Homework Policy

- Assignments will be finalized 1 hour after class ends: do not start the assignment until then

- This gives me time to edit the assignment in case some concepts were not covered sufficiently in during class

- The Syllabus has been updated to display this info.

Describing data with charts: histograms

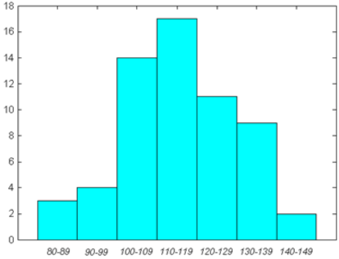
General quantitative data: A **histogram** is a *continuous* barplot for *ranges* of a variable.

TABLE 1.1

IQ test scores for 60 randomly chosen fifth-grade students

145	139	126	122	125	130	96	110	118	118
101	142	134	124	112	109	134	113	81	113
123	94	100	136	109	131	117	110	127	124
106	124	115	133	116	102	127	117	109	137
117	90	103	114	139	101	122	105	97	89
102	108	110	128	114	112	114	102	82	101

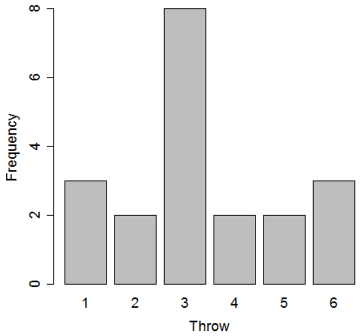
Class	Count
80 – 89	3
90 – 99	4
100 – 109	14
110 – 119	17
120 – 129	11
130 – 139	9
140 - 149	2

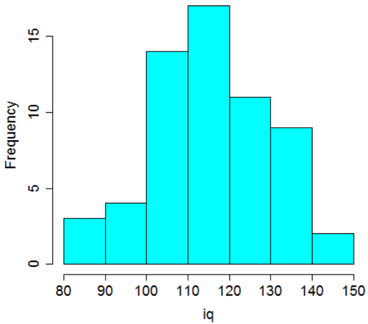


Describing data with charts: bar graphs vs. histograms

Things to notice:

- Bar graphs separate by *value*; histograms separate by *range*.
-
-

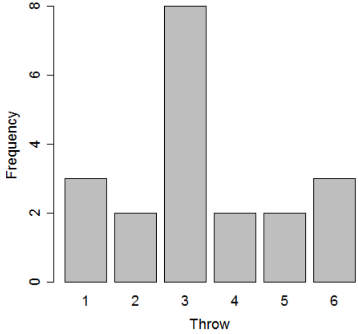


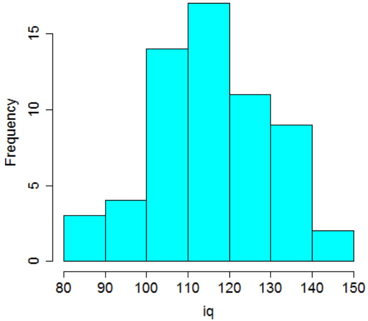


Describing data with charts: bar graphs vs. histograms

Things to notice:

- Bar graphs separate by *value*; histograms separate by *range*.
- Bar graphs have spaces between columns; histograms do not
-

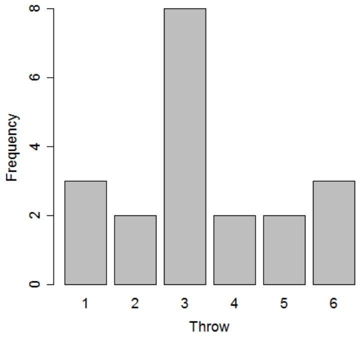


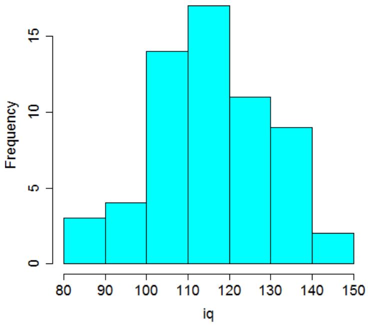


Describing data with charts: bar graphs vs. histograms

Things to notice:

- Bar graphs separate by *value*; histograms separate by *range*.
- Bar graphs have spaces between columns; histograms do not
- Both have frequency versions

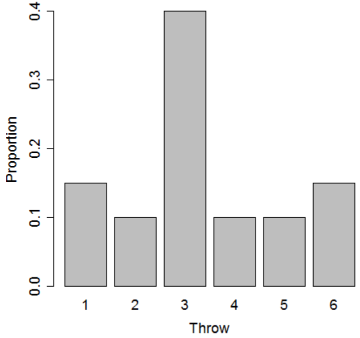


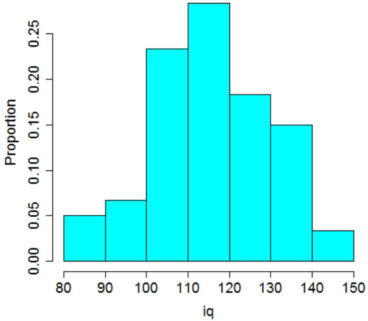


Describing data with charts: bar graphs vs. histograms

Things to notice:

- Bar graphs separate by *value*; histograms separate by *range*.
- Bar graphs have spaces between columns; histograms do not
- Both have frequency versions and both have proportion versions



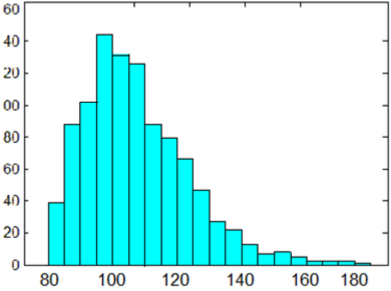


Histograms: distribution shape

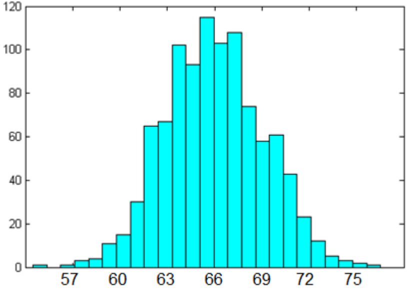
Terminology: *Skewed* data vs. *symmetric* data

- ⋆ Skew is in the direction of the "longer" side

Right-skewed data:

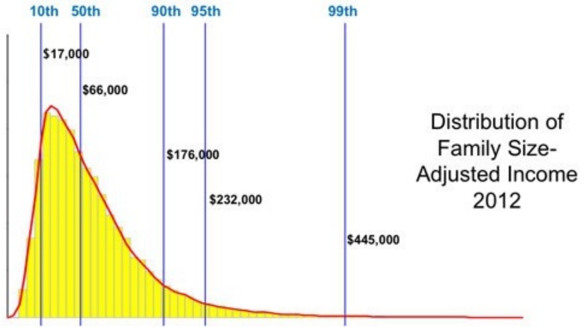


Symmetric data:



Describing densities (Section 1.4)

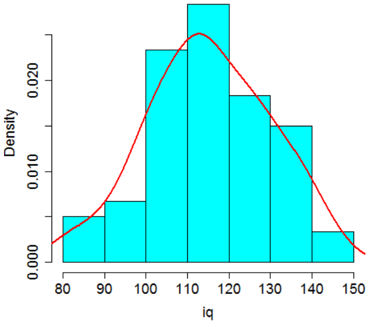
- For symmetric densities, mean and median are the same
- For skewed densities, mean is pulled in direction of skew
- Example: Median below is 66K; mean is much higher



Describing data with densities (Section 1.4)

Density (roughly): a curve which describes data and where it falls

We can find a density that well-approximates a histogram:

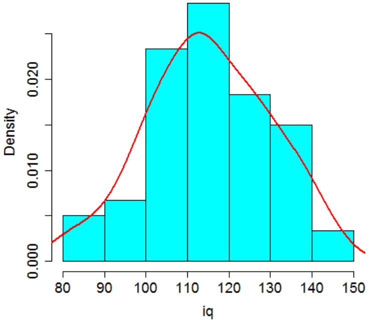


Note: this is a density histogram; *area* under bars is 1

Density definition (Section 1.4)

Density (exactly): a positive line that has area exactly area 1 between it and the horizontal axis.

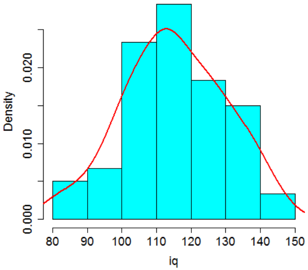
For any two numbers, we can find the area under a density between them. It will always be less than or equal to 1.



Describing data with densities (Section 1.4)

Use/purpose of a density:

- Consider: every histogram represents a sample from a larger population
- A density is like our best guess at the true distribution of the population, given the sample
- For any 2 numbers, area under the density between them is our best guess at the true % between them in the population



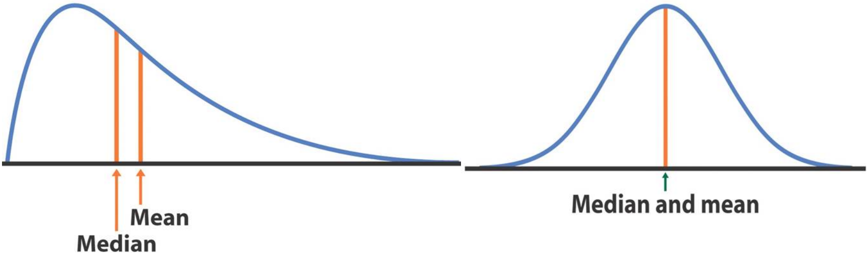
Describing densities (Section 1.4)

Densities have many properties of histograms:

- Median is the point with 50% of the area to the left (and right)
- p-th percentile is the point with p% of area to left
  - Q1 is 25th percentile; Q3 is 75th percentile
- Mode is the highest point of the curve (may not be unique)
- Mean is the center of mass (balance point)
- Right/left skew are analgous

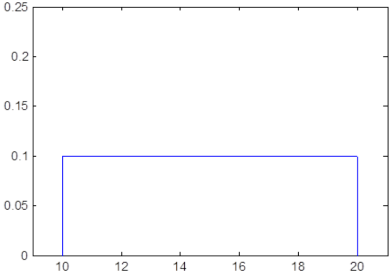
Describing densities (Section 1.4)

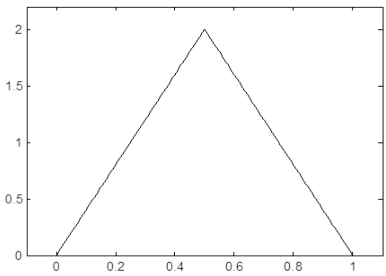
- For symmetric densities, mean and median are the same
- For skewed densities, mean is pulled in direction of skew



Simple densities

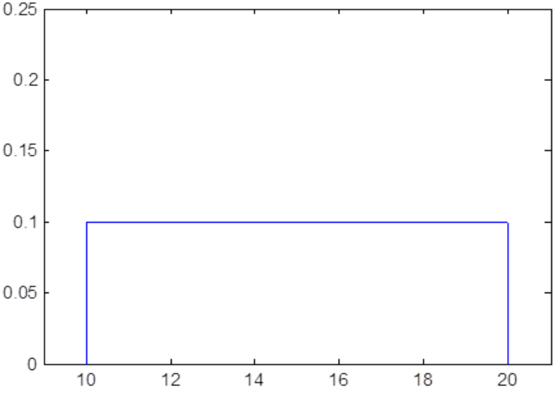
- Densities don't have to be curvy.
- Both of these are densities because the area underneath is 1.
- Left side: **uniform** density. All equal-length intervals take up the same proportion of the population.





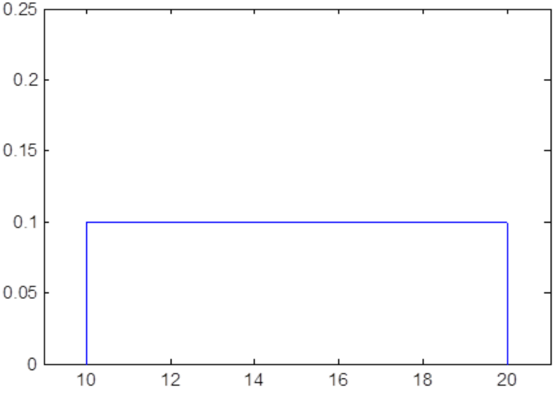
In-class exercise

What is the median of this density? mean? Q1?



In-class exercise

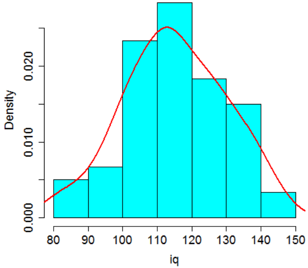
What is the median of this density? mean? Q1?

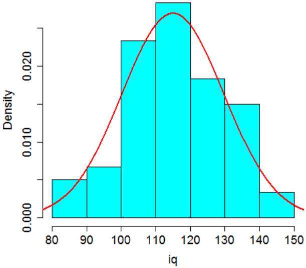


Answer: (15, 15, 12.5)

Describing data with densities (Section 1.4)

- Many different "reasonable" densities
- Not all are mathematically convenient
- Sometimes, worse fitting density is chosen for convenience.
- Left: we fit the "best-fitting" density to the histogram
- Right: we fit the Normal density:



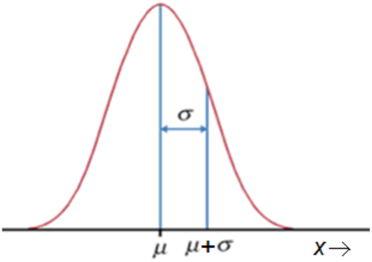


Normal densities (Section 1.4)

- Symmetric, unimodal, and bell-shaped
- Center and spread are controlled by two *parameters*:

$$\mu \text{ the mean, and } \sigma \text{ the standard deviation}$$

- Parameters are like mean and standard error of real data.
- σ extends to "inflection point" of curve

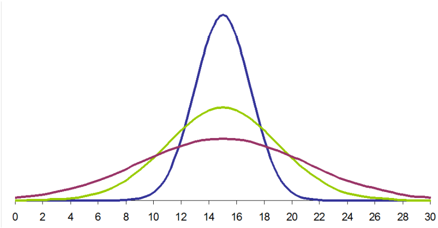


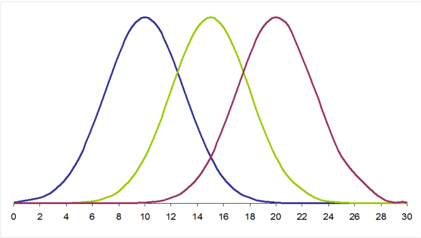
Normal densities (Section 1.4)

Center and spread are controlled by:

μ the mean, and σ the standard deviation

When you change μ and σ, you change the density:

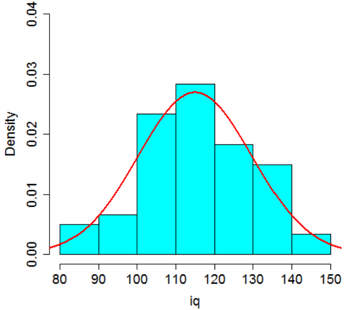




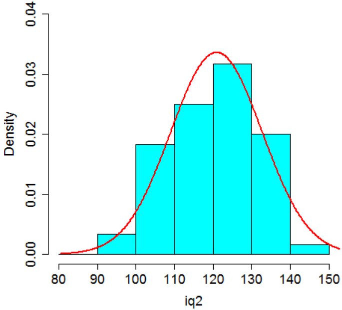
Normal densities (Section 1.4)

- To fit a Normal density to data, set μ and σ to the sample mean and standard error.

$\bar{x} = 115.0$, $s = 14.8$



$\bar{x} = 121.9$, $s = 11.9$



So the right population is "smarter", and with less variance!

Normal densities (Section 1.4)

Why use the Normal density?

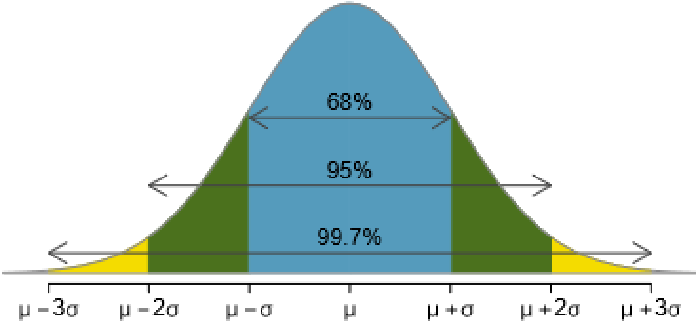
- Normal densities look like many chance outcomes (e.g. coin flip counts)
- ... therefore, many real data sets *are* closely Normal
- Convenience: many stat methods work well w/Normal
- Convenience2: has handy properties to describe data (next)

But be careful! Some data sets are obviously non-Normal. Important to recognize when this occurs (later in course).

Describing data with the Normal (Section 1.4)

68-95-99.7 Rule: under a Normal density with mean μ and standard deviation σ, there is:

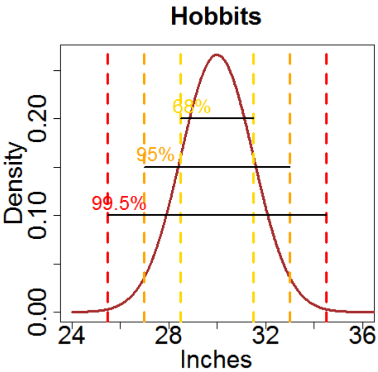
- 68% of the data within 1σ of μ
- 95% of the data within 2σ of μ
- 99.7% of the data with 3σ of μ



Describing data with the Normal (Section 1.4)

Example of 68-95-99.7. Suppose heights of Hobbits follow a Normal density with $\mu = 30$ inches and $\sigma = 1.5$ inches. Then:

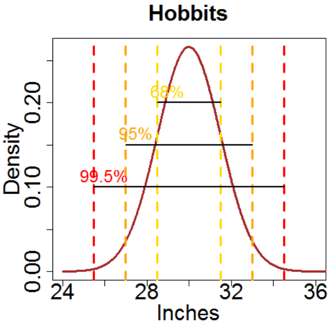
- 68% of Hobbits are within 1.5 inches of 30 inches
- 95% of Hobbits are within 3 inches of 30 inches
- 99.7% of Hobbits are within 4.5 inches of 30 inches



In-class thought exercise

Rule: $(68, 95, 99.7)\%$ of data is within $(1, 2, 3)\sigma$ of μ

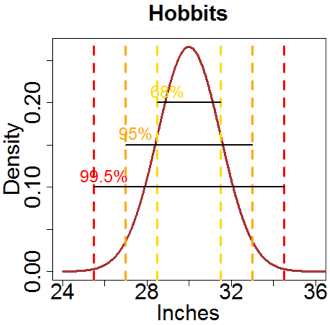
Heights of Hobbits are $\mathcal{N}(30, 1.5)$. Suppose Frodo is 33 inches tall. What proportion of Hobbits are shorter than Frodo?



In-class thought exercise

Rule: $(68, 95, 99.7)\%$ of data is within $(1, 2, 3)\sigma$ of μ

Heights of Hobbits are $\mathcal{N}(30, 1.5)$. Suppose Frodo is 33 inches tall. What proportion of Hobbits are shorter than Frodo?



Answer: 97.5%

In-class thought exercise

Rule: $(68, 95, 99.7)\%$ of data is within $(1, 2, 3)\sigma$ of μ

Heights of Hobbits are $\mathcal{N}(30, 1.5)$. Suppose Frodo is 33 inches tall. What proportion of Hobbits are shorter than Frodo?

Answer: 97.5%

Heights of Elves are $\mathcal{N}(72, 3)$. Suppose Legolas is 78 inches tall (6-foot-6!). What proportion of Elves are shorter than Legolas?

In-class thought exercise

Rule: $(68, 95, 99.7)\%$ of data is within $(1, 2, 3)\sigma$ of μ

Heights of Hobbits are $\mathcal{N}(30, 1.5)$. Suppose Frodo is 33 inches tall. What proportion of Hobbits are shorter than Frodo?

Answer: 97.5%

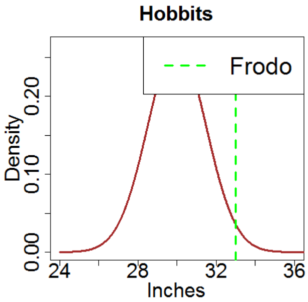
Heights of Elves are $\mathcal{N}(72, 3)$. Suppose Legolas is 78 inches tall (6-foot-6!). What proportion of Elves are shorter than Legolas?

Answer: 97.5%

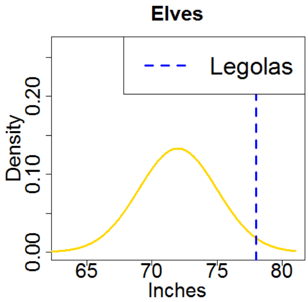
In-class thought exercise

Rule: $(68, 95, 99.7)\%$ of data is within $(1, 2, 3)\sigma$ of μ

Hobbits are $\mathcal{N}(30, 1.5)$:



Elves are $\mathcal{N}(72, 3)$:

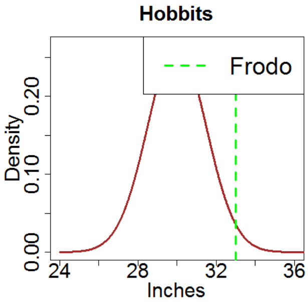


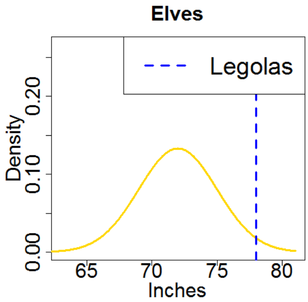
In-class thought exercise

Rule: (68, 95, 99.7)% of data is within $(1, 2, 3)\sigma$ of μ

Hobbits are $\mathcal{N}(30, 1.5)$:

Elves are $\mathcal{N}(72, 3)$:





- Point: Frodo and Legolas are at the same *percentile*
- Is there a *standardized* unit that could show this?

Normal z-scores (Section 1.4)

For a point x from a $\mathcal{N}(\mu, \sigma)$ population, the z-**score** is defined

$$z = \frac{x - \mu}{\sigma}$$

Data at same percentile have the same z-score (& vice-versa)

Hobbits are $\mathcal{N}(30, 1.5)$, Frodo is 33 inches tall. His z-score is

$$\frac{33 - 30}{1.5} = 2$$

Elves are $\mathcal{N}(72, 3)$, Legolas is 78 inches tall. His z-score is

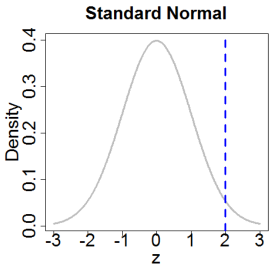
$$\frac{78 - 72}{3} = 2$$

$\star\star\star$ So a z-score **counts sigmas** between x and μ!

Normal z-scores (Section 1.4)

$$z = \frac{x - \mu}{\sigma}$$

- z values are have a Normal $\mu = 0$, $\sigma = 1$ density (called "Standard Normal")
- ... thus, the 68 - 95 - 99.7 rule applies to z-scores too
- Notice $z = 2$ is just 2σ when $\sigma = 1$
- ... and 97.5% of z-values are below $z = 2$

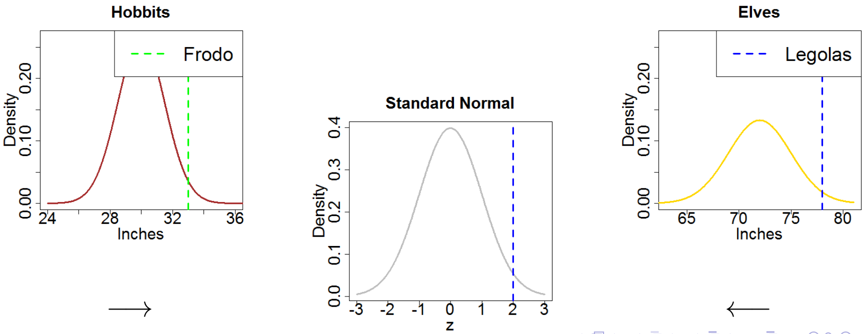


Normal z-scores (Section 1.4)

- $z = \frac{x-\mu}{\sigma}$

Frodo's z-score: $\dfrac{33-30}{1.5} = 2$

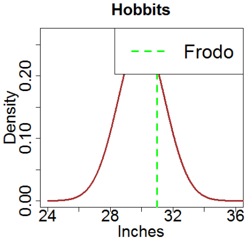
Legolas's z-score: $\dfrac{78-72}{3} = 2$



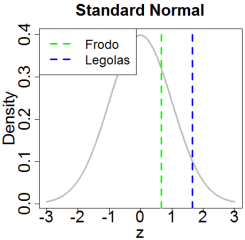
Normal z-scores (Section 1.4)

- What happens when things aren't as easy?

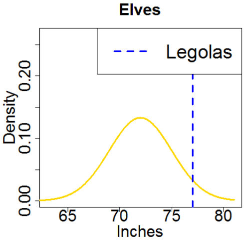
Frodo's z-score: $\dfrac{31-30}{1.5} \approx 0.66$ Legolas's z-score: $\dfrac{77-72}{3} \approx 1.66$



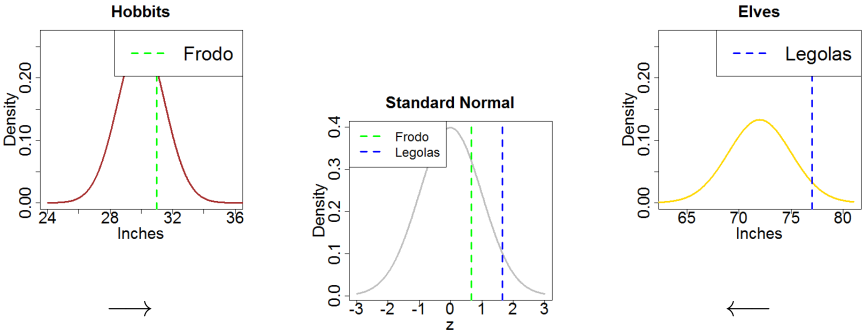
→



←



Normal z-scores (Section 1.4)



- Frodo and Legolas are now different percentiles of their populations
- How do we know what percentiles they are?
- Can't use 68-95-99.7 rule: their z-scores aren't integers

Normal tables (Section 1.4)

- Every *z*-score has a **cumulative** proportion before it, given by the Standard Normal density
- *z* proportions cannot be computed directly
- Need to use a table (Table A in your textbook):

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	[illegible]	0.6026
0.3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.6368	0.6406
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	[illegible]	0.6772
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515

- *z*-scores in the margins
- Proportions in the table
- Top margin is "completion" of side margin

Normal tables (Section 1.4)

- z-scores can also be negative
- If a Hobbit is $26\frac{1}{4}$ inches, the z-score is $\frac{26.25-30}{1.5} = \frac{-3.75}{1.5} = -2.5$. What % of Hobbits are shorter?

z	.00	.01	.02	.03	.04
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618