Probability Density functions

Probability DENSITY Functions: Pdf : works on interval
Probability MASS Function: Pmf : Works on discrete variables – you can assign probabilities to each value of x and sum the probabilities to equal 1

PROPERTIES of Probability Density function:

- For ALL x, F(x) >= 0
- The area under f(x) and x=(a,b) , the x-axis gives P(a ≤ x ≤ b) = P(a ≤ x < b) = P(a < x ≤ b), because P(x=integer) = 0

If x is a CONTINUOUS RANDOM VARIABLE, the probability that x is equal to any SINGLE number is 0
P(x=a)=0 (i.e. P(x=5) = 0)

UNIFORM RANDOM VARIABLE: x has UNIFORM(a,b)

F(x) = $\frac{1}{b-a}$ if a ≤ x ≤ b, and 0 otherwise

If X has **UNIFORM[0,2]**
P(c <= x <= d) = (d-c)/b-a therefore P(1 <= x <= 1.5) = (1.5-1)/(2-0)

EXPONENTIAL RANDOM VARIABLE/DISTRIBUTION
Memoryless – (forget about the past)
Related to Poisson Random variable
Exp(μ) or Exponential(μ)

We say that x has an exponential distribution with mean **μ**, IF it has $f(x) = \begin{cases} \frac{1}{\mu} e^{-\frac{x}{\mu}} & if \ x \geq 0 \\ 0 \ Otherwise \end{cases}$

$P(x > a) = e^{-\frac{a}{\mu}}$

$P(a <= x <= b) = e^{-\frac{a}{\mu}} - e^{-\frac{b}{\mu}}$

Eg. The lifetime of a particular brand of LED TV is EXPONENTIAL with mean μ= 4 years
If you know that a certain LED TV of the same brand, has been working for at least 3 years, what is the probability that the TV's lifetime will be more than 8 years?
X= Lifetime of an LED TV

$P(X \geqq 8 \mid X \geqq 3) = \frac{p(x \geqq 8 \cap x \geqq 3)}{P(x \geq 3)} = \frac{p(x \geqq 8)}{P(x \geq 3)} = \frac{e^{-\frac{8}{4}}}{e^{-\frac{3}{4}}} = e^{-\frac{5}{4}} = P(x \geq 5)$

The intersection of two sets where one is inside the other, is always the probability of the smaller set

Probability DENSITY FUNCTION (PDF)
Where X has a NORMAL DISTRIBUTION N(μ,σ^2)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$ **; -infinity < x < + infinity**

STANDARD NORMAL DISTRIBUTION: Z has standard normal (normal zero, one) if PDF N(0,1) - **Expected Value (μ) = 0, and variance =1**

$$f(z) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{(Z)^2}{2}}$$

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}$$ **, you can determine that σ^2 = 1, and μ = 0 -- STANDARD NORMAL DISTRIBUTION. X has N(0,1)**

CUMULATIVE DISTRIBUTION FUNCTION
P(X≤k)
x has N(μ,σ^2)
P(X ≤ k) eg. P(X ≤ 1.5)

P (z <= 1.46) = 0.9279
P (z > 1.52) = 1- P(z <= 1.52) = 1 – 0.9357 =
P (1.55 < Z < 1.68) = P(z < 1.68) – P(z < 1.55) - we are dealing with a continuous random variable – it does not matter whether or not it is < or ≤ - the result is the same.

What if x has N(μ,σ^2)? P(x ≤ a) = P(z ≤ (a-μ)/σ)
TRANSFORM it to STANDARD NORMAL and then you can use the table

Can only use STANDARD NORMAL APPROXIMATION for BINOMIAL PROPORTIONS if the following TWO CONDITIONS ARE TRUE

1) **If np > 5 and nq > 5 (where q = 1-p)**
2) **p should not be close to 0 or 1 – best is when p is close to .5**

If so, you can approximate P(x ≤ k) ≈ $P\left(z \leq \frac{k+0.5-np}{\sqrt{npq}}\right)$
Adding .5 of a unit is the CONTINUITY CORRECTION. It is absolutely vital. This is because X is a DISCRETE random variable and Z is a CONTINUOUS random variable. Adding .5 is necessary to adjust for this discrepancy.

PROPERTIES OF THE SAMPLE MEAN $\bar{x}$:

1) $\bar{x}$ is an UNBIASED ESTIMATOR for μ; E($\bar{x}$) = μ (regardless of sample size)
2) The standard deviation of $\bar{x}$ is $\frac{\sigma}{\sqrt{n}}$ where n is the sample size
3) The CENTRAL LIMIT THEOREM APPLIES: if the sample size n is LARGE, then the sampling distribution of $\bar{x}$ is approximately NORMAL with mean of μ and stddev of $\frac{\sigma}{\sqrt{n}}$

Sufficiently large samples of a RANDOM distribution converge to a NORMAL distribution.

if n >= 30 then $\bar{x}$ is APPROXIMATELY $\mathbf{N(\mu, \frac{\sigma^2}{n})}$

If $\bar{x}$ is APPROXIMATELY Normal, THEN $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is APPROXIMATELY STANDARD NORMAL, N(0,1)

$\hat{p}$ (sample proportion) is approximately $\mathbf{N(P, (P(1-p))/n)}$

IF $\mathbf{np > 5}$ and $\mathbf{n(p-1) > 5}$ then $z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$

Rules for determining if SAMPLE SIZE n is LARGE enough to use the CENTRAL LIMIT THEOREM:

1) If the original population is normally distributed, then I can use the CLT for any size of *n*, and the approximation becomes EXACT.
2) If the population distribution is fairly symmetrical, then for a relatively small sample size, you can use the CLT
3) If we the population distribution is skewed or unknown, we must use a sample size of at least 30 to use CLT.

Inferential statistics:

Methods of inference

1) **Estimation** – estimating or predicting the value of the unknown parameter. Can be POINT or INTERVAL estimation.
2) **Hypothesis testing** – process of making a decision about a parameter based on some preconceived idea about it.
   Read examples 8.1 and 8.2 in the book

Types of estimators:

An ESTIMATOR is a FORMULA that tells us how to calculate an estimation for an unknown parameter

1) POINT ESTIMATOR – Based on a sample, one single value is computed as an estimator for an unknown parameter
2) INTERVAL ESTIMATOR – based on a sample, TWO values are computed/calculated to form an interval that will contain the unknown parameter with high probability

Unbiased ESTIMATOR for the parameter of interest (θ)

We say the estimator $T(x_1,\ldots,x_n)$ is UNBIASED for θ if the EXPECTED VALUE, $E(T(x_1,\ldots,x_n)) = \theta$

Choose estimators which satisfy the following conditions:

1) **Pick the ones who are UNBIASED - the sampling distribution is CENTERED over the parameter of interest**
2) **Choose (among the UNBIASED) estimators that have the SMALLEST Variance (or SD)**

Proportion is $\hat{p} = \frac{\sum_{i=1}^{n} y_i}{n}$ **error is** $|\hat{p} - p|$

Error of ANY estimation is $|\hat{\theta} - \theta|$

MARGIN OF ERROR (will use 95% margin)

For POPULATION MEAN, $\mu = 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$ **but we may not have population STDDEV, so use sample stddev** $= 1.96\left(\frac{s}{\sqrt{n}}\right)$

For POPULATION PROPORTION, P $= 1.96\sqrt{\frac{p(1-p)}{n}}$ **- but use sample** $= 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

IF $n\hat{p} > 5 \;\&\; n(1-\hat{p}) > 5$

RANGE ≈ 4σ

To construct Confidence interval we replaced σ with S, because we didn't know σ…. THIS Implies a MARGIN OF ERROR because we used the SAMPLE standard deviation.

The desired margin of error for capturing the population mean using $\bar{x}$ should be $Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq B$

(Where B **is the DESIRED margin of error)**

So… $n > \left(\frac{Z_{\frac{\alpha}{2}}}{B}\right)^2 \sigma^2$

There are two ways to calculate n for a confidence interval, if you don't have access to σ:

1) **you can use S, if you trust the data:** $n > \left(\frac{Z_{\frac{\alpha}{2}}}{B}\right)^2 s^2$
2) **you can use the** $\left(\frac{RANGE}{4}\right)^2$ **:** $n > \left(\frac{Z_{\frac{\alpha}{2}}}{B}\right)^2 \left(\frac{RANGE}{4}\right)^2$

ALWAYS ROUND UP!

Choosing a SAMPLE SIZE n to study a POPULATION PROPORTION, P.

To construct a 100(1-α)% CI for P, we would use $\hat{p} \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}}$

But we don't have P, so use :

$Z_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}} < B$ **therefore** $n > \left(\frac{Z_{\frac{\alpha}{2}}}{B}\right)^2 pq$

$f(p) = P(1-P), where\ 0 < P < 1$; IF P = ½ , MAX $(f(p)) = 1/4$

$\boldsymbol{t = \frac{(\bar{x}-\mu)}{s/\sqrt{n}}}$ Can be described as "has t-distribution with n-1 degrees of freedom" or "has t(n-1)"

PROPERTIES OF THE t-Distribution with *r* degrees of freedom (*df*) (where *r ≥ 1*)

1. It is BELL SHAPED, with heavier tails than a standard normal distribution
   As the degrees of freedom, r, increases, the t-distribution approaches N(0,1). It becomes normal when r ≥ 30
2. The t-distribution is identified by its *df*, *r*
3. *t-distribution ONLY APPLIES if the ORIGINAL POPULATION is NORMALLY DISTRIBUTED*

ONE SIDED:

$H_0: \mu = \mu_0$

$H_a: \mu > \mu_0$

Test statistic: $\boldsymbol{\mathcal{T} = \frac{(\bar{x}-\mu_0)}{s/\sqrt{n}}}$ **critical value at level α, $t_\alpha(n-1)$**

P-value = $P(t_\alpha(n-1) > \mathcal{T})$

Conclusion:

- **critical value approach**: Reject H_0 if $\mathcal{T} > t(n-1)$
- **P-value:** Reject H_0 if **P-VALUE ≤ α**

Two-sided:

$H_0: \mu = \mu 0$

$H_a: \mu \neq \mu 0$

Test statistic $\boldsymbol{\mathcal{T} = \frac{(\bar{x}-\mu_0)}{s/\sqrt{n}}}$ **critical value at level α, $-t_{\alpha/2}(n-1)$ and $t_{\alpha/2}(n-1)$**

P-value = $P(t(n-1) < |\mathcal{T}|) + P(t(n-1) > |\mathcal{T}|) =$ **$2\,P(t_\alpha(n-1) > |\mathcal{T}|)$**

Conclusion:

- critical value approach: Reject H_0 if $\mathcal{T} < -t_{\alpha/2}(n-1)$ **OR** $\mathcal{T} > t_{\alpha/2}(n-1)$
- **P-value:** Reject H_0 **IF P-VALUE ≤ α**

Confidence intervals based on n< 30 : $\boldsymbol{\bar{x} \pm t_{\frac{\alpha}{2}}(n-1)\frac{s}{\sqrt{n}}}$

Test statistic: $\mathcal{T} = \frac{\bar{x}_1 - \bar{x}_2 - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ **Pooled variance:** $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$

Case 1: Directional hypothesis

H_0 : $\mu_1 - \mu_2 = D_0$

H_a : $\mu_1 - \mu_2 > D_0$

Critical value at level alpha t_α(n1 +n2 -2)
P-value = p(t_α(n1 +n2 -2) > $\mathcal{T}$)
Conclusion:

- **Critical value approach: Reject H_0 IF $\mathcal{T}$> t_α(n1 +n2 -2)**
- **P-value approach: Reject H_0 IF P-value <= α**

Case 2: Opposite direction:

H_0 : $\mu_1 - \mu_2 = D_0$

H_a : $\mu_1 - \mu_2 < D_0$

Critical value at level α: $-t_\alpha$(n1 +n2 -2)
P-value = p(t_α(n1 +n2 -2) < $\mathcal{T}$)
Conclusion:

- **Critical value approach: Reject H_0 IF $\mathcal{T}$ < $-t_\alpha$(n1 +n2 -2)**
- **P-value approach: Reject H_0 IF P-value <= α**

Case 3: TWO SIDED

H_0 : $\mu_1 - \mu_2 = D_0$

H_a : $\mu_1 - \mu_2 \neq D_0$

Test statistic $\mathcal{T} = \frac{\bar{x}_1 - \bar{x}_2 - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$

Critical value at level α: $-t_{\alpha/2}$(n1 +n2 -2), $+t_{\alpha/2}$(n1 +n2 -2)

P-value = P($t_{\alpha/2}$(n1 +n2 -2) < $-|\mathcal{T}|$ + $t_{\alpha/2}$(n1 +n2 -2) > $|\mathcal{T}|$) = **2P($t_{\alpha/2}$(n1 +n2 -2) > $|\mathcal{T}|$)**

Conclusion:

- **Critical value approach: Reject H_0 IF $\mathcal{T}$ < $-t_{\alpha/2}$(n1 +n2 -2) OR $\mathcal{T}$ > $t_{\alpha/2}$(n1 +n2 -2)**
- **P-value approach: Reject H_0 IF P-value <= α**

CONFIDENCE INTERVALS when n1 and n2 are small….

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}(n_1+n_2-2)} \sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

CI for n > 30: $\bar{x} + z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)$

CI for n < 30 : $\bar{x} + t_{\frac{\alpha}{2}(n-1)}\left(\frac{s}{\sqrt{n}}\right)$

Small sample inference for the difference between two population means

- **TWO SAMPLES** must be **INDEPENDENT**
- Populations must be **NORMALLY DISTRIBUTED**
- **If even one sample is less than 30, *you must use the t-test***
- **The VARIANCES** of the two populations **MUST coincide** $\sigma_1^2 = \sigma_2^2$
    - **Test this by:** $\frac{Larger\ S^2}{Smaller\ S^2} \leq 3$ **means we can conclude** $\sigma_1^2 = \sigma_2^2$

Situation where TWO SAMPLES are NOT independent: PAIRED DIFFERENCE INFERENCE

Take differences – treat as a sample of n coming from ONE population:

H_0 : $\mu_A - \mu_B = 0$ - replace with H_0 : $\mu_d = 0$

H_a : $\mu_A - \mu_B \neq 0$ - replace with H_a : $\mu_d \neq 0$

With α = 0.05, α/2 = 0.025

Test statistic: $\mathcal{T} = \frac{(\bar{d}-0)}{s_d/\sqrt{n}}$

For TWO-TAILED:

P-Value = $2P(t_{\alpha/2(n-1)} > |\mathcal{T}|)$

Conclusion:

- **Critical Value Approach: Reject H_0 if** $\mathcal{T} < -t_{\alpha/2(n-1)}$ or $\mathcal{T} > t_{\alpha/2(n-1)}$
- **P-value approach: Reject H_0 if** P-value ≤ α

Z-TEST (Large Samples, Standard Normal)

To test a hypothesis on a large sample you need:

1. Null and alternative hypotheses
2. Test Statistic (computed based on the data)
3. Critical point(s) are identified by the so-called level of the test α – (and α should be something small, i.e. 0.01, 0.05, 0.1), or by p-value
    a. P-value is the minimum value of α that results in rejection of the null hypothesis. (obtained from the data)
4. Conclusion: Either REJECT $H_{\emptyset}$ in favor of H_a, OR DO NOT reject $H_{\emptyset}$

	REJECT $H_{\emptyset}$	**Fail to REJECT $H_{\emptyset}$**
$H_{\emptyset}$ is TRUE	Type I error	CORRECT decision
$H_{\emptyset}$ is NOT TRUE	CORRECT decision	Type II error

Cannot control both types of errors at the same time so we generally try to control Type I error.
α = P(Type I Error)

POWER OF TEST = 1-β = P(Reject H_0 when H_a is true)
= P (Truly rejecting H_0)

Case 1 DIRECTIONAL HYPOTHESIS

$H_{\emptyset} : \mu = \mu_0$

$H_a : \mu > \mu_0$

The test statistic is $\mathcal{Z} = \frac{(\bar{x}-\mu_0)}{s/\sqrt{n}}$

α is the maximum P-value you can accept
Minimum Critical value you can accept at level α is Z_{α}
P-value = P(Z > $\mathcal{Z}$) = 1 – P(Z < $\mathcal{Z}$) (from tables)
Where Z is Z standard normal
Conclusion:

- **Critical value approach says that we can reject H_0 if $\mathcal{Z} > Z_{\alpha}$**
- **P-Value approach says REJECT H_0 IF P-Value is <= α**

MARGIN OF ERROR

$|\hat{p} - p| \leq Z_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}}$ → **100%(1-α) MARGIN of error for P**

$|\bar{x} - \mu| \leq Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ → **100%(1-α) MARGIN of error for μ**

Directionality is crucial – pay attention to the direction of the inequality

Hypothesis testing for TWO SAMPLES from TWO POPULATIONS

When working with two SAMPLES from two populations, they must be INDEPENDENT

One sided:

$H_0: \mu_1 - \mu_2 = D_0$

$H_a: \mu_1 - \mu_2 > D_0$

Test Statistic $\mathcal{Z} = \dfrac{\bar{x}_1 - \bar{x}_2 - D_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

At the level α, the critical value = Z_α

P-value = $P(Z > \mathcal{Z}) = 1 - P(Z < \mathcal{Z})$

Conclusion:

- Critical Value approach: **Reject H_0 IF $\mathcal{Z} > Z_\alpha$**
- P-value approach: **Reject H_0 IF P-value $\leq \alpha$**

If the direction of the evaluation for H_a reverses : $\mu_1 - \mu_2 < D_0$ then

P-value = $P(Z < \mathcal{Z})$

Conclusion:

- Critical Value approach: Reject H_0 IF $\mathcal{Z} < -Z_\alpha$
- P-value approach: Reject H_0 IF P-value $\leq \alpha$

Two SIDED:

$H_0: \mu_1 - \mu_2 = D_0$

$H_a: \mu_1 - \mu_2 \neq D_0$

Calculate Z. look up probability for Z from tables, and multiply by 2 , as this is the probability

P-value = $P(Z > |\mathcal{Z}|) + P(Z < -|\mathcal{Z}|) = \mathbf{2P(Z > |\mathcal{Z}|) = 2(1-(Z < |\mathcal{Z}|)}$

Conclusion:

- Critical value approach: **Reject H_0 IF $\mathcal{Z} < -Z_{\alpha/2}$ OR $\mathcal{Z} > Z_{\alpha/2}$**
- P-Value approach: **Reject H_0 IF P-value $\leq \alpha$**

Hypothesis testing for Population Proportion

One SIDED:

H_0: P = P_0

H_a: P > P_0

Test Statistic $\mathcal{Z} = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o q_o}{n}}}$

CRITICAL VALUE at level α is Z

P-VALUE $= p(Z > \mathcal{Z})$

Two sided approach:

H_0: P = P_0

H_a: P ≠ P_0

CRITICAL VALUE at level alpha is $Z_{\alpha/2}$

P-VALUE =$P(Z > |\mathcal{Z}|) + P(Z < -|\mathcal{Z}|)$ **= 2P(Z >** $|\mathcal{Z}|$)

Critical value approach: Reject H_0 IF $\mathcal{Z} < -Z_{\alpha/2}$ OR $\mathcal{Z} > Z_{\alpha/2}$

P-Value approach: Reject H_0 IF P-value $\leq \alpha$

Hypothesis testing for DIRECTIONAL DIFFERENCE between TWO POPULATION PROPORTIONS

$H_0 : \hat{p}_1 - \hat{p}_2 = 0$

$H_a: \hat{p}_1 - \hat{p}_2 > 0$

POOLED PROPORTION: $\hat{p} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$; $\hat{q} = 1 - \hat{p}$

TEST STATISTIC: $\mathcal{Z} = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\hat{p}\hat{q}\frac{1}{n_1} + \frac{1}{n_2}}}$

MUST CHECK CONDITIONS!: $n_1\hat{p}_1, n_1\hat{q}_1, n_2\hat{p}_2, n_2\hat{q}_2 >= 5$

Reject if test statistic is LESS THAN Z_α

Conclusion for two sided uses $Z_{\alpha/2}$

CONCLUSION STATEMENT (if rejecting H_0): The proportion of X with (condition) is higher than that Y to a significance level of α=0.05

How to construct a HYPOTHESIS TEST using CONFIDENCE INTERVALS

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$ **-> 100(1-alpha)% CI for μ_1 - μ_2**

H_0 : $\mu_1 - \mu_2 = 0$
H_a : $\mu_1 - \mu_2 \neq 0$ at level α

Eg… calculate a 95% CI for μ_1 - μ_2….
IF the value of H_0 (i.e. $\mu_1 - \mu_2 = 0$) is WITHIN THE CONFIDENCE INTERVAL, ***you must ACCEPT H_0,***
Otherwise REJECT H_0

Poisson distribution

A random variable that is associated with a certain event in a specific location.
i.e. number of accidents at an intersection at a certain time of day
number of earthquakes in Ottawa in a year
number of cases coming into the emergency ward on a Saturday night

When X counts the number of events in a period of time and/or space so that the average, μ of these events are expected to happen, then X has a Poisson(μ)

Probability distribution of Poisson (μ)

PMF – Probability MASS Function $P(x) = \frac{e^{-\mu}\mu^{k}}{k!}$, $k = 0, 1, 2, 3, \ldots$

Poisson is a discrete random variable
e = 2.71828

eg. If the # of car accidents at the intersection of Carleton U and colonel by has poisson distribution with mean or average of 6 accidents per year.

a) What is the probability that next year, there will be exactly 5 accidents?
   $P(5) = \frac{e^{-6}6^{5}}{5!}$
b) What is the probability that there will be at most 2 accidents next year?

   $P(x \le 2) = P(0) + P(1) + P(2) = \frac{e^{-6}6^{0}}{0!} + \frac{e^{-6}6^{1}}{1!} + \frac{e^{-6}6^{2}}{2!}$

   (POISSON DISTRIBUTION TABLES IN BOOK provide these numbers)

c) What is the probability of having exactly 1 car accident in the next **6 MONTHS**?
   Divide **μ/2 = 6/2 = 3**

   P(x=1) = $\frac{e^{-3}3^{1}}{1!}$

Mean and variance of a V.V. X which has poisson (μ):
Mean = E(x) = μ = np
Variance = σ^2 = μ
Std dev = σ = $\sqrt{\mu}$

(Note that text has Poisson distribution tables – column is **μ,** rows are **x<=n**)

P(x = 2) = P(x <= 2) – P(x <= 1)
P(x >= 2) = 1- P(x <= 1)
P(2 <= x <= 5) = P(x<=5) – P(x<=1)
P(2 < x <= 5) = P(x <=5) – P(x <= 2)

Be comfortable with being able to do these kinds of permutations and calculations

Approximating a BINOMIAL distribution by Poisson:

Let X have a binomial distribution, $\mathbf{BINOMIAL(n, p)}$

P(x <= k) - you can get either x or k or both from the Cumulative Binomial Probability tables (table stops at n=20 or 25, with jumps of 5 numbers), or use the POISSION APPROXIMATION.

Use POISSON APPROXIMATION instead of binomial calculation, **IF n is large AND n*p = < 7 then**

P(x = k) is $\approx \frac{e^{-np}(np)^k}{k!}$ **And** $\mathbf{P(x \leq k) \approx P(Y \leq k)}$ **when Y has POISSON(np)**

(For poisson approximation, μ = np, then use tables)

Hyper Geometric Distributions

Box of N Balls, M are red, N-M are blue
Take a sample of size n, let x = # of red balls in the sample of size n

P(k) = P(x=k) $= \dfrac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$ k = 0,1,2,… k is restricted to min(n, M)

(in how many ways can you choose n out of N? C $\binom{N}{n}$)

n= sample size
M = total SUCCESSES in POPULATION
N = total POPULATION size
k = # of successes in sample

Mean of HYPER DISTRIBUTION = $n\left(\frac{M}{N}\right)$
Variance of HYPER DISTRIBUTION = $n\left(\frac{M}{N}\right)\left(\frac{N-M}{N}\right)\left(\frac{N-n}{N-1}\right)$

Hyper geometry from binomial:

Binomial probabilities don't change (with replacement)

In Hyper geometries the probability changes with each sample (without replacement)

Eg. A box of 25 lamps, with 6 defective. Take a sample of n=3, WITHOUT replacement.
What is the probability of having 2 defective lamps in the sample set of 3?

P(k) = P(x=2) $= \dfrac{\binom{6}{2}\binom{25-6}{3-2}}{\binom{25}{3}} = \dfrac{\binom{6}{2}\binom{19}{1}}{\binom{25}{3}}$ =

Male/female in large population can follow hypergeometric distribution.