CE 311S Final Review

Probability and Statistics for Civil Engineers

Carlin Liao and Priyadarshan Patil
Teaching Assistants

The University of Texas at Austin

Spring 2020

Standard disclaimer and a note

This review isn't meant to be exhaustive of all the material covered by the course or what might be on the exam, but hopefully it's helpful to you!

I and Prof. Boyles recommend that you handwrite your cheat sheet instead of printing something like this out, since (1) handwriting is proven to improve recall and understanding of material, (2) there may be typos, and (3) there is at least one poorly drawn diagram in these slides that Carlin would be embarrassed for other people to see.

Good luck on the final!

311S in thirds

1. Fundamentals of probability, conditional probability, discrete distributions, CDFs, descriptive statistics

2. Continuous distributions, joint distributions, linear combinations, point estimation

3. Confidence intervals, hypothesis testing, linear regression

Fundamentals of probability, conditional probability, discrete distributions, CDFs, descriptive statistics

Definitions

(This review assumes you know how to find the mean, median, mode, and variance.)

Probability

We know how a random process works, and investigate how it would play out in the world (i.e., in a sample).

Statistics

We see how a random process plays out in the world and try to reconstruct what random process it came from.

Set theory

Union

If $C = A \cup B$ then every element in **either** A **OR** B is in C.

Intersection

If $C = A \cap B$ then every element in **both** A **AND** B is in C.

Complement

A^c denotes the all elements **NOT** in A.

De Morgan's Laws

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

Difference

If $C = A - B$ then C contains all elements of A not in B.

Disjoint

If $A \cap B = \emptyset$ then A and B are *mutually exclusive* or *disjoint*.

Exhaustive

A collection of sets is *exhaustive* if their union covers the entire sample space.

Partition

A collection of sets is a *partition* of the sample space if their union is exhaustive **and** they are all mutually exclusive.

Probability fundamentals

Sample space

A **sample space** $\mathcal{S}$ is the set of all possible **outcomes**, or results of an experiment. An **event** is a subset of the sample space.

Probability

The **probability** of an event A, denoted $P(A)$, is the precise chance that A will occur. It will range between 0 and 1.

Axioms of probability

1. For any event, $P(A) \geq 0$
2. $P(\mathcal{S}) = 1$
3. If $A = \{A_i\}$ is a collection of disjoint events, then $P(\cup A_i) = \sum_i P(A_i)$

Inclusion-exclusion principle

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

Conditional probability and the multiplication rule

The conditional probability of A given B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

This can be arranged to get the multiplication rule:

$$P(A \cap B) = P(A|B)P(B)$$

Tree diagrams are very helpful for solving conditional probability problems.

Law of Total Probability

If $B_1, \ldots, B_k$ are a partition, the probability of any event A is

$$P(A) = \sum_i^k P(A|B_i)P(B_i)$$

Independence

Independence

Two events A and B are independent if $P(A|B) = P(B)$ (equivalently, $P(A \cap B) = P(A)P(B)$) and dependent otherwise.

Law of Total Probability

If $B_1, \ldots, B_k$ are a partition, the probability of any event A is

$$P(A) = \sum_{i}^{k} P(A|B_i)P(B_i)$$

Bayes' Theorem

For any events A and B with $P(B) > 0$,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

Combinatorics

Ordered with replacement

The number of ways to choose k objects from a set of size n is n^k.

Ordered without replacement

Given a set A, a **permutation** is an *ordered* subset of A. The number of permutations of size k from a set of size n is

$$P_k^n = \frac{n!}{(n-k)!}$$

Unordered with replacement

The number of unordered sets of size k with repetition allowed from a set of size n is

$$\binom{n+k-1}{k} = \frac{(n+k-1)!}{k!(n-1)!}$$

Unordered without replacement

Given a set A, a **combination** is an *unordered* subset of A. The number of combinations of size k for a set of size n is

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

Discrete vs. Continuous

A **discrete** random variable is one where all outcomes can be put into a (finite or infinite) list, like the list of all integers. **Continuous** random variables' outcomes can't.

Probability Mass Function (PMF)

A function $P(X = x)$ or $P_X(x)$ that describes the probability of each individual outcome x. A valid PMF must satisfy the following:

1. $P_X(x) \geq 0$ for all possible x
2. $\sum_x P_X(x) = 1$ over all possible x

Cumulative Distribution Function (CDF)

A function $F_X(x)$ that describes the probability of all outcome less than or equal to x.

$$F_X(x) = P(X \leq x) = \sum_{z \leq x} P_X(z)$$

Discrete random variables

Expected value

The mean or expected value of a discrete random variable X is

$$E[X] = \mu_X = \sum_x x P_X(x)$$

Variance and standard deviation

As always, the standard deviation $\sigma_X = \sqrt{V[X]}$ where

$$\begin{aligned} V[X] &= E[(X - \mu_x)^2] \\ &= \sum_x g(x)(x - \mu_X)^2 P_X(x) \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

Discrete random variables

Linear transformations

Given that a and b are not dependent on X,

$$E[aX+b] = aE[x] + b$$
$$V[aX+b] = a^2V[X]$$

Law of the Unconscious Statistician (LOTUS)

For a nonlinear transformation of x, $g(x)$

$$E[g(X)] = \sum_x g(x)P_X(x)$$

Special discrete distributions

Name of special distribution (X)	**Properties**	**What X represents and its range (R_X)**	**Params**	**Probability $P(X=x)$**	**Expected value $E(X)$**	**Variance $Var(X)$**
Binomial distribution	○ You perform $\boldsymbol{n}$ **identical** experiments and you know $\boldsymbol{n}$ before starting ○ Each experiment or trial is **independent** of each other ○ The probability of success, $\boldsymbol{p}$ (or seeing a particular event) is the **same** for every trial	The number of successes observed $\boldsymbol{R_X}$: $\{0,1,2,\ldots,n\}$	n,p	$\binom{n}{x}p^x(1-p)^{n-x}$	np	$np(1-p)$
Hypergeometric distribution	○ I sample k objects from a **finite** population with b success and r failures ○ Each sample of k objects is **equally likely** to be chosen	The number of successes observed $\boldsymbol{R_X}$: $\{0,1,\ldots,min\{b,k\}\}$	k,b,r	$\frac{\binom{b}{x}\binom{r}{k-x}}{\binom{b+r}{k}}$	$\frac{kb}{b+r}$	$\frac{kbr(b+r-k)}{(b+r)^2(b+r-1)}$
Negative binomial distribution	○ You perform identical experiments and you **don't know** n before starting, rather you keep performing experiments until you observe m **successes** ○ Each experiment or trial is **independent** of each other ○ The probability of success, p is same for every trial	The number of trials before m-th success is observed $\boldsymbol{R_X}$: $\{m, m+\boldsymbol{1},\ldots,\infty\}$	m,p	$\begin{cases}\binom{x-1}{m-1}p^m(1-p)^{x-m} \\ 0\end{cases}$	$\frac{m}{p}$	$\frac{m(1-p)}{p^2}$
Poisson distribution	○ An event occurs at an average rate of λ such that: ○ Occurrences of the event are **independent** of each other ○ More than one of these events **can't** occur simultaneously	The number of events $\boldsymbol{R_X}$: $\{0,1,2,\ldots,\infty\}$	λ	$\frac{e^{-\lambda}\boldsymbol{\lambda}^x}{x!}$	λ	λ

Continuous distributions, joint distributions, linear combinations, point estimation

Continuous distributions

Probability density functions (PDF)

$$P(a \leq x \leq b) = \int_a^b f(x)dx$$

The density function must be nonnegative for all x and can take values greater than 1 at any point, but $\int_{-\infty}^{\infty} f(x)dx$ must equal 1. The probability that x takes any one specific value is zero.

Cumulative distribution function (CDF)

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

This is the area to the left of x in the PDF. The PDF is the derivative of the CDF.

Continuous distributions

Mean

Given some random variable X and its PDF $f(x)$,

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

For $Y = g(X)$,

$$E[Y] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

Percentile

Let $0 < p < 1$. The $(100p)$th percentile $\eta(p)$ is the value such that $F(\eta(p)) = p$. The median of a continuous random variable is the 50th percentile.

We can find a PDF for some random variable X by

1. Finding the CDF and deriving it
2. Using the method of transformations on a known PDF

Method of transformations

Given $Y = g(X)$ and $g(x)$ is strictly increasing or decreasing,

$$f_Y(y) = \frac{f_X(x)}{|g'(x)|}$$

where x is the value of X that corresponds with y. If no such value exists, $f_Y(y) = 0$. If $g(x)$ is not strictly increasing or decreasing, you can cut up the function into n sections where it is, then sum them together.

$$f_Y(y) = \sum_{i=1}^{n} \frac{f_X(x_i)}{|g'(x_i)|}$$

Special continuous distributions

Name of special distribution (X)	**Properties**	**Range ($\boldsymbol{R_X}$) and Parameters**	**PDF (it is not probability X=x)** $\boldsymbol{f(x)}$	**CDF** $\boldsymbol{F(x) = P(X \leq x)}$	**Expected value** $\boldsymbol{E(X)}$	**Variance** $\boldsymbol{Var(X)}$
Uniform distribution	○ If all possible values of X are equally likely to occur ○ And all values fall in the range $[a, b]$	$\boldsymbol{R_X}: [\boldsymbol{a}, \boldsymbol{b}]$ Parameters: a, b	$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & otherwise \end{cases}$	$F(x) = \begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x \geq b \end{cases}$	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$
Exponential distribution	○ X represents the waiting time between Poisson distribution events ○ X has property of being memoryless which means $P(X \geq x + x_0 \mid X \geq x_0) = P(X \geq x)$	$\boldsymbol{R_X}: [\boldsymbol{0}, \infty)$ Parameters: λ	$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & otherwise \end{cases}$	$F(x) = 1 - e^{\{-\lambda x\}}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normal distribution	○ When the pdf of X has a ________-curve shape ○ t-distribution is the normal distribution but wider (reflecting more uncertainty)	$\boldsymbol{R_X}: [-\infty, \infty)$ Parameters: μ, σ (and degrees of freedom for t-dist)	$f(x) = \frac{1}{\sqrt{\{2\pi\}}\sigma} e^{-(x-\mu)^2/2\sigma^2}$	Use z-table, t-table, or R	μ	σ^2
Gamma distribution (Chi-squared distribution)	○ Has two special cases: exponential distribution ($\alpha = 1$) and chi-squared distribution ($\alpha = \frac{\nu}{2}, \lambda = \frac{1}{2}$) ○ Gamma function ○ $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$ ○ $\Gamma(n) = (n-1)!$ If pos integer ○ $\Gamma(1/2) = \sqrt{\pi}$	$\boldsymbol{R_X}: [0, \infty)$ Parameters: α, λ (Just ν for X^2)	$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x^{\alpha-1} e^{-\lambda x} & x \geq 0 \\ 0 & otherwise \end{cases}$	Use R or chi-squared table	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$

Jointly distributed random variables (discrete)

The joint PMF or joint distribution will be given as a table with a random variable and its possible values on each axis.

Marginal PMF

We find the PMF of X using the joint distribution (X, Y) with

$$P_X(x) = \sum_y P_{XY}(x, y)$$

We can use this to find the expected value of X.

Independence

Two variables in a joint distribution are independent if

$$P_{XY}(x, y) = P_X(x) P_Y(y)$$

for every possible x and y.

Jointly distributed random variables (continuous)

Joint density function

$f_{XY}(x, y)$ is valid if

1. $f_{XY} \geq 0$ for all x and y
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) dx dy = 1$

Marginal density function

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy$$

Jointly distributed random variables (continuous)

Independence

Two variables in a joint distribution are independent if

$$f_{XY}(x,y) = f_X(x)f_Y(y)$$

for all x and y.

Expectation

For some $Z = g(x,y)$

$$E[Z] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f_{XY}(x,y)dydx$$

Covariance

Covariance

The covariance of two random variables X and Y is

$$\begin{aligned}\text{Cov}(X, Y) &= E[(x - E[X])(y - E[Y])] \\ &= E[XY] - E[X]E[Y]\end{aligned}$$

Properties

1. $\text{Cov}(X, X) = \text{Var}[X]$
2. X, Y independent means $\text{Cov}(X, Y) = 0$ (but not always the other way around)
3. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
4. $\text{Cov}(aX, Y) = a\text{Cov}(X, Y)$
5. $\text{Cov}(X + c, Y) = \text{Cov}(X, Y)$
6. $\text{Cov}(X + Z, Y) = \text{Cov}(X, Y) + \text{Cov}(Z, Y)$

Correlation coefficient

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y}$$

Properties

1. $\rho_{XY} \in [-1, 1]$
2. If $\rho_{XY} = 1$, then $Y = aX + b$ for some $a > 0, b$
3. If $\rho_{XY} = -1$, then $Y = aX + b$ for some $a < 0, b$
4. If $\rho_{XY} = 0$, there's no *linear* correlation between X and Y
5. If $\rho_{XY} \in (0, 1)$ there's a positive linear relationship between X and Y
6. If $\rho_{XY} \in (-1, 0)$ there's a negative linear relationship

Linear combinations

For any random variables $X_1, \ldots, X_n$

$$E[a_1X_1 + \cdots + a_nX_n] = a_1E[X_1] + \cdots + a_nE[X_n]$$

$$V[a_1X_1 + \cdots + a_nX_n] = \sum_{i=1}^{n}\sum_{j=1}^{n} a_ia_j\mathrm{Cov}(X_i, X_j)$$

Variance of a sum

For a special case of linear combinations,

$$V[aX + bY] = a^2V[X] + b^2V[Y] + 2ab\mathrm{Cov}(X, Y)$$

Statistics

What is a statistic?

Consider a sample of n elements, and let X_i describe the variable of the i-th member of the sample. A statistic is a random variable Y which is determined from the random variables $X_1, \ldots, X_n$.

Random samples

The random variables $X_1, \ldots, X_n$ are a random sample if they are independent and identically distributed (iid). If each X_i has mean μ and variance σ^2, then the expected value and variance of the mean of the random sample $\overline{X}$ are μ and σ^2 as well.

Central limit theorem

Given the random sample $X_1, \ldots, X_n$, if n is sufficiently large $\overline{X}$ has an approximately normal distribution with the same mean and variance as each X_i regardless of the distribution of X_i.

Sample variance

Sample variance

$$S^2 = \frac{\sum(x_i - \overline{x})^2}{n-1}$$

Note the $n-1$ is necessary to prevent bias.

Confidence intervals, hypothesis testing, linear regression

Note on standard deviations

When we find the mean $\bar{x}$ and standard deviation s of some sample, we're estimating the population mean μ and standard deviation σ.

In hypothesis testing, we're usually testing if the population mean μ is likely to be what we think it is (the null hypothesis μ_0) based on our data. That requires us to check the likelihood of pulling the sample mean $\bar{x}$. We're testing a mean and taking a mean squeezes the variance, which is why we divide by $\sqrt{n}$. Assuming μ_0 is true, $\bar{x}$ comes from the distribution $N\left(\mu_0, \frac{\sigma}{\sqrt{n}}\right)$ (replacing σ with s if we aren't given σ and $n > 40$) .

Note on sample size requirements

When we're directly applying the Central Limit Theorem, we require $n > 30$ before we allow ourselves to use the z-table instead of the t-table. At this threshold, the sample mean $\overline{x}$ is usually a good estimate of the population mean μ.

This requirement jumps up to $n > 40$ when we start doing confidence intervals and hypothesis testing because, in addition to assuming that the sample mean is a good estimate of the population mean, we have to make the **additional assumption** that the sample standard deviation s is a good estimate of the population standard deviation σ. Because of this additional condition we require more samples before we start assuming that the normal distribution is a good approximation.

Interpretation of the confidence interval

There is a $(1 - \alpha)$ chance that the confidence interval contains the true population metric.

There is **NOT** a $(1 - \alpha)$ chance that the true population metric is in your confidence interval.

The true population metric can't move. It's not random. Your estimate and your confidence interval do move because your sample is random. (Wider intervals and more confident because your net is bigger so it can move further and still catch the correct, fixed value.)

Confidence intervals

A confidence interval (CI) gives a plausible range for a **population** parameter p by using outcomes from a sample

$(1-\alpha)100\%$ two-sided CI finds LB and UB such that $\boldsymbol{P(LB \le p \le UB) = 1-\alpha}$ and the CI is denoted by (LB, UB)

$(1-\alpha)100\%$ one-sided lower bound CI finds LB such that $\boldsymbol{P(LB \le p) = 1-\alpha}$ and the CI is denoted by (LB, ∞)

$(1-\alpha)100\%$ one-sided upper bound CI finds LB such that $\boldsymbol{P(p \le UB) = 1-\alpha}$ and the CI is denoted by $(-\infty, UB)$

Given: $X_1, X_2, \dots, X_n$ are random samples i.e. they are **independent and identically distributed**

Assumption	X_i's are from a **normal distribution** with unknown mean and **known** variance	X_i's are from **any distribution** with unknown mean and unknown variance and n>40	X_i's are from **normal distribution** with unknown mean and unknown variance and n<40
Population parameter to be estimated	Mean	Mean	Mean
$(1-\alpha)100\%$ two-sided confidence interval	$\left(\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$	$\left(\bar{x} - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right)$	$\left(\bar{x} - t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}\right)$
Sample size for CI of width w	$\left(\frac{2z_{\alpha/2}\sigma}{w}\right)^2$	$\left(\frac{2z_{\alpha/2}s}{w}\right)^2$	N/A
$(1-\alpha)100\%$ one-sided upper bound CI	$\left(-\infty, \bar{x} + z_{\alpha} \frac{\sigma}{\sqrt{n}}\right)$	$\left(-\infty, \bar{x} + z_{\alpha} \frac{s}{\sqrt{n}}\right)$	$\left(-\infty, \bar{x} + t_{\alpha,n-1} \frac{s}{\sqrt{n}}\right)$
$(1-\alpha)100\%$ one-sided lower bound CI	$\left(\bar{x} - z_{\alpha} \frac{\sigma}{\sqrt{n}}, \infty\right)$	$\left(\bar{x} - z_{\alpha} \frac{s}{\sqrt{n}}, \infty\right)$	$\left(\bar{x} - t_{\alpha,n-1} \frac{s}{\sqrt{n}}, \infty\right)$

Fancier confidence intervals

Given: $X_1, X_2, \dots, X_n$ are random samples i.e. they are **independent and identically distributed**

A confidence interval (CI) gives a plausible range for a **population** parameter p by using outcomes from a sample

A prediction interval (PI) gives a plausible range for a **single future prediction value**

A tolerance interval (TI) gives a plausible range which **contains at least k%** of the entire population

Assumption	X_i's are from **normal distribution** with unknown mean and unknown variance	X_i's are from **normal distribution** with unknown mean and unknown variance	X_i's are from **normal distribution** with unknown mean and unknown variance and n<40
Interval	Confidence interval on population variance	Prediction interval for X of a single individual	Tolerance interval containing at least k% of population
Notations	n: sample size $\chi^2_{\alpha,n-1}$: chi-squared value for α significance with n-1 df s: sample standard deviation	$\bar{x}$: sample mean n: sample size s: sample standard deviation $t_{\frac{\alpha}{2},n-1}$: t-value, α/2 sig, n-1 df	$\bar{x}$: sample mean n: sample size s: sample standard deviation $C_{\alpha,k}$: C-value, α sig, k of pop
$(1-\alpha)100\%$ two-sided interval	Lower limit: $(n-1)s^2/\chi^2_{\frac{\alpha}{2},n-1}$ Upper limit: $(n-1)s^2/\chi^2_{1-\frac{\alpha}{2},n-1}$	Lower limit: $\bar{x} - t_{\frac{\alpha}{2},n-1}\, s\sqrt{1+\frac{1}{n}}$ Upper limit: $\bar{x} + t_{\frac{\alpha}{2},n-1}\, s\sqrt{1+\frac{1}{n}}$	Lower limit: $\bar{x} - C_{\alpha,k}s$ Upper limit: $\bar{x} + C_{\alpha,k}s$ (two-sided C values)
$(1-\alpha)100\%$ one-sided upper bound interval	Lower limit: $-\infty$ Upper limit: $(n-1)s^2/\chi^2_{1-\alpha,n-1}$	Lower limit: $-\infty$ Upper limit: $\bar{x} + t_{\alpha,n-1}\, s\sqrt{1+\frac{1}{n}}$	Lower limit: $-\infty$ Upper limit: $\bar{x} + C_{\alpha,k}s$ (one-sided C value)
$(1-\alpha)100\%$ one-sided lower bound interval	Lower limit: $(n-1)s^2/\chi^2_{\alpha,n-1}$ Upper limit: $+\infty$	Lower limit: $\bar{x} - t_{\alpha,n-1}\, s\sqrt{1+\frac{1}{n}}$ Upper limit: $+\infty$	Lower limit: $\bar{x} - C_{\alpha,k}s$ Upper limit: $+\infty$

Hypothesis testing

Population Parameter	Population mean
Null hypothesis	$\mu = \mu_0$
Test statistic	$\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ OR $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ based on if population variance is known
Case 1	$n > 40$ OR $n < 40$ AND population distribution is normal AND population variance is known
Case 2	$n < 40$ AND population distribution is normal AND variance not known
Otherwise	Consult a knowledgeable statistician

Alternate Hypothesis test type	Rejection region Case 1	Rejection region Case 2
$H_a\colon \mu > \mu_0$ Or $H_a\colon p > p_0$	(z_α, ∞)	$(t_{\alpha,n-1}, \infty)$
$H_a\colon \mu < \mu_0$ Or $H_a\colon p < p_0$	$(-\infty, -z_\alpha)$	$(-\infty, -t_{\alpha,n-1})$
$H_a\colon \mu \neq \mu_0$ Or $H_a\colon p \neq p_0$	$(-\infty, -z_{\frac{\alpha}{2}}) \cup (z_{\frac{\alpha}{2}}, \infty)$	$(-\infty, -t_{\frac{\alpha}{2},n-1}) \cup (t_{\frac{\alpha}{2},n-1}, \infty)$

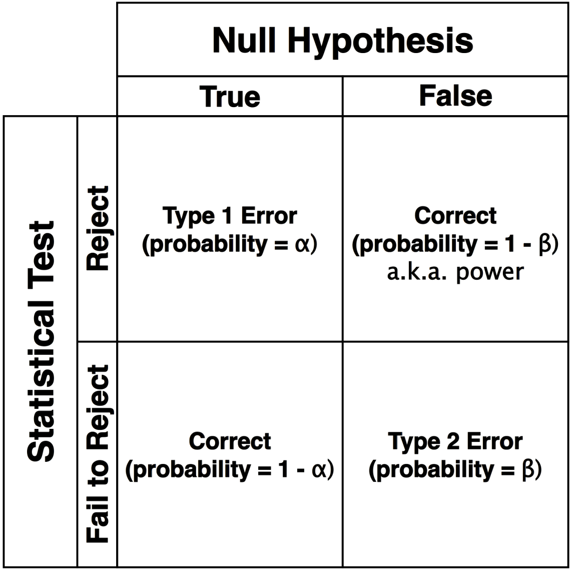
Reject if test statistic belongs to rejection region

Population Parameter	Sample proportion
Null hypothesis	$p = p_0$
Test statistic	$\frac{p' - p_0}{\sqrt{p_0(1-p_0)/n}}$
Case 1	$np_0 > 10$ AND $n(1-p_0) > 10$
Otherwise	Consult a knowledgeable statistician

Hypothesis test type	P-values Case 1
$H_a\colon \mu > \mu_0$ Or $H_a\colon p > p_0$	$1 - \phi(z)$
$H_a\colon \mu < \mu_0$ Or $H_a\colon p < p_0$	$\phi(z)$
$H_a\colon \mu \neq \mu_0$ Or $H_a\colon p \neq p_0$	$2[1 - \phi(\lvert z\rvert)]$

Reject if $\alpha > P$ value

Type I and Type II errors



		Null Hypothesis	
		True	False
Statistical Test	Reject	Type 1 Error (probability = α)	Correct (probability = 1 - β) a.k.a. power
	Fail to Reject	Correct (probability = 1 - α)	Type 2 Error (probability = β)

Calculating Type II error

Notice that calculating Type II error has *nothing to do* with the sample mean $\overline{x}$ or sample standard deviation s. Instead you need a null hypothesis mean μ_0, a population standard deviation σ, another mean you're assuming to be the true mean (let's call it μ_t), and the sample size n.

We start by looking at the distribution centered around μ_0.

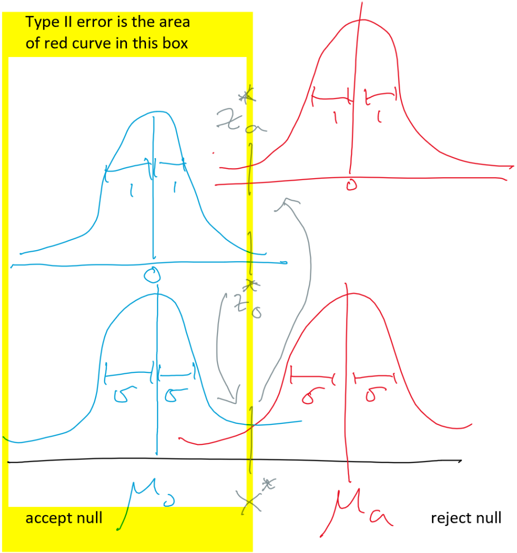
1. Take the confidence level α and turn it into a critical z-score z_0^* (or t-score t_0^*) (or two if it's a two-sided confidence interval). Identify your accept-the-null region and reject-the-null region(s) based on which sides of your z^* score(s) is okay.
2. Convert the z_0^* from the $N(0,1)$ scale into a real value using μ_0 and *sigma*, i.e. $z^* \times \frac{\sigma}{\sqrt{n}} + \mu_0$. Call this value x^* and remember which sides around it are the acceptance and rejection regions. We just did a unit conversion from z-scores to whatever your sample is measured in.

Calculating Type II error

Now we can define our acceptance and rejection regions in terms of real x values. To finish finding Type II error, we need to find the area under the distribution centered at μ_t that overlaps with the acceptance region. (Recall that a Type II error happens when we accept the null hypothesis when we should reject it.)

4. Next we're going to convert x^* from real units back into $N(0, 1)$ scale, but around our assumed mean μ_t instead of μ_0. Call this value z_a^*. Find this by $z_a^* = \frac{x^* - \mu_t}{\frac{\sigma}{\sqrt{n}}}$.
5. Now that we have a z_a^* and we know which sides of z_a^* are accept or reject, we can use the z-table to calculate the area under the μ_t curve that falls into the accept-the-null region. That is the probability of a Type II error.

Calculating Type II error



Type II error is the area of red curve in this box

accept null

reject null

P values

P values are the likelihood that, given that the null hypothesis is true (i.e., that you're in the distribution centered around the null hypothesis μ_0) and you've draw a sample as large as the one you have n, you would draw a sample with a mean as or more extreme than the one you actually got $\overline{x}$.

To find P values, instead of using an area α to find some critical z^* (or t^*), we turn our $\overline{x}$ into a z-score (or t-score) and find the area on the other side of this z. (For one sided hypothesis tests, that's just the area under the distribution between z or t and the closest infinity. For two-sided, multiply that by 2 thanks to symmetry.)

We reject the null hypothesis if the P value is less than α (but that doesn't mean the P value is the probability that the null hypothesis is true or that it's the probability that you made an error). This is basically another way to look at hypothesis testing. Instead of converting the significance level α from proportion/area units into z-score units as the crtical z-value z^* so we can compare it with the z-score for our sample mean, we convert the z-score of our sample mean into proportion/area units p.

Linear regression preliminaries

$$S_{xy} = \sum_i x_i y_i - \frac{\left(\sum_i x_i\right)\left(\sum_i y_i\right)}{n}$$

$$S_{xx} = \sum_i x_i^2 - \frac{\left(\sum_i x_i\right)^2}{n}$$

Sum of squared errors (SSE)

Given some model we use to predict y-values from x-values $f(x)$, the sum of squared errors is

$$SSE = \sum_i [y_i - f(x_i)]^2$$

Linear regression model

Given an independent variable x and a dependent variable y, we model a linear relationship between the two as

$$Y = \beta_0 + \beta_1 x + \epsilon$$

Where $\epsilon \approx N(0, \sigma^2)$ is some "random deviation." (More on this later.)

We model this linear relationship as

$$\hat{y} \approx f(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

These $\hat{\beta}$s were found to construct the function that minimizes the SSE.

For nonlinear relationships we can try transforming the variables before applying linear regression (e.g., $y = ae^{xb} \rightarrow \log y = \log a + bx$). Multiple linear regression with more than one independent variable is also possible (e.g., $y = ax^2 + bx \rightarrow x_1 = x^2, x_2 = x$).

Hypothesis testing the linear regression model

β_1 hypothesis test

We estimate the variance of the random deviation as

$$\sigma^2 = \frac{SSE}{n-2}$$

where n is the size of the sample. This gives us a t-score for β_1,

$$t = \frac{\hat{\beta}_1}{S/\sqrt{S_{xx}}}$$

This can be used to test the null hypothesis that $\beta_1 = 0$, implying that their is no (linear) relationship between x and y. The critical t value(s) will be subject to some α of your choosing and $n-2$ degrees of freedom.

Note: Correlation ($\beta_1 \neq 0$) doesn't imply causation but it does suggest that something might be going on.

Goodness of fit

Total sum of squares

The total sum of squares (SST) is what the error would be if we just predicted a constant (the mean of y) for all values of x.

$$SST = S_{yy} = \sum_i (y_i - \bar{y}])^2 = \sum_i y_i^2 - \frac{(\sum_i y_i)^2}{n}$$

Coefficient of determination

The coefficient of determination (r^2) basically measures how well your linear model does compared to just guessing the average $\bar{y}$ value for all inputs x.

$$r^2 = 1 - \frac{SSE}{SST}$$

Higher is better. $r^2 = 1$ means your linear model perfectly predicts y. Negative values means that your model does worse than guessing a constant.