p = prob of success
q = prob of fail = 1-p

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

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

$P(A \cap B) = P(A|B)P(B)$
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
$P(|3| > 2) = p(3 < -2) + p(3 > 2)$
$P(|x-2| < 3) = P(-3 < x-2 < 3)$

If independent, $P(A \cap B) = P(A)P(B)$
$P(A \mid B) = P(A)$
If mutually exclusive, $P(A \cup B) = P(A) + P(B)$

Bayes = $P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$

$E(aX + b) = aE(X) + b$
$E(X +/- Y) = E(X) + E(Y)$
$var(aX + b) = a^2\, var(X)$
$var(X +/- Y) = var(X) +/- var(Y)$, if ind. X,Y

param = numerical characteristic
of population

Discrete rv
pmf = f(x) = P(X=x) for all x in X
1. f(x) ≥ 0 for all x in X
2. Σf(x) = 1

cdf = F(x) = P(X ≤ x) = Σf(x)
mean = Σxf(x)
var = $\Sigma(x-\mu)^2 f(x)$

Continuous rv
pdf = P(a<x<b) = ∫(a->b) f(x)dx
1. f(x) ≥ 0 for all x
2. ∫(-∞ -> ∞) f(x)dx = 1

cdf = F(x) = P(X ≤ x) = ∫(-∞ -> x) f(t)dt
mean = ∫(-∞ -> ∞) x f(x) dx
var = ∫(-∞ -> ∞) $(x-\mu)^2$ f(x)

Uniform
pdf = f(x) = 1/(b-a), a ≤ x ≤ 1
μ = (a+b) / 2
$\sigma^2 = (b-a)^2 / 12$

Exponential
pdf = f(x) = $\lambda e^{-\lambda x}$
μ = 1/λ
$\sigma^2 = 1/\lambda^2$

Sampling distribution
μ = μ
σ = σ/sqrt(n) = standard err

Central limit theorem (sample mean/avg)
$X_1,\ldots,X_n$ is random sample of population w/
μ, σ^2 when n ≥ 20,
$\bar{x} = X_1,\ldots,X_n / n \sim N(\mu, \sigma^2/n)$

Poisson (discrete) X ~ Po(λ) where λ>0
Prob of events occurring in a fixed interval of time/space
Events occur w/ a known constant rate and independently
X = # of occurrences in a given interval/space has a poisson
λ = rate of occurance
pmf = P(X=x) = $\lambda^x e^{-\lambda}/x!$ *non-overlapping events*
μ = λ
$\sigma^2 = \lambda$

Poisson approximation to Binomial
if n ≥ 20 and np < 5
Bin(n,p) ~ Pos(np)
np = λ

if CLT asks for
sum not avg…
μ = nμ
$\sigma^2 = n\sigma^2$

Normal approximation to Binomial
If np ≥ 5 and n(1-p) ≥ 5
X ~ N(np, np(1-p))
μ = np
σ^2 = np(1-p)

original: X~N(μ, σ^2)

Poisson exponential distribution
T is time b/w 2 consecutive occurrences
pdf = f(t) = $\lambda e^{-\lambda t}$ t ≥ 0
cdf = F(t) = 1 - $e^{-\lambda t}$ t ≥ 0
μ = 1/λ
$\sigma^2 = 1/\lambda^2$

Normal approximation to Poisson
If λ ≥ 20
X~Po(λ) -> X ~ N(λ, λ)
μ = λ
$\sigma^2 = \lambda$

Bernoulli distribution
X(success) = p = 1
X(fail) = q = 1-p = 0
pmf = p(X=x) = $p^x(1-p)^{1-x}$ where x = 0,1
μ = p
σ^2 = p(1-p)

Binomial distribution X~Bin(n,p)
# of success for n independent trials
X ~ Bin(num of trails n, success probability p)
pmf = p(X=x) = $\binom{n}{k} p^k (1-p)^{n-k}$ $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
μ = np
σ^2 = np(1-p)

Continuity correction
normal -> poisson
discrete -> continuous
P(X ≤ 5) = P(X ≤ 5.5)
P(X ≥ 5) = P(X ≥ 4.5)

Geometric distribution X~Geo(p)
# of individual trials needed until the first success occurs w/ success probability of each trial p
pmf = f(x) = P(X=x) = $p(1-p)^{x-1}$
cdf = F(x) = 1 - $(1-p)^x$, x ≥ 1
μ = 1/p
σ^2 = 1 - p / p^2

cdf -> pdf
∫f(x) = F(x)

pdf -> cdf
F’(x) = f(x)

CI for population μ
true σ is known?
test stat = $z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ ~ N(0,1)

true σ is unknown? or n < 30?
test stat = $t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$ ~ t_{n-1}

- if n≥30, you can replace t_{n-1} to Z
- inc. width CI = inc (1-α) or n

$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$, $\bar{x}$ = point estimator
RHS = margin of error

CI for population $\mu_1 - \mu_2$

$$(\bar{x}_1 - \bar{x}_2) \pm t^*(s_p)\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$$

$$t^* = t_{\left(\frac{\alpha}{2}, df\right)}, \; df = n_1 + n_2 - 2$$

- assume random samples, independent, normal or big n for CLT, $\sigma^2_1 = \sigma^2_2$

Left tail
$H_0: \mu_1 - \mu_2 \geq 0$
$H_a: \mu_1 - \mu_2 < 0$

Left tail
$H_0: \mu_1 - \mu_2 \leq 0$
$H_a: \mu_1 - \mu_2 > 0$

Two tail
$H_0: \mu_1 - \mu_2 = 0$
$H_a: \mu_1 - \mu_2 \neq 0$

large test stat = small p-val = good = H_a is likely true
p-val is calculated based on H_0 is true

μ_0 is μ_{true}

CV approach
α = 0.05 (0.025 for 2 tail)

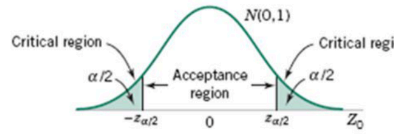
if |test stat| > $t_{\alpha,\ n-1}$ reject H_0

* remember to replace t with Z if n ≥ 30 or population is normal, Z_α -> (1-α))

p-value approach
p-val = $2p(\bar{x} \le x_{sample})$, 2 tail
p-val = $1-p(\bar{x} \le x_{sample})$, right tail
p-val = $p(\bar{x} \le x_{sample})$, left tail
x_{sample} = test stat = z, t, etc

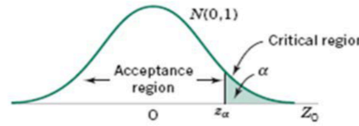
find test stat value within t_{n-1} row then its column is your p-value

if p-val ≤ α, reject H_0



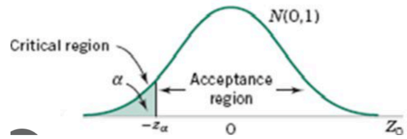
- Two tail test:

$H_0 : \mu = \mu_0$
$H_1 : \mu \ne \mu_0$



- Upper tail test

$H_0 : \mu \le \mu_0$
$H_1 : \mu > \mu_0$



- Lower tail test

$H_0 : \mu \ge \mu_0$
$H_1 : \mu < \mu_0$

t-test assumptions
- data is randomly sampled from normally distributed, independent

ANOVA
overall variability = b/w grp SS + w/i grp SS
True = Between
Error = Within

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares (MS)	F
Within	$SSW = \sum_{i=1}^{a}(n_i-1)s_i^2$	$df_w = k-1$	$MSW = \frac{SSW}{df_w}$	$F = \frac{MSB}{MSW}$
Between	$SSB = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$	$df_b = n-k$	$MSB = \frac{SSB}{df_b}$	
Total	SST = SSW + SSB	$df_t = n-1$		

combined mean = $\overline{x_{12}} = \frac{n_1\overline{x_1} + n_2\overline{x_2}}{n_1+n_2}$

$d_1 = \overline{x_{12}} - \overline{x_1}$; $d_2 = \overline{x_{12}} - \overline{x_2}$

combined variance = $\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1+n_2}}$

if $f_{obs/calculated} > F_{Trdf,\ Erdf,\ \alpha}$ reject H_0
num, denom, α

Non-parametric tests

useful when pop. dist. is unknown or sample size is too small for CLT. Requires no assumption about distribution of data

Sign test
- one-sample, paired
- assume data is independent
- H_0: median = Θ (postulated median)
- H_a: median ≠ Θ => two tail test
  median > Θ => one tail test
- under null hypothesis, T~Bin(n, 0.5) => median has p=0.5
- paired tests? get difference of pair and use difference's signs

1. Test stat t= 4, there were 4 values bigger than Θ
2. p-val = P(T ≤ t) = P(T ≤ 4) = 0.0059, with α=0.05, reject H_0

Wilcox rank sum test
- two-sample, use if sample size is small, non-normal
- assume data is independent, symmetric (under H_0), continuous
- H_0: $f_A = f_B$ *implies c = 0*
- H_a: $f_A(x) = f_B(x - c)$ where c ≠ 0 => two tail test
  c > 0 => one tail test

1. rank all data (x,y) in order smallest to largest
   - to rank, order the absolute values and assign rank from 1 to n (assign mean rank when there are ties)
2. calculate sums for each group
3. calculate each group's test stat = sum of rank - n(n+1) / 2
4. test stat = min test stat, compare w lookup

- p-value by hand:

1. W_y = sum of ranks of y = 22 + 23 = 45 (there are 24 ranks)
2. Assuming, one-sided hypothesis, how many ranks are ≥45?
3. (22,23), (21,24), (22,24), (23,24) => 4 wavs
4. How many ways to rank 2 y and 22 x? $\binom{24}{2} = \binom{24}{22} = 276$
5. 276 equally likely possibilities under H_0 => 4/276 = p-val

Sign rank test
- one-sample, paired
- assume distribution is symmetric and continuous
- paired two-sample:
  1. calculate $y_i - x_i$ for i = 1,…,n
  2. assign ranks based on their $|y_i - x_i|$ for i = 1,…,n
     - if ties, assign average of the ranks
     - if diff is 0, remove that pair and reduce n by one
  3. sum all ranks for which $y_i > x_i$ (don't add if $y_i < x_i$)
  4. test stat = sum, compare w lookup
- one-sample:
  1. calculate xi - postulated median
  2. assign ranks
  3. sum ranks of positives and sum ranks of negatives
  4. test stat = min of rank sums, compare w lookup

to that

Kruskal-Wallis test
- test statistic = H = $H = \frac{12 \sum_{i=1}^{k} n_i(\bar{R}_i - \bar{R})^2}{n(n+1)}$

group rank mean
overall rank mean
\# group observations
\# total observations

- non-parametric ANOVA
- compare test stat chi sq w/ $\chi^2_{k-1,\alpha}$ dist, k = # of groups
- overall rank mean = sum / n = n(n+1) / 2n

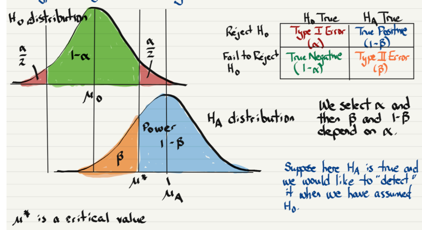
Power of a test
- tests for 1-β: reject H_0 when we should
- we want low T1E and T2E but we can't have both…
  hence, we often go with small α in exchange for large β

Permutation test

- assume observations are ind. and data is from same distribution

1. compute test stat for observed data
2. re-group data and re-calculate test stat on the new groups for all possible groups
3. count how many of the re-groupings produce a test statistic as extreme or more so than the observed one
4. calculate the total number of groupings possible
5. what proportion were extreme?

- test whether distributions are the same or not
  test whether means are the same or not
  test whether difference in means is 0/set value or not



The true $p = 0.75$. H_0 is that $p = 0.5$, and $\alpha = 0.01$; $n = 150$.
We will reject H_0 if

$$z = \frac{\hat{p} - p_0}{SD(\hat{p})} = \frac{\hat{p} - 0.5}{0.0408} > z^* = 2.33 \quad \text{or} \quad \hat{p} > 0.595$$

$$P(\hat{p} > 0.595 \text{ when } p = 0.75) \qquad SD(\hat{p})^{HA} = \sqrt{\frac{(0.75)(0.25)}{150}}$$

$$= P\left(\frac{\hat{p} - p}{SD(\hat{p})} > \frac{0.595 - p}{SD(\hat{p})}\right) \qquad = 0.0354$$

$$= P\left(z > \frac{0.595 - 0.75}{0.0354}\right) = P(z > -4.38) = 0.99999$$

Hence, over 150 trials if the true success rate is 75% we are nearly certain of making the right decision if we reject H_0 when $\hat{p} > 0.595$.

1. $\hat{p}$ = ratio of the number of successes
2. $sd(\hat{p}) = \sqrt{p_0(1-p_0)/n}$
3. $z = (\hat{p} - p_0) / sd(\hat{p}) = 0.1$
4. $p(z > 0.8)$