Master Meriadoc

```
import scipy.stats as st
```

Math Help

$a^{-n} = \frac{1}{a^n}$, $a^{m/n} = n\sqrt{a^m}$,
$a^{1/n} = n\sqrt{a}$, $a^{mn} = (a^m)^n$
$\log(x^{-2}) = -2\log(x), \log_a a^n = n,$
$\log(bc) = \log(b) + \log(c)$, $\log(\frac{1}{b}) = -log(b)$
$n^0 = 1$, $1! = 1$, $0! = 0$,

Descriptive Stats - central tendency

Arithmetic Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Median: if n is odd then its the $\left(\frac{n+1}{2}\right)^{th}$ term otherwise the average of the $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n+1}{2}\right)^{th}$ terms

Mode: most freq occurring value Freq data mean: $\bar{x} = \frac{1}{\sum f_i}\sum_{i=1}^{n} f_i x_i$

Example

value	freq	relative freq	cum rel freq
24	5	0.0746	0.0467
25	10	0.1493	0.2239
26	28	0.4179	0.6418
27	24	0.3582	1.000

Mode is 27. Median is 26 because 50% of items are less than or equal to 26 Mean: Each value is multiplied by its frequency then the products are summed

$$\bar{x} = \frac{1746}{67} = 26.0597$$

	mean	median	mode
unique	yes	yes	no
simple	yes	yes	yes
scale	numerical	ordinal	nominal
outliers	affected	not aff	not aff

Symmetrics of Distns:
positively or right skewed distn (i.e. mass left) indicates that $med < mean$ where negatively skewed distn indicates $mean < med$.

Descriptive Stats - dispersion

Range: Max value - Min value

Percentiles: The p^{th} percentile is the value V_p s.t. p percent of the distribution is $\leq V_p$. i.e. 50% of the distn is $\leq$ the median. The median is the 50th percentile.

The p^{th} percentile is defined as,
the $(k+1)^{th}$ largest item in the sample if $np/100$ is not an integer (where k is the largest integer less than $np/100$

Also it is the average of the $(np/100)^{th}$ and $(np/100+1)^{th}$ largest item in the sample if $np/100$ is an integer

Example: n = 9 and we want the 20th percentile (V_{20}). $9(0.2) = 1.8$. V_{20} = the $(1+1)^{th} = 2^{nd}$ largest sample point. Example: n = 20 and we want the 10th percentile (V_{10}). $20(0.1) = 2$. The 10th percentile is the average of the 2nd and 3rd ordered sample points.

Sample variance:
$s^2 = \frac{\sum x_i^2 - \left((\sum x_i)^2/n\right)}{n-1} = \frac{SS}{df}$
variance measures variability about the sample mean.

Standard Deviation: $s = \sqrt{s^2}$
Standard deviation is in the same units as the variable

```
np.array([1,3,5,7]).var(ddof=1)
6.667
np.array([1,3,5,7]).std(ddof=1)
2.582
```

Coefficient of variation:
$c.v. = \frac{s}{\bar{x}} * 100$

Grouped Data

Tabulate or display in frequency tables or histograms

Grouped Mean: $\bar{x}_g = \frac{\sum_{i=1}^{k} f_i m_i}{\sum_{i=1}^{k} f_i}$
where m_i is the midpoint

Grouped variance:

$$s_g^2 = \frac{\sum_{i=1}^{k} f_i m_i^2 - \left[\left(\sum_{i=1}^{k} f_i m_i\right)^2/n\right]}{n-1}$$

Grouped Standard Dev: $s_g = \sqrt{s_g^2}$

Graphical Methods

Bar Graphs: Used for displaying grouped data. Data divided into groups and frequency is determined within each group.

Histograms: Widely used for displaying grouped data.

Stem and Leaf Plot:
data = c(5.2, 5.8, 3.8, 2.8, 4.1)

Stem	Leaf
2	8
3	8
4	1
5	28

Stem is usually all but right most digit. Separate ea. data point into stem and leaf. When writing stems include those w/o values. Stem and leaf plots are easier to construct, don't need to order or group data and we actually see the points (as oppose to histograms).

Box plots: Graphical technique to compare mean and median. Describes skewness using 25th and 75th percentiles. Give a feel for the spread of the data, and helps identify outliers.

Probability

The sample space if the set of all possible outcomes. An event is any set of outcomes of interest. The probability of an even is the relative freq of this set of outcomes over an indefinitely large number of trials. If outcomes A and B cannot happen at the same time then $P(AorB) = P(A) + P(B)$ (mutually exclusive).

P(Two heads in three tosses) = P(HHT or HTH or THH) = 1/8 + 1/8 +1/8 = 3/8

Two events are said to be independent in case the occurrence of one does not influence the prob of the occurance of the other. If A and B are independent then P(A and B) = P(A) P(B)

Union: $(A \cup B)$ is the event that either A or B occurs or both. $A = (X < 90), B(90 \leq X \leq 95)$, then A or $B = (A \cup B) = X < 95$

Intersect: $(A \cap B)$ is the event that both A and B occur simultaneously. example: $A = (250 \leq chol \leq 299), B = (chol \geq 300), C = (chol \leq 280)$

A and B are mutually exclusive. A and C are not. If $P(A) = 0.2$, and $P(B) = 0.1$. then $P(chol \geq 250) = P(250 \leq chol \leq 299) + P(chol \geq 300) = 0.2 + 0.1 = 0.3$. Also $(A \cap C) = (250 \leq chol \leq 280)$.

Law of Probability: P(A or B) = P(A) + P(B) - P(A and B)

Conditional Probability
$P(A|B)\frac{P(A \cap B)}{B}$
if $P(B|A) = P(B) = P(B|\text{not}A)$ then the events are independent.

Relative Risk: The RR of B given A is,
$RR = \frac{P(B|A)}{P(B|A_c)}$
Example: Let A represent positive TB test and B represent having TB. 1 in 10,000 negative tests has TB, P(B— not A) = 0.0001. 1 person in 100 that have positive skin tests have TB, P(B—A) = 0.01. The RR is thus $\frac{0.01}{0.0001} = 100$. People with positive skin tests are 100 times more likely to have TB as those with negative skin tests.

Event Probability: For any events A and B. $P(B) = P(B|A)P(A) + P(B|A_c)P(A_c)$

Example: A is skin test. B is TB. $P(A^+) = 0.01, P(B|A^-) = 0.0001, P(B|A^-) = 0.01$. What is prob randomly selected person will have TB?

$$P(B) = P(B|A^+)P(A^+) + P(B|A^-)P(A^-)$$
$$= 0.01(0.01) + (0.0001)(0.99) = 0.0002$$

We interpret this as approx 2 persons in each 10,000 will have TB.

Note: In the general case P(B) = $\sum_{i=1}^{n} P(B|A_i)P(A_i)$ assuming the events are exclusive and exhaustive

predictive value positive: Prob person has disease given the test was positive. $PV^+ = P(D^+|T^+)$

predicitve value negative: Prob person does not have diease given test was negative $PV^- = P(D^-|T^-)$

Sensitivity: Prob that test positive given person has disease $P(T^+|D^+)$

Specificity: Prob that test negative given person does not have disease $P(T^-|D^-)$

False Negative: Person who tests negative but actually has the disease. $P(T^-|D^+) =$ 1 - sensitivity

False Positive: Person who tests positive but is actually does not have disease. $P(T^+|D^-) = 1$ - specificity

Prevalance: $d = P(D^+)$

Bayes Rule:
$PV^+ = P(D^+|T^+)$
$= \frac{P(D^+)P(T^+|D^+)}{P(D^+)P(T^+|D^+)+P(D^-)P(T^+|D^-)}$
$= \frac{d*sensitivity}{d*sensitivity+(1-d)*(1-specificity)}$
Example: Given sensitivity = 0.84, specificity = 0.77, prevalence = 0.20
$PV^+ = \frac{(0.2)(0.84)}{(0.2)(0.84)+(0.8)(0.23)} = 0.48$
$PV^- = \frac{(0.8)(0.77)}{(0.8)(0.77)+(0.2)(0.16)} = 0.95$

Discrete Probability Distributions

Probability Mass Fn: rule that assigns a prob to each possible values of a discrete r.v. X.

Expected Value:

$$E(X) = \mu = \sum_{i=1}^{k} x_i P(X = x_i)$$

X	$P(X = x_i)$
0	0.008
1	0.076
2	0.265
3	0.411
4	0.240

An example pmf.
The expected value is,
$E(X) = \mu = (0)(0.008) + (1)(0.076) + (2)(0.265) + (3)(0.411) + (4)(0.24) = 2.80$

Population Variance:

$$var(X) = \sigma^2 = \sum_{i=1}^{k} (x_i - \mu)^2 p(X = x_i)$$
$$= \sigma^2 = E(X - \mu)^2$$
$$= \sum_{i=1}^{k} X^2 P(X = x_i) - \mu^2$$

The population standard deviation is the sqrt of the variance.

Cumulative Distn Fn: $F(x) = p(X \leq x)$

Permutation - order matters. For all permutations of size k from a set A with n

elements (order matters)
$\frac{nPk}{P_k^n} = \frac{n!}{(n-k)!}$
Combination - set containing a certain num of objects of another set (order does not matter) to find nCk combinations of k things taken from a set of n things,
$nCk = \binom{n}{k} = \frac{n!}{(n-k)k!}$

```
from scipy.misc import comb
comb(5,2)}
```

Binomial Distribution:
$\sim \mathbf{Bin}(n,p) = \binom{n}{k}p^k(1-p)^{n-k}; k = 0...n$, mean $= p(X=k) = np$, var $= np(1-p)$
note: pmf is prob of exactly k successes
Example1: prob of success = 0.6. n = 5. The p(of exactly 2 successes)
$= \binom{5}{2}0.6^2(1-0.6)^{5-2} =$ Calculator: [2nd] [DISTR] [binompdf] – binompdf(n,p,[x])
Example2: What is prob of 2 boys out of five children? p = 0.51? binompdf(5,0.51,2) = .306
Example3: n = 20, prob of disease = 0.05 Out of 20 how often do we expect at least 3 to develop the disease? binomcdf(20,0.05,2) = 0.0754
Poisson Distribution:
$\sim \mathbf{Pois}(\theta) = p(X=x) = \frac{e^{-\theta}\theta^x}{x!}$;
$\theta = 0, ..\infty; x = 1, 2...$ and mean $= \theta =$ var
$\theta = \lambda\Delta t$. We can think Δt as a small unit of time and P(1 death in $\Delta t = \lambda\Delta t$ for some constant λ. The prob of observing 0 deaths over δt is approx $1 - \lambda\Delta t$. And the prob of observ more than 1 over this time interval is essentially 0. We assume that (1) the prob of a new death from typhoid in any one day is very low and (2) that the number of cases reported in any two distinct periods are independent r.v.'s. So in this case the expected num of events in 1 year is $\theta = \Delta t = (4.6)(1)$ For a six month period than we would have $\theta = 2.3$. Calculator: [2nd] [DISTN] [poissonpdf]$(theta, x)$
Example: for a 6 month period where $\lambda = 4.6$ what is the prob of some number of deaths i.e. 0 or 1? poissonpdf(2.3,0) = 0.1 or poissonpdf(2.3,1) = .231.
Poisson Recursion Rule:
If p(X=x) is pois prob of observing k events w/ underlying parameter θ then
$p(X = x+1) = [\theta/(k+1)]p(X = x)$
Also note that the binom distn with large n and small p can be approx with a pois distn with paramter $\theta = np$.

Continuous Probability Distributions

Prob Density fn (pdf): A fn is the pdf, $f(x)$, of the continuous r.v., X, when (1) $f(x) \geq 0$ for all x. (2) the total area under the curve is 1.0. (3) the prob that X falls between a and b is equl the area under the curve btn a and b.
Gaussian Distn:
mean $= \mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$ and $Var(X) = E(X^2) - \mu^2$
Gaussian cdf: The cdf is $\phi(a) = p(X \leq a)$
Inverse Gaussian: $100(\times u)^{th} = Z_u$ This is defined by, $P(X < Z_u) = u$, where X is N(0,1)
Standard Normal: $N(\mu, \sigma^2) = N(0,1)$
Calculator:
[2nd][DISTN][normalcdf](lower,upper)
Examples: $P(-1 < X < 1) = 0.68$ – normalcdf(-1,1) = 0.68 $P(-2 < X < 2) = 0.95, P(0 < X < 1) = 0.34$
Rules: P(a ¡ X ¡ b) = P(X ¡ b) - P(X ¡ a), $\phi(-x) = P(X \leq -x) = P(X \geq x) = 1 - P(X \leq x) = 1 - \phi(x)$
More Examples: $P(X < 1) =$ normalcdf(-1E99,1) $= 0.84, P(X > 1) = 1 - P(X < 1)$
Inverse Examples: Suppose $u = 0.975$ Then $Z_{0.975} = 1.96$.
Calc: [2nd] [DISTN][invNorm](u)
i.e. $\phi(-1.96) = 1 - \phi(1.96) = 0.025$ and $Z_{0.025} = invNorm(0.025) = -1.96$ or $\phi(1.645) = 0.95$ and $Z_{0.95} = 1.645$
Standardization to N(0,1): If X is $N(\mu, \sigma^2)$ and $Z = (X - \mu)/\sigma$ then Z is N(0,1). or,

$$P(a < X < b) = P(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}) = \phi[\frac{b-\mu}{\sigma}] - \phi[\frac{a-\mu}{\sigma}]$$

Example: $P(90 < X < 95)$ when $\mu = 80$ and $\sigma^2 = 144$, normalcdf(90,95,80,12) = 0.097.
Normal Approx to binom Distn: If X is a binom r.v. with parameters n and p then $E(X) = np$ and $Var(X) = np(1-p)$. If n is large relative to p then it is well approximated. i.e. $np \geq 5$ and $n(1-p) \geq 5$
Example: $X \sim bin(n = 25, p = 0.4)$. What is $P(7 \leq X \leq 12)$? so $X \sim N(10, 6)$ Remember to round appropriatly!! so $P(7 \leq X \leq 12) = P(6.5 \leq X \leq 12.5)$?? normalcdf(6.5,12.5,10,$\sqrt{6}$) = 0.7698.
Normal Approx to Pois Distn:
This approx is used for $\mu \geq 10$. Example: Prob of obs x bacteria per area A. $\lambda = 0.1$ and $A = 100cm^2$. $\mu = \sigma^2 = \lambda A = 10$. And like for binom ajust the bounds and take sqrt if necessary.

Estimation

Populations are often too large to fully sample so we turn to the area of statistical inference, which consists of estimation and hypothesis testing. Estimation conscern predicting specific population parameters.
Sample Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Sample mean is an unbiased i.e $E\hat{\theta} = \theta$ estimator of μ. Another property of a good estimator is a small SD

Standard Error: The SD of the sampling distn is the standard error of the mean.
$SD(\bar{X}) = \sqrt{var(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{n}$
Note: variance is σ^2/n And as n becomes large SE becomes small.

Central Limit Thm: If $X_1, ...X_n$ is a random sample $\sim N(\mu, \sigma^2)$ then (1) $E(\bar{X}) = \mu$ (2) $Var(\bar{X} = \sigma^2/2)$ and if n is suff large then $\bar{X} \sim N(\mu, \sigma^2/n)$ so (it says) even if sampling distn if not normal if you take enough samples the sampling distn of the sampling mean will be normal.
Example: if we calc some prob like $P(98 \geq \bar{X} \geq 126)$ Give $\mu = 112$, SD = 20.6 and if sample 10 times what is the prob? $normalcdf(98, 126, 112, (20.6/\sqrt{10})) = .968$ This means that random samples of size 10 will fall btn 98 and 126 96.8 % of the time.

Confidence Interval - Known Variance:
In general % CI: $\left(\bar{X} \pm Z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right)$
For 95 % CI: $\left(\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}\right)$
Example: $\left(97.2 \pm 1.96\frac{0.2}{\sqrt{10}}\right) = (97.08, 97.32)$
We can say we are 95% confident that the true mean value for this pop lies between ... Basically if we darw the random sample, compute the interval and repeat a large number of times $100(1-\alpha)\%$ of the intervals constructed in this way will contain the true pop parameter μ.
Other common CI values. $Z_{0.90} = 1.645$, $Z_{0.95} = 1.96$ and $Z_{0.995} = 2.576$
Thing that affect the length of a CI. (1) sample size n. (2) standar dev σ and (3) CI coeff $(100\% \times (1-\alpha)$.

t-distribution: When the standard deviation is not known we cannot use the normal
if $X_1, ...X_n \sim N(\mu, \sigma^2)$ then $\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t(df = n-1)$ where s is the sample standard deviation. Example: $t_{20,0.95} = 1.725$.(Table 5 appendix)

Sample Standard Deviation:
$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
s^2 is an unbiased estimator of σ^2.
Confidence Interval - UnKnown Variance: sample variance, s^2 can be replaced for σ^2. A 100% $\times(1-\alpha)$ CI for mean μ of a normal distn w/ unknown var,
$\left(\bar{X} \pm t_{n-1,1-\alpha/2} \times \frac{s}{\sqrt{n}}\right)$
Example: Suppose an $n = 10$ yeilded a mean of 116.9 and a sample standard dev of 21.70. a 95% CI would be,
$\left(116.9 \pm 2.262 \times \frac{21.70}{\sqrt{10}}\right) = (101.38, 132.42)$
Chi-Square Distn: If $X_1...X_n \sim N(0,1)$ then $G = \sum_{i=1}^{n} x_i^2 \sim \chi^2(df = n)$
Confidence Interval - population variance:
$\left[\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}\right]$
Binomial Distn Point Estimation:
Let X be a binomal r.v. w/ parmaeter n and p. $\hat{p} = \frac{\sum_{i=1}^{n} x_n}{n}$ is an unbiased estimator or p. The standard error is given by $\sqrt{p(1-p)/n}$ and estimated by $\sqrt{\hat{p}(1-\hat{p})/n}$. The just said does not apply if we cannot apply the central limit thm i.e. (np(1-p)) has to be at least 5. Example if we have 10,000 women and 400 have breast cancer. the best estimate for prevalence is $\hat{p} = \frac{400}{10000} = 0.04$ using this result we can compute a CI...
Confidence interval for population prop:
$\left[\hat{p} \pm Z_{1-\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$
$\left[0.04 \pm 1.96 \times \sqrt{\frac{0.04(0.96)}{10000}}\right] = (0.036, 0.044)$

Testing Hypotheses - One Sample

Don't write hypotheses in terms of sample statistics. General protocol: (1) State Hypotheses, (2) State test to be used, (3) State Assumptions, (4) calculate test statistic.
Type I, II Error:
$\alpha =$ Type I error: Reject H_0 when H_0 true (false pos)
$\beta =$ Type II error: Fail to reject H_0 when H_0 false (f. neg)
$\alpha =$ significance level and $1 - \beta =$ reject H_0 when H_0 is false = Power. Each can be expressed in terms of probability too.
One-tailed test: A test in which the values of the parameter being studied under the alternative hypothesis, are allowed to be either greater than or less than the values of the parameter under the null hypothesis, but no both.
We don't every truly accept – only fail to reject
Sample Size: Factors affecting sample size: (1) as sample size increases σ^2 increases. (2) sample sizes increases as sig level is made smaller (3) sample size increases as required power increases (4) sample size de-

creases as $|\mu_0 0\mu_1|$ increases.
P-value: is the prob under the null hypothesis of obtaining a test stat as extreme as or more extreme than the observed test statistic, where in the case of a two-sided hypothesis extremeness is measured by the absolute value of the test statistic.
Z - Test:
Use: One Sample Test for Pop Mean - Var known
Hypothesis: $H_0 : \mu = \mu_0,\ \sigma = \sigma_0$
Test Stat: Z = $\frac{\bar{x}-\mu_0}{\sigma_0/\sqrt{n}}$
Test Distn: Z-distn
95% CI: $\bar{x} \pm \left(Z_{1-\alpha/2}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$
accept/reject: If $Z < -Z_{a/2} or if Z > Z_{1-a/2}$ then we reject H_0
p-value: If $Z < 0$ then $p = Z\phi(Z)$, if $Z > 0$ then $p = Z(1-\phi(Z))$
Power: Power = $\phi\left[Z_{\alpha/2} + \frac{\mu_0-\mu_1}{\sigma}\sqrt{n}\right]$
The power of a test tells us how likely we are to find a significant difference given H_1. Factors affecting power: (1) significance level (2) $|\mu_0 - \mu_1|$ (3) standard deviation (4) sample size.
Sample Size: $n = \frac{\sigma^2\left(z_{1-\beta}+Z_{1-\alpha/2}\right)^2}{(\mu_0-\mu_1)^2}$
Example1: $\mu_0 = 170, \sigma_0 = 10, \bar{x} = 165$ and $n = 25$.

$$p = p(\bar{x} < 165|H_0 \text{ is true})$$
$$= P\left(\frac{165-170}{10/\sqrt{25}}\right) = \phi(-2.5)$$
$$= 1 - \phi(-2.5) = 1 - \phi(2.5) = 1 - 0.994 = 0.006$$

$Z = -2.5$. normalcdf(-1E99,-2.5) = 0.006. Because p ¡ 0.05 we will reject at the 5% level of significnce. We would also reject at the 1% level. For this one-sided test the prob of observing a sample mean as small as 165 is 0.0062.
Now suppose the true mean and var is $\sim N(160, 4)$ because $\frac{\sigma}{\sqrt{25}} = 2$ We were to reject in case the $\bar{x} < 166.7$.

$$\beta = P(\bar{x} \geq 166.7|\mu = 160, \sigma - 10)$$
$$= P\left(\frac{\bar{x}-160}{2} \geq \frac{166.7-160}{2}\right)$$
$$= P(Z \geq 3.35) = 1 - P(Z < 3.35) = 0.0004$$
$$\text{power} = 1 - \beta = P(\bar{x} < 166.7|\mu = 160, \sigma = 10)$$
power(160 = 1 − β = 1 − 0.00041 = 0.99959.

i.e. solve to get 3.35. 1 - normalcdf(-1E00,3.35)) = 0.0004 so the power is 1 - that. If the true mean were 160 we would reject the null hypothesis 99.96% of the time.
Example2: $X \sim N(190, 1600)$ This is a two tailed example. $\bar{x} = 181.52$.

$$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} = \frac{181.52-190}{40/\sqrt{100}} = -2.12$$

For a 5% level test $Z_{0.975} = 1.96$ and $Z_{0.025} = -1.96$. So we reject H_0.
p-value = 2*normalcdf(-1E99,-2.12) = 0.034
Example3: When testing mean of normal distn. $\sigma = 10$. And we want to detect a difference as small as 5 units. Also we want to use a 5% level, two-sided test that guarantee's at least 80% power.

$$n = \frac{\sigma^2(Z_{1-\beta} + Z_{1-\alpha/2})^2}{(\mu_0 - \mu_1)^2}$$
$$= \frac{10^2(1.96 + 0.84)^2}{5^2} = 31.36$$

Thus we need a minimum of 32 subject.
One Sample t - Test Normal Distn:
Use: 1 Sample Test for Pop Mean - Var unknown
Hypothesis: $H_0 : \mu = \mu_0,\ \sigma = \sigma_0$
Test Stat: t = $\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$
Test Distn: t-distn
95% CI: $t_{n-1,1-\alpha/2} \pm \frac{s}{\sqrt{n}}$
accept/reject: If $t < t_{n-1,a/2}$ or if $t > t_{n-1,1-a/2}$ we reject. If $t_{n-1,a/2} \leq t \leq t_{n-1,1-a/2}$ then we accept H_0
p-value: 2× [area to left under t_{n-1} distn] if $t > 0$, or 2× [area to rt under t_{n-1} distn], if $t \leq 0$
Example1: n = 15. is the mean of x different than 120? $\bar{x} = 96$, $s = 35$.

$$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{96-120}{35/\sqrt{15}} = -2.66$$

We reject H_0 b/c $t < -2.1448$. If it is a two-sided test p = 0.0187 or if it were a one-sided test p = 0.0093. Either way we reject H_0. tcdf(-1E99,-2.66,14) = 0.0093
χ^2 - Test for Var of a normal:
Use: One Sample Test for var of normal distn
Hypothesis: $H_0 : \sigma^2 = \sigma_0^2$
Test Stat: $X^2 = \frac{(n-1)s^2}{\sigma_0^2}$
Test Distn:χ^2_{n-1}
95% CI ??: $\left(\bar{X} \pm t_{n-1,1-\alpha/2} \times \frac{s}{\sqrt{n}}\right)$
accept/reject: if $X^2 < \chi^2_{n-1,\alpha/2}$ or $X^2 > \chi^2_{n-1,1-\alpha/2}$ then reject H_0. If $\chi^2_{n-1,\alpha/2} \leq X^2 \leq \chi^2_{n-1,1-\alpha/2}$ then H_0 is accepted.
p-value: If $s^2 \leq \sigma_0^2$ then p-value = 2× (area left of X^2 under χ^2_{n-1}) otherwise it is area to the right
comments
Example1: to calc p-value if $s^2 = 8.178$ and $n = 10$ then $X^2 = \frac{9(8.178)}{35} = 2.103$. The critical regions are $\chi^2_{9,0.025} = 2.70$ and $\chi^2_{9,0.975} = 19.02$. Because $X^2 = 2.103$ is less than 2.70, H_0 is rejected using a 2-sided test at the $\alpha = 0.05$ level. p = $2 \times P(\chi^2_9 < \chi^2_9)$ calculator χ^2cdf(0,2.103,9) = 0.0102. So p = 0.0205.
One sample test for binom prop- normal theory: In survey of 300, 123 wear seat belts. Can we conclude the prop who wear seat belts is not 0.50? $p = p_0$ is the sample proportion. $\hat{p} \sim N(p_0, \frac{p_0(1-p_0)}{n})$. We can use the usual one-sample normal tests to test hypotheses about p because,

$$z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1)$$
$$= \frac{0.41-0.50}{\sqrt{\frac{(0.5)(0.5)}{300}}} = \frac{-0.09}{0.0289} = -3.11$$

Critical values are -1.96 and 1.96 so we reject H_0. if $\hat{p} < p_0$, p-value = $2\phi(z)$ elseif $\hat{p} > p_0$ p-value = $2(1-\phi(z))$ under $\sim N(0,1)$. Note: This procedure only works $np(1-p) \geq 5$

$$\text{Power} = \phi\left(\sqrt{\frac{p_0q_0}{p_1q_1}}\left[Z_{\alpha/2} + \frac{|p_0-p_1|\sqrt{n}}{\sqrt{p_0q_0}}\right]\right)$$
$$n = \frac{p_0q_0\left(Z_{1-\alpha/2} + Z_{1-\beta}\sqrt{\frac{p_1q_1}{p_0q_0}}\right)^2}{(p_1-p_2)^2}$$

One sample test for binom prop- exact theory: The two sided alternative. We know that $\hat{p} = \frac{x}{n}$. if $p \leq p_0$ then, p = $2 \times P(\leq x$ successes in n trials $|H_0$ is true). elseif $p > p_0$ then, p = $2 \times P(\geq x$ successes in n trials $|H_0$ is true) Example $H_0 : p = 0.20$ which is the national norm for cancer deaths. 5 deaths out of 13 are cancer related in some group. Note: npq = 2.1. Because $\hat{p} > p_0$ we use i.e. P(5 or more succesess). binomcdf(13,0.2,4) = 0.9009. 1 - ans = 0.0991. and because it is two sided: p-value = 2 × 0.0991 = 0.1983. Fail to reject.
One Sample t-Test:
Use: One Sample Test for dif in Pop Means
Hypothesis: $H_0 : \Delta = 0, H_1 : \Delta \neq 0$
Test Stat:
$t = \frac{\bar{d}-0}{s_d/\sqrt{n}}$, $\bar{d} = \sqrt{\frac{\sum d_i}{n}}$, $s_d = \sqrt{\frac{\sum(d_i-\bar{d})^2}{n-1}}$
Test Distn:t_{n-1}
95% CI: $\bar{d} \pm t_{n-1,1-\alpha/2} \times \frac{s_d}{\sqrt{n}}\Big)$
accept/reject: If $t > t_{n-1,1-\alpha/2}$ or $t < -t_{n-1,\alpha/2}$ then we reject H_0
p-value: if $t > t_{n-1,\alpha/2}$ or if $t < -t_{n-1,1-\alpha/2}$ reject H_0 otherwise if $t_{n-1,\alpha/2} \leq t \leq -t_{n-1,1-\alpha/2}$ accept H_1
Testing Hypotheses- 2 sample
Paired vs Independent Two samples are said to be paired when ea. data point of the first sample is matched with or is related to a unique data point of the second sample. They are independent when the data points in one sample are unrelated to those in the second sample. In a longitudinal study (paired) the same individuals are being measured twice. In a cross-sectional study (independent), two different groups of individuals are being compared.
Paired t-test:
Use: Two sample test for dif in pop means
Hypothesis: $H_0 : \Delta = 0$ vs $H_1 :\neq 0$
Test Stat: $t = \frac{\bar{d}}{s_d/\sqrt{n}}$ where $\bar{d}$ is the mean and s_d is the sample standard deviation of the observed differences and n is the num of matched pairs.
$s_d = \sqrt{\frac{\left[\sum_{i=1}^n d_i^2 - \left(\sum_{i=1}^n d_i\right)^2/n\right]}{n-1}}$
Test Distn: t_{n-1}
95% CI: $\bar{d} \pm t_{n-1,1-\alpha/2} \times \frac{s_d}{\sqrt{n}}\Big)$
accept/reject: if $t > t_{n-1,1-\alpha/2}$ or if $t < -t_{n-1,1-\alpha/2}$ reject H_0
p-value: If $t < 0$, $p = 2\times$ [the area to the left of t under a t_{n-1} distn] elseif $t \geq 0$, $p = 2\times$ [the area to the right]
Two sample t-test - equal variance:
Use: Two sample test for dif in pop means
Hypothesis: $H_0 : \mu_1 = \mu_2$
Test Stat:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$
$$s = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

Test Distn: $t_{n_1+n_2-2}$
95% CI:
$(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2,1-\alpha/2}\left(s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$
accept/reject: if $t > t_{n_1+n_2-2,1-\alpha/2}$ or if $t < -t_{n_1+n_2-2,1-\alpha/2}$ reject H_0
p-value: If $t \leq 0$, $p = 2\times$ [the area to the left of t under a $t_{n_1+n_2-2}$ distn] elseif $t > 0$, $p = 2\times$ [the area to the right]
Sample Sizes: two samples of equal size.

$$n = \frac{(\sigma_1^2 + \sigma_2^2)\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{\Delta^2}$$

where $\Delta = \mu_1 - \mu_2$
Two sample t-test - unequal variance:
Use: Two sample test for dif in pop means

Hypothesis: $H_0 : \mu_1 = \mu_2$
Test Stat:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

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

$$d' = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^2/n_1}{n_1 - 1} + \frac{s_2^2/n_2}{n_2 - 1}}$$

Test Distn:
$t_{n_1+n_2-2,1-\alpha/2}$ or $-t_{n_1+n_2-2,1-\alpha/2}$
95% CI:
$(\bar{x}_1 - \bar{x}_2) \pm t_{d'',1-\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
accept/reject: if $t > t_{d'',1-\alpha/2}$ or $t < -t_{d'',1-\alpha/2}$ reject H_0
p-value: if $t \leq 0$ then $p = 2\times$ (area to left of t under $t_{d''}$ distn) elseif area to right.
sample size: for two normally distd samples of unequal size

$$n_1 = \frac{(\sigma_1^2 + \sigma_2^2/k)(Z_{1-\alpha/2} + Z_{1-\beta})^2}{\Delta^2}$$

$$n_2 = \frac{k(\sigma_1^2 + \sigma_2^2)(Z_{1-\alpha/2} + Z_{1-\beta})^2}{\Delta^2}$$

Where $\Delta = |\mu_2 - \mu_1|; (\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2)$ are the means and variances of the two respective groups and $k = n_2/n_1 =$ the projected ratio of the two samples.
Comments: (1) Compute test stat,t. (1) Compute approximating df, d'. (3) round d' to the nearest integer d''.
F-test:
Use: Two sample test for dif in pop variation
Hypothesis: $\sigma_1 = \sigma_2$
Test Stat: $F = \frac{S_1^2}{S_2^2}$
Test Distn: F_{n_1-1,n_2-1}
95% CI: NEED **accept/reject**: if $F > F_{n_1-1,n_2-1,1-\alpha/2}$ or if $F < F_{n_1-1,n_2-1,\alpha/2}$ reject H_0
p-value: if $F \geq 1, p = 2\times p(F_{n_1-1,n_2-1} > F)$ elseif $F < 1, p = 2\times p(F_{n_1-1,n_2-1} \leq F)$
Example1: $F_{5,9,0.99} = 6.06$
Example2: $n_1 = 13, n_2 = 10$; $\bar{x}_1 = 63.538, \bar{x}_2 = 63.9$; $s_1 = 7.944, s_2 = 9.171$
CR: reject H_0 if $F > F_{9,12,0.975} = 3.44$
$F = \frac{(9.171)^2}{(7.944)^2} = 1.333$ Fail to reject H_0 b/c $F < 3.44$ The two variances are the same.
power for comparing means:

$$\text{Power} = \phi\left(-Z_{1-\alpha/2} + \frac{\Delta}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}\right)$$

Where $\Delta = |\mu_1 - \mu_2|$
Non-Parametric Methods
textcolorredSign-test: Set up as a paired test and the signs of differences are noted. Then and exact Binom test can be set up. Note: the magnitude of the difference is not used.
Wilcoxon Signed Rank test: This is similar to the 2 sample paired t-test. Rank absolute diffs (assign avg ranks too). Compute rank sum of positive diffs. Compute test stat T. Then depending on ties use appropriate formula.
Wilcoxon Rank-Sum Test: Analogous to the independent samples t-test. Combine data and rank it. Assign avg ranks. Compute rank sum. Compute Test stat where depending on ties use appropriate formula.
Categorical Data
Cardinal Data are on a scale where it is meaningful to measure distance btn possible data values. For these data if the zero point arbitrary the data are on an interval scale. If the zero point is fixed then the data are on a ratio scale.
Ordinal Data Can be ordered but do not have specific numeric values
Nominal Scale Data are on a scale such that different data values can be classified into categories but the categories have no specific ordering.
Normal Theory Method
Use: 2 sample comparison indep binomial prop
Hypothesis: $H_0 : p_1 = p_2$
Test Stat: $Z = \frac{|\hat{p}_1 - \hat{p}_2| - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$
where $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$ where $x_1 and x_2$ are the num of event in the 1st and 2nd samples. The 2nd term in the numerator of the test stat is the continuity correction.
Test Distn: $Z \sim N(0, 1)$
95% CI:
$(\hat{p}_1 - \hat{p}_2) \pm Z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2}{n_2}}$
accept/reject: N(0,1)
p-value: N(0,1)
Power: ????
Sample Size: $n_1, n_2 = \ldots$

$$= \frac{\left[\sqrt{\bar{p}\bar{q}(1 + \frac{1}{k})}Z_{1-\alpha/2} + \sqrt{p_1 q_1 + \frac{p_2 q_2}{k}}Z_{1-\beta}\right]^2}{\Delta^2}$$

$$= kn_1$$

where p_1 and p_2 are projected true probabilities of success. $\Delta = |p_1 - p_2|$, $\bar{p} = \frac{p_1 + kp_2}{1+k}$
Comments: must have $n_1 p_1 q_1 \geq 5$ and $n_2 p_2 q_2 \geq 5$
Contingency Table Method
If we test independence of row variable and column variable it is a test of association or test of independence. If we test prop of column variables between row variables it is a test of homogeneity of binomial proportions.
Hypothesis: $H_0 : p_1 = p_2$
Test Stat: $X^2 = \sum^R \sum^C \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
where E_{ij} is (row total)(col total) / (total total). And the sum is over all cells in the table. **Test Distn**: $X^2 \sim \chi_1^2$
The Yates corr simply subtracts 0.5 for ea sum from $|O - E|$. An easier form is,
$X^2 = \frac{n(|ad-bc|-n/2)^2}{(a+b)(c+d)(a+c)(b+d)}$
R x C Contingency Table Method
Hypothesis: $H_0 : p_{ij} = p_i p_j$ for all i and j H_1 : one is dif
Test Stat: $X^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ where $E_{ij} = n_{i.} n_{.j}/n_{..}$
Test Distn: $X^2_{(R-1)(C-1),1-\alpha}$
accept/reject: reject null if $X^2 \geq \chi^2_{(R-1)(C-1),1-\alpha}$
p-value: $p = P(\chi^2_{(R-1)(C-1)} > X^2$
Use the test only if no more than 20% of cells have exp values < 5 and no cell < 1
Fisher's Exact Test
If one of the exp values is less than 5 we should use this test. Basically you ennumerate all possible tables, compute exact probs of ea., the calc p-value based on cummulative prob of all tables.
McNemar's Test for Correlated props (Normal Theory Test)
Use: Comparison of 2 paired binom prop
Hypothesis: $H_0 : p_1 = \frac{1}{2}$
Test Stat: NEED
Test Distn: NEED
95% CI: NEED
accept/reject: NEED
p-value ??: NEED
Power: NEED
Sample Size: NEED
McNemar's Test for Correlated proportions (Exact Method)
Use: Comparison of 2 paired binom prop
Hypothesis: $H_0 : p_1 = \frac{1}{2}$
Test Stat: NEED
Test Distn: NEED
95% CI: NEED
accept/reject: NEED
p-value ??: NEED
Power: NEED
Sample Size: NEED
In an [a][b] / [c] [d] table we can say that b and c are the discordant pairs. Concordant Pair Is a matched pair in which the outcome is the same for each member of the pair
Discordant Pair Is a matched pair in which the outcomes are different for the two members of the pair
Type A discordant pair treatment A member of the pair has the event and treatment B does not
Type B discordant pair treatment B member of the pair has the event and treatment A does not
Kappa Statistic: used to measure reproducibility between surveys
K > 0.75 : denotes excellent reproducibility
$0.4 \leq K \leq 0.75$: denotes good reproducibility
$O \leq K < 0.4$: denotes marginal reproducibility
Regression and Correlation