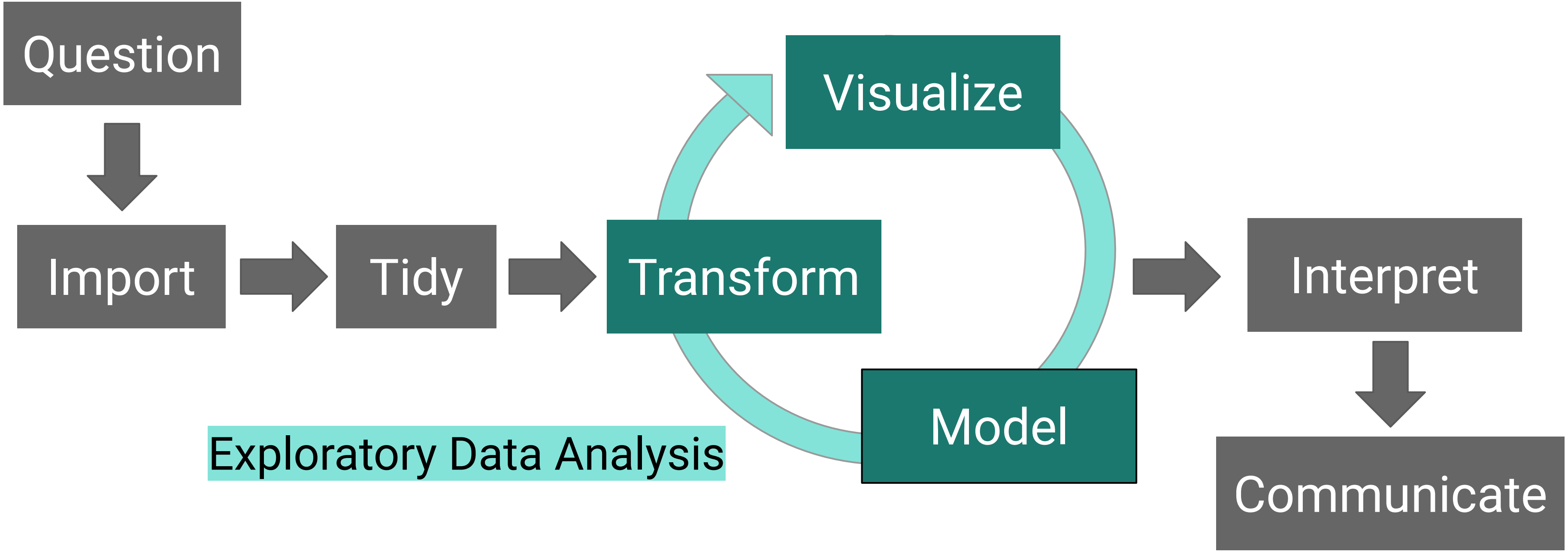
Statistical Inference

Lecture 10

Motivation

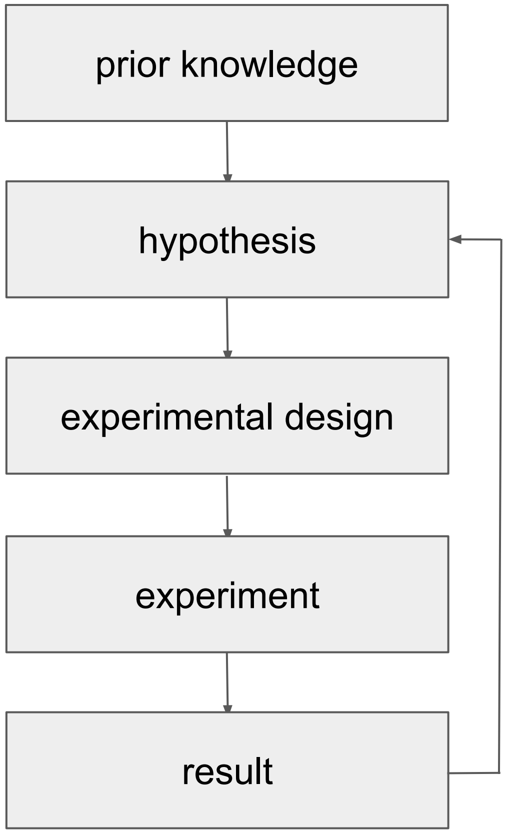


Modified from https://r4ds.had.co.nz/explore-intro.html#explore-intro

Scientific method

- Hypothesis-driven

Based on observations and published results
Testable (to prove/disprove hypothesis)
New information or knowledge



Statistical analysis

- Inference (hypothesis testing)
- Prediction

What is the purpose of statistical analysis?

The scientific method assumes that the “truth” exists and it can be tested/proven/investigated.

- To make sense of data (summary statistics, trends, patterns, spread)
- To test hypotheses, compare groups
- To determine associations, correlations
- To predict or estimate outcomes (evaluate errors, uncertainties, forecasts)

Descriptive statistics

<u>Measures of central tendency</u>

Mean – sum of all observations divided by the number of observations; fulcrum to balance histogram

Median – rank ordering all observations and choosing the middle term for which 50% of the values lay above/below (50th percentile)

less resistant to outliers since it finds the middle distribution that includes extreme values

Mode – most commonly occurring value in a sampled distribution

Descriptive statistics

Measure of dispersion

description of how far the data are spread out about the center (cluster or scatter)

Range – difference between the largest and smallest value; not robust and sensitive to outliers

Variance – spread in the distribution can be measured by assessing how far each individual values differ from the mean

Standard deviation (SD) - square root of variance; average separation of data from the mean

mean ± SD

median interquartile range (IQR)

Standard error (SE)

- standard error of the mean (SEM) is common in basic science but it is NOT a measure of dispersion

- SEM is a measure of how accurately the population mean has been estimated

- **Standard error (SE)** – represents variability of estimate in a collection of measurements derived from multiple runs or different groups

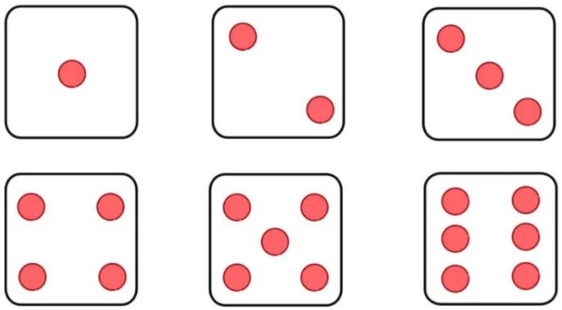
$$SE = \frac{SD}{\sqrt{n}}$$

Summarizing outcomes

Data	Statistics
Descriptive	
continuous	sample size (n)
	mean ± SD
	median and IQR
categorical	sample size (n)
	relative frequency (%)
Group comparisons	
continuous	mean ± SE for each group
categorical	proportion (%) and SE for each group

Probability

- The **probability** of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.
- Tossing a coin:
  - Outcome: head, tail
  - P(head) = ½ = 50%
  - P(tail) = ½ = 50%
- Rolling a die:
  - Outcome: 1,2,3,4,5,6
  - P(rolling a 5) = ⅙ = 0.167 = 16.7%
  - P(rolling an odd number) = P(1,3,or 5) = 3/6 = 50%



Independence

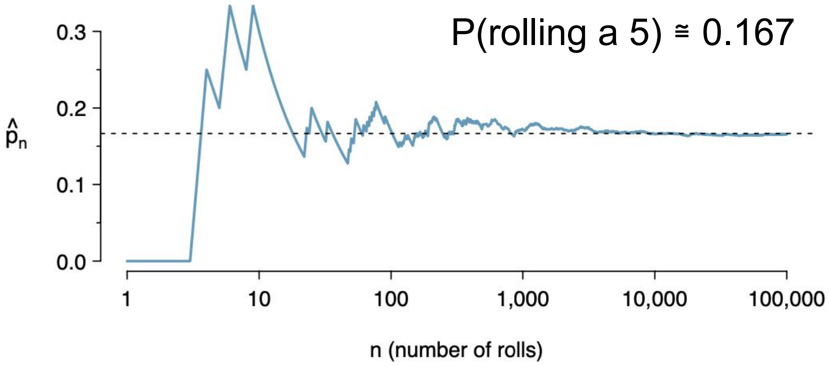
- Two events are **independent** if the outcome of one provides no useful information about the outcome of the other.
- Flipping a coin and rolling a die are two independent process

P(A and B) = P(A) * P(B)

P(male and left-handed) = P(male) * P(left-handed)
P(male and left-handed) = 0.5 * 0.09 = 0.045
P(male and left-handed) = 4.5%

Law of large numbers

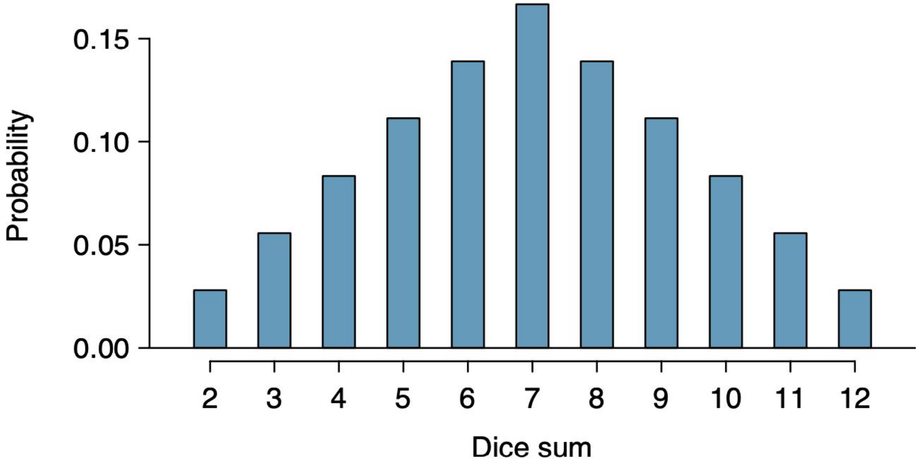
- As more observations are collected, the proportion of occurrences with a particular outcome converges to the probability of the outcome.
- Casinos always make money in the long run.



https://en.wikipedia.org/wiki/Craps

Probability distributions

- A probability distribution is a list of all outcomes and their associated probabilities.

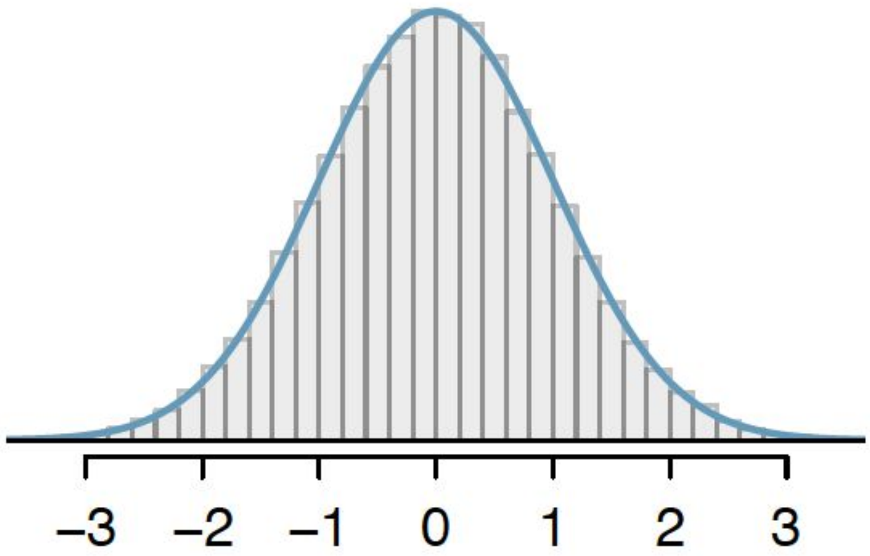


Dice sum	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Probability distributions

Normal distribution

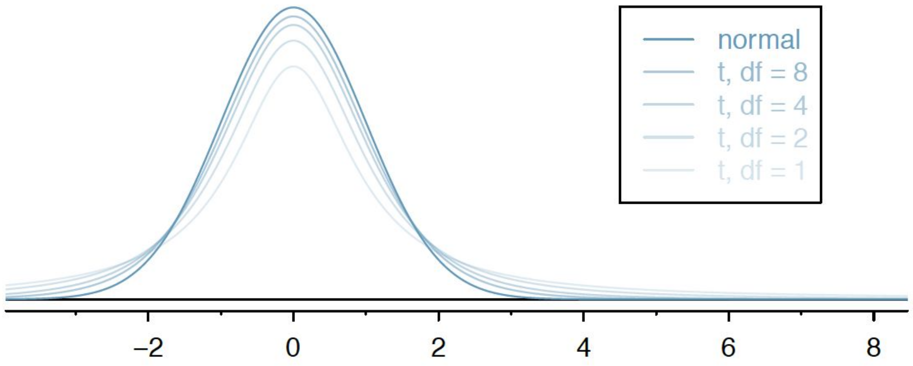
- Symmetric, unimodal, bell-shaped curve
- 2 parameters: **mean** and **SD** describe the shape of a normal curve



Distributions of variables

Student’s t distribution

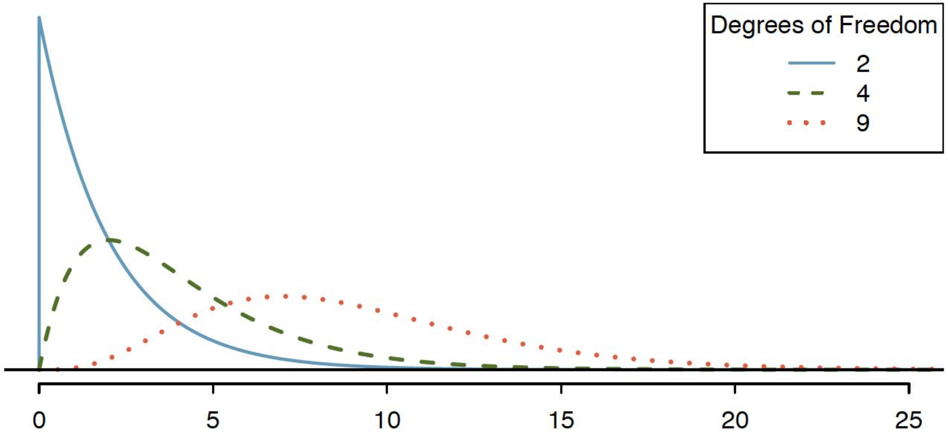
- Use to estimate the mean of a normally distributed continuous variable when n is small
- Use to test difference between two sample means or confidence intervals for small sample sizes
- Centered at zero with 1 parameter: **degrees of freedom**



Distributions of variables

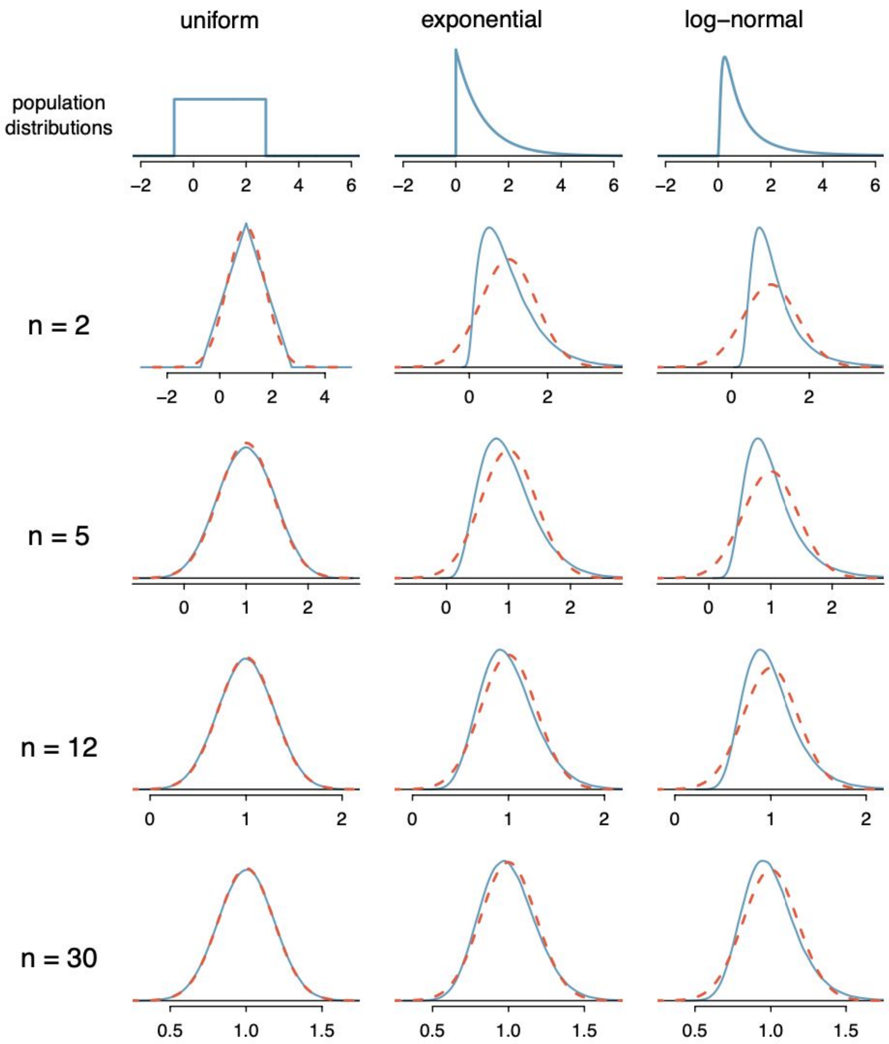
Chi-square distribution

- Use to characterize categorical variables that are always positive and usually right skewed
- 1 parameter: **degrees of freedom** dictate the shape of chi-square curve



Central Limit Theorem

- A large, properly drawn sample will resemble the population from which it was drawn.

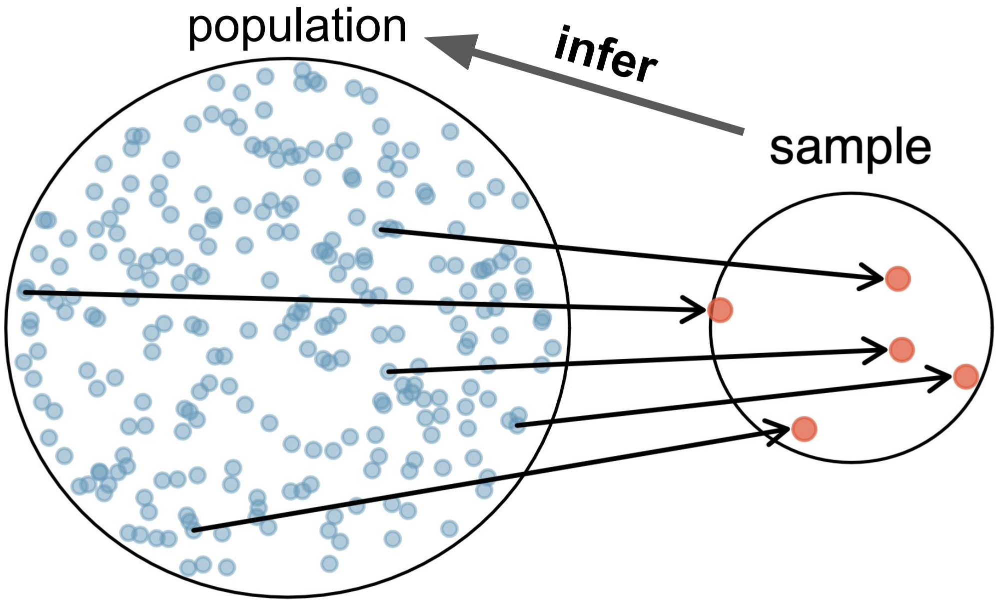


Central Limit Theorem

- Given a detailed information about some population, then it is possible to infer about any properly drawn sample from the population.
- Given a detailed information about a properly drawn sample (mean and SD), then it is possible to infer about the population from which that sample was drawn.
- Given a data describing a particular sample and data on a particular population, then it is possible to infer whether or not that sample is consistent with a sample that is likely to be drawn from that population.
- Given the underlying characteristics of two samples, then it is possible to infer whether or not both samples were likely drawn from the same population.

Statistical inference

- testing of hypothesis is performed to draw **inference** from the data



Statistical inference

- Independent vs repeated measurements
  - independent - one observation, sampling, or treatment per unit
  - dependent - repeated measurements taken on the same set of experiment unit under differing conditions
- Parametric vs non-parametric tests
  - parametric statistical methods that rely on estimation of parameters (mean, variance)
  - assume that the distribution is normally distributed (do a test for normality e.g. Kolmogorov-Smirnov normality test, Shapiro-Wilk's test)
  - normality refers to the distribution of the population and not the sample
  - perform non-parametric tests if distribution is not normally distributed

How to select test statistics?

Outcome Variable	Group Structure	Assumptions for Parametric test	Parametric Test	Nonparametric Test	Regression Models
Continuous	2 Independent	Independence of observations, normality, large samples, homogeneity of variances	Student t-test	Mann-Whitney U or Wilcoxon rank sum test	Univariate or Multivariate Linear Regression
	2 Dependent	Independence of pairs, normality, large samples, and homogeneity of variances	Paired Student t-test	Wilcoxon signed rank test	
	>2 Independent	Independence of observations, normality, large samples, homogeneity of variances	ANOVA	Kruskal-Wallis test	
	>2 Dependent	Repeated measures in independent observations, normality, large samples, homogeneity of variances	Repeated-measures ANOVA	Friedman test	
Categorical	≥2 Independent	Independence of observations, expected count >5 in each cell	Chi-square test	Fisher's exact test	Binomial or Multinomial Logistic Regression
	≥2 Dependent	Independence of pairs	McNemar test		
Time-to-event	≥2 Groups	Non-informative censoring, sufficient follow-up time and number of events	Parametric Proportional Hazards	Kaplan-Meier, Log-rank test	Cox Proportional-Hazards Regression

Steps in a hypothesis testing

1. Statement of the question to be answered
2. Formulation of the null and alternative hypotheses
3. Decision for a suitable statistical test
4. Specification of the level of significance ($\alpha = 0.05$)
5. Performance of the statistical test analysis (p-value)
6. Statistical decision
   - If $p > 0.05$, then accept the null hypothesis
   - If $p < 0.05$, reject the null and accept the alternative hypothesis
7. Interpretation of test result

Babies dataset

- The Child Health and Development Studies (USA) investigated pregnancy in women in San Francisco, 1960-19967. Studied the relationship between mothers who smoked and weight of their babies.
- 1,236 observations x 8 variables

variable	description
case	id number
bwt	birthweight, ounces
gestation	length of gestation, days
parity	binary indicator for a first pregnancy (0=first pregnancy)
age	mother's age, years
height	mother's height, inches
weight	mother's weight, pounds
smoke	binary indicator whether the mother smoked, 1=smoker

Babies dataset

```
# Summary statistics
> summary(babies)
```

```
      case             bwt              gestation           parity
 Min.   :   1.0   Min.   : 55.0    Min.   :148.0      Min.   :0.0000
 1st Qu.: 309.8   1st Qu.:108.8    1st Qu.:272.0      1st Qu.:0.0000
 Median : 618.5   Median :120.0    Median :280.0      Median :0.0000
 Mean   : 618.5   Mean   :119.6    Mean   :279.3      Mean   :0.2549
 3rd Qu.: 927.2   3rd Qu.:131.0    3rd Qu.:288.0      3rd Qu.:1.0000
 Max.   :1236.0   Max.   :176.0    Max.   :353.0      Max.   :1.0000

      age             height            weight             smoke
 Min.   :15.00    Min.   :53.00    Min.   : 87.0      Min.   :0.0000
 1st Qu.:23.00    1st Qu.:62.00    1st Qu.:114.8      1st Qu.:0.0000
 Median :26.00    Median :64.00    Median :125.0      Median :0.0000
 Mean   :27.26    Mean   :64.05    Mean   :128.6      Mean   :0.3948
 3rd Qu.:31.00    3rd Qu.:66.00    3rd Qu.:139.0      3rd Qu.:1.0000
 Max.   :45.00    Max.   :72.00    Max.   :250.0      Max.   :1.0000
 NA's   :2        NA's   :22       NA's   :36         NA's   :10
```

Babies dataset

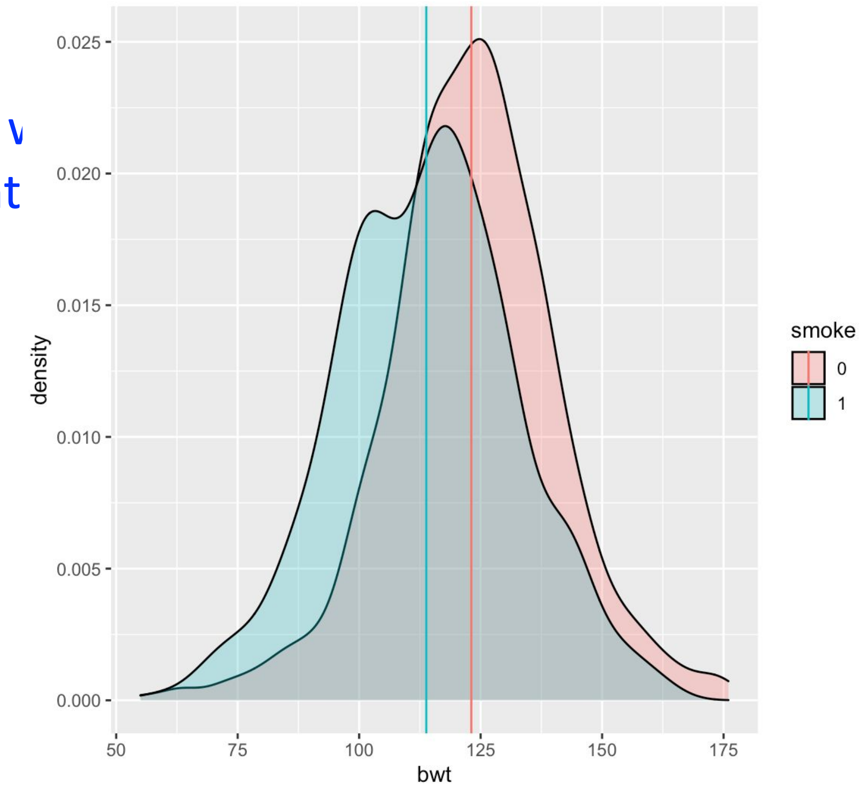
```
# Compute mean and SD between 2 groups (smoke)
> babies %>%
    group_by(smoke) %>%
    summarize(n=n(), mean=mean(bwt), SD=sd(bwt))
```

	smoke	n	mean	SD
1	0	715	123.	17.4
2	1	459	114.	18.3

1. Statement of the Problem

"Is there a relationship between mothers w
smoked during pregnancy and birthweight
their newborns?"

	smoke	n	mean	SD
1	0	715	123.	17.4
2	1	459	114.	18.3



2. Formulation of null and alternative hypotheses

- null : There is no difference in mean birthweight for newborns from mothers who did and did not smoke during pregnancy

  $\mu_s - \mu_n = 0$

- alternative: There is a difference in mean newborn birthweights from mothers who did and did not smoke during pregnancy

  $\mu_s - \mu_n \neq 0$

$\mu_s - \mu_n = 114 - 123 = -9$

	smoke	n	mean	SD
1	0	715	123.	17.4
2	1	459	114.	18.3

3. Decision for the test statistic

- Outcome variable: Continuous
- Group structure: 2 Groups (independent)
- Assumptions:
    - independence of observations
    - normality
    - large samples
    - homogeneity of variances

T-test

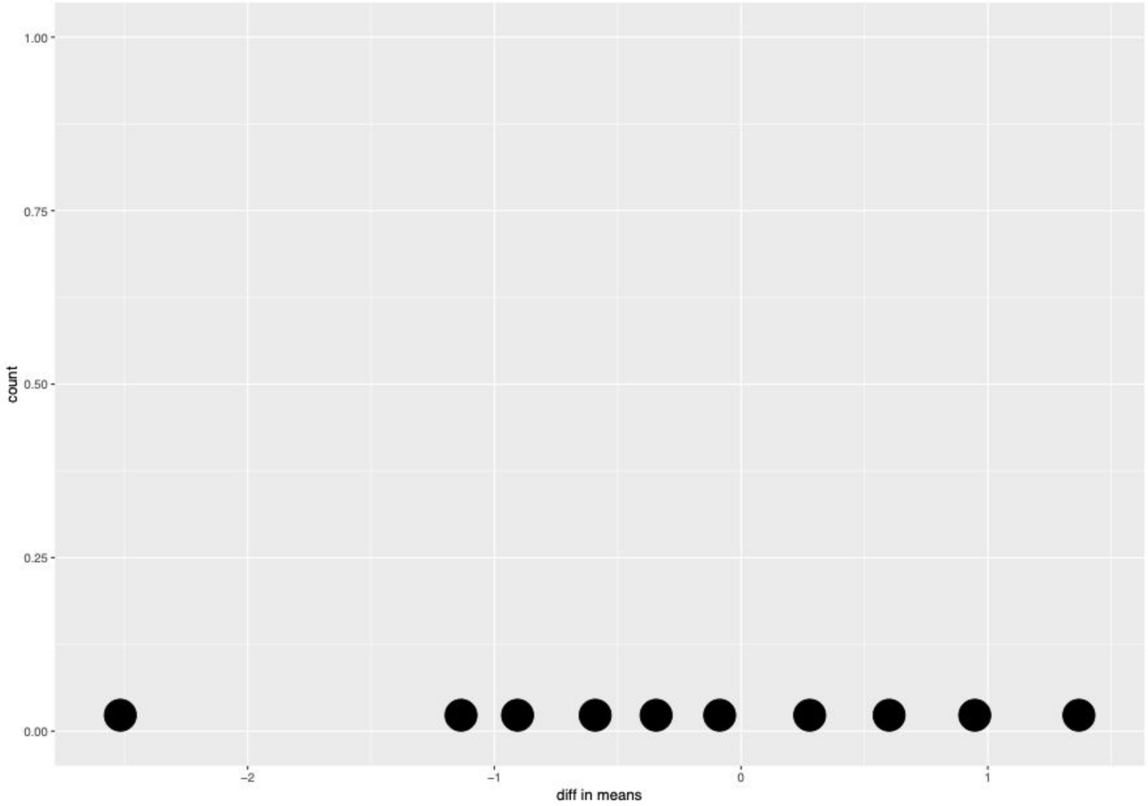
- **t-statistic** is the test statistic for inference on the difference of two sample means where σ_1 and σ_2 are unknown.

$$t_{df} = \frac{point\ estimate\ -\ null\ value}{SE}$$

where

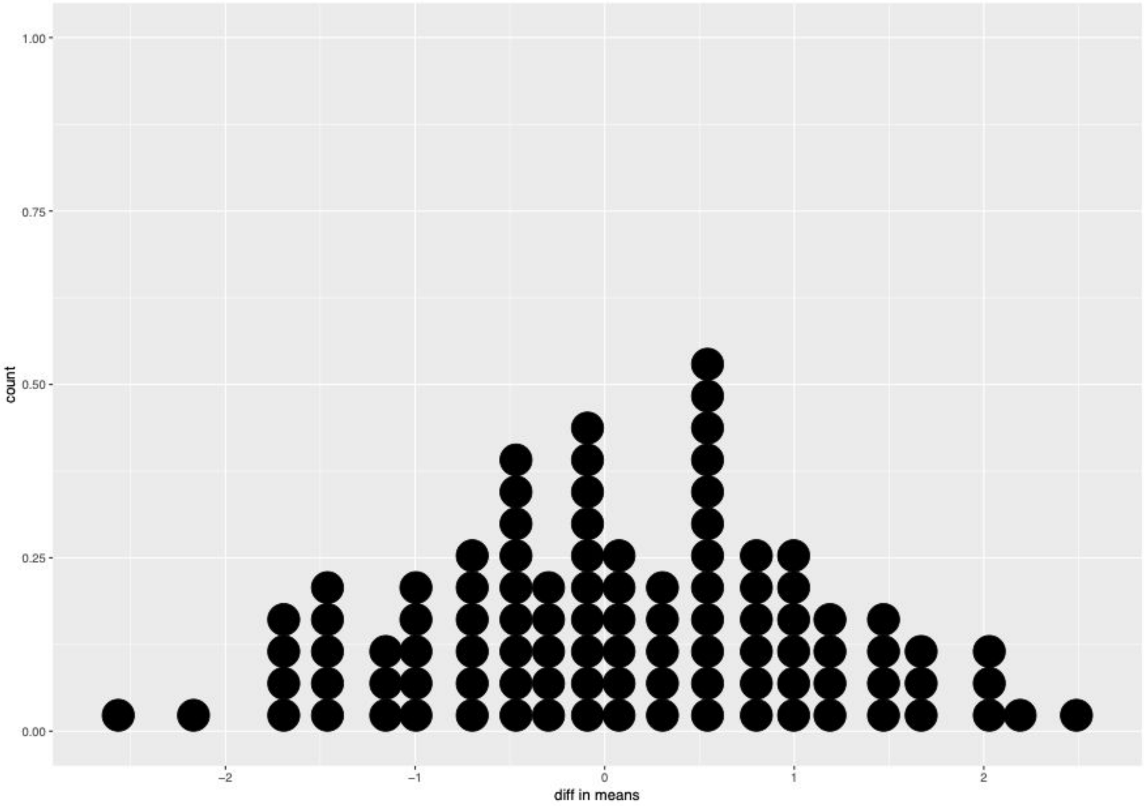
$$SE = \sqrt{\frac{{s_1}^2}{n1} + \frac{{s_2}^2}{n2}} \text{ and } df = \min(n_1 - 1, n_2 - 1)$$

Null distribution



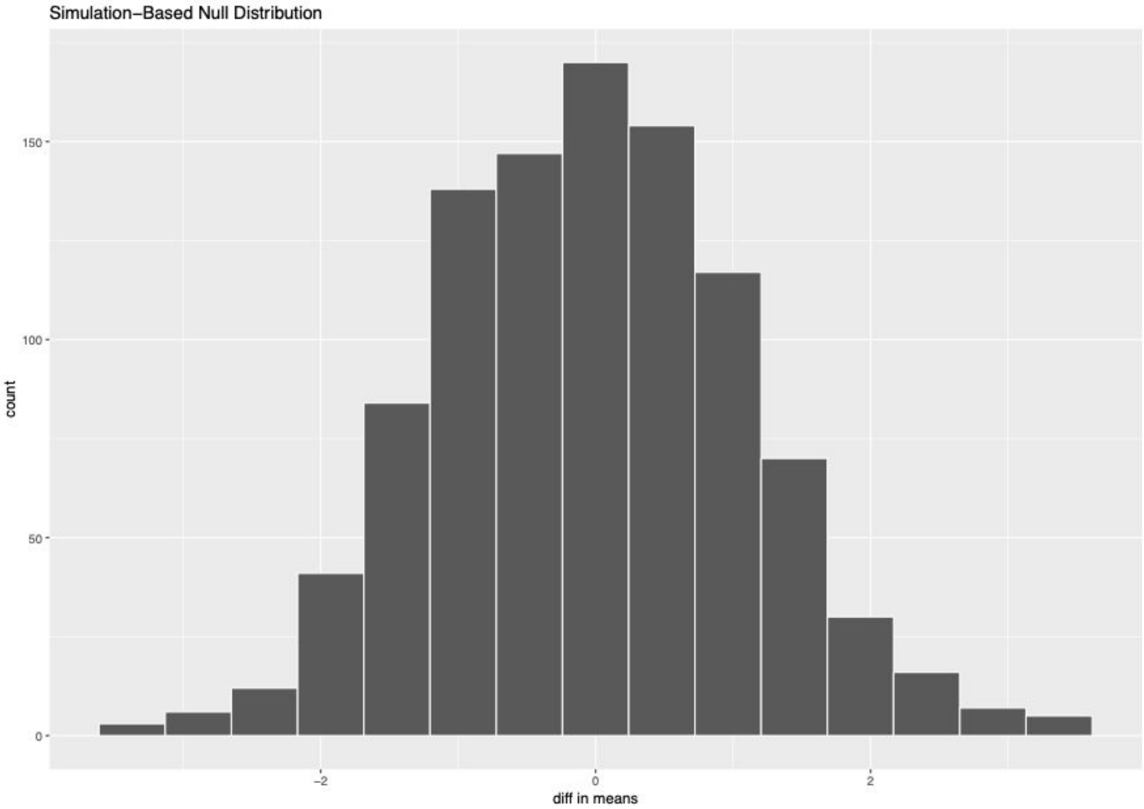
$n_{sampling}$ = **10**

Null distribution



$n_{sampling}$ **= 100**

Null distribution



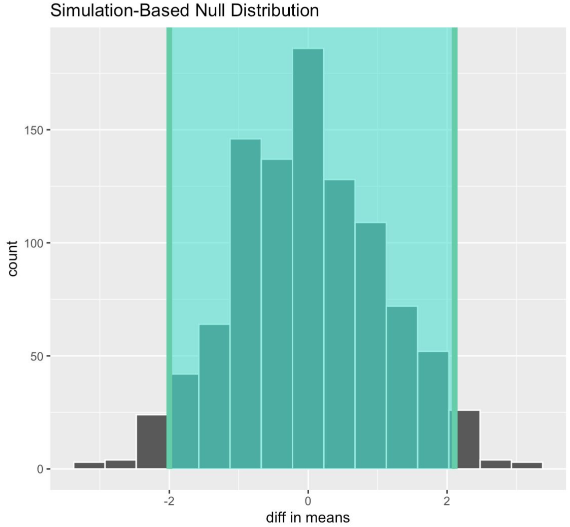
$n_{sampling}$ = **1000**

4. Level of significance (α = 0.05)

- probability of rejecting the null hypothesis when it is true
- a significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference

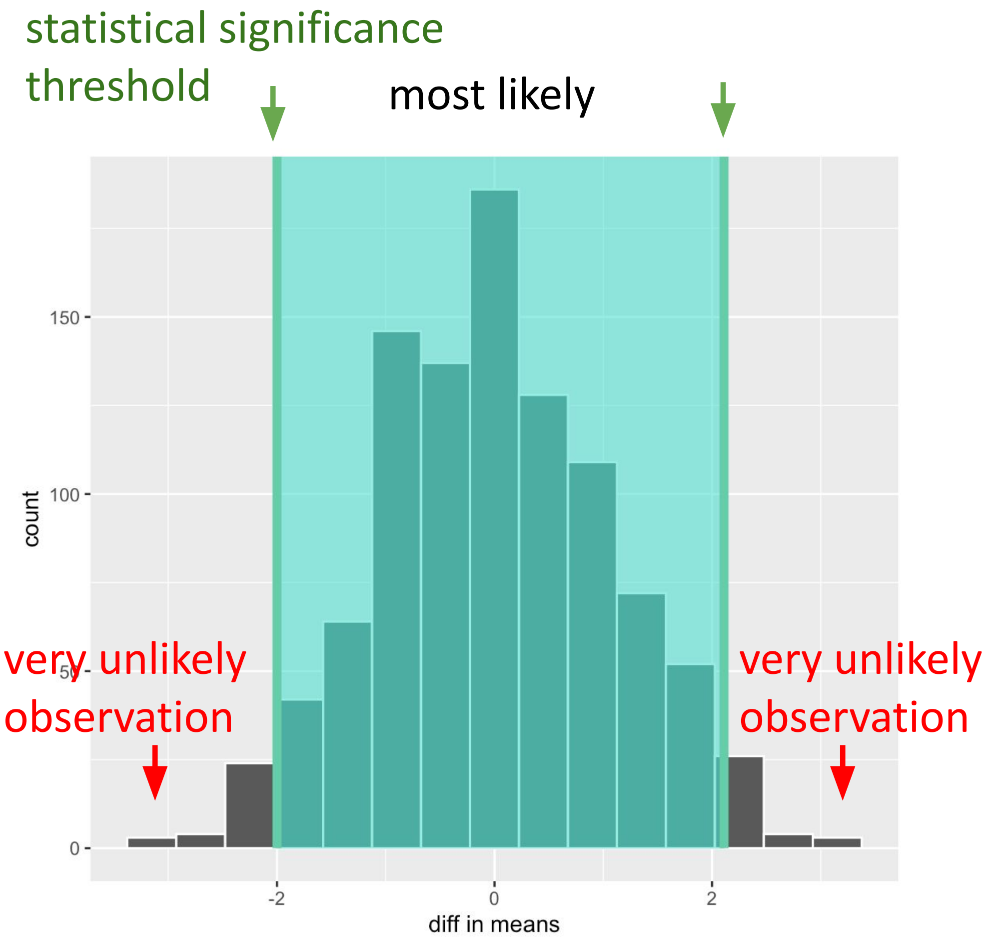
most likely

statistical significance threshold



p-value

- what is the probability that a specific assertion is right or wrong?
- probability of observing data as more extreme than actually obtained given that the null hypothesis is true



5. Statistical test analysis

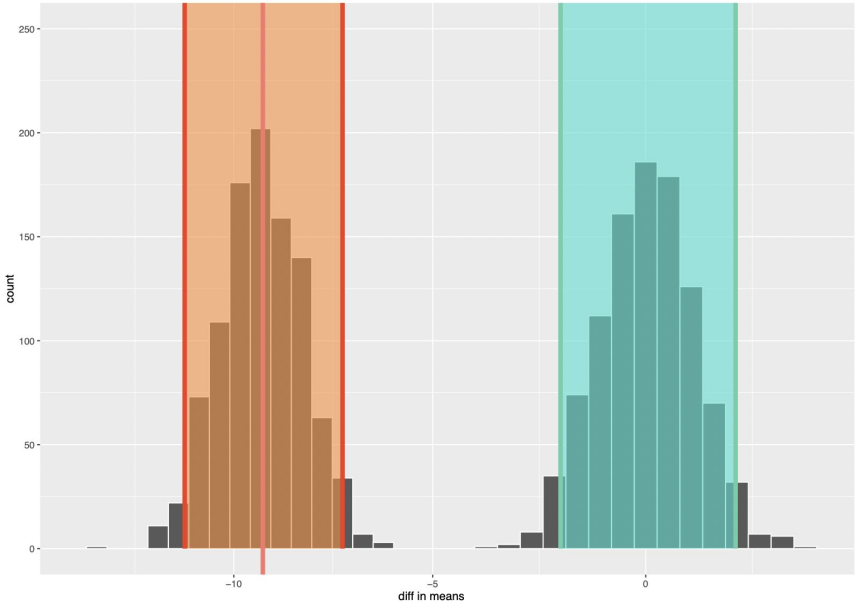
```
# Calculate t-test
> t.test(bwt ~ smoke, data = babies)
```

```
        Two Sample t-test

data:  bwt by smoke
t = 8.7188, df = 1172, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  7.180973 11.351312
sample estimates:
mean in group 0 mean in group 1
       123.0853        113.8192
```

6. Statistical decision

- p-value < 2.2e-16
- since p-value < 0.05, reject null and accept alternative
- There is a significant difference in average weight of newborns from mothers who smoke during pregnancy than mothers who did not smoke.



7. Interpretation of test result

- Birthweight of newborns from mothers who smoked during pregnancy was about 9 oz. (95% CI: 7.2-11.4, p-value < 0.05) lighter on average than mothers who did not smoke.

```
        Two Sample t-test

data:  bwt by smoke
t = 8.7188, df = 1172, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  7.180973 11.351312
sample estimates:
mean in group 0 mean in group 1
       123.0853        113.8192
```

Take-away message

- Hypothesis testing is useful when determining how sure are you that the sample estimate (e.g. mean, difference of means or proportions) you obtained is near to the true population value.