Introduction to Statistics I

Textbook: Elementary Statistics (4th Edition, by Navidi and Monk), and Workshop Statistics (4th Edition, by Rossman and Chance).

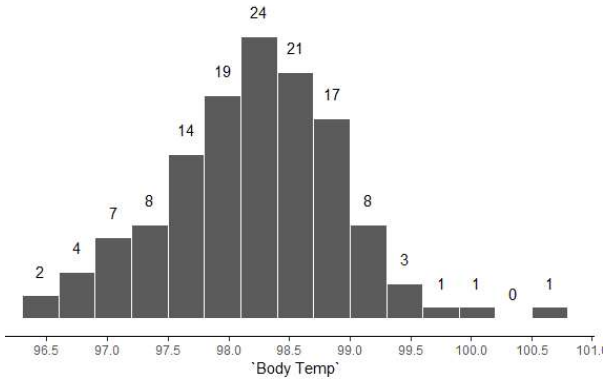
Previous Lecture

- Calculate residual $e = y - \hat{y}$ between data and regression line $\hat{y}$
- Fit *least squares line* to data by minimizing SSE: $e_{11}^2 + e_{12}^2 + \ldots + e_n^2$
- Interpret slope/intercept
- Make predictions

§7.1 & §7.2: Normal Curves

Normal distributions are distrs that are mound shaped, and occur frequently in real life.

Example: Body Temperatures. Below you find a frequency histogram of the body temps of 130 people. Because your temp can be anything (92, 91.3, 94.75, etc.), and is not part of a predetermined discrete set of numbers (like the 6 results of a die roll), temp is a **continuous random variable** (as opposed to a **discrete random variable**).



$n = 130$

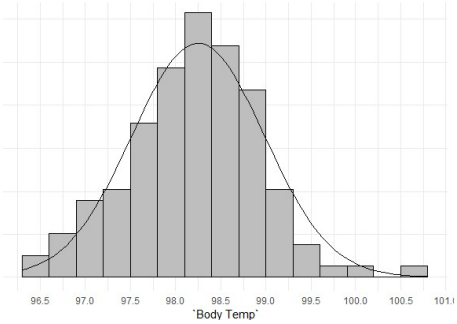
If I take my temperature, and it's 97.5° or 99°. Are these concerning?

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

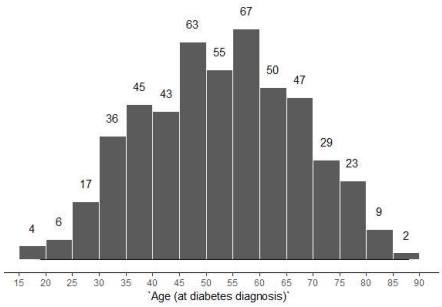
Using our sample estimates leads to **overfitting**, where we assume the population behaves identically to our sample. Instead, let's fit our data to a **mathematical model**.

not *that* kind of model

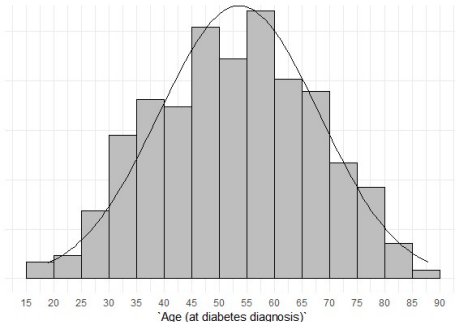
The graphed curve is a function/model that "fits" the data.



Body Temps w/a Mathematical Model



Diabetes Diagnosis vs Age



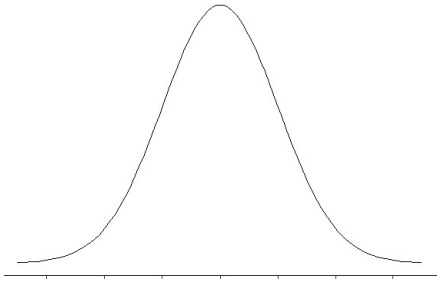
Diabetes Diagnosis vs Age w/Mathematical Model

The curves/distr for diabetes and body temps look similar!

Normal Curve/Distr

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

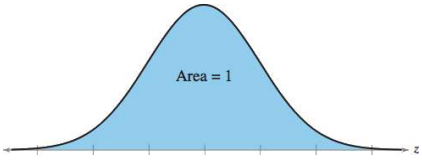
Carl Gauss



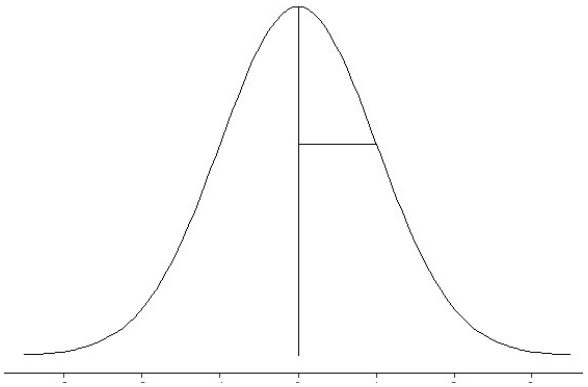
Normal Distribution

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.



Definition: This type of curve, which describes the distr of a continuous random variable, is a *probability density curve*. The area under the curve (once scaled) near a value (temp), is roughly the probability that a random observation will give a similar value (temp).



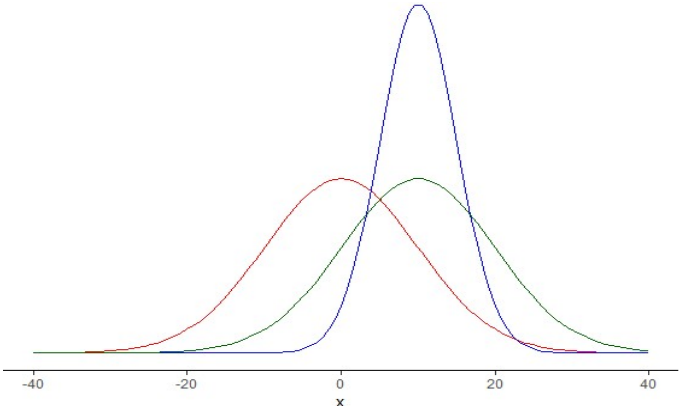
There's only one normal distr for every **mean** & **SD**.

What's different between:

red & green distrs?

red & blue distrs?

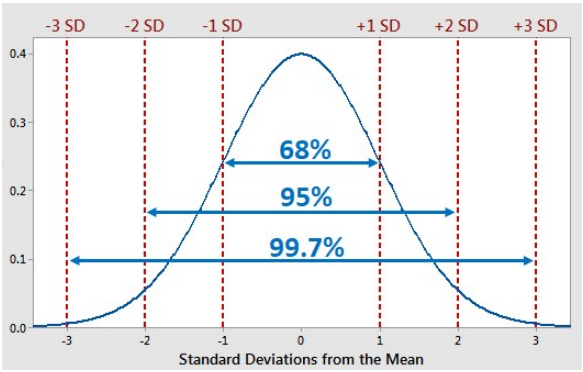
blue & green distrs?



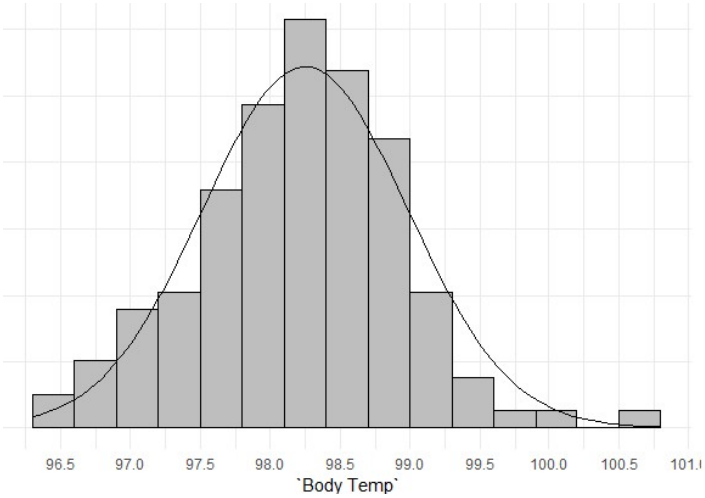
Normal Distrs.

Properties of Normal Distr

- One Mode
- Symmetric around the mode
- Mean and median are equal to the mode
- Extends to both positive and negative infinity
- Follows the Empirical Rule



Back to Our Example:

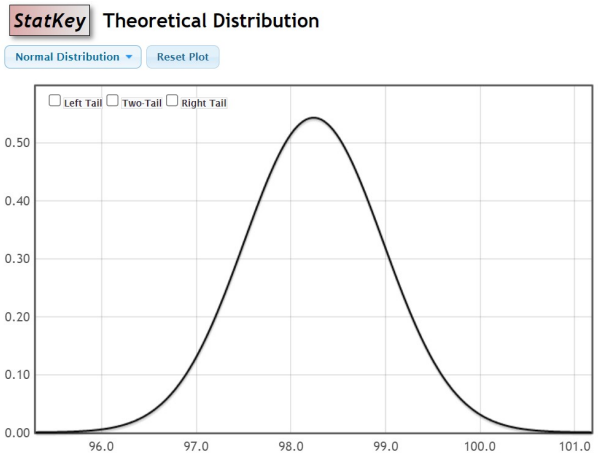


bit.ly/introstatsdata

Applets: Normal Distr Generator

Percent of people in this the *population* below 97.5°?

Using a Normal Model



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

Edit applet's params: Use our sample mean of 98.24. SD is 0.734.

Choose "left tail" and set to 97.5.

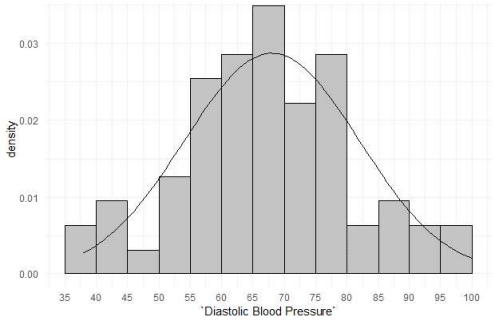
Applet says the area under the curve below 97.5 is 0.157.

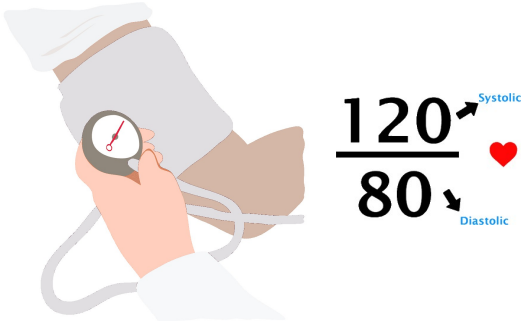
Interpreting Probabilities

What does this 0.157 mean?

- ≈ 15.7% of population's 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 a randomly selected person has temp $< 97.5°$ is ≈ 15.7%. Or, $P(\text{temp} < 97.5°) = 0.157$.

Blood Pressure Example





In your samples, you find mean diastolic of 68 & SD of 13.9. One patient has diastolic of 54.

Probability of observing someone w/diastolic less than 54?

Using the applet: set the mean (68) & SD (13.9) to use the normal model.

bit.ly/introstatsdata

Applets: Normal Dist Generator

Conclusion:

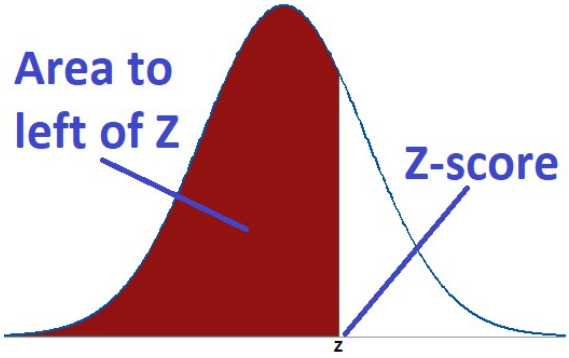
- Probability of observing someone w/diastolic < 54 is 0.157.

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

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

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

Recall z-Scores



bit.ly/introstatsdata

Normal Distr Generator

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??

Recall: Body temp mean is 98.24 & SD is 0.734. What's the 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. What's the z-score for 54?

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

This process of calculating the z-score is called *standardization*.
If you applied the process to all your sample data, the data's new distr would be centered at zero, w/a SD of one.

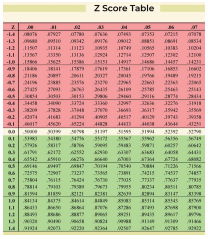
Standard Normal: A normal distr with mean 0 and SD of 1.

So, standardization alter's your data to be similar to a std normal distr.

And since we know the probabilities (probs) associated w/the std normal distr (using online calculator), we can use the z-score to determine the probs associated with a particular data pt from our sample.

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

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



Incomplete z-score table

Prob associated w/z-score found by using std normal prob table.

The **table value** for that z-score gives the prob of observing that value or lower.

To find prob of observing less/more than a certain value x using the normal model:

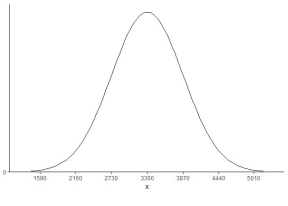
- Find the z-score
- Use z-score calculator to discover area to the left or right of the z-score.

bit.ly/introstatsdata

Calculators: z-score calculator

Example: US birthweights in 2004 are modeled by a normal distr w/mean 3300 g (≈ 7.3 lbs) and SD 570 g (≈ 1.3 lbs).

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



Birthweights

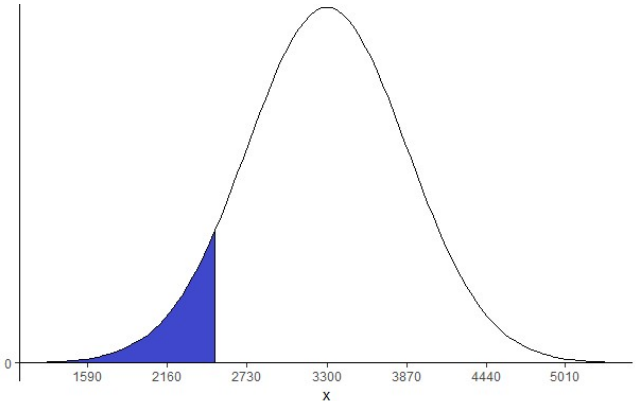
P(birthweight < 2500)? (use a z-score calculator)

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 8.08%.

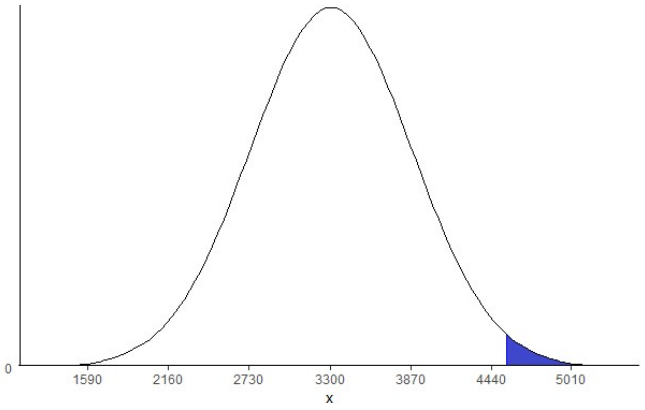
P(birthweight < 2500) = 0.0808.

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

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

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

0.0150.

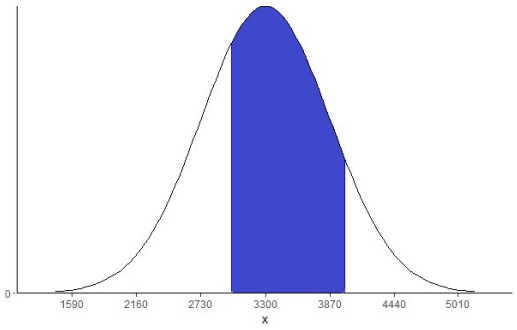


High Birthweights

Mid-Weight Babies

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

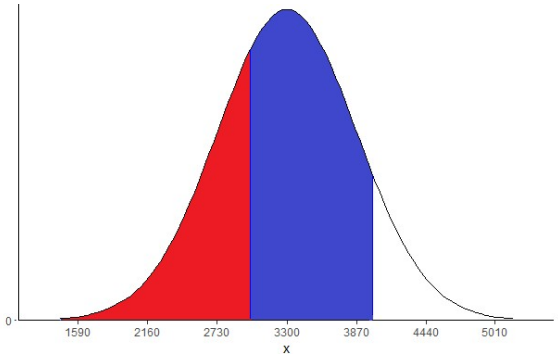
Recall the mean is 3300, SD is 570. ...



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

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

How to find prob 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, prob 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 calculating probs**:

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

Activities: 7.2a

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 value
- *z*-scores, areas under normal curves