Week 02: Review of inference logic & classical tests

Foundations & reproducibility (Part 1)

Bartosz Maćkiewicz

Logic of NHST

1. Start by formulating a **research hypothesis**.
2. Set up the **null hypothesis**.
3. Construct the **sampling distribution** of the particular statistic on the assumption that H_0 is true.
4. **Collect** some data.
5. **Compare** the sample statistic to that distribution.
6. Reject or retain H_0, depending on the probability, under H_0, of a sample statistic as extreme as the one we have obtained.

Binomial test: introduction

Binomial distribution

- discrete (not continuous)
- a specific number of **independent** trials
- each trial results in one of two mutually exclusive outcomes
- probability of success is constant across trials
- the distribution describes the probabilities of varying numbers of successes
    - "success" and "failure" are only labels! you can model a wide variety of phenomena using this distribution
- can be used to test hypotheses

Binomial test: working with binomial distribution

The binomial distribution is associated with several key functions:

- `dbinom` (or more generally `d...`) - probability density/mass function
- `pbinom` (or more generally `p...`) - cumulative distribution
- `qbinom` (or more generally `q...`) - quantile function
- `rbinom` (or more generally `r...`) - random generation

A coin is tossed 10 times at random. What is the probability of getting

- exactly one head
- at most three heads
- at least three heads
- four, five or six heads

Binomial test: working with binomial distribution

- **Quantile function** - specifies the value of the random variable such that the probability of the variable being less than or equal to that value equals the given probability.

A coin is tossed 10 times at random. What is the number of heads that you will get less or equal than 90% of times

Binomial test: hypothesis testing

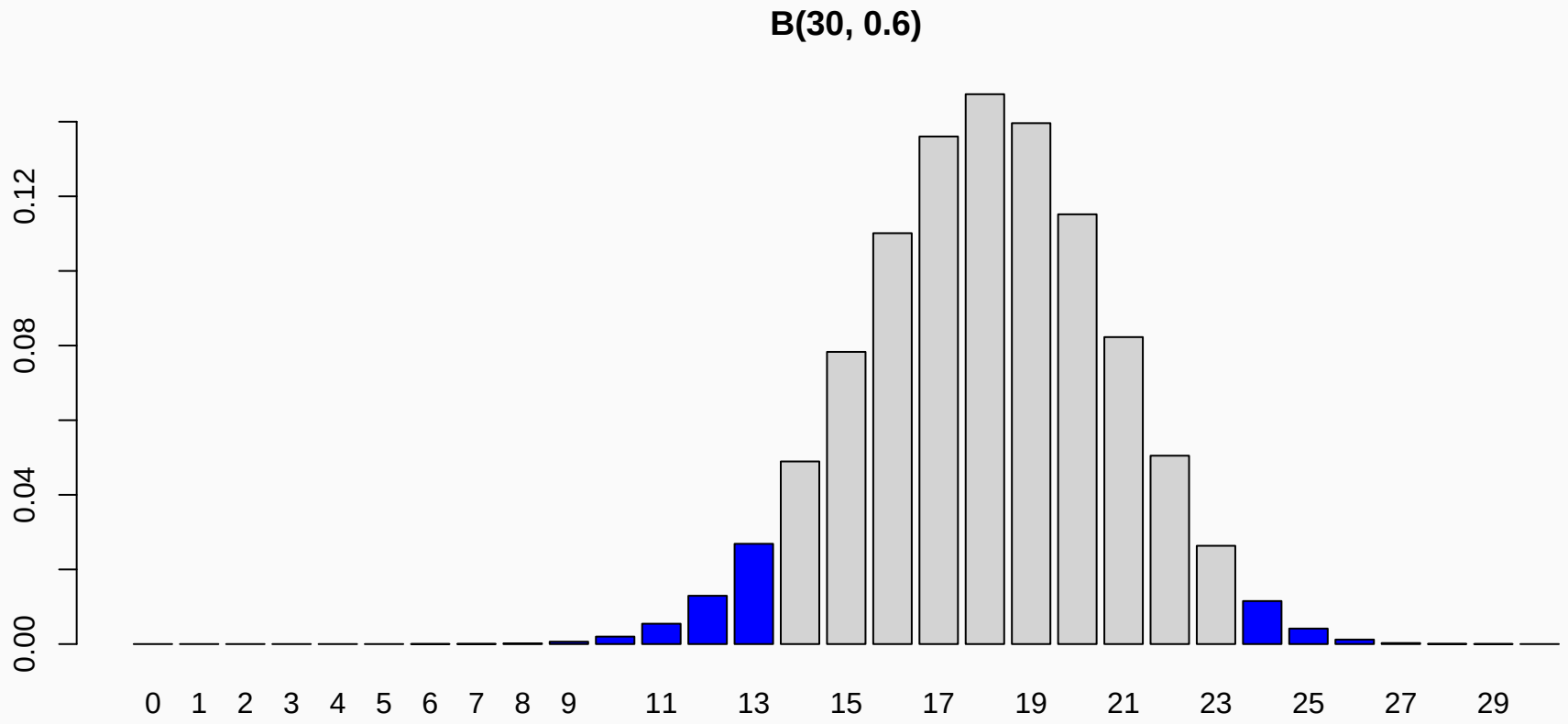
Consider a scenario where we design a driving test, which, based on solid evidence, we expect 60% of all drivers to pass. We administer it to 30 drivers, and 22 pass it. Is the result sufficiently large to cause us to reject H_0: $p = .60$?

You can calculate the result using either the:

- `pbinom` function or
- `binom.test` function

Binomial test: directionality

- **one-tailed or directional test** - we reject H_0 only for the lowest or highest values; our rejection region is located in only one tail of the distribution
- **two-tailed or nondirectional test** - we reject extremes in both tails



Binomial test: Example 1

Under (the assumption of) simple Mendelian inheritance, a cross between plants of two particular genotypes produces progeny 1/4 of which are "dwarf" and 3/4 of which are "giant", respectively.

In an experiment to determine if this assumption is reasonable, a cross results in progeny having 243 dwarf and 682 giant plants. If "giant" is taken as success, the null hypothesis is that $p = 3/4$ and the alternative that $p \neq 3/4$.

Chi-squared goodness-of-fit test

The Chi-squared goodness of fit test is a statistical hypothesis test used to determine whether a variable is likely to come from a specified distribution or not.

Alternative hypothesis

$$H_A : \neg(X \sim F)$$

Null hypothesis

$$H_0 : X \sim F$$

- X empirical distribution
- F theoretical distribution

Chi-squared goodness-of-fit test

- The observed frequencies, as the name suggests, are the frequencies you actually observed in the data.
- The expected frequencies are the frequencies you would expect if the null hypothesis were true.
- You need to assume that the observations are independent of each other.
- The distribution of the χ^2 statistic is approximated by the χ^2 distribution of $k-1$ degrees of freedom where k is the number of categories.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

- $\sum$ sum
- O - observed frequencies
- E - expected frequencies

Chi-squared goodness-of-fit test: Example 1

Many psychologists are particularly interested in how people make decisions, and they often present their subjects with simple games. Psychologists use such games to see how human behavior compares with optimal behavior. We are going to look at the presence or absence of optimal behavior in an universal children's game of "rock/paper/scissors".

While players are expected to throw each symbol with equal frequency, our data indicate an unexpectedly high occurrence of 'Rock' throws.

Rock	Paper	Scissors
30	21	24

However, this may just be a random deviation due to chance. Even if you are deliberately randomizing your throws, one is likely to come out more frequently than others. What we want is a goodness-of-fit test to ask whether the deviations from what would be expected by chance are large enough to lead us to conclude that they were really throwing Rock at greater than chance levels.

Chi-squared test of independence

The expected frequencies in a contingency table represent those frequencies that we would expect if the two variables forming the table were independent.

- H_A: variables are dependent
- H_0: variables are independent

This formula for the expected values is derived directly from the formula for the probability of the joint occurrence of two independent events.

$$E_{ij} = \frac{R_i C_i}{N}$$

Degrees of freedom (dfs) are given by the formula:

$$df = (R-1)(C-1)$$

where R and C = the number of rows and columns in the contingency table.

Chi-squared test of independence: Example 1

There have been a number of studies over the years looking at whether the imposition of a death sentence is affected by the race of the defendant (and/or the race of the victim). Peterson (2001) reports data on a study by Unah and Borger (2001) examining the death penalty in North Carolina in 1993–1997. The data in `unahborger2001.csv` show the outcome of sentencing for white and non-white (mostly black and Hispanic) defendants when the victim was white.

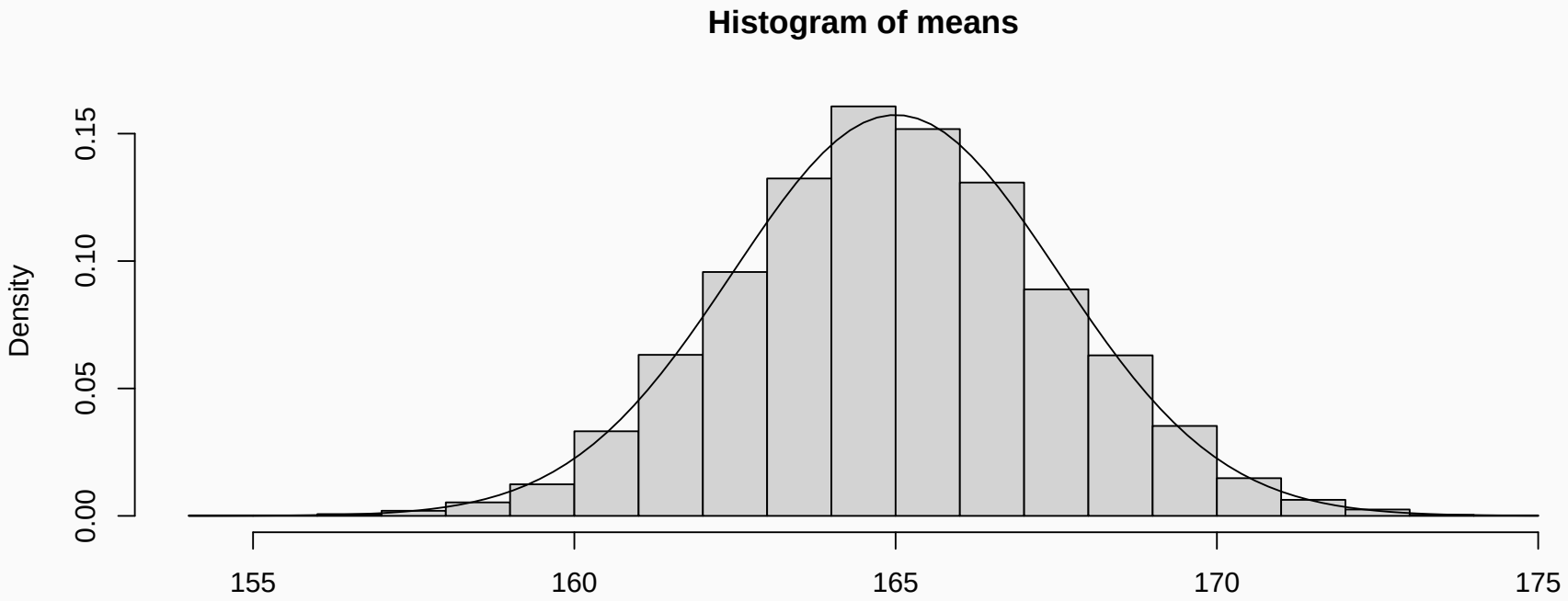
Central Limit Theorem: theory

Given a population with mean μ and variance σ^2, the sampling distribution of the mean (the distribution of sample means) will have a mean equal to μ (i.e., $\mu_X = \mu$), a variance ($\sigma^2_{\bar{X}}$) equal to σ^2/n, and a standard deviation ($\sigma_{\bar{X}}$) equal to $\sigma_X/\sqrt{n}$. The distribution will approach the normal distribution as n, the *sample size*, increases, regardless of the population's distribution shape, given sufficiently large sample size.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Central Limit Theorem: practice

```r
n <- 35; mu <- 165; sigma <- 15
means <- replicate(10000, mean(rnorm(n, mu, sigma)))
mean(means)
## [1] 165.0047
sd(means)
## [1] 2.525481
sigma / sqrt(n)
## [1] 2.535463
```



z test

Williamson (2008) had 166 children from homes in which at least one parent had a history of depression. Because there is evidence in the psychological literature that stress in a child's life may lead to subsequent behavior problems, Williamson expected that a sample of children of depressed parents would show an unusually high level of behavior problems.

These children all completed the Youth Self-Report, and the sample mean was 55.71 with a standard deviation of 7.35. This is a convenient example here because we know the actual population mean and standard deviation of YSR scores — they are 50 and 10.

We want to test the null hypothesis that these children come from a normal population.

z test: NHST in action

1. Start by formulating a **research hypothesis**.
   - Children from homes in which at least one parent had a history of depression display more behavior problems than the general population (H_A: $\mu > 50$).
2. Set up the **null hypothesis**.
   - Children from homes in which at least one parent had a history of depression display the same number of behavior problems as the general population (H_0: $\mu = 50$).
3. Construct the **sampling distribution** of the particular statistic on the assumption that H_0 is true.
   - From Central Limit Theorem we know that sample means are distributed normally: $N(50, 10/\sqrt{166})$.
4. Collect some data.
   - The sample mean was 55.71 with a standard deviation of 7.35
5. Compare the sample statistic to that distribution.
   - What is the probability of getting 55.71 or more from $N(50, 10/\sqrt{166})$?
6. Reject or retain H_0, depending on the probability, under H_0, of a sample statistic as extreme as the one we have obtained.
   - R to the rescue!

z test in R

Direct calculation of probability

```
s_mean <- 55.71
n <- 166
mu <- 50
sigma <- 10
pnorm(s_mean, mu, sigma / sqrt(n), lower.tail = F)
## [1] 9.417126e-14
```

Using z-score (standardized difference)

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

```
z <- (s_mean - mu) / (sigma / sqrt(n))
pnorm(z, lower.tail = F)
## [1] 9.417126e-14
```

z test: Example 1

A principal at a school claims that the students in his school are above average intelligence.

A random sample of thirty students' IQ scores has a mean score of 112.5. Is there sufficient evidence to support the principal's claim?

The mean population IQ is 100 with a standard deviation of 15.

Unknown variance

The preceding example was chosen deliberately from among a fairly limited number of situations in which the population standard deviation (σ) is known.

In the general case, we rarely know the value of σ and usually have to estimate it by way of the sample standard deviation (s).

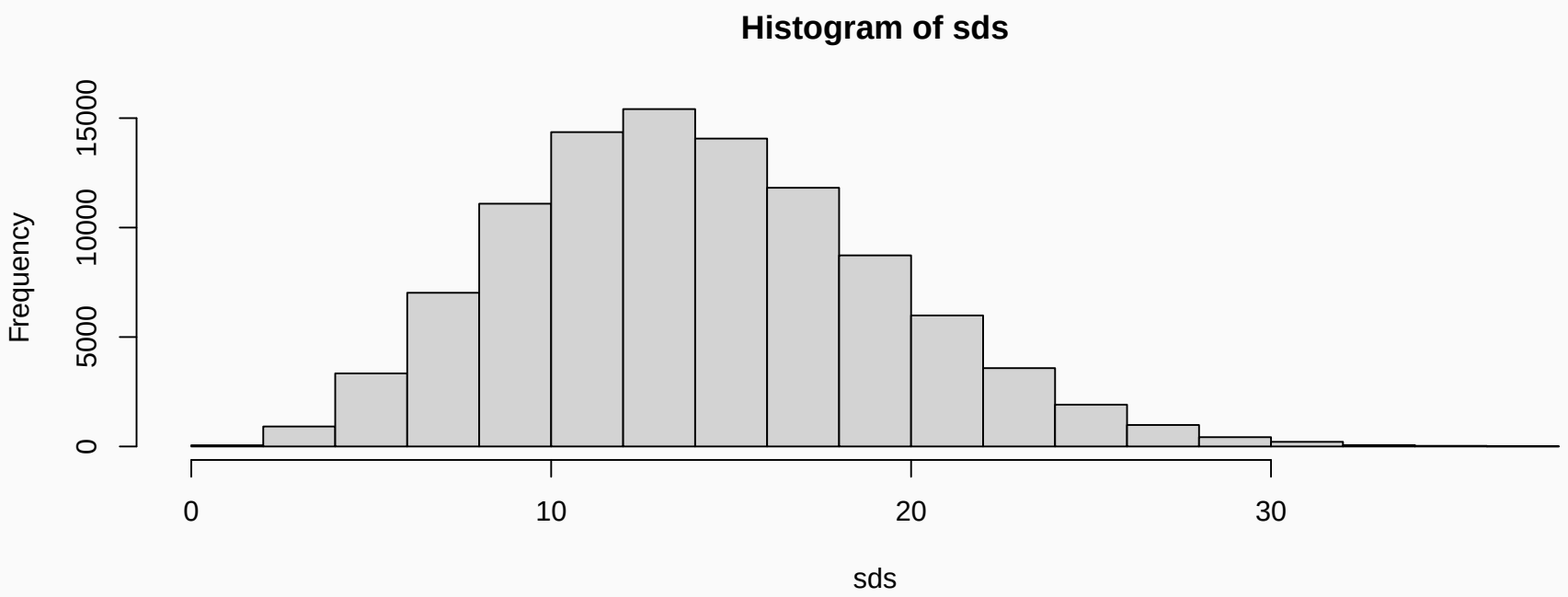
When we replace σ with s in the formula, however, the nature of the test changes.

We can no longer declare the answer to be a z score and evaluate it using the normal distribution. Instead, we will denote the answer as t and evaluate it using t distribution, which is different from normal.

The reasoning behind the switch from z to t is actually related to the sampling distribution of the sample variance.

The sampling distribution of s^2

```r
n <- 5; mu <- 165; sigma <- 15
sds <- replicate(100000, sd(rnorm(n, mu, sigma)))
hist(sds)
```



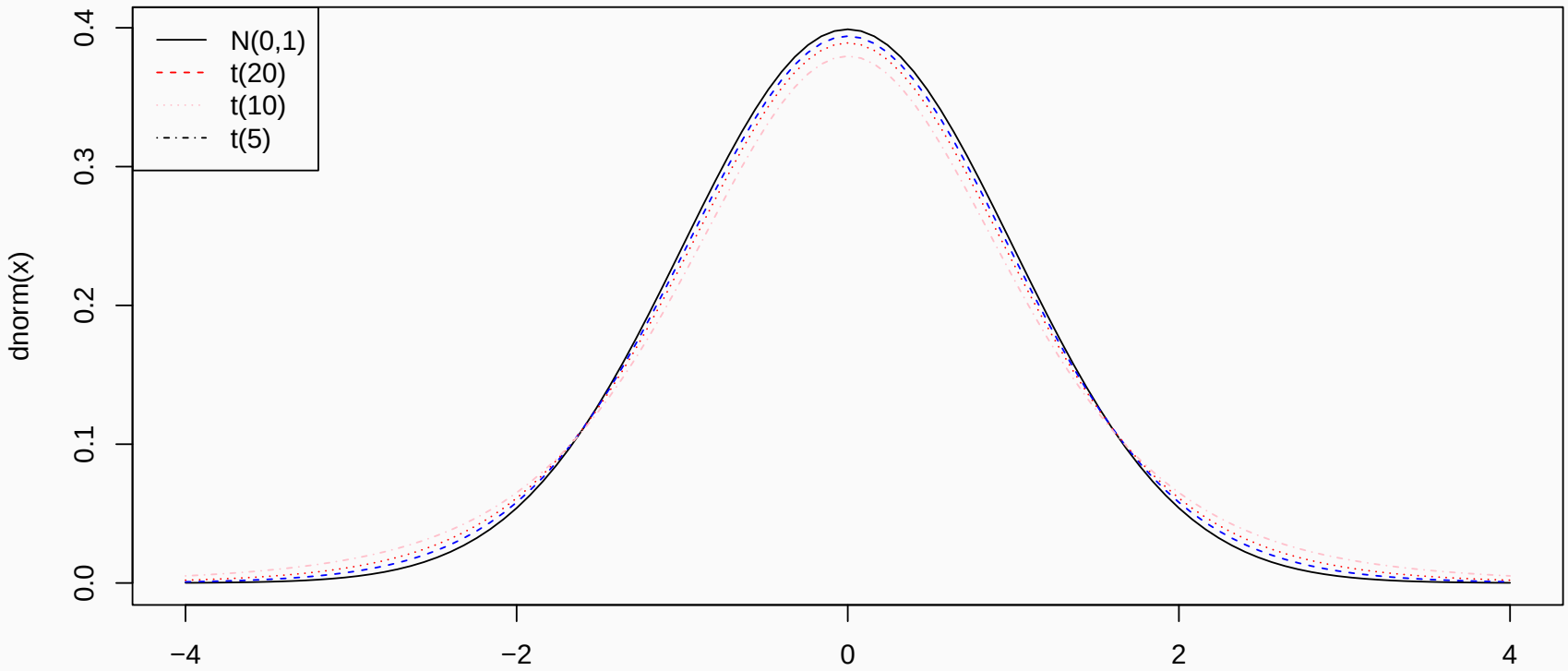

```r
mean(sds)
## [1] 14.09224
mean(sds < sigma)
## [1] 0.59373
```

t statistic and t distribution

t statistic

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

t distribution



t test for one sample

Compas and others (1994) were surprised to find that young children under stress actually report fewer symptoms of anxiety and depression than we would expect. But they also noticed that their scores on a Lie Scale (a measure of the tendency to give socially desirable answers) were higher than expected. The population mean for the Lie scale on the Children's Manifest Anxiety Scale (Reynolds and Richmond, 1978) is known to be 3.87. Scores of 36 children under stress can be found in the `compas1994.csv` file.

How would we test whether this group shows an increased tendency to give socially acceptable answers?

t test for one sample: Example 1

Katz, Lautenschlager, Blackburn, and Harris (1990) examined the performance of 28 students who answered multiple-choice items on the SAT without having read the passages to which the items referred. Scores can be found in the `katz1990.csv` file. Random guessing would have been expected to result in 20 correct answers.

- Were these students responding at better-than-chance levels?
- If performance is statistically significantly better than chance, does it mean that the SAT test is not a valid predictor of future college performance?

t test for independent samples

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

We can construct t statistic by substituting known variance with variance estimated from the samples. Because the null hypothesis is generally the hypothesis that $\mu_1 - \mu_2 = 0$, we will drop that term from the equation and write

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Two sample variances (s_1^2 and s_2^2) have gone into calculating t. Each of these variances is based on squared deviations about their corresponding sample means, and therefore each sample variance has $n_j - 1$ df. Across the two samples, therefore, we will have $(n_1 - 1) + (n_2 - 1) = (n_1 + n_2 - 2)$ df. Thus, the t for two independent samples will be based on $n_1 + n_2 - 2$ degrees of freedom.

t test for independent samples: Example 1

Adams, Wright, & Lohr (1996) were interested in some basic psychoanalytic theories that homophobia may be unconsciously related to the anxiety of being or becoming homosexual.

They administered the Index of Homophobia to 64 heterosexual males and classed them as homophobic or nonhomophobic on the basis of their score. They then exposed homophobic and nonhomophobic heterosexual men to videotapes of sexually explicit erotic stimuli portraying heterosexual and homosexual behavior and recorded their level of sexual arousal.

Adams et al. reasoned that if homophobia were unconsciously related to anxiety about one's own sexuality, homophobic individuals would show greater arousal to homosexual videos than would nonhomophobic individuals.

In this example, we will examine only the data from the homosexual video. The data in `adams1996.csv` file were created to have the same means and variance as the data that Adams collected.

The dependent variable is the degree of arousal at the end of the 4-minute video, with larger values indicating greater arousal.

Correlation and Regression

We will explore two interrelated topics: correlation and regression.

Statisticians often distinguish between these two techniques, although the distinction is not always observed in practice. However, it is significant enough to merit brief consideration.

In problems of simple correlation and regression, the data consist of two observations from each of N subjects, one observation on each of the two variables under consideration. If we were interested in the correlation between the running speed of mice in a maze (Y) and the number of trials to reach some criterion (X) (both common measures of learning), we would obtain a running-speed score and a trials-to-criterion score from each subject.

Similarly, if we were interested in the regression of running speed (Y) on the number of food pellets per reinforcement (X), each subject would have scores corresponding to his speed and the number of pellets he received.

The difference between these two situations illustrates the statistical distinction between correlation and regression.

In both cases, Y (running speed) is a *random variable*, beyond the experimenter's control. We don't know what the mouse's running speed will be until we carry out a trial and measure the speed.

In the former case, X is also a random variable, because the number of trials to criterion depends on how fast the animal learns, and this, too, is beyond the control of the experimenter. Put another way, a replication of the experiment would leave us with different values of both Y and X. In the food pellet example, however, X is a fixed variable. The number of pellets is determined by the experimenter (for example, 0, 1, 2, or 3 pellets) and would remain constant across replications.

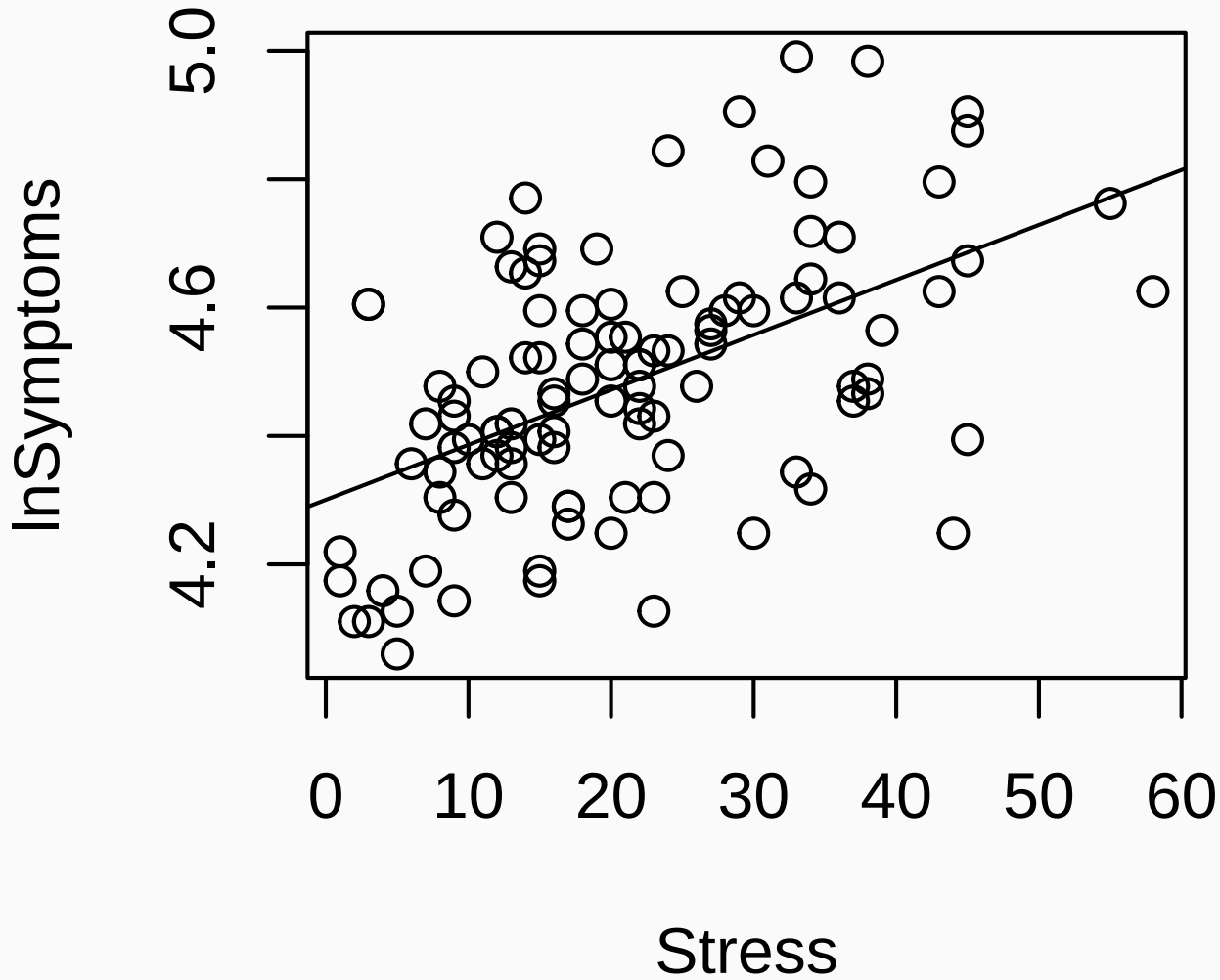
To some statisticians, the word **regression** is reserved for those situations in which **the value of X is fixed or specified by the experimenter before the data are collected**. In these situations, no sampling error is involved in X, and repeated replications of the experiment will involve the same set of X values. The word **correlation** is used to describe the situation in which **both X and Y are random variables**.

Example 1: Stress and mental health (`wagner1988.csv`)

Wagner, Compas, and Howell (1988) investigated the relationship between stress and mental health in first-year college students. Using a scale they developed to measure the frequency, perceived importance, and desirability of recent life events, they created a measure of negative events weighted by the reported frequency and the respondent's subjective estimate of the impact of each event. This served as their measure of the subject's perceived social and environmental stress. They also asked students to complete the Hopkins Symptom Checklist, assessing the presence or absence of 57 psychological symptoms.

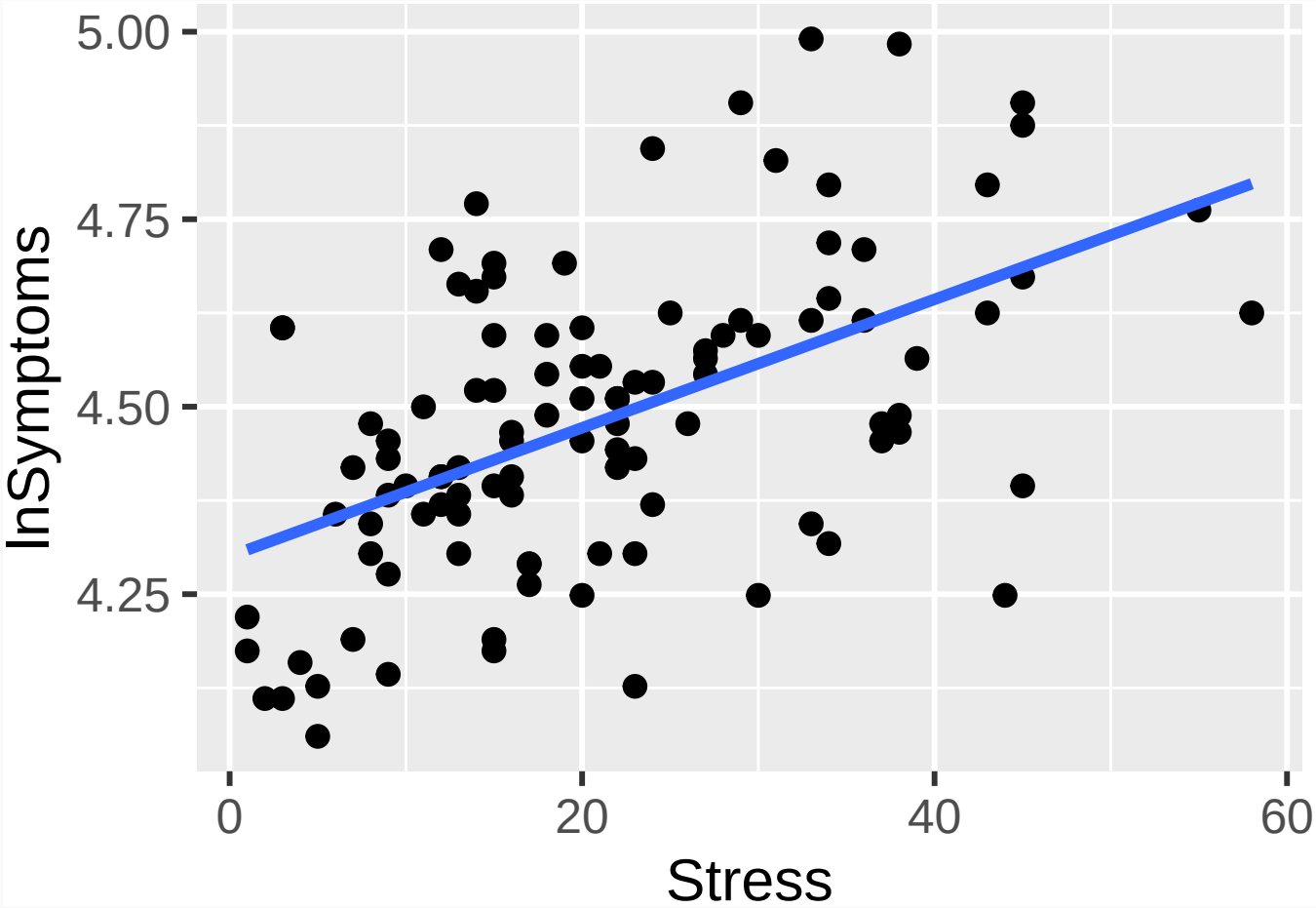
Scatterplot: visualization of bivariate data - base R

```
wagner <- read.csv("wagner1988.csv")
plot(lnSymptoms ~ Stress, data = wagner)
fit <- lm(lnSymptoms ~ Stress, data = wagner) # fit a linear model
abline(fit) # add regression line
```



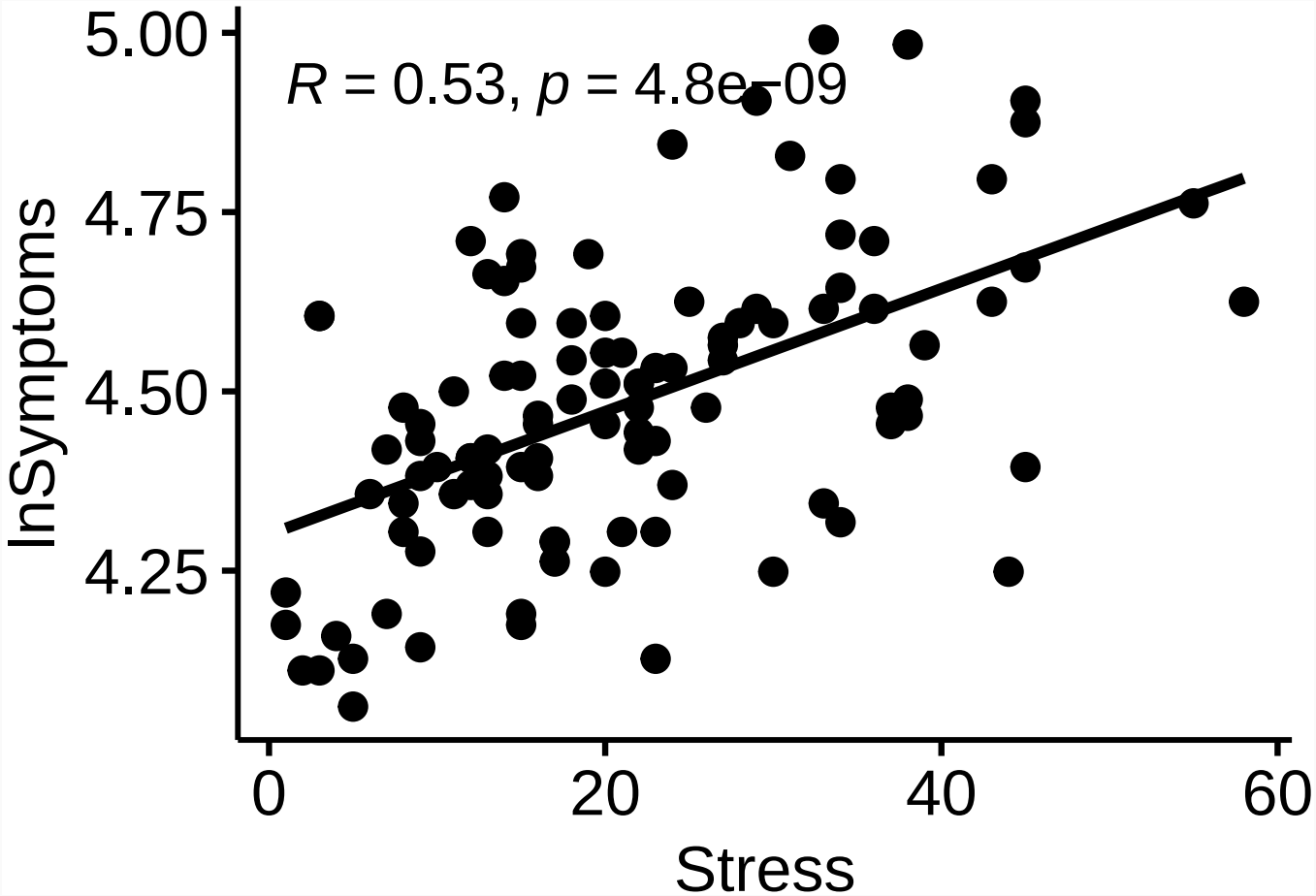
Scatterplot: visualization of bivariate data - ggplot2

```
library(ggplot2)
ggplot(data = wagner) +
  aes(x = Stress, y = lnSymptoms) +
  geom_point() +
  geom_smooth(method = lm, se = F) # add regression line
```



Scatterplot: visualization of bivariate data - ggpubr

```
library(ggpubr)
ggscatter(data = wagner, x = "Stress", y = "lnSymptoms",
          add = "reg.line", cor.coef = T)
```



The Covariance

The covariance is a number that reflects the degree to which two variables vary together.

$$cov_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{N - 1}$$

The covariance is similar in form to the variance. If we changed all the Ys in the equation to Xs, we would have s_X^2 ; if we changed the Xs to Ys, we would have s_Y^2.

“Mechanics” of the Covariance

$$cov_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{N - 1}$$

Case I: Positive Relationship

For the data on `Stress` and `lnSymptoms` we would expect that high stress scores will be paired with high symptom scores. Thus, for a stressed participant with many problems, both $(X - \bar{X})$ and $(Y - \bar{Y})$ will be positive and their product will be positive.

For a participant experiencing little stress and few problems, both $(X - \bar{X})$ and $(Y - \bar{Y})$ will be negative, but their product will again be positive. Thus, the sum of $(X - \bar{X})(Y - \bar{Y})$ will be large and positive, giving us a large positive covariance.

$$cov_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{N - 1}$$

Case II: Negative Relationship

The reverse would be expected in the case of a strong negative relationship. Here, large positive values of $(X - \bar{X})$ most likely will be paired with large negative values of $(Y - \bar{Y})$, and *vice versa*. Thus, the sum of the products of the deviations will be large and negative, indicating a strong negative relationship.

$$cov_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{N - 1}$$

Case III: No relationship

Finally, consider a situation in which there is no relationship between X and Y. In this case, a positive value of $(X - \bar{X})$ will sometimes be paired with a positive value and sometimes with a negative value of $(Y - \bar{Y})$. The result is that the products of the deviations will be positive about half of the time and negative about half of the time, producing a near-zero sum and indicating no relationship between the variables.

The Covariance in R

```
sum(
  (wagner$Stress - mean(wagner$Stress)) *
  (wagner$lnSymptoms - mean(wagner$lnSymptoms))
  ) /
(nrow(wagner) - 1)
```

```
[1] 1.336434
```

```
cov(wagner$Stress, wagner$lnSymptoms)
```

```
[1] 1.336434
```

The Pearson Correlation Coefficient

$$r = \frac{cov_{XY}}{s_X s_Y}$$

Because the maximum value of cov_{XY} can be shown to be $s_X s_Y$, it follows that the limits on r are ± 1.00. One interpretation of r, then, is that it is a measure of the degree to which the covariance approaches its maximum.

The Pearson Correlation Coefficient in R

```
(sum(
  (wagner$Stress - mean(wagner$Stress)) *
  (wagner$lnSymptoms - mean(wagner$lnSymptoms))
  ) /
 (nrow(wagner) - 1)
) /
(sd(wagner$Stress) * sd(wagner$lnSymptoms))
```

```
[1] 0.5286565
```

```
cor(wagner$Stress, wagner$lnSymptoms)
```

```
[1] 0.5286565
```

Intepretation of r

Rules for interpretation of r ale domain specific.

According to Funder and Ozer (2019) (see `effectsize::interpret_r` for more details and different sets of rules):

- $r < 0.05$ - Tiny
- $0.05 \leq r < 0.1$ - Very small
- $0.1 \leq r < 0.2$ - Small
- $0.2 \leq r < 0.3$ - Medium
- $0.3 \leq r < 0.4$ - Large
- $r \geq 0.4$ - Very large

Guess the correlation

http://guessthecorrelation.com/ - an useful tool for developing intuition about correlation

Statistical test for correlation

The most common hypothesis that we test for a sample correlation is that the correlation between X and Y in the population, denoted ρ (rho), is zero. This is a meaningful test because the null hypothesis being tested is really the hypothesis that X and Y are linearly independent. Rejection of this hypothesis leads to the conclusion they are not independent and there is some linear relationship between them.

It can be shown that when $\rho = 0$, for large N, r will be approximately normally distributed around zero. (It is not normally distributed around its mean when $\rho \neq 0$) A legitimate t test can be formed from the ratio

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

which is distributed as t on $N-2\ df$.

Statistical test for correlation in R

```r
covariance <- sum(
  (wagner$Stress - mean(wagner$Stress)) *
  (wagner$lnSymptoms - mean(wagner$lnSymptoms))
                  ) / (nrow(wagner) - 1)
covariance
## [1] 1.336434
correlation <- covariance /
               (sd(wagner$Stress) * sd(wagner$lnSymptoms))
correlation
## [1] 0.5286565
t_stat <- correlation * sqrt(nrow(wagner) - 2) /
          sqrt(1 - correlation**2)
t_stat
## [1] 6.381819
pval <- pt(t_stat, nrow(wagner) - 2, lower.tail = F)
pval
## [1] 2.413743e-09
```

Statistical test for correlation in R

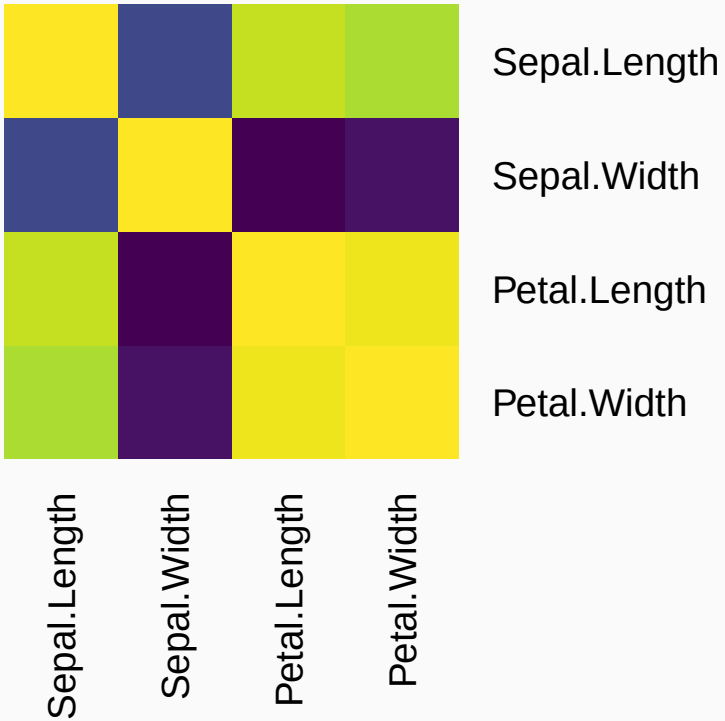
```r
cor.test(~ Stress + lnSymptoms, data = wagner)
```

```
    Pearson's product-moment correlation

data:  Stress and lnSymptoms
t = 6.3818, df = 105, p-value = 4.827e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3765970 0.6529758
sample estimates:
      cor
0.5286565
```

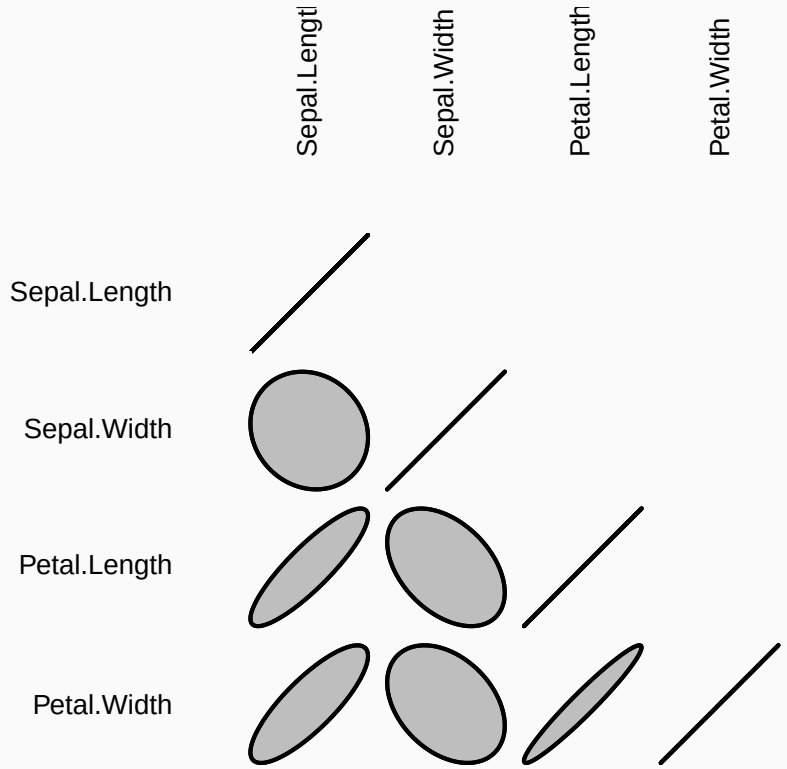
Visualization of correlations in multivariate datasets - base R

```r
library(viridis)
correlations <- cor(iris[, 1:4])
heatmap(correlations, symm = T, Rowv = NA, col = viridis(256),
        cexRow = 0.7, cexCol = 0.7) # graphical parameters, ignore
```



Visualization of correlations in multivariate datasets - `ellipse`

```
library(ellipse)
plotcorr(correlations, type = "lower", diag = T,
         cex.lab = 0.4, cex = 0.2, mar = c(4,4,4,4)) # ignore
```



Visualization of correlations in multivariate datasets - `corrplot`

```
library(corrplot)
corrplot.mixed(correlations, lower = "ellipse", upper = "number",
               cl.cex = 0.7, number.cex = 0.7, tl.cex = 0.5) # ignore
```