

# Stats 3Y03 Summary

---

Note: R might be on the final :\$

## Chapter 1

**Categorical variable:** qualitative variable, such as funny; limited number of options

- e.g. Blood type, Political party
- It can still be a number if the number doesn't describe a quantity
- **Ordinal:** Values that can be ordered, such as academic grade
- **Nominal:** Values that cannot be ordered, such as brand name

## Types of variables

**Numerical variable:** quantitative variable, such as position

- **Continuous:** decimals
- **Discrete:** integer

**Univariate Data:** single variable

**Bivariate Data:** 2 variables (not required in this course)

**Multivariate Data:** more than 2 variables

**Probability:** average of population is from average of sample

**Inferential statistics:** average of sample is from average of population

**Sampling Frame:** list of things in a list that can be sampled

- telemarketers' sample frame is the people with a phone number in the phone book/phone archive of the company
- when doing a culture study of farms, the sample frame could even be a map

**Enumerative study:**

- identifiable goal
- well-defined, unchanging sample frame
- enumerate (explain, evaluate, describe) a condition that exists with the existing population

**Analytic study:**

- focused on improvement of the process which created the results and which will continue creating results in the future
- no well-defined sampling frame

## Target population

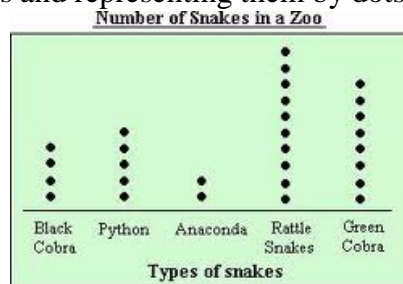
- population you want to be collecting data from
- **sample population** is the population you are collecting data from
- sample population is usually subset of target population
- sample population is useful when the target population is too large
- sometimes it is not the same as the sample population
  - e.g., when informing factory workers that their productivity is being observed, they'll act differently

**Simple random sample:** from entire population

**Stratified random sample:** from a sub-population (1 from each row)

**Convenience sample:** not entirely random; what is easy to obtain (first row)

**Dot plot:** quantifying increments and representing them by dots



**Mean:** average

**Median:** middle value; if length of set is even, average of  $(n+1)/2$  and  $n/2$ ; if length of set is odd,  $(n+1)/2$

**Mode:** common number

**Unimodal:** 1 peak

**Bimodal:** 2 peaks

**Multimodal:** more than 2 peaks

Graphs can also be **symmetric** or **asymmetric**, which is when the top half of the boxplot looks similar to the bottom half.

**Left skew:** mostly on right side

**Right skew:** mostly on left side

Graphs can also be **unskewed**.

**Outliers:**

- values that must be mistakes or abstract exceptions
- $> 1.5 \times$  forth spread (see below) beyond closest quartile
- **extreme outlier** is  $> 3 \times$  forth spread

Each data set is split up into 4 **quartiles**.

Q1: median of bottom half (includes middle number if odd length)

Q2:

Q3: median of top half (includes middle number if odd length)

### **The Five-Number Summary:**

1. Minimum
2. Q1
3. Q2
4. Q3
5. Maximum

The range, minimum, and maximum can include outliers

**range:**  $\text{max} - \text{min}$

### **Variance**

**Variance:** distribution of range

$N$  is [target population](#) size

$x_i$  are the values

$n$  is [sample population](#) size

**Trimmed mean:** mean calculated by trimming away a given percentage of elements (relative to number of elements) from the top and bottom. If the percentage gives a non-discrete number of elements, you have to calculate multiple trimmed means and find the mean of the 2 trimmed means

**Population mean:** expected outcome of mean of [target population](#), i.e. average given a theoretically infinite amount of measurements; a.k.a. true mean, expected value

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

**Sample mean:** average given finite number of inputs; an estimate of the population mean

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

**Sample median:**  $\tilde{x}$

**Sample variance:**  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

**Population variance:**  $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$

**Spread:** interquartile range

## Standard Deviation

s.d.

- Average distance from the mean
- Larger s.d. means more spread
- i.e. when all values are the same, s.d. = 0
- Square root of variance =  $\sqrt{s^2}$

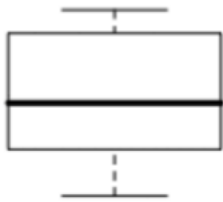
**Degrees of freedom:**  $n - 1$

Another measure of spread is **interquartile range** or **forth spread**. ( $Q_3 - Q_1$ )

**Whiskers:** minimum and maximum points of the range that does not include outliers

**Boxplot:**

- Top and bottom lines are whiskers
- Box surrounds forth spread
- Middle line is median
- Can be vertical or horizontal
- Outliers are still placed on boxplots, using circles (o) or stars (\*)
- (a.k.a. Boxplot-and-whisker plot)



## Chapter 2

This is similar to the logic course [SFWR ENG 2FA3](#).  
Probability is between 0 and 1

**Sample space:** all possible outcomes

The size of the sample space is: outcomes<sup>events</sup>.

N: number of outcomes for an event

N(A): number of outcomes in sample space, A

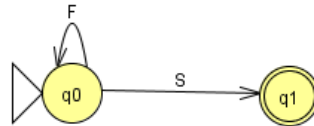
**Relative frequency probability:** events that occur frequently, such as rolling dice or buying lottery tickets

**Relative frequency** of a value =  $\frac{\text{occurrences of value}}{\text{observations in data set}}$

**Personal probability:** events that cannot be repeated or non-random events with unknown quantities that is based on belief of an individual

**Coherent:** personal probability of one event does not contradict personal probability of another

Sometimes you can have an **infinite number of possible outcomes**. For example, if you are testing something until failure, you will repeat testing until success {S, FS, FFS, ...}



If there are a given number of outcomes, such as 1 through 6 for a dice, and a sample space, A, such as containing all odd outcomes, A', the **complement**, contains everything A does not, such as all even outcomes. Therefore,  $P(A) + P(A') = 1$

**Simple Event:** Only one way to get each outcome

**Compound Event:** Multiple ways to get the same outcome

## Replacement

**With replacement:** e.g. if you are picking names out of a hat and you put the names back after each pick; independent

**Without replacement:** when you use each option only once; dependent

## Mutually-Exclusive Events

**Mutually exclusive** (a.k.a. disjoint event): 2 outcomes cannot occur simultaneously;  $A \cap B = \emptyset$ ; e.g. rolling a dice can either be 3 or 5—not both, whereas it being 3 or odd is not mutually exclusive

The **probability** is the sum of the probability of each individual event:

$$P(A_1 \cup A_2 \cup \dots \cup A_k) = \sum_{i=1}^{i=k} P(A_i)$$

$$P(A) = \frac{N(A)}{N}$$

For ordered pairs, number of possible arrangements is:  $N!$

**Permutations** are ordered sequences that are made up by  $k$  elements that are a subset of a set of  $n$  elements.

The notation for **number of permutations** is:  $P_{k,n} = \frac{n!}{(n-k)!}$

**Bayes's Theorem:**  $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

## Non-mutually exclusive events

For non-mutually exclusive events, there can be overlap, so:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

For ordered pairs, number of possible arrangements for k events is:  $\prod_{i=1}^k N(A_i)$

Unordered permutations are known as **combinations** (n choose k). They are denoted:

$$C_{k,n} = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

For unordered pairs, number of combinations is:  $\frac{n!}{k!(n-k)!}$ , where n is the number of objects and

k is the size of the group (pick k, 5, players for the team from n, 8 people. number of permutations?)

**Dependent:** you can't put it back

**Independent:** you can put it back

**Conditional probability:** Probability of A given B:  $P(A|B) = \frac{P(A \cap B)}{P(B)}$

## Chapter 3

### Random variables

rv

- function whose domain is the sample space and whose range is the set of real numbers, but is subject to random variations
- denoted by a capital letter, whereas its values have the same letter as the rv, but lower-case
- can either be [continuous or discrete](#)
- x is a particular value of a [random variable](#)

**Bernoulli:** binary output; can only be either a 0 or a 1

**Probability Mass Function (pmf):** a function that gives the probability that a [discrete random variable](#) is exactly equal to some value

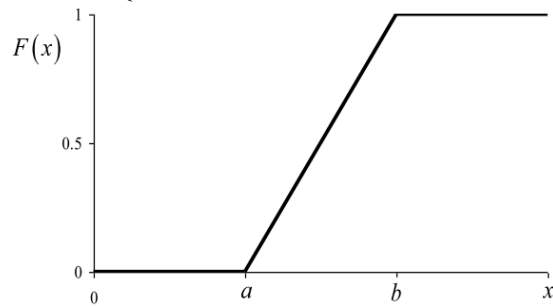
### Cumulative Distribution Function

**CDF:** add up all probabilities within a given range

$$F(x) = P(X \leq x)$$

$$= \int_{-\infty}^x f(y) dy$$

$$= \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ 1, & x > b \end{cases}$$



### Expected Value

- mean using probability of [discrete](#) rv's
- gives same result as population mean
- use if you're not given data, but given probability

$$E(X) = \mu_x$$

- $$= \sum_{x \in D} xp(x)$$
- $$= \int_{-\infty}^{\infty} xf(x) dx$$

- **Variance:** 
$$V(X) = \sum_{x \in D} (x - \mu)^2 p(x) = E[(X - \mu)^2]$$

- **General Expectation formula:** 
$$E(x) = \sum xP(x)$$

### Variance of CDF

$$\text{Var}(X) = E[(X - \mu)^2]$$

$$= \sigma^2$$

$$= \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

$$= \int x^2 f(x) dx - \mu^2$$

### Binomial experiment

1. fixed trial
2. 2 outcomes—success or failure
3. Trials are independent ([with replacement](#))
4. Probability of each outcome is the same for each trial

- If the sample size is at most 5% of the population size, the experiment can be analyzed as though it were a binomial experiment ([with replacement](#)).
- $n$ : repetitions of trials
- $p = P(\text{success in single trial})$
- $q = P(\text{fail in single trial})$
- $x$ : total number of successes
- $$b(x; n, p) = \begin{cases} \binom{n}{x} p^x \underbrace{(1-p)^{n-x}}_q, & x = 0..n \\ 0, & \text{else} \end{cases}$$
- Note: the above notation can be read, where  $x$  is a variable in  $b$  and  $n$  and  $p$  are constants

**Hypergeometric (H.D.):** same as [binomial](#), but dependent ([without replacement](#))

- $N$ : number of items in population
- $M$ : number of successes in population
- $n$ : number of items in sample
- $x$ : number of successes in sample

$$P(X = x) = h(x; n, M, N) = \frac{\overbrace{\binom{M}{x}}^p \overbrace{\binom{N-M}{n-x}}^q}{\underbrace{\binom{N}{n}}_{\substack{\text{removes redundancy} \\ \text{since order isn't important}}}}$$

- $E(X) = n \cdot \frac{M}{N}$  (same as binomial)
- $V(X) = \left( \frac{N-n}{N-1} \right) \cdot n \cdot \frac{M}{N} \cdot \left( 1 - \frac{M}{N} \right)$

**Negative Binomial Distribution:**

- $n$  is fixed in [binomial](#), whereas *here*,  $n$  is random
- trials repeated until success we want
- $r$  is the number of successes you want
- If  $r = 1$ , this is known as a **geometric distribution**

**Poisson distribution:**

- discrete pdf
- number of occurrences of an event in a given interval, given average rate and time (independent), since last event
- $$p(x; \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$$
- $x$ : you are determining the probability that  $x$  things will happen
- $\lambda$  (or  $\mu$ ): average occurrences given population (multiply average rate by population)



- mean = variance =  $\lambda$ , so [S.D.](#) =  $\sqrt{\lambda}$
- $\alpha$  – expected number of events during unit interval
- $t$  – time interval length
- $\lambda = \alpha t$
- $P_k(t) = \frac{e^{-\alpha t} \cdot (\alpha t)^k}{k!}$

**Exponential:** time between events, whereas poisson is more the number of events; continuous distribution

Expected value:  $\frac{1}{\lambda} = \mu$

$$p(x) = \lambda e^{-\lambda x} = \frac{1}{\mu} e^{-\frac{x}{\mu}}$$

For ranges,  $p(a < x < b) = \int_a^b \frac{1}{\mu} e^{-\frac{x}{\mu}} dx$

## Chapter 4

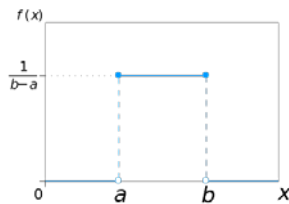
### Probability Density Function

**PDF:** a function that gives the probability that a [continuous](#) random variable is exactly equal to some value, such that:  $P[a \leq X \leq b] = \int_a^b f(x) dx$

Area under whole curve = 1

**Uniform Distribution:** if a [continuous](#) random variable,  $X$ , has a [pdf](#),  $f(x; a, b)$ :

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{else} \end{cases}$$



Note:  $a$  and  $b$  do not represent the entire range of the [PDF](#). Just look at the  $f(x)$  formula above!

To get [pdf](#) from [cdf](#), take the derivative of the [cdf](#).

$$F'(x) = f(x)$$

### Percentile

**Percentile:** percentage of data below you; relative to other data in the range

- $p$ : percentile
- $\eta$ : percentile function
- $p = F(\eta(p)) = \int_{-\infty}^{\eta(p)} f(y) dy$   $z_{0.9}$

$$\mu_x = E(X)$$

$$E(X) = \alpha\beta$$

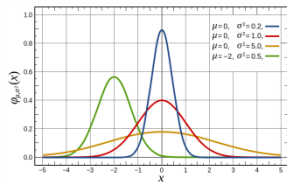
This can be used to determine the probability

## Normal Distribution

A.k.a. population normality

symmetric; mean = median = mode

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$



**Bell curve (a.k.a. Gaussian curve):** normal curve, normal distribution; **Central Limit Theory** says the sampling distribution of sample means will be bell-shaped; s.d. = population s.d./√sample size

## Z-Tables

A.K.A. Standard Normal Cumulative Probability Table

**Z-function:** a standardized [cdf](#) that you use to predict data

- $z_c$ : critical value; this is also the area of the graph from 0 to c, where c is a point on the z-graph
- $z_{c < x} = \frac{x - \mu}{\sigma}$
- It's horizontal units are s.d.'s

If you're given a probability (or percentile), you find the value on the z-table, where the probability represents  $\alpha$  and choose the values at the location. If you cannot find the value on the z-table, find the two closest ones and find the weighted average.

$$E(x) = z_c \frac{\sigma}{\sqrt{n}}$$

**Standardized Score:** a.k.a. "z-score"  $\frac{\text{observed value} - \text{mean}}{\text{s.d.}}$

$\alpha$ -level is the area of the graph of a normal distribution curve  $\alpha = P(Z \geq z_\alpha)$

$Z_\alpha$ : for the standard normal distribution

When trying to find the  $\alpha$  based on a z, make sure you round to the preferred sig figs

**Empirical rule:** you can identify that your data has normal distribution by using the rule that:

- 68% of data is within 1 s.d.'s from mean
- 95% of data is within 2 s.d.'s from mean
- 99.7% is within 3 s.d.'s from mean

- there are 3 s.d.'s from the mean

## Chapter 5

$p \leftarrow$  discrete

$f \leftarrow$  continuous

$$p_x(x) = \sum_y p(x, y), p_y(y) = \sum_x p(x, y)$$

Mean of sum of joint pdf (discrete):  $E(x + y) = \sum_{x, y} (x + y) p(x, y)$

Mean of sum of joint pmf (continuous):  $E(x + y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) \cdot f(x, y) dx dy$

**Covariance:** variance for multiple variables;  $\text{Cov} = E(XY) - \mu_x \cdot \mu_y$

Independent and Identically Distributed (IID):

- form a simple, random sample of size  $n$
- $X_i$ 's are independent r.v.'s
- $X_i$ 's all have same probability distribution

**Multinomial distribution:** represented by the pmf,  $f(x_{1..k}; n, p_{1..k}) = \prod_{i=1}^n \frac{i}{x_i!} p_i^{x_i}$

**Marginal pdf** (continuous):  $f_x(x) = \int_{-\infty}^{\infty} f(x, y) dy, -\infty < x < \infty$   
 $f_y(y) = \int_{-\infty}^{\infty} f(x, y) dx, -\infty < y < \infty$

Conditional probability of joint pdf:  $f(x | y) = \frac{f(x, y)}{f_y(y)}, -\infty < x < \infty$

**Correlation coefficient:**  $\rho_{x, y} = \frac{\text{Cov}(x, y)}{\sigma_x \cdot \sigma_y}$

## Chapter 6

$\theta$  represents the parameter of interest

$\hat{a}$ : variable with a hat means it is an estimate

$\hat{\theta} = \theta + \text{error of estimation}$

- a function of the sample, i.e. rv

**Point estimate:** mean from multiple estimate(s), using the standard error, where  $\theta$  represents parameter of interest (e.g.  $\mu$  or  $\sigma$ ), where you estimate  $\hat{\theta}$ .

**Bias of  $\hat{\theta}$ :**  $E(\hat{\theta}) - \theta$

**Unbiased:**  $E(\hat{\theta}) = \theta$

**Estimator:** the formula

- Should be unbiased (0 avg. error)
  - $\hat{\sigma}^2 = S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$
  - $E(S^2) = \sigma^2$
- Should have minimum variance (i.e. little spread)
- Summary for good estimators: Minimum Variance Unbiased Estimator (MVUE)
  - Unbiased is not always better than minimum variance

**Estimate:** value obtained from the formula after data has been inputted

What is point estimate for each  $\theta$ :

- $\mu : \bar{x}$
- Estimated chance of success  $p : \hat{p} = \frac{\bar{x}}{n}$

**True value:** mean of the population (instead of sample)

Trimmed means will result in **robust estimator**.

**Robust estimators** are less affected by outliers

Standard error of an estimator,  $\hat{\theta}$  is its standard deviation,  $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$

Estimated standard error:  $\hat{\sigma}_{\hat{\theta}}$  or  $s_{\hat{\theta}}$

## Bootstrapping

**Bootstrapping:** fabricating multiple samples from one sample with replacement

- Only works for independent, equally-distributed, random samples
  - Not useful if small data set, lots of outliers (remove outliers first), dependence structures (data based on changing time, etc.)
  - $n^*$  depends on computing capacity, type of problem, and complexity
  - Computed bootstrap value is indicative of the accuracy of your sample. If it is higher than sample, sample is probably higher than actual; if lower than sample, sample is probably lower than actual
1. Compute  $x^*$ , which is from  $x$ , sampled with replacement
  2. Compute  $\hat{\theta}^*$  from  $x^*$

3. Estimate standard error,  $Se_B(\hat{\theta}) = \sqrt{\frac{\sum_{i=1}^B (\theta_i^* - \bar{\theta}^*)^2}{B-1}}$

## e.g. Bootstrapping

Given a sample of:

Given	0.5	1.5	2.5	3.5	4.5
-------	-----	-----	-----	-----	-----

Fabricated Range	0-1	1-2	2-3	3-4	4-5
------------------	-----	-----	-----	-----	-----

Now that you've established a range, you use a random number generator to generate 5 new points.

I randomly generated: 4.5290, 0.6349, 4.5669, 3.1618, 0.4877

Now tally how many are within each range:

Quantity	2	0	0	1	2
----------	---	---	---	---	---

Multiply this quantity by the initial value of the range, pretend that's the new point, and add it up:

$$\begin{aligned}\mu_{\text{boot}} &= 2 \times 0.5 + 0 \times 1.5 + 0 \times 2.5 + 1 \times 3.5 + 2 \times 4.5 \\ &= 13.5/5 \\ &= 2.7\end{aligned}$$

Whereas, the sample average was actually 2.5

**Parametric bootstrap:** note: parameter refers to the population

## Point Estimation

**Point estimation:** 2 main methods: a method of inferring a value for a large population,  $\theta$ , based on a small IID random sample,  $X$ , by calculating standard error

- [Method of moments](#)
- [Maximum Likelihood estimation](#)

Minimum Variance Unbiased Estimator (MVUE)

Estimators

Population mean, $\mu$	Sample mean $\bar{x}$
Population s.d., $\sigma$	Sample s.d., $s$

## Method of Moments

$\alpha$  and  $\beta$  are unknown parameters that yield the estimator

$$k^{\text{th}} \text{ sample moment of } f(x) \text{ is } E(X^k) = \frac{1}{n} \sum_{i=1}^n X_i^k$$

$$\text{Method of Moments Estimator (MME): } \lambda = \frac{1}{E(X^k)} = \frac{1}{\bar{X}}$$

## -Maximum Likelihood Estimation

**Joint pdf:** pdf governing occurrences of A & B, not just one (i.e. pdf of occurrences of multiple potential events), like for regular pdf's

Likelihood function: PMF = PDF =  $P(X;\theta) = f(X;\theta)$  [ $f \Leftrightarrow P$ ]

$$P(X_{i=1..n};\theta) = \prod_{i=1}^n P(X_i;\theta)$$

More popular, easier

Results in normal distribution

**Maximum Likelihood Estimator:** (MLE)  $\hat{\theta} = \max (P(X_i))$  is the random sample, X, with the highest probability of being an appropriate estimator for the population,  $\theta$

How to find:

1. Find  $\ln(P(X_{1..n}; \theta))$ .
2. Find  $\frac{d}{d\theta} [\ln(P)]$ .
3. Equate to 0.
4. Solve for  $\theta$ .

*e.g.*

e.g. for exponential distribution  $f(x_1, \dots, x_n; \lambda) = (\lambda e^{-\lambda x_1}) \dots (\lambda e^{-\lambda x_n}) = \lambda^n e^{-\lambda \sum x_i}$  Note that at \*, the product becomes a summation of the  $x_i$  values

## Chapter 7

Most important chapter for the midterm!

### Confidence interval

**Confidence level:** measures reliability of confidence interval; most popular confidence levels: 90, 95, and 99%; the percent of all samples that will give correct results,  $CL = P(CI)$

**Confidence interval:** interval where certain where data is reliable

- Precision is width of confidence interval
- First determine confidence level
- use the [z tables](#) to find
- In order for this to work:
  - Population distribution is [normal](#)
  - [s.d.](#) given
- Actual mean  $\mu$  does not necessarily have to be in the interval even if the estimated mean is in it

Sample mean  $\pm 1.96$  standard errors

$$P(CI) = 100\% (1 - \alpha)$$

$$CI = \left( \bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right)$$

$$CI = \left( \bar{x} - t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}} \right)$$

The bound of the error is half the width, i.e. if estimate is within 1% of the true percentage, the 1% represents the bound of the error, so the width is  $0.01 \times 2$ .

$$\text{Sample size: } n = \left( 2z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{w} \right)^2 \text{ OR } n = \left( 2z_{\frac{\alpha}{2}} \cdot \frac{1}{w} \right)^2 \hat{p}\hat{q}$$

$$\text{CI: } \left( \hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \right)$$

A larger sample size gives a narrower confidence interval.

A smaller sample size gives a wider confidence interval.

**Standard error:** conversion of standard deviation (total population) to sample distribution

$$(\text{sample population}) = \frac{\text{s.d.}}{\sqrt{n}}$$

**Statistical Inference:** a method of inferring certain statistical characteristics of a population based off a smaller sample, where characteristics could include things, such as sample mean or sample portion

**Sampling variability:** a concept in statistical inference, where even though you are inferring from a sample, each sample's inferred population characteristics can vary from sample-to-sample; the smaller the standard error, the less the sampling variability; the larger the sample size, the smaller the standard error of the mean

**T-Table:** Z-Table, but for s, instead of  $\sigma$ , but you can still use z if sample size  $> 40$ ; uses 2 parameters: degrees of freedom and probability level

## Chapter 8

The point of this to see if the error in the sample mean is low enough to make the sample valid/satisfactory.

**Statistical hypothesis:** assumption about a population characteristic; 2 types:

- [Null Hypothesis](#)
- [Alternative Hypothesis](#)
- Choose the hypothesis based on the [level of significance](#)
  - for lower level, choose Type I / Null
  - for higher level, choose Type II / Alternative

### Null Hypothesis

- $H_0: \mu = \mu_0$ , where  $\mu_0$  is the given value of  $\mu$
- proof by contradiction
- assume it is the thing you think it isn't and prove that wrong
- think *equality*
- If you reject it, the evidence is **statistically significant**

### Alternative Hypothesis

- $H_A$  OR  $H_a$  OR  $H_1$ 
  - $\mu > \mu_0, z \geq z_\alpha$

- $\mu < \mu_0, Z \leq Z_\alpha$
- $\mu \neq \mu_0, \dots$
- specified *range*
- think  $>, <, \text{ or } \neq$
- if you only choose one inequality, it is called a **one-sided hypothesis test**

## Errors

- **Type I:** say something is right when it's wrong
  - **Level of significance** ( $\alpha$ ): P(Type I error)
  - Proving [null hypothesis](#) true
  - Since null hypothesis is a value, P has one value
- **Type II:** say something is wrong when it's right
  - P(Type II error) =  $\beta$
  - Proving [alternative hypothesis](#) true
  - Since alternative hypothesis is a range, P is a range

## Case I

$\sigma$  given (not  $s$ ), normal distribution

$$z = \frac{\bar{x} - \mu_0}{\sigma} \sqrt{n}$$

## Case II

For large  $n$  (i.e.  $n > 40$ ),  $s$  is close to  $\sigma$

$$z = \frac{\bar{x} - \mu_0}{s} \sqrt{n}$$

## Case III

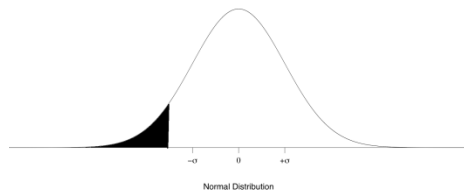
normal dist,  $s$  given

$$\text{Test statistic value: } t = \frac{\bar{x} - \mu_0}{s} \sqrt{n}$$

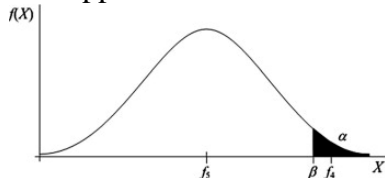
## Hypothesis Test

**One tail:**  $z_\alpha$

Use lower-tail when the alternative hypothesis is:  $\mu < H_a$



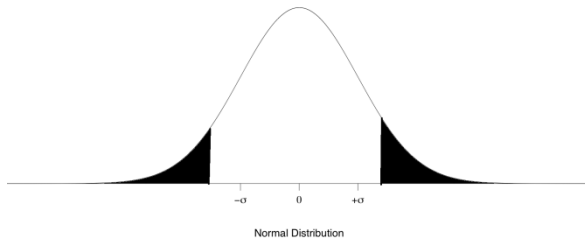
Use upper-tail when the alternative hypothesis is:  $\mu > H_a$





**Two tails:**  $z_{\alpha/2}$

Use this when alternative hypothesis is  $\mu \neq H_a$



The rejection region is the dark part of these graphs. If in rejection region, reject the null hypothesis.

## P-Value

**P-Value:** observed level of significance

**Level of significance ( $\alpha$ ):** a percentage or decimal that represents the cut-off value

- If P-value  $< \alpha$ , reject the null hypothesis and accept the alternative hypothesis
- If p-value  $> \alpha$ , don't reject the null hypothesis and there is not enough information to determine whether or not to accept the alternative hypothesis
  - Just because one thing is wrong, doesn't make the null hypothesis right. Thus, instead of accepting the null hypothesis, you refuse to reject the null hypothesis
- It is different for each region
  - $\Phi(z_\alpha) = \alpha$
  - Upper tailed:  $P = 1 - P(z < z_c)$
  - Lower tailed:  $P = P(z < z_c)$
  - Two-tailed:  $P = 2(1 - P(z < z_c))$

## Chapter 9 – Test Statistics

**Degrees of Freedom:** number of samples – 1

For normal populations with known variances, test statistic value:  $z = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$

$\Delta_0$  is usually 0

**Null hypothesis:**  $|\mu_1 - \mu_2| = \mu_D = \Delta_0$

**Alternative Hypothesis:**  $\begin{cases} \mu_D > \Delta_0 \\ \mu_D < \Delta_0 \\ \mu_D \neq \Delta_0 \end{cases}$

3 Cases (conditions stay the same as before):

Note: n's must be the same, use  $\bar{d} = |\bar{x}_1 - \bar{x}_2|$ , instead of  $\bar{x}$ , use  $\Delta_0$  instead of  $\mu$ ;  $\sigma_D = |\sigma_1 - \sigma_2|$ , and  $s_D = |s_1 - s_2|$

1. [Case I](#)
2. [Case II](#)
3. [Case III](#)

Round down to the nearest integer

Pooled  $t$  happens when  $\sigma_1^2 = \sigma_2^2$

**Margin of Error:**  $E = t_{\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}}$

**$f$  distribution:**

- pdf distribution is too difficult, so we will work with tables
- Assumptions:
  - 2 populations independent
  - Simple random samples
  - Normally distributed
  - Test statistic for test hypothesis, given two variances is:  $f = \frac{s_1^2}{s_2^2}$
- Demonstrates the difference between the two variances
- Determines whether or not the rejection region is too high or not
- Inputs:
  - Significance level,  $\alpha$
- Null Hypothesis:  $\sigma_1^2 = \sigma_2^2$
- Alternative Hypothesis:  $\sigma_1^2 < \sigma_2^2$

## Chapter 12

**Correlation:** a measure of the relationship of your independent variable to your dependent variable

- Defines the relationship between the data points and the equation only true within  $-1$  to  $+1$
- **-1:** implies perfectly negative correlation, i.e. inversely proportional
- **0:** no linear correlation
- **1:** proportional relationship
- **Weak correlation:**  $-0.5 \leq r \leq 0.5$
- **Strong correlation:**  $r \leq -0.8$  OR  $r \geq 0.8$
- **Moderate:**  $-0.8 < r < -0.5$  OR  $0.5 < r < 0.8$

**Causation:** Correlation does not mean causation.

Determine a line with the least variance

**Deterministic Relationship:**  $y = \beta_0 + \beta_1 x$

one variable can be found in terms of the other variable

**Linear:** a first order polynomial example of a deterministic relationship (i.e.  $y = mx + b$ )

**Statistical:** non-deterministic; relies on probability

**Regression Analysis:** looks at correlations between two things by removing other variables

**Model equation:**  $y = \beta_0 + \beta_1 x + \varepsilon$

$\varepsilon$ : measure of variation; error in data

**Principle of least squares:** gives minimum error

$$s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$s_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} =$$

$$\boxed{b_0 = \bar{y} - b_1 \bar{x}}$$

**Point Prediction:** plugging in values of  $x$  into the regression equation

**Residual:** error; vertical deviation from estimated line ( $y - y_0$ )

**Extrapolation:** usually doesn't work, though

library (MASS)

summary() gives 5-number summary

**Sum of Squares for Errors (SSE):**  $SSE = \sum (y_i - \hat{y})^2 = \sum y_i^2 - \hat{B}_0 \sum y_i - \hat{B}_1 \sum x_i y_i$

## Chapter 10

**ANOVA:** ANalysis Of VAriance

**Factor:**

**levels of the factor:**

The number of populations being compared is  $I$ .

$X_{i,j}$  represents the random variable for the  $j^{\text{th}}$  experiment for the  $i^{\text{th}}$  population

$$\text{Total Sum of Squares: (SST)} \sum_{i=1}^I \sum_{j=2}^J x_{ij}^2 - \frac{1}{IJ} \underbrace{x_{..}^2}_{\sum_{i=1}^I \sum_{j=1}^J x_{ij}} = \text{SSTr} + \text{SSE}$$

Total sum squares = treatment + error

**Treatment Sum of Squares: (SSTr)**

**Mean Square Treatment: MSTr** = SSTr / (I - 1), I = # populations (think d.o.f.)

**Mean Square Error: MSE** = SSE / (I(J-1))

$$\text{F-ratio} = \frac{\text{MSTr}}{\text{MSE}}$$

This is your  $\alpha$  for  $F_{\frac{\alpha}{2}}$

## Hypothesis

$H_0 : \mu_1 = \mu_2 = \dots = \mu_I$ , i.e. the  $\mu$  is the same for all populations

$H_a$  : at least 2 values of  $\mu_i$

$H_0$ : E(MSTR) = E(MSE)  $\leftarrow$  rejected if  $F_{\text{calc}} < F_{\frac{\alpha}{2}}$

$H_a$ : E(MSTR) > E(MSE)

For the  $j^{\text{th}}$  value of the  $i^{\text{th}}$  population

$$E(X_{i,j}) = \mu_i$$

$$V(X_{i,j}) = \sigma^2$$