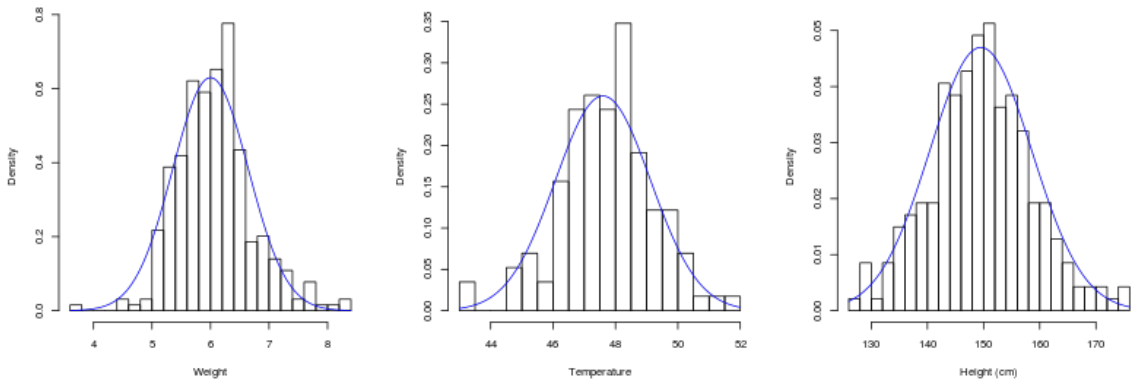
Normal Distributions

Supplement to Section 1.5: Inference for a Single Proportion

Stacey Hancock

Look at these three different data sets. Each histogram is overlaid with a "curve":

A B C



A) Weights (g) of newly born lab rat pups. B) Mean annual temperatures ($^\circ F$) in Ann Arbor, Michigan. C) Heights (cm) of 14 year old boys in Oxford, England.

What differs between these three distributions? What characteristics are similar? Many distributions we look at have a shape similar to those above: most of the data lies close to the mean, and the left and right sides are symmetric. We call it "bell-shaped" and the best example is called the "Normal" distribution. Normal distributions all have the same shape; they differ only in mean μ and standard deviation σ.

Important fact:
Statistics vary from sample to sample, and the pattern is predictable. For many statistics, the pattern of the sampling distribution resembles a normal distribution with a bell-shaped curve.

Studying the normal distribution will allow us to find probabilities for statistical inference which do not require running simulations.

1 Properties of Normal Distributions

The **Standard Normal Distribution** has mean $\mu = 0$ and standard deviation $\sigma = 1$. We can "standardize" any normal distribution to make it have mean 0 and standard deviation 1 by subtracting μ and dividing by σ. If a random variable, X, has a normal distribution with mean μ and standard deviation σ, then

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

has a Standard Normal Distribution. We use the standardized versions to say how many standard deviations (σ's) an observation is from the mean (μ). For example, suppose birth weights of full term

babies have a normal distribution with mean $\mu = 3000$ grams and standard deviation $\sigma = 700$ grams. Then a baby who weighs 3500 grams is $(3500 - 3000)/700 = 0.714$ standard deviations above the mean birth weight.

The **Empirical Rule** states that for variable X that has a normal distribution with mean μ and standard deviation σ, approximately

- 68% of the values of X fall within 1 standard deviation of the mean in either direction $(\mu \pm \sigma)$
- 95% of the values of X fall within 2 standard deviations of the mean in either direction $(\mu \pm 2\sigma)$
- 99.7% (almost all) of the values of X fall within 3 standard deviations of the mean in either direction $(\mu \pm 3\sigma)$.

For example, approximately 95% of babies have birth weights between $3000 - 2(700) = 1600$ grams and $3000 + 2(700) = 4400$ grams.

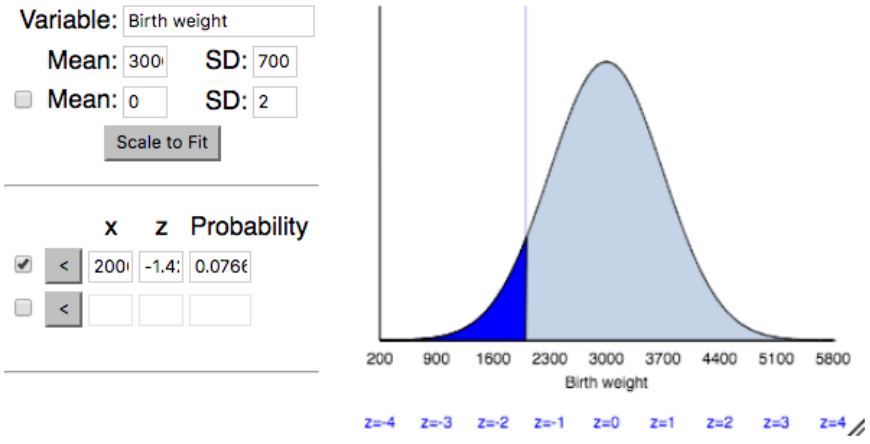
We calculate **probabilities** of a normally distributed random variable by finding areas under the normal curve. The applet

http://www.rossmanchance.com/applets/NormCalc.html

will calculate these areas for us. Use the applet to answer the following questions about the distribution of birth weights specified previously (mean 3000 grams and standard deviation 700 grams). For all questions below, enter "Birth weight" for the Variable name, 3000 for the Mean, and 700 for the SD. Then click "Scale to Fit." (You can ignore the second row of Mean and SD – this would add another normal distribution to the plot.

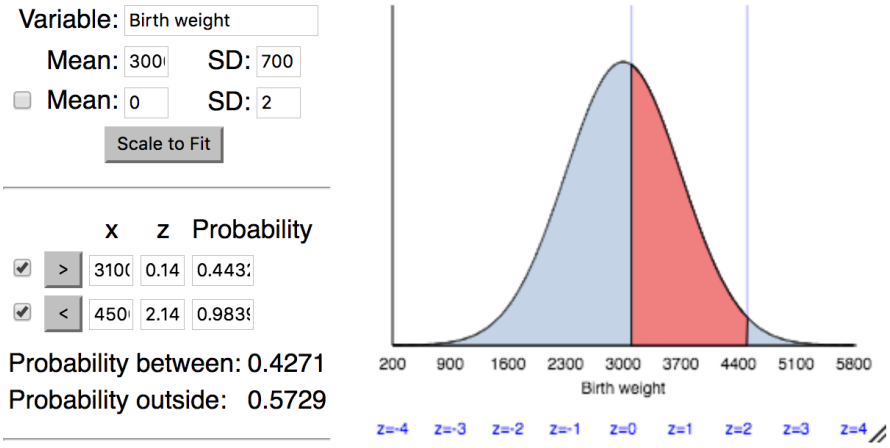
1. What proportion of babies have birth weights below 2000 grams?

   Answer: 0.0766. Check the box next to the first row of probability calculations and select "<". Input 2000 for X.



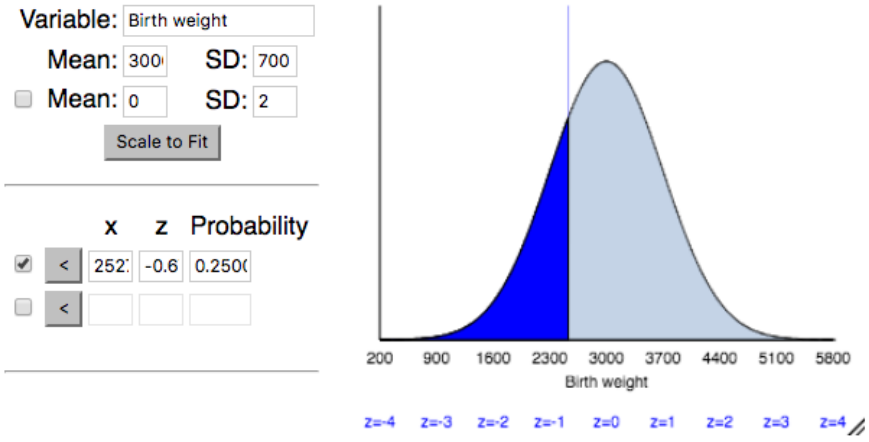
2. What is the probability that a randomly selected baby will have a birth weight between 3100 and 4500 grams?

   Answer: 0.4271. Check both boxes of probability calculations. Input 3100 for X in the first row and 4500 for X in the second row.



3. What is the 25th percentile of birth weights? That is, at what birth weight do 25% of babies fall below and 75% of babies fall above?

   Answer: 2527.856 grams. Check the box next to the first row of probability calculations and select "<". Input 0.25 for Probability.



2 Sampling Distributions of Statistics

A **sampling distribution** is a *probability distribution of a sample statistic*. For large sample sizes n, the sampling distribution of a sample proportion $\hat{p}$ has an approximate normal distribution with a

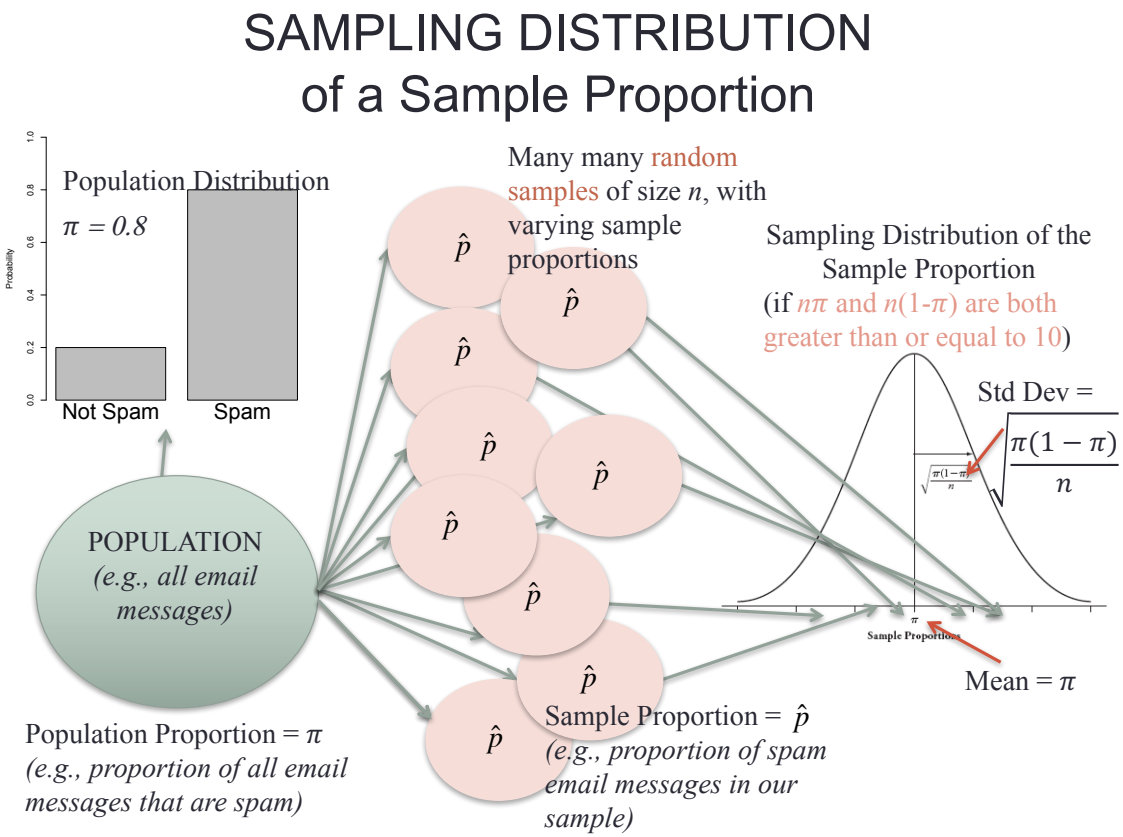
mean equal to the population proportion π and standard deviation

$$SD(\hat{p}) = \sqrt{\frac{\pi(1-\pi)}{n}}.$$

The sample size n is large enough for this approximation to be valid if both $n\pi$ and $n(1-\pi)$ are at least 10 (or if the number of successes and the number of failures in your sample are both at least 10). Example: Suppose in the population of all email messages, 80% are spam, and 20% are not spam. We plan on taking a simple random sample of n emails and measuring the sample proportion of emails that are spam. Then for large n, the sampling distribution of sample proportions is approximately normal with mean $E(\hat{p}) = 0.80$ and standard deviation $SD(\hat{p}) = \sqrt{\frac{0.8(1-0.8)}{n}}$.

For instance, if $n = 50$, then $n\pi = 50(0.8) = 40$ and $n(1-\pi) = 50(0.2) = 10$ are both at least 10, so our sample size is large enough and the sampling distribution of sample proportions will be approximately normal with mean 0.80 and standard deviation $\sqrt{\frac{0.8(1-0.8)}{50}} = 0.0566$.

The figure below visualizes this process: We start with the population of all emails, for which 80% are spam and 20% are not spam. (Note that each observational unit is one email. The variable is whether an email is spam or not spam, which is categorical.) Imagine selecting a sample of n emails and calculate the sample proportion of emails that are spam, $\hat{p}$. We repeat this process many times, each time giving us a new value for $\hat{p}$ – these many samples are represented by pink circles. Lastly, we plot all these values of $\hat{p}$ on a dotplot, which will have the shape of the plot on the right as long as we have a large enough sample size.



We use this property of sample proportions when we assess statistical significance using a standardized statistic. In Exploration 1.2, you learned how to calculate a standardized sample proportion as:

$$\frac{\text{observed statistic} - \text{mean of null distribution}}{\text{SD of null distribution}}.$$

The standard deviation of the null distribution can be calculated through simulation, or through the formula above. If the center of the null distribution is π, this standardization matches

$$z = \frac{\hat{p} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}.$$

3 References

- Robison-Cox, J. (2016) *Stat 216 Course Pack Fall 2016: Activities and Notes*. License: Creative Commons BY-SA 3.0.
- Utts, J. M., & Heckard, R. F. (2015). *Mind on Statistics*, 5th ed., Chapters 7 and 8. Stamford, CT: Cengage Learning.