Basics

	Sample statistic	Population Parameter
Mean	$\bar{x}$	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.

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

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

Covariance and Correlation

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

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

Most general combination of random variables:

$$
\begin{aligned}
E((a+bX)+(c_dY)) =& a + bE(X) + c + dE(Y) \\
Var((a+bX)+(c+dY)) =& b^2Var(X) + d^2Var(Y) + 2bdCov(X,Y)
\end{aligned}
$$

Probabilities

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

$$
\begin{aligned}
P(\bar{A}) =& 1 - P(A) \\
P(A \text{ or } B) =& P(A) + P(B) - P(A \text{ and } B) \\
P(A|B) =& \frac{P(A \text{ and } B)}{P(B)}
\end{aligned}
$$

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

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

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

All Things χ^2

$$
\begin{aligned}
\chi_1^2 =& Z^2 \\
\chi_n^2 =& \sum_{i=1}^{n} Z^2 \\
t_n =& \frac{Z}{\sqrt{\chi_n^2/n}} \\
F_{n,m} =& \frac{\chi_n^2/n}{\chi_m^2/m} \\
\chi_{n+m}^2 =& \chi_n^2 + \chi_m^2 \\
\frac{(n-1)S^2}{\sigma^2} \sim& X_{df=n-1}^2
\end{aligned}
$$

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.

$$E(\bar{x}) = \sum_{i=1}^{\infty} x_i p_i = \mu$$

$$E(s^2) = \sigma^2$$

$$
\begin{aligned}
E[X+c] =& E[X] + c \\
E[X+Y] =& E[X] + E[Y] \\
E[aX] =& aE[X] \\
E[aX+bY+c] =& aE[X] + bE[Y] + c
\end{aligned}
$$

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

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

Hypothesis Test - General

The purpose of hypothesis testing is to help the researcher reach a conclusion about a population by examining the data contained in a sample.

H_0 is default position, the status quo. It requires significant evidence to be disproven.

Hypotheses	Decision Rule
$H_0 : \mu = \mu_0$ $H_a : \mu \neq \mu_0$	If $\|t_{stat}\| > 1.96$, 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

We use 1.96 because it is 2.5% on either side. We use 1.64 because it is 5% on a single side.

Hypothesis Test - Mean

Calculation by hand using a t test:

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

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

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.

Comparing Two Sets - General

The null hypothesis is always $H_0 : p_1 = p_2$.

Hypotheses	Decision Rule	Stata Diff $(p_1 - p_2)$
$H_0 : p_1 = p_2$ $H_a : p_1 \neq p_2$	If $\lvert T \rvert > 1.96$, reject H_0	$H_a :$ diff $\neq 0$
$H_0 : p_1 = p_2$ $H_a : p_1 < p_2$	If $T < -1.64$, reject H_0	$H_a :$ diff < 0
$H_0 : p_1 = p_2$ $H_a : p_1 > p_2$	If $T > 1.64$, reject H_0	$H_a :$ diff > 0

If the interval is all positive then $\hat{p}_1 > \hat{p}_2$. If the interval is all negative then $\hat{p}_1 < \hat{p}_2$. 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.

Comparing Two Proportions

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

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 Means

Requirements:

1. σ_1 and σ_2 are unknown. No assumption made about their equality.
2. The two samples are independent.
3. Both samples are simple random samples.
4. The two samples size are both large (ie. > 30) or both populations have normal distributions.

A confidence interval for $(\mu_1 - \mu_2)$ is

$$(\bar{x}_1 - \bar{x}_2) \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Comparing Two Normal Distributions

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

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

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

then Z-score

Matched Pairs

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. To account for this, take the difference between $\overline{X}_1 - \overline{X}_2$ and then do a hypothesis test on the *difference.*

$$\begin{aligned} H_0 :& \mu_D = 0 \\ H_a :& \mu_D > 0 \end{aligned}$$

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

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

Binomial Distribution

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

$$\begin{aligned} X \sim& B(n, p) \\ \binom{n}{x} =& \frac{n!}{x!(n-x)!} \\ P(X = x) =& \frac{n!}{x!(n-x)!}p^x q^{(n-x)} \\ \mu_x = E(X) =& n \cdot p \\ \sigma^2 = Var(X) =& n \cdot p \cdot q = n \cdot p \cdot (1-p) \end{aligned}$$

Shape of distribution depends on p, n.

Small p, left skewed. Large p, right skewed