Notes-pdf

```r
library(stat1201)
library(lattice)
```

01 - Intro

Sources of Variability

- Natural Variability
  - Something that we expect to be different such as the height of a person
- Measurement Variability
  - Differences in how people measure a certain thing

Diffrent Types of Variables

Quantitative Numerical value the represents measurements.

- Discrete
  - Variable that can have only whole counting numbers (integers)
- Continuous
  - Variable that is measurable, can have any value over some range, includes numerical values with decimal placed and can be counting numbers.

Categorical Represents groups of objects with a particular characteristic.

- Nominal
  - The groups do not have an order
- Ordinal
  - The groups have an order

Observational Study

- The researcher observes part of the population and measures the characteristics of interest
- Makes conclusions based on the observations but does not influence to change the existing conditions or does not try to affect them.
- E.g. Examine the effect of smoking on lung cancer on those who already smoke.

Experimental Study

- The researcher assigns subjects to groups and applies some treatments to groups and the other group does not receive the treatment.
- Can be designed as blind (participants don't know what group they are in).
- Can be designed as double-blind (participants and the researcher doesn't know the groups).
- When an experiment involves both comparison and randomization we call it a randomized comparative experiment.
- E.g. Examine the effect of caffeinated drinks on blood pressure.

Hypothesis Testing

Null hypothesis

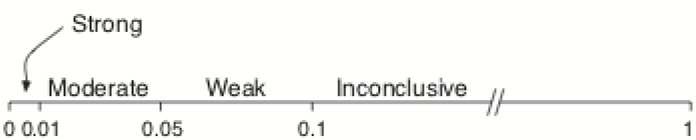
- Denoted H_0
- A statement of no effect.
- Either reject or do not reject H_0
- E.g. H_0: Caffeinated drinks has no effect on the mean change in pulse rate among young adults

Alternative Hypothesis

- Denoted H_1
- A statement of an effect
- If we reject H_0 we conclude there is sufficient evidence to accept the alternative hypothesis
- E.g. H_1: Caffeinated drinks increase the mean change in pulse rate among young adults

p-value

- We use the concept of p-value to reject or do not reject the null hypothesis



- p < 0.01 – Strong evidence against H_0
- 0.01 <= p < 0.05 – Moderate Evidence against H_0
- 0.05 <= p < 0.1 – Weak evidence against H_0
- p >= 0.1 – No evidence against H_0

02 - Exploratory Data Analysis

```
survey = read.csv("data/M2Survey.csv")
```

Central Tendency

Provides information about the center, or middle part of a quantitative variable.

Mode The most frequently occurring value in a set of data.

```
mode_stats(survey$Weight)
#> [1] 59
```

Median The middle value in ordered data and can be used to measure the center of the distribution.

- 50% of the observations are to the left of the median.
- If the number of observations is odd, the median is the middle number.
- If the number of observations is even, the median is the average of the two middle numbers.

```
median(survey$Weight)
#> [1] 65
```

Mean The average of a set of numbers.

```
mean(survey$Weight)
#> [1] 67
```

Measures of Location

Percentiles

- Measures of location
- Percentiles divide a set of ranked data so that a certain fraction of data is falling on or below this location
- E.g. 10th percentile is the value such that 10% of the data is equal to or below that value.

Quantiles

- Are labeled between the values 0 to 1.
- 10th percentile is the same as the 0.1 quantile

```
# 13th percentile
quantile(survey$Weight, probs = 0.13)
#> 13%
#>  54
```

Quartiles

- Divide a set of ranked data into four subgroups of parts. (Q_1, Q_2, Q_3)
- Q_1 separates the first 25% of ranked data to its left.
  - Same as the 25th percentile. Or 0.25 quantile.
- Q_2 separates the first 50% of ranked data to its left.
  - Same as the 50th percentile. Or 0.5 quantile.
  - Also the median
- Q_3 separates the first 75% of ranked data to its left.
  - Same as the 75th percentile. Or 0.75 quantile.

```
quantile(survey$Weight, probs = c(0.25, 0.5, 0.75))
#>   25%   50%   75%
#> 58.75 65.00 74.25
```

Measures of Variability

The variability measures can be used to describe the spread or the dispersion of a set of data. The most common measures of variability are range, the interquartile range (IQR), variance and standard deviation.

Range

- Range = Max - Min
- Range is affected by extreme values (outliers)

```
max(survey$weight) - min(survey$Weight)
#> Warning in max(survey$weight): no non-missing arguments to max; returning -Inf
#> [1] -Inf
```

Interquartile Range (IQR)

- IQR measures the distance between the first and third quartiles.
- This is the range of the middle 50% of the data
- IQR = Q_3 - Q_1

```
IQR(survey$Weight)
#> [1] 15.5
```

Variance and Standard Deviation

- Considers how far each data value is from the mean
- SD is the square root of variance
- SD is the most useful and most important measure of variability.

```
aggregate(Weight~Sex, survey, sd)
#>      Sex    Weight
#> 1 Female  8.872169
#> 2   Male 13.017779
```

Five Num Summary

Gives a compact description of a distribution including a rough picture of its shape.
Min, Q_1, Median, Q_3, Max

```
fivenum(survey$Height)
#> [1] 155 167 173 178 193
```

Skewed Distriubtions

Skewness measures the shape of a distribution.

Left or Negatively skewed: A greater number of observations occur in the left tail of the distribution (Mean < Median).

Right or Positively skewed: A greater number of observations occur in the right tail of the distribution (Mean > Median).

Outliers

The causes of outliers come from different ways.

- Data entry or measurement errors
- Sampling problems and unusual conditions
- Natural variation

Detecting

- IQR can be used to find outliers.
- Observation < Q_1 - 1.5 * IQR
- Observation > Q_3 + 1.5 * IQR

```
outliers(167, 178)
#> Observation < 150.5
#> Observation > 194.5
```

03 - Randomness and Probability

Population Parameters Vs Sample Statistics

	Population Param	Sample Stat
Size	N	n
Mean	μ	$\overline{x}$
Variance	σ^2	s^2
SD	σ	s
Proportion	p	$\hat{p}$

Sampling Error

- Sampling error is an unavoidable consequence of being able to observe only a subset of the elements in the population.
- Sampling errors can be reduced by increasing the sample size, and sometimes by using a different sampling selection approach.

Probability

- How likely that a particular event will happen.
- Probabilities to outcomes can be assigned in three ways
  - Subjective probability (reflects on an individual's belief)
  - Calculated or theoretical probability (based on prior knowledge)
  - Empirical probability (outcome is based on observed data).

Key Concepts

- Sample Space (Ω)
  - Set of all possible outcomes that might be observed in a random process.
- Event (A)
  - A subset of sample space. If an event occurs one of the outcomes in it occurs,
- Complement ($\overline{A}$)
  - The set of all outcomes in Ω not in A
- Union ($A \cup B$)
  - The set of all outcomes in A, or in B, or in both.
- Intersection ($A \cap B$)
  - The set of outcomes in both A and B
- If the two events are disjoint then,
  - $P(A \cup B) = P(A) + P(B)$

Conditional Probability

- Probability of event A occurring if B has already occurred.
- $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Independent Events

- Two events are independent if one event occurs and it does not affect the probability of the other event occurring.
- Only if A and B are independent events the probability of A occuring, given B has already occured, will be the same as just the probability of A.

$$
\begin{aligned}
P(A|B) &= P(A) \\
P(B|A) &= P(B) \\
P(A \cap B) &= P(A) \times P(B)
\end{aligned}
$$

Discrete Probability Distribution

- The listing of all possible values of a discrete random variable X along with their associated probabilities
- A random variable that has a countable number of possible values.
  - Usually things which are counted, and not measured.
- Example:

```
discrete_dist(0:3, c(0.21, 0.45, 0.23, 0.11))
#>  x P(X=x)
#>  0   0.21
#>  1   0.45
#>  2   0.23
#>  3   0.11
#>
#> Discrete Probability Distribution
#>
#>  E(X) Var(X)     sd(X)
#>  1.24 0.8224 0.9068627
```

Expected Value (Mean) and Variance

- Long run average of a random variable.

$$
\begin{aligned}
E(X) &= \mu = \sum xP(X = x) \\
Var(X) &= \sigma^2 = \sum P(X = x)(x - \mu)^2 \\
SD(X) &= \sigma = \sqrt{Var(X)}
\end{aligned}
$$

Continuous Probability Distribution

- A random variable that takes values at every time over a given interval
  - Usually things which are measured, not counted
- Can not be presented in a table or histogram as there is an uncountable number of possible outcomes.
- The probability of any individual outcome is zero.
  - $P(X = x) = 0$
- We always calculate the probability for a range of the continuous random variable X.
  - $P(X > a)$
  - $P(a <= X <= b)$

Expected Value and Variance of Combined Variables

- Rule 1:
  - Suppose X is a random variable and a is a constant

$$
\begin{aligned}
Y &= aX \\
E(Y) &= aE(X) \\
Var(Y) &= a^2 Var(X)
\end{aligned}
$$

- Rule 2:
  - Suppose X is a random variable and a and b are constants.

$$
\begin{aligned}
Y &= aX + b \\
E(Y) &= aE(X) + b \\
Var(Y) &= a^2 Var(X) \\
SD(Y) &= aSD(X)
\end{aligned}
$$

- Rule 3:
  - Suppose X_1 and X_2 are two independent random variables.

$$
\begin{aligned}
Y &= X_1 + X_2 \\
E(Y) &= E(X_1) + E(X_2) \\
Var(Y) &= Var(X_1) + Var(X_2)
\end{aligned}
$$

- Rule 4:
  - Suppose X_1 and X_2 are two independent random variables.

$$
\begin{aligned}
Y &= X_1 - X_2 \\
E(Y) &= E(X_1) - E(X_2) \\
Var(Y) &= Var(X_1) - Var(X_2)
\end{aligned}
$$

04 - Probability and Sampling Distributions

Binomial Distribution

Important discrete probability distribution.
We use the concept of Bernoulli Trial to describe the Binomial Distribution.

- A Bernoulli Trial is a random process with only two possible outcomes.
- These outcomes are *success* and *failure*
- Let X be the number of successes from n number of independent Bernoulli trials and P(Success) = p.
- X has a Binomial distribution with parameters n and p
  - X ~ Binom(n, p)

Mean and SD of X

$$
\begin{aligned}
X &\sim Binom(n, p) \\
E(X) &= np \\
Var(X) &= np(1-p) \\
SD(X) &= \sqrt{Var(X)}
\end{aligned}
$$

Example

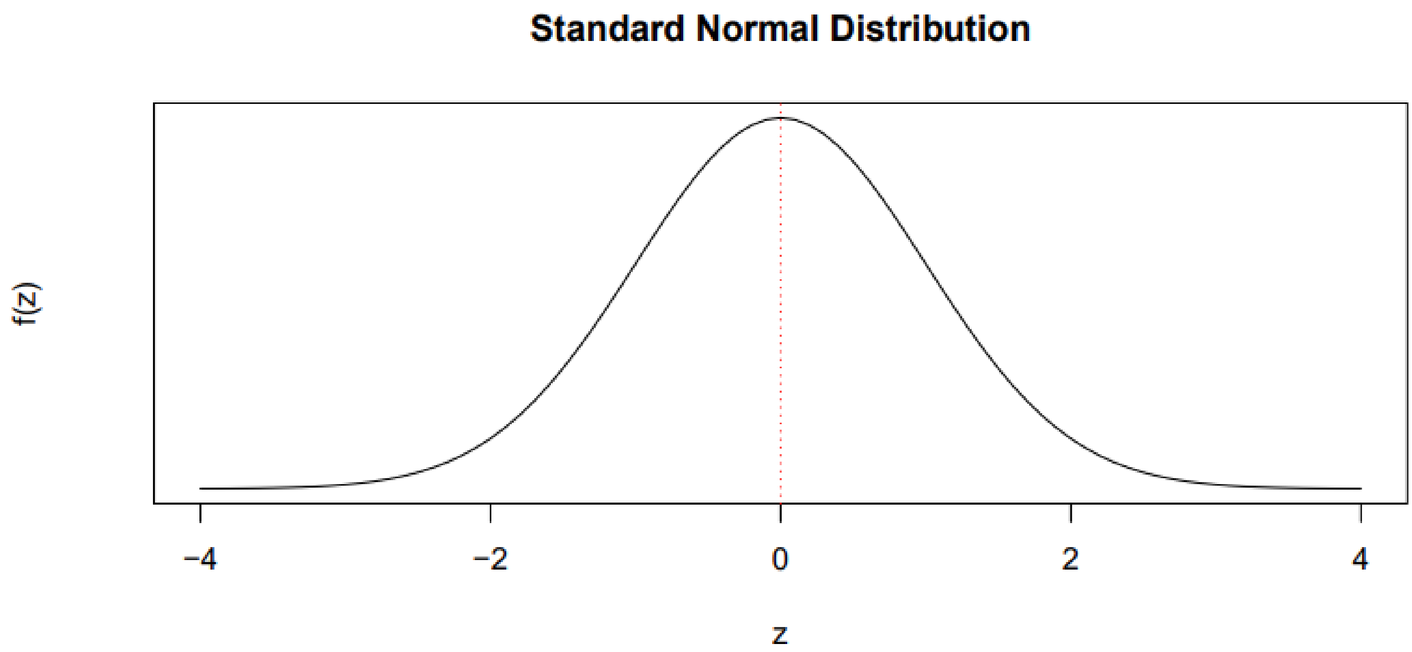
```r
# Let X be the number of lizards whose length is above the mean. (60%)
n = 5
p = 0.6
# Then X ~ Binom(5, 0.6)

# P(X=2)
dbinom(2, 5, 0.6)
#> [1] 0.2304

# P(X < 2)
sum(dbinom(0:1, 5, 0.6))
#> [1] 0.08704

# P(X >= 2) == 1 - P(X < 2)
1  - sum(dbinom(0:1, 5, 0.6))
#> [1] 0.91296

binom_dist(5, 0.6)
#> X ~ Binom(5, 0.6)
#>
#> Binomial Distribution (n, p)
#>
#>  E(X) Var(X)    sd(X)
#>     3    1.2 1.095445
```

Normal Distribution

- Also called Gaussian Distribution
- Normal Distribution is a continuous probability distribution with two parameters, μ and σ
- Let X be a continuous random variable. If X has a Normal distribution we can write,
  - X ~ Normal(μ, σ)
- Bell shaped and symmetrical about μ
- Location is determined by μ
- Spread is determined by σ
- The random variable X has an infinite theoretical range ($-\infty$ to $+\infty$).

Probability Calculations

- The area under the Normal density curve is 1.
- Rough rule to calculate the areas.
  - Within 1 SD of the mean is 68%
  - Within 2 SD of the mean is 95%
  - Within 3 SD of the mean is 99.7%
- We Transform the Normal Distribution to a Standard Normal Distribution.
  - If X ~ Normal(μ, σ)
  - Then $Z = \frac{X-\mu}{\sigma}$
  - and Z ~ Normal(0, 1)



Sampling Distribution of the Sample Mean

- The distribution of all possible sample means using the same sample size, selected from a population.
- Suppose we have a population of 1000 people's heights.
  - $\mu = 162.1504$
  - $\sigma = 8.147348$
- We can then take 20 samples each of size n from the population
- We can treat the sample means ($\overline{X}$) as a random variable and calculate the mean and the standard deviation of the 20 sample means.

Sample Size (n)	E($\overline{X}$)	sd($\overline{X}$)
4	161.95	4.619
16	162.55	2.095
25	162.53	1.521
100	162.36	0.780

- We can observe that the mean of the sample means closes in on the population mean.
- The standard deviation of the sample means becomes smaller.
- The ratio of the standard deviation of the sample means to the population standard deviation is $\frac{1}{\sqrt{n}}$
- If the population is normally distributed, the sampling distribution of the sample means ($\overline{X}$) is normally distributed.
- Therefore the distribution of $\overline{X}$ can be summarized as follows:

$$E(\overline{X}) = \mu$$
$$Var(\overline{X}) = \frac{\sigma^2}{n}$$
$$sd(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$
$$\therefore \overline{X} \sim Norm(\mu, \frac{\sigma}{\sqrt{n}})$$

Central Limit Theorem

- As the sample size increases, the sampling distribution of the sample means becomes approximately normally distributed regardless of the shape of the population variable distribution

Example

```r
sampling_dist_mean(50, 8, 4)
#> Xbar ~ Norm(50, 8/sqrt(4))
#>
#> Sampling Distribution of the Sample Mean
#>
#>  E(Xbar) Var(Xbar) sd(Xbar)
#>       50        16        4
```

Sampling Distribution of the Sample Proportions

- The sample proportion ($\hat{p}$)
- Define p as the population proportion of students whose height is less than or equal to 155cm
- $\hat{p} = \frac{x}{n}$ where x is the number of students in the sample whose height is less than or equal to 155cm
- Provided that n is large such that $np > 5$ and $n(1-p) > 5$ we can show that,

$$E(\hat{p}) = p$$

$$sd(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

$$\therefore \hat{p} \sim Norm(p, \sqrt{\frac{p(1-p)}{n}})$$

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

$$Z \sim Norm(0, 1)$$

Example

```r
sampling_dist_prop(0.1, 10)
#> phat ~ Norm(0.1, 0.0948683)
#>
#> Sampling Distribution of the Sample Proportions
#>
#>  E(p.hat) Var(p.hat)  sd(p.hat)
#>       0.1      0.009 0.09486833
```