431 Class 06

thomaselove.github.io/431

2021-09-09

Today's R Packages

```
library(janitor)
library(knitr)
library(magrittr)
library(patchwork)
library(tidyverse) # always load tidyverse last

theme_set(theme_light()) # other TEL option: theme_bw()
```

- As mentioned in Class 5, I used `{r, message = FALSE}` in the code chunk header to suppress messages about conflicts between R packages.

Ingesting Today's Data

```
dm431 <- read_csv("data/dm_431.csv",
                  show_col_types = FALSE) %>%
  clean_names() %>%
  mutate(across(where(is.character), as_factor)) %>%
  mutate(class5_id = as.character(class5_id))
```

- This is the same approach we wound up with in Class 5.

dm431 codebook (part 1)

- This sample includes 431 female adults living with diabetes in Cuyahoga County at ages 30-64, and who have complete data on all of the variables listed in this codebook.

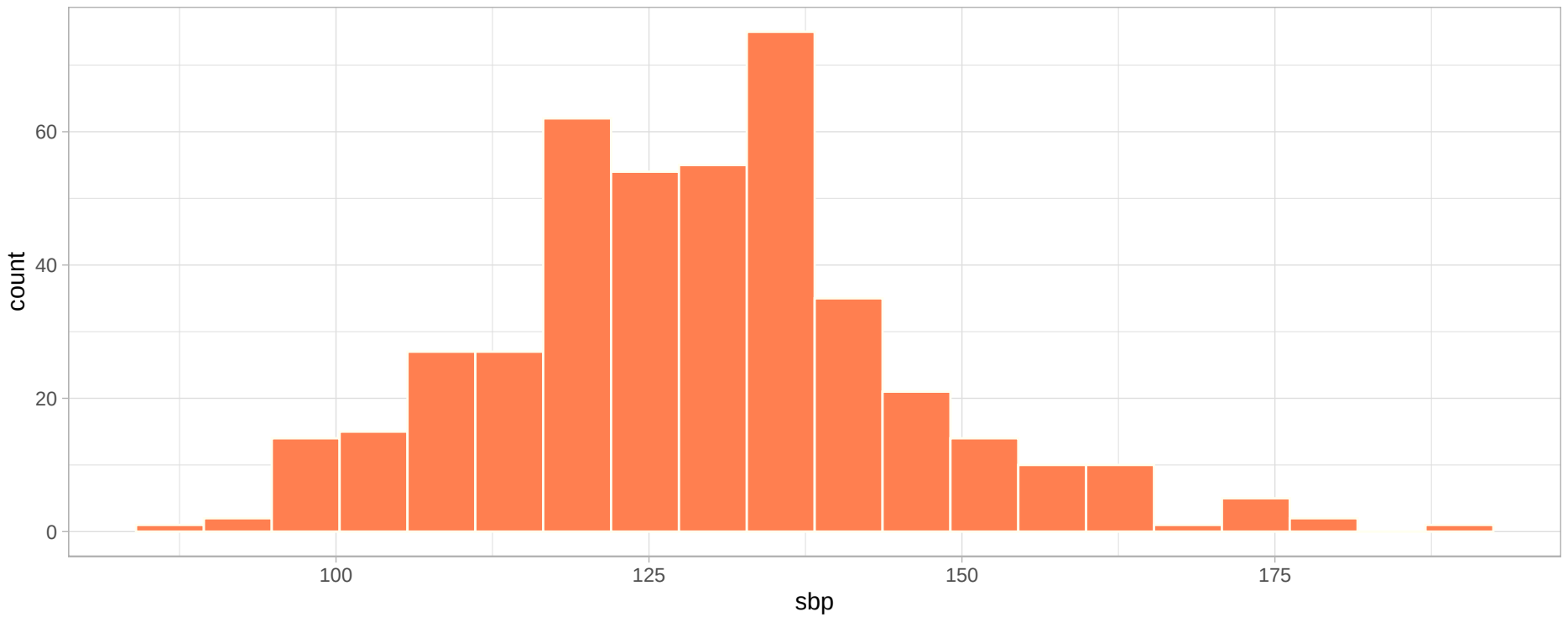
Variable	Description
class5_id	subject code (S-001 through S-431)
age	subject's age, in years
sbp	most recent systolic blood pressure (mm Hg)
dbp	most recent diastolic blood pressure (mm Hg)
n_income	neighborhood median income, in $
ldl	most recent LDL cholesterol level (mg/dl)
a1c	most recent Hemoglobin A1c (%, with one decimal)
insurance	primary insurance, 4 levels
statin	1 = prescribed a statin in past 12m, 0 = not

Remainder of `dm431` codebook (part 2)

Variable	Description
`ht`	height, in meters (2 decimal places)
`wt`	weight, in kilograms (2 decimal places)
`tobacco`	most recent tobacco status, 3 levels
`eye_exam`	1 = diabetic eye exam in past 12m, 0 = not
`race_ethnicity`	race/ethnicity category, 3 levels
`sex`	all subjects turn out to be Female
`county`	all subjects turn out to be in Cuyahoga County

Histogram of 431 sbp values

```
ggplot(data = dm431, aes(x = sbp)) +
  geom_histogram(bins = 20, fill = "coral", col = "ivory") +
  labs("Systolic Blood Pressure for 431 women with diabetes")
```



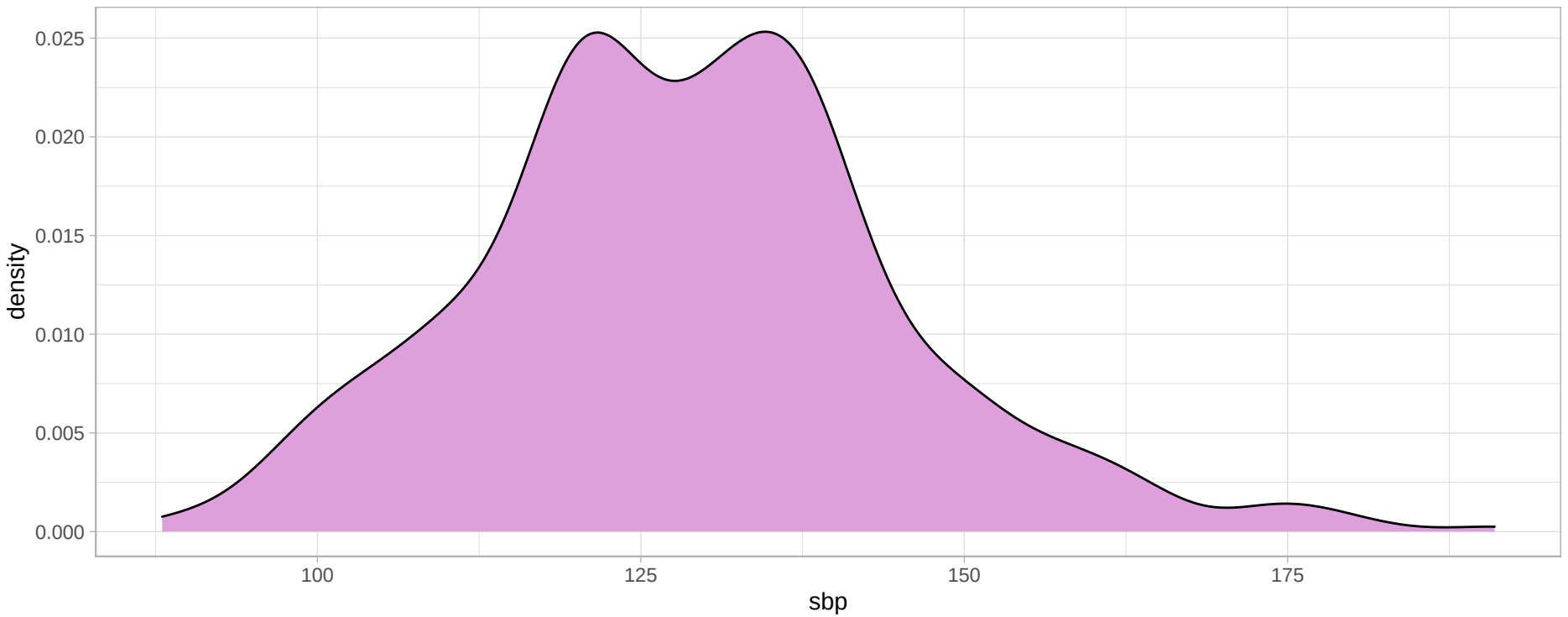
Does a Normal model fit our data "well enough"?

We evaluate whether a Normal model fits sufficiently well to our data on the basis of (in order of importance):

1. Graphs (**DTDP**) are the most important tool we have
   - There are several types of graphs available that are designed to (among other things) help us identify clearly several of the potential problems with assuming Normality.
2. Planned analyses after a Normal model decision is made
   - How serious the problems we see in graphs need to be before we worry about them changes substantially depending on how closely the later analyses we plan to do rely on the assumption of Normality.
3. Numerical Summaries are by far the least important even though they seem "easy-to-use" and "objective".

Density Plot of the 431 sbp values

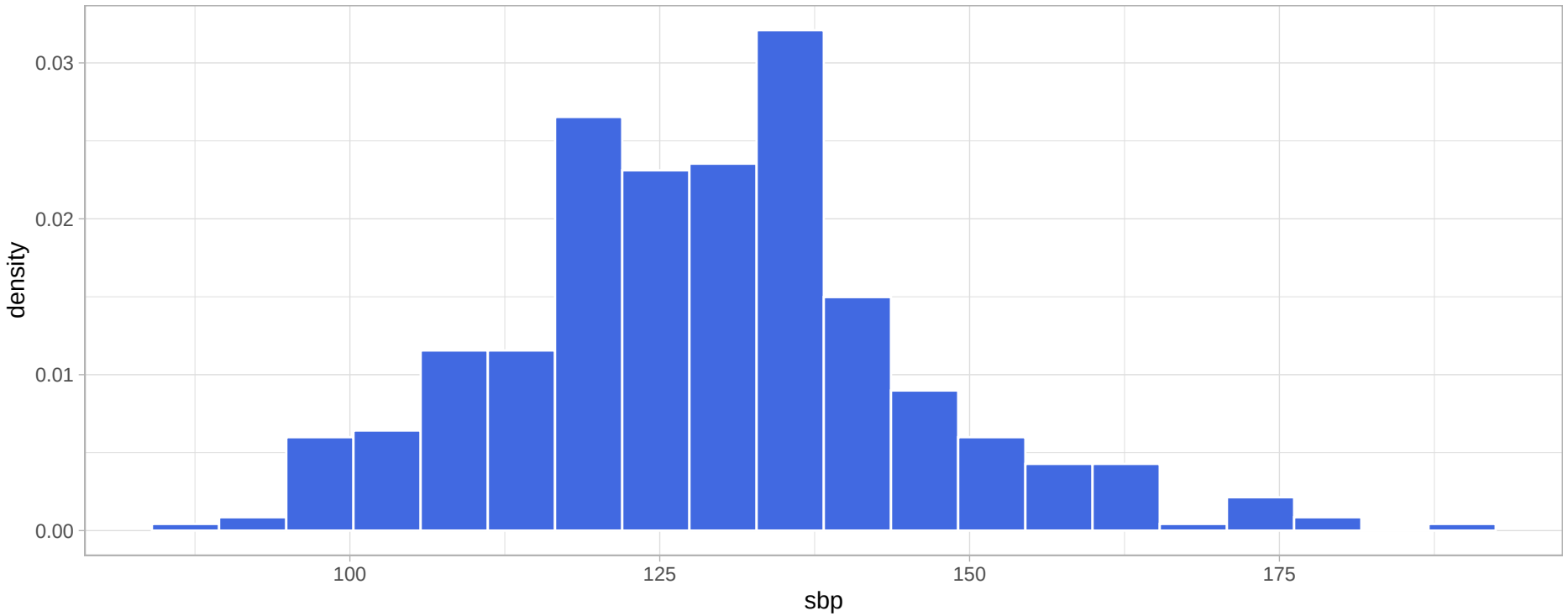
```
ggplot(data = dm431, aes(x = sbp)) +
  geom_density(fill = "plum") +
  labs("Systolic Blood Pressure for 431 women with diabetes")
```



Rescale dm431 SBP histogram as density

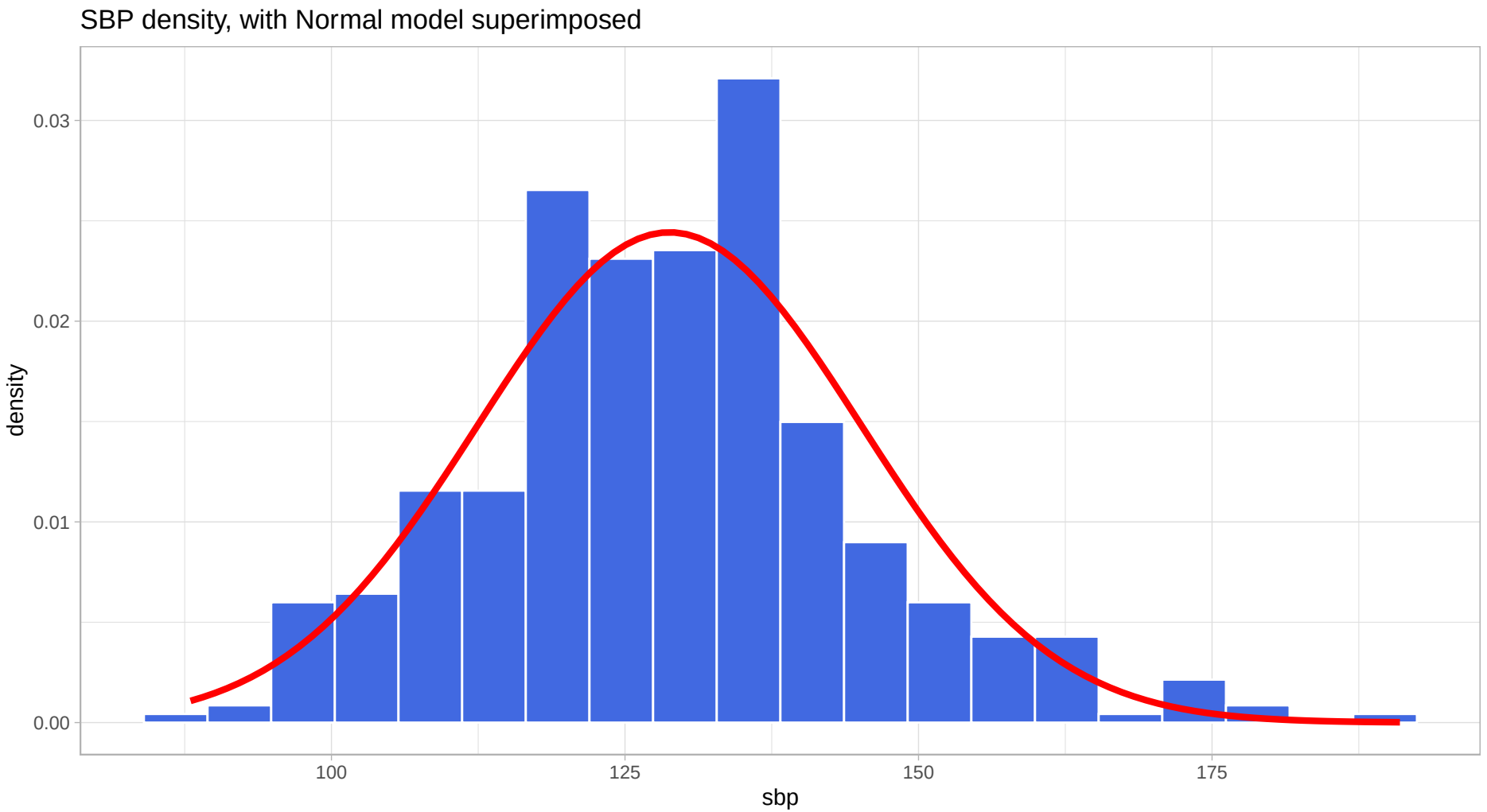
Suppose we want to rescale the histogram counts so that the bar areas integrate to 1. This will let us overlay a Normal density onto the results.

```
ggplot(dm431, aes(x = sbp)) +
  geom_histogram(aes(y = stat(density)), bins = 20,
                 fill = "royalblue", col = "white")
```



Density Function, with Normal superimposed

Now we can draw a Normal density curve on top of the rescaled histogram.

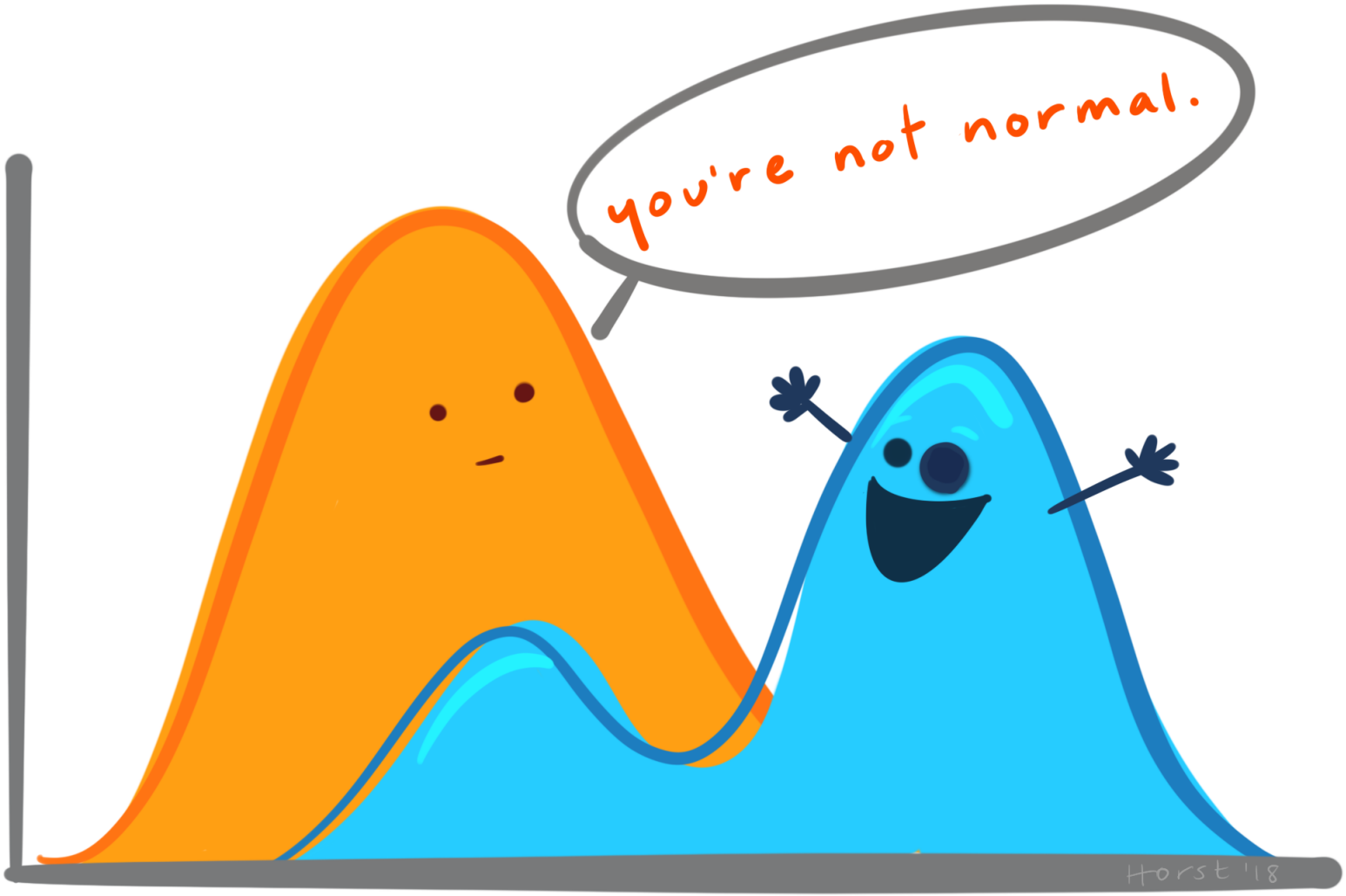


Code for plotting Histogram as Density function

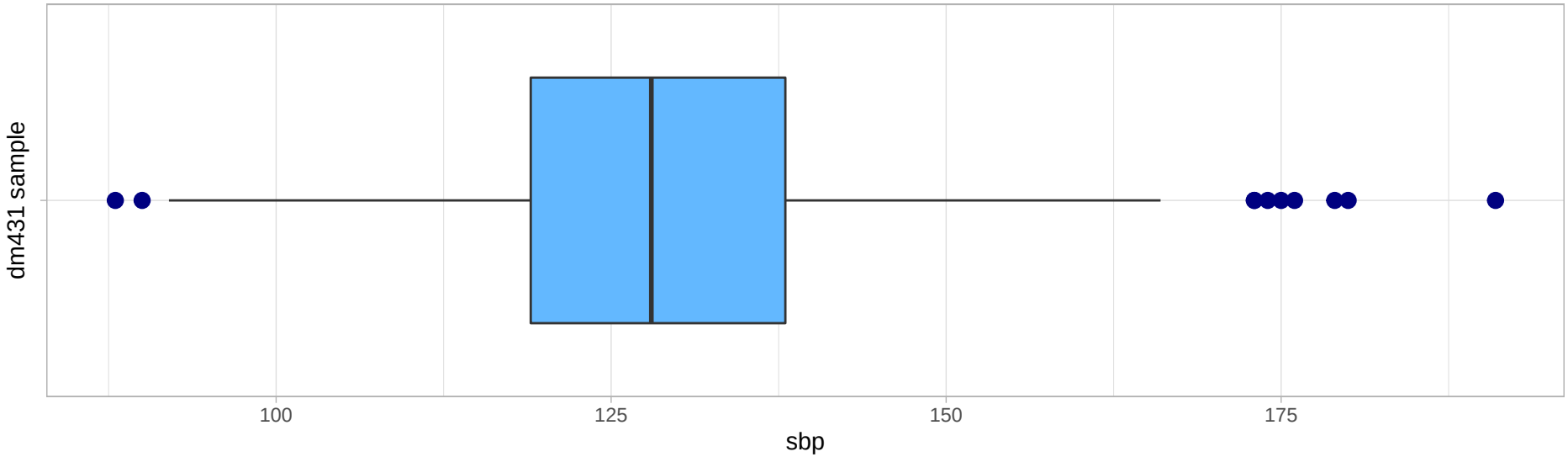
Including the superimposition of a Normal density on top of the histogram.

```
ggplot(dm431, aes(x = sbp)) +
  geom_histogram(aes(y = stat(density)), bins = 20,
                 fill = "royalblue", col = "white") +
  stat_function(fun = dnorm,
                args = list(mean = mean(dm431$sbp),
                            sd = sd(dm431$sbp)),
                col = "red", lwd = 1.5) +
  labs(title = "SBP density, with Normal model superimposed")
```

Graphs are our most important tool!



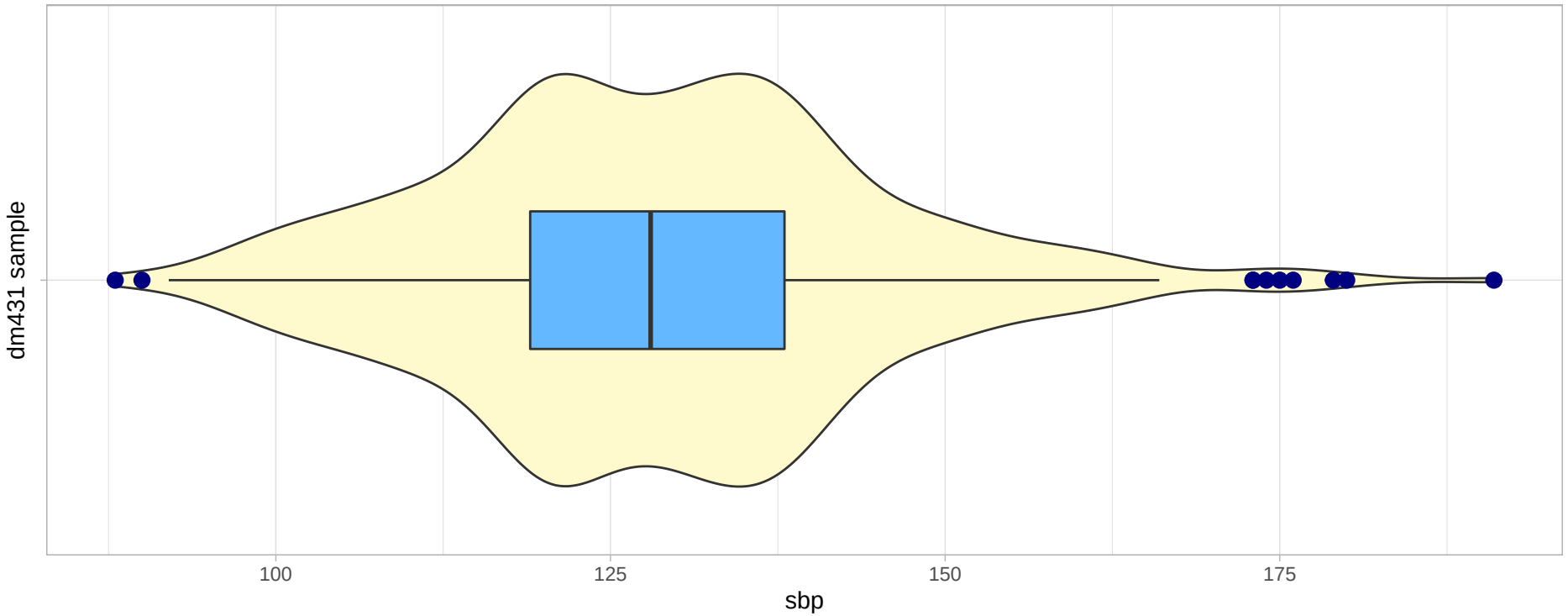
Boxplot for `dm431` Systolic BP values



min	Q1	median	Q3	max	mean	sd	n	missing
88	119	128	138	191	128.8	16.3	431	0

- Box located at the quartiles (Q1 and Q3), with central line at median
- IQR = interquartile range = Q3 - Q1 = width of the box
- Fences identifying outlier candidates at Q1 - 1.5 IQR and Q3 + 1.5 IQR
- Center, Spread, Shape?

Adding a Violin Plot of `dm431` Systolic BPs



min	Q1	median	Q3	max	mean	sd	n	missing
88	119	128	138	191	128.8	16.3	431	0

- What does the violin plot suggest about the shape?

Code for Previous Slide

```
ggplot(dm431, aes(x = "", y = sbp)) +
  geom_violin(fill = "lemonchiffon") +
  geom_boxplot(width = 0.3, fill = "royalblue",
               outlier.size = 3,
               outlier.color = "royalblue") +
  coord_flip() +
  labs(x = "dm431 sample")

mosaic::favstats(~ sbp, data = dm431) %>%
  kable(digits = 1)
```

- Remember that you have the R Markdown code for every slide!

What would a sample of 431 systolic blood pressures from a Normal distribution look like?

New simulated sample from Normal distribution

Simulate a sample of 431 observations from a Normal distribution that has mean and standard deviation equal to the mean and standard deviation of our dm431 systolic blood pressures.

```
dm431 %>% summarize(mean(sbp), sd(sbp)) %>% kable(dig = 2)
```

mean(sbp)	sd(sbp)
128.79	16.33

```
set.seed(20210909) # note change from Class 05
sim_data <- tibble(
  sbp = rnorm(n = 431,
              mean = mean(dm431$sbp),
              sd = sd(dm431$sbp)))
```

Observed & Simulated Numerical Summaries

```
mosaic::favstats(~ sbp, data = dm431) %>% kable(dig = 1)
```

min	Q1	median	Q3	max	mean	sd	n	missing
88	119	128	138	191	128.8	16.3	431	0

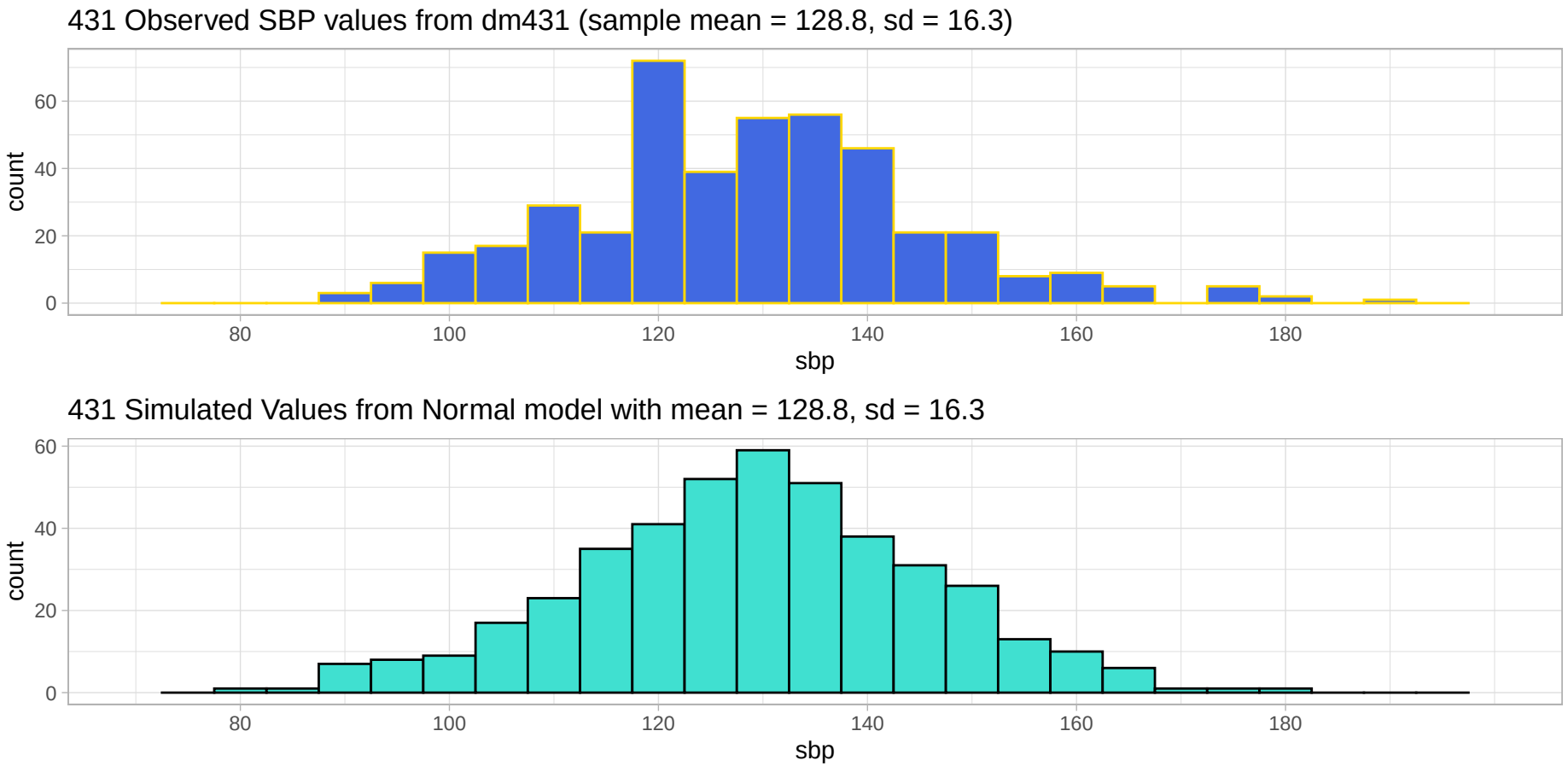
```
mosaic::favstats(~ sbp, data = sim_data) %>% kable(dig = 1)
```

min	Q1	median	Q3	max	mean	sd	n	missing
80.5	118.9	129.6	139.6	178.8	129	16.5	431	0

- The first time you use `mosaic::favstats` in an R Markdown file, R will throw a message about a function that `ggplot2` and `mosaic` share.
- I suppress it with `{r, message = FALSE}` in the code chunk header.

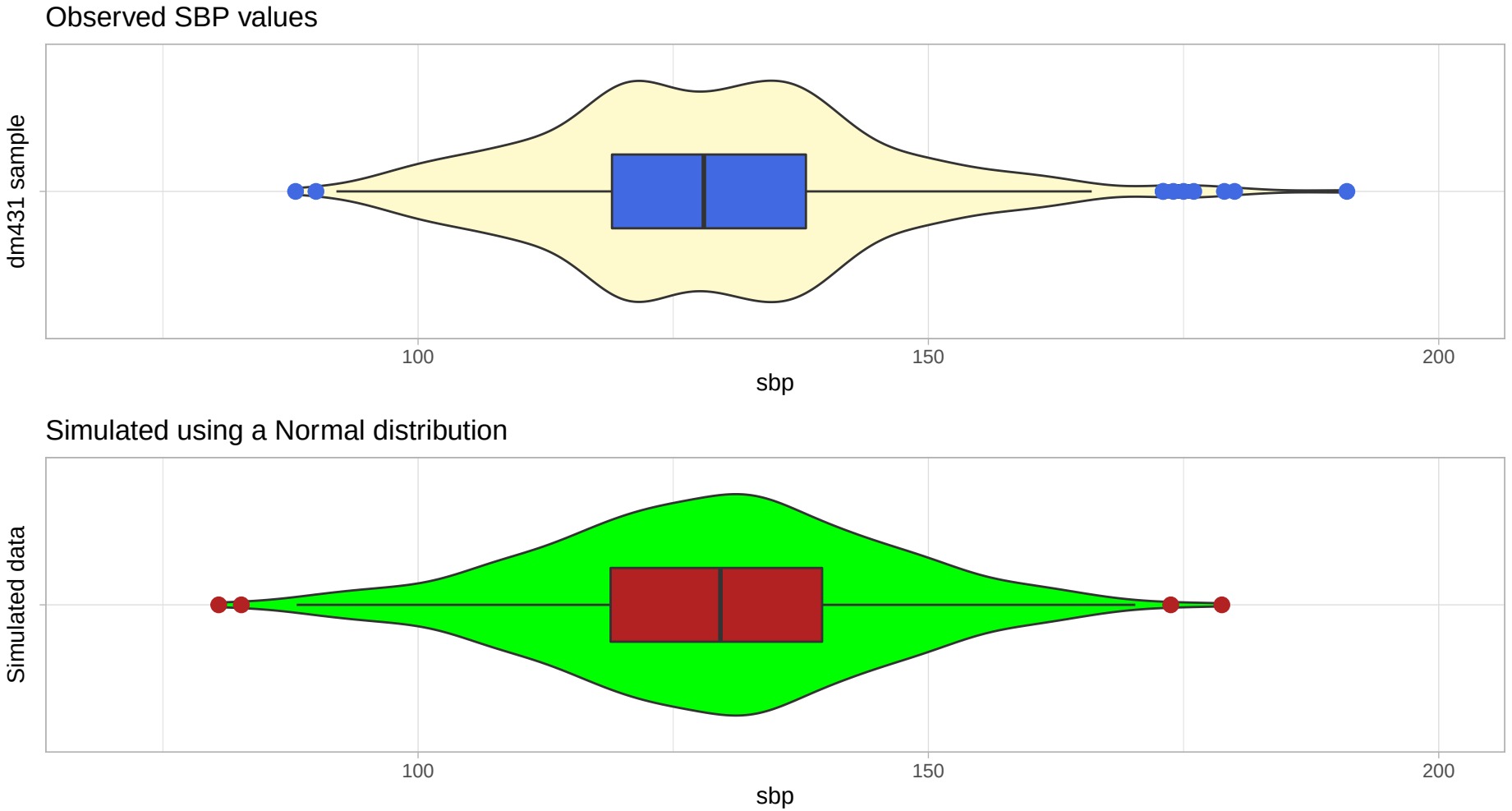
So, do the `sbp` values we observe look like our simulated sample from a Normal distribution?

Comparing histograms of dm431 and simulated SBP



- Does a Normal model look appropriate for describing the SBP in dm431?

Observed vs. Simulated Systolic BPs

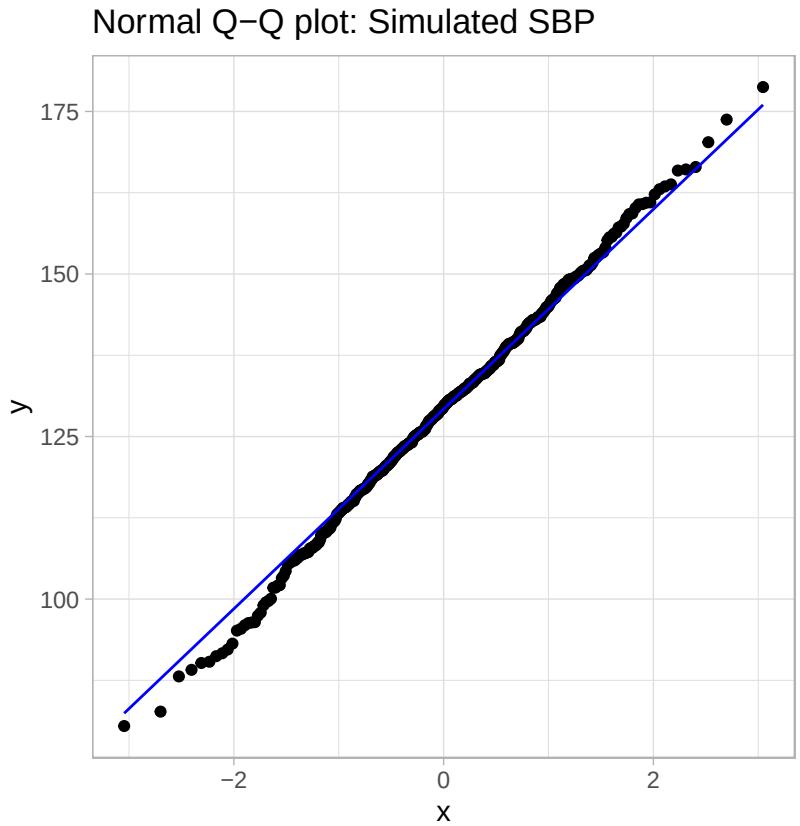


- Does a Normal model look appropriate for describing the SBP in dm431?

Using a Normal Q-Q plot to assess Normality of a batch of data

Normal Q-Q plot of our simulated data

Remember that these are draws from a Normal distribution, so this is what a sample of 431 Normally distributed data points should look like.



The Normal Q-Q Plot

Tool to help assess whether the distribution of a single sample is well-modeled by the Normal.

- Suppose we have N data points in our sample.
- Normal Q-Q plot will plot N points, on a scatterplot.
  - Y value is the data value
  - X value is the expected value for that point in a Normal distribution

Using the Normal distribution with the same mean and SD as our sample, R calculates what the minimum value is expected to be, given a sample of size N, then the next smallest value, and so forth all the way up until the maximum value.

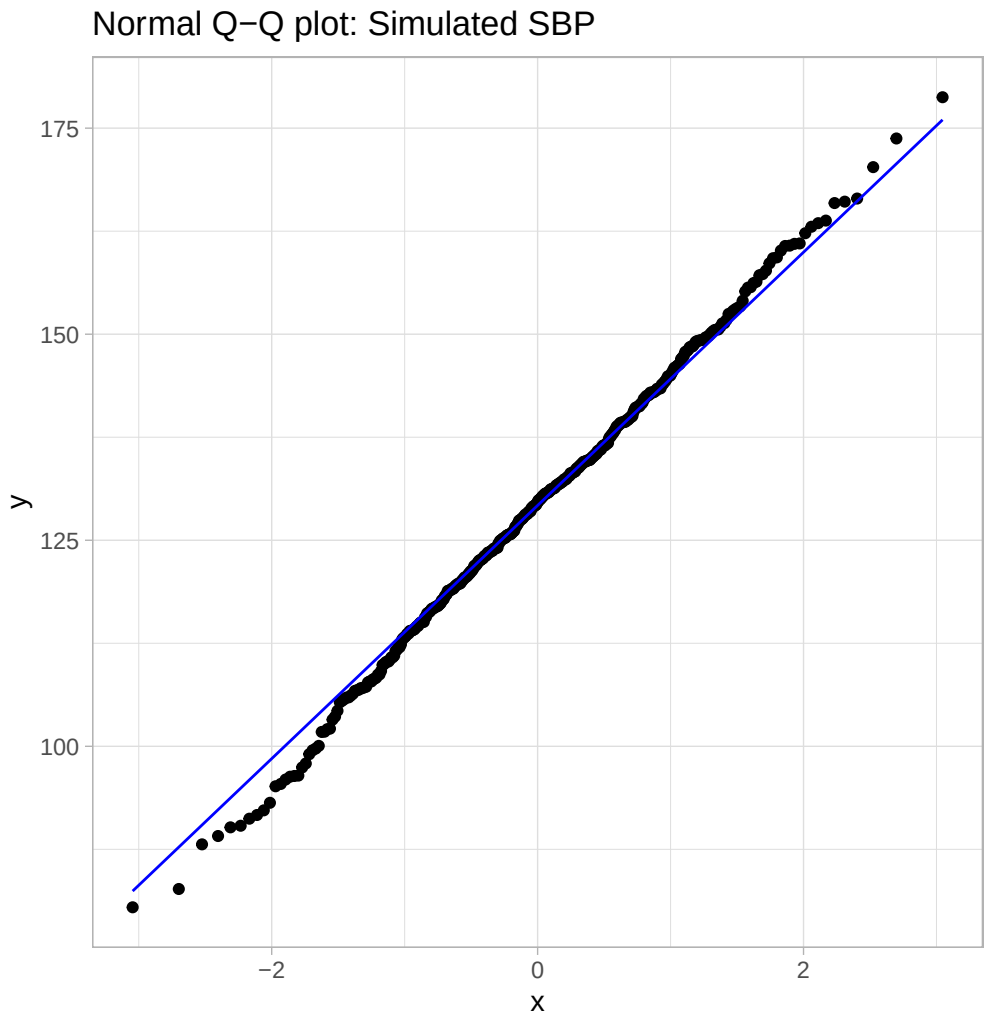
- X value in the Normal Q-Q plot is the value that a Normal distribution would take for that rank in the data set.
- We draw a line through Y = X, and points close to the line therefore match what we'd expect from a Normal distribution.

How do we create a Normal Q-Q plot?

For our simulated blood pressure data

```
ggplot(sim_data, aes(sample = sbp)) +
  geom_qq() + # plot the points
  geom_qq_line(col = "blue") + # plot the Y = X line
  theme(aspect.ratio = 1) + # make the plot square
  labs(title = "Normal Q-Q plot: Simulated SBP")
```

Result, again. . .



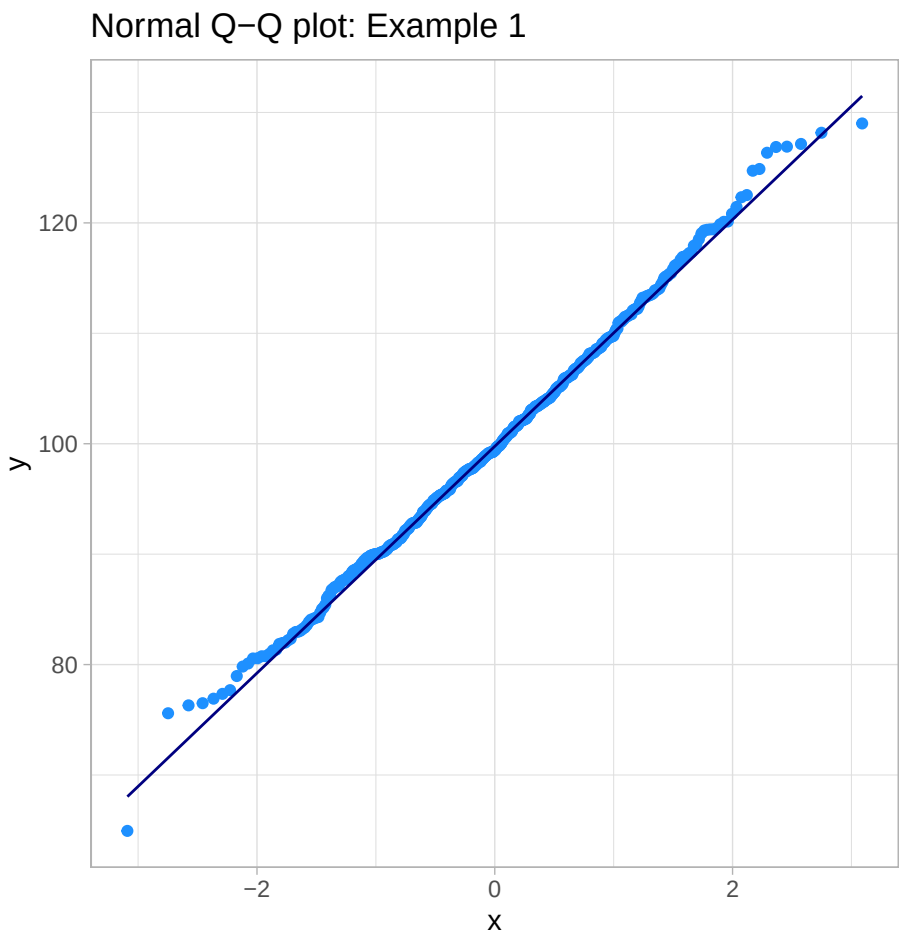
Interpreting the Normal Q-Q plot?

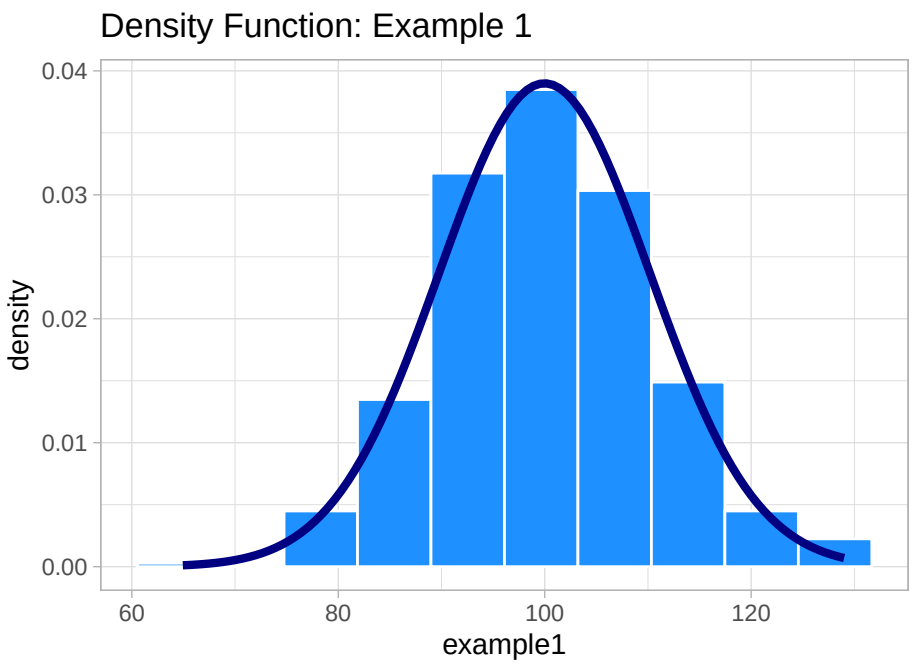
The Normal Q-Q plot can help us identify data as well approximated by a Normal distribution, or not, because of:

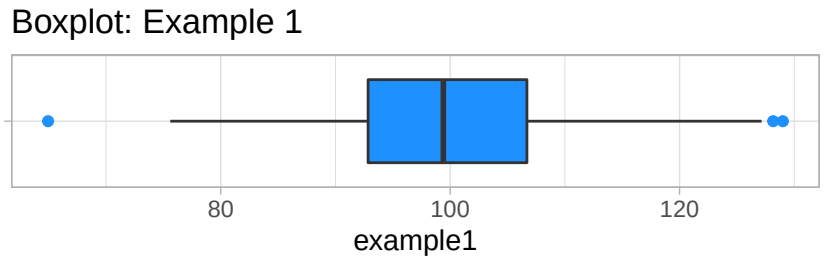
- skew (including distinguishing between right skew and left skew)
- behavior in the tails (which could be heavy-tailed [more outliers than expected] or light-tailed)

1. Normally distributed data are indicated by close adherence of the points to the diagonal reference line.
2. Skew is indicated by substantial curving (on both ends of the distribution) in the points away from the reference line (if both ends curve up, we have right skew; if both ends curve down, this indicates left skew)
3. An abundance or dearth of outliers (as compared to the expectations of a Normal model) are indicated in the tails of the distribution by an "S" shape or reverse "S" shape in the points.

Example 1: Data from a Normal Distribution





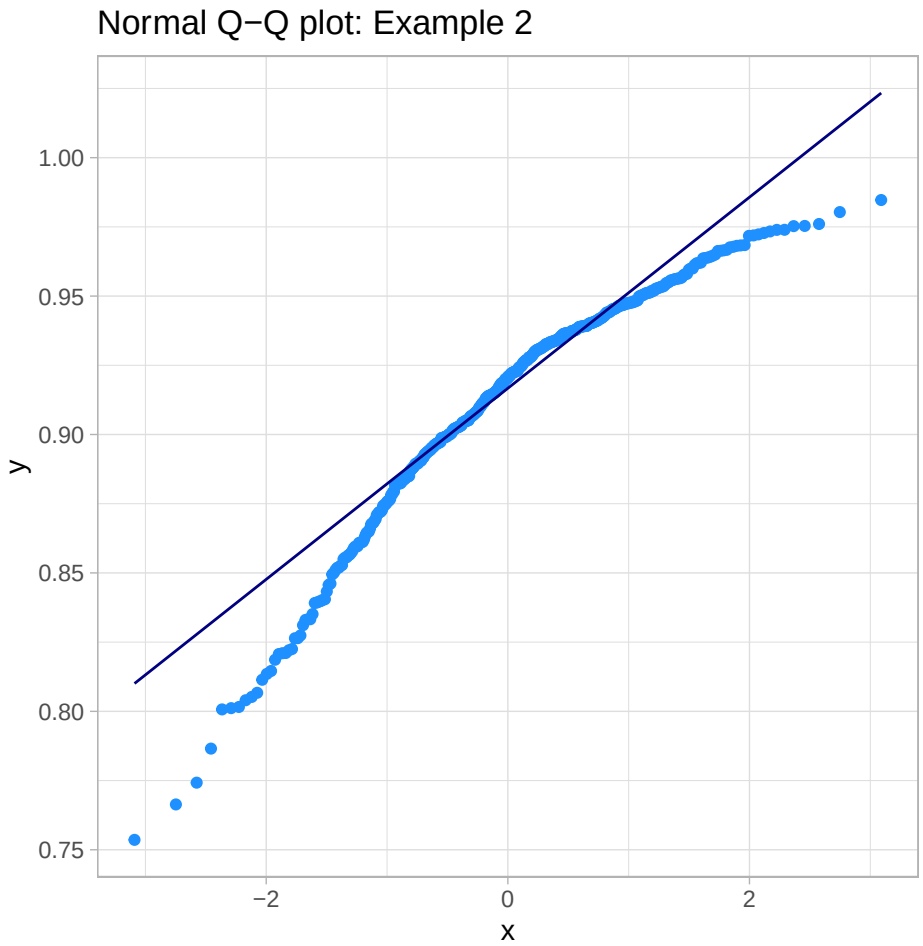


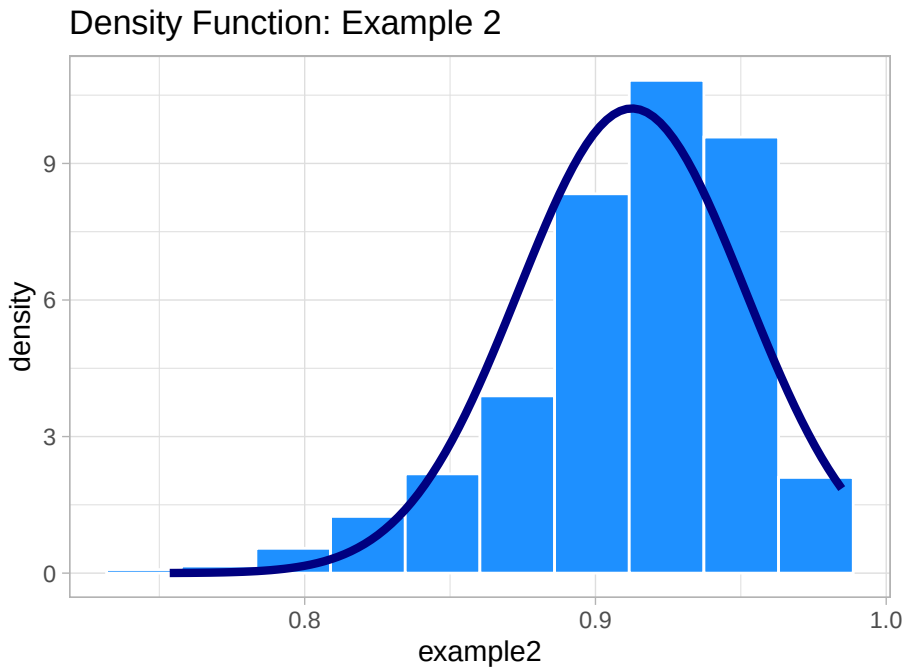
min	Q1	median	Q3	max	mean	sd	n	missing
64.9	92.8	99.4	106.7	129	100	10.2	500	0

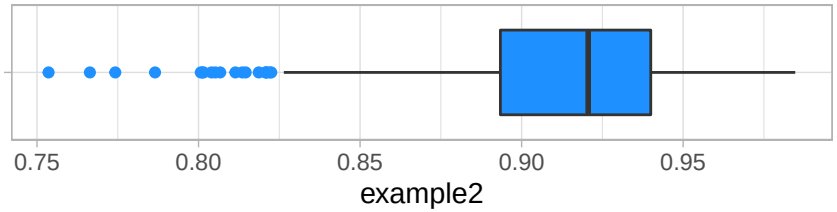
Does a Normal model fit well for my data?

1. Is a Normal Q-Q plot showing something close to a straight line, without clear signs of skew or indications of lots of outliers (heavy-tailedness)?
2. Does a boxplot, violin plot and/or histogram also show a symmetric distribution, where both the number of outliers is modest, and the distance of those outliers from the mean is modest?
3. Do numerical measures match up with the expectations of a normal model?

Example 2: Data from a Left-Skewed Distribution

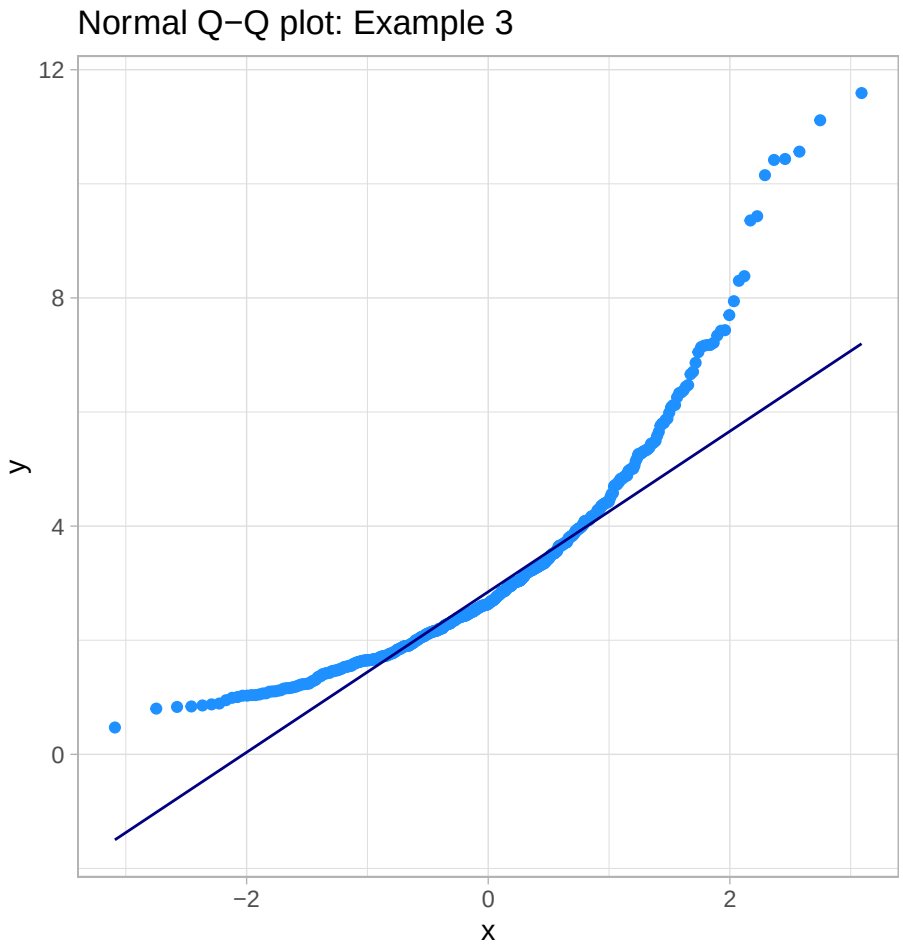


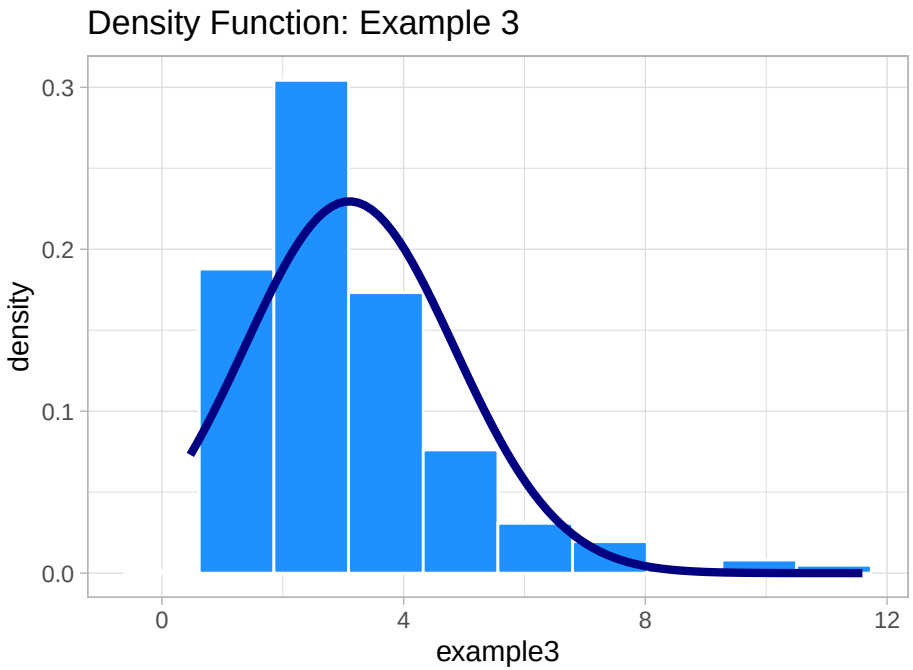


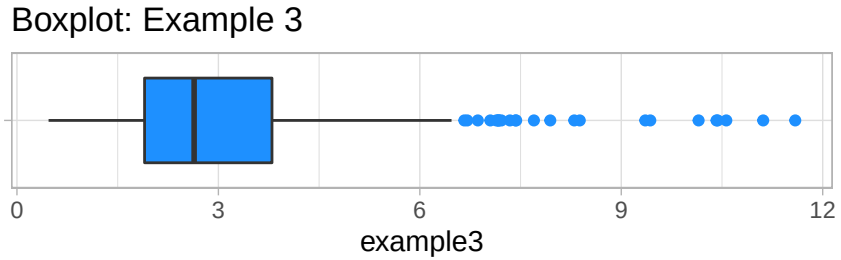


min	Q1	median	Q3	max	mean	sd	n	missing
0.8	0.9	0.9	0.9	1	0.9	0	500	0

Example 3: Data from a Right-Skewed Distribution

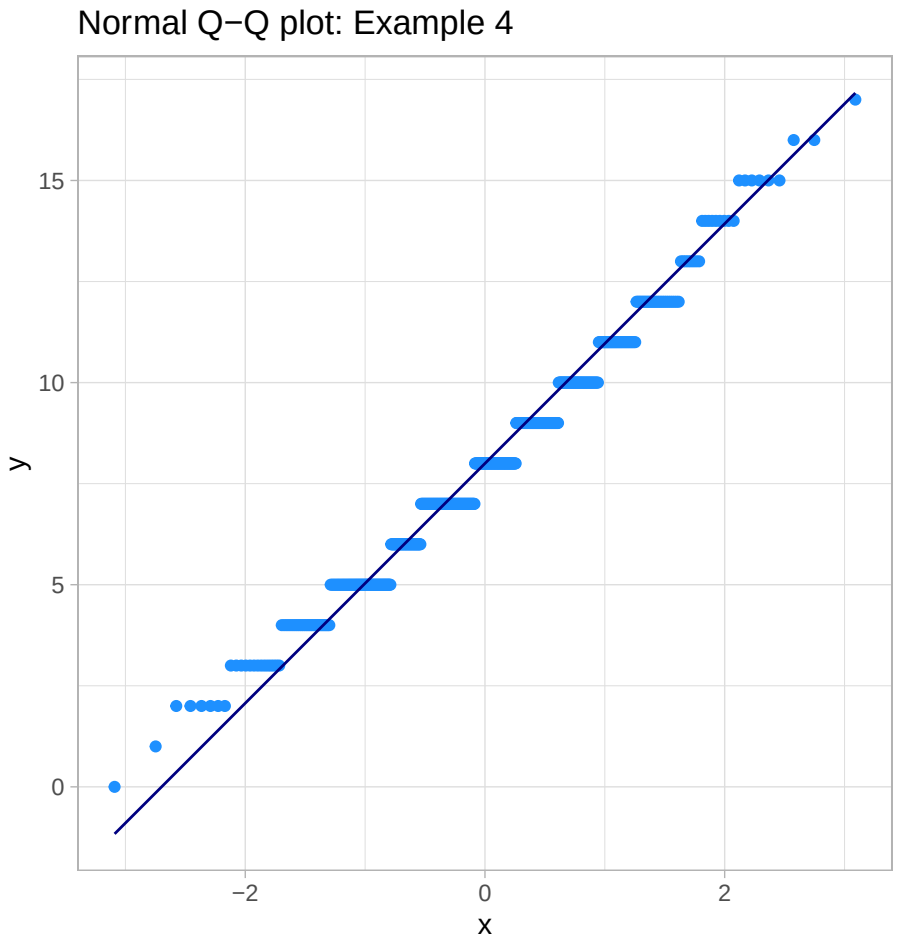


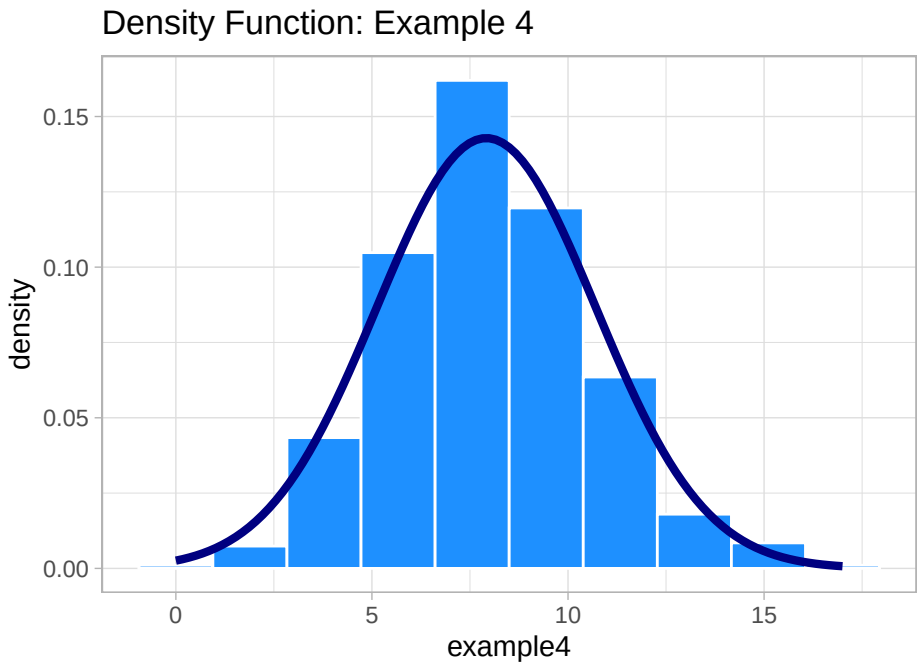


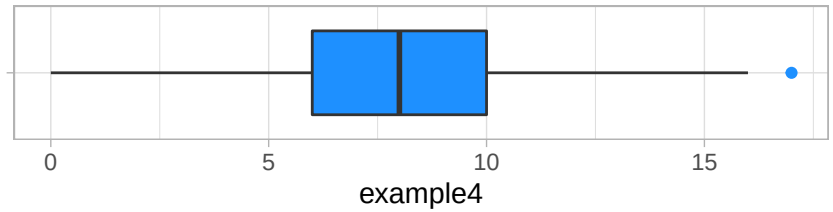


min	Q1	median	Q3	max	mean	sd	n	missing
0.5	1.9	2.6	3.8	11.6	3.1	1.7	500	0

Example 4: Discrete "Symmetric" Distribution

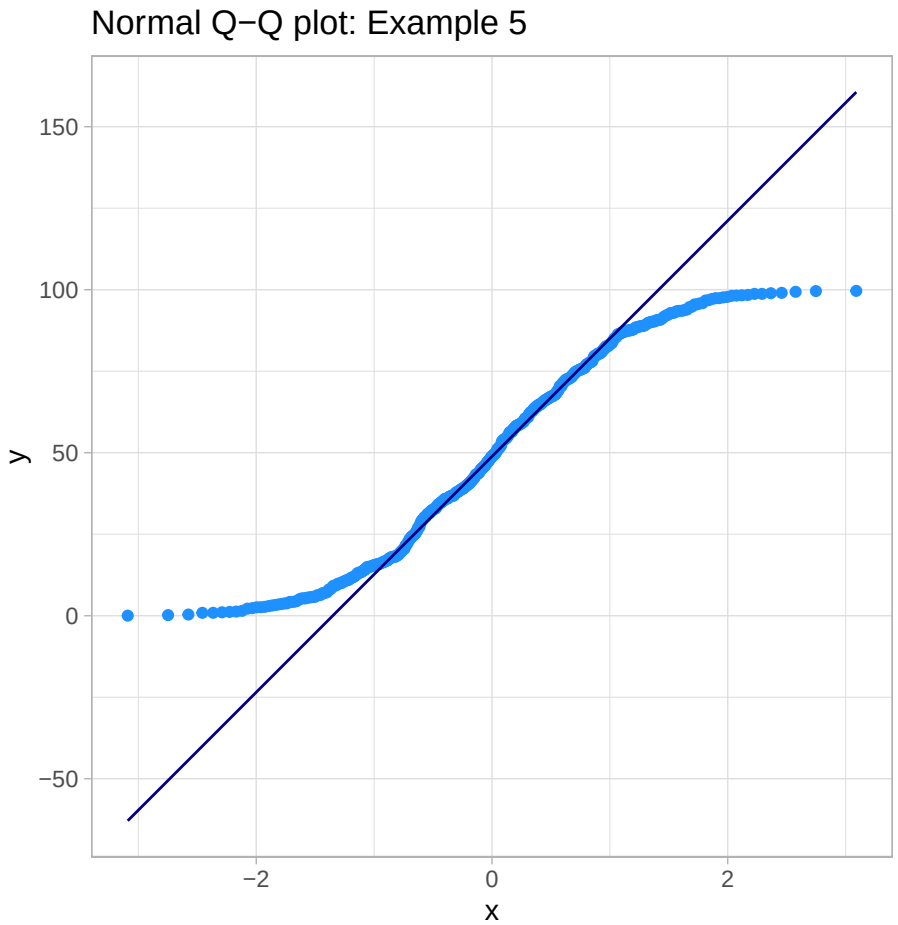


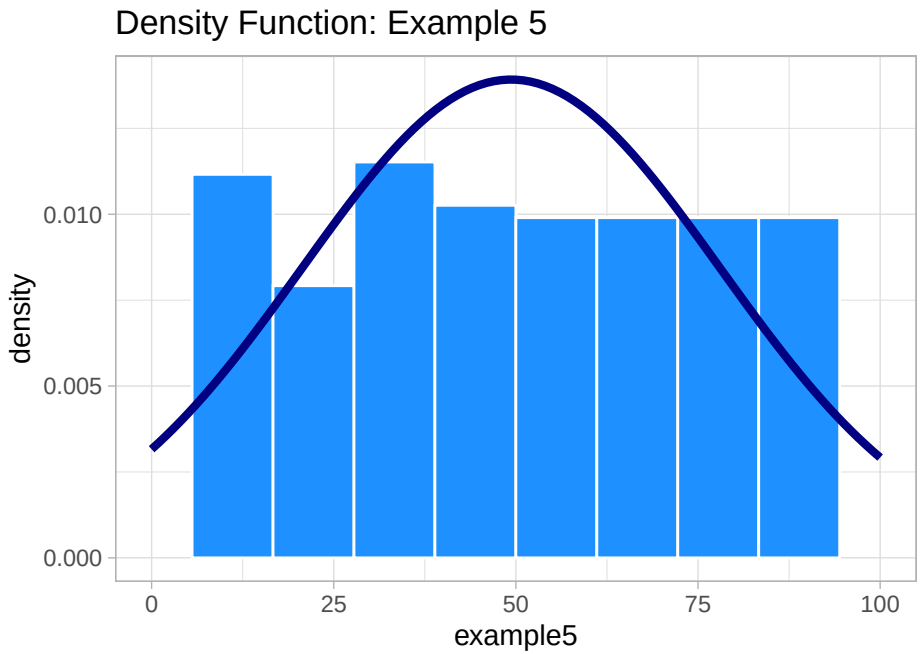


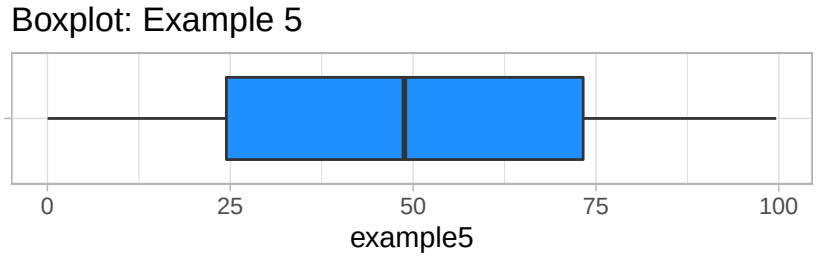


min	Q1	median	Q3	max	mean	sd	n	missing
0	6	8	10	17	7.9	2.8	500	0

Example 5: Data from a Uniform Distribution

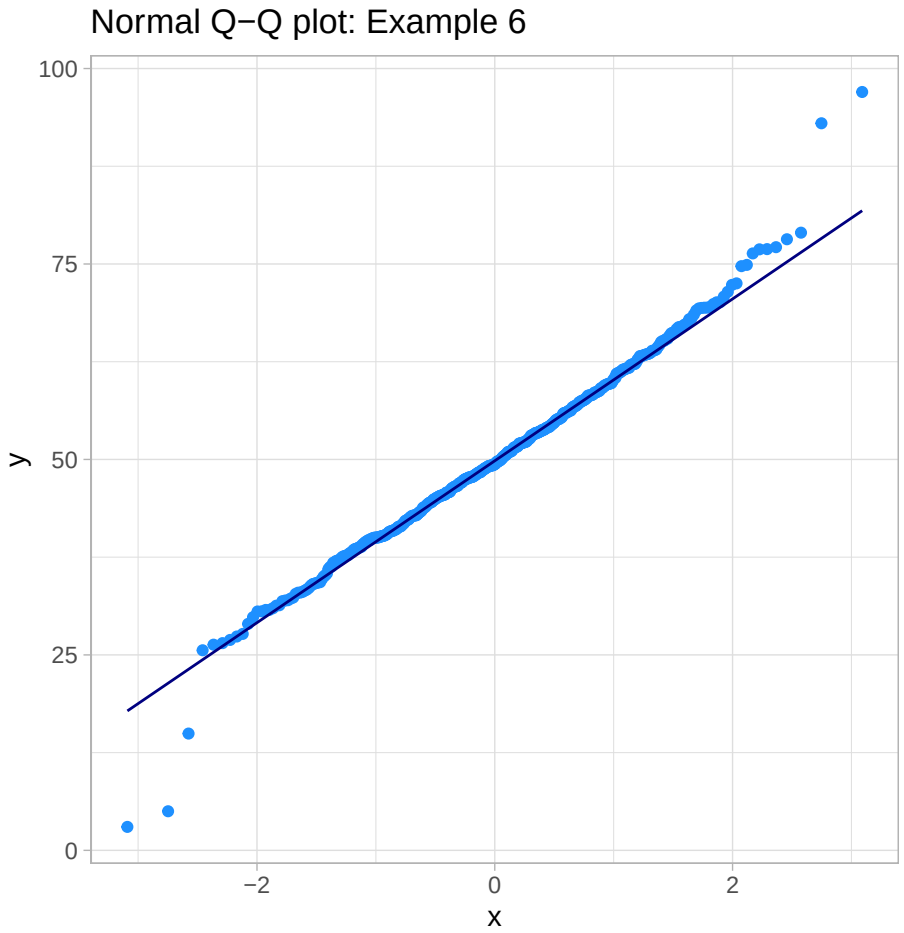


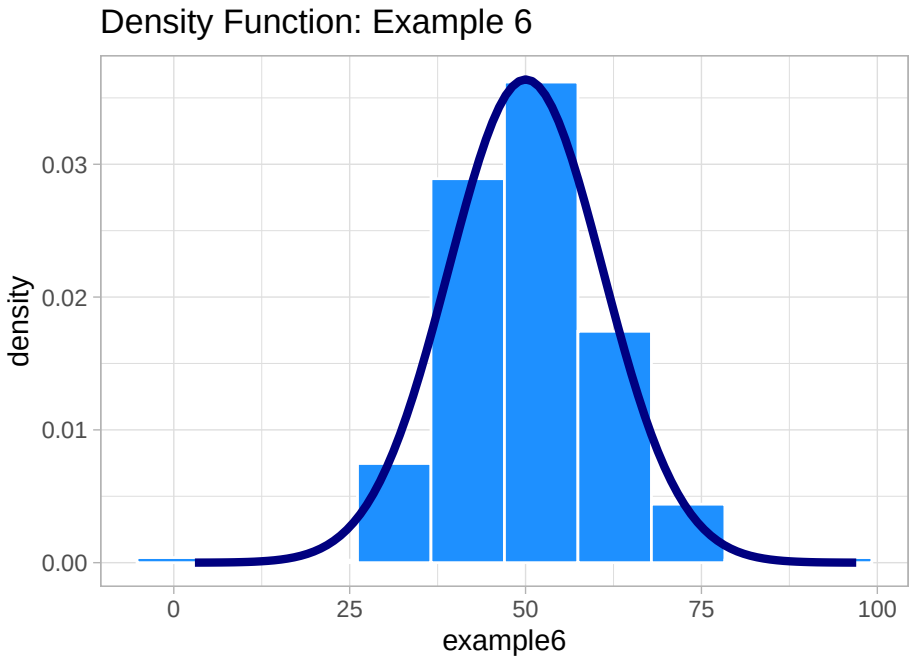


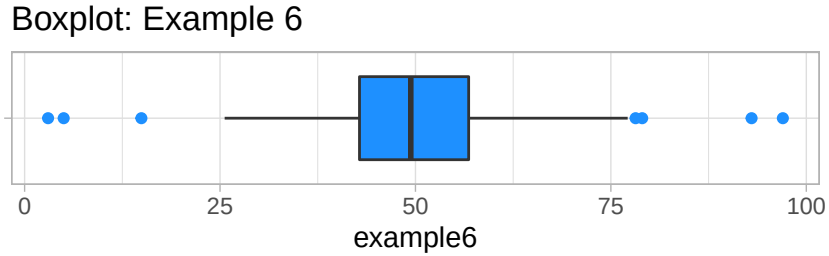


min	Q1	median	Q3	max	mean	sd	n	missing
0	24.5	48.8	73.2	99.6	49.3	28.7	500	0

Example 6: Symmetric data with outliers

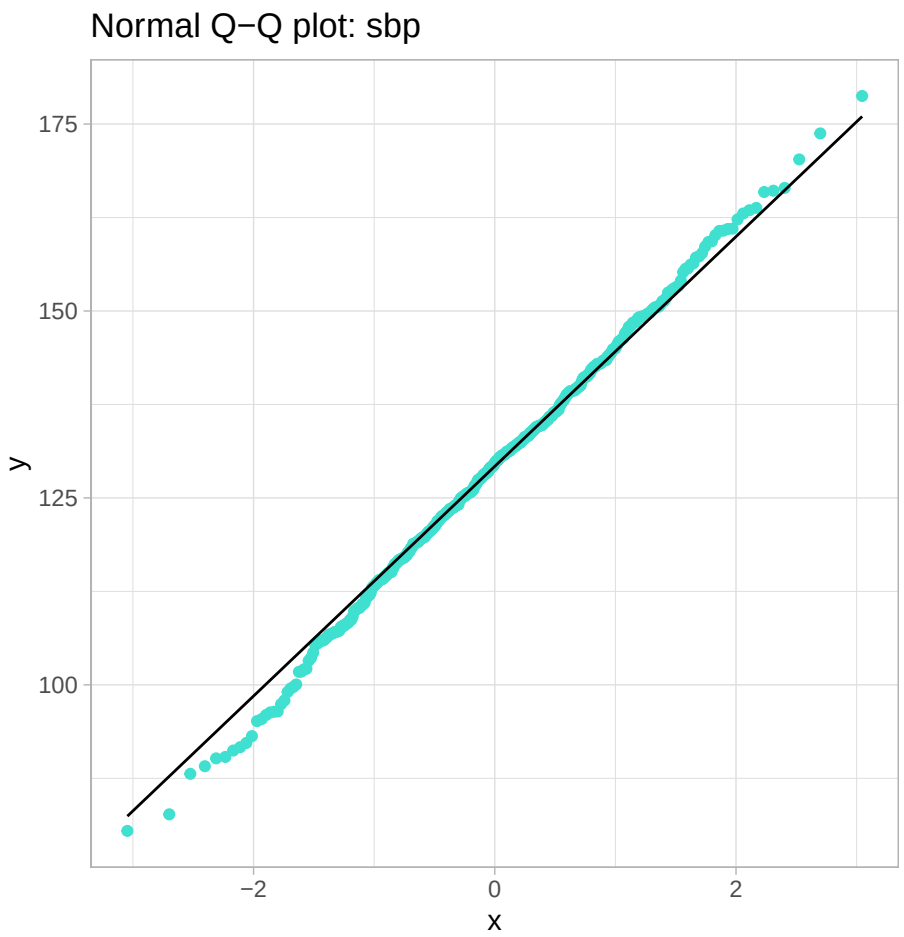


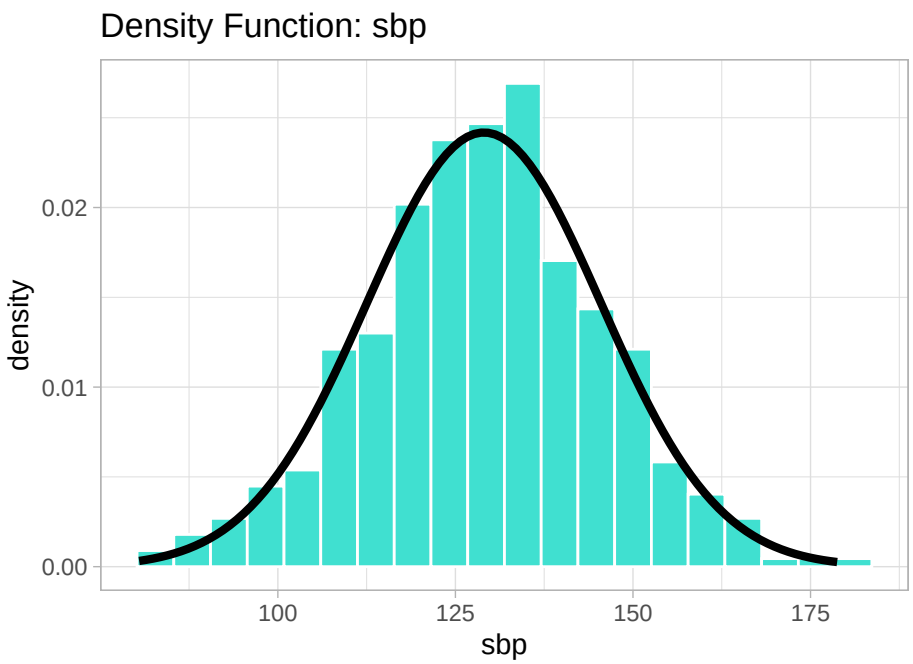


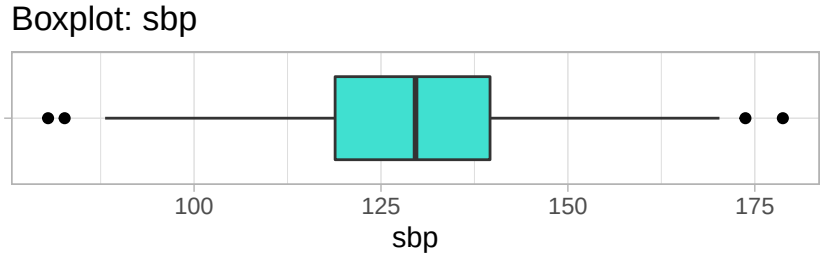


min	Q1	median	Q3	max	mean	sd	n	missing
3	42.8	49.4	56.8	97	50	11	500	0

Our 431 simulated Systolic Blood Pressures

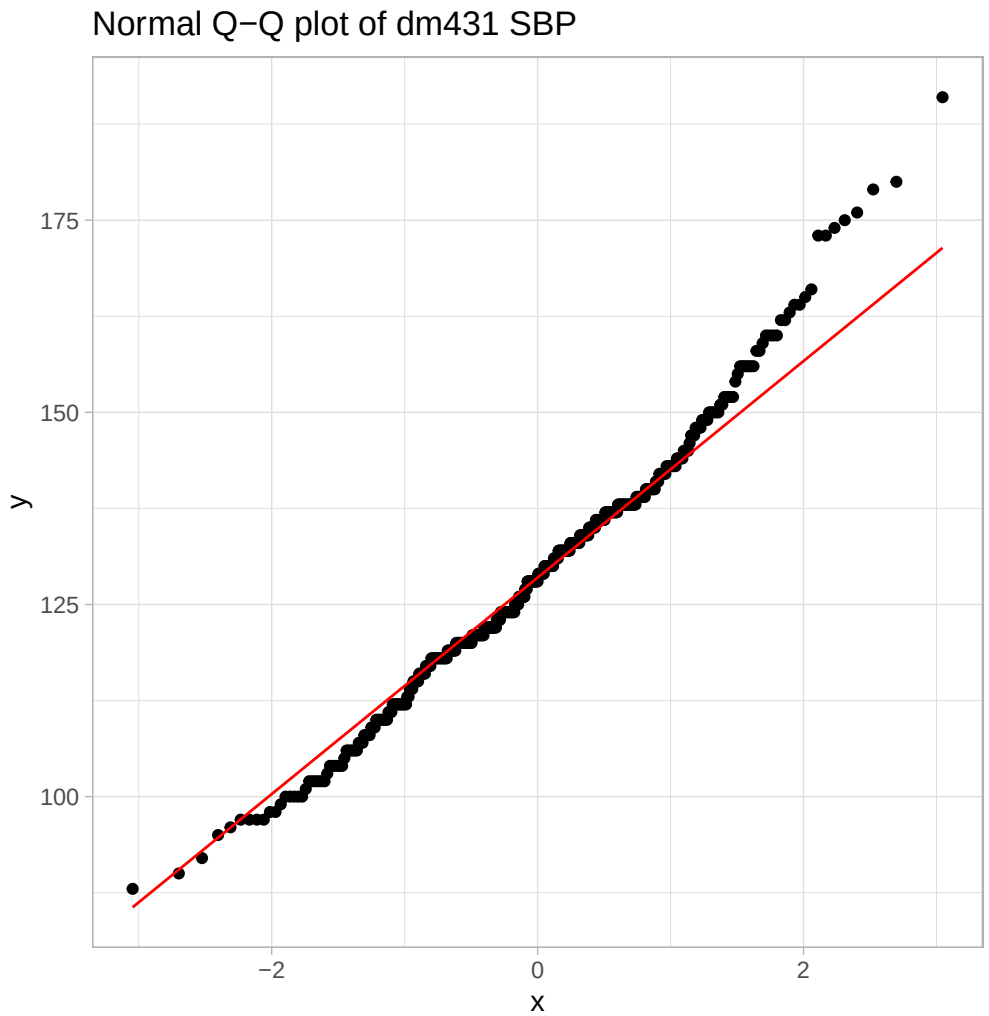




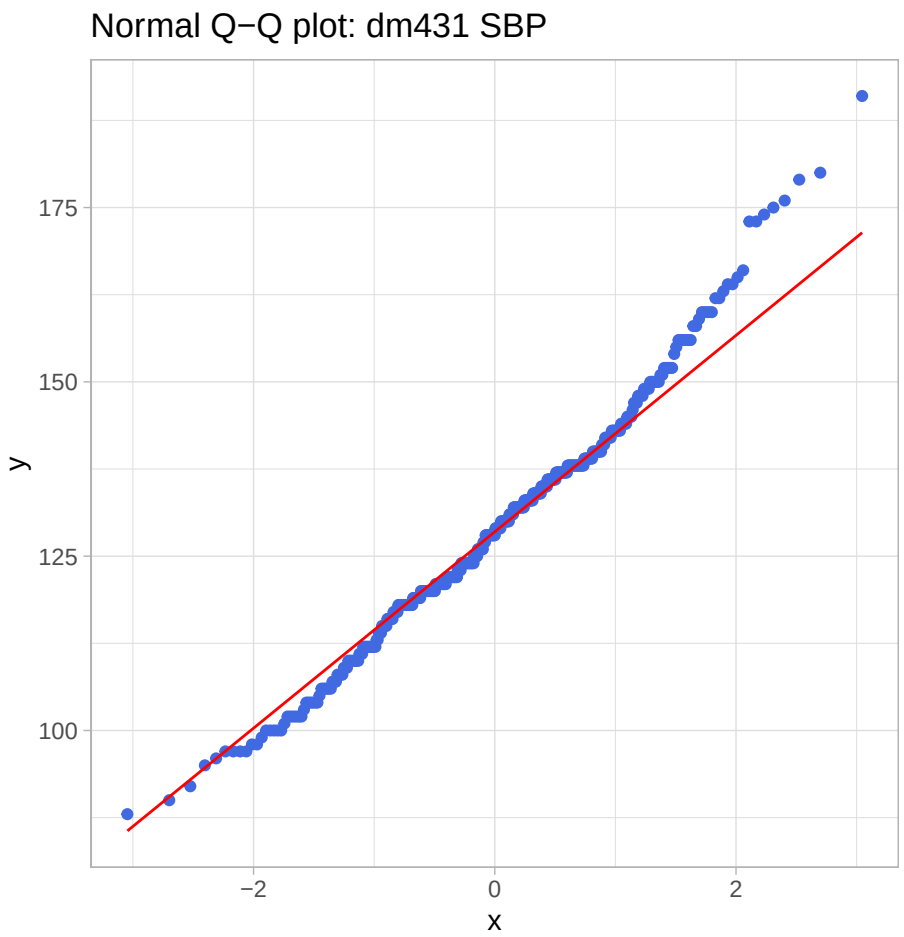


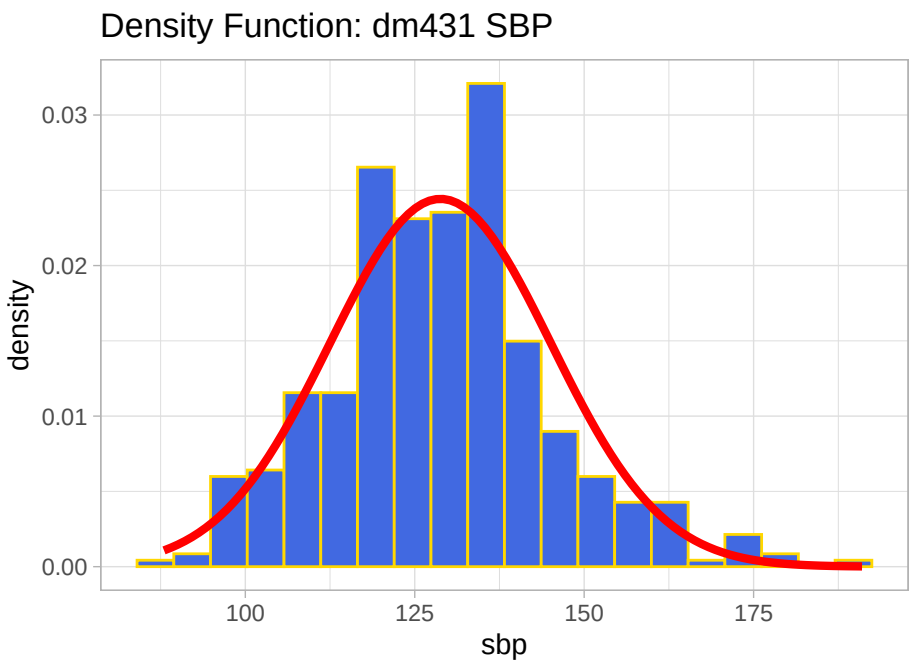
min	Q1	median	Q3	max	mean	sd	n	missing
80.5	118.9	129.6	139.6	178.8	129	16.5	431	0

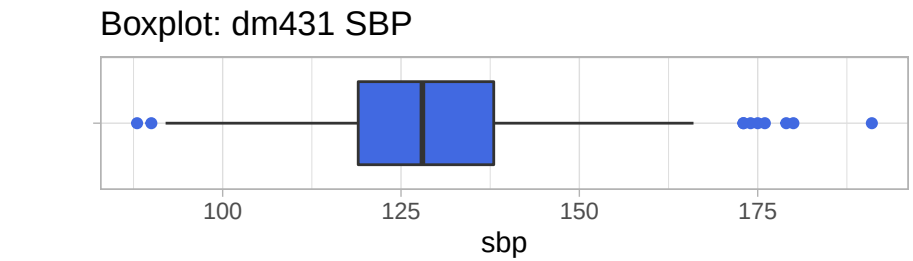
A Normal Q-Q Plot of the dm431 SBP data



How do we build this slide?







min	Q1	median	Q3	max	mean	sd	n	missing
88	119	128	138	191	128.8	16.3	431	0

Code for sbp in dm431 (First of Three Plots)

```
p1 <- ggplot(dm431, aes(sample = sbp)) +
  geom_qq(col = "royalblue") +
  geom_qq_line(col = "red") +
  theme(aspect.ratio = 1) +
  labs(title = "Normal Q-Q plot: dm431 SBP")
```

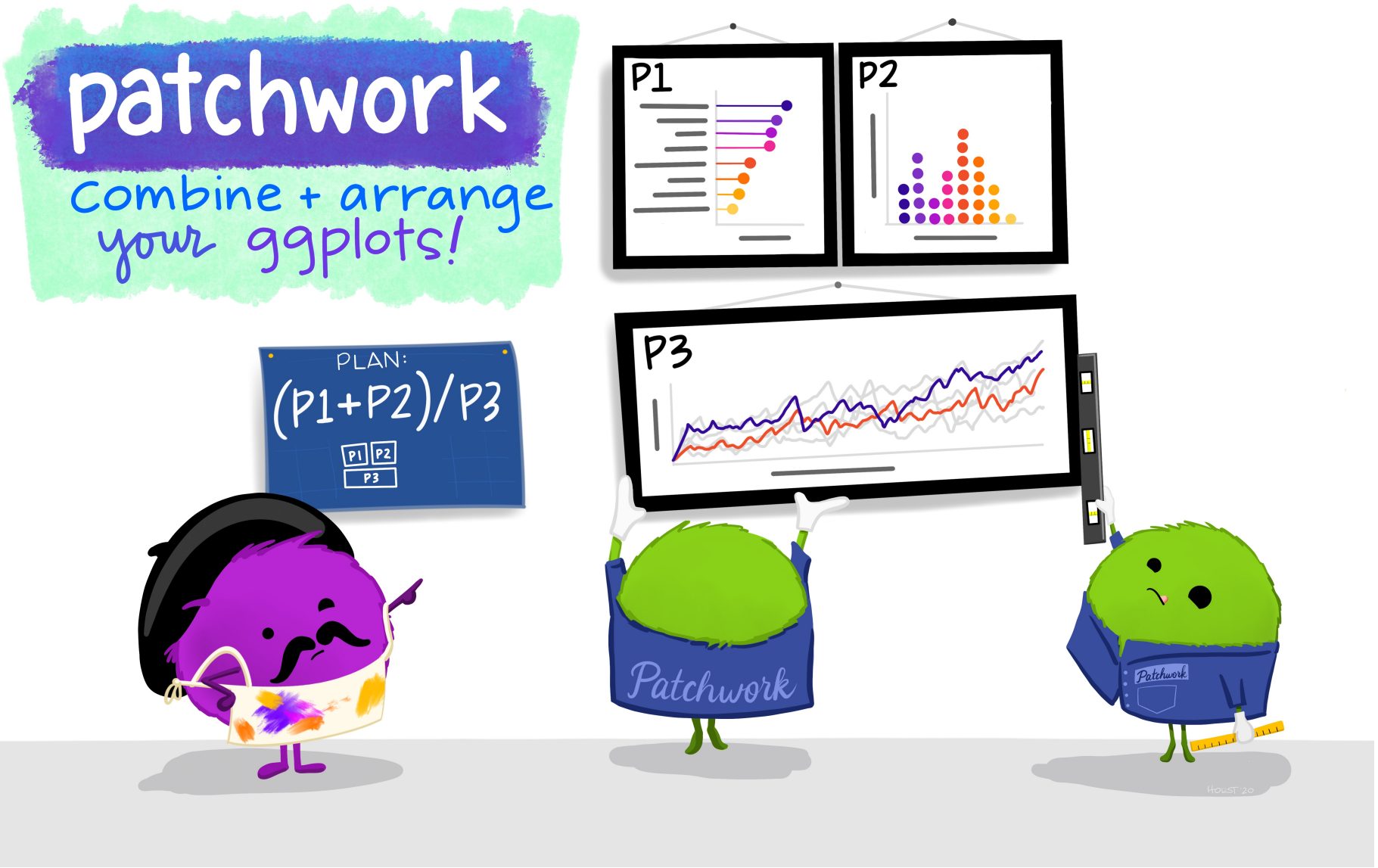
Code for sbp in dm431 (Second of Three Plots)

```
p2 <- ggplot(dm431, aes(x = sbp)) +
  geom_histogram(aes(y = stat(density)),
                 bins = 20,
                 fill = "royalblue", col = "gold") +
  stat_function(fun = dnorm,
                args = list(mean = mean(dm431$sbp),
                            sd = sd(dm431$sbp)),
                col = "red", lwd = 1.5) +
  labs(title = "Density Function: dm431 SBP")
```

Code for `sbp` in `dm431` (Third of Three Plots)

```
p3 <- ggplot(dm431, aes(x = sbp, y = "")) +
  geom_boxplot(fill = "royalblue",
               outlier.color = "royalblue") +
  labs(title = "Boxplot: dm431 SBP", y = "")
```

Putting the plots together...



Using patchwork

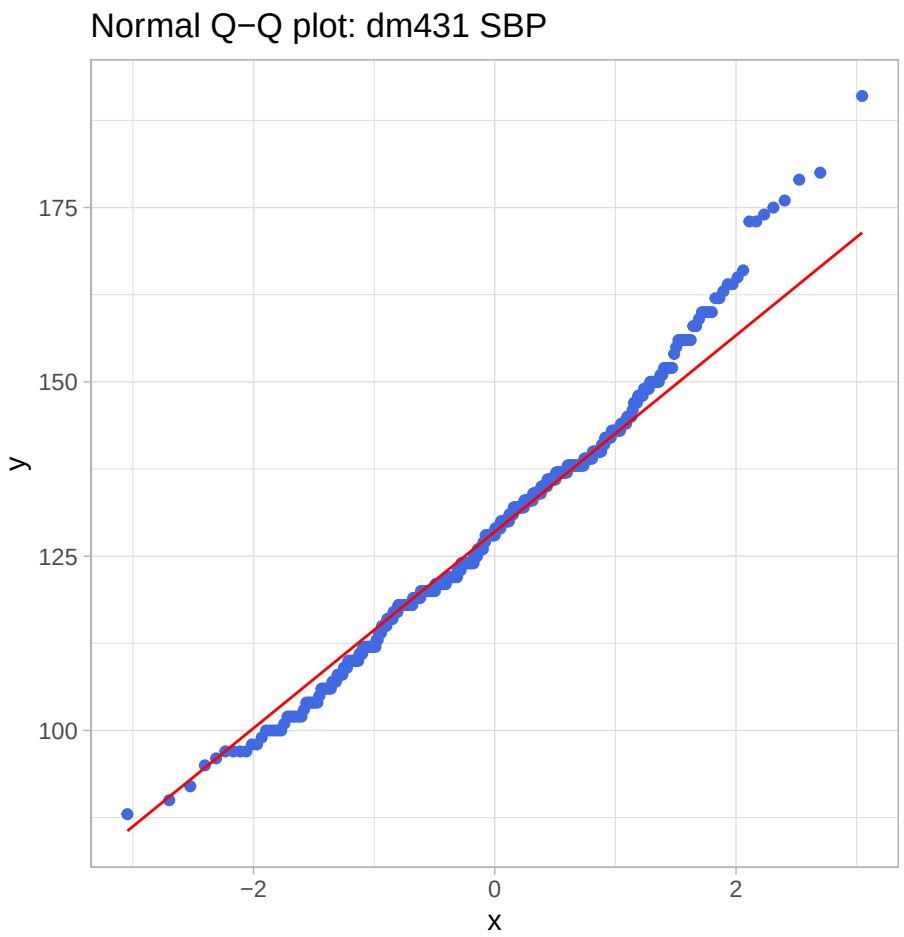
```
p1 + (p2 / p3 + plot_layout(heights = c(4,1)))
```

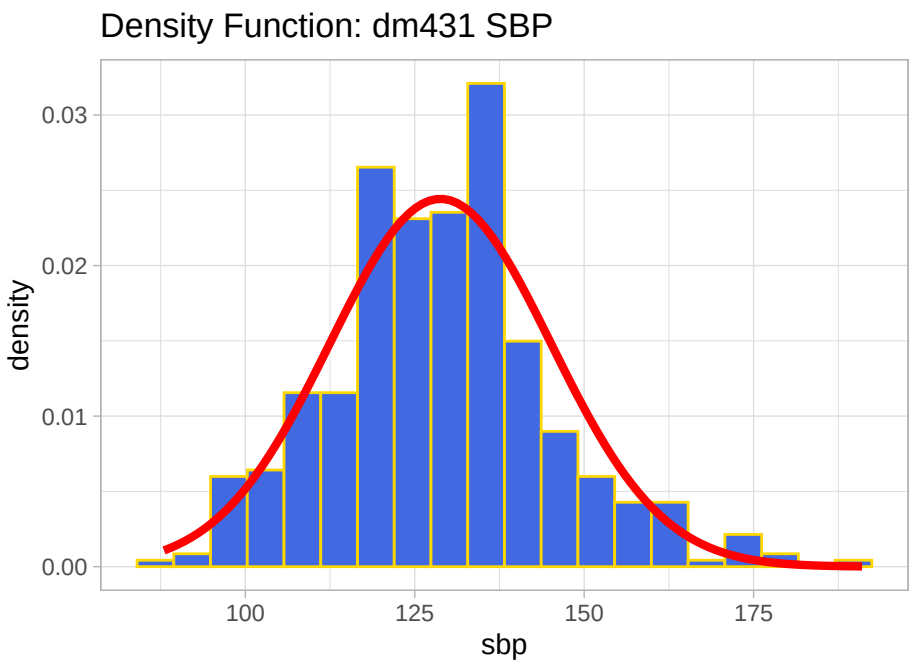
Also added. . .

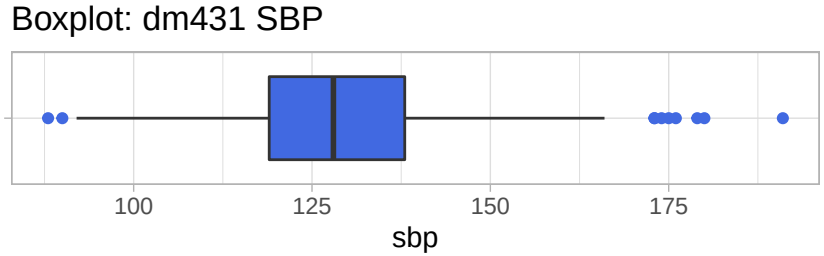
```
mosaic::favstats(~ sbp, data = dm431) %>% kable(digits = 1)
```

min	Q1	median	Q3	max	mean	sd	n	missing
88	119	128	138	191	128.8	16.3	431	0

Result: 431 observed Systolic BP values







min	Q1	median	Q3	max	mean	sd	n	missing
88	119	128	138	191	128.8	16.3	431	0

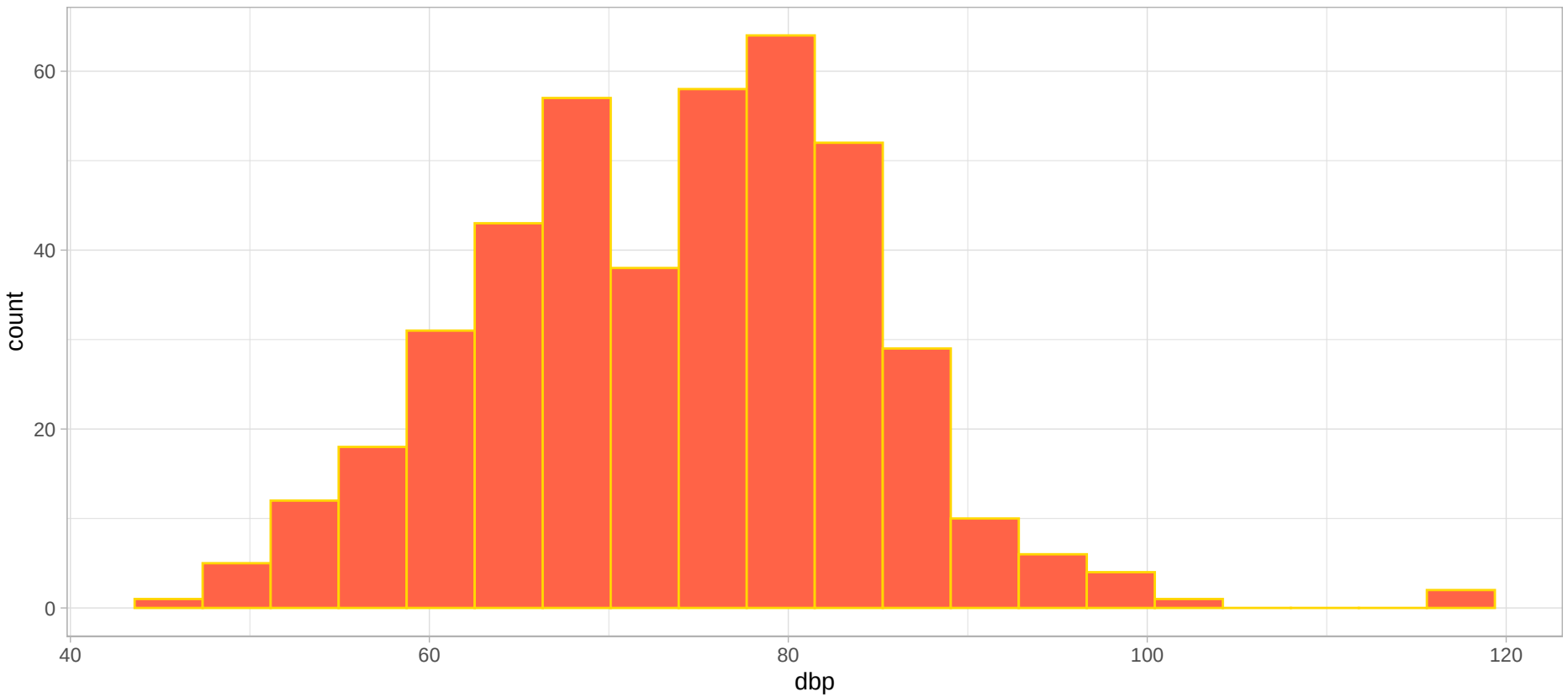
What Summaries to Report

It is usually helpful to focus on the shape, center and spread of a distribution. Bock, Velleman and DeVeaux provide some useful advice:

- If the data are skewed, report the median and IQR (or the three middle quantiles). You may want to include the mean and standard deviation, but you should point out why the mean and median differ. The fact that the mean and median do not agree is a sign that the distribution may be skewed. A histogram will help you make that point.
- If the data are symmetric, report the mean and standard deviation, and possibly the median and IQR as well.
- If there are clear outliers and you are reporting the mean and standard deviation, report them with the outliers present and with the outliers removed. The differences may be revealing. The median and IQR are not likely to be seriously affected by outliers.

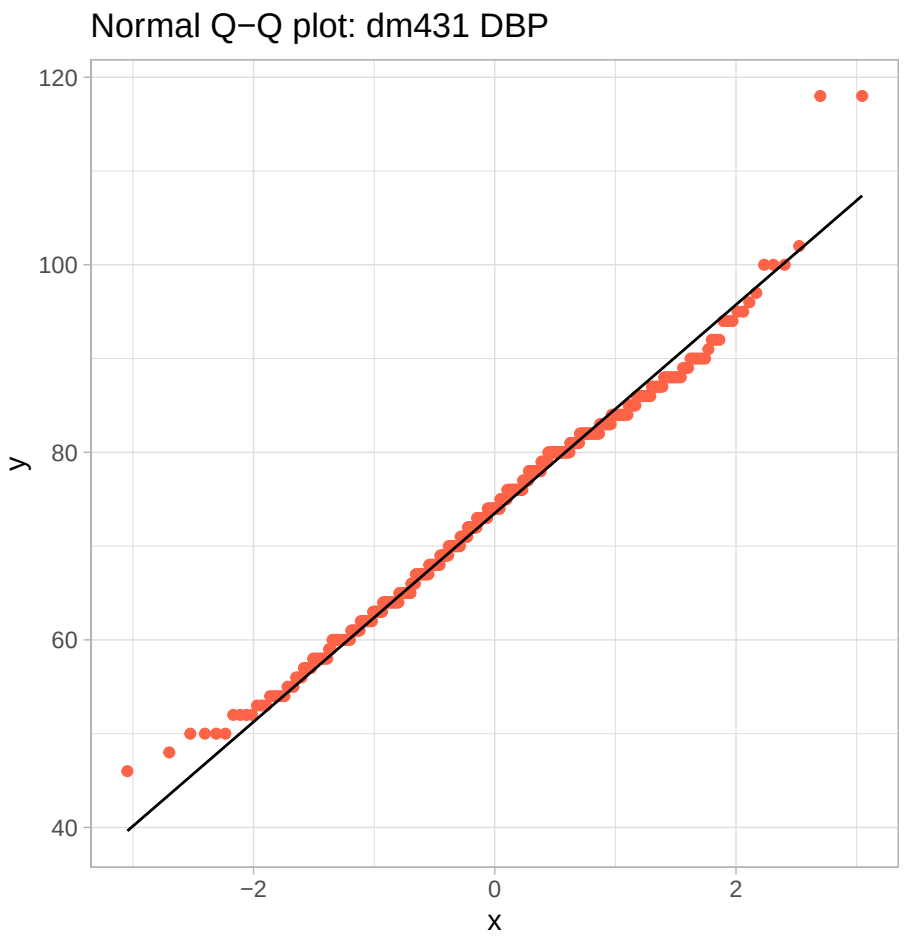
OK, what about Diastolic Blood Pressure?

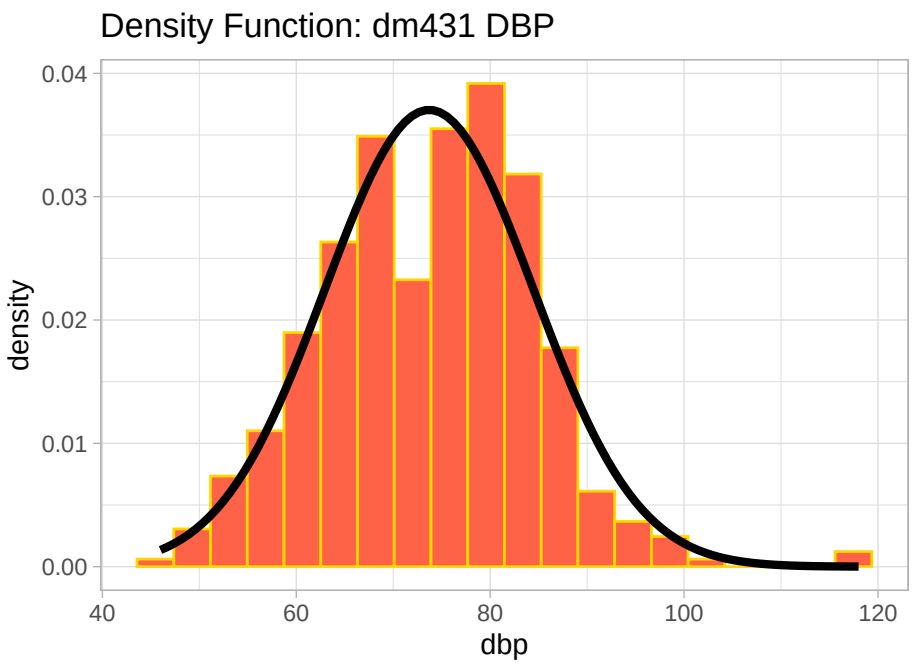
```
ggplot(data = dm431, aes(x = dbp)) +
  geom_histogram(bins = 20, fill = "tomato", col = "gold")
```

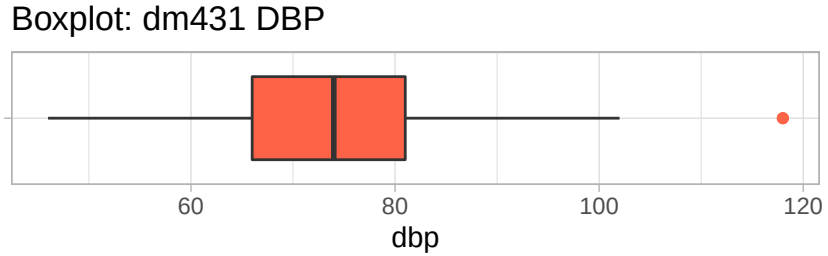


- We can generate the set of plots we've been using. . .

DBP in dm431: Center/Spread/Outliers/Shape?







min	Q1	median	Q3	max	mean	sd	n	missing
46	66	74	81	118	73.7	10.8	431	0

Does a Normal model fit well for my data?

1. Is a Normal Q-Q plot showing something close to a straight line, without clear signs of skew or indications of lots of outliers (heavy-tailedness)?
2. Does a boxplot, violin plot and/or histogram also show a symmetric distribution, where both the number of outliers is modest, and the distance of those outliers from the mean is modest?
3. Do numerical measures match up with the expectations of a normal model?

Hmisc::describe for dbp?

```
dm431 %$% Hmisc::describe(dbp)
```

```
dbp
       n  missing distinct     Info     Mean      Gmd
     431        0       51    0.999    73.71    12.08
     .05      .10      .25      .50      .75      .90
      56       60       66       74       81       86
     .95
      90

lowest :  46  48  50  52  53, highest:  96  97 100 102 118
```

What is a plausible diastolic blood pressure?

Stem-and-Leaf of dbp values?

```
stem(dm431$dbp, scale = 0.6, width = 55)
```

```
  The decimal point is 1 digit(s) to the right of the |

   4 | 68
   5 | 000022223334444555666777788888888899
   6 | 000000000001111111222222222333333333444+58
   7 | 00000000000000011111111122222222222223333+83
   8 | 0000000000000000000000001111111111122222+64
   9 | 00000012224445567
  10 | 0002
  11 | 88
```

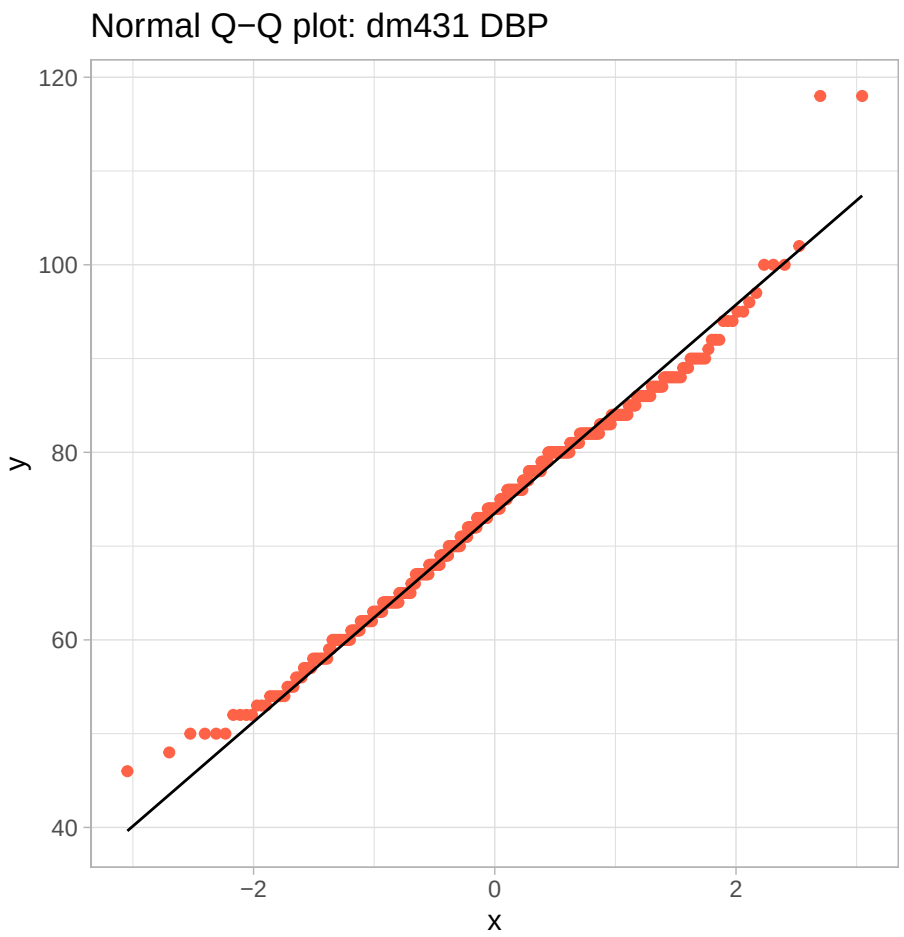
- I've specified `scale` and `width` just for this slide.

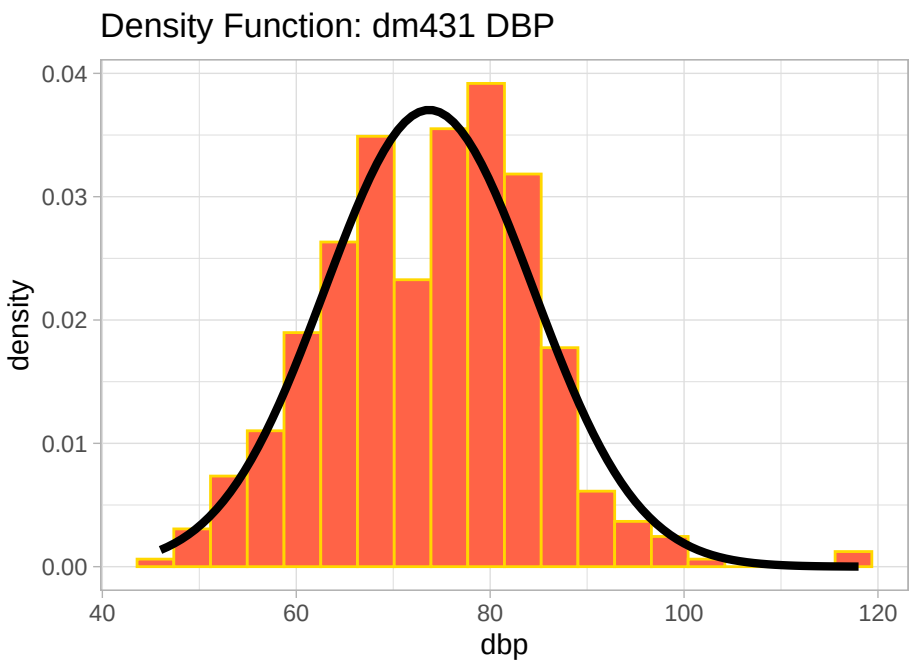
Who are those people with extreme dbp values?

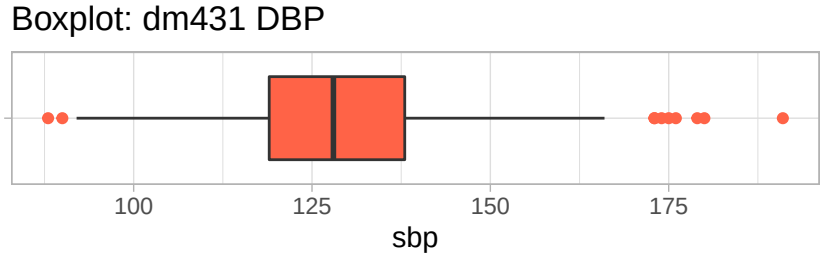
```
dm431 %>%
  filter(dbp < 50 | dbp > 110) %>%
  select(class5_id, sbp, dbp)
```

```
# A tibble: 4 x 3
  class5_id   sbp   dbp
  <chr>     <dbl> <dbl>
1 S-005       156   118
2 S-202       124    46
3 S-219       120    48
4 S-240       158   118
```

dm431: Diastolic Blood Pressure







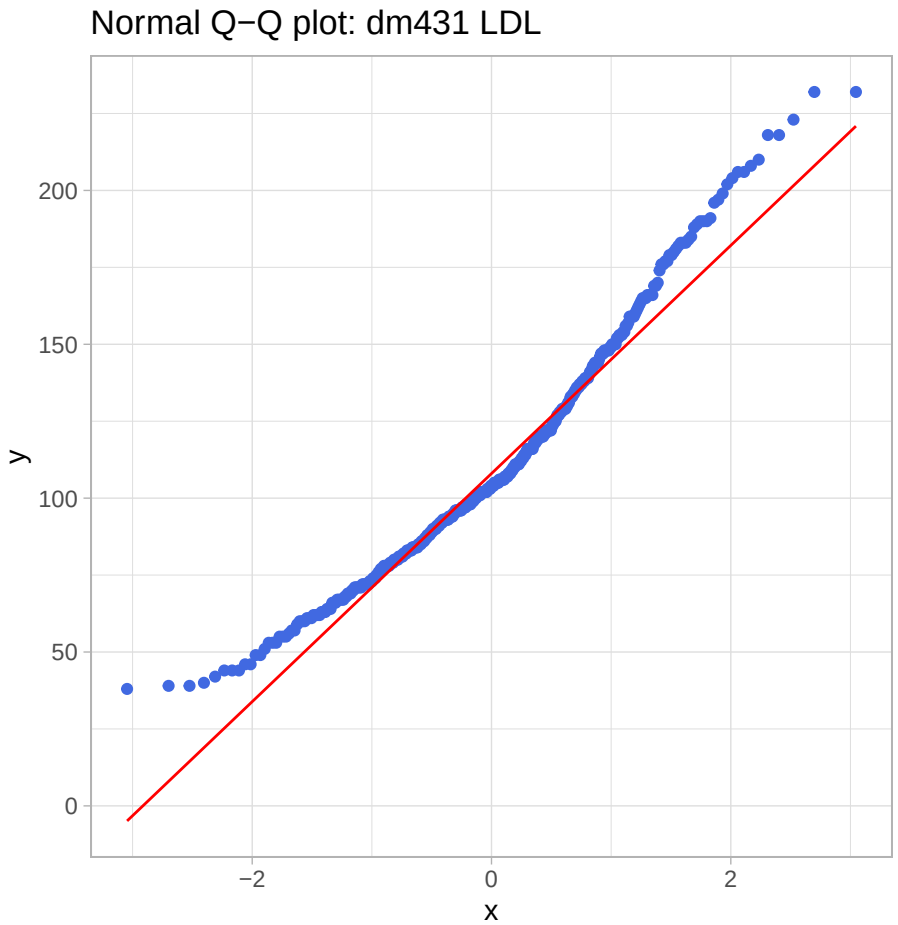
min	Q1	median	Q3	max	mean	sd	n	missing
46	66	74	81	118	73.7	10.8	431	0

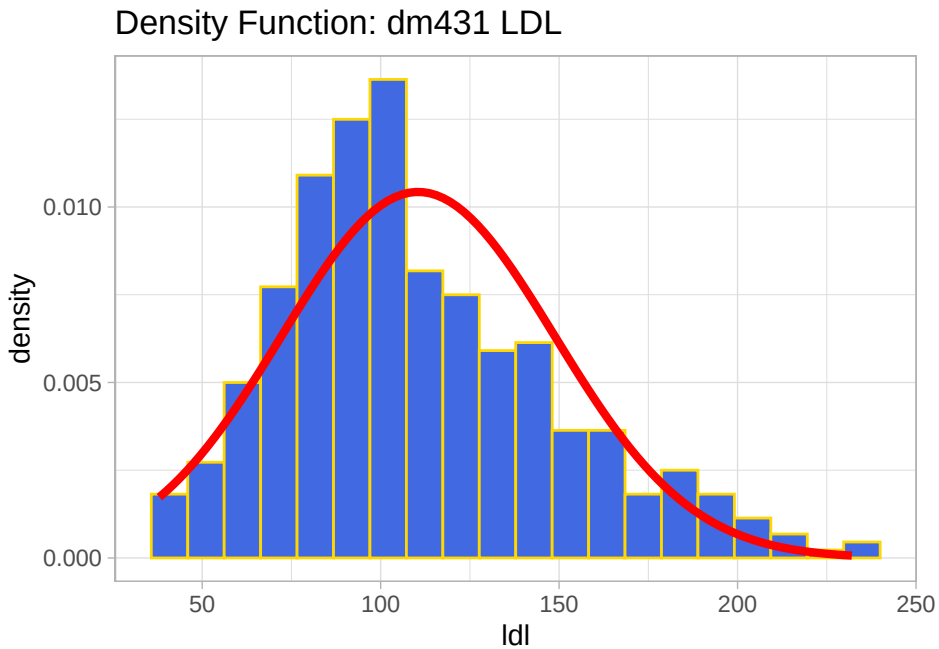
Making Decisions: Does a Normal Model fit well?

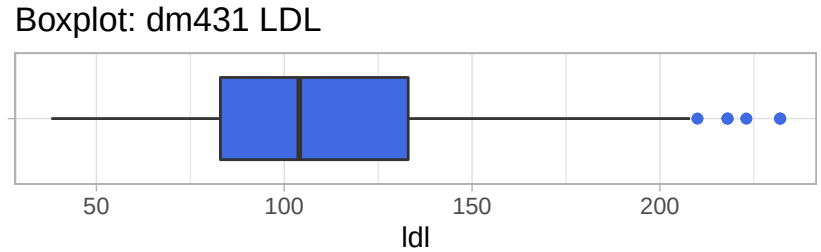
If a Normal model fits our data well, then we should see the following graphical indications:

1. A histogram that is symmetric and bell-shaped.
2. A boxplot where the box is symmetric around the median, as are the whiskers, without a serious outlier problem.
3. A normal Q-Q plot that essentially falls on a straight line.

dm431: LDL Cholesterol

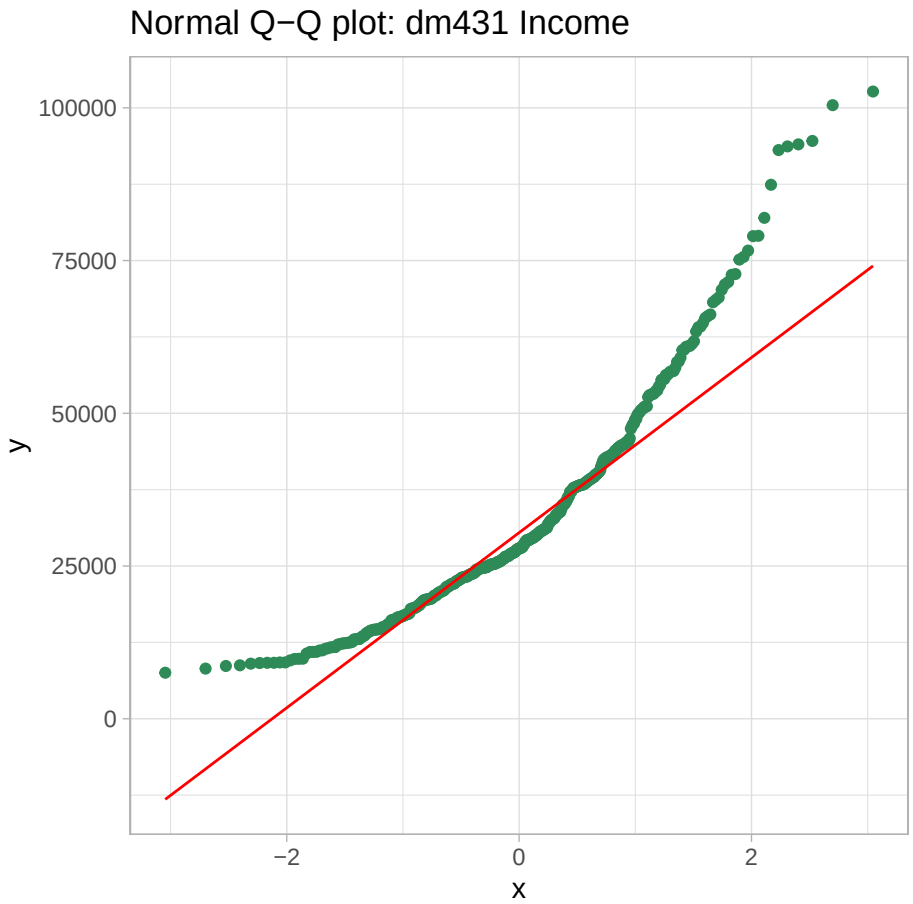


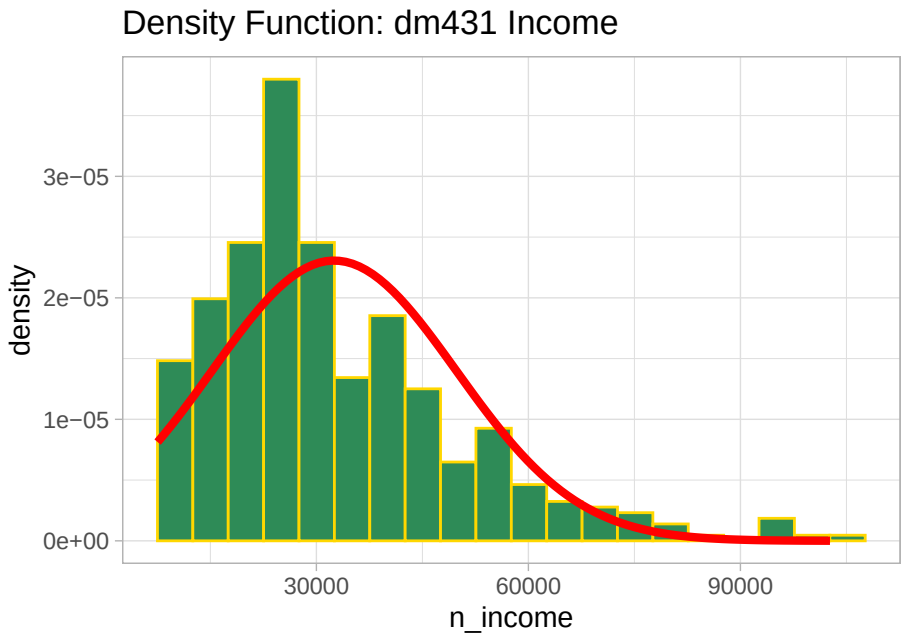


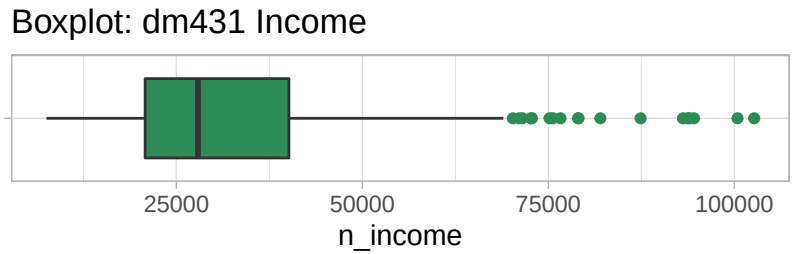


min	Q1	median	Q3	max	mean	sd	n	missing
38	83	104	133	232	110.5	38.3	431	0

dm431: Neighborhood Income

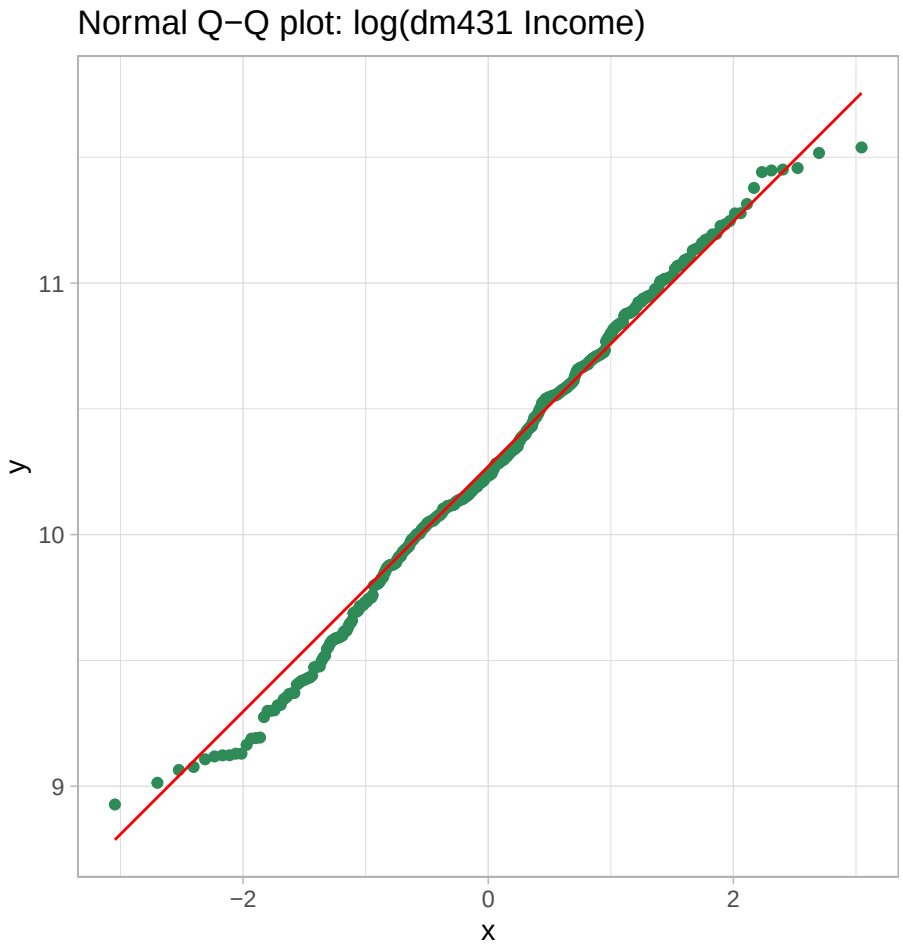


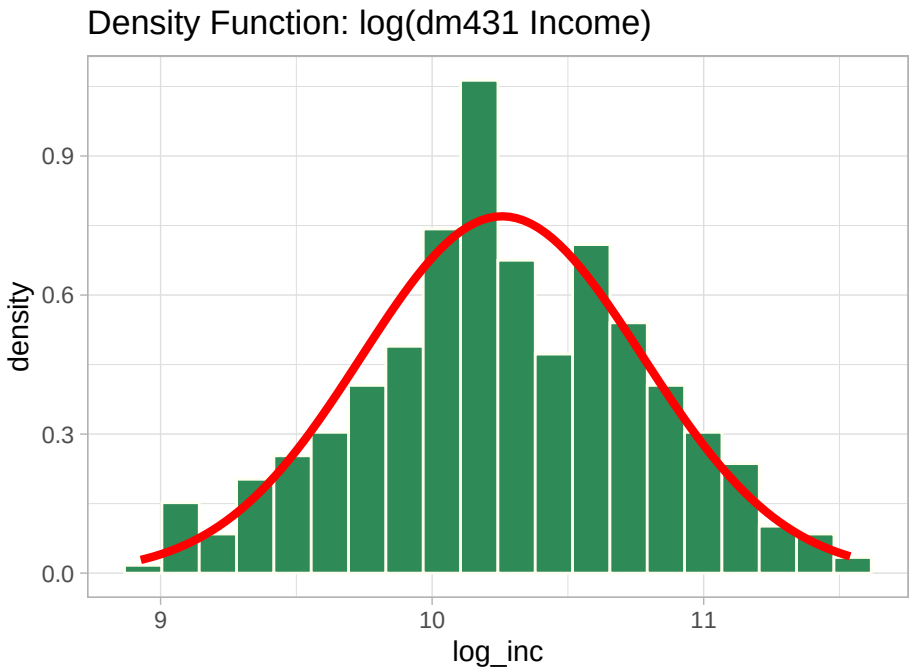


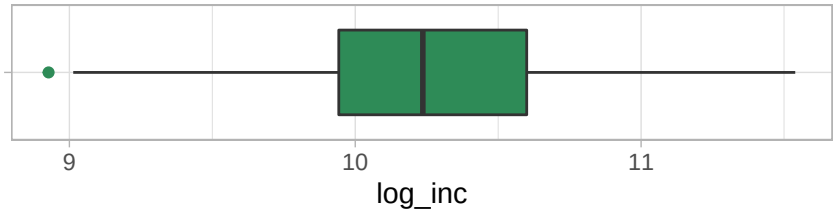


min	Q1	median	Q3	max	mean	sd	n	missing
7534	20794	27903	40128	102672	32514	17295	431	0

dm431: Natural Logarithm of Nbhd. Income

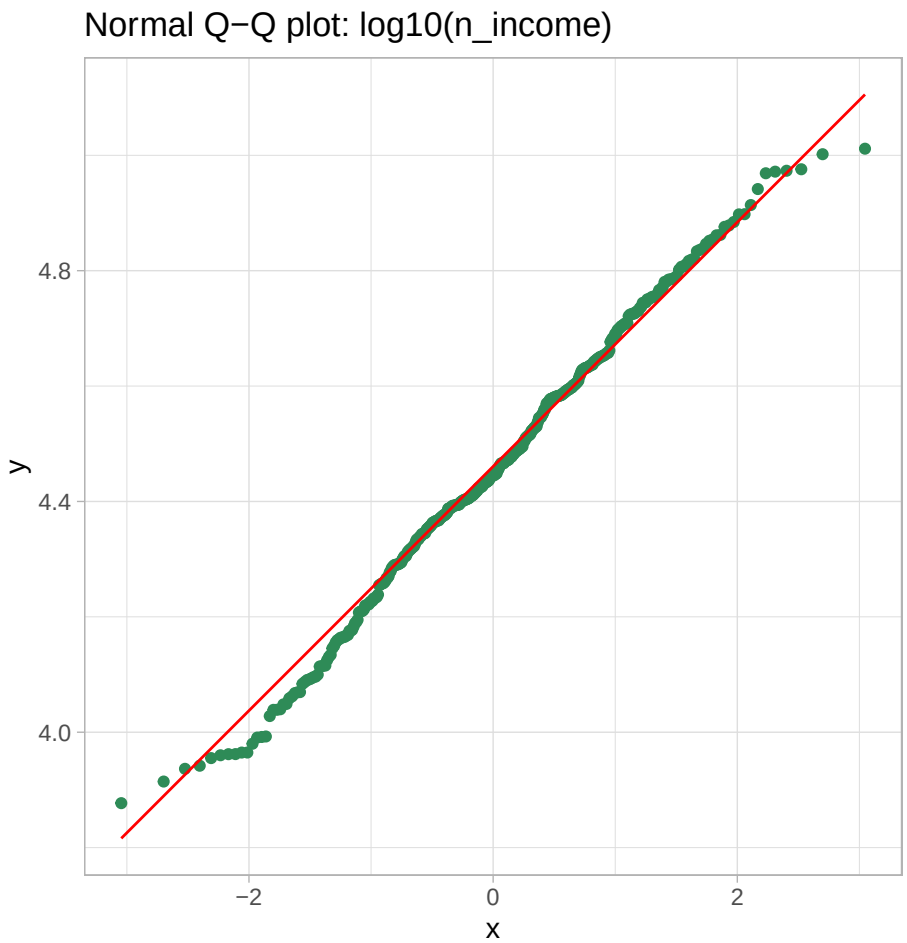


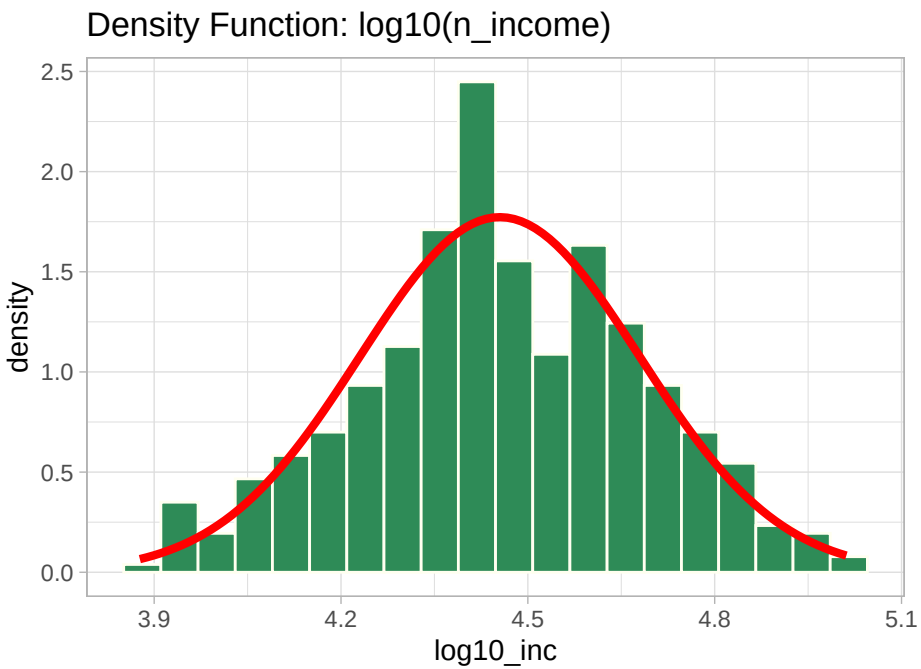




min	Q1	median	Q3	max	mean	sd	n	missing
8.93	9.94	10.24	10.6	11.54	10.26	0.52	431	0

dm431: Base-10 Logarithm of Nbhd. Income





min	Q1	median	Q3	max	mean	sd	n	missing
3.88	4.32	4.45	4.6	5.01	4.45	0.23	431	0

Using Numerical Summaries to Assess Normality: A Good Idea?

Does a Normal model fit well for my data?

The least important approach (even though it is seemingly the most objective) is the calculation of various numerical summaries.

Semi-useful summaries help us understand whether they match up well with the expectations of a normal model:

1. Assessing skewness with $skew_1$ (is the mean close to the median?)
2. Assessing coverage probabilities (do they match the Normal model?)

Quantifying skew with $skew_1$

$$skew_1 = \frac{mean - median}{standard\ \ deviation}$$

Interpreting $skew_1$ (for unimodal data)

- $skew_1 = 0$ if the mean and median are the same
- $skew_1 > 0.2$ indicates fairly substantial right skew
- $skew_1 < -0.2$ indicates fairly substantial left skew

Measuring skewness in the SBP values: dm431?

```
mosaic::favstats(~ sbp, data = dm431)
```

```
 min  Q1 median  Q3 max     mean       sd   n missing
  88 119    128 138 191 128.7889 16.33058 431       0
```

```
dm431 %>%
  summarize(skew1 = (mean(sbp) - median(sbp))/sd(sbp))
```

```
# A tibble: 1 x 1
   skew1
   <dbl>
1 0.0483
```

What does this suggest?

How about our other variables?

```
dm431 %>%
  summarize(
    dbp_skew1 = (mean(dbp) - median(dbp))/sd(dbp),
    ldl_skew1 = (mean(ldl) - median(ldl))/sd(ldl),
    ninc_skew1 = (mean(n_income) - median(n_income))/
                          sd(n_income))
```

```
# A tibble: 1 x 3
  dbp_skew1 ldl_skew1 ninc_skew1
      <dbl>     <dbl>      <dbl>
1   -0.0271     0.170      0.267
```

- How do these results match up with our plots?

Empirical Rule for a Normal Model

If the data followed a Normal distribution, perfectly, then about:

- 68% of the data would fall within 1 standard deviation of the mean
- 95% of the data would fall within 2 standard deviations of the mean
- 99.7% of the data would fall within 3 standard deviations of the mean

Remember that, regardless of the distribution of the data:

- Half of the data will fall below the median, and half above it.
- Half of the data will fall in the Interquartile Range (IQR).

How many SBPs are within 1 SD of the mean?

```
dm431 %>%
  count(sbp > mean(sbp) - sd(sbp),
        sbp < mean(sbp) + sd(sbp)) %>%
  kable()
```

sbp > mean(sbp) - sd(sbp)	sbp < mean(sbp) + sd(sbp)	n
FALSE	TRUE	70
TRUE	FALSE	55
TRUE	TRUE	306

- Note that $306/431 = 0.71$, approximately.
- How does this compare to the expectation under a Normal model?

SBP and the mean ± 2 standard deviations rule?

The total sample size here is 431

```
dm431 %>%
  count(sbp > mean(sbp) - 2*sd(sbp),
        sbp < mean(sbp) + 2*sd(sbp)) %>%
  kable()
```

sbp > mean(sbp) - 2 * sd(sbp)	sbp < mean(sbp) + 2 * sd(sbp)	n
FALSE	TRUE	5
TRUE	FALSE	15
TRUE	TRUE	411

- Note that 411/431 = 0.95, approximately.
- How does this compare to the expectation under a Normal model?

Coverage Probabilities for our other variables

- Here are the observed percentages of our data from `dm431` that fall in each of the limits specified by the Empirical Rule.

Variable	Mean	SD	Mean ± 1 SD	Within 2 SD	Within 3 SD
`sbp`	128.8	16.3	71%	95.4%	99.3%
`dbp`	73.7	10.8	71%	95.8%	99.5%
`ldl`	110.5	38.3	68.4%	95.4%	99.5%
`n_income`	32514	17295	72.2%	95.1%	98.4%

- What does this suggest about the effectiveness of a Normal distribution as a model for each of these variables?

Should we use hypothesis tests to assess Normality?

Hypothesis Testing to assess Normality

Don't. Graphical approaches are **far** better than hypothesis tests. . .

```
dm431 %$% shapiro.test(sbp)
```

```
    Shapiro-Wilk normality test

data:  sbp
W = 0.98636, p-value = 0.0004525
```

The very small p value suggests **against** adopting a Normal model.

Variable	p value from Shapiro test
dbp	9.8×10^{-4}
ldl	0
n_income	0

- Exciting, huh? But not actually useful in any real sense.

Why not test for Normality?

There are multiple hypothesis testing schemes (Kolmogorov-Smirnov, etc.) and each looks for one specific violation of a Normality assumption. None can capture the wide range of issues our brains can envision, and none by itself is great at its job.

- With any sort of reasonable sample size, the test is so poor at detecting non-normality compared to our eyes, that it finds problems we don't care about (this is the main problem) and (also, sometimes) ignores problems we do care about.
- And without a reasonable sample size, the test is essentially useless.

Whenever you *can* avoid hypothesis testing and instead actually plot the data, you should plot the data. Sometimes you can't plot (especially with really big data) but the test should be your very last resort.

Summing Up: Does a Normal Model fit well?

If a Normal model fits our data well, then we should see the following graphical indications:

1. A histogram that is symmetric and bell-shaped.
2. A boxplot where the box is symmetric around the median, as are the whiskers, without a serious outlier problem.
3. A normal Q-Q plot that essentially falls on a straight line.

As for numerical summaries, we'd like to see

4. The mean and median within 0.2 standard deviation of each other.
5. No real evidence of too many outlier candidates (more than 5% starts to get us concerned about a Normal model)
6. No real evidence of individual outliers outside the reasonable range for the size of our data (we might expect about 3 observations in 1000 to fall more than 3 standard deviations away from the mean.)

Next Time

Building Confidence Intervals and making comparisons