The Normal Distribution

- The Normal Distribution is the most important distribution of continuous random variables.
- The normal density curve is the famous symmetric, bell-shaped curve.
- The central limit theorem is the reason that the normal curve is so important. Essentially, many statistics that we calculate from large random samples will have approximate normal distributions (or distributions derived from normal distributions), even if the distributions of the underlying variables are not normally distributed.
- This fact is the basis of most of the methods of statistical inference we will study in the last half of the course.
- Chapter 4 introduces the normal distribution as a probability distribution.
- Chapter 5 culminates in the central limit theorem, the primary theoretical justification for most of the methods of statistical inference in the remainder of the textbook.
The Normal Density

Normal curves have the following bell-shaped, symmetric density.

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

**Parameters:** The parameters of a normal curve are the mean \( \mu \) and the standard deviation \( \sigma \).

Here is an example of a normal curve with \( \mu = 100 \) and \( \sigma = 20 \).

### Standardization

- All normal curves have the same essential shape.
- Every normal curve can be drawn in exactly the same manner, just by changing labels on the axis.
- The **standard normal curve** is the normal curve with mean \( \mu = 0 \) and standard deviation \( \sigma = 1 \).
- The **standardization formula** is
  \[ Z = \frac{Y - \mu}{\sigma} \]
- Every problem that asks for an area under a normal curve is solved by first finding an equivalent problem for the standard normal curve.
- The justification for this comes from calculus. An area under a normal curve is a definite integral. The integral is simplified by using the standardization formula.
The 68–95–99.7 Rule

For every normal curve, 68% of the area is within one SD of the mean, 95% of the area is within two SDs of the mean, and 99.7% of the area is within three SDs of the mean.

---

The Standard Normal Table

- The standard normal table lists the area to the left of $z$ under the standard normal curve for each value from $-3.49$ to $3.49$ by $0.01$ increments.
- The normal table is on the inside front cover of your textbook.
- Numbers in the margins represent $z$.
- Numbers in the middle of the table are areas to the left of $z$.

R can do this for general values of $z$, and R can do the standardization for you.
Example Calculations

Suppose that egg shell thickness is normally distributed with a mean of 0.381 mm and a standard deviation of 0.031 mm. Here are a large number of example calculations.

Example Area Calculation

Area to the left. Find the proportion of eggs with shell thickness less than 0.34 mm.

> gnorm(0.381, 0.031, b = 0.34)

Normal Distribution
mu = 0.381 , sigma = 0.031
Example Area Calculation

**Area to the right.** Find the proportion of eggs with shell thickness more than 0.36 mm.

> gnorm(0.381, 0.031, a = 0.36)

![Probability density graph](image1)

\[ P(X > 0.36) = 0.7509 \]

Example Area Calculation

**Area between two values.** Find the proportion of eggs with shell thickness between 0.34 and 0.36 mm.

> gnorm(0.381, 0.031, a = 0.34, b = 0.36)

![Probability density graph](image2)

\[ P(0.34 < X < 0.36) = 0.1561 \]
\[ P(X < 0.34) = 0.093 \]
Example Area Calculation

**Area outside two values.** Find the proportion of eggs with shell thickness smaller than 0.32 mm or greater than 0.40 mm.

\[ > \text{gnorm}(0.381, 0.031, a = 0.32, b = 0.4) \]

![Normal Distribution](image)

\[ P(0.32 < X < 0.4) = 0.7055 \]
\[ P(X < 0.32) = 0.0245 \]
\[ P(X > 0.4) = 0.27 \]

Example Area Calculation

**Central area.** Find the proportion of eggs with shell thickness within 0.05 mm of the mean.

\[ > \text{gnorm}(0.381, 0.031, a = 0.381 - 0.05, b = 0.381 + 0.05) \]

![Normal Distribution](image)

\[ P(0.331 < X < 0.431) = 0.8932 \]
\[ P(X < 0.331) = 0.0534 \]
\[ P(X > 0.431) = 0.0534 \]
Example Area Calculation

Two-tail area. Find the proportion of eggs with shell thickness more than 0.07 mm from the mean.

> gnorm(0.381, 0.031, a = 0.381 - 0.07, b = 0.381 + 0.07)

Quantiles

Quantile calculations ask you to use the normal table backwards. You know the area but need to find the point or points on the horizontal axis.
Example Quantile Calculations

Percentile. What is the 90th percentile of the egg shell thickness distribution?

\[ \text{> gnorm(0.381, 0.031, quantile = 0.9)} \]

![Normal Distribution](image)

\( P( X < 0.4207 ) = 0.9 \)

Example Quantile Calculations

Upper cut-off point. What value cuts off the top 15% of egg shell thicknesses?

\[ \text{> gnorm(0.381, 0.031, quantile = 0.85)} \]

![Normal Distribution](image)

\( P( X < 0.4131 ) = 0.85 \)
Example Quantile Calculations

Central cut-off points. The middle 75% egg shells have thicknesses between which two values?

\[
> \text{gnorm(0.381, 0.031, quantile = 0.875)}
\]

Using R

Previous commands show how to use R and the new function \texttt{gnorm} to graph normal distributions where the graphs include some probability calculations. You can also make the same calculations without graphing within R using the functions \texttt{pnorm} and \texttt{qnorm}. The following lists the commands for all of the previous computations.

Area to the left. Find the proportion of eggs with shell thickness less than 0.34 mm.

\[
> \text{pnorm(0.34, 0.381, 0.031)}
\]

\texttt{[1] 0.09298744}

Area to the right. Find the proportion of eggs with shell thickness more than 0.36 mm.

\[
> 1 - \text{pnorm(0.36, 0.381, 0.031)}
\]

\texttt{[1] 0.75093}
Using R

**Area between two values.** Find the proportion of eggs with shell thickness between 0.34 and 0.36 mm.

```r
> pnorm(0.36, 0.381, 0.031) - pnorm(0.34, 0.381, 0.031)
[1] 0.1560825
```

**Area outside two values.** Find the proportion of eggs with shell thickness smaller than 0.32 mm or greater than 0.40 mm.

```r
> pnorm(0.32, 0.381, 0.031) + 1 - pnorm(0.4, 0.381, 0.031)
[1] 0.2945190
```

**Central area.** Find the proportion of eggs with shell thickness within 0.05 mm of the mean.

```r
> 1 - 2 * pnorm(0.381 - 0.05, 0.381, 0.031)
[1] 0.8932345
```

**Two-tail area.** Find the proportion of eggs with shell thickness more than 0.07 mm from the mean.

```r
> 2 * pnorm(0.381 - 0.07, 0.381, 0.031)
[1] 0.02394164
```

**Percentile.** What is the 90th percentile of the egg shell thickness distribution?

```r
> qnorm(0.9, 0.381, 0.031)
[1] 0.4207281
```

**Upper cut-off point.** What value cuts off the top 15% of egg shell thicknesses?

```r
> qnorm(0.85, 0.381, 0.031)
[1] 0.4131294
```

**Central cut-off points.** The middle 75% egg shells have thicknesses between which two values?

```r
> qnorm(c(0.25/2, 1 - 0.25/2), 0.381, 0.031)
[1] 0.3453392 0.4166608
```

Statistics 371, Fall 2004 10
Normal Probability Plots

• A standard question begins, “Assuming that variable Y has a normal distribution,…” But how do we know the distribution is approximately normal?
• Sometimes we can rely on the central limit theorem.
• If given data, there are tests for normality, but there are reasons not to do these.
• It is generally better to make a plot that sheds light on the question, “Is the data so far from normality as to bias a method that assumes normality?”
• There is no easy answer to the question, but a normal probability plot is much more informative than the result of a test.

What is a Normal Probability Plot?

• A normal probability plot is a plot of the sorted sample data versus something close to the expected z-score for the corresponding rank of a random normal sample of the same size.
• For example, the expected z-score of the minimum from a random sample of size 10 from a normal population is about –1.54.
• If the plotted points are close to a straight line, there is evidence that the distribution is close to normal.
• If the plotted points are far from a straight line, there is evidence of non-normality.
Normal Example

All three plots are normal with $n = 50$.

Normal Example

All three plots are normal with $n = 500$. 
Non-normal Example

The first plot is not normal, the second two are with \( n = 50 \).

The Continuity Correction

- The normal distribution can also be used to compute approximate probabilities of discrete distributions.
- These calculations are more accurate when the boundaries of normal calculations match midpoints between possible values of the discrete distribution.
- Approximations to the binomial distribution use a normal curve with \( \mu = np \) and \( \sigma = \sqrt{np(1-p)} \).
- When approximating a binomial probability, the approximation is usually pretty good when \( \sqrt{np(1-p)} > 3 \).

Here is an example with the binomial distribution with \( n = 15 \) and \( p = 0.4 \) where we want to compute \( \Pr\{3 \leq Y \leq 7\} \).

Algebraically,

\[
\Pr\{3 \leq Y \leq 7\} = \Pr\{2 < Y < 8\}
\]
The Continuity Correction

Which endpoints should we use? We get a better answer if we use 2.5 and 7.5.

The exact calculation and accurate approximation are

\[
\text{Exact calculation: } pbinom(7, 15, 0.4) - pbinom(2, 15, 0.4)
\]

\[
\text{Approximation: } \text{pnorm}(7.5, \mu, \sigma) - \text{pnorm}(2.5, \mu, \sigma)
\]

This is more accurate than the following choices.

\[
\text{Choice 1: } \text{pnorm}(7, \mu, \sigma) - \text{pnorm}(3, \mu, \sigma)
\]

\[
\text{Choice 2: } \text{pnorm}(8, \mu, \sigma) - \text{pnorm}(2, \mu, \sigma)
\]

Continuity Correction

Here is a picture that shows why the continuity correction helps.
**Example**

If $Y$ has a binomial probability with $n = 500$ and $p = 0.2$, what is the probability of 110 or more successes?

**Exact:**

```r
> 1 - pbinom(109, 500, 0.2)
[1] 0.1443539
```

**Normal approximation:**

```r
> mu = 500 * 0.2
> sigma = sqrt(500 * 0.2 * 0.8)
> 1 - pnorm(109.5, mu, sigma)
[1] 0.1440878
```

With R, do the exact calculation. With a calculator, use the normal approximation.