Introduction to Statistics I

Instructor: Jodin Morey moreyj@lemoyne.edu

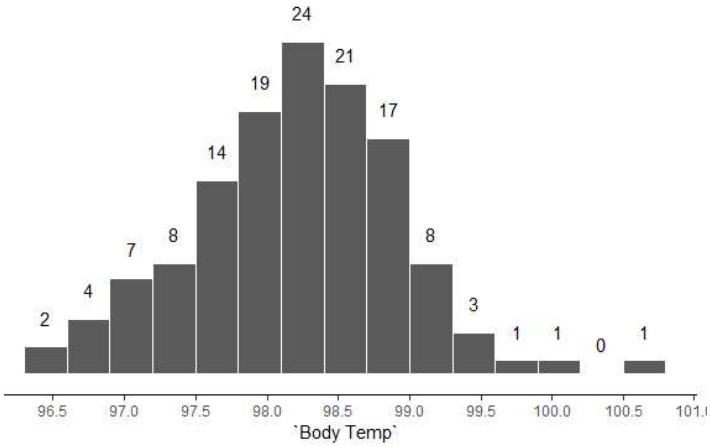
Previous Lecture

- Five Number Summary (FNS)
- Box plot 
- Outliers ●●● ●● ○
- Modified Box Plot

Topic 12: Normal Distributions

What are normal distributions?

Example: Body Temperatures



$n = 130$

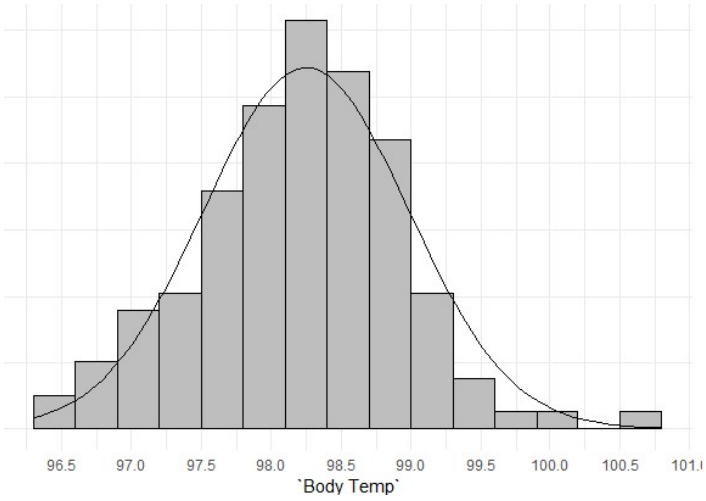
I take my temperature, and it's 97.5° or 99°. Are these concerning?

- % in this sample below 97.5°?
- % in this sample above 99.0°?
- Is this a histogram of a sample or the population (people of the world)?
- How do we estimate proportions in a **population** when we have proportions from a **sample**?

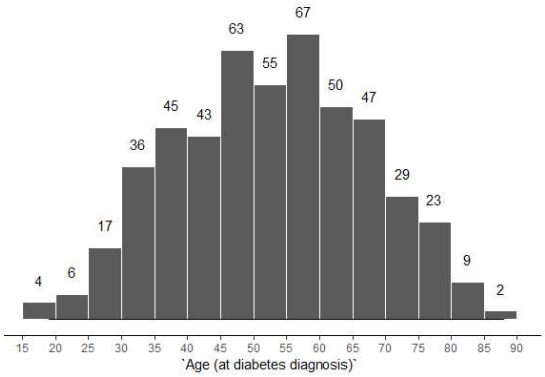
Using exact sample numbers leads to **overfitting**, where we assume the population behaves identically to our sample.

Instead, let's fit a mathematical model.

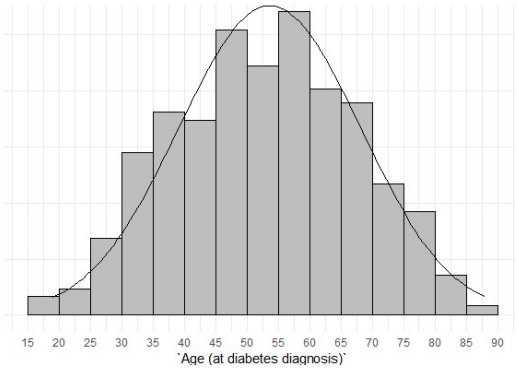
The curve is a function/model that "fits" the data.



Body Temps w/a Mathematical Model



Diabetes Diagnosis vs Age



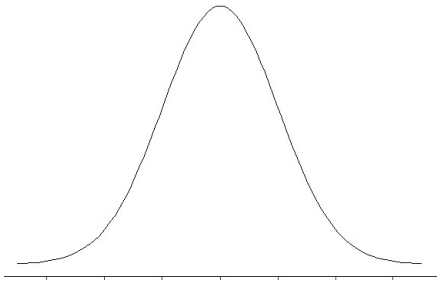
Diabetes Diagnosis vs Age w/Mathematical Model

The curves for diabetes/body temps look similar!

Normal Curve

The **Normal Dist**. was discovered in 1809; in an attempt to locate the dwarf planet Ceres.

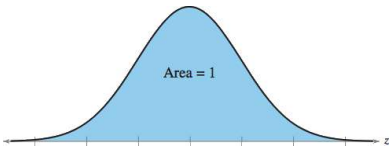
Carl Gauss

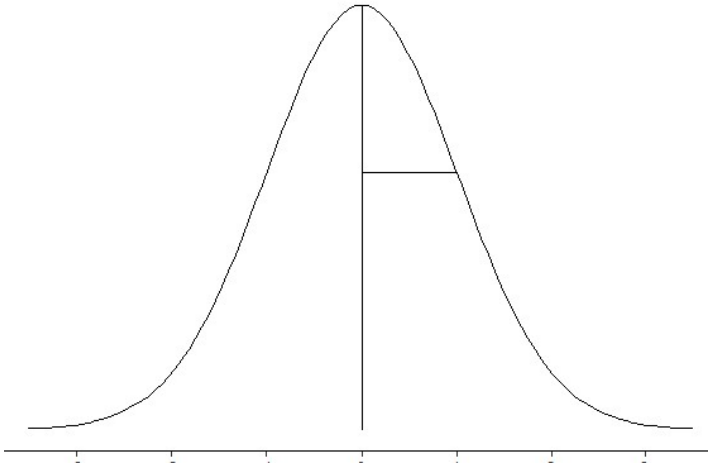


Normal Distribution

Ceres

The only difference between the body temps curve, and a normal distr is that to generate a normal distr, you must scale the curve so that the area under the curve is equal to one.





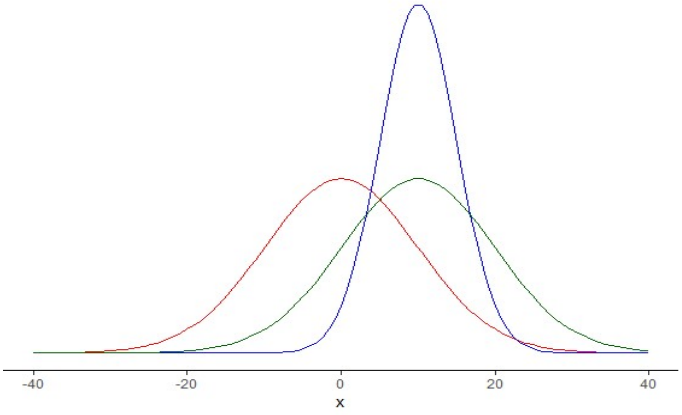
There's only one normal distr for every **mean** & **SD**.

What's different between:

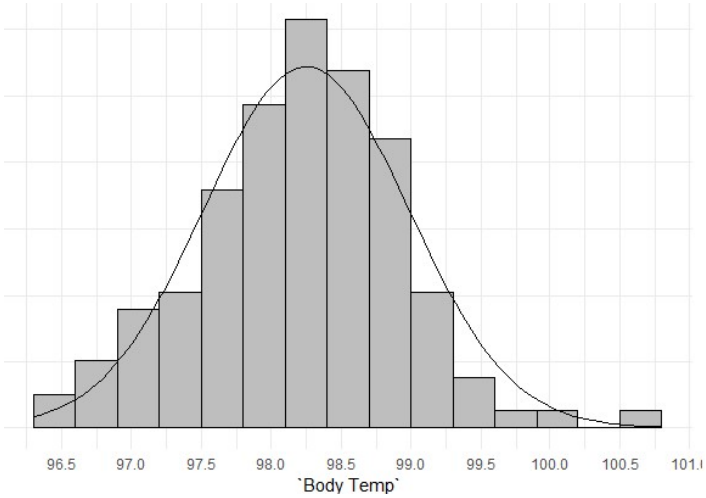
red & green dists?

red & blue dists?

blue & green dists?



Normal Distrs.

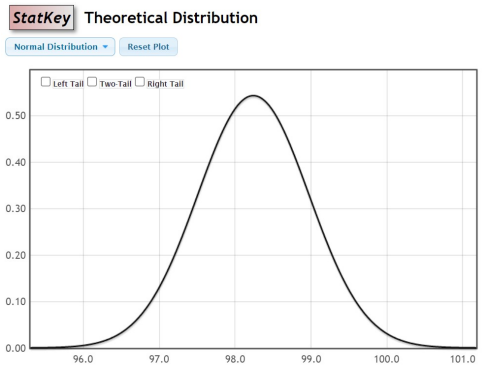


bit.ly/introstatsdata

Applets: Normal Distr Generator

Percent of people in this sample below 97.5°?

Using a Normal Model



To draw a normal curve, we need mean and SD.

Mean is 98.24. SD is 0.734.

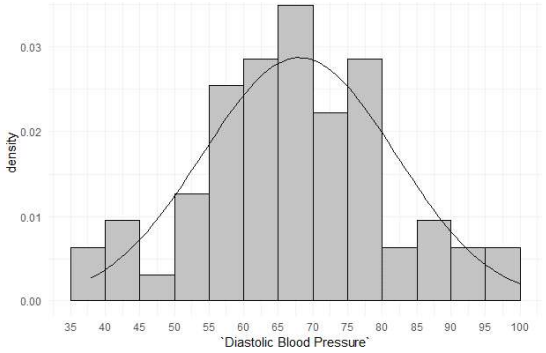
So (using applet), area under the curve below 97.5 is 0.157.

Interpreting Probabilities

What does this 0.157 mean?

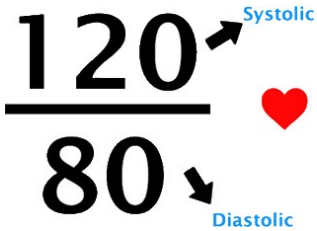
- ≈ 15.7% of body temps are less than 97.5.
- Repeated sampling would find ≈ 15.7% of people in each sample would have body temps less than 97.5.
- Probability that randomly selected person has temp $< 97.5°$ is ≈ 15.7%. Or, $P(\text{temp} < 97.5°) = 0.157$.

Blood Pressure Example



Probability of observing someone w/diastolic less than 54?

We need the mean (68) & SD (13.9) to use the normal model.



bit.ly/introstatsdata

Applets: Normal Dist Generator

Conclusion:

- Probability of observing someone w/diastolic < 54 is 0.157.

  $P(\text{BP} < 54) = 0.157$.

- Proportion of population w/diastolic < 54 is 0.157.

  Eg: 15.7% of people have diastolic BPs below 54.

---

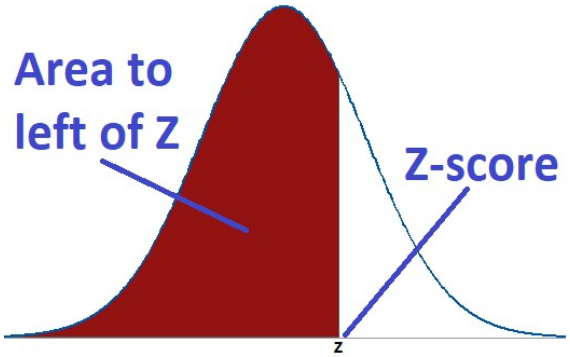
z-scores

Temps & BP

Recall: Probability of observing temp < 97.5 is 0.157.

Probability of observing diastolic < 54 is 0.157.

Why are these both the same??



bit.ly/introstatsdata

Normal Dist Generator

Recall: Body temp mean is 98.24 & SD is 0.734. z-score for 97.5?

$$z = \frac{97.5-98.24}{0.734} \approx -1.01.$$

Recall: BP mean is 68 & SD is 13.9. z-score for 54?

$$z = \frac{54-68}{13.9} \approx -1.01.$$

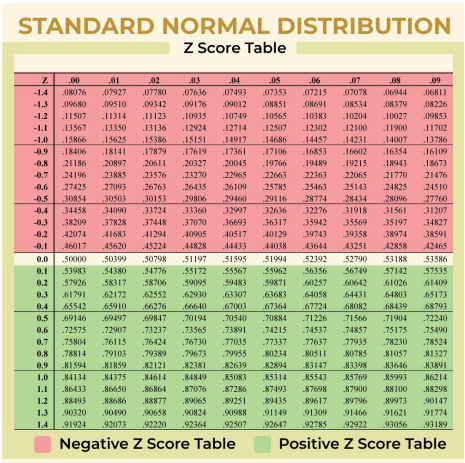
Observations w/same z-scores give same probabilities, regardless of context.

We use **z-score calculators** to calculate probabilities, (in the old days, z-tables).

Standard Normal: normal dist. with mean 0 and SD of 1.

Probability associated w/z-score found by using std normal probability table.

Table value (TV) for that z-score gives probability of observing that value or lower.



Incomplete z-score table

To find probability of observing less than a certain value x using normal model:

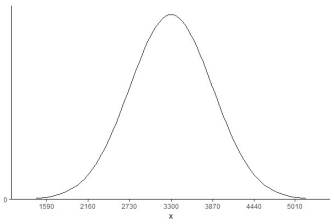
- Find the z-score
- Use z-score calculator to discover area to the left or right of the z-score.

bit.ly/introstatsdata

Calculators: z-score calculator

Example: US Birthweights in 2004 modeled by normal distr.

w/mean 3300 g (≈ 7.3 lbs) and SD 570 g (≈ 1.3 lbs).



Birthweights

Babies weighing less than 2500 g (≈ 5.5 lbs) are **low birthweight**.

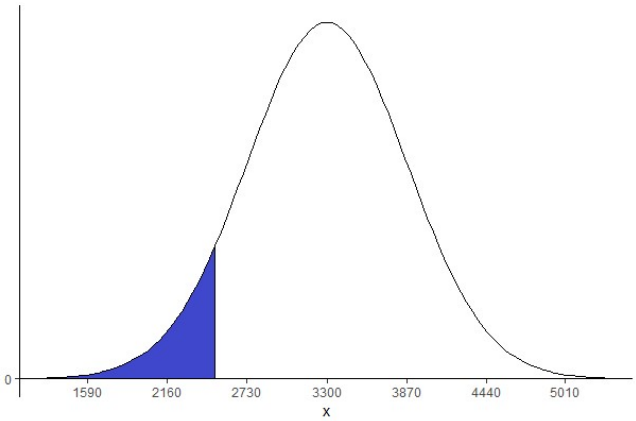
P(birthweight < 2500)? (use a z-score calculator)

z-score for 2500:

$$z = \frac{2500-3300}{570} \approx -1.40.$$

Area to the left for $z = -1.40$:

0.0808



Low birthweights

Probability of low birthweight is 0.0808:

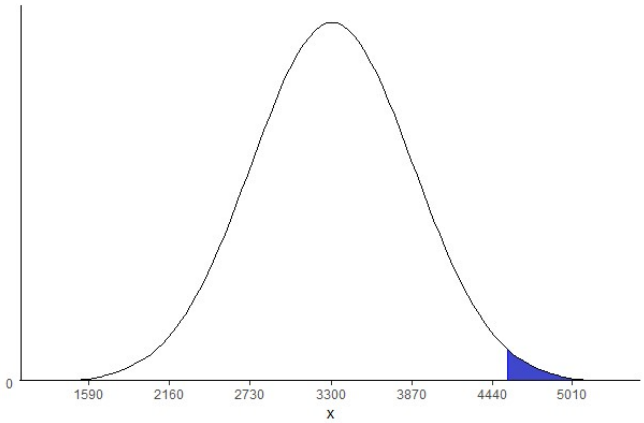
P(birthweight < 2500) = 0.0808.

Proportion of babies we predict to weigh more than 10 lbs (4536 g)?

$$z = \frac{4536-3300}{570} \approx 2.17.$$

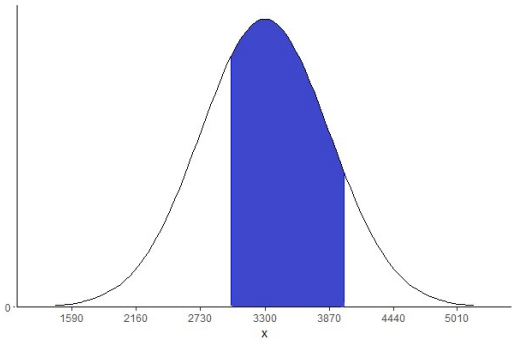
Area to the right for $z = 2.17$ is:

0.0150



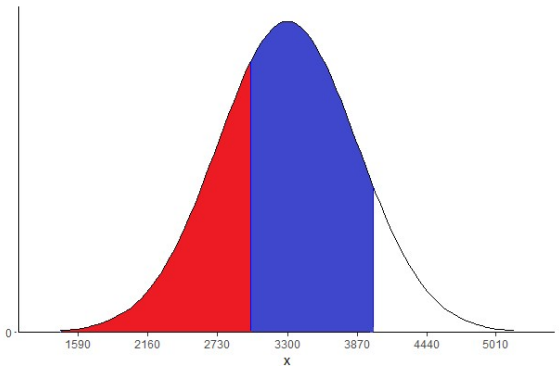
High Birthweights

Mid-Weight Babies



The probability that a randomly selected baby weighs between 3000 and 4000 g?

Recall the mean is 3300, SD is 570.



Need: z-scores to find $P(\text{birthweight} < 3000)$.

Also, z-scores to find $P(\text{birthweight} < 4000)$.

How to find probability of being between them?

$$z_{3000} = \frac{3000-3300}{570} \approx -0.5263.$$

Area to the left:

0.2993.

The z-score for 4000?

$$z_{4000} = \frac{4000-3300}{570} \approx 1.2281.$$ Area to the left: 0.8903

So, probability of between 3000 g and 4000 g:

$0.8903 - 0.2993 = 0.591$

Proportion of babies born w/weights between 3000 and 4000 g is 0.591.

Putting it all together, **when calculation probabilities**:

- Calculate z-score
- Look-up z-score area (either to the left or right)
- If probability is between two scores, calculate the z-scores and areas to the left, and find the difference (larger minus smaller).

Activities: **12-2**

bit.ly/introstatsdata

Calculators: *z*-score calculator

What did we learn?

- Normal distr for each mean and SD.
- Probability of observing data pt less/greater than some other value
- *z*-scores, areas under normal curves