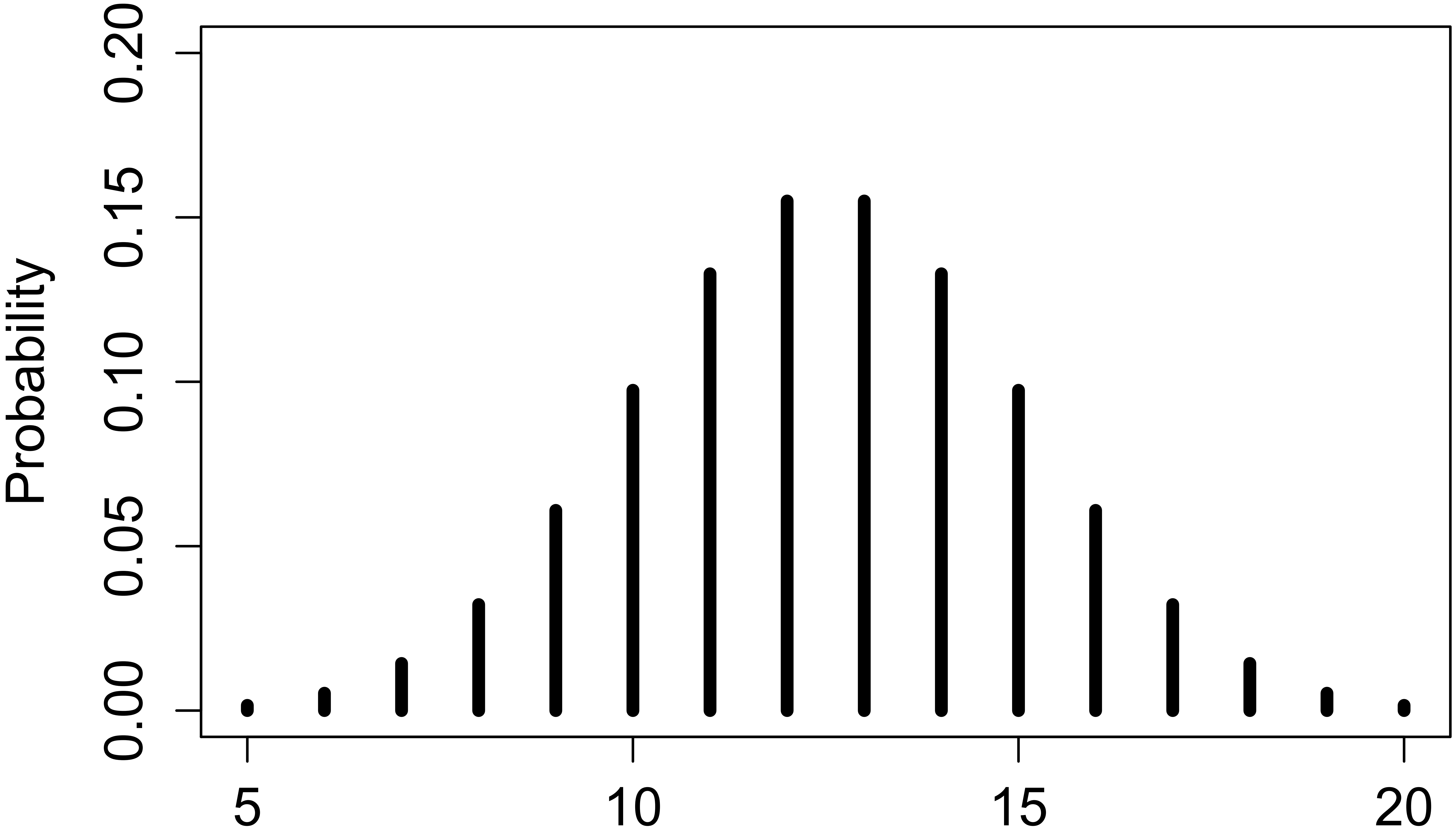
Binom(3 , 0.5) $\mu=1.5, \sigma=0.87$



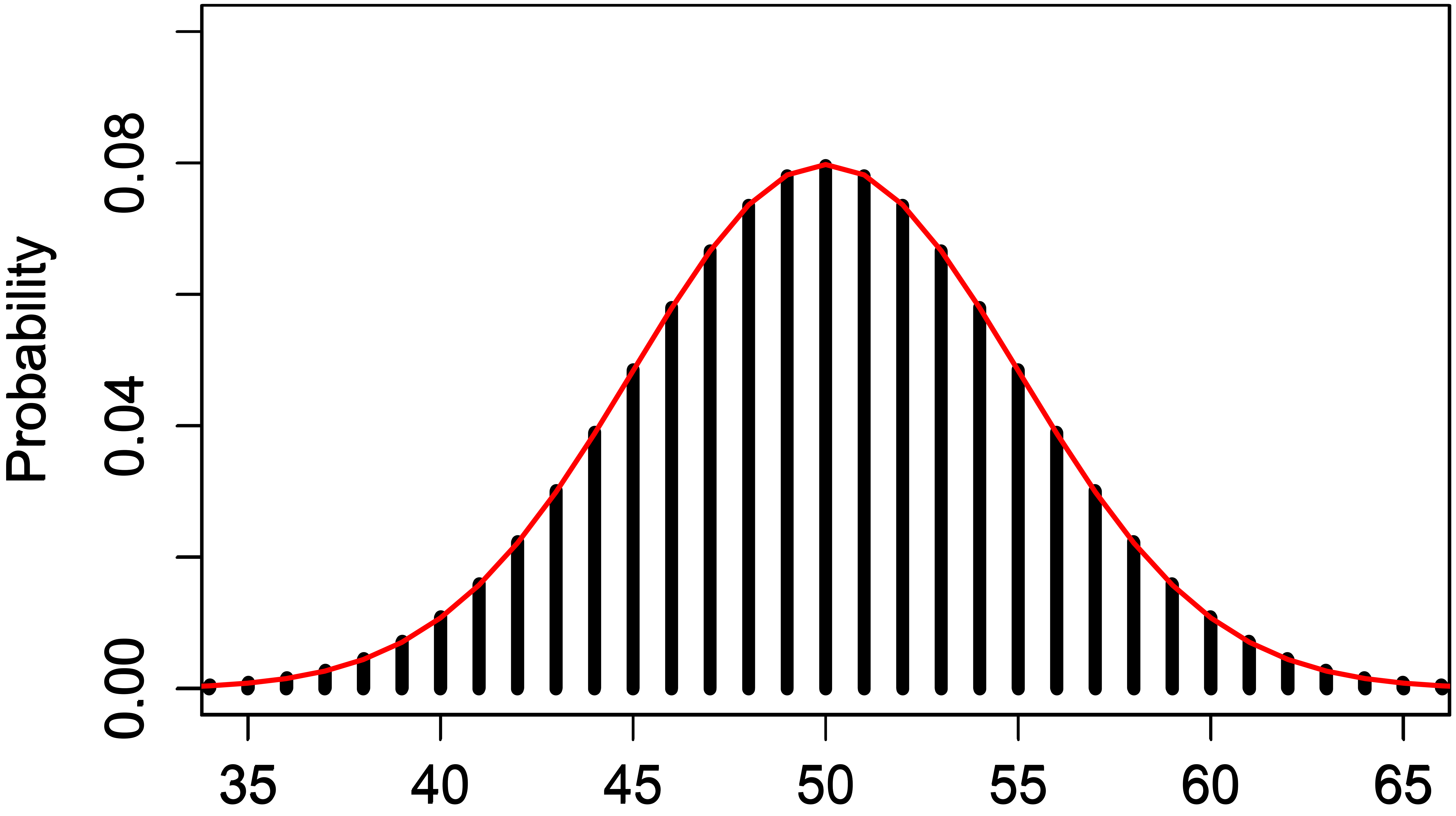
Binom(10 , 0.5) $\mu=5, \sigma=1.6$



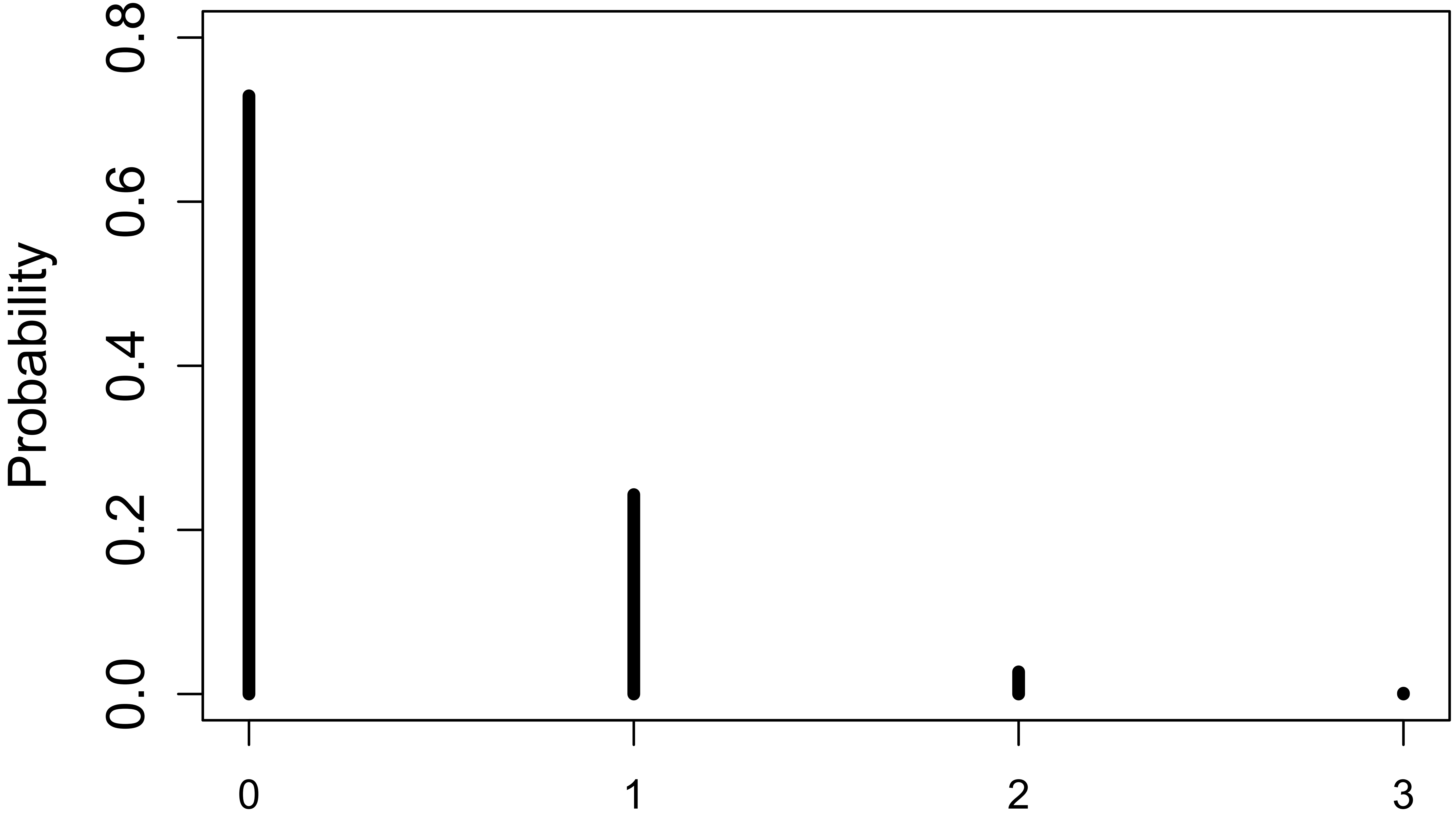
Binom(25 , 0.5) μ=12.5, σ=2.5



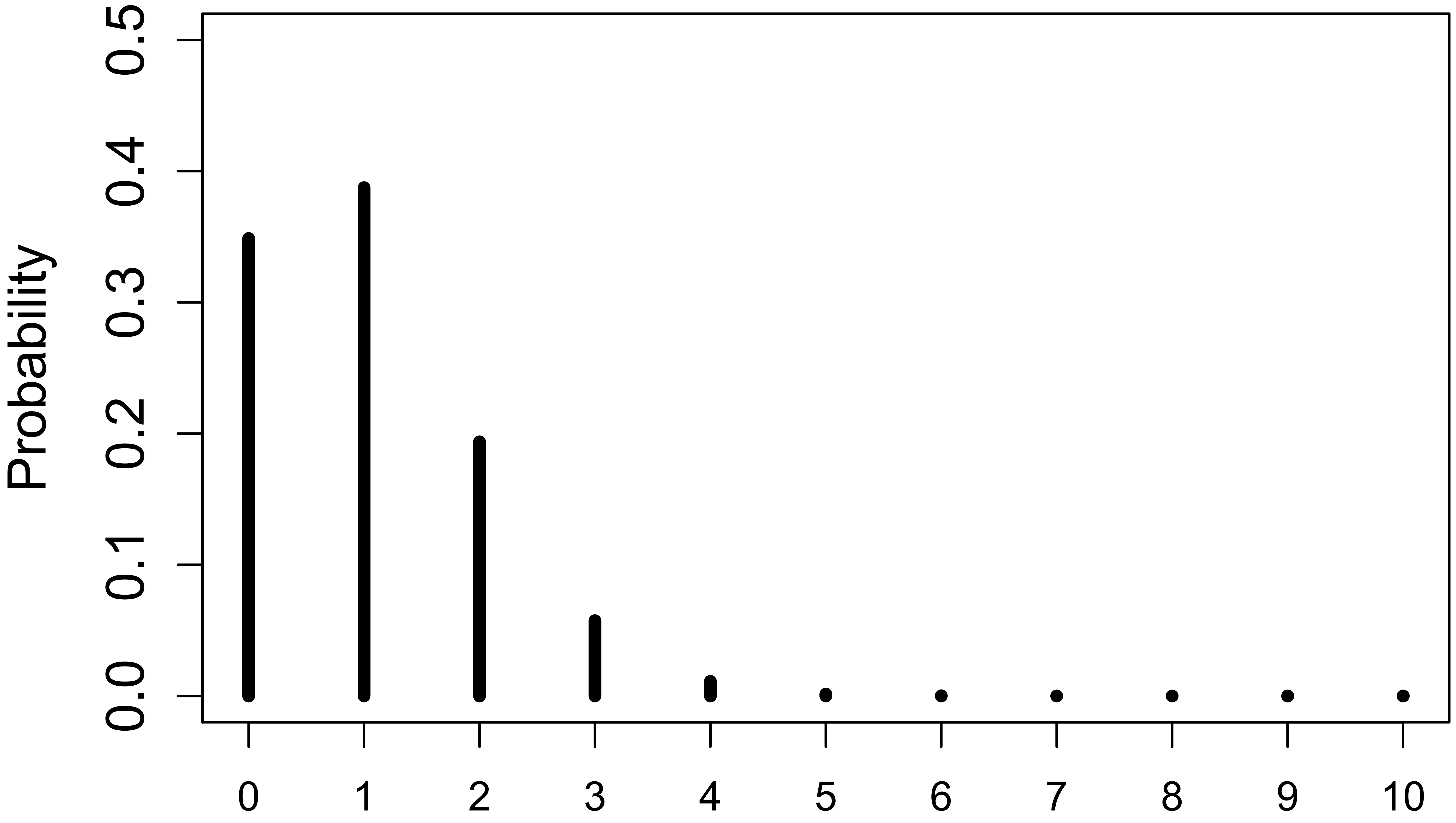
Binom(100 , 0.5) μ=50, σ=5



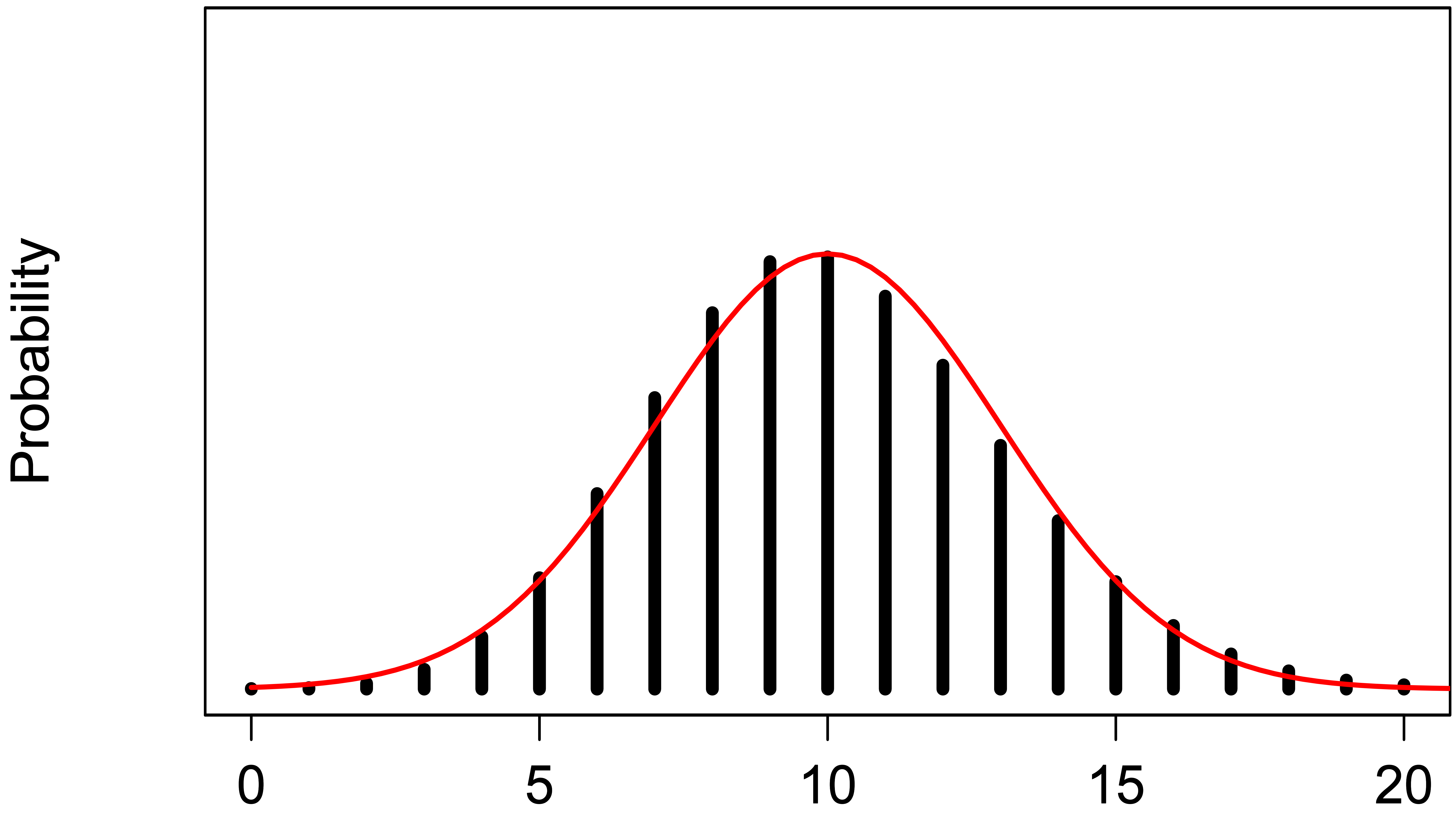
Binom(3 , 0.1) μ=0.3, σ=0.52



Binom(10 , 0.1) $\mu=1, \sigma=0.95$



Binom(100 , 0.1)   μ=10, σ=3



Example

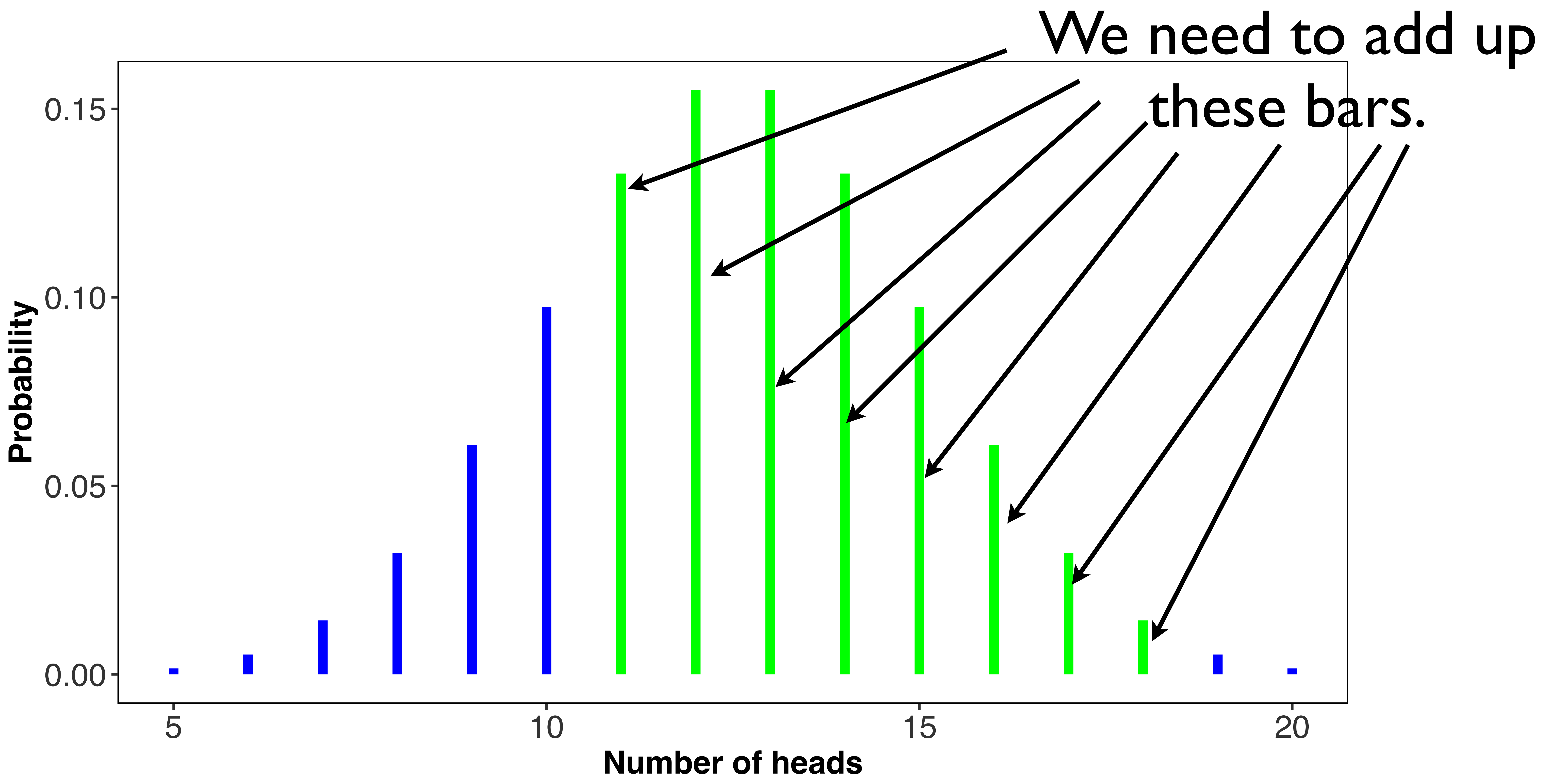
Flip 25 coins. What is the probability that the number of heads is between 11 and 18, using the normal approximation?

X=# heads~Bin(25,0.5).

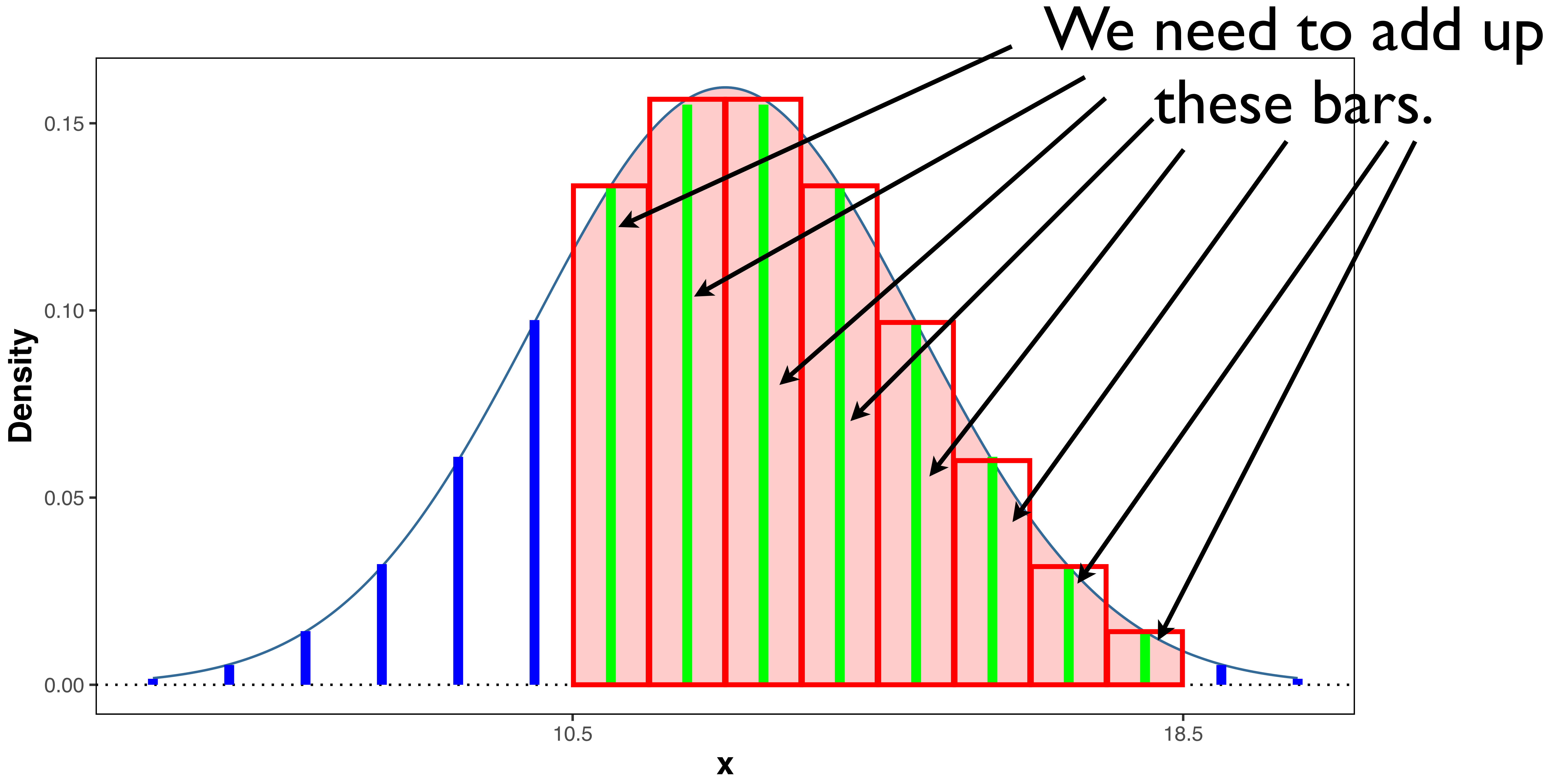
Clarification: Do we mean INCLUDING 11 and 18?

Let's say we do. So we want
P(X=11, 12, 13, 14, 15, 16, 17, or 18).

Binom(25 , 0.5)

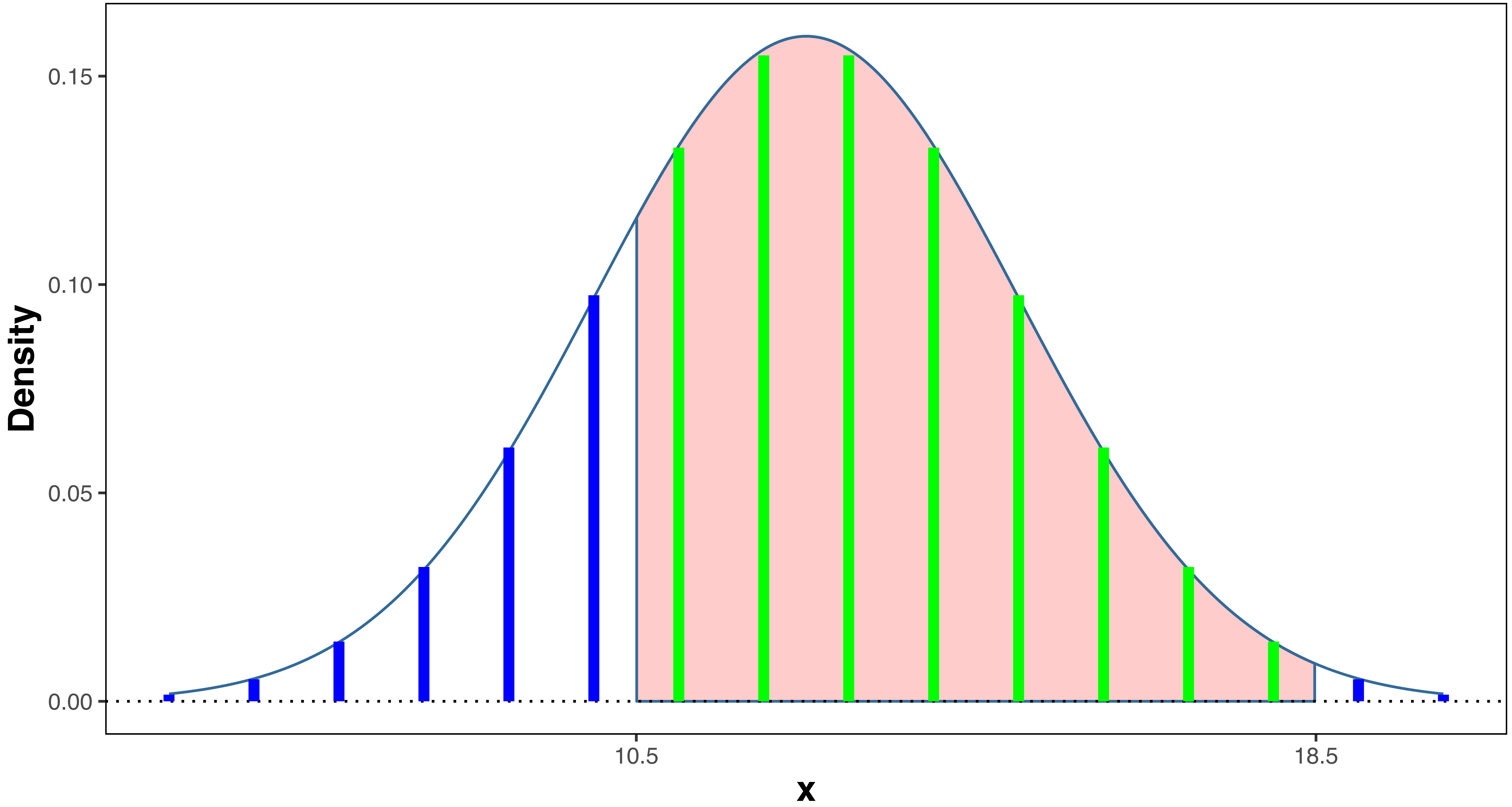

We need to add up these bars.
0.15
0.10
0.05
0.00
Probability
5
10
15
20
Number of heads

Binom(25 , 0.5)



Which is like adding up the areas of these rectangles

Binom(25 , 0.5)

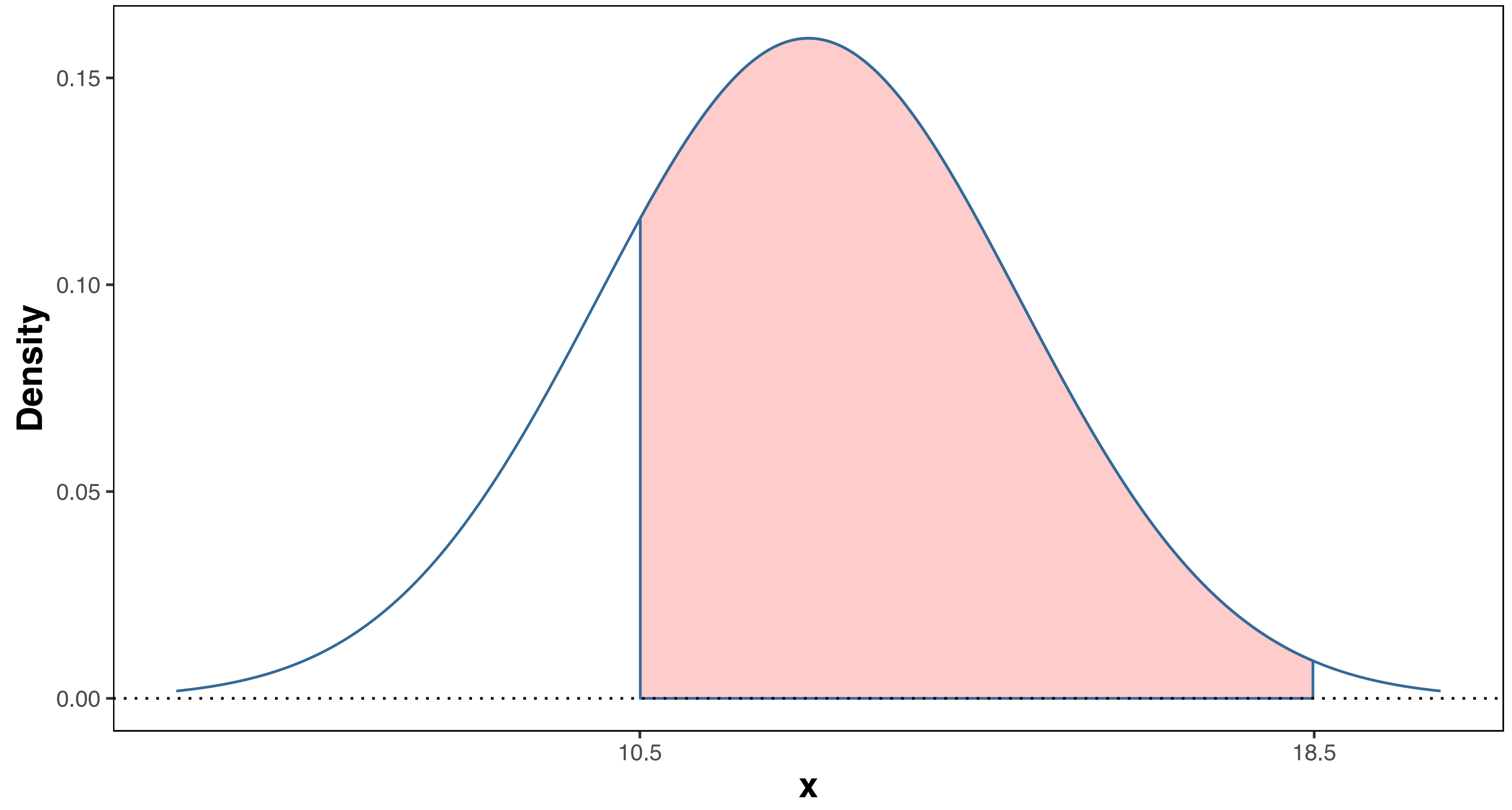


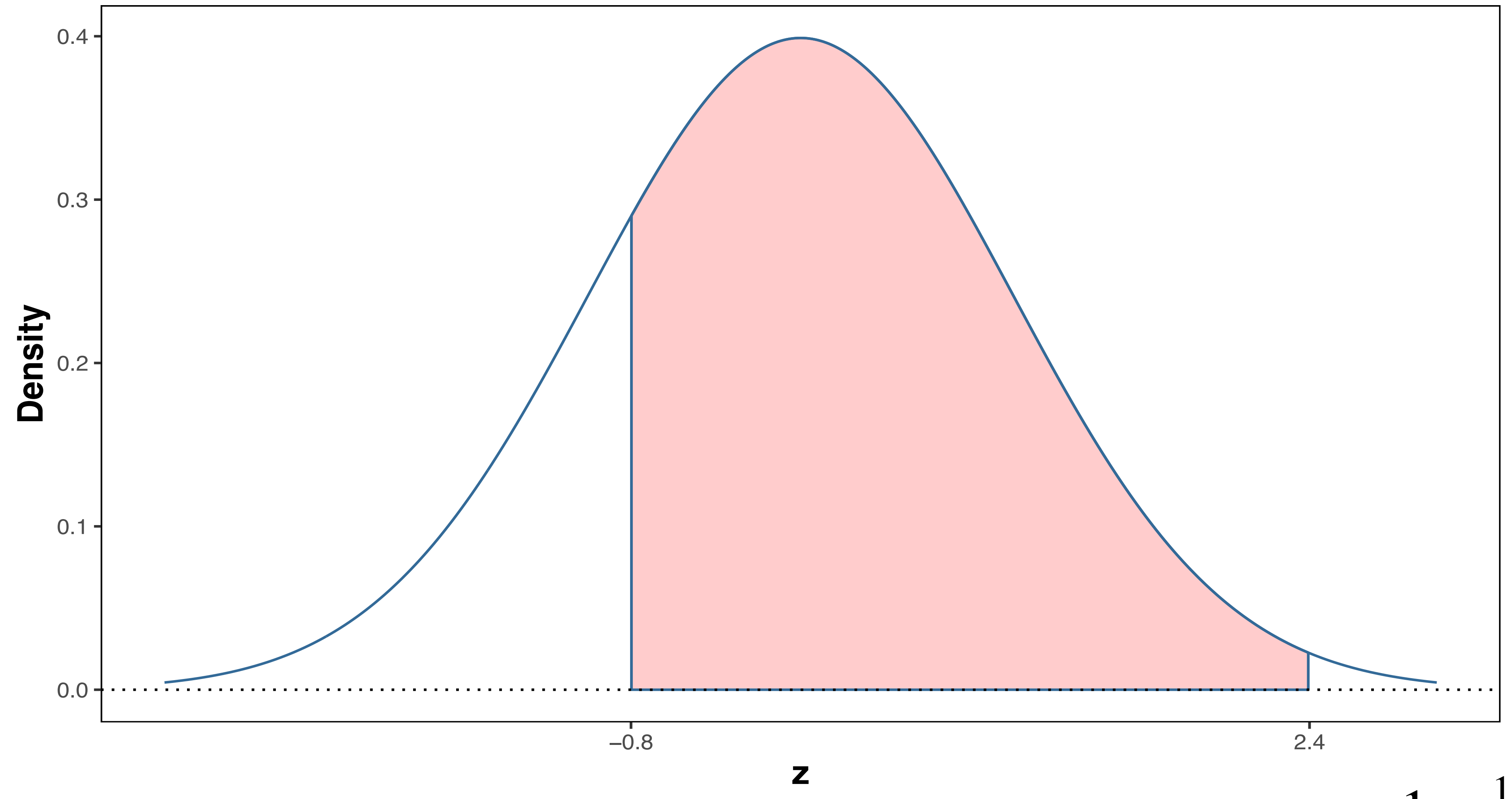
Which is like the area under the normal density curve from 10.5 to 18.5

This is called the “continuity correction”. You have to do this when a continuous distribution approximates a discrete one.

Flip 25 coins. What is the probability that the number of heads is between 11 and 18, using the normal approximation?

X=# heads∼Bin(25,0.5).





$$\mu = 25 \times 0.5 = 12.5$$

$$\sigma = \sqrt{25 \times 0.5 \times 0.5} = 2.5$$

In standard units, $z = \dfrac{x - \mu}{\sigma}$

$$z_1 = \frac{10.5 - 12.5}{2.5} = -0.8$$

$$z_2 = \frac{18.5 - 12.5}{2.5} = 2.4$$

So P(11≤X≤18) is about the same as P(-0.8<Z<2.4), where Z∼N(0,1).

$$P(-0.8 \le Z \le 2.4) = P(Z \le 2.4) - P(Z \le -0.8)$$
$$= \Phi(2.4) - \Phi(-0.8)\,.$$

```
> pnorm(2.4) -pnorm(-.8)
[1] 0.7799471
```

Exact $\dfrac{1}{2^{25}} \sum_{x=11}^{18} \binom{25}{x}$

```
> pbinom(18,25,.5)-pbinom(10,25,.5)
[1] 0.7805052
```

Joint distributions

- Whenever we have multiple random variables on a single probability space they define a **joint distribution**.

- Example: The probability space of all outcomes of 10 fair coin flips.

  - X = number of heads on first 5 flips, Y = number of heads on last 5 flips. These are independent random variables: $\mathbb{P}(X \in A \cap Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$

  - X = number of heads on first 7 flips, Y = number of heads on last 7 flips. These are not independent.

- Example: A probability space where Z is a standard normal random variable, W=|Z|.

  - X = W, Y = sgn(Z). These are independent.

  - $X = \lfloor W \rfloor$ (the integer part), $Y = \{W\} = W - \lfloor W \rfloor$ (the fractional part). Not independent.

Describing a joint distribution: Discrete

- Discrete random variables: Joint probability mass function $p_{X,Y}(x, y) = \mathbb{P}(X = x \cap Y = y)$.
- **Marginal distributions** $\mathbb{P}(X = x) = p_X(x) = \sum_y p_{X,Y}(x, y)$ **,**

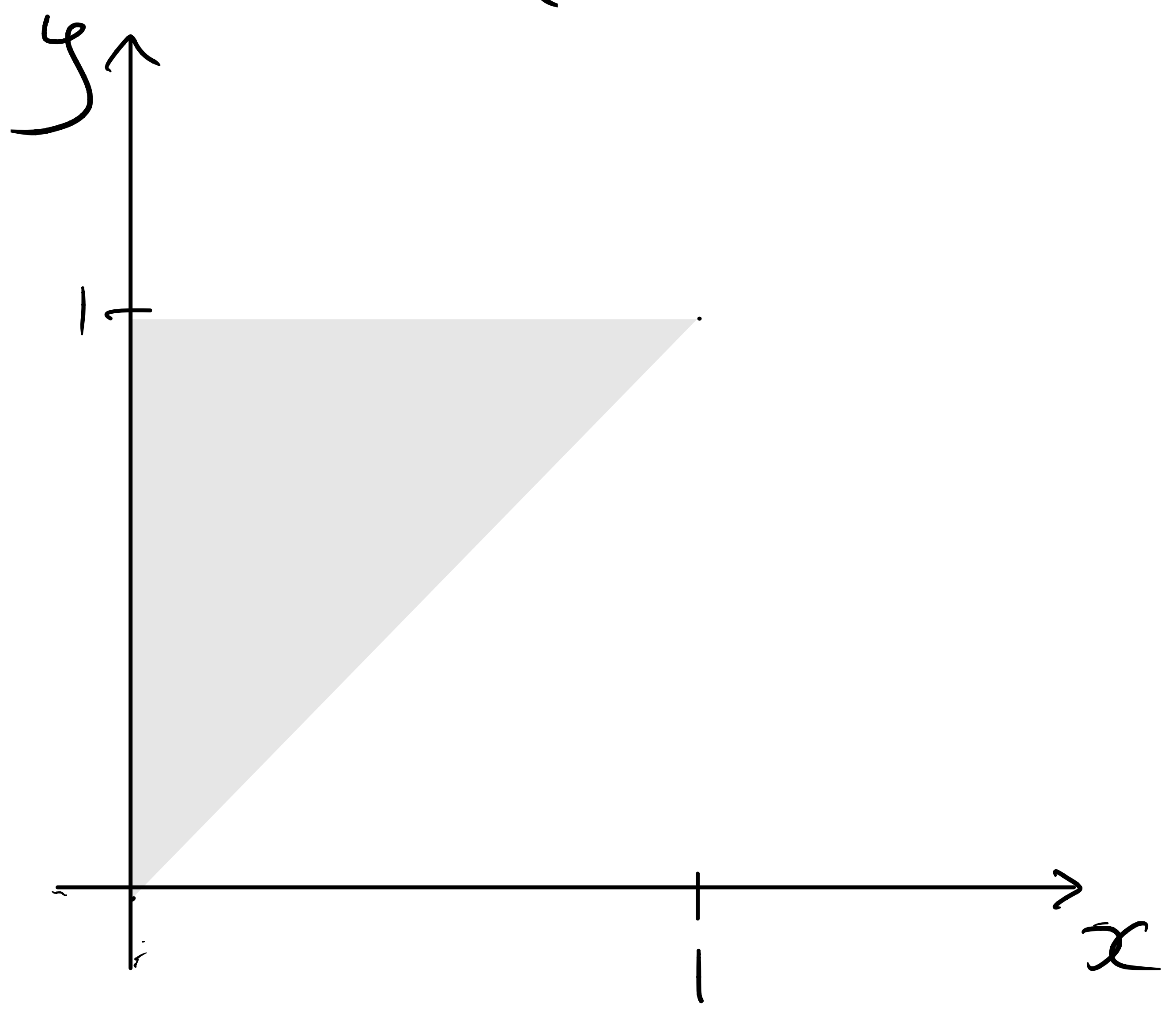
  $\mathbb{P}(Y = y) = p_Y(y) = \sum_x p_{X,Y}(x, y)$ **.**
- Independence: X and Y are independent when $p_{X,Y}(x, y) = p_X(x)p_Y(y)$.
- Conditional distribution: $p_{Y|X=x}(y) = \mathbb{P}(Y = y \,|\, X = x) = \dfrac{p_{X,Y}(x, y)}{p_X(x)}$, $p_{X|Y=y}(x) = \dfrac{p_{X,Y}(x, y)}{p_Y(y)}$.
- This definition extends obviously to more than two random variables.

Describing a joint distribution: Continuous

- Continuous random variables: Joint density $f_{X,Y}(x, y)$ is a nonnegative function with $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x, y)\mathrm{d}x\mathrm{d}y = 1.$
- $\mathbb{P}(a \leq X \leq b \ \& \ c \leq Y \leq d) = \int_c^d \int_a^b f_{X,Y}(x, y)\mathrm{d}x\mathrm{d}y$.
- **Marginal densities** $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\mathrm{d}y$ **,** $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\mathrm{d}x$ **.**
- Independence: X and Y are independent when $f_{X,Y}(x, y) = f_X(x)f_Y(y)$.
- **Conditional densities**: $f_{Y|X=x}(y) = \dfrac{f_{X,Y}(x, y)}{f_X(x)}$, $f_{X|Y=y}(x) = \dfrac{f_{X,Y}(x, y)}{f_Y(y)}$.
- This definition also extends obviously to more than two random variables.

Example

$$f_{X,Y}(x,y) = \begin{cases} 2 & \text{if } 0 < \mathrm{x} < \mathrm{y} < 1, \\ 0 & \text{otherwise}. \end{cases}$$



$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y = \int_0^1\int_0^y 2\mathrm{d}x\mathrm{d}y = \int_0^1 2y\mathrm{d}y = 1.$$

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\mathrm{d}y = \int_x^1 2\mathrm{d}y = 2 - 2x\,.$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\mathrm{d}x = \int_0^y 2\mathrm{d}x = 2y\,.$$

$$f_{X|Y=.4}(x) = \frac{f_{X,Y}(x,.4)}{f_Y(.4)} = \frac{1\{x < .4\}}{.8} = \begin{cases} 2.5 & \text{if } 0 < \mathrm{x} < .4, \\ 0 & \text{otherwise}. \end{cases}$$

Conditioned on *Y*=*y*, X is uniformly distributed on (0,y).

Example

$$f_{X,Y}(x,y) = \begin{cases} \lambda\mu \mathrm{e}^{-\lambda x - \mu y} & \text{if x > 0 \textbf{and} y > 0,} \\ 0 & \text{otherwise}\,. \end{cases}$$

X and Y are independent exponential random variables: X~Exp(λ), Y~Exp(μ).

$$\mathbb{P}(X > Y) = \int_{-\infty}^{\infty}\int_{y}^{\infty} f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y = \lambda\mu\int_{0}^{\infty}\int_{y}^{\infty} \mathrm{e}^{-\lambda x - \mu y}\mathrm{d}x\mathrm{d}y = \mu\int_{0}^{\infty}\mathrm{e}^{-(\lambda+\mu)y}\mathrm{d}y = \frac{\mu}{\lambda+\mu}\,.$$

Let Z = min(X,Y), W= max(X,Y). Change of variables formula (see probability lectures).

$$f_{Z,W}(z,w) = \begin{cases} \lambda\mu\left(\mathrm{e}^{-\lambda w - \mu z} + \mathrm{e}^{-\mu w - \lambda z}\right) & \text{if w > z > 0,} \\ 0 & \text{otherwise}\,. \end{cases}$$

$$f_Z(z) = \int_z^{\infty} \lambda\mu\left(\mathrm{e}^{-\lambda w - \mu z} + \mathrm{e}^{-\mu w - \lambda z}\right)\mathrm{d}w = \mu\mathrm{e}^{-\lambda z - \mu z} + \lambda\mathrm{e}^{-\mu z - \lambda z} = (\lambda+\mu)\mathrm{e}^{-(\lambda+\mu)z} \text{ for } z > 0.$$

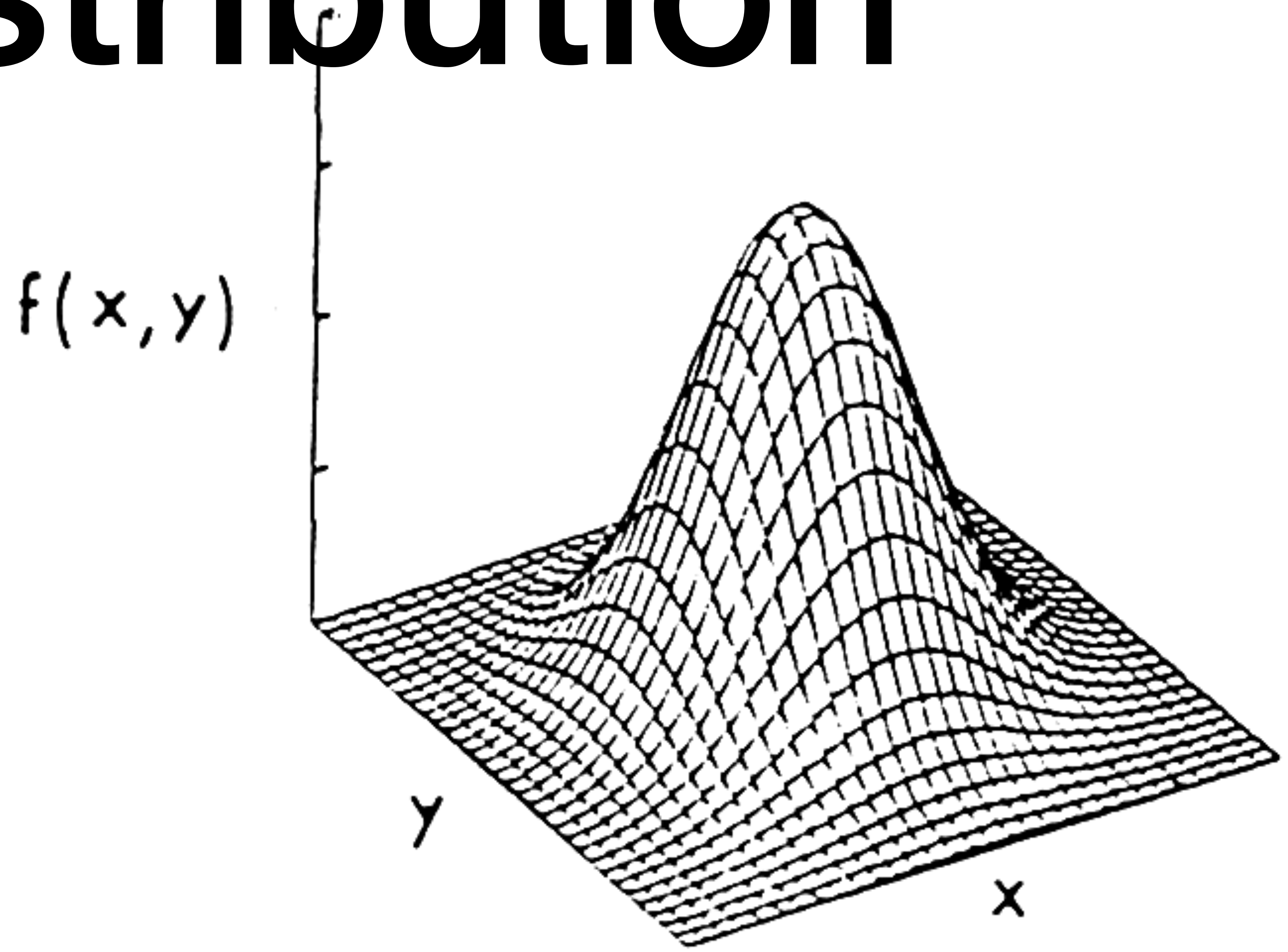
$$f_W(w) = \int_0^{w} \lambda\mu\left(\mathrm{e}^{-\lambda w - \mu z} + \mathrm{e}^{-\mu w - \lambda z}\right)\mathrm{d}z = \lambda\mathrm{e}^{-\lambda w}\left(1-\mathrm{e}^{-\mu w}\right) + \mu\mathrm{e}^{-\mu w}\left(1-\mathrm{e}^{-\lambda w}\right) = \lambda\mathrm{e}^{-\lambda w} + \mu\mathrm{e}^{-\mu w} - (\lambda+\mu)\mathrm{e}^{-(\lambda+\mu)z} \text{ for } w > 0.$$

Note: Pairs like (X,Z) are **not** jointly continuous, don't have a joint density.
Descriptions of such variables are done ad hoc, or require more advanced mathematics.

Covariance and correlation

- Covariance $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
  - $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$.
  - Measures extent to which above-average X tends to come with above-average Y.
  - But not scale invariant. e.g. Doubling X also doubles Cov(X,Y).
- Correlation $\mathrm{Cor}(X, Y) = \dfrac{\mathrm{Cov}(X, Y)}{\mathrm{SD}_X\,\mathrm{SD}_Y}$. Always between -1 and +1.

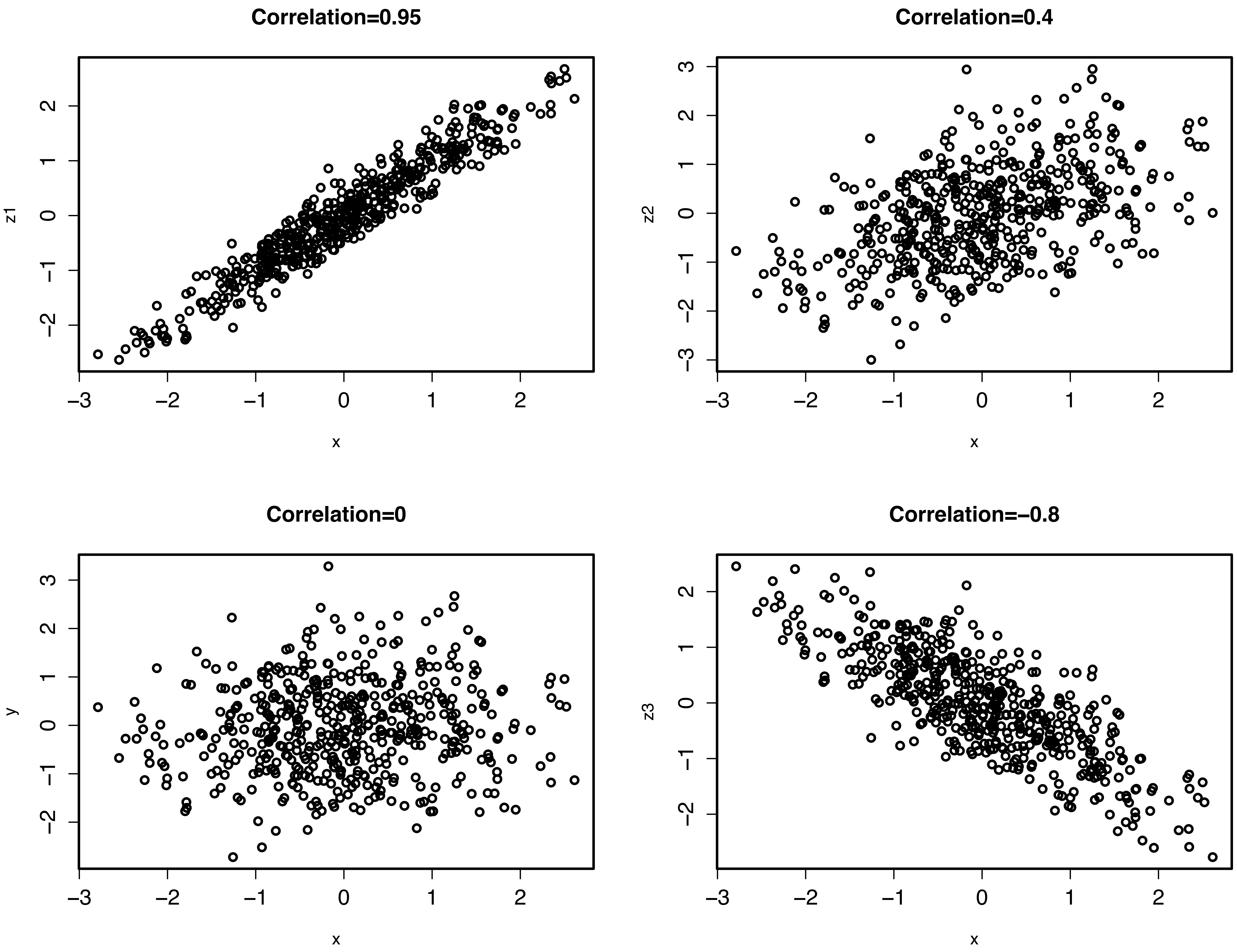
Bivariate Normal distribution



- Five parameters: Means μ_X, μ_Y , Variances σ_X^2, σ_Y^2 , Correlation ρ .
- Correlation is a number between -1 and +1, $\rho = \dfrac{\mathrm{Cov}(X, Y)}{\mathrm{SD}_X \mathrm{SD}_Y}$ where covariance $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$.
- Joint density

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right)$$

- Very important in statistical applications as model for pairs of outcomes.
- Generalises to arbitrary numbers of quantities: Multivariate normal.


Correlation=0.95
Correlation=0.4
Correlation=0
Correlation=−0.8
z1
z2
y
z3
x
−3
−2
−1
0
1
2
3

Example: Heights

Question: Given a randomly chosen US male-female married couple, what is the probability that the woman is taller? Assume as before

Men	Women
mean(heights)=1754mm	mean(heights)=1616mm
SD(heights)=75.8mm	SD(heights)=73.3mm
$\mathcal{N}(1754, 75.8^2)$	$\mathcal{N}(1616, 73.3^2)$

Correlation $\rho = 0.5$. Cov$(X, Y) = \rho \mathrm{SD}_X \mathrm{SD}_Y = 0.5 \cdot 75.8 \cdot 73.3$.

X = random man's height Y = random woman's height

$$\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(X) - 2\mathrm{Cov}(X, Y) = 75.8^2 + 73.3^2 - 2 \cdot 0.5 \cdot 75.8 \cdot 73.3 = 5562 = 74.6^2$$

mean $= 138$mm SD74.6mm $\mathbb{P}(X - Y < 0) = \mathtt{pnorm}(0, \mathtt{mean} = 138, \mathtt{sd} = 74.6) = 0.032.$

Alternative: Standardise $Z = \dfrac{\text{Height difference - 138}}{74.6}$ has standard normal distribution.

difference $< 0 \Leftrightarrow Z < \dfrac{0 - 138}{74.6} = -1.85$ $\mathbb{P}(Z < -1.85) = \mathtt{pnorm}(-1.85) = 0.032.$