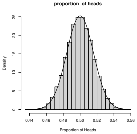
Chapter 11 Notes

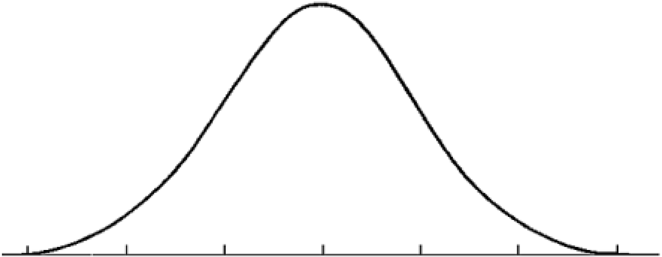
The Normal Distribution
STP-231

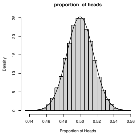
Arizona State University



The normal Curve

- The normal density curve has a symmetric “bell-shaped” curve
- It is symmetric and unimodal





The Normal Distribution

- If a random variable Y follows a normal distribution, then it is distributed as

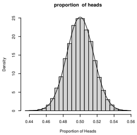
  $$Y \sim N(\mu, \sigma)$$

  Where N means the distribution is normal. More formally,

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

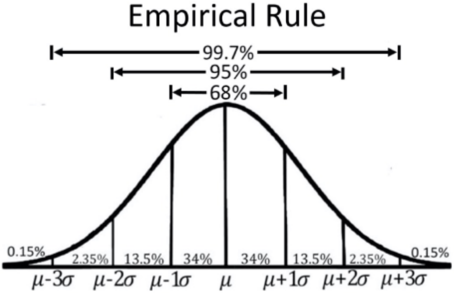
- μ represents the population mean, which can be positive, negative, or zero. This shifts where the peak is left or right

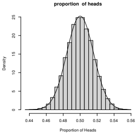
- σ represents the standard deviation, which is always greater than zero. The widens/thins our curve



Basis for the Empirical Rule

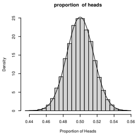
- Recall the empirical rule, that 68% of data is within 1 standard deviation, 95% within 2, 99.7 within 3





Mean and Median

- They are the same! The normal curve is symmetric and unimodal
- The highest peak is at the mean. That is, the mode=median=mean.
- The distribution is symmetric to the mean divides the distribution in half
- Since it is symmetric around the 50%-ile, that is why the mean=median



Standardization and Z-score

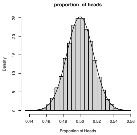
- The Z-score represents values in relation to μ and σ

- How many standard deviations above or below the mean is a particular value

- Unit of measurement does not affect the z-score value

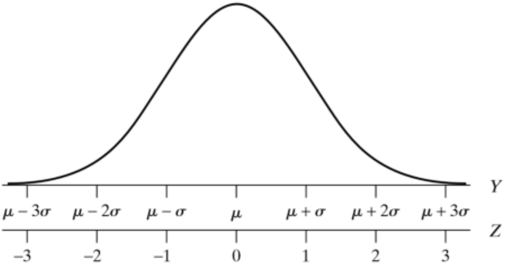
- The transformation is of the form:

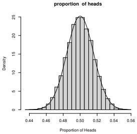
$$z = \frac{y - \mu}{\sigma}$$



Y vs Z

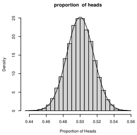
- If we recall from earlier chapter, a linear transformation does not affect the shape of distribution





Calculating Probability from Z-scores

- Draw a standard normal curve
- Label the z-score(s) on the curve
- Shade in the region of interest
- Determining the corresponding area under the standard normal curve using a Z-table



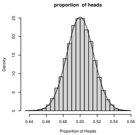
Example

Find the area to the left of specified z-scores

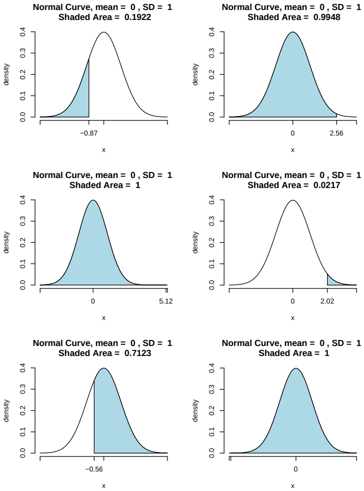
- -0.87
- 2.56
- 5.12

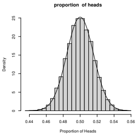
Find the area to the right of specified z-scores

- 2.02
- -0.56
- -4



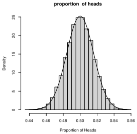
Example Answer





New Example: Area between curves

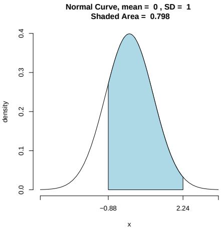
Determine the area under the standard normal curve that lies between -0.88 and 2.24

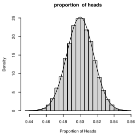


New Example: Area between curves

Determine the area under the standard normal curve that lies between -0.88 and 2.24

$$
\begin{aligned}
\Pr(-0.88 \leq Z \leq 2.24) &= \Pr(Z \leq 2.24) - \Pr(Z \leq -0.88) \\
&= \Pr(Z < 2.24) - \Pr(Z < -0.88) \\
&\approx -0.9875 - 0.1894 = 0.7984
\end{aligned}
$$



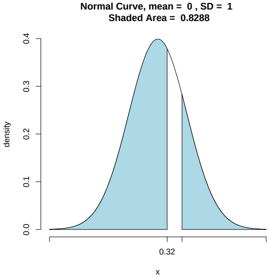


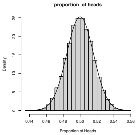
Another Example

Find the area below 0.32 and above 0.83 under the curve

$$\Pr(Z < 0.32) + \Pr(Z > 0.83) \approx 0.6255 + (1 - \Pr(Z < 0.83) \approx 0.8288$$

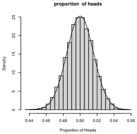
$$\text{OR } 1 - \Pr(0.32 < Z < 0.83)$$





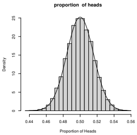
Probability for a Normally Distributed Variable

1. Sketch the normal curve associated with the variable
2. Shade the region of interest and mark its delimiting y-value(s)
3. Find the z-score(s) for the delimiting y-value(s) found in step 2
4. Use the table to find the area under the standard normal curve delimited by the z-score(s) found in step 3.



Another Example

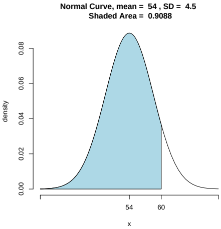
In a certain population of the herring Pomolobus aestivalis, the lengths of the individual fish follow a normal distribution. The mean length of the fish if 54.0 mm, and the standard deviation is 4.5mm. What percentage of fish are less than 60 mm long?

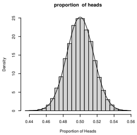


Another Example

In a certain population of the herring Pomolobus aestivalis, the lengths of the individual fish follow a normal distribution. The mean length of the fish if 54.0 mm, and the standard deviation is 4.5mm. What percentage of fish are less than 60 mm long?

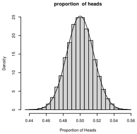
$$z = \frac{y - \mu}{\sigma} = \frac{60 - 54}{4.5} = 1.33$$





More Examples

A variable is normally distributed with mean 68 and standard deviation 10. Find the percentage of all possible values of the variable that

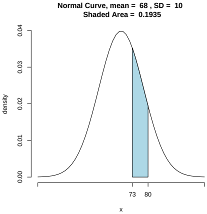


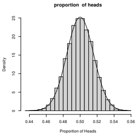
More Examples

A variable is normally distributed with mean 68 and standard deviation 10. Find the percentage of all possible values of the variable that

1. Lie between 73 and 80

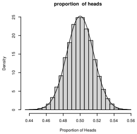
$$\Pr(73 < Y < 80) - \Pr\left(\frac{73-68}{10} < Z < \frac{80-68}{10}\right)$$
$$= \Pr(Z < 1.2) - \Pr(Z < 0.5) \approx 0.1935$$





More Examples

A variable is normally distributed with mean 68 and standard deviation 10. Find the percentage of all possible values of the variable that

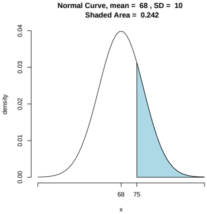

proportion of heads

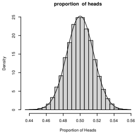
More Examples

A variable is normally distributed with mean 68 and standard deviation 10. Find the percentage of all possible values of the variable that

2. Are at least 75

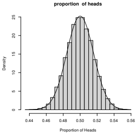
$$\Pr(Y > 75) = 1 - \Pr(Y < 75) = 1 - \Pr(Z < 0.7)$$
$$= \approx 0.242$$


Normal Curve, mean = 68 , SD = 10
Shaded Area = 0.242



More Examples

A variable is normally distributed with mean 68 and standard deviation 10. Find the percentage of all possible values of the variable that



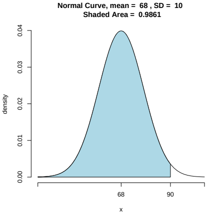
More Examples

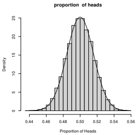
A variable is normally distributed with mean 68 and standard deviation 10. Find the percentage of all possible values of the variable that

3. Are at most 90

$$\Pr(Y < 90) = \Pr\left(Z < \frac{90 - 68}{10}\right)$$

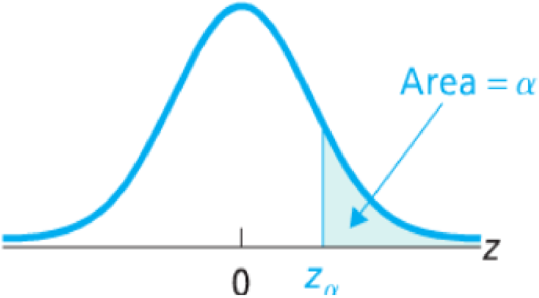
$$= \Pr(Z < 2.2) \approx 0.9861$$

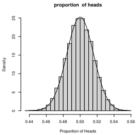




Z_α Notation

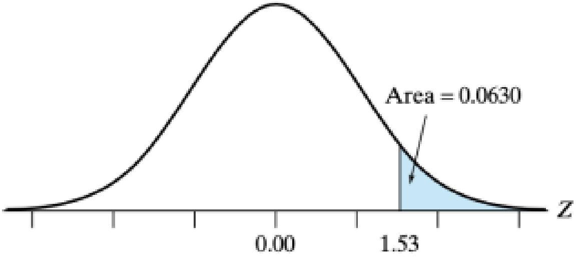
- Z_α is used to denote the z-score that has an area of α to the right under the standard normal curve
- α is a probability. Z_α is a z-score, but not a probability...why is that?

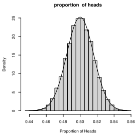




Example

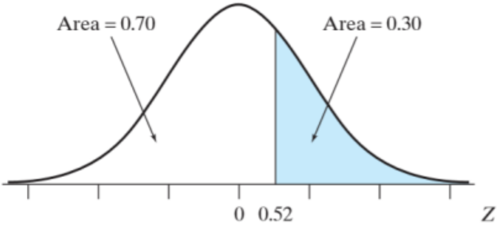
- $Z_{0.0630} = 1.53$

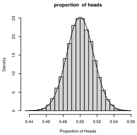




Example

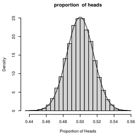
- $Z_{0.30} = 0.52$





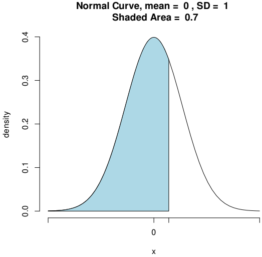
Percentiles

- Percentiles divide the distribution into 100 equal parts.
- Indicates the value below which a given percentage of observations fall
- We can compare to $Z_{\alpha'}$, but here we consider area of α to the <u>left</u>

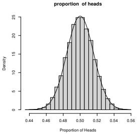


Percentiles Example

Suppose we want to find the 70th percentile of a standard normal distribution. We want to find the z-value that divides the bottom 70% from the top 30%. What is the value?

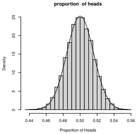


This is at $Z_{0.30} = 0.524$.



%-iles for any Normally Distributed Variable

1. Sketch the normal curve associated with the variable
2. Shade the region of interest
3. Use the table to find the z-score(s) for the delimiting region found in step 2
4. Find the y-value(s) having the z-score(s) found in step 3



How to quickly generate in R

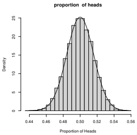
Y is normally distributed with mean 68 and standard deviation 10. Find the value of the 99th percentile:

- We note the 99th percentile is also

$$Z_{0.01} = 2.325$$

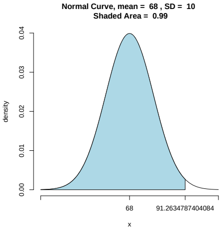
on the Z-scale. So we transform Y to Z:

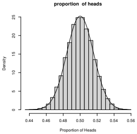
$$z = \frac{y - 68}{10} \implies Y_{.01} = 91.25$$



Graphically

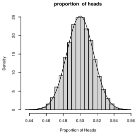
```
library(tigerstats)
#use qnorm() for percentile, pnorm() for area
#code to get 99th percentile
pnormGC(qnorm(0.99), region="below", mean=0,
sd=1,graph=TRUE)

```





Where do quartiles fit in?

Again assume a variable is normally distributed with mean 68 and standard deviation 10. Find the quartiles:



Where do quartiles fit in?

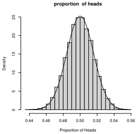
Again assume a variable is normally distributed with mean 68 and standard deviation 10. Find the quartiles:

$$Q_1 = Z_{0.75} = -0.675 \xrightarrow{\text{Transform Y}} Q_1[Y] = -0.675 \cdot (10) + 68 = 61.26$$

$$Q_2 = Z_{0.50} = 0 \xrightarrow{\text{Transform Y}} Q_2[Y] = 68$$

$$Q_3 = Z_{0.25} = 0.674 \xrightarrow{\text{Transform Y}} Q_3[Y] = 0.675 \cdot (10) + 68 = 74.74$$

Find the value that 85% of all possible values of the variable exceed



Where do quartiles fit in?

Again assume a variable is normally distributed with mean 68 and standard deviation 10. Find the quartiles:

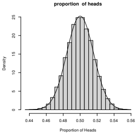
$$Q_1 = Z_{0.75} = -0.675 \xrightarrow{\text{Transform Y}} Q_1[Y] = -0.675 \cdot (10) + 68 = 61.26$$

$$Q_2 = Z_{0.50} = 0 \xrightarrow{\text{Transform Y}} Q_2[Y] = 68$$

$$Q_3 = Z_{0.25} = 0.674 \xrightarrow{\text{Transform Y}} Q_3[Y] = 0.675 \cdot (10) + 68 = 74.74$$

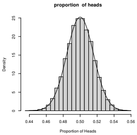
Find the value that 85% of all possible values of the variable exceed If 85% exceed then 15% don't. So we want

$Z_{0.85} = -1.35$. Which means $Y_{0.15} = -1.35 \cdot 10 + 58 = 54.5$



Example continued

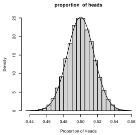
Again assume a variable is normally distributed with mean 68 and standard deviation 10. Find the two values that divide the area under the corresponding normal curve into a middle area of 0.90 and two outside areas of 0.05.



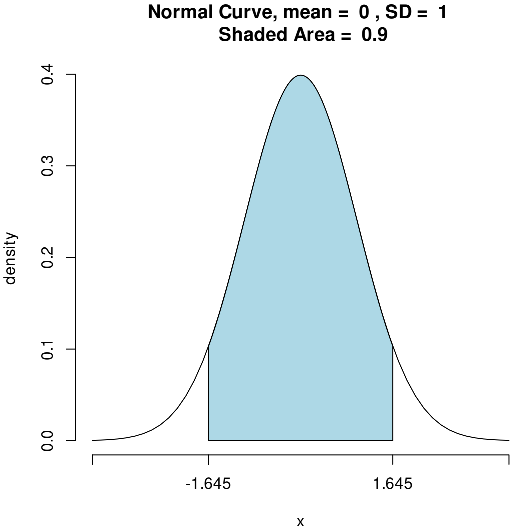
Example continued

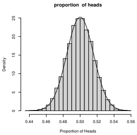
Again assume a variable is normally distributed with mean 68 and standard deviation 10. Find the two values that divide the area under the corresponding normal curve into a middle area of 0.90 and two outside areas of 0.05.

Note because of symmetry $\Pr(X \geq a) = \Pr(X \leq -a)$ for some a, so its also true $Z_\alpha = -Z_{1-\alpha}$. We see that in this example. We want the 5th and 95th percentile, which are respectively -1.645 and 1.645.



Pictoral Representation



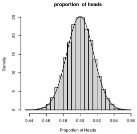


Word Problem

The salinity, or salt content, in the ocean is expressed in parts per thousand (ppt). The number varies due to depth, rainfall, evaporations, river runoff, and ice formation. During January and February, the mean salinity in a region of the northeast continental shelf was 34.08 ppt. The distribution of salinity is normal w/ standard deviation 0.52 ppt. Suppose a random sample of ocean water from this region is obtained.

- What is the probability the salinity is more than 35 ppt?
- A certain species of fish can only survive if the salinity is between 33pt and 35 ppt. What is the probability this species can survive in a randomly selected area?
- Find the salinity that corresponds to the 65th percentile

Assessing Normality



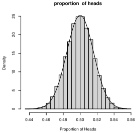
Verify Empirical Rule

Any normally distributed random variable has the following properties:

- 68% of all possible observations lie within one standard deviation to either side of the mean, that is, between $\mu - \sigma$ and $\mu + \sigma$

- 95% of all possible observations lie within two standard deviation to either side of the mean, that is, between $\mu - 2\sigma$ and $\mu + 2\sigma$

  99.7% of all possible observations lie within 3 standard deviations to either side of the mean, that is, between $\mu - 3\sigma$ and $\mu + 3\sigma$

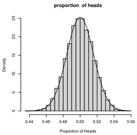


Example: Number of Siblings

Number of Siblings	Frequency	$\Pr(Y = y_i)$
0	4	$\frac{4}{29}$
1	9	$\frac{9}{29}$
2	10	$\frac{10}{29}$
3	3	$\frac{3}{29}$
4	1	$\frac{1}{29}$
5	0	$\frac{0}{29}$
6	0	$\frac{0}{29}$
7	2	$\frac{2}{29}$

How well does the empirical rule estimate this data? $\mu_Y = 1.931$ and $\sigma_Y = 1.680$

- 75% of the data lies within 1 standard deviation from the mean, i.e. $(0.251, 3.611)$
- 93% of the data lies within 2 standard deviations from the mean, i.e. $(-1.43, 5.29)$
- 93% of the data lies within 3 standard deviations from the mean, i.e. $(-3.11, 6.97)$



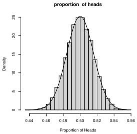
Assessing Samples (Large vs Small)

If we have a large data set:

- Plot a histogram
- Analyze the shape
- Is it normal or approximately normal looking?

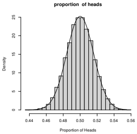
For a small data set

- Make a normal probability plot (aka the quantile plot), because the histogram may not be entirely useful if data-set too small



Normal Quantile Plots (QQ Plots)

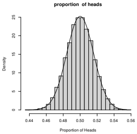
- Statistical graphs that assess normality
- Compare observations with the values we would expect for a standard normal distribution
- Focus on proportions for percentiles rather than the empirical rule
- Normal score: the data points we expect to obtain if our data was normal



QQ Plot Example

- Height of 11 women has mean 65.5 inches and standard deviation 2.9 inches. The smallest observation is 61 inches tall, which means our sample predicts 1/11th of women are 61 inches or shorter in population, i.e. 9.09th percentile

[1] We will use adjusted percentiles in reality



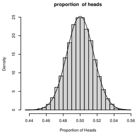
QQ Plot Example

- Height of 11 women has mean 65.5 inches and standard deviation 2.9 inches. The smallest observation is 61 inches tall, which means our sample predicts 1/11th of women are 61 inches or shorter in population, i.e. 9.09th percentile
- If data were truly from a normal distribution (our big assumption), then[1]

$$\mu + Z_{1-0.0909}\sigma = 65.5 - 1.34 \cdot 2.9 = 61.6$$

- Repeat this calculation for rest of the data. Then plot the actual quantiles vs the expected
- Normal score: the data points we expect to obtain if our data was normal

[1] We will use adjusted percentiles in reality

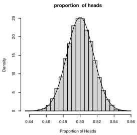


How to Construct a Quantile Plot

1. Rank the observed values from smallest to largest
2. Percentiles=100×$\frac{i}{n}$, where i is the ranked observation index and n is the sample size

   $$\text{We solve for the adjusted percentiles } \left(\frac{i-0.5}{n}\right) \times 100$$

   Where we adjust because the max of the sample should not equate to the 100th percentile
3. Find the corresponding expected normal scores (or z-scores) for the adjusted percentiles
4. Plot the observed values on the y-axis and the normal scores (or z-score) on the x-axis
5. If the plot is roughly linear, you can assume that the plot is approximately normally distributed

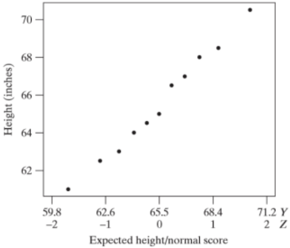


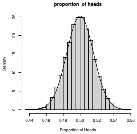
Example

We observe the heights of the 11 women:

$$61, 62.5, 63, 64, 64.5, 65, 66.5, 67, 68, 68.5, 70.5$$

i	1	2	3	4	5	6	7	8	9	10	11
Observed Height	61.0	62.5	63.0	64.0	64.5	65.0	66.5	67.0	68.0	68.5	70.5
Adjusted percentile	4.55	13.64	22.73	31.82	40.91	50.00	59.09	68.18	77.27	86.38	95.45
z	-1.69	-1.10	-0.75	-0.47	-0.23	0.00	0.23	0.47	0.75	1.10	1.69
Theoretical height	60.6	62.3	63.4	64.1	64.8	65.5	66.2	66.9	67.6	68.7	70.4





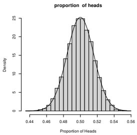
Interpret QQ Plots

- The plotted points fall along an imaginary straight line through (0,mean) when comparing z-scores to normal scores, (0,0) when comparing z-scores to z-scores, or (mean, mean) when comparing normal scores to normal scores

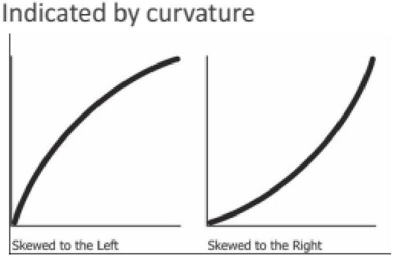
- If the plot is roughly linear, we conclude our data is roughly normally distributed

- Interpret loosely for small samples; adhere strictly for large samples

- Check for curvature at ends (i.e. different tails) and for outliers

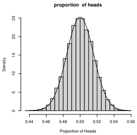


Skewness


Indicated by curvature

The transformations $\sqrt{y}$ or $\ln(y)$ have mild skewness, whereas $\frac{1}{\sqrt{y}}$ and $\frac{1}{y}$ have extreme skew

Normal Approximation of Binomial

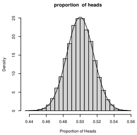


Normal Approximation

- As the number of observations n gets larger, the binomial distribution gets close to normal distribution
- Consider a binomial random variable X. Recall this has n observations, p as a success probability, an expected value and standard deviation of

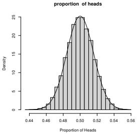
$$\mu_X = np \text{ and } \sigma_X = \sqrt{np(1-p)}$$

- If n is large, then X is approximately $N(np, \sqrt{np(1-p)})$. That is, with enough trials, our outcome of interest ends up having a normal curve describing its potential values
- We consider a sample large enough if $np \geq 10$ and $n(1-p) \geq 10$.



Example

Let's say 1/10 people have been to Antarctica (probably way too high, but let's go with it). What is the probability you know 1 (or more) person who has been to Antarctica, assuming you know 200 people.



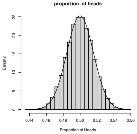
Example

Let's say 1/10 people have been to Antarctica (probably way too high, but let's go with it). What is the probability you know 1 (or more) person who has been to Antarctica, assuming you know 200 people.

- Binomial calculation: Let X be the random variable denoting how many people you know in Antarctica

$$\Pr(X \geq 1) = 1 - \Pr(X = 0) = 1 - \binom{200}{0} 0.1^0 (1-p)^{200} \approx 1$$

So not super helpful...

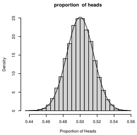


Example Continued

- But the approximation (check the conditions are satisfied) is helpful, because we can now use our tables to calculate probabilities. We can approximate the number of people we know who have been to Antarctica as

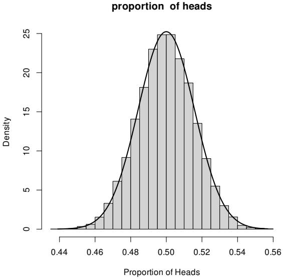
$$X \sim N(.1 * 200, \sqrt{20(1-.1)})$$

$$\begin{aligned}\Pr(Y<0) &= \Pr\left(Z < \frac{0-20}{\sqrt{18}}\right)\\ &= \Pr(Z < -4.71) \approx 1.21 \times 10^{-6}\end{aligned}$$

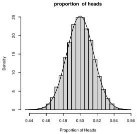


Coin Flip Example

Say we flip a fair coin 1000 times (some large number). We tabulate the number of heads we get, and repeat this experiment 10000 times. (These numbers need not be the same)



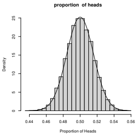
Note, when plotting, we converted the scale from total to relative frequency for the sake of visual clarity



Code to Simulate this Example

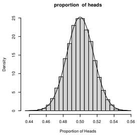
This code runs in base R with no necessary packages. Takes a few seconds to run however.

```
n=1000
p=0.5
sig=sqrt(n*p*(1-p))
aa=sapply(1:10000, function(i) mean(rbinom(1000, 1, .5)))
#b/c mean is same as proportion of 1’s
h <- hist(aa, 40,
col = "lightgray", xlab = "Proportion of Heads", main = "proportion  of heads", freq=F)
xfit <- seq(min(aa), max(aa), length = n)
yfit <- dnorm(xfit, mean = n*p/n, sd = sig/n)
lines(xfit, yfit, col = "black", lwd = 2)

```



New Example

500 guests will attend a dinner event, and it is estimated that 10% of those attendees require a vegan option. What is the probability that the caterers need to have at least 55 plates designated for the vegan option. Use the estimation method: Note, the actual binomial probability is 0.2477



New Example

500 guests will attend a dinner event, and it is estimated that 10% of those attendees require a vegan option. What is the probability that the caterers need to have at least 55 plates designated for the vegan option. Use the estimation method: Note, the actual binomial probability is 0.2477

- We can approximate as $N(500 \cdot 0.1, 500 \cdot 0.1(1-0.1))$. Then,

$$\begin{aligned}\Pr(Y \geq 55) &= 1 - \Pr\left(Z < \frac{55-50}{\sqrt{45}}\right) \\ &= 1 - \Pr(Z < 0.745) = 0.228\end{aligned}$$