Regression and Linear Models (37252)

Lecture 1 - Statistics Review

Lecturer: Joanna Wang
Notes adopted from Scott Alexander

School of Mathematical and Physical Sciences, UTS

Autumn 2024

Acknowledgement

These notes are based, in part, on earlier versions prepared by Dr Ed Lidums and Prof. James Brown.

1. Lecture outline

Topics:

- subject outline
- types of RVs
- analytical tools for numerical RVs
  - PMF
  - PDF
- descriptions of RVs
- graphical descriptions of RVs
  - bar chart
  - pie chart
  - boxplot
  - histogram
  - scatter plot
  - grouped boxplots

1. Lecture outline

Topics (continued):

- population statistics
  - mean
  - variance
  - quantile
  - mode
  - covariance
  - correlation
- sample statistics
  - mean and variance
  - confidence interval
  - descriptive statistics
  - correlation
- descriptions of RVs – summary
- hypothesis testing
  - one-sample T-test
  - two-independent-sample T-test
  - F-test (introduction)

1. Subject Outline

Subject outline is available on Canvas and contains important information regarding the schedule and assessments for this subject.

Weekly schedule:

- one 2-hour lecture
- one 2-hour lab

Assessment tasks:

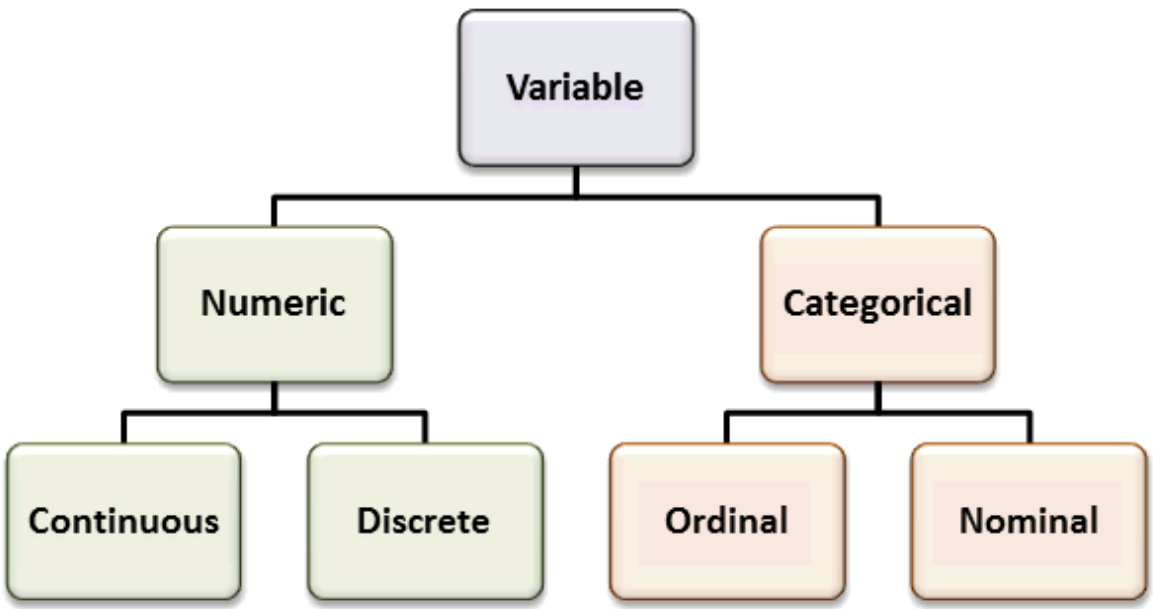
- weekly lab worksheet (20%)
- group assignment (30%)
- 2-hour final exam (50%)

1. Types of RVs

A **random variable (RV)** is a variable whose value is determined by some chance experiment.

We can think of a random variable as a special type of function.

RVs can be classified according to the type of values they take.



Source: www.abs.gov.au

1. Analytical tools for numerical RVs – PMF

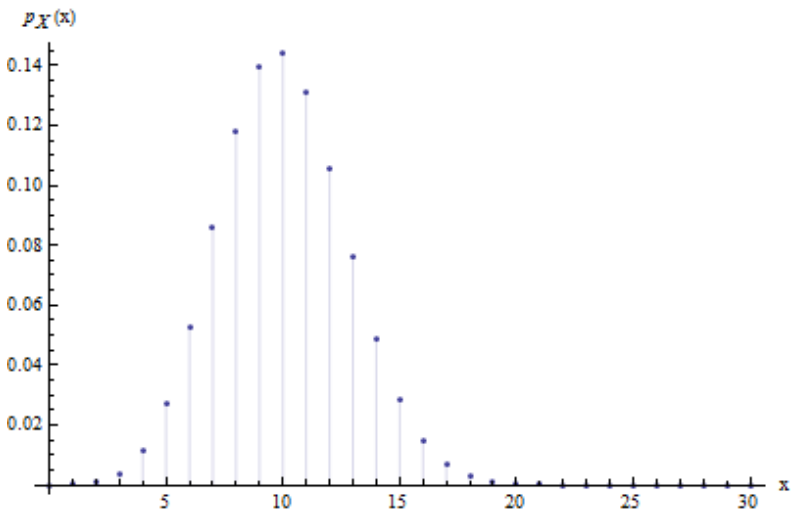
Let X be a discrete RV taking values in $\mathbb{Z}$, the set of integers. (Discrete RVs can be defined on **countable** sets of real numbers)

Associated with such a RV X is its **probability mass function (PMF)** p_X given by

$$\text{Prob}(X = x) = p_X(x),$$

with the property $\sum_{x=-\infty}^{\infty} p_X(x) = 1$ (probabilities must sum to one).

Example: PMF of $X \sim B(40, 1/4)$ RV



1. Analytical tools for numerical RVs – PDF

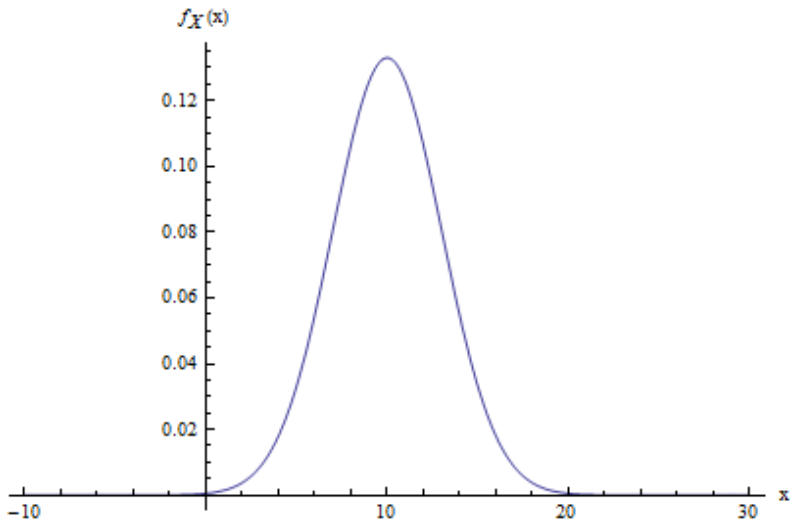
Let Y be a continuous RV taking values in $\mathbb{R}$, the set of real numbers. (Continuous RVs can be defined on any interval.)

If one can write

$$\text{Prob}(a \leq Y \leq b) = \int_a^b f_Y(y)\,\mathrm{d}y,$$

then f_Y is called the **probability density function (PDF)** of the RV Y, which possess the property $\int_{-\infty}^{\infty} f_Y(y)\,\mathrm{d}y = 1$ (probabilities sum to one).

Example: PDF of $Y \sim N(10, 3)$ RV



1. Descriptions of RVs

We often have to deal with large amounts of data, sometimes many thousands or even millions of data points.

This means it is virtually impossible to inspect the actual data and have any hope of extracting meaningful information.

In statistics we instead make use of various tools to describe characteristics of the data.

These tools tend to be either **graphical** or **numerical** in nature.

Graphical tools can be used to describe all RVs, but numerical tools are used only on numerical RVs.

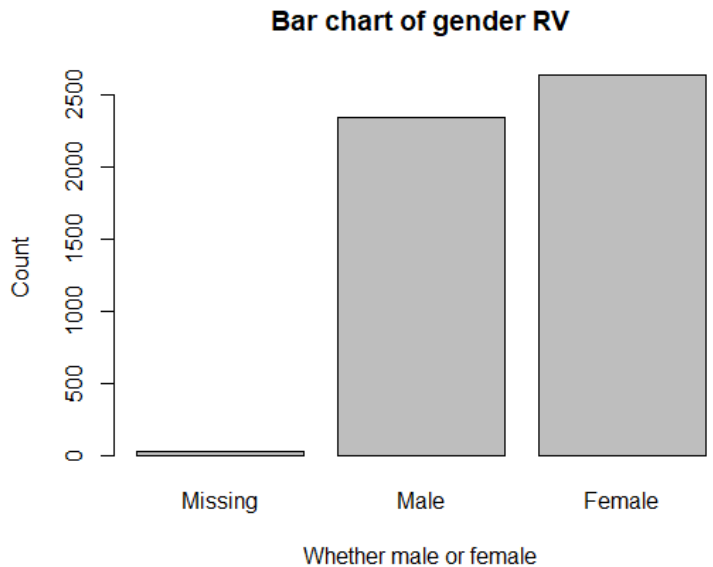
1. Graphical descriptions of RVs – bar chart

Consider the data set health.csv (see Week 2 folder on Canvas).

One of the RVs in this set is the categorical variable *gender*.

We can construct a **bar chart** of this RV using the R command

```
counts <- table(health$gender)
barplot(counts, main="Bar chart of gender RV", xlab="Whether male or female",
ylab="Count", names.arg=c("Missing","Male","Female"))
```

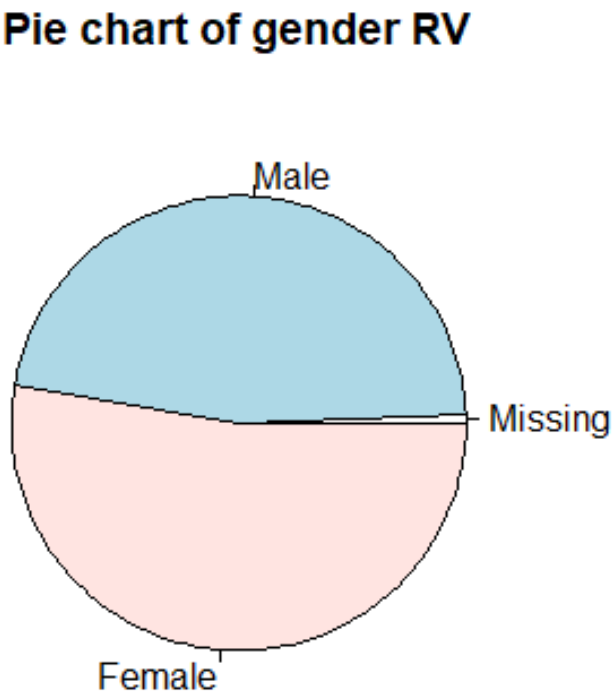


Q. What is the **modal category**? A. Female (more on the **mode** shortly).

1. Graphical descriptions of RVs – pie chart

An alternative is to use a **pie chart**, again in R using

```
pie(counts,labels=c("Missing","Male","Female"),main="Pie
chart of gender RV")
```

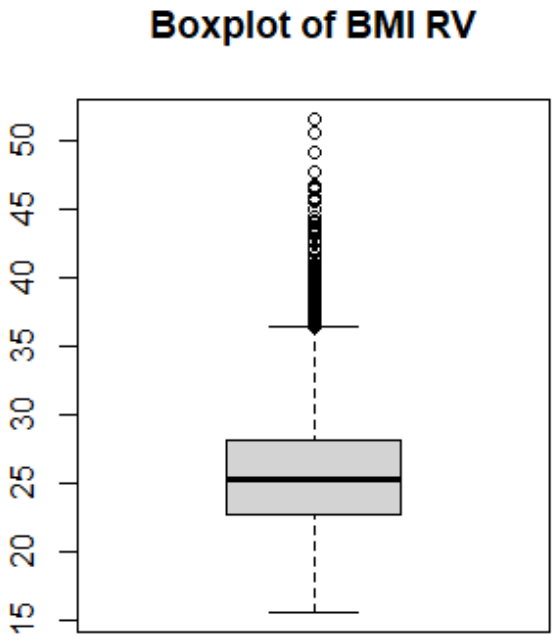


Both the bar and pie charts are easy to interpret as the RV *gender* can take only two values: **female** or **male** (categorical variable).

1. Graphical descriptions of RVs – boxplot

Below is a **boxplot** of the variable *bmi*, also from the same data set.

```
boxplot(health$bmi, main="Boxplot of BMI RV")
```

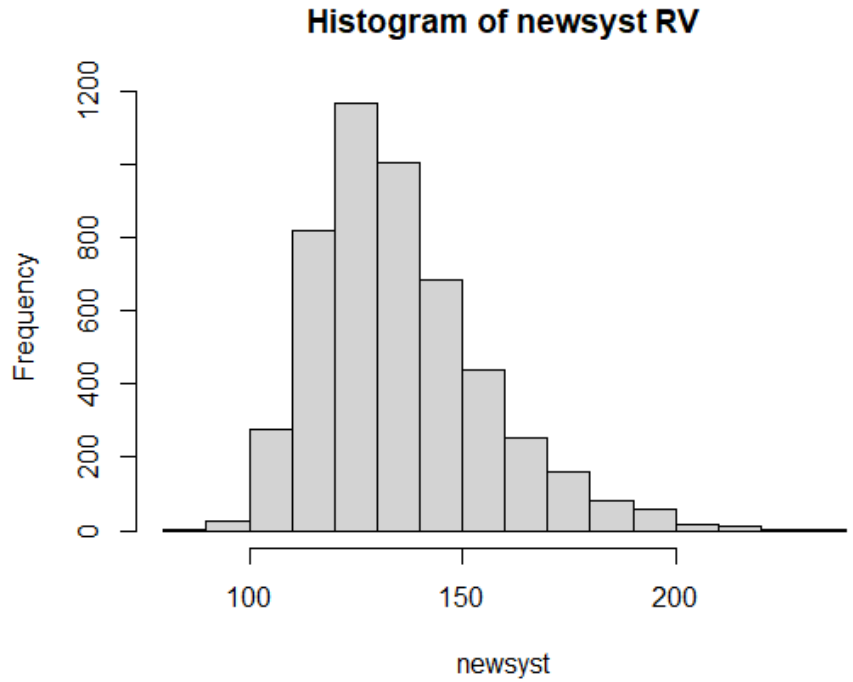


The box height is the **the interquartile range (IQR)**, the line in the box identifies the **median**, each fence and whisker is one-and-a-half times the IQR and the **outliers** are marked with either a circle or asterisk.

1. Graphical descriptions of RVs – histogram

Another way to get a feel for the **distribution** of a numerical RV is with a **histogram**.

```
hist(health$newsyst,xlab="newsyst",main="Histogram of newsyst RV")
```

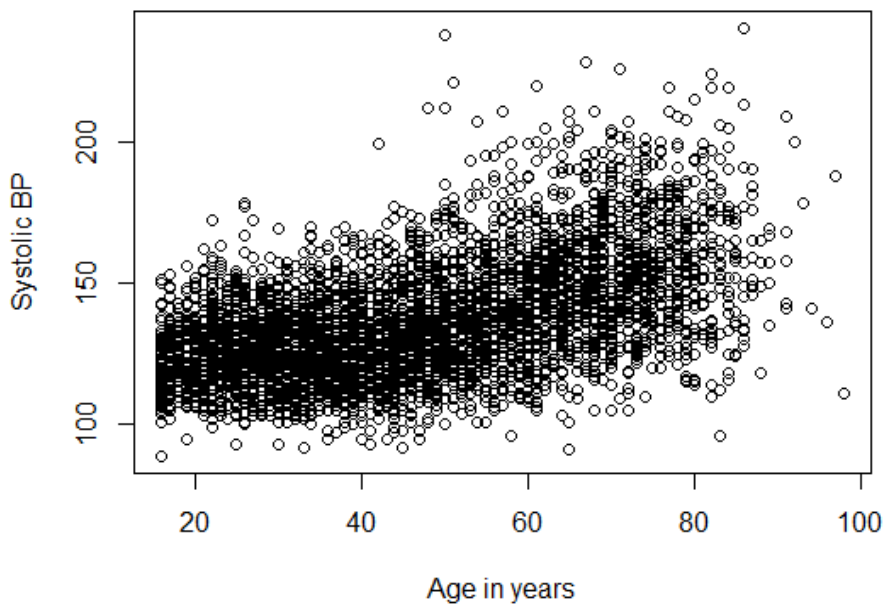


We will see that a histogram is related to a very important property of a RV, its density function or mass function.

1. Graphical descriptions of RVs – scatter plot

When we have a bivariate RV, both numerical, we need to consider possible **dependence** between them. A **scatter plot** can convey this visually.

```
plot(health$age, health$newsyst,xlab="Age in years",
ylab="Systolic BP")
```

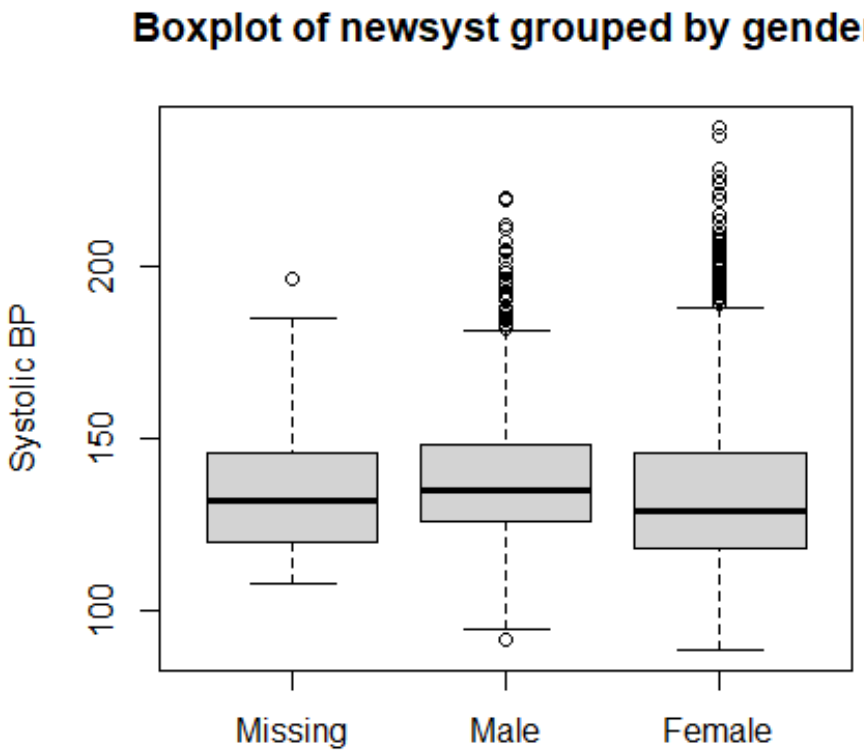


Here we see some suggestion of a quadratic relationship between the RVs.

1. Graphical descriptions of RVs – grouped boxplots

When we have a bivariate RV, one numerical and the other categorical, a grouped boxplot can be generated using the commands

```
boxplot(health$newsyst~health$gender,ylab="Systolic BP",
names=c("Missing","Male","Female"),xlab="",
main="Boxplot of newsyst grouped by gender")
```



1. Population statistics – mean

We can use the PMF or PDF of a RV to calculate **population statistics**.

The **mean**, average or **expected value** of the discrete RV X is defined as

$$\mu_X := \mathbb{E}[X] = \sum_{x=-\infty}^{\infty} x p_X(x)$$

while for the continuous RV Y as

$$\mu_Y := \mathbb{E}[Y] = \int_{-\infty}^{\infty} y f_Y(y)\,\mathrm{d}y.$$

This is an example of a **location parameter**.

N.B. If you haven't studied calculus fear not - we won't be evaluating integrals in this course.

1. Population statistics – variance

The PMF or PDF can also be used to calculate **variance**, a measure of the variability of a RV.

The **variance** for the discrete RV X is

$$\text{var}(X) \equiv \sigma_X^2 := \mathbb{E}[(X - \mu_X)^2] = \sum_{x=-\infty}^{\infty} (x - \mu_X)^2 p_X(x)$$

and the continuous RV Y is

$$\text{var}(Y) \equiv \sigma_Y^2 := \mathbb{E}[(Y - \mu_Y)^2] = \int_{-\infty}^{\infty} (y - \mu_Y)^2 f_Y(y)\,\mathrm{d}y.$$

This is an example of a **scale parameter**.

(Recall variance as the square of the **standard deviation**.)

1. Population statistics – quantile

Another useful statistic is the **α-quantile**, $0 \leq \alpha \leq 1$.

For the continuous RV Y, the α-quantile y_α satisfies

$$\text{Prob}(Y < y_\alpha) = \alpha.$$

For the discrete RV X, the α-quantile x_α satisfies

$$\text{Prob}(X < x_\alpha) \leq \alpha.$$

We can use α-quantiles to define other statistics.

The **median** of the discrete RV X is given by the $1/2$-quantile

$$\text{median}(X) := x_{1/2}$$

and the **inter-quartile range (IQR)** by the interval

$$\text{IQR}(X) := x_{3/4} - x_{1/4}.$$

1. Population statistics – mode

Another location parameter is the **mode** which for the discrete random variable X, can be defined in terms of its PMF p_X as the solution of

$$\operatorname*{argmax}_{x \in \mathbb{Z}} p_X(x),$$

i.e. the value of $x \in \mathbb{Z}$ which maximises $p_X(x)$.

Similarly, the **mode** of the continuous RV Y with PDF f_Y is the solution of

$$\operatorname*{argmax}_{y \in \mathbb{R}} f_Y(y).$$

1. Population statistics – covariance

Consider the discrete random variables X_1, X_2 taking values in $\mathbb{Z}$. When dealing with multivariate RVs one is often interested in any **dependence** between them.

One such measure is the **covariance**, defined in the discrete case as

$$\begin{aligned}\text{covar}(X_1, X_2) \equiv \sigma^2_{X_1,X_2} &:= \mathbb{E}[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})] \\ &= \sum_{x_2=-\infty}^{\infty} \sum_{x_1=-\infty}^{\infty} (x_1 - \mu_{X_1})(x_2 - \mu_{X_2}) p_{X_1,X_2}(x_1, x_2)\end{aligned}$$

where $\mu_{X_1} := \mathbb{E}[X_1]$, $\mu_{X_2} := \mathbb{E}[X_2]$ and the **bivariate PMF** is defined via the relationship

$$\text{Prob}(X_1 = x_1, X_2 = x_2) = p_{X_1,X_2}(x_1, x_2).$$

Of course, the PMF must satisfy $\sum_{x_2=-\infty}^{\infty} \sum_{x_1=-\infty}^{\infty} p_{X_1,X_2}(x_1, x_2) = 1$.

1. Population statistics – covariance

Now consider the continuous random variables Y_1, Y_2 taking values in $\mathbb{R}$.

The **covariance** in this case is given by

$$\begin{aligned}\text{covar}(Y_1, Y_2) \equiv \sigma^2_{Y_1,Y_2} &:= \mathbb{E}[(Y_1 - \mu_{Y_1})(Y_2 - \mu_{Y_2})] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y_1 - \mu_{Y_1})(y_2 - \mu_{Y_2}) f_{Y_1,Y_2}(y_1, y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2,\end{aligned}$$

where $\mu_{Y_1} := \mathbb{E}[Y_1]$, $\mu_{Y_2} := \mathbb{E}[Y_2]$ and the **bivariate PDF** is defined via the relationship

$$\text{Prob}(a_1 \leq Y_1 \leq b_1, a_2 \leq Y_2 \leq b_2) = \int_{a_2}^{b_2}\int_{a_1}^{b_1} f_{Y_1,Y_2}(y_1, y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2.$$

Once more we have the condition $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2 = 1$.

1. Population statistics – correlation

An alternative measure of dependence between two RVs is **Pearson's correlation coefficient**

$$\text{corr}(X_1, X_2) \equiv \rho_{X_1,X_2} := \frac{\text{covar}(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}},$$

where σ_{X_1} and σ_{X_2} are the standard deviations of the discrete RVs X_1 and X_2 respectively.

This measure transforms covariance onto the scale

$$-1 \leq \rho_{X_1,X_2} \leq 1,$$

with -1 indicating perfect negative (linear) dependence and $+1$ perfect positive (linear) dependence.

The correlation parameter ρ_{Y_1,Y_2} for the continuous RVs Y_1 and Y_2 is defined in the same way.

1. Sample statistics

The population statistics for discrete and continuous RVs just described are properties of the **distribution** of these RVs and are theoretical in nature.

However, when working with real data we will never be sure which PMF or PDF to use and therefore can't be certain what the population statistics are.

Instead we use **sample statistics** and from these infer information about their population counterparts.

1. Sample statistics – mean and variance

Suppose we have drawn a sample $X_1, \ldots, X_n$ of n independent observations of some numerical random variable X.

The **sample mean** of this RV is defined as

$$\overline{X} := \frac{1}{n} \sum_{i=1}^{n} X_i$$

and the **sample variance** as

$$S^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2,$$

which is the square of the **sample standard deviation** S.

These are **unbiased estimators** as it is easy to show that $\mathbb{E}[\overline{X}] = \mathbb{E}[X]$ and $\mathbb{E}[S^2] = \mathrm{var}(X)$.

Sample versions of other population statistics can be constructed similarly.

1. Sample statistics – confidence interval

A **confidence interval** for the population mean $\mu = \mathbb{E}[X]$ can be constructed from these sample statistics.

If X follows a **normal distribution** with mean μ and standard deviation σ, then the RV

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

follows a **normal distribution** with mean zero and standard deviation of one.

If σ is unknown, consider instead the RV

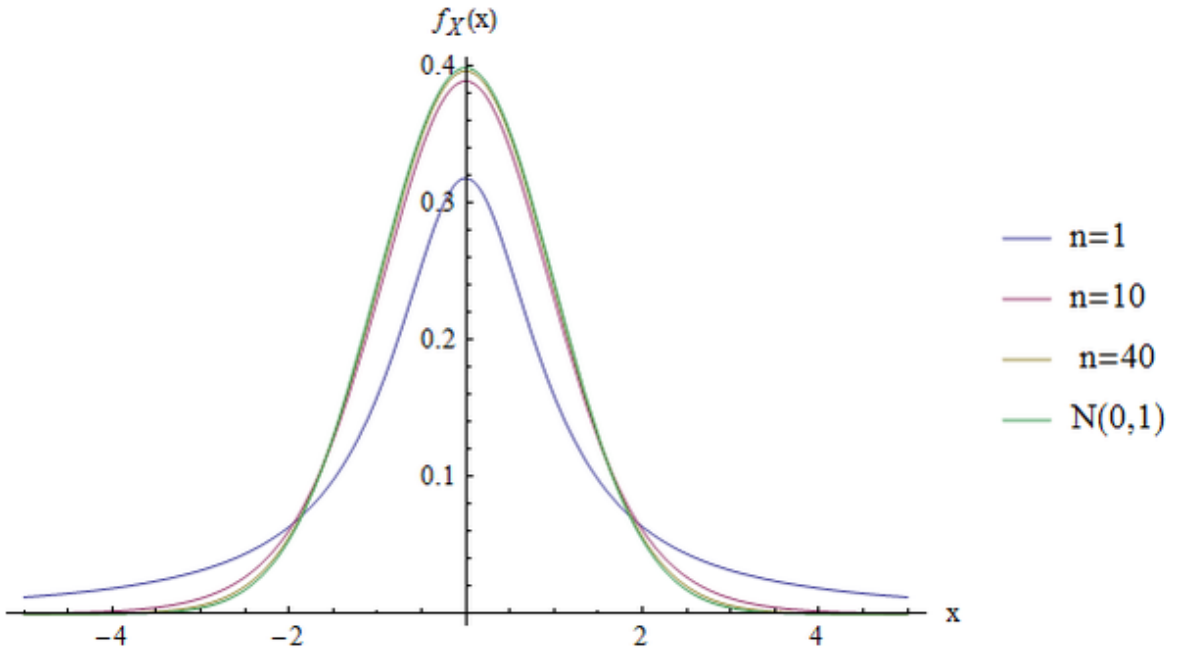
$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \tag{1}$$

which follows a **Student's T distribution** with $n-1$ degrees of freedom.

For large n these statements hold approximately true irrespective of the distribution of X (by CLT).

1. Sample statistics – confidence interval

The PDF for a Student's T-distributed RV for a range of degrees of freedom.



Example: PDF of $X \sim \mathrm{T}(n)$ RV with N(0,1) for comparison

The Student's T-distribution converges to the standard normal distribution as $n \to \infty$.

1. Sample statistics – confidence interval

By calculating the value $t_{1-\alpha/2}$, $0 < \alpha < 1/2$, such that

$$\begin{aligned} \text{Prob}(t_{\frac{\alpha}{2}} \leq T \leq t_{1-\frac{\alpha}{2}}) &= \text{Prob}(-t_{1-\frac{\alpha}{2}} \leq T \leq t_{1-\frac{\alpha}{2}}) \\ &= 1-\alpha \end{aligned} \tag{2}$$

we have by (1)

$$\text{Prob}\left(-t_{1-\frac{\alpha}{2}} \leq \frac{\overline{X}-\mu}{S/\sqrt{n}} \leq t_{1-\alpha/2}\right) = 1-\alpha \tag{3}$$

or with $100(1-\alpha)\%$ confidence

$$\overline{X} - t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}. \tag{4}$$

This **confidence interval** is itself random. If we constructed 100 different samples of X we would expect $100(1-\alpha)$ of the confidence intervals to contain the true mean μ.

1. Sample statistics – R descriptive statistics

```
> summary(health$bmi,na.rm = T)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  15.68   22.71   25.30   25.83   28.20   51.47     177
> var(health$bmi,na.rm = T)
[1] 19.68836
> sd(health$bmi,na.rm = T)
[1] 4.437157
> IQR(health$bmi, na.rm = T)
[1] 5.491445
> library('moments')
> skewness(health$bmi,na.rm = T)
[1] 0.9979072
> kurtosis(health$bmi, na.rm = T)
[1] 4.989826
```

1. Sample statistics – R correlation results

An numerical analysis of dependence between RVs can be generated with the R command

```
> cor.test(health$age, health$newsyst, method = "pearson")

Pearson's product-moment correlation

data:  health$age and health$newsyst
t = 44.7, df = 4998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5143147 0.5539288
sample estimates:
      cor
0.5344152
```

The Pearson correlation coefficient of $+1$ indicates perfect positive (linear) dependence. It will not identify any quadratic dependence.

1. Descriptions of RVs – summary

Graphical tools can be used to describe both categorical and numerical RVs.

Bar charts and pie charts and best reserved for categorical RVs, or discrete RVs that only take a limited number of values. Boxplots and histograms only work for numerical RVs.

Numerical tools provide descriptions via statistics and only make sense when applied to numerical RVs.

Most of these statistics work equally well for both discrete and continuous RVs, although mode is used more frequently to describe discrete RVs.

1. Hypothesis testing – one-sample T-test

Consider a sample $X_1, \ldots, X_n$ of independent observations of some normally-distributed random variable with (unknown) mean μ.

The **one-sample T-test** is used to test a hypothesised value μ^* of μ.

The **null hypothesis** for this test is

$$H_0\colon\ \mu = \mu^*$$

while the **alternative hypothesis** may be any of

$$H_A\colon\ \mu > \mu^* \ (\textbf{upper-tail test})$$
$$H_A\colon\ \mu \neq \mu^* \ (\textbf{two-tail test})$$
$$H_A\colon\ \mu < \mu^* \ (\textbf{lower-tail test}).$$

The **test statistic** is calculated as

$$t^* = \frac{\overline{X} - \mu^*}{S/\sqrt{n}}, \tag{5}$$

with $\overline{X}$ and S^2 the mean and variance of the sample described above.

It can be considered an **observation** of the RV T given in (1).

1. Hypothesis testing – one-sample T-test

Rejection of null hypothesis (upper-tail test)

H_0 is **rejected** in favour of H_A at **significance level** $0 < \alpha < \frac{1}{2}$ if

$$t^* = \frac{\overline{X} - \mu^*}{S/\sqrt{n}} > t_{1-\alpha}$$

where $t_{1-\alpha}$ is the **quantile** satisfying

$$\text{Prob}\big(T > t_{1-\alpha}\big) = \alpha.$$

Equivalently, H_0 is rejected if μ^* falls outside the one-sided $100(1-\alpha)\%$ CI for μ given by

$$\overline{X} - \frac{S}{\sqrt{n}} t_{1-\alpha} \leq \mu < \infty$$

or if the **p-value**

$$p = \text{Prob}(T > t^*) < \alpha.$$

The null hypothesis H_0 is **retained** in any other case.

1. Hypothesis testing – one-sample T-test

Rejection region (two-tail test)

H_0 is **rejected** in favour of H_A at **significance level** $0 < \alpha < \frac{1}{2}$ if

$$|t^*| = \left| \frac{\overline{X} - \mu^*}{S/\sqrt{n}} \right| > t_{1-\alpha/2}$$

where $t_{1-\alpha/2}$ is the **quantile** satisfying

$$\text{Prob}\big(|T| > t_{1-\alpha/2}\big) = \alpha.$$

Equivalently, H_0 is rejected if μ^* falls outside the two-sided $100(1-\alpha)\%$ CI for μ given by

$$\overline{X} - \frac{S}{\sqrt{n}} t_{1-\alpha/2} \leq \mu \leq \overline{X} + \frac{S}{\sqrt{n}} t_{1-\alpha/2}$$

or if the **p-value**

$$p = 2 \times \text{Prob}(T > |t^*|) < \alpha.$$

The null hypothesis H_0 is **retained** in any other case.

1. Hypothesis testing – one-sample T-test

Rejection of null hypothesis (lower-tail test)

H_0 is **rejected** in favour of H_A at **significance level** $0 < \alpha < \frac{1}{2}$ if

$$t^* = \frac{\overline{X} - \mu^*}{S/\sqrt{n}} < t_\alpha$$

where t_α is the **quantile** satisfying

$$\text{Prob}\big(T < t_\alpha\big) = \alpha.$$

Equivalently, H_0 is rejected if μ^* falls outside the one-sided $100(1-\alpha)\%$ CI for μ given by

$$-\infty < \mu \leq \overline{X} + \frac{S}{\sqrt{n}} t_{1-\alpha}$$

or if the **p-value**

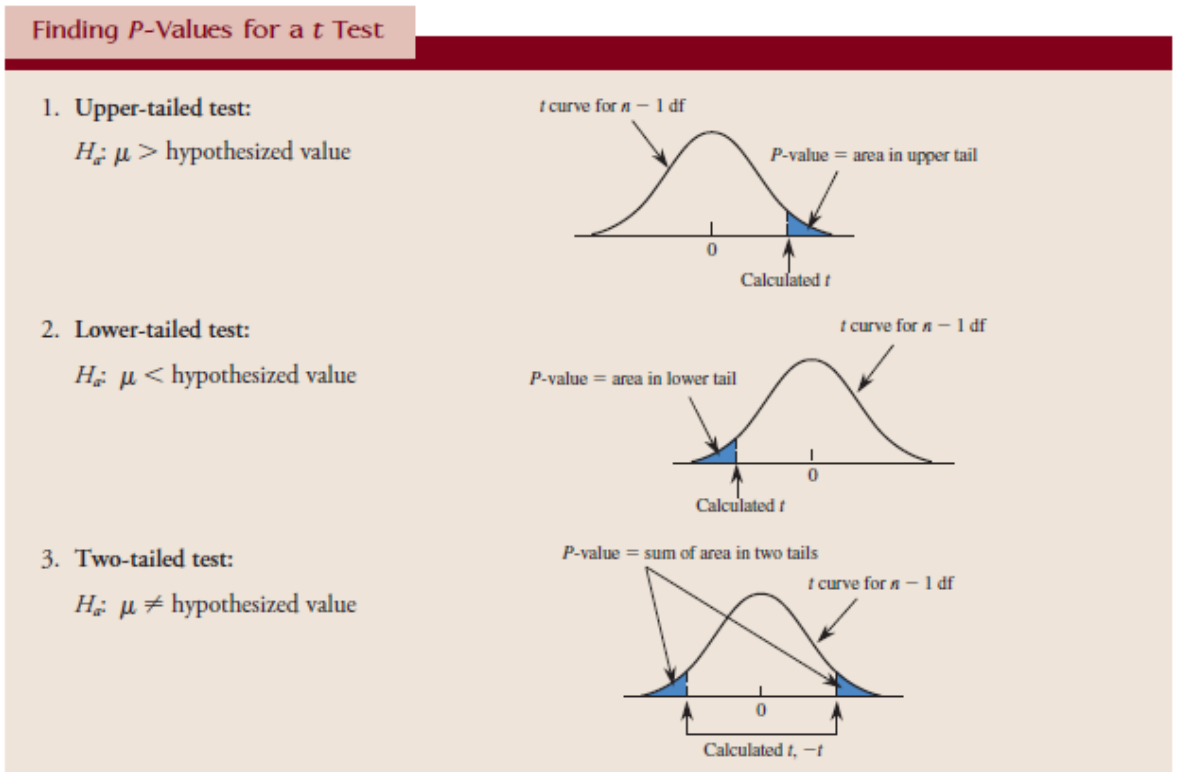
$$p = \text{Prob}(T < t^*) < \alpha.$$

The null hypothesis H_0 is **retained** in any other case.

1. Hypothesis testing – one-sample T-test

The following diagram illustrates the rejection regions for the three versions of this test.

Rejection regions for T-test



Source: Peck et al. (2012), page 604

1. Hypothesis testing – one-sample T-test

As an example, we perform a two-sided one sample T-test on the hypothesised value $\mu^* = 25$ of the (unknown) population mean μ of the RV *bmi* in the data set health.xlsx.

(Assumption of independence sounds reasonable, what about assumption of normality?)

We define the hypotheses

$$H_0: \ \mu = 25$$
$$H_A: \ \mu \neq 25$$

and test at the five-percent significance level ($\alpha = 0.05$).

This test can be conducted using the R command

```
> t.test(health$bmi, mu = 25)
```

1. Hypothesis testing – one-sample T-test

```
One Sample t-test

data:  health$bmi
t = 12.945, df = 4822, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 25
95 percent confidence interval:
 25.70181 25.95232
sample estimates:
mean of x
 25.82707
```

With $p < \alpha$ the null hypothesis H_0 is rejected in favour of H_A.

1. Hypothesis testing – one-sample T-test

It is a simple matter to also perform this test manually, using the following statistics.

```
> length(health$bmi)-sum(is.na(health$bmi))
[1] 4823
> mean(health$bmi, na.rm = T)
[1] 25.82707
> sd(health$bmi, na.rm = T)
[1] 4.437157
```

Using this output we calculate the test statistic (5) with $\mu = 25$ as

$$\frac{\overline{bmi} - \mu^*}{S/\sqrt{n}} = \frac{25.82707 - 25}{4.437157/\sqrt{4823}} = 12.945$$

which gives $p = 0.000$.

1. Hypothesis testing - two-independent-sample T-test

Suppose we have two samples (sizes n_1 and n_2) of independently selected normally-distributed RVs with means μ_1 and μ_2.

We want to test for differences between μ_1 and μ_2.

We set up a two-tailed version of the test with the following null and alternative hypotheses

$$H_0: \ \mu_1 - \mu_2 = \mu_1^* - \mu_2^*$$
$$H_A: \ \mu_1 - \mu_2 \neq \mu_1^* - \mu_2^*$$

with μ_1^* and μ_2^* the hypothesised values of μ_1 and μ_2 respectively.

(Upper and lower-tail versions of the alternative hypothesis can also be specified.)

See Chapter 11 of Peck et al. (2012) for details.

1. Hypothesis testing – two-independent-sample T-test

Let $\overline{X}_1$, $\overline{X}_2$ be the means and S_1^2, S_1^2 the variances of the two samples.

The test statistic is defined differently for the cases of equal and unequal population variances, namely

$$t^* = \begin{cases} \dfrac{\overline{X}_1 - \overline{X}_2 - (\mu_1^* - \mu_2^*)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}, & \sigma_1 \neq \sigma_2 \\ \dfrac{\overline{X}_1 - \overline{X}_2 - (\mu_1^* - \mu_2^*)}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1+n_2)(n_1+n_2-2)}}}, & \sigma_1 = \sigma_2 \end{cases}, \tag{6}$$

and is an observation of a Students's T-distributed RV with degrees of freedom given by

$$\mathrm{df} = \begin{cases} n_1 + n_2 - 2, & \sigma_1 = \sigma_2 \\ \dfrac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}, & \sigma_1 \neq \sigma_2 \end{cases}.$$

The reject/retain decision is analogous to that for the one-sample T-test.

1. Hypothesis testing – two-independent-sample T-test

As an example we test the means for the RV *newsyst* grouped by the categorical RV *gender* against a hypothesised difference of zero.

(What about the assumptions?)

The hypotheses take the form

$$H_0: \ \mu_{\mathsf{m}} - \mu_{\mathsf{f}} = 0$$
$$H_A: \ \mu_{\mathsf{m}} - \mu_{\mathsf{f}} \neq 0.$$

1. Hypothesis testing – two-independent-sample T-test

```
> library(car)
> leveneTest(health$newsyst, as.factor(health$gender))
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    1  61.236 6.181e-15 ***
      4803
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> t.test(health$newsyst~health$gender, var.equal = F)
Welch Two Sample t-test
data:  health$newsyst by health$gender
t = 6.8574, df = 4907.3, p-value = 7.88e-12
alternative hypothesis: true difference in means between group 1 and
group 2 is not equal to 0
95 percent confidence interval:
 2.757764 4.965847
sample estimates:
mean in group 1 mean in group 2
       138.0598        134.1980
```

1. Hypothesis testing – two-independent-sample T-test

From the first table returned we see that Levene's test of the null hypothesis that the population variances are equal is rejected.

So we use the test statistic and p-value from the unequal variances version of the test.

With $p = 0.000$ we reject the null hypothesis H_0 and conclude that there is significant evidence that the population means μ_{m} and μ_{f} are unequal.

We can also manually calculate out test statistic as

$$t^* = \frac{\overline{newsyst}_{\text{m}} - \overline{newsyst}_{\text{f}} - 0}{\sqrt{\frac{(n_{\text{m}}-1)S_{\text{m}}^2+(n_{\text{f}}-1)S_{\text{f}}^2}{(n_{\text{m}}+n_{\text{f}})(n_{\text{m}}+n_{\text{f}}-2)}}} = 6.857.$$

1. Hypothesis testing – F-test

Now let there be m independent samples of some normally-distributed RVs and consider the problem of testing the means $\mu_1, \ldots, \mu_m$.

Specifically, we test the hypothesis that at least one of the means μ_i is different from the others.

Define the following null and alternative hypotheses:

H_0: $\mu_1 = \mu_2 = \cdots = \mu_m$;

H_A: at least one of μ_i differs from the others.

The test we use is an F-test and the procedure is called **analysis of variance (ANOVA)**.

We leave the details until required in our study of regression; here it suffices to say that the test statistic follows an appropriately parameterised F-distribution.

1. Hypothesis testing – F-test

We can run ANOVA with the R command **aov**.

```
> health$marstat <- replace(health$marstat, health$marstat==9, NA)
> res.aov <- aov(health$newsyst ~ as.factor(health$marstat))
> summary(res.aov)
                            Df  Sum Sq Mean Sq F value Pr(>F)
as.factor(health$marstat)    5  225511   45102   124.3 <2e-16 ***
Residuals                 4987 1809242     363
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
7 observations deleted due to missingness

> leveneTest(health$newsyst, as.factor(health$marstat))
Levene’s Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    5  20.229 < 2.2e-16 ***
      4987
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

With a $p = 0.000$ we reject H_0 and conclude that there is significant evidence that at least one of the means differs from the others.

References I

Peck, R., Olsen, C., and Devore, J. (2012). *Introduction to statistics & data analysis*. Brooks/Cole, 4th edition.