Stat 101 **Lia Liu**

webpage: www.uic.edu/~yliu

Ch1-3: Categorical data

List of topics:

1. **Statistics**
2. **Data: categorical, numerical**
3. **Survey (Year(Freshman,...), gender, party affiliation, number of siblings, height in inches, shoe size, sleep pattern, average sleeping time in hours per day, average exercise time in minutes per day)**
4. **Terms: variables, frequency table, relative frequency, distribution,**
5. **How to display categorical variable?**
6. **One variable: Bar chart and Pie chart**
7. **Two variable:**
   two-way table (contingency table)
   marginal distribution
   conditional distribution

8. Simpson'd Paradox:

Example: Moe and Jill are two pilots. Here's their on-time record broken down by the time (day or night) they flew:

	Day	Night	Overall
Moe	90/100=90%	10/20 =50%	100/120=83%
Jill	19/20=95%	75/100=75%	94/120=78%

Chapter 4-5 Display Numerical data

List of topics:

1. **5-number summary (median, quartiles, max, min)**
2. **Range, inter-quartile-range, percentiles**
3. **Boxplot, identify outliers by 1.5IQRs (Build a fence:**
   Lower fence=$Q_1 - 1.5IQR$**,**
   Upper fence=$Q_3 + 1.5IQR$**.**
   Any data outside of the fence are outliers.)
4. **Histogram**
5. **Arrange data in order (sort or using stem-and-leaf plot)**
6. **Discribe center of distribution: median,mode,mean**
7. **Describe spread:population vs sample, variance, standard deviation**

Chapter 6 Normal Model

List of topics:

1. **Normal $N(\mu, \sigma)$, Standard Normal $N(0, 1)$**
2. **Empirical rule (68-95-99.7 rule)**
3. **How to use table and calculator to find probability? Left tail, right tail, between two numbers**
4. **Z score (standardized score):** $z = \frac{X-\mu}{\sigma}$, μ **=mean,** σ **=standard deviation**
5. **4 types of calculations:**
   - **For Standard normal N(0,1): given a,b, find** $P(a < Z < b)$ **(TI-83/84: Distr** $\rightarrow$ **2. Normalcdf(a,b)).**
     EX: $P(-1 < Z < 2), P(Z > 2.1), P(Z < -2)$
   - **For Standard normal N(0,1): given left tail, find the z score. (TI-83/84: Distr** $\rightarrow$ **3.invnorm(left tail).**
     Example: Find the z score for top 5%**. invnorm(.95)=1.645.**
     Find the z score for 25 percentile. invnorm(0.25)= -0.6745
   - **For general normal model,** $N(\mu, \sigma)$**: Given a,b, find** $P(a < X < b)$**. Standardize.**
     Ex. If X **has a normal model** $N(\mu = 100, \sigma = 15)$**, Find** $P(70 < X < 85)$
     Standardize X by subtract μ**, then divide by** σ**:**
     $P(70 < X < 105) = P(\frac{70-100}{15} < \frac{X-\mu}{\sigma} < \frac{105-100}{15}) = P(-2 < Z < 0.3333) = 0.6078$
   - **For general normal model,** $N(\mu, \sigma)$**: Inverse normal (give probability, find z-score, then find x)**
     Ex: If IQ test score satisfies a normal model $N(\mu = 100, \sigma = 15)$**, find the cut-off score for top** 10%**.**
     Step 1: Find the z score for top 10%**: z=invnorm(.90)=1.28.**
     Step 2: Use $z = \frac{x-\mu}{\sigma}$ **to solve for** x**:** $1.28 = \frac{x-100}{15}$**,** $x = 100 + 1.28 * 15 = 119$**.**

6. **When we are not sure the population is Normal, check "nearly normal" conditions:**

   check symmetric, unimodal, free of outliers.

Chapter 8 & 9 Linear Regression

List of topics:

1. **Always draw scatter plot! Look for direction, form, strength and outlier.**
2. **Correlation coefficient:** $r = \dfrac{\Sigma z_x z_y}{n-1}$
3. **Residual=data-model:** $e = y - \hat{y}$
4. **Linear regression:** $\hat{y} = a + bx$**, where** $b = r\frac{s_y}{s_x},\ a = \overline{y} - b\overline{x}$
5. **How to identify outliers for 2-variable data?**
   - **High leverage: a data point with x-values far from the mean of the x-values.**
     What can happen? A high leverage point not on the regression line may pull the line towards the point. A high leverage point close to the line may strengthen the relationship by inflating the r and r^2.
   - **Large residual: a data point with y-values far from the cluster of the y-values.**
6. **A data point is *influential* if omitting it gives a very different model (line or r).**
7. **r^2 gives the fraction of the data's variation accounted for by the model. If r^2 is small, choosing a linear model is probably wrong. However even if r^2 is large, it may still be wrong to choose a linear model. A single outlier or data that separate into two groups can make r^2 seem large. On the other hand, an outlier that pulls a roughly linear cloud of data can reduce the r^2 value. Always draw a scatter plot.**

8. What can go wrong?

- **Don't fit a straight line to a nonlinear relationship; always check with scatter plot of the data, and then the residual plot.**
- **Be aware of the effect of outliers, leverage, and influence;**
- **Don't invert regression;**
- **Be careful with extrapolation;**
- **Don't infer that x causes y; beware of lurking variables.**

Chapter 12 Sample Survey

List of topics:

1. **Population: parameters: mean μ, standard deviation σ, correlation coefficient ρ, etc.**
2. **Sample: statistics: sample mean $\overline{x}$, sample standard deviation s_x, correlation coefficient r, etc.**
3. **Census**
4. **SRS (Simple random sample): Each element is equally likely to be selected.**
5. **How to generate a random sample by Calculator? Math $\rightarrow$ Prob $\rightarrow$5:randInt(lower, upper, how many)**
6. **Sample frame: a list of individuals from which a sample is drawn (can be reached).**
7. **Strata: The population is divided into homogeneous groups, called strata.**
8. **Stratified random sampling: EX: If men and women are likely to have different views on an issue, a university has 75% females and 25% males. It may not be a good idea to choose a simple random sample of 100. Instead, first divide the population to 2 stratas: Male and female. Then choose a SRS of 75 from the females, and another SRS from the males.**

9. **Cluster: If you take one typical cluster of tomatoes, it usually contains big, small, raw and ripe ones. A good representative of the whole plant.**
10. **Cluster sampling: First split the population into representative clusters. Then just pick one or a few clusters and perform census.**
11. **Multistage sample: Combination of several sampling methods.**
12. **Systematic sample: e.g. choose every 7th person.**

What can go wrong?

1. **A biased sample is useless.**
2. **Voluntary response bias**
3. **Nonrespondent bias**
4. **Undercoverage bias**
5. **Question wording**

Chapter 13 Experimental Design

List of topics:

1. **Observational study (no treatment)**
2. **Experiment (must have at least one treatment).**
3. **Factors- explanatory variables**
4. **Response variable(s)**
5. **Experimental units-objects**
6. **Treatments(including control group), levels,**
7. **4 Principles: Control, randomize, replicate, block.**
8. **Single blind, double blind**
9. **Placebo effect**
10. **Confounding**

Chapter 14-15 Probability

List of topics:

1. **The Probability of an event is the number of outcomes in the event divided by the total number of possible outcomes.**
   $P(A) = \dfrac{n(A)}{n(total)}$
2. **Def: Probability for a discrete sample space S:**
   P1: $P(A) \geq 0$, **for any event** A **in S;**
   P2: $P(S) = 1$**;**
   P3: If $A_1, A_2, ...$ **is a sequence of mutually exclusive events of S, then** $P(A_1 \cup A_2 \cup ...) = P(A_1) + P(A_2) + ...$
3. **Equally likely model**
4. **Complement:** $P(A) = 1 - P(A^C)$
5. **Inclusion-exclusion principle:** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
6. **DeMorgan's rule:**
   $A^C \cap B^C = (A \cup B)^C$
   $A^C \cup B^C = (A \cap B)^C$
7. **Venn Diagram**
8. **Conditional Probability:** $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$
9. **Product rule:**
   $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$
10. **Independent: Two events are** *independent* **if**
    $P(A \cap B) = P(A)P(B)$
    which means $P(A|B) = P(A)$ **and** $P(B|A) = P(B)$
11. **Tree Diagram**

12. Bayes' Theorem: If $A_1, A_2, ..., A_k$ form a partition, then for any event B,

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{P(B)}$$

In particular, $$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}$$

Chapter 16 Random Variables

List of topics:

1. **Random variable(r.v.): a** *random variable* **is a function from sample space to real numbers.**
2. **Probability distribution function for discrete r.v.(pdf)**
   1) $P(X = k) \geq 0$**;**
   2) $\Sigma_k P(X = k) = 1$**.**
3. **How to denote pdf (probability model)?**
   By table;
   By formula;
   Histogram;
   Relative frequency table.
4. **Expectation (mean, expected value) of a discrete random variable:**
   $\mu = E(X) = \Sigma_{\textbf{all } k} k \cdot P(X = k)$
5. **Expectation of a function of a discrete random variable:**
   $Eu(X) = \Sigma_{\textbf{all } k} u(k) \cdot P(X = (k)$
6. **Expectation is linear:** $E(aX + bY) = aE(X) + bE(Y)$ **for any constants** a, b**, and any random variables** X, Y**.**
7. **Variance of X:**
   $\sigma^2 = V(X) = E(X - \mu)^2$ **(Good for understanding)**
   $\sigma^2 = V(X) = E(X^2) - \mu^2 = (\Sigma_{\textbf{all } k} k^2 \cdot P(X = k)) - \mu^2$ **(Good for computation)**
8. **Variance is not linear.**
   i.e. $V(aX + bY) \neq aV(X) + bV(Y)$
   Actually $V(aX + bY) = a^2V(X) + b^2V(Y) + 2abCov(X, Y)$
   Ex: Find the mean and variance of
   a constant, $X - Y, X + Y, \frac{X_1+X_2+X_3}{3}$ **given independence.**
9. **Standard deviation of** X**:** $\sigma = \sqrt{V(X)}$

Chapter 17 Special Probability Models

List of topics: Binomial Distribution

- **Geometric model**
  1. **Toss a coin until 1st head,** $P(H) = p, P(T) = q, q = 1 - p$**.**
     Let X= number of tosses until the 1st head.
     pdf $P(X = k) = pq^{k-1},\ k = 1, 2, 3, ...$
  2. **Mean** $E(X) = \mu = 1/p$
  3. **Variance** $V(X) = \sigma^2 = q/p^2$
  4. **Standard deviation:** $\sigma = \sqrt{q/p^2}$
- **Binomial Model**
  1. **Binomial expansion:** $(x + y)^n = \Sigma_{k=0}^{n} C(n, k)x^{n-k}y^k = C(n, 0)x^n + C(n, 1)x^{n-1}y + C(n, 2)x^{n-2}y^2 + ... + C(n, r)x^r y^{n-r} + ... + C(n, n - 1)xy^{n-1} + C(n, n)y^n$
  2. **Binomial coefficients:** $C(n, k)$
     1) By Yang's triangle
     2) By combination: $C(n, k) = \dfrac{n!}{k!(n - k)!}$
  3. **Bernoulli trials:**
  4. **Binomial distribution: Toss a coin n times.**
     $P(H) = p, P(T) = q, q = 1 - p$**.**
     X= number of heads.
  5. **p.d.f.** $P(X = k) = C(n, k)p^k q^{n-k},\ k = 0, 1, 2, ..., n$
  6. **Expected value** $E(X) = np$**; or** $\mu = np$
     Variance $V(X) = npq$**; or** $\sigma^2 = npq$
     Standard deviation $\sigma = \sqrt{npq}$
- **Poisson Model**
  1. **Model: Counting number of occurences**
  2. **pdf:** $P(X = k) = \dfrac{e^{-\lambda}\lambda^k}{k!},\ k = 0, 1, 2, ...$
  3. **The mean and variance are both** λ

- **Use Normal Distribution to approximate**

  Recall

  1. **How to compute probability of a Standard Normal rv using Table Z ?**
  2. **How to compute probability of any Normal rv using Table Z? Standardize**
  3. **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%;
  4. **How to use Normal to approximate Binomial distribution? Half unit correction**

Chapter 18 Sampling distributions

List of topics:

1. **a random sample $X_1, X_2, ..., X_n$ of size n from a population with mean μ and standard deviation σ. (independent r.v with the same distribution)**
2. **The sample mean is $\overline{X} = \frac{X_1+X_2+...+X_n}{n}$. What is the expectation, $E(\overline{X})$, and the standard deviation of $\overline{X}$?**
3. **The sample sum is $S_n = X_1 + X_2 + ... + X_n$. What is the expectation and the standard deviation of S_n?**
4. **CLT (Central Limit Theorem) (some call it the Fundamental Theorem of Statistics):**

   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** $n \to \infty$
5. **Conditions for CLT: (1)"independence" (when it is too hard to check, use"Randomization" and "10% condition" instead)**

   (2) Sample size must be large enough. Say $n \geq 30$. When population distribution is roughly symmetric, less n is OK. Binomial can be really skewed when p or q is close to 0. Check $np \geq 10$ and $nq \geq 10$ to make sure n is large enough.
6. **Application to Binomial B(n,p). $X =$ the number of heads out of n tosses. $E(X) = np,\ SD(X) = \sqrt{npq}$**
7. **Application to proportion $\hat{p} = X/n$, $E(\hat{p}) = p,\ SD(\hat{p}) = \sqrt{pq/n}$**

Chapter 19 Confidence Intervals

List of topics:

1. **Confidence interval: estimate $\pm$ Margin of Error**
2. **Confidence Interval for the population proportion p is :**
   $\hat{p} \pm z * SE(\hat{p})$,
   where z depends on the confidence level, $SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$.
3. **If you want to reduce the margin of error by a half, how large should the sample size be?**
4. **Explain the meaning of CI, check conditions of independence(randomization and 10% condition if not independent), sample size.**

Chapter 20 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.**

Chapter 21 More on Hypothesis Testing

List of topics:

1. **Rule of Thumb: If the significance level is not given, reject the null hypothesis if the P-value is less than 5%.**
2. **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".**
3. **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.**
4. **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.
5. **The probability of making a type I error is α.**
6. **The probability of making a type II error is β, which is not easy to calculate.**
7. **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)=$1-\beta$
8. **Is it possible to reduce both Type I and Type II errors at the same time?**

   Yes, by increasing the sample size.
9. **Graph**

Chapter 23 Inferences about μ

List of topics:

1. **Conditions to check: 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 $P(l \leq T \leq u)$

Chapter 22 Two Proportions

List of topics:

1. **Recall: When population is binomial, $E(\hat{p}) = p$ and $V(p) = \frac{pq}{n}$.**
2. **Problem: to compare two population proportions. (Men vs. women, treatment group vs placebo group)**
3. **Bernoulli trials: under proper assumptions,**
   $\hat{p_1} - \hat{p_2}$ has mean $p_1 - p_2$ and SD $\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$
4. **Standard deviation is usually the trouble-maker because it is usually unknown. We shall use various defined standard errors to approximate.**
5. **To find confidence interval for $p_1 - p_2$, use "hats" to approximate:**
   $SE(\hat{p_1} - \hat{p_2}) = \sqrt{\frac{\hat{p_1}\hat{q_1}}{n_1} + \frac{\hat{p_2}\hat{q_2}}{n_2}}$
6. **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.

Chapter 24 Two Means

List of topics:

1. **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.**
2. **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 $s^2_{pooled} = \frac{(n_1-1)s_x^2+(n_2-1)s_y^2}{(n_1+n_2-2)}$
3. **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. (see footnote on Page 619). 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.**