Density Curves

We’ve defined probabilities for discrete variables (such as coin tossing). Probabilities for continuous or measurement variables also are evaluated using relative frequencies. When sampling (i.e. SRS), probabilities of events are evaluated relative to a population histogram.

Example: I “manufactured” 1 million observations like those in Example 3.27 page 93 of SW. That’s certainly enough values to be considered a whole population. I then made a histogram with 20 classes of all 1 million values, then a second with 100 classes. The second one looks like a smooth curve, so we’ll summarize a set of population values more succinctly with a smooth curve (the third plot below).

Now suppose in this population that 42% of all BG levels are in the interval 100-150, and 2% are greater than 200. If we randomly select 1 individual from the population, and define the random variable

\[ Y = \text{BG level for selected person} \]

then

\[ \Pr\{100 \leq Y \leq 150\} = .42 \quad \text{and} \quad \Pr\{Y > 200\} = .02 \]

The vertical scale on the population histograms gets reset so that the area of a bar over an interval is the population proportion in that interval. With this rescaling, probabilities are obtained by summing areas of bars. A conceptually similar process, but one which is easier to visualize, is to first approximate the rescaled histogram by a smooth curve, called the density curve from which areas (same as probabilities) are evaluated. The third curve above then is called the density.

Interpretation of density

For any two numbers a and b,

the area under the curve between a and b \[ \quad = \text{population proportion of values between a and b} \]

\[ \quad = \text{probability of randomly selecting someone with a value between a and b} \]
Area = $\Pr\{a \leq Y \leq b\}$

$1 = \text{Total Shaded Area}$

$.42 = \Pr\{100 \leq Y \leq 150\}$ for BG problem

Example: SW example 3.30 p. 95
Binomial Distribution

This is a special discrete probability distribution. We need to understand it to understand statistical techniques such as logistic regression.

Independent Trials Model for Binomial Experiment

- A series of $n$ trials
- Each trial results in 1 of 2 outcomes, called Success and Failure (generic names)
- The probability of success, called $p$, is the same for each trial
- The outcomes of the trials are assumed to be independent

Example: Toss a coin 5 times
Outcome for each trial (toss)  
H (call it success) 
T (call it failure)

Suppose coin is fair, so $p = \Pr\{H\} = .5$

Example: SW example 3.42 p. 103

The binomial distribution specifies a formula for evaluating the probability distribution associated with the number of successes in a binomial experiment. In the previous problem, $n = 2$ and, if we define $Y =$ No. of Albino Children, we can directly evaluate this distribution, but it is much harder to do with trees if $n$ is large (too many branches).

Definitions

- For any positive integer $x$ define $x! = 1\cdot 2\cdot 3\cdots (x-1)\cdot x$. This is called x-factorial. What is 2!? 3!? 5!?
- Also define $0! = 1$
- The binomial coefficient ${n \choose j}$ is defined as ${n \choose j} = \frac{n!}{j!(n-j)!}$ (note: we will always have $n \geq j$).

Table 2 of SW p. 674 gives binomial coefficients. Notice how fast they get large.

The Binomial Distribution Formula

For a binomial experiment, the probability of exactly $j$ successes in $n$ trials (so $n-j$ failures) is $\Pr\{j \ Successes\} = {n \choose j} p^j (1-p)^{n-j}$, $j = 0,1,2,...,n$

Example: 3.45 p. 106 of SW

Example: 3.47 p. 108 of SW
Mean and Standard Deviation of the Binomial

If Y = No. Successes in a binomial experiment with n trials and probability of success = p, then the mean (or expected value) is \( \mu = np \) and the standard deviation is \( \sigma = \sqrt{np(1-p)} \). If we were to repeat the experiment a huge number of times, each time recording the number of successes, we would have a huge collection of integers between 0 and n. The preceding formulas give the mean and standard deviation we would calculate from that huge collection.

Example: 3.47 p.108 revisited. Here n = 6 and p = .85, so expected no. of Rh+ is \( 6(.85) = 5.10 \) and the standard deviation is \( \sqrt{6(.85)(.15)} = .87 \). Note that the expected value is not necessarily something that is observed!

The Normal Distribution

SW Chapter 4, Sections 1-4

The normal or bell-shaped distribution plays an important role as an approximation to certain discrete and continuous distributions, but is also useful as an approximate density curve for certain populations.

Example: SW example 4.1 p. 119.

Example: SW example 4.2 p. 120.

Normal Curves

The preceding examples illustrate two important properties of normal distribution or curves

- There are many distinct normal curves
- The normal curve that approximates a population histogram is indexed by the mean and standard deviation of the population, labeled \( \mu \) and \( \sigma \), respectively.

Note that normal curves are centered at \( \mu \) and the spread of the distribution is controlled by the size of \( \sigma \) (i.e. larger \( \sigma \) implies less probability near \( \mu \)).
The functional form for a normal density curve is

\[ f(y) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{y - \mu}{\sigma} \right)^2} \]

We will see presently how to evaluate probabilities for normal distributions. The following diagram does show some of the most important areas (approximately).

### Standard Normal Distribution

The standard normal curve corresponds to a normal distribution with mean 0 and standard deviation 1. The letter Z (for Z-score) is used to identify a standard normal variable.

All normal distributions can be related to the standard normal through the so-called z-score transformation, which allows one to calculate probabilities for arbitrary normal distributions as areas under the standard normal curve.
Areas under the standard normal curve

SW Table 3 p. 675-6, but more conveniently on the inside front cover, gives **lower-tailed areas** under the Z-curve of the form

![Diagram of normal distribution with shaded area under curve for all values ≤ z]

In the table z ranges from -3.49 to 3.49, although it is positive above.

Example: If \( z = 0.47 \), then

This picture also tells you the area to the right of 0.47. How?

It also tells you the area to the right of -0.47. How?

It also tells you the area to the left of -0.47. How?

We exploit symmetry extensively to use these tables, i.e. lower tailed areas can be used to compute **upper tailed areas**. We also use the fact that the total area under the curve is 1.

Example:

Area under curve right of -1.47 can be computed two ways. It is the same as the area to the left of +1.47, or .9292. It also is the 1-area left of -1.47, or 1-.0708, which also is .9292. The first way probably is most direct; the second can be easiest sometimes.
We find central areas by subtracting:

\[ \text{Desired area} = .8907 - .158 = .7320 \]

**Comments**

1. Always draw a picture
2. Recognize symmetry about 0 (area to left of 0 = area to right of zero = .5)
3. Areas are probabilities. If \( Z \) is a standard normal variable then we computed in the three examples

   (a) \( \Pr\{Z \leq .47\} = .6808 \)
   (b) \( \Pr\{Z \geq -1.47\} = .9292 \)
   (c) \( \Pr\{-1 \leq Z \leq 1.23\} = .7320 \)

Example: compute \( \Pr\{Z \geq 2.12\} \). Draw a picture!

Using the standard normal table we can show, for a \( Z \), that the following areas are true. These imply the more general result on page 5. We will return to this shortly.
Areas for General Normal Curves

Suppose for a certain population of herring, the lengths of individual fish follow a normal distribution with mean 54 mm and standard deviation 4.5 mm (population mean and std. dev.).

What percent of fish has length less than or equal to 60 mm?

Let $Y =$ length of randomly selected fish. Then the percent of interest is just $\Pr\{Y \leq 60\}$. The picture is

To get the area of interest we use the $Z$-score transform $Z = \frac{Y - \mu}{\sigma}$ to create a standard normal variable from $Y$. We then convert the limits of the area from the $Y$-scale to the $Z$-scale in the same manner. That is for $y = 60$, $z = \frac{60 - \mu}{\sigma} = \frac{60 - 54}{4.5} = 1.33$. We then compute the corresponding area under the $Z$-curve using Table 3,
Symbolically,

\[ \Pr\{Z \leq 1.33\} = \Pr\{Y \leq 60\} = \text{proportion of fish with length} \leq 60 \text{ mm} = .9082 = 90.82\%. \]

What percentage of fish is more than 45 mm long?

Follow steps 1-2-3

\[ 1.33 \]

\[ 0 \]

\[ 45 \]

\[ 54 \]

\[ -2 \]

\[ 0 \]

\[ 2 \]

\[ -2 \]

\[ .9772 \]

\[ 1 \]

\[ 2 \]

\[ 3 \]

\[ 1.0228 \]

\[ .9772 \]

\[ \text{i.e. } \Pr\{Z \geq -2\} = \Pr\{Y \geq 45\} = \text{proportion of fish} \geq 45 \text{ mm} = .9772 = 97.72\%. \]

Let us work out, together, the following problem: What proportion of fish is between 56 and 62 mm long?

### 1–2–3 SD rule

The Z-score tells us how many standard deviation (SD) units a value is from the mean. For example, \( y = \mu + 2\sigma \) is a value 2 standard deviations above \( \mu \), so \( z = \frac{\mu + 2\sigma - \mu}{\sigma} = \frac{2\sigma}{\sigma} = 2 \).

Similarly \( y = \mu - \sigma \) is a value 1 standard deviation below \( \mu \), so \( z = \frac{\mu - \sigma - \mu}{\sigma} = \frac{-\sigma}{\sigma} = -1 \). The 1–2–3 SD rule says that for a normal population,
• 68% of values are within 1 std dev of $\mu$ (i.e. in $(\mu - \sigma, \mu + \sigma)$)
• 95% 2 (i.e. in $(\mu - 2\sigma, \mu + 2\sigma)$)
• 99.7% 3 (i.e. in $(\mu - 3\sigma, \mu + 3\sigma)$)

This follows from a direct transformation of interval endpoints to $z$-scores.

In essence, this result implies

1. That the distance from the center is best measured in terms of multiple of standard deviations (i.e. $Z$-scores).
2. The standard deviation determines how “extreme” or unusual a value is.

Example: Fish lengths. Here $\mu = 54$ mm and $\sigma = 4.5$ mm
- 68% of fish have lengths between $54 - 4.5$ and $54 + 4.5$, or 49.5 and 58.8
- 95% $54 - 2(4.5)$ and $54 + 2(4.5)$, or 45 and 63
- 99.7% $54 - 3(4.5)$ and $54 + 3(4.5)$, or 40.5 and 67.5

Example: SW exercise 4.3 (a) – (c) p. 131.
**Percentiles of Normal Distributions**

These are “inverse problems” where you are given an area and need to find a value that will produce it (instead of the “direct” problem of finding an area like we have been doing).

The $p^{th}$ percentile for a probability distribution is the value such that there is probability $p$ of being less than or equal to that value, and probability $1-p$ of being greater. The picture below identifies the placement of the $p^{th}$ percentile for a normal distribution.

To compute this percentile, we first compute the $p^{th}$ percentile for a standard normal

This can be obtained by an inverse process from what we’re used to. Given the standard normal percentile we solve for the $y$-value that gave rise to it:

$$z = \frac{y^* - \mu}{\sigma} \Rightarrow y^* = \mu + z\sigma$$

Examples (fish lengths, revisited):
- What is the $40^{th}$ percentile of the fish length distribution?
- What length must a fish be so that only 5% have longer lengths?

Example: SW ex 4.16 p. 133
**Sampling Distributions**

Read SW Chapter 5, Sections 1-4

Suppose that I wish to know what the *mean* is for a specific population (target or study population). The following diagram summarizes the sampling from the population.

![Sampling Diagram](image)

Population of interest
Mean μ (unknown)
The pop. is what we’re interested in but cannot see completely.

Random sample
SRS of n

**Important points**

- The sample mean \( \overline{Y} \) estimates \( \mu \), but there is error associated with the estimate (we would need to access the whole population to exactly calculate \( \mu \)).
- Different samples give different \( \overline{Y} \)’s, so the size of the error depends upon the sample. Of course in practice we don’t know if we got a “good” sample or a “bad” one (i.e. we don’t know if \( \overline{Y} \) is close to \( \mu \) or not).

The value of \( \overline{Y} \) cannot be predicted prior to the sample being selected, so before collecting data we think of \( \overline{Y} \) as being a *random variable*. On a conceptual level we can envision listing *all possible* samples of size \( n \) and the \( \overline{Y} \) that results. The collection of sample means that is obtained can be collected and plotted as a histogram, leading to what is commonly called the *sampling distribution* of the sample mean \( \overline{Y} \). The sampling distribution of \( \overline{Y} \) is, in essence, a *probability distribution* of \( \overline{Y} \) that specifies the probability of observing specific ranges of values for \( \overline{Y} \) when we sample the population.

The sampling distribution is real but is mostly a conceptual construct. We can list the sampling distribution only if we know the population values. If we knew the population distribution we would not need to sample it! Nonetheless, the idea of the sampling distribution is crucial to understanding the accuracy of \( \overline{Y} \).
The following properties of the sampling distribution of $\bar{Y}$ can be shown *mathematically*:

If we have a SRS from a population with mean $\mu$ and standard deviation $\sigma$, then

1. The average or mean in the sampling distribution of $\bar{Y}$ is the population mean $\mu$. In symbols, $\mu_{\bar{Y}} = \mu$ (the mean of the mean is the mean?)

2. The standard deviation of the sampling distribution of $\bar{Y}$ is the population standard deviation divided by the square root of the sample size $n$. In symbols, $\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$.

3. **Shape of sampling distribution**
   
   a) If the population distribution is normal, then the sampling distribution of $\bar{Y}$ is normal, regardless of $n$.
   
   b) **Central Limit Theorem**: If $n$ is large, the sampling distribution of $\bar{Y}$ is normal, even if the population distribution is not normal.

**Important points**

1. The “typical error” in $\bar{Y}$ as an estimate of $\mu$ is 0.

2. The accuracy of $\bar{Y}$ increases with sample size (i.e. smaller variability in sampling distribution).
3. For a given sample size, $\bar{Y}$ will be less accurate for populations with larger standard deviation $\sigma$.

4. Knowing that the shape of the sampling distribution is normal, or approximately normal, allows one to evaluate probabilities for $\bar{Y}$.

These ideas are fairly subtle but important. Let us go through a number of examples in the text.

Examples: SW example 5.8 p. 158 5.11 p. 162 5.14 p.168
SW example 5.9 p. 159 5.12 p. 163
SW example 5.10 p. 161 5.13 p.167

**Dichotomous observations**

Dichotomous = binary (successes or failures)
Suppose we select a SRS of size $n$ from a population and count the number of “successes” or individuals in the sample with a given characteristic. We know this has a Binomial distribution. As an alternative way to summarize the data, we might consider the sample proportion $\hat{p}$, where $\hat{p} = \# \text{successes} / n$. If the population proportion of successes is $p$, we can show that the sampling distribution of $\hat{p}$ has

1. Mean = $p$
2. Standard deviation = $\sqrt{\frac{p(1-p)}{n}}$
3. Shape that is approximately normal for large $n$

These results can be used to approximate binomial probabilities using a normal distribution.

Example: SW example 5.16 p. 171.

I do not want to overemphasize probability calculations based on sampling distribution. The important place for sampling distributions is where they directly lead to procedures for statistical inference.
Confidence Intervals
SW Chapter 6
Course notes p 20-26

This chapter introduces a standard statistical tool, the confidence interval, which will be used to estimate unknown population means $\mu$ and proportions $p$.

**Standard Errors of the Mean**

Given a SRS of size $n$ from a population with unknown mean $\mu$ and standard deviation $\sigma$, our best guess for $\mu$ is the sample mean $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$. The accuracy of $\bar{Y}$ as an estimate of $\mu$ is dictated by the size of the standard deviation of the sampling distribution of $\bar{Y}$, $\sigma_\bar{Y} = \frac{\sigma}{\sqrt{n}}$. In practice, this quantity is unknown (we don’t know $\sigma$), but can be estimated by using the sample standard deviation $S$ as an estimate of $\sigma$. This leads to the so-called *standard error of $\bar{Y}$*, $SE_\bar{Y} = \frac{S}{\sqrt{n}}$ as an estimate of $\sigma_\bar{Y}$.

Example: SW ex 6.3 p 181
SW ex 6.5 p.183-4

**The Student’s t-distribution**

If the population standard deviation $\sigma$ were known, then a Z-score transformation of the sampling distribution of $\bar{Y}$, i.e. $Z = \frac{\bar{Y} - \mu}{\sigma_\bar{Y}} = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$ could be used to generate confidence intervals for $\mu$, and this idea is illustrated by SW on p. 186. Because $\sigma$ is unknown in practice, we need an alternative approach and that is to standardize $\bar{Y}$ by using $SE_\bar{Y}$ instead of $\sigma_\bar{Y}$, computing a t-score instead of a z-score, $t = \frac{\bar{Y} - \mu}{SE_\bar{Y}} = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$. If the population we sampled from is normal, then $Z$ is standard normal and $t$ has the so-called t-distribution.

The density curve for a t-distribution looks like a bell-shaped curve, but has more probability in the tails than the Z-distribution. Furthermore, there are many t-distribution, but the family or collection of t-distributions is indexed by one parameter called the degrees of freedom (df). For our problem the df = n-1. As df increases the t-density curve more closely resembles a standard normal curve. For df = $\infty$ (infinity) the t-curve is the standard normal curve. See Figure 6.7 p. 187 of SW.
For our applications we will need upper tailed critical values from the t-distribution of the form

Table 4 (inside back cover) gives upper tail critical values.

Example: for df = 10 find \( t_{0.05}, t_{0.025}, t_{0.01} \) Repeat for df = 3

Question: For a given upper tail area, how do critical values change as df increases? What do you get when df = \( \infty \)?