STAT381 **Lia Liu**

Chap 3: Continuous Probability Models

List of topics: Section 3.1

1. **Continuous random variable: a random variable that takes values of an entire interval.**
2. **A function $f(y)$ is a *probability density function* of a continuous random variable (in short pdf) for some random variable Y if it satisfies two conditions:**

   1) $f(y) \geq 0$;

   2) $\int_{-\infty}^{\infty} f(y)dy = 1$
3. **Distribution function of a random variable(in short, DF):**

   $F(y) = P(Y \leq y),\ -\infty < y < \infty$

   Note: $F(-\infty) = 0, F(\infty) = 1$, **F is increasing(nondecreasing).**
4. **Probability Density function of a continuous random variable (in short pdf):**

   $f(y) = F'(y)$
5. **How to compute probability?**

   1) using pdf: $P(a \leq Y \leq b) = \int_a^b f(y)dy$

   2) using DF: $P(a \leq Y \leq b) = F(b) - F(a)$

   Note the difference between discrete and continuous r.v.: If Y is a continuous r.v., then

   $P(a \leq Y \leq b) = P(a < Y \leq b) = P(a < Y < b) = P(a \leq Y < b)$ **because** $P(X = a) = 0$**(area of a line segment).**
6. **Expected value** $\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$

   Variance $\sigma^2 = V(X) = E(X - \mu)^2 = E(X^2) - \mu^2$

   Standard deviation $\sigma = \sqrt{V(X)}$

Sections 3.2

List of topics: Normal Distribution

1. **A continuous rv is said to have a** *normal distribution* **with parameters mean μ and standard deviation σ, where $-\infty < \mu < \infty$ and $\sigma > 0$, if the pdf of X is**

   $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}, \ -\infty < x < \infty$

2. **Standard normal: when $\mu = 0$ and $\sigma = 1$.**

3. **How to compute probability of a Standard Normal rv using Table C4? TI-83 or TI-84?** $P(a < Z < b) = normalcdf(a, b)$

4. **How to compute probability of any Normal rv using Table C4? Standardize or** $P(a < X < b) = normalcdf(a, b, \mu, \sigma)$

5. **The Empirical Rule (68-95-99.7 rule):**

   If a rv Y satisfies a Normal Distribution with mean μ and standard deviation σ, then

   1) The probability that Y is within one standard deviation from the mean

   =$P(|Y - \mu| \leq \sigma)$ **is roughly** 68%**;**

   2) The probability that Y is within two standard deviation from the mean

   =$P(|Y - \mu| \leq 2\sigma)$ **is roughly** 95%**;**

   3) The probability that Y is within three standard deviation from the mean

   =$P(|Y - \mu| \leq 3\sigma)$ **is roughly** 99.7%**;**

6. **How to use Normal to approximate Binomial distribution? Half unit correction**

Sections 3.3

List of topics: Other Cont. Distribution

First let us look at a type of questions concerning the life time of something(e.g. human life, lifetime of a light bulb, TV set, etc.). We'll derive the formula so you can understand where the pdf came from.

Let a continuous rv X represent the length of life of a component(e.g. human life, lifetime of a light bulb, TV set, etc.), and assume X has pdf $f(x)$ and CDF $F(x)$.

From Calculus, we have

$$P(x < X < x + \Delta x | X > x) = \frac{P(x < X < x + \Delta x)}{P(X > x)} \approx \frac{f(x)\Delta x}{1 - F(x)}$$

Let $\lambda(x) = \dfrac{f(x)}{1 - F(x)}$, $\lambda(x)$ is called the *failure rate function.*

Let's see how the failure rate function is related to the pdf and CDF.

Take a integral to $\lambda(x)$, we have

$\int_0^x \lambda(x)dx = \int_0^x \dfrac{f(x)}{1 - F(x)}dx$ **by substitution,**

$\int_0^x \lambda(x)dx = -\ln[1 - F(x)]$**. Now solve for F:**

$F(x) = 1 - e^{-\int_0^x \lambda(x)dx}$**.**

So $f(x) = \lambda e^{-\int_0^x \lambda(x)}dx$**.**

Now we can find out what the pdf is if we know what is the failure rate function.

If the parts "wear out", the lifetime deteriorates over time(e.g. human life, lifetime of a light bulb, TV set, etc.), then the failure rate $\lambda(x)$ is increasing according to time x. What kind of increasing function do you know? Power, exponential, log, etc.

When we pick a power function, we call the distribution"Weibull".

when we pick an exponential function, we call it"Gompertz".

If the old is just as good as new, we say it has no memory. Then the Failure rate is a constant. Our distribution is then called Exponential distribution. We use

1. **3.3.1. Weibull Distribution: A continuous rv X is said to have a *Weibull Distribution* with parameters α and β, if the pdf of X is**

   $f(x;\alpha,\beta) = \frac{\alpha}{\beta^\alpha} x^{\alpha-1} e^{-(x/\beta)^\alpha},\ x \geq 0$

2. **3.3.2. Gompertz distribution:**

   If the parts has no memory, then $\lambda(x) =$ constant. The distribution is exponential.

   If it increase like a power function, match the constant, we get a Weibull distribution (so $\lambda(x) = \alpha x^{\alpha-1}/\beta^\alpha$)

   Human mortality (failure rate) increases exponentially once a person reaches mid twenties.

   $\lambda(x) = ce^{bx}$

   The distribution is called Gompertz.

3. **3.3.4. Gamma Distribution: A continuous rv X is said to have a *gamma distribution* with parameters α and β, if the pdf of X is**

   $f(x;\alpha,\beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-(x/\beta)},\ x \geq 0$

4. **$\Gamma(\alpha)$ is called gamma function which is defined by**

   $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x},\ \alpha > 0$

5. **Properties of a gamma function:**

   - **Recursive: For any $a > 1, \Gamma(a) = (a-1)\cdot\Gamma(a-1)$**
   - **For any positive integer n, $\Gamma(n) = (n-1)!$**
   - $\Gamma(1/2) = \sqrt{\pi}$

- **The mean and variance of r.v X with a gamma distribution is**
  $E(X) = \mu = \alpha\beta;\ V(X) = \sigma^2 = \alpha\beta^2$

  The gamma distribution is used to model the waiting time until kth occurrence. If λ is the mean number of occurrences in an interval of length 1, then the waiting time until kth occurrence is Gamma distribution with $\alpha = k, \beta = 1/\lambda$

6. **Special case I: Exponential distribution: Waiting time between any two occurrences is Gamma distribution with $\alpha = 1, \beta = 1/\lambda$**

   pdf of an exponential rv, X, is

   $f(x;\lambda) = \lambda e^{-\lambda x},\ x \geq 0$

   The mean waiting time is $1/\lambda$.

   We usually use exponential distribution to model waiting time between two occurrences, if the process is *memoryless*.

7. **Special case II: 3.3.5. Chi-square distribution: is Gamma distribution with $\alpha = v/2, \beta = 2$. We shall see it later on.**

8. **3.3.6. Lognormal distribution: A continuous rv X is said to have a *Lognormal Distribution* with parameters μ and σ, if the pdf of X is**

   $$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-(\ln(x)-\mu)^2/(2\sigma^2)},\ x \geq 0$$

   If Y has a Normal distribution $N(\mu, \sigma^2)$, then $X = e^Y$ has a lognormal distribution.

   How to derive the pdf of X?

   $$F(x) = P(X \leq x) = P(e^Y \leq x) = P(Y \leq \ln x) = \int_{-\infty}^{\ln x} ce^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

   $f(x) = \frac{d}{dx}F(x)$

9. **Model that relates Poisson distribution, Gamma distribution and Exponential distribution:**

If the process is "memoryless", i.e. the number of occurrences depends only on the length of the time interval, not the beginning point; in another word, the distribution of the number of occurrences over $[0, x]$ is the same as $[k, k + x]$ for any k:

Another way to derive it:

- **Then let** $W =$ *the number of occurrences during* $[0, x]$**, let** $\lambda =$ **average rate of occurrences during unit time interval.**

  Then W **has a Poisson distribution with mean** $E(W) = \lambda x$

  (i.e. the pdf of W **is** $f_W(w) = e^{-\lambda x}\frac{(\lambda x)^w}{w!},\ w = 0, 1, 2, ...$**)**
- **Let** X **be** *the waiting time between any two occurrences*

  The X **has an Exponential distribution with mean** $1/\lambda$**.**

  Proof: $F_X(x) = P(X \leq x) = 1 - P(X > x)$

  $= 1 - P($**zero occurrence during** $[0, x])$

  $= 1 - P(W = 0) = 1 - e^{-\lambda x}$

  $f_X(x) = F' = \lambda e^{-\lambda x},\ x > 0$
- **Let** Y **be** *the length of time until rth occurrence***. Then** Y **satisfies a Gamma distribution.**

  Proof: Same idea: $F_Y(y) = P(Y \leq y) = 1 - P(Y > y)$

  $= 1 - P(W \leq r - 1) = 1 - \Sigma_{w=0}^{r-1} e^{(-\lambda y)}\frac{(\lambda y)^w}{w!}$

  Take derivative term by term:

  $f_Y(y) = \frac{\lambda^r}{(r-1)!} y^{r-1} e^{-(\lambda y)}, y > 0$

Sections 3.5 Dist of 2 cont. r.vs

Print out the extra problems for double integral on my webpage.

List of topics:

1. **Continuous case: Joint density**
   $f_{X,Y}(x,y)$
   Note: This f is not a probability. However,
   (1) $f \geq 0$;
   (2) $\int\int_{R^2} f(x,y)dxdy = 1$
   (3) $P[(x,y) \in A] = \int\int_A f(x,y)dxdy$
2. **CDF of X and Y:**
   $F_{X,Y}(x,y) = P[(X \leq x) \cap (Y \leq y)]$
3. **Continuous case: Joint density**
   $f_{X,Y}(x,y) = \dfrac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}$
   (2nd mixed partial derivative)
4. **Marginal pdfs:** $f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$
5. **Expectations:**
   $E(u(X,Y)) = \int\int_{R^2} u(x,y)f(x,y)dxdy$
6. **Conditional pdf is the ratio of joint pdf to marginal pdf.**
   Continuous case: $f_{X|Y}(x|y) = \dfrac{f_{X,Y}(x,y)}{f_Y(y)},\ f_Y(y) > 0$
7. **Conditional expectation:** $E(u(Y)|x) = \int_{-\infty}^{\infty} u(y)f_{Y|X}(y|x)dy$ **(which is a function of x.)**
   Hence $E(Y|X)$ is a random variable of X.
8. **Covariance of X and Y:**
   $cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$ **(easier to understand)**
   $cov(X,Y) = E(XY) - \mu_X\mu_Y$ **(easier for calculation)**

9. Correlation coefficient $\rho = \dfrac{cov(X,Y)}{\sigma_X \sigma_Y}$

How can you tell X and Y are independent r.v.s?

1. **Equivalent conditions: X and Y are independent iff**

   (1) By pdf: $f_{X,Y}(x,y) \equiv f_X(x) f_Y(y)$ **(note the regions.)**

   (2) By CDF: $F(x,y) = F_X(x) F_Y(y)$ **for all** (x,y)

   (3) By MGF: $M(t_1,t_2) = M(t_1,0)M(0,t_2)$ **(Not covered in stat381)**

2. **How can you tell two r.v. X and Y are independent by joint density?**

   Region and function

3. **What happens to covariance(and corelation coefficient) if X and Y are independent?**

   If X and Y are independent, then $Cov(X,Y) = 0$, hence $\rho = 0$.

   However, if $Cov(X,Y) = 0$, or $\rho = 0$, X and Y may not be independent.

   Facts:

   - $-1 \le \rho \le 1$
   - $\rho = \pm 1$ **implies"perfect linear relationship between X and Y."**
   - $\rho = 0$ **implies "No linear relationship between X and Y (but can be quadratic, for example)."**
   - $\rho > 0$ **means X and Y are positively related (If X increases, so is Y).**
   - $\rho < 0$ **means X and Y are negatively related (If X increases, Y decreases).**
   - **If X and Y are independent, then $Cov(X,Y) = 0$, hence $\rho = 0$.**
     However, if $Cov(X,Y) = 0$, or $\rho = 0$, X and Y may not be independent.

- **Special case: If X and Y are normal, then $Cov(X, Y) = 0$, or $\rho = 0$ implies "X and Y are independent."**

Section 4.1 Sampling distribution and the CLT

List of topics:

Given n rv's $X_1, X_2, ..., X_n$ (not necessarily random sample). We consider linear combination

$Y = a_1X_1 + a_2X_2 + ... + a_nX_n$

1. **Expectation is linear:** $E(Y) = a_1E(X_1) + a_2E(X_2) + ... + a_nE(X_n)$
2. **Variance is not linear, but we still have**
   $V(Y) = \Sigma_{i=1}^n \Sigma_{j=1}^n a_i a_j Cov(X_i, X_j)$
   $= {a_1}^2 V(X_1) + {a_2}^2 V(X_2) + ... + {a_n}^2 V(X_n) + 2a_1a_2Cov(X_1, X_2) + ... + 2a_{n-1}a_nCov(X_{n-1}, X_n)$
3. **In case X_i's are independent, all covariances are zero.**
   $V(Y) = {a_1}^2 V(X_1) + {a_2}^2 V(X_2) + ... + {a_n}^2 V(X_n)$
4. **Example:** $E(X_1 - X_2) = E(X_1) - E(X_2)$;
   $V(X_1 - X_2) = V(X_1) + V(X_2) - 2Cov(X_1, X_2)$;
   If X_1, X_2 are independent, $V(X_1 - X_2) = V(X_1) + V(X_2)$
5. **Example: X_1 is a rv with Binomial B(n,p);**
   X_2 is a rv with Poisson λ;
   X_3 is a rv with Normal $N(\mu, \sigma^2)$;
   Suppose X_i's are independent for $i = 1, 2, 3$.
   Find $E(X_1 + 2X_2 - 3X_3)$; $V(X_1 + 2X_2 - 3X_3)$;
   $E(X_1 + X_2 + X_3)$; $V(X_1 + X_2 + X_3)$;
   $E(\frac{X_1 + X_2 + X_3}{3})$; $V(\frac{X_1 + X_2 + X_3}{3})$;
6. **Stable distributions:**
   Normal, Binomial, Poisson, Chi-square.
   Not stable: Exponential
7. **Independent and identically distributed random variables(i.i.d.)**

8. **random sample (iid)**

9. **Sample mean** $\overline{X}$**,** $E(\overline{X}), V(\overline{X})$

10. **Sample sum** $S_n = X_1 + X_2 + ... + X_n$**, what is** $E(S_n)$**?** $V(S_n)$**?**

11. **Central Limit Theorem: Roughly speaking, no matter what is the population distribution, as long as the sample size is large, the sample mean and the sample sum are normal.**

    Let $X_1, X_2, ...$ **be iid's, each with mean** μ **and standard deviation** σ**, (both finite), then**

    Part 1) $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \to N(0,1)$ **as** $n \to \infty$

    Part 2) $\dfrac{S_n - n\mu}{\sigma\sqrt{n}} \to N(0,1)$ **as** $n \to \infty$

Sections 4.2-4.6 Confidence Intervals

List of topics: Given a random sample $X_1, X_2, ..., X_n$ from the same distribution with unknown parameters. What can we say about the unknown parameter?

- **confidence interval:** *estimate* $\pm$ *Margin of error*
  How?
- **Choice of test statistic. Why?**
- **What can/cannot you say?**

1. **Given a random sample from a Normal population with unknown μ (Given standard deviation σ), what can you say about μ?**
   Use z test: $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$
   The $1-\alpha$ CI for μ is $\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
   The estimate for μ is $\overline{X}$, $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is the margin of error. Notice that the margin of error is $z*$ times the standard deviation of $\overline{X}$.
   What about 2 samples?
2. **Given a random sample from a Normal population with unknown μ and unknown standard deviation σ, , what can you say about μ?**
   Use t test: $T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ has t distribution with $n-1$ degree of freedom.
   The $1-\alpha$ percent CI for μ is $\overline{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$
   What about 2 samples? $\mu_1 - \mu_2$
3. **Large Population: Normal with unknown μ and unknown σ. ($n > 40$)**
   Use z test: $Z \approx \frac{\overline{X}-\mu}{S/\sqrt{n}}$

The $1-\alpha$ percent CI for μ is $\overline{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$

4. **Population: Binomial with unknown population proportion p.**

   When n is large ($np \geq 10, nq \geq 10$), Choose test statistics: $\hat{Y} = Y/n$, where Y is the number of heads out of n tosses.

   Use Z- test: $Z \approx \frac{\hat{p}-p}{\sqrt{\hat{p}\hat{q}/n}}$

5. **Two populations: $p_1 - p_2$:Bernoulli trials: under proper assumptions,**

   $\hat{p_1} - \hat{p_2}$ has mean $p_1 - p_2$ and SD $\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$

6. **Standard deviation is usually the trouble-maker because it is usually unknown. We shall use various defined standard errors to approximate.**

7. **To find confidence interval for $p_1 - p_2$, use "hats" to approximate:**

   $SE(\hat{p_1} - \hat{p_2}) = \sqrt{\frac{\hat{p_1}\hat{q_1}}{n_1} + \frac{\hat{p_2}\hat{q_2}}{n_2}}$

8. **To do hypothesis testing, e.g**

   $H_0 : p_1 - p_2 = 0$

   Use pooled standard error

   $SE_{pooled}(\hat{p_1} - \hat{p_2}) = \sqrt{\hat{p}_{pooled}\hat{q}_{pooled}(\frac{1}{n_1} + \frac{1}{n_2})}$,

   where $\hat{p}_{pooled} = \frac{n_1\hat{p_1}+n_2\hat{p_2}}{n_1+n_2}$

   When n is large enough to use Z-test, the z-score is $Z = \frac{(\hat{p_1}-\hat{p_2})-0}{SE_{pooled}(\hat{p_1}-\hat{p_2})}$. Then according to H_A hypothesis, compute the corresponding P-value.

9. **Given a random sample from a Normal population with unknown μ and unknown standard deviation σ, , what can you say about σ^2?**

   Use χ^2 test: $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ has χ^2 distribution with $n-1$ degree of freedom.

The $1-\alpha$ CI for σ^2 is $(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}})$

Construction process:

1. **Z-test: Let** $P(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) = 1-\alpha$
   Solve the inequality for μ.
2. **t-Test:**
   - **What is a t-distribution?**
     If a rv Z has N(0,1), V has χ^2_n distribution, then $T_n = \frac{Z}{\sqrt{V/n}}$ has t-distribution with df n.
   - **The pdf of a t-distribution with df n is**
     $f(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \frac{1}{(1+\frac{t^2}{n})^{(n+1)/2}}, \ -\infty < t < \infty$ **.**
   - $E(T_n) = 0, V(T_n) = \frac{n}{n-2}$ **What happens if n is large?**
   - **Since $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ is N(0,1) and $V = \frac{(n-1)S^2}{\sigma^2}$ has χ^2_{n-1} distribution,**
     Hence $T_{n-1} = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}}} = \frac{\overline{X}-\mu}{s/\sqrt{n}}$ has t-distribution with df $n-1$.
     Proof: T_n^2 has an F-distribution F(1,n).
3. **For Binomial distribution, X= number of heads out of n tosses.**
   $E(X) = np, V(X) = npq$**.**
   - **Let $\hat{p} = X/n$, then $E(\hat{p}) = p, V(\hat{p}) = \frac{pq}{n}$.**
   - **When n is large ($np \geq 10, nq \geq 10$), use Normal:**
     $P(-z_{\alpha/2} < \frac{\hat{p}-p}{\sqrt{pq/n}} < z_{\alpha/2}) = 1-\alpha$
     Now p, q are unknown, one way to overcome the difficulty is to use $\hat{p}$ to replace p, we get the confidence interval for p:
     $\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$

- **Sometimes, we are given the accuracy requirement, we need to find the sample size n.**
  $P(-d \leq \hat{p} - p \leq d) = 1 - \alpha$**, minimum value of** pq **is** $1/4$**, solve for n:** $n = \frac{z^2_{\alpha/2}}{4d^2}$

4. χ^2_n **distribution:**

- χ^2_n **distribution is a special case of the Gamma distribution with** $\alpha = n/2, \beta = 2$**.**
  Hence $E(\chi^2_n) = \alpha\beta = n; V(\chi^2_n) = \alpha\beta^2 = 2n$**.** n **is called the degree of freedom.**
- **If** Y_1 **has** χ^2_n**,** Y_2 **has** χ^2_m**,** Y_1 **and** Y_2 **are independent, then** $Y_1 + Y_2$ **has** χ^2_{n+m}**. In another word,** χ^2 **distribution is stable.**
- **If** $Z_1, Z_2, ..., Z_n$ **is a random sample from N(0,1), then** $\Sigma_{i=1}^n {Z_i}^2$ **has** χ^2_n**. (Note the degree of freedom)**
- **If** $Y_1, Y_2, ..., Y_n$ **is a random sample from** $N(\mu, \sigma^2)$**, then** $\frac{(n-1)S^2}{\sigma^2} = \Sigma_{i=1}^n (\frac{Y_i - \overline{Y}}{\sigma})^2$ **has** χ^2_{n-1} **distribution. (Note the loss of 1 degree of freedom).**
- **Using Table C5, we pick the confidence interval with each tail** $\alpha/2$**,**
  $P(\chi^2_{1-\alpha/2,n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2,n-1}) = 1 - \alpha$

Case Study: Coral are decline worldwide, possibly because of pollution or changes in sea temperature. One kind of coral, the sea fan, looks like a plant growing from the sea floor, but is actually an animal. In June 2000, Dr. Drew Harvell's lab randomly sampled some sea fans at the Las Redes Reef in Akumal, Mexico, at a depth of 40 ft. They found that 54 of the 104 sea fans they sampled were infected with disease.

The sample proportion $\hat{p} = 54/104 = 51.9\%$. What can we say about the population proportion p?

$\sqrt{\hat{p}\hat{q}/n} = 4.9\%$, hence 95% confidence interval for p is $(42.1\%, 61.7\%)$

What can we say about p?

True or false:

1. "51.9% of all sea fans on the Las Redes Reef are infected."

2. " It is probably true that 51.9% of all sea fans on the Las Redes Reef are infected."

3. "We don't know exactly what proportion of sea fans on the Las Redes Reef are infected but we know that it is within $(42.1\%, 61.7\%)$.

4. "We don't know exactly what proportion of sea fans on the Las Redes Reef are infected but we know that it is probably within $(42.1\%, 61.7\%)$.

What you cannot say:

5. **Don't suggest that the parameter varies.**
6. **Don't claim that other samples will agree with yours.**
7. **Don't be certain about the parameter. "Between a and b of sea fans are infected"**
8. **Don't forget: it is the parameter. "I am 95% confident that $\hat{p}$ is between so and so"**
9. **Don't claim to know too much. (generalize)**

10. Do take responsibility. Confidence interval is about uncertainty. You are the one who is uncertain, not the parameter. Not all the intervals you compute will capture the true value of the parameter.

Section 4.5 Hypothesis Testing

List of topics:

1. **Compare to a jury trial: A defendant is accused of robbery**

   Step 1: Null Hypothesis: The defendant is assumed innocent (until proven guilty);

   Step 2: The prosecutor present the evidence

   Step 3: Review the evidence: The jury consider"could these data have happened by chance if the null hypothesis were true?" "How unlikely is unlikely?" $50\%,\ 5\%?,\ 1\%$**?**

   Step 4: Verdict: Make a decision based on the evidence "beyond a reasonable doubt".

2. **How to choose the alternative hypothesis?**

3. **Present the evidence–Choose a test statistic**

4. **Review the evidence: Compute the tail probability according to the alternative hypothesis(p-values), assume null hypothesis were true.**

5. **Verdict: Accept or reject the null hypothesis.**

   Ex: Many people have trouble setting up all the features of their cell phones, so a company has developed what it hopes will be easier instructions. The goal is to have at least 95% **of customers succeed. The company tests the new system on 200 people, of whom 188 were successful. Is this strong evidence that the new system fails to meet the company's goal?**

   $H_0 : p = 0.95$

   $H_A : p < 0.95$

   SRS, $np \geq 10, nq = 200 * 0.05 = 10 \leq 10$ **OK to use Z-test.**

   $\frac{x}{n} = \frac{188}{200} = .94$. **Assume null hypothesis is true, then** $E(\hat{p}) = p_0 = 0.95,\ SD(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}} = 0.015$

P-value: $P(\hat{p} < .94) = P(Z < \frac{.94-.95}{0.015}) = 0.252$

Not enough evidence to reject H_0. Not enough evidence to claim the new system fails the goal.

6. **P-Values are the conditional probability of the tail(s) if the null hypothesis were true:**

   (i) If $H_A : p < p_0$, the P-value is

   $P(\hat{p} < \frac{x}{n} | H_0$ **is true); or standardize**

   $$P = P(Z < \frac{\frac{x}{n} - p_0}{\sqrt{\frac{p_0 q_0}{n}}})$$

   (ii) If $H_A : p > p_0$, the P-value is

   $P(\hat{p} > \frac{x}{n} | H_0$ **is true); or standardize**

   $$P = P(Z > \frac{\frac{x}{n} - p_0}{\sqrt{\frac{p_0 q_0}{n}}})$$

   (iii) If $H_A : p \neq p_0$, the P-value is

   $$2P(Z > |\frac{\frac{x}{n} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}| | H_0 \text{ is true}).$$

7. **The smaller the P-value is, the stronger the evidence is against the null hypothesis.**

8. **If given a significance level α, then when P-value $\leq \alpha$, reject the H_0.**

9. **Rule of Thumb: If the significance level is not given, reject the null hypothesis if the P-value is less than 5%.**

10. **If the P-value is larger than the preset significance level α, say "Not enough evidence to reject the null hypothesis", instead of "Accept the null hypothesis". (compare to "Not enough evidence to prove the defendant is guilty", instead of "the defendant is proven innocent".**

11. **If the significance level, α, is given, instead of finding the P-value, you have the alternative way to make a decision, by finding the critical value, $z*$, according to α.**

12. **Two types of errors: (Type I: The defendant is innocent, but the jury find him guilty. Type II: The defendant is guilty, but the jury find him innocent.)**

    Type I: The null hypothesis is true, but we made a wrong decision to reject it.

    Type II: The null hypothesis is false, but we failed to reject it.

13. **The probability of making a type I error is α, which is called the** *significance level of the test.*

14. **The probability of making a type II error is β, which is not easy to calculate.**

15. **When we choose some values of the parameter, and calculate the type II error, the curve is called** *the operating characteristic curve.* **e.g.** $OC(p)$ **or** $OC(\mu)$

16. **A test's ability to detect a false null hypothesis is called the** *power* **of the test. The power of a test is the probability that it correctly rejects a false null hypothesis, i.e.**

    P(Reject $H_0|H_0$ **is false)=P(Reject** $H_0|H_A$ **is true)=**$1-\beta$

17. **Is it possible to reduce both Type I and Type II errors at the same time?**

    Yes, by increasing the sample size.

18. **Graph**

Case 1: Inferences about μ

List of topics:

1. **Conditions to check: One population, SRS, nearly normal population**

2. If the population standard deviation σ is given, or if σ is unknown but the sample size is larger than 30,

use Z-test: Recall $E(\bar{X}) = \mu, SD(\bar{X}) = \sigma/\sqrt{n}$

When n is large, $SD(\bar{X}) \approx s/\sqrt{n}$

Confidence interval:

1) If σ is given, $\bar{x} \pm z * \frac{\sigma}{\sqrt{n}}$

2) If σ is unknown but n is large, $\bar{x} \pm z * \frac{s}{\sqrt{n}}$

Hypothesis testing: use Z-test

3. **If σ is unknown but the sample size is less than 30, use t-test.**

   $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ **has a t_{n-1}-distribution**

   Confidence interval

   3) If σ is unknown but n is small, Use Table to find t*, then $\bar{x} \pm t * \frac{s}{\sqrt{n}}$

   Hypothesis testing: To find P-value, you can use calculator tcdf(lower, upper, df) returns the $P(l \leq T \leq u)$

Case 2: Two Populations

1. **Recall: If X, Y are independent, then** $E(X - Y) = E(X) - E(Y)$,

   $SD(X - Y) = \sqrt{V(X) + V(Y)}$ **and**

   If 1) $X_1, X_2, ..., X_{n_1}$ **is a random sample from population I with mean μ_1 and standard deviation σ_1;**

   2) $Y_1, Y_2, ..., Y_{n_2}$ **is a random sample from population II with mean μ_2 and standard deviation σ_2;**

   3) X_i**'s and** Y_j**'s are independent.**

   4) $\overline{X} = \frac{1}{n_1}\Sigma X_i,\ \overline{Y} = \frac{1}{n_2}\Sigma Y_i$

   then, $E(\overline{X} - \overline{Y}) = \mu_1 - \mu_2$**,**

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

2. **Problem I: When the populations are near Normal, the independence assumptions are met, we want inferences**

about the two means, then use Z-test if sample sizes are large, use t-test if sample sizes are small.

3. **Standard deviation could still be a problem. How to work around it?**
   - **Use standard error:** $SE(\overline{X} - \overline{Y}) = \sqrt{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}$
   - **If, in addition, we can assume the two groups have the same variances, then it is more accurate to use the pooled standard error:**
     $SE_{pooled}(\overline{X} - \overline{Y}) = s_{pooled}\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}$,
     where $s_{pooled}^2 = \frac{(n_1-1)s_x^2+(n_2-1)s_y^2}{(n_1+n_2-2)}$
4. **Ex: If sample sizes are not large enough, we should use T-test. However, if there is reason to believe that the two groups do not have the same (similar) variances, the degree of freedom is tricky. if the two groups should have the same (similar) variances, the pooled SE isn't easy to memorize. You can use STAT software to solve.**
5. **Problem II: to compare two population proportions. (Men vs. women, treatment group vs placebo group)**
6. **Bernoulli trials: under proper assumptions,**
   $\hat{p_1} - \hat{p_2}$ **has mean** $p_1 - p_2$ **and SD** $\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$
7. **Standard deviation is usually the trouble-maker because it is usually unknown. We shall use various defined standard errors to approximate.**
8. **To find confidence interval for** $p_1 - p_2$**, use "hats" to approximate:**
   $SE(\hat{p_1} - \hat{p_2}) = \sqrt{\frac{\hat{p_1}\hat{q_1}}{n_1} + \frac{\hat{p_2}\hat{q_2}}{n_2}}$
9. **To do hypothesis testing, e.g**
   $H_0 : p_1 - p_2 = 0$
   Use pooled standard error

$SE_{pooled}(\hat{p_1} - \hat{p_2}) = \sqrt{\hat{p}_{pooled}\hat{q}_{pooled}(\frac{1}{n_1} + \frac{1}{n_2})}$,

where $\hat{p}_{pooled} = \frac{n_1\hat{p_1}+n_2\hat{p_2}}{n_1+n_2}$

When n is large enough to use Z-test, the z-score is $Z = \frac{(\hat{p_1}-\hat{p_2})-0}{SE_{pooled}(\hat{p_1}-\hat{p_2})}$. Then according to H_A hypothesis, compute the corresponding P-value.