Statistical Inference

Bhaskar

26/12/2020

Introduction to statistics and Data Analysis by Prof Christian Heumaan

Distribution	Example
Uniform	Rolling a die (discrete) Waiting for a train (continuous)
Bernoulli	Any binary variable such as gender
Binomial	Number of "heads" when tossing a coin n times
Poisson	Number of particles emitted by a radioactive source entering a small area in a given time interval
Multinomial	Categorical variables such as "party voted for"
Geometric	Number of raffle tickets until first ticket wins
Hypergeometric	National lotteries; Fisher's test
Normal	Height or weight of women (men)
Exponential	Survival time of a PC
2	Sample variance; 2 tests
t	Confidence interval for the mean
F	Tests in the linear model

Statistical tests

Case	Statistical Test	R Function
Test for the Mean When the Variance is Known	One-Sample Gauss Test	z.test{BSDA}, zsum.test{BSDA}
Test for the Mean When the Variance is Unknown	One-Sample T test	t.test{stats}, tsum.test{BSDA}, t_test{infer}, t_test{mosaic}
Comparing means of two groups (Two independent Samples, Variances are known)	Two Sample Gauss test	z.test{BSDA}, zsum.test{BSDA}
Comparing means of two groups (Two independent Samples, Variances are unknown)	Two Sample T test	t.test{stats}, tsum.test{BSDA}, t_test{infer}
Comparing the Means of Two Dependent Samples (paired data)	Paired t-Test	t.test with argument paired = TRUE
Comparing means across more than two groups	ANOVA test	

Case	Statistical Test	R Function
Test for the variance	One-Sample Chi-Squared Test on Variance	varTest{EnvStats}
Comparing variances of two populations	F test (with normality assumptions), Bartlett tests & Levene tests (for non normal distributions)	var.test{stats},bartlett.test{stats}, levene{car}

Sample Size for a One- or Two-Sample t-Test

tTestN {EnvStats}
Compute the sample size necessary to achieve a specified power for a one- or two-sample t-test, given the scaled difference and significance level.

Introduction to statistical thinking in R without calculus

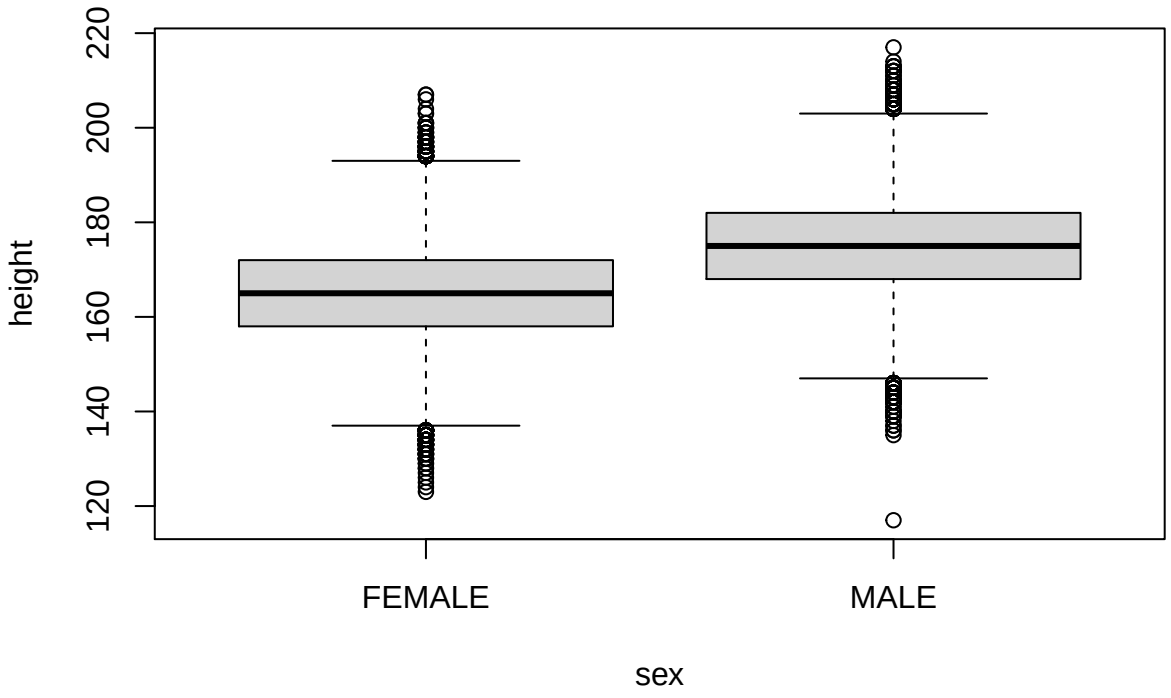
```
filepath <- paste("http://pluto.huji.ac.il/~msby/StatThink/Datasets/pop1.csv")

library(readr)
pop1 <- read.csv(filepath)
tibble::glimpse(pop1)
```

```
## Rows: 100,000
## Columns: 3
## $ id     <int> 5696379, 3019088, 2038883, 1920587, 6006813, 4055945, 926326...
## $ sex    <chr> "FEMALE", "MALE", "MALE", "FEMALE", "MALE", "FEMALE", "FEMAL...
## $ height <int> 182, 168, 172, 154, 174, 176, 193, 156, 157, 186, 143, 182, ...
```

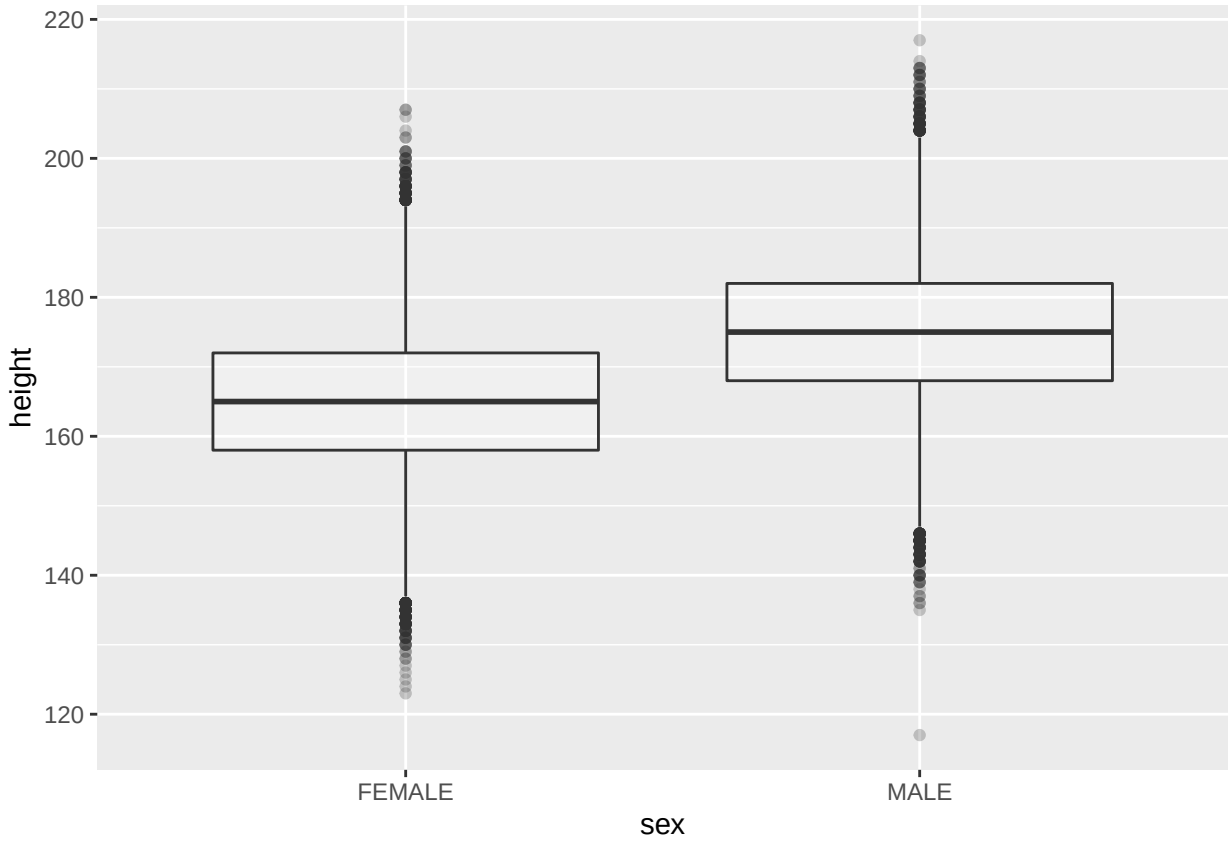
```
summary(pop1)
```

```
##        id              sex                height
##  Min.   :1000082   Length:100000      Min.   :117
##  1st Qu.:3254220   Class :character   1st Qu.:162
##  Median :5502618   Mode  :character   Median :170
##  Mean   :5502428                      Mean   :170
##  3rd Qu.:7757518                      3rd Qu.:178
##  Max.   :9999937                      Max.   :217
```

```
boxplot(height ~ sex, data = pop1 )

library(ggplot2)
```

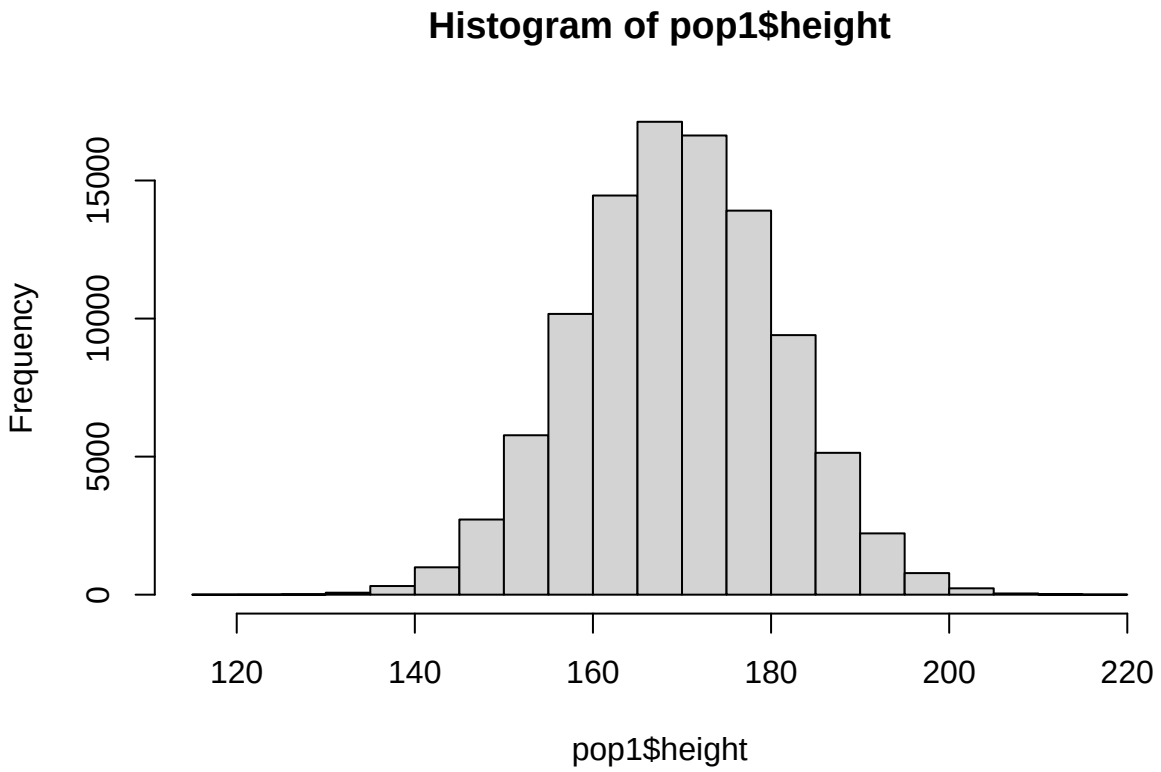



```
ggplot(pop1, aes(x = sex, y = height))+
  geom_boxplot(alpha = 0.25)
```

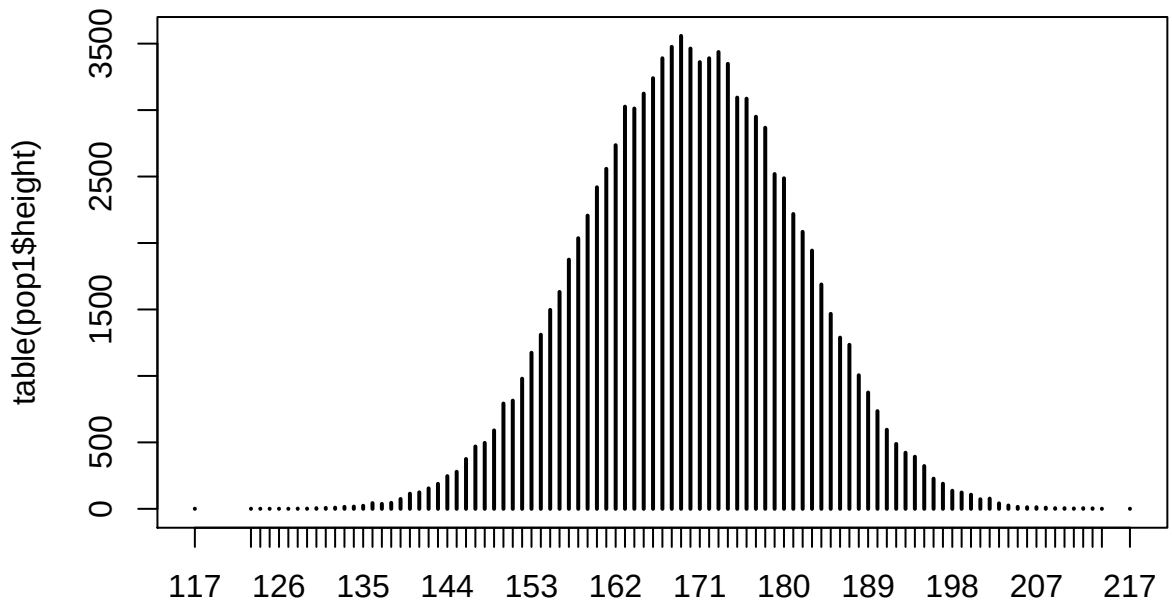


Histogram of the population

```r
hist(pop1$height)
```




```r
plot(table(pop1$height))
```



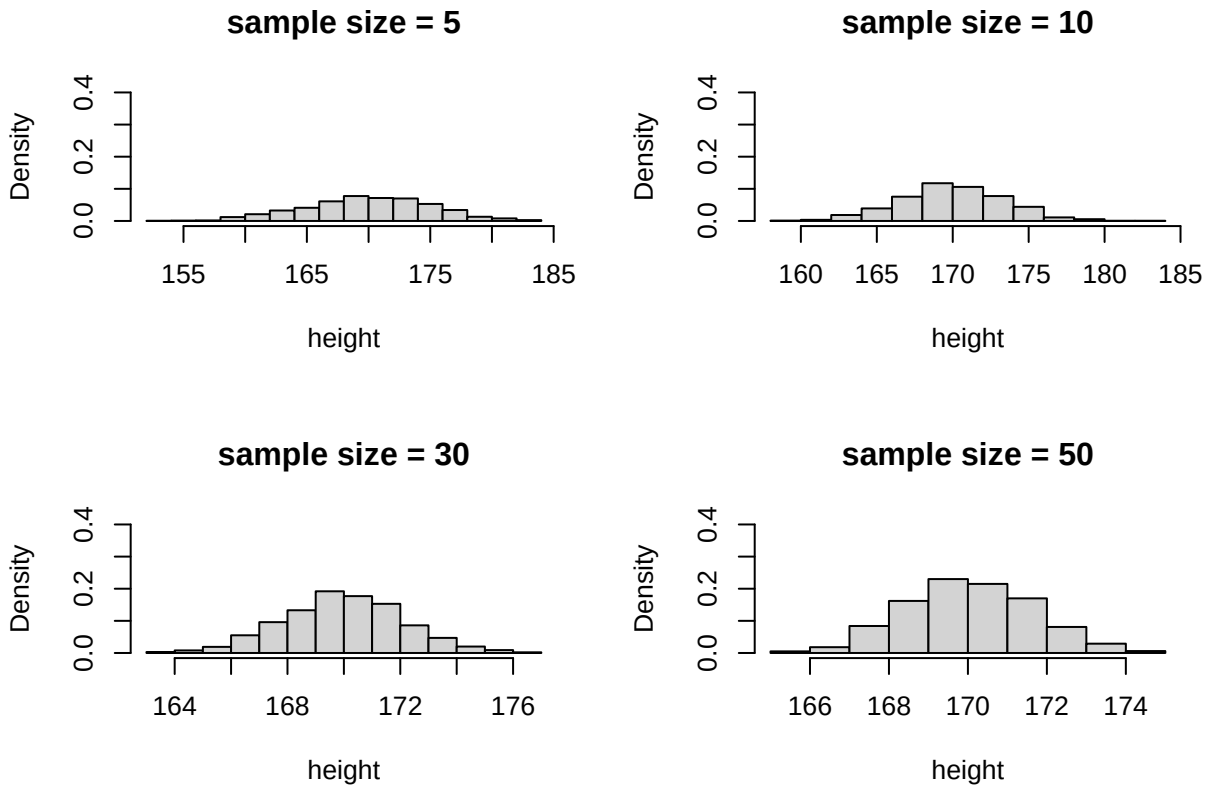

```r
#https://stackoverflow.com/questions/32712301/create-empty-data-frame-with-column-names-by-assigning-a-
sample_dist <- data.frame(matrix(nrow = 1000, ncol = 4))
colnames(sample_dist) <- c("size_5","size_10","size_30","size_50")
size <- c(5,10,30,50)
```

```r
#https://laurakgray.weebly.com/uploads/7/3/6/2/7362679/20_-_for_loops_in_r.pdf


for (i in 1:4)
{ for (j in 1:1000)
  {sample_dist[j,i] <- mean(sample(pop1$height,size[i]))}
}

head(sample_dist)
```

```
##   size_5 size_10  size_30 size_50
## 1  166.8   172.7 167.7333  167.10
## 2  177.6   173.6 172.0333  171.70
## 3  171.8   172.3 168.9333  173.34
## 4  158.6   164.3 169.5333  168.90
## 5  170.2   169.5 166.0000  171.76
## 6  168.4   164.7 169.3000  171.38
```

Using for-loops to plot the sampling distribution

```r
par(mfrow = c(2,2))
for (i in c(5,10,30,50))
{
 hist(eval(parse(text = (paste0("sample_dist$size_",i)))),
      main = paste0("sample size = ", i), prob=TRUE, ylim=c(0,0.4), xlab="height")
}
```

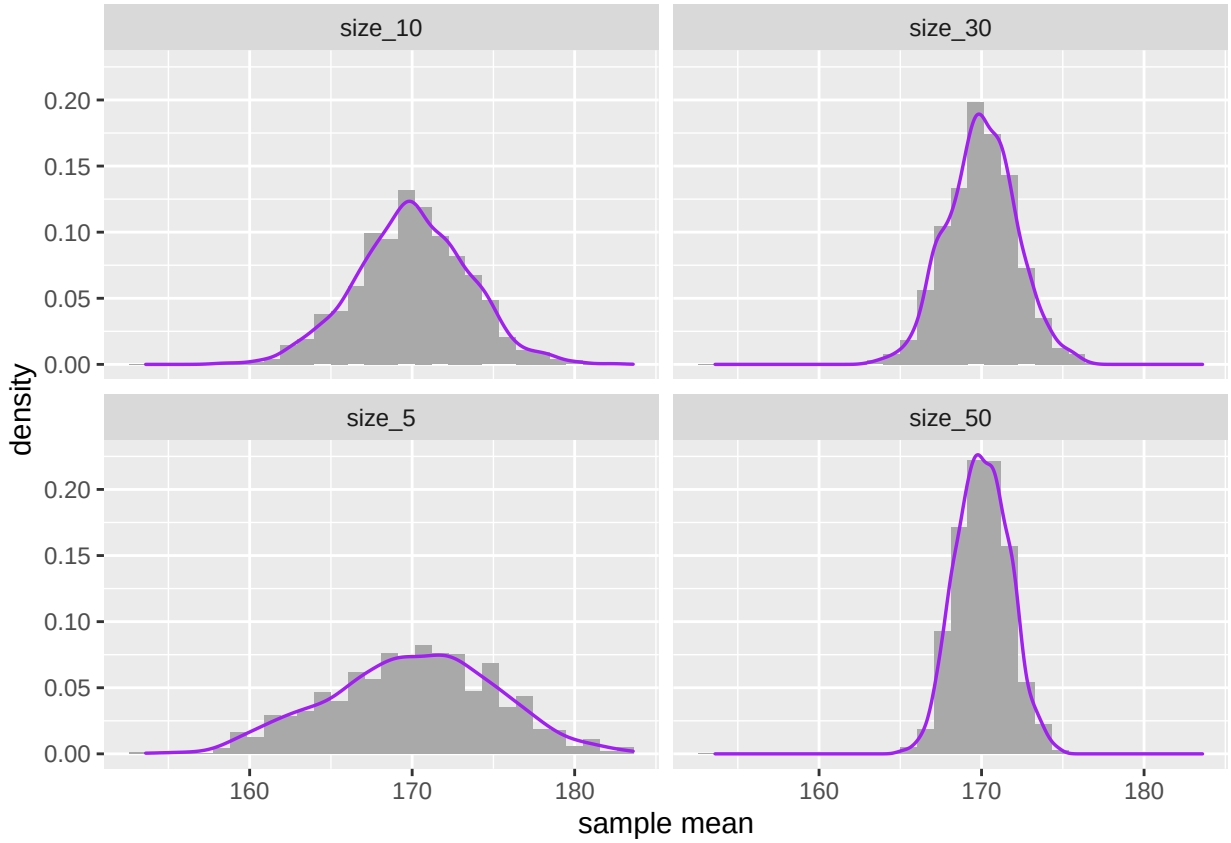


```
# hist(sample_dist$size_5)
```

Using ggplot with facet wrap to plot the sampling distributions of different sample sizes

Tidying the data

```
library(tidyr)
library(ggplot2)
sample_dist %>%
gather(key = "sample size", value = "sample mean", c(size_5:size_50))%>%
ggplot(aes(x = `sample mean`))+
geom_histogram(aes(y = ..density..),bins = 30, fill = "dark grey")+
geom_density(col = "purple", size = .6)+
facet_wrap(~`sample size`)
```




```
#http://rstudio-pubs-static.s3.amazonaws.com/374857_5a23bad9783a43c1b102aa80aa5c1a7c.html
```

```
filepath <- paste("http://pluto.huji.ac.il/~msby/StatThink/Datasets/pop2.csv")
library(readr)
pop.2 <- read.csv(filepath)
tibble::glimpse(pop.2)
```

```
## Rows: 100,000
## Columns: 7
## $ id        <int> 3695908, 5778095, 5138370, 2109892, 4132609, 9681961, 495...
## $ sex       <chr> "FEMALE", "FEMALE", "MALE", "FEMALE", "FEMALE", "MALE", "...
## $ age       <int> 34, 33, 32, 35, 34, 29, 29, 34, 45, 32, 38, 36, 39, 29, 4...
## $ bmi       <dbl> 28.78903, 18.91321, 27.66339, 26.30668, 21.78160, 28.9040...
## $ systolic  <dbl> 112.5887, 122.9261, 128.3985, 124.0975, 121.3278, 128.652...
## $ diastolic <dbl> 64.84949, 78.71555, 86.57248, 79.18808, 78.51906, 85.0172...
## $ group     <chr> "NORMAL", "NORMAL", "NORMAL", "NORMAL", "NORMAL", "NORMAL...
```

```
summary(pop.2)
```

```
##        id              sex                 age              bmi
##  Min.   :1000050   Length:100000      Min.   :20.00   Min.   : 9.986
##  1st Qu.:3227516   Class :character   1st Qu.:32.00   1st Qu.:22.081
##  Median :5479268   Mode  :character   Median :35.00   Median :24.819
##  Mean   :5482739                      Mean   :34.98   Mean   :24.984
##  3rd Qu.:7721878                      3rd Qu.:38.00   3rd Qu.:27.704
##  Max.   :9999889                      Max.   :54.00   Max.   :46.232
##     systolic        diastolic         group
##  Min.   : 73.37   Min.   : 24.77   Length:100000
##  1st Qu.:116.33   1st Qu.: 72.62   Class :character
##  Median :124.64   Median : 81.27   Mode  :character
##  Mean   :125.02   Mean   : 81.67
##  3rd Qu.:133.22   3rd Qu.: 90.30
##  Max.   :191.65   Max.   :152.34
```

Our goal in this question is to investigate the sampling distribution of the sample average of the variable "bmi". We assume a sample of size n = 150.

1. Compute the population average of the variable "bmi".

```
mean(pop.2$bmi)
```

```
## [1] 24.98446
```

2. Compute the population standard deviation of the variable "bmi".

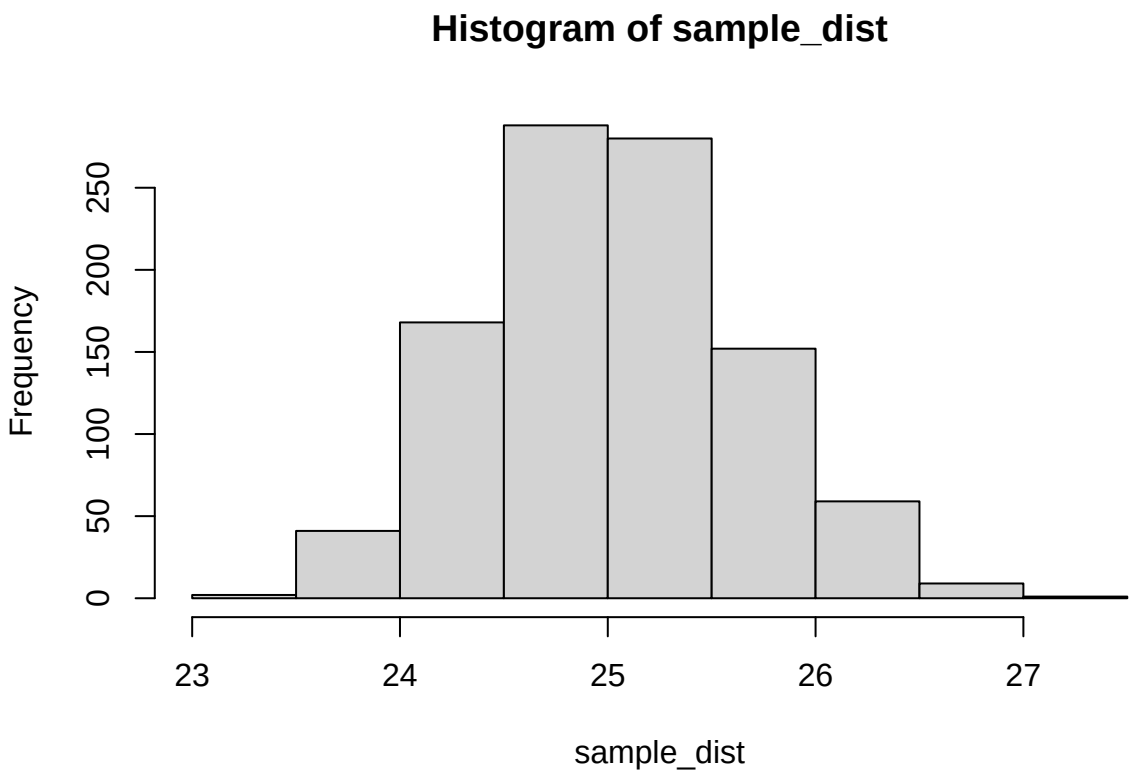
```
sd(pop.2$bmi)
```

```
## [1] 4.188511
```

3. Compute the expectation of the sampling distribution for the sample average of the variable.

Creating a simulation to produce (an approximation) of the sampling distribution of the sample average.

```
sample_dist <- c()
for (i in 1:1000){
  rsample = sample(pop.2$bmi,50)
  sample_dist[i] = mean(rsample)
}

hist(sample_dist)
```




```
mean(sample_dist)
```

```
## [1] 25.01473
```

4. Compute the standard deviation of the sampling distribution for the sample average of the variable.

```
sd(sample_dist)
```

```
## [1] 0.6216151
```

5. Identify, using simulations, the central region that contains 80% of the sampling distribution of the sample average.

```
quantile(sample_dist, probs = c(0.1,0.9))
```

```
##      10%      90%
## 24.21630 25.82594
```

6. Identify, using the Central Limit Theorem, an approximation of the central region that contains 80% of the sampling distribution of the sample average.

```
Llimit <- qnorm(0.1, mean = mean(sample_dist), sd = sd(sample_dist))
ULimit <- qnorm(0.9, mean = mean(sample_dist), sd = sd(sample_dist))

print(paste("(",Llimit,",", ULimit,")"))
```

```
## [1] "( 24.2180947713497 , 25.8113582768941 )"
```

```
qnorm(c(0.1,0.9), mean(sample_dist), sd(sample_dist))
```

```
## [1] 24.21809 25.81136
```

Question 7.2. A subatomic particle hits a linear detector at random locations. The length of the detector is 10 nm and the hits are uniformly distributed. The location of 25 random hits, measured from a specified endpoint of the interval, are marked and the average of the location computed.

1. What is the expectation of the average location?

This is a uniform distribution with hits equally distributed across the length of linear detector. The exepectation of average location will be at the center of the detector.

2. What is the standard deviation of the average location?
3. Use the Central Limit Theorem in order to approximate the probability the average location is in the left-most third of the linear detector.

4. The central region that contains 99% of the distribution of the average is of the form 5 $\pm$ c. Use the Central Limit Theorem in order to approximate the value of c.

Moments

{Introduction to Probability, Second Edition - Joseph K. Blitzstein and Jessica Hwang, Chapter 6}

The nth moment of an r.v. X is E(X_n).

A useful way to study a distribution is via its moments. The first 4 moments are widely used as a basis for quantitatively describing what the distribution looks like, though many other descriptions are also possible. In particular, the first moment is the mean, the second central moment is the variance, the third standardized moment measures skew (asymmetry), and the fourth standardized moment minus 3 is a measure of how heavy the tails are.

Moment generating functions

https://bookdown.org/probability/beta/moment-generating-functions.html

8.3.3 Example 3

Suppose that a market research analyst for a cellular phone company conducts a study of their customers who exceed the time allowance included on their basic cellular phone contract. The analyst finds that for those customers who exceed the time included in their basic contract, the excess time used follows an exponential distribution with a mean of 22 minutes. Consider a random sample of 80 customers and find
1. The probability that the average excess time used by the 80 customers in the sample is longer than 20 minutes.
2. The 95th percentile for the average excess time for samples of 80 customers who exceed their basic contract time allowances.

```r
pexp(q = 20, rate = 1/22,lower.tail = FALSE)
```

```
## [1] 0.4028903
```

5 core statistics concepts that show up in data science interviews:
{Eric Weber}

1. T-tests. Know their assumptions. Know how the t distribution relates to the normal distribution.
2. Central Limit Theorem. Know what it means for the distribution of the mean. Understand the magic!
3. Regression assumptions. Know why independence of observations matters and what IID means.
4. Confidence intervals and relationship to hypothesis tests. Discuss relationship to credible intervals.
5. The implications of Bayesian priors vs frequentist perspective. What does it mean for practical decision making?

When do we need to use these different types of statistical tests?
I realize the importance of understanding the relationships between variables before building predictive models, so here are some layman descriptions for stat tests:
Pearson Correlation
The strength of the association between 2 continuous variables.
Chi-Square Test
The strength of the association between two categorical variables.
Spearman Correlation
The strength of the association between two ordinal variables. (Does not rely on normally distributed data)
ANOVA Test
The difference between group means after any other variance in the outcome variable is accounted for.
Paired T-Test
The difference between two variables from the same population. (ex: a pre and post-test score)
Independent T-Test
The difference between the same variable from different populations. (ex: comparing boys to girls)
Simple Regression
How change in the predictor variable predicts the level of change in outcome variable.
Multiple Regression
How changes in the combination of two or more predictor variables predict the level of change in outcome variables.

Comparison of numeric data
A. Parametric Tests
– T tests
– Dependant (paired)
– independent (student's t test)
– One Way ANOVA

B. Non Parametric Tests
– Wilcoxon signed rank test (unpaired Student's t-test)
– Mann Whitney U test (paired Student's t-test)
– Kruaskal Wallis Test (nonparametric equivalent of ANOVA)

Normality Tests
In statistics, normality tests are used to determine if a data set is well-modeled by a normal distribution and to compute how likely it is for a random variable underlying the data set to be normally distributed.

Kolmogorov-Smirnov (K-S) normality test and Shapiro-Wilk's test. Prerequisite testing for normallity before many of the other tests can be used. Very important to use preleminary tests to make sure that the test assumptions are met. All tests assume some certain characteristics about the data. Some tests require the

data to follow a normal distribution or Gaussian distribution, others are used when the data is skewed, etc. Failure to check that data meets the prerequisites can (will) result in false results which won't be noticed.

If the assumptions for parametric tests are not met, there are nonparametric alternatives for comparing data sets. These include Mann–Whitney U-test as the nonparametric counterpart of the unpaired Student's t-test, Wilcoxon signed-rank test as the counterpart of the paired Student's t-test, Kruskal–Wallis test as the nonparametric equivalent of ANOVA and the Friedman's test as the counterpart of repeated measures ANOVA.

Statistical terminology for model building and validation
Machine learning terminology for model building and validation
Machine learning model overview

Statistical terminology for model building and validation
– Descriptive statistics – Inferential statistics

Machine learning – Supervised learning
– Classification Problem – Regression Problem – Unsupervised learning – Reinforcement learning

Predictive Modelling Statistical learning

Different methods of estimating parameters which are expected to provide estimators having some of these important properties. Commonly used methods are: 1. Method of moments 2. Method of maximum likelihood 3. Method of minimum v2 4. Method of least squares {P.K. Sahu et al., Estimation and Inferential Statistics}

Bayesian Statistics Philosophers of science are aware that many scientists use classical rather than Bayesian statistical methods, associated with the names of Fisher, Neyman and Pearson.

{Introduction to Statistics and Data Analysis by Christian Heumann · Michael Schomaker Shalabh} How to use different measures of association: ✓ nominal variables → Pearson's 2 , relative risks, odds ratio, Cramer's V , and Ccorr 2 ordinal variables → Spearman's rank correlation coeffi- cient, , c 2 continuous variables → Pearson's correlation coefficient, Spearman's correlation coefficient For two variables which are measured on different scales, for exam- ple continuous/ordinal or ordinal/nominal, one should use measures of association suitable for the less informative of the two scales. Another graphical representation of both a continuous and discrete vari- able is stratified confidence interval plots (error plots

p Value

Comparing the observed value of the statistic (here the obtained t-value) with the corresponding distribution (the t-distribution), we can find the likelihood that a value as extreme as or more extreme than the observed one is found by chance. This is the so-called p-value.

If the p-value is $p < 0{:}05$, we reject the null hypothesis, and speak of a statistically significant difference. If a value of $p < 0{:}001$ is obtained, the result is typically called highly significant. The critical region of a hypothesis test is the set of all outcomes which cause the null hypothesis to be rejected.
In other words, the p-value states how likely it is to obtain a value as extreme or more extreme by chance alone, if the null hypothesis is true.
The value against which the p-value is compared is the significance level, and is often indicated with the letter α. The significance level is a user choice, and typically set to 0:05.

This way of proceeding to test a hypothesis is called statistical inference.
Remember, p only indicates the likelihood of obtaining a certain value for the test statistic if the null hypothesis is true—nothing else!