Class 13

DATA1220-55, Fall 2024

Sarah E. Grabinski

2024-09-27

Chapter 4 - Distributions

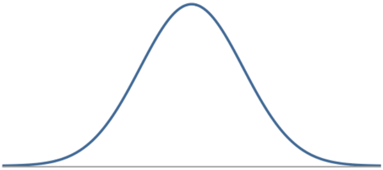
- We will *only* be covering Chapter 4.1 on the normal distribution in your textbook
- If you have an interest in math or statistics, you may want to read the rest of Chapter 4
  - 4.2 - Geometric distribution
  - 4.3 - Binomial distribution
  - 4.4 - Negative binomial distribution
  - 4.5 - Poisson distribution

Chapter 4 Objectives

- Identify and describe the standard normal and normal distributions
- Standardize normal distributions and calculate Z-scores
- Calculate percentiles and exact probabilities
- Apply the 68-95-99.7 Rule
- Read a QQ-Plot (not in book)

The Normal Distribution

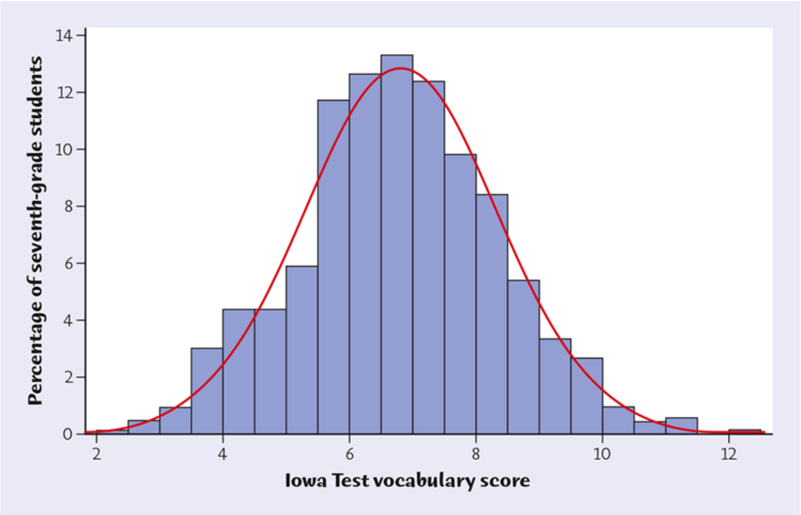
- Symmetric, unimodal, “bell-shaped”
- Not as common as people think in real data
- Strong assumption in small sample sizes ($ 20)
- Powerful statistical tests available when outcome approximates normal distribution



Notation

- μ (Greek letter mu) represents the mean
- σ (Greek letter sigma) represents the standard deviation of the mean
- $N(\mu, \sigma)$ stands for a normal distribution with mean μ and standard deviation σ

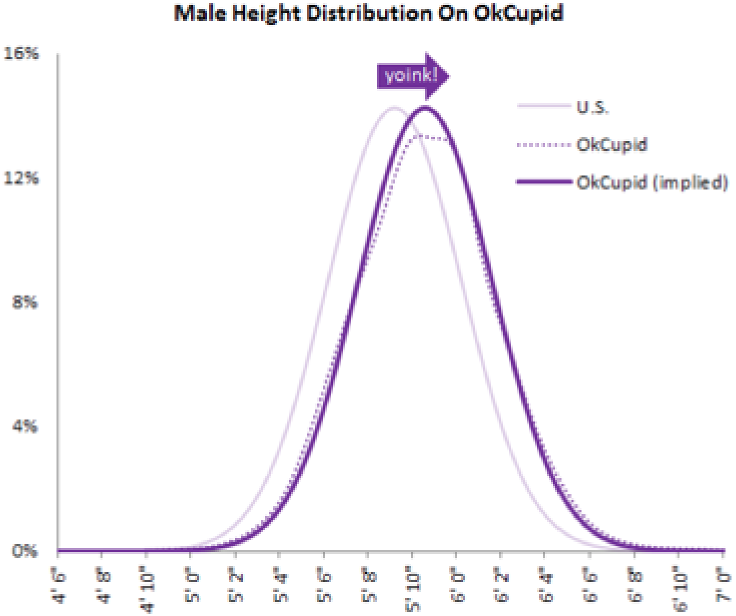
Histograms and Density Curves



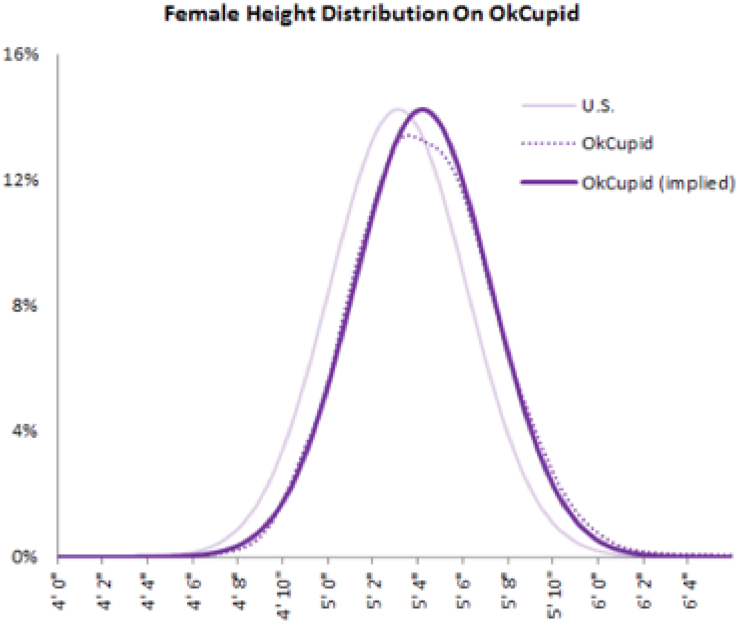
Moore/Notz, *The Basic Practice of Statistics*, 9e, © 2021 W. H. Freeman and Company

Figure 1: Vocabulary scores for 947 seventh-graders. Both histograms and density curves can be helpful in identifying normal distributions.

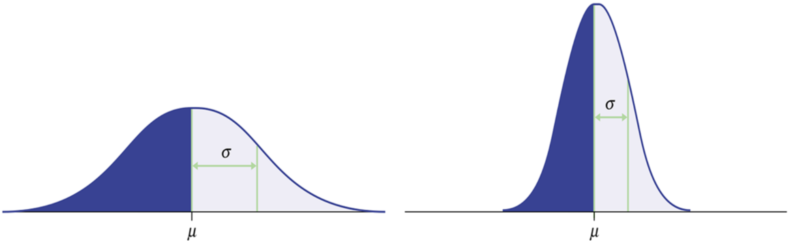
Example: OkCupid, Heights of Males



Example: OkCupid, Heights of Females



The shape of a normal distribution varies by location (mean) and scale (standard deviation)



Moore/Notz, *The Basic Practice of Statistics*, 9e,
© 2021 W. H. Freeman and Company

Figure 4: Changing the mean shifts the “center” of the distribution. Changing the standard deviation alters the “width” of the distribution (i.e. variability).

Standardizing Normal Distributions with Z-Scores

- A ***Z-score*** is the number of standard deviations a value falls above (when positive) or below (when negative) the mean of the data
- Z-scores standardize a normal distribution by...
  - Centering the data at 0 by subtracting the mean from each score
  - Scaling the units of the data to 1 by dividing the centered data by the standard deviation

Calculating the Z-Score

$$\begin{aligned} Z &= \frac{\mathrm{observed value} - \mathrm{mean}}{\mathrm{standard deviation}} \\ &= \frac{x - \mu}{\sigma} \end{aligned}$$

What does “centering” the data mean?

- The numerator of the Z-Score $x - \mu$ calculates how many units an observed value is from the mean of the normal distribution
- When $x_i \approx \mu$, $x_{\text{centered}} \approx 0$
- The units of random variable X_{centered} are the same as the units for the original variable X

Properties of Centered Normal Distributions

For a given random variable X with a normal distribution, you ***center*** the data by calculating $x_i - \mu$ for each value of X such that...

- When $x_i > \mu$, $x_{\text{centered}} > 0$ and is interpreted "x_{centered} is $x_i - \mu$ units *greater than* the mean μ"
- When $x_i = \mu$, $x_{\text{centered}} = 0$
- When $x_i < \mu$, $x_{\text{centered}} < 0$ and is interpreted "x_{centered} is $\mu - x_i$ units *less than* the mean μ"

What does “scaling” the data mean?

- Dividing the numerator of the Z-score $x - \mu$ by the denominator σ converts the units of the centered data to standard deviations
- Converts “x_{centered} is $x_i - \mu$ units greater/lesser than the mean μ” to “x_{scaled} is $x_i - \mu$ units greater/lesser than the mean
- When $x_i - \mu \approx \sigma$, $x_{\mathrm{centered}} \approx 1$
- For scaled data, $1\,\mathrm{unit} = 1\,\mathrm{standarddeviation}$

Example: Test Scores

- SAT scores are normally distributed with $\mu = 1500$ and $\sigma = 300$ ($N(\mu = 1500, \sigma = 300)$)
- ACT scores are normally distributed with $\mu = 21$ and $\sigma = 5$ ($N(21, 5)$)

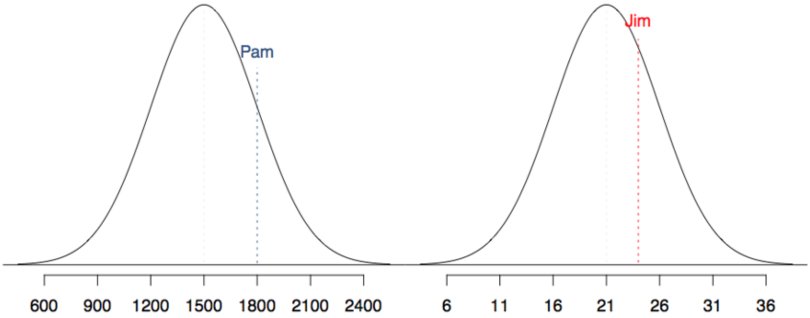


Figure 5: How do we compare normal distributions with different locations and scales? Is Pam more above average than Jim? Vice versa?

Zooming in

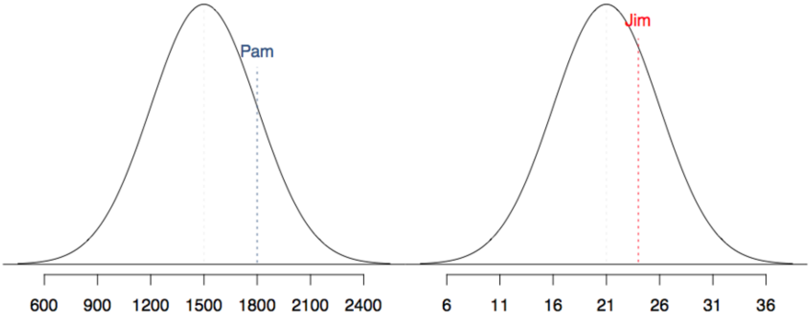


Figure 6: If both Pam and Jim applied to John Carroll, who would be the better student to admit?

Example: Calculating Pam's SAT Z-Score

If SAT scores have the distribution $N(\mu = 1500, \sigma = 300)$ and Pam's SAT score is 1800, then Pam's Z-score is...

$$\begin{aligned}\text{Z-Score} &= \frac{x - \mu}{\sigma} \\ &= \frac{1800 - 1500}{300} \\ &= 1\end{aligned}$$

Pam's SAT Z-score is 1, so Pam's SAT score is ***1 standard deviation greater than the mean***.

Example: Pam's SAT Z-Score in R

```r
pam_mean <- 1500
pam_sd <- 300

pam_centered <- 1800 - pam_mean

print(pam_centered)
```

```
[1] 300
```

```r
pam_z <- pam_centered / pam_sd

print(pam_z)
```

```
[1] 1
```

Example: Pam's SAT Z-Score in R

```
(1800 - 1500) / 300
```

```
[1] 1
```

Example: Calculating Jim's ACT Z-Score

If ACT scores have the distribution $N(\mu = 21, \sigma = 5)$ and Jim's ACT score is 24, then Jim's Z-score is...

$$\begin{aligned}\text{Z-Score} &= \frac{x-\mu}{\sigma} \\ &= \frac{24-21}{5} \\ &= 0.6\end{aligned}$$

Jim's ACT Z-score is 0.6, so Jim's ACT score is ***0.6 standard deviations greater than the mean***.

Putting it together...

- Pam's SAT Z-Score is 1
- Jim's ACT Z-Score is 0.6
- Pam's SAT score is more above average than Jim's ACT score

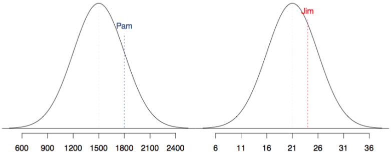


Figure 7: Pam's SAT score and Jim's ACT score on a standardized scale, with center = 0 and 1 unit = 1 standard deviation.

The Standard Normal Distribution

- The ***standard normal distribution*** is a normal distribution with $\mu = 0$ (*centered*) and $\sigma = 1$ (*scaled*)
- The standard normal distribution is written $N(\mu = 0, \sigma = 1))$
- Units of the standard normal distribution are standard deviations (Z-scores) (i.e. 1 unit $=$ 1 SD)
- Observations that are $2+$ standard deviations from the mean are considered unusual

The 68-95-99.7 Rule

When data is (nearly) normally distributed...

- ~68% of the observations are within 1 standard deviation of the mean ($\mu \pm \sigma$)
- ~95% of the observations are within 2 standard deviations of the mean ($\mu \pm 2\sigma$)
- 99.7% of the observations are within 3 standard deviations of the mean ($\mu \pm 3\sigma$)

The 68-95-99.7 Rule

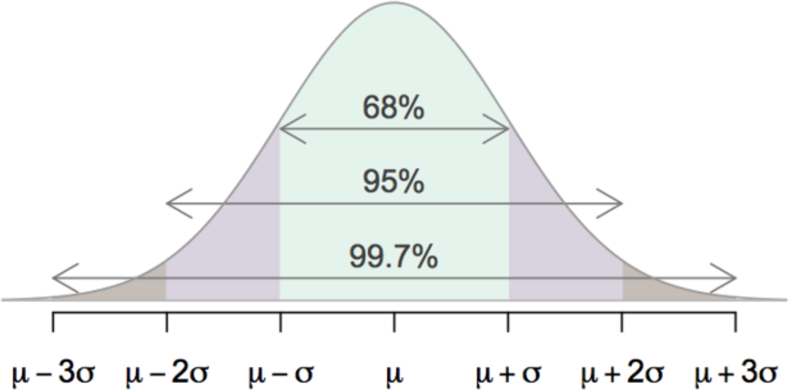


Figure 8: The 68-95-99.7 Rule describes approximately what proportion of the observations should lie within 1, 2, and 3 standard deviations of the mean respectively, if the data is normally distributed

Example: Test Scores

- SAT scores have the distribution $N(1500, 300)$
- $\sim$68% of scores will be 1200-1800
- 95% of scores will be 900-2100
- 99.7% of scores will be 600-2400

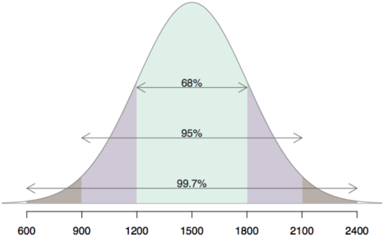


Figure 9: The 68-95-99.7 Rule describes approximately what proportion of the observations should lie within 1, 2, and 3 standard deviations of the mean respectively, if the data is normally distributed

Proportions, Probabilities, and Percentiles

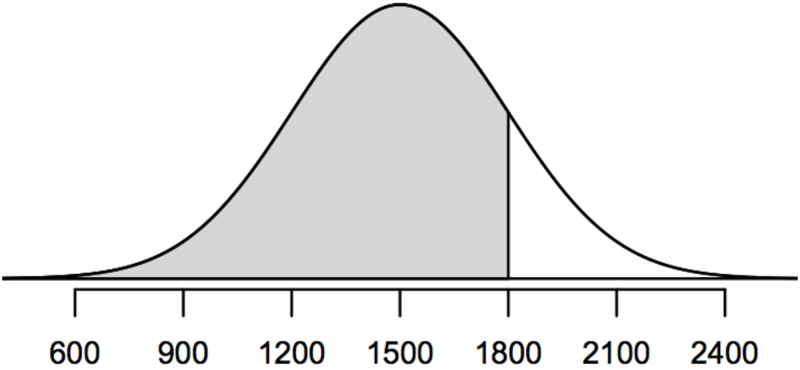


Figure 10: A ***percentile*** is the proportion or percentage of observations that fall *below* a given threshold in a distribution.

Proportions, Probabilities, and Percentiles

$$\begin{aligned}\text{Percentile}(X = x_i) &= \frac{\text{count}(\text{observations} \leq x_i)}{\text{count}(\text{totalobservations})} \\ &= \text{Proportion}(\text{observations} \leq x_i) \\ &= \text{Probability}(\text{anyobservation} \leq x_i)\end{aligned}$$

Probability Density and Cumulative Density Functions

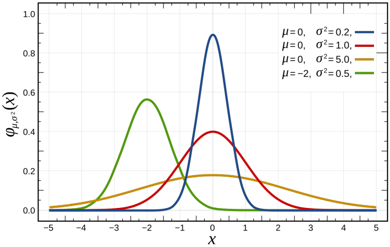


Figure 11: We can calculate the exact probability for a particular value or range of values in a normal distribution.

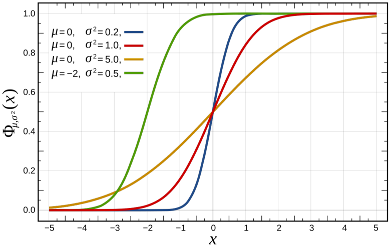


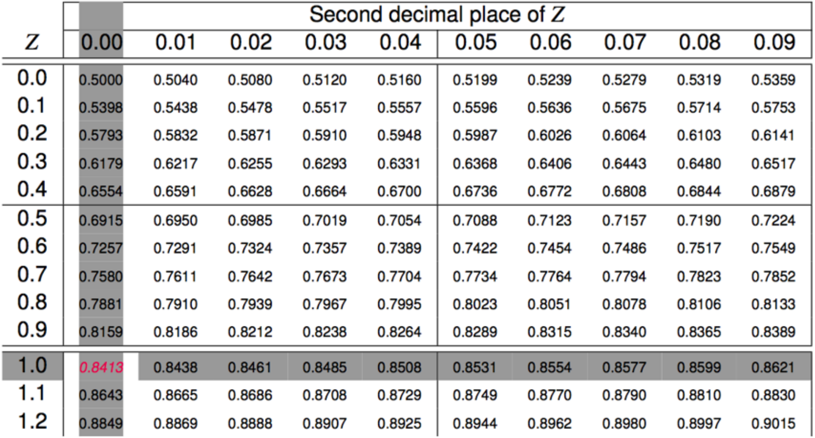
Figure 12: We can calculate the cumulative probability (the percentile) of a variable being less than a given value in a normal distribution.

Probability Density Function for Normal Distributions

You too can calculate probabilities for continuous numeric variables!

$$P(X = x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$

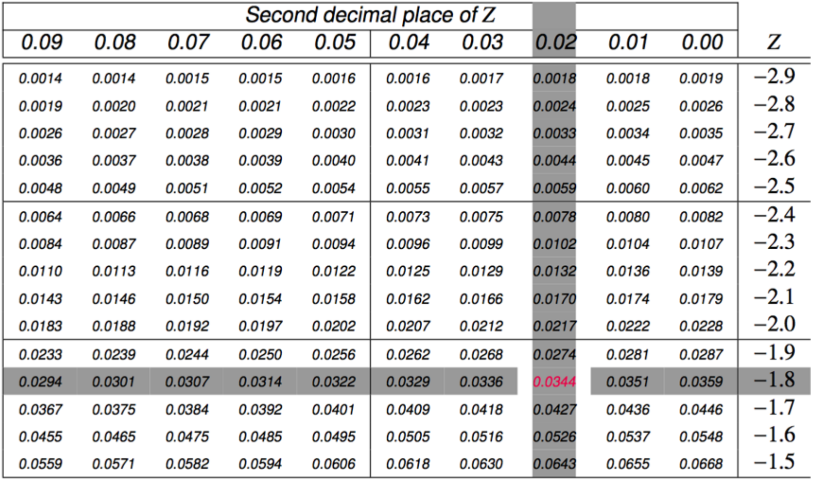
Calculating Percentiles with Z-Score Tables



	Second decimal place of 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	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015

Figure 13: You can use a Z-Score Table to look up the percentile that corresponds to a particular Z-Score for a standard normal distribution.

Calculating Probabilities with Z-Score Tables



Second decimal place of Z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	−2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	−2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	−2.7
0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	−2.6
0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	−2.5
0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	−2.4
0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	−2.3
0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	−2.2
0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	−2.1
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	−2.0
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	−1.9
0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359	−1.8
0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	−1.7
0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	−1.6
0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	−1.5

Figure 14: You can use a Z-Score Table to look up the probability that an observed Z-Score is less than or equal to a given Z-Score (i.e. threshold) for a standard normal distribution.

Calculating Percentiles in R

```
# Parameter 1 = value to look up
# Pam's SAT Z-Score --> Percentile
pnorm(1, mean = 0, sd = 1)
```

```
[1] 0.8413447
```

```
# Parameter 1 = value to look up
# Jim's ACT Z-Score --> Percentile
pnorm(0.6, mean = 0, sd = 1)
```

```
[1] 0.7257469
```

What if we want to know the percent ABOVE a threshold?

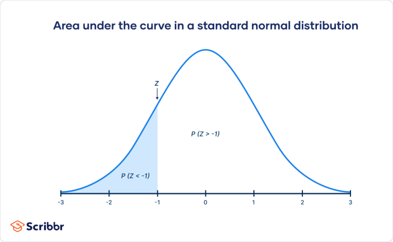


Figure 15: The shaded area under this normal probability distribution is the proportion of observations which are ***less than*** a given threshold.

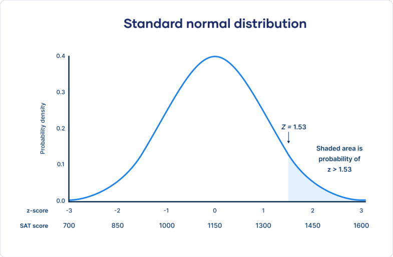


Figure 16: The shaded area under this normal probability distribution is the proportion of observations which are ***greater than*** a given threshold.

Probabilities above a threshold

- Area under a probability curve = 1 (i.e. ***sample space***)
- Probability above a threshold = 1 - percentile of threshold
- $P(X \leq x_i) = 1 - P(X \nleq x_i)$
- $P(X > x_i) = 1 - P(X \leq X_i)$

Easy to find the probability of the complement in R

```r
pnorm(1, mean = 0, sd = 1, lower.tail = F)
```

```
[1] 0.1586553
```

```r
1 - pnorm(1, mean = 0, sd = 1)
```

```
[1] 0.1586553
```

Other: Discrete Numeric Variables

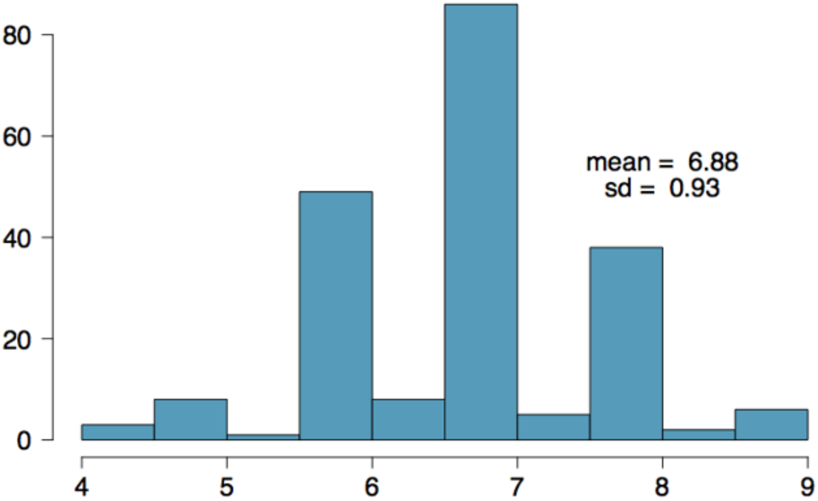


Figure 17: Sometimes the normal distribution is an acceptable approximation of a discrete numeric variable, but other distributions may be more appropriate.

Other: QQ-Plot

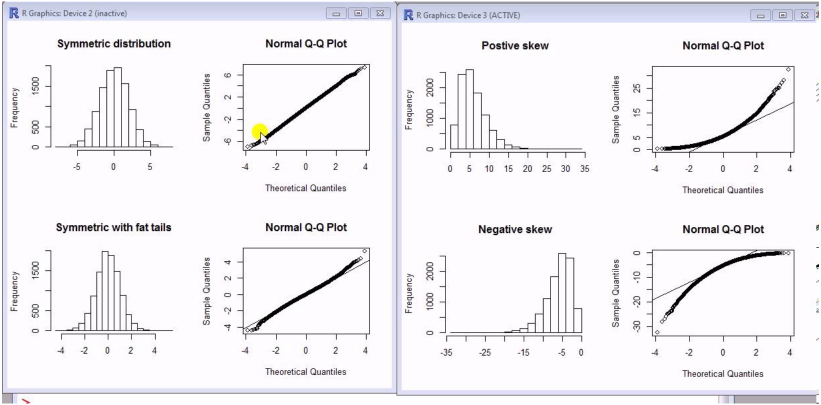


Figure 18: Quantile-Quantile (QQ) Plots can help easily identify when you can and cannot assume normality.