Basics

	Sample statistic	Population Parameter
Mean	$\overline{X}$	μ
Variance	S^2	σ^2
Correlation	r	ρ
Noise	e_i	ϵ_i
	Guess	True, but unknown

The Mean, Variance and StdDev are subject to outliers.

$$\begin{aligned}
\overline{X} =& \frac{1}{n}\sum_{i=1}^{n} x_i \\
\mu = E(W) =& E(a+bX) = a + bE(X) \\
\sigma^2 = Var(X) =& E[X^2] - \mu_x^2 \\
S^2 =& \sum_{all\ i} \frac{(X_i - \overline{X})^2}{n-1} \\
\sigma = StdDev(X) =& \sqrt{\sigma^2} \\
Var(X) =& E[(X-\mu)^2] = E(X^2) - \mu^2 \\
=& \sum_{allx} (x-\mu_x)^2 P(X=x) \\
=& E(X^2) - E(X)^2 = E(X - E(X))^2 \\
Var(X+c) =& Var(X) \\
Var(cX) =& c^2 Var(X) \\
Var(\bar{X}) =& \frac{\sigma^2}{n}
\end{aligned}$$

If X,Y are dependent

$$\begin{aligned}
Var(X+Y) \neq& Var(X) + Var(Y) \\
Var(X+Y) =& Var(X-Y) \\
Var((a+bX)+(c+dY)) =& b^2 Var(X) + d^2 Var(Y) + 2bdCov(X,Y) \\
Var(aX+bY) =& a^2 Var(X) + b^2(Y) + 2abCov(X,Y) \\
=& a^2 Var(X) + b^2(Y) + 2ab\sigma_x\sigma_y\rho_{xy}
\end{aligned}$$

Quartiles split the data into 4 equal groups by number of values, or 25% percentiles. Q2 is the median.

Study Design

Types of variables: categorical vs. quantitative, dummy, discrete vs continuous. Observational vs Experiments. Only experimentals can show causal effects. Key factors to an experiment: control group, randomized, and replicated. Usually 30 replications is sufficient. Avoid confounding by having cross-over groups.
Scope of Inference: **Interval Validity**: low when 1) unaccounted confounding factors, 2) ignored missing data 3) noncompliance 4) unverified assumptions 5) suboptimal method of analysis. Applicable when allocation of units to groups is random. **External Validity**: high when can be generalized to population. Applicable when selection of units is random.

Covariance and Correlation

Covariance gives direction.
Correlation gives direction and strength.
Both are *linear*.

$$\begin{aligned}
Cov(X,Y) =& \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) \\
Cor =& \frac{Cov}{\sigma_x \cdot \sigma_y}
\end{aligned}$$

Probabilities

$0 \leq$ All probabilities ≤ 1
Mutually exclusive and Exhaustive

$$\begin{aligned}
P(\overline{A}) =& 1 - P(A) \\
P(A \text{ or } B) =& P(A) + P(B) - P(A \text{ and } B) \\
P(A|B) =& \frac{P(A \text{ and } B)}{P(B)} \\
P(B) =& \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B|A_i)((A_i) \\
\binom{n}{k} =& \frac{n!}{k!(n-k)!}
\end{aligned}$$

Joint vs. Marginal probabilities
Independent if $P(A|B) = P(A)$
For Independent only: $P(E \text{ and } F) = P(E) \cdot P(F)$

	B	$\overline{B}$
A	$P(AandB) = P(B)P(A\|B)$	$P(Aand\overline{B}) = P(\overline{B})P(A\|\overline{B})$
$\overline{A}$	$P(\overline{A}andB) = P(\overline{A})P(B\|\overline{A})$	$P(\overline{A}and\overline{B}) = P(\overline{A})P(\overline{B}\|\overline{A})$

all success	p^n
all failure	$(1-p)^n$
at least one failure	$1-p^n$
at least one success	$1-(1-p)^n$

Random Variables

$$\begin{aligned}
P(X \leq x) \rightarrow& CDF \\
E(cX) =& c \cdot E(X) \\
E(X+c) =& E(X) + c \\
E(X+Y) =& E(X) + E(Y)
\end{aligned}$$

If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ then $\bar{X} \sim N(\mu_{\bar{X}} = \mu, \sigma^2_{\bar{X}} = \sigma^2/n)$.
If pop dist is not normal, S^2 will eventually be Normal but may not be independent.

Bias of an Estimator

In practice n has to relatively much larger like > 100.
Guesses should be *unbiased* and have *minimum variance*.
MVUE (Minimum Variance, Unbiased Estimates).

$$bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Unbiased if $bias = 0$ (expected value equals true, not a particular value of $\bar{x}$).
For samples, we divide by $n-1$ instead of n to make an unbiased estimator. The guess would otherwise be too low.

$$\begin{aligned}
E(X) =& \sum_{allx} xP(X=x) \\
E(\overline{X}) =& \sum_{i=1}^{\infty} x_i p_i = \mu \\
E(S^2) =& \sigma^2
\end{aligned}$$

$$\begin{aligned}
E[X+c] =& E[X] + c \\
E[X+Y] =& E[X] + E[Y] \\
E[aX] =& aE[X] \\
E[aX+bY+c] =& aE[X] + bE[Y] + c \\
E((a+bX)+(c_dY)) =& a + bE(X) + c + dE(Y)
\end{aligned}$$

Example: Roulette has a \$1 bet with a \$35 payoff for $\frac{1}{38}$ odds.

$$E[\text{gain from a \$1 bet}] = -\$1 \cdot \frac{37}{38} + \$35 \cdot \frac{1}{38} = -\$0.0526$$

Uniform Distribution

$$\begin{aligned}
E(X) =& \frac{(a+b)}{2} \\
Var(X) =& \frac{(b-a)^2}{12}
\end{aligned}$$

y axis should be a fraction to make area $= 1$

Normal Distribution

$$\begin{aligned}
X \sim& \mathcal{N}(\mu, \sigma^2) \\
Z =& \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1) \\
X =& \sigma Z + \mu \\
\Phi(z) =& \text{distribution} \\
P(a \leq X \leq b) =& P[(a-\mu) \leq (X-\mu) \leq (b-\mu)] \\
=& P[\frac{(a-\mu)}{\sigma} \leq \frac{(X-\mu)}{\sigma} \leq \frac{(b-\mu)}{\sigma}] \\
=& P[\frac{(a-\mu)}{\sigma} \leq Z \leq \frac{(b-\mu)}{\sigma}]
\end{aligned}$$

Linear combination of normals is normal. Used by CLT.

Binomial Distribution

- n independent trials
- binary result
- same probability of success
- total number of successes

$$\begin{aligned} X \sim & Bin(n,p) \\ \binom{n}{x} = & \frac{n!}{x!(n-x)!} \\ P(X=x) = & \frac{n!}{x!(n-x)!}p^x q^{(n-x)} \\ \mu_x = E(X) = & n \cdot p \\ \sigma^2 = Var(X) = & n \cdot p \cdot q = n \cdot p \cdot (1-p) \\ X \sim & N(\mu = np, \sigma = \sqrt{np(1-p)}) \\ \hat{p} \sim & N(\mu = p, \sigma\sqrt{p(1-p)/n}) \end{aligned}$$

Shape of distribution depends on p, n.
Small p, left skewed. Large p, right skewed

Hypothesis Test - General

A **scientific hypothesis** makes a testable statement about the observable universe.
A **statistical hypothesis** is more restricted. It concerns the behavior of a measurable (or observable) random variable. It is often a statement or claim about a parameter of a population or distribution.

1. Formulate hypotheses (H_0, H_a, α)
2. Calculate test statistic and reference distrib.
3. Calculate p-value based on reference distribution of test statistic $|H_0$. p **value is NOT the probability that the null hypothesis is true.** It is probablity of seeing our data on the null hypothesis.
4. Determine conclusion and scope of inference

Note: t values are measures of *distance*. p values are measure of *probability*. Assumptions? Independence? Distribution of Observations? Parametric vs. non-parametric?

Hypotheses	Decision Rule
$H_0 : \mu = \mu_0$ $H_a : \mu \neq \mu_0$	If $\lvert t_{stat}\rvert > 1.96$, reject H_0
$H_0 : \mu = \mu_0$ $H_a : \mu < \mu_0$	If $t_{stat} < -1.64$, reject H_0
$H_0 : \mu = \mu_0$ $H_a : \mu > \mu_0$	If $t_{stat} > 1.64$, reject H_0

Assumptions:

1. Observations are **independent**. This is true if the sample is randomly selected, but false if there is bias in the sampling.
2. **Normal Distributions** of the observations. This happens when the sample is large enough (via the Central Limit Theorm) or it is known that the population is normally distributed.

If these assumptions are not correct, none of the t-tests will work correctly.

Hypothesis Test - Mean

Calculation by hand using a t test:

$$t_{stat} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Fisher's Randomization Test

For casual inference in experiments. No assumption as to the underlying distribution has to be made. Generally, is $\mu_1 = \mu_2$? The only assumption is Additive Treatment Effect (δ):

$$Y_{c,i} = Y_{v,i} + \delta$$

$H_0 : \delta = 0$, ie zero treatment effect for all units. $H_a : \delta \neq 0$ non-zero treatment effect for ALL units. Reference distrib built through simulation. Non-parametric.
Assumptions:

- Random assignment to groups
- Under H_0, *independence, interchangability* of study units.

Test Statistic: Difference of outcomes of the two groups: (could also use difference between the medians)

$$\hat{\delta} = \overline{Y}_c - \overline{Y}_v$$

The maximum number of possible simulations is the binomial distribution:

$$\binom{t}{s}$$

where t is the total number of study units and s is the number where the effect was shown.
Calculating p value: The p value is proportion of values above or below the observed test statistic.

$$\begin{aligned} \hat{\delta} = & 0.5 \\ p = & P(\overline{Y}_c - \overline{Y}_v \geq \bar{y}_c - \bar{y}_v) \\ = & P(\overline{Y}_c - \overline{Y}_v \geq 0.5) = 0.029 \end{aligned}$$

so there is sufficient evidence to reject the null.
Scope of inference: Internal Validity maybe not if the same hospital. External validity: possibly not, as these were volunteers.

Permutation Test

For observational studies, comparing two groups - for generalization. Means of observations are presumed the same $E(Y_{1,i}) = \mu_1, E(Y_{2,j}) = \mu_2 . H_0 : \mu_1 = \mu_2$.
Test stat: $\hat{D} = \bar{Y}_i - \bar{Y}_1$. Reference distrib by simulation. Assumes independence of study units and groups, nonparametric, similar shapes and spreads.

All Things χ^2

$$\begin{aligned} \chi^2_1 = & Z^2 \\ \chi^2_n = & \sum_{i=1}^{n} Z_i^2 \\ t_n = & \frac{Z}{\sqrt{\chi^2_n/n}} \\ F_{n,m} = & \frac{\chi^2_n/n}{\chi^2_m/m} \\ \chi^2_{n+m} = & \chi^2_n + \chi^2_m \\ \frac{(n-1)S^2}{\sigma^2} \sim & X^2_{df=n-1} \\ S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2 \sim & \frac{\sigma^2}{n-1}\chi^2_{df=n-1} \end{aligned}$$

The t distribution

The t distribution has a slightly wider spread in the tails than Standard Normal. Also known as student-t distribution. This happens when you don't know the real σ and are forced to use S, which is less reliable.

$$\overline{X} \approx N(\mu, \frac{\sigma^2}{n})$$

allows us to Z-score

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

but we don't know σ so in practice we use

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

(which has two random variables ($\overline{X}$ and S). Now, take that previous definition of T and add $\sigma/\sqrt{n}$ to the top and bottom:

$$\begin{aligned} T = & \frac{\overline{X} - \mu}{S/\sqrt{n}} \\ = & \frac{(\overline{X} - \mu)/(\sigma/\sqrt{n})}{(S/\sqrt{n})/(\sigma/\sqrt{n})} \\ = & \frac{Z}{\sqrt{S^2/\sigma^2}} \\ = & \frac{Z}{\sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}}} \\ = & \frac{Z}{\sqrt{\chi^2/df}}, \text{ since } \chi^2 = \frac{(n-1)S^2}{\sigma^2} \end{aligned}$$

Assumptions: 1) Observations are independent (randomly selected), 2) Normal distrib of observations (either > 30 via CLT or pop. is known $\sim N(\mu, \sigma^2)$).

Confidence Intervals

CI is Estimate ± margin of error or Estimate ± (z value) × (SD of estimate).

$$\mu_0 = \overline{X} \pm t \cdot s/\sqrt{n}$$

If the interval spans 0, then one is not significantly bigger than the other (or cannot be determined). As long as $n > 30$, it doesn't matter if the sample size is different between the random variables.

To find the 95% Confidence Internval for the **true mean** μ:

$$\bar{x} \pm t^*_{df=n-1}(s/\sqrt{n})$$

t^* has to be looked up based on the reference distrib.

Types of Errors

Type I the null hypothesis is rejected when it is true

Type II the null hypothesis is accepted when it is false

Power the probability of correctly rejecting the null when the alternative is actually true.

α is *level of significance* - probability of making a Type I error. The greater the cost of an error, the smaller α should be. β is the probability of making a Type II error. There is an inverse relation between Type I and II errors. Reducing one increases the other. The only way to reduce both is to increase n the sample size.

Pooled Two Sample t-Test

assumes equal variances, different means, all observations are independent, random, > 30 or known normal distrib.

$$H_0 : \mu_1 - \mu_2 = 0$$
$$T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2 | H_0)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$
$$T \sim t_{df=n_1+n_2-2}$$
$$CI : (\bar{x}_1 = \bar{x}_2) \pm t^* \left(S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$
$$S_p^2 = \frac{(n_1 - 1)S_1^2) + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Do not use if $\frac{S_1^2}{S_2^2} \geq 2$.

Unpooled Two Sample t-Test

assumes unequal variances, all observations independent, random, > 30 or known normal distrib.

$$H_0 : \mu_1 - \mu_2 = 0$$
$$T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2 | H_0)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$
$$T \sim t_{df=min(n_1,n_2)-1}$$
$$CI : (\bar{x}_1 = \bar{x}_2) \pm t^* \left(S_p \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right)$$

Advantage Pooling: slightly more power to detect a difference. Disadvantage pooling: extra assumption could be wrong, with a high type I error rate.

Paired Test

This when there are two samples that are **not** independent, e.g. Weight Watchers, Before / After or matched, shared characteristics. Is the data matched or independent? If we don't take into account the match, the results are wrong. Applies when there are two obs for each study unit.
To account for this, take the difference between $\overline{X}_1 - \overline{X}_2$ and then do a hypothesis test on the *difference*.
Basically it is a one-sample test on the differences.
$H_0 : \mu_D = 0$. $T \sim t_{df=n_D-1}$

Comparing Two Proportions

(Not really covered in this class, but may be useful).

$$Var(\hat{p}_1 - \hat{p}_2) = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}$$

The 95% confidence interval for $p_1 - p_2$ is:

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Decision Rules for Testing Two Proportions:

$$T = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}, \text{ where } \hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

$\hat{p}$ is called the *pooled proportion.*

Comparing Two Normal Distributions

$$A \sim \mathcal{N}(\mu_A, \sigma_A^2)$$
$$B \sim \mathcal{N}(\mu_B, \sigma_B^2)$$

Find $P(A < B + c)$:

$$P(A - B - c < 0)$$
$$(A - B - c) \sim \mathcal{N}(\mu_A - \mu_B - c, \sigma_A^2 + \sigma_B^2 - 2Cov(A, B)$$
$$P(X = A - B - c > 0)$$

then Z-score

Assumptions

Robustness: able to be useful when assumptions are violated. Check for violations graphically (QQplot, box plot, density) and look at the study design. Under the null and correct assumptions, p-values should be uniformly distributed. t-test is very robust when sample sizes are large, or both groups have the same size. Not so good when one group is very skewed. Fix with unpooled t test, data transformations, or non-parametric tests.

F-Distribution and F-tests

$$R = \frac{X/n_x}{Y/n_y} \sim F_{n_x,n_y}$$

Hypothesis: $H_0 : \sigma_x^2 = \sigma_y^2, H_a : \sigma_x^2/\sigma_y^2 \neq 1$.

Test statistic:

$$F = \frac{S_X^2}{S_Y^2} \sim F_{n_x-1,n_y-1}$$

Transformations

Purpose: make data more symmetric

If Right skewed, use: $1/Y, log(Y), \sqrt{Y}$. Left skewed: $Y^2, e^Y, (n - Y)$

If done on log scale: 1) be cautious about ≤ 0. Use ratios of medians as exp(mean) will not be useful. Consider rank transform.

Rank Sum Test

non-parametric, 2 indep. groups. Rank ALL data (2 groups combined) and sum up the ranks in one of the groups. Average ties. H_0 : medians are the same. Reference distrib: small n simulate like permutation test; large n approx normality if $n_j \geq 10$:

$$T \sim N(n_x\bar{R}, S_R^2\frac{n_x n_y}{n_x + n_y})$$

Sign Test - Paired Data