Midterm 1

1.1:
Mutually Exclusive: no overlap; P(AnB)=0;
Exhaustive: complete coverage

2.1: Random Variables: Sample space S, function X: S -> R associates a number X(s) to each outcome
pmf: gives the probability for each value of x

1.2: Sample counting
Ordered: With replacement: n^r; Without replacement (order important): nPr = n!/[(n–r)!]; **Unordered**: nCr = n!/((n-r)!r!)
Distinct permutations of n things, k kinds: n!/(n1!n2!...nk!) 1.3 BELOW

1.4:
P(A|B)=P(AnB)/P(B):**conditional** prob that A willoccur given that B occurs
P(A|B) = N(AnB)/N(B) equally likely model, propo. of outcomes from B that satisfy criteria for A
P(AnB)=P(**B|A)**P(A)=P(A|B)P(B)
P(B) = P(AnB) + P(A'nB) = P(B|A)P(A)+ P(B|A')P(A')
P(A'nB) = 1 – P(A|B); P**(B|A)** = P[(A|B)P(B)]/P(A)
If E_1, ... E_k are M.E. and exhaustive, for any event B and any i: (Bayes')
$P(E_i|B) = P(B|E_i)P(E_i) / [P(B|E_1)P(E_1) + ... + P(B|E_k)P(E_k)]$
independent if: P(B|A) = P(B) or P(AnBnCnD) = P(A)P(B)P(C)P(D)

2.2 Discrete:
pmf: f(x)=P(X=x); cdf:F(x)=P(X<=x)=$\Sigma_{y<=x}f(y)$
μ=E(X)=$\Sigma_x xf(x)$=center of distributon mass=mean of X's probability distribution; To get, SUM[E(X)*p(x)]; (If it's a function E(g(X)), multiply that function's value at x by p(x))
σ^2=$E[(X-\mu)^2]=\Sigma_x(x-\mu)^2f(x)=E[X^2]-(E[X])^2=\Sigma_x(x^2)f(x)-(\mu^2)$; to get, SUM[$(x-\mu)^2$ * p(x)]; prob multiplied by (distance from mean)^2
Prob. dist.: a table with possible x values and their p(x) values

2.3 Linear Transformation
Y = a+bX is the format; ★ E(a+bX)=a+bE(X);
★ $Var(a+bX)=b^2Var(X)$; ★ SD(a+bX)=|b|SD(X)

2.4 Continuous:
pmf becomes **pdf**=$f_X(x)=d/dx\ [F_X(x)]$;
CDF of any r.v. X: $F_X(x)=P(X<=x)$
$P(a<X<=b)=F_X(b)-F_X(a)$
If we take a linear function a+bX of a c.r.v, formulas ★ are the same.
Continuous Random Variable (M2)
pdf f(x): height of the curve at x, probabilities are areas under the curve. Total area=1. Integrate with an x in a bound logically to get cdf, use to calculate probs on intervals. Prob at one exact point=0 (area), so must be an interval a to b on cdf.
μ: integrate with extra *x on [a,b]: $_{inf}S^{inf}[xf_X(x)]dx$;
σ^2: integrate with extra*x^2 on [a,b], then subtract μ ^2: $E(X^2) - [E(X)]^2$
Properties of pdf f(x):
Single values have no mass, if a<b then:
P(a<X<b)=P(a<=X<b)=P(a<=X<=b)
$=P(a<X<=b)=_aS^b(f(x)dx)=F(b)-F(a)$ <- CDF

3.1 Binomial Dist.: [X ~ B(n,p)]
x = number of out of n indie bernie (pass/fail) trials, x has Binomial dist w/ replacement (indie); **Pmf:** $(nCx)(p^x)(1-p)^{n-x}$ for x=0,1,2..,n
Cdf: $P(X<=x) = SUM_{i\ to\ x}((_nC_i)(p^i)(1-p)^{n-i})$; **μ** = np; **$\sigma^2$** = np(1-p)
Bernoulli: Special case with n=1, x=0, 1;

3.2 Geometric Dist.:
x: #trials needed to get 1st success out of many (indie) bernie trials
Pmf: $(1-p)^{x-1}$ * p for x=1,2,3... (p is success and 1-p is fail)
x=1 is always the most likely, so $P(X>x) = (1-p)^x$ and...
Cdf: $P(X<=x) = 1 - (1-p)^x$; **μ** = 1/p; **σ^2** = $(1-p)/(p^2)$

3.3 Negative Binomial Dist.:
x: number of indie bern (pass/fail) trials needed to get r^{th} success; P(success)=p; P(fail)=1-p; x=trial# of the r^{th} success.
Pmf: $(_{x-1}C_{r-1})(p^r)(1-p)^{x-r}$ for x=r,r+1,...; **μ** = r/p; **σ^2**= [r(1-p)]/p

2.5 Discrete Joint Distribution:
X and Y are both d.r.v.'s – their joint pmf: $f_{XY}(x,y)$ = P(X=x,Y=y) = P(X=x and Y=y);
Range of random vector (X,Y): R_{XY}={(x,y): $f_{XY}(x,y)$!= 0};
P((X,Y) element of A) = $\Sigma_{(x,y)\ element\ of\ A\ n\ R(sub\ xy)}\ f_{XY}(x,y)$
Marginal pmf of X (same with Y):
$f_X(x)=P(X=x)=\Sigma_y f_{XY}(x,y)$
Independant if $f_{XY}(x,y) = f_x(x)f_Y(y)$ for all x and y

3.4 Poisson Dist.: (# warhorses killed in 30 days)
x: #occurences of an indie event in a period; rate is constant.
Pmf: $(\lambda^x * e^{-\lambda})/(x!)$, for x=0,1,2...; $\mu=\lambda=\sigma^2$;
λ= (events/second) * (#seconds in period)
Poisson Approximation to the Binomial Dist.:
large n and small p; μ=np;
Pdf = $(nCx)(p^x)(1-p)^{n-x} \approx e^{-\mu}(\mu^x/(x!))$
good approx if n>=20 & p<=0.05.
Poisson Process of λ if: # changes in separate regions are indie; P(1 change in int. length h) ≈ λh; prob of 2 or more changes in a short interval is ≈ 0. Counting changes in a fixed length interval: X is # changes in interval of length t in a poisson process of rate α. X has a poisson distribution w/ param μ, μ = λt. If a poi proc. exists w/ rate λ, X is the length of interval needed to see a change. The cdf of X: $F(x) = 1-e^{-\lambda x}$ where x>=0 (see exponential distribution)

2.6 Covariance and Correlation
X and Y both r.v.'s; expectation of h(X,Y) called E|h(X,Y)| is the ave. expected value w/ repeated trials.
E[aX + bY] = aE[X] + bE[Y]; V[X + Y] = V[X] + V[Y] + 2E[(X - μ_X)(Y – μ_Y)]
Covariance between X and Y: σ_{XY} = E[(X – μ_X)(Y – μ_Y)] = E[XY] – $\mu_X\mu_Y$
Corr. coeff. is between X and Y: $\rho_{XY} = \sigma_{XY}/(\sigma_X\sigma_Y)$; -1<=ρ<=1; if points on (X,Y) are a line then ρ_{XY}= either 1 or -1 (slope +/-); if indie then $\sigma_{XY} = \rho_{XY} = 0$ (reverse not true); if !=0 then correlated and indie; if =0 then uncorrelated but ? Indie.

Erlang Distribution: X~Erlang(r, n)
x: interval length for r events with rate λ.
cdf x>0: P(X<=x)=1-P(X>x)=1-P("at most r-1 changes in [0,x]")= $1-\ _{k=0}\Sigma^{r-1}\ [(e^{-\lambda x} * (\lambda x)^k)/(k!)]$; **μ**=r/λ; **$\sigma^2$**=$r/(\lambda^2)$

parameter	estimator
μ	$\bar{X} = \sum_{i=1}^{n} X_i/n$
σ^2	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
σ	$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$
p	$\hat{p} = \#\text{of successes}/n$

Midterm 2

Uniform Distribution
pdf = 1/(b-a) on [b,a]; find probs w/ rectangle areas. Mean: **μ**=E[X]=(a+b)/2; Var: **σ²**=V[X]=[(b-a)²]/12
cdf: (x-a)/(b-a) w/ a<=x<=b.

Exponential Distribution: X~exp(λ)
x: interval length to see change w/ rate λ; **pdf**= $\lambda e^{-\lambda x}$, **cdf**=$1-e^{-\lambda x}$. **μ** = 1/λ; **σ²**=$1/(\lambda^2)$

Normal Distribution: X~N(μ, σ2)

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

Cdf: Φ(z)=P(Z<=z)=$\int_{-\infty}^{z} \phi(t)\,dt = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$

Standard Normal: Z~N(0,1)

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

phi(z) = prob that a standard normal var assumes a value on [0, z].

Convert to std. norm: $Z = \frac{X - E[X]}{\sqrt{V(X)}} = \frac{X-\mu}{\sigma}$

$$F_X(x) = P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = P\left(Z \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

$$P(a \le X \le b) = F_X(b) - F_X(a) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Using the table:
What you get from finding a value: The probability to the left of it.
To get P(a<Z<b) you do z(b)-z(a); To get P(Z>a), do P(Z<(1-a)).
P(a<Z<b) = 0.95, 1=0.95 = 0.05, 0.05/2 =0.0025, find in table!
"n^{th} percentile" → find 0.n in body of table
Ex: P(X<180.0)? = P((x-mu)/sig < (180-mu)/sig) = P(Z<dec.ans.), find z(dec.ans.) = answer.

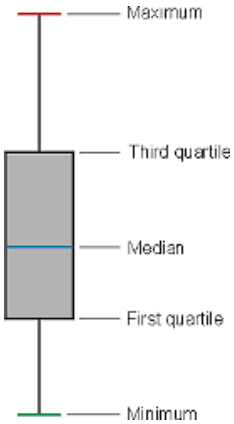
Sampling distribution and data descriptions
Characteristics of the pop. are known as variables like X,Y, etc. We take n sample objects. Thus $x_1,x_2,..x_n$ are n indie observations of X. A function of the random sample $X_1,X_2,...X_n$ is a **statistic**. We use the stat ^Θ=h($X_1,X_2,....X_n$) to estimate pop param θ. ^Θ is a point estimator of θ, is a rv, and its observed value of the rv ^θ =h($x_1,x_2,..x_n$) is called a point estimate of θ.
Sampling dist of Xbar is the dist of Xbar in all possible samples of a certain size from the population. Sampling dist of sample mean: r.v. Xbar is the mean of n indie observations of a pop with mean μ, sd σ
Mean of the sample mean's sampling distribution is equal to the population mean (μ_{Xbar} = μ). σ_{Xbar} = σ/root(n)

Sampling Distributions: Central Limit Theorem
Sample mean will be ~= normally distributed for large sample sizes (n>30). Xbar: sample mean μ of n indie observations. Sample mean is = population mean. Standev of sampling distribution of Xbar: **σ_{Xbar}=σ/root(n)**

Measures of Dispersion/Variability
The sample Standard Deviation is s and s^2 is the sample variance: $s = \sqrt{\frac{(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2/n}{n-1}}$
The Sample range, r = max(x_i)-min(x_i) = $y_n - y_1$; Quartile: Q_1 = 25th percentile. IQR = $Q_3 - Q_1$, measures dispersion.

Expectations and Variances of a Linear Function of the Random Sample
consider rv's $X_1,..X_n$ and constants $c_0,..c_n$.Y=$c_0+c_1X_1+...+c_nX_n$;
E[Y]=$c_0+c_1E[X_1]+...+cnE[Xn]$. Assume all X indie, Var $\sigma_Y = c_1^2\sigma_{X1}^2 +...+c_n^2\sigma_{Xn}^2$. Let X be rv, a&b constant; E[aX+b] = aE[X]+b; V[aX+b]=$a^2\sigma_X^2$. Sample mean has μ_{Xbar}=μ mean and $\sigma_{Xbar}^2 = \sigma^2/n$

Boxplot:



Measures of center
Sample mean: avg of the n values, denoted by **xbar**. Median: 50th percentile, position is (n+1)*(1/2); to get it integrate pdf (f)x from bottom to m, must=1/2, solve for m. Percentile: find %th percentile of f(x) by finding x at P(x)=0.%. Position of %th percentile is (n+1)*(k/100)=m(whole)+p(fractional). The kth percentile is y_m if p=0, and is equal to $y_m + p(y_{m+1} - y_m)$ if p is fractional.

Independent samples
2 indie random samples of populations X_1, . . . , X_{n1}, Y_1, . . . , Y_{n2} each with mean μ and variance σ². If they're both normal, N(μ_1, σ_1^2) and N(μ_2, σ_2^2), then also

$$\overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

If we assume n_1 and n_2 are both >=30, we can show that Z has approx. a standard normal dist.:

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

Estimators and point estimates:
^p=sample proportion=estimate of pop. proportion p; sometimes sampling distros are ~= normal; thus their conf intervals are in form Point estimate +- margin of err. We use estimator, point estimate, sample mean(Xbar), sample variance(S^2), sample standard deviation (S), and sample proportion ^P = X/n (X is # items satisfying attribute among n items).
Let ^θ be an estimator for θ. ^θ is an unbiased estimator if E[^θ] = θ. A binomial experiment with ^P = X/n be the sample proportion from a pop with P(success) = p (population proportion), X is successes among n trials. E[^P] = p; $\sigma^2_{\hat{}P}$ = [p(1-p)]/n. Standev of a point estimator is the standard error. serr of the mean and sample proportion are: σ_{Xbar} = σ/root(n) = s/root(n), $\sigma_{\hat{}p}$=root([p(1-p) / n]=root[(^p(1-^p)) / n]

Confidence Interval for the mean
Conf. int.: within +- errmargin, represented by a % between (bla,bla). n indie obs, normal dist, mean μ=σ=σ_{Xbar}=σ/root(n). Std.Normal Z: P(-$z_{alpha/2}$ < Z < $z_{alpha/2}$) = 1-alpha. Conf int for μ: X_{bar}+- $z_{alpha/2}$*(σ/root(n)), where alpha is 1-(%confidence), ie 95% con-> alpha is 0.05; the +- is the margin of error.

Parameter	Estimator	Estimate
μ	$\bar{X}$	$\bar{x}$
σ^2	S^2	s^2

Midterm 2+

Confidence Interval for the proportion

X_{bar} +- [$t_{alpha/2}$ * (s/root(n))]; n>40 is good, 15 to 40 is less so, <15 is a no go.

If we have a sample of size n from the population, and we denote by Y the total number of individuals (in the sample) who possess the desired characteristic, then a *point estimator* for p is: $\hat{P} = Y/n$.

If n is large, ^p is distributed normally with mean: p and var: p*(1-p)/n.

$$Z = \frac{\hat{P} - p}{\sqrt{p(1-p)/n}} \text{ is approximately a standard normal random variable}$$

For large n, 100(1-alpha)% conf int for p: $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Thus, if $\hat{p}$ is an estimate for p, we are at least $100(1-\alpha)\%$ confident that the error $|\hat{p} - p|$ will not be greater than E with a sample size

$$n \geq \frac{1}{4}\left(\frac{z_{\alpha/2}}{E}\right)^2$$

Hypothesis Testing:

Left-sided(H_1: mu < mu_0), right-(H_1: mu > mu_0), two-(H_1: mu != mu_0);

Type 1 Error: Rejecting H_0 when it is actually true;

Type 2 Error: Not rejecting H_0 when it is false.

Type 1: P(Type I error | H_0 is true) = alpha = significance level; Type 2: Beta. **Power**= 1-P(Type2error)=1–Beta

Hypothesis Testing, σ known:

Z = (orange); If H_0 (null hyp.) is true, you should get a random sample from the standard normal.

Rejection regions:

alpha=sig.level, probability of rejecting null hyp. if H_0 is true); Find rejection region, reject H_0 if the test falls in area.

For H_1: mu != mu_0, alpha =0.05:

e.g. H_0: mu=mu_0, H_1: mu!=mu_0; reject outside center region (95% conf. interval area, i.e. areas outside of z(alpha/2), the "critical value"). So if Z value is >= 1.96 or <= than -1.96, reject H_0.

For H_1: mu < mu_0, alpha =0.05:

Reject null hypothesis if Z <= -1.645

For H_1: mu > mu_0, alpha =0.05:

Reject null hypothesis if Z >= 1.645

P-value method:

P-value: probability of getting the observed value or a value w/better evidence against H_0, if H_0 is true; **i.e.** it is a measure of how strong the evidence against the null hypothesis. is.

Smaller p-value → greater evidence against H_0.

If p-val < alpha, the evidence against H_0 is significant at the alpha level of significance, so **we reject H_0.**

ex: P-value = P(H_1) **zx** = (Xbar – mu_0)/(sig/root(n)), **P[**Z <= zx] = ztable(zx)

i.e. **if H_1: mu < mu_0**, pval=P[Z <= zx] (finding area to left of crit. value)

and **if H_1: mu > mu_0**, pval= 1- P[Z <= zx] (finding area to right of crit. value)

Confidence Interval for the mean μ; σ known:

Zalpha/2: Find alpha, divide by 2, use logic, find the z of that% in table. (body of table is the %) "A sample of n people gave an average value of gh" means that gh = Xbar. For a "%" conf int, % of the "%" conf ints calculated would capture the true value.

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

$$1 - \alpha = P(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = P(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

a $100(1-\alpha)\%$ confidence interval for μ is

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = [\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

If $\bar{X}$ is the size-n-sample mean from a normal population with known variance σ^2, the *exact* $(1-\alpha)$ C.I. for μ:

$$\bar{X} \pm z_{\alpha/2}\sigma/\sqrt{n};$$

Precision: $\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} - (\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 2\frac{z_{\alpha/2}\sigma}{\sqrt{n}}$

Sample Size: If $\bar{x}$ is used as a point estimate for μ, then we are $100(1-\alpha)\%$ confident that the error $|\bar{x} - \mu|$ will not be greater than E, if the sample size satisfies

(where E is Error estimation) $n \geq \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$

Confidence Interval for the mean μ; σ unknown:

Conf interval: Xbar +- [$t_{alpha/2}$ * (s/root(n))]; where s = StandErr(Xbar) = s/root(n); t on table with n-1 degrees of freedom v.

if σ is not known, then

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

where t_{n-1} is a Student's t-distribution with $n-1$ degrees of freedom. Consequently,

$$P\left(-t_{\alpha/2,n-1} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2,n-1}\right) = 1 - \alpha,$$

where $t_{\alpha/2,n-1}$ is the $100(1-\alpha/2)$th percentile of the Student distribution with $n-1$ degrees of freedom. **Student's *t*-distribution** It is perhaps not

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

T has Student's t-dist with n-1 degrees of freedom df.

a $100(1-\alpha)\%$ confidence interval for μ is

$$\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right) \quad (1)$$

where t is the value that we read in Table at level $\nu = n-1$, such that $P(-t \leq T \leq t) = 1-\alpha$, i.e $P(T \leq t) = 1 - \alpha/2$.

If $\bar{X}$ is the size-n-sample mean from a normal population with unknown variance, the *exact* $(1-\alpha)$ C.I. for μ:

$$\bar{X} \pm t_{\alpha/2,n-1}S/\sqrt{n};$$

If $\bar{X}$ is the size-n-sample mean from a normal or non-normal population withunknown variance and n is 'big', the approximate $(1-\alpha)$ C.I. is:

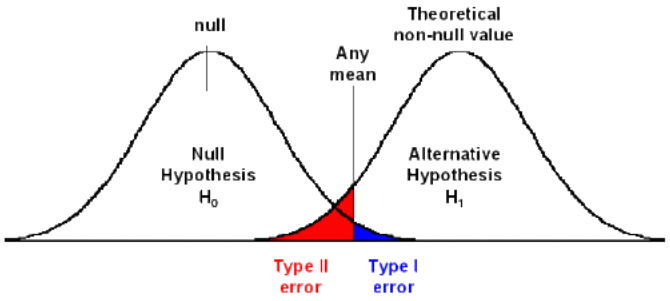
$$\bar{X} \pm z_{\alpha/2}S/\sqrt{n};$$

←The confidence interval is simply the values of mu_0 for which we do not reject the null hypothesis.

1.3 Subjective: 0<=P(E)<=1; Equally likely: P(E)=N(E)/N(S)= #good/#total; Axioms: Positivity: P(E) >= 0; Certainty: P(S) = 1; Additivity: add up prob. of mut.excl. events; P(E') = 1 – P(E); P(A) = P(AnB) + P(AnB')

P(AuB) = P(A) + P(B) – P(AnB)

P(AuBuC)=P(A)+P(B)+P(C)-P(AnB)-P(AnC)-P(BnC)+P(AnBnC)



Parameter name	Population parameter symbol	Sample statistic
Number of cases	N	n
Mean	μ (mu)	$\bar{x}$ (Sample mean)
Proportion	π (Pi)	P (Sample proportion)
Variance	σ^2 (Sigma-square)	s^2 (Sample variance)
Standard deviation	σ (Sigma)	s (sample standard deviation)
Correlation	ρ (rho)	r (Sample correlation)
Regression Coefficient	β (beta)	b (sample regression coefficient)

Regression:

Predictor: X (independent); Response: Y (dependent). Line of best fit = regression line.

Notation of fitted/estimated line: E(X)=^y = ^alpha + (^Beta)*x = yintercept + ystretch*x; Fitted value of the i^{th} observation: $\hat{y}_i$ = ^alpha + (^Beta)*x_i; Difference between fitted value/observed response y_i is called the i^{th} residual observed error: $e_i = y_i - \hat{y}_i = y_i - (\hat{\alpha} + \hat{\beta}x_i)$ = observed –predicted.

Residual sum of squares: $SSE = \sum_{i=1}^{n}\left[y_i - (\hat{\alpha} + \hat{\beta}x_i)\right]^2$; Total sum of squares: $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$; $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$

Regression sum of squares: $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$; Error sum of squares: $SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2$; ^Beta = s_{xy}/s_x^2

Point $(x, \hat{\alpha} + \hat{\beta}x)$ is always on the regression line. $y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$ $\text{total deviation} = \text{explained deviation} + \text{unexplained deviation}$

r: correlation between y and x, r^2: coeff. of determination, r between +- 1 with exactly positive / and negative \ slope respectively being ~= +-1. (r_{xy} close to zero means no correlation)

$R^2 = r^2 = \frac{SSR}{SST}$ = $s_{xy}^2/(s_x^2 s_y^2)$; $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$; $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$; $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$; $r = \frac{s_{xy}}{s_x s_y}$

standerr(^Beta) = sqrt(^sig^2/s_x^2); ^sig^2=SSE/(n-2)

Other distributions:

<u>Hypergeometric:</u> without replacement -> not independant.

If you are sampling <5% of pop, you can approximate this using binomial.

Example calculation:

12D, 24R, 8I. Pick 6 out of the total (42). What is P(3D 2R 1I)?

[(12C3) * (24C2) * (8C1)]/(44C6) = 0.0688

<u>Multinomial:</u> $\mathbf{x_i}$ = number of occurrences of outcome i out of n indie trials, k m.e. outcomes

For each trial, these k outcomes occur with probabilities p1,..pk, SUM(p1,..pk)=1

Pmf: P(X1=x1,..,Xk=xk) = [n!/(x1!*...*xk!)]($p1^{x1}$ * ... * pk^{xk}) for xi=0,1,..,n , (x1 != 1)

E(Xi) = n*p_i ; Var(Xi) = n*p_i*(1-p_i); SUM(xi) = n