Exploring numerical variables

Histograms

In [1]:

```
# Setup
%matplotlib inline
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib.ticker as ticker

sns.set_style("white")
# Custom colors
blue = "#3F83F4"
blue_dark = "#062089"
blue_light = "#8DC0F6"
blue_lighter = "#BBE4FA"
grey = "#9C9C9C"
grey_dark = "#777777"
grey_light = "#B2B2B2"
orange = "#EF8733"
colors_blue = [blue, blue_light]
```

Import data

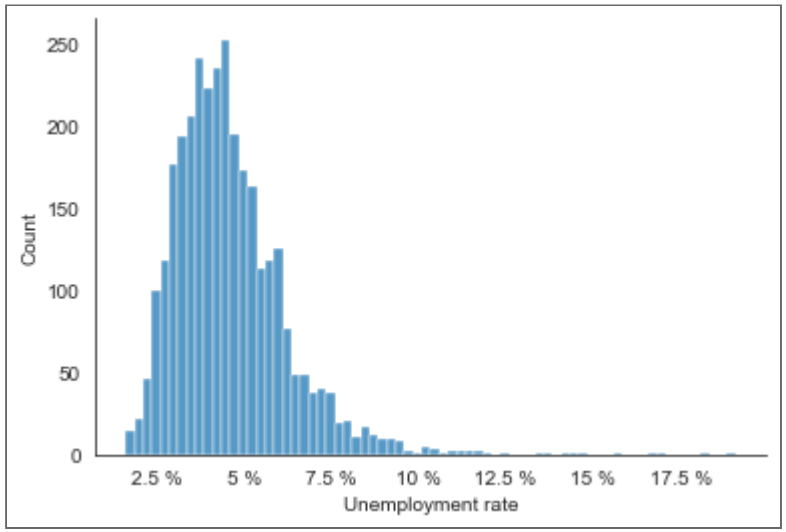
In [2]:

```
ROOT = "https://raw.githubusercontent.com/kirenz/modern-statistics/main/data/"
DATA = "county.csv"

df = pd.read_csv(ROOT + DATA)
```

Histogram

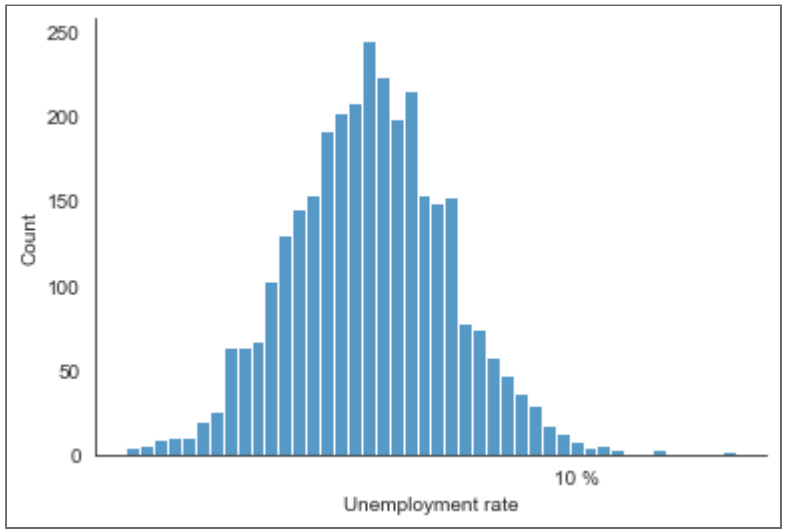
In [3]:

```
# A histogram of the percentage of unemployed in all US counties.
fig, ax = plt.subplots()

sns.histplot(data=df, x= "unemployment_rate", palette=colors_blue)

ax.xaxis.set_major_formatter(ticker.EngFormatter('%'))
plt.xlabel("Unemployment rate")
sns.despine();
```



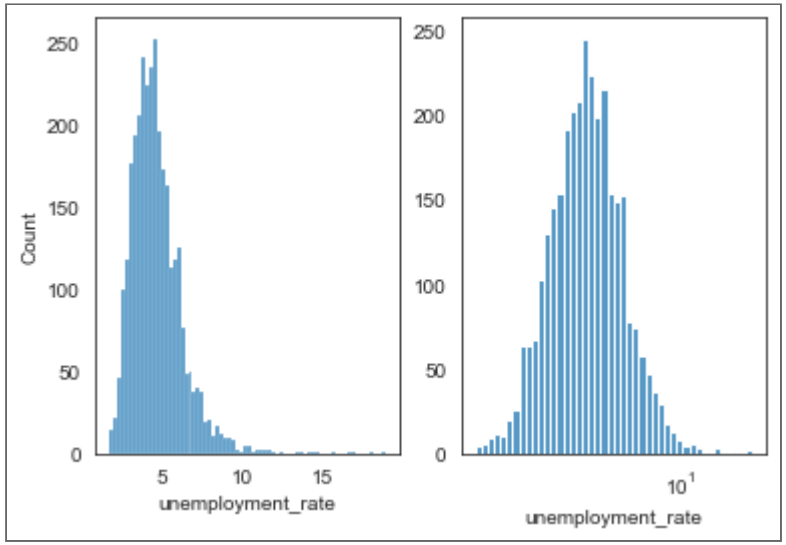
In [4]:

```
# A histogram of log10-transformed unemployed percentages
fig, ax = plt.subplots()

sns.histplot(data=df, x= "unemployment_rate", palette=colors_blue, log_scale=True)

ax.xaxis.set_major_formatter(ticker.EngFormatter('%'))
plt.xlabel("Unemployment rate")
sns.despine();
```



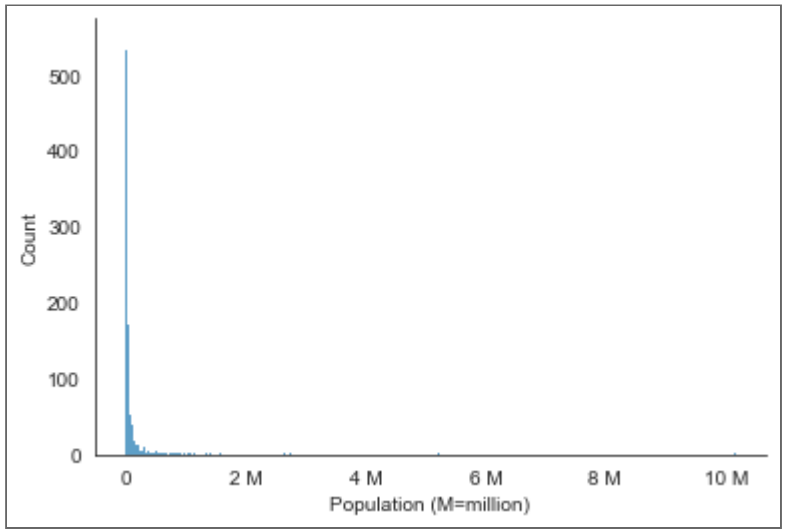
In [5]:

```
# Both plots
fig, ax =plt.subplots(1,2)
sns.histplot(data=df,  x= "unemployment_rate", palette=colors_blue, ax=ax[0])
sns.histplot(data=df,  x= "unemployment_rate", palette=colors_blue, log_scale=True, ax=ax[1]);

plt.ylabel("");
```



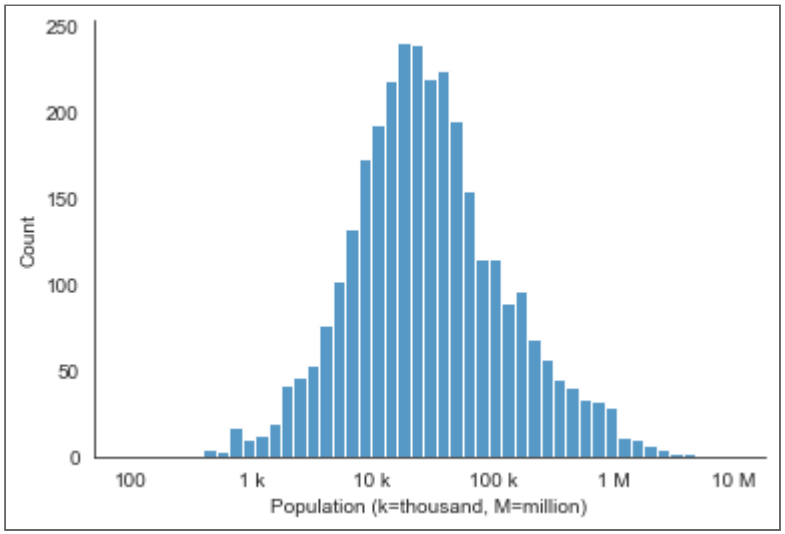
In [6]:

```
#  A histogram of population in all US counties in 2017
fig, ax = plt.subplots()

sns.histplot(data=df,  x= "pop2017", palette=colors_blue)

ax.xaxis.set_major_formatter(ticker.EngFormatter(''))
plt.xlabel("Population (M=million)")
sns.despine();
```



In [7]:

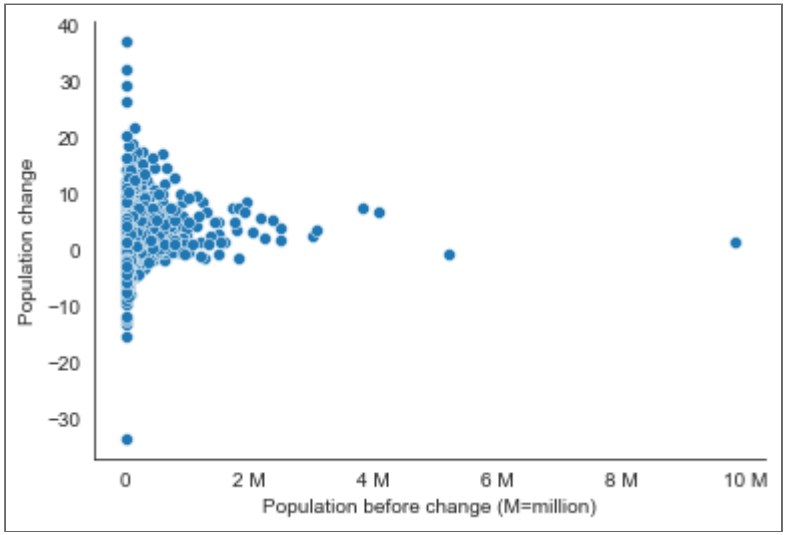
```
# A histogram of log10-transformed values
fig, ax = plt.subplots()

sns.histplot(data=df, x= "pop2017", palette=colors_blue, log_scale=True)

ax.xaxis.set_major_formatter(ticker.EngFormatter(''))
plt.xlabel("Population (k=thousand, M=million)")
sns.despine();
```



Scatterplot

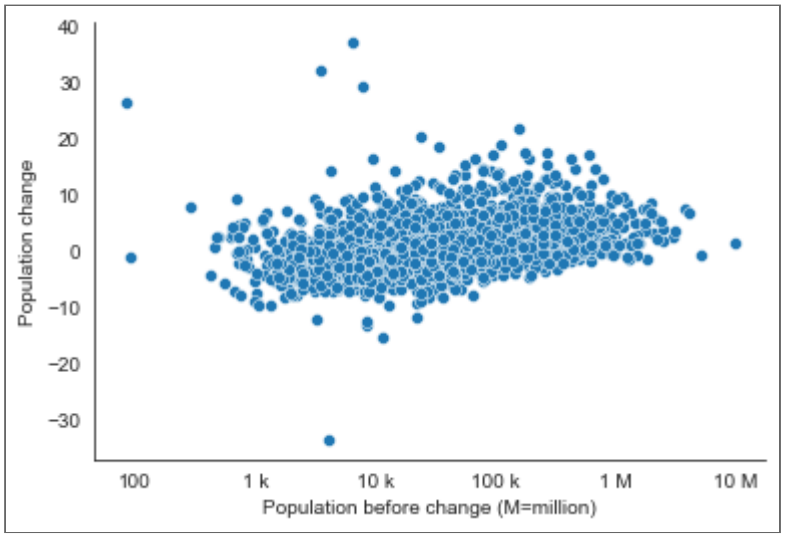
In [8]:

```
# Scatterplot of population change against the population before the change
fig, ax = plt.subplots()

sns.scatterplot(data=df, x="pop2010", y="pop_change")

ax.xaxis.set_major_formatter(ticker.EngFormatter(''))
plt.xlabel("Population before change (M=million)")
plt.ylabel("Population change")
sns.despine();
```



In [9]:

```
#  population size has been log-transformed.
fig, ax = plt.subplots()

ax = sns.scatterplot(data=df, x="pop2010", y="pop_change")

ax.set_xscale('log')
ax.xaxis.set_major_formatter(ticker.EngFormatter(''))
plt.xlabel("Population before change (M=million)")
plt.ylabel("Population change")
sns.despine();
```



Skew

In [10]:

```
df[['unemployment_rate', 'pop2017']].agg(['skew']).transpose()
```

Out[10]:

	skew
unemployment_rate	1.877885
pop2017	13.715670