Applied stats - quick overview

August 9, 2022

1 Discrete distributions

1.1 Bernoulli distribution

- Notation: Ber(p)
- E[X] = 1*p + 0*(1 - p) = p
- Var(X) = p(1 - p)

1.2 Geometric distribution

- $E[X] = \frac{1}{p}$ (p. 93)

1.3 Binomial distribution

- Any random variable $X \sim Bin(n, p)$ can be represented as $X = R_1 + \cdots + R_n$ where $R_i \sim Ber(p)$ (p. 138)
- E[X] = np (p. 138)
- Var(X) = np(1 - p) (p. 141)
- Let $X \sim Bin(n, p)$ and $Y \sim Bin(m, p)$. Then $Z \sim Bin(n + m, p)$ where $Z = X + Y$ (p. 153)
- For sufficiently large N, binomial distribution can be approximated by **standard normal distribution**, see concrete example on page 201.

2 Continuous distributions

2.1 Uniform distribution

- $E[X] = \frac{1}{2}(b + a)$
- PDF:

$$f(x) = \begin{cases} \frac{1}{\beta-\alpha}, & \text{if } x \in [\alpha, \beta] \\ 0, & \text{otherwise} \end{cases}$$

- CDF:

$$F(x) = \begin{cases} \frac{x-\alpha}{\beta-\alpha}, & \text{if } x \in [\alpha, \beta] \\ 0, & \text{if } x < \alpha \\ 1, & \text{if } x > \beta \end{cases}$$

2.2 Normal distribution

- $E[X] = \mu$ (p. 94)
- $Var(X) = \sigma^2$ (p. 97)
- For $X \sim N(\mu, \sigma^2)$, and $Y = g(X) = rX + s$ for any s and $r \neq 0$, we can write that Y has an $N(r\mu + s, r^2\sigma^2)$ distribution. (p. 106). This is also how you can derive standard normal distribution.
- If X and Y are independent random variables with normal distribution, then their sum $Z = X + Y$ is also normally distributed (p. 158)

2.3 Exponential distribution

- $E[X] = \frac{1}{\lambda}$ (p. 93)
- Sum of n independent random variables $X_i \sim Exp()\lambda$ has distribution $Gam(n, \lambda)$ (p. 157)

2.4 Gamma distribution

- PDF - see page 157

2.5 Cauchy distribution

- PDF - see page 161
- Let X and Y be two random variables from standard normal distribution. Let $Z = \frac{X}{Y}$, then $Z \sim Cau(0, 1)$

3 Useful formulas & concepts - PROBABILITY

3.1 Continuous random variable

- $F(b) = \int_{-\infty}^{b} f(x)dx$
- $f(x) = \frac{d}{dx}F(x)$

3.2 Conditional probability

Bayes theorem

- $P(X|Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{P(Y|X)P(X)}{P(Y)}$
- $P(Y) = P(Y \cap X) + P(Y \cap X^c) = P(Y|X)P(X) + P(Y|X^c)P(X^c)$

3.3 Expectation

- Discrete: $E[X] = \sum_i a_i * P(X = a_i)$ (p. 90)
- Continuous: $E[X] = \int_{-\infty}^{\infty} x * f(x)dx$ (p. 91)
- Both continuous and discrete:
  - Linearity of the Expectation (p. 137):

$$E[\alpha f(X) + \beta g(X) + t] = \alpha E[f(X)] + \beta E[g(X)] + t$$

- If you apply function g on the random variable X, then on page 96, find what happens to expectation (in general). See below a specific case for g:
  - Let $Y = g(X) = rX + s$, then $E[Y] = E[g(X)] = E[rX + s] = rE[X] + s$ (p. 98) This also implies that $E[g(X)] = g(E[X])$ (p. 106 - bottom). In other words, $E[Y] = g(E[X])$. In words, to you can obtain expected value of Y simply by applying $g(X) = rX + s$ on $E[X]$. Note, that this holds only for the above mentioned function g. If the function would be different, it may not apply!
- In case, you apply function g on more than one random variable, for instance $g(X, Y)$, then look on page 136 to see how to compute the expected value of the new random variable Z, obtained such that $Z = g(X, Y)$.
- If X and Y are **independent** random variables then $E[XY] = E[X]E[Y]$
- Let $X_1, \ldots, X_n$ be a sequence of **identically distributed** and **independent** (i.i.d.) random variables with distribution function F, expected value μ and standard deviation σ. Then $E[\overline{X}_n] = \mu$. (p. 182)

3.4 Variance

3.4.1 One variable

- $Var(X) = E[X^2] - (E[X])^2$ (p. 97)
- $Var(rX + s) = r^2 Var(X)$ (p. 98)

3.4.2 More variables

For simplicity, examples will be made on just two random variables X, Y.

- **Covariance:** To some extent expresses how two variables X and Y influence each other. $Cov(X, Y) = E[XY] - E[X]E[Y]$ We speak of three kinds of covariance: positively and negatively correlated, and uncorrelated. For detail, see page 139. Notice from this formula one thing, if $E[XY] = E[X][Y]$, then $Cov(X, Y) = 0$. This happens only if X and Y are independent. (p. 140) Note, that just because two variables are **dependent**, it does not mean that they are **correlated**. There are examples of dependent random variables X and Y which are uncorrelated, i.e., $Cov(X, Y) = 0$. So the bottom line is **If X and Y are independent random variables, then they are uncorrelated.** (On the opposite site, you CAN NOT say: If two random variables X and Y are uncorrelated, then they are independent)
- **Correlation coefficient**
  - Covariance under change of units: $Cov(rX+s, tY+u) = rtCov(X, Y)$. This rule simply shows that covariance value is influenced by change of units and thus, it is not very reliable to look at. (p. 141) For example the choice of unit for a random variable X, e.g., centimeters vs. inches, implies different covariance: $Cov(X_{cm}, Y) \neq Cov(X_{inches}, Y)$
  - Because of the above reason, standardized version of covariance is introduced - **correlation coefficient** - see page 142.
- **Variance of a sum** (p. 140)
  - X and Y are dependent: $Var(X+Y) = Var(X)+Var(Y)+2Cov(X, Y)$
  - X and Y are independent: $Var(X + Y) = Var(X) + Var(Y)$

3.4.3 Variance of an average

To use the below formula, make sure that $\overline{X}_n$ is the average of n independent random variables with the same expectation μ and variance σ^2

- $Var(\overline{X}_n) = \frac{\sigma^2}{n}$ (p. 182)

- Note, that this is an **important implication** for the Law of large numbers. Namely, notice that $SD(\overline{X}_n) = \frac{\sigma}{\sqrt{n}}$. This means that as n grows, SD is decreasing by a factor of $\sqrt{n}$. We know that expected value of $\overline{X}_n$ is μ. So if you put these two information together, you can see that as n grows larger, you have larger and larger probability that you will actually obtain μ via $\overline{X}_n$.

3.5 Change of variable - impact on distribution

This precisely means that if we have a random variable X, and apply on it a function g, we obtain a new random variable Y, such that $Y = g(X)$. (p. 103)

- Discrete: There is no formula, see example on page 103, where they show how to approach such problem.
- Continuous: In general, there is no formula. However, in the book, they show following two specific cases:
  - When $Y = g(X) = \frac{1}{X}$, and you know f_X, then for $y \neq 0$

    $$f_Y(y) = \frac{1}{y^2} f_X(\frac{1}{y})$$

    , usually for $y = 0$, $f(0) = 0$. (p. 105)
  - Let $Y = g(X) = rX + s$ where $r, s \in \mathbb{R}$. Then $F_Y(y)$ is:

    $$F_Y(y) = F_X(\frac{y-s}{r})$$

    and for $f_Y(y)$, it applies that:

    $$f_Y(y) = \frac{1}{r} f_X(\frac{y-s}{r})$$

    , all can be found on page 106.
  - Distribution of **Maximum** and **Minimum**: In this problem, you have several **independent** random variables, which have same distribution function F and you apply on them a function g, which is either max or min. For example, for max: $Z = g(X_1, \ldots, X_n) = max(X_1, \ldots, X_n)$. You are interested, what will be the distribution function of Z. For formulas, see the page 109.

3.6 Sum, product and quotient of two random independent variables

This section assumes that you

- have two independent random variables X and Y
- you are interested what will be the distribution when you sum, multiply or divide the above mentioned random variables

3.6.1 Sum

Discrete random variables

- $p_Z(c) = \sum_c p_X(a_i)p_Y(b_i)$ where $c = a_i + b_i$ (p. 152)

Continuous random variables

- See page 156.

3.6.2 Product and quotient

See page 160 and 161.

3.7 Jensen's inequality

I am not quite sure where to apply this concept apart from proving something. The exact definition can be found on page 107. The key takeaways should be:

- If g is a convex function, see page 107, how to check that, and X is a random variable, then $g(E[X]) \leq E[g(X)]$. In words, this means that you know that if you apply function g on the expected value of X, then the result will be smaller or equal to the expected value of a new variable Y, which you obtained such that $Y = g(X)$.

3.8 Chebyshev's inequality

Chebyshev's inequality gives you the **bound** for a probability that random variable X will be outside the interval $(E[X] - a, E[X] + a)$ where $a > 0$. Here are the important versions of the Chebyshev's inequality derived from the book:

- $P(|X - E[X]| \geq \alpha) \leq \frac{Var(X)}{\alpha^2}$ (p. 183) Description of what this says is in the intro. Note, that this formula gives you an **upper bound** for the probability
- $P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$ (p. 185) This gives you a **lower bound** that realization of random variable X is within k standard deviations from the expected value of X, i.e. μ.

3.9 Joint distribution

When do we talk about a joint distribution? Whenever there is more than one random variable in play. Consider these examples:

- Throw of a dice six times in a row modeled by $X_1, \ldots, X_6$
- Sampling without replacement modeled by $X_1, \ldots, X_n$

Here is a very good summary of this topic. Key takeaway from this article is that in order to be able to compute the joint probability distribution, we need to know whether the variables are independent or dependent. Note, that in many problems in the book, you are already given for example joint CDF, and you are supposed to for example compute joint PDF and marginal PDF, and then based on that determine whether the two variables are dependent or independent. (Lecture 6 - Ex. 3)

3.10 Independence of random variables

Examples will be made for simplicity on two random variables X, Y, but the rules can be then applied to more than two variables indeed.

- **Discrete** random variables - check that for all x, y holds true that $P(X = x, Y = y) = P(X = x)P(Y = y)$. In words, check that joint probability of particular events can be computed by multiplying marginal probabilities of those events.
- **Continuous** random variables - check that for all x, y holds true that $f(x, y) = f(x)f(y)$

3.10.1 Propagation of independence

The idea is very simple, if you have a set of independent random variables X_i, and you transform them using a function h_i, then the resulting set of variables Y_i will be also independent. The core rule here for independence is that all Y must be based on X.

Let's say that for instance, you have a set of random variables $X \sim U(0, 1)$. You also have a function $g(X) = 2X^2$. Then, for every even realization of X, i.e., X_{2i}, you will produce set of random variables T such that $T_i = g(X_{2i})$. Similarly, for every odd realization of X, you will produce set of random variables S such that $S_i = g(Y_{2i+1})$. Then, you know that $S_1, T_1, S_2, T_2, ...$ is a set of **independent** random variables.

3.11 The law of large numbers

It says that as n goes to infinity, the probability that difference between average of n i.i.d. random variables X and the true expected value of X, i.e., μ, is 0.

- $\lim_{n\to\infty} P(|\overline{X}_n - \mu| > \epsilon) = 0$. (p. 186) Note that this a weak version of a law of large numbers.
- Strong version is: $P(\lim_{n\to\infty} \overline{X}_n = \mu) = 1$ (p. 187)

3.11.1 Application of the law of large numbers

The book says that you can derive any property of probability distribution of X using the Law of large numbers. The book makes following examples:

Recovering probability of an event

You have a sample of i.i.d. random variables $X_1, \ldots, X_n$, and you want to know $p = P(X = C)$ where C is an event. In order to get a valid estimate for p, you can count number of times X hits the event C within the sequences and divide that by n. The larger the sample is, the better chance you have that the estimate will be accurate.

Recovering the probability density function

If you have a continuous random variable, then the procedure to obtain an estimate for probability of an event is same as above. But even more important here is to realize that if you use **histogram**, then this is actually a good estimation of the underlying distribution. (page 190)

3.12 The central limit theorem

Central limit theorem serves you as a tool to **approximate** probability distribution of either:

- **Averages** of i.i.d. random variables X, i.e., distribution of $\overline{X}_n$. The following equation is important here: $Z_n = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}$. (p. 197)
- **Sum** of i.i.d. random variables X. Here, following equation is important $Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$.

Note that as $n \to \infty$, $Z_n \sim N(0,1)$ and that is why central limit theorem is useful. On the other hand, have to be critical about the approximations you get. One factor is the size of n. Another factor is for example whether the distribution where X comes from is discrete, since then you ary trying to approximate discrete distribution by continuous.

4 Useful formulas and concepts - STATISTICS

4.1 Bootstrap principle

There are two kinds of bootstrap:

- **Empirical bootstrap**: Use when you are given data from unknown distribution F. Here is an application example from the class. You decided to estimate μ of F using the sample mean, i.e., $\overline{x}_n$.The problem with this approach is that if you got a new sample from F, your approximation would be different. Therefore, by using bootstrap, you simulate the distribution of **centered mean** which tells you how likely it is that other means will differ by a certain distance. Note, you have to use centered mean as this has been proved to be unbiased.

- **Parametric bootstrap**: When you use empirical bootstrap, you approximate F by F_n. In this case, however, you approximate F by $\hat{F}_\theta$. What is $\hat{F}_\theta$? Let's say you looked at histogram of sample data, and they seem to be normally distributed. Therefore, based on the sample, you approximate μ by $\overline{x}_n$ and σ^2 by S_n^2, and then you will bootstrap the sample using $\hat{F}_\theta = N(\overline{x}_n, S_n^2)$

4.2 Unbiased estimators

4.2.1 Unbiased estimators for mean and variance

Estimators is unbiased if $E[T] = \theta$ where θ is the parameter of our interest. The difference $E[T] - \theta$ is called *bias*.

- Mean of the sample and variance of the sample are UNBIASED estimators for the expectation and variance of the **true** distribution. **NOTE:** There is not unbiased estimator for standard deviation.

- Unbiasedness does not carry over. If T is an unbiased estimator, it does NOT necessary mean that g(T) will be also unbiased estimator. There is once exception to this: $g(T) = aT + b$ for $a, b \in R$

4.2.2 Maximum likelihood estimator

For context, maximum likelihood estimator is important since it has some very useful properties which makes is well applicable in practice. Namely, it is

- Asymptotically unbiased

- Asymptotically minimum variance

4.2.3 The method of least squares

Here is a typical setting:

- You are given a bi-variate data set (x_i, y_i) such that X_i is a non-random variable and Y_i is a random variable.
- $Y_i = \alpha + \beta x_i + U_i$ where U_i is a random variable with zero expectation and variance σ^2. (represent a simple linear regression model)
- Your goal is to estimate parameters α, β, σ^2 such that we come up with a line that best fits the points. To do so, you can use **least square method**
- Finally, notice that since you have no knowledge regards to the distribution of Y_i, it is not possible to use maximum likelihood estimator.

Least squares method - formula
You want to find parameters α, β such that:

$$S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

is minimal. (p. 329)

Least squares method - unbiased estimators
In the book on page 331, it is then derived that **unbiased** estimators for α, β are:

$$\hat{\alpha} = \overline{Y}_n - \hat{\beta}\overline{x}_n \quad \hat{\beta} = \frac{n\sum x_i Y_i - (\sum x_i)(\sum Y_i)}{n \sum x_i^2 - (\sum x_i)^2}$$

In addition, unbiased estimator for σ^2 is:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

4.3 Confidence interval

When using estimators for estimating different features of a distribution from a sample, we give a single number. The idea of confidence intervals is to give interval within which we are confident at a certain level that the feature of interest is. Things to be aware of:

- You can say only: we are 100γ % (e.g. 95 %) **confident** that $\theta \in$ *some interval*
- It would be incorrect to say that with probability γ, $\theta \in$ *some interval.* This is because θ is just a number not a random variable, thus we can not speak about probability. (See more precise info on page 342)
- See precise definition of confidence intervals on page 343.

4.3.1 Find two sided confidence interval for the mean of sample

For the later use, it is useful to know the definition of **studentized mean of a normal random sample**:

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}$$

Based on the definition in the book from page 349, if $X \sim N(\mu, \sigma^2)$, then studentized mean, as defined above, has $t(n-1)$ distribution regardless of the values of μ, σ.

Normal distribution: known variance
The formula is as follows (p. 347):

$$(\overline{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

where $\alpha = 1-\gamma$. And z_p can be computed as $P(Z \geq z_p) = p$ where $Z \sim N(0,1)$. In R, you can use the following command to compute $z_{\alpha/2}$:

```
qnorm(a/2, mean = 0, sd = 1, lower.tail = FALSE)
```

Note, here you make an assumption that α is between both tails (lower, upper) distributed equally. In more general terms, the above formula is:

$$(\overline{x}_n - c_u\frac{\sigma}{\sqrt{n}}, \overline{x}_n - c_l\frac{\sigma}{\sqrt{n}})$$

where c_l and c_u are lower and upper critical values respectively. Let's say $\alpha = 0.05$, but you want to obtain lower critical value such that $c_l = P(Z \leq z) = 0.01$ and upper critical value such that $c_u = P(Z \geq z) = 0.04$. In this case, you can use R as follows:

```
c.l = qnorm(0.01, mean = 0, sd = 1, lower.tail = TRUE)
c.u = qnorm(0.04, mean = 0, sd = 1, lower.tail = FALSE)
```

Normal distribution: UN-known variance
The formula is as follows:

$$(\overline{x}_n - t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}}, \overline{x}_n + t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}})$$

There are following two differences compare to the previous formula:

- Instead of σ, you use s_n (standard deviation of the given sample)
- Because of the above change, you need to obtain **critical values** from a **t-distribution**, not standard normal distribution. Why **t-distribution**? Because above formula was computed using studentized mean, see the above definition of it and implications regards to its distribution.

To obtain critical values where $\alpha = 0.05$ and we split α evenly between the tails, you can write in R to obtain $t_{n-1,\alpha/2}$:

```
qt(a/2, df = n - 1, lower.tail = FALSE)
```

In the case, where we decide to distribute α not evenly within the tails (assume the similar example as in previous section), we can write:

```
c.l = qt(0.01, df = n - 1, lower.tail = TRUE)
c.u = qt(0.04, df = n - 1, lower.tail = FALSE)
```

and then use it in the equation:

$$(\overline{x}_n - c_u \frac{s_n}{\sqrt{n}}, \overline{x}_n - c_l \frac{s_n}{\sqrt{n}})$$

Any distribution using bootstrap
Use this method if:

- You doubt the normality of data (or just can not assume the data is normal)
- The sample size is not large enough

Compare to previous case, we no longer know, what is the distribution of **studentized mean** since we can not assume the normality. For this reason, we can use bootstrap to approximate distribution of **studentized mean** and then via this distribution obtain **critical values**. (see the procedure on page 351)
Note, **advantage of bootstrap** compare to the previous two methods is that it adapts to the shape of distribution and it reflects that in the obtained intervals. See the page 353 for more detail.

Any distribution using assumption about large sample size
If you can assume that your sample size is large enough, you can use the central limit theorem which states that as n goes to infinity, distribution of **studentized mean** can be approximated by $N(0, 1)$. What is large enough? There is no rule. You can obtain the confidence interval as follows:

$$(\overline{x}_n - z_{\alpha/2} \frac{s_n}{\sqrt{n}}, \overline{x}_n + z_{\alpha/2} \frac{s_n}{\sqrt{n}})$$

where $\alpha = 1-\gamma$. And z_p can be computed as $P(Z \geq z_p) = p$ where $Z \sim N(0, 1)$. In R, you can use the following command to compute $z_{\alpha/2}$:

```
qnorm(a/2, mean = 0, sd = 1, lower.tail = FALSE)
```

4.3.2 Find ONE sided confidence interval for the mean of sample

Essentially, the only difference between this sub chapter and the previous one is that here we are interested in **one sided confidence interval**, i.e., we need to compute only one **critical value**. So for clarity, here are the concrete examples for above mentioned cases:

- **Normal distribution: known variance**

$$\text{Lower: } (\overline{x}_n - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty) \text{ Upper: } (-\infty, \overline{x}_n + z_\alpha \frac{\sigma}{\sqrt{n}})$$

- **Normal distribution: UN-known variance**

$$\text{Lower: } (\overline{x}_n - t_{n-1,\alpha} \frac{s_n}{\sqrt{n}}, \infty) \text{ Upper: } (-\infty, \overline{x}_n + t_{n-1,\alpha} \frac{s_n}{\sqrt{n}})$$

- **Any distribution: bootstrap**

$$\text{Lower: } (\overline{x}_n - c_u^* \frac{s_n}{\sqrt{n}}, \infty) \text{ Upper: } (-\infty, \overline{x}_n - c_l^* \frac{s_n}{\sqrt{n}})$$

4.3.3 Find sample size needed to obtain confidence interval of given width w

- **Data comes from a normal distribution with known variance** σ^2:

$$n \geq (\frac{2z_{\alpha/2}\sigma}{w})^2$$

- **Data comes from a normal distribution with unknown variance** σ^2:

4.4 Testing hypothesis: General approach

This section covers the general intro and all important terms for hypothesis testing to which we were introduced in lecture 12.

P-value
= probability of the given event itself + probability of events which are equally likely + probability of events that are even rarer (more extreme)

Critical region and critical values

- To avoid computing P-value, you can compute critical region based on the **significance** level. You want to find value(s) c such that $P(T \geq c)$ or $P(T \leq c)$ is equal to the significance level where T is your test statistic.

Interpretation of P-value
See Verzani page 296, 2nd paragraph, it is explained really well how to interpret the results.

Type I and type II error

- **Type I error**: you decide to reject null hypothesis, you made either a correct decision or type I error. You can control how likely you are to make type I error by setting a **significance level** α. Note, that you never know whether you made the error, you just know using the significance level, how likely this is to happen. In other words, when you conclude to reject the null hypothesis, there is still α chance that you made a wrong decision.

- **Type II error**: you decide not to reject null hypothesis, you made either a correct decision or type II error. In this case, there is no way how you can control how large the type II error will be. Why? Since it depends on the parameters of test statistic which are in turn dependent on alternative hypothesis. Recall, that alternative hypothesis is usually something like $\mu > 120$. So there are many μ and thus many possible ways for probability of type II error.

4.5 Testing hypothesis: Significance test for population proportion

Here is the setting of the problem:

- You are given a data set which represents a random sample from $X_i, \ldots, X_n$
- You are given some known proportion, denote this by p_0 (e.g. approval rate of a politician)
- You design your hypothesis as follows: $H_0 : p = p_0$ where p represents the new unknown proportion. Via your null hypothesis, you assume that old proportion is the same as the new one. As your alternative hypothesis, you can choose $H_1 : p \neq p_0$, $p < p_0$ or $p > p_0$.

4.6 Testing hypothesis: T-test

4.6.1 One sample

Assume, you have a sample which is a realization of random variables $X_1, \ldots, X_n$. Your null hypothesis is $H_0 : \mu = \mu_0$ and alternative hypothesis is $H_1 : \mu \neq \mu_0$. Usually the question is framed like *Is the true value of the distribution from which sample comes equal to* μ_0? Your test statistic T is given by $T = \frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}}$ (this is **studentized mean**) Choose appropriate method based on whether the data comes from normal distribution or not:

- **Normal data: t-test** - In the case where $X \sim N(\mu, \sigma^2)$, you know that $T \sim t(n-1)$ (p. 349).
- **Non - normal data: SND or Bootstrap** In this case you have two options:
  - If the sample size is large, i.e., n is sufficiently large, you can approximate the distribution of T by **standard normal distribution**: $T \sim N(0, 1)$

- Bootstrap method - to approximate the distribution of T, use the bootstrap method described on page 351. (Or see lecture 11 - exercise 6)

4.6.2 Two samples

Here are the assumptions for the examples below:

- You have two samples: $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$
- Each sample has its own distribution: F_X, F_Y, with expected values μ_X and μ_Y
- Your null hypothesis is $H_0 : \mu_X = \mu_Y$ and alternative hypothesis $H_1 : \mu_X \neq \mu_Y$. In words, you want to test, whether the two distributions have same expected values. Alternative hypothesis can be also $H_1 : \mu_X < \mu_Y$ or $H_1 : \mu_X > \mu_Y$.
- Both variances σ_X^2 and σ_Y^2 are unknown.
- Your test statistic is $T = \overline{X}_n - \overline{Y}_m$. Note, that T needs to be standardized. This will be done and explained in the below sections based on whether the variances are or are not equal.

Samples have equal variances
Since samples have the same variances, you will standardize the T by *pooled* standard deviation S_p. It is important to notice that T_p represents **pooled studentized mean difference** See page 417 for the formula.

Normal data - 2 sample t-test

- Based on the normality assumption, you know that $T_p \sim t(n+m-2)$ (p. 417)

Non-Normal data - bootstrap

- You can NOT longer assume that $T_p \sim t(n+m-2)$ and for this reason, you need to approximate the distribution of T using a **bootstrap** method which is described on page 418.

Samples have unequal variances
Since samples have DIFFERENT variances, you will standardize T by *non-pooled* standard deviation S_d. It is important to notice that T_d represents **non-pooled studentized mean difference** See page 420 for the formula.

Normal data - Welsch's test

Based on Sami's slide you can write $T_d \sim t(v)$ where

$$v = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{S_X^4}{n^2(n-1)} + \frac{S_Y^4}{m^2(m-1)}}$$

Non-Normal data - bootstrap

Unfortunately, in this case, the only option is to **approximate using bootstrap method** distribution of T_d. The procedure is same as in previous section, the only difference is that now at each iteration you compute non-pooled studentized mean difference.

Large samples

If n and m are large enough, then you can approximate both T_p and T_d by standard normal distribution. Indeed, you have to first decide whether the samples have equal or different variance, but after that, especially in cases, where you need to bootstrap the distribution, it is useful to use this method instead assuming the sample sizes are large enough.