

NORMAL DISTRIBUTION



PARANORMAL DISTRIBUTION

Lecture 05

10.6.2021

Refresher Quiz

Wavy hair in mice is a recessive genetic trait. If mice with wavy hair are mated with straight haired (heterozygous) mice, each offspring has a probability of 0.5 of having wavy hair. Consider a large number of such matings, each producing a litter of 5 offspring. What percentage of the litters will consist of two wavy-haired and three straight-haired offspring?

Wavy hair in mice is a recessive genetic trait. If mice with wavy hair are mated with straight haired (heterozygous) mice, each offspring has a probability of 0.5 of having wavy hair. Consider a large number of such matings, each producing a litter of 5 offspring. What percentage of the litters will consist of two wavy-haired and three straight-haired offspring?

Wavy or straight

B - I - n - S

n = 5

P(wavy) = 0.5

Binary - Independent - n - Same p

Wavy hair in mice is a recessive genetic trait. If mice with wavy hair are mated with straight haired (heterozygous) mice, each offspring has a probability of 0.5 of having wavy hair. Consider a large number of such matings, each producing a litter of 5 offspring. What **percentage of the litters will consist of two wavy-haired and three straight-haired offspring?**

Prob = relative frequency

$${}_n C_j p^j (1 - p)^{n-j}$$

Wavy or straight

n = 5

P(wavy) = 0.5


B - Y - n - S

Binary - Independent - n - Same p

Wavy hair in mice is a recessive genetic trait. If mice with wavy hair are mated with straight haired (heterozygous) mice, each offspring has a probability of 0.5 of having wavy hair. Consider a large number of such matings, each producing a litter of 5 offspring. What **percentage of the litters will consist of two wavy-haired and three straight-haired offspring?**

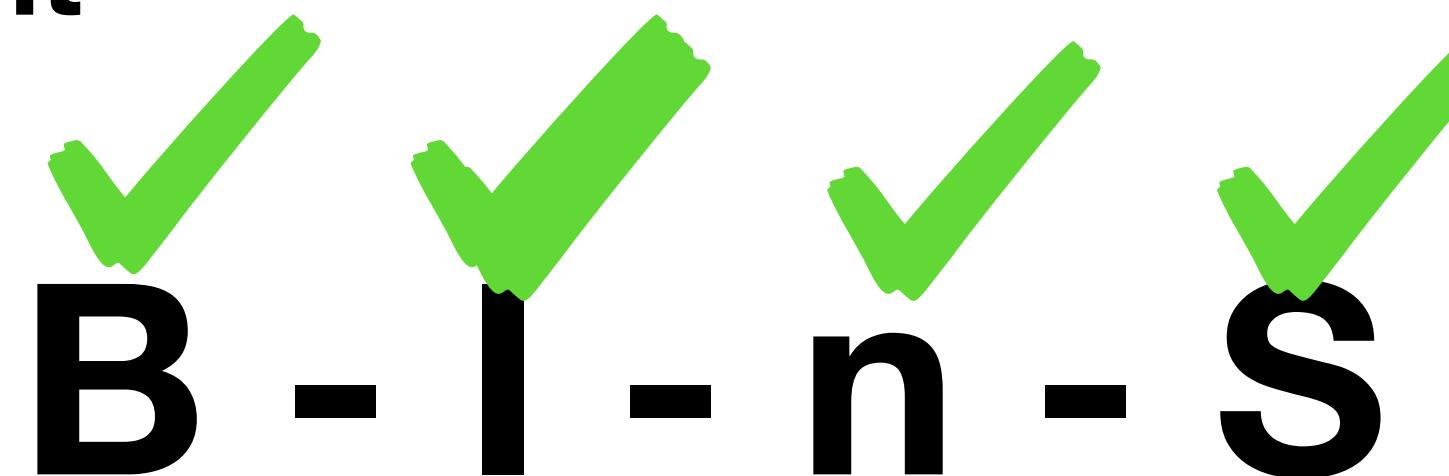
Prob = relative frequency

$$5C_2(0.5)^2(0.5)^3 = 0.3125$$

Wavy or straight

n = 5

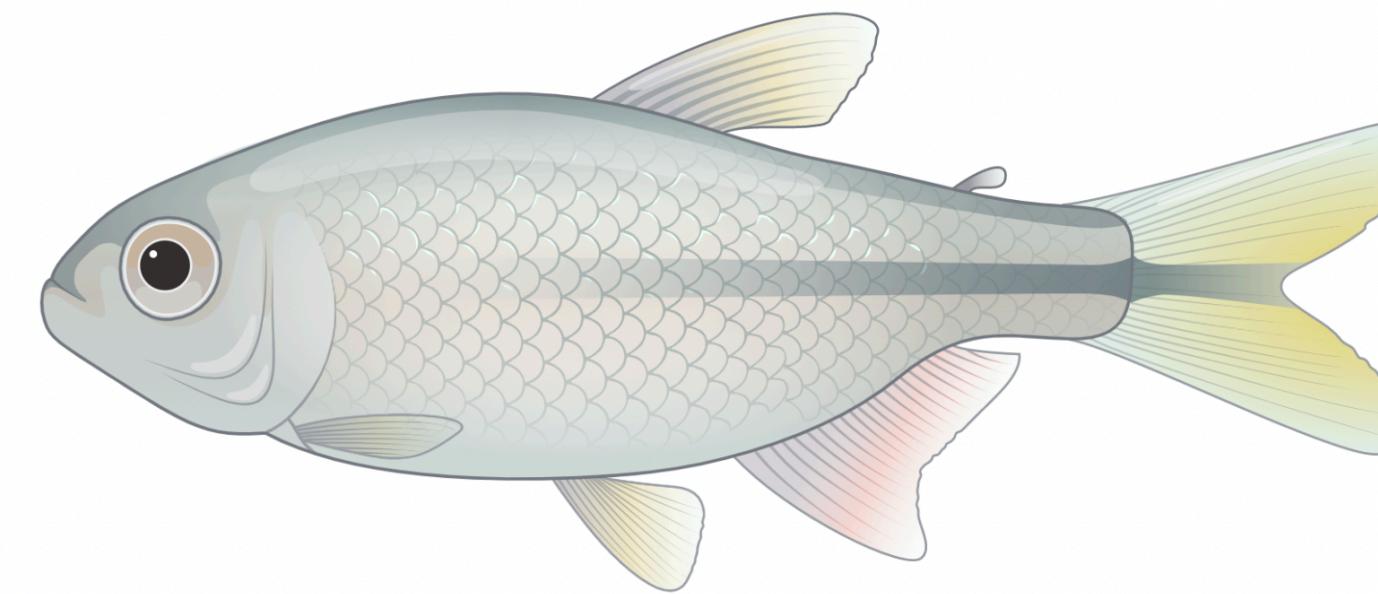
P(wavy) = 0.5


B - Y - n - S

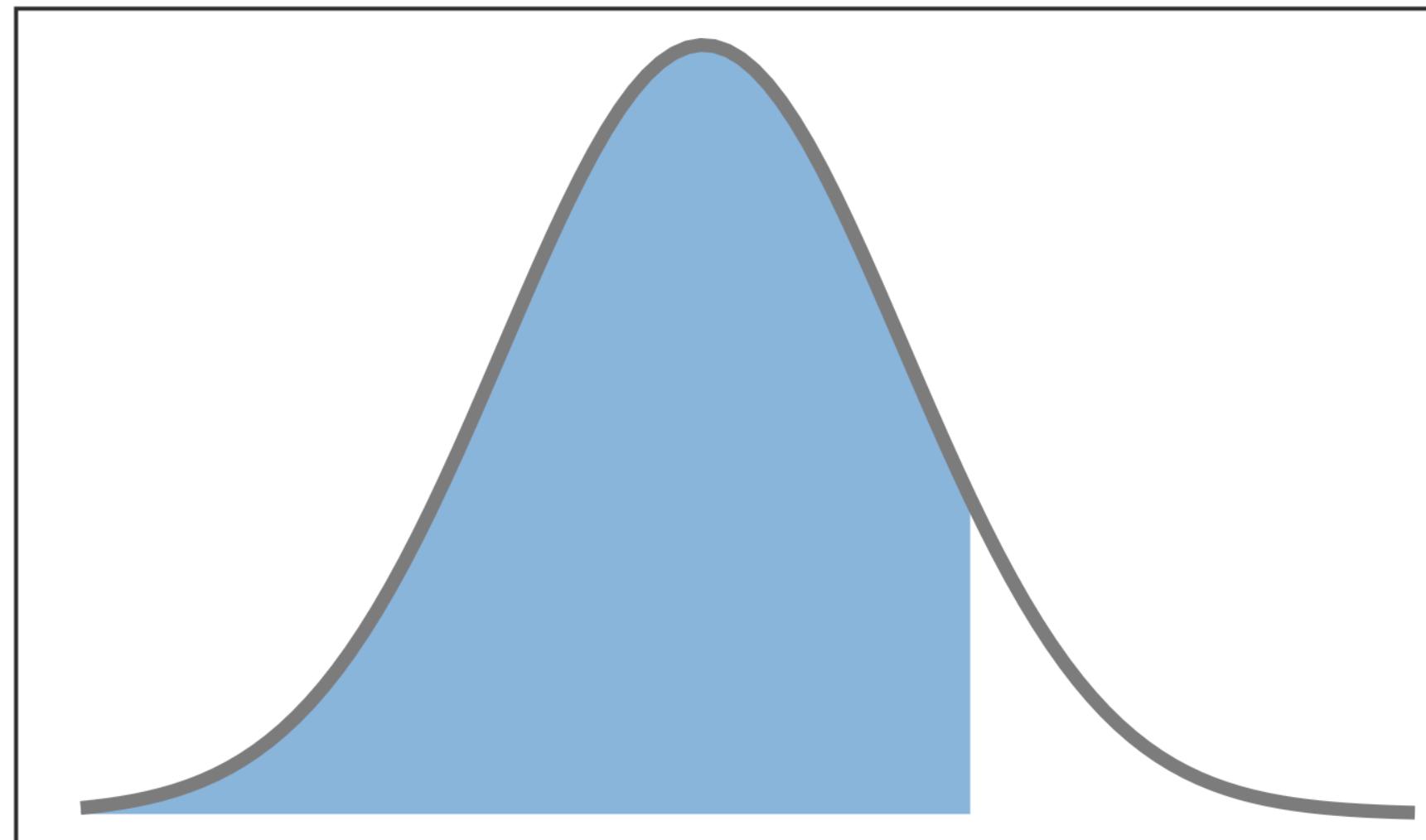
Binary - Independent - n - Same p

In a certain population of fish, individual lengths follow a normal distribution. The mean length of the fish is 54.0 mm, and the standard deviation is 4.5 mm. What percentage of the fish are shorter than 60 mm?

*Even if you use R to calculate p-value,
ALWAYS smart to make
a quick sketch*

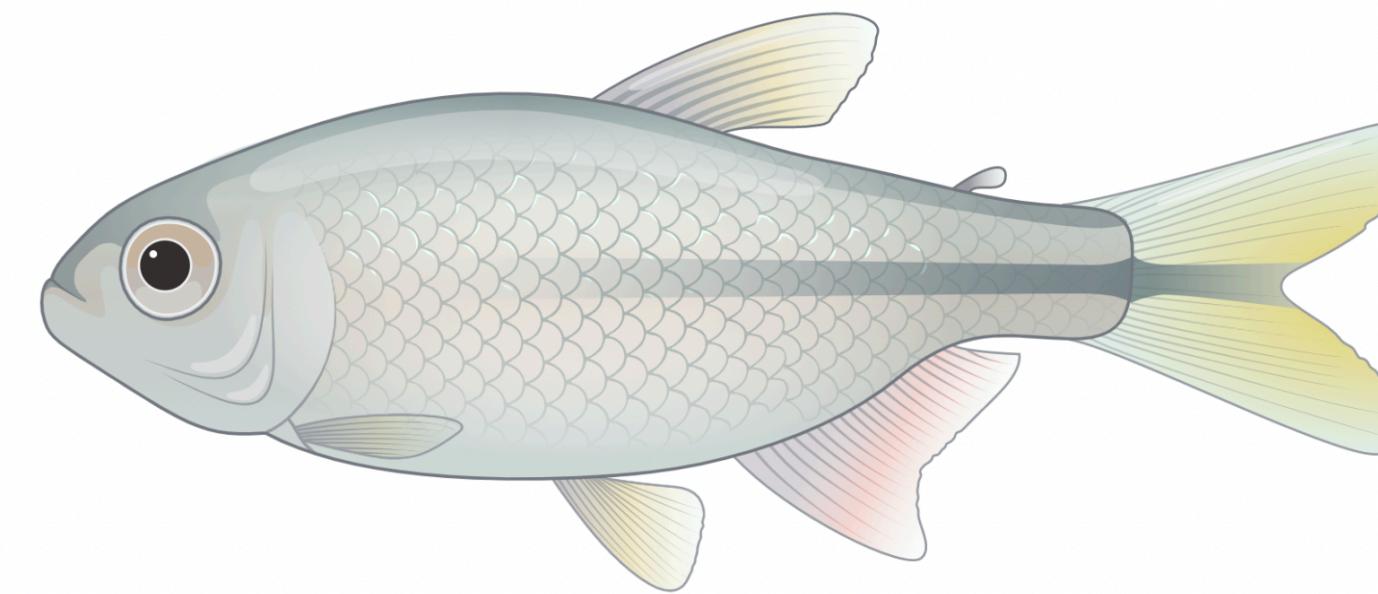


$$Z = \frac{y - \mu}{\sigma}$$

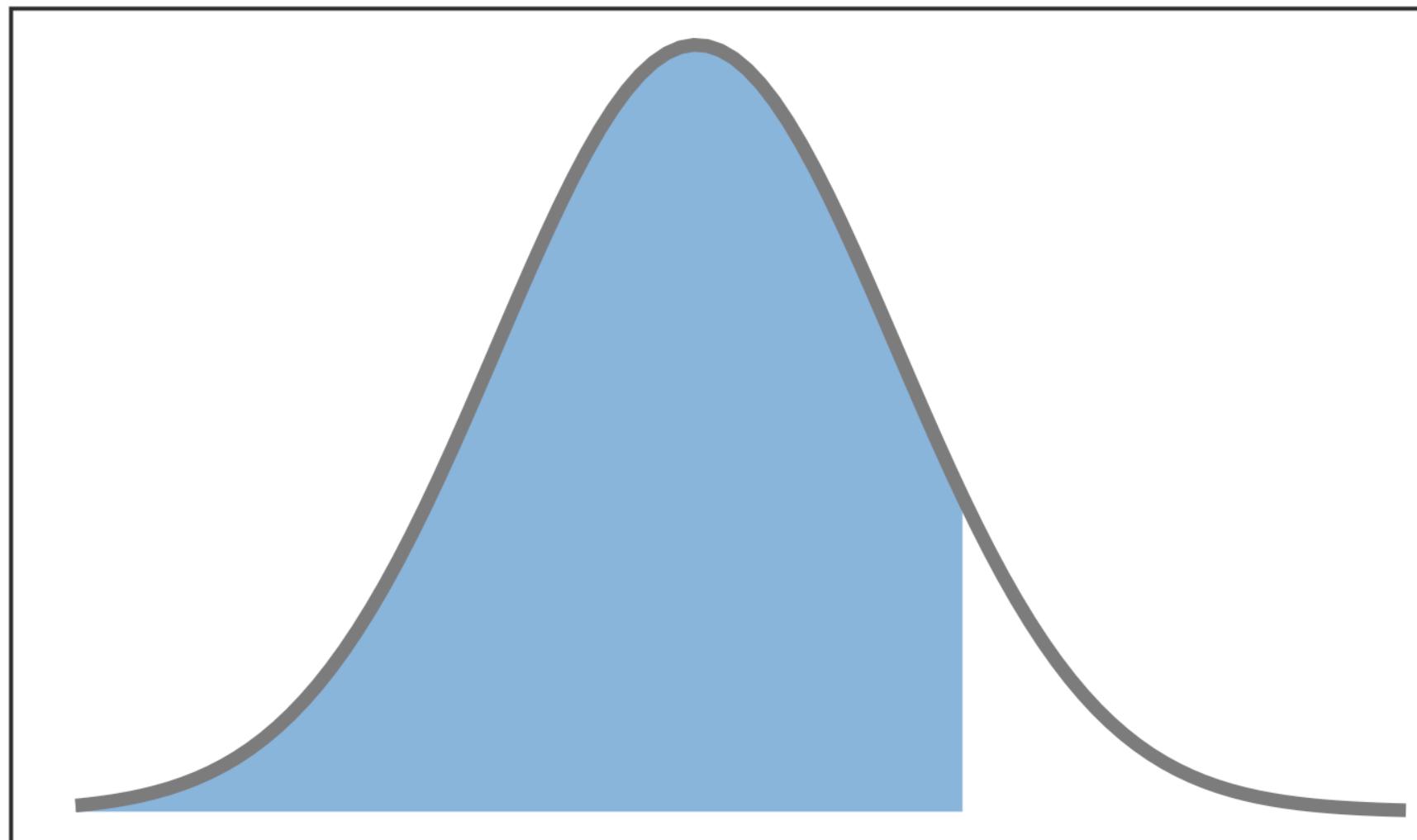


In a certain population of fish, individual lengths follow a normal distribution. The mean length of the fish is 54.0 mm, and the standard deviation is 4.5 mm. What percentage of the fish are shorter than 60 mm?

*Even if you use R to calculate p-value,
ALWAYS smart to make
a quick sketch*



$$Z = \frac{60 - 54}{4.5} = 1.33$$



Tables of the Normal Distribution



Z = 1.33

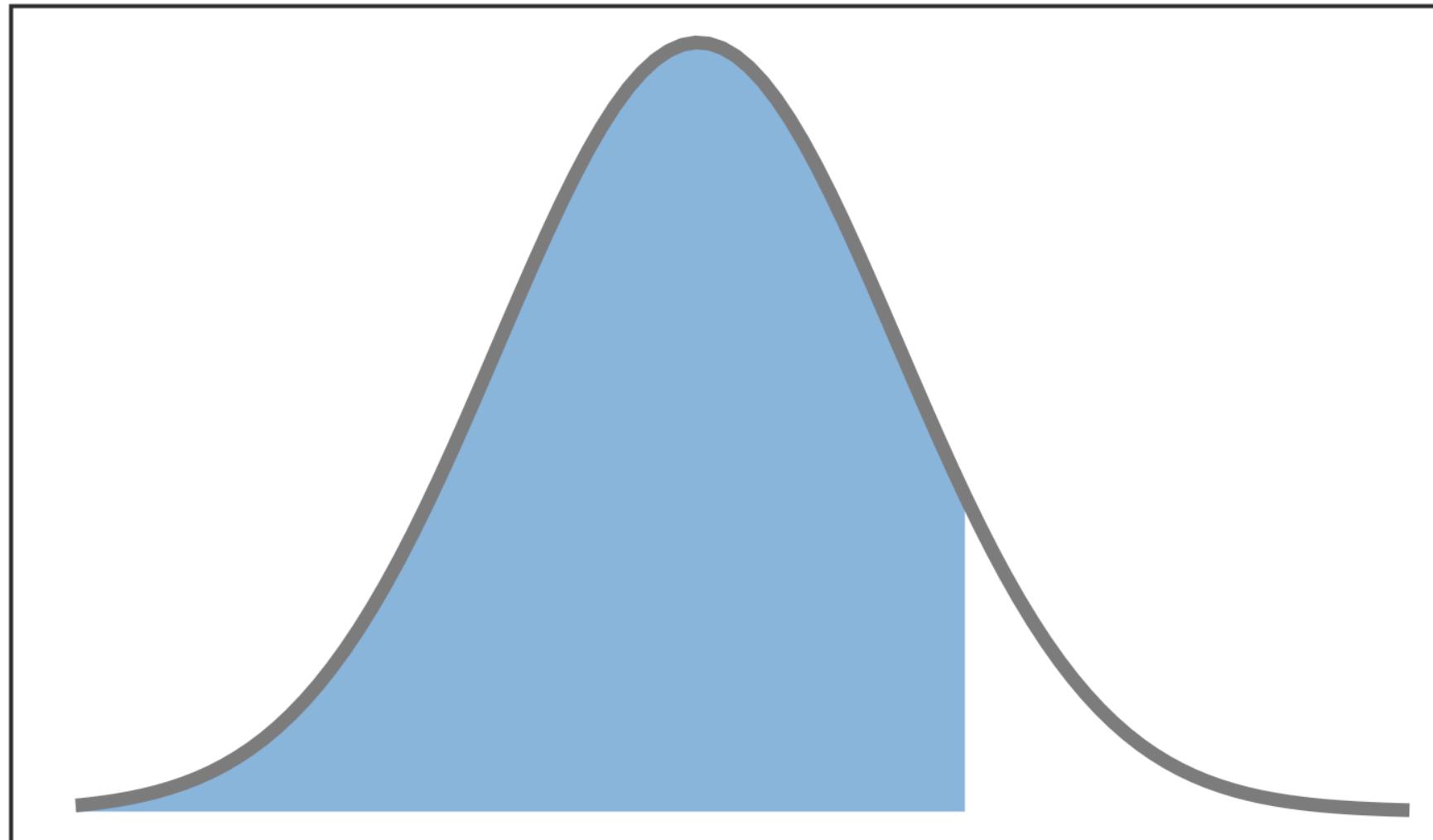
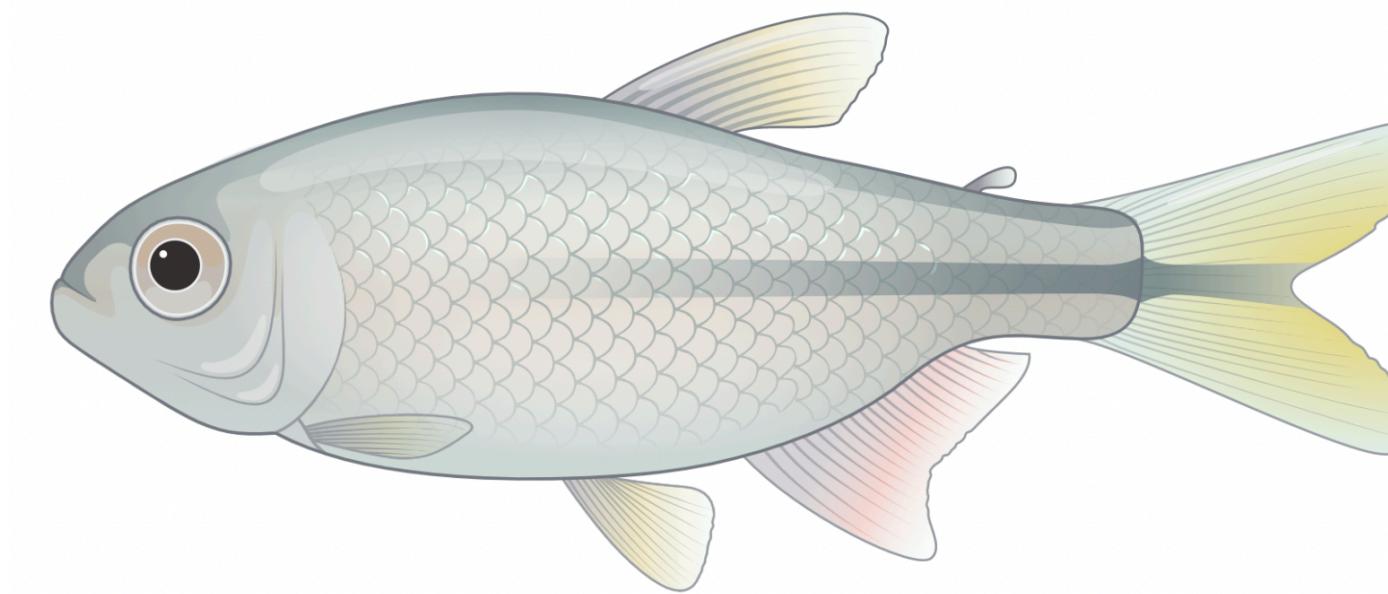
Area = 0.9082

Probability Content from -oo to Z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8906	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9237	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

In a certain population of fish, individual lengths follow a normal distribution. The mean length of the fish is 54.0 mm, and the standard deviation is 4.5 mm. What percentage of the fish are shorter than 60 mm?

*Even if you use R to calculate p-value,
ALWAYS smart to make a quick sketch*



$$Z = \frac{60 - 54}{4.5} = 1.33$$

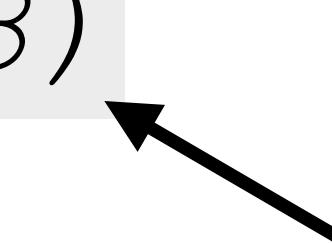
Area = 0.9082

```
> pnorm(z, mean, sd)
```

```
> pnorm(1.33, 0, 1)
```

```
> pnorm(1.33)
```

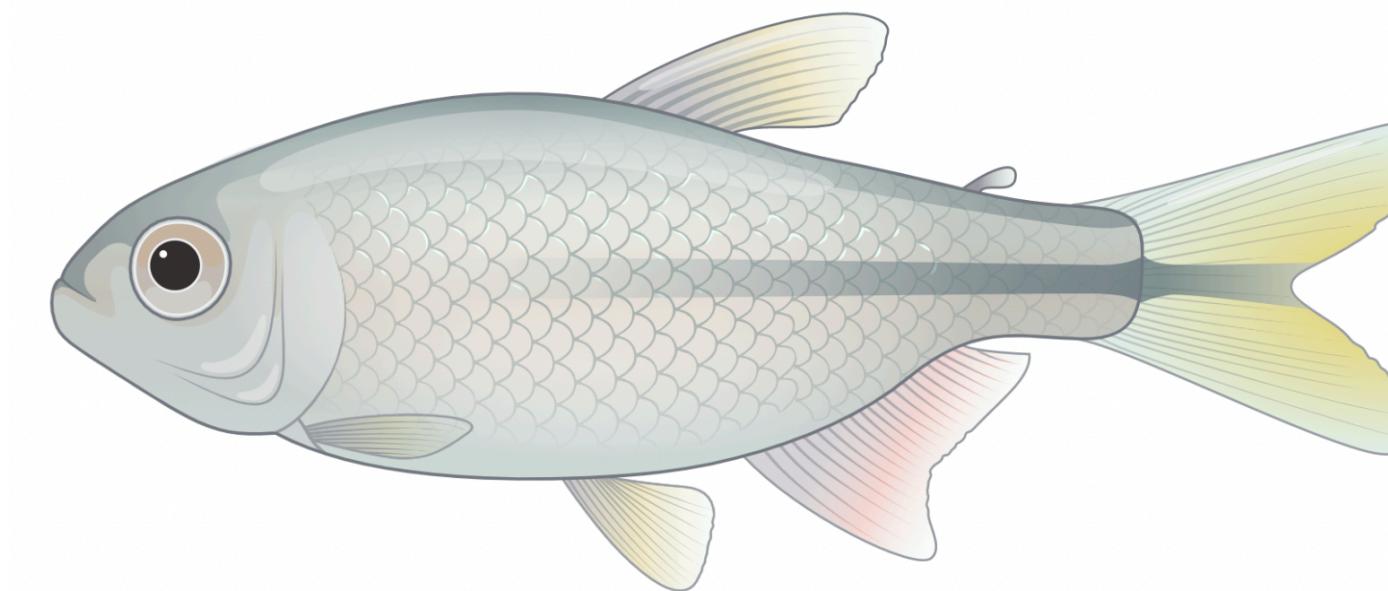
```
> 0.908
```



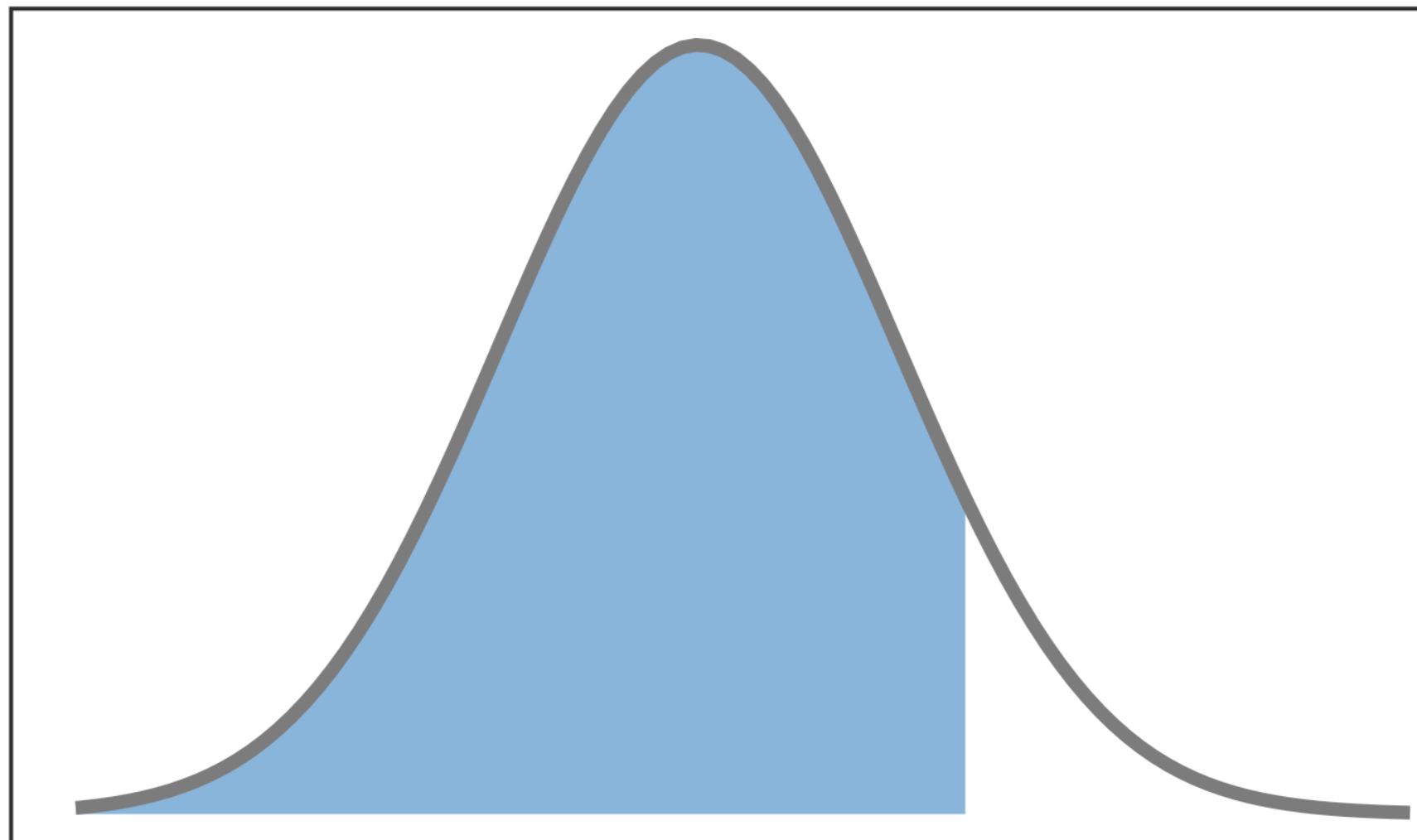
Default is mean = 0, sd = 1

In a certain population of fish, individual lengths follow a normal distribution. The mean length of the fish is 54.0 mm, and the standard deviation is 4.5 mm. What percentage of the fish are shorter than 60 mm?

*Even if you use R to calculate p-value,
ALWAYS smart to make
a quick sketch*



$$Z = \frac{60 - 54}{4.5} = 1.33$$



```
> pnorm(z, mean, sd)
```

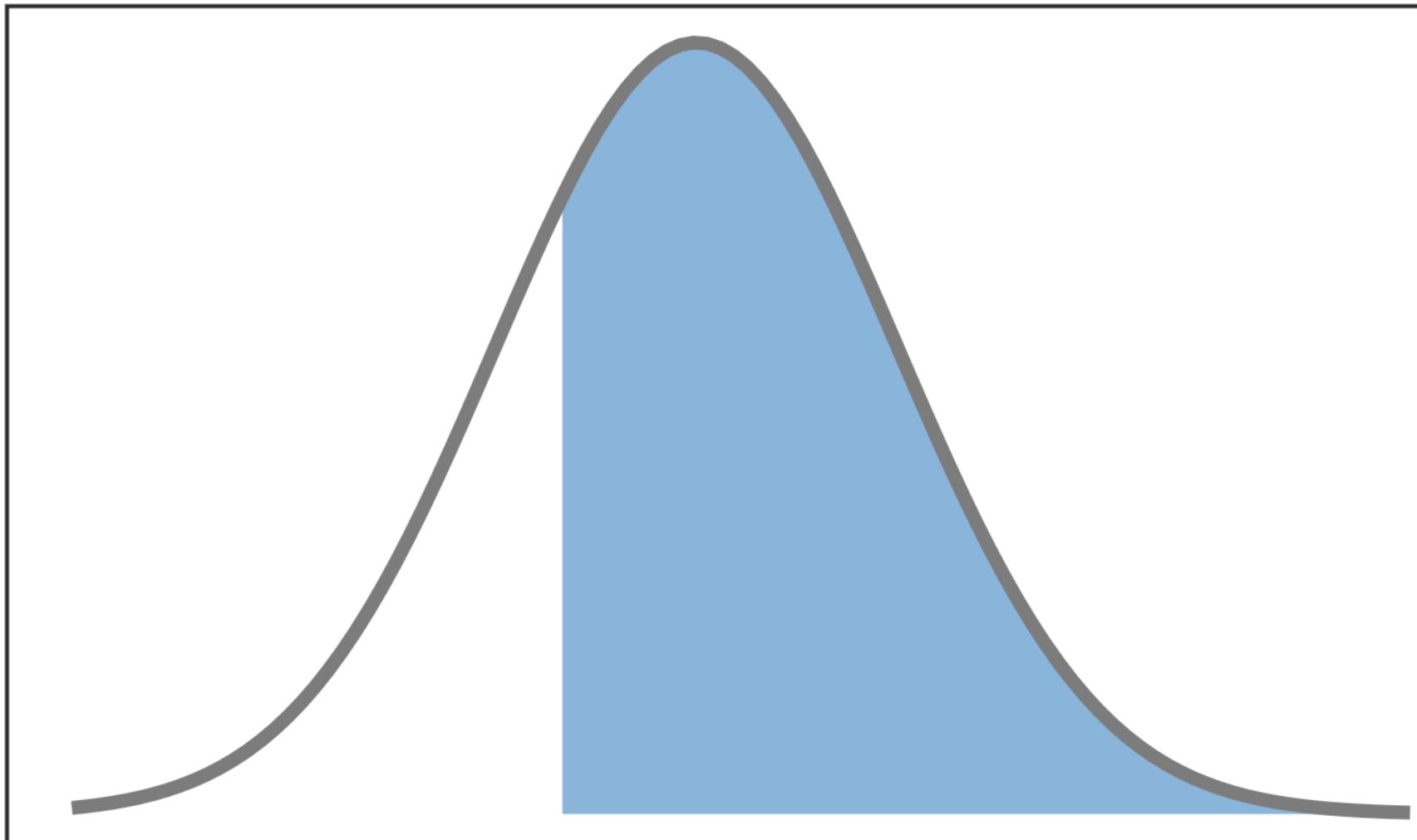
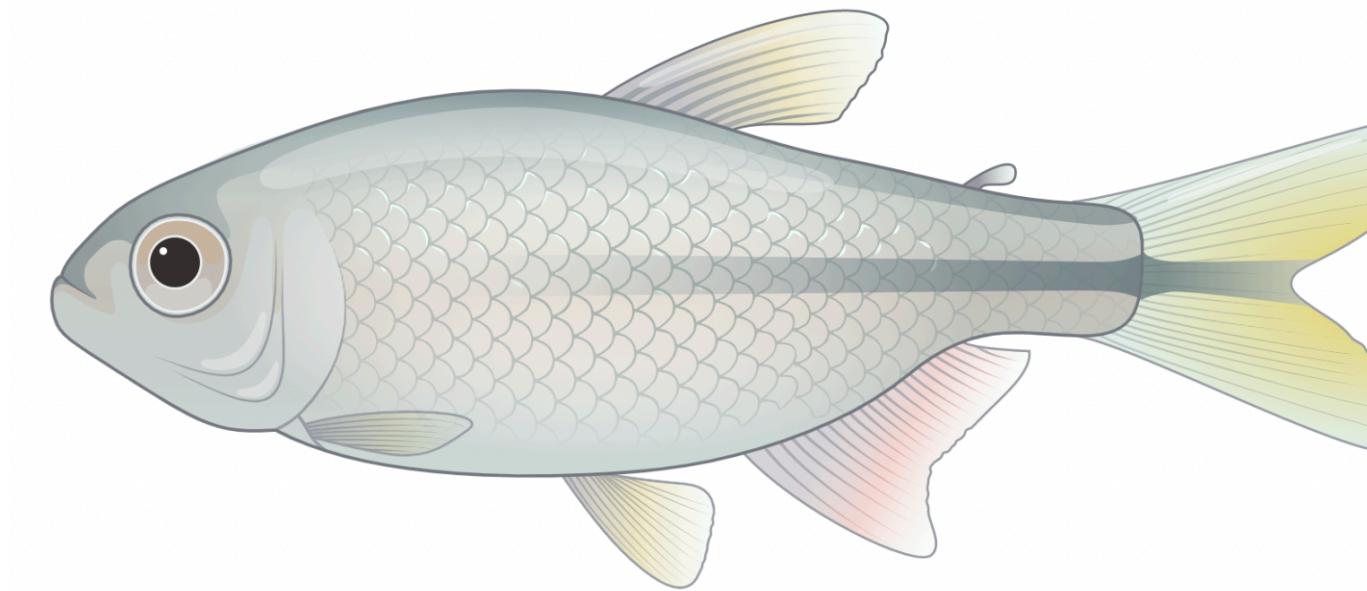
```
> pnorm(60, 54, 4.5)
```

```
> 0.908
```



In a certain population of fish, individual lengths follow a normal distribution. The mean length of the fish is 54.0 mm, and the standard deviation is 4.5 mm. What percentage of the fish are longer than 51 mm?

*Even if you use R to calculate p-value,
ALWAYS smart to make
a quick sketch*



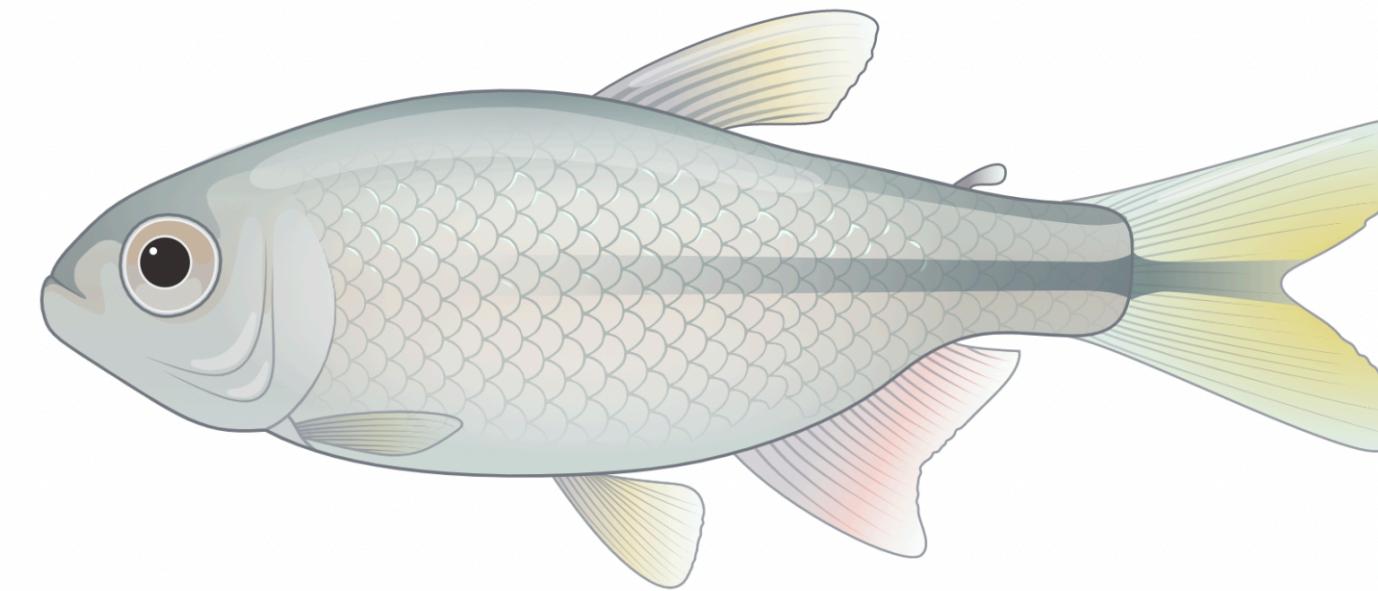
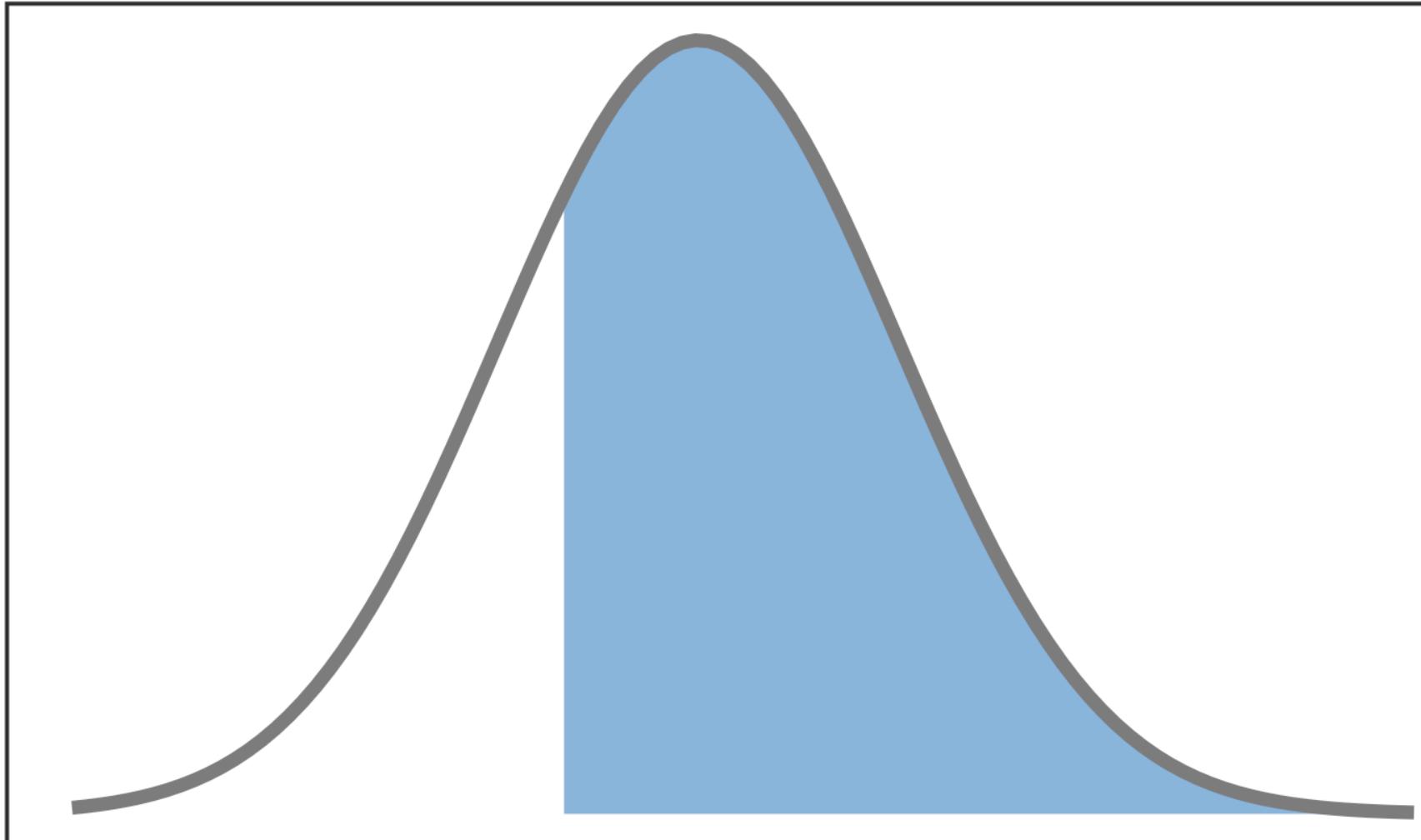
```
> 1 - pnorm(z, mean, sd)
```

```
> 1 - pnorm(51, 54, 4.5)
```

```
> 0.747
```

In a certain population of fish, individual lengths follow a normal distribution. The mean length of the fish is 54.0 mm, and the standard deviation is 4.5 mm. What percentage of the fish are longer than 51 mm?

*Even if you use R to calculate p-value,
ALWAYS smart to make a quick sketch*



$N(54, 4.5)$

(How we describe this distribution)

```
> pnorm(z, mean, sd)
```

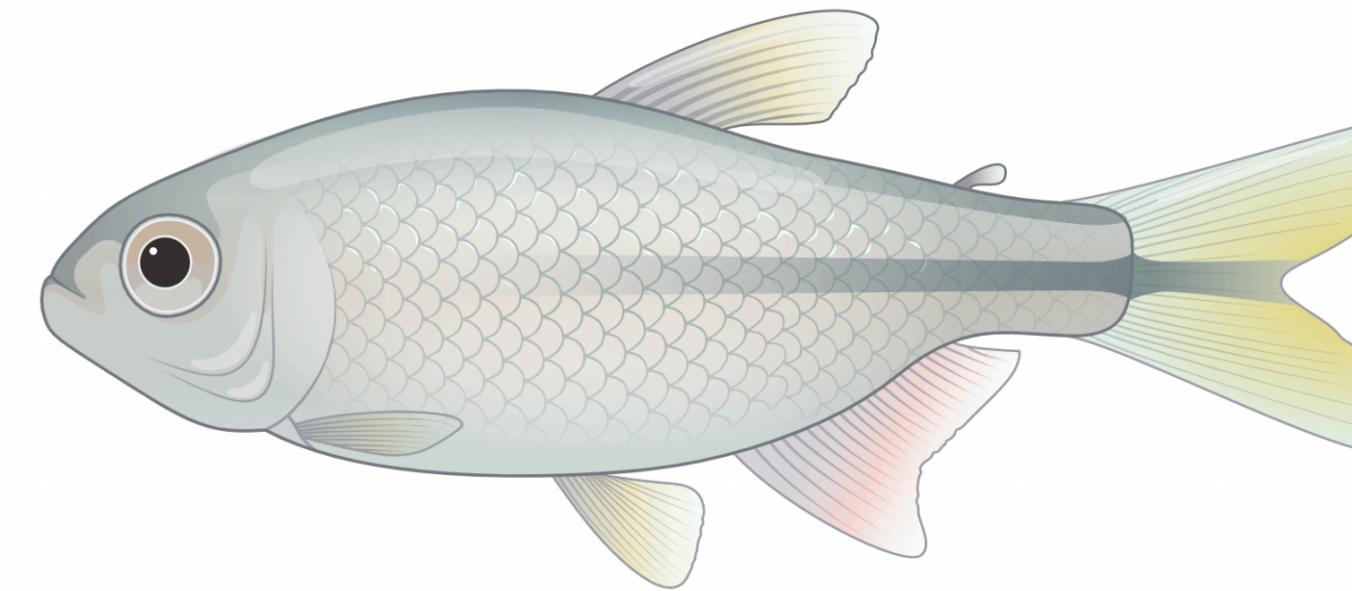
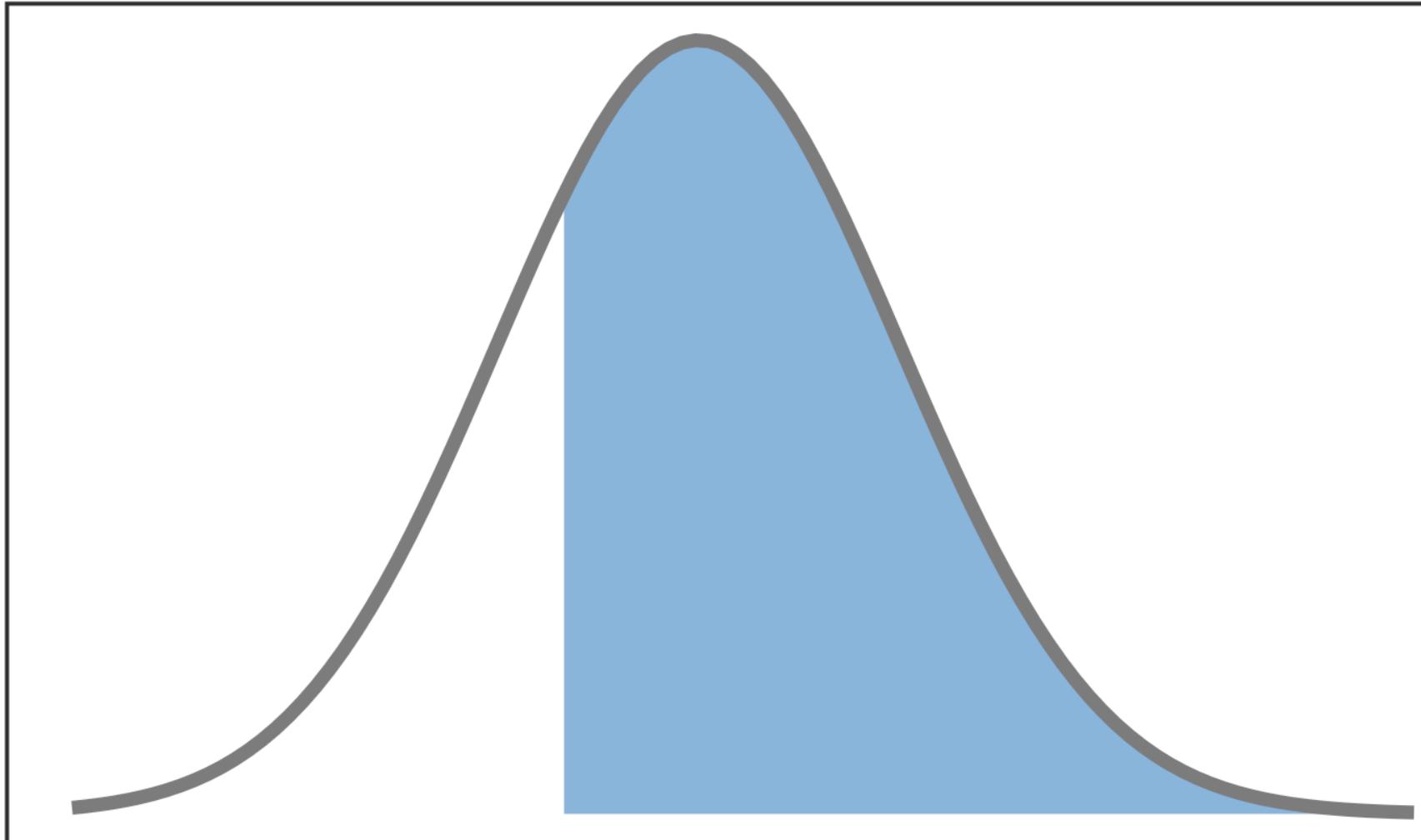
```
> 1 - pnorm(51, 54, 4.5)
```

```
> pnorm(51, 54, 4.5, lower.tail = F)
```

```
> 0.747
```

In a certain population of fish, individual lengths follow a normal distribution. The mean length of the fish is 54.0 mm, and the standard deviation is 4.5 mm. What percentage of the fish are longer than 51 mm?

*Even if you use R to calculate p-value,
ALWAYS smart to make
a quick sketch*



$$x \sim N(54, 4.5) > 51$$

How we could write this question

```
> pnorm(z, mean, sd)
```

```
> 1 - pnorm(51, 54, 4.5)
```

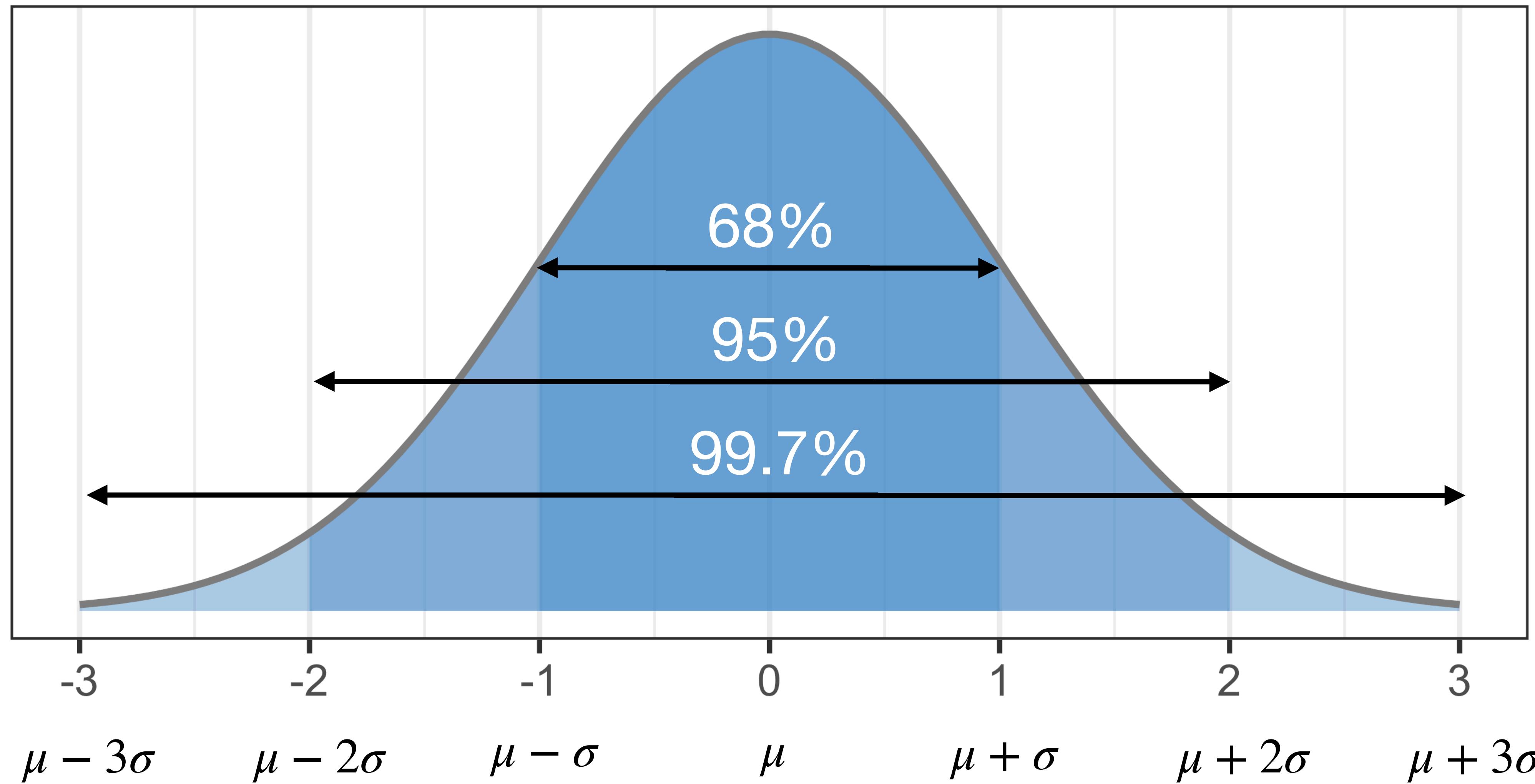
```
> pnorm(51, 54, 4.5, lower.tail = F)
```

```
> 0.747
```

How can we assess normality?

- As we will see, many statistical tests are based on having data from a normal distribution, but how do you know if your data follows a normal distribution?
 - And what should you do if it doesn't?

How can we assess normality?



How can we assess normality?

```
> summary(fish)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
46.21	52.41	55.13	54.82	57.46	62.83

1. Calculate the mean +/- 1 SD

```
> mean(fish) + sd(fish)
```

```
[1] 58.65478
```

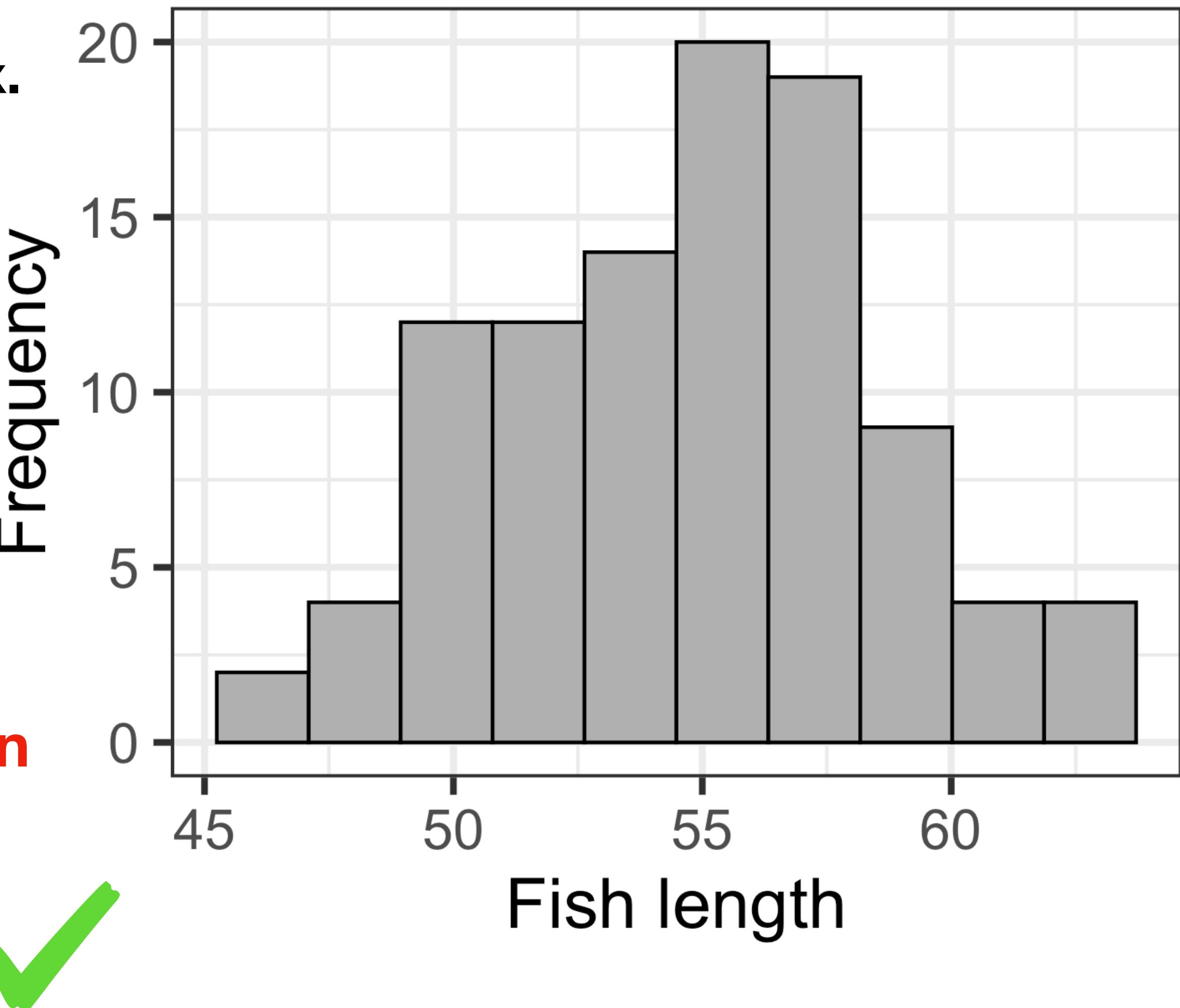
```
> mean(fish) - sd(fish)
```

```
[1] 50.97822
```

2. What percent of the population is within these values?

```
> sum(fish < 58.65 & fish > 50.97)
```

```
[1] 65
```



How can we assess normality?

```
> summary(fish)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
46.21	52.41	55.13	54.82	57.46	62.83

1. Calculate the mean +/- 2 SD

```
> mean(fish) + 2*sd(fish)
```

```
[1] 62.49306
```

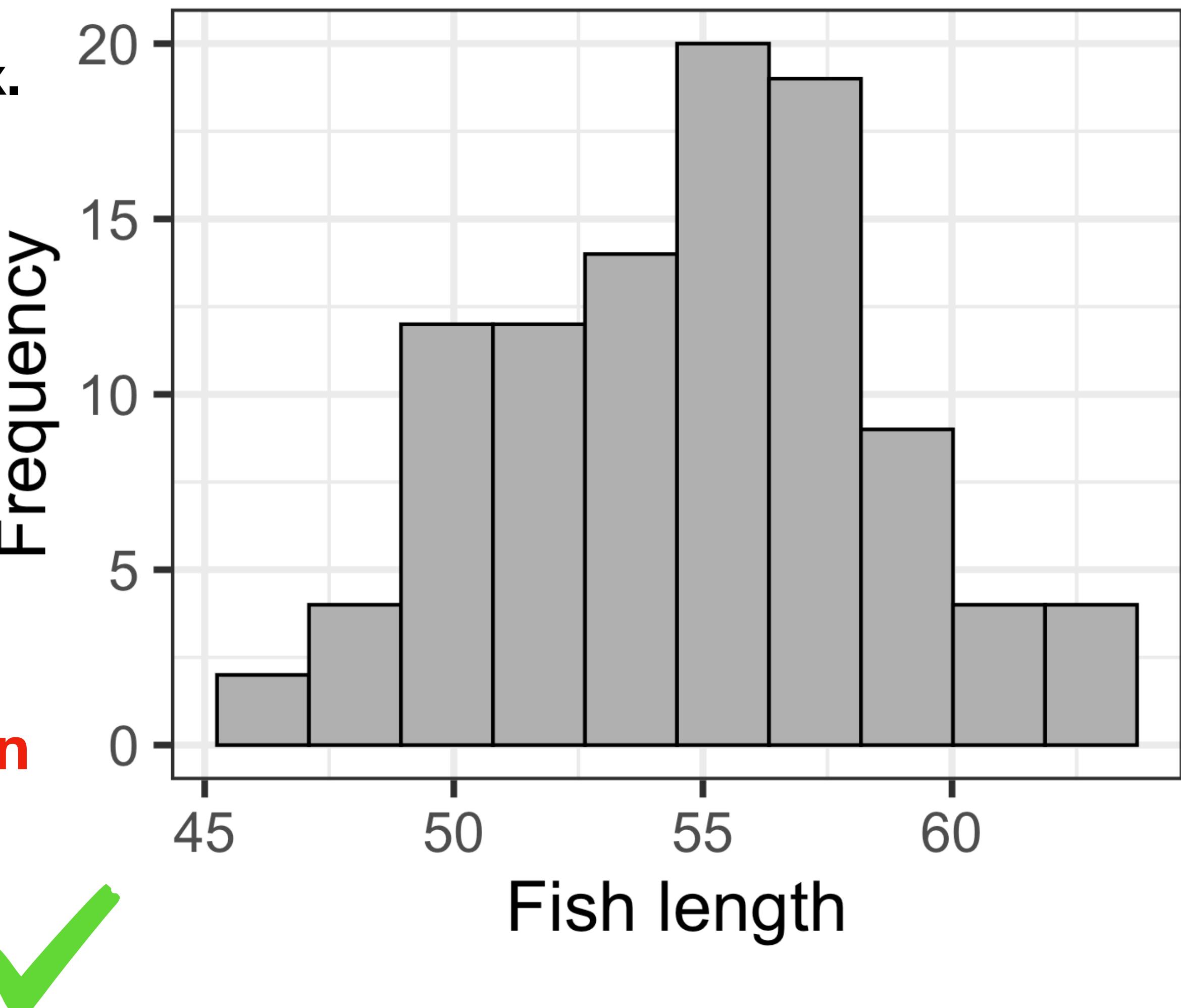
```
> mean(fish) - 2*sd(fish)
```

```
[1] 47.13994
```

2. What percent of the population is within these values?

```
> sum(fish < 62.49 & fish > 47.13)
```

```
[1] 96
```



How can we assess normality?

```
> summary(fish)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
46.21	52.41	55.13	54.82	57.46	62.83

1. Calculate the mean +/- 3 SD

```
> mean(fish) + 3*sd(fish)
```

```
[1] 66.33134
```

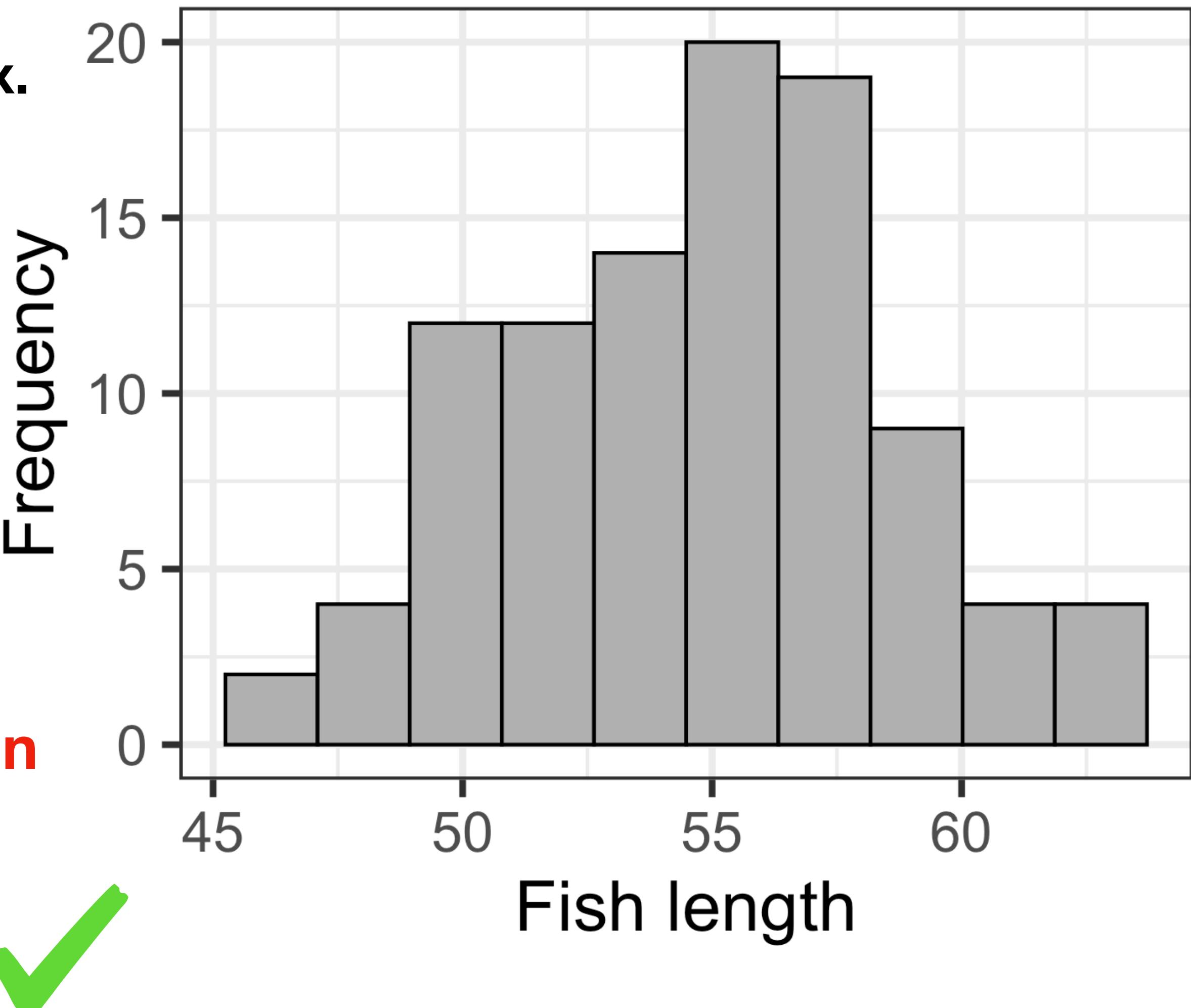
```
> mean(fish) - 3*sd(fish)
```

```
[1] 43.30166
```

2. What percent of the population is within these values?

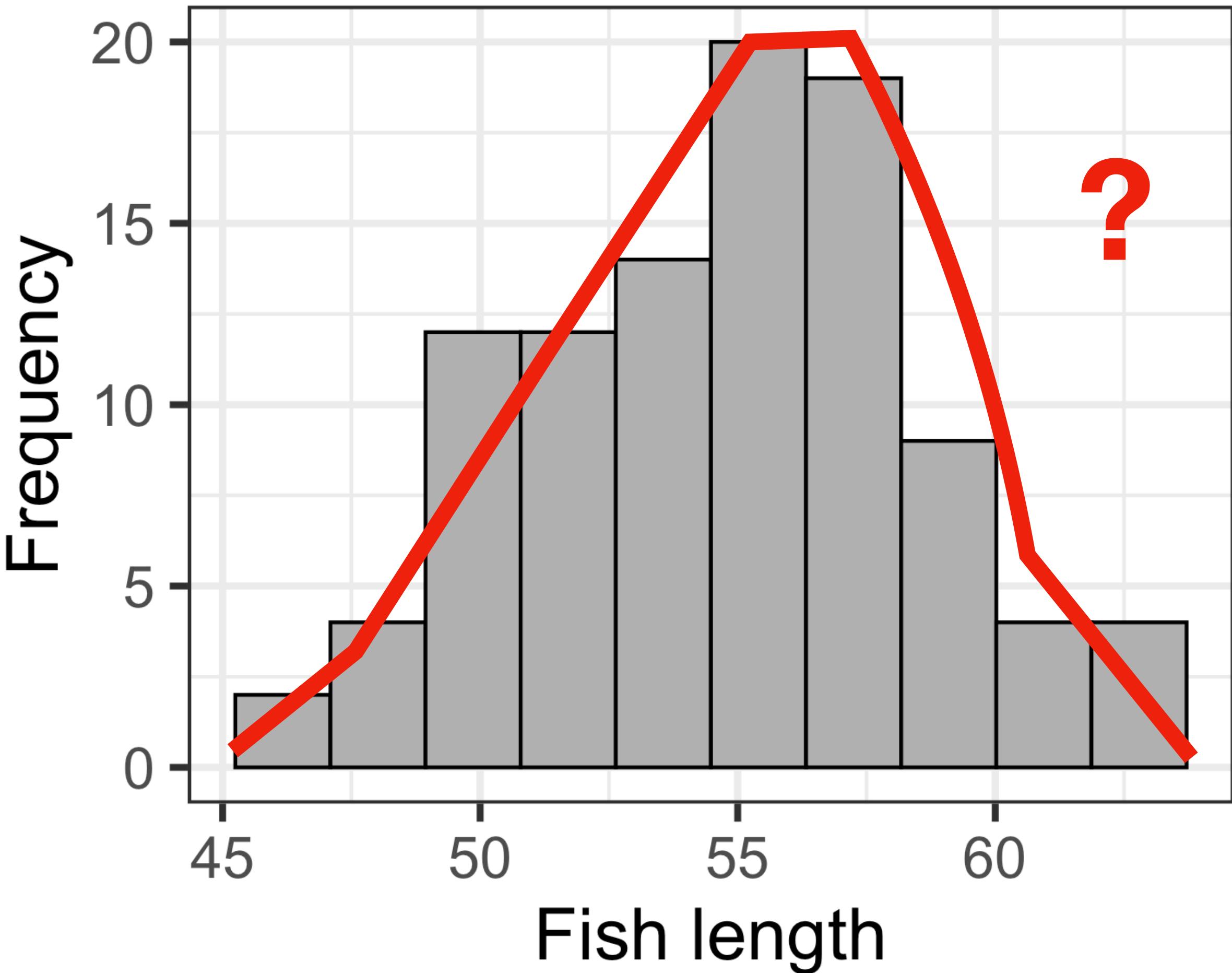
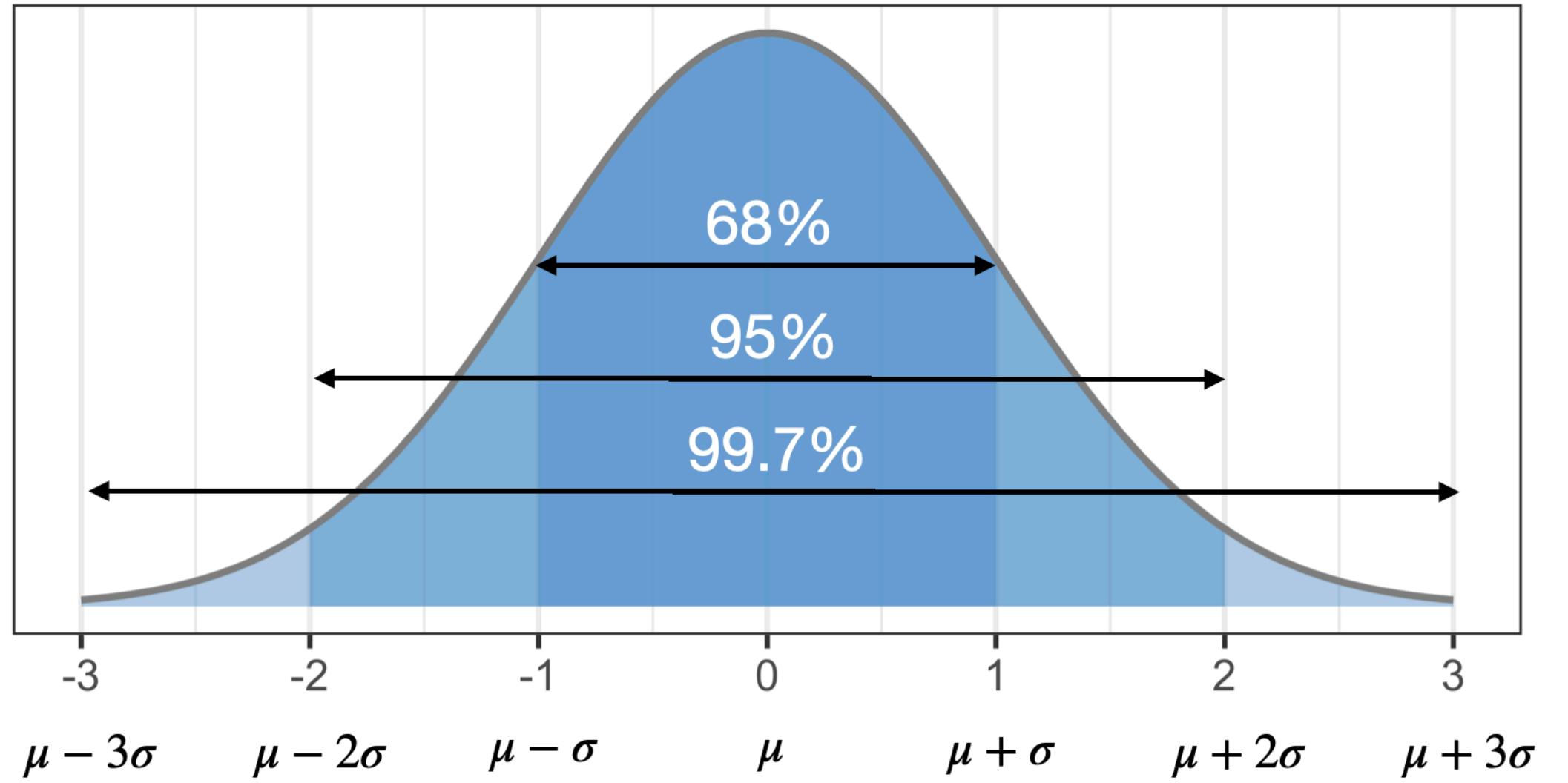
```
> sum(fish < 66.33 & fish > 43.30)
```

```
[1] 100
```

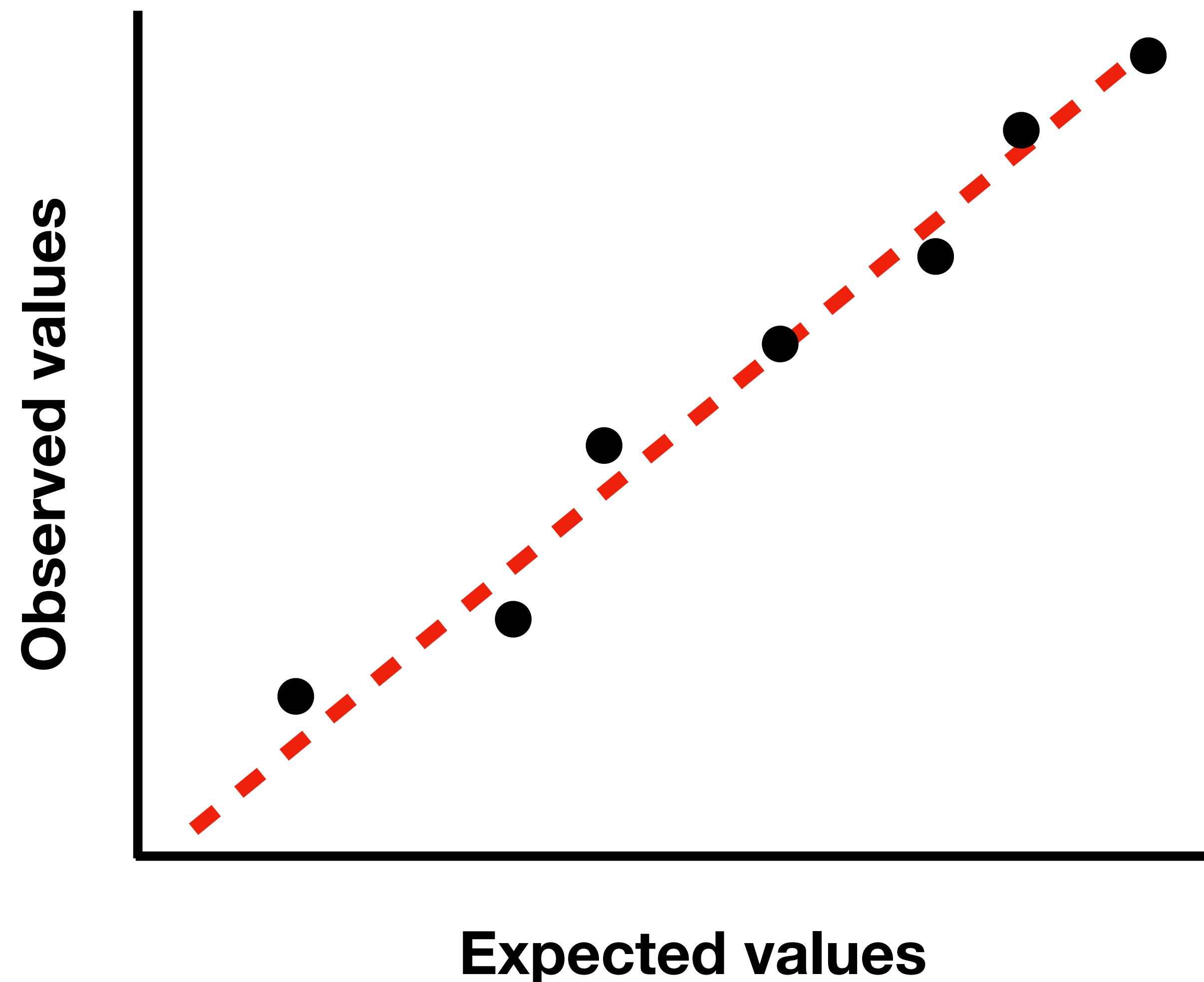


How can we assess normality?

+/- 1 SD = 65%
+/- 2 SD = 95%
+/- 3 SD = 100%



Normal Quantile Plots

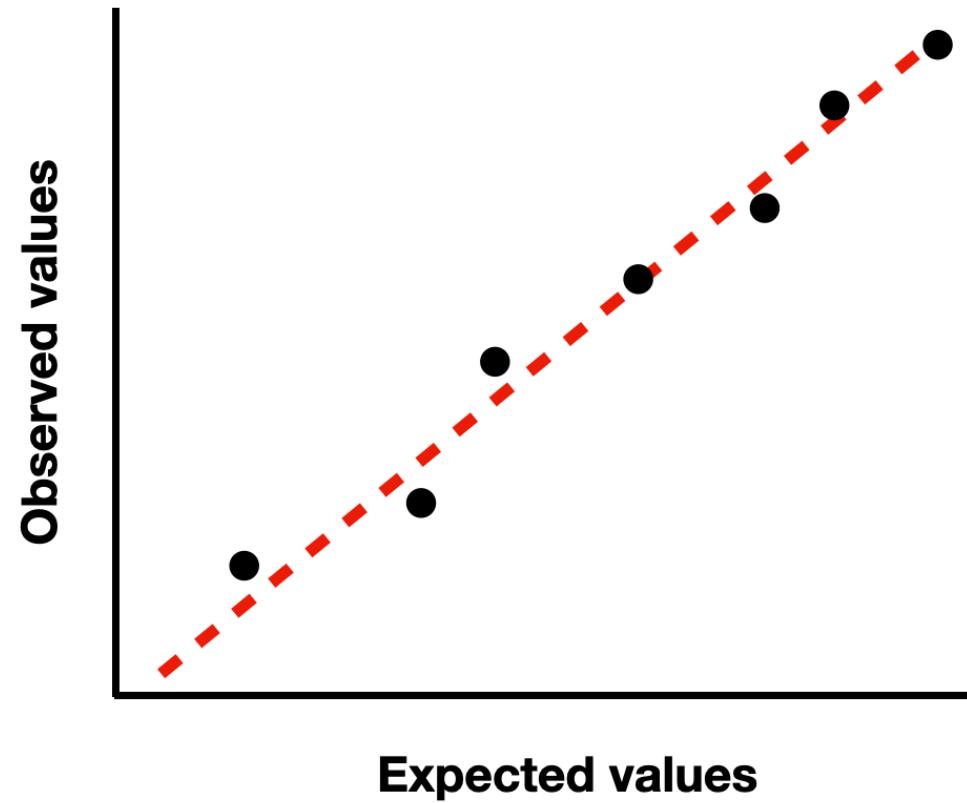


$$Z = \frac{y - \mu}{\sigma}$$



$$y = \sigma z + \mu$$

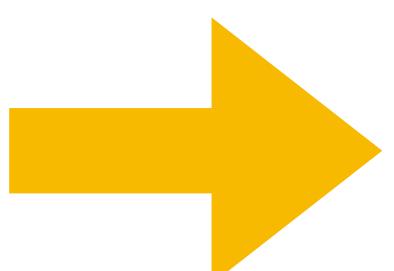
$$(y = mx + b)$$



Normal Quantile Plots

- 1. Calculate (adjusted) percentile of each data point**
- 2. Calculate the z score from the percentile**
- 3. Calculate the theoretical y (expected value) from z score**
- 4. Plot expected vs. observed**

	1	2	3	4	5	6	7	8	9	10	11
Observed height	61.0	62.5	63	64	64.5	65	66.5	67	68	68.5	70.5



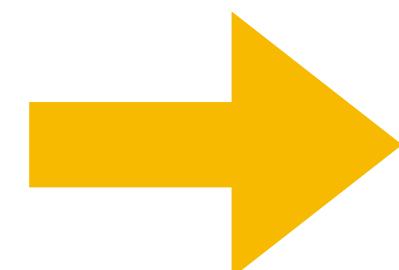
1. Calculate (adjusted) percentile of each data point

2. Calculate the z score from the percentile

3. Calculate the theoretical y (expected value) from z score

4. Plot expected vs. observed

	1	2	3	4	5	6	7	8	9	10	11
Observed height	61.0	62.5	63	64	64.5	65	66.5	67	68	68.5	70.5



1. Calculate (adjusted) percentile of each data point

$$1/11 = 9.09 \%$$

$$11/11 = 100 \%$$

2. Calculate the z score from the percentile

3. Calculate the theoretical y (expected value) from z score

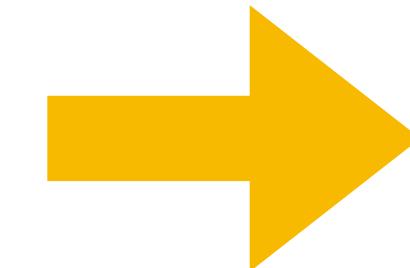
4. Plot expected vs. observed

	1	2	3	4	5	6	7	8	9	10	11
Observed height	61.0	62.5	63	64	64.5	65	66.5	67	68	68.5	70.5



1. Calculate (adjusted) percentile of each data point

$$100\left(i - \frac{\frac{1}{2}}{n}\right) = 100\left(1 - \frac{\frac{1}{2}}{11}\right) = 4.55\% \quad (11 - 0.5)/11 = 95.45\%$$



2. Calculate the z score from the percentile

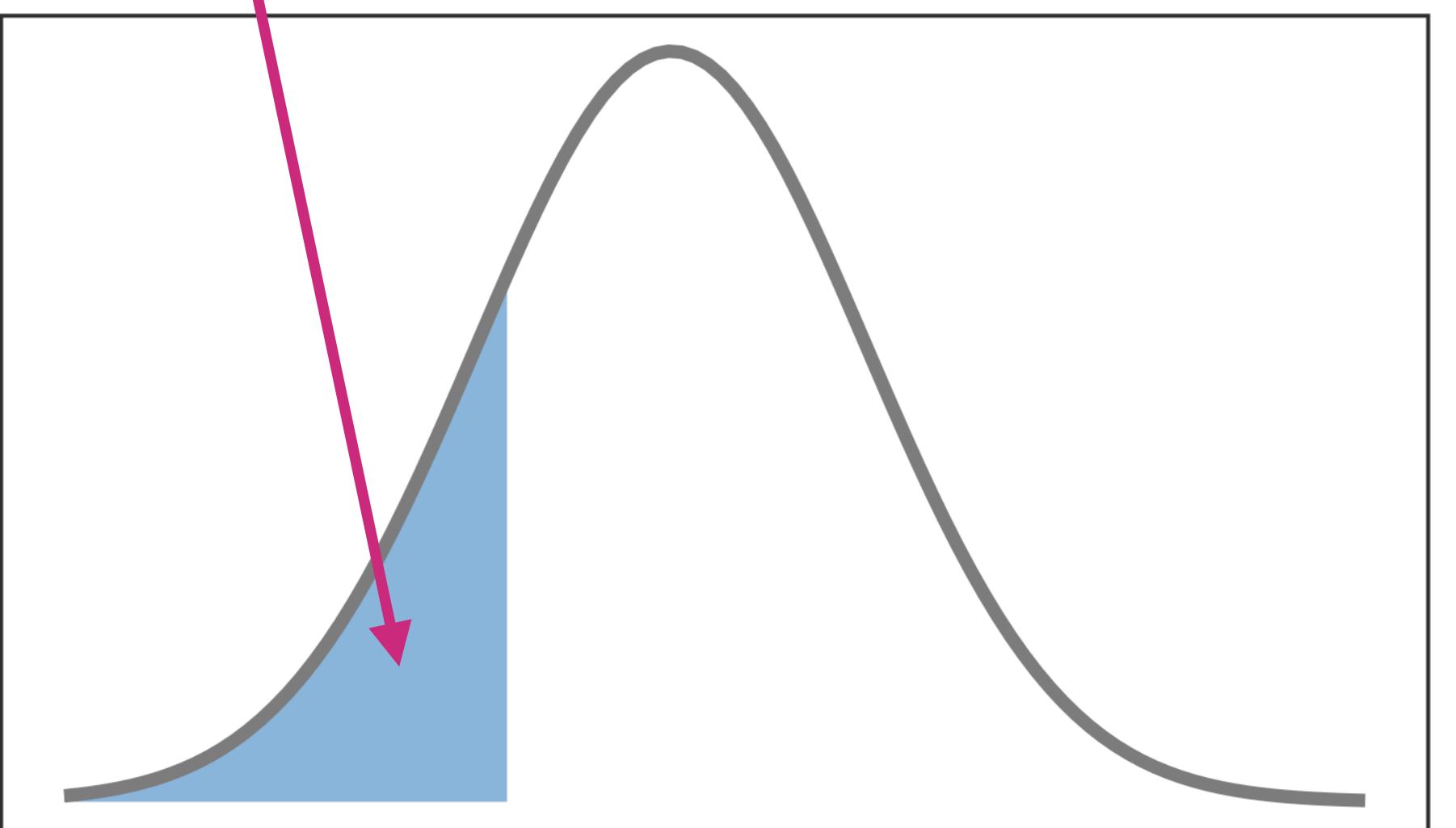
3. Calculate the theoretical y (expected value) from z score

4. Plot expected vs. observed



4.55%

(Area under curve)



$z = -1.69$

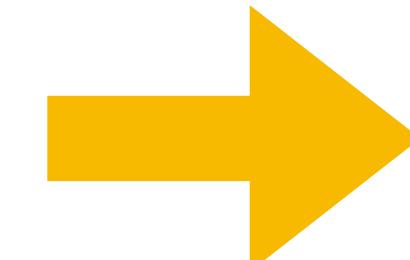
<i>z</i>	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

	1	2	3	4	5	6	7	8	9	10	11
Observed height	61.0	62.5	63	64	64.5	65	66.5	67	68	68.5	70.5



1. Calculate (adjusted) percentile of each data point

$$100\left(i - \frac{\frac{1}{2}}{n}\right) = 100\left(1 - \frac{\frac{1}{2}}{11}\right) = 4.55\% \quad (11 - 0.5)/11 = 95.45\%$$



2. Calculate the z score from the percentile

```
> qnorm(0.0455) = -1.69
```

3. Calculate the theoretical y (expected value) from z score

4. Plot expected vs. observed

	1	2	3	4	5	6	7	8	9	10	11
Observed height	61.0	62.5	63	64	64.5	65	66.5	67	68	68.5	70.5



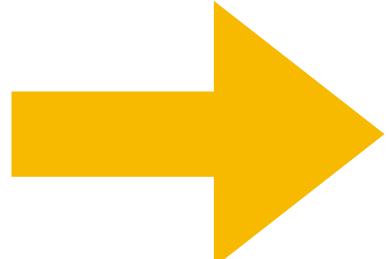
1. Calculate (adjusted) percentile of each data point

$$100\left(i - \frac{\frac{1}{2}}{n}\right) = 100\left(1 - \frac{\frac{1}{2}}{11}\right) = 4.55\% \quad (11 - 0.5)/11 = 95.45\%$$



2. Calculate the z score from the percentile

```
> qnorm(0.0455) = -1.69
```



3. Calculate the theoretical y (expected value) from z score

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

4. Plot expected vs. observed



	1	2	3	4	5	6	7	8	9	10	11
Observed height	61.0	62.5	63	64	64.5	65	66.5	67	68	68.5	70.5



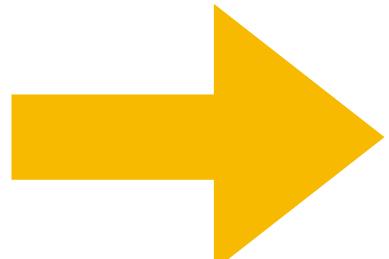
1. Calculate (adjusted) percentile of each data point

$$100\left(i - \frac{\frac{1}{2}}{2}\right)/n = 100\left(1 - \frac{\frac{1}{2}}{2}\right)/11 = 4.55\% \quad (11 - 0.5)/11 = 95.45\%$$



2. Calculate the z score from the percentile

```
> qnorm(0.0455) = -1.69
```



3. Calculate the theoretical y (expected value) from z score

$$Z = \frac{y - \mu}{\sigma} \longrightarrow y = z\sigma + \mu$$

4. Plot expected vs. observed



	1	2	3	4	5	6	7	8	9	10	11
Observed height	61.0	62.5	63	64	64.5	65	66.5	67	68	68.5	70.5



1. Calculate (adjusted) percentile of each data point

$$100\left(i - \frac{\frac{1}{2}}{2}\right)/n = 100\left(1 - \frac{\frac{1}{2}}{2}\right)/11 = 4.55\% \quad (11 - 0.5)/11 = 95.45\%$$



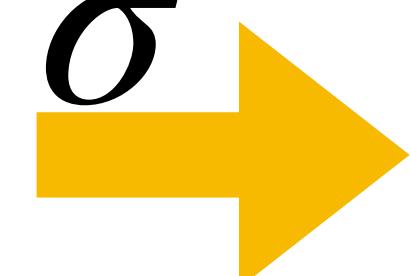
2. Calculate the z score from the percentile

```
> qnorm(0.0455) = -1.69
```



3. Calculate the theoretical y (expected value) from z score

$$Z = \frac{y - \mu}{\sigma} \longrightarrow y = -1.69 * 2.87 + 65.5 = 60.6$$



4. Plot expected vs. observed



	1	2	3	4	5	6	7	8	9	10	11
Observed height	61.0	62.5	63	64	64.5	65	66.5	67	68	68.5	70.5

✓ 1. Calculate (adjusted) percentile of each data point

$$100\left(i - \frac{\frac{1}{2}}{n}\right) = 100\left(1 - \frac{\frac{1}{2}}{11}\right) = 4.55\% \quad (11 - 0.5)/11 = 95.45\%$$

→ 2. Calculate the theoretical y (expected value) from percentile

```
> qnorm(0.0455, mean(height), sd(height))
```

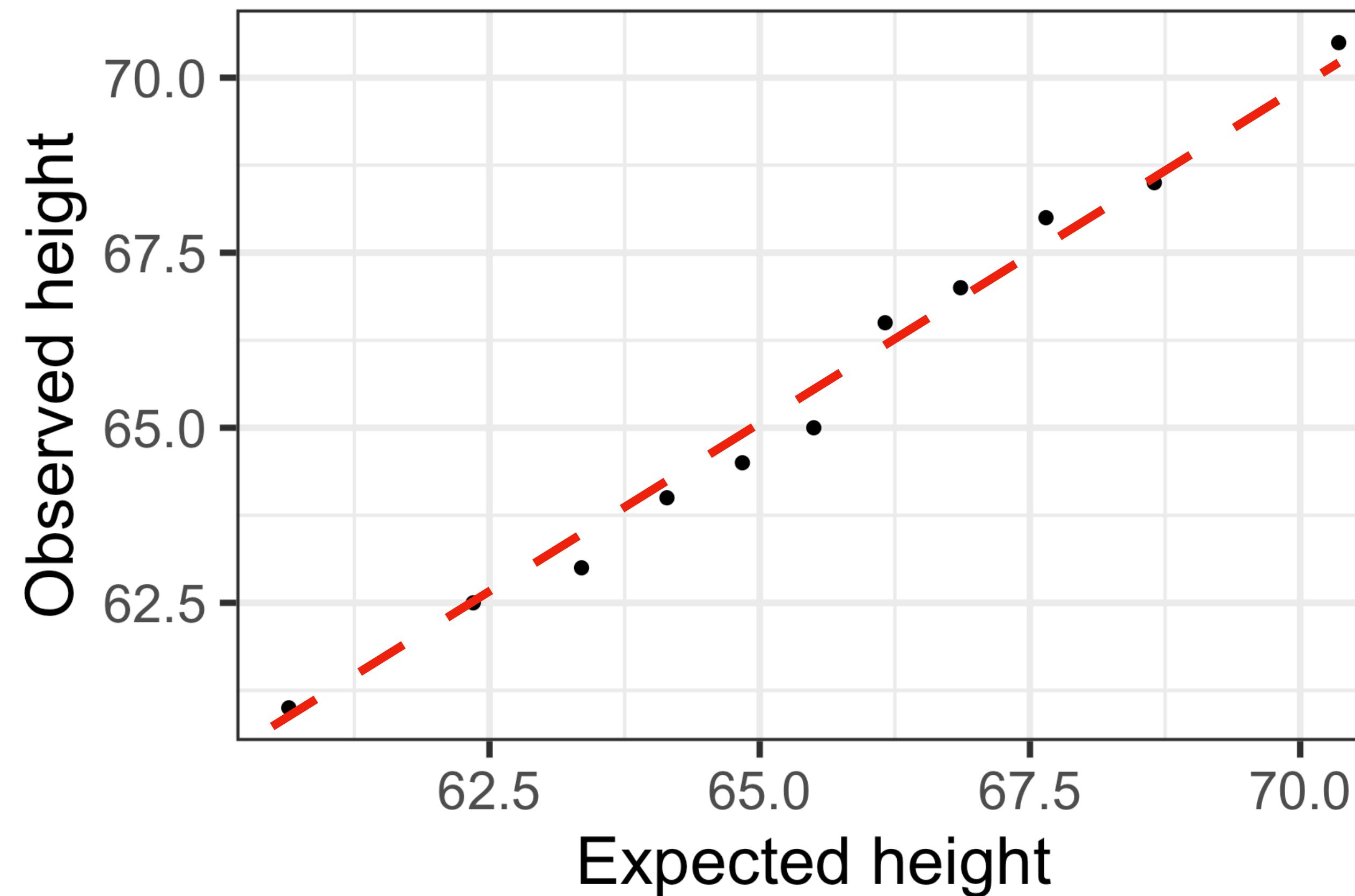
```
> qnorm(0.0455, 65.5, 2.87)
```

= 60.06 ✓

4. Plot expected vs. observed

	1	2	3	4	5	6	7	8	9	10	11
Observed height	61.0	62.5	63	64	64.5	65	66.5	67	68	68.5	70.5
Percentile	0.00	18.18	27.27	36.36	45.45	54.55	63.64	72.73	81.82	90.91	100
Adjusted Percentile	4.55	13.64	22.73	31.82	40.91	50	59.09	68.18	77.27	86.36	95.45
Z score	-1.69	-1.10	-0.75	-0.47	-0.23	0	0.23	0.47	0.75	1.10	1.69
Theoretical height	60.6	62.3	63.4	64.1	64.8	65.5	66.2	66.9	67.6	68.7	70.4

	1	2	3	4	5	6	7	8	9	10	11
Observed height	61.0	62.5	63	64	64.5	65	66.5	67	68	68.5	70.5
Theoretical height	60.6	62.3	63.4	64.1	64.8	65.5	66.2	66.9	67.6	68.7	70.4



Shapiro-Wilk test to assess non-normality

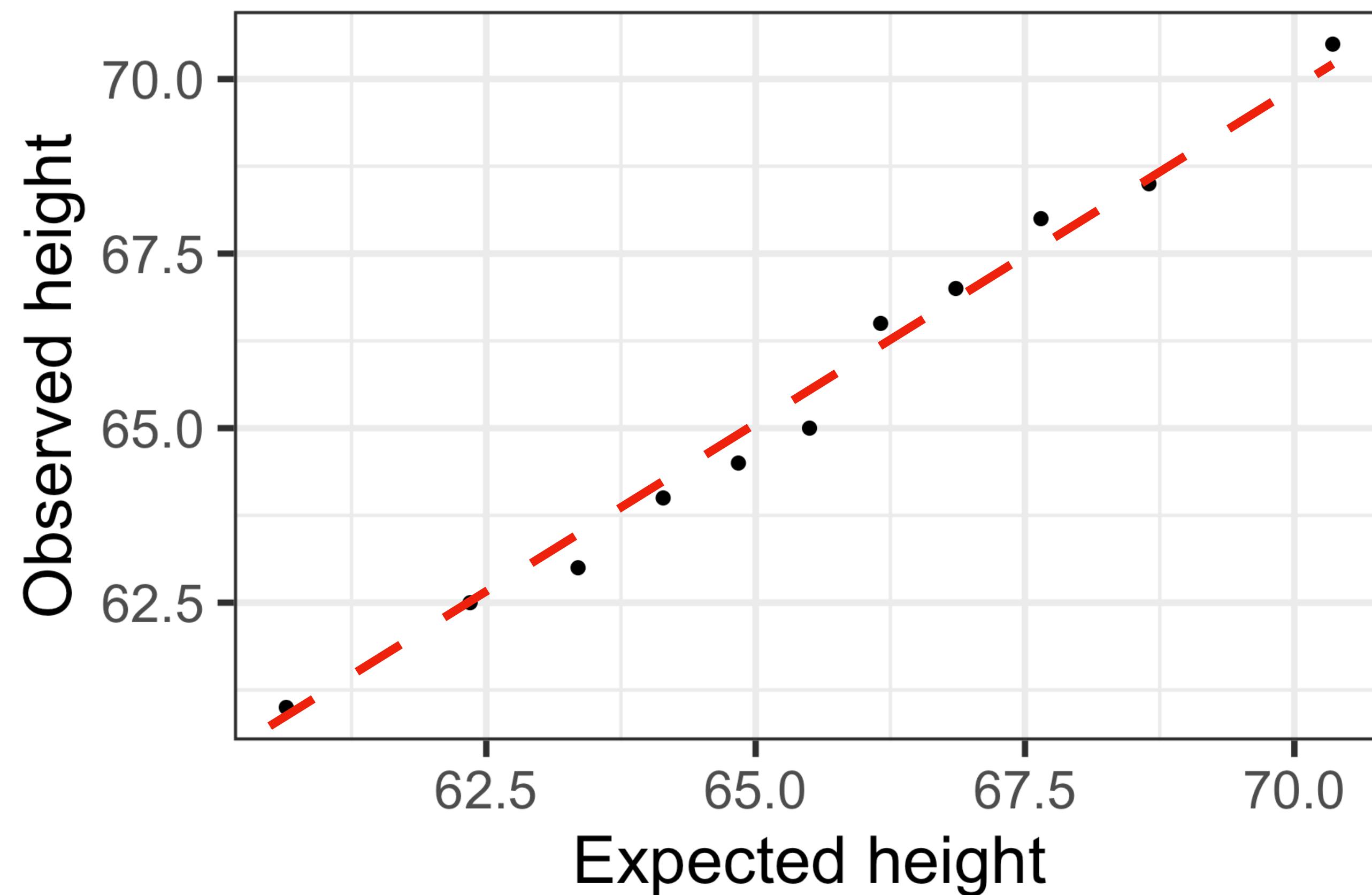
```
> shapiro.test(height)
```

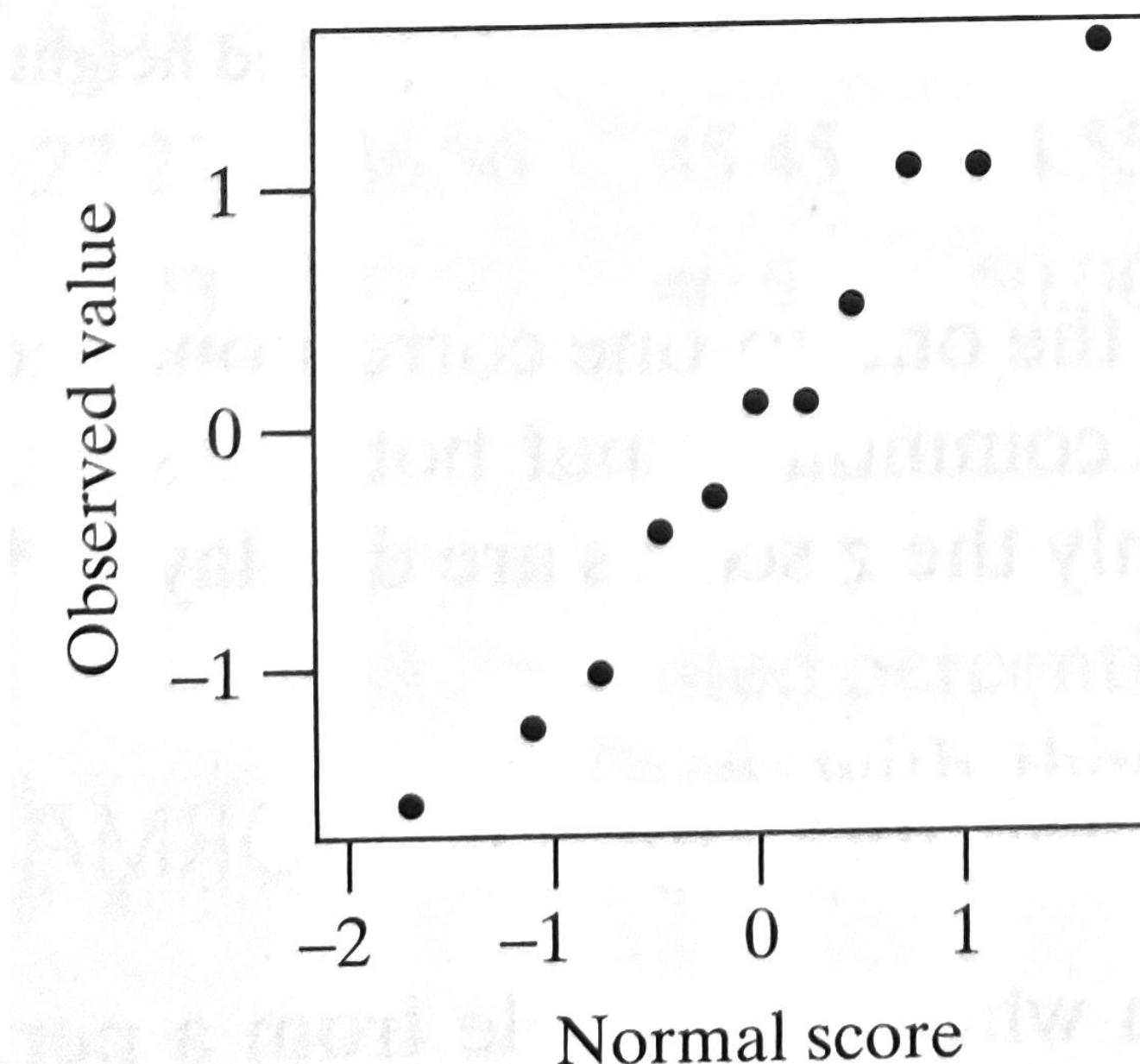
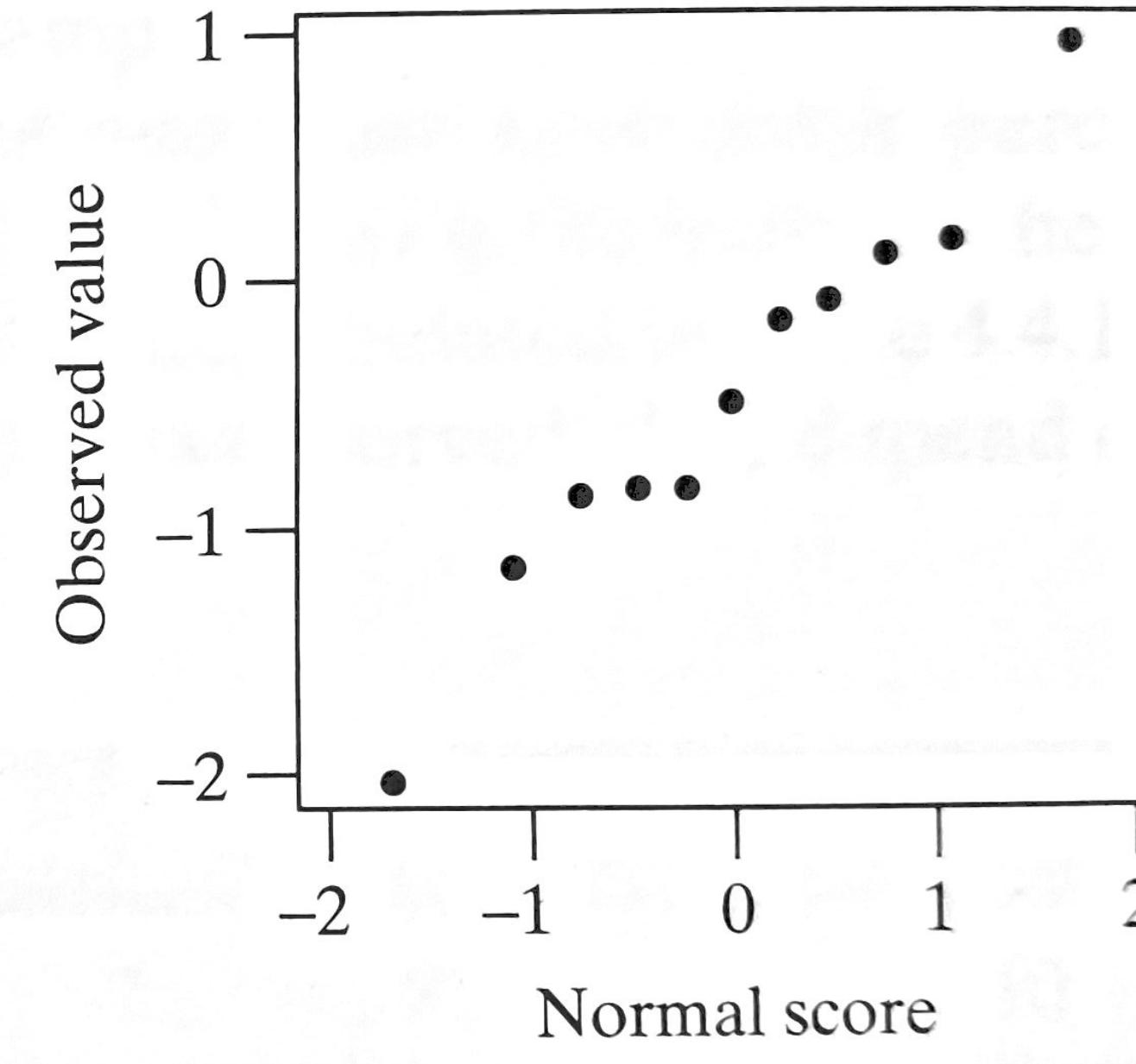
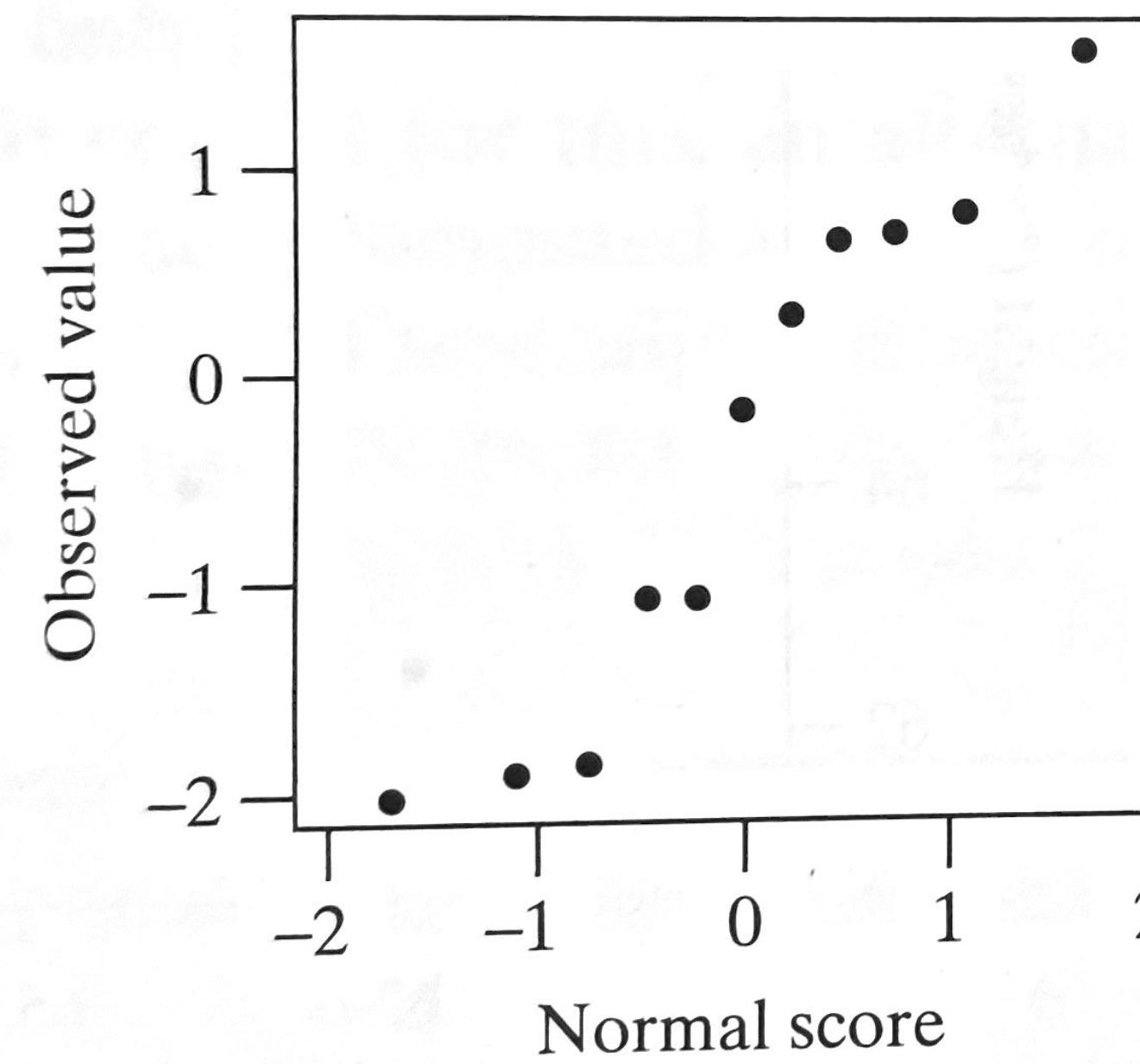
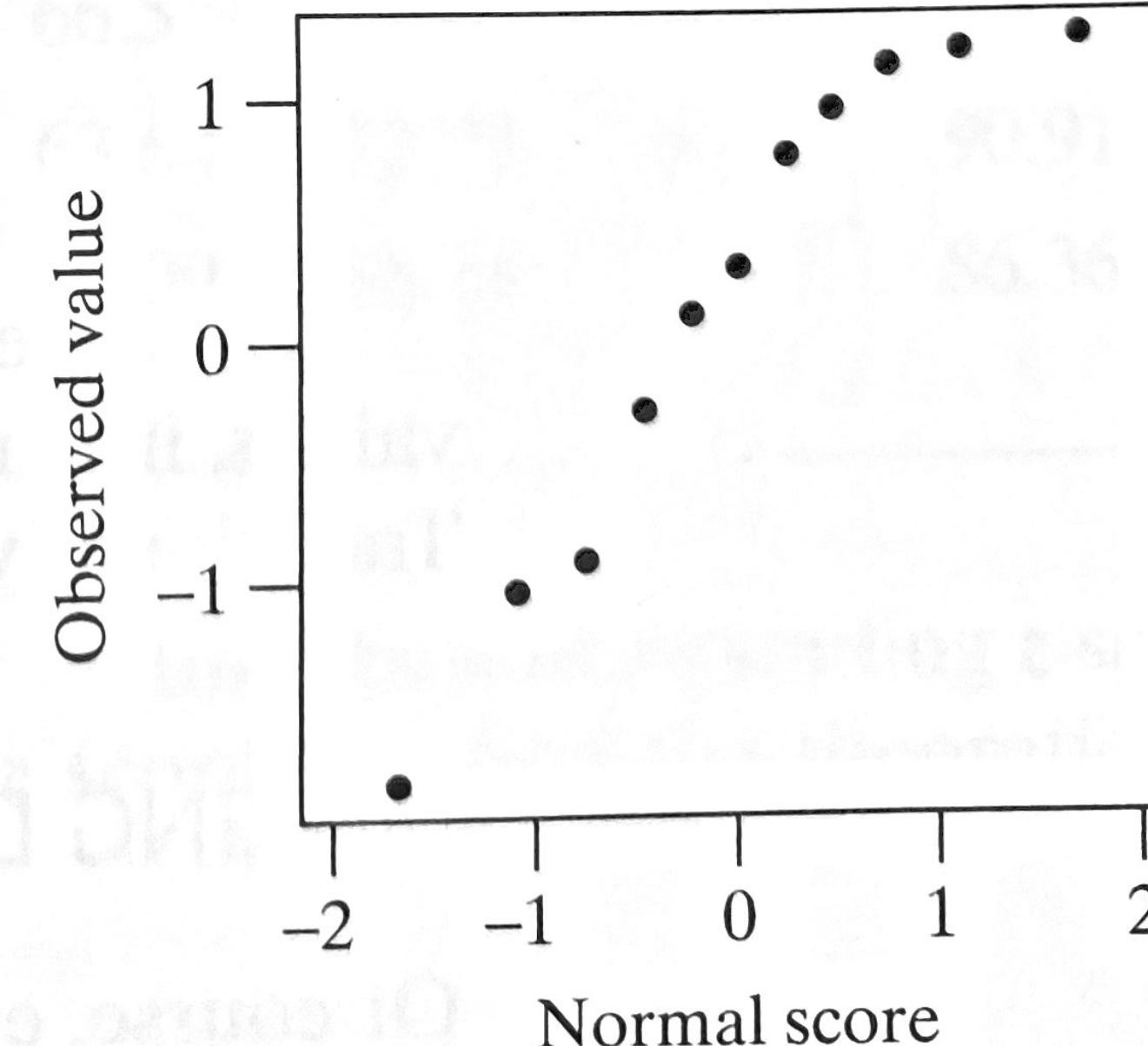
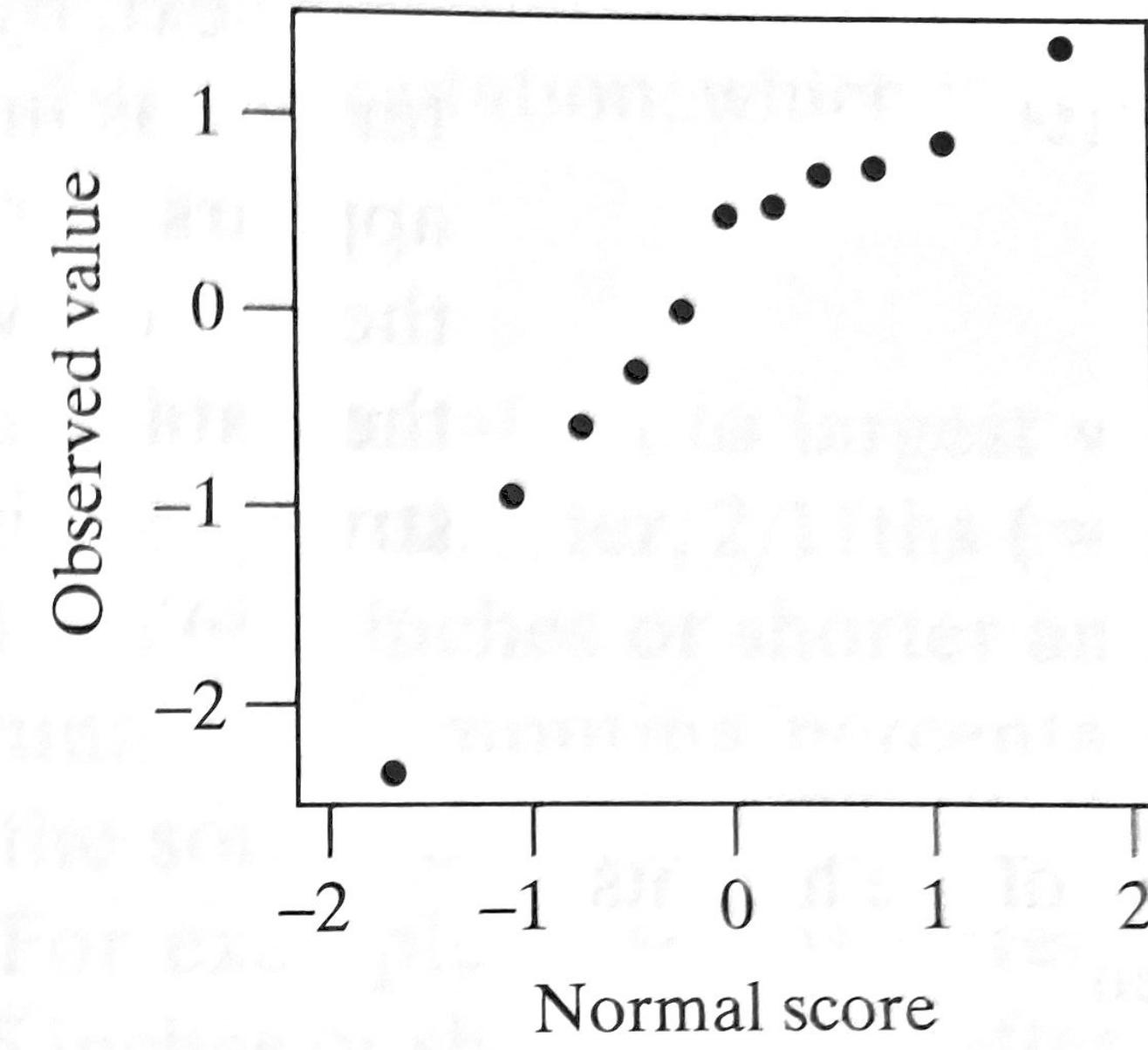
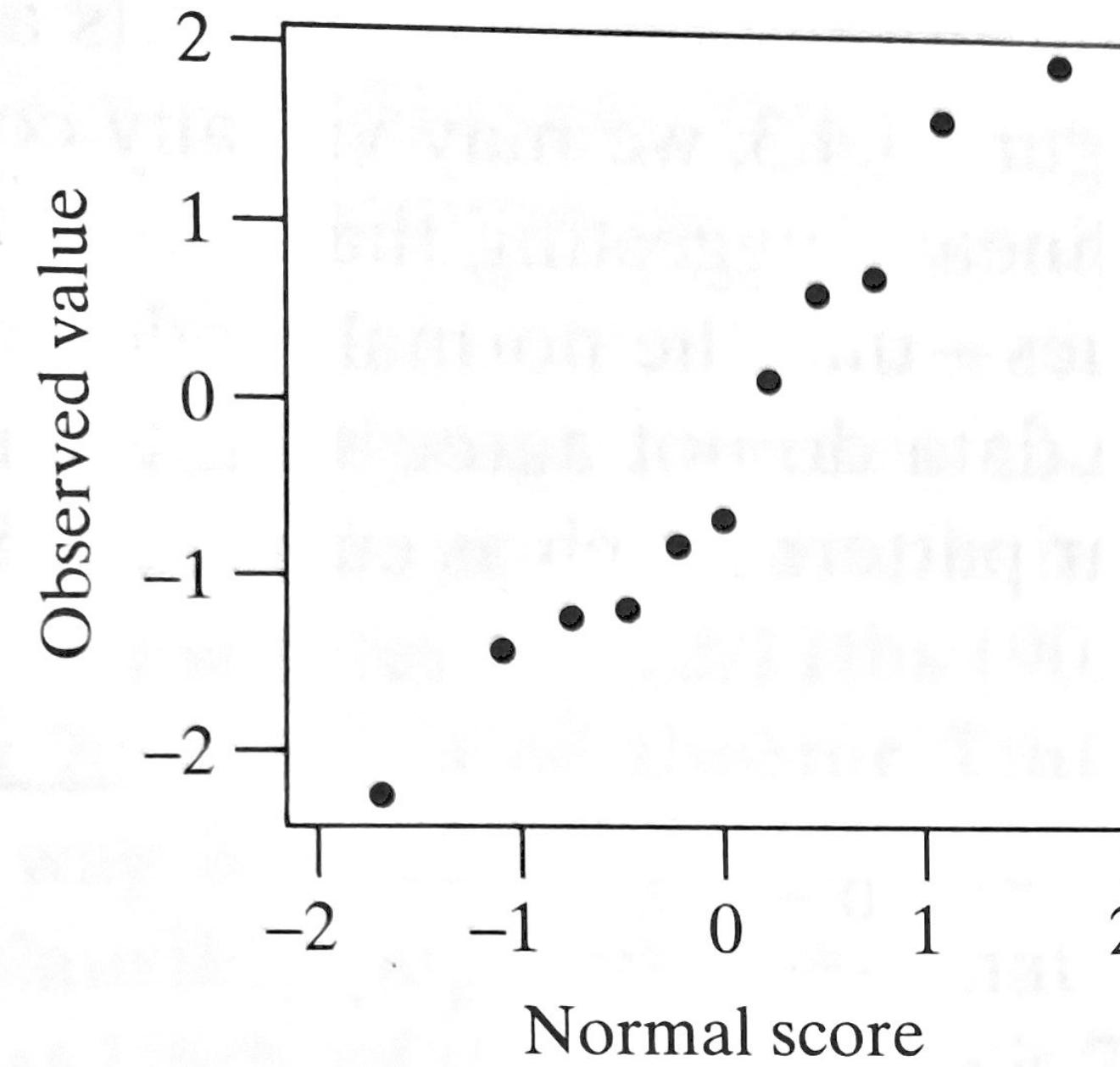
Shapiro-Wilk normality test

data: height
W = 0.98371, p-value = 0.9833

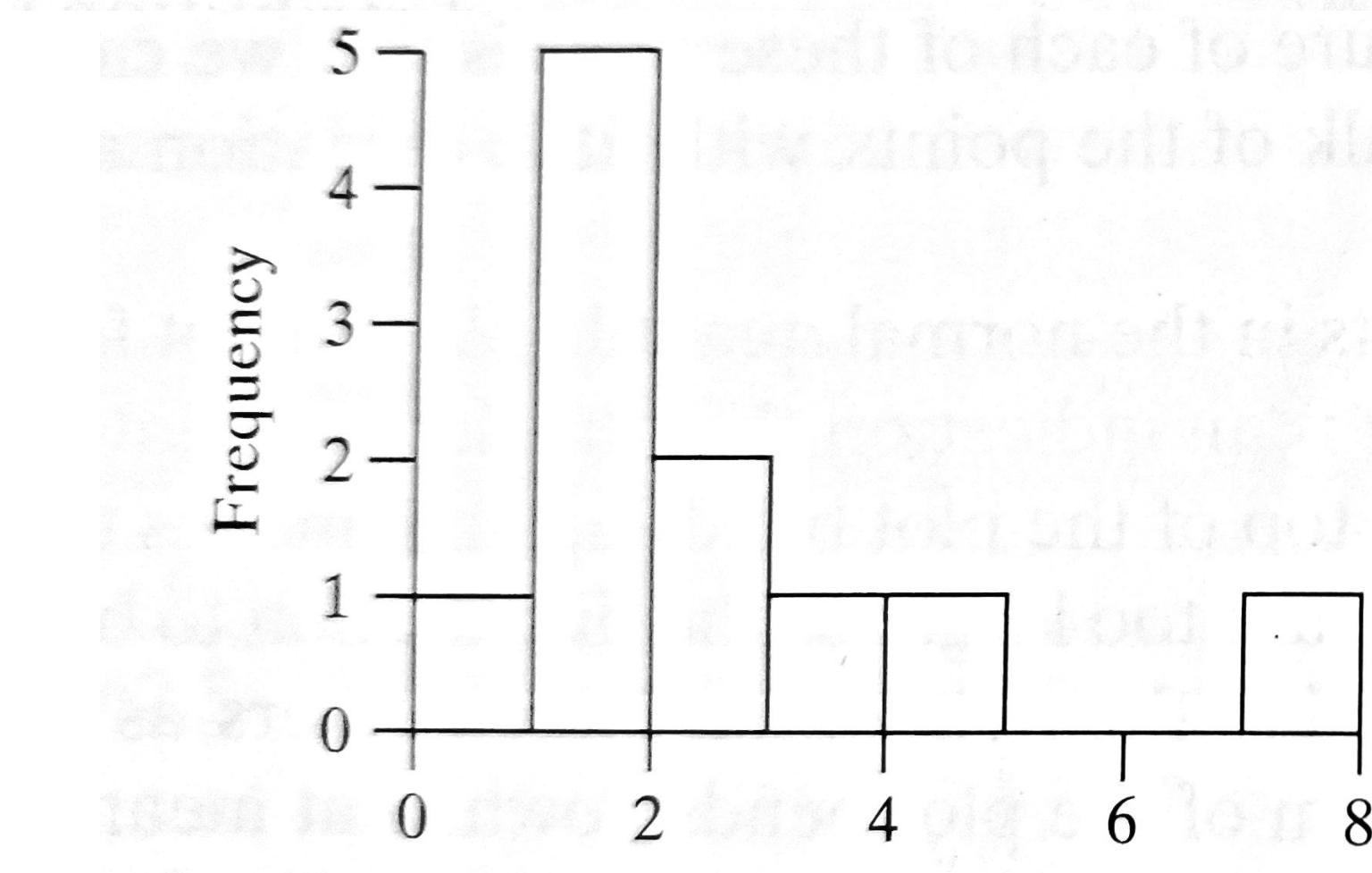
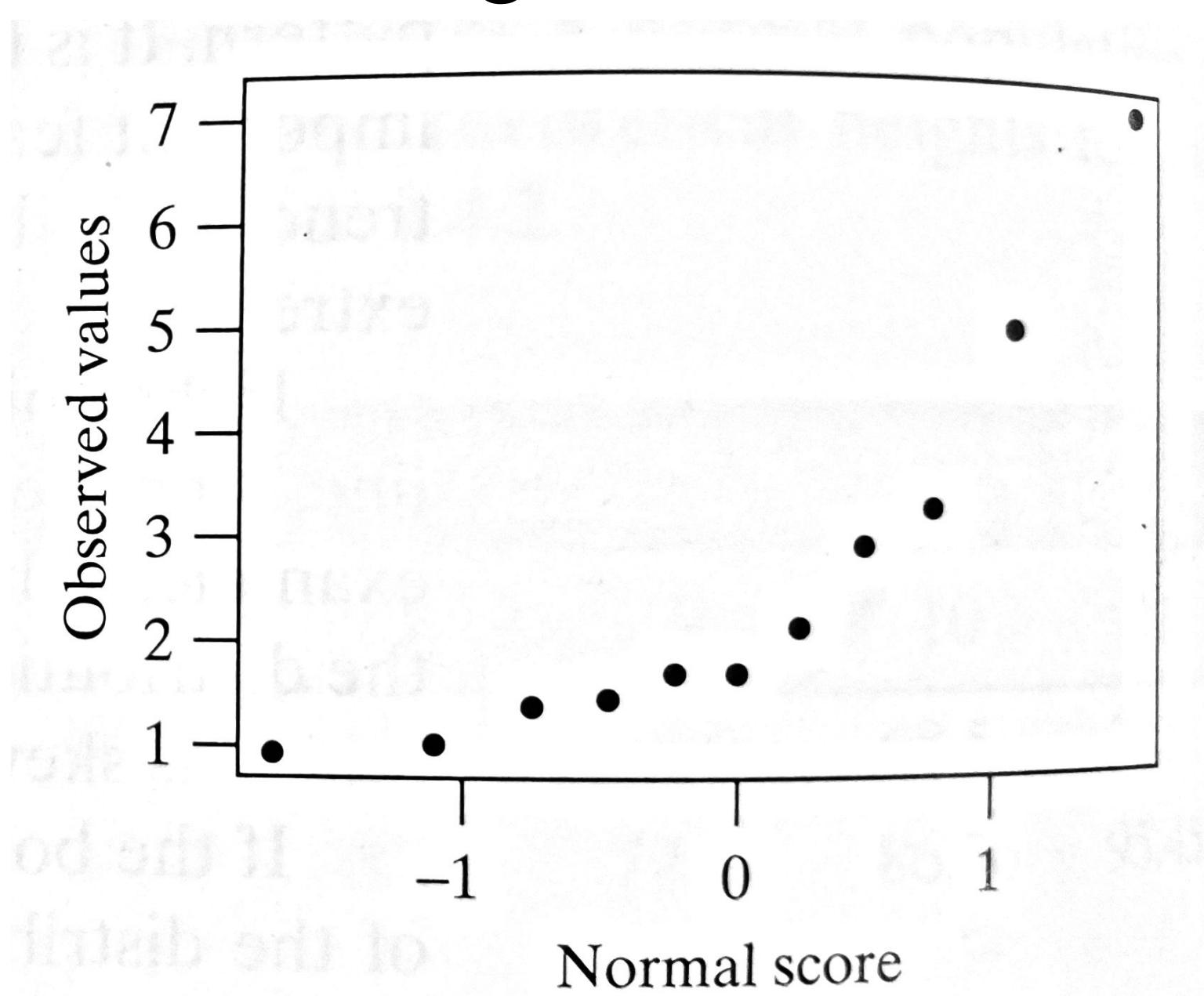
Smaller the p-value = stronger evidence for NON-normality

Larger the p-value, stronger evidence for normality

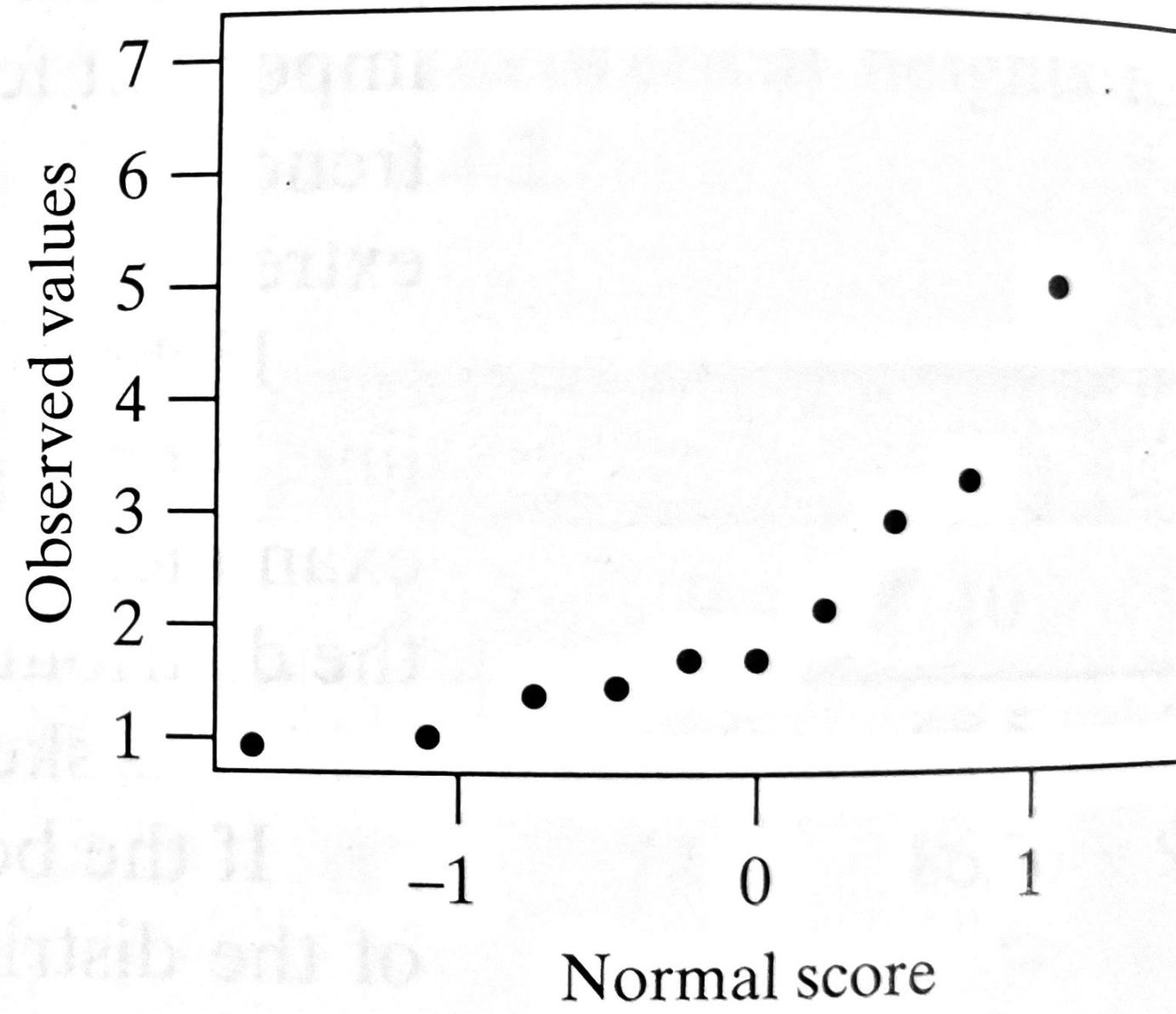




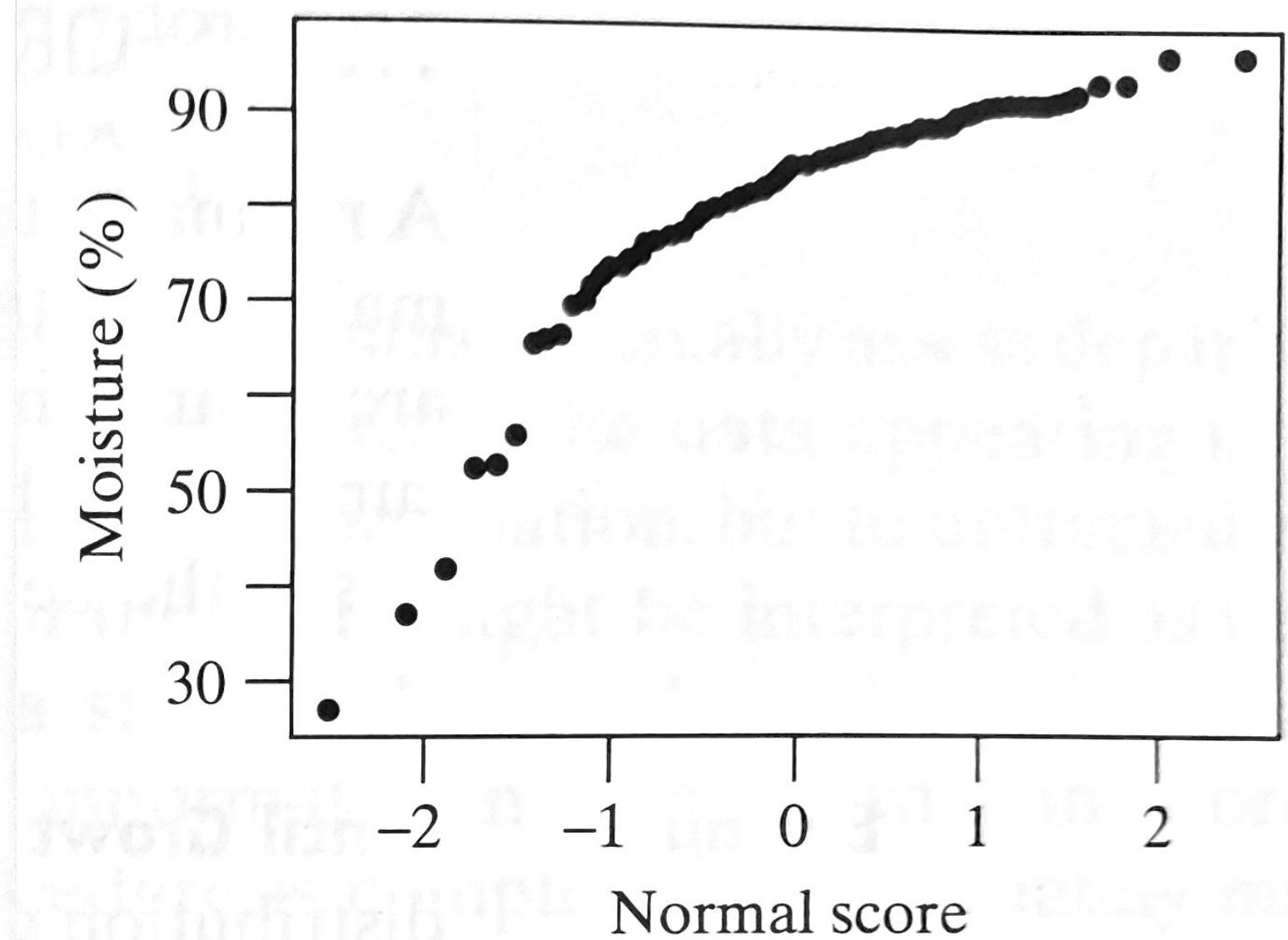
Right skew



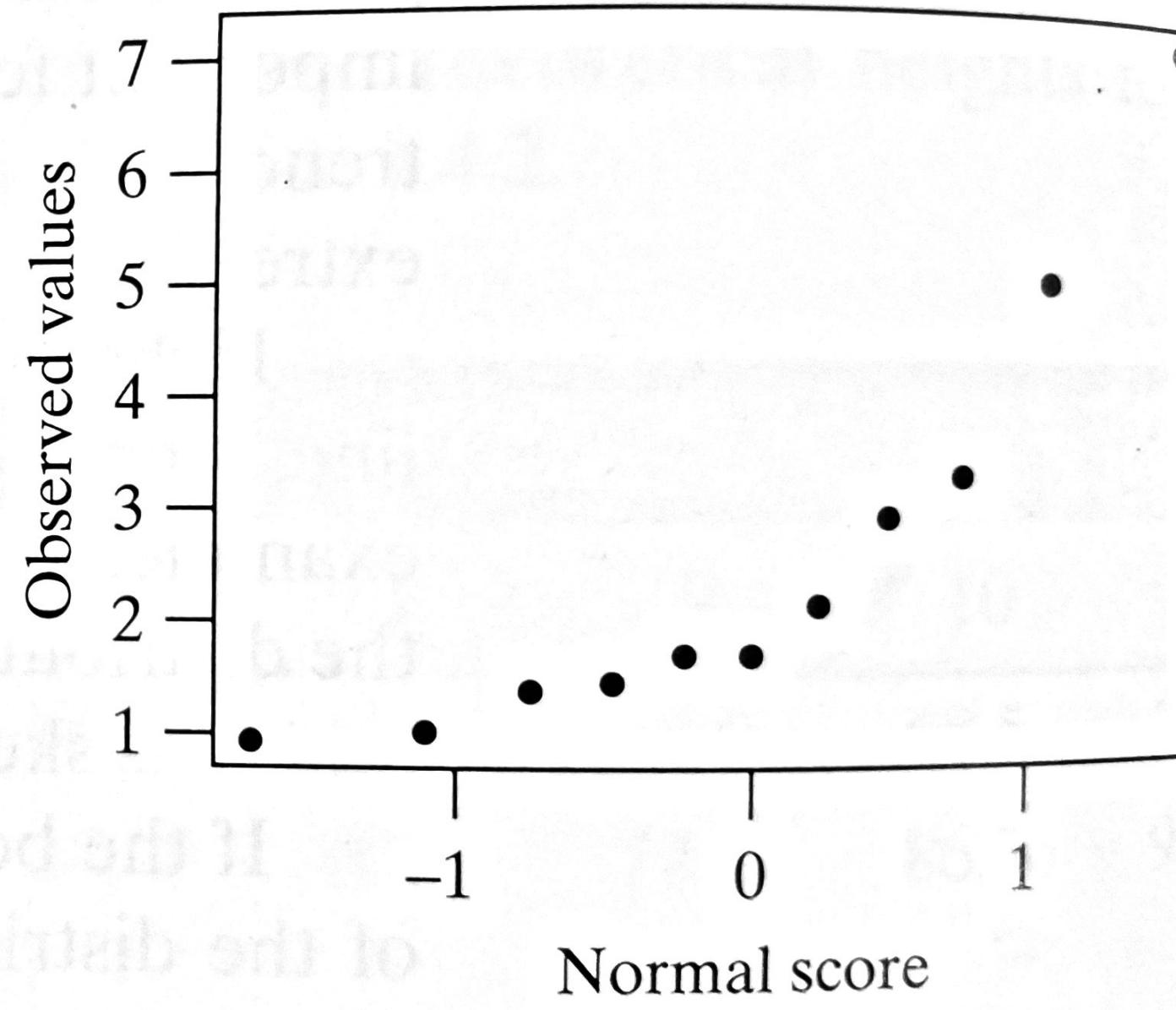
Right skew



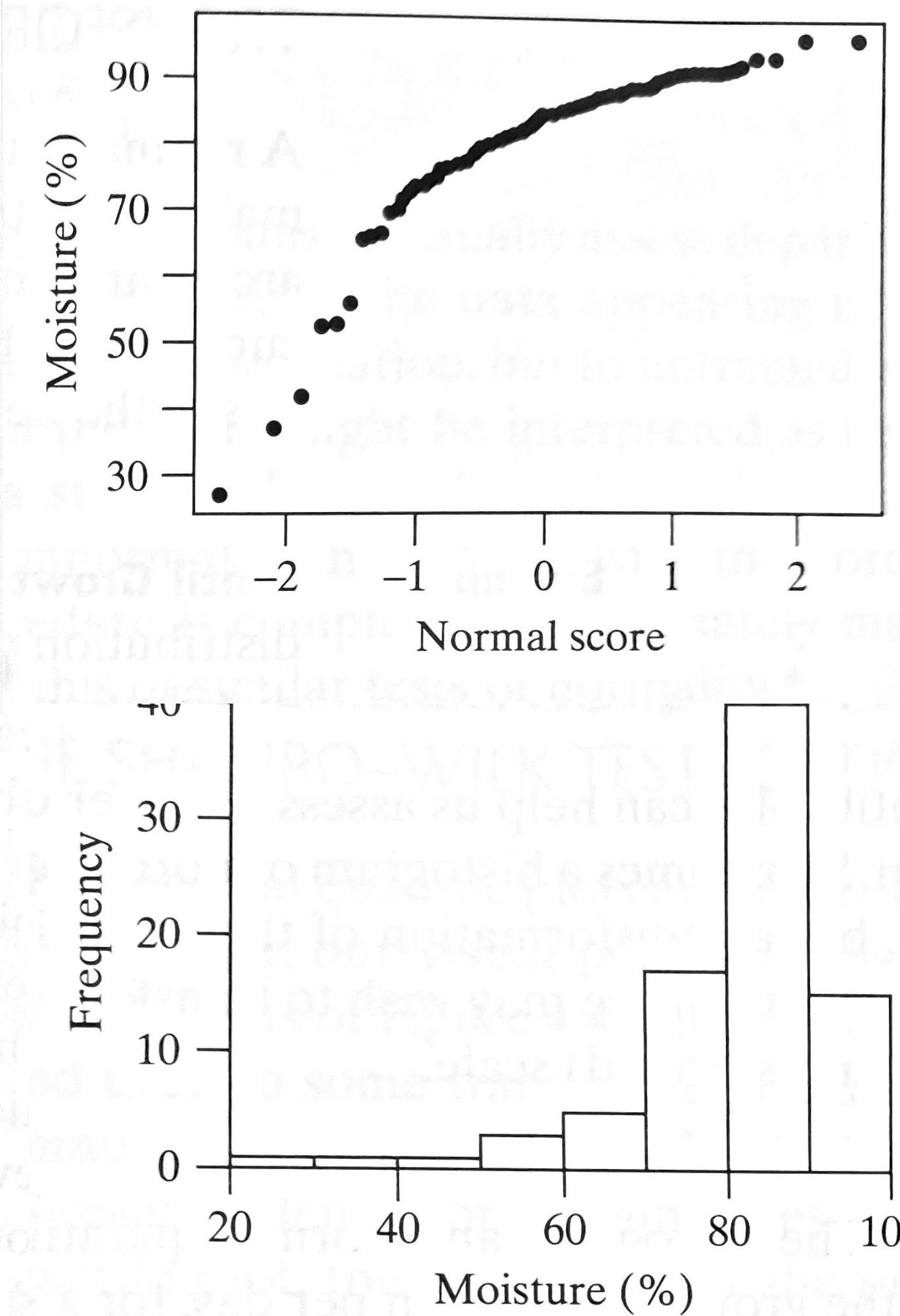
Left skew



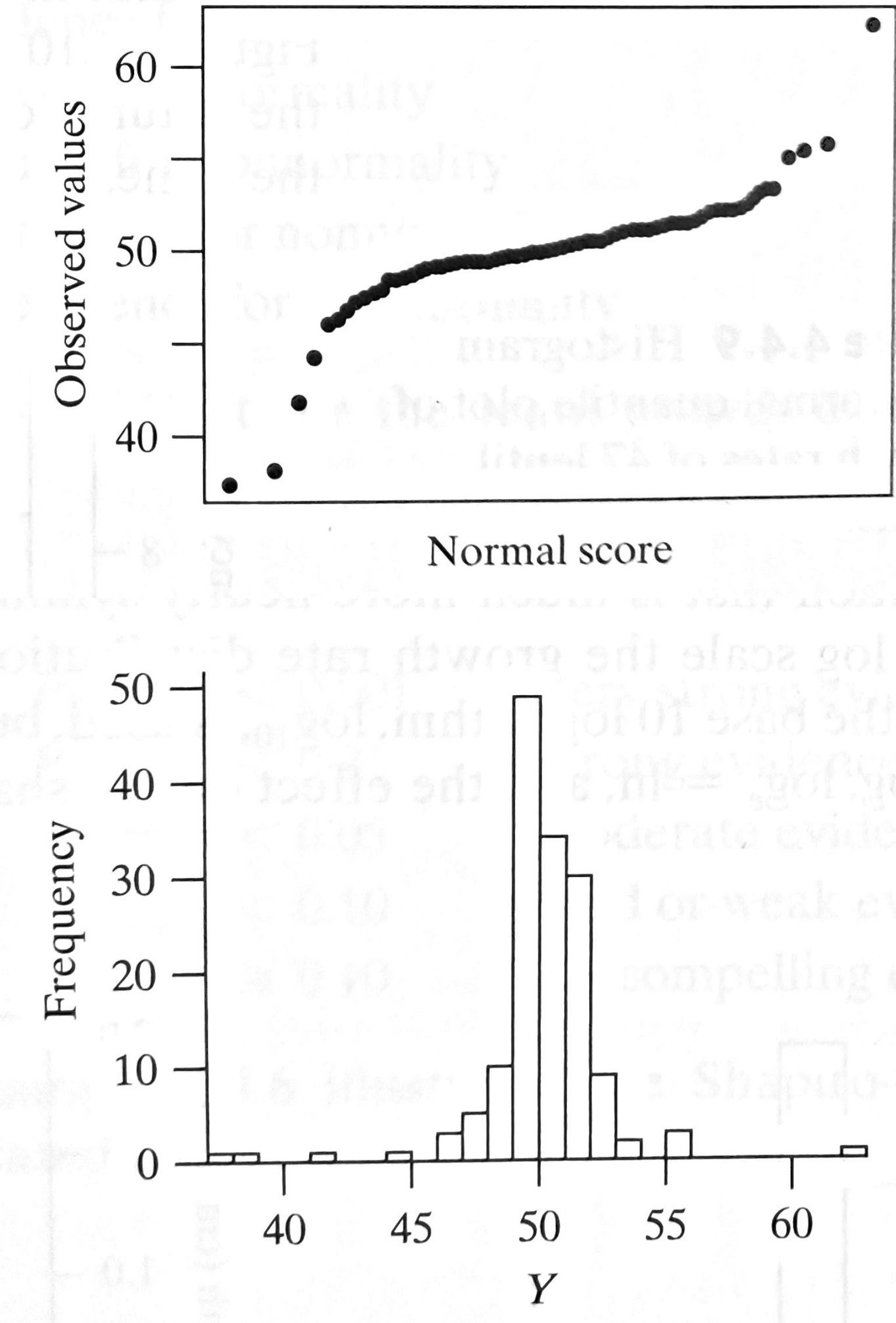
Right skew



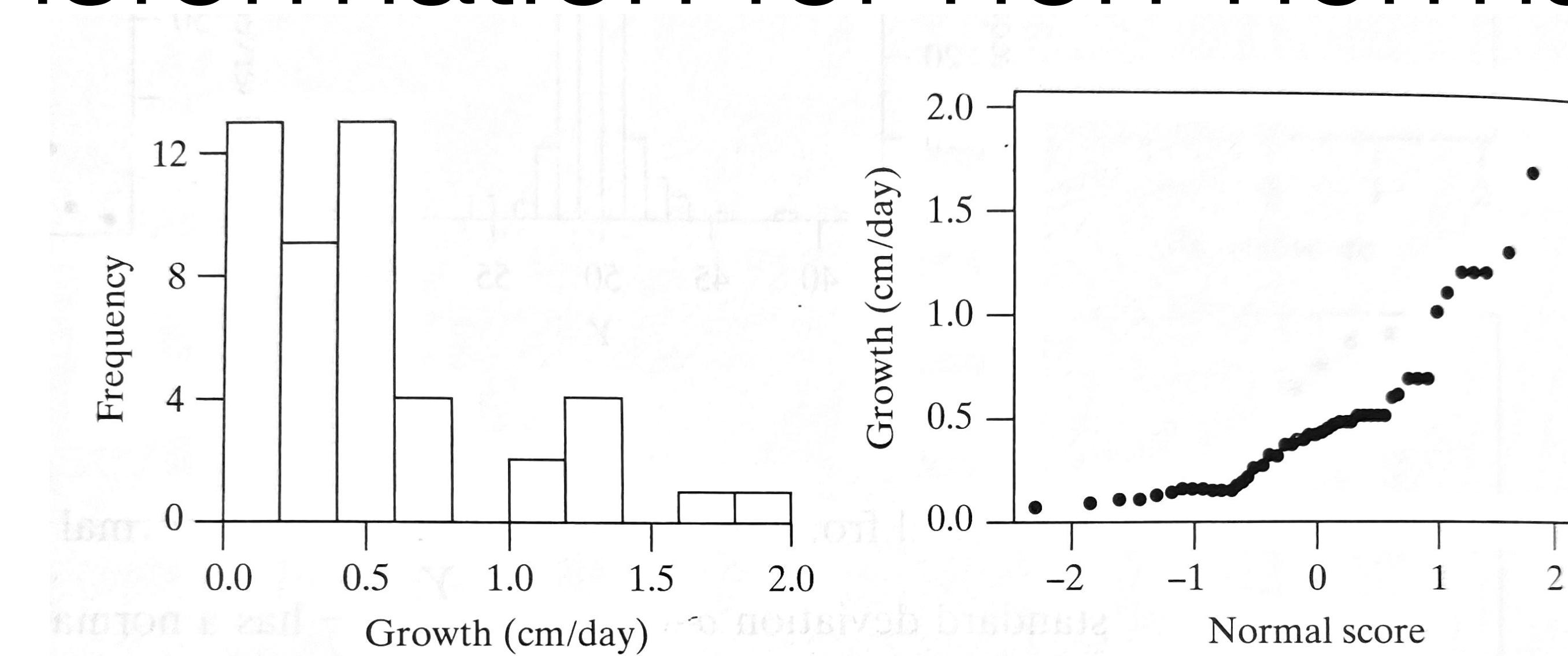
Left skew



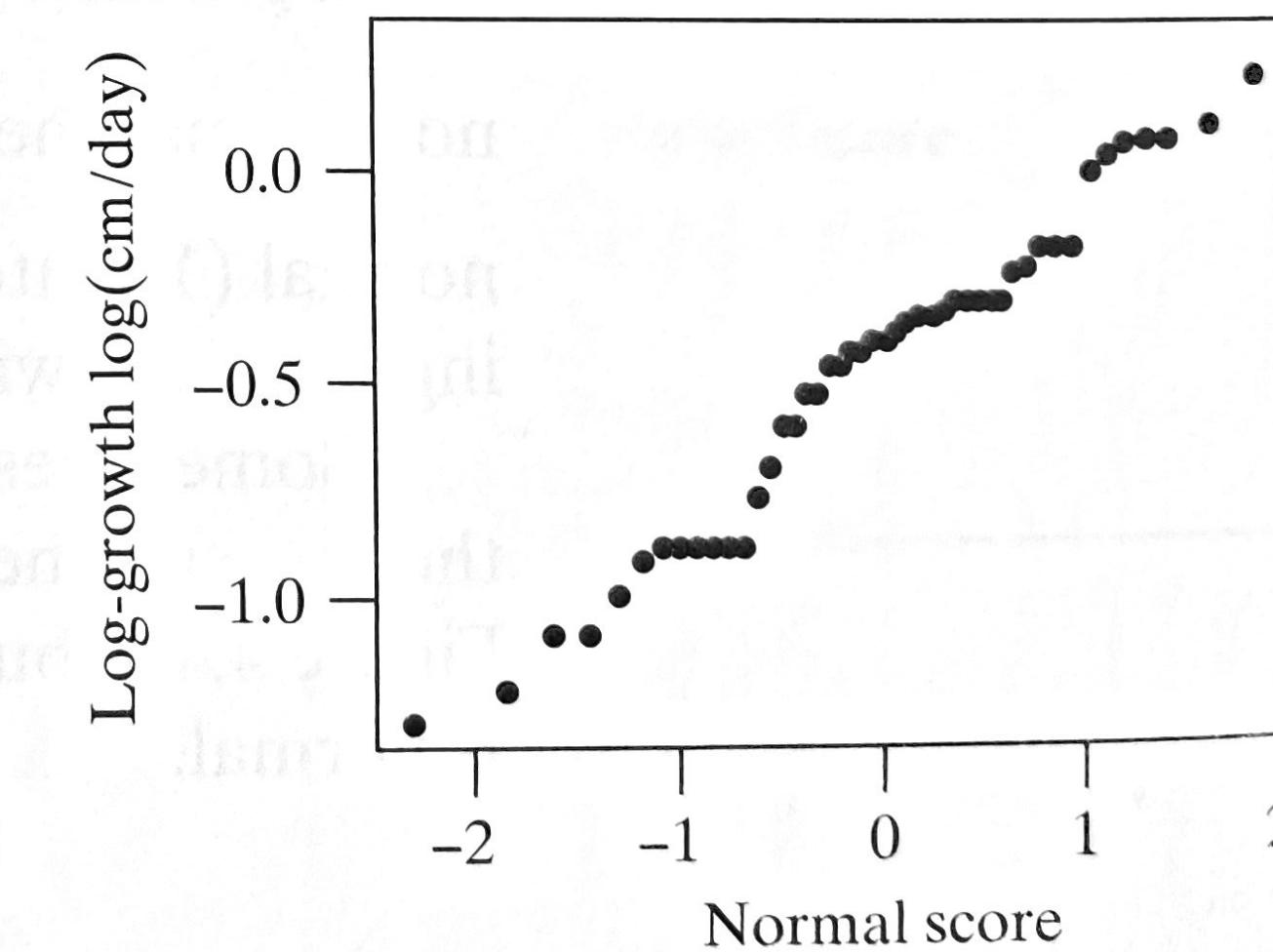
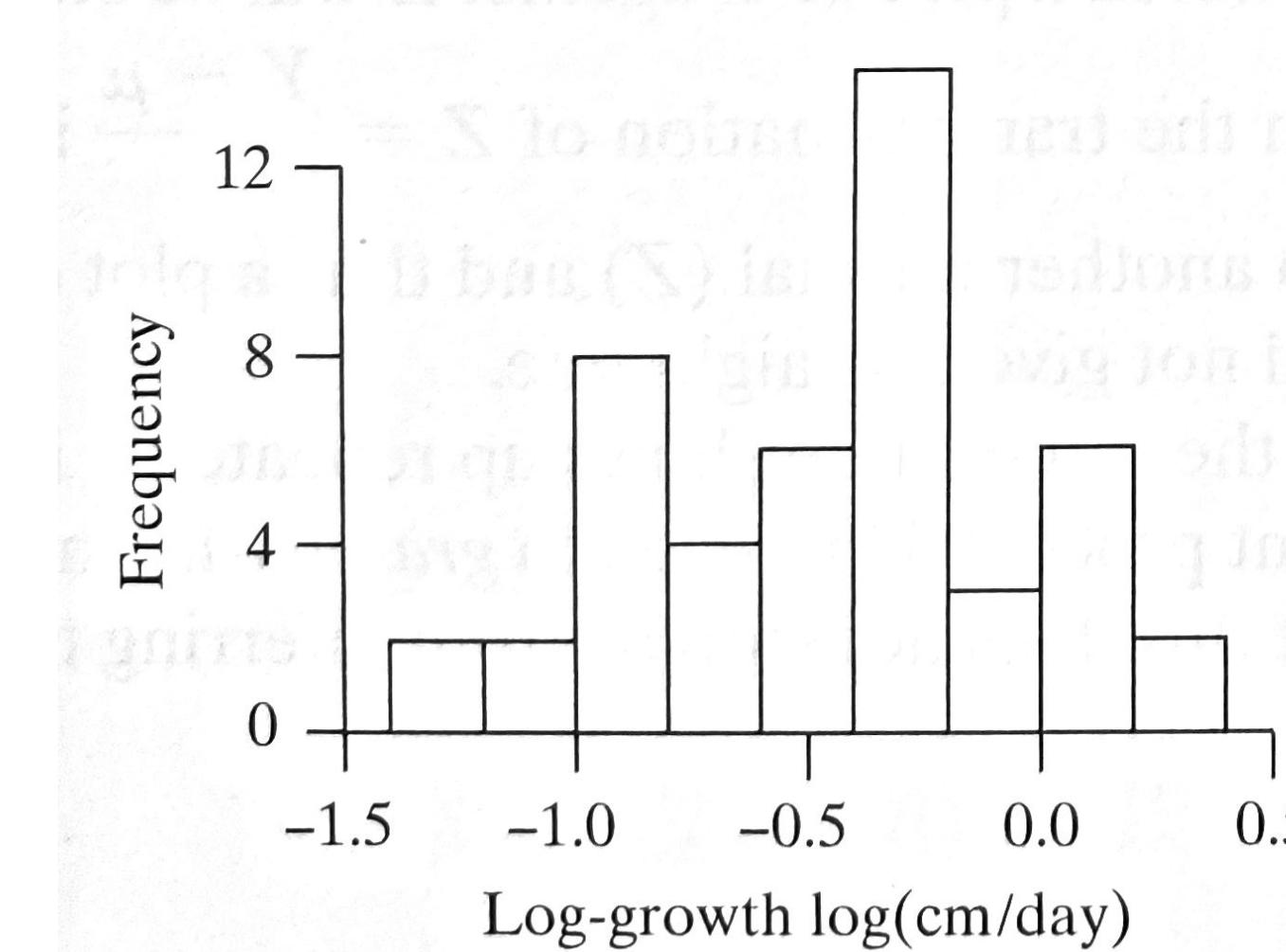
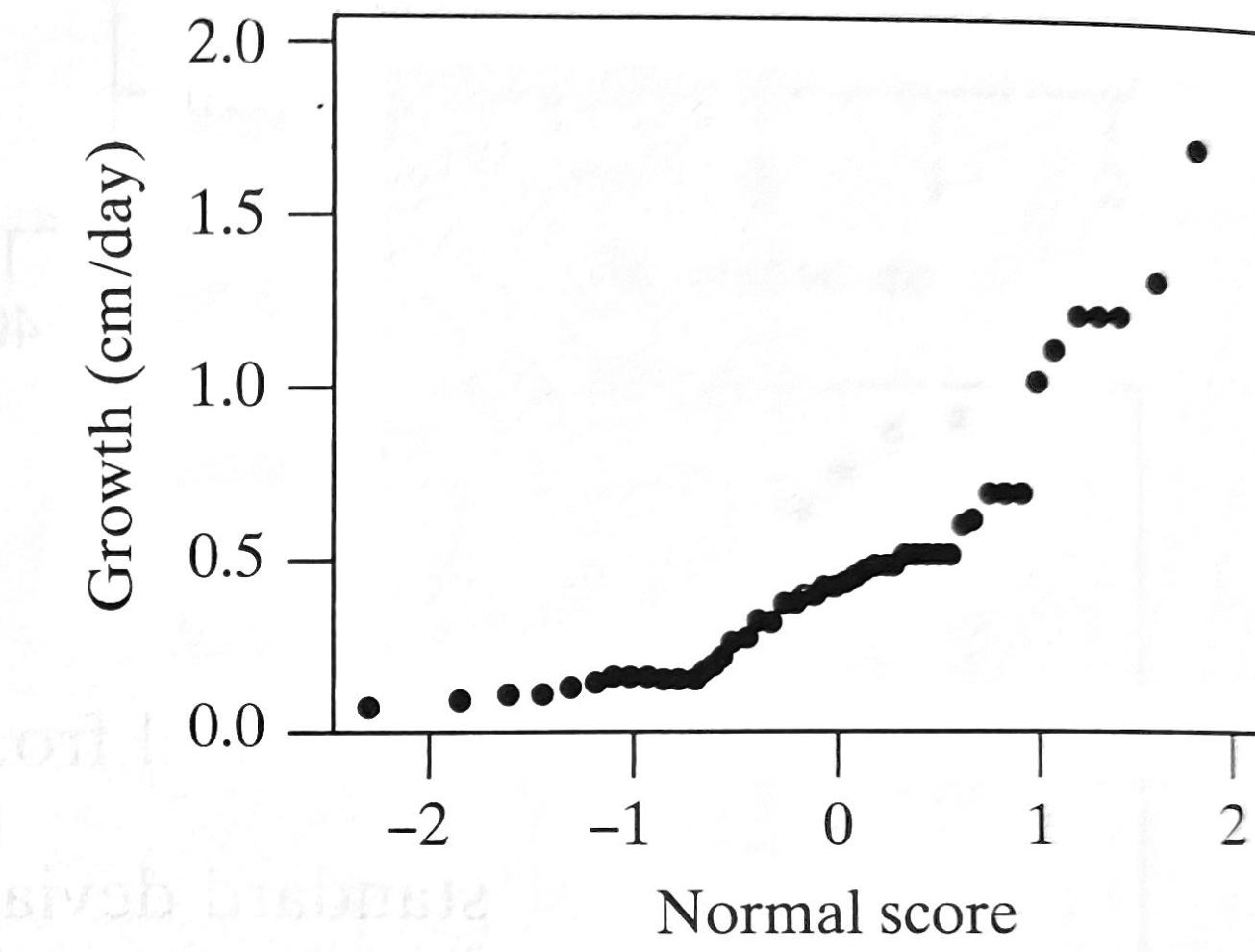
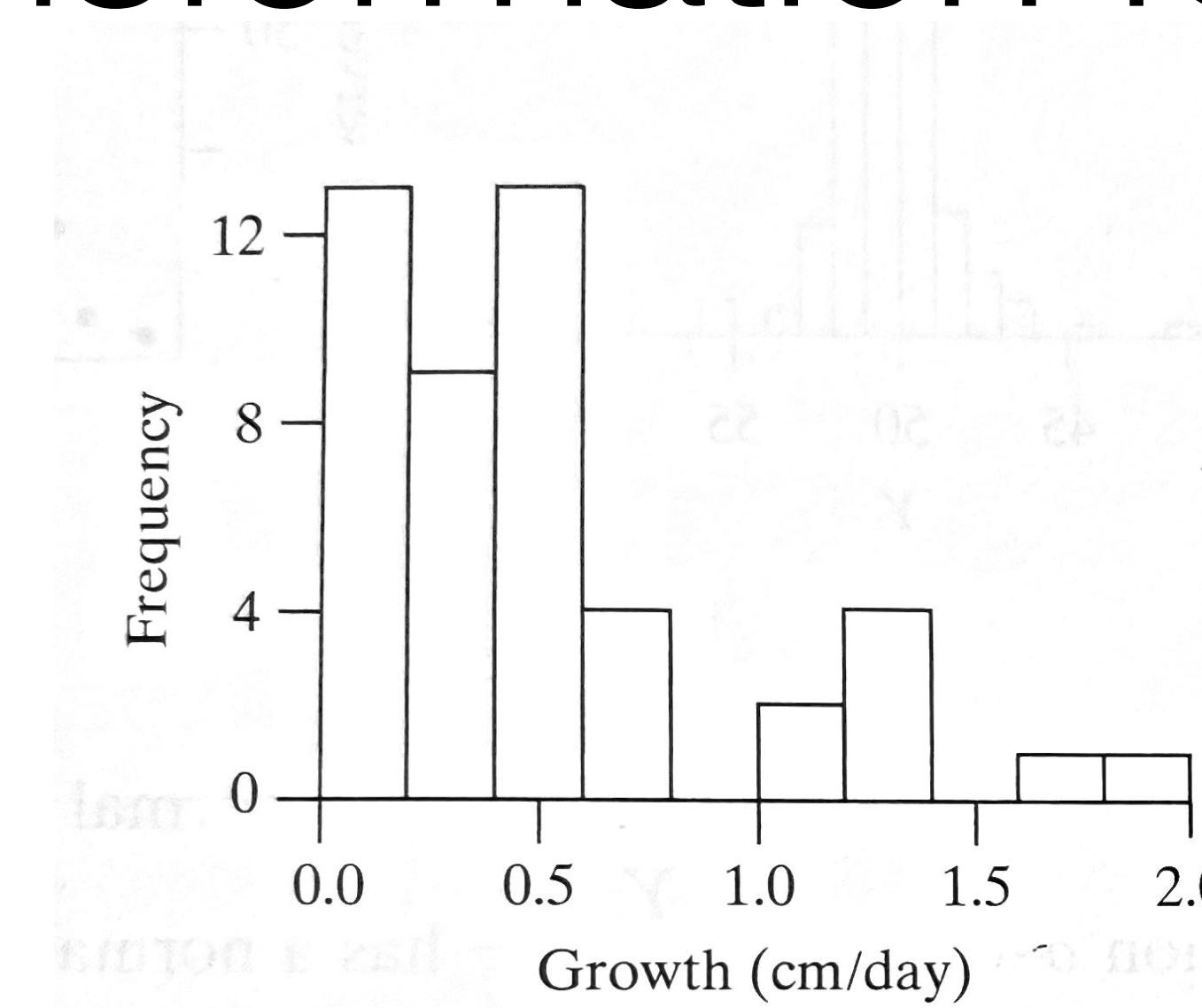
Normal, long tails



Transformation for non-normal data

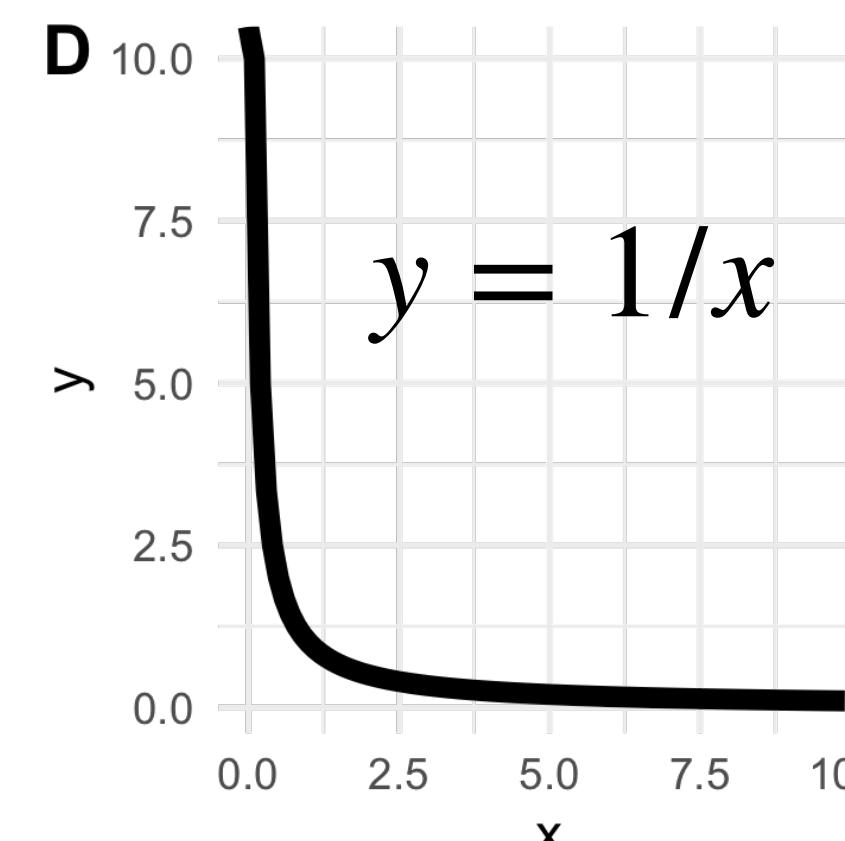
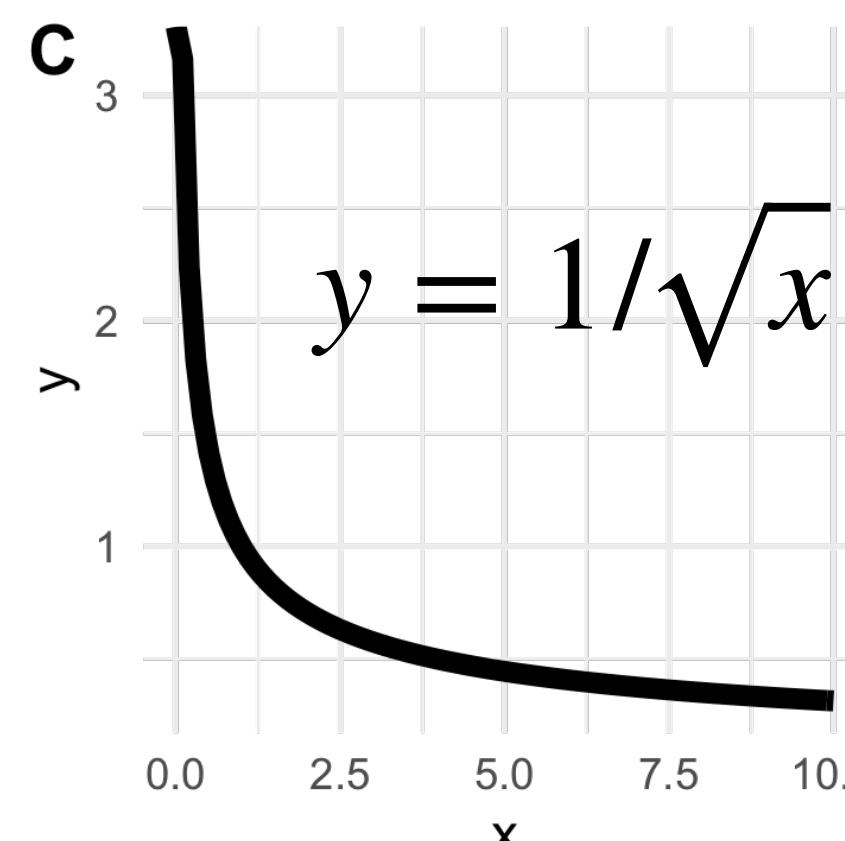
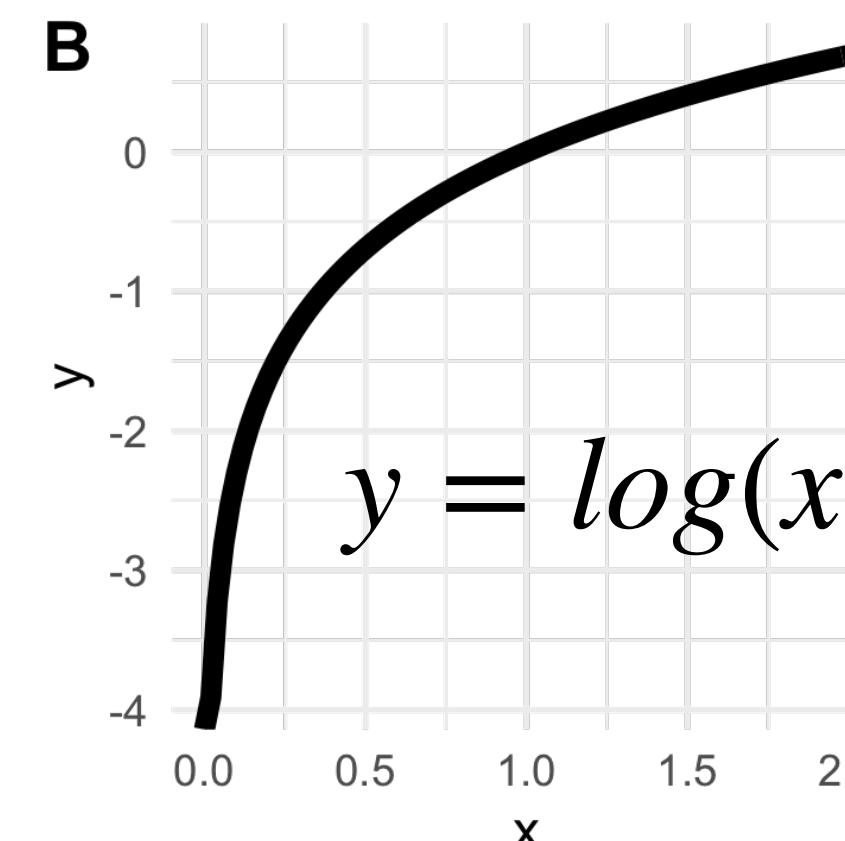
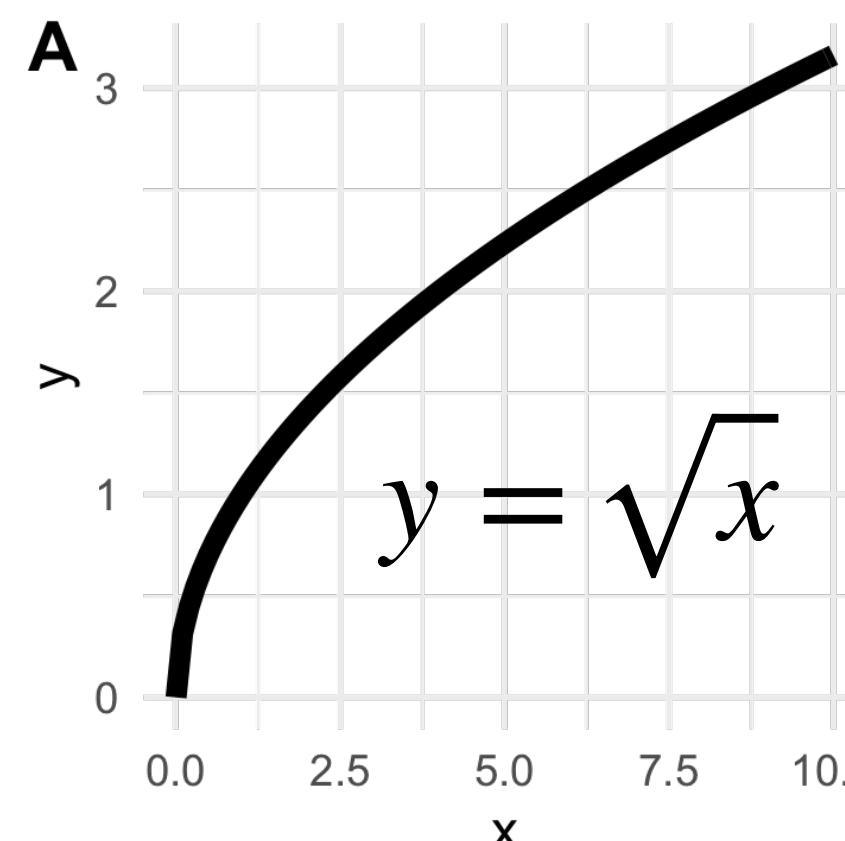


Transformation for non-normal data

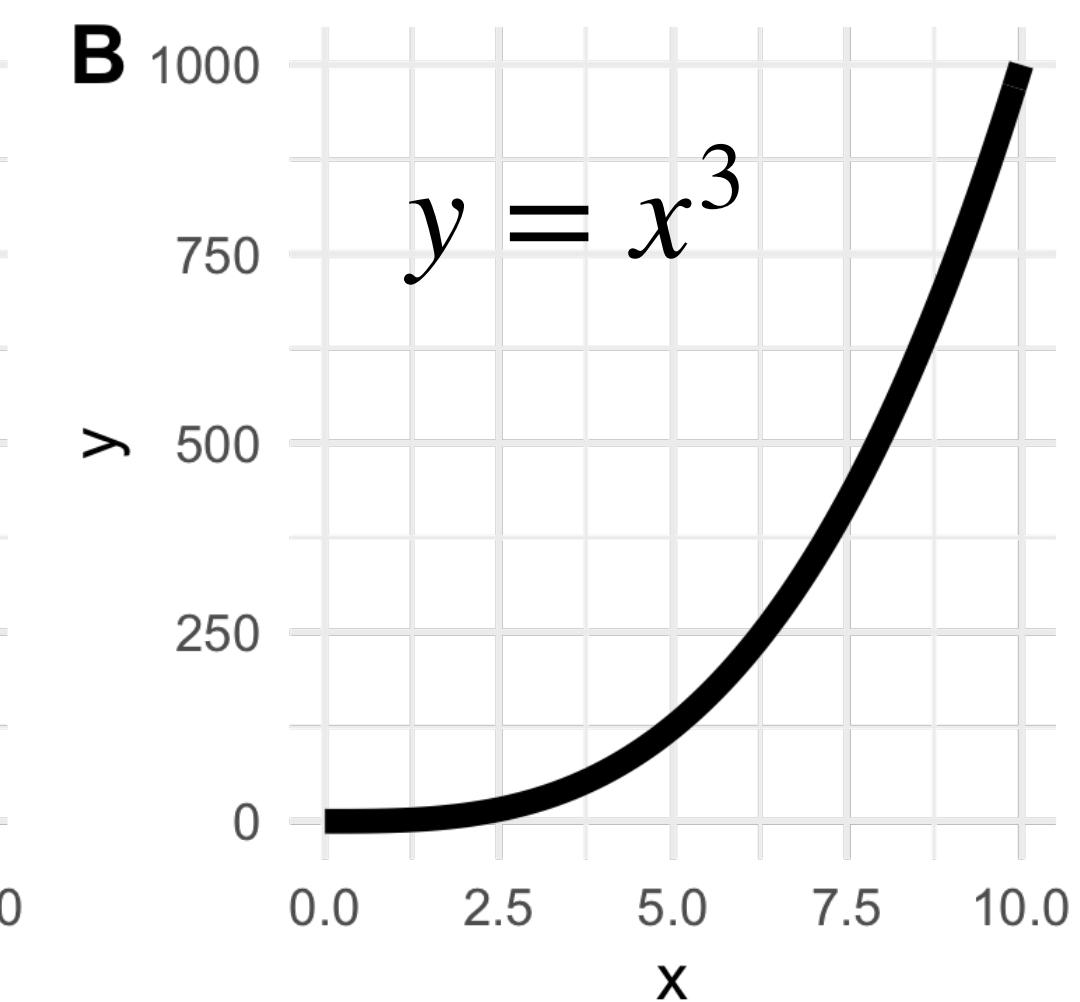
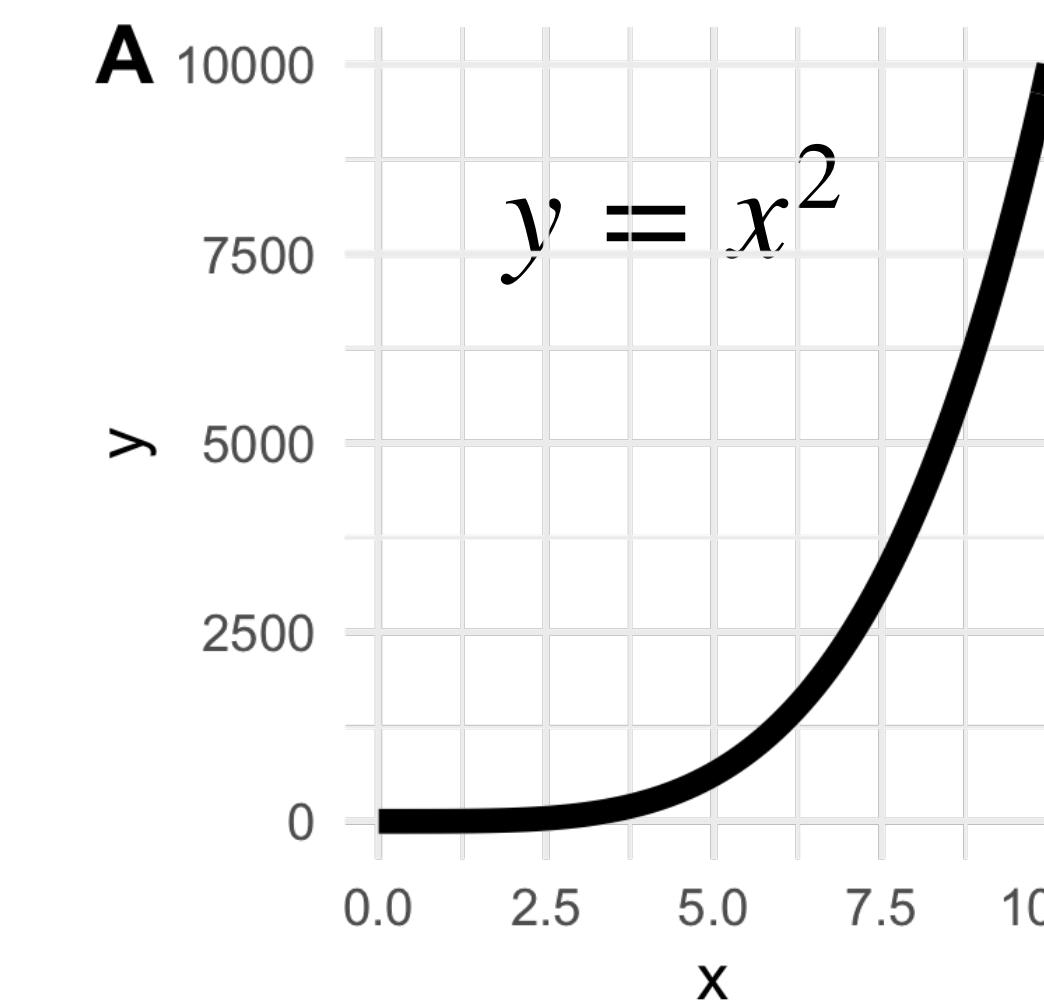


Transformation for non-normal data

RIGHT SKEW



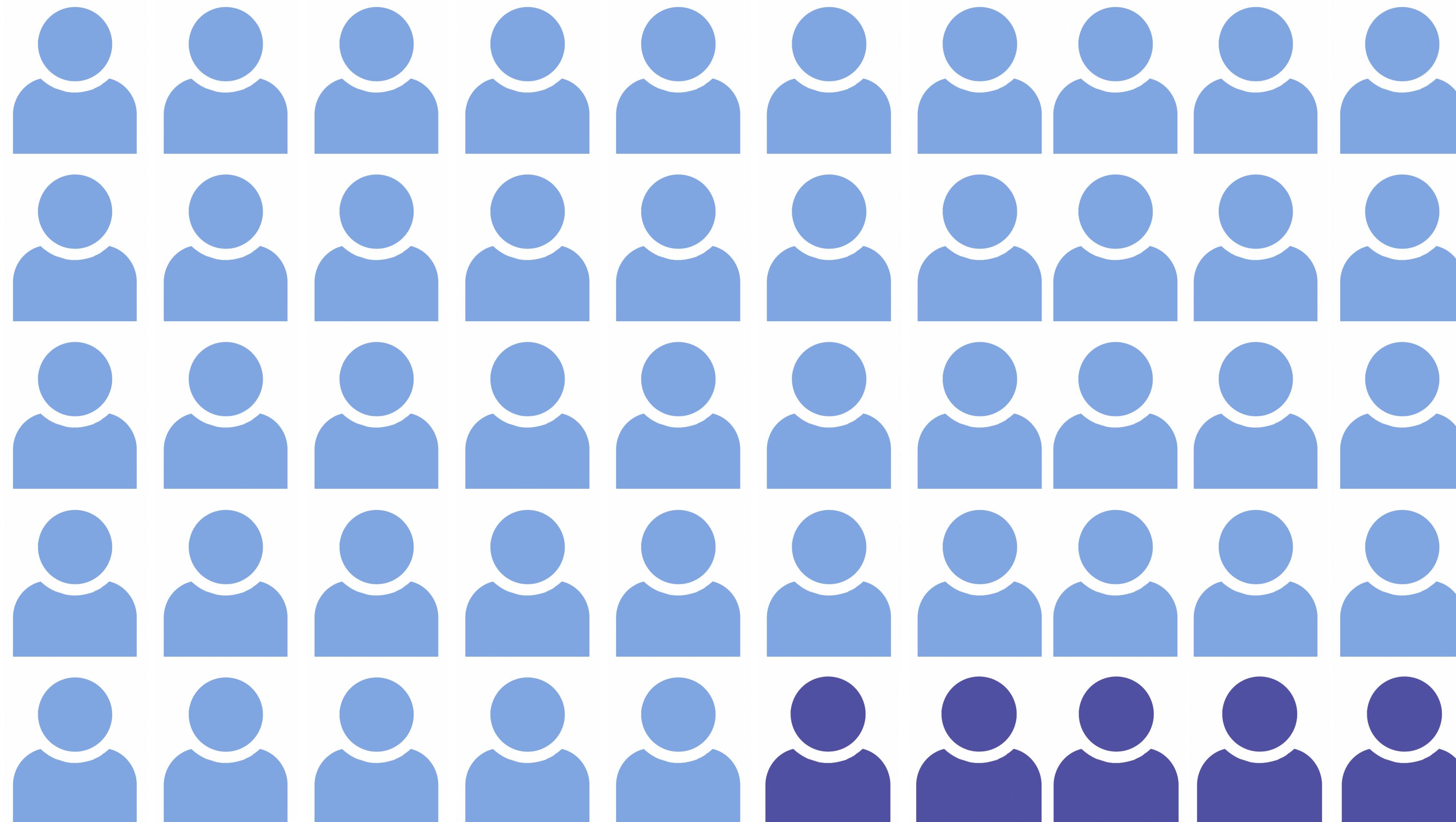
LEFT SKEW



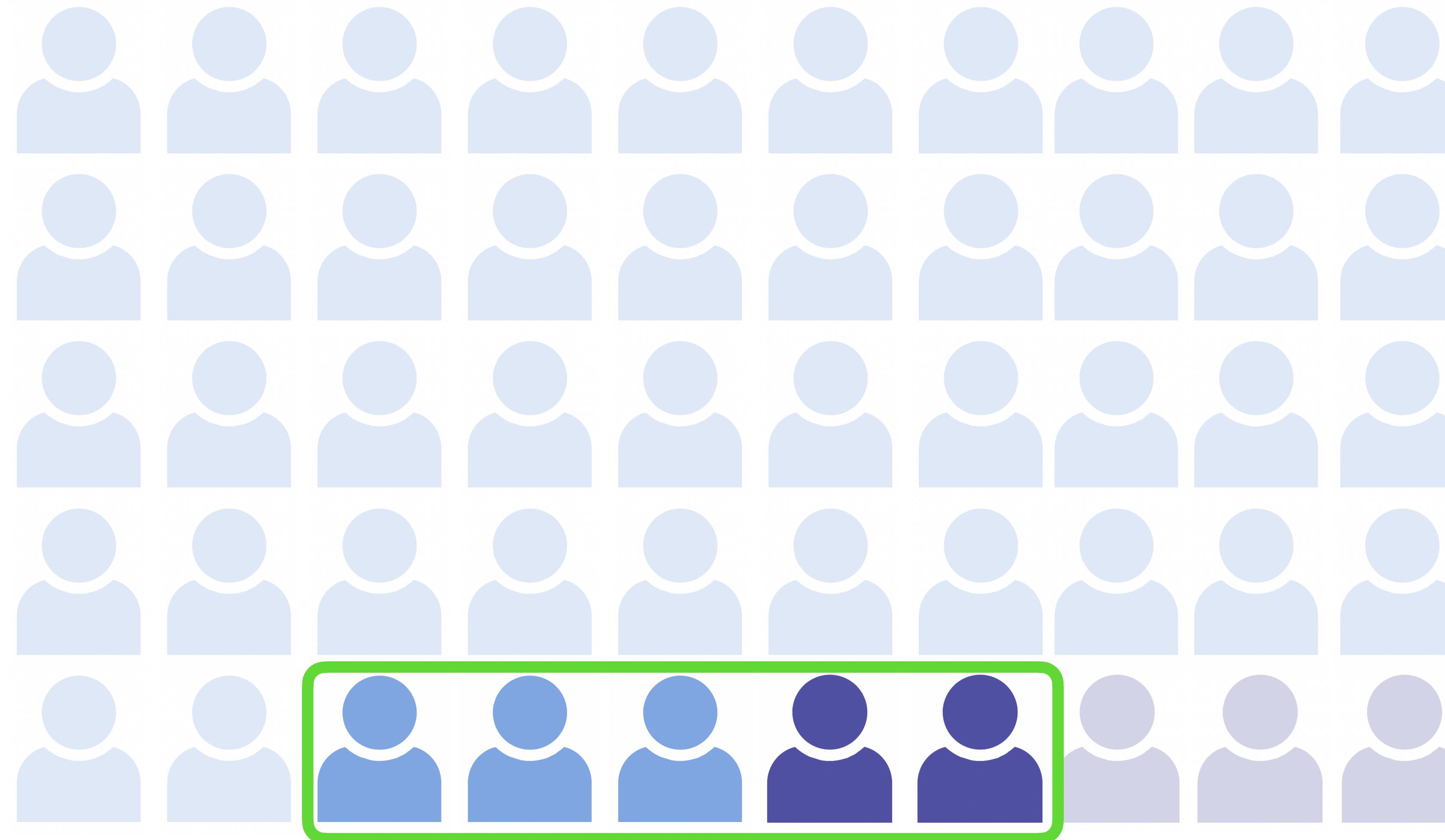
How can we assess ~~normality~~? *Gaussian distribution*

- As we will see, many statistical tests are based on having data from a normal distribution, but how do you know if your data follows a normal distribution?
 - You can make a quantile plot (Q-Q plot) comparing expected and observed values (*straight line = normally distributed*)
 - You can also do a Shapiro-Wilk test of non-normality (*higher p-value = more evidence of a normal distribution*)
- And what should you do if it doesn't?
 - You can transform your data (*i.e. $\log(x)$ for right skewed data and x^2 for left skewed data*)
 - Don't freak out! Common for biological data to not have a normal distribution...
(and we will learn soon that sampling distributions are often normal...)

Population vs. Sample

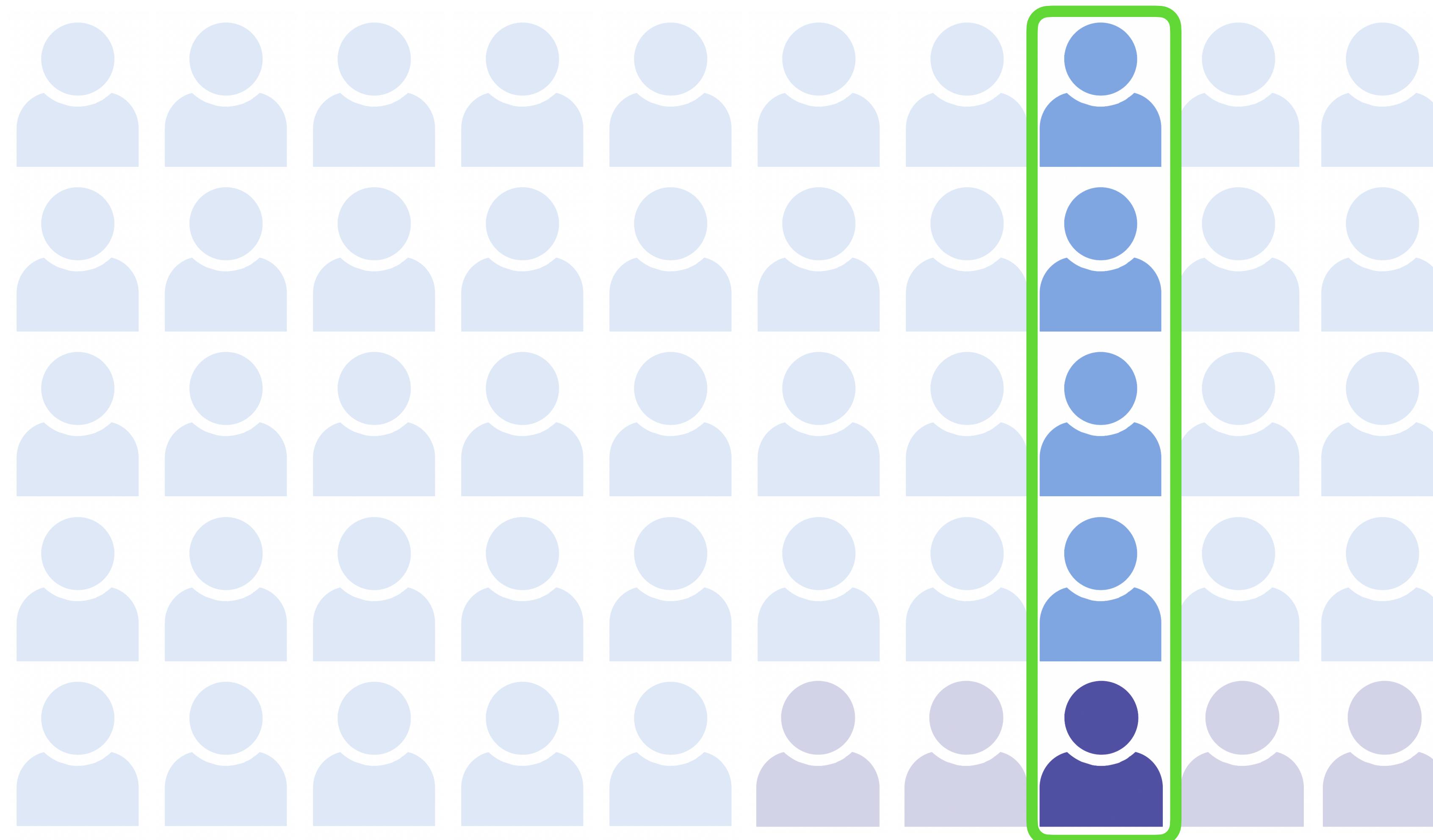


Population vs. Sample



Sample 1: 40%

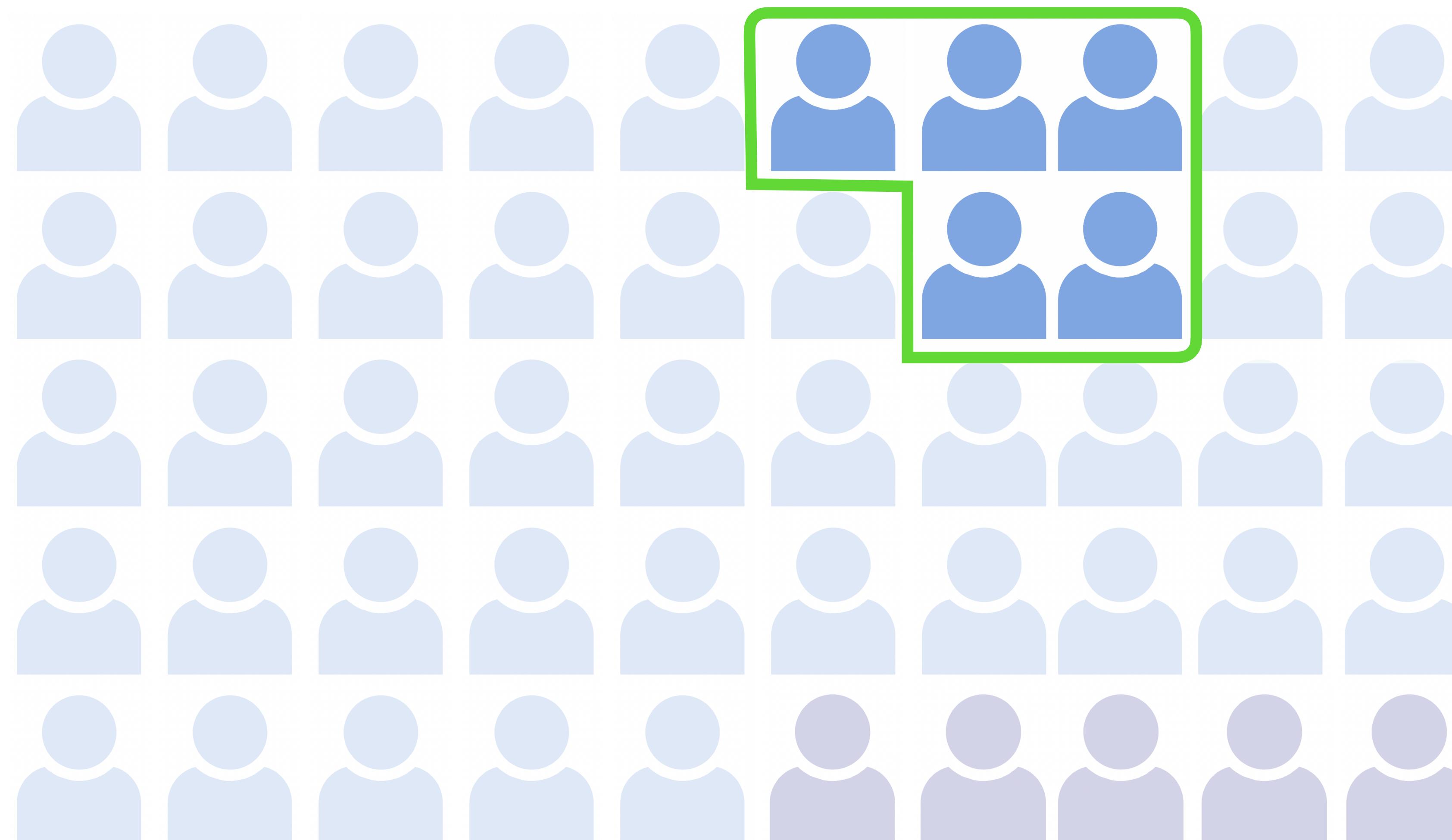
Population vs. Sample



Sample 1: 40%

Sample 2: 10%

Population vs. Sample

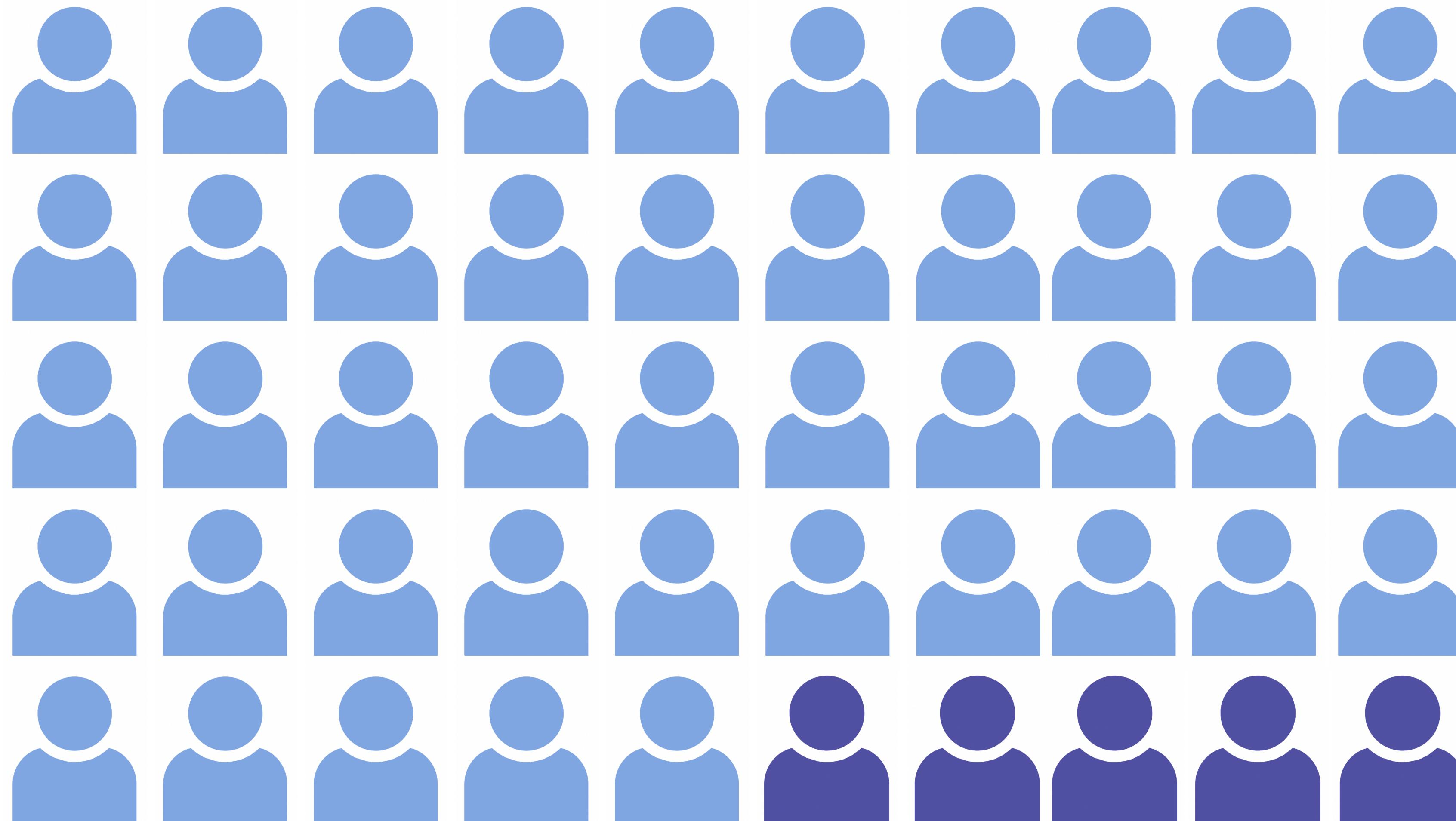


Sample 1: 40%

Sample 2: 10%

Sample 3: 0%

Population vs. Sample



Sample 1: 40%

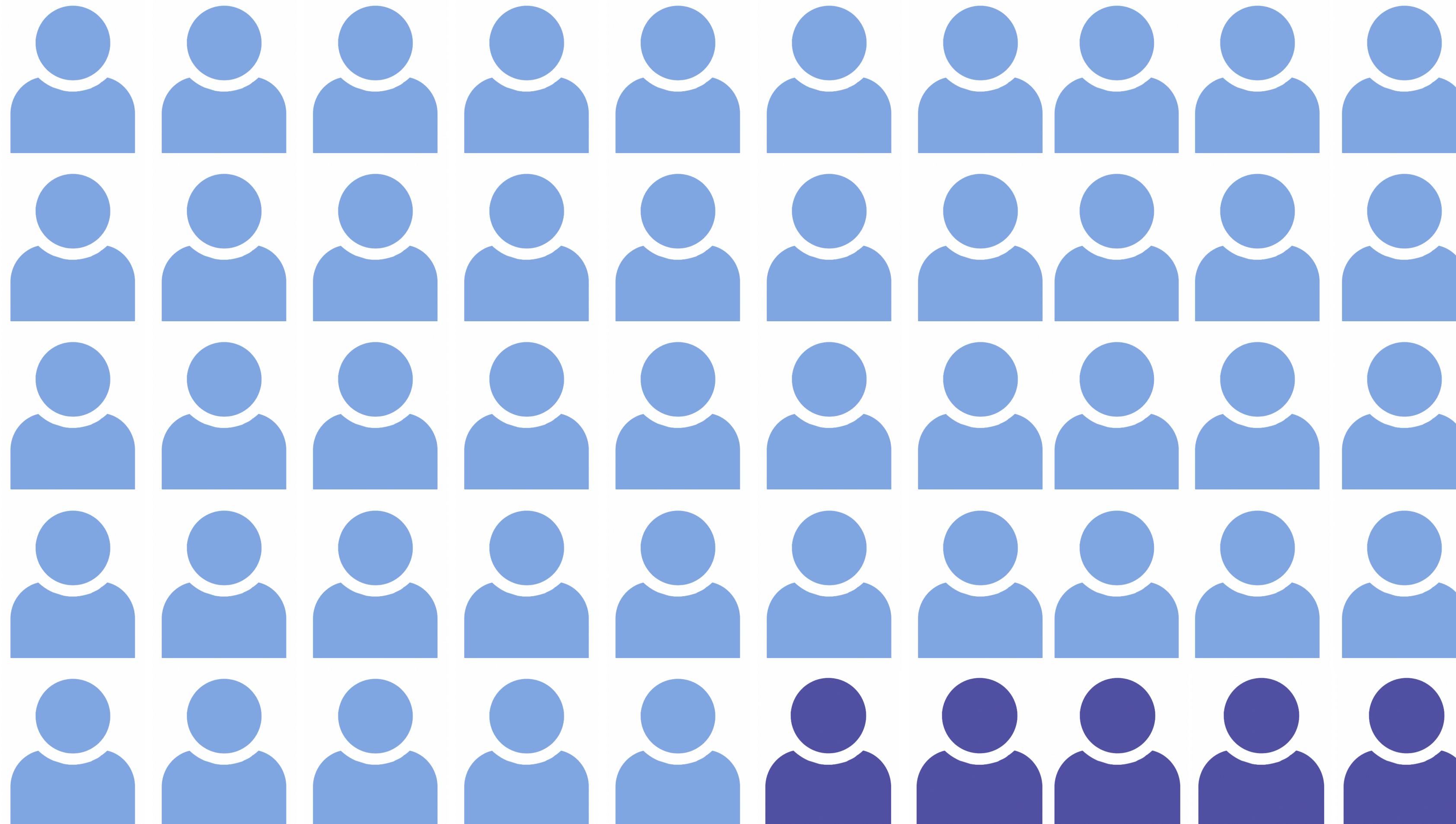
Sample 2: 10%

Sample 3: 0%

Population: 10%

“Sampling variability”

Population vs. Sample



“Meta-study”

Sample 1: 40%

Sample 2: 10%

Sample 3: 0%

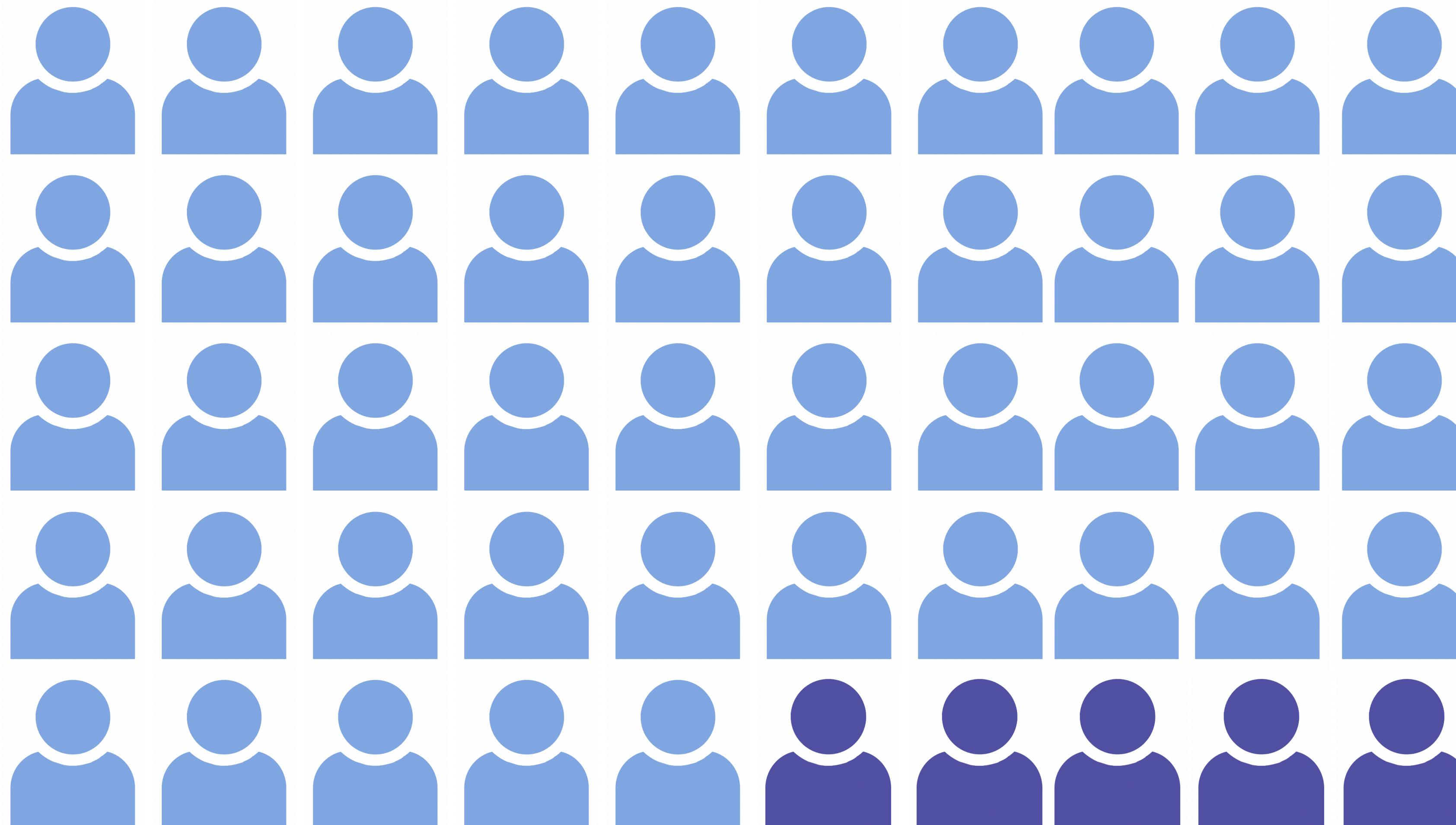
... ?

Sample n: 10%

Population: 10%

($50C5$)

Population vs. Sample



“Meta-study”

Sample 1: 40%

Sample 2: 10%

Sample 3: 0%

...
 $(50C5)$

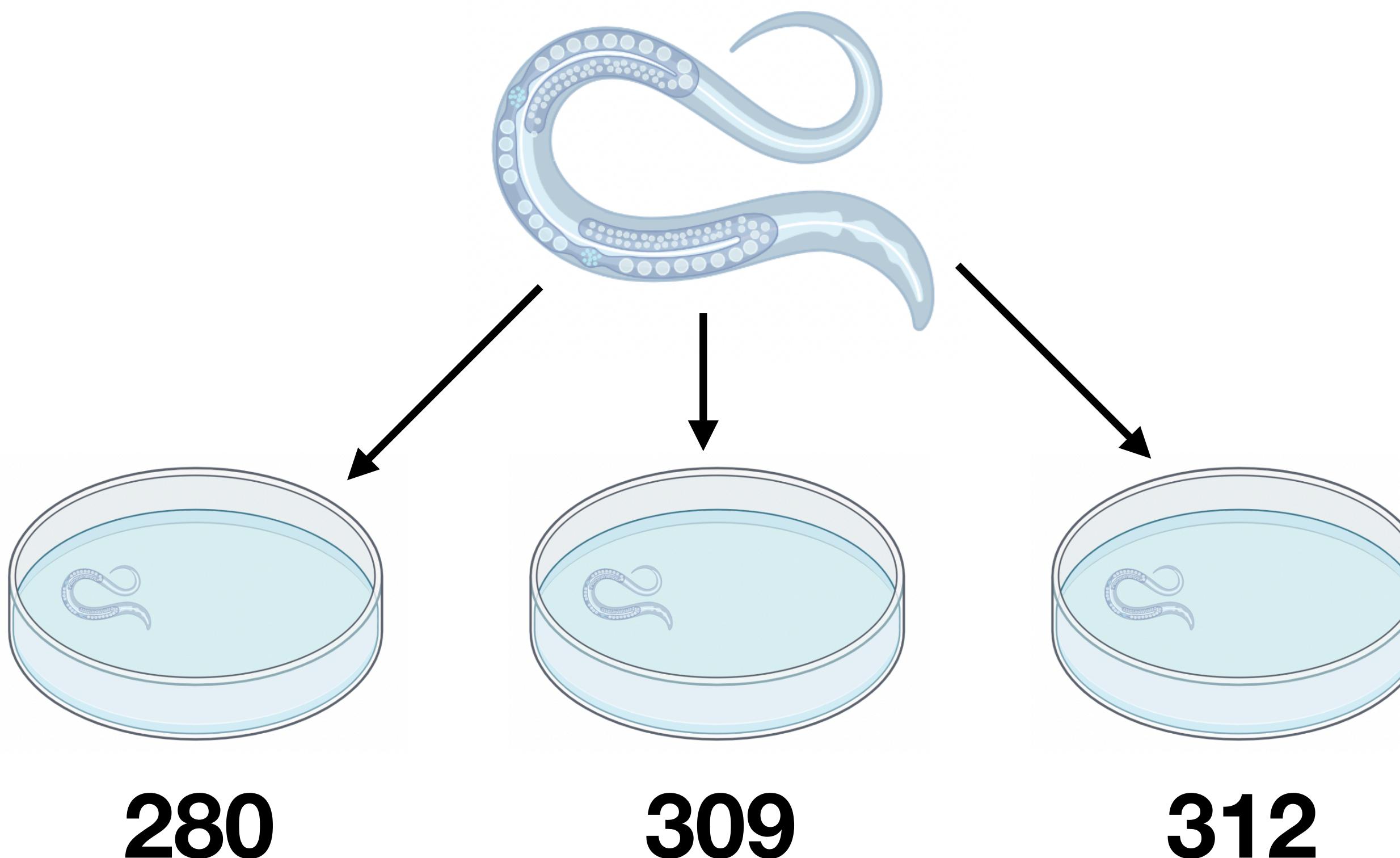
Sample n: 10%

Sample mean: 10%

Probabilities of random sample = relative frequencies in meta-study

Population vs. Sample

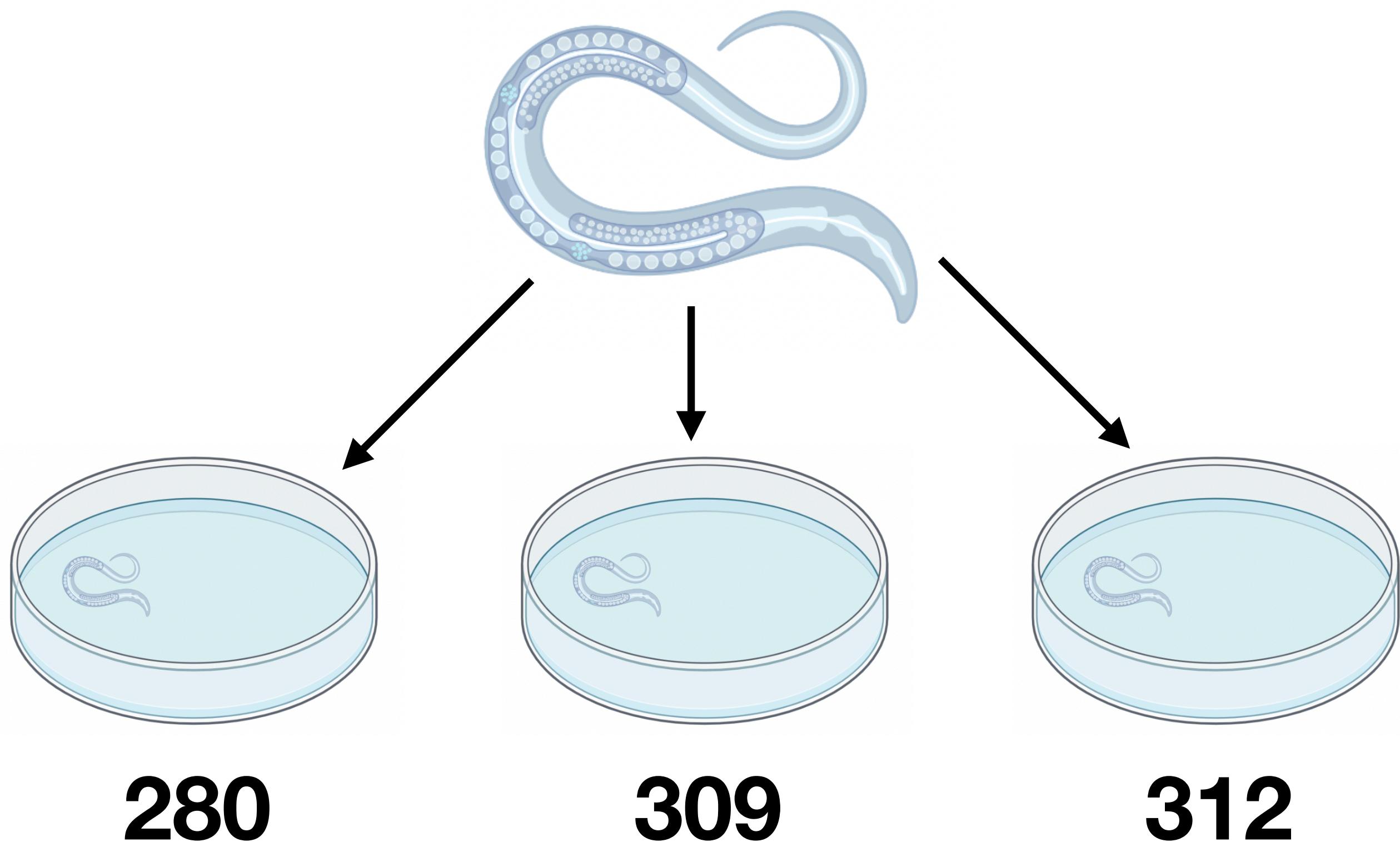
On average, the laboratory strain of *C. elegans* has 300 progeny. You pick three individuals each to their own plate and choose two plates to count broods. What is the sampling distribution?



$$(1/3)(1/2)^2 \text{ ways} = 1/3$$

Population vs. Sample

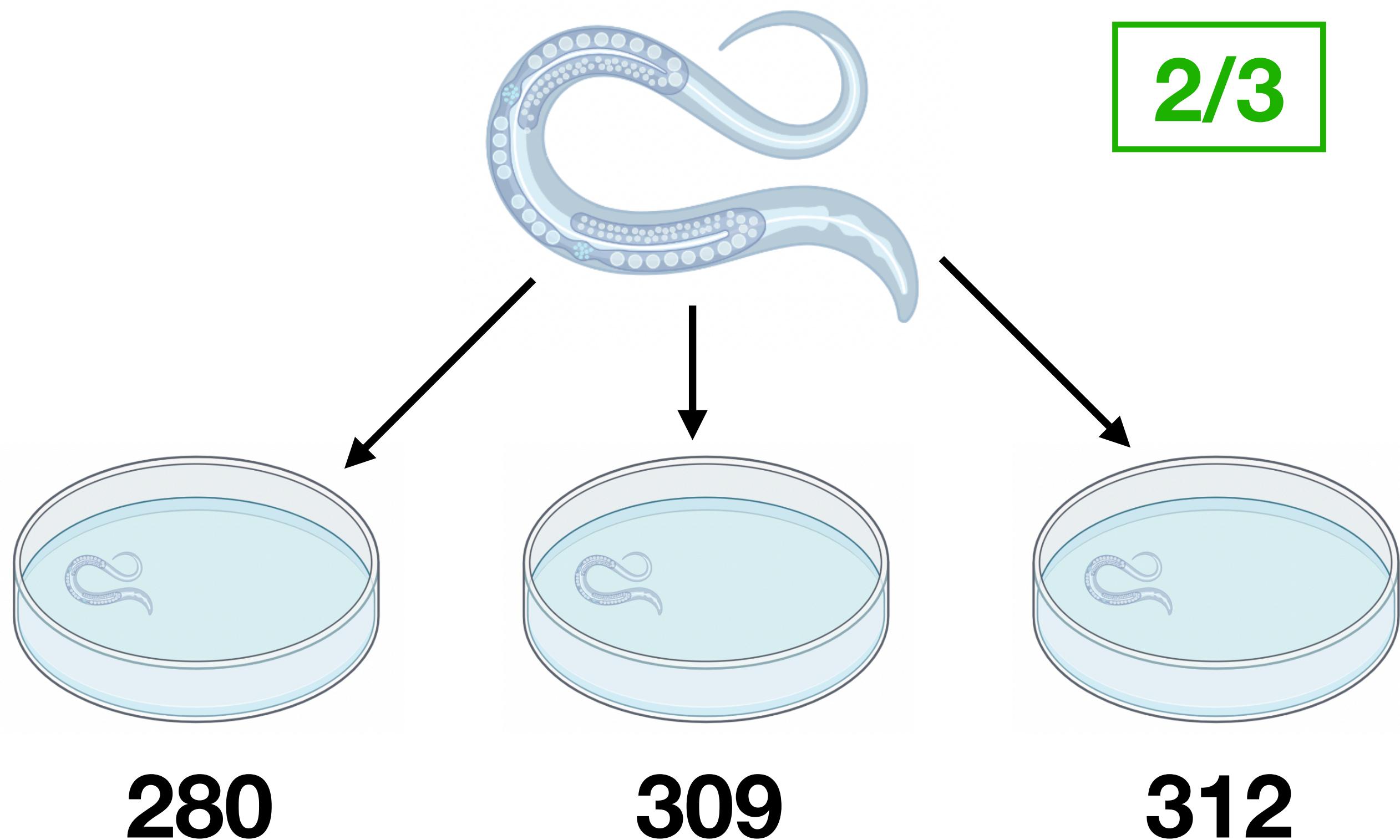
On average, the laboratory strain of *C. elegans* has 300 progeny. You pick three individuals each to their own plate and choose two plates to count broods. What is the sampling distribution?



<u>SAMPLE</u>	<u>MEAN</u>	<u>PROB</u>
280, 309	(294.5)	1/3
280, 312	(296)	1/3
309, 312	(310.5)	1/3

Population vs. Sample

On average, the laboratory strain of *C. elegans* has 300 progeny. You pick three individuals each to their own plate and choose two plates to count broods. What is the probability the average sample brood < 300?

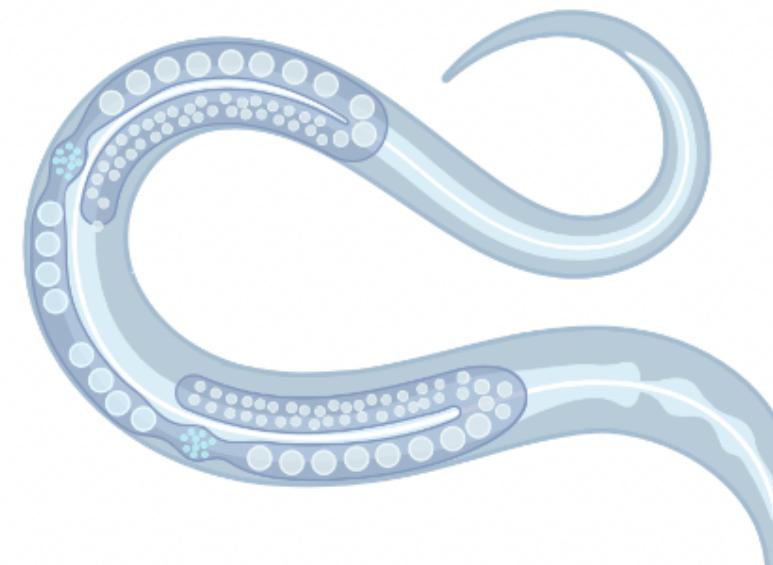


2/3	<u>SAMPLE</u>	<u>MEAN</u>	<u>PROB</u>
	280, 309	(294.5)	1/3
	280, 312	(296)	1/3
	280, 312	(310.5)	1/3

Sampling distribution of the sample mean

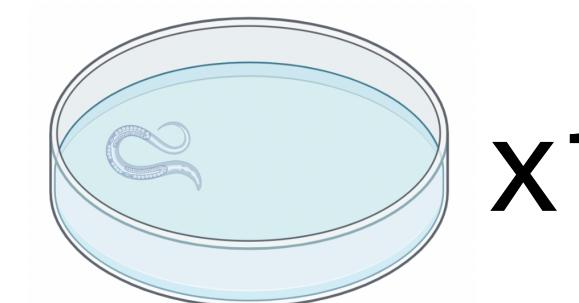
Let's take our *C. elegans* example and expand our sampling:

- Count broods for 10 individuals
- Repeat this experiment on four different days

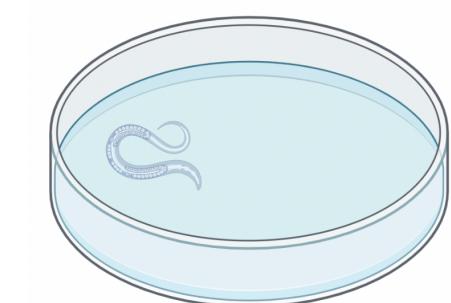


Mean: 300

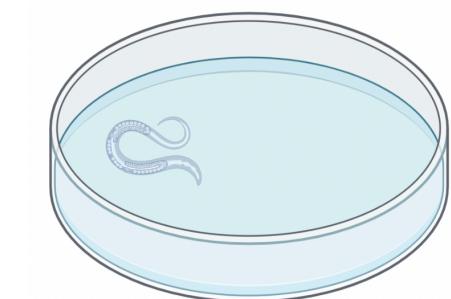
Sd: 20



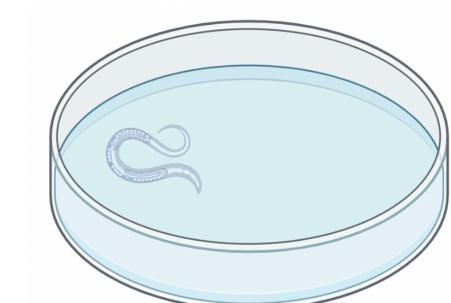
x10



x10



x10



x10



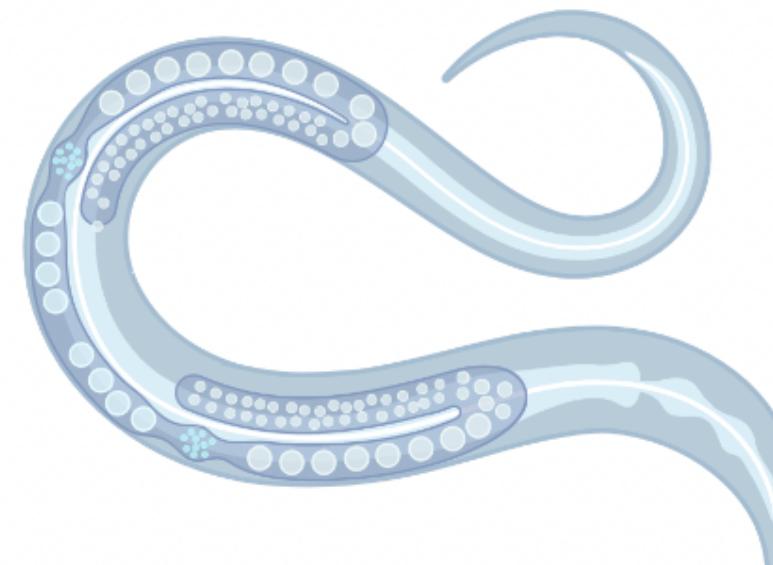
290, 304, 293, 315, 304, 298,
287, 302, 267, 291

mean: 295.4, sd: 12.7

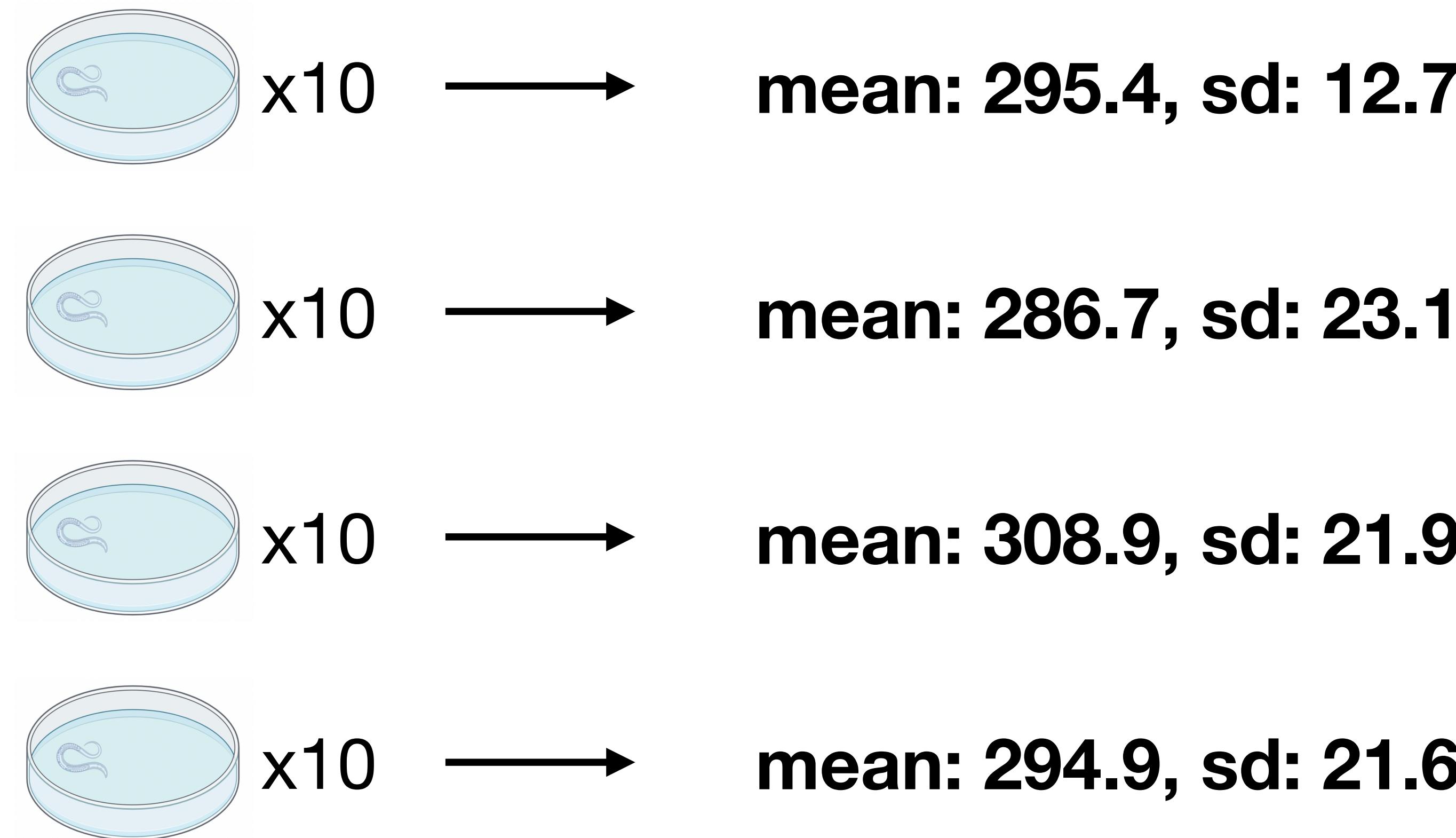
Sampling distribution of the sample mean

Let's take our *C. elegans* example and expand our sampling:

- Count broods for 10 individuals
- Repeat this experiment on four different days



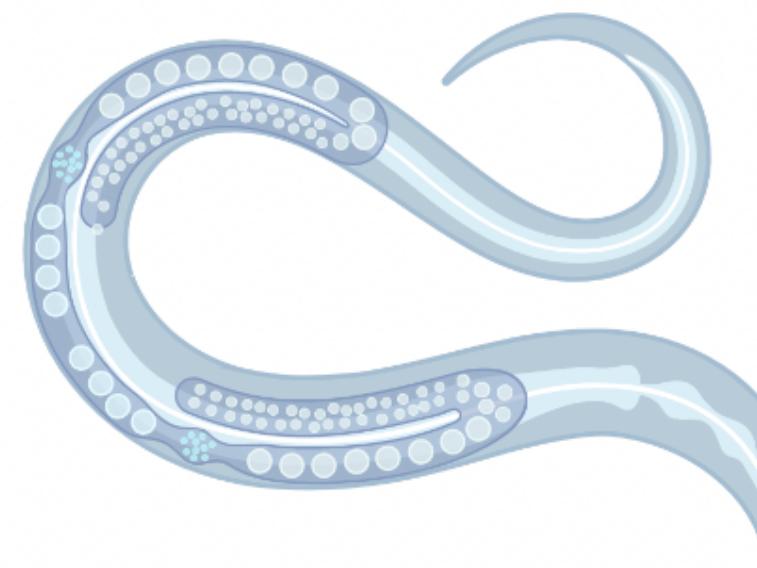
Mean: 300
Sd: 20



Sampling distribution of the sample mean

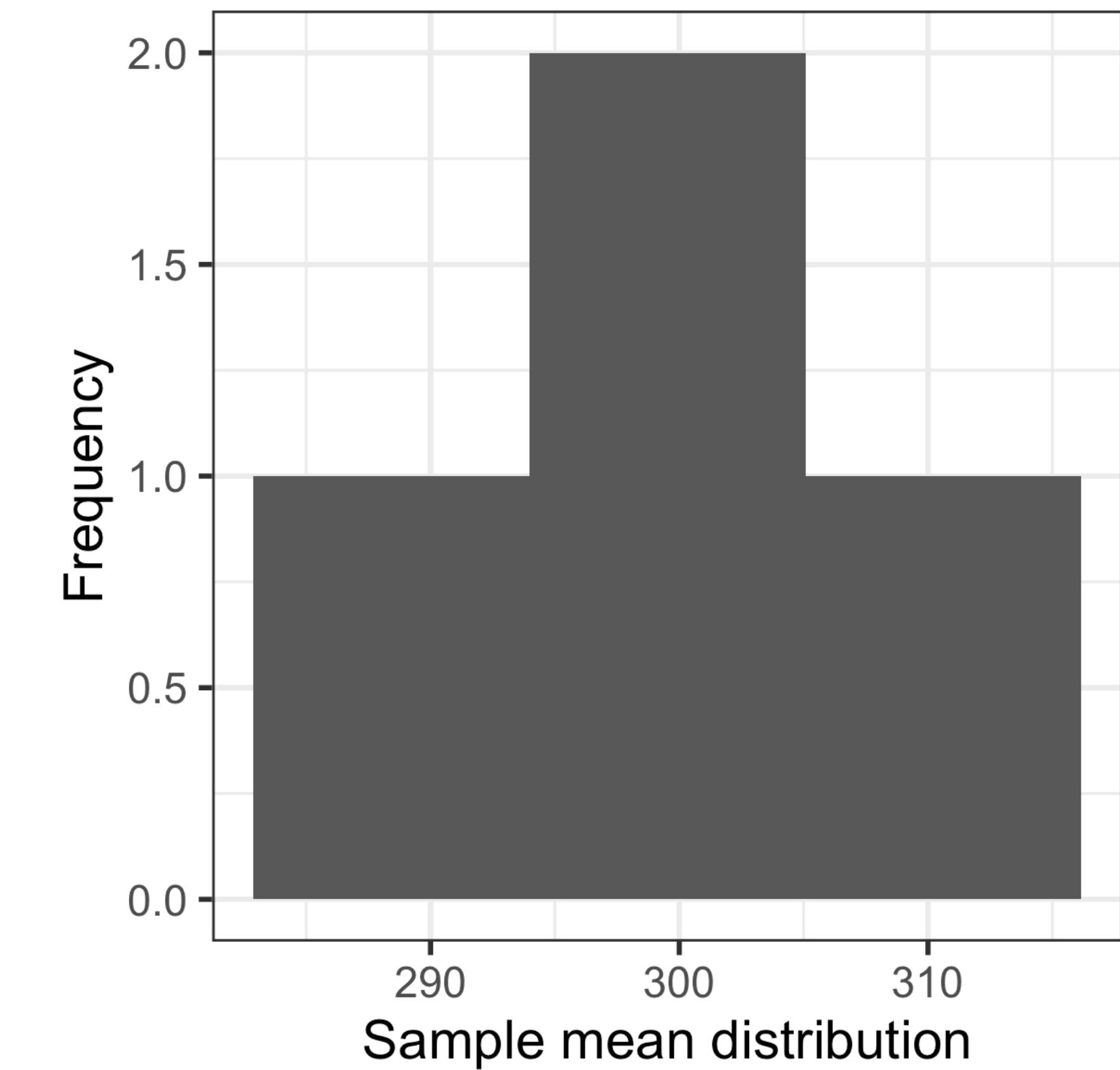
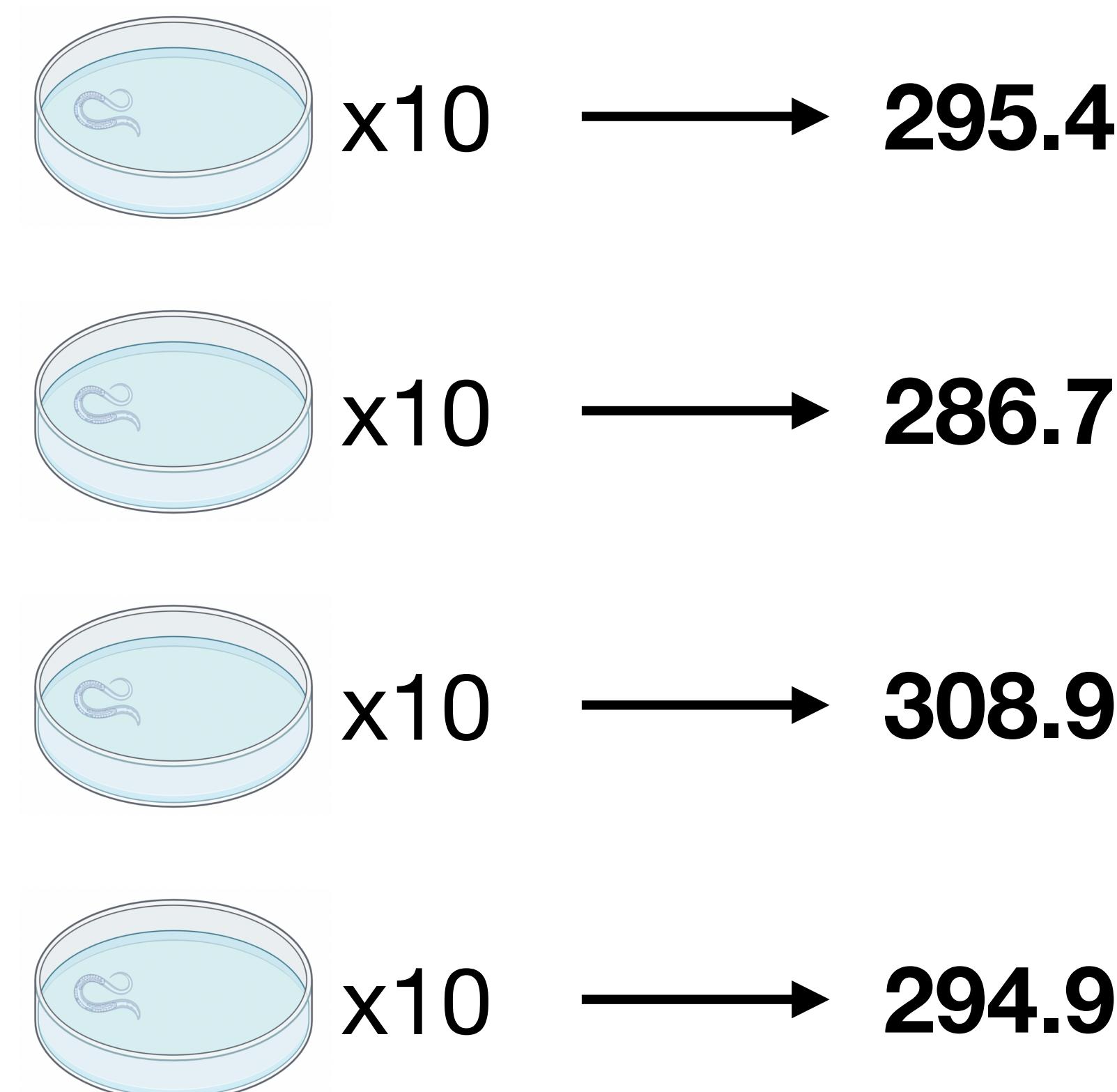
Let's take our *C. elegans* example and expand our sampling:

- Count broods for 10 individuals
- Repeat this experiment on four different days

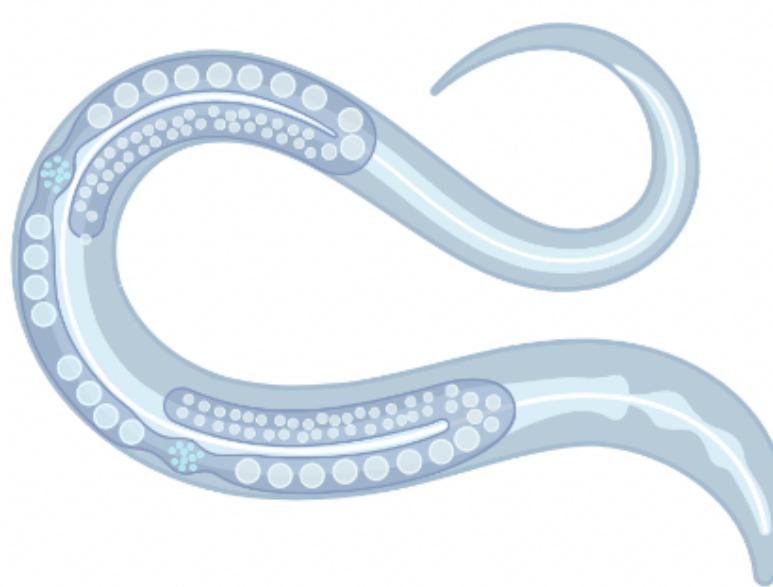


Mean: 300

Sd: 20

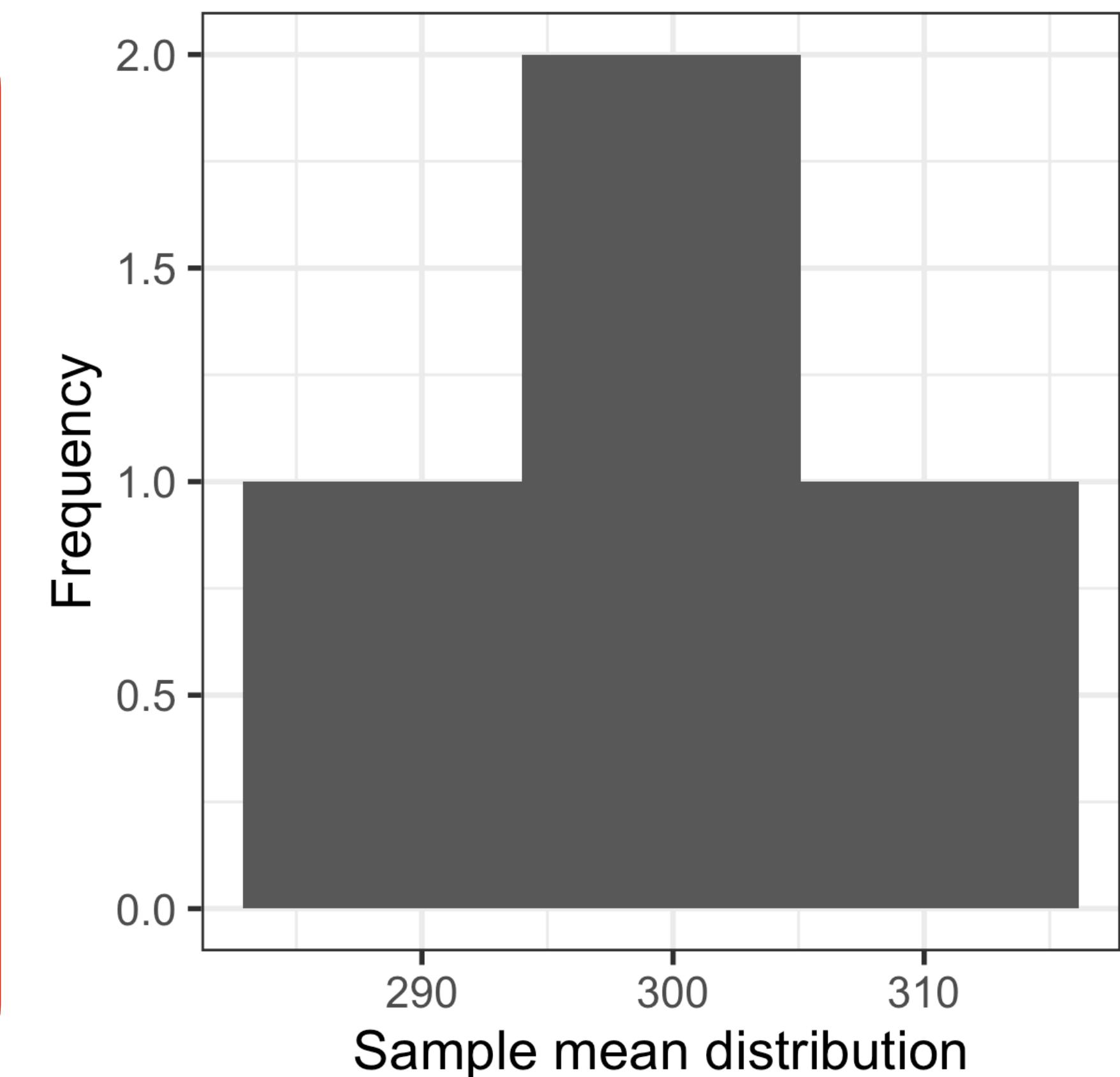
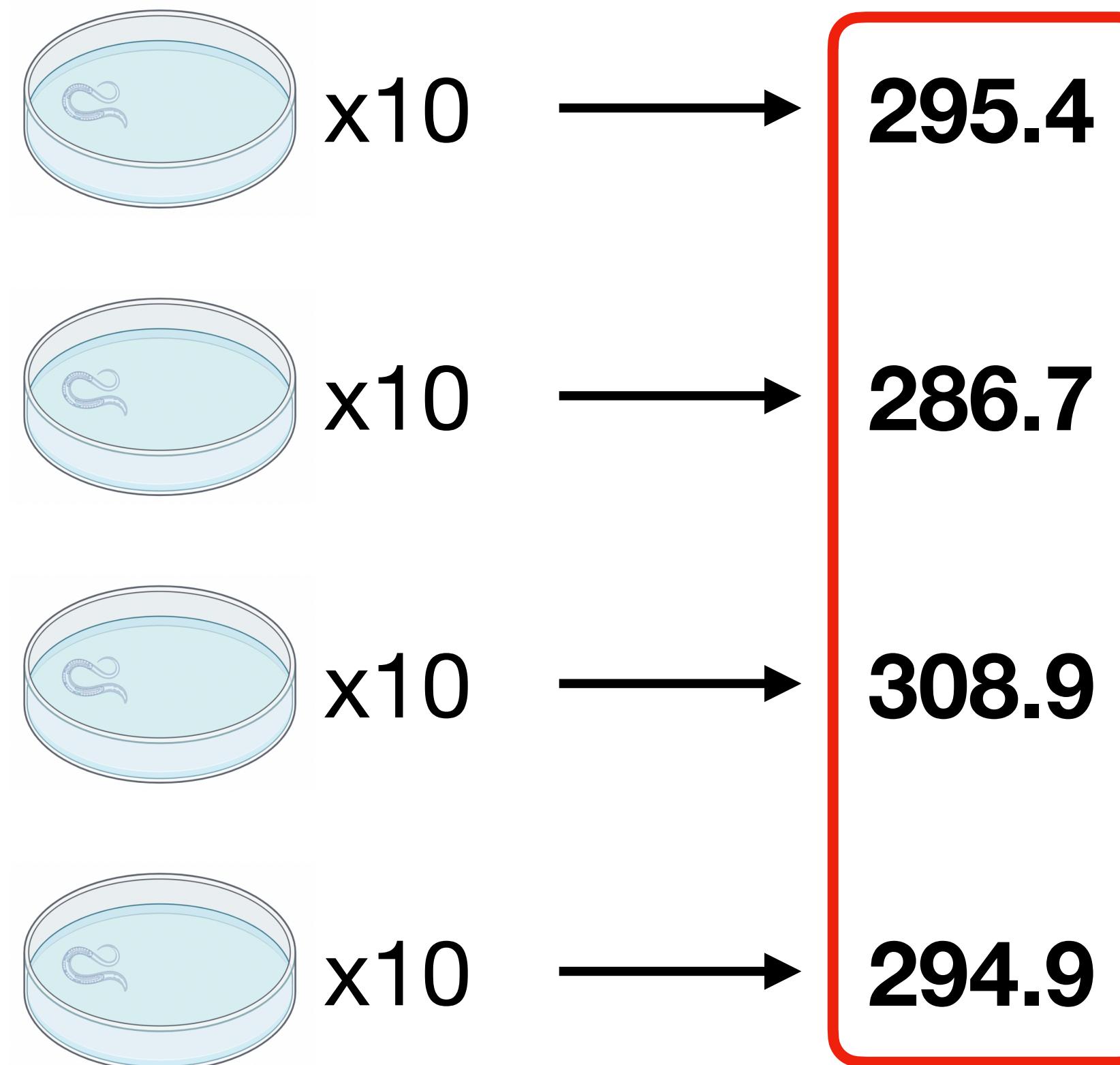


Sampling distribution of the sample mean



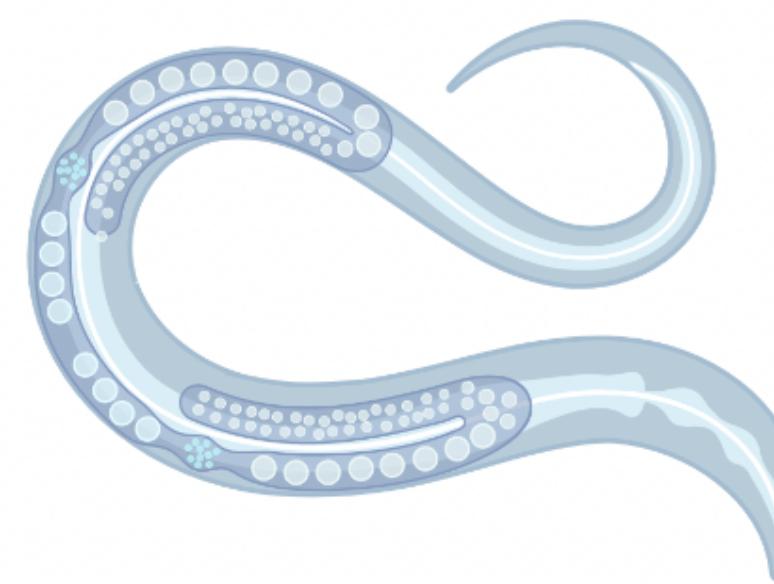
Mean: 300
Sd: 20

Sample mean: 296.5



1. The sample mean = the population mean

The sample standard variation

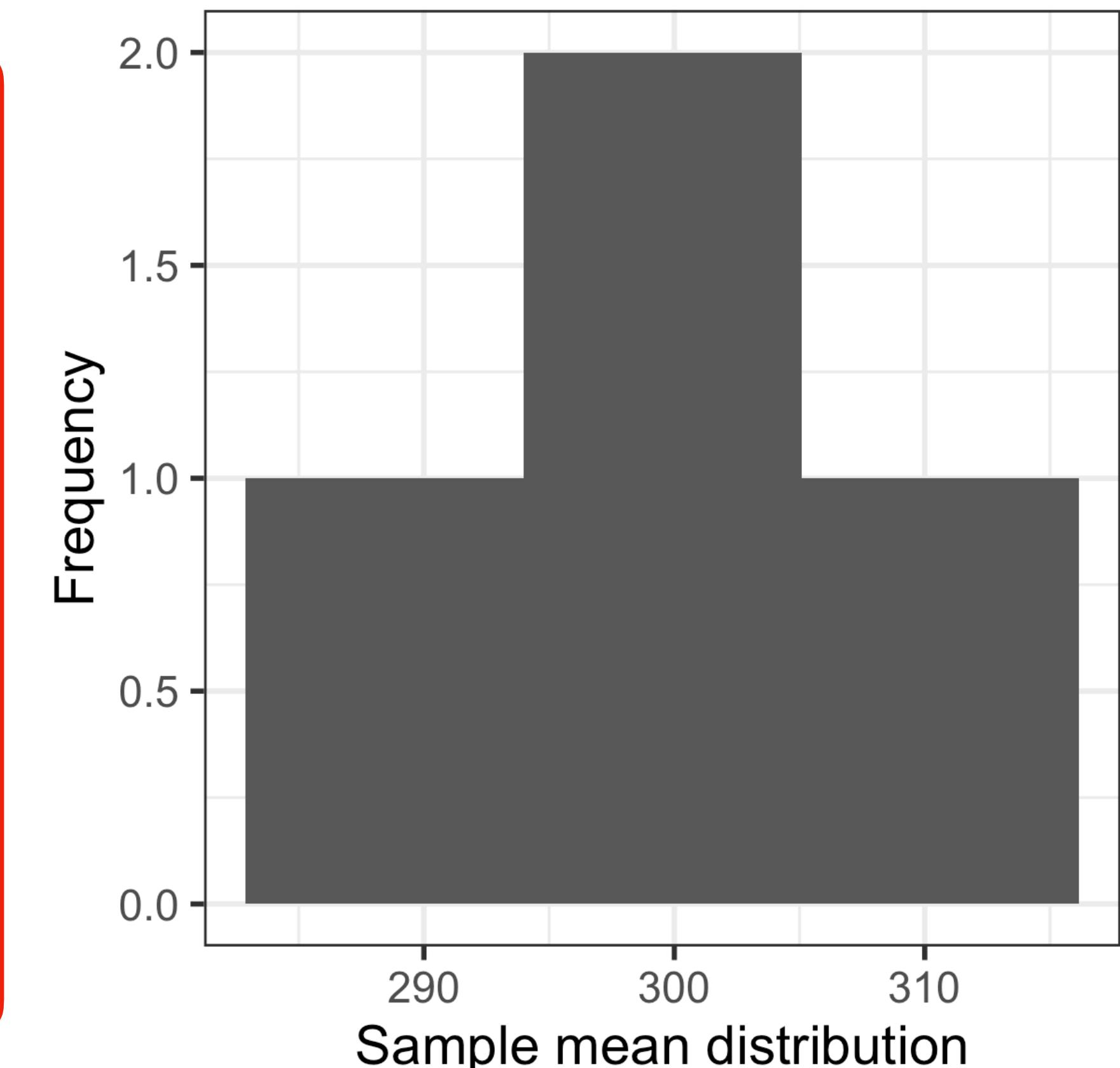
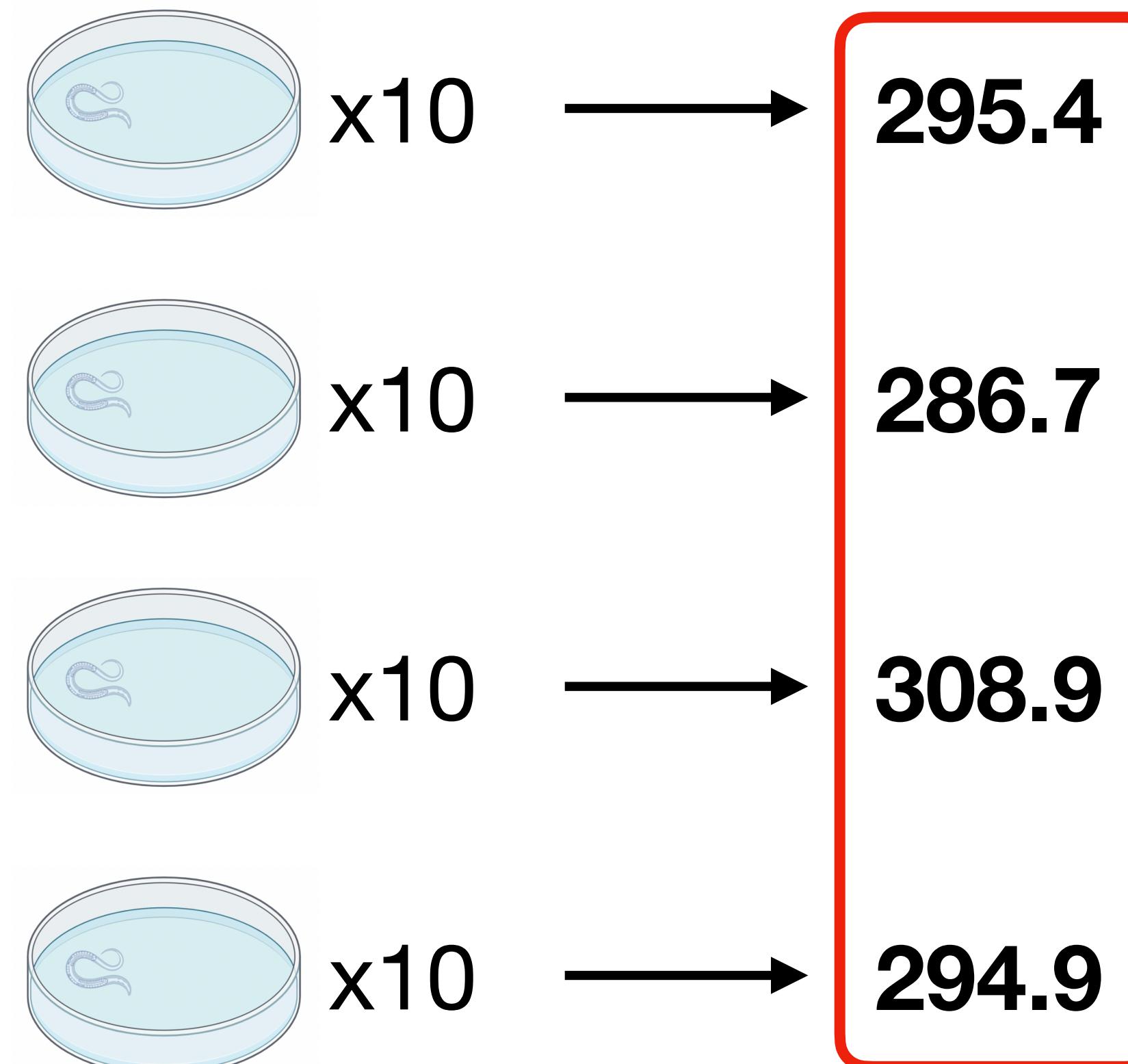


Mean: 300

Sd: 20

Sample mean: 296.5

Sample sd: 9.19



1. The sample mean = the population mean

2. The sample sd = population sd / sqrt(n)

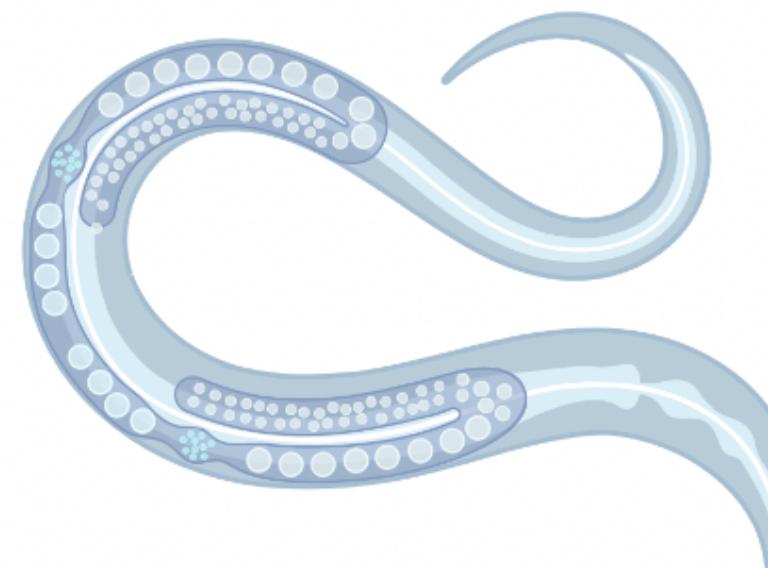
$$20/\sqrt{4} = 10$$



The sample standard variation

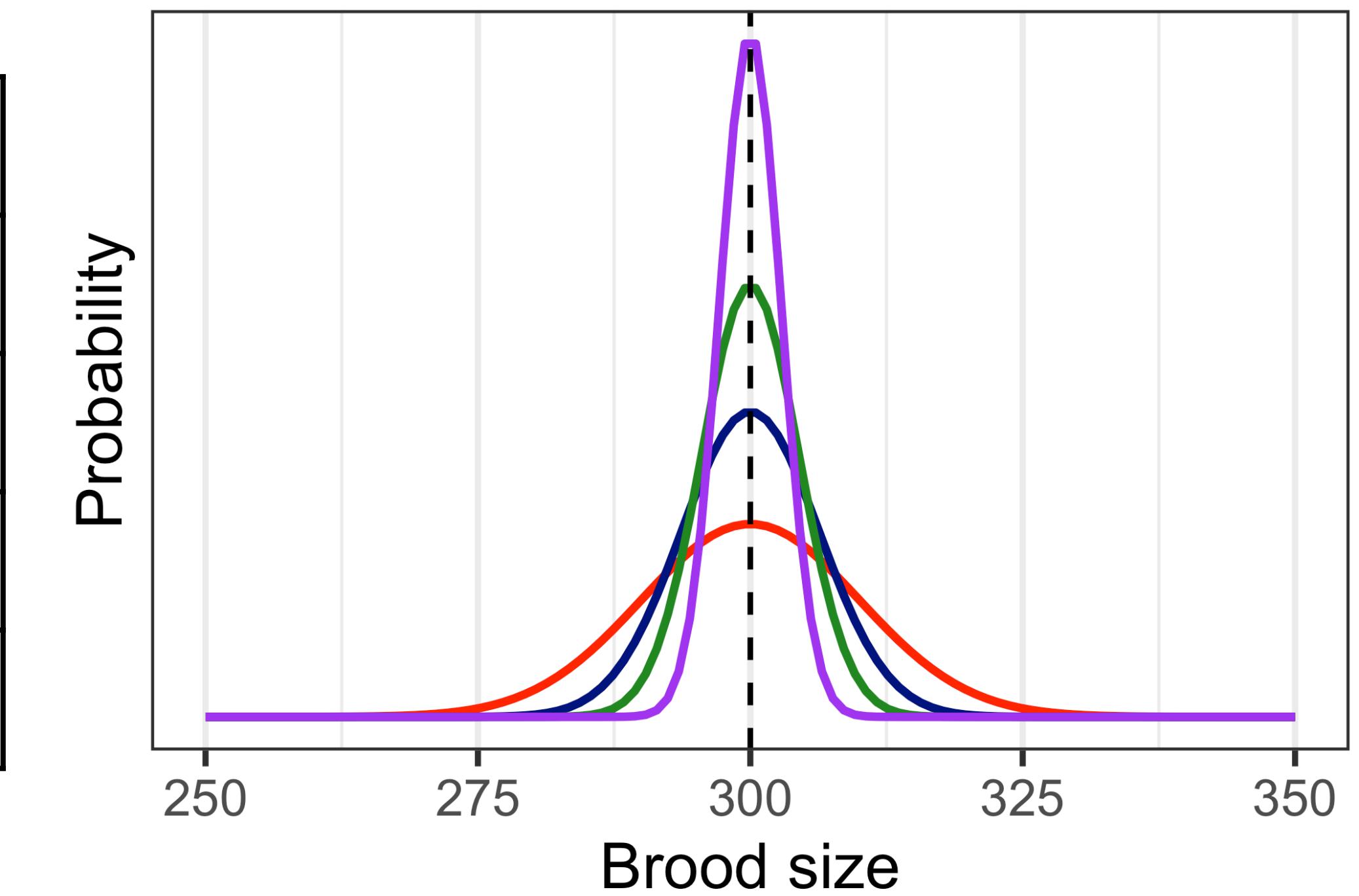
Let's take our *C. elegans* example and expand our sampling:

- Count broods for 10 individuals
- Repeat this experiment on 10 different days



Mean: 300
Sd: 20

n	mean	s
4	300	10
10	300	6.32
20	300	4.47
50	300	2.82

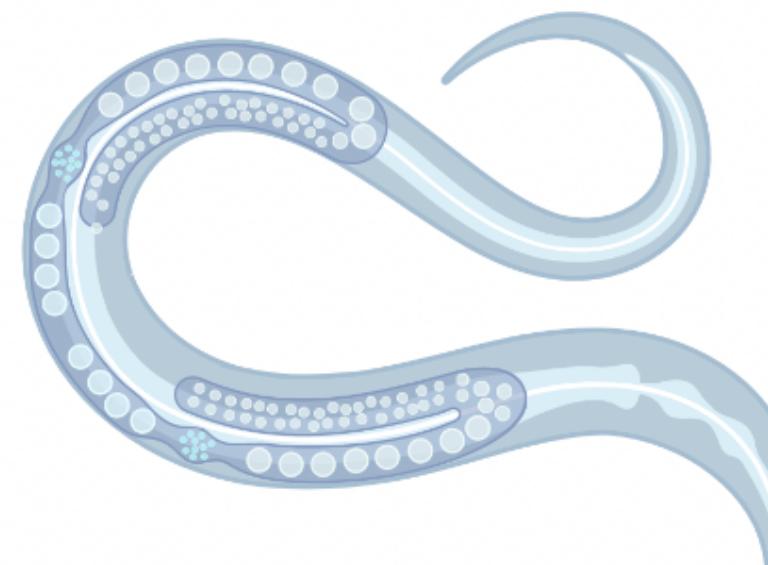


2. The sample sd = population sd / sqrt(n)

Shape of sampling distribution

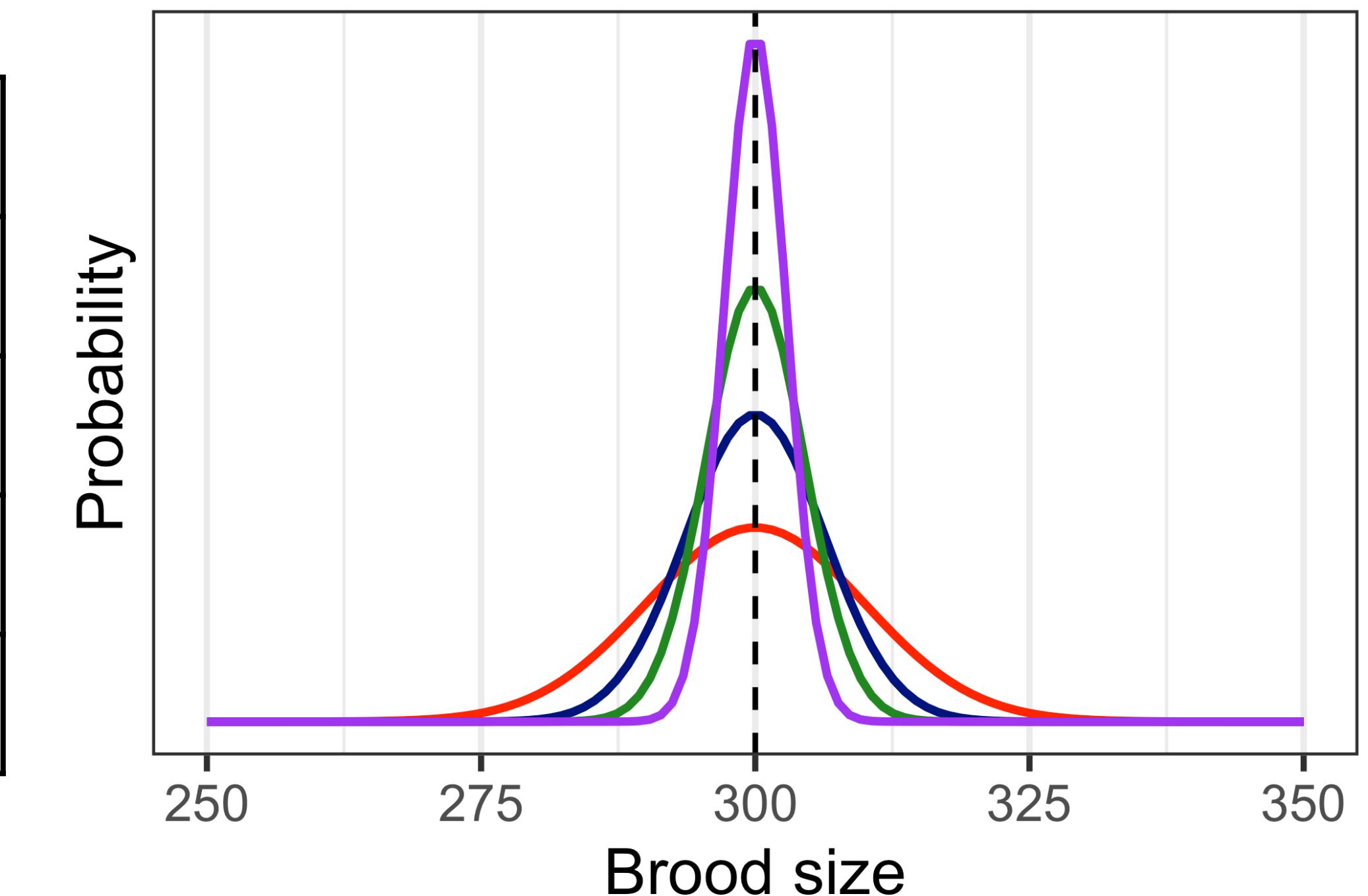
Let's take our *C. elegans* example and expand our sampling:

- Count broods for 10 individuals
- Repeat this experiment on **10** different days



Mean: 300
Sd: 20

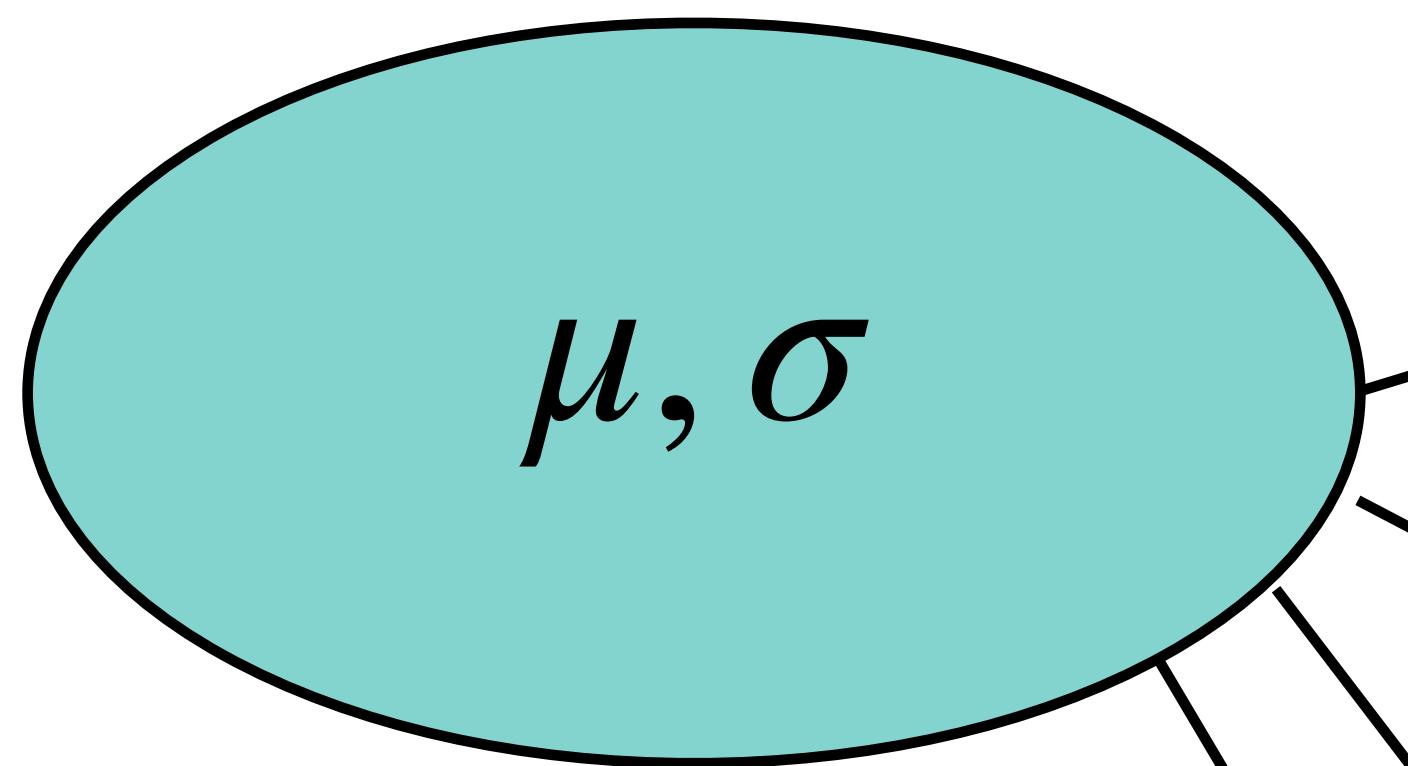
n	mean	s
4	300	10
10	300	6.32
20	300	4.47
50	300	2.82



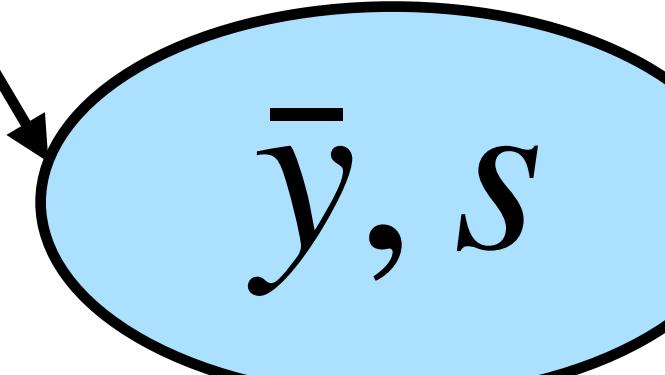
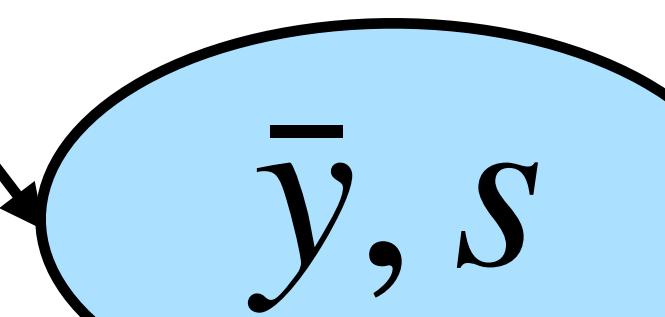
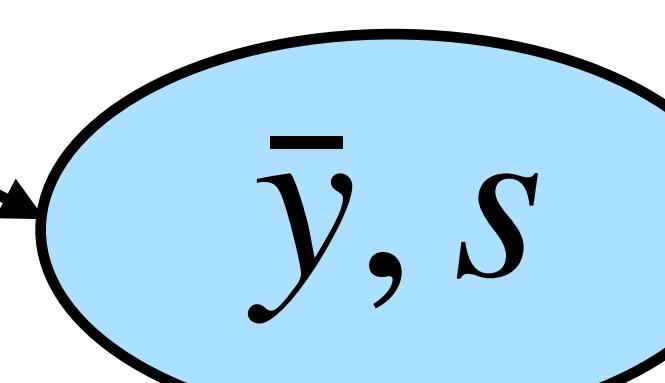
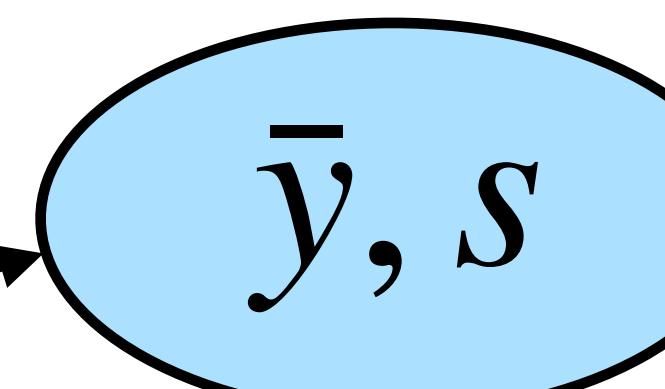
3. If the population distribution is normal, the sampling distribution will be normal

Summary: sampling distribution

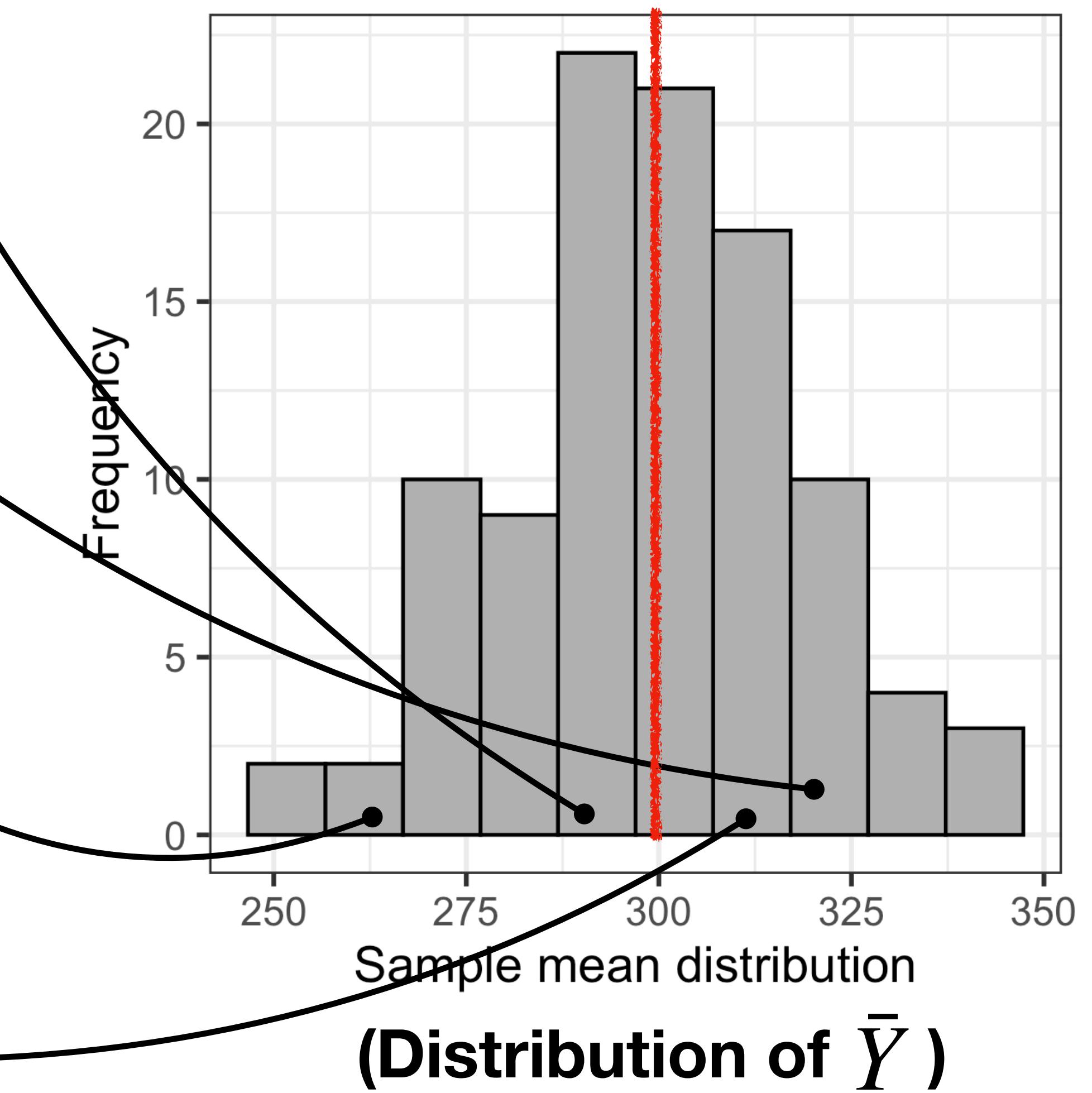
Population



Samples



$$\mu_{\bar{Y}} = \mu$$
$$\sigma_{\bar{Y}} = \sigma / \sqrt{n}$$



Example

A large population of seeds ($\mu = 500\text{mg}$; $\sigma = 120\text{mg}^2$) are weighed. If you take a random sample of four seeds, what is the probability that the mean weight of the four seeds will be greater than 550 mg?

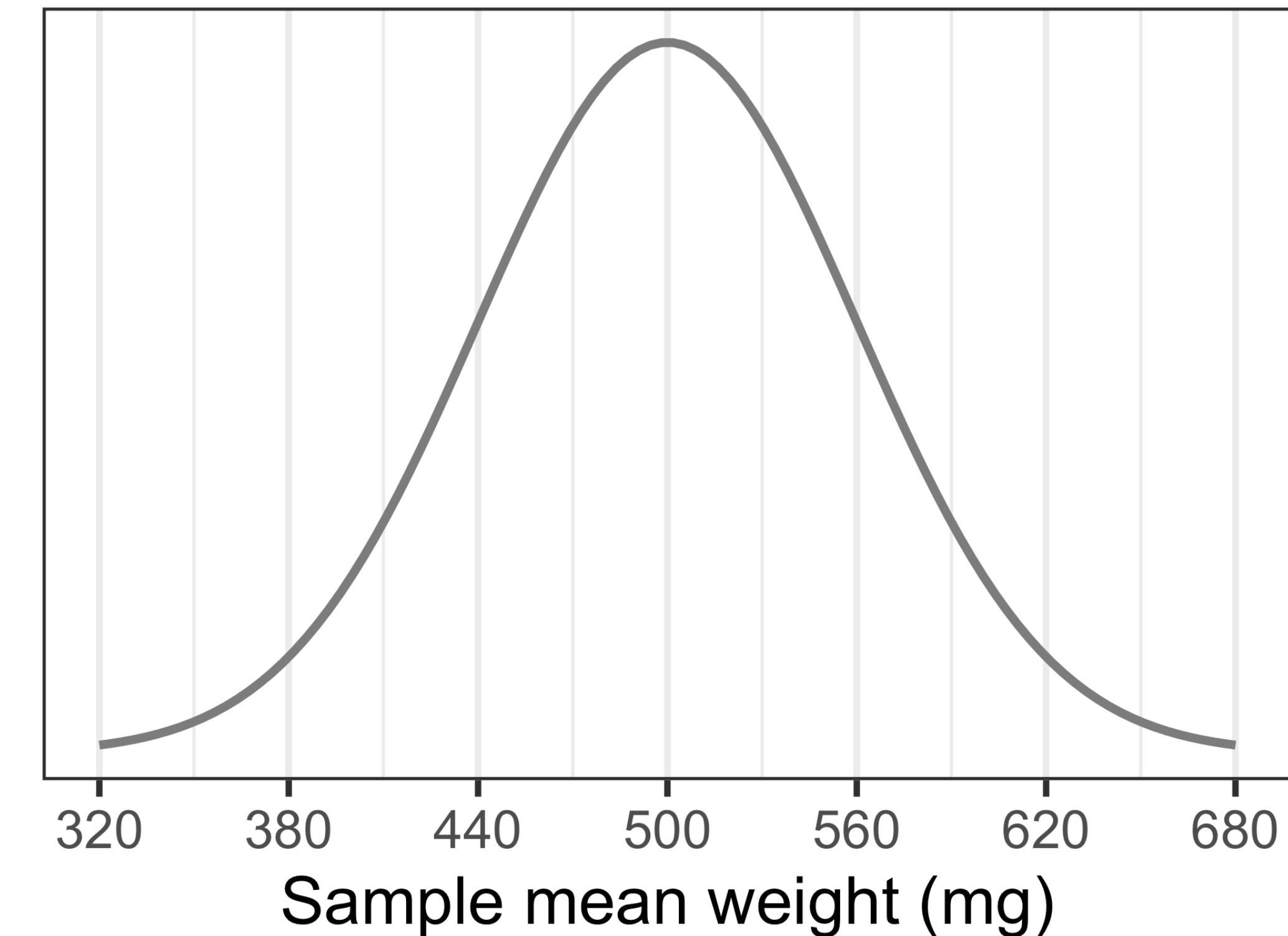
$$\mu_{\bar{Y}} = \mu \quad 500 \text{ mg}$$

$$\sigma_{\bar{Y}} = \sigma/\sqrt{n} \quad 120 / \sqrt{4} = 60 \text{ mg}$$

Example

A large population of seeds ($\mu = 500\text{mg}$; $\sigma = 120\text{mg}^2$) are weighed. If you take a random sample of four seeds, what is the probability that the mean weight of the four seeds will be greater than 550 mg?

$$\sigma_{\bar{Y}} = 60\text{mg}$$



Example

A large population of seeds ($\mu = 500\text{mg}$; $\sigma = 120\text{mg}^2$) are weighed. If you take a random sample of four seeds, what is the probability that the mean weight of the four seeds will be greater than 550 mg?

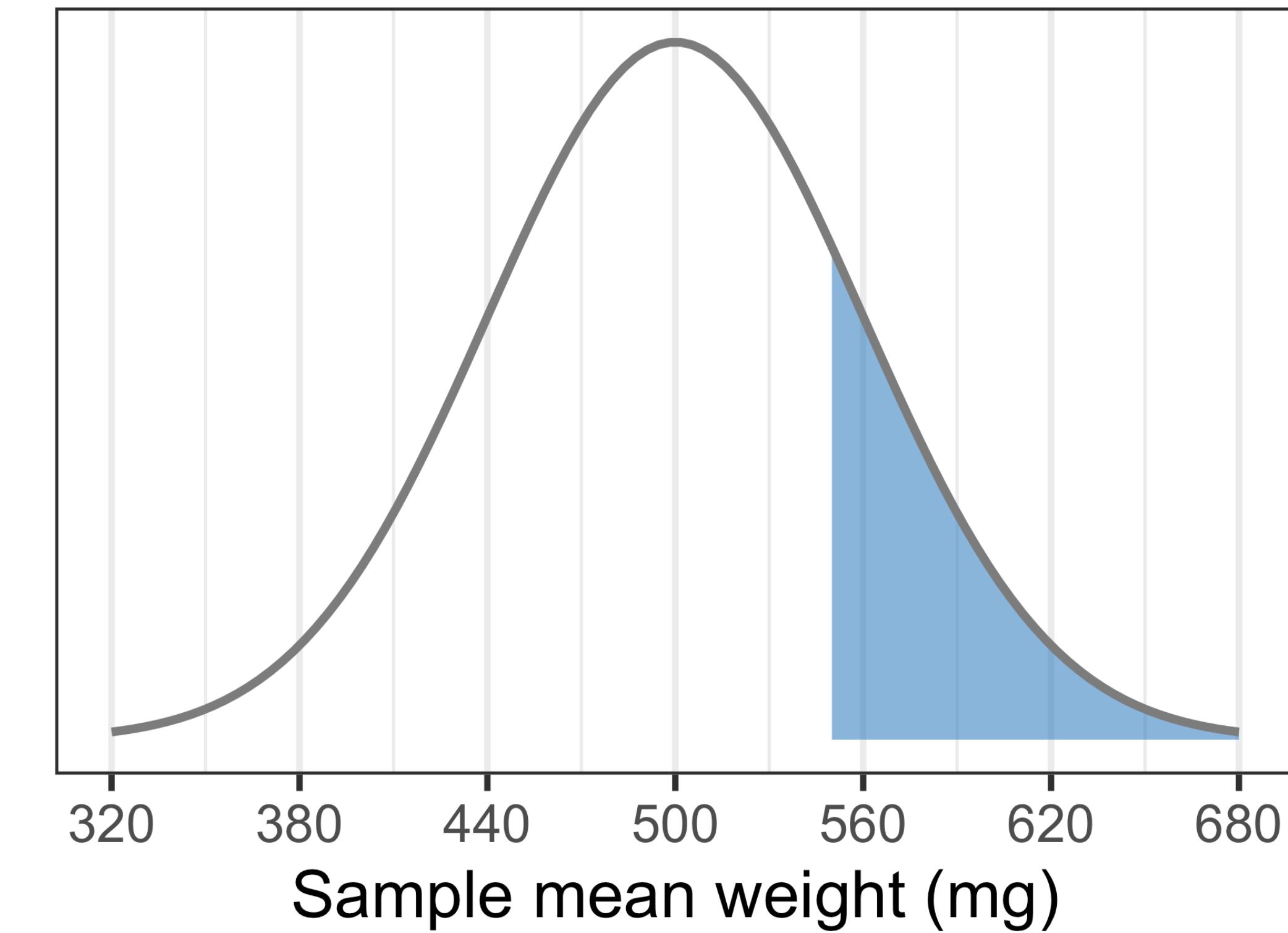
$$\mu = 500\text{mg}$$

$$\sigma_{\bar{Y}} = 60\text{mg}$$

$$z = \frac{\bar{y} - \mu_{\bar{Y}}}{\sigma_{\bar{Y}}}$$

$$z = \frac{550 - 500}{60}$$

$$z = 0.83$$



Example

A large population of seeds ($\mu = 500\text{mg}$; $\sigma = 120\text{mg}^2$) are weighed. If you take a random sample of four seeds, what is the probability that the mean weight of the four seeds will be greater than 550 mg?

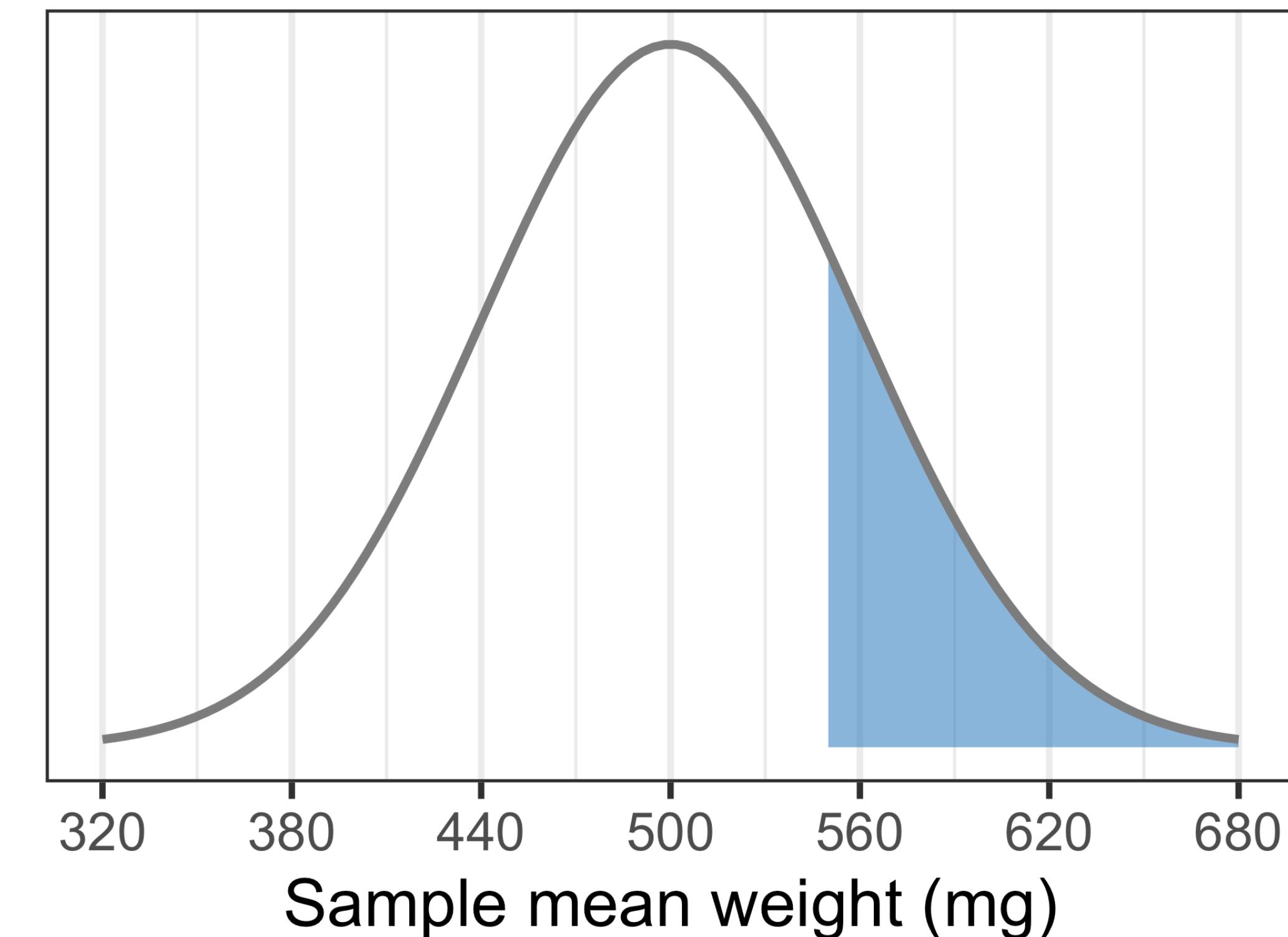
$$\mu = 500\text{mg}$$

$$\sigma_{\bar{Y}} = 60\text{mg}$$

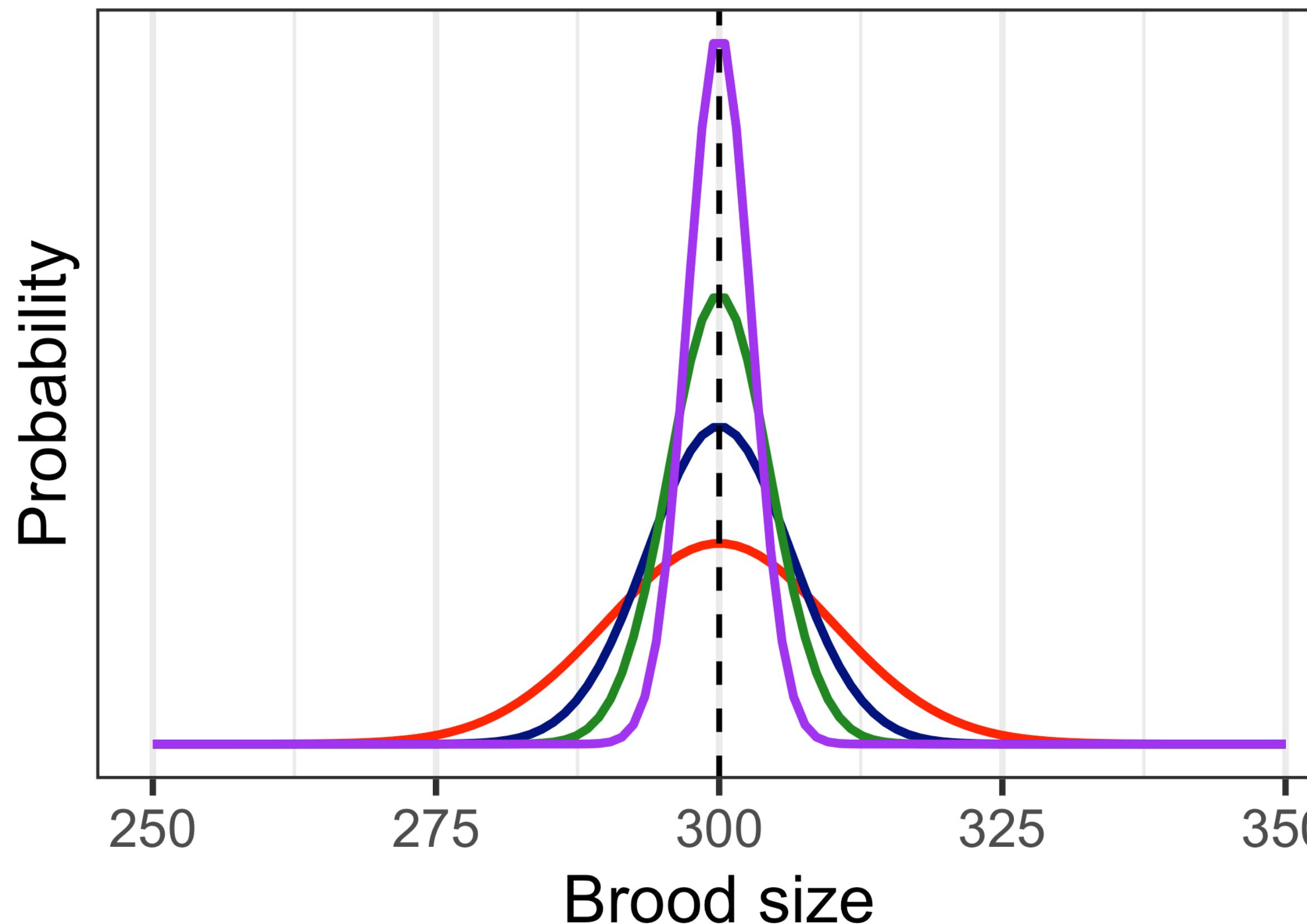
$$z = 0.83$$

$$1 - \text{pnorm}(0.83) = 0.20$$

$$1 - \text{pnorm}(550, 500, 60)$$



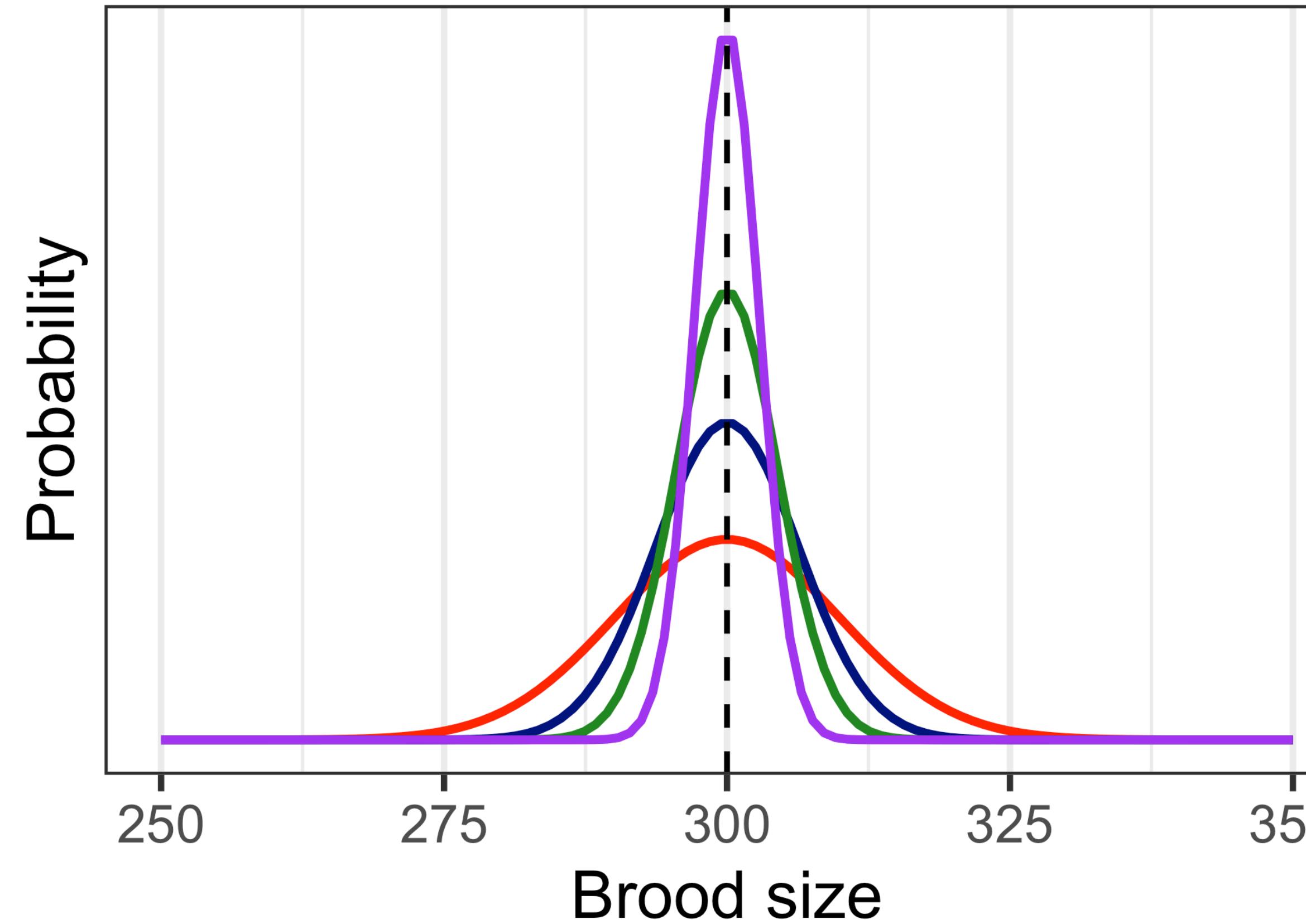
Central limit theorem



3. If the population distribution is normal, the sampling distribution will be normal

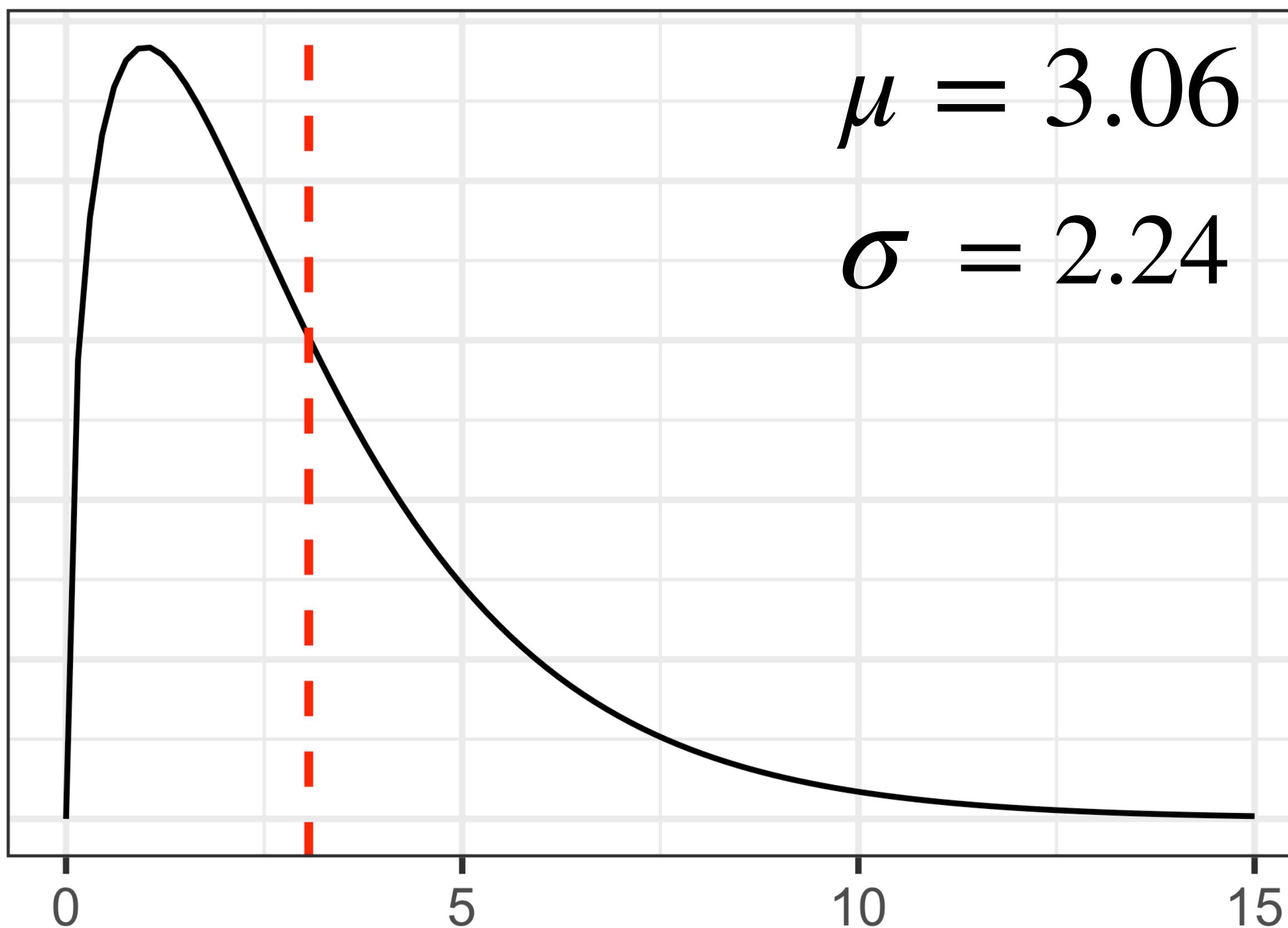
Central limit theorem

How large?
Large enough...



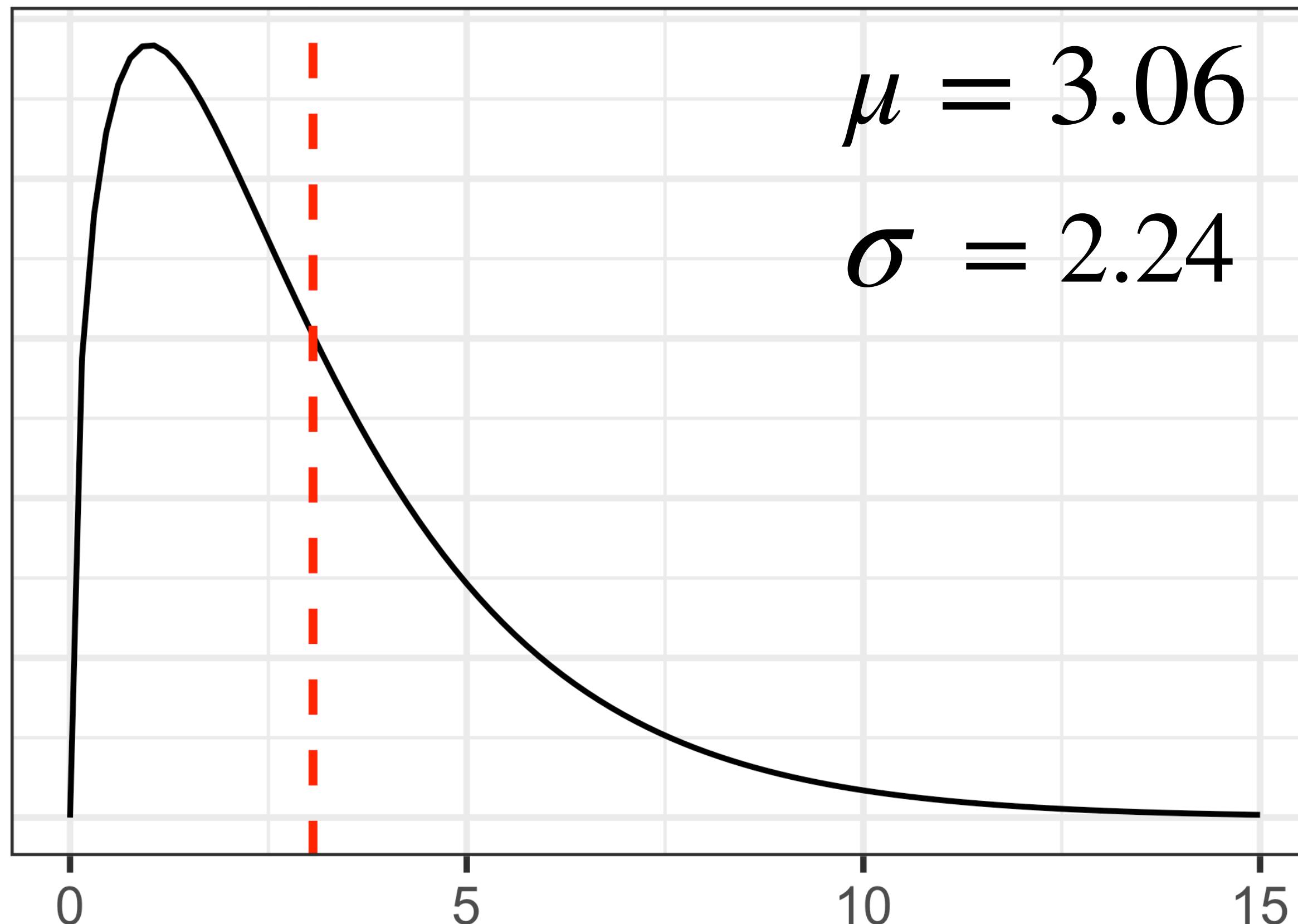
4. If the population distribution is NOT normal, but n is LARGE, the sampling distribution will be normal

Central limit theorem



4. If the population distribution is NOT normal, but n is LARGE, the sampling distribution will be normal

Central limit theorem



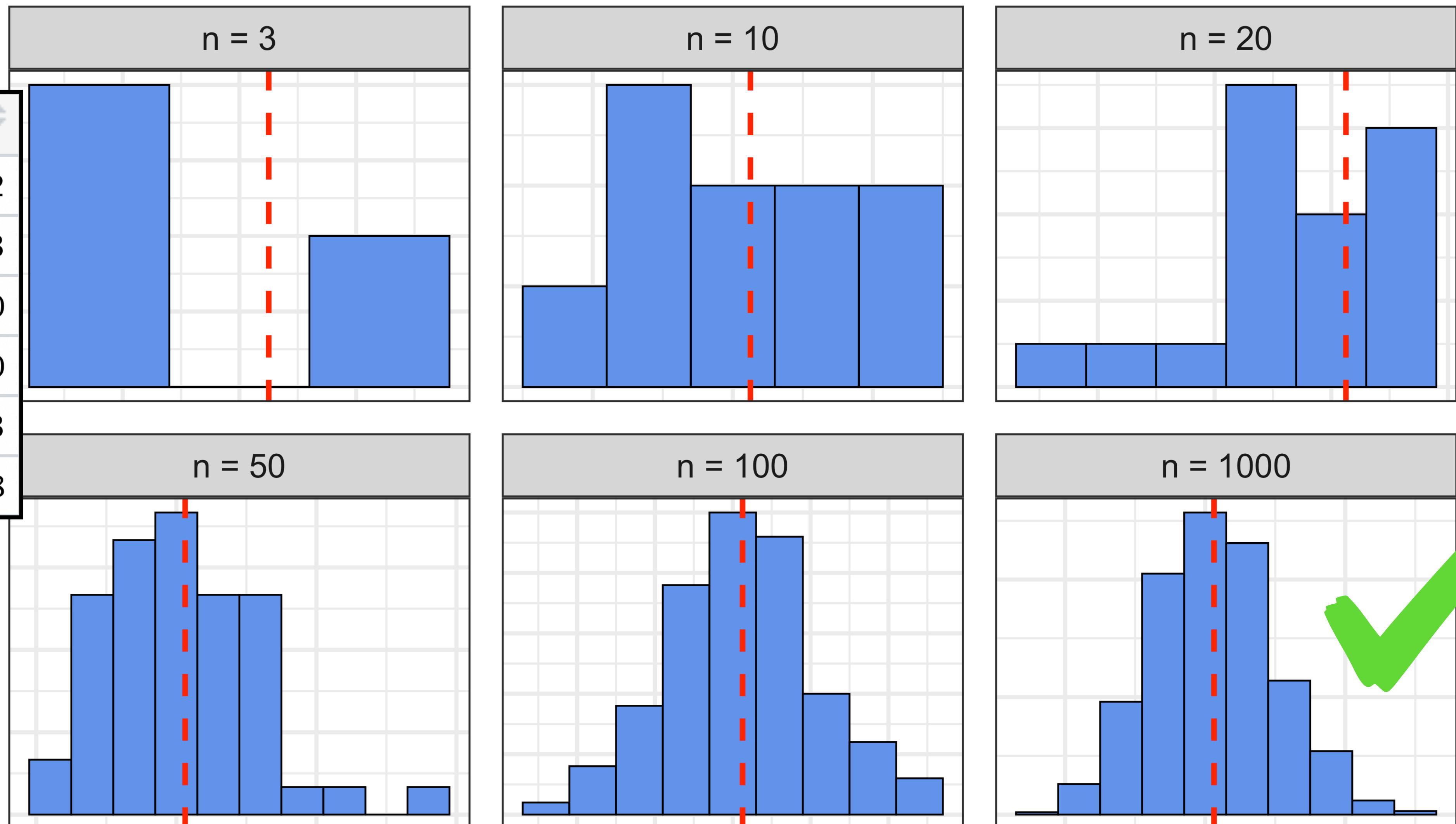
- Take a sample of 20 individuals
- Calculate the mean of the sample
- Repeat this sampling 3, 10, 20, 50, 100, or 1000 times
- Plot the distribution of the means across samples

4. If the population distribution is NOT normal, but n is LARGE, the sampling distribution will be normal

Central limit theorem

$$\mu = 3.06$$

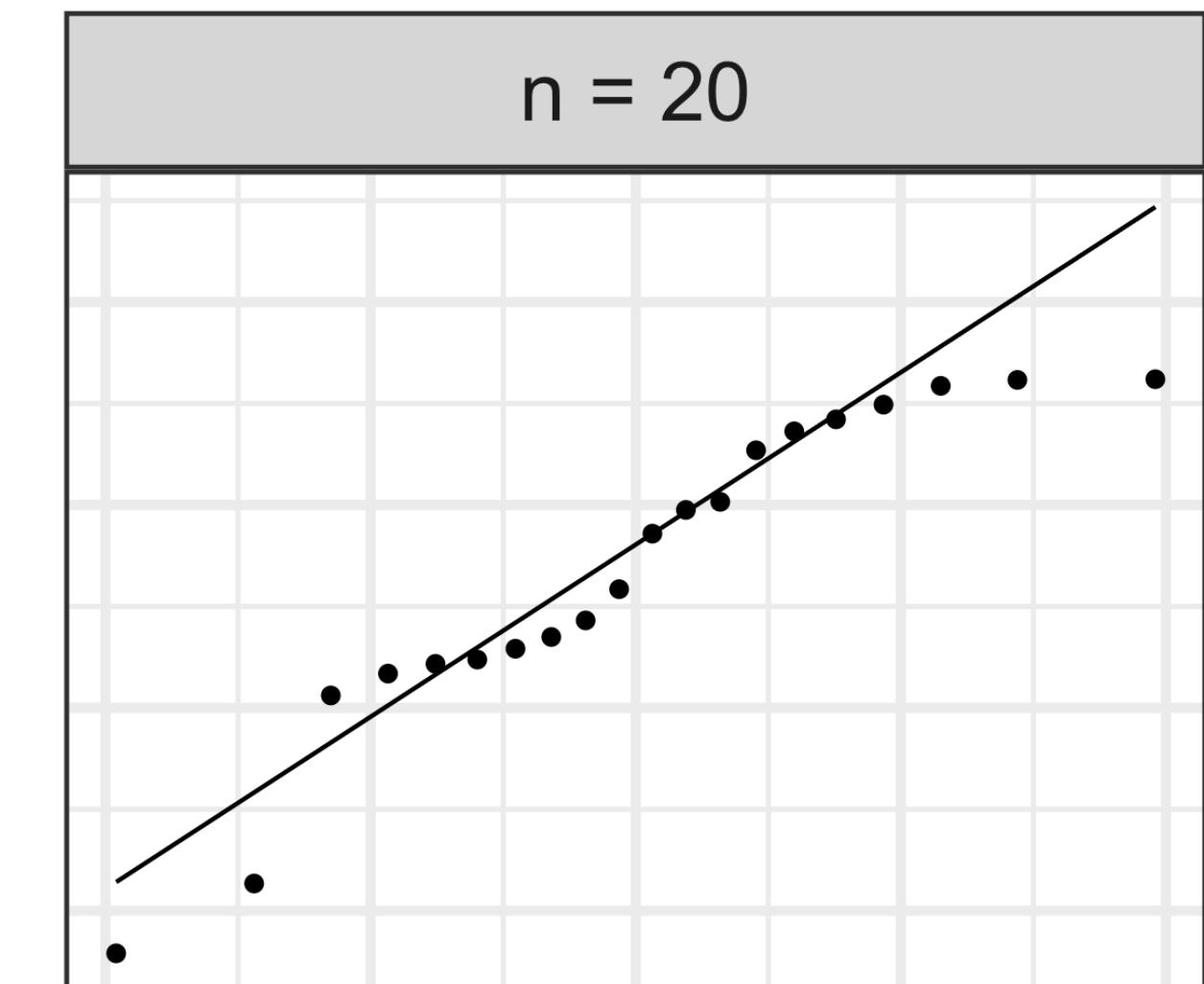
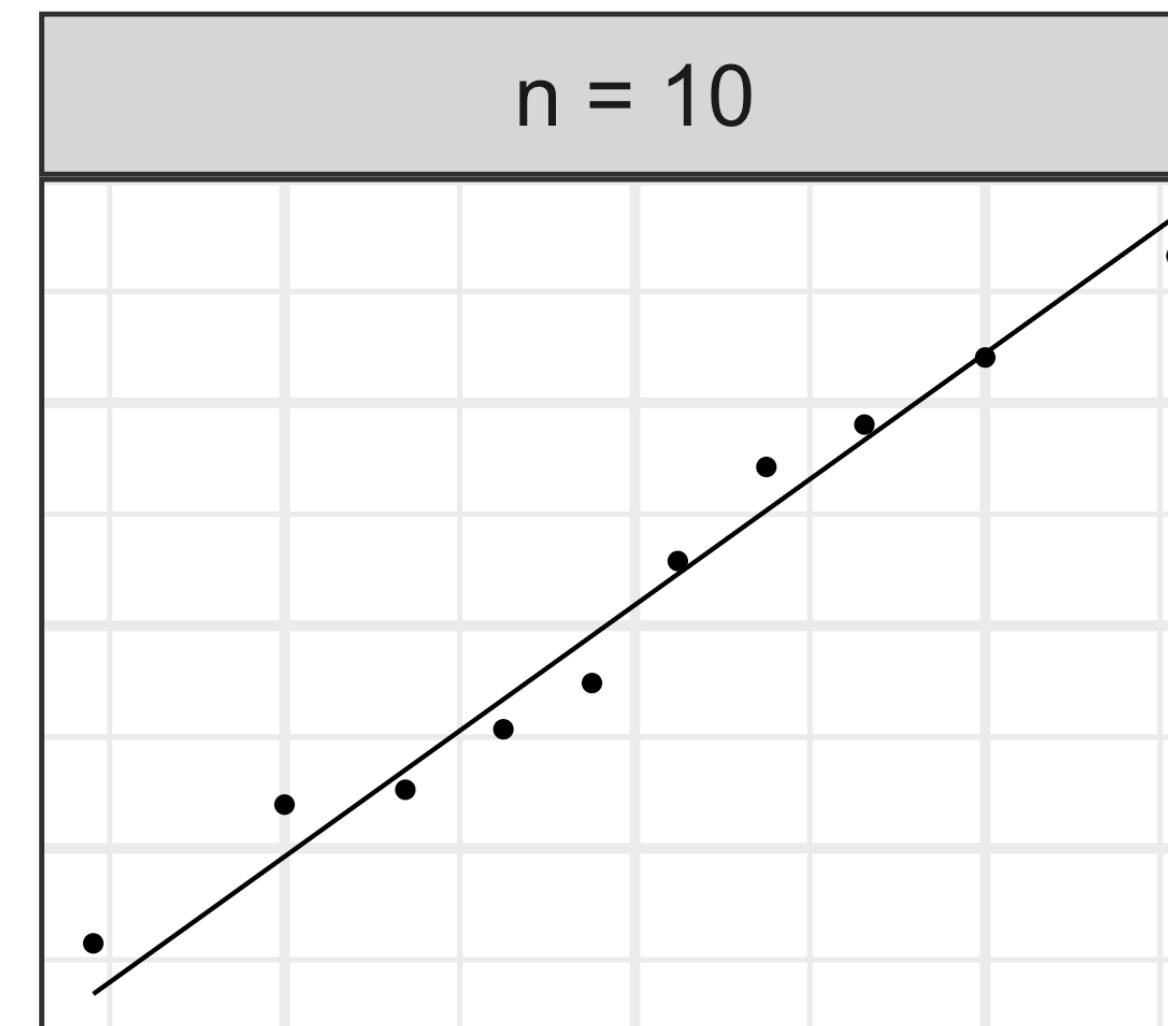
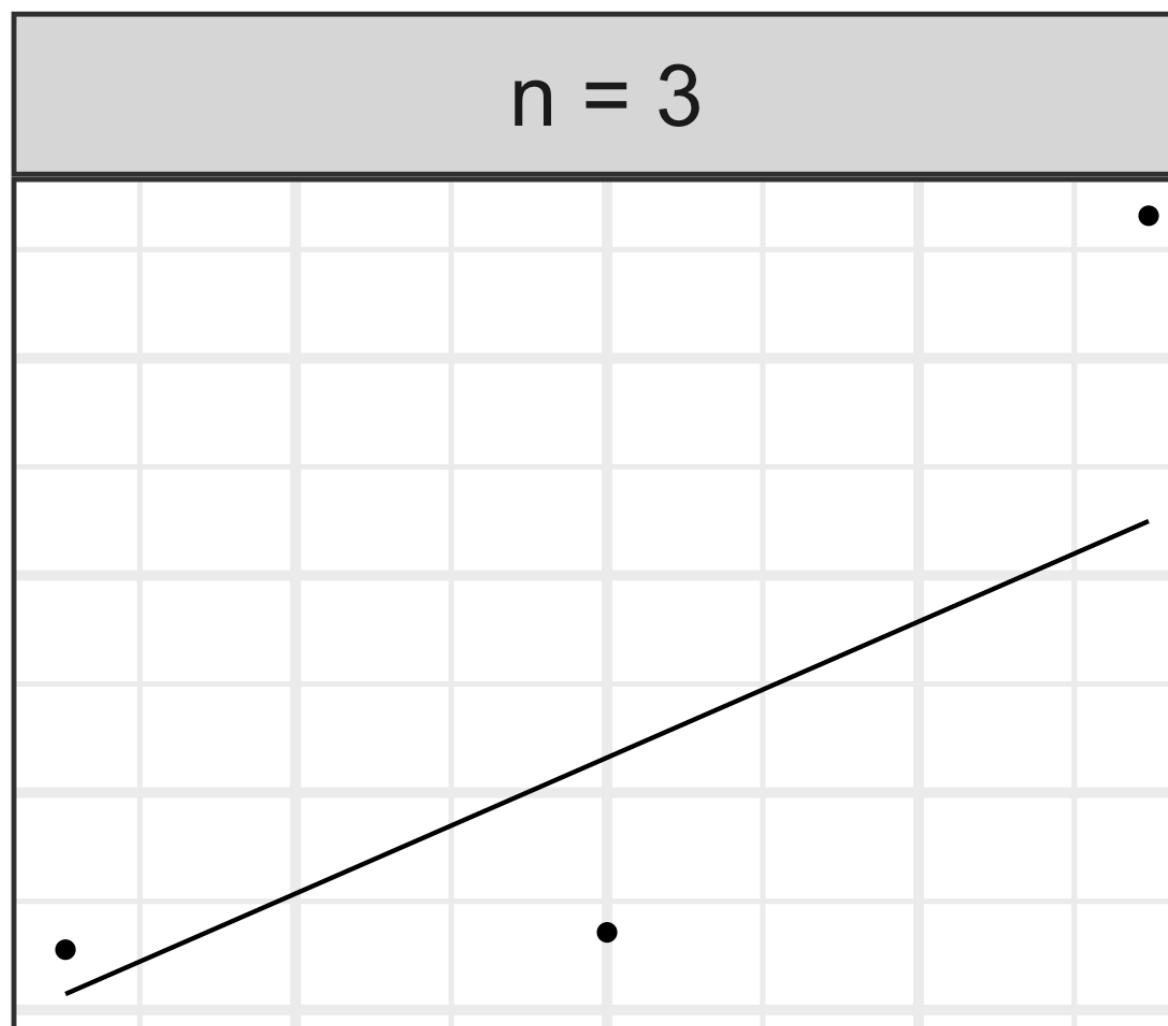
samples	sample_mean
n = 3	2.857522
n = 10	3.043083
n = 20	2.837200
n = 50	3.011640
n = 100	3.053503
n = 1000	3.061708



Central limit theorem

```
qqnorm(data)  
qqline(data)
```

data = means of samples



```
ggplot2::ggplot(data) +  
  ggplot2::aes(sample = mean) +  
  ggplot2::stat_qq() +  
  ggplot2::stat_qline()
```

R does the calc + plot at same time!

