Review

STAT-UB.0001 Statistics for Business Control

Ningshan Zhang

IOMS Department
nzhang@stern.nyu.edu

Aug 7, 2018

* Office Hours: Wed & Fri 10:00 - 11:30 AM, KMC 8-174

Final Exam

- Aug 9, 10:00 - 12:00 AM, Tisch UC19.
- Open book and notes. No cellphone or laptop.
- Bring a calculator (make sure it cake take square root).
- Covers all the topics; focuses on the second half.

Populations vs Samples

"Statistics is using a *sample* to make a statement about a *population*."

- Population: The set of items or individuals that we are interested in studying and drawing conclusions about.
- Sample: A subset of items or individuals from the population.
  - Unbiased sample: every member of the population has an equal chance of being included in the sample.

Descriptive Statistics

Descriptive Statistics: types of statements.

- Center of the distribution: mean, median.
- Spread of the distribution: range, standard deviation.
- Shape of the distribution: histogram, boxplot.

Center of the Distribution: Mean vs Median

- Mean: the average of the observations:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n\right).$$

- Median: the middle value in a *sorted* dataset.
  - When n is odd, take "true" middle value.
  - When n is even, take the average of the two middle values.
- Skewness
  - Positive/right skew: mean - median > 0 , mean is to the right of the median.
  - Negative/left skew: mean - median < 0, mean is to the left of the median.

Spread of the Distribution

- Range:

$$\max(\{x_1, \cdots, x_n\}) - \min(\{x_1, \cdots, x_n\})$$

- Variance (s^2) and standard deviation (s):

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$s = \sqrt{s^2}$$

Probability: Terminology

- Random experiment: the process of observation leading to an outcome that cannot be predicted with certainty.
- Sample point: a possible outcome of an experiment.
- Sample space of experiments: the set of all sample points, denoted by Ω, or S.
- Event: a set of sample points.

Example: flip a coin; roll a 6-sided dice.

Probability

- Given a sample space, $\Omega = \{e_1, e_2, \cdots, e_n\}$. A probability $\mathbb{P}$ is a function with two properties:

$$\mathbb{P}(e_i) \geq 0, \quad \mathbb{P}(e_1) + .. + \mathbb{P}(e_n) = 1.$$

- Probability of an event: If $A = \{e_1, \ldots, e_m\}$, then

$$\mathbb{P}(A) = \mathbb{P}(e_1) + \cdots + \mathbb{P}(e_m).$$

- Interpretations of probability: long-run relative frequency; when an experiment is repeated n times (n is large),

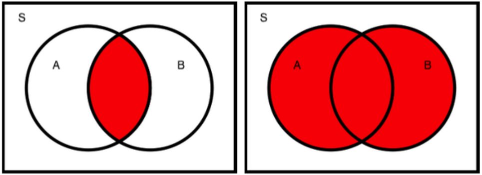
$$\mathbb{P}(A) \approx (\text{no. of times } A \text{ occured})/n.$$

Compound Events: Union and Intersections

A and B are two events.

- Union ($A \cup B$, "A or B"): event A or event B occurs, or both occur.
- Intersection ($A \cap B$, "A and B"): event A and event B both occur.

Figure: Left: $A \cap B$. Right: $A \cup B$



- Mutually exclusive events: A and B cannot occur together, $\mathbb{P}(A \cap B) = 0$.

Conditional Probability and Independence

- Conditional probability $\mathbb{P}(A \mid B)$: the probability of event A, given that event B occurred. It is formally defined as

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

- Statistical independence: A and B are independent events if the occurrence of A does not affect the probability that B occurs:

$$\mathbb{P}(A \mid B) = \mathbb{P}(A) \iff \mathbb{P}(B \mid A) = \mathbb{P}(B)$$

Rules for Computing with Probability

- Additive rule:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B). \qquad \text{(only when A and B are ME)}$$

- Complement rule:

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A).$$

- Multiplicative rule:

$$\mathbb{P}(A \cap B) = \mathbb{P}(B)\mathbb{P}(A \mid B) = \mathbb{P}(A)\mathbb{P}(B \mid A).$$

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \quad \text{(only when A and B are independent)}$$

Bayes' Rule: relates $\mathbb{P}(A \mid B)$ to $\mathbb{P}(B|A)$

Given k mutually exclusive events $B_1, B_2, \ldots, B_k$ such that $\mathbb{P}(B_1) + \mathbb{P}(B_2) + \cdots + \mathbb{P}(B_k) = 1$, then

$$\begin{aligned}\mathbb{P}(B_i \mid A) &= \frac{\mathbb{P}(A \cap B_i)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}{\mathbb{P}(A \mid B_1)\mathbb{P}(B_1) + \cdots + \mathbb{P}(A \mid B_k)\mathbb{P}(B_k)}\end{aligned}$$

- Bayes' rule can be derived from additive rule and multiplicative rule.

Counting Rules

When all sample points are equally likely,

$$\mathbb{P}(A) = \frac{\text{no. of sample points in } A}{\text{no. of sample points in } \Omega}.$$

- Permutations rule: The number of ways to arrange k out of n objects is

$$P(n,k) = n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}.$$

- Combinations rule: Number of ways to pick *unordered* k out of n objects is

$$\begin{aligned} C(n,k) &= \frac{\text{no. of ways to pick \textit{ordered} } k \text{ objects out of } n}{\text{no. of ways to order } k \text{ objects}} \\ &= \frac{P(n,k)}{k!} = \frac{n!}{k!(n-k)!} \end{aligned}$$

Random Variable

- Random Variable: A variable whose value depends uniquely on the outcome of a random experiment.
- Properties of a discrete random variable X:
  - Probability Distribution Function (PDF):

$$p(x) = \mathbb{P}(X = x).$$

  - Expected value/mean/expectation (μ, μ_X):

$$\mathbb{E}(X) = \sum_x x \cdot p(x)$$

  - Variance (σ^2, σ_X^2) and standard deviation (σ, σ_X):

$$\text{var}(X) = \sum_x (x - \mu)^2 p(x), \quad \text{sd}(X) = \sqrt{\text{var}(X)}$$

Continuous Random Variables

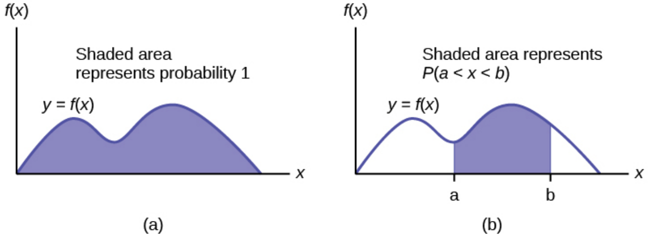
X is a continuous random variable.

- The probability of X taking any *individual value* is 0.

$$\mathbb{P}(X = x) = 0, \text{ for any value of } x.$$

- The probability of X *within a range of values* is defined by probability density function (pdf):

Figure: The pdf and the area under the curve.



Properties of Expected Value

1. (Aaffine transformation) Let a, b be two constants, and let X be a random variable. Then,

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$$

2. (Sum) Let X and Y be two random variables. Then,

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

Applications:

$$\mathbb{E}(-X) = -\mathbb{E}(X), \quad \mathrm{var}(aX) = a^2\mathrm{var}(X)$$

The Binomial Distribution

Binomial experiment:

- It consists of a fixed number n of statistically independent trials;
- each trial has the same probability of success p;
- we want to count the number of successes.

Let X = the number of successes. Then X is a *binomial random variable* that has *binomial distribution*, written as $X \sim B(n, p)$. The PDF, mean and standard deviation are:

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

$$\mathbb{E}(X) = np, \quad \mathrm{var}(X) = np(1-p).$$

The Poisson Distribution

Let X = the number of events that occur in a fixed interval of time, space, etc. Assume that

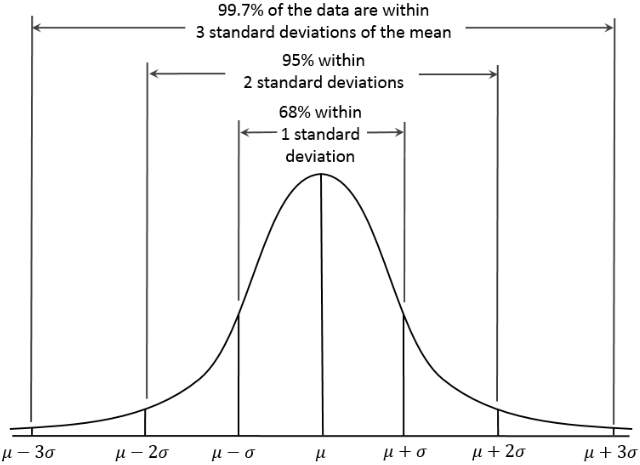
- Events occur with a known constant rate.
- The events occur independently of the time since the last event.

Then X follows a *Poisson distribution*. The PDF, mean and standard deviation are:

$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},$$
$$\mathbb{E}(X) = \text{var}(X) = \lambda.$$

The Normal Distribution

Figure: The pdf of a normal distribution with mean μ and variance σ^2.



The Normal Distribution

- Standard normal distribution Z: a normal distribution with $\mu = 0$ and $\sigma = 1$.
- We use z-tables to find the areas under the curve for Z.
  - Given z_0, look for $\mathbb{P}(Z \leq z_0)$.
  - Given p_0, look for z_0 such that $\mathbb{P}(Z \leq z_0) = p_0$.
- If X is a normal with mean μ and standard deviation σ, then

$$\frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

The Central Limit Theorem (CLT)

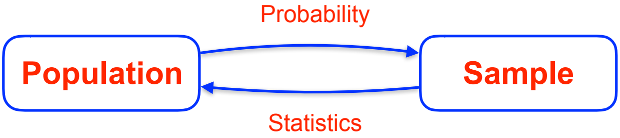
Suppose $X_1, X_2, \ldots, X_n$ are sampled independently from a population with mean μ and standard deviation σ. Let $\bar{X}$ be the sample mean,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then,

- $\mu_{\bar{X}} = \mathbb{E}(\bar{X}) = \mu$,
- $\sigma_{\bar{X}} = \mathrm{sd}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$,
- If n is sufficiently large ($n \geq 30$), then $\bar{X}$ is approximately normal.
- (Not by CLT) If population is normal, then $\bar{X}$ is normal for any $n \geq 0$.

Probability and Statistics



- Probability: CLT, how does $\bar{X}$ relate to the population.
- Statistics: estimation with confidence intervals, hypothesis testing.

Overview of Estimation and Hypothesis Testing

Parameter	Estimate	$\mathbb{E}$(Estimate)	sd(Estimate)
μ	$\bar{X}$	μ	$\sigma/\sqrt{n}$
p	$\widehat{p}$	p	$\sqrt{p(1-p)/n}$
$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$	$\mu_1 - \mu_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

- A $1-\alpha$ CI for Parameter: Estimate $\pm\, (z_{\alpha/2})$ sd(Estimate).
- Hypothesis test H_0: Parameter$=\mu_0$ v.s. H_A: Parameter$\neq \mu_0$,

$$T = \frac{\text{Estimate} - \mu_0}{\text{sd(Estimate)}},$$

and compute p-value from there on.

* Note: use $t_{\alpha/2,n-1}$ instead of $z_{\alpha/2}$ when necessary.

CI for the Mean

	σ known	σ unknown
$n \geq 30$	$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$	$\bar{X} \pm z_{\alpha/2} \frac{S}{\sqrt{n}}$
$n < 30$, pop. is normal	$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$	$\bar{X} \pm t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}$
$n < 30$, pop. isn't normal	N.A.	N.A.

Hypothesis Test for the Mean

- $H_0 : \mu = \mu_0, \quad H_A : \mu \neq \mu_0.$
- Test statistic:
$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}.$$
- p-value: given the observed test staistic t,
$$\text{p-value} = \mathbb{P}(|t_{n-1}| \geq |t|),$$
where t_{n-1} is a t-distribution with df$= n-1$.
- Given significance level α,
  - Reject H_0 if p-value $\leq \alpha$.
  - Equivalently, reject H_0 if observe test statistic t such that
$$t \notin (-t_{\alpha/2,n-1}, t_{\alpha/2,n-1}).$$

CI for the Proportion

- Use $\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$ as an approximation of $\sigma_{\widehat{p}} = \sqrt{\frac{p(1-p)}{n}}$.
- A $1-\alpha$ confidence interval for population proportion p is

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

CI for the Difference of Means

- Use $\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ as an approximation of $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.
- The $1 - \alpha$ confidence interval for difference of population means $\mu_1 - \mu_2$ is

$$(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}.$$

Hypothesis Test for the Difference of Means

- $H_0 : \mu_1 = \mu_2, \quad H_A : \mu_1 \neq \mu_2.$
- Test statistic:
$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$
- p-value: given the observed test staistic t,
$$\text{p-value} = \mathbb{P}(|Z| \geq |t|),$$
where Z is the standard normal random variable.
- Given significance level α,
  - Reject H_0 if p-value $\leq \alpha$.
  - Equivalently, reject H_0 if observe test statistic t such that
$$t \notin (-z_{\alpha/2}, z_{\alpha/2}).$$

When are CI and Htest valid?

The sample must satisfy

1. Observations $X_1, \cdots, X_n$ are drawn randomly and independently from the population.

2. We can reason about the distribution of the estimate:
   - $\bar{X}$: population is normal, or $n \geq 30$.
   - $\widehat{p}$: $np \geq 15$ and $n(1-p) \geq 15$.
   - $\bar{X}_1 - \bar{X}_2$: $n_1 \geq 30$ and $n_2 \geq 30$.

Interpretations of CI and Htest

- Confidence interval: $1 - \alpha$ is the probability, or proportion of the time, that a interval constructed by this procedure would cover true parameter.
- Hypothesis test: α is the probability, or proportion of the time, that a test of this kind would reject H_0 when H_0 is true.
- Caution: there is nothing random about the true parameter or null hypothesis!