STAT 210P

Lecture 0

Sevan Koko Gulesserian

University of California, Irvine

Programming language R:

- Go to `https://cran.r-project.org/`
- At the top of the page, where it says "Download and Install R", download the correct R for your operating system.
- Install the downloaded file.

R Studio

R Studio IDE (integrated development environment):

- Go to `https://posit.co/download/rstudio-desktop/`
- Scroll down to the "All Installers and Tarballs" sections and download the appropriate installer for your operating system.
- R Studio can be used as a graphic interface to R which can help using R, and R Studio is used for R Markdown (creating pdf files in conjugation with R).

5 number summary.

- Say we are given a string of n many numbers called "data".
- In R, we obtain the 5 number summary by doing `summary(x)`.

The output will look like:

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.86600 -0.61900 -0.06351 -0.02148  0.45940  2.55000
```

5 number summary.

- Min. is the minimum of the numbers and Max. is the maximum of the numbers.
- 1st Qu. is the 25-th quartile. 25% of the observations are below this number (which means 75% are above).
- 3rd Qu. is the 75-th quartile. 75% of the observations are below this number (which means 25% are above).
- Median is the 50-th quartile. 50% of the observation are below this number (which means 50% are above).

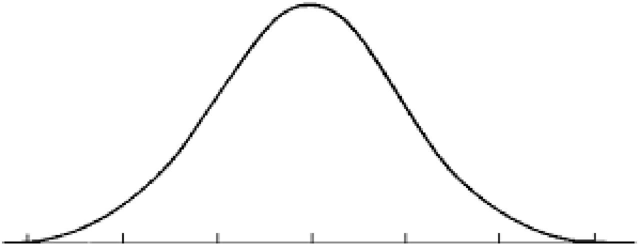
Sample mean and sample variance.

- *Mean* is the average of all the numbers. The population mean (usually unknown) is μ and the sample mean (which we always know) is $\bar{x}$
    - The formula for the sample mean was $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ where x_i are the individual observations.
- *Variance* is the average of the squared deviations of the observations from the mean. The population variance is σ^2 and the sample variance is s^2.
    - The formula for the sample variance was $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
    - Remember that σ was the population *standard deviation* and s was the sample standard deviation.

Histogram.

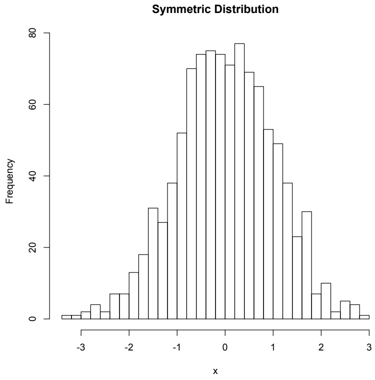
- A histogram plots the data (a single string of numbers) by grouping them into bins (say from 0-4.99, 5-9.99, and so on).
- You can examine a histogram to see if the data is left skewed, right skewed, or symmetric.
- This can be used to assess assumptions of normality.

Remember a Normally distributed random variable will have probability density curve that is similar to what follows:

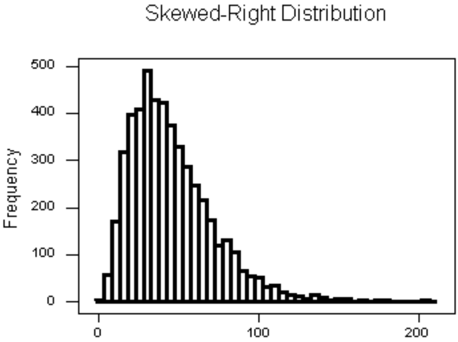


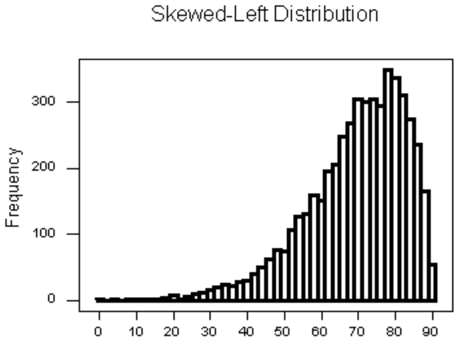
Will compare histograms of the empirical data to the shape of the normal curve.

Brief Statistics Review



Brief Statistics Review




Skewed-Left Distribution
Frequency
300
200
100
0
0
10
20
30
40
50
60
70
80
90

Brief Statistics Review

Review of discrete random variables.

- $f(x)$ is the probability mass function of a random variable X.
- Denote the *support* of X as $\mathbb{S}_X$.
- The *support* of X is the space of values which X has a positive probability of occurring.
- The input is x, which is a specified value of X from its support.
- The output is $P(X = x)$, the probability that X is equal to what we specified, x.
- Thus $f(x) = P(X = x)$.
- Since it is a probability, $0 \leq f(x) \leq 1$.
- To be a valid probability mass function (pmf), need $\sum\limits_{x \in \mathbb{S}_X} f(x) = 1$.

Review of discrete random variables.

- The cumulative distribution function (cdf) of X is not just $P(X = x)$ but $P(X \leq x)$.
- We denote cdf as F(x) (where the pmf is f(x)).
- $F(x) = P(X \leq x) = \sum_{\tilde{x} \leq x} f(\tilde{x}) = \sum_{\tilde{x} \leq x} P(X = \tilde{X})$.
- The cumulative distribution function is the sum of several probability mass functions.
- Note that $F(x) = P(X \leq x) = 1 - P(X > x)$.

Continuous random variable.

- $f(x)$ is now called the *probability density function*.
- $f(x)$ does not represent $P(X = x)$ anymore.
- $P(X = x) = 0$ for all x in $\mathbb{S}_X$.
- This is to say the probability that a continuous random variable is equal to a single fixed number is 0.
- Although $f(x)$ is not a probability when it comes to continuous random variables, in a loose manner we can view it as something proportional to probability (values with higher density values are more likely to occur compared to values with low density values).

Expectation.

- The *expectation* of X is denoted as $\mathrm{E}(X)$.
- For discrete: $\mathrm{E}(X) = \sum_{x \in \mathbb{S}_X} xP(X = x) = \sum_{x \in \mathbb{S}_X} xf(x)$.
- For continuous: $E(X) = \int_{\mathbb{S}_X} x\, f(x) dx$.
- Can be viewed as averaging over all possible X values while weighting each possible value by its probability.
- For ease of notation, we let μ be the population expected value of X. That is to say $\mathrm{E}(X) = \mu$.

Variance,.

- The *variance* of X is defined to be the average squared deviation from the mean/expected value/average (which is E(X)).
- $\text{var}(X)=\text{E}[(X-\text{E}(X))^2]=\text{E}[(X-\mu)^2]$.
- We denote the variance of X as σ^2.
- Since it is the expected value of a squared random variable, $\sigma^2 > 0$.
- $\sigma = \sqrt{\text{var}(X)}$ is known as the *standard deviation* of X.

Probability.

- The probability of an event is a number between 0 and 1.
- If we have two disjoint events, A and B, and no other possible events, then P(A or B)=1 (you have to be either A or B).
- We denote the probability of A occurring given or conditional on B having already occurred as:

$$P(A|B) = \frac{P(A\ and\ B)}{P(B)}$$

- If A and B are independent, then knowing A has occurred has no bearing on the probability that B will occur.
- This is to P(B|A)=P(B) or similarly P(A|B)=P(A).
  - Independence between A and B also means that $P(A\ and\ B) = P(A)P(B)$.

Normal Distribution

Normal/Gaussian distribution.

- Now, we come to what is knows as the *normal or Gaussian* distribution.
- This is a continuous random variable that has support $\mathbb{S}_X = (-\infty, \infty)$.
- Its probability density function is determined by two parameters, μ and σ.

Normal Distribution

- The density is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.
- Support is $\mathbb{S}_X = (-\infty, \infty)$.
- Parameters are μ in $(-\infty, \infty)$ and σ^2 in $(0, \infty)$.
- Note that $f(\mu + a) = f(\mu - a)$ for all $a > 0$ (it is a symmetric function about μ).

Properties of the Normal distribution.

- $\int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1.$
- $E(X) = \int\limits_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \mu.$
- $var(X) = \int\limits_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \sigma^2.$
  - And so the standard deviation is $\sqrt{var(X)} = \sigma$

Normal Distribution

Properties of the Normal distribution.

- There is no closed form solution to the cdf $F(x) = P(X \leq x)$.
- In integral form, this is $\int\limits_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(u-\mu)^2}{2\sigma^2}} du$.
- For all computations, we use approximations.
- In R, all computations are of the form:

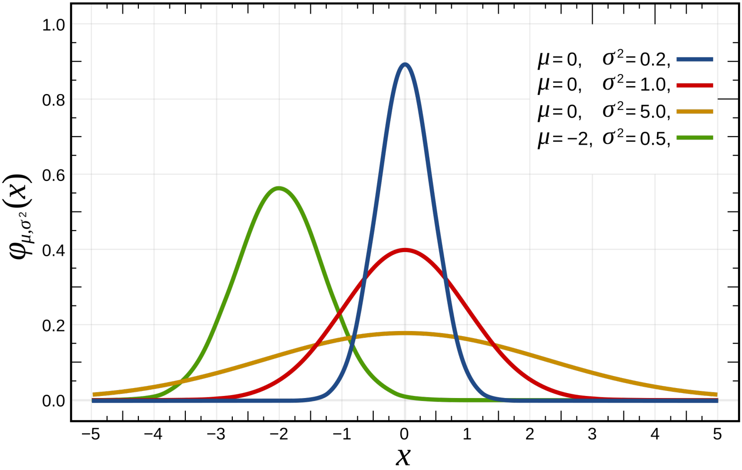
  $P(X \leq x)$=pnorm(X=x, expectation, standard deviation)

  or

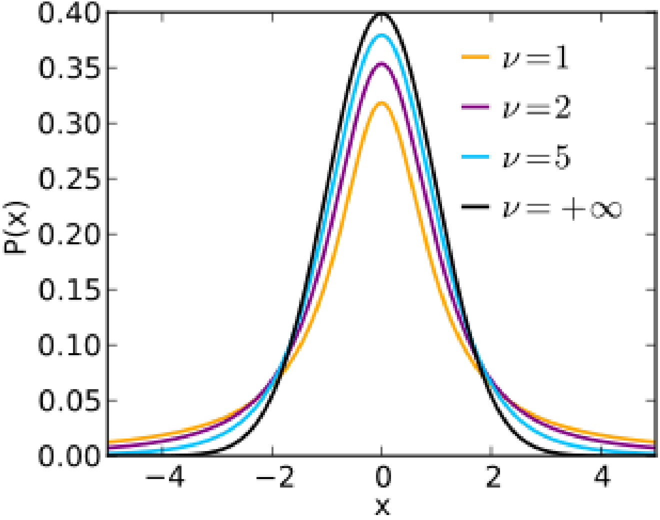
  $P(X \leq x)$=pnorm(X=x, μ, σ)

Normal Distribution

Graph of the Normal distribution density curve.



Graph of the Student t-distribution density curve.



Correlation (linear) coefficient: Say we now have two strings of numbers of equal length, X and Y. Can think of X as being your height and Y being the height of your sibling.

- The linear correlation coefficient, ρ (the population value) or r (the sample value), is a measure of the linear association between X and Y and is a number between -1 and 1.
- Negative number imply a negative association, and positive number implies positive association.
- The closer ρ is to -1 or 1 implies a strong association.

Correlation (linear) coefficient:

- The formula for the population correlation coefficient was $\rho = \frac{cov(X,Y)}{\sigma_x \sigma_y}$, where cov(X,Y) is the covariance between X and Y, and σ_x and σ_y are the standard deviations for X and Y respectively.
- Remember we had the following definitions: $E(X) = \mu_x$, $E(Y) = \mu_y$, $\sigma_x^2 = E[(X - \mu_x)^2]$, $\sigma_y^2 = E[(Y - \mu_y)^2]$ and we have

$$cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)] = E(XY) - E(X)E(Y)$$

- The formula for the sample correlation coefficient is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Confidence Interval

Confidence interval.

- The general form of a $(1-\alpha) * 100\%$ confidence interval for an unknown population parameter is:

$$\hat{\theta} \pm z_{1-\frac{\alpha}{2}} \widehat{SE}(\theta)$$

- Where $\hat{\theta}$ is the sample based estimate of θ, $z_{1-\frac{\alpha}{2}}$ is the multiplier from the asymptotic/approximating distribution of $\hat{\theta}$, and $\widehat{SE}(\theta)$ is the estimated standard error/deviation of θ.

Confidence Interval

- Given a value of α, can find out the value of z.
- Need the area under the curve between -z and z to be 1-α.
- Will need the area above z to be $\frac{\alpha}{2}$ (so area below is $1-\frac{\alpha}{2}$) .
- Can us qnorm in R to get these values (z=qnorm($1-\frac{\alpha}{2}$, 0 ,1)).
- Some useful values are:
  - When $\alpha = 0.05$ then z=1.96 ,$P(-1.96 < Z < 1.96) = 0.95$.
  - When $\alpha = 0.01$ then z=2.576 ,$P(-2.57 < Z < 2.57) = 0.99$.
  - When $\alpha = 0.10$ then z=1.645 ,$P(-1.65 < Z < 1.65) = 0.90$.

Confidence Interval

Confidence interval.

- One sample mean confidence interval (for μ):

  $$\bar{X} \pm t_{1-\frac{\alpha}{2}} s/\sqrt{n}$$

  where the t multiplier is with n-1 degrees of freedom.

- Two sample mean confidence interval (for $\mu_x - \mu_y$):

  $$\bar{X} - \bar{Y} \pm t_{1-\frac{\alpha}{2}} s_p \sqrt{1/n + 1/m}$$

  where the t multiplier is with n-1 degrees of freedom.

Confidence Interval

Confidence interval.

- We will fail to reject (accept) the null hypothesis for all values in the interval.
- And we will reject the null for all values outside the interval.
- Will fail to reject all null hypothesis of the form $H_0 : \theta = \theta_0$ for values of θ_0 in the confidence interval.
- Will reject all null hypothesis of the form $H_0 : \theta = \theta_0$ for values of θ_0 outside the confidence interval.

Confidence Interval

Confidence interval example.

- Take a two sample t-test (with equal variance) confidence interval for $\mu_x - \mu_y$ is (1.5,3.7).
  - We will reject the null hypothesis $H_0 : \mu_x - \mu_y = 0$ since 0 is outside of the interval.
- Take a one sample t-test confidence interval for μ is (11.3, 20.2).
  - We will reject the null hypothesis $H_0 : \mu = 10$ since 10 is outside of the interval.
  - We will fail to reject the null hypothesis $H_0 : \mu = 12$ since 12 is inside the interval.

Confidence Interval

Confidence interval and hypothesis testing relation derivation.

- Let us assume a one sample mean problem, where we are testing the null $H_0 : \mu = \mu_0$ vs. $H_A : \mu \neq \mu_0$.
- Specify a certain level of significance. For ease of notation, let us set $\alpha = 0.05$.
- We will show that we reject all nulls of the form $H_0 : \mu = \mu_0$ at significance level α if and only if μ_0 is not in the confidence interval.

Confidence Interval

Confidence interval and hypothesis testing relation derivation.

- Say we reject based on a critical value, and the critical value here is approximately 1.96 ($t_{1-\frac{\alpha}{2}} \approx 1.96$ for large degrees of freedom and $\alpha = 0.05$).
- Rejecting the null means our computed test statistic t^* is either larger than 1.96 or smaller than -1.96 ($|t^*| > t_{1-\frac{\alpha}{2}}$).
- Say $t^* > t_{1-\frac{\alpha}{2}}$, this means that $\frac{\bar{X}-\mu_0}{s/\sqrt{n}} > 1.96$ or $\frac{\bar{X}-\mu_0}{s/\sqrt{n}} < -1.96$
- This implies that:

  $$\bar{X} - 1.96s/\sqrt{n} > \mu_0 \text{ or } \bar{X} + 1.96s/\sqrt{n} < \mu_0$$

- The previous bullet point implies that μ_0 is not in the confidence interval (first piece says the lower limit of the interval is above μ_0 and second piece says upper limit of interval is below μ_0).

Confidence Interval

Confidence interval and hypothesis testing relation derivation.

- Now say our confidence interval does not contain μ_0.
- Let our interval be (a, b) (a is lower limit and b is upper limit).
- For μ_0 to not be in this interval means $\mu_0 < a$ or $b < \mu_0$.
- Say $b < \mu_0$, this means that $\bar{X} + 1.96s/\sqrt{n} < \mu_0$.
- This implies that $\frac{\bar{X}-\mu_0}{s/\sqrt{n}} < -1.96$ (and $\mu_0 < a$ implies $\frac{\bar{X}-\mu_0}{s/\sqrt{n}} > 1.96$).
- Both of these mean our test statistic is in the critical region ($|t^*| > 1.96$).

Confidence Interval

One sided confidence intervals.

- A one sided confidence interval is of the form $(b, \inf)$ or $(-\inf, a)$ (where a and b are numeric finite values).
- Note again to find the critical value for an alternative of the form $H_A : \theta > \theta_0$ is to find the value from the test statistics distribution (normal, t, chi square,...) such that the are above this value is α (or area below is $1 - \alpha$).
- To find the critical value for an alternative of the form $H_A : \theta < \theta_0$ is to find the value from the test statistics distribution (normal, t, chi square,...) such that the are below this value is α.

Confidence Interval

One sided confidence intervals.

- A two sided confidence interval at confidence level $1 - \alpha$ had the general form:

$$(\hat{\theta} - z_{1-\frac{\alpha}{2}}\hat{SE}(\theta), \hat{\theta} + z_{1-\frac{\alpha}{2}}\hat{SE}(\theta))$$

- An **upper** one sided confidence interval is of the form:

$$(\hat{\theta} - z_{1-\alpha}\hat{SE}(\theta), \infty)$$

- An **lower** one sided confidence interval is of the form:

$$(-\infty, \hat{\theta} + z_{1-\alpha}\hat{SE}(\theta))$$

One sided confidence intervals.

- The primary difference between one sided and two sided confidence intervals is the multiplier (and the fact that one of the limits is infinite).
- For confidence level $1-\alpha$, our multiplier value needs area below to be $1-\alpha$ (instead of $1-\alpha/2$ like for 2-sided intervals).
- We can now use these one sided intervals to test one sided alternative hypothesis.

Brief Statistics Review

A few of the statistics topics we will need to know for this class.

- Probability distribution function (density for continuous random variable and mass for discrete random variable).
- Expectation and variance of a random variable.
- Sample mean, sample variance, and sampling distribution.
- Covariance and correlation between two random variables.
- The normal (Gaussian) distribution (and we will review the Student's t-distribution).
- Hypothesis testing (null and alternative hypothesis, test statistic, and p-value).