

## Course outline through exam 3.

- Describing data (chapters 1-3)
  - graphically
  - with summary statistics
- Estimation – point estimates and confidence intervals of the mean. (chapter 4)
- Probability rules (chapter 5)
- Hypothesis testing (chapter 6)
  1. binomial (chapter 7)
  2.  $\chi^2$  goodness of fit (chapter 8)
    - Contingency analysis (chapter 9)
  3. Tests of continuous parameters
    - Background on the normal distribution (chapter 10)
    - Tests assuming normality (chapter 11)
    - Comparing means (chapter 12)
    - Violations of assumptions (chapter 13)

Has fish size changed in a salmon hatchery?

Consider a case in which very extensive sampling (huge  $n$ ) in 1985 gives us essentially error-free estimates of the fish sizes.

In 2009 we measure a hatchling and a 3-week old fish.

$H_0$ : The population mean hatchling size has not changed since 1985.

$H_A$ : The population mean hatchling size has changed since 1985.

$$H_0: \mu_{h2009} = \mu_{h1985}$$

$$H_A: \mu_{h2009} \neq \mu_{h1985}$$

Another  $H_0$ : The population mean size at 3 weeks has not changed since 1985.  $\mu_{2009} = \mu_{1985}$

The other  $H_A$ :  $\mu_{2009} \neq \mu_{1985}$

Steps to hypothesis testing:

1. Ask a scientific question.
2. State your question in terms of  $H_0$  and  $H_A$
3. Collect a random sample
4. Calculate descriptive statistics
5. Determine  $P$ -value:
  - (a) What values of the test statistic are expected under  $H_0$ ?
  - (b) How does the observed test statistic differ from these expectations?
  - (c) What is the probability of observing a value of a test statistic this extreme or more extreme if the  $H_0$  is true? – this is the  $P$ -value.
6. Make a decision about  $H_0$  and  $H_A$
7. Answer your question and report the results.

Mass of Salmon hatchlings:

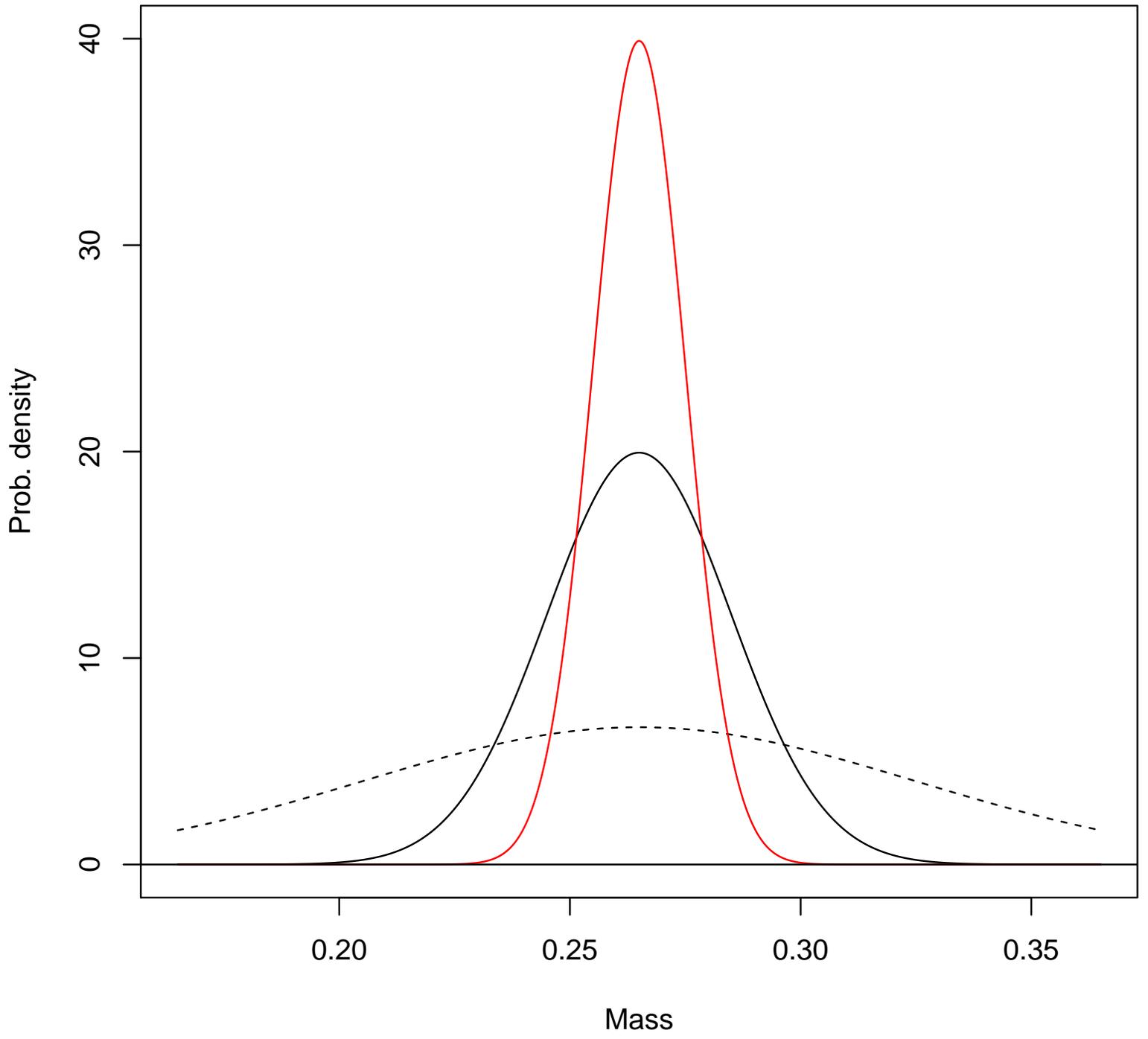
$$\mu_{h1985} = 0.265\text{g}$$

$$Y_{h2009} = 0.190\text{g}$$

Mass at 3 weeks:

$$\mu_{1985} = 0.625\text{g}$$

$$Y_{2009} = 0.7\text{g}$$



$H_0$ : Hatchling size is described by a normal distribution, and the population mean and variance of hatchling size has not changed since 1985.

$H_A$ : otherwise.

$H_0$ :  $\mu_{h2009} = \mu_{h1985}$  and  $\sigma_{h2009} = \sigma_{h1985}$

$H_A$ :  $\mu_{h2009} \neq \mu_{h1985}$  or  $\sigma_{h2009} \neq \sigma_{h1985}$

Mass of Salmon hatchlings:

$$\mu_{h1985} = 0.265\text{g}$$

$$\sigma_{h1985} = 0.02\text{g}$$

$$Y_{h2009} = 0.190\text{g}$$

Hypothetical mass at 3 weeks:

$$\mu_{1985} = 0.625\text{g}$$

$$\sigma_{1985} = 0.0456\text{g}$$

$$Y_{2009} = 0.7\text{g}$$

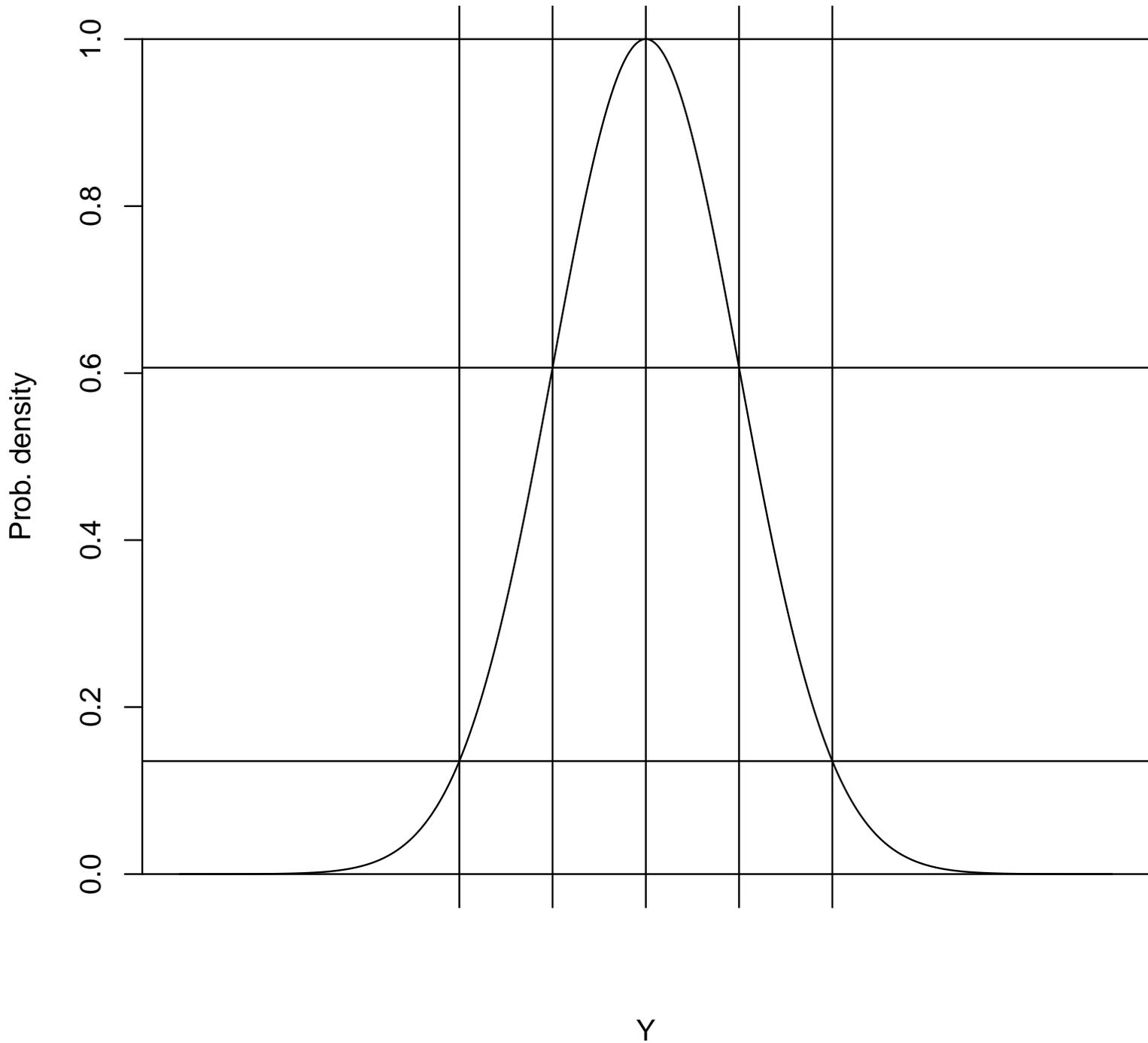
$$f(Y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y-\mu)^2}{2\sigma^2}}$$

optional drawing hints:

$$e^{-0} = 1 \quad Y = \mu \quad e^{-\frac{(\mu-\mu)^2}{2\sigma^2}} = e^{-0}$$

$$e^{-0.5} \approx 0.61 \quad Y = \mu \pm \sigma \quad e^{-\frac{(\mu+\sigma-\mu)^2}{2\sigma^2}} = e^{-0.5}$$

$$e^{-2} \approx 0.13 \quad Y = \mu \pm 2\sigma \quad e^{-\frac{(\mu+2\sigma-\mu)^2}{2\sigma^2}} = e^{-2}$$



Properties of the Normal distribution:

- continuous probability distribution,
- symmetric
- mode = mean = median

null hypotheses:

$$Y_{h2009} \sim \text{Normal}(\mu = \mu_{h1985}, \sigma = \sigma_{h1985})$$

$$Y_{2009} \sim \text{Normal}(\mu = \mu_{1985}, \sigma = \sigma_{1985})$$

$D$ -statistic (I just made this up):

$$D = Y - \mu$$

$$D_{h2009} = 0.190\text{g} - 0.265\text{g} = -0.075\text{g}$$

$$D_{2009} = 0.7\text{g} - 0.625\text{g} = 0.075\text{g}$$

null hypotheses:

$$D_{h2009} \sim \text{Normal}(\mu = 0, \sigma = \sigma_{h1985})$$

$$D_{2009} \sim \text{Normal}(\mu = 0, \sigma = \sigma_{1985})$$

$D$ -statistic (I just made this up):

$$D = Y - \mu$$

$$D_{h2009} \sim \text{Normal}(\mu = 0, \sigma = \sigma_{h1985})$$

$Z$ -statistic (this is what we actually use):

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

$$Z \sim \text{Normal}(\mu = 0, \sigma = 1)$$

$$Z_{h2009} = -0.075\text{g}/0.02\text{g} = -3.75$$

$$Z_{2009} = 0.075\text{g}/0.0456\text{g} = 1.64$$

A  $Z$  statistic is a measure of how far your observation differs from the expected value, measured in units of “number of standard deviations”

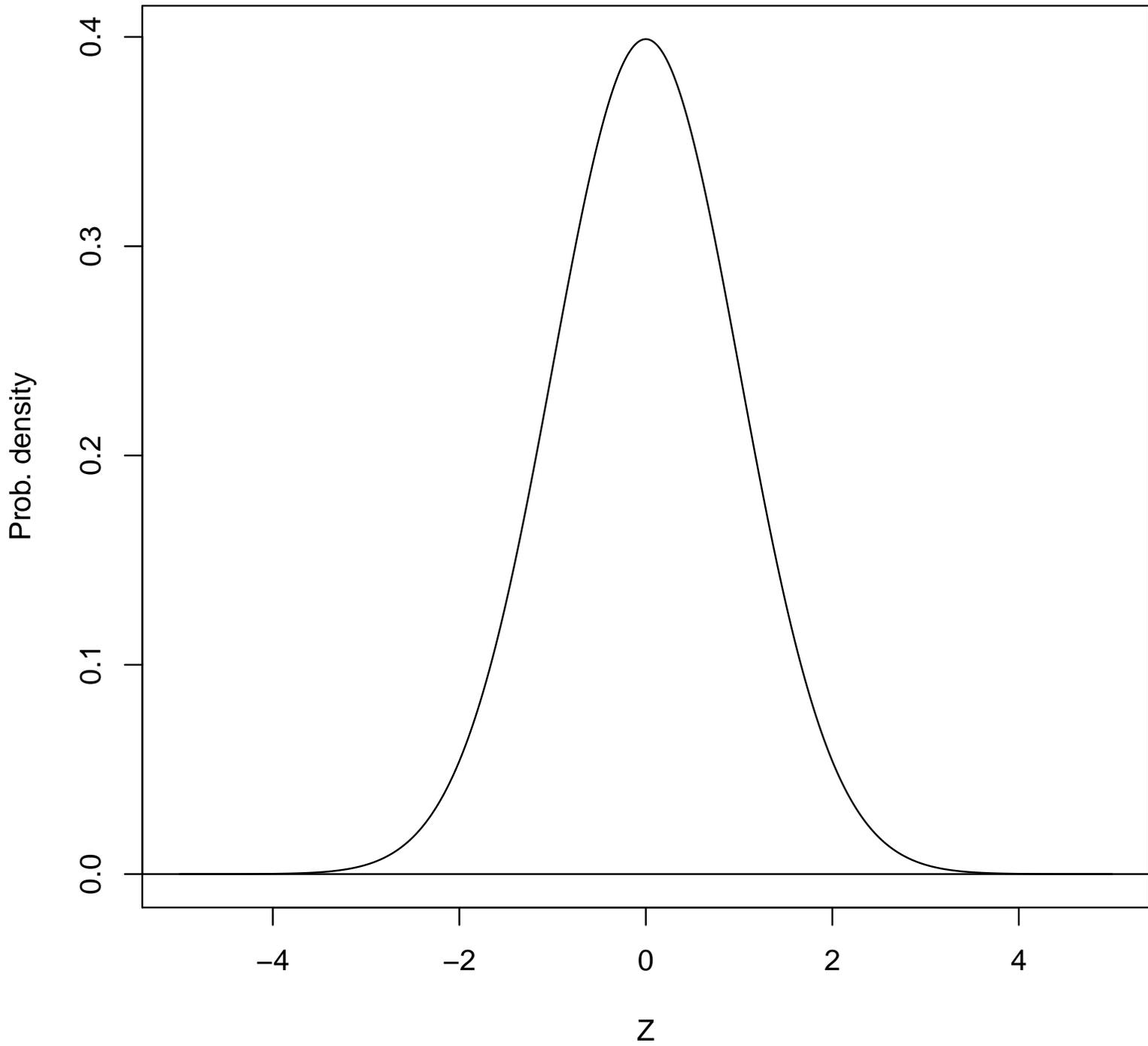
$Z_{h2009} = -3.75$  means

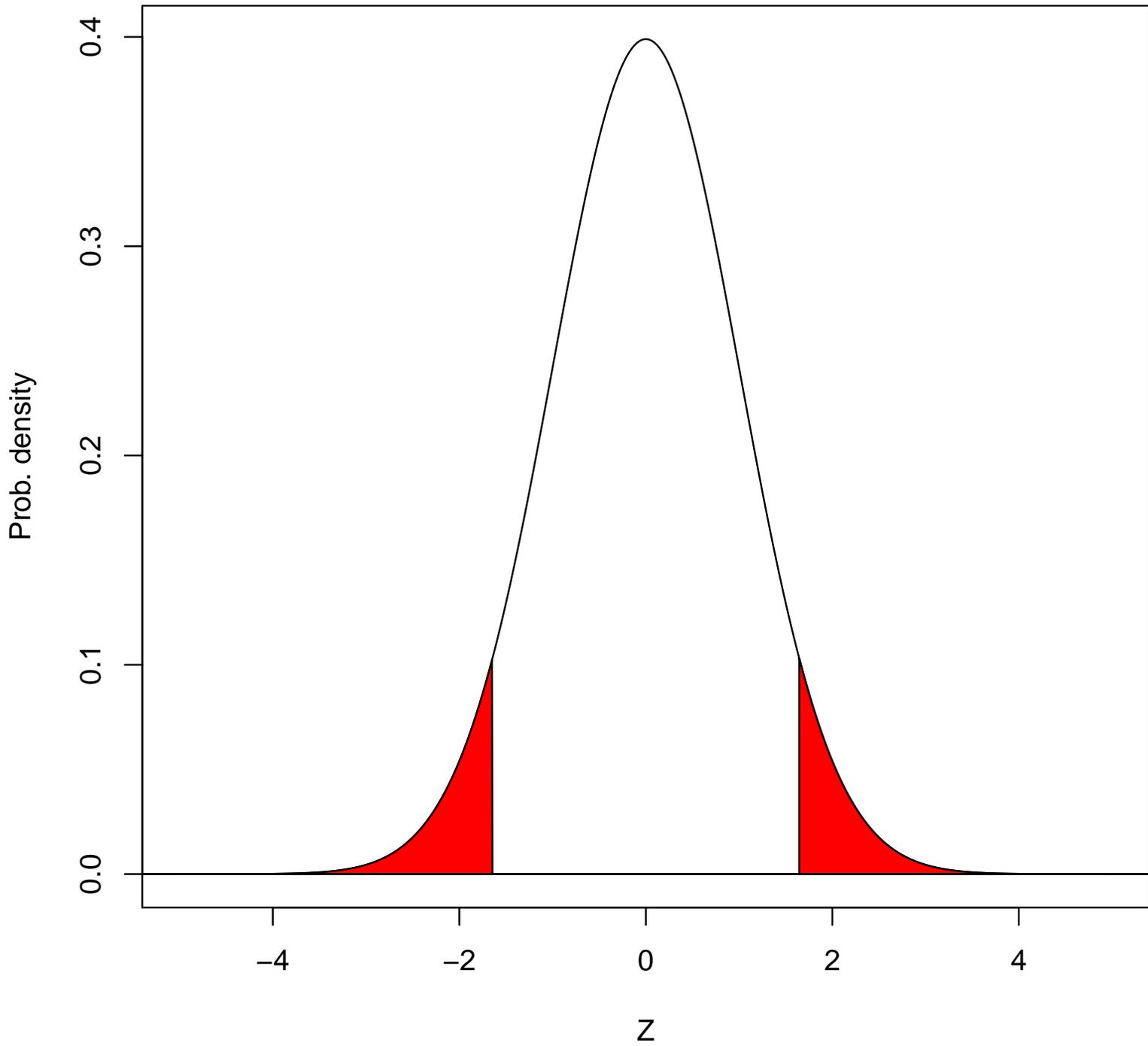
Our hatchling mass was 3.75 standard deviations smaller than the expected value.

$Z_{2009} = 1.64$  means

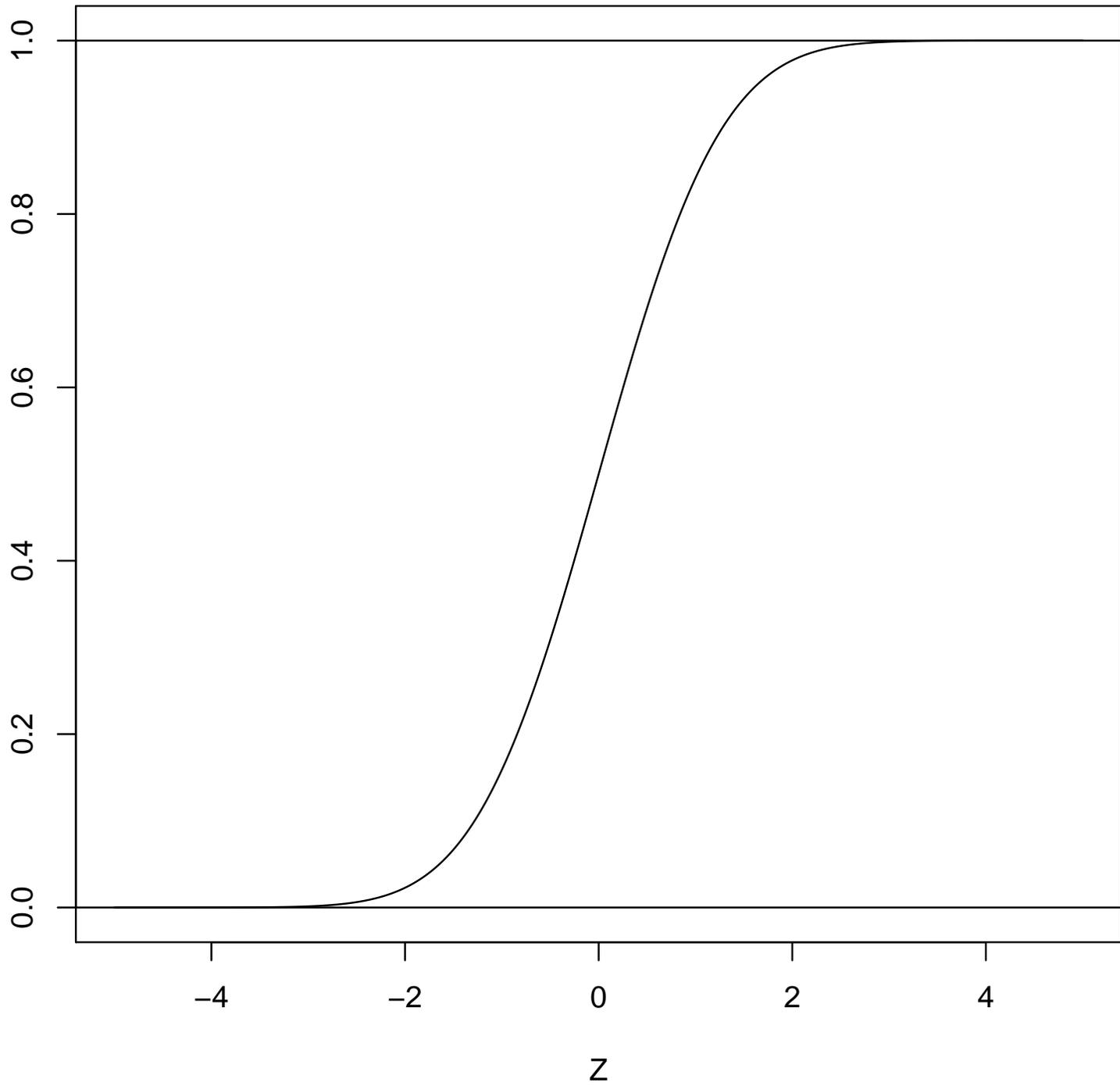
The mass at three weeks was 1.64 standard deviations larger than the expected value.

We are “standardizing” for different variances  
→ the “standard normal distribution”





Cumulative Prob.



Right Tail Prob.

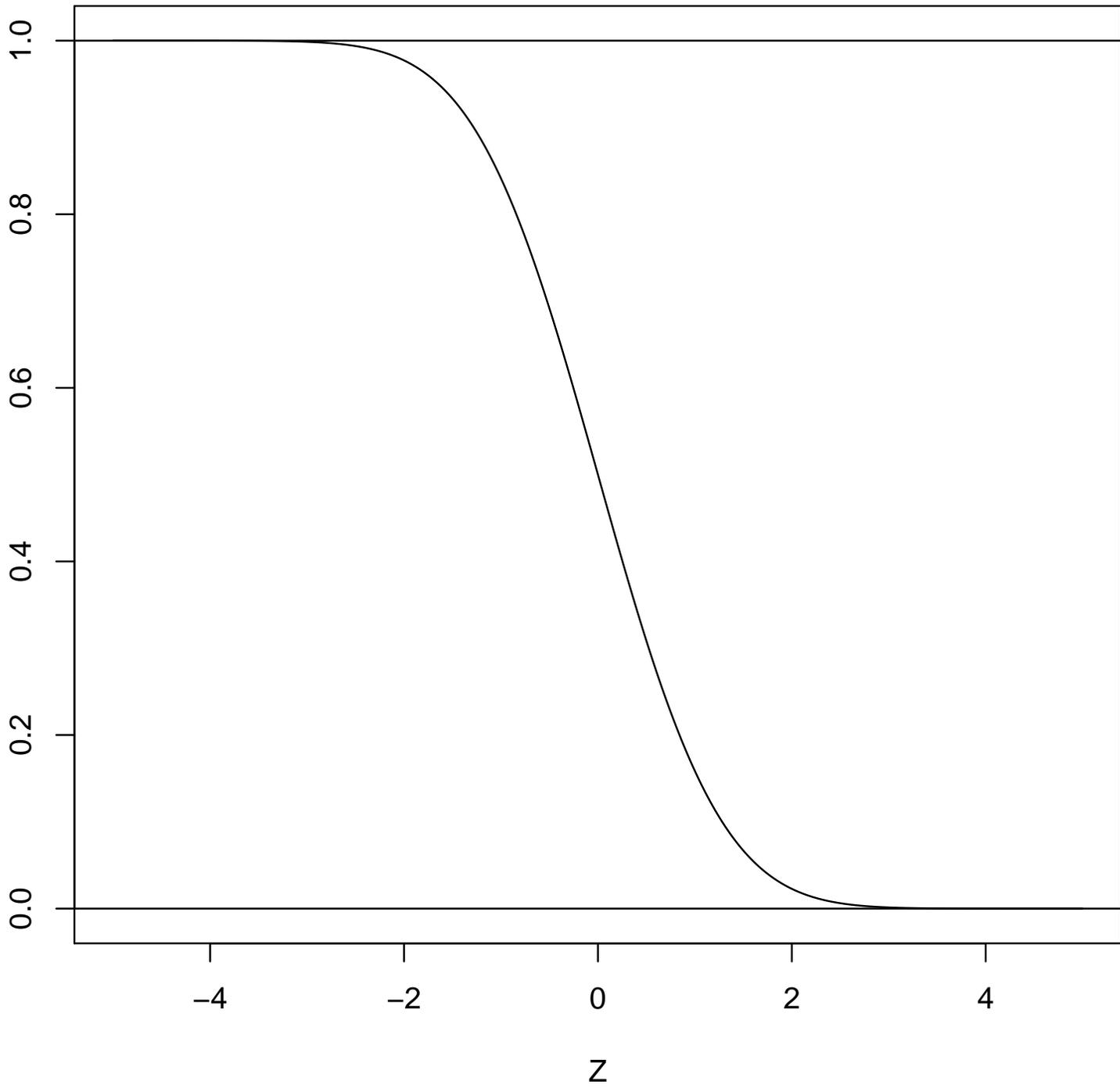


Table 1: The standard normal ( $Z$ ) distribution

First two digits of a.bc	Second digit after decimal(c)									
	0	1	2	3	4	5	6	7	8	9
0.0	0.50000	0.49601	0.49202	0.48803	0.48405	0.48006	0.47608	0.47210	0.46812	0.46414
0.1	0.46017	0.45620	0.45224	0.44828	0.44433	0.44038	0.43644	0.43251	0.42858	0.42465
0.2	0.42074	0.41683	0.41294	0.40905	0.40517	0.40129	0.39743	0.39358	0.38974	0.38591
0.3	0.38209	0.37828	0.37448	0.37070	0.36693	0.36317	0.35942	0.35569	0.35197	0.34827
0.4	0.34458	0.34090	0.33724	0.33360	0.32997	0.32636	0.32276	0.31918	0.31561	0.31207
0.5	0.30854	0.30503	0.30153	0.29806	0.29460	0.29116	0.28774	0.28434	0.28096	0.27760
0.6	0.27425	0.27093	0.26763	0.26435	0.26109	0.25785	0.25463	0.25143	0.24825	0.24510
0.7	0.24196	0.23885	0.23576	0.23270	0.22965	0.22663	0.22363	0.22065	0.21770	0.21476
0.8	0.21186	0.20897	0.20611	0.20327	0.20045	0.19766	0.19489	0.19215	0.18943	0.18673
0.9	0.18406	0.18141	0.17879	0.17619	0.17361	0.17106	0.16853	0.16602	0.16354	0.16109
1.0	0.15866	0.15625	0.15386	0.15151	0.14917	0.14686	0.14457	0.14231	0.14007	0.13786
1.1	0.13567	0.13350	0.13136	0.12924	0.12714	0.12507	0.12302	0.12100	0.11900	0.11702
1.2	0.11507	0.11314	0.11123	0.10935	0.10749	0.10565	0.10383	0.10204	0.10027	0.09853
1.3	0.09680	0.09510	0.09342	0.09176	0.09012	0.08851	0.08691	0.08534	0.08379	0.08226
1.4	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07215	0.07078	0.06944	0.06811
1.5	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592
1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551
1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673
1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938
1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330
2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
2.5	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494	0.00480
2.6	0.00466	0.00453	0.00440	0.00427	0.00415	0.00402	0.00391	0.00379	0.00368	0.00357
2.7	0.00347	0.00336	0.00326	0.00317	0.00307	0.00298	0.00289	0.00280	0.00272	0.00264
2.8	0.00256	0.00248	0.00240	0.00233	0.00226	0.00219	0.00212	0.00205	0.00199	0.00193
2.9	0.00187	0.00181	0.00175	0.00169	0.00164	0.00159	0.00154	0.00149	0.00144	0.00139
3.0	0.00135	0.00131	0.00126	0.00122	0.00118	0.00114	0.00111	0.00107	0.00104	0.00100
3.1	0.00097	0.00094	0.00090	0.00087	0.00084	0.00082	0.00079	0.00076	0.00074	0.00071
3.2	0.00069	0.00066	0.00064	0.00062	0.00060	0.00058	0.00056	0.00054	0.00052	0.00050
3.3	0.00048	0.00047	0.00045	0.00043	0.00042	0.00040	0.00039	0.00038	0.00036	0.00035
3.4	0.00034	0.00032	0.00031	0.00030	0.00029	0.00028	0.00027	0.00026	0.00025	0.00024
3.5	0.00023	0.00022	0.00022	0.00021	0.00020	0.00019	0.00019	0.00018	0.00017	0.00017
3.6	0.00016	0.00015	0.00015	0.00014	0.00014	0.00013	0.00013	0.00012	0.00012	0.00011
3.7	0.00011	0.00010	0.00010	0.00010	0.00009	0.00009	0.00008	0.00008	0.00008	0.00008
3.8	0.00007	0.00007	0.00007	0.00006	0.00006	0.00006	0.00006	0.00005	0.00005	0.00005
3.9	0.00005	0.00005	0.00004	0.00004	0.00004	0.00004	0.00004	0.00004	0.00003	0.00003
4.0	0.00003	0.00003	0.00003	0.00003	0.00003	0.00003	0.00002	0.00002	0.00002	0.00002

Has fish size changed in a salmon hatchery?

Consider a case in which very extensive sampling (huge  $n$ ) in 1985 gives us essentially error-free estimates of the fish sizes.

In 2009 we measure a hatchling and a 3-week old fish.

$H_0$ : Hatchling size is described by a normal distribution, and the population mean and variance of hatchling size has not changed since 1985.

$H_A$ : otherwise.

$H_0$ :  $\mu_{h2009} = \mu_{h1985}$  and  $\sigma_{h2009} = \sigma_{h1985}$

$H_A$ :  $\mu_{h2009} \neq \mu_{h1985}$  or  $\sigma_{h2009} \neq \sigma_{h1985}$

Mass of Salmon hatchlings:

$$\mu_{h1985} = 0.265\text{g}$$

$$\sigma_{h1985} = 0.02\text{g}$$

$$Y_{h2009} = 0.190\text{g}$$

Hypothetical mass at 3 weeks:

$$\mu_{1985} = 0.625\text{g}$$

$$\sigma_{1985} = 0.0456\text{g}$$

$$Y_{2009} = 0.7\text{g}$$

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

$$Z_{h2009} = \frac{0.190\text{g} - 0.265\text{g}}{0.02\text{g}} = -3.75$$

$$Z_{2009} = \frac{0.7\text{g} - 0.625\text{g}}{0.0456\text{g}} = 1.64$$

A  $Z$  statistic is a measure of how far your observation differs from the expected value, measured in units of “number of standard deviations”

$Z_{h2009} = -3.75$  means

Our hatchling mass was 3.75 standard deviations smaller than the expected value.

$Z_{2009} = 1.64$  means

The mass at three weeks was 1.64 standard deviations larger than the expected value.

Hatchling: The mass of the single hatching (0.19g) was significantly smaller than expected. We can reject the hypothesis that the mass of hatchlings is normally distributed with a mean of 0.265g and standard deviation of 0.02g (which were the descriptive statistics from the well-studied 1985 fishery).  $Z = -3.75$  and  $P\text{-value} \approx 1.8 \times 10^{-4}$  for our test.

3-week old: The mass of the single fish we measured at 3-weeks (0.7g) was *not* significantly different than expected. We cannot reject the hypothesis that the mass of 3-week old fish is normally distributed with a mean of 0.625g and standard deviation of 0.0456g (which were the descriptive statistics from the well-studied 1985 fishery).  $Z = 1.64$  and our  $P\text{-value} \approx .101$  for the test.

Z-test assumptions:

1. random sample

2. null hypothesis:

(a) a normal distribution,

(b) with a specific mean, and,

(c) with a specific standard deviation.

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

## **The Central Limit Theorem:**

if you add a *sufficiently large* number of independent variables, then the sum will be Normally distributed.

If the some of the terms have a much larger magnitude than others, then “sufficiently large” will mean “very large.”

British spies must be  $< 180.3\text{cm}$  tall.

For men,  $h \sim \text{Normal}(\mu = 177.0, \sigma = 7.1)$

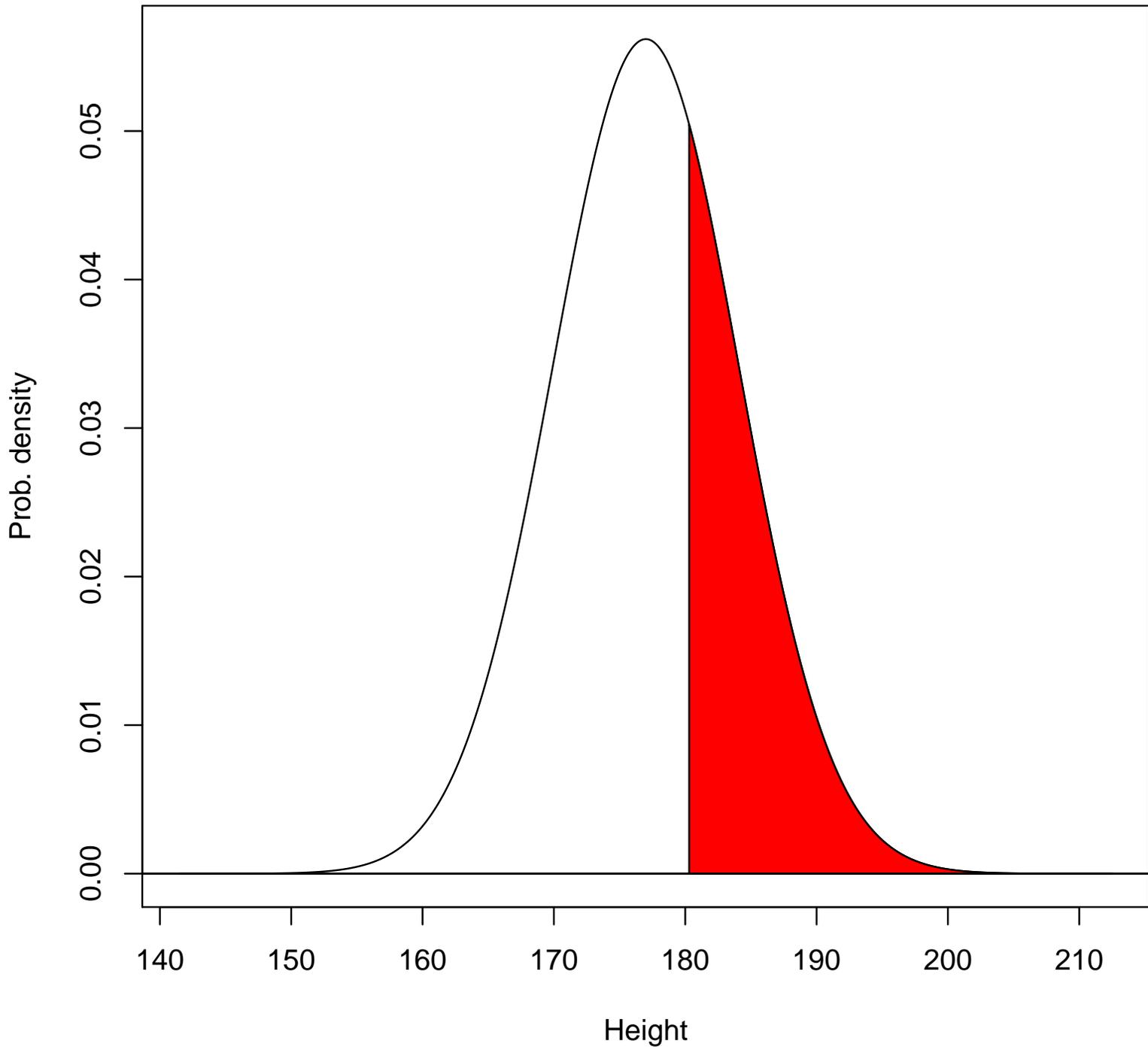
What proportion of men are excluded by this criterion?

$$\begin{aligned}\Pr(\text{excluded}) &= \Pr(h > 180.3) \\ &= \Pr\left(Z > \frac{180.3 - 177.0}{7.1}\right) \\ &= \Pr(Z > 0.4647) \\ &\approx 0.32276\end{aligned}$$

Where the last step comes from looking up  $Z = 0.46$  in the table B in the back of your text. That is on row 0.4 and column 6 of the table.

If we had asked “What proportion of men are allowed to be spies according to this criterion” then we would calculate:

$$\begin{aligned}\Pr(\text{included}) &= 1 - \Pr(\text{excluded}) \\ &\approx 1 - .032276 \\ &= 0.67724\end{aligned}$$



Height in British men is described as:

$$h \sim \text{Normal}(\mu = 177.0, \sigma = 7.1)$$

What is the probability of a British man having a height between 172 cm and 180.3cm?

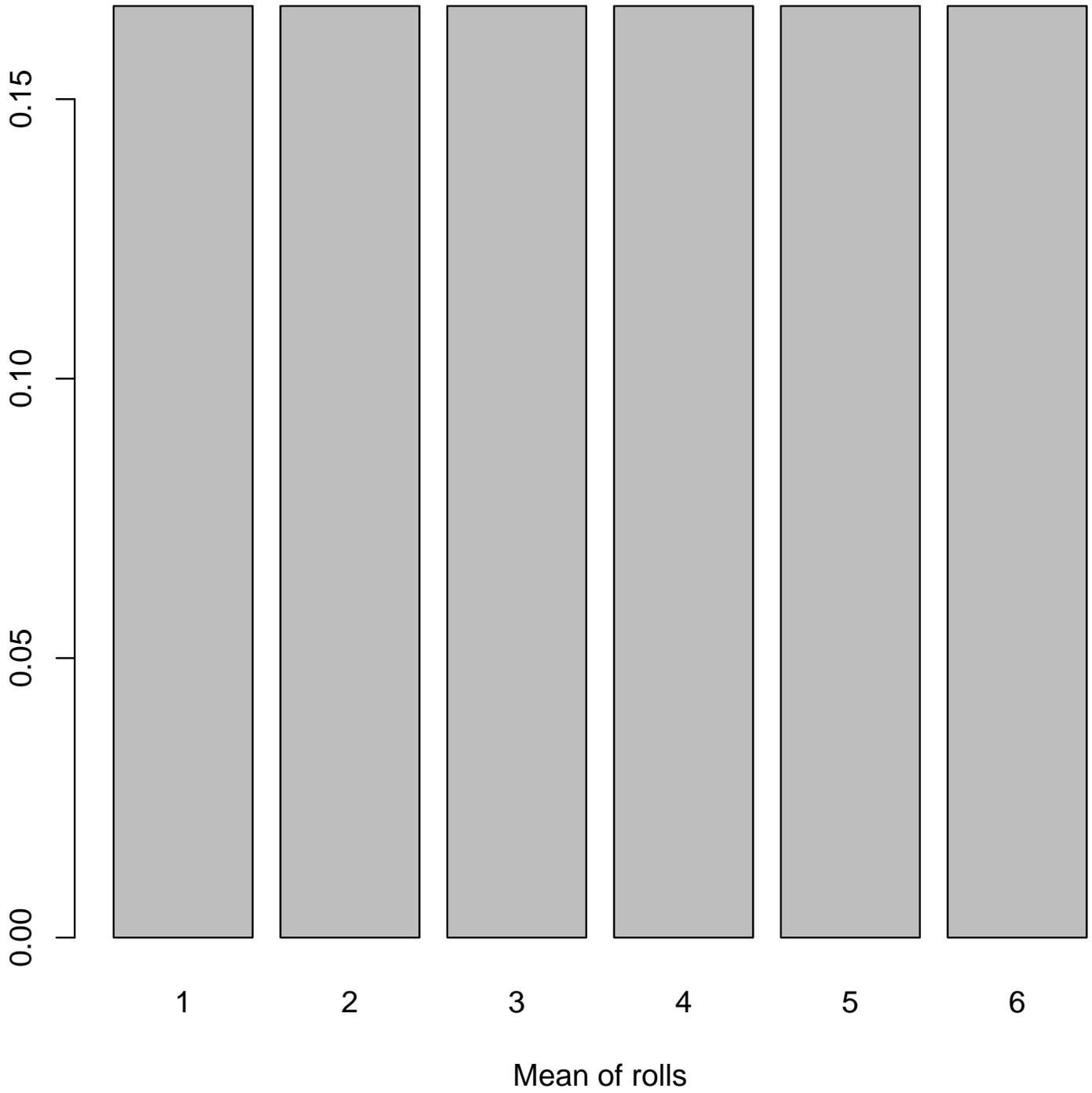
## The Central Limit Theorem:

if you add a *sufficiently large* number of independent variables, then the sum will be Normally distributed.

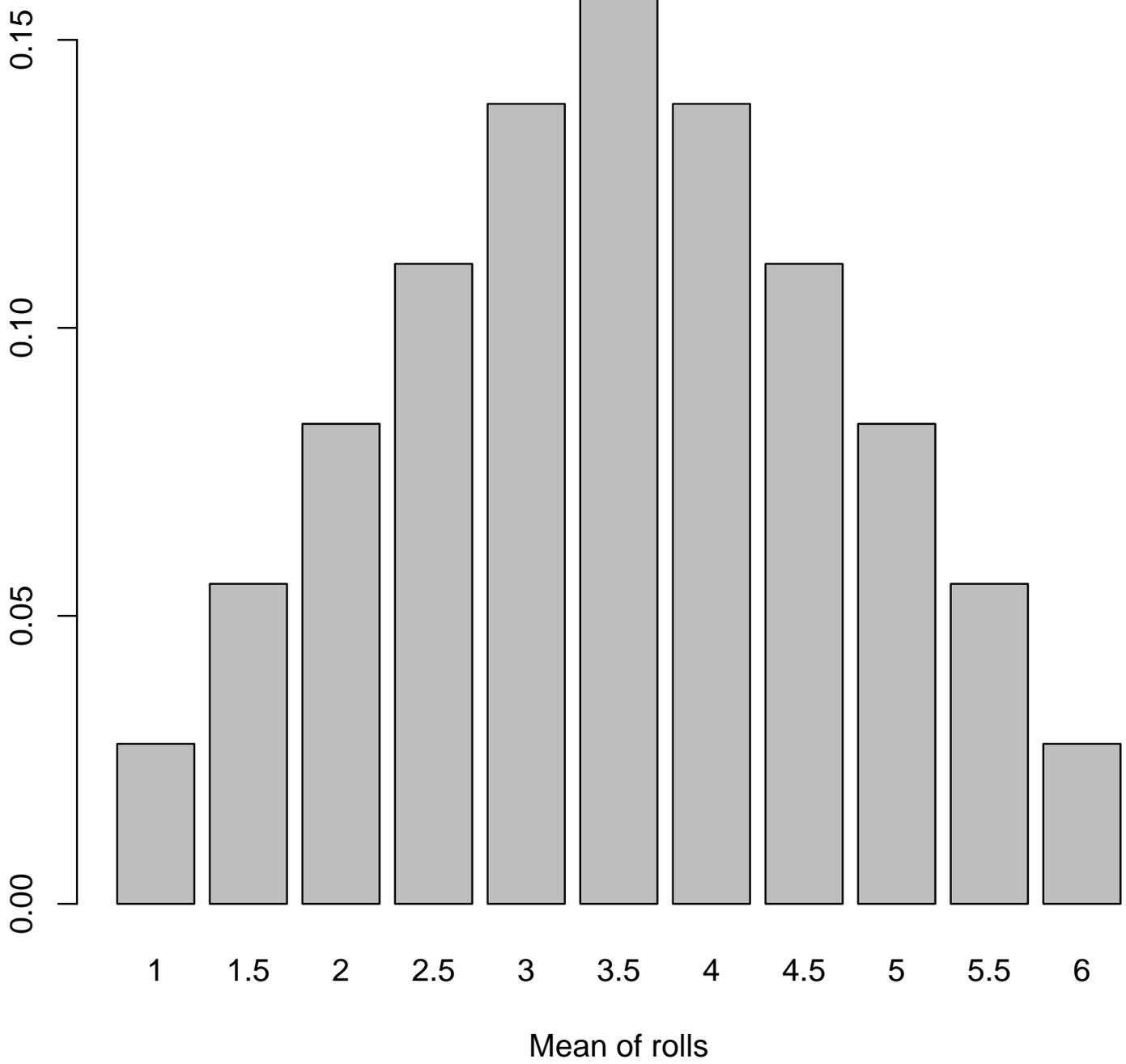
If the some of the terms have a much larger magnitude than others, then “sufficiently large” will mean “very large.”

Even if we cannot assume Normal we can often solve problems by treating the distribution as a **mixture of Normals** (because the total is the sum of a few large effects and many small ones).

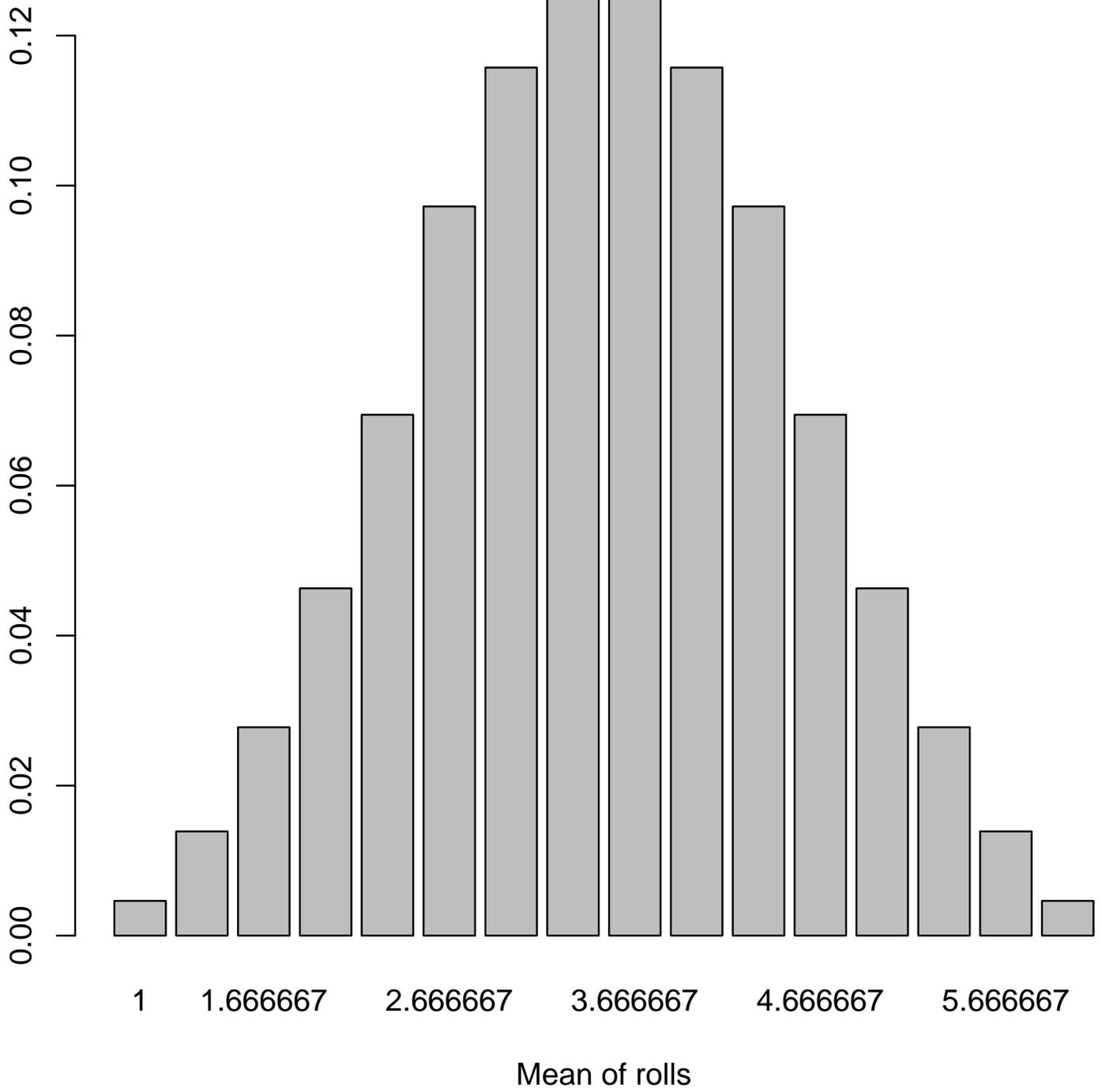
### Distribution of the mean of 1 rolls



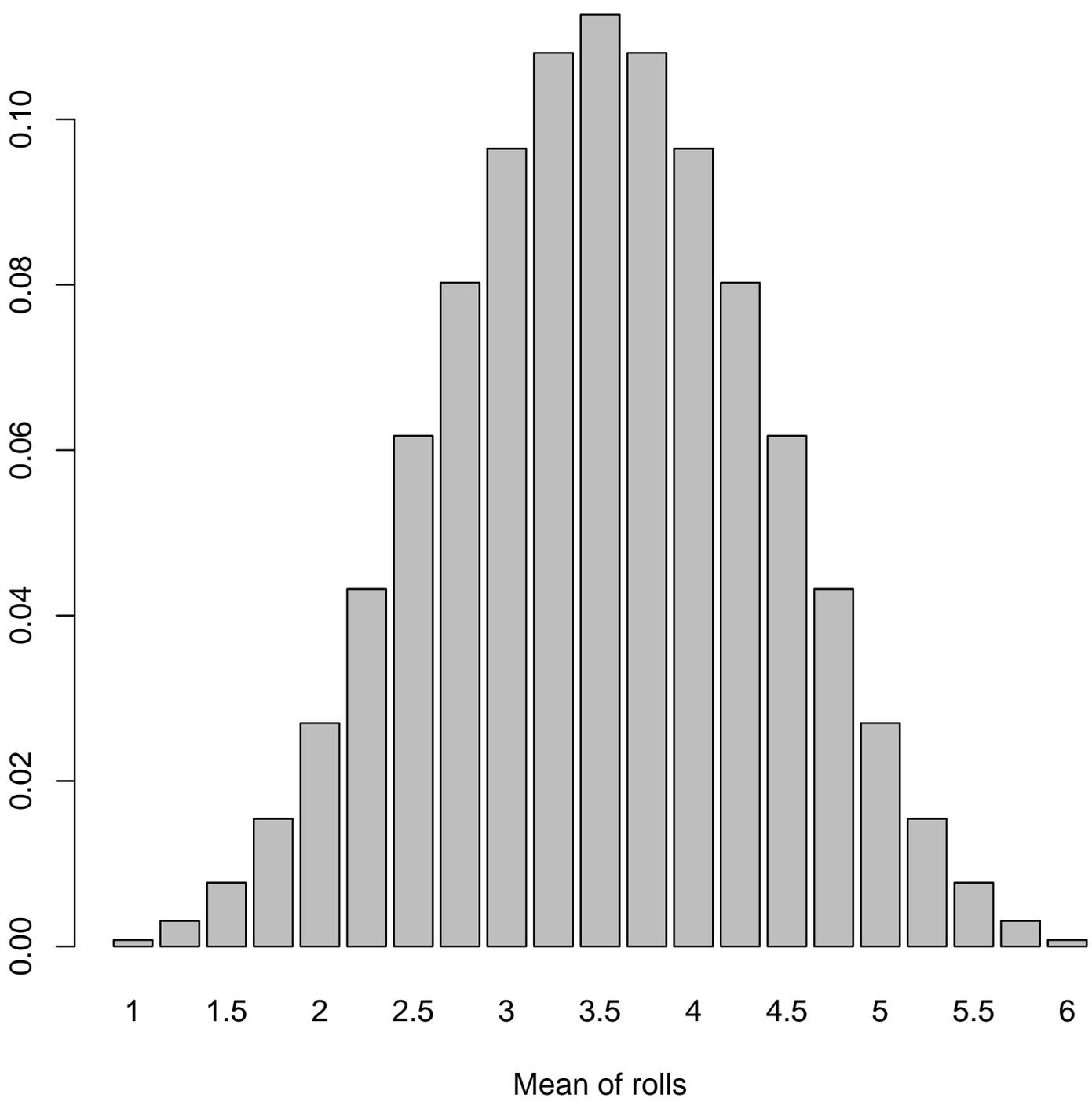
### Distribution of the mean of 2 rolls



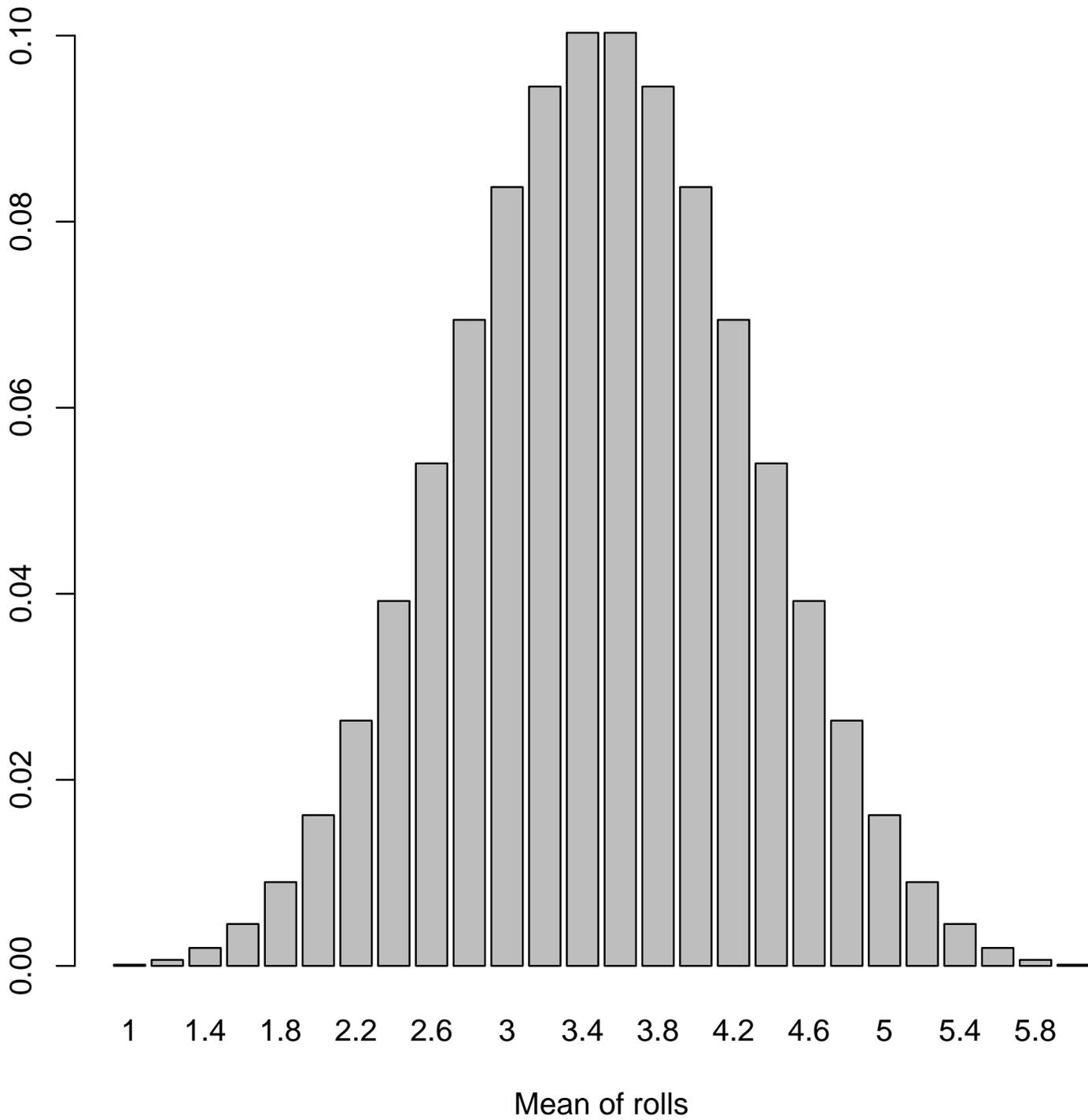
### Distribution of the mean of 3 rolls



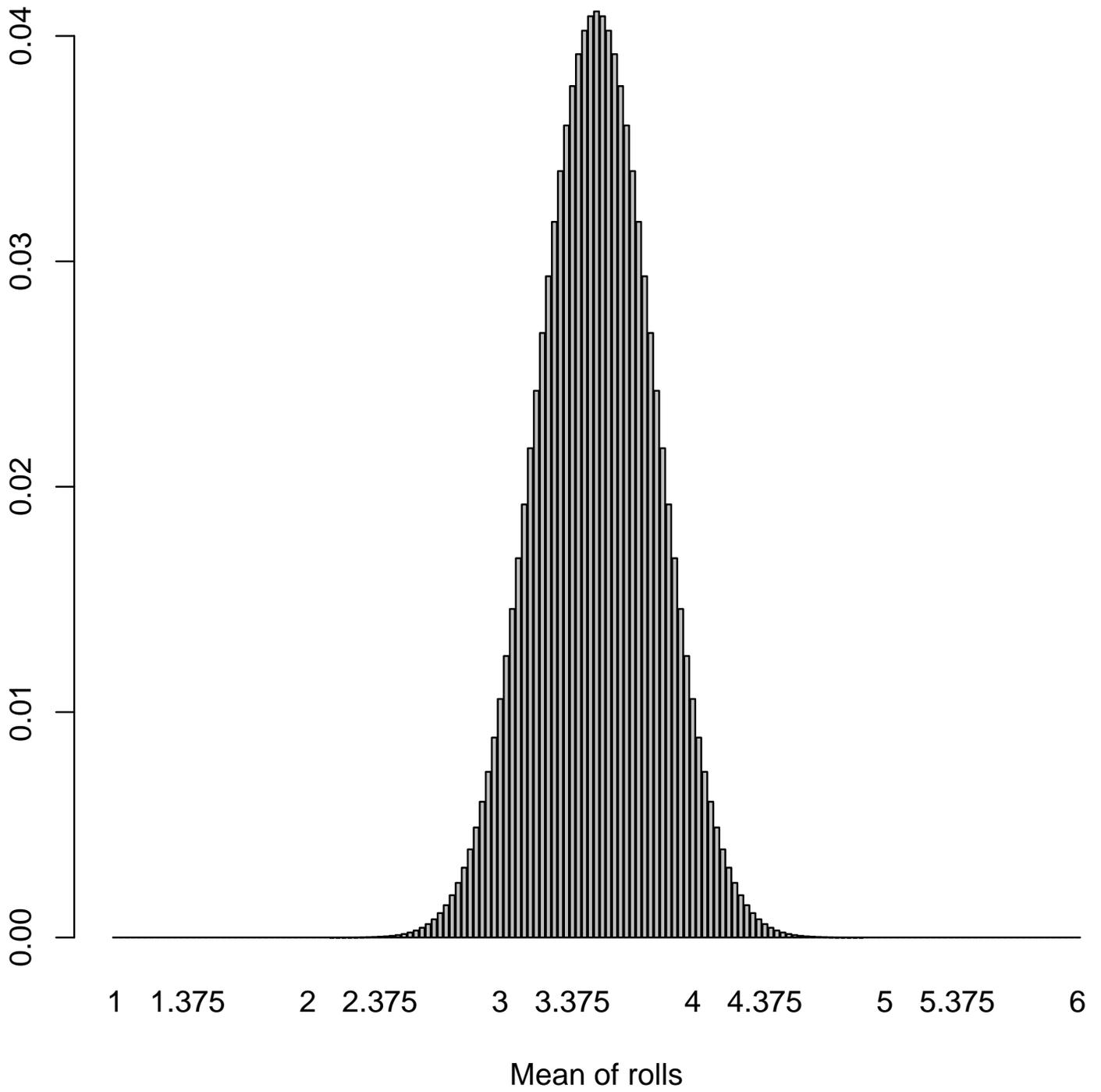
### Distribution of the mean of 4 rolls



### Distribution of the mean of 5 rolls

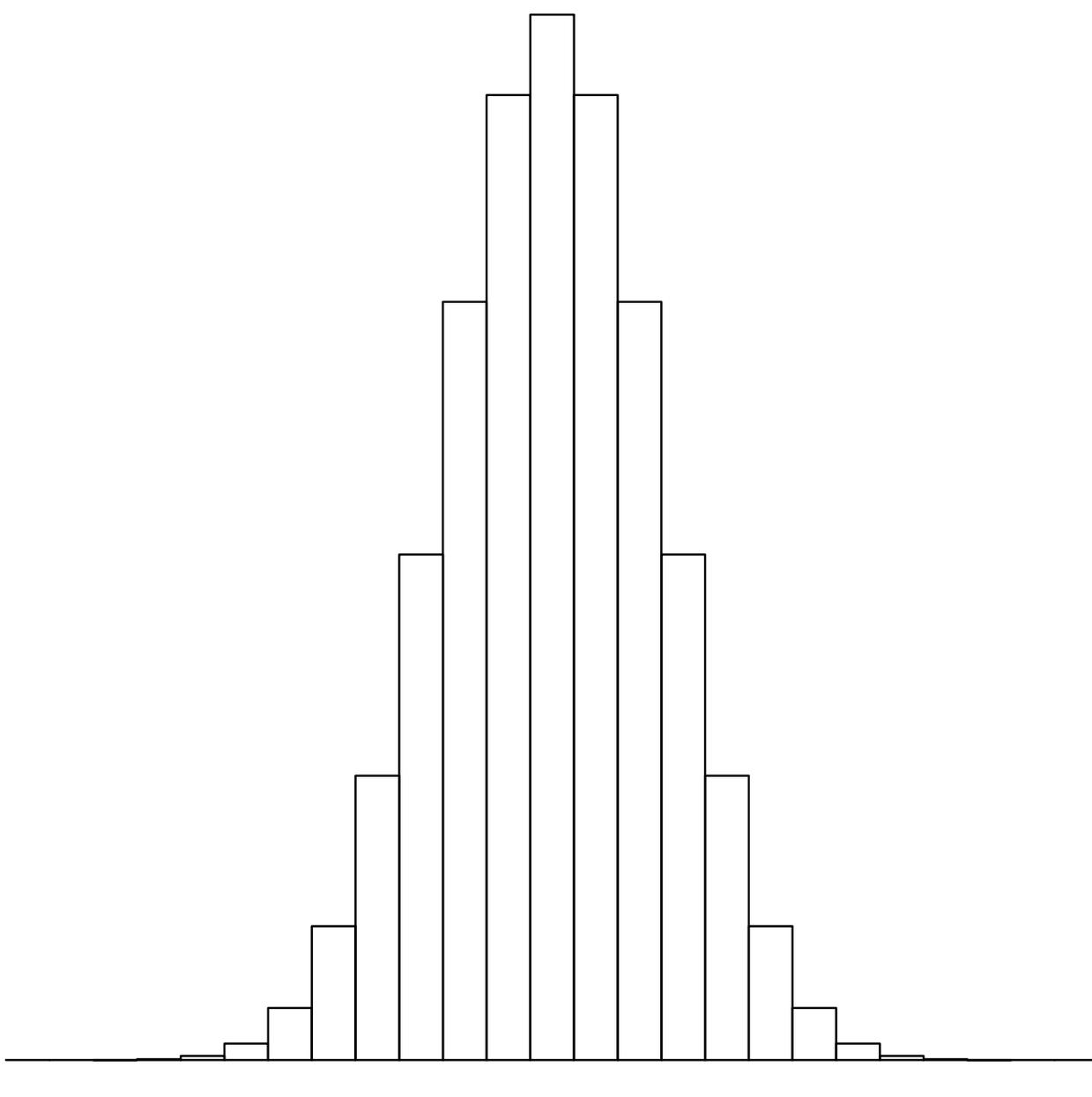


# Distribution of the mean of 32 rolls

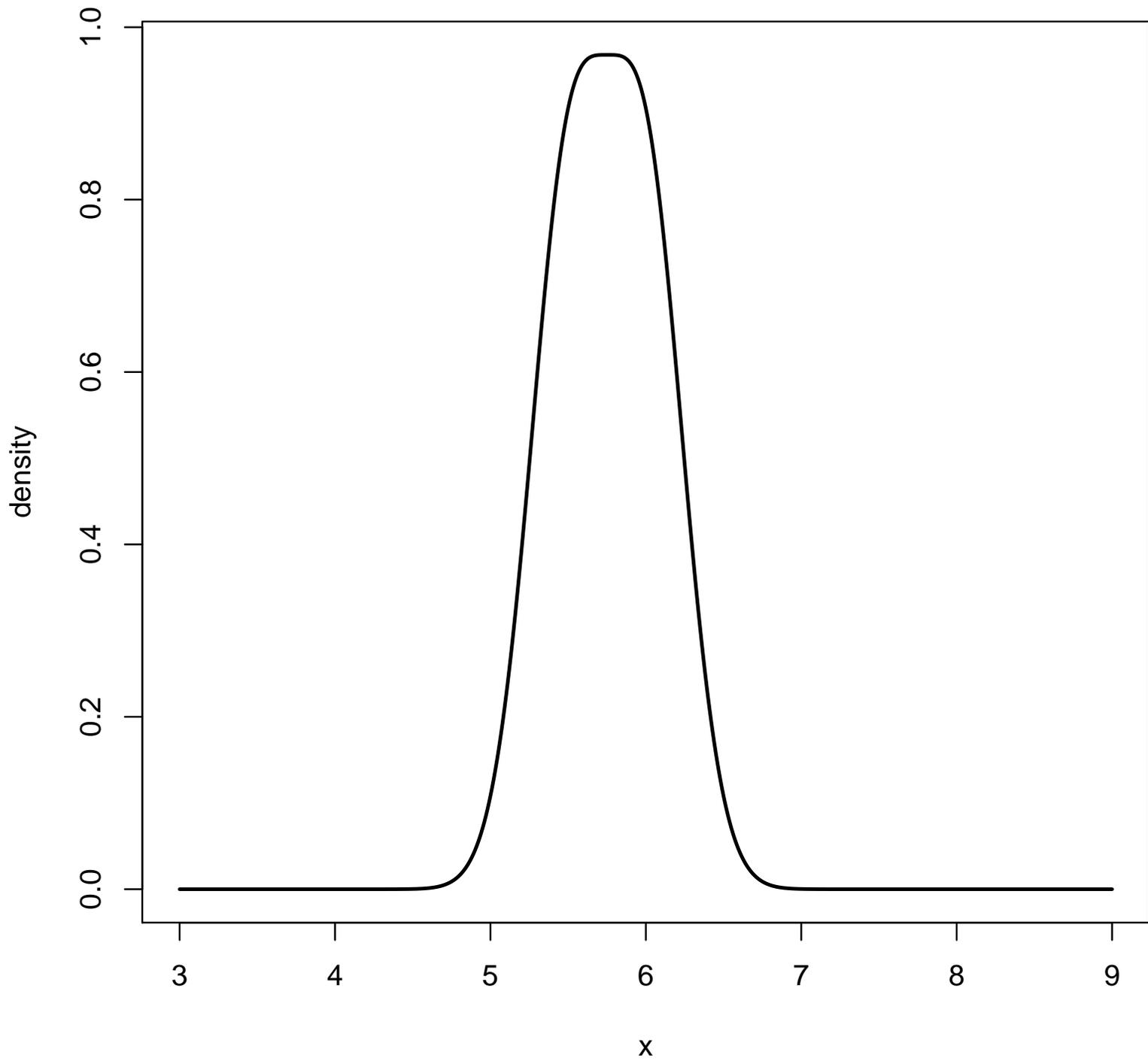


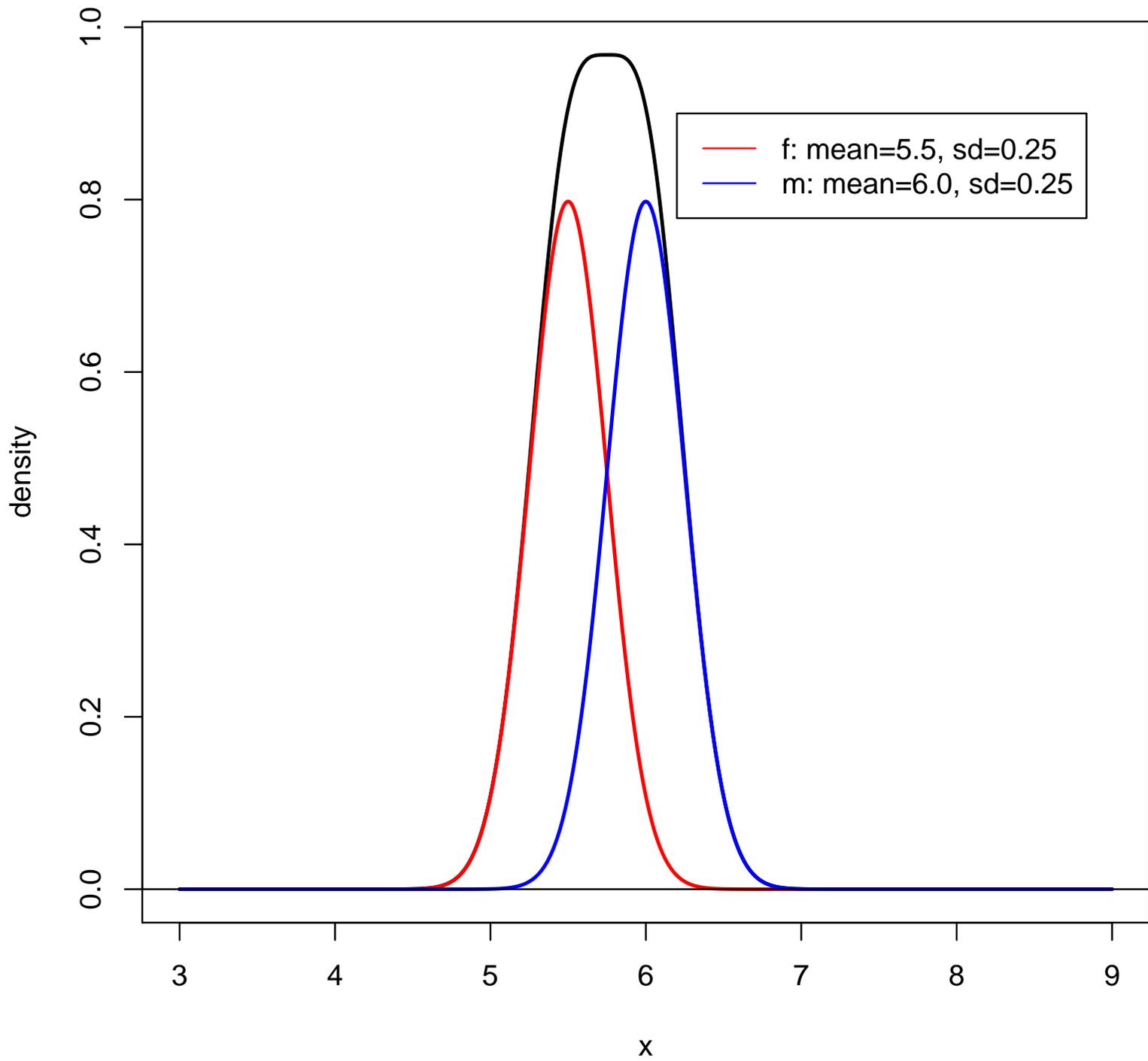
expected relative frequency

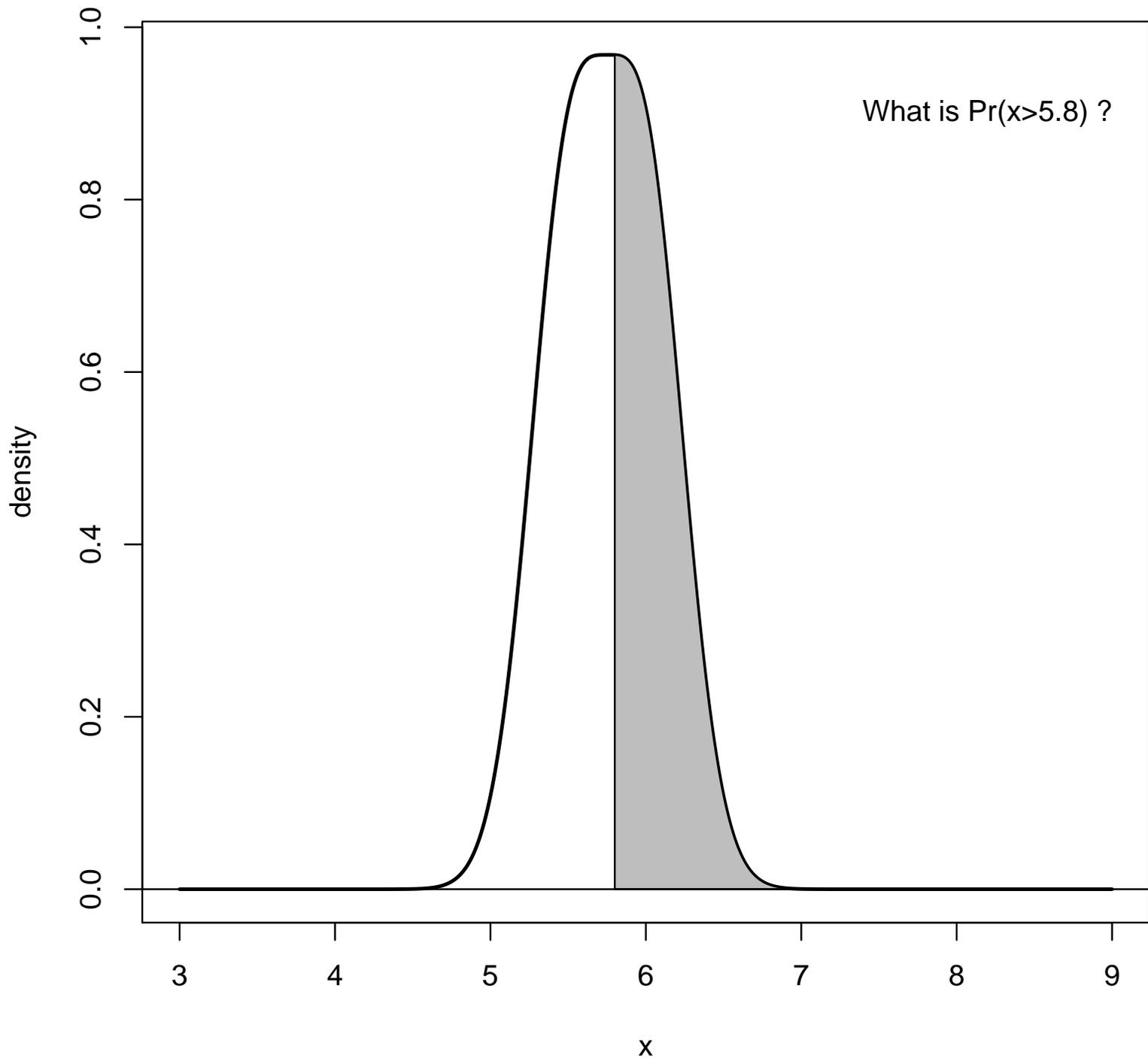
0.15  
0.10  
0.05  
0.00

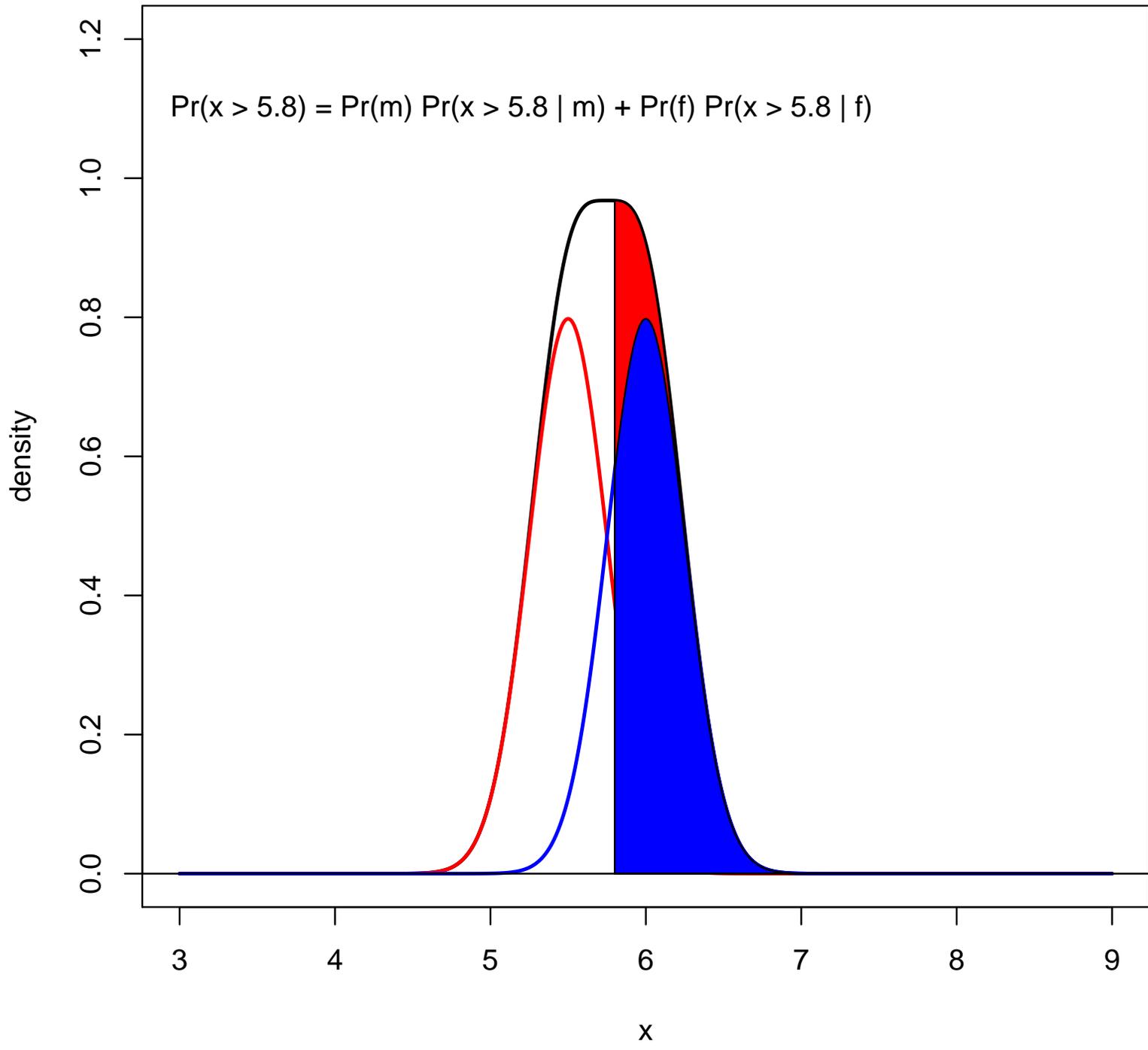


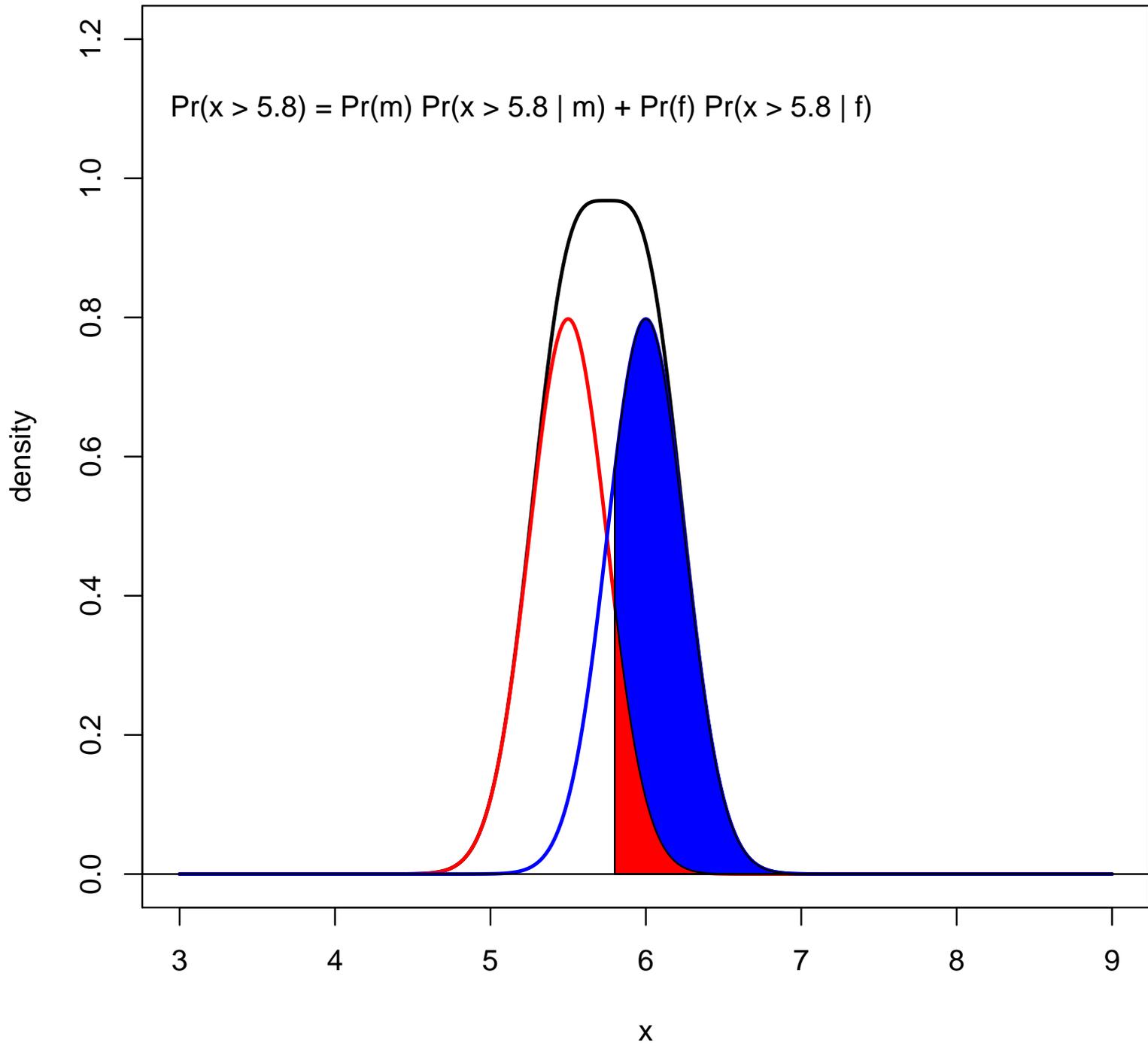
estimate of mean # of anthers/individual (N=24)











From chapter 4:

$$\bar{Y} \sim \text{Normal} \left( \mu, \frac{\sigma}{\sqrt{n}} \right)$$

$$\mu_{1985} = 0.625\text{g}$$

$$\sigma_{1985} = 0.0456\text{g}$$

$$\bar{Y}_{2009} = 0.7\text{g}$$

What if  $n = 9$  for the 2009 study?

$$Z = \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

For large  $n$  and  $p$  not close to 0 or 1:

Binomial with  $n$  and  $p$  can be approximated by:

Normal( $\mu = np, \sigma = \sqrt{np(1 - p)}$ )

*Note:* Your book presents a more accurate version of using the Normal to approximate the binomial on page 291 (in the paragraph starting with “However, we can improve...”). We are not going to worry about the more accurate “continuity correction.”

People between 60 and 64 with certain risk factors have a 16% chance of having a heart attack within a time window (this is known prior to the study). A drug company is trying out a new treatment. Only 15 of their 300 trial subjects using the drug have a heart attack in the given time window. Does the drug alter the probability of a heart attack?

$$\bar{Y} = \left(\frac{1}{n}\right) \sum_i^n Y_i$$

$$SE_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

$$\bar{Y} \sim \text{Normal} \left( \mu, \frac{\sigma}{\sqrt{n}} \right)$$

Sampling distribution of  $\hat{p}$ :

$$\hat{p} = \left(\frac{1}{n}\right) \sum_i^n \begin{cases} 1 & \text{if trial } i = \text{success} \\ 0 & \text{if trial } i = \text{failure} \end{cases}$$

$$\hat{p} = \left(\frac{1}{n}\right) \sum_i^n I(Y_i = \text{success})$$

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

$$\hat{p} \sim \text{Normal} \left( p, \sqrt{\frac{p(1-p)}{n}} \right)$$

Where does  $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$  come from?

$$SE_{\hat{p}} = \mathbb{E} [\text{std. dev. of } \hat{p} \text{ from many studies}]$$

$$\sigma^2 = \mathbb{E} [\text{squared deviations}]$$

$$\text{success} \rightarrow \text{deviation} = 1 - p$$

$$\text{failure} \rightarrow \text{deviation} = 0 - p = -p$$

$$\mathbb{E} [s^2] = p(1-p)^2 + (1-p)(-p)^2$$

$$= p(1-p)^2 + (1-p)(p)^2$$

$$= p(1-p)(1-p+p)$$

$$= p(1-p)$$

$$\sigma = \sqrt{\mathbb{E} [s^2]} = \sqrt{p(1-p)}$$

For a sample of size  $n$ :

$$SE_{\hat{p}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}$$

Note: When analyzing proportion data we use the closely related Agresti-Coull formulae, not the formula above.

Where does  $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$  come from?

$$\sigma^2 = \mathbb{E} [\text{squared deviations}]$$

$$\text{success} \rightarrow \text{deviation} = 1 - p$$

$$\text{failure} \rightarrow \text{deviation} = 0 - p = -p$$

$$\begin{aligned}\mathbb{E} [s^2] &= \Pr(\text{success})(1 - p)^2 + \Pr(\text{fail})(1 - p)(-p)^2 \\ &= p(1 - p)^2 + (1 - p)(-p)^2\end{aligned}$$

$$= p(1 - p)^2 + (1 - p)(p)^2$$

$$= p(1 - p)(1 - p + p)$$

$$= p(1 - p)$$

$$\sigma = \sqrt{\mathbb{E} [s^2]} = \sqrt{p(1 - p)}$$

For a sample of size  $n$ :

$$SE_{\hat{p}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1 - p)}{n}}$$

$$\bar{Y} \sim \text{Normal} \left( \mu, \frac{\sigma}{\sqrt{n}} \right)$$

$$\hat{p} \sim \text{Normal} \left( p, \sqrt{\frac{p(1-p)}{n}} \right)$$

What about the sampling distribution of  $X$ , where  $X$  is the number of successes?

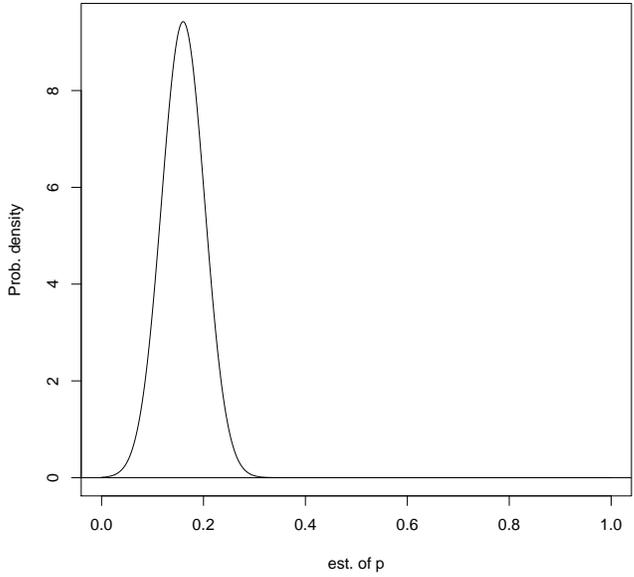
$X$  is the just  $n\hat{p}$  because  $\hat{p} = \frac{X}{n}$ .

Remember if we multiply a distribution of values by  $c$  then the mean and s.d. are scaled by a factor of  $c$ .

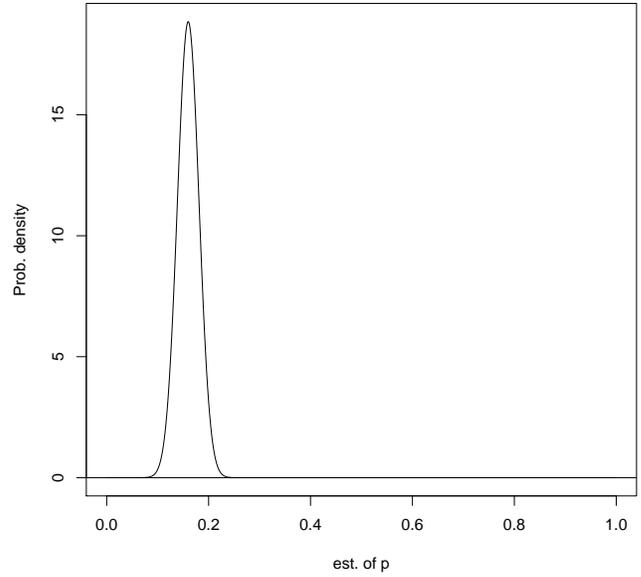
$$X \sim \text{Normal} \left( np, n\sqrt{\frac{p(1-p)}{n}} \right)$$

$$X \sim \text{Normal} \left( np, \sqrt{np(1-p)} \right)$$

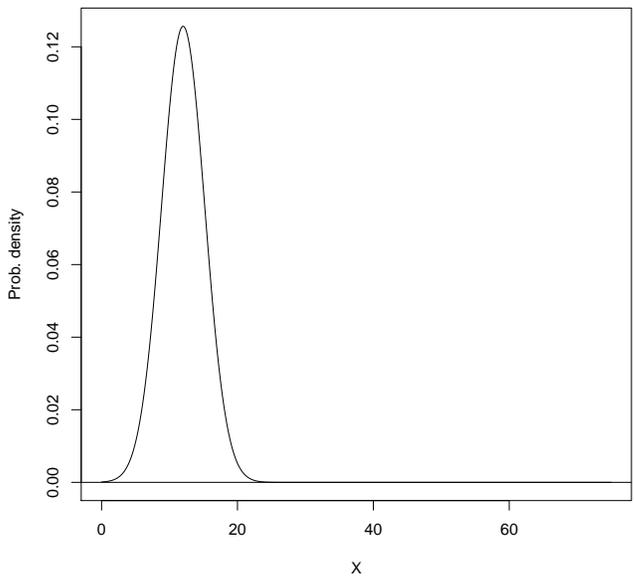
**Samp. dist. of  $\hat{p}$ : N=75**



**Samp. dist. of  $\hat{p}$ : N=300**



**Samp. dist. of X: N=75**



**Samp. dist. of X: N=300**

