1 Even when things are running smoothly, 5% of the parts produced by a certain manufacturing process are defective.

a) If you select 10 parts at random, what is the probability that none of them are defective?

Suppose you have a quality control procedure for testing parts to see if they are defective, but that the test procedure sometimes makes mistakes:

- If a part is good, it will fail the quality control test 10% of the time.
- 20% of the defective parts go undetected by the test.

b) What percentage of the parts will fail the quality control test?

c) If a part passes the quality control test, what is the probability that the part is defective?

d) The parts that fail inspection are sold as "seconds". If you purchase a "second", what is the probability that it is defective?

2 Numbers in a hat. In a hat are slips of paper with the number 1 – 100 on them.

a) If you reach in and grab one slip of paper, what is the probability that the number will be greater than 80?

b) If you reach in and grab one slip of paper, what is the probability that it is either greater than 80 or an odd number?

3 A manufacturing process produces parts that might have one of two faults. Let's call them type 1 and type 2. 1% of parts have a type 1 fault, and 2% of faults have a type 2 fault.

a) If having the two types of faults are **independent** events, what is the probability that a part has a fault (of either type)?

b) If having the two types of faults are **mutually exclusive events**, what is the probability that a part has a fault (of either type)?

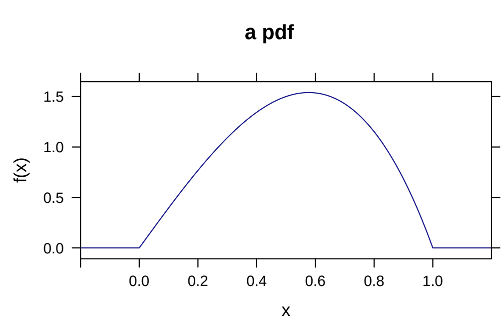
4 Lucky Number. The Lucky Number games works as follows. You pick a lucky number (1 – 6) and then roll three dice. If none of the dice match your lucky number, you get nothing. If one of the dice match your lucky number, you get \$3. If two of the dice match your lucky number, you get \$10. If all three of the dice match your lucky number, you get \$50.

a) Let X be your winnings in one play of the game. What is $\mathrm{E}(X)$?

b) What does $\mathrm{E}(X)$ tell you about this game?

5 The **pdf** for a continuous random variable X is

$$f(x) = \begin{cases} 4(x - x^3) & \text{when } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$



a) Determine $\mathrm{P}(X \le \frac{1}{2})$.

b) Determine the mean and variance of X.

6 The **kernel** of a continuous distribution is given by $k(x) = 4 - x^2$ on the interval $[-2, 2]$.

a) Determine the pdf of the distribution.

b) Compute the mean and variance of the distribution.

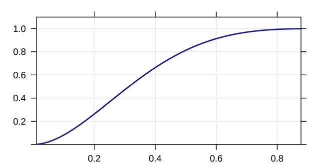
7 The **cdf** for a random variable X is given by $F(x) = x^2$ on $[0, 1]$.

a) What is $\mathrm{P}(X \leq \frac{1}{2})$?

b) What is the median of X?

c) What are the expected value and variance of X?

8 Below is the graph of the **cdf** of a random variable Y.



a) Use the graph to estimate $\mathrm{P}(Y \leq 1/2)$

b) Use the graph to estimate the median of Y.

c) Sketch the pdf for Y.

d) Which is larger, the median of Y or $\mathrm{E}(Y)$?

9 The `mtcars` data frame has some information from Motor Trend about cars.

```
head(mtcars)

##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Write down R commands to do the following:

a) Compute the mean weight (`wt`) of these cars.

b) Compute the mean weight of these cars, separately for 4-cylinder, 6-cylinder, and 8-cylinder vehicles. (How many different ways can you do this?)

c) Create a scatter plot of fuel efficiency (`mpg`) vs. weight.

d) Create side by side scatter plots of fuel efficiency (`mpg`) vs. weight separately for 4-cylinder, 6-cylinder, and 8-cylinder vehicles.

e) Make up additional numerical and graphical summaries you could make from these data and write down the R commands you would use to create them.

10 Claire and Donald conduct an experiment to measure the strength of paper plates. Using the same data, Claire creates a 98% confidence interval for the mean paper plate strength and Donald creates a 95% confidence interval for the mean paper plate strength. Whose interval will be wider? Explain your reasoning.

11 Here are some terms we have used in class: statistic, parameter, sample, population, estimate, estimand, estimator, confidence interval, standard uncertainty, relative uncertainty, standard error, margin of error, critical value.

a) Explain each of the terms in your own words. Where this makes sense, give (or make up) a specific example.

b) Find pairs or groups of terms that are related somehow and explain how they are related. (Have fun with this, see how many relationships you can find.)

12 Beth and Sarah are bowlers. Bowling scores are discrete (whole numbers), but we can model them with a continuous distribution. As a first approximation, let's model their bowling scores as with normal distributions. Use the summary information below (based on their last 25 games) to answer some questions about their bowling performance, assuming that their performances are independent (a pretty good model for bowling, since players do not directly affect each other).

```
##   response bowler min  Q1 median  Q3 max mean   sd  n missing
## 1    score   Beth 109 154    175 196 241  175 31.8 25       0
## 2    score  Sarah 137 166    180 194 223  180 20.9 25       0
```

a) Who is the more consistent bowler? How do you know?

b) What is the probability that Sarah scores 200 or more?

c) What is the probability that Beth scores 200 or more?

d) If they bowl together as a team, what is the probability that their total score is 400 or more?

e) If they bowl against each other, what is the probability that Beth will win (i.e., have a higher score than Sarah)?

13 A sample of 25 pieces of laminate used in the manufacture of circuit boards was tested for warpage (measured in mm) under specified conditions. The mean warpage for the sample was 2.42 mm and the sample standard deviation was 0.26 mm.

a) Compute the standard error of the mean for this sample.

b) Compute a 95% confidence interval for the mean warpage of all circuit boards under these conditions.

14 A 1996 study of Scotch pine specimens reported a 95% confidence interval for the mean elasticity of $14,500 \pm 1100$ MPa for the mean modulus of elasticity obtained 1 minute after applying a certain load.

True or False. This means that the researchers believe that 95% of Scotch pine specimens will have an elasticity between 13,400 and 15,600 MPa.

Explain your reasoning.

15 Below is the summary output for a simple linear model for predicting the length of a child's foot from the width (both in cm).

```
msummary(model)

##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.8172     2.9381   3.341  0.00192 **
## width          1.6576     0.3262   5.081  1.1e-05 ***
##
## Residual standard error: 1.025 on 37 degrees of freedom
## Multiple R-squared:  0.411,Adjusted R-squared:  0.3951
## F-statistic: 25.82 on 1 and 37 DF,  p-value: 1.097e-05
```

a) Give a 95% confidence interval for the slope in this model.

b) What is the predicted response if the predictor variable has a value of of 8.4 cm?

c) Here is the data for David.

```
   name birthmonth birthyear length width sex biggerfoot domhand
1 David          5        88   24.4   8.4   B          L       R
```

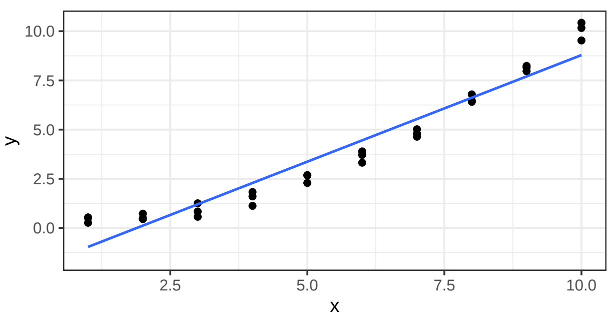
What is David's residual?

16 In linear modeling, what is the difference between a confidence interval and a prediction interval for the response? How do you get R to calculate them for you?

17 An experimenter measures the distance an object moves and how long it takes to move that far and records the measurements in the table below. Use this information to estimate the average speed of the object. Report your answer with the appropriate uncertainty to complete the table.

quantity	estimate	uncertainty
distance (m)	28.15	0.03
time (sec)	6.51	0.02
average speed (m/sec)		

18 The plot below shows some data that indicate a nonlinear relationship between `y` and `x`.



Sketch what the residual plot for this data will look like. Be sure to label the axes.

Solutions

1

a) $\text{P(none defective)} = \text{P(all are good)} = 0.95^{10} = 0.599 = 59.9\%$

b) Even though only 5% are defective, nearly 14% fail the quality control:

$$\begin{aligned}\text{P(fail test)} &= \text{P(good and fail)} + \text{P(bad and fail)} \\ &= \text{P(good)}\,\text{P(fail} \mid \text{good)} + \text{P(bad)}\,\text{P(fail} \mid \text{bad)} \\ &= 0.95(0.10) + 0.05(.80) = 0.095 + 0.04 = 0.135 = 13.5\%\end{aligned}$$

c) If a part passes QC, the probability that it is defective drops from 5% to just over 1%:

$$\begin{aligned}\text{P(bad} \mid \text{pass)} &= \frac{\text{P(bad and pass)}}{\text{P(pass)}} \\ &= \frac{0.05(0.20)}{0.865} = 0.01156\end{aligned}$$

The cost to get this improvement in quality is the cost of the QC test plus the cost of discarding 10% of good parts tested in the QC process.

d)

$$\text{P(bad} \mid \text{fail)} = \text{P(bad and fail)} = \text{P(fail)} = 0.04/(0.04 + 0.095) = 0.2962963$$

```
0.04 / (0.04 + 0.095)

## [1] 0.2962963
```

2

a) By the equally likely rule: $\frac{20}{100}$ since there are 20 such numbers out of 100 in the hat.

b) Let A be the event that the number is odd, and let B be the event that the number if big (bigger than 80). Then $P(B \text{ or } A) = P(A) + P(B) - P(A \text{ and } B) = 0.5 + 0.2 - 0.10 = 0.60$.

3

a) $P(B \text{ or } A) = P(A) + P(B) - P(A \text{ and } B) = 0.01 + 0.02 - (0.01)(0.02) = 0.0298$.

b) $P(B \text{ or } A) = P(A) + P(B) - P(A \text{ and } B) = 0.01 + 0.02 - 0 = 0.03$

4

```
p <- 1/6; q <- 1-p
vals <- c(0, 3, 10, 50)
probs <- setNames(c(q^3, 3 * p * q^2, 3 * p^2 * q, p^3), vals)
probs

##          0          3         10         50
## 0.57870370 0.34722222 0.06944444 0.00462963

# this should be 1
sum(probs)

## [1] 1

# E(X)
sum(vals * probs)

## [1] 1.967593
```

The value of the game is just under \$2. If the person running the game charges \$2 to play, they will in the long run make just over 3 cents per play.

5

```
  f <- makeFun( 4*(x-x^3) ~ x )
  F <- antiD( f(x) ~ x )
 xF <- antiD( x * f(x) ~ x )
xxF <- antiD( x^2 * f(x) ~ x )
F(1) - F(0)                   # should be 1

## [1] 1

m <- xF(1) - xF(0); m         # mean

## [1] 0.5333333

xxF(1) - xxF(0) - m^2         # variance

## [1] 0.04888889
```

6

```
F <- antiD( 4 - x^2 ~ x )
F(2) - F(-2)

## [1] 10.66667
```

We can divide by this value to get the pdf.

```
f <- makeFun( (4 - x^2) / C ~ x, C = F(2) - F(-2) )
integrate(f, -2, 2)

## 1 with absolute error < 1.1e-14

F <- antiD( f(x) ~ x)
F(2) - F(-2)

## [1] 1

xF <- antiD( x*f(x) ~ x )
m <- xF(2) - xF(-2)            # mean
xxF <- antiD( x*x*f(x) ~ x )
xxF(2) - xxF(-2) - m^2         # variance

## [1] 0.8
```

7

a) $\mathrm{P}(X \leq \frac{1}{7}2) = F(1/2) = (1/2)^2 = 1/4$.

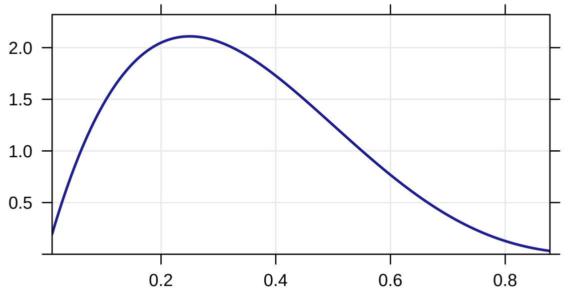
b) The median m satisfies $1/2 = \mathrm{P}(X \leq m) = F(m) = m^2$. So $m = \sqrt{1/2} \approx 0.7071$.

c) $f(x) = 2x$ on $[0, 1]$. We can easily integrate by hand, but here is a computer solution.

```
f <- function(x, k) x^k * 2 * x
# this should be 1
integrate(f, 0, 1, k = 0)

## 1 with absolute error < 1.1e-14

# this is the expected value
integrate(f, 0, 1, k = 1)

## 0.6666667 with absolute error < 7.4e-15

# this is E(X^2)
integrate(f, 0, 1, k = 2)

## 0.5 with absolute error < 5.6e-15

# variance: Var(X) = E(X^2) - E(X)^2
value(integrate(f, 0, 1, k = 2)) - value(integrate(f, 0, 1, k = 1))^2

## [1] 0.05555556
```

8

```
## P(Y <= 1/2)
##         0.8
## median
##    0.3
```

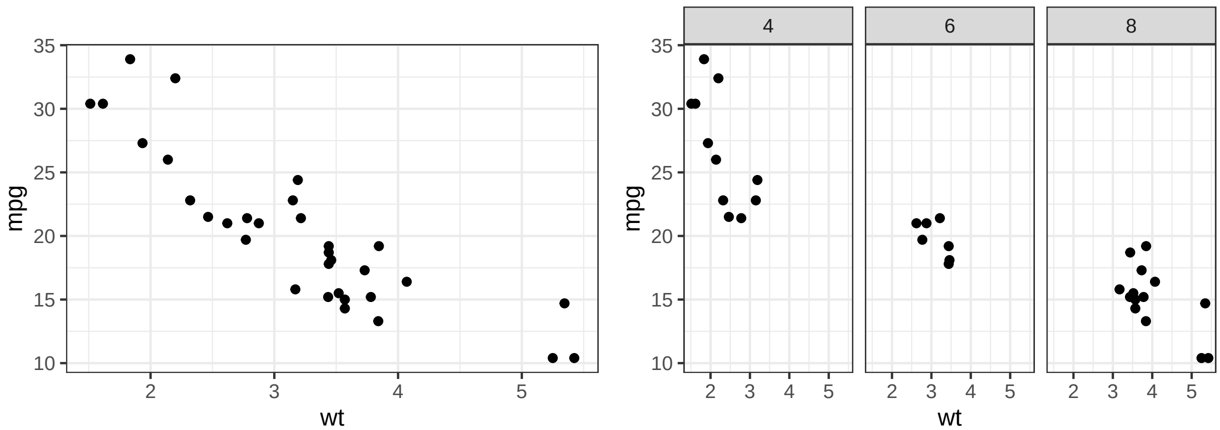



```
##      mean
## 0.3333333
```

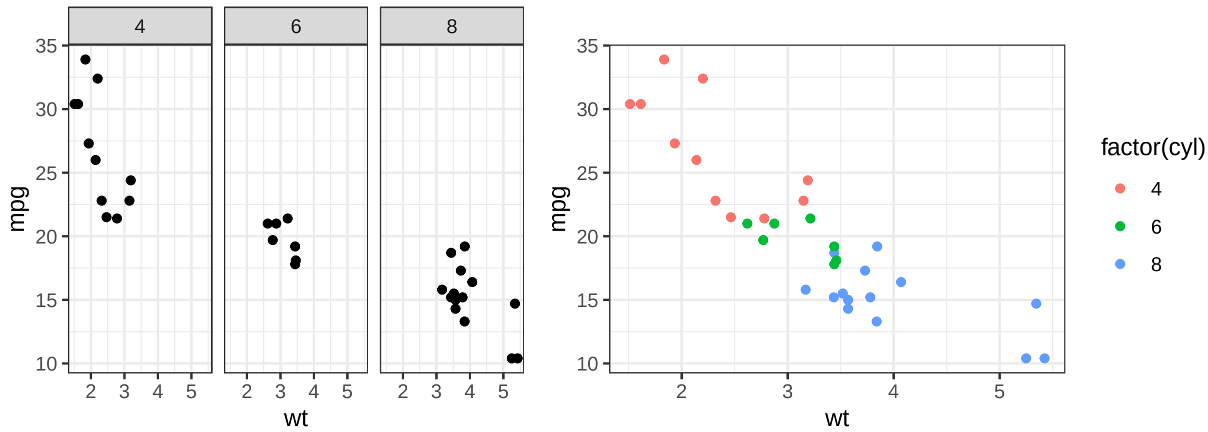
You can't work out the mean so precisely from your sketch of the pdf, but you can know the mean is larger because the distribution is skewed right.

9

```
mean(~wt, data = mtcars)

## [1] 3.21725

mean(~wt | cyl, data = mtcars)

##        4        6        8
## 2.285727 3.117143 3.999214

mean(wt ~ cyl, data = mtcars)

##        4        6        8
## 2.285727 3.117143 3.999214

mean(~ wt, groups = cyl, data = mtcars)

##        4        6        8
## 2.285727 3.117143 3.999214

gf_point(mpg ~ wt, data = mtcars) |
gf_point(mpg ~ wt | cyl, data = mtcars)
```




```
gf_point(mpg ~ wt | factor(cyl), data = mtcars) |
gf_point(mpg ~ wt, color = ~ factor(cyl), data = mtcars)
```



10 Claire's interval will be wider. A larger confidence level requires a wider interval because we must cover the true parameter value for a higher proportion of samples. In the formula, this can be seen because t_* will be larger.

11 The key here is not to memorize certain phrases but to make sure you have a solid understanding of the concepts.

12

```
tibble(
   m1 = 175,
   m2 = 180,
   s1 = 30,
   s2 = 20,
   over200A = 1 - pnorm( 200, m1, s1),
   over200B = 1 - pnorm( 200, m2, s2),
   over400 = 1 - pnorm( 400, m1+m2, sqrt(s1^2 + s2^2)),
   win     = 1 - pnorm( 0, m1-m2, sqrt(s1^2 + s2^2))
) %>% as.data.frame()

##    m1  m2 s1 s2  over200A  over200B   over400       win
## 1 175 180 30 20 0.2023284 0.1586553 0.1060017 0.4448535
```

13

```
##    n x_bar    s   t_star    SE        me        lo       hi
## 1 25  2.42 0.26 2.063899 0.052 0.1073227 2.312677 2.527323
```

14 False. Confidence intervals are about estimating an unknown parameter (in this case the mean elasticity in the population of all Scotch pine trees), not about containing a large fraction of the population.

Note: It is also important to understand what the confidence level measures. A 95% confidence interval makes a claim not about all possible samples in the situation at hand. 95% of samples should produce intervals that cover the estimand (and 5% should fail to do so). We won't know for any particular interval whether it contains the estimand or not.

15

```
t.star <- qt(0.975, df = 37); t.star

## [1] 2.026192

1.6576 - t.star * 0.3262   # lower end of CI

## [1] 0.996656

1.6576 + t.star * 0.3262   # upper end of CI

## [1] 2.318544
```

16 For a confidence interval, the estimand is the mean response (for a given value of the predictor variable). For a prediction interval, the estimand is the response value for a single new observation (for a given value of the predictor variable).

17

$$u_v = \sqrt{\left(\frac{\partial v}{\partial s}\right)^2 u_s^2 + \left(\frac{\partial v}{\partial t}\right)^2 u_t^2}$$

where $v = \frac{s}{t}$, $\frac{\partial v}{\partial s} = \frac{1}{t}$, and $\frac{\partial v}{\partial t} = \frac{-s}{t^2}$.

(It is also possible to do this problem using relative uncertainty, but make sure you convert back to absolute uncertainty at the end.)

```
##       s  u_s    t  u_t        v        u_v
## 1 28.15 0.03 6.51 0.02 4.324117 0.01406113
```

18 Two versions:

```
## `geom_smooth()` using formula 'y ~ x'
```