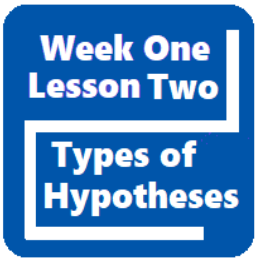
Statistical Inference and Hypothesis Testing in Data Science Applications

DTSA 5003 offered on Coursera

by the University of Colorado, Boulder

Instructor: J.N. Corcoran

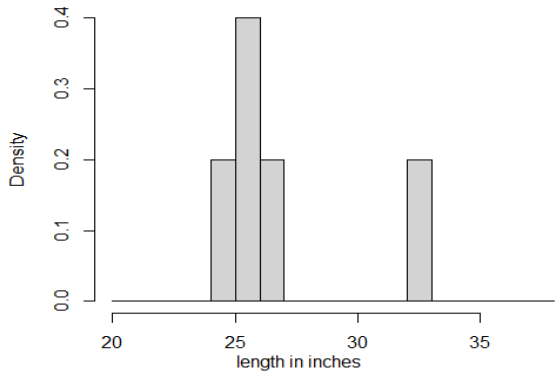


The Shape of Data: Sample versus Population versus Theoretical Distribution

Suppose that we are measuring the length, in inches, of sockeye salmon at a hatchery in the Pacific Northwest. Let's select 5 salmon at random and measure them. Suppose that the measurements, in inches, are as follows.

$$24.05 \quad 32.12 \quad 26.02 \quad 25.85 \quad 25.45$$

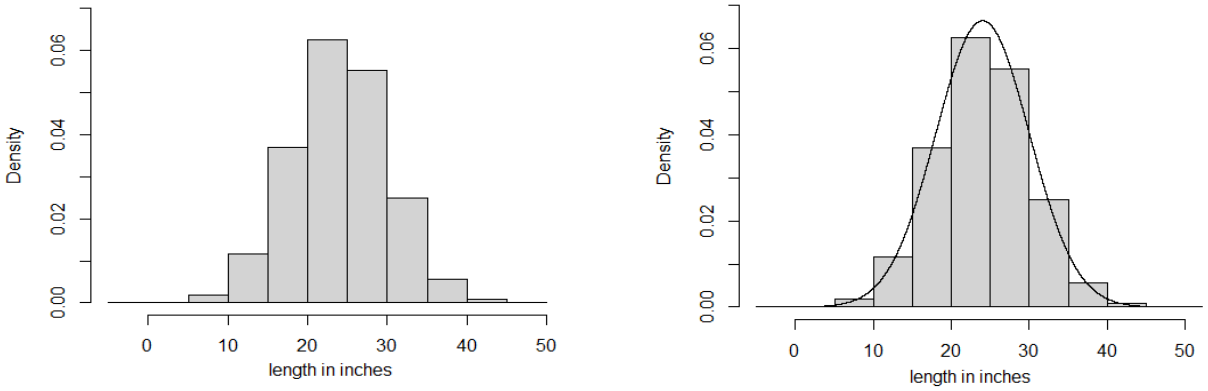
Let's make a histogram.



Here, the x-axis is cut up into 1 inch segments or "bins". Looking at the relative height of the bars, we see that one salmon from our sample was between 24 and 25 inches, two were between 25 and 26 inches, one was between 26 and 27 inches, and there was one really large fish coming in between 32 and 33 inches! You might notice that the y-axis is labeled "Density". It is possible to make a histogram where we put count data on the y-axis. The heights of the bars would them be 1, 2, 1, and 1, respectively. Another possibility is for us to put *relative frequency* on the y-axis. Relative frequencies are simply proportions. The heights of the bars would be $1/5$, $2/5$, $1/5$, and $1/5$, respectively. A third option is to use *density* on the y-axis. This requires that the height of the bars be such that the area of each bar is equal to the proportion of measurements that were observed in each bin. Since the width of the

bars happens to be 1 for this histogram, density is the same as relative frequency here. If the first bin was chosen to be, for example, from 23 to 25 (a base width of 2 inches), the height of the associated bar would stay the same if we were using relative frequency, whereas it would come down to half the height when using density. Density is about the visual impact of the area of the bars.

With a much larger sample of $10,000$ salmon, we get a histogram that is really shaping up nicely and probably looks much more like it would look if we had measurements for the entire population.



The data appear to be following a normal distribution. In the second histogram above, we have overlaid the normal probability density function (pdf)

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

using values for μ and σ^2 that have been estimated from the larger data set. The pdf is a curve under which area represents probability. Overlaying the pdf like this only makes sense if we use density, where areas of the bars represent proportions, on the y-axis for our histogram.

The normal distribution pdf is defined for all x between $-\infty$ and ∞. Clearly fish lengths are not going to take on negative values, but with the area under the tails of this normal pdf being so negligible, it seems to still be a decent model to use to analyze this data set.

In summary, we have a sample that gives us a rough (or not, depending on the sample size) picture of what is going on with the population that we never really get to see, and we have a distribution that we believe models the population.

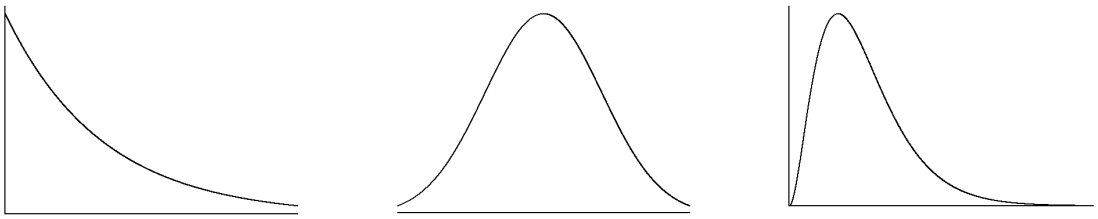
Random Samples

Suppose that we don't have our fish measurements yet but are planning to go out, "randomly sample" n fish, and measure them. The results will be numbers but, before the sample is taken, they are unknown and subject to the randomness of our selection process.

Using the convention, in statistics, of capital letters for random variables, we will denote this random sample as

$$X_1, X_2, \ldots, X_n.$$

We likely don't know the exact distribution that models the population but may have a basic idea of its shape. Here are some common shape examples.



These are probability density functions (pdfs) for an exponential, normal, and gamma distribution, respectively. Based on histogram for the larger sample shown in the previous section, it seems that we may be sampling salmon lengths from a normal distribution (population). Let's assume that this is the case for now. Note that we have still not completely specified the distribution because it depends on two parameters: a mean μ and a variance σ^2.

Definition

The random variables $X_1, X_2, \ldots, X_n$ are a **random sample** from a distribution if they are

- independent, and
- "identically distributed".

"Identically distributed" means that they all come from the the same exact distribution.

The "exact same distribution" means not only that all of the X_i come from, for example, a normal distribution or all from an exponential distribution. It also implies that the parameters are identical. For example, X_1 and X_2 might both have normal distributions but if $X_1 \sim N(0,1)$ and $X_2 \sim N(0,2)$, the distributions are not identical. The phrase "independent and identically distributed" is abbreviated by "iid". We write

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$$

if we are talking about a random sample from the normal distribution with mean μ and variance σ^2.

Hypotheses

Let's return to our sockeye salmon measurement data. Suppose that we are trying to figure out whether the true mean length in inches is less than or equal to 28 inches or greater that 28 inches.

It seems natural to want to look at the sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

For our 5 point fish example, we have an observed sample mean of $\overline{x} = 26.698$. (Note that we have made "x bar" lowercase since it has been observed and is no longer random.) Is this enough to say that the true mean μ is less than 28? What if your friend also took a random sample of size 5 and they observed a sample mean of $\overline{x} = 29.3$? Things are still unclear at this point!

The way that we have set things up, there are two "hypotheses" up for consideration– we either have $\mu \leq 28$ or $\mu > 28$. We are going to assume that one of these is true and reject the idea if the data gives strong evidence to the contrary. These hypotheses have names.

Definition

The **null hypothesis** is a hypothesis that is assumed to be true. We denote it with an H_0. An example of specifying a null hypothesis is

$$H_0 : \mu \leq 28.$$

Definition

The **alternative hypothesis** is a hypothesis that we are looking for evidence for or "out to show". We denote it with an H_1. An example of specifying an alternative hypothesis is

$$H_1 : \mu > 28.$$

(Note: Some people use the notation H_a here.)

What does "out to show" mean? It all depends on our perspective. Suppose that we run the salmon hatchery and claim to produce salmon that are, on average, greater than 28 inches in length. A local salmon conservation group, protesting outside of our facility, claims that salmon produced in captivity and released into the wild are less robust. They claim that we are weakening the salmon population as measured by several criteria. In particular, they believe that we produce salmon that are, on average, shorter than 28 inches in length. We randomly sample some hatchery salmon for measurement and allow the conservation group

to do the same. Since we are trying to show that the average length for the entire popultaion is greater than 28 inches, we would test the hypotheses

$$H_0 : \mu \leq 28 \qquad \text{versus} \qquad H_1 : \mu > 28.$$

On the other hand, if the conservation group is trying to make the point that the average length is actually below 28 inches, they would consider the hypotheses

$$H_0 : \mu \geq 28 \qquad \text{versus} \qquad H_1 : \mu < 28.$$

Simple and Composite Hypotheses

Definition

A **simple hypothesis** is one that completely specifies the distribution.

A hypothesis that is not simple is called a **composite hypothesis.**

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the normal distribution with mean μ and variance 1. The hypothesis

$$H_0 : \mu = 28$$

is a simple hypothesis. If it is true, we know that the random sample comes from the $N(28, 1)$ distribution.

The hypothesis

$$H_0 : \mu \leq 28$$

is a composite hypothesis, If it is true, the random sample might come from the $N(28, 1)$ distribution, or the $N(25.73, 1)$ distribution, or any of an uncountably infinite number of normal distributions. Knowing that H_0 is true does not tell us exactly which one it is.

If we are testing

$$H_0 : \mu = 28 \qquad \text{versus} \qquad H_1 : \mu \neq 28,$$

the null hypothesis is simple and the alternate hypothesis is composite.

It is a common misconception to say that a hypothesis "with an equals sign" is a simple hypothesis. This is not necessarily true. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the normal distribution with mean μ and variance σ^2. If both μ and σ^2 are unknown, the hypothesis

$$H_0 : \mu = 28$$

is no longer simple. If H_0 is true, we know that the sample comes from a normal distribution with mean 28. However, the variance is unknown. The sample may be from a $N(28, 1)$

distribution or a $N(28, 2.11)$ distribution or any of an uncountable infinite number of normal distributions all having mean 28.